Optimisation of MERFISH Data Analysis and Visualisation

Naureen Ghani and Yoh Isogai

May 21, 2019

Abstract

We studied the cellular diversity of hypothalamic tissue using Multiplexed, error-robust fluorescence in situ hybridization (MERFISH). The hypothalamus controls essential social behaviours and homeostatic functions. MERFISH provides precise quantitative gene expression but suffers from two major disadvantages: (1) low signal-to-noise ratio of single molecule signals and (2) big data interpretation challenge. To overcome these challenges, we benchmarked open-source single molecule localization algorithms in 2D vs 3D (i.e. 3D-daoSTORM and starfish) and developed a data visualization platform in JavaScript, respectively. This promotes data interpretation and generation of new hypotheses with MERFISH. We used machine learning techniques and GPU-computing to optimise MERFISH data analysis and generate cellular atlases of brain tissue.

Introduction

To determine the circuitry of the brain, we need an understanding of its components. Genomic-scale imaging is a way to obtain a systems-level view of neuroscience. Brain cells can be distinguished by their gene expression profiles. Moreover, RNA is the first level of gene expression. Single molecule fluorescent in situ hybridization (FISH) labels individual RNA molecules of a type within a tissue [1]. We can characterize the spatial distribution of RNA molecules and identify cell types based on RNA expression. With precise quantitative gene expression, we can test models of gene regulatory networks at molecular resolution [1]. However, we are limited by the number of spectrally distinct fluorophores that will allow us to see single molecules. The laboratory of Xiaowei Zhuang developed a method to solve this problem by sequential labelling and stripping of fluorophores to a fixed sample [2]. Multiplexed, error-robust FISH (MERFISH) exploits principles of digital communications to do robust quantification of gene expression.

In a landmark paper, Moffitt et al used MERFISH to reveal the organization of the hypothalamic pre-optic area [3]. This brain region is necessary for fundamental social behaviours (parenting and mating) and homeostatic functions (thirst, feeding, and sleep). Traditionally, cellular organization research has focused on the cortex since it is involved in higher-level cognition. However, we would be able to live without a cortex but not without a hypothalamus. Moffitt et al define a cellular atlas for this important brain region and identify ˜70 previously unknown neurons and their behavioural functions.

Moffitt et al studied >1 million cells in a 2mm x 2mm x 0.6mm block of brain [4]. They used hierarchical clustering to identify ˜70 neuronal sub-types. These sub-types corresponded to distinct combinations of neuromodulators that underlie social behaviours. For example, clusters with high levels of oxytocin receptor expression (Oxtr) were activated by pup-directed aggression by virgin males. Oxtr is an important hormone released during childbirth and sexual reproduction. This study highlights a novel role of Oxtr in social behaviours towards infants.

MERFISH is a way to build cellular atlases of the brain and define not only which cells are involved in behaviours but where these cells are. Spatial context is critical for understanding the brain, which has a unique topological arrangement. While some organs have symmetric cellular organization (i.e. the liver), the brain has microscopic and macroscopic anatomical organization that are the results of differential gene expression during development [5]. However, there are two major disadvantages of MERFISH. The first is a low signal-to-noise ratio of single molecule signals. There are spurious background signals from foreground RNA signals and single molecule signals with smaller areas tend to be low brightness [6]. This poses significant challenges for robust single molecule detection in tissues. The second is a big data interpretation challenge. MERFISH datasets consist of millions of RNA dots scatted across tissue sections. It is difficult to visualize, browse, and interpret such large-scale data.

10x Genomics EX2011
NanoString v. Harvard
IPR2024-00526

To overcome these challenges, we developed an analysis pipeline in 3D by leveraging many open-source image processing tools [7-11]. MERFISH analysis is currently done in 2D, where the z-data is discarded after generating maximum intensity projections to decrease computation time at the cost of sub-optimal single molecule localisation. We hypothesized that using machine learning techniques and GPU-computing would minimise this trade-off and accelerate 3D MERFISH analysis. We benchmarked a variety of single molecule localisation algorithms developed to date (i.e. 3D-daoSTORM and starfish) to optimise single molecule detection in tissues [10,11]. We then combined the improved single molecule localisation method with an enhanced MERFISH protocol that produces ~10x signals and is thus largely tissue-compatible. In addition, we developed a data visualisation platform based on JavaScript, in which a user can browse the data in an interactive fashion using a web browser. This facilitates data interpretation and generation of new hypotheses.

Results

We developed a scalable analysis pipeline of image-based transcriptomics data that relies on GPU-computing and published Python libraries. The input is a series of image stacks that represents a small panoramic view. We transform this into a table that lists the X, Y, Z positions of each RNA molecule detected. To do so, the steps are as follows:

Imaging Corrections

We did basic pre-processing to account for errors specific to the microscope optics independent of any experiment.

Illumination Correction

We used an acquisition system where the light source was perpendicular to the centre of the tissue sample. Thus, the field-of-view was not uniformly illuminated, and the periphery suffered. We implemented illumination correction to produce images such that the centre and periphery were equally lit. Since many spot-finding algorithms rely on a global threshold for peak intensities across an image, spots can be incorrectly excluded in low-illumination areas. This error can propagate to incorrect conclusions about the spatial distribution of RNA molecules.

The core of illumination correction is background subtraction. We averaged a series of images captured with no sample and no light (dark image) to calibrate the microscope. We multiply this averaged image with the experimental data to correct for any uneven illumination.

Chromatic Aberration

We used two wavelengths of light (561 nm and 647 nm) for imaging Cy3 and Cy5, respectively. Since the lens refractive index depends on the light wavelength, different colours are focused in different positions. Thus, image is blurred due to misaligned light paths. In **Figure 4**, the red and green colour channels are shifted with respect to one another. This is especially clear in the string of four dots at the bottom of the image. Channels that show different scaling or exhibit rotational misalignment would also need chromatic aberration correction.

