Interpretable Deconvolution from Spatial Transcriptomic Sequencing

Sarah Liu[1] and Will Specht[1]
1. Department of Biomedical Engineering, Columbia University, New York, NY

Abstract

Deconvolution of image-based spatial transcriptomics can serve to help uncover tissue niches and cell-to-cell communication, but current reference-based computational deconvolution approaches are limiting in cost and sample availability. Reference-free approaches, such as Starfysh[1], offer an alternative approach using manually and automatically defined gene signatures of present cell types, but while incorporating histology. Instead, we propose an integrated variational autoencoder based model using explicitly extracted morphological features and gene expression to predict cell type proportions for improved model performance, reduced memory usage, and a more interpretable integration of histology. This proposed model seemed to align with assumptions about the distribution of morphological features, and led to similar performance as Starfysh with histology integration, but with the added advantage of greatly reduced memory usage.

Introduction

Since its initial breakthrough in 2009 by Tang et al., single-cell RNA sequencing (scRNA-Seq) has become an invaluable method for understanding cellular activity and dissecting distinctions within and across tissue and disease[2]. scRNA-Seq allows scientists a high-dimensional snapshot of the current transcriptomic activity of cells, which provides strong indication of a cell's downstream function, can help reveal heterogeneity within previously established subsets, and can reveal upstream mechanisms driving differences between conditions. Advancements in dissociation-based single cell technologies, which strip the spatial context of cells, have been coupled by advances in Spatial Transcriptomic Sequencing (stRNA-seq), offering a means to profiling cell-cell communication, localization, and dysregulation of tissue niches in disease[3,4,5].

One approach to spatial sequencing are "image-based" technologies, which use fluorescently labeled probes to illuminate the presence of individual transcripts at subcellular resolution[6], but can only image hundreds of genes at a time and can struggle to differentiate gene-expression levels in abundant transcripts[7,8,9]. Another approach builds on Next-Generation Sequencing (NGS) through array- or spot-based sequencing approaches, in which non-specific RNA probes are placed and barcoded by location-specific tags before PCR amplification and conventional sequencing[3]. This approach often allows for a greater read depth and unbiased profiling of transcripts, but each spot may contain multiple cells, losing single-cell resolution[10]. Computational approaches have been developed to separate the expression profiles from each cell within these spots[1,11], by relying on paired scRNA-Seq, snRNA-Seq or known sets of gene markers. These approaches are able to predict cell proportions within spots, but require the user to select a cell density (which may vary across tissue) and fail to fully leverage available histological information. Morphological metrics—such as nuclear surface area, convexity, and perimeter—have provided insight into cell identity even before the development of spatial transcriptomics and could serve to enhance existing single-cell deconvolution techniques. Several computational approaches have been developed to dissect cellular morphology for clinical purposes and have already been successfully applied to samples of microglia and breast cancer tumor cells[12,13].

This new approach is built on a new reference-free approach Starfysh, which leverages automatically and manually determined gene signatures as reference and optionally uses histological images of the sequenced spot and its neighbors to improve cell proportion prediction of each spot. However, histology is currently incorporated in the model as a matrix of pixel intensities in a square section of the histology image surrounding the spot, which may lead to increased computational costs and model complexity, as well as low interpretability of different cell types' morphological traits and the morphological drivers of cell type classification. By building on the work of Starfysh and instead leveraging explicit morphological feature extraction, we were able to achieve comparable deconvolution performance with lower peak memory usage.

Methods and Algorithm

Archetypal Analysis

A key aspect of spot deconvolution is the creation of references of the cell types present in the dataset to define the signatures into which the overall gene expression signature can be decomposed. For reference- based approaches, these references are defined using clustering or other classification methods in a single cell, before deconvolution is performed into these component cell types.

By contrast, archetypal analysis[14] fits a convex polytope to the data in principal components space of gene expression (PCs=30), which minimizes the projection distance of points to the polygon, defined by *k* extrema or vertices. These extrema were shown by Starfysh to correspond with the most homogeneous spots in the tested spatial transcriptomic data, and spots between these extrema or anchors are assumed to have cell proportions that correspond to linear combinations of these pure spots. These spots are then used to define the gene signature that defines each cell type.

