WOP Title

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. Robert Holt

Dr. Sara Mostafavi

Chair

???

Examination Date

June 19, 2015

Contents

List of Figures

List of Tables

1 Motivation and Introduction

The brain is a remarkably heterogeneous organ with 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 known, most large scale expression studies that focus on neurological/neurodegenerative disorders use whole tissue samples to examine the effects of diseases[1–3]. Even though this approach remains popular due to its relative ease and low cost, it complicates the analysis of the results by obfuscating the source of observed differences (eg. which cell type(s) is effected by the changes) and making harder to detect changes in less abundant cell types due to signal dilution[4].

Studies focusing on the expression profiles of single cell types do exist. These studies attempt to observe how a single cell type is effected by different conditions such as disease[5,6] or development[7,8], or they attempt to characterize cell types by finding their unique properties among a small subset of cells of interest such as marker genes or electrophysiological properties.[9,10].

The studies mentioned above however leave significant gaps to be filled in the field which will be the focus of this work. **1)** They do not tell what are the marker genes of a cell type in the scope of the entire brain region they reside in since they do not examine all cell types in that region. Discovery of new markers are important since not all neuron types have known unique markers, and many have few numbers of markers. This is a problem because cell unique 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 etc.). 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)** They do not to tell if there are changes in type proportions which is a common occurence in neurological diseases such as Parkinson's and Alzheimer's. **3**) They do not tell which gene expression changes occur in which cell types.

Trying to access this knowledge by common laboratory methods would be expensive and labor intensive. Finding definitive marker genes require isolation of all known cell types from a brain region, detecting cell type specific differences of all cells types 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 formed a highly inclusive dataset of cell type specific expression profiles from a variate of resources to help with the problems described above. This dataset, along with various validation methods, will be used to detect marker genes that that best represents a given cell type in their brain regions. 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 be using 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 be using these surrogate proportions as covariates in statistical tests in order to improve statistical power of differential expression analyses and be able to tell which cell types are effected by specific changes.

The pipeline 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 common research that focuses on cell types isolates few related cell types based based on the lab's interests and try to characterize the cells in relation to other cells they are working on[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 envalues our approach of choosing marker genes using a dataset of cell type expression profiles gathered across independent studies to be as inclusive as possible.

This creates the opportunity for taking a more comprehensive approach using the already available data in the literature.

2.1.2 Are mouse marker genes applicable to humans?

Most available data in the literature on isolated cell types are coming from mouse cells. Whereas 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 brain, using our methodology.

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 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 their markers. Due the non straightforward 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 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. Gathering high quality gene expression data representing brain cell types
2. Employing quality control measures to minimize intake of flawed data
3. Making the data available in a web application

2.2.2 Aim 2: Identification and verification of marker gene sets

1. Detecting cell type marker genes in a region dependent 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 the whole tissue samples as a basis enumerate the relative amounts of given cell types in the 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 and independent dataset of isolated blood cell types and manually enumerated blood samples to repeat and validate the enumeration method.
5. Use enumeration information in improve the accuracy of differential expression analyses.
6. Create an R package for easy application of the method by third parties.

3 Background

3.1 Expression profiling

3.1.1 Microarrays

RNA microarray is the most common way to quantify RNA in a high-throughput manner and it's development have been a transformative force in many branches of biology[14]. Microarrays are built by planting single strand probes, that are specific to a location on a target genome, in high density, to a known location on a solid surface. These probes will later be hybridized to a labelled complementary DNA (cDNA) acquired by reverse transcription of a target transcriptome. The amount of cDNA that hybridized to a probe is later quantified by staining the label attached to the cDNA molecules[15]. Often, multiple probes target a close-by region on the genome which are then summarized to make up the signal for a single gene[16].

There are a multitude of microarray platforms for researchers to chose from. These platforms primarily differ in the probes they use that can cover a different number of genes and/or cover the same genes using different sequences.

Microarrays are extensively used in neurobiology as a default tool for expression analysis. A wide array of data is available both in tissue[1,17,18] and isolated cell type[6,7,9] level.

A short summary of microarray methodology can be found at Figure 2.

3.1.2 RNA sequencing

RNA sequencing (RNA seq) is a much more recent method of RNA quantification. RNA seq is performed by sequencing of cDNA molecules that is acquired from the reverse transcription of a 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. The 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[19]. In general RNA seq is more prone to technical artifacts due to the stochasticity of the sequencing process. This effect is particularly powerful for low expressed genes which often make up the majority of the data[20].

Recently RNA sequencing of single cells are becoming increasingly popular[21]. 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 is more prominent[21].

RNA-seq analysis, especially single cell studies are starting to gain popularity in neuroscience[12,22]. Due to it's heterogeneous structure brain is a prime target for single cell studies that allows differentiation of individual cell types with much less concern of isolating heterogeneous samples.

A short summary of RNAseq methodology can be found at Figure 2.

3.1.3 Analysis of RNA quantification results

Most common use of RNA quantification is to observe differential expression between two or more groups. This is done in order to find what genes are effected by certain conditions or expression is different between different cell or tissue types to a degree of statistical significance[16,22,23]. Another common method of analysis is coexpression, where researchers attempt to identify genes that show similar changes in expression across different samples[24]. This information is often used to derive functional relationships between the genes[25,26].

A higher level analysis of RNA quantification results uses more complex methods to deconvolute cell type proportions as discussed in a later section.

3.2 Major cell types of the brain

Brain hosts a variety of cell types and is one of the most heterogeneous organs in the mammalian body. Alongside the major groups of cell types that is associated with it: neurons and glia, it also contains blood and endothelial cells. Below are short summaries of descriptions of major neuron and glia types that are included in our cell type specific expression database.

3.2.1 Glia

- **Astrocytes** are star shaped cells present throughout the brain[27]. 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[27]. Relatively recently it was also discovered that they play roles in signal transduction by regulating intracellular ion concentrations and releasing gliaotransmitters[28]. Astrocytes often proliferate in diseased brain[27].

- **Oligodendrocytes** are responsible for forming the mylein sheet around neurons to insulate their axons[29]. A single oligodendrocyte ensheats multiple neuronal cells as a result of a highly coordinated process effected by axon size, neuronal activity and molecular signalling[29]. Most myelination occurs during early differentiation process. Alongside myelination, oligodendrocytes also have neuroprotective functions and are known to dysfunction in several neurodegenerative diseases[30].

- **Microglia** are resident macrophages of the brain[31]. They clear apaptotic cells and are involved in maintenance of synapses[31], they are aksi the antigen presenting cells of the brain[32]. In most neurological diseases they activate and proliferate[31] and their proliferation is associated with degredation of neuronal cells[32].

3.2.2 Neurons

Compared to glial cells, neurons are more closely related to each other but they have distinct functions in a neurotransmitter and region dependent manner. Neurons act under a wide variety of neurotransmitters that allow information transfer between neurons and gives us clues about their function. GABA for instance is an inhibitory neurotransmitter[33]. Different subtypes of GABAergic cells with unique firing properties and morphologies regulate different properties of the brain. Fast spiking cells function as the main inhibitory system in the neocortex by providing fast and powerful inhibition whereas SST expression GABAergic cells receive facilitating signals to provide inhibitory feedback[33].

Pyramidal Cells are prominent in areas of the brain associated with high cognitive functions such as cerebral cortex and hippocampus[34]. They are recognized by their triangular cell body and short basal dendrites[34]. Pyramidal cells of different brain regions and layers has differences in structure and gene expression[34]. They can receive inhibitory gabaergic inputs through soma and axon. Excitatory signals are delivered through the dendrites[34]. Dendrites of pyramidal cells are covered with high number of spines that act as synapse sites for glutamatergic synapses[34].

Cortical gabaergic neurons are local inhibitory neurons that control pyramidal cell firing and generation of cortical rhythms. In cortex 3 major groups are defined. PV expressing, SST expressing and 5HT3a expressing[33]. These subtypes of gabaergic cells localize into specific layers of the neocortex[33]. PV expressing neurons are fast spiking cells with low input resistance[33]. 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[33]. Finally 5HT3aR interneurons are a more heterogeneous subgroup of interneurons which differ in function and morphology[33].

Midbrain dopaminergic neurons are primarily responsible for motor functions, emotion and reward[35]. During development of the midbrain hey separate into 3 distinct clusters of cells[35]. Cells making up the A9 cluster, that forms substantia nigra are lost during the prognosis of Parkinson's Disease[35].

Purkinje cells are gabaergic cells local to cerebellum's purkinje layer[36]. They receive signals from nearby granule cells and send inhibitory signals to other purkinje cells towards deep cerebellar nuclei[36]. They have functions in motor learning[36].

Cholinergic cells have functions in memory formation[37]. Particularly, they are associated with the Alzheimer's Disease[37]. They are also responsible for maintaining the circadian rhythm, observable by a high acetylcholine release during awake periods[38].

3.3 Cell type markers and their applications

Marker genes are useful in many ways to understand the biology of the cell type they are specific to. Primarily they can be used to identify cells of interests in whole tissue samples for a variety of purposes such as counting and purifying cells. Aside from their uses in the wet-lab, marker genes are also powerful tools in computational settings. They can be used as features in deconvolution of complex tissue samples as explained the the following sections.

Many brain cell types have specific and well known markers that makes them easy to identify. Many of these genes have known functions that is directly attributable to the cell type. They include Mog that is an

oligodendrocyte marker[39] with roles in myelination, and Aif1, a microglia marker with roles in inflammation[40] for glial cells and Th that is responsible for dopamine synthesis in dopaminergic cells[35], or Gad1/2 that catalyses production of GABA in gabaergic cells[33]. 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 harder.

3.4 Cell type isolation

Isolation of single cell types is necessary as a precursor to their proper characterization or to analyze specific cells in different conditions such as their response to diseases or chemicals. There are multiple ways to isolate the cell types of interest with varying degrees of precision and quality. Most commonly, such methods rely on one or more marker genes that is 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 using in situ hybridization or labelled antibodies respectively. Cells are then gated according to specific conditions (eg. expression of a gene, absence of a gene) (Figure 3 A)[41]. Another established method of isolation is Immunopanning where antibodies layered to a plate are used to hold the cell that express a specific surface marker (Figure 3 B)[42]. A relatively recent marker based isolation is Translating Ribosome Affinity Purification. This method combines the promoter of a marker gene with the coding region of L10a ribosomal subunit fused with green fluorescent protein (GFP) . The tissue is degraded in mass and after fixation, ribosomes marked with GFP are isolated, ensuring only translating RNAs from the target cell type are isolated (Figure 3 C)[43,44]. Alternatively, based on visible characteristics or expression of known markers, cells can be visually located on the tissue and isolated by manually picking or laser capture microdissection (LCM) (Figure 3 D)[45].

