Statistical Transfer Learning with Generative Encoding for Spatial Transcriptomics

David Banh
AskExplain
david.b@askexplain.com

Abstract

A new method proposes to align single cell reference datasets to spatial transcriptomic genes and tissue images. The technique can transfer single cells to unmeasured histology tissue by first aligning a single cell reference dataset to known Spatial Transcriptomic tissue, and learn from the alignment to predict gene expression on new histology. The model can invert the alignment transformation to generate new histology images from gene expression vectors, allowing for in-silico perturbation analyses through dynamically altering the levels of gene expression. Leveraging the cell atlas can lead to annotation of pathology and clinical specimens, enabling a mapping from the cellular and transcriptomic level to imaging tissue.

1 Introduction

One application of an interpretable statistical transfer learning technique is through Spatial Transcriptomic technologies (ST-tech), Method of the Year (2021) in Nature Methods [1]. The platform measures gene expression spatially on a histological tissue slide, linking gene expression to a tissue in a spatial context. Prior to the advent of ST-tech, there were single cell RNA sequencing technologies (scRNA-seq). The ability to sequence the gene transcription to count the expression and quantify the abundance at a single cell level enables the improved understanding of biological heterogeneity at a cellular level [2] [3] [4].

An important question is how to align single cells to the region of interest (ROI) on ST-tech that already contains gene transcripts within a spatial context. Another question is how to align single cells to regions outside of the ROI (outROI), so as to transfer information from the scRNA-seq to the outROI, thus predicting gene expression on regions without prior measurements. One benefit of the alignment using parametric models is the generation of new tissue images from the modelling of the gene expression vector. This would allow biologists to assess the perturbation of a gene by computationally adjusting its' expression, then via a generative transformation, convert the perturbed gene expression vector into a histology image for spatial context.

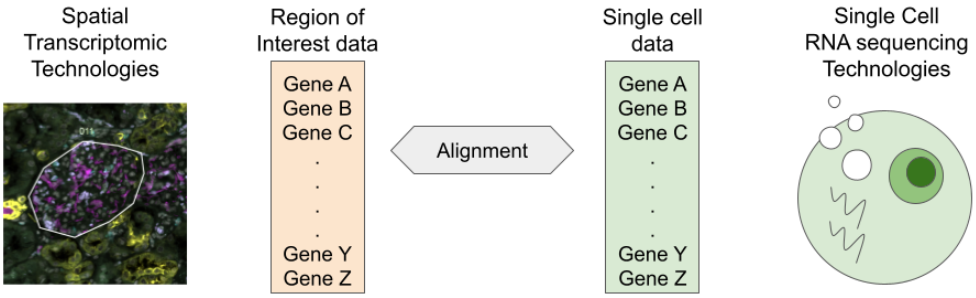

Figure 1: Aligning single cell data from a high-quality reference to Spatial Transcriptomic technologies (here the image used is from NanoString GeoMx data). Through learning from the alignment, the transfer of single cell data to histology images (without measured gene expression) is realisable.

The contributions of this paper on Generative Modelling are made through two aims. By modelling scRNAseq and ST-tech to transfer learn gene expression from ROI to outROI:

- Predict gene expression in histology that does not have any previous transcript measurements
- Generate new histology images based on gene expression vectors via generative modelling

The github code can be found at: https://github.com/AskExplain/gcproc

The github analysis code can be found at: https://github.com/AskExplain/gcproc$_analysis$

An interactive dashboard can be found at: https://board.askexplain.com/genecode

2 Data

The dataset consists of four main parts: 1) the scRNA-seq reference dataset containing gene expression, the ST-tech dataset containing 2) gene expression in conjunction with 3) the spatial histology tissue corresponding to the gene expression, and 4) the spatial histology tissue without gene expression to be evaluated.

2.1 Main Data

The scRNA-seq data is sourced from a single cell kidney reference atlas [5]. The dataset consists of 12,423 single cells, each with 17,682 genes that overlap with the spatial transcriptomic dataset. There are 33 cell type from the adult non-PT parenchyma and immune kidney datasets. There are also 44 cell types from the same atlas, but of fetal nephron, stroma and immune kidney tissue (only the fetal nephron tissue is used).

