Resolving the tissue topology of the Alzheimer's brain

Shimmy Balsam

Advisor: Dr. Naomi Habib

Abstract

Alzheimer's disease (AD) is a neurodegenerative disease known to damage and degenerate neural cells. Previous research has shown crucial changes which occur to other surrounding non-neuronal cells as well, specifically, astrocyte cells. A new disease associated astrocyte (DAA) state was found in AD mouse models, by a recent research done with Single Nucleus RNA Sequencing (sNuc-seq). Yet much remains unknown regarding the role of DAAs in AD.

Single Cell/Nuclei experiments, and analysis lack the spatial information, the location of cells within the tissue and their cellular microenvironment, since the tissue is dissociated into single cells. To better understand the cellular neighborhoods of DAAs, their localization and cell-cell interactions in the AD brain, we use image analysis of Fluorescent In-Situ Hybridization (FISH) for RNA and Immunohistochemistry (IHC) for proteins. We built a pipeline for cleaning the images and retrieving both the amount and location of cells and the number of RNA molecules of the genes expressed per cell. Analysis of the FISH images showed another way to quantify the differences between AD and WT mice using marker cells present in different astrocyte states, complimentary to the sNuq-seq method and validating its results. The analysis of the IHC images of DAAs markers and AD pathological hallmark Amyloid Beta (Aβ) plaques, revealed that there is a significant change in the proximity and density of Gfap, a marker of reactive astrocytes, to and around Aβ plaques in AD mouse brains.

Introduction

Alzheimer's disease (AD) is an age-related neurodegenerative disease affecting cognitive and behavioral practice. While AD causes remain undetermined, much research is underway to better understand the disease mechanism. AD is known to damage neural cells and a pathological

hallmark of AD is the appearance of beta-amyloid (Aβ) plaques outside the neurons (Alzheimer's Association, 2019).

One area of the brain particularly vulnerable to known AD symptoms such as neuronal damage is the hippocampus (Holtzman, D.M., et al., 2011; Fig 1a). The hippocampus is a key contributor to navigation skills and to both short and long-term memory. Within the hippocampus lays the Dentate gyrus (DG) which too is known to be linked to memory formation. The DG is easily visible and distinguishable using microscopic and fluorescent methods, due to its unique shape (Fig 1b).

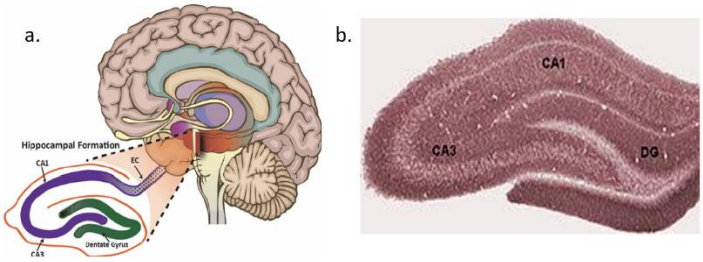

Fig 1. Hippocampus and Dentate Gyrus (DG). a) Schema and representation of the Hippocampus area within the brain (Vineyard C. M., et al., 2012). b) Visualization of the DG and its surrounding area within the Hippocampus.

An important undefined impact is the effect that different cell types within the brain and their cell-cell interactions have on AD progression (Moor A.E., et al. 2017). Previous research has shown that crucial changes occur to other nearby cells besides neurons, such as astrocytes (Habib N., et al., 2017; Strooper, B.D. and Karran, E., 2016).

Astrocytes are a type of glia cells and are the largest and most common cells in the central nervous system. They are located in the brain and spinal cord and play multiple roles such as: biochemical support of endothelial cells that build the blood-brain barrier, supply of nutrients to nerve tissue, preservation of extracellular ion balance, and play a role in the repairing and scarring processes after a trauma injury (Strooper B.D. and Karran E., 2016; Attwell D., et al., 2010). Despite their importance in maintaining brain homeostasis, their connection to AD has not been studied much.

Single Cell/Nucleus RNA sequencing are methods in which we can dissociate a tissue and profile the different cells based on each individual cell/nucleus gene expression. A recent sNuc-seq based research done at the Habib lab has shown numerous changes amongst astrocytes in AD and has defined a new state of disease associated astrocytes (DAA). They have defined three states in astrocyte cells, two found in WT and AD mice, *homeostatic* and *active* states, and the third, *DAA*, is unique to the AD brain. The DAA state is defined by high expression of the *Gfap* gene, together with several other marker genes which appear/increase in the disease state, such as *Vim*, *Osmr*, and *Slc38a1*. Habib *et al.* also showed the difference between the DAAs to the Homeostatic state which is *Gfap* negative and shows an increase in the marker gene *Mfge8*. Whereas the Active state also highly expresses *Gfap* and the marker genes *Myoc* and *Slc38a1* (Habib N. et al., 2020; Fig 2).

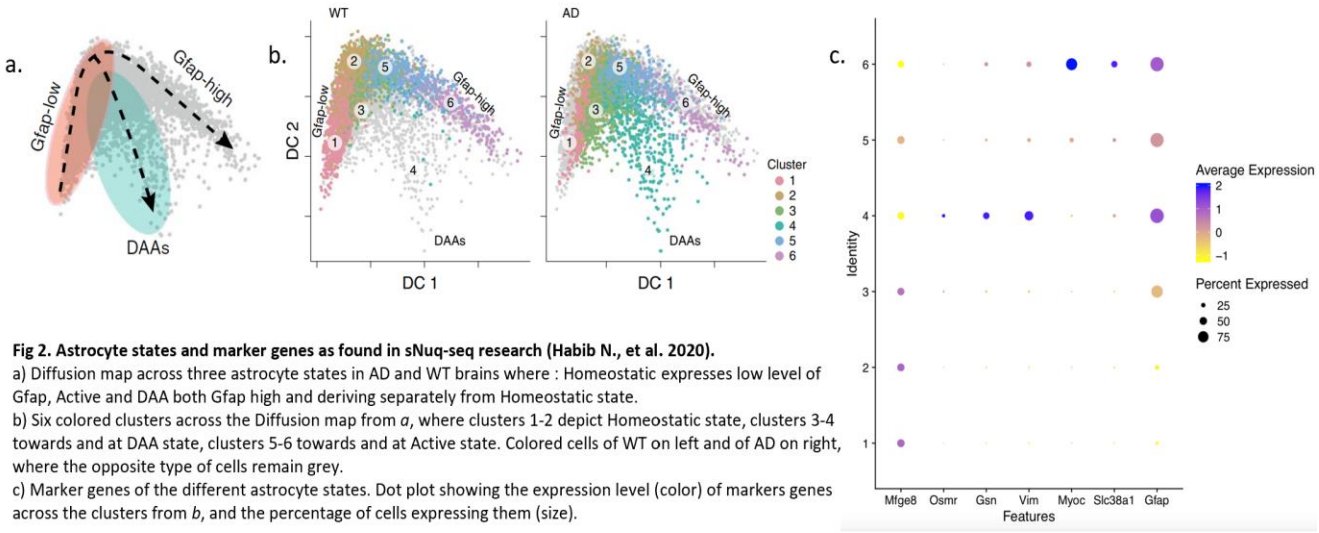

Fig 2. Astrocyte states and marker genes as found in sNuq-seq research (Habib N., et al. 2020).
a) Diffusion map across three astrocyte states in AD and WT brains where : Homeostatic expresses low level of Gfap, Active and DAA both Gfap high and deriving separately from Homeostatic state.
b) Six colored clusters across the Diffusion map from *a*, where clusters 1-2 depict Homeostatic state, clusters 3-4 towards and at DAA state, clusters 5-6 towards and at Active state. Colored cells of WT on left and of AD on right, where the opposite type of cells remain grey.
c) Marker genes of the different astrocyte states. Dot plot showing the expression level (color) of markers genes across the clusters from *b*, and the percentage of cells expressing them (size).