The resulting samples from each of these methods have varying purity and data quality. Study by Okaty el al[46] evaluated the relative quality of different purification methods and found that TRAP has the worst record in purity while PAN causes most stress to the cells.

3.5 Cell type deconvolution

Expression levels obtained from whole tissues contain signals from multiple cell types. Expression profile of complex tissues can be modelled as below

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

where X_{ij} is the expression value from a complex sample for genes j and sample i and W_{ik} is a matrix containing cell type proportions for sample i and cell type k, and h_{kj} is the cell type specific gene expression of cell type k and gene j and e_{ij} representing random error. This model enables usage of various methods to attempt to acquire information about the matrix W or h. Two main classes of deconvolution methods exist: Reference based and reference free deconvolution methods and will be explained in the next sections. In mammals, deconvolution methods are commonly applied to blood data due to ease of access to both mixed samples and isolated cell types[4,47,48].

While deconvolution attempts in brain is not a new idea, these attempts remained at a relatively superficial level. Early attempts that attempted to deconvolute human brains estimated proportions of neurons as a single group along side astrocytes, oligodendrocytes and microglia[49]. Later, more in depth deconvolution was performed in human cortex and cerebellum that estimated cerebellar neuron types separately while leaving cortical neurons as a single group[50]. 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 end are available in the literature.

A reference based deconvolution (see below) of 64 distinct cell types was performed on whole tissue expression profiles across various brain regions[51]. In this study, proportion estimations of most cell types fit the literature but they also reported paradoxical results such as detecting high levels of purkinje cells in thalamus instead of cerebellum[51]. Also no attempt was done to deconvolute cell types in samples from the same regions but under different conditions (disease models, developmental stages, eg).

3.5.1 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. Most naively, researchers try to estimate the W (matrix of cell type proportions) in solving the equation 1 by minimizing the e (error)[51]. 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, **2)** the reference dataset has all cell types represented in the mixed sample, which can be violated by existence of previously unkown cell types in the

region. To combat such problems different methods of feature selection that aim to identify most informative parts the reference expression matrix can be used which in turn makes the estimation process more robust[48].

3.5.2 Reference free deconvolution

In cases where cell type expression profiles is 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 might the the better alternative. A common method is to use expression of certain marker genes as a surrogate for cell type proportions[4,46,47,50]. Even though the marker genes themselves are often acquired from a reference expression dataset, deconvolution is independent of their expression in the reference dataset. Often the first principle component (PC1) of the genes in the whole tissue samples are used as a surrogate[4,50,52]. This assumes that most of the used marker genes are not differentially regulated between samples and the main source of variation is the difference between the cell type proportions in between samples.

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

The first aim of the project, that also lays the groundwork of all the later ones, 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 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 its human counterparts.

4.1 Data acquisition and preprocessing

The bulk of the dataset is based on a previous compilation made by Okaty et al.[46] for a study attempting to compare different cell type isolation methods. This initial dataset had data obtained using Affymetrix Mouse Expression 430A Array (430A) and Affymetrix Mouse Genome 430 2.0 Array (430.2). Data from two platforms is straightforward to combine since 430A array contains a subset of the probesets in 430.2 array. Due to high availability of the data collected 430A and 430.2 arrays and to keep processing of data easy, we decided to populate our database with datasets from these platforms only. We queried Gene Expression

Omnibus(GEO) for isolated cell types from mouse samples. To be able to pre-process the entire database all together, we acquired raw data files (CEL format) for each sample. Samples from 430.2 array were stripped of the extra probesets they contained and merged with the data from 430A array samples. The resulting dataset is pre-processed and normalized using Robust Multichip Average (RMA) method[53,54]. Potentially due to the technical differences between studies,we observed significant differences in distribution of the probeset level signal distribution after RMA normalization. To make samples comparable to each other, we used a second quantile normalization after RMA[55]. 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 involves ensuring expression of known cell type markers (markers from literature and markers that were used to isolate the cell type) and making sure 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 since other samples represented their subtypes. For instance samples representing Htr3a positive gabaergic cells removed due to the existence of, VIP positive cells that are their subtypes[33]. When we concluded some cell types are too similar to each other to detect meaningful differences between them, they are grouped together in a single cell type. Drd1 and Drd2 positive spiny neurons exemplify this as their expression were too simmilar to each other to detect different markers for both. The resulting dataset has 31 cell types isolated from 11 regions gathered from 24 studies isolated with a variety of methods (Table 1). We are still looking at newly published papers in order to add more cell types when it is possible

4.2 Presentation of the data in a web application

Upon collection of the database, we created a web application to facilitate access to the database The web application allows third parties to easily visualize expression of chosen genes in individual cell types in their respective regions (Figure 4). The application also allows grouping of cells together in a hierarchical manner. Every sample shown also 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 do quick differential expression analyses between cell types and a gene set enrichment tool that will allow researchers to check their hitlists for cell type specific enrichment (see below). The application increases the value of our database by making its use much easier for researchers who are not from a computational background.

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 with 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 said region. To accomplish this we separated samples into regions based on the metadata acquired from the original source. We had to generalize certain regions to make the marker gene selection biologically relevant. For instance brainstem is taken as a single region because 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 cell types represented in each region by a clustering based method. For any given cell type, we designated a gene as a marker gene if:

- There is 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.
- Separating samples into 2 clusters, samples representing the target cell type and all others, and defining the distance between samples as difference of expression of the target gene, the silhouette coefficient of the resulting clusters must be 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 is why we choose our gene selection method instead of a simple differential expression analysis

due to the uncorrectable batch effects that potentially resides in the database. Since batch effects were potentially prominent, we wanted to choose genes that have drastically different expression (first condition) to the rest of the samples which would reduce the effect of batch effects on our decisions. We also aimed to look for genes that reliably separate samples from each other rather that simply being highly expressed (second condition).

As a result, we selected marker genes across 31 cell types isolated from 11 regions. Number of marker genes greatly vary 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 problematic due to artifacts caused by potential differences in mouse strains and 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 make sure that they act as marker genes in 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 ensuring the sensitivity and specificity of our marker genes. This will be done 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[56] to confirm our findings. While a powerful resource including thousands of ISH images, since every slice is labelled by a probe specific to a single gene, it is not possible to conclusively decide if the signal is coming from the same cell types unless the cell type is highly concentrated in a specific structure in the brain. Granule cells of dentate gyrus and purkinje cells of cerebellum fitted this criteria so we were able to confirm some of the marker genes through Allen Brain Atlas 6[56] .

We are currently collaborating with Dr. Etienne Sibille to validate the markers by dual labelling. They were 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 potentially 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 not possible to apply biological validation methods to all marker genes that we selected, we are using recently published single cell RNA-sequencing datasets[12,13,22] to validate our finding. These data sets are coming from a recent outburst of various labs' efforts to characterize 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 are assuming that they have the same cell types that our database represents. Yet due to the high granularity of the clusters resulting from clustering of the single cell data, it is not straightforward to match which individual cells should match the cell types defined in our data. To combat this, 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.

Since single cell expression analysis is still in it's infancy, often the transcript counts for most genes end up being very low, which makes the exact expression value an unreliable measure. To combat this, instead of using the full expression values, we converted the data into a binary matrix where 0 meant no expression of the gene and 1 meant any expression of the gene. This approach potentially biases the results against us since we do not chose genes based on their exclusivity to a single cell type, but its heightened expression in the cell type. Yet applying the method to a mouse[12] and a human[22] 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 the null distribution, and in the human dataset a majority (7/10) of the marker gene sets showed significantly more coexpression.

Next step is to repeat this analysis using a more recently published RNA-seq study done on mouse brains that isolates single cells in a more cell type specific manner[13]. This dataset is of higher quality and have a better coverage of cortical cell types.

5.3.3 Validation of marker genes in human whole tissue data

As another validation approach, we used a human dataset containing 16 brain regions from pathalogically healthy subjects[57]. For all marker gene sets we obtained, we analysed the coexpression of marker genes in brain regions relevant to the cell type. Since marker gene sets are cell type specific, in a complex tissue, variation in the amount of the said cell type is determinant of their expression (eg. 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 will 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 in between these samples to coexpression levels of randomly selected genes. The results showed heightened coexpression for most of the cell type marker sets (11/15) (Figure 9).

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

The high frequency of recently published RNA-seq studies create a large output of data that are very similar to each other by nature. All of them are cells from the brains of the same organism, hence in ideal conditions, the cell types they identify should overlap with each other and our microarray dataset. Due to the inherent differences in the data structure, assessing repeatability of such single cell studies is not straightforward to do. 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 of them in an expression independent manner. For microarray data we will be using absence presence calls based on an expression threshold and for RNA-seq, we will be looking to see if the gene is captured at all in the sample. We are hoping that this analysis will provide sufficient information to correlate the samples to and allow us to identify which samples from each dataset correspond to each other. If we cannot reliably group samples from independent sources together, we will group the single cell data according to their designated groups in their respective papers. We are hoping to find out if these independent studies really identify the same cell types or if the cell types are not fully equivalent potentially due to experimental methods or differences between mouse strains. A plan of the expected analysis can be found at Figure 10.

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 relatively well established use of marker genes is the estimation of cell type proportions in whole tissue samples using their expression. As mentioned in the introduction, two types of deconvolution methods dominate the field: reference based and reference free deconvolution. In our case, we choose to use reference free deconvolution due to the fact that the reference expression profiles we are using are from mice and we often want to do proportion estimation in human brains, and the exact level of expression in our dataset might not be reliable since we are attempting to deconvolute human samples and the expression profiles are 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, and it is sensible to assume that sufficient amount of marker genes will be preserved across species. Our aforementioned validation in human RNA-seq and whole tissue data, along with further validation of the pipeline that will be explained later this section

confirms that this is not an unreasonable assumption to make.

To estimate the relative amount of cell types between samples, we used the expression of marker genes in the samples as a proxy. This was done by calculating the first principle component of marker gene expression in individual samples and using it as a proxy for cell type proportion. Adopted by multiple other groups that attempts to do deconvolution in whole tissues[4,47,50], the idea behind the method is that most of the variation 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, that represents the relative amount of a given cell type compared to other samples. This number cannot be used to compare two different cell types.

