# Identification cell type marker genes of the brain and their use in identification 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**

???

**Examination Date**

June 19, 2015

# Contents

# List of Figures

# List of Tables

# 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 this approach remains popular due to its relative ease and low cost, 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]. The aim of this project is to address this problem by estimating cell type proportions in whole tissue studies.

Studies which do use single cell types are typically examining only one type at a time to compare conditions such as disease[5,6] or 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]. However such studies does not address several critical points. **1)** What are the marker genes of a cell-type in the scope of the entire brain region they reside in? Not all cell-types in the region of interest are examined by these studies. Discovery of new markers are important since not all neuron types have known unique markers, and many have few numbers of markers. This complicates analysis 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 if 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. **2)** Are there changes in type proportions which is a common occurrence in neurological diseases such as Parkinson's and Alzheimer's? **3**) which gene expression changes occur in which cell-types?

Addressing these issues by common laboratory methods s expensive and labor intensive. Finding definitive marker genes requires isolation of all known cell-types from a brain region; detecting cell-type specific differences of all cells types would require all of them to be labelled and counted; testing for expression changes in all cell-types require isolation and expression profiling of all cell-types.

For this work, we are forming a comprehensive database of cell-type-specific expression profiles from a variate of resources to help with the problems described above. This dataset, along with in siloco and in vitro validation methods, will be used to detect marker genes that best represent a particular cell-type in it's associated brain region. We are hoping this list of cell-type specific marker genes will be useful to scientific community. To make further use of the discovered marker genes we will consider their expression levels

in whole tissue samples as surrogates for their abundance in the sample. This will allow us to understand the fate of cell-type populations under specific diseases and conditions. Finally, we 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 for 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[12,13] 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 a comprehensive dataset 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 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 well 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 cell-types of the brain[12,13]. 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 genes' expression in whole tissue samples as a basis, enumerate the relative amounts of particular cell-types in these samples in 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 enumeration method
4. Use an independent dataset of isolated blood cell-types and manually enumerated blood samples to repeat and validate the enumeration method.
5. Use enumeration 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 Major cell-types of the brain

The brain is composed of a variety of cell-types and is one of the most heterogeneous organs in the mammalian body. Alongside the major groups of cell-types associated with it: neurons and glia, it also contains blood and endothelial cells. The major neuron and glia types that are included in our cell-type specific expression database are described briefly below

### 3.1.1 Glia

- **Astrocytes** are star shaped cells present throughout the brain[14]. They have roles in preserving chemical balance in synapses, regulating blood flow by controlling vessel diameters and by providing support to endothelial cells forming the blood brain barrier[14]. Relatively recently it was also discovered that they play roles in signal transduction by regulating intracellular ion concentrations and releasing gliaotransmitters[15]. Astrocytes often proliferate in diseased brain[14].
- **Oligodendrocytes** are responsible for forming the mylein sheet around neurons to insulate their axons[16]. A single oligodendrocyte ensheats multiple neuronal cells as a result of a highly coordinated

process affected by axon size, neuronal activity and molecular signalling[16]. Most myelination occurs early in the differentiation process. Alongside myelination, oligodendrocytes also have neuroprotective functions and are known to dysfunction in several neurodegenerative diseases[17].

- **Microglia** are resident macrophages of the brain[18]. They clear apaptotic cells and are involved in maintenance of synapses[18], and are the antigen presenting cells of the brain[???]. In most neurological diseases, they activate and proliferate[18], and their proliferation is associated with degradation of neuronal cells[???].

### 3.1.2 Neurons

**Pyramidal Cells** are prominent in areas of the brain associated with high cognitive functions such as the cerebral cortex and hippocampus[19]. They are recognized by their triangular cell body and short basal dendrites[19]. Pyramidal cells of different brain regions and layers have differences in structure and gene expression[19]. They can receive inhibitory GABAergic inputs through soma and axon, while excitatory signals are delivered through the dendrites[19]. Dendrites of pyramidal cells are covered with spines that act as synapse sites for glutamatergic synapses[19].

**Cortical GABAergic neurons** are local inhibitory neurons that control pyramidal cell firing and generation of cortical rhythms. In the cortex, three major groups are defined: PV expressing, SST expressing and 5HT3a expressing[20]. These subtypes of gabaergic cells localize into specific layers of the neocortex[20]. PV expressing neurons are fast spiking cells with low input resistance[20]. They are thought to be the dominant inhibitory system in the cortex. SST positive cells often receive facilitating signals from pyramidal neurons to provide inhibitory feedback[20]. Finally 5HT3aR interneurons are a more heterogeneous subgroup of interneurons which differ in function and morphology[20].

**Midbrain dopaminergic neurons** are primarily thought to regulate motor functions, emotion and reward[21]. During development of the midbrain they separate into three distinct clusters of cells[21]. Cells making up the A9 cluster, which forms substantia nigra are lost during the progression of Parkinson's disease[21].

**Purkinje cells** are GABAergic cells local to cerebellum's purkinje layer[22]. They receive signals from nearby granule cells and send inhibitory signals to other purkinje cells towards deep cerebellar nuclei[22]. They have functions in motor learning[22].

**Cholinergic cells** are involved in memory formation[23]. In particular, they are associated with Alzheimer's Disease[23]. They are also responsible for maintaining the circadian rhythm, by a high acetylcholine release during awake periods[24].

## 3.2 Cell-type markers and their applications

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[25] with roles in myelination and Aif1, a microglia marker with roles in inflammation[26] for glial cells; Th that is responsible for dopamine synthesis in dopaminergic cells[21], and Gad1/2 that catalyzes production of GABA in gabaergic cells[20] for neurons. Many specific neurons, however, do not have known markers. For instance while Gad1/2 are a useful marker of gabaergic neurons, subtypes of gabaergic neurons lack known specific markers which makes accessing them more challenging.

## 3.3 Cell-type isolation

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)[27]. 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)[28]. A relatively recent marker based isolation is Translating Ribosome Affinity Purification[29]. This method combines the promoter of a marker gene with the coding region of L10a ribosomal subunit fused with green fluorescent protein (GFP)[29]. 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)[29,30]. 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)[31]. The resulting samples from each of these methods have varying purity[32] which is a potential confound on studies that use samples acquired by multiple methods.

## 3.4 Cell-type deconvolution

Expression levels obtained from whole tissues contain signals from multiple cell-types. 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, deconvolution methods are commonly applied to blood data due to ease of access to both mixed samples and isolated cell-types[4,33,34].

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 deconvolute human brains estimated proportions of neurons as a single group alongside astrocytes, oligodendrocytes and microglia[35]. 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[36]. 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 (see next subsection) of 64 distinct cell-types was performed on whole tissue expression profiles across various brain regions[37]. In this study, 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[37]. 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)[37]. 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[34].

### 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, usage of 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[4,32,33,36]. 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 principle component of the genes in the whole tissue samples is used as a surrogate[4,36,38]. 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

### 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[39]. Microarrays are built by fixing single stranded probes, that are specific to a location on a target genome, at a known location 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[40].

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 go to tool for expression analysis. A wide array of data is available at both in the tissue[1,41,42] 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[43]. In general, RNA-seq is more prone to technical artifacts due to the stochasticity of the sequencing process. This effect is particularly powerful for genes with lower expression levels which often make up the majority of the data[44].

Recently, RNA sequencing of single cells is becoming increasingly popular[45]. 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 artifacts resulting from amplification and sequencing are more prominent[45].

RNA-seq analysis, especially single cell studies are starting to gain popularity in neuroscience[12,46]. 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 and make it 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 dataset is collected from the Gene Expression Omnibus (GEO) and through personal communications. We also make the data available via a web application that allows easy browsing of the data. Mouse cell-type specific data 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.[32] 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 only. 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[47,48]. 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[49]. 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 Okaty dataset 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 also lead to removal of some samples whose associated subtypes were already represented by another sample. For instance samples representing Htr3a positive GABAergic cells removed due to the presence of, VIP positive cells (their subtypes[20]) in another study. When cell-types wee 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 (Table 1 - 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 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 principle use of the comprehensive database we created is to find gene sets that show highly enriched expression in single cell-types. 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. To accomplish, this we separated samples into regions based on the metadata acquired from the original source. We generalized certain regions to improve the biological relevance of the marker gene selection. For instance, the brainstem is taken as a single region while there are samples taken from specific regions of brainstem such as midbrain, either the definitions of exact origins of cell-types were not clear, or attempting to find marker genes in that subregion would not be useful due to lack of other cell-types. Oligodendrocyte and astrocyte samples isolated from cortex were added to other regions from cerebrum since these cell-types are known to be prevalent across the brain.

