

Shuo Jiao

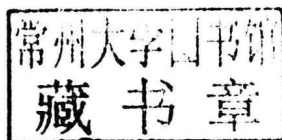
# MICROARRAY DATA ANALYSIS

DETECTING DIFFERENTIALLY EXPRESSED  
GENES WHILE CONTROLLING THE FALSE  
DISCOVERY RATE

Shuo Jiao

# MICROARRAY DATA ANALYSIS

DETECTING DIFFERENTIALLY EXPRESSED  
GENES WHILE CONTROLLING THE FALSE  
DISCOVERY RATE



VDM Verlag Dr. Müller

## **Impressum/Imprint (nur für Deutschland/ only for Germany)**

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Coverbild: [www.purestockx.com](http://www.purestockx.com)

Verlag: VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG

Dudweiler Landstr. 99, 66123 Saarbrücken, Deutschland

Telefon +49 681 9100-698, Telefax +49 681 9100-988, Email: [info@vdm-verlag.de](mailto:info@vdm-verlag.de)

Zugl.: Lincoln, University of Nebraska at Lincoln, 2009

Herstellung in Deutschland:

Schaltungsdienst Lange o.H.G., Berlin

Books on Demand GmbH, Norderstedt

Reha GmbH, Saarbrücken

Amazon Distribution GmbH, Leipzig

ISBN: 978-3-639-23809-9

## **Imprint (only for USA, GB)**

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this works is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Cover image: [www.purestockx.com](http://www.purestockx.com)

Publisher:

VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG

Dudweiler Landstr. 99, 66123 Saarbrücken, Germany

Phone +49 681 9100-698, Fax +49 681 9100-988, Email: [info@vdm-publishing.com](mailto:info@vdm-publishing.com)

Copyright © 2010 by the author and VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG and licensors

All rights reserved. Saarbrücken 2010

Printed in the U.S.A.

Printed in the U.K. by (see last page)

ISBN: 978-3-639-23809-9

**Shuo Jiao**

**MICROARRAY DATA ANALYSIS**

**Shuo Jiao**

# **MICROARRAY DATA ANALYSIS**

**DETECTING DIFFERENTIALLY EXPRESSED  
GENES WHILE CONTROLLING THE FALSE  
DISCOVERY RATE**

**VDM Verlag Dr. Müller**

## Table of Contents

Acknowledgements	3
List of Tables	7
List of Figures	9
Chapter 1. INTRODUCTION	11
1.1. Background	11
1.2. Problem Statement	13
1.3. Research Objectives	15
Chapter 2. LITERATURE REVIEW	16
2.1. Detecting DE genes	16
2.2. Estimating FDR	23
Chapter 3. THE $T$ -MIXTURE MODEL APPROACH FOR DETECTING DIFFERENTIALLY EXPRESSED GENES IN MICROARRAYS	28
3.1. Introduction	28
3.2. Methods	30
3.3. Results	35
3.4. Discussion	40

Chapter 4. A MIXTURE MODEL BASED APPROACH FOR ESTIMATING THE FDR IN REPLICATED MICROARRAY DATA	44
4.1. Introduction	44
4.2. Methods	46
4.3. Results	52
4.4. Discussion	53
Chapter 5. ON CORRECTING THE OVERESTIMATION OF THE PERMUTATION- BASED FDR ESTIMATOR	58
5.1. Introduction	59
5.2. Methods	61
5.3. Results	67
5.4. Discussion	72
Chapter 6. ESTIMATING THE PROPORTION OF EQUIVALENTLY EXPRESSED GENES IN MICROARRAY DATA BASED ON TRANSFORMED TEST STATISTICS	79
6.1. Introduction	80
6.2. Methods	82
6.3. Results	90
6.4. Discussion	95
6.5. Appendix	95
Chapter 7. CONCLUSION	100
7.1. Summary	100

7.2. Future work	101
References	103
APPENDIX	108
a. Codes for fitting a $t$ -mixture model	108
b. Codes for comparison between the model based FDR and the empirical FDR	113
c. Codes for comparison between the two-step FDR estimator and the standard method	114
d. Codes for estimating $\pi_0$ using our method	120