To verify that the method work as expected, we estimated relative cell proportions in different brain regions with known differences between cell type proportions and we got results that would be expected based on the literature. In a dataset of different brain regions from healthy donors[57] it was possible to observe an increase in glial population and decrease in most neuronal populations between white matter and grey matter (Figure 11) and purkinje cells were found to be exclusive to cerebellum (Figure 12). 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[58], 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 13).

We will be analyzing more datasets from neurological diseases and brains under different conditions to increase our confidence in the dataset and attempt to use it as a discovery tool. There are hundreds of brain samples available in Gemma database[59] for expression profiles with well annotated metadata to allow for their fast analysis. These datasets include brains under many conditions such as neurological diseases and disease models, gene knock downs, etc. We are hoping this will reveal many condition specific cell type proportion changes. This step will also be another useful validation since we it will allow us to see how if we can detect the same proportion changes in different studies working on similar condition.

6.2 Repeating the whole analysis pipeline with blood cells and tissue to validate the method

Estimation of brain cell type proportions is hard to validate since there are no expression datasets couples with cell type counts. Any result we find is unverifiable other than the expected differences between groups. To asses the accuracy our method, expression data from whole tissues that is paired with cell type counts is 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 samples are coupled with cell counts, acquired through well established methods[60,61]. Cell type reference datasets for blood cell types are present in the literature for both mouse and men. This allowed us to construct a similar database to our brain database for mouse and human blood cell types. We subjected both these databases to marker gene selection steps. To see 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 appeared to be specific to the cell type they were selected from in different species(Figure 14). 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 compare our results with a recently published reference based estimation method[48]. When the cell type definitions are kept at a relatively general (eg. B cells, T cells) level (eg. not differentiating cell types at different activation stages), mouse genes performed than human genes (Figure 15), potentially due to difference of quality between the datasets. Whereas attempting to estimate finely defined cell (eg. activated/deactivated CD4 cells) types with mouse genes returned poor correlation to actual counts (Figure 16 - 17). Shay et al.[62] 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 which 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

Basic differential expression analyses on whole tissues have problems due to 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, making it harder to get significance. Previous work suggests that it might be possible to increase the power of differential expression analysis by adding estimated cell type proportions as covariates [chikina_cellcode_2015]. The authors also show observed effects can be localized to their cell types by using

the estimated proportions by using interaction models. In neuroscience, where sample sizes are often small and data quality is not top notch, this approach had the potential to increase the value of the existing data to a great extent by increasing the statistical power of the analyses. We are hoping to validate this 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 single cell type study to that cell type in the study that uses on whole tissue samples.

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

The pipeline for gene selection and enumeration is while relatively simple, it is time consuming to deal with the magnitude steps aiming to fine tune the process. By turning creating an R package we aim for our process to be reproducible by third parties. The package will include streamlined functions to select and validate the marker genes, along with functions used in enumeration process. The package will be publicly available on Bioconductor, CRAN 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 Hoheisel JD. Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews Genetics* 2006; **7**: 200–210.

15 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.

16 Gautier L, Cope L, Bolstad BM, Irizarry RA. Affyanalysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004; **20**: 307–315.

17 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.

18 Torrey EF, Webster M, Knable M, Johnston N, Yolken RH. The Stanley Foundation brain collection and

Neuropathology Consortium. *Schizophrenia Research* 2000; **44**: 151–155.

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

20 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.

21 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.

22 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.

23 Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* 2014; **11**: 163–166.

24 Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD *et al.* Functional Discovery via a Compendium of Expression Profiles. *Cell* 2000; **102**: 109–126.

25 Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM *et al.* A Gene Expression Map for Caenorhabditis elegans. *Science* 2001; **293**: 2087–2092.

26 Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* 2003; **302**: 249–255.

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

28 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.

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

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

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

32 Graeber MB, Streit WJ. Microglia: Biology and pathology. *Acta Neuropathologica* 2010; **119**: 89–105.

33 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.

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

35 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.

36 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.

37 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.

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

39 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.

40 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.

41 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.

42 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.

43 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.

44 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.

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

46 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.

47 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.

48 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.

49 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.

50 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.

51 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.

52 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.

53 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.

54 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.

55 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.

56 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.

57 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.

58 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.

59 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.

60 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.

61 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.

62 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
Cholin						✓	2
DentateGranule		✓					1
Dopaminergic		✓					2
Ependymal	✓						1
GabaPV			✓				3
GabaReln			✓				1
GabaRelnCalb			✓				1
GabaSSTReln			✓				1
GabaVIPReln			✓				1
Gluta			✓				1
Golgi						✓	1
Hypocretinergic						✓	1
Microglia	✓						1
MotorCholin						✓	1
Oligo				✓		✓	4
Purkinje		✓	✓			✓	6
PyramidalCorticoThalam			✓				1
Pyramidal_Glt_25d2						✓	1
Pyramidal_S100a10						✓	1
Pyramidal_Thy1			✓				1
Serotonergic						✓	1
Spiny						✓	4
Th_positive_LC			✓				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.

