UTS UNIVERSITY OF TECHNOLOGY SYDNEY

A hierarchical model to detect differential gene expression distributions, and their investigation as a reflection of dysregulation in cancer

by Aedan Roberts

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Professor Paul Kennedy and Associate Professor Daniel Catchpoole

University of Technology Sydney
Faculty of Engineering and Information Technology

March 2021

Certificate of Original Authorship

I, Aedan Roberts, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

[Aedan Roberts]

DATE: 12th August, 2021

PLACE: Sydney, Australia

Acknowledgements

I would like to thank my supervisors, Prof Paul Kennedy and A/Prof Daniel Catchpoole, for their support and advice throughout the course of my PhD. Their encouragement, guidance, and the many discussions on the direction of my work have been invaluable.

My endless gratitude goes to my wife Beatrice, for supporting, encouraging and tolerating me, not necessarily in that order. A nearly equal amount of gratitude goes to Waffles and Clio, who really didn't offer any useful advice to be honest, but whose presence was immensely helpful nonetheless.

I would also like to thank the support staff at the School of Computer Science for their help throughout the process, in particular Margot Kopel and Janet Stack. Thanks also go to Thomas Lysaght and George Mundackal, who helped to optimise my R code as part of their Computing Science Studio 1 project.

I am grateful for the financial support provided by the NSW Health PhD Scholarship and the Australian Government Research Training Program.

Finally, I thank the anonymous donors whose tissue donations to GTEx, TCGA and the GEO repository make this and so much other research possible.

Contents

List of Figures xi

List of Tables xv

List of Publications xix

List of Abbreviations and Symbols xxi

Abstract xxv

1 Introduction 1

1.1 Background 1

1.2 Research questions 3

1.3 Contributions to knowledge 4

1.4 Thesis structure 7

2 Literature review 9

2.1 Gene expression data and its analysis 9

2.1.1 Microarray data 10

2.1.2 RNA-seq data 12

2.1.3 The negative binomial distribution 14

2.1.4 Shrinkage estimators and hierarchical models in RNA-seq data analysis 16

2.1.5 Alternatives to the negative binomial model 19

2.2 Biological significance and analysis of changes in gene expression distributions 20

2.2.1 Biological significance of expression variability 20

2.2.2 Measures of variability 22

2.2.3 Identification of differentially variable genes 23
2.2.4 Classification based on differential variability 25
2.2.5 Differential distribution 30
2.3 Research gaps 32

3 Differential variability and distribution are informative for cancer subtype classification 35
3.1 Introduction 35
3.2 Methods 36
3.2.1 Datasets and classification problems 36
3.2.2 Feature selection and engineering 37
3.2.3 Classification 39
3.3 Results 41
3.3.1 Colon adenoma versus cancer 41
3.3.2 Lung cancer prognosis 43
3.3.3 ALL subtype classification 46
3.4 Discussion 50
3.4.1 Differential variability feature selection for predicting cancer progression and prognosis 50
3.4.2 Cancer subtype classification using differential variability feature selection 52
3.4.3 Differential expression, variability and distribution 53
3.4.4 Feature transformation and considerations for classification methods 53
3.5 Conclusion 55

4 A hierarchical model to detect differential gene expression distributions in RNA-seq data 57
4.1 Introduction 57
4.2 Initial model development and implementation 58
4.2.1 Selection of priors 60
4.2.2 Posterior inference for means and dispersions 62
4.2.3 Markov chain Monte Carlo algorithm 65
4.3 Mixture model for differential distribution 69
4.4 MCMC algorithm optimisation 74

4.4.1 Joint parameter updates 75
4.4.2 Independence sampler 75
4.4.3 Non-symmetric proposal densities 77
4.4.4 Adaptive proposal densities 78
4.4.5 Final MCMC implementation 79
4.5 Posterior inference from the hierarchical model 80
4.6 Alternative model specification 82
4.7 Conclusion 86