However, single cell/nucleus RNA-sequencing data is missing the spatial information regarding cells, i.e. their position within the tissue and their cellular neighbors, seeing that it is retrieved from tissues which dissociated to single cells.

Single molecule mRNA Fluorescent in Situ Hybridization (smFISH) is a method to enable visualization of gene expression amongst cells (Frickmann H., 2017), usually allowing the staining of up to 3 different colored probes and as a result marking up to 3 different genes, per tissue (Fig 3a). Applying this method can provide imaging and quantification of multiple genes in their original position within both WT and AD brains.

Immunohistochemistry (IHC) is a method to identify specific proteins within tissue sections using a primary antibody specific to the antigen. In addition, a secondary antibody specific to the primary antibody labeled with a fluorescent marker is added (Clifton P.D. and Wun-Ju S. 2015; Fig 4a). IHC can be used to detect target proteins, allowing protein staining and localization (Maity B., et al. 2013).

The FISH imaging data can be combined with the single nucleus RNA-seq data to connect to known cellular populations, analyze cellular neighborhoods within the brain and specific populations of interests and their neighboring cells. Each FISH image data set contains two FISH images, one from an AD mouse and one from WT, taken around the DG subregion of the hippocampus brain region. Each of these images contains nuclei with a blue fluorescent signal and three genes out of seven marker genes found in the previous Single Nucleus research, each with a specific fluorescent signal of either red, green or cyan. Each data set was created in parallel, with the same 3 target genes in both WT and AD images (Fig 3b-g).

The IHC imaging data provides further information on the protein level, allowing spatial analysis of the target proteins together with Aβ plaques, which are detected on the protein level, and can not be detected by FISH . Each IHC image shows GFAP stained in green (Fig 4b) and Aβ plaques stained in red (Fig 4c), taken from AD mice only.

The spatial analysis can validate the single nuclei data and also advance our understanding of the cell-cell interactions and their effect on each cell's functionality and disease progression. A spatial map of such neighborhoods can hopefully shed light on much more of the causes and aspects of AD.

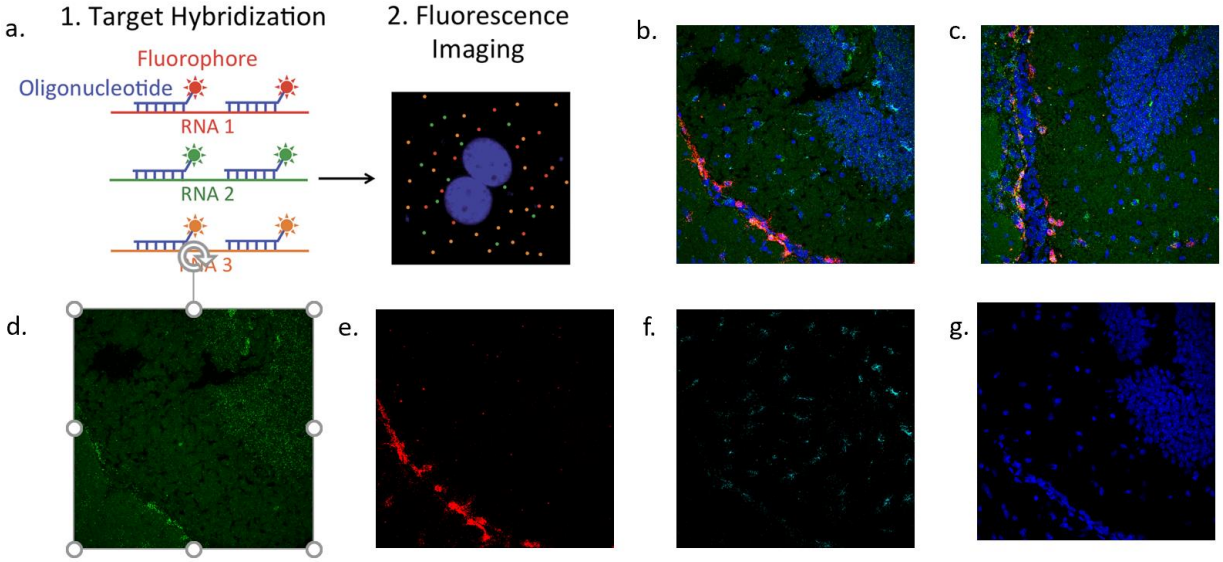

Fig 3. smFISH method and implementation on WT and AD mouse brains with targeted genes
a) Schema of the fluorescent hybridization to cells and mRNA strands of three different genes. Resulting in a microscopic image of each cell and target gene dyed in their representative color.
An example of the smFISH method on b) AD and c) WT mouse brain with DAPI (Nuclei) stained in blue and targeted genes Slc38a1 (Active/DAAS marker) in green, Myoc (Active marker) in red and Mfge8 (Homeostatic marker) in cyan.
AD Mouse single channel representation of d) green channel with only Slc38a1 (Active/DAAS marker), e) red channel with only Myoc (Active marker), f) cyan channel with only Mfge8 (Homeostatic), g) blue channel with only DAPI (Nuclei).

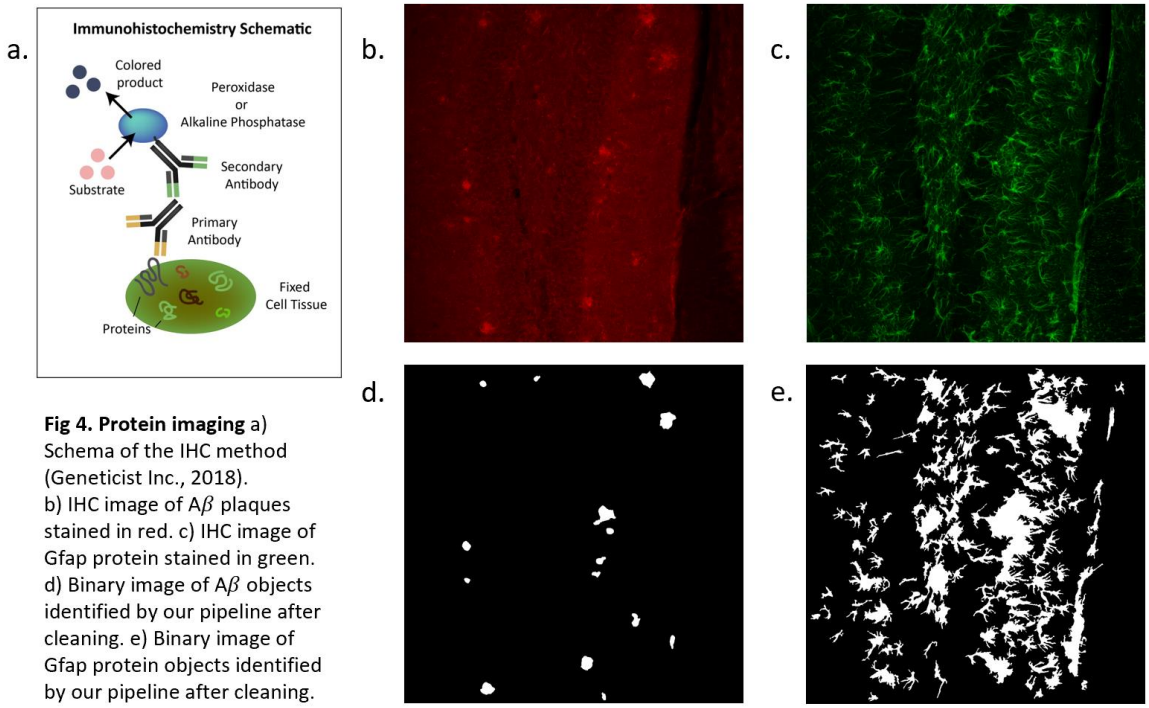

Fig 4. Protein imaging a) Schema of the IHC method (Geneticist Inc., 2018). b) IHC image of Aβ plaques stained in red. c) IHC image of Gfap protein stained in green. d) Binary image of Aβ objects identified by our pipeline after cleaning. e) Binary image of Gfap protein objects identified by our pipeline after cleaning.