8 Figures

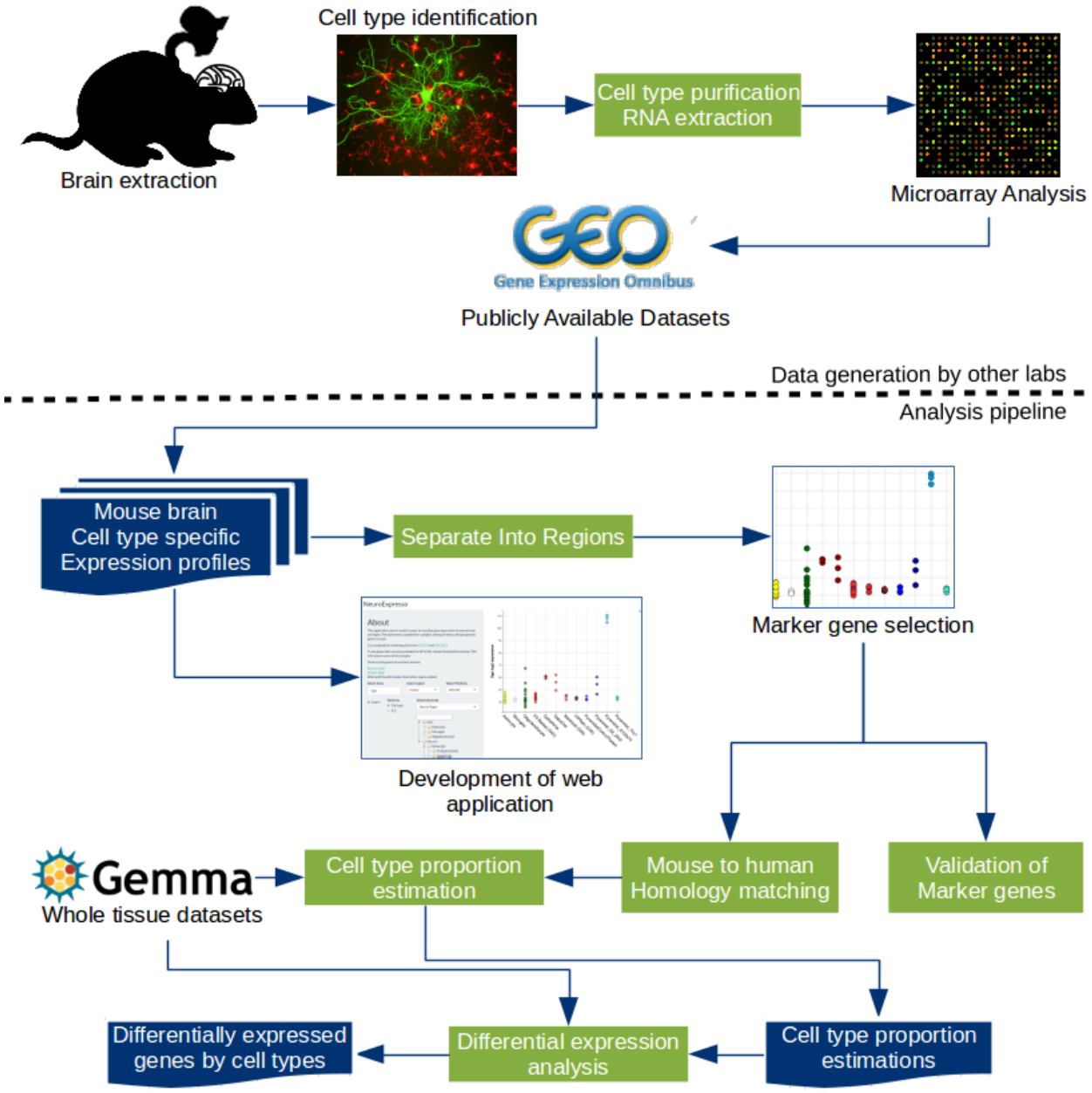

Figure 1: Workflow of the project

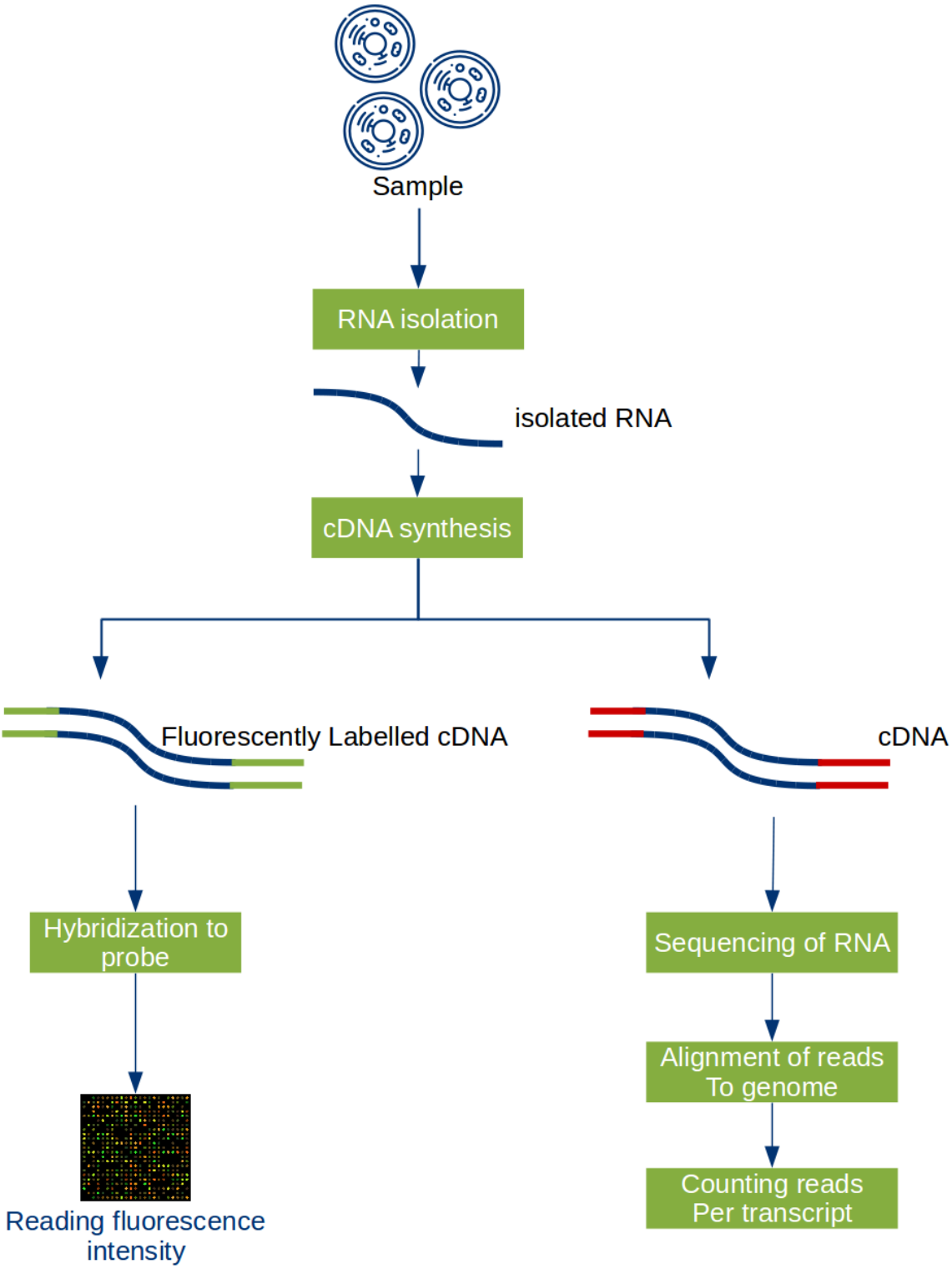

Figure 2: A short summary of microarray (right) and RNA sequencing (left) methods.

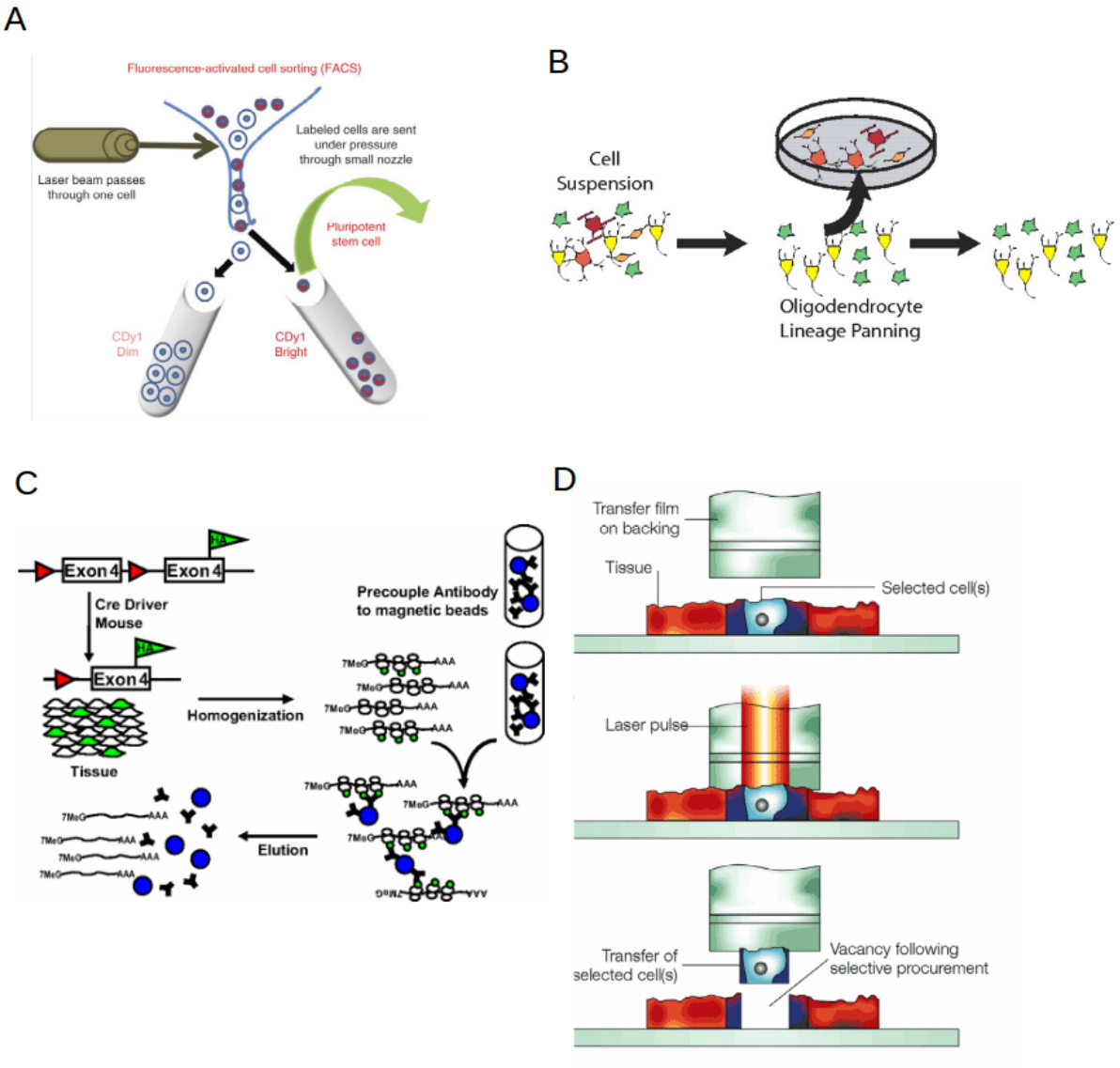

Figure 3: Example representations of cell type isolation techniques. A. Adapted from Kang et al. 2011. An example application of FACS. A fluorescently active molecule is used to label specific cell in the population. Upon detection of fluorescence, a charge is placed on the droplet whose path is later 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.

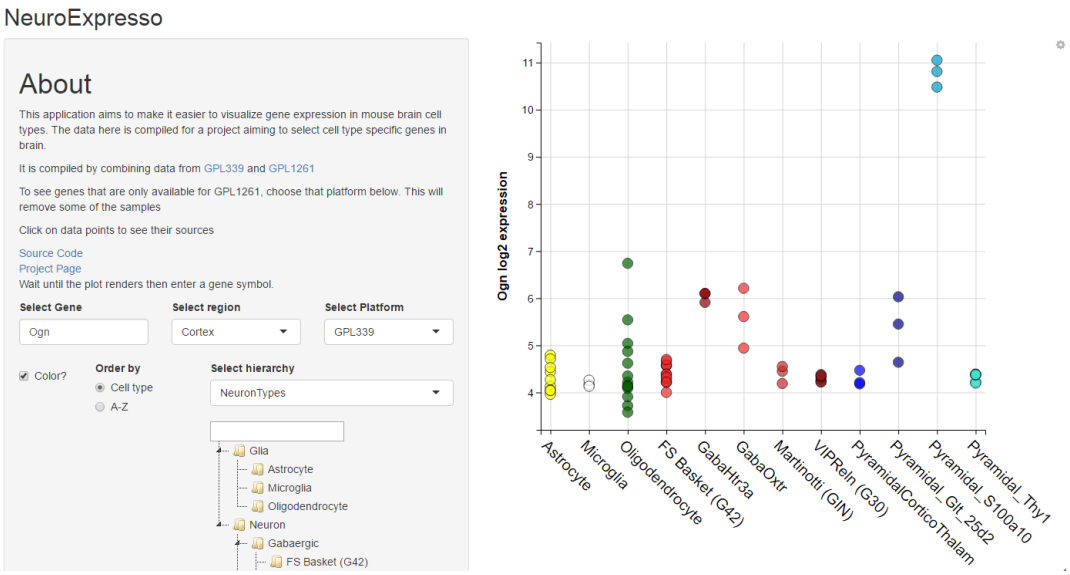

Figure 4: A screenshot of the NeuroExpresso web application: A tool to visualize gene expression in the cell types of our database.

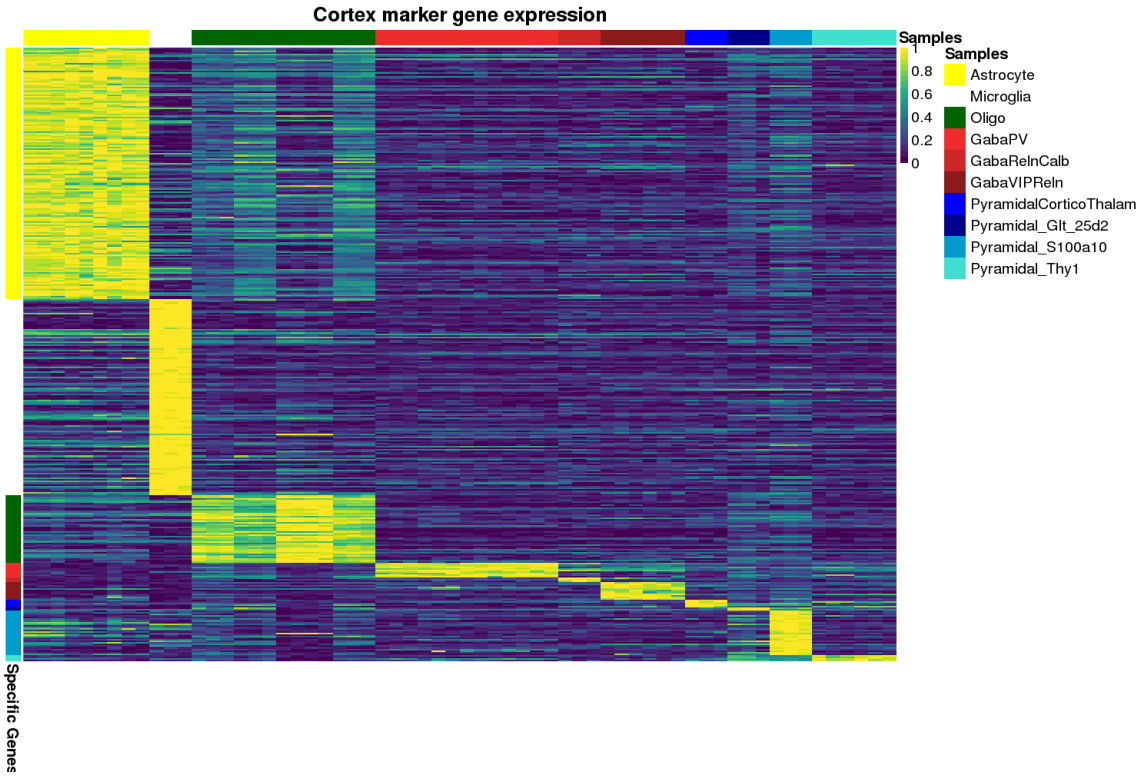

Figure 5: Expression of marker genes detected from cortex cell types. Values are scaled to be between 0 and 1, 0 representing the lowest observed expression level for the gene while 1 representing the highest. Samples and genes follow the same order of cell types to emphasize the specificity of the selected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