The ST-tech data is sourced from the NanoString Spatial Omics Hackathon at DevPost. The data used was a subset of what was provided, namely, the gene expression data and the stained histology tissue. The gene expression data consists of 18,504 genes with 231 ROI. The ROI are tiled into 32 by 32 pixel regions, each with the 3 standard RGB colour channels. These are flattened into vectors for further analysis. The outROI consist of a larger image of size 1280 by 1280 pixels, tiled into 1600 pixel frames of 32 by 32 pixels, again with the 3 standard RGB channels.

Importantly, the ST-tech provided is stained with 3 different biomarkers: magenta or pink for WT1, a marker of the podocytes in the glomeruli which are the filtration units of the kidney, yellow for Pan-cytokeratin, a marker of renal tubules which are the tubes that connect the glomeruli to the excretory regions of the kidney, and blue for CD45, a marker for immune cell types.

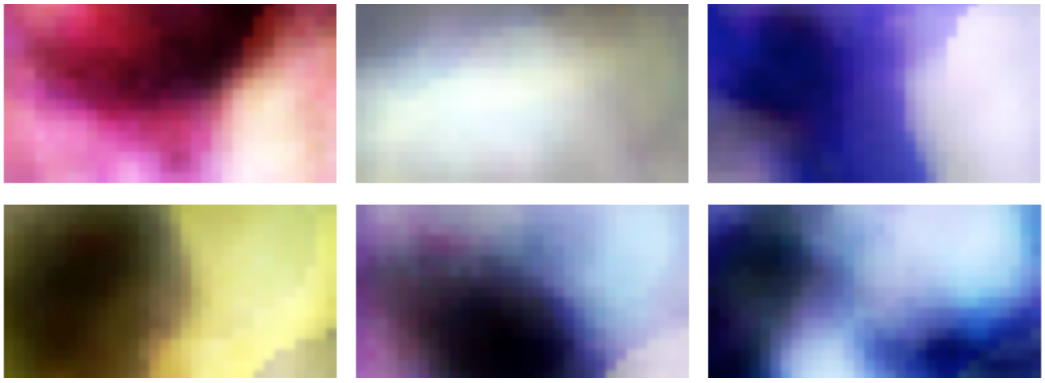

Figure 2: The learned top 6 components (32 by 32 pixel square tiles cropped in half) in a 70 dimensional pixel space highlight the most prominent features - a pink tile for WT1 staining of podocytes, a yellow tile for PanCK staining of renal tubules, three blue tiles for CD45 staining of immune cells and a gray-scale tile to shade and highlight other tiles when the tile vectors are linearly composed.

2.2 Perturbation datasets

The perturbation dataset involves only the fetal nephron scRNA-seq dataset from the Kidney Cell Atlas, and does not involve Spatial Transcriptomic data (albeit, model parameters learned using the ST-tech are involved). All images are generated from gene expression data, with added perturbations from a simulated continuous-time gene regulatory dynamics.

In the analysis, only a single cell vector is taken, composed of all 17,682 genes. Five genes are selected (WT1, PODXL, SIX2, NPHS1, NPHS2) to be perturbed simultaneously by complemeting the observed expression with an increase at every time step via a logistic sigmoid curve (parameters for the curves are randomly sampled).

The logistic sigmoid function is given as:

$$x(t) = \frac{L}{1 + exp(-k(t - t_0))}$$

where L determines the maximum gene expression level, t_0 defines the intercept of maximum change in the sigmoid curve, and k determines the overall rate at which the gradient changes. Here t is the time step, and $x(t)$ is the gene expression, which is a function of time at t.

2.3 Simulation datasets

The simulation dataset uses Splatter, and does not contain any prior gene expression information, other than those needed to estimate the parameters for Splatter [6]. The simulated samples are transformed via the learned parameters to fit within the structure of a kidney scRNA-seq dataset, but does not involve measured observations form the Kidney Cell Atlas.