## 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.
- 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.

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 4).

## 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 ensure 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)[50] 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. Granule cells of dentate gyrus and purkinje cells of cerebellum fit this criteria, so we were able to confirm some of the marker genes through ABA 5[50] .

We are currently collaborating with Dr. Etienne Sibille 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 6). 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 infeasible to apply biological validation methods to all marker genes that we selected, we are using recently published single cell RNA-sequencing datasets[12,13,46] 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 surmised 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[12] and a human[46] 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, and in the human dataset a majority (7/10) of the marker gene sets showed significantly more coexpression.

The next step is to repeat this analysis using a more recently published RNA-seq study performed on mouse brains, that isolates single cells in a more cell-type specific manner[13]. This dataset is of higher quality and 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 a human dataset containing sixteen brain regions isolated from pathologically healthy subjects[51]. 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 given cell-type, expression of the marker genes for that cell-type will be all higher. This is expected to result in increased coexpression of marker genes in whole tissue datasets since all samples will have some variability in cell-type proportions. We compared the overall level of marker gene coexpression between these samples to coexpression levels of randomly selected genes. The results showed heightened coexpression for the majority of the cell-type marker sets (11/15) (Figure 8).

### 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. 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 dataset. Due to the inherent differences in the data structure, assessing repeatability of such single cell studies is not straightforward. 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 expected analysis can be found in Figure 9.

## 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. Marker genes, on the other hand, are less sensitive to fine changes in expression in the reference datasets due to our stringent selection criteria. Therefore it is sensible to assume that sufficiently many marker genes will be preserved across species. 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[4,33,36]. 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[51] we observed an increase in glial population and decrease in most neuronal populations between white matter and grey matter (Figure 10), and figuring that purkinje cells were exclusive to the cerebellum (Figure 11). To assess the usefulness of the method in the context of neurological diseases, we acquired a dataset of substantia nigra expression from healthy donors and Parkinson's disease patients[52], characterized by loss of dopaminergic cells in the region. Our analysis was able to show a marked decrease in dopaminergic cells in Parkinson's disease patients (Figure 12).

We will continue to analyze more datasets from neurological diseases 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[53] for expression profiles with well-annotated metadata to facilitate fast analysis. These datasets include brains under many conditions, such as neurological diseases, disease models, and gene knock-downs, etc. We expect this analysis will 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 with 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 asses 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[54,55]. 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 3) and human blood cell-types[34]. 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 13). 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[34]. 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 14), 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 15 - 16). Shay et al.[56] 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. 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 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.

13 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.

14 Sofroniew MV, Vinters HV. Astrocytes: Biology and pathology. *Acta Neuropathologica* 2010; **119**: 7–35.

15 Jahn HM, Scheller A, Kirchhoff F. Genetic control of astrocyte function in neural circuits. *Frontiers in Cellular Neuroscience* 2015; **9**. doi:10.3389/fncel.2015.00310.

16 Bradl M, Lassmann H. Oligodendrocytes: Biology and pathology. *Acta Neuropathologica* 2010; **119**: 37–53.

17 Bankston AN, Mandler MD, Feng Y. Oligodendroglia and neurotrophic factors in neurodegeneration. *Neuroscience bulletin* 2013; **29**: 216–228.

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

19 Spruston N. Pyramidal neurons: Dendritic structure and synaptic integration. *Nature Reviews Neuroscience* 2008; **9**: 206–221.

20 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.

21 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.

22 Ito M. Historical Review of the Significance of the Cerebellum and the Role of Purkinje Cells in Motor

Learning. *Annals of the New York Academy of Sciences* 2002; **978**: 273–288.

23 Baxter MG, Bucci DJ, Gorman LK, Wiley RG, Gallagher M. Selective immunotoxic lesions of basal forebrain cholinergic cells: Effects on learning and memory in rats. *Behavioral Neuroscience* 2013; **127**: 619–627.

24 Hut RA, Van der Zee EA. The cholinergic system, circadian rhythmicity, and time memory. *Behavioural Brain Research* 2011; **221**: 466–480.

25 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.

26 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.

27 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.

28 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.

29 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.

30 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.

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

32 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.

33 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.

34 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.

35 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.

36 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.

37 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.

38 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.

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

40 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.

41 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.

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

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

44 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.

45 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.

46 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.

47 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.

48 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.

49 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.

50 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.

51 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.

52 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.

53 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.

54 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.

55 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.

56 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.

| 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 3: Sources used in generation of the mouse blood cell type database.

