University of Adelaide

Doctoral Thesis

BMEA: Bayesian Modelling For Exon Array Data

Author:
Stephen Martin Pederson

Supervisors:
Prof Simon Barry
A/Prof Gary Glonek

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

in the

Molecular Immunology Group
School of Medicine
Discipline of Paediatrics

November 2018

SUB CRUCE LUMEN

Declaration of Authorship

I, Stephen Martin Pederson, author of this thesis titled, 'BMEA: Bayesian Modelling For Exon Array Data':

- certify that the work is original and has not been accepted for the award of any other degree or diploma in any university or other tertiary institution and, contains no material previously published or written by another person, except where due acknowledgement is made in the text;
- certify that no part of the work will be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of the degree except where due reference has been made in the text;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- give consent for the thesis to be made available for loan and photocopying after it has been examined and placed in the library, subject to the provisions of the Copyright Act 1968
- give consent for the digital version of the thesis to be made available on the web, via the University's digital research repository, the library catalogue and also through web search engines
- acknowledge the support I have received for this research through the provision of an Australian Government Research Training Program Scholarship

Signed: ________________________________

Date: ________________________________

“Most of us discover where we are headed when we arrive. At that time, we turn around and say, yes, this is obviously where I was going all along. It’s a good idea to try to enjoy the scenery on the detours, because you’ll probably take a few.”

Bill Watterson

UNIVERSITY OF ADELAIDE

Abstract

Faculty of Health Sciences
School of Medicine
Discipline of Paediatrics

Doctor of Philosophy

BMEA: Bayesian Modelling For Exon Array Data

by Stephen Martin Pederson

The development of Affymetrix Exon Arrays was a significant step forward from 3' Microarray technology, however detection of alternate splicing events proved challenging. In this work a novel method, *Bayesian Modelling for Exon Arrays* (BMEA), is described which shows an improvement in performance over previous approaches, and fits a more appropriate model for each gene using an MCMC process. Applying BMEA to an in-house dataset contrasting resting and stimulated $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ cells, shed significant new light into key mechanisms involved in regulation of the T cell activation response.

Acknowledgements

A large number of people have played vital roles over the many years this work has taken to complete. Firstly, my deep gratitude is given to Prof Simon Barry and A/Prof Gary Glonek for their enthusiasm, wisdom and encouragement during the course of this work. Thank you for giving me the freedom to roam intellectually, and importantly, for not giving up on me when completion became increasingly difficult. Since moving into academia during the latter stages this project, the strong professional relationships I continue to share with both are highly valued.

The members of the Barry Lab, particularly from 2008-2013, were a ideal set of workmates and my thanks are extended to Bridget Wilkinson, Chris Wilkinson, Suzanne Bresatz-Atkins, Cheryl Brown, Danika Hill, Natasha Huxtable, Nicola Eastaff-Leung, Tessa Mattiske, Tessa Gargett, Tzu Ying Yap, John Welch and Liz Melville. Particular thanks and appreciation go to Dr Tim Sadlon for performing all of the lab work associated with this thesis, and for the countless PCR reactions he ran.

In the latter years of this work, as the demands of full-time work slowed progress to a near halt, a large amount of support, understanding and encouragement were extended to me by Prof David Adelson and Dr Dan Kortschak. The importance of this has not been lost on me and is greatly appreciated. The time taken by Dr Jimmy Breen and Dr Rick Tearle to read and review this thesis was also a significant aid in reaching the finish line. And to all the other members of the Bioinformatics Hub who have even been remotely interested in this work, a genuine thank you for your interest.

To Dr Jono Tuke, thank you for early advice, for setting such a high standard for R code and for teaching me so many new tricks.

To my dear friends Mel Horsman, Mandy Lumsden, Beck Kennedy and Jen Lush, who regularly stepped in with wine, coffee, chocolate and important personal support during some very difficult times, thank you.

To Peter Arthur & Justin Slater, with whom I have been making music for most of my life, thanks for waiting. A new album is long overdue. If you're interested in a new song idea, I've got a really dull story about a scientist who doesn't know when to quit. And to my more recent collaborators, Max McHenry, Carla Lippis, Kelly Menhennett and the many associated musicians, thanks for being such great company and a vital counterpoint to my

scientific life.

To all others who have taken an interest over the years, including, but not limited to, Suzanne Wilson, Jenny Simpson, Michael Robertson, Nicole Thompson, Justine Whitham, Craig Poole, Alf Bennett, Ali and Mike Coates, thanks for not asking too many questions.

Finally, to my Mother and Father who have waited patiently for me to finally finish, but would have been proud regardless, thank you for listening to my ramblings over and over again, thank you for your financial and moral support, thank you for providing the safety net I have needed far too often, and thank you for believing in me as a person.

Contents

List of Figures

List of Tables

Abbreviations

APC	**A**ntigen **P**resenting **C**ell
AG	**A**nti**G**enomic (probe)
AS	**A**lternate **S**plicing
BG	**B**ack**g**round
BGX	**B**ayesian **G**ene E**x**pression
BMEA	**B**ayesian **M**odelling for **E**xon **A**rrays
BUGS	**B**ayesian inference **U**sing **G**ibbs **S**ampling
CB	**C**ord **B**lood
CD	**C**luster of **D**ifferentiation
ChIP	**Ch**romatin **I**mmuno**p**recipitation
CPI	**C**entral **P**osterior **I**nterval
DABG	**D**etection **A**bove **B**ack**g**round
DE	**D**ifferentially **E**xpressed
DNA	**D**eoxyribo**n**ucleic **A**cid
ENA	**E**uropean **N**ucleotide **A**rchive
FACS	**F**luorescence **A**ctivated **C**ell **S**orting
FC	**F**old **C**hange
FIRMA	**F**inding **I**soforms using **RMA**
FOXP3	**F**orkhead B**ox** **P3**
GEO	**G**ene **E**xpression **O**mnibus
GC-RMA	**GC**-content dependent version of **RMA**
IL	**I**nter**l**eukin
IQR	**I**nter-**Q**uartile **R**ange
IPA	**I**ngenuity **P**athway **A**nalysis
IRLS	**I**teratively **R**e-weighted **L**east **S**quares
LB	**L**ower **B**ound

logFC	log **F**old-**C**hange
MAD	**M**edian **A**bsolute **D**eviation
MAS5.0	**M**icroarray **A**nalysis **S**uite v**5.0**
MAT	**M**odel-based **A**nalysis of **T**iling arrays
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MEE	**M**utually **E**xclusive **E**xons
MHC	**M**ajor **H**istocompatibility **C**omplex
MM-BGX	**M**ulti-**M**apping **BGX**
MM	**M**is**m**atch (probe)
ncRNA	**n**on-**c**oding **RNA**
NSB	**N**on-**S**pecific **B**inding
NUSE	**N**ormalised, **U**nscaled **S**tandard **E**rrors
PLIER	**P**robe **L**ogarithmic **I**ntensity **Er**ror
PLM	**P**robe-**L**evel **M**odel
PM	**P**erfect **M**atch (probe)
PSR	**P**robe **S**election **R**egion
RAR	**R**etinoic **A**cid **R**eceptor
RLE	**R**elative **L**og **E**xpression
RMA	**R**obust **M**ulti-chip **A**nalysis
RNA	**R**ibo**n**ucleic **A**cid
ROR$\boldsymbol{\alpha}$	**R**etinoic Acid Receptor-Related **O**rphan **R**eceptor α
ssDNA	**S**ingle **S**tranded **DNA**
$\mathbf{T_{conv}}$	**Conv**entional (i.e. non-regulatory) **T** Cell
$\mathbf{T_h}$	**H**elper (i.e. non-regulatory) **T** Cell
$\mathbf{T_{reg}}$	**Reg**ulatory **T** Cell
TCR	**T** **C**ell **R**eceptor
TGF$\boldsymbol{\beta}$	**T**ransforming **G**rowth **F**actor β
TSS	**T**ranscription **S**tart **S**ite
UCSC	**U**niversity of **C**alifornia, **S**anta **C**ruz
WT	**W**hole **T**ranscript

Chapter 1

Introduction

1.1 Overview

Beginning with the Mendelian concept of a gene as a simple but unknown unit of inheritance, to our current understanding as a physically identifiable region of genomic DNA, the field of genetics has undergone a rapid and profound expansion. The continually emerging complexity of gene regulation and gene expression has demanded the development of new genome-wide technologies which are able to reveal new insights into the many layers of gene regulation, which in turn require new analytic methods. The development of microarray technology as a high throughput approach to the quantification of gene expression, marked a significant milestone in our ability to assess gene behaviour within a whole genome context.

Beginning with the ability to quantify and compare gene expression levels (Lockhart et al., 1996; Schena et al., 1995), subsequent generations of arrays were developed to be theoretically capable of revealing additional insight into the structure of the expressed transcripts (Gardina et al., 2006). Accessing this finer level of detail proved to be a great challenge and in this body of work a new model will be investigated as an improvement for extracting this deeper level of information, along with an R package which will enable the application of this approach to any suitable dataset. However, since the commencement of this body of work, RNA-Seq (Mortazavi et al., 2008) and tools such as *kallisto* (Bray et al., 2016) have rapidly become the preferred platform for addressing these questions, potentially marking this work as more of a footnote than a breakthrough.

Whilst the focus of this work has been primarily on algorithm development and data analysis, the dataset being investigated targets the specific role within the immune system of Regulatory T Cells ($\mathrm{T}_{\mathrm{reg}}$). In this chapter, we will first look at the underlying technologies and statistical methods, before presenting a brief introduction to the biological context, with the primary dataset introduced in the second chapter. The subsequent analysis and development of the BMEA model is given in detail from Section 3.1 onwards.

The Microarray dataset under primary consideration in this work is available from the Gene Expression Omnibus (`https://www.ncbi.nlm.nih.gov/geo/`) under the accession GSE20934. The FOXP3 ChIP-Chip data included as an additional dataset is available under the accession GSE20995.

All analytic code, including the LaTeX code used to write this thesis, is available at `https://bitbucket.org/steveped/thesis`. The R Package developed as part of this work is available at `https://github.com/steveped/BMEA`.

1.2 Microarray Technology

1.2.1 Biological motivation

A fundamental biological dogma is that genomic DNA is transcribed into mRNA, which is in turn translated into protein. Although biological reality is often far more complex, the relative quantification of a transcribed mRNA in a specific biological context can yield vital information as to the underlying mechanisms associated with cellular function and cellular fate. This desire to assess relative gene expression levels and understand the regulatory processes on a genome-wide basis led to the development of microarrays and other high-throughput technologies.

A single gene can be expressed in a variety of isoforms which arise from a multitude of control mechanisms, with a recent study concluding that up to 95% of multi-exon genes exhibit some degree of alternate splicing (Pan et al., 2008). A gene may have multiple initiation sites, where transcription begins in a different genomic location, or the well-established intron/exon structure of a transcript can yield multiple protein isoforms from the same transcript as a result of post-transcriptional mRNA splicing (Figure 1.1).

Beyond quantifying the expression level of a gene, identification of expressed transcripts is a biological question of great importance, as differing isoforms of a gene can vary greatly in function. For example, two isoforms of the nuclear co-repressor 2 (NCOR2) protein arise through alternately spliced transcripts and have identical affinities for their partner retinoic acid receptor (RAR) proteins which promote gene transcription (Goodson et al., 2005). However, they have vastly different binding affinities for additional partner transcriptional repressor (TR) proteins, which results in preferential expression or repression of this subset of genetic pathways in response to the relative quantities of the two isoforms.

In addition to *protein-coding* transcripts, any number of transcripts from a given gene may be *non-coding* transcripts, with the role of the vast majority of these being poorly understood. Recently, a non-coding transcript of *ASCC3* was shown to functionally oppose the protein transcribed by the same gene (Williamson et al., 2017), highlighting the potential importance of these relatively uncharacterised transcripts.

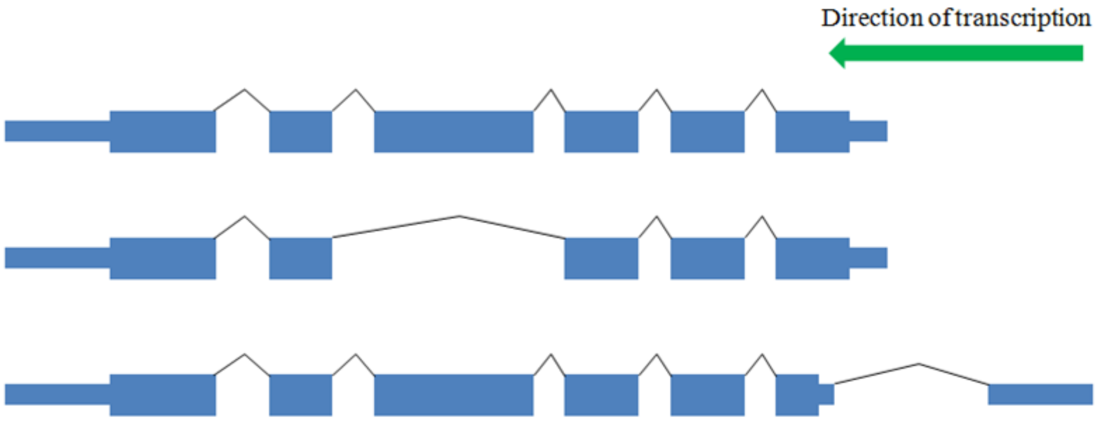


Figure 1.1 – *Example of a hypothetical gene with three possible transcripts. The top two have the same transcription initiation site, however the fourth exon is spliced out in the second transcript. The bottom transcript has an alternate transcription initiation site which results in a first exon which is different to the other two transcripts. The protein products will thus be three unique isoforms.*

1.2.2 3' Microarrays

Emerging quickly as the dominant platform amongst researchers, the early generations of Affymetrix GeneChip® Microarrays (Lockhart et al., 1996) were designed to quantify the abundance of mRNA transcripts within a single biological sample. Each array consisted of millions of spatially-identified, 25-mer single-stranded DNA (ssDNA) probes which are perfectly complementary to a corresponding target genomic sequence, known as perfect match (PM) probes. Cellular mRNA from each sample is fluorescently labelled during amplification, then hybridised to a separate array. Each array is then scanned using a laser to excite the fluorescent dye, with the fluorescence intensity at each predefined location being proportional to the abundance of the target sequence specifically associated with the PM probe at that location, acting as a proxy for gene expression. Each PM probe was additionally paired with a mismatch (MM) probe for which the middle (13th) base is changed, which are used for estimation of background signal (Lockhart et al., 1996) within the observed PM probe intensities.

Target sequences for the PM probes on these first generation of Affymetrix arrays were restricted to the 3' region of an expressed transcript (Lockhart et al., 1996), ensuring that only intact mRNA molecules were measured for a given gene. PM-MM probe pairs were designed for 11-20 unique sequences within the 3' region of each target transcript, and these were combined to form a transcript-specific *probeset* (Lockhart et al., 1996). The signal across all probes within a probeset can then be used to obtain an estimate of the expression level for the corresponding transcript within each sample. By using multiple arrays across experimental conditions, any change in expression level in response to the treatment can be assessed for each transcript targeted by the array.

1.2.3 Whole Transcript Microarrays

With the advent of the Affymetrix Gene and Exon Arrays, probes were instead designed to target sequences throughout the length of each transcript. Up to four probes were designed for each “probe selection region” (PSR), which roughly corresponds to an exon (Affymetrix, 2005e) and can be considered as an exon-level probeset (Figure 1.2). Multiple layers of annotation were also introduced with exon-level probesets being grouped into transcript clusters, and being classified as *core*, *extended*, *full*, *free* or *ambiguous* depending on the strength of the supporting evidence for expression of each sequence (Affymetrix, 2005c).

The strict PM-MM probe pairing was no longer adhered to for whole transcript arrays, and two unique sets of probes were instead included for background correction. The set of *genomic background probes* represent mismatch (MM) probes to a subset of PM probes, selected from genomic regions thought to be poorly expressed. The additional set of *antigenomic background probes* contain probes with no match to any known expressed sequence (Affymetrix, 2005e) at the time of array design. As a result, the antigenomic set of probes will not be sensitive to the binding of the PM target sequence, as occurs with conventional MM probes.

With the ability to generate transcript-level and exon-level expression estimates, these arrays promised to identify which specific transcripts of a gene are being expressed within a sample, and to detect changes in mRNA structure across multiple samples. Whilst providing significantly more information about gene expression, utilising this extra information proved a substantial analytic challenge (Purdom et al., 2008; Turro et al., 2010).

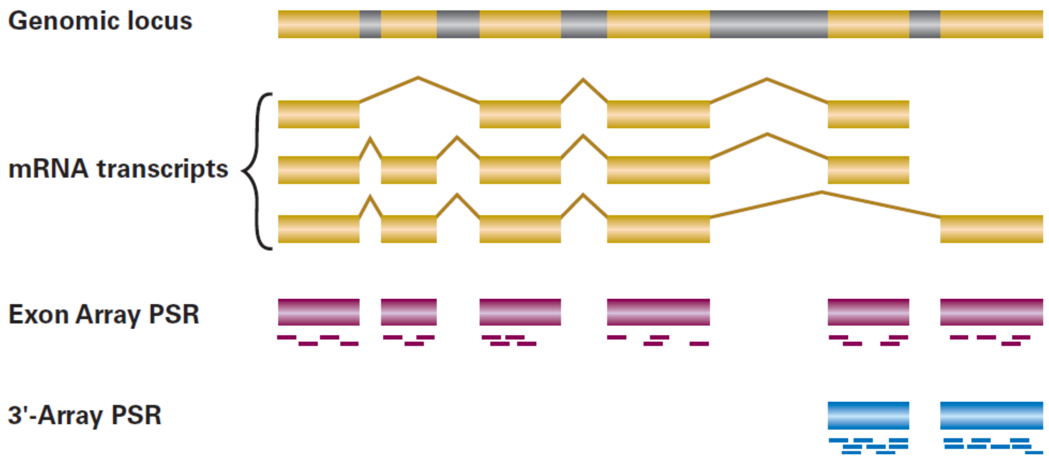


Figure 1.2 – *Comparison of 3' Array & Exon Array design and probe layout. A 3' Array would be able to differentiate between the third transcript and the previous two, whilst an Exon Array would theoretically be capable of differentiating between the all three transcripts. Figure taken from Affymetrix Technical Note 702026 (Affymetrix, 2005e).*

1.2.4 Chip Description Files

The probe intensity data from each type of GeneChip® is stored using the Affymetrix .CEL file format, and each specific type of array has an associated *Chip Description File* (CDF). The CDF maps the physical location of the probes on the array using X & Y co-ordinates, to the appropriate probeset. Although no sequence information is specifically contained in these files, this is publicly available from the manufacturer (`www.affymetrix.com`).

Probeset assignment for each probe is typically performed at design-time, and is indicative of known expressed sequences in genomic databases at that particular point in time. As the rate of re-annotation in genomic databases is high, the constantly changing genomic information can be incorporated into an analysis via the creation of custom CDF files, which have been shown to improve the accuracy of any analysis (Dai et al., 2005; Sandberg & Larsson, 2007). During construction of these custom files, probe sequences can be checked against current genome builds to ensure accurate mapping of probe sequence to transcript, and to ensure that sequences are detecting signal from unique targets.

1.3 Statistical Approaches

As any high-throughput analysis will be heavily dependent on the use of statistical tools and methods, a brief summary of some integral approaches is essential before discussing the technical challenges of specific to microarray analysis.

1.3.1 Classical Statistics

Classical (or frequentist) statistical approaches have been the dominant methodology for decades. Under these approaches, every experiment is considered to be a random sample from a larger body of independent experiments, and inference for population-level parameters is performed using estimates obtained from each specific sample. Two common population-level parameters of interest are the population mean (μ) and variance (σ^2) which are estimated using the *sample mean* ($\overline{y}$) and *sample variance* (s^2) respectively. For a set of normally distributed observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, it can be shown that the sample mean

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.1}$$

is the value for the unknown estimate of the population mean ($\hat{\mu}$) which minimises the sum of squares

$$f(y) = \sum_{i=1}^{n}(y_i - \hat{\mu})^2 \tag{1.2}$$

and is consequently the best linear, unbiased estimator.

For classical statistical inference, a suitable hypothesis is formulated and a *p-value* obtained which represents the probability of observing data at least as extreme as the given data, if the hypothesis is true. This is traditionally framed in terms of a *null hypothesis* (e.g. $H_0 : \theta = 0$) and an *alternate hypothesis* (e.g. $H_A : \theta \neq 0$), where θ represents a quantity or set of interest. For example, a null hypothesis may be that the mean expression level of a gene (μ) is the same in two cell types, and thus the quantity of interest would be the difference in the population means ($\theta = \mu_1 - \mu_2$). If the obtained p-value is below a specified threshold (α), the null hypothesis is rejected and the alternate hypothesis accepted. Importantly, under this approach no prior knowledge about any of the model parameters is assumed.

1.3.2 Multiple Testing Considerations

A common threshold of significance for any p-value obtained as above is $\alpha = 0.05$, and this value acts as an upper bound for obtaining a *false positive*, where the null hypothesis is incorrectly rejected in a single hypothesis test. As the number of hypotheses tested within an experiment increase in number, the probability of at least one false positive, also known as the family-wise error rate (FWER), increases such that across an experiment involving m hypothesis tests, this can be simply observed as:

$$P(\text{at least one false positive}) = FWER \leq m\alpha \, . \tag{1.3}$$

Across a microarray experiment, where thousands of statistical tests are performed in parallel, methods of controlling the error rate are a necessity. Methods which strictly control the FWER, such as those suggested by Bonferroni or Holm (Holm, 1979), are generally very stringent but can lack statistical power. Thus a common approach is to instead control the *false discovery rate* (FDR), in which a number of false discoveries are permitted and estimated within a group of results. The most commonly-used FDR approach is that proposed by Benjamini & Hochberg (Benjamini & Hochberg, 1995), and this method has also been shown to control the FDR under positive correlation between test statistics (Benjamini & Yekutieli, 2001).

1.3.3 Robust Statistics

Statistical methods can also be used to obtain parameter estimates, with the calculation of a single expression estimate for a microarray probeset being a prime example. In this particular case, we would commonly have 11-20 probe-level signal estimates within each probeset, and a single probeset-level estimate is required for each array, which is in turn used for comparisons across multiple arrays and conditions. A simple approach to this would be to use the *sample mean* for the set of probes on each array, which is an unbiased and efficient estimator for normally distributed data. However, the sample mean can be sensitive to any departures from normality such as the presence of outliers, as are common with microarray probe intensities. *Robust estimators* offer practical alternatives to the sample mean which less sensitive to presence of outliers in the data.

Whereas the sample mean is the value that minimises the sum of squares in equation 1.2, an alternate class of estimators, known as *M-estimators*, is obtained by minimising a suitable objective function $\rho(u)$ (Huber, 2005). For the value

$$u_i = \frac{y_i - \hat{\mu}}{\hat{\sigma}} \tag{1.4}$$

with $\mathbf{y}$ as defined as in section 1.3.1, and the desired location and scale estimates indicated by $\hat{\mu}$ and $\hat{\sigma}$ respectively, the sum of squares in equation 1.2 can be rewritten as solving $min_{\hat{\mu}} \sum_{i=1}^{n} \rho(u_i)$ where $\rho(u) = u^2$. In contrast, *Huber's M-estimator* (Huber, 2005) minimises the function

$$\rho(u) = \begin{cases} u^2 & \text{if } |u| < c \\ c(2|u| - c) & \text{otherwise} \end{cases} \tag{1.5}$$

for some constant c (commonly $c = 1.345$), whilst *Tukey's bisquare* (Hoaglin et al., 1983) minimises the function

$$\rho(u) = \begin{cases} \frac{1}{6}[1 - (1 - u^2)^3] & \text{if } |u| \leq 1 \\ \frac{1}{6} & \text{if } |u| > 1 \end{cases}. \tag{1.6}$$

A closed form solution does not exist for these M-estimators, and the Iteratively Re-weighted Least Squares (IRLS) algorithm is commonly used solve the minimisation (Holland & Welsch, 1977). Both Tukey's bisquare and Huber's M-estimator are alternatives to the sample mean used for estimating an array-specific probeset-level expression estimate as described above, and will be less susceptible to outlier probes as are commonly found in microarray data. Robust measures of variability are also in widespread use with the most common being the *Median Absolute Deviation* (MAD)

$$\text{MAD}(\mathbf{y}) = \text{median}(\mid y_i - \hat{\mu} \mid) \tag{1.7}$$

where $\hat{\mu}$ here represents the sample median.

1.3.4 Bayesian Statistics

Whilst robust statistics essentially build on classical statistical approaches, an alternate analytic framework is that of *Bayesian statistics.* Under the Bayesian framework, prior knowledge about the unknown parameter θ is incorporated into the model via a *prior* distribution $f(\theta)$. Given the observed data $\mathbf{y} = (y_1, y_2, \ldots, y_i)$, a *posterior* distribution is obtained for $f(\theta|\mathbf{y})$ using Bayes' theorem

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{\int(\mathbf{y}|\theta)f(\theta)d\theta} \tag{1.8}$$

with inference made from this posterior distribution. This makes inference such as $p(\theta > 0)$ far more intuitive than under the frequentist approach.

Hierarchical Bayesian models are an extension of the above principles where the parameters of the prior distribution are no longer fixed, but are instead given their own set of prior distributions. For example, the set of observations $\mathbf{y}$ may be defined as having a prior distribution with unknown mean, i.e. $y_i \sim \mathcal{N}(\theta, \sigma)$. The parameter θ would be given a prior distribution such as $\theta \sim \mathcal{U}(a, b)$, referred to as a *hyperprior* with the fixed values defining parameters a and b referred to as *hyperparameters* (Gelman et al., 2004).

In complex models, obtaining the posterior distribution can be analytically overwhelming and simulation techniques such as Markov Chain Monte Carlo (MCMC) can be utilised (Gelman et al., 2004). Under this iterative approach, random draws for each parameter are sampled to form posterior distributions, which are subsequently used for posterior inference about suitable hypotheses, such as $p(\theta < 0)$. This method of analysis has only recently become feasible with the advent of increasingly fast and powerful computers, and is being applied with growing frequency across many fields of statistical analysis.

As an alternative to MCMC strategies, Empirical Bayesian methods allow for specification of hyperparameters by estimating suitable values from the observed data (Efron & Morris, 1972). In the case of microarrays, where thousands of genes are analysed as individuals within a larger set, hyperparameters for a hierarchical model defined for an individual gene, can be estimated by inspecting the complete set of data. Posterior point estimates of model parameters may then be obtained where possible, as introduced in Section 1.6.

1.4 Technical Considerations for Microarrays

1.4.1 Optical Correction

Before the individual observed probe intensities can be combined as a probeset-level estimate of expression, some key non-biological artefacts must be taken into account at the individual probe level. The signal observed at a given location will not only contain fluorescence from any bound sequence, but may also contain optical noise (O) from nearby probes and from the surface itself. Whilst not being strictly uniform across an array surface, for mathematical simplicity it is often assumed to be so and defined as the lowest observed probe intensity on an array (Irizarry et al., 2003).

1.4.2 Non-specific Binding

Despite stringent hybridisation and washing steps, some of the observed signal at each probe is due to *non-specific binding* (NSB) of an off-target sequence. The MM probes on 3' Arrays were included for estimation of this component of the signal. However, the ability of the MM probes to also bind the target sequence for the PM probe, and in some cases provide a stronger signal than the PM probe, has been well documented (Irizarry et al., 2003). This has led to numerous approaches to background correction and estimation of the *true* signal, as detailed in section 1.5.3.

1.4.3 Detection Above Background

Some probes within a probeset may also be consistently unresponsive (Sanchez-Graillet et al., 2008) and can bias any probeset level expression estimates downwards, and as such, removal of these probes can be of great benefit to analysis. A common method for assessment of the presence of signal (S) as opposed to purely background noise (B), is Detection Above Background (DABG) (Clark et al., 2007). Under this approach the probe signals are compared to the signal distribution from background probes with a similar sequence structure, and probes with a low probability of containing signal, e.g. $Pr(S > 0) < 0.05$, are removed from the analysis.

1.4.4 Quantile Normalisation

A further source of variation between arrays can be caused by subtle differences in the amount of sample bound to the surface of each array, or other non-biological variability which leads to some arrays having differing intensities (see Figure 2.2 for an example). Quantile normalisation is the process by which the overall signal level for each array is given the same distribution which minimises the impact of this variation in signal intensities which is likely non-biological in origin (Irizarry et al., 2003). In brief, the individual probes with the lowest signal on each array are given the between-array average of the observed signals for those probes, and so forth with the second and third dimmest probes until all probes have been normalised. Notably, this is performed regardless of probeset membership, and is based purely on the observed signal at the probe level. In general, this step is performed after some background correction procedures have been performed, as described in Section 1.5, but before probeset-level expression estimates are obtained through the probe-level models detailed in Equation 1.11.

1.5 Background Correction Methods

1.5.1 The Additive Background Model

The fundamental premise behind background correction methods for Affymetrix arrays, is that the observed PM intensity is a combination of true signal (S) from the target sequence, and background signal (B) where B contains both optical signal and non-specific binding, such that

$$PM = B + S\,. \tag{1.9}$$

The aim from most existing background correction approaches is to simply provide a single probe-level estimate of the true signal ($\hat{S}$) via some corrective process, under the clear restriction that $S > 0$, in cases where DABG procedures are not used.

1.5.2 Affymetrix Approaches

For the strict PM-MM pairings of 3' arrays, the MAS5.0 background correction method was introduced by Affymetrix, which utilised localised optical correction for each probe (Liu et al., 2002) along with a modified MM value (Hubbell et al., 2002). This modified-MM is taken as the estimate of background signal and subtracted from the observed PM value. Probeset-level expression estimates are then obtained using Tukey's bisquare (Section 1.3.3) on the $\log_2$-transformed, background-corrected PM signal values. Whilst known to be a strongly performing algorithm for highly expressed genes, a high level of variance instability is introduced for more moderate to low expressed genes, primarily due to the separate variance components of the observed PM and the observed MM signals (Zakharkin et al., 2005). The PLIER algorithm (Affymetrix, 2005d) was introduced as an alternative, and whilst still being reliant on PM-MM pairings, utilised a model-based approach to background correction with considerably improved performance (Irizarry et al., 2006).

For whole transcript arrays the PM-MM pairings are no longer available, rendering this data incompatible with both MAS5.0 and PLIER. Background probes are instead placed into bins based on GC content, and the initial method proposed (Affymetrix, 2005b) was to simply subtract the median observed value from the corresponding GC bin for each PM probe.

1.5.3 RMA and GC-RMA

RMA background correction (Irizarry et al., 2003) was developed to provide a single probeset-level summary using observed PM intensities independently of any MM observations, hence removing the variability and noise introduced by the MM values. A theoretical background signal distribution is used to obtain the expected value for true signal:

$$\hat{S}_{ik} = E(S_{ik}|PM_{ik}) \tag{1.10}$$

without reference to probe sequence content. A single probeset-level value ($\hat{c}_i$) is then obtained for each array by fitting the probe-level model (PLM):

$$\log_2 \hat{S}_{ik} = c_i + p_k + \varepsilon_{ik} \,. \tag{1.11}$$

In contrast to MAS5.0, this model incorporates a probe affinity term p_k which represents the ability of each probe ($k = 1, \ldots, K$) to bind its target sequence, under the constraint $\sum_{k=1}^{K} p_k = 0$. The expression level for each array ($i = 1, \ldots, I$) is represented by the chip effects term c_i with errors ε_{ik} assumed to independent identically distributed with mean 0. The model is fitted robustly using Tukey's median polish (Tukey, 1977) method, bypassing any assumptions of normality for ε_{ik}.

Whilst RMA directly resolved the variance instability observed under MAS5.0, a subsequent refinement was to incorporate probe sequence information to estimate background signal, via GC-RMA (Wu et al., 2004). Under GC-RMA, the contribution of each base along the length of the probe is taken into account and an affinity curve estimated for the position of each base (Figure 1.3). Whilst the final *probeset-level* expression estimate is still obtained robustly, *probe-level* signal estimates are obtained using an empirical Bayesian point estimate, based on observed PM and MM intensities, as well as the estimated affinity curves. The original GC-RMA approach utilised the PM-MM pairings, and an alternative was also proposed in which a training set of background probes were used to estimate model parameters (Wu et al., 2004), with minimal reduction in performance.

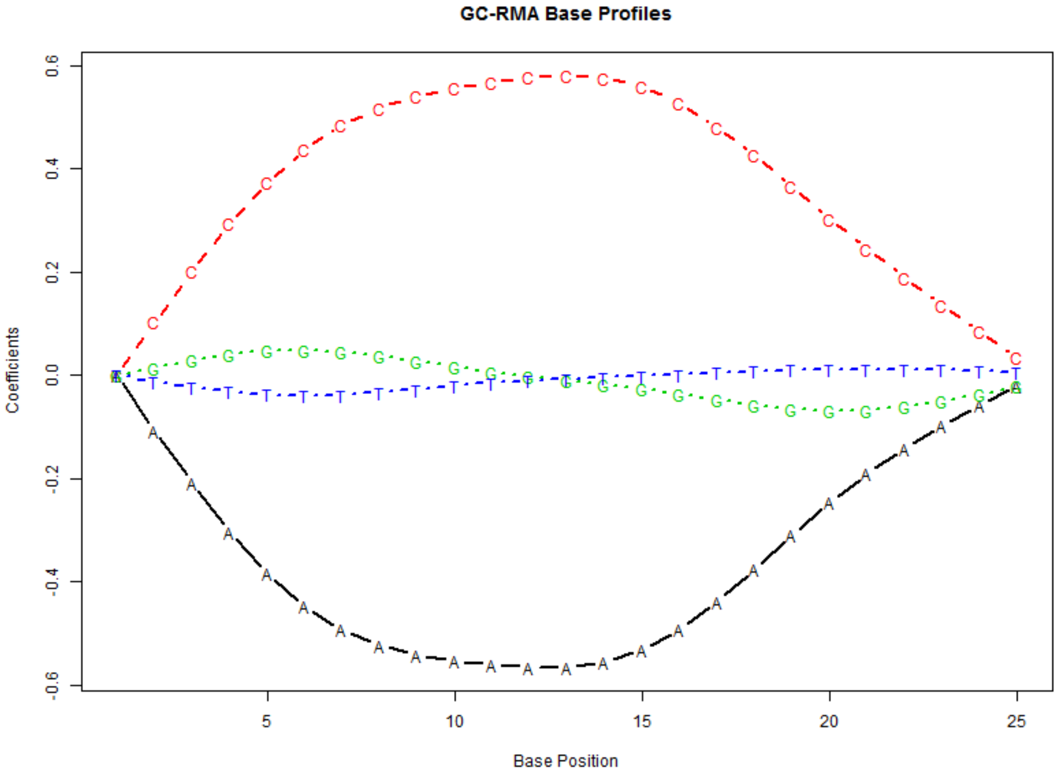


Figure 1.3 – *Default GC-RMA base profiles. These are estimated for 3' Arrays from a non-specific binding experiment, in which no gene specific binding is expected. Values taken from Wu, Irizarry, MacDonald, & Gentry, 2011.*

1.6 Differential Gene Expression

1.6.1 Moderated t-statistics

Regardless of the background correction method, a single summary value ($\hat{c}_i$) will be obtained for each probeset on array i, which is assumed as proportional to the abundance of the target transcript (or gene). These values can then be fitted across the entire dataset, using an appropriate statistical model, to obtain an estimate of differential expression between biological conditions under investigation. Suitable ranking methods are then used to obtain a list of candidate genes for differential expression.

A common ranking method is to use a t-statistic, however as some genes will have a very low variability between samples and others may have excessive variability, the former may be ranked highly in a final list whilst the latter may be ranked lower than is appropriate given true biological behaviour. A *moderated* $\tilde{t}$-statistic (Smyth, 2004) can instead be calculated to account for these artefacts which are likely non-biological in origin. Using the observed behaviour across all genes on a set of arrays to estimate the hyperparameters of the prior distribution (Smyth, 2004), an empirical Bayesian approach is used to provide a "moderated" point estimate for the variance of each gene. This value is substituted into the conventional calculations for the t-statistic to create a moderated $\tilde{t}$-statistic, which has been shown to reduce the number of false positives (Smyth, 2004). The resultant set of p-values can then be used to identify differentially expressed (DE) genes, after accounting for multiple testing considerations using methods such as the Benjamini-Hochberg FDR (Benjamini & Hochberg, 1995).

1.7 Exon Array Analysis

1.7.1 Unique Design Properties of Exon Arrays

Instead of a single probeset targeting the 3' end of a transcript, probes on an Exon Array are arranged into exon-level probesets (or *groups*) with up to four probes per group, which are contained within gene-level "metaprobesets" (or *units*). This dual-layer structure significantly complicates the analysis of whole transcript arrays, as the implicit assumption of the PLM in equation 1.11, is that each probe within a probeset will be detecting the same level of signal. This will largely remain valid at the exon level, but is no longer valid at the gene-level as some exons may not be included in all transcripts.

1.7.2 The Splicing Index and Related Approaches

An early and intuitive measure for detection of splice variation was the Splicing Index (SI) (Clark et al., 2002)

$$SI_{ij} = \frac{\hat{e}_{ij}}{\hat{c}_i}, \tag{1.12}$$

in which a gene-level signal estimate ($\hat{c}_i$) is first obtained for each array ($i = 1, \ldots, I$). The splicing index is then calculated by normalising the exon-level signal ($\hat{e}_{ij}$) to the gene-level signal estimate for each exon ($j = 1, \ldots, J$).

An extension of the Splicing Index approach is Microarray Detection of Alternate Splicing (MIDAS) (Affymetrix, 2005a). The MIDAS approach assumes that the exon-level signal can be described by

$$e_{ij} = \alpha_j p_{ij} c_i, \tag{1.13}$$

with exon and gene-level signals as above, but where $\alpha_j = argmax_i(SI_{ij})$, and p_{ij} represents the proportionate expression of each exon within each sample, such that $0 \leq p_{ij} \leq 1$ with $p_{ij} = 1$ in the sample from which α_j is derived. Taking logarithms, this simplifies to

$$\log(SI_{ij}) = \log(e_{ij}) - \log(c_i) = \log(\alpha_j) + \log(p_{ij}). \tag{1.14}$$

The MIDAS procedure incorporates an error term into equation 1.14 and at the exon-level is a one-way ANOVA testing for $\log(p_{ij}) = 0$, or $p_{ij} = 1$ as complete exon inclusion.

Another improvement to the Splicing Index was to incorporate probe sequence effects via the Corrected Splicing Index (COSIE) approach (Gaidatzis et al., 2009). Under COSIE, the non-linear nature of probe response is modelled and a corrected splicing index calculated after taking these effects into account, successfully reducing the effects of gene-level fold-change on alternate splice detection. However, it should be noted that in all SI related approaches, the gene-level signal ($\hat{c}_i$) is fitted by robustly incorporating signal from all probes, which assumes that the majority of probes are detecting the same level of signal.

1.7.3 FIRMA

A subsequent method of splice detection was proposed as *Finding Isoforms using RMA* (FIRMA) (Purdom et al., 2008), in which probe-level signal estimates $y_{ijk} = \log_2 E(S_{ik}|PM_{ik})$ are first obtained using RMA background correction, with gene-level expression estimates obtained using the standard PLM from Equation 1.11. Residuals r_{ijk} are obtained for each probe ($k = 1, \ldots, K$) within each exon ($j = 1, \ldots, J$) on each chip ($i = 1, \ldots, I$)

$$r_{ijk} = y_{ijk} - \hat{c}_i - \hat{p}_k, \tag{1.15}$$

with the chip-specific FIRMA score for each exon F_{ij} defined as

$$F_{ij} = \text{median}_{k \in \text{exon} j} \frac{r_{ijk}}{s} \, . \tag{1.16}$$

This is effectively the median residual for each exon scaled by s, which is defined as the median absolute deviation (MAD) of all residuals within each gene.

An extreme FIRMA score for an exon will be indicative of poor model fit, and provide candidate exons for alternate splicing events. Once again, the gene-level expression estimate ($\hat{c}_i$) is obtained with no explicit modelling of exon-level effects, and by assuming that the robust fitting will minimise any effects due to missing exons.

1.7.4 MM-BGX

A subsequent and unique approach to the analysis of whole transcript arrays was Multi-Mapping Bayesian Gene Expression (MM-BGX) (Turro et al., 2010) in which each transcript of a gene was treated as a multiple mapping event for any common probes, and the proportion of signal due to each transcript was estimated using an MCMC approach. Notably, this was the first analytic approach in which the gene-level signal is not fitted during the first stage of the analysis. A strong improvement in performance was shown over FIRMA and SI-based approaches, however, the reliance on existing transcripts restricts the analysis to those which have been previously identified, and novel splicing events will not be detected.

1.8 Chromatin Precipitation Arrays

The microarray dataset under primary consideration in this work was partially analysed in conjunction with a complementary ChIP-Chip dataset. As such, a brief introduction to the molecular and analytic processes behind this type of data is necessary.

1.8.1 Chromatin Immunoprecipitation (ChIP)

Microarrays are able to detect the relative abundance of mRNA transcripts within and between samples, but are unable to shed light on the molecular mechanisms behind any changes in gene expression. The ability to form cross-links between proteins and DNA in vitro led to the development of Chromatin Immunoprecipitation (ChIP), in which the DNA-binding of a protein of interest can be observed (Gilmour & Lis, 1985). Cross-linking is performed within a sample and cellular DNA is fragmented before a primary antibody to the protein of interest is added. The solution is then immunoprecipitated, using a secondary antibody to extract only the elements of the sample containing the primary antibody, and hence only the fragments of DNA to which the protein of interest is bound. The cross-links are then reversed and the precipitated DNA fragments can be analysed using techniques such as ChIP-Seq (D. S. Johnson et al., 2007) or ChIP-Chip (Blat & Kleckner, 1999; Kapranov et al., 2002).

1.8.2 ChIP-Chip Analysis

Using the same approach as for expression arrays, a DNA sample obtained from a ChIP experiment can be amplified and hybridised to a Genomic Tiling Array (Carroll et al., 2006), with the process becoming known as *ChIP-Chip*. Whilst retaining the 25 bp design, PM probes on the arrays are no longer restricted to expressed genomic sequences, but instead provide coverage for the vast majority of the genome. Rather than comparing between samples to detect changes in gene expression, tiling arrays are able to reveal which genomic regions the protein of interest has bound to in the initial sample, and can be used to identify over-represented transcription factor (TF) binding motifs in these regions, or to define a previously unknown consensus binding motif for a given transcription factor (Li et al., 2005).

1.8.3 The MAT Background Correction Model

As with expression arrays, the analysis of tiling array data presents many statistical challenges, such as the assessment of background signal and quantification of bound DNA fragments. The MAT model (W. E. Johnson et al., 2006)

$$\log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k \in (A,C,G)} \beta_{jk} I_{ijk} + \sum_{k \in (A,C,G,T)} \gamma_k n_{ik}^2 + \delta \log(c_i) + \varepsilon_i \,, \tag{1.17}$$

as proposed for Affymetrix Tiling Arrays, defines the baseline behaviour of each probe based on sequence composition and quantifies the presence of each target sequence utilising a least squares fit. The parameters in the above model are defined by:

- PM_i as the observed intensity value of PM probe i;
- k represents one of the four nucleotides A, C, G or T;
- n_{ik} is the count of nucleotide $k : k \in (A, C, G, T)$ within PM probe i;
- α is the intercept (or constant) based on the number of T nucleotides within a probe;
- I_{ijk} is a binary indicator function such that $I_{ijk} = 1$ if the nucleotide at position j is k in probe i, and $I_{ijk} = 0$ otherwise;
- β_{jk} is the effect of each nucleotide k at each position j, with the exception of $k =$T;
- γ_{ijk} is the effect of nucleotide count squared;
- c_i is the number of times that the sequence of probe i appears in the genome;
- δ is the effect of the log of the probe copy number; and
- ε_i is the probe-specific error term, assumed to follow $\mathcal{N}(0, \sigma)$

Once the baseline intensity has been estimated for each probe, a sliding window approach is used to detect genomic regions where probes consistently exhibit higher than the expected baseline intensities. These regions are candidates for enrichment and implicate binding by the protein of interest.

1.9 Software

For the processing of the data formats as introduced above some key software tools are available. These have been used significantly in this body of work for both data analysis and during development of the final model.

1.9.1 The R Statistical Software Environment

Affymetrix provide the free Expression Console software as part of their suite of analysis packages, however, a great deal of data processing and analysis is performed using the open-source statistical environment R (R Development Core Team, 2017), and the many associated packages and projects. A repository tailored to analysis of biological data is the Bioconductor project (Gentleman et al., 2004) which hosts packages for the implementation of RMA and GC-RMA, as well as for quantile normalisation and other relevant algorithms for the analysis of microarray data.

With the advent of Whole Transcript and ChIP-Chip arrays, data storage requirements were significantly increased and management of these was included as part of the aroma-project (Bengtsson et al., 2008). The *aroma.affymetrix* package was designed to enable analysis of large genomic datasets by retaining most information on hard disk instead of resident memory, enabling fast data access via caching.

1.9.2 WinBUGS

Developed specifically for the implementation of Bayesian models under the Windows operating system, the BUGS language enables flexible model specification and posterior simulation via MCMC methods for a wide variety of models, without the need for the analytic derivation of posterior distributions (Gilks et al., 1994). Utilising the relative efficiency of the C programming language, the Windows implementation of the language (WinBUGS) enables model development and testing as well as complete analysis, and is able to be accessed from R via the *R2WinBUGS* package (Sturtz et al., 2005b).

1.10 FOXP3 and Regulatory T Cells

The microarray dataset under primary investigation in this work centres on regulatory T cells (T_{reg}), helper T cells (T_h) and their joint response to immune activation. In order to provide context for the subsequent analysis, a brief introduction to some key biological processes and molecules is presented below, with a full description of the experimental outline and sample preparation in Chapter 2.

1.10.1 T Cells and the T-cell Receptor

Deriving from haematopoietic stem cells in the bone marrow, immature T cells, or *thymocytes*, migrate to the thymus where they undergo multiple stages of development and selection, before being released into the peripheral tissues as mature T cells. All thymocytes exhibit a small degree of self-reactivity via their T-cell receptor (TCR) and during thymic development, approximately 98% of cells die as a result of unsuitable TCR expression (Janeway et al., 2001b).

Unlike direct binding of the native-state antigen, as occurs with antibodies and B-cell receptors, antigen recognition by the TCR is dependent on display by an antigen presenting cell (APC), such as a dendritic cell. Short, contiguous amino-acid sequences are recognised when partially unfolded and displayed by a Major Histocompatibility Complex (MHC) molecule on an APC (Janeway et al., 2001a). Once the TCR has recognised and bound both the antigen and the MHC molecule, and the appropriate co-stimulatory signals are received from the APC, the T cell becomes activated and undergoes *clonal expansion*. This process results in the rapid proliferation of T cells specific to the recognised antigen, and is further accompanied by the expression of pro-inflammatory cytokines such as IL-2. After 4-5 days of proliferation, the activated T cells develop into armed effector T cells which are capable of destroying any cell displaying the antigen without the need for further co-stimulation (Janeway et al., 2001c), thus enabling clearance of infected cells.

In additional to variability within the TCR, T cells are a highly heterogeneous cell population encompassing many specific cell-types for distinct functional roles, with many of these defined by the expression of characteristic surface proteins. Well-defined surface molecules are assigned a Cluster of Differentiation (CD) number (Zola et al., 2007) and a common delineator of T cell function is the presence of either CD4 or CD8 on the cell surface. These molecules are simultaneously expressed during thymocyte development, but denote separate lineages once mature T cells are released into the periphery. $CD4^{-}$ $CD8^{+}$ T cells represent the broad class of Cytotoxic T cells, whilst the $CD4^{+}$ $CD8^{-}$ population contains both Helper (T_{h} or T_{conv}) and Regulatory T cells (T_{reg}). Within the $CD4^{+}$ subset a further division can be made using CD25, which is the α-subunit of the IL-2 receptor (IL-2R) and is highly expressed upon activation whilst being relatively absent on resting $CD4^{+}$ cells. This makes CD25 a useful delineator between *resting* ($CD4^{+}$ $CD25^{-}$) and *activated* ($CD4^{+}$ $CD25^{+}$) cells.

1.10.2 The Regulatory T cell subset

Regulatory T cells (T_{reg}) are a subset of T cells involved in the suppression of the previously described immune response, playing an important role towards the latter stages of infection to prevent tissue damage (Belkaid, 2007), and playing a vital role by preventing any inappropriate immune response to self tissue (Sakaguchi, 2004). When this latter function fails, autoimmune disease such as Type I Diabetes or Multiple Sclerosis may result.

A key hallmark of the regulatory phenotype is the suppression of clonal expansion of conventional (i.e. non-regulatory) T_{h} cells, acting as a direct block to the inflammatory response. Numerous subsets of cells with a regulatory phenotype have been identified, with the best characterised being those found enriched in the sub-population of $CD4^{+}$ T cells which also express the *highest levels* of CD25. These T_{reg} are commonly referred to as $CD4^{+}$ $CD25^{\mathrm{hi}}$ cells and specifically constitute the type under investigation here. However, as this delineation is imprecise, isolation of pure T_{reg} populations is difficult given the similarity of CD25 profiles, and was the source of much controversy over the very existence of T_{reg} until recent times (Sakaguchi, 2004). The recent addition of $CD127^{\mathrm{lo}}$ as a third characteristic for T_{reg} has significantly improved cell isolation strategies (Liu et al., 2006).

Whereas activated T cells utilise IL-2 expression and IL-2R signalling as an integral part of clonal expansion, T_{reg} do not express or secrete IL-2, but are instead dependent on binding of *exogenous* IL-2 to the IL-2R for cellular survival (Setoguchi et al., 2005). Whilst removal of IL-2 from the local micro-environment of pro-inflammatory T cells is one of the many suppressive functions of T_{reg}, the molecular basis for suppression is still poorly understood in many key biological scenarios.

1.10.3 FOXP3 is the Genetic Master Switch

Located on the X chromosome, the gene *Forkhead box P3* (*FOXP3*) encodes a transcription factor, which produces 6 known splice variants, derived from 11 coding exons (Figure 1.4). Two of these transcripts have received experimental characterisation, whilst the remaining four have been identified through predictive methods (The Ensembl Project, 2012). The Forkhead domain is responsible for the DNA-binding of the transcription factor and when in complex with RORα and RORγt, FOXP3 is able to antagonize these transcriptional activators (Du et al., 2008; Zhou et al., 2008). However, the protein encoded by the *FOXP3Δ2* isoform lacks the sequences to interact with these transcription factors which indicates a separate role for this isoform. Whilst both isoforms are found in similar amounts in $\mathrm{T_{reg}}$, the exact role of the shorter transcript, and the dynamics of the two isoforms are yet to be clearly identified.

The *scurfy* mouse strain contains mutations in the *Foxp3* gene (Brunkow et al., 2001), are lacking $\mathrm{T_{reg}}$, and suffers from a lethal autoimmune condition (Godfrey et al., 2005) which is a model for the human disease IPEX (immunodysregulation polyendocrinopathy enteropathy X-linked syndrome) (Wildin et al., 2001). Injection of $\mathrm{CD4^+}$ $\mathrm{CD25^+}$ $\mathrm{T_{reg}}$ to scurfy mice was shown to rescue the condition and permit long-term survival (Smyk-Pearson et al., 2003), leading to the identification of this gene as the genetic master switch for the development of $\mathrm{CD4^+}$ $\mathrm{CD25^+}$ $\mathrm{T_{reg}}$. Likewise, mutations in *FOXP3* were identified as being the source of the human disease IPEX (Wildin et al., 2002) playing a parallel role as the genetic master switch (Yagi et al., 2004). During thymic development, *FOXP3* expression in immature thymocytes leads to them becoming $\mathrm{T_{reg}}$, and continued *FOXP3* expression is essential for maintaining the suppressive phenotype (Williams & Rudensky, 2007). $\mathrm{T_{reg}}$ which develop via this thymic pathway have a stable suppressive phenotype and are commonly referred to as natural $\mathrm{T_{reg}}$ ($\mathrm{nT_{reg}}$).

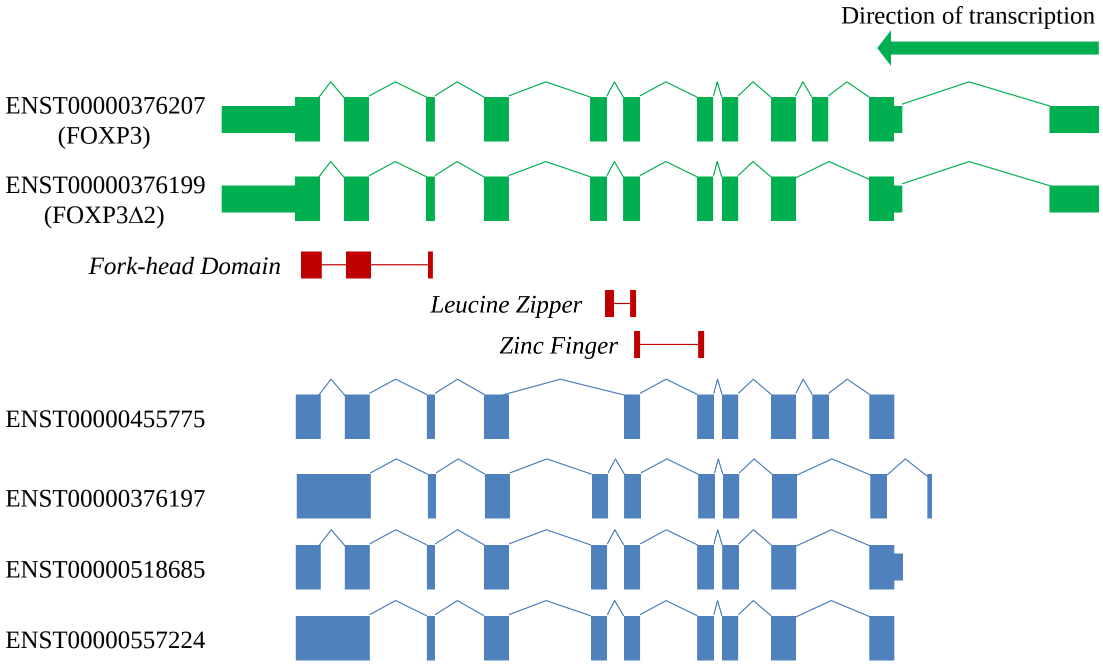


Figure 1.4 – *Known FOXP3 transcripts. The well characterised transcripts are shown in green as the full length transcript (ENST0000037607) and as FOXP3Δ2 (ENST00000376199), along with the known protein domains in red. The more putative transcripts, which all use alternative initiation sites, are shown in blue. Two transcripts contain a retained intron, whilst various combinations of the cassette exons 2 and 7 are also evident.*

1.10.4 The iT_{reg} subset

An additional type of T_{reg} within the larger $CD4^+$ $CD25^+$ subset, are induced in the periphery (i.e. iT_{reg}), possess a more transient regulatory phenotype (Bluestone & Abbas, 2003), but still utilise the FOXP3 pathway (Chen et al., 2003). Deriving from conventional $CD4^+$ $CD25^-$ (T_h) cells (Vukmanovic-Stejic et al., 2006), the mechanisms which give rise to this subset in vivo are still largely unknown, although in vitro treatment with IL-2 and TGF-β is able to generate iT_{reg} from a starting population of naive $CD4^+$ $CD25^-$ cells (Davidson et al., 2007). As *FOXP3* is also transiently expressed in $CD4^+$ $CD25^-$ cells upon TCR stimulation (Walker et al., 2003), the model describing the decision to become an iT_{reg} or remain a T_h cell is yet to be clearly defined.

1.10.5 Cell Surface Markers

A common method of isolating cell populations is Fluorescence Activated Cell Sorting (FACS®), in which an antibody to the surface protein of interest is fluorescently labelled and added to the cells in need of sorting (Julius et al., 1972). Cells on which the antibody is detected are isolated into a separate population via the use of a laser for fluorescence detection, and a physical gating mechanism which responds to the appropriate intensity signals. As such, definitive isolation and identification of T cell subsets requires unique cell-surface molecules acting as biomarkers, and in the case of human T_{reg} these have remained elusive. Further complicating the picture is the fact that CD25 is a marker of T cell activation. A T cell expressing low levels of CD25 can be described as being in a resting state, but these levels will increase rapidly upon TCR stimulation. Thus cells expressing high levels of CD25 will be enriched for T_{reg}, but will additionally contain a number of activated T_h cells. The IL-7 receptor (CD127) has been shown to inversely correlate with *FOXP3* expression (Liu et al., 2006) thus $CD4^+$ $CD25^{hi}$ $CD127^-$ cells are found to be further enriched for T_{reg}, and is the current best practice for isolation of a pure T_{reg} population. However, a single cell-surface marker which can uniquely define a pure population of T_{reg} is yet to be discovered.

1.10.6 Cord Blood

As blood is the primary source for isolating T cell populations, an optimal source for isolating the most pure $\mathrm{nT_{reg}}$ cell populations is umbilical cord blood (Godfrey et al., 2005). The lack of immune exposure to external antigens results in fewer activated, pro-inflammatory $\mathrm{CD4^{+}}$ $\mathrm{CD25^{+}}$ cells and a more distinct $\mathrm{CD4^{+}}$ $\mathrm{CD25^{hi}}$ population when compared to adult blood (Figure 1.5), along with a lack of $\mathrm{iT_{reg}}$ which are generated in the periphery. The selection of these cells with the highest levels of CD25 (i.e. $\mathrm{CD25^{hi}}$) contributes to a greater purity in the final $\mathrm{nT_{reg}}$ populations, and those isolated in this manner have a potent suppressor function (Godfrey et al., 2005).

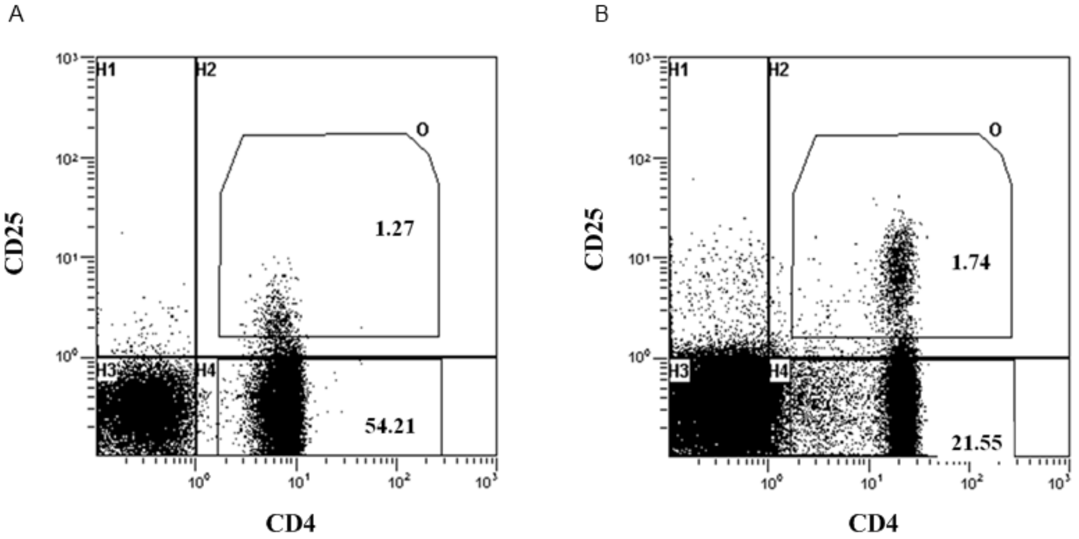


Figure 1.5 – *Representative FACS plots demonstrating the difference in $CD4^{+}$ $CD25^{+}$ populations for samples obtained from from A) adult peripheral blood and B) cord blood. Note the more clearly defined $CD4^{+}$ $CD25^{hi}$ population in the cord blood sample. Image provided by A/Prof Simon Barry.*

1.11 Existing Data

1.11.1 Previous Array Data

Regulatory T cells have been an active area of research since their re-emergence over twenty years ago (Sakaguchi et al., 1995) with genome-wide approaches being utilised to define the unique T_{reg} "signature" of gene expression as the technology became available. A small number of analyses have been performed on human T_{reg} using 3' microarrays (Ocklenburg et al., 2006; Pfoertner et al., 2006) with a much larger pool of array data being generated using murine T_{reg} (Williams & Rudensky, 2007; Pfoertner et al., 2006; Fontenot et al., 2005). There are many differences between murine and human immune biology (Mestas & Hughes, 2004) and as such, the relevance of these studies to human processes may be limited. Prior to commencement of this analysis, no human datasets were available which utilised whole transcript microarrays, and any potential difference in transcript structure between T_{reg} and T_h cells remained an important, but unanswered question. Likewise, RNA-Seq was still an expensive and immature transcriptomic platform.

The first T_{reg} microarray dataset made publicly available through the Gene Expression Omnibus (GEO) repository, was a simple, four-array experiment using a transgenic mouse strain where the *Foxp3* gene had been modified to create a functional Foxp3-GFP chimeric protein (Fontenot et al., 2005). Whilst no biological or technical replicates were included, cells were sorted on $CD25^{hi}/CD25^{lo}$ status and GFP+/GFP- status as an indicator of *Foxp3* expression, with *Foxp3* expression taken as a T_{reg} indicator. Whilst this paper declared over 1100 genes as differentially expressed, subsequent re-analysis (not shown) failed to achieve statistical significance for differential expression of even *Foxp3* itself. Other early datasets possessed a similarly low statistical power (Ocklenburg et al., 2006), were not available as raw data (Pfoertner et al., 2006) or were analysed in a related but highly specific context (Williams & Rudensky, 2007).

1.11.2 FOXP3 ChIP-Chip Data

In addition to the Exon Array dataset which will form an important subject of this work, an in-house ChIP-Chip dataset was generated and made available (Sadlon et al., 2010), in which the binding of human FOXP3 to genomic regions was assessed *in vitro* using expanded cord blood T_{reg}. Data processing and peak detection was performed independently of this thesis by Dr Bridget Wilkinson using the MAT algorithm (W. E. Johnson et al., 2006), and FOXP3 binding peaks were detected with an estimated FDR of 0.5%. Binding peaks were based on genomic co-ordinates, and gene accessions were attributed to a binding site if it was located within 20kb either side of the transcription start site (TSS) or transcription end site for an annotated gene.

1.12 Closing Comments

At the commencement of this body of work, both FIRMA and MM-BGX were yet to be published, and upon publication FIRMA quickly became the most widely adopted approach for detection of alternate splicing between samples. Unclear results from the FIRMA algorithm, as presented in Section 2.5.2, became the motivation for the development of the BMEA model as will be outlined in Section 3.1 and beyond. The publication of MM-BGX occurred when the vast majority of BMEA model development had been completed, and as such was not considered for the initial analysis of the available dataset.

The desire for an unbiased approach, without restriction to previously identified transcripts additionally contributed to the model development stage, as the complexity and range of transcripts associated with some genes may potentially limit the power of approaches such as MM-BGX.

This introductory chapter provides both the mathematical and biological framework for subsequent chapters, and summarises the leading analytic approaches for investigating Exon Array data at the time this work was commenced. The limitations in the models and approaches applied to this type of data, as described in Section 1.7, motivated the subsequent development of the BMEA model from Chapter 3 onwards.

Chapter 2

Data Inspection and Preliminary Analysis

2.1 Introduction

Before development of the BMEA model (Chapter 3), a conventional analysis of the primary Exon Array dataset was undertaken, using RMA/PLM-based strategies at the gene level and FIRMA at the exon level. The results under FIRMA provided much of the motivation for subsequent development of BMEA. This chapter presents this initial analysis, an early version of which was published in Sadlon *et al “Genome-Wide Identification of Human FOXP3 Target Genes in Natural Regulatory T Cells”* (Sadlon et al., 2010) as part of the ChIP-Chip analysis.

Whilst the difference between $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ had been previously addressed by limited datasets in earlier work (see Section 1.11), the key differences making this research unique were four-fold. Firstly, analysis of differential expression will be powered correctly in order to draw a more complete picture of the transcriptome across the relevant cell-types. Secondly, the effects of immune stimulation on $\mathrm{T_{reg}}$ were to be addressed alongside the effects on $\mathrm{T_h}$, which will enable clear identification of the unique $\mathrm{T_{reg}}$ activation signature, and the shared activation signature. Thirdly, the analysis would be not just at the gene-level, but detection of alternate transcript expression was to be an important part of the process. Finally, the relationship of FOXP3 binding sites to differential expression was to be assessed by incorporating the ChIP-Chip data from section 1.11.2. For the remainder of this work, this dataset introduced below will primarily be referred to as the “$\mathrm{T_{reg}}$ dataset”.

2.1.1 The Four Cell Types

For this analysis, both $\mathrm{T_{reg}}$ ($\mathrm{CD4^+}$ $\mathrm{CD25^{hi}}$) and $\mathrm{T_h}$ ($\mathrm{CD4^+}$ $\mathrm{CD25^-}$) cells were prepared when resting and after immune stimulation for 2 hours, in order to capture the early stages of the stimulation response. Cell populations for all four conditions were obtained from the same donors, enabling comparison of any changes in expression within individuals that were consistent across multiple donors, despite expression levels which may be variable across donors. This essentially gave four comparisons (Figure 2.1): a) $\mathrm{T_{reg}}$ vs $\mathrm{T_h}$ (Resting); b) $\mathrm{T_{reg}}$ vs $\mathrm{T_h}$ (Stimulated); c) Stimulated vs Resting ($\mathrm{T_{reg}}$); and d) Stimulated vs Resting ($\mathrm{T_h}$). In order to restrict the $\mathrm{T_{reg}}$ population to the thymically-derived $\mathrm{nT_{reg}}$, cord blood was used as the source biological material.

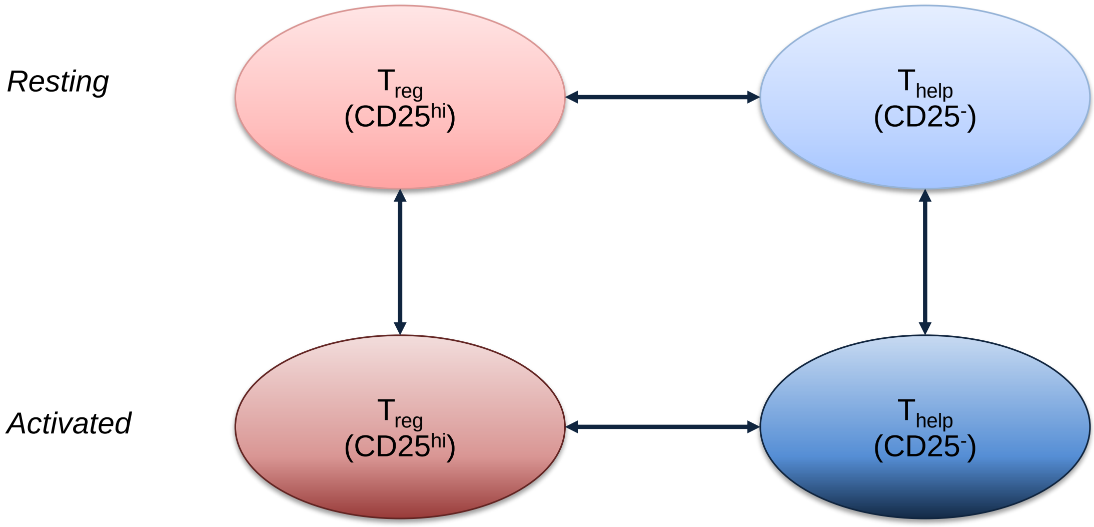


Figure 2.1 – *The four cell types under investigation. Both T_{reg} ($CD4^{+}$ $CD25^{hi}$) and T_h ($CD4^{+}$ $CD25^{-}$) cells were analysed in the resting state ($n = 4$) and after stimulation ($n = 5$) for 2 hours with ionomycin. Differences in sample sizes are described in Section 2.1.2. All cells were sourced from cord blood.*

2.1.2 Cell Preparation

All sample preparation described below and prior to labelling was performed by Dr Timothy Sadlon. Cord blood was obtained for five donors with the approval from the Children's, Youth and Women's Health Service Research Ethics Committee. Following standard protocols (Bresatz et al., 2007), CD4$^+$ CD25$^{\mathrm{hi}}$ and CD4$^+$ CD25$^-$ cell populations were obtained from purified mononuclear cells using a Regulatory CD4$^+$ CD25$^+$ T Cell Kit (Dynabeads; Invitrogen, Carlsbad, CA). Both cell populations were expanded (Sadlon et al., 2010) using CD3/CD28 T cell expander beads (Dynabeads; Invitrogen; catalog no. 111-41D) in order to obtain suitable cell numbers, then rested for 60 hours. Prior to RNA extraction, cells were treated for 2 hours with ionomycin to mimic stimulation or with the empty vehicle (DMSO) to maintain the resting state. During early protocol development, cells from one of the donors (Expansion 41) were treated *with ionomycin only*, providing 5 donor-matched $\mathrm{T_{reg}/T_h}$ comparisons for stimulated cells, and but only 4 donor-matched $\mathrm{T_{reg}/T_h}$ comparisons for resting cells (Sadlon et al., 2010).

2.1.3 RNA Extraction and Hybridisation

RNA isolation was performed using QIAshredder and a miRNeasy Mini Kit (Qiagen), and RNA quality was assayed using an Agilent Systems Bioanalyzer (Santa Clara, CA). Labelling and hybridization to Affymetrix Human Exon 1.0 ST arrays was carried out according to the manufacturer's protocols at the Biomolecular Resource Facility (John Curtin School of Medical Research, Australian National University). Raw data files are lodged on Gene Expression Omnibus (`www.ncbi.nlm.nih.gov/geo/`; accession no. GSE20934) as denoted in (Sadlon et al., 2010) with the final set of .CEL files as described in Table 2.1.

Table 2.1 – *Cell treatments and array designations. Each expansion number corresponds to the same donor. The resting state is sometimes referred to in the text as the "control" state and these terms can be considered as interchangeable in this context.*

Cell Type	**Expansion**	**Treatment**	**CEL File**
T_{reg} (CD25^{hi})	41	Stimulated	TrE41Stim.CEL
	43	Stimulated	TrE43Stim.CEL
		Resting	TrE43Cont.CEL
	86	Stimulated	TrE86Stim.CEL
		Resting	TrE86Cont.CEL
	87	Stimulated	TrE87Stim.CEL
		Resting	TrE87Cont.CEL
	88	Stimulated	TrE88Stim.CEL
		Resting	TrE88Cont.CEL
T_h (CD25$^-$)	41	Stimulated	ThE41Stim.CEL
	43	Stimulated	ThE43Stim.CEL
		Resting	ThE43Cont.CEL
	86	Stimulated	ThE86Stim.CEL
		Resting	ThE86Cont.CEL
	87	Stimulated	ThE87Stim.CEL
		Resting	ThE87Cont.CEL
	88	Stimulated	ThE88Stim.CEL
		Resting	ThE88Cont.CEL

2.2 Data Pre-Processing

2.2.1 Background Correction

Analysis was conducted using the R statistical software environment (R Development Core Team, 2017) and the *aroma.affymetrix* package (Bengtsson et al., 2008), alongside the *limma* (Smyth, 2005) package from the Bioconductor repository (Gentleman et al., 2004). All data pre-processing and gene-level analysis was conducted using v11 of an EntrezGene based CDF file sourced from the University of Michigan (`http://brainarray.mbni.med.umich.edu/www/data-analysis/custom-cdf/`). This CDF contained no exon-level mappings, instead mapping each PM probe to a single gene-level probeset, with ambiguous or poorly mapping probes excluded (Dai et al., 2005). Gene definitions were as defined by the EntrezGene database at the time the CDF was generated (11-Mar-2008) giving 23,536 unique EntrezGene identifiers. No background probes were included on this CDF, as RMA background correction is able to be performed independently of BG probes. The complete dataset (Figure 2.2) was quantile normalised then background corrected using RMA (Irizarry et al., 2003). As no exon-level annotations were included on this CDF, gene-level expression estimates (c_i) were generated using the PLM approach of Equation 1.11.

2.2.2 Quality Assessment

After initial processing, arrays were assessed for quality using Normalised Unscaled Standard Errors (Bolstad et al., 2004) (NUSE) and Relative Log Expression (RLE). Under the NUSE approach, any array with a median NUSE > 1.05 is taken as an indicator of possible poor quality (Silkworth et al., 2008). All arrays passed this criteria (Figure 2.3A).

Inspection using RLE revealed that median values were all ≈ 0, however a subtle trend towards positive values was noted for the 9 T_{reg} arrays (Figure 2.3B), with this being the most notable in the arrays containing Stimulated samples. The inverse of this was also in the set of 9 T_{h} arrays, however no further unusual spread of values was noted, and no arrays were marked for removal.