In the analysis, four cell vectors are taken, each composed of all 17,682 genes. Eight genes are selected (WT1, PODXL, NPHS1, NPHS2, VEGFA, VIM, CDH1, AQP2) to be perturbed simultaneously by complemeting the observed expression with an increase at every time step via a continuous function.

The four histology image tiles from the inverse transformation of the cell vectors are combined by adjusting the colour and pixel gradient between image tiles using the expression only on overlapping pixels:

$$Tile_1 \times \alpha + Tile_2 \times (1 - \alpha)$$

Here, α ranges from $[0, 1]$ at discrete intervals. The full-sized image tiles are of 32 by 32 pixels, and overlapped by 8 pixels at each intersecting border. See Figure 3 for more details.

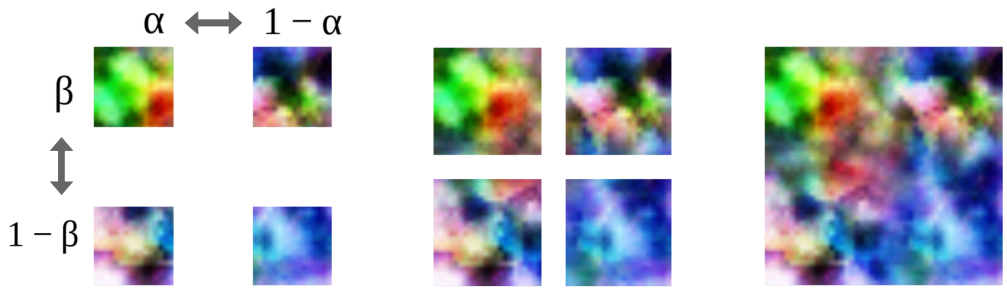

Figure 3: Simulation of spatial transcriptomic datasets (including histology images with gene expression) involving four scRNA-seq cells. To map between tiles that are inverse transformed from scRNA-seq data, overlapping pixels at intersecting tiles are filled in using a probabilistic gradient. Note that all images are generated from gene expression, and does not impact the genes via a feedback loop.

3 Methods

3.1 Transfer Learning method

To integrate data from one set of observations to another into an aligned and reduced dimensional space for both the samples and features via a transformation with matrix projections.

$$K_{k,i}Y_{i,l}v_{l,c} = J_{k,i}X_{i,j}u_{j,c}$$

The K and J parameter matrix transforms the samples into the same reduced dimensional plane along the sample space. This gives the benefit that JX and KY are now ready to be analysed together along the same dimensions.

Furthermore, the v and u parameters transform the features across the datasets into the same space, allowing Yv and Xu to be analysed along the same gene dimensions.

3.2 Aligning the gene expression of the tiles with the pixel information of the corresponding tiles

The model for this step begins with the ST-tech gene expression tiles $Y_{i,l}$ of the ROI and the ST-tech pixels from the expression tiles $P_{a,l}$, where a are the pixels, l are the ROI tiles, and i are the genes:

$$A_{k,i}Y_{i,l}q_{l,c} = B_{k,a}P_{a,l}r_{l,c}$$

3.3 Transfer Learning the encoding via shared spaces

Using the pixels as anchors, the parameters learned in the encoded alignment can be transferred to X the scRNA-seq reference via R the ST-tech pixels from histology tissue of unmeasured gene expression:

$$C_{k,i}X_{i,j}s_{j,c} = B_{k,a}R_{a,h}t_{l,c}$$

Notice that only the parameter $B_{k,a}$ for R is fixed as learned in the first stage. It is used as an anchor to pivot the model so the relevant parameters can be learned to enable a similar encoding for the scRNAseq data, X.

3.4 Transferred alignment enables prediction of gene expression on new histology tissue

Given the relation between the scRNA-seq dataset X, and R the pixel information from histology images without measured transcripts, a simple solution to aligning gene expression information to the pixel tiles is via:

$$Z_{i,j} = ((C_{k,i}^T)(C_{k,i}))^{-1}(C_{k,i}^T)^{-1}B_{k,a}R_{a,h}$$

3.5 Generative modelling to transform cells into new tissue images

Given the relation between X the scRNA-seq dataset and R the pixels from the outROI, it is possible to now generate images based on gene expression vectors alone. The model to generate new samples of tissue images is given by:

$$Q_{a,j} = ((B_{k,a})^T(B_{k,a}))^{-1}(B_{k,a}^T)^{-1}C_{k,i}X_{i,j}$$

This is notable given the benefit of interpretability for biologists to better understand how genes affect the histology in a spatial context.

