AD_____

Award Number: W81XWH-08-1-0420

TITLE: Brain Region and Cell Type Transcripts for Informative Diagnostics

PRINCIPAL INVESTIGATOR: Leroy E. Hood, Ph.D., M.D.

CONTRACTING ORGANIZATION: Institute for Systems Biology
Seattle WA 98103

REPORT DATE: September 2010

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
Sep 1 2010	Annual	16 JUN 2009 - 9 AUG 2010

4. TITLE AND SUBTITLE
Brain Region and Cell Type Transcripts for Informative Diagnostics

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-1-0420

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Leroy E. Hood, PhD, MD

Email: LHOOD@SYSTEMSBIOLOGY.ORG

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Systems Biology
Seattle, WA 98103

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
US Army Medical Research and Material Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our study will provide a set of candidate markers that can be used not only for the diagnosis of brain trauma, such as traumatic exposure to explosions, but also diseases such as brain cancer, including glioblastoma, and neural degenerative diseases, such as Alzheimer's disease. This year we used our EigenBrain approach to identify specific regions that are highly expressed in each cell type. In addition, we discovered a strong candidate set of brain-specific and cell-type specific transcripts. Moreover, we applied the region-based clustering method to the in situ hybridization of cell type specific genes. This revealed dramatic spatial consistency of neuron-specific genes, sufficient to recapitulate most anatomical brain regions from gene expression alone. We also analyzed brain RNAseq data to measure transcripts present in both human and mouse brains. This allowed us to detect the presence of many brain-specific genes (798 transcripts) that improved our candidate selection.

15. SUBJECT TERMS
brain-region specific markers

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		42	USAMRMC
					19b. TELEPHONE NUMBER *(Include area code)* (301)-619-2254, Lisa Sawyer, Contract Specialist

Reset

Table of Contents

Page

Annual Report for DoD Grant (Award Number: W81XWH-08-1-0420)

1. Introduction

The identification of cell-type specific or brain-region specific transcripts is important to enable the use of peripheral biomarkers to identify disease perturbations in the brain that can then be located to region and cell type(s) affected by the disease. Thus, the identification and characterization of these cell-type and brain region specific transcripts will be of high utility for the eventual diagnosis of a wide variety of brain diseases and brain trauma, such as traumatic exposure to explosions of soldiers in war. In this year, we developed the novel classification method -"Eigen-Brain" approach- to identify the new candidate cell-type specific transcripts and the biological interpretations have been done for the results. In addition, region-based clustering for spatial expression patterns of the cell-type specific genes has been applied and it revealed unique anatomic expression patterns from each cell-type specific genes. We also focused in a deep analysis of the brain transcriptome using high throughput sequencing technologies.

2. Body

2.1. Brain Image Preprocessing

Our first task was to preprocess *in situ* hybridization (ISH) brain images which were provided by Allen Brain Atlas (ABA) in order to remove the side effects such as noise, background, and inconsistent orientation etc. To deal with these issues, we applied three preprocessing steps: *Brain Extraction*, *Noise Image Removal*, and *Image Registration* and this preprocessing step helps us to have robust ISH image training set which can be used to identify cell-type specific genes.

2.1.1. Brain Extraction

In most of brain images, mouse brain itself has been located in the center of images by the automatic image capture system of the ABA. However, they have different margin background, and it need to be removed. For the brain extraction, we remove rows or columns that don't include any information in terms of pixels.

2.1.2. Noise Image Removal

There are some expression images that barely have expression pattern. It is not possible to use these images as training set for an effective learning. Thus, the ISH images only having expression patterns with less than 5% of area and less than 10% of density are removed.

2.1.3. Image Registration

The ISH brain images could have been located in different position or orientation. For an exact and fair comparison, we apply image registration method proposed by [1] to align the images into the same position and orientation. This method considers a subset of affine transformation in

which straight lines remain straight without using any curvature or perspective distortion. Affine transformation is called as a linear transformation with operations such as shifts, rotations, and scaling. For the image registration, we perform the only coarse registration step from the method [2]. This step allows us to align the ISH brain images with the reference image (Figure 1 and Figure 2). However, since many ISH gene expression images hardly have recognizable pattern and in this case, it is hard to compare these expression images with reference images. Thus, we apply two stage image registration procedures. In the first stage, we have used the original ISH images for image registration and find the optimal parameters for shift, rotation, and scaling operations, which maximize the entropy and mutual information between original brain image and reference image. Once we find the optimal parameters for the image registration, we perform the same transformation into the ISH expression images (bottom left and right). Figure 1 and Figure 2 shows the original brain image before image registration (top left) and after image registration (top right) for coronal section and sagittal section.

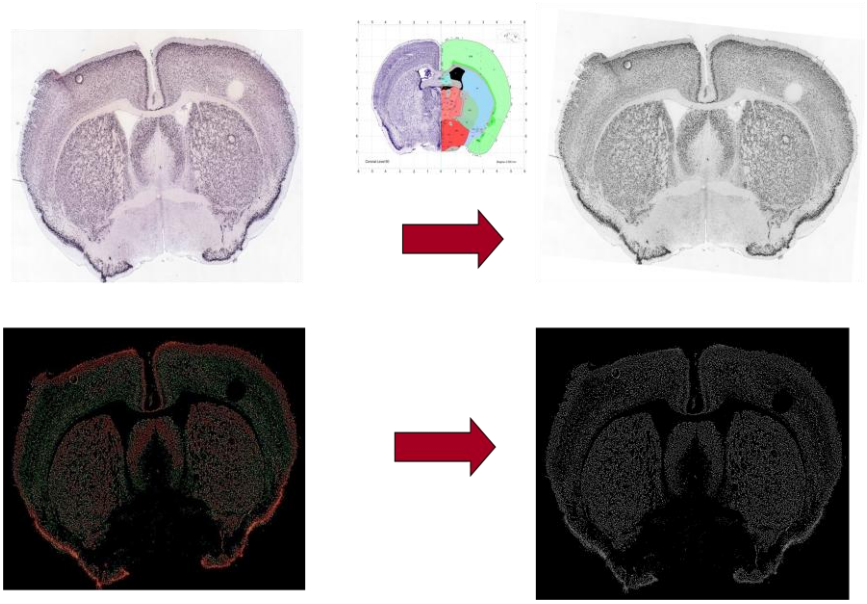


Figure 1. Image registration example for the coronal section

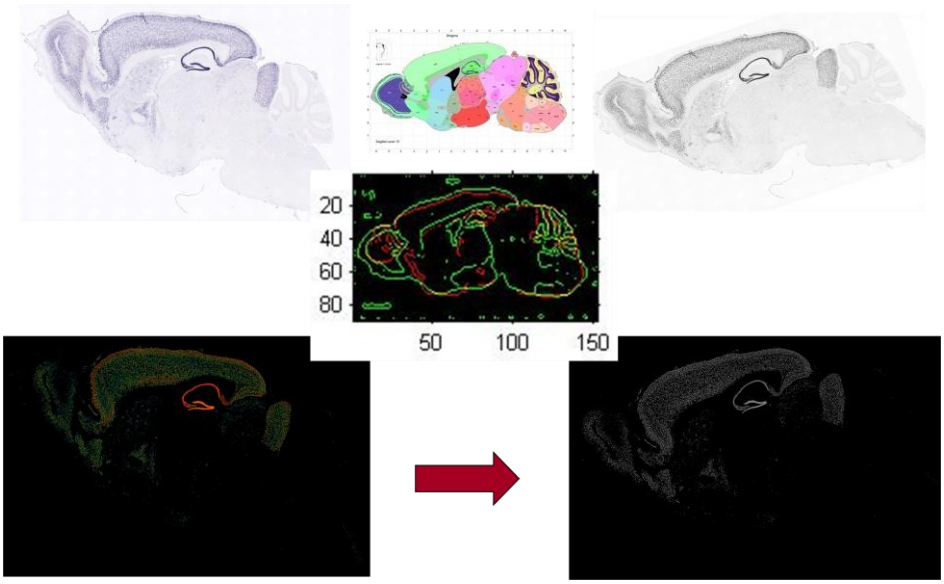


Figure 2. Image registration example for sagittal section (Center image shows the procedure to align the original image with the reference image)

2.2. Feature (i.e. intensity and density) extraction

To classify the brain expression images of cell-type specific genes, we start to divide each brain expression image with fixed number of patch N (e.g. N = 100) and extract the two representative features: intensity (or brightness) and density per each patch.

$$\text{intensity}_{\text{Patch}} = \frac{\sum_{p \in Patch} brightness(p)}{|Patch|} \qquad \text{density}_{\text{Patch}} = \frac{\sum_{p \in Patch} I(p)}{|Patch|}$$

where *I(p)* is an indicator function which has 1 if p has an expression value, otherwise 0.

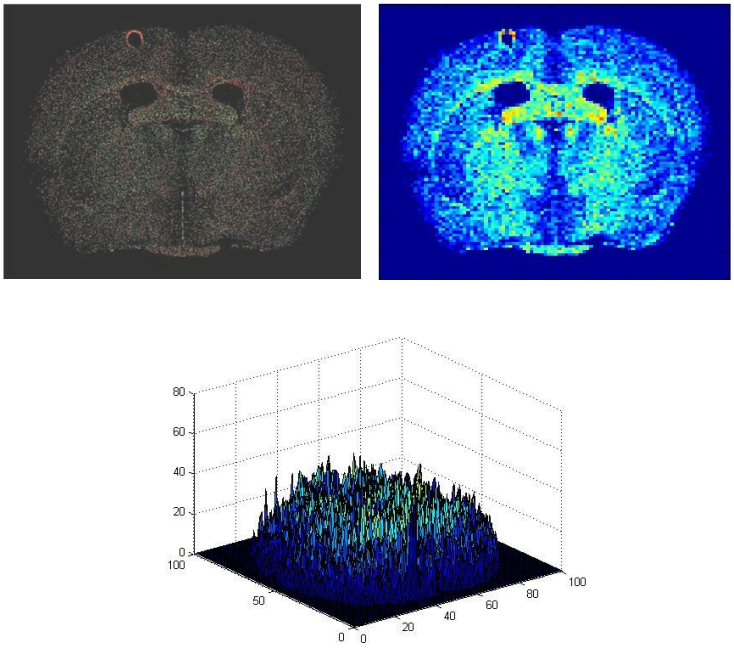


Figure 3. The original and recovered images with a summarized intensity feature vector (upper) and the distribution of intensity feature over patches in brain expression image (bottom).

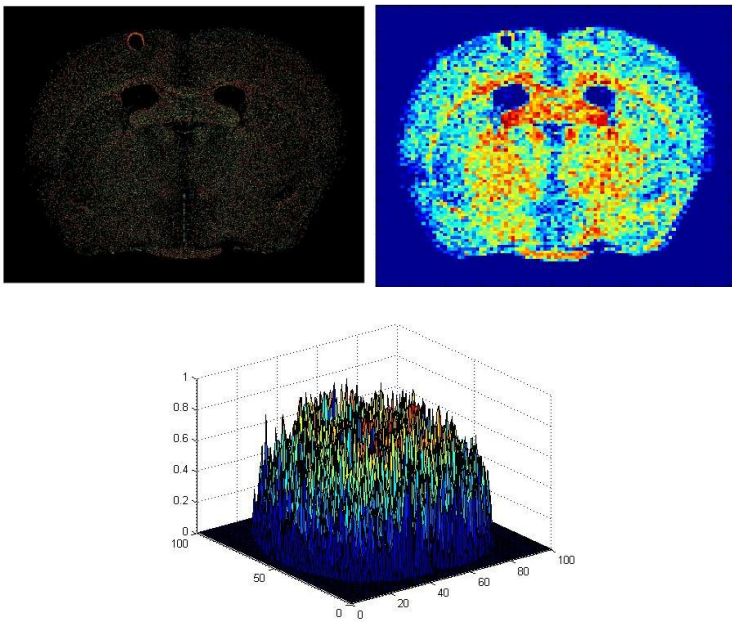


Figure 4. The original and recovered images with a summarized density feature vector (upper) and the distribution of density feature over patches in brain expression image (bottom).

The distribution of intensity and density features over patches in brain expression image has been shown in Figure 3 and Figure 4. Recovered image with a summarized feature vector (i.e. intensity and density) confirmed that extracted feature vector represent the original image enough well.

2.3. Training set for cell-type specific gene classification

We compiled three different cell-type specific gene lists such as oligodendrocytes-enriched genes, astrocytes-enriched genes, and neuron enriched gene from literature [3]. There are two different brain images depending on cutting section of brain: coronal section and sagittal section. Table 1 shows the number of training set per each cell class. Some noisy images have been removed through the following preprocessing step.

Table 1. The number of training set per cell class

	Coronal Section	Sagittal Section
Oligodendrocytes	75	182
Astrocytes	95	231
Neurons	338	496
Total	508	909

2.4. Cell-type specific gene classification with SVM (support vector machine)

To validate the usefulness of extracted features (e.g. intensity and density), we first performed a classification experiment. We applied a standard SVM (i.e. *libsvm* package [4]) for multi-class classification. We experimented with various kernels including the linear kernel, polynomial kernel, radius kernel, and sigmoid kernels and parameters for each kernel in SVM are optimized using a 10-fold cross validation. We compared the classification result with different kernels and different brain sections (Table 2 and Table 3).

Table 2. 10-fold cross validation accouracy over different kernel and different brain sections

CV accuracy	Linear Kernel	Polynomial Kernel	Radius Kernel	Sigmoid Kernel
Coronal section	78.3465%	67.8161%	66.6667%	65.5172%
Sagittal section	70.7371%	65.7866%	60.176%	42.4642%

Table 3. Sensitivity, specificity, and precision for coronal section and sagittal section with SVM

	Coronal			Sagittal		
	Oligodendrocytes	Astrocytes	Neuron	Oligodendrocytes	Astrocytes	Neuron
Sensitivity	63%	53%	89%	57%	71%	76%
Specificity	95%	92%	65%	91%	81%	82%
Precision	70%	62%	84%	62%	56%	83%

2.5. Eigen-Brain Approach

In this approach, we decompose brain expression images into a small set of characteristic brain expression image patterns called "*Eigen-Brain*". This technique has been firstly introduced for face

recognition in [2]. *Eigen-Brain* can be thought as a visualization or ghost image of principal components over training set and thus, these *Eigen-Brains* explain the variation of brain expression pattern of our brain images and represent the major characteristics of expression image patterns. Each individual brain image can be represented as a linear combination of the *Eigen-Brain*s. These *Eigen-Brains* are calculated using our training set, keeping only *M Eigen-Brains* that correspond to the highest eigen values. The detail of method will be explained in a following section.

These *M Eigen-brains* define the new feature space and we project all brain expression images in our training set into this new feature space (e.g. *Eigen-Brain* space). It gives not only large dimension reduction benefits but also clear expression pattern for cell type specific genes, which will be further helpful to characterize and define the cell-type specific expression patterns. After projecting into the new feature space, the weight vector for each image is calculated, which describes the contribution of each *Eigen-Brain* for the representation of the image. It also can be thought as new coordinates of image in *Eigen-Brain* space. Once we have these weight vectors, they will be used to classify the unknown expression image by comparing these weight vectors.

2.5.1. Method

Let the brain expression image of training set $\Gamma_1, \Gamma_2, \ldots, \Gamma_N$. The average brain image of the training set is $\Psi = \frac{1}{N}\sum_{n=1}^{N}\Gamma_n$ and each brain image is different from the average image by $\Phi_i = \Gamma_i - \Psi$. The eigen vector μ_k for covariance matrix *C* of our training set is calculated by (1).

$$C = \frac{1}{N}\sum_{n=1}^{N}\Phi_n\Phi_n^T \qquad (1)$$

Once eigen vectors are calculated, M *Eigen-Brains* are used to define the new feature space. All images are transformed into this "*Eigen-Brain space*" (2).

$$\omega_k = \mu_k^T(\Gamma - \Psi) \text{ , where } k = 1, \ldots, M \qquad (2)$$

Thus, the weight vector (Ω^T) for new image (Γ) is $\Omega^T = [w_1, w_{2,\ldots,} w_M]$ and this weight vector is compared with weight vectors in training set. The decision for the cell type specific gene classification is made by finding the most similar gene expression image's class (cell) label by borrowing a KNN concept. For this experiment, we have applied unanimity vote scheme, which guarantee that the new data set are classified with high confidence and it only returns very accurate cell-type specific gene.

2.5.2. Classification Result

We applied the *Eigen-Brain* approach into our training set and the accuracy and specificity for the sagittal section has been reported in Figure 5 and Figure 6. The result for coronal section has been reported in Figure 7 and Figure 8. The x-axis in Figures represents the degree of variance covered by *Eigen-Brains* and the larger number of variance means that we are using many *Eigen-Brains*

as a new feature space. It shows that the degree of variance or number of *Eigen-Brains* doesn't affect the classification accuracy. Moreover, as we can see, our *Eigen-Brain* approach achieves very high accuracy and high specificity that represents that our approach can classify the cell-type specific gene with high confidence. Even though the specificity for neuron cell-type specific gene in coronal section is relatively less than others, it is still high with 80% specificity and it can be covered by high accuracy. In all other cases, it shows more than 95% specificity and around 99% accuracy. Especially, these high specificities guarantee the low false positive rate and it allows us to identify real cell-type specific genes.