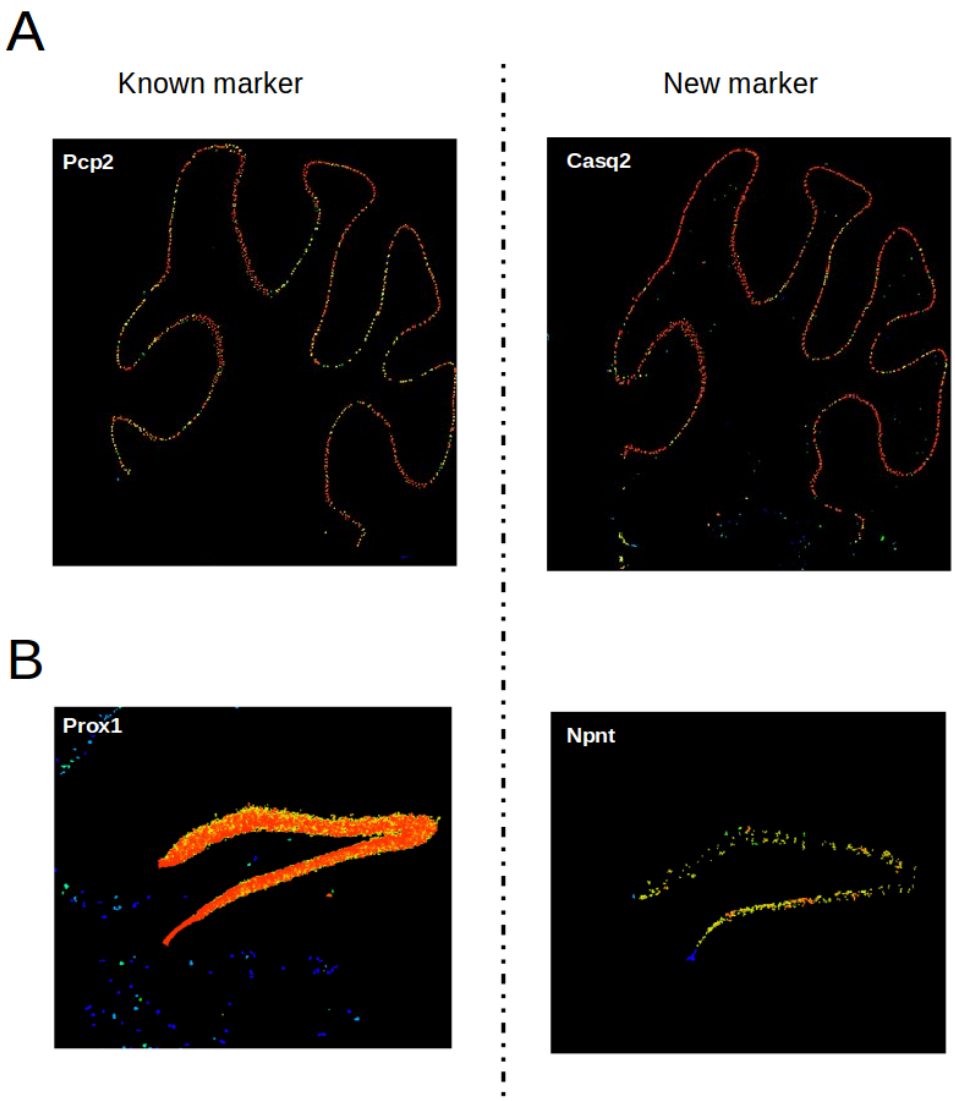

Figure 6: Expression of known marker genes and newly discovered marker genes in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. A. Expression of new and known markers of purkinje cells in cerebellum. B. Expression of new and known markers of granule cells in dentate gyrus,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

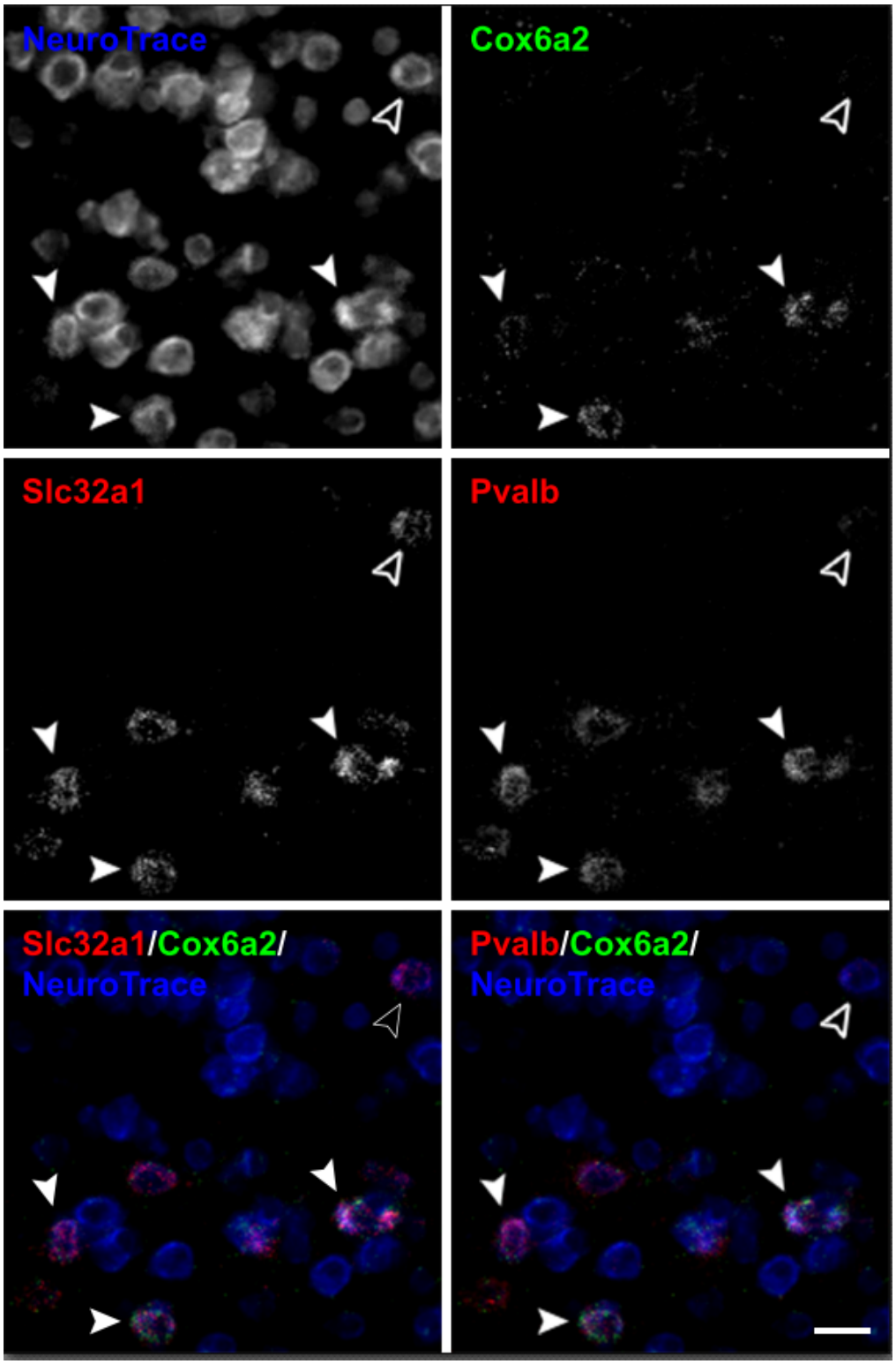

Figure 7: Triple labeling of fast spiking gabaergic cells in mouse cortex. NeuroTrace is a general neuronal marker. Slc32a1 and Pvalb are known markers of fast spiking pyramidal genes. Cox6a2 is a marker gene discovered through our analysis. Last row shows superimposition of known markers and Cox6a2 which appear in the same cells.

