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

Dr. Steven J.M. Jones

Examination Date

June 19, 2015

Contents

List of Figures

1 Motivation and Introduction

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). Characterization of these cell types is an ongoing challenge in neuroscience. A common problem with characterization attempts stems from the fact that they often focus on a subset of cell types that a research group is interested in. Such studies often include electrophysiological measurements coupled with expression data and attempt to find defining characteristics of a given cell type [@sugino_molecular_2006]. Major issue with such studies is their limited scope. Such that a characteristic thought to be specific to a cell type in that study might be commonplace in other cell types that were not a part of the study.

Another important problem is that even though the heterogeneity of brain is well known, most large scale expression studies that focus on neurological/neurodegenerative disorders use whole tissue samples to examine the effects of the diseases [@iwamoto_molecular_2004;@hokama_altered_2014;@maycox_analysis_2009]. 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 [@chikina_cellcode_2015].

We aim to solve the former problem by creating a highly inclusive database of cell type specific expression profiles. This database will be used to detect marker genes that that best represents a given cell type in a particular brain region that is likely to be biologically relevant. The genes that were discovered this way can shed light into characteristics of given cell types and be useful to neuroscientists who want to work on the cell type by being useful biomarkers. Through the markers we found, we partially want to solve the latter problem by using expression of marker genes as a surrogate proportion of given cell type. This will enable us to make previously unknown inferences from the whole tissue data to have a better understanding of the disease. Also we are hoping to use recently established methods [@chikina_cellcode_2015] to use the estimated proportions to get more out of differential expression data.

A large amount of cell type specific data gathered by independent groups is available in GEO. Many of these samples are collected independently, using different methods from different mouse strains. Due inherent problems of using a dataset compiled from many different sources, significant effort has to be allocated to make sure the marker genes that were found are reliable. We will do this by analyzing independent datasets to make sure marker genes behave as expected. Namely, in a whole tissue dataset, we will show that they are more co-expressed than the background and in single cell datasets they tend to occur in the same cells more

frequently than a randomly selected gene set.

We also hope to make the entire dataset more accessible by creating a web app to visualize the data. The application will allow users to create plots showing the expression of selected genes in cell types and allow them to do quick differential expression analyses.

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 [@okaty_transcriptional_2009;@sugino_molecular_2006]. Relatively few studies [@zeisel_cell_2015;@tasic_adult_2016] attempt to characterize cell types in the context of other known cell types of the brain. 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 [@chikina_cellcode_2015]. 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 [@zeisel_cell_2015;@tasic_adult_2016]. 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

Proteins are the main functioning units in a cell. They are the ultimate products of the central dogma the way it is traditionally described. DNA transcribing RNA and RNA is translated into proteins. Quantification of specific proteins is, while possible, much more difficult than quantification of specific RNA molecules and often not scalable to the same degree. RNA quantification on the other hand have been historically used much more frequently on a high throughput fashion particularly with the rise of microarrays and more recently RNA sequencing. A common concern about RNA quantification is that they do not always correlate with protein levels [@maier_correlation_2009]. Therefore care should be taken when considering their biological significance.

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 [@hoheisel_microarray_2006]. Microarrays by fabricated 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 cDNA library acquired by amplification of a target transcriptome. The amount of cDNA that hybridized to a probe is later quantified by staining the label attached to the cDNA library [@hegde_concise_2000]. Often, multiple probes target a close-by region on the genome to form a probeset that target a single gene. The standard output for most normalization techniques is a summarization of the probes that make up a probeset [@gautier_affyanalysis_2004].

There are a multitude of microarray chips for researchers to chose from. These chips primarily differ in the probesets 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 [@iwamoto_molecular_2004;@kang_spatio-temporal_2011;@torrey_stanley_2000] and isolated cell type [@okaty_transcriptional_2009;@sugino_molecular_2006;@dougherty_disruption_2013] level. Since the level of heterogeneity in the brain is not fully understood, samples aiming to feature single cell types often have a risk of contamination by unintended cell types.

