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

lilah: what you saic here doesn't make much sense. regulation obviously matters since it wil introduce large errors in the matrix e. RNA seq and microarray signals do not necesarily follow the same distributions

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 either attempt to observe how a single cell type is effected from a diseases[5,6] and conditions [7,8], or they attempt to characterize cell types by finding their unique properties amongst their peers[9,10]. Such studies however are limited in their scope. It is prohibitevely expensive and difficulty to isolate all known cell types in order to see how they react to disease or what their characteristics are. Even studies which aim to create comprehensive libraries of brain expression profiles end up covering a limited portion of cell types by themselves[9,10]. The studies that focus on disease/condition effects on the other hand, are while useful at detecting changes to gene expression within the cell, are not useful if there are changes in cell type proportions (eg. as a result of neurodegeneration or astrogliosis.) unless cells are counted which adds to the cost and difficulty of the experiment and omitted from most studies.

Condition means normal stuff like development or sleep. Would accept a recommendation for a better word.

For this work, we formed a highly inclusive dataset of cell type specific expression profiles from a variaty of resources to help with the proplems 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 **1)** as tools to isolate specific cell types with greater ease, and **2)** as potential clues to the functional characteristics of given cells. Discovery of new markers are important since not all neuron types have their own unique markers, and if the number of markers are low, it becomes probable for them to be regulated under different conditions, making the use of available marker genes impossible[11]. 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: Enumaration 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].

The source is still the best one I could find that describes array building which are not effected by the fact that is is two color or not

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 metholodology 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] (Figure 3).

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 metholodology 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 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. Their most straightforward use is to enable researchers to differentiate between the cell type and the rest of the tissue, thus enabling further research on the cell type . A researcher can use RNA probes or protein antibodies to uniquely label a cell type[27]. Knowing a gene to be specific also allows researchers to genetically modify the cell and add foreign sequences that will only be expressed in the specific cell type.

Aside from their uses in the wet-lab, marker genes are also powerful tools in computational settings. They are often use as features in deconvolution of complex tissue samples as explained the the following sections.

3.3 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 4 A)[28]. 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 4 B)[29]. 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 flourescent 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 4 C)[30,31]. 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 4 D)[32].

The resulting samples from each of these methods has varying purity and data quality. TRAP stands out with having the worst record in purity while PAN seems to be inducing the most level of stress to the cells[33].

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 below

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

where X_{ij} is the expression value from a complex sample for genes j and sample i 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[4,34,35].

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[36]. Later, more in depth deconvolution is performed in human cortex and cerebellum that estimates cerebellar neuron types seperately while leaving cortical neurons as a single group[37]. Deconvolution of human brains is difficult due to the absence of human cell type specific expression profiles from human brain cell types as deconvolution is often aided by such expression profile data. Mouse brain on the other hand doesn't suffer this drawback. A reference based deconvolution (see below) of 64 distinct cell types was performed on whole tissue expression profiles accross various brain regions[38], though estimations of cell types in different brain regions are reported to fit the literature, no attempt was done to deconvolute cell types in samples from the same regions but under different conditions (disease models, developmental stages, eg).

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. Most naively, researchers try to estimate the W (matrix of cell type proportions) in solving the equation 1 by minimizing the e (error)[38]. Without any feature selection, this requires the expression profiles at hand to be a very good match to actual expression of the cell types in the mixture. This assumption might not be true due to differences between the cell type dataset and mixed tissue dataset such as gene regulation, incidence of noise or or changes in the RNA quantification method. Also potential effects of cell to cell interaction to gene expression in whole tissue samples can contribute to inacuracy of deconvolution. 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[35].

lilah: what you sai here doesn't make much sense. regulation obviously matters since it wil introduce large errors in the matrix e. RNA seq and microarray signals do not necesarily follow the same distributions so it is harder to make linear relationships between them. Even having different microarra platforms is problematic since affinity changes individual genes will mess your results up. I would definitely no say unkown cells is the main source of error here since they don't violate the assumptions of the model. everything else does.

3.4.2 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[4,33,34,37]. 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[4,37,39]. 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.

3.5 Major cell types of the brain

Brain hosts a variety of cell types that in some cases highly similar to each other while in others highly differentiated. Neurons and glias, the two main classes of cells are very different from each other in both function and morphology. Whereas compared to highly differentiated astrocyte cell types, neuron types tend to be more similar to each other.

3.5.1 Glia

- **Astrocytes** are star shaped cells present throughout the brain[40]. They highly outnumber neurons and 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[40]. More recently, it was also discovered that astrocytes play roles in synapse function by neurotransmitter release in response to changes in neuronal activity[40]. Astrocytes often proliferate in diseased brain[40]. This increase is often coupled with upregulation of certain genes, GFAP which is often used as the go to astrocyte marker by neurologists[40].

- **Oligodendrocytes** are responsible for forming the mylein sheet around neurons to insulate their axons[41]. A single oligodendrocyte ensheats multiple neuronal cells as a result of a highly coordinated process effected by axon size, neuronal activity and molecular signalling[41]. Most myelination occurs during early differentiation process. Oligodendrocytes are often susceptible to cell death due to oxidative damage due their high metabolic rates[41].

- **Microglia** are differ from other glia in terms of their place of origin. Resting microglia has functions in detection of damaged synapses and removal of damaged cells. They are also the antigen presenting cells of the brain. Microglia switch to activated stage as a result of stress or damage with distinct morphology and expression pattern. Persistent activation of microglia can result in degeneration of both microglia and neuronal cells. Hence microglia is often found to be associated with nervous system disorders[42].

3.5.2 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[43]. 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[43].

In a neurological disease, often a subset of these cells are effected. Parkinson's disease for instance is known for dopaminergic cell loss it causes in substantia nigra[44].

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 of non-overlapping cell types. The database is a valuable resource since it allows comparison of all available cell types to each other, allowing us to find specific properties. 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.

4.1 Data acquisition and preprocessing

The bulk of the dataset is based on a previous compilation made by Okaty et al.[33] 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 probesets is straightforward to combine since 430A array is a subset of 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. By the use of a custom script, samples from 430.2 array are 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[45,46]. Due to the fact that the database included samples from a large number of datasets, RMA normalization, that performs quantile normalization at a probeset level still resulted in a observable asymmetry in probeset level signal distribution (Figure 5). 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. To equalize the signals based on the common assumption of equal amount of total RNA in between samples, we used quantile normalization[47] to make the samples comparable to each other (Figure 6). 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 are used to isolate the cell type) and making sure samples are not contaminated by other cell types by looking for expression of foreign markers. At the end of the cleanup process, cell types are separated into non overlapping groups which again lead to removal of samples representing multiple cell types from the main analysis. The resulting dataset has 25 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 dataset, we created a web application to facilitate other researchers' access to the dataset. The web application allows third parties to easily visualize expression of chosen genes in individual cell types in their respective regions (Figure 7). 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. We are also planning to embed an enrichment tool that will enable researchers to check their hitlists for cell type specific enrichment (see below) and do quick differential expression analyses between cell types. The application will increase 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 cortex is taken as a single region even though many samples are taken from specific regions of cortex since most biological samples of whole tissue are taken from whole cortex, and such fine divisions will leave many cell types alone, making the marker gene selection process meaningless. Certain cell types were assigned to multiple regions other than regions they were isolated from because the the cell type is known to exist in said regions. 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.
- 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

We used our method instead of 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 10 regions from 25 cell types. Number of marker genes greatly vary from one cell type to another depending on the presence of closely related cell types in the dataset

(Figure 8).

5.3 Validation of marker genes

Finding marker genes using independent datasets is problematic due to potential differences in mouse strains and batch effects. To ensure the reliability of our genes, testing is required to make sure that they act as marker genes in biological and computational settings. It is uncertain if marker genes detected using mouse cell types will apply to human cell types.

5.3.1 Validation of marker genes via in situ hybridization

In situ hybridization (ISH), while expensive and time consuming is a reliable 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 will be using the ISH data available from the Allen Brain Atlas[48] 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 9[48] .

We are currently collaborating with another lab to validate the markers by dual labelling. We were able to validate Cox6a2 as a marker of fast spiking gabaergic cells in mice. We will be expanding the number of validated genes through this method and potentially apply it to human samples as well.