This algorithm was used unmodified by our project, but is described in their paper as minimizing $X' - (WH)^2 = X' - (WBX')^2$. $W \in \mathbb{R}^{SxD}$ is a positive valued matrix whose rows sum to 1, $B \in \mathbb{R}^{SxD}$ is a positive valued matrix whose rows sum to 1. $H = BX' \in \mathbb{R}^{DxP}$, where D is the number of archetypes, P is the number of principal components (30), and S is the number of spots[1]. Fisher Separability Analysis was used to estimate the intrinsic dimensionality k, and D was calculated by iterating from 1 to *k* until the variance explained by the polytope converged. They then implemented a resolution parameter to merge archetypes within a specified Euclidean distance, before neighborhoods of spots around each archetype, before a Wilcoxon rank sum test was used between the neighborhood of spots vs all other spots to find the top 30 differentially expressed markers, which are then used as gene signatures[1]. Finally, these gene signatures are used to define priors for cell proportions $A \in \mathbb{R}^{Sxk}$, by alignment between spots and each reference gene signature.

Spot and Feature Extraction

To incorporate morphological features on a spot-level basis, 144 pixel rectangular sections centered on each spot area were isolated from the high-resolution H&E image corresponding to the sequenced slice. CellProfiler was then used to identify individual cell nuclei, filter out low quality segmentations, and extract morphology-based metrics, such as area, eccentricity, stain intensity and number of nuclei per image (Figure 1). Features were selected qualitatively based on potential biological relevance to cell identity. These per cell features were then converted to per spot features by taking the average and standard deviation of all events in each spot. Spots with no identified cells were removed from further analysis. To reduce computational complexity, only the first 500 spots with identified cells were used for analysis.

Variational Autoencoder Architecture and Assumptions

Variational autoencoder architecture and model assumptions are consistent with those described by vanilla Starfysh, with modification to the encoding and decoding-related features. As described as above, image morphological features are extracted and compressed into averages and standard deviations and form $M \in \mathbb{R}^{SxF}$, where F is the number of features extracted (18 features, from which average and standard deviation are extracted as well as cells per spot, for a total of 37 features). Each morphological feature for a cell type is assumed to be drawn from normal distribution and it is further assumed that the linear combination of these cell level features in each spot can be described by a single normal distribution representing each spot.

These morphological features are integrated with expression in a z-dimensional (32) latent space using Product of Experts distributions[15]. For each spot, morphological features are approximated by $q_{\theta 2}(z_i \mid M_i) = \text{Normal}(\mu_2, \sigma_2^2)$ and expression is approximated by $q_{\theta 1}(z_i \mid x_i, c_i) = \text{Normal}(\mu_1, \sigma_1^2)$, where x_i and c_i are the expression and cell proportions of spot i. The morphological and expression independent decoding is drawn from each of those distributions respectively. Joint modeling through the Product of Experts

occurs by calculating an overall $q_\theta(z_i \mid c_i, x_i, M_i) = \frac{\mu_1/\sigma_1^2 + \mu_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}$. The ELBO method can with an information bottleneck to optimize view-specific and joint views objective functions through a single loss function: $\mathcal{L}_{total} = \mathcal{L}_{joint} + \alpha * (\mathcal{L}_{GEX} + \mathcal{L}_{morphology})$, where α is a weighting parameter between loss functions set to 5. These joint and independent objective functions were then optimized using Starfysh's default implementation and parameters in PyTorch (200 epochs, 0.001 learning rate, an exponential scheduler learning decay rate of 0.98, and neural network initializations using Kaiming initialization)[1].

Benchmarking of Algorithm

This algorithm was benchmarked against Xenium annotations alongside Starfysh's default implementation with histology integration. Both models were scaled down in complexity (32-dimensional latent space) in order to reduce overall model size and to maintain an accurate comparison between the two modalities.

Xenium and Visium H&E stains were aligned using SIFT[16] registration, which was then used to define a transformation matrix between the paired Xenium and Visium datasets. This transformation matrix could then be used to map cell's locations and labels from the Xenium dataset to their corresponding Visium coordinates. The cell identities and locations were then used to calculate per-spot cell proportions. Performance was compared between expert Xenium annotations and our predictions through using the residual across all spots, normalized by the number of spots tested. Pearson correlations were also determined between archetypes and annotations to partially account for ambiguous annotations.

Assignment of archetypes to cell types

Archetypes are defined by differential gene expression spots outside the archetype neighborhood. These gene signatures were used to manually assign archetypes to cell types, using the same naming conventions and granularity as Xenium annotations. Ambiguous archetypes were assigned to the predicted cell type to which the archetype most closely aligned, then it was validated that the gene signatures aligned with that predicted cell type. Archetypes had previously been validated by Starfysh, and concordance in the most popular cell types between default Starfysh and our algorithm indicates that this method of assignment does not invalidate the benchmarking results.

