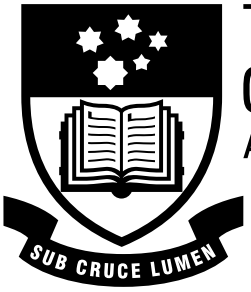

THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

School of Mathematical Sciences

Statistical equivalence in gene expression studies

PhD Thesis

Submitted November 2012

Simon Jonathan Tuke

Signed Statement

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Simon Jonathan Tuke and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: _____

Date: _____

Contents

Abstract

The goal of this thesis is to develop methods that enable researchers to identify genes with a pre-specified profile in gene expression studies. The motivation for these methods is a gene expression study performed at the University of Adelaide to identify the genes associated with pluripotency in mouse embryonic stem cells.

The method developed, *gene profiling*, utilises intersection-union tests that combine tests both for equivalent and differential expression. The theoretical basis of gene profiling and its application to the stem cell data are discussed in the thesis, as well as a comparison of gene profiling to alternative methods described in the literature for identifying genes corresponding to a pre-specified profile.

In the second part of this thesis, 'strength of evidence' measures for equivalence and gene profiling are developed. A P-value for equivalence is shown to have undesirable properties and a posterior probability of equivalence is proposed. The calculation of the posterior probabilities for equivalence using a hierarchical model is described and extended to gene profiling.

A method to calculate q-values from the posterior probabilities is described. The resultant q-values are compared to the posterior probabilities.

Acknowledgements

First I would like to thank Professor Patty Solomon and Associate Professor Gary Glonek for their patience and support as my PhD supervisors. Without you these past few years would not have been the same. You have shown me how to research and then communicate our ideas in clear and concise prose. I can only hope that in our future work I can repay the debt.

To the friends I have made at the School of Mathematical Sciences, thank you for making the journey so enjoyable. To name a few: David Butler, David Roberts, Ray Vozzo, Sam Cohen, Matt Finn, Matt Roughan, Trent Mattner, and many more. I hope to continue drinking coffee and chewing the fat with you.

To Eric Parsonage, thanks. It is nice to have a friend who in times of need is willing to tell you that you are an idiot and should 'suck it up, fat boy'. There would probably not have been a PhD without coffees with you. I know that I will always be barely an acquaintance, but you will always be my friend.

Finally, thanks to Alison: I dedicate this PhD to you. Sorry it took so long, but I got distracted. Is that not a great motto for life?

Note on Original Work

This thesis presents original results and proofs, but it also reviews existing results by other authors. Results and proofs due to other authors will be clearly marked with a citation such as (Storey, 2003). Any result not marked in this way is an original result the proof or exposition of which is due to the author of this thesis.

The key results on gene profiling in Chapter 2 have been published in the journal *Biostatistics*; see Tuke, Glonek, and Solomon (2009).