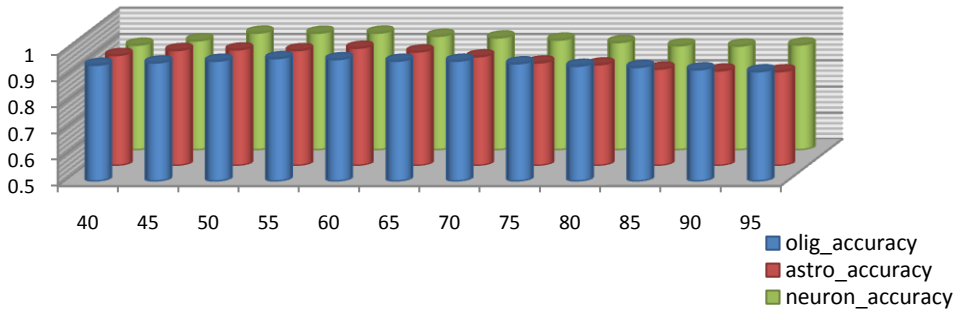


Figure 5. Accuracy per cell class classification in sagittal section

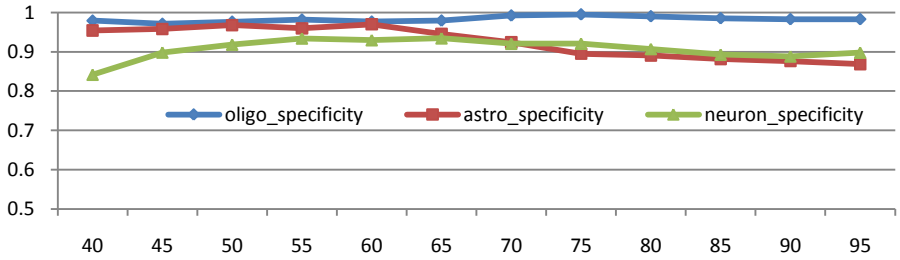


Figure 6. Specificity for cell class classification in sagittal section

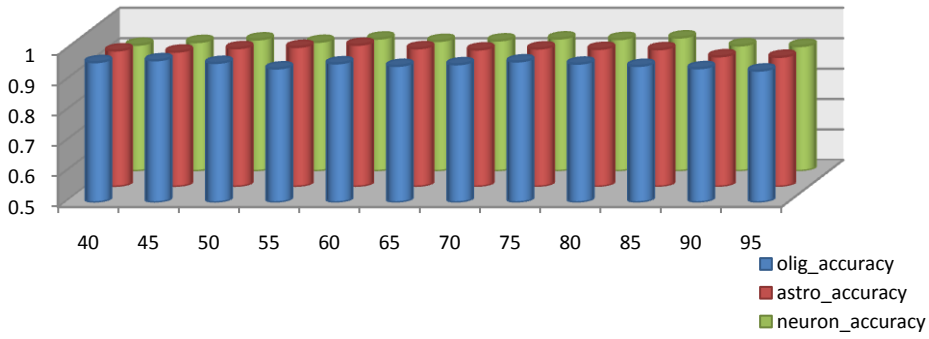


Figure 7. Accuracy per cell class classification in coronal section

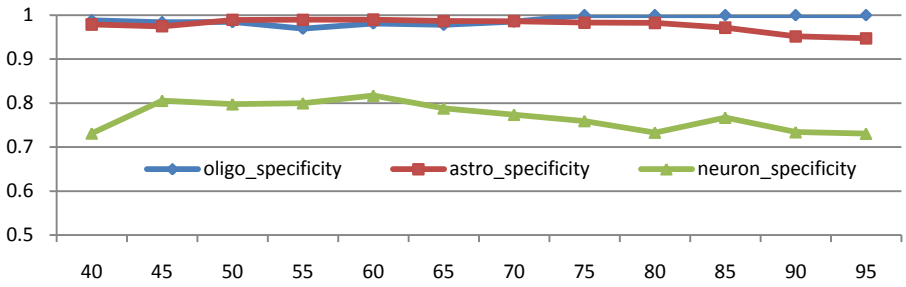


Figure 8. Specificity for cell class classification in coronal section

2.5.3. EigenBrains per each cell type

In each *Eigen-Brain* images, highly expressed brain regions are appeared as Figure 9 illustrates the *Eigen-Brains* for neuron cell-type specific genes in coronal section. The red circle region in Figure 9 represents the *VL lateral ventricle* region that is related with neurological condition and is on average larger in patients with schizophrenia and bipolar disorder. The blue circle region represents the *cerebral cortex* (CTX) region that is a sheet of neural tissue that is outermost to the cerebrum of the mammalian brain and takes a key role in memory, attention, perceptual awareness, thought, language, and consciousness.

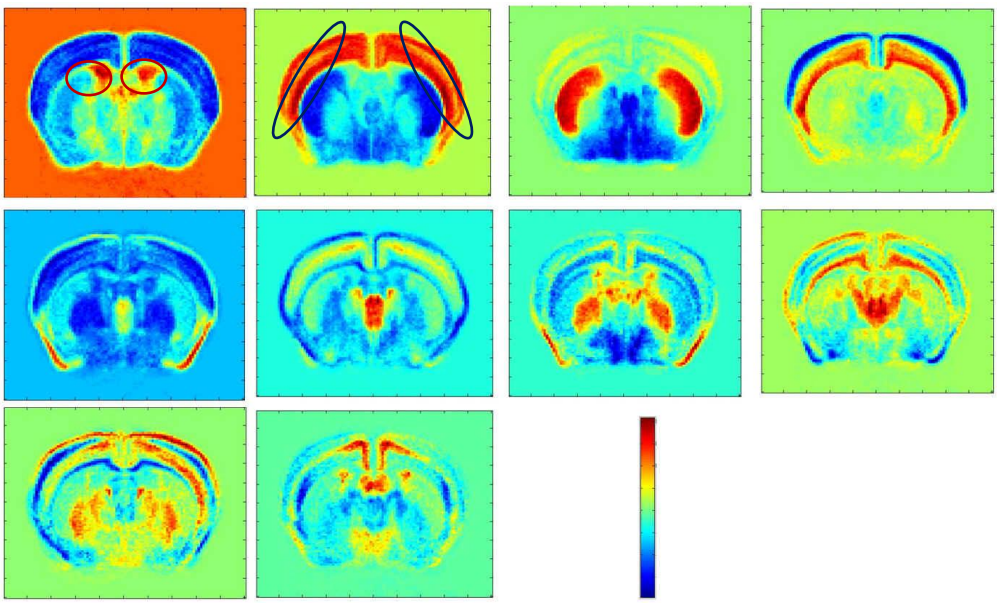


Figure 9. Eigen-Brains for neuron specific genes in coronal section

The red circle region in Figure 10 represents the *optic chiasm* (och) region that allows for the right visual field to be processed. Similar analysis can be applied to other *Eigen-brains* to identify the specific region especially responsible to characterize the specific cell type.

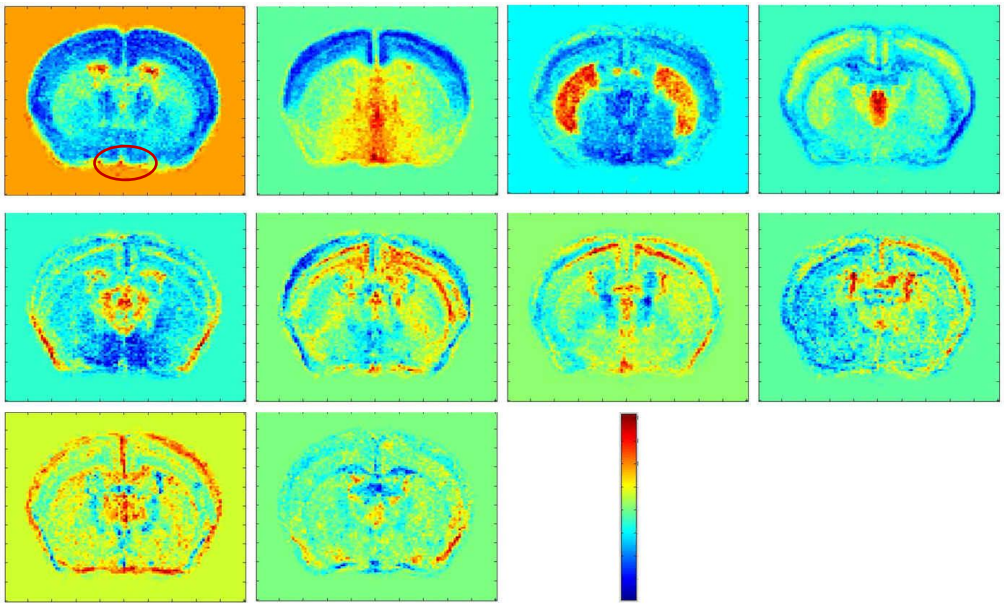


Figure 10. Eigen-Brains for astrocytes specific genes in coronal section

Black circle in Figure 11 illustrates the nucleus of the lateral olfactory (NLOT) region, which is highly expressed in oligodendrocytes specific genes.

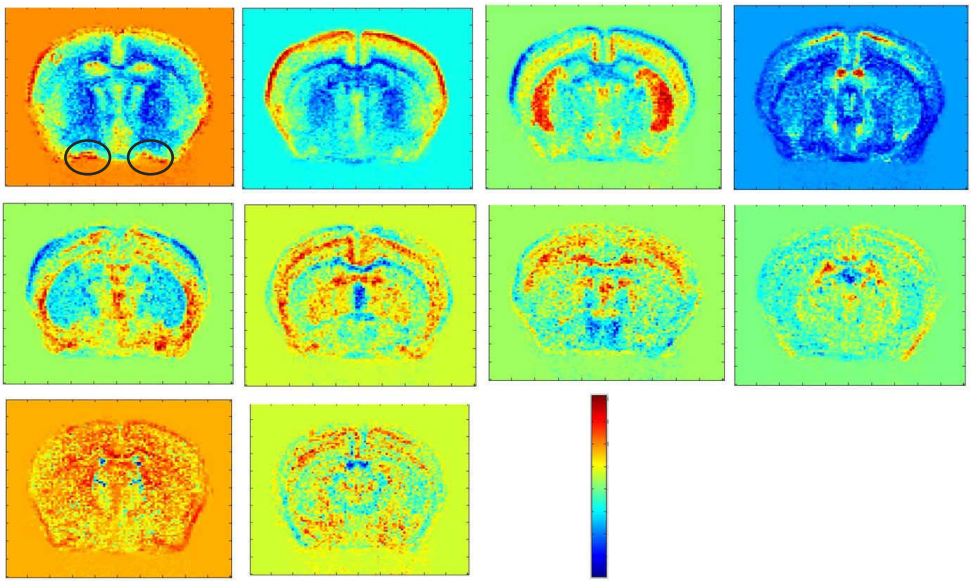


Figure 11. Eigen-Brains for oligodendrocytes specific genes in coronal section

Figure 12, Figure 13, and Figure 14 represent the Eigen-Brains per each cell type in sagittal section.

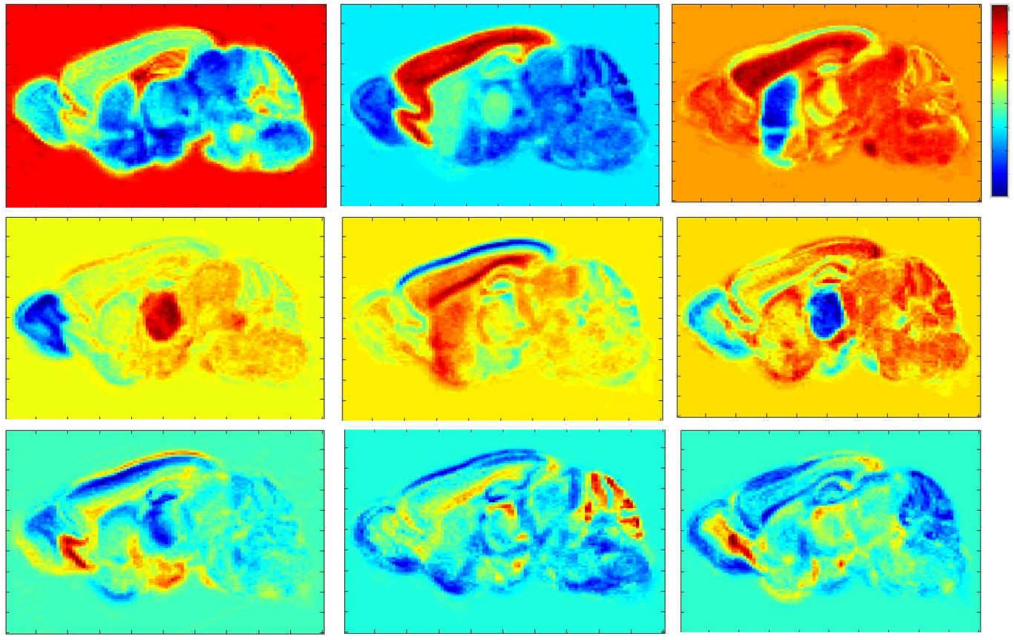


Figure 12. Eigen-Brains for neuron specific genes in sagittal section

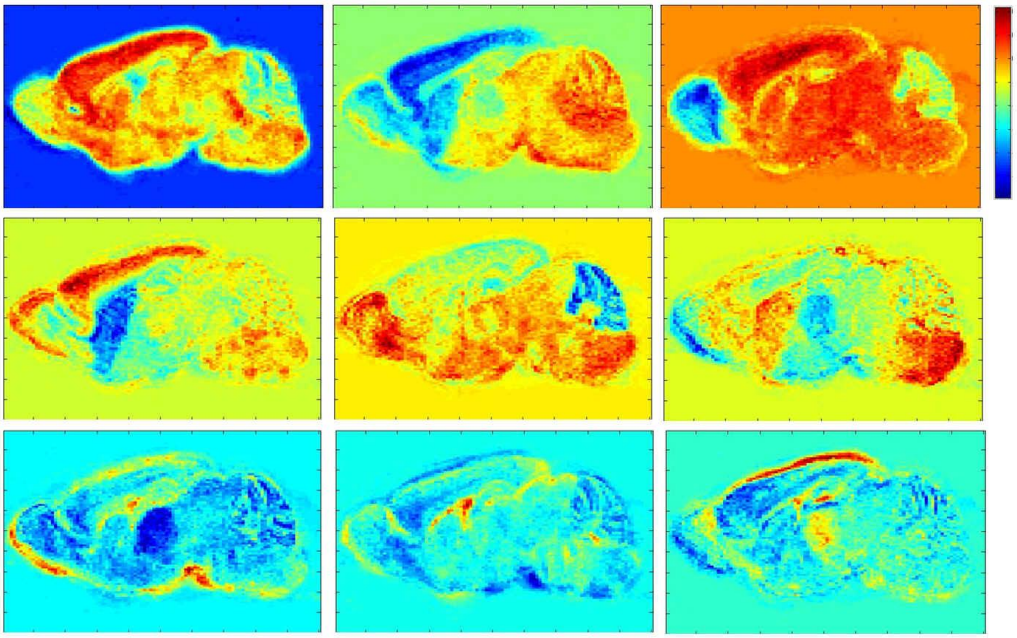


Figure 13. Eigen-Brains for astrocytes specific genes in sagittal section

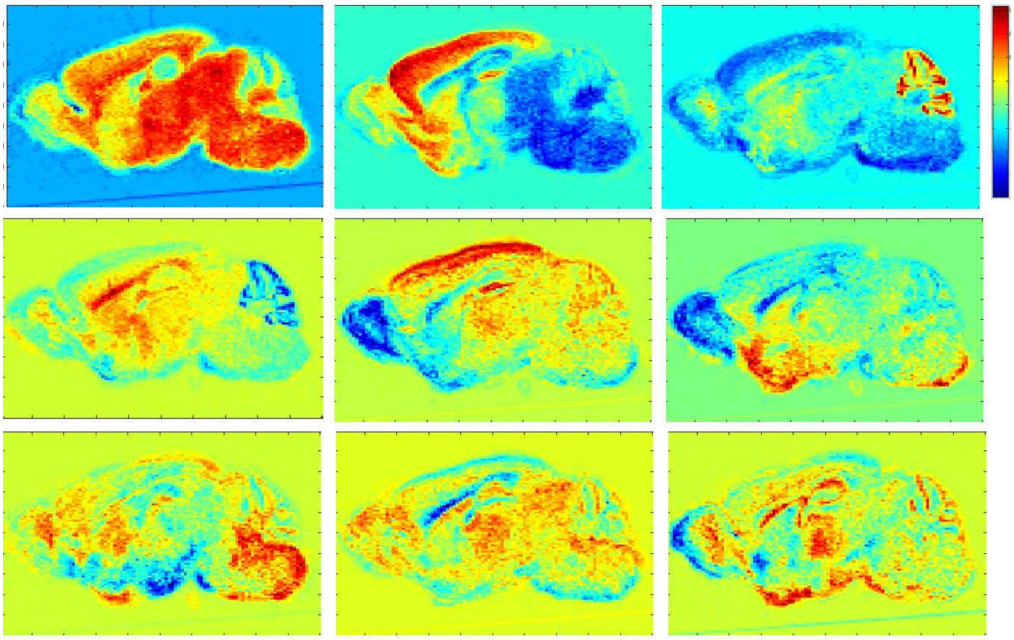


Figure 14. Eigen-Brains for oligodendrocytes specific genes in sagittal section

2.6. Test data set for cell-type specific gene identification

We downloaded all the original mouse brain images and their corresponding gene expression images using a Perl script, which has been written during our first quarter of last year. The script parsed the XML files to extract the associated image file information and download the correspondent image. Table 4 shows the statistics of test data set that will be used for our cell type specific gene identification. In average, each gene has about 2-3 ISH brain images from the section we are focusing on. Thus, the number of genes in our test data set is less than the total number of ISH brain images.

Table 4. Statistics of Test data set

	# of image files	# of genes	Total size
Coronal Section	7341	~ 3600	27 Gigabyte
Sagittal Section	29112	~12000	145 Gigabyte

2.6.1. Image registration for Test data set

The ISH brain images have not been aligned in the same position or orientation. In addition, the image size itself is also various. Thus, for a fair comparison, alignment of the orientation and adjustment of the size of images are necessary. Here, we apply the image registration technique [1] into the *in situ* hybridization (ISH) brain images that were provided by Allen Brain Atlas (ABA) in order to remove the side effects such as noise, background, and inconsistent orientation etc. This method considers a subset of affine transformation in which straight lines remain straight without using any curvature or perspective distortion. Affine transformation is called as a

linear transformation with operations such as shifts, rotations, and scaling. We apply two stage image registration procedures to transform the ISH expression images. In the first stage, the original ISH images have been used for image registration and the optimal parameters for shift, rotation, and scaling operations have been estimated, which maximize the entropy and mutual information between original brain image and reference image. In the second stage, these estimated optimal parameters have been used to transform the ISH expression images.

For handling the test dataset, new challenging problem is occurred by the huge number of images and their size (Table 4). In average, image registration process takes about 4 minutes to complete one image registration, which means that it would take more than 3 months to finish this image registration process for our test data set if this process is applied sequentially. Thus, we have used our cluster machine composed of 10 nodes with 38 CPUs, which allows us to finish this process in two weeks.

2.6.2. Classification using EigenBrain approach

In this project, we have proposed a new approach to identify the new candidate cell type specific genes using what we call the *EigenBrains*. Here, we briefly summarize the overall procedure (Figure 15). First, we applied an image registration technique to align all the brain expression images. After this image registration then the specific features (e.g. density per each patch) are extracted. This procedure results in very high-dimensional data. Thus, to reduce data dimensionality, we applied our *EigenBrain* approach that transforms the original feature vector into the new *EigenBrain* space (projected into low dimensional brain space). This transformation helps to reduce the high dimensional feature space into the low dimensional feature space – and is shown to greatly improve the accuracy of our algorithm to successfully identify cell-type specific transcripts.

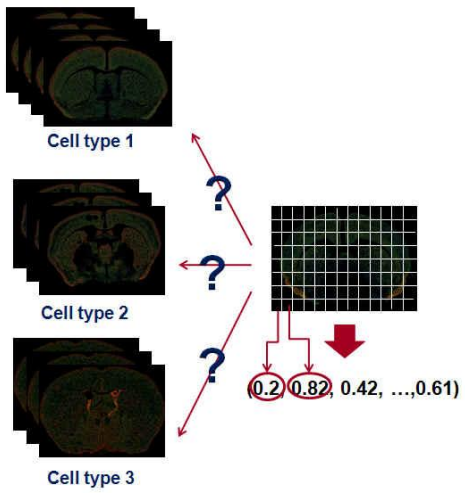


Figure 15: Overall procedure for classification

This reduced new feature set is made more biologically meaningful by representing the regions instead of independent patches (Figure 16). Once the original features were transformed into the new

feature vector, we applied a K-neighest neighbors (KNN) algorithm with unanimous vote scheme to identify the candidate cell-type specific genes.

