



COPYRIGHT AND USE OF THIS THESIS

This thesis must be used in accordance with the provisions of the Copyright Act 1968.

Reproduction of material protected by copyright may be an infringement of copyright and copyright owners may be entitled to take legal action against persons who infringe their copyright.

Section 51 (2) of the Copyright Act permits an authorized officer of a university library or archives to provide a copy (by communication or otherwise) of an unpublished thesis kept in the library or archives, to a person who satisfies the authorized officer that he or she requires the reproduction for the purposes of research or study.

The Copyright Act grants the creator of a work a number of moral rights, specifically the right of attribution, the right against false attribution and the right of integrity.

You may infringe the author's moral rights if you:

- fail to acknowledge the author of this thesis if you quote sections from the work
- attribute this thesis to another author
- subject this thesis to derogatory treatment which may prejudice the author's reputation

For further information contact the University's Director of Copyright Services

sydney.edu.au/copyright

STATISTICAL METHODS FOR THE ANALYSIS
AND INTERPRETATION OF RNA-SEQ DATA

ELLIS PATRICK

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
The University of Sydney
8 April 2014

ABSTRACT

In the post-genomic era, sequencing technologies have become a vital tool in the global analysis of biological systems. RNA-Seq, the sequencing of messenger RNA, in particular has the potential to answer many diverse and interesting questions about the inner workings of cells. Despite the decreasing cost of sequencing data, the majority of RNA-Seq experiments are still suffering from low replication numbers. The statistical methodology for dealing with low replicate RNA-Seq experiments is still in its infancy and has room for further development. Incorporating additional information from publicly accessible databases may provide a plausible avenue to overcome the shortcomings of low replication. Not only could this additional information improve on the ability to find statistically significant signal but this signal should also be more biologically interpretable.

This thesis is separated into three distinct statistical problems that arise when processing and analysing RNA-Seq data. Firstly, the use of experimental data to customise gene annotations is proposed. When customised annotations are used to summarise read counts, the corresponding measures of transcript abundance include more information than alternate summarisation approaches and offer improved concordance with qRT-PCR data. A moderation methodology that exploits external estimates of variation is then developed to address the issue of small sample differential expression analysis. This approach performs favourably against existing approaches when comparing gene rankings and sensitivity. With the aim of identifying groups of miRNA-mRNA regulatory relationships, a framework for integrating various databases of prior knowledge with small sample miRNA-Seq and mRNA-Seq data is then outlined. This framework appears to identify more signal than simpler approaches and also provides highly interpretable models of miRNA-mRNA regulation. To conclude, a small sample miRNA-Seq and mRNA-Seq experiment is presented that seeks to discover miRNA-mRNA regulatory relationships associated with loss of Notch2 function and its links to neurodegeneration. This experiment is used to illustrate the methodologies developed in this thesis.

PUBLICATIONS AND PRESENTATIONS

Some of the methods, concepts, analyses and results in this thesis have appeared previously in the following:

PUBLICATIONS

Ellis Patrick, Michael Buckley, David Ming Lin, and Yee Hwa Yang. Improved moderation for gene-wise variance estimation in RNA-Seq via the exploitation of external information. *BMC Genomics*, 14 Suppl 1:S9, 2013.

Ellis Patrick, Michael Buckley, and Yee Hwa Yang. Estimation of data-specific constitutive exons with RNA-Seq data. *BMC Bioinformatics*, 14:31, 2013.

Pengyi Yang, Ellis Patrick, Shi-Xiong Tan, Daniel Fazakerley, James Burchfield, Christopher Gribben, Matthew Prior, David James and Yee Hwa Yang. Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. *Accepted in Bioinformatics*, 2013.

PRESENTATIONS

Optimising a genomic annotation for the analysis of RNA-Seq data. *Australasian Microarray and Associated Technologies Association (AMATA)*, 2010, Hobart, TAS.

Improved moderation for gene-wise variance estimation in RNA-Seq. *Australian Statistical Conference (ASC)*, 2012, Adelaide, SA.

Improved moderation for gene-wise variance estimation in RNA-Seq. *Asia Pacific Bioinformatics Conference (APBC)*, 2013, Vancouver, BC.

An integrative analysis of miRNA-Seq and mRNA-Seq data. *Western North American Region of the International Biometrics Society (WNAR)*, 2013, Los Angeles, CA.

An integrative analysis of miRNA-Seq and mRNA-Seq data. *The 57th Annual Meeting of the Australian Mathematical Society (AustMS)*, 2013, Sydney, NSW.

"A bird doesn't sing because it has an answer, it sings because it has a song."

- Maya Angelou

ACKNOWLEDGMENTS

Through the course of my studies I have been incredibly fortunate to have been surrounded by so many amazing personalities. While I could not possibly list them all, these amazing people have shaped my understanding of statistics, biology and life in general.

I would like to start by thanking my supervisor Jean for everything she has done for me. She has been an inspiration, incredibly patient and always made me feel like my opinions have value. I would not be where I am now, for so many reasons, without her. I am so glad that she took me on five years ago and could never effectively express my appreciation for sharing all her experience, ideas and time.

I would like to also thank all of the following: My associate supervisor Mike for being a constant source of knowledge, reassurance and encouragement. Dave Lin for being such an awesome bloke, always having time for me and having such an articulate knowledge of biology. Also, the rest of his lab for making me feel so welcome. Anna Campain for all our conversations and making me feel very much like the younger sibling. Garth Tarr for simply putting up with me, he's made the past four years so much easier. Neville Weber for not only sparking my interest in statistics but always finding time to offer me guidance and advice. John Ormerod for becoming both a friend and mentor. Samuel Mueller for always leaving his door open. Everyone that has participated in Monday lab meetings and all the faculty on the 8th floor for any advice, wisdom or laughs they have shared with me. Pengyi for coping with my bluntness. David James for appreciating my sense of humour. Terry Speed for his valuable opinions and advice. Jennifer Chan and Ross Sparks for giving me a chance to cut my teeth.

Of course I could never have made it through my studies without the emotional and financial support of my family. I have never felt any pressure to be anything other than myself or do anything other than those things I'd like to do. I'd also like to thank my fiance Tanya for falling in love with this poor student and for all her patience and support.

CONTENTS

1	Introduction	1
1.1	Background	2
1.1.1	The biology of a cell	2
1.1.2	Technologies for measuring gene expression	4
1.1.3	Common biological questions	6
1.1.4	A common RNA-Seq pipeline	6
1.2	Motivational data	9
1.3	Outline of the thesis	10
2	Summarisation – The estimation of data-specific constitutive exons with RNA-Seq data	11
2.1	Genes, alternative splicing and isoforms	12
2.2	Summarisation of read counts	15
2.2.1	Estimation of constitutive exons	16
2.2.2	Processing exon annotation	18
2.3	exClust - Estimate data-specific constitutive exons through clustering	20
2.4	Evaluation study	23
2.4.1	Data	23
2.4.2	Evaluation criterion and results	25
2.5	Conclusions and further discussion	32
3	Differential expression – Improved moderation for gene-wise variance estimation in RNA-Seq via the exploitation of external information	34
3.1	Modelling RNA-Seq data	35
3.2	Tshrink+	39
3.3	Evaluation study	40
3.3.1	Data	41
3.3.2	Evaluation strategies and results	42
3.4	Conclusions and further discussion	50
4	Functional, network & pathway analysis – Using pathway information to help integrate small sample miRNA-Seq and mRNA-Seq data	52
4.1	miRNA	53
4.2	Combining p-values	55
4.2.1	Simulation	56
4.3	pMiM - Pathway, microRNA and mRNA integration	60
4.4	Evaluation Study	65
4.4.1	Data	65
4.4.2	Evaluation strategies and results	66
4.5	Conclusions and further discussion	71
5	Case study of the Lin data	73
5.1	Design	74
5.2	Mapping	75
5.3	Summarisation	77
5.4	Normalisation	80
5.5	Differential expression	82

5.6	Functional, network & pathway analysis	87
5.6.1	Pathway analysis	87
5.6.2	Integration of miRNA and mRNA data	88
5.7	Conclusions and further discussion	89
6	Conclusion	91
A	Additional information for Chapter 2	95
A.1	Detection of differential alternative splicing	96
A.2	Additional figures and tables	96
B	Additional information for Chapter 3	98
B.1	Normalisation for Bottomly dataset	99
B.1.1	GC content	99
B.1.2	Other Technical effects	99
B.2	Additional Figures	100
C	Additional information for Chapter 4	103
C.1	Additional Figures	104
	BIBLIOGRAPHY	105

LIST OF FIGURES

Figure 1.1	The Central Dogma of Molecular Biology – An illustration of the Central Dogma of Molecular Biology, a simplified conceptualisation of the flow of genetic information within the cell. Information stored in DNA can be transcribed into messenger RNA which can then move outside of the nucleus and be translated into a functional protein.	3
Figure 1.2	RNA-Seq analysis pipeline – A flow chart describing a typical RNA-Seq analysis pipeline. This pipeline consists of two broad steps, data processing and data analysis. Data processing includes aligning reads to a genome (mapping), summarising how many of these reads lie in particular regions of the genome (summarisation) and correcting for any systematic technical variation (normalisation). Data analysis consists of identifying genes that have changed in expression between two conditions (differential expression) and some higher level analysis to improve the interpretability of the results (functional, network & pathway analysis).	7
Figure 2.1	Transcription and alternative splicing – A toy example demonstrating how the information stored in one gene can be transcribed, spliced and translated to form multiple distinct proteins. Gene 1, contains four exons, three introns and a 5' and 3' UTR. This information can be transcribed to form a pre-mRNA that has had a cap added to its 5' end. The introns in the pre-mRNA are spliced out to form a mature mRNA. This pre-mRNA can be alternatively spliced to form a mature mRNA that contains all four exons and a mature mRNA that has had its third exon removed. These two mature mRNA can then move outside of the nucleus of the cell into the cytoplasm to be translated into Protein A and Protein B.	13
Figure 2.2	Effect of differential alternative splicing on gene counts – A toy example of a gene with two isoforms is considered. The number of reads that aligned to each exon of the gene are provided for three biological samples. The sums of these exon counts are also included for each sample.	17

- Figure 2.3 **Processing exon annotation** – A graphic describing how the annotation of two overlapping genes is processed into an exon annotation appropriate for the use of exClust. The isoform annotation can be used to define a set of disjoint exon regions that could be rejoined to describe any of the known isoforms of the gene. It is these disjoint exon regions that are used as the exon annotation in exClust. Exon regions which overlap multiple genes are ignored. The set of UI exons are also shown for these two genes and are simply the exons that are present in all the annotated isoforms. 19
- Figure 2.4 **Identifying constitutive exons** – Plot of exons selected by exClust for a particular gene. A clustering dendrogram of the exons is formed by apply Ward’s linkage hierarchical clustering to the distance matrix $1 - \Sigma_g^E$. Cutting the dendrogram at the dashed red line results in the creation of three subgroups of exons (each box here contains a subgroup). For each subgroup the average coverage of the two exons in that subgroup with the highest coverage is calculated. The subgroup with the highest average coverage (the shaded subgroup) is selected to represent the DSC exons for this gene. 22
- Figure 2.5 **Concordance Plot** – Concordance plot with the RNA-Seq log fold changes on the y-axis and qRT-PCR log fold changes on the x-axis. For the RNA-Seq data we use the union of all exons within a gene to summarise our counts where a value of one is added to the count of every gene. The black circles are those genes for which the UI definition is non-empty. The blue triangles are the 386 genes for which the UI definition is empty. The red dots are those genes that our method identified as having a change in isoforms and had a non-empty UI definition. 28
- Figure 2.6 **Residual Plot** – After fitting a straight line through the plot in Figure 2.5, this figure plots on the y-axis the residuals for the genes identified as having a change in isoforms for three different annotations, union of all exons (black dots), UI definition (blue triangles) and exClust (red circles), and cufflinks (purple cross) ordered by qRT-PCR fold change. 29
- Figure 3.1 **The mean-variance relationship in RNA-Seq data on a log-log scale** – This figure demonstrates the strong mean-variance relationship observed in RNA-Seq data. For each gene from the ten Bottomly B6 samples described in Section 3.3 the mean and variance are calculated. These means and variances are then plotted against each other on a log-log scale. The blue line corresponds the line $y=x$, when the mean equals the variance. 36

Figure 3.2	Effect of utilising different sources of information on the estimation of λ – Variance estimates from the external datasets (Table 3.1) and gene length are used to aid in the estimation of the common variance functions of one hundred comparisons of n B6 and n D2 mouse striatum samples. The average λ value is plotted for each n comparison and information source for n ranging from two to five. The parameter λ is the ratio of the expected and average squared error of the gene sample variance to the common variance. The information source “None” corresponds to using no extra information, “Striatum” the RNA-Seq samples from Polymenidou et al (2011) and “Striatum Microarray” the microarray striatum samples from Bottomly et al (2011). The information sources have been sorted by their λ values for n equals two.	44
Figure 3.3	Comparing six DE methods on a 4 vs 4 comparison – One hundred random comparisons of four B6 and four D2 mouse striatum samples for six DE methods. Average TP and FP are calculated for the full range of p-value cut-offs. The TPR and FPR are plotted against each other in a) to form ROC curves and displayed in the region for FPR less than 0.01 as this is most relevant for calling DE. For any given FPR a method with a larger TPR is deemed to have ranked the genes better. In b) the number of TP (in bold) and FP are plotted for a range of p-value cut-offs. The x-axis is in log-scale. The grey dashed vertical line corresponds to a Bonferroni adjusted cut-off of 0.05.	47
Figure 3.4	Partial AUCs for a range of n vs n comparisons – One hundred random comparisons of n B6 and n D2 mouse striatum samples a performed for six DE methods for n ranging from two to five. For each method and n , partial areas under the ROC curves (partial AUC) are calculated for the regions of FPR less than 0.01	48
Figure 3.5	The number of True and False Positives for a range of n vs n comparisons – One hundred random comparisons of n B6 and n D2 mouse striatum samples a performed for six DE methods for n ranging from two to five. For each method and n , the conservative Bonferroni adjusted cut-off of 0.05 is used to calculate the average number of (a) True Positives and (b) False Positive are counted.	49
Figure 4.1	p-value cut-offs for various combination methods – A plot illustrating a p-value cut-off of 0.05 for various p-value combination methods in a two dimensional setting. The p-value cut-off is plotted in the negative z-score space so that a small p-value corresponds to a large positive z-score. The combination methods under consideration are Fisher (red), Stouffer (blue), maxP (pink) and OSP (green).	57

Figure 4.2	Data matrices and summary statistics – A visual representation of the input data matrices which include the matched mRNA- and miRNA-Seq data, a pathway database and the miRNA target binding predictions. Also represented are the corresponding summary statistics used which are the statistics from a moderated two-sample t-test performed on the mRNA-Seq data, the statistics from a moderated two-sample t-test performed on the miRNA-Seq data and the cross correlations or probability of observing the cross correlations between the miRNA- and mRNA-Seq data.	61
Figure 4.3	Number of mir-pathways that contain n genes before and after randomisation – For $n = 1, \dots, 20$ the average number of mir-pathways that contain n genes are plotted on a log-scale. These are plotted for mir-pathways calculated using TargetScan and KEGG, randomised TargetScan and KEGG, TargetScan and randomised KEGG and both randomised TargetScan and randomised KEGG.	68
Figure 4.4	Number of mir-pathways that contain 3, 4 or 5 genes before and after randomisation – For $n = 3, 4$ and 5 the average number of mir-pathways that contain n genes are plotted. These are plotted for mir-pathways calculated using TargetScan and KEGG, randomised TargetScan and KEGG, TargetScan and randomised KEGG and both randomised TargetScan and randomised KEGG.	69
Figure 4.5	TP vs FP from PubMed search – True Positives (TP) are plotted against False Positives (FP) for the small FPR region of Figure C.1. The plotted lines are for four methods, miRNA DE (black), cMimDE (blue), pMimDE (green) and pMimCor (red).	71
Figure 5.1	Number of mapped mRNA – (a) A histogram of the number of reads that were sequenced for each mRNA sample. (b) A bar plot of the percentage of these reads that mapped uniquely, failed to map and mapped to multiple regions on the genome.	76
Figure 5.2	Number of base pairs cut by cutAdapt – A histogram describing the number of reads that had a certain number of base pairs cut by cutAdapt.	78
Figure 5.3	Number of mapped miRNA – (a) A histogram of the number of reads that were sequenced for each miRNA sample. (b) A bar plot of the percentage of these reads that mapped uniquely, failed to map and mapped to multiple regions on the genome.	79
Figure 5.4	Plot of summarised miRNA counts vs mRNA counts – Gene counts from the mRNA and miRNA enriched samples are plotted against each other on a log-log scale. The hollow circles are annotated genes. The solid red circles are the annotated miRNA.	81

Figure 5.5	MA-plot illustrating TMM normalisation – An MA-plot generated for one of the WT mRNA samples. The y-axis is the log ratio of one of the WT samples and the average across all samples (M) and the x-axis are the average gene counts across all samples (A). The solid blue line is $M = 0$ and the dashed red line is the line fitted by TMM.	83
Figure 5.6	P-value histograms for four different methods – P-value histograms were generated from the comparison of WT and NCN mice using a two sample t-test (T), a two sample t-test using a fitted common variance (Tcommon), a moderated t-test (Tshrink) and a moderated t-test using variance information from the Borchelt mice (Tshrink+).	86
Figure A.1	Verification of the Poisson assumption for the MAQC and Wang data – The squared standardised residuals are plotted against the average for each gene in the a) MAQC and b) Wang datasets. The blue line is the $y = x$ line. The red circles correspond to the fitted points found using local smoothing.	97
Figure A.2	Gene exon counts – For two genes, a) ENSG00000103769 and b) ENSG00000076662, the exon counts for each sample are plotted. The first seven samples are brain and the second seven are UHR. A line is drawn between the points to make the behaviour of each exon easier to follow. Highlighted in red are the UI exons and dashed blue are the exClust exons.	97
Figure B.1	ROC of normalised data using arrays as truth – Average TPR and FPR are calculated from 100 random four B6 vs four D2 mouse striatum comparisons for four normalisation methods using results from an a) Affymetrix and b) Illumina array as truth. These are plotted against each other to form ROC curves. For any given FPR a method with a larger TPR is deemed to have ranked the genes better.	100
Figure B.2	Boxplots of log variances – Boxplots of the log variance of the within sample gene ranks for four normalisation methods. All normalisation methods on average reduce the variance of the ranking of the genes.	101
Figure B.3	Residual plots for the MAQC and Wang data – Average TPR and FPR are calculated from a) 100 random four B6 vs four D2 mouse striatum comparisons and b) 100 random five vs five D2 mouse striatum comparisons for six DE methods. These are calculated using results from an Affymetrix array experiment as truth. The TPR and FPR are plotted against each other to form ROC curves and displayed in the region for FPR less than 0.1 as this is most relevant for calling DE. For any given FPR a method with a larger TPR is deemed to have ranked the genes better. T and Tshrink both improve in performance relative to edgeR and DESeq when moving from the four vs four comparison to the five vs five comparison.	102

Figure C.1	ROC plot from PubMed search – ROC curves are plotted for various methods as described in <i>Literature search strategy</i> . The search term used is neurodegeneration. True Positive Rates (TPR) are plotted against False Positive Rates (FPR) for four methods, miRNA DE (black), cMimDE (blue), pMimDE (green) and pMimCor (red).	104
------------	--	-----

LIST OF TABLES

Table 2.1	MAQC correlations – A table showing two subsets of genes from the MAQC data: differential alternatively spliced genes (DAS) and not differentially alternatively spliced genes (non-DAS). For each set of genes the correlations between Union, UI, exClust and Cufflinks log fold changes are given. The given correlations are only calculated on the subset of genes for which the UI definition is non-empty and have finite log fold change.	31
Table 2.2	Wang correlations – A table showing two subsets of genes from the Wang data: differential alternatively spliced genes (DAS) and not differentially alternatively spliced genes (non-DAS). For each set of genes the correlations between Union, UI, exClust and Cufflinks log fold changes are given. The given correlations are only calculated on the subset of genes for which the UI definition is non-empty and have finite log fold change.	32
Table 3.1	Additional information sources – Variance estimates from these three datasets are be used to improve the estimation of the common variance function in the main analysis dataset.	42
Table 3.2	Using D2 variance estimates to estimate common variance of four B6 samples – The average λ values calculated using a random n D2 mouse striatum samples to estimate the variance of a random four B6 mouse striatum samples from one hundred simulations.	45
Table 4.1	Results for five simulations in evaluating the performance of the four p-value combination methods at testing H_A and H_B . The percentage of combined p-values less than 0.05 over 1 000 000 simulations (rounded to two decimal places) are reported.	59
Table 4.2	The number of significant mir-pathways – A table of the average number of significant mir-pathways calculated at various arbitrary p-value cut-offs. Significance is calculated for mir-pathways estimated using TargetScan and KEGG, randomised TargetScan and KEGG, TargetScan and randomised KEGG and both randomised TargetScan and randomised KEGG. Results are shown for both up-regulated miRNA and down-regulated miRNA.	70

Table 5.1	Design of the experiment – This table describes the design of the experiment performed by the Lin lab. Listed are the genotype, age and sex of the mice as well as whether they were treated with BSO. Also included is the lane of the flowcell that each sample was sequenced on.	75
Table 5.2	Number of reads summarised – Tabulated are the number of reads summarised by the Union, UI and exClust approaches. For each approach the number of genes with average counts greater than zero and twenty are also included.	79
Table 5.3	Table of results from mRNA DE – The number of DE genes are reported from the comparison of WT and NCN mice by four DE methods for an arbitrary 0.05 p-value cut-off; a two sample t-test (T), a two sample t-test using a fitted common variance (Tcommon), a moderated t-test (Tshrink) and a moderated t-test using variance information from the Borchelt mice (Tshrink+). The number of DE genes are also reported after adjusting for multiple testing using the Benjamini-Hochberg and Bonferroni methods.	84
Table 5.4	Table of results from miRNA DE – The number of DE miRNA are reported from the comparison of WT and NCN mice by four DE methods for an arbitrary 0.05 p-value cut-off; a two sample t-test (T), a two sample t-test using a fitted common variance (Tcommon), a moderated t-test (Tshrink) and a moderated t-test using variance information from the Borchelt mice (Tshrink+). The number of DE miRNA are also reported after adjusting for multiple testing using the Benjamini-Hochberg and Bonferroni methods.	85
Table 5.5	Table of results Goseq – Goseq was used to perform a pathway over-representation test on the list DE genes from Tshrink+. Listed are the pathways that had a p-value less than 0.05.	87
Table 5.6	Table of results from pMimCor for over expressed miRNA. – Listed are the top ten mir-pathways found using pMimCor whose miRNA were over expressed in the NCN samples. The corresponding miRNA and KEGG pathway are listed, as well as, the p-value from pMimCor, the miRNA Tshrink+ p-value and the number of genes in the mir-pathway.	89
Table 5.7	Table of results from pMimCor for under expressed miRNA. – Listed are the top ten mir-pathways found using pMimCor whose miRNA were under expressed in the NCN samples. The corresponding miRNA and Kegg pathway are listed, as well as, the p-value from pMimCor, the miRNA Tshrink+ p-value and the number of genes in the mir-pathway.	90
Table A.1	Log fold changes for different summarisation methods – For two genes, ENSG00000103769 and ENSG00000076662, log fold changes are reported for four summarisation methods and qRT-PCR from the MAQC dataset.	96

ABBREVIATIONS

3' UTR	Three prime untranslated region
5' UTR	Five prime untranslated region
bp	Basepair
BSO	Pro-oxidant buthionine sulphoximine
C6	C57BL/6J mouse strain
D2	DBA/2J mouse strain
DAS	Differentially alternatively spliced
DNA	Deoxyribonucleic acid
DE	Differentially expressed
DSC	Data-specific constitutive
FDR	False discovery rate
FP	False positives
FPKM	The number of fragments per kilobase of exon per million fragments that were mapped
FPR	False positive rate
mRNA	Messenger-RNA
mRNA-Seq	Sequencing of mRNA molecules
miRNA	Micro-RNA
miRNA-Seq	Sequencing of miRNA molecules
NCN	Nestin-cre/N2 flox
qRT-PCR	Quantitative real time polymerase chain reaction
RNA	Ribonucleic acid
RNA-Seq	Sequencing of RNA molecules
ROC	Receiver operating characteristic
pAUC	Partial area under the receiver operator curve
TMM	Trimmed means of M-values
TP	True positives
TPR	True positive rate
WT	Wild type

INTRODUCTION

1.1 BACKGROUND

Since the announcement in May 2007 of the sequencing of James Watson's entire genome using 454 Life Sciences, next-generation sequencing technologies made headlines around the world, demonstrating the advancement in genome sequencing. These technologies saw the beginning of projects such as the \$1000 genome project, the 1000 genome project and the ENCODE project, all of whose completion was expected to make huge impacts on the understanding of human health. These projects have highlighted how complex the biology of a cell and all its regulatory mechanisms are. While these new technologies have ushered in a torrent of new biological knowledge, they have also underlined that we have a lot more knowledge to acquire.

Next-generation (or second-generation) high-throughput sequencing produces datasets that are incredibly exciting for statisticians. While these technologies can be used to answer a vast array of biological questions, many of the datasets that are produced lie squarely in the domain of very small n , extremely large p . That is, it is not unusual to see datasets with three or fewer biological replicates in each condition making measurements on tens of thousands of genes. This framework flies in the face of classical statistics and provides an exciting new world for statisticians to explore.

1.1.1 *The biology of a cell*

Over the past ten years, the scientific community has begun to appreciate how complex the biological system within a cell really is. Just as the community is on the verge of cracking a problem that is thought to provide a significant leap in understanding and treatment of human health, another level of complexity in the cell is discovered. Epigenetics was one such paradigm shift. Biological mechanisms such as alternate splicing, methylation, histone modification and non-coding RNA, have all provided insights into the complexity of the cell and its regulation. It has become apparent that we are not just a product of our DNA.

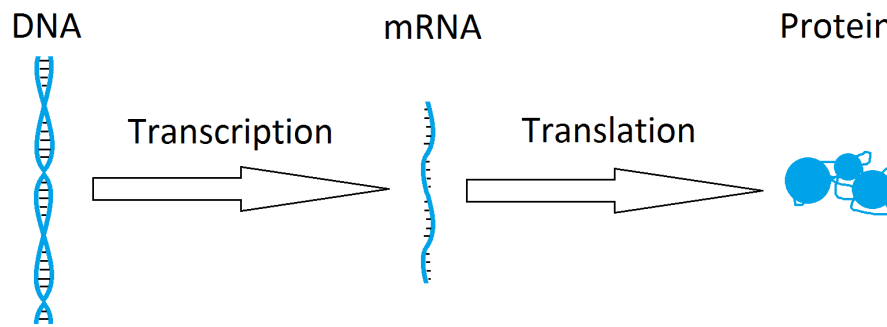


Figure 1.1: **The Central Dogma of Molecular Biology** – An illustration of the Central Dogma of Molecular Biology, a simplified conceptualisation of the flow of genetic information within the cell. Information stored in DNA can be transcribed into messenger RNA which can then move outside of the nucleus and be translated into a functional protein.

When first learning about cell biology it is often useful to consider a simplified flow of information (Figure 1.1). The Central Dogma of Molecular Biology (Crick, 1970) states that in the nucleus of your cells, information contained on your DNA (often called genes) can be transcribed into messenger RNA which can move outside of the nucleus and be translated into a functional protein. This flow of information is generally referred to as gene expression. How it is regulated, inhibited and can be manipulated is of great interest to the scientific community.

Deoxyribonucleic acid (DNA) is located in the nucleus of the cell and consists of two complementary strands of sequences of nucleotides that are arranged in a double helix. A strand of DNA consists of a three billion long sequence of the four nucleotides adenine (A), thymine (T), cytosine (C) and guanine (G). The two strands are complementary in the sense that if you have an adenine in a position on one strand then you know that there is a thymine in the same position on the other strand. Likewise if there is a cytosine in a position on one strand then there is a guanine in the same position on the other strand. While the DNA within a cell has a very complex structure, it is generally annotated as a single linear sequence of base-pairs (A-T, T-A, C-G, G-C) which is split up into chromosomes (23 for humans).

Around 1–2% of human DNA contains information, or blueprints, for building proteins. These regions are generally referred to as genes. Transcription is a process that copies the information of a gene into a complementary single stranded Ribo-

nucleic acid (RNA) molecule, messenger RNA (mRNA). RNA again consists of four nucleotides adenine (A), uracil (U), cytosine (C) and guanine (G) with uracil replacing its DNA counterpart thymine. After being transcribed, mRNA undergoes post-transcriptional modifications and then moves outside of the nucleus into the cytoplasm of the cell to be translated into a protein. Some of these post-transcriptional modifications are referred to as alternative splicing and allow the information of a gene to be arranged in many different ways (isoforms), such that the information stored in one gene could be translated to make many different proteins.

1.1.2 *Technologies for measuring gene expression*

There are many ways to measure gene expression. This thesis contains gene expression data generated with qRT-PCR, microarrays and next generation high-throughput sequencing. These approaches are outlined in the following.

1.1.2.1 *qRT-PCR*

Quantitative real time polymerase chain reaction (qRT-PCR) is often seen as a gold standard for measuring gene expression and will often be used to validate a small set of genes. The effectiveness of qRT-PCR relies on the design of a 12 – 500 basepair probe that is complementary and specific to a portion of the mRNA of interest. A fluorescent dye can be attached to these probes making it possible to quantify how many of them bind to sequence of interest in the sample. QRT-PCR is a cyclic process of iteratively doubling the amount of product in the sample via PCR while measuring the fluorescence emitted by the labelled probes. By tracking this fluorescence and comparing the rate that it doubles to a control gene, the relative quantity of a mRNA of interest can be measured.

1.1.2.2 *Microarrays*

Microarrays facilitated the high-throughput measurement of gene expression making it possible to measure the expression of tens of thousands of genes simultaneously. They achieved this by arranging thousands of probes spatially on a slide, either printed

on directly in spots or coded onto beads. After amplification, fluorescent labels are attached to the mRNA samples which are then washed across the slides and treated so that the mRNA will hybridise (or bind to) any complementary probes on the array. Relative quantities of mRNA can then be determined by exciting the fluorescent labels with a laser and observing the intensity of light from different spots on the array.

One of the key disadvantages of microarrays is their reliance on probes. These probes have to be known in advance and generally selected in such a way so as to cover as much of an organisms transcriptome as possible. Some probes may also suffer from cross-hybridisation, that is, mRNA binding to probes that were not specifically designed to detect them.

