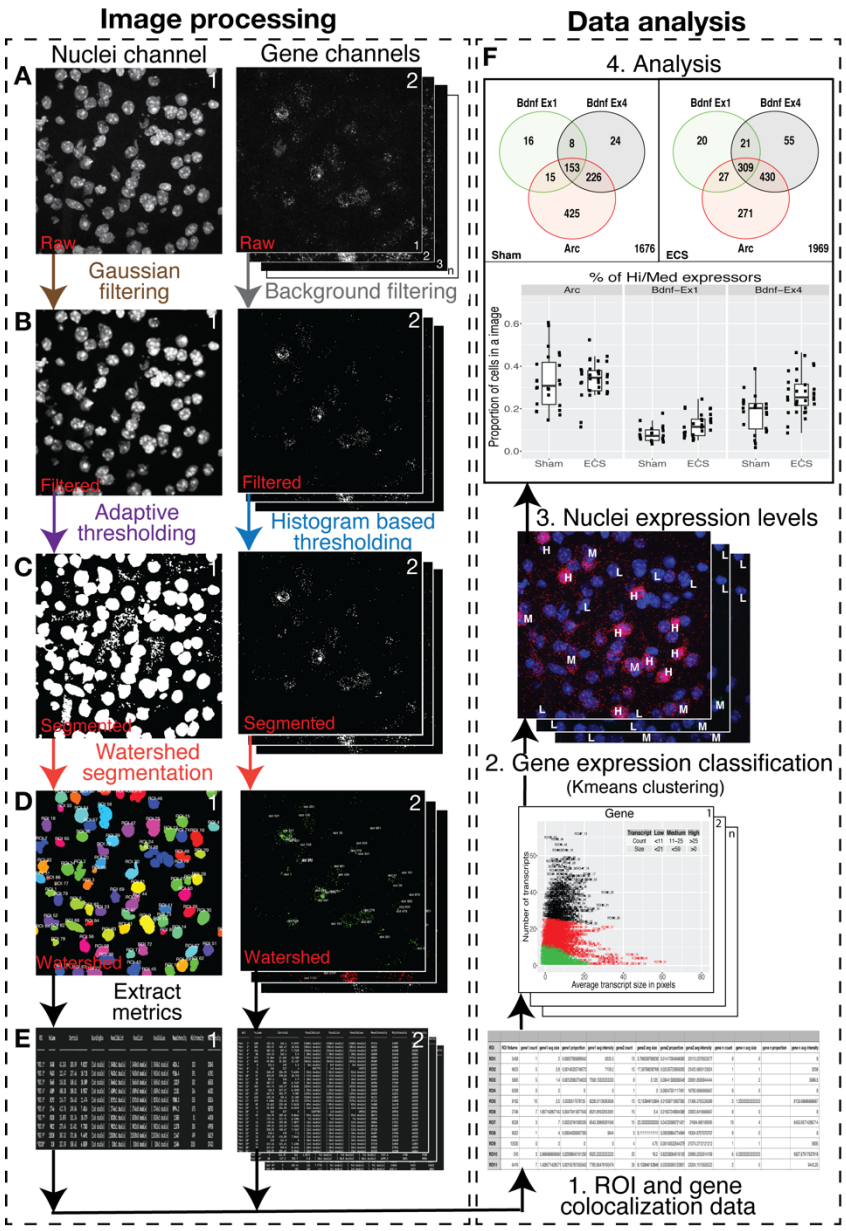
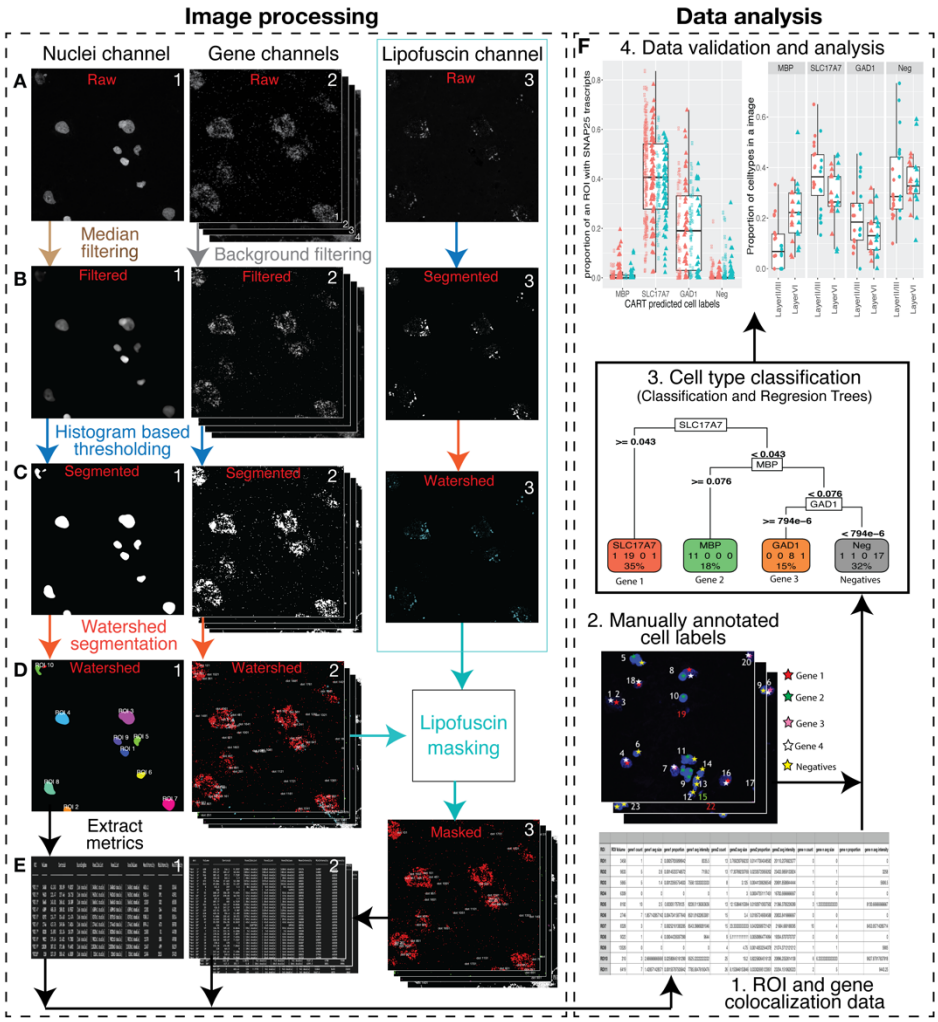
Supplementary Material

"Dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues"