# 8 Figures

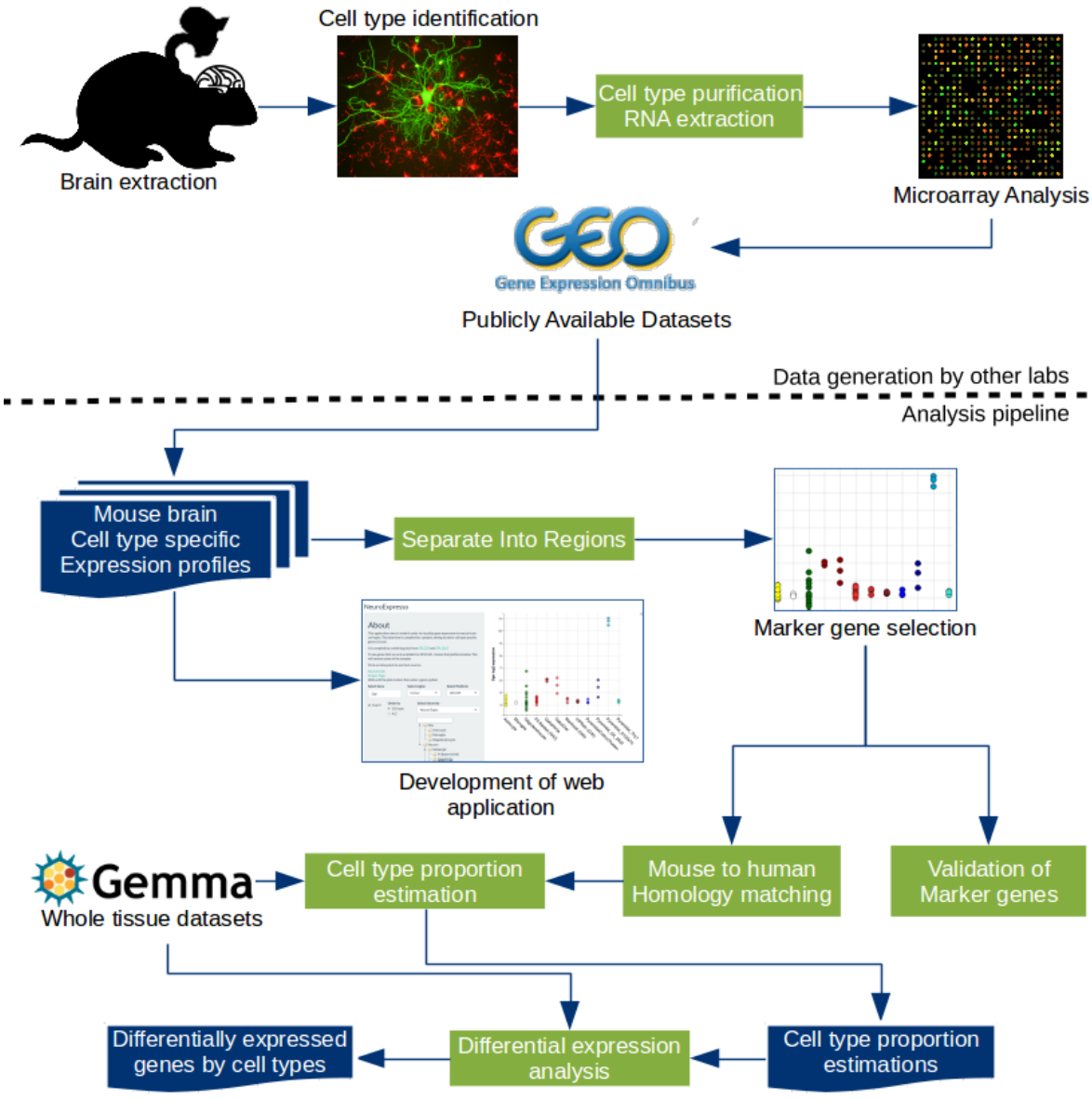


Figure 1: Workflow of the project

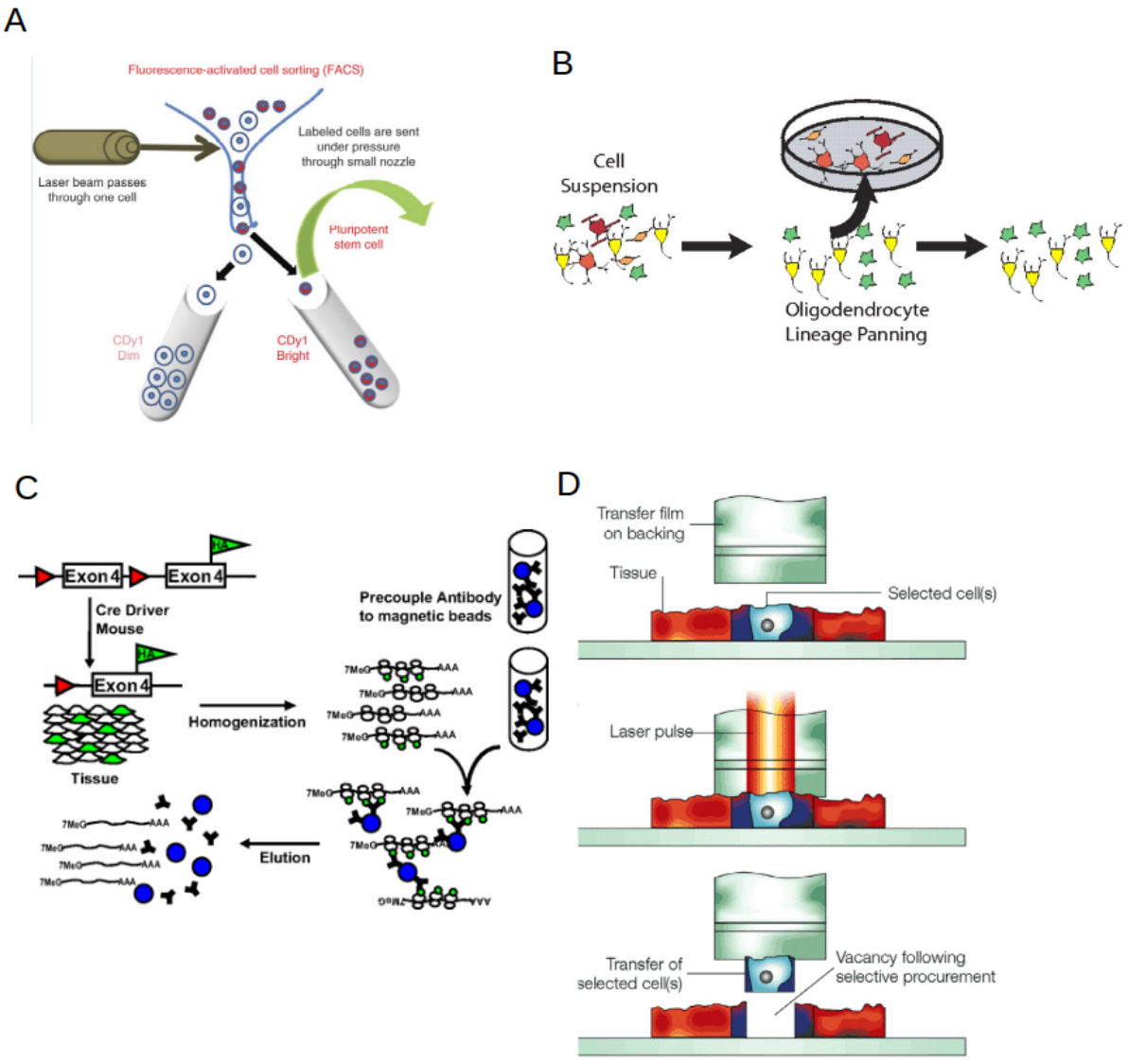


Figure 2: Example representations of cell type isolation techniques. A. Adapted from Kang et al. 2011. An example application of FACS. A fluorescently active molecule is used to label specific cell in the population. Upon detection of fluorescence, a charge is placed on the droplet whose path is later 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.

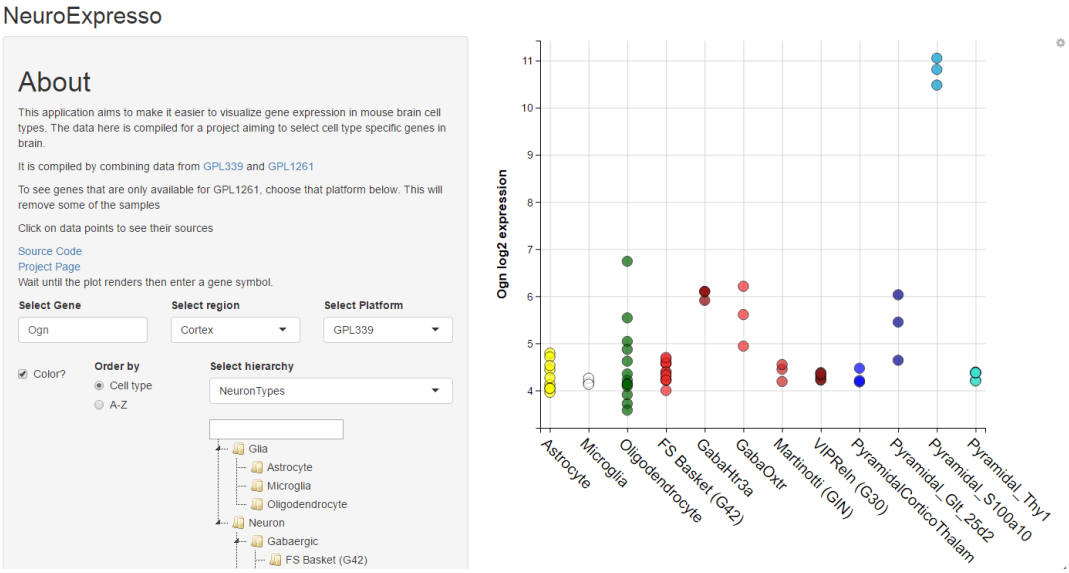


Figure 3: A screenshot of the NeuroExpresso web application: A tool to visualize gene expression in the cell types of our database.

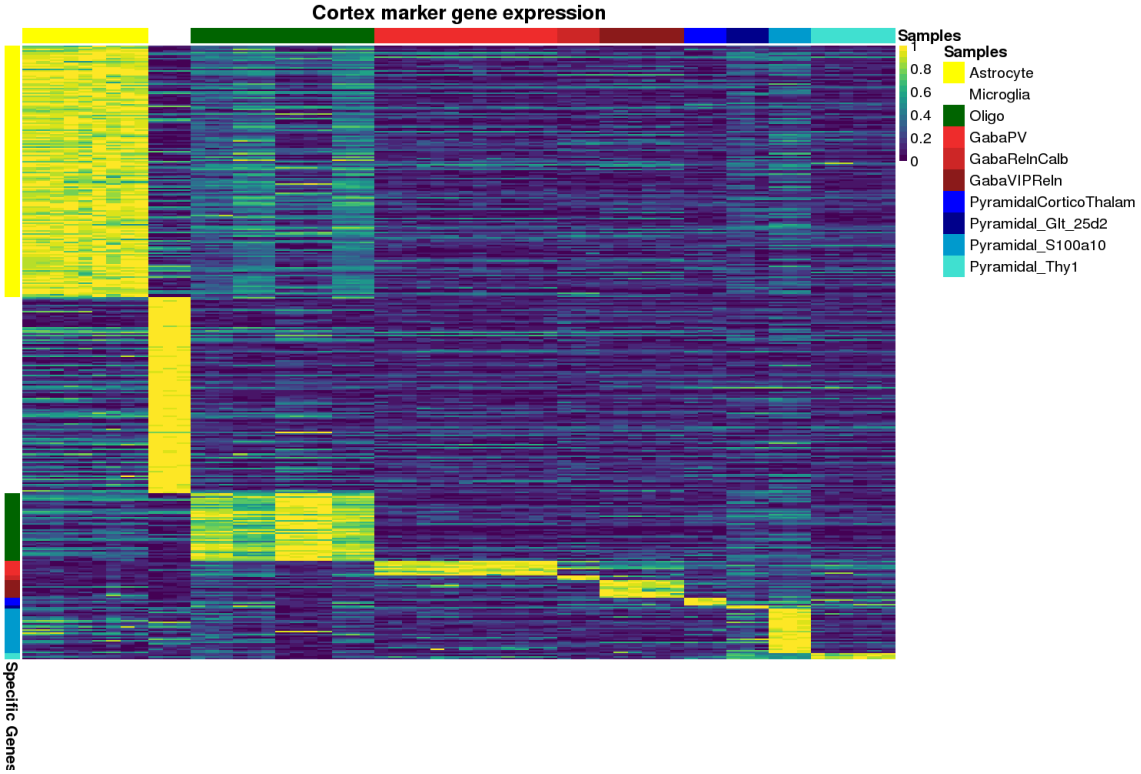


Figure 4: Expression of marker genes detected from cortex cell types. Values are scaled to be between 0 and 1, 0 representing the lowest observed expression level for the gene while 1 representing the highest. Samples and genes follow the same order of cell types to emphasize the specificity of the selected genes.