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 amplified cDNA that is acquired from a target transcriptome. Unlike microarrays they do not target specific locations on the genome hence can give a much comprehensive picture of the transcriptome including gene discovery and splicing variant identification. The quantification 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 [@wang_rna-seq_2009]. In general RNA seq is more prone to technical artifacts due to the stochasticity of the sequencing process **citation**. This effect is particularly powerful for lowly expressed genes which often make up the majority of the data [@labaj_characterization_2011] Figure 2. Increasing the sequencing depth, read lengths and using unique labels for molecules before any amplification are ways developed to alleviate such artifacts.

With development of microfluidics it became possible for RNA sequencing to be performed on single cells **citations**. Due to scarcity of the starting product, technical artifacts resulting from amplification and sequencing is more prominent **citation**.

RNA-seq analysis, especially single cell studies are starting to gain popularity in neuroscience [@zeisel_cell_2015;]. 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.

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 [@gautier_affyanalysis_2004;@islam_quantitative_2014;@darmanis_survey_2015]. Another common method of analysis is coexpression, where researchers attempt to identify genes that show similar changes in expression across different samples **citation**. This information is often used to derive functional relationships between the genes. **citation**

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 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). 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. Finally, 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 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 **a figure for all methods. add citations** [@heiman_cell_2014]. Alternatively cells can be isolated either by relying on markers as in the previous methods or known physical characteristics by manually picking them, or by Laser Capture Microdissection (LCM).

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 [@okaty_quantitative_2011].

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. 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 **citation**. 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.

3.4 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

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

(from Shenn-Orr et. al. [-@shen-orr_cell_2010-1])

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. Due to ease of access to both mixed samples and isolated cell types, blood has been a principle focus of deconvolution experiments [@westra_cell_2015;@newman_robust_2015;@chikina_cellcode_2015]. While deconvolution attempts in brain has been performed [@xu_cell_2013;@grange_cell-typebased_2014], deconvolution of human brains remained at a relatively superficial level that focuses on more generalized cell type groups.

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. This requires researchers to have data from individual cell types that make up the mixture. Most naively, researchers try to estimate the W in solving the equation 1 by minimizing the e [@grange_cell-typebased_2014]. 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 platform. To combat such problems feature selection can be used to use the most informative parts the reference expression matrix which in turn makes the estimation process more robust [@newman_robust_2015].

3.4.2 Reference free deconvolution

In cases where cell type expression profiles is not available or is 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. 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 [@xu_cell_2013;@chikina_cellcode_2015]. 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 the samples.

3.5 Cell types of the central nervous system

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 througout the brain. 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. More recently, it was also discovered that astrocytes play roles in synapse function by neurotransmitter release in response to changes in neuronal activity. Astrocytes often proliferate in diseased brain. This increase is often coupled with upregulation of certain genes, GFAP which is often used as the go to astrocyte marker by neurologists [@sofroniew_astrocytes_2010].

- **Oligodendrocytes** are responsible for forming the mylein sheet around neurons to insulate their axons. A single oligodendrocyte ensheats multiple neuronal cells as a result of a highly coordinated process effected by axon size, neuronal activity and molecular signalling. Most myelination occurs during early differentiation process. Oligodendrocytes are often susceptible to cell death de to oxidative damage due their high metabolic rates [@bradl_oligodendrocytes_2010].

- **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 associtated with nervous system disorders [@graeber_microglia_2010].

3.5.2 Neurons

Pyramidal Cells are prominent in areas of the brain associated with high cognitive functions such as cerebral cortex and hippocampus. They are recognized by their triangular cell body and short basal dendrites. Pyramidal cells of different brain regions and layers has differences in structure and gene expression. They can receive inhibitory gabaergic inputs through soma and axon. Excitory signals are delivered through the dendrites. Dendrites of pyramidal cells are covered with high number of spines that act as synapse sites for glutamatergic synapses [@spruston_pyramidal_2008].

Cortical gabaergic neurons are local inhibitory neurons that control pyramidal cell firing and generation of cortical rhytms. In cortex there seems to be 3 major groups. PV expressing, SST expressing and 5HT3a expressing. These subtypes of gabaergic cells localize into specific layers of the neocortex. PV expressing

neurons are fast spiking cells with low input resistence. 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. Finally 5HT3aR interneurons are a more heterogeneous subgroup of interneurons which differ in function and morphology. They are modulated with serotonin receptors [@rudy_three_2011].