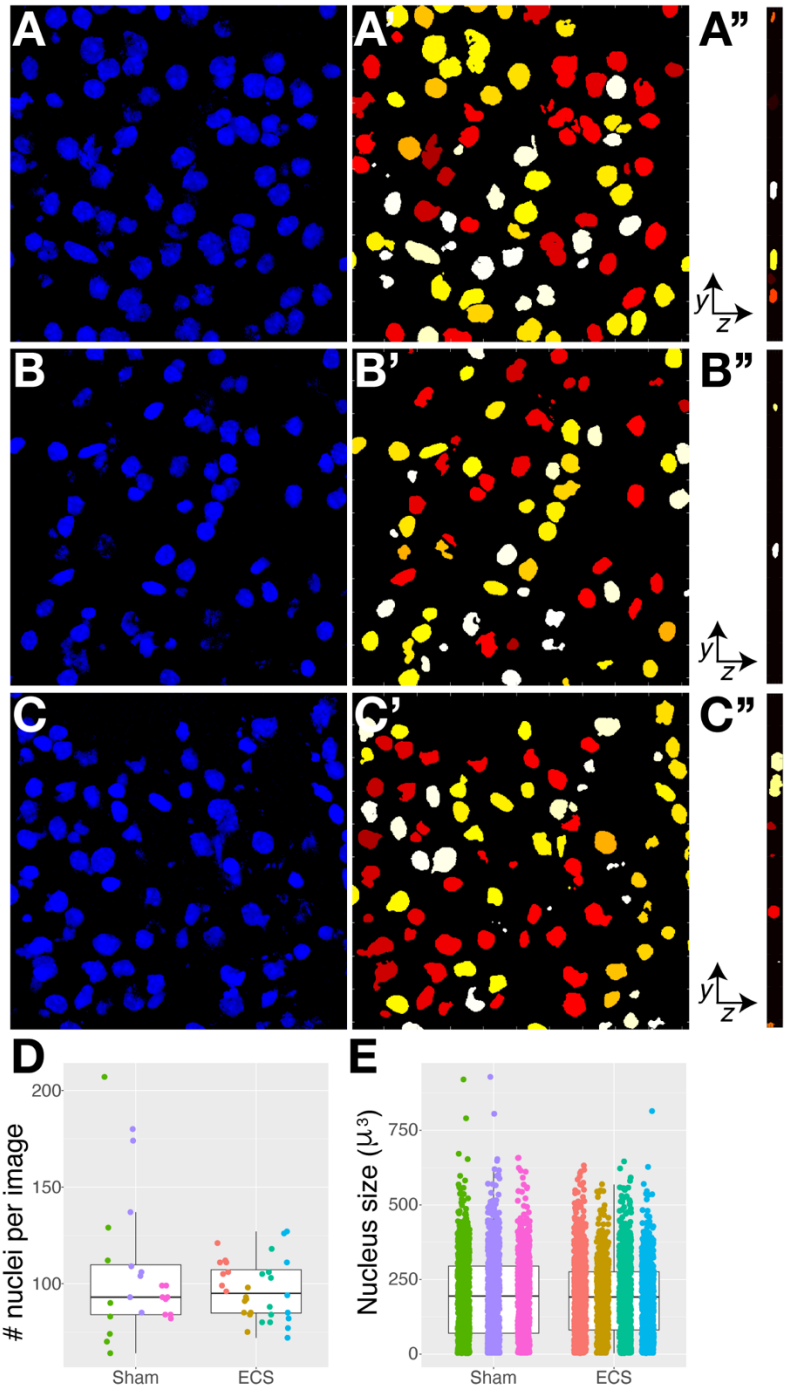
12 figures and corresponding legends (Figures S1-S12)
1 table (Table S1)



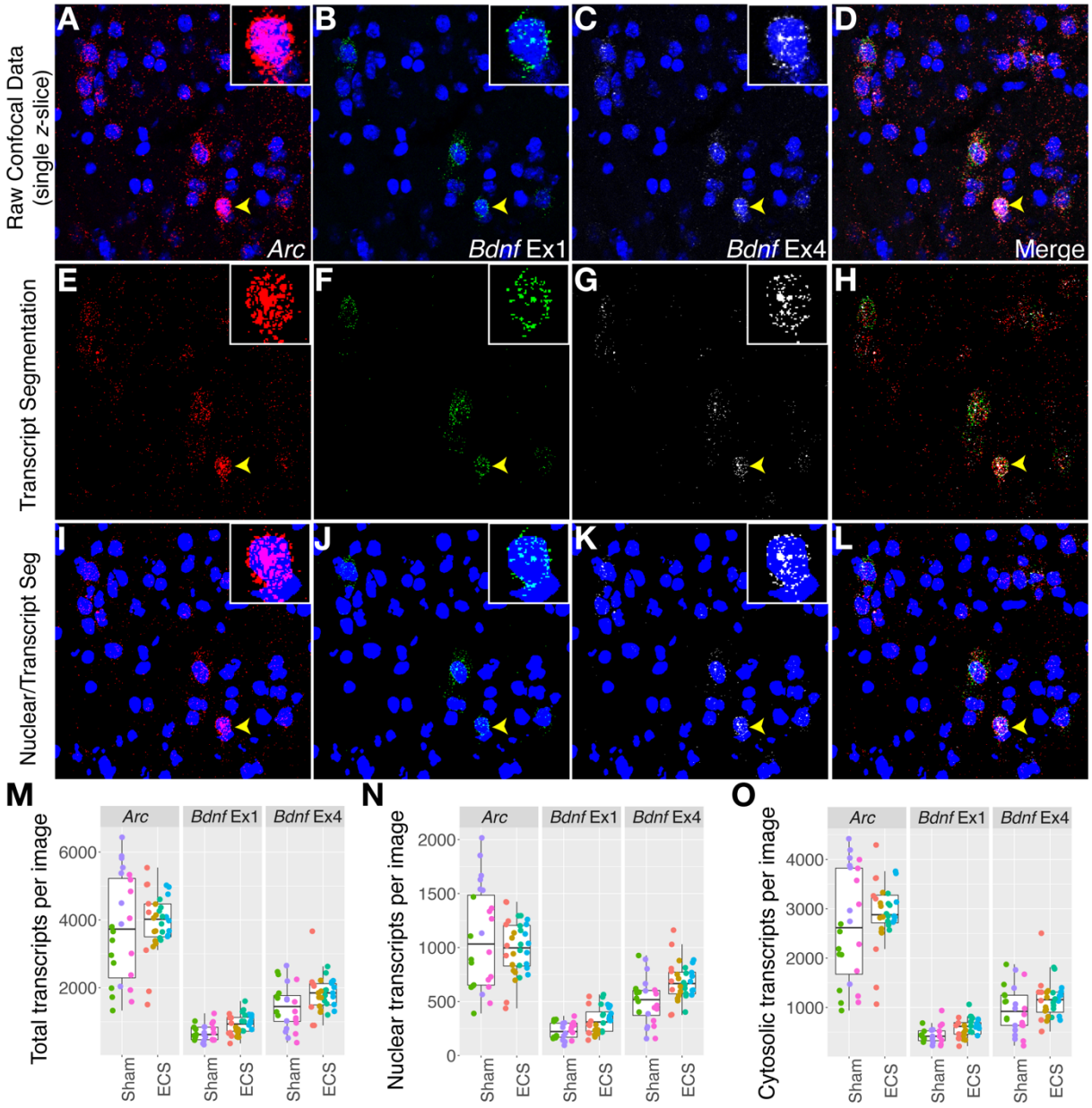
Supplementary Figure 1. *Dotdotdot* image processing and data analysis workflows for smFISH data in mouse tissue. A-E, Image processing workflow. A) Raw *".czi"* images of mouse nuclei (DAPI) and gene channels (i.e., *Bdnf* Ex1 [Opal520], *Arc* [Opal570], and *Bdnf* Ex4 [Opal690]). B) Processed images of nuclei and transcript channels (gaussian smoothened (B1) and background filtered (B2)). C) Segmented binary images of nuclei (adaptive thresholding) and transcript (image histogram-based thresholding) channels. D) Final watershed segmentation of nuclei and transcript channels. E) Metrics showing ID, volume (in pixels), location (centroid, bounding box, indices), intensities (mean, minimum and maximum) of each segmented region (i.e. nuclei and transcripts) per channel. F, Data analysis workflow**.** 1) Location metrics from E1 and E2 are used to find ROIs/nuclei and transcript colocalization information. Table shows each ROI/nucleus with its transcript quantification (count, average size, average intensity, proportion of nuclei volume occupied) per gene channel. 2-3) K-means clustering of ROIs/nuclei into high, medium and low expressers per gene channel based on transcript count and average transcript size of nuclei. 4) Data analysis based on thresholds from k-means clustering.