We used Chromagnon, an open-source image correction software for chromatic shifts in fluorescent microscopy, developed by Matsuda et al [7]. The algorithm relies on calculating the translation, rotation, and magnification from reference images to warp the target image. In addition to global transformations, images are sub-divided to do phase correlation between channels and correct local chromatic shifts. We implemented Chromagnon on the High-Performance Cluster to correct several thousand images. An example frame is shown in **Figure 4** where the red and green colour channels fully overlap, and the dots are thus labelled in yellow.

Deconvolution of Optical Point Spread Functions

Deconvolution is a technique to improve the contrast and resolution of images captured using an optical microscope. It assumes that the path of light through the microscope is non-ideal only because it is convolved with the point spread function. By deconvolving the image and the point spread function, any distortion introduced by the microscope can be removed.

To measure the Point Spread Function (PSF) of our widefield microscope, we captured 3D images of sub-resolution beads. Since a single bead image will contain photon noise, we increased our SNR by aligning and averaging many beads. We reconstructed the experimental PSF from the averaged bead image [**Figure 7 and Table 1**]. We used microscopy parameters identical to those in MERFISH experiments to restore datasets in 3D. Since the experimental PSF includes all the physical deviations from the ideal model, we hypothesized that it would perform better than the theoretical PSF in image restoration. However, the experimental PSF data in Table 1 was questionable. We expected a greater FWHM in [647nm] than in [561nm] since it has a broader emission spectrum and higher spectral width. We stipulate that the bead used in [647nm] may have been an aggregate of two beads, which would decrease the FWHM.

We did not use the measured PSF to deconvolve because it is much larger (100 x 100 x 100) in the axial direction than the target image (2048 x 2048 x 40). To resolve this, we can pad the target image with zeros out to 100 z-slices, but this will increase processing time. Alternatively, we could remove slices in the measured PSF. However, we found that 40 z-planes was not enough to fit the PSF. We resolved this issue by regenerating the theoretical PSF with 40 z-slices.

We used PSF Generator, an open-source Java software package to generate realistic 3D microscope PSFs, developed by Kirshner and Sage [12]. We generated a scalar-based Born and Wolf diffractive model based on the acquisition system optics [**Figures 1-3**]. As expected, the FWHM for [647nm] was greater than that of [561nm] for the theoretical PSFs [**Figure 3**]. The thinner the peak, the better signal-to-noise ratio we can achieve.

We used Flowdec, an open-source library containing TensorFlow (TF) implementations of image and signal deconvolution algorithms, developed by Czech et al [8]. Flowdec constructs and executes TF graphs in Python and uses Richardson-Lucy deconvolution. We implemented Flowdec on the High-Performance Cluster to execute deconvolution against multiple GPUs in parallel. The data analysis is still in process and we plan to write a Snakemake workflow to split jobs.

To process a 2048 x 2048 x 40 3D volume with a PSF of the same dimensions, the execution times were:

- CPU-only solution: **2.5 hours**
- Flowdec/TensorFlow with partial GPU acceleration: **10 minutes**
- Flowdec/TensorFlow with full GPU acceleration: **20 seconds**

Partial GPU acceleration means that the fast Fourier transform (FFT) and inverse FFT (iFFT) operations run on GPUs while all other operations run on the CPU. We plan to process ~1600 (100 volumes x 16 rounds) and expect deconvolution will take **10 hours**. With full GPU acceleration, the processing time increases immensely by **30x**. An example of pre- and post- deconvolution is shown in **Figures 5-6**. The dots are much sharper even in hazy dense regions. This will greatly simplify single molecule localisation and increase the number of RNAs detected.

Feature Identification and Assignment

After the tissue volumes are pre-processed, we can use single molecule localisation algorithms to resolve X, Y, Z coordinates of RNA molecules. We are currently testing two code repositories for MERFISH data: 3D-daoSTORM and starfish. This data is not ready yet at the time of submission.

Image Registration and Decoding

Images of the same field-of-view in different imaging rounds are registered using the fiducial beads of each round by affine transformation [2,17]. Moffitt et al found that the RNA signal shifts ~100nm from round-to-round. To correct for this, they applied a low-pass filter with a kernel size of 100nm radius. The pixel intensities in all 16 rounds of imaging were then used to form a 16-D vector, which was normalized to unit amplitude. This vector was compared to the unit vectors defined by all valid barcodes. By pairwise comparisons of Euclidean distance, each pixel was assigned to a given barcode if the distance between them was less than the distance defined by a single-bit error. The algorithm then extended a line down each point of the 16-D vector as if skewering a kabob to identify contiguous pixels. Contiguous pixels that matched to same barcode were combined as a single RNA molecule. We expect single-molecule localisation to be the rate-limiting step of MERFISH analysis. We plan to use multiple cores on a computer cluster to perform it.

Segmenting Cells

Cell segmentation is a great challenge in MERFISH analysis. Although many algorithms exist for segmenting *nuclei*, there are few if any automated approaches to segment cells [11]. In a recent paper by the laboratory of Long Cai, segmentation for ~3000 cells were performed manually using the ROI tool in ImageJ [14]. This is neither efficient nor scalable. Current methods rely on classic techniques such as watershed segmentation and thresholding. The starfish code repository inverts the intensities of the nuclei and spots and sets a threshold ("water line"), such that each basin is treated as its own entity ("cell") [11]. This approach may work well for sparse cells but will fall short on densely-labelled tissues.

Assigning Spots to Cells

The 3D-daoSTORM segmentation algorithm exploits how the RNA density drops significantly at cell edges. To generate cell boundaries, Moffit et al created mosaics where each field-of-view was surrounded by its 8 nearest neighbours [17]. They segment nuclei by normalizing the DAPI signal to the maximum observed value and thresholding it. To assign spots to cells, they calculated the local density of barcodes in each composite image by binning decoded barcodes into 2x2 µm bins. These binned images were smoothed with a Gaussian filter of 2 µm and restored to their original size by bicubic interpolation. Segmentation was done as described above and used to find RNA counts per cell.