4 Results

There are numerous papers on kidney development, as well as, kidney organoid and stem cell research which will aide in validating the computational results.

4.1 Modelling gene regulation from alignment of scRNAseq to histology

The first set of results in Figure 4 involves the same single cell vector of 17,000 genes with perturbations by three different gene sets (Set A: Immune markers - IGKC, CD74, B2M; Set B: Tubule markers - CDH1, PAX2, GATA3, Set C: Progenitor Nephron markers - SIX2, NOTCH1, WNT4, FGF8, FGFR1). The perturbations by modelling an increase in the expression of the gene sets takes the same cell down three different paths (Set A: Increasing blue cells, Set B: Increasing yellow cells, Set C: Cells of more defined and smaller shapes).

The second set of results in Figure 4 involves the alignment of gene expression scRNA-seq to ST-tech histology images, thus transferring information on all 17,000 genes spatially onto the tissue. With the main tissue in the center of Figure 4, the predicted gene expression can be seen in red on the sides. Notice SIX2 is expressed in the middle horizontally where podocytes are expression WT1, stained fluroescently pink. Furthermore, the three human leukocyte antigen genes HLA-A, HLA-B, HLA-C are expressed where the immune cells are, stained blue for the CD45 marker. Notice this is transferred from a pre-trained model, learned on a relatively small sample size of 231 tiles.

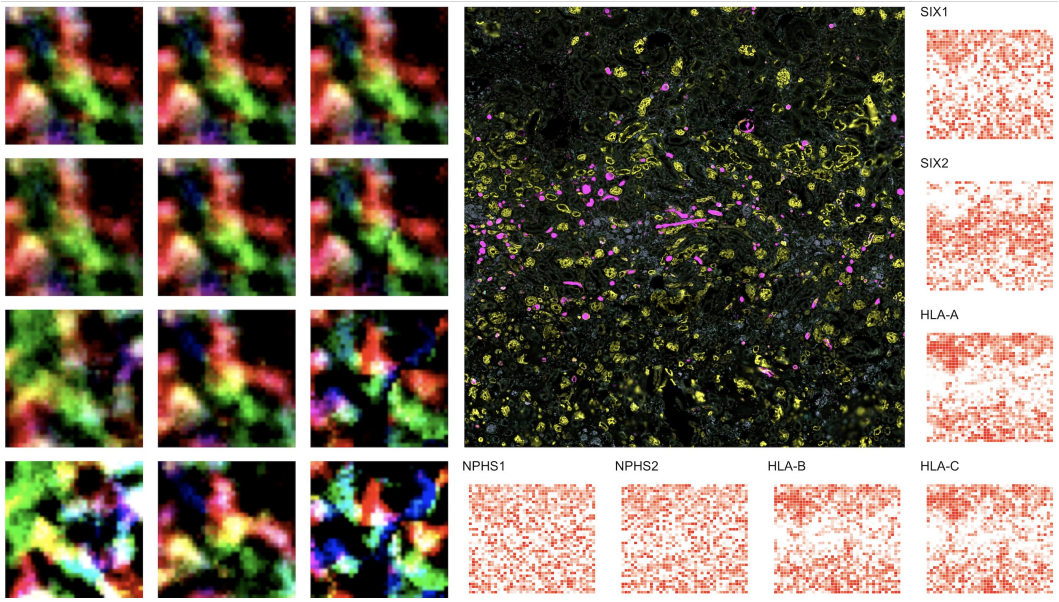

Figure 4: A) All images are gene expression vectors (a single cell) transformed into a histology tile. Each column represents the same single cell vector that is perturbed with increasing expression from top to bottom - increasing immune markers (blue, left column), increasing tubule markers (yellow, middle column) and increasing podocyte markers, and developmental transcription factors (pink, right column). B) Main histology slide is of a renal tissue showing WT1 (pink) in the middle region, CD45 at the top region (blue) and PanCK (yellow) from middle to bottom regions of the slide. Corresponding heatmap of gene expression alignment to histology for SIX1, SIX2, NPHS1, NPHS2, HLA-A, HLA-B, HLA-C.