5.3.2 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 single cell RNA-sequencing datasets 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. The studies attempt to define cell types from the ground up by using a variety of clustering methods. 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 genes with similar prevalence in the dataset. For most gene sets this gave favourable results. Since single cell expression analysis is still in it's infancy, often the transcript counts for most genes end up being really 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.

We used this method in single cell RNA-seq datasets from mouse[12] and human[22] and in both cases, we were able to see enriched coexpression in a majority of our gene sets ((Figure 10)). We will later add a more recently published RNA-seq study done on mouse brains that isolates single cells in a more cell type specific manner[13].

5.3.3 Validation of marker genes in human whole tissue data

As another validation method, we analysed the coexpression of marker gene sets in whole tissue data in a dataset[49] containing tissues that houses 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 coexpression in between these samples to coexpression levels of randomly selected genes to see if it is higher. The results (Figure 11) were comparable to the ones we got from human RNA-seq data.

5.4 Asses 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 captured 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 12.

6 Aim 3: Enumaration of cell type proportions

6.1 Enumeration 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 is not very reliable due to batch effects that is potentially present in the data. Marker genes on the other hand are less sensitive to fine changes in expression in the reference datasets due to out 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 is done by taking their first principle component in individual samples. The idea 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 (removal of genes with negative contribution to the 1st principle component, seeking consensus across experimental group to rule out differential regulation) 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 enumerated relative cell proportions in different brain regions with known cell type proportions and we got results that would be expected from a typical brain. For instance in a dataset of different brain regions from healthy donors[49] it was possible to observe an increase in glial population and decrease in most neuronal populations between white matter and grey matter (Figure 13) and purkinje cells were found to be exclusive to cerebellum (Figure 14). 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[44], which is a disease 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 15.

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. Current plans include analyzing an Alzheimer's disease cohort where samples across different brain regions are collected from patients and controls and seeking for potential confounds in healthy donors to asses the effect of factors such as aging and sex.

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

Enumeration of brain cell types poses problems due to unavailability of cell counts from the brain. Any result we find is unverifiable other than the expected differences between groups. To asses the real accuracy of the study, expression data from whole tissues that is paired with cell type counts is required. 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. 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. 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 were behaving like marker genes in the species they were not selected from (Figure 16). To see how lesser quality of marker genes effects proportion estimations we attempted to enumerate cell type proportions in whole blood cell types and compare our results with a recently published reference based enumeration method[35]. When the cell types are kept at a relatively general level, mouse genes returned better estimates than human genes

(Figure 17), potentially due to difference of quality between the datasets. Whereas attempting to enumerate finely defined cell types with mouse genes returned poor correlation to actual counts (data not shown).

Future work will focus on characterizing the genes that do not perform well across species and assessing if their properties are generalizible to brain cell type markers. With our current resources, it is impossible to tell if certain brain cell type markers are working or not. We will be attempting to characterize such genes by analyzing their individual performance at marker gene validation steps

6.3 Use enumeration information 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 often specific to cell types, having unaffected cell types in the sample will reduce the observed difference, making it harder to get significance.. Previous work shows that it is possible to increase the power of differential expression analysis by adding estimated cell type proportions as covariates [chikina_cellcode_2015]. They 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.

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 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.

3 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.

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 Speicher MR, Carter NP. The new cytogenetics: Blurring the boundaries with molecular biology. *Nature Reviews Genetics* 2005; **6**: 782–792.

28 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.

29 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.

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 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.

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

33 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.

34 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.

35 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.

36 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.

37 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.

38 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.

39 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.

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

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

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

43 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.

44 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.

45 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.

46 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.

47 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.

48 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.

49 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.

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

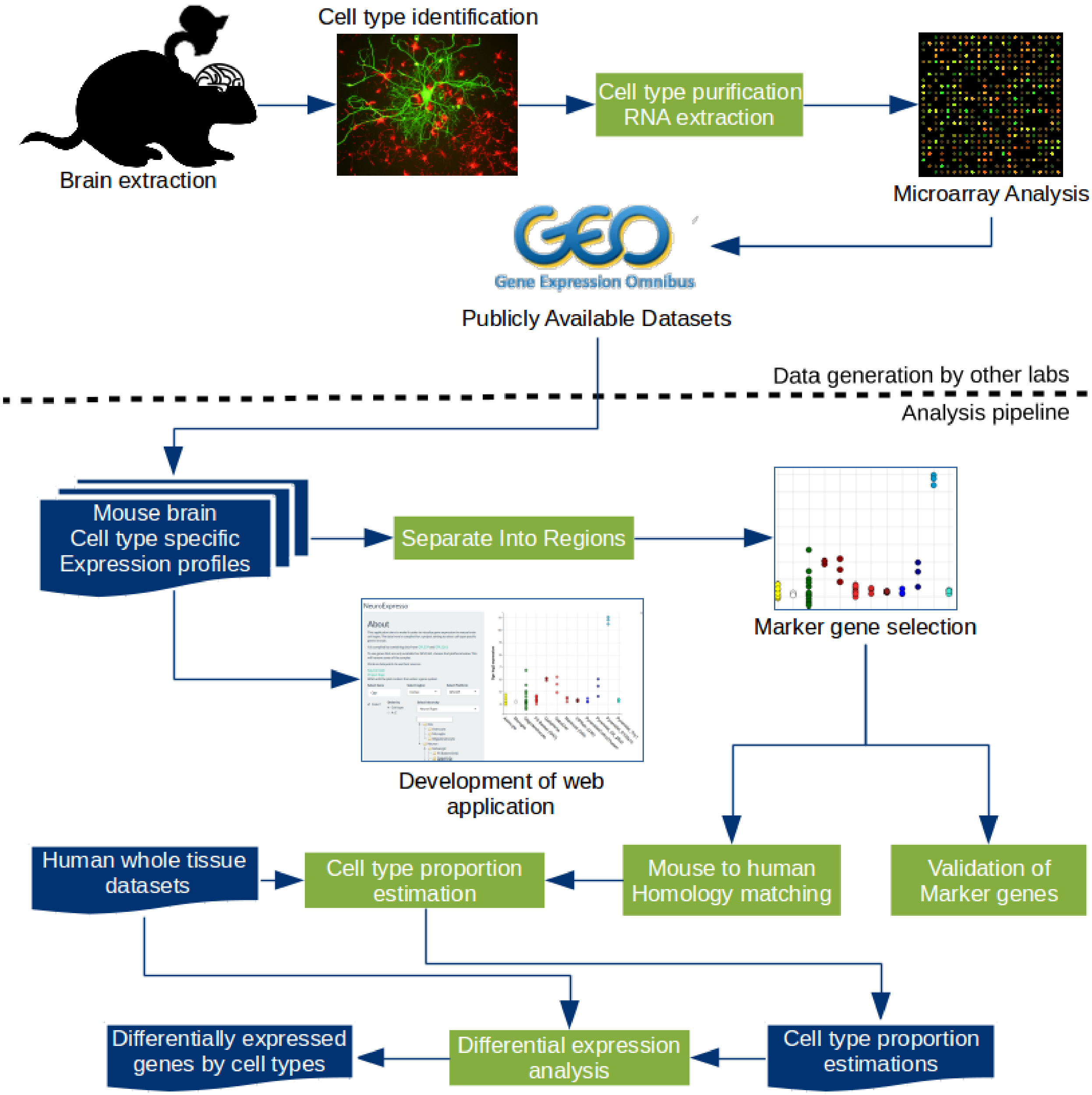

Brain extraction
Cell type identification
Cell type purification
RNA extraction
Microarray Analysis
Gene Expression Omnibus
Publicly Available Datasets
Data generation by other labs
Analysis pipeline
Mouse brain
Cell type specific
Expression profiles
Separate Into Regions
Marker gene selection
NeuroExpresso
About
Development of web application
Human whole tissue datasets
Cell type proportion estimation
Mouse to human Homology matching
Validation of Marker genes
Differentially expressed genes by cell types
Differential expression analysis
Cell type proportion estimations