Methods

Image processing: noise filtering, cell identification, quantification

The noise filtering and signal identification algorithms were executed using CellProfiler (version 4.0), a biological image analysis software.

Images stained using fluorescent signaling tend to be noisy, where some of the noise may be due to biological diversity, yet most is technical noise introduced during the experimental pipeline and the fluorescent imaging. To reduce such noise before image analysis we first filter the images with the following steps: first, to reduce general smudging we blur the image using the Cell Profiler module *CorrectIlluminationCalulate* with *GaussianFilter* as the smoothing method parameter, followed by *CorrectIlluminationApply* module with *subtract* as the illumination function parameter. Gaussian filter blurs the image by changing each pixel to a weighted average of the pixels near it, based on a normal distribution (Fig 5a-c). *Subtract* deducts the result of the Gaussian calculation from the original image (Fig 5d). Next, we remove the remaining fine noise using Non-Local Means (NLM) Algorithm via Cell Profiler's *ReduceNoise* module (Fig 5e). This algorithm normalizes each pixel as well, this time as a weighted average of all pixels in the image, where the weight of each pixel is defined by how similar that pixel is to the target pixel. Formally, for each

target pixel p with intensity value $v(p)$, we calculate its normalized value to be $u(p) = \frac{1}{C(p)} \Sigma_q v(q) f(p,q)$ where $C(p) = \Sigma_q f(p,q)$ and $f(p,q)$ is the similarity function between pixels p, q.

FISH images required both noise filters to be applied given the high level of noise. For IHC images we only applied NLM without using the gaussian filter beforehand. Additional noise reduction was done as part of the identification of cells and mRNA molecules, by applying a threshold over the size and intensity.

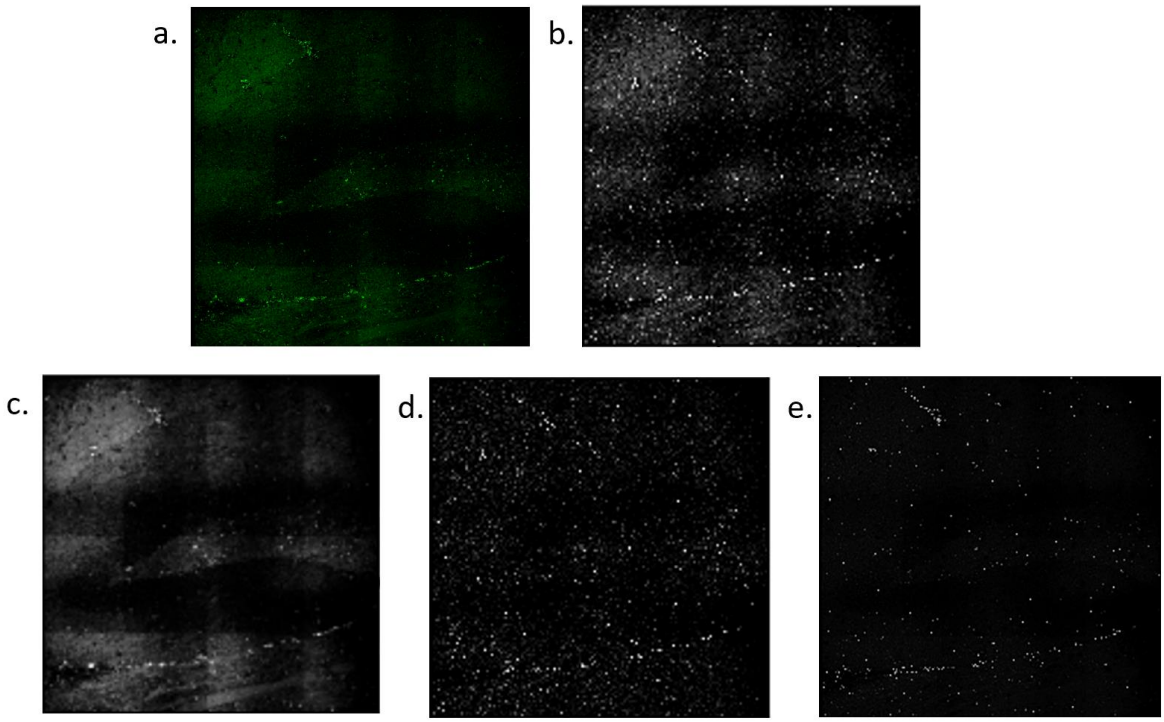

Fig 5. Image preprocessing – noisy image cleaning using Gaussian Filter and NLM algorithm
a) original single channel image of Osmr target gene stained in green, together with much background cell-less tissue stained in green due to technical noise cased during fluorescent imaging. b) A greyscale version of original Osmr green channel. c) Blurred greyscale image as a result of Gaussian Filter used by CellProfiler's *CorrectIlluminationCalulate* module. d) First level cleaning of greyscale Osmr image, as a result of CellProfiler's *CorrectIlluminationApply* module which subtracted the blurred image *c* from the original greyscale image *b*. e) Clean image, after applying second level of cleaning with NLM algorithm, using CellProfiler's *ReduceNoise* module, on the first level cleaned image *d*.

Given the cleaned FISH and IHC images we next needed to identify and quantify cells, mRNA molecules and proteins. Finally, each mRNA molecule or protein needs to be assigned to the cell from which it originated. Cells segmentation is done in two stages:

First, we find the cell nucleus, using the Cell Profiler module *IdentifyPrimaryObjects* (with parameters: *Object Diameters range = 30-80, Threshold Strategy=Global, threshold method=Otsu, smoothing scale=1.3, correction factor=1.5, Distinguishing clumped objects= Intensity)*, which finds the nucleus based on intensity difference between object and background pixels using a global threshold across the entire image. The threshold is learned from the data with the *Otsu* method which will optimize the threshold that minimizes the variance in both groups. Differentiating between neighboring nuclei and classifying ambiguous pixels to the correct nucleus is done by calculating the intensity distances to each nucleus' intensity center of mass together with weighing the surrounding pixels affiliation (Fig 6a-b).

The second stage is finding the area of the cell, which is a non-trivial task (see discussion), which we decided to do by expanding the nucleus area by a defined radius. This was done using Cell Profiler's *IdentifySecondaryObjects* module with *identification method= Distance-N* and *number of expansion pixels=15,* as parameters. This module takes a radius of size *r* (*r* is given as a parameter, in our case we defined *r*=15) and defines all pixels within the radius around the already defined nucleus, as the cell area (Fig 6c). Ambiguous pixels due to multiple neighboring cells' proximity are classified in a similar manner as done in the previous module in which we found the nuclei.

Identifying mRNA molecules to cells is done using Cell Profiler's *IdentifyPrimaryObjects* module, same as in the nucleus, only with a smaller size diameter range of 10-40, for noise reduction (Fig 6d-f). Quantification of the mRNA molecules of each gene per cell was done using Cell Profiler's *RelateObjects* module. This module counts the amount of mRNA molecules identified which fall in the area of each specific cell, then assigns each of these molecules as children of their now parent cell. Resulting in a count of how many mRNA molecules each cell holds, per gene (Fig 6g). Furthermore, this module outputs the location of each of the "parent" cells and that of each of their mRNA "children" molecules. All calculations are outputted to a *.csv* file using CellProfiler's *ExportToSpreadsheet* module.

