Identification cell type marker genes of the brain and their use in estimation of cell type proportions

Thesis Proposal for Doctor of Philosophy(PhD) Degree

UBC bioinformatics Graduate Program

Ogan Mancarci, B.Sc

Thesis Supervisor

Dr. Paul Pavlidis

Committee Members

Dr. Clare Beasley

Dr. Shernaz Bamji

Dr. Sara Mostafavi

Chair

Dr. Ryan Brinkman

Examination Date

June 19, 2015

Contents

Contents . . . . . **ii**

List of Figures . . . . . **iii**

List of Tables . . . . . **vi**

1 Motivation and Introduction **1**

2 Research questions and specific aims **2**

2.1 Research questions 2

2.1.1 What are the specific marker genes of brain cell types? 2

2.1.2 Are mouse marker genes applicable to humans? 2

2.1.3 How accurately can cell type proportions be predicted with the use of marker genes? . 3

2.1.4 How do cell type proportions change accross neurological diseases? 3

2.1.5 Can cell type specific regulatory events be detected using cell type proportion information? 3

2.1.6 How do recent single cell experiments correlate with each other and cell type specific microarray samples in the literature? 3

2.2 Specific aims 3

2.2.1 Aim 1: Compilation of cell type specific expression database and make it available to third parties 3

2.2.2 Aim 2: Identification and verification of marker gene sets 4

2.2.3 Aim 3: Estimation of cell type proportions 4

3 Background **4**

3.1 Cell type specific effects in neurological conditions 4

3.2 Cell type markers and their applications 5

3.3 Methods for cell type purification 5

3.4 Methods for estimation of cell type proportions from expression patterns 6

3.4.1 Reference-based deconvolution 7

3.4.2 Reference-free deconvolution 7

3.5 Expression profiling 8

3.5.1 Microarrays 8

3.5.2 RNA sequencing 8

4 Aim 1: Compilation of cell type specific expression database available to third parties 9

4.1 Data acquisition and preprocessing . . . 9

4.2 Presentation of the data in a web application . . . 10

5 Aim 2: Identification and Validation of Marker Gene Sets . . . 10

5.1 Separation of samples into brain regions . . . 10

5.2 Selection of marker genes . . . 11

5.3 Validation of marker genes . . . 12

5.3.1 Validation of marker genes via in situ hybridization . . . 12

5.3.2 Computational validation of marker genes in mouse and human single cell data . . . 12

5.3.3 Validation of marker genes in human whole tissue data . . . 13

5.4 Assess condordance of single cell RNA-seq studies with each other and to microarray samples in our database . . . 14

6 Aim 3: Estimation of cell type proportions . . . 15

6.1 Estimation of cell type proportions in whole tissue samples using the marker gene sets . . . 15

6.2 Validation of the pipeline using isolated blood cell types and whole blood data . . . 17

6.3 Use proportion estimations to improve accuracy of differential expression analysis . . . 17

6.4 Create an R package for easy application of the method by third parties. . . . 18

References . . . 18

7 Tables . . . 25

8 Figures . . . 32

List of Figures

1 Workflow of the project 32

2 Example representations of cell type isolation techniques. A. Adapted from Kang et al. 2011. An example application of FACS. A fluorescently active molecule is used to label specific cell in the population. Upon detection of fluorescence, a charge is placed on the droplet whose path is later manipulated by an electrical field to separate the cells. B. Adapted from Cahoy et al. 2008. An example application of TRAP. A cell suspension is placed into a plate with bounded antibodies binding to a specific cell type. Removing suspended cells removes the cell type from the population. C. Adapted from Sanz et al. 2009. A schematic representation of the TRAP method. A cell type specific promoter driven expression of a labelled ribosome component causes certain cells to contain labelled ribosomes. The tissue is homogenized as a whole and fixed. Labelled ribosomes that carry RNAs from specific cells are isolated. Removal of ribosomes leaves cell type specific RNA samples behind. D. Adapted from Liotta et al. 2000. A schematic representation of LCM method. Cells are visually identified on the slide and marked. A laser then cuts the marked part and separates it from the rest of the samples. 33

3 A screenshot of the NeuroExpresso web application: A tool to visualize gene expression in the cell types of our database. 34

4 Hierarchy of brain regions used to separate the cell types into groups representing different regions of the brain. Cell types isolated from the regions in the lower nodes of the hierarchy are added to the higher nodes connected to them, while cell types isolated from the regions in the higher nodes are inlcuded in the lower nodes. 34

5 Expression of top 5 marker genes detected from cortex cell types. Values are scaled to be between 0 and 1, 0 representing the lowest observed expression level for the gene while 1 representing the highest. Samples and genes follow the same order of cell types to emphasize the specificity of the selected genes. 35

6 Expression of known marker genes and newly discovered marker genes in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. A. Expression of new and known markers of purkinje cells in cerebellum. B. Expression of new and known markers of granule cells in dentate gyrus, granule cell layer 36

7 Single-plane image of mouse sensorimotor cortex labeled for Pvalb, Slc32a1, and Cox6a2 mRNAs and counterstained with NeuroTrace. Arrows depict a Slc32a1+/Pvalb+ neuron that is Cox6a2+ (solid arrowhead), a Slc32a1+/Pvalb- neuron that lacks Cox6a2 mRNA (open arrowhead) and a Slc32a1- cell that lacks Cox6a2 mRNA (arrow). Bar = 5 µms. 37
8 Pipeline for the upcoming analyisis on concordance of different cell type based analysis studies. 37
9 Estimations of cortical cell types in frontal cortex and white matter. Values are normalized to be between 0 and 1. Estimations appropriately reflect expected differences between white and gray matter for the most part. It is also possible to see some unexpected increase of some pyramidal subtypes. 38
10 Estimations of purkinje cells in different brain regions. Values are normalized to be between 0 and 1. Purkinje cells are specific to the cerebellum. 39
11 Estimations of dopaminergic cells in different substantia nigra of male parkinson's disease patients. Values are normalized to be between 0 and 1. Dopaminergic cell loss is an expected consequence of Parkinson's Disease 40
12 A-B. Expression of the genes selected from a species in the samples used for isolation from the same species. A shows human genes in human cell type specific expression profile dataset while B is mouse genes in mouse cell type specific expression profile dataset. C-D. Expression of homologues of the genes selected from a species in cell type specific expression profile dataset of the other species. C shows human marker gene expressio in mouse samples while D shows mouse marker gene expression in human samples. 41
13 Estimations done by our method (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. A. Estimations done using marker genes selected from human cell type expression profiles. B. Estimations done using marker genes selected from mouse cell type expression profiles. 42
14 Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. 43

15 Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. Our estimations are much worse for these cells, in the case of memory B cells there is strong negative correlation and we failed to detect enough genes to make an estimation for resting memory T cells. 44
16 Expression of purkinje markers discovered in the study in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. 45
17 Expression of dentate granule cell markers discovered in the study in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. 55

List of Tables

1 A summarization of the datasets collected. Check marks show the methods used to isolate cell types. Number of studies that contain the cell type are given on the right. 25

2 Sources used in generation of the mouse brain cell type database. 26

3 Enriched co-existence of marker genes in cortical single cell samples from mouse brains. Correlation of marker genes in a binary matrix of gene expression is compared to the correlation of randomly selected genes with matching prevalence across the dataset 27

4 Enriched co-existence of marker genes in cortical single cell samples from human brains. Correlation of marker genes in a binary matrix of gene expression is compared to the correlation of randomly selected genes with matching prevalence across the dataset 27

5 Enriched coexpression in cortical cell types. For each marker gene set, random gene sets are created and their co-expression levels are compared to the co-expression levels of the genes in the marker gene set. Above are the p values from a wilcox test that compares the co-expression levels between the gene sets. P values are calculated only for gene sets with more than 2 genes. For Stanley institute datasets, names provided by the institute are used to describe the datasets 28

6 Enriched coexpression in cerebellar cell types. For each marker gene set, random gene sets are created and their co-expression levels are compared to the co-expression levels of the genes in the marker gene set. Above are the p values from a wilcox test that compares the co-expression levels between the gene sets. P values are calculated only for gene sets with more than 2 genes. 28

7 Enriched coexpression in hippocampal cell types. 28

8 Enriched coexpression in cell types of the substantia nigra. 29

9 Enriched coexpression in thalamic cell types. 29

10 Enriched coexpression in cell types found in all brain regions. 29

11 Coefficients found by the linear mixed-effect model. The estimate indicates the estimated effect size by the model. Disease State:PD is the estimated effect of having the Parkinson's disease, Region:Medial is the estimated effect of the sample being from medial substantia nigra and Sex:Male is the estimated effect of the sample being from a male patient. The later rows describe the interaction effect that describes the estimated effect of a sample satifying multiple conditions (being a male parkinson's disease sample for instance). 30

12 Sources used in generation of the mouse blood cell type database. 31

1 Motivation and Introduction

The brain is a heterogeneous organ composed of a wide variety of cell types that can be very closely related (eg. neuronal subtypes) or highly differentiated from each other (eg. neurons and glial cells). While this heterogeneity is well studied, most large scale expression studies that focus on brain disorders use whole tissue samples to examine the effects of diseases[1–3]. Though examination of whole tissue remains popular due to its relative ease and low cost, it does not reveal which cell types are affected by the changes and complicates detection of changes in less abundant cell types due to signal dilution[4]. Studies which do use single cell types are typically examining only one type at a time to compare conditions such as disease[5,6], development[7,8], or they are designed to discover unique properties, such as marker genes or electrophysiological properties, among a small subset of cell types of interest[9,10].

The current state of the literature on neurological diseases and brain cell types leaves several questions unanswered:

1. **What are the marker genes of a cell type in the scope of the entire brain region they reside in?**

   Discovery or marker genes in more general scopes is important because unique cell type markers are needed to identify the cell type in whole tissue samples in many types of experiments (eg. cell type specific microarrays, in-situ hybridization). Even when a cell type has a well defined marker gene, whole experimental design can be compromised if that gene is regulated under the condition tested by the experiment[11]. Having as many marker genes as possible for a specific cell makes this less likely since researchers can swap the marker gene they are using with another one if needed. Marker genes were also shown to be useful in computational settings[4,12,13].

2. **How do cell type proportion's change in neurological diseases?**

   While there are many known cases where the cell type proportions change in the brain such as Parkinson's or Alzheimer's diseases, in many other conditions, changes are much less clear. Experiments that aim to reveal cell type proportion changes require the cell types to be labeled and counted via methods like in situ hybridization or immunohistochemistry. These experiments are lengthy, expensive and often assume that a single marker gene is stably expressed in all samples regardless of the condition.

3. **Under neurologically relevant conditions, which gene expression changes occur in which cell types?**

   Apart from the fact that it is harder to detect differentially expressed genes in whole tissue studies, it is

also uncertain which cell types are effected by the changes when the changes are detected. Acquiring this information requires isolation and expression profiling of the cell types in different conditions and is not practical to perform for a high number of conditions and a high number of cell types.

For this work, I formed a comprehensive database of cell type-specific expression profiles from a variate of resources in an attempt to resolve the problems described above. This database, along with in siloco and in vitro validation methods, is used to detect marker genes that best represent a particular cell type in it's associated brain region. We believe that this list of cell type specific marker genes will be useful to scientific community in characterization of cell types and understanding neurological diseases.

To make further use of the discovered marker genes I will consider their expression levels in whole tissue samples as surrogates for their abundance in the sample. This will allow me to understand the fate of cell type populations under specific diseases and conditions. Finally, I will use these surrogate proportions as covariates in statistical tests in order to improve the statistical power of differential expression analyses and to tell which cell types are affected by specific changes.

The schematic of the project can be found in Figure 1.

2 Research questions and specific aims

2.1 Research questions

2.1.1 What are the specific marker genes of brain cell types?

Cell types of the brain, particularly neurons are loosely defined in terms of their marker genes and properties. Most research focusing on cell types isolates a small number of related cell types and characterizes the cells in relation to each other[7,9]. Relatively few studies[14,15] attempt to characterize cell types in the context of other known cell types of the brain. Absence of such a comprehensive approach in the literature motivates our approach of choosing marker genes using an inclusive database of cell type expression profiles gathered from multiple independent studies to be as inclusive as possible.

2.1.2 Are mouse marker genes applicable to humans?

Most available data in the literature on isolated cell types originates from mouse cells. Ideally researchers would like to have information about human marker genes as well. It is necessary to assess how well marker genes detected in mice can be applied to humans.

2.1.3 How accurately can cell type proportions be predicted with the use of marker genes?

Since marker genes are specific to a cell type by nature, their expression in whole tissue samples can be used as a surrogate for cell type proportion. Even though this is not a new approach, it is necessary to show how accurate it is for the brain.

2.1.4 How do cell type proportions change accross neurological diseases?

It is known that many diseases of the CNS are neurodegenerative in nature. Computational prediction of cell type proportion will allow us to show which cell types are effected in any given condition

2.1.5 Can cell type specific regulatory events be detected using cell type proportion information?

Enumeration of cell types in a sample allows these values to be used as covariates in other models. This information was previously used to improve accuracy of differential expression studies and assign differentially expressed genes to cell types[4]. Applying this method to neurological diseases may uncover cell type specific changes to gene expression.

2.1.6 How do recent single cell experiments correlate with each other and cell type specific microarray samples in the literature?

There has been a recent surge in single cell RNA sequencing experiments attempting to characterize the cell types of the brain[14,15]. Such studies often use different sequencing and clustering methods to define cell types and find marker genes. Due the complex nature of cell type determination and incompleteness of RNA-seq data, it is important to know how well the results correlate with each other and with pre-existing microarray studies working on the same cell types.

2.2 Specific aims

2.2.1 Aim 1: Compilation of cell type specific expression database and make it available to third parties

1. Gather high quality gene expression data representing brain cell types
2. Employ quality control measures to maximize data integrity

3. Make the data available in a web application for easy access.

2.2.2 Aim 2: Identification and verification of marker gene sets

1. Detect cell type marker genes in a region based on the localization of their expression
2. Verify marker genes in independent datasets and by in situ hybridization
3. Asses the concordance of single cell RNA seq data with each other and with the cell types in our database

2.2.3 Aim 3: Estimation of cell type proportions

1. Using marker gene expression, estimate the relative proportions of particular cell types in whole tissue samples an a variety of conditions.
2. Look for generalizable effects in conditions such as neurological diseases.
3. Use datasets on neurological diseases with known effects on cell type composition as positive controls to validate estimation method.
4. Use an independent dataset of isolated blood cell types and manually enumerated blood samples to repeat and validate the estimation method.
5. Use estimated relative proportion information to improve the accuracy of differential expression analyses.
6. Create an R package for easy application of the method by other researchers.

3 Background

3.1 Cell type specific effects in neurological conditions

A major focus of this my thesis will be the detection of cell type specific changes. The brain is composed of a variety of cell types and is one of the most heterogeneous organs in the mammalian body. In neurological diseases and conditions, it is common to see that certain effects are cell type specific. During development in rats, different cell types mature and proliferate at different rates, neurons in general developing earlier than astrocytes and oligodendrocytes[16]. Inflammation, which is observed in many neurological diseases[17], and it is known to hinder neurogenesis[18] and cause neuronal degradation[17]. Neurodegenerative diseases can cause degredation of specific cell types such as Parkinson's Disease with dopaminergic cells[19]. Detection of such changes is possible by using common laboratory methods can be laborious and often impossible due to

scarcity of the samples, as explained in section 3.4, several computational methods were developed to reveal this information from whole tissue expression data.

Cell type specific effects are not limited to proportion changes. Gene expression of individual cell types can also be subject to change under specific conditions. Inflammation for instance, causes significant changes in microglia transcriptome[20] and specific neurological diseases can cause gene expression changes in individual cells. In Parkinson's disease for instance dopaminergic cells both decrease in numbers and experience changes in expression levels in a large number of genes[21]. Just like detecting cell type proportion changes, cell type specific expression changes are difficult by common laboratory methods since it requires isolation and expression profiling of the target cell types. On the other hand whole tissue analysis coupled by cell counts or estimation of cell type proportions have been suggested to allow access to this information[4].

3.2 Cell type markers and their applications

A product of my work is a list of cell type markers. Marker genes are useful in many ways to understand the biology of their associated cell type. Primarily, they can be used to identify cells of interest in whole tissue samples for purposes such as counting and purifying cells. Marker genes are also powerful tools in computational experiments. For example, they can be used as features in deconvolution of complex tissue samples as described in following sections.

Some brain cell types have specific and well known markers that makes them easy to identify. Many of these genes have known functions that are directly related to the cell type. These include Mog, an oligodendrocyte marker[22] with roles in myelination and Aif1, a microglia marker with roles in inflammation[23] for glial cells; Th that is responsible for dopamine synthesis in dopaminergic cells[19], and Gad1/2 that catalyzes production of GABA in gabaergic cells[24] for neurons. Many neuron types, however, do not have known markers. For instance while Gad1/2 are useful markers of gabaergic neurons, many finer subtypes of gabaergic neurons lack known specific markers which makes studying them more challenging[24].

3.3 Methods for cell type purification

In my work, detection of cell type markers will rely on cell type specific gene expression profiles, acquired through isolation of cell types. Isolation of single cell types is necessary as a precursor to their proper characterization, or analysis of specific cells in different conditions such as to diseases or chemicals. There are multiple ways to isolate the cell types of interest with which vary in precision and quality. Most commonly, such methods rely on one or more marker genes specific to the cell type, selectively isolating cells that

express the marker. A well established method is Fluorescence Activated Cell Sorting (FACS) where one or more protein or RNA is labelled to be fluorescently active, either through genetic manipulation or labelled antibodies respectively. Cells are then gated according to specific conditions (eg. expression or absence of a gene) (Figure 2 A)[25]. Another established method of isolation is immunopanning where antibodies layered onto plate are used to hold the cells that express a specific surface marker (Figure 2 B)[26]. A relatively recent marker based isolation is Translating Ribosome Affinity Purification[27]. This method combines the promoter of a marker gene with the coding region of L10a ribosomal subunit fused with green fluorescent protein (GFP)[27]. The tissue is degraded en masse and after fixation, ribosomes marked with GFP are captured, ensuring only translating RNAs from the target cell type are isolated (Figure 2 C)[27,28]. Alternatively, based on visible characteristics or expression of known markers, cells can be visually located on the tissue and isolated by manual extraction or laser capture microdissection (LCM) (Figure 2 D)[29]. The resulting samples from each of these methods have varying purity[13] which is a potential confound on studies that use samples acquired by multiple methods.

3.4 Methods for estimation of cell type proportions from expression patterns

Part of my efforts will be estimation of relative cell type proportions in complex tissue samples in a reference-free manner} Generally, expression profile of complex tissues can be modelled as

$$X_{ij} = \sum_{k=1}^{K} W_{ij} h_{kj} + e_{ij} \tag{1}$$

where X_{ij} is the expression value from a complex sample for genes j and sample i, W_{ik} is a matrix containing cell type proportions for sample i and cell type k, h_{kj} is the cell type specific gene expression of cell type k and gene j and e_{ij} represents random error. Various methods can be applied to acquire information about the matrices W and h. Two main classes of deconvolution methods exist; namely, reference-based and reference-free methods which will be discussed in the next subsections. In mammals, proportion estimation (deconvolution) methods are commonly applied to blood data due to ease of access to both mixed samples and isolated cell types[4,12,30].

While the use of deconvolution in brain is not a new idea, so far, applications have been restricted to a superficial level. Early studies that attempted to estimate cell type proportions in human brains grouped all neurons together as a single cell type alongside astrocytes, oligodendrocytes and microglia[31]. Later, more in depth deconvolution was performed in human cortex and cerebellum which estimated cerebellar neuron types separately while leaving cortical neurons as a single group[32]. Deconvolution of human brains is difficult due

to the absence of human cell type specific expression profiles from human brain cell types which prevents proper use of reference-based deconvolution methods and lack of high numbers of reliable marker genes which prevents reduces the reliability of reference-based deconvolution methods. Expression profiles of the cell types of the mouse brain on the other hand are available in the literature.

A reference-based deconvolution of 64 distinct cell types was performed on whole tissue expression profiles across various brain regions was perfomed by Grange et al.[33]. In the study by Grange et al., proportion estimations of most cell types agreed with the literature, but the authors also reported paradoxical results such as detecting high levels of purkinje cells in thalamus instead of cerebellum[33]. Also no attempt was made to deconvolute cell types in samples from the same regions but under different conditions (eg. disease models, developmental stages).

3.4.1 Reference-based deconvolution

Reference-based deconvolution methods assume we have accurate information about the matrix h: expression profiles of the cell types in the tissue. At the most basic level, researchers try to estimate the W (matrix of cell type proportions) in solving the equation 1 by minimizing the sum of squares of e (error)[33]. This approach assumes that **1)** reference expression profiles are good matches to the actual expression of the cell types in the mixed sample, which can be violated due to noise or differences in RNA extraction methods, and **2)** the reference dataset has all cell types represented in the mixed sample, which can be violated by the presence of previously uncharacterized cell types in the region. To combat such problems, different methods of feature selection that aim to identify the most informative parts the reference expression matrix can be used which makes the estimation process more robust[30].