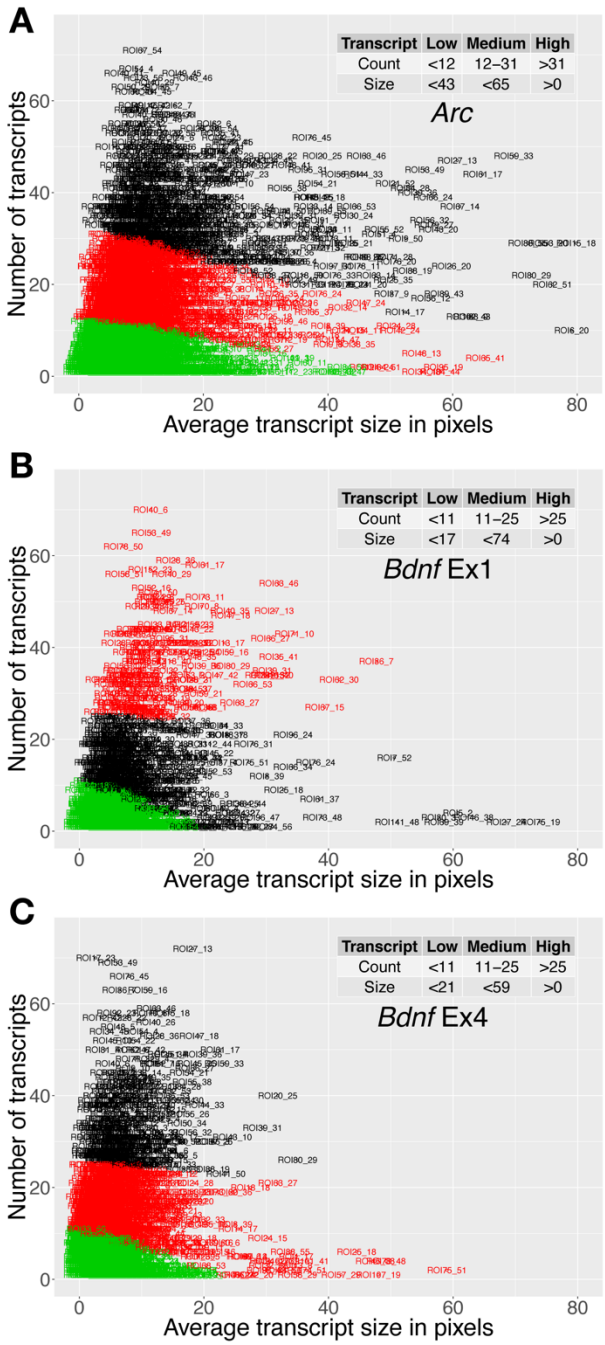
Supplementary Figure 2 . Image processing and data analysis workflows for smFISH data in human tissue. A-E, Image processing workflow**.** A1, A2, A3) Raw *".czi"* images of human nuclei (DAPI), gene channels (*MBP* [Opal520], *SLC17A7* [Opal570], *GAD1* [Opal620], and *SNAP25* [Opal690]) and the lipofuscin channel. B1, B2) Processed images of nuclei and transcript channels (3D Median filtered/smoothened (B1) and background filtered (B2)). There was no processing step on the lipofuscin channel to retain all signals. C1, C2, B3) Histogram-based threshold-ed binary images of nuclei, transcript and lipofuscin channels. D1, D2, C3) Final watershed segmentation of nuclei, transcript (overlaid with lipofuscin autofluorescence pixels in cyan) and lipofuscin channels. E3) Segmented and lipofuscin-masked image of each gene channel. E1,E2) Final extracted metrics showing ID, volume (in pixels), location (centroid, bounding box, indices), intensities (mean, minimum and maximum) of each segmented region (i.e. nuclei and transcripts after lipofuscin masking) per channel. F) Data analysis workflow. 1) Location metrics from E1 and E2 are used to find ROIs/nuclei and transcript colocalization information. Table shows each ROI/nucleus with its transcript quantification (count, average size, average intensity, proportion of nuclei volume occupied) per gene channel. 2) Manually annotated images show ROIs/nuclei positive for gene targets (color labelled as red-*SLC17A7*, green-*MBP,* white-*SNAP25*, pink-*GAD1* and yellow-negative) based on qualitative analysis. 3) Classification and regression tree (CART) model produced by using the proportion of nuclei occupied by each gene and its manually annotated cell labels (Gene1 (*SLC17A7*), Gene2 (*MBP*), Gene3 (*GAD1*) and negative). 4) Based on CART predictions, nuclei from all images are classified into predefined cell types and data analysis is performed. To validate CART predictions, we show that *SNAP25* is enriched in nuclei that are positive for *SLC17A7* and *GAD1*.