Data acquisition

All data used in this study was obtained from paired slices of human breast cancer tumors from 10x Genomics sequenced on their Visium spot-based sequencing and Xenium probe-based platforms. Both datasets are publicly available on 10x Genomics and are linked below
Xenium: https://www.10xgenomics.com/products/xenium-in-situ/human-breast-dataset-explorer
Visium: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE243280 (GEO: GSE243280)

Implementation

Spot Extraction

Spot extraction was performed on the high resolution Visium H&E stain by taking square sections 144x144 pixels from the sample. The large radius samples were taken in order to account for RNA diffusion, which may have led to spots picking up RNA from outside their spots radius and for cells partially inside a spot, which may have relevant global morphology even if RNA is captured primarily from one portion of the cell. The pixel coordinates of the spots were also rescaled in order to align with the captured spot coordinates. Examined spots appeared to contain 0–15 cells based on visual examination of nuclear staining.

Morphological Feature Extraction with CellProfiler

During the feature extraction, several considerations were taken into account. Hematoxylin and eosin (H&E) staining is intended to provide good contrast for nuclei as compared to background

cytoplasm and therefore does not explicitly color other parts of the cell. As such, achieving successful identification and segmentation of cell boundaries from the input image would be unlikely. When initially attempting to segment cell boundaries from identified nuclei in CellProfiler, there were many discrepancies in the results and segmentation appeared only tentative at best. This is in line with the Xenium annotations, which also seem to infer cell boundaries surrounding the nucleus, rather than clearly defined boundaries that are visible in the stain (Figure 2). Therefore, although the project was initially proposed with overall cell morphology in mind, only nuclear features and cell counts were used in downstream analysis.

Nuclei were identified using CellProfiler, based on inverted, grayscale H&E images of individual spots and designated target radius dimensions. These segmented nuclei were then analyzed for various features that were thought to be relevant to cell classification. These features were then aggregated to single means and standard deviations for each spot, then mapped back to their corresponding spot barcodes, which were rearranged due to the method by which Cellprofiler orders images for processing.

Data Loading and Model Adjustments

Due to computational limitations—primarily RAM, Visium data loading had to be adjusted when testing the original model with histology integration. Most significantly, the image files were stored only once and other duplicate information was stripped from the normalized and unnormalized object when they were not needed. Furthermore, the image and spot information from the datasets were subsetted to only include 500 events to further reduce memory constraints, while maintaining the high resolution necessary to extract relevant morphological information. The overall model also had to be reduced in size, namely the number of variables in the latent space. Finally, throughout this process excess variables were continuously cleared out to further reduce memory usage. Overall, this process served to highlight an advantage of using morphology rather than the full histology image to train our model, which is that image processing can be performed separately, reducing peak memory usage and training complexity. Additionally, when aligning Visium and Xenium images to perform benchmarking, the full image files could not be loaded, so partial loading and other optimizations were needed.

VAE Model Changes

The primary novelty of this approach is to use prior tools and biological knowledge to define explicit morphological features rather than feeding pixel intensity information directly into the jointly trained model. We aligned the morphological features as a ".uns" dataframe and inputted them in as features to their own variational autoencoder, in place of the histology slices. We similarly modified decoding to optimize for reconstruction of these features. We assumed that the features were drawn from a normal distribution defined by cell type identity and that the spots were also normally distributed.

Validation

Validation was performed by aligning H&E images between the Visium and Xenium datasets, then using that transformation to adjust cell locations to align with Visium spot locations. The cell counts of each annotated type were then normalized to cell proportions and compared to the archetype annotations by Xenium. For Xenium multi-annotated cells (e.g. stromal/immune/tumor) each cell was assigned to an equal proportion of each cell type.

Results

This new approach of explicit morphological feature extraction prior to model training led to comparable model performance compared to the default Starfysh implementation without histological training. This tool uses explicit morphological features with learned feature weights in order to facilitate reduced computational complexity and more interpretable feature importances.

Nuclei Morphology Extraction

The first major question of this project was whether meaning morphological features could be extracted from the Visium H&E images. We intuitively believed that cell membranes would be the most informative in defining cell type specific morphological features, however, qualitative analysis of cell segmentation of both CellProfiler extracted cell boundaries and Xenium annotations did not align with the weakly-defined borders between cells in the H&E stain (Fig 2). Therefore, CellProfiler was instead used to segment cell nuclei and define potentially relevant features (Fig 1). This segmentation seemed to align to qualitative analysis of the H&E sections, but may have slightly undercounted the number of nuclei due to overly stringent thresholds.