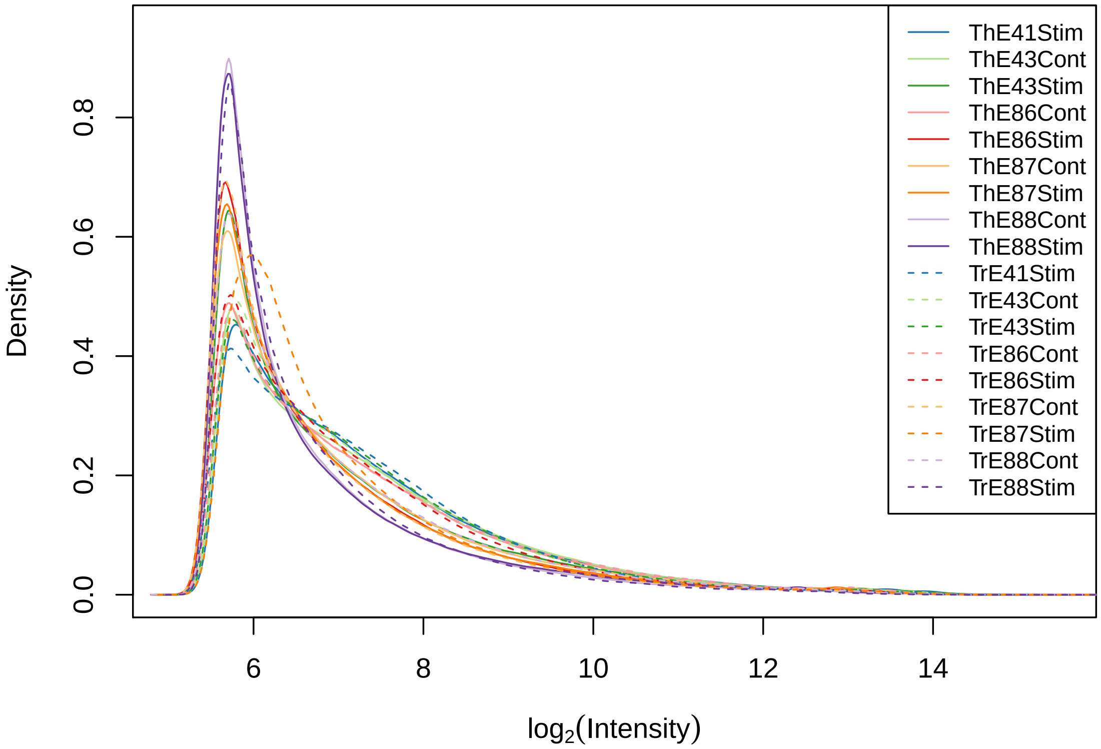


Figure 2.2 – *Perfect match (PM) probe intensity histograms before quantile normalisation, using the full set of PM probes contained on the EntrezGene CDF. T_{reg} and T_h samples are shown as paired samples by colour, with T_{reg} samples shown as dashed lines. Intensity data is shown on the* $\log_2$ *scale.*

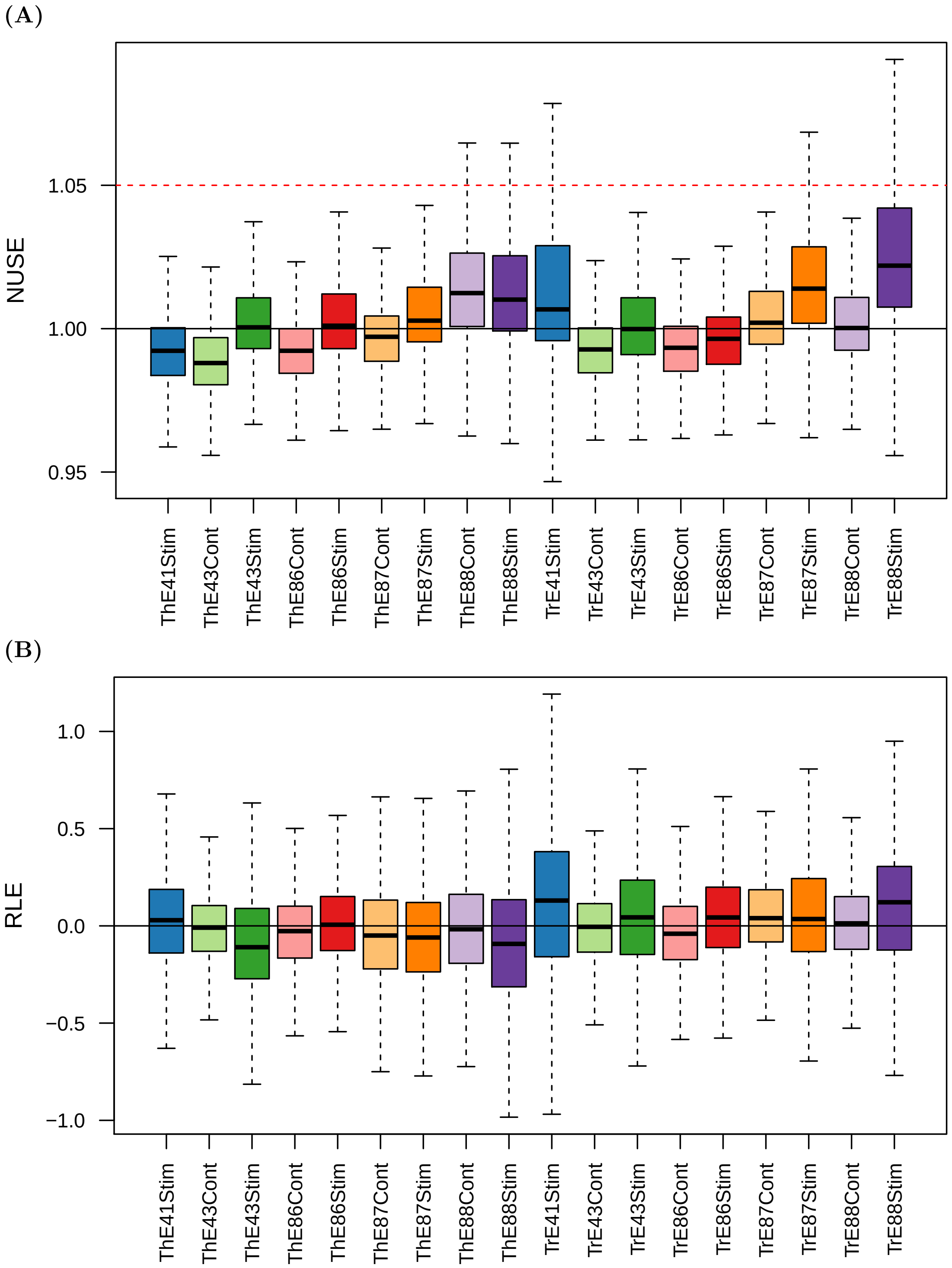


Figure 2.3 – *Quality Control panels using A) NUSE and B) RLE. The median NUSE value for all arrays was below the chosen exclusion threshold of 1.05, and all RLE values appears similarly distributed around zero, with a roughly consistent spread through the inter-quartile ranges. All arrays were deemed to be of acceptable quality.*

2.2.3 Formation of Synthetic Difference-Chips

As the dataset essentially consisted of a series of matched pairs, four sets of data were formed by taking the differences in gene-level expression estimates obtained during RMA fitting (Section 2.2.1), within each donor pair across each of the four comparisons (Section 2.1.1). However, instead of containing expression level data, these datasets contained differences in expression at the probeset-level across the entire array, and are referred to below as difference-chips, and all subsequent gene-level analysis was performed on these sets of data. Identical sets of average expression values were also formed for each pairwise comparison within donors. In essence, these were similar to conventional MA values used in the original two-colour array designs.

2.2.4 Power Calculations

The *sizepower* package (Qiu et al., 2006) was used to assess the statistical power of this dataset (Lee & Whitmore, 2002) to detect an absolute $\log_2$ fold-change greater than 1, under the null and alternate hypotheses:

$$H_0 : \log FC = 0 \;\;\&\;\; H_A : \log FC \neq 0\,.$$

Under the assumptions that 10% of the genes will be truly differentially expressed and that an FDR of 5% will be desired, the value σ_d was estimated as the critical value giving 90% power. Gene-wise standard deviations below σ_d will give a power >90%, whilst standard deviations above σ_d will see a reduction in power.

This value was estimated for sample sizes of both $n = 4$ and $n = 5$, given the varying numbers of donors within the comparisons with $\hat{\sigma}_d = 0.493$ and $\hat{\sigma}_d = 0.551$ for values of $n = 4$ and 5 respectively. Observed standard deviations were plotted against these values (Figure 2.4), with the vast majority of gene-wise standard deviations being below these critical values in all comparisons. The dataset was thus determined to be large enough for the desired results and analysis could proceed.

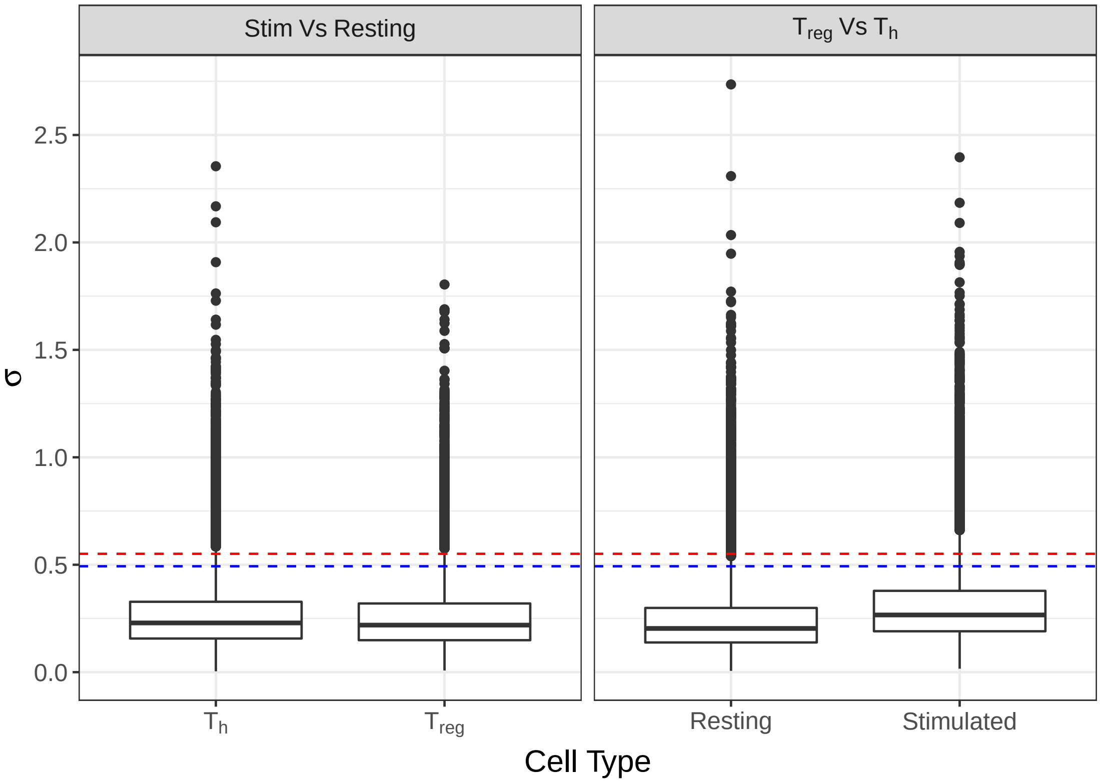


Figure 2.4 – *Boxplots showing standard deviations between donors for estimates of fold-change using the sets of difference chips described in Section 2.2.3. Critical values for giving 90% power are shown in red for* $n = 5$ *(*$\sigma_d = 0.551$*) and blue for* $n = 4$ *(*$\sigma_d = 0.493$*). Only the stimulated* T_{reg} *Vs* T_h *comparison contained the full set of* $n = 5$ *donors, with the remaining comparisons being derived using the remaining sets of* $n = 4$ *donors.*

2.3 Differential Gene Expression

2.3.1 Linear Model Fitting

As they have been shown to improve performance during analysis, array-level weights (Ritchie et al., 2006) for the synthetic difference-chips were calculated within each comparison separately, and a weighted least squares model was fit to obtain the differences in gene expression within each comparison. Moderated $\tilde{t}$-statistics (Smyth, 2004) were calculated for each gene in each comparison as implemented in *limma*.

Inspection of the distributions of these $\tilde{t}$-statistics revealed a strong positive bias in both comparisons involving stimulated $\mathrm{T}_{\mathrm{reg}}$ (Figure 2.5), as may have been expected considering Figure 2.3B. Taking the difference-chips as the M values and using the average expression values within donors as the A values for each comparison, loess normalisation (Dudoit et al., 2002) was conducted within each comparison to correct this apparent bias (Figure 2.5). Array-level weights were recalculated post-normalisation (Figure 2.6) and the analysis was repeated. After this normalisation step, moderated $\tilde{t}$-statistics more closely resembled a symmetric distribution around zero.

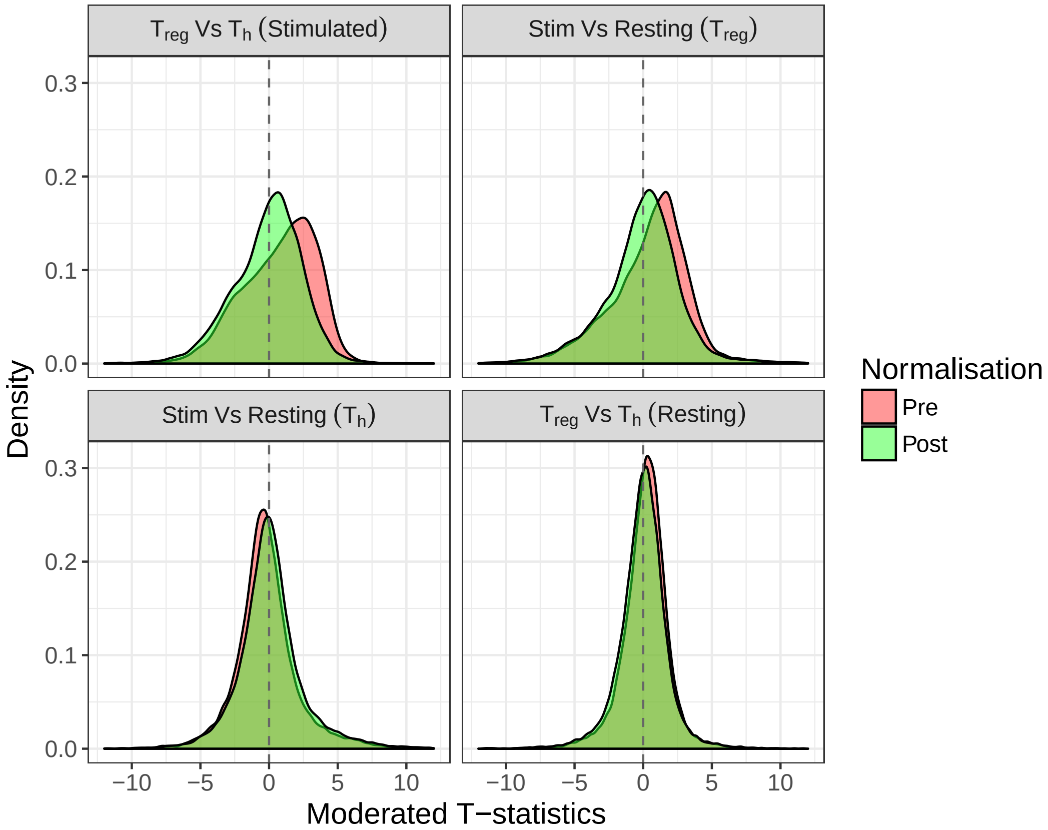


Figure 2.5 – *Moderated $\tilde{t}$-statistics for each comparison before (red) and after (green) loess normalisation. Statistics beyond the range ±12 were omitted to better show the central regions of each distribution. The positive bias seen in both comparisons involving stimulated T_{reg} was strongly reduced by this procedure. Zero is indicated with the vertical grey line in all comparisons.*

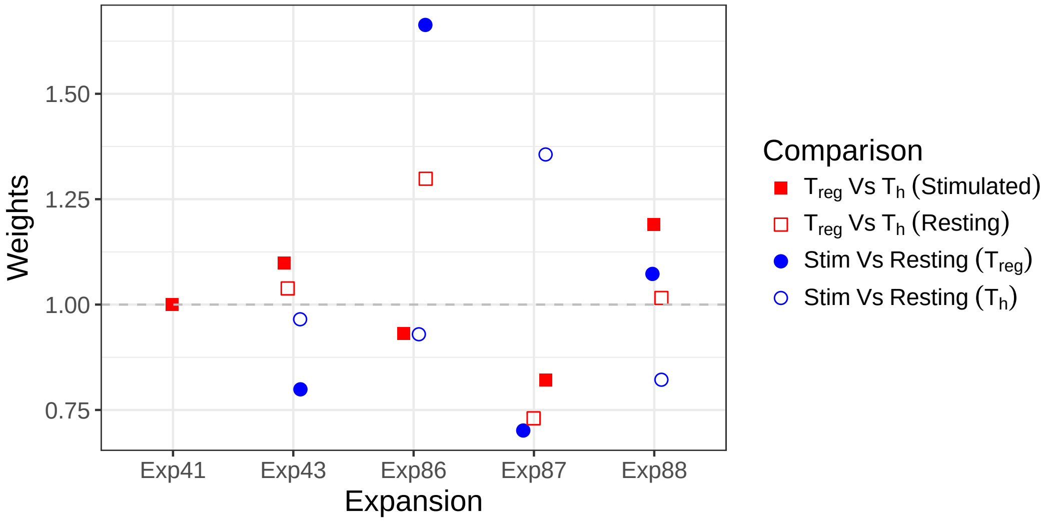


Figure 2.6 – *Weights for each donor across all four comparisons, after loess normalisation. The ideal for equal weighting ($w_i = 1$) is indicated in grey. Jitter has been added along the x-axis.*

2.3.2 Detection of Significant Genes

The p-values obtained from the moderated $\tilde{t}$-statistics were used to generate ranked gene lists for each comparison. Using an FDR-adjusted $p < 0.05$ as indicative of differential expression, large gene lists were obtained in three of the four comparisons, and as such a two-step selection method was used to define differentially expressed (DE) genes.

For each comparison, an initial cut-off p-value (p^0) was chosen based on visual inspection of volcano plots (Figure 2.7). Genes with a raw p-value below this threshold, i.e. $p_g < p^0$, were declared significant regardless of the observed fold-change, and the FDR estimated for p^0. Genes were then additionally included as DE using the dual criteria of an FDR-adjusted p-value < 0.05 and $\log_2$ fold-change (logFC) estimate with an absolute value > 0.7. Under this method, the comparison between resting $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ produced very few results, and as such, the criteria for being considered as DE in this comparison was relaxed. For this comparison the initial inclusion criteria was simply an FDR-adjusted p-value < 0.05, with the additional logFC threshold set at 0.5 for genes with an FDR-adjusted p-value < 0.1.

Using this initial screening method a total of 2805 unique genes were selected across the four comparisons for downstream analysis. The breakdown of DE genes across all four comparisons is presented individually in Table 2.2 with the overlaps between initial lists shown as an UpSet plot (Lex et al., 2014) in Figure 2.8.

Table 2.2 – *Total genes considered as DE in each comparison.*

Comparison	Total
Treg Vs Th (Stimulated)	1201
Treg Vs Th (Resting)	505
Stim Vs Resting (Treg)	1391
Stim Vs Resting (Th)	1107

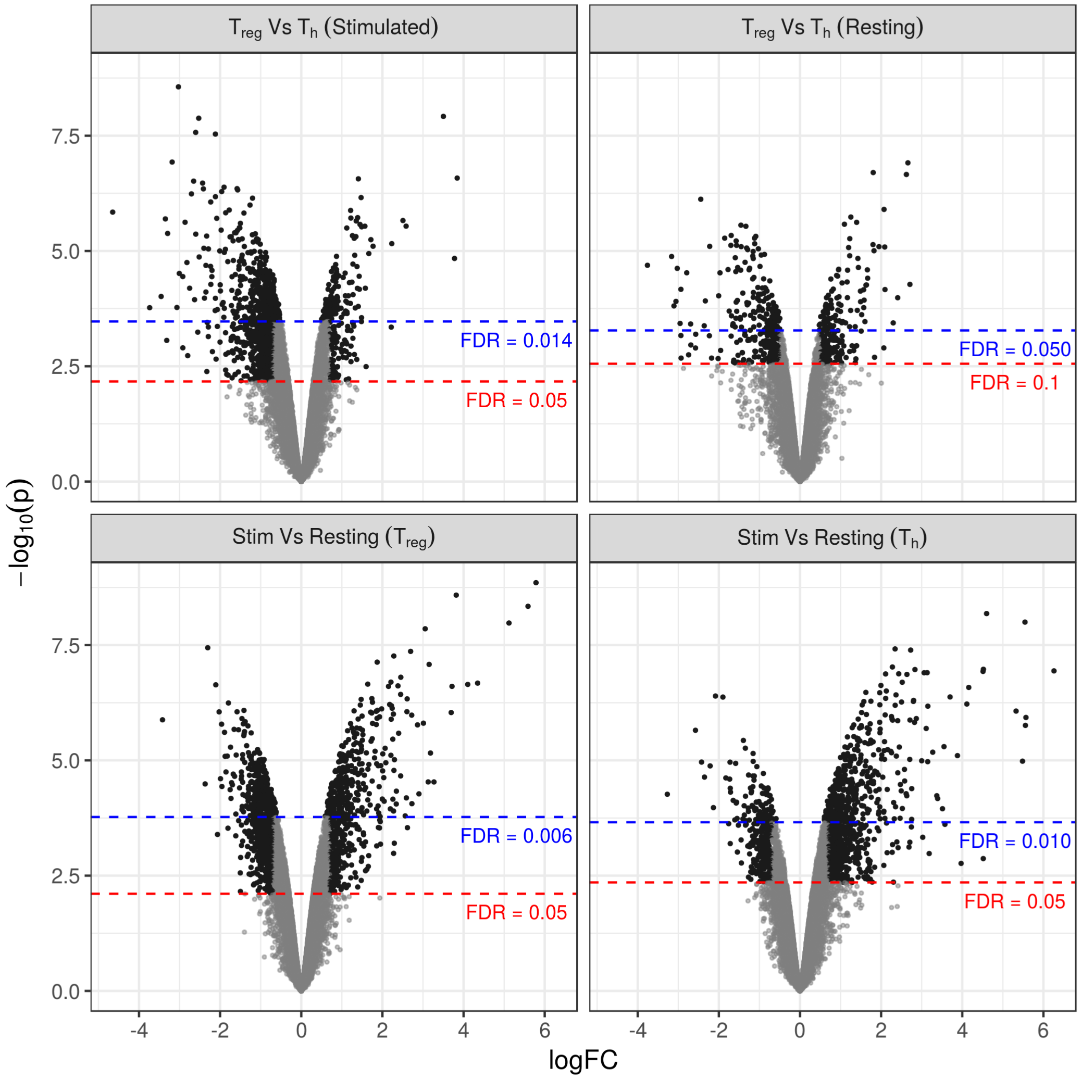


Figure 2.7 – *Volcano plots for all comparisons with genes initially considered as DE indicated in black. The y-axis represents the raw p-value taken to* $\log_{10}$*, thus a higher position on the plot corresponds with a lower p-value. Genes above the first horizontal lines (blue) were automatically included, whereas genes between the blue & red horizontal lines were subject to secondary inclusion criteria based on logFC.*

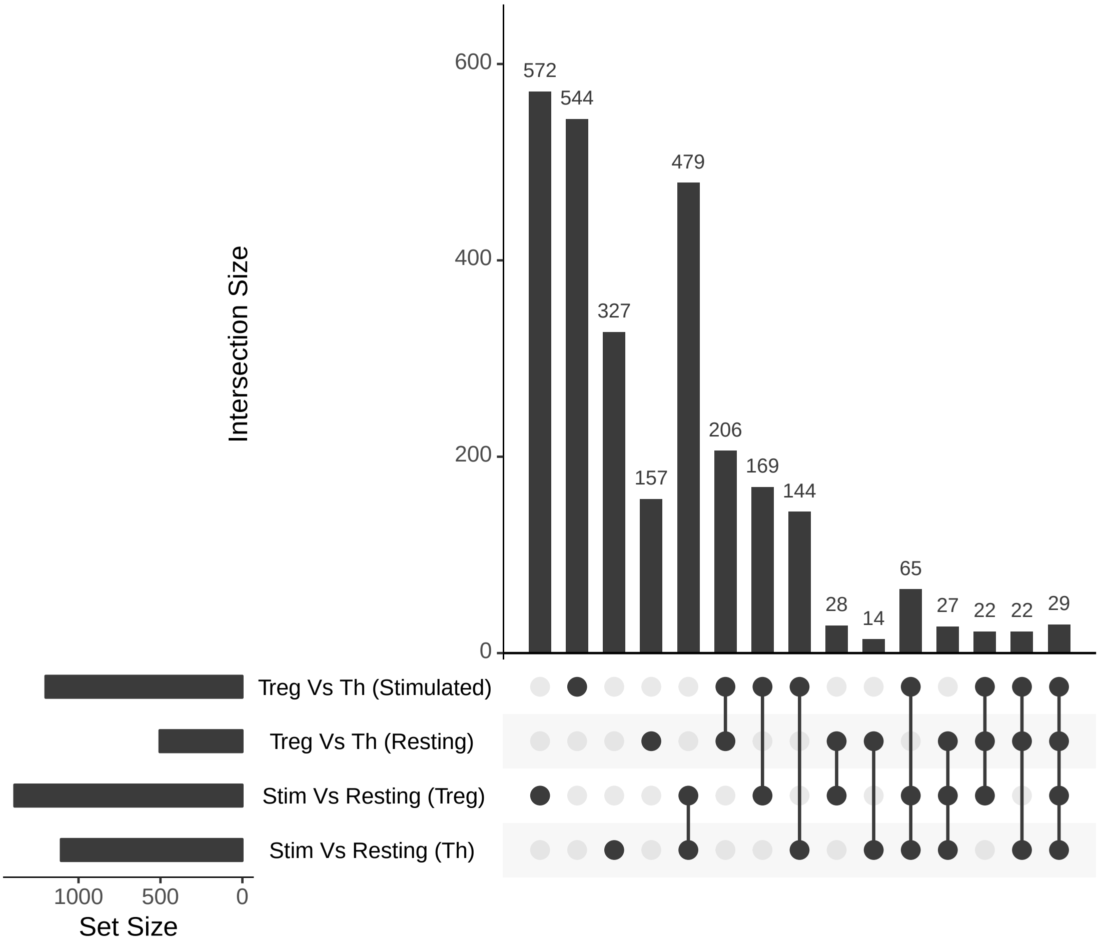


Figure 2.8 – *UpSet plot showing genes initially selected for downstream analysis under the two-stage method described in the text. This represented a total of 2805 unique genes considered as DE in at least one comparison.*

2.3.3 Group Assignment

If a gene is truly differentially expressed in one or more cell types compared to others, this should result in the detection of differential expression across at least two comparisons. For example, if a gene is up-regulated in stimulated $\mathrm{T_{reg}}$ only (Figure 2.9D), this gene should appear to be DE in both the Stimulated Vs Resting $\mathrm{T_{reg}}$ comparison, and the Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison. In the initial UpSet plot (Figure 2.8) there were large numbers of genes considered as significant in only one comparison, which is biologically implausible given this observation. Whilst these genes could have been simply discarded, much of this discrepancy was considered as likely to be the artefact of using hard cut-off values in each comparison. In order to retain the maximum biological information, two steps were taken to address this issue:

1. Genes considered as DE in the initial round of selection had the logFC restrictions removed in all other comparisons
2. For genes still considered as DE in only one comparison, the FDR restriction was loosened and genes were considered as DE in a secondary comparison by choosing an additional comparison with the lowest FDR-adjusted p-value < 0.1

This process left 431 genes which were still only considered as DE in one comparison, and this set of genes was not considered for inclusion into behavioural groups due to a lack of clearly defined expression patterns. The remainder of the genes following more expected patterns (Figure 2.10), with the vast majority showing differential expression across two comparisons, making for simple biological interpretation.

This final dataset (Figure 2.10) was then used to classify the remaining set of 2374 candidate genes into mutually-exclusive groups which defined specific behaviours, with example expression patterns presented in Figure 2.9. In addition to group classification, genes considered as being FOXP3 targets in the ChIP-Chip dataset (Sadlon et al., 2010) (Section 1.11.2) were noted within each group. As described in Table 2.3, these groups were defined as:

- **T1**: The *Common T_{reg} Signature* (Figure 2.9A) is the group of genes DE in both $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparisons, incorporating the 242 DE in only these comparisons, with the addition of another 160 showing activation effects in at least one Stimulated Vs Resting comparison.

- **T2**: The *Resting T_{reg} Signature* (Figure 2.9B) is the group of 105 genes DE in the Resting $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison, but *not* in the Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison. Activation effects were not considered for this group of genes.

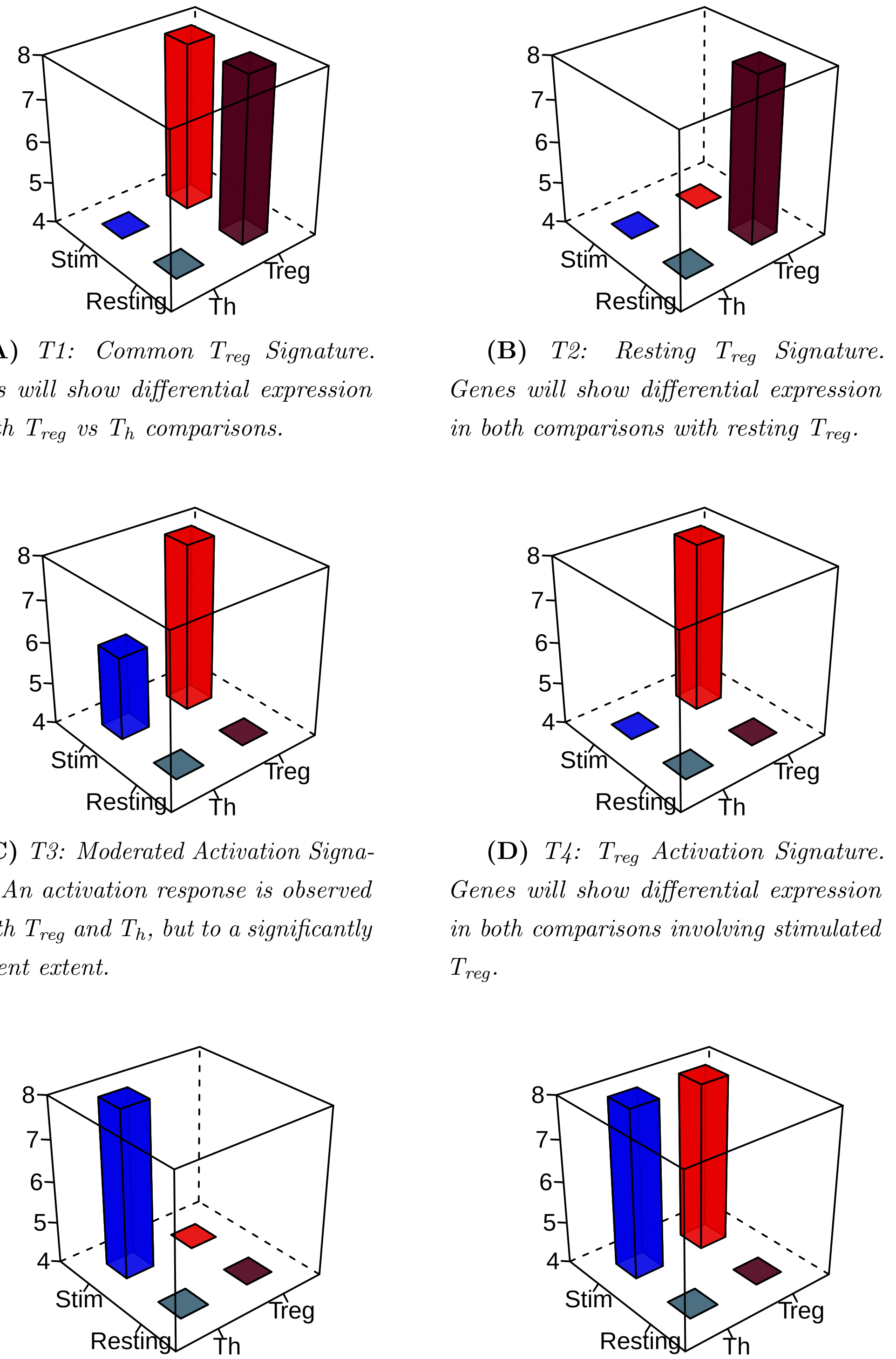
- **T3**: The *Moderated Activation Response* (Figure 2.9C) is the set of genes considered as DE in both Stimulated Vs Resting comparisons, but in which the effect size is significantly different between $\mathrm{T_{reg}}$ and $\mathrm{T_h}$, as defined by genes being additionally DE in one or both of the $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparisons. Any genes previously assigned to groups T1 or T2 were not included here.

- **T4**: The *T_{reg} Activation Signature* (Figure 2.9D) is the set of 581 genes with no $\mathrm{T_h}$ activation effects, but DE between both Stimulated and Resting $\mathrm{T_{reg}}$, and Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$.

- **T5**: The *T_h Activation Signature* (Figure 2.9E) is the set of 345 genes considered as DE in Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ and in Stimulated Vs Resting $\mathrm{T_h}$.

- **A1**: The *Common Activation Signature* (Figure 2.9F) is the shared activation response in both $\mathrm{T_{reg}}$ and $\mathrm{T_h}$, with no significant differential expression in either $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison

The entire set of DE genes as classified into groups is summarised as a heatmap in Figure 2.11, with the top 10 genes from each individual comparison given in Tables 2.4 to 2.7.



(A) *T1: Common T_{reg} Signature. Genes will show differential expression in both T_{reg} vs T_h comparisons.*

(B) *T2: Resting T_{reg} Signature. Genes will show differential expression in both comparisons with resting T_{reg}.*

(C) *T3: Moderated Activation Signature. An activation response is observed on both T_{reg} and T_h, but to a significantly different extent.*

(D) *T4: T_{reg} Activation Signature. Genes will show differential expression in both comparisons involving stimulated T_{reg}.*

(E) *T5: T_h Activation Signature. Genes will show differential expression in both comparisons involving stimulated T_h.*

(F) *A1: Common Activation Signature. Genes will show differential expression in both comparisons of stimulated vs resting.*

Figure 2.9 – *Idealised expression patterns for each group as defined in Table 2.3, and presented in Figure 2.11. Whilst only up-regulation is used for these examples, down-regulation is clearly an equally expected scenario.*

Table 2.3 – *Summary of behavioural groups along with the comparison to the additional set of FOXP3 ChIP-Chip targets.*

Group	Description	# Genes	ChIP Hits	% ChIP
T1	Common T_{reg} Signature	402	207	51.5%
T2	Resting T_{reg} Signature	105	53	50.5%
T3	Moderated Activation Response	177	88	49.7%
T4	T_{reg} Activation Signature	581	177	30.5%
T5	T_h Activation Signature	345	127	36.8%
A1	Common Activation Signature	764	286	37.4%

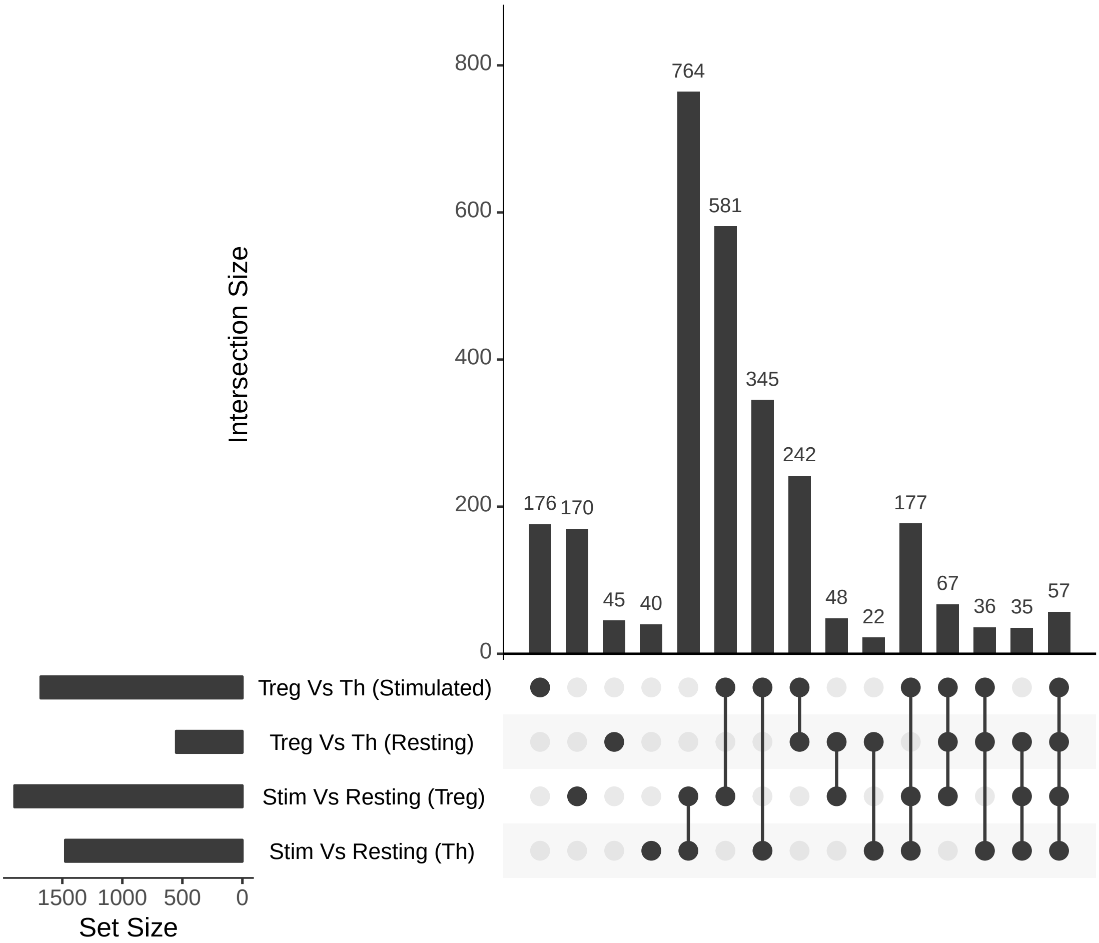


Figure 2.10 – *UpSet plot showing genes after additional steps to correctly describe differential expression across the four comparisons.*

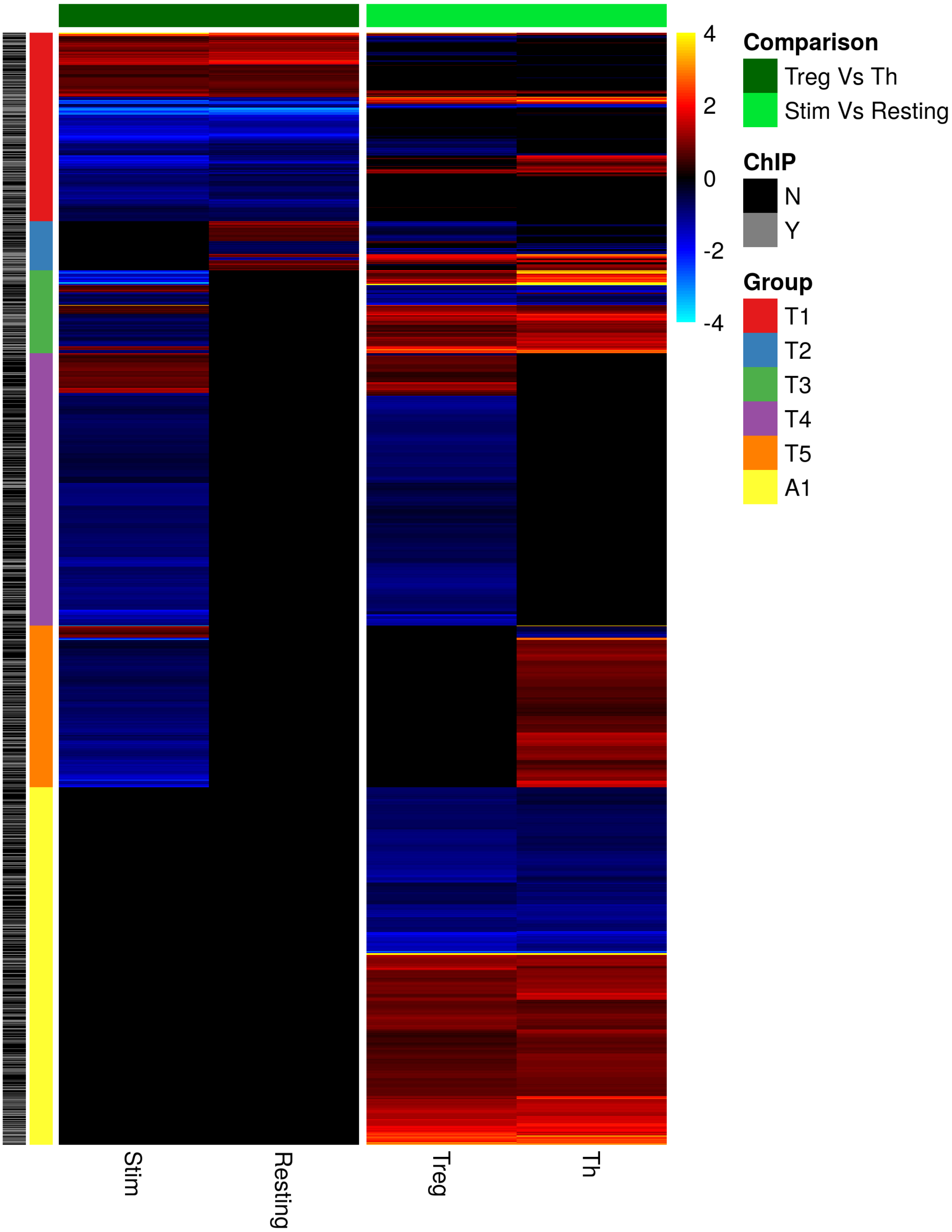


Figure 2.11 – *Heatmap showing group classification for differentially expressed genes, in combination with data from the FOXP3 ChIP-Chip experiment. Group classification criteria is given in Table 2.3. Genes are clustered within each group using the hclust algorithm based on logFC across all four comparisons. Extreme values for log fold-change (i.e. > 4 or <-4) were truncated to ±4 for simplicity of display. Genes and comparisons not considered as DE are shown in black.*

Table 2.4 – *Top 10 DE genes from the Stimulated T_{reg} Vs T_h comparison.*

Gene	Name	Group	ChIP	logFC	T	P	FDR
255231	*MCOLN2*	T1	✓	-3.03	-28.75	2.75e-09	6.48e-05
50943	*FOXP3*	T1	✓	3.50	23.81	1.21e-08	1.04e-04
54806	*AHI1*	T1	✓	-2.53	-23.54	1.32e-08	1.04e-04
3575	*IL7R*	T1	✓	-2.60	-21.49	2.68e-08	1.37e-04
55605	*KIF21A*	T1		-2.12	-21.27	2.91e-08	1.37e-04
79895	*ATP8B4*	T1	✓	-3.18	-17.76	1.18e-07	4.62e-04
2615	*LRRC32*	T1		3.84	16.01	2.62e-07	7.93e-04
51599	*LSR*	T1		1.41	15.93	2.72e-07	7.93e-04
10800	*CYSLTR1*	T1	✓	-2.65	-15.70	3.06e-07	7.93e-04
225	*ABCD2*	T1	✓	-2.43	-15.50	3.37e-07	7.93e-04

Table 2.5 – *Top 10 DE genes from the Resting T_{reg} Vs T_h comparison.*

Gene	Name	Group	ChIP	logFC	T	P	FDR
50943	*FOXP3*	T1	✓	2.66	26.30	1.23e-07	1.72e-03
9734	*HDAC9*	T1	✓	1.80	24.31	1.99e-07	1.72e-03
55930	*MYO5C*	T1	✓	2.62	23.94	2.19e-07	1.72e-03
255231	*MCOLN2*	T1	✓	-2.45	-19.60	7.54e-07	4.44e-03
157506	*RDH10*	T2	✓	2.07	18.05	1.25e-06	5.90e-03
4162	*MCAM*	T1		1.25	16.96	1.84e-06	6.86e-03
3684	*ITGAM*	T1		1.40	16.24	2.39e-06	6.86e-03
3559	*IL2RA*	T1	✓	1.10	15.99	2.63e-06	6.86e-03
92	*ACVR2A*	T1	✓	-1.46	-15.87	2.76e-06	6.86e-03
54103	*GSAP*	T1	✓	-1.33	-15.73	2.91e-06	6.86e-03

Table 2.6 – *Top 10 DE genes from the T_{reg} Stimulated Vs Resting comparison.*

Gene	Name	Group	ChIP	logFC	T	P	FDR
4929	*NR4A2*	A1	✓	5.79	41.71	1.40e-09	3.06e-05
1959	*EGR2*	T3	✓	3.82	38.12	2.60e-09	3.06e-05
8013	*NR4A3*	A1	✓	5.59	35.16	4.54e-09	3.56e-05
3164	*NR4A1*	A1		5.12	31.14	1.05e-08	6.16e-05
84807	*NFKBID*	A1		3.05	29.87	1.40e-08	6.58e-05
474344	*GIMAP6*	T1	✓	-2.31	-26.03	3.60e-08	1.41e-04
1326	*MAP3K8*	T1	✓	2.70	25.35	4.32e-08	1.45e-04
196383	*RILPL2*	A1		2.28	24.51	5.44e-08	1.60e-04
80223	*RAB11FIP1*	T2	✓	1.87	23.44	7.39e-08	1.93e-04
6385	*SDC4*	A1		3.15	23.05	8.27e-08	1.95e-04

Table 2.7 – *Top 10 DE genes from the T_h Stimulated Vs Resting comparison.*

Gene	Name	Group	ChIP	logFC	T	P	FDR
3164	*NR4A1*	A1		4.60	37.33	6.53e-09	1.18e-04
8013	*NR4A3*	A1	✓	5.55	34.98	1.00e-08	1.18e-04
959	*CD40LG*	T1	✓	2.34	28.50	3.82e-08	2.29e-04
196383	*RILPL2*	A1		2.73	28.25	4.05e-08	2.29e-04
80223	*RAB11FIP1*	T2	✓	2.28	24.83	9.41e-08	2.29e-04
84002	*B3GNT5*	T3	✓	4.52	24.40	1.06e-07	2.29e-04
152559	*PAQR3*	T3		2.83	24.35	1.07e-07	2.29e-04
4345	*CD200*	T3		6.26	24.10	1.14e-07	2.29e-04
1959	*EGR2*	T3	✓	4.51	24.00	1.17e-07	2.29e-04
3565	*IL4*	T3	✓	3.13	23.77	1.25e-07	2.29e-04

2.4 Downstream Gene-Level Analysis

2.4.1 FOXP3 Targets

A total of 5578 genes had been identified with FOXP3 bound in the ChIP-Chip dataset (Sadlon et al., 2010), with 4673 of these included on the CDF used for array analysis. These genes were noted during group assignment, and the relative enrichment of these "ChIP Hits" (i.e. FOXP3 targets) within each group was then assessed, with the aim of providing a deeper insight into the role of FOXP3 within each set of DE genes. A logistic regression model (Equation 2.1) was applied, setting the genes not considered as DE as the genomic background, or control set of genes:

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = y_{ij} = \beta_0 + \boldsymbol{\beta^T x}. \tag{2.1}$$

Under this model, the probability π of gene i in group j being a ChIP hit is modelled for the control set of genes (β_0), with group membership from Section 2.3.3 denoted by the set of binary indicator variables in the vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_j)$. Any difference in probability based on group membership is captured by the vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_j)$. A positive coefficient in the set of fitted values ($\hat{\boldsymbol{\beta}}$) will correspond to an increase on the probability of a gene being a ChIP hit for that group in comparison to the background set. All groups were found to be significantly enriched for FOXP3 targets in comparison to the background set of genes (Table 2.8) confirming the important role of FOXP3 in this biological context.

Table 2.8 – *Results of logistic regression analysis, testing enrichment of ChIP Hits against the genomic background (Intercept). Adjusted p-values are shown after using the Bonferroni correction.*

Coefficient	**Group**	**Estimate**	**Std. Error**	**Z-statistic**	**p-value**
β_0	(Intercept)	-1.565	0.018	-85.23	0.00E+00
β_1	T1	1.625	0.101	16.01	1.00E-57
β_2	T2	1.584	0.196	8.08	6.40E-16
β_3	T3	1.554	0.151	10.26	1.05E-24
β_4	T4	0.740	0.092	8.05	8.61E-16
β_5	T5	1.025	0.113	9.06	1.30E-19
β_6	A1	1.052	0.077	13.66	1.70E-42

Simultaneous 95% Confidence Intervals were obtained for the true probability of being a ChIP Hits within each of the groups (Figure 2.12) with clearly similar results obtained for genes in groups T1 to T3, showing nearly half of all genes in these groups to be FOXP3 targets. Surprisingly, the set of genes considered to be the $\mathrm{T_{reg}}$ activation signature (T4), appeared to show slightly lower enrichment for FOXP3 targets than the $\mathrm{T_h}$ activation and Common activation signatures (T5 and A1), and significantly lower enrichment in comparison to groups T1 to T3.

A direct comparison between groups was also performed using simultaneous 95% CIs and omitting the genomic background (Figure 2.13). Groups T1 and T3 were both enriched for ChIP Hits in reference to to T4, T5 and A1, as seen by confidence intervals which excluded zero. Whilst T2 only formally achieved significance in comparison to T4, the direction was again consistent in the T5 and A1 comparisons.

Directional Bias Amongst FOXP3 Targets

As FOXP3 is traditionally regarded as a transcriptional repressor, any directional bias amongst the potential FOXP3 targets (i.e. ChIP hits) was of additional interest. The choice of which comparison to use for consideration as up- or down-regulated is not immediately obvious, and as such, the direction of fold-change was assessed for the following groups and comparisons: 1) T1, both $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparisons; 2) T2, Resting $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$; 3) T3, Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$; 4) T4, Stimulated $\mathrm{T_{reg}}$ Vs Resting $\mathrm{T_{reg}}$; 5) T5, Stimulated $\mathrm{T_h}$ Vs Resting $\mathrm{T_h}$; and 6) A1, both Stimulated Vs Resting comparisons. A simple 2×2 table was constructed for each comparison counting Up/Down regulated genes and ChIP hits against non-ChIP hits. Fisher's Exact Test was then used to test for any association between ChIP hits (i.e. FOXP3 targets) and the direction of fold-change within the specified comparison. All returned p-values were > 0.05 with the exception of A1 ($p = 0.0204$), which in the context of multiple testing was not considered as significant, and no direct adjustment was required. Thus it was concluded that no evidence was found in this dataset for the specific activity of FOXP3 as a transcriptional repressor or activator.

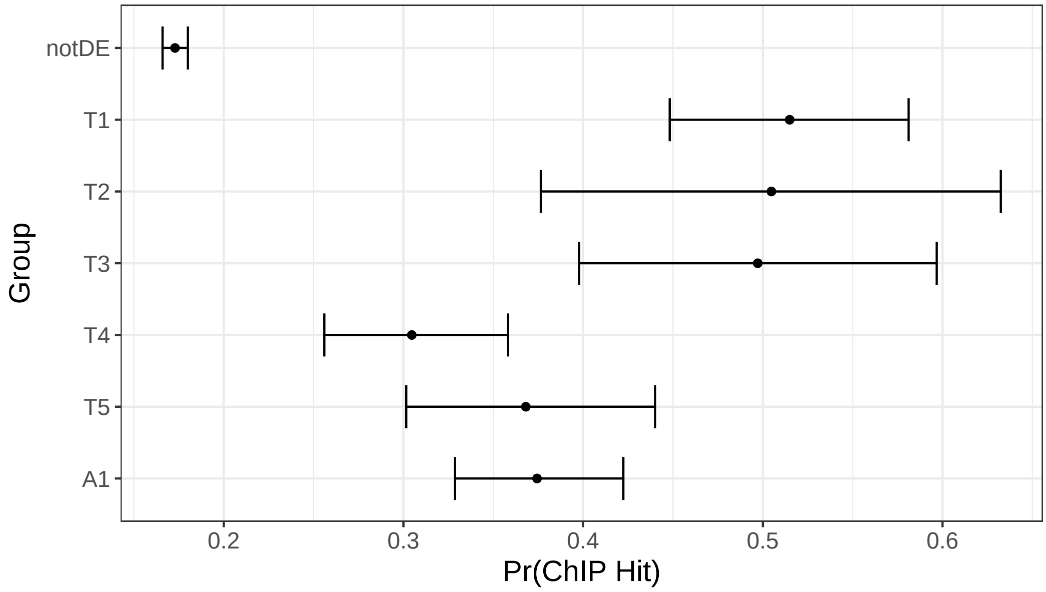


Figure 2.12 – *Simultaneous 95% Confidence Intervals for the probability of a gene being a ChIP hit based on group membership. Simultaneous Intervals were generated using the package multcomp (Hothorn et al., 2008) in order to account for multiple testing considerations.*

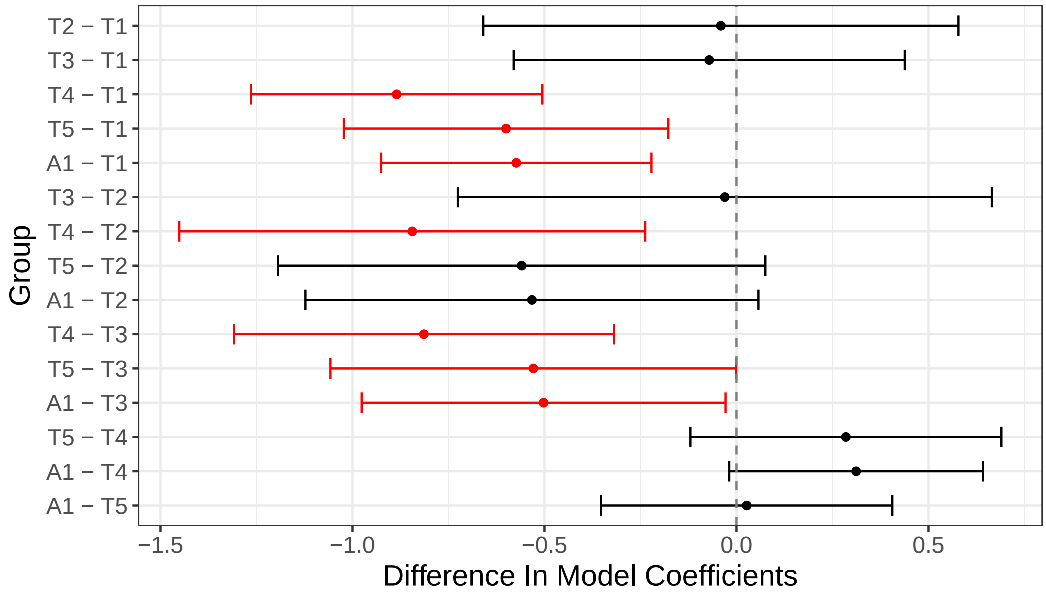


Figure 2.13 – *Simultaneous 95% Confidence Intervals comparing coefficients from Table 2.8 for the probability of a gene being a ChIP hit based on group membership. Simultaneous Intervals were generated using the package multcomp in order to account for multiple testing considerations.*

2.4.2 Gene Ontology Analysis

Gene Ontology Analysis can often produce a large set of highly redundant terms which can be difficult to comprehend from a biological perspective. As an alternative to "term by term" tests for enrichment, targeted search queries were instead used to investigate specific biological behaviours of interest amongst the groups defined in Section 2.3.3. These search queries were motivated by the wider research interests of collaborators and investigated 1) potential cell surface molecules; 2) transcription-related genes; 3) genes with secreted products; 4) genes associated with signalling; and 5) genes associated with RNA metabolism. This enabled characterisation of the defined gene groups in a simple manner. All electronically inferred annotations (IEA) were removed from consideration for these analyses.

Detection of Potential Cell Surface Molecules

Due to the lack of a uniquely-identifying, cell-surface marker for T_{reg}, potential cell surface molecules were identified by searching for those with an appropriate GO identifier amongst the differentially expressed genes. The well-defined surface molecules with CD identifiers (Zola et al., 2007) were used to form a minimal list of 3 GO terms which could be used as a basis for searching the list of DE genes. Of the 64 CD molecules included on the arrays, it was found that a "match-any" search using the terms *GO:0005886 (plasma membrane)*, *GO:0016020 (membrane)* and *GO:0009986 (cell surface)* was sufficient to identify all 64 CD molecules from the complete list.

A total of 6,680 genes were identified as potential surface molecules using this search criterion and 706 had been included as differentially expressed in one of groups (T1-5) or in the common activation group (A1). As a direct result of this work, the important T_{reg} surface molecule PI16 (T1) was identified (Sadlon et al., 2010) and investigations into its role in T_{reg} biology are ongoing (Nicholson et al., 2012; Mohandas et al., 2014). This has since been assigned the additional identifier of CD364.

In addition to the list of potential surface molecules, gene expression groups were assessed for enrichment of genes matching this search. A logistic regression model based on Equation 2.1 was fitted using the same set of genes as the genomic baseline as for Section 2.4.1, with changes in the probability of a gene matching this query assessed for each group relative to baseline (Table 2.9). The impact of a gene being a FOXP3 target (i.e. ChIP Hit) was also included in the model using an interaction term, however interaction terms were not found to be significant and were removed from the model. In comparison to the genomic background set of genes, a gene being included in the common T_{reg} signature (T1) strongly increased the probability of a gene matching this search, as did being a FOXP3 target. In contrast the T_{reg} activation signature was strongly *reduced* in genes matching this query.

Fitted terms for each group were compared to each other using a series of simultaneous 95% Confidence Intervals (Figure 2.14). T1 was enriched for potential surface genes in comparison to all other groups except the Resting T_{reg} Signature (T2), whilst T4 was reduced in these genes in comparison to both the T_{h} and Common Activation signatures (T5 and A1).

Table 2.9 – *Logistic regression results for genes matching the set of GO terms describing potential surface molecules. Results correspond to changes in the logit-transformed probability of a gene being a potential surface molecule in comparison to the genomic background (Intercept). P-values are provided as raw and after Bonferroni adjustment, with significance indicated using asterisks as per the standard conventions of R.*

Term	**Estimate**	**Std. Error**	**z value**	p**-value**	p_{adj}	
(Intercept)	-0.962	0.017	-57.15	0.00E+00	0.00E+00	***
T1	0.558	0.103	5.44	5.46E-08	4.37E-07	***
T2	0.105	0.209	0.51	6.13E-01	1.00E+00	
T3	-0.161	0.170	-0.95	3.43E-01	1.00E+00	
T4	-0.508	0.105	-4.85	1.21E-06	9.71E-06	***
T5	-0.013	0.119	-0.11	9.12E-01	1.00E+00	
A1	-0.041	0.082	-0.50	6.15E-01	1.00E+00	
ChIP	0.315	0.036	8.74	2.31E-18	1.84E-17	***

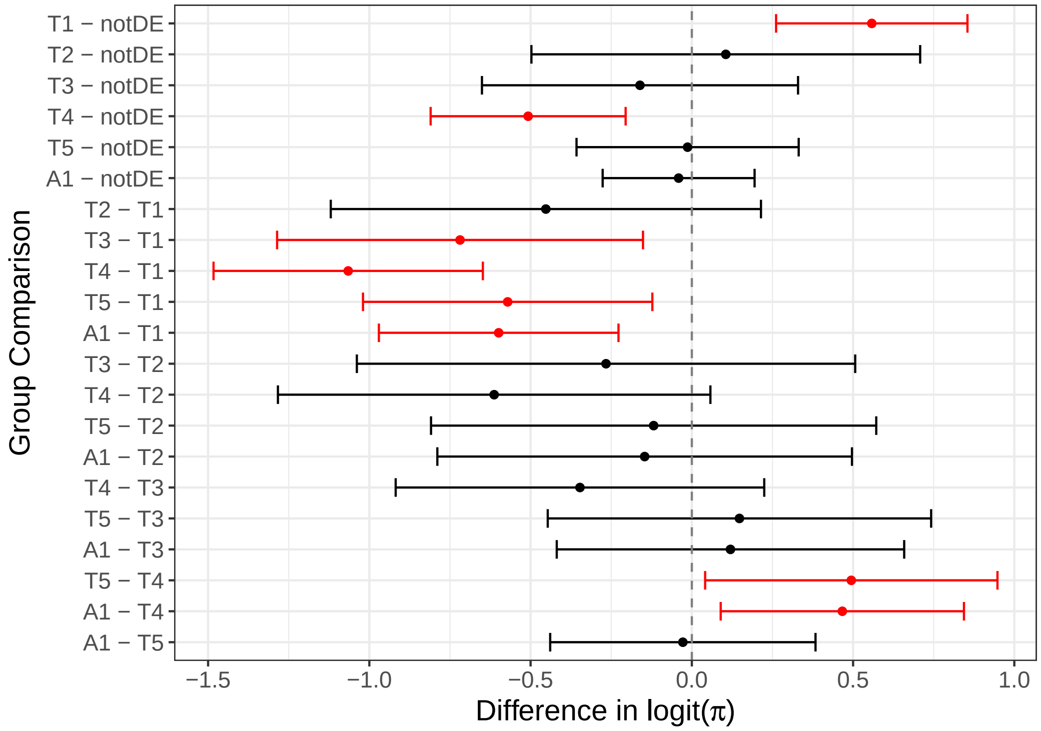


Figure 2.14 – *Simultaneous 95% Confidence Intervals for the probability of a gene being a potential surface molecule based on group membership. Simultaneous Intervals were generated using the package multcomp in order to account for multiple testing considerations. Intervals which exclude zero are considered as supportive of a difference between the two groups, and are shown in red.*

Transcription Related Genes

Using the same principles as above, an alternate search query was formulated to identify genes associated with transcriptional regulation. This contained 8 search terms (Table 2.10), obtained by finding matches to the word transcription amongst all GO term descriptions, then curating manually to obtain a minimal query. A total of 5933 genes matching this query were represented on the EntrezGene CDF.

A logistic regression model was once again fitted assessing FOXP3 targets and groups (Table 2.11). The Moderated Activation (T3), T$_{reg}$ Activation (T4) and Common Activation (A1) Signatures were all enriched for genes matching this query in comparison to the genomic background, as were FOXP3 targets.

Table 2.10 – *GO search terms used for transcription-related genes.*

GO ID	Ontology	Description
GO:0006351	BP	transcription, DNA-templated
GO:0006325	BP	chromatin organization
GO:0016570	BP	histone modification
GO:0005634	CC	nucleus
GO:0003700	MF	transcription factor activity, sequence-specific DNA binding
GO:0043565	MF	sequence-specific DNA binding
GO:0003677	MF	DNA binding
GO:0000988	MF	transcription factor activity, protein binding

Table 2.11 – *Logistic regression results for genes matching the query describing transcription-related molecules. Results correspond to changes in the logit-transformed probability of a gene being a transcription-related in comparison to the genomic background (Intercept). P-values are provided as raw and after Bonferroni adjustment.*

Term	Estimate	Std. Error	z value	p-value	p_{adj}	
(Intercept)	-1.219	0.018	-68.18	0.00E+00	0.00E+00	***
T1	0.198	0.109	1.81	7.06E-02	5.65E-01	
T2	0.161	0.213	0.76	4.50E-01	1.00E+00	
T3	0.478	0.157	3.04	2.35E-03	1.88E-02	*
T4	0.475	0.089	5.36	8.51E-08	6.81E-07	***
T5	0.242	0.118	2.05	4.06E-02	3.25E-01	
A1	0.455	0.078	5.85	4.86E-09	3.89E-08	***
ChIP	0.526	0.036	14.45	2.44E-47	1.95E-46	***

Genes with Secreted Protein Products

The same approach was again used to investigate any potential enrichment of genes with secreted protein products. A search query using the two terms *GO:0005576 (extracellular region)* and *GO:0005125 (cytokine activity)* identified 3807 genes which were present on the CDF.

The results of the logistic regression indicated that only T1 was significantly enriched for this set of genes (Table 2.12), whilst there was a significant *under-representation* of these genes in the activation related groups T4 and A1. The set of putative FOXP3 targets were also comparatively more enriched for genes which matched this search query.

Table 2.12 – *Logistic regression results for genes matching the set of GO terms describing potentially secreted molecules. Results correspond to changes in the logit-transformed probability of a gene being a potentially secreted molecule in comparison to the genomic background (Intercept). P-values are provided as raw and after Bonferroni adjustment.*

Term	**Estimate**	**Std. Error**	**z value**	***p*-value**	p_{adj}	
(Intercept)	-1.646	0.020	-80.38	0.00E+00	0.00E+00	***
T1	0.468	0.117	4.00	6.32E-05	5.06E-04	***
T2	-0.041	0.260	-0.16	8.75E-01	1.00E+00	
T3	-0.094	0.205	-0.46	6.46E-01	1.00E+00	
T4	-0.622	0.140	-4.45	8.64E-06	6.91E-05	***
T5	-0.413	0.166	-2.49	1.29E-02	1.03E-01	
A1	-0.402	0.112	-3.58	3.39E-04	2.71E-03	**
ChIP	0.213	0.044	4.85	1.23E-06	9.83E-06	***

Genes Associated With Signalling

Any group differences for genes associated with signalling were investigated using the 4367 genes matching the single term *GO:0007165 (signal transduction)*. The same model was again fitted with results presented in Table 2.13. Enrichment was found for genes matching this query within groups T1 to T3, as well as the Common Activation Signature (A1). A significant lack of genes matching this query was also noted in the T_{reg} Activation Signature (T4).

Table 2.13 – *Logistic regression results for genes matching the search query for signal transduction. Results correspond to changes in the logit-transformed probability of a gene being a potentially signalling molecule in comparison to the genomic background (Intercept). P-values are provided as raw and after Bonferroni adjustment.*

Term	**Estimate**	**Std. Error**	**z value**	***p*-value**	p_{adj}	
(Intercept)	-1.582	0.020	-79.31	0.00E+00	0.00E+00	***
T1	0.675	0.108	6.23	4.57E-10	3.65E-09	***
T2	0.569	0.213	2.68	7.37E-03	5.89E-02	.
T3	0.502	0.166	3.02	2.55E-03	2.04E-02	*
T4	-0.374	0.121	-3.09	2.00E-03	1.60E-02	*
T5	-0.037	0.139	-0.26	7.92E-01	1.00E+00	
A1	0.396	0.085	4.68	2.80E-06	2.24E-05	***
ChIP	0.442	0.040	11.00	3.82E-28	3.05E-27	***

Genes Associated With RNA Metabolism

The final search query used the terms *GO:0051252 (regulation of RNA metabolic process)*, *GO:0016070 (RNA metabolic process)* and *GO:0019219 (regulation of nucleobase-containing compound metabolic process)*, in order to capture genes associated with RNA metabolism. A total of 3513 genes matched this search query, 483 of which were included amongst the defined groups. The same logistic regression model was again fitted (Table 2.14). In addition to enrichment amongst FOXP3 targets, an enrichment for genes matching this query was noted for the T_h and Common Activation Signatures (T5 and A1), as well as the Moderated Activation Signature (T3).

Table 2.14 – *Logistic regression results for genes matching the search query for RNA metabolism. Results correspond to changes in the logit-transformed probability of a gene being involved in RNA metabolism in comparison to the genomic background (Intercept). P-values are provided as raw and after Bonferroni adjustment.*

Term	**Estimate**	**Std. Error**	**z value**	p**-value**	p_{adj}	
(Intercept)	-1.859	0.022	-84.77	0.00E+00	0.00E+00	***
T1	0.039	0.134	0.29	7.70E-01	1.00E+00	
T2	0.155	0.251	0.62	5.37E-01	1.00E+00	
T3	0.502	0.177	2.85	4.42E-03	3.54E-02	*
T4	0.245	0.108	2.26	2.37E-02	1.90E-01	
T5	0.388	0.133	2.92	3.48E-03	2.78E-02	*
A1	0.369	0.091	4.05	5.17E-05	4.14E-04	***
ChIP	0.475	0.043	11.02	3.04E-28	2.43E-27	***

2.5 Exon-Level Analysis

After the gene-level analysis above, attention was turned to the exon-level analysis. For this stage the exon-level "*groups*" within each gene-level "*unit*" were considered, with each *group* corresponding to a probe selection region as defined by Affymetrix, and with this terminology used to match the defaults of the *aroma.affymetrix* package (Bengtsson et al., 2008). A *group* can be loosely considered to be representative of an exon. A maximum of four probes were present in each group with no limit to the number of groups (exons) within each unit (gene).

2.5.1 CDF Selection and Pre-Processing

The EntrezGene CDF used for the gene-level analysis contained no group-level annotation and as such could not be used for this section of the analysis. A CDF was sourced on-line (http://www.aroma-project.org/chipTypes/HuEx-1_0-st-v2) which utilised the same group structure as the unsupported Affymetrix CDF, however these were mapped to units based on Ensembl annotations (Genome Build 49) instead of the Affymetrix-defined transcript clusters. The limit of 4 probes per group was maintained for this CDF, and as opposed to the EntrezGene CDF, rigorous probe-level quality control was not undertaken during the construction of the file. As per section 2.2.1, the dataset was quantile normalised and background corrected using RMA, with transcript-level estimates being obtained using probe-level modelling.

2.5.2 FIRMA Analysis

FIRMA scores were obtained for each group using the algorithm as implemented in the *aroma.affymetrix* package. Whilst FIRMA was designed primarily as a ranking tool for detection of alternate splicing (AS) events, a model was fitted using the same matched-pairs approach as performed at the gene-level. For each of the four comparisons of interest (Figure 2.1), the differences between the group-level FIRMA scores were calculated and a weighted least-squares model was applied to each set of differences using the *limma* package (Smyth, 2005). Moderated $\tilde{t}$-statistics were obtained for each comparison, and FDR-adjusted p-values calculated.

After eliminating any groups containing fewer than 4 probes, a large number of candidate exons for alternate splicing were detected in the Stimulated T_{reg} Vs T_h comparison using a FDR cut-off set to 5%. However the results for the remaining three comparisons were less compelling (Table 2.15) and shifting the FDR threshold to 10% failed to reveal any candidate AS events for the Resting T_{reg} Vs T_h comparison. This analysis was not pursued any further, but instead became the motivation for development of BMEA in subsequent chapters.

Table 2.15 – *Number of candidate AS groups with adjusted p-values below key FDR thresholds after exclusion of groups with fewer than 4 probes. P-values were adjusted using the Benjamini-Hochberg method and are indicative of the expected FDR.*

Comparison	**FDR** ≤ 0.05	**FDR** ≤ 0.10
T_{reg} Vs T_h (Stimulated)	1980	>2000
T_{reg} Vs T_h (Resting)	0	0
Stim Vs Resting (T_{reg})	0	102
Stim Vs Resting (T_h)	6	137

2.6 Discussion

The above sections have detailed the quality control processes and the retrospective power calculations undertaken to assure an adequate sample size of high quality data. Subsequent analysis revealed large number of genes able to be detected as differentially expressed across three of the four comparisons. These were classified into groups based on their observed behaviour which revealed not only the steady-state T_{reg} signatures (T1 and T2), but the specific activation signature of T_{reg} (T4), the T_h-specific activation signature (T5), the common T_{reg} and T_h activation signature (A1), along with more subtle behaviours (T3).

The Resting T_{reg} Vs T_h comparison was notably less powerful than the other three comparisons, which was surprising given the estimates of gene-level variance shown in Figure 2.4. The significant overlap between the T_{reg} phenotype and the T_h activation response has been well documented (Hyatt et al., 2006) and as the expansion process undertaken in all cell populations involved a level of stimulation, the after effects of this stage may have partially masked the physical distinction between these two cell-types. Unfortunately, due to the low numbers of nT_{reg} obtained in most biological samples, expansion is a required step for increasing total cell numbers to a large enough quantity for analysis.

Integration of the ChIP-Chip results reinforced the vital role of FOXP3 in the T_{reg} phenotype. The Common T_{reg} Signature as well as the Resting T_{reg} Signature were both found to be highly enriched for FOXP3 targets in comparison to the T_{reg} Activation Signature (T4), as was the moderated T_{reg} activation response of group T3. This is of much biological significance, as the T_{reg} Activation signature exclusively consisted of genes considered as DE in the Stimulated T_{reg} Vs T_h comparison (Figure 2.11), which would conventionally be considered as driven by FOXP3. The clear clustering of FOXP3 targets towards the groups defined with more of a phenotypic maintenance role (T1 to T3) was very clear. Whilst this observation strongly reinforces the known role of FOXP3, it suggests that the T_{reg} activation response is not primarily driven by FOXP3 but may instead be orchestrated by one or more as yet undetermined genes.

The targeted GO analysis (Section 2.4.2 and Table 2.16) revealed more interesting insights into T_{reg} biology. The set of Common T_{reg} genes (T1) were clearly enriched for potential surface markers, suggesting that the surface profile of a T_{reg} may indeed be different to that

Table 2.16 – *Summary of Directed GO Term Analysis. Groups enriched for genes corresponding to each search query are indicated with an up-arrow, whilst those exhibiting fewer genes than the genomic background are indicated with a down-arrow. Groups not significantly different to background are indicated with a dot.*

Group	Surface	Transcription	Secreted	Signalling	RNA Metabolism
ChIP Hits	↑	↑	↑	↑	↑
T1	↑	↑	↑	↑	·
T2	·	·	·	↑	·
T3	·	·	·	↑	↑
T4	↓	↑	↓	↓	·
T5	·	·	·	·	↑
A1	·	↑	↓	↑	↑

of a $\mathrm{T_h}$. However, the continuing search for a single defining surface marker suggests that this difference is likely combinatorial, as opposed to being a single defining marker. It is also clear from these results that the response of a $\mathrm{T_{reg}}$ to activation (T4) does not result in a widespread rearrangement of the surface profile (Table 2.9).

Unsurprisingly, many groups were enriched for genes associated with transcription in comparison to the genomic background. This simply indicates that the various cell types and response are driven by clear changes within the transcriptome.

The variable enrichment for genes associated with secretion was perhaps a little surprising, especially considering that two activation-related groups (T4 and A1) were lacking in these genes. However, it may be worth hypothesising that in the case of secreted molecules, many may already exist within the cell either as mRNA or in nascent form, and may only be secreted after a change in the transcriptional profile of the cell, or after changes to the intracellular transport network, enabling faster response to activation. Whilst recent progress has been made, the exact methods of cytokine secretion as a response to activation still remain unclear, and recent work does support this possibility (Gomez & Billadeau, 2008) (Huse et al., 2008). Additionally, this finding does not exclude that possibility that it is a small number of secreted molecules which change in response to activation, and that have vast physiological consequences.

Genes associated with signalling processes were enriched in all groups except the $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ Activation signatures (Table 2.16). Again, the underlying biology behind this is unclear, but this may imply that cell-type specific responses to immune stimulus may not be driven by wide-spread changes to the signalling pathways, but that specific pathways are key players in this response.

The role of RNA metabolism as an immune response is an emerging area of research (Chang & Pearce, 2016), and the enrichment of associated genes across multiple groups reinforces this as an interesting area. The three groups found to be enriched for this set of GO terms (T3, T5 and A1) all have an activation component, again showing consistency with existing research.

An unexpected finding in the above analyses was the reduced enrichment of the $\mathrm{T_{reg}}$ Activation Signature (T4) for genes matching three of the five targeted queries. Whether this represents a true biological phenomenon, or whether the genes in this group represent a poorly characterised set of genes remains an open question. As this work is the first to define this specific expression pattern, the latter may indeed be a possibility. Of additional note, was the observation that FOXP3 target genes were consistently found to be enriched for all GO search queries (Table 2.16). Whilst the implications are not immediately clear, it does reinforce the importance of FOXP3 in orchestrating a wide variety of T cell responses.

Turning to the exon-level analysis, the inconsistent results for the detection of alternate splicing in Section 2.5.2 were cause for much reflection and motivated the analytic model as detailed in subsquent chapters. The methods used in section 2.5.2 were the most advanced available at the time, but failed to reveal any significant changes in transcript structure beyond the Stimulated $\mathrm{T_{reg}}$ vs. $\mathrm{T_h}$ comparison. Whether this discrepancy between the comparisons was a reflection of true alternate splicing remained an unanswered question. If the nearly 2000 events detected in first comparison were an accurate reflection of the underlying biology, then the remaining comparisons had much information which could potentially be revealed. Alternatively, if the number of AS events was as few as had been detected in the remaining comparisons, then the approach utilised had clearly produced many spurious results in the first comparison and was in need of much refinement. As development of the BMEA model became the primary focus, the potential AS events detected above were not pursued any further.