1.1.2.3 *High-throughput sequencing*

Next generation high-throughput sequencing breaks the reliance on having to design probes for an experiment. While high-throughput sequencing platforms differ in their chemistry and protocols, their processed outputs are generally similar. The sequencing platforms take a sample of fragmented RNA as input and then read off 25–400 base pair regions at the ends of these fragments. The output of these sequenced regions, sequences of base pairs, are referred to as *reads*. These reads are used to infer the presence and quantity of RNA in the sample.

The development of high-throughput sequencing technologies has made it possible to sequence the transcriptome at a much higher resolution and coverage than was previously available. Sequencing of mRNA samples (RNA-Seq) has a dynamic range larger than that of microarrays (Wang *et al.*, 2009). This, combined with its high level of reproducibility (Mortazavi *et al.*, 2008) and falling cost, makes high-throughput sequencing technologies an increasingly attractive alternative to microarrays for transcriptome analysis. The three most prominent sequencing platforms are 454 Life Sciences, Illumina and Applied Biosystems SOLiD, all having their own distinct advantages.

1.1.3 Common biological questions

There are many uses for high-throughput sequencing, with the technology being used to address various biological problems. These problems can be broken up into two main categories:

1. reassembling what we measured to find out what was present (de novo assembly, SNP identification, motif finding) and
2. quantifying and comparing how much of a product was present (differential expression, tissue profiling).

High-throughput sequencing can also be used to sequence various preselected subsets of RNA and/or DNA to sequence. These include:

- DNA-Seq** — genome sequencing,
- RNA-Seq** — transcriptome sequencing,
- miRNA-Seq** — microRNA sequencing and
- ChIP-Seq** — chromatin immunoprecipitation sequencing.

All of these various biological problems and applications have their own technical intricacies which have themselves generated many interesting and challenging computational and statistical problems. In this thesis we will consider some of the analytical problems that arise when detecting differentially expressed genes with RNA-Seq data.

1.1.4 A common RNA-Seq pipeline

Many RNA-Seq experiments are generated with the common biological goal of identifying differentially expressed genes or transcripts – that is, to determine if the transcription of any genes is different between any two phenotypically distinct cellular populations. A typical RNA-Seq data analysis workflow with this focus consists of many steps (Oshlack *et al.*, 2010); these steps generally consist of mapping, summarisation, normalisation, differential expression and systems biology (functional, network & pathway analysis). This workflow is illustrated in Figure 1.2. We will consider mapping

and summarisation as *data processing* steps and differential expression and functional, network & pathway analysis as *data analysis* steps.

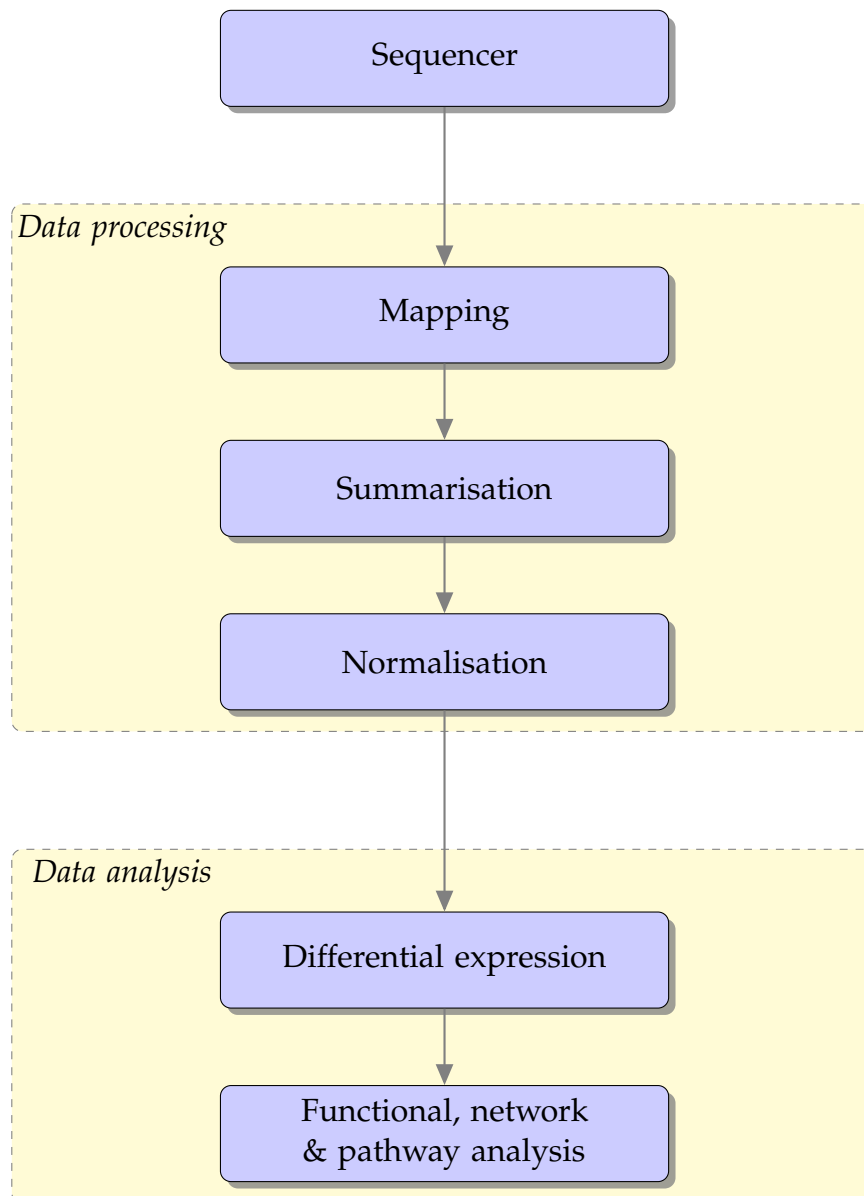


Figure 1.2: **RNA-Seq analysis pipeline** – A flow chart describing a typical RNA-Seq analysis pipeline. This pipeline consists of two broad steps, data processing and data analysis. Data processing includes aligning reads to a genome (mapping), summarising how many of these reads lie in particular regions of the genome (summarisation) and correcting for any systematic technical variation (normalisation). Data analysis consists of identifying genes that have changed in expression between two conditions (differential expression) and some higher level analysis to improve the interpretability of the results (functional, network & pathway analysis).

Mapping is the process of aligning reads to a reference genome (or transcriptome) and inferring from where they may have been transcribed. Most sequencing technologies are limited in the length of the read they can sequence and hence sequence limited intervals of fragmented transcripts. Mapping millions, billions or trillions of reads back to reference genomes that have billions of base pairs is both computationally and statistically intensive (Langmead *et al.*, 2009). The process of identifying and aligning various splice variants only further adds to this bioinformatics burden (Bona *et al.*, 2008; Trapnell *et al.*, 2009; Bryant *et al.*, 2010; Wang *et al.*, 2010b).

Summarisation is the process of simplifying the mapping information into read counts (or expression values) for all genes (or isoforms) of interest. While identifying the presence of an isoform is difficult, as many of these transcript fragments are present in multiple isoforms, it is also a statistically challenging problem to quantify expression of these isoforms (Jiang and Wong, 2009; Li *et al.*, 2010; Trapnell *et al.*, 2010). To avoid inferring isoform specific expression, an alternative approach is to count how many reads lie within either exonic or genomic regions (Bullard *et al.*, 2010). This will produce a large matrix of read counts generally having dimension in the order of tens of thousands of rows (genes) by hundreds, tens or even less than ten columns (samples).

Similar to other complex molecular datasets, before analysing RNA-Seq data, it is important to normalise or adjust for any systematic technical variation that may have arisen in the measurement process. The largest abnormality, generally observed in a RNA-Seq experiment, is that different samples may have been sequenced to different depths (library sizes) and hence can have very different amounts of total reads mapped (Bullard *et al.*, 2010; Robinson and Oshlack, 2010). This can occur due to differences in the total amount of material sequenced, lane effects or biological processing steps such as ribosome and adaptor read removal. The GC content of a gene (the abundance of cytosine and guanine in the sequence of the gene) can also affect the amplification process during sequencing and may often cause biases between lanes or batches (Benjamini and Speed, 2012).

A gene is called differentially expressed (DE) if its expression has changed between conditions, e.g. treatment vs control. There are many methods that have been optim-

ised to identify differentially expressed genes in RNA-Seq data (Robinson *et al.*, 2010; Anders and Huber, 2010; Hardcastle and Kelly, 2010; Srivastava and Chen, 2010; Li *et al.*, 2010). As many RNA-Seq experiments have small sample sizes, most methods share information between genes; this is referred to as moderation. Methods also often vary in the way that they model the strong mean-variance relationship observed in RNA-Seq data.

Functional, network or pathway analysis is often performed to improve the interpretability or power of differential expression results. The large number of genes analysed in most experiments can be both a blessing and a curse. The large number of genes, often combined with small sample sizes, creates problems with multiple testing and hence the ability to detect statistical significant differences in expression. Conversely, the large number of genes also creates the possibility of finding a large number of differentially expressed genes which can make interpretation of results quite difficult. Analysing genes in terms of pathways, networks or function groupings can alleviate both of these issues.

1.2 MOTIVATIONAL DATA

This thesis was motivated by an ongoing collaboration with Associate Professor David Lin of Cornell University. The Lin Lab primarily studies the development and degeneration of the nervous system using the mouse olfactory system as a model. Once neurons are born, they are exposed to a variety of environmental insults that must be properly dealt with to avoid degeneration. Neurodegenerative disorders, such as Alzheimer's disease, are thought to arise in part due to a failure to deal with this increased stress.

Guided by previous research we designed an experiment to measure both mRNA and miRNA transcription in the brains of mice that are exposed to various stressors using RNA-Seq. The design of this experiment is outlined in Section 5.1. Due to the costs associated with both obtaining and sequencing samples, our experimental design has very few biological replicates. Producing reliable and interpretable results in situations of low sample size is a significant statistical challenge. This in itself has provided

ample inspiration for the development of the novel statistical methodology presented in this thesis.

1.3 OUTLINE OF THE THESIS

In this thesis, several statistical issues associated with the processing and analysis of RNA-Seq data are proposed as well as the techniques for evaluating their effectiveness in practice. Some of these methods, concepts, analyses and results have already been published (or are currently under review) by the author.

The scaffold of this thesis has been structured to replicate that of the RNA-Seq analysis pipeline outlined in Figure 1.2 on page 7. It contains three chapters on the summarisation, differential expression and functional, network & pathway analysis steps of this pipeline in addition to a case study. The first of these, Chapter 2, proposes a novel way of approaching summarisation which uses experimental data to customise gene annotations and includes work published in Patrick *et al.* (2013a). Chapter 3 outlines a novel moderation methodology to improve small sample differential expression analysis by exploiting external estimates of variation and includes work published in Patrick *et al.* (2013b). Chapter 4 develops a novel framework for integrating various databases of prior knowledge with small sample miRNA-Seq and mRNA-Seq data to build meaningful and interpretable models of miRNA-mRNA regulation. Finally, Chapter 5 demonstrates the proposed methods from the previous chapters on our motivating dataset in a case study.

SUMMARISATION – THE ESTIMATION OF DATA-SPECIFIC
CONSTITUTIVE EXONS WITH RNA-SEQ DATA

RNA-Seq has the potential to answer many diverse and interesting questions about the inner workings of cells. However, estimating changes in the overall transcription of a gene is not always straight forward. In this chapter we will describe the difficulties associated with the summarisation step of the RNA-Seq analysis pipeline. Following this we will propose the concept of data-specific constitutive exons and a methodology for estimating these, exClust. When applied on two real datasets, exClust includes more than three times as many reads as the standard UI method, improves concordance with qRT-PCR data and is shown to produce robust estimates of overall gene transcription.

2.1 GENES, ALTERNATIVE SPLICING AND ISOFORMS

A simplified explanation of the Central Dogma of Molecular Biology was introduced in Section 1.1.1 and Figure 1.1. Figure 1.1 illustrated how the information stored in DNA can be transcribed into a messenger RNA (mRNA) and then for many genes translated into a protein. In the following we will expand on our previous explanation to include the concept of alternative splicing and discuss the implications this process may have on a RNA-Seq analysis. However, in short, alternative splicing is a process that allows different proteins to be coded for from the same genetic region (gene).

A gene is commonly seen as a fundamental unit in mRNA biology. While the term gene is commonly used, its usage and meaning has changed over time as our knowledge of the genome, its transcription and regulation has increased. We see it appropriate to use the definition that *a gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products* (Gerstein *et al.*, 2007). This definition allows for a gene to be transcribed into many products that may have different or even contrary functions (Latchman, 1996). This definition could in itself steer the direction of an analysis as one must decide whether the activity of a genomic region or its products is of primary interest. Figure 2.1 illustrates a toy example of Gene 1. This gene generally consists of many sub-components such as exons and introns. Exons contain information that can be translated to form amino acids, the building blocks of proteins. While introns are regions that lie between the exons and not included in a mature mRNA. The five prime and three prime untranslated regions (5' and 3' UTR) are regions on

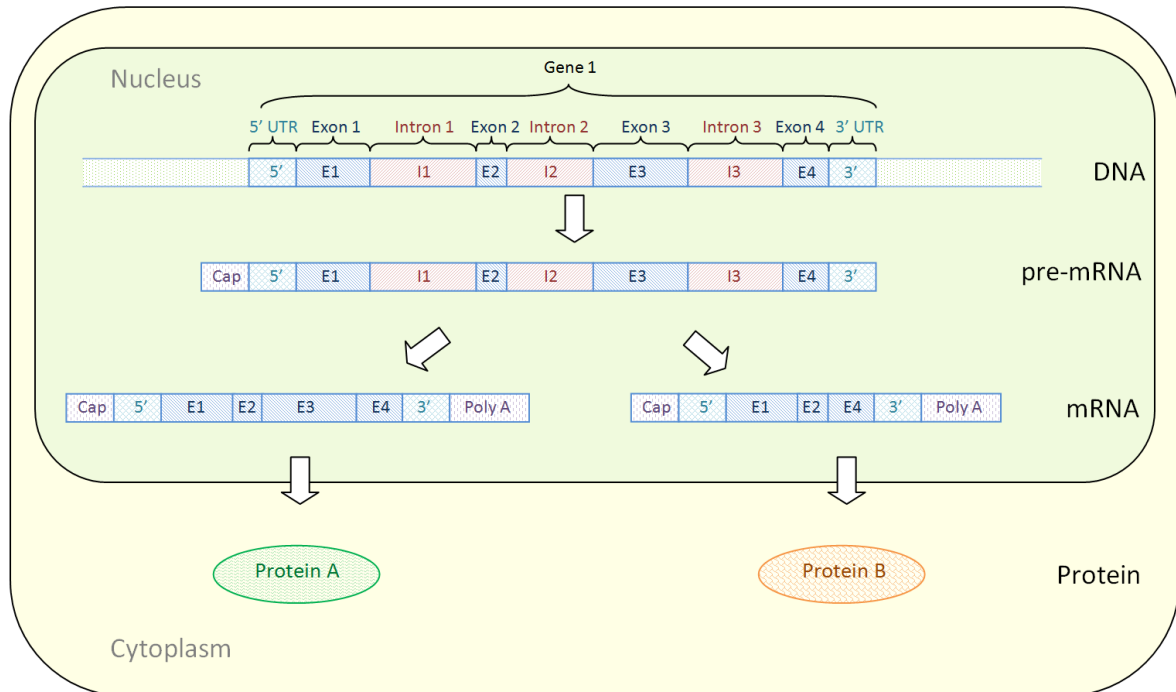


Figure 2.1: **Transcription and alternative splicing** – A toy example demonstrating how the information stored in one gene can be transcribed, spliced and translated to form multiple distinct proteins. Gene 1, contains four exons, three introns and a 5' and 3' UTR. This information can be transcribed to form a pre-mRNA that has had a cap added to its 5' end. The introns in the pre-mRNA are spliced out to form a mature mRNA. This pre-mRNA can be alternatively spliced to form a mature mRNA that contains all four exons and a mature mRNA that has had its third exon removed. These two mature mRNA can then move outside of the nucleus of the cell into the cytoplasm to be translated into Protein A and Protein B.

the 5' and 3' ends of a mRNA that do not translate into protein. These regions generally assist the translational and regulatory machinery of the cell.

Alternative splicing is a biological mechanism to expand protein diversity from a limited gene pool (Maniatis and Tasic, 2002). The implications of this mechanism are explored in Figure 2.1. In this figure information stored in Gene 1 can be translated into two distinct proteins, Proteins A and B. In the nucleus of the cell Gene 1 can be transcribed to form a pre-mRNA. This pre-mRNA contains exons, introns and the 5' and 3' UTR. The pre-mRNA has had a cap added at its 5' end which will ensure the mRNA's stability while it undergoes translation. The pre-mRNA is then spliced to form a mature mRNA that does not contain the non-coding introns, and multiple adenine bases are added to its 3' end (poly A tail). Different exons may also be spliced

from the pre-mRNA to give different mature mRNA transcripts. This is referred to as alternative splicing. In Figure 2.1 the mature mRNA that codes for Protein A retains all its exons while the third exon is spliced out of mRNA which codes for Protein B. These alternatively spliced mRNA molecules can then generally travel outside the nucleus of the cell into the cytoplasm where they are translated into unique proteins.

Alternative splicing generally refers to the inclusion of different exons in mature mRNA. Other alternative splicing events may include intron retention or alternative usage of 3' or 5' splice sites. These changes often lead to modifications in the encoded proteins and have been shown to play a critical role in development and disease (Lopez, 1998; Blencowe, 2000; Black, 2003). For simplicity, in the following we consider alternative splicing to be all mechanisms by which multiple and distinct mRNAs can be created from a single gene region including both alternative transcription start and alternative polyadenylation. The term *isoform* is used to refer to the blue-print of a distinct mRNA created from a particular gene region and *transcript* to refer to an actual mRNA molecule within a cell – an instance of the corresponding isoform.

Alternative splicing needs to be taken into consideration when analysing RNA-Seq data as it occurs ubiquitously within mammalian transcriptomes (Kim *et al.*, 2007). It is estimated in early studies that 50–80% of the approximately 25,000 human protein-coding genes are subject to alternative splicing (Modrek *et al.*, 2001; Johnson *et al.*, 2003; Lander *et al.*, 2001). This is further highlighted in a recent RNA-Seq study, where it is estimated that 86% of genes were found to be alternately spliced with a minor isoform frequency greater than 15% (Wang *et al.*, 2008).

Sequencing technologies produce reads of limited length, so each read is of a limited interval of a fragmented transcript. Sequencing only fragments of transcripts creates a significant bioinformatics burden in both the mapping and summarisation steps of the data analysis workflow. The longer an observed read, the higher the likelihood that it will span a splice junction. Identifying and aligning such reads is both computationally and statistically difficult as the number of possible splice junctions is large (Bona *et al.*, 2008; Trapnell *et al.*, 2009; Bryant *et al.*, 2010; Wang *et al.*, 2010a). Identifying the presence of a splice junction is only the first challenge; many of these transcript fragments are

present in multiple isoforms and it is a statistically challenging problem to estimate isoform-specific expression (Jiang and Wong, 2009; Li *et al.*, 2010; Trapnell *et al.*, 2010).

There are many biological questions that may be addressed with RNA-Seq data. A typical focus of RNA-Seq analysis is to identify differential expressed isoforms (Jiang and Wong, 2009; Li *et al.*, 2010; Trapnell *et al.*, 2010). However, there is still interest in studying RNA-Seq data at a gene level. That is, rather than estimating the abundance of each different isoform of a gene, it may be preferable instead to estimate the overall or total abundance of all the different isoforms of a gene. This may be of interest in itself, may be needed in cross-species or cross-platform comparisons and studies (Cox *et al.*, 2009), when there may be a lack of confidence in the quality of the organism's annotation, or where sequencing depth may not be sufficient to make inferences about the abundance of different isoforms within a gene. Many pathway annotations such as KEGG (Kanehisa *et al.*, 2012) are still annotated at gene level. Furthermore, such analyses avoid inferring transcript-specific expression, as the key focus is to count the number of reads that lie within either the region of exons or of genes.

2.2 SUMMARISATION OF READ COUNTS

Gene expression levels in RNA-Seq experiments reflect the number (or the amount) of mRNA that is within the samples. In a typical RNA-Seq experiment we can count the number of reads that map back to any given gene and associate this count with the amount of mRNA that gene produced. This is known as *summarisation*. For a given gene, this read count is a function of the abundance of its transcripts in the cell and the length of those transcripts. Our main interest is in the abundance of transcripts created from a gene, not the number of reads produced by the gene. This subtle difference is driven by the fact that a longer isoform will produce more reads than a shorter isoform if both are expressed at the same abundance. Due to alternative splicing, a gene can produce transcripts of different lengths. Thus, if the overall transcription of a gene does not change between conditions but the splicing does, this can result in a change of count. Accounting for this change in length using a method such as FPKM (the number of fragments per kilobase of exon per million fragments that were mapped)

(Trapnell *et al.*, 2010) would be appropriate if isoforms were mutually exclusive. Unfortunately there is often evidence of multiple isoforms for a gene being present. If the abundance of these isoforms could be accurately estimated (Trapnell *et al.*, 2010) it may be possible to estimate the rate of transcription by summing the FPKM of all isoforms of a gene. However, if only regions of the gene that were conserved across isoforms were considered, the changing lengths of transcripts would have no effect on the summarised count. These exons that are present in all isoforms within a gene are referred to as *constitutive exons* as they are common to all isoforms of a gene.

Figure 2.2 illustrates an example to demonstrate the effect differential alternative splicing can have on gene counts. In this example a gene with two isoforms is considered. Based on the observed exon counts it may be reasonable to assume that sample 1 and 2 only contain transcripts from isoform 1, while sample 3 only contains transcripts from isoform 2. It would probably be reasonable to assume that samples 1, 2 and 3 all contain a similar number of transcripts for this gene. If this gene's expression were measured only using the counts from the first exon, this gene would not be considered differentially expressed in any sample. However, if the expression of a gene is measured as the sum of its exon counts then sample 3 would generally be considered as differentially expressed from sample 1 and 2. This differential expression is driven primarily by the change in isoform length as opposed to a change in the number of transcripts created by the gene. Hence, if estimating transcript abundance the choice of summarisation method can influence conclusions.

2.2.1 Estimation of constitutive exons

In order to focus on the overall expression of a gene, rather than isoform-specific expression, the Union-Intersection (UI) (Bullard *et al.*, 2010) method is commonly used to define a set of constitutive exons for each annotated gene (Figure 2). The UI method produces a gene region consisting of all exons which are common to all known isoforms of the gene, excluding the regions which overlap with other genes (Bullard *et al.*, 2010). The UI definition is simple and conceptually relevant, but it is derived entirely from the collection of known isoforms which are present in an annotation database.



Isoform annotation					
Isoform 1					
Isoform 2					
Exon counts					
					Sum
Sample 1	251	50	460	28	789
Sample 2	246	53	472	24	795
Sample 3	254	49	0	26	329

Figure 2.2: **Effect of differential alternative splicing on gene counts** – A toy example of a gene with two isoforms is considered. The number of reads that aligned to each exon of the gene are provided for three biological samples. The sums of these exon counts are also included for each sample.

In general, there will be differences between this collection of annotated isoforms and the collection of isoforms actually present in the samples in the current experiment. In particular, for any given gene,

- the annotation may include isoforms which are not present in the current samples, and
- the current samples may include isoforms which are not present in the annotation.

In the first case, the UI definition selects exons which are conserved across the isoforms present in the data but may exclude some exons which are also conserved across isoforms present in the data but not across all isoforms in the annotation. This is an issue as the number of reads summarised for a gene can affect the sensitivity of tests for differential expression (Oshlack and Wakefield, 2009). Excluding exons unnecessarily would reduce the number of summarised reads for a gene and hence the power we have to estimate gene expression or detect changes in gene expression. In the second case, the UI definition may include an exon that is not conserved across all isoforms of a gene present within the current samples. The UI definition would then not give an accurate representation of the overall transcription of that gene. These two points

not only highlight the deficiencies in the UI method but also highlight the need for an alternate concept of exon conservation. As more transcripts are discovered and annotated, fewer exons can then be considered as constitutive. While constitutive exons may still have a nice interpretation with respect to the importance of those exons for the function of the gene, they will become less relevant when attempting to measure the rate of gene transcription.

To address these issues we propose in Section 2.3 a new method, exClust, inspired by work on exon arrays (Xing *et al.*, 2006) to estimate data-specific constitutive (DSC) exons using both annotation and experimental data. We will show that this new procedure retains two to three times more reads than the very conservative UI method, and hence extracts much more useful information from the data set. The new procedure also generates estimates of gene transcription which are independent of isoform composition, and potentially gives insights into gene annotation.

This chapter develops a methodology for identifying the DSC exons within a gene between two or more conditions. These methods are then evaluated on two publicly available datasets (M. A. Q. C. Consortium, 2006; Wang *et al.*, 2008). The estimates of differential gene expression produced by exClust are similar to that of the UI method when there has been a change in isoform composition. Our method performs consistently well on both datasets including more genes and more reads in the analysis than the UI method, and also offering improved concordance with qRT-PCR data.

2.2.2 Processing exon annotation

We assume that, for the organism of interest, at least one set of transcript annotation exists (it may be derived *de novo* or a combination of multiple annotations) and that annotation source has been selected for use in the analysis. From this annotation, we define for each gene what we call *exon regions*. These approximately correspond to the exons of the gene, but are in fact something subtly different: a set of disjoint exon regions that could be rejoined to describe any of the known isoforms of the gene. Some of the exon regions are whole exons; in other cases, exons may be divided into two or more pieces. This process is illustrated in Figure 2.3. In the remainder we will

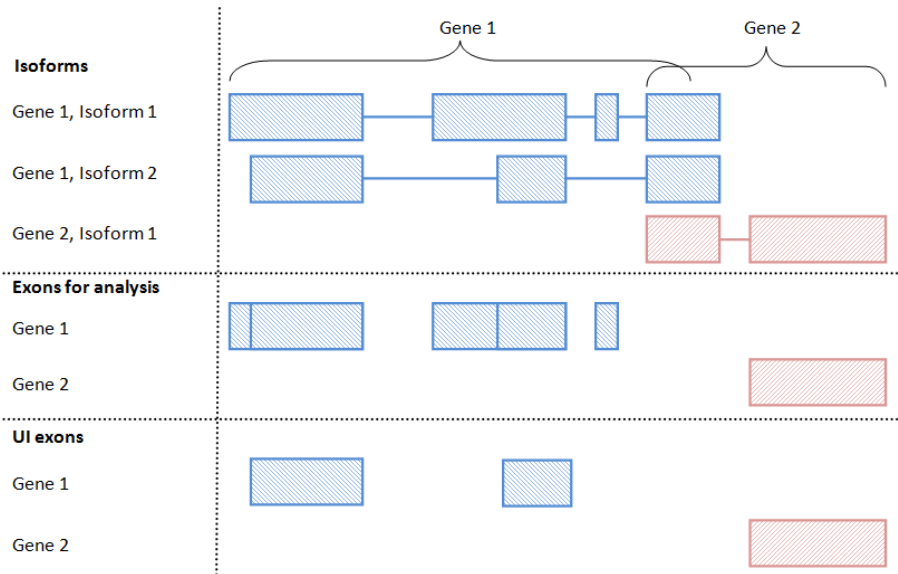


Figure 2.3: **Processing exon annotation** – A graphic describing how the annotation of two overlapping genes is processed into an exon annotation appropriate for the use of exClust. The isoform annotation can be used to define a set of disjoint exon regions that could be rejoined to describe any of the known isoforms of the gene. It is these disjoint exon regions that are used as the exon annotation in exClust. Exon regions which overlap multiple genes are ignored. The set of UI exons are also shown for these two genes and are simply the exons that are present in all the annotated isoforms.

ignore this distinction and use the term *exon* to refer to exon regions. If we ignore the distinction between exons and exon regions, or assume that all exon regions are whole exons, we are effectively using only the exon definitions from the annotation source, and not the isoform definitions. This is a key distinction between our approach and the UI method which depends heavily on the known annotated isoforms of each gene. The UI exons are those exons which are present in all the annotated isoforms. In the same way as the UI method, we also, as a final step, ignore any exon regions that overlap with multiple genes.

2.3 EXCLUST - ESTIMATE DATA-SPECIFIC CONSTITUTIVE EXONS THROUGH CLUSTERING

Let x_{ij} be the observed read count for the i^{th} exon of the j^{th} sample in the experiment. Furthermore, let the i^{th} exon come from gene $g(i)$ and the j^{th} sample be treated by treatment condition $t(j)$.

Define $m_{ij} = E(X_{ij})$ as the expected count for exon i in sample j , and use a log-linear model for m_{ij} . One appropriate model is

$$\log m_{ij} = \beta_{g(i)}^G + \beta_{g(i)i}^{GE} + \beta_{t(j)j}^{TS} + \beta_{g(i)j}^{GS} + \beta_{g(i)t(j)}^{GT} \quad (2.1)$$

Here G stands for gene, E for exon, T for treatment and S for sample. Exons are nested with genes, and samples within treatments. The variables $\beta_{g(i)j}^{GS}$ and $\beta_{g(i)t(j)}^{GT}$ correspond to differential expression of gene j between samples and treatments respectively. The parameters $\beta_{g(i)}^G$ and $\beta_{g(i)i}^{GE}$ correspond to the average expression of each gene and each exon within each gene whilst $\beta_{t(j)j}^{TS}$ reflects the library size or sequencing depth of each sample.

Assuming the count data follows a Poisson distribution then due to the nestedness of samples within treatments and exons within genes and by conditioning on $N = \sum_{ij} m_{ij}$, the maximum likelihood estimate of m_{ij} can be written as

$$\log \hat{m}_{ij} = \frac{\sum_{k=1}^{n_s} x_{ik} \sum_{h|g(h)=g(i)} x_{hj}}{\sum_{k=1}^{n_s} \sum_{h|g(h)=g(i)} x_{hk}},$$

where n_s is the number of samples (Bishop, 1971). As we have assumed that the count data is Poisson distributed then the data could be standardised using the Anscombe transform (Anscombe, 1948) as follows:

$$Z_{ij} = 2 \left(\sqrt{X_{ij} + \frac{3}{8}} - \sqrt{\hat{m}_{ij} + \frac{3}{8}} \right).$$

The Anscombe transform will stabilise the variances if the data is Poisson and make Z_{ij} approximately standard normal and is a slight extension on the usual square root variance stabiliser. If there is evidence that the data is not Poisson another variance

stabiliser should be used. The next step is to estimate the covariance matrix, Σ_g^E , of the exon counts within gene g . Let Z_g be a subvector of Z which contains only the exons from gene g then we can estimate Σ_g^E as

$$\hat{\Sigma}_g^E = \frac{Z_g Z_g^T}{n_s}.$$

We expect the diagonal elements of $\hat{\Sigma}_g^E$ to be close to one and the off-diagonals to be close to zero if there is no differential alternative splicing between samples.