Figure 1: Workflow of the project

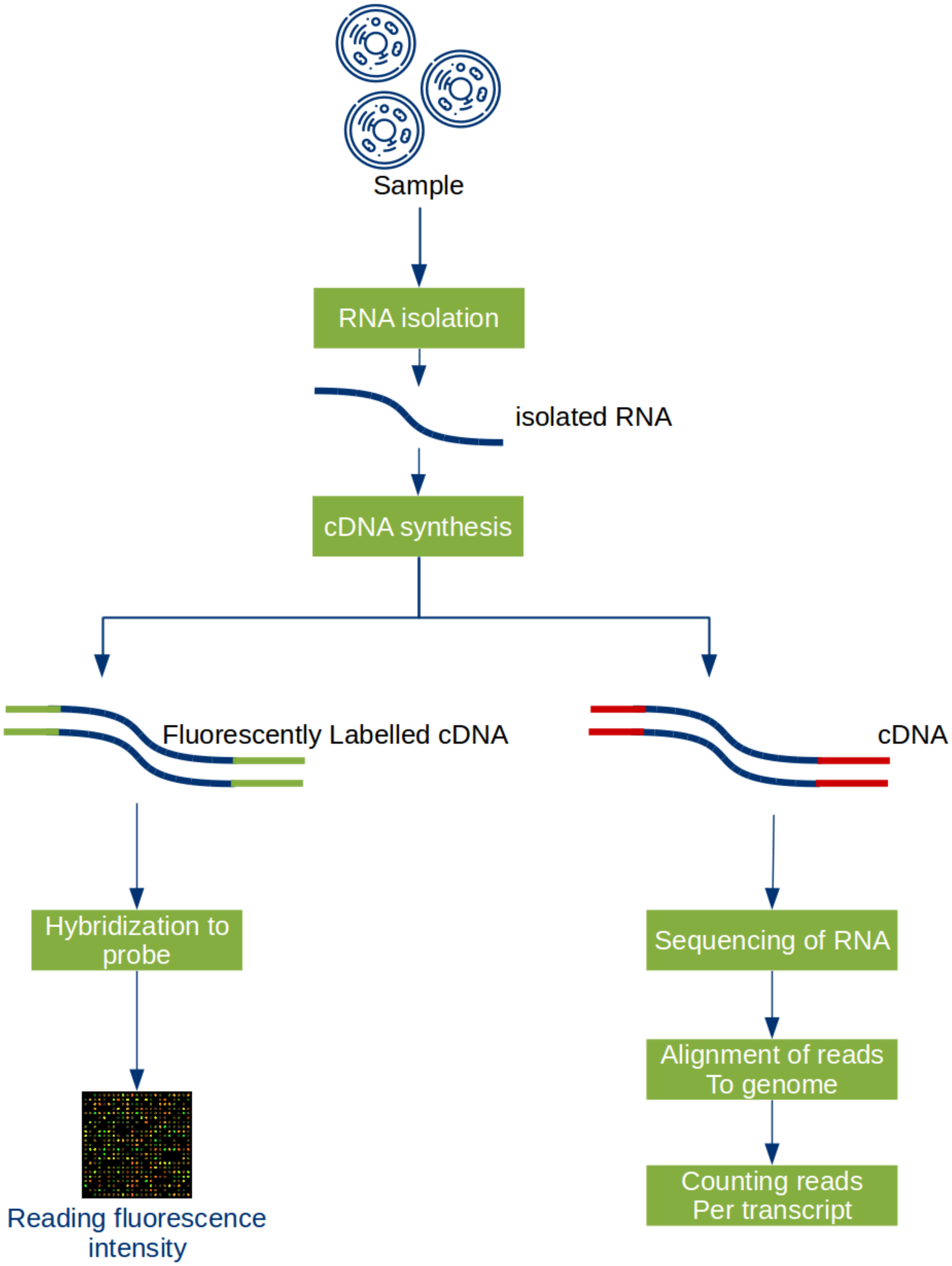

Sample
RNA isolation
isolated RNA
cDNA synthesis
Fluorescently Labelled cDNA
cDNA
Hybridization to probe
Sequencing of RNA
Alignment of reads To genome
Counting reads Per transcript
Reading fluorescence intensity

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

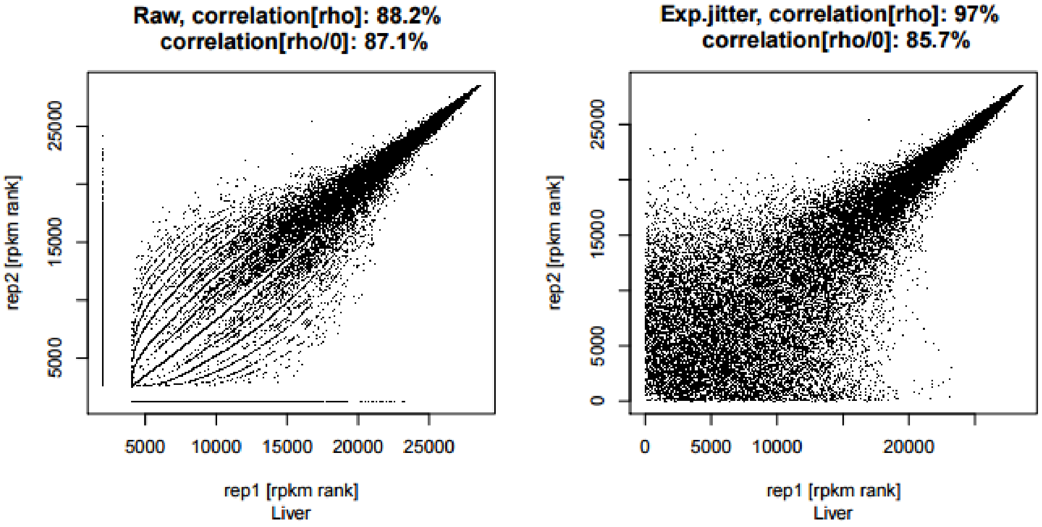


Figure 3: Taken from RNA sequencing shows low repeatability at low levels (Łabaj et al. 2011): Ranks of expression values for two indepent replicates are shown. Left figure left figure shows unmodified ranks. Right figure adds jitter to the points to reveal overplotting on the edges. The sample is taken from liver tissue

