Modeling Spatial Correlation of Transcripts With Application to Developing Pancreas

Ruishan Liu[1*], Marco Mignardi[236*†], Robert Jones[2], Martin Enge[2], Seung K Kim[4], Stephen R Quake[36], James Zou[56†]

[1] *Department of Electrical Engineering, Stanford University, 450 Serra Mall, Stanford CA 94305*

[2] *Department of Bioengineering and Applied Physics, Stanford University, 450 Serra Mall, Stanford CA 94305*

[3] *Department of Information Technology, Uppsala University, Lgerhyddsvgen 2, Uppsala, Sweden SE-751 05*

[4] *Department of Developmental Biology, Stanford University, 279 Campus Drive, Stanford CA 94305, USA*

[5] *Department of Biomedical Data Science, Stanford University, 450 Serra Mall, Stanford CA 94305*

[6] *Chan-Zuckerberg Biohub, 499 Illinois St., San Francisco CA 94158*

*These authors contributed equally to this work.

[†] *Corresponding authors, please send correspondence to: jamesz@stanford.edu and marco.mignardi@czbiohub.org*

Abstract

Recently high-throughput image-based transcriptomic methods were developed and enabled researchers to spatially resolve gene expression variation at the molecular level for the first time. In this work, we develop a general analysis tool to quantitatively study the spatial cor-

relations of gene expression in fixed tissue sections. As an illustration, we analyze the spatial distribution of single mRNA molecules measured by in situ sequencing on human fetal pancreas at three developmental time points 80, 87 and 117 days post-fertilization. We develop a density profile-based method to capture the spatial relationship between gene expression and other morphological features of the tissue sample such as position of nuclei and endocrine cells of the pancreas. In addition, we build a statistical model to characterize correlations in the spatial distribution of the expression level among different genes. This model enables us to infer the inhibitory and clustering effects throughout different time points. Our analysis framework is applicable to a wide variety of spatially-resolved transcriptomic data to derive biological insights.

Introduction

The spatial heterogeneity of gene expression has attracted much attention in disease, medicine and developmental studies. Understanding transcriptional heterogeneity provides critical information to interpret biological processes and to develop clinical therapies [1,2]. For decades, immunohisto-chemistry has been the workhorse for studying the protein expression in tissue samples. Although robust, this method is limited to the study of few proteins at time and it is sometimes hampered by the poor performance of the antibody used. On the contrary, transcriptome measurements are now performed genome-wide either with bulk measurements of the tissue of interest or by analysis of single cells extracted from the tissue. The spatial resolution is lost in both approaches [3,4].

Recently, in situ sequencing and other fluorescent in situ hybridization (FISH) - based meth-ods were developed, which enabled high-resolution spatially resolved transcriptomic studies [5–9]. These technologies image and detect RNA molecules directly in tissue samples, thus maintain-ing the spatial information with high resolution. In contrast to the rapid growth of in situ tran-scriptomic technologies, the computational analysis on the spatial transcriptomic data is still in its infancy. Most studies are carried out in a non-quantitative manner or only provide prelimi-nary statistics [10,11]. Recent methods such as SpatialDE aim to identify individual genes that are spatially varying but do not model gene-gene spatial correlations [12]. Many important spatial char-acteristics remain unexplored, and the poor quantification becomes a severe problem, especially when comparison across different time points is required like in developmental studies. Therefore computational methods to explore these novel datasets are needed.

We develop a general analysis tool to explore and quantitatively study the spatial distribution of gene expression data generated by in situ transcriptomic methods. We demonstrate our approach by exploring spatial transcriptomic data generated by in situ RNA sequencing of human fetal pancreas tissues of different ages 80, 87 and 117 days post-fertilization. A density profile-based method is developed to capture the relation between gene expression and other biological targets such as cell nuclei and forming pancreatic islets of the Langerhans. A statistical model is built to characterize the spatial interactions among the expression of different genes. This new tool allows us to model and measure inhibition or clustering effects between transcripts expressed by different cells in the tissues. As a broadly new perspective in development studies, we show that our method can be used as an exploratory tool to identify spatial gene interactions of potential importance in the development of the pancreas. Our tool is publicly available at `https://github.com/RuishanLiu/Gene-Spatial`.