3.4.2 Reference-free deconvolution

In cases where cell type expression profiles are not available or are likely to have high level of error compared to the real expression of the cell types in the mixed sample, reference-free deconvolution methods provides an alternative. A common method is to use expression of certain marker genes as a surrogate for cell type proportions[4,12,13,32]. Even though the marker genes themselves are often acquired from a reference expression dataset, deconvolution is independent of their expression in the reference. Often the first principal component of the genes in the whole tissue samples is used as a surrogate[4,32,34]. This assumes that most of the used marker genes are not differentially regulated between samples and the main source of variation is the difference in the cell type proportions across samples.

3.5 Expression profiling

Expression profiles are the major component of my work. Microarrays and RNA-seq are the two competing methods for high-throughput expression profiling. I work on expression profiling datasets that are extracted using both of these methods, therefore it is important to make note of properties and limitations of these methods.

3.5.1 Microarrays

RNA microarray is the most common way to quantify RNA in a high-throughput manner and its development have been a transformative force in many branches of biology[35]. Microarrays use single standed probes, specific to a location on the target genome to quantify RNA. These probes are placed on known locations on a solid surface. These probes are later hybridized to a labelled complementary DNA (cDNA) acquired by reverse transcription of a target transcriptome. The amount of cDNA that hybridizes to a probe is quantified by staining the label attached to the cDNA molecules[36].

There are several of microarray platforms available for researchers to choose from. These platforms primarily differ in the probes they use, which can cover a different number of genes and/or cover the same genes using different sequences.

Microarrays are extensively used in neurobiology as a tool for expression analysis. A wide array of data is available at both in the tissue[1,37,38] and isolated cell type[6,7,9] level.

3.5.2 RNA sequencing

RNA sequencing (RNA-seq) is a much more recent method of RNA quantification. RNA seq is performed by sequencing of cDNA molecules acquired from the reverse transcription of an entire target transcriptome. Unlike microarrays, they do not target specific genes, hence can give a much comprehensive picture of the transcriptome including gene discovery and splicing variant identification. Quantification is done by normalizing the number of reads found from a single transcript to the length of the transcript and to the total number of transcripts[39]. A major limitation of RNA-seq is it's accuracy drops significantly in lower expression levels. It was shown that less than 30% of the genes were able to be quantified with high accuracy[40].

Recently, RNA sequencing of single cells is becoming increasingly popular[41]. While single cell RNA-seq is a powerful tool that allows characterization of individual cells in the population, due to scarcity of the starting product, technical biases resulting from amplification and sequencing are more prominent[41]. Single

cell studies are starting to gain popularity in neuroscience[14,42]. Due to its heterogeneous structure, the brain is a prime target for single cell studies that allows differentiation of individual cell types with much less concern of isolating heterogeneous samples.

4 Aim 1: Compilation of cell type specific expression database available to third parties

The first aim of the project, which also provides the groundwork for later stages, is to compile a comprehensive database of cell type specific expression profiles. The database is a valuable resource since it allows comparison of all available cell types to each other, allowing us to find specific expression patterns. The database is collected from the datasets available through Gene Expression Omnibus (GEO) and through personal communications. We made the data available via a web application that allows easy browsing of the data. Mouse cell type specific data being is used due to its higher quality and abundance compared to that available for humans.

4.1 Data acquisition and preprocessing

The bulk of the dataset is based on a previous compilation made by Okaty et al.[13] for a study comparing different cell type isolation methods. This initial dataset was obtained using Affymetrix Mouse Expression 430A Array (430A) and Affymetrix Mouse Genome 430 2.0 Array (430.2). Data from these two platforms is straightforward to combine since the 430A array contains a subset of the probesets in the 430.2 array. Due to the high availability of the data collected 430A and 430.2 arrays and to simplify processing of data, we decided to populate our database with datasets from these platforms. We queried GEO for isolated cell types from mouse samples. In order to pre-process the database, we acquired raw data files (CEL format) for each sample. Samples from the 430.2 array were stripped of the extra probesets they contained and merged with the data from 430A array samples. The resulting dataset was pre-processed and normalized using Robust Multichip Average (RMA) method[43,44]. We observed significant differences in the distribution of the probeset level signal distribution after RMA normalization, potentially due to the technical differences between studies. To make samples comparable to each other, we used a second quantile normalization after RMA[45]. In ideal conditions, batch correction would have been desirable, but since datasets were composed of independent sources with non overlapping cell types, this was not possible. All samples including the ones used in Okaty et. al. study, passed through a quality control phase that involved ensuring expression of known cell type

markers (markers from literature and markers that were used to isolate the cell type) and confirming samples were not contaminated by other cell types by looking for expression of foreign markers. At the end of the cleanup process, cell types were separated into non overlapping groups. This lead to removal of some samples whose associated subtypes were already represented by another sample. For instance samples representing Htr3a positive GABAergic cells were removed due to the presence of VIP positive cells (their subtypes[24]) in another dataset. In addition, when cell types were too similar to each other to detect meaningful differences between them, they were grouped together in a single cell type. For example Drd1 and Drd2 positive spiny neurons' expression profiles were too similar to each other to detect different markers for both. The current dataset has 31 cell types, isolated from 11 regions gathered from 24 studies, and isolated with a variety of methods (Tables 1 and 2). We are still looking at newly published papers in order to add more cell types.

4.2 Presentation of the data in a web application

We created a web application to facilitate access to the cell type database The web application allows users to easily visualize expression of chosen genes in individual cell types in their respective regions (Figure 3). The application also allows grouping of cells together in a hierarchical manner. Every sample shown links to the original data source if it is a publicly available dataset. Future modifications will add the ability to group samples based on sources along with other visualization options. We will also be embedding tools to perform rapid differential expression analyses between cell types, and a gene set enrichment tool that will allow researchers to check a list of genes for cell type specific enrichment. The application increases the usability of our database especially for researchers who are not from a computational background less familiar with computational tools.

5 Aim 2: Identification and Validation of Marker Gene Sets

5.1 Separation of samples into brain regions

A principal use of the comprehensive database we created is to find gene sets with highly enriched expression in single cell types. We reasoned that since most biological samples are from specific brain regions, for marker genes to be biologically and computationally relevant, they should be unique to a single cell type in the context of the associated region. For instance Pvalb is a marker of fast spiking gabaergic cells in cortex but it is a purkinje marker in the cerebellum. To accomplish this, we created groups representing various brain regions. Each cell type in the database is assignetd to a group based on the brain region it was isolated

from and a hierarchy of brain regions (Figure 4). In this hierarchy, cell types that are isolated from regions represented by the lower nodes are included in regions represented by the higher nodes connected to them, for instance samples isolated from cortex are added to the brain regions higher in the hierarchy: cerebrum and whole brain (All). Similarly, cell types isolated from regions represented by higher nodes are included in the regions, represented by lower nodes. For instance, microglia that are isolated from whole brain extracts are added to all other regions lower in the hierarchy. Only exception to this procedure are the astrocyte and oligodendrocyte samples isolated from cortex. Since these cell types are well known to be prevalent across the brain and we did not have samples isolated from other regions, oligodendrocytes are considered to be originated from whole tissue, while astrocytes are added to all regions except cerebellum where they are replaced by Bergmann glia.

5.2 Selection of marker genes

Upon separation of regions we chose specific marker genes for the cell types represented in each region by a clustering based method. For a given cell type, we designated a gene as a marker gene if the following conditions were met:

- There was more than 10 fold change between the median expression of the gene in samples representing the cell type and all other samples from the same region. This condition ensures that there is a large enough difference between the cell type and the rest of the samples to reduce the number of genes selected due to differences that can be study or strain specific.
- When samples were separated into two clusters, those representing the target cell type and all others, with the distance between samples defined as the difference of expression of the target gene, the silhouette coefficient of the resulting clusters was higher than 0.5. This condition prioritizes genes that separate the cell types best.
- By randomly removing samples and equilizing the number of samples chosen from each individual study, we make sure that studies with more samples do not unfairly influence the silhouette coefficient and reduce the effect of outliers.

Since the samples are gathered across multiple samples and cells are isolated by different extraction methods, it is inevitable for there to be technical artifacts, making sample to sample comparison difficult. This was the main advantage for gene selection method over a simple differential expression analysis.

Using those conditions, we selected marker genes across 31 cell types isolated from 11 regions. The number of marker genes greatly varied from one cell type to another depending on the presence of highly similar cell

types in the dataset (Figure 5).

5.3 Validation of marker genes

Finding reliable marker genes using independent datasets is challenging due to artifacts caused by differences in mouse strains, isolation methods and batches. It is also uncertain if marker genes detected using mouse cell types will apply to human cell types. To establish the reliability of our genes, we used the validation methods described below to verify that they act as marker genes in both biological and computational settings.

5.3.1 Validation of marker genes via in situ hybridization

In situ hybridization (ISH), is a well accepted way of assessing the sensitivity and specificity of marker genes, by ensuring that the expression of newly discovered markers and markers from the literature are co-localized to the cells. When possible, we used the ISH data available from the Allen Brain Atlas (ABA)[46] to confirm our findings. While ABA is a powerful resource including thousands of ISH images, every slice is labelled by a probe specific to a single gene. Thus, it is not possible to conclusively decide if the signal is coming from the same cell types unless that cell type is highly concentrated in a specific structure in the brain and are thus identifiable based on location. Granule cells of dentate gyrus and purkinje cells of cerebellum fit this criteria. This allowed us to validate the marker gene specificity for these cell types. The majority of the marker genes with ISH slides available in ABA was shown to be cell type specific (Figures 6, 16 and 17)[46].

We are currently collaborating with Dr. Etienne Sibille (University of Toronto) to validate the markers by dual labelling. Dr. Sibelle's group was able to validate Cox6a2 as a marker of fast spiking gabaergic cells in mice (Figure 7). We will be expanding the number of validated genes through this method, and will apply it to human samples as well.

5.3.2 Computational validation of marker genes in mouse and human single cell data

Since it is unfeasible to apply biological validation methods to all marker genes that we selected, we are using recently published single cell RNA-sequencing datasets[14,15,42] to validate our findings. These data sets have been generated by a recent proliferation of studies aimed at characterizing the cell types of the brain. These studies attempt to define cell types from the ground up by using a variety of clustering methods. Based on descriptions of the identified cell types in the papers, and that the cells are randomly selected from cortices of mouse and men, we have assumed that they have identified the same cell types that our database represents.

Unfortunately, due to the high granularity of the clusters of the single cell data, it is not straightforward to match which individual cells correspond to the cell types defined in our data. To combat this problem, we tried to validate our marker genes in a cell type-agnostic way. For all of our marker gene sets, we checked to see if the genes are more co-expressed than average based on a null distribution of random genes with similar prevalence in the dataset. For human samples, we used the homologues of the marker genes for the same purpose.

Since single cell expression analysis is still in its infancy, often the transcript counts for most genes are being very low, which makes the exact expression value an unreliable measure. Therefore, instead of using the full expression values, we converted the data into a binary matrix where 0 indicated no expression of the gene and 1 indicated any expression of the gene. This approach may be too conservative since we do not chose genes based on their exclusivity to a single cell type, but its heightened expression in the cell type. Despite this, applying the method to a mouse[14] and a human[42] single cell dataset from cortex returned favourable results. In mouse all marker gene sets for cortical cell types were found to be significantly more coexpressed than expected under the null distribution (Table 3), and in the human dataset a majority (7/10) of the marker gene sets showed significantly more coexpression (Table 4).

The next step is to repeat this analysis using a more recently published single cell RNA-seq study performed on mouse brains[15]. This dataset has a better coverage of cortical cell types.

5.3.3 Validation of marker genes in human whole tissue data

As another validation approach, we analysed several human datasets containing multiple brain regions isolated from pathologically healthy humans. The first dataset included samples from 16 different brain regions[47]. Of these 16 regions cell types from, various cortical sub-regions, hippocampus, substantia nigra, thalamus and cerebellum were included in our cell type specific expression dataset. The second dataset included samples from two different cortical regions (Brodmann areas 11 and 47)[48]. Finally, we used 4 datasets that analysed the cortex samples from the Stanley Array Collection. In these studies, we discarded the samples from bipolar disorder and schizophrenia patients.

For all marker gene sets we obtained, the coexpression of marker genes in brain regions relevant to the cell type were analysed. Since marker gene sets are cell type specific, in a complex tissue, variation in the amount of a cell type is determinant of the associated markers' expression. That is, if a sample has higher amount of a cell type, expression of all marker genes for that cell type will be higher. This is expected to result in increased coexpression of marker genes in whole tissue datasets due to the variability in cell type

proportions. We compared the overall level of marker gene coexpression between these samples to coexpression levels of randomly selected genes with similar expression levels. The results showed significantly heightened coexpression for the majority of the cell type marker sets in all datasets analysed (15/22) while PV^+ gabaergic cells and VIP^+ $Reln^+$ cabaergic cells were shown to have a significantly heightened co-expression in the majority of the datasets (Tables 5 to 10).

5.4 Assess condordance of single cell RNA-seq studies with each other and to microarray samples in our database

The high volume of recently published RNA-seq studies has created a large output of data that are very similar to each other in terms of design. All of them are derived from cells of the brains of the same organism, hence in ideal conditions, the cell types they identify should overlap with each other and with our microarray database. Due to the inherent differences in the data structure, assessing repeatability of such single cell studies is not straightforward. The definitions of cell types in our database is based on characterization of the cell types by experts while single cell studies use various clustering methods to seperate their cells into groups and retroactively decide which groups represent which cell types based on the expression of known markers.

We aim to capture similarities between individual cells from RNA-seq datasets and samples from the microarray database by using common genes that are detected by all the datasets in an expression independent manner. For microarray data we will be looking to see if expression of a gene is above the background level. For RNA-seq, we will simply determine whether the gene is captured at all in the sample. This analysis should provide sufficient information to correlate the samples to and allow us to identify which samples from each dataset correspond to each other. However, if we cannot reliably group samples from independent sources together, we will group the single cell data according to their designated groups as identified by the original investigators. We hope to determine whether these independent studies really identify the same cell types or if the cell types are not fully equivalent due to experimental methods or differences between mouse strains. A plan of the proposed analysis can be found in Figure 8.

6 Aim 3: Estimation of cell type proportions

6.1 Estimation of cell type proportions in whole tissue samples using the marker gene sets

A well-established use of marker genes is the estimation of cell type proportions in whole tissue samples using their expression. As discussed in the introduction, two types of deconvolution methods dominate the field: reference-based and reference-free deconvolution. We choose to use reference-free deconvolution because the fact that the reference expression profiles we are using were derived from mice, but we often want to do proportion estimation in human brains. Hence the exact level of expression in our dataset might not be reliable since we are attempting to deconvolute human samples using expression profiles extracted from mouse brains. Genes highly enriched in cell types however are more likely to be functionally relevant to the cell type, hence they are are likely to remain as markers across species. Human astrocytes for instance was shown to have enriched expression of 52% of genes enriched in mouse astrocytes[49]. Therefore it is sensible to believe that sufficiently many marker genes will be preserved across species for our purposes. Our aforementioned validation in human RNA-seq and whole tissue data, along with further validation of our experimental pipeline that will be explained in the next subsection confirms that this is not an unreasonable assumption to make.

To estimate the relative amount of cell types between samples, we used the first principal component of expression of marker genes in the samples as a proxy for cell type proportions. This method has been adopted by multiple other groups performing deconvolution in whole tissues[4,12,32]. The idea behind the method is that most of the variation in marker gene expression will be explained by changes in the cell type proportions. We have implemented countermeasures against genes that do not behave as marker genes or show variation independent from cell type proportion in between samples, such as calculation of rotations based on the control samples, to ensure that genes that do not act as markers do not interfere with the estimation. The result from the analysis is a unitless number per cell type, representing the relative amount of that cell type compared to other samples. This number is used to compare samples only and cannot be used to compare two different cell types.

To check that the method works as expected, we estimated relative cell proportions in different brain regions with known differences between cell type proportions. Our results were concordant with the literature. In a dataset of different brain regions from healthy donors[47] we observed an increase in glial population and decrease in most neuronal populations between white matter and grey matter (Figure 9), and demonstrating

that purkinje cells were exclusive to the cerebellum (Figure 10).

To assess the potential use of the method in the context of brain related disorders, we acquired datasets of substantia nigra expression from healthy donors and Parkinson's disease patients[50,51], characterized by loss of dopaminergic cells in the region. In the first dataset[50] that included samples from healthy donors and Parkinson's disease patients, our initial analysis only showed a significant difference between healthy donors and Parkinson's disease patients in male samples (Figure 11). Further examination of the metadata revealed that female healthy donors had a large age difference (mean age of control females: 85.75, median age of Parkinson's disease females: 71.00). While it was not possible to add the age and other metadata components to a model, due to the way metadata is presented in the original source, age is known to effect dopaminergic cell counts[52]. Apart from that, female patients are also thought to be more resistant to dopaminergic cell loss[53]. The second dataset we used included samples from both lateral and medial substantia nigra[51]. Previous work suggested that medial substantia nigra is less affected by the disease than lateral substantia nigra[54]. To test the region effect and the effect of sex we fitted a linear mixed-effects model that included sex, brain region, disease state of the samples and the interactions between the factors. Patient ID was added as a random effect. It was not possible to add the age to the model because it confounded with the disease, namely, Parkinson's Disease patients were on average, older than controls (Mean age of controls: 69.3, mean age of patients: 80.7). The model confirmed that along with the disease state, that has a large effect size on the estimates for the dopaminergic cells, interaction between the disease state and brain region was also shown to be effective, supporting previous findings that medial substantia nigra has less severe loss of dopaminergic cells[54]. While sex, by itself did not show a large effect size, its interaction with brain region and disease state did, indicating that males might have more cell loss specifically in medial substantia nigra (Table 11). While this can be a novel finding, current literature says little to support or discredit this claim and the sample size is limited.

We will continue to analyze more datasets from brain related disorders and brains under different conditions to increase our confidence in the database apply it as a discovery tool. There are thousands of studies of rodent or human brain tissue samples available in Gemma database[55] for expression profiles with well-annotated metadata to facilitate fast analysis. These datasets include brain samples under various conditions, such as different developmental stages, disorders, treatments, etc. We expect this analysis to reveal many condition specific cell type proportion changes. This step will also provide further validation, since we it will allow us to compare our ability to detect the same proportion changes in different studies working on similar conditions.

6.2 Validation of the pipeline using isolated blood cell types and whole blood data

Estimation of brain cell type proportions is challenging to validate, since, to our knowledge, there are no expression datasets coupled with cell type counts. Any result we find is unverifiable, other than the expected differences between groups. Therefore, to assess the accuracy our method, expression data from whole tissues paired with cell type counts are required. Isolation of blood cell types is much more straightforward than isolation and brain cell types, and can be done more easily without harming the subject. While this data is virtually absent for brain tissue samples, a wide array of blood expression panels are coupled with cell counts, acquired through well established methods[56,57]. cell type reference datasets for blood cell types are present in the literature for both mouse and men. We used these data to construct a similar database to our brain database for mouse (Table 12) and human blood cell types[30]. We subjected both these databases to the same marker gene selection steps. To examine expression changes of marker genes between species, we checked if homologues of genes selected for one species behave as marker genes in other. As expected, not all genes were specific to the cell type they were selected from in different species(Figure 12). To evaluate the ability of marker genes selected for mouse cell types to correctly estimate the relative proportions, we estimated cell type proportions in whole blood cell types and compared our results with a recently published reference-based estimation method[30]. When the cell type definitions were kept at a relatively general level(eg. B cells, T cells) not differentiating cell types at different activation stages, mouse genes performed better than human genes (Figure 13), potentially due to difference of quality between the datasets. On the other hand attempting to estimate finely defined cell (eg. activated/deactivated CD4 cells) types with mouse genes yielded poor correlation to actual counts (Figure 14 - 15). Shay et al.[58] showed that while most lineage specific gene expression is conserved between mouse and human, there are still significant expression changes between a considerable number of genes. This might explain the poor quality of estimations when attempting to estimate the finer subtypes.

6.3 Use proportion estimations to improve accuracy of differential expression analysis

Differential expression analyses on whole tissues are complicated by the heterogeneity of the sample. Since effects are likely to be specific to cell types, having unaffected cell types in the sample will reduce the observed difference, reducing statistical power. Also changes in the cell type proportions can generate false positives. Previous work suggests that it might be possible to increase the power of differential expression analysis by

adding estimated cell type proportions as covariates[4]. The authors also show observed effects can be localized to their cell types by using the estimated proportions in interaction models. In neuroscience of human brains, where sample sizes are often small and data quality is relatively poor, this approach had the potential to increase the value of the existing data by increasing the statistical power of the analyses. We are hoping to validate this approach by finding studies that isolate cell types under certain conditions and controls, paired with other studies that work on the same condition using whole tissue samples. We will assess our ability to assign the differentially expressed genes detected in the single cell type study to that cell type in the study that uses whole tissue samples.