5 Assessment of the hierarchical model 87
5.1 Introduction 87
5.2 Methods 88
5.2.1 Data simulation and processing 89
5.2.2 Introducing expression changes in real data 91
5.2.3 Methods to compare 93
5.2.4 Metrics for comparison 95
5.3 Results: simulated data 97
5.3.1 Dispersion estimation 97
5.3.2 Differential expression 99
5.3.3 Differential dispersion 103
5.3.4 Differential distribution 108
5.4 Results: Real data with artificially introduced differences in expression 113
5.4.1 Differential expression 113
5.4.2 Differential dispersion 118
5.4.3 Differential distribution 121
5.5 Discussion 124
5.5.1 Differential expression 125
5.5.2 Differential dispersion 125
5.5.3 Differential distribution 126
5.6 Conclusion 127

6 An investigation into dysregulation of gene expression in cancer 129
6.1 Introduction 129
6.2 Methods 131
6.2.1 Datasets and processing 131

6.2.2 Differential expression, dispersion and distribution analyses 134

6.2.3 Gene set enrichment analysis 135

6.3 Results 136

6.3.1 Differential dispersion identifies different sets of genes from differential expression 136

6.3.2 Differential dispersion and differential distribution identify cancer-related genes 138

6.3.3 Differential expression and differential dispersion identify functionally distinct sets of genes 143

6.4 Discussion 145

6.5 Conclusion 148

7 Conclusion 149

7.1 Cancer classification using differential variability and distribution 150

7.2 Detection of differentially variable and differentially distributed genes in RNA-seq data 152

7.3 Differential variability and distribution in cancer 154

7.4 Conclusion 156

Appendix A R code for the hierarchical model 157

Appendix B R code for methods comparisons 207

Appendix C Gene set enrichment analysis results 213

C.1 Breast adenocarcinoma 214

C.2 Clear cell renal cell carcinoma 216

C.3 Thyroid carcinoma 218

C.4 Hepatocellular carcinoma 220

C.5 Lung squamous cell carcinoma 222

C.6 Prostate adenocarcinoma 224

C.7 Colon adenocarcinoma 226

Bibliography 229

List of Figures

Figure **Page**

2.1 Illustration of classification problems using differential expression and differential variability. 27
2.2 Solving classification problems using differential expression and differential variability. 29

3.1 Combinations of feature selection and engineering methods and classification methods used. 40
3.2 Boxplots of AUCs from hyperdiploidy comparison on 100 random samples of training data. 49

4.1 Schematic illustration of the hierarchical model. 59
4.2 Assessment of prior forms for means. 61
4.3 Density histograms of 100,000 log-transformed draws from log-normal priors for the mean and dispersion resulting from 100,000 random draws from each hyperprior. 63
4.4 Traces of 50,000 posterior samples of the mean for 50 randomly selected genes, with three independent runs overlaid in different colours. . . . 67
4.5 Traces of 50,000 posterior samples of the dispersion for 50 randomly selected genes, with three independent runs overlaid in different colours. 68
4.6 Traces of 50,000 posterior samples of the log-normal prior scale parameters for the mean and dispersion, with five independent runs using different initial values overlaid in different colours. 69

5.1 Boxplots of MSEs for dispersion estimation from 50 simulated datasets with 4, 10, 20 and 40 samples. 97

5.2 Boxplots of MSEs for dispersion estimation from 50 simulated datasets with 2, 5, 10 and 20 samples per group, with 5% of genes differentially expressed. 99
5.3 Boxplots of AUCs for differential expression analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 100
5.4 Boxplots of partial AUCs for FPR < 0.05 for differential expression analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 100
5.5 Boxplots of FDRs at nominal 0.05 level for differential expression analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 101
5.6 Boxplots of TPRs at nominal FDR 0.05 for differential expression analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 102
5.7 False discovery curves for differential expression analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 103
5.8 Boxplots of AUCs for differential dispersion analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 104
5.9 Boxplots of partial AUCs for FPR < 0.05 for differential dispersion analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 104
5.10 Boxplots of FDRs at nominal 0.05 level for differential dispersion analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 105
5.11 Boxplots of TPRs at nominal FDR 0.05 for differential dispersion analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 106
5.12 False discovery curves for differential dispersion analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 107
5.13 False discovery curves for differential dispersion tests at minimum LFC of 1.44 on simulated data with 2, 5, 10, 20 and 50 samples per group. 108
5.14 Boxplots of AUCs for differential distribution analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 110
5.15 Boxplots of partial AUCs for FPR < 0.05 for differential distribution analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 110
5.16 Boxplots of FDRs at nominal 0.05 level for differential distribution analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 111
5.17 Boxplots of TPRs at nominal FDR 0.05 for differential distribution analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 112