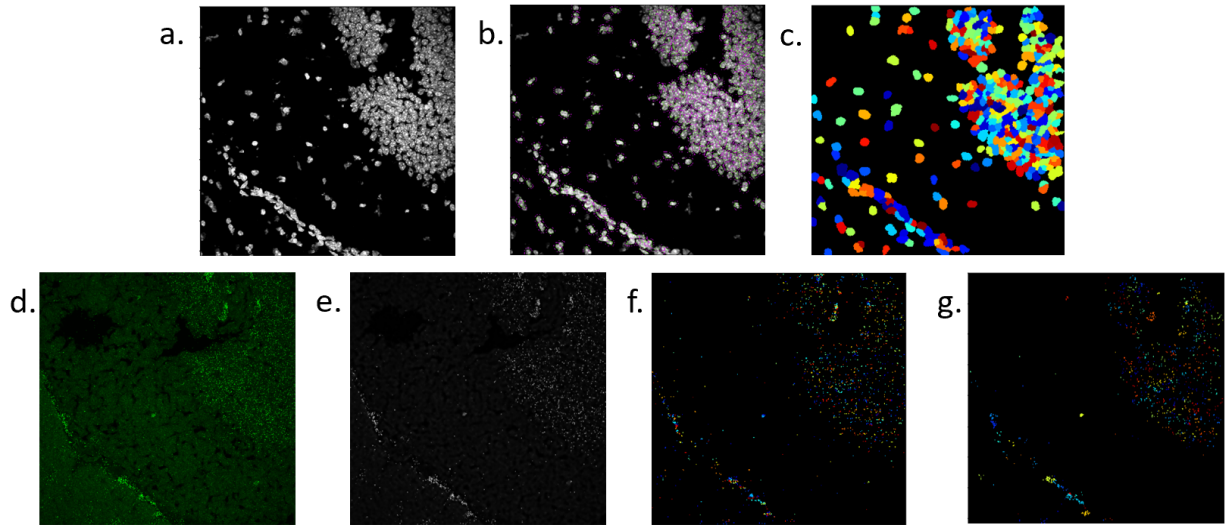

Fig 6. Identification of cells and mRNA molecules a) Greyscale image of DAPI (nuclei) single channel from AD brain originally taken with target genes Slc38a1, Myoc and Mfge8. b) Result of CellProfiler's *IdentifyPrimaryObjects* module in which each nucleus from *a* is identified and objectified separately. c) Result of CellProfiler's *IdentifySecondaryObjects* module in which each cell is identified and objectified separately, based on the primary nuclei objects found in *b,* where a radius of given *r* was added from each pixel of the nuclei circumference. d) Original green channel of Slc38a1 target gene from same AD data set as *a,* before cleaning. e) clean version of *d* after applying Gaussian filter and NLM cleaning methods as part of pipeline. f) Result of Cell Profiler's *IdentifyPrimaryObjects* module in which each mRNA molecule from *e* is identified and objectified separately. g) Result of Cell Profiler's *Relate Objects* module in which each mRNA molecule which was identified in *f* is defined as a child object of a specific cell from *c,* as it is located within the cell's area.

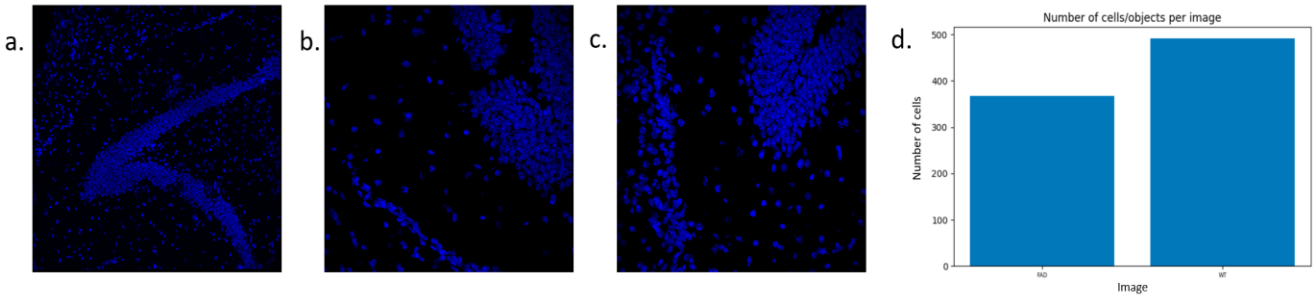

Fig 7. Cell count differentiation between images a) Dapi channel image of AD mouse taken from collective image with the targeted genes Osmr, Myoc and Mfge8. This Dapi image shows a larger area of the DG with many more cells in the image, yet smaller as it is more zoomed out, than in b) AD Dapi channel of AD mouse and c) WT mouse both taken from collective image with targeted genes Slc38a1, Myoc and Mfge8.
d) Bar plot depicting the difference of cell count between images *b* (AD) and *c* (WT). Although the images seem to be similar in resolution, as opposed to that of image *a*, the bar plot shows a difference of over 100 cells in between the images.

Analysis and Single Nuclei RNA-seq validation: single double and triple counts, correlation

We quantify gene expression in several different ways, for single genes, pairs and triplets and compare these values for AD and WT mice brain slices. For each gene, we quantify the **expression intensity,** by counting the sum of its molecules in both AD and WT images. Next, we quantify the **number of cells expressing the gene**, by counting the number of cells which exhibit at least 1 molecule of this gene, normalized by the total number cells per image (see discussion; Fig 7), and compare.

For each pair of genes, we quantify the number of cells expressing the genes, by counting the number of cells which exhibit at least 1 molecule of each gene in WT and AD, normalized by the total number cells per image.

To test the relation between two genes, we calculate the correlation between the two genes, using a linear regression prediction and $pearson's\ r$ which is the correlation coefficient, calculated by the covariance of the two variables divided by the product of their standard deviations.

Finally, we count the number of cells which exhibit at least 1 molecule of all 3 genes, normalized by the total number cells per image, and compare.

Spatial Analysis of general localization and distribution

To find the localization of genes and cellular populations within the tissue structure, for each FISH image we quantified the intensity level of each gene, by dividing the image to a 16X16 grid where in each block of the grid we sum the intensities of all the pixels within that block's area, in the original image. This is done on each gene separately within each image. Python's *seaborn.heatmap* function is then applied on the grid to visualize the calculation throughout the image area..

Protein co-localization: density and minimum distance

Using the clean IHC images (as explained above) as input, we applied Cell Profiler's *IdentifyPrimaryObjects* module, same as in the FISH images, then saving them as binary images where each pixel in an object (of either GFAP or Aβ plaques) has an intensity equals to 1 while background pixels have intensity which equals to zero.

To calculate the co-localization of proteins in the binary version of the IHC images we calculated and compared between the density of GFAP positive pixels in areas near plaques vs. the density of GFAP pixels in areas without plaques. The plaque areas were found by going over the Aβ plaques image twice, first to find a radius R as an identifier for the size of the plaques' areas – it

was calculated as the average of all the counts of consecutively positive pixels in the plaques image. Second time, we defined each plaque area to be a square taken from the left corner above the intensity mass center location, given from CellProfiler's pipeline output. We define a square area as, an area that holds $4R^2$ pixels. We calculated those squares' pixels in Gfap image and counted the number of positive Gfap pixels divided by sum of pixels in that area.We also calculated the density in all squares in which the center of mass of any plaque does not fall. Finally, we compared between the densities of square areas with and without plaques, using Kolmogorov–Smirnov (KS) test to find the $p - value$ of colocalization of GFAP and Ab-plaques. KS is a non-parametric test comparing between two distributions, without assuming that any of the values come from a normal distribution. It was applied using python's *scipy.stats.ks_2samp* function.

In order to test if GFAP is significantly in proximity to the Aβ plaques, minimum distance between plaques and GFAP was calculated and then compared to both normal and random distributions as follows: For each plaque pixel we found the closest GFAP pixel to it. We did so by finding all the x, y values of pixels which are positive in GFAP and in plaques and then calculate for every x,y GFAP pair its distance from all plaque X, Y pairs, choosing the minimum distance found. The equation is as such: $d = \sqrt{(x - X)^2 + (y - Y)^2}$.
To calculate the statistical significance of the distance between GFAP and plaques, we plot the distribution of all the minimal distances between plaques and GFAP and compare it to a Normal distribution. Next, we randomly define spots and areas in the image as fake plaques and calculate the minimum distances from them to GFAP and see their distribution as well, defining this as the random distribution. We compare between the distribution of measured distnaces from GFAP to plaques and the normal and random distributions using KS test to find $p - value$, again using python's *scipy.stats.ks_2samp* function.