In Situ RNA Sequencing

In situ techniques enable us to spatially resolve gene expression by performing molecular reactions directly in fixed cells and tissue sections [13]. The techniques achieve high multiplexing by two main strategies: combinatorial decoding or sequencing-based readout. Combinatorial decoding methods, typically exploited by fluorescent in situ hybridization (FISH) assays, use fluorescently labeled probes in multiple combinations to distinguish a large number of different targets, each one corresponding to a specific color combination in a predetermined color-coding scheme [14,15]. Sequencing-based readouts for in situ assays build on biochemical methods developed for parallel

DNA sequencing in next-generation sequencing (NGS) platforms and apply them to a molecular substrate that is generated directly in fixed tissue [5,6].

The gene expression data analyzed in this work are generated by a combination of these methods. Specifically, single RNA molecules are amplified as previously described by Ke et al. (In situ sequencing) [5] using a gene-tiling approach. Each gene is targeted by 1 to 13 different cDNA primers which hybridize at different positions along the length of the mRNA. This increases the probability to successfully reverse transcribe the gene. Each primer is coupled with a gene-specific barcoded padlock probe which is ligated to the cDNA and subsequently amplified via rolling-circle amplification (RCA). The molecular barcodes associated with each transcript are then decoded by sequential hybridization of fluorescence probes following a combinatorial decoding scheme. The protocol used to stain the tissues is detailed in the Supplementary Note 1 along with the list of targeted genes and the probe sequences.

Every round of hybridization is carried out using four oligonucleotide probes, each one labeled with a distinct fluorophore, which are hybridized to the amplified cDNA molecules directly on a section of pancreatic tissue. The total barcoding space results in $4^3 = 64$ possible combinations. Here we assign 25 combinations to transcripts from 25 different genes, and leave the remaining 39 combinations as negative controls to assess sequencing quality. The targeted genes list comprises a number of marker genes for endocrine cells (alpha, beta and delta cells), transcription factors implicated in differentiation of the endocrine cells and genes expressed in mesenchymal cells at different levels during pancreas development. The data are collected in samples from three

developmental ages 80, 87 and 117 days post-fertilization.

The RNA molecules of an entire tissue section undergo three rounds of staining and imaging as diffraction-limited spots in its native cellular context together with a nuclear staining (DAPI). The collected images are processed as described previously [5,16] and a detailed description of the image processing can be found in Supplementary Note 1. The intensity values are extracted from each individual diffraction-limited signal, where the fluorescent probes appear as bright round spots. The raw data quality is potentially affected by the influence of neighboring fluorescence, misalignment between the three rounds of imaging and camera noise. For example, 7% of detected RNA molecules (11,611 out of 159,716) are labeled by the 39 negative control combinations at day 87. To carry out quality control, we define a quality metric as the averaged confidence of fluorescence and filter out all the detected transcripts with a quality lower than 55%. After the processing, 32% of the data are discarded and as low as 2% of the rest (2,647 out of 108,430) have meaningless labels. This is in accordance with the accuracy of the method as previously described [5].

For every image, the position (x, y coordinates) of each segmented nuclei and detected transcript as well as the transcript identity are recorded and can be plotted like in Figure 1a where the spatial distribution of three mRNA transcripts somatostatin (SST), glucagon (GLUC) and insulin (INS) is shown in 2D coordinates (x, y). At day 117, for example, 159,716 RNA molecules for the 25 types of genes are detected. The slice of pancreas has 50,147 cells in total, and the nucleis positions are illustrated in Figure 1b. We first focus our computational analysis on data from day 117,

since that represents the highest quality data. Then in the *Temporal Analysis* Section, we discuss how we integrate data from earlier time points to model temporal differences in spatial expression.

Computational Analysis

Identification of Pancreatic Islets The pancreas is composed of a hormone-producing compartment (endocrine pancreas) and a digestive enzyme-producing one (exocrine pancreas). In the developed organ, the endocrine compartment is organized in discrete units, known as islets of Langerhans. These are clusters of hormone- producing cells mostly alpha, beta and delta cells which produce respectively Glucagone (GLUC), Insulin (INS) and Somatostatin (SST). Endocrine and exocrine areas have different physiological functions and cell type composition, thus the study of spatial properties requires identification of the morphological context.

Here, endocrine islets in the process of formation are identified using a clustering algorithm that we developed (Algorithm 1 in Supplementary Note 3). SST, GLUC and INS transcripts are used as marker genes for identification of endocrine cells and pancreatic islets. For the convenience of computation thereafter, all the endocrine islets are assumed to be circular. Real boundaries for islets could be approximated by circles. As shown in Figure 1a, the algorithm is able to identify large islets as well as single cell exocrine regions. The distribution of identified islets size is provided in Supplementary Figure 2. The wide variation in islets diameter has been reported in early as well as more recent studies. However, in the fetal samples the maximum diameter size of islets is smaller than previously reported in the adult pancreas (300 μm) [17,18].