6.4 Create an R package for easy application of the method by third parties.

The pipeline we have developed for gene selection and cell type estimation is time consuming to set up with the magnitude steps aiming to fine tune the process. By creating an R package we will make for our process to be reproducible by other researchers. The package will include streamlined functions to select and validate the marker genes, along with functions used in the estimation process. The package will be publicly available on Bioconductor, CRAN and/or Github platforms.

References

1 Iwamoto K, Kakiuchi C, Bundo M, Ikeda K, Kato T. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Molecular Psychiatry* 2004; **9**: 406–416.

2 Maycox PR, Kelly F, Taylor A, Bates S, Reid J, Logendra R *et al.* Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Molecular Psychiatry* 2009; **14**: 1083–1094.

3 Hokama M, Oka S, Leon J, Ninomiya T, Honda H, Sasaki K *et al.* Altered expression of diabetes-related genes in Alzheimer's disease brains: The Hisayama study. *Cerebral Cortex (New York, NY: 1991)* 2014; **24**: 2476–2488.

4 Chikina M, Zaslavsky E, Sealfon SC. CellCODE: A robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics* 2015;: btv015.

5 Heiman M, Heilbut A, Francardo V, Kulicke R, Fenster RJ, Kolaczyk ED *et al.* Molecular adaptations of

striatal spiny projection neurons during levodopa-induced dyskinesia. *Proceedings of the National Academy of Sciences* 2014; **111**: 4578–4583.

6 Dougherty JD, Maloney SE, Wozniak DF, Rieger MA, Sonnenblick L, Coppola G *et al.* The Disruption of Celf6, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors. *The Journal of Neuroscience* 2013; **33**: 2732–2753.

7 Okaty BW, Miller MN, Sugino K, Hempel CM, Nelson SB. Transcriptional and electrophysiological maturation of neocortical fastspiking GABAergic interneurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 2009; **29**: 7040–7052.

8 Bellesi M, Pfister-Genskow M, Maret S, Keles S, Tononi G, Cirelli C. Effects of Sleep and Wake on Oligodendrocytes and Their Precursors. *The Journal of Neuroscience* 2013; **33**: 14288–14300.

9 Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C *et al.* Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience* 2006; **9**: 99–107.

10 Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G *et al.* Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types. *Cell* 2008; **135**: 749–762.

11 Sommeijer J-P, Levelt CN. Synaptotagmin-2 Is a Reliable Marker for Parvalbumin Positive Inhibitory Boutons in the Mouse Visual Cortex. *PLoS ONE* 2012; **7**: e35323.

12 Westra H-J, Arends D, Esko T, Peters MJ, Schurmann C, Schramm K *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet* 2015; **11**: e1005223.

13 Okaty BW, Sugino K, Nelson SB. A Quantitative Comparison of Cell-Type-Specific Microarray Gene Expression Profiling Methods in the Mouse Brain. *PLoS ONE* 2011; **6**: e16493.

14 Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, Manno GL, Juréus A *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015; **347**: 1138–1142.

15 Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience* 2016; **19**: 335–346.

16 Sauvageot CM, Stiles CD. Molecular mechanisms controlling cortical gliogenesis. *Current Opinion in Neurobiology* 2002; **12**: 244–249.

17 Aguzzi A, Barres BA, Bennett ML. Microglia: Scapegoat, Saboteur, or Something Else? *Science* 2013; **339**: 156–161.

18 Ekdahl CT, Kokaia Z, Lindvall O. Brain inflammation and adult neurogenesis: The dual role of microglia.

Neuroscience 2009; **158**: 1021–1029.

19 Hegarty SV, Sullivan AM, O'Keeffe GW. Midbrain dopaminergic neurons: A review of the molecular circuitry that regulates their development. *Developmental Biology* 2013; **379**: 123–138.

20 Holtman IR, Noback M, Bijlsma M, Duong KN, van der Geest MA, Ketelaars PT *et al.* Glia Open Access Database (GOAD): A comprehensive gene expression encyclopedia of glia cells in health and disease. *Glia* 2015;: n/a–n/a.

21 Simunovic F, Yi M, Wang Y, Macey L, Brown LT, Krichevsky AM *et al.* Gene expression profiling of substantia nigra dopamine neurons: Further insights into Parkinson's disease pathology. *Brain* 2009; **132**: 1795–1809.

22 Linington C, Bradl M, Lassmann H, Brunner C, Vass K. Augmentation of demyelination in rat acute allergic encephalomyelitis by circulating mouse monoclonal antibodies directed against a myelin/oligodendrocyte glycoprotein. *The American Journal of Pathology* 1988; **130**: 443–454.

23 Imai Y, Ibata I, Ito D, Ohsawa K, Kohsaka S. A novel gene iba1 in the major histocompatibility complex class III region encoding an EF hand protein expressed in a monocytic lineage. *Biochemical and Biophysical Research Communications* 1996; **224**: 855–862.

24 Rudy B, Fishell G, Lee S, Hjerling-Leffler J. Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Developmental Neurobiology* 2011; **71**: 45–61.

25 Kang N-Y, Yun S-W, Ha H-H, Park S-J, Chang Y-T. Embryonic and induced pluripotent stem cell staining and sorting with the live-cell fluorescence imaging probe CDy1. *Nature Protocols* 2011; **6**: 1044–1052.

26 Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *The Journal of Neuroscience* 2008; **28**: 264–278.

27 Heiman M, Kulicke R, Fenster RJ, Greengard P, Heintz N. Cell typespecific mRNA purification by translating ribosome affinity purification (TRAP). *Nature Protocols* 2014; **9**: 1282–1291.

28 Sanz E, Yang L, Su T, Morris DR, McKnight GS, Amieux PS. Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proceedings of the National Academy of Sciences* 2009; **106**: 13939–13944.

29 Liotta L, Petricoin E. Molecular profiling of human cancer. *Nature Reviews Genetics* 2000; **1**: 48–56.

30 Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y *et al.* Robust enumeration of cell subsets

from tissue expression profiles. *Nature Methods* 2015; **12**: 453–457.

31 Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature Methods* 2011; **8**: 945–947.

32 Xu X, Nehorai A, Dougherty JD. Cell type-specific analysis of human brain transcriptome data to predict alterations in cellular composition. *Systems Biomedicine* 2013; **1**: 151–160.

33 Grange P, Bohland JW, Okaty BW, Sugino K, Bokil H, Nelson SB *et al.* Cell-typebased model explaining coexpression patterns of genes in the brain. *Proceedings of the National Academy of Sciences* 2014; **111**: 5397–5402.

34 Tan PPC, French L, Pavlidis P. Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain. *Neurogenomics* 2013; **7**: 5.

35 Hoheisel JD. Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews Genetics* 2006; **7**: 200–210.

36 Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R *et al.* A concise guide to cDNA microarray analysis. *BioTechniques* 2000; **29**: 548–550, 552–554, 556 passim.

37 Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M *et al.* Spatio-temporal transcriptome of the human brain. *Nature* 2011; **478**: 483–489.

38 Torrey EF, Webster M, Knable M, Johnston N, Yolken RH. The Stanley Foundation brain collection and Neuropathology Consortium. *Schizophrenia Research* 2000; **44**: 151–155.

39 Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 2009; **10**: 57–63.

40 Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* 2011; **27**: i383–i391.

41 Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* 2015; **58**: 610–620.

42 Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* 2015; **112**: 7285–7290.

43 Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*

2003; **4**: 249–264.

44 Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* 2003; **31**: e15.

45 Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* 2003; **19**: 185–193.

46 Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 2007; **445**: 168–176.

47 Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nature Communications* 2013; **4**. doi:10.1038/ncomms3771.

48 Chen C-Y, Logan RW, Ma T, Lewis DA, Tseng GC, Sibille E *et al.* Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 2016; **113**: 206–211.

49 Zhang Y, Sloan SA, Clarke LE, Caneda C, Plaza CA, Blumenthal PD *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* 2016; **89**: 37–53.

50 Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M *et al.* A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease. *PLoS genetics* 2007; **3**: e98.

51 Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RKB, Graeber MB. Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. *Neurogenetics* 2006; **7**: 1–11.

52 Costa KM. The Effects of Aging on Substantia Nigra Dopamine Neurons. *The Journal of Neuroscience* 2014; **34**: 15133–15134.

53 Gillies GE, Pienaar IS, Vohra S, Qamhawi Z. Sex differences in Parkinson's disease. *Frontiers in Neuroendocrinology* 2014; **35**: 370–384.

54 Duke DC, Moran LB, Pearce RKB, Graeber MB. The medial and lateral substantia nigra in Parkinson's disease: mRNA profiles associated with higher brain tissue vulnerability. *Neurogenetics* 2007; **8**: 83–94.

55 Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JRC, Van Rossum T *et al.* Gemma: A resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics (Oxford, England)* 2012; **28**: 2272–2273.

56 Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data

identifies cellular activation patterns in systemic lupus erythematosus. *PloS One* 2009; **4**: e6098.

57 Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* 2011; **144**: 296–309.

58 Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, Feng T *et al.* Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proceedings of the National Academy of Sciences of the United States of America* 2013; **110**: 2946–2951.

7 Tables

	FACS	LCM	Manual	PAN	PAN.FACS	TRAP	Studies
Astrocyte	✓				✓		2
Basket			✓			✓	2
Bergmann						✓	1
CerebGranule						✓	1
Cholinergic						✓	2
DentateGranule		✓					1
Dopaminergic		✓					2
Ependymal	✓						1
FS Basket (G42)			✓				3
GabaReln			✓				1
GabaSSTReln			✓				1
Gluta			✓				1
Golgi						✓	1
Hypocretinergic						✓	1
Martinotti (GIN)			✓				1
Microglia	✓						1
MotorCholin						✓	1
Oligodendrocyte				✓		✓	1
Purkinje		✓	✓			✓	4
PyramidalCorticoThalam			✓				6
Pyramidal_Glt_25d2						✓	1
Pyramidal_S100a10						✓	1
Pyramidal_Thy1			✓				1
Serotonergic						✓	1
Spiny						✓	1
Th_positive_LC			✓				4
VIPReln (G30)			✓				2

Table 1: A summarization of the datasets collected. Check marks show the methods used to isolate cell types. Number of studies that contain the cell type are given on the right.

Cell type	PMID
Doyle et al., 2008	19013282
Cahoy et al., 2008	18171944
Sugino et al., 2006	16369481
Okaty et al., 2009	19474331
Anandasabapathy et al., 2011	21788405
Rossner et al., 2006	17005859
Chung et al., 2005	15888489
Unpublished	NA
Beckervordersandforth et al 2010	21112568
Perrone-Bizzozero NI et al. 2011	22004431
Maze et al 2014	24584053
Heiman et al 2014	24599591
Tan et al 2013	24311694
Schmidt et al 2012	22632977
Dalal et al 2013	23431030
Fomchenko et al 2011	21754979
Bellesi et al 2013	24005282
Paul et al 2012	22754500
Galloway et al 2014	24986919
Dougherty et al 2013	23407934
Zamanian et al 2012	22553043
G??rlich et al 2013	24082085
Sugino et al. 2014	25232122
Phani et al. 2015	20462502

Table 2: Sources used in generation of the mouse brain cell type database.

Cell Types	p-value
Astrocyte	p<0.001
Microglia	p<0.001
Oligo	p<0.001
GabaPV	0.001
GabaRelnCalb	0.015
GabaVIPReln	p<0.001
PyramidalCorticoThalam	p<0.001
Pyramidal_Glt_25d2	0.039
Pyramidal_S100a10	p<0.001
Pyramidal_Thy1	0.024

Table 3: Enriched co-existence of marker genes in cortical single cell samples from mouse brains. Correlation of marker genes in a binary matrix of gene expression is compared to the correlation of randomly selected genes with matching prevalence across the dataset

Cell Types	p-value
Astrocyte	p<0.001
Microglia	p<0.001
Oligo	p<0.001
GabaPV	p<0.001
GabaRelnCalb	p<0.001
GabaVIPReln	p<0.001
PyramidalCorticoThalam	0.494
Pyramidal_Glt_25d2	0.382
Pyramidal_S100a10	0.003
Pyramidal_Thy1	0.494

Table 4: Enriched co-existence of marker genes in cortical single cell samples from human brains. Correlation of marker genes in a binary matrix of gene expression is compared to the correlation of randomly selected genes with matching prevalence across the dataset

Cell Types	UCL Dataset	Sibille Dataset	Stanley - AltarA	Stanley - Bahn	Stanley - Dobrin	Stanley - Kato
GabaPV	0.205	p<0.001	0.116	p<0.001	p<0.001	p<0.001
GabaVIPReln	p<0.001	p<0.001	0.268	0.021	p<0.001	0.023
PyramidalCorticoThalam	0.724	0.875	0.8	0.875	0.809	0.635
Pyramidal_S100a10	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001
Pyramidal_Thy1	0.038	0.04	genes<3	genes<3	0.679	genes<3

Table 5: Enriched coexpression in cortical cell types. For each marker gene set, random gene sets are created and their co-expression levels are compared to the co-expression levels of the genes in the marker gene set. Above are the p values from a wilcox test that compares the co-expression levels between the gene sets. P values are calculated only for gene sets with more than 2 genes. For Stanley institute datasets, names provided by the institute are used to describe the datasets

Cell Type	p-value
Basket	0.11
Bergmann	p<0.001
CerebGranule	0.268
Golgi	0.679
Purkinje	p<0.001

Table 6: Enriched coexpression in cerebellar cell types. For each marker gene set, random gene sets are created and their co-expression levels are compared to the co-expression levels of the genes in the marker gene set. Above are the p values from a wilcox test that compares the co-expression levels between the gene sets. P values are calculated only for gene sets with more than 2 genes.

Cell Type	p-value
DentateGranule	p<0.001
GabaSSTReln	p<0.001
Pyramidal_Thy1	p<0.001

Table 7: Enriched coexpression in hippocampal cell types.

Cell Type	p-value
BrainstemCholin	0.008
Dopaminergic	p<0.001
Serotonergic	0.001

Table 8: Enriched coexpression in cell types of the substantia nigra.

Cell Type	p-value
GabaReln	p<0.001
Hypocretinergic	0.015
ThalamusCholin	0.001

Table 9: Enriched coexpression in thalamic cell types.

Cell Type	p-value
Astrocyte	p<0.001
Microglia	p<0.001
Oligo	p<0.001

Table 10: Enriched coexpression in cell types found in all brain regions.

	Estimate	Std..Error
Intercept	0.6527065	0.1775530
Disease State:PD	-0.3784387	0.1957226
Region:Medial	-0.3040865	0.1019972
Sex:Male	0.0510080	0.1926069
Disease State:PD, Region:Medial	0.4180618	0.1138480
Disease State:PD, Sex:Male	-0.0374649	0.2213898
Region:Medial, Sex: Male	0.3578188	0.1116042
Disease State:PD, Region:Medial, Sex:Male	-0.4197041	0.1323141

Table 11: Coefficients found by the linear mixed-effect model. The estimate indicates the estimated effect size by the model. Disease State:PD is the estimated effect of having the Parkinson's disease, Region:Medial is the estimated effect of the sample being from medial substantia nigra and Sex:Male is the estimated effect of the sample being from a male patient. The later rows describe the interaction effect that describes the estimated effect of a sample satifying multiple conditions (being a male parkinson's disease sample for instance).

Reference	PMID
Toker et al. 2013	23420886
Maruyama et al. 2012	23200825
Yao et al. 2014	24394418
Haldar et al.	Unpublished
Menssen et al. 2009	19265543
Lotem et al. 2013	24236182
Tanaka et al. 2014	25236782
Li et al. 2015	25526089
Yao et al 2015	25527787
Lindvall et al. 2006	16764821
Moriyama et al. 2014	24913235
Berrien-Elliot et al. 2015	25516478
Tartey et al. 2014	25107474
McKinstry et al. 2014	25369785
Kramer et al. 2013	25931581
Kramer et al. 2014	25931582
Kramer et al. 2015	25931583
Vahl et al. 2014	25464853
Wang el al.	Unpublished
Holmes et al. 2015	25398911
Nakano et al. 2015	25769922
Ortutay et al. 2015	25926688
Yang et el al. 2015	26390156
Fehniger et al. 2007	17540585
Tomayko et al. 2008	18566367
Zietara et al. 2013	23345431
Cao et al.	Unpublished
Somervaille et al. 2009	19200802
Ingersoll et al. 2010	19965649
Guo et al. 2010	20703300
Laird et al. 2010	20974990
Konuma et al. 2011	21540074
Kuczma et al. 2011	21642545
Kaji et al. 2012	23027924
Jung et al. 2013	23248261
Shen et al. 2014	24572363
Petersen et al. 2012	22543263
Luckey et al. 2006	16492737
Baranek et al. 2012	23084923

Table 12: Sources used in generation of the mouse blood cell type database.