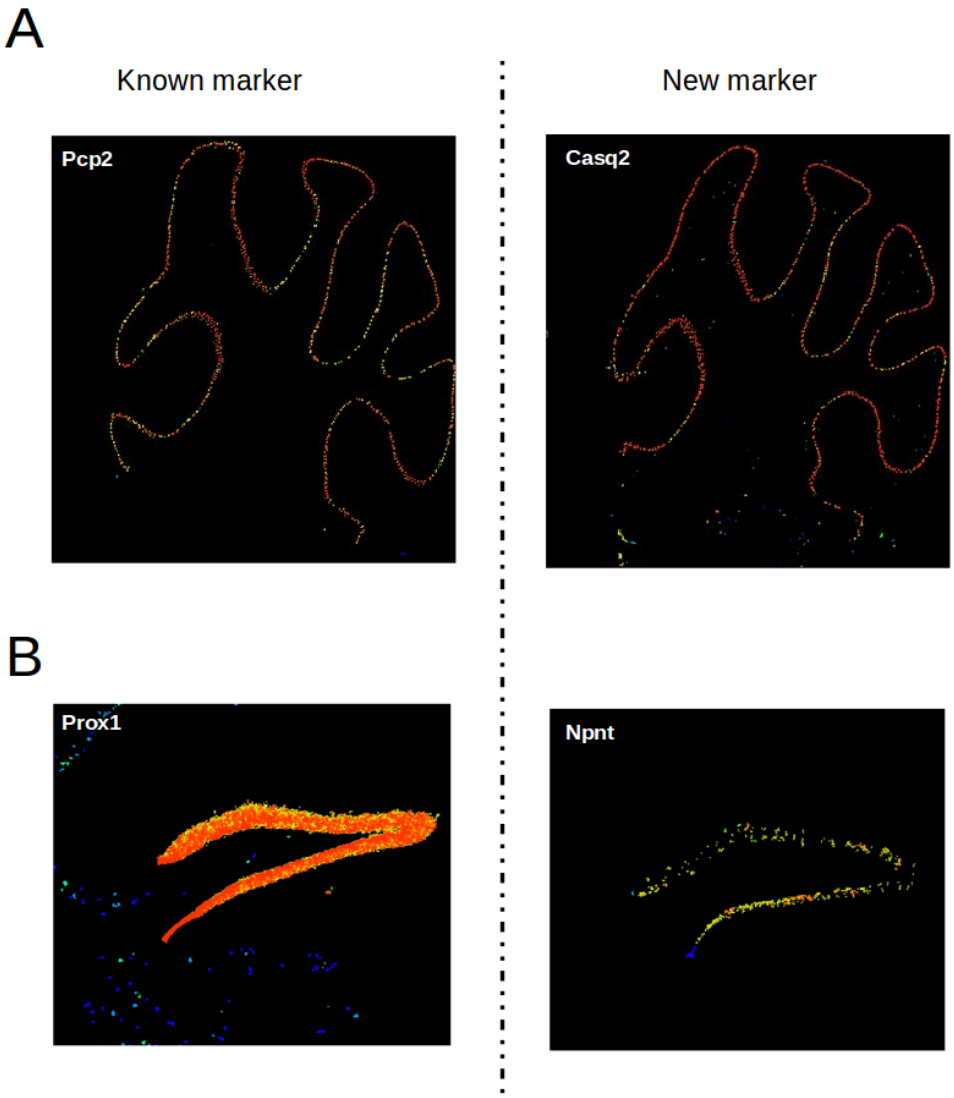


Figure 5: Expression of known marker genes and newly discovered marker genes in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. A. Expression of new and known markers of purkinje cells in cerebellum. B. Expression of new and known markers of granule cells in dentate gyrus, granule cell layer

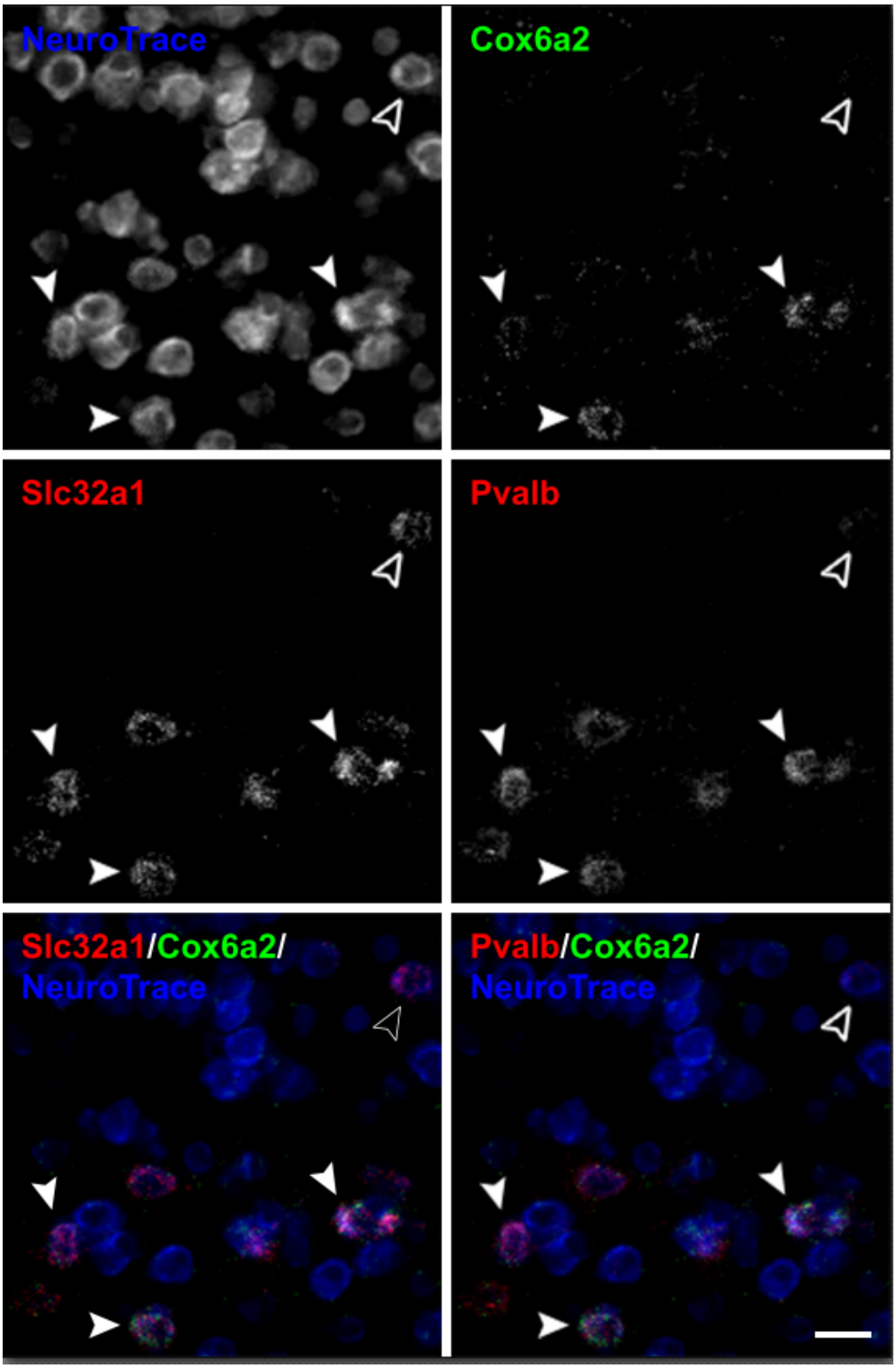


Figure 6: Triple labeling of fast spiking gabaergic cells in mouse cortex. Experiments performed by Dr. Sibelle's group. NeuroTrace is a general neuronal marker. Slc32a1 and Pvalb are known markers of fast spiking pyramidal genes. Cox6a2 is a marker gene discovered through our analysis. Last row shows superimposition of known markers and Cox6a2 which appear in the same cells.