Results

A new computational pipeline for filtering, segmentation and quantification of FISH images

Measuring the localization and quantification of RNA molecules *in situ* by FISH images based on cell area is highly challenging as only the nuclei are stained. As such a new pipeline had to be created. We thus developed such a pipeline which succeeded in this very specific type of quantification. While we applied a simplified scheme to infercell areas, our pipeline still provides relatively detailed information on gene expression with single cell resolution, and is more precise

than a general image analysis based on a grid of general areas, blind to the location and the shape of each specific nucleus. Even more so, our pipeline can handle noisy images and retrieve sufficient data from them (Fig 5), proved accurate by the validation of sNuc-seq results by comparing results together with manual eye inspection and previous knowledge about the hippocampus and gene patterns. Finally, running the pipeline on IHC images called for minimal changes, showing that this indeed is a robust detection and quantification pipeline for different types of Bio-stained images, as needed.

In situ RNA hybridization analysis validates astrocytes states in AD and WT mice

Once retrieving the output from the image processing pipeline we start the first level of analysis in which we want to quantify and compare the gene count between AD and WT, this will also work as a validation to the results of the previous sNuc-seq research. sNuc-seq found three distinct populations of astrocyte cells, one of which a novel state showing significant increase in proportion in AD. Each population had specific gene expression patterns which we could use as marker genes for FISH. Our data is based on seven marker genes (Gfap, Osmr, Gsn, Vim, Slc38a1, Myoc, Mfge8) of these three different states, where we expect to see DAA state markers (Gfap, Osmr, Gsn, Vim, and partially Slc38a1) to be expressed higher in AD images as opposed to both Homeostatic marker (Mfge8) and Active states markers (Gfap, Slc38a1, Myoc) which we expect to have higher expression level in WT.

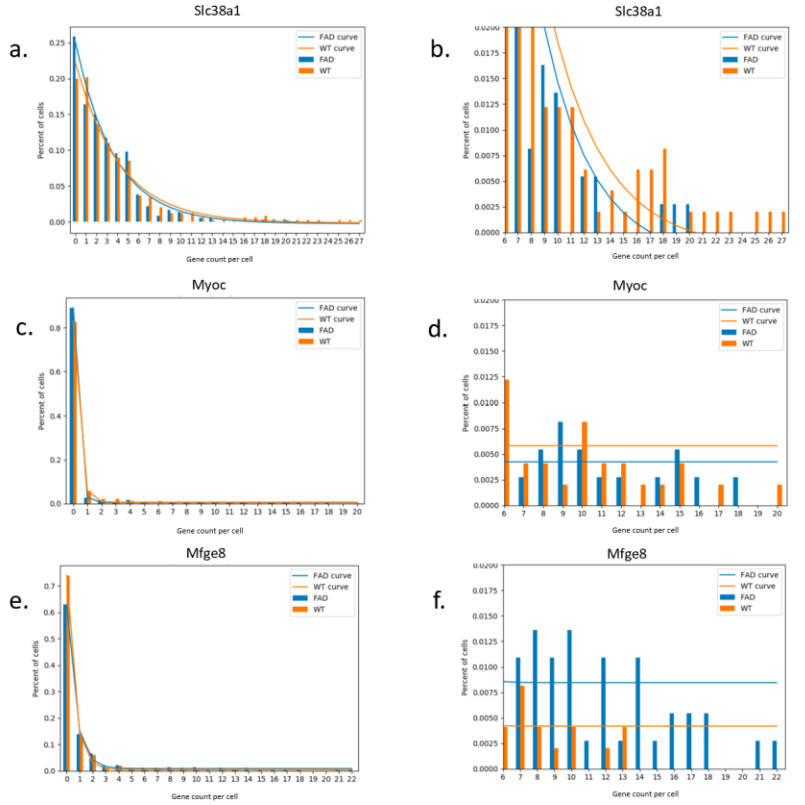

Fig 8. Gene count per cell distribution comparison between AD and WT, depicting the percent of cells which contain each specific amount of mRNA molecules per gene.

a) Comparison of the DAA/Active marker gene Slc38a1 distribution in WT and in AD, showing a near exponential decline in both AD and WT.
b) A zoom in on *a* in which we better see the differences of the trends in the higher count values, showing a slightly higher trend in WT yet quickly equalizing to zero together with AD.
c) Comparison of the Active marker gene Myoc distribution in WT and in AD, showing an immediate decline in both AD and WT.
d) A zoom in on *c* in which we better see the differences of the trends in the higher count values, showing a constant higher trend in WT.
e) Comparison of the Homeostatic marker gene Mfge8 distribution in WT and in AD, showing an immediate decline in both AD and WT.
f) A zoom in on *e* in which we better see the differences of the trends in the higher count values, showing a constant higher trend in AD.

For each data set of WT and AD images stained with the same three marker genes, we compared the percentage of expression of each gene in WT and AD and then compare to the sNuc-seq results. At a birds eye view, the distribution plots (Fig 8) show small differences between AD and WT, but when looking at the high tail of the histogram, count per cell levels, and plotting the density function of the distribution, we can see a more definitive but still subtle difference, per gene. For more accurate quantification, we calculated the percent of cells that express each gene and show a bar plot of the difference of each gene's result in AD vs. WT. While the number of samples didn't enable us to perform a statistical test, we used the fold change of $\frac{loglog\ (geneAD)}{log\ (geneWT)}$, with log base 2, to quantify the differences (fig 9a-c).

After comparison of all genes, we found that the genes Gfap, Vim and Osmr, which are all DAA markers according to the sNuc-seq results, are expressed higher in AD than in WT. Slc38a1, a marker for both DAA and Active states, is expressed near equal in AD and WT. Myoc, a physiological Active state marker, is expressed higher in WT than in AD. All in accordance with sNuc-seq results. On the other hand, Mfge8, a Homeostatic state marker, expected to be expressed higher in WT, yet in our results was higher in AD (Fig 9c). A closer examination of the expression pattern of this gene in sNuc-seq data showed that it's expressed also in DAA, although higher in homeostatic astrocytes, which could account for this discrepancy (see discussion).

Next, we wanted to test the co-expression patterns of genes. For every pair of or triplet of genes measured simultaneously on the same brain slice, we calculated the percentage of cells which express both or all three genes and compared the difference between WT and AD (fig 9d-g). We also calculated the correlation between the cells levels for every pair of genes (fig 10).Our analysis strengthened our results from the single gene count analysis, showing co-expression and higher correlation of the Active markers Myoc and Slc38a1 in WT, and the DAA markers Gfap and Vim in AD. The triplet Gfap, Vim and Slc38a1 showed higher expression in AD as well, which demonstrates Slc38a1's diversity as a marker for both Active and DAA states.

Overall, we were able to validate the sNuc-seq results using a different method of image analysis, where in return, the sNuc-seq results validate our image processing pipeline. Once validated we could now turn to test the spatial positions of each marker, to better understand their localization and the neighboring effects of AD.