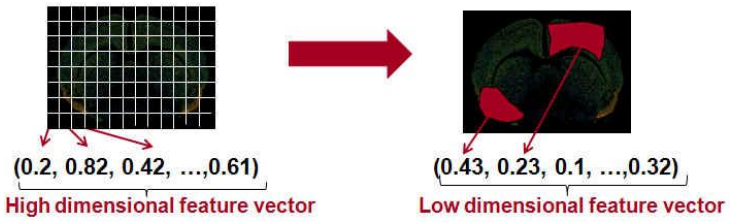
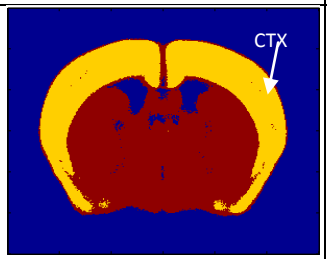
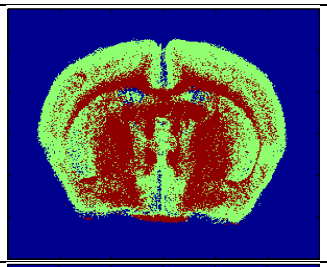
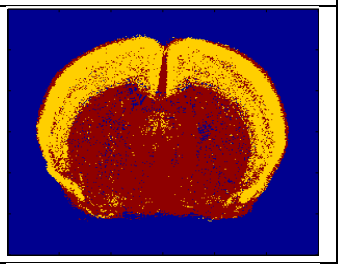
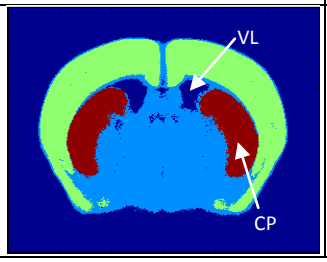
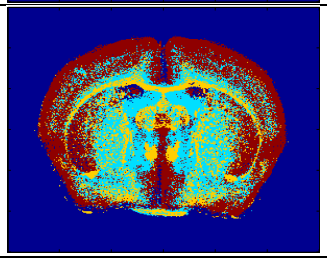
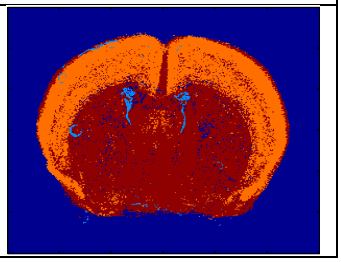


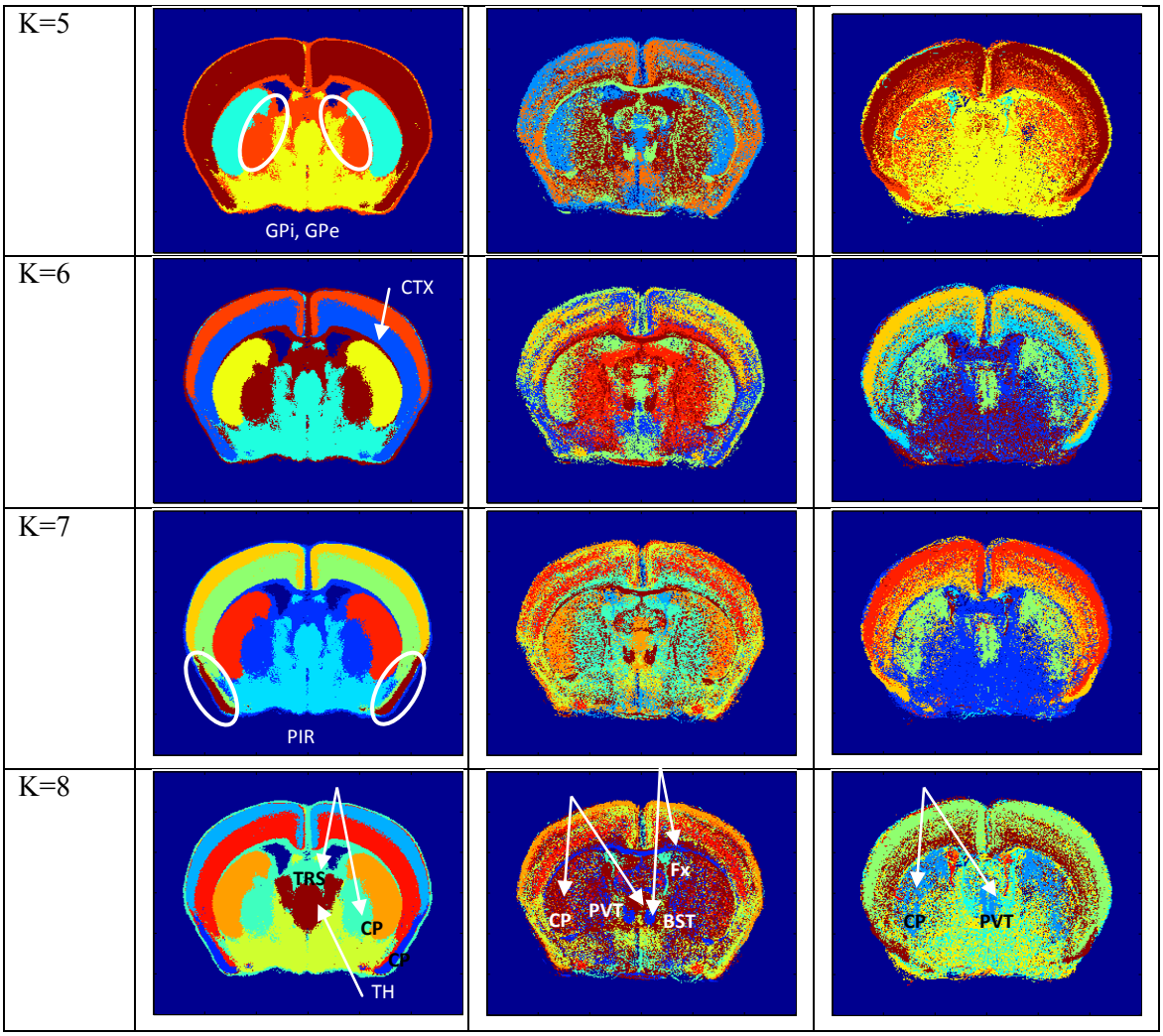
Figure 16: Transformation of original feature into the EigenBrain space

2.7. Region based clustering for different cell-type specific genes

To characterize the cell-type specific genes, we applied region-based clustering to the spatial gene expression. For computational efficiency, we first reduced the original *in situ* hybridization images into 300×300 pixels using bicubic interpolation, the output is a weighted average of pixels in the nearest 4×4 neighborhood. Thus, each pixel represents the averaged expression values of approximately 600 pixels on a coronal section and 1100 pixels on a sagittal section. Then, we applied the K-means clustering algorithm to group summarized pixels based on their expressions across all cell-type specific genes within same cell specificity and the results for cell-type specific gene expression are reported.

Table 5: Region-based clustering based on coronal section of brain images

Number of Cluster	Neuron	Oligodendrocytes	Astrocytes
K=3			
K=4			



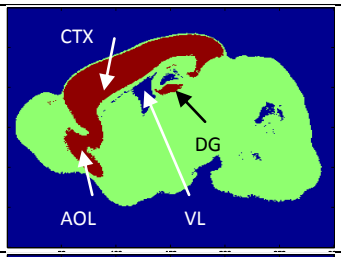
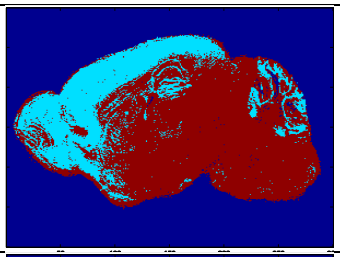
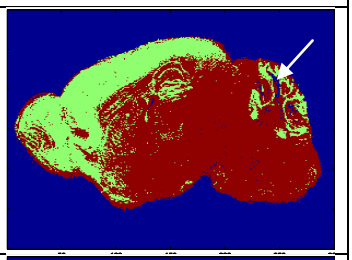
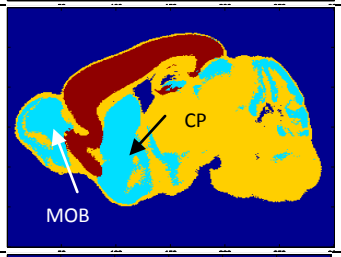
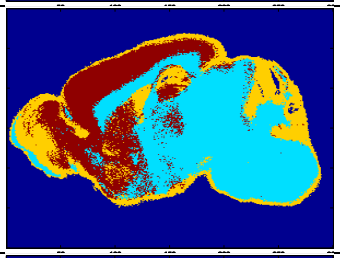
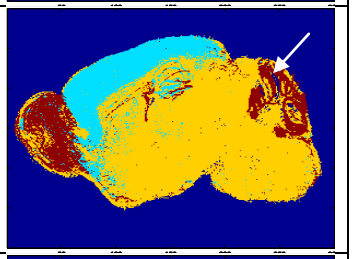
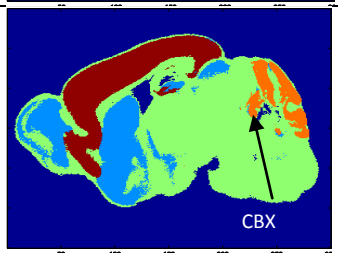
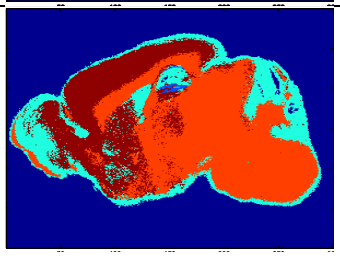
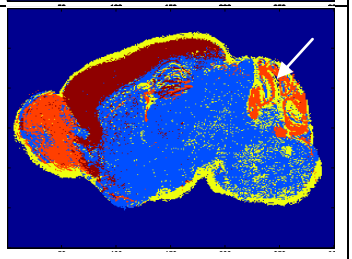
2.8. The discovery of brain anatomical region on coronal and sagittal sections

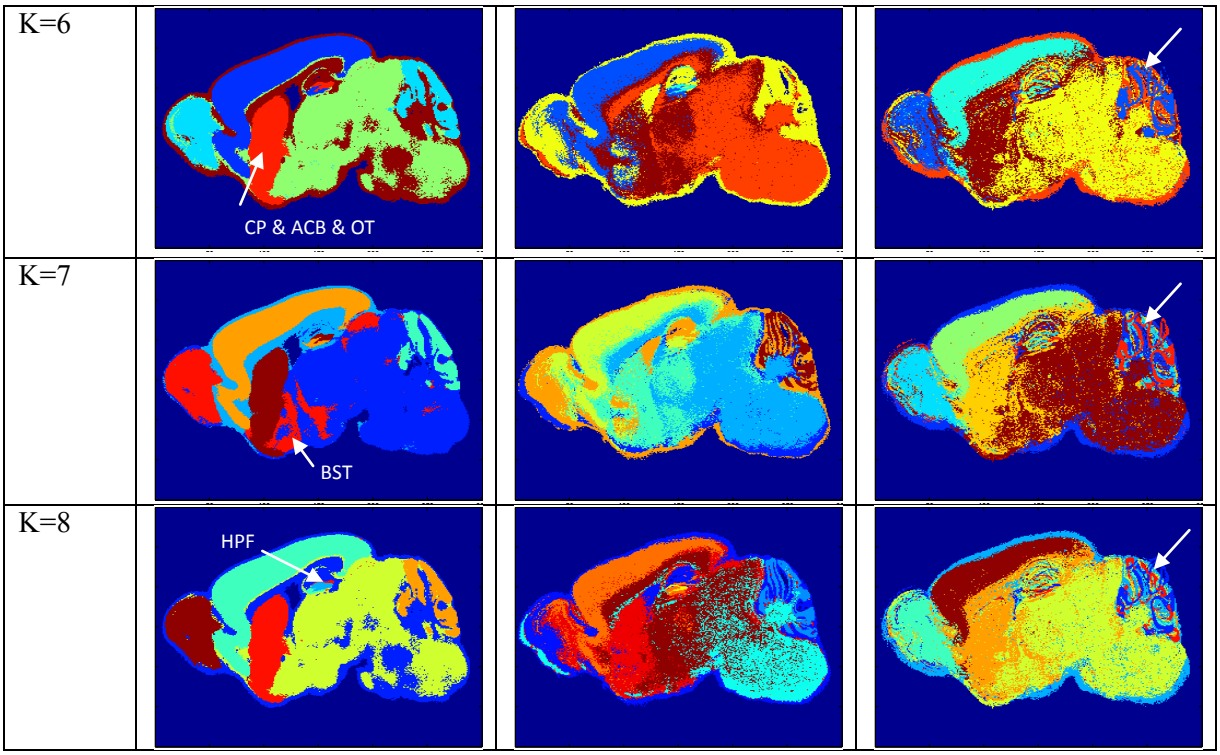
Table 5 shows the region-based clustering results on the coronal section for cell-type specific genes of three different cells (i.e. neuron, astrocytes, oligodendrocytes). The most intriguing characteristic of this spatial gene expression patterns was discovered from neuron cell-type specific genes (left column in Table 1). As can be seen, the region-based clustering result reveals dramatically clear patterns from neuron cell-type specific genes, consistent with classical anatomic brain structure. In particular, by increasing the number of clusters, the new anatomic brain regions are detached from previous clustering results. As shown in the left column of Table 5, cerebral cortex (CTX) region called gray matter is distinctly separated from white matter region in the brain at the beginning. As K is increased to 8, the caudoputamen (CP), Globus pallidus internal and external segment (GPi, GPe), inner cerebral cortex (CTX), piriform cortex (PIR), corpus callosum (CC), fimbria (Fi), and thalamus (TH) etc are clearly separated from white matter region of the brain in the order named. This reveals overall concordance between spatial gene expression patterns with the anatomical regions of mouse brain. For example, after the separation of the cerebral cortex region, the caudoputamen region

distinctly comes out. This caudoputamen region is known to be related with neurogenesis process whose role is important for learning and memory. Thus, the clear separation of this brain region in neuron cell-type specific genes can be explained by its functional connection with neuron cells. However, such a clear anatomical pattern could not be detected for other cell types (e.g. oligodendrocytes and astrocytes) and instead, more lousy patterns are observed comparing to the neuron cell-type specific genes.

Table 6 represents the similar anatomical structure from the region-based clustering of spatial gene expressions on a sagittal section. In particular, the cerebral cortex (CTX) and anterior olfactory nucleus (AON) are separated very clearly. In addition, the dentate gyrus (DG), the part of the hippocampal formation (HP), also comes out distinctly, which is implicated in new memory. This region is known to be related with high rates of neurogenesis in adult human [5]. Such a clear separation of this anatomical brain region is revealed especially in the neuron specific genes on both sagittal section and coronal section. Furthermore, astrocyte specific genes show striking region-specific expression pattern in the cerebellar cortex (CBX) region.

Table 6: Region-based clustering for sagittal section brain images

Number of Cluster	Neuron	Oligodendrocytes	Astrocytes
K=3			
K=4			
K=5			



2.9. Identification of uniquely co-regulated brain regions for cell type

Another interesting observation we found is the discovery of co-expressed brain anatomical regions per each cell type. At K=8, the caudoputamen (CP) region is clustered together with the paraventricular nucleus of the thalamus (PVT) region on both oligodendrocytes and astrocytes cell-specific genes (marked with white arrow in Table 5). The columns of the fornix (Fx) region are also clustered with the bed nuclei of the stria terminallis (BST) region in oligodendrocyte specific genes (marked with white arrow in Table 5). For neuron specific genes, the lateral septal nucleus (Lsc), caudal or caudodorsal part, triangular nucleus of septum (TRS), and septofimbrial nucleus (SF) regions have co-expressed together with the glubus pallidus external segment (GPe) region. Furthermore, unlike other cells, the lateral ventricle (VL) region never expressed for neuron cell-type specific genes across all Ks on both brain sections. This VL region is known to be increased with age and enlarged in a number of neurological conditions. Furthermore, this region is usually larger for schizophrenia, bipolar disorder, and Alzheimer's disease patients than normal people. Regardless of K values, there are several co-expressed anatomical regions that are co-expressed on the sagittal section too (Table 6). For example, the cerebral cortex (CTX) and anterior olfactory nucleus (AON), caudoputamen (CP) and nucleus accumbens (ACB) show similar expression patterns clustered together for neuron cell type specific genes.

2.10. Correlation Matrix for Region based clustering

Figure 17 demonstrates the correlation matrix between left and right hemispheres. The left bottom corner represents the correlation of center position of two hemispheres.