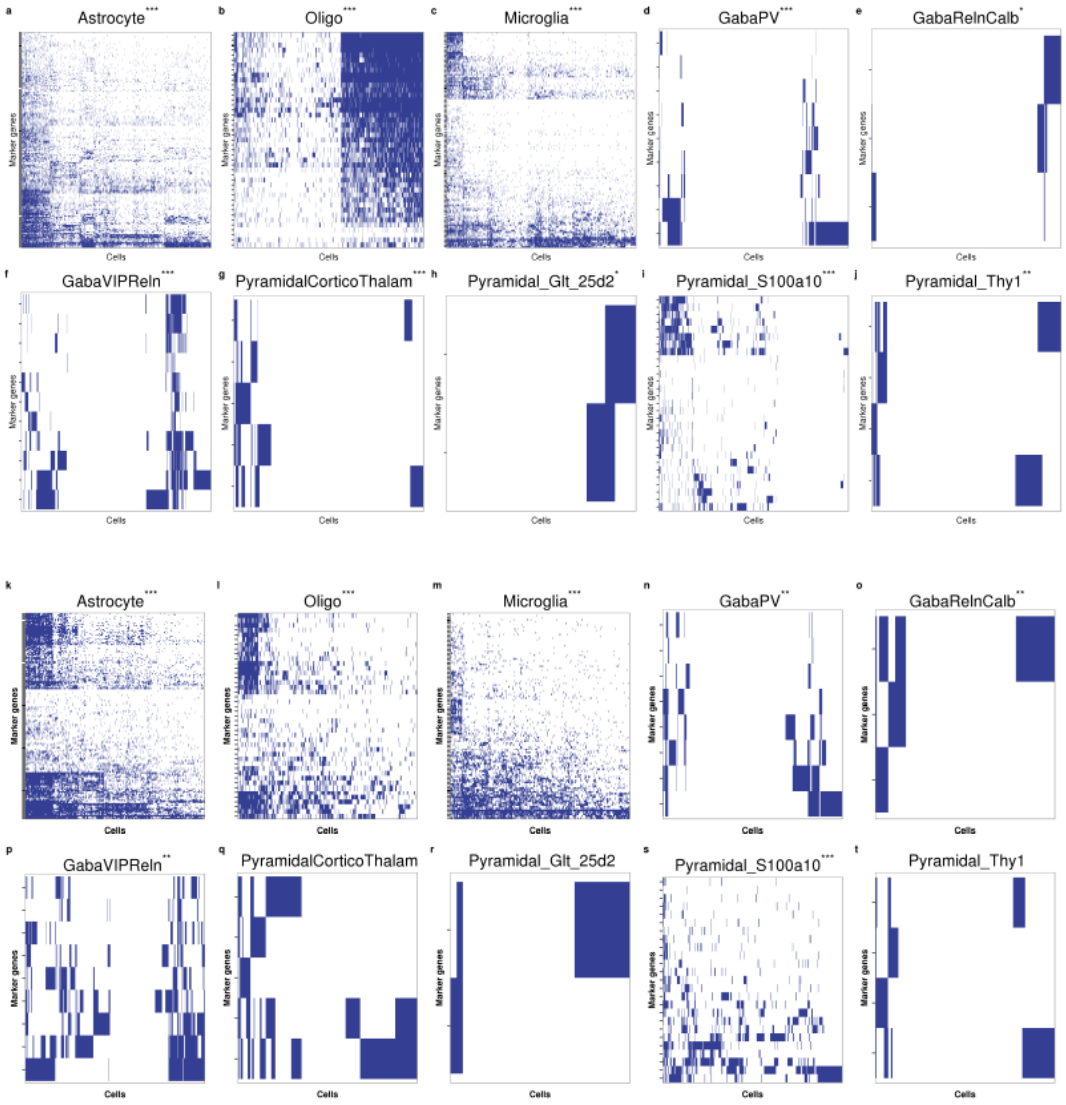


Figure 7: Binary heatmaps representing the expression of marker genes in single human and mouse cells. Blue shows the existence of a gene while white shows its absance. Significance stars represent the difference between coexistance of the genes and randomly selected gene sets with similar prevelance in the dataset. a-j shows the expression of marker genes in mouse single cells (Zeisel et al. 2015). k-t shows the expression of marker genes in single human cells (Darmanis et al. 2015). Since the data is collected specifically from frontal cortex, only cortex cell types are tested.

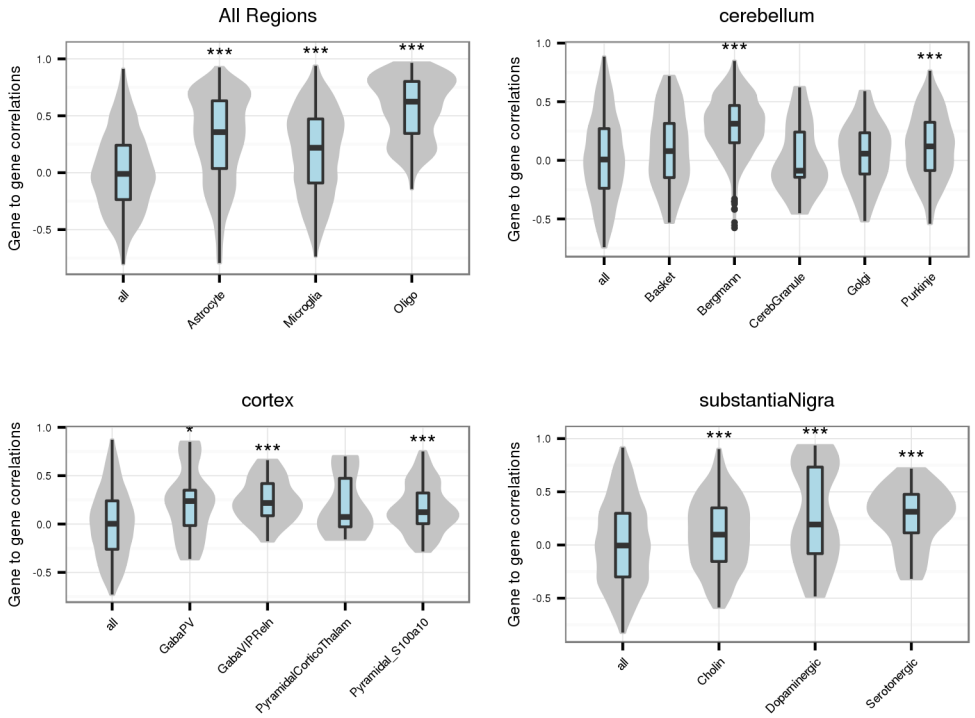


Figure 8: In between coexpression levels of marker gene sets. Significance markers show significantly higher co-expression than co-expression between all genes

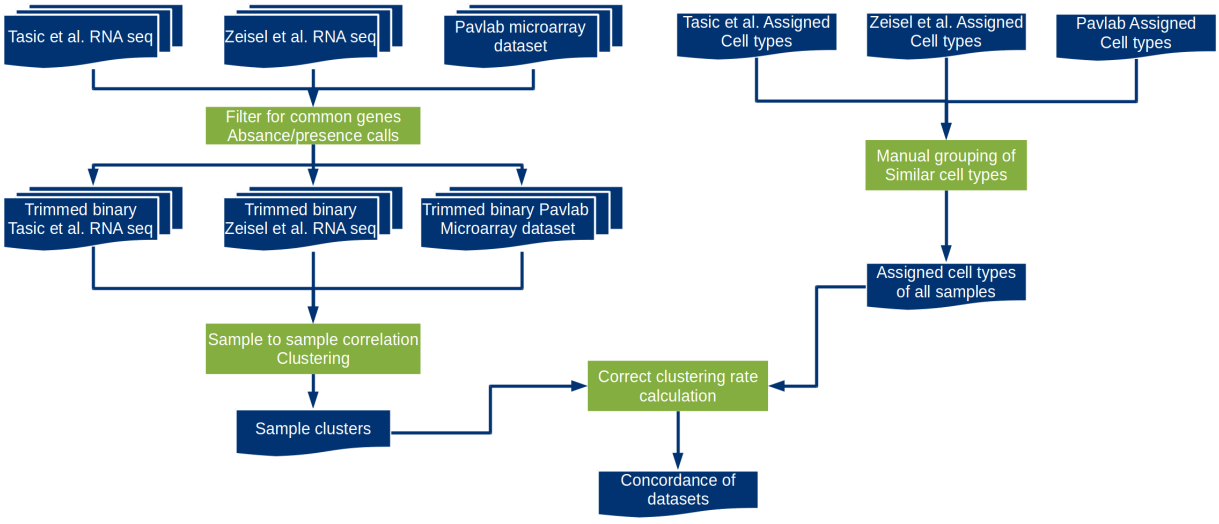


Figure 9: Pipeline for the upcoming analyisis on concordance of different cell type based analysis studies.

