

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

| | |
|--|-----------|
| Contents | i |
| List of Figures | ii |
| 1 Motivation and Introduction | 1 |
| 2 Research questions and specific aims | 2 |
| 2.1 Research questions | 2 |
| 2.1.1 What are the specific marker genes of central nervous system (CNS) cell types? | 2 |
| 2.1.2 Are mouse marker genes applicable to humans? | 2 |
| 2.1.3 How accurately can cell type proportions be predicted with the use of marker genes? . | 2 |
| 2.1.4 How cell type proportions change accross neurological diseases? | 2 |
| 2.1.5 Can cell type specific regulatory event be detected using cell type proportion information? | 3 |
| 2.2 Specific aims | 3 |
| 2.2.1 Aim 1: Compilation of cell type specific expression database and make it available to third parties | 3 |
| 2.2.2 Aim 2: Identification and verification of marker genes | 3 |
| 2.2.3 Aim 3: Enumaration of cell type proportions | 3 |
| 3 Background | 3 |
| 3.1 Expression profiling | 3 |
| 3.1.1 Microarrays | 4 |
| 3.1.2 RNA sequencing | 4 |
| 3.1.3 Analysis of RNA quantification results | 5 |
| 3.2 Cell type isolation | 5 |
| 3.3 Cell type markers and their applications | 6 |
| 3.4 Cell type deconvolution | 6 |
| 3.4.1 Reference based deconvolution | 7 |
| 3.4.2 Reference free deconvolution | 7 |
| 3.5 Cell types of the central nervous system | 7 |
| 4 References | 8 |
| 5 Figures | 9 |

List of Figures

| | | |
|---|-----------------------------------|---|
| 1 | Workflow of the project | 9 |
|---|-----------------------------------|---|

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 unknwn 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 analysing 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 accesible 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 central nervous system (CNS) 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 the entire 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?

Many diseases of the CNS

2.1.5 Can cell type specific regulatory event be detected using cell type proportion information?

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 genes

1. Detecting cell type marker genes in a region dependent based on the localization of their expression
2. Verify marker genes in independent datasets

2.2.3 Aim 3: Enumeration of cell type proportions

1. Using marker genes' expression

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 its development have been a transformative force in many branches of biology [hoheisel_microarray_2006]. Microarrays are 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 labeled 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 closeby 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 choose 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 its 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 labeled to be fluorescently active, either through genetic manipulation or using in situ hybridization or labeled 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 wetlab, 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 modeled as below

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

(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, reseachers 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 glia, 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 myelin 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 extracellular space, and take active roles in memory formation and regulation of synaptic transmission **citation**.

Complete later

4 References

5 Figures



Figure 1: Workflow of the project