Open Science Software Initiative Proposal – EASI-FISH pipeline

MultiFISH Project Team

Introduction

Determining the three-dimensional spatial organization and morphological characteristics of molecularly defined cell types is essential to understand the architecture underpinning brain function. While methods have been developed to map molecularly defined cell types in the brain, these methods are usually in thin tissue sections and lack a robust data analysis workflow. Profiling in large and thick tissue volumes is desirable to reveal the 3D spatial molecular tissue organization, but it also poses challenges for data storage (>10TB) and processing. Additionally, methods to correlate molecular information with functional measurement of neuronal types are missing.

To address these issues, the multiFISH project team developed an EASI-FISH (Expansion assisted iterative-Fluorescence in situ hybridization) method and analysis pipeline to acquire and analyze gene expression in large tissue volumes. EASI-FISH is based on expansion microscopy, where biomolecules (in this case RNA) are anchored in place into a solvent-expandable polymer network, allowing the biomolecules to expand along with the network. Taking advantage of SPIM microscopy, EASI-FISH enables imaging of tissue volumes close to or greater than 1 × 1 × 0.3 mm in dimension. The utility of this method was demonstrated in one of the hypothalamic regions in mouse brain, the lateral hypothalamus area (LHA), where dozens of marker genes were profiled in close to 100,000 cells, revealing molecular and morphological diversities and 3D molecular organizations in this brain region (Wang et al., 2021).

The EASI-FISH analysis pipeline (**Fig. 1**) handles image data in the range of 10s of terabytes (TB) and consists of image 1) stitching, 2) registration, 3) 3D segmentation and 4) spot extraction steps that through parallel data processing, rapidly extract spatial, molecular, and morphological information from cells in the imaged tissue volume.

Stitching: For EASI-FISH in expanded thick tissue samples, multiple sub-volumes (tiles) were sequentially acquired, followed by computational stitching into a single large image (n5 format). We used an Apache Spark-based high-performance stitching pipeline, called stitching-spark (Gao et al., 2019) (https://github.com/saalfeldlab/stitching-spark), which includes a flat-field correction step, followed by globally optimal translation for each tile.

Registration: Next, image volumes from each round of FISH were aligned using cytosolic contours of cytoDAPI using an automated, Python-based, non-rigid 3D registration pipeline, called Bigstream (https://github.com/GFleishman/bigstream), which is more than 10-times faster than other deformable registration methods (e.g., ANTs) (Yushkevich, 2016).

Cell segmentation: In EASI-FISH samples, cytoDAPI provided a cytosolic signal for generating cell segmentation masks. We developed a deep learning-based, automated 3D segmentation algorithm, called *Starfinity* (https://github.com/stardist/stardist/tree/refinement), that creates a segmentation mask with an accuracy greater than 93%.

Spot detection: We adapted Airlocalize (Lionnet et al., 2011) for parallel spot detection, which we termed hAirlocalize (high-throughput spot detection based on Airlocalize). For cells with very high gene expression, where single spots cannot be resolved even with 2× expansion, we measured the signal intensity for each cell and converted the intensity to spot counts based on measured well-isolated single spot intensities for these genes.