Density Profile-based Analysis To capture the relation between transcripts and other morphological features of the tissue such as nucleis position or developing pancreatic islets, we carried out a density profile-based analysis. The density profiles are calculated based on kernel density estimation with linear combination correction [19]. The difference between two density profiles is characterized by KL divergence. See Supplementary Note 2 for more details.

First, we focused on the spatial relationship between endocrine islets and the transcripts which are outside the islets, in order to identify genes whose expression resulted enriched in proximity of the forming islets. These genes may be directly involved in the differentiation of endocrine cells or be constitutively expressed in the cells nearby. Only the most abundant genes which have at least 100 counts are examined (17 out of 25 genes). At day 117, these are genes which contribute at least 0.1% of the total reads. As an example, the density profile of some transcripts with respect to their distance to the closest islets boundary on day 117 is illustrated in Figure 2a. For each gene pair, the KL divergence of density profiles indicates the difference between the spatial distributions of two genes outside endocrine islets and is plotted in Figure 2b. The larger the difference is, the more distinct the two density profiles are.

Based on the KL divergence of density profiles we identify two groups of genes with distinct density distribution profile from each other. The two groups are highlighted in Figure 2b, where MUC6, EPCAM, GLUC, PROM1, SST, INS and PDX1 form Group 1, marked in bold, and the rest of genes form Group 2. Within Group 1, INS, GLUC and SST are markers for endocrine cells and are expected to be found within or in proximity of pancreatic islets. PDX1 is a transcriptional

activator of several genes, including insulin and somatostatin, and is involved in the early development of the pancreas, which plays a major role in glucose-dependent regulation of insulin gene expression [20]. EPCAM is an antigen expressed in epithelial cells and in stem cells [21], and PROM1 (CD133) is a surface antigen found in progenitor and stem cells in the mouse and human pancreas [22,23]. Their distribution profile clustered together with endocrine cell markers which may indicate a potential role for these genes in the differentiation of progenitor cells into endocrine cells. The MUC6 gene transcribes a glycoprotein belonging to the mucin family, a class of protein which are found in many epithelial tissues. Increasing expression of MUC6 has been observed during development of several human organs including pancreas, but its role has not been well defined yet [24,25].

Similarly our spatial analysis of gene expression could be carried out on other tissue features such as the nuclei. In this case, the density profile captures how likely it is to find a transcript as we move further away from the cell nucleus. As an example, the density profiles of some transcripts with respect to the closest nucleus at day 117 are plotted in Figure 3. However, because automatic nuclei and cell segmentation is particularly difficult in our data and in in situ methods in general, the retrieved nuclei locations may not be accurate.

Temporal Analysis. We then asked whether the transcriptional density profile observed for sample aged 117 days differs from profiles of samples at earlier time points. A difference could be indicative of transient gene expression in the vicinity of the endocrine cells and thus identify genes involved in development of specific cell types at specific time points.

We found that the distinction between the two groups of transcripts identified in sample 117 day shows a temporal trend, becoming larger at later time points — distinction is small at day 80 in Figure 4a, moderate at day 87 in Figure 4b, and most obvious at day 117 in Figure 2b. The density profiles of marker genes which demarcate forming islets cluster more and more during development and the identified groups of genes separate markedly from each other. Thus as the tissue structures (pancreatic islets) become more and more evident with time so does their gene expression profile distribution.

In addition our analysis can reveal temporal changes in expression distribution for individual genes, which could also be of potential biological interest. To do so we rank the averaged difference between one gene and the genes in one of the two groups. For instance, MUC6 is found to become closer to Group 1 during development, as shown in Figure 4c. At day 80, MUC6 is farthest to Group 1 among 17 genes, but is the fifth closest to Group 1 at day 117. This suggests that MUC6 might play a particularly dynamic role in islet development. PROM1 is found to become more distinct to Group 2 across the time, as depicted in Figure 4d.