Following a similar method described for exon arrays (Xing *et al.*, 2006) we define our method for identifying data-specific constitutive (DSC) exons as follows for each gene g separately:

1. Apply Ward's linkage hierarchical clustering (Ward, 1963) to the exons with gene g using $1 - \hat{\Sigma}_g^E$ as a distance metric.
2. Cut the clustering dendrogram, determining the cut-off height as below.
3. Evaluate all the resulting clusters using a scoring metric—again, see below.
4. Identify the cluster with the highest score. The exons in this cluster are the DSC exons for gene g .

This process is illustrated in Figure 2.4.

Deciding at what height to cut the clustering dendrogram is not trivial. As we are analysing well-annotated organisms we would like our method to perform similarly to the UI definition. To this end we choose to cut the dendrogram at a value that maximises the correlation of the exClust log fold changes with the UI log fold changes. A value of 2 maximised this correlation for the Bullard dataset following a grid search and may be a reasonable choice for poorly annotated data where a similar strategy would not be appropriate.

There are also many potential choices of scoring metric that could be used to select the subcluster of DSC exons. As DSC exons should be present in all isoforms of a gene, the DSC exons of a gene should have the highest number of reads mapping to them per base pair relative to the non DSC exons. Choosing the subcluster of exons with

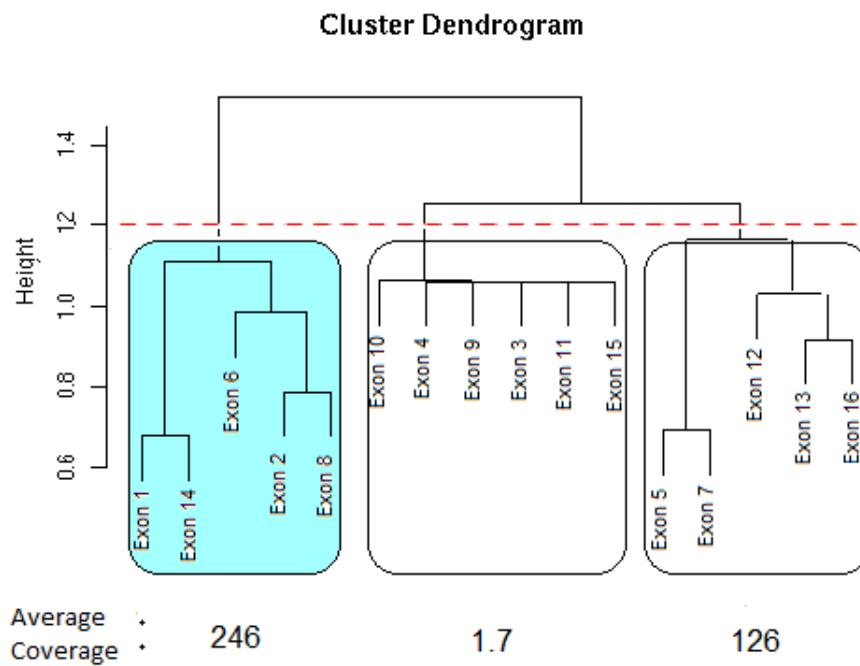


Figure 2.4: **Identifying constitutive exons** – Plot of exons selected by exClust for a particular gene. A clustering dendrogram of the exons is formed by applying Ward's linkage hierarchical clustering to the distance matrix $1 - \Sigma_g^E$. Cutting the dendrogram at the dashed red line results in the creation of three subgroups of exons (each box here contains a subgroup). For each subgroup the average coverage of the two exons in that subgroup with the highest coverage is calculated. The subgroup with the highest average coverage (the shaded subgroup) is selected to represent the DSC exons for this gene.

highest average coverage (the average number of reads mapped per base pair to each exon) may then be an appealing scoring metric. However, this scoring metric can be affected detrimentally if a subcluster has a lowly expressed exon that was included by chance. An alternative scoring procedure may then be to choose the subcluster that has the single exon with the highest coverage. However, the efficiency of the sequencing and mapping process can be influenced by artefacts such as exon length, GC content or whether the exon is an initial, internal or terminal exon (Griebel *et al.*, 2012). As a compromise between these two metrics we select the subcluster that has the largest average coverage of the two exons in each subcluster with largest coverage.

2.4 EVALUATION STUDY

In this study, we evaluate our proposed method for estimating data-specific constitutive exons, *exClust*. The performance of *exClust* will also be compared with three other summarisation approaches, Union, UI (Bullard *et al.*, 2010) and Cufflinks (Trapnell *et al.*, 2010). These approaches will be compared by evaluating their behaviour in comparison to qRT-PCR data in both a qualitative and quantitative fashion.

2.4.1 Data

We will evaluate our method for identifying constitutive exons on two publicly available datasets (M. A. Q. C. Consortium, 2006; Wang *et al.*, 2008). These were chosen as both were well studied and clearly annotated. Both datasets have a relatively high amount of replication. The MAQC data also has qRT-PCR for a selected set of genes which aids in our evaluation by providing an accurate alternate estimation of gene expression.

2.4.1.1 MAQC Data

The data consists of two mRNA-Seq datasets from the MicroArray Quality Control Project (M. A. Q. C. Consortium, 2006). In this project, Illumina's Genome Analyser II

high-throughput sequencing system was used to generate 35 bp reads from two cell line mRNA samples: Ambion’s human brain reference RNA (Brain) and Stratagene’s human universal reference RNA (UHR). Both Brain and UHR were assayed in seven lanes which we treat here as technical replicates. Fastq files were downloaded from the NCBI short read archive, submission number SRA010153. All reads were mapped to the human genome (GRCh37 assembly) using bowtie (Langmead *et al.*, 2009) ignoring all splice junction and multi-mapping reads. Using the Ensembl human exon annotation (Hubbard *et al.*, 2009), we can summarise how many reads lie within each exon of each gene for each sample. We say a read lies within an exon if its left most base pair lies within that exon. As we have ignored splice junction reads this should not introduce any bias. Processing of the data results in a matrix of counts where each row corresponds to an exon for a gene and each column corresponds to one of the 14 (7 replicates \times 2 conditions) samples. Accompanying this data set is qRT-PCR data from MAQC-1 which consists of four observations for both Brain and UHR over 1021 genes. For each gene these values were logged, averaged over the four replicate observations, and then differenced to give a single qRT-PCR log-fold-change value for each of the 1021 genes.

2.4.1.2 Wang Data

The Wang data (Wang *et al.*, 2008) consists of ten diverse human tissues and five mammary epithelial or breast cancer cell lines where 32 bp reads were obtained using Illumina’s Genome Analyser. We analyse seven samples of heart and seven samples of skeletal muscle tissue. All samples originated from the same donor and are treated as technical replicates. Fastq files were downloaded from the NCBI short read archive, submission number SRA008403. The sequenced reads were processed in the same way as the MAQC data described earlier. Processing of the data results in a matrix of counts where each row corresponds to an exon for a gene and each column corresponds to one of the 14 (7 replicates \times 2 conditions) samples.

2.4.2 Evaluation criterion and results

In this section we will primarily use the MAQC data to evaluate the effectiveness of our method for identifying constitutive exons. To do this, we will assess the concordance between the qRT-PCR data and the RNA-Seq data when summarising the RNA-Seq data using four different methods:

- **Union** the union of the exons,
- **UI** the UI definition (Bullard *et al.*, 2010),
- **Cufflinks** sum of the FPKM values of all isoforms estimated by Cufflinks for each gene (Trapnell *et al.*, 2010) and
- **exClust** the union of the exons selected by the clustering method.

Cufflinks was implemented following a standard pipeline (Trapnell *et al.*, 2012) and setting the segment length flag in Tophat to 18 for the MAQC data and 16 for the Wang data.

2.4.2.1 Number of reads included

For the three count summarisation methods, each method retains the following number of reads for the MAQC data

- 62 850 300 reads for the Union summarisation,
- 49 191 469 for the exClust summarisation, and
- 15 249 893 for the UI summarisation.

There is a successive loss of reads as each method makes increasingly stricter assumptions. The union method includes over four times as many reads as UI and exClust includes over three times as many as UI. This behaviour is also observed in the Wang data where each method includes

- 13 949 371 reads for Union summarisation,
- 10 892 133 for exClust summarisation, and

- 4 138 796 for UI summarisation.

Again, large differences between the number of reads summarised by each method are observed.

2.4.2.2 Concordance with qRT-PCR log fold change

QRT-PCR is often considered a gold standard for gene expression measurement, even though it is highly reliant on primer choice. If the primer probes for the qRT-PCR data were generally chosen in DSC regions of the genes, we expect that a better summarisation method will show higher concordance with the qRT-PCR results. In particular, as the quantification of the qRT-PCR is independent of transcript lengths, a summarisation method that removes the bias of differing transcript lengths should offer improved concordance with the qRT-PCR data. We will use two criteria to assess this concordance. Both methods rely on the detection of differential alternatively spliced (DAS) genes. A gene will be called DAS if it has a Bonferroni corrected DASI p-value less than 0.05 (Richard *et al.*, 2010) (this is described in more detail in Appendix A.1).

Of the 1021 genes in the RNA-Seq data which had matched qRT-PCR data all 1021 genes had a non-empty Union and exClust summarisation and 635 genes had a non-empty UI summarisation. The word "empty" is used to refer to the situation when a summarisation method chooses no exons in a gene to summarise over. Assuming a gene isn't completely overlapped by another gene, the Union and exClust methods always select at least one exon for each gene.

Log fold change values for each summarisation method are calculated as follows. For each gene and summarisation method, when at least one exon is deemed to be constitutive, counts are summed over the set of selected exons and over replicates to produce a total count for each of the two tissue types. The log ratio between the totals for each tissue type is then used as the log fold change estimate for each gene and method. Any gene with a log fold change of positive or negative infinity for any method is ignored. Log fold changes for each gene were estimated for Cufflinks using the difference of the log mean of the isoform FPKM values of each condition.

The first criterion for comparing the summarisation methods with the qRT-PCR data is described as follows:

CRITERION 1: Log fold change values from the given method are regressed against corresponding qRT-PCR values. Residuals for all genes against this fitted line are then computed. The top 20 DAS genes are ordered by log qRT-PCR fold change, and their residuals are plotted. An effective summarisation method should be unaffected by the length bias produced from differential alternative splicing and hence changes in residuals should be seen with the Union summarisation for these DAS genes but not the UI and exClust summarisations.

In Figure 2.5 we plot the log fold changes of the RNA-Seq data (y-axis) against the log fold changes given by qRT-PCR (x-axis). It is through these points that we fit a regression line. There is clearly a strong relationship between the log fold changes of the RNA-Seq data and those of the qRT-PCR data; this has been seen in previous analysis (Bullard *et al.*, 2010).

Figure 2.5 also provides an opportunity to examine the conceptual links between differential alternative splicing and differential expression. Highlighted are the 127 genes that DASI suggests as being differentially alternatively spliced (red) and the genes whose UI definition is empty (triangle). Of the 127 DAS genes, 42 had a non-empty UI definition. Of the genes that were identified as being differentially spliced, around one fifth of these (26 out of 127) had an absolute log fold change greater two (up or down regulated by a fold change of four). For these genes, if summarising using the Union method, these fold changes may possibly be driven by a change in the lengths of the transcripts due to splicing rather than a change in the overall transcription rate of the gene. Represented by triangles, there are a large number of genes whose UI definition is empty, with a reasonable proportion of these potentially being differentially expressed as well. Many of these have not been identified as being DAS and are potentially being excluded by the UI method unnecessarily. The omission of such a large number of genes could potentially lead to the omission of relevant biological signal.

In Figure 2.6 we focus on the residuals of the top 20 DAS genes calculated after fitting a straight line through the points in Figure 2.5. Assuming all transcripts are annotated, the UI method should always select a set of constitutive exons for a gene if that gene has exons that are conserved across all transcripts. With this in mind,

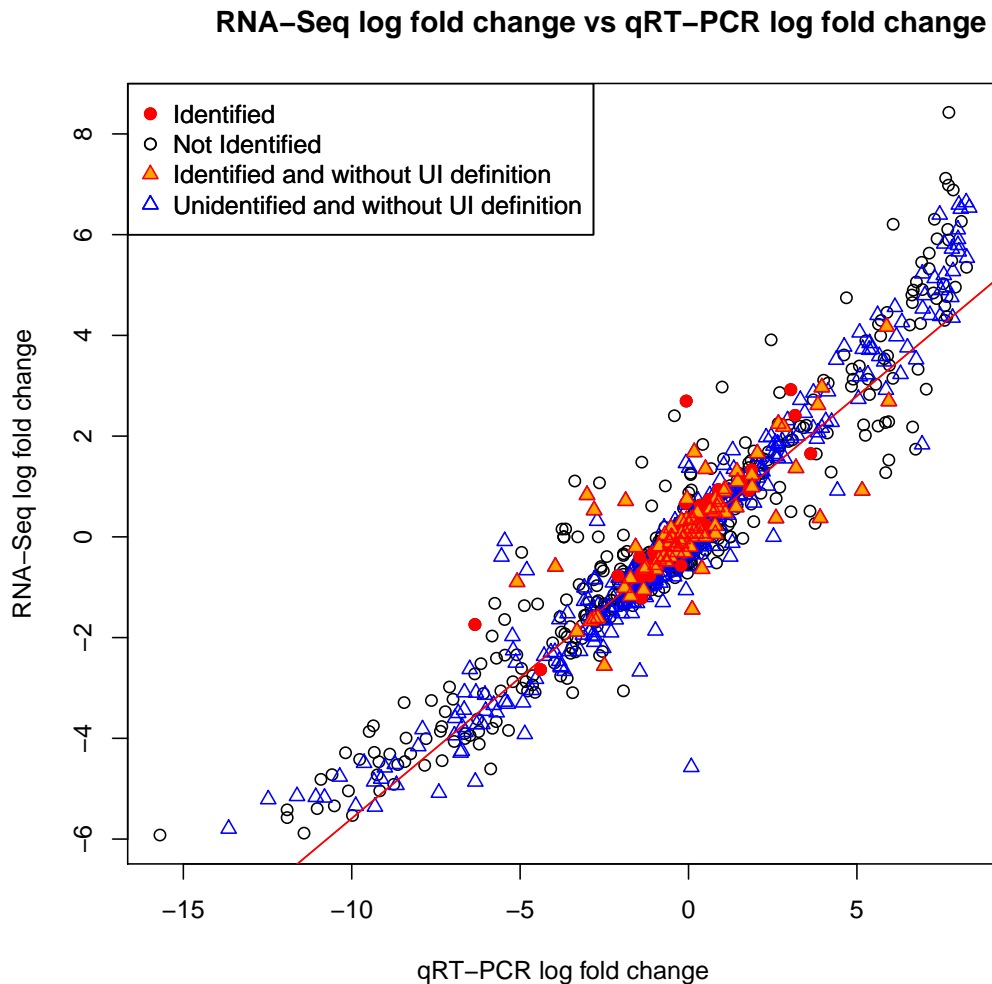


Figure 2.5: **Concordance Plot** – Concordance plot with the RNA-Seq log fold changes on the y-axis and qRT-PCR log fold changes on the x-axis. For the RNA-Seq data we use the union of all exons within a gene to summarise our counts where a value of one is added to the count of every gene. The black circles are those genes for which the UI definition is non-empty. The blue triangles are the 386 genes for which the UI definition is empty. The red dots are those genes that our method identified as having a change in isoforms and had a non-empty UI definition.

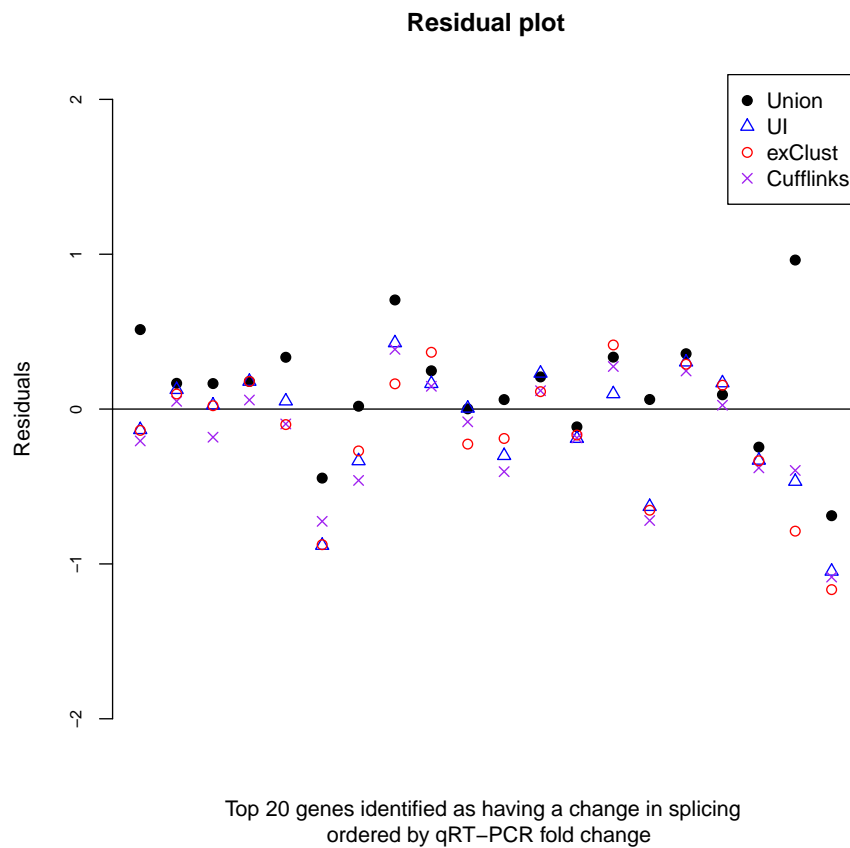


Figure 2.6: **Residual Plot** – After fitting a straight line through the plot in Figure 2.5, this figure plots on the y-axis the residuals for the genes identified as having a change in isoforms for three different annotations, union of all exons (black dots), UI definition (blue triangles) and exClust (red circles), and cufflinks (purple cross) ordered by qRT-PCR fold change.

whenever there is a large change in residuals of the UI summarisation compared to the Union summarisation, this change is also observed with exClust and Cufflinks. Due to this similarity in behaviour, both exClust and Cufflinks appear to be selecting a similar set of exons as those selected by UI for these genes. These 20 genes demonstrate the impact of summarising using the UI or exClust summarisations as opposed to simply using the Union.

Figure A.2 and Table A.1 in Appendix A provide examples of genes for which the UI summarisation appears to not be selecting DSC exons. While neither of these genes provide conclusive evidence against the UI summarisation, the log fold changes of the exClust summarisation are closer to both the qRT-PCR and Cufflinks log fold changes than the log fold changes of UI are.

As quantification by qRT-PCR is independent of transcript lengths, a summarisation method that removes the bias due to differing transcript lengths should offer improved correlation with the qRT-PCR data. Correlation will be used as a quantitative criterion for comparing the summarisation methods with the qRT-PCR data and is calculated as follows:

CRITERION 2: In this second criterion we compute the Pearson correlations between log fold change values from the Union, UI and exClust summarisations, the sum of the isoform FPKM of Cufflinks and the qRT-PCR value. This is done separately for

- the DAS genes, and
- the non-DAS genes,

where only genes with a non-empty UI definition are used. An effective method will produce a high Pearson correlation score in all cases.

A numerical summary of the correlations from the second criterion are presented in Table 2.1. As we would expect, correlations with qRT-PCR are higher for non-differentially alternatively spliced (non-DAS) genes than for differentially alternatively spliced (DAS) genes for all methods. For the DAS genes the Union summarisation appears to be affected adversely by the change in transcript lengths in comparison to the UI, Cufflinks and exClust summarisations. When there are differential alternat-

DAS	qRT-PCR	Union	UI	exClust	Cufflinks
qRT-PCR	1.0000	0.8292	0.8462	0.8651	0.8578
Union		1.0000	0.9373	0.9208	0.9322
UI			1.0000	0.9868	0.9764
exClust				1.0000	0.9777
Cufflinks					1.0000
non-DAS	qRT-PCR	Union	UI	exClust	Cufflinks
qRT-PCR	1.0000	0.9435	0.9416	0.9442	0.9360
Union		1.0000	0.9917	0.9995	0.9868
UI			1.0000	0.9917	0.9806
exClust				1.0000	0.9869
Cufflinks					1.0000

Table 2.1: **MAQC correlations** – A table showing two subsets of genes from the MAQC data: differentially alternatively spliced genes (DAS) and not differentially alternatively spliced genes (non-DAS). For each set of genes the correlations between Union, UI, exClust and Cufflinks log fold changes are given. The given correlations are only calculated on the subset of genes for which the UI definition is non-empty and have finite log fold change.

ive splicing events, exClust performs in a similar way to UI though in the absence of these events, exClust is similar to the Union summarisation. This makes the performance of the exClust summarisation more robust, performing well on all tested sets of genes. Cufflinks performs worst when compared to qRT-PCR for the non-DAS genes. While this is probably hindered by our unconventional implementation of Cufflinks, this lack of performance is driven mostly by genes with low counts in one condition. This puts Cufflinks at a disadvantage on two fronts; (i) estimation of transcripts is difficult in these situations of low expression; and (ii) due to the low expression the log fold changes for the isoforms of these genes are unstable and hence the aggregation of these isoforms is also unstable.

Similar outcomes were observed in the Wang dataset. However, for the Wang dataset qRT-PCR data is not available. For this data set, correlations were only calculated between the three summarisation methods and Cufflinks and can be found in Table 2.2. For the differentially alternatively spliced genes the correlation between exClust and the Union summarisation decreases to 0.968 from 0.999 for the non-DAS genes. This suggests that the Union summarisation is affected by differing transcript lengths. The correlation between the Union and UI summarisations is 0.952 for the non-DAS genes which suggests that either there are still a large number of DAS gene in this set which

DAS	Union	UI	exClust	Cufflinks
Union	1.0000	0.9488	0.9684	0.9884
UI		1.0000	0.9252	0.9488
exClust			1.0000	0.9675
Cufflinks				1.0000
non-DAS	Union	UI	exClust	Cufflinks
Union	1.0000	0.9522	0.9992	0.8990
UI		1.0000	0.9520	0.8753
exClust			1.0000	0.8986
Cufflinks				1.0000

Table 2.2: **Wang correlations** – A table showing two subsets of genes from the Wang data: differentially alternatively spliced genes (DAS) and not differentially alternatively spliced genes (non-DAS). For each set of genes the correlations between Union, UI, exClust and Cufflinks log fold changes are given. The given correlations are only calculated on the subset of genes for which the UI definition is non-empty and have finite log fold change.

were not detected or that the log fold changes of the UI summarisation have become less stable due to the large reduction in included reads. Cufflinks is less concordant with the Union summarisation in the set of non-DAS genes (0.8990) than the DAS genes (0.9884). Again, while this has probably not been helped by our unconventional implementation of Cufflinks, this lack of concordance in the non-DAS genes appears to be driven mostly by genes with low counts in one condition.

2.5 CONCLUSIONS AND FURTHER DISCUSSION

We have developed a method to improve the preprocessing of RNA-Seq data, specifically the summarisation of reads into gene counts. When working at a gene level, between-treatment differential alternative splicing can cause problems with an expression analysis. The concept of constitutive exons helps to resolve these problems by finding exons which are common to all isoforms of a gene. Our novel approach, exClust, estimates the constitutive exons in a gene using both empirical and annotated data. Importantly, we allow constitutive exons to be data-specific. That is, we defined data-specific constitutive exons as exons which are common to all the isoforms of a gene which are present *in the current experimental samples*. This new approach allows for more accurate quantification of gene expression.

In the datasets examined, a more complex modelling of the variances than what is assumed by the Poisson is not required. This is shown in Figure A.1. While the technical variability between samples should be Poisson, most experiments have an element of biological variability as well and hence RNA-Seq data is often modelled as an over-dispersed Poisson. A more sophisticated methodology would model this over-dispersion and standardise accordingly (Robinson *et al.*, 2010; Anders and Huber, 2010). However, as our model does fit an interaction term between gene and sample effects, a large amount of the biological variability typically observed in a differential expression analysis may be accounted for.

Our approach, exClust, for empirically estimating the data-specific constitutive exons within a gene can be seen to perform favourably when compared with the current alternatives. In summary, exClust had higher correlation with qRT-PCR data compared to other approaches. This favourable performance comes without a dramatic decrease in total read count, with exClust including three times as many reads as UI.

DIFFERENTIAL EXPRESSION – IMPROVED MODERATION FOR
GENE-WISE VARIANCE ESTIMATION IN RNA-SEQ VIA THE
EXPLOITATION OF EXTERNAL INFORMATION

The identification of differential expressed (DE) genes and transcripts is still a key question of interest in many biological studies. However, even with the lowering cost of sequencing data, the majority of RNA-Seq experiments are still suffering from low replication numbers which makes confidently calling DE genes difficult. To date, there are many methods that provide a test of whether a gene is DE or not (Pachter, 2011), including cufflinks (Trapnell *et al.*, 2010), DESeq (Anders and Huber, 2010) and edgeR (Robinson *et al.*, 2010). A feature in all of these methods is moderation of gene-wise variance estimates to improve DE inference. Moderation is important in small samples size comparisons, increasing both the power and accuracy of a DE test (Smyth, 2004). The key differences between these methods are the extent of the moderation and their *common variance* estimate—the variance estimate that the procedure is moderating towards.

DESeq and edgeR account for the heteroscedasticity observed in the read counts of genes by modelling the relationship between expected value of the count and its variability. In this chapter we will propose using additional information, such as gene length and variance estimates from external datasets, as explanatory variables to further model the heterogeneity seen in the observed gene variances. Combining these improved models of gene variance with a moderation method (Opgen-Rhein and Strimmer, 2007) creates a robust tool for estimating gene variances and hence calling differential gene expression. When evaluated on publicly available data this tool offers both improved gene ranking and power of detection when compared to DESeq and edgeR.

3.1 MODELLING RNA-SEQ DATA

In the following we will describe how RNA-Seq data has generally been characterised and modelled. This will mainly focus on the models used to describe the strong relationship observed between the average expression of genes in RNA-Seq data and the variances of this expression. Figure 3.1 illustrates this mean-variance relationship by comparing the gene means and variances from a dataset described in Section 3.3 of this chapter.

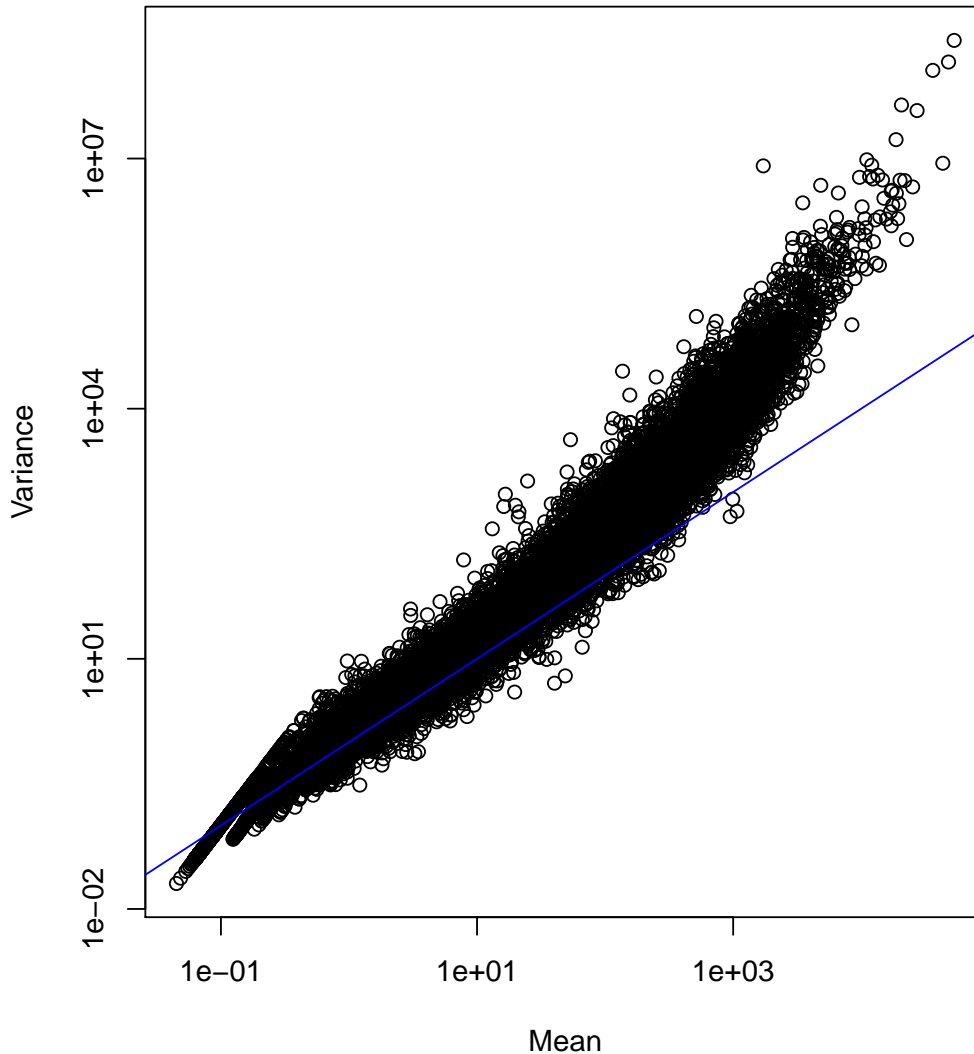


Figure 3.1: **The mean-variance relationship in RNA-Seq data on a log-log scale** – This figure demonstrates the strong mean-variance relationship observed in RNA-Seq data. For each gene from the ten Bottomly B6 samples described in Section 3.3 the mean and variance are calculated. These means and variances are then plotted against each other on a log-log scale. The blue line corresponds to the line $y=x$, when the mean equals the variance.