8 Figures

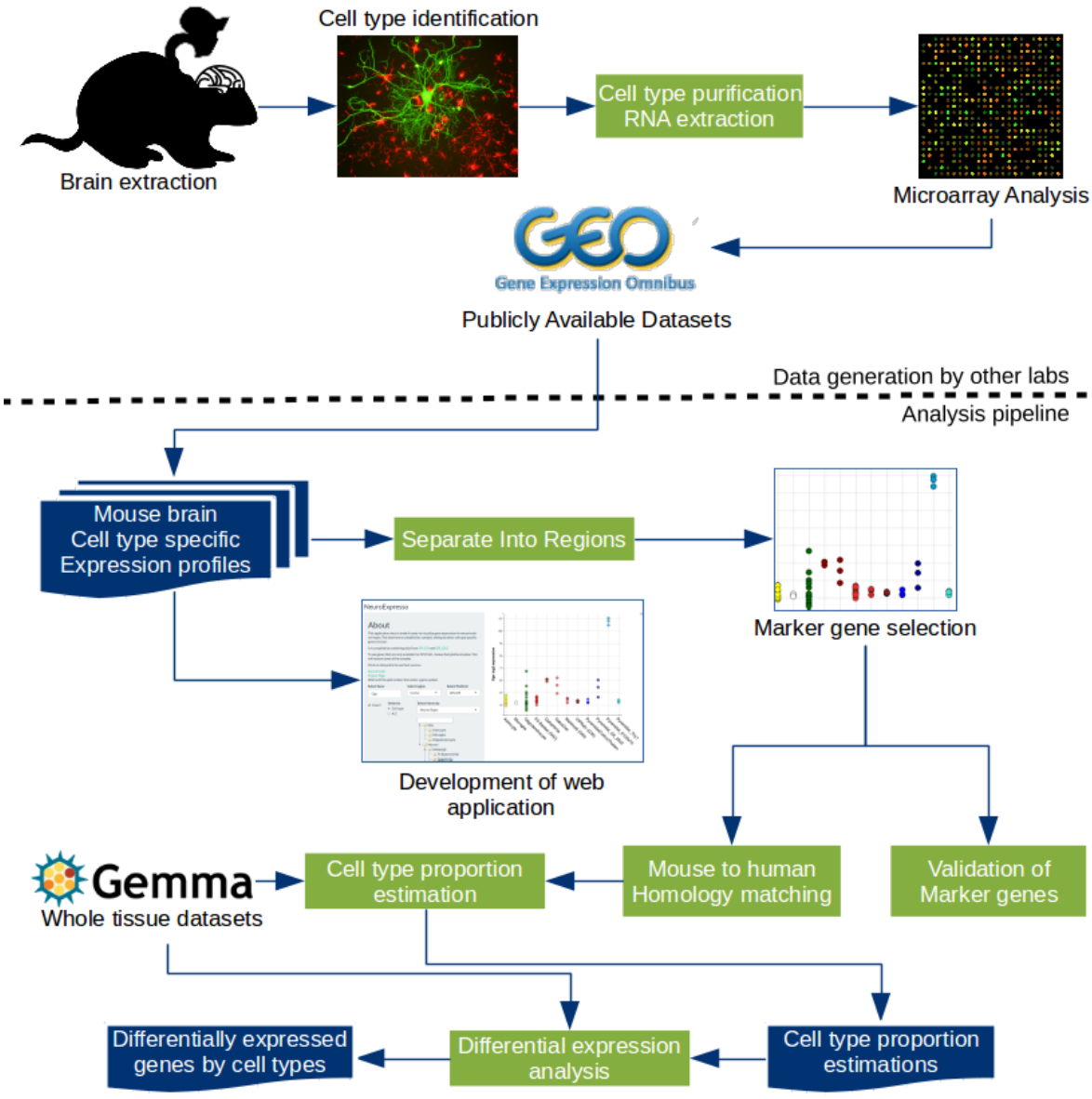


Figure 1: Workflow of the project

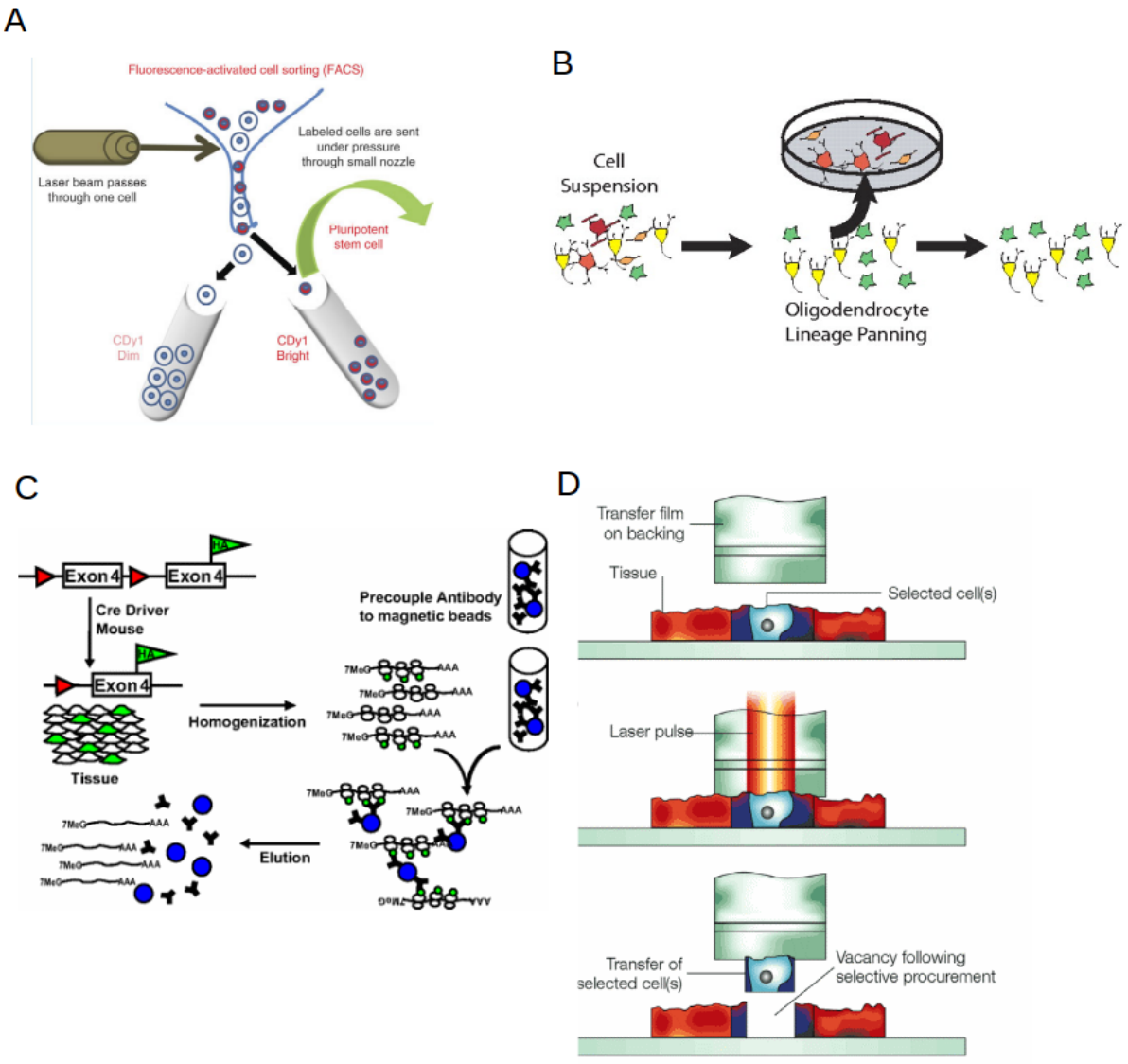


Figure 2: Example representations of cell type isolation techniques. A. Adapted from Kang et al. 2011. An example application of FACS. A fluorescently active molecule is used to label specific cell in the population. Upon detection of fluorescence, a charge is placed on the droplet whose path is later manipulated by an electrical field to separate the cells. B. Adapted from Cahoy et al. 2008. An example application of TRAP. A cell suspension is placed into a plate with bounded antibodies binding to a specific cell type. Removing suspended cells removes the cell type from the population. C. Adapted from Sanz et al. 2009. A schematic representation of the TRAP method. A cell type specific promoter driven expression of a labelled ribosome component causes certain cells to contain labelled ribosomes. The tissue is homogenized as a whole and fixed. Labelled ribosomes that carry RNAs from specific cells are isolated. Removal of ribosomes leaves cell type specific RNA samples behind. D. Adapted from Liotta et al. 2000. A schematic representation of LCM method. Cells are visually identified on the slide and marked. A laser then cuts the marked part and separates it from the rest of the samples.

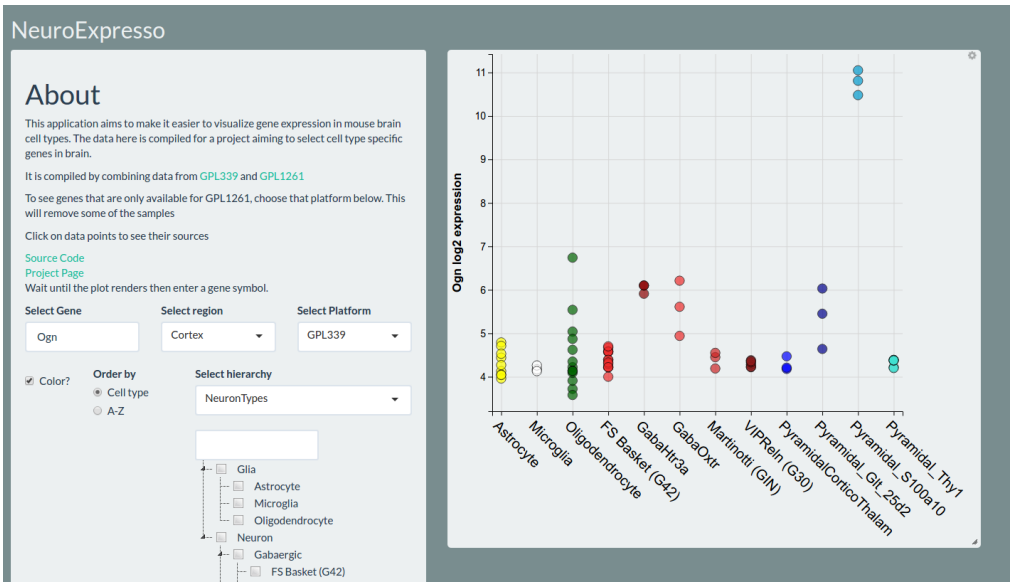


Figure 3: A screenshot of the NeuroExpresso web application: A tool to visualize gene expression in the cell types of our database.

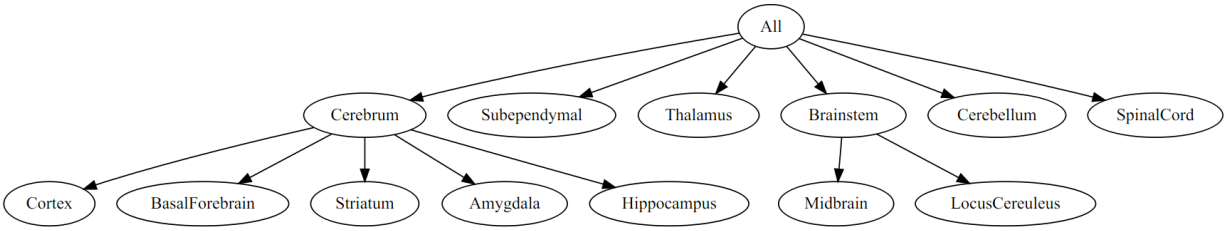


Figure 4: Hierarchy of brain regions used to separate the cell types into groups representing different regions of the brain. Cell types isolated from the regions in the lower nodes of the hierarchy are added to the higher nodes connected to them, while cell types isolated from the regions in the higher nodes are inlcuded in the lower nodes.

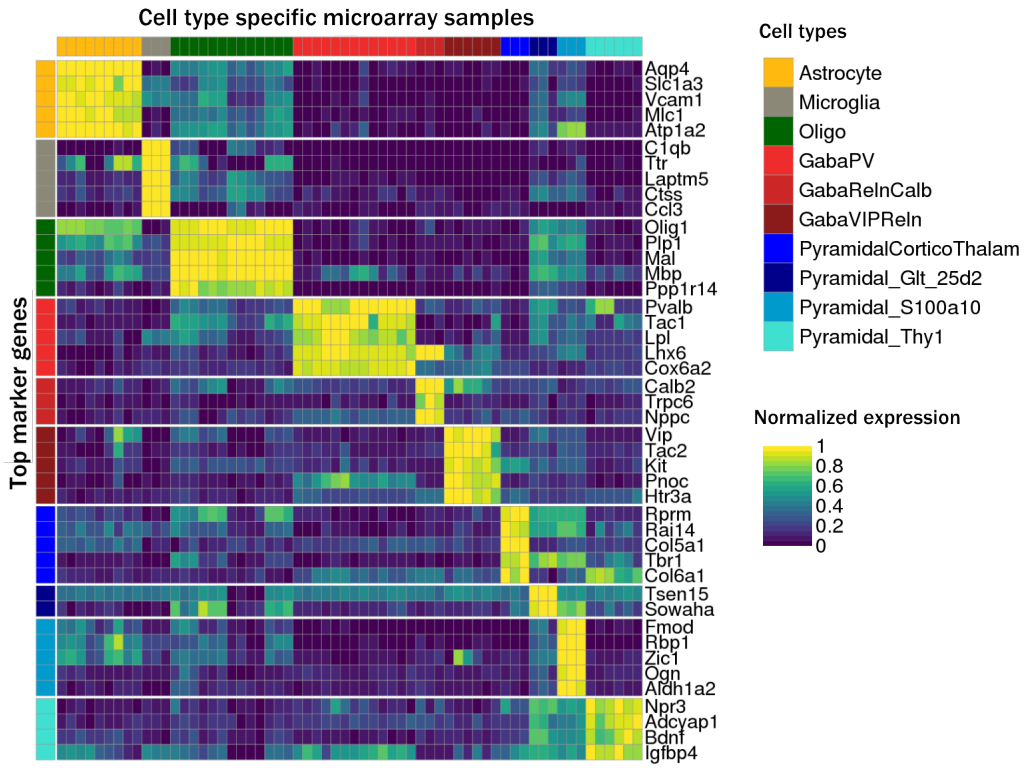


Figure 5: Expression of top 5 marker genes detected from cortex cell types. Values are scaled to be between 0 and 1, 0 representing the lowest observed expression level for the gene while 1 representing the highest. Samples and genes follow the same order of cell types to emphasize the specificity of the selected genes.

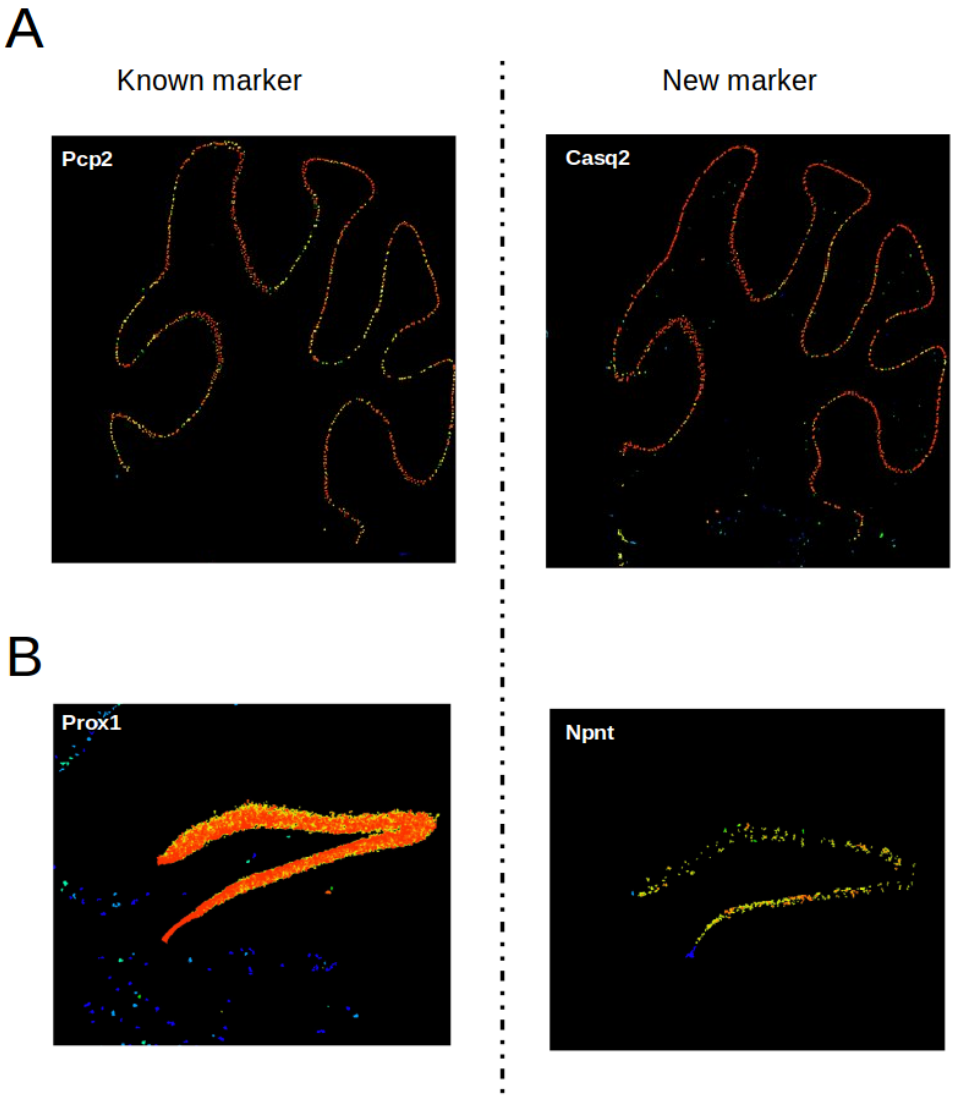


Figure 6: Expression of known marker genes and newly discovered marker genes in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. A. Expression of new and known markers of purkinje cells in cerebellum. B. Expression of new and known markers of granule cells in dentate gyrus, granule cell layer

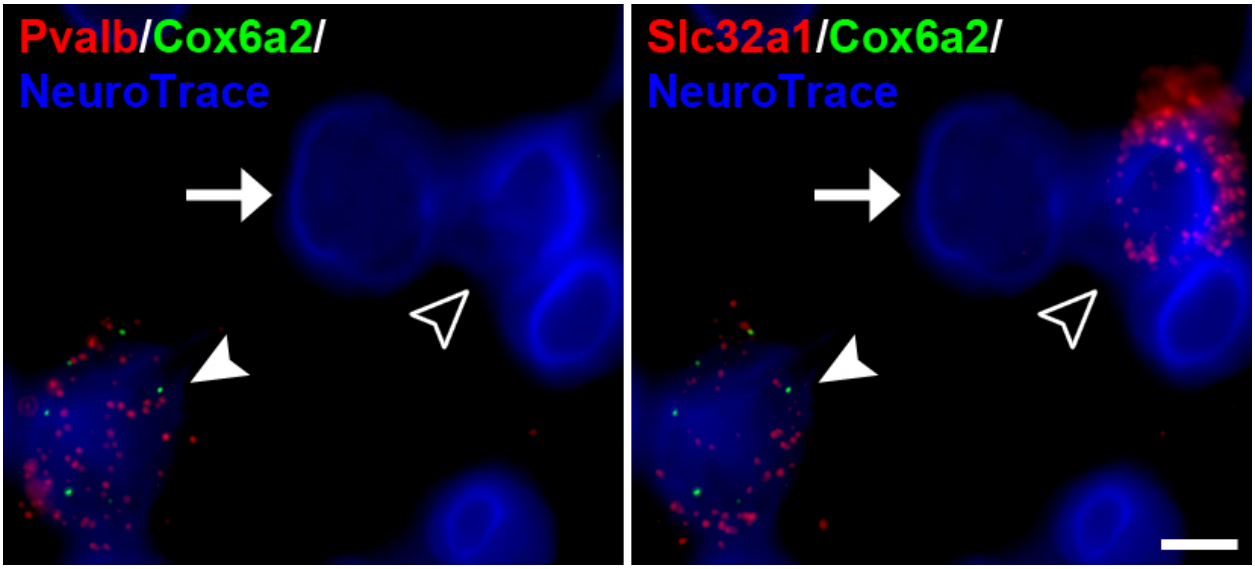


Figure 7: Single-plane image of mouse sensorimotor cortex labeled for Pvalb, Slc32a1, and Cox6a2 mRNAs and counterstained with NeuroTrace. Arrows depict a Slc32a1+/Pvalb+ neuron that is Cox6a2+ (solid arrowhead), a Slc32a1+/Pvalb- neuron that lacks Cox6a2 mRNA (open arrowhead) and a Slc32a1- cell that lacks Cox6a2 mRNA (arrow). Bar = 5 µms.

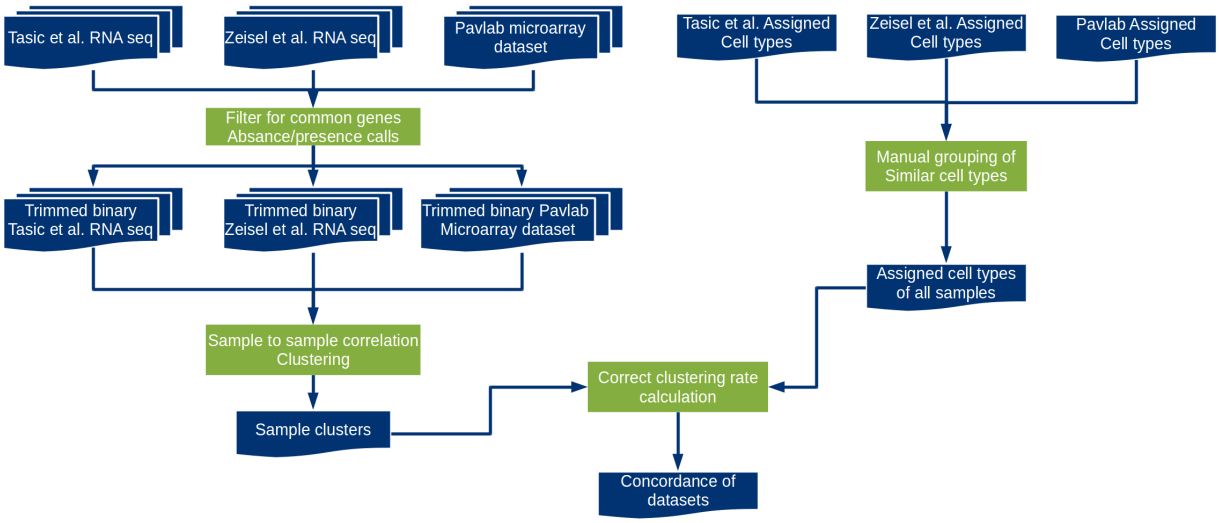


Figure 8: Pipeline for the upcoming analyisis on concordance of different cell type based analysis studies.

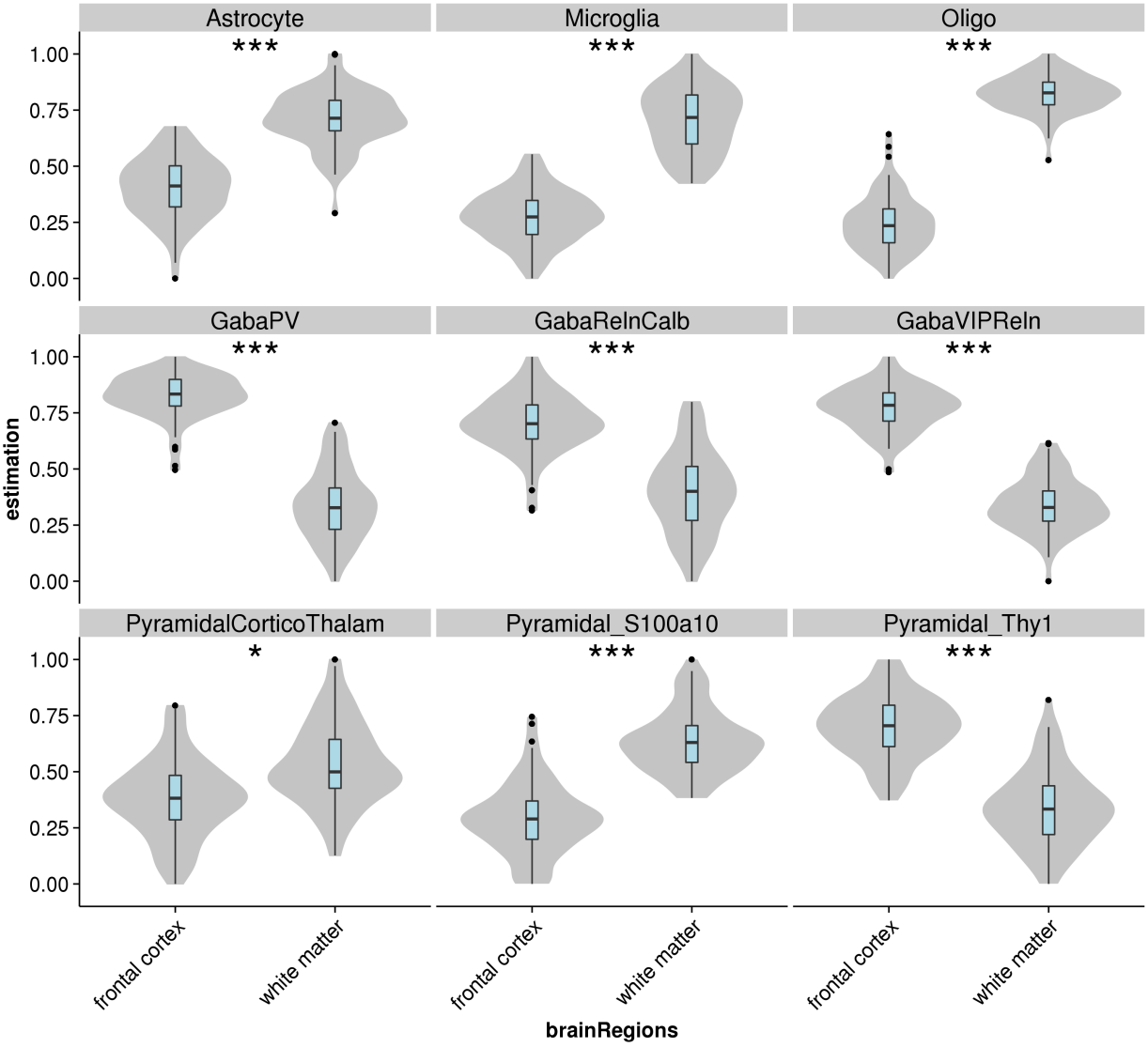


Figure 9: Estimations of cortical cell types in frontal cortex and white matter. Values are normalized to be between 0 and 1. Estimations appropriately reflect expected differences between white and gray matter for the most part. It is also possible to see some unexpected increase of some pyramidal subtypes.