Statistical Model for Spatial Correlations To characterize the spatial distribution of the expression level among different genes, we carried out the analysis based on a statistical model. Within the range of imaged tissue, the spatial transcriptome is characterized by the likelihood ratio $l(\mathcal{Z})$, where $\mathcal{Z} = \{(x_1, y_1), ..., (x_n, y_n)\}$ is the set of transcripts positions. Here we use the multitype Strauss process model [26,27]

$$l(\mathcal{Z}) = \alpha \prod_i \prod_j \beta_i^{n_i} \gamma_{ij}^{s_{ij}}, \tag{1}$$

where α is a normalizing constant, β_i indicates the intensity of type i transcripts, γ_{ij} denotes the spatial correlations between type i and type j transcripts, n_i is the number of type i transcripts in $\mathcal{Z}$ and s_{ij} is the number of type j transcripts in the neighbor of type i within radius r. The correlations are fully described by γ_{ij}. The case $0 < \gamma_{ij} < 1$ indicates an inhibition effect between the expression of type i and type j genes, and $\gamma_{ij} > 1$ represents a clustering effect. If $\gamma_{ij} = 1$ for all i and j, Eq. (1) gives a Poisson process with intensity β_i for type i transcripts. To increase the resolution of our analysis we applied this model within each cluster of endocrine cells to test for clustering or inhibition effects among these cells. To capture the short-range interactions, the radius is set to be $20 \ \mu m$, twice as the averaged nuclei spacing likely describing interactions between neighboring cells or genes co-expressed in the same cell. Only genes detected at a threshold level of 500 transcripts within endocrine islets are analyzed and the results are summarized in Table 1.

Most spatial correlations γ_{ij} among genes within endocrine islets are fitted to be close to 1, i.e., showing almost no correlation consistently over the three time points. Typical results are illustrated in Table 2. One plausible explanation for the observed lack of correlation is that the selected genes are distinctive of different cell types. Within the forming pancreatic islets at this developmental stage, there seem to be no evident clustering effect between distinct cells types bringing them physically close to each other.

For three pairs of genes a positive spatial correlation was measured at all three time points: EPCAM $\leftrightarrow$ PROM1, MUC6 $\leftrightarrow$ EPCAM and MUC6 $\leftrightarrow$ PROM1. The values are listed in Table 2. As described above, both EPCAM and PROM1 (CD133) are markers of stemness and thus it

11

is not surprising to see their expressions highly correlated spatially, perhaps on the same cell. The positive spatial correlation observed between MUC6 and stem cell markers is certainly interesting because it may indicate a potential role of MUC6 in the differentiation of precursor cells into endocrine cells.

Discussion

Most in situ transcriptomic studies have so far focused on identification and localization of specific cell types in different organs, mapping data obtained by single-cell RNA sequencing back to tissue sections. Fewer studies have focused at identifying the relations in gene expression between cell types or to other structural and morphological features of the tissue [12, 28, 29]. We describe a general analysis tool for spatial correlations of gene expression and carry out temporal study of in situ sequencing data on human fetal pancreas at three developmental ages. We increase the efficiency of the method by probing multiple sites on each transcript and adopting a combinatorial hybridization readout.

A density profile-based method is proposed to study the distribution of transcripts in relation to tissue structures and a statistical model is built to study the spatial correlation between transcripts. The difference between the profiles of each transcript allows us to identify two groups of genes. Notably, we are able to analyze the profiles at different time points and observe how clusters of genes markedly separates from each other. Analyzing samples at three time points, we are able to capture the temporal distribution of single genes within the clusters. We show that MUC6 distri-

bution profile becomes more similar to the group of genes containing endocrine markers and this may indicate a previously unknown role of this gene in the development of pancreatic endocrine cells. The role of mucins genes in the fetal development of several human organs is already known [30]. Also, MUC6 expression has been identified as an early event in certain pancreatic cancers [31,32]. Our spatial analysis shows that MUC6 distribution positively correlates with other stemness genes and its gene expression clusters with forming endocrine islets following a temporal trend. Altogether these observations identify MUC6 as a candidate marker gene of endocrine differentiation. Notably, other genes of the mucins gene family are present in our panel, but none show strong spatial correlation with endocrine cell or stemness markers. Our density profile-based method is a powerful tool to identify genes of interest at a whole-tissue level.

We show that we can increase the resolution of the spatial analysis by applying our statistical model to genes expressed within clusters of endocrine cells. We find that most gene expressions within identified clusters of endocrine cells are not correlated with each other at the examined time points. Among the pairs of genes with strongest correlations we find epithelial and stemness-related markers EPCAM and PROM1, and MUC6, reinforcing the hypothesis of a role of this gene in cellular differentiation. Because in our analysis we specifically looked for short-range interactions (20 μm) it is possible that these genes are co-expressed or expressed from a niche of progenitor cells. On the contrary, hormones secreting cells identified by expression of GLUC, INS or SST show no correlation with each other at this distance, as expected.