## List of Tables

1.1	Outcome of a microarray data analysis with $n$ genes.	14
3.1	Comparison of the MMM and TMM in Type I error rates at given gene specific levels of significance.	37
3.2	Comparison of the MMM and TMM in Type I error rates with respect to different number of permuted sets of null scores when all the genes are EE.	39
3.3	Comparison of the MMM and TMM in Type I error rates with respect to different number of permuted sets of null scores and with the existence of DE genes.	39
3.4	Comparison of the results from the TMM and TMM at given levels of significance for the Leukaemia data.	41
3.5	List of DE genes identified by both the TMM and MMM when genome-wide significance level is 0.0005.	41
5.1	Comparison of estimated false positive numbers and the true false positive numbers using the SAM, mean and $t$ -statistics. $\widehat{FP}_p$ is the estimated FP number with 150 predicted DE genes removed; $\widehat{FP}_t$ is the estimated FP number with 150 true DE genes removed.	69

5.2	Comparison of the performance of FDR estimator when the ratio of induced and repressed genes changes.	71
5.3	Comparison of the performance of $\widehat{FDR}(d)_2$ and $\widehat{FDR}(d)_0$ using microarray data from Zhong <i>et al.</i> (2004).	72
6.1	Comparison of $\pi_0$ Estimates from our method, BUM, SPLOSH, QVALUE and LBE for the Golub <i>et al.</i> (1999) and Hedenfalk <i>et al.</i> (2001) data.	91
6.2	Comparison of the mean and bias of the $\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for set-up (a), EE, DE genes well separated; (b), EE, DE genes not well separated; and (c), Mimic the real data. The values outside and inside parenthesis are mean and bias, respectively.	93
6.3	Comparison of the standard error of the $\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for set-up (a), EE, DE genes well separated; (b), EE, DE genes not well separated; and (c), Mimic the real data.	93
6.4	Comparison of the mean squared error of the $\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for set-up (a), EE, DE genes well separated; (b), EE, DE genes not well separated; and (c), Mimic the real data.	94

## List of Figures

3.1	Plot of the comparison between TMM and MMM.	43
4.1	Comparison of the true FDR, the empirical FDR estimator $\widehat{FDR}$ and the model based FDR estimator $\widehat{FDR}_1$ for two sample microarray data. 5 replicates are listed. Total number of significant genes is decreasing from 100 to 1 (left to right) for each replicate.	55
4.2	Comparison of the true FDR, the empirical FDR estimator $\widehat{FDR}$ and the model based FDR estimator $\widehat{FDR}_1$ for two sample microarray data. 5 replicates are listed. Total number of significant genes is decreasing from 150 to 1 (left to right) for each replicate.	56
4.3	Comparison of the empirical FDR estimator $\widehat{FDR}$ and the model based FDR estimator $\widehat{FDR}_1$ for Leukemia microarray data.	57
5.1	The FDR curves of different estimation methods using the SAM, mean, and $t$ -statistics. There are 400 DE genes among 4000 genes. The number of claimed significant gene ranges from 100 to 200. $\hat{\pi}_0^{sam}$ is used as the estimate of $\pi_0$ . Our method 1 is the estimator $\widehat{FDR}(d)_1$ from (5.9).	74
5.2	The FDR curves of different estimation methods using the SAM, mean, and $t$ -statistics. There are 400 DE genes among 4000 genes. The number	

of claimed significant gene ranges from 500 to 600. Our method 1 is the estimator  $\widehat{FDR}(d)_1$  from (5.9). 75

5.3 The FDR curves of different estimation methods using the SAM, mean, and  $t$ -statistics. There are 150 DE genes among 4000 genes. The number of claimed significant gene ranges from 20 to 400.  $\hat{\pi}_0^{sam}$  is used as estimate of  $\pi_0$ . Our methods 1 and 2 are the estimators  $\widehat{FDR}(d)_1$  from (5.9) and  $\widehat{FDR}(d)_2$  from (5.10), respectively. 76

5.4 The FDR curves of different estimation methods using the SAM, mean, and  $t$ -statistics. There are 150 DE genes among 4000 genes. The number of claimed significant gene ranges from 20 to 400. The true  $\pi_0 = 3850/4000$  is used as estimate of  $\pi_0$ . Our methods 1 and 2 are the estimators  $\widehat{FDR}(d)_1$  from (5.9) and  $\widehat{FDR}(d)_2$  from (5.10), respectively. 77

5.5 The FDR curves of different estimation methods using the SAM, mean, and  $t$ -statistics. Mimicking the real data. There are 150 DE genes among 4000 genes. The number of claimed significant gene ranges from 20 to 400. Our methods 1 and 2 are the estimators  $\widehat{FDR}(d)_1$  from (5.9) and  $\widehat{FDR}(d)_2$  from (5.10), respectively. 78