Following from the previous chapter we will assume that the reads from an RNA-Seq experiment can be summarised into a matrix of gene counts. Let y_{ij} be the observed read count for the i^{th} gene in the j^{th} sample where sample j belongs to treatment $t(j) = 1, 2$. For ease of presentation we will assume that all effects that are generally normalised for or modelled, such as library sizes and GC content, remain constant

across samples. Let σ_i^2 and μ_i be the variance and mean read count for gene i . The technical variability (the variability introduced by the sampling process of the sequencer) for a gene count in RNA-Seq can be modelled quite reliably as Poisson (Bullard *et al.*, 2010; Marioni *et al.*, 2008). This is attractive in situations of low replication as one parameter can be estimated to describe both the mean and variance of a gene. Modelling the data as Poisson will give a very reliable estimate of which genes have changed in expression between any two samples. However many experiments are not simply focused on the detection of gene expression differences between any two samples focusing instead on the differences between any two types of cells for example. This distinction is important as it requires us to not only model the technical variability of the experiment but to also model the biological variability of a particular cell type (or experimental condition).

An over-dispersed Poisson, a discrete distribution with dispersion greater than a Poisson, can be modelled using a Negative Binomial. A negative binomial random variable, Y , can be parametrised with probability mass function

$$P(Y = y) = \binom{r + y - 1}{y} p^r (1 - p)^y. \quad (3.1)$$

This standard formulation is generally referred to as NB2. Under this formulation, the biological variability of the expression of a gene is modelled as a quadratic function of its mean expression μ :

$$\sigma^2 = \mu + b\mu^2, \quad (3.2)$$

where as $b = \frac{1}{r}$ gets small the negative binomial will approach a Poisson. The parameter b has been referred to as the coefficient of biological variation. A negative binomial is generally parametrised as a function of r and p . However, by parametrising a negative binomial in terms of its mean μ and variance σ^2 where

$$r = \frac{\mu^2}{\sigma^2 - \mu} \quad (3.3)$$

$$p = \frac{\mu}{\sigma^2} \quad (3.4)$$

and $\sigma^2 > \mu$, a negative binomial can then be used to model counts that have untraditional mean-variance relationships. This relationship is generally expressed as

$$\sigma^2 = \mu + f(\mu) \quad (3.5)$$

where $f(\mu)$ explains the biological variability can be fitted by some form of nonparametric regression (Anders and Huber, 2010). This formulation highlights that σ^2 should always be greater than or equal to μ .

In current RNA-Seq experiments it is still quite common to see experiments with very little biological replication. Estimating variances from a small number of observations is typically unstable (Cui and Churchill, 2003). To improve the stability and accuracy of these variance estimates there have been many methods proposed to shrink the variances to some common value for microarrays (Cui and Churchill, 2003) and RNA-Seq (Pachter, 2011). We will refer to this as moderation. By stabilising the variances and sharing information moderation also increase the power of a statistic as this increases the degrees of freedom of a variance estimate (Smyth, 2004).

3.1.0.3 *Heterogeneous gene variances*

It is well accepted that some genes have a higher variance than other genes (Cheung *et al.*, 2003). That is, some genes vary in expression more from cell to cell, person to person, or treatment to treatment in comparison to other genes. In RNA-Seq datasets, genes with larger average expression have on average larger observed variances (see Figure 3.1). Instead of shrinking the estimate of a genes variance towards some common value (as is often done in microarrays) to improve stability (Cui and Churchill, 2003), edgeR and DESeq shrink the estimate towards some fitted curve describing the relationship between mean and variance. We refer to this fitted curve as the *common variance*. In doing this they are making the strong assumption (although not an unreasonable one) that genes with a similar average count should have a similar variance.

3.2 TSHRINK+

We propose using local regression (Loader, 2010) to fit a smoothed surface through any number of variables ($\gamma_{(1)}, \gamma_{(2)} \dots$) that may help to explain the observed pooled sample variances $\hat{\sigma}_{\text{gene}}^2 = s^2$. We estimate the common variances σ_{common}^2 as

$$\hat{\sigma}_{\text{common}}^2 = \mu + f(\mu, \gamma_{(1)}, \gamma_{(2)}, \dots). \quad (3.6)$$

When using variance estimates from other RNA-Seq experiments, these variances will also have a very strong mean-variance relationship. For use as an explanatory variable we normalise the external variance estimates in such a way that they have mean zero and variance one for all ranges of expression.

To illustrate how this improved common variance can aid in moderation we propose using a quasi-empirical Bayes moderation method (Opgen-Rhein and Strimmer, 2007), where the variance is moderated as

$$\hat{\sigma}_{\text{shrink}}^2 = \lambda \hat{\sigma}_{\text{common}}^2 + (1 - \lambda) \hat{\sigma}_{\text{gene}}^2, \quad (3.7)$$

and $\hat{\sigma}_{\text{shrink}}^2$, $\hat{\sigma}_{\text{common}}^2$ and $\hat{\sigma}_{\text{gene}}^2$ are the moderated, common and sample variances. Without making distributional assumptions, the shrinkage parameter λ can be estimated by the equation

$$\lambda = \min \left(1, \frac{\sum_{k=1} \text{Var}(\sigma_{k(\text{gene})}^2 / \sigma_{k(\text{common})}^2)}{\sum_{k=1} (\hat{\sigma}_{k(\text{gene})}^2 / \hat{\sigma}_{k(\text{common})}^2 - 1)^2} \right). \quad (3.8)$$

The parameter λ is the ratio of the expected and average squared error of the common variance estimate. Due to the large amount of smoothing that is used in estimating the common variance, we will make the assumption that the data, standardised using the common variance estimate, is approximately standard normal. This normality assumption will only be appropriate for genes with larger average expression. Under

this assumption the variance of $\sigma_{k(\text{gene})}^2/\sigma_{k(\text{common})}^2$ (which has $n - 2$ degrees of freedom) is $2/(n - 2)$. Our estimate of λ then becomes

$$\lambda = \min \left(1, \frac{2n}{(n-2) \sum_{k=1} (\hat{\sigma}_{k(\text{gene})}^2/\hat{\sigma}_{k(\text{common})}^2 - 1)^2} \right). \quad (3.9)$$

A Wald test for each gene is then performed using the statistic

$$\frac{\bar{y}_{t_1} - \bar{y}_{t_2}}{\sqrt{\sigma_{\text{shrink}}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}, \quad (3.10)$$

where we utilise the Welch-Satterthwaite equation (Satterthwaite, 1946; Welch, 1947) to estimate its degrees of freedom $\hat{\nu}$. We have assumed earlier that the degrees of freedom corresponding to common variance is $\nu_{\text{common}} = \infty$ and can thus estimate ν_k as

$$\hat{\nu}_k = \frac{(\lambda \hat{\sigma}_{k(\text{common})}^2 + (1 - \lambda) \hat{\sigma}_{k(\text{gene})}^2)^2}{\frac{(1 - \lambda)^2}{\nu_{\text{gene}}} \hat{\sigma}_{k(\text{gene})}^4} \quad (3.11)$$

where $\nu_{\text{gene}} = n - 2$. For simplicity, rather than using a different ν for each gene we instead use one degrees of freedom estimate, ν_{shrink} , for all genes, taken as the mean of the $\hat{\nu}_k$'s.

3.3 EVALUATION STUDY

In this study we evaluate Tshrink+, our proposed method for improving variance estimation for differential gene expression analysis by using additional external information. This evaluation consists of two components, assessing the capacity of a common variance estimate to explain the observed gene sample variances, and evaluating how improving this common variance estimate can aid in the detection of differentially expressed genes. The performance of Tshrink+ will also be compared with two commonly used packages, edgeR and DESeq. This evaluation study is built upon one main dataset, the Bottomly data, and three datasets which are used for additional information.

3.3.1 Data

3.3.1.1 Bottomly Dataset

The Bottomly data (Bottomly *et al.*, 2011) was used as the main analysis dataset for evaluation and was chosen because of its relatively large number of biological replicates. The pre-processed RNA-Seq data comparing ten C57BL/6J (B6) and eleven DBA/2J (D2) mouse striatum was downloaded from the ReCount project (Frazee *et al.*, 2011) as a matrix of counts. For simplicity only the first ten DBA/2J samples were used. All data used in the analysis are normalised counts as DESeq and edgeR do not accept gene-wise normalisation factors. To model the disparate library sizes and biases of PCR amplification observed in the data, a cyclic robust linear model was used. Using the first sample in the dataset as a reference, M values were calculated for each gene in the remaining samples and a straight line was fitted through the M-values using GC-content as an explanatory variable. The M-values were then normalised to this line such that the average M-value was zero over the range of GC-content. After this normalisation there were still batch and other sample specific effects evident in the data. These were normalised out using a cyclic loess (Cleveland *et al.*, 1992) strategy as described in Appendix B. This normalisation appeared to be more suitable than RUV (Gagnon-Bartsch and Speed, 2012) and SVA (Leek *et al.*, 2012) improving concordance with microarray results as seen in Figure B.1. The data is filtered to only include genes with average expression greater than twenty.

3.3.1.2 External datasets

Sample variances from three datasets were used as sources of additional information to aid in the estimation of the common variance. These are described in Table 3.1. All RNA-Seq data were mapped to the mm9 mouse genome using bowtie (Langmead *et al.*, 2009) and normalised for GC content bias and library size differences as the Bottomly dataset was. The microarray data were read and processed using the R packages Affy (Gautier *et al.*, 2004) and gcrma (Wu *et al.*, 2013).

Species	Tissue	Replicates	Platform	Source	GEO accession
C57BL/6J mouse	Liver	6	RNA-Seq	Keane <i>et al.</i> (2011)	GSE30617
	Spleen	6			
	Thymus	6			
	Lung	6			
	Heart	6			
	Hippocampus	6			
C57BL/6J mouse	Striatum	4	RNA-Seq	Polymenidou <i>et al.</i> (2011)	GSE27218
C57BL/6J mouse	Striatum	10	microarray	Bottomly <i>et al.</i> (2011)	GSE26024

Table 3.1: **Additional information sources** – Variance estimates from these three datasets are be used to improve the estimation of the common variance function in the main analysis dataset.

3.3.2 Evaluation strategies and results

3.3.2.1 The estimation of the common variance

We begin by examining the effect of using information from different additional sources to help explain the variances observed in the Bottomly Data. This is explained in Strategy 1 below. We aim to assess the effectiveness of using information, in addition to the average expression of a gene, to estimate a common variance function. In order to assess the capacity of a common variance estimate to explain the observed gene sample variances we will use the shrinkage coefficient λ , which is described in Equation 3.9, as a statistic. The parameter λ is the ratio of the expected and average squared error of the common variance estimate. As λ is proportional to the reciprocal of the average squared error of the variance estimates, a larger λ corresponds to a better estimate of the common variance. A λ value of one implies that the common variance is representative of the population variance. A λ of zero suggests that the common variance estimate is failing to describe the observed gene sample variances.

STRATEGY 1 Variance estimates from the external datasets described in Table 3.1 and also gene length are used to aid in the estimation of the common variance functions of one hundred random comparisons of n samples of B6 mouse striatum tissue with n samples of D2 mouse striatum tissue. This is performed for one additional dataset at a time. By only consider one additional dataset at a time we

do not consider how they might interact. The average λ value is calculated for each n comparison and information source using only the genes that are present in all data sources.

The results from Strategy 1 can be found in Figure 3.2. When used to help fit the common variance surface, using information from any of the additional data sources improve the estimate of the common variance. This is observed through all of the average λ 's being higher when using additional information when compared to using only the mean. Where again, a larger λ should correspond to a better estimate of the common variance.

The more relevant the information contained in the additional data source, the greater the improvement seen in the common variance estimate. As is perhaps expected, either of the two striatum tissue datasets (RNA-Seq and microarray) when used to estimate the common variance produce the largest λ , with microarray striatum and RNA-Seq striatum only slightly out performing hippocampus. Spleen and lung both increase λ highlighting that information can still be gained from unrelated tissue types, however, liver and heart barely increase λ at all. This can mostly be explained by the use of liver and heart resulting in the variance of one gene, transthyretin, being severely under-estimated. If this gene is excluded the λ generated by using liver and heart are much similar to that of spleen and lung. Including information on gene length also has the potential to improve variance information. Although, this appears to relatively decrease as the sample size n increases.

Improving the accuracy of the sample variance decreases λ and improving the accuracy of the common variance increases λ . As the sample size n increases, λ decreases. This is because as n increases the accuracy of the gene sample variance estimates increase. As the estimation of the gene sample variances improves, the inability of the common variance to describe the gene variances becomes more clear.

We can then further demonstrate that improving the information content of an additional information source improves the estimation of the common variance. This will be achieved by using variance estimates from the D2 mice to aid in the estimation of a common variance function of the B6 mice.

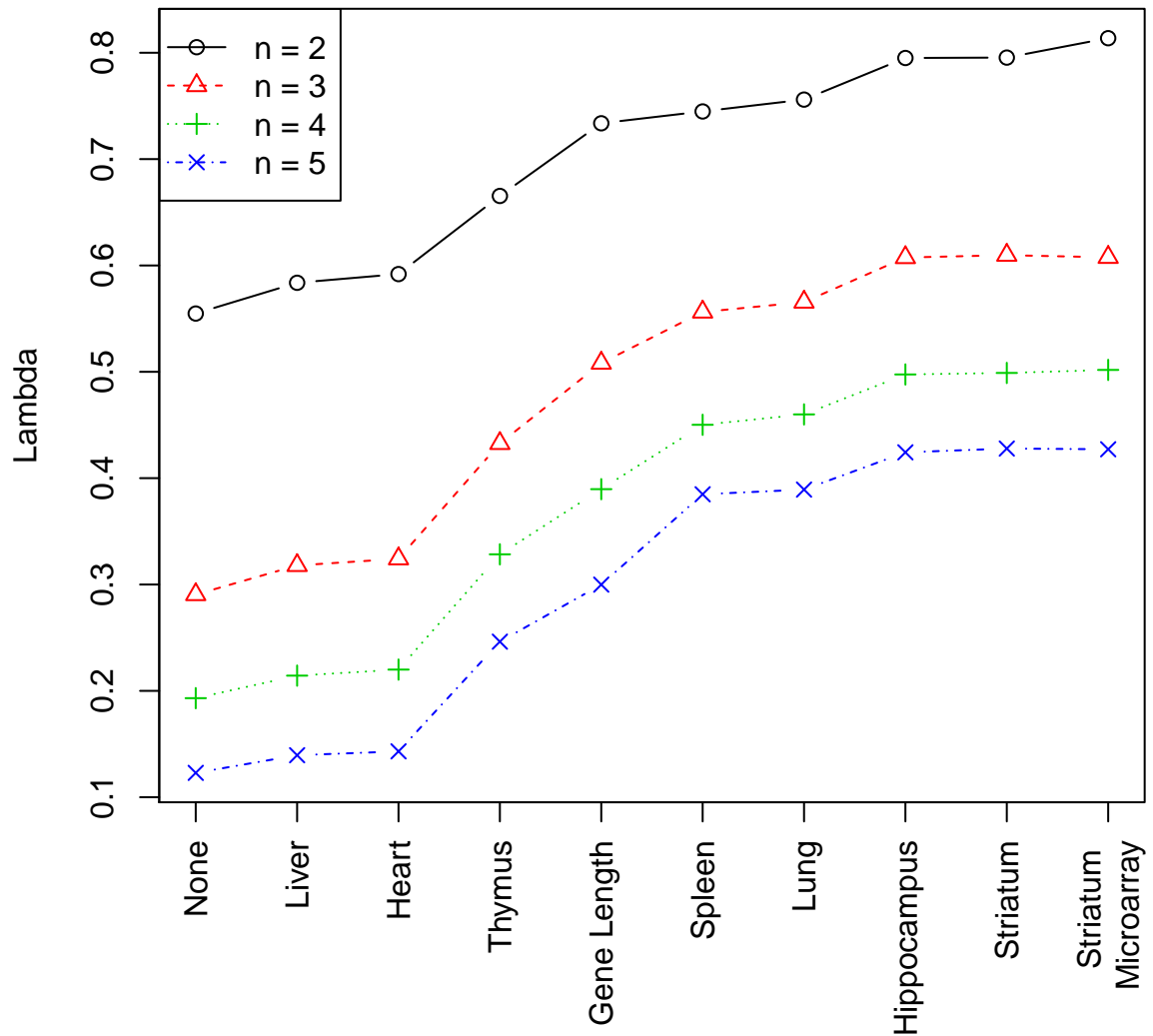


Figure 3.2: **Effect of utilising different sources of information on the estimation of λ** – Variance estimates from the external datasets (Table 3.1) and gene length are used to aid in the estimation of the common variance functions of one hundred comparisons of n B6 and n D2 mouse striatum samples. The average λ value is plotted for each n comparison and information source for n ranging from two to five. The parameter λ is the ratio of the expected and average squared error of the gene sample variance to the common variance. The information source “None” corresponds to using no extra information, “Striatum” the RNA-Seq samples from Polymenidou et al (2011) and “Striatum Microarray” the microarray striatum samples from Bottomly et al (2011). The information sources have been sorted by their λ values for n equals two.

n D2 samples	0	2	3	4	5	6	7	8	9	10
λ	0.35	0.45	0.50	0.55	0.58	0.65	0.68	0.72	0.75	0.77

Table 3.2: **Using D2 variance estimates to estimate common variance of four B6 samples** – The average λ values calculated using a random n D2 mouse striatum samples to estimate the variance of a random four B6 mouse striatum samples from one hundred simulations.

STRATEGY 2 The variance estimates from a random n D2 mouse samples are used to estimate the common variance function of a random four B6 mouse, this is repeated one hundred times and average λ values are calculated.

The results of this can be seen in Table 3.2. As the information content of the additional data source improves, i.e. the variance estimates from D2 mice calculated with increasing sample sizes, the ability of the common variance to describe the observed gene variances, calculated from four replicates of B6 mice, also improves. The estimated value of λ is doubled by using ten replicates of D2 mouse as opposed to nothing, that is, the average squared error of the common variance is halved.

3.3.2.2 *The impact of moderation on inferring differential expression.*

The aim of the remainder of the evaluation is to assess the influence of using additional information and moderation on the detection of differentially expressed (DE) genes. To do this we compare

1. a t-test (T),
2. a moderated t-test (Tshrink) and
3. a moderated t-test using additional information (Tshrink+).

These will also be compared to

4. DESeq using only the common variance (DESeqCommon),
5. DESeq using the maximum of the common variance and sample variance (DESeqMax) and
6. edgeR using a trended common variance and empirical Bayes to shrink the gene sample variances towards the common variance (edgeR).

For the additional data source used by Tshrink+, the four striatum RNA-Seq samples (Keane *et al.*, 2011) in Table 3.1 were chosen as they gave the second highest λ value but were not generated from the same lab as the analysis dataset (as the microarray data were). The most unusual comparison here is the t-test (T). A standard two-sample t-test is not usually used in RNA-Seq due to small sample sizes (n), which can benefit from moderation approaches, and a normality approximation being inappropriate for genes with small average expression (μ). As we are only considering genes with μ greater than twenty, this comparison will hopefully demonstrate the benefits of moderation in the presence of small sample sizes.

To assess the effectiveness of the six DE methods the following strategy was used:

STRATEGY 3 A standard t-test was performed comparing ten B6 and ten D2 mouse striatum samples. In all of the following, the results of this t-test are taken to be the "truth". From this t-test a gene is conservatively called "truly" DE if it has a Bonferroni adjusted p-value of less than 0.05. A gene is called "truly" not DE if it has an unadjusted p-value greater than 0.05. We will then evaluate the ability of the DE methods to recover the information in the comparison of ten B6 samples with ten D2 samples by smaller comparisons of n B6 samples and n D2 samples, for n ranging from two to five. This is done by comparing a random set of n B6 and n D2 mouse striatum samples one hundred times and then

- generating Receiver Operating Characteristic curves (ROC curve, a curve describing each methods True Positive Rate as a function of its False Positive Rate for a complete range of p-value cut-offs),
- calculating partial areas under the ROC for FPR less than 0.01 and
- calculating True Positives (TP) and False Positives (FP) using a Bonferroni adjusted p-value cut-off of 0.05.

We can assess how the use of moderation affects inference on differential gene expression. This is done by assessing the impact of moderation on both gene ranking and sensitivity. Moderation is used to both increase the sensitivity of a test, by increasing the degrees of freedom of the variance estimate, and to improve the ranking of a test,

by improving the accuracy of the variance estimate. We will start by simply comparing the t-test (T), moderated t-test (Tshrink) and a moderated t-test using additional information (Tshrink+).

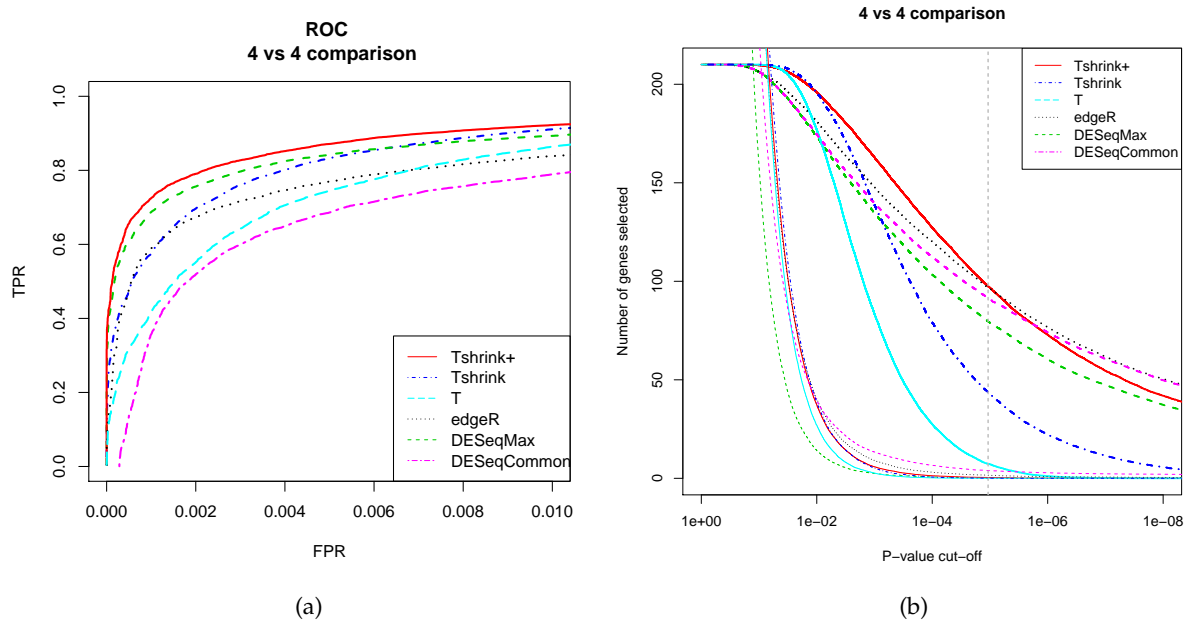


Figure 3.3: **Comparing six DE methods on a 4 vs 4 comparison** – One hundred random comparisons of four B6 and four D2 mouse striatum samples for six DE methods. Average TP and FP are calculated for the full range of p-value cut-offs. The TPR and FPR are plotted against each other in a) to form ROC curves and displayed in the region for FPR less than 0.01 as this is most relevant for calling DE. For any given FPR a method with a larger TPR is deemed to have ranked the genes better. In b) the number of TP (in bold) and FP are plotted for a range of p-value cut-offs. The x-axis is in log-scale. The grey dashed vertical line corresponds to a Bonferroni adjusted cut-off of 0.05.

By first considering only four vs four comparisons, the ability of moderation to improve gene ranking is illustrated in Figure 3.3a where a partial average ROC curve from one hundred four vs four comparisons of B6 and D2 mouse striatum is plotted for each method. This curve shows each methods TPR for a range of FPR, where a method is deemed to have ranked genes better than another at a given FPR if its TPR is higher. Here we see that Tshrink (dark blue) performs better than T (light blue) for all FPR less than 0.01. Tshrink+ (red) offers a similar improvement again on top of that of Tshrink nearly doubling the improvement of Tshrink to T.

Moderation improves gene ranking. Furthermore, improving what a method moderates towards can improve gene ranking further. This is again illustrated in Figure 3.4, where the partial area under the ROC curve is plotted for a range of n vs n comparisons. A value of 1 corresponds to a perfect ranking and a value of zero corresponds to the most imperfect ranking. For all n considered Tshrink+ appears to double the improvement of Tshrink when compared to T. The relative improvements decrease as n increases as the information in the sample variance increases in comparison to the common variance.

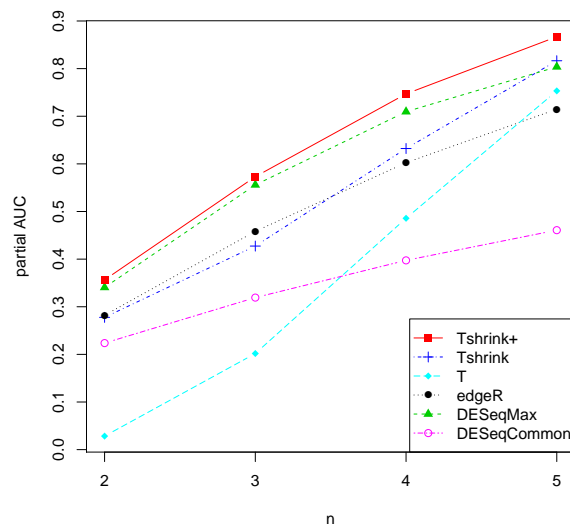


Figure 3.4: **Partial AUCs for a range of n vs n comparisons** – One hundred random comparisons of n B6 and n D2 mouse striatum samples a performed for six DE methods for n ranging from two to five. For each method and n , partial areas under the ROC curves (partial AUC) are calculated for the regions of FPR less than 0.01

Moderation can also improve the sensitivity of a test for differential expression as seen in Figure 3.3b. Figure 3.3b plots the average number of True Positive genes called at varying p-value cut-offs for one hundred four vs four tests. At a Bonferroni adjusted p-value cut-off of 0.05 (the grey dashed line) T calls 8 TP, Tshrink 47 TP and Tshrink+ 108 TP. These improvements are seen at very little cost the the number of False Positives called.

The number of TP and FP called at a Bonferroni adjusted p-value cut-off of 0.05 for n ranging from two to five are plotted in Figures 3.5a and 3.5b respectively. Here

we see the number of TP called for Tshrink+ increases as n increases and the number FP decreasing as n increases. While the number of TP called also increase for T, it decreases for Tshrink over this range of n . The number of TP called by Tshrink will decrease until Tshrink converges to T when it will continue to increase. Tshrink may be over-zealous in its calling of TP calling a relatively large amount of FP as well for small n .

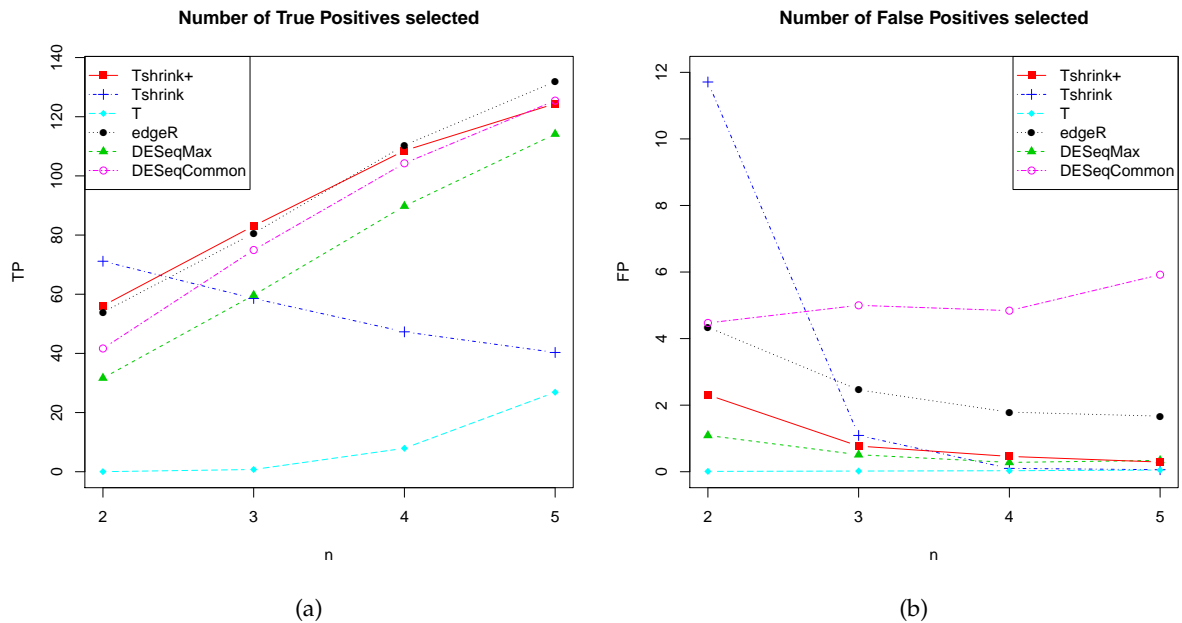


Figure 3.5: **The number of True and False Positives for a range of n vs n comparisons** – One hundred random comparisons of n B6 and n D2 mouse striatum samples a performed for six DE methods for n ranging from two to five. For each method and n , the conservative Bonferroni adjusted cut-off of 0.05 is used to calculate the average number of (a) True Positives and (b) False Positive are counted.

3.3.2.3 Comparison with edgeR and DESeq

Tshrink+ performs favourably when compared to both DESeq and edgeR when considering gene ranking. When assessing gene ranking using Figure 3.4, Tshrink+ performs marginally better than DESeqMax (green) which is better than edgeR (black) and DESeqCommon (pink). The relative performance of Tshrink+ over DESeqMax increases as n increases. For n equal to five edgeR performs worse than T. It could be argued that this is because T is becoming closer to the t-test that was used as “truth”.

However, this behaviour is also observed when using the results from microarray array data (Bottomly *et al.*, 2011) as “truth” as seen in Figure B.3 in Appendix B. This performance could also be explained by edgeR over moderating to a common variance that is become decreasingly relevant as n increases.

Tshrink+ compares comparably to edgeR and DESeq when assessing sensitivity. T selects a similar number of TP at the cut-off when compared edgeR but selects less FP as seen in Figures 3.5a and 3.5b. While DESeqMax does not select as many TP for the given cut-off as DESeqCommon it selects dramatically less FP.

3.4 CONCLUSIONS AND FURTHER DISCUSSION

Using additional information improves the estimation of the common variance and the detection of differentially expressed genes. Our differential expression test, Tshrink+ which incorporates information from additional datasets, showed marked improvement in both gene ranking and sensitivity over a moderated t-test, Tshrink, and a standard t-test. Tshrink+ also performed favourably against edgeR and DESeq when comparing gene ranking and comparably when assessing sensitivity.