Nuclei Morphology are Normally Distributed

To investigate our assumption of a normal distribution of spot morphology, we examined the distribution of morphological features across all spots in breast cancer dataset and found that both feature averages and standard deviation were roughly normally distributed (Fig 4). Some features fit more closely to a binomial distribution and exhibited cutoffs, which were artifacts of CellProfiler thresholding, but across all features analyzed the assumption of normally distributed feature expression between spots appeared to hold.

Performance comparisons to Starfysh and Xenium

Our algorithm displayed a slightly improved normalized residual compared to default Starfysh (0.460 vs 0.467), but both algorithms display low correlation between any spot value and annotation (Fig 5,6). These low correlations held whether comparing cell type labeled archetypes to "ground-truth" annotations or the archetypes alone for both the default and modified implementations of Starfysh.

Discussion

The low reported performance of both approaches may have been due to low annotation quality, as many of the Xenium annotated cells had "None" labels or had multiple cell type labels. Although we randomly split these multi-assigned, this still likely led to poor accuracy measurements. This low annotation quality is supported by the continued improvement of the objective function during training and past performance of the unmodified archetypal analysis, indicating that priors were likely informative and that training was at least partially successful.

The lack of a large reduction in the residual between the normal and modified Starfysh may be an indication of a low quality ground truth, but may also indicate that nuclear morphology is not a useful additional modality to improve the performance of Starfysh. This lack of improvement may be due to artifacts of under-tuning the cell segmentation in CellProfiler, which would indicate that automated morphology extraction from H&E to augment deconvolution may prove to be unreliable.

Therefore, any future use of explicit cell morphology for improved reference-free deconvolution could investigate the construction of alternative priors for per spot feature extraction. Currently, per-cell features are modeled by a single mean and standard deviation, however, if the features associated with each cell type are normally distributed and a spot contains a combination of those cells, then per-spot information may better be modeled as a Guassian mixture of cell types. Additionally, morphology features could be incorporated into other facets of the model, including by helping to refine the prior before training or by updating the library size parameter. Finally, this method could be better validated by testing it across multiple annotation depths.

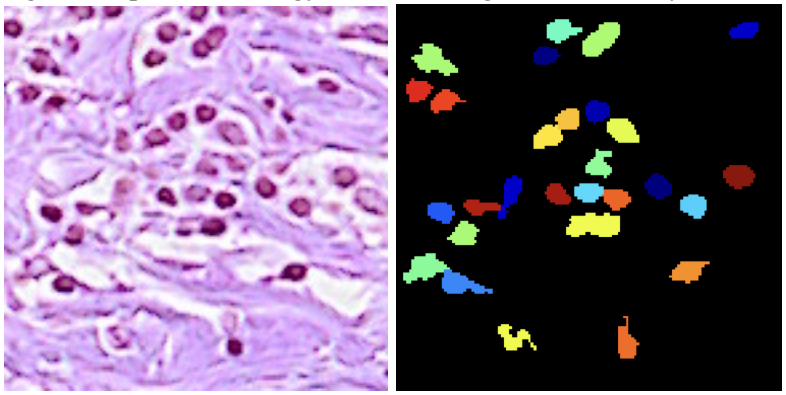
Contributions

Sarah and Will both performed background research on spatial transcriptomics as a technology and current deconvolution tools. Sarah focused on researching and performing cell morphology extraction using CellProfiler. Will researched prior work done by Starfysh and proposed algorithmic changes to better incorporate morphological features. Will performed memory optimization on the current Starfysh package and data loading approaches to make the problem computationally feasible. Sarah performed

aligned Xenium histology and performed performance analysis on modified and original Starfysh. Will and Sarah both made figures and wrote and edited the manuscript.

