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[1]. 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[2].

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 unkown inferences from the whole tissue data to have a better understanding of the disease. Also we are hoping to use recently established methods[2] 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 Hypotheses and specific aims

2.1 Hypotheses

2.1.1 Hypothesis 1:

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 3: 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 4: Enumaration of cell type proportions

1. Estimate cell types
2. later

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[3]. 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[4]. 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 labeled cDNA library acquired by amplication of a target transcriptome. The amount of cDNA that hyrbidized to a probe is later quantified by staining the label attached to the cDNA library[5]. Signal for a gene is computed by summarizing the signal from mutliple probes that target closeby regions[6].

Talk about data

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[7]. In general RNA seq is more prone to technical artifacts due to the schotasticity of the sequencing process **citation**. This effect is particularly powerful for lowly expressed genes which otfen make up the majority of the data[8].

Increasing the sequencing depth, read lengths and using unique labels for molecules before any amplification are ways developed to aleviate 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**. This can readily be observed by the high prominence of low signals in most single cell RNA-seq experiments **figure**

Ta

8

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.

3.2 Cell type isolation

3.3 Cell type deconvolution

3.3.1 Reference based deconvolution

3.3.2 Reference free deconvolution

References

1 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.

2 Chikina M, Zaslavsky E, Sealfon SC. CellCODE: A robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics* 2015;: btv015.

3 Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Letters* 2009; **583**: 3966–3973.

4 Hoheisel JD. Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews Genetics* 2006; **7**: 200–210.

5 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.

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

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

8 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.

4 Figures

Figure 1: Workflow of the project