Data Visualisation

Big data demands scalable visualization tools to assist in data exploration and hypothesis generation. In MERFISH datasets, there are millions of spots scatted across tissue sections where each spot represents an RNA molecule. We hope to study: (1) the spatial distribution of different combinations of genes and (2) the spatial gene expression distribution within cell types. To do so, we developed a data visualization platform in JavaScript based on React and D3. This enables zooming, panning, and rotating across fields of views.

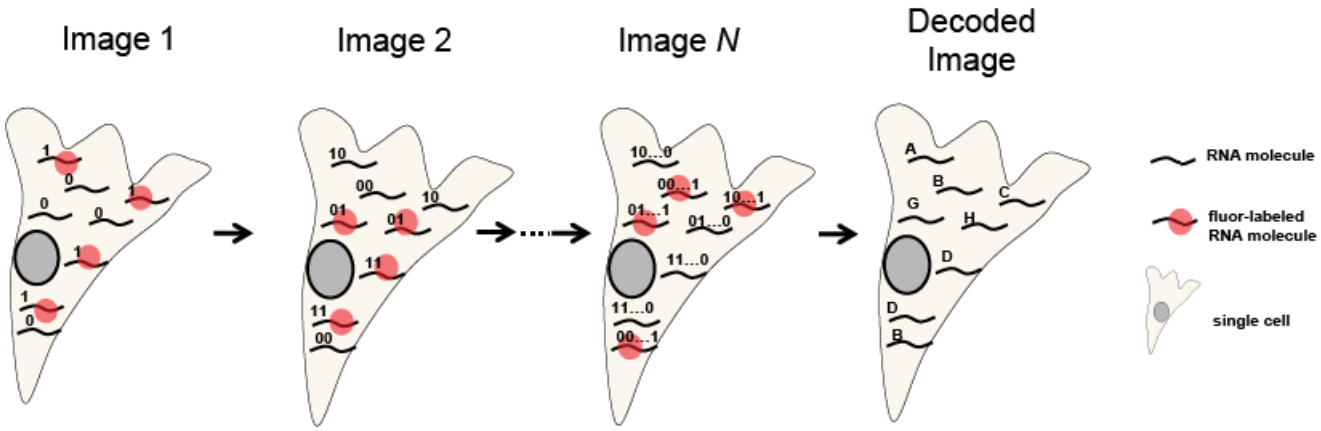
Spatially-resolved mRNAs can be combined with cell staining to enable segmentation into cells. Each segment is a cell with many mRNAs. Cells can be clustered using single-cell analysis into transcriptionally distinct subpopulations or cell types. We imaged across tissue layers in addition to many fields of views.

We plan to use MERmaid (developed by Jean Fan) to go beyond spatial transcriptomics and visualize our single-cell RNA-sequence data in 3D tSNE space [18]. In this tool, each point is an mRNA and mRNA species are represented by different colours. Imaging is done across many fields of views and tiled. We have a tiled image of the pre-optic hypothalamic area where users can zoom, pan, and rotate [**Figure 10**]. We hope to integrate this with MERmaid to visualise segmented cells and corresponding mRNA counts.

Methods

Multiplexed, error-robust FISH (MERFISH)

MERFISH uses a two-step labelling approach and combinatorial labelling [2]. First, multiple RNA species are labelled and imaged with the same fluorophore. The position of each label is recorded, where 1 = "on" and 0 = "off". Each RNA species encodes one of N different readout sequences probed with N hybridization rounds [15]. Second, the labels are stripped away. This process is repeated over successive rounds to reconstruct RNA types. In other words, the probes used for labelling have binding sequences and readout sequences. Sequential rounds of hybridisation target different readout sequences. Sequential rounds of imaging after hybridisation reveal RNA barcodes.



MERFISH Schematic. Reprinted with permission from *Boettiger, A and Zhuang, X. (2016)* [1-2]

Each hybridization round targets a specific readout sequence and the multiplexing capacity is 2^{N-1}. With 16 hybridization rounds, Chen et al could differentiate $2^{16-1} = 65,535$ different RNAs in theory. However, the error rate also increases exponentially with successive hybridization rounds. To correct for this, Chen et al applied a modified Hamming distance error correction code with a minimum distance of 4 [2,15]. Only a fraction of the 65,535 different RNAs met the minimum Hamming distance, limiting the multiplexing capacity. The MERFISH approach was further optimised by introducing chemical cleavage of fluorophores and multi-colour imaging (8 hybridisation rounds with 2 colour channels = 16) to reduce the number of hybridisation rounds and thus the imaging time [6,15-16].

Code

All code is available at: https://github.com/isogailab/stella/tree/master/SLURM_scripts

References

1. Boettiger, A. N. (2016, October 7). MERFISH. Retrieved from: http://alistairboettiger.info/610

2. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science, 348(6233).

3. Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez, J. D., ... & Zhuang, X. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science, 362 (6416).

4. Reuell, P. (2018, November 1). Cellular atlas of brain region leads to discoveries. Retrieved from: https://news.harvard.edu/gazette/story/2018/11/key-region-of-brain-examined-through-cellular-atlas/

5. Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., ... & Guennel, T. (2011). Spatio-temporal transcriptome of the human brain. Nature, 478(7370), 483.

6. 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.

7. Matsuda, A., Schermelleh, L., Hirano, Y., Haraguchi, T., & Hiraoka, Y. (2018). Accurate and fiducial-marker-free correction for three-dimensional chromatic shift in biological fluorescence microscopy. Scientific reports, 8(1), 7583.