Dopaminergic neurons

Purkinje cells are gabaergic cells local to cerebellum [@ito_historical_2002].

Cholinergic cells

Serotonergic cells

Complete later

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. [-@okaty_quantitative_2011] 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 querried 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 costum 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 [@irizarry_exploration_2003;@irizarry_summaries_2003]. 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 assymetry in probeset level signal distribution **pre-quantile normalization figure**. 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 [@bolstad_comparison_2003] to make the samples comparable to each other **after quantile normalization figure**. 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 **x** studies **cell type table**. 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: basic screenshot**. 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 publically 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 differencial 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 relevent, they should be unique to a single cell type in the context of said region. To acomplish 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 mutliple 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 accross 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 seperate samples from each other rather that simply being highly expressed (second condition).

As a result, we selected marker genes accross 10 regions **table** from **x** cell types. Number of marker genes grealty

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 colocalized to the cells.

When possible, we will be using the ISH data available from the Allen Brain Atlas [@lein_genome-wide_2007] to confirm our findings. While a powerful resource including thousands of ISH images, since every slice is labeled 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 **figure**.

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 coexpressed than average based on a null distribution of genes with similar prevelance in the dataset. For most gene sets this gave favorable 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 exlusivity to a single cell type, but its hightened expression in the cell type.

We used this method in single cell RNA-seq datasets from mouse [@zeisel_cell_2015] and human [@darmanis_survey_2015] and in both cases, we were able to see enriched coexpression in a majority of our gene sets **RNA-seq coexpression figures**. 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 [@tasic_adult_2016].

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 [@trabzuni_widespread_2013] 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** 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, assesing repeatability of such single cell studies is not straigtforward to do. We aim to capture similarities between individual cells from RNA-seq datasets and samples from the microarray database by using commong genes that are captured by all of them in an expression independent manner. For microarray data we will be using absance 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 anaylsis 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 designadet groups in their respective papers **(pipeline figure for this mess)**. 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.

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 accross 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 concensus accross 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 [@trabzuni_widespread_2013] it was possible to observe an increase in glial population and decrease in neuronal populations between white matter and gray matter **figure** and purkinje cells were found to be exclusive to cerebellum **figure**. To assess the usefulnes 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 [@lesnick_genomic_2007], 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**.

We will be analysing 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 analysing an Alzheimer's disease cohort where samples accross different brain regions are collected from patients and controls and seeking for potential confounds in healty 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

Enumaration 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 genes selected for one species behave as marker genes in other. As expected there were mixed results depending on the resolution of cell type definitions **figure**. Then we attempted to enumarete cell type proportions in whole blood cell types and compare our results with a recently published reference based enumeration method [@newman_robust_2015]. While we got comparable results for human marker gene sets, mouse gene sets performed poorly compared to the published method when the cell types are finely defined **figure**.

Future work will focus on characterizing the genes that do not perform well accross species and assesing 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 analysing 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 heterogenicity of the sample. Since effects are often specific to cell types, having uneffected 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.