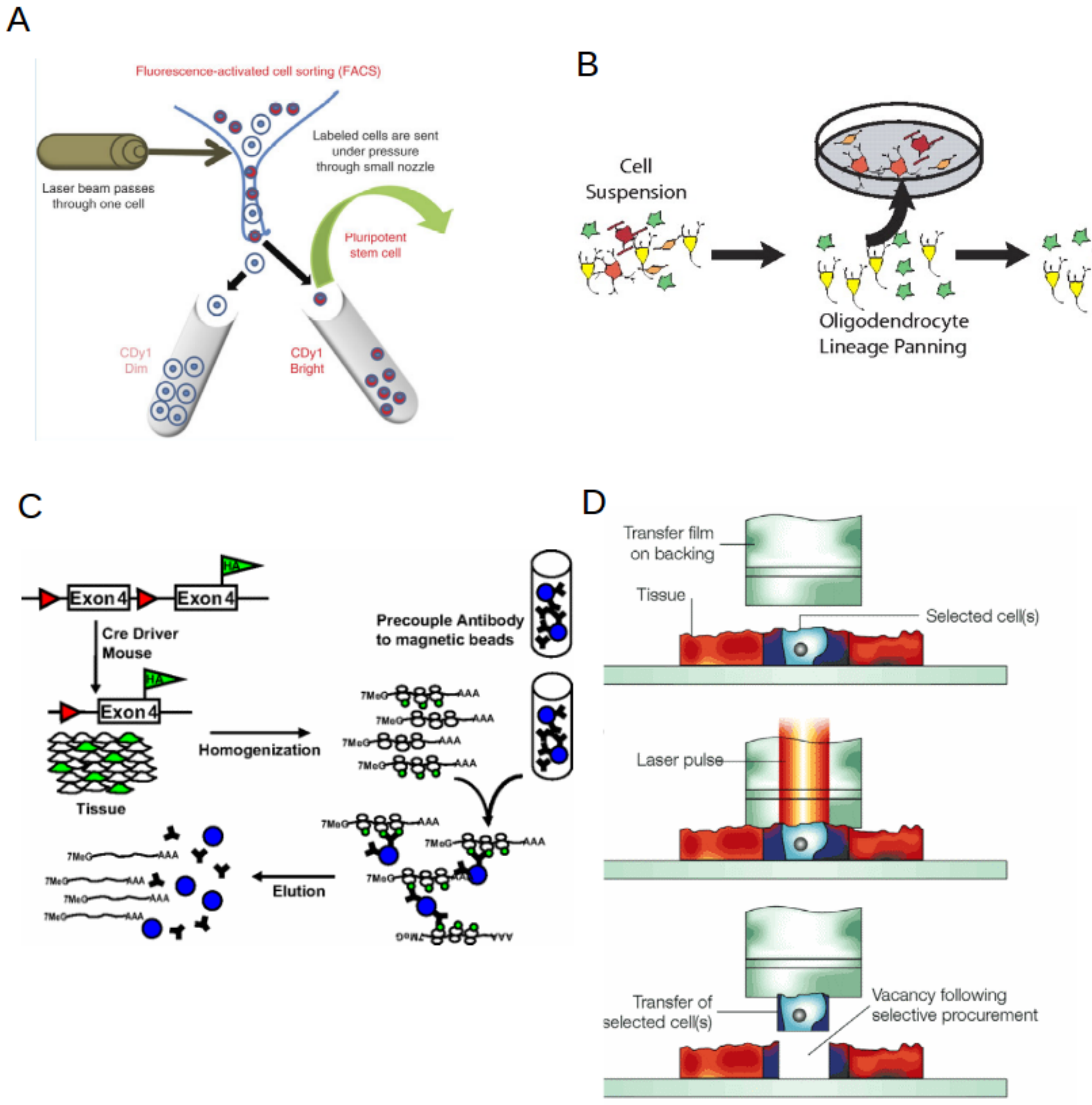

A
Fluorescence-activated cell sorting (FACS)
Laser beam passes through one cell
Labeled cells are sent under pressure through small nozzle
Pluripotent stem cell
CDy1 Dim
CDy1 Bright
B
Cell Suspension
Oligodendrocyte Lineage Panning
C
Exon 4
Exon 4
HA
Cre Driver Mouse
Exon 4
HA
Tissue
Homogenization
Precouple Antibody to magnetic beads
7MeG
AAA
Elution
D
Transfer film on backing
Tissue
Selected cell(s)
Laser pulse
Transfer of selected cell(s)
Vacancy following selective procurement

Figure 4: Example representations of cell type isolation techniques. A. Adapted from Kang et al. 2011. An example application of FACS. CDy1 positivity is used to isolate the target cell types, which are pluripotent stem cells. B. Adapted from Cahoy et al. 2008. Immunopanning is used to remove oligodendrocytes from the cell mixture. C. Adapted from Sanz et al. 2009. A schematic representation of the TRAP method. D. Adapted from Liotta et al. 2000. A schematic representation of LCM method.

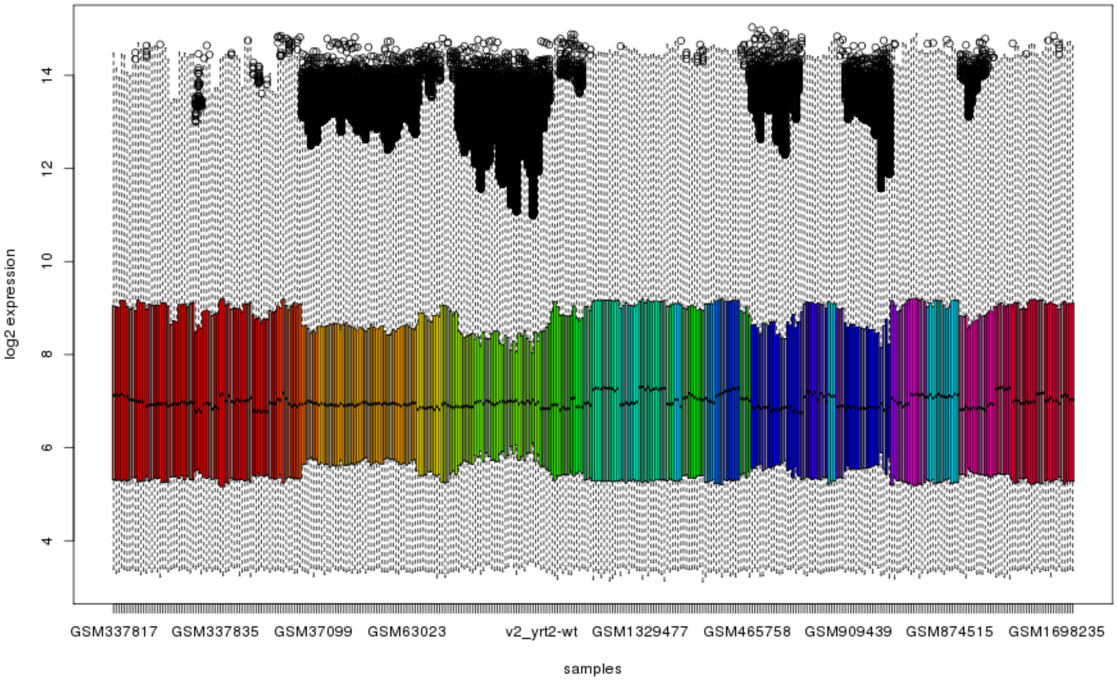


Figure 5: Distribution of expression levels in between samples before quantile normalization. Different colors represent different studies included in the dataset

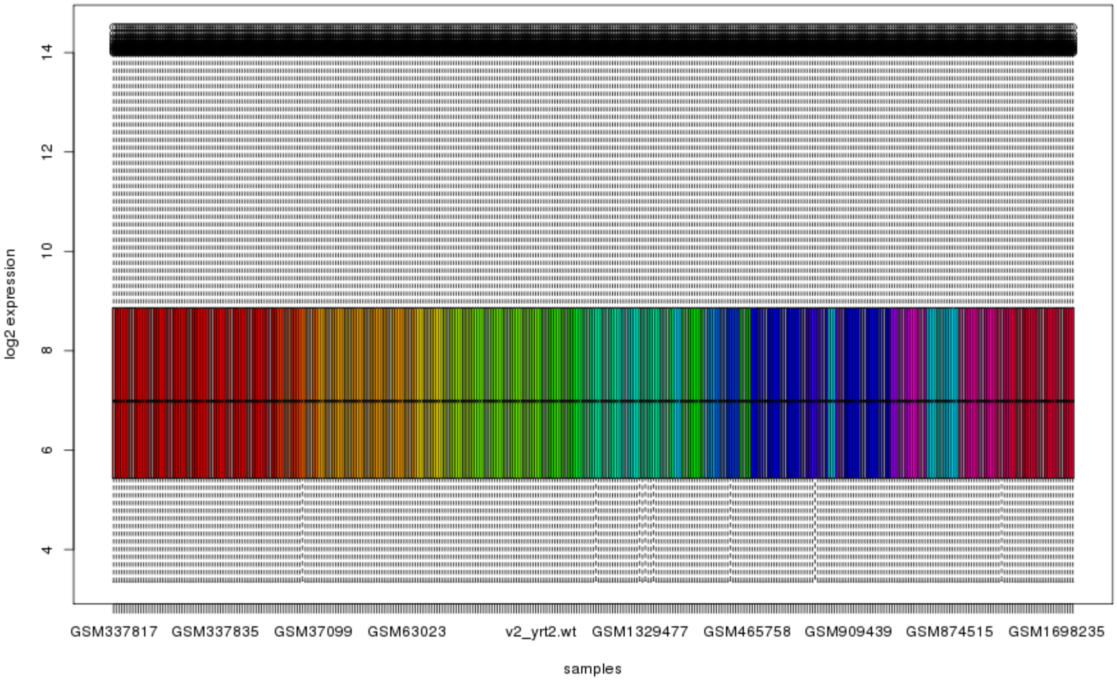
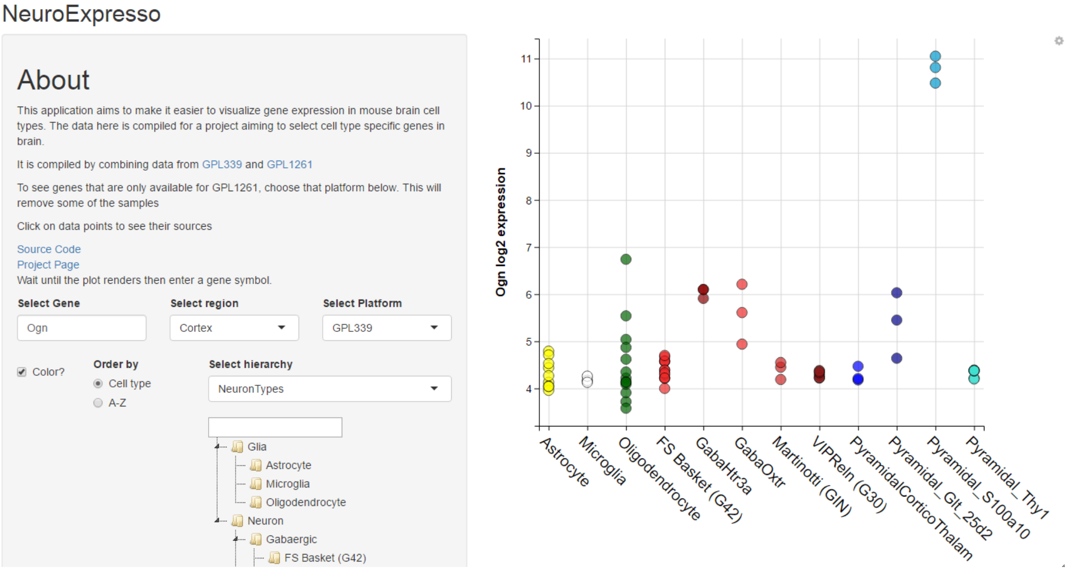


