




starfish: scalable pipelines for image-based transcriptomics

Shannon Axelrod¹, Matthew Cai¹, Ambrose J. Carr¹, Jeremy Freeman¹, Deep Ganguli¹, Justin T. Kiggins¹, Brian Long², Tony Tung¹, and Kevin A. Yamauchi³

¹ Chan Zuckerberg Initiative ² Allen Institute for Brain Science ³ Chan Zuckerberg Biohub

DOI: [10.21105/joss.02440](https://doi.org/10.21105/joss.02440)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Charlotte Soneson](#) 

Reviewers:

- [@giovp](#)
- [@shazanfar](#)
- [@vals](#)

Submitted: 29 June 2020

Published: 04 May 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

The exploding field of single cell transcriptomics has begun to enable deep analysis of gene expression and cell types, but spatial context is lost in the preparation of tissue for these assays. Recent developments in biochemistry, microfluidics, and microscopy have come together to bring about an “alphabet soup” of technologies that enable sampling gene expression in situ, with varying levels of spatial resolution, sensitivity, and genetic depth. These technologies promise to permit biologists to ask new questions about the spatial relationships between cell type and interactions between gene expression and cell morphology. However, these assays generate very large microscopy datasets which are challenging to process using general microscopy analysis tools. Furthermore, many of these assays require specialized analysis to decode gene expression from multiplexed experimental designs.

Statement of Need

starfish is a Python library for processing images generated by microscopy-based spatial transcriptomics assays. It lets biologists build scalable pipelines that localize and quantify RNA transcripts in image data generated by any hybridization- or sequencing-based *in situ* transcriptomics method, from classic RNA single-molecule FISH to combinatorial barcoded assays. Image processing of an experiment is divided into fields of view (FOV) that correspond to the data produced by a microscope at a single location on a microscope slide. *starfish* lets users register and pre-process images in each FOV, localize spots representing tagged RNA molecules in 3D, decode the identity of those molecules according to the experimental design, segment cells, assign the spots to cells, then aggregate spots into a cell x gene expression matrix. This spatially-annotated gene expression matrix can then be analyzed and visualized in downstream tools for single-cell biology, such as Seurat ([Stuart et al., 2019](#)), Bioconductor ([Huber et al., 2015](#)), Scanpy ([Wolf et al., 2018](#)), and cellxgene ([Megill et al., 2020](#)).

To enable large scale processing of these data, *starfish* leverages a 5-dimensional imaging data model (x, y, z, round, channel) backed by the cloud-friendly *spacex-format* file format, and *slicedimage*, an interface for lazy, distributed loading of *spacex-format* datasets. Furthermore, *starfish* implements comprehensive logging of all data processing steps for provenance tracking, reproducibility, and transparency. *starfish* is built on top of popular Python tools like *xarray* ([Hoyer et al., 2016](#)) and *scikit-image* ([van der Walt et al., 2014](#)).

There are a number of other tools which support localization and quantification of spots in fluorescent microscopy images, including ImageJ and CellProfiler, however these tools do not support multiplexed decoding of gene targets necessary for many assays. Other tools

which are designed more specifically to handle the kinds of assays that starfish supports include [dotdotdot](#) (Maynard et al., 2020), a MATLAB toolbox designed for RNAscope assays; [pysmFISH](#), a Python package designed for smFISH assays; and [SMART-Q](#), a fork from an earlier development release of starfish adding support for immunostaining and other features (Yang et al., 2020).

starfish requires a working knowledge of Python and fluorescent image analysis for a user to create an analysis pipeline. To help new users get started and support the broader single cell biology community in learning how to work with these data, *starfish* maintains example datasets and reference implementations ported from published assays, including MERFISH (Moffitt et al., 2016), In Situ Sequencing (Ke et al., 2013), osmFISH (Codeluppi et al., 2018), BaristaSeq (Chen et al., 2017), smFISH (Long et al., 2018), DARTFISH (Cai & Zhang, 2019), STARmap (Wang et al., 2018), and seqFISH (Shah et al., 2018). To take advantage of starfish's support for large scale processing, users must have familiarity with cluster or cloud computing.

starfish was developed alongside the [SpaceTx project](#), a CZI-funded effort to compare spatial transcriptomics methods in the context of determining cell types in the brain (Lein et al., 2018). *starfish* is currently in use by multiple research groups, including the [Allen Institute for Brain Science](#), the [Chan Zuckerberg Biohub](#), and the [Zhang Lab at UC San Diego](#). These groups support multiple large-scale projects profiling *in situ* gene expression, including the SpaceTx consortium, the [Human Cell Atlas](#), the [BRAIN Initiative Cell Census Network](#), and the [HuBMAP Consortium](#).

Acknowledgements

We would like to acknowledge the following contributions, which have been invaluable to the development of starfish.

For direct contributions of code and documentation, we thank Olga Botvinnik, Gökçen Eraslan, Kira Evans, Marcus Kinsella, Nicholas Mei, Josh Moore, Nicholas Sofroniew, and Ola Tarkowska.