4.2 Modelling dynamic gene regulatory networks

Modelling gene expression to continuously change as a function of time using logistic sigmoidal curves can be used to improve the interpretability of scRNAseq data. The main aim would be to benchmark continuous-time dynamically coupled gene expression networks to visualise how gene perturbations affects the histology in Figure 5.

Given the novelty of the method, the limitation is that each gene has a 1 to 1 correspondence to a image tile, and thus dynamic network perturbations are essentially a linear interpolation of these gene-based image tiles. This means that the method is more focused on interpretation to generate new hypotheses for biologists, rather than to discover new biology.

4.3 Simulating multiple cells as image tissues spatially and temporally

Multiple cells inverse transformed to image tiles can be stitched together to overlap and create a new image tile from the smaller sub-tiles. The new larger image tile can be used to model gene expression spatially. By dynamically altering gene expression through time, it is possible to assess this temporally.

In Figure 6, four cells are inverse transformed to generate the corner image tiles, of which are then stitched together to generate a larger image tile. The four cells containing genes are simulated from Splatter [6], and transformed to fit the shape of data from the Kidney Cell Atlas (in particular, it is of fetal nephron tissue). Notice that no real observations are used to generate the Spatial Transcriptomics simulated data.

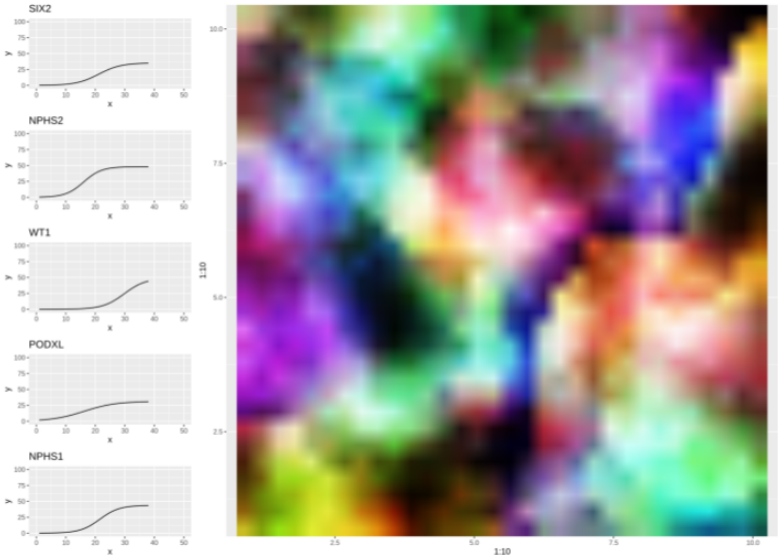

Figure 5: Modelling the dynamic continuous-time gene expression by adjusting five genes (SIX2, NPHS2, WT1, PODXL, NPHS1) in a vector of 17,000 genes representing a single cell. The single cell vector is inverse transformed into a histology image tissue that alters visual structure based on the gene expression modelling.

5 Extensions to develop an open-source toolbox

With the generation of the Gene Ontology database, computational methods have had a tool to which models can be benchmarked to assess how well the methods capture biological variation and signals. By extending this to integrate with current methods such as Splatter (see Figure 6), as in SplatPop [7] can enable computational scientists a working tool in the coming years of growing Spatial Transcriptomic modelling.

Furthermore, it can also be used as a tool to benchmark dynamically coupled gene expression regulatory networks (see Figure 5). This new tool can improve upon the interpretability of these models to help life-scientists and biologists draw more precise and detailed hypotheses.