Figures

Figure 1: Spot 17 histology and nuclei segmentation analysis



Total Identified Nuclei: 28		
Measurement	**Mean**	**Standard Deviation**
Nuclei Area (pixels)	76.23	21.30
Nuclei Eccentricity	0.75	0.15
Nuclei FormFactor	0.82	0.23
Nuclei Perimeter (pixels)	35.71	10.01

Figure 2: Xenium labeled cell boundaries

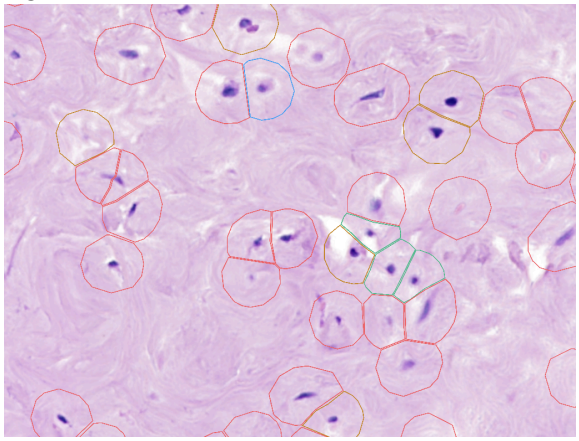


Figure 3: Distribution of Morphological Features. Average area of segmented nuclei (left) and average perimeter of segmented nuclei (right)

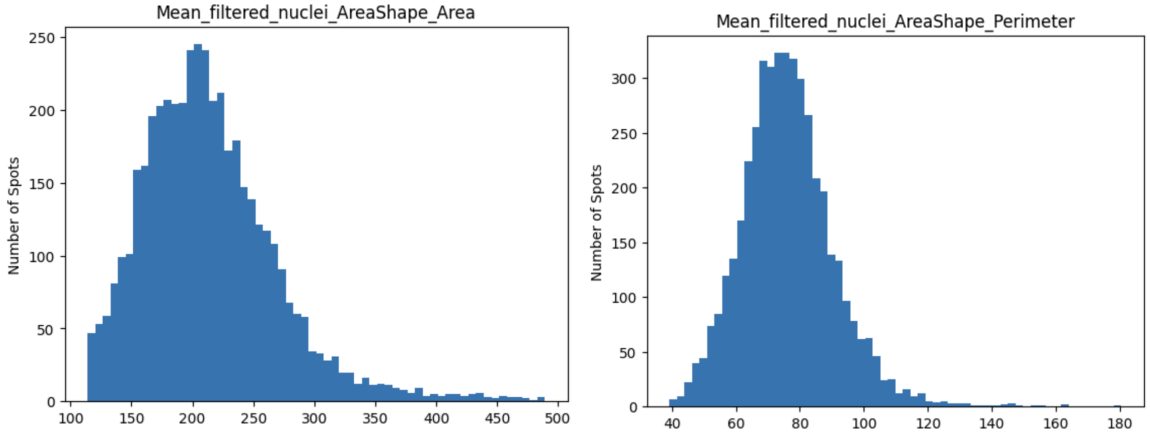


Figure 4: Heatmap of Pearson Correlation between Xenium cell type annotations (x-axis) and Starfysh inferred archetypes (y-axis)

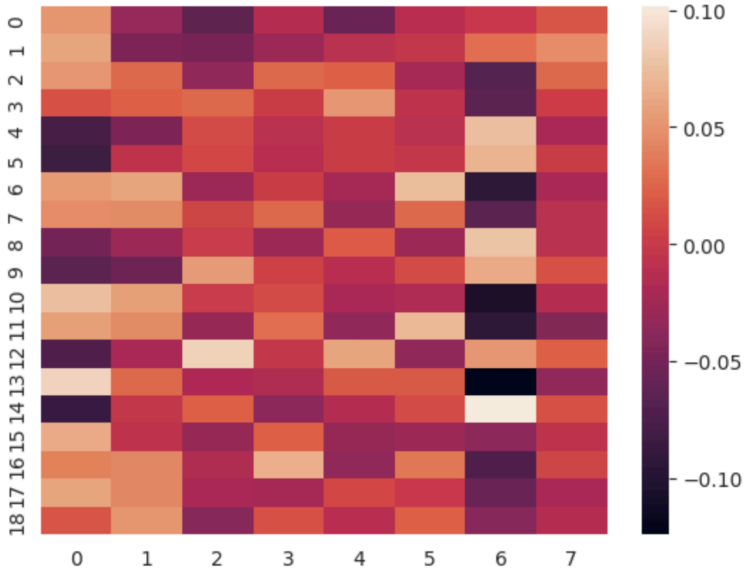
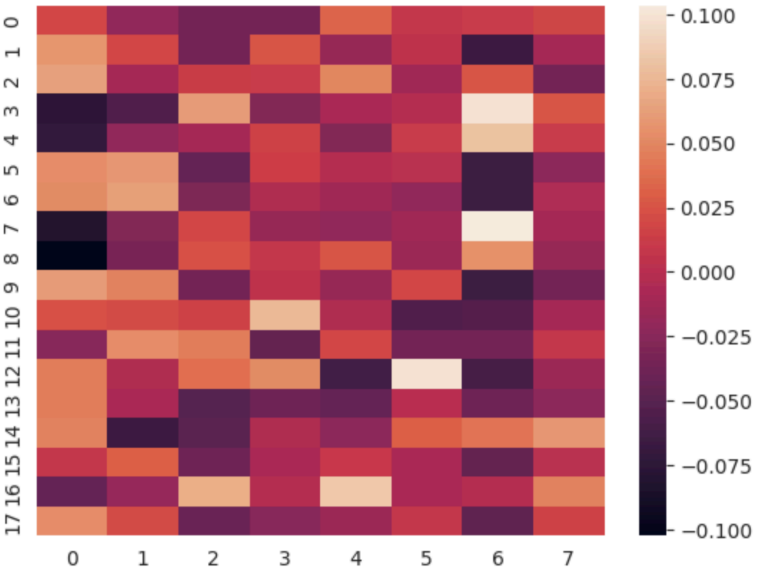