**DETECTING DIFFERENTIALLY EXPRESSED GENES WHILE  
CONTROLLING THE FALSE DISCOVERY RATE FOR  
MICROARRAY DATA**

by

Shuo Jiao

A DISSERTATION

Presented to the Faculty of  
The Graduate College at the University of Nebraska  
In Partial Fulfillment of Requirements  
For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professors Shunpu Zhang and Stephen D. Kachman

Lincoln, Nebraska

December, 2009

# **DETECTING DIFFERENTIALLY EXPRESSED GENES WHILE CONTROLLING THE FALSE DISCOVERY RATE FOR MICROARRAY DATA**

Shuo Jiao, Ph.D.

University of Nebraska, 2009

Advisers: Shunpu Zhang and Stephen D. Kachman

Microarray is an important technology which enables people to investigate the expression levels of thousands of genes at the same time. One common goal of microarray data analysis is to detect differentially expressed genes while controlling the false discovery rate. This dissertation consists with four papers written to address this goal. The dissertation is organized as follows: In Chapter 1, a brief introduction of the Affymetrix GeneChip microarray technology is provided. The concept of differentially expressed genes and the definition of the false discovery rate are also introduced. In Chapter 2, a literature review of the related works on this matter is provided. In Chapter 3, a  $t$ -mixture model based method is proposed to detect differentially expressed genes. In Chapter 4, a  $t$ -mixture model based false discovery rate estimator is proposed to overcome several problems of the current empirical false discovery rate estimators. In Chapter 5, a two-step false discovery rate estimation procedure is proposed to correct the overestimation of the false discovery rate caused by differentially expressed genes. In Chapter 6, a novel estimator is developed to estimate the proportion of equivalently expressed genes, which is an important component of the false discovery rate estimators. In Chapter 7, a summary of the dissertation will be given along with some possible directions for the future work.

## **Acknowledgements**

The completion of this dissertation is impossible without the support from many people. I would like to give my deepest gratitude to my advisor Dr. Shunpu Zhang. He directed me into the area of my dissertation, gave me insightful advices, and encouragingly supported my ideas. I would also like to thank my co-advisor Dr. Stephen D. Kachman for always being there to listen and discuss. I learned a lot from his way of thinking. I would like to thank Dr. Kent M. Eskridge and Dr. Istvan Ladunga for serving on my PhD. supervisory committee. Their careful proofreading of the dissertation proposal helps me improve my writing skills and I am grateful to them for holding me to a high research standard.

I am also indebted to Dr. Walter W. Stroup, Dr. Jim Lewis, and Dr. Ruth Heaton for providing the financial support to me, which was crucial for my PhD program. I want to give a special thanks to Dr. Yuannan Xia for letting me participate in his microarray experiment.

I dedicate this work to my parents, my fiancée, and our family who have been supportive all the time.

## Table of Contents

Acknowledgements	3
List of Tables	7
List of Figures	9
Chapter 1. INTRODUCTION	11
1.1. Background	11
1.2. Problem Statement	13
1.3. Research Objectives	15
Chapter 2. LITERATURE REVIEW	16
2.1. Detecting DE genes	16
2.2. Estimating FDR	23
Chapter 3. THE $T$ -MIXTURE MODEL APPROACH FOR DETECTING DIFFERENTIALLY EXPRESSED GENES IN MICROARRAYS	28
3.1. Introduction	28
3.2. Methods	30
3.3. Results	35
3.4. Discussion	40



Chapter 4. A MIXTURE MODEL BASED APPROACH FOR ESTIMATING THE FDR IN REPLICATED MICROARRAY DATA	44
4.1. Introduction	44
4.2. Methods	46
4.3. Results	52
4.4. Discussion	53
Chapter 5. ON CORRECTING THE OVERESTIMATION OF THE PERMUTATION- BASED FDR ESTIMATOR	58
5.1. Introduction	59
5.2. Methods	61
5.3. Results	67
5.4. Discussion	72
Chapter 6. ESTIMATING THE PROPORTION OF EQUIVALENTLY EXPRESSED GENES IN MICROARRAY DATA BASED ON TRANSFORMED TEST STATISTICS	79
6.1. Introduction	80
6.2. Methods	82
6.3. Results	90
6.4. Discussion	95
6.5. Appendix	95
Chapter 7. CONCLUSION	100
7.1. Summary	100