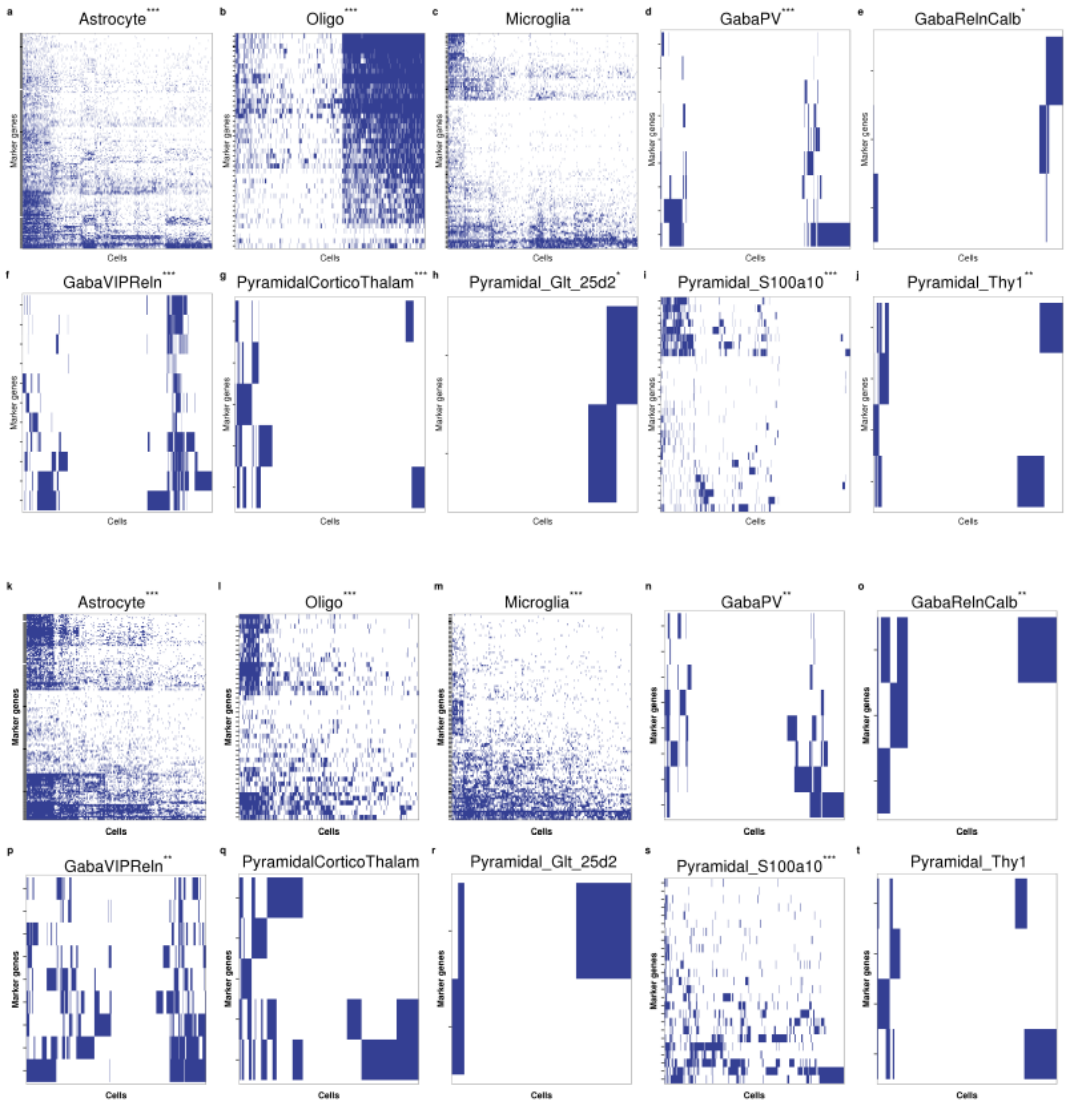

Figure 8: Binary heatmaps representing the expression of marker genes in single human and mouse cells. Significance stars represent the difference between coexistance of the genes and randomly selected gene sets with similar prevelance in the dataset. a-j shows the expression of marker genes in mouse 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.

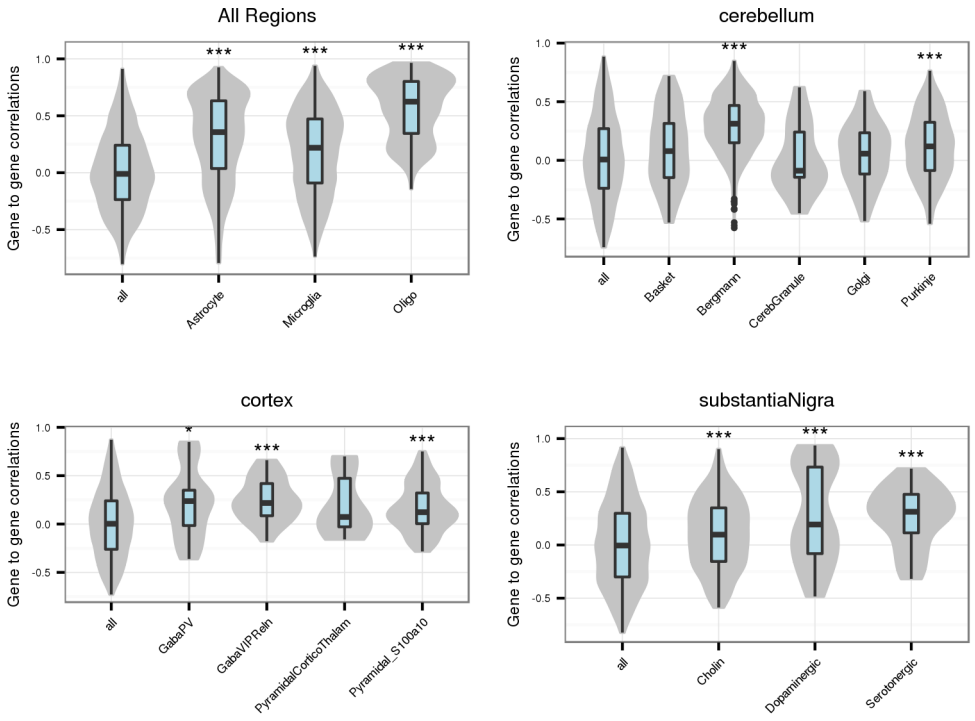

Figure 9: In between coexpression levels of marker gene sets. Significance markers show significantly higher co-expression than co-expression between all genes

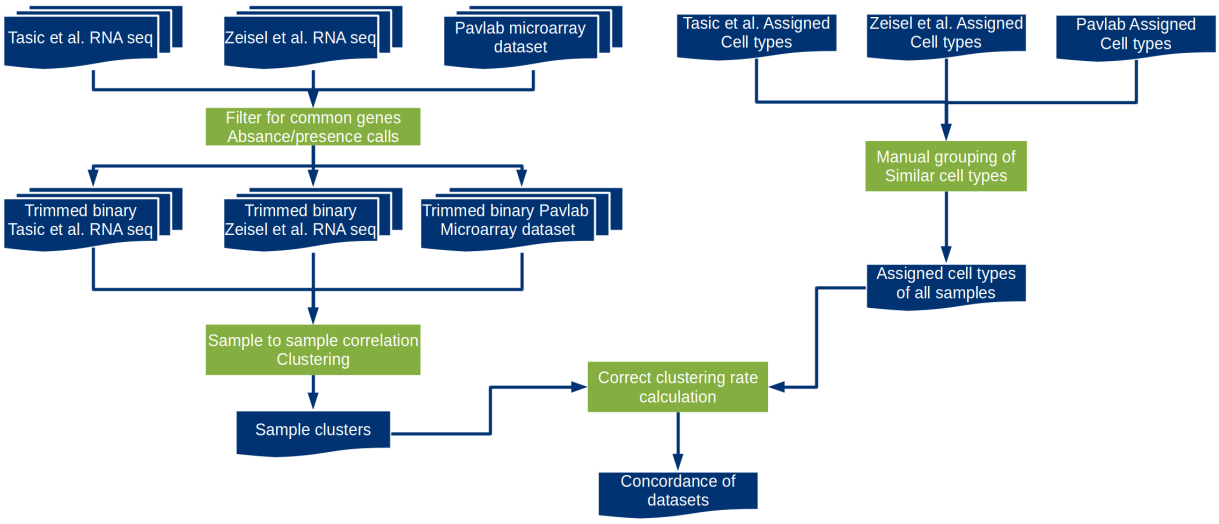

Figure 10: Pipeline for the upcoming analyisis on concordance of different cell type based analysis studies.

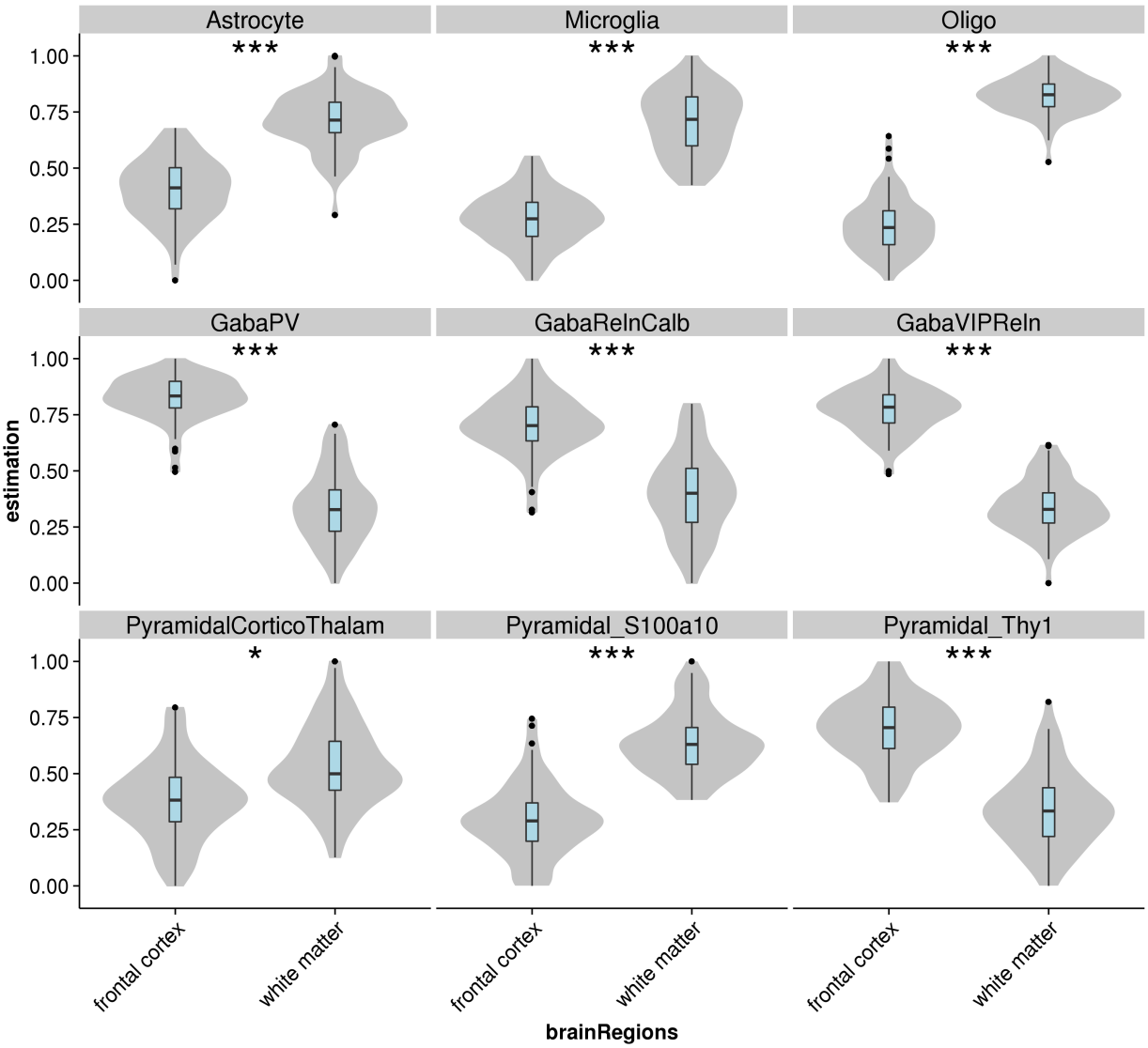

Figure 11: Estimations of cortical cell types in frontal cortex and white matter. Values are normalized to be between 0 and 1. Estimations appropriately reflect expected differences between white and gray matter for the most part. It is also possible to see some unexpected increase of some pyramidal subtypes.