Chapter 3

Development of the BMEA Model

3.1 Development of A New Analytic Model

3.1.1 Context and Motivation

After analysing the dataset at the gene level, attention was turned to the exon (i.e. transcript) level analysis, and this chapter documents the model development from initial explorations to the complete model specification in Figure 3.13.

The FIRMA model as applied to the $\mathrm{T_{reg}}$ dataset (Section 2.5.2) produced less than compelling results. For the three comparisons in which there were four matched samples, the numbers of outlier exons as indicated by FIRMA analyses were very small, suggesting that either there is minimal alternate splicing between the cell types, or that there was not enough statistical power to detect any alternate splicing events. However, the extreme divergence in the number of outlier exons detected in the Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison suggested that there is in fact considerable alternate splicing between the two cell types, and that the addition of one extra paired sample resulted in a marked increase in statistical power. Considering the requirement for significance in two comparisons (Section 2.3.3) these results were unsatisfactory and no experimental validation was considered.

In essence, the approach taken under FIRMA is to fit a model which breaks down under alternate transcript usage, then to look for evidence of a poor model fit, declaring outliers as points of potential interest. Neglected under this approach is the impact of poor model fit on gene-level expression estimates. Any exons included in the CDF annotation, but not present in the biological sample will effectively lead to downward-biased estimates of gene expression as the only contribution brought by these probes will be 100% background signal, or noise.

Given the above results, and the fitting of an often inappropriate model under FIRMA, a more cohesive approach was developed and is described below as *Bayesian Modelling for Exon Arrays* (BMEA) in this body of work.

3.1.2 The Biological Framework

The existence of multiple transcripts presents a challenge for whole-transcript, array-based approaches with many genes expressed as multiple different transcripts within the same cell. A relevant example is FOXP3 which is known to exist in (at least) two isoforms in human $\mathrm{T_{reg}}$, with the primary difference being the omission of a single exon (Du et al., 2008). If both isoforms are equally present in a cell, the skipped exon would be expected to be present in about half of any mRNA sampled. Similarly if the cell contained only one isoform, the exon would be either entirely present or virtually absent, depending on which isoform was present. It is highly plausible that subtle shifts in relative concentrations of an isoform could have marked effects on cellular phenotype, and a model was sought which could capture this type of information by attempting to model the proportion of mRNA transcripts containing each exon, and detecting any changes in these across cell-types.

3.1.3 Modelling Using Exon Proportions

The "additive signal" model in Equation 1.9 is commonly used to correct probe intensities and estimate expression levels in a microarray experiment. In the case of a skipped exon, the observed intensity at a probe targeting that exon would be expected to be zero, thus a suitable term would need to be introduced into the model in *linear space* (prior to log transformation) to allow for this possibility. For a given PM probe, a possible description of the observed signal is

$$PM = B + \phi S\,. \tag{3.1}$$

where $0 \leq \phi \leq 1$ representing the proportion of transcripts containing the relevant exon, and where S represents the "true" signal as would be observed under 100% exon inclusion. If an exon is 100% absent, the observed PM value would reduce to pure background signal ($\phi = 0$), whereas if an exon is contained within every transcript, this will simplify to the initial model of Equation 1.9 ($\phi = 1$).

Conventional probe-level modelling (Equation 1.11) could then be applied to the background-corrected signal term, giving the equation

$$S_{ijk} = \phi_{ij} e^{c_i + p_k + \varepsilon_{ijk}}\,. \tag{3.2}$$

where i and k denote the array and probe respectively, but with the additional subscript j denoting each exon-level grouping within a gene-level meta-probeset. Thus the term 'ϕ_{ij}' represents the proportion of transcripts from a given sample i containing exon j, and the entire signal term can be expressed in log space where $0 < \phi_{ij} \leq 1$ as:

$$\log S_{ijk} = \log \phi_{ij} + c_i + p_k + \varepsilon_{ijk} \,. \tag{3.3}$$

Whilst this shares much with the MIDAS approach from Equation 1.13, MIDAS fits the model first, then normalises to the probe with maximum signal to calculate the Splicing Index. This is in direct contrast to BMEA which will attempt to obtain accurate measures of expression and exon-inclusion by estimating ϕ_{ij} during the model fitting stage. Notably, methods such as FIRMA have already been shown to outperform the MIDAS approach (Purdom et al., 2008).

If the above model were fitted after background correction using a least-squares approach, it would be possible to obtain an estimate for $\log \phi_{ij}$ which is > 0 (i.e. $\phi_{ij} > 1$) and not representative of the true parameter, leaving options to normalise the model as in the MIDAS approach (Affymetrix, 2005a). Additionally, background correction methods such as RMA or GC-RMA assume that the signal component is non-zero, whilst the PLM approach was developed where all probes within a probeset measure a broadly similar amount of signal, as for 3' arrays. The development of a model in a hierarchical, Bayesian context can conceptually overcome all of these shortcomings.

Equation 3.3 can alternatively be expressed in a Bayesian framework by

$$\log S_{ijk} \sim \mathcal{N}(\eta_{ijk}, \sigma_S) \tag{3.4}$$

with mean $\eta_{ijk} = \log \phi_{ij} + c_i + p_k$, as defined in equation 3.3.

In the model specifications below, the terms *"exon-level probeset"* and *"exon"* are used interchangeably. Similarly, the term *"probeset"* is occasionally used to denote exon-level probesets. As all specifications are given within a single gene, gene-level meta-probesets are the assumed context, with any subscripts referring to this level (i.e. g) commonly suppressed.

3.2 Initial Simulations

3.2.1 A Prior Distribution for ϕ_{ij}

In order to begin the specification of the complete Bayesian model, a prior distribution for ϕ was required, with the restriction that $0 \leq \phi \leq 1$. A flexible set of distributions which satisfy this property, minus the boundary points, are the Beta family distributions.

The Beta Distribution

The two shape defining parameters for the Beta distribution are commonly defined as α & β. Setting both of these to $\alpha = \beta = 1$ gives the Uniform distribution on the interval (0, 1), whilst increasing both together and maintaining equality gives an increasingly narrow bell-shaped distribution around the value 0.5. Increasing the parameter $\alpha > 1$ whilst holding $\beta = 1$ yields an increasing probability weight towards 1, with the reverse of this shifting the probability weight towards zero (Figure 3.1).

Selection of a Beta Prior

One approach under early consideration was to incorporate the information within genomic databases, and increase the value of the parameter α as the number of defined transcripts containing an exon increases. This would weight the probabilities towards 1 for exons which are contained in the majority of transcripts, but provide a more even spread across the range for those more frequently spliced out.

Alternatively, some exons may even be considered as *"constitutive"* if present in every transcript of a gene, whilst others may be considered as potentially-spliced, or *"non-constitutive"* exons. Constitutive exons could be given a beta prior, based on the number of known transcripts, or could even be given a prior strictly equal to the value 1. Non-constitutive exons could then be given a beta prior based on the number of transcripts they are identified in, or more conservatively, could be given a Uniform prior across the range (0, 1). Five possible combinations of these were explored during model development as summarised in Table 3.1.

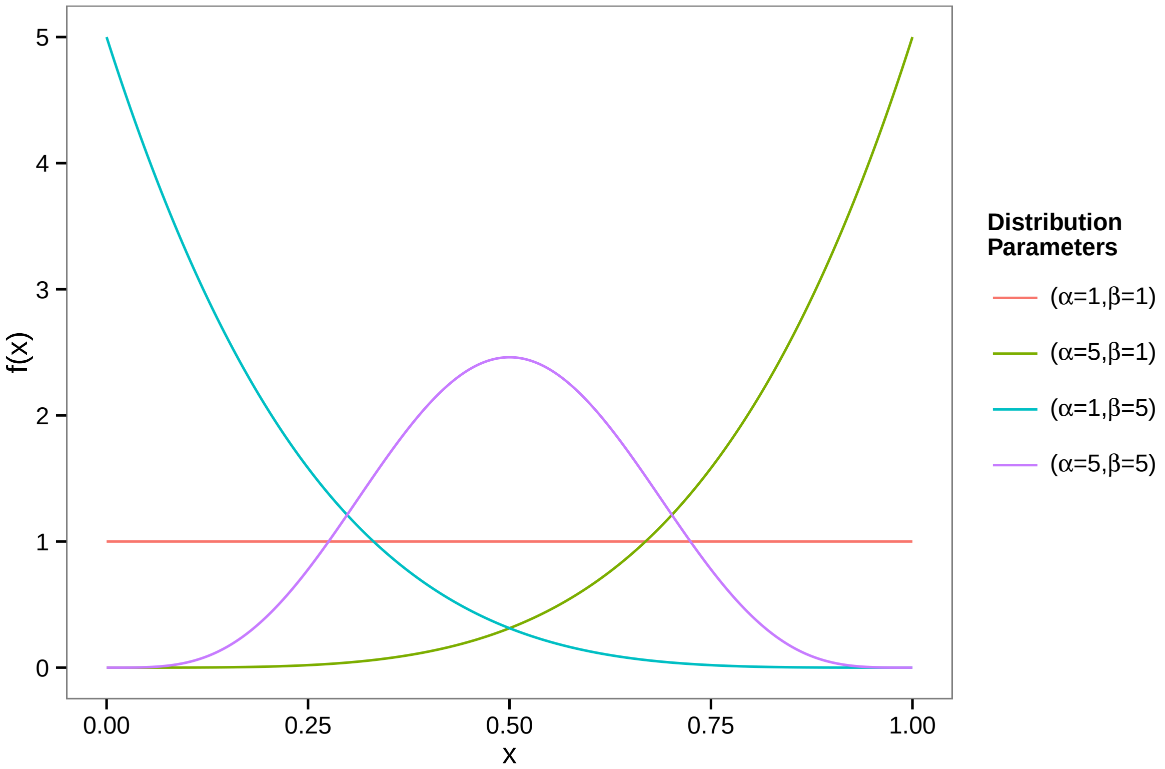


Figure 3.1 – *Examples of probability distribution functions for the Beta distribution under varying combinations of the shape parameters α and β.*

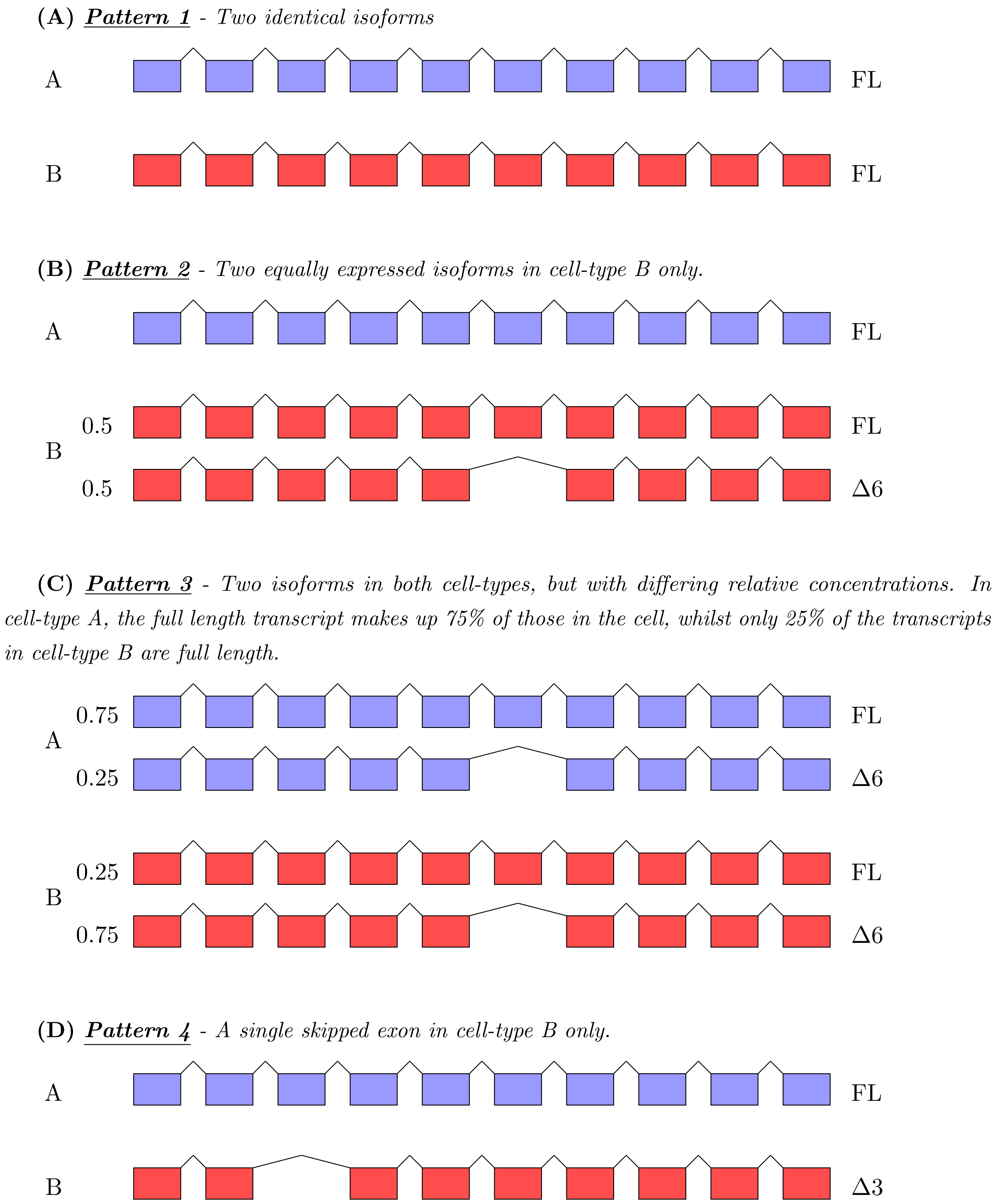
Table 3.1 – *The set of alternative specifications of the prior distributions for the exon-level terms the model presented in Equation 3.2. The value N represents the number of known transcripts for a gene, whilst the value n_j represents the number of known transcripts containing exon j. For an exon defined as constitutive $N = n_j$, whilst for a non-constitutive exon, $N > n_j$.*

Abbreviation	**Constitutive Exons**	**Non-Constitutive Exons**
1B	1	$\alpha = n_j + 1; \beta = N - n_j + 1$
BB	$\alpha = N + 1; \beta = 1$	$\alpha = n_j + 1; \beta = N - n_j + 1$
1U	1	$\alpha = 1, \beta = 1$
BU	$\alpha = N + 1, \beta = 1$	$\alpha = 1, \beta = 1$
UU	$\alpha = 1, \beta = 1$	$\alpha = 1, \beta = 1$

3.2.2 Splicing Patterns For Simulated Data

In order to test the priors from Table 3.1 a set of simulated data was generated for 1000 genes. A hypothetical 10-exon gene (Figure 3.2) with 8 different splicing patterns across two cell-types was included as part of the simulation model. These patterns were designed to test a variety of alternate splicing possibilities such as changes in the concentration of two very similar isoforms (3.2B & 3.2C), single-exon skips (3.2D, 3.2E & 3.2F) and large truncations (3.2G & 3.2H). A pattern was also included with full length transcripts in both cell-types (3.2A) as a negative control. As the relative levels of the two primary FOXP3 isoforms are still poorly characterised in $\mathrm{T_{reg}}$, patterns 2 & 3 were of specific relevance to the underlying biology in this body of work.

This set of splicing patterns gave a total of 6 simulated transcripts, which would allow testing of various priors during the model fitting steps. Data were simulated with no variation in alternate splicing patterns within a condition, effectively enabling the sample-specific value for exon-inclusion proportions ϕ_{ij} to be simulated in a condition-specific manner as ϕ_{hj}.

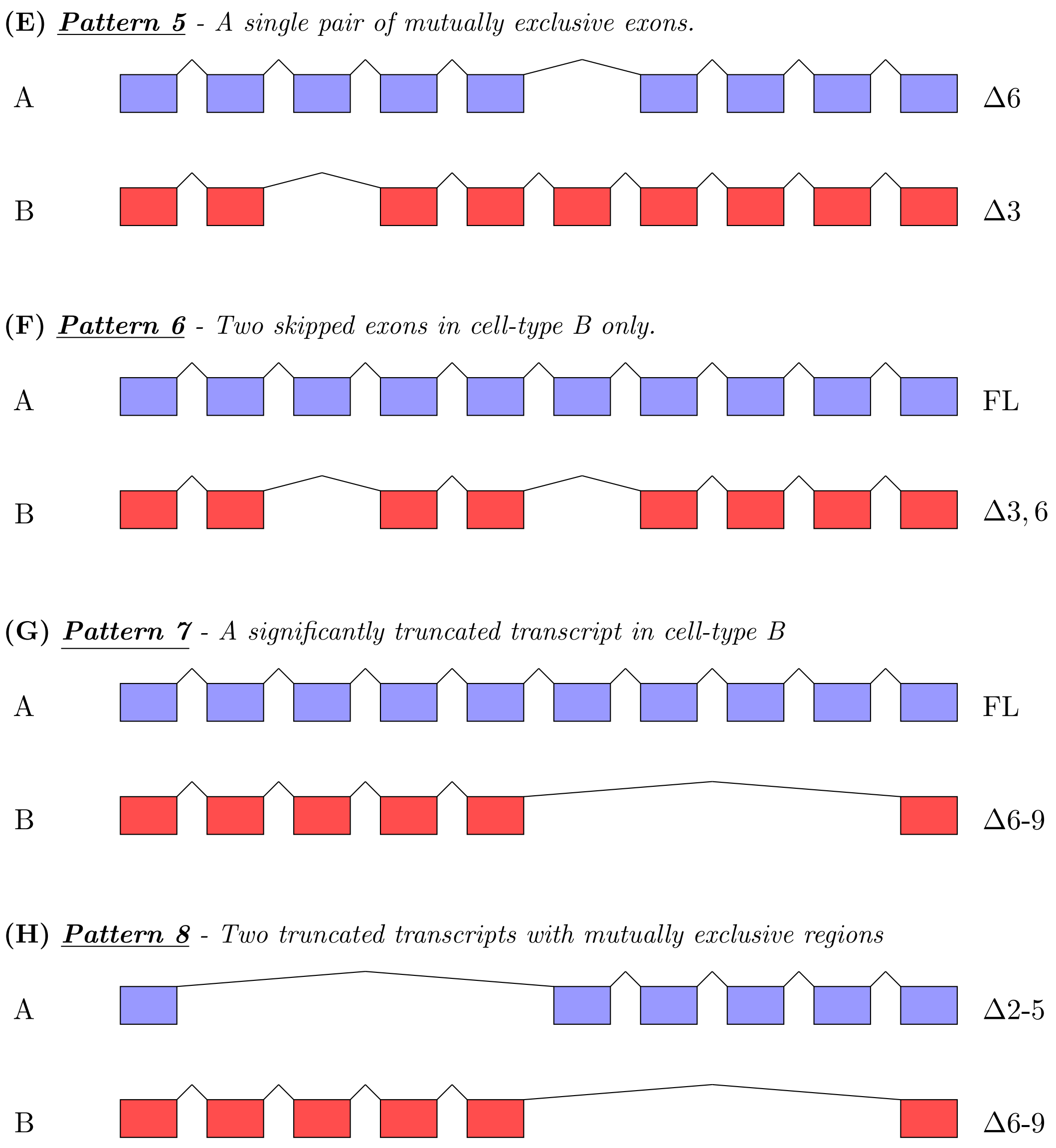


(A) ***Pattern 1*** - *Two identical isoforms*

(B) ***Pattern 2*** - *Two equally expressed isoforms in cell-type B only.*

(C) ***Pattern 3*** - *Two isoforms in both cell-types, but with differing relative concentrations. In cell-type A, the full length transcript makes up 75% of those in the cell, whilst only 25% of the transcripts in cell-type B are full length.*

(D) ***Pattern 4*** - *A single skipped exon in cell-type B only.*



(E) ***Pattern 5*** - *A single pair of mutually exclusive exons.*

(F) ***Pattern 6*** - *Two skipped exons in cell-type B only.*

(G) ***Pattern 7*** - *A significantly truncated transcript in cell-type B*

(H) ***Pattern 8*** - *Two truncated transcripts with mutually exclusive regions*

Figure 3.2 – *Hypothetical splicing patterns for the 10-exon gene used in simulated datasets. The two theoretical cell-types are shown in blue (cell-type A) and red (cell-type B). A total of 6 different transcripts were defined (FL; Δ3; Δ6; Δ3,6; Δ2-5 & Δ6-9) for the specific splicing behaviours under investigation.*

3.2.3 Additional Parameters for Initial Simulations

In addition to the splicing patterns described above (Figure 3.2) each gene was defined as containing four probes per exon, in keeping with the design of Affymetrix Exon Arrays. To replicate the structure of the T cell dataset under investigation, data was simulated as four complete sets of matched pairs ($n = 4$). Data for the 1000 simulated genes was then generated using a broadly similar parametrisation to that in the presentation of the FIRMA model (Purdom et al., 2008), where values were simulated for background signal with $\log_2 B_{ijk} \sim \mathcal{N}(5, 0.35)$. The overall expression level for each array was generated using $c_i \sim \mathcal{U}(6, 12)$, probe effects were derived from $p_k \sim \mathcal{N}(0, 3)$ and general noise added as $\varepsilon_{ijk} \sim \mathcal{N}(0, 0.7)$. Data were generated on the $\log_2$ scale using the following algorithm for each simulated gene to give correlations as might be observed in paired samples, as is under investigation in this work.

Simulation Algorithm

1. Randomly sample an initial expression value μ between 6 and 12 using

$$\mu \sim \mathcal{U}(6, 12)$$

2. Randomly sample a value σ_B for biological variability from

$$\sigma_B^2 \sim \text{Scaled Inv-}\chi^2(3.6, 0.3)$$

3. Randomly sample pair-specific expression values v_i for each pair $i = 1, 2, 3, 4$ using

$$v_i \sim \mathcal{N}(\mu, \sigma_B); 0 < v_i < 16$$

4. Randomly sample log fold-change (logFC $= \Delta\mu$) from

$$\Delta\mu = \begin{cases} 2 & \text{with prob.} = 0.1 \\ 1 & \text{with prob.} = 0.1 \\ 0 & \text{with prob.} = 0.7 \\ -1 & \text{with prob.} = 0.1 \end{cases}$$

5. Using the value $\sigma_T = 0.25$ to represent technical variability within a donor pair, randomly sample pair-specific expression estimates c_{hi} for treatments A ($h = 1$) & B ($h = 2$) from

$$\begin{aligned} c_{1i} &\sim \mathcal{N}(v_i + \Delta\mu, \sigma_T); \quad && 0 < c_{1i} < 16 \\ c_{2i} &\sim \mathcal{N}(v_i, \sigma_T); && 0 < c_{2i} < 16 \end{aligned}$$

6. Sample probe affinities p_k for $k = 1, 2, \ldots, 40$ from

$$p_k \sim \mathcal{N}(0, 3)$$

7. Sample one of the 8 possible splicing patterns as described in Figure 3.2, with equal probability to provide values for ϕ_{hj} for $h = 1, 2; j = 1, 2, \ldots, 10$, from the set $P = \{0, 0.25, 0.5, 0.75, 1\}$ as appropriate for the sampled splicing pattern.

8. Using the value $\sigma_S = 0.7$ to represent general noise, obtain final simulated values for true signal S_{hijk} using:

$$\log_2 S_{hijk} \sim \mathcal{N}(\eta_{hijk}, \sigma_S); \quad 0 < \log_2 S_{hijk} < 16$$

where $\eta_{hijk} = \log_2 \phi_{hj} + c_{hi} + p_k$.

9. Sample background signal B_{hijk} from:

$$\log_2 B_{hijk} \sim \mathcal{N}(5, 0.35)$$

10. Obtain final simulated PM_{hijk} values

$$PM_{hijk} = B_{hijk} + S_{hijk}$$

3.2.4 Fitting Initial Simulated Data

After generation of the initial simulated dataset, several priors for ϕ were investigated using the stand-alone software package *WinBUGS* (Lunn et al., 2000a) and the R package *R2WinBUGS* (Sturtz et al., 2005a). The splicing patterns defined in Figure 3.2 had specified six hypothetical transcripts, i.e. $N = 6$, giving each exon a different number of transcripts (n_j) to which it belonged. Priors were assigned to each ϕ_{ij} as described in Table 3.2 following the principles described in Table 3.1, with two additional *misspecified models* included to assess results if known genomic annotations were inaccurate or incomplete. These models reversed the priors for constitutive and non-constitutive exons (model UC), or followed the BU model but without knowledge of the Δ2-5 transcript (BU245). This latter model would erroneously classify exons 2, 4 & 5 as constitutive, but would still correctly define exons 3 & 6-9 as non-constitutive.

For the purposes of model exploration, the remaining model parameters were assigned priors as defined in Table 3.3. The values PM_{hijk} were fitted as representative of background corrected data, with simulated background signal considered to be representative of residual noise after conventional background correction methods such as RMA or MAS5.0.

Posterior distributions for the parameters c_i, p_k and $\log_2 \phi_{ij}$ were obtained after WinBUGS was run with 2 chains for 20,000 iterations, with the first 10,000 being discarded as the burn-in period. Model convergence was checked using $\hat{r}$ (Gelman et al., 2004) for the set of fitted parameters, with the mean $\hat{r}$ value for each simulation being < 1.06 in 99% of runs across all models, indicating relatively strong convergence.

To provide an initial guide to model performance, the posterior means of the expression estimates c_i were used to obtain estimates of $\log_2$ fold-change. In order to provide a comparison to the proposed models, the full set of simulated data was also fitted using probe-level modelling (Equation 1.11), giving a comparative set of $\log_2$ fold-change estimates.

Likewise, the posterior means of $\log_2 \phi_{ij}$ were used to obtain estimates of the difference in exon inclusion rates. For PLM fitted data, FIRMA scores were calculated for each exon and estimates of the differences in FIRMA scores were also calculated. Both approaches were fit at the gene and exon levels using *limma* to obtain moderated $\tilde{t}$-statistics, FDR-adjusted p-values.

Table 3.2 – *Prior distributions assigned to each exon under the 7 different models under investigation. Under the correct model specifications, only exons 1 & 10 are constitutive, whilst all other exons are alternate splicing candidates. The number of transcripts containing each exon are given as n_j. Model abbreviations are as provided in Table 3.1 with the addition of the incorrectly specified models UC and BU245. As specified in the text, these models respectively represent a switching of constitutive and non-constitutive exons, or the omission of the Δ2-5 transcript from genomic databases.*

		Correct Priors					**Incorrect Priors**	
Exon	$\mathbf{n_j}$	**1B**	**BB**	**1U**	**BU**	**UU**	**UC**	**BU245**
1	6	1	B(7, 1)	1	B(7, 1)	B(1, 1)	B(1, 1)	B(7, 1)
2	5	B(6, 2)	B(6, 2)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(7, 1)
3	3	B(4, 4)	B(4, 4)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(1, 1)
4	5	B(6, 2)	B(6, 2)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(7, 1)
5	5	B(6, 2)	B(6, 2)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(7, 1)
6	3	B(4, 4)	B(4, 4)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(1, 1)
7	5	B(6, 2)	B(6, 2)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(1, 1)
8	5	B(6, 2)	B(6, 2)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(1, 1)
9	5	B(6, 2)	B(6, 2)	B(1, 1)	B(1, 1)	B(1, 1)	B(7, 1)	B(1, 1)
10	6	1	B(7, 1)	1	B(7, 1)	B(1, 1)	B(1, 1)	B(7, 1)

Table 3.3 – *Priors assigned to model parameters beyond those described in Table 3.2*

Model Term	**Prior**
c_i	$c_i \sim \mathcal{U}(0, 16)$
p_k	$p_k \sim \mathcal{N}(0, \sigma_p)$
σ_p	$\sigma_p = 3$
σ_S	$\sigma_S = 0.7$

3.2.5 Simulation Results

Log Fold-change Estimates

Of the 1000 simulated genes, 709 were randomly simulated with no fold-change, and for this subset of data points estimates of fold-change were plotted for each combination of priors (Figure 3.3). A clear bias to logFC estimates was observed as the transcripts begin to diverge in length. Patterns 4, 6 & 7 contain transcripts in Treatment B which were respectively 1, 2 & 4 exons shorter than Treatment A, and this clearly presents as an increase in expression estimates for the samples with the longer transcripts.

The two models utilising Beta Priors for the spliced exons (1B & BB) also showed a similar positive bias in logFC estimates that was evident for the PLM fitted values. The two models which demonstrated the least bias with the increasing divergence in transcript length were those with Uniform Priors for spliced exons (1U & BU). Whilst this bias was still evident in the more extreme patterns (6 & 7), it was considerably less noticeable than all other approaches.

The correctly specified priors for constitutive exons for the 1U or BU combinations both demonstrated the least overall bias in the logFC results and initially appeared to be the preferred approaches. However, estimates obtained when one transcript in Pattern 8 was unknown (BU245, last panel Figure 3.3); or when constitutive exons are misspecified (UC) clearly highlight that these approaches are not robust in the presence of unknown transcripts. Notably, the estimates from the misspecified UC model most closely resembled those obtained using the PLM approach. This leaves the specification of Uniform priors for all exons (UU) as the model which is the most robust to the presence of previously unknown transcripts. Whilst the positive bias in the presence of diverging transcript lengths is still evident under this approach, it is much less marked than the PLM model. Thus the general specification of Uniform priors for all exons appears on this evidence to be the preferred approach of the models tested in terms of introducing bias to estimates of logFC.

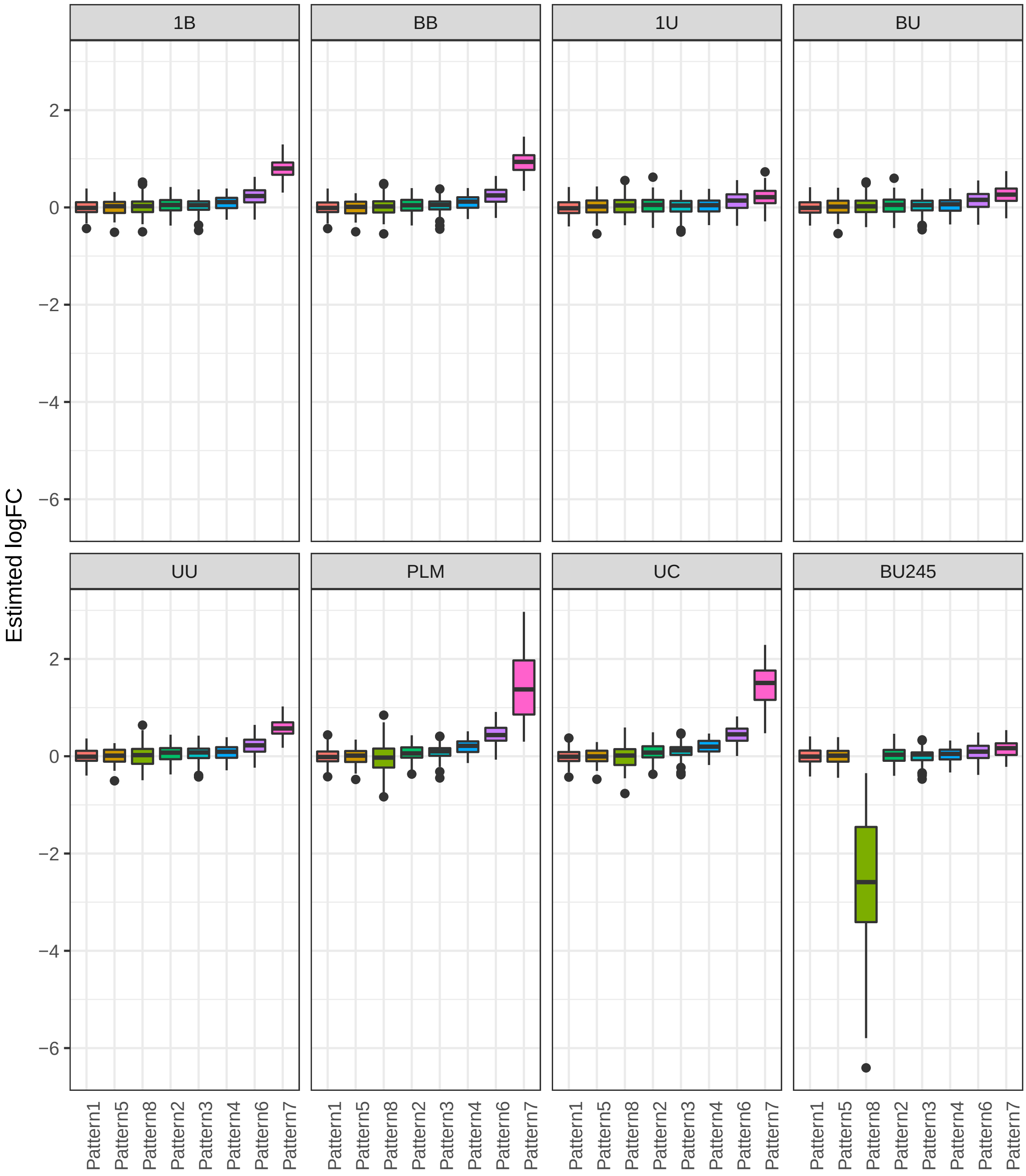


Figure 3.3 – *Comparison of estimated logFC using posterior means for the BMEA approaches, and conventional PLM methods for simulation where no logFC was specified. Prior specifications for ϕ_{hj} were as described in Table 3.2. Splicing patterns are shown in increasing order of the difference in transcript lengths. Patterns 1, 5 and 8 all contain the identical number of exons, with patterns 2, 3, 4, 6 and 7 gradually increasing in the difference between exon numbers.*

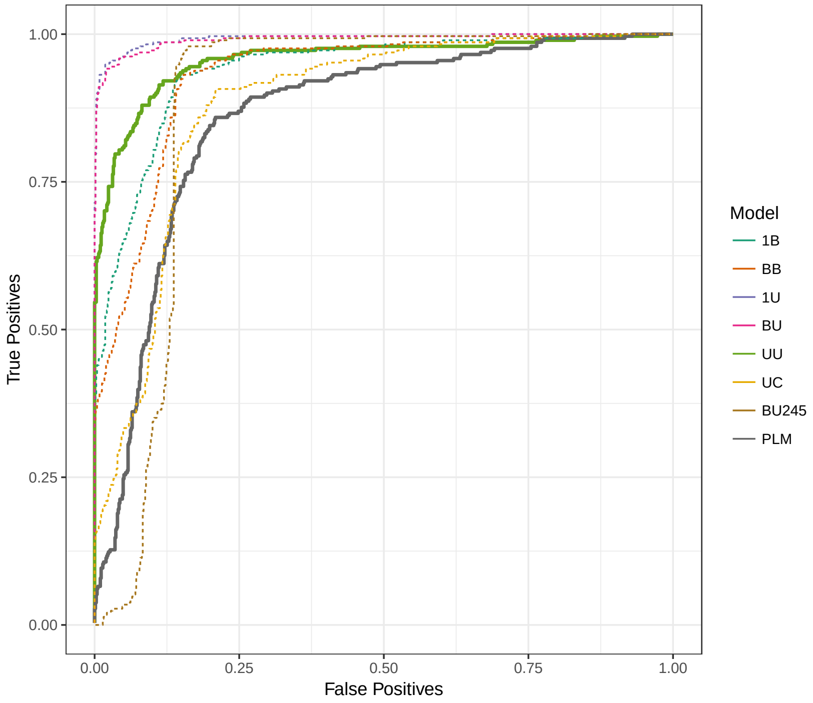


Figure 3.4 – *ROC curves for logFC using the approaches under investigation. The two models of primary interest are the PLM approach and the dual-Uniform (UU) priors, and these are shown as solid lines. All other models tested are shown as dotted lines.*

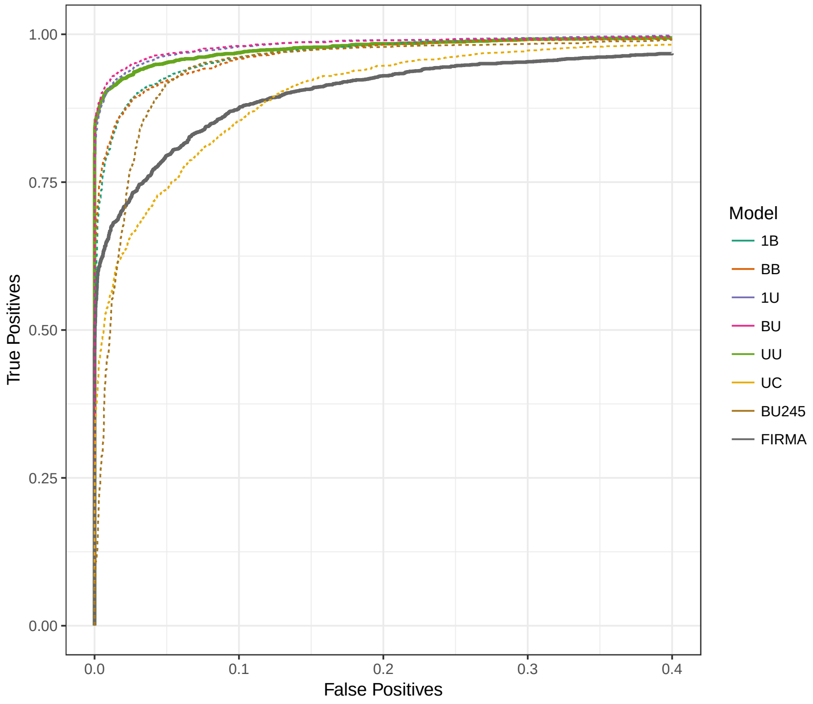


Figure 3.5 – *ROC curves for detection of alternate splicing. The two models of primary interest are FIRMA and the dual-Uniform (UU) priors, and these are shown as solid lines. Other models with flaws noted in the previous section are shown as dotted lines.*

ROC curves

As well as comparing the obtained estimates of log fold-change, ROC curves were compared using the various models and specified priors (Figure 3.4). For both detection of logFC and alternate splicing (Figure 3.5), the UU model clearly out-performed the conventional PLM/-FIRMA approach. Best performing of all models were those with the correct specification of constitutive exons (1U & BU), however under incorrect specification of constitutive exons (BU245), this approach was the worst performing, rendering this approach non-viable unless all splice variants are known *a priori* with 100% accuracy.

The performance of the UU and PLM models was also assessed for the 130 simulations where *no splice variation* was included (splicing pattern 1 from Figure 3.2). 33 of these simulated genes were simulated with non-zero fold-change, whilst 97 were given zero fold change. Both models were virtually indistinguishable with near-perfect performance in this much smaller data set. The UU model ranked all simulations with fold-change above those with none, whilst the PLM approach ranked all but one simulation correctly, with this final simulation being ranked behind two of those with no-fold-change.

3.2.6 Discussion Of Initial Simulation Results

The above set of simulated data was designed to contain a large amount of splicing variation as the performance of the various approaches under splice variation was of specific interest. The bias in the estimates of fold-change has been clearly shown in the presence of transcripts of varying length. However, the impact of this in a true research context will be widely variable and difficult to discover using conventional array-based approaches. Two highly similar cell-types may only contain minimal splice variation, whilst two less related cell-types may conceivably contain a much greater degree of splice variation.

In the absence of splice variation both models were comparable for detection of fold-change, with the conventional PLM approach clearly having a computational advantage. However, in the presence of splice variation, the BMEA approach appears to have a greater accuracy for estimation of fold-change, as well as for alternate splice detection. Additionally, the BMEA approach has the advantage of modelling an exon-level term with a value that directly correlates with a biological behaviour, as opposed to a FIRMA score which is a scaled residual and has a generally abstract interpretation.

The above results were also obtained using posterior means for each term, which implicitly assumes normality for each term, and fails to exploit the true Bayesian nature of the BMEA model. However, for the purposes of initial explorations this provided a computationally simple approach and allowed a preliminary comparison between analytic models. Subsequent stages of model development were designed to better capture the true Bayesian nature of BMEA parameters.

3.3 The Complete BMEA Model For True Signal

The previous section explored a simple Bayesian approach for the *"true"* signal component of the observed array intensities (Equation 1.9), with prior distributions only assigned to a subset of possible model parameters. In order to apply this to experimental data, a more complete specification of the set of parameters was required. In addition, the background signal component of observed probe intensities needs to be incorporated into the model to more accurately represent observed PM signal. This was performed by breaking the observed intensities into the separate components as defined in Equation 1.9, and giving each component a distinct set of priors:

$$PM_{hijk} = B_{hijk} + S_{hijk}\,, \tag{3.5}$$

where

$$\log B_{hijk} \sim \mathcal{N}(\lambda_{hijk}, \delta_{hijk});\quad B_{hijk} \leq 2^{16} \tag{3.6a}$$

and

$$\log S_{hijk} \sim \mathcal{N}(\eta_{hijk}, \sigma_S);\quad S_{hijk} \leq 2^{16} \tag{3.6b}$$

The components of the mean for the true signal component (η_{hijk}) are as defined in Equation 3.4 with the addition of the subscript $h = \{1, 2, \ldots, H\}$, to denote the *treatment or cell-type*. All other subscripts remain as above in Equation 3.2. The hyperparameters λ_{hijk} & δ_{hijk} for the background component can be estimated using the sequence information for each probe, as detailed below in Section 3.4. As above, each term is discussed in the context of a single gene with possible subscript g suppressed, as no parameters are shared across genes under BMEA.

3.3.1 The Overall Signal Level

In conventional microarray analysis, data is transformed onto the the $\log_2$ scale and model fitting as applied using Equation 1.11 assumes normality on this scale. An analogous approach was taken for BMEA, such that the *natural logarithm* of the observed signal estimate would be normally distributed around mean η_{hijk}, with standard deviation σ_S as defined in Equation 3.6b. Whilst conventional PLM analysis is performed using $\log_2$ as the default, natural logarithms were utilised for iterative convenience, with conversion to the $\log_2$ being a trivial matter when required.

Additionally, the Affymetrix scanner has an upper detection limit of 2^{16} and a small proportion of probes can be reasonably expected to show saturation at this level. Thus under BMEA, the observed signal is defined with a truncated log-normal prior distribution, bound by the saturation level above at 2^{16}. The hyperparameter η_{hijk} is a composite term derived from Equation 3.3 such that

$$\eta_{hijk} = c_{hi} + p_{jk} + \log\phi_{hj}\,, \tag{3.7}$$

where both the chip effects (c_{hi}) and exon-proportion term (ϕ_{hj}) are now considered in a cell-type or treatment dependent context. The standard deviation is assumed to be constant within an entire gene across all samples, and is given the Jeffreys Prior (Jeffreys, 1946):

$$\sigma_S^2 \propto \frac{1}{\sigma_S^2}\,. \tag{3.8}$$

Expression-level terms

The chip-specific, expression-level component of η_{hijk} is the hyperparameter c_{hi} where $i = (1, 2, \ldots, I_h)$ and $I = \sum_{h=1}^{H} I_h$ to allow for different sizes within treatment groups. The term c_{hi} is defined as being normally distributed around a treatment-specific expression level, μ_h, with standard deviation σ_μ which is assumed to be common across treatments. This distribution is also truncated at the upper limit of Affymetrix scanner resolution:

$$c_{hi} \sim \mathcal{N}(\mu_h, \sigma_\mu); c_{hi} \leq \log 2^{16}\,. \tag{3.9}$$

The 2^{nd}-level hyperparameter μ_h was given a locally uniform hyperprior from zero to the Affymetrix scanner saturation-level ($\log 2^{16}$).

$$\mu_h \sim \mathcal{U}(0, \log 2^{16})\,. \tag{3.10}$$

Whilst values below zero are technically possible, these were considered unlikely as the true signal at this level would be ≈ 1 on the linear scale, and would effectively be swamped by background signal. Additionally, the chip-specific term c_{hi} is free to fall below this value.

The standard deviation around the treatment-specific mean, σ_μ, was given a uniform hyperprior over a range of values, such that $\mu_h \pm 2\sigma_\mu$ would comfortably cover most of the possible range for μ_h:

$$\sigma_\mu \sim \mathcal{U}(0, 5)\,. \tag{3.11}$$

This parameterisation explicitly assumes that the sample-based variability, i.e. biological variability, around the cell-type specific mean is equal between cell-types.

The above parameterisation effectively defines *log fold-change* (logFC) as the difference between μ_a & μ_b (i.e. $\Delta\mu$), where a and b define any two cell-types, with the exception that the natural logarithm is used here instead of the more common $\log_2$ scale. Thus the posterior distribution for logFC can be directly sampled during any MCMC iterative process.

Probe-level terms

The next component of the hyperparameter η_{hijk} is the probe affinity term p_{jk}, which is assumed to be normally distributed around a zero-mean with a common variance across all probes (σ_p^2):

$$p_{jk} \sim \mathcal{N}(0, \sigma_p)\,. \tag{3.12}$$

Each probe is formally nested within an exon j such that $k = 1, 2, \ldots, K_j$ with $K = \sum_{j=1}^{J} K_j$ being equal to the total number of probes. However, unlike the nested term c_{hi}, the exon j is not specifically modelled in any associated hyperparameters for p_{jk} and as such is simply referred to as p_k.

The variance hyperparameter σ_p is given a locally uniform hyperprior over an interval which is wide enough to be non-restrictive, but within reasonable limits for the purposes of simple estimation:

$$\sigma_p \sim \mathcal{U}(0, 10)\,. \tag{3.13}$$

Exon-level terms

As the distribution of sample-specific exon-inclusion rates around any cell-type specific value would be difficult to simply define, this term was instead specified at the *cell-type* or *treatment* level. This also allows easy sampling of the differences in exon proportions between any two cell-types under investigation ($\Delta \log \phi_j$), which will be analogous to the sampled logFC parameter discussed above. The Uniform prior as defined in Section 3.2.1 would then become

$$\phi_{hj} \sim \mathcal{U}(0,1)\,. \tag{3.14}$$

Under this specification neither boundary point of the $\mathcal{U}(0,1)$ distribution will ever be sampled, and a Mixture prior was also proposed to allow for these possibilities. This was done by defining a vector of binary indicator variables $\boldsymbol{\xi_{hj}} = (\xi^1_{hj}, \xi^2_{hj}, \xi^3_{hj})$, with the two alternate specifications being referred to as either the Uniform or Mixture models in subsequent model testing:

$$\phi_{hj} \sim \begin{cases} 1 & \text{if } \ \xi^1_{hj} = 1 \\ \mathcal{U}(0,1) & \text{if } \ \xi^2_{hj} = 1 \\ 0 & \text{if } \ \xi^3_{hj} = 1 \end{cases} . \tag{3.15}$$

For the Mixture model the indicator vector $\boldsymbol{\xi_{hj}}$ was defined with a multinomial hyperprior to ensure that only one term in the vector can take the value one:

$$\boldsymbol{\xi_{hj}} \sim \text{Multinom.}(3, \boldsymbol{q_{hj}})\,. \tag{3.16}$$

The associated probability vector q_{hj} was then assigned a Dirichlet(1, 1, 1) hyperprior distribution

$$\boldsymbol{q_{hj}} \sim \text{Dirichlet}(1,1,1)\,. \tag{3.17}$$

A complete summary of all components of the hyperparameter η_{hijk} is presented in Table 3.4.

Table 3.4 – *Summary of the components of the hyperparameter η_{hijk} broken down by component and associated hyperpriors.*

1^{st} **Level Hyperparameters**		**Higher Level Hyperparameters**	
Parameter	**Prior**	**Parameter**	**Prior**
c_{hi}	$c_{hi} \sim \mathcal{N}(\mu_h, \sigma_\mu); c_{hi} \leq \log 2^{16}$	μ_h σ_μ	$\mu_h \sim \mathcal{U}(0, \log 2^{16})$ $\sigma_\mu \sim \mathcal{U}(0, 5)$
p_{jk}	$p_{jk} \sim \mathcal{N}(0, \sigma_p)$	σ_p	$\sigma_p \sim \mathcal{U}(0, 10)$
ϕ_{hj}	1. $\phi_{hj} \sim \mathcal{U}(0,1)$ or 2. $\phi_{hj} \sim \begin{cases} 1, & \xi^1_{hj} = 1 \\ \mathcal{U}(0,1), & \xi^2_{hj} = 1 \\ 0, & \xi^3_{hj} = 1 \end{cases}$	ξ_{hj} q_{hj}	$\xi_{hj} \sim$ Multinom$(3, q_{hj})$ $q_{hj} \sim$ Dirichlet$(1, 1, 1)$

3.3.2 Ranking of Results

As the posterior distributions for log fold-change and any change in exon proportions can be sampled directly for each comparison, a suitable ranking method needs to be defined. As distribution of the parameter of interest (θ) would be of interest in relationship to zero, i.e. $p(\theta > 0)$ or $p(\theta < 0)$, a possible statistic would be

$$B = \log \frac{N^+}{N^-}\,, \tag{3.18}$$

where N^+ and N^- represent the number of sampled values retained for θ either above or below zero respectively.

Under the proposed models, it would not be uncommon to obtain results where either N^+ or N^- were zero, so a correction factor of 0.5 was added to both lines, giving the corrected version of the B-statistic

$$B = \log \frac{0.5 + N^+}{0.5 + N^-}\,. \tag{3.19}$$

Due to the finite number of retained samples in any MCMC process, it is also worth noting that this statistic will effectively be a discrete statistic.

3.3.3 Simulated Data for the Complete True Signal Term

Due to the computational complexity of the model specified by the Mixture prior for ϕ_{hj}, a smaller set of 500 simulated genes was generated to enable fitting in a feasible time frame. Data for each simulated gene was generated using the algorithm below as based on the model specification.

1. Define all parameters for the splicing patterns in Figure 3.2 with 2 cell-types and 4 samples from each cell type, giving $h = (1, 2)$ and $i = (1, 2, 3, 4)$ for both cell-types
2. Randomly select $\mu_1 = \mu_2$ from $\mu = (4, 6, 8)$ to represent low, medium or high expression
3. For 100 of the simulations add ± 1.5 to μ_2 to provide data points with fold-change
4. Simulate c_{hi} from $\mathcal{N}(\mu_h, 0.4)$.
5. Simulate p_{jk} from $\mathcal{N}(0, 2.5)$
6. Randomly select a splicing pattern and set ϕ_{hj} accordingly
7. Simulate general noise (ε_{hijk}) from $\mathcal{N}(0, 0.7)$
8. Set $S_{hijk} = \phi_{hj} \exp(c_{hi} + p_{jk} + \varepsilon_{hijk})$
9. Simulate $\log B_{hijk}$ from $\mathcal{N}(4, 0.4)$
10. Create a final matrix $PM_{hijk} = \min(B_{hijk} + S_{hijk}, 2^{16})$

Fitting Simulated Data in R

As fitting the model using the Mixture prior for ϕ_{hj} was a computationally complex task, WinBUGS proved to be inadequate and the MCMC sampling algorithms were written in R. This required manual derivation of the posterior distributions and sampling strategies for each parameter, and these are included as Appendix A.

Updates during each iteration were performed using the Metropolis-Hastings Algorithm (Hastings, 1970) for variance parameters, and the Gibbs Sampler (Geman & Geman, 1984) for all remaining parameters. Posterior distributions were obtained using 4 chains and 10000 iterations, discarding the first 5000 as the burn-in period. 1000 iterations for each posterior distribution were retained from each chain. Chains were run in parallel using the R package *snow* (Tierney et al., 2009). Priors for each parameter were as specified in Table 3.4, with background signal given the prior $\log B_{hijk} \sim \mathcal{N}(4, 0.4)$ corresponding to the distribution it was drawn from. On a quad-core desktop machine with 16GB RAM, fitting took ∼1 minute per simulated gene for the Uniform model, and approximately 30 minutes per gene for the Mixture model, giving a total run time of over a week for this latter model.

The posterior distribution for logFC ($\mu_2 - \mu_1$) was sampled directly at each retained iteration, as were changes in $\log \phi_j$ ($\Delta \log \phi_j$). The B-statistic as defined in Equation 3.19 was used to rank results for both parameters of interest, with ties broken by the absolute value of the relevant parameter. Posterior means were taken as a point estimate for each parameter for direct comparison with PLM/FIRMA results.

To obtain fitted values under the PLM/FIRMA approach, the set of simulated genes was background corrected using RMA, then fitted values for expression levels were used to obtain estimates of fold-change and FIRMA scores. Results for the PLM/FIRMA model were ranked based on FDR-adjusted p-values obtained from moderated $\tilde{t}$-statistics.

3.3.4 Inspection of Results

Comparison of Fold-Change Estimates

Initial inspection of logFC estimates for simulations with no specified fold-change (Figure 3.6) again showed the same pattern of bias in the presence of differing transcript lengths (e.g. Patterns 6 & 7) under both the PLM and Uniform approaches. However, this bias was not evident for simulations fitted using the Mixture prior (Equation 3.17).

For simulations where fold-change was included, simulations with splicing patterns 1,5 and 8 were considered to be the same length and estimates were grouped together for visual comparison between models (Figure 3.7). Simulations with other splicing patterns were considered to be of differing lengths between cell types, and these estimates were also grouped together.

For transcripts of the same length, estimates of log fold-change were generally conservative compared to the true values (Figure 3.7), with no model showing any clear improvement in accuracy. However, in the presence of truncated transcripts within cell-type B, the expected positive bias was again evident except under the Mixture model.

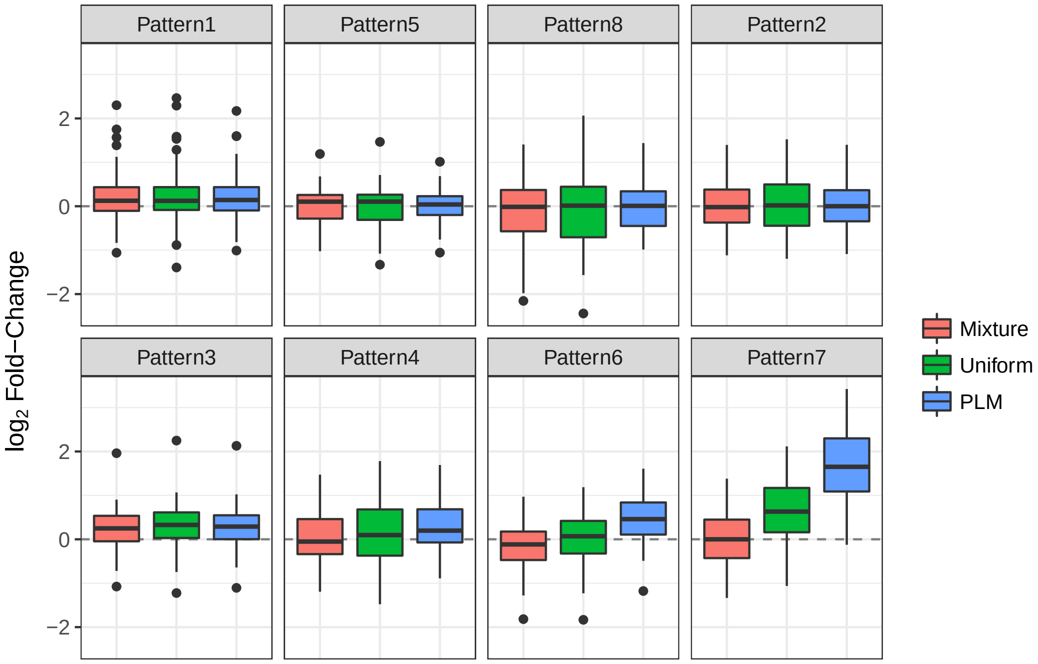


Figure 3.6 – *Comparison of logFC estimates for simulations in which no fold-change was specified. Posterior means were taken as point estimates for BMEA results, whilst fitted values are plotted for the PLM approach. Estimates of logFC are shown on the* $\log_2$ *scale. Boxplots are presented in order of increasing differences in transcript length. Splicing patterns 1, 5 and 8 contain transcripts of the same length across both cell-types, whilst the remaining five contain increasingly shorter transcripts in cell-type B, beginning with changing transcript ratios in patterns 2 and 3, and moving through 1, 2 or 4 missing exons in the remaining patterns.*

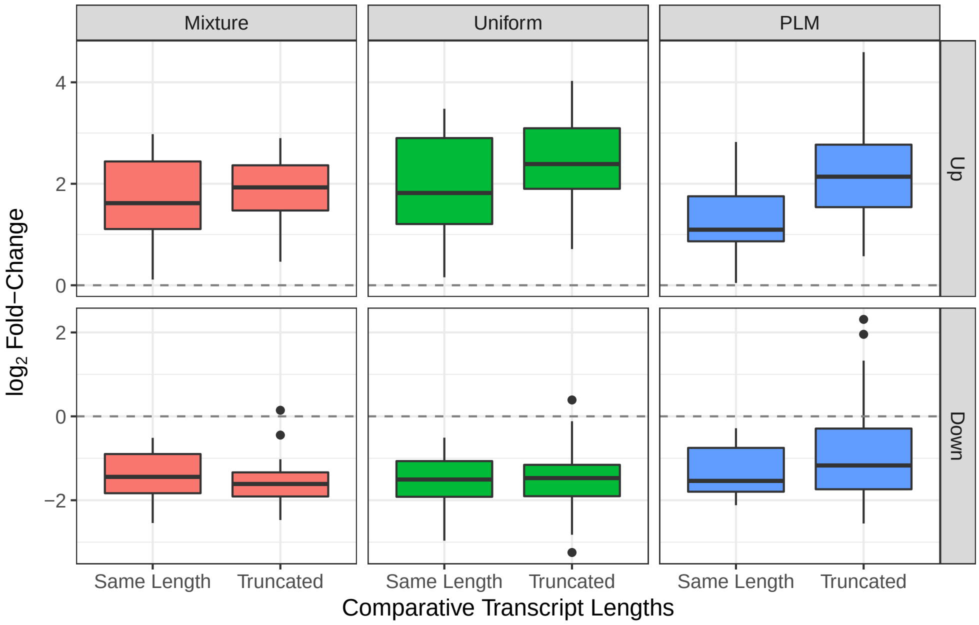
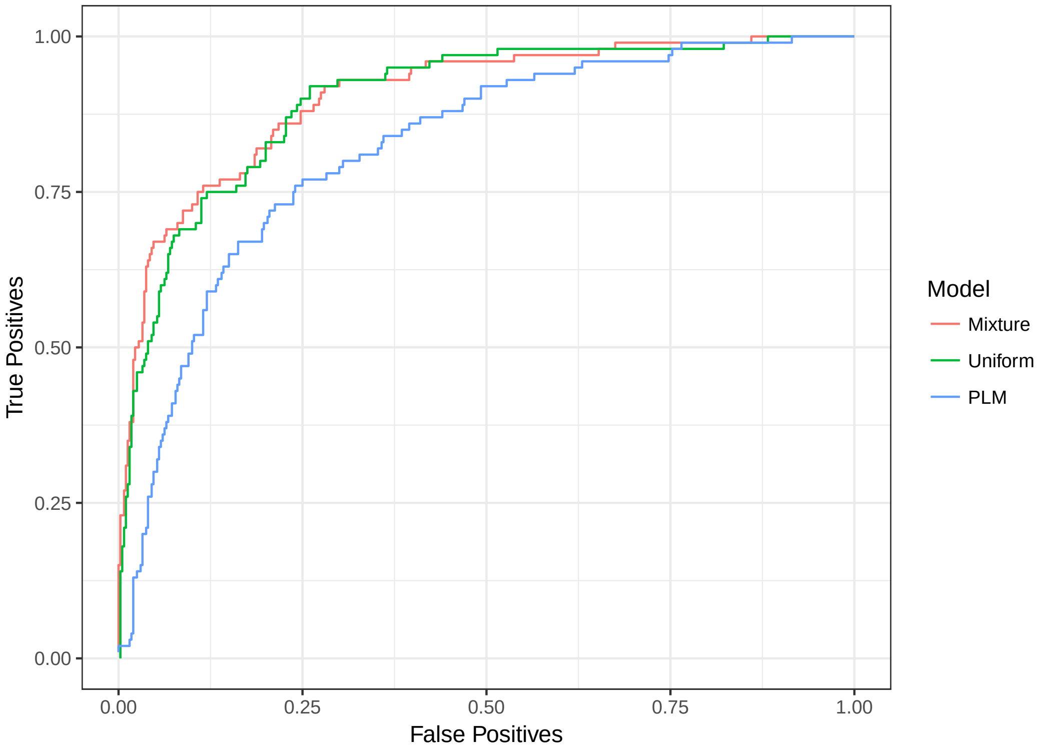


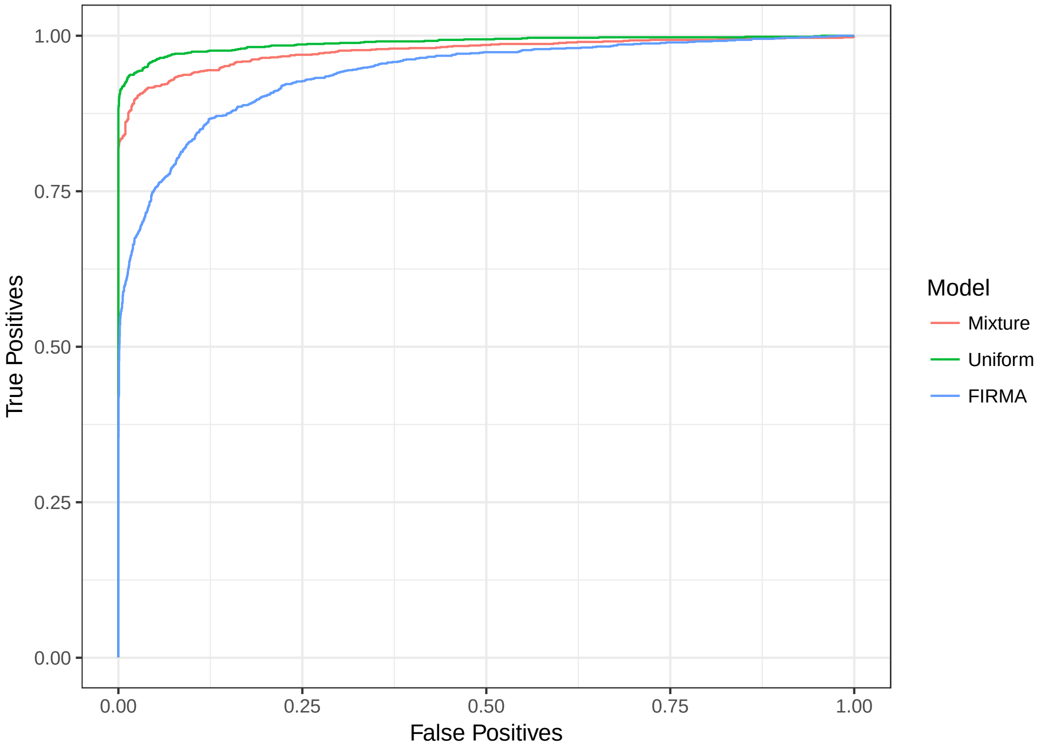
Figure 3.7 – *Comparison of logFC estimates for simulations in which fold-change of* ±2 *was specified. Posterior means were taken as point estimates for the Bayesian Models (Uniform & Mixture Priors), whilst the fitted values are plotted for the PLM approach. Simulations which included splicing patterns 1,5 & 8 were considered to contain transcripts of the same length in both simulated cell-types, and are grouped together. The varying reductions in exon inclusion rates for the remaining splicing patterns were considered to be truncated in cell-type B and are also shown as grouped together. The simulated values of logFC were* $\approx \pm 2$ *using* $\log_2$.

Comparison of Ranked Lists

The above comparisons of fitted values to true values indicated that the Mixture model may be preferable due to the comparative lack of bias, and ROC curves were fitted based on the ranking statistics for each model (Figure 3.8). For detection of fold-change (Figure 3.8A), both Bayesian approaches performed more strongly than the PLM model, with the Mixture model showing a narrow advantage over the BMEA approach using Uniform priors. However, for detection of differing exon-inclusion rates (Figure 3.8B), the Uniform model was clearly the best performing model of the three, with both of the BMEA approaches still outperforming FIRMA. The combination of these results, along with 30-fold increase in fitting time for the Mixture priors led to the adoption of Uniform priors for ϕ_{hj} as the only approach for all subsequent model development.



(A) *ROC curves for fold-change.*



(B) *ROC curves for splice detection.*

Figure 3.8 – *ROC curves for (A) fold-change and (B) alternate splice detection in simulated data, using the PLM/FIRMA, Uniform & Mixture models.*

3.4 Prior Specification for Background Signal

In order to complete the specification of the full BMEA model for any true PM intensities, an appropriate specification of the background signal is required, and a method was sought in which probe sequence information can be incorporated. As introduced in Section 1.2.3, the *anti-genomic* (AG) set of BG probes are designed specifically to not match any known sequence at the time of array design, whilst the *genomic* background set of probes are analogous to the mismatch (MM) probes in traditional 3' arrays (Section 1.2.2). Whilst neither will perfectly capture the NSB properties of PM probes due to competitive binding, the AG probes will be most likely to capture the binding properties of probes for less related sequences, given that the MM set of probes will additionally bind any targets of their corresponding PM probe.

3.4.1 Three Approaches for Modelling Background Signal

GC Bins

A simple method for specifying the parameters for background signal would be to use the bins defined by Affymetrix based on counts of GC nucleotides within the probe sequence. The observed intensities for the selected set of genomic (MM), or anti-genomic (AG) background probes could then be used to obtain estimates for any model parameters associated with sequences matching the GC-content for each bin. However, it has been shown (Kapur et al., 2007) that alternative approaches provide better estimates of the NSB for each probe.

GC-RMA

The GC-RMA background correction method (Wu et al., 2004) developed for 3' Arrays became widely used and the R package *gcrma* comes with an estimate of the positional effects of each base obtained across numerous experiments on this platform (Wu et al., 2011). However, fitting of these parameters using anti-genomic background probes from the Exon Array dataset from Chapter 2, revealed a distinctly different range of parameter estimates from the default estimates as calculated for 3' Arrays (Figure 3.9A).

A selection of publicly available Exon Array datasets were also used to fit GC-RMA parameter estimates (not shown), with all showing a high-level of similarity to the estimates shown in Figure 3.9. The choice of the genomic (MM) background probes also had a clear impact on the parameter estimates for the GC-RMA model (Figure 3.9B). As the source of this variation was beyond the scope of this work, these observations were simply noted and not explored further. However, this does provide a clear caution when using the default GC-RMA settings for background correction of Exon Array data.

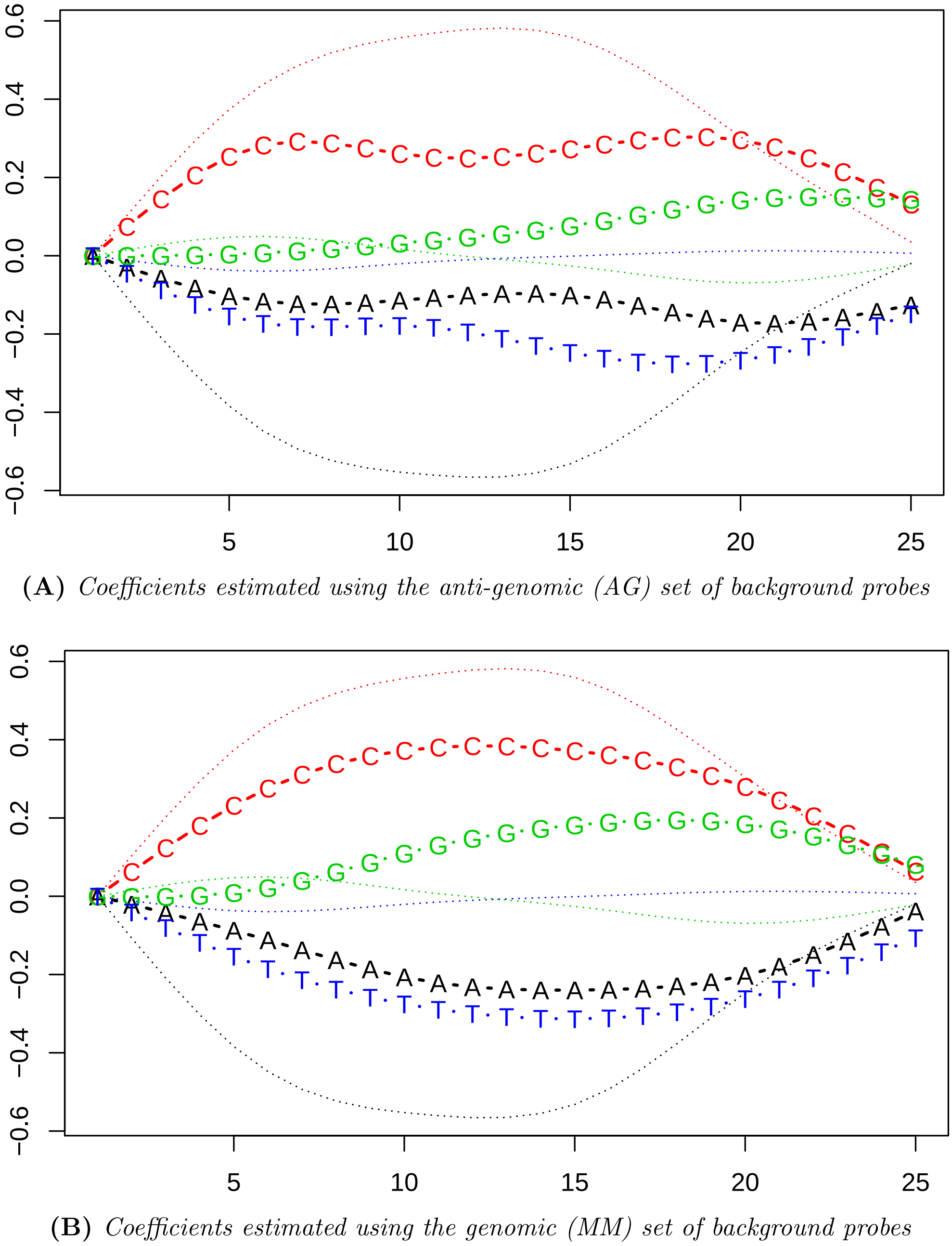
MAT

A less ubiquitous model for providing NSB estimates would be to use the modified MAT model (Kapur et al., 2007), which is able to provide more representative estimates of BG signal than using GC bins alone, or than under GC-RMA. For a given probe under this model, NSB can be modelled as:

$$\log B = \alpha n_T + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{jk} I_{jk} + \sum_{k \in \{A,C,G,T\}} \gamma_k n_k^2 + \varepsilon \,. \tag{3.20}$$

where n_k is the count of nucleotides of type k in the probe sequence, I_{jk} is the binary indicator variable for a nucleotide of type k at position j. The fitted parameters of the model are α, β_{jk} and γ_k, whilst ε is a general error term. This represents the same model as Equation 1.17 with the removal of the terms associated with copy number.

Using the AG set of probes, fitted values for expected intensities were obtained for both the MAT and GC-RMA approaches, using the entire set of arrays to fit the models. Comparison of these fitted values with the observed probe intensities (Figure 3.10) also confirmed that the MAT model better models the NSB properties of these probes than GC-RMA.



(A) *Coefficients estimated using the anti-genomic (AG) set of background probes*

(B) *Coefficients estimated using the genomic (MM) set of background probes*

Figure 3.9 – *Comparison of the model coefficients obtained under the GC-RMA model using the AG or MM sets of background probes. Estimates obtained using the dataset under investigation here are shown with bases labelled, and were obtained by fitting the entire set of 18 arrays. Estimates as given in the R package gcrma are shown as unlabelled dotted lines, with the colours representing the same bases as for the locally fitted coefficients.*

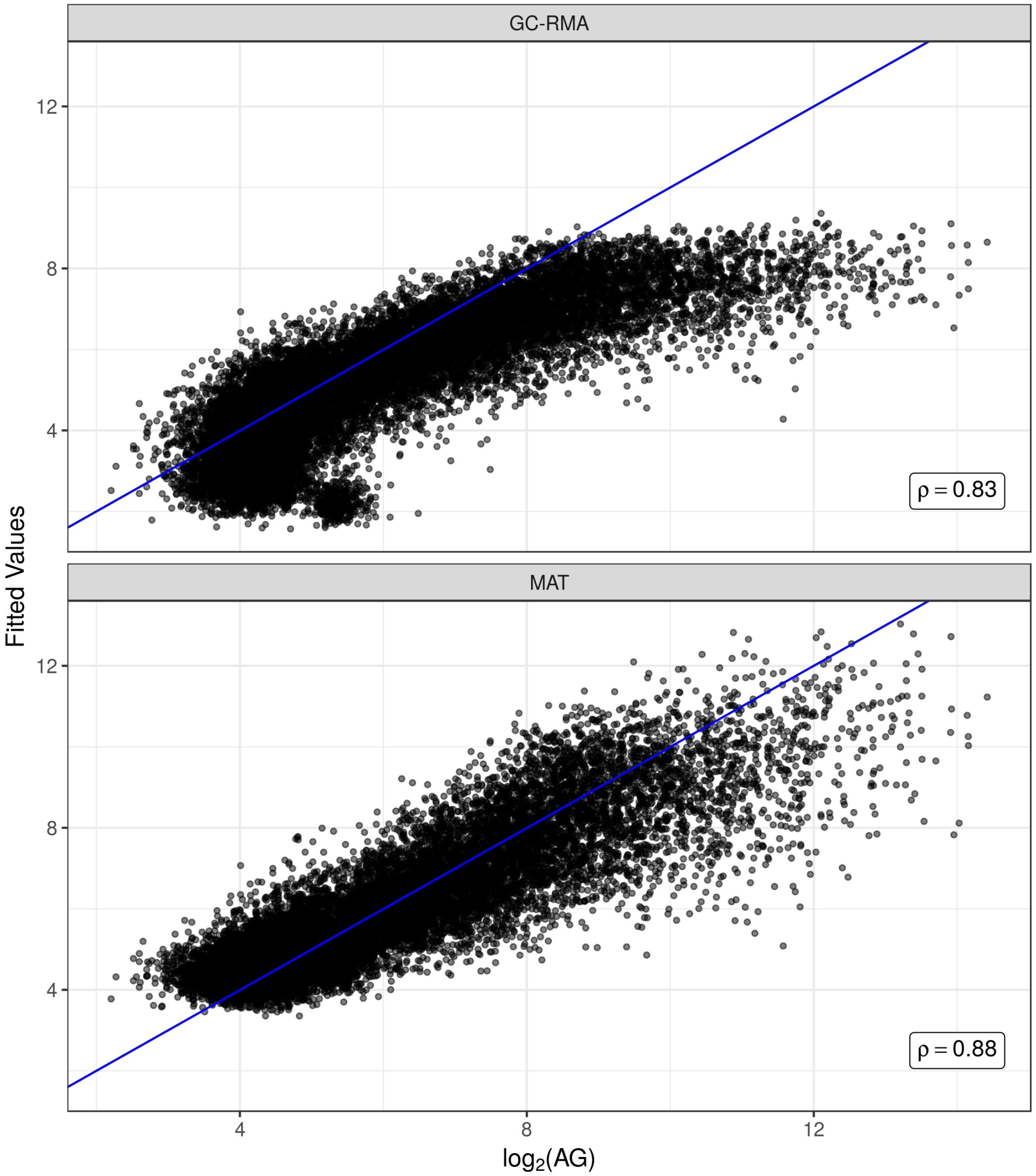


Figure 3.10 – *Fitted and observed values using MAT and GC-RMA on the set of antigenomic (AG) background probes. Observed values from the sample ThE41Stim are shown, with similar patterns being observed across all arrays. Correlations are shown in the bottom right, and the line $y = x$ is shown in blue to indicate a perfect fit. The unusual cluster of points near the bottom of the GC-RMA derived plot were observed across all T_{reg} arrays, but the underlying reason for this poor fit was not explored in this work.*

3.4.2 Using Bins to Define Background Signal

Placing BG probes in approximately equal-sized bins based on the fitted values under either MAT, GC-RMA or the GC content alone, would allow for simple estimation of a separate mean and standard deviation for each bin (l), using observed values for the set of BG probes. These values can then be simply assigned in an array-specific manner as the mean (λ_i) and standard deviation (δ_i) for a log-normal distribution, which defines the expected background signal for each probe:

$$\log B_{il} \sim \mathcal{N}(\lambda_{il}, \delta_{il}) \,, \tag{3.21}$$

where bin membership is denoted using the subscript l, and the subscript i continues to represent the array.

