Permutation Test in Marker gene analysis

by

Xinyi Jin

Student number: 955712

A thesis submitted in total fulfillment for the degree of Master of Data Science

School of Mathematics and Statistics
THE UNIVERSITY OF MELBOURNE

September 2021

THE UNIVERSITY OF MELBOURNE

Abstract

School of Mathematics and Statistics

Master of Data Science

by Xinyi Jin

Student number: 955712

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Declaration of Authorship

I, Xinyi Jin, declare that this thesis titled, 'THESIS TITLE' and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work towards the NAME OF AWARD except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Signed: ____

Date: ____

Preface

Where applicable, the following information must be included in a preface:

- a description of work towards the thesis that was carried out in collaboration with others, indicating the nature and proportion of the contribution of others and in general terms the portions of the work which the student claims as original;
- a description of work towards the thesis that has been submitted for other qualifications;
- a description of work towards the thesis that was carried out prior to enrolment in the degree;
- whether any third party editorial assistance was provided in preparation of the thesis and whether the persons providing this assistance are knowledgeable in the academic discipline of the thesis;
- the contributions of all persons involved in any multi-authored publications or articles in preparation included in the thesis;
- the publication status of all chapters presented in article format using the descriptors below;
  - Unpublished material not submitted for publication
  - Submitted for publication to [publication name] on [date]
  - In revision following peer review by [publication name]
  - Accepted for publication by [publication name] on [date]
  - Published by [publication name] on [date]
- an acknowledgement of all sources of funding, including grant identification numbers where applicable and Australian Government Research Training Program Scholarships, including fee offset scholarships.

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor…

Contents

List of Figures

List of Tables

Abbreviations

LAH	**L**ist **A**bbreviations **H**ere

Constants

Speed of Light & c & = & $2.997\ 924\ 58 \times 10^8$ ms$^{-\mathrm{S}}$ (exact)

Symbols

t	t statistic	
t_{mod}	moderated t statistic	
t_{treat}	treat t statistic	
$\log \text{FC}$	log fold-change	β
N	number of permutations	
W	a defined statistic, $\log \text{FC} \times (1 - p)$	
p	raw p-value	
p_{adj}	adjusted p-value	

Chapter 1

Literature Review

Single cell sequencing has increasing popularity in recent years. It is a technology that provides gene counts for each cell present in the sequencing. As the special count data , The analysis of single cell RNA seq has its own workflows and challenges. A typical workflow starts from pre-processing steps of the raw count matrix. Poor quality cells and genes are filtered in the quality control steps. Single-cell data has technical noise and is therefore noisy.

Genes and cells that pass the quality control can be used for downstream analysis. One common cell-level analysis is to find cell types in the data. Cells are often clustered based on similarities and different methods are used to annotate the clusters. The computed clusters are useful for further cell type annotations.

There are two main approaches to annotate the computed clusters. approach is to identify marker genes that are representative for each cluster, and further analysis with marker gene sets will reveal more insights about the cell identity. Although the automatic cluster annotation methods exist, manual annotation remains its significance when the experiment is performed under different conditions and the reference cells are not applicable. As the number of marker genes tend to be large, the obtained marker genes are typically examined as a group. The group of marker genes is tested against many well-known gene sets to identify over-represented biological pathways or processes involved in. The biological pathways will aid the understanding of existing cell identities.

Marker analysis Manual cell type annotation involves finding marker genes that are highly regulated for the cluster of interest. Marker genes are identified by performing differential expression testing between the cells in the interested cluster versus all the rest cells. A considerable amount of literature has been published on marker-gene identification. There are various statistical tests used in DE, ranging from basic statics such as t-statistic to . The literature comparing thirty-six DE methods has highlighted that current tools have significant variations in terms of the number of detected marker genes and their significances. Tools developed originally for bulk RNA sequence, such as Limma, edgeR are argued to be competitive to those designed specific for single cell RNA sequencing. Gene set testing Different methods exist in the literature regarding gene set testing.

Permutation-based methods As argued in , commonly used DE tools tend to give underestimated p-vlaue estimates, resulting in an overestimation in the number of up-regulated genes. Single cell RNA seq usually contains large amount of cells in different cell identities, and permutation-based methods have the potential to approximate empirical distribution without specific distribution assumptions. P-value for each gene is expected to be more accurate with permutation testing, and therefore control the false positive rate in marker genes.

Similarly for gene set testing, what many available tools tend to show is the underestimated p values for the genes. There is a relatively small body of literature that is concerned with applying nonparametric methods to find marker genes [] []. However, the time and design complexity for those methods possibly limits the practical applications.

The key purpose of this project is to develop a permutation-based framework that performs marker identification to discover underlying cell types in the data. To make permutation-based framework work efficiently and effectively, different test statistics will be explored for marker gene analysis and gene set testing. It is aimed to provide more accurate p-value estimates with the consideration of cell-to-cell and sample-to-sample variabilities in a computationally efficient framework. The final permutation framework will be implemented in the speckle R package.

1.1 Standard workflow

1.2 Literature Review

Chapter 2

Assumptions

2.1 Assumption

The experiments are conducted with simulated count data from Splatter which makes assumptions about the expected behaviour of true DEs. Splatter fits the real count matrix with a lognormal distribution and allows . The parameter de.facloc ad de.facscale control the mean and standard deviation of the assigned multipliers to the ramdomly selected true DEs. Although two-sided tests are popular in practice, one sided test will be used to identify up-regulated genes. The purpose of marker genes analysis is to identify up-regulated genes that can be represntative for specific cell types so that more clues can be made about existing cell types. Two-sided statistical test will include both up-regulated and down-regulated genes, and those down-regualted genes will cause ambigurity for further interpretation.

2.2 Limma Framework

Limma fits a linear model for each gene. The adjusted p-values are believed to be over-optimistic.

2.3 Permutation Framework

For any statistic $\mathbf{T}_{obs}$ calculated from original count matrix, the same statistic $\mathbf{T}_{perm}$ will be constructed from each permutated count matrix. The one-sided p-value will be evaluated based on the probability of having a larger permutated statitic than the observed statistic.

Let N denote the number of permutations, the raw p-value p for each gene will be

$$p = \frac{\mathbf{I}_{\mathbf{T}_{perm} > \mathbf{T}_{obs}} + 1}{N + 1}$$

The key advantage of permutation test is the distribution-free property, which makes a wide range of statistics be and comparable. As the efficiency of limma framework, the permutation framework is implementated on top of the limma package, which makes the comparision simpler.

2.4 statistic

t-statistic (t) and moderated t-statistic $t_{\text{mod}}$are widely used for identifying marker genes in scRNA-seq. It is agued that t_{mod}should be prefera treat.t One additional statistic, $\mathbf{W}$, $\mathbf{W} = \log \text{FC} \times (1 - p)$ where p is the raw p-value of the t-statistic. An ideal marker gene is expected to have a larger logFC, and a smaller p-value, which will generate a larger W statistic.

Appendix A

Real data analysis script

Bibliography