In conclusion, we present a novel method to analyze spatially-resolved transcriptomic dataset

which is widely applicable to different technologies and applications. We describe a novel way to explore gene expression data which can be now produced in high throughput by a number of imaged-based techniques. For instance, we demonstrate our method on in situ sequencing data, but the same analysis is applicable to other FISH-based assays. Developmental biology is an ideal application for spatially and temporal- resolved transcriptomic analysis and we demonstrate that our tool can be used to explore and identify potentially novel gene expression patterns and temporal changes.

Code Availability We have released a Python implementation of our profile-based method and statistical model analysis on GitHub (`https://github.com/RuishanLiu/Gene-Spatial`). Most figures in this paper can be reproduced with the codes and datasets in the GitHub repository.

1. Janiszewska, M. *et al.* In situ single-cell analysis identifies heterogeneity for pik3ca mutation and her2 amplification in her2-positive breast cancer. *Nature genetics* **47**, 1212 (2015).

2. Grundberg, I. *et al.* In situ mutation detection and visualization of intratumor heterogeneity for cancer research and diagnostics. *Oncotarget* **4**, 2407 (2013).

3. Huang, S. Non-genetic heterogeneity of cells in development: more than just noise. *Development* **136**, 3853–3862 (2009).

4. OHuallachain, M., Karczewski, K. J., Weissman, S. M., Urban, A. E. & Snyder, M. P. Extensive genetic variation in somatic human tissues. *Proceedings of the National Academy of Sciences* **109**, 18018–18023 (2012).

5. Ke, R. *et al.* In situ sequencing for rna analysis in preserved tissue and cells. *Nature methods* **10**, 857 (2013).

6. Lee, J. H. *et al.* Highly multiplexed subcellular rna sequencing in situ. *Science* **343**, 1360–1363 (2014).

7. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed rna profiling in single cells. *Science* **348**, aaa6090 (2015).

8. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).

9. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).

10. La Manno, G. *et al.* Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).

11. Mignardi, M., Ishaq, O., Qian, X. & Wählby, C. Bridging histology and bioinformaticscomputational analysis of spatially resolved transcriptomics. *Proceedings of the IEEE* **105**, 530–541 (2017).

12. Svensson, V., Teichmann, S. A. & Stegle, O. Spatialde: identification of spatially variable genes. *Nature methods* **15**, 343 (2018).

13. Crosetto, N., Bienko, M. & Van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics* **16**, 57 (2015).

14. Valm, A. M. *et al.* Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proceedings of the National Academy of Sciences* **108**, 4152–4157 (2011).

15. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature methods* **9**, 743 (2012).

16. Darmanis, S. *et al.* Single-cell rna-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell reports* **21**, 1399–1410 (2017).

17. Hellman, B. Actual distribution of the number and volume of the islets of langerhans in different size classes in non-diabetic humans of varying ages. *Nature* **184**, 1498 (1959).

18. Bosco, D. *et al.* Unique arrangement of α-and β-cells in human islets of langerhans. *Diabetes* **59**, 1202–1210 (2010).

19. Jones, M. C. Simple boundary correction for kernel density estimation. *Statistics and Computing* **3**, 135–146 (1993).

20. Zhu, Y., Liu, Q., Zhou, Z. & Ikeda, Y. Pdx1, neurogenin-3, and mafa: critical transcription regulators for beta cell development and regeneration. *Stem cell research & therapy* **8**, 240 (2017).

21. Ng, V. Y., Ang, S. N., Chan, J. X. & Choo, A. B. Characterization of epithelial cell adhesion molecule as a surface marker on undifferentiated human embryonic stem cells. *Stem cells* **28**, 29–35 (2010).

22. Sugiyama, T., Rodriguez, R. T., McLean, G. W. & Kim, S. K. Conserved markers of fetal pancreatic epithelium permit prospective isolation of islet progenitor cells by facs. *Proceedings of the National Academy of Sciences* **104**, 175–180 (2007).

23. Hori, Y. Prominin-1 (cd133) reveals new faces of pancreatic progenitor cells and cancer stem cells: current knowledge and therapeutic perspectives. In *Prominin-1 (CD133): New Insights on Stem & Cancer Stem Cell Biology*, 185–196 (Springer, 2013).

24. Bartman, A. E. *et al.* The muc6 secretory mucin gene is expressed in a wide variety of epithelial tissues. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **186**, 398–405 (1998).

25. Reid, C. J. & Harris, A. Expression of the muc 6 mucin gene in development of the human kidney and male genital ducts. *Journal of Histochemistry & Cytochemistry* **47**, 817–821 (1999).