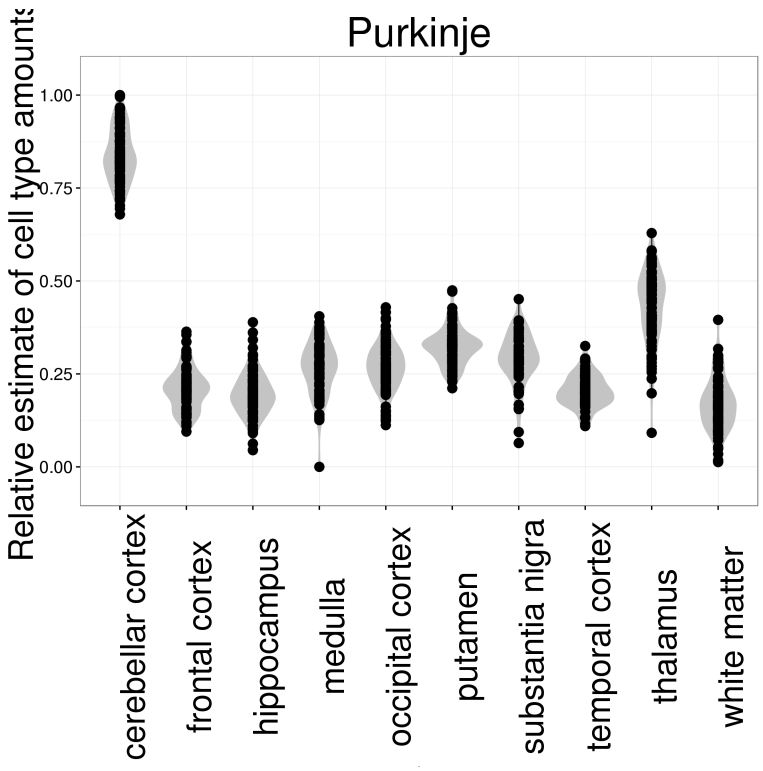


Figure 10: Estimations of purkinje cells in different brain regions. Values are normalized to be between 0 and 1. Purkinje cells are specific to the cerebellum.

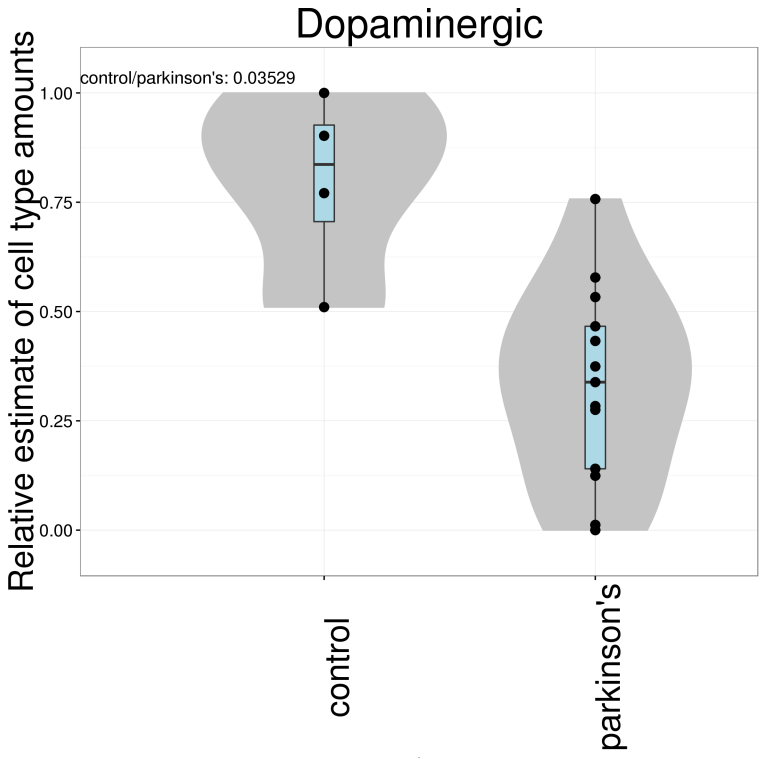


Figure 11: Estimations of dopaminergic cells in different substantia nigra of male parkinson's disease patients. Values are normalized to be between 0 and 1. Dopaminergic cell loss is an expected consequence of Parkinson's Disease

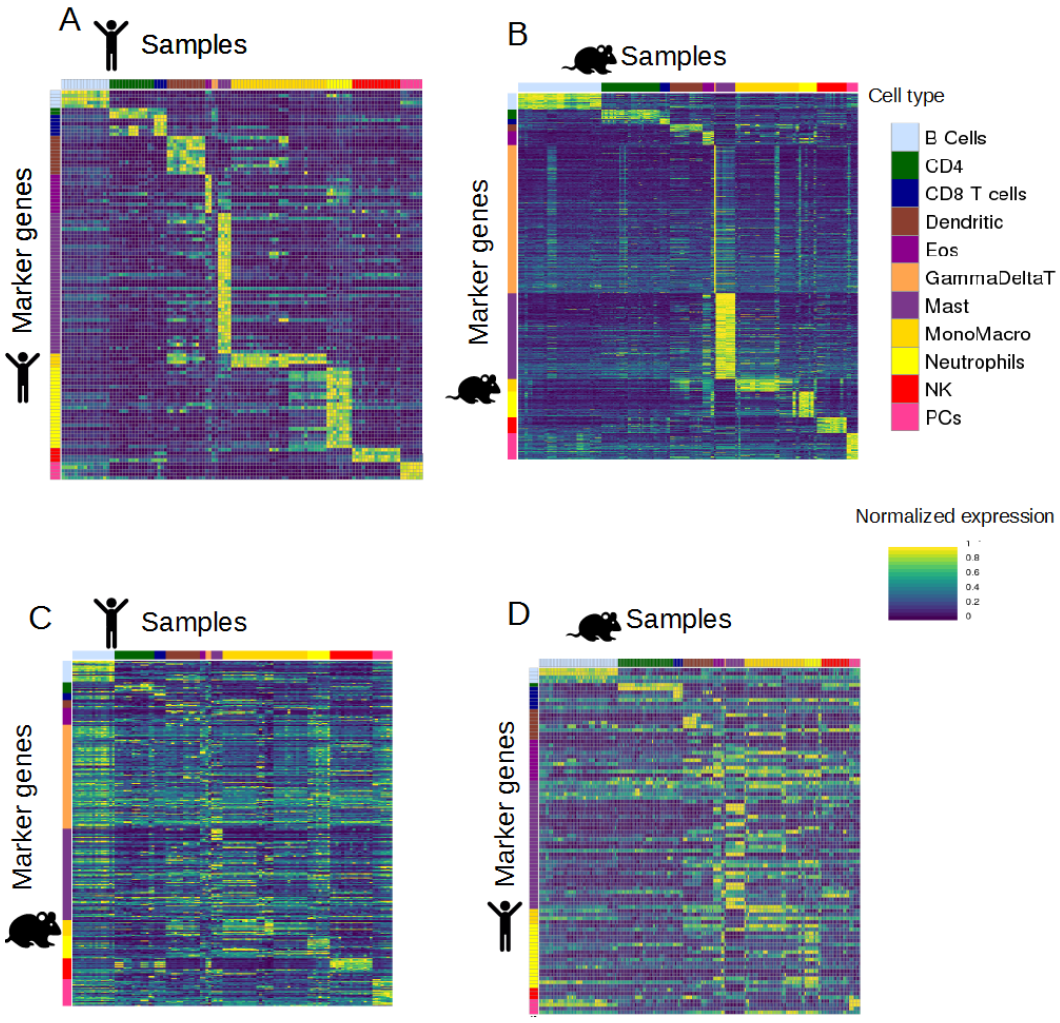


Figure 12: A-B. Expression of the genes selected from a species in the samples used for isolation from the same species. A shows human genes in human cell type specific expression profile dataset while B is mouse genes in mouse cell type specific expression profile dataset. C-D. Expression of homologues of the genes selected from a species in cell type specific expression profile dataset of the other species. C shows human marker gene expressio in mouse samples while D shows mouse marker gene expression in human samples.

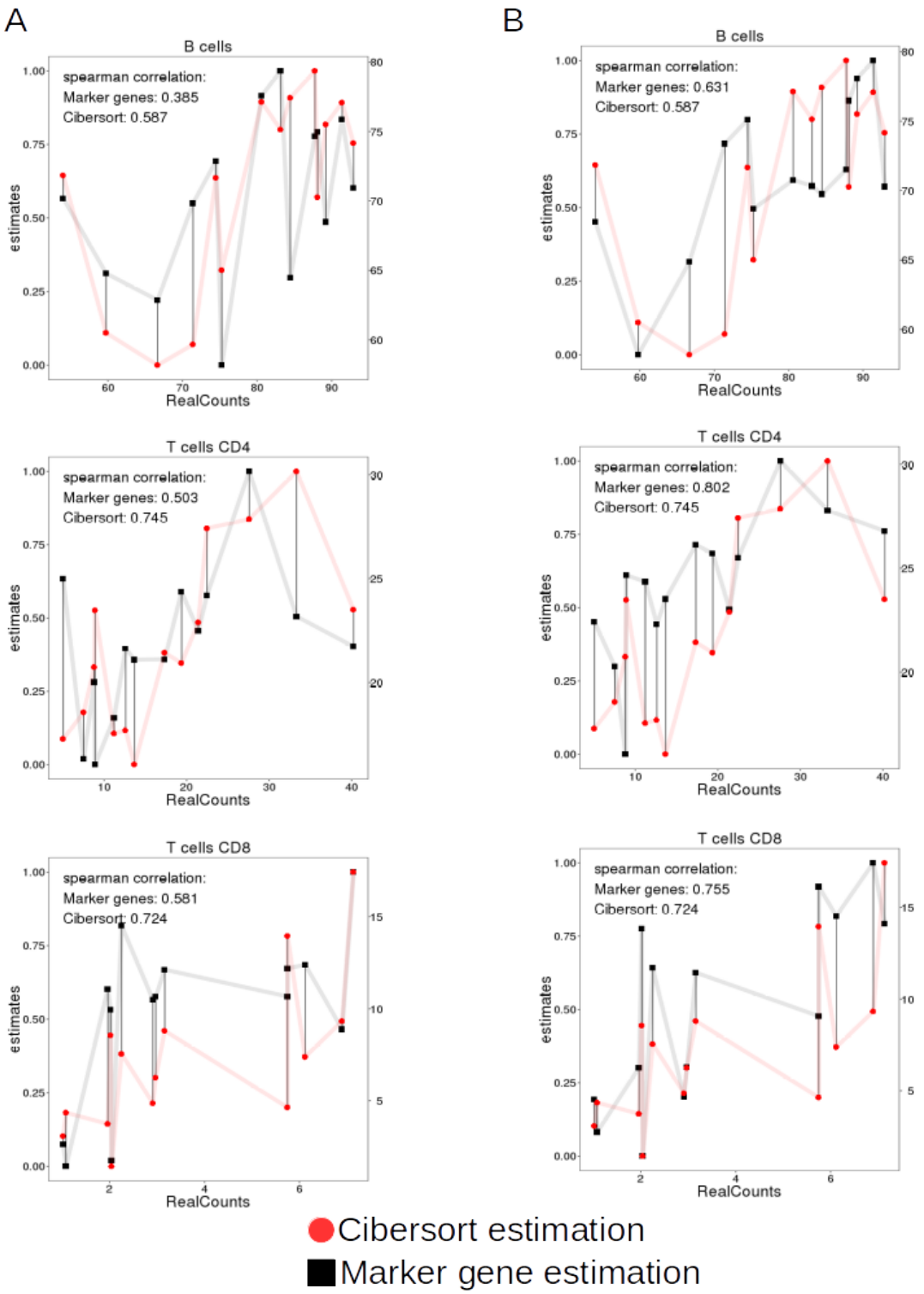


Figure 13: Estimations done by our method (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. A. Estimations done using marker genes selected from human cell type expression profiles. B. Estimations done using marker genes selected from mouse cell type expression profiles.

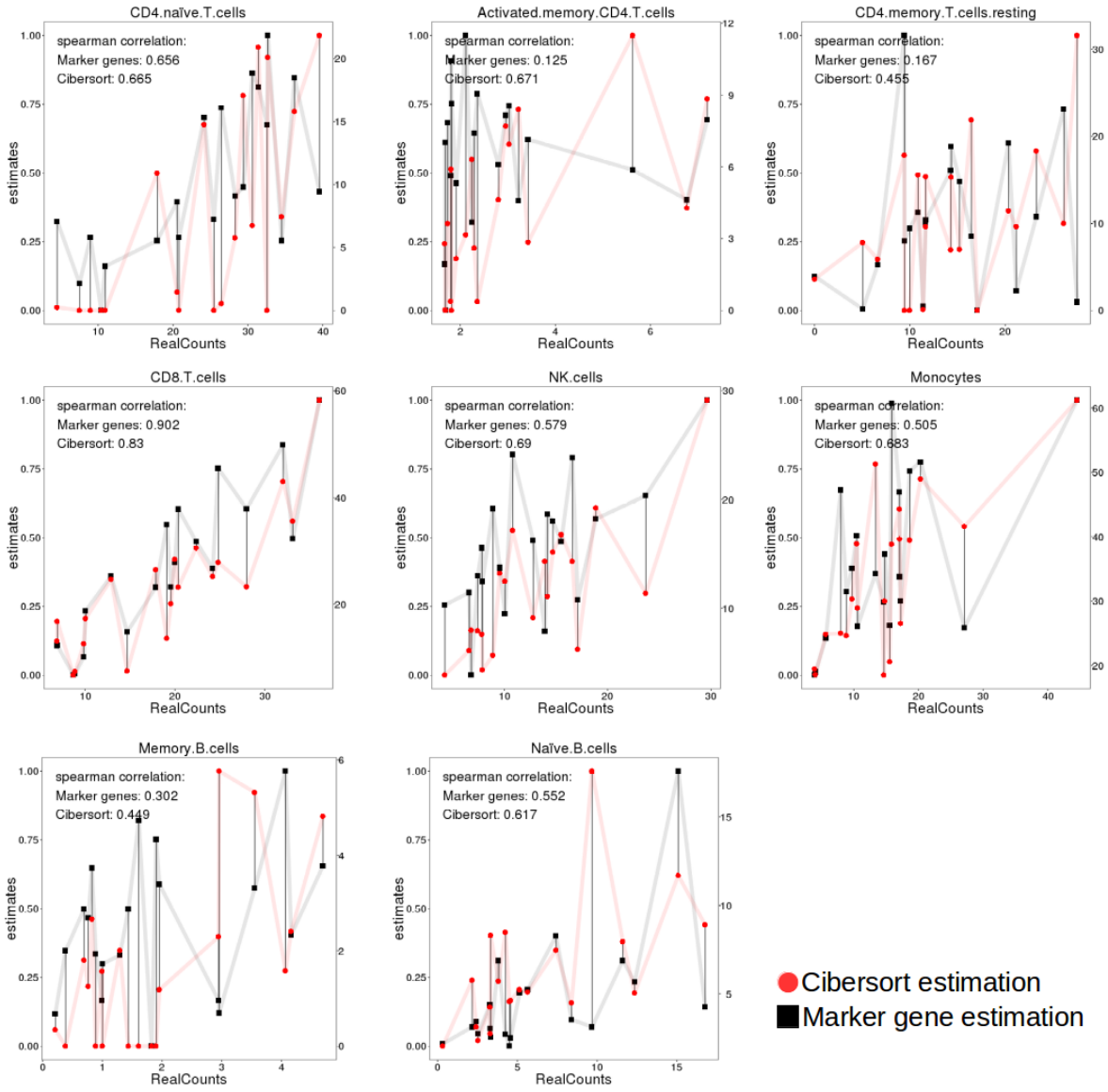


Figure 14: Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage.