5.18 False discovery curves for differential distribution analysis on simulated data with 2, 5, 10, 20 and 50 samples per group. 113

5.19 Boxplots of AUCs for differential expression analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 114

5.20 Boxplots of partial AUCs for FPR < 0.05 for differential expression analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 115

5.21 Boxplots of FDRs at nominal 0.05 level for differential expression analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 115

5.22 Boxplots of TPRs at nominal FDR 0.05 for differential expression analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 116

5.23 False discovery curves for differential expression analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 117

5.24 Boxplots of AUCs for differential dispersion analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 118

5.25 Boxplots of partial AUCs for FPR < 0.05 for differential dispersion analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 119

5.26 Boxplots of FDRs at nominal 0.05 level for differential dispersion analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 119

5.27 Boxplots of TPRs at nominal FDR 0.05 for differential dispersion analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 120

5.28 False discovery curves for differential dispersion analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 120

5.29 False discovery curves for differential dispersion tests at minimum LFC of 1.44 on GTEx data with 2, 5, 10, 20 and 50 samples per group. . . 121

5.30 Boxplots of AUCs for differential distribution analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 122

5.31 Boxplots of AUCs for FPR < 0.05 for differential distribution analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 122

5.32 Boxplots of FDRs at nominal 0.05 level for differential distribution analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 123

5.33 Boxplots of TPRs at nominal FDR 0.05 for differential distribution analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. . 124

5.34 False discovery curves for differential distribution analysis on GTEx data with 2, 5, 10, 20 and 50 samples per group. 124

6.1 Spearman correlation and p-values for hypothesis tests for negative correlation between gene lists for differential expression and differential dispersion for breast adenocarcinoma, with varying thresholds for calling differential expression or dispersion. 137

6.2 Spearman correlations and p-values for hypothesis tests for negative correlation between gene lists for differential expression and differential dispersion for TCGA data, with varying thresholds for calling differential expression or dispersion. 138

6.3 Number of lung adenocarcinoma- and prostate adenocarcinoma-related genes identified with varying thresholds for calling differential expression or dispersion. 142

6.4 Number of cancer-related genes identified with varying thresholds for calling differential expression or dispersion for eight cancer types. . . 143

List of Tables

Table **Page**

3.1 Datasets and classification problems. 37

3.2 Combinations of feature selection and engineering methods and classification methods used. 40

3.3 AUCs for classifying colon adenoma versus cancer using feature selection following Dinalankara and Corrada Bravo (2015). 42

3.4 AUCs for classifying colon adenoma versus cancer using features selected by differential variance between classes. 43

3.5 AUCs for classifying colon adenoma versus cancer using feature selection by differential expression or combining differential expression and differential variance. 43

3.6 AUCs for classifying low versus high risk lung cancer patients using universal anti-profile feature selection following Dinalankara and Corrada Bravo (2015). 44

3.7 AUCs for classifying low versus high risk lung cancer patients using features selected by differential variance between classes. 44

3.8 AUCs for classifying low versus high risk lung cancer patients using feature selection by differential expression or combining differential expression and differential variance. 45

3.9 AUCs for classifying low versus high risk lung cancer patients using features selected by differential variance between classes, for stage I and II carcinomas with complete tumour resection and no prior chemotherapy. 46

3.10 AUCs for classifying low versus high risk lung cancer patients using feature selection by differential expression or combining differential expression and differential variance, for stage I and II carcinomas with complete tumour resection and no prior chemotherapy. 46
3.11 AUCs for ALL classification problems using features selected by differential variance between classes. 47
3.12 AUCs for ALL classification problems using feature selection by differential expression or combining differential expression and differential variance. 48
3.13 Mean AUCs from hyperdiploidy comparison on 100 random samples of training data. 50