26. Diggle, P. J. *Statistical analysis of spatial and spatio-temporal point patterns* (Chapman and Hall/CRC, 2013).

27. Baddeley, A., Rubak, E. & Turner, R. *Spatial point patterns: methodology and applications with R* (CRC Press, 2015).

28. Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nature methods* (2018).

29. Schapiro, D. *et al.* histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods* **14**, 873 (2017).

30. Su, L. *et al.* Expression of mucins in the human fetal and neonatal stomach. *Acta Histochemica et Cytochemica* **37**, 163–172 (2004).

31. Nagata, K. *et al.* Mucin expression profile in pancreatic cancer and the precursor lesions. *Journal of Hepato-Biliary-Pancreatic Sciences* **14**, 243–254 (2007).

32. Ringel, J. & Löhr, M. The muc gene family: their role in diagnosis and early detection of pancreatic cancer. *Molecular cancer* **2**, 9 (2003).

Acknowledgements

Author Contributions

J.Z., M.M and S.R.Q. supervised the research. J.Z., S.R.Q and M.E. conceived the study. R.L. designed the analysis tool, implemented the algorithm and carried out the computational study. M.M. performed experiments and image analysis. R.J. contributed to the interpretation of the results. S.K.K. procured the samples. R.L., M.M. and J.Z. wrote the manuscript with input from all the authors. All of the authors reviewed the manuscript.