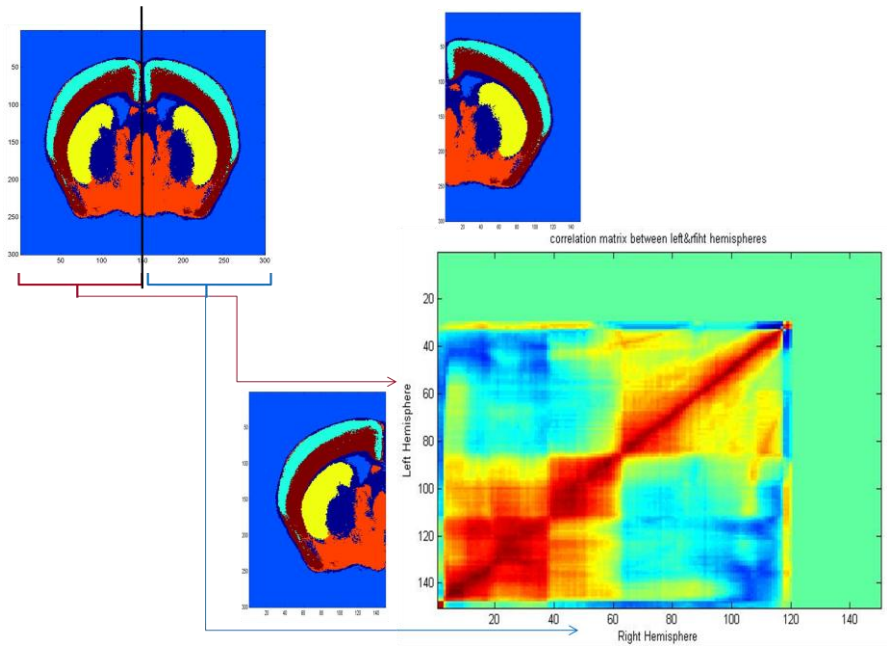
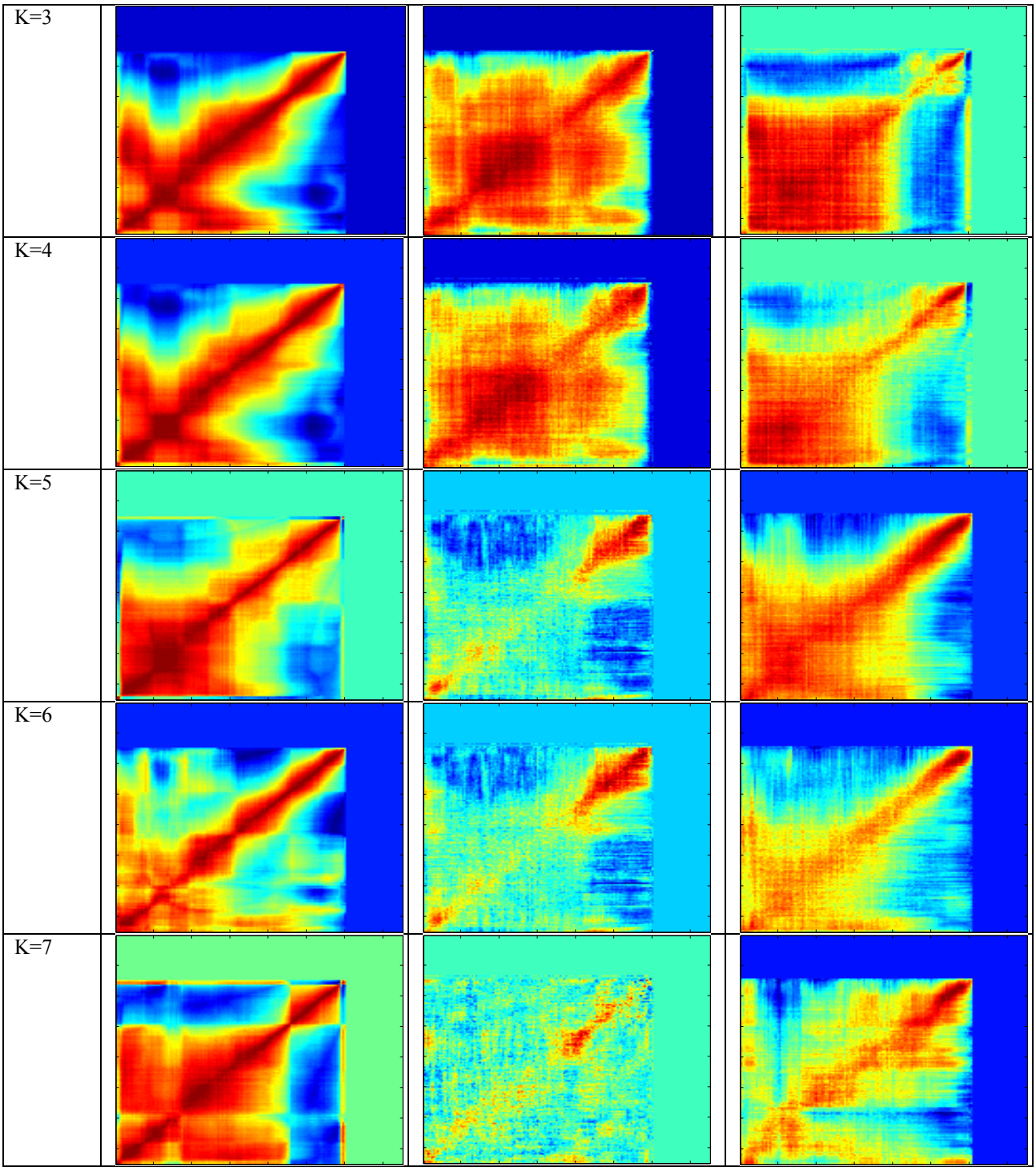


Figure 17: Correlation Matrix between left and right hemisphere

Table 7 shows the correlation matrix between left hemisphere and right hemisphere for region-based clustering results on coronal sections (Table 5). The diagonal line is mapped to the exact symmetry position in the both hemispheres. In the human brain, unique functions seem to be controlled in the left or right hemisphere. For example, language ability is usually predominated by a left hemisphere, while spatial recognition is controlled by a right hemisphere [6]. However, this fact is somewhat counter-intuitive compared to what we observed from ISH spatial expressions. Based on the gene expression or cell distribution, we can discover that there is no clear difference between left and right hemisphere at the perspective of spatial gene expression. Furthermore, such a clear symmetric pattern is identified more clearly in neuron cell-type specific genes (the first column of Table 7). Unlike other cell-type specific genes, diagonal line in correlation matrix of the neuron cell-type specific genes is marked as a red color, which means the clear symmetry between left and right hemisphere regardless of K values. Especially, oligodendrocyte specific genes show the least symmetric pattern comparing to other cell types. Even though there are many research have been done to discuss the asymmetry of brain functions and expression (Sun et al., 2005), they also reported that such an asymmetry is only detected at earlier stage in the fetal brain and clear left-right expression difference is diminished at the latest stage of brain (19-week -old human brain) [7]. Since Allen Brain Atlas (ABA) *in situ* hybridization images are obtained from 56 days adult mouse brain, such an asymmetric expression pattern cannot be detected unlike a fetal brain. It is supported by investigating the *in situ* hybridization images of genes known for asymmetrically expressed genes [7] in Figure 18.

Table 7: Correlation Matrix between left and right hemisphere

Number of Cluster	Neuron	Oligodendrocytes	Astrocytes


K=3
K=4
K=5
K=6
K=7

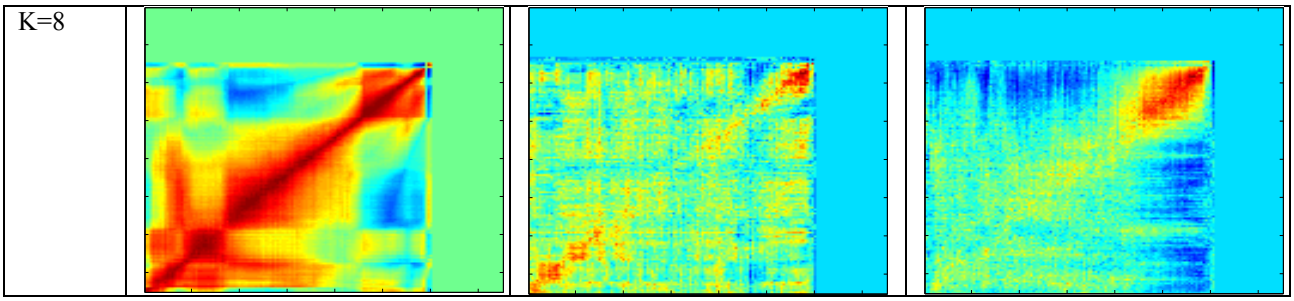
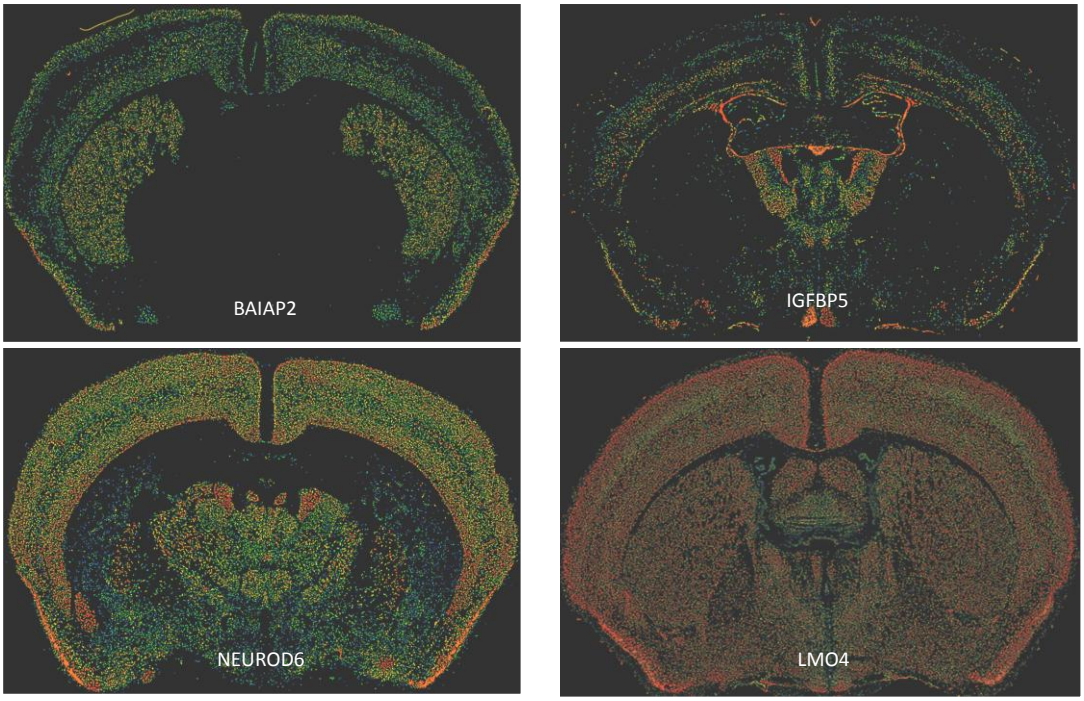


Figure 18 represents the *in situ* hybridization images of representative genes known for asymmetrically expressed gene by [6]. Three genes (i.e. BAIAP2, NEUROD6, SH3GL2) in the left column of Figure 18 were known as highly expressed ones in the left hemisphere, while IGFBP5, LMO4 and STMN4 (right column of Figure 19) were verified as differentially expressed ones in the right hemisphere of 12-week-old human fetal brains through either real-time reverse transcription (RT) - PCR or *in situ* hybridization [6]. However, as we can see from Figure 18, those genes didn't reveal clear asymmetrical expression pattern from *in situ* hybridization images from ABA. As we pointed out, such an asymmetry clearly tends to appear in early fatal brains. However, it may not necessarily be present for adult brains. Furthermore, the other reason is that the subtle expression difference between left and right hemispheres appeared only in small number of genes can be faded away because we are averaging the expression patterns of all cell-type specific genes for region-based clustering.



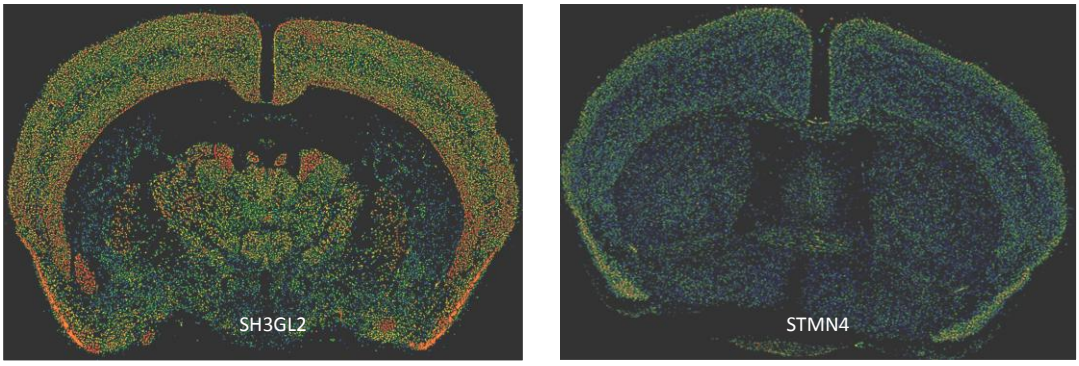


Figure 18: *in situ* hybridization images known for asymmetrically expressed gene

In order to compare the degree of correlation in different cell-type specific genes, we calculate the Pearson correlation between left and right hemisphere (diagonal line in Table 7) and Figure 19 shows correlation across all K values. As shown, neuron specific genes reveal prominent symmetric expression patterns, while oligodendrocyte specific genes show the least symmetric expression pattern.

2.11. T-test to identify differentially expressed regions for each cell type

To investigate the highly/lowly expressed brain regions for an each cell type, we applied unpaired t-test between different cell-type specific genes at the different levels: intensity and density. At the gene expression intensity level, we reduced the original *in situ* hybridization images into 300×300 pixels using bicubic interpolation as we did for K-means clustering and applied quantile normalization. Then, t-test has been applied to identify the highly differentially expressed region (at a

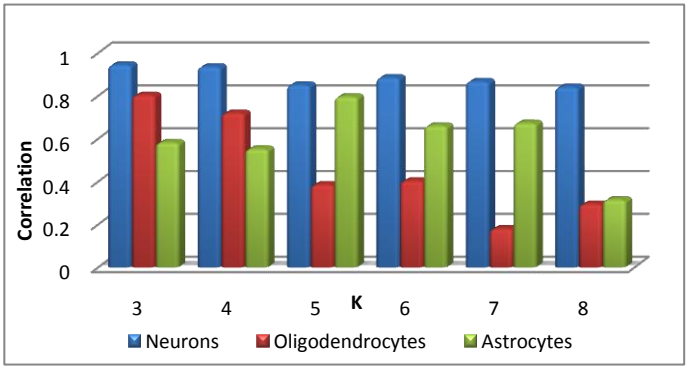


Figure 19: Correlation between left and right hemispheres

P value < 0.05). The unique anatomic brain region that is highly expressed for specific cell could be discovered using this test. In addition, the similar analysis has been done at the density level. For this test, original images have been divided by 100×100 patches and each patch value is represented as a density value of this patch, which shows how many pixels in this patch are expressed. Density level based t-test helps us to identify the highly differentially expressed anatomical brain regions and this result demonstrates the relation of cell distribution in the brain.

2.12. Microarray analysis

We have studied the gene expression in human and mouse "atlas" from the GNF data sets (NCBI GEO GSE1133), and we expanded our knowledge with other microarray studies, (Table 8).

Table 8. Microarray datasets utilized

NCBI GEO	Specie	# Tissues	Reference
GSE1133	Human	79	[8]
GSE3526	Human	65	NA
GSE2361	Human	36	[9]
GSE1133	Mouse	61	[8]
GSE9954	Mouse	22	[10]
GSE10246	Mouse	96	[11]

All the datasets selected are public for academic usage. Few samples are shared, because each author select with different criteria in the sampling. We basically look for a genomic chip (Affymetrix in particular) and one or more brain-related sample in the dataset. After raw data (CEL files) reading, background correction and RMA normalization in R/bioconductor:affy, we collected the expression patterns of almost all the genes in the human and mouse genome in a local database. Then we used the same criteria to select brain-specific genes, this is 10-fold enrichment in one brain sample contrasted with the maximal expression level detected in the rest of non-brain samples. Also the probe must have a P-value < 0.05. We processed each dataset separately to reduce noise.

We compared the genes selected in each dataset, but we don't see too much overlap (Figure 20).

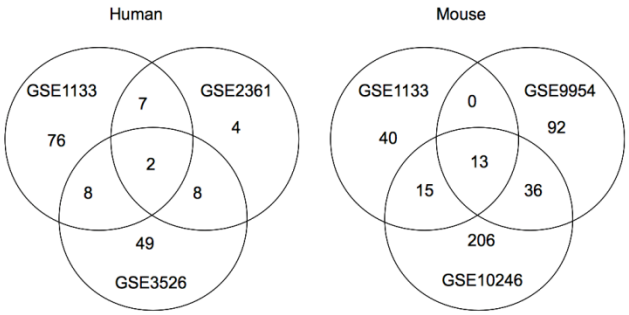


Figure 20. Brain-specific genes overlap between microarray experiments.

We have already predicted a "secretable" probability using SignalP program [12], but we noted that many isoforms of a gene have quite different probabilities to be secreted. We are now integrating RNA-seq data to undercover the isoforms expressed in brain samples.

2.13. RNA-seq analysis

Other exploration in the brain-specific genes was to include RNA-seq data. Our interest is to discover the transcript-specific patterns in brain, as we mentioned before, we used the datsets from NCBI GEO GSE12946 and GSE13652 [13,14], both in combination measure the mRNA levels for 12 human tissues. We mapped and aligned the reads for each dataset to the human reference genome and the gene models (hg18 of UCSC Genome DB) with Bowtie [15], a short-read aligner (Table 9).

Table 9. human reference genome and the gene models

GEO	description	total reads	% unique	% genome	% genes
GSM325476	**brain**	17,246,957	57.35%	62.62%	18.68%
GSM325477	liver	18,517,121	45.60%	57.05%	15.53%
GSM325478	heart	20,169,301	41.81%	55.04%	21.66%
GSM325479	skeletal muscle	22,640,454	46.70%	60.60%	20.93%
GSM325480	colon	28,435,996	48.65%	60.45%	19.58%
GSM325481	adipose	27,752,231	53.47%	62.07%	18.32%
GSM325482	testes	27,303,938	56.58%	67.68%	18.20%
GSM325486	breast	16,120,746	61.41%	66.31%	15.09%
GSM325483	lymph node	27,492,254	50.40%	61.94%	13.98%
GSM343512	**cerebral cortex**	31,940,303	68.61%	62.63%	9.22%
GSM343515	lung	25,862,064	62.77%	59.12%	8.37%
GSM343511	**brain**	17,246,964	57.35%	62.62%	6.99%

In the RNA-seq analysis we excluded reads with more than one hit in the genome or in the gene models, we recovered reads with perfect match (no-mismatches), then we counted the reads per kbp for each gene model. Because the total number of reads is variable in each sample, we use a global normalization to obtain standard values for each gene model (Figure 21).

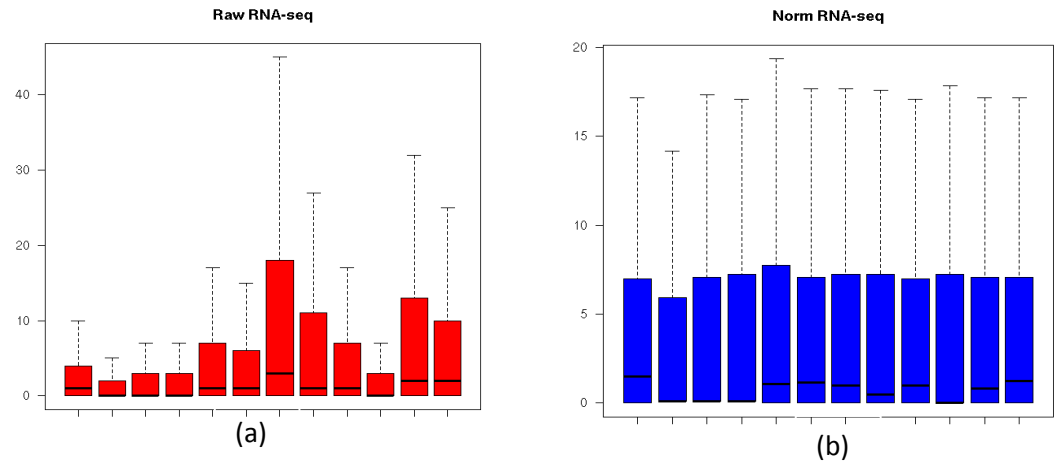


Figure 21. RNA-seq read counts (a) raw data and (b) after global normalization.