Figure 6: Distribution of expression levels in between samples after quantile normalization. Different colors represent different studies included in the dataset



NeuroExpresso

About

This application aims to make it easier to visualize gene expression in mouse brain cell types. The data here is compiled for a project aiming to select cell type specific genes in brain.

It is compiled by combining data from GPL339 and GPL1261

To see genes that are only available for GPL1261, choose that platform below. This will remove some of the samples

Click on data points to see their sources

Source Code
Project Page
Wait until the plot renders then enter a gene symbol.

Select Gene: Ogn

Select region: Cortex

Select Platform: GPL339

☑ Color?

Order by
● Cell type
○ A-Z

Select hierarchy: NeuronTypes

Figure 7: A screenshot of the NeuroExpresso web application

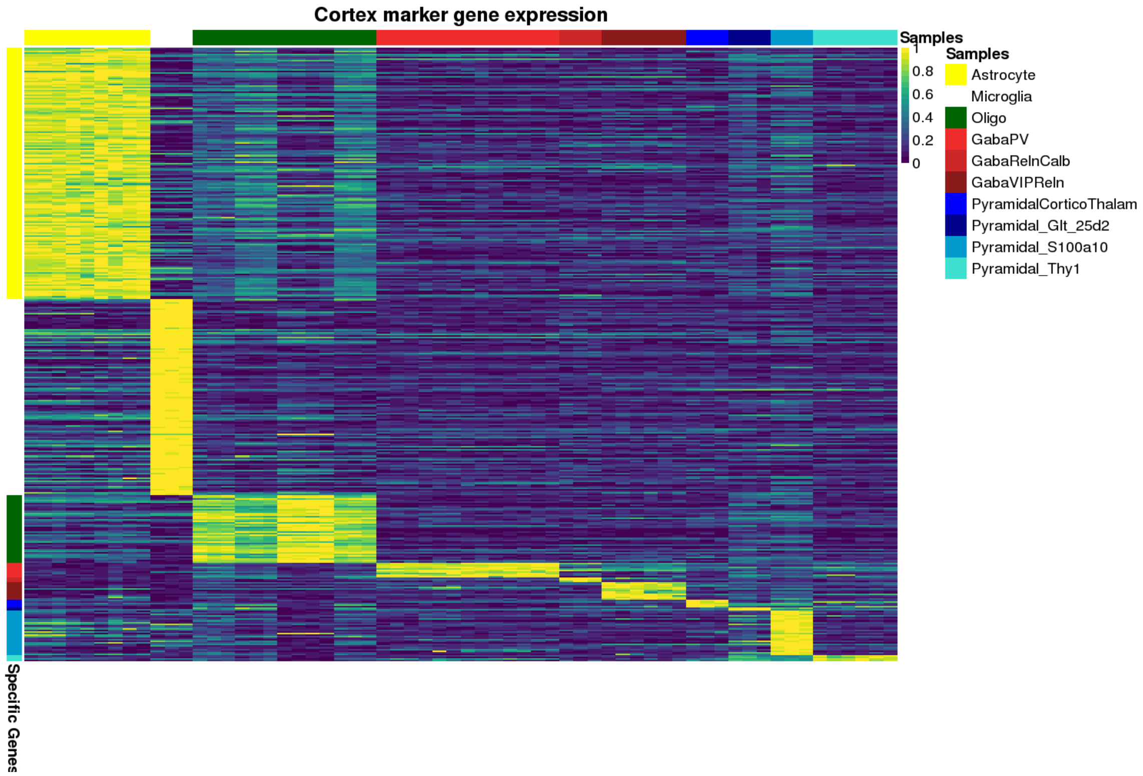


Figure 8: Expression of marker genes detected from cortex cell types. Values are scaled to be between 0 and 1, 0 representing the lowest observed expression level for the gene while 1 representing the highest. Samples and genes follow the same order of cell types to emphasize the specificity of the selected genes

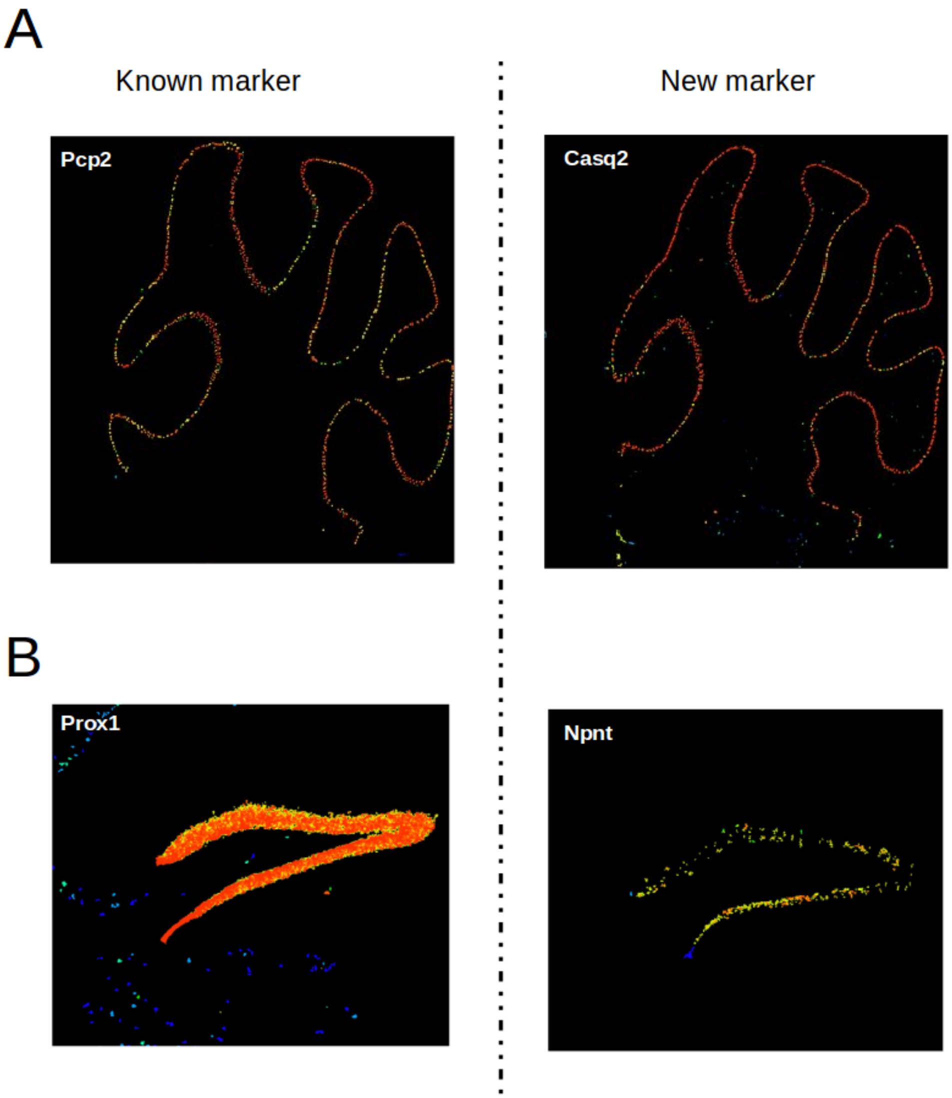


Figure 9: Expression of known marker genes and newly discovered marker genes in Allen Brain Atlas (Lein et al. 2007) mouse brain in situ hybridization database. A. Expression of new and known markers of purkinje cells in cerebellum. B. Expression of new and known markers of granule cells in dentate gyrus, granule cell layer