Whilst Tshrink+ can offer improvements to a differential expression analysis it also provides insight into avenues for further research. The moderation used in Tshrink+ (Opgen-Rhein and Strimmer, 2007) can be drastically affected by genes with unusual variances. A more sophisticated methodology which manages the influence of these genes on moderation could offer potentially large improvements. While using local regression to fit the common variance when incorporating one additional dataset is easy to implement, it does not scale well to the use of multiple information sources. A parametric based approach may make the integration of multiple data sources feasible.

Using external data to improve the estimation of the common variance for a particular problem highlights the significance of access to public data repositories like the gene expression omnibus (GEO) (Barrett *et al.*, 2011). These repositories have the ability to actualise improved inference lending both confidence and power to results. Projects like ReCount (Frazee *et al.*, 2011) aid in this process by providing access to pre-

processed data that avoids the duplication of the computationally intensive procedure of both downloading and processing large datasets.

This methodology should be considered as a complement, not a replacement, for meta-analysis when similar studies to the RNA-Seq study of interest exist. Tshrink+ leverages only the variance estimates from external datasets to improve the variance estimation in the study of interest. If information exists on the changes of expression between conditions as well, a researcher may be remiss to not utilise this information through the use of existing meta-analysis methodologies. However, when similar studies are not available, Tshrink+ provides a unique way of utilising information from any related studies that are available.

FUNCTIONAL, NETWORK & PATHWAY ANALYSIS – USING
PATHWAY INFORMATION TO HELP INTEGRATE SMALL
SAMPLE MIRNA-SEQ AND MRNA-SEQ DATA

Functional, network or pathway analysis is the process of leveraging external annotation of functional gene groups, networks or pathways with statistical techniques to highlight biologically significant signal in a dataset. It is often performed after a differential expression analysis to improve the interpretability of the results. While simultaneously analysing 20 000 genes does provide opportunities for sharing information via moderation, it can also create problems. Performing differential expression tests on a large number of genes creates problems with multiple testing and hence the ability to detect statistically significant differences in expression. Conversely, if a large number of genes are identified as statistically different, interpretation of results can quickly become difficult. Analysing genes in terms of pathways, networks or other annotated measures of function can alleviate both of these issues in addition to creating opportunities to investigate data in alternative ways.

In this chapter we propose using pathway information to create a supervised framework, pMimCor, for integrating two forms of RNA-Seq data, micro-RNA (miRNA) sequencing and messenger-RNA (mRNA) sequencing. We demonstrate that the KEGG pathway annotation is enriched for miRNA-mRNA relationships and that these relationships appear to be present in our motivational dataset. We also show that even when ignoring the interpretational benefits of our approach, pMimCor appears to outperform approaches that do not utilise pathway information.

4.1 MIRNA

Micro-RNAs (miRNA) are a class of small non-coding RNA molecules which down regulate gene expression. Through pairing, miRNAs down-regulate the expression of genes by inhibiting their translation or promoting the degradation of their target messenger-RNAs (mRNA). Dysregulation of miRNAs can lead to a variety of human diseases with miRNAs being shown to play a critical regulatory role in many cellular pathways and functions such as developmental timing, cell death, cell proliferation, immunity, and patterning of the nervous system (Mo, 2012).

There exist many computational algorithms for predicting the target genes of miRNAs, such as TargetScan (Lewis *et al.*, 2005), miRBase (Griffiths-Jones *et al.*, 2008) and PicTar

(Krek *et al.*, 2005). These algorithms essentially attempt to identify whether a mRNA contains the binding motif of a miRNA. The output of these methods can generally be reduced to a large relatively sparse binary matrix, after appropriate thresholding. This remains an active area of research as there is generally not a large consensus between algorithms (Jayaswal *et al.*, 2009).

There is also biological interest in identifying changes in miRNA expression and their relationship with phenotypic outcome. Current state of the art methods attempt to identify groups of genes that are potentially being regulated by a miRNA or group of miRNAs. Such methods include canonical correlation analysis (Witten and Tibshirani, 2009), multivariate random forests (Jayaswal *et al.*, 2011) and integrative Bayesian analysis (Wang *et al.*, 2013). Statistically this can be thought of as a having high dimensional multivariate response and high dimensional multivariate covariates. Unfortunately, even if constrained by the output of a target binding prediction algorithm, all such methods rely quite strongly on having reasonable sample sizes, as they are attempting to estimate large complex networks. In addition to this, due to the high dimensional nature of the problem, the outputs from many of these methods can quickly become intimidating and quite difficult to interpret.

In situations of small sample sizes the biological focus must be slightly different. One common approach is to identify a differentially expressed (DE) miRNA and then use predicted binding target information to perform an enrichment analysis on the DE genes or miRNA-mRNA correlations (Xu and Wong, 2013). While this approach is statistically more satisfying than simply identifying significant pair-wise correlations between miRNA and mRNA (Havelange *et al.*, 2011; Li *et al.*, 2011), for which correcting for multiple testing is often ignored, it also has a couple of disadvantages. Firstly, there is generally little effort made to combine the significance from the test on the miRNA with the test on its targets. This means that any results may be overly conservative as two p-value cut-offs are performed instead of one. This could result in biologically significant signal being missed. Secondly, as miRNAs can potentially target hundreds or even thousands of genes, if a miRNA and its targets are identified as changed then further pathway enrichment analysis is often performed to make any results biologic-

ally interpretable. This adds another level of statistical complication and yet another p-value cut-off.

By integrating biological network information (pathways, protein-protein interaction networks) into a simple analysis framework, we may be able to directly test for and identify miRNAs that target a group of genes from a specific biological pathway. This has two clear benefits; we are now free of the statistically demanding task of identifying clusters of genes and our predefined clusters should be functionally more interpretable. However, this comes at the cost of being reliant on the accuracy of the target prediction algorithms and the quality of the biological annotation.

4.2 COMBINING P-VALUES

In order to construct and describe our analysis framework we first must have a solid understanding of different methods for combining measures of significance. There are many methodologies for combining information across studies or within pathways and the key discriminating differences between many of these methods are their assumed alternate hypotheses (Tseng *et al.*, 2012). Let n be the number of tests and τ_j , for $j = 1, 2, \dots, n$, the test statistics. Assume the null hypothesis that none of the features measured by these test statistics have changed. Li and Tseng (2011) propose two broad classes of alternative hypotheses H_A and H_B . The first class of alternate hypotheses, H_A , are used to detect a series of tests in which *all* the test statistics show change. The corresponding test can be expressed as

$$\begin{cases} H_0 : \mathbb{E}\tau_j = 0, \text{ for all } j = 1, 2, \dots, n. \\ H_A : \mathbb{E}\tau_j > 0, \text{ for all } j = 1, 2, \dots, n. \end{cases} \quad (4.1)$$

The second class of alternative hypotheses, H_B , are used to detect a series of tests in which *any* of the test statistics show change. The corresponding test can be expressed as

$$\begin{cases} H_0 : \mathbb{E}\tau_j = 0, \text{ for all } j = 1, 2, \dots, n. \\ H_B : \mathbb{E}\tau_j > 0, \text{ for at least one } j \text{ in } 1, 2, \dots, n. \end{cases} \quad (4.2)$$

Here we consider four methods of p-value combination that could be used to perform the previously described tests:

Fisher: Fisher's method is defined as $P\left(\mathcal{U}_j > -2 \sum_{j=1}^n \log(p_j)\right)$ (Fisher, 1925).

Stouffer: Stouffer's method is defined as $\Phi\left(\frac{\sum_{j=1}^n \Phi^{-1}(p_j)}{n}\right)$ (Stouffer *et al.*, 1949).

maxP: The maximum of the p_j for $j = 1, 2, \dots, n$ (Wilkinson, 1951).

OSP: A one-sided version of Pearson's method $P\left(\mathcal{U}_j < -2 \sum_{j=1}^n \log(1 - p_j)\right)$ (Pearson, 1934).

where Φ is the probability distribution function for the standard normal and \mathcal{U}_j is distributed as a chi-squared distribution with $2n$ degrees of freedom.

The combined p-values for Fisher's and Stouffer's methods approach zero if any one of the p_j also approach zero making them appropriate for testing H_0 against the alternative H_B . For maxP or OSP to approach zero all p_j must approach zero, thus making them appropriate for testing H_0 against the alternative H_A .

The properties of the four methods are further illustrated in Figure 4.1 from a two dimensional perspective. While the arbitrary cut-offs of maxP and OSP are quite different their overall topologies are quite similar. Fisher is also seen to be quite sensitive to any change. If one of the z-scores is larger than approximately 2.4, then regardless of the sign of the other z-score, the combined p-value will be less than 0.05.

4.2.1 Simulation

To illustrate and distinguish the performance of the four p-value combination methods in relationship to the two classes of alternative hypotheses (H_A and H_B), we perform a simulation study to assess how each method combines information from three test statistics. The distributions of the three test statistics were chosen to describe situations of no change, mild change and strong change in none, some or all of the statistics. In our simulation study, we simulate directly from the standard normal distribution, this is equivalent to simulating test statistics (potentially from a two-sample t-test) and transforming them into z-scores. In practice, these test statistics have most likely

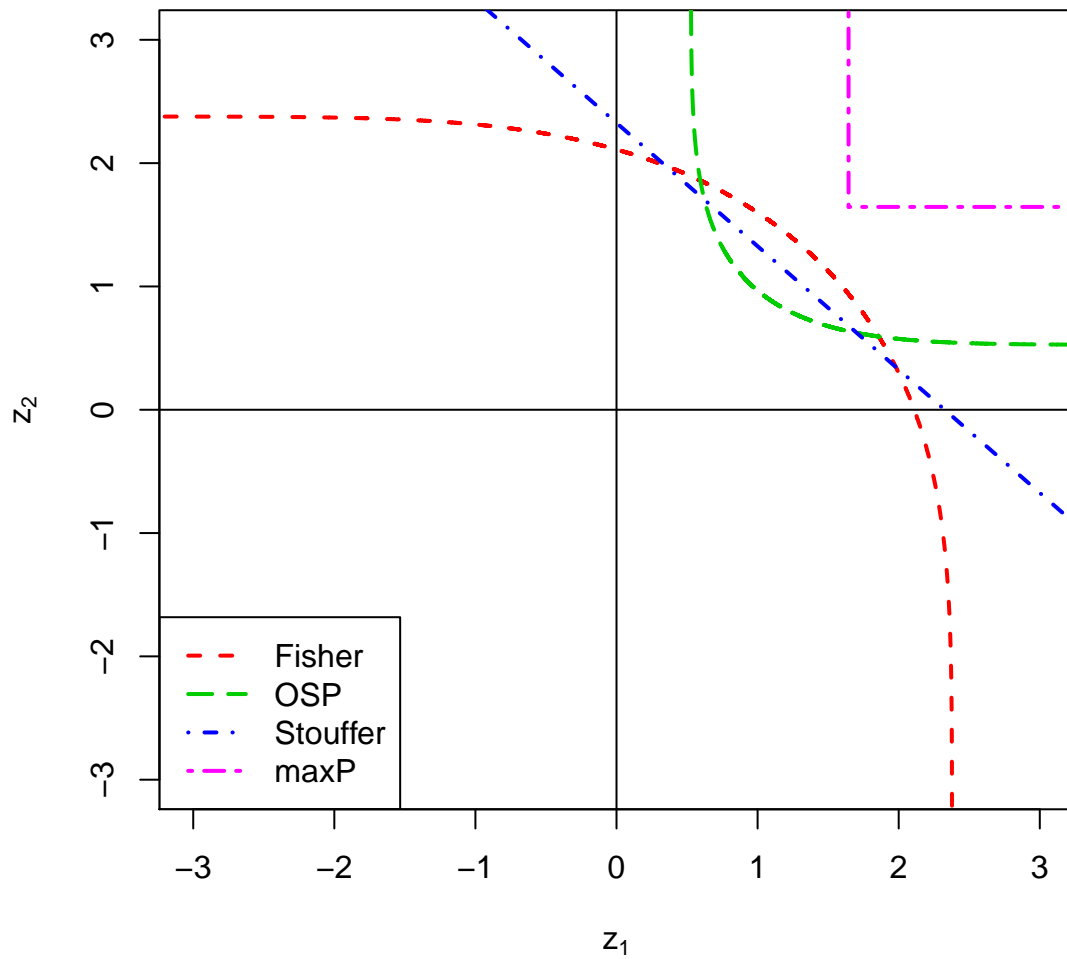


Figure 4.1: **p-value cut-offs for various combination methods** – A plot illustrating a p-value cut-off of 0.05 for various p-value combination methods in a two dimensional setting. The p-value cut-off is plotted in the negative z-score space so that a small p-value corresponds to a large positive z-score. The combination methods under consideration are Fisher (red), Stouffer (blue), maxP (pink) and OSP (green).

come from multiple two-sample t-tests but could be other statistics such as regression coefficients.

In each simulation we generated 1 000 000 observations of the three test statistics from the multivariate normal distribution $\mathbf{X}_i \sim N(\mu_i, I)$ where $\mu_1 = (0, 0, 0)$. We call this simulation I and it represents the initial test results obtained from three different sets of experiments. With all three components of μ_1 equal zero, this represents the situation where test statistics were generated from comparisons with no change. We have also examined different values of μ in different simulations.

- $\mu_2 = (2, 0, 0)$ in simulation II;
- $\mu_3 = (2, 2, 0)$ in simulation III;
- $\mu_4 = (4, 0, 0)$ in simulation IV and
- $\mu_5 = (2, 2, 2)$ in simulation V.

Notice, that simulation V represents the situation where all three test statistics from three experiments were generated from comparisons with change. Simulations II, III and IV represent situations where at least one of the three statistics were generated from comparisons with change.

We then applied each p-value combination method to each observation in each simulation. An observation was called significant if its overall significance (combined p-value) was less than an arbitrary cut-off of 0.05. The percentage of called significance from all observations was used to characterise each method's power in performing hypothesis testing under the two classes of alternative hypotheses H_A and H_B .

In general a method with a low percentage of significance in simulation I (as this simulation is consistent with the null hypothesis of no change) and high percentage in simulations II, III, IV and V (as these are simulations representing some change) would be a good method for testing H_0 against H_B ; that an observation is changed in *any* tests. In contrast, a method with a low percentage of significance in simulations I, II, III and IV (as these are simulations where not all statistics have changed) but high percentage in simulation V (a simulation where all the statistics have changed) would be more suitable for testing H_0 against H_A that an observation is changed in *all* tests.

Table 4.1: Results for five simulations in evaluating the performance of the four p-value combination methods at testing H_A and H_B . The percentage of combined p-values less than 0.05 over 1 000 000 simulations (rounded to two decimal places) are reported.

Simulation Method	I $\mu_1 = (0, 0, 0)$	II $\mu_2 = (2, 0, 0)$	III $\mu_3 = (2, 2, 0)$	IV $\mu_4 = (4, 0, 0)$	V $\mu_5 = (2, 2, 2)$
Fisher	0.05	0.43	0.80	0.95	0.95
Stouffer	0.05	0.31	0.75	0.75	0.97
maxP	0.00	0.00	0.02	0.00	0.26
OSP	0.05	0.17	0.46	0.20	0.93

Results from the simulation study can be seen in Table 4.1. Focusing on the results from simulation I, Fisher, Stouffer and OSP all call five percent of the observations significant. As this simulation represents the situation where test statistics were generated from comparisons with no change, all the observations called significant are false positives. As an arbitrary cut-off of 0.05 was used, it is comforting to see that the false positive rates of Fisher, Stouffer and OSP are consistent with this.

The set of simulations demonstrates that OSP is the most suitable method at testing H_0 against H_A . OSP is competitive with Fisher and Stouffer in simulation V but calls much less significance in simulation II, III and IV compared to the other two methods.

When considering the class of alternative hypothesis H_B , Stouffer is most powerful in detecting observations that have changed in all three tests (simulation V). It also has higher power in detecting changes from simulations II, III, and IV when compared to OSP. In comparison, Fisher has the highest percentage in simulations II, III and IV. These results suggest that both Stouffer and Fisher are most suitable for testing H_0 against H_B . However, while Fisher appears to be more sensitive to any changes, having the highest percentage in simulations II, III and IV, Stouffer is relatively more conservative.

In our application, we will select Stouffer to combine statistics within a pathway so that any pathway results are less likely to be driven by a single protein. These simulations demonstrate the importance of having a clearly defined alternate hypothesis in mind when analysing data.

4.3 PMIM - PATHWAY, MICRORNA AND MRNA INTEGRATION

In the following we will propose a general framework for integrating various data sources to identify regulatory miRNAs and their targets. This framework will be utilised to form three specific methods cMimDE, pMimDE and pMimCor. We will propose cMimDE as a method to identify regulatory miRNAs and their targets in an experiment. We will then introduce the concept of mir-pathways and use these to develop pMimDE and pMimCor as methods for identifying regulatory miRNAs that may be targeting a group of genes that share a common biological function.

Before these methods can be described in detail let us first introduce the data sources and their corresponding notations; a visual representation of the data sources is given in Figure 4.2 and described in more detail in the following. Let \mathbf{Y} be a p_g by n matrix corresponding to gene expression estimates for p_g genes in n samples, where the n samples are divided into two conditions such that $\gamma(k) = 1$ or 2 . Similarly let \mathbf{Z} be a matrix of expression estimates for p_{mi} miRNAs and the same matched n samples. Assume there also exists some p_{mi} by p_g matrix \mathbf{M} , where M_{ij} is equal to one if the i^{th} miRNA is predicted to bind to the j^{th} gene and is zero otherwise. Furthermore, assume there exists some functional or pathway annotation of the genes and represent this as a p_f by p_g matrix \mathbf{F} , where F_{ij} is equal to one if the j^{th} gene is in the i^{th} group and zero otherwise.

Prior to integration the information from matrices \mathbf{Y} and \mathbf{Z} are reduced to summary statistics. We would like to combine information from three statistics; t -statistics from gene level differential expression, t -statistic from miRNA level differential expression and the correlations between a miRNA and a gene. We will calculate these as described in the following paragraphs.

A moderated t-statistic

In the following we will calculate a test statistic for differential expression in the miRNA or gene level data with a mildly moderated t -test. Assuming some data matrix

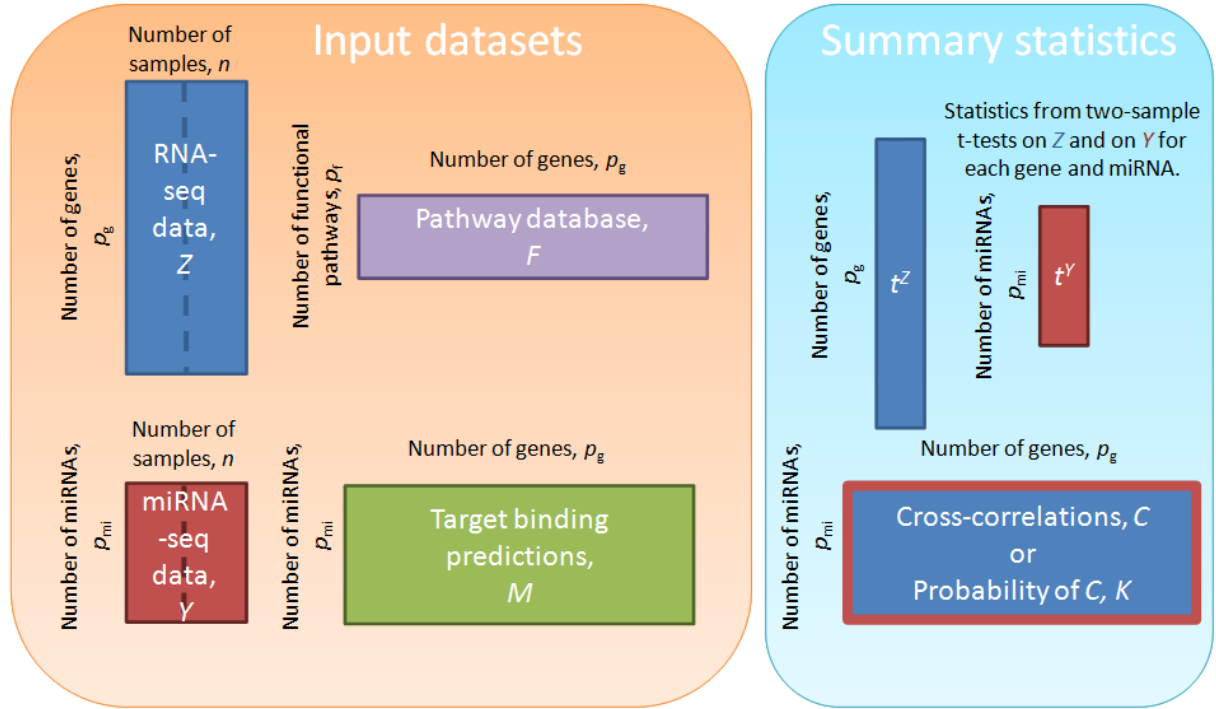


Figure 4.2: **Data matrices and summary statistics** – A visual representation of the input data matrices which include the matched mRNA- and miRNA-Seq data, a pathway database and the miRNA target binding predictions. Also represented are the corresponding summary statistics used which are the statistics from a moderated two-sample t-test performed on the mRNA-Seq data, the statistics from a moderated two-sample t-test performed on the miRNA-Seq data and the cross correlations or probability of observing the cross correlations between the miRNA- and mRNA-Seq data.

X , this statistic can be calculated as follows: Estimate the mean expression value of each condition as

$$\mu_{i1} = \frac{\sum_{j:\gamma(j)=1} X_{ij}}{\sum_{j:\gamma(j)=1} 1} \quad (4.3)$$

$$\mu_{i2} = \frac{\sum_{j:\gamma(j)=2} X_{ij}}{\sum_{j:\gamma(j)=2} 1} \quad (4.4)$$

The moderated variance will be estimated as the maximum of the sample variance and the average expression

$$\sigma_i^2 = \max \left(\frac{\mu_{i1} + \mu_{i2}}{2}, \frac{1}{n-2} \left(\sum_{j:\gamma(j)=1} (X_{ij} - \mu_{i1})^2 + \sum_{j:\gamma(j)=2} (X_{ij} - \mu_{i2})^2 \right) \right) \quad (4.5)$$

The test statistic for this moderated t-test is then

$$t_i^X = \sqrt{n} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right) \quad (4.6)$$

and will have approximately $n - 2$ degrees of freedom.

Correlation between genes and miRNA

Define the entries of the p_{mi} by p_g cross correlation matrix C between the miRNA and genes as

$$C_{ij} = \frac{\sum_{k=1}^n (Y_{jk} - \mu_j^Y)(Z_{ik} - \mu_i^Z)}{\sqrt{\sum_{k=1}^n ((Y_{jk} - \mu_j^Y)^2 \sum_{k=1}^n (Z_{ik} - \mu_i^Z)^2)}}, \quad (4.7)$$

where μ_j^Y is the average expression for gene j and μ_i^Z is the average expression for miRNA i . These correlations could be transformed using the Fisher transformation to be approximately standard normal. Define these transformed correlations as the matrix \mathbf{K} , where the entries of \mathbf{K} can then be calculated as

$$K_{ij} = \sqrt{n-3} \tanh^{-1}(C_{ij}). \quad (4.8)$$

If n is small (e.g $n < 10$) then the Fisher transformation will be inappropriate. As C is a very large matrix we will use a probit rank transformation to make the correlations approximately standard normal. We will define the elements of \mathbf{K} as

$$K_{ij} = \Phi^{-1} \left(\frac{r_{ij} - 0.5}{p_{mi}p_g} \right) \quad (4.9)$$

where r_{ij} is the rank of the corresponding element in the p_{mi} by p_g cross-correlation matrix.

4.3.0.1 *cMimDE - Classic microRNA and mRNA integration using DE*

As mentioned earlier a common approach for highlighting interesting miRNA-mRNA relationships is to first identify a differentially expressed (DE) miRNA and then use predicted binding target information to perform an enrichment analysis on the DE

genes or miRNA-mRNA correlations (Xu and Wong, 2013). This is similar to taking the maximum of the DE miRNA p-value and the enrichment analysis p-value. As demonstrated in the earlier simulations (Section 4.2.1) taking the maximum of two p-values is generally not ideal. In the following we will propose a method, cMimDE, which is both a formalisation and improvement of this approach.

We would like to test whether a miRNA is DE and its target genes are DE in the opposite direction. From the datasets \mathbf{Y} and \mathbf{Z} we can calculate the vectors of test statistics \mathbf{t}^Y and \mathbf{t}^Z for whether genes and miRNAs, respectively, are DE between conditions. First consider finding miRNAs with negative log fold change and genes with positive log fold change. We can then perform a gene set test for whether the target genes of the i^{th} miRNA are up-regulated using Stouffer's method

$$s_i = \Phi \left(\frac{\sum_{j:M_{ij}=1} \Phi^{-1}(P(t_j^Y > 0))}{\sqrt{\sum_{j:M_{ij}=1} 1}} \right). \quad (4.10)$$

This gene set test could also have been performed using any method that is favoured, such as by using Fisher Method, a Wilcoxon rank-sum test, over-representation test or gene set enrichment analysis (GSEA).

To identify those miRNA that are negatively DE and whose target genes are positively DE we can then combine the vectors \mathbf{t}^Z and \mathbf{s} using OSP, where

$$p_i^{cMimDE} = P(\chi_4^2 < -2 \log((1 - P(t_i^Z < 0))(1 - s_i))). \quad (4.11)$$

This test will rank similarly to taking the maximum of the p-values from the miRNA DE and its corresponding gene set test.

4.3.0.2 *pMimDE - Pathway, microRNA and mRNA integration using DE*

In addition to finding miRNA that are regulating their target genes it may be of interest to find miRNA that are regulating a set of genes that also share some common biological function or outcome. We can test this by performing gene set tests on the genes that lie in the intersection of the binding target predictions of a miRNA and a biological pathway. These intersections will be referred to as mir-pathways. We can

again use Stouffers method to find evidence of whether a set of genes in a particular mir-pathway (ie. genes that are both targeted by the i^{th} miRNA and also belong to the k^{th} functional pathway) are up-regulated. We will store these gene-set tests in the matrix \mathbf{S} as

$$S_{ik} = \Phi \left(\frac{\sum_{j: M_{ij}=1 \& F_{kj}=1} \Phi^{-1}(P(t_j^Y > 0))}{\sqrt{\sum_{j: M_{ij}=1 \& F_{kj}=1} 1}} \right). \quad (4.12)$$

S_{ik} whose corresponding mir-pathways contain less than two intersecting genes will be defined as empty.

We could again combine these gene set tests \mathbf{S} with the miRNA test statistics t^Z using OSP. However, as using OSP is similar to taking the maximum of two p-values, then as multiple gene set tests are being performed on the target genes of each miRNA, we might expect at least one of the p-values of these gene set test to be smaller than the miRNA p-value due to multiple testing. This would result in the ranking of the minimum combined p-values for each miRNA being very close to the ranking of the miRNA p-values. To account for these multiple testing issues we will apply a multiple testing correction to each row of \mathbf{S} before combining using OSP. Many of the mir-pathways may contain similar genes and hence their expression may be correlated. We will be conservative and ignore this correlation performing a row-wise Benjamini-Hochberg (Benjamini and Hochberg, 1995) correction to \mathbf{S} and calling this \mathbf{S}^{fdr} .

To identify those miRNAs that are negatively DE and whose target genes are both positively DE and share some common biological function the elements of the matrix \mathbf{S}^{FDR} can now be combined with the elements of the vector t^Z using OSP, where

$$p_{ik}^{\text{pMimDE}} = P(\chi_4^2 < -2 \log((1 - P(t_i^Z < 0)) (1 - S_{ik}^{\text{fdr}}))). \quad (4.13)$$

and $\mathbf{P}^{\text{pMimDE}}$ is a p_{mi} by p_{f} matrix.

4.3.0.3 *pMimCor - Pathway, microRNA and mRNA integration using correlation*

By finding sets of genes that are DE in the opposite direction to a miRNA we were essentially requiring the expression of genes to be negatively correlated with the ex-

pression of their corresponding miRNA between conditions. In the following we will extend pMimDE to simply require that the expression of a miRNA and gene be negatively correlated. As we expect the correlation between the expression of a miRNA and its targets to be negative the test described for pMimDE can be modified by replacing the definition of S_{ik} with

$$S_{ik} = \Phi \left(\frac{\sum_{j: M_{ij}=1 \& F_{kj}=1} K_{ij}}{\sqrt{\sum_{j: M_{ij}=1 \& F_{kj}=1} 1}} \right). \quad (4.14)$$

4.4 EVALUATION STUDY

In this study, we evaluate our proposed framework for integrating small sample miRNA-Seq and mRNA-Seq with pathway information, pMimCor. This evaluation consists of two components. The first component assesses the validity of the information in both the KEGG pathway annotation (Kanehisa *et al.*, 2008) and the TargetScan binding target predictions. The second component evaluates the performance of our proposed integration approach on our motivational dataset.

4.4.1 Data

We will evaluate pMimCor on a conditional Notch2 knockout experiment described in more detail in Chapter 5. This experiment compares matched mRNA and miRNA enriched samples in three wild type (WT) mice and three conditional Notch2 knockout (NCN) mice. The samples were sequenced with an Illumina HiSeq 2000 which provided 51bp reads. Adaptor sequencers were cut from the miRNA enriched samples using cutAdapt (Martin, 2011).

Bowtie (Langmead *et al.*, 2009) was used to map the samples back to the mm9 genome (Church *et al.*, 2009). The mRNA read alignments were allowed two mismatches while the miRNA alignments were allowed one mismatch. Reads that aligned to multiple regions on the genome were ignored. Reads were then summarised into gene or miRNA counts using the Union method (Bullard *et al.*, 2010) and the Ensembl gene

annotation (Hubbard *et al.*, 2009). Exons that overlapped multiple genes were included for the miRNA enriched samples.

4.4.2 Evaluation strategies and results

The key strength of pMimCor is to rank both miRNAs and pathways concurrently, ie ranking the top mir-pathways. As such, this is a favourable approach for integration as it facilitates interpretation and addresses the biological question directly. This ranking is demonstrated in Tables 5.6 and 5.7 in the next chapter where a qualitative assessment of them is given. The following approaches assess these rankings in a more quantitative manner.