8. Czech, E. (2019). TensorFlow Deconvolution for Microscopy Data. https://github.com/hammerlab/flowdec

9. Moffitt, J. & Zhuang, X. (2016). MATLAB software for the analysis of MERFISH data and the construction of MERFISH probes. https://github.com/ZhuangLab/MERFISH_analysis

10. Babcock, H. (2017). 3D-daoSTORM: Software for analyzing high-density 3D STORM data using a fitting based approach. https://github.com/ZhuangLab/storm-analysis

11. Chan Zuckerberg Initiative. (2018). starfish: Standardized pipelines for image-based transcriptomics. https://github.com/spacetx/starfish

12. H. Kirshner, F. Aguet, D. Sage, M. Unser, 3-D PSF Fitting for Fluorescence Microscopy: Implementation and Localization Application, Journal of Microscopy, vol. 249, 2013.

13. Vicidomini, G. (2008). Three-Dimensional Image Restoration in Fluorescence Microscopy. 10.13140/RG.2.1.1784.0728.

14. Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., ... & Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 1.

15. Strell, C., Hilscher, M. M., Laxman, N., Svedlund, J., Wu, C., Yokota, C., & Nilsson, M. (2019). Placing RNA in context and space–methods for spatially resolved transcriptomics. *The FEBS journal*, 286(8), 1468-1481.

16. 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*, 113(50), 14456-14461.

17. Moffitt, J. R., & Zhuang, X. (2016). RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). In Methods in enzymology (Vol. 572, pp. 1-49). Academic Press.

18. Fan, J. (2019). WebGL-based viewer for spatially-resolved transcriptomics data. https://github.com/JEFworks/MERmaid

Figures

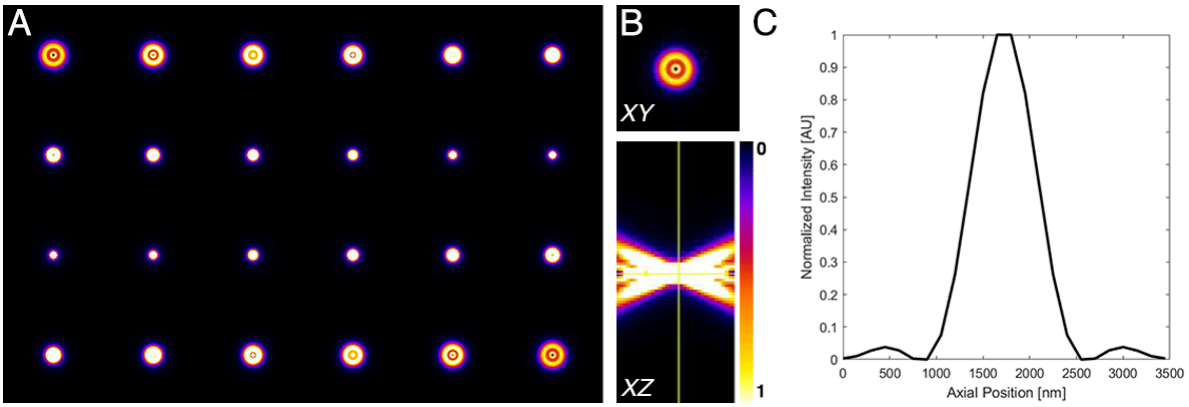


Figure 1: **Born and Wolf Model of PSF [561 nm].** Theoretical point spread function of a wide-field microscope. Different representations can be used to show a 3D PSF. **(A)** Montage of through-focus series. Sections at 150 nm z-intervals are shown. Section 20 at the top-right corner represents the central plane, which coincides with the plane of sharp focus. **(B)** Lateral and axial views. If the origin of the Cartesian reference system is placed on the centre of the max of the PSF, the lateral view is the plane such that z = 0 and the axial view is the plane such that y = 0. **(C)** Axial (z = 0 and y = 0) intensity profile.

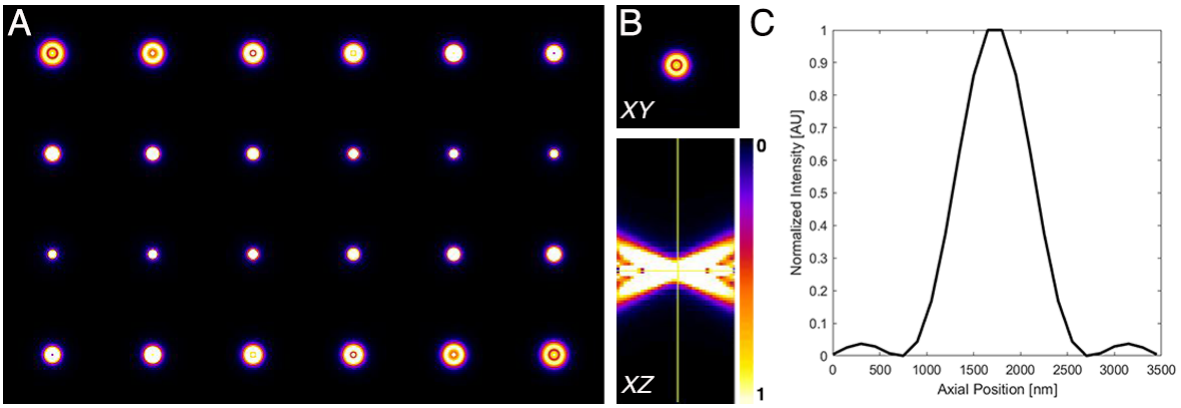


Figure 2: **Born and Wolf Model of PSF [647 nm].** Identical to **Figure 1** but for [647 nm]. **Note**: **Figures 1-2** descriptions are based on *Vicodomini, 2008* [13].