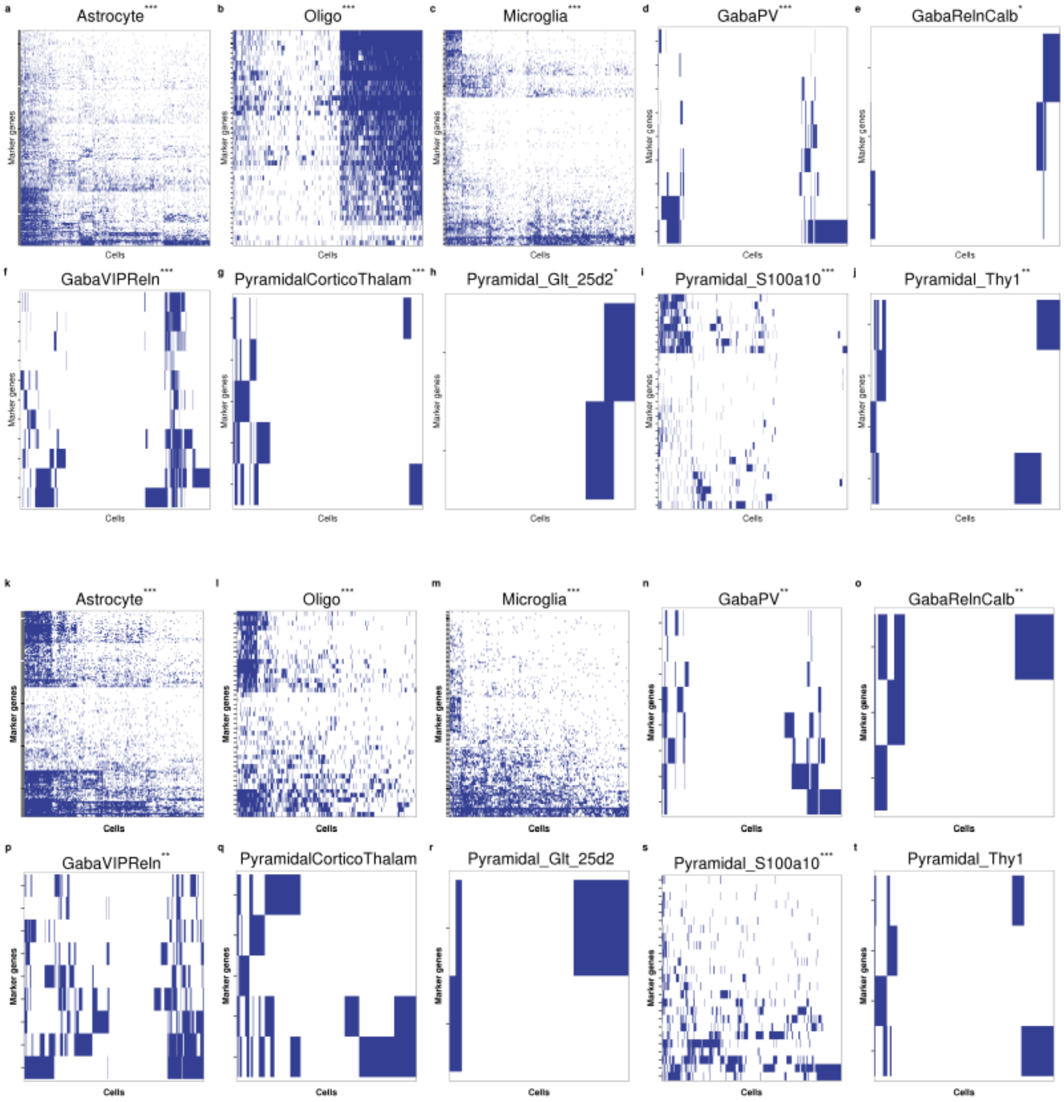


Figure 10: Binary heatmaps representing the expression of marker genes in single human and mouse cells. Significance stars represent the difference between coexistance of the genes and randomly selected gene sets with similar prevelance in the dataset. a-j shows the expression of marker genes in mouse 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.

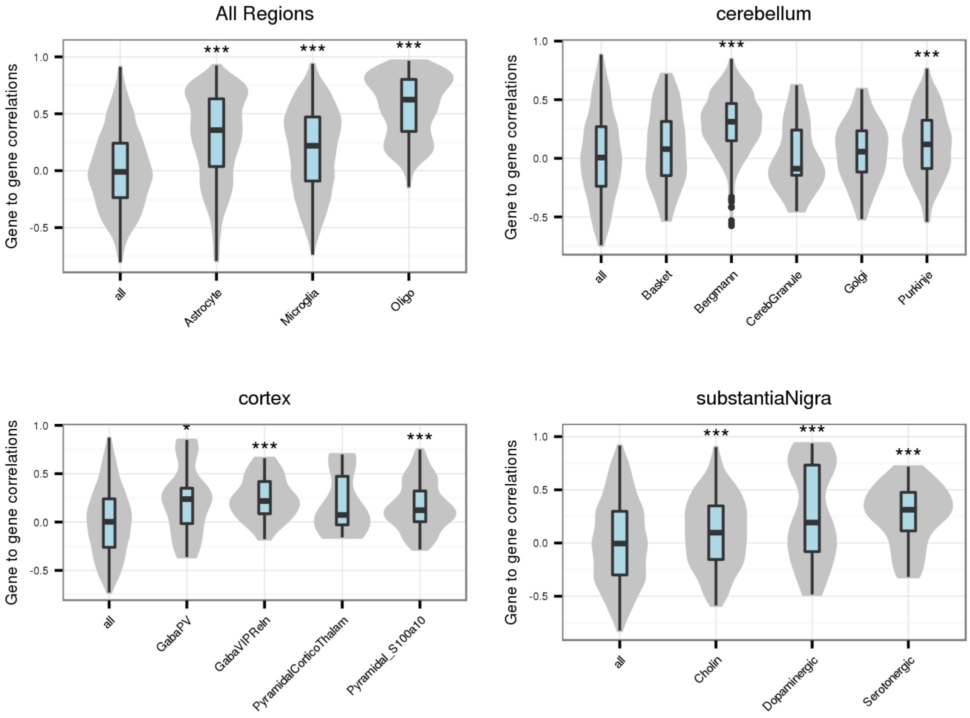


Figure 11: In between coexpression levels of marker gene sets. Significance markers show significantly higher co-expression than co-expression between all genes