As the MAT model appeared to provide the strongest relationship with observed values (Figure 3.10), this was chosen as the preferred model for estimating NSB parameters. The three alternative approaches for generation of bins above are presented in Figure 3.11, with minimal differences found for the set of AG probes, beyond possible under-estimates or over-estimates of BG signal at the extreme ends of the range using GC Counts alone. Boxplots in Figure 3.11 also appeared broadly consistent with assumptions of normality, supporting this as a suitable specification for a prior distribution.

On the Human Exon 1.0 ST Array, there are 16,493 antigenomic (i.e. AG) probes, with 20 bins giving about 850 probes in each bin. This would also be the default number of bins available using the GC-content alone, as these values range between 4 and 23 in the set of PM probes.

After obtaining estimates of all MAT model parameters by using a suitable set of background probes, the known sequences of PM probes can be used to obtain fitted values for each PM probe, allowing each PM probe to be assigned to a suitable bin based on these fitted values. The observed values for the fitted set of background probes across the set of arrays, are thus able to provide estimates of the location (λ_{il}) and scale (δ_{il}) parameters for the prior distributions of background signal.

The range of estimates for both λ and δ obtained under this method for the $\mathrm{T_{reg}}$ dataset are shown in Figure 3.12. Notably, the expected signal for the first few bins is consistently low, with relatively low and stable estimates of the scale parameter. However, as the expected NSB signal increases, the variability of the subsequent prior distributions (δ_{il}) similarly increases, particularly towards the high end of the expected signal.

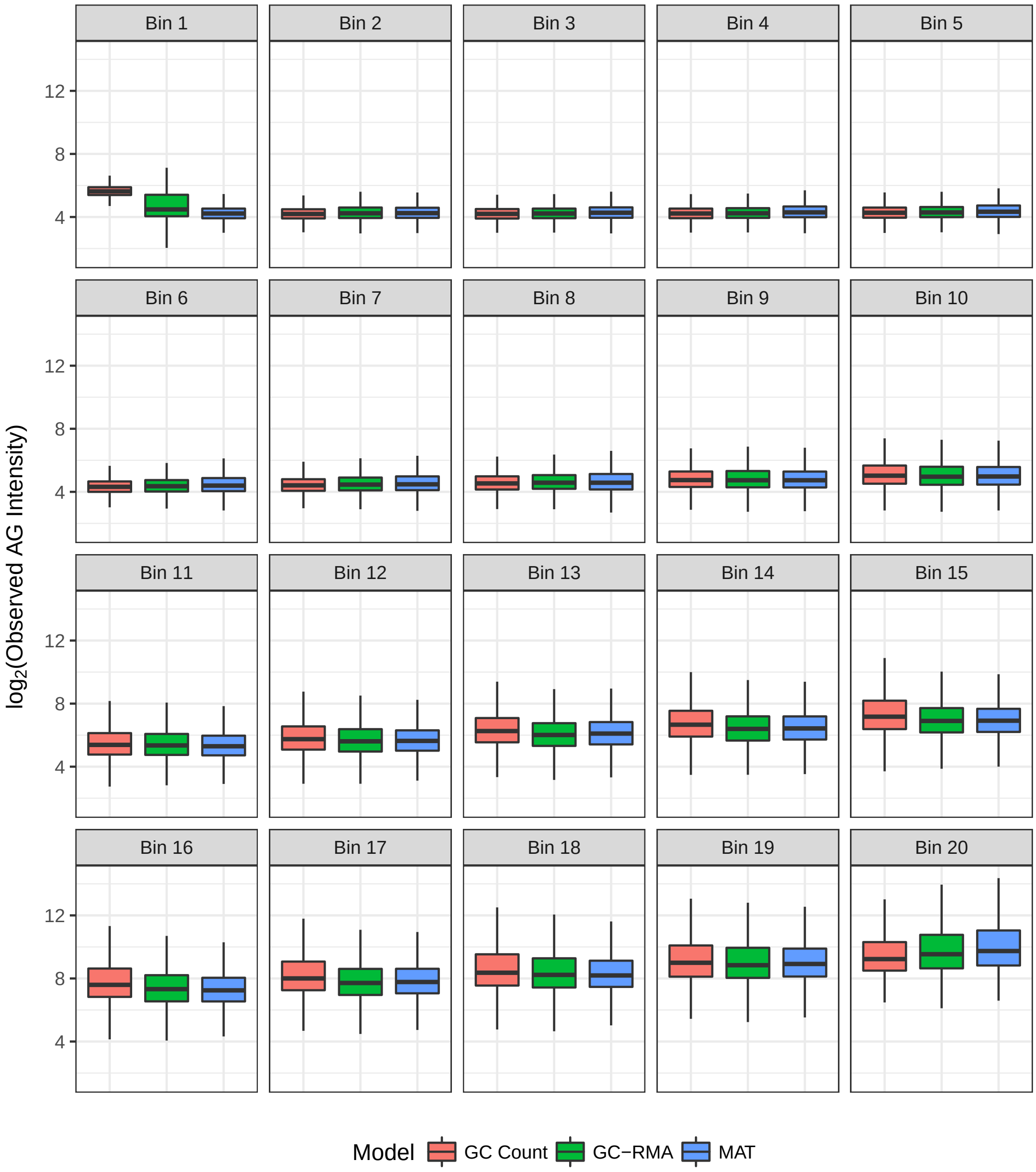


Figure 3.11 – *Comparison of all observed AG probe intensities, across all arrays in the T_{reg} dataset, shown on the* $\log_2$ *scale. Probes are broken in 20 bins based on a) GC-count, b) Fitted GC-RMA values, or c) Fitted MAT values. Bin sizes are near identical for GC-RMA & MAT, whilst the extreme bins are considerably smaller for those formed using the GC-count alone. Whiskers extend to 1.5 times the interquartile range, and points beyond this are not shown.*

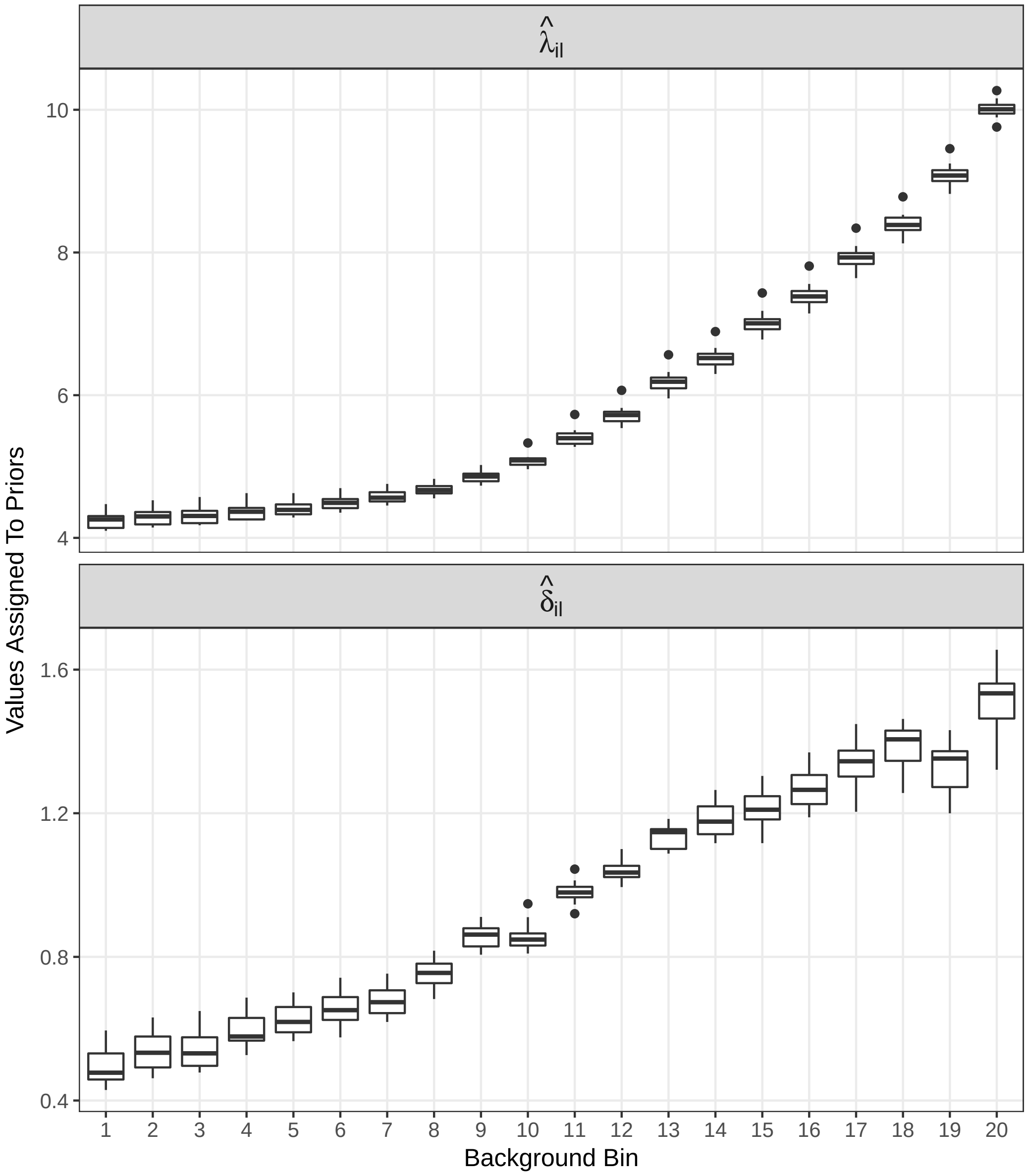


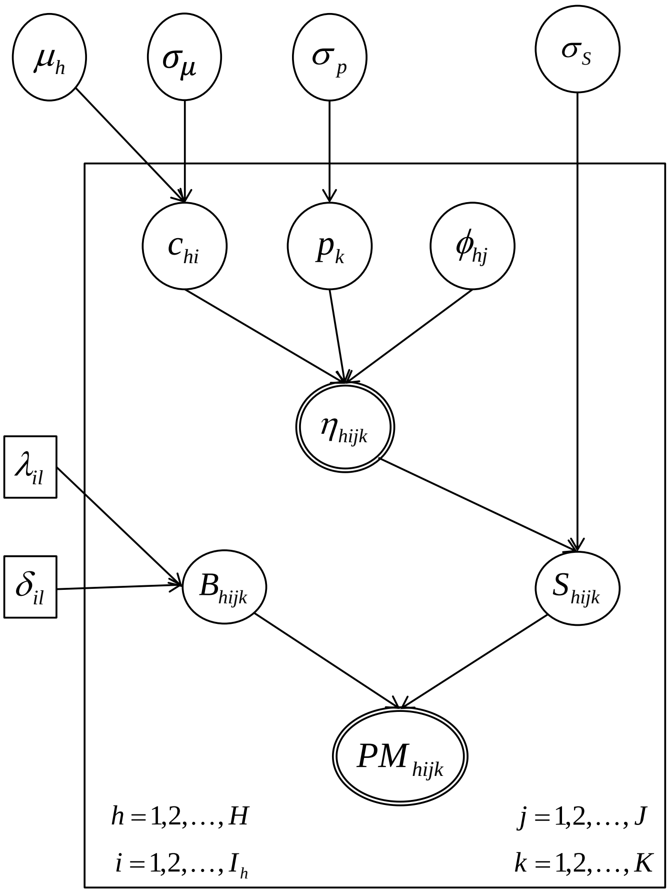
Figure 3.12 – *The complete set of values for location ($\hat{\lambda}_l$) and scale ($\hat{\delta}_l$) parameters obtained for each bin using the MAT approach for the complete set of T_{reg} arrays. All values were estimated using the anti-genomic set of background probes. Values are shown on the* $\log_2$ *for easier comparison with Figure 3.11. When fitting the BMEA model, these values will be returned to the natural logarithmic scale.*

3.5 Model Summary and Discussion

After the simulations and specifications above, the complete BMEA model can be presented as a Directed Acyclic Graph (DAG) as seen in Figure 3.13, following the standard DAG methodology for hierarchical Bayesian models (Lunn et al., 2000b). Notably, the constants for the background signal component B_{hijk} are derived *a priori* from the probe sequence information, and each PM probe is assigned a bin $l \in 1, 2, \ldots, L$ based on this information, with values λ_{il} and δ_{il} supplied in an array-specific manner. All other parameters have been given as uninformative priors as possible to avoid biasing posterior distributions.

The above initial explorations showed much promise for the BMEA approach. Whilst only the signal component of the model had been tested in this chapter, testing of the full model requires scaling up to larger simulated datasets. In order to perform this the entire process was developed into an R package using C for the MCMC sampling procedure, as discussed in the following chapter.

An important potential drawback which must be noted is that the second round of simulations (Section 3.3.3) no longer incorporated the paired nature of the $\mathrm{T_{reg}}$ dataset. Instead, data was simulated based on the BMEA model itself and the behaviours of signal under this model are not strictly as defined for analysis under PLM/RMA. As such, a decreased performance using PLM/RMA may not be unrealistic. Only testing on real data will ensure that the simulation algorithm has not created an implicit bias towards the BMEA model.



$$\mu_h \sim U(0, \log 2^{16})$$

$$\sigma_\mu \sim U(0, 5)$$

$$\sigma_p \sim U(0, 10)$$

$$\sigma_S \propto \sigma_S^{-1}$$

$$c_{hi} \sim N(\mu_h, \sigma_\mu)$$

$$p_k \sim N(0, \sigma_p)$$

$$\phi_{hj} \sim U(0, 1)$$

$$\eta_{hijk} = c_{hi} + p_k + \log \phi_{hj}$$

$$\log B_{hijk} \sim \mathrm{N}(\lambda_{il}, \delta_{il})$$

$$\log S_{hijk} \sim \mathrm{N}(\eta_{hijk}, \sigma_S)$$

$$PM_{hijk} = B_{hijk} + S_{hijk}$$

Figure 3.13 – *The full BMEA model expressed as a directed acyclic graph (DAG). In addition to the distributions specified, the terms S_{hijk}, B_{hijk} and c_{hi} are truncated at the Affymetrix scanner saturation level of 2^{16}. The constants λ_{il} & δ_{il} are assigned in an array-specific manner before any MCMC processes are begun.*

Chapter 4

BMEA Software Implementation and Validation

4.1 Building an R Package

Given the simulation results from the previous chapter, the BMEA process was built as an R package to enable processing of experimental datasets (`https://github.com/steveped/BMEA`). As the Uniform model had advantages in terms of computational time, and comparable analytic performance to the Mixture model, only this model was implemented in the package, given the working title *BMEA*. The technical description of the package implementation is given Appendix B, along with diagnostics such as parameter recovery and convergence. This chapter instead focusses on the correct detection of fold-change and alternate splicing using a large set of simulated data, which more closely resembled a true dataset in size.

4.2 Simulating Data for Package Testing

4.2.1 Using Observed Values for the Background Signal

In order to test the model on a more realistic dataset, one final set of simulated data was generated containing data for 10,000 simulated genes. To more closely resemble observed data, observed background signal from anti-genomic (AG) probes was incorporated into simulated data points as the background signal component. Eight arrays were randomly sampled from the Tissue Mixture dataset, described more completely in Section 5.4, and observed values from these arrays were used to randomly sample observed BG signal. The process for sampling this signal was as follows, with steps 1 and 2 being performed prior to any data simulation:

1. Define the bins for background signal using the AG probes:
   (a) Fit the full set of AG probes on all arrays ($i = 1, 2, \ldots, 8$) using the modified MAT model (Figure 4.1)

   (b) Assign each AG probe to one of 20 equally sized bins

   (c) Estimate λ_{li} and δ_{li} for each of $l = 1, 2, \ldots, 20$ bins for each array i
2. Find the true distribution of all PM probes across bins for BG signal:
   (a) Calculate fitted values for the full set of PM probes on each selected array, and assign them to the appropriate bins

   (b) Determine the probability $\pi_l : \sum_{l=1}^{2} 0\pi_l = 1$ of a PM probe belonging to each bin (Figure 4.2)
3. For each simulated PM probe:
   (a) Assign to a bin with probability π_l

   (b) Randomly sample an observed value from an AG probe in the same bin

   (c) Assign the values λ_{li} and δ_{li} to the simulated PM probe, as obtained in step 1c).

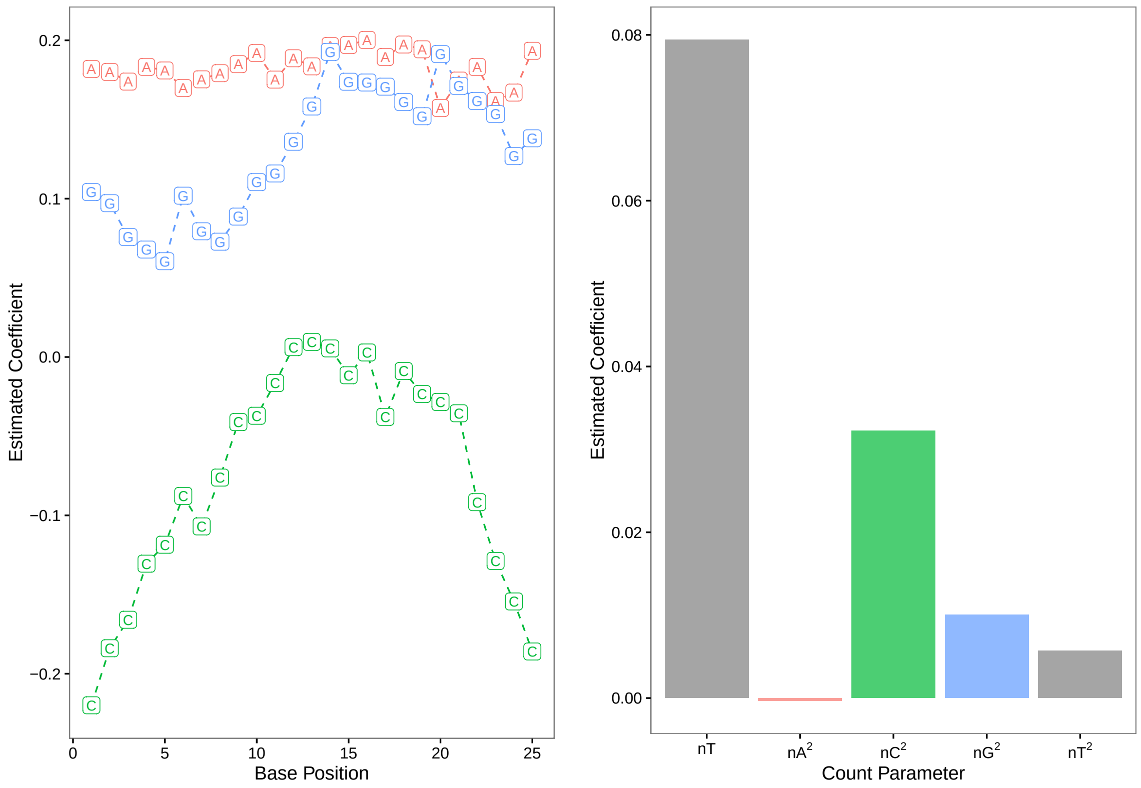


Figure 4.1 – *Parameter estimates obtained from fitting the MAT model in Equation 3.20 on the anti-genomic probes from 8 randomly selected arrays in the Tissue Mixture dataset.*

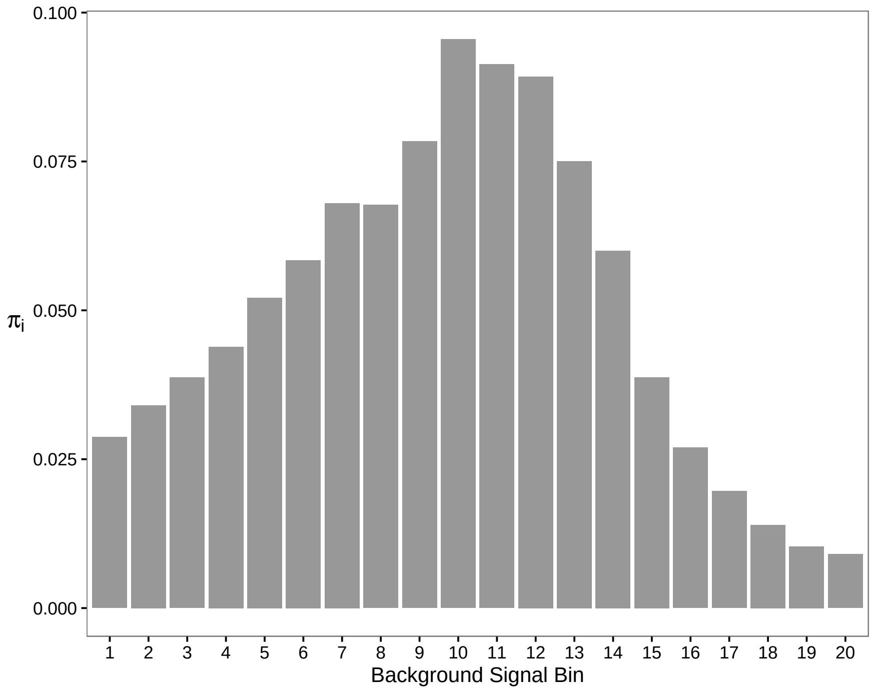


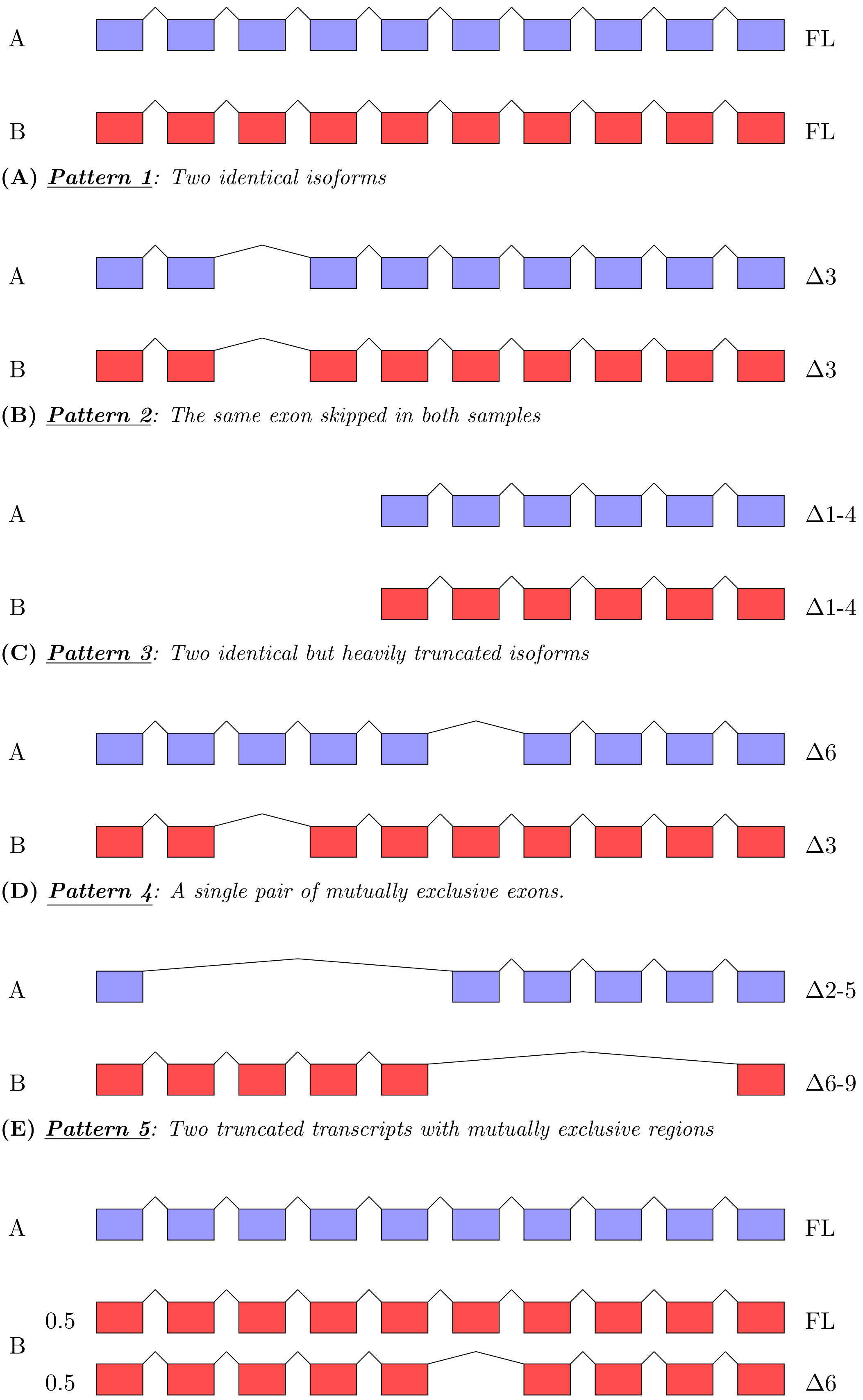
Figure 4.2 – *Probabilities (π_l) of a PM probe belonging to each BG signal bin. These were used for assigning prior distributions for background signal, and for sampling observed signal from AG probes during simulation of data points.*

4.2.2 Simulating the Signal Component

As noted in Section 3.3.4, changes in the length of transcripts between samples can impact estimates of fold-change. Previous simulations did not explicitly consider the impact of both cell types containing truncated transcripts beyond the mutually exclusive exon-groupings of pattern 8 (Figure 3.2). As a result, two further patterns (Figures 4.3B and 4.3C) were introduced to the set of simulations, in which the transcripts in both samples contained the same missing exon, and the same missing group of exons. Splicing patterns were renumbered as a result, and were more clearly grouped into patterns of the same length (Patterns 1-5), and patterns where sample B effectively contained a shorter transcript (Patterns 6-10). The complete set of 10 splicing patterns is given in Figure 4.3.

It should also be noted that if each pattern is selected with equal probability, $\sim 30\%$ of simulations will have no splice variation, whilst the remaining $\sim 70\%$ will contain some level of variability between samples. How truly this reflects the level of splice variation between two cell types will vary based on a genuine experiment. Analysis of two disparate tissues may indeed have a similar level of splice variation, whilst analysis of two highly related tissues will likely contain data with a much lower level of splice variation.

For generation of simulated data, each of the 10 splicing patterns as in Figure 4.3 were sampled with equal probability. Simulation parameters are summarised in Table 4.1, with genes being simulated with a 0.6 probability of no fold change. Saturation at $PM = 2^{16}$ was additionally permitted as part of the simulation procedure. Distributions of key simulated parameters are given in Figure 4.4.



(A) ***Pattern 1***: *Two identical isoforms*

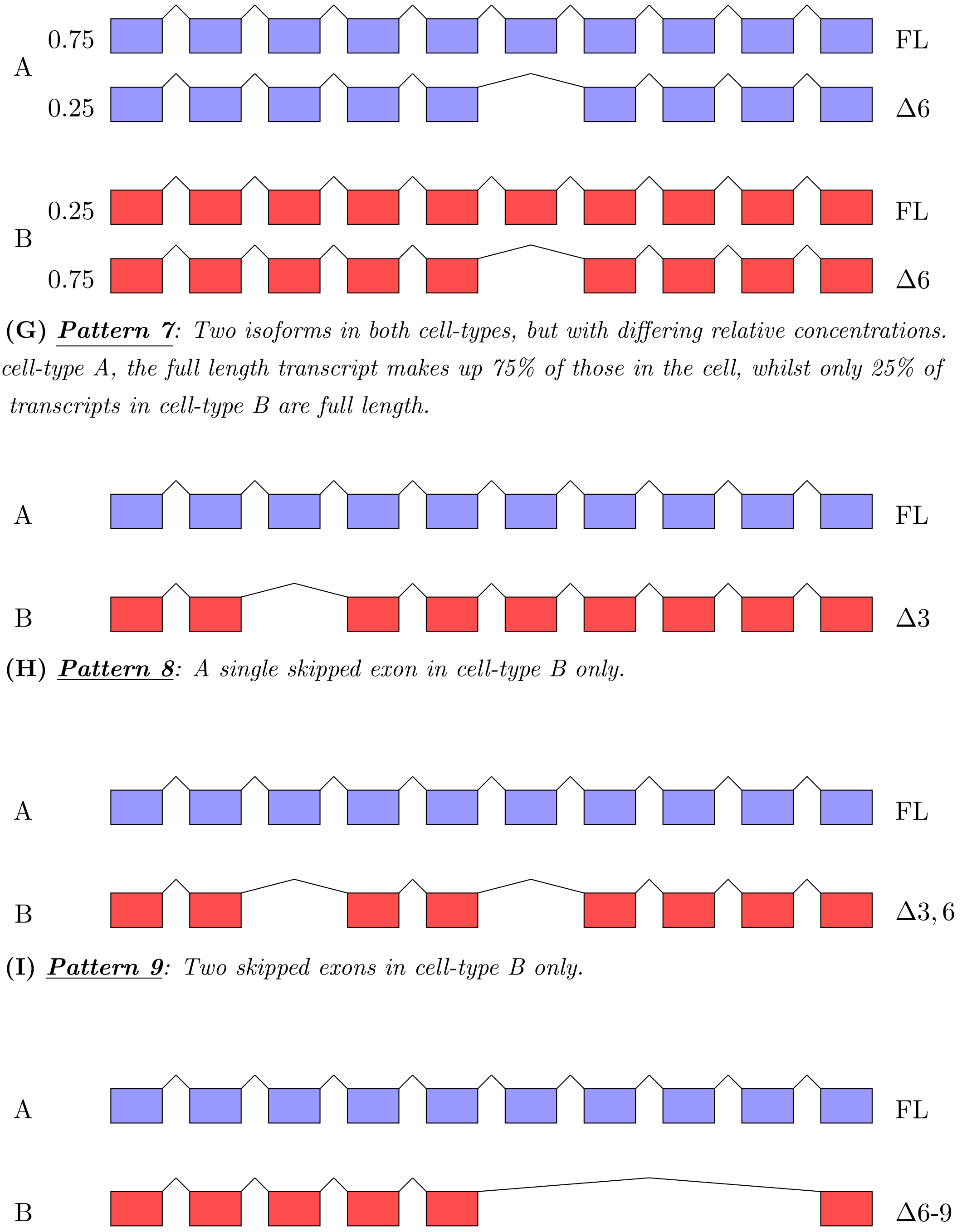
(B) ***Pattern 2***: *The same exon skipped in both samples*

(C) ***Pattern 3***: *Two identical but heavily truncated isoforms*

(D) ***Pattern 4***: *A single pair of mutually exclusive exons.*

(E) ***Pattern 5***: *Two truncated transcripts with mutually exclusive regions*

(F) ***Pattern 6***: *Two equally expressed isoforms in cell-type B only.*



(G) ***Pattern 7***: *Two isoforms in both cell-types, but with differing relative concentrations. In cell-type A, the full length transcript makes up 75% of those in the cell, whilst only 25% of the transcripts in cell-type B are full length.*

(H) ***Pattern 8***: *A single skipped exon in cell-type B only.*

(I) ***Pattern 9***: *Two skipped exons in cell-type B only.*

(J) ***Pattern 10***: *A significantly truncated transcript in cell-type B*

Figure 4.3 – *Transcript usage patterns for the 10-exon gene specified in simulated datasets for testing the R package BMEA. The two simulated cell types are shown in blue (cell-type A) and red (cell-type B). Patterns 1-5 contain transcripts of the same length, with no splice variation between cell types in patterns 1-3. In patterns 6-10, transcripts in cell-type B are consistently shorter than in cell-type A.*

Table 4.1 – *Parameters used for the simulation of 10,000 genes to be analysed using BMEA as implemented in an R package. In the case of μ_h, values were resampled until both values satisfied the boundary point criteria. The final step for simulating PM_{hijk} incorporates saturation, as would be observed in real data from an Affymterix scanner.*

Parameter	**Sampling Method**
logFC	$\text{logFC} \sim \begin{cases} -\log(4), & \pi_1 = 0.05 \\ -\log(2), & \pi_2 = 0.15 \\ 0, & \pi_3 = 0.6 \\ \log(2), & \pi_4 = 0.15 \\ \log(4), & \pi_4 = 0.05 \end{cases}$
	$\theta \sim \begin{cases} 2, & \pi_1 = 0.5 \\ 4, & \pi_2 = 0.4 \\ 6, & \pi_3 = 0.1 \end{cases}$
μ_1	$\mu_1 \sim \mathcal{N}(\theta, 1)\,; 0 < \mu_1 < \log 2^{16}$
μ_2	$\mu_2 = \mu_1 + \text{logFC};\ 0 < \mu_2 < \log 2^{16}$
σ_μ	$\log \sigma_\mu \sim \mathcal{N}(-1, 0.2)\,;\ \sigma_\mu < 5$
c_{hi}	$c_{hi} \sim \mathcal{N}(\mu_h, \sigma_\mu)\,;\ 0 < c_{hi} < \log 2^{16}$
σ_p	$\log \sigma_p \sim \mathcal{N}(\log 2, 0.2)\,;\ \sigma_p < 10$
p_k	$p_k \sim \mathcal{N}(0, \sigma_p)$
ϕ_{hj}	As determined by random selection of splicing patterns 1-10 (Figure 4.3) with equal probability.
σ_S	$\log \sigma_S \sim \mathcal{N}(\log 0.7, 0.2)\,;\ \sigma_S < 10$
η_{hijk}	$\eta_{hijk} = c_{hi} + p_k + \log \phi_{hj}$
S_{hijk}	$\log S_{hijk} \sim \mathcal{N}(\eta_{hijk}, \sigma_S)$
B_{hijk}	As described in Section 4.2.1
PM_{hijk}	$PM_{hijk} = \min\left(B_{hijk} + S_{hijk}, 2^{16}\right)$

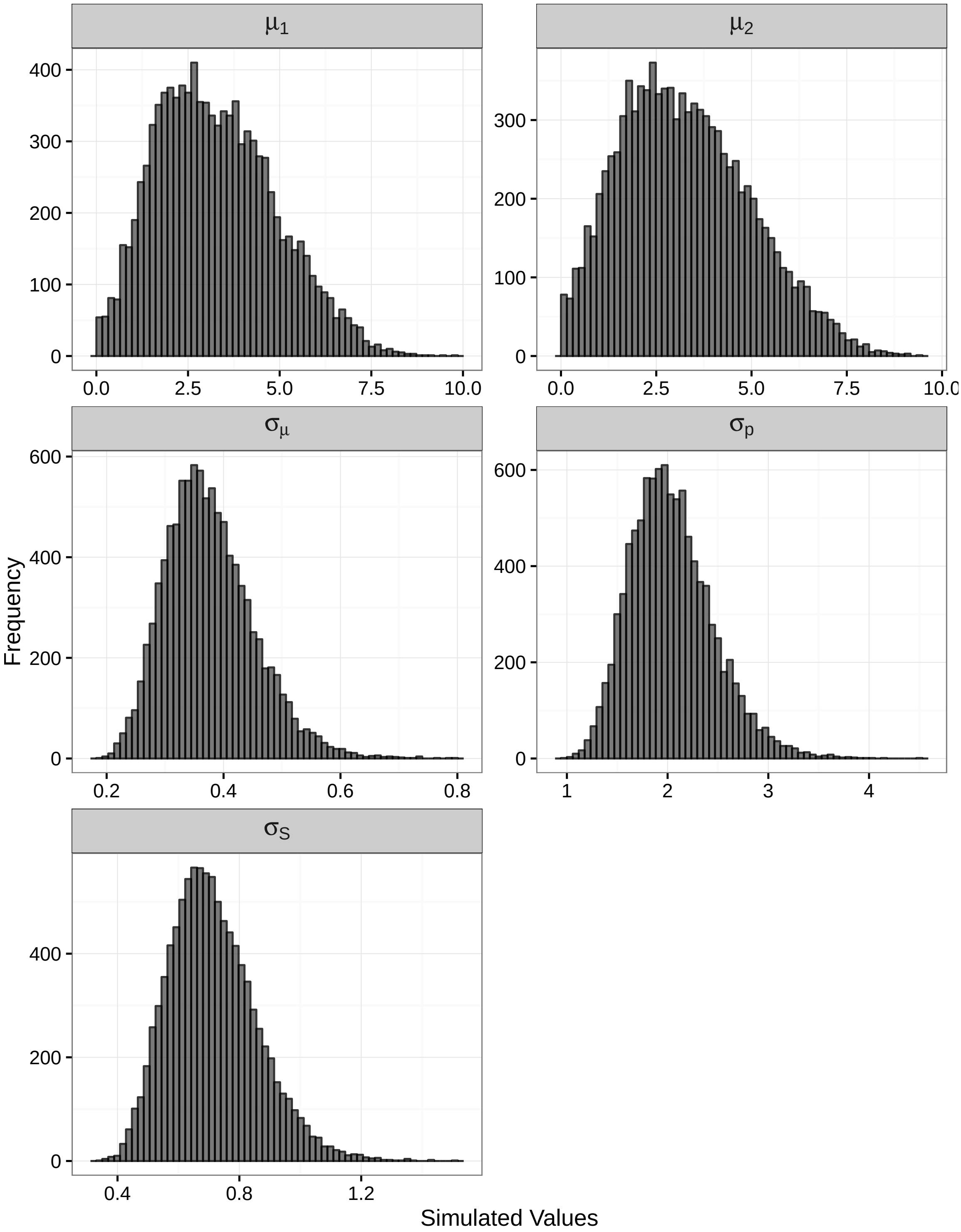


Figure 4.4 – *Key values generated for the 10,000 simulated genes, by the simulation procedure in Section 4.2.2 and Table 4.1*

4.3 Analysing Simulated Data Using the Package BMEA

4.3.1 DABG using Z-Scores

As the Mixture model of Section 3.3.1 proved to be a computationally complex undertaking, the ascribing of Log-Normal priors to the background signal component of the observed PM intensities still gives an opportunity to remove probes which contain no “true” signal prior to the model-fitting stages, but without the computational complexity of the Mixture model. This can be performed using a DABG-style (Detection Above Background) (Liu et al., 2002) approach, by comparing an observed PM intensity to it’s prior distribution for background signal. This relies on the property that if $S_{hijk} = 0$, then $PM_{hijk} = B_{hijk}$ and should be drawn from the relevant prior distribution.

A simple Z-score can be obtained for each PM observation by calculating

$$Z_{hijk} = \frac{\log PM_{hijk} - \lambda_{hijk}}{\delta_{hijk}}, \tag{4.1}$$

where λ_{hijk} represents the expected value of the Normal distribution for the bin to which the probe PM_{hijk} has been assigned. The set of Z-scores for a given gene (Z_g) or exon (Z_j) can then be combined using Stouffer’s Method (Demerath, 1949)

$$Z_j = \frac{\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{k=1}^{K_j} Z_{hijk}}{\sqrt{IK_j}} \tag{4.2a}$$

$$Z_g = \frac{\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J}\sum_{k=1}^{K_j} Z_{hijk}}{\sqrt{IK}} \tag{4.2b}$$

to give a simple assessment of the null hypothesis $H_0 : S = 0$, with alternative $H_A : S > 0$.

During fitting of this set of simulated data, the BMEA model was applied both with and without DABG filtering, as it was hypothesised that using a DABG approach may remove potentially spurious results in the case of shared transcript truncation across both cell types. This would occur most commonly in splicing pattern 3 (Figure 4.3C) and the expectation of this being a common occurrence when working in a true biological context is not unreasonable

The Z_j-scores obtained for all exons in the set of simulated data are shown in Figure

4.5, with 91.5% of exons simulated as missing receiving a score below the 95th quantile of the standard normal distribution ($Z = 1.645$). Exons which were only simulated as present in one condition were given $Z_j < 1.645$ 10.3% of the time, whilst constitutive exons were given $Z_j < 1.645$ only 4.5% of the time. As this seemed a suitable balance between Type I and Type II errors during DABG filtering, $Z_j < 1.645$ value was considered an appropriate criteria which would be expected to remove $> 90\%$ of non-expressed exons whilst retaining $> 90\%$ of alternately-spliced and constitutive exons.

Applying the Z_g scores revealed that even when using the 99.999th quantile of the standard normal ($Z = 4.265$), only 0.27% of genes would be excluded in this simulated dataset. Although all genes in this simulated dataset were simulated to contain signal, this suggests that when applied to real data, applying the Z-score at the gene-level will result in very few genes being inappropriately removed.

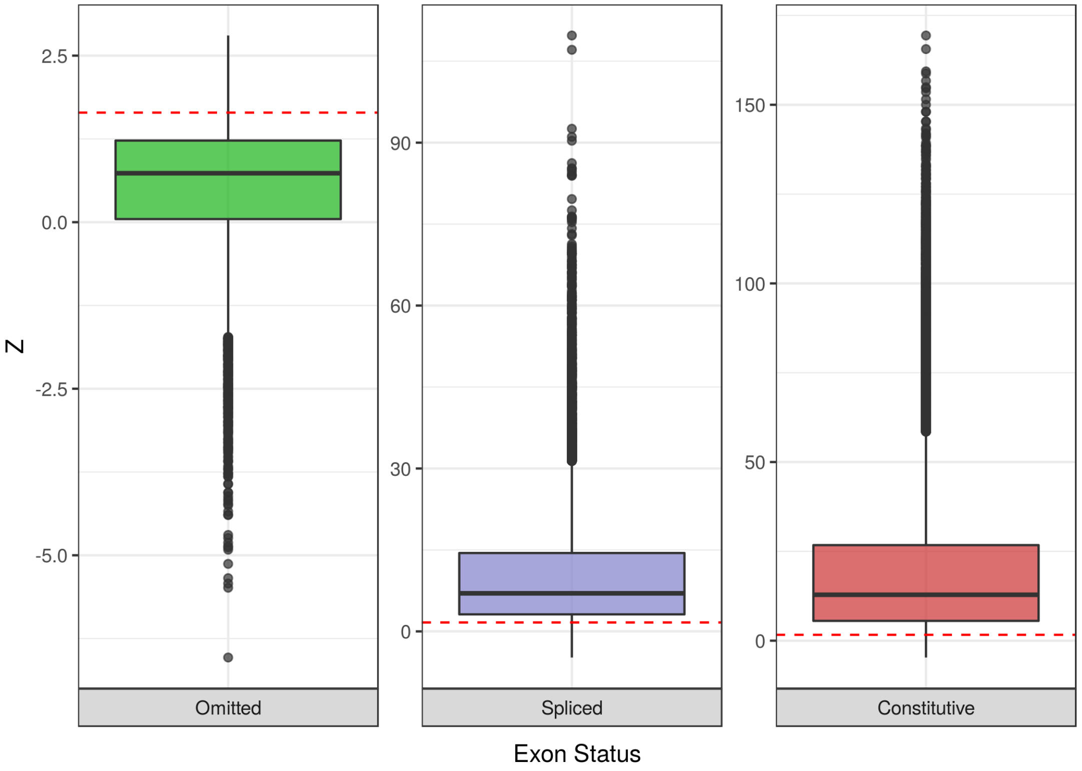


Figure 4.5 – *Z_j scores obtained for all simulated exons. Exons simulated as missing are shown in the left panel, whilst those simulated as having some splice variation are shown in the middle panel. Exons which were included in all samples are shown in the right panel. The 95^{th} quantile ($Z_j = 1.645$) of the standard normal distribution is indicated in red.*

4.3.2 Fitting the Complete Model

The BMEA model was applied to the simulated data using 3 independent MCMC chains and 12,000 iterations, discarding the first 6,000 as the burn-in stage. The thinning parameter was set to 6, giving 2000 retained MCMC iterations for each chain. Testing (not shown) showed this to be a good trade-off between computational time and parameter convergence.

In addition to fitting the BMEA model with and without Z-scores, simulated data was additionally analysed using RMA background correction and PLM/FIRMA approaches. Data was initially fit using the standard methods, with an additional analysis excluding exons for which $Z_j < 1.645$, prior to estimation of all PLM/FIRMA values . The standard methodology of *limma* (Smyth, 2005), including the use of moderated $\tilde{t}$-statistics (Smyth, 2004) was followed.

4.4 Results From Analysis of Simulated Data

4.4.1 Detection of Differentially Expressed Genes

The dataset of 10,000 simulated genes included 4,045 with non-zero fold-change, with the remainder not being simulated as differentially expressed. Analyses using *limma* provided estimates of the false discovery rate (FDR), with the threshold of $\alpha = 0.05$ yielding lists of 667 and 796 genes for the analyses without and with DABG (i.e. Z_j-score) exon filtering respectively. Within each analysis, the percentages of correctly identified DE genes were 76.2% and 81.5%, with the remainder being Type I errors. The true FDR was clearly > 0.05 in both analyses, with the analysis incorporating Z-score exon removal clearly identifying a higher number of DE genes, at a higher level of accuracy.

In order to address the higher than expected FDR in both analyses, the distributions of $\tilde{t}$-statistics for each RMA analysis were plotted (Figure 4.6), for the 5,955 simulations in which no fold-change was specified. These were separated by splicing pattern with some patterns having a clear impact on the observed distributions. Simulations with splicing patterns 1, 2 and 4 generally returned t-statistics in keeping with the expected distribution under H_0, as seen by the good fit between the IQR lines and the IQR seen in the boxplots. Pattern 3 also returned t-statistics as expected after removal of omitted exons with the Z-score, however, this distribution was far more restricted than expected when DABG filtering was not applied. Without the removal of exons, Pattern 5 returned t-statistics in keeping with the expected distribution, however after removal of undetectable exons the returned distribution under H_0 was a more heavy-tailed distribution than expected.

For all patterns in which Condition B contained a shorted transcript, the positive bias in $\tilde{t}$-statistics was as expected, given the positive bias in estimates of logFC previously noted (Section 3.2.5). In the most extreme pattern (i.e. 10) approximately 75% of $\tilde{t}$-statistics were beyond the 97.5$^{\text{th}}$ quantile of the theoretical distribution regardless of Z-score filtering. These findings reinforce the flaws in many previous approaches for analysis of Exon Array data in the presence of unknown transcript variation, where all parameters are fit without regard to this possibility.

An approximate null distribution for the B-statistic in this dataset was determined from the B-statistics returned in simulations using splicing pattern 1, and with no fold-change. The complete distributions of B-statistics were then compared to the IQR and central 95% region from this null distribution across all splicing patterns (Figure 4.7). Simulations with transcripts of the same length appeared to return B-statistics in keeping with the null distribution, however, simulations in which cell-type B contained a shorter transcript again showed the positive bias previously seen under PLM/RMA. Despite this bias, violations of the IQR for patterns 9 and 10 were much less prominent than was observed in the RMA/PLM analysis (Figure 4.6). The 25th percentile of the B-statistic from pattern 10 only just exceeded the 75th percentile of the null B-distribution, whereas the 25th percentile of the $\tilde{t}$-statistics under RMA/PLM exceeded the 97.5th percentile of the null $\tilde{t}$-distribution. No significant difference in the distribution of B-statistics was noted between analyses with or without the use of Z-scores for exon or gene removal.

A common approach for selection of DE genes is to initially select genes to a given FDR, then additionally filter on logFC beyond a range such as ±1 on the $\log_2$ scale, corresponding to a fold-change greater than 2. A similar strategy was applied to both sets of RMA and both sets of BMEA results for easy comparison between the approaches. For both RMA/PLM analyses, genes were sorted based on p-value. Due to the discrete nature of the BMEA B-statistic for these analyses genes were sorted by $|B|$ with ties broken by |logFC| for both BMEA analyses, using posterior means as point estimates. After sorting, all lists were filtered to only contain those with estimated $\log_2$FC beyond the range ±0.4 which corresponds to a fold-change greater than 1.5. Given the simulation parameters, this provided a degree of variability around the discrete values for true fold change (±2 and ±4) as simulated in this dataset.

The top 500 genes from these filtered lists were selected as the candidate DE genes, which would be roughly analogous to selecting the top 1000 in a true biological dataset. This corresponded to an expected FDR of 0.050 and 0.068 for the PLM/RMA analyses with and without Z_j-score exon filtering respectively. The totals in each list which were true positives are given in Table 4.2, along with the observed FDR. Both BMEA analyses performed identically, and returned more accurate results than both RMA/PLM approaches. As expected, Z-score filtering gave a slightly improved performance when using RMA/PLM, although the observed FDR was again much higher than predicted.

Table 4.2 – *Accuracy of differential expression results from the most highly ranked 500 genes as obtained under the four approaches, after filtering for genes with fold-change estimates beyond ±1.5. The number of genes correctly identified is indicated, along with the True Positive Rate (TPR) with observed and expected False Discovery Rates within the sets of results. As no FDR calculations are included in BMEA, these values are given as dashes.*

			FDR	
Method	**True Positives**	**TPR**	**Observed**	**Expected**
BMEA	494	0.988	0.012	-
BMEA (Z)	494	0.988	0.012	-
RMA/PLM	421	0.842	0.158	0.068
RMA/PLM (Z)	429	0.858	0.142	0.050

Given the noted bias as a result of differing transcript lengths (Figures 4.6 and 4.7), ROC curves for the four analytic approaches were generated for simulations where there was no difference in transcript lengths, and for simulations where condition B contained the shorter transcript (Figure 4.8). BMEA strongly outperformed RMA/PLM for both sets of simulations, with the versions excluding undetectable exons generally performing more weakly than the analyses for which Z-scores were not used. The exception to this was for RMA/PLM in the presence of different transcript lengths (Figure 4.8B).

As a final measure of gene-level performance of each analytic approach, the point was found at which the ranked lists reached an observed FDR of 0.05 (Table 4.3). Again, BMEA performed more strongly than RMA/PLM, with the best performance coming from the analysis without Z-score based exon removal. For the set of simulations containing transcripts of different lengths, the 3$^{\text{rd}}$ gene in both RMA/PLM lists was a Type I error, giving an observed FDR of 0.33 at this point. In all cases, models performed far more strongly when transcripts were the same length in both conditions.

Table 4.3 – *Number of genes detected as DE before the observed FDR in this simulated dataset exceeded 0.05.*

Transcript Type	**BMEA**	**BMEA (Z)**	**RMA/PLM**	**RMA/PLM (Z)**
Same Length	987	926	518	176
Shorter in Condition B	586	491	2	2

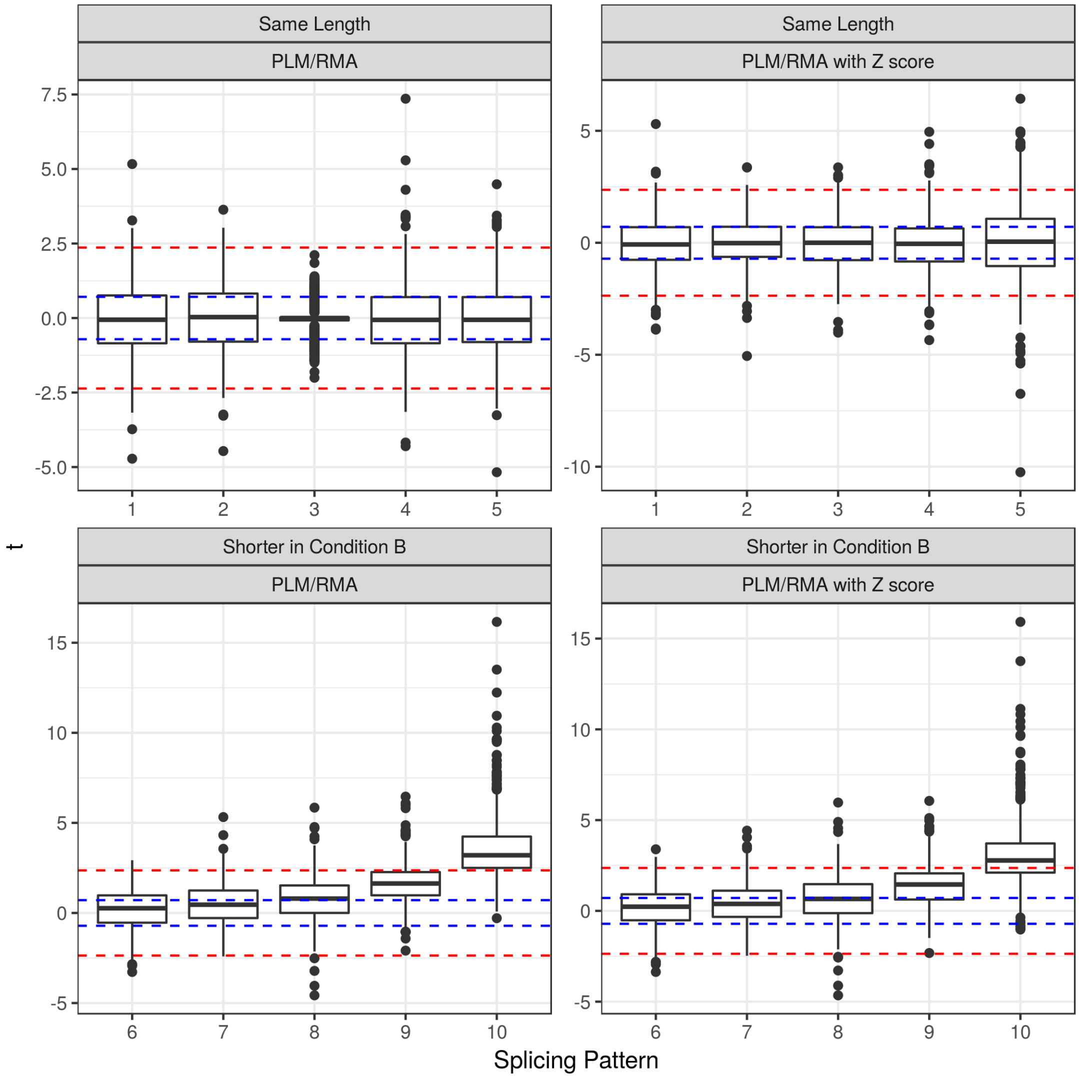


Figure 4.6 – *Moderated $\tilde{t}$-statistics for simulated genes with no specified fold-change, separated by splicing patterns. The IQR for a T distribution with 7 degrees of freedom is indicated by the dashed blue line to provide a visual guide for the IQR of the observed T-statistics and the expected theoretical IQR. The central 95% region for T_7 is indicated by the dashed red line.*

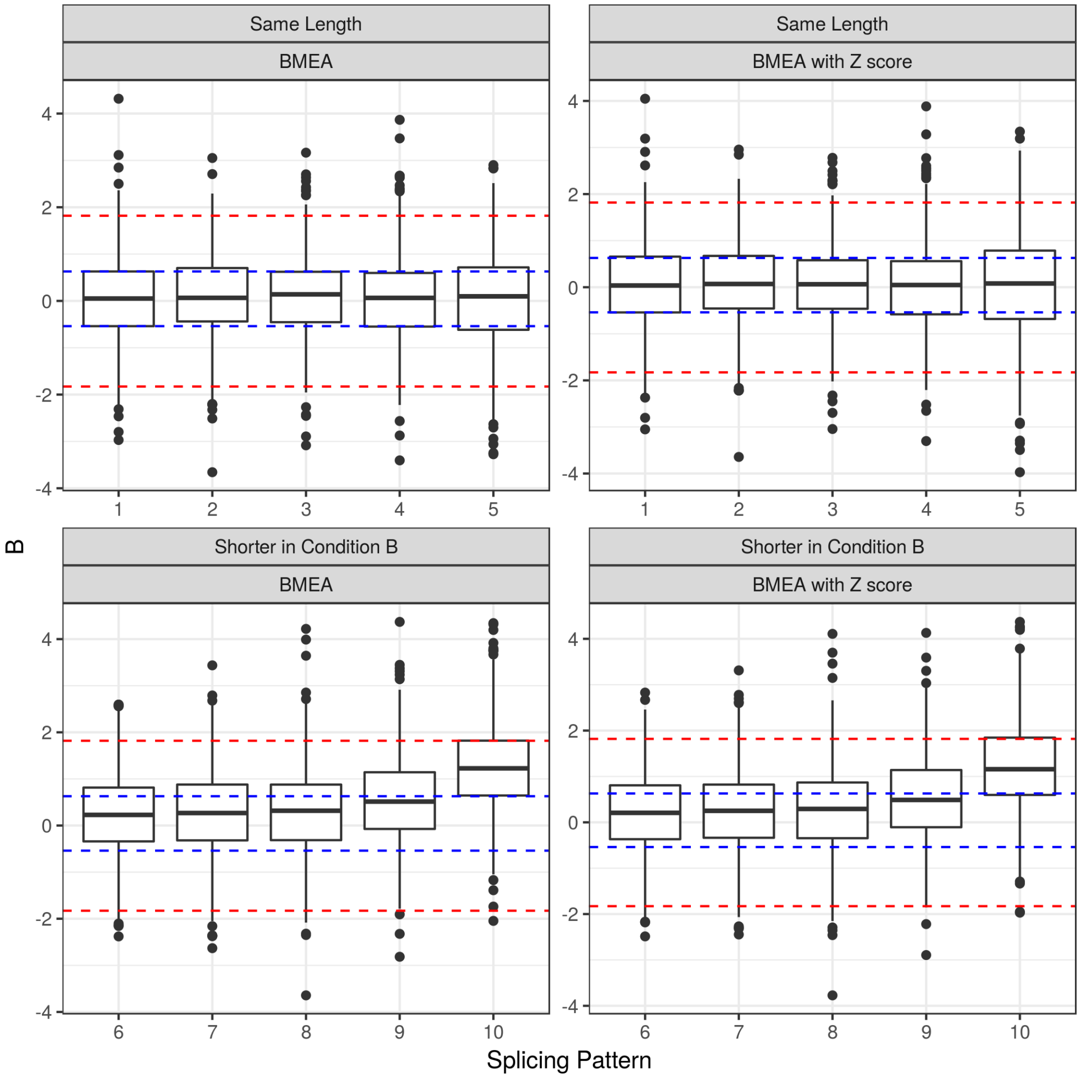
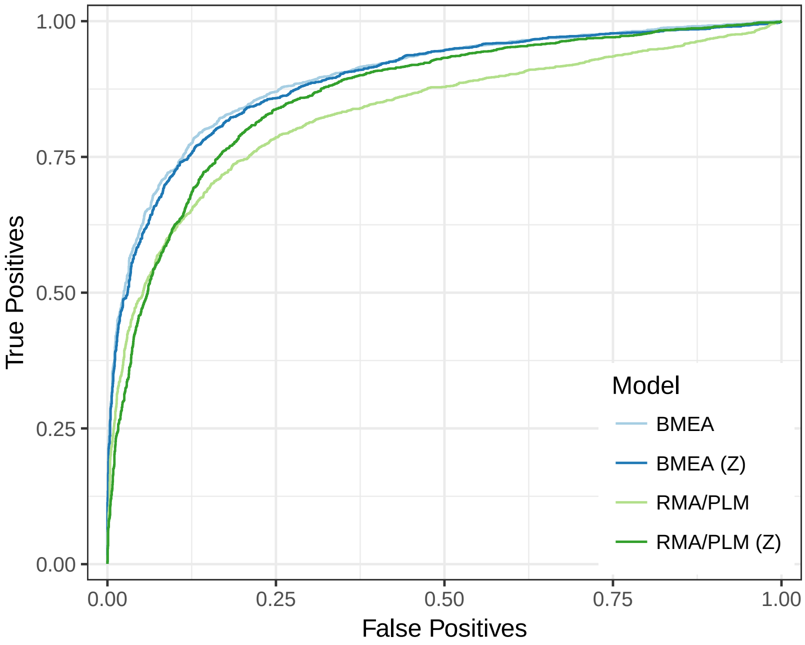
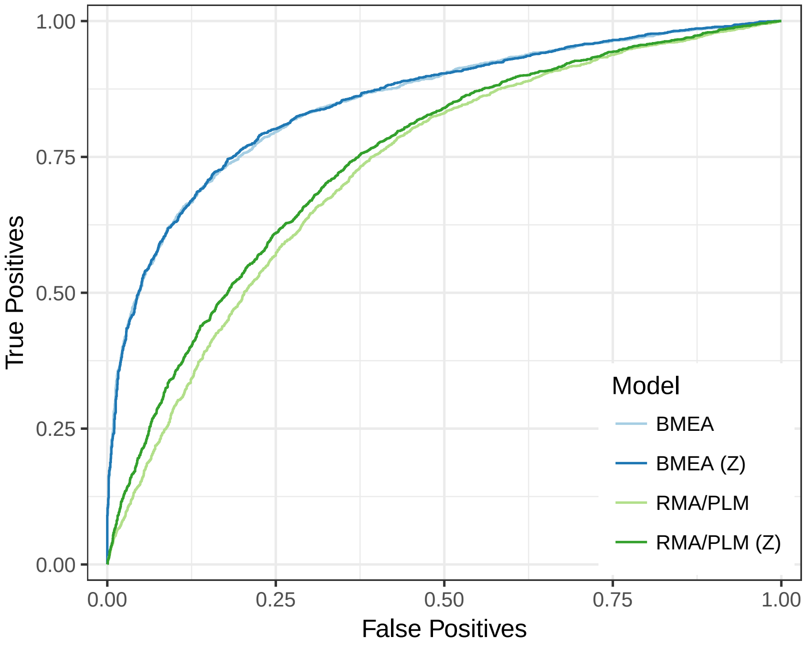


Figure 4.7 – *B-statistics for simulated genes with no specified fold-change, separated by splicing patterns. An empirical null distribution for the B-statistic was determined from the set of simulations using splicing pattern 1, and with no added logFC. The IQR from the empirical null B distribution is indicated by the dashed blue line to provide a visual guide for the expected theoretical IQR. The central 95% region for the null distribution is indicated by the dashed red line.*



(A) *ROC curve for detection of DE genes only including simulations with no difference in transcript length between conditions A and B.*



(B) *ROC curve for detection of DE genes only including simulations where condition B contains a shorter transcript.*

Figure 4.8 – *ROC curves for detection of DE genes using BMEA or the RMA/PLM models, both with and without the additional filtering of undetectable exons using the Z-score. Curves are plotted for A) simulations with no difference in transcript lengths between conditions, and B) simulations where condition B contains a shorter transcript.*

4.4.2 Detection of Alternately Spliced Exons

The dataset was generated with 19,064 of the 100,000 exons being simulated as being alternately spliced (AS) exons. For the PLM fitted datasets, FIRMA scores were obtained for each exon across all samples and moderated $\tilde{t}$-statistics were used to obtain FDR-adjusted p-values. An FDR < 0.05 was used as the initial criteria for declaring a candidate AS exon under FIRMA, and the two PLM fits were assessed for accuracy (Table 4.4), noting that the use of Z_j scores reduced the total number of exons under consideration by 11.1%. Again, the observed FDR was far higher than the expected FDR, with the conventional FIRMA model outperforming the one which incorporated DABG-removal of undetectable exons.

In order to compare the results across all four approaches, the 5,000 most highly ranked exons were selected and the identification rates of correct AS events were compared (Table 4.5). BMEA results were ranked using $|B|$, with ties broken using posterior means from $\Delta \log \phi$. The observed rate of correctly identified AS exons was clearly higher using both BMEA approaches, however the use of Z-scores to remove undetectable exons had a minimal impact on the rate of correctly identified events.

Table 4.4 – *Results for detection of alternatively spliced (AS) exons using FIRMA with and without Z-scores. Exons considered as significantly AS under FIRMA are listed as Detected AS Exons, with the true number within this set also given as True AS Exons. Significance was determined by an FDR-adjusted p-value* < 0.05.

Method	**Detected AS Exons**	**True AS Exons**	**TPR**	**FDR**
FIRMA	17,976	14,373	0.800	0.200
FIRMA (Z)	17,200	13,526	0.786	0.214

Table 4.5 – *Correctly identified AS events from the most highly ranked 5,000 exons using all four approaches. The use of the Z-score for DABG filtering of exons is indicated by the presence of a Z in brackets. The true FDR is provided in the final column.*

Method	**True AS Exons**	**TPR**	**FDR**
BMEA	5000	1.0000	0.0000
BMEA (Z)	5000	1.0000	0.0000
FIRMA	4975	0.9950	0.0050
FIRMA (Z)	4987	0.9974	0.0026

The $\tilde{t}$-statistics for exons simulated with no splice variation were inspected for both the FIRMA analyses, separated by whether logFC was included in the simulated data points (Figure 4.9). When no fold-change was included in the simulation, distributions more closely resembled the expected T_7 distribution, however, distributions under H_0 were clearly more heavy tailed in the presence of fold-change. Whilst 3-5 points were expected beyond the $1/10000^{\text{th}}$ quantile, the numbers of points falling beyond this range were slightly above these numbers in all analyses except for the sets of statistics returned with no logFC included in simulations. A subtle downwards bias was also noted in all sets of $\tilde{t}$-statistics with mean values ranging between -0.243 and -0.230. Whilst no theoretical distribution was defined for the BMEA B-statistic, a similar downwards bias was noted, with mean values ranging from -0.206 to -0.178 (not shown). The 95% central region for the B-statistics ranged between -2.59 to 2.49 from the mean, which was not dissimilar to values expected under T_ν with $\nu \approx 5$.

An ROC curve across the entire set of genes was also plotted for each of the four approaches (Figure 4.10A), with the removal of exons using Z-scores leading to a clear improvement in performance for the BMEA analysed data, and a subtle improvement for the FIRMA analysed data. Both BMEA approaches strongly outperformed the FIRMA approaches.

The impact of non-zero fold-change was also assessed on model performance (Figure 4.10B). Separate ROC curves were generated for simulations including fold-change and those without. The presence of non-zero fold-change led to a small reduction in performance for BMEA using Z-scores, but had an indistinguishable effect on BMEA without DABG filtering. However the difference under PLM/FIRMA was very stark. For simulations without any fold-change, the performance of FIRMA was still weaker than BMEA, but appeared relatively competitive with the version of BMEA which didn't include DABG exon-filtering. For simulations which included non-zero fold-change, FIRMA showed significantly reduced performance.

To assess whether the overall expression level also had an impact on the accuracy of AS detection, separate ROC curves were plotted for each quantile of genes, ranked from highest expression level (Q1) to lowest (Q4) (Figure 4.11). Clearly different behaviours were shown when breaking the data down in this manner, with BMEA still generally outperforming FIRMA. Both BMEA approaches were indistinguishable in the most highly expressed genes, with the addition of the Z_j-score filtering having a greater impact on the results through the lower expressed genes. In the quantile of genes with the lowest expression level, the curves for both FIRMA approaches overtook the BMEA curves as the False Positive Rates approached 0.025 and 0.05, indicating FIRMA as the preferable approach for genes with low expression levels.

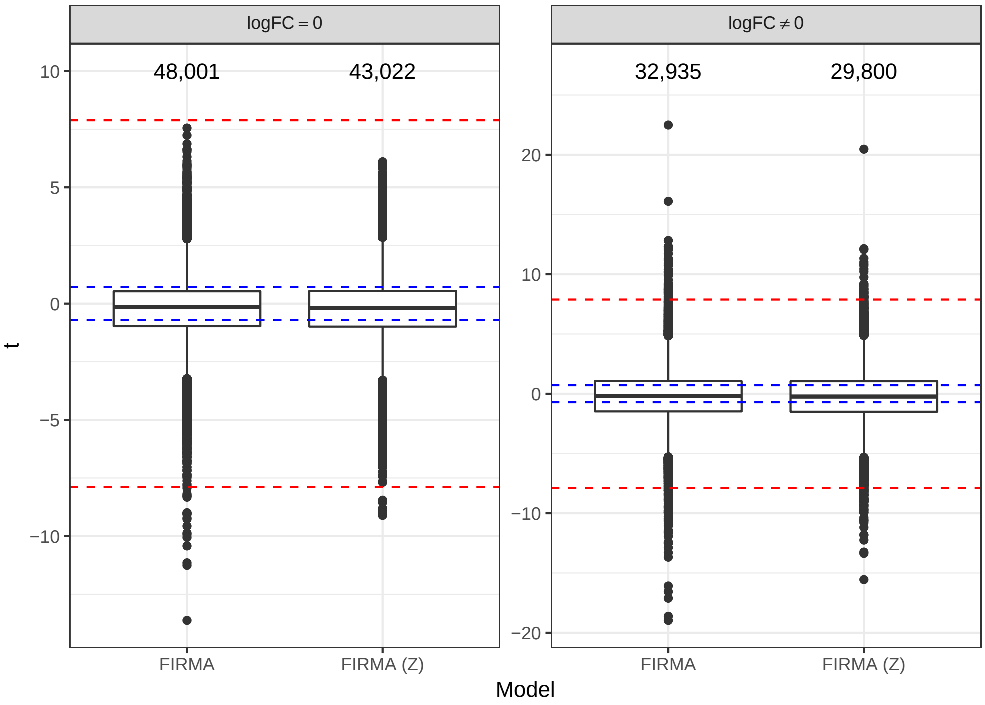
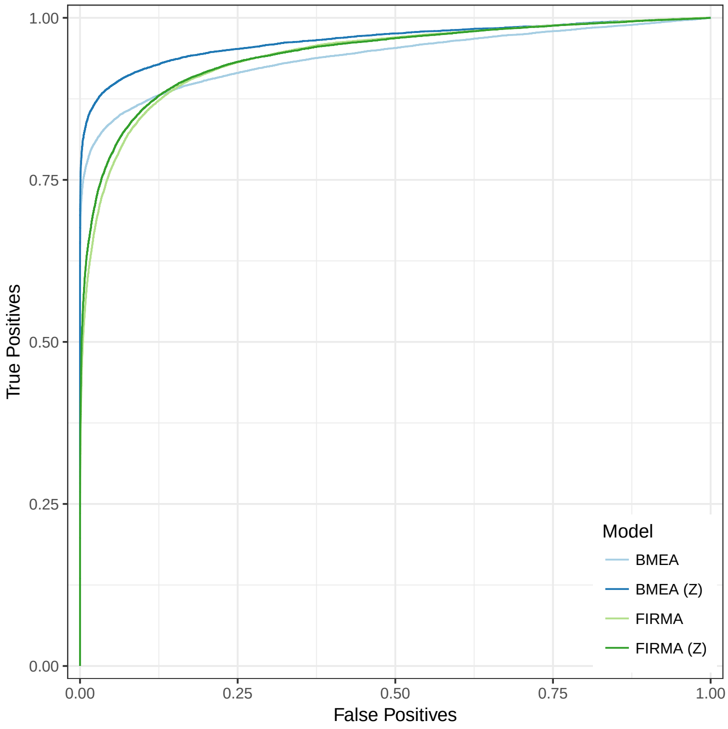
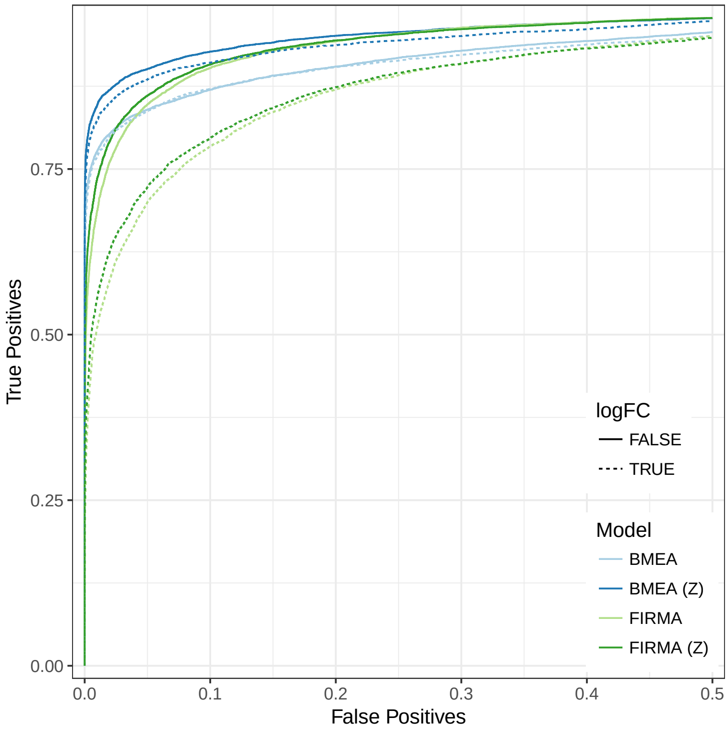


Figure 4.9 – *Moderated $\tilde{t}$-statistics for exons simulated with no splice variation, separated by genes simulated with and without logFC. The IQR for a T-distribution with 7 degrees of freedom is indicated by the dashed blue line to provide a visual guide for the IQR of the observed T-statistics and the expected theoretical IQR. The 1/10,000th quantile is indicated by the dashed red lines, with the numbers of exons in each category given above each boxplot to provide a guide as to how many $\tilde{t}$-statistics would be expected outside of this range under H_0.*



(A) *Overall ROC curve for detection of simulated AS events.*



(B) *ROC curves for detection of simulated AS events broken down by presence or absence of fold-change. Dashed lines indicate simulations which contained non-zero fold-change, whilst solid lines indicate simulations with no simulated fold-change.*

Figure 4.10 – *ROC curves for detection of simulated AS events for the 10,000 genes from Section 4.2. The A) overall curves are presented along with B) curves separated by the presence/absence of fold-change.*

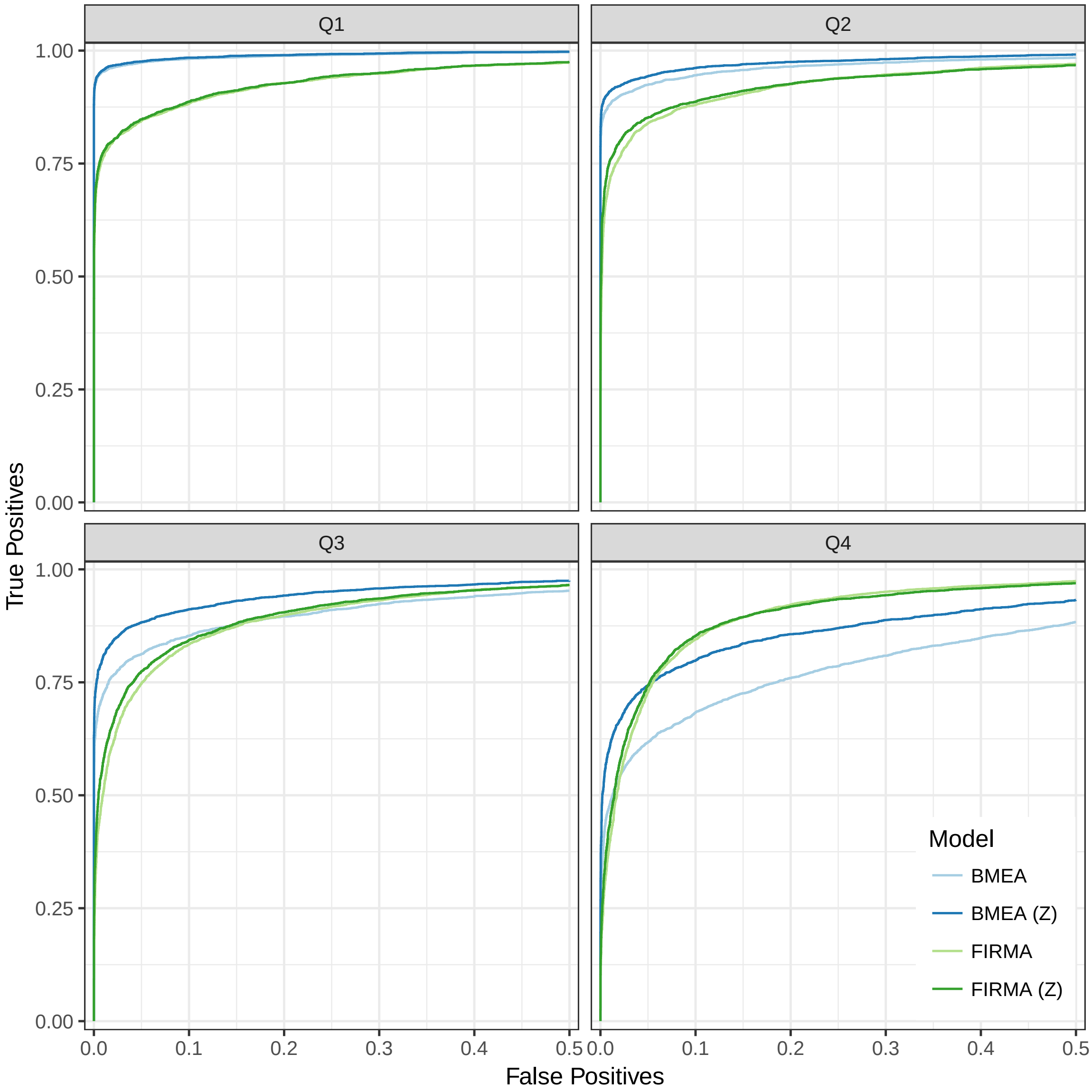


Figure 4.11 – *ROC curve for detection of simulated AS events broken down by expression level quartiles, ranked for highest to lowest expressed. The data was the 10,000 simulated genes from Section 4.2, as analysed using BMEA and FIRMA approaches, both with and without Z-scores based on DABG.*

4.5 Discussion

In the results obtained from this simulated dataset, BMEA outperformed RMA-based approaches both in detection of DE genes (Section 4.4.1) and detection of AS exons (Section 4.4.2). Whilst not offering any tangible improvement to the detection of DE genes under the BMEA approach, exon filtering using Z_j-scores offered a very slight improvement to the detection of AS exons. However, it should be noted that ~11.1% of the truly AS exons were removed from the dataset by this filtering approach, giving even a perfect analysis a statistical power of 89.9%. Still, the removal of genuinely undetectable exons remains a conceptually sound strategy during analysis.