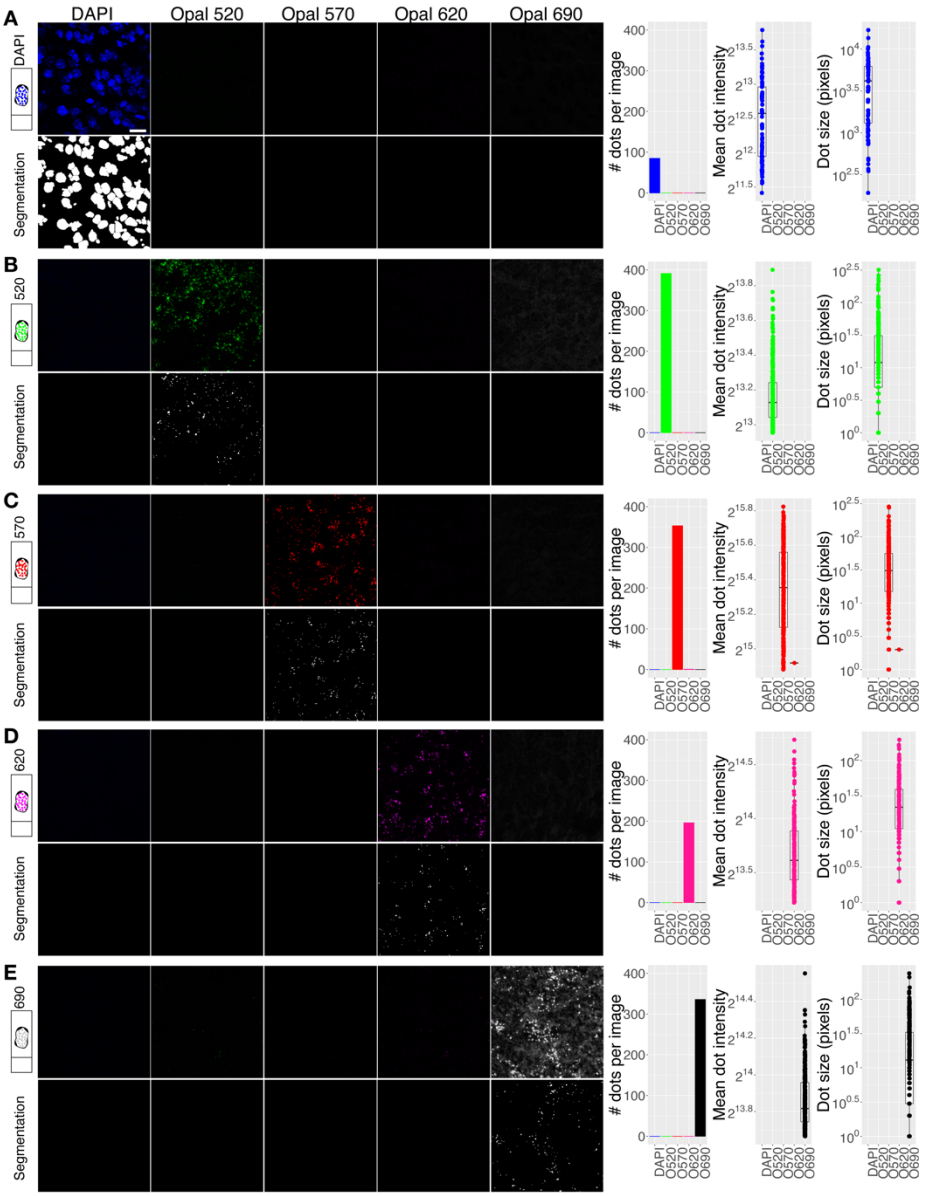
Supplementary Figure 3. Nuclei segmentation defines regions of interest (ROIs) in mouse tissue. A-C) DAPI staining depicting individual nuclei in *x*,*y*-dimensions of a single confocal *z*-plane from three representative areas in the mouse piriform cortex. A'-C') Corresponding nuclear segmentation in *x*,*y*-dimensions with each nucleus (yellow, red, orange, or white) representing a single ROI. A''-C'') Nuclear segmentation in *y*, *z*-dimensions. For *z*-stacks, nuclear segmentation is performed in each *z*-plane and ROIs are reconstructed in 3 dimensions. D) Number of segmented nuclei per field in images acquired from Sham and ECS treated mice (n= 24 images, 3 mice and n= 32 images, mice, respectively). Each dot represents an image and each color represents a different mouse. E) Average size of segmented nucleus in Sham compared to ECS images (n= 56 images and 5645 nuclei).