Competing Interests

The authors declare no competing financial or non-financial Interests.
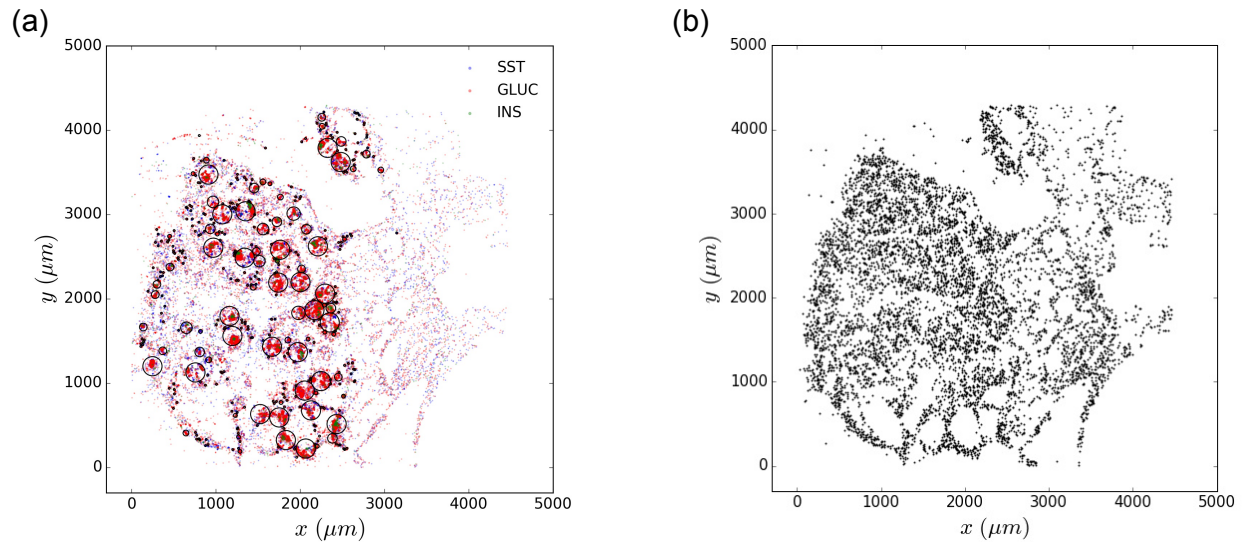
Figure 1: **In situ sequencing.** The sample is from fetal pancreas at age 117 days post fertilization. (a) Detected SST, GLUC and INS transcripts are plotted on xy coordinates. Computationally identified pancreatic islets are identified by black circles. (b) Identified and segmented nuclei are plotted on xy coordinates.
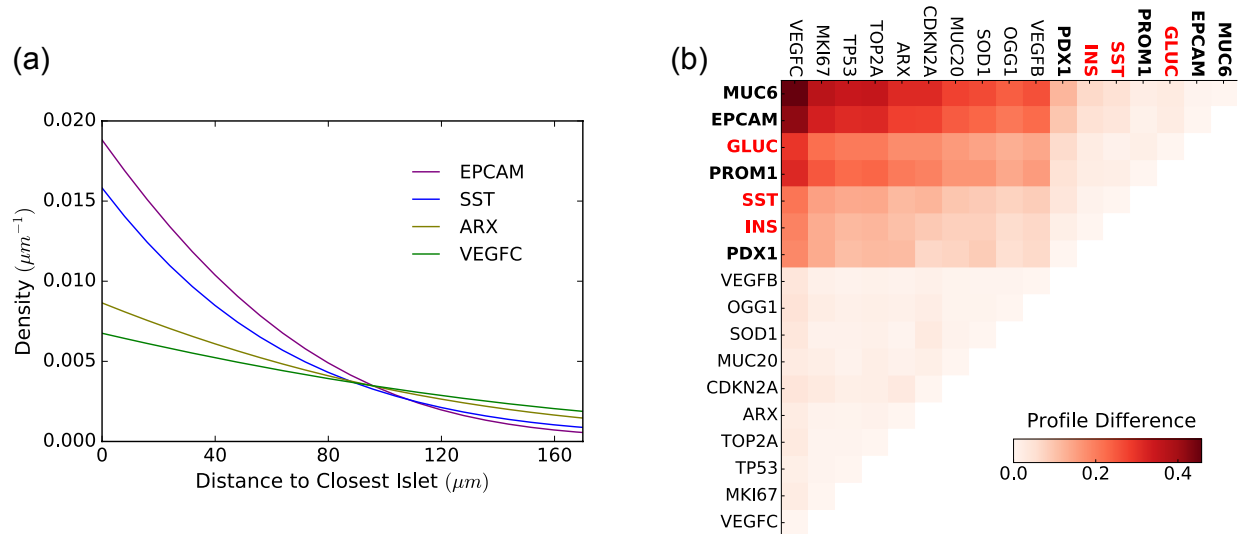
Figure 2: **Islets-related density profile.** Here the sample is collected at age 117 days post fertilization. (a) An example of two different density profiles for four selected genes in respect to pancreatic islets. EPCAM and SST show a higher density closer to islets compared to ARX and VEGFC. (b) The difference between the density profiles is calculated and plotted as heatmap. Two groups of genes can be identified. In bold are the genes belonging to group one. In red are the genes used to identify the islets and therefore expected to be found closer to them.
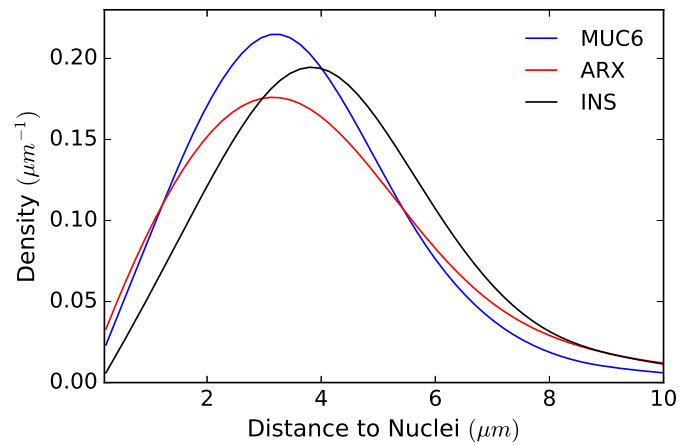
Figure 3: **Nuclei-related density profile.** An example of three density profiles for three genes in respect to their closer nuclei. Genes are assigned to the closer nucleus identified by segmentation of the DAPI staining images. Here the sample is collected at age 117 days post fertilization.
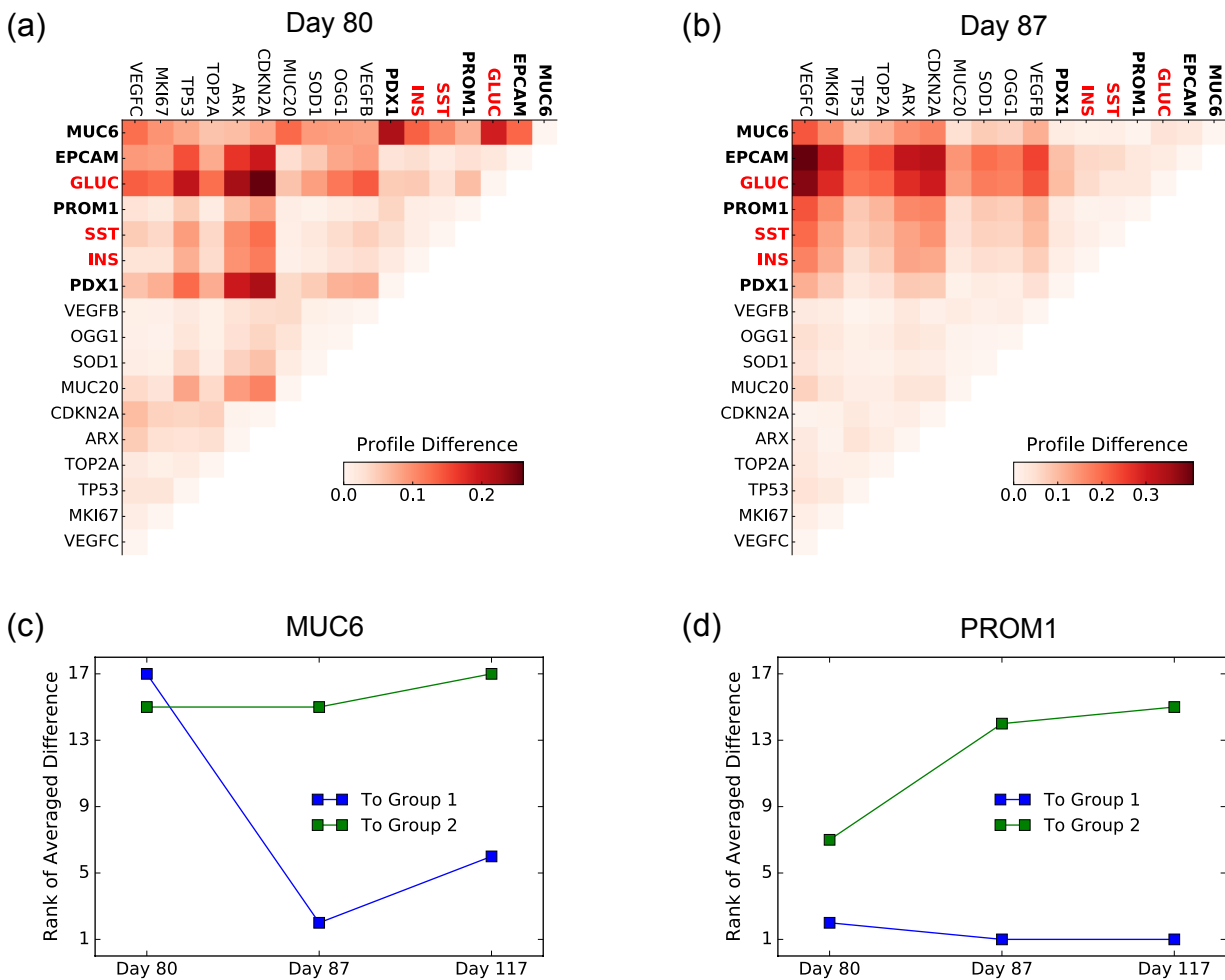
Figure 4: **Islets-related temporal analysis of density profiles.** (a-b) The difference between the density profiles for samples age 80 and 87 days after fertilization is calculated and plotted as heatmap. The two groups of genes identified in sample age 117 days are still evident but to a lesser extent. In bold are the genes belonging to group one. In red are the genes used to identify the islets and therefore expected to be found closer to them. The rank of average difference from the two groups can be plotted for each single gene. Here the difference at the three time points is shown for (c) MUC6 and for (d) PROM1.