It was also noted that for all approaches, the best performance was observed for the most highly expressed genes, and filtering based on expression levels may also be an important consideration when analysing experimentally-derived data. These genes will also be less likely to suffer from incorrectly filtered exons which are expressed in only a subset of samples. Similarly restriction to genes with near-zero fold-change was able to offer a small improvement in the performance of the model.

A small improvement after Z_j-score filtering was noted in the detection of DE genes under the PLM approach. Whilst not strictly being relevant to the BMEA model, this has implications for the wider analysis of whole transcript arrays and could become an important part of these protocols for all researchers.

It should also be noted that 70% of genes in this simulated dataset contained an AS event, and this may be an over-representation of this phenomenon compared to a true experimental dataset. This also equated to ~19% of exons, which again is likely to be an over-representation. Whilst this may have served to exaggerate the advantages of BMEA over PLM/FIRMA, the difference between the two approaches is still clear. As mentioned in Section 3.5, data analysed here has also been simulated using the same model which underlies BMEA. Whilst this model shares a great deal with that for RMA/PLM, any bias introduced to model comparisons by this strategy is difficult to directly quantify.

When directly comparing the two BMEA approaches, Z-score filtering brought subtle improvement across all measured values, including computational time. The removal of some exons within transcripts at the low end of the range of expression values is the primary drawback for this approach, however as this would lessen the chance of false discoveries overall, Z-score filtering appears to be a useful improvement to the BMEA model, and goes some way to incorporating an aspect of the Mixture Model from Section 3.3.1.

As well as the general performance of the BMEA approach, several key behaviours of analysis using RMA/FIRMA were brought to light by the above dataset. Firstly, the increased variability in $\tilde{t}$-statistics for logFC in the presence of *shared missing* exons (Figure 4.6) has implications for both Exon and Gene Arrays, and for the selection of appropriate CDF for probe-to-gene mappings. Secondly, the bias introduced as a result of differing transcript lengths across treatment groups was made clear and becomes an important consideration for all whole-transcript arrays. Whilst BMEA seems to be less significantly affected by this phenomenon (Figure 4.7), it remains an issue of note. Comparisons involving treatment groups with a high expected rate of splice variation may indeed not be suitable for analysis using whole transcript arrays.

In addition to the impact of splice variation on estimates of fold-change and the associated $\tilde{t}$-statistics, conventional analysis using FIRMA appears to generate a more heavy-tailed distribution than may be expected under the null hypothesis (Figure 4.9). The assumptions of normality may not hold for this statistic and as such, this approach is likely to lead to numerous spurious results, confirming the caution raised in Section 2.5.2.

In conclusion, this dataset was large enough to robustly test the capacity of the model and the package designed to implement BMEA as an analytic strategy. The use of observed NSB from the set of AG probes additionally ensured that this component of the simulated data was relatively realistic, notwithstanding any impact of competitive binding by the true target sequences of a PM probe.

Chapter 5

Application of BMEA to Experimental Datasets

5.1 Introduction To Experimental Datasets

For true assessment of the BMEA algorithm, experimentally-derived data is a necessity as results will not contain any implicit bias from simulation algorithms, and all intensities will be truly representative of genuine data. This chapter presents an assessment of BMEA using two reference datasets, before returning to the T_{reg} dataset in Chapter 6, and attempting to uncover this fine-level of detail in the biological context of primary interest. The two reference datasets, described immediately below, were also used for assessment of MMBGX (Turro et al., 2010), and this enabled a degree of comparison to FIRMA and MMBGX, with both of these shown to be improvements on previous approaches such as the Splicing Index and MIDAS.

The Gardina Dataset is publicly available from `www.affymetrix.com`, and is taken from a set of 20 donor-matched Tumour-Normal comparisons in colon cancer (Gardina et al., 2006). The published analysis assessed alternate splicing (AS) events using methodologies introduced by Affymetrix, such as MIDAS and the Splicing Index (Section 1.7.2). A selection of putative AS events were assessed individually using PCR gels, giving a set of genes for which the splicing status is known and can be confirmed visually.

The Tissue Mixture Dataset represent combinations of Brain and Heart Tissues in a series of changing proportions (Table 5.1). Comparisons between the 100% tissue samples can be used to obtain a list of putative DE genes and AS exons, with remaining samples available for verification of changes across the changing mixture levels, in accordance with those predicted by the initial comparison. In addition, the 3 sets of technical replicates with an identical 50:50 mixture of Brain and Heart Tissue provide an effective set of negative controls, in which no differential expression or alternate splicing is expected to be present. Raw data is also available from `www.affymetrix.com`.

Table 5.1 – *Mixture concentrations for the Tissue Mixture Exon Array Dataset. Three technical replicate arrays were present at each mixture level, with the exception of Mix 5, where three replicate mixtures (a, b & c) were performed, each with three technical replicates, giving nine replicates for this mixture.*

Tissue	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5a/b/c	Mix 6	Mix 7	Mix 8	Mix 9
Brain	.00	.05	.10	.25	.50	.75	.90	.95	1.00
Heart	1.00	.95	.90	.75	.50	.25	.10	.05	.00

5.2 Additional Preparations for Experimental Dataset

5.2.1 Alternate Ranking Methods

All previous testing had been performed using the BMEA B-statistic in conjunction with estimates of fold-change ($\Delta\mu$ or $\Delta \log \phi_j$) to provide suitable candidates. Due to the fixed number of retained MCMC iterations across all genes, the BMEA B-statistic is essentially discrete and is limited in it's ability to differentiate extreme results. A more Bayesian approach would be to use the Central Posterior Interval (CPI), which has been shown to control the FWER in an analogous way to the frequentist confidence interval (Gelman et al., 2004), such that use of a 1-α% CPI retains an experiment-wide error rate at the level α. Thus the use a 95% CPI which excludes zero as an alternative selection criteria should maintain a 5% error rate across a set of parallel comparisons at either the gene or exon level.

This methodology can be further extended by using a 95% CPI which excludes *any interval*, such as symmetrical region around zero $[-\kappa, \kappa]$, where $\kappa \in \mathbb{R}$. This can be alternatively parametrised as a 95% CPI which excludes zero, and with a lower bound (CPI-LB) $> \kappa$, where $\kappa = \min(|\theta_\alpha|, |\theta_{1-\alpha}|)$ for $\alpha = 0.025$, and where subscripts denote quantiles of the ordered statistic θ.

5.2.2 Creation Of A Custom CDF

In previous analyses two CDF files had been utilised, serving different purposes. For gene-level analysis in Section 2.3, a CDF with no exon-level structure was used, but with all genes mapped to Entrez Gene IDs after strict QC processes. The second, publicly-available CDF (Section 2.5) did not take advantage of these QC processes and instead used exon-level probesets (or groups) corresponding to the original groups defined on the unsupported Affymetrix CDF, with groups assigned to Ensembl IDs as the primary unit instead of the "transcript clusters" defined by Affymetrix.

As the number of probes for an exon are considerably smaller than for a gene, ensuring that observed exon-level signal was from the intended target is even more essential, and a custom CDF was designed to incorporate the additional quality filtering. This custom CDF was created utilising v14.1 of two separate CDF files available at the BrainArray website (`http://brainarray.mbni.med.umich.edu/www/data-analysis/custom-cdf/`), the first of which mapped the probes to Ensembl exon (ENSE) identifiers as gene-level probesets, whilst the second file mapped each probe to an Ensembl gene (ENSG) identifier, again at the gene-level. Due to the nature of ENSE identifiers, many probes in this CDF were mapped to multiple identifiers which defined overlapping but distinct splice variants within the same gene. All annotations were based on the hg19 genome build.

In brief, the common set of probes was found between the two CDF files, and mappings to units were assigned based on ENSG identifiers, whilst the mappings to groups were based on probes with shared ENSE identifiers. This gave a CDF with 1,470,651 unique probe-to-genome mappings, and with 289,999 exon level probe groups making up 35,202 gene units. In addition to the quality control stages provided by the use of the BrainArray CDFs, this process gave a structure in which the strict mapping of up to four probes to a given exon was *no longer enforced* (Figure 5.1). Whilst 55.7% of exon-level probesets still contained four or fewer probes, this gave a considerable proportion of probesets on the array which can be expected to provide more stable signal estimates across an exon, or the targeted region of the relevant transcript.

It was also noted that a small number of exon-level probesets contained an unexpectedly large number of probes with identical sequence information, that were allocated to separate groups on the original Affymetrix CDF and were located at distinct physical points on the array. Any reasons for this were not explored further and these were simply assumed to provide more stable signal estimates over the course of the inevitable averaging during model fitting. This custom CDF was then used for any subsequent analytic steps, as presented below.

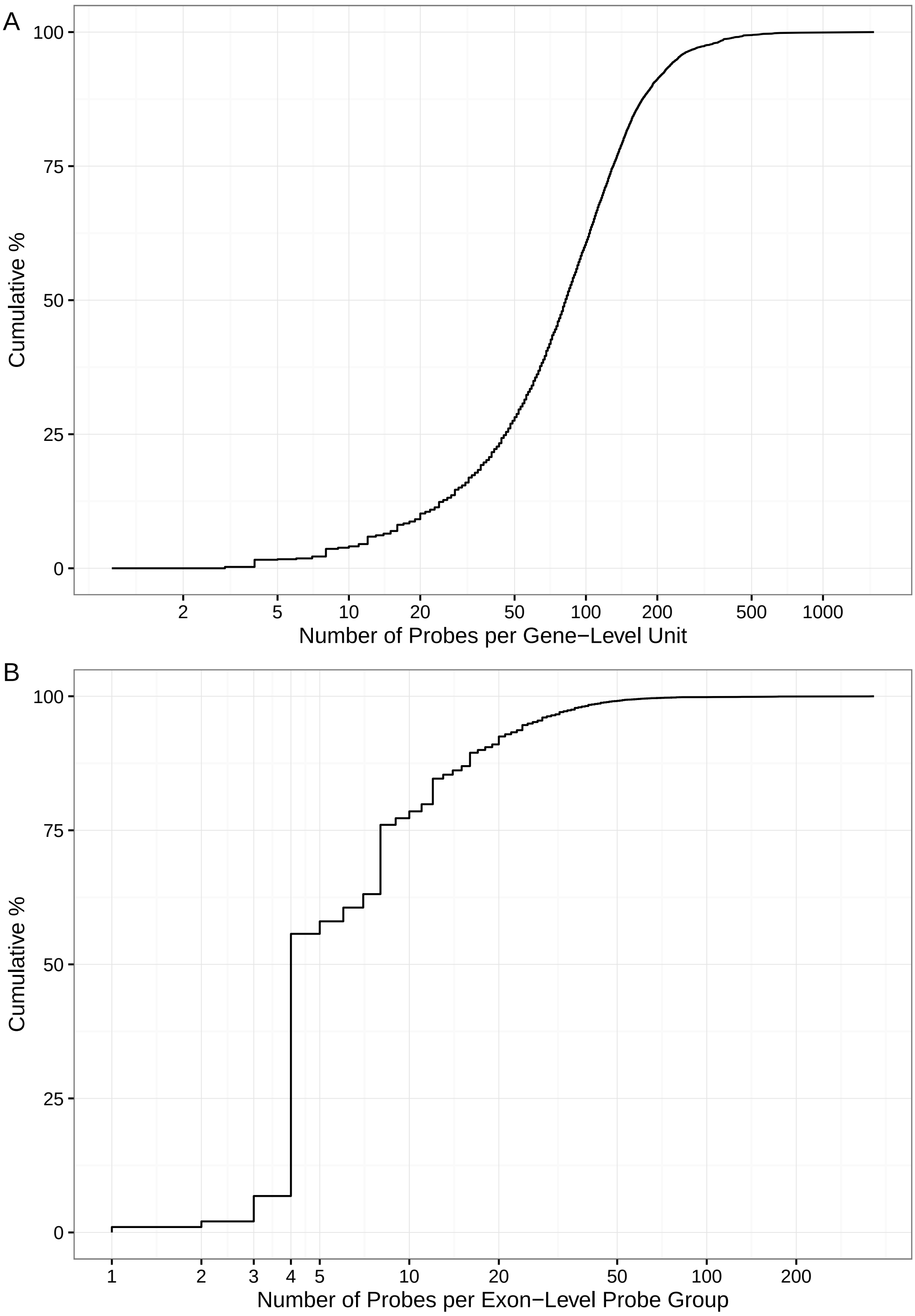


Figure 5.1 – *Cumulative distribution of probe numbers assigned to A) each gene-level "unit", and B) each exon-level probeset "group" on the custom CDF generated for all subsequent analyses. Under this new design, 44.3% of exon-level probe groups were able to contain* > 4 *probes, whilst* $< 10\%$ *of groups contained fewer than 4 probes.*

5.3 Analysis of the Gardina Dataset

5.3.1 Published Analysis

The original workflow (Gardina et al., 2006) exclusively utilised Affymetrix-proposed approaches and software, such as PLIER (Affymetrix, 2005d), MIDAS (Affymetrix, 2005a) and the software ExACT, as well as using the default annotations for genes and exons. Patient 3 was removed from their analysis after inspection using PCA. In brief, expression values were fitted separately for genes and exons using PLIER-based background correction. Multiple filtering criteria were applied based on DABG p-values and expression levels, before merging the parallel analyses for detection of AS events using MIDAS. Further unspecified filtering of results was then applied based on expression-levels and Splicing Index (SI) based fold-change, with a final list of putative AS events being determined by $p \leq 0.005$ from the SI analysis. No formal estimates of FDR or FWER control procedures were provided. All putative events were inspected visually and selected manually to provide a list of 189 candidate AS events across 162 genes.

In order to verify the putative AS events, details were provided for 73 PCR gels investigating 95 exon-level probesets, with potential AS events across 49 genes. PCR gel images were provided for 13 of these genes, with some images providing visual confirmation for single exon inclusion/omission events (e.g. Exon 21 for *ATP2B4*), whilst other gels provided indication of multiple splicing events (e.g the complex patterns of exon inclusion for exons 12, 13 and 14 for *CD44*). By implication, the additional probeset 3569827 from *ACTN1* was able to be visually assessed from the PCR gel for probeset 3569830 as this was the mutually exclusive exon (MEE) for the two isoforms. 16 of the tested splice events were drawn from the list of 189 candidate events from the defined workflow, whilst an additional 14 were splicing events previously reported in colon cancer. The remaining 66 tested probesets were assumed to be from a general workflow *not specified* in the published text. The gels as published are summarised in Table 5.2.

In addition to the gel images, a table of results indicated that moderate evidence was detected for splicing events in the genes *GK* and *MAST2*, however no images were provided for these genes. Evidence was classified by the authors as “Weak” for AS events in *LGR5*, *ZAK* and *FXYD6* with no gel images provided. Of these, only *ZAK* was included amongst the list of top 189 probesets, and other genes, along with *CTTN* and *SLC3A2*, were assumed to be from the unspecified workflow. For comparative analysis with BMEA, probesets denoted as showing “Weak” or stronger evidence were considered as confirmed, given the low precision of manual gel inspection over more accurate quantitative methods such as qPCR.

This gave a set of 71 probesets with unconfirmed AS events (i.e. True Negatives), with gel images available for 5 of these. However, it should be noted that this does not unequivocally eliminate AS events at these sites, as the resolution of the gels may have been inadequate for detection of these events. The remaining 24 probesets had supporting evidence for 23 alternate splicing events (i.e. True Positives), and 16 of these were able to be independently confirmed by inspection of the PCR gel images.

Table 5.2 – *Results as presented by Gardina et al., 2006 in Additional File 4, with PCR gel images published. Results are classified as originally defined, with NC indicating "Not Confirmed". In addition to the original results, probesets are classified based on the origins of the tested AS event. The source of each tested AS event is indicated as being from the published list of 189 (Top189), in the list of previously reported events (PR) or from the unspecified general workflow (GW). Probeset 3569827 from ACTN1 was tested as a Mutually Exclusive Exon (MEE) for probeset 3569830, as was Probeset 3597388 from TPM1. These additional probesets were not included in the initial Top189.*

Gene	Probeset ID	Result	Known AS Event	Source
ACTN1	3569830	Good	Ex19 (Ex20 is MEE)	Top189
	3569827	Good	Ex20 (Ex19 is MEE)	MEE
ATP2B4	2375766	Good	Ex21 is CE	Top189
CALD1	3025632	Good	extended Ex5; Ex6 is CE	Top189
COL6A3	2605386	Good	Ex6 is CE	Top189
	2605390	Good	Ex4 is CE	Top189
	2605391	Good	Ex3 is CE	Top189
CTTN	3338589	Good	Ex11 is CE	GW
FN1	2598321	Good	Ex25 is CE	Top189
ITGB4	3735208	Good	Ex35 is CE	PR
SLC3A2	3333718	Good	Ex2, 3, 4 are CEs	GW
TPM1	3597384	Good	Ex7 (Ex8 is MEE)	Top189
	3597388	Good	Ex8 (Ex7 is MEE)	MEE
VCL	3252129	Good	Ex19 is CE	Top189
CD44	3326711	Good (some)	Ex 12 is CE	PR
	3326712	Good (some)	Ex 13 is CE	PR
	3326714	Good (some)	Ex 14 is CE	PR
MST1R	2674945	NC	Ex11 is CE	PR
SIAHBP1	3157834	NC	Ex5 is CE	PR
(*PUF60*)	3157838	NC	Ex2 is CE	PR
VEGF	2908196	NC	Ex6 is CE	PR
	2908200	NC	Ex7 is CE	PR
RAC1	2989068	Weak	Ex4 is CE	PR

5.3.2 Analysis Using BMEA

Methods

In order to directly compare the published results with BMEA, the CDF with Affymetrix-defined exon-level groups was used (Section 2.5), but with these assigned to the relevant Ensembl ID. Eleven probesets under consideration were omitted from this CDF due to failure of the QC stages during construction of the individual source CDF files. This reduced the total to 84 probesets across 47 unique Ensembl IDs, for which the AS status was known. One of the missing probesets was 3422185 (*LGR5*) which showed weak evidence for an AS event, however the probeset 3422189 targeting the same event was taken as an alternate probeset. Whilst no gel image was provided for this AS event, this still gave 23 confirmed and 61 unconfirmed AS events.

In contrast to the published analysis, no PCA artefact was detected for patient 3 and this patient was retained for all downstream analysis. The entire set of arrays was quantile normalised and the BMEA process was performed on these 47 genes incorporating Z-score filtering (Section 4.3.1), using the default MCMC settings of 3 chains, 12000 iterations and 6000 burn-in iterations, with 1000 retained samples from each chain. Probesets were ranked based on the absolute value of the B-statistic for visualisation, and exons were considered as candidate AS events if the 95% CPI excluded zero.

Results

Of the 23 AS events confirmed by Gardina *et al*, 20 were detected by BMEA (Table 5.3). This contrasts with the 11 from this list which were included in their list of 189 candidates, with the remaining events included either as previously reported or due to the unspecified workflow. None of the three AS events which were undetected by BMEA were from the published list of 189, and were included in the list of confirmed events as either a mutually exclusive exon (3597388 - *TPM1*), as a previously reported event (3326714 - *CD44*) or from the unspecified workflow (3422189 - *LGR5*).

However, Table 5.4 also revealed 23 false positives amongst the results, and due to the unclear methodology provided by Gardina et al, direct comparison of Type 1 errors is difficult. Only a small proportion of the top 189 AS events were tested via PCR gel, and if considering the general workflow as the true selection methodology, this yielded a far higher level of false positives than BMEA.

During inspection of these results, it also became clear that altering the BMEA criteria for detection of an AS event to a 95% CPI excluding the interval [-0.5, 0.5] would eliminate 16 of the 23 false positives, and would retain 14 of the 20 detected and confirmed AS events, shifting true to false positives to 14:7 as opposed to the much higher ratio observed without this step. However, this was simply an initial observation, to be explored more thoroughly during analysis of the Tissue Mixture dataset of Section 5.4.

Table 5.3 – *AS events considered as confirmed in the published Gardina Analysis. All were considered as significant under BMEA with the exceptions of 3326714 (CD44), 3422189 (LGR5) and 3597388 (TPM1), as indicated by asterisks next to the 95% CPI. Posterior medians are shown as point estimates for $\Delta \log \phi_j$. Both CPI and estimates of $\Delta \log \phi_j$ are given using the $\log_2$ scale. The reason for inclusion by Gardina et al is indicated as being in the Top 189, previously reported (PR), mutually exclusive exons (MEE) or from the unspecified general workflow (GR).*

Gene	**Probeset**	**Source**	**Gel Image**	$\widehat{\mathbf{\Delta \log \phi_j}}$	**95% CPI**	**B**
ACTN1	3569830	Top189	✓	1.45	[1.13,1.78]	8.70
	3569827	MEE	✓	-0.85	[-1.18,-0.52]	-8.70
ATP2B4	2375766	Top189	✓	-1.05	[-1.46,-0.65]	-8.70
CALD1	3025632	Top189	✓	-1.27	[-1.72,-0.78]	-8.70
CD44	3326711	PR	✓	0.70	[0.14,1.29]	5.14
	3326712	PR	✓	0.71	[0.22,1.26]	5.86
	3326714	PR	✓	0.29	[-0.13,0.74]*	2.18
COL6A3	2605386	Top189	✓	1.03	[0.72,1.34]	8.70
	2605390	Top189	-	1.21	[0.92,1.51]	8.70
	2605391	Top189	-	1.09	[0.72,1.44]	8.70
CTTN	3338589	GW	✓	0.55	[0.16,0.95]	5.75
FN1	2598321	Top189	✓	-0.59	[-0.83,-0.35]	-8.70
GK	3972987	GW	-	1.16	[0.62,1.73]	8.70
ITGB4	3735208	PR	✓	-1.61	[-2.23,-1.05]	-8.70
LGR5	3422189	GW	-	-0.68	[-1.93,0.83]*	-1.59
MAST2	2334499	GW	-	-0.97	[-1.34,-0.62]	-8.70
RAC1	2989068	PR	✓	0.76	[0.18,1.38]	5.14
SLC3A2	3333718	GW	✓	-0.49	[-0.99,-0.01]	-3.74
TPM1	3597384	Top189	✓	1.46	[0.91,2.00]	8.70
	3597388	MEE	✓	-0.37	[-0.85,0.12]*	-2.70
VCL	3252129	Top189	✓	-0.96	[-1.33,-0.58]	-8.70
ZAK	2516001	Top189	-	1.18	[0.57,1.79]	8.70
	2516011	Top189	-	2.17	[1.58,2.79]	8.70

Table 5.4 – *AS events considered as not confirmed in the published Gardina Analysis, but which were considered as significant under BMEA. Posterior medians are shown as point estimates for $\Delta \log \phi_j$. Both CPI and estimates of $\Delta \log \phi_j$ are given using the* $\log_2$ *scale. All probesets were included via the unspecified general workflow. Gardina et al ascribe the results for COL11A1, PSD and PTK7 to fold-change at the gene level.*

Gene	**Probeset**	$\widehat{\boldsymbol{\Delta}\log\phi_{\mathbf{j}}}$	**95% CPI**	**B**
CABIN1	3939720	0.57	[0.16,1.00]	6.30
CDH11	3694667	-0.87	[-1.49,-0.20]	-5.40
COL11A1	2425849	2.46	[1.88,3.05]	8.70
	2425850	2.86	[2.28,3.47]	8.70
	2425837	1.47	[0.80,2.16]	8.70
ENOSF1	3795922	0.58	[0.02,1.15]	3.90
FAM44B	2887648	1.09	[0.21,2.00]	4.68
FIP1L1	2727236	0.64	[0.07,1.24]	4.40
LRP8	2413242	-0.88	[-1.64,-0.18]	-4.93
	2413218	0.93	[0.25,1.56]	5.75
MUC4	2712246	0.72	[0.27,1.19]	7.60
NCAM1	3349365	-0.78	[-1.53,-0.05]	-4.01
	3349371	0.88	[0.26,1.49]	6.13
NME1	3726945	-1.25	[-1.78,-0.69]	-8.70
PFKP	3232406	-1.02	[-1.49,-0.53]	-8.70
PSD	3304306	-0.95	[-1.53,-0.38]	-8.70
PTK7	2907705	0.85	[0.22,1.46]	5.86
PTK9L	2676011	-0.74	[-1.22,-0.25]	-6.75
STK25	2607278	-0.66	[-1.15,-0.16]	-5.33
TENS1	3049621	0.88	[0.37,1.44]	8.70
TNS	2599194	0.75	[0.34,1.17]	7.60
	2599195	1.22	[0.79,1.66]	8.70
	2599199	0.86	[0.55,1.18]	8.70

5.3.3 Analysis Using FIRMA

Methods

The complete Gardina dataset was also re-analysed using FIRMA, setting the same CDF as for BMEA. As this approach effectively relies on all genes on the array, the entire set of 22,035 genes was fitted. Arrays were quantile normalised and background corrected using RMA, before obtaining FIRMA scores for each exon-level probeset using the implementation in *aroma.affymetrix*.

To assess alternate splicing, differences in FIRMA scores were calculated between tumour and normal samples *within each donor*. Donor-based weights were then assigned using the function *arrayWeights* from the package *limma* on the sets of differences in FIRMA scores, and weighted models were fit with moderated $\tilde{t}$-statistics. An initial FDR estimate was obtained for each probeset using the p-values across the entire set of 326,983 exon-level probesets.

Results

Using the conventional FDR cutoff of $\alpha = 0.05$, 6 confirmed AS events were detected (Table 5.3) with 3 false positives also being retained (Table 5.4). Extending the FDR to 10% gave an additional 6 confirmed AS events at the cost of 3 additional false positives, maintaining the observed FDR at 33.3%.

Table 5.5 – *FIRMA results for confirmed AS events in the Gardina Dataset. The logFC column indicates the estimated change in the FIRMA score.*

Gene	**EnsemblID**	**Probeset**	**logFC**	**$\tilde{t}$-statistic**	**p-value**	**FDR**
ACTN1	ENSG00000072110	3569830	2.01	9.88	7.78e-08	0.005
ZAK	ENSG00000091436	2516011	1.45	8.64	4.18e-07	0.016
COL6A3	ENSG00000163359	2605386	1.89	8.61	4.37e-07	0.016
CALD1	ENSG00000122786	3025632	-1.18	-7.50	2.28e-06	0.025
COL6A3	ENSG00000163359	2605390	1.89	7.34	2.97e-06	0.026
LGR5	ENSG00000139292	3422189	-1.11	-6.95	5.49e-06	0.029
VCL	ENSG00000035403	3252129	-1.30	-5.92	3.18e-05	0.051
MAST2	ENSG00000086015	2334499	-1.01	-5.79	4.01e-05	0.053
ATP2B4	ENSG00000058668	2375766	-2.08	-5.50	6.86e-05	0.059
GK	ENSG00000198814	3972987	0.85	5.03	1.65e-04	0.078
FN1	ENSG00000115414	2598321	-1.30	-4.84	2.38e-04	0.086
ACTN1	ENSG00000072110	3569827	-1.27	-4.53	4.34e-04	0.099
COL6A3	ENSG00000163359	2605391	1.22	4.41	5.45e-04	0.106
RAC1	ENSG00000136238	2989068	1.43	3.85	1.66e-03	0.130
TPM1	ENSG00000140416	3597388	-0.62	-3.13	7.06e-03	0.191
ZAK	ENSG00000091436	2516001	0.51	2.95	0.010	0.210
TPM1	ENSG00000140416	3597384	0.65	2.39	0.031	0.293
CTTN	ENSG00000085733	3338589	0.37	2.31	0.036	0.307
ITGB4	ENSG00000132470	3735208	-0.69	-1.85	0.085	0.407
SLC3A2	ENSG00000168003	3333718	-0.31	-1.33	0.206	0.561
CD44	ENSG00000026508	3326711	0.27	1.20	0.251	0.604
CD44	ENSG00000026508	3326712	0.41	1.12	0.280	0.630
CD44	ENSG00000026508	3326714	0.26	0.92	0.373	0.702

Table 5.6 – *FIRMA results for unconfirmed AS events in the Gardina Dataset. The logFC column indicates the estimated change in the FIRMA score. Probesets with a raw p-value* > 0.01 *are not shown.*

Gene	**EnsemblID**	**Probeset**	**logFC**	t**-statistic**	p**-value**	**FDR**
COL11A1	ENSG00000060718	2425850	1.89	10.06	6.20e-08	0.005
CST2	ENSG00000170369	3901398	1.21	7.22	3.56e-06	0.026
COL11A1	ENSG00000060718	2425849	1.41	7.17	3.86e-06	0.027
PTK7	ENSG00000112655	2907705	0.69	5.61	5.57e-05	0.057
BOD1	ENSG00000145919	2887648	0.86	5.60	5.75e-05	0.057
PSD	ENSG00000059915	3304306	-0.94	-5.49	6.97e-05	0.060
LRP8	ENSG00000157193	2413218	0.62	4.37	5.87e-04	0.108
NCAM1	ENSG00000149294	3349371	0.62	4.33	6.32e-04	0.109
TNS1	ENSG00000079308	2599194	1.10	4.18	8.55e-04	0.115
TWF2	ENSG00000173366	2676011	-1.01	-3.93	1.41e-03	0.127
NME2	ENSG00000011052	3726958	0.69	3.87	1.58e-03	0.129
PTK7	ENSG00000112655	2907698	-0.70	-3.70	2.26e-03	0.142
TNS1	ENSG00000079308	2599195	1.26	3.59	2.78e-03	0.148
TNS1	ENSG00000079308	2599200	1.00	3.59	2.83e-03	0.149
CDH11	ENSG00000140937	3694667	-0.66	-3.53	3.19e-03	0.155
TNS1	ENSG00000079308	2599199	1.06	3.52	3.22e-03	0.155
COL11A1	ENSG00000060718	2425837	0.66	3.43	3.85e-03	0.164

5.3.4 Analysis Using MMBGX

The initial MMBGX publication (Turro et al., 2010) also compared the results of their approach to those presented by Gardina *et al*, giving a further benchmark against which BMEA can be compared. Under MMBGX, transcripts are organised into clusters based around genes, or related genes, and the proportion of each transcript within the cluster is estimated for the given treatment group. Although being a transcript-based model as opposed to the exon-based approaches of BMEA, FIRMA and Gardina *et al*, each exon-level probeset was verified based on the estimated fold-change in the relevant transcripts.

Careful inspection of their published results revealed four points of note:

1. For *ITGB4* the transcripts were labelled based on exon 33 instead of exon 35, and the results marked as contradictory. When checked against the correct exon the results from MMBGX in fact confirmed the PCR gel;

2. For *TPM1* all relevant transcripts were presented as including Exon7 but not Exon8, when these transcripts all included Exon8 but not Exon7. The cumulative fold-change predicted by their model again accurately reflected the gel instead of being inconclusive as reported;

3. Although the AS events for *COL6A3* satisfied their selection criteria, it was noted that the direction of the predicted AS event was incorrect, and this was subsequently deemed to be an incorrect call.

4. No indication was given as to the exclusion of patient 3 for direct comparison with the Gardina analysis, with an implication in the text that the full dataset was fitted.

In order to assess which of the AS events would be detected based on this method, the simplest approach in keeping with their methodology was decided as 1) a minimum of one transcript per probeset being identified with their custom statistic $\max(p_t, 1 - p_t) > 0.9$, and, 2) which was not filtered out for low expression. A summary of confirmed AS events which satisfied this criteria is given in Table 5.7 along with a summary of all other methods.

A list of the gels not confirming AS events from the original Gardina paper was also included as supplementary data, with results for MMBGX added for these probesets. A further 3 false positives were noted for *FAM44B*, *GBA* and *CDH11*. Any AS events declared as "Incompatible" on this list were taken as true negatives.

5.3.5 Comparison of All Methods Used for the Gardina Dataset

As is clear from Figure 5.2 and Table 5.7, BMEA was able to detect the highest number of confirmed AS events, however the observed FDR was unacceptably high at >50% using the initial approach. The extension of the 95% CPI method to exclude the arbitrarily chosen interval [-0.5,0.5] (BMEA + 0.5) immediately reduced the observed FDR to 33%, whilst still being able to detect a greater number of confirmed AS events than other approaches. FIRMA using an expected FDR of 10% showed a true FDR of 33%, equivalent to BMEA + 0.5, but detecting two fewer confirmed events. For both MMBGX and the Gardina workflow, the observed FDR was ≥40%, and with lower numbers of detected AS events.

Direct comparison to the Gardina methodology was compounded by a lack of transparency, where only a small number of the top-ranked 189 exons were tested via PCR gels. AS events for genes such as *CTTN*, *SLC3A2*, *GK*, *MAST2*, *LGR5*, *COL11A1*, and numerous others were additionally tested for reasons not directly specified. However, it was noted that under both BMEA and FIRMA, *COL11A1* returned strongly significant results, and it is possible that similar observations by the authors also led to inclusion of this gene. The strong AS signals were explained away by Gardina *et al* as due to fold-change at the gene level ($\log_2$FC ≈ 1.5), however without a gel image, the truth behind this claim is difficult to ascertain. Indeed the approach used throughout this publication (i.e. visual inspection and densitometric analysis of a gel image) falls well short of the standard for quantification by established methods such as qPCR, leaving many questions unable to be answered satisfactorily.

Whilst it appears that both BMEA and FIRMA outperform MMBGX, contrary to the results of Turro *et al*, it should be noted that the matched pairs approach was not used in their assessment of FIRMA and the increased power of the paired approach may explain this disparity. In addition, the approach used by Turro *et al* for estimation of the FDR followed Storey's methodology (Storey, 2002), as opposed to the Benjamini-Hochberg method used here. Both BMEA and FIRMA assess changes in exon-level probesets, using the CDF format. The method of assessment under MMBGX is vastly different and indeed has many appealing qualities, such as the direct estimation of transcript abundances, and association probes to clusters of related transcripts. However, given the relative ease of direct comparison between BMEA and FIRMA, and the out-performance of FIRMA over both MMBGX and SI-based analysis, BMEA and FIRMA were the only two methods compared in subsequent analyses.

The discrepancy between the observed error rate under BMEA using the 95% CPI selection method, and the expected error rate was also surprising. It was assumed to be largely due to excessive noise inherent to array data, and the small number of data points which are used to make inferences at the exon level (≤ 4 per sample). This assumption, whilst unsatisfying, is strongly supported by the observed FDR under FIRMA also being $\approx$33% for both expected FDR thresholds of $\alpha = 0.05$ and 0.1, and which represents an inflation of up to 6-fold in the expected error rate.

Table 5.7 – *Summary of all AS events confirmed by PCR Gel in Gardina et al. Probesets presented as confirmed, but for which no gel was provided are marked with an asterisk. Detection by each method is indicated by a plus or minus, with the criteria for detection by BMEA being that the 95% CPI excluded zero. The criteria for inclusion as detected by MMBGX was as described in Section 5.3.4. AS events not reported by the MMBGX authors are indicated with NA. FIRMA is presented using FDR thresholds of both 0.05 and 0.10.*

				FIRMA		**Gardina**
Gene	**Probeset**	**BMEA**	**MMBGX**	**FDR05**	**FDR10**	**Top189**
ACTN1	3569830	+	+	+	+	+
	3569827	+	NA	-	+	-
ATP2B4	2375766	+	+	-	+	+
CALD1	3025632	+	+	+	+	+
CD44	3326711	+	-	-	-	-
	3326712	+	-	-	-	-
	3326714	-	-	-	-	-
COL6A3	2605386	+	-	+	+	+
	2605390*	+	-	+	+	+
	2605391*	+	-	-	-	+
CTTN	3338589	+	-	-	-	-
FN1	2598321	+	-	-	+	+
GK	3972987*	+	NA	-	+	-
ITGB4	3735208	+	+	-	-	-
LGR5	3422185	NA	-	NA	NA	-
	3422189	-	-	+	+	-
MAST2	2334499*	+	NA	-	+	-
RAC1	2989068	+	+	-	-	-
SLC3A2	3333718	+	-	-	-	-
TPM1	3597384	+	+	-	-	+
	3597388	-	-	-	-	-
VCL	3252129	+	-	-	+	+
ZAK	2516001*	+	NA	-	-	-
	2516011*	+	NA	+	+	-
Total Detected		20	6	6	12	9

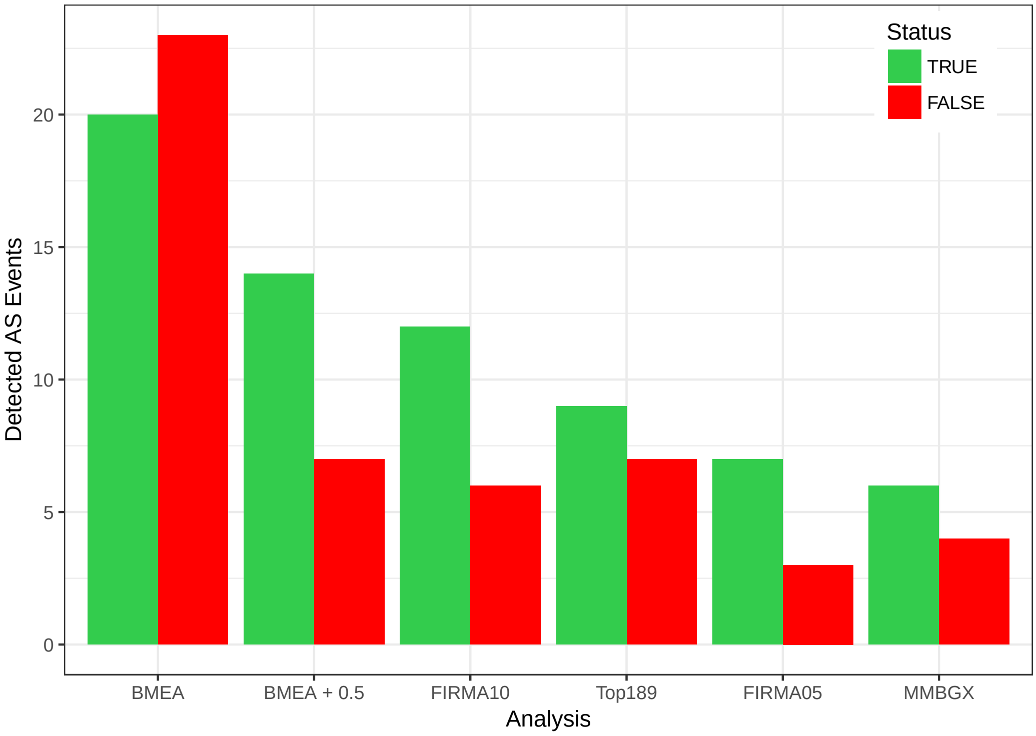


Figure 5.2 – *Comparison of detected AS events for the Gardina Dataset using all approaches. BMEA is shown using the 95% CPI which excludes zero, and using the 95% CPI which excludes the interval [-0.5, 0.5] (BMEA + 0.5), whilst FIRMA is shown for two FDR thresholds (0.05 and 0.10).*

5.4 Tissue Mixture Dataset

5.4.1 Initial Model Fitting

The Tissue mixture dataset, introduced in Section 5.1 was then investigated using BMEA. The entire set of 33 arrays was quantile normalised, then the BMEA model was applied to the subset of 15 arrays consisting of the 100% Brain ($n = 3$), 50:50 Brain/Heart ($n_a = n_b = n_c = 3$) and the 100% Heart ($n = 3$) tissues. Notably, the 50:50 Mixture levels contained 3 sets of 3 technical replicates of alternate mixture preparations which were able to be used as negative controls with no expected splice variation or differentially expressed genes. All genes were fitted using the custom CDF described in Section 5.2.2, with 5808 genes excluded under the DABG criteria of $Z_g < 4.265$ as described in Section 4.3.1. The gene *TITIN* (*ENSG00000155657*) was also omitted for convenience as it contains >350 exons. MCMC parameters were set with the default values as per Section 5.3.2. This reduced set of 15 arrays was also fitted using RMA background correction and conventional PLM/FIRMA approaches. Condition-specific expression estimates, log fold-change estimates and FIRMA scores were obtained using array-level weights.

5.4.2 Negative Control Analysis

PLM/FIRMA Results

In order to assess BMEA in comparisons where no splice variation or differential expression is present, the three sets of technical replicates which contained the 50:50 tissue mixture were compared to each other as 3 sets of two-way comparisons (a vs b; a vs c; b vs c). The same groups were also compared to each other using PLM/FIRMA, using moderated $\tilde{t}$-tests and FDR-adjusted p-values. Using a threshold of $\alpha = 0.05$, one gene was considered as DE under PLM modelling, and no exon-level probesets were considered as a candidate AS event. Correlations between average expression levels across each set of comparisons were $0.992 < \rho < 0.994$. As all arrays were replicates containing the same mixture levels, these results effectively provide a benchmark under conditions when the null hypothesis can reasonably be expected to be true across all genes and comparisons.

BMEA at the Gene Level

Under BMEA, correlations at the gene level were slightly lower than for PLM ($0.989 < \rho < 0.990$), but still at the high end of the possible range (Figure 5.3). Using a 95% CPI which excludes zero as the sole criteria for a DE gene, 11 genes were considered as DE across the three comparisons. Setting the DE criteria as a CPI-LB $> \kappa$ for $\kappa = \log(1.1)$ reduced this to a single gene, which was an equivalent result to analysis under PLM, albeit with each method producing a different false positive. The 95% CPI-LB for this gene was 0.189, which is approximately log(1.2). This supports the idea that using a 95% CPI-LB with $\kappa > 0$, would be expected to limit the number of false positives within a set of putatively DE genes.

Comparison between BMEA and PLM-derived expression estimates gave correlations of $0.929 < \rho < 0.933$ which was lower than the correlations within each method, but still showed a high level of consistency (Figure 5.4).

BMEA at the Exon Level

Using the most inclusive criterion, an exon was initially considered as a candidate AS event if the 95% CPI simply excluded the value $\kappa = 0$. Despite detection of no potential AS events under FIRMA, this produced more than 4000 candidate AS exons in each comparison, which were all considered to be false positives as all samples were replicates. In order to determine the underlying source of this error, these erroneous AS events were investigated using the average expression level $\bar{\mu}_h$, exon-level Z_j scores and the value for κ which would reduce this number.

Each 95% CPI-LB was plotted using average expression levels broken up into quantiles, and taking the posterior mean as a point estimate for μ_h (Figure 5.5B) and grouped by Z_j quantiles in Figure 5.5A. No clear relationship between the overall expression level and the CPI-LB was found, whilst it was very clear that exons with lower Z_j scores consistently gave lower bounds which were further from zero. As these were all erroneous and should contain zero, it is clear that probesets with higher Z_j scores will produce the fewest errors. These will represent probesets for which true signal strongly dominates background across a significant number of PM probes, making this a relatively intuitive observation.

A simple selection method of restricting consideration to probesets with a Z_j score above the median ($Z_j > 14.3$) would halve the original set of 242,666 probesets. Using the 95% CPI-LB with $\kappa = \log(1.5)$ on this reduced set would have yielded between 4 and 16 erroneous candidate AS events from each of the three comparisons, which is an error rate $< 0.01\%$ of the total considered. Increasing the value κ consistently reduced these errors, however in a setting with true AS events, this would increasingly reduce the power of detection by removing an increasing number of probesets from consideration. To completely eliminate errors from these particular comparisons a value of $\kappa = \log(6)$ would need to be chosen. However, adjusting the threshold for consideration to $Z_j > 15.3$ would change the minimum value for an error-free analysis to $\kappa = \log(2.2)$ which would considerably increase the power of true AS detection than for $Z_j > 14.3$, although this would be on a reduced set of exons. Thus choosing a suitable combination of thresholds for Z_j and κ provides numerous options for restricting errors and these can be chosen as appropriate. A further exploration of choosing suitable values in subsequent sections.

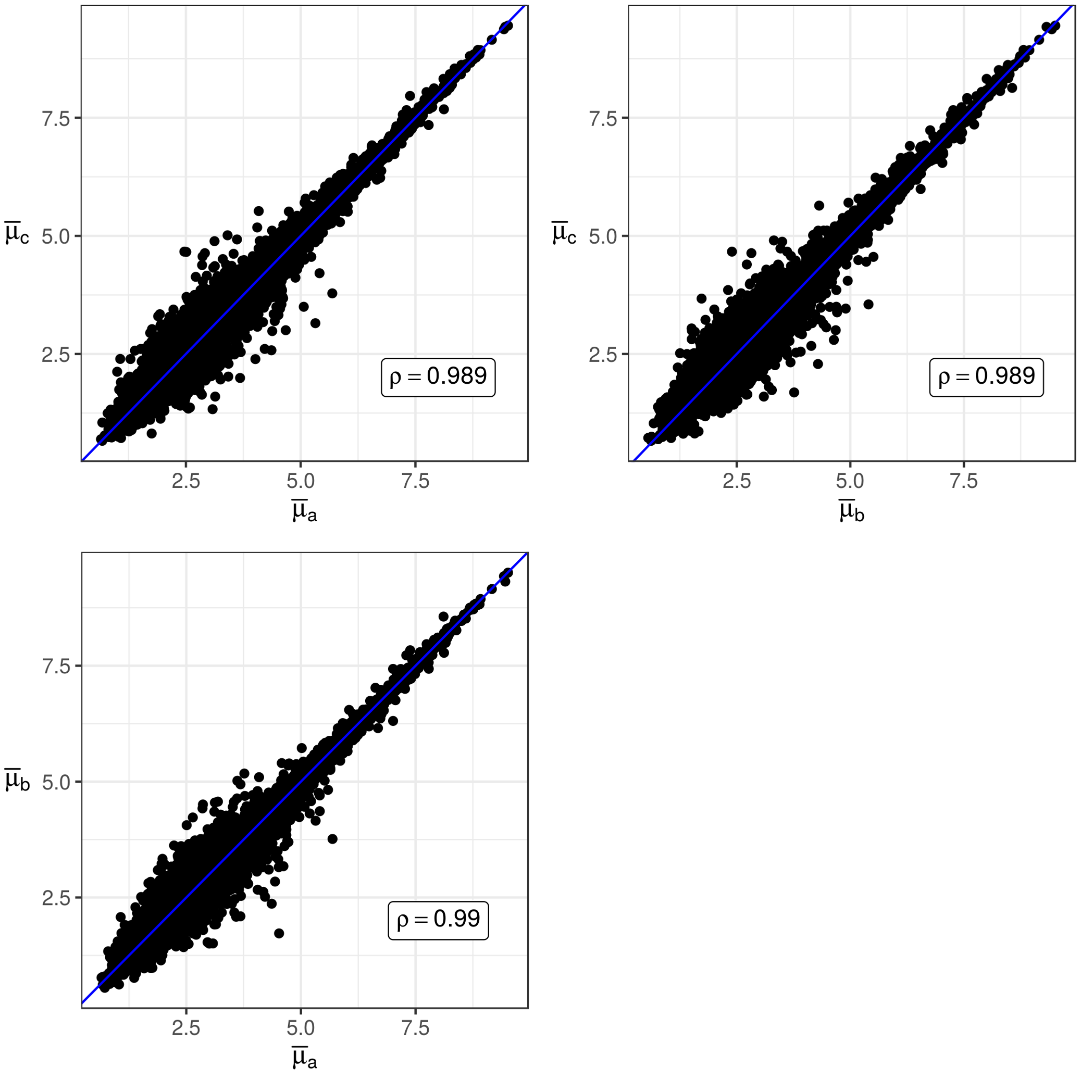


Figure 5.3 – *Comparison of posterior means for μ_h between each set of triplicate 50:50 tissue mixtures. No genes are truly DE.*

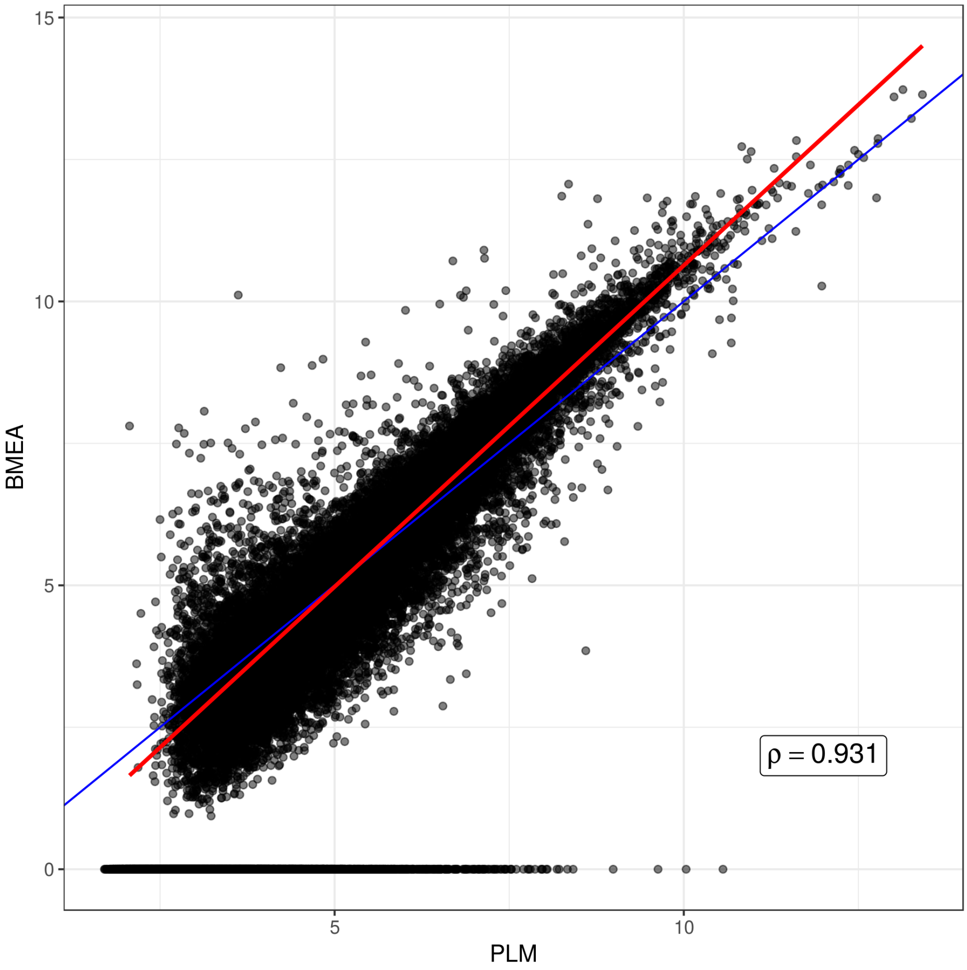


Figure 5.4 – *Average expression estimates for genes in the 50:50 Tissue Mixture samples. Values are shown for the set of replicates from TisMix_5a. RMA/PLM derived values are shown on the x-axis, whilst posterior means for μ_h as obtained under BMEA are shown on the y-axis. BMEA derived values have been converted to the $\log_2$ scale. The line $y = x$ is shown in blue, whilst the regression line for the model $y_i = \beta_0 + \beta_1 x$ is shown in red. The correlation between the two sets of values is also given.*

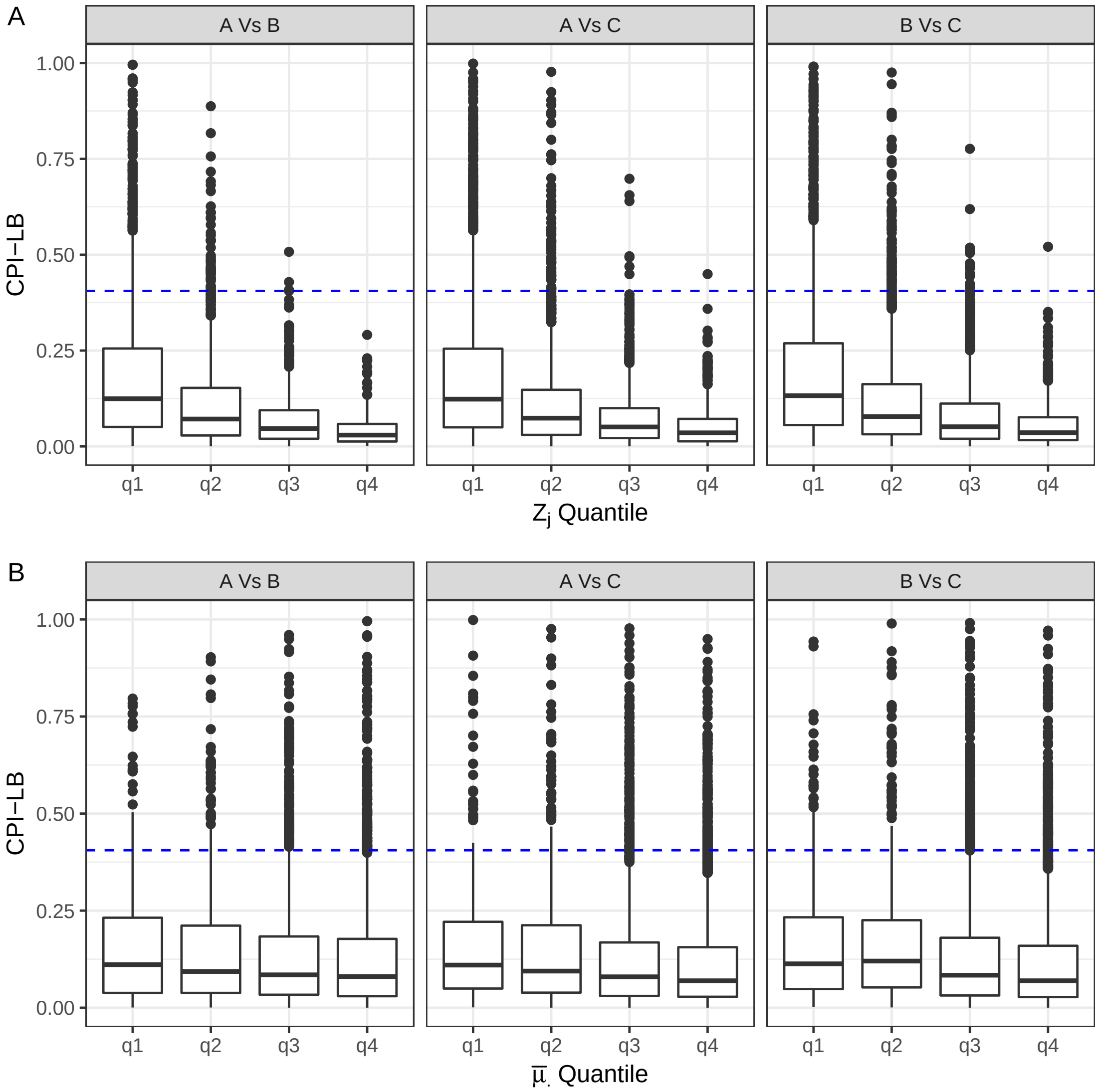


Figure 5.5 – *Distributions of the CPI-LB for all exons with a 95% CPI which excludes zero. Values are separated by comparison within each A) quantile of Z_j scores, and B) quantile of average expression levels. y-axes are truncated at the value 1 to reduce outlier points. The value* $\log(1.5)$ *is shown as the horizontal blue line.*

5.4.3 Differential Gene Expression

A set of differentially expressed genes was then defined from each of the two methodologies using the 100% Heart and 100% Brain samples. For the RMA/PLM-fitted data, significant genes were defined as those with a estimated $\log_2$ fold-change beyond ±1, and an adjusted p-value < 0.05. As this comparison involves two highly differentiated tissues, a large number of differentially expressed genes is highly likely, and would be expected to represent true biological signal. Thus, adjusted p-values were obtained using Holm's method on the raw p-values obtained from the moderated t-statistics. The estimated FDR using this cutoff value was 1.36×10^{-5}, corresponding to a raw p-value of 1.60×10^{-6}, and giving a list of 3375 differentially expressed genes under the RMA/PLM approach. For BMEA-fitted data, genes were considered as DE using the 95% CPI-LB with $\kappa = \log(1.3)$, as this gave a set of 3118 DE genes and was of comparable size to the set defined under RMA/PLM. This was also more conservative than the value chosen in Section 5.4.2 which resulted in similar error rates between PLM and BMEA.

A high degree of concordance between the two gene sets was observed with 2583 genes being identified in both lists (Figure 5.6). Genes are shown ranked by the BMEA B-statistic in Figure 5.7 with ties broken by logFC, and with the moderated $\tilde{t}$-statistics shown on the y-axis. A broad similarity of trend was observed between the two models, confirming that BMEA produced results which are comparable to this widely accepted approach. Of the 535 genes defined as significantly DE under BMEA but not under RMA/PLM, 521 would have been included using an FDR-adjusted p-value without any filter of logFC. The remaining 14 genes were either single-exon genes, or were fitted as single-exon genes after removal of undetectable exon-level probesets, and this largely accounted for this remaining discrepancy.

Of the 792 genes considered as DE under RMA/PLM but not under BMEA, 489 would have been included as DE under BMEA if filtering based on a 95% CPI which excluded zero, with the remainder not being considered as DE under BMEA. For many of these genes, the moderated variance as calculated for the moderated $\tilde{t}$-statistics, had received considerable shrinkage and as BMEA contains no analogous procedure, this we considered to be a key contributing factor to this discrepancy.

Point estimates of fold-change (Figure 5.8) were also broadly consistent between the two approaches ($\rho = 0.932$), however the increased dynamic range of BMEA expression estimates (Figure 5.4) generally impacted the range of logFC estimates in a similar fashion. It was also evident that some genes which showed logFC values between ±1 under RMA/PLM were detected as differentially expressed under BMEA, whilst only one gene considered as DE under RMA/PLM was undetectable using Z-score filtering. As the returned values were strongly concordant with those obtained under RMA/PLM no further verification across the differing mixture levels was undertaken.

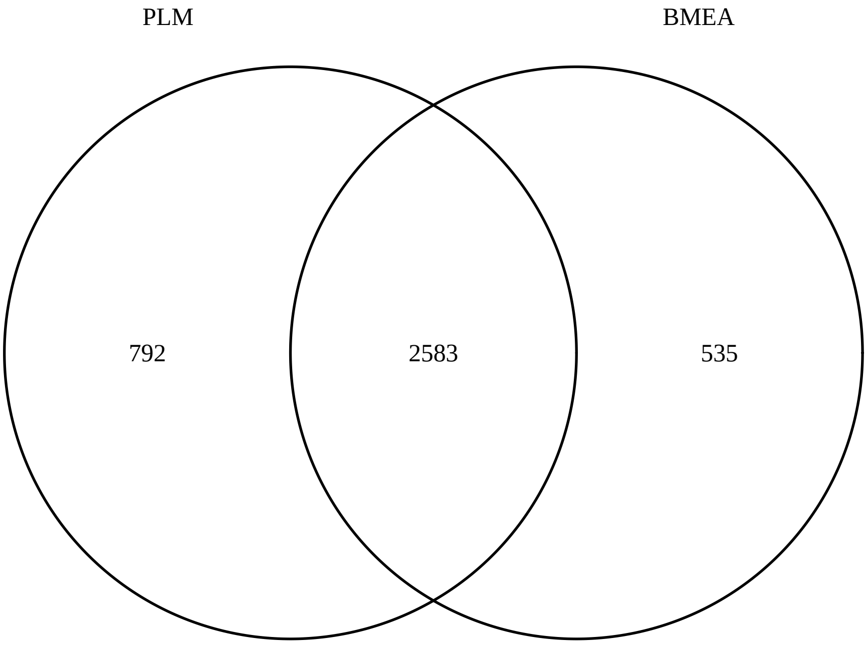


Figure 5.6 – *Comparison of genes defined as differentially expressed under RMA and BMEA based approaches.*

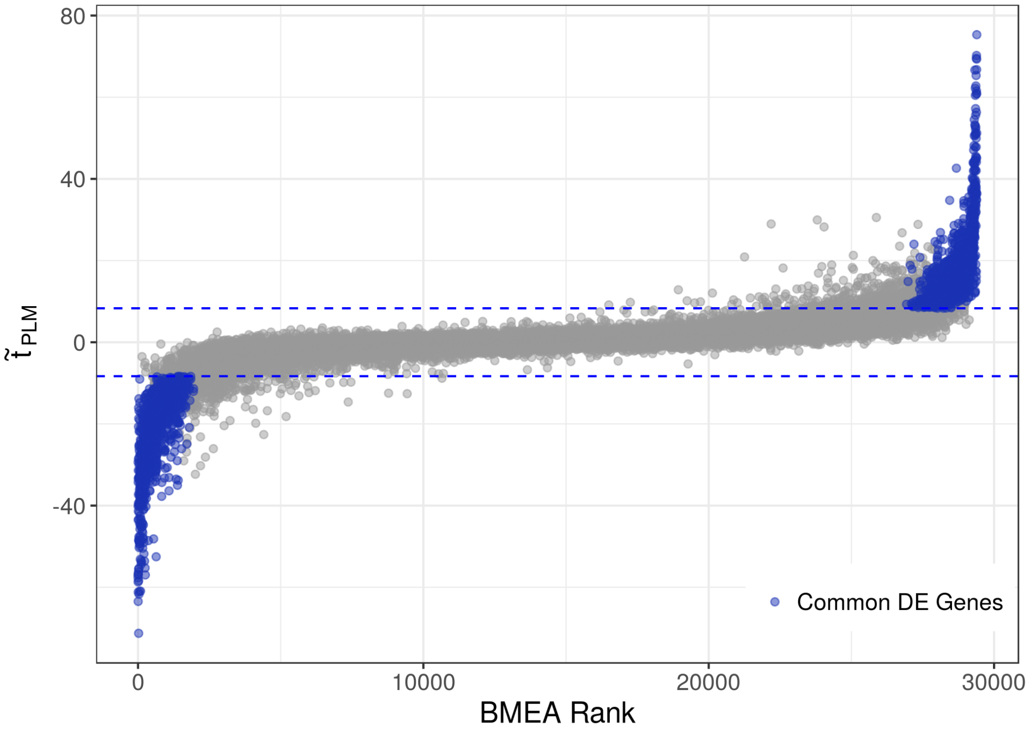


Figure 5.7 – *Moderated $\tilde{t}$-statistics obtained under PLM with genes ordered by the BMEA B-statistic. Ties in the B-statistic were broken by logFC. Genes considered DE under both approaches are shown in blue. The t-statistic cutoffs equivalent to an adjusted p-value of 0.05 under PLM, are shown as horizontal blue lines.*

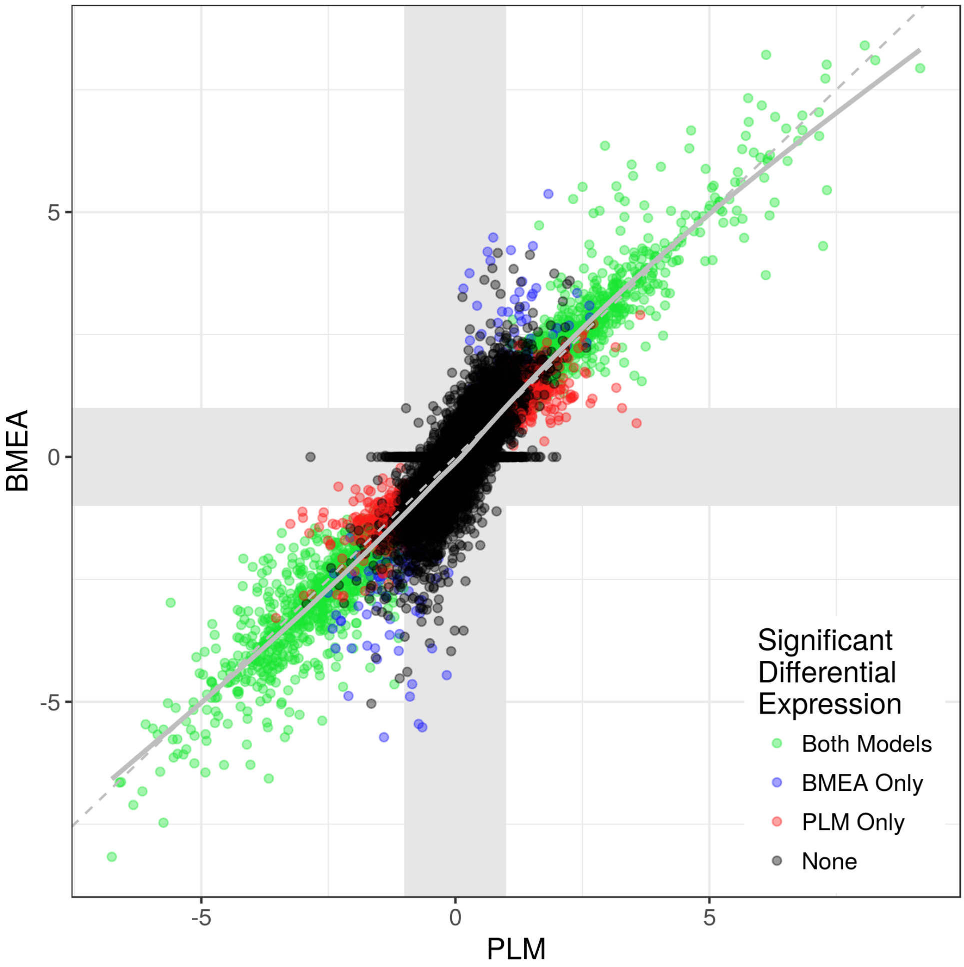


Figure 5.8 – *Comparison of point estimates for logFC under both models. Genes are coloured based on which model they were found to be differentially expressed in. The line $y = x$ is shown as the dotted grey line, whilst the smoothed line generated using generalized additive models (Wood & Augustin, 2002) is shown as the solid grey line. The grey bands represent the region $-1 < logFC < 1$ for each model and signify the region where genes will not be defined as DE, regardless of the level of purely statistical significance.*

5.4.4 Initial Selection of Candidate AS Exons

Whilst the performance of BMEA in the absence of AS events was explored in Section 5.4.2, the accuracy of the model needed to be tested in the presence of AS events. In order to assess the accuracy of the BMEA model for detection of tissue-specific exons, candidate probesets in this dataset need to have the expectation of tracking linearly with the tissue mixture proportion, and any potential effects due to changing gene abundances need to be minimised. To illustrate the importance of this, if an exon is included in brain transcripts only, but the gene itself is exactly twice as highly expressed in heart tissue, at the 50:50 mixture level, only 33.3% of the transcripts will contain this exon, whilst the 50% level would not be expected to be observed until a 2:1 mixture level. Thus for this dataset, the value ϕ_{hj} will not track the mixture levels in a linear fashion if a gene is differentially expressed. Candidate genes for detection of alternate splicing were thus restricted to those in the lowest quartile of genes when ranked by posterior means for logFC between the two 100% tissue mixtures. These candidate genes were further restricted to those in the upper quartile of expressed genes, as defined by the average of all three posterior means for μ_h in the three 50:50 tissue mixture replicates, in keeping with the findings of Section 4.4.2. A total of 1124 genes matched these criteria and were classed as good candidates for detection of alternate splicing, giving 17,756 eligible exon-level probesets in this test set.

Exon-level probesets from this set of candidate genes were then selected as initial candidates for AS detection if $|B_j| > 6$, as determined by the distribution of B-statistics (Figure 5.9). This gave 675 putative *brain-specific* exons and 599 putative *heart-specific* exons (1274 total) (Figure 5.10) across 469 genes. Whilst not following the CPI-LB or Z_j approaches of previously (Section 5.4.2), this will allow assessment of the performance of the 95% CPI by using a more inclusive initial selection method. The BMEA model was then applied for these 469 genes, across the subset of 27 arrays *excluding* the 100% Brain and 100% Heart samples. For this smaller analysis, all nine 50:50 tissue mixture samples were considered as simple replicates of the same mixture level. This gave posterior summaries for ϕ_{hj} across all mixture levels which were able to be independently used to support or reject the predictions from the comparison between pure Heart or Brain tissue samples.

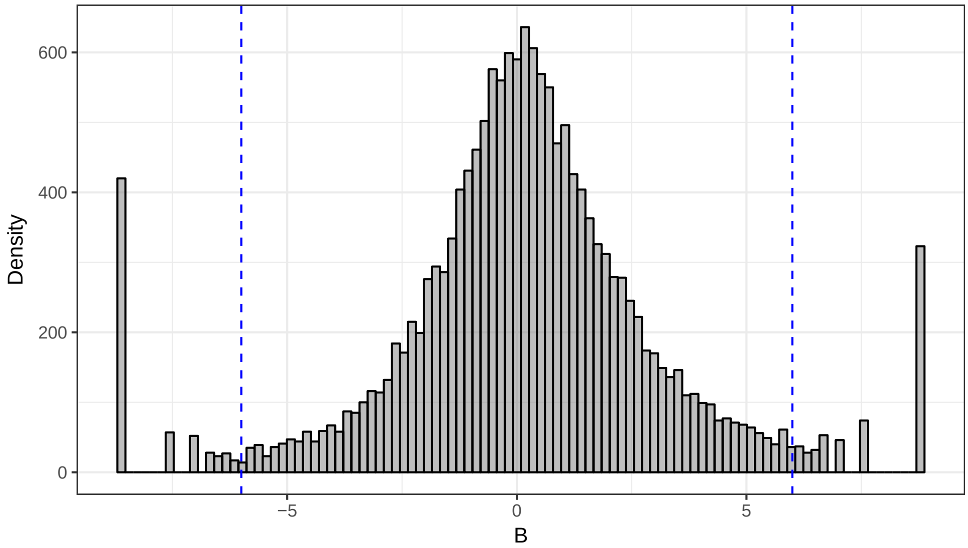


Figure 5.9 – *Histogram of exon-level B-statistics for the 1124 genes classed as the best candidates for detection of alternate splice events. The cutoff values used for potential significance (±6) are shown as dashed blue lines.*

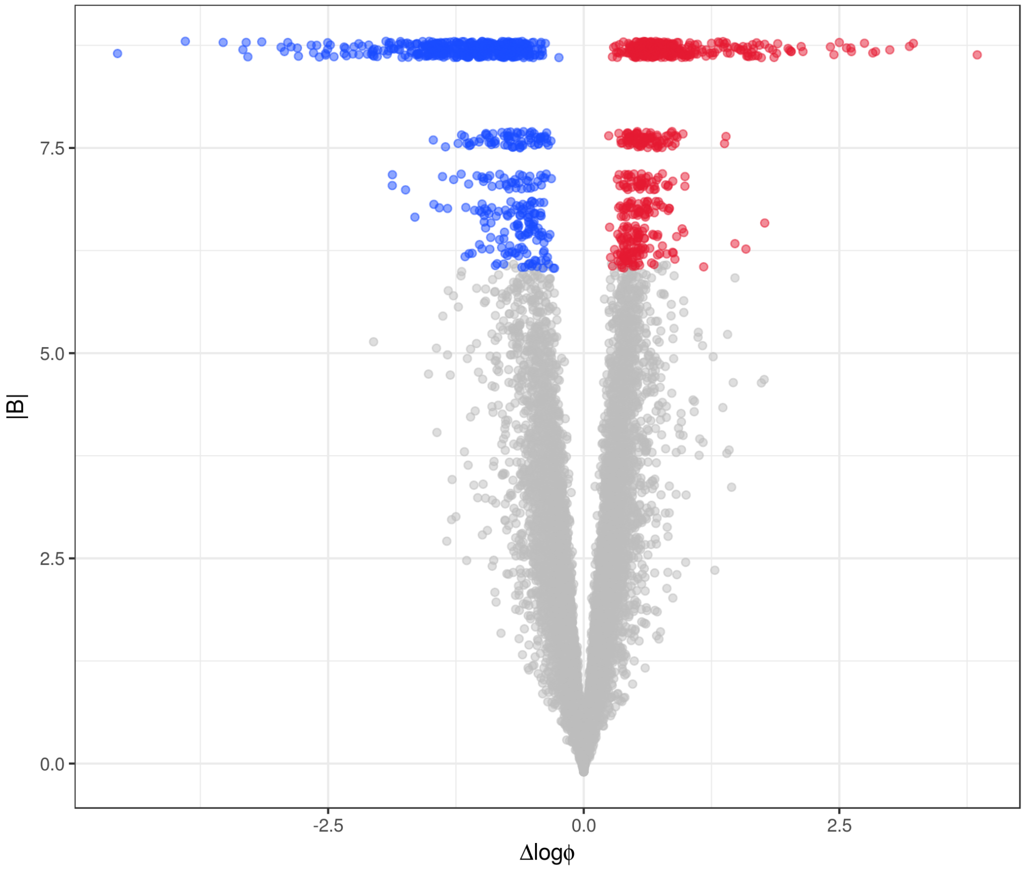


Figure 5.10 – *Volcano plot for exon-level B-statistics against changes in logϕ based on the 100% Heart Vs 100% Brain comparison. Points are only shown for the 1124 genes classed as candidates for detection of alternate splice events. Putative heart-specific exons are shown in red, whilst putative brain-specific exons are shown in blue. Jitter has been added on the vertical axis.*

5.4.5 Confirmation of AS Events Using Linear Regression

In order to provide confirmation of the candidate AS events, posterior median values for ϕ_{hj} were taken as point estimates, and a simple linear regression across mixture levels was fitted, using the proportion of heart tissue in the mixture as the continuous predictor. By omission of the 100% tissue samples when refitting, this effectively gave a prediction set of arrays and a confirmatory set which were independent. A candidate AS event was considered as confirmed if a non-zero slope was returned across mixture levels, with a zero-slope or contradictory direction indicating a false positive. Standard t-tests for $H_0 : \beta_1 = 0$ were conducted where β_1 represents the slope of the regression line through the mixture levels. One candidate exon-level probeset was considered as not-detectable using $Z_j < 1.645$ when fitting across the independent test dataset and this was subsequently removed from all downstream analysis.