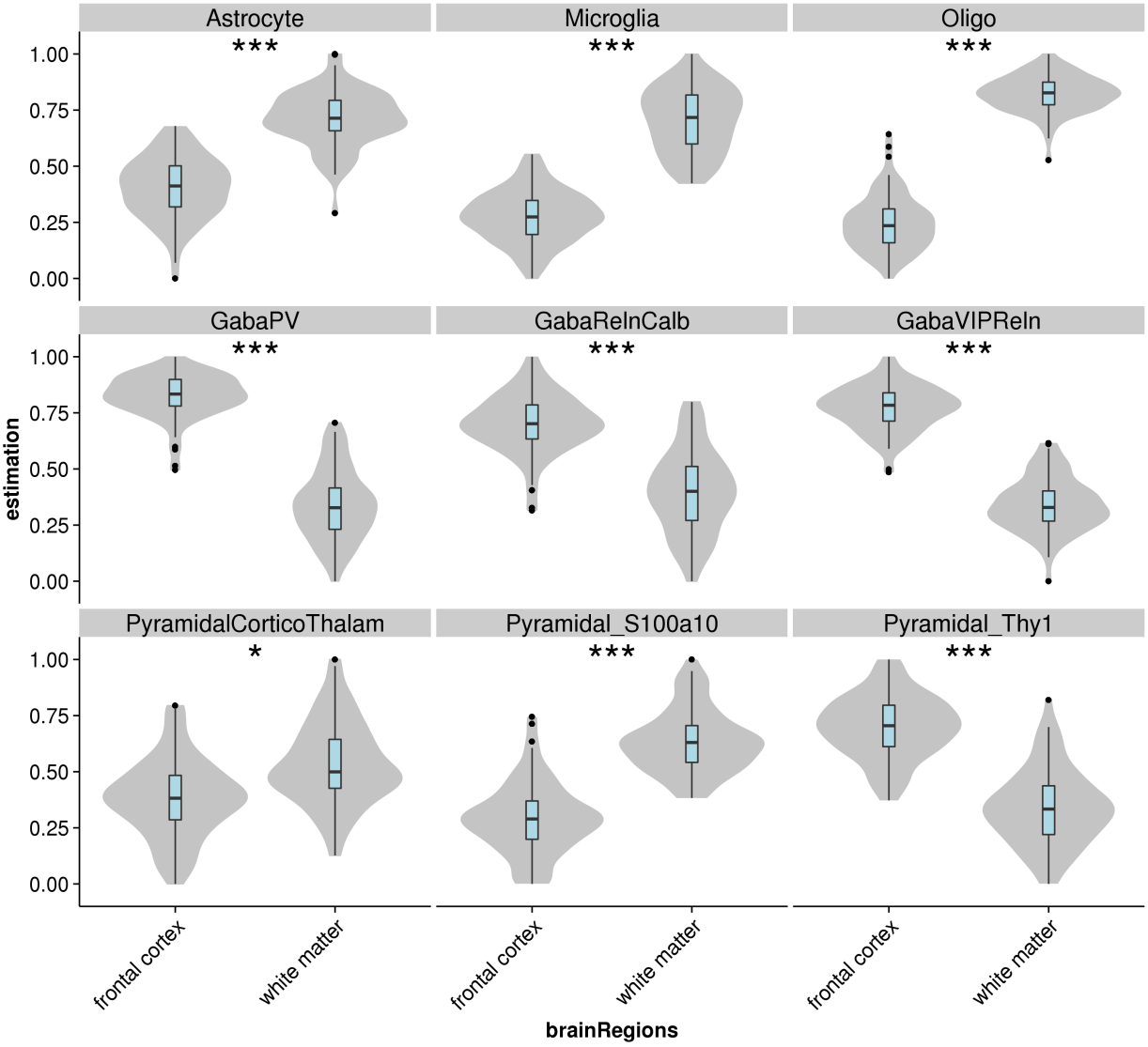


Figure 10: Estimations of cortical cell types in frontal cortex and white matter. Values are normalized to be between 0 and 1. Estimations appropriately reflect expected differences between white and gray matter for the most part. It is also possible to see some unexpected increase of some pyramidal subtypes.

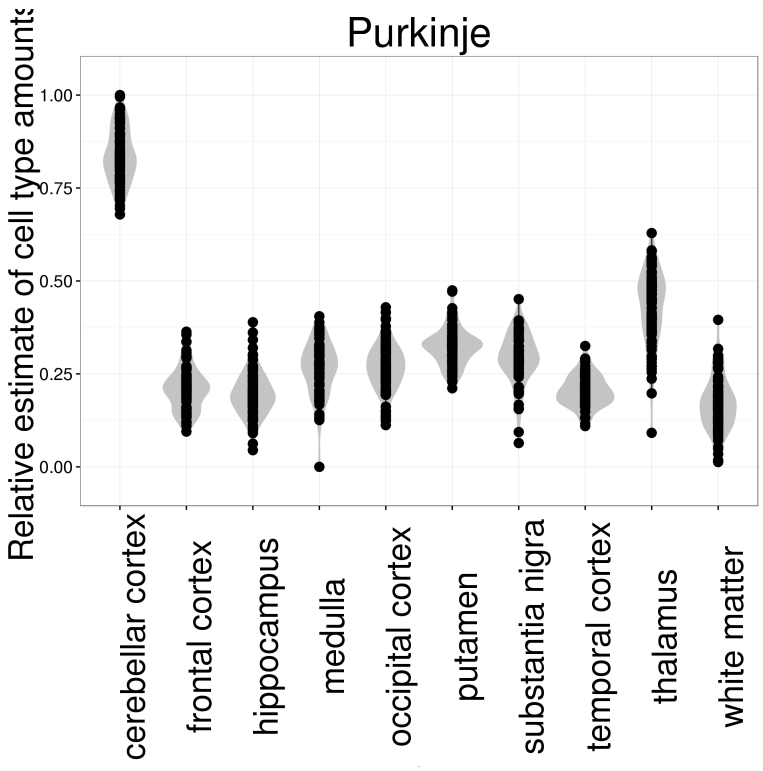


Figure 11: Estimations of purkinje cells in different brain regions. Values are normalized to be between 0 and 1. Purkinje cells are specific to the cerebellum.