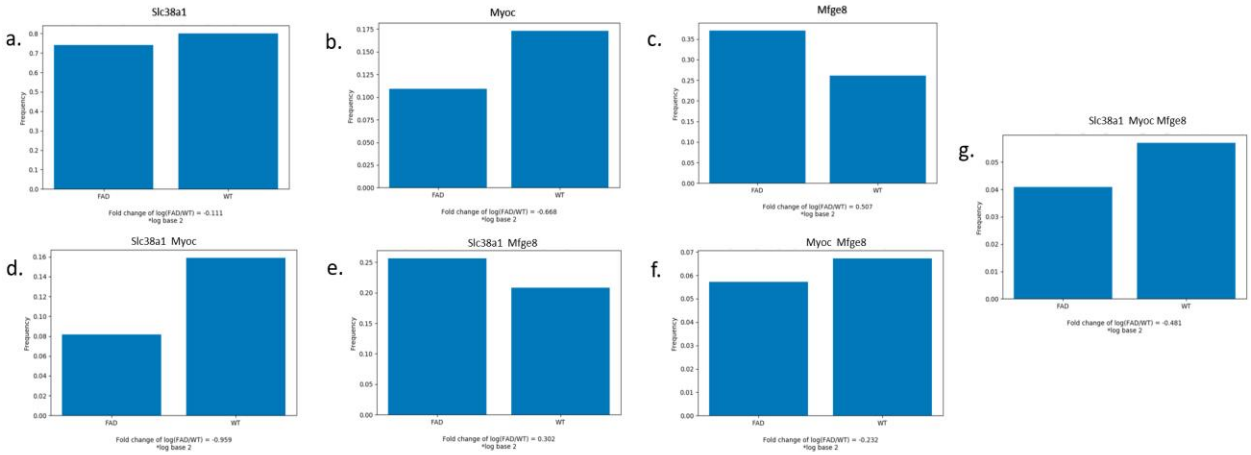

Fig 9. Bar plots comparing the percentage of single, double and triple counts of marker genes expressed across the cells in an image, with Fold change as a significance exhibitor.

a) Comparison of the single counts of Slc38a1, Active/DAA marker gene, in AD and WT, shows a near to equal percentage of cells expressing a similar amount of the gene. b) Comparison of the single counts of Myoc, Active marker gene, in AD and WT, shows a higher percentage of expression amongst WT cells. c) Comparison of the single counts of Mfge8, Homeostatic marker gene, in AD and WT, shows a higher percentage of expression amongst AD cells. d) Comparison of double counts of Slc38a1 and Myoc, both potentially Active markers, shows a higher percentage of expression amongst WT cells. e) Comparison of double counts of Slc38a1 and Mfge8, shows a higher percentage of expression amongst AD cells. f) Comparison of double counts of Myoc and Mfge8, shows a higher percentage of expression amongst WT cells. g) Comparison of all three genes counts, Slc38a1, Myoc and Mfge8, shows a higher percentage of expression amongst WT cells.

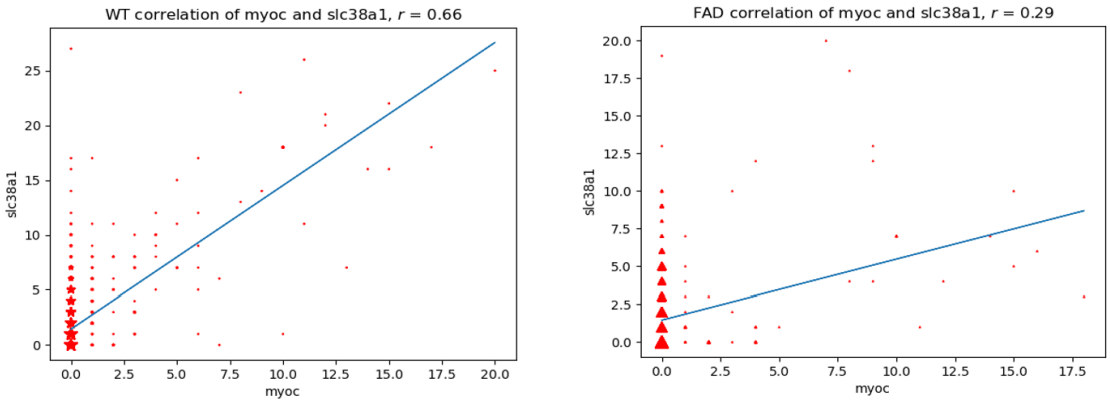

Fig 10. Correlation of gene pairs' counts Comparison between WT (left, stars) and AD (right, triangles) of the pair of genes Myoc (x-axis), Active marker, and Slc38a1 (y-axis), both DAA and Active marker. Each element in graph depicts cells containing the x-axis count it is on of Myoc and y-axis count it is on of Slc38a1. The size of each element reflects the amount of cells with such counts. Blue linear function line shows the liner regression prediction which together with Pierson's *r* explain how well correlated the two genes' expressions are.
In this example we see that Myoc and Slc38a1, both potentially Active state markers of WT brains, show both a better regression prediction and a higher *r* value in WT.

Distinct positions in the hippocampus for astrocyte states

We began our spatial analysis by looking at how each astrocyte population (according to relevant marker genes) is distributed within the hippocampus brain area, around the DG sub-region. We wanted to test if the distributions of the different astrocyte populations were uniform or different

and with distinct localizations. Furthermore, we wanted to see if any of the marker genes are expressed in any specific niches in the hippocampus/DG.

To do so, for every single gene image we calculated its intensity along the image in a 16X16 block grid. This showed us both general localization and the distribution of the expression amongst the image (Fig 11).

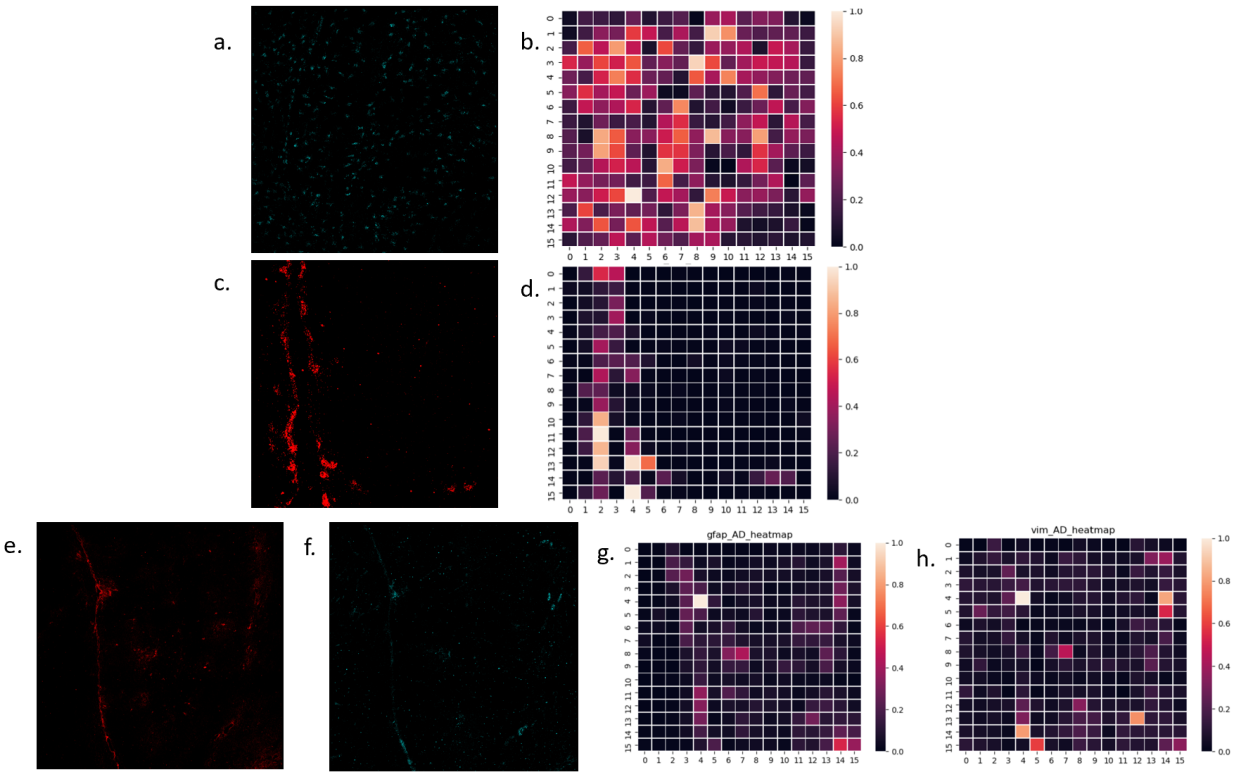

Fig 11. Spatial distribution of different state markers a) Mfge8, Homeostatic marker, single channel image from WT mouse, together with b) heatmap of *a* show a sparse and uniform distribution across the image plane.
c) Myoc, Active marker, single channel image from WT mouse together with d) heatmap of *c* show distribution across a specific area localizing towards the left outline of the image, near but not on the DG.
e) Gfap and f) Vim, DAA markers, single channel images from AD mice, together with g) and h) heatmaps of *e* and *f* accordingly, show distribution across a specific areas and localization surrounding the DG and the Hilus.