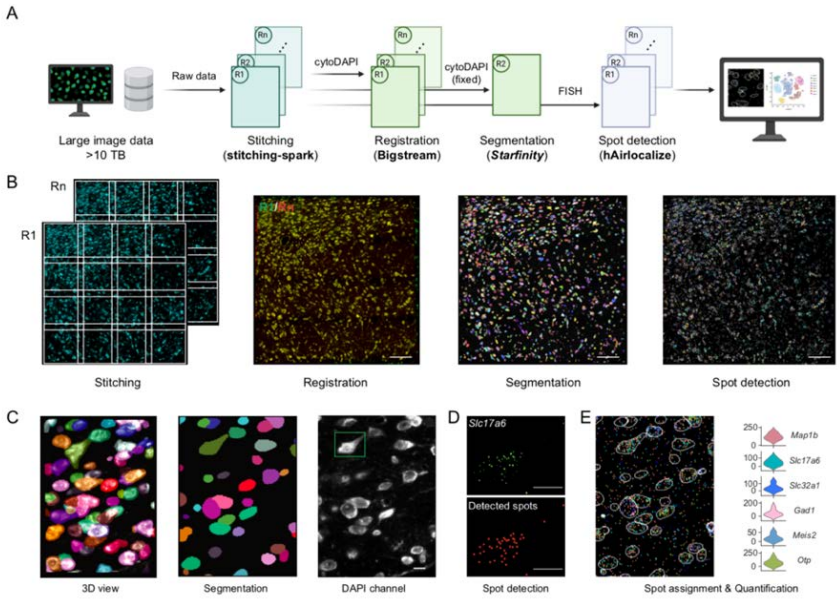

Fig. 1 EASI-FISH analysis pipeline (**A**) EASI-FISH data processing workflow. (**B**) Representative images showing stitching, registration, segmentation, and spot detection in large image volumes. Scale bar: 100 µm. (**C**) Example of *Starfinity* segmentation from cytoplasmic DAPI. (**D**) Representative hAirlocalize-enabled spot detection in cell highlighted with green square in C. Scale bar: 10 µm. (**E**) Spot counts per neuron measured with EASI-FISH.

Impact

With the help from Janelia Scientific Computing (SciComp), these computational modules were combined into a self-contained, platform-agnostic computational pipeline (https://github.com/JaneliaSciComp/multifish) for end-to-end EASI-FISH data analysis on high-performance workstations, computing clusters (such as LSF) and cloud-based platforms (e.g., AWS). EASI-FISH pipeline has attracted many potential users in and outside Janelia (>7 labs/teams in Janelia and >9 labs outside). An incomplete map of EASI-FISH pipeline current and potential users is presented on the right.

The EASI-FISH preprint has been downloaded 5,013 times from bioArxiv, since being posted in March. The GitHub repositories receive a steady stream of visitors, with 34 unique visitors over the past 2 weeks. A total of 26 users have read the user manual since it was posted on GitHub Pages in October.

Specific Aims

The goal for this proposal is to both maintain an active support of current EASI-FISH pipeline (Aim 1-3) and extend the EASI-FISH applications (Aim 4-6). Here we explain in depth our specific aims

and provide an estimated effort to achieve these goals. Each of the tasks represents an independent milestone, so this proposal can be scaled with the availability of funding.

1) Maintenance of current EASI-FISH pipeline (Estimated effort: **1.0 FTE** month)
Lack of maintenance is one major problem with many pipelines developed in the academia setting. Even worse, many software projects die with the projects that created them. A modest commitment towards active maintenance of the current pipeline, responding to bug reports, GitHub issues, and user feedback, as well as periodical upgrade of software modules, would increase the sustainability of the EASI-FISH pipeline for community usage. We have received great support from SciComp (Konrad Rokicki, Cristian Goina) on these issues and hope to continue this level of support.

2) Flexible input data format (Estimated effort: **0.5 FTE** month)
- Metadata reading: Currently, EASI-FISH pipeline accepts CZI files with MVL metadata. The MVL file needs to be saved separately through a Zeiss plugin. The stringent requirement for this separate MVL metadata file limits the utility of EASI-FISH pipeline. It is desirable to have the metadata parsed directly from the CZI files (XML format) and converted to JSON that can be read directly into the stitching module.
- Image data format: It would also be desirable if the pipeline accepts data collected as (or converted to) other generally accepted data formats, specifically h5 and TIFF-OME. Additionally, to accept single tile images rather than multi-tile images as input would further broaden the utility of EASI-FISH.
Please note that this proposed aim mainly deals with interaction between raw image data and the stitching module (first step of EASI-FISH pipeline). As image stitching is a general image analysis process and is currently used for many other applications (e.g., cleared whole brain samples, ExM samples, live samples), thus this work will be generally useful for the biomedical imaging community.