Inspection of the regression lines (Figure 5.11) showed good support for all probesets with a raw p-value < 0.05, although some lines with relatively slight slopes were noted at the lower end of both tissues. As minimal differences existed between the probesets using raw and FDR-adjusted p-values, probesets were considered as having strong evidence for an AS event if adjusted p-values were < 0.05 (Holm's method), whilst for remaining probesets support was considered as moderate if a raw p-value was < 0.05. Additionally, the primary source of the difference between probesets with FDR-adjusted p-values > 0.05 but a raw p-value < 0.05, appeared to be variability between mixture levels. As this was considered likely to be technical rather than biological variability, using raw p-values for confirmation was justifiable. Under this method, only 65 (5.1%) of the 1273 candidate probesets were considered with strongly confirmed non-zero slopes. However, the less stringent criteria gave 829 (65%) of the tested probesets with moderate support for non-zero slopes (Table 5.8), indicating the likely presence of AS events in the corresponding exons, and suggesting that the initial selection criteria had yielded an observed FDR near 35%.

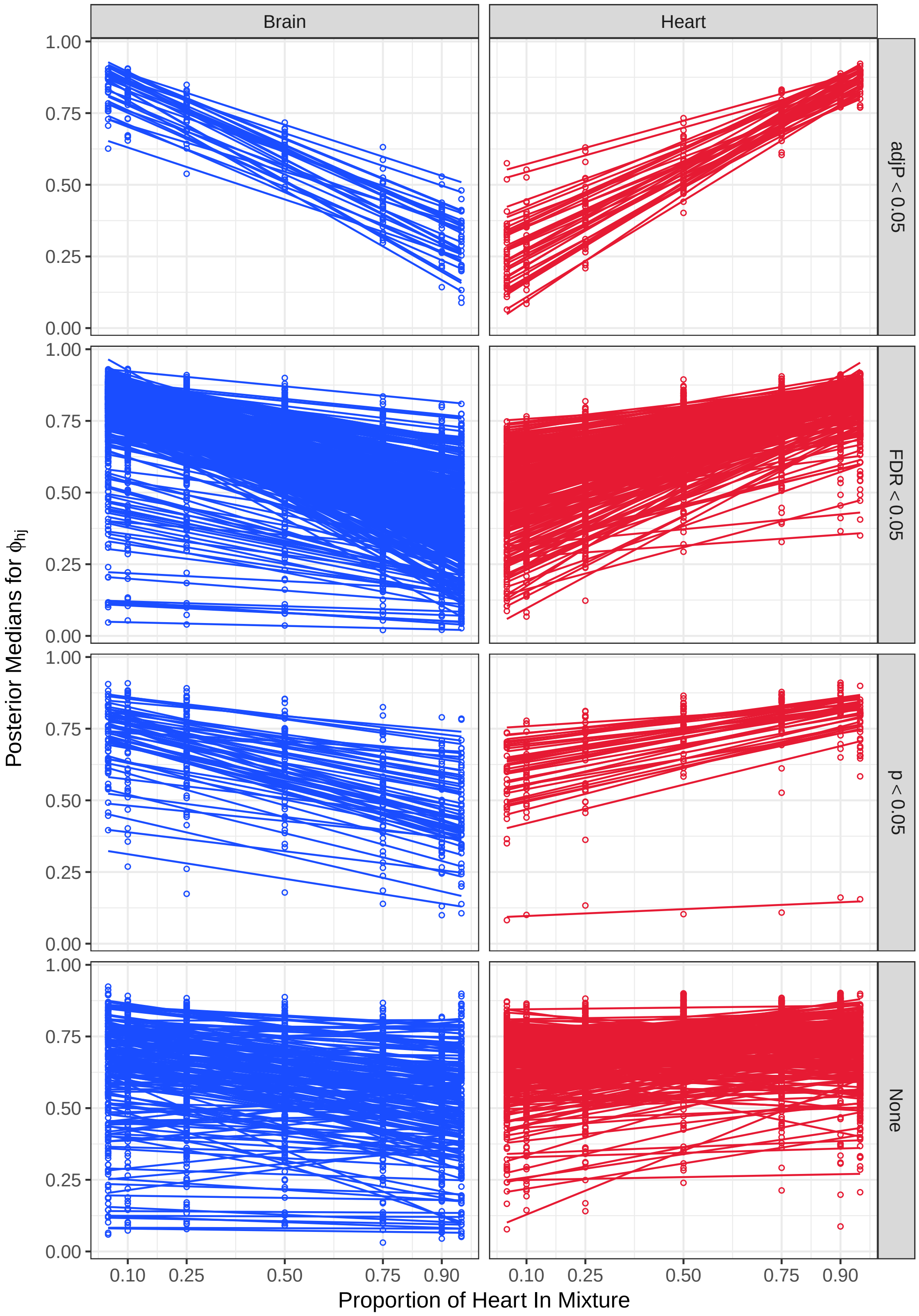


Figure 5.11 – *Posterior means for ϕ_{hj} plotted against the proportion of brain in the tissue mixture. Posterior means are shown as points with regression lines for each probeset being shown. Values are separated by those with an adjusted p-value < 0.05 using Holm's method, to an FDR of 0.05, using p-values with no adjustment, and with no supporting evidence to reject $H_0 : \beta_1 = 0$.*

Table 5.8 – *Probesets with confirmed non-zero slopes across the tissue mixtures. These are indicative of probesets for which the AS events predicted using BMEA were confirmed by linear regression analysis across all tissue mixture samples. Two possible criteria of adjusted p-values using Holm's method, or a raw p-value cutoff are shown. The observed FDR represents the proportion of unconfirmed AS events in the initial list of candidate exons, using raw* $p < 0.05$ *as supportive of an AS event.*

		Number Confirmed		
Tissue	**Tested Exons**	$\mathbf{p_{adj} < 0.05}$	$\mathbf{p_{raw} < 0.05}$	**Observed FDR**
Brain	675	29	501	0.26
Heart	598	36	328	0.45
Overall	1273	65	829	0.35

Development Of A Filtering Approach

In order to determine any contributing factors to the weaker than expected performance of BMEA, a filtering approach was sought to increase accurate detection of true AS events. Using a raw $p < 0.05$ for rejection of $H_0 : \beta_1 = 0$ as the sole criteria for a confirmed AS event, four data characteristics were assessed using logistic regression. Under this model, any impact on the probability of a confirmed AS event was investigated for: 1) the number of probes in a probeset; 2) 95% CPI-LB; 3) the Z_j-score; and, 4) the absolute value of the B_j-statistic using the model:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j}\,. \tag{5.1}$$

where

- y_j represents $\text{logit}(\pi_j)$, where π_j is the probability of a confirmed AS event (i.e. rejection of H_0 for exon j using $p < 0.05$)
- x_{1j} is the number of probes in probeset j
- x_{2j} is the lower bound for the 95% CPI for $\Delta \log \phi_j$, i.e. CPI-LB. This was taken as CPI-LB $= \min(|\theta_{0.025}|, |\theta_{0.975}|)$, where θ represents the values from the posterior distribution for $\Delta \log \phi_j$. By inclusion in this list, all 95% CPIs excluded zero. Posterior medians were again taken as point estimates of $\Delta \log \phi_j$
- x_{3j} is the initial Z_j score obtained when fitting the 100% tissue samples
- x_{4j} is the absolute value of the BMEA B_j-statistic obtained when fitting the 100% tissue samples

Thus β_0 represents the baseline probability of a confirmed slope (on the logit scale), whilst the remaining terms represent any changes to this probability due to the respective predictor variables.

Results from this analysis (Table 5.9) suggested that the original ranking statistic had minimal effect on the probability of a confirmed non-zero slope, which was unsurprising given this was only present in a small number of discrete levels, as a result of the initial selection criteria. All other terms investigated were found to have a significant impact on the probability of a confirmed slope and by implication a true AS event (all $p < 0.006$). Given the findings of Section 5.4.2, this was not entirely unexpected.

Table 5.9 – *Results from fitting the logistic regression model in Equation 5.1, in which the factors impacting the probability of accurate AS detection are being considered. The lower bound of the 95% CPI as defined in Equation 5.1 is indicated as CPI-LB. Significant p-values are indicated with asterisks as per the standard conventions of* R. *No multiple testing considerations were incorporated.*

Term	**Interpretation**	**Estimate**	**Std. Error**	**Z**	**Pr(> |Z|)**	
β_0	(Intercept)	-1.963	0.557	-3.526	4.22E-04	***
β_1	nProbes	0.057	0.021	2.750	5.96E-03	**
β_2	LB	2.103	0.316	6.656	2.82E-11	***
β_3	Z	0.005	0.001	6.267	3.68E-10	***
β_4	$\|B\|$	0.120	0.079	1.520	1.28E-01	

Probesets were then separated into four groups based on quartiles of Z_j scores observed across all expressed genes, and the minimum value found for the CPI-LB, at which the True Positive Rate was 95% (Figure 5.12). As expected, the lower the Z_j-score, the higher this minimum value was found to be, with values as shown. Thus, a high degree of confidence in putative AS events could be expected if restricting candidate exons to those in the upper two quartiles based on Z_j scores, and requiring the 95%CPI to exclude the region contained by $[-\kappa, \kappa]$ for $\kappa = \log(1.5)$. This was highly consistent with Section 5.4.2 and became the suggested methodology for selection of putative AS events, giving a set of 288 high-confidence candidate AS events if applied retrospectively to the initial comparison of 100% Heart and 100% Brain tissues. 272 of these were considered as confirmed via a non-zero slope of the regression line (Table 5.10). Although this showed a high true positive rate, given that 829 probesets were confirmed with an AS event from Table 5.8, this represented a relatively low power of detection. Whilst the number of probes was formally found to contribute to the likelihood of a correctly identified AS event (Table 5.9), this was not included in the candidate selection methodology as minimal improvements in performance were noted.

Whilst all confirmation above was obtained by fitting probesets across mixture levels, all candidate probesets from the genes with a probeset in the 10 most highly ranked candidate AS events were inspected manually (Appendix C). Three separate criteria were used for this inspection using 1) raw probe intensities which were supportive; 2) posterior medians for ϕ_{hj} which tracked consistently across all mixture levels, and 3) the list of transcripts which could satisfactorily explain the candidate AS event. Good overall support for the confirmed results were found, especially for probesets considered to represent brain-specific isoforms.

Table 5.10 – *Results for the high-confidence set of exons selected as candidate AS events using the upper two Z_j quantiles and setting $\kappa = \log(1.5)$. Probesets were considered as confirmed AS events if the direction of the slope was consistent, and the raw p-value for $H_0 : \beta_1 = 0$ lead to a rejection of H_0.*

Tissue	**Total**	**Confirmed**	**TPR**
Brain	188	187	0.99
Heart	100	85	0.85
Total	288	272	0.94

For many heart-specific isoforms, supporting evidence appeared less conclusive, and whilst this is not directly addressed here, subsequent explorations revealed the potential origins of this, as will be discussed in Section 6.2.

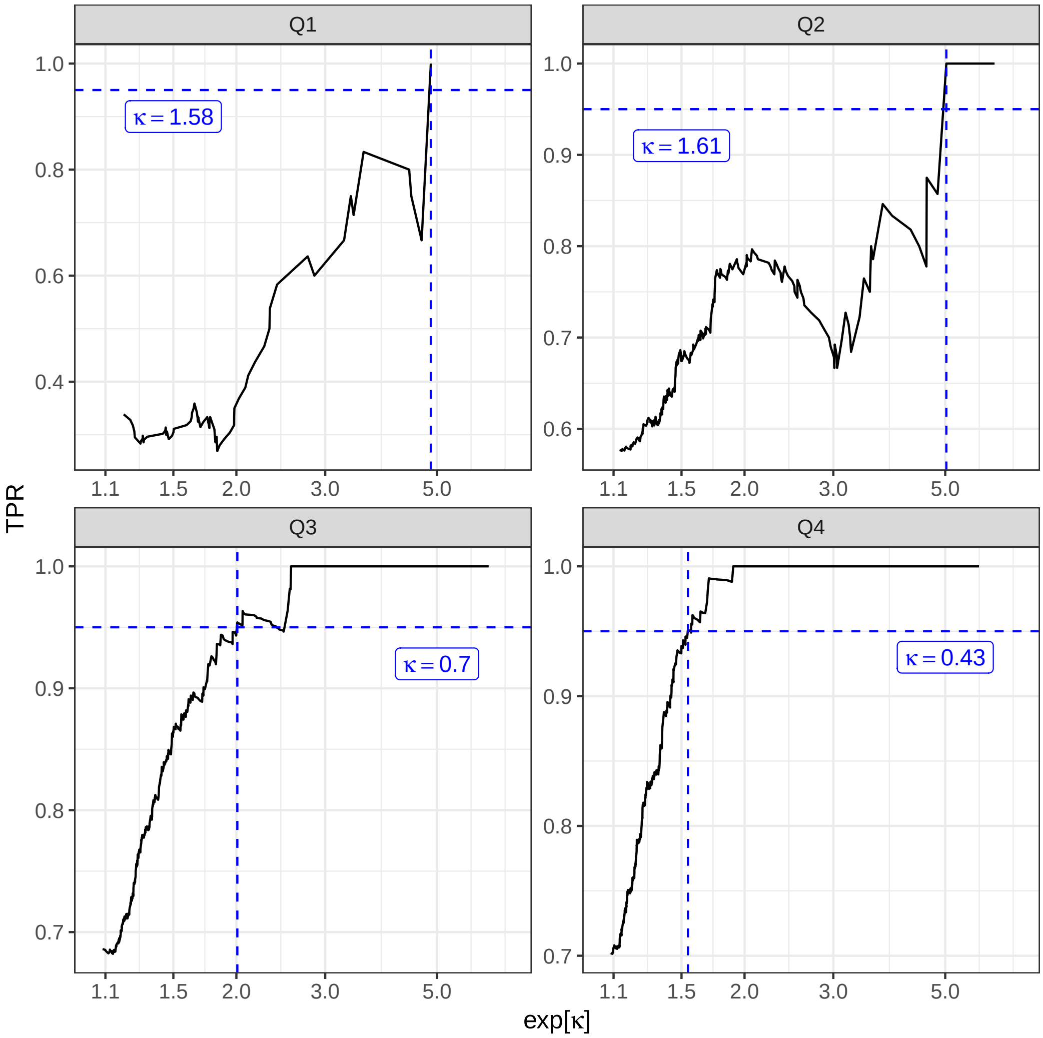


Figure 5.12 – *Minimum values for κ required for a 95% true positive rate (TPR). Values are presented by quantiles based on Z_j scores, with probesets being more consistently accurate for those with higher Z_j scores. Threshold values (κ) shown as labels are given on the log scale. For easier interpretation, these are presented as untransformed fold-change on the x-axis.*

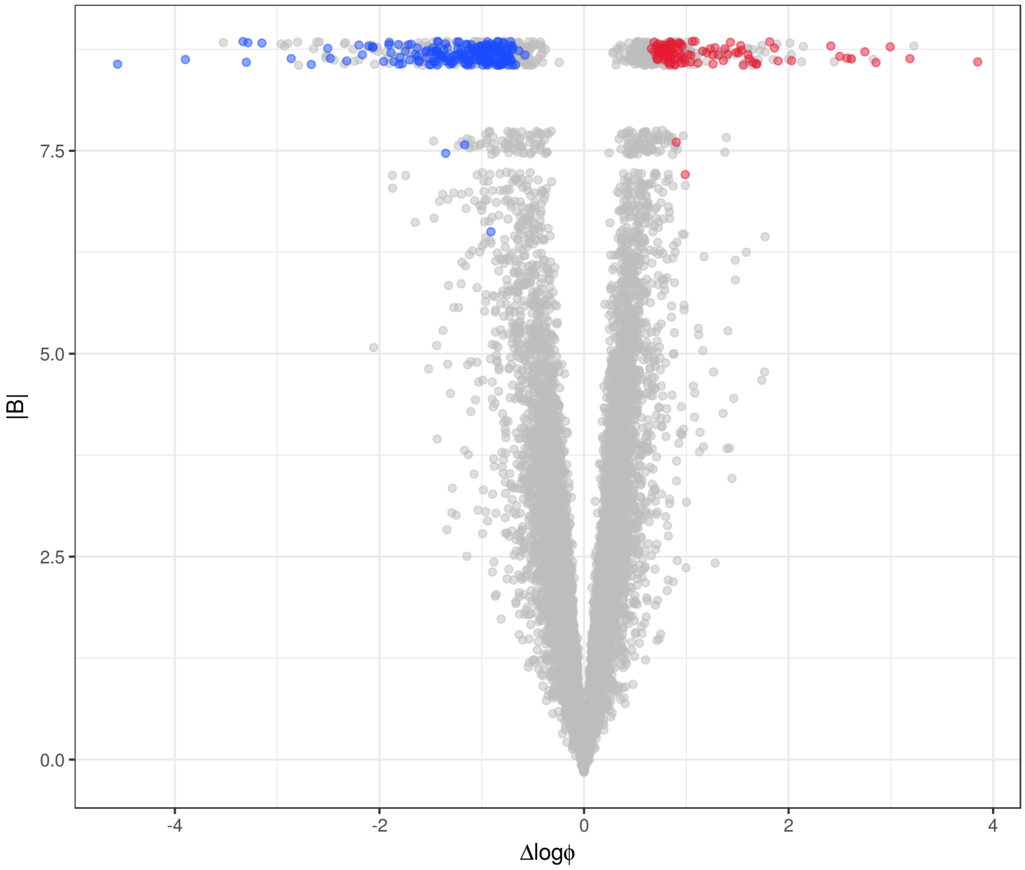


Figure 5.13 – *Volcano plot for exon-level B-statistics against* $\Delta \log \phi_j$ *based on the 100% Heart Vs 100% Brain comparison. Posterior median values for* $\Delta \log \phi_j$ *are used as point estimates. Points are only shown for the 1124 genes which provided the original candidates for detection of alternate splice events. Coloured probesets denote those in the high-confidence set after the CPI-LB and* Z_j*-score inclusion criteria as defined above. Putative heart-specific exons are shown in red, whilst putative brain-specific exons are shown in blue. Jitter has been added on the vertical axis.*

Comparison With FIRMA

Given the confirmed status of multiple probesets in the above section, the results from analysis using FIRMA were compared to those under BMEA. Probesets were considered as significant under FIRMA if receiving an FDR-adjusted p-value < 0.05 for the probesets within the tested 1124 genes. As the FIRMA score does not correspond directly to any true parameter, no further filtering was done. Of the 272 confirmed AS events from the list of 288 high-confidence candidates, 266 were also considered significant under FIRMA, showing that this model is also strongly capable of detecting AS events, particularly given the robust signal-to-noise ratio within this smaller list via the filtering based on Z_j. Of the 16 within the high-confidence list for which the AS event was not confirmed, 11 were also considered as significant under FIRMA, which was again highly-comparable to BMEA

Of the initial 1273 probesets for which AS status was assessed by linear regression in Section 5.4.5, 987 were considered as candidate AS events under FIRMA using an FDR of 5%. 717 were supported as true positives, using raw p-values to reject H_0 at $\alpha = 0.05$. The remaining 270 were not confirmed as AS events, giving the unfiltered FIRMA analysis an FDR $\approx 27\%$. Whilst filtering of candidate probesets under FIRMA is also plausible, this was beyond the scope of this body of work and not investigated further.

5.5 Discussion

Taking RMA-based approaches as the industry standard for differential expression, BMEA parameter estimates were highly comparable at the gene level (Section 5.4.2). The additional incorporation of probe sequence information and the use of DABG procedures removed some genes, for which the true signal component of the probe intensity was indistinguishable from the background signal, and generally gave a greater dynamic range for the estimates of fold-change (Figure 5.8). An additional filtering step which may be advantageous for analysis using BMEA could include checks for the number of remaining exon-level probesets in the final gene-level estimate. As some genes had considerable numbers of these removed, these may require some level of scrutiny before being passed on to researchers for follow-up analysis.

At the exon-level, the intuitive interpretation of model parameters under BMEA allowed for the development of a robust filtering approach during selection of candidate AS events. Whilst a large number were considered as confirmed during manual inspection (Appendix C), many of those initially considered as candidates but not included in the high-confidence group, were considered inconclusive reinforcing the selection methods detailed above. Many of these were additionally drawn from complex sets of transcripts within each gene and resolution by inspection without access to the source tissue, may be difficult . However, the strong improvement in accuracy of results using the filtering steps reinforced the effectiveness of Z-scores as a valuable measure of true signal strength above background signal.

The stringent filtering method described above resulted in a surprising lack of power under BMEA, and although the accuracy rate was encouraging, the exclusion of so many candidates under the filtering approach was far short of expectations. Comparison with FIRMA revealed that many of the confirmed candidates were also detected under this approach, however, without an equivalent filtering and selection of high-confidence candidates under FIRMA, a direct comparison between the two methods is not easily drawn .

Chapter 6

BMEA Analysis of the $\mathbf{T_{reg}}$ Dataset

6.1 Initial BMEA Analysis of the $\mathbf{T_{reg}}$ Dataset

6.1.1 Fitting the Dataset

The BMEA fitting procedure was then applied to all 18 arrays previously analysed in Chapter 2 as the $\mathrm{T_{reg}}$ dataset. Arrays were first quantile normalised, and all BMEA parameters were set to default values. Anti-genomic (AG) probes were used for estimation of λ_{il} and δ_{il} for the background signal component. The custom CDF described in Section 5.2.2 was used for mapping probes to exons and genes, as for Section 5.4. Initial fitting on a quad-core, Windows workstation completed within 5 days, whilst subsequent runs using an Linux workstation with 20 threads completed in 17 hours.

To enable comparisons with the PLM/FIRMA methods, the normalised arrays were fit using RMA background correction and probe-level models, with FIRMA scores (Purdom et al., 2008) calculated as implemented in `aroma.affymetrix`. Instead of the comparisons within donors as performed in Chapter 2, donors were included as blocking variables with correlations estimated (Smyth et al., 2005) as per the `limma` Users Guide. Sample weights were calculated at the unit (gene) and group (exon) level separately using the function `arrayWeights` (Ritchie et al., 2006) (Figure 6.1). Moderated $\tilde{t}$-statistics (Smyth, 2004) were calculated and FDR-adjusted p-values were used to obtain the final ranked lists. The complete workflow for analysis using PLM/FIRMA is able to be performed using a single CPU core in under two hours.

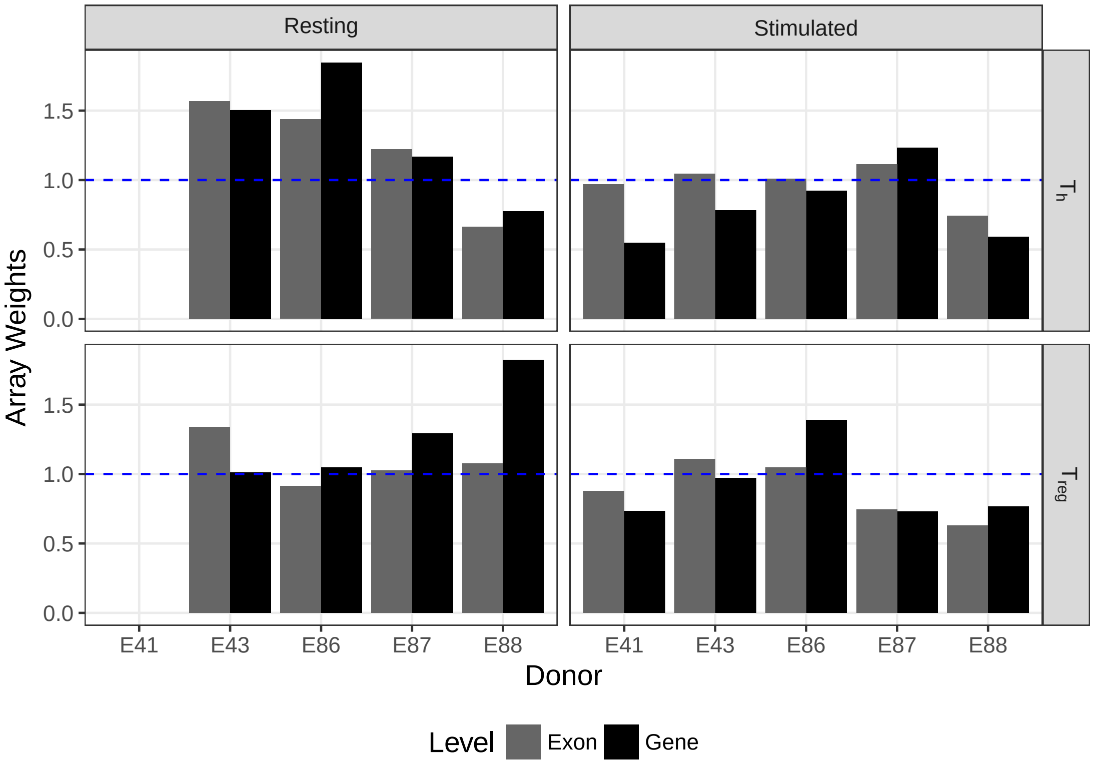


Figure 6.1 – *Array-level weights obtained for fitting using PLM/FIRMA approaches. The idealised equal-weighting value of $w_i = 1$ is highlighted as the dotted blue line. As described in Chapter 3, the expansion denoted as E41 was unsuccessful in resting cells and these samples were not run on any arrays.*

6.1.2 Performance at the Gene Level

Comparison of BMEA Values to PLM Values

Using $Z_g < 4.265$ to identify genes which were not DABG (Section 4.3.1), 3697 genes were excluded under BMEA, leaving 31,504 for further analysis. Using posterior mean values for c_{hi} converted to the $\log_2$ scale as directly comparable to PLM-derived chip effects, correlations between the two sets of expression estimates ranged from 0.924 to 0.951, consistent with the Tissue Mixture dataset (Section 5.4.3). The four resting $\mathrm{T_{reg}}$ samples are shown as examples in Figure 6.2. Posterior mean values for $\log_2$FC were also compared to PLM-derived estimates (Figure 6.3). In contrast to the Tissue Mixture dataset, correlations after exclusion of undetectable genes were considerably lower than for simple expression values, with the lowest correlation in the $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison for stimulated cells ($\rho = 0.640$). The highest correlation was in the Stimulated Vs Resting comparison for $\mathrm{T_h}$ cells ($\rho = 0.880$).

In general, BMEA-derived estimates of fold-change were higher than those obtained under PLM (Figure 6.3), with a considerable number of points in each comparison receiving BMEA-derived estimates > 2, whilst being below the common cut-off value of logFC $= 1$ under the PLM approach. Additionally, some genes considered undetectable under BMEA received estimates of logFC > 1 when using PLM. When comparing this common-use threshold for "biologically significant" fold-change using the line of best fit (Figure 6.3), it was noted that an estimated logFC $= 1$ under PLM corresponded on average to between 1.5 and 1.8 under BMEA. Similarly a logFC estimate of 1 under BMEA translated to logFC ≈ 0.5 under PLM. Whilst considerable variation is to be expected, the generally larger point estimates under BMEA are consistent with those observed in Figure 5.8.

Ranking statistics for differential expression were additionally compared (Figure 6.4) taking the moderated $\tilde{t}$-statistics obtained under PLM and the BMEA-derived B-statistic. A similarity of trend was again noted amongst the more highly-ranked genes, however, it was clear that amongst the genes lower down either list, significant variability may be observed between DE gene lists. Example threshold values for consideration as significantly DE are shown, with a theoretical value for the B-statistic given as $|B| > 3.66$. This corresponds to the value at which a 95% central posterior interval (CPI) for the sampled posterior distribution for logFC would be likely to exclude zero, under the default BMEA parameters.

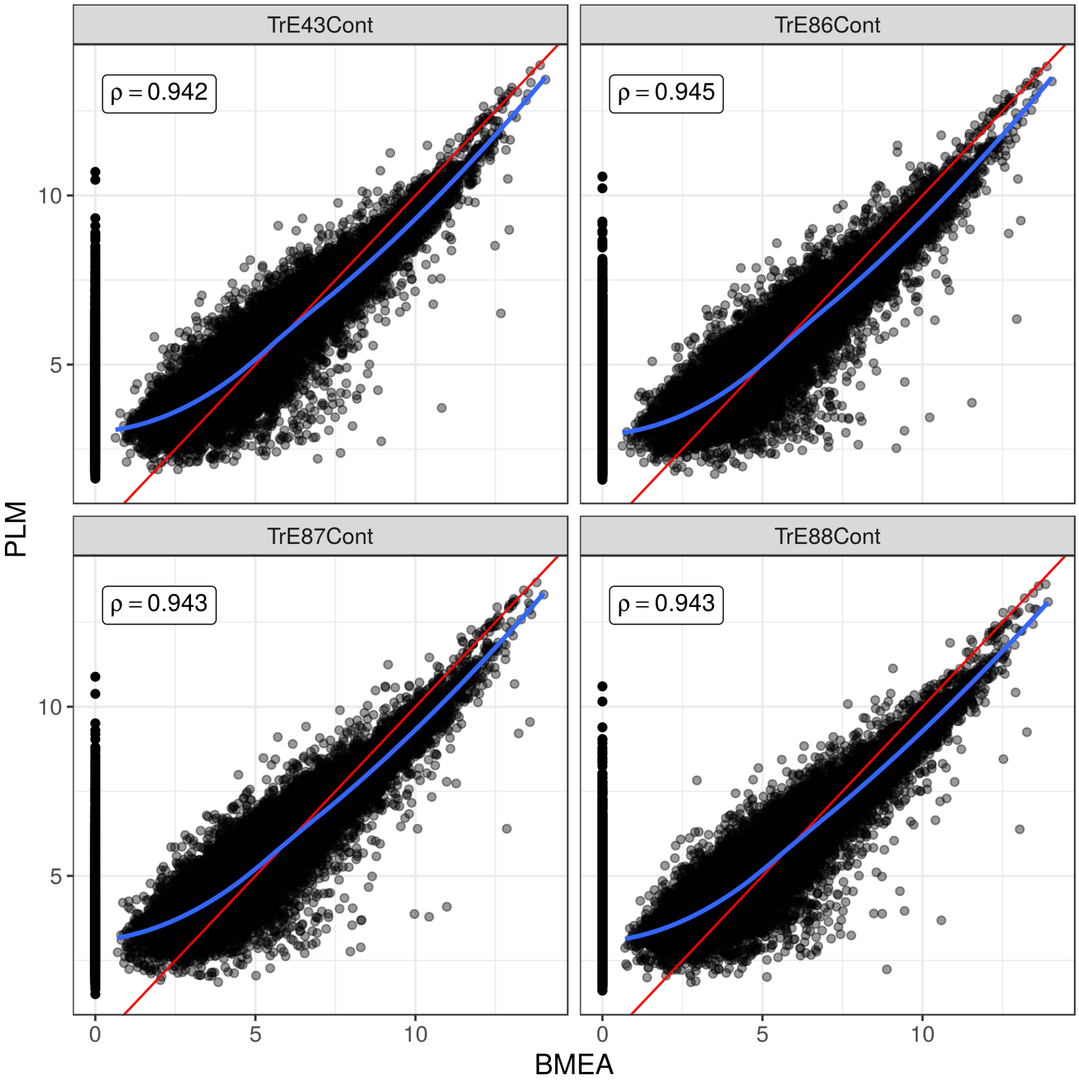


Figure 6.2 – *Comparison between posterior means for c_{hi} under the BMEA approach and PLM-derived chip effects. All resting T_{reg} samples are shown as examples ($n = 4$). BMEA values are transformed to the $\log_2$ scale. The line $y = x$ is shown in red, with loess line of best fit shown in blue, as calculated excluding genes not considered as DABG under BMEA. Undetectable genes are clearly visible as the solid column of points in the LHS of each frame. Correlations between the two sets of values are given in the top left corner for detectable genes only.*

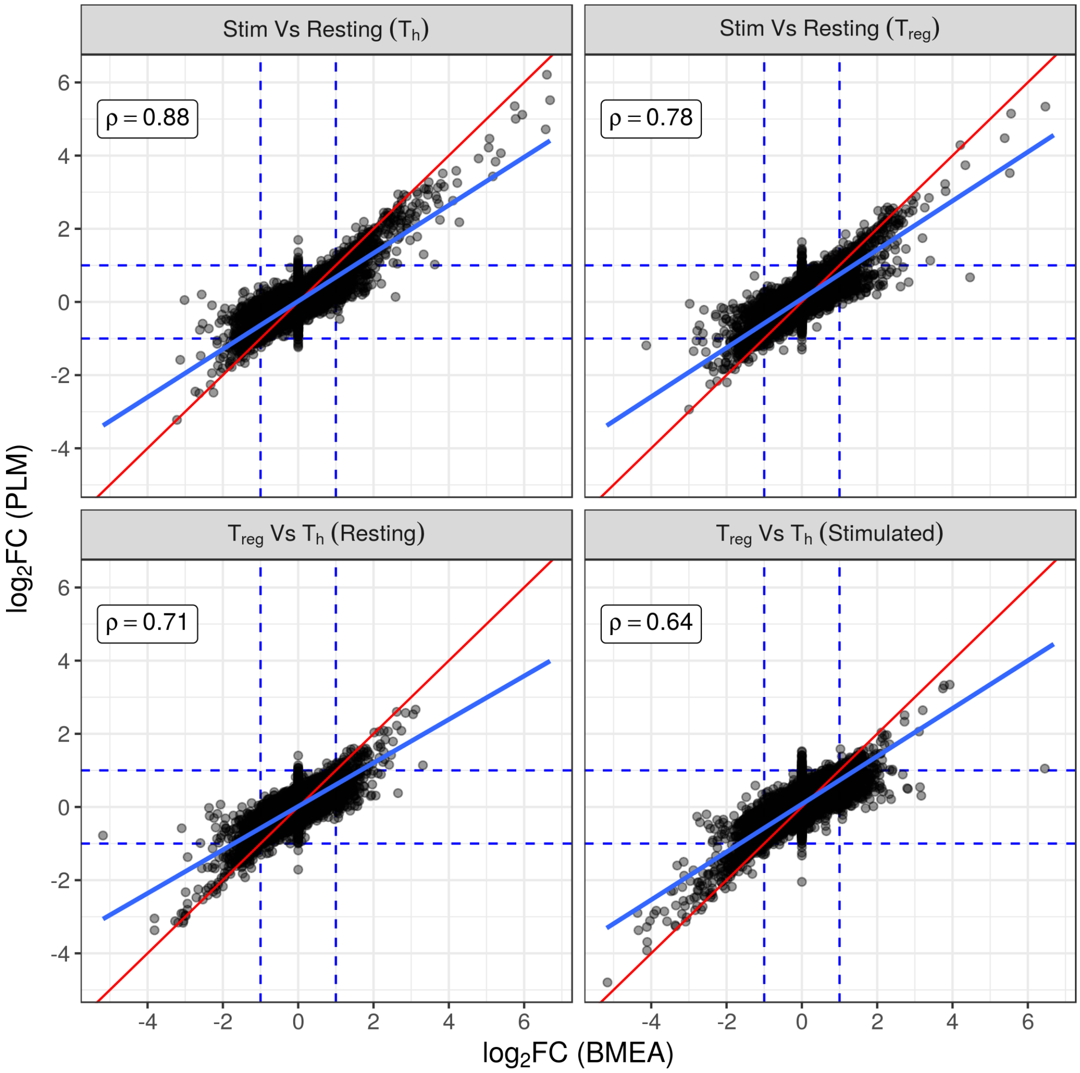


Figure 6.3 – *Comparison of logFC estimates between BMEA and PLM approaches, with the common-use threshold of $logFC = \pm 1$ indicated by the dashed blue line. The line $y = x$ is shown in red, with the regression line of best fit shown in light blue. Correlations between the two sets of values are given in the top left of each frame.*

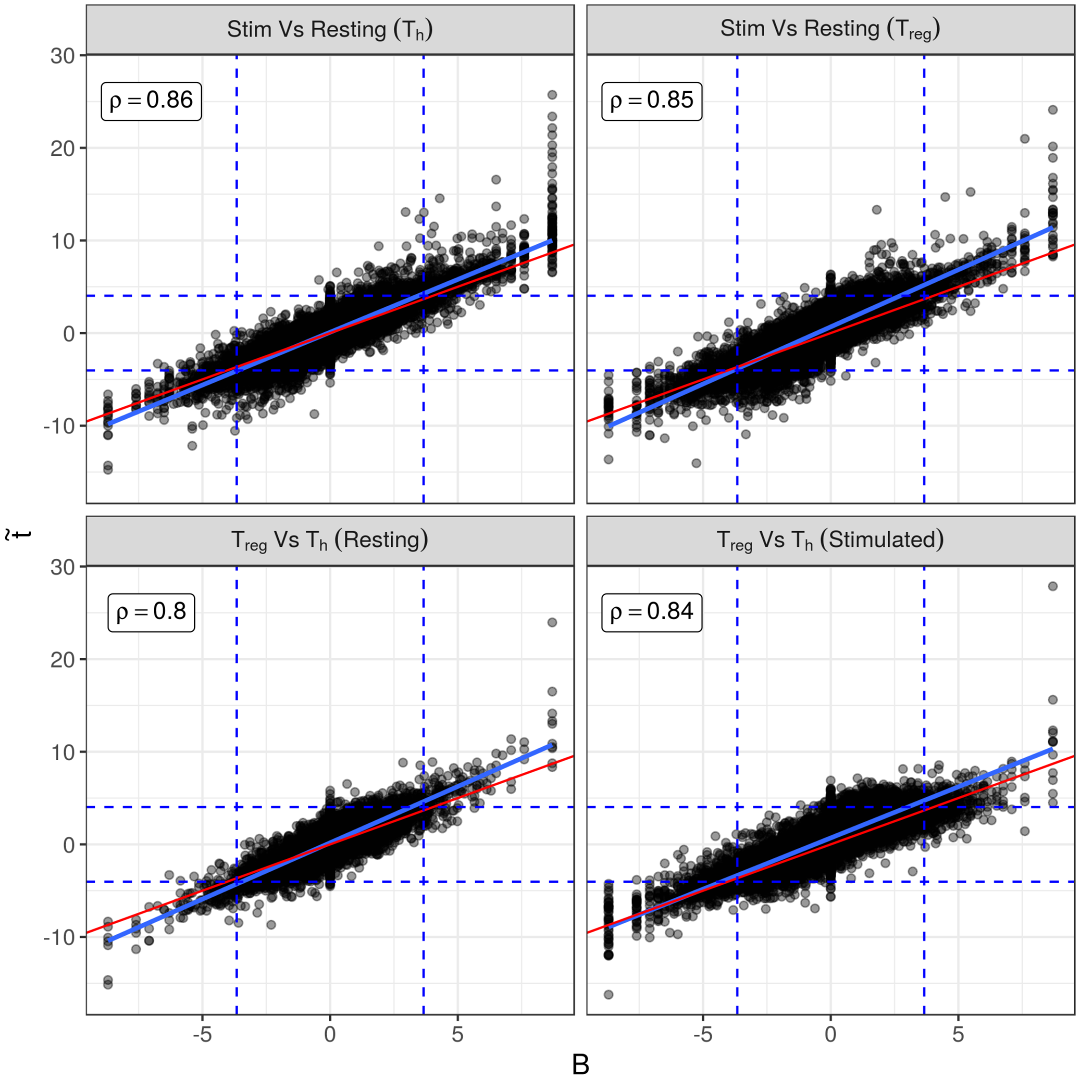


Figure 6.4 – *Comparison of ranking statistics between BMEA and PLM approaches. The BMEA-derived B-statistic is shown on the x-axis, whilst the moderated $\tilde{t}$-statistics obtained under PLM methods are shown on the y-axis. The line $y = x$ is shown in red, with the regression line of best fit shown in light blue. The dashed blue lines indicate example threshold values for consideration as DE without consideration of the magnitude of fold-change, such as $|\tilde{t}| > 4.03$ as the 0.995 percentile of a t_5 distribution, and $|B| > 3.66$. This value for the B-statistic represents the value at which a 95% CPI for logFC is likely to exclude zero. Correlations between the two sets of values are given in the top left of each panel.*

Selection of DE Genes Using BMEA

Under the PLM model, DE genes were simply defined as those receiving an FDR-adjusted p-value < 0.05 and with an estimated $|\text{logFC}| > 1$, giving the total numbers of DE genes in Table 6.1. The skew in $\tilde{t}$-statistics observed in Section 2.3.1 was not addressed as this would either be comparable across methods, or if one method better compensated for this, it would be an important difference between approaches. Under the the CPI-LB approach, there is no set rule for choosing a value κ such that a gene would be considered as DE if the 95% CPI excluded the region [-κ, κ]. Three candidate values were tested (Table 6.2). Choosing $\kappa = 0$ yielded vary large lists of DE genes, whilst setting $\kappa = 0.2$ and using posterior estimates of logFC to the $\log_2$ scale yielded DE gene lists of a generally comparable size to those generated under the PLM approach. This value corresponded to $\kappa \approx \log_2(1.15)$ and was slightly more conservative than the value chosen in Section 5.4.2 to achieve a comparable error rate to PLM approaches, but less conservative than that used in Section 5.4.3.

Table 6.1 – *Number of significant genes for each comparison under the PLM approach, using an FDR of 0.05 and* $|\log FC| > 1$ *as the criteria for differential expression. The number of DE genes without filtering based on fold-change (i.e. FDR Only), is also given. The raw p-value* ($p_{\max}$) *corresponding to an FDR of 0.05 is given in the final column.*

		DE Genes		
Comparison	**Cell Type**	**FDR Only**	**$\|$logFC$\|$ > 1**	$\mathbf{p}_{\max}$
T_{reg} Vs T_h	Resting	537	182	7.31e-04
T_{reg} Vs T_h	Stim	7979	412	0.011
Stim Vs Resting	T_{reg}	4626	462	0.006
Stim Vs Resting	T_h	2614	521	0.004

Table 6.2 – *Number of significant genes for each comparison under the BMEA model, varying the threshold (κ) for the 95% CPI-LB in steps of 0.1. The totals considered as DE under the selection methods for the PLM model (Table 6.1) are also given.*

Comparison	**Cell Type**	$\kappa = 0$	$\kappa = 0.1$	$\kappa = 0.2$	**PLM**
T_{reg} Vs T_h	Resting	322	227	148	182
T_{reg} Vs T_h	Stim	1433	780	458	412
Stim Vs Resting	T_{reg}	1225	787	536	462
Stim Vs Resting	T_h	1193	802	580	521

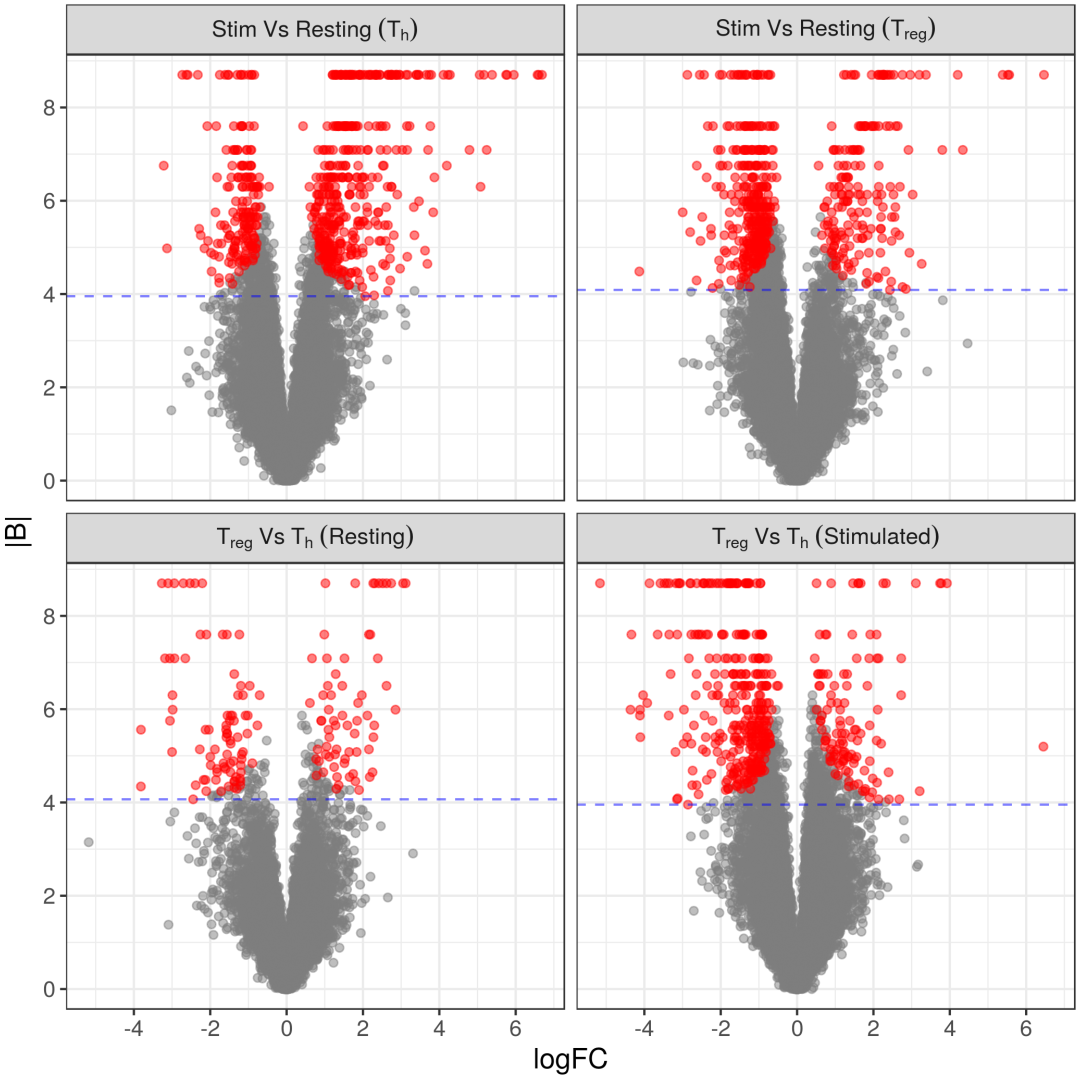


Figure 6.5 – *Volcano plots for each of the four T_{reg} comparisons using BMEA. Genes considered as significantly DE using the CPI-LB method with $\kappa = 0.2$ are indicated in red. Posterior means are shown as point estimates for fold-change, and are shown on the $\log_2$ scale. The minimum value for $|B|$ amongst each set of DE genes is shown as blue horizontal lines.*

Comparison of DE Genes Found Under BMEA and PLM

Comparison of the DE genes under each approach revealed differences in gene lists which were surprisingly large (Table 6.3). However, removal of the restriction on logFC for PLM, and setting $\kappa = 0$ during selection (Table 6.4) showed that >80% of the genes considered as DE under BMEA were also detected under PLM. This similarity between lists suggests that the hard thresholds used during selection of DE genes was a significant source of the discrepancy between the two approaches.

Inspection of genes considered as DE under BMEA, using PLM-derived values in volcano plots (Figure 6.6), revealed that most of the genes unique to BMEA were excluded from the PLM-derived lists due to $|\widehat{\text{logFC}}| < 1$, with the vast majority receiving FDR-adjusted p-values < 0.05. However, a small number of genes in each comparison were significantly DE under BMEA, but yielded estimates of logFC near zero on the PLM-derived lists, along with p-values well beyond the threshold for consideration as DE.

Table 6.3 – *Summary of DE genes using both PLM and BMEA approaches for each comparison, including the use of a filter on fold-change to exclude candidate genes. Genes considered DE under PLM, but which were not detectable under BMEA, are included in the totals and given in brackets.*

Comparison	Cell Type	Common Genes	BMEA Only	PLM Only
T_{reg} Vs T_h	Resting	103	45	79 (4)
T_{reg} Vs T_h	Stim	234	224	178 (37)
Stim Vs Resting	T_{reg}	262	274	200 (20)
Stim Vs Resting	T_h	342	238	179 (6)

Table 6.4 – *Comparison of DE genes under both approaches without filtering on logFC. Genes considered as DE under both approaches are included as Common Genes, whilst those only considered as DE under one approach are included in the respective columns. Numbers of genes considered as DE under PLM, but which were not considered as DABG under BMEA are also shown in brackets.*

Comparison	Cell Type	Common Genes	BMEA Only	PLM Only	
T_{reg} Vs T_h	Resting	260	62	277	(12)
T_{reg} Vs T_h	Stim	1247	186	6732	(347)
Stim Vs Resting	T_{reg}	1121	104	3505	(159)
Stim Vs Resting	T_h	1067	126	1547	(48)

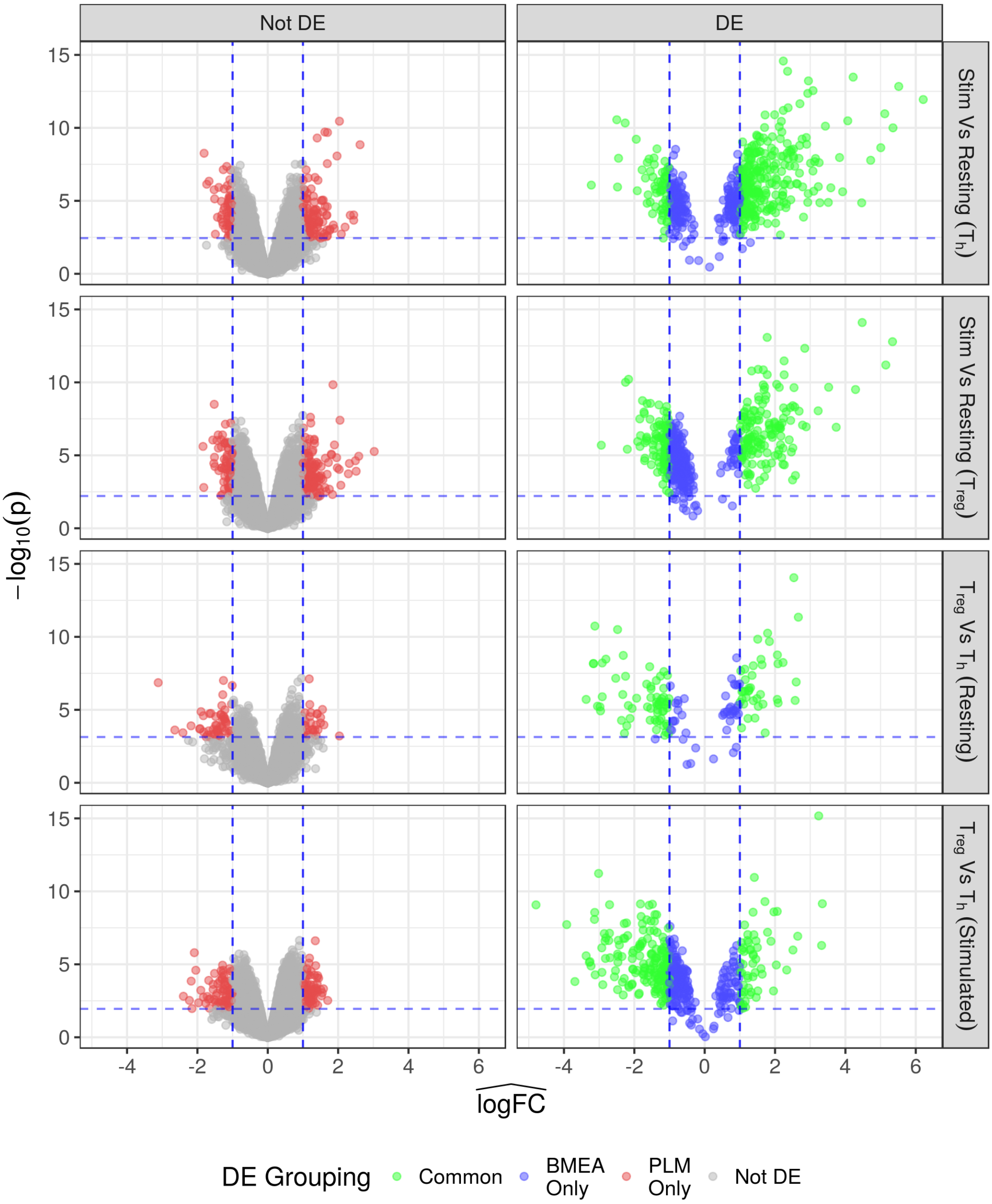


Figure 6.6 – *Volcano plots for each of the four T_{reg} comparisons showing estimates of logFC and p-values obtained using the PLM approach. Genes considered DE under BMEA are shown in the right panels, whilst those in the left were not considered DE under BMEA. Colours denote significance under both models for each comparison. Cutoffs used for classification as DE under the PLM approaches are shown as the dashed blue lines.*

The reverse visualisation, overlaying genes considered as DE under PLM using BMEA-derived estimates of logFC and B-statistics (Figure 6.7), clearly showed the genes considered DE under PLM, but undetectable under BMEA (Table 6.3) as the red dots at (0, 0). Considerable numbers of DE genes under PLM were also ranked well below the minimum $|B|$ statistics for BMEA. The vast majority of PLM-only DE genes also received estimates of $|\text{logFC}| > 1$ under BMEA, however a considerable number of BMEA-only DE genes were highly ranked using the B-statistic, with point estimates for $|\text{logFC}| < 1$. Much of this is likely due to the larger dynamic range of estimates for fold-change under BMEA than under PLM, as well as the differing selection methods.

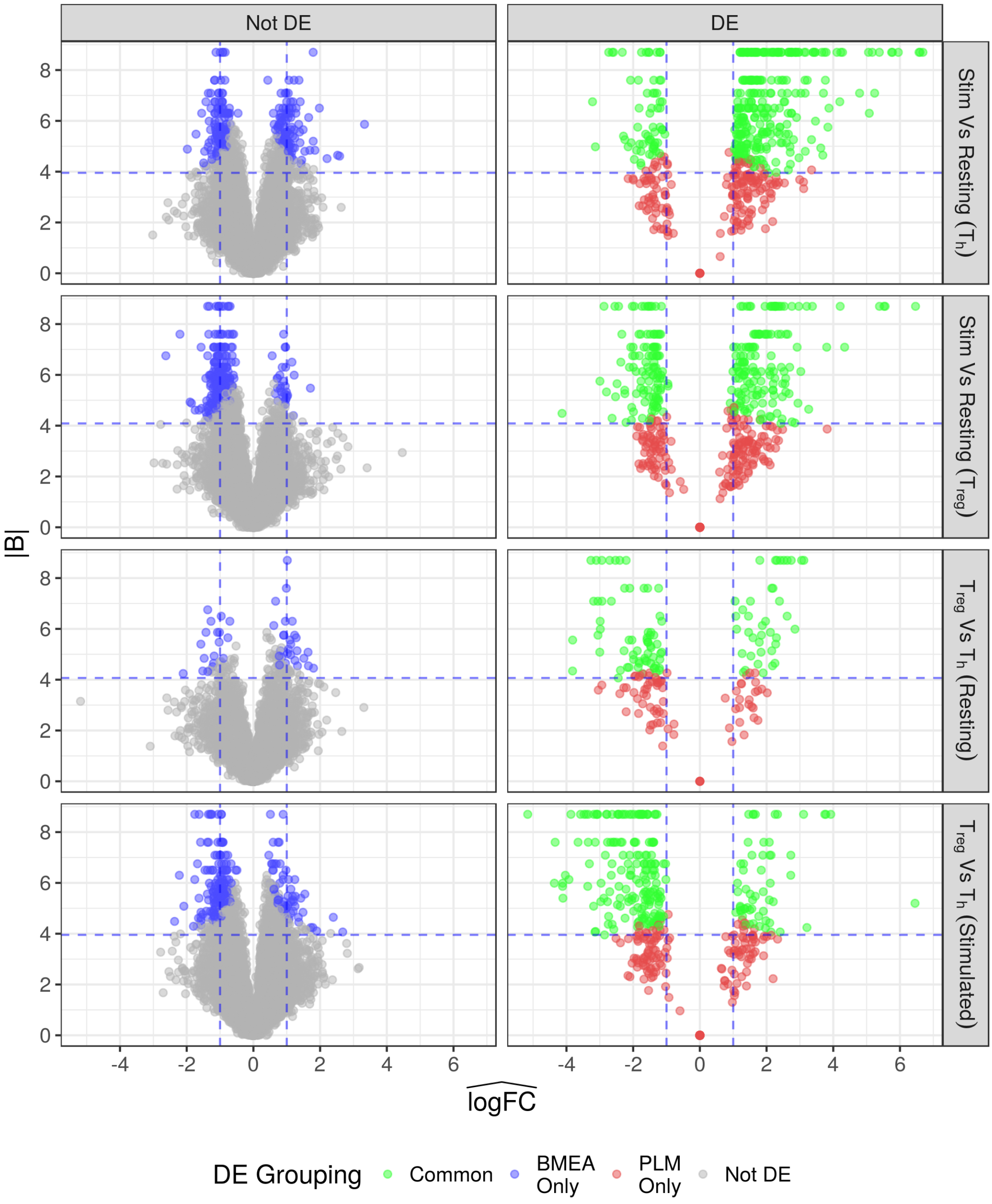


Figure 6.7 – *Volcano plots for each of the four T_{reg} comparisons using the BMEA approach. Genes considered as significantly DE under the PLM method are shown on the right panels, with those not DE under PLM are included on the left. Colours denote significance under both models for each comparison. Point estimates of logFC are represented using posterior means. The minimum $|B|$ statistics observed under the selection approaches above are indicated using a horizontal blue line, whilst the values indicating $|\log FC| > 1$ are shown as the vertical blue lines.*

Genes Unique to PLM

With the exception of genes not DABG under BMEA (Table 6.3), many genes declared as DE only under PLM were ranked considerably lower by BMEA (Figure 6.7). Point estimates of fold-change for the majority of these genes were beyond the ±1 values, but below the minimum B-statistic obtained under BMEA. As such, the primary reason for these differences was likely to be the breadth of the 95% CPI. An important cause of this higher ranking under PLM was found to be the posterior estimates for residual variance ($\tilde{s}_g$) for each gene, generated using the function `eBayes()` from the package `limma`. This process will either increase or shrink the initial estimates (s_g) to provide moderated $\tilde{t}$-statistics (Smyth, 2004) ($\tilde{t}_g$). If genes are highly variable, the shrunken estimates $\tilde{s}_g$ will be lower than the initial values, giving more extreme values for $\tilde{t}_g$ than without moderation, and in turn these genes become more highly ranked. BMEA includes no comparable procedure, however this may be a possibility in future iterations of the approach.

In Figure 6.8, genes which were detected as DE exclusively under PLM in any comparison, and which were excluded under BMEA due to a 95%CPI containing zero, were considered as "Unique To PLM". The $\log_2$ ratios of $\tilde{s}_g$ and s_g are shown for these genes, with the remaining genes being considered as significant under BMEA if detected at least one comparison, or "Never DE" if they were not considered as DE in any of the four comparisons. These ratios clearly showed that the genes uniquely detected as DE under PLM generally had strongly shrunken posterior values $\tilde{s}_g$ in comparison to other sets of genes.

Confirmation of this was obtained when inspecting the standard deviations of the sampled posterior distributions for logFC in each comparison (Figure 6.9). Genes unique to PLM in each comparison clearly showed greater variability than genes considered as DE under BMEA, as well as in comparison to the wider set of genes not considered DE. This was then considered as the likely source of this discrepancy between DE gene lists, and despite being a slight weakness of the BMEA approach, was not considered as a significant enough drawback to invalidate results under BMEA.

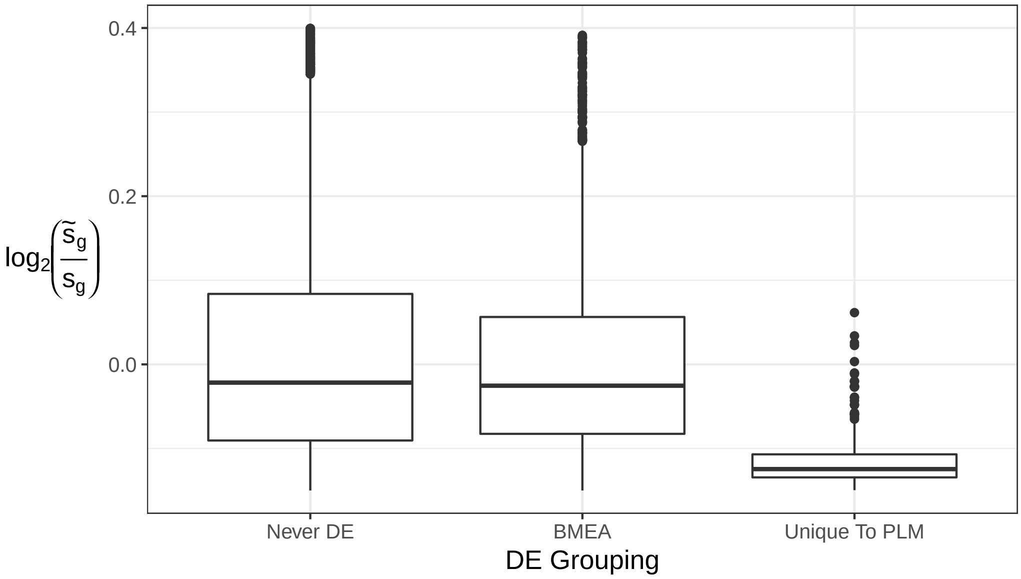


Figure 6.8 – *Ratios of the initial s_g values and the posterior estimates $\tilde{s}_g$, as obtained from the function* `eBayes()` *in the package* `limma`*. Genes are grouped into those never classified as DE (Never DE), those included as DE under BMEA in at least one comparison (BMEA), and those only declared as significant only under analysis using PLM (Unique To PLM). Only genes excluded from BMEA results by a 95% CPI which includes zero are denoted as unique to PLM. The upper limit of the y-axis has been truncated for easier visualisation.*

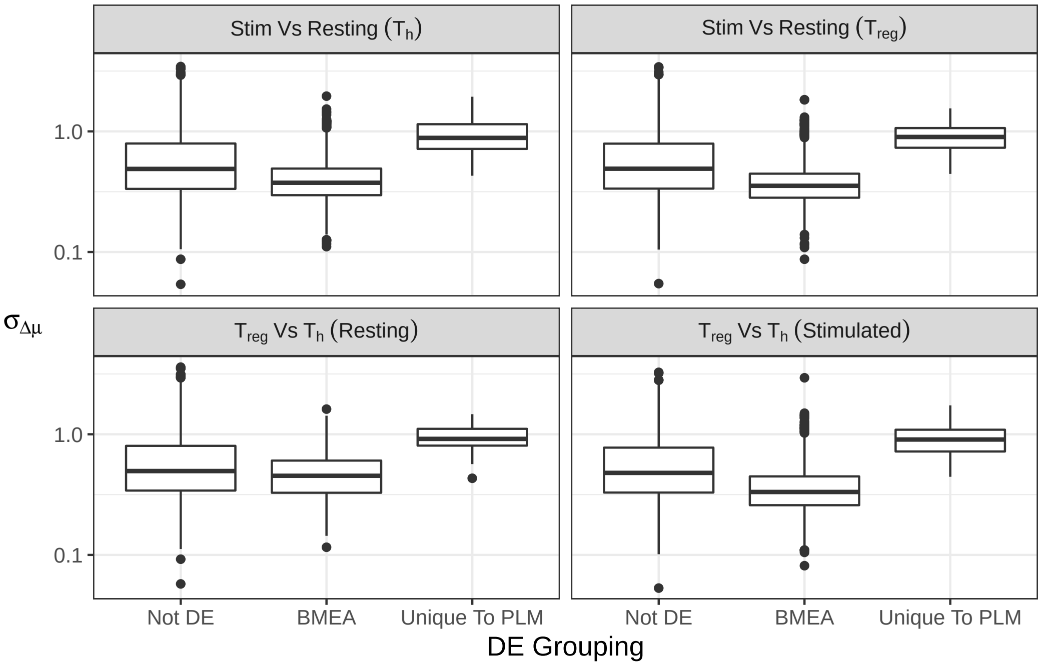


Figure 6.9 – *Standard deviations ($\sigma_{\Delta\mu}$) of the sampled posterior distributions for logFC in each comparison. Genes are grouped by whether they were considered as differentially expressed under neither approach (Not DE), under BMEA or uniquely under PLM. The y-axis is shown using the* $\log_{10}$ *scale.*

Genes Unique to BMEA

A large number of the differences between gene lists are due to genes being near the inclusion threshold under each approach, with values falling on either side under the different approaches. However, there were still a number of genes considered DE under BMEA, but not considered as DE under PLM, with the primary reason being a p-value well above the threshold for inclusion. These genes can be seen as blue points on the right panel in Figure 6.6, lying below the horizontal blue line. To ensure only genes unique to BMEA which were not as a result of the hard cutoff were investigated, genes with an FDR-adjusted p-value > 0.1 were checked for the number of exon-level probesets they contained, and if the DABG process had played any role in these differences. As seen in Table 6.5 and Figure 6.10, the majority of these were left with only a single exon under BMEA. Due to the low probe numbers, single-exon genes may be highly sensitive to the effects of modelling of background signal, and as such this may not be unexpected. For some of these genes, biased estimates of c_i will have been obtained under PLM by including exons which are absent. As this was an initial concern motivating the development of BMEA, this was also in keeping with expectations.

Five genes considered as DE under BMEA in a $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison, but with FDR-adjusted p-values > 0.1 under PLM were inspected at the raw probe intensity level (Figure 6.11). Estimates of logFC under both approaches are also given, along with the DABG Z-scores. For the genes with high Z-scores ($Z_g > 20$; Figures 6.11D & 6.11E) consistent differences between probe intensities were observed, despite a surprisingly high number of probes yielding intensities within the range expected by BG signal alone. However, for genes with lower Z-scores a greater proportion of genes were within the expected range for BG signal alone, and these differences in probe intensities were not as clear. These plots also

Table 6.5 – *Summary of exons remaining after DABG for genes considered as DE under BMEA but not under PLM, where the reason for exclusion under PLM was an FDR-adjusted p-value* > 0.1. *The total number of genes is first given, with the breakdown of those with one or more exons given in the subsequent columns.*

Comparison	Cell Type	Total Genes	Single Exon	Multiple Exons
Stim Vs Resting	$\mathrm{T_h}$	6	6	0
Stim Vs Resting	$\mathrm{T_{reg}}$	7	7	0
$\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$	Resting	6	5	1
$\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$	Stim	11	9	2

confirm that Z-scores are highly sensitive to any true signal, and exclusion of genes under this approach will only remove genes for which the vast majority of probes lie within the expectations for BG signal. The otherwise high degree of concordance between DE genes under BMEA and PLM, along with the similarities from Section 5.4.3 gives a high degree of confidence in the results obtained under BMEA.

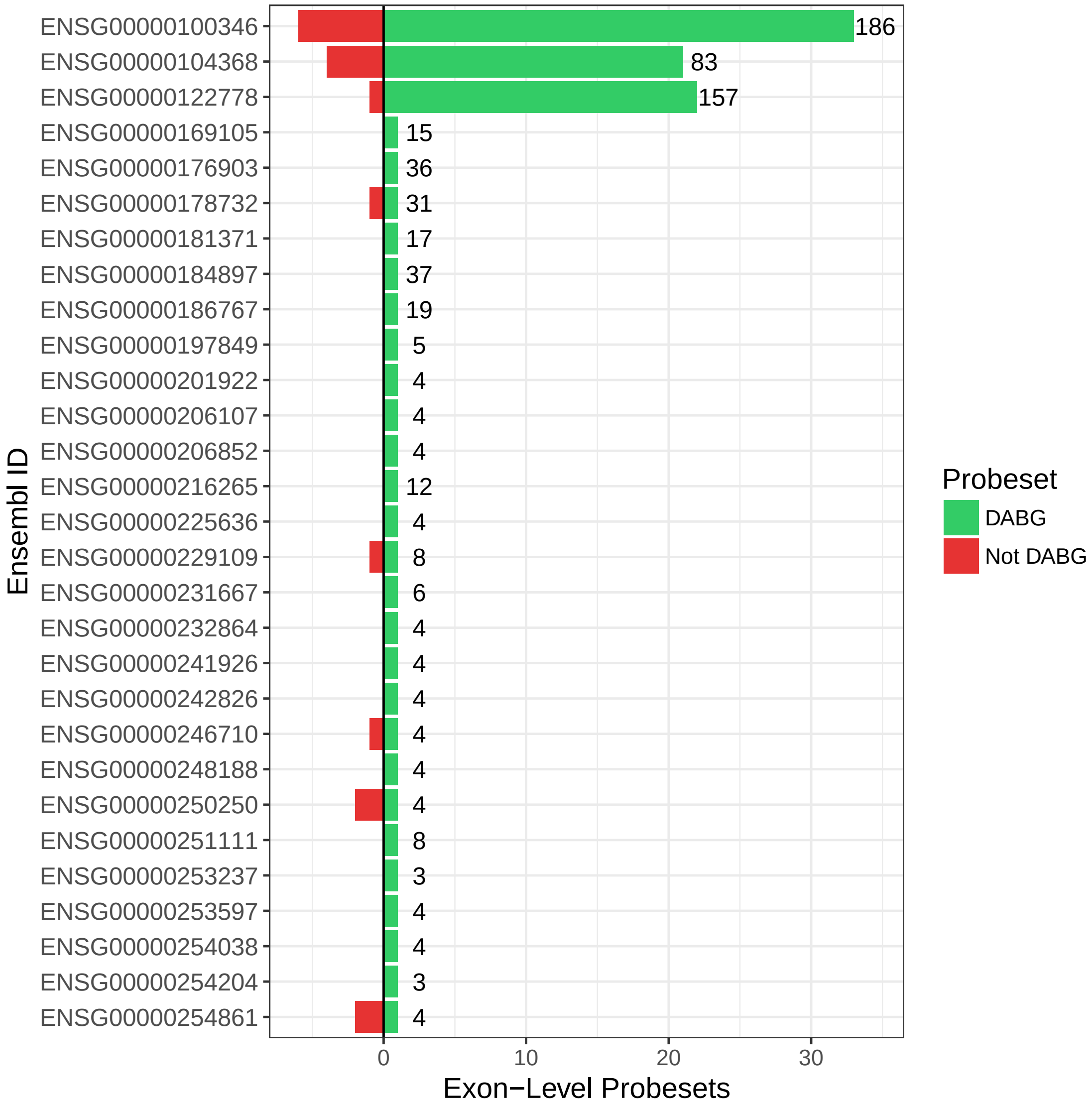


Figure 6.10 – *Probeset structure for genes considered as DE under BMEA, but with FDR-adjusted p-values*>0.1 *under PLM. The number of exon-level probesets retained after DABG-filtering are shown on the right of zero, with the number of not-DABG probesets being shown as negative numbers. The number of probes in all probesets retained after DABG is also given, with some single probeset genes showing a surprisingly high number of probes.*

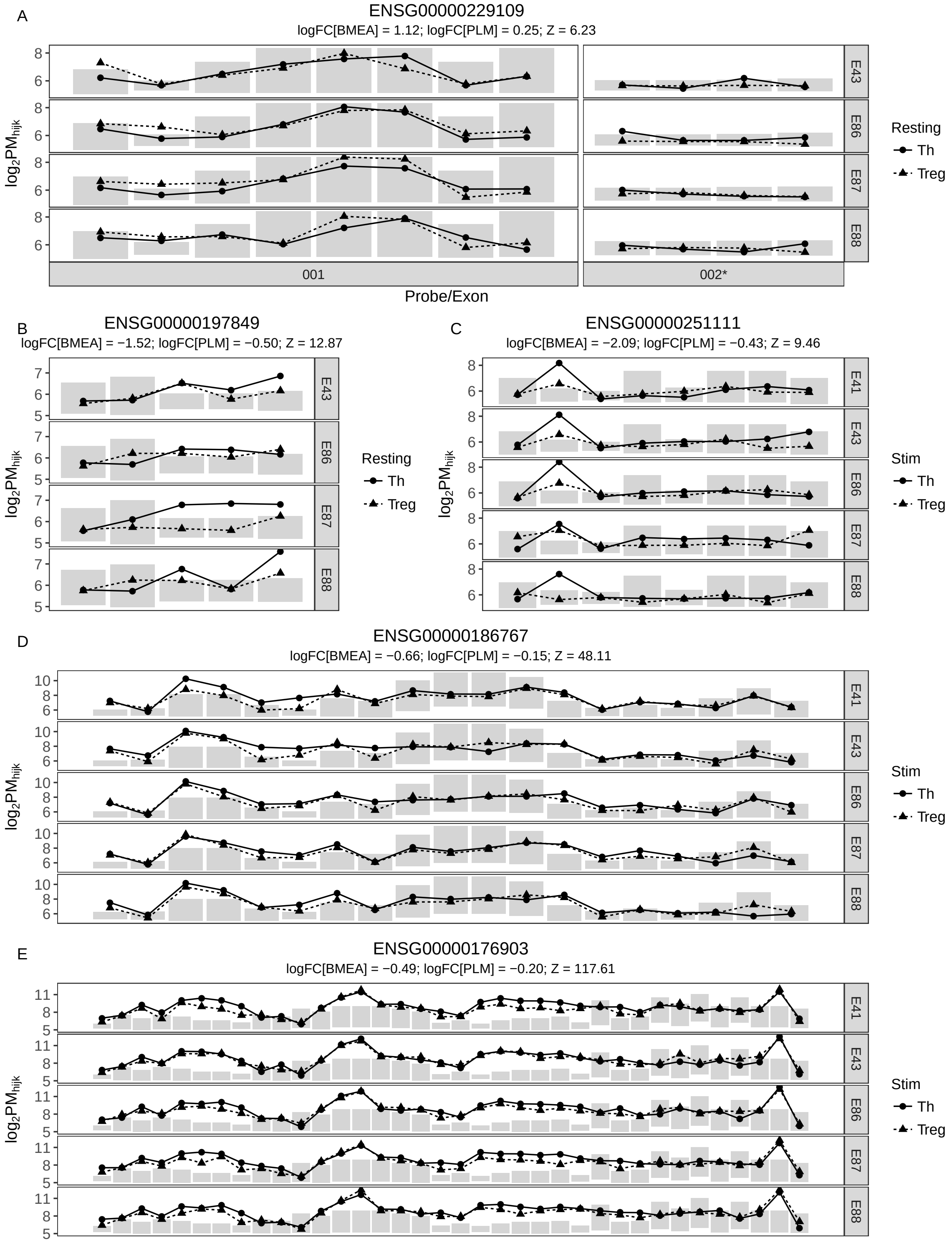


Figure 6.11 – *Raw probe intensities for 5 genes considered as DE under BMEA, but with low ranking under PLM. Only genes considered as DE in one T_{reg} Vs T_h comparison are shown. Probes are shown in order of genomic position, separated by donor. Grey rectangles indicate the range of intensities which would be expected for BG signal alone, in 95% of observations. Only cell types relevant to the specific comparison are shown. In sub-figure A, probeset 002 was excluded under DABG, as indicated by the asterisk.*

6.1.3 Selection of Candidate AS Exons

A set of genes was first formed to draw candidate exon-level probesets for AS detection. The average expression level ($\bar{\mu}.$) across all four cell types was calculated using the average of all four posterior means for μ_h, and similarly the average extent of fold-change ($|\overline{\Delta\mu.}|$) was calculated using the absolute value of the posterior means of logFC distributions for each comparison. A list of 2121 candidate genes for AS detection was then defined as those in the upper quartile based on $\bar{\mu}.$ (as per Figure 4.11) and in the lower quartile of $\overline{\Delta\mu.}$. Whilst no impact on AS detection was expected (Figure 4.10B), the lower quartile based on fold-change was added to enable simpler verification via qPCR, as changes in exon proportions will be clearer to detect if not confounded with changes in expression level.