We found that Mfge8, a Homeostatic marker, shows sparse expression in a uniform distribution across the entire area of the WT and AD images, both in the DG and around it (Fig 11a-b). Myoc, an Active state marker, shows expression only in a specific subregion in WT. It suggests that active astrocytes are localized in a group of cells which lay in the white matter above the DG. The

distribution of the intensities of the Myoc marker is dense only in this defined stretch of cells and non-existent in the other parts of the DG and its surrounding (Fig 11c-d). We hypothesize this area to be proximal to a blood vessel (see discussion; Fig 12). Finally, Gfap and Vim, both DAA markers, seem to co-localize on cells within the hilus region of the DG, in AD brains (Fig 11e-h). This interesting result clearly showed that DAA are not uniformly distributed in the hippocampus, but rather centered in a specific area. Given the potential role of DAA in AD, we hypothesize that this localization might be dependant on the position of Aβ plaques.

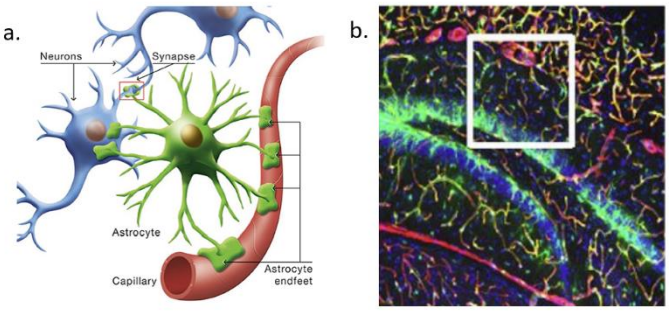

Fig 12. Active state marker on blood vessel a) Schema of astrocyte binding to a blood vessel in a process called endfeet, as to its role in mediating between the neuron and the vessel (Demetrius L. A., et al., 2015).
b) Fluorescent imaging of astrocyte cells on a blood vessel in adult Hippocampus (Licht T., et al., 2016).

DAAs positioned in proximity to Aβ plaques

To test our hypothesis, we next analyzed IHC protein images, to see if DAA are positioned near Aβ plaques and localize around them. These images showing GFAP protein, a DAA marker, together with the plaques, were cleaner than the FISH images and the proteins were found easily through our pipeline, with minor changes only (Fig 4b-e). This gave us an additional way to view our DAA state, allowing us to overcome the limitations we had in the FISH data. Such as not enough marker genes together in the same image, very noisy images which may have harmed some of the data, small area in each image giving a limited view for localization.

Calculating the density of GFAP protein around Aβ plaques *vs.* the density of Gfap in areas with no Aβ plaques (see methods) showed a significance difference in which we found that around Aβ plaques the density is significantly higher (Fig 13a)). We calculated the median density in areas around a plaque to be 0.56, while the median in areas without plaques was only 0.15, with $p - value = 4.52\dot{e}^{-5}$ (done using the Kolmogorov–Smirnov non-parametric test).

Calculating the minimum distance between Gfap proteins to Aβ plaques shows that the distances distribute closer to each other in comparison to both normal distribution (around same mean; Fig 13b) and also in comparison to a random distribution in which we randomly placed Aβ plaques around the image area and recalculated minimum distance to Gfap (Fig 13c). Both the normal and the random distributions were found to be significantly different by using the Kolmogorov–Smirnov non-parametric test and resulting with $p - value = e^{-323}$.

Together, the density and minimum distance calculations show how DAA localize specifically around Aβ plaques, suggesting an important role for DAA in Alzheimer's disease.

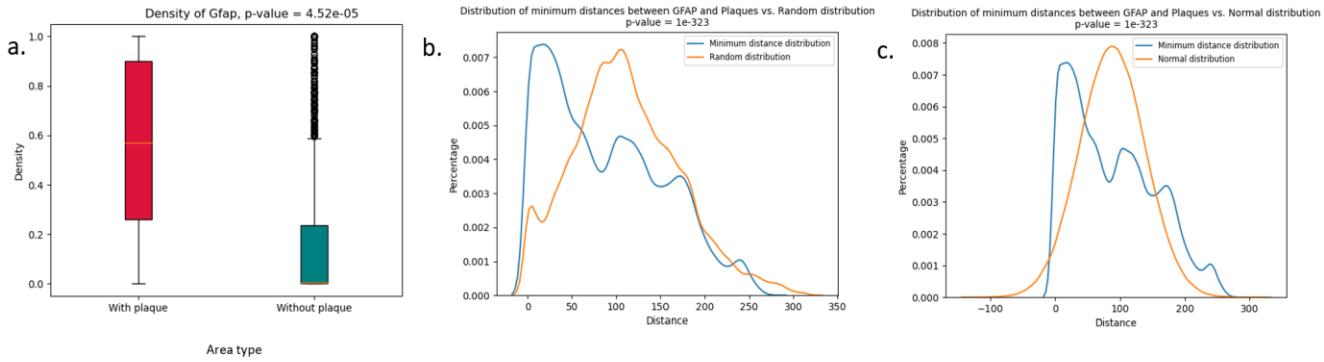

Fig 13. Spatial Analysis – Proteins
a) Box plot of GFAP pixels density in AD mouse, comparing areas with Aβ plaques to areas without plaques. A significant difference is seen between the medians and across almost all the values showing higher density in areas with plaques. $p - value = 4.52e^{-5}$ acquired via KS test.
b) Distribution of minimum distance between GFAP pixels to those of Aβ plaques in AD mouse compared to Normal distribution shows a that the calculated AD distances our significantly shorter, meaning that GFAP and plaques seem to reside specifically close to each other rather than distribute normally throughout the DG. $p - value = e\text{^} - 323$ acquired via KS test.
c) As it is not clear that plaques nor GFAP localize across the DG normally, comparison of the previous minimum distances calculated is compared to a random distribution which was calculated by randomly replacing the plaques throughout the image area and recalculating the minimum distance from them to GFAP pixels once more. This comparison continues to show significantly shorter distances amongst the real data rather than the simulated random data. Here too $p - value = e\text{^} - 323$ acquired via KS test.

Discussion

The initial imagery data given were the four FISH data sets each containing three marker genes that by comparing to the Single Nucleus research (Fig 2c) allowed us to mark them as such:

1. Slc38a1 – DAAs/Active marker, stained in green; Myoc – Active marker, stained in red; Mfge8 – Homeostatic marker, stained in cyan.
2. Slc38a1 – DAAs/Active marker, stained in green; Gfap – DAAs/Active Marker, stained in red; Vim – DAAs Marker, stained in cyan.
3. Slc38a1 – DAAs/Active marker, stained in green; Gsn – DAAs Marker, stained in red; Mfge8 – Homeostatic marker, stained in cyan.
4. Osmr – DAAs marker, stained in green; Myoc – Active Marker, stained in red; Mfge8 – Homeostatic marker, stained in cyan.

Each image contains Nuclei died in blue together with the three genes. The image processing and analysis was done on these single channel (single color of specific gene) images as they give us better visibility and less ambiguity of each gene's data.

Due to technical noise caused by the FISH method during the microscopic fluoresce and imagery we first had to clean the images, where the green channel was always the most noisy due to a technical effect of the green fluorescent, coloring a lot of background empty tissue and causing much smearing. Cleaning the green channel was done in two steps, the first being a Gaussian filter, reducing the background noise intensity, followed by the Non-Local Means Algorithm for more specific noise reduction. This resulted in very clean images with much more clarity of the mRNA molecules. The noise in other channels was mostly defined by small specks of color with low intensities. This was filtered out using a size and intensity threshold, disregarding lone and dim pixels.