We compared the values for brain samples and the global expression in all the datasets (Figure 22 and Figure 23).

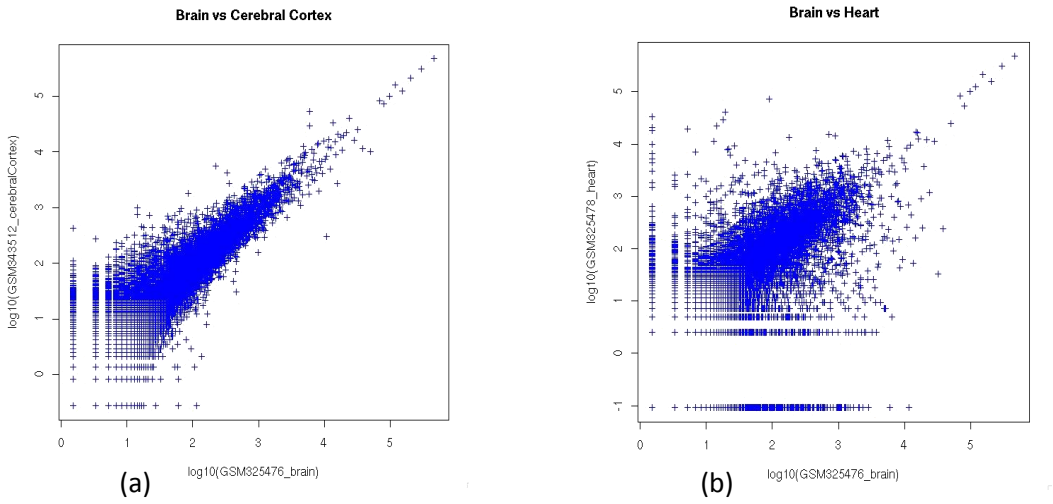


Figure 22. RNA-seq gene expression values for (a) brain and cerebral cortex and (b) brain and heart. All values are log10 scaled.

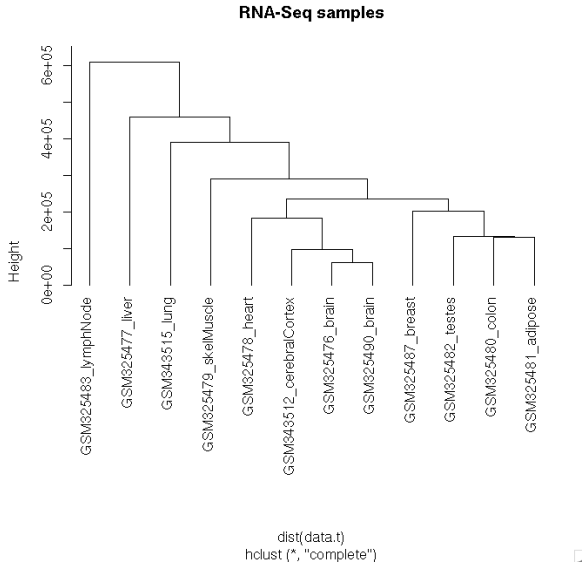


Figure 23. Cluster dendogram for all the RNA-seq datasets.

We used the same method for brain specificity in the RNA-seq data, we obtained a list of 798 transcripts which are potentially brain-specific.

Table 10: RNA-seq samples

ID	SAMPLE	MODEL	EQUIPMENT	READ SIZE	TOTAL READS	Gbp/sample
NHA	Normal astrocytes	Human	SOLiD	50	15,133,796	0.76
H683	GBM cell line	Human	SOLiD	50	7,693,226	0.38
LN18	GBM cell line	Human	SOLiD	50	13,707,723	0.69
LN229	GBM cell line	Human	SOLiD	50	13,129,539	0.66
MJ	GBM cell line	Human	SOLiD	50	12,756,653	0.64
MK	GBM cell line	Human	SOLiD	50	13,145,984	0.66
T98	GBM cell line	Human	SOLiD	50	2,746,698	0.14
U87	GBM cell line	Human	SOLiD	50	30,998,618	1.55
S01	GBM tumor	Human	Illumina	75	22,388,016	1.68
S02	GBM tumor	Human	Illumina	75	21,456,618	1.61
S03	Normal brain	Human	Illumina	75	21,258,049	1.59
S04	GBM tumor	Human	Illumina	75	22,058,350	1.65
S06	GBM tumor	Human	Illumina	75	22,430,149	1.68
S07	GBM tumor	Human	Illumina	75	21,513,273	1.61
S08	GBM tumor	Human	Illumina	75	21,667,433	1.63
B01	GBM tumor	Mouse	SOLiD	50	47,603,332	2.38
B02	GBM tumor	Mouse	SOLiD	50	56,857,949	2.84
B03	GBM tumor	Mouse	SOLiD	50	60,794,006	3.04
B04	GBM tumor	Mouse	SOLiD	50	54,238,308	2.71
B05	GBM tumor	Mouse	SOLiD	50	58,347,341	2.92
B06	GBM tumor	Mouse	SOLiD	50	45,976,615	2.3
B07	GBM tumor	Mouse	SOLiD	50	50,157,968	2.51
B08	GBM tumor	Mouse	SOLiD	50	50,668,644	2.53
				TOTAL	686,728,288	38.16

We started with public data from the original RNA-seq reports, but now we are incorporating fresh data from our own lab as well, in particular from normal samples and glioblastoma cell-lines and tumors. For example, we integrated large-scale data from two different technologies, ABI SOLiD and Illumina. Both technologies require similar approaches in analysis, but the analysis and interpretation of RNA-seq is a current challenge for bioinformatics due the quantity of data (on the order of several GB per run), event detection (e.g. exon expression, exon-junctions detection), normalization between samples and technical error detection. Currently, we are in the evaluation and testing of some of the most recent tools for RNA-seq analysis and in particular with alternative isoform detection.

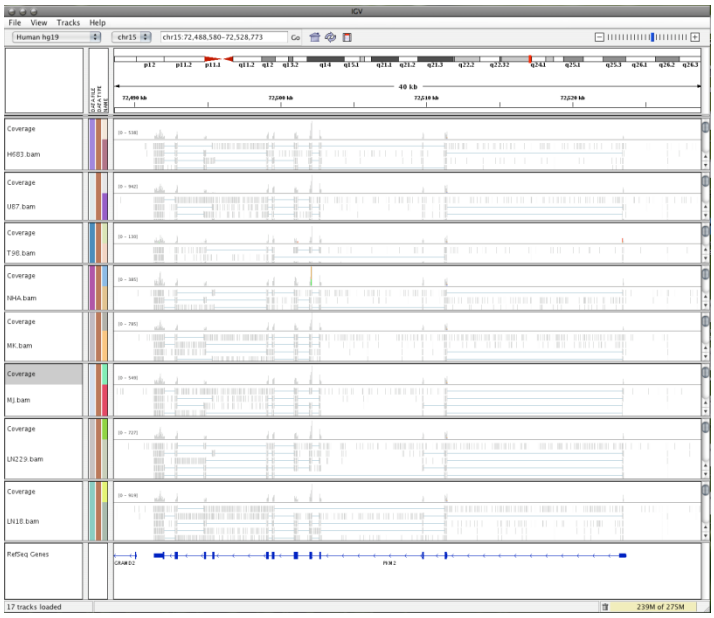


Figure 24: Expression levels for gene PKM2 in the SOLiD RNA-seq samples

2.14. Quantification of gene expression in human brain

We focused on the integration of other sources for gene expression quantification like RNA-seq technology (Figure 25 and Figure 26). We developed a general pipeline to analyze this type of data independently of the technology (Illumina or ABI/SOLiD). We also expanded our brain expression data sets with data from Illumina Inc., we'll have experimental data from normal brain, and other tissues in high coverage with single and paired reads.

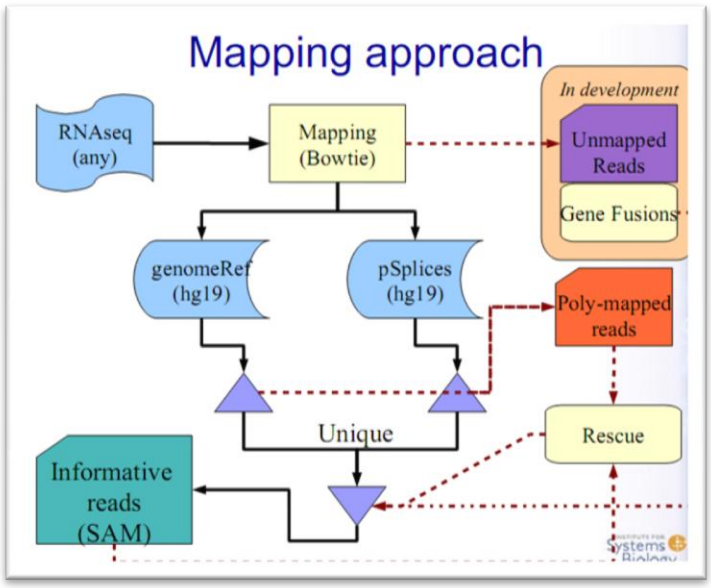


Figure 25: General view of the RNA-seq pipeline, mapping modules

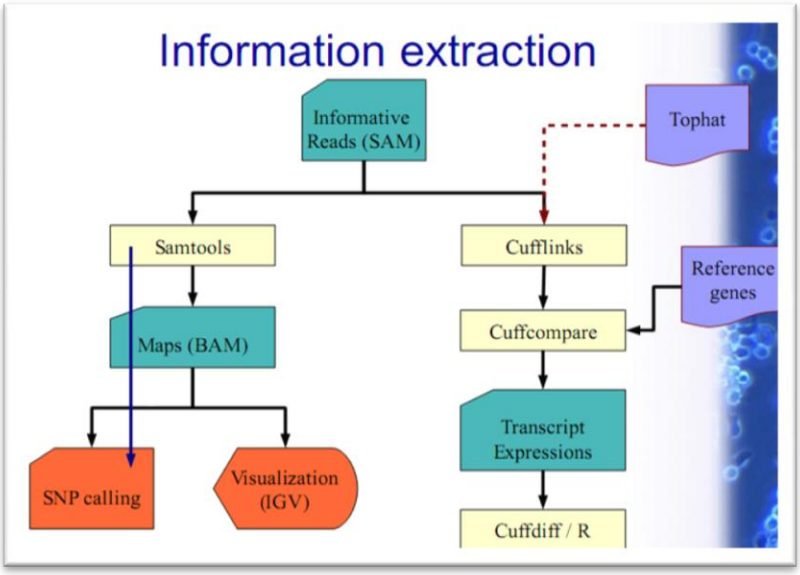


Figure 26: General view of the mining modules of the RNA-seq pipeline

3. Key Research Accomplishments

3.1. Biological investigation of EigenBrain image for each cell type

- **Oligodendrocytes**

Oligodendrocytes cells are responsible for the insulation of axons in the central nervous system. Figure 27 and Figure 28 shows the highly expressed brain regions in the *EigenBrains* computed from the Oligodendrocyte-specific markers for both coronal section and sagittal section.

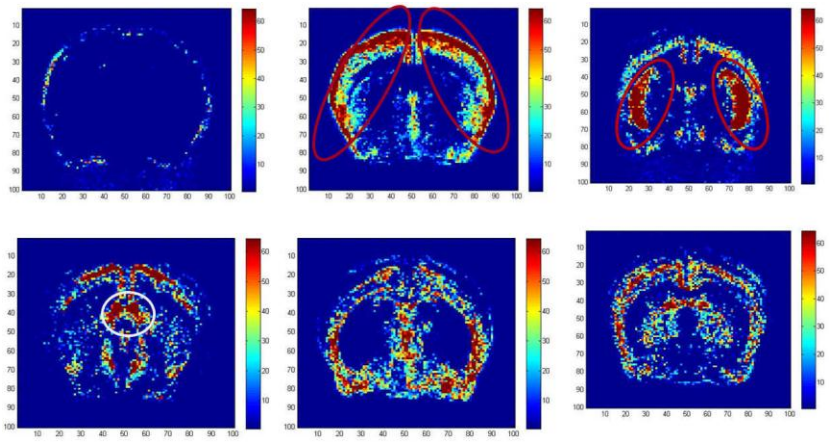


Figure 27: EigenBrain image for astrocytes enriched genes in coronal section

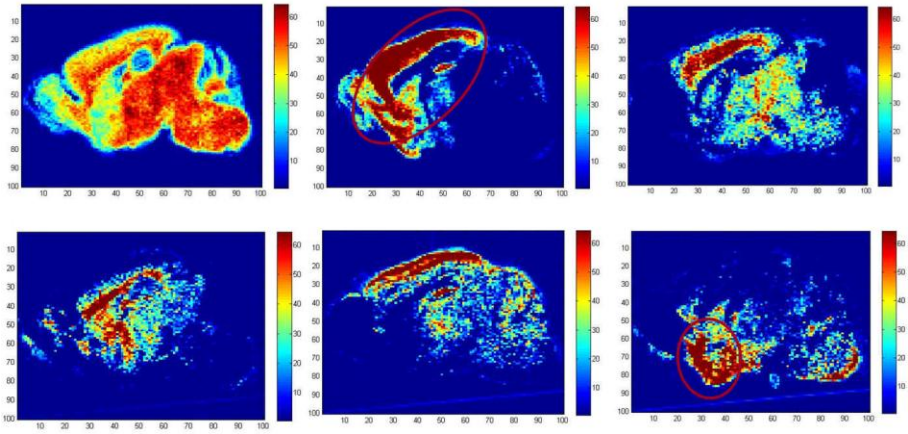


Figure 28: EigenBrain image for Astrocytes enriched genes in sagittal section

- Astrocytes

Astrocytes are known for providing the critical role of biochemical support to endothelial cells that form the blood-brain barrier. They also provide nutrients to the nervous tissue, and help in the repair and scarring process of the brain following traumatic injuries. The *EigenBrain* images in Figure 29 and Figure 30 also show the highly expressed brain regions whose roles are closely related with astrocytes. The circled regions in Figure 8 indicate the OCH (Optic Chiasm) region allowing for the right visual field to be process, the HY (Hypothalamus) region linking the

nervous system to the endocrine system via the pituitary gland, and the CP (Caudoputamen) region relating to cognition and working memory, respectively.

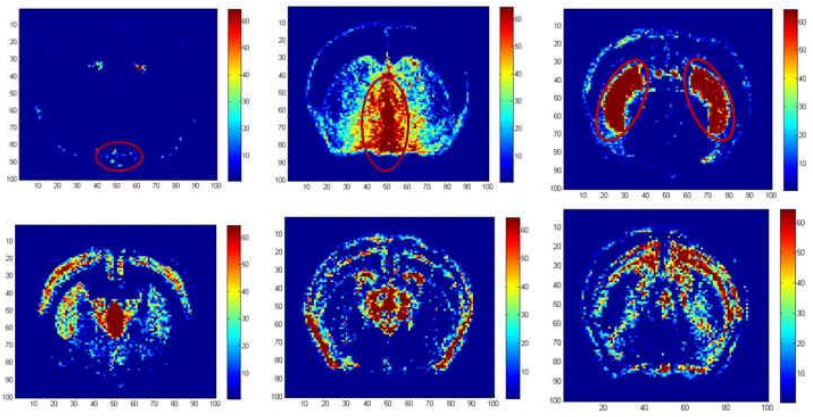


Figure 29: EigenBrain Image for Astrocytes enriched genes in coronal section

Figure 30 also reveals highly expressed regions in sagittal section of astrocyte-enriched genes : CTX (Cerebral Cortex) and MDRN (Medullary Retricular Nucleus). The CTX region is a sheet of neural tissue that is outermost to the cerebrum of the mammalian brain and plays a role in memory, attention, perceptual awareness, thought, language, and consciousness. The MDRN region is also responsible for controlling several major autonomic functions of the body.

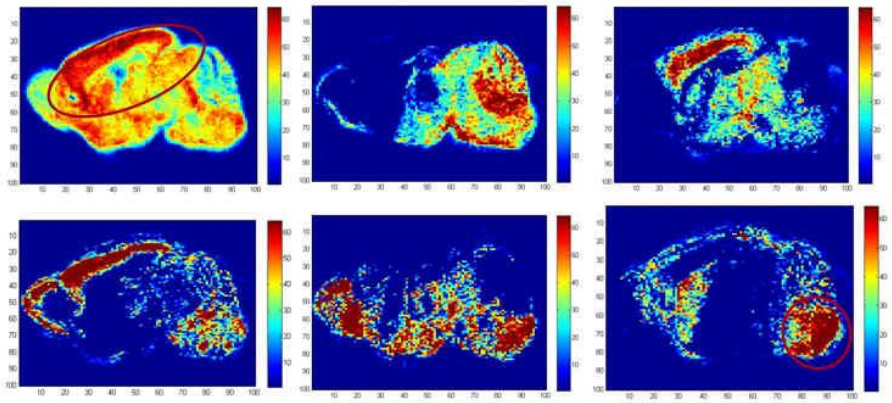


Figure 30: EigenBrain Image for Astrocytes enriched genes in sagittal section

- Neurons

Neuron cells are core components of the nervous system. Our *EigenBrain* approach displayed the biologically meaningful patterns in different brain regions. Figure 31 and Figure 32 show the highly expressed brain region for neuron cells in coronal and sagittal sections. In Figure 31, marked regions are revealed as representative regions such as VS (ventricular systems), CTX (cerebral cortex), HPF (Hippocampus Formation), PIR (piriform cortex), and TH (Thalamus). More specifically, the VS region is usually increased with age and enlarged in a number of neurological conditions. The HPF region has key role in long term memory and spatial navigation and in Alzheimer's disease, and this region is often one of the first regions of the brain to suffer damage. The PIR and TH regions are related with sensory system, story linking image

and smell, and responsible for the regulation of consciousness, sleep and alertness, respectively. These discovered regions are consistent with the ones from sagittal regions (Figure 32).