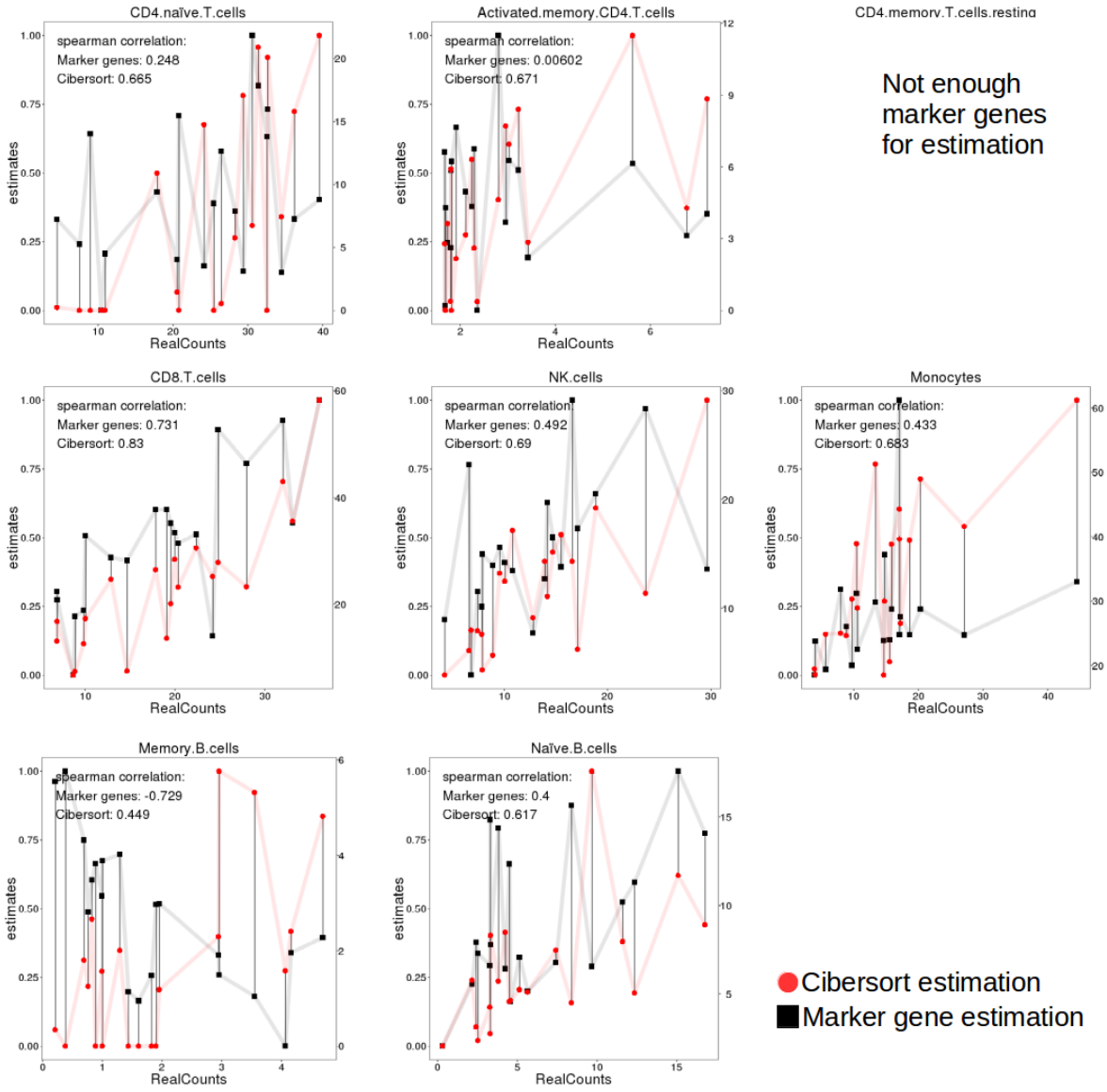
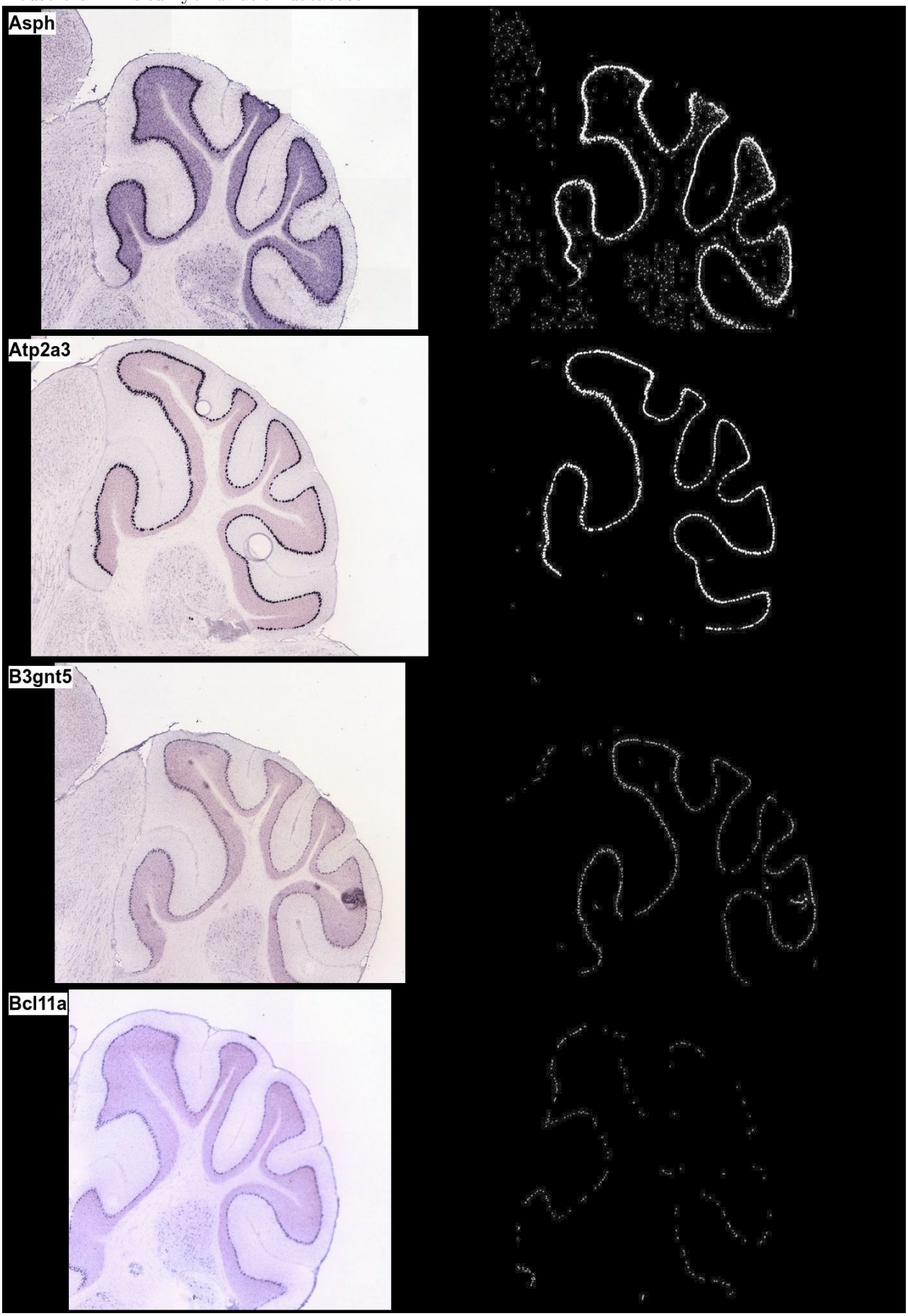
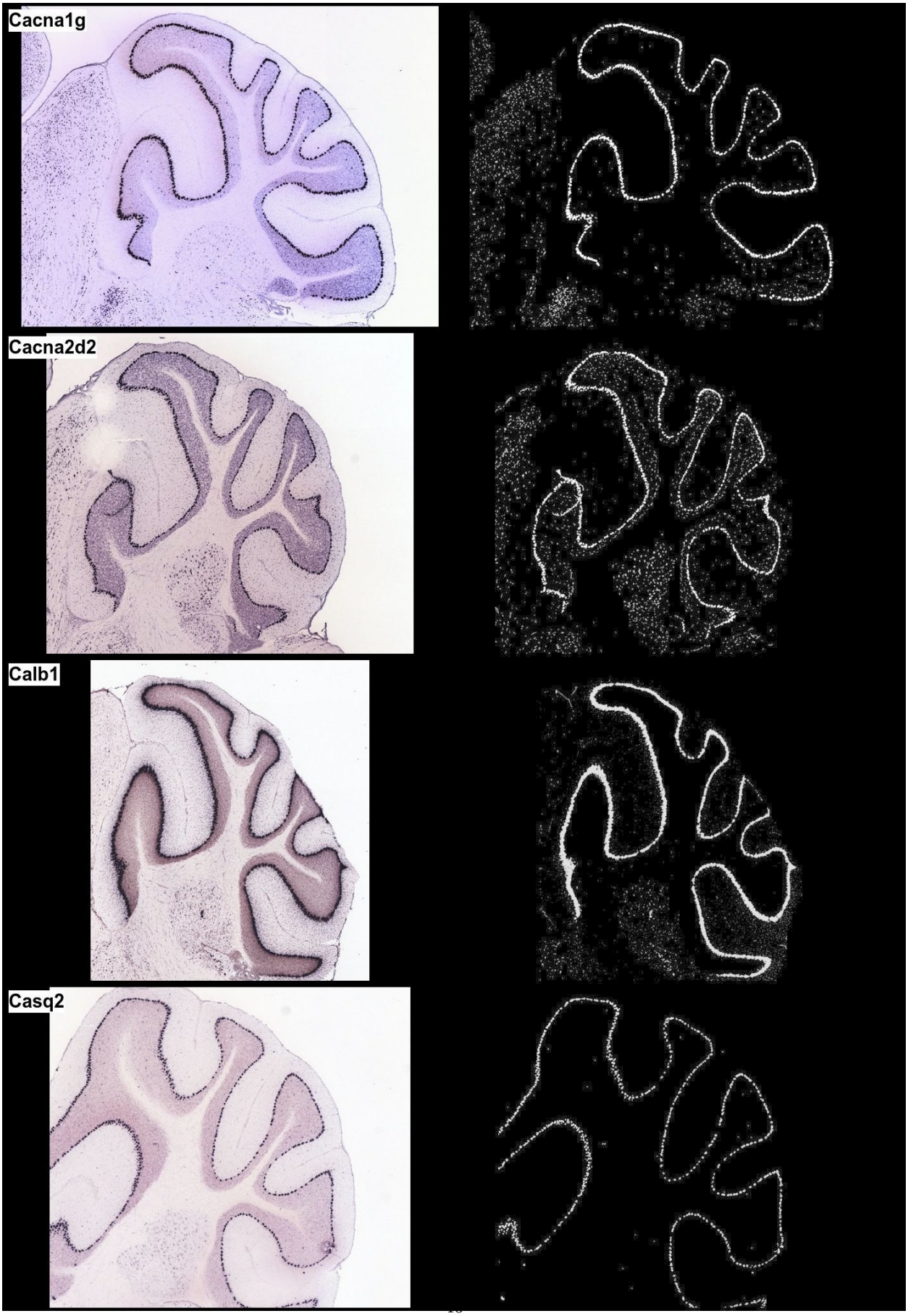
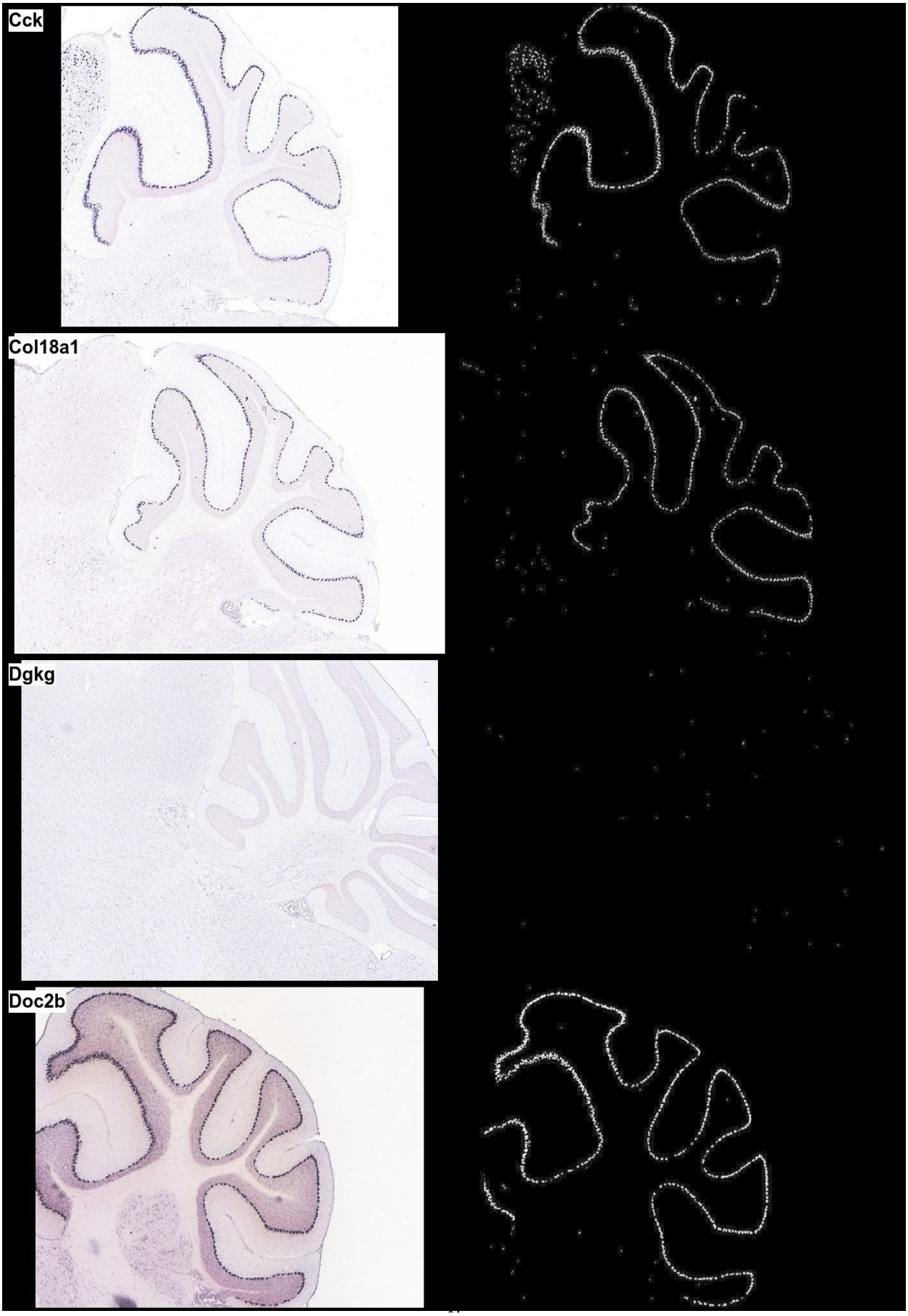


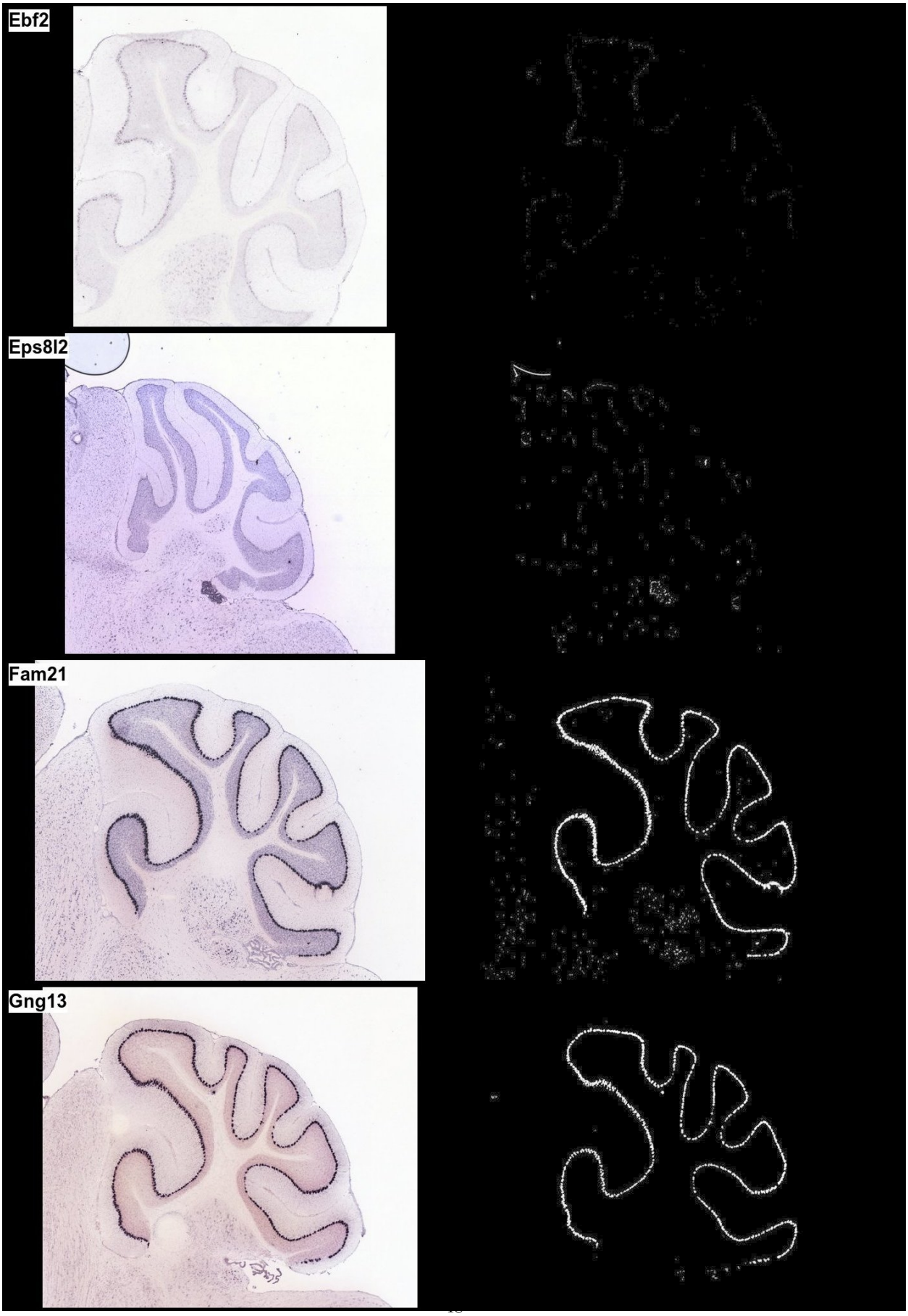
Figure 15: Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. Our estimations are much worse for these cells, in the case of memory B cells there is strong negative correlation and we failed to detect enough genes to make an estimation for resting memory T cells.

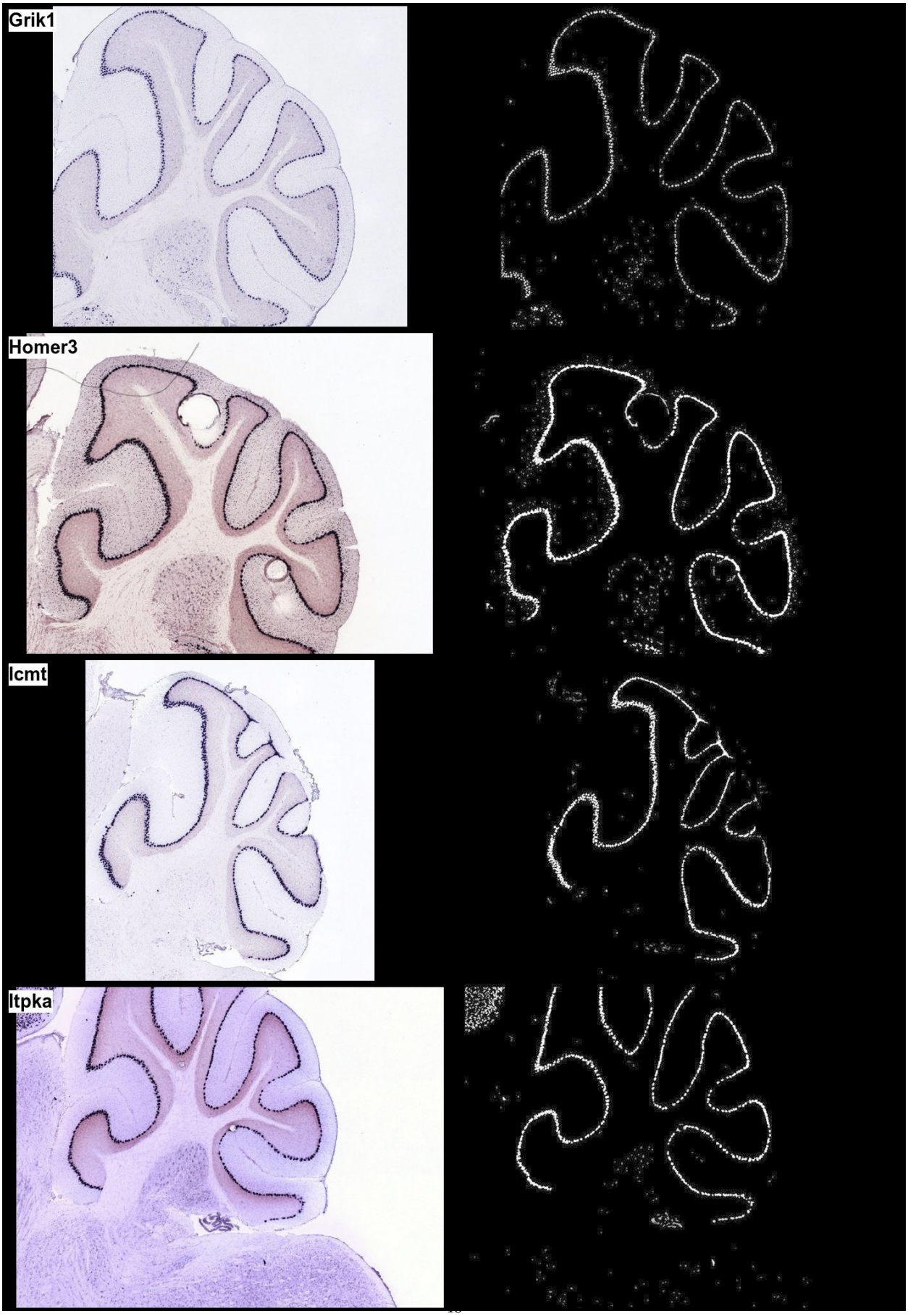
Figure 16: Expression of purkinje markers discovered in the study in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database.