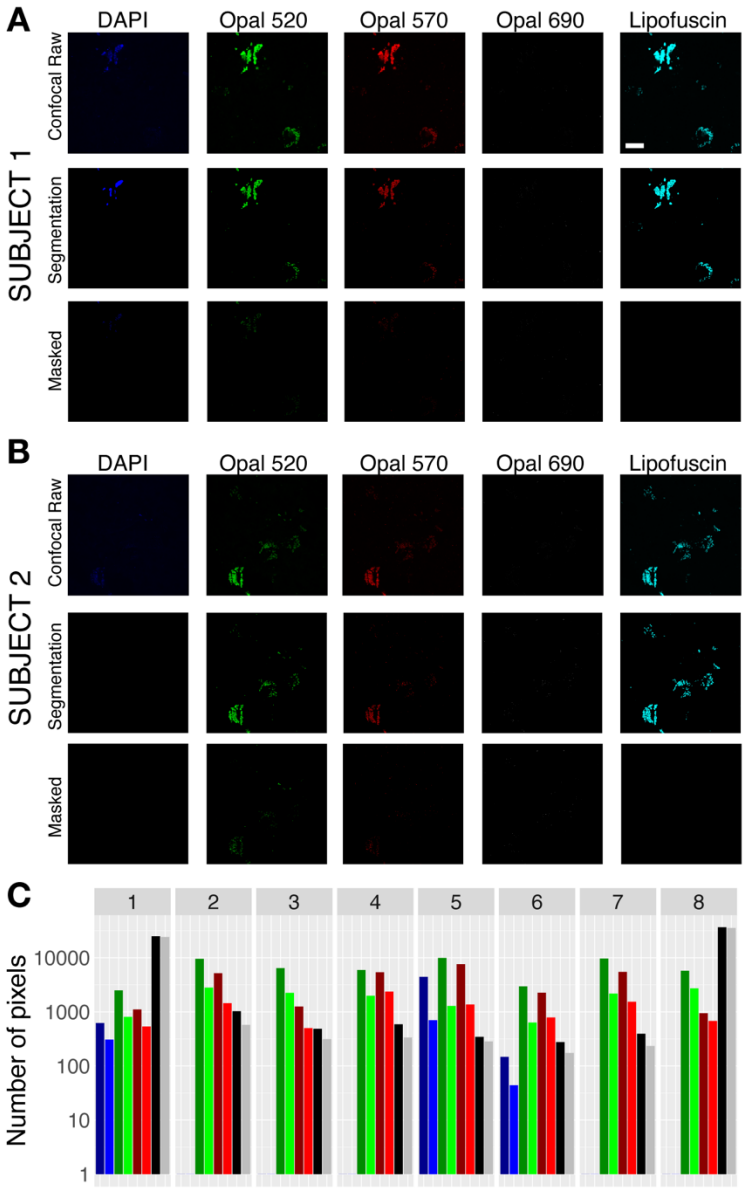
Supplementary Figure 4. Three-dimensional dot segmentation and feature extraction delineates individual probe signals in multiplex images. A-D) Representative confocal *z*-plane of mouse piriform cortex depicting nuclear DAPI staining (blue) and single transcripts for *Arc* (A), *Bdnf Ex1* (B), *Bdnf Ex4* (C), and merged (D). E-H) Corresponding dot segmentation for *Arc* (E), *Bdnf Ex1* (F), *Bdnf Ex4* (G), and merged (H). I-L) Overlay of nuclear and dot segmentation used for identification of ROIs and quantification of dot/transcript features (size, number, and intensity) for *Arc* (I), *Bdnf* Ex1 (J), *Bdnf* Ex4 (K), and merged (L). M) Total number of *Arc*, *Bdnf* Ex1, and *Bdnf* Ex4 transcripts per field in images acquired from Sham and ECS treated mice (n= 3 mice, 24 images, 2543 nuclei and n=4 mice, 32 images, and 3102 nuclei, respectively). Each dot represents an image and each color represents a different mouse. N) Total number of nuclear transcripts per field for *Arc*, *Bdnf* Ex1, and *Bdnf* Ex4 in Sham compared to ECS. O) Total number of cytosolic transcripts per field for *Arc*, *Bdnf* Ex1, and *Bdnf* Ex4 in Sham compared to ECS. Yellow arrows depict positive cells.



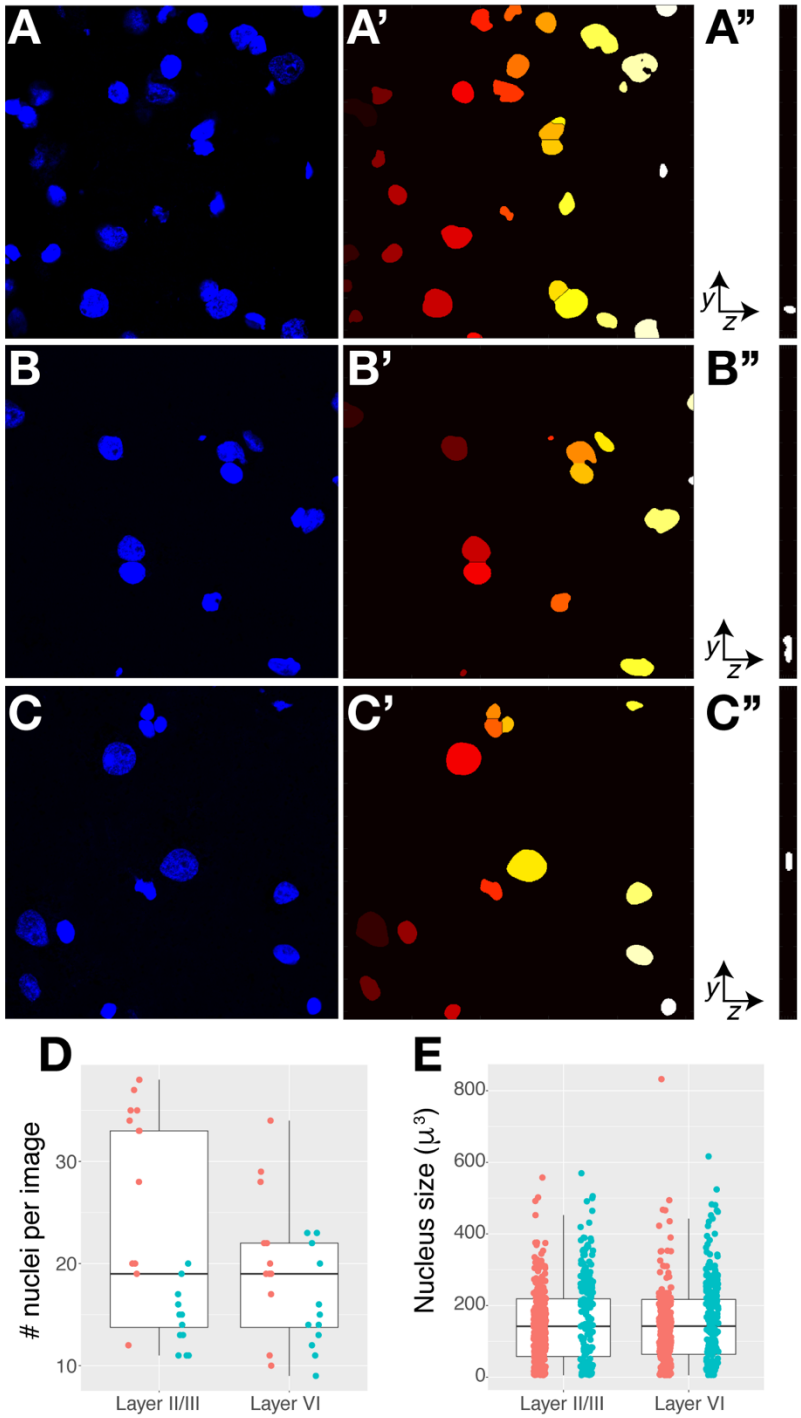
Supplementary Figure 5. K-means cluster analysis uses transcript (dot) count and size per ROI to group ROIs as low, medium, or high expressers for individual activity-regulated genes. A) Plot depicting clustering of 4,967 ROIs with low, medium, and high *Arc* expression based on number of transcripts (low: <12; medium: 12-31; high: >31) and average transcript size (low: <43; medium: <65; high: >0). B) Plot depicting clustering of 2,538 ROIs with low, medium, and high *Bdnf* Ex1 expression based on number of transcripts (low: <11; medium: 11-25; high: >25) and average transcript size (low: <17; medium: <74; high: >0). C) Plot depicting clustering of 3,983 ROIs with low, medium, and high *Bdnf* Ex4 expression based on number of transcripts (low: <11; medium: 11-25; high: >25) and average transcript size (low: <21; medium: <59; high: >0). Nuclei with <1 transcript are directly assigned to the low group.