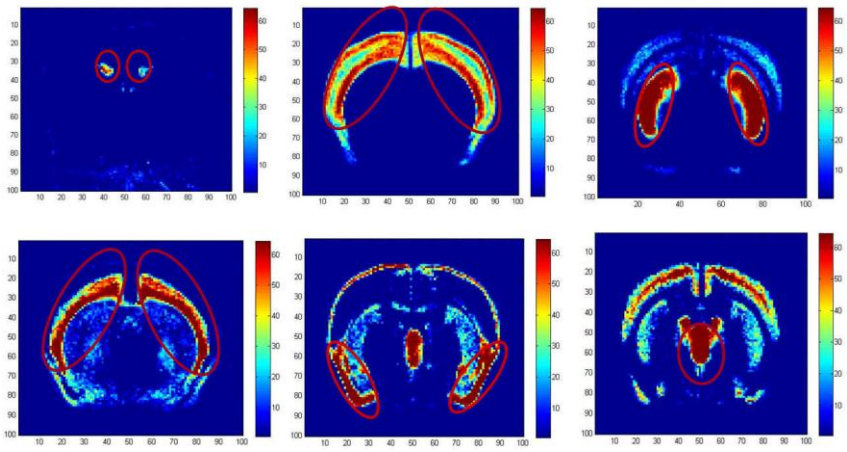


Figure 31: EigenBrain Image for Neuron enriched genes in coronal section

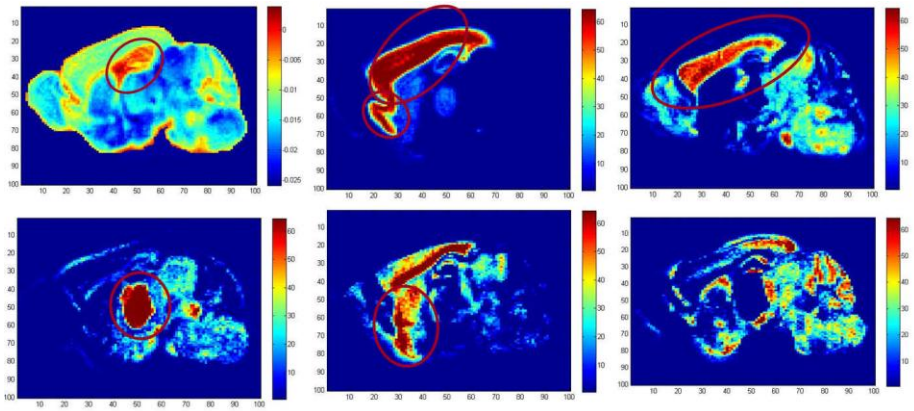


Figure 32: EigenBrain Image for Neuron enriched genes in sagittal section

3.2. More investigation about EigenBrain approach

3.2.1. Symmetry pattern in EigenBrain image

In order to check whether the symmetry expression patterns in *EigenBrain* images come from the bias of *EigenBrain* approach itself or not, we applied the following test: all original images were rotated into 45 degree and we applied *EigenBrain* approach to see whether the symmetry in gene expression comes from the bias of the method itself or not. As can be seen from Figure 33, even though we applied the rotations to the original images, the symmetry expression patterns in *EigenBrain* images can be detected, which means those symmetries were not the bias of method itself.

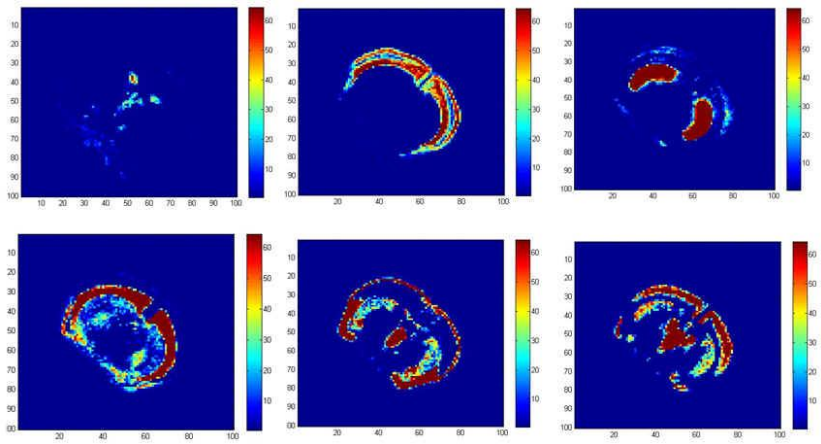


Figure 33: *EigenBrain* image from rotated brain images

3.2.2. Applying the EigenBrain approach to the test dataset

The *EigenBrain* approach has been applied to the large volume of test data (Table 11). As can be seen, the number of *in situ* hybridization (ISH) brain images are larger than the number of genes, which means that usually each gene has 1 or 2 corresponding brain images.

Table 11: Statistics of Test dataset

	# of image files	# of genes
Coronal section	7340	4034
Sagittal section	29110	19446

We applied our *EigenBrain* approach to the test dataset to identify the cell type specific genes. These genes are discovered from both coronal and sagittal sections as cell-type specific genes. We included a subset of the candidate cell-type specific genes in Table 12.

Table 12: Candidate cell type specific gene

	# of candidate cell type specific genes	Candidate cell type specific gene list
Oligodendrocytes	37	Adamts4, Anln, Arrdc3, BC030477, Car2, Cldn11, Edg2, Efnb3, Elovl1, Enpp6, Fa2h, Galnt6, Gatm, etc
Astrocytes	23	Acaa2, AI987712, C230095G01Rik, Capsl, Cldn1, Decr1, Dip3b, Dlx6os1, E030013G06Rik, etc
Neuron	363	1700010C24Rik, 1700020C11Rik, 2010004A03Rik,A830018L16Rik, A930041I02Rik, Aacs, Abhd6, Adcy1, Adcy9, Ap3s1, Arf3 etc

3.3. Identification of candidate cell type specific genes

3.3.1. Oligodendrocytes

From applying our algorithm to the analysis of the brain images, we identified EFNB3 as one of our candidate oligodendrotye enriched genes (Figure 34). As can be seen from Figure 34, these figures reveal the highly expressed patterns especially in *Alveus* and *Fimbria*, which are known as important regions of heavy oligodendorocyte expression in the cell [16]. In [3] also confirmed the consistent result though the microarray expression data. *EFNB3* is a member of the *ephrin* gene family and very important in brain development as well as its maintenance, particularly in the nervous system. The right panel in Figure 34 shows the related GO information with gene *EFNB3*. "Axon Guidance" process is detected as one of main GO processes enriched for the oligodendrocyte cell.

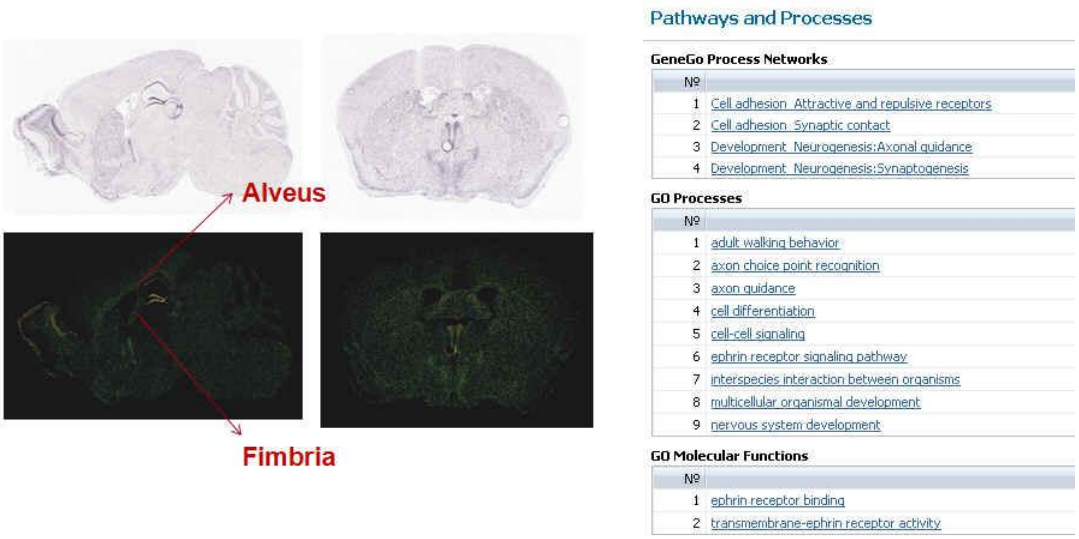


Figure 34: ISH brain image and corresponding expression image for oligodendorocytes specific gene : EFNB3 (Ephrin-B3) and its' related Gene Ontology pathway and process

Figure 35 shows the enrichment of GO categories for candidate ologodendrocyte specific genes. Axon escheatment, escheatment of neurons, myelination, and lipid biosynthetic process are identified as a critical pathways with P-value < 0.01. These GO categories also have been known as a major functionalities for oligodendrocyte cell though the literature [3].

GO_ID	TERM	NB_IN_REF	FREQ_IN_REF	NB_IN_SET	FREQ_IN_SET	P_VALUE	GENES_IN_SET
GO:0008366	axon ensheathment	33	0.0023	3	0.1034	3.60E-05	Mbp,Cldn11,Ugt8a
GO:0007272	ensheathment of neurons	33	0.0023	3	0.1034	3.60E-05	Mbp,Cldn11,Ugt8a
GO:0019228	regulation of action potential in neuron	37	0.0025	3	0.1034	5.09E-05	Mbp,Cldn11,Ugt8a
GO:0001508	regulation of action potential	45	0.0031	3	0.1034	9.17E-05	Mbp,Cldn11,Ugt8a
GO:0006633	fatty acid biosynthetic process	86	0.0059	3	0.1034	0.000615	Elovl1,Fa2h,Ptgds
GO:0016053	organic acid biosynthetic process	93	0.0063	3	0.1034	0.00077	Elovl1,Fa2h,Ptgds
GO:0046394	carboxylic acid biosynthetic process	93	0.0063	3	0.1034	0.00077	Elovl1,Fa2h,Ptgds
GO:0042391	regulation of membrane potential	105	0.0072	3	0.1034	0.001088	Mbp,Cldn11,Ugt8a
GO:0042552	myelination	31	0.0021	2	0.069	0.001666	Mbp,Ugt8a
GO:0008610	lipid biosynthetic process	276	0.0188	4	0.1379	0.001828	Elovl1,Fa2h,Ptgds,Ugt8a
GO:0015670	carbon dioxide transport	1	0.0001	1	0.0345	0.001978	Car2
GO:0007399	nervous system development	797	0.0544	6	0.2069	0.00336	Mbp,Sema6a,Cldn11,Sox10,Efnb3,Ugt8a
GO:0022410	circadian sleep/wake cycle process	2	0.0001	1	0.0345	0.003949	Ptgds
GO:0050802	circadian sleep/wake cycle, sleep	2	0.0001	1	0.0345	0.003949	Ptgds
GO:0006601	creatine biosynthetic process	2	0.0001	1	0.0345	0.003949	Gatm
GO:0042749	regulation of circadian sleep/wake cycle	2	0.0001	1	0.0345	0.003949	Ptgds
GO:0019695	choline metabolic process	2	0.0001	1	0.0345	0.003949	Enpp6
GO:0045187	regulation of circadian sleep/wake cycle, sleep	2	0.0001	1	0.0345	0.003949	Ptgds
GO:0006643	membrane lipid metabolic process	57	0.0039	2	0.069	0.005449	Fa2h,Ugt8a
GO:0006631	fatty acid metabolic process	188	0.0128	3	0.1034	0.005451	Elovl1,Fa2h,Ptgds
GO:0006600	creatine metabolic process	3	0.0002	1	0.0345	0.005912	Gatm
GO:0042745	circadian sleep/wake cycle	3	0.0002	1	0.0345	0.005912	Ptgds
GO:0042396	phosphagen biosynthetic process	3	0.0002	1	0.0345	0.005912	Gatm
GO:0006629	lipid metabolic process	664	0.0453	5	0.1724	0.007391	Elovl1,Enpp6,Fa2h,Ptgds,Ugt8a
GO:0051642	centrosome localization	4	0.0003	1	0.0345	0.007868	Sema6a
GO:0016198	axon choice point recognition	4	0.0003	1	0.0345	0.007868	Efnb3
GO:0006873	cellular ion homeostasis	218	0.0149	3	0.1034	0.008071	Mbp,Cldn11,Ugt8a
GO:0055082	cellular chemical homeostasis	225	0.0153	3	0.1034	0.008766	Mbp,Cldn11,Ugt8a
GO:0019752	carboxylic acid metabolic process	441	0.0301	4	0.1379	0.009002	Gatm,Elovl1,Fa2h,Ptgds
GO:0006082	organic acid metabolic process	442	0.0302	4	0.1379	0.009068	Gatm,Elovl1,Fa2h,Ptgds
GO:0019226	transmission of nerve impulse	233	0.0159	3	0.1034	0.009599	Mbp,Cldn11,Ugt8a
GO:0042752	regulation of circadian rhythm	5	0.0003	1	0.0345	0.009816	Ptgds
GO:0048512	circadian behavior	5	0.0003	1	0.0345	0.009816	Ptgds
GO:0030431	sleep	5	0.0003	1	0.0345	0.009816	Ptgds
GO:0048484	enteric nervous system development	5	0.0003	1	0.0345	0.009816	Sox10
GO:0006599	phosphagen metabolic process	5	0.0003	1	0.0345	0.009816	Gatm
GO:0042439	ethanolamine and derivative metabolic process	5	0.0003	1	0.0345	0.009816	Enpp6
GO:0032787	monocarboxylic acid metabolic process	248	0.0169	3	0.1034	0.011275	Elovl1,Fa2h,Ptgds
GO:0050801	ion homeostasis	248	0.0169	3	0.1034	0.011275	Mbp,Cldn11,Ugt8a
GO:0048013	ephrin receptor signaling pathway	6	0.0004	1	0.0345	0.011757	Efnb3
GO:0009247	glycolipid biosynthetic process	7	0.0005	1	0.0345	0.01369	Ugt8a
GO:0007622	rhythmic behavior	7	0.0005	1	0.0345	0.01369	Ptgds
GO:0007411	axon guidance	100	0.0068	2	0.069	0.015604	Sema6a,Efnb3
GO:0006575	cellular amino acid derivative metabolic process	102	0.007	2	0.069	0.016178	Gatm,Enpp6
GO:0019725	cellular homeostasis	293	0.02	3	0.1034	0.017172	Mbp,Cldn11,Ugt8a
GO:0050910	detection of mechanical stimulus involved in sens	9	0.0006	1	0.0345	0.017534	Slc12a2
GO:0051592	response to calcium ion	9	0.0006	1	0.0345	0.017534	S100a16

Figure 35: Partial result for gene enrichment test of GO category for candidate oligodendrocyte specific genes

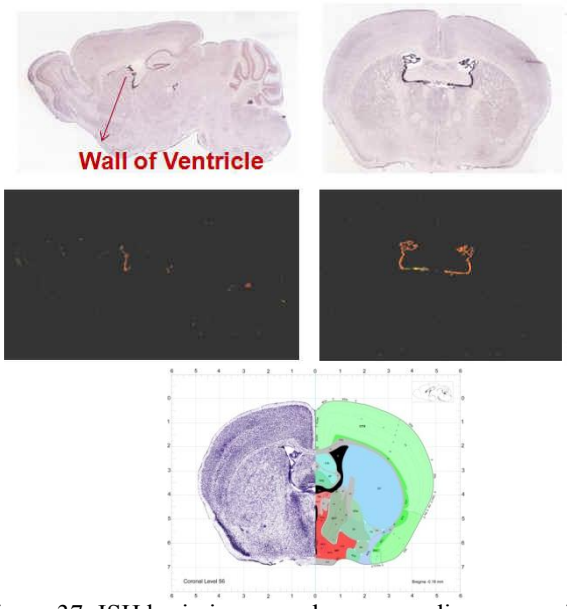
We also identified relevant pathways in which each of the candidate oligodendrocyte cells were involved. As we can see from Figure 36, axon guidance pathway was again detected as a highly relevant pathway with oligodendrocyte cell specific genes. Currently, we are still investigating other pathways or GO functions.

Rank in list	Symbol in list	Symbol in pathway	Pathways
5	Car2	CA2	NITROGEN_METABOLISM
6	Cldn11	CLDN11	HSA04670_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION
7	Edg2	EDG2	SMOOTH_MUSCLE_CONTRACTION
8	Efnb3	EFNB3	HSA04360_AXON_GUIDANCE
10	Enpp6	ENPP6	HSA00565_ETHER_LIPID_METABOLISM
12	Galnt6	GALNT6	O_GLYCAN_BIOSYNTHESIS
13	Gatm	GATM	UREA_CYCLE_AND_METABOLISM_OF_AMINO_GROUPS
14	Gpr37	GPR37	PARKINPATHWAY
17	Map2k6	MAP2K6	TOLLPATHWAY
19	Mbp	MBL2	HSA04610_COMPLEMENT_AND_COAGULATION_CASCADES
26	Plxnb3	PLXNB3	HSA04360_AXON_GUIDANCE
27	Ptgds	PTGDS	PROSTAGLANDIN_SYNTHESIS_REGULATION
29	Sema6a	SEMA6A	HSA04360_AXON_GUIDANCE
36	Unc5b	UNC5B	HSA04360_AXON_GUIDANCE
37	Vegfb	VEGFB	HSA05219_BLADDER_CANCER

Figure 36: Pathway analysis for candidate oligodendrocyte specific genes

3.3.2. Astrocytes

Figure 37 shows ISH brain image and corresponding expression image of the gene *Acaa2* (*acetype-Coenzyme A acyltransferase 2*). This gene is identified as one of the candidate astrocytes- specific genes by our *EigenBrain* approach. As we can see, the wall of the ventricle region is very clearly expressed in both sections – a clear indication of astrocyte specific genes. *Acaa2* encodes protein catalyzing the last step of the mitochondrial fatty acid beta-oxidation spiral and this function is also confirmed from Gene Ontology processes revealing fatty acid metabolic process as a major related process [3]. Additionally, we found that this gene is related with leucune, isoleucine and valine metabolism as might be expected in astrocyte cells.



Pathways and Processes

GeneGo Pathway Maps

№	
1	Bile Acid Biosynthesis
2	Bile Acid Biosynthesis / Rodent version
3	Butanoate metabolism
4	Leucune, isoleucine and valine metabolism.p.2
5	Leucune, isoleucine and valine metabolism/ Rodent version
6	Lysine metabolism
7	Lysine metabolism/ Rodent version
8	Mitochondrial ketone bodies biosynthesis and metabolism
9	Mitochondrial long chain fatty acid beta-oxidation
10	Mitochondrial unsaturated fatty acid beta-oxidation
11	Peroxisomal branched chain fatty acid oxidation
12	Phenylalanine metabolism
13	Phenylalanine metabolism/ Rodent version
14	Propionate metabolism p.2
15	Tryptophan metabolism
16	Tryptophan metabolism/ Rodent version
17	Tyrosine metabolism p.2 (melanin)

GO Processes

№	Name
1	cholesterol biosynthetic process
2	fatty acid metabolic process
3	lipid metabolic process
4	metabolic process

Figure 37: ISH brain image and corresponding expression ir e-Coenzyme A acyltransferase 2) and its' related Gene Ontology pathway and process

Figure 38 shows the highly enriched gene ontology terms among candidate astrocyte specific genes. With P-value < 0.01, there were 19 GO terms potentially related with astrocyte cells. We are planning to investigate more about these GO terms to discover the more biological insights in the coming quarter.