We would like to thank the members of the SpaceTx consortium for working closely with us to understand their image processing steps, port pipelines into starfish, and contribute example datasets. We especially would like to thank Jeff Moffitt and Xiaowei Zhuang for their help with the MERFISH pipeline and example data, Marco Mignardi and Mats Nilsson for their help with the In Situ Sequencing pipeline and example data, Simone Codeluppi and Sten Linnarsson for their help with the osmFISH pipeline and example data, Xioayin Chen and Anthony Zador for their help with the BaristaSeq pipeline and example data, Richard Que and Kun Zhang for their help with the DARTFISH pipeline and example data, Xiao Wang, William Allen, and Karl Deisseroth for their help with the STARmap pipeline and example data, Dan Goodwin and Ed Boyden for their help with the ExFISH pipeline and example data, Nico Pierson, Sheel Shah, and Long Cai for their help with the seqFISH pipeline and example data, and Denis Shapiro for their help with the Imaging Mass Cytometry example data. Brian Long would like to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement and support. Finally, we would like to thank the Science and Technology teams and the Single Cell Program at the Chan Zuckerberg Initiative for their support of this work.

References

Cai, M., & Zhang, K. (2019). *Spatial mapping of single cells in human cerebral cortex using DARTFISH: A highly multiplexed method for in situ quantification of targeted RNA transcripts* [PhD thesis, UC San Diego]. <https://escholarship.org/uc/item/5bq3128f>

- Chen, X., Sun, Y.-C., Church, G. M., Lee, J. H., & Zador, A. M. (2017). Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research*, 46(4), e22. <https://doi.org/10.1093/nar/gkx1206>
- Codeluppi, S., Borm, L. E., Zeisel, A., Manno, G. L., van Lunteren, J. A., Svensson, C. I., & Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osm-FISH. *Nature Methods*, 15(11), 932–935. <https://doi.org/10.1038/s41592-018-0175-z>
- Hoyer, S., Fitzgerald, C., Hamman, J., & others. (2016). *Xarray: v0.8.0*. <https://doi.org/10.5281/zenodo.59499>
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., ... Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2), 115–121. <https://doi.org/10.1038/nmeth.3252>
- Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., & Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9), 857–860. <https://doi.org/10.1038/nmeth.2563>
- Lein, E., Long, B. R., Ganguli, D., Carr, A., Tung, T., & Freeman, J. (2018). **fish: Developing a computational pipeline for spatial transcriptomics in the brain*. <https://www.abstractsonline.com/pp8/#!/4649/presentation/25389>
- Long, B. R., Close, J. L., Tasic, B., Levi, B. P., Garren, E. J., Maltzer, Z., Nguyen, T., Thomsen, E., Bakken, T., Miller, J. A., Nicovich, P. R., Lein, E., & Zeng, H. (2018). *Exploring neuronal cell types in mouse and human brain using multiplex fluorescence in situ hybridization*. <https://www.abstractsonline.com/pp8/#!/4649/presentation/25387>
- Maynard, K. R., Tippani, M., Takahashi, Y., Phan, B. N., Hyde, T. M., Jaffe, A. E., & Martinowich, K. (2020). dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues. *Nucleic Acids Research*, 48(11), e66. <https://doi.org/10.1093/nar/gkaa312>
- Megill, C., Martin, B., Weaver, C., Bell, S., Badajoz, S., Weiden, M., Kiggins, J., Freeman, J., fionagriffin, bmccandless, Kinsella, M., Snyk bot, Prete, A., P., von Muhlen, M., Taylor, J., Virshup, I., Eraslan, G., Haliburton, G., & Wolf, A. (2020). *Chanzuckerberg/cellxgene: Release 0.15.0* (Version 0.15.0) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.3710410>
- Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39), 11046–11051. <https://doi.org/10.1073/pnas.1612826113>
- Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C.-H. L., Koulana, N., Cronin, C., Karp, C., Liaw, E. J., Amin, M., & Cai, L. (2018). Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell*, 174(2), 363–376.e16. <https://doi.org/10.1016/j.cell.2018.05.035>
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177, 1888–1902. <https://doi.org/10.1016/j.cell.2019.05.031>
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., & the scikit-image contributors. (2014). Scikit-image: Image processing in Python. *PeerJ*, 2, e453. <https://doi.org/10.7717/peerj.453>
- Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F.-A., & Deisseroth, K. (2018).

- Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), eaat5691. <https://doi.org/10.1126/science.aat5691>
- Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1). <https://doi.org/10.1186/s13059-017-1382-0>
- Yang, X., Bergenholtz, S., Maliskova, L., Pebworth, M.-P., Kriegstein, A. R., Li, Y., & Shen, Y. (2020). SMART-Q: An integrative pipeline quantifying cell type-specific RNA transcription. *PLOS ONE*, 15(4), 1–11. <https://doi.org/10.1371/journal.pone.0228760>