Given the small and limited size of the current dataset (n = 231), it would be of great benefit to assess how well the model can capture inherent gene expression signal on larger datasets. It is hypothesised that it could provide clues to more biological knowledge on how two similar cells with similar gene

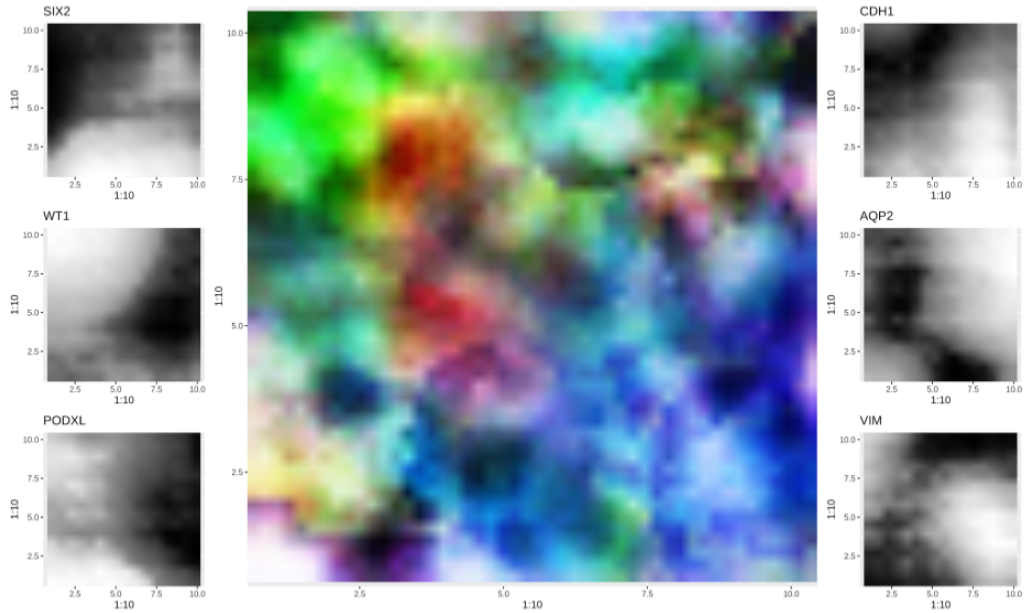

Figure 6: The simulation of four cell samples are combined into a large spatial transcriptomic map with both histology image (middle), and spatial gene expression (6 plots by the side - SIX2, WT1, PODXL, CDH1, AQP2, VIM). Genes can be modelled to change dynamically through time and space. Heatmaps highlight which regions of the spatial tissue is expressing high levels of that gene.

expression can be physically different - as well as distant in the principal component space, using the Generative Encoding modelling technique and transform between gene expression and tissue image.

6 Related work

Alignment of datasets

The projection of a dataset into the same space can occur by transforming one dataset into the space of another, similar to Procrustes analysis [8], or into the same reduced dimensional space, akin to Generalised Principal Component Analysis [9].

Alignment of cells to spatial transcriptomics

The alignment of cells to ST-tech to impute spatial gene expression involves maximising the aligned gene correlation between cell and ST-tech regions. There exist several methods that optimise the correlation between cell and ST-tech regions, such as Tangram that uses cosine correlation [10], or cell2location [11] which assigns cell type similarity based on a negative binomial regression weight.

Prediction of gene expression

There are a number of methods that aim to predict gene expression on histology tissue without prior measurements using ST-tech data to train on. For example, stNet looks at using a CNN to extract features and predict gene expression of around 200 genes in a fully integrated deep learning model [12]. Another method using data from the The Cancer Genome Atlas Program (TCGA) uses tissue RNA-seq data measured on patients with histology samples and predicts the gene expression at a bulk-level [13].

Generative image modelling

From reviewing the current literature, the generation of tissue images from gene expression vectors is a novel method for ST-tech and scRNAseq data. It is possible to sample from the manifold of likeliy activations based on the weights in a deep learning method to visualise the features learned in a Convolutional Neural Network [14]. The deep mixtures of factor analysers by Tang et al, from Geoffrey Hinton's group, also produced generative models that created sample images, although using face data [15].