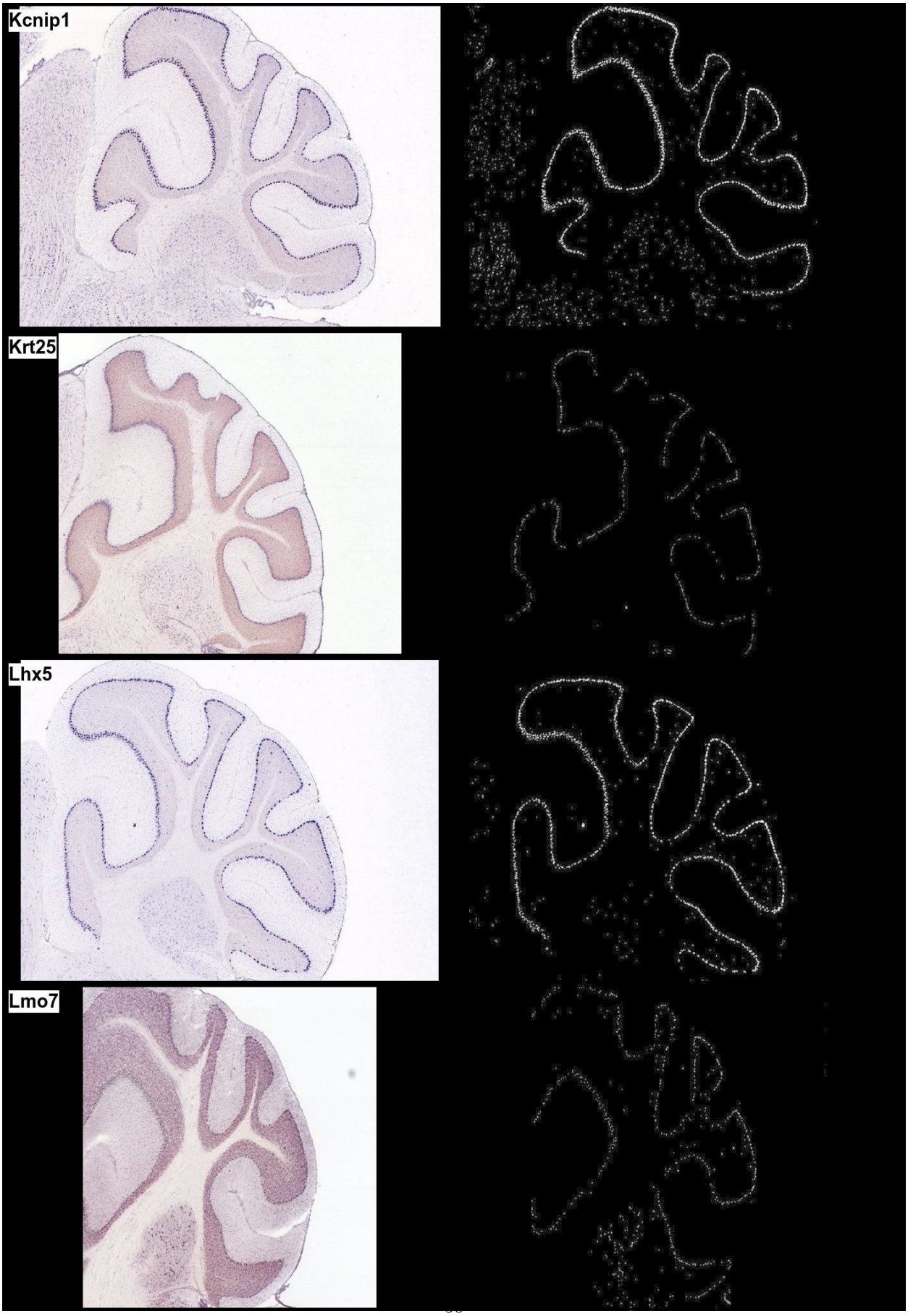


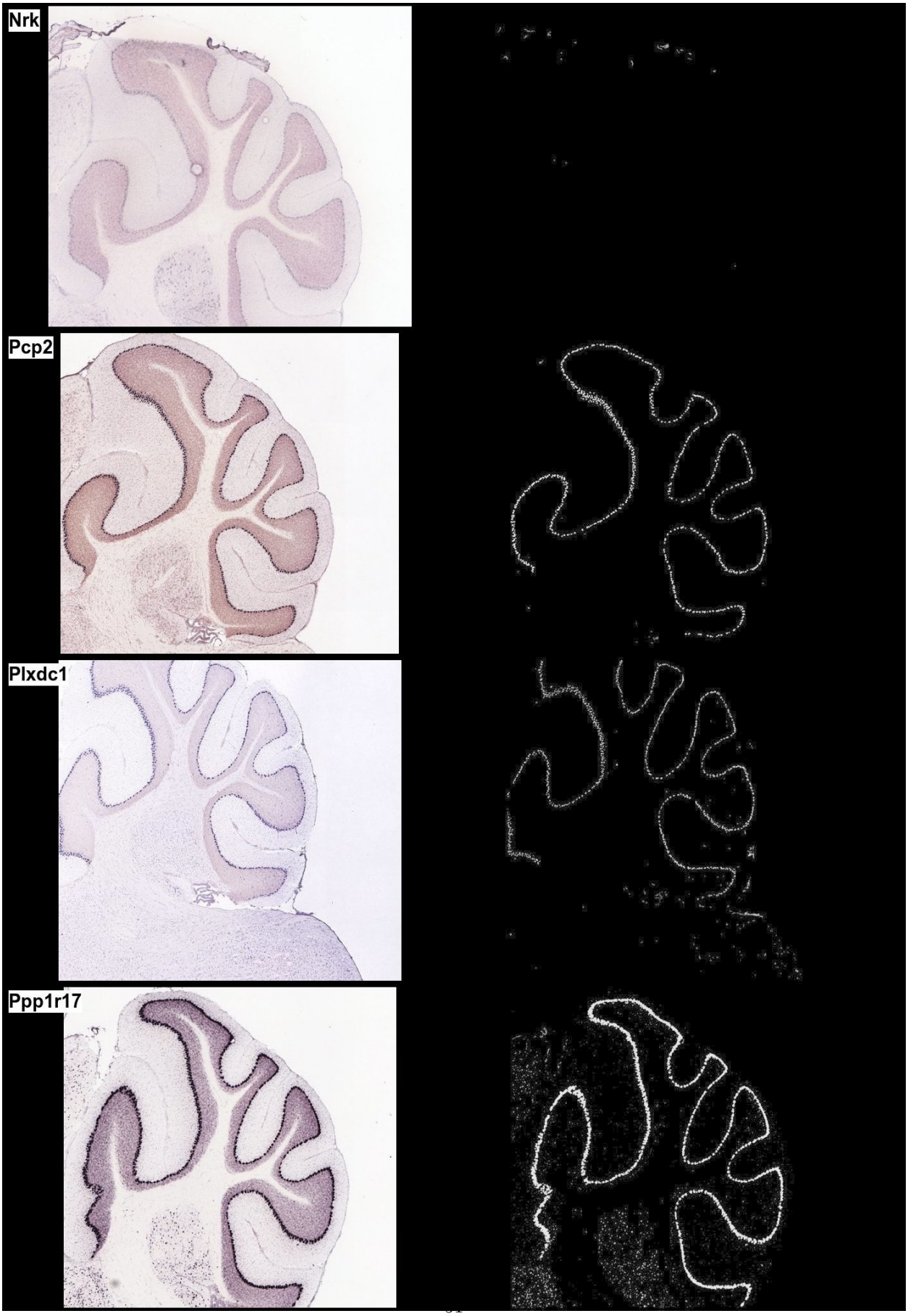

Cacna1g
Cacna2d2
Calb1
Casq2

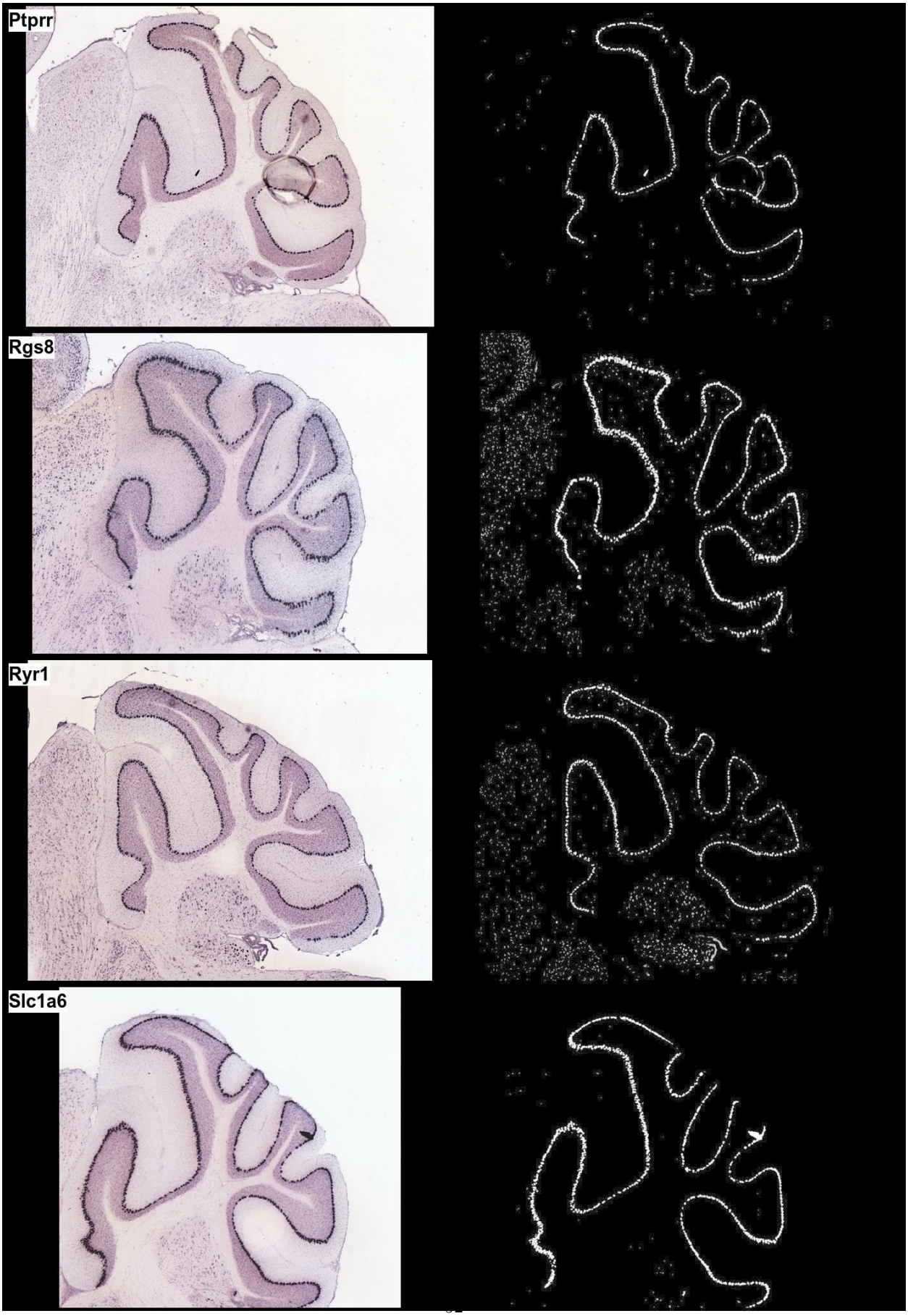

Cck
Col18a1
Dgkg
Doc2b

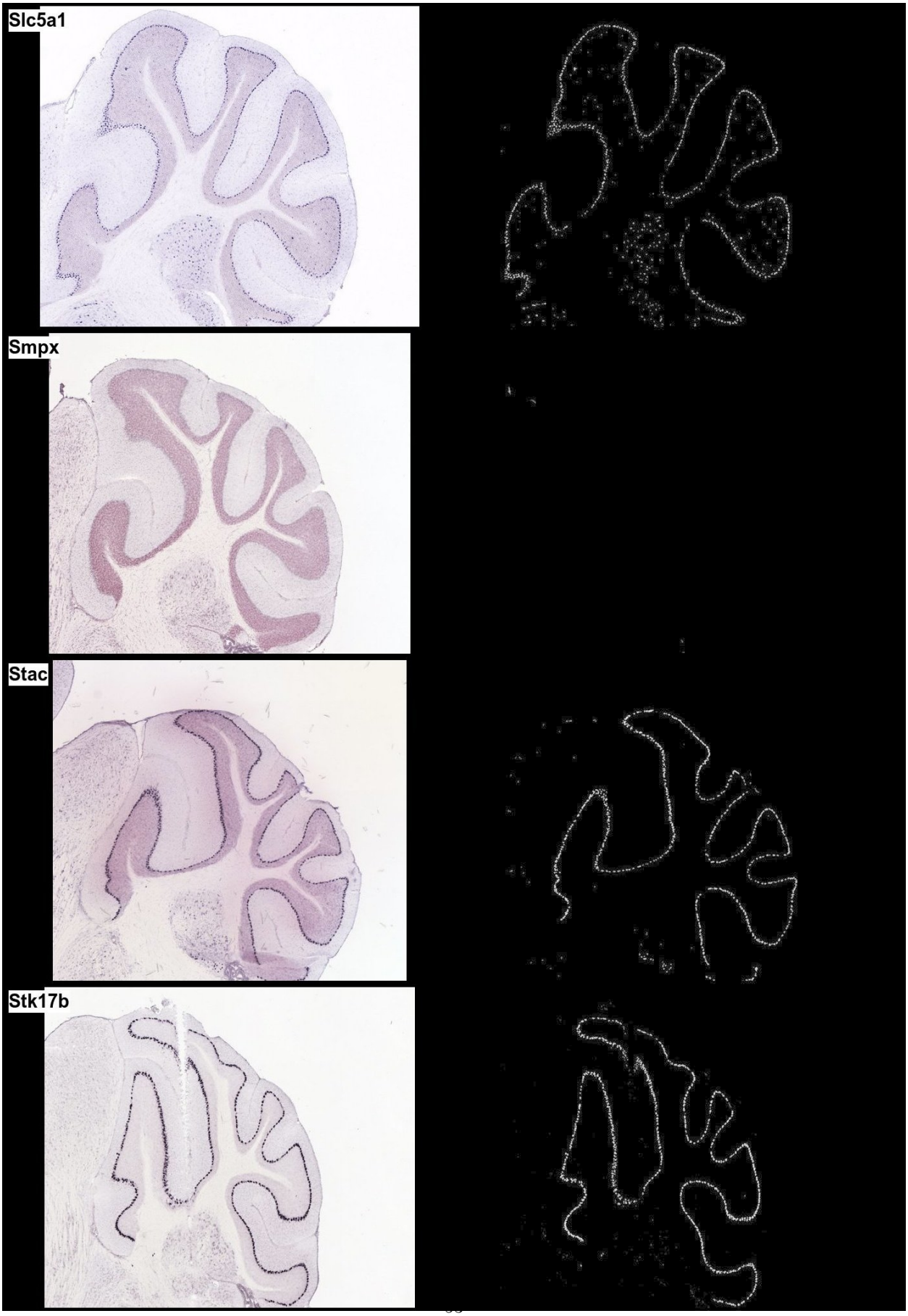

Ebf2
Eps8l2
Fam21
Gng13


Grik1
Homer3
Icmt
Itpka


Kcnip1
Krt25
Lhx5
Lmo7


Nrk
Pcp2
Plxdc1
Ppp1r17


Ptprr
Rgs8
Ryr1
Slc1a6


Slc5a1
Smpx
Stac
Stk17b

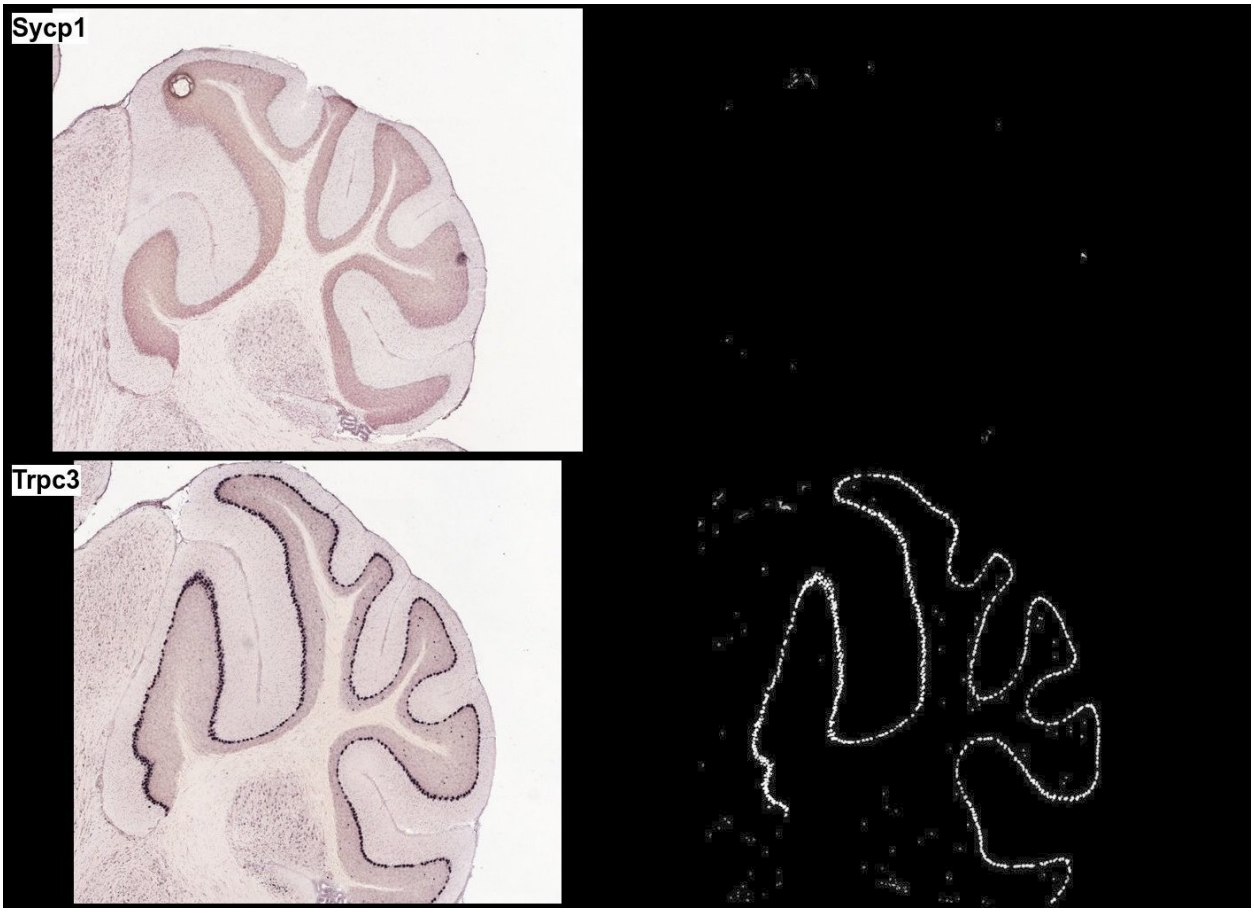
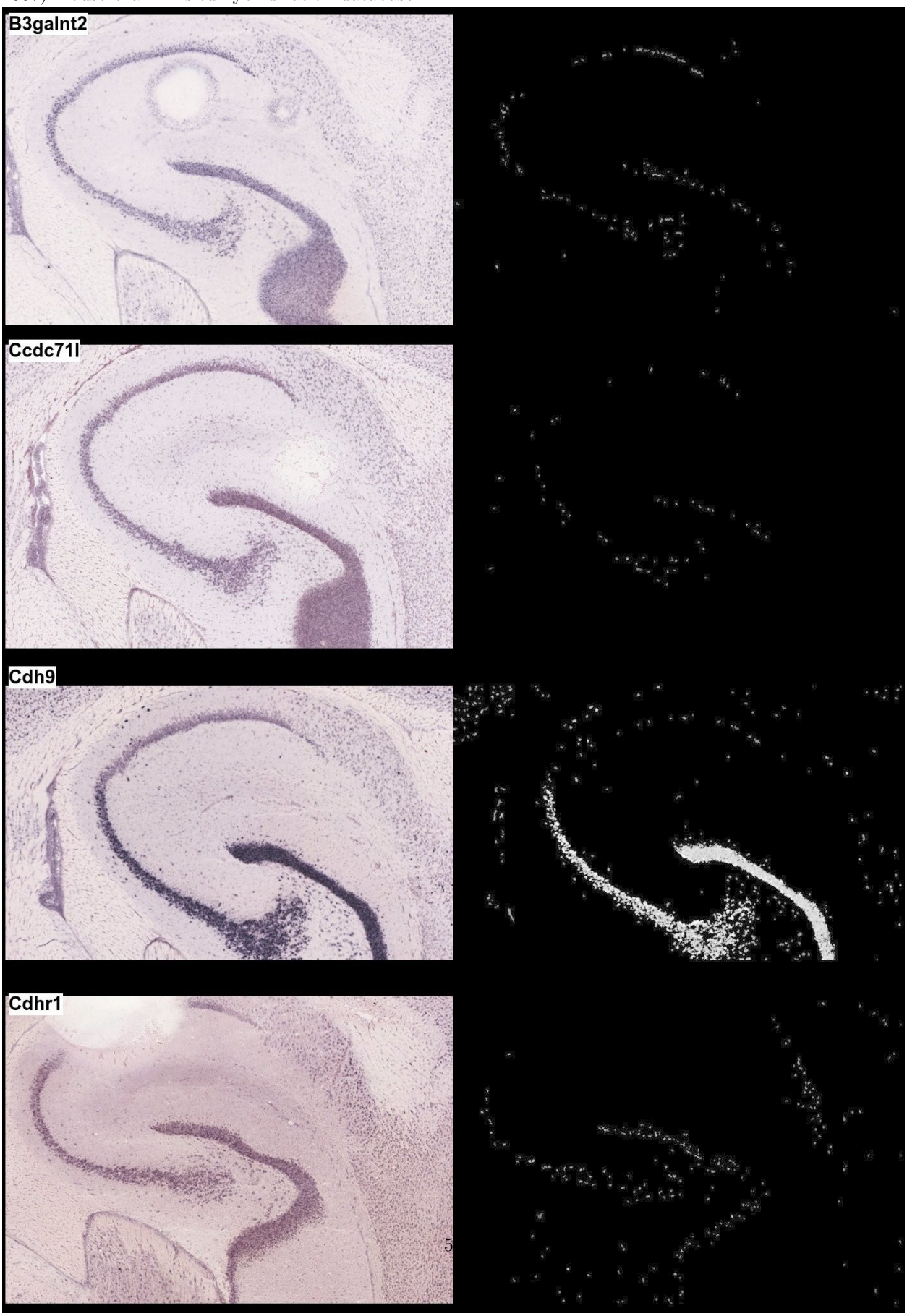
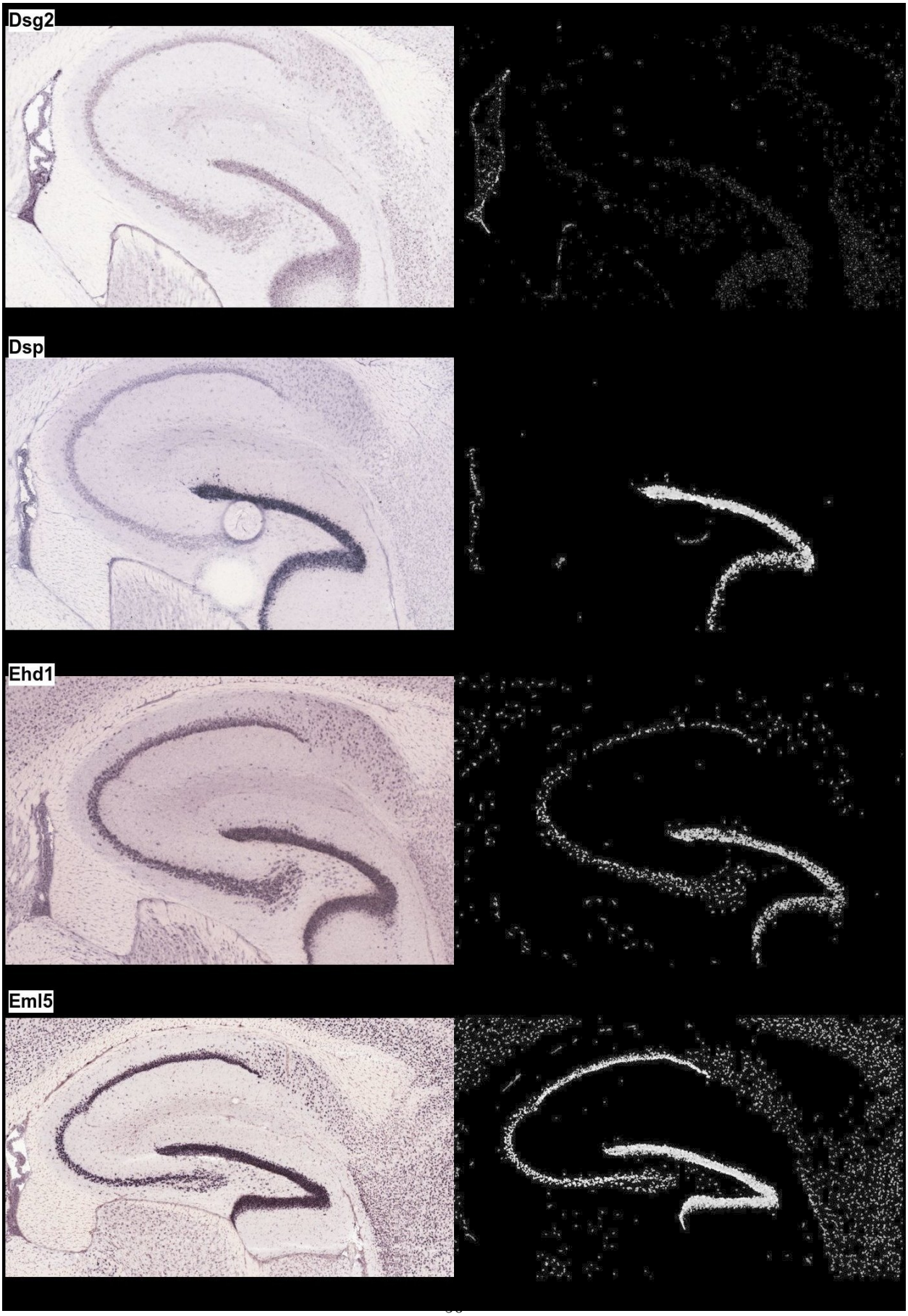