Table 1: Number of transcripts inside endocrine islets for 7 genes at age 117 days post fertilization.

	GLUC	**SST**	**INS**	MUC6	EPCAM	PROM1	ARX
Total Number in Islets	37171	11611	1904	4665	1056	774	697

Table 2: Spatial correlation γ_{ij} $(mean \pm std)$ at age 80, 87 and 117 days after fertilization.

Correlation Intensity		Day 80	Day 87	Day 117
Typical	**SST** $\leftrightarrow$ **INS**	1.00 ± 0.02	1.00 ± 0.01	1.00 ± 0.01
	INS $\leftrightarrow$ MUC6	1.00 ± 0.04	1.00 ± 0.03	0.94 ± 0.03
	INS $\leftrightarrow$ ARX	1.00 ± 0.16	1.00 ± 0.04	0.89 ± 0.05
	ARX $\leftrightarrow$ MUC6	1.00 ± 0.03	1.07 ± 0.03	1.00 ± 0.04
Strongest	EPCAM $\leftrightarrow$ PROM1	1.26 ± 0.08	1.26 ± 0.07	1.33 ± 0.09
	MUC6 $\leftrightarrow$ EPCAM	1.15 ± 0.03	1.17 ± 0.08	1.12 ± 0.02
	MUC6 $\leftrightarrow$ PROM1	1.09 ± 0.02	1.13 ± 0.09	1.19 ± 0.04