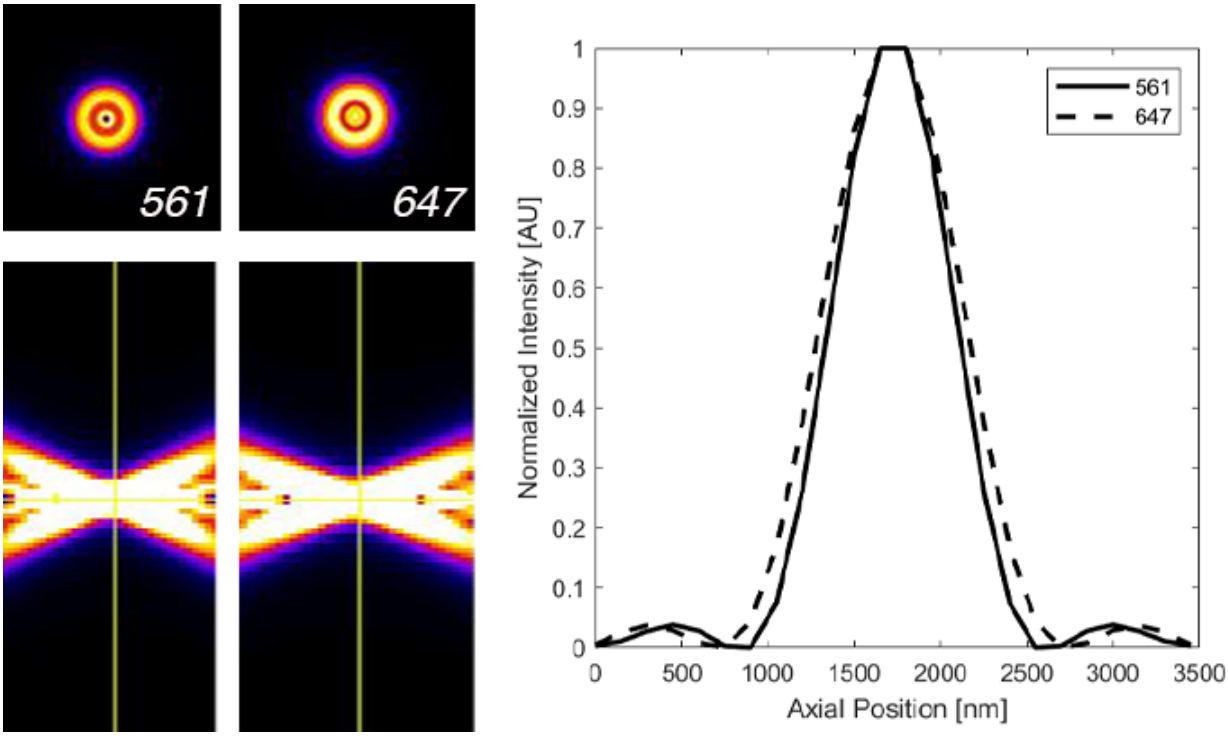


Figure 3: **Comparison of Theoretical PSF for [561nm] and [647nm]. (Left) Lateral and axial views of theoretical PSF for [561 nm] and [647 nm]. (Left)** Lateral and axial views of theoretical PSF for [561nm] and [647nm]. **(Right)** Axial (z = 0 and y = 0) intensity profiles. The FWHM is greater for [647 nm] than 561 nm], which is expected. Since [647 nm] has a broader emission spectrum than [561 nm], it has a higher spectral width.

Name	Lateral FWHM [µm]	Axial FWHM [µm]
Ideal Widefield	0.22	0.56
Measured Bead [561 nm]	4.13	0.53
Measured Bead [647 nm]	3.30	0.42
Aggregate Bead [561 nm]	4.05	0.84

Table 1: **Full-width-half-maximum (FWHM) Comparison in Experimental PSF.** FWHM comparison between [561 nm] and [647 nm] experimental PSFs in wide-field microscope. The FWHM was found by taking the maximum of the transmission peak and identifying the points on either side of the peak where the transmission is half the maximum. The full-width is the distance along the wavelength axis between those two points.

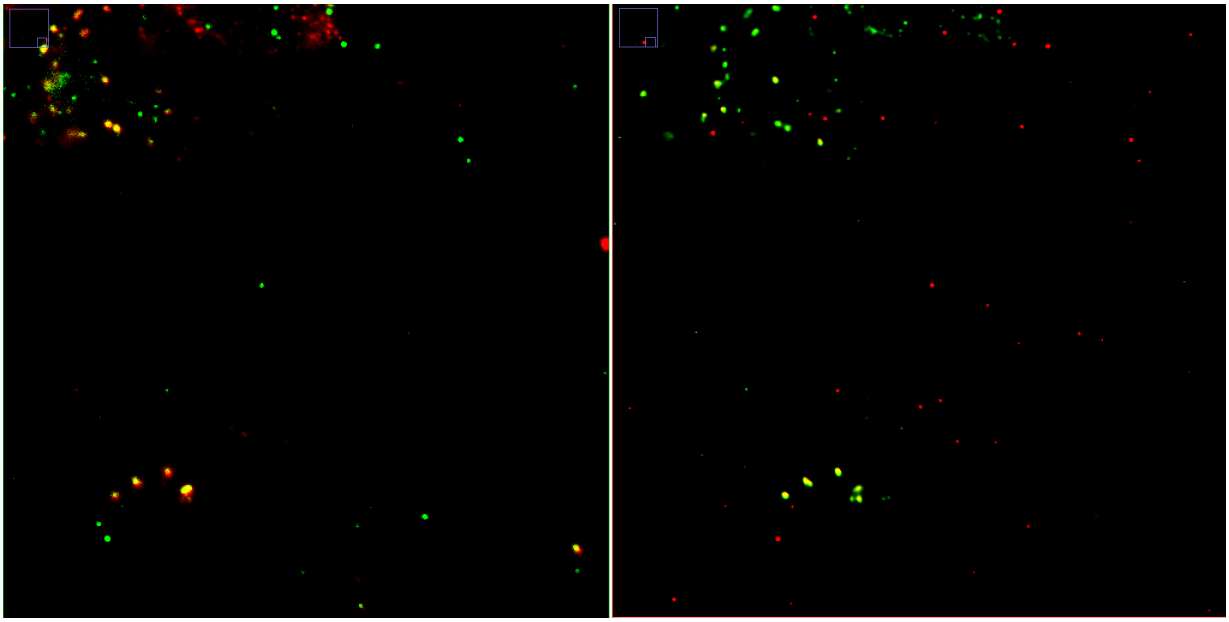


Figure 4: **Chromatic Aberration Correction. (Left)** Pre-processed and **(Right)** Post- processed frame of 2048 x 2048 x 40 stack that illustrates chromatic aberration correction. Each dot represents an RNA molecule. The two-colour channels (green = 561nm; red = 647nm) represent Cy3 and Cy5 staining of one hybridization round in MERFISH, respectively. Note the misaligned red and green channels in the left image (fringe) that is gone in the right image.