In order to select candidate exons for AS events, an initial list of the best candidate probesets was obtained. A selection criteria for the CPI-LB method of $Z_j > 42$ (i.e. the upper quartile of all Z_j values) along with $\kappa = \log_2(1.8)$ gave a list of 13 probesets from 12 unique genes satisfying these conditions in at least one comparison. In order to better characterise these exons across all comparisons, and in keeping with the experimental design (Figure 2.9), the value of $\kappa = \log_2(1.2)$ was used to consider a probeset as a candidate for an AS event in subsequent comparisons. Probesets were only retained for verification if considered as significant in at least two comparisons, with one from the initial list failing this requirement and reducing candidates to 12 probesets across 11 genes.

Probesets were then placed into one of four basic groupings based on inclusions rates across all comparisons (Figure 6.12): 1) Common $\mathrm{T_{reg}}$ AS events; 2) AS events specific to $\mathrm{T_{reg}}$ Activation; 3) Common Activation AS events, and; 3) AS events specific to $\mathrm{T_h}$ Activation. The trend of lower than expected values for $\hat{\phi}_{hj}$ in stimulated $\mathrm{T_{reg}}$ was noted as the dominant $\mathrm{T_{reg}}$ activation signature, but this was not investigated in detail prior to qPCR. Each probeset was inspected at the probe level in a similar fashion to Appendix C, and transcripts which supported the putative AS event were identified (Table 6.6). It was also noted during inspection than none of the probesets predicted as exclusively reduced in stimulated $\mathrm{T_{reg}}$ mapped to known transcripts or splicing patterns. Of those presented in Figure 6.12, ENSG00000115677_048 (*HDLBP*) was not included for testing via qPCR as this targeted a small region within an exon containing multiple promoters, which was not feasible for primer design. With additional limitations on source biological material, only 8 of the 12 putative AS events were assessed using qPCR, as indicated.

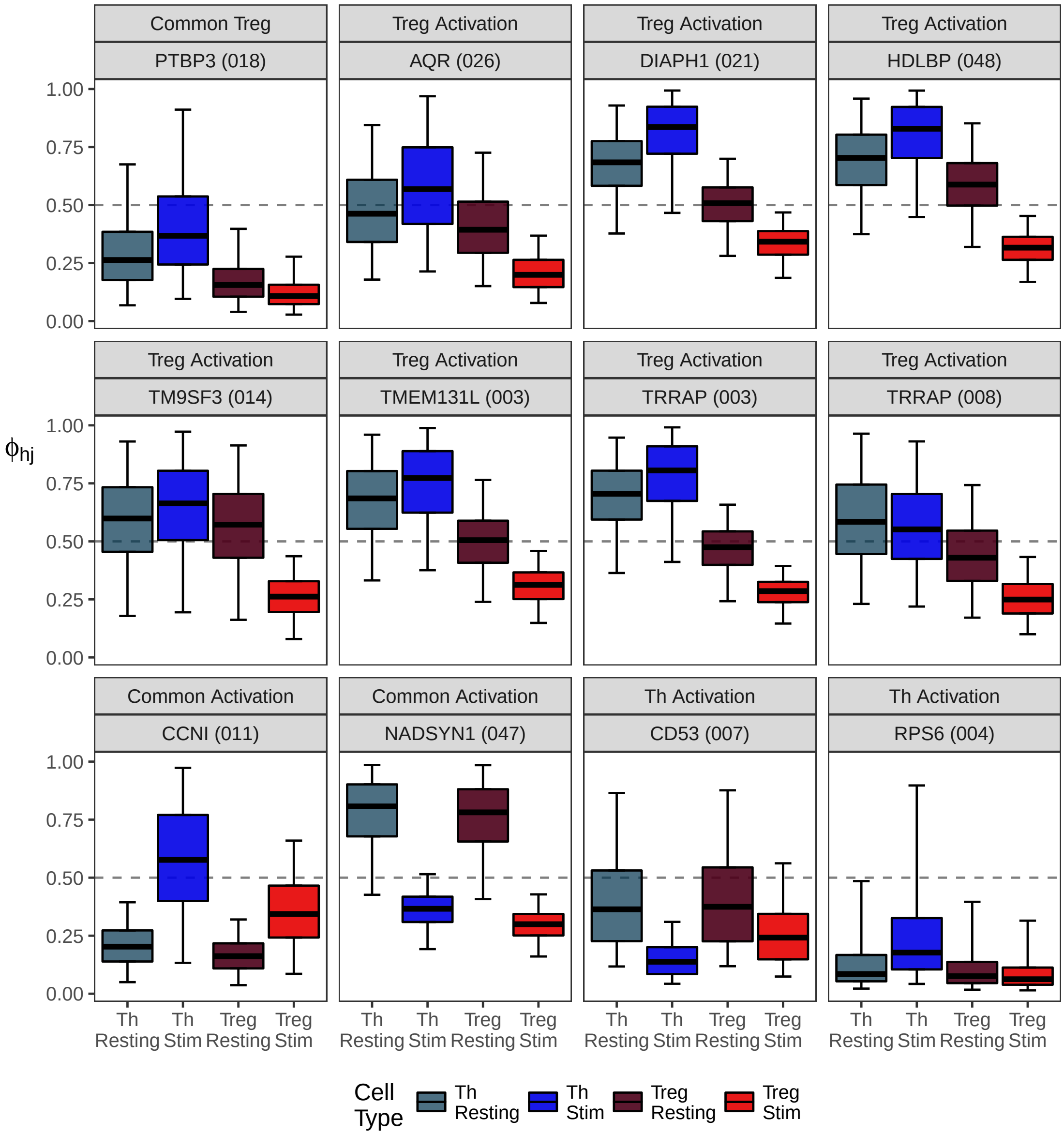


Figure 6.12 – *Boxplots of posterior distributions for ϕ_{hj} in candidate exon-level probesets. Whiskers extend to the 97.5 and 2.5 percentiles, with boxes indicating the IQR.*

Table 6.6 – *Genes and exon-level probesets selected for verification using qPCR. Genomic co-ordinates are given using hg19. No events predicted as the T_{reg} activation signature were able to be explained by combinations of existing transcripts, whilst all others corresponded to known events. Those for which qPCR was performed are indicated in the final column.*

Gene ID	Probeset	Gene	Group	Probeset Region	Known Event	qPCR
ENSG00000119314	018	PTBP3	Common $\mathrm{T_{reg}}$	chr9:115092722-115092754	Cassette Exon	✓
ENSG00000021776	026	AQR	$\mathrm{T_{reg}}$ Activation	chr15:35226773-35226820		✓
ENSG00000077147	014	TM9SF3	$\mathrm{T_{reg}}$ Activation	chr10:98325062-98325168		✓
ENSG00000121210	003	KIAA0922	$\mathrm{T_{reg}}$ Activation	chr4:154471226-154471251		
ENSG00000131504	021	DIAPH1	$\mathrm{T_{reg}}$ Activation	chr5:140955833-140955860		
ENSG00000196367	003	TRRAP	$\mathrm{T_{reg}}$ Activation	chr7:98479599-98479643		✓
ENSG00000196367	008	TRRAP	$\mathrm{T_{reg}}$ Activation	chr7:98493400-98493443		✓
ENSG00000118816	011	CCNI	Common Activation	chr4:77996323-77996416	Alt. Promoter	✓
ENSG00000172890	047	NADSYN1	Common Activation	chr11:71210454-71210599	ENST00000530831	✓
ENSG00000137154	004	RPS6	$\mathrm{T_h}$ Activation	chr9:19376651-19376772	ENST00000498815	
ENSG00000143119	007	CD53	$\mathrm{T_h}$ Activation	chr1:111435158-111435338	ENST00000471220	✓

6.1.4 Verification of Candidate Exons

Primer design, sample preparation and qPCR reactions for verification were performed by Dr Tim Sadlon. In short, primers were designed for each putative AS event as specified in Appendix D. RNA from the original sample denoted as E43 was available and used for this analysis, along with two cord blood samples collected at a similar point in time and frozen for the same duration, giving a total of $n = 3$ for each gene. PCR amplification was performed for 45 cycles using a Rotor-Gene Q, with *RPL13A* used as the housekeeping gene. Triplicate C_t values for each gene were obtained using the Rotor-Gene Q Software v2.0.3.2. The $\Delta\Delta C_t$ method was then used taking the robust mean of triplicate measurements within each sample, using Huber's M-estimator (Huber, 2005). Statistical analysis was performed by fitting the mixed effects model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta_k + \varepsilon$$

where,

- y_{ijk} is the normalised C_t value (ΔC_t) for the gene of interest in cell type i, under treatment j for donor k
- μ represents the baseline ΔC_t value, in this case representing that for resting $\mathrm{T_h}$
- α_i is the change from baseline due to cell type $i = 1, 2$, with $i = 1$ denoting $\mathrm{T_h}$ and α_1 set to 0
- β_j is the change from baseline due to stimulation for $j = 1$, with $j = 1$ indicating resting and β_1 set to 0
- γ_{ij} is an interaction term capturing any change from baseline in Stimulated $\mathrm{T_{reg}}$ not already captured by the previous terms.
- δ_k represents the donor specific change from baseline such that $\delta_k \sim \mathcal{N}(0, \sigma_d)$
- ε indicates a general error term, $\varepsilon \sim \mathcal{N}(0, \sigma)$

The interaction term (γ_{ij}) was never considered significant and was removed from the model for all genes. Results are shown using raw p-values in Figure 6.13 with *CD53*, *CNNI*, *NADSYN1* and *PTBP3* all being supportive of common activation effect. After adjustment using Holm's method, strong support was still found for the predicted common activation effect of *NADSYN1* ($p_{adj} < 0.001$), with *CD53* also remaining significant for a common activation effect, instead of the predicted $\mathrm{T_h}$-specific activation effect. Inspection of posterior distributions in Figure 6.12 suggested this was not an inconsistent result. In the case of *NADSYN1* the transcript which was down-regulated was a poorly characterised non-coding transcript (ENST00000530831), and similarly, the down-regulated *CD53* transcript was also non-coding (ENST00000471220)

Only moderate support ($p_{adj} = 0.09$) was found for the expected common $\mathrm{T_{reg}}$ effect for *PTBP3*, with the unexpected common activation effect remaining significant ($p_{adj} = 0.002$). Given the low power due to the small number of samples, this was still considered as supportive of the initial prediction, with the additional activation effect not being evident in Figure 6.12, and this was specifically absent in $\mathrm{T_{reg}}$. From these results, it appears that the cassette exon is increasingly included on activation, with lower initial inclusion rates in both resting and stimulated $\mathrm{T_{reg}}$.

Results for *CNNI* failed to remain significant after p-value adjustment ($p_{adj} = 0.26$), however the initial results were again considered as supportive based on the direction, and the larger standard errors than for other genes. This finding represents probable alternate promoter usage on activation in both $\mathrm{T_{reg}}$ and $\mathrm{T_h}$.

Predictions were not confirmed for the remaining four events, with all of these being AS events which could not be explained in terms of known transcripts (Table 6.6). Additionally, these were all considered to be exon omissions in stimulated $\mathrm{T_{reg}}$. This set of predictions appeared unusual in comparison to the other predictions and the complete lack of verification for these predictions raises the possibility of some unidentified artefact in the data, which was unique to the stimulated $\mathrm{T_{reg}}$ samples. The lack of an increase for stimulated $\mathrm{T_{reg}}$ in the arrays but which was suggested by qPCR for *PTBP3* and *CNNI*, also follows this pattern of lower estimates in stimulated $\mathrm{T_{reg}}$, indicating these lower estimates may not be representative of the true biology. An investigation of this follows in the subsequent section.

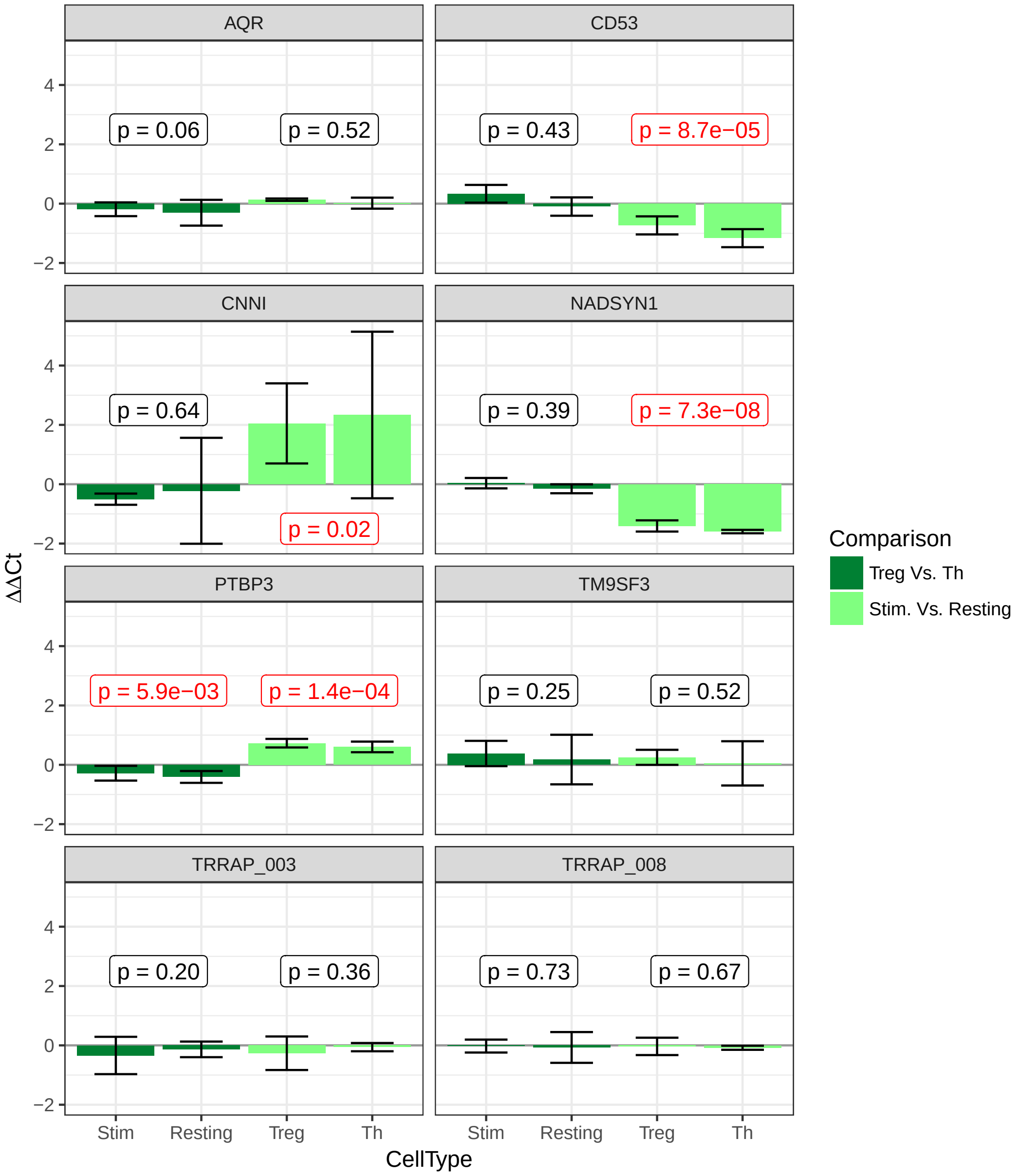


Figure 6.13 – *qPCR Results for candidate AS exons from the T_{reg} dataset. P-values are not adjusted, and those with $p < 0.05$ are shown in red.*

6.2 The Impact of Uneven Hybridisation

6.2.1 Hybridisation Patterns that Correspond to Sample Groups

For the two comparisons involving stimulated $\mathrm{T_{reg}}$, taking only the exons with the most extreme B-statistic would predict 93% (Stimulated $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$) and 83% (Stimulated Vs Resting $\mathrm{T_{reg}}$) of candidate exons as absent/reduced in stimulated $\mathrm{T_{reg}}$. This predicted level of exon omission supported the possibility that there was an undetected artefact in the data, connected to this cell type as potentially identified in Section 6.1.4. Higher than expected levels of exon omission imply a systemic downward bias in estimation of ϕ_{hj}, which would occur if the true background component within a PM intensity is *lower* than that specified by the prior distribution in Equation 3.21. This would occur if the means of the prior distributions (i.e. λ_{il}) were specified at a higher level than is truly contained within the data.

For each background signal bin created across the set of AG probes, prior means for each cell-type are presented as boxplots in Figure 6.14, with values for λ_{il} showing a clear upward bias for stimulated $\mathrm{T_{reg}}$ across bins 9 to 20. These bins represent the probes expected to show the highest level of background signal (Figure 3.12), and correspond to about 60% of PM probes on this set of arrays. This positive bias in expected BG signal was considered as the most likely source of the high-level of spurious exon omissions in this cell type. In order to ascertain whether this behaviour was unique to the $\mathrm{T_{reg}}$ dataset, the values for λ_{il} were also inspected for the Tissue Mixture dataset (Figure 6.15) grouped by mixture level. A similar pattern was again found for Bins 1-11, corresponding to the set of probes expected to contain lower levels of background signal and representing 56.6% of the PM probes in this second dataset. Instead of being restricted to a single mixture level, a very clear gradient was observed with a positive bias tracking with increased amounts of heart tissue in the mixture. A gradual increase in values corresponding to the amount of heart tissue in the mixture was also noted for bins 13-20, although this was not as pronounced as for the lower bins. No plausible biological explanation is apparent, however as these samples represent a serial dilution in technical triplicates, the possibility of a systemic technical source is not able to be excluded. Within the three replicates of the 50:50 mixture a level of variation was also noted for some bins, and this may have been partially responsible for the unexpectedly high-level of false positives detected in Section 5.4.2.

As for the $\mathrm{T_{reg}}$ dataset, this upwards bias in estimates of λ would result in a downward bias in estimates for ϕ_{hj} in the 100% heart mixture samples, and by implication a general increase in the numbers of putative brain-specific exons predicted. This is evident in Figure 5.9 and Table 5.8. However, this gradient will have additionally influenced the results from Section 5.4.5, as this behaviour tracks with mixture levels and will have biased the slopes fitted through the varying mixture levels. This additionally explains the observation in Appendix C that some putative exons were less easily explained biologically, and considered inconclusive. No immediate resolution to this issue is possible, however the good levels of confirmation via plausible transcripts for the majority of exons do suggest that the influence of this artefact may be less than feared.

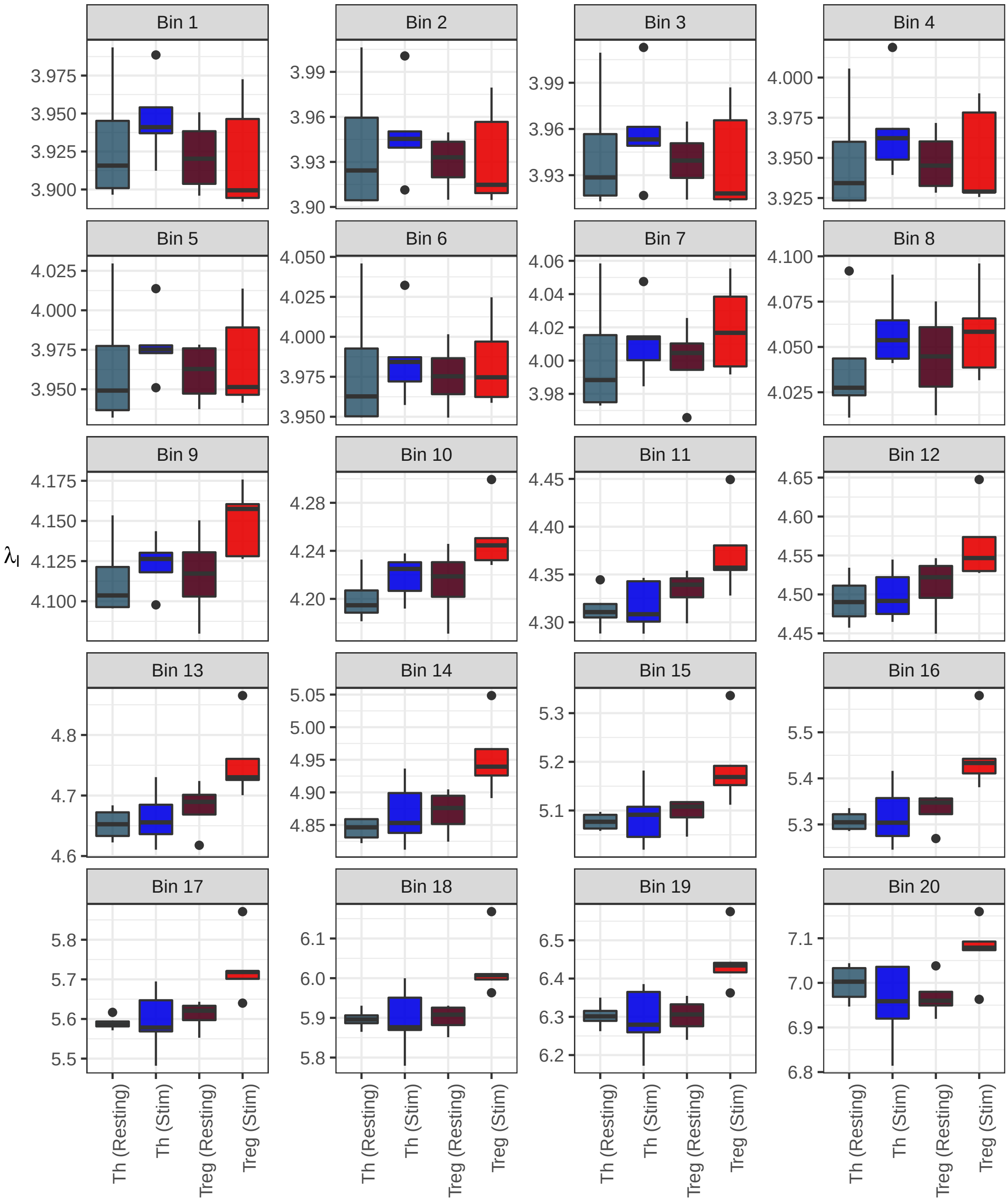


Figure 6.14 – *Boxplots showing means of the distributions specified for each bin of the background signal component of the BMEA model (λ_{il}),for the T_{reg} dataset. Values are grouped by cell type, with increased values for Stimulated T_{reg} evident across bins 9-20, indicating higher expected amounts of non-specific binding for these samples.*

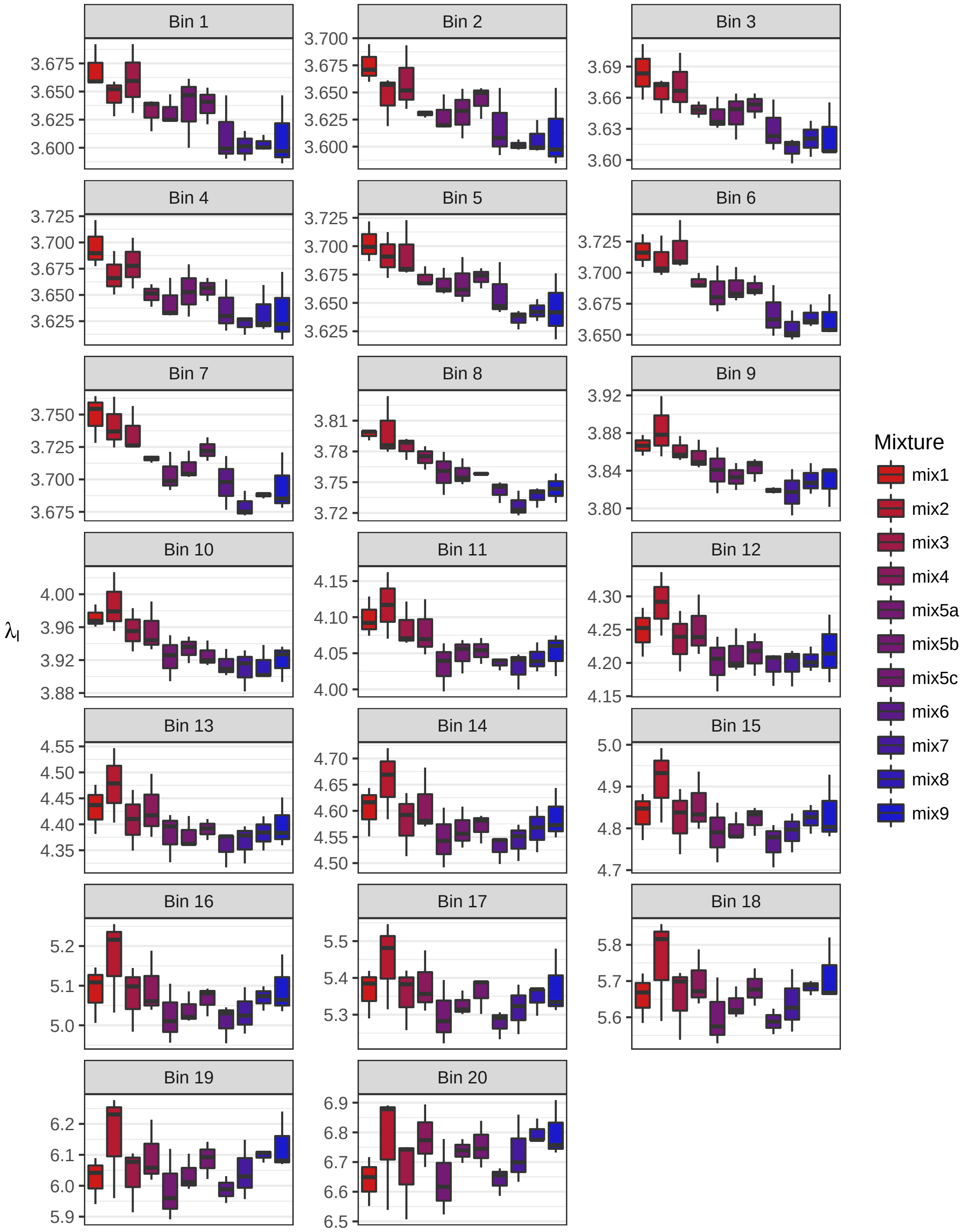


Figure 6.15 – *Boxplots showing means of the distributions specified for each bin of the background signal component of the BMEA model (λ_{il}) for the Tissue Mixture dataset. Values are grouped by mixture level, with increased values tracking with increased proportions of heart tissue in the mixture evident across bins 1-11, indicating higher expected amounts of non-specific binding for these samples.*

6.2.2 Quantile Normalisation for BMEA

Because of the need to assign priors for background signal, the entire set of probes (i.e. PM + AG or MM) had been quantile normalised prior to analysis. The post-normalised distributions of PM and AG probes are shown separately for the $\mathrm{T_{reg}}$ dataset in Figure 6.16, with a subtle increase in the AG probe density clearly observable between the values 8 and 10 for stimulated $\mathrm{T_{reg}}$ samples (Figure 6.16A; dashed, red lines). A corresponding decrease is also able to be seen in the PM probe density for these samples between the values 8 and 12 (Figure 6.16B). Although the combined intensity distributions for all probes are identical by definition, this internal variation within probe types, within samples is in keeping with the observations in the previous section.

In order to ascertain whether this was an artefact of quantile normalisation, or was inherent in the data within each array, probes at the 25, 50 and 75th intensity percentiles were selected from the PM and both the AG and MM (i.e. genomic) background probes *before* quantile normalisation. The $\log_2$ ratio of PM probes to both sets of BG probes was then calculated within each sample at each percentile (Figure 6.17) for both the $\mathrm{T_{reg}}$ and Tissue Mixture datasets. Patterns seen in the previous section were repeated, with lower PM/BG ratios observed within both sets of BG probes for the sample groups corresponding to the previous patterns of inflated estimates for λ_{il}, implying that intensities for BG probes were far higher in these samples than for the other samples. As this ratio was internally calculated with no background correction or normalisation, this confirmed that these samples contained an inherent structure within the probe intensities which lead to inflated estimates for λ_{il} for these sample groups.

No convincing biological explanation is able to be offered for this observation, as AG probes specifically target no known genomic sequences. A technical source for these observations is the most realistic alternative, with different stages of the protocol such as washing steps, hybridisation times being the most likely source. This does raise the possibility that some steps may have been performed using arrays grouped together by cell type or mixture level. By way of example, it is not difficult to imagine a researcher systematically preparing arrays for all pure brain samples, followed by all 95:05 mixtures, and so on, in order to successfully keep track of their progress. This would lead to small variations in the samples which were systemic in nature and may produce the results observed here. Unfortunately this is not able to be confirmed, but if true, would only reinforce the importance of randomisation during *all* experimental procedures, as has been long advocated by the statistical community.

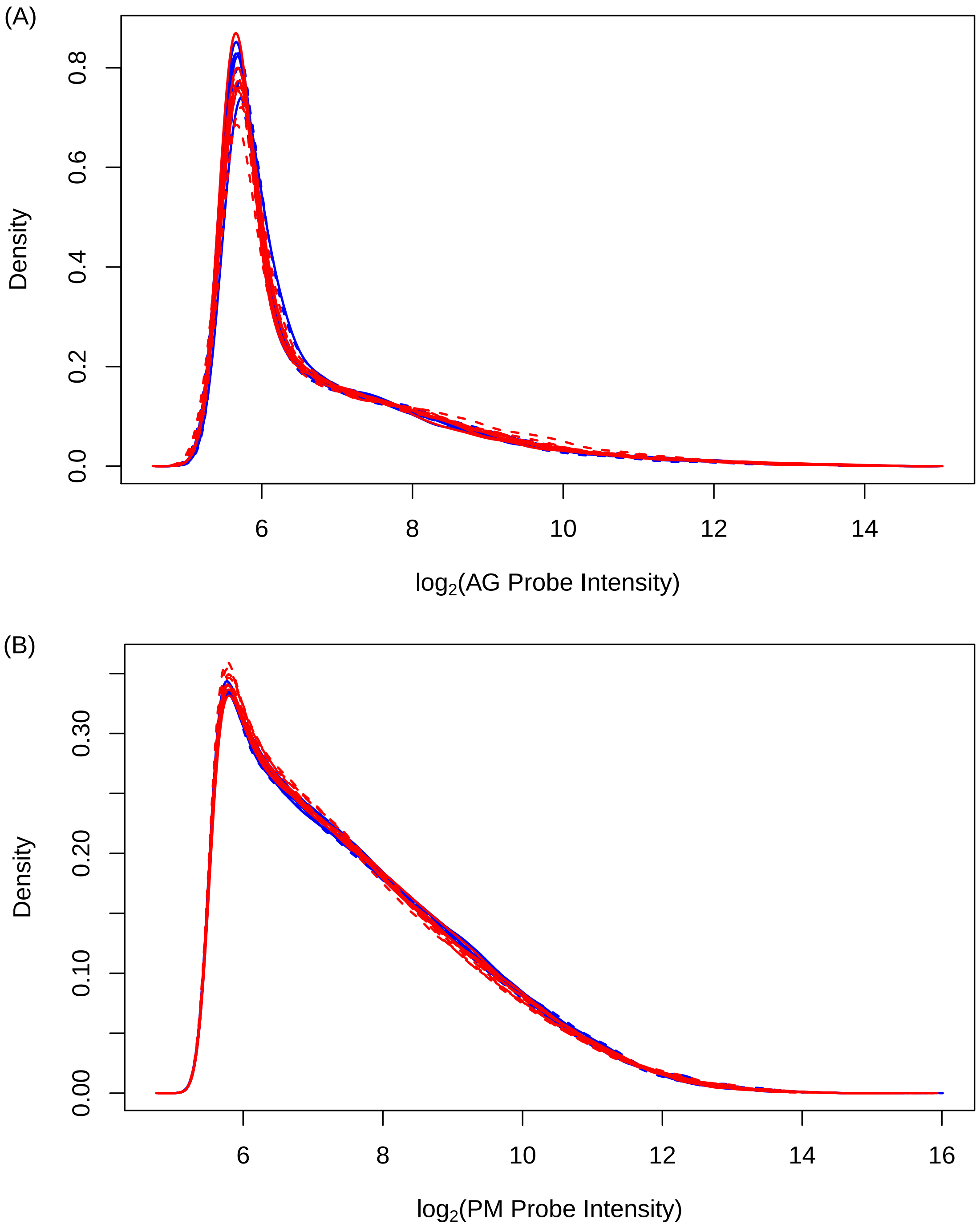
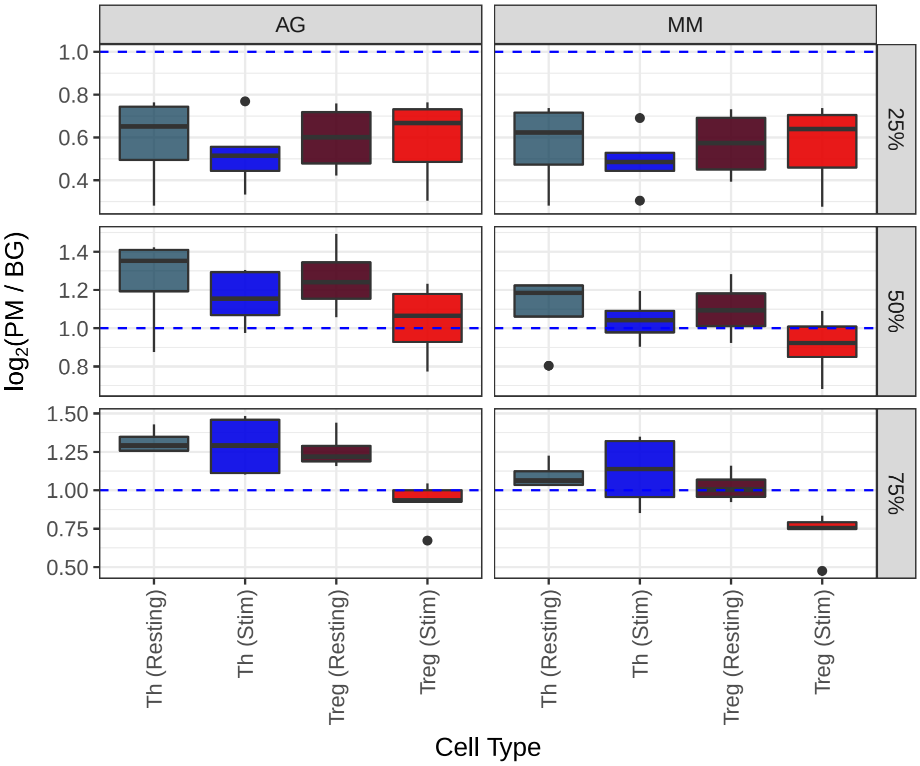
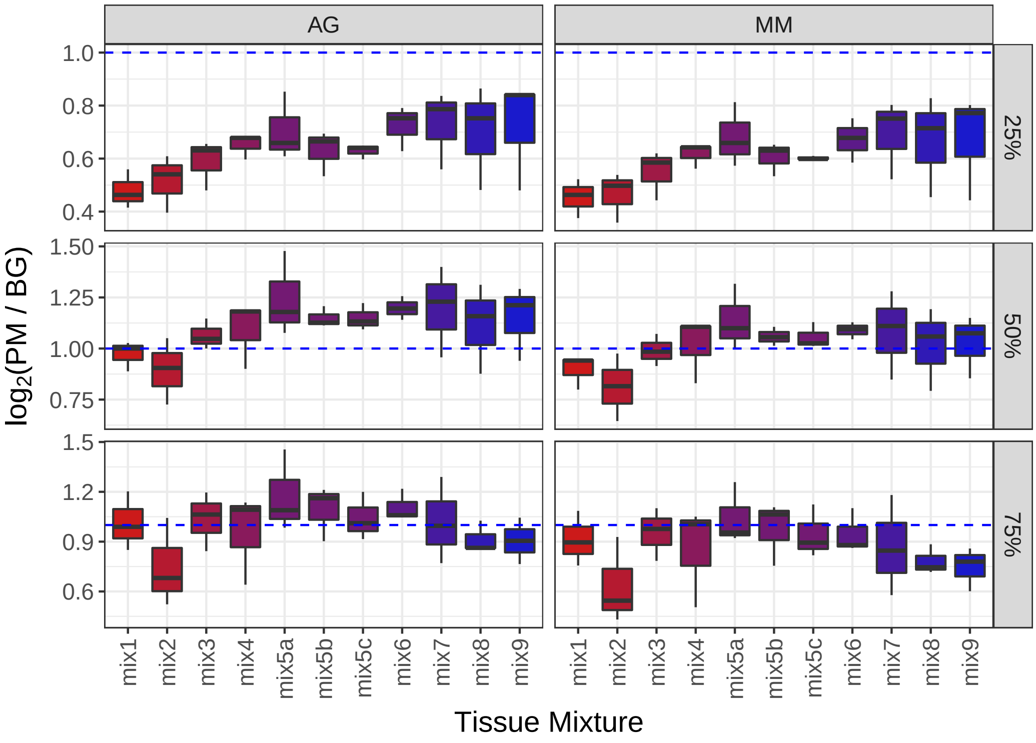


Figure 6.16 – *Probe intensities separated into A) Background (antigenomic) and B) Perfect Match (PM). T_{reg} samples are shown in red,with T_h samples shown in blue. Stimulated samples are shown with dashed lines. The right skew for anti-genomic probes in Stimulated T_{reg} is balanced by the left skew in the PM probe intensities for Stimulated T_{reg}.*



(A) *T_{reg} Dataset*



(B) *Tissue Mixture Dataset*

Figure 6.17 – *Comparison of BG and PM probe intensities at each quartile boundary. Values are the* $\log_2$ *ratios of PM/AG or PM/MM probes at the 25, 50 and 75*th *percentiles within each set of probes.*

6.3 Final BMEA Analysis of the $\mathrm{T_{reg}}$ Dataset

6.3.1 Detection of AS Events

After identification of the technical issues underlying AS detection within Stimulated $\mathrm{T_{reg}}$ arrays, the process of Section 5.4.4 was repeated in a more inclusive fashion. The same set of candidate genes was used, but the Z_j threshold for inclusion was lowered to 16.6 as the 50^{th} percentile of the distribution, following the initial analysis of the Tissue Mixture dataset. The exclusion region ($[-\kappa, \kappa]$) for CPI-based candidate selection was set to $\kappa = \log(1.5)$ for those with Z_j in the upper quartile, and $\kappa = \log(2)$ for the remainder of probesets (Figure 6.18).

After this initial selection step, candidate AS events were checked in all four comparisons using the value $\kappa = \log(1.2)$ to define a putative AS event in any subsequent comparisons. Probesets were only retained if detecting an AS event across at least two comparisons, in keeping with Section 2.3.3. Potential AS events which indicated exon omission exclusively in stimulated $\mathrm{T_{reg}}$ were removed from the list of candidates. Using the posterior median as a point estimate ($\hat{\phi}_{hj}$), remaining probesets were retained only if $\hat{\phi}_{hj} > 0.25$ in at least one comparison, ensuring probesets were only included for exons which were relatively abundant in at least one cell type. This gave a total of 20 exon-level probesets across 18 genes with candidate AS events (Table 6.7).

Each probeset was checked for a plausible explanation given the known transcripts. The AS event from *TUBA3C* didn't match any known alternate transcripts, and the gene itself is not predicted to have any alternate isoforms. This probeset contained a single probe with $Z_j = 18.8$ being near the 50^{th} percentile of all Z_j scores ($Z_{0.5} = 16.6$) and was excluded from further consideration. Probe intensities from the four probesets with the lowest Z_j scores were manually inspected (Figure 6.19) with probes within probeset 061 from *HDLBP* appearing to contain mostly background signal. No supporting evidence was observed in donor E43 with only one or two probes from the remaining donors being weakly supportive and this was also excluded as a candidate. Intensity patterns amongst the remaining probesets were considered supportive and these were retained.

Candidate probesets were then assigned to groups (Figure 6.20) based on inclusions patterns: 1) $\mathrm{T_h}$ Resting Inclusion; 2) $\mathrm{T_h}$ Activation Response; 3) Complex patterns across cell types; 4) Common to either $\mathrm{T_h}$ or $\mathrm{T_{reg}}$; and 5) Common Activation response in both $\mathrm{T_h}$ and $\mathrm{T_{reg}}$. The additional inclusion of the $\mathrm{T_h}$ activation response gave a degree of confidence to exons with putative omission as a $\mathrm{T_{reg}}$ activation response and a degree of useful information was able to be gained from the Stimulated $\mathrm{T_{reg}}$ samples, with some probesets still indicating *increased* exon-level inclusion rate in Stimulated $\mathrm{T_{reg}}$. AS events already confirmed via qPCR (Section 6.1.4) were manually re-assigned to the correct group as this was considered known biology.

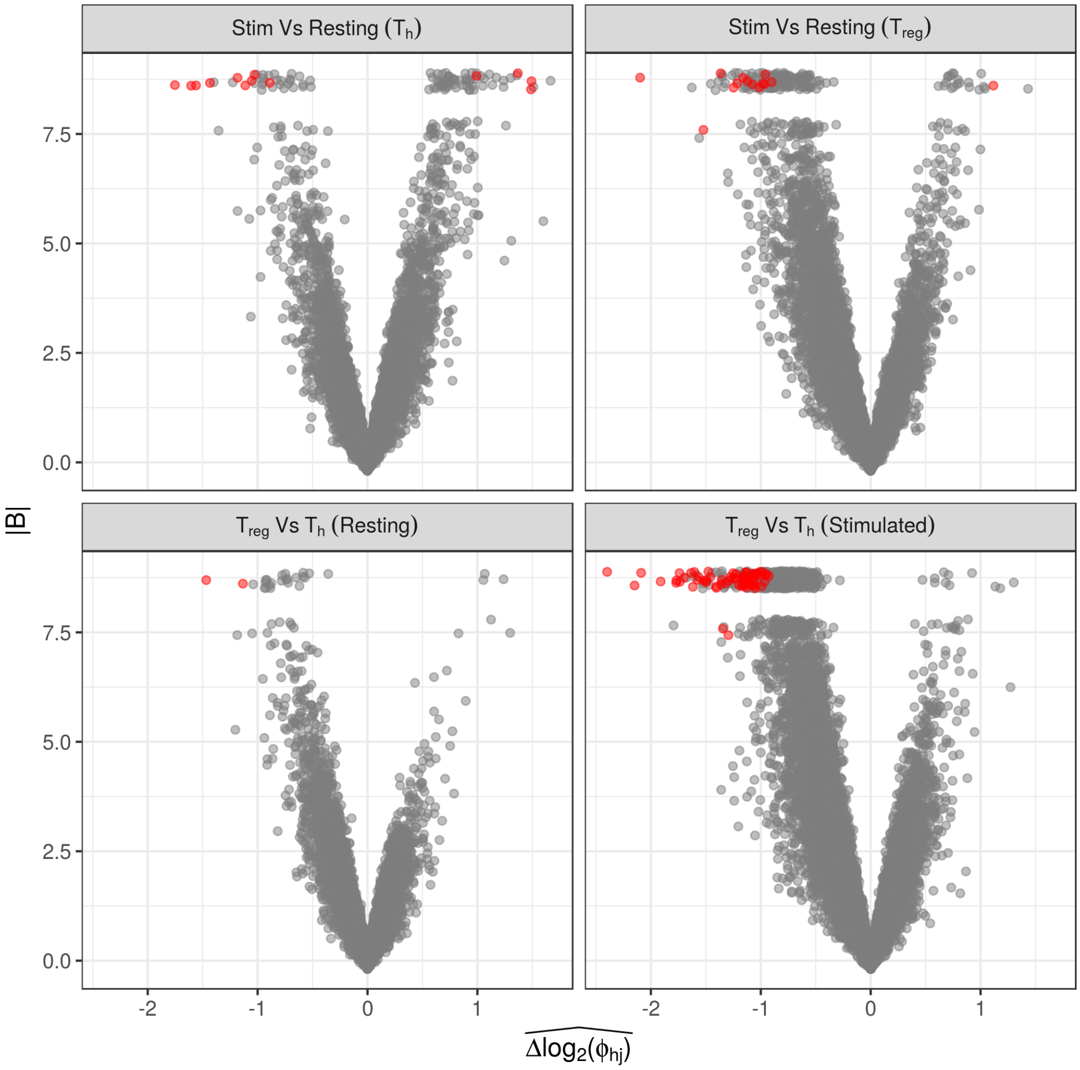


Figure 6.18 – *Candidate exons for the final list of AS events in the T_{reg} dataset. Those passing the initial selection criteria in each comparison are shown in red. The noted downwards bias in comparisons involving Stimulated T_{reg} are clear in the two right-most panels. Jitter has been added on the y-axis.*

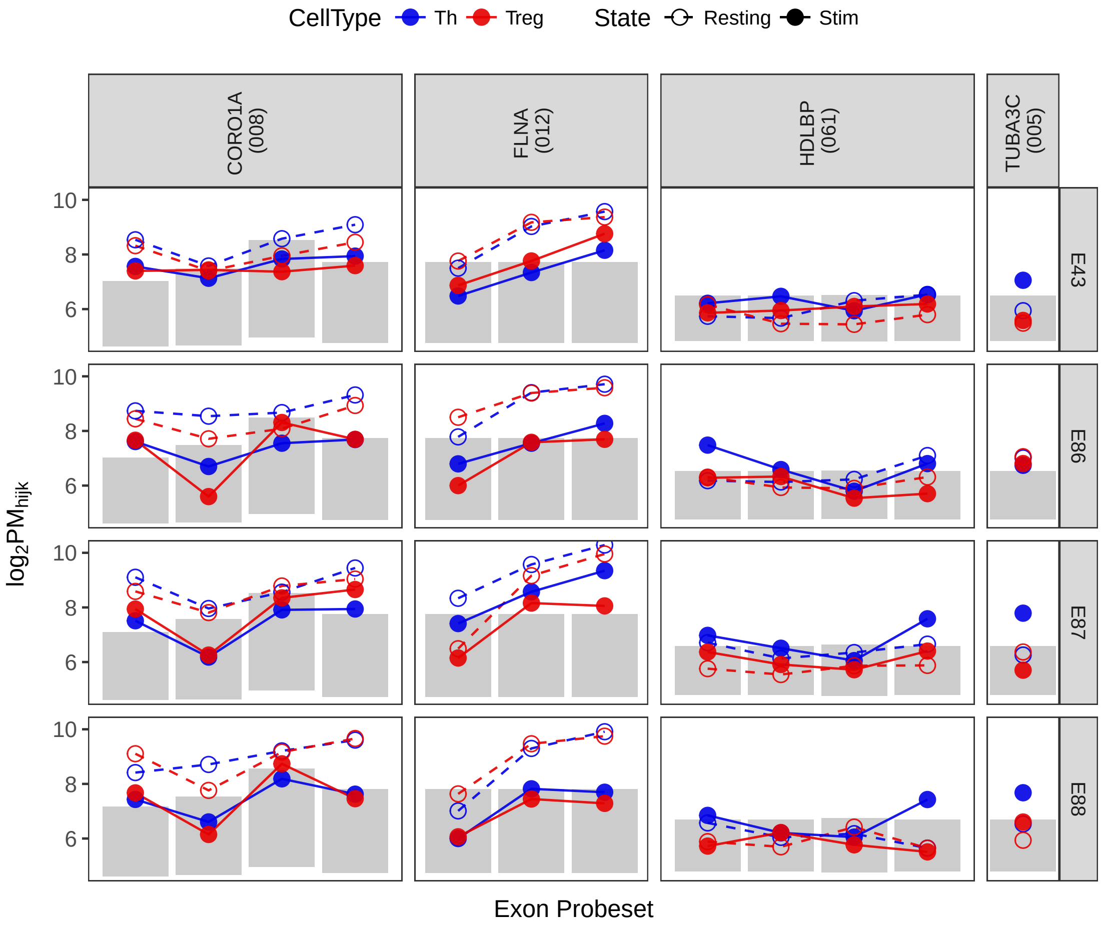


Figure 6.19 – *Probe intensities for the four candidate exons with the lowest Z_j scores. Expected BG signal is indicated in grey as $\lambda_{\cdot ijk} + 2\delta_{\cdot ijk}$ averaged within each donor across all cell types. Donor E41 was omitted due to the lack of stimulated samples.*

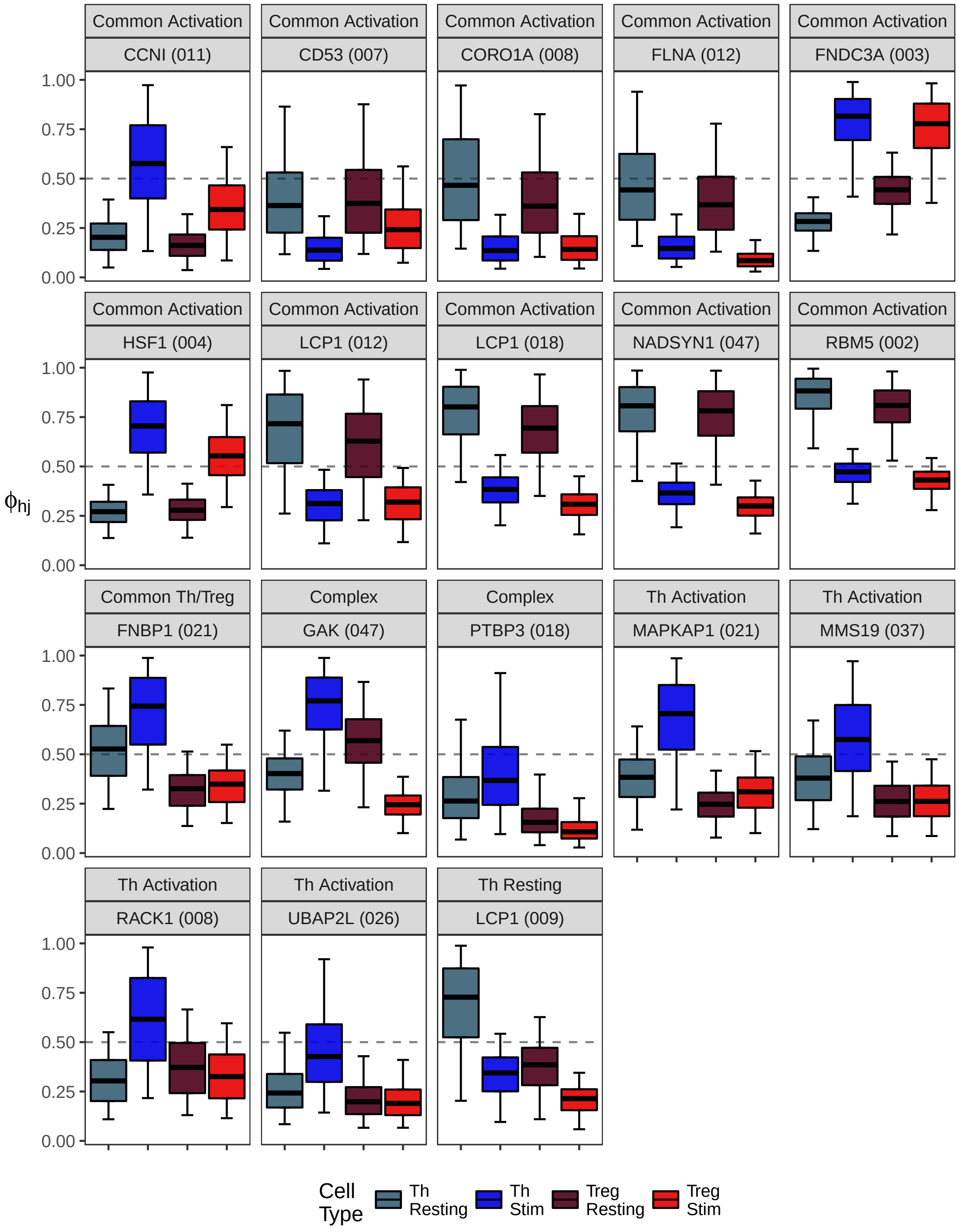


Figure 6.20 – *Posterior distributions for candidate probesets in the final list of candidate AS events in the T_{reg} dataset. Groups defined based on inclusion patterns are indicated in the panel strips. All probesets were considered as a candidate in at least 2 comparisons. AS events verified by qPCR in Section 6.1.4 have been reclassified manually into the correct groups Probesets within each gene are indicates in brackets.*

Table 6.7 – *Description of final AS events from the T_{reg} dataset. Exon-level probesets are given in brackets after the Ensembl ID. AS events with multiple possible transcripts are shown in italics. Transcript biotypes are denoted as PC (Protein Coding); NMD (Nonsense Mediated Decay); RI (retained intron) and PT (Processed Transcript). Events for CCNI, NADSYN1 and PTBP3 were confirmed in Figure 6.13. The AS event for CD53 was confirmed as down in both Stimulated cells, as opposed to the prediction of down in T_h Stim only. All events confirmed using qPCR are indicated with an asterisk.*

Ensembl ID	**Gene**	**Key Transcript**	**Direction**
ENSG00000003756 (002)	*RBM5*	ENST00000469838 (PC)	↓ Stim
ENSG00000102531 (003)	*FNDC3A*	ENST00000484074 (NMD)	↑ Stim
ENSG00000102879 (008)	*CORO1A*	*ENST00000567034 (RI)* *ENST00000564768 (RI)*	↓ Stim
ENSG00000143119 (007)	*CD53**	ENST00000471220 (PT)	↓ Stim
ENSG00000118816 (011)	*CCNI**	ENST00000513774 (PC)	↑ Stim
ENSG00000136167 (012) (018)	*LCP1*	ENST00000494531 (PT) ENST00000460190 (PT)	↓ Stim ↓ Stim
ENSG00000172890 (047)	*NADSYN1**	ENST00000530831 (RI)	↓ Stim
ENSG00000185122 (004)	*HSF1*	ENST00000528988 (NMD)	↑ Stim
ENSG00000196924 (012)	*FLNA*	ENST00000498491 (PT)	↓ Stim
ENSG00000187239 (021)	*FNBP1*	ENST00000491905 (PT)	↓ T_{reg}
ENSG00000178950 (047)	*GAK*	*ENST00000511983 (ncRNA)* *ENST00000505819 (NMD)*	↑ T_h Stim ↓ T_{reg} Stim
ENSG00000119314 (018)	*PTBP3**	Cassette Exon *(Multiple PC transcripts)*	↑ T_h ↑ Stim
ENSG00000119487 (021)	*MAPKAP1*	ENST00000468896 (PC)	↑ T_h Stim
ENSG00000143569 (026)	*UBAP2L*	ENST00000489076 (PT)	↑ T_h Stim
ENSG00000155229 (037)	*MMS19*	Cassette Exon *(Multiple PC transcripts)*	↑ T_h Stim
ENSG00000204628 (008)	*RACK1*	ENST00000514183 (RI)	↑ T_h Stim
ENSG00000136167 (009)	*LCP1*	ENST00000494531 (PT)	↑ T_h Resting

6.3.2 Biological Implications

If the changes in transcript expression from Table 6.7 were random, no enrichment for biological process would be expected to be detected. Tests for enrichment of GO terms were performed for these 16 genes, using the subset of GO terms annotated to these genes, and excluding those with fewer than 3 steps back to the ontology root nodes. Taking the initial list of 2121 candidate genes as the reference set, enrichment was tested using Fisher's Exact test (Table 6.8), with terms containing only one gene omitted from the results.

Although only a small number of genes were in the test set, 3 of these (*CORO1A*, *LCP1* and *RACK1*) were assigned to the term GO:0001891 (phagocytic cup), with only four being in the reference set of genes. Even under Holm's adjustment ($p = 0.001$) this was considered statistically significant. GO:0051764 (actin crosslink formation) also retained significance under Holm's method ($p = 0.033$), with both *LCP1* and *FLNA* being associated with this term. Both of these genes, with the addition of *CORO1A* and *MAPKAP1* also make up the genes associated with the remaining actin-associated GO terms. Given the known connection between actin filaments and the phagocytic cup (Gerisch, 2010) this represents $> 30\%$ of the genes with putative AS events mapping to this biological behaviour. Additionally, all of these candidate AS events, with the exception of *MAPKAP1*, represent differential regulation of non-coding transcripts, as either retained introns or the undefined "processed transcripts".

The immunological synapse (GO:0001772) which forms between T cells and an antigen presenting cell (APC), has been described as "*reminiscent of a frustrated phagocytosis*" (Alarcón & Martínez-Martín, 2012). Both *CORO1A* and *CD53* are known to play a role in this process reinforcing this possibility, providing a high-quality set of candidate genes and processes for further research, either through computational, or laboratory-based approaches. Given the additional presence of GO:1990778 in the list (protein localization to cell periphery; *FLNA*, *GAK* and *RACK1)*, which may indicate inter-cellular signalling, it is easy to speculate about a wide ranging and interconnected set of cellular responses which are driven by changes in which specific transcripts are expressed, as opposed to an overall change in expression level, so commonly referred to as differential expression. The same three genes are also those associated with GO:0072659 (protein localization to plasma membrane) and GO:0007009 (plasma membrane organization). All genes associated with these processes

Table 6.8 – *The most enriched GO terms within the set of 16 genes with candidate AS events. The number of genes mapped to the term in the complete list is given in the column N, whilst the number in the list of AS genes follows. The Benjamini-Hochberg FDR is shown with all terms being within an FDR* < 0.05. *Terms with only one gene were not included.*

GO ID	Term	Ontology	N	AS	p	FDR
GO:0001891	phagocytic cup	CC	4	3	2.26e-06	0.00
GO:0051764	actin crosslink formation	BP	2	2	7.34e-05	0.02
GO:0005884	actin filament	CC	12	3	1.19e-04	0.02
GO:0051015	actin filament binding	MF	19	3	5.05e-04	0.05
GO:1901215	negative regulation of neuron death	BP	23	3	9.03e-04	0.05
GO:0030036	actin cytoskeleton organization	BP	54	4	9.90e-04	0.05
GO:0099106	ion channel regulator activity	MF	6	2	1.08e-03	0.05
GO:0098609	cell-cell adhesion	BP	56	4	1.14e-03	0.05
GO:0072659	protein localization to plasma membrane	BP	26	3	1.30e-03	0.05
GO:0051130	positive regulation of cellular component organization	BP	154	6	1.35e-03	0.05
GO:0009299	mRNA transcription	BP	7	2	1.50e-03	0.05
GO:0048872	homeostasis of number of cells	BP	28	3	1.63e-03	0.05
GO:1990778	protein localization to cell periphery	BP	28	3	1.63e-03	0.05
GO:0005938	cell cortex	CC	29	3	1.80e-03	0.05
GO:0001772	immunological synapse	CC	8	2	1.99e-03	0.05
GO:0007009	plasma membrane organization	BP	30	3	1.99e-03	0.05

(*CORO1A, CD53, FLNA, GAK* and *RACK1*) specifically involve changes in non-coding transcripts and it is again, very easy to hypothesise that non-coding transcripts are playing an active role in regulation of these processes. However, this remains largely speculative at this point, as defining the biological roles of these transcripts is very much a field in it's infancy (Schmitz et al., 2017).

Importantly, the expected transcript for *CD53* (ENST00000471220) was verified using qPCR (Section 6.1.4), reinforcing that these changes in transcript expression patterns do have independent support in at least one instance. Of the other verified alternate transcript usage, no enrichment for GO terms was detected. However *CD53* and *PTPB3*, along with the predicted *CORO1A, LCP1, MAPKAP1*, mapped to GO terms associated with the immune response again reinforcing the plausibility of the above predictions.

6.3.3 Alternate Splicing For *FOXP3*

As the transcript usage for *FOXP3* was of interest for this work, any candidate AS events or changes in transcript proportions were also explored, despite the exclusion of this gene from the initial list of candidate genes based on significant logFC. Using the above inclusion criteria of $Z_j > 16.6$ and a CPI-LB $> \log_2(1.5)$, probeset 015 would potentially considered a candidate in both $\mathrm{T_{reg}}$ vs $\mathrm{T_h}$ comparisons (Figure 6.21). Similarly probeset 006 passed these criteria in simulated $\mathrm{T_{reg}}$ vs $\mathrm{T_h}$, and would have been considered a candidate using $\kappa = \log_2(1.2)$ in the secondary resting $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison. Whilst the $\Delta \log_2 \phi$ posterior distribution for probeset 003 was a considerable distance from zero, this probeset received $Z_j = 13.5$ and was not a viable candidate.

An AS event for probeset 015 would implicate alternate promoter usage in $\mathrm{T_h}$, although clearly identifying transcripts which this would correspond to was not clear, as this implies that one of the *FOXP3* $\Delta 2$ isoforms (ENST00000376197) would be expressed in preference to full length transcripts. Raw probe intensities were inspected (Figure 6.22) and intensities for this probeset in $\mathrm{T_h}$ samples appeared consistent with other probesets, and if anything appeared to be detecting more signal than some other probesets. This potential AS event was thus considered to be a spurious detection due to the large fold-change for *FOXP3* between $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ cells. No qPCR was performed on this potential AS event.

In addition to the above, probeset 006 would have also passed the initial inclusion criteria ($Z_j = 35.3$) indicating decreased usage in $\mathrm{T_{reg}}$. However, this did not map to any known transcript, and probe intensities did not support this prediction (Figure 6.22) with signal being clearly detected in both $\mathrm{T_{reg}}$ cell-types. The observations here are readily explained by changes in the expression of *FOXP3* between the two cell types, with the lower levels of *FOXP3* in $\mathrm{T_h}$ leading to higher estimates of ϕ_{hj} for this cell type. No qPCR was performed on this candidate AS event as this too was not considered plausible. As such it was considered that beyond the initial detection of fold-change between $\mathrm{T_{reg}}$ and $\mathrm{T_h}$, no changes in transcript usage for *FOXP3* were able to be detected.

FOXP3 is one of the most significantly DE genes in most comparisons between $\mathrm{T_{reg}}$ and $\mathrm{T_h}$, and the interpretations above suggest that in this case of highly significant fold-change, BMEA has difficulty accurately capturing the true biology, and reinforces the restriction of candidate exons to those genes with minimal fold-change between cell-types, as was used in previous sections.

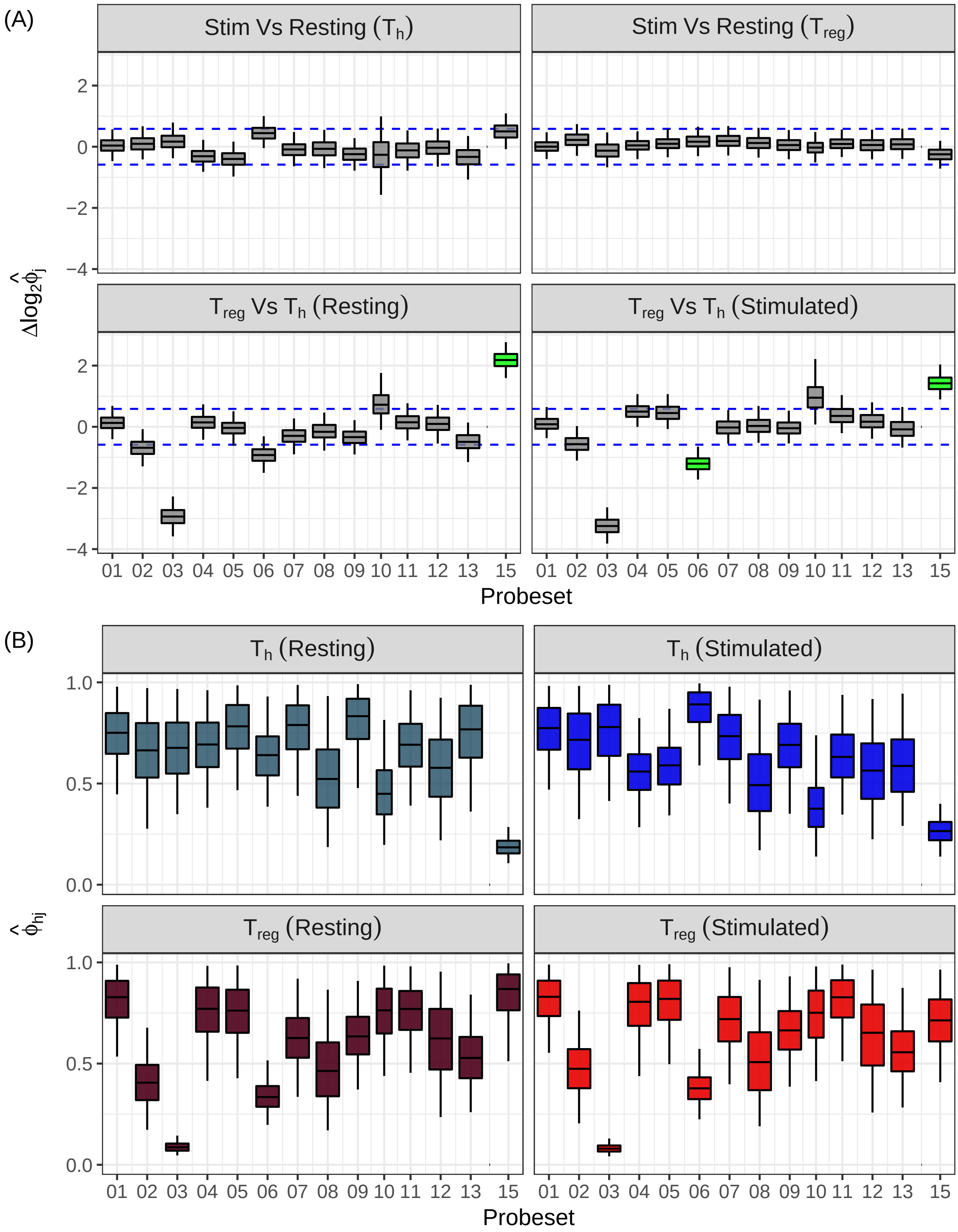


Figure 6.21 – *Posterior distributions for A)* $\Delta\log_2\hat{\phi}_j$ *and B)* $\hat{\phi}_{hj}$ *for FOXP3 . Probesets which were candidates for an AS event in any comparison are highlighted in green (A), with blue horizontal lines indicating* $\kappa = \log_2(1.5)$*. Probeset 015 corresponds to the most commonly used promoter for FOXP3. Probeset 012 corresponds to the exon skipped in the* $\Delta 2$ *transcript. Probeset 14 was not detectable above background.*

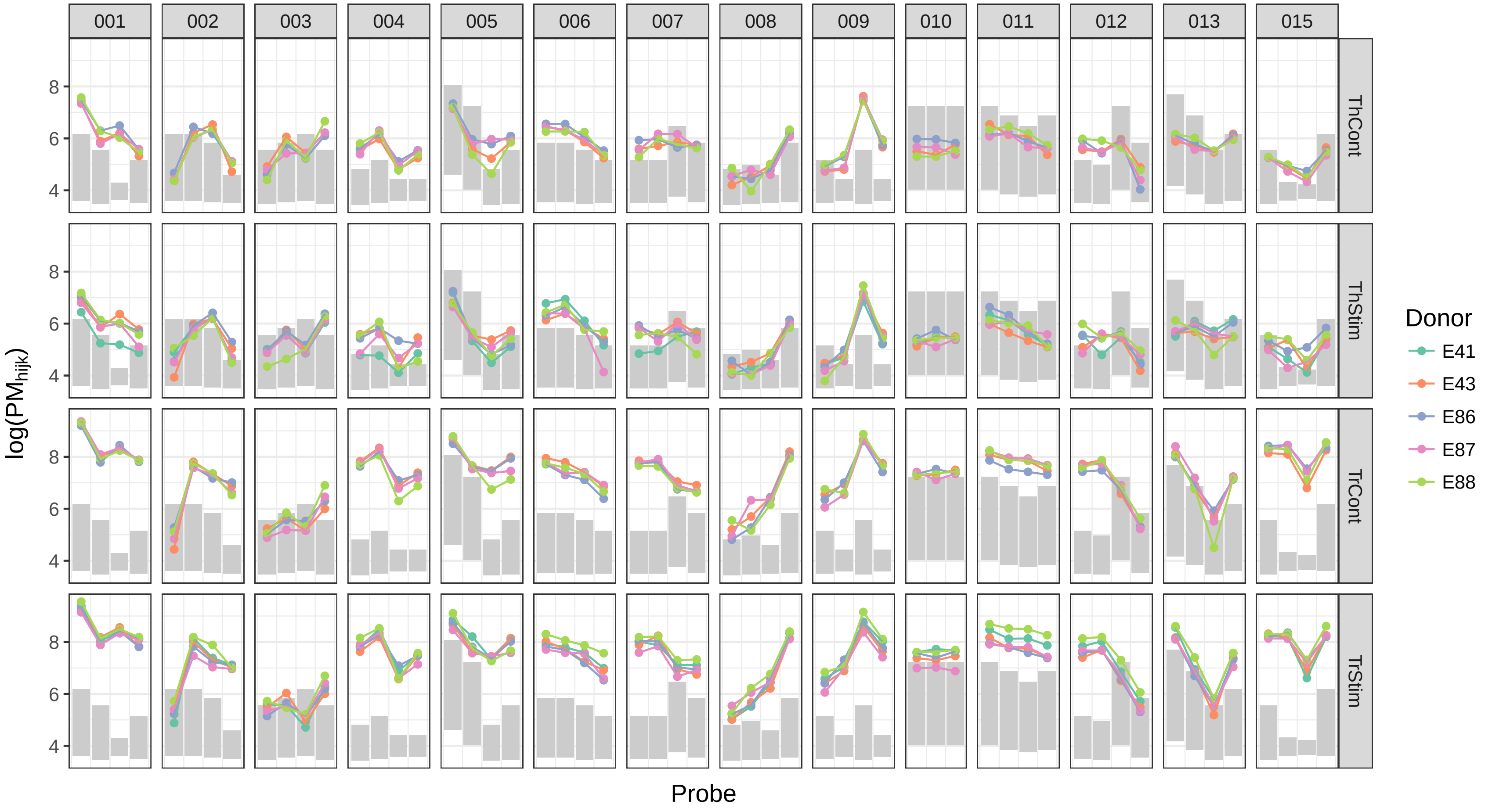


Figure 6.22 – *Probe intensities for FOXP3 shown by cell type. Probes and probesets are shown along the x-axis corresponding to increasing position along the X chromosome. Probeset 015 corresponds to the most commonly used promoter for FOXP3. Probeset 012 corresponds to the exon skipped in the* $\Delta 2$ *transcript. The expected range for background signal is shown as grey rectangles using* $\lambda_{h \cdot jk} \pm 2 * \delta_{h \cdot jk}$*, The fold-change known for FOXP3 is clearly evident with nearly all probes in* T_{reg} *samples being beyond the range of BG signal, but with many probes in* T_h *samples falling with this range.*

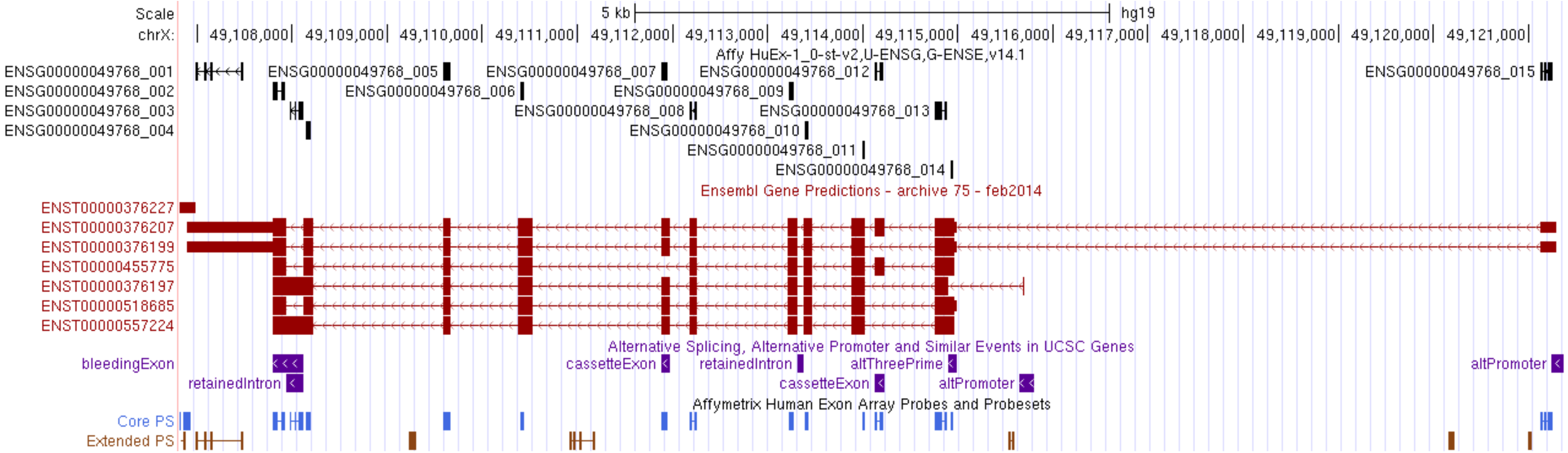


Figure 6.23 – *Probesets for FOXP3 as shown on the UCSC browser. Probeset 012 corresponds to the commonly spliced exon separating the full-length and* $\Delta 2$ *transcripts. Probeset 014 was undetectable.*

6.4 Discussion

The final analysis of the $\mathrm{T_{reg}}$ dataset using BMEA was able to overcome a considerable structural flaw, and the parameterisation of the BMEA model itself was specific enough to identify this artefact, allowing analysis to proceed with requisite caution. The presence of a similar bias in the Tissue Mixture dataset was unexpected, but reinforces the difficulty presented by exon-level analysis of Exon Array data. As seen for *CD53*, the predicted groupings may have missed vital information as a result of the unusual hybridisation observed in Stimulated $\mathrm{T_{reg}}$, however, the use of two comparisons across the four provides a small degree of checks and balances for candidate selection.