4.4.2.1 Assessing the information in KEGG and TargetScan

A resampling scheme will be used to assess the information content within KEGG and TargetScan. This scheme is driven by the hypothesis that if miRNA regulation plays a part in phenotypic outcome then a database of genes associated with phenotypic outcomes (such as KEGG) should contain phenotypes that have more associated genes that are predicted miRNA binding targets than expected. Define the resampling scheme as follows:

RESAMPLING SCHEME 1 Mir-pathways are defined using KEGG and TargetScan. The number of mir-pathways that contain n genes, for $n = 1, \dots, 20$, are calculated. This is then repeated, taking the average number of mir-pathways that contain n genes, after randomly reassigning the genes for each pathway in KEGG one hundred times, reassigning the target genes for each miRNA in TargetScan one hundred times, and, reassigning genes in both KEGG and TargetScan one hundred times.

The term mir-pathway corresponds to the intersecting genes of a particular KEGG pathway and miRNA. A conservative approach to resampling is taken, with the pathway and target information being resampled only from the genes that are in both KEGG and TargetScan. The top thirty percent, by score, of the TargetScan predicted miRNA-

mRNA interactions were used to construct a binary miRNA-mRNA target prediction matrix.

The results from the resampling scheme can be seen in Figure 4.3. The number of mir-pathways with n genes decreases if either the KEGG information or TargetScan information is randomised. As seen in Figure 4.4, there are about 50% more mir-pathways defined by KEGG and TargetScan that contain $n = 3, 4$ and 5 genes than expected by chance. This would appear to lend weight to the argument that miRNA regulation may play a role in some of the annotated KEGG pathways. These results are all exaggerated further if a less conservative approach is taken to the resampling, resampling the pathway and target information from all genes as opposed to only those present in both KEGG and TargetScan.

4.4.2.2 *Evaluating signal in the data*

One of the key frustrations of working with small sample sizes is the inability to use sampling methods, such as cross validation, to improve confidence about results. However, the experimental data is not the only input into pMimCor. As pMimCor also relies on large pathway and target matrices, these could be resampled instead.

Resampling the pathway and target matrices may introduce some bias into an evaluation of signal. As was demonstrated in scheme 1, resampled pathway or target matrices produce less mir-pathways and mir-pathways with less genes. However, as pMimCor performs a Benjamini-Hochberg correction on the mir-pathway p-values for each miRNA, this should over correct for any differences in the number of mir-pathways due to resampling. In this situation the Benjamini-Hochberg correction could be considered as overly conservative as it does not account for any of the mir-pathways being correlated.

A method for evaluating the output of pMimCor when run on the Lin data is described as follows:

RESAMPLING SCHEME 2 pMimCor is run using KEGG and TargetScan as inputs. The number of significant mir-pathways are calculated at various arbitrary p-value cut-offs. This is then repeated, finding the average number of significant mir-pathways, after randomly reassigning the genes for each pathway in KEGG one

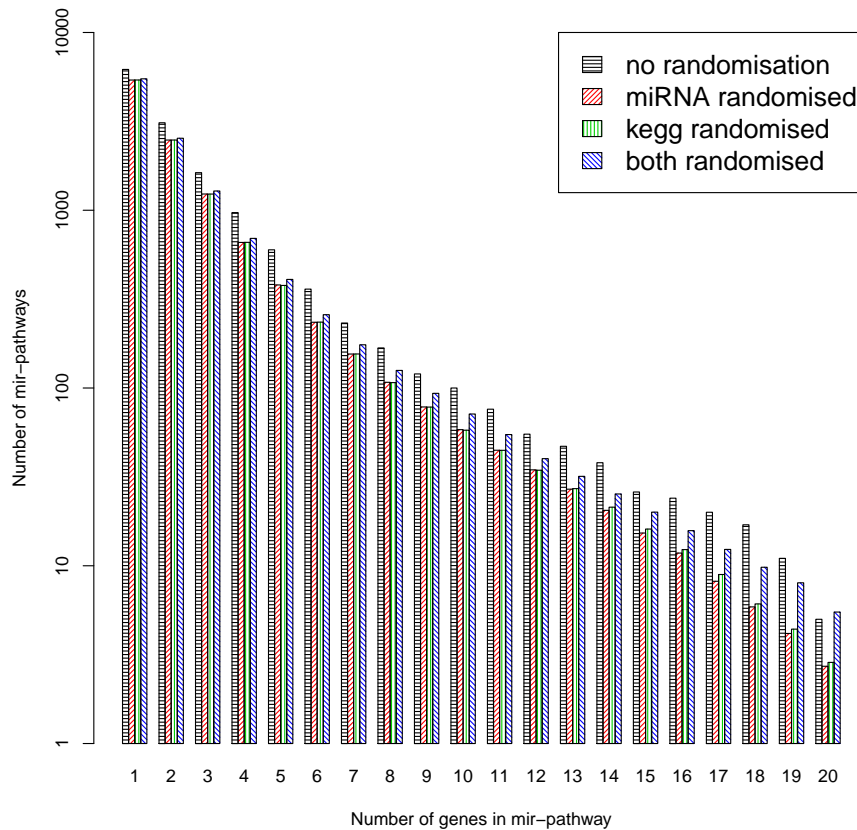


Figure 4.3: **Number of mir-pathways that contain n genes before and after randomisation** – For $n = 1, \dots, 20$ the average number of mir-pathways that contain n genes are plotted on a log-scale. These are plotted for mir-pathways calculated using TargetScan and KEGG, randomised TargetScan and KEGG, TargetScan and randomised KEGG and both randomised TargetScan and randomised KEGG.

hundred times, reassigning the target genes for each miRNA in TargetScan one hundred times, and, reassigning genes in both KEGG and TargetScan one hundred times.

The p-values have not been adjusted for multiple testing across miRNA. If there is signal related to miRNA regulation in the data, the binding target predictions are accurate and there exists annotated phenotypes that are associated with miRNA regulation then resampling the pathway and/or target matrices should reduce the number significant mir-pathways found.

The results from the second resampling scheme can be found in Table 4.2. Randomising either the KEGG or TargetScan matrices reduces the number of mir-pathways

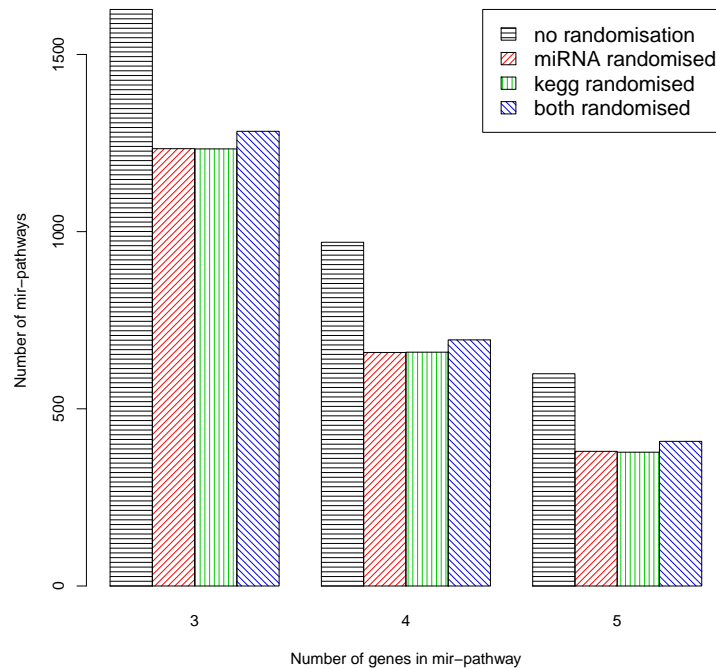


Figure 4.4: **Number of mir-pathways that contain 3, 4 or 5 genes before and after randomisation** – For $n = 3, 4$ and 5 the average number of mir-pathways that contain n genes are plotted. These are plotted for mir-pathways calculated using TargetScan and KEGG, randomised TargetScan and KEGG, TargetScan and randomised KEGG and both randomised TargetScan and randomised KEGG.

that are called significant. This suggests that both matrices may contain information pertinent to miRNA regulation. This also suggests that miRNAs may be regulating gene expression in response to loss of Notch2 function. It would also appear that there are more up-regulated miRNA, and hence miRNA down-regulated genes, in response to loss of Notch2.

4.4.2.3 Evaluation via literature search

We will evaluate the performance of our various proposed integration methods on our data, by assessing their concordance with results from a literature search. The Lin experiment was designed to study the loss of Notch2 function in the brain in the context neurodegeneration. With this in mind, we will use PubMed to identify miRNA that have been association with neurodegeneration. This information will allow us to

	Up-regulated				Down-regulated			
	0.001	0.005	0.01	0.05	0.001	0.005	0.01	0.05
No Randomisation	3.0	13.0	32.0	181.0	0	1.0	14.0	171.0
TargetScan Randomised	2.0	10.3	19.7	95.9	0.6	3.3	9.0	56.9
Kegg Randomised	0.6	6.5	14.1	76.4	0.5	1.4	3.5	22.9
Both Randomised	1.5	5.6	11.7	61.7	0.3	1.6	4.4	29.9

Table 4.2: **The number of significant mir-pathways** – A table of the average number of significant mir-pathways calculated at various arbitrary p-value cut-offs. Significance is calculated for mir-pathways estimated using TargetScan and KEGG, randomised TargetScan and KEGG, TargetScan and randomised KEGG and both randomised TargetScan and randomised KEGG. Results are shown for both up-regulated miRNA and down-regulated miRNA.

concurrently verify the relationships between miRNA regulation, the loss of Notch2 function and neurodegeneration and test the effectiveness of our analysis framework.

Our strategy for performing and evaluating the literature search is outlined as follows:

LITERATURE SEARCH STRATEGY A PubMed search was performed for each miRNA that was observed in our data. Each miRNA was included in the following analysis if it had at least one publication referring to it in the abstract. A second batch of searches were performed searching for each miRNA and the word “neurodegeneration”. A miRNA was then classified as being associated with neurodegeneration if it had at least one hit on this search. This information was then used to calculate the number of true positives (TP) and false positives (FP) for the following methods:

- miRNA DE - performing a moderated t-test on just the miRNA data.
- cMimDE - a classic microRNA and mRNA integration using just miRNA and gene DE.
- pMimDE - Pathway, microRNA and mRNA integration using DE.
- pMimCor - Pathway, microRNA and mRNA integration using correlation.

As pMimDE and pMimCor have many p-values for each miRNA, the minimum of these for each miRNA was taken to provide a rank for the miRNA.

This approach will be biased towards previous research and/or methodology.

Figure 4.5 plots the TP against FP for the region of small FP. For this region, pMimCor performs better than all other approaches with quite dramatic improvements to just using the miRNA data. Five of the top seven miRNA ranked by pMimCor have been associated with neurodegeneration in the literature. The performance of these top ranked pMimCor miRNA should warrant some optimism.

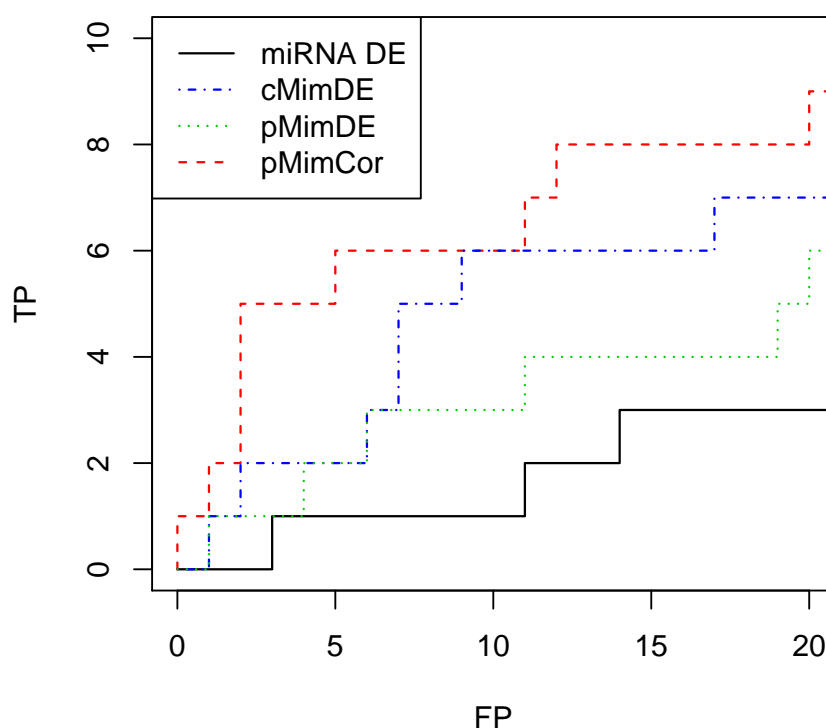


Figure 4.5: **TP vs FP from PubMed search** – True Positives (TP) are plotted against False Positives (FP) for the small FPR region of Figure C.1. The plotted lines are for four methods, miRNA DE (black), cMimDE (blue), pMimDE (green) and pMimCor (red).

4.5 CONCLUSIONS AND FURTHER DISCUSSION

We proposed a general framework for integrating miRNA-Seq and mRNA-Seq data. This framework included three methods cMimDE, pMimDE and pMimCor. In particular, pMimCor, appears to be a favourable approach for integrating miRNA-Seq and

mRNA-Seq in the presence of small sample sizes. Not only did this approach identify more signal than expected by chance, but the identified signal was also associated with the observed phenotype.

Our integration approach has raised some statistically interesting questions that should warrant further investigation. To transcend the typical modularised analysis framework, our approaches make use of p-value combination methods to combine significance. By combining significance from a differential expression test of one miRNA with tests on multiple gene sets, pMimCor runs into an issue related to correction for multiple comparisons. While there has been some research into correction for multiple comparisons in the presence of correlated pathways (Holmans *et al.*, 2009), this methodology could be extended further to account for having correlated summary statistics from multiple miRNA. It would also be interesting to establish whether it is more appropriate to correct for multiple testing before or after utilising a p-value combination method like Fisher or OSP.

Our framework relies heavily on the hypothesis that databases like KEGG may contain pathways associated with miRNA regulation. This was shown to be highly plausible. While our framework is constrained to the accuracy of the pathway annotation, this constraint does add the benefit of making the results from pMimCor highly interpretable.

CASE STUDY OF THE LIN DATA

In collaboration with the Lin lab of Cornell University we designed an experiment to measure both mRNA and miRNA transcription in the brains of mice that are exposed to various stressors using RNA-Seq. This chapter will use the data from this experiment to illustrate the various methodologies proposed earlier in this thesis.

5.1 DESIGN

The Lin lab primarily studies the development and degeneration of the nervous system using the mouse olfactory system as a model. Once neurons are born, they are exposed to a variety of environmental insults that must be properly dealt with to avoid degeneration. Neurodegenerative disorders, such as Alzheimer's disease, are thought to arise in part due to a failure to deal with this increased stress.

Table 5.1 outlines a prototype experiment that was designed to provide insight into how the mouse brain responds to various stressors. For each sample in this experiment matched mRNA and miRNA enriched samples were sequenced with an Illumina HiSeq 2000. The induced stressors included treatment with pro-oxidant buthionine sulphoximine (BSO) to induce oxidative stress, a mouse strain with a conditional Notch2 knockout (NCN) and Borchelt mice, a strain of mouse with a chimeric human/mouse amyloid precursor protein with the "Swedish" mutation as well as mutant human presenilin 1 (Jankowsky *et al.*, 2004). Due to some of the difficulties associated with obtaining mice with a conditional knockout mutation of Notch2, the design of this experiment is less than favourable.

Notch2 is an essential gene for development and hence Notch2 mutant mice do not survive beyond embryogenesis. In order to study the function of Notch2 in adult animals, we employed a conditional knockout mutation of Notch2. However, the genetics of this conditional knockout are complicated. First, a homozygous floxed line must be crossed to the cre driver. This generates a double heterozygous mouse (cre/+, flox/+). These mice must then be crossed out to the parental floxed line (flox/flox), this generation contains controls (cre/+;flox/+ or +/+;flox/+) and the (cre/+;flox/flox) mutants. However, only one out of eight mice in this second generation will be controls and one in four will be mutants. So the proportion of the animals needed for the experiment

Genotype	Treatment	Age	Sex	Lane
WT	BSO	6 months	male	5
WT	BSO	6 months	male	2
WT(+)	BSO	6 months	male	3
NCN	BSO	6 months	male	4
NCN	BSO	6 months	male	1
NCN	BSO	6 months	male	6
WT	none	7 months	female	1
WT	none	7 months	female	6
WT	none	15 months	male	4
NCN	none	15 months	male	2
NCN	none	15 months	male	3
NCN	none	15 months	male	5
Borchelt	none	15 months	male	7
Borchelt	none	15 months	male	7
Borchelt	none	15 months	male	8
Borchelt	none	15 months	male	8

Table 5.1: **Design of the experiment** – This table describes the design of the experiment performed by the Lin lab. Listed are the genotype, age and sex of the mice as well as whether they were treated with BSO. Also included is the lane of the flowcell that each sample was sequenced on.

is very low, requiring a bit of luck or a lot of mating cages. This was a big problem as we would often get a mutant but no controls, so it was hard to get age matched and litter matched mice. In Table 5.1 the wild type (WT) mice are $+/+;flox/+$ and the WT+ mice are $cre/+;flox/+$. NCN is used to refer to the mice with the conditional knockout mutation of Notch2 where NCN stands for nestin-cre/Notch2 flox.

The design of this experiment is quite complex. For simplicity, this case study will focus on the comparison of untreated WT and NCN mice.

5.2 MAPPING

Mapping is the process of aligning reads to a reference genome (or transcriptome) and hence inferring from where they may have been transcribed. This process is not trivial. Aligning trillions of reads to a reference genome that is over three billion base pairs long is a computationally demanding task. As reads represent limited intervals of fragmented transcripts, the process of mapping reads back to a genome is further

complicated by reads that may span splice junctions. In this case study we ignore such reads.

Mapping of mRNA samples

Sequencing of the mRNA samples generated 1.3 trillion reads. These reads were 51 base pairs long. The number of reads that were sequenced for each sample can be seen in Figure 5.1a. While there is a large amount of variability in the number of reads sequenced per sample, there does not appear to be any samples which are extremely under-sequenced.

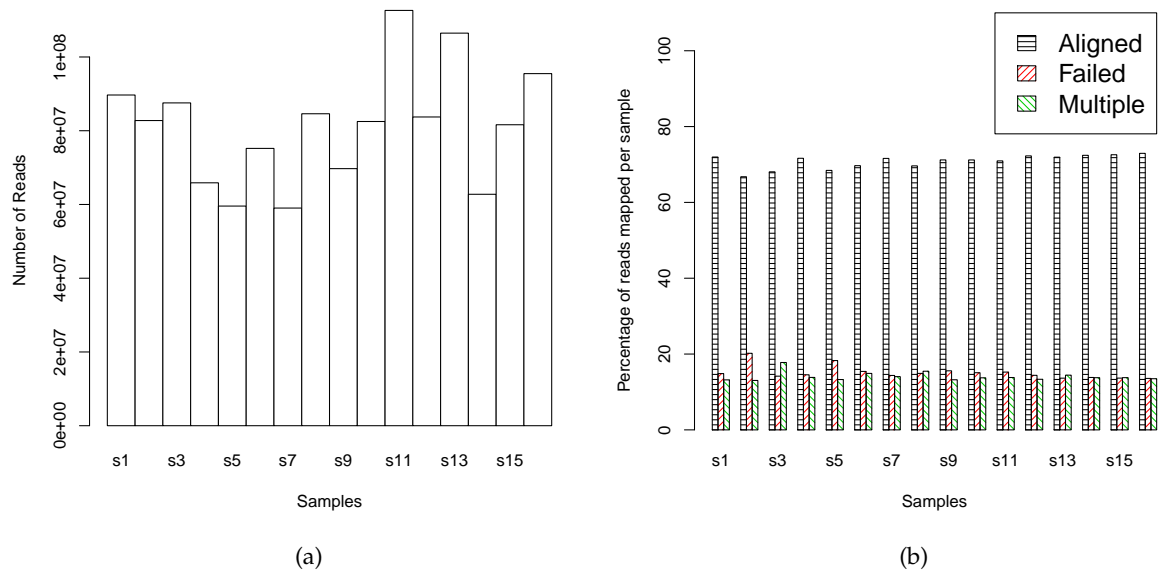


Figure 5.1: **Number of mapped mRNA** – (a) A histogram of the number of reads that were sequenced for each mRNA sample. (b) A bar plot of the percentage of these reads that mapped uniquely, failed to map and mapped to multiple regions on the genome.

Bowtie (Langmead *et al.*, 2009) was used to map the samples back to the mm9 genome (Church *et al.*, 2009). Alignments were allowed two mismatches. Reads that aligned to multiple regions on the genome were ignored. A breakdown for each sample of the percentage of reads that were mapped, unmapped and had multiple mappings can be seen in Figure 5.1b. Overall 71% of the reads mapped uniquely back to the genome.

Mapping of miRNA samples

The protocol for sequencing miRNA involves a size selection step. This essentially enriches the samples for RNA fragments that are the expected length of a mature miRNA (22 nucleotides). However, the reads we received from the sequencer were of length 51. This means that if we have selected and sequenced a mature miRNA, the majority of the base pairs of a read should be unnecessary.

A large majority of the unnecessary base pairs on the reads correspond to those from adaptor reads that have joined to the miRNA. Before the reads can be mapped back to the genome these adaptor sequences must be cut from the reads. This cutting was performed using cutAdapt (Martin, 2011). Figure 5.2 visually represents the number of reads that had a certain number of base pairs cut. The vast majority of reads had 29 base pairs cut. This would imply that our reads were enriched with initial fragments of length 22, the expected length of a mature miRNA.

Sequencing of the miRNA samples generated 156 million reads. The number of reads that were sequenced for each sample can be seen in Figure 5.3a. Again we observe a large amount of variability in the number of reads sequenced per sample.

Bowtie was used to map the samples back to the mm9 genome. Alignments were allowed only one mismatch. Reads that aligned to multiple regions on the genome were ignored. A breakdown for each sample of the percentage of reads that were mapped, unmapped and had multiple mappings can be seen in Figure 5.3b. Overall, only 36% of the reads mapped uniquely back to the genome and 54% of the reads mapped to multiple regions on the genome. These numbers could potentially have been improved by mapping with no mismatches before mapping with one mismatch.

5.3 SUMMARISATION

Summarisation is the process of summarising the mapping information into read counts (or expression values) for all genes of interest. This will produce a large matrix of read counts.

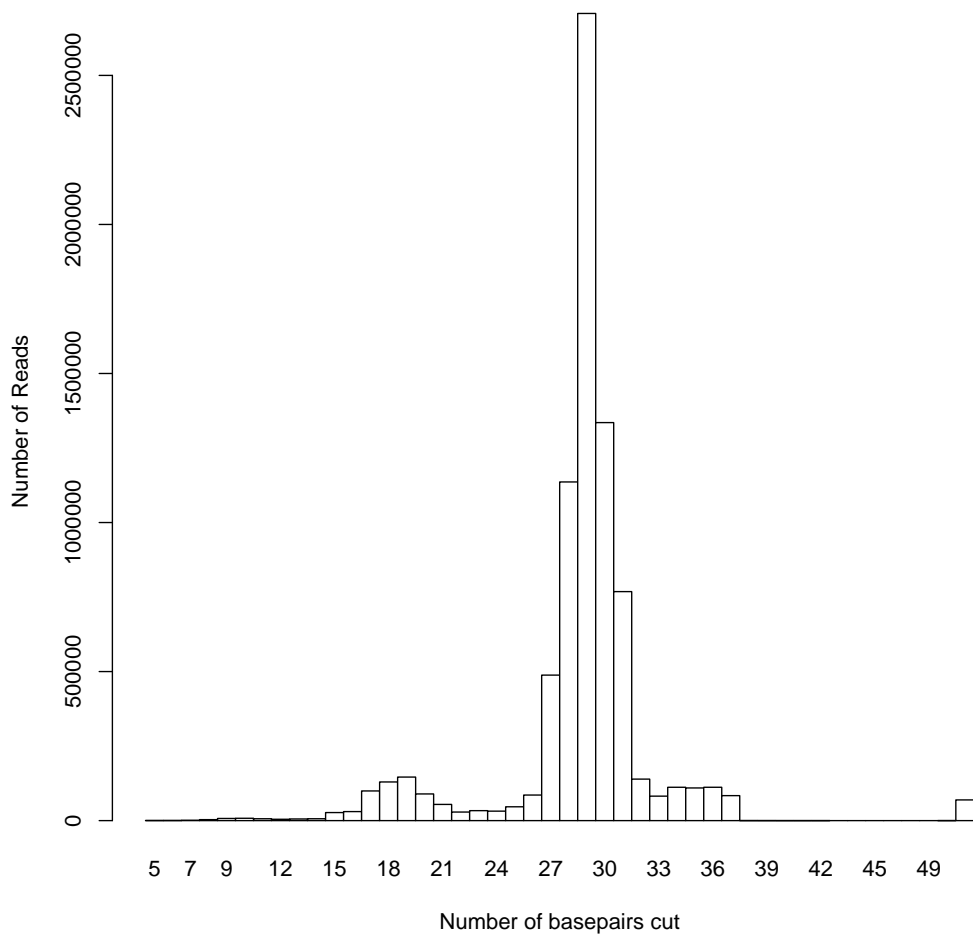


Figure 5.2: **Number of base pairs cut by cutAdapt** – A histogram describing the number of reads that had a certain number of base pairs cut by cutAdapt.

Summarisation of mRNA samples

We have chosen to analyse the mRNA data at the gene level. The mapped reads were summarised using the three summarisation approaches described in Chapter 2 in conjunction with the Ensembl gene annotation. These are Union, UI and exClust (see Chapter 2).

The number of reads that each method summarised can be seen in Table 5.2. These numbers appear concordant with those seen in the datasets presented in Chapter 2. With exClust including more than twice as many reads as the UI definition of a gene. It is also interesting to see that Union, the least stringent approach, still only included

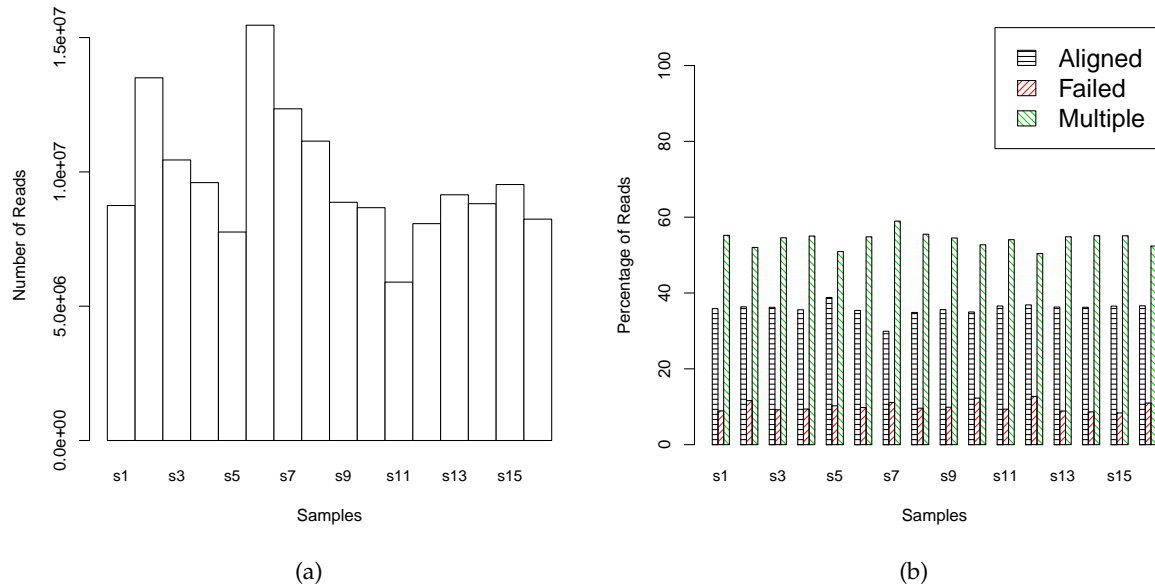


Figure 5.3: **Number of mapped miRNA** – (a) A histogram of the number of reads that were sequenced for each miRNA sample. (b) A bar plot of the percentage of these reads that mapped uniquely, failed to map and mapped to multiple regions on the genome.

Method	Total count	Number of genes with counts greater than zero	Number of genes with counts greater than twenty
Union	722 111 653	28 725	16 513
UI	170 042 046	18 068	8 747
exClust	393 618 731	28 722	16 454

Table 5.2: **Number of reads summarised** – Tabulated are the number of reads summarised by the Union, UI and exClust approaches. For each approach the number of genes with average counts greater than zero and twenty are also included.

about half the number of reads that were reported to have mapped back to the genome. For the remainder of the study we will use the exClust summarised mRNA data.

Summarisation of miRNA samples

The miRNA enriched samples were summarised to the Union definition of a gene using the Ensembl gene annotation. In this situation the Union definition was quite literal as we did not ignore exon regions which overlapped multiple genes. We then extracted the counts for the annotated miRNA and continued with these for the remainder of the

miRNA analysis. In total there were 23 168 700 reads that aligned to annotated genes on the mouse genome with 21 593 357 of these aligning to annotated miRNA. There were 393 miRNA with at least one aligned read and 236 with an average of twenty reads or greater.

As summarisation was performed using all genes, we are able to capture the background expression of genes whose mRNA made it through the size selection step in the miRNA enriched samples. This can be observed in Figure 5.4 where the gene counts from a mRNA sample are plotted against the gene counts from a miRNA enriched sample. The hollow circles are annotated genes and the solid red circles are annotated miRNA. The background expression of the genes in the miRNA enriched samples appears to be correlated with the gene expression from the mRNA samples. The annotated miRNA in general have very low expression in the mRNA samples but are quite highly expressed in the miRNA enriched samples. This indicates that the size selection for miRNA appears to have worked as the annotated miRNA are highly expressed in the miRNA enriched samples.

5.4 NORMALISATION

Before analysing RNA-Seq data it is important to normalise or model for any systematic technical variation that may have arisen in the measurement process. The largest abnormality generally observed in a RNA-Seq experiment is differences in library sizes. That is, different samples can have very different amounts of total reads mapped.