7 References

8 Figures

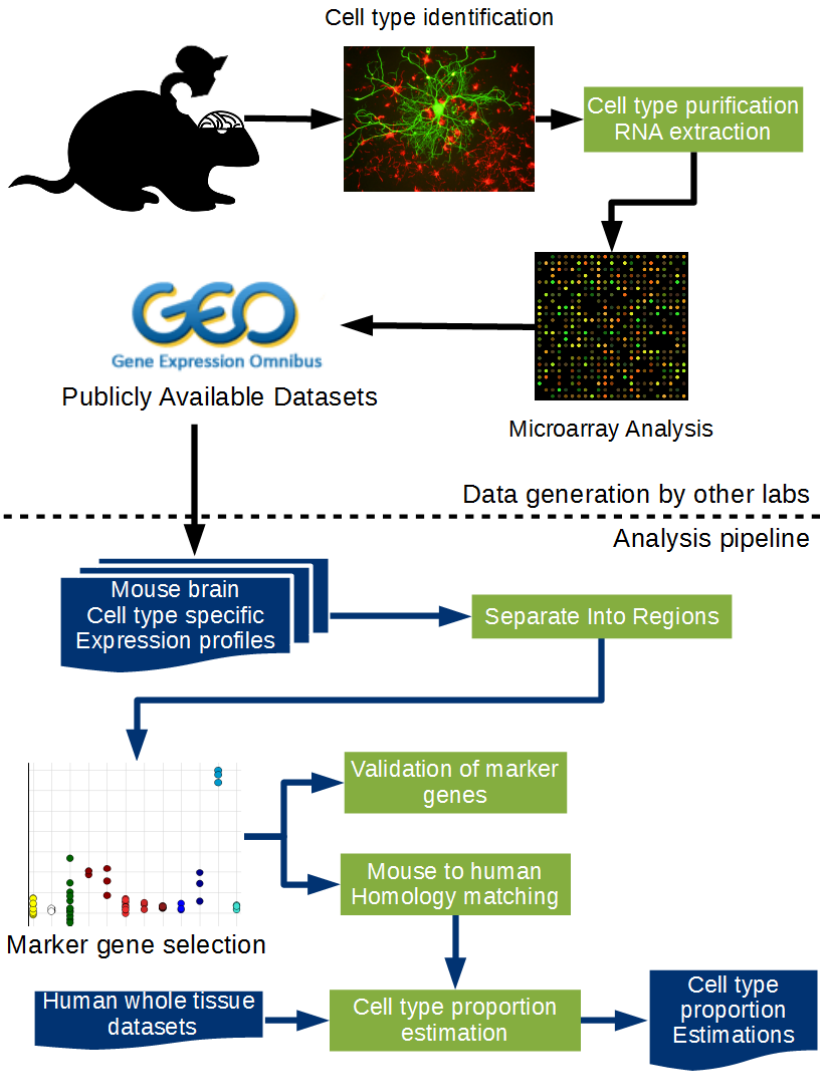

Figure 1: Workflow of the project

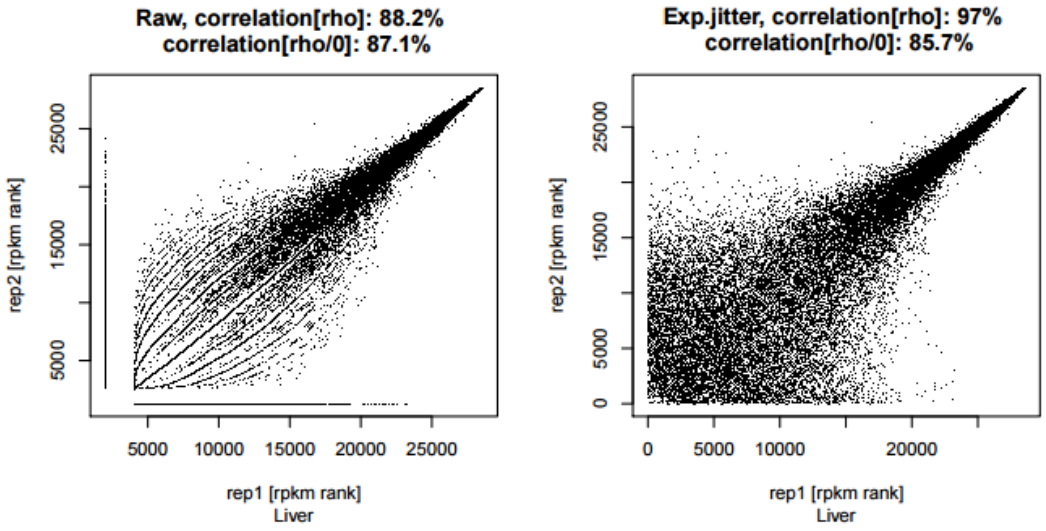

Figure 2: Taken from RNA sequencing shows low repeatability at low levels (Łabaj et al. 2011): Ranks of expression values for two indepent replicates are shown. Left figure left figure shows unmodified ranks. Right figure adds jitter to the points to reveal overplotting on the edges.