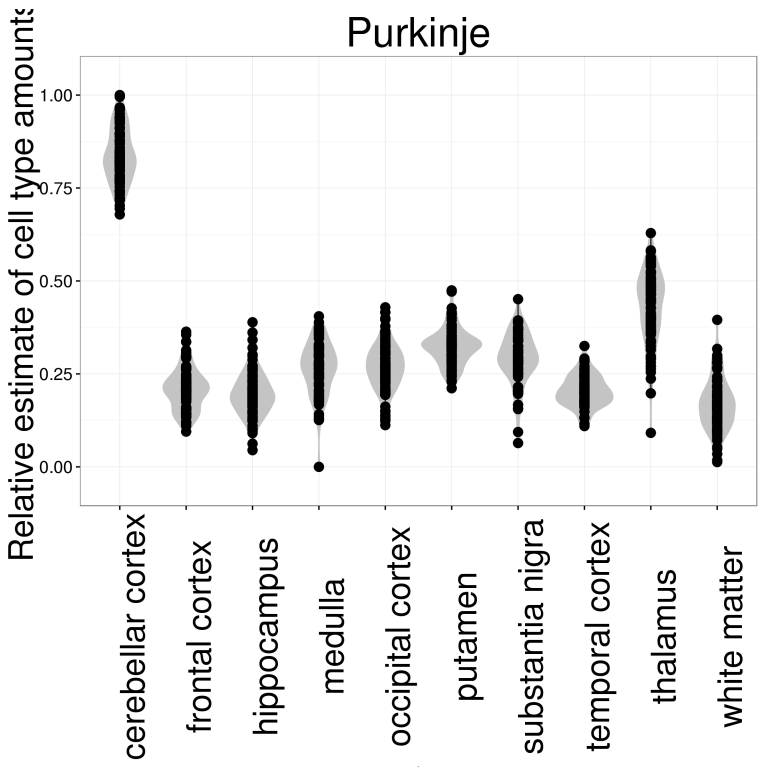

Figure 12: Estimations of purkinje cells in different brain regions. Values are normalized to be between 0 and 1. Purkinje cells are specific to the cerebellum.

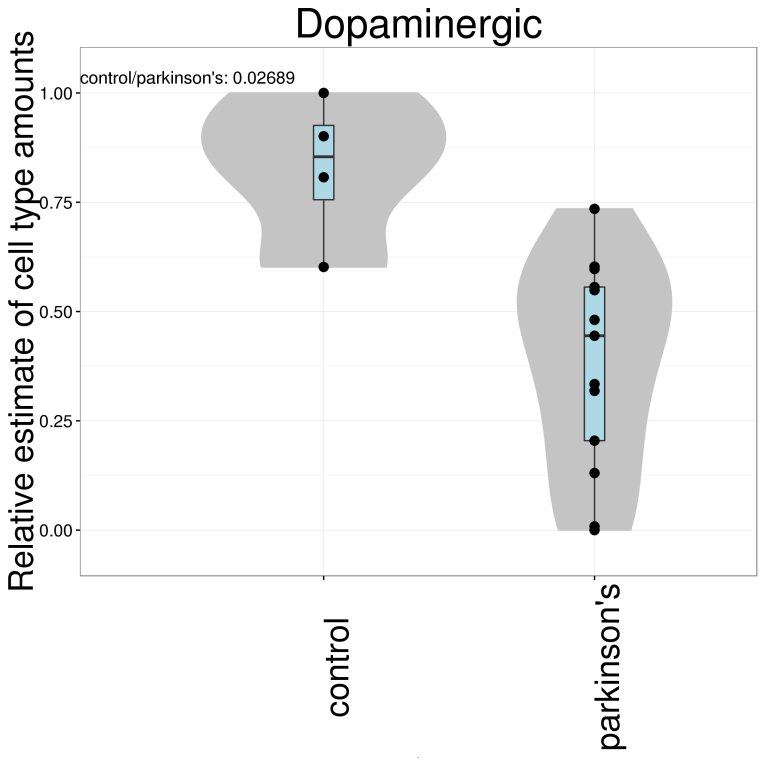

Figure 13: Estimations of dopaminergic cells in different substantia nigra of male parkinson's disease patients. Values are normalized to be between 0 and 1. Dopaminergic cell loss is an expected consequence of Parkinson's Disease

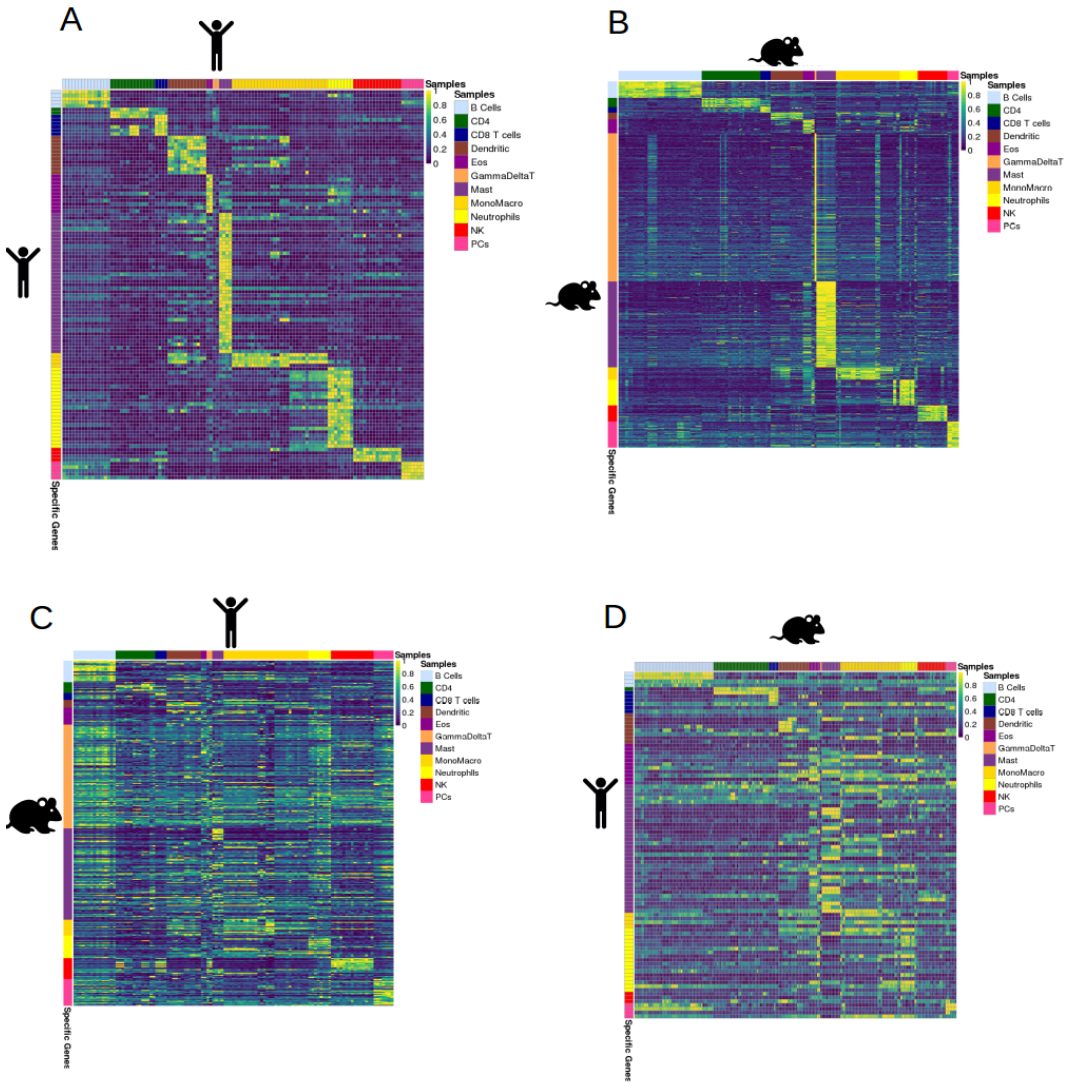

Figure 14: A-B. Expression of the genes selected from a species in the samples used for isolation from the same species. A shows human genes in human cell type specific expression profile dataset while B is mouse genes in mouse cell type specific expression profile dataset. C-D. Expression of homologues 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.