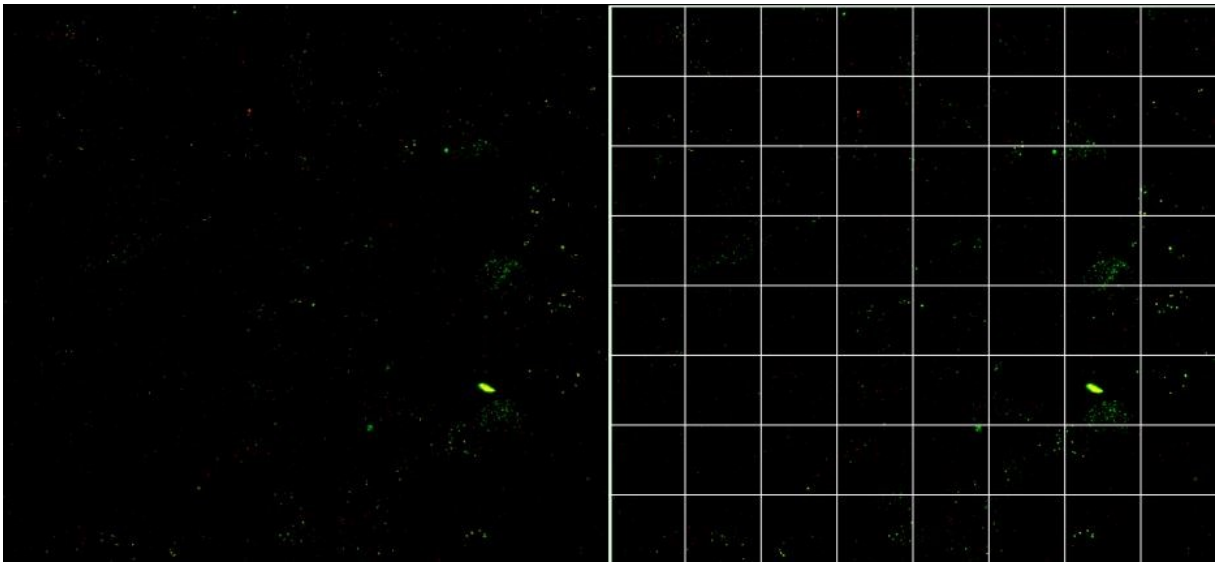


Figure 5: **3D Sparse Deconvolution. (Left)** Single frame of 3D sparse deconvolution applied to data described in **Figure 4. (Right)** Schematic of sparse deconvolution: a 2048 x 2048 input is divided into squares of 256 x 256 to identify any squares that do not contain any dots (black background). These squares are not deconvolved, decreasing computation time. Note that in dense images this will have no effect as no squares will be removed.

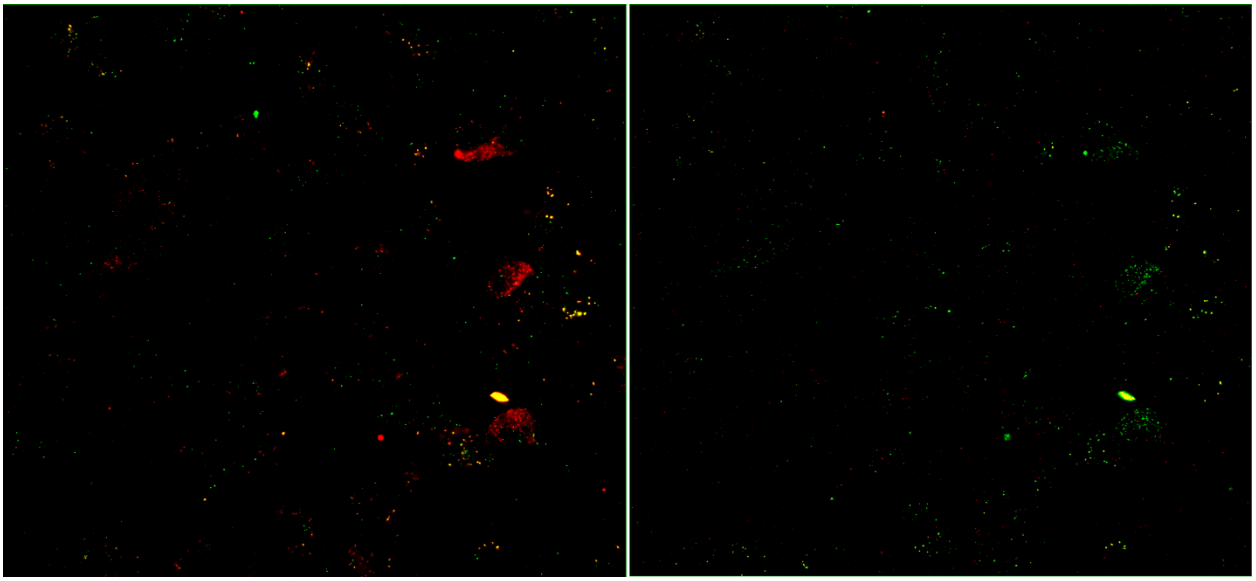


Figure 6: **Pre and Post-Processed Data. (Left)** Pre-processed and **(Right)** Post- processed frame of data described in **Figure 4** following: illumination correction, chromatic aberration correction, and 3D sparse deconvolution. Compared to the left, the right image is much sharper and clearly labels single molecules in dense regions.