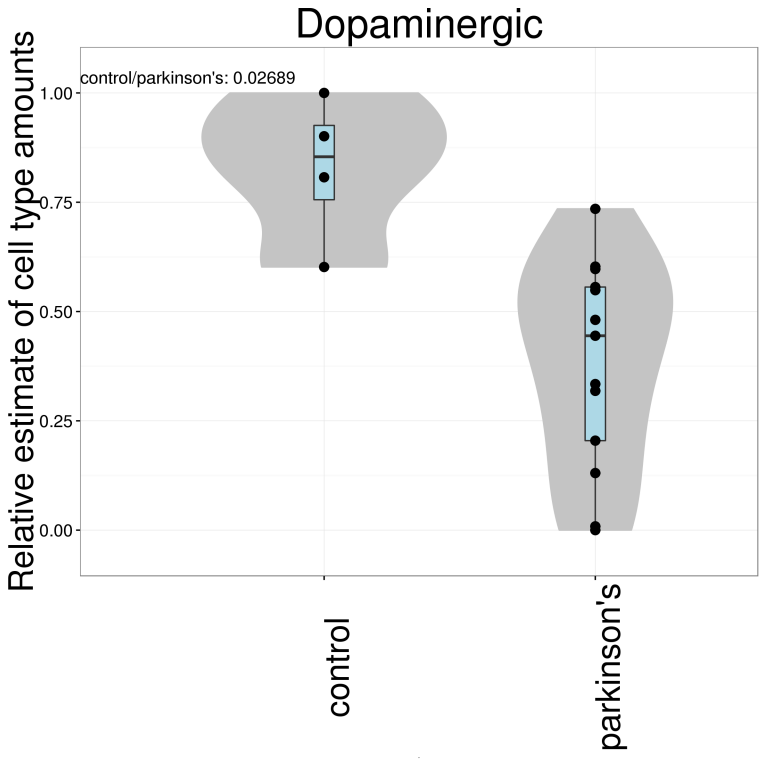


Figure 12: Estimations of dopaminergic cells in different substantia nigra of male parkinson's disease patients. Values are normalized to be between 0 and 1. Dopaminergic cell loss is an expected consequence of Parkinson's Disease

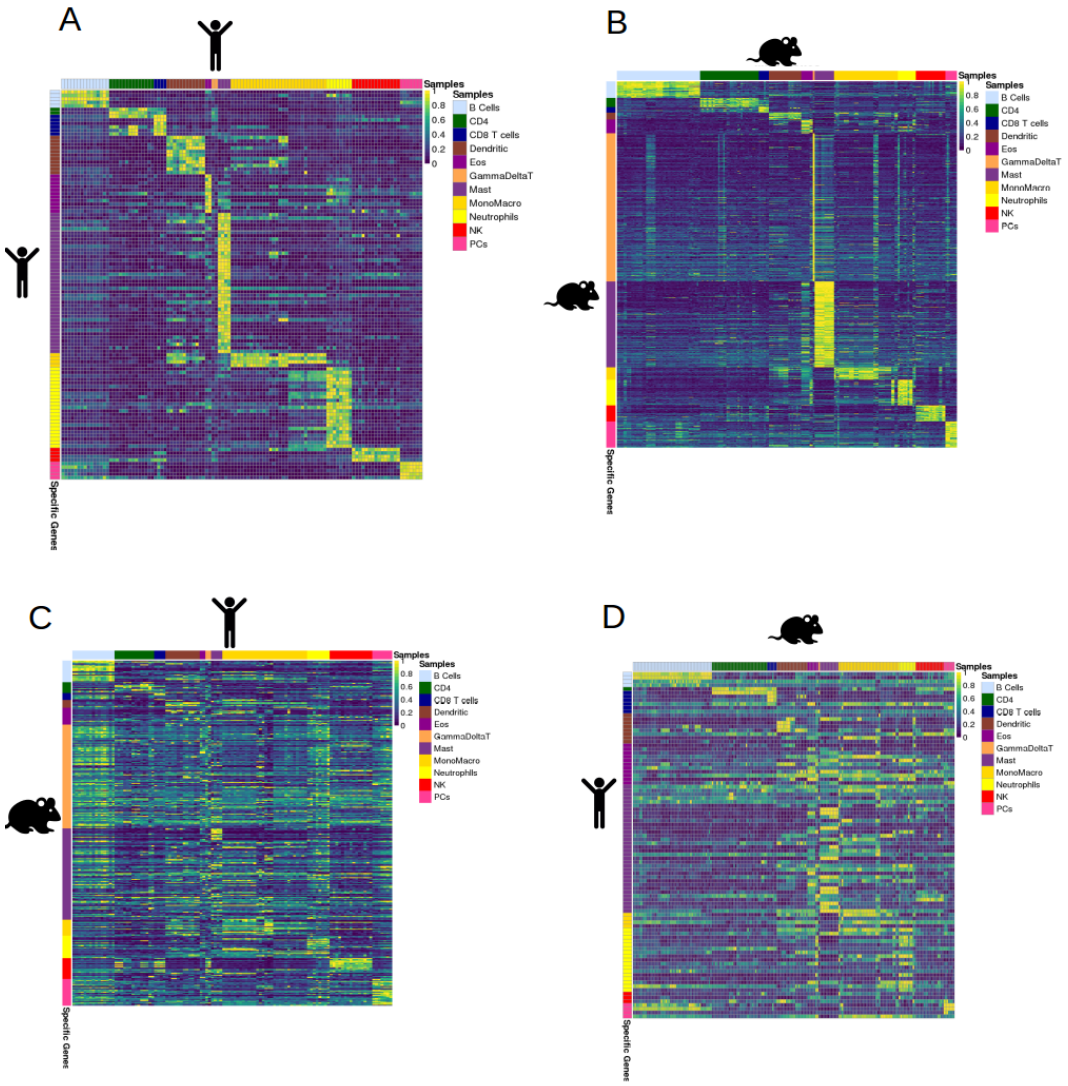


Figure 13: A-B. Expression of the genes selected from a species in the samples used for isolation from the same species. A shows human genes in human cell type specific expression profile dataset while B is mouse genes in mouse cell type specific expression profile dataset. C-D. Expression of homologues 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.

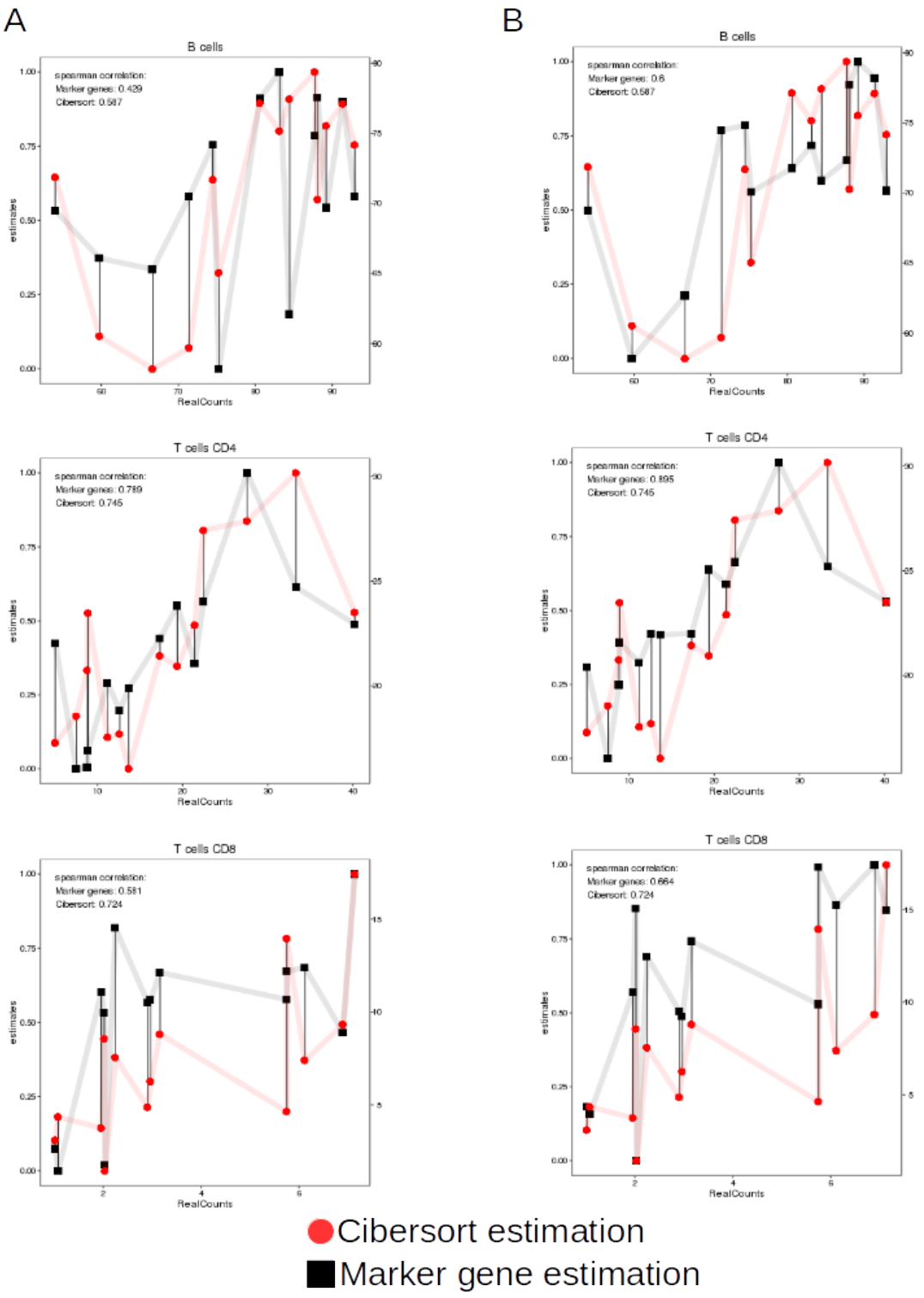


Figure 14: Estimations done by our method (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. A. Estimations done using marker genes selected from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