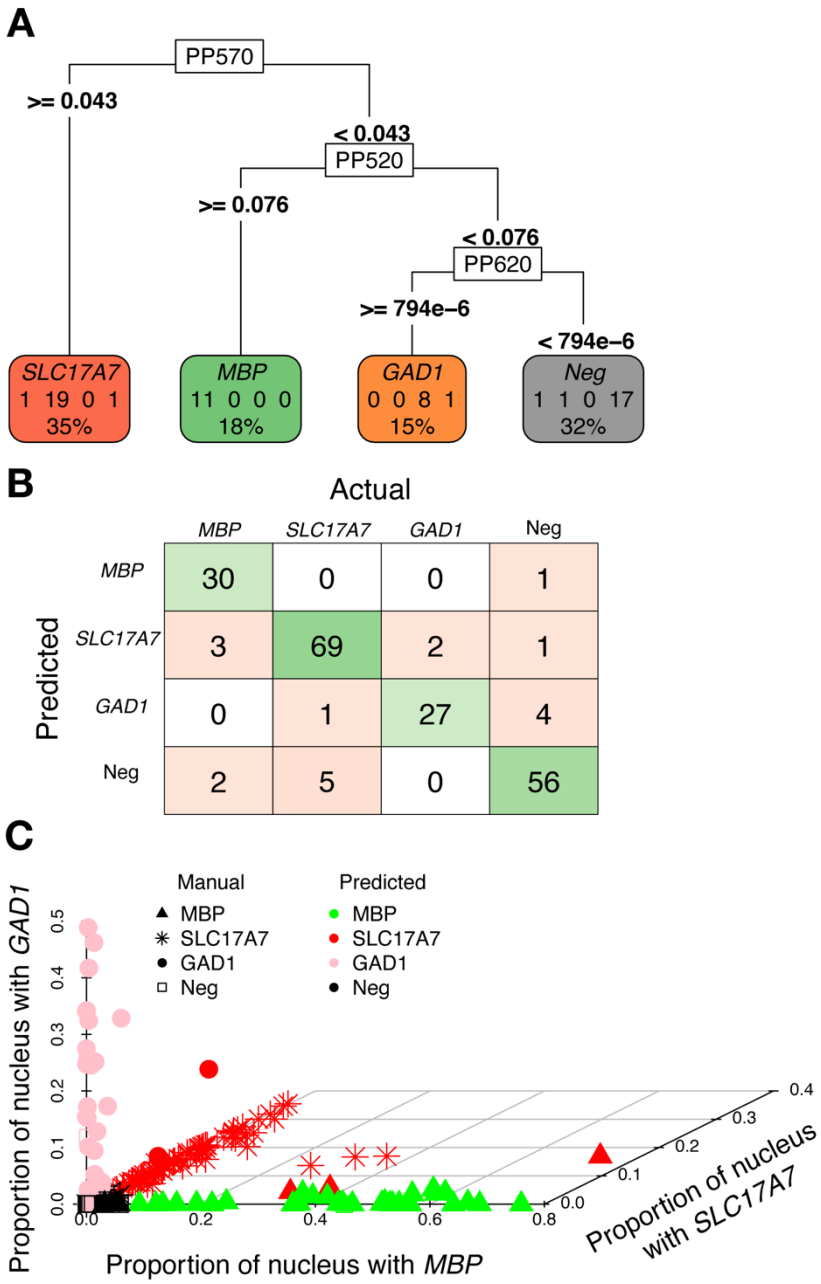
Supplementary Figure 6. Validation of reference emission spectral profiles ("fingerprints") used for linear unmixing. A) A reference emission spectral profile, or "fingerprint," was created in Zen software for DAPI using mouse brain tissue subjected to pretreatment conditions, but no additional probe labeling. Top panel shows linear unmixing with the DAPI fingerprint specifically recognizes DAPI signal. Bottom panel depicts segmentation for ROI identification. B-E) Reference emission spectral profiles were created in Zen software for each of the Opal dyes (referred to as Opal (O)520, Opal (O)570, Opal (O)620, and Opal(O) 690) using a "single positive" slide of mouse brain tissue hybridized with a positive control probe against a house-keeping gene, *POLR2A*, and visualized with the respective Opal dye (Opal520 in B, Opal570 in C, Opal620 in D, and Opal690 in E). Top panels show linear unmixing with DAPI, Opal520, Opal570, Opal620, and Opal690 for each single positive slide. Bottom panels depict segmentation for dot detection. Plots to right of images show *POLR2A* dot number, size, and average intensity after linear unmixing of respective single positive images. Importantly, reference emission spectral profiles are highly specific (i.e. when *POLR2A* is labeled with Opal570 in C, the spectrally similar Opal620 fingerprint does not detect Opal570 signal, etc.). Similarly, dot features are only detected after unmixing with the relevant reference emission spectral profile. Scale bar is 20um.