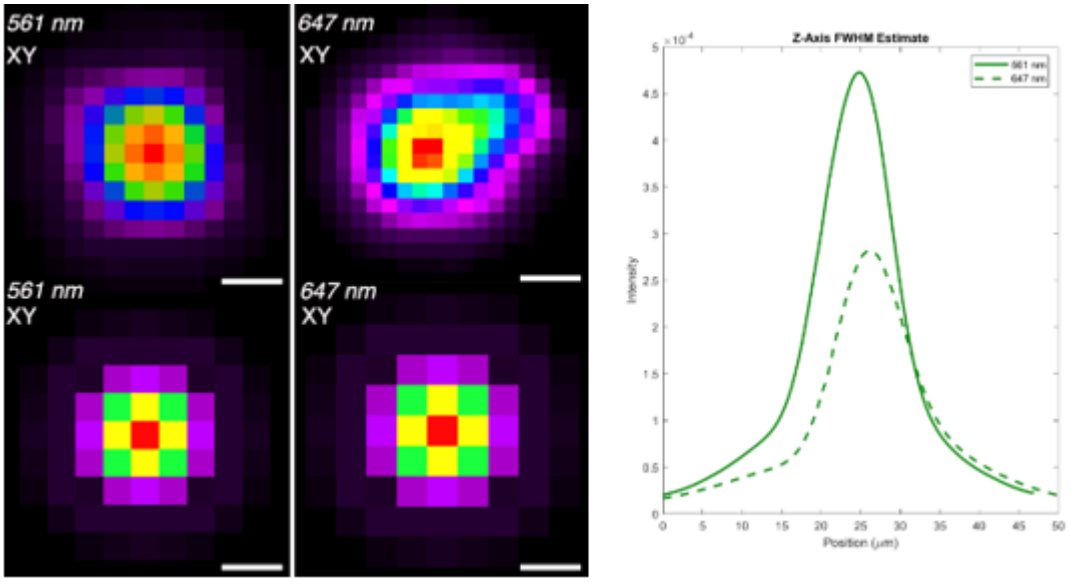


Figure 7: **Pre and Post-Processed Data. (Left)** Experimental PSF found by measured beads in [561nm] and [647nm] on top. The top-left corner PSF is less circular because it was done on an aggregate bead (2 beads). Theoretical PSF found by Born and Wolf model in [561nm] and [647nm] on bottom. Note that the deviations from ideal in experimental PSFs are representative of the microscope optics. **(Right)** Axial ($z = 0$ and $y = 0$) intensity profile of experimental PSFs. The FWHM for [647nm] is less than that of [561nm], which is incorrect (See **Table 1**). This may be due to using an aggregate bead for [647nm].

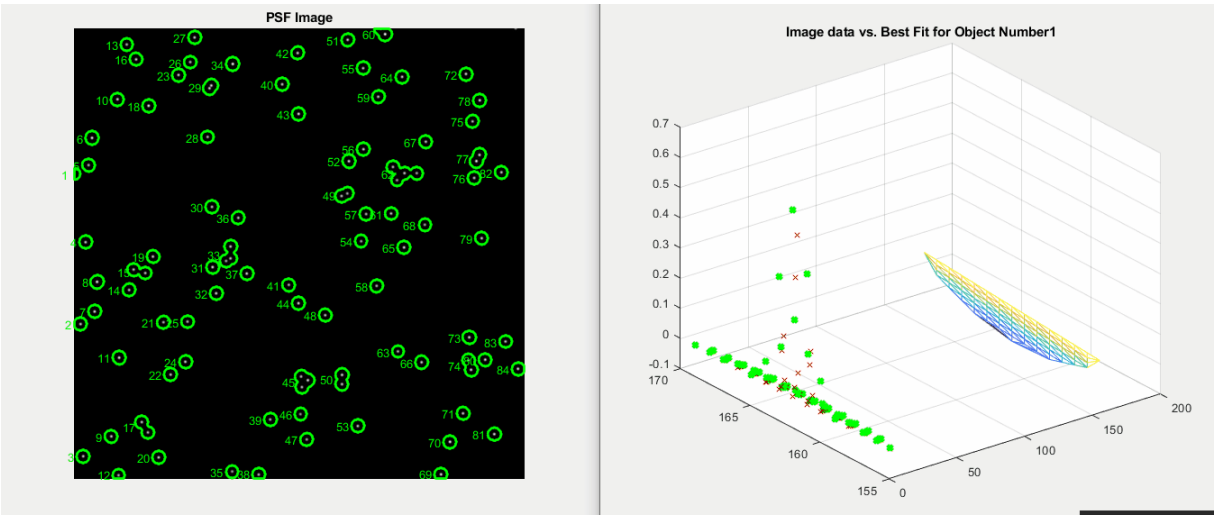


Figure 8: **Pre and Post-Processed Data. (Left)** Synthetic DAPI image generated by providing inputs for microscope optics (i.e. numerical aperture, magnification, and electrons per count) where green circles correspond to different beads **(Right)** PSF fitting of a bead in left image.