The trimmed means of M-values (TMM) method (Robinson and Oshlack, 2010) was used to model differences in library sizes between the mRNA samples. In this case, for each gene, an M-value is the log ratio of the gene count from one sample with the average count across all samples. For all genes with average count greater than one hundred, the thirty percent trimmed mean is taken of the M-values. This can be used to model differences in library sizes and should be robust to differentially expressed genes. Instead of carrying through these modelled differences into further analysis, we choose to use them to normalise the data. This is not ideal as it may influence the

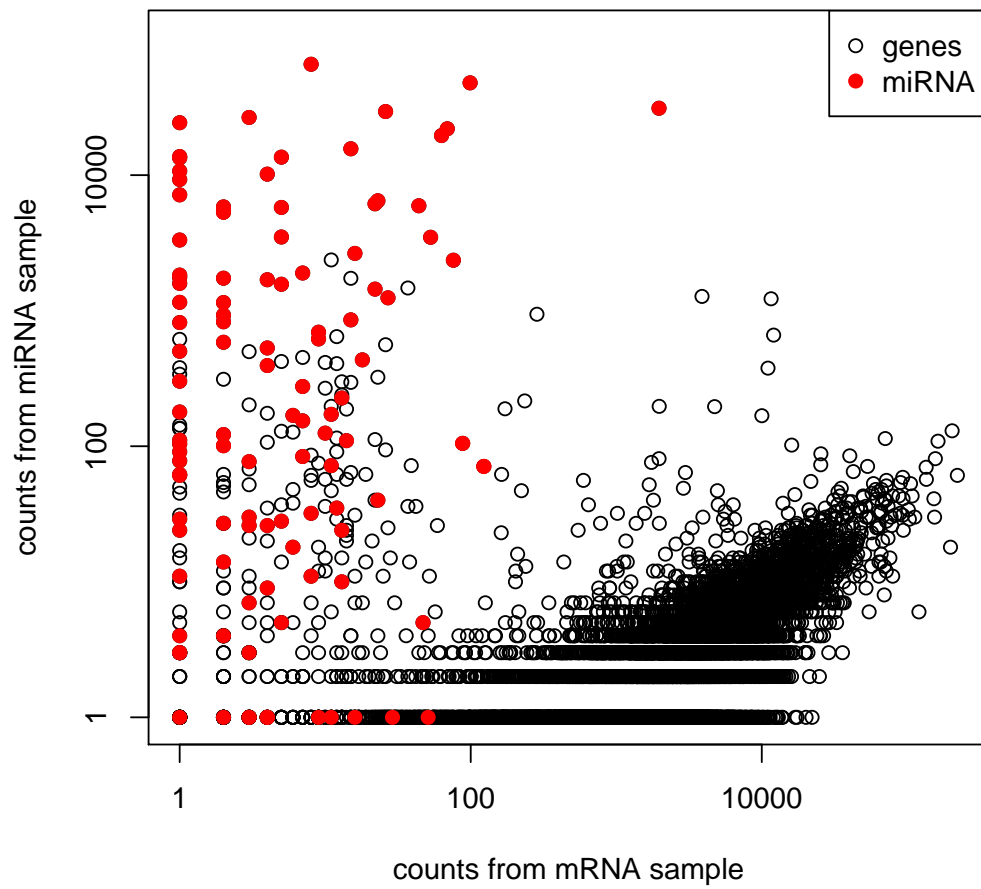


Figure 5.4: **Plot of summarised miRNA counts vs mRNA counts** – Gene counts from the mRNA and miRNA enriched samples are plotted against each other on a log-log scale. The hollow circles are annotated genes. The solid red circles are the annotated miRNA.

variance estimation of genes, particularly those with smaller average count. However, it does simplify further analysis.

Figure 5.5 demonstrates the modelled difference for the seventh sample. As observed in Figure 5.1 on page 76, the seventh sample has less total reads than most other samples. This is consistent with sample seven also having smaller M-values on average.

The miRNA enriched samples were also normalised using TMM. However, only the annotated miRNA were used to perform the normalisation.

5.5 DIFFERENTIAL EXPRESSION

Identifying genes that have changed in expression between conditions is one of the main aims of many RNA-Seq experiments. As many RNA-Seq experiments have small sample sizes, most methods for identifying differentially expressed (DE) genes perform some form of moderation to share information between genes. As our experiment has small sample sizes, we will use it to demonstrate some of the methods outlined in Chapter 3. This provides an opportunity to observe how the use of these methods may affect the results seen by an end user.

Four methods are used to identify differentially expressed genes between wild type (WT) and conditional Notch2 knockout (NCN) mice. The four methods are T, Tcommon, Tshrink and Tshrink+ (see Chapter 3). T is simply a two sample t-test, Tcommon is a t-test using the fitted common variance as a gene variance estimate, Tshrink uses a quasi-empirical Bayes moderation method to shrink the gene variances towards the common variance and Tshrink+ extends on Tshrink using the variance estimates from the four Borchelt mice to help fit the common variance before moderation.

The methods Tshrink and Tshrink+ both fit a shrinkage parameter λ . The parameter λ is inversely proportional to the mean squared error of the fitted common variance and can thus offer an indication of how well the fitted common variance is modelling the observed gene variances. When used on our data Tshrink produced a λ of 0.097 and Tshrink+ produced a λ of 0.227. As λ is higher for Tshrink+ this indicates that the fitted common variance of Tshrink+ is doing a better job of modelling the observed

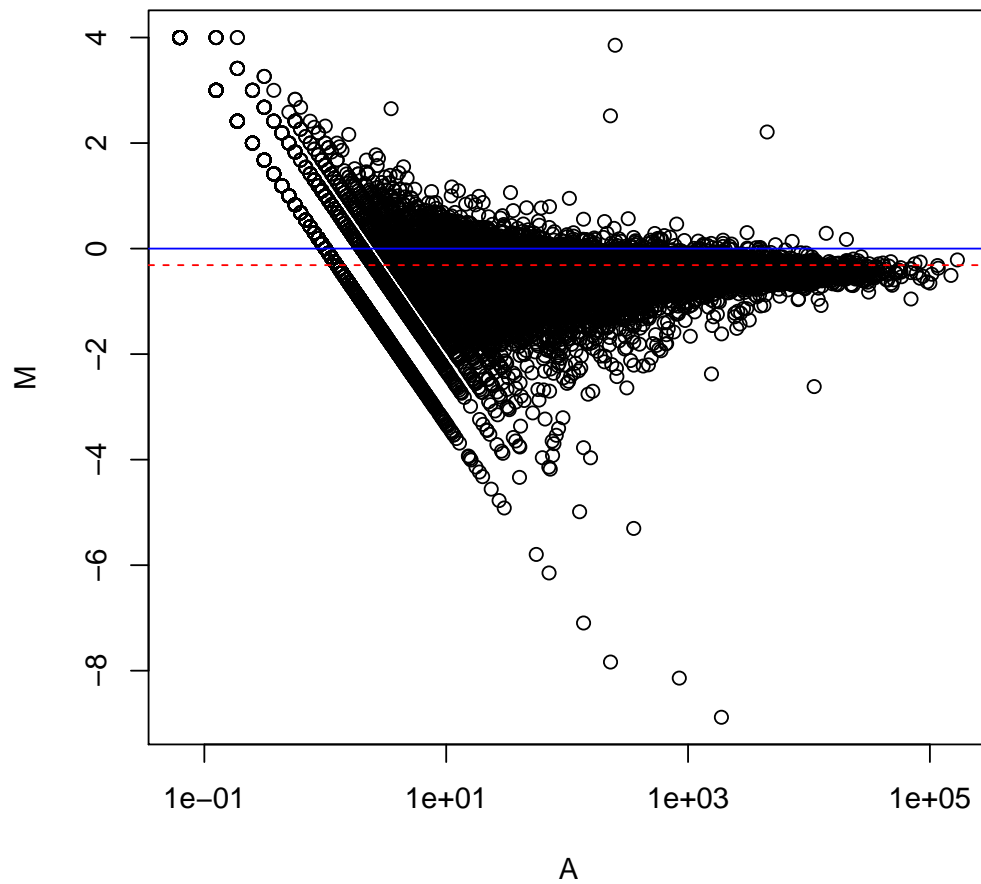


Figure 5.5: **MA-plot illustrating TMM normalisation** – An MA-plot generated for one of the WT mRNA samples. The y-axis is the log ratio of one of the WT samples and the average across all samples (M) and the x-axis are the average gene counts across all samples (A). The solid blue line is $M=0$ and the dashed red line is the line fitted by TMM.

gene variances than Tshrink. This also indicates that the gene variances observed in the Borchelt samples are informative for the WT and NCN mice.

The number of DE genes found by each method can be observed in Table 5.3. For each of the four methods, the number of DE genes are listed for a 0.05 cut-off on the unadjusted p-values, Benjamini-Hochberg adjusted p-values (Benjamini and Hochberg, 1995) and Bonferroni adjusted p-values (Bonferroni, 1936). The Benjamini-Hochberg and Bonferroni corrections adjust the p-values to account for the thousands of tests that have been performed and hence control the false discovery rate and family-wise error rate respectively. Bonferroni is more stringent than Benjamini-Hochberg.

	Unadjusted	Benjamini-Hochberg	Bonferroni
T	1027	0	0
Tcommon	1169	252	74
Tshrink	1352	7	1
Tshrink+	1446	64	8

Table 5.3: **Table of results from mRNA DE** – The number of DE genes are reported from the comparison of WT and NCN mice by four DE methods for an arbitrary 0.05 p-value cut-off; a two sample t-test (T), a two sample t-test using a fitted common variance (Tcommon), a moderated t-test (Tshrink) and a moderated t-test using variance information from the Borchelt mice (Tshrink+). The number of DE genes are also reported after adjusting for multiple testing using the Benjamini-Hochberg and Bonferroni methods.

Table 5.3 demonstrates the impact that moderation can have on statistical power. No genes will be statistically identified as DE if a two sample t-test (T) is used and Benjamini-Hochberg or the more stringent Bonferroni correction are utilised to account for multiple comparisons. As Tshrink+ uses more of the fitted common variance than Tshrink it identifies more DE genes, with Tcommon (the most moderated) identifying the most DE genes. Of course, while moderation improves the power of a test it could also lead to higher false positives if any of the underlying assumptions in the moderation approach are compromised.

Figure 5.6b suggests that Tcommon may be failing to model the gene variances appropriately. If there is no signal in the data the corresponding p-value histogram should be flat. If there is signal, the p-value histogram should exhibit some form of peak on the left of the plot. This is observed for T, Tshrink and Tshrink+ in Figure 5.6a, 5.6c and 5.6d. A p-value histogram that has some form of peak, or slant, on the right im-

	Unadjusted	FDR	Bonferroni
T	20	0	0
Tcommon	24	7	3
Tshrink	17	0	0
Tshrink+	21	0	0

Table 5.4: **Table of results from miRNA DE** – The number of DE miRNA are reported from the comparison of WT and NCN mice by four DE methods for an arbitrary 0.05 p-value cut-off; a two sample t-test (T), a two sample t-test using a fitted common variance (Tcommon), a moderated t-test (Tshrink) and a moderated t-test using variance information from the Borchelt mice (Tshrink+). The number of DE miRNA are also reported after adjusting for multiple testing using the Benjamini-Hochberg and Bonferroni methods.

plies an issue with the test. This is seen in the p-value histogram of Tcommon, Figure 5.6b. This is most likely caused by still having heterogeneous gene variances even after conditioning on the average expression of a gene. Larger than expected gene variances will disproportionately inflate the common variance estimate. Hence, if the common variance estimate is solely used to model the gene variances, the majority of the modelled gene variances will be too large creating the peak on the right side of the p-value histogram. Conversely, the common variance estimate will most likely underestimate the gene variances for the genes with large variances. This should generate scepticism over the peak on the left of the histogram.

When testing for DE miRNA Tshrink fits a λ of 0.17 and Tshrink+ fits a λ of 0.58. As with the mRNA data this indicates that the miRNA variances observed in the Borchelt samples are informative for the WT and NCN mice. Table 5.4 lists the number of DE miRNA found by T, Tcommon, Tshrink and Tshrink+. Unfortunately T, Tshrink and Tshrink+ are unable to identify any statistically DE miRNA after correcting for multiple comparisons. This may imply that there is no biological signal in the data or we do not have enough samples to identify statistically significant differences.

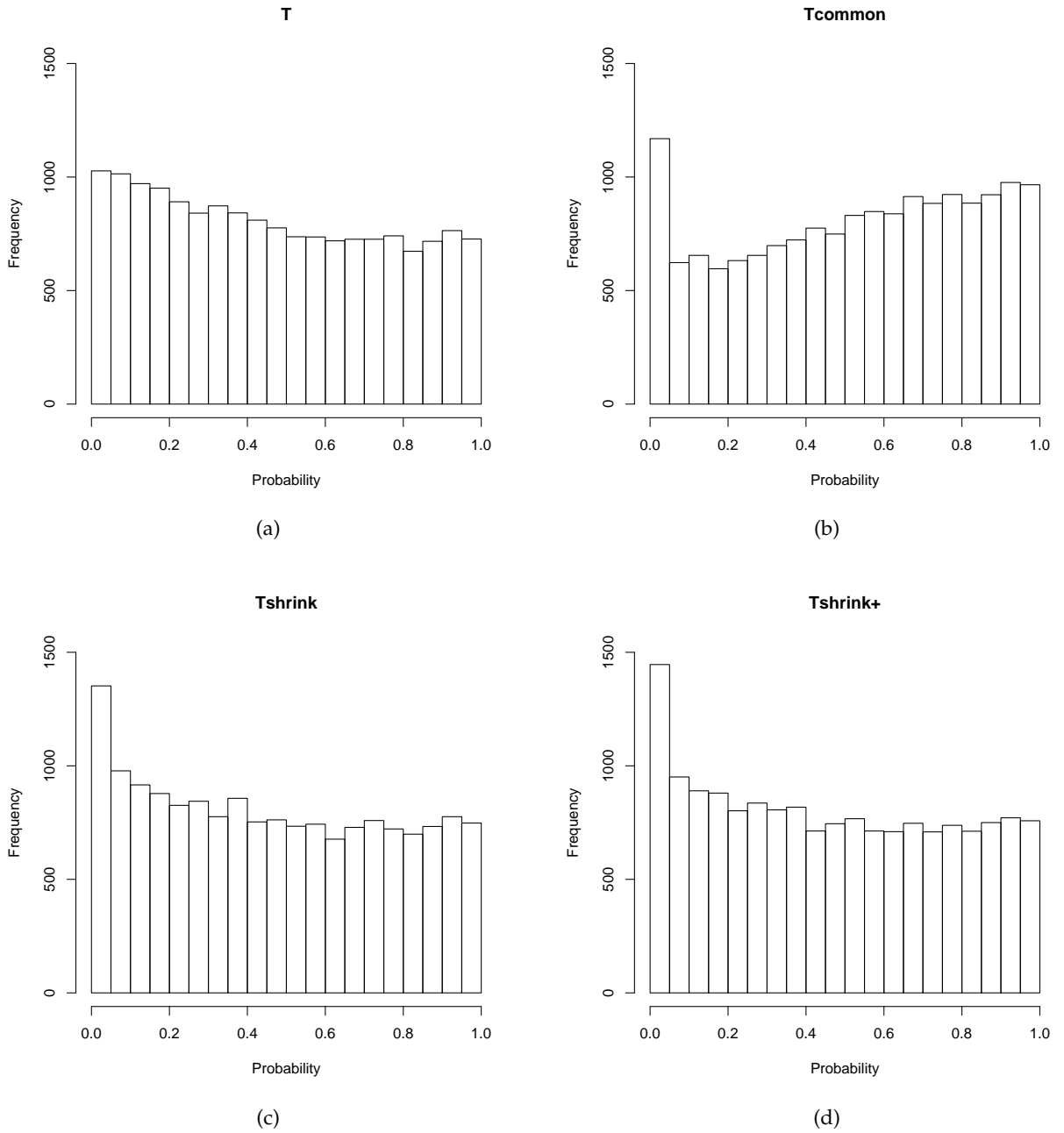


Figure 5.6: **P-value histograms for four different methods** – P-value histograms were generated from the comparison of WT and NCN mice using a two sample t-test (T), a two sample t-test using a fitted common variance (Tcommon), a moderated t-test (Tshrink) and a moderated t-test using variance information from the Borchelt mice (Tshrink+).

5.6 FUNCTIONAL, NETWORK & PATHWAY ANALYSIS

5.6.1 *Pathway analysis*

Pathway analysis is often performed to improve the interpretability or power of differential expression results. We applied Goseq (Young *et al.*, 2010) to test if the genes in any KEGG (Kanehisa *et al.*, 2008) pathways were over-represented in our list of DE genes. Tshrink+ was used to generate a list of DE genes where we called a gene DE if it had a Benjamini-Hochberg corrected p-value of less than 0.05.

Goseq performs an over-representation test while accounting for the length bias in the DE results. Longer genes on average have more reads and hence on average more power for detecting changes in expression. When performing a competitive test, ie an over-representation test, these tests can artificially favour pathways that contain longer genes. Testing if a pathway is over-represented in a list of genes is similar to testing a two-by-two table for independence. This two-by-two table would summarise how many genes are DE or not DE and in a pathway or not in a pathway.

Pathway	p-value
Protein digestion and absorption	0.0018
Other types of O-glycan biosynthesis	0.0054
Caffeine metabolism	0.0139
Gastric acid secretion	0.0297
Sulfur metabolism	0.0300
Phosphatidylinositol signaling system	0.0425
ECM-receptor interaction	0.0481
GnRH signaling pathway	0.0491

Table 5.5: **Table of results Goseq** – Goseq was used to perform a pathway over-representation test on the list DE genes from Tshrink+. Listed are the pathways that had a p-value less than 0.05.

The results from the over-representation test can be observed in Table 5.5. The pathways listed in the table are consistent with what is known about the Notch signalling pathway. Notch is known to regulate cellular metabolism and the switch between oxidative phosphorylation and glycolysis in tumor cells (Landor *et al.*, 2011). While nothing on this list demands further investigation, it does provide confidence that the experiment has signal.

5.6.2 Integration of miRNA and mRNA data

As a further exploratory analysis, we are interested in identifying sets of genes that share a similar functional outcome and are potentially being regulated by a miRNA. Our proposed approach, pMimCor, has been applied to do just this. In the following we will discuss some of the output from the application of pMimCor.

As discussed in Chapter 4, five of the top seven miRNA ranked by pMimCor for a WT and NCN comparison have been associated with neurodegeneration in the literature. While the search performed was not very specific, this does provide some optimism that there may be some signal in our data. In addition to this, it is hoped that some of the miRNA we are identifying are playing an active regulatory role in response to loss of Notch2 function. We identified 10 and 33 mir-pathways whose miRNA appeared to be over and under expressed respectively in the NCN samples. These mir-pathways were chosen with an arbitrary p-value cut-off of 0.01. Tables 5.6 and 5.6 list the top ten mir-pathways for each direction.

Table 5.6 lists the top ten mir-pathways whose miRNA appeared to be over expressed in the NCN samples. In general it would not be surprising to see the genes in these pathways down-regulated when Notch2 has been knocked out. Notch signalling generally promotes growth and inhibits differentiation and as such is implicated in many cancers (Wang *et al.*, 2006; Santagata *et al.*, 2004; Balint *et al.*, 2005). Notch signalling has also been associated with amyotrophic lateral sclerosis (Praline *et al.*, 2010) and is known to regulate JAK-STAT (Kamakura *et al.*, 2004) and MAPK (Ghai *et al.*, 2010).

Table 5.7 lists the top ten mir-pathways whose miRNA appeared to be over expressed in the NCN samples. Notch has been associated with immune response with the Notch signalling pathway being implicated in the activation and proliferation of T cells and the generation of MZ B cell precursors and mature MZ B cells (Yuan *et al.*, 2010). The vascular endothelial growth factor (VEGF) signalling pathway and the ErbB signalling pathway, whose founding member is the epidermal growth factor receptor, are both pathways that are associated with growth. As Notch signalling generally promotes growth, this may provide evidence that these pathways have been activated to promote growth in the absence of Notch. In addition to this, VEGF is thought to be a possible

miRNA	pMimCor p-value	Tshrink+ p-value	Number of targets	Pathway
mmu-mir-760	0.0025	0.053	2	Pancreatic cancer
mmu-mir-760	0.0025	0.053	2	Chronic myeloid leukemia
mmu-mir-760	0.0029	0.053	2	Amyotrophic lateral sclerosis (ALS)
mmu-mir-760	0.0045	0.053	2	Jak-STAT signalling pathway
mmu-mir-760	0.0046	0.053	8	Pathways in cancer
mmu-mir-130a	0.0058	0.044	2	MAPK signalling pathway
mmu-mir-3475	0.0064	0.062	2	Oocyte meiosis
mmu-mir-29b	0.0075	0.1	3	Endometrial cancer
mmu-mir-23b	0.0083	0.056	2	Inositol phosphate metabolism
mmu-mir-760	0.0085	0.053	2	Acute myeloid leukemia

Table 5.6: **Table of results from pMimCor for over expressed miRNA.** – Listed are the top ten mir-pathways found using pMimCor whose miRNA were over expressed in the NCN samples. The corresponding miRNA and KEGG pathway are listed, as well as, the p-value from pMimCor, the miRNA Tshrink+ p-value and the number of genes in the mir-pathway.

treatment for amyotrophic lateral sclerosis (Rothstein, 2009). As such this may also provide evidence that this pathway has been activated in response to the loss of Notch function and its down-stream signals.

Our methodology, pMimCor, can only be used as an exploratory tool from a biological sense. As it uses correlations it can only test for associations between miRNA and their target genes. Hence, it becomes difficult to know if the expression of the genes in these pathways have been altered directly by known Notch signalling mechanisms or if the identified miRNA are playing an active role in their regulation. Regardless, pMimCor has potentially highlighted the mir-gene relationships that make functional sense and warrant further investigation. Many of the gene pathways and associated miRNAs having been associated with Notch signalling or neurodegeneration.

5.7 CONCLUSIONS AND FURTHER DISCUSSION

In this chapter we used a matched miRNA-Seq and mRNA-Seq experiment to illustrate the three statistical methodologies; exClust, Tshrink+ and pMimCor; that were proposed in this thesis. As this particular experiment was exploratory in nature, there does not exist any related gold standard experiments to use for comparison. However, the behaviour of the three proposed methods when applied to this experiment ap-

miRNA	pMimCor p-value	Tshrink+ p-value	Number of targets	Pathway
mmu-mir-200a	0.0014	0.019	2	Fc epsilon RI signalling pathway
mmu-mir-200a	0.0014	0.019	3	GnRH signalling pathway
mmu-mir-200a	0.0017	0.019	4	ErbB signalling pathway
mmu-mir-200a	0.0017	0.019	3	Gap junction
mmu-mir-96	0.0018	0.029	3	VEGF signalling pathway
mmu-mir-96	0.0018	0.029	4	Natural killer cell mediated cytotoxicity
mmu-mir-96	0.0018	0.029	4	T cell receptor signalling pathway
mmu-mir-96	0.0018	0.029	3	B cell receptor signalling pathway
mmu-mir-96	0.0018	0.029	7	Long-term potentiation

Table 5.7: **Table of results from pMimCor for under expressed miRNA.** – Listed are the top ten mir-pathways found using pMimCor whose miRNA were under expressed in the NCN samples. The corresponding miRNA and Kegg pathway are listed, as well as, the p-value from pMimCor, the miRNA Tshrink+ p-value and the number of genes in the mir-pathway.

peared generally consistent with the behaviours observed in the evaluations in their respective chapters. When applied to this experiment exClust included more than twice as many reads as UI, Tshrink+ found more DE genes than Tshrink and pMimCor produced results that were both interpretable and appear consistent with current knowledge of Notch signalling. In addition to this, this study indicates that there is biologically relevant signal in this experiment that warrants further attention.

CONCLUSION

Developing methodologies that concurrently identify features that are statistically and biologically relevant is a challenge faced in the fields associated with high-throughput biotechnologies. This challenge is exacerbated by the large number of genes measured by RNA-Seq along with small sample sizes, commonly referred to as a large-p small-n problem. This thesis contributed to the development of such methodologies with respect to RNA-Seq experiments, providing practical and biologically interpretable results in a low replication setting. The methodological solutions proposed further bridge the gap between the notions of statistical significance and biological significance. This is achieved by integrating data from several public repositories and annotation databases related to the experimental data of interest.

This thesis was motivated by a collaboration with the Lin lab of Cornell University and can be separated into three distinct statistical problems that arise when processing and analysing RNA-Seq data. These include methods for the summarisation and differential expression steps of a typical analysis, as well as, a framework for integrating small sample miRNA-Seq and mRNA-Seq data. While the amount of publicly available RNA-Seq data has been increasing dramatically, there is a lack of gold standard datasets making evaluation of methodology quite difficult. As a result many of the evaluation approaches contained in this thesis are novel in their own right.

The concept of using experimental data to customise gene annotations was the foundation of developing a novel summarisation method for gene counts in RNA-Seq data. Estimating changes in the overall transcription of a gene with high throughput sequencing is not straightforward and was addressed in Chapter 2. For example, changes in overall gene transcription can easily be confounded with changes in exon usage, which alter the lengths of transcripts produced by a gene. We proposed measuring the expression of data-specific constitutive exons – exons which are consistently conserved after splicing in the given dataset – as an unbiased estimation of the overall transcription of a gene. This is in contrast to only measuring the expression of exons which are annotated as constitutive in public databases.

We demonstrated that data-specific constitutive exons can be estimated using our clustering-based method, exClust. When the exons from exClust were used to summarise the reads from two real datasets, they summarised more than three times as

many reads as the standard UI method (Bullard *et al.*, 2010) and also improved concordance with qRT-PCR data. Our method is shown to produce robust estimates of overall gene transcription when compared with other methods. The results from this chapter have been published in (Patrick *et al.*, 2013a).

Basic molecular biology experiments focusing on identifying differentially expressed genes are still commonly faced with small sample sizes. Estimating robust variances of gene expression estimates, in order to call differential expression, is challenging in the presence of low replication. We demonstrated in Chapter 3 that the wealth of publicly available gene expression data generated over the past ten to fifteen years can be leveraged to make improvements in the variance estimation of genes when combined with our novel moderation methodology, Tshrink+. The general concept of data integration is a popular concept in theory but challenging in its implementation. Furthermore, the selection of appropriate external information is a challenge in itself. Our proposed moderation methodology provides both a means of integrating and assessing the utility of external information.

Using biological data we demonstrated that incorporating additional external information can improve the modelling of the common variance and hence the calling of differentially expressed genes. These sources of additional information include gene length and gene-wise sample variances from other RNA-Seq and microarray datasets, of both related and seemingly unrelated tissue types. Demonstrating that seemingly unrelated tissue types still can contribute information exemplifies the value of having an approach for assessing the appropriateness of external information. It also opens up the opportunity of utilising wider sources of external information in more situations. The results were promising, with our differential expression test, Tshrink+, performing favourably when compared to existing methods such as DESeq (Anders and Huber, 2010) and edgeR (Robinson *et al.*, 2010) when considering both gene ranking and sensitivity. These results have been published in (Patrick *et al.*, 2013b).

Extracting information from multiple data sources is a statistically challenging problem. In Chapter 4 a framework for integrating small sample miRNA-Seq and mRNA-Seq data was presented. The mouse genome has about 24000 mRNA families (genes) and 1000 annotated microRNA (miRNA) making any analysis and interpretation of the

interaction between these two classes of RNA molecules a daunting task. We developed a supervised framework for integrating various databases of prior knowledge with experimental data to build meaningful and interpretable models of miRNA-mRNA regulation. This framework was used to develop three methodologies; cMimDE, pMimDE and pMimCor.

The pMimDE and pMimCor methods provide novel approaches for the joint estimation and ranking of interesting miRNAs, their target genes and their associated biological functions. It was demonstrated that the KEGG pathway database (Kanehisa *et al.*, 2012) is enriched for genes that are predicted to be targets of miRNAs by TargetScan (Lewis *et al.*, 2005). By using these databases to pre-define potential regulatory relationships we showed that they identify more differentially expressed relationships than expected by chance in a matched miRNA- and mRNA-Seq experiment. These differentially expressed relationships contain many miRNA that have been associated with the experimental conditions in other literature. We conclude that not only does pMimCor identify more interesting miRNA than other approaches but the miRNA-mRNA relationships are inherently more interpretable. Various results from this chapter were published in (Yang *et al.*, 2013) and have been presented at WNAR 2013 and AustMS 2013.

We concluded the thesis by implementing our proposed methodologies on our own experiment. In collaboration with the Lin lab of Cornell University we designed an experiment to assess how miRNA and mRNA may interact in a conditional knockout mouse model. Due to the complications with breeding these mice, our experiment suffers from low replication and is consistent with the primary motivation for this thesis. When applied to this experiment, the behaviours of our proposed methodologies were concordant with those observed in their respective chapters. In addition to this, using these methods we were able to identify interpretable signal in the experiment that is functionally consistent with domain knowledge. In summary, despite the presence of low replication, our proposed statistical methodologies are able to identify biologically relevant signal within a RNA-Seq experiment.



ADDITIONAL INFORMATION FOR CHAPTER 2

A.1 DETECTION OF DIFFERENTIAL ALTERNATIVE SPLICING

For the purpose of evaluating our method it would be useful to know if the relative abundances of gene isoforms has changed in two conditions. It is in this situation that comparing the overall expression of a gene in two conditions will be confounded by the changes in lengths of the isoforms. In comparisons across samples and/or conditions, it is standard to test for changes between the samples or conditions in total gene expression; that is, to test for “differential expression” of each gene. When we consider alternative splicing and the multiple isoforms this can produce, it is also of interest to test a gene for changes between the samples or conditions in the relative abundances of its isoforms. We will adopt the terminology used in (Xing *et al.*, 2008) and call such tests, tests for differential alternative splicing. One such test is the Differential Alternative uSage Index (DASI) described in (Richard *et al.*, 2010) which equates to a Fisher’s exact test. DASI takes as input the exon counts for a gene and tests for independence between condition and relative exon expression and is appropriate for Poisson distributed data.

A.2 ADDITIONAL FIGURES AND TABLES

log fold changes	qRT-PCR	Union	UI	exClust	Cufflinks
ENSG00000103769	-0.26	0.17	-0.12	-0.26	-0.26
ENSG00000076662	-6.34	-1.77	-1.77	-2.04	-2.78

Table A.1: **Log fold changes for different summarisation methods** – For two genes, ENSG00000103769 and ENSG00000076662, log fold changes are reported for four summarisation methods and qRT-PCR from the MAQC dataset.