3) Big data compression, decompression, transfer and storage (Estimated effort: **1.0 FTE** month)
For labs that regularly process EASI-FISH data, in the range of 10-100TB in size, data transfer and storage can become a challenge. It also creates barriers for data reproducibility. Therefore, we hope to join the Open Data effort at Janelia and we think that developing data handling procedures, including fast data compression for storage and transfer and data decompression for inspection, visualization and reproducibility, would allow for wider accessibility of acquired data and also more efficient data storage. Currently, raw EASI-FISH datasets are stored in CZI format. Converting these files to N5, TIFF-OME, or HDF5 format with fast lossless compression (gzip, lz4, blosc, etc.) will be ideal. In addition, while compressing these files, we need to make sure the decompressed files are still compatible with the EASI-FISH pipeline, for example, can be recognized by the stitching module. Furthermore, the lossy compression schemes, such as isolate foreground from background should also be explored.

4) Continued integration of EASI-FISH modules (Estimated effort: **1.0-2.0 FTE** month)
With the continued development of imaging methods and processing tools, we envision that the performance and speed of EASI-FISH pipeline could be further improved by the continued integration of new and fully developed analysis modules. This will provide users more flexibility in image analysis that is tailored towards their data type. For example, SciComp is already integrating the newly developed

spot detection algorithm RS-FISH into the EASI-FISH pipeline. EASI-FISH could also benefit from other segmentation methods, such as the generalist 3D segmentation algorithm, Cellpose, to allow for broader tissue type segmentation without additional training data. Additionally, as stitching, registration and segmentation are generally useful image processing tools, this would broaden the application of EASI-FISH pipeline for analysis of other data types. For example, enable spark-based parallelization with BigStitcher, which would allow for non-rigid stitching of EXPAND samples.

5) Registration support for EASI-FISH (Estimated effort: **1.0-2.0 FTE** month)
Image registration is a key component of the EASI-FISH pipeline. To allow for higher-throughput gene expression profiling with EASI-FISH, spot-spot registration between image rounds will be necessary. While progress has been made on the experimental side, development and incorporation of analysis workflow (mainly registration related) for spot decoding will be necessary. In addition, registration of EASI-FISH data (*post hoc, ex vivo*) with other experimental modalities (such as *ex vivo* confocal imaging, and *in vivo* Ca^{2+} imaging) will be highly desirable to correlate molecular information with functional measurement in the brain. We would also like to upgrade the pipeline to take advantage of the newly developed Dask-based version of Bigstream, which performs better than the current version and will be maintained in the future.

6) Big Data visualization support (Estimated effort: **1.0 FTE** month)
Currently visualization of EASI-FISH data uses BigDataViewer in Fiji (raw data) and custom-scripting with Napari (processed). Despite efforts in this direction (e.g. RS-FISH commands), a user-friendly interface that would allow for visualization of raw and processed data (segmentation masks, registered image from multiple rounds and extracted spots) at multi-scale with BigDataViewer in Fiji or through lazy loading with Dask and Napari, while allowing users to query specific cell/area by ID and position (the query part is already solved) would be highly desirable.

References

Gao, R., Asano, S.M., Upadhyayula, S., Pisarev, I., Milkie, D.E., Liu, T.L., Singh, V., Graves, A., Huynh, G.H., Zhao, Y., et al. (2019). Cortical column and whole-brain imaging with molecular contrast and nanoscale resolution. Science 363.
Wang, Y., Eddison, M., Fleishman, G., Weigert, M., Xu, S., Henry, F.E., Wang, T., Lemire, A.L., Schmidt, U., Yang, H., et al. (2021). Expansion-Assisted Iterative-FISH defines lateral hypothalamus spatio-molecular organization. bioRxiv, 2021.2003.2008.434304.