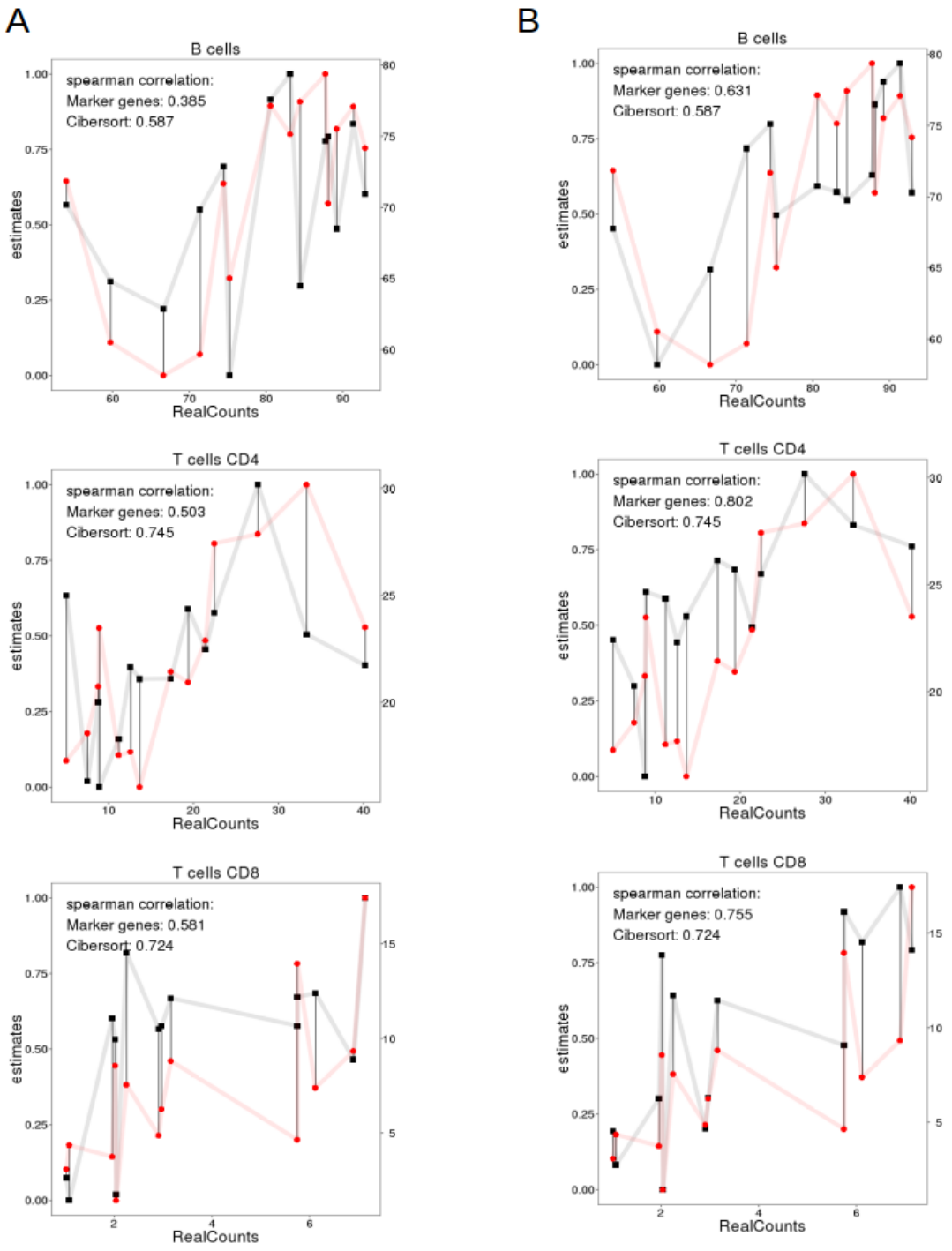

Figure 15: Estimations done by our method (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. A. Estimations done using marker genes selected from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

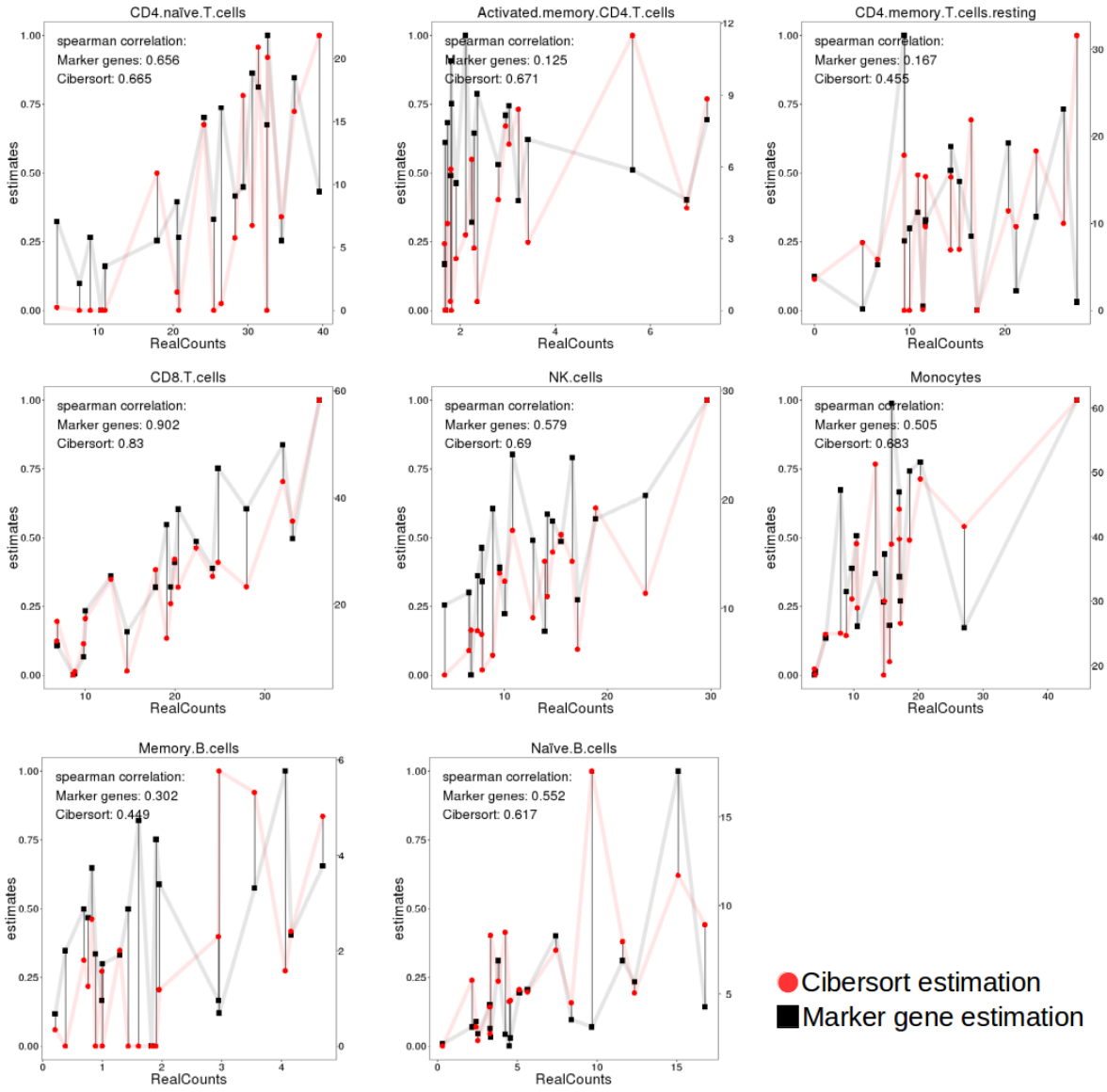

Figure 16: Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is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

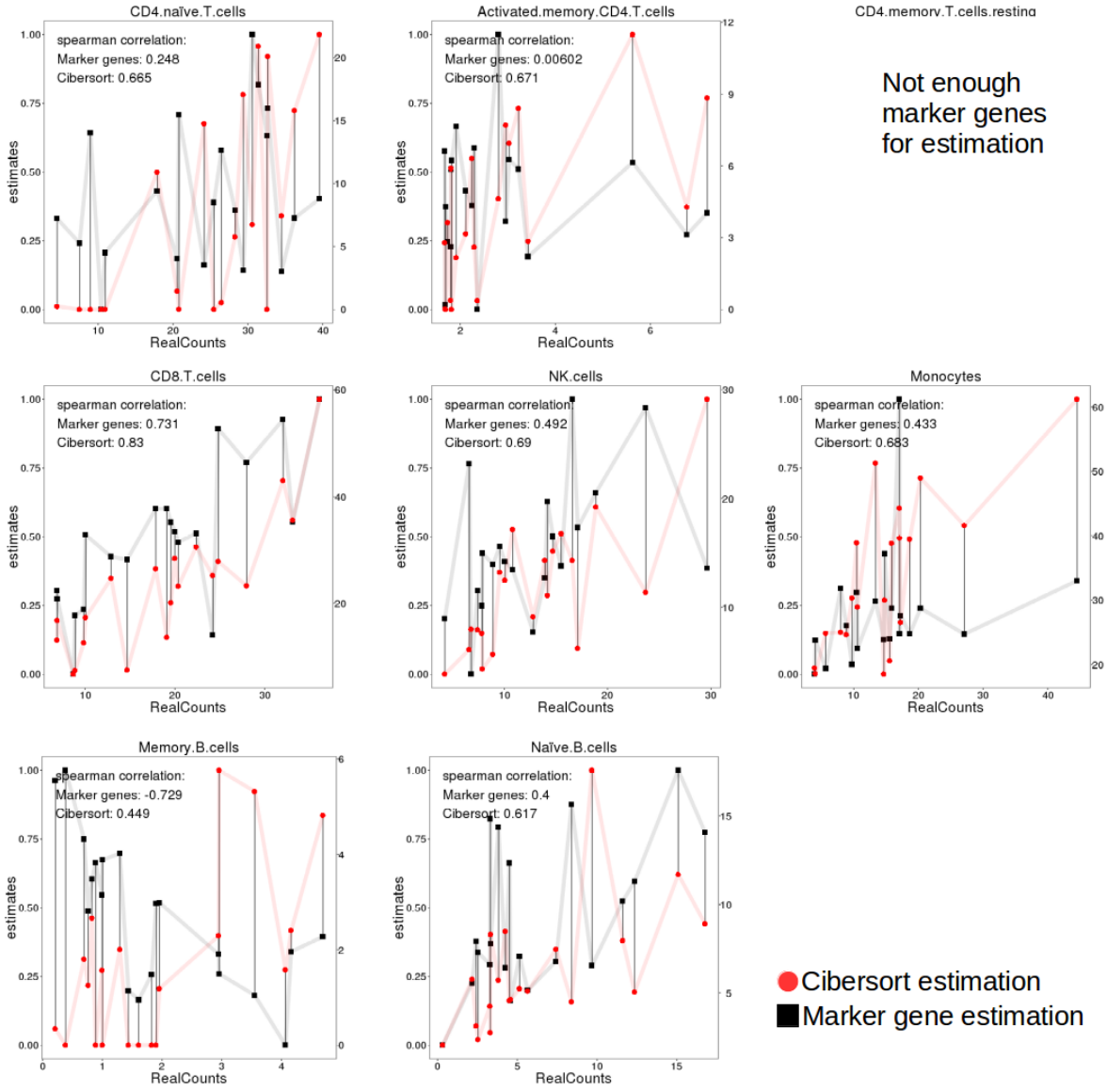

Figure 17: Estimations of finer subtypes done by our method using marker genes selected from human samples (black) and Cibersory (red) are plotted against the real cell counts from the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is 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.