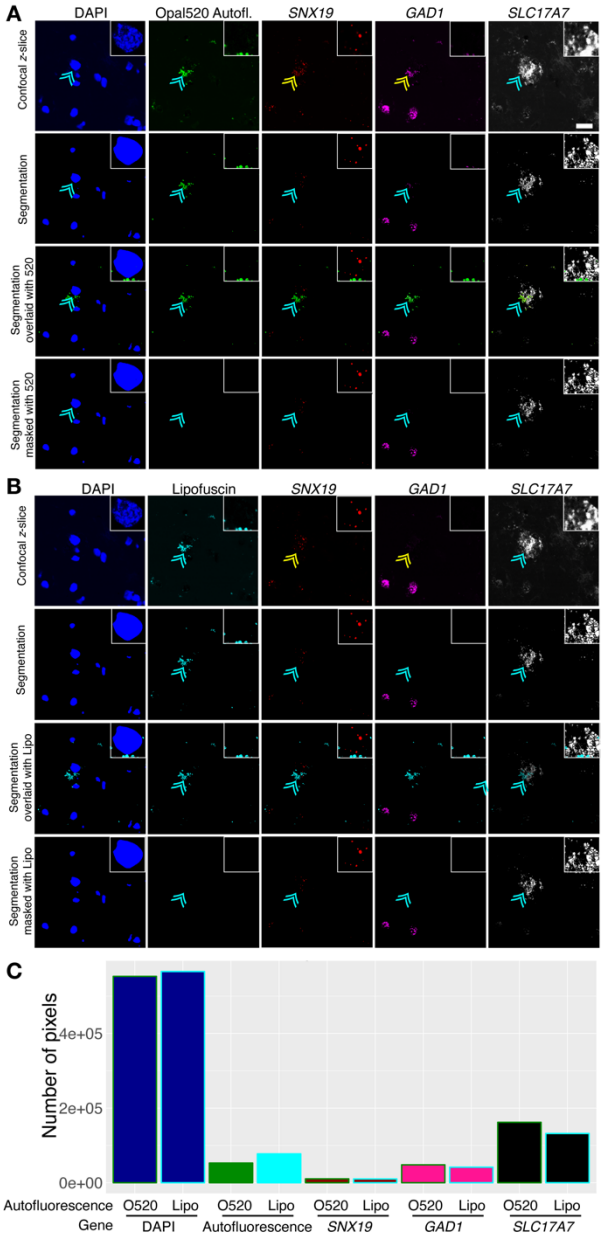
Supplementary Figure 7. A common spectral signature for lipofuscin can be used across multiple subjects to identify and exclude autofluorescence. A-B) A reference emission spectral profile, or "fingerprint," was created in Zen software for lipofuscin autofluorescence using DLPFC from a representative subject hybridized with a negative control probe for the bacterial gene *dapB*. For 2 different human subjects (A and B), top panels show a single confocal z-slice after linear unmixing with DAPI, Opal520, Opal570, Opal60, and lipofuscin fingerprints. Middle panels show corresponding segmentation. Bottom panels show segmentation after masking with lipofuscin signals. C) Plot depicting the number of transcript pixels detected for DAPI (blue bars), Opal520 (green bars), Opal570 (red bars), Opal690 (black/gray bars) before and after masking with lipofuscin (dark bars=before, light bars=after) across 8 images derived from 4 different subjects (Images 5-6 from Subject 1, Images 7-8 from Subject 2, Images 1-2 from Subject 3, Images 3-4 from Subject 4,). The reference emission spectral profile for lipofuscin is equally effective across subjects and successfully masks and excludes pixels confounded by autofluorescence across the electromagnetic spectrum. Scale bar 20um.



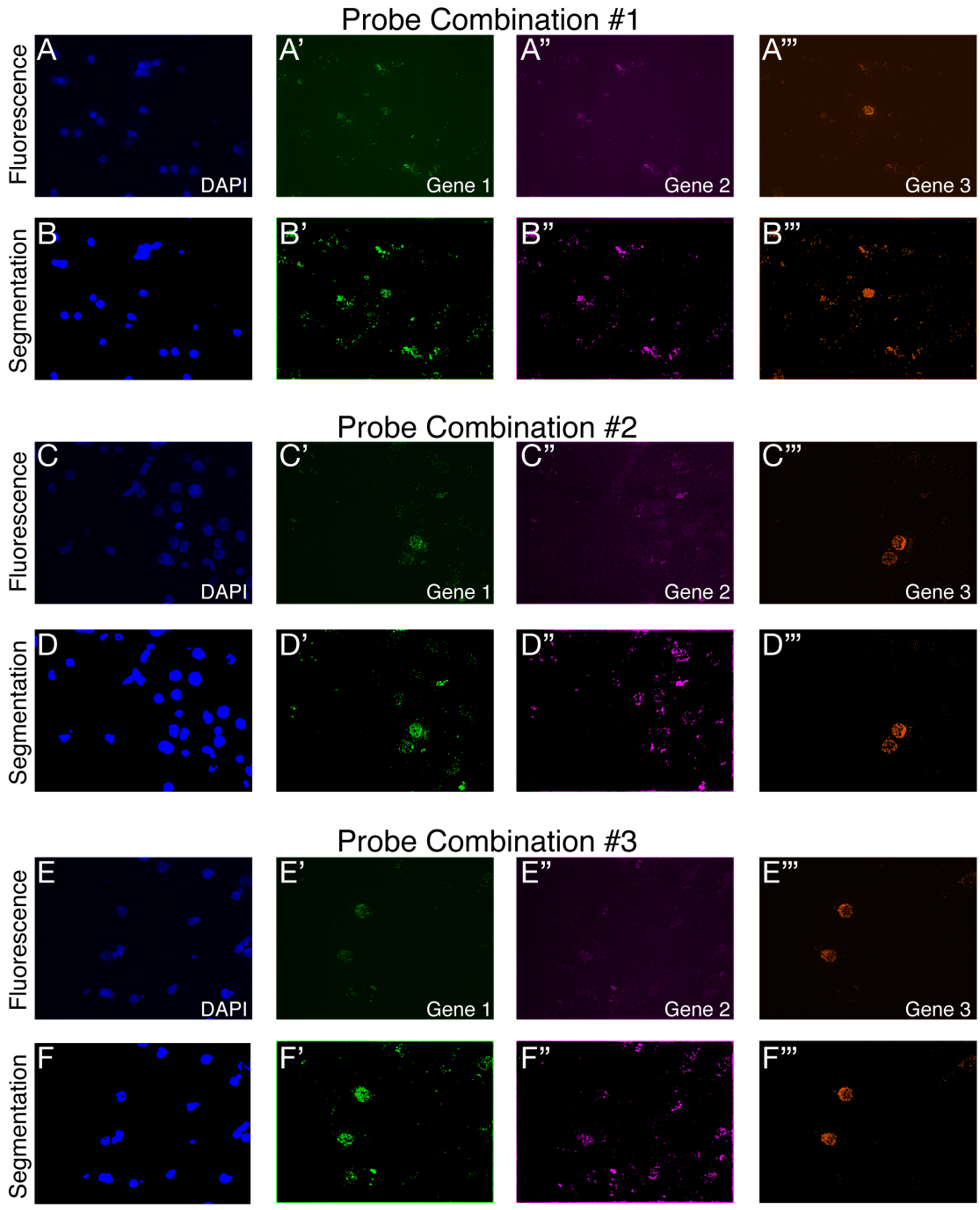
Supplementary Figure 8. Nuclei segmentation defines regions of interest (ROIs) in human brain. A-C) DAPI staining depicting individual nuclei in *x*,*y*-dimensions of a single confocal *z*-plane from three representative areas in postmortem human DLPFC. A'-C') Corresponding nuclear segmentation in *x*,*y*-dimensions with each nucleus (yellow, red, orange, or white) representing a single ROI. A''-C'') Nuclear segmentation in *y*, *z*-dimensions. For *z*-stacks, nuclear segmentation is performed in each z-plane and ROIs are reconstructed in 3 dimensions. D) Number of segmented nuclei per field in images acquired from cortical layers II/III and VI (n=2 subjects, 2 cortical strips per subject, 24 images). E) Average size of segmented nucleus in layer II/III compared to layer VI (n= 519 and 442 nuclei, respectively).