7 Conclusion

The notion of communicating information across disciplines is important, the same idea can be applied to datasets. Aligning datasets together to make use of both in a synergistic way can improve modelling.

In this computational work, Single Cell RNA-sequencing is aligned to Spatial Transcriptomic technologies. This allows both the alignment of gene expression to histology tissue and the inverse through transforming histology tissue to gene expression.

This modelling technique of encoding both datasets into the same space (similar to an Autoencoder) can learn interpretable parametric weights that have biological meaning. The correspondence between gene expression and image tiles, would allow the benchmarking of computational methods that either simulate or generate gene expression for improved interpretations for biologists.

An open-source community-based tool can be developed and integrated into existing methods to benefit the single cell field.

8 Acknowledgements

Thanks to Ryan Deslandes for preparing the interactive dashboard and providing advice on the database and software platforms.

References

[1] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature Methods*, 18 (1):9–14, Jan 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01033-y. URL https://doi.org/10.1038/s41592-020-01033-y.

[2] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190): 1396, Jun 2014.

[3] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138, Mar 2015.

[4] Angela R. Wu, Norma F. Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E. Rothenberg, Francis M. Mburu, Gary L. Mantalas, Sopheak Sim, Michael F. Clarke, and Stephen R. Quake. Quantitative assessment of single-cell rna-sequencing methods. *Nature Methods*, 11(1):41–46, Jan 2014.

[5] Benjamin J. Stewart, John R. Ferdinand, Matthew D. Young, Thomas J. Mitchell, Kevin W. Loudon, Alexandra M. Riding, Nathan Richoz, Gordon L. Frazer, Joy U. L. Staniforth, Felipe A. Vieira Braga, Rachel A. Botting, Dorin-Mirel Popescu, Roser Vento-Tormo, Emily Stephenson, Alex Cagan, Sarah J. Farndon, Krzysztof Polanski, Mirjana Efremova, Kile Green, Martin Del Castillo Velasco-Herrera, Charlotte Guzzo, Grace Collord, Lira Mamanova, Tevita Aho, James N. Armitage, Antony C. P. Riddick, Imran Mushtaq, Stephen Farrell, Dyanne Rampling, James Nicholson, Andrew Filby, Johanna Burge, Steven Lisgo, Susan Lindsay, Marc Bajenoff, Anne Y. Warren, Grant D. Stewart, Neil Sebire, Nicholas Coleman, Muzlifah Haniffa, Sarah A. Teichmann, Sam Behjati, and Menna R. Clatworthy. Spatiotemporal immune zonation of the human kidney. *Science*, 365(6460):1461–1466, 2019. doi: 10.1126/science.aat5031. URL https://science.sciencemag.org/content/365/6460/1461.

[6] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1), September 2017. doi: 10.1186/s13059-017-1305-0. URL https://doi.org/10.1186/s13059-017-1305-0.

[7] Christina B. Azodi, Luke Zappia, Alicia Oshlack, and Davis J. McCarthy. splatPop: simulating population scale single-cell RNA sequencing data. June 2021. doi: 10.1101/2021.06.17.448806. URL https://doi.org/10.1101/2021.06.17.448806.

[8] Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis.

[9] Ren Vidal, Yi Ma, and S. S. Sastry. *Generalized Principal Component Analysis*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 0387878106.

[10] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Asa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with tangram. *bioRxiv*, 2020. doi: 10.1101/2020.08.29.272831. URL https://www.biorxiv.org/content/early/2020/09/24/2020.08.29.272831.

[11] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Artem Lomakin, Veronika Kedlian, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. *bioRxiv*, 2020. doi: 10.1101/2020.11.15.378125. URL https://www.biorxiv.org/content/early/2020/11/17/2020.11.15.378125.

[12] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8): 827–834, Aug 2020. ISSN 2157-846X. doi: 10.1038/s41551-020-0578-x. URL `https://doi.org/10.1038/s41551-020-0578-x`.

[13] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature Communications*, 11(1):3877, Aug 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17678-4. URL `https://doi.org/10.1038/s41467-020-17678-4`.

[14] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

[15] Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Deep mixtures of factor analysers. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1123–1130, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.