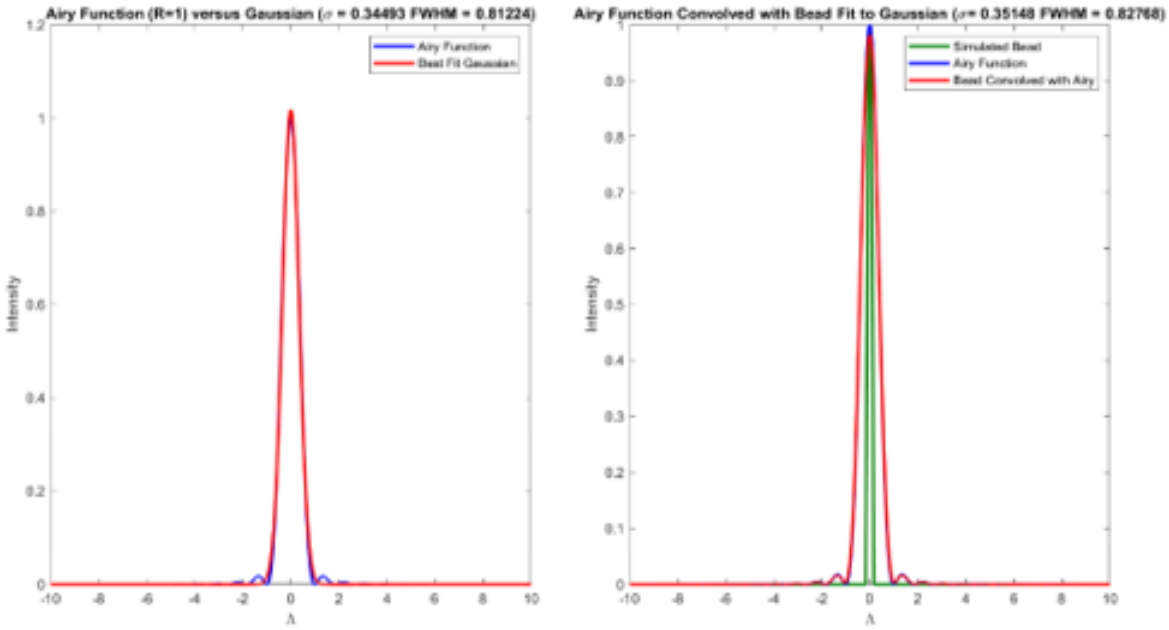


Figure 9: **Pre and Post-Processed Data. (Left)** Simulated airy disk (blue) with best-fit Gaussian (red). **(Right)** Simulated airy disk (blue) convolved with bead fit (green) to Gaussian (red). Note these simulations can be used to simulate common optics issues such as chromatic aberration (see **Figure 4**) or explore the effect of bead size on the PSF.

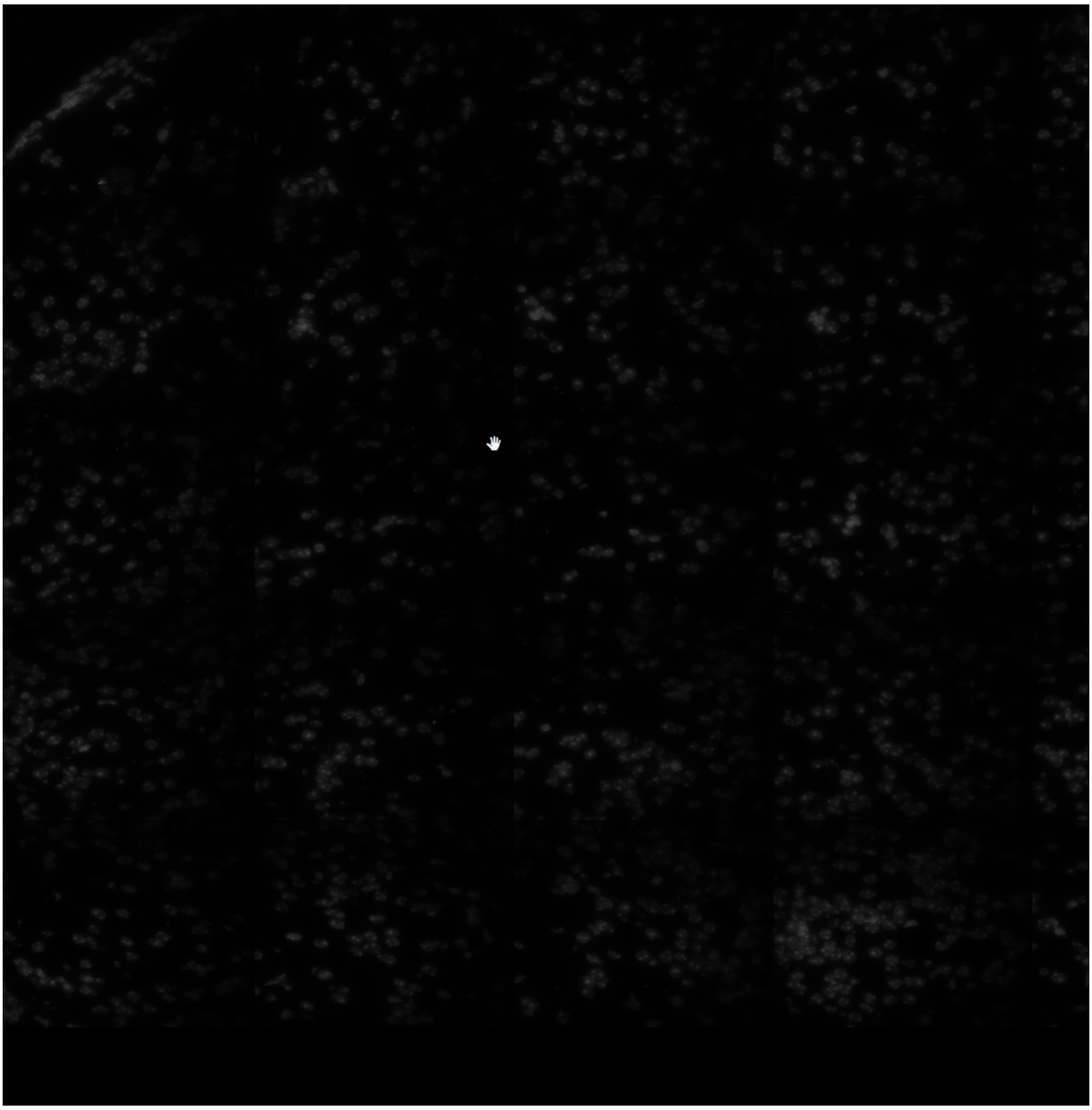


Figure 10: **Development of Data Visualisation.** Tiled images of pre-optic hypothalamic area (4 x 5). In-progress data visualisation development where zoom, pan, and rotate functions are available. We plan to overlay segmented cells and mRNA counts. We used JavaScript and the D3 plug-in.