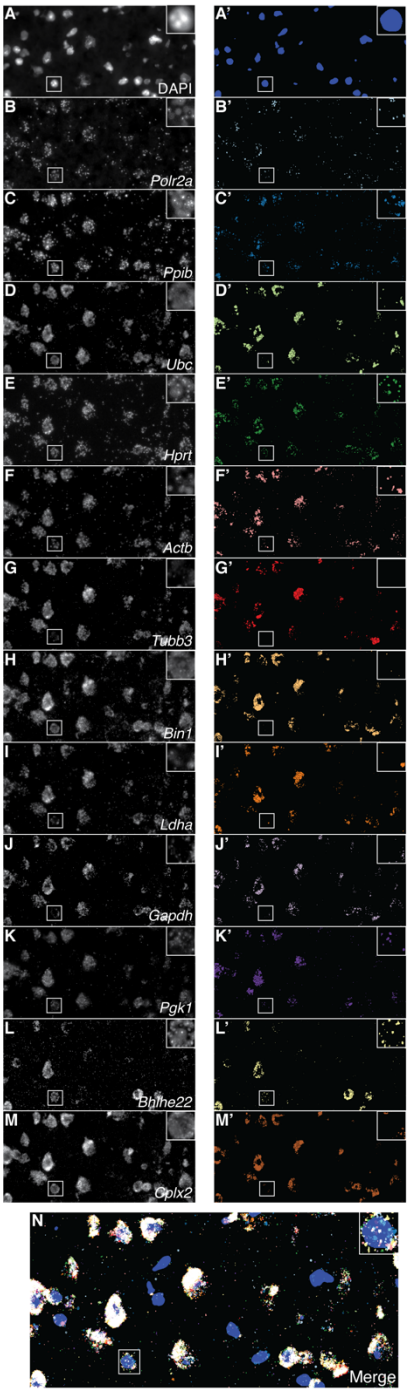
Supplementary Figure 9. Classification and regression tree (CART) model for accurately predicting cell types in DLPFC. A) Classification tree built from 60 random ROIs from 5 manually annotated images using the rpart algorithm with termination criteria of 4 major classes: *SLC17A7* (excitatory neuron), *MBP* (oligodendrocyte), *GAD1* (inhibitory neuron), and triple negative (Neg; likely microglia and astrocytes). For example, a cell is defined as an *SLC17A7*+ excitatory neuron when the volume of the ROI is at least 4.3% *SLC17A7* (given that *MBP* covers less than 7.6% of the ROI and *GAD1* covers less than 0.0794% of the ROI). For these 60 ROIs, the classifier predicts 35% of the ROIs are SLC17A7, 18% are MBP, 15% are GAD1, and 32% are triple negative. B) Confusion matrix for 201 manually annotated ROIs comparing predicted and actual cell type. C) Plot showing manual and predicted cell type for each ROI plotted against the proportion of the nuclei positive for *GAD1*, *SLC17A7*, and *MBP*. Prediction accuracy of the CART model is 91%.



Supplementary Figure 10. Flexibility for using autofluorescence detected in a narrow spectral range to mask lipofuscin autofluorescence. A) Top panel shows confocal z-slice after linear unmixing with DAPI, Opal520, Opal570, Opal620, and Opal690 fingerprints to detect *SNX19* (Opal570), *GAD1* (Opal620), and *SLC17A7* (Opal690). As no probe was used with the Opal520 dye, signal in this spectral range can be attributed to lipofuscin autofluorescence. Second panel shows segmentation of fluorescent signals. Third panel shows segmentation overlaid with autofluorescence captured by unmixing with the Opal520 fingerprint. Bottom panel shows segmentation for each gene after masking with Opal520 autofluorescence. Cyan and yellow arrows show lipofuscin autofluorescence. B) Using the same lambda stack, top panel shows confocal z-slice after linear unmixing with DAPI, lipofuscin, Opal570, Opal620, and Opal690 fingerprints to detect *SNX19* (Opal570), *GAD1* (Opal620), *SLC17A7* (Opal690), and lipofuscin autofluorescence. Second panel shows segmentation of fluorescent signals. Third panel shows segmentation overlaid with autofluorescence captured by unmixing with the lipofuscin fingerprint. Bottom panel shows segmentation for each gene after masking with lipofuscin autofluorescence. Lipofuscin fingerprint captures additional autofluorescence compared to Opal520 fingerprint (compare yellow arrows in A and B). C) Plot comparing the number of pixels detected for DAPI or each gene when lipofuscin autofluorescence is unmixed using the Opal520 (O520; green border bars) versus lipofuscin (Lipo; cyan border bars) fingerprints. The lipofuscin fingerprint captures more autofluorescent pixels than the Opal520 fingerprint (cyan vs. green bar) thereby reducing the number of pixels attributed to autofluorescence in other channels. Scale bar is 20um.



Supplementary Figure 11. *dotdotdot* is compatible with diverse file formats generated using different microscope systems. A-F) Raw confocal fluorescence (A, C, E) and corresponding nuclear/transcript segmentation (B, D, F) from 3 unique RNAscope images acquired in postmortem human brain tissue utilizing separate probe combinations. Imaging processing with *dotdotdot* was performed on Nikon "*.nd2*" files.



Supplementary Figure 12. *dotdotdot* segments 12 gene targets labeled using RNAscope Hi-plex assay. A-M) Raw image fluorescence (left) and corresponding transcript segmentation (right) for DAPI and 12 positive control (house-keeping) genes in mouse brain tissue. N) Merge of segmented images.

Table S1. Comparison between *dotdotdot* and alternative platforms

Platform	Software	Reference	Differences compared to *dotdotdot*
ImageJ	Count nuclear foci	https://imagej.net/ImageJ	- Needs manual adjustment for smoothing and thresholding - Limitations for batch processing - Each channel should be an individual image - Quantification output is limited
Matlab	FISH-quant	PMID: 23538861	- GUI based - Requires images in .tiff format - Requires each channel to be saved as a separate image - Nuclear segmentation is performed in 2D - Automated nuclei segmentation requires other software (CellProfiler) - If batch processed, there are multiple steps such that nuclei segmentation for all images is processed first followed by transcript segmentation
Leica Biosystems	Aperio	Commercial	- Not open source (only commercially available)
Indica Labs	Halo	Commercial	- Not open source (only commercially available)
IDL (Interactive Data Language)	Localize	PMID: 29785002	- Limited details and documentation are available to run the software - Does not provide any cell segmentation
Python	Allen Cell Structure Segmenter	*Biorxiv*: https://doi.org/10.1101/491035	- One arm is based on machine learning and deep learning techniques that require ground truths to be generated by the researchers - Need to create custom scripts to obtain quantification metrics
Python	Starfish	PMID:31427807	- Needs data to be converted to SpaceTx format - No fixed pipeline is specified; user needs to build one based on individual needs - Needs human intervention for cell segmentation