6.1 TCGA tumour–normal sample pairs used. 132
6.2 Tumour types from the CGC list used to identify related genes for each cancer type. 133
6.3 Number of related genes identified for each cancer type from each database and overall. 134
6.4 p-values from Wilcoxon rank-sum tests for ranking cancer-related genes above other genes, for differential expression (DE) using limma–voom, differential dispersion (Dϕ) using the hierarchical model, and differential distribution using the hierarchical mixture model (HMM) or hybrid method (Hyb) for paired tumour–normal TCGA RNA-seq datasets. 139
6.5 p-values from Wilcoxon rank-sum tests for ranking cancer pathway genes above other genes, for differential expression (DE) using limma–voom, differential dispersion (Dϕ) using the hierarchical model, and differential distribution using the hierarchical mixture model (HMM) or hybrid method (Hyb) for paired tumour–normal TCGA RNA-seq datasets. 140
6.6 Top 10 enriched GO terms in each ontology for lung adenocarcinoma based on differential expression. 144
6.7 Top 10 enriched GO terms in each ontology for lung adenocarcinoma based on differential dispersion. 145

C.1 Top 10 enriched GO terms in each ontology for breast adenocarcinoma based on differential expression. 214
C.2 Top 10 enriched GO terms in each ontology for breast adenocarcinoma based on differential dispersion. 215
C.3 Top 10 enriched GO terms in each ontology for clear cell renal cell carcinoma based on differential expression. 216
C.4 Top 10 enriched GO terms in each ontology for clear cell renal cell carcinoma based on differential dispersion. 217
C.5 Top 10 enriched GO terms in each ontology for thyroid carcinoma based on differential expression. 218
C.6 Top 10 enriched GO terms in each ontology for thyroid carcinoma based on differential dispersion. 219
C.7 Top 10 enriched GO terms in each ontology for hepatocellular carcinoma based on differential expression. 220
C.8 Top 10 enriched GO terms in each ontology for hepatocellular carcinoma based on differential dispersion. 221
C.9 Top 10 enriched GO terms in each ontology for lung squamous cell carcinoma based on differential expression. 222
C.10 Top 10 enriched GO terms in each ontology for lung squamous cell carcinoma based on differential dispersion. 223
C.11 Top 10 enriched GO terms in each ontology for prostate adenocarcinoma based on differential expression. 224
C.12 Top 10 enriched GO terms in each ontology for prostate adenocarcinoma based on differential dispersion. 225
C.13 Top 10 enriched GO terms in each ontology for colon adenocarcinoma based on differential expression. 226
C.14 Top 10 enriched GO terms in each ontology for colon adenocarcinoma based on differential dispersion. 227

List of Publications

Listed below are the publications and other outputs associated with the research presented in this thesis.

Roberts, A. G. K., Catchpoole, D. R. & Kennedy, P. J., 2021, 'Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in expression mean and variability'. (Submitted) Available as a preprint: https://www.biorxiv.org/content/10.1101/2021.02.15.431343v2.

Roberts, A. G. K., 2021, 'DiffDist', https://github.com/aedanr/DiffDist. R package.

Roberts, A. G. K., Catchpoole, D.R. & Kennedy, P.J., 2018, 'Variance-based Feature Selection for Classification of Cancer Subtypes Using Gene Expression Data', *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro.

List of Abbreviations and Symbols

Abbreviation	Description
ALL	Acute lymphoblastic leukaemia
AP	Anti-profiles
AUC	Area under the ROC curve
BFDR	Bayesian false discovery rate
BRCA	Breast invasive adenocarcinoma
CGC	Cancer Gene Census
COAD	Colon adenocarcinoma
CV	Coefficient of variation
DE	Differential expression
DMD	Differences in means and deviations
Dϕ	Differential dispersion
FDR	False discovery rate
FPR	False positive rate
GAMLSS	Generalised additive models for location, scale and shape
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype–Tissue Expression
HM	Hierarchical model
HMM	Hierarchical mixture model
HPD	Highest posterior density
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney renal clear cell carcinoma
LDA	Linear discriminant analysis
LFC	Log fold change
LIHC	Liver hepatocellular carcinoma
LSVM	Linear SVM
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma

Abbreviation	Description
MAD	Median absolute deviation
MCMC	Markov chain Monte Carlo
mRNA	messenger RNA
MSE	Mean squared error
NB	Negative binomial
NCBI	National Center for Biotechnology Information
NOS	Not otherwise specified
NSCLC	Non-small cell lung carcinoma
PRAD	Prostate adenocarcinoma
PSVM	SVM with polynomial kernel
RBF	Radial basis function
RF	Random forest
RIN	RNA integrity number
RLE	Relative log expression
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
RSVM	SVM with RBF kernel
SAGE	Serial analysis of gene expression
SAM	Significance Analysis of Microarrays
SVM	Support vector machine
TCGA	The Cancer Genome Atlas
THCA	Thyroid carcinoma
TMM	Trimmed mean of M-values
TPR	True positive rate
TRF	Random forest with distance-to-median transformed features

Symbol	Description
g	Number of genes represented in a gene expression dataset
m_μ	Location hyperparameter for log-normal prior on mean
m_ϕ	Location hyperparameter for log-normal prior on dispersion
n	Number of samples in a dataset
n_A	Number of samples in group A
n_B	Number of samples in group B
$\hat{R}$	Gelman–Rubin diagnostic
s	Sample standard deviation
s^2	Sample variance
v_μ	Scale hyperparameter for log-normal prior on mean
v_ϕ	Scale hyperparameter for log-normal prior on dispersion
y	Set of observed RNA-seq counts for all genes and samples in a dataset
$\bar{y}$	Sample mean of y
y_{ij}	Observed count for gene j in sample i
z_j	Mixture component indicator for gene j
$\Gamma(\cdot)$	Gamma function
γ	Set of all hyperparameters
γ_μ	Set of hyperparameters for prior on mean
γ_ϕ	Set of hyperparameters for prior on dispersion
θ	Set of means and dispersions for all genes: $(\mu_j, \phi_j)\,,\ j = 1, \ldots, g$
θ_j	Set of mean and dispersion for gene j: (μ_j, ϕ_j)
λ	Proportion of differentially distributed genes in HMM or Poisson rate parameter
μ	Mean
ϕ	Negative binomial dispersion
σ	Standard deviation
σ^2	Variance

Abstract

Data from genome-wide gene expression studies provides a wealth of information on diseases such as cancer, which can lead to insights into disease mechanisms and advances in diagnosis and treatment. Analysis of expression data is most commonly aimed at identifying genes whose mean expression levels are increased or decreased in disease compared to normal tissue, or between disease subtypes – differential expression analysis. However, there is strong evidence that changes in the variability of gene expression, without a difference in mean, can also be relevant. Genes related to cancer have been shown to have changes in the variability of their expression between normal and tumour tissue, and these differentially variable genes have also been found to be informative for diagnostic and prognostic cancer classification. The research presented in this thesis addresses several aspects of research on differential gene expression variability, and the broader concept of differential distribution, defined as any difference in the distribution of expression values between groups.

This work makes three contributions to knowledge, relating to cancer classification, identification of differentially variable or distributed genes, and the biology of differential variability and distribution in cancer. Contribution 1 extends previous work by demonstrating that genes identified by differential variability or distribution can be used to classify closely related cancer subtypes, rather than purely diagnostic or prognostic classification. Contribution 2 is a Bayesian hierarchical model for RNA-seq data that provides tests for differential expression, variability and distribution. The performance of each test is compared with existing methods on simulated data and on real RNA-seq datasets modified to artificially introduce changes in expression between groups. The differential expression test is competitive with state-of-the-art methods, and the differential variability test improves on existing methods, particularly for small sample sizes. The differential distribution test is the first such test available for RNA-seq data. Contribution 3 builds on previous work by providing the first clear demonstration that differential variability and differential distribution analyses can identify cancer-related genes, and that differential expression and differential variability analyses identify distinct sets of cancer-related genes, each with different biological functions.

Overall, this research confirms and extends previous findings showing that changes in expression variability and distribution in cancer are both of biological significance and informative for classification. As well as further demonstrating the need to look beyond differential expression to a comprehensive assessment of changes in gene expression distributions, this work provides a method that enables the identification of these differentially distributed genes.