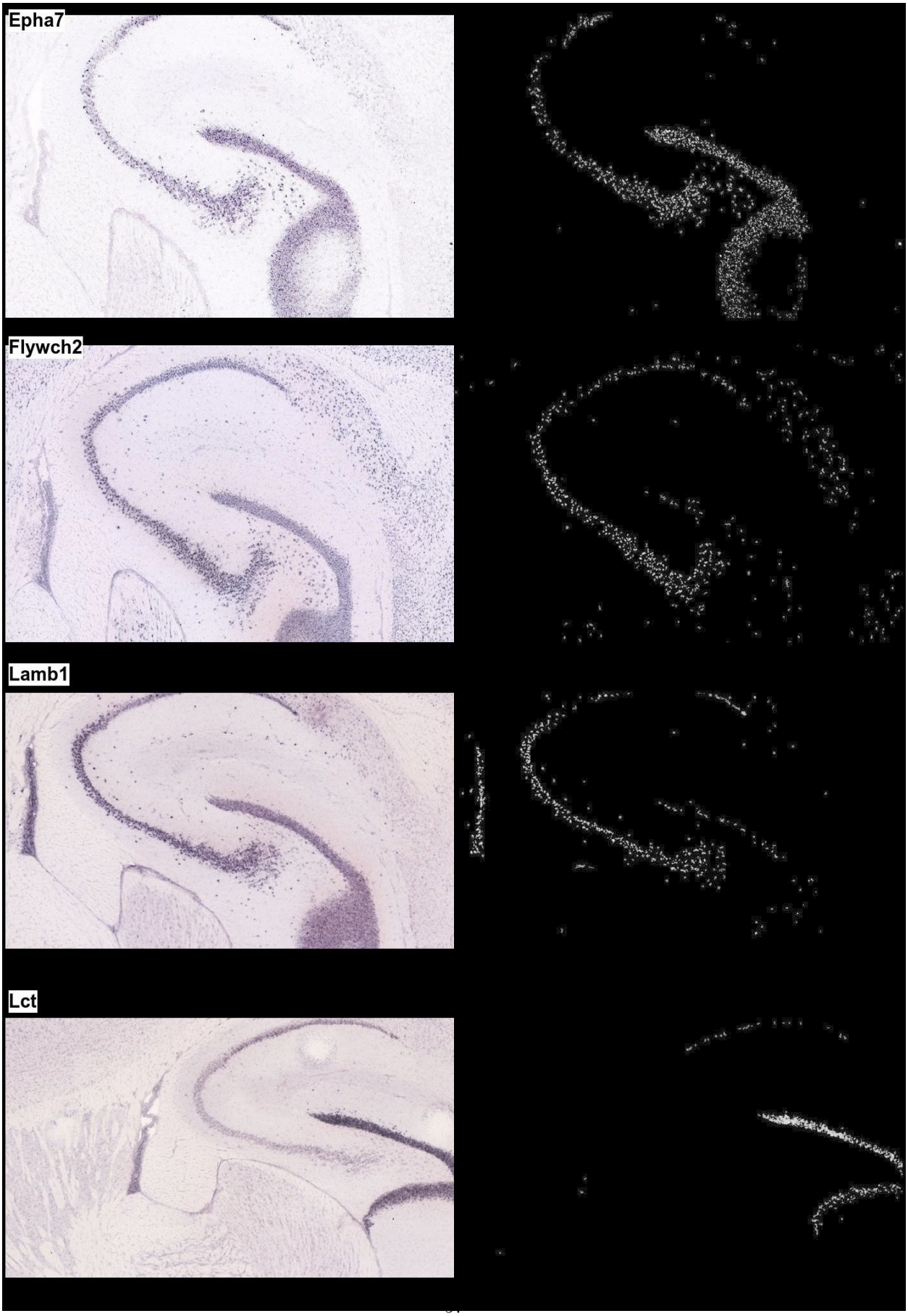
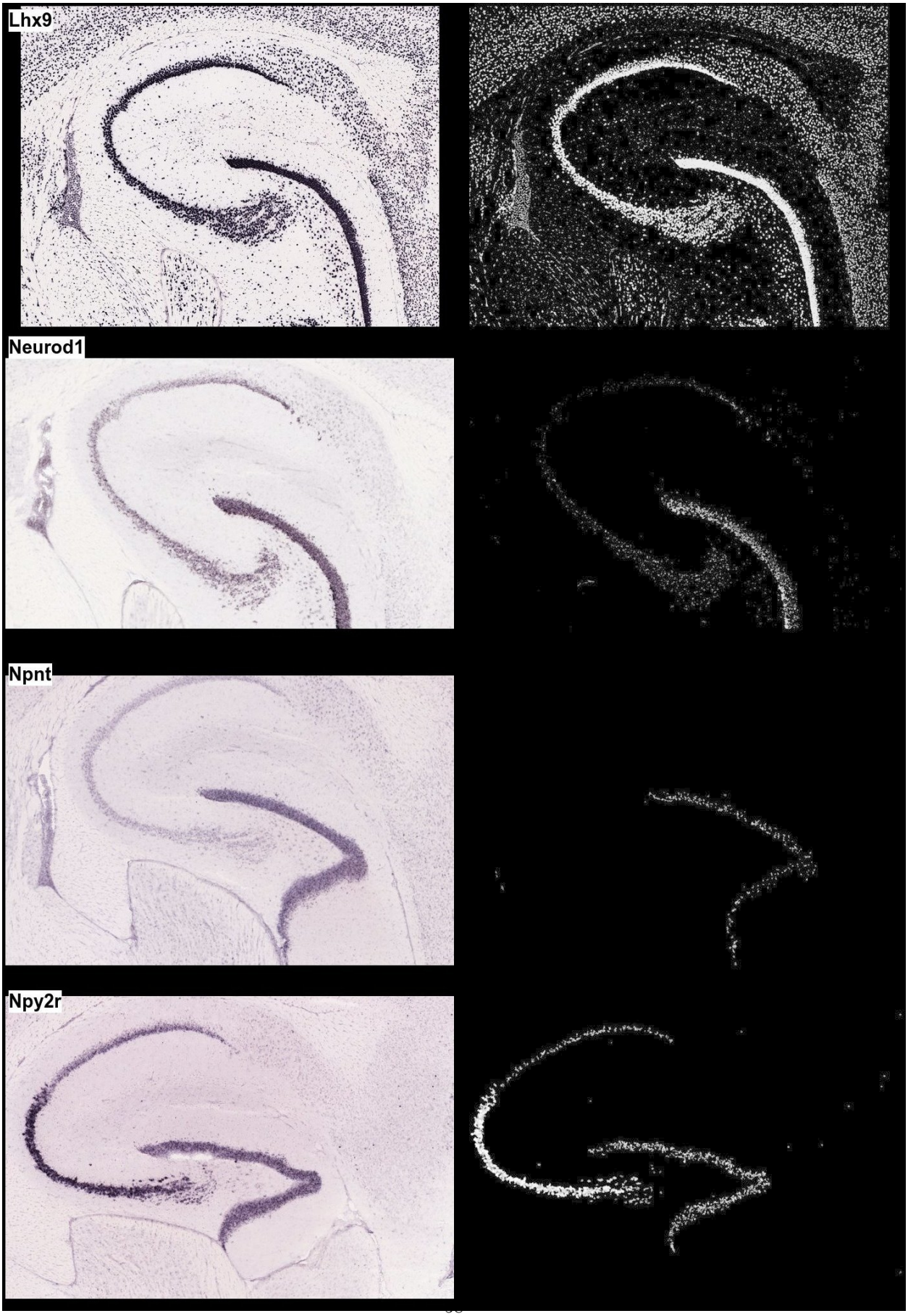

Sycp1
Trpc3

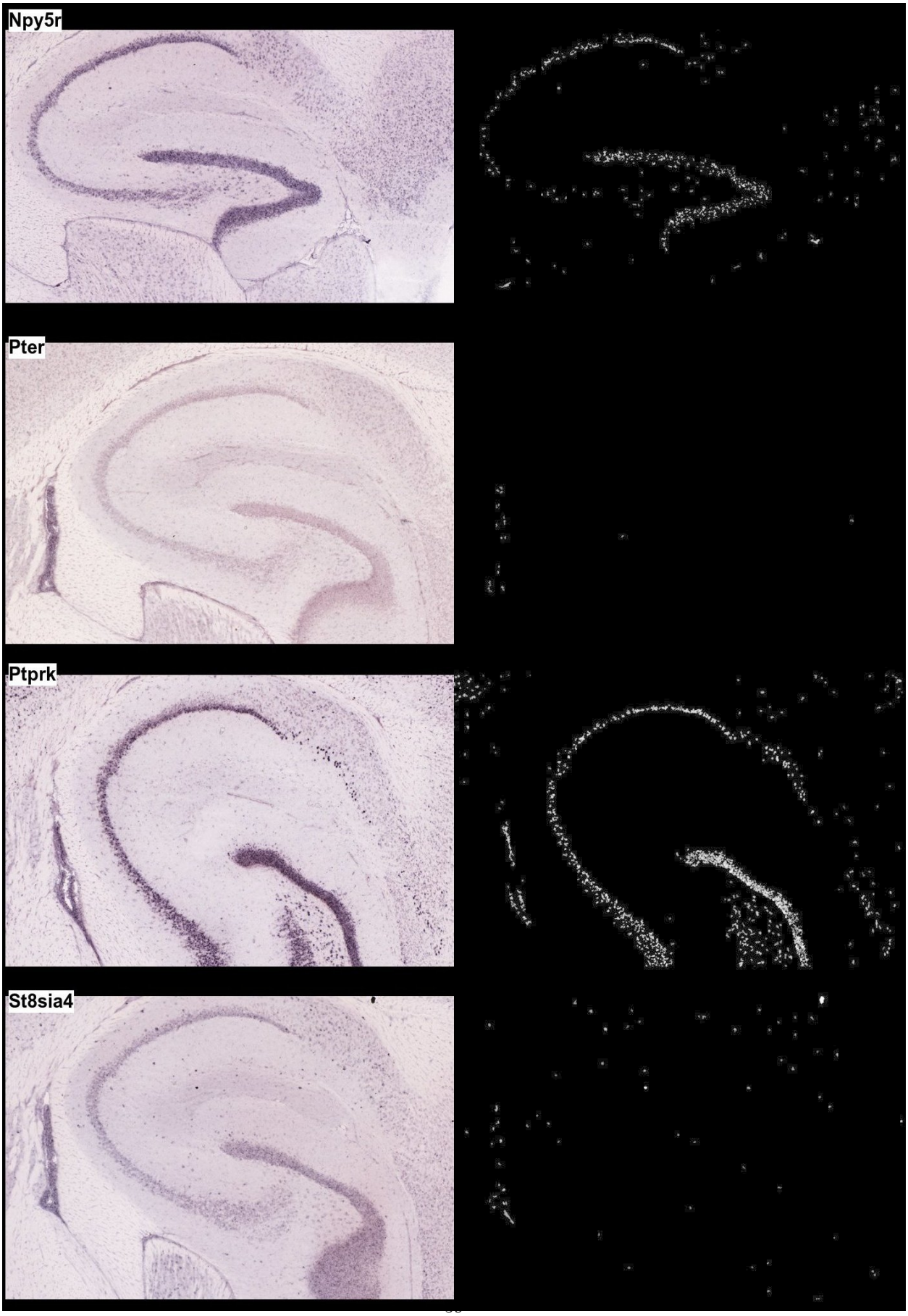
Figure 17: Expression of dentate granule cell markers discovered in the study in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database.

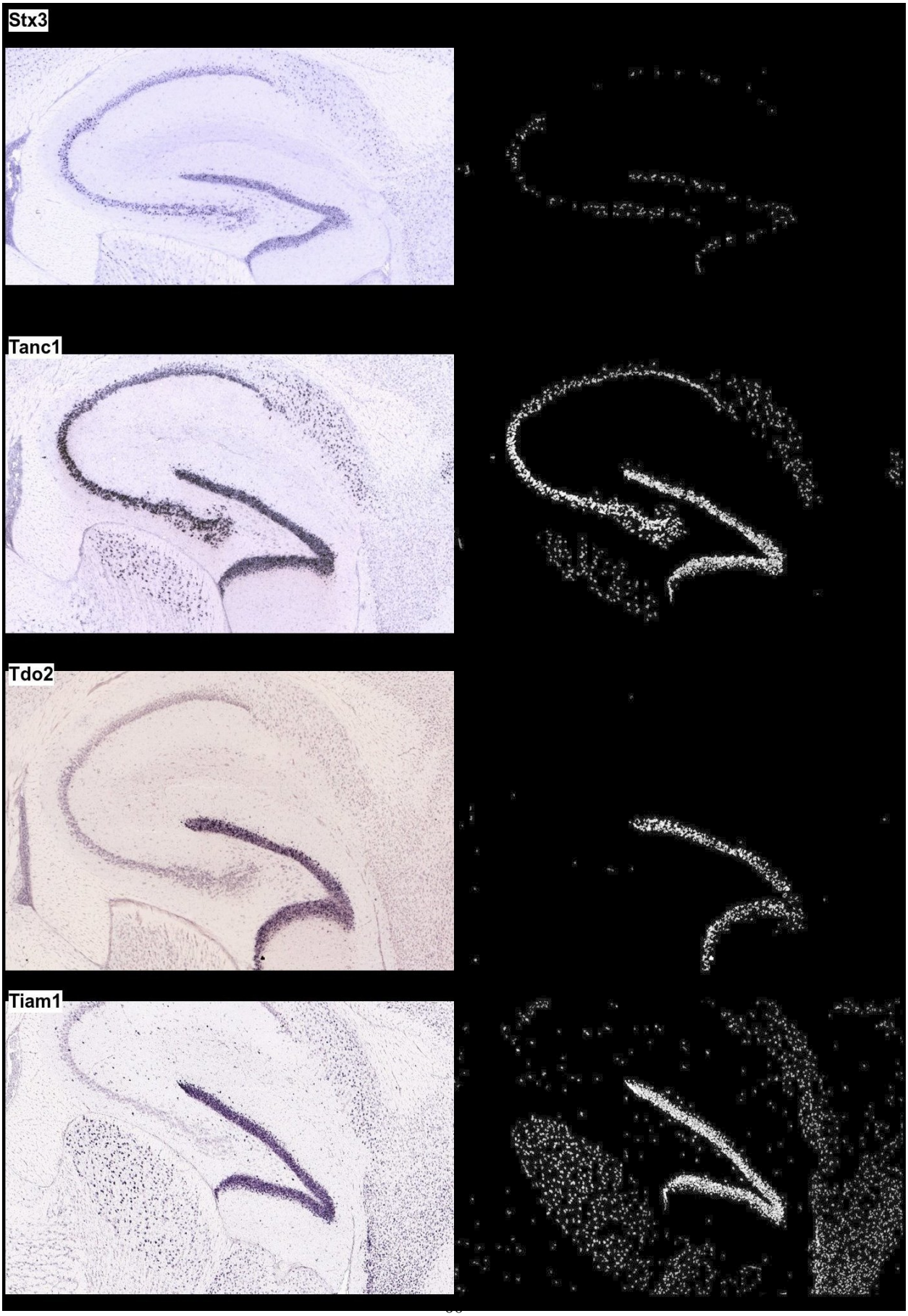


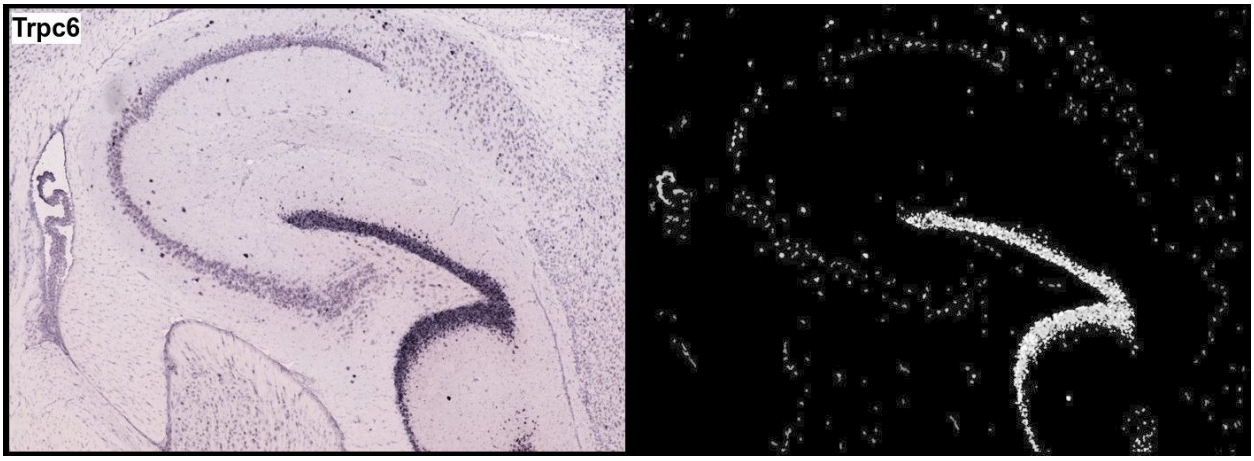

Dsg2
Dsp
Ehd1
Eml5


Epha7
Flywch2
Lamb1
Lct


Lhx9
Neurod1
Npnt
Npy2r


Npy5r
Pter
Ptprk
St8sia4


Stx3
Tanc1
Tdo2
Tiam1


Trpc6