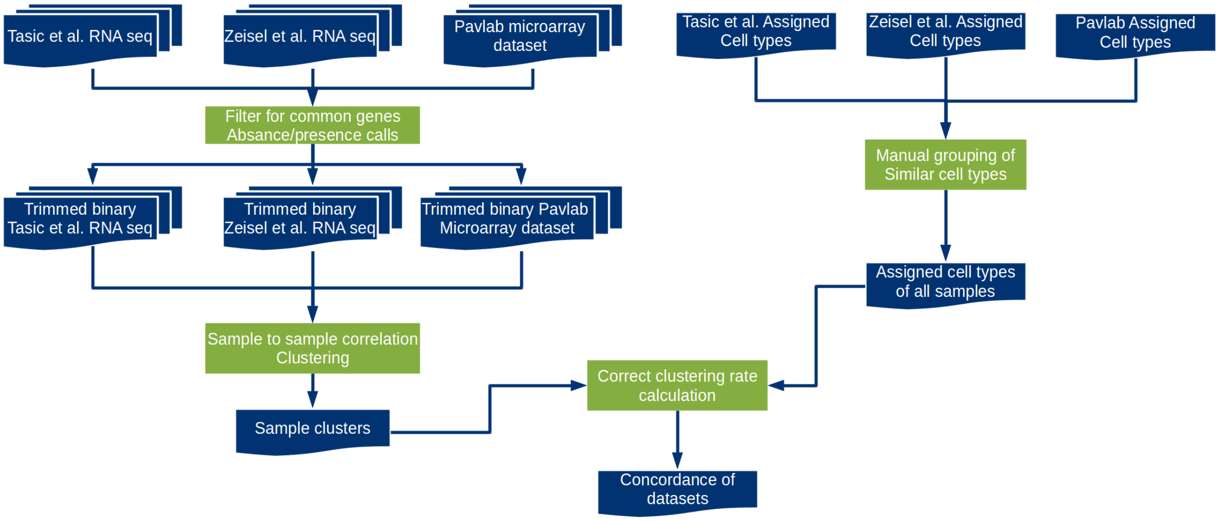

Tasic et al. RNA seq
Zeisel et al. RNA seq
Pavlab microarray dataset
Filter for common genes Absance/presence calls
Trimmed binary Tasic et al. RNA seq
Trimmed binary Zeisel et al. RNA seq
Trimmed binary Pavlab Microarray dataset
Sample to sample correlation Clustering
Sample clusters
Tasic et al. Assigned Cell types
Zeisel et al. Assigned Cell types
Pavlab Assigned Cell types
Manual grouping of Similar cell types
Assigned cell types of all samples
Correct clustering rate calculation
Concordance of datasets

Figure 12: Pipeline for the upcoming analiysis on concordance of different cell type based analysis studies.

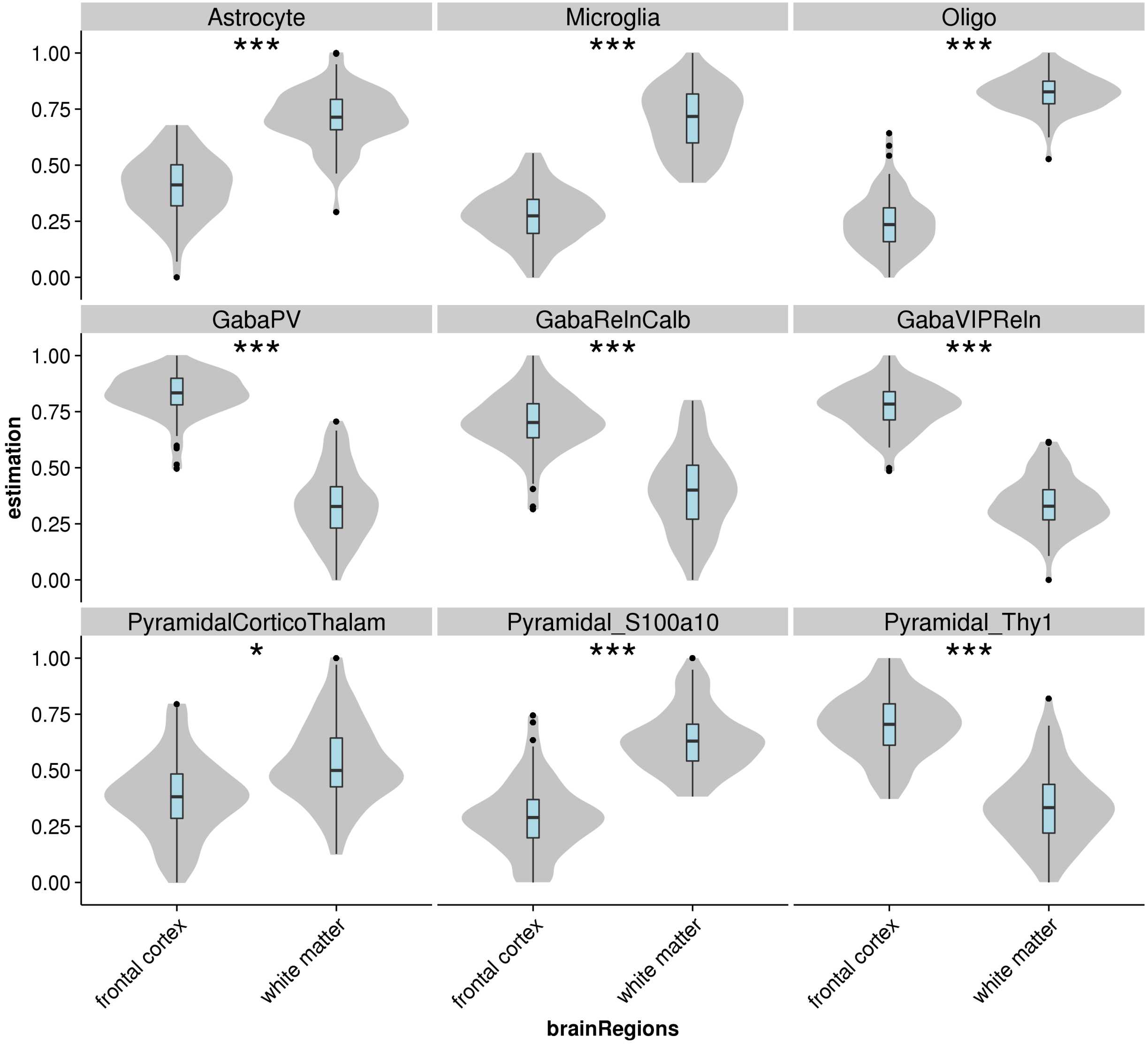


Figure 13: Estimations of cortical cell types in frontal cortex and white matter. Values are normalized to be between 0 and 1

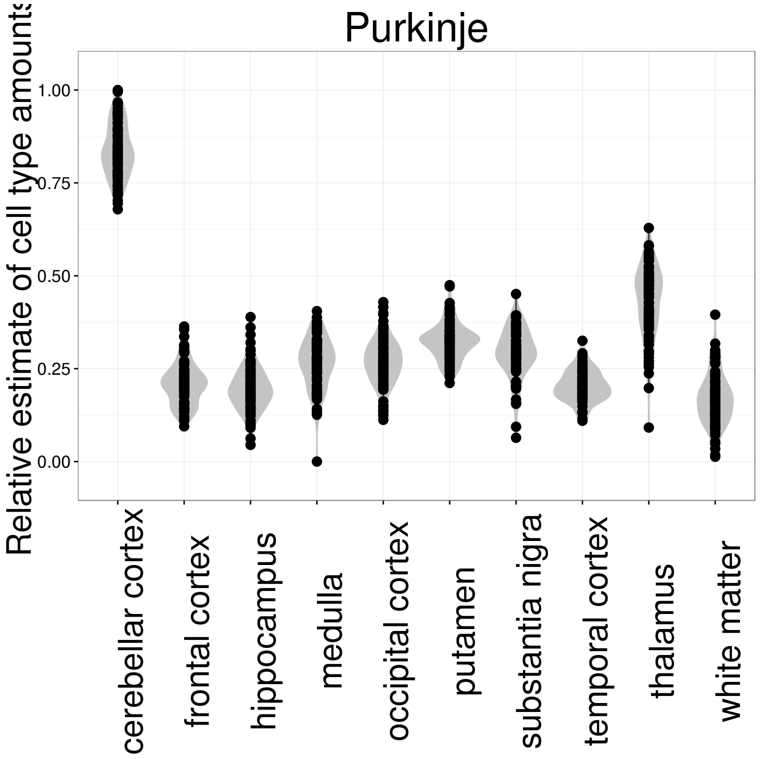


Figure 14: Estimations of purkinje cells in different brain regions. Values are normalized to be between 0 and 1