Once our images were sufficiently cleaned, we continued to identify each cell in the images and quantify the mRNA molecules of each gene belonging to each cell. Identification of a cell cannot be done in an optimal manner when only the nuclei are stained in the image, as the cell's area is not an exact circle around the nucleus with a predefined radius, rather it can vary between multiple shapes and sizes (Carpenter A.E., et al., 2006). As such, a sub-optimal solution we used is to first identify the nucleus area and then choose a radius around each pixel of the nucleus' perimeter, keeping the general shape of each nucleus and an approximated size relative to that of the nucleus. Furthermore, this solution can also deal with dense areas of neighboring nuclei, in which defining which pixels belong to which nucleus can be ambiguous. To do so we calculated the intensity difference between any ambiguous pixel and each of the nuclei's already defined center of mass. The smallest difference calculated determines to which nucleus the pixel belongs (Carpenter A.E., et al., 2006).

To quantify the mRNA molecules per cell we first needed to identify the mRNA molecules. These were all stained quite clearly after the preprocessing of the clean we applied, so we were able to identify them using only intensity difference to differentiate them from the background and from each other. Quantifying was done by simply counting per gene how many mRNA molecules fall within our previously defined area of each cell. In addition, we also retrieved the location of each cell and RNA molecule. Initial analysis of the quantification results was done as a two-way validation. Firstly, validating the results of the previous sNuc-seq research. The results matching also validated our imaging-based method of retrieving the data, allowing us to continue using this new data for spatial analysis.

Before commencing analysis, we realized that due to the varying amount of cells in the different images (fig 7) we cannot compare any raw count between two images even in the same dataset as they are taken from different mice, at a different zoom. As such, all our comparisons were done between percentages of cells per image. We started our analysis by comparing each gene count separately. We looked at the distribution of the percentage of cells containing different gene

counts, comparing the WT and AD images (from the same dataset). This comparison showed mild differences.

We moved on to a different perspective, looking at the percent of cells in the image which are "positive" as in have a larger count than a predefined threshold of the gene in question. This threshold was calculated by looking at the previous distribution and found to be larger than one. This analysis gave us a clearer understanding of the differences between AD and WT per gene (Fig 9). We then used this same method to compare the difference between the percent of cells which are positive to either a pair or triplet of genes (Fig 9).

To be more accurate with our comparisons of gene pairs co-expression, we added a correlation analysis to differentiate between cells with different counts per gene but the same sum of counts per pair of genes. As in, if a cell expressing two genes A and B, has a count of X of gene A and a count of Y of gene B, whereas another cell which has a count of X-Z of gene A and a count of Y+Z of gene B, then the total for both cells is the same, X+Y. But the larger Z is, the more significant the difference between these two types of cells. As such the correlation allowed us to better determine if a pair of genes truly co-express (Fig 10).

Many of our results proved that our data matches the sNuc-seq data. We found one troubling result of Mfge8 which is a Homeostatic marker, yet in our results showed higher expressions in AD, opposite of our expectations. This may be explained by the fact that Mfge8 is a general astrocyte marker, expressed in all states, although supposedly expressed higher in Homeostatic state. When looking into the data from sNuc we found that first, DAA state emerges from the homeostatic state, and second that the percentage of cells expressing Mfge8 in the different states does rise in Homeostatic, rather only the expression level per cell (Fig 2c). In contrast, in our calculations we looked only at positive cells in a binary form, justifying our results. This together with the previous successful results of the other states' marker genes, brought us to conclude that our data adheres to the same results as sNuc, allowing us to move on to spatial analysis.

When trying to understand how and where around the DG each gene is expressed, we looked at the overall expression distribution amongst the image area of each gene (Fig 11). We found that Mfge8, a Homeostatic marker, is expressed sparsely across the entire image in a uniform manner, which meets our expectation as this is also a general marker for astrocytes so all the cells we marked should express. Myoc, an Active marker seemed to localize around a group of cells near but not on the DG. We didn't know what this area is and so after looking through literature we found that astrocytes, which play a role in the regulation of the blood brain barrier, do localize around blood vessels (Demetrius L. A., et al., 2015; Licht T., et al., 2016; Fig 12). We then hypothesized this area to be as such, seeing that Active is still a healthy WT state. Furthermore,

when re-examining the sNuc-seq data we found that active astrocytes express Aqp4, a marker of astrocytes endfeet that contact the BBB. Finally, Gfap and Vim, both DAAs markers, seemed to express fully around the DG. This seemed interesting because we knew that the DG is connected to memory and we saw two markers of DAAs co-localizing around this area, raising a wonder of what is in the area around the DG. We hypothesized that perhaps it is Aβ plaques.

As such, we continued onto the protein level which allowed us an additional view with cleaner data and most of all allows us to compare the location of Gfap to that of Aβ plaques.

Once receiving the protein images via the IHC method and tweaking our image processing pipeline to run on such images, we were able to retrieve the location data of GFAP proteins and Aβ plaques. We wanted to check if the Aβ plaques localize near the GFAP proteins, as we hypothesized earlier. Seeing that we didn't have any WT images to compare to we decided to calculate the distribution of minimum distance between GFAP and Aβ plaques and compare it to both normal and random distribution. Next, we calculated the density of Gfap around plaques vs the density in the area with no plaques. The result of both the density and the minimum distance calculations showed significantly high differences, proving that GFAP and Aβ plaques co-localize as we hypothesized. This also shows that Aβ plaques, which have been known as hallmark pathology of AD but never known its full connection to the disease, grow near to a specific gene marker which can be targeted. As such we believe that more of the previously mentioned DAA marker genes should be analyzed on their protein level, together with Aβ plaques as well. Examining where they localize, if they also co-localize with the plaques or with each other, can ultimately help build a bank of targetable genes and proteins in AD brains.

References

Attwell, D., et al. (2010). Glial and neuronal control of brain blood flow. *Nature* **468**: 232-243.

Alzheimer's Association (2019). Alzheimer's Disease Facts and Figures. *Alzheimers Dement* **15**(3): 87-321.

Carpenter A.E., et al. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* **7**.

Clifton P.D. and Wun-Ju S. (2015). Current Laboratory Techniques in Rabies Diagnosis, Research and Prevention, **2**.

Demetrius L. A., et al. (2015). Alzheimer's disease: thea myloid hypothesis and the Inverse Warburg effect. *Frontiers in Phisiology* **5** (522).

Frickmann H. (2017). Fluorescence in situ hybridization (FISH) in the microbiological diagnostic routine laboratory: a review. *Critical Reviews in Microbiology* **43**(3): 263-293.

Geneticist Inc. (2018). The gold standard for Immunohistochemistry. *Genetics Insider.*

Habib N., et al. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**: 925-928.

Habib N., et al. (2017). Massively-parallel single nucleus RNA-seq with DroNc-seq. *Nature Methods* **14**: 955-958.

Habib N., et al. (2020). Disease-associated astrocytes in Alzheimer's disease and aging. *Nature Neuroscience* **23**: 701-706.

Holtzman, D.M., et al. (2011). Alzheimer's disease: the challenge of the second century. *Science translational medicine* **3**(77).

Licht T., et al. (2016). VEGF preconditioning leads to stem cell remodeling and attenuates age-related decay of adult hippocampal neurogenesis. *PNAS* **113** (48) E7828-E7836.

Lubeck E., et al. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature* **11**: 360–361.

Maity B., et al. (2013). Immunostaining: detection of signaling protein location in tissues, cells and subcellular compartments. *Methods Cell Biol* **113**: 81-105.

Moor A.E., et al. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. *Current Opinion in Biotechnology* **46**: 126-133.

Strooper, B.D., and Karran, E. (2016). The Cellular Phase of Alzheimer's Disease. *Cell* **164**: 603-615.

Stuart T. and Satija R. (2019). Integrative single-cell analysis. *Nature* **20**: 257–272.

Vineyard C. M., et al. (2012). A multi-modal network architecture for knowledge discovery. *Security Informatics* **1** (20).