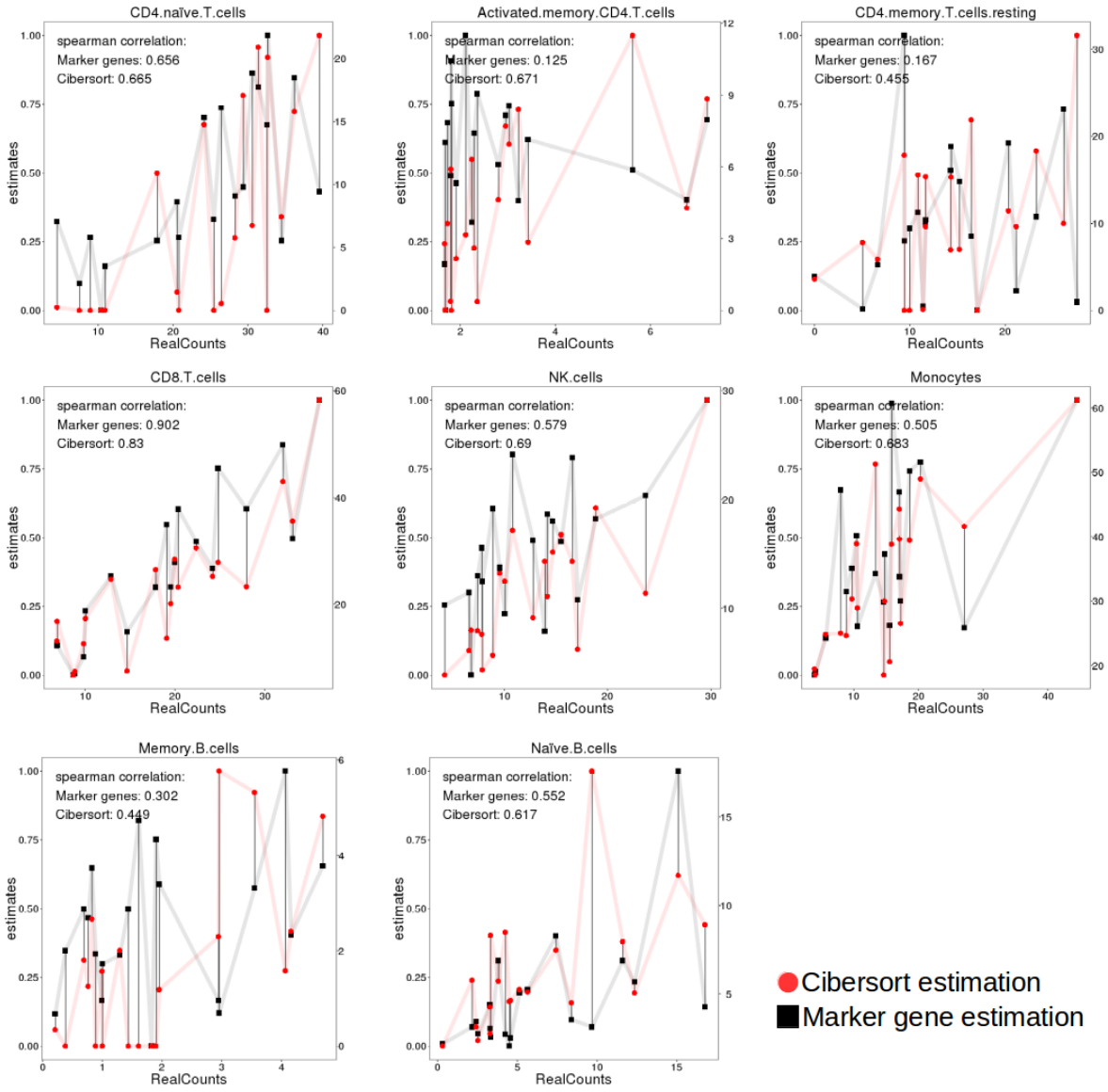


Figure 15: Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage.

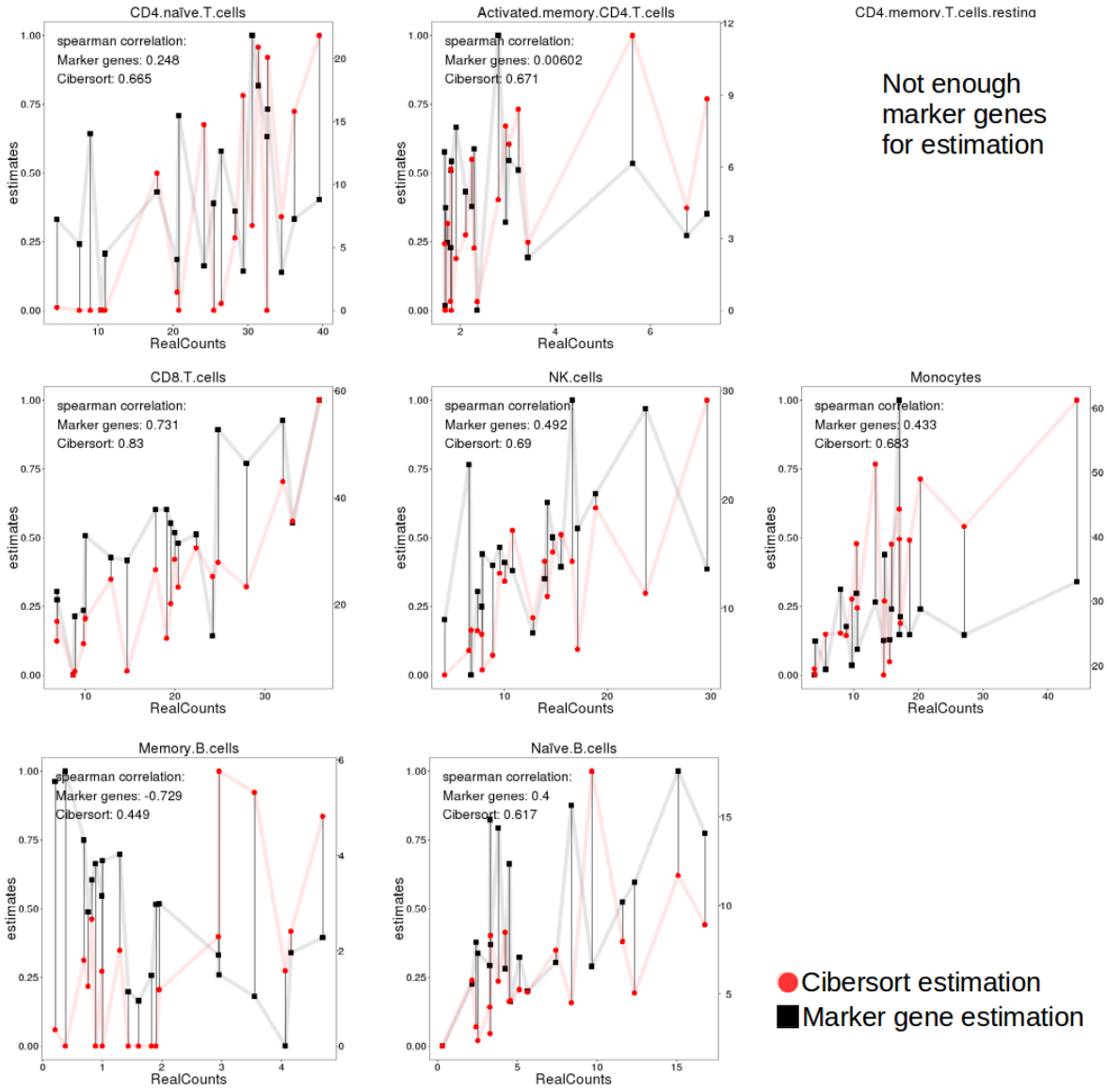


Figure 16: Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is 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.