# 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 **(citations)**. 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 users will be able to see expression of genes in the cell types represented in the dataset.

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.

#### 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.

#### 2.2.4 Aim 4: Find cell type specific regulatory events with the help of enumaration information

1. Use enumeration information as covariates to make better use of differential expression data
2. Assign differentially expressed genes to cell types using correlation of differentially expressed genes to cell type proportions.

#### 2.2.5 Aim 5: Assess the concordance of single cell RNA seq data with each other and microarray data

1. Assess coexistence of marker genes detected from each source in each other to verify that they remain to be markers in independent datasets
2. Correlate expression data from RNA-seq studies to find if same subtypes can be detected in different studies
3. Attempt deconvolution of cell type specific microarray samples based on single cell data to assess the composition of manually isolated cell groups.

# 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 **citation?**. 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] **add figure (S8) from the paper**. 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. Traditionally glial cells are credited as the supporter cells of the brain **citation**. For the greater part this generalization still holds true. Oligodendrocytes are responsible for mylein formation and provide trophic support to neurons while microglia acts as the immune cells of the central nervous system. Astrocytes on the other hand both act as general purpose maintenance cells, taking up duties as maintainers of the blood brain barrier and extracelullar space, and take active roles in memory formation and regulation of synaptic transmission **citation**.

**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 aren't 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

We are currently in process of validating the marker genes we found by in situ hybridization on mouse brains. We will be using double labelling in situ hybridization to see if some of your marker genes co-exist with known markers of certain cell types. This will ensure that, at least a subset of our marker genes are useful as biomarkers of specific cell types.

### 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.

# 6 References

# 7 Figures

Figure 1: Workflow of the project