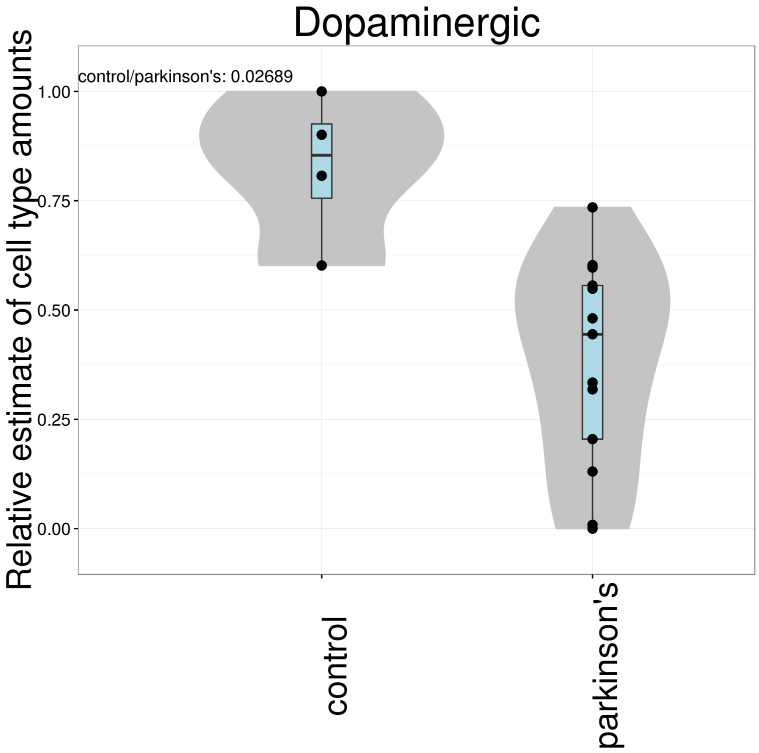


Figure 15: Estimations of dopaminergic cells in different substantia nigra of male parkinson's disease patients. Values are normalized to be between 0 and 1

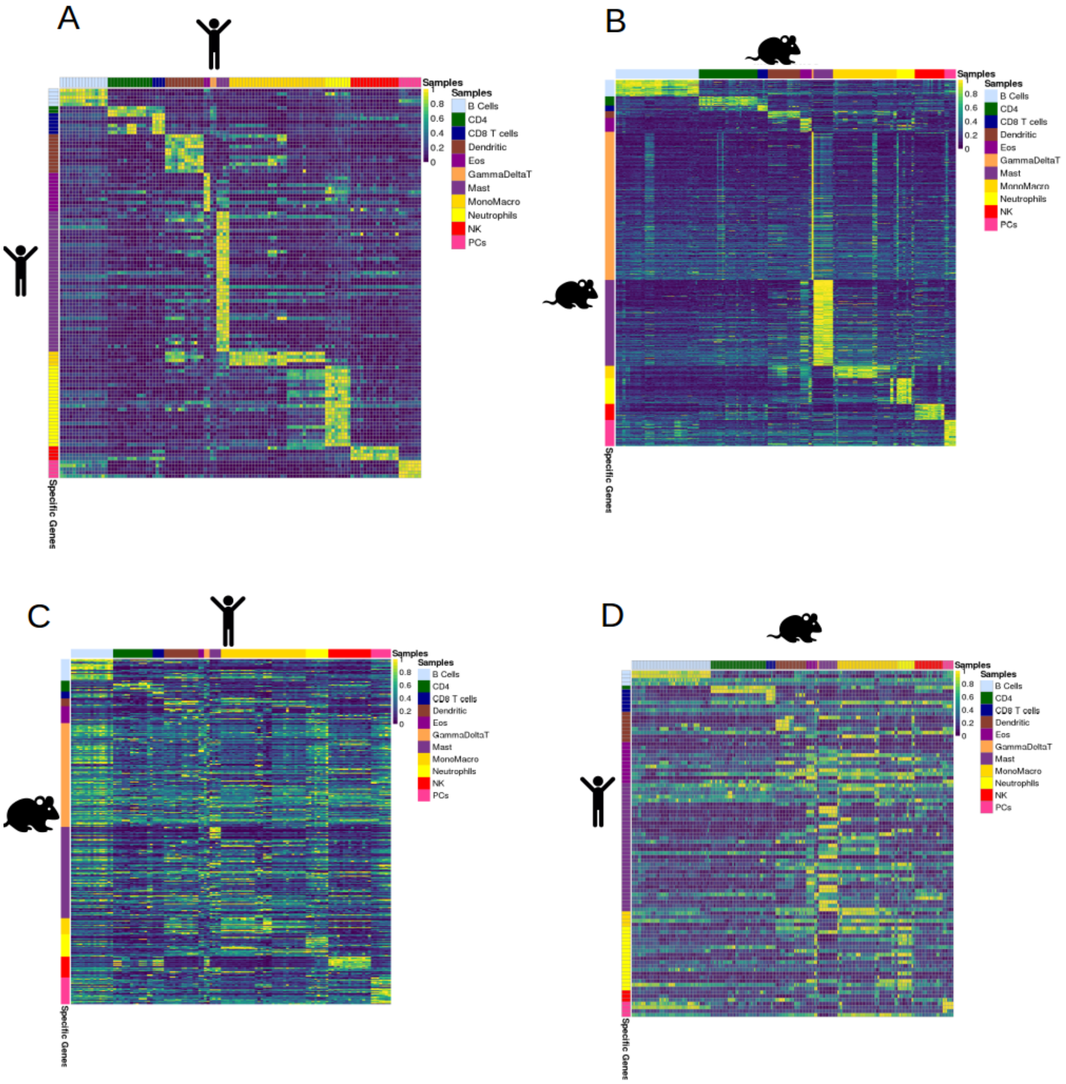


Figure 16: A-B. Expression of the genes selected from a species in the samples used for isolation from the same species. A shows human genes in human cell type specific expression profile dataset while B is mouse genes in mouse cell type specific expression profile dataset. C-D. Expression of homologues 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.

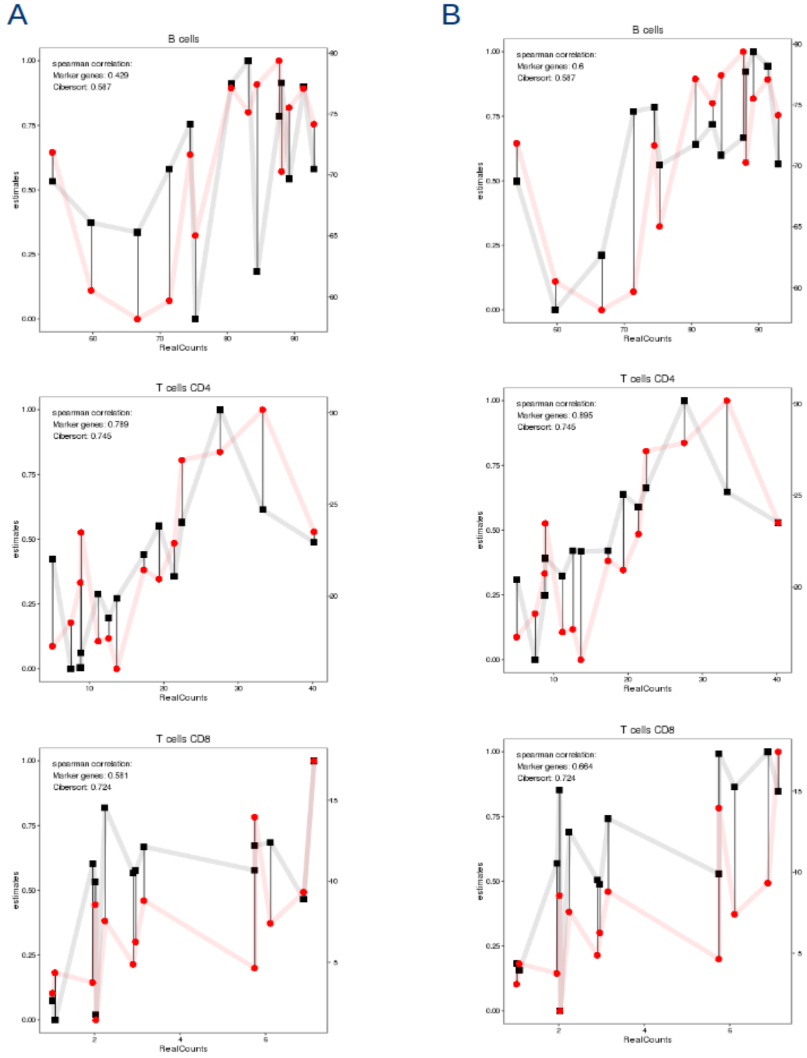


Figure 17: Correlations of estimations done by marker genes (black) and Cibersort (red) to real cell counts of the samples. Left axis shows our estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. A. Estimations done using marker genes selected from human cell type expression profiles. B. Estimations done using marker genes selected from mouse cell type expression profiles.