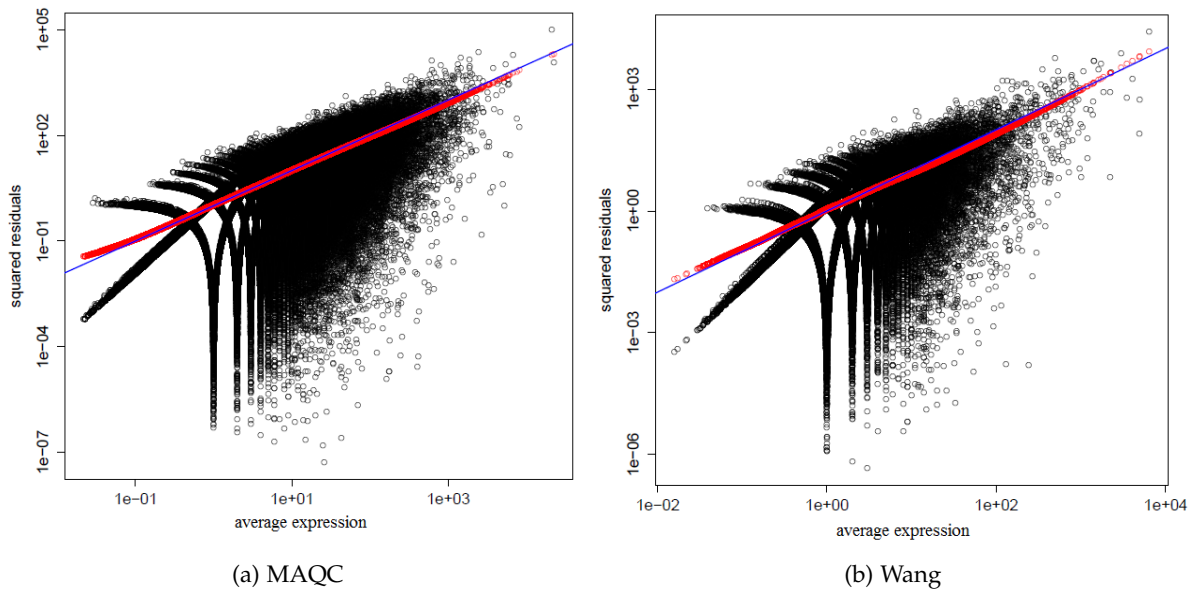


Figure A.1: **Verification of the Poisson assumption for the MAQC and Wang data** – The squared standardised residuals are plotted against the average for each gene in the a) MAQC and b) Wang datasets. The blue line is the $y = x$ line. The red circles correspond to the fitted points found using local smoothing.

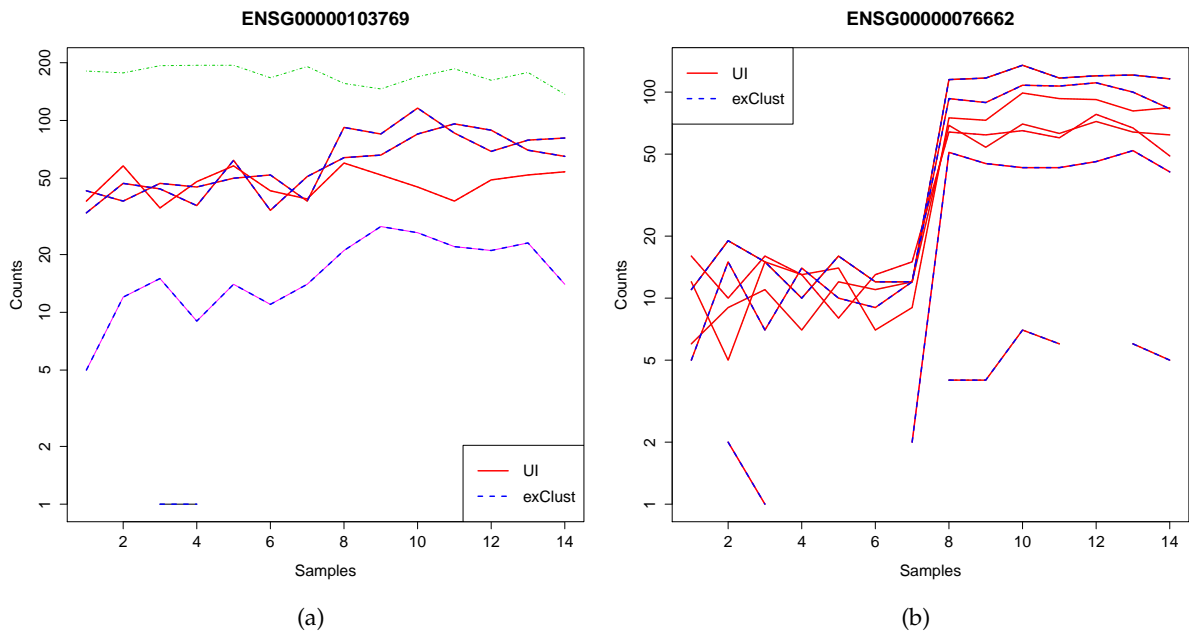


Figure A.2: **Gene exon counts** – For two genes, a) ENSG00000103769 and b) ENSG00000076662, the exon counts for each sample are plotted. The first seven samples are brain and the second seven are UHR. A line is drawn between the points to make the behaviour of each exon easier to follow. Highlighted in red are the UI exons and dashed blue are the exClust exons.

ADDITIONAL INFORMATION FOR CHAPTER 3

B.1 NORMALISATION FOR BOTTOMLY DATASET

B.1.1 *GC content*

In order to model the disparate library sizes observed in the data and the biases of PCR amplification driven by the GC content of the sequences, a cyclic robust linear model was used. Using the first sample in the dataset as a reference, M values were calculated for each gene in the remaining samples and a straight line was fitted through the M-value vs GC-content space. The M-values were then normalised to this line.

B.1.2 *Other Technical effects*

In order to normalise out any remaining technical effects the following approach is taken. In principal this approach is similar to that described for normalising out the GC content bias. The first six steps essentially estimate those sets of genes that may be influenced by some technical component of the analysis (ie high GC content genes or low GC content genes). This is done in such a way as to hopefully avoid selecting genes associated with the biological conditions of interest. The last step corrects for this technical effect.

- The pooled correlations of the counts for house-keeping genes are calculated.
- Hierarchical clustering is performed using the distance $1 - \text{correlation}$.
- Tree is cut into k groups (we arbitrarily choose k equal to four).
- For all genes, within condition residuals of the log counts are calculated.
- LDA trained on house-keeping genes using groups from clustering, this is then used to classify all genes into groups.
- The posterior probability of a gene belonging to a particular group is converted to a quantile value of the normal distribution (Q-values).

- Genes are then normalised to the loess curve that is fitted through the pair-wise MQ plots. (M-values vs Q-values) using the first sample as a reference.

Figure B.1 illustrates the performance of our normalisation method (loess) compare to using RUV (Gagnon-Bartsch and Speed, 2012), SVA (Leek *et al.*, 2012) and no normalisation. Results from an experiment using Affymetrix arrays and illumina arrays are used as truth. While not conclusive our method has the highest area under the ROC for both array types. Figure B.2 plot the log variance of the within sample gene ranks for the normalisation approaches. All normalisation approaches reduce the variance of the ranks.

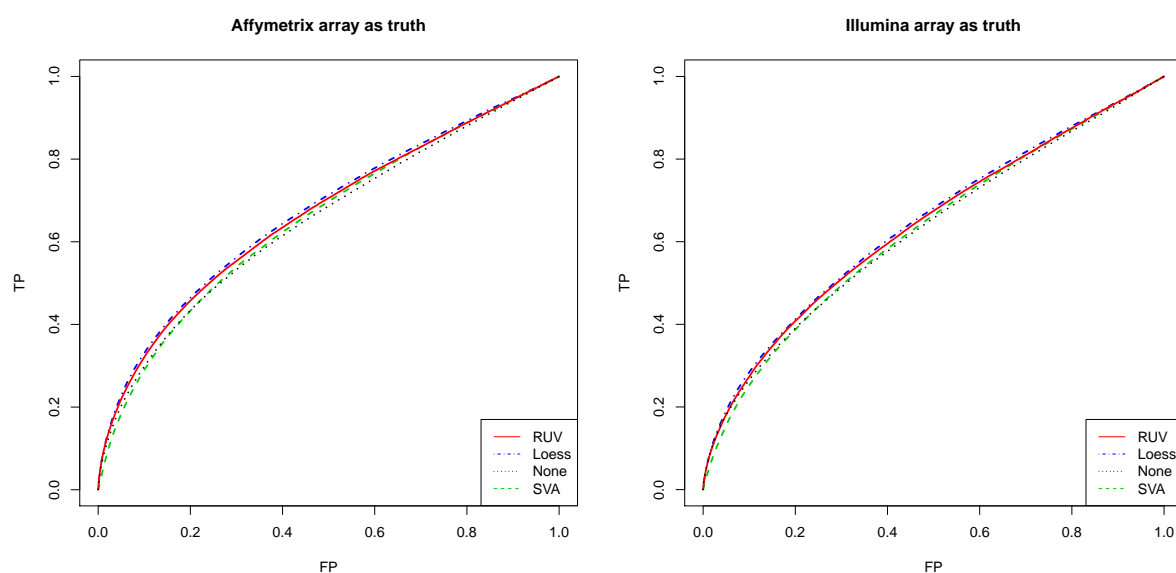


Figure B.1: **ROC of normalised data using arrays as truth** – Average TPR and FPR are calculated from 100 random four B6 vs four D2 mouse striatum comparisons for four normalisation methods using results from an a) Affymetrix and b) Illumina array as truth. These are plotted against each other to form ROC curves. For any given FPR a method with a larger TPR is deemed to have ranked the genes better.

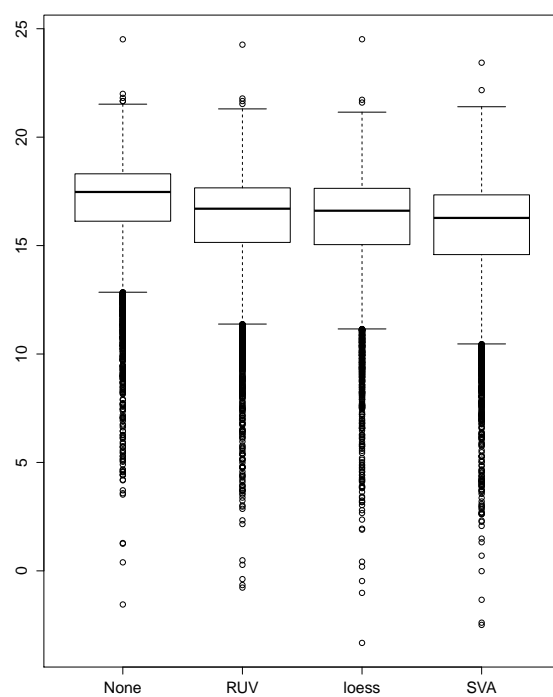


Figure B.2: **Boxplots of log variances** –Boxplots of the log variance of the within sample gene ranks for four normalisation methods. All normalisation methods on average reduce the variance of the ranking of the genes.

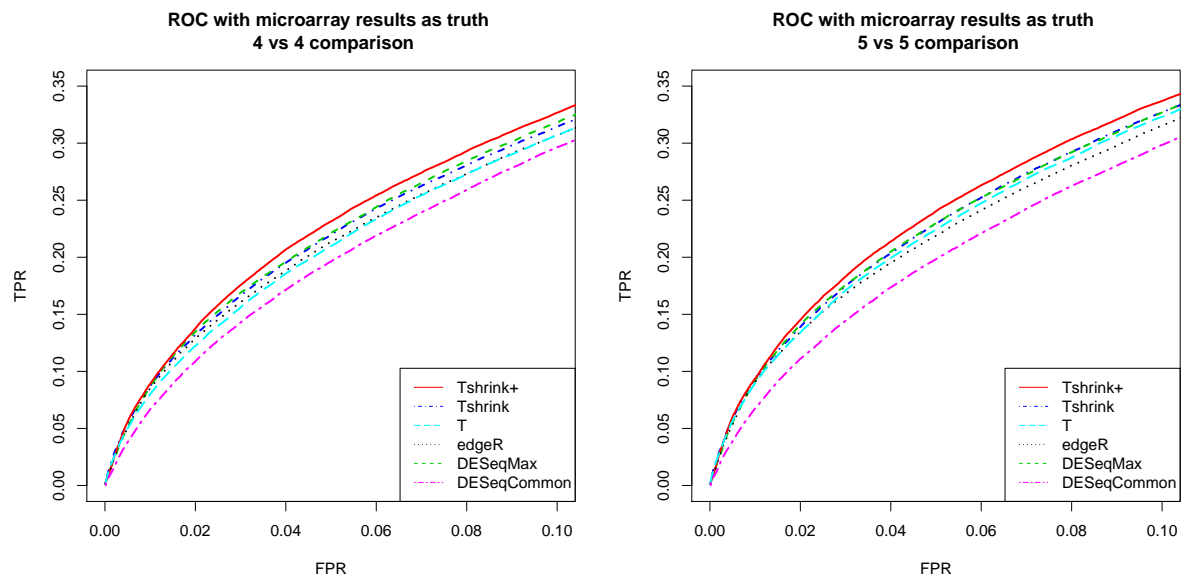


Figure B.3: **Residual plots for the MAQC and Wang data** – Average TPR and FPR are calculated from a) 100 random four B6 vs four D2 mouse striatum comparisons and b) 100 random five vs five D2 mouse striatum comparisons for six DE methods. These are calculated using results from an Affymetrix array experiment as truth. The TPR and FPR are plotted against each other to form ROC curves and displayed in the region for FPR less than 0.1 as this is most relevant for calling DE. For any given FPR a method with a larger TPR is deemed to have ranked the genes better. T and Tshrink both improve in performance relative to edgeR and DESeq when moving from the four vs four comparison to the five vs five comparison.

ADDITIONAL INFORMATION FOR CHAPTER 4

C.1 ADDITIONAL FIGURES

The ROC curves generated from the literature search can be seen in Figure C.1. While the full ROC plots may not be assuring, it is important to remember that our measure of truth is approximate in nature. However, all three integration methods (cMimDE, pMimDE and pMimCor) perform better for small FPR when compared to just using the miRNA data. The results from pMimCor are better than random for most of the curve, the area under the curve being 0.62.

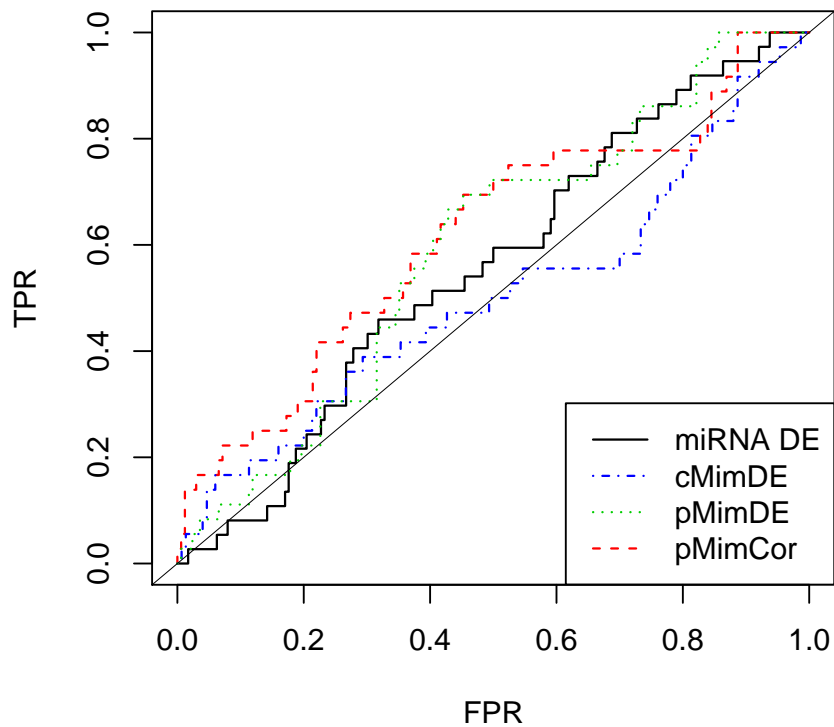


Figure C.1: **ROC plot from PubMed search** – ROC curves are plotted for various methods as described in *Literature search strategy*. The search term used is neurodegeneration. True Positive Rates (TPR) are plotted against False Positive Rates (FPR) for four methods, miRNA DE (black), cMimDE (blue), pMimDE (green) and pMimCor (red).

BIBLIOGRAPHY

- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106. (Cited on pages 9, 33, 35, 38, and 93.)
- Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika*, **35**(3/4), pp. 246–254. (Cited on page 20.)
- Balint, K., Xiao, M., Pinnix, C. C., Soma, A., Veres, I., Juhasz, I., Brown, E. J., Capobianco, A. J., Herlyn, M., and Liu, Z.-J. (2005). Activation of notch1 signaling is required for beta-catenin-mediated human primary melanoma progression. *J Clin Invest*, **115**(11), 3166–3176. (Cited on page 88.)
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muetter, R. N., Holko, M., Ayanbule, O., Yefanov, A., and Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*, **39**(Database issue), D1005–D1010. (Cited on page 50.)
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, pages 289–300. (Cited on pages 64 and 84.)
- Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, **40**(10), e72. (Cited on page 8.)
- Bishop, Y. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics*, **27**(3), 545–562. (Cited on page 20.)
- Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291–336. (Cited on page 14.)
- Blencowe, B. J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci*, **25**(3), 106–110. (Cited on page 14.)
- Bona, F. D., Ossowski, S., Schneeberger, K., and Rättsch, G. (2008). Optimal spliced alignments of short sequence reads. *Bioinformatics*, **24**(16), i174–i180. (Cited on pages 8 and 14.)
- Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Libreria internazionale Seeber. (Cited on page 84.)
- Bottomly, D., Walter, N. A. R., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*, **6**(3), e17820. (Cited on pages 41, 42, and 50.)
- Bryant, D. W., Shen, R., Priest, H. D., Wong, W.-K., and Mockler, T. C. (2010). Supersplat—spliced RNA-seq alignment. *Bioinformatics*, **26**(12), 1500–1505. (Cited on pages 8 and 14.)

- Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**(1), 94+. (Cited on pages 8, 16, 23, 25, 27, 37, 65, and 93.)
- Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K.-Y., Morley, M., and Spielman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, **33**(3), 422–425. (Cited on page 38.)
- Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., Bult, C. J., Agarwala, R., Cherry, J. L., DiCuccio, M., Hlavina, W., Kapustin, Y., Meric, P., Maglott, D., Birtle, Z., Marques, A. C., Graves, T., Zhou, S., Teague, B., Potamosis, K., Churas, C., Place, M., Herschleb, J., Runnheim, R., Forrest, D., Amos-Landgraf, J., Schwartz, D. C., Cheng, Z., Lindblad-Toh, K., Eichler, E. E., Ponting, C. P., and M. G. S. C. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol*, **7**(5), e1000112. (Cited on pages 65 and 76.)
- Cleveland, W., Grosse, E., and Shyu, W. (1992). Local regression models. *Statistical models in S*, pages 309–376. (Cited on page 41.)
- Cox, B., Kotlyar, M., Evangelou, A. I., Ignatchenko, V., Ignatchenko, A., Whiteley, K., Jurisica, I., Adamson, S. L., Rossant, J., and Kislinger, T. (2009). Comparative systems biology of human and mouse as a tool to guide the modeling of human placental pathology. *Mol Syst Biol*, **5**, 279. (Cited on page 15.)
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563. (Cited on page 3.)
- Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4**(4), 210. (Cited on page 38.)
- Fisher, R. (1925). *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd. (Cited on page 56.)
- Frazeo, A. C., Langmead, B., and Leek, J. T. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449. (Cited on pages 41 and 50.)
- Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**(3), 539–552. (Cited on pages 41 and 100.)
- Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**(3), 307–315. (Cited on page 41.)
- Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbil, J. O., Emanuelson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? history and updated definition. *Genome Res*, **17**(6), 669–681. (Cited on page 12.)
- Ghai, K., Zelinka, C., and Fischer, A. J. (2010). Notch signaling influences neuroprotective and proliferative properties of mature müller glia. *J Neurosci*, **30**(8), 3101–3112. (Cited on page 88.)

- Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* (Cited on page 23.)
- Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**(Database issue), D154–D158. (Cited on page 53.)
- Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422. (Cited on page 9.)
- Havelange, V., Stauffer, N., Heaphy, C. C. E., Volinia, S., Andreeff, M., Marcucci, G., Croce, C. M., and Garzon, R. (2011). Functional implications of microRNAs in acute myeloid leukemia by integrating microRNA and messenger RNA expression profiling. *Cancer*, **117**(20), 4696–4706. (Cited on page 54.)
- Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A. R., Purcell, S. M., Sklar, P., , W. T. C.-C. C., Owen, M. J., O'Donovan, M. C., and Craddock, N. (2009). Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet*, **85**(1), 13–24. (Cited on page 72.)
- Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Res.* **37**(suppl 1), D690–D697. (Cited on pages 24 and 66.)
- Jankowsky, J. L., Fadale, D. J., Anderson, J., Xu, G. M., Gonzales, V., Jenkins, N. A., Copeland, N. G., Lee, M. K., Younkin, L. H., Wagner, S. L., Younkin, S. G., and Borchelt, D. R. (2004). Mutant presenilins specifically elevate the levels of the 42 residue beta-amyloid peptide in vivo: evidence for augmentation of a 42-specific gamma secretase. *Hum Mol Genet*, **13**(2), 159–170. (Cited on page 74.)
- Jayaswal, V., Lutherborrow, M., Ma, D. D. F., and Hwa Yang, Y. (2009). Identification of microRNAs with regulatory potential using a matched microRNA-mRNA time-course data. *Nucleic Acids Res.* **37**(8), e60. (Cited on page 54.)
- Jayaswal, V., Lutherborrow, M., Ma, D. D. F., and Yang, Y. H. (2011). Identification of microRNA-mRNA modules using microarray data. *BMC Genomics*, **12**, 138. (Cited on page 54.)
- Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**(8), 1026–1032. (Cited on pages 8 and 15.)
- Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**(5653), 2141–2144. (Cited on page 14.)

- Kamakura, S., Oishi, K., Yoshimatsu, T., Nakafuku, M., Masuyama, N., and Gotoh, Y. (2004). Hes binding to STAT3 mediates crosstalk between Notch and JAK-STAT signalling. *Nat Cell Biol*, **6**(6), 547–554. (Cited on page 88.)
- Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, **36**(Database issue), D480–D484. (Cited on pages 65 and 87.)
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, **40**(Database issue), D109–D114. (Cited on pages 15 and 94.)
- Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunção, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**(7364), 289–294. (Cited on pages 42 and 46.)
- Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, **35**(1), 125–131. (Cited on page 14.)
- Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat Genet*, **37**(5), 495–500. (Cited on page 54.)
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brothier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M.,

- Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowski, J., and Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921. (Cited on page 14.)
- Landor, S. K.-J., Mutvei, A. P., Mamaeva, V., Jin, S., Busk, M., Borra, R., Grönroos, T. J., Kronqvist, P., Lendahl, U., and Sahlgren, C. M. (2011). Hypo- and hyperactivated Notch signaling induce a glycolytic switch through distinct mechanisms. *Proc Natl Acad Sci USA*, **108**(46), 18814–18819. (Cited on page 87.)
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**(3), R25. (Cited on pages 8, 24, 41, 65, and 76.)
- Latchman, D. S. (1996). Activation and repression of gene expression by POU family transcription factors. *Philos Trans R Soc Lond B Biol Sci*, **351**(1339), 511–515. (Cited on page 12.)
- Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**(6), 882–883. (Cited on pages 41 and 100.)
- Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**(1), 15–20. (Cited on pages 53 and 94.)
- Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**(4), 493–500. (Cited on pages 8, 9, and 15.)
- Li, J. and Tseng, G. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann Appl Stat*, **5**(2A), 994–1019. (Cited on page 55.)
- Li, X., Chen, J., Hu, X., Huang, Y., Li, Z., Zhou, L., Tian, Z., Ma, H., Wu, Z., Chen, M., Han, Z., Peng, Z., Zhao, X., Liang, C., Wang, Y., Sun, L., Chen, J., Zhao, J., Jiang, B., Yang, H., Gui, Y., Cai, Z., and Zhang, X. (2011). Comparative mRNA and microRNA

- expression profiling of three genitourinary cancers reveals common hallmarks and cancer-specific molecular events. *PLoS One*, **6**(7), e22570. (Cited on page 54.)
- Loader, C. (2010). *locfit: Local Regression, Likelihood and Density Estimation*. R package version 1.5-6. (Cited on page 39.)
- Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet*, **32**, 279–305. (Cited on page 14.)
- M. A. Q. C. Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**(9), 1151–1161. (Cited on pages 18 and 23.)
- Maniatis, T. and Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**(6894), 236–243. (Cited on page 13.)
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**(9), 1509–1517. (Cited on page 37.)
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1), 10–12. (Cited on pages 65 and 77.)
- Mo, Y.-Y. (2012). MicroRNA regulatory networks and human disease. *Cell Mol Life Sci*, **69**(21), 3529–3531. (Cited on page 53.)
- Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*, **29**(13), 2850–2859. (Cited on page 14.)
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**(7), 621–628. (Cited on page 5.)
- Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol*, **6**, Article9. (Cited on pages 35, 39, and 50.)
- Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, **4**, 14. (Cited on page 17.)
- Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From rna-seq reads to differential expression results. *Genome Biol*, **11**(12), 220. (Cited on page 6.)
- Pachter, L. (2011). Models for transcript quantification from RNA-Seq. *Arxiv preprint arXiv:1104.3889*. (Cited on pages 35 and 38.)
- Patrick, E., Buckley, M., and Yang, Y. H. (2013a). Estimation of data-specific constitutive exons with RNA-Seq data. *BMC Bioinformatics*, **14**, 31. (Cited on pages 10 and 93.)
- Patrick, E., Buckley, M., Lin, D. M., and Yang, Y. H. (2013b). Improved moderation for gene-wise variance estimation in RNA-Seq via the exploitation of external information. *BMC Genomics*, **14 Suppl 1**, S9. (Cited on pages 10 and 93.)
- Pearson, K. (1934). On a new method of determining "goodness of fit". *Biometrika*, **26**(4), 425–442. (Cited on page 56.)

- Polymenidou, M., Lagier-Tourenne, C., Hutt, K. R., Huelga, S. C., Moran, J., Liang, T. Y., Ling, S.-C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J. P., Shiue, L., Bennett, C. F., Yeo, G. W., and Cleveland, D. W. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat Neurosci*, **14**(4), 459–468. (Cited on page 42.)
- Praline, J., Limousin, N., Vourc'h, P., Pallix, M., Debiais, S., Guennoc, A.-m., Andres, C. R., and Corcia, P. (2010). CADASIL and ALS: a link? *Amyotroph Lateral Scler*, **11**(4), 399–401. (Cited on page 88.)
- Richard, H., Schulz, M. H., Sultan, M., Nürnbergger, A., Schrunner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S. A., and Yaspo, M.-L. (2010). Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res*, **38**(10), e112. (Cited on pages 26 and 96.)
- Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(3), R25. (Cited on pages 8 and 80.)
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140. (Cited on pages 9, 33, 35, and 93.)
- Rothstein, J. D. (2009). Current hypotheses for the underlying biology of amyotrophic lateral sclerosis. *Ann Neurol*, **65 Suppl 1**, S3–S9. (Cited on page 89.)
- Santagata, S., Demichelis, F., Riva, A., Varambally, S., Hofer, M. D., Kutok, J. L., Kim, R., Tang, J., Montie, J. E., Chinnaiyan, A. M., Rubin, M. A., and Aster, J. C. (2004). JAGGED1 expression is associated with prostate cancer metastasis and recurrence. *Cancer Res*, **64**(19), 6854–6857. (Cited on page 88.)
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, **2**(6), 110–114. (Cited on page 40.)
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**, Article3. (Cited on pages 35 and 38.)
- Srivastava, S. and Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res*. (Cited on page 9.)
- Stouffer, S., Suchman, E., DeVinney, L., Star, S., and Williams Jr, R. (1949). The American soldier: adjustment during army life.(studies in social psychology in world war ii, vol. 1.). (Cited on page 56.)
- Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9), 1105–1111. (Cited on pages 8 and 14.)
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**(5), 511–515. (Cited on pages 8, 15, 16, 23, 25, and 35.)

- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat Protoc*, **7**(3), 562–578. (Cited on page 25.)
- Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, **40**(9), 3785–3799. (Cited on page 55.)
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221), 470–476. (Cited on pages 14, 18, 23, and 24.)
- Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. (2010a). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, **38**(18), e178. (Cited on page 14.)
- Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010b). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**(1), 136–138. (Cited on page 8.)
- Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**(2), 149–159. (Cited on page 54.)
- Wang, Z., Zhang, Y., Banerjee, S., Li, Y., and Sarkar, F. H. (2006). Notch-1 down-regulation by curcumin is associated with the inhibition of cell growth and the induction of apoptosis in pancreatic cancer cells. *Cancer*, **106**(11), 2503–2513. (Cited on page 88.)
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1), 57–63. (Cited on page 5.)
- Ward, Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, **58**, 236–244. (Cited on page 21.)
- Welch, B. L. (1947). The generalisation of student's problems when several different population variances are involved. *Biometrika*, **34**(1-2), 28–35. (Cited on page 40.)
- Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychol Bull*, **48**(3), 156–158. (Cited on page 56.)
- Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, **8**, Article28. (Cited on page 54.)
- Wu, J., Irizarry, R., MacDonald, J., and Gentry, J. (2013). *gcrma: Background Adjustment Using Sequence Information*. R package version 2.26.0. (Cited on page 41.)
- Xing, Y., Kapur, K., and Wong, W. H. (2006). Probe selection and expression index computation of Affymetrix Exon Arrays. *PLoS One*, **1**, e88. (Cited on pages 18 and 21.)

- Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D. L., and Wong, W. H. (2008). MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA*, **14**(8), 1470–1479. (Cited on page 96.)
- Xu, J. and Wong, C.-W. (2013). Enrichment analysis of miRNA targets. *Methods Mol Biol*, **936**, 91–103. (Cited on pages 54 and 63.)
- Yang, P., Patrick, E., Tan, S.-X., Fazakerley, D. J., Burchfield, J., Gribben, C., Prior, M. J., James, D. E., and Yang, Y. H. (2013). Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. *Accepted in Bioinformatics*. (Cited on page 94.)
- Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*, **11**(2), R14. (Cited on page 87.)
- Yuan, J. S., Kousis, P. C., Suliman, S., Visan, I., and Guidos, C. J. (2010). Functions of notch signaling in the immune system: consensus and controversies. *Annu Rev Immunol*, **28**, 343–365. (Cited on page 88.)