Fig 6: Heatmap of Pearson Correlation between Xenium Cell type percentage annotations (x-axis) and Starfysh inferred archetype proportions with morphological features (y-axis)



References

1. He, S. et al. Starfysh reveals heterogeneous spatial dynamics in the breast tumor microenvironment. 2022.11.21.517420 Preprint at https://doi.org/10.1101/2022.11.21.517420 (2022).
2. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377-82. doi: 10.1038/nmeth.1315. Epub 2009 Apr 6. PMID: 19349980.

3. Zollinger DR, Lingle SE, Sorg K, Beechem JM, Merritt CR. GeoMx™ RNA Assay: High Multiplex, Digital, Spatial Analysis of RNA in FFPE Tissue. Methods Mol Biol. 2020;2148:331-345. doi: 10.1007/978-1-0716-0623-0_21. PMID: 32394392.
4. Janesick, A., Shelansky, R., Gottscho, A.D. et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nat Commun 14, 8353 (2023). https://doi.org/10.1038/s41467-023-43458-x.
5. Wang F, Flanagan J, Su N, Wang LC, Bui S, Nielson A, Wu X, Vo HT, Ma XJ, Luo Y. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. J Mol Diagn. 2012 Jan;14(1):22-9. doi: 10.1016/j.jmoldx.2011.08.002. PMID: 22166544; PMCID: PMC3338343.
6. O'Conner C. (2008) Fluorescence *In Situ* Hybridization (FISH). Nat Educ, 1(1), 171. https://www.nature.com/scitable/topicpage/fluorescence-in-situ-hybridization-fish-327/.
7. Xia, C., Babcock, H.P., Moffitt, J.R. et al. Multiplexed detection of RNA using MERFISH and branched DNA amplification. Sci Rep 9, 7721 (2019). https://doi.org/10.1038/s41598-019-43943-8.
8. Moffitt, J. R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C., & Zhuang, X. (2016). High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proceedings of the National Academy of Sciences of the United States of America, 113(50), 14456–14461. https://doi.org/10.1073/pnas.1617699113.
9. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science (New York, N.Y.), 348(6233), aaa6090. https://doi.org/10.1126/science.aaa6090.
10. Williams, C.G., Lee, H.J., Asatsuma, T. et al. An introduction to spatial transcriptomics for biomedical research. Genome Med 14, 68 (2022). https://doi.org/10.1186/s13073-022-01075-1
11. Kleshchevnikov, V., Shmatko, A., Dann, E. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol 40, 661–671 (2022). https://doi.org/10.1038/s41587-021-01139-4.
12. Leyh J. (2021). Classification of Microglial Morphological Phenotypes Using Machine Learning. Cell Neurosci. https://doi.org/10.3389/fncel.2021.701673.
13. Masood S. (2016). Breast cancer subtypes: morphologic and biologic characterization. Women's health (London, England), 12(1), 103–119. https://doi.org/10.2217/whe.15.99
14. Cutler, A. & Breiman, L. Archetypal Analysis. Technometrics vol. 36 338–347 Preprint at https://doi.org/10.1080/00401706.1994.10485840 (1994).
15. Lee, C. & van der Schaar, M. A Variational Information Bottleneck Approach to Multi-Omics Data Integration. (2021) doi:10.48550/arXiv.2102.03014.
16. Lowe, D. G. Object recognition from local scale-invariant features. https://ieeexplore.ieee.org/document/790410.