GO_ID	TERM	NB_IN_REF	FREQ_IN_REF	NB_IN_SET	FREQ_IN_SET	P_VALUE	GENES_IN_SET
GO:0050878	regulation of body fluid levels	88	0.006	3	0.1875	0.000204	Trp73,F5,F3
GO:0042060	wound healing	98	0.0067	3	0.1875	0.000277	Gja1,F5,F3
GO:0009611	response to wounding	341	0.0233	4	0.25	0.00099	Gja1,Trp73,F5,F3
GO:0033326	cerebrospinal fluid secretion	1	0.0001	1	0.0625	0.002315	Trp73
GO:0007596	blood coagulation	69	0.0047	2	0.125	0.003327	F5,F3
GO:0007599	hemostasis	70	0.0048	2	0.125	0.003419	F5,F3
GO:0050916	sensory perception of sweet taste	2	0.0001	1	0.0625	0.003471	Itpr3
GO:0043508	negative regulation of JUN kinase activity	2	0.0001	1	0.0625	0.003471	Trp73
GO:0050817	coagulation	72	0.0049	2	0.125	0.003607	F5,F3
GO:0050917	sensory perception of umami taste	3	0.0002	1	0.0625	0.004625	Itpr3
GO:0031638	zymogen activation	3	0.0002	1	0.0625	0.004625	Mmp14
GO:0048636	positive regulation of muscle development	3	0.0002	1	0.0625	0.004625	Gja1
GO:0045844	positive regulation of striated muscle development	3	0.0002	1	0.0625	0.004625	Gja1
GO:0009605	response to external stimulus	565	0.0385	4	0.25	0.005777	Gja1,Trp73,F5,F3
GO:0043403	skeletal muscle regeneration	6	0.0004	1	0.0625	0.008079	Gja1
GO:0016338	calcium-independent cell-cell adhesion	6	0.0004	1	0.0625	0.008079	Cldn1
GO:0045793	positive regulation of cell size	7	0.0005	1	0.0625	0.009227	Trp73
GO:0008015	blood circulation	119	0.0081	2	0.125	0.009316	Gja1,F5
GO:0003013	circulatory system process	119	0.0081	2	0.125	0.009316	Gja1,F5
GO:0002347	response to tumor cell	8	0.0005	1	0.0625	0.010374	Trp73
GO:0042246	tissue regeneration	9	0.0006	1	0.0625	0.01152	Gja1
GO:0002070	epithelial cell maturation	10	0.0007	1	0.0625	0.012664	Gja1
GO:0014706	striated muscle tissue development	145	0.0099	2	0.125	0.01347	Gja1,Tnc
GO:0060537	muscle tissue development	154	0.0105	2	0.125	0.015061	Gja1,Tnc
GO:0007512	adult heart development	13	0.0009	1	0.0625	0.016088	Gja1
GO:0060070	Wnt receptor signaling pathway through beta-catenin	13	0.0009	1	0.0625	0.016088	Rspo1
GO:0007589	body fluid secretion	14	0.001	1	0.0625	0.017227	Trp73
GO:0001836	release of cytochrome c from mitochondria	14	0.001	1	0.0625	0.017227	Trp73
GO:0035050	embryonic heart tube development	16	0.0011	1	0.0625	0.0195	Gja1
GO:0030574	collagen catabolic process	16	0.0011	1	0.0625	0.0195	Mmp14
GO:0044243	multicellular organismal catabolic process	17	0.0012	1	0.0625	0.020635	Mmp14
GO:0031099	regeneration	18	0.0012	1	0.0625	0.021768	Gja1
GO:0043506	regulation of JUN kinase activity	18	0.0012	1	0.0625	0.021768	Trp73
GO:0030308	negative regulation of cell growth	18	0.0012	1	0.0625	0.021768	Trp73
GO:0002064	epithelial cell development	19	0.0013	1	0.0625	0.022899	Gja1
GO:0043407	negative regulation of MAP kinase activity	19	0.0013	1	0.0625	0.022899	Trp73
GO:0007517	muscle organ development	194	0.0132	2	0.125	0.02302	Gja1,Tnc
GO:0050913	sensory perception of bitter taste	20	0.0014	1	0.0625	0.02403	Itpr3
GO:0032963	collagen metabolic process	21	0.0014	1	0.0625	0.025158	Mmp14
GO:0021766	hippocampus development	22	0.0015	1	0.0625	0.026286	Trp73
GO:0007528	neuromuscular junction development	22	0.0015	1	0.0625	0.026286	Tnc
GO:0008637	apoptotic mitochondrial changes	22	0.0015	1	0.0625	0.026286	Trp73
GO:0044259	multicellular organismal macromolecule metabolic process	22	0.0015	1	0.0625	0.026286	Mmp14
GO:0001947	heart looping	24	0.0016	1	0.0625	0.028537	Gja1
GO:0045792	negative regulation of cell size	24	0.0016	1	0.0625	0.028537	Trp73
GO:0035295	tube development	235	0.016	2	0.125	0.032556	Gja1,Mmp14
GO:0048634	regulation of muscle development	28	0.0019	1	0.0625	0.033022	Gja1

Figure 38: Partial result for gene enrichment test of GO category for candidate astrocyte specific genes

Figure 39 shows the critical pathways related with candidate astrocytes specific genes. Valine leucine and isoleucine degradation pathway is also known to play a critical role in astrocytes [3].

Rank in list	Symbol in list	Symbol in pathway	Pathways	
1	Acaa2	ACAA2	VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	
5	Cldn1	CLDN1	HSA05131_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EPEC	HSA04514_CELL_ADHESION_MOLECULES
11	F3	CNTN1	HSA04514_CELL_ADHESION_MOLECULES	
12	F5	F5	INTRINSICPATHWAY	
13	Gja1	GJA1	GSK3PATHWAY	SMOOTH_MUSCLE_CONTRACTION
14	Itpr3	ITPR3	SMOOTH_MUSCLE_CONTRACTION	HSA04540_GAP_JUNCTION
17	Mmp14	MMP14	HSA04912_GNRH_SIGNALING_PATHWAY	RECKPATHWAY
18	Prkg2	PRKG2	HSA04540_GAP_JUNCTION	HSA04730_LONG_TERM_DEPRESSION
22	Tnc	TNC	HSA01430_CELL_COMMUNICATION	HSA04512_ECM_RECEPTOR_INTERACTION

Figure 39: Pathway analysis for candidate astrocyte specific genes

3.3.3. Neurons

A candidate neuron-specific gene we have identified, is *Grin1* (glutamate receptor, ionotropic, NMDA1(zeta 1)), shown in Figure 18. In particular, we see high expression in a "G-shaped" region (Hippocampal subfields: CA1, CA2, CA3 and DG (dentate gyrus)) in the sagittal cross-section. *Grin 1* is a NMDA receptor subtype of glutamate-gated ion channels and possesses high calcium permeability and voltage-dependent sensitivity to magnesium. In particular, it encodes for a protein that plays a role in synaptic plasticity, synaptogenesis, excitotoxicity, memory

acquisition and learning. This function also can be confirmed from GO enrichment analysis (Figure 40).

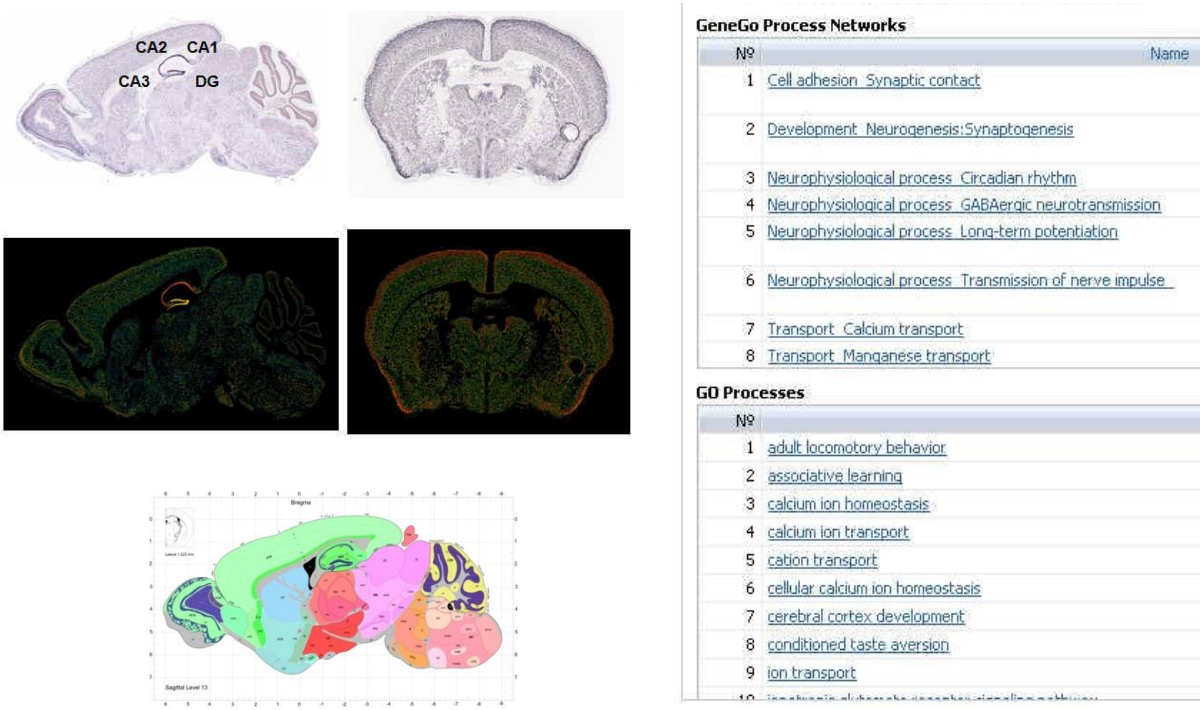


Figure 40: ISH brain image and corresponding expression image for neuron specific gene: Grin1 (glutamate receptor, ionotropic, NMDA1(zeta 1)) and its' related Gene Ontology pathway and process

From this data, we identified many significant GO categories that were enriched in neuron-specific genes such as synaptic transmission, regulation of synaptic plasticity, and neurotransmitter transport. Figure 41 lists a subset of the enriched GO terms and biological validation is needed for further analyses.

GO_ID	TERM	NB_IN_REF	FREQ_IN_REF	NB_IN_SET	FREQ_IN_SET	P_VALUE	GENES_IN_SET
GO:0051179	localization	2671	0.1822	103	0.3962	2.16E-16	Rbp4,Kcnk2,Kcnf1,Ica1,Slc2a3,Slit3,Vegfa,Ndufa10,Got2,Kif5c,Sh3gl2,Slc36a2,Osbpl8,G
GO:0006810	transport	2263	0.1544	91	0.35	2.42E-15	Rbp4,Kcnk2,Kcnf1,Ica1,Slc2a3,Ndufa10,Got2,Sh3gl2,Slc36a2,Osbpl8,Grik2,Kpna1,Rab3
GO:0051234	establishment of localization	2276	0.1553	91	0.35	3.44E-15	Rbp4,Kcnk2,Kcnf1,Ica1,Slc2a3,Ndufa10,Got2,Sh3gl2,Slc36a2,Osbpl8,Grik2,Kpna1,Rab3
GO:0007268	synaptic transmission	194	0.0132	20	0.0769	2.14E-10	Ica1,Grik2,Sv2b,Snap25,Pclo,Gabrg2,Egr3,Nrxn1,Grin1,Gipc1,Slc24a2,Park2,Slc17a7,Sy
GO:0032940	secretion by cell	217	0.0148	20	0.0769	1.53E-09	Rbp4,Ica1,Rab3c,Sv2b,Scrn1,Snap25,Pclo,Scg5,Nrxn1,Gipc1,Pfkl,Doc2a,Park2,Rims3,Vgf
GO:0006836	neurotransmitter transport	83	0.0057	13	0.05	2.04E-09	Ica1,Sv2b,Snap25,Pclo,Nrxn1,Slc6a7,Slc32a1,Park2,Slc17a7,Rims3,Syn2,Stx1a,Nrxn3
GO:0007267	cell-cell signaling	353	0.0241	25	0.0962	3.21E-09	Ica1,Grik2,Sv2b,Snap25,Pclo,Gabrg2,Egr3,Scg5,Nrxn1,Grin1,Gipc1,Pfkl,Slc24a2,Park2,Sl
GO:0019226	transmission of nerve impulse	233	0.0159	20	0.0769	5.17E-09	Ica1,Grik2,Sv2b,Snap25,Pclo,Gabrg2,Egr3,Nrxn1,Grin1,Gipc1,Slc24a2,Park2,Slc17a7,Sy
GO:0003001	generation of a signal involved in cell-cell signaling	111	0.0076	14	0.0538	8.69E-09	Ica1,Sv2b,Snap25,Pclo,Scg5,Nrxn1,Gipc1,Pfkl,Park2,Vgf,Pfkm,Syn2,Rapgef4,Nrxn3
GO:0006811	ion transport	682	0.0465	35	0.1346	1.03E-08	Kcnk2,Kcnf1,Slc36a2,Grik2,Kcnn2,Kcnmb4,Gabrb2,Kcnj4,Slc9a1,Slc12a3,Gabrg2,Kcnip3
GO:0046903	secretion	248	0.0169	20	0.0769	1.47E-08	Rbp4,Ica1,Rab3c,Sv2b,Scrn1,Snap25,Pclo,Scg5,Nrxn1,Gipc1,Pfkl,Doc2a,Park2,Rims3,Vgf
GO:0001505	regulation of neurotransmitter levels	67	0.0046	10	0.0385	2.41E-07	Ica1,Sv2b,Snap25,Pclo,Nrxn1,Park2,Slc17a7,Syn2,Gad2,Nrxn3
GO:0007269	neurotransmitter secretion	42	0.0029	8	0.0308	5.76E-07	Ica1,Sv2b,Snap25,Pclo,Nrxn1,Park2,Syn2,Nrxn3
GO:0015672	monovalent inorganic cation transport	297	0.0203	19	0.0731	1.15E-06	Kcnk2,Kcnf1,Slc36a2,Kcnn2,Kcnmb4,Kcnj4,Slc9a1,Slc12a3,Kcnip3,Kcns2,Kctd1,Slc17a7
GO:0030001	metal ion transport	431	0.0294	23	0.0885	1.92E-06	Kcnk2,Kcnf1,Kcnn2,Kcnmb4,Kcnj4,Slc9a1,Slc12a3,Kcnip3,Kcns2,Grin1,Slc24a2,Kctd1,Sl
GO:0007214	gamma-aminobutyric acid signaling pathway	15	0.001	5	0.0192	4.26E-06	Gabrg2,Gabra5,Gabra1,Gabrb3,Gabra3
GO:0006812	cation transport	499	0.034	24	0.0923	6.51E-06	Kcnk2,Kcnf1,Slc36a2,Kcnn2,Kcnmb4,Kcnj4,Slc9a1,Slc12a3,Kcnip3,Kcns2,Grin1,Slc24a2
GO:0006813	potassium ion transport	151	0.0103	12	0.0462	1.30E-05	Kcnk2,Kcnf1,Kcnn2,Kcnmb4,Kcnj4,Kcnip3,Kcns2,Kctd1,Kcnip4,Kcnq3,Hcn1,Kctd16
GO:0016311	dephosphorylation	127	0.0087	11	0.0423	1.34E-05	Ptprj,Dusp1,Ppm2c,Dusp6,Ppm1l,Ptpn5,Mtmr12,Ptprk,Ptprs,Mtmr7,Ppp2ca
GO:0051046	regulation of secretion	95	0.0065	9	0.0346	4.01E-05	Rbp4,Ica1,Rab3c,Pclo,Scg5,Pfkl,Park2,Pfkm,Rapgef4
GO:0006470	protein amino acid dephosphorylation	103	0.007	9	0.0346	7.45E-05	Ptprj,Dusp1,Ppm2c,Dusp6,Ppm1l,Ptpn5,Ptprk,Ptprs,Ppp2ca
GO:0050804	regulation of synaptic transmission	68	0.0046	7	0.0269	0.0001704	Ica1,Grik2,Grin1,Gipc1,Slc24a2,Park2,Cpeb1
GO:0051649	establishment of localization in cell	626	0.0427	24	0.0923	0.0001976	Rbp4,Ica1,Grik2,Kpna1,Rab3c,Ipo4,Sv2b,Scrn1,Snap25,Pclo,Scg5,Nrxn1,Gipc1,Pfkl,Doc2
GO:0051641	cellular localization	666	0.0454	25	0.0962	0.0002005	Rbp4,Ica1,Grik2,Kpna1,Rab3c,Ipo4,Sv2b,Scrn1,Snap25,Pclo,Plxna2,Scg5,Nrxn1,Gipc1,Pf
GO:0006887	exocytosis	94	0.0064	8	0.0308	0.0002194	Rab3c,Sv2b,Scrn1,Pclo,Doc2a,Rims3,Rapgef4,Stx1a
GO:0051969	regulation of transmission of nerve impulse	71	0.0048	7	0.0269	0.0002218	Ica1,Grik2,Grin1,Gipc1,Slc24a2,Park2,Cpeb1
GO:0031644	regulation of neurological system process	75	0.0051	7	0.0269	0.0003087	Ica1,Grik2,Grin1,Gipc1,Slc24a2,Park2,Cpeb1
GO:0048167	regulation of synaptic plasticity	35	0.0024	5	0.0192	0.0003238	Grik2,Grin1,Gipc1,Slc24a2,Cpeb1
GO:0046879	hormone secretion	60	0.0041	6	0.0231	0.0005716	Pclo,Scg5,Pfkl,Vgf,Pfkm,Rapgef4
GO:0016486	peptide hormone processing	10	0.0007	3	0.0115	0.0005849	Pcsk5,Scg5,Pcsk2
GO:0031175	neuron projection development	230	0.0157	12	0.0462	0.0006042	Slit3,Mtap2,Kif5c,Grin1,Nefl,Slitrk1,Nrn1,Stmn1,Fezf2,Pak1,Cck,Ntng2
GO:0009914	hormone transport	61	0.0042	6	0.0231	0.000623	Pclo,Scg5,Pfkl,Vgf,Pfkm,Rapgef4
GO:0051049	regulation of transport	201	0.0137	11	0.0423	0.000684	Rbp4,Ica1,Rab3c,Pclo,Scg5,Pfkl,Park2,Nedd4l,Pacsin1,Pfkm,Rapgef4
GO:0030073	insulin secretion	43	0.0029	5	0.0192	0.0008341	Pclo,Pfkl,Vgf,Pfkm,Rapgef4
GO:0016192	vesicle-mediated transport	419	0.0286	17	0.0654	0.0008385	Sh3gl2,Rab3c,Sv2b,Elmo1,Scrn1,Pclo,Sorl1,Arf3,Doc2a,Rims3,Icam5,Gata2,Pacsin1,Rin
GO:0065008	regulation of biological quality	1045	0.0713	32	0.1231	0.0008575	Rbp4,Kcnk2,Ica1,Vegfa,Grik2,Pcsk5,Sv2b,Gucy1a3,Snap25,Slc9a1,Pclo,Scg5,Nrxn1,Tmsb
GO:0031111	negative regulation of microtubule polymerization o	12	0.0008	3	0.0115	0.001035	Mtap2,Mapt,Stmn1
GO:0010817	regulation of hormone levels	152	0.0104	9	0.0346	0.0011845	Rbp4,Pcsk5,Pclo,Scg5,Pfkl,Vgf,Pfkm,Rapgef4,Pcsk2
GO:0006796	phosphate metabolic process	938	0.064	29	0.1115	0.001242	Ptprj,Dusp1,Ppm2c,Mtap2,Dusp6,Uqcrh,Erbb4,Ppm1l,Ptpn5,Mtmr12,Nlk,Ptprk,Pak7,Rhc
GO:0006793	phosphorus metabolic process	938	0.064	29	0.1115	0.001242	Ptprj,Dusp1,Ppm2c,Mtap2,Dusp6,Uqcrh,Erbb4,Ppm1l,Ptpn5,Mtmr12,Nlk,Ptprk,Pak7,Rhc
GO:0044057	regulation of system process	154	0.0105	9	0.0346	0.0012908	Ica1,Grik2,Gucy1a3,Grin1,Gipc1,Slc24a2,Park2,Thrb,Cpeb1
GO:0007019	microtubule depolymerization	13	0.0009	3	0.0115	0.0013218	Mtap2,Mapt,Stmn1
GO:0007409	axonogenesis	187	0.0128	10	0.0385	0.0013645	Slit3,Kif5c,Grin1,Nefl,Slitrk1,Nrn1,Stmn1,Fezf2,Cck,Ntng2
GO:0030072	peptide hormone secretion	49	0.0033	5	0.0192	0.0014867	Pclo,Pfkl,Vgf,Pfkm,Rapgef4
GO:0018107	peptidyl-threonine phosphorylation	14	0.001	3	0.0115	0.0016528	Mtap2,Nlk,Mapk8
GO:0018210	peptidyl-threonine modification	14	0.001	3	0.0115	0.0016528	Mtap2,Nlk,Mapk8
GO:0006091	generation of precursor metabolites and energy	261	0.0178	12	0.0462	0.0016662	Ndufa10,Dlst,Uqcrh,Ndufs2,Pfkl,Ndufv2,Vgf,Uqcr,Pfkm,Txn2,Pfkp,Sdhb

Figure 41: Partial result for gene enrichment test of GO category for candidate neuron specific genes

Figure 42 describes the pathways where neuron enriched genes are belonging. Red marked pathways such as calcium signaling pathway, mapK signaling pathway, Gaba pathway, and long-term depression pathways etc., are confirmed pathways through the literatures.

Rank	Symbol in l	Symbol in g	Pathways				
20	Hpca	CACNA1A	HSA04742_TASTE_TRANSDUCTION	HSA04020_CALCIUM_SIGNALING_PATHWAY	HSA04730_LONG_TERM_DEPRESSION	HSA04010_MAPK_SIGNA	HSA04930_TYPE_II_DIABE
26	Camkk2	CAMKK2	CACAMPATHWAY	HSA04920_ADIPOCYTOKINE_SIGNALING_PATHWAY			
28	Cckbr	CCKBR	HSA04020_CALCIUM_SIGNALING_PATHW	PEPTIDE_GPCRS	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION		
53	Dusp1	DUSP1	HSA04010_MAPK_SIGNALING_PATHWAY	PPARAPATHWAY	TNFR2PATHWAY	ST_P38_MAPK_PATHWA	NTHIPATHWAY
57	Dusp6	DUSP6	HSA04010_MAPK_SIGNALING_PATHWAY	ST_ERK1_ERK2_MAPK_PATHWAY			
70	Erbb4	ERBB4	HSA04020_CALCIUM_SIGNALING_PATHW	HSA04320_DORSO_VENTRAL_AXIS_FORMATION	ERBB4PATHWAY	HSA04012_ERBB_SIGNA	SIG_PIP3_SIGNALING_IN_C
79	Gabra1	GABRA1	HSA04080_NEUROACTIVE_LIGAND_RECE	GABAPATHWAY			
80	Gabrb2	GABRB2	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION				
81	Gabrb3	GABRB3	HSA04080_NEUROACTIVE_LIGAND_RECE				
82	Gabrg2	GABRG2	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION				
98	Gria2	GRIA2	HSA04730_LONG_TERM_DEPRESSION	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTE	FOSBPATHWAY	HSA04720_LONG_TERM_	POTENTIATION
99	Gria3	GRIA3	HSA04730_LONG_TERM_DEPRESSION	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION			
100	Grik2	GRIK5	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION				
102	Grin1	GRIN1	HSA04020_CALCIUM_SIGNALING_PATHW	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTE	HSA04720_LONG_TERM_POTENTIATIO	NOS1PATHWAY	
104	Gucy1a3	GUCY1A3	HSA04730_LONG_TERM_DEPRESSION	PURINE_METABOLISM	SMOOTH_MUSCLE_CONTRACTION	HSA04540_GAP_JUNCTION	
145	Map3k5	MAP3K5	HSA04010_MAPK_SIGNALING_PATHWAY	ST_P38_MAPK_PATHWAY	P38MAPKPATHWAY	MAPKPATHWAY	ST_JNK_MAPK_PATHWAY
147	Mapk11	MAPK11	HSA04010_MAPK_SIGNALING_PATHWAY	ST_P38_MAPK_PATHWAY	NTHIPATHWAY	SIG_CD40PATHWAYMAI	MAPKPATHWAY
148	Mapk8	MAPK8	HSA04010_MAPK_SIGNALING_PATHWAY	HSA04930_TYPE_II_DIABETES_MELLITUS	HSA04920_ADIPOCYTOKINE_SIGNALIN	SIG_CD40PATHWAYMAI	ST_DIFFERENTIATION_PAT
149	Mapt	MAPT	HSA04010_MAPK_SIGNALING_PATHWAY	BIOPEPTIDESPATHWAY	HSA05010_ALZHEIMERS_DISEASE	HSA01510_NEURODEGE	P35ALZHEIMERSPATHWAY
151	Ncam2	NCAM2	HSA04514_CELL_ADHESION_MOLECULES				
162	Ndufa10	NDUFA10	UBIQUINONE_BIOSYNTHESIS	HSA00190_OXIDATIVE_PHOSPHORYLATION	OXIDATIVE_PHOSPHORYLATION		
169	Nefl	NEFL	HSA05030_AMYOTROPHIC_LATERAL_SCLEROSIS				
208	Plcb4	PLCB4	HSA04020_CALCIUM_SIGNALING_PATHW	HSA04730_LONG_TERM_DEPRESSION	HSA04720_LONG_TERM_POTENTIATIO	HSA04540_GAP_JUNCT	HSA04912_GNRH_SIGNAL
224	Ppp2ca	PPP2CA	HSA04730_LONG_TERM_DEPRESSION	HSA04530_TIGHT_JUNCTION	KERATINOCYTEPATHWAY	HSA04310_WNT_SIGNA	P35ALZHEIMERSPATHWAY
225	Ppp2r2c	PPP2R2C	HSA04730_LONG_TERM_DEPRESSION	HSA04530_TIGHT_JUNCTION	HSA04310_WNT_SIGNALING_PATHWA	HSA04350_TGF_BETA_SIGNALING_PATHWAY	
226	Prdx5	PRDX6	HSA00650_BUTANOATE_METABOLISM	PHENYLALANINE_METABOLISM	STILBENE_COUMARINE_AND_LIGNIN_BIOSYNTHESIS		
228	Pak1	PRKCL1	AKAPCENTROSOMEPATHWAY	MYOSINPATHWAY			
234	Hrmt1l1	PRMT2	HSA00450_SELENOAMINO_ACID_METAB	HSA00350_TYROSINE_METABOLISM	HSA00626_NAPHTHALENE_AND_ANTH	RACENE_DEGRADATION	
236	Psmc2	PSMC2	HSA03050_PROTEASOME				
238	Ptdss1	PTDSS1	HSA00564_GLYCEROPHOSPHOLIPID_METABOLISM				
239	Ptpn5	PTPN5	HSA04010_MAPK_SIGNALING_PATHWAY				
299	Slc9a1	SLC9A1	HSA04810_REGULATION_OF_ACTIN_CY	G_PROTEIN_SIGNALING			
301	Slit3	SLIT3	HSA04360_AXON_GUIDANCE				
317	Stx1a	STX1A	HSA05020_PARKINSONS_DISEASE	BOTULINPATHWAY			

Figure 42: Pathway analysis for candidate neuron specific genes

3.4. Differentially expressed region for each cell type

Figure 43 shows results of unpaired t-test between different cell type specific genes to indentify the highly (lowly) expressed region in brain based on the gene expression intensity level. As seen, oligodendrocyte specific genes seem highly expressed around cerebral cortex region comparing to astrocytes or neuron cell type specific genes. Likewise, neuron-specific genes tend to be more highly expressed in the thalamus (TH) and hippocampal formation (HPF) than other cell type specific genes. However, there can be one point we need to catch up. Since we are focusing on the gene expression intensity in this experiment, we could not conclude that such a high expression is a necessary consequence of cell distribution. Thus, we applied same test at the density level of original spatial gene expression images (Figure 44). Density level test helps to remove the possibility that such a high expression patterns comes from relative expression difference in a particular region. For this test, each expression image is divided into 100 by 100 patches. For an each patch, density value is calculated representing how many pixels are expressed. Surprisingly, Figure 43 and Figure 44 demonstrate almost same results. T-test at the density level reveals that those highly expressed (dense) region is enriched for those specific cell distribution. For example, cerebral cortex (CTX) is highly enriched for astrocytes specific gene expressions, which suggest that astrocytes cells might be highly distributed in this region. In the same analogy, it is implicated that neuron cells are densely distributed in the thalamus (TH) and hippocampal formation (HPF).

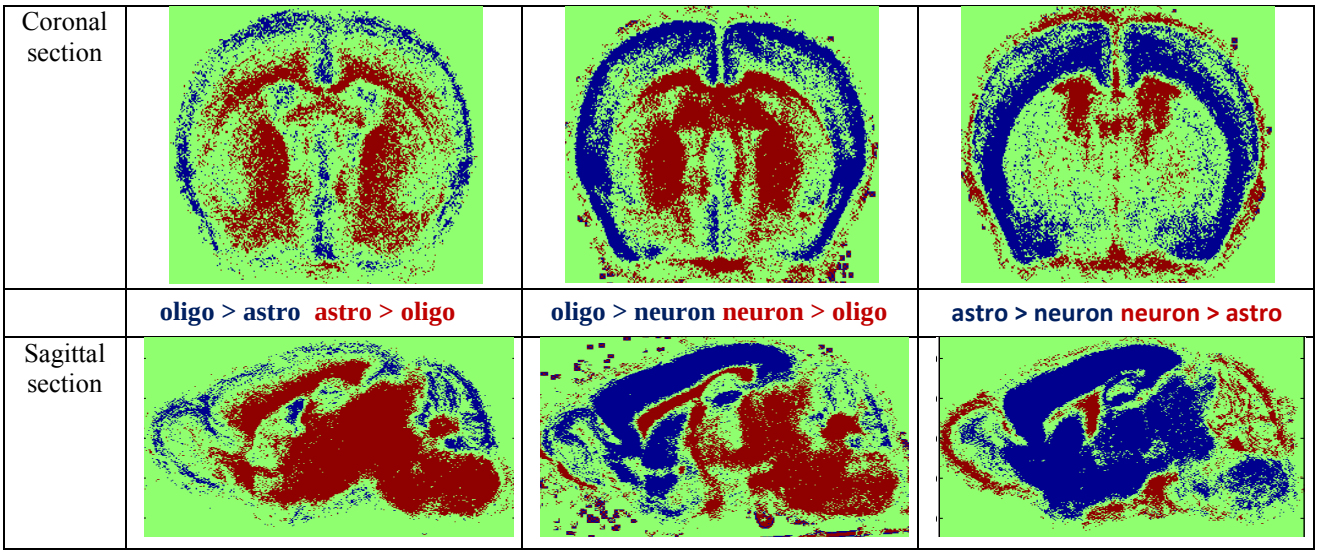


Figure 43 : Unpaired t-test between different cell type specific gene images at the intensity level. Results are thresholded at $p < 0.05$.

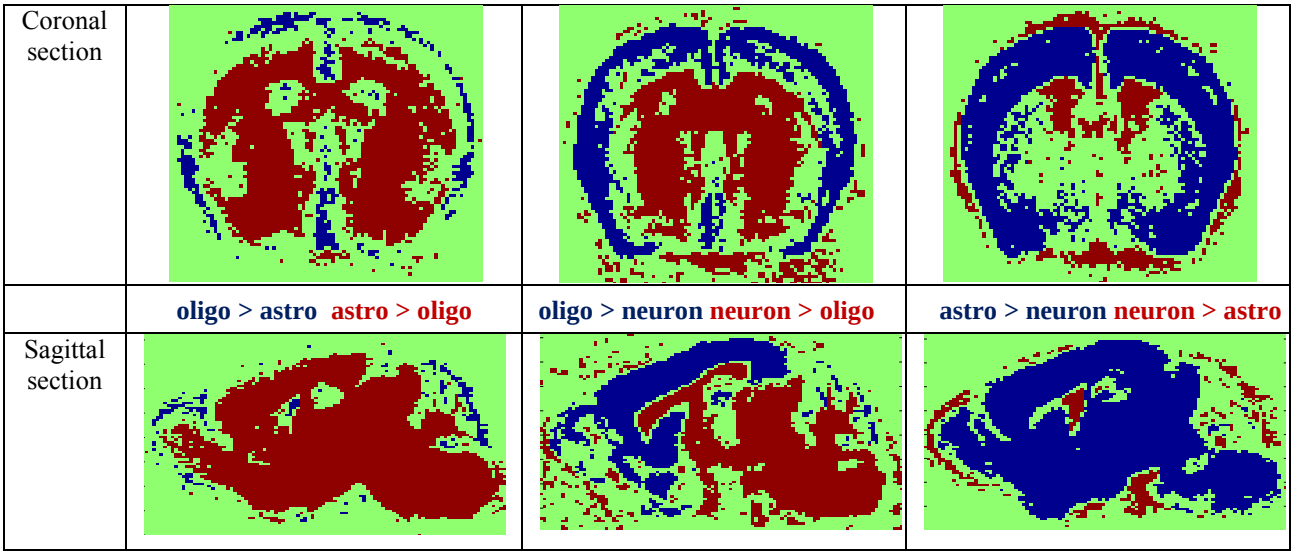


Figure 44 : Unpaired t-test between different cell type specific gene images at the density level. Results are thresholded at $p < 0.05$.

4. Reportable Outcomes

Papers:

Ko, Y., Cabellero, J., Glusman, G., Hood, L., and Price, N.D., Spatial expression patterning in cell-type specific genes, In preparation (2010)

Invited talks this year:

NP: Seminar, Department of Bioengineering, University of Illinois, Urbana-Champaign, IL, September 2, 2010

NP: Invited talk, 8th International Aegean Conference on Pathways, Networks, and Systems Medicine, Rhodes, Greece, July 12, 2010

NP: Panelist Speaker, Personalized Medicine Symposium, Research Triangle Park, Durham, NC, June 15, 2010

NP: Translation Biomedical Research Seminar, University of Illinois, Systems approaches to disease diagnosis and prognosis, April 5, 2010

NP: Seminar, Genome Institute of Singapore, Jan. 21, 2010, Systems approaches to disease stratification, Jan 21, 2010

NP: Seminar, Department of Genetics, Case Western Medical School, Systems medicine approaches to disease diagnosis and prognosis, Dec. 9, 2009

JC: "Computer prediction of blood biomarkers for neurological diseases" presented in The Allen Institute for Brain Science Data Integration Workshop, March 16th-17th, 2010.

JC: "Computer prediction of blood biomarkers for neurological diseases" presented in the Amgen Mini-Symposium, March 26th, 2010.

5. Conclusion

During this year, we applied our *EigenBrain* approach to identify candidate cell-type specific genes in the set of 20,000 mouse genes represented in the Allen Brain Atlas dataset. The *EigenBrain* approach identified specific regions that are highly expressed in each cell type and will provide a basis for further biological insight relating cell-type-specific expression with different brain regions. We investigated these highly expressed patterns in brain regions for each cell type – and further biological insights and these results are reported. In addition, as a result, we discovered a strong candidate set of brain-specific and cell-type specific transcripts. Moreover, we applied the region-based clustering method into the *in situ* hybridization of cell type specific genes. Region-based clustering method reveals dramatic spatial consistency of neuron-specific genes, sufficient to recapitulate most anatomical brain regions from gene expression alone. Furthermore, we also applied unpaired t-test between different cell-type specific genes at the different levels: intensity and density. This result helped to characterize the highly expressed brain region for specific cells and also understand cell distribution in conjunction with a density feature. We also performed the analysis of brain RNAseq data to measure transcripts present in both human and mouse brains. We have already detected the presence of many brain-specific genes (798 transcripts) that improve our candidate selection.

References

1. Jagalur M, Pal C, Learned-Miller E, Zoeller RT, and Kulp D (2007) Analyzing *in situ* gene expression in the mouse brain with image registration, feature extraction and block clustering. BMC Bioinformatics 8 Suppl **10**:S5.
2. Turk M and Pentland A (1991) Eigenfaces for Recognition. J Cognitive Neuroscience 3:71-86.
3. Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A.*, et al.* (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci ***28***, 264-278.
4. Fan RE, Chen PH and Lin CJ (2005) Working set selection using second order information for training SVM. J of Machine Learning Res 6, 1889-1918.
5. Cameron, H.A., and McKay, R.D. (2001). Adult neurogenesis produces a large pool of new granule cells in the dentate gyrus. J Comp Neurol ***435***, 406-417.
6. Sun, T., and Walsh, C.A. (2006). Molecular approaches to brain asymmetry and handedness. Nat Rev Neurosci ***7***, 655-662.
7. Sun, T., Patoine, C., Abu-Khalil, A., Visvader, J., Sum, E., Cherry, T.J., Orkin, S.H., Geschwind, D.H., and Walsh, C.A. (2005). Early asymmetry of gene transcription in embryonic human left and right cerebral cortex. Science ***308***, 1794-1798.
8. Su AI, Wiltshire T, Batalov S, Lapp H et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101(16):6062-6067.
9. Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y et al. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 2005 **86**(2):127-141.
10. Thorrez L, Van Deun K, Tranchevent LC, Van Lommel L et al. (2008) Using ribosomal protein genes as reference: a tale of caution. PLoS One 3(3):e1854.
11. Lattin JE, Schroder K, Su AI, Walker JR et al. (2008) Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res. **4**(1):5.
12. Bendtsen, J. D., H. Nielsen, *et al*. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.,* **340**(4): 783-95.
13. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. **40**, 1413–1415.
14. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.
15. Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**:R25.
16. Lein, E. S., M. J. Hawrylycz, *et al*. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature,* **445**(7124): 168-76.