A conservative approach was taken for selection of candidate genes and subsequent AS exons, which limited the power of the approach for this dataset. Whilst not revealing a list of thousands of candidate AS events, the candidate AS events can be expected to provide more biologically informative results than the original FIRMA results obtained in Section 2.5, which mostly corresponded to detection associated with stimulated $\mathrm{T_{reg}}$, and would clearly consist of large numbers of erroneous predictions. The Tissue Mixture dataset (Section 5.4) yielded hundreds of candidate AS events, which given the disparate nature of the two tissues is not surprising. Conversely, the $\mathrm{T_{reg}}$ dataset contains four highly-related cell types and the numbers of detected events is likely to be considerably smaller.

Importantly, however, the recovery of biological behaviours specifically associated with stimulation of T cells (Section 6.3.1), even amongst such a small set of AS events, supports the idea that AS detection under BMEA has captured genuine biological signal, and has shed preliminary light into as yet unknown regulatory processes in T cell biology.

The unconfirmed results for alternate transcript usage within *FOXP3* supported that the restriction of candidate AS events to genes in the lowest percentile for fold-change was a prudent strategy. Whilst this analysis does not eliminate the possibility of alternate transcript usage for *FOXP3* , it remains beyond the capacity of this approach to confidently detect any changes within the set of transcripts know to be expressed in both $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ cells.

Chapter 7

Conclusion

7.1 The Wider Context

By the arrival of Whole Transcript (WT) Arrays as a research platform, detection of differentially expressed genes via 3' arrays was a relatively mature field. Platforms such as Exon Arrays promised to target more genes and by capturing information across the length of the gene, information about transcript-specific expression seemed viable. However, as seen in this study, reliably obtaining this level of information presented many challenges.

The difficulties underlying transcript identification is not restricted to WT Arrays, but has also been a significant challenge for RNA-Seq analysis. Only recently have approaches such as `kallisto` (Bray et al., 2016) begun to produce consistent results, although issues such as RNA degradation, GC-bias (Love et al., 2016), length-bias, sequencing depth and complex expression patterns still leave the most confident calls to more highly expressed genes (Zhang et al., 2017). RNA-Seq brings the advantage of direct target measurement from the obtained data, as opposed to array technology which measures both direct target and off-target molecules via non-specific binding, and is largely restricted to predefined gene models. RNA-Seq also enables use of tools such as StringTie (Pertea et al., 2015) for construction of previously unidentified transcripts, however these are also dependent on the choice of aligners and their ability to produce accurate alignments, particularly across gene families with high levels of homology. Moreover, the meaning of differential expression itself, and how this is measured, has been under review since the advent of these technologies (Trapnell et al., 2012).

At the commencement of this body of work, these techniques and tools were a long way off, and RNA-Seq itself was still prohibitively expensive for many research groups. This left smaller groups eagerly producing data via array technology, in the hope of producing high-quality results. Whilst this was viable at the gene-level, detection of AS events using array data proved to be a significant challenge, which this analysis potentially addresses in a greater depth than any previous work. The vast body of WT array data which exists in public data repositories such as GEO (Barrett et al., 2013) and Array Express (Kolesnikov et al., 2015) remains a resource which is still largely untapped at the exon and transcript level.

7.2 Key Findings

7.2.1 Background Signal

Beyond the development of a full Bayesian model for detection of AS events, this body of work contains some important additional observations. The first of these is the discrepancy between the fitted model coefficients for GCRMA across multiple datasets, and those calculated using 3' arrays (Figure 3.9). RMA remains the preferred background correction strategy for WT arrays due primarily to difficulties in the implementation of GCRMA using `aroma.affymetrix`. This may have been fortuitous for researchers as naïve use of GCRMA would be inappropriate for BG correction of WT arrays due to the clear differences in probe behaviour. The distinct characteristics of AG and MM probes included on Exon Arrays further complicate this picture in comparison to 3' arrays.

The strategy proposed under BMEA is essentially agnostic to the choice of MAT or GCRMA for estimation of bin-specific values $\hat{\lambda}_l$ and $\hat{\delta}_l$, with this model only playing a role in assignment of probes to a bin. The parameter estimates themselves derive from the *observed values* for BG probes within each bin, with these values showing minimal difference between bins defined under either the MAT or GCRMA model (Figure 3.11). The MAT model was chosen for the majority of analysis as the model itself better captures the true nature of data from BG probes than GCRMA (Figure 3.10).

A strong advantage of the approach taken under BMEA was that the previously unidentified disparity in NSB behaviour between samples was able to be clearly identified as a confounding effect which is technical, not biological in nature (Section 6.2). The comparison between quartiles of BG and PM probes (Figure 6.17) provides a simple QC check for any similar behaviour within datasets under analysis, and guidelines as to the interpretation of any exon-level predictions obtained. Whilst not formally investigated, the observed skew in the distribution of $\tilde{t}$-statistics in Figure 2.5 involved both comparisons against stimulated $\mathrm{T_{reg}}$, and this detected effect is likely to have played a role in this behaviour. The approach taken to rectify this in Section 2.3.1 remains an appropriate remedy to this problem. In addition, this artefact may have been a strong contributor to the disparate behaviours in Section 2.5, especially considering that the overwhelming majority of candidate AS events detected under FIRMA were sourced from a comparison involving stimulated $\mathrm{T_{reg}}$.

Although not raised in previous sections, a clear alternative strategy for assigning the values λ_{hijk} and δ_{hijk} would have been to use a sliding window approach and using the values obtained for some number (e.g. $n = 1000$) of probes which were the closest in value to the PM probe of interest. The use of bins provided a simple and clear way to identify these NSB artefacts, which may have remained more difficult to detect under an alternative approach. Finally, the assignment of PM probes to bins based on expected BG signal, provided a simple methodology for calculation of Z-scores for both whole-gene and exon-level probesets. Whilst initially conceived as a test for detection of true signal above background (DABG) and exclusion of probesets which only contain *off-target* signal, these scores provided measures for the overall strength of the *on-target* signal component over background. These scores proved to be an invaluable tool in assessment of suitable candidates for differential transcript usage, and restriction to probesets for which $S \gg B$ strongly improved the accuracy of predicted AS events.

Beyond analysis under BMEA, the use of Z-scores is possible for exclusion of genes which are not considered as expressed. Even when performing a gene-level analysis using RMA, the removal of these genes will clearly limit the number of false positives and increase the power of an experiment from the perspective of multiple testing considerations. When performing downstream analysis, such as GO or TFBS enrichment, this enables clear definition of a set of genes considered as expressed with the tissue of interest, and can provide a more suitable set of background genes than just the wider genome. If using the wider genome as the background or control dataset, terms which simply define the cell or tissue of experimental interest will commonly appear in lists of results, as opposed to terms associated with *the specific treatment within* that tissue. By using a set of reference genes which are specifically expressed within the cell type of interest, many of these terms which simply describe the cell type will not be included as enriched and a more informative set of results can be obtained.

7.2.2 The BMEA Model

After completion of much development and detection of confounding factors, the BMEA successfully identified an important biological process for which alternate transcript usage appears to play a role (Section 6.3.2). The immune synapse itself has been implicated in the exchange of signalling molecules between cells (Mittelbrunn et al., 2011) (Choudhuri et al., 2014), and as such the identification of non-coding transcripts which center around this process may be of much interest. Importantly, one of these transcripts was validated with qPCR (*CD53 - ENST00000471220*), confirming that at least one ncRNA associated with this process is indeed differentially regulated in response to activation. However, current association of this gene and process has been performed primarily at the protein level (Draber et al., 2011) and association of this transcript with the same process is still speculative at this point. Although being far from a definitive description of associated processes, this work has laid the foundation for potentially exciting research.

The majority of candidate AS events detected under BMEA are associated with the common activation signature, and in general appear to be a relatively small number of events compared to the levels of differential expression observed at the gene level. Three important factors are likely to have played a role in this: 1) The observed artefact for stimulated $\mathrm{T_{reg}}$ samples (Section 6.2.1); 2) Lower sample numbers in the Resting $\mathrm{T_{reg}}$ Vs $\mathrm{T_h}$ comparison; and, 3) The underlying biology.

The background signal artefact observed in stimulated $\mathrm{T_{reg}}$ samples in this dataset, is a type of technical effect not previously characterised and is difficult to easily ascribe to a protocol variation, beyond the general hybridisation and washing steps. The fact that a similar effect was found in the Tissue Mixture dataset suggests that this type of effect may be more common than previously realised, and does account for some of the anecdotal difficulty associated with AS detection using Exon Arrays, and the ease with which RNA-seq rendered the technology relatively obsolete. In particular, this effect largely explains the unsatisfactory FIRMA results in Section 2.5.2, confirming that many of these were likely to be false positives. The characterisation of this effect allowed for exclusion of candidate AS events which exhibited the predicted behaviour, however this directly reduced the predictive power of this dataset by effectively excluding two comparisons. For candidate AS events detected in other comparisons, the ability of stimulated $\mathrm{T_{reg}}$ to act as the second, confirmatory comparison was also significantly reduced. The confirmation of alternate transcript usage for *CD53* as a common activation effect, rather than the predicted $\mathrm{T_h}$-specific activation response reinforces this, and it is entirely possible that many of the predicted AS events from other comparisons may not be correctly classified as resting $\mathrm{T_{reg}}$ or stimulated $\mathrm{T_h}$ effects only.

The four-way layout of the T_{reg} dataset, and the changing mixture levels made identification of this type of effect simple to detect. In a more simple, but common experimental design, with only two cell types, this may prove more of a challenge. There will only be two conditions for each bin, and identifying whether one cell type has an positive or negative bias for estimation of λ_{il} will be less clear. A clear alternative may be to average across an experiment to arrive at values for $\lambda_{\cdot l}$ and $\delta_{\cdot l}$, however as noted in Figure 6.16 this impact may also be evident within the overall set of intensities, including the PM intensities. If they are also impacted, this may potentially be a difficult effect to resolve beyond taking a cautious and careful approach to interpretation of results.

When selecting candidate AS events, the restriction on comparisons involving stimulated T_{reg} clearly reduced the power of this dataset, with only the resting T_{reg} Vs T_{h} ($n = 4$) and stimulated vs resting T_{h} ($n = 5$) comparisons being able to offer their full capability. It is also of key importance that the initial power calculations for this dataset (Section 2.2.4) were performed by estimating fold-change under the common model assumptions used by RMA/PLM. The detection of AS events is a fundamentally different question, and as such would require a different set of calculations. As minimal information regarding exon-level analysis was available before analysis began, this dataset is likely to be under-powered for this level of analysis. The far smaller numbers of probes utilised in exon-level probesets is likely to lead to greater instability of estimates than at the gene-level, where far greater probe numbers are able to be used to provide more stable estimates of gene-level terms. This smaller number of probes within an exon-level probeset also gives less opportunity for the data to overcome any influence the priors have on final results. The alternate specifications used for generation of the custom CDF may have slightly enabled an improved performance by creating at least a limited number of probesets with increased probe numbers.

The Tissue Mixture dataset was able to detect hundreds of AS events from 3 samples in each of the 100% tissue groups, lending confidence to the capacity of the BMEA approach under low sample numbers. How many of the true AS events this represents remains a difficult question to answer, and with these being two highly differentiated tissues, it is likely that detected events represent only a small proportion of the true AS events which exist between the tissues. Comparisons between the four cell types in the T_{reg} dataset involve four very similar cell types, and as such, a far lower number of AS events may be expected than in the Heart Vs Brain comparison. Thus, given the underlying biological similarities, a far smaller pool of true AS events would have been present in the initial samples, and the much lower numbers of candidate AS events was not unexpected.

In testing using simulated data BMEA appeared to strongly outperform the FIRMA approach and it is not unreasonable to expect this behaviour to continue when working with true experimental data. The addition of filtering steps during selection of candidate AS events under BMEA makes a direct comparison of approaches difficult, as development of similar strategies for analysis using FIRMA was not of primary interest. However, a key advantage of BMEA remains the inclusion of intuitive model terms, where the proportion of transcripts containing each exon is directly modelled, making results far easier to interpret and communicate with researchers. The simpler interpretation of Bayesian results when compared to the conventional frequentist approach of rejecting or accepting H_0, additionally makes interpretation by researchers an easier task.

7.3 Future Directions

7.3.1 The BMEA Model

Whilst successfully providing strong leads for future experiments, BMEA has much scope for further improvements, notwithstanding the wider shift towards sequencing technologies. The $\mathrm{T}_{\mathrm{reg}}$ dataset contained matched samples from within the same donor, and the model detailed under BMEA does not directly take this data structure into account. This is clearly an area which could be further explored by the definition of a model in which the change in exon-inclusion rates ($\Delta\phi_{hj}$) and overall fold-change ($\Delta\mu_h$) are defined in this manner. This may have additional advantages for specifying other nesting structures within sample groups such as sibships for murine experiments or other similar experimental designs.

The exon-level term ϕ_{hj} was specifically defined to sidestep difficulties defining the distribution of this term across samples within a cell-type. Conceptually, this does allow for detection of candidate AS events which are not consistent within each cell type, however the general frequentist approach of detection of average differences would behave in a similar manner. Incorporation of a sample-specific term at the exon level, with a relationship to ϕ_{hj} in keeping with the structure of chip-effects (c_{hi}) and the cell-type specific expression level (μ_h), would be a further area of potential future development.

The assignment of sample-weights has become a common-place procedure when analysing 3' array data and for RNA-seq data when utilising the *voom* method (Law et al., 2014). Considering the observations regarding uneven hybridisation, the development of a model incorporating sample weights may be an additional development which strengthens this approach.

A key refinement to the model which would be the clear next step would be to develop an approach which better models normalisation. All analyses used quantile normalised data, as that appeared to be the most appropriate for gene-level results. The appropriateness of this for exon-level analysis is less clear, as the fundamental requirement for equal amounts of total hybridised sample is no longer relevant. A superior approach may be to normalise data by the incorporation of an additional term in the model, as is used by approaches such as RUV (Gagnon-Bartsch & Speed, 2012). This may better allow for accurate assessment at the gene-level, without compromising the power of the approach to detect changes at the exon-level, as noted in Section 6.2.2. Additionally finding a method to more satisfactorily manage any unexpected bias in $\hat{\lambda}_{il}$ may also prove strongly advantageous.

Whilst Figure 3.11 showed distributions which were a reasonable approximation of normally distributed data, the assumption of normality for the background signal, is also an area which may require deeper exploration. At each iteration t of the MCMC process, BMEA samples $\hat{S}^t = PM - \hat{B}^t$, with $\hat{B}^t$ drawn from the prior $\mathcal{N}(\lambda, \delta)$. An alternative may be to sample observed values from AG or MM probes with similar properties to each PM probe, or to construct an empirical distribution based on observed values. However, neither of these approaches would correctly deal with competitive binding of target and off-target sequences.

The correlations between replicates for values such as μ_h also showed an increase in variability for genes at the low end of expression values which is reminiscent of that for replicates under MAS5.0 (Zakharkin et al., 2005), although this was not as extreme as has been noted under MAS5.0. BMEA essentially introduces variance for BG signal and variance for true signal, as does MAS5.0 and this is to be expected to a certain extent. An alternative model-based approach for $E(S|PM)$ in this context may be a further avenue to explore to better manage this. In particular, the representation of BG signal offered by MAT may prove advantageous and methods of incorporating this into a model-based approach such as for GC-RMA may be feasible.

An unplanned development which arose from this work was the definition of an alternate CDF, which was able to partially overcome the limitation of ≤ 4 probes per exon-level probeset. The definitions used in Section 5.2.2 utilised the common probes for Ensembl-based genes and exons. Probes which mapped to common sets of exons were considered as probesets, to allow for the potential detection of novel AS events. Considering that no novel AS events were detected in Section 6.3, an alternative may have been to define probesets based on probes which target common transcripts. If using this approach, the annotation file could then be defined to contain the appropriate transcript identifiers associated with the probesets considered as detecting a putative AS event. This would effectively render them less as exon-level probesets, more as alternate-transcript detecting probesets, and these probesets would generally be far larger than 4 probes. Whilst it is expected that these may provide more stable estimates at both the gene and alternate-transcript level, this would limit the capacity of an experiment to detect novel AS events. Interestingly, this would contradict the implied design of “Exon” Arrays, which may subtly bias the researcher into approaching these analyses by considering the exon as the important unit, as opposed to the true goal of detecting alternate transcript usage.

The CDF defined in Section 5.2.2 was also based on the Ensembl gene models which contain numerous predicted transcripts with minimal supporting evidence. Whilst this proved advantageous for the detection of many of the events in Section 6.3.1, this may also increase the number of false positives due to the presence of many small probesets targeting subtle changes in exon definitions. A more conservative approach may be to define transcripts or exon-level probesets using more stringent gene models such as those contained in the RefSeq database. As these are generally better characterised transcripts than those in the Ensembl database, this may provide an additional stringency on the biological relevance of any candidate alternate transcript usage.

Whilst Exon Array usage is in clear decline, thousands of experiments using this technology already reside in the GEO database (`https://www.ncbi.nlm.nih.gov/geo/`). As discussed above, the ability to redefine the CDF used for analysis also opens the possibility of considering Affymetrix Gene Arrays as candidates for analysis using BMEA, as these also contain probes targeting the length of each gene. Thousands of these additional datasets exist in public repositories. Whilst many of these datasets will suffer from a lack of experimental power, the development of BMEA will enable a degree of retrospective mining of these datasets, and the exploitation of an existing resource has thus been enabled to a far greater extent than previously.

7.3.2 T Cell Research

Surprisingly, the majority of the detected AS events tracked more closely with activation than with differences between $\mathrm{T_{reg}}$ and $\mathrm{T_h}$ cells. Whether this is an artefact of the limited power, or representative of the underlying biology is unclear, however this research has shed important light on an important area of T cell biology. Much of the $\mathrm{T_{reg}}$-APC communication process is still poorly understood, and this research may provide insights into key genes involved in this process. Further study into all isoforms of interest is very much a blank slate for biologists as most of the predicted and experimentally verified transcripts detected here are still poorly understood. The surprising number of non-coding transcripts detected as differentially expressed provides an exciting additional opportunity for further investigation in this rapidly growing area of biological research.

Appendix A

MCMC Sampling Procedures

A.1 True Signal Component

Under BMEA the background signal B and true signal S are considered as independent. In reality, given the nature of competitive binding, this assumption may be violated to some extent. However, in terms of the modelling parameters this assumption can be considered to hold. Thus, the equation

$$h(PM) = h(B, S) = f(B).g(S) \tag{A.1}$$

is the basis for model specification.

Given the specification of B_{hijk} in Equation 3.6a, and temporarily ignoring the truncation point, we can see that the background signal for each probe can be modelled using the log-normal probability density function

$$f(B_{hijk}) = \frac{1}{B_{hijk}\delta_{hijk}\sqrt{2\pi}} \exp\left[-\frac{(\log B_{hijk} - \lambda_{hijk})^2}{2\delta^2_{hijk}}\right]. \tag{A.2}$$

Using the same notation as 3.3, the subscripts h, i, j & k respectively denote the cell-type, chip, exon and probe. The values λ_{hijk} & δ_{hijk} are assigned to each probe before BMEA model fitting, and are based on the probe sequence, as described in Section 3.4.

A.1.1 Uniform Model

Noting that $B_{hijk} = PM_{hijk} - S_{hijk}$, Equation A.2 can be equivalently expressed as

$$\begin{aligned} f(B_{hijk}) &= f(PM_{hijk} - S_{hijk}) \\ &= \frac{1}{(PM_{hijk} - S_{hijk})\delta_{hijk}\sqrt{2\pi}} \exp\left[-\frac{(\log(PM_{hijk} - S_{hijk}) - \lambda_{hijk})^2}{2\delta^2_{hijk}}\right]. \end{aligned} \tag{A.3}$$

Similarly, as specified in Equation 3.6b, the true signal term S_{hijk} will have the density function

$$g(S_{hijk}) = \frac{1}{S_{hijk}\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(\log S_{hijk} - \eta_{hijk})^2}{2\sigma^2_S}\right] \tag{A.4}$$

where

$$\eta_{hijk} = \log\phi_{hj} + c_{hi} + p_{jk}\,.$$

This will give the conditional density function of interest as

$$g(S|PM) = \frac{f(PM - S)g(S)}{\int_0^{PM} f(PM - S)g(S)dS}\,. \tag{A.5}$$

MCMC Updating for S Under the Uniform Model

Temporarily suppressing subscripts for each given probe PM_{hijk}, equation A.5 can be simplified to

$$\begin{aligned} g(S|PM, \eta, \lambda, \delta, \sigma_S) &\propto f(PM - S)g(S) \\ &\propto \frac{1}{S(PM - S)} \exp\left[-\frac{(\log(PM - S) - \lambda)^2}{2\delta^2} - \frac{\log S - \eta}{2\sigma^2_S}\right], \end{aligned} \tag{A.6}$$

and

$$\log g(S|PM, \eta, \lambda, \delta, \sigma_S) \propto -\log\{S(PM - S)\} - \frac{\{\log(PM - S) - \lambda\}^2}{2\delta^2} - \frac{\log S - \eta}{2\sigma^2_S}\,. \tag{A.7}$$

At iteration '$t : t \geq 1$', S can be updated using a Metropolis-Hastings step. A new value S^* can be proposed by sampling B^* from the prior, truncated at the observed value PM

$$\log B \sim \mathcal{N}(\lambda, \delta); 0 \leq B \leq PM\,. \tag{A.8}$$

Then, relying on the property $S^* = PM - B^*$ the value is accepted for S^t with probability r using

$$r = \min\left(\frac{g(S^*|PM, \eta, \lambda, \delta, \sigma_S)}{g(S^{t-1}|PM, \eta, \lambda, \delta, \sigma_S)}, 1\right) . \tag{A.9}$$

A.1.2 Mixture Model

Under the mixture model, the complete specification of S requires the 3-component mixture model

$$S_{hijk} \sim \begin{cases} \text{Log-}\mathcal{N}(c_{hi} + p_{jk}, \sigma_S) & \phi_{hj} = 1 \\ \text{Log-}\mathcal{N}(c_{hi} + p_{jk} + \log\phi_{hj}, \sigma_S) & 0 < \phi_{hj} < 1 \\ 0 & \phi_{hj} = 0 \end{cases} \cdot \tag{A.10}$$

Group membership for each probe & exon can be represented using the indicator variable

$$\boldsymbol{\xi}_{hj} = (\xi_{hj}^1, \xi_{hj}^2, \xi_{hj}^3) \sim \text{Multinomial}(1, \boldsymbol{q}_{hj})$$

where

$$\boldsymbol{q}_{hj} = (q_{hj}^1, q_{hj}^2, q_{hj}^3); \quad \sum_{n=1}^{3} q_{hj}^n = 1\,.$$

For notational convenience, let the vector of known values be represented by

$$\boldsymbol{\psi} = (\boldsymbol{c}, \boldsymbol{\mu}, \sigma_\mu, \boldsymbol{p}, \sigma_p, \sigma_S, \boldsymbol{\lambda}, \boldsymbol{\delta})\,,$$

where

$$\begin{aligned}
\boldsymbol{c} &= (c_{hi}; h = 1, \ldots, H; i = 1, \ldots, I_h) \\
\boldsymbol{\mu} &= (\mu_h; h = 1, \ldots, H) \\
\boldsymbol{p} &= (p_{jk}; j = 1, \ldots, J; k = 1, \ldots, K_j) \\
\boldsymbol{\lambda} &= (\lambda_{hijk}; h = 1, \ldots, H; i = 1, \ldots, I_h; j = 1, \ldots J; k = 1, \ldots, K_j) \\
\boldsymbol{\delta} &= (\delta_{hijk}; h = 1, \ldots, H, i = 1, \ldots, I_h, j = 1, \ldots J, k = 1, \ldots, K).
\end{aligned}$$

The true signal component of the intensity for each probe will have pdf

$$g(S_{hijk}|\boldsymbol{\psi}, \boldsymbol{\xi}_{hj}, \phi_{hj}) = \begin{cases} g^1(S_{hijk}|\boldsymbol{\psi}, \phi_{hj}) & \xi^1_{hj} = 1 \\ g^2(S_{hijk}|\boldsymbol{\psi}, \phi_{hj}) & \xi^2_{hj} = 1 \\ g^3(S_{hijk}|\boldsymbol{\psi}, \phi_{hj}) & \xi^3_{hj} = 1 \end{cases} \tag{A.11}$$

where

$$\begin{aligned} g^1(S_{hijk}|\boldsymbol{\psi}, \phi_{hj}) &= \frac{1}{S_{hijk}\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(\log S_{hijk} - c_{hi} - p_{jk})^2}{2\sigma^2_S}\right] \\ g^2(S_{hijk}|\boldsymbol{\psi}, \phi_{hj}) &= \frac{1}{S_{hijk}\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(\log S_{hijk} - c_{hi} - p_{jk} - \log\phi_{hj})^2}{2\sigma^2_S}\right] \\ g^3(S_{hijk}|\boldsymbol{\psi}, \phi_{hj}) &= \begin{cases} 1 & \text{when} S_{hijk} = 0 \\ 0 & \text{elsewhere.} \end{cases} . \end{aligned}$$

Noting that only $g^2(S_{hijk}|\boldsymbol{\psi}, \phi_{hj})$ is dependent on ϕ_{hj}, the marginal probability for $g^2(S_{hijk}|\boldsymbol{\psi})$ can be obtained by solving

$$g^2(S_{hijk}|\boldsymbol{\psi}) = \int_0^1 \frac{1}{S_{hijk}\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(\log S_{hijk} - c_{hi} - p_{jk} - \log\phi)^2}{2\sigma^2_S}\right] d\phi\,. \tag{A.12}$$

Firstly, let

$$u = \log S_{hijk} - c_{hi} - p_{jk}$$

and

$$e^x = \phi \Rightarrow d\phi = e^x dx\,.$$

Substituting these into equation A.12 gives

$$\begin{aligned} g^2(S_{hijk}|\boldsymbol{\psi}) &= \int_{-\infty}^0 \frac{1}{S_{hijk}\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(u-x)^2}{2\sigma^2_S}\right] e^x dx \\ &= \frac{1}{S_{hijk}} \int_{-\infty}^0 \frac{1}{\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(u-x)^2}{2\sigma^2_S} + x\right] dx\,. \end{aligned}$$

Taking the terms in the exponent and completing the square:

$$\begin{aligned}
-\frac{(u-x)^2}{2\sigma_S^2}+x &= \frac{-u^2+2ux-x^2+2\sigma_S^2 x}{2\sigma_S^2} \\
&= \frac{-u^2+2(u+\sigma_S^2)x-x^2}{2\sigma_S^2} \\
&= \frac{-(u+\sigma_S^2)^2+2(u+\sigma_S^2)x-x^2}{2\sigma_S^2}+\frac{(u+\sigma_S^2)^2-u^2}{2\sigma_S^2} \\
&= \frac{-(x-(u+\sigma_S^2))^2}{2\sigma_S^2}+\frac{(u+\sigma_S^2)^2-u^2}{2\sigma_S^2} \\
&= \frac{-(x-(u+\sigma_S^2))^2}{2\sigma_S^2}+u+\frac{\sigma_S^2}{2}\,.
\end{aligned}$$

Returning to the main equation again:

$$\begin{aligned}
g^2(S_{hijk}|\boldsymbol{\psi}) &= \frac{1}{S_{hijk}}\int_{-\infty}^{0}\frac{1}{\sigma_S\sqrt{2\pi}}\exp\left[\frac{-(x-(u+\sigma_S^2))^2}{2\sigma_S^2}+u+\frac{\sigma_S^2}{2}\right]dx \\
&= \frac{1}{S_{hijk}}\exp\left(u+\frac{\sigma_S^2}{2}\right)\int_{-\infty}^{0}\frac{1}{\sigma_S\sqrt{2\pi}}\exp\left[\frac{-(x-(u+\sigma_S^2))^2}{2\sigma_S^2}\right]dx\,.
\end{aligned}$$

Now let

$$t=\frac{x-(u+\sigma_S^2)}{\sigma_S}\Rightarrow dx=\sigma_S dt\,.$$

Substituting t into the equation gives

$$\begin{aligned}
g^2(S_{hijk}|\boldsymbol{\psi}) &= \frac{1}{S_{hijk}}\exp\left[u+\frac{\sigma_S^2}{2}\right]\int_{-\infty}^{-\frac{u+\sigma_S^2}{\sigma_S}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2}\right]dt \\
&= \frac{1}{S_{hijk}}\exp\left[u+\frac{\sigma_S^2}{2}\right]\Phi\left[-\frac{u+\sigma_S^2}{\sigma_S}\right]
\end{aligned}$$

where $\Phi(x)$ denotes the standard normal cumulative distribution function.

Returning the original parameters to the model and simplifying the S_{hijk} terms

$$g^2(S_{hijk}|\boldsymbol{\psi})=\exp\left[\frac{\sigma_S^2}{2}-c_{hi}-p_{jk}\right]\Phi\left(-\frac{\log S_{hijk}-c_{hi}-p_{jk}+\sigma_S^2}{\sigma_S}\right)\,. \tag{A.13}$$

Thus the full set of marginal probabilities can be written as

$$g(S_{hijk}|\boldsymbol{\psi}) = \begin{cases} \begin{cases} 0 & \text{when } S_{hijk} = 0 \\ \frac{1}{S_{hijk}\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(\log S_{hijk} - c_{hi} - p_{jk})^2}{2\sigma_S}\right] & \text{elsewhere} \end{cases} \\ \\ \begin{cases} 0 & \text{when } S_{hijk} = 0 \\ \exp\left[\frac{\sigma_S^2}{2} - c_{hi} - p_{jk}\right] \Phi\left(-\frac{\log S_{hijk} - c_{hi} - p_{jk} + \sigma_S^2}{\sigma_S}\right) & \text{elsewhere} \end{cases} \\ \\ \begin{cases} 1 & \text{when } S_{hijk} = 0 \\ 0 & \text{elsewhere.} \end{cases} \end{cases} . \tag{A.14}$$

Noting in the above that

$$g^1(S_{hijk}|\boldsymbol{\psi}) \equiv S_{hijk} \sim \log\mathcal{N}(c_{hi} + p_{jk}, \sigma_S^2)\,.$$

We also know that since $PM = B + S$, where B has the pdf as defined in Equation A.2, for each of the mixture components $n = 1, 2, 3$ the pdf will be of the form

$$f^n\left(PM_{hijk}|\boldsymbol{\psi}\right) = \int_0^{PM_{hijk}} f\left(PM_{hijk} - S_{hijk}|\boldsymbol{\psi}\right) g^n\left(S_{hijk}|\boldsymbol{\psi}\right) dS\,.$$

This can be expressed more explicitly as:

$$f(PM_{hijk}|\boldsymbol{\psi}) = \begin{cases} \int_0^{PM_{hijk}} \frac{1}{S(PM_{hijk}-S)\delta_{hijk}\sigma_S 2\pi} \exp\left[-\frac{\left(\log(PM_{hijk}-S)-\lambda_{hijk}\right)^2}{2\delta^2_{hijk}} - \frac{\left(\log S-c_{hi}-p_{jk}\right)^2}{2\sigma^2_S}\right] dS & \xi^1_{hj}=1 \\ \int_0^{PM_{hijk}} \frac{1}{(PM_{hijk}-S)\delta_{hijk}\sqrt{2\pi}} \exp\left[\frac{\sigma^2_S}{2} - c_{hi} - p_{jk} - \frac{(\log(PM_{hijk}-S)-\lambda_{hijk})^2}{2\delta^2_{hijk}}\right] \Phi\left(-\frac{\log S_{hijk}-c_{hi}-p_{jk}+\sigma^2_S}{\sigma_S}\right) dS & \xi^2_{hj}=1 \\ \frac{1}{PM_{hijk}\delta_{hijk}\sqrt{2\pi}} \exp\left[-\frac{(\log(PM_{hijk}-S)-\lambda_{hijk})^2}{2\delta^2_{hijk}}\right] & \xi^3_{hj}=1 \end{cases}, \tag{A.15}$$

and subsequently

$$g(S_{hijk}|PM_{hijk},\boldsymbol{\psi},\boldsymbol{\xi_{hj}}) = \begin{cases} \begin{cases} 0 & \text{for} S_{hijk}=0 \\ \frac{1}{S_{hijk}(PM_{hijk}-S_{hijk})\delta_{hijk}\sigma_S 2\pi} \exp\left[-\frac{\left(\log(PM_{hijk}-S)-\lambda_{hijk}\right)^2}{2\delta^2_{hijk}} - \frac{\left(\log S-c_{hi}-p_{jk}\right)^2}{2\sigma^2_S}\right] & \text{elsewhere} \end{cases} & \xi^1_{hj}=1 \\ \begin{cases} 0 & \text{for} S_{hijk}=0 \\ \frac{1}{(PM_{hijk}-S)\delta_{hijk}\sqrt{2\pi}} \exp\left[\frac{\sigma^2_S}{2} - c_{hi} - p_{jk} - \frac{(\log(PM_{hijk}-S)-\lambda_{hijk})^2}{2\delta^2_{hijk}}\right] \Phi\left(-\frac{\log S_{hijk}-c_{hi}-p_{jk}+\sigma^2_S}{\sigma_S}\right) & \text{elsewhere} \end{cases} & \xi^2_{hj}=1 \\ \begin{cases} 1 & \text{for} S_{hijk}=0 \\ 0 & \text{elsewhere} \end{cases} & \xi^3_{hj}=1\,. \end{cases} \tag{A.16}$$

For the set of probes belonging to condition h and exon j with observed values

$$\boldsymbol{PM_{hj}} = \{PM_{h1j1}, PM_{h1j2}, \ldots, PM_{h1jk}, PM_{h2j1}, \ldots, PM_{hijk}\}$$

and corresponding signal estimates

$$\boldsymbol{S_{hj}} = \{S_{h1j1}, S_{h1j2}, \ldots, S_{h1jk}, S_{h2j1}, \ldots, S_{hijk}\}\,,$$

assuming independence between the signal estimates, the pdf for $\boldsymbol{S_{hj}}$ becomes

$$g(\boldsymbol{S_{hj}}|\boldsymbol{PM_{hj}}, \boldsymbol{\psi}, \boldsymbol{\xi_{hj}}) = \prod_{i=1}^{I_h}\prod_{k=i}^{K_j} g_{ik}(S_{hijk}|\boldsymbol{PM_{hj}}, \boldsymbol{\psi}, \boldsymbol{\xi_{hj}})\,.$$

Thus,

$$g(\boldsymbol{S_{hj}}|\boldsymbol{PM_{hj}}, \boldsymbol{\psi}, \boldsymbol{\xi_{hj}}) = \begin{cases} \prod_{i=1}^{I_h}\prod_{k=i}^{K_j} g^1(S_{hijk}|PM_{hijk}, \boldsymbol{\psi}, \boldsymbol{\xi_{hj}}) & \xi^1_{hj} = 1 \\ \\ \prod_{i=1}^{I_h}\prod_{k=i}^{K_j} g^2(S_{hijk}|PM_{hijk}, \boldsymbol{\psi}, \boldsymbol{\xi_{hj}}) & \xi^2_{hj} = 1 \\ \\ \prod_{i=1}^{I_h}\prod_{k=i}^{K_j} g^3(S_{hijk}|PM_{hijk}, \boldsymbol{\psi}, \boldsymbol{\xi_{hj}}) & \xi^3_{hj} = 1 \end{cases} . \tag{A.17}$$

MCMC Updating for S Under the Mixture Model

The parameter S can then updated at iteration 't' by the following Metropolis-Hastings approach:

1. Generate proposal values $\boldsymbol{S^*_{hj}|PM_{hj},\xi^t_{hj},\psi^{t-1}}$

2. Sample $\boldsymbol{S^t_{hj}}$ from $(\boldsymbol{S^*_{hj}},\boldsymbol{S^{t-1}_{hj}})$ with probability $r_{hj}=\{r_{h1j1},r_{h1j2},\ldots,r_{h1jk},r_{h2j1},\ldots,r_{hijk}\}$, where

$$r_{hijk}=\min\left[1,\frac{g(S^*_{hijk}|\boldsymbol{PM_{hj},\xi^t_{hj},\psi^{t-1}})f(S^{t-1}_{hijk}|S^*_{hijk})}{g(S^{t-1}_{hijk}|\boldsymbol{PM_{hj},\xi^t_{hj},\psi^{t-1}})f(S^*_{hijk}|S^{t-1}_{hijk})}\right].$$

Generation of proposal values for $\boldsymbol{S^*_{hj}|PM_{hj},\xi^t_{hj}\neq(0,0,1),\psi^{t-1}}$ can be performed using the prior for B_{hijk} truncated at the observed value PM_{hijk} as for the Uniform model. It should be noted however, that automatic acceptance of $\boldsymbol{S^*_{hj}}$ will occur when $\boldsymbol{\xi^{t-1}_{hj}}=(0,0,1)$ & $\boldsymbol{\xi^t_{hj}}\neq(0,0,1)$.

A.2 The Expression Related Terms c_{hi}, μ_h & σ_μ

Taking all other terms $(S_{hijk}, p_{jk}, \phi_{hj}, \sigma_S)$ as being known, we can see that given

$$\log S_{hijk} = c_{hi} + p_{jk} + \log \phi_{hj} + \varepsilon_{hijk} \tag{A.18}$$

for

$$\begin{aligned} x_{hijk} &= \log S_{hijk} - p_{jk} - \log \phi_{hj} \\ \Rightarrow x_{hijk} &\sim \mathcal{N}(c_{hi}, \sigma_S^2) \\ \Rightarrow \bar{x}_{hi\cdot\cdot} &\sim \mathcal{N}(c_{hi}, \sigma_{hi}^2) \end{aligned}$$

where

$$\begin{aligned} \bar{x}_{hi\cdot\cdot} &= \frac{1}{K_h} \sum_{\{j \in J : \phi_{hj} \neq 0\}} \sum_{k=1}^{K_j} x_{hijk} \\ \sigma_{hi}^2 &= \frac{\sigma_S^2}{K_h} \\ K_h &= \sum_{\{j \in J : \phi_{hj} \neq 0\}} K_j \,. \end{aligned}$$

Noting that under the Uniform model $\phi_{hj} \neq 0$ and thus $K_h = K$.

A.2.1 Updating the Chip Effects Term (c_{hi})

For the set of H treatment groups, the posterior for

$$\boldsymbol{\theta} = (c_{11}, \ldots, c_{1I_1}, c_{21}, \ldots, c_{2I_2}, \ldots, c_{H1}, \ldots, c_{HI_H})$$

can thus be written as

$$p\left(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma_\mu^2 | \boldsymbol{x}\right) \propto p\left(\boldsymbol{\mu}, \sigma_\mu^2\right) p\left(\boldsymbol{\theta} | \boldsymbol{\mu}, \sigma_\mu^2\right) p\left(\boldsymbol{x} | \boldsymbol{\theta} \boldsymbol{\mu}, \sigma_\mu^2\right) \,.$$

From the above, and noting that $\mu_h \sim \mathcal{U}(0, \log 2^{16})$, the three components can then be defined as:

$$\begin{aligned} p\left(\boldsymbol{x}|\boldsymbol{\theta}\boldsymbol{\mu}, \sigma_\mu^2\right) &= \prod_{h=1}^{H}\prod_{i=1}^{I_h} \mathcal{N}\left(c_{hi}, \sigma_{hi}^2\right) \\ p\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \sigma_\mu^2\right) &= \prod_{h=1}^{H}\prod_{i=1}^{I_h} \mathcal{N}\left(c_{hi}|\mu_h, \sigma_\mu^2\right) \\ p\left(\boldsymbol{\mu}, \sigma_\mu^2\right) &= p(\boldsymbol{\mu}|\sigma_\mu^2)p(\sigma_\mu^2) \propto p(\sigma_\mu^2)\,. \end{aligned}$$

The full conditional posterior is thus:

$$\begin{aligned} p\left(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma_\mu^2|\boldsymbol{x}\right) &\propto p(\sigma_\mu^2)\prod_{h=1}^{H}\prod_{i=1}^{I_h} \mathcal{N}\left(c_{hi}|\mu_h, \sigma_\mu^2\right)\prod_{h=1}^{H}\prod_{i=1}^{I_h} \mathcal{N}\left(c_{hi}, \sigma_{hi}^2\right) \\ &= p(\sigma_\mu^2)\prod_{h=1}^{H}\prod_{i=1}^{I_h} \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{(c_{hi}-\mu_h)^2}{2\sigma_\mu^2}\right] \frac{1}{\sqrt{2\pi\sigma_{hi}^2}} \exp\left[-\frac{(\bar{x}_{hi\cdot\cdot}-c_{hi})^2}{2\sigma_{hi}^2}\right]. \end{aligned}$$

As per Gelman et al., 2004, p. 135, this simplifies to

$$c_{hi}|\boldsymbol{\mu}, \sigma_\mu, \boldsymbol{x} \sim \mathcal{N}(\hat{c}_{hi}, V_{hi})$$

where

$$\hat{c}_{hi} = V_{hi}\left(\frac{\bar{x}_{hi\cdot\cdot}}{\sigma_{hi}^2} + \frac{\mu_h}{\sigma_\mu^2}\right) \text{and} \quad V_{hi} = \frac{1}{\frac{1}{\sigma_{hi}^2} + \frac{1}{\sigma_\mu^2}}\,.$$

Noting that in our model construction $\sigma_{hi} = \dfrac{\sigma_S}{\sqrt{K_h}}$, and K_h is constant across all arrays in condition H, this can be rewritten as

$$\hat{c}_{hi} = V_h\left(\frac{K_h\bar{x}_{hi\cdot\cdot}}{\sigma_S^2} + \frac{\mu_h}{\sigma_\mu^2}\right) \text{and} \quad V_h = \frac{1}{\frac{K_h}{\sigma_S^2} + \frac{1}{\sigma_\mu^2}} = \frac{\sigma_\mu^2\sigma_S^2}{K_h\sigma_\mu^2 + \sigma_S^2}\,.$$

Thus for each treatment condition

$$c_{hi}|\bar{x}_{hi}, \mu_h, \sigma_\mu^2, \sigma_S^2 \sim \mathcal{N}\left(V_h\left(\frac{K_h\bar{x}_{hi\cdot\cdot}}{\sigma_S^2} + \frac{\mu_h}{\sigma_\mu^2}\right), V_h\right) \tag{A.19}$$

and using this distribution, the c_{hi} terms can be updated using a Gibbs Sampler.

A.2.2 Updating the Condition Specific Expression-Level Terms (μ_h)

As expressed in Gelman et al., 2004, p. 136, the posterior for $\boldsymbol{\mu}|\sigma_\mu^2, \boldsymbol{x}$ can be written as

$$p(\boldsymbol{\mu}|\sigma_\mu^2, \boldsymbol{x}) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{V_\mu})$$

for the vectors $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_h)$ and $\boldsymbol{V_\mu} = (V_{\mu_1}, V_{\mu_2}, \ldots, V_{\mu_h})$, with $h = 1, \ldots, H$, where

$$\begin{aligned}
\hat{\mu}_h &= V_{\mu_h} \sum_{i=1}^{I_h} \frac{\bar{x}_{hi\cdot\cdot}}{\sigma_{hi}^2 + \sigma_\mu^2} \\
V_{\mu_h}^{-1} &= \sum_{i=1}^{I_h} \frac{1}{\sigma_{hi}^2 + \sigma_\mu} = \frac{I_h}{\sigma_{hi}^2 + \sigma_\mu^2} \\
\sigma_{hi}^2 &= \frac{\sigma_S^2}{K_h}\,.
\end{aligned}$$

Noting once again that $\sigma_{hi}^2 = \frac{\sigma_S^2}{K_h}$, the values $\hat{\mu}_h$ & V_h can thus be simplified to

$$\begin{aligned}
V_{\mu_h} &= \frac{\sigma_S^2 + K_h \sigma_\mu^2}{I_h K_h} \\
\hat{\mu}_h &= \frac{1}{I_h} \sum_{i=1}^{I_h} \bar{x}_{hi\cdot\cdot} = \bar{x}_{h\cdot\cdot\cdot}
\end{aligned}$$

and

$$\boldsymbol{\mu}|\boldsymbol{x}, \sigma_\mu^2, \sigma_S^2 \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{V_\mu})\,. \tag{A.20}$$

The vector $\boldsymbol{\mu}$ can be updated using a Gibbs Sampler.

A.2.3 Updating the Expression-Level Variance Term σ_μ

Again as expressed in Gelman et al., 2004, p. 136, the posterior for the variance term σ_μ^2 is defined by

$$\begin{aligned}
p(\sigma_\mu^2|\boldsymbol{x}, \sigma_S) &\propto \frac{p(\sigma_\mu^2|\sigma_S) \prod_{h=1}^H \prod_{i=1}^{I_h} \mathcal{N}(\bar{x}_{hi\cdot\cdot}|\mu_h, \sigma_{hi}^2 + \sigma_\mu^2)}{p(\boldsymbol{\mu}|\sigma_\mu, \boldsymbol{x}, \sigma_S)} \\
&\propto \frac{p(\sigma_\mu^2|\sigma_S) \prod_{h=1}^H \prod_{i=1}^{I_h} \left[2\pi(\sigma_{hi}^2 + \sigma_\mu^2)\right]^{-\frac{1}{2}} \exp\left[-\frac{(\bar{x}_{hi\cdot\cdot} - \mu_h)^2}{2(\sigma_{hi}^2 + \sigma_\mu^2)}\right]}{\prod_{h=1}^H (2\pi V_{\mu_h})^{-\frac{1}{2}} \exp\left[-\frac{(\mu - \hat{\mu}_h)^2}{2V_{\mu_h}}\right]}\,.
\end{aligned}$$

Noting that the denominator holds for all values of μ, set $\mu = \hat{\mu}_h$ and simplifying gives

$$\begin{aligned} p(\sigma_\mu^2|(x), \sigma_S) &\propto p(\sigma_\mu^2) \prod_{h=1}^{H} \left(\frac{\sigma_S^2}{K_h} + \sigma_\mu^2 \right)^{-\frac{I_h - 1}{2}} \prod_{i=1}^{I_h} \exp\left[-\frac{(\bar{x}_{hi\cdot\cdot} - \mu_h)^2}{2(\sigma_{hi}^2 + \sigma_\mu^2)} \right] \\ &= p(\sigma_\mu^2) \prod_{h=1}^{H} \left(\frac{\sigma_S^2}{K_h} + \sigma_\mu^2 \right)^{-\frac{I_h - 1}{2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{I_h} \frac{(\bar{x}_{hi\cdot\cdot} - \mu_h)^2}{(\sigma_{hi}^2 + \sigma_\mu^2)} \right] . \end{aligned}$$

With the uniform prior $p(\sigma_\mu^2)$

$$p(\sigma_\mu^2|\boldsymbol{x}, \sigma_S) \propto \prod_{h=1}^{H} \left(\frac{\sigma_S^2}{K_h} + \sigma_\mu^2 \right)^{-\frac{I_h - 1}{2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{I_h} \frac{(\bar{x}_{hi\cdot\cdot} - \mu_h)^2}{(\sigma_{hi}^2 + \sigma_\mu^2)} \right]$$

and

$$\begin{aligned} \log p(\sigma_\mu^2|(x), \sigma_S) &\propto \sum_{h=1}^{H} \left[-\frac{(I_h - 1)\log(\frac{\sigma_S^2}{K_h} + \sigma_\mu^2)}{2} - \sum_{i=1}^{I_h} \frac{(\bar{x}_{hi} - \mu_h)^2}{2(\sigma_{hi}^2 + \sigma_\mu^2)} \right] \\ &= -\frac{1}{2} \sum_{h=1}^{H} \left[(I_h - 1)\log(\frac{\sigma_S^2}{K_h} + \sigma_\mu^2) + SS_h \right] , \end{aligned}$$

where

$$SS_h = \sum_{i=1}^{I_h} \frac{(\bar{x}_{hi\cdot\cdot} - \mu_h)^2}{\sigma_{hi}^2 + \sigma_\mu^2} .$$

Thus σ_μ^2 can be updated at step t using a Metropolis-Hastings step with the proposal generating distribution for σ_μ^* being a Gaussian distribution centred at σ_μ^{t-1}, and truncated at the limits of the Uniform prior, i.e.

$$\sigma_\mu^* \sim \mathcal{N}(\sigma_\mu^{t-1}, \tau_\mu); 0 < \sigma_\mu^* < \max[p(\sigma_\mu^2)] .$$

The variance of the proposal generating distribution τ^2 must be adapted during the burn-in period of the MCMC process to ensure suitable coverage of the posterior probability space.

A.3 The Probe-Level Terms

The probe-level term p_{jk} is defined as being nested within each exon j, however as exon j is not explicitly included in any associated hyperparameter, the subscript j is omitted in all derivations below for increased simplicity, leaving $k = 1, 2, \ldots, K$.

For the probe-level terms p_k and σ_p, all other terms can be taken as known. Given the original equation

$$\log S_{hijk} = c_{hi} + p_k + \log \phi_{hj} + \varepsilon_{hijk} \,,$$

we can redefine our dummy variable to be

$$\begin{aligned} x_{hijk} &= \log S_{hijk} - c_{hi} - \log \phi_{hj} \\ \Rightarrow x_{hijk} &\sim \mathcal{N}(p_k, \sigma_S^2) \\ \Rightarrow \bar{x}_{\ldots k} &\sim \mathcal{N}(p_k, \sigma_k^2) \,, \end{aligned}$$

where

$$\begin{aligned} \bar{x}_{\ldots k} &= \frac{1}{I_k} \sum_{h \in H : \phi_{hj} \neq 0} \sum_{i=1}^{I_h} x_{hijk} \\ \sigma_k^2 &= \frac{\sigma_S^2}{I_k} \\ I_k &= \sum_{h \in H : \phi_{hj} \neq 0} I_h \,, \end{aligned}$$

additionally noting that under the Uniform Model $I_k = I$.

A.3.1 Updating the Probe Effects (p_k)

Following the same principles as Section A.2.1, and as per Gelman et al., 2004, pp. 132–136, given the prior $p_k \sim \mathcal{N}(0, \sigma_p^2)$ the posterior distribution for our unknown vector $\boldsymbol{\theta} = (p_1, p_2, \ldots, p_k); k = 1, 2, \ldots, K$ can be expressed as

$$p_k | \sigma_p, \boldsymbol{x} \sim \mathcal{N}(\hat{p}_k, V_k) \,, \tag{A.21}$$

where

$$\begin{aligned} V_k &= \frac{1}{\frac{1}{\sigma_k^2}+\frac{1}{\sigma_p^2}} \\ &= \frac{\sigma_p^2\sigma_S^2}{I_k\sigma_p^2+\sigma_S^2} \end{aligned}$$

and

$$\begin{aligned} \hat{p}_k &= \frac{\frac{\bar{x}_{\cdots k}}{\sigma_k^2}}{\frac{1}{\sigma_k^2}+\frac{1}{\sigma_p^2}} \\ &= \frac{I_k V_k \bar{x}_{\cdots k}}{\sigma_S^2}\,. \end{aligned}$$

Thus each value $p_k \in \boldsymbol{\theta}$ can be updated using a Gibbs Sampler.

A.3.2 Updating the Probe-Level Variance (σ_p^2)

Firstly, we know that the following equality holds

$$p(\sigma_p^2|\boldsymbol{x}) \propto p(\sigma_p^2)p(\boldsymbol{x}|\sigma_p^2)\,.$$

We can also see that since

$$\bar{x}_{\cdots k}|p_k \sim \mathcal{N}(p_k, \sigma_k^2)$$

and since

$$\begin{aligned} p_k &\sim \mathcal{N}(0, \sigma_p^2) \\ \Rightarrow \bar{x}_{\cdots k}|\sigma_p^2 &\sim \mathcal{N}(0, \sigma_k^2+\sigma_p^2)\,, \end{aligned}$$

this leads to the posterior

$$\begin{aligned} p(\sigma_p^2|\boldsymbol{x}) &\propto p(\sigma_p^2)\prod_{k=1}^{K}\mathcal{N}(0,\sigma_k^2+\sigma_p^2) \\ &= p(\sigma_p^2)\prod_{k=1}^{K}\frac{1}{\sqrt{2\pi(\sigma_k^2+\sigma_p^2)}}\exp\left[-\frac{\bar{x}_{\cdots k}^2}{2(\sigma_k^2+\sigma_p^2)}\right]\,. \end{aligned}$$

Assuming a Uniform Prior for $p(\sigma_p^2)$ then gives

$$p(\sigma_p^2|\boldsymbol{x}) \propto \prod_{k=1}^{K} \frac{1}{\sqrt{\sigma_k^2 + \sigma_p^2}} \exp\left[-\frac{\bar{x}_{\cdots k}^2}{2(\sigma_k^2 + \sigma_p^2)}\right] .$$

Thus the posterior $p(\sigma_p^2|\boldsymbol{x})$ can be updated using a Metropolis-Hastings step with log-posterior

$$\begin{aligned}\log p(\sigma_p^2|\boldsymbol{x}, \sigma_S^2) &\propto \sum_{k=1}^{K} \left[-\frac{\log(\sigma_k^2 + \sigma_p^2)}{2} - \frac{\bar{x}_{\cdots k}^2}{2(\sigma_k^2 + \sigma_p^2)}\right] \\ &= -\frac{1}{2} \sum_{k=1}^{K} \left[\log(\sigma_k^2 + \sigma_p^2) + \frac{\bar{x}_{\cdots k}^2}{\sigma_k^2 + \sigma_p^2}\right] .\end{aligned}$$

A proposal value σ_p^* can be generated at iteration t using a Gaussian Random Walk centred at σ_p^{t-1}, truncated at the limits of the Uniform Prior $p(\sigma_p^2)$, .i.e.

$$\sigma_p^* \sim \mathcal{N}(\sigma_p^{t-1}, \tau_p); 0 < \sigma_p^* < \max\left[p(\sigma_p^2)\right] .$$

The variance (τ_p) of the proposal generating distribution must also be adapted during the burn-in period to ensure proper exploration of the probability space for this term.

A.4 The Exon-Level Terms (ϕ_{hj}, ξ_{hj} & q_{hj})

A.4.1 The Mixture Model

Under the Mixture Prior for ϕ_{hj} as detailed in Equation A.10 an exon can fall into one of three groups, i.e.

1. $\boldsymbol{\xi_{hj}} = (1,0,0) \Rightarrow \phi_{hj} = 1$
2. $\boldsymbol{\xi_{hj}} = (0,1,0) \Rightarrow \phi_{hj} \sim \mathcal{U}(0,1)$
3. $\boldsymbol{\xi_{hj}} = (0,0,1) \Rightarrow \phi_{hj} = 0$.

Thus before the term ϕ_{hj} can be updated, the indicator variable $\boldsymbol{\xi_{hj}}$ and the probabilities of group membership $\boldsymbol{q_{hj}} = (q^1_{hj}, q^2_{hj}, q^3_{hj})$ must be updated. This can be done in the following steps, for $n = 1, 2, 3$

1. Calculate $\boldsymbol{\hat{\gamma}_{hj}} = (\hat{\gamma}^1_{hj}, \hat{\gamma}^2_{hj}, \hat{\gamma}^3_{hj})$ where

$$\begin{aligned}\hat{\gamma}^n_{hj} &= E\left(\xi^n_{hj} | \boldsymbol{PM_{hj}}, \boldsymbol{q_{hj}}, \boldsymbol{\psi}\right)\\ &= \frac{q^n_{hj} \prod_{i=1}^{I_h} \prod_{k=1}^{K_j} f^n(PM_{hijk}|\boldsymbol{\psi})}{\sum_{n=1}^{3} q^n_{hj} \prod_{i=1}^{I_h} \prod_{k=1}^{K_j} f^n(PM_{hijk}|\boldsymbol{\psi})}\end{aligned}$$

2. Sample $\boldsymbol{\xi^t_{hj}} \sim$ Multinomial$(1, \boldsymbol{\hat{\gamma}_{hj}})$
3. Sample new values for $\boldsymbol{q^t_{hj}} \sim$ Dirichlet $(\boldsymbol{\xi^t_{hj}} + 1)$.

The functions $f^n(PM_{hijk}|\boldsymbol{\psi})$ are defined in Section A.1.2 as Equation A.15.

A.4.2 Updating the Exon Proportion Term (ϕ_{hj})

Under the Mixture Model, the value ϕ_{hj} can only take the values 0 or 1 if $\xi^2_{hj} = 0$. However, if $\xi^2_{hj} = 1$, the updating method for the Mixture Model and the Uniform Model become identical. In all following equations for this updating step, this will be presented in the context of the Uniform model.

From Equation A.18, we can see that

$$\begin{aligned} x_{hijk} &\sim \mathcal{N}\left(\log \phi_{hj}, \sigma_S^2\right) \\ \bar{x}_{h\cdot j\cdot} &\sim \mathcal{N}\left(\log \phi_{hj}, \sigma_{hj}^2\right) \end{aligned}$$

where

$$\begin{aligned} x_{hijk} &= \log S_{hijk} - c_{hi} - p_k \\ \bar{x}_{hj} &= \frac{1}{I_h}\sum_{i=1}^{I_h}\frac{1}{K_j}\sum_{k=1}^{K_j} x_{hijk} \\ \sigma_{hj}^2 &= \frac{\sigma_S^2}{I_h K_j}\,. \end{aligned}$$

Thus for $\theta = -\log \phi_{hj} \sim$ Exponential(1),

$$\begin{aligned} f(\theta|\bar{x}_{hj}) &\propto f(\theta)f(\bar{x}_{hj}|\theta) \\ &\propto \exp\left(-\theta\right)\exp\left(-\frac{(\bar{x}_{hj}+\theta)^2}{2\sigma_{hj}^2}\right) \\ &= \exp\left(\frac{-1}{2\sigma_{hj}^2}\left[2\sigma_{hj}^2\theta + (\bar{x}_{hj}+\theta)^2\right]\right)\,. \end{aligned}$$

Taking the exponent terms and completing the square

$$\begin{aligned} 2\sigma_{hj}^2\theta + (\bar{x}_{hj}+\theta)^2 &= \bar{x}_{hj}^2 + 2\bar{x}_{hj}\theta + \theta^2 + 2\sigma_{hj}^2\theta \\ &= \bar{x}_{hj}^2 + 2(\bar{x}_{hj}+\sigma_{hj}^2)\theta + \theta^2 + 2\bar{x}_{hj}\sigma_{hj}^2 - 2\bar{x}_{hj}\sigma_{hj}^2 + (\sigma_{hj}^2)^2 - (\sigma_{hj}^2)^2 \\ &= (\bar{x}_{hj}+\sigma_{hj}^2)^2 + 2(\bar{x}_{hj}+\sigma_{hj}^2)\theta + \theta^2 - 2\bar{x}_{hj}\sigma_{hj}^2 - (\sigma_{hj}^2)^2 \\ &\propto (\bar{x}_{hj}+\sigma_{hj}^2)^2 + 2(\bar{x}_{hj}+\sigma_{hj}^2)\theta + \theta^2 \\ &= (\bar{x}_{hj}+\sigma_{hj}^2+\theta)^2 \\ &= (\theta - (-\bar{x}_{hj}-\sigma_{hj}^2))^2\,. \end{aligned}$$

Returning to the full equation gives

$$f(\theta|\bar{x}_{hj}) \propto \exp\left(\frac{-1}{2\sigma_{hj}^2}(\theta - (-\bar{x}_{hj}-\sigma_{hj}^2))^2\right)$$

which we can recognise as the kernel of a normal distribution. Thus

$$\theta|\bar{x}_{hj} \sim \mathcal{N}(-\bar{x}_{hj} - \sigma^2_{hj}, \sigma^2_{hj}).$$

Substituting $-\log \phi_{hj}$ for θ, and taking note of the range restriction for $\log \phi_{hj}$ gives

$$-\log \phi_{hj}|\bar{x}_{hj} \sim \mathcal{N}(-\bar{x}_{hj} - \sigma^2_{hj}, \sigma^2_{hj}); 0 < -\log \phi_{hj} < \infty\,.$$

By symmetry, we can see

$$\log \phi_{hj}|\bar{x}_{hj} \sim \mathcal{N}(\bar{x}_{hj} + \sigma^2_{hj}, \sigma^2_{hj}); -\infty < \log \phi_{hj} < 0$$

and a Gibbs Sampler can be used to update this term for both the Uniform Model, and the Mixture Model given $\xi^2 = 1$.

A.5 Signal Variance

Assuming all other terms are known, by Equations 3.6b and 3.7 we know that

$$p(S|\eta, \sigma_S) = \frac{1}{S\sigma_S\sqrt{2\pi}} \exp\left[-\frac{(\log S - \eta)^2}{2\sigma_S^2}\right]$$

with the prior for σ_S^2 being defined as the Jeffreys Prior in Equation 3.8. Thus, the posterior for $\sigma_S^2|\boldsymbol{S}, \boldsymbol{\eta}$ can be seen to be

$$p(\sigma_S^2|\boldsymbol{S}, \boldsymbol{\eta}) \propto p(\sigma_S^2) \prod_{h=1}^{H} \prod_{i=1}^{I_h} \prod_{\{j \in J : \phi_{hj} \neq 0\}} \prod_{k=1}^{K_j} \frac{1}{S_{hijk}\sqrt{2\pi\sigma_S^2}} \exp\left[-\frac{1}{2\sigma_S^2}(\log S_{hijk} - \eta_{hijk})^2\right] 7.. \tag{A.22}$$

Defining the values

$$\begin{aligned} \nu &= \sum_{h=1}^{H} I_h \sum_{\{j \in J : \phi_{hj} \neq 0\}} K_j \\ s^2 &= \frac{1}{\nu} \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{\{j \in J : \phi_{hj} \neq 0\}} \sum_{k=1}^{K_j} (\log S_{hijk} - \eta_{hijk})^2, \end{aligned}$$

Equation A.22 then can be seen to be

$$\begin{aligned} p(\sigma_S^2|\boldsymbol{S}, \boldsymbol{\eta}) &\propto p(\sigma_S^2)(\sigma_S^2)^{\frac{\nu}{2}} \exp\left[-\frac{\nu s^2}{2\sigma_S^2}\right] \\ &= (\sigma_S^2)^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu s^2}{2\sigma_S^2}\right] \end{aligned}$$

which can be recognised as a Scaled Inverse χ^2 distribution. Thus

$$\sigma_S^2|\boldsymbol{S}, \boldsymbol{\eta} \sim \text{Scaled Inv-}\chi^2(\nu, s^2) \tag{A.23}$$

and σ_S^2 can be updated using a Gibbs Sampler.

A.6 Parameter Initialisation

Before any parameter updating is able to take place, initial values must be generated for each of the above parameters. Considering the reliance of some parameters on other values, the order of initialisation is also important. In the following description, the superscript ^{0} indicates the initial value for each parameter.

A.6.1 S_{hijk}

For each probe, set the initial signal estimate to be

$$S^0_{hijk} = PM_{hijk} - B^0_{hijk}\,. \tag{A.24}$$

The initial values for background signal can be sampled directly from the prior using

$$\log B^0_{hijk} \sim \mathcal{N}(\lambda_{hijk}, \delta_{hijk}); \log B_{hijk} < \log PM_{hijk}\,.$$

A.6.2 σ_S

Take a random sample from a locally Uniform distribution

$$\sigma^0_S \sim \mathcal{U}(0, 10)\,. \tag{A.25}$$

A.6.3 c_{hi}

For each array $i : 1 \leq i \leq I$, randomly sample an integer $M : 1 \leq M \leq K$ then set

$$c^0_{hi} = \log S^0_{hijM}\,. \tag{A.26}$$

A.6.4 μ_h

For each condition $h : 1 \leq h \leq H$ set

$$\mu^0_h = \frac{1}{I_h} \sum_{i=1}^{I_h} c^0_{hi}\,. \tag{A.27}$$

A.6.5 σ_μ

Set

$$\sigma_\mu^0 = \sqrt{\frac{1}{I-1}\sum_{h=1}^{H}\sum_{i=1}^{I_h}(c_{hi}^0 - \mu_h^0)^2}\,. \tag{A.28}$$

A.6.6 ϕ_{hj}

For each exon $hj : 1 \leq h \leq H; 1 \leq j \leq J$, take a random sample from the prior

$$\phi_{hj}^0 \sim \mathcal{U}(0,1)\,. \tag{A.29}$$

Note that this method was chosen even for the Mixture Model, for simplicity. The tendency of the model towards convergence was relied upon for resolving any MCMC runs where the true nature of the data was highly dissimilar to these initial values.

A.6.7 p_k

For each probe $k : 1 \leq k \leq K$ set

$$p_k^0 = \frac{1}{I}\sum_{h=1}^{H}\sum_{i=1}^{I_h}\log S_{hijk}^0 - c_{hi}^0 - \log\phi_{hj}^0\,. \tag{A.30}$$

A.6.8 σ_p

Set

$$\sigma_p^0 = \frac{1}{K-1}\sqrt{\sum_{k=1}^{K}(p_k^0)^2}\,. \tag{A.31}$$

A.6.9 q_{hj}

For the Mixture model only, the probabilities of group membership $\boldsymbol{q_{hj}^0}$ must also be initialised for each combination of $hj : 1 \leq h \leq H; 1 \leq j \leq J$. This is done by sampling from the prior

$$\boldsymbol{q_{hj}^0} \sim \text{Dirichlet}(3, \boldsymbol{\alpha}),\text{ where } \boldsymbol{\alpha} = (1,1,1)\,. \tag{A.32}$$

A.7 MCMC Process Summary

A.7.1 Uniform Model

After initialisation, the Uniform model was simply run by stepping through each of the parameters using the sampling methods described in Sections A.1 to A.5. The order of updating is not important.

A.7.2 Mixture Model

The process for the Mixture model requires sampling of group membership at the beginning of each iteration as described in Section A.4.1, and as such, the process of updating parameters at iteration $t : t \geq 1$ is:

1. Beginning at $h = 1; j = 1$,calculate $\boldsymbol{\hat{\gamma}_{hj}} = (\hat{\gamma}^1_{hj}, \hat{\gamma}^2_{hj}, \hat{\gamma}^3_{hj})$ where

$$\begin{aligned}\hat{\gamma}^n_{hj} &= E(\xi^n_{hj}|PM, \boldsymbol{q^{t-1}_{hj}}, \boldsymbol{\psi^{t-1}}) \\ &= \frac{q^n_{hj}\prod_{i=1}^{I_h}\prod_{k=1}^{K_j} f^n(PM_{hijk}|\boldsymbol{\psi^{t-1}})}{\sum_{n=1}^{3} q^n_{hj}\prod_{i=1}^{I_h}\prod_{k=1}^{K_j} f^n(PM_{hijk}|\boldsymbol{\psi^{t-1}})}\end{aligned}$$

2. Sample $\boldsymbol{\xi^t_{hj}} \sim \text{Multinomial}(1, \boldsymbol{\hat{\gamma}_{hj}})$

3. Sample new values for $\boldsymbol{q^t_{hj}} \sim$ Dirichlet $(3, \boldsymbol{\xi^t_{hj}} + 1)$

4. Update all values of $(S^t_{hijk}|PM, \boldsymbol{\xi^t_{hj}}, \boldsymbol{\psi^{t-1}})$ (for $h = 1; j = 1$)

5. Update $(\phi^t_{hj}|S^t, \boldsymbol{\xi^t_{hj}}, \boldsymbol{\psi^{t-1}})$

6. Repeat for all values of $(h : 1 < h \leq H)$ & $(j : 1 < j \leq J)$

7. Update all remaining values as described in Sections A.1 to A.5.

Appendix B

BMEA Package Design and Performance

B.1 Package Design

B.1.1 The MCMC Process

The MCMC process itself was written in the language C, leaving R as the user interface via the package *BMEA*, giving an $\sim$ 8-fold improvement in computational time when compared to the pure R code of Section 3.3.3. Instead of running separate chains in parallel, the package was written to break the dataset into batches of genes, with each batch being fitted in parallel. All parameter updates were as defined in Appendix A.

B.1.2 Aroma Affymetrix Conventions

As the BMEA process is based around generating posterior distributions for the parameters of interest, this equates to a huge amount of data. Given the number of genes is in the tens of thousands, and the number of parameters fitted for a given gene will often be > 100 with many thousands of sampled values during the MCMC process, great care was required to determine what should be retained in the output of the process.

The architecture of the *aroma.affymetrix* package (Bengtsson et al., 2008) was taken

```
parentDirectory
 |__annotationData
 |   |__chipTypes
 |       |__HuEx-1_0-st-v2
 |           |__HuEx-1_0-st-v2.CDF
 |__rawData
     |__experimentName
         |__HuEx-1_0-st-v2
             |__myData1.CEL, myData2.CEL etc.
```

Figure B.1 – *Minimal initial directory structure required to begin an analysis using the package* `aroma.affymetrix`*, with data generated on Human Exon 1.0 ST arrays. Directory and file names given in italics are able to be changed as required for an individual analysis, with the remaining paths being the rigid directory structure required.*

as the basis for *BMEA* and extended to store the required information. This package uses a rigid directory structure, with a minimal directory structure shown in Figure B.1. Under this structure, the CDF file for HuEx-1_0-st-v2 arrays is placed in the `annotationData/chipTypes/HuEx-1_0-st-v2` folder, and so on for every chip type under investigation. The raw data, in the form of .CEL files, is likewise stored in the folder `rawData/experimentName/HuEx-1_0-st-v2`, where `experimentName` is the working name of the current experiment. The name of the chip-type (HuEx-1_0-st-v2) can also be changed, but must match the name of the supplied CDF exactly.

In addition to this directory structure, a reduced .CEL file type is used by this package, referred to as a *monocell* .CEL file. A corresponding monocell CDF is generated in the initial stages of an analysis, and written to the folder `annotationData/chipTypes/ HuEx-1_0-st-v2`. This file structure allows only one entry (i.e. cell) per exon, making it very efficient for writing summary values at the probeset (i.e. gene or exon) level, and giving a much smaller file size than the full .CEL file structure, which contains probe-level information. This file format is used by *aroma.affymetrix* to write gene-level expression estimates to disk using the directory structure `plmData/exptName/HuEx-1_0-st-v2`.

These structures and file-types are stored in R as objects using the `S3` class `AffymetrixCelSet`, which is used to read information from a collection of .CEL files into R. As array analysis often involves multiple processing steps, the use of *tags* is utilised by *aroma.affymetrix* by the addition of a tag after a comma as the directory `exptName,tag`, to denote any steps that have been performed on the .CEL files in the lower-level directories.

B.1.3 BMEA Extensions to Aroma Affymetrix Structures

The above conventions were extended in BMEA in order to store values for background priors, and the key values from posterior distributions. Values for the Background Signal priors were treated in the same manner as the raw data files, and were stored in a probe-specific manner, using .CEL files matching the original "rawData" directory structure, using the directories named `backgroundPriors/experimentName,lambda` and `backgroundPriors/experimentName,delta/` (Figure B.2). All values were estimated in an array specific manner as described in Section 3.4.1, effectively allowing the notation λ_{hijk} and δ_{hijk}, instead of requiring the additional indexing subscript l to indicate which bin the probe belongs to. As the location (λ) and scale (δ) parameters were able to be defined as separate `AffymetrixCelSet` objects, the new informal `S3` class `AffymetrixCelSetList` was introduced to collect these related parameter files into single objects within the R environment.

For posterior distributions, all values are written to the directory `bmeaData`, with two subdirectories `bmeaData/modelData` and `bmeaData/contrastData`. Instead of saving all MCMC sampled values, by default only the $0.025, 0.25, 0.5, 0.75$ and 0.975 quantiles were saved, along with the posterior mean and the convergence statistic $\hat{r}$ (Gelman et al., 2004), for any model parameter chosen for saving (Figure B.2). All values for a saved parameter (e.g. μ) were written to the directory `bmeaData/modelData/experimentName,parameterName`, again using the tag system of naming directories. The same sets of summary statistics for any contrasts were written to the directories `bmeaData/contrastData/experimentName,logFC` and `bmeaData/contrastData/experimentName,phiLogFC`, with the addition of the B-statistic. In all cases, the `AffymetrixCelSetList` object type was used to manage the relationship between the directory structure and the R environment.

For chip, condition, exon or any other grouped term, *monocell* .CEL files are used to store the summary statistics, with only probe-level values requiring the full .CEL file structure. All model parameters are able to be saved, with only the chip effects (c_i), condition-specific expression levels (μ_h), exon proportions ϕ_{hj} saved by default, in the interests of minimising disk storage, and minimising disk writing times whilst the process was running. Both logFC ($\Delta\mu$) and phiLogFC ($\Delta \log \phi_j$) are also saved by default.

```
parentDirectory
├── annotationData
│   └── chipTypes
│       └── cdfName
│           └── cdfName.CDF
├── rawData
│   └── experimentName
│       └── cdfName
│           └── myData1.CEL, myData2.CEL, ... etc
├── backgroundPriors
│   ├── experimentName,lambda
│   │   └── cdfName
│   │       └── myData1.CEL, myData2.CEL ... etc
│   └── experimentName,delta
│       └── cdfName
│           └── myData1.CEL, myData2.CEL ... etc
└── bmeaData
    ├── modelData
    │   ├── experimentName,c
    │   │   └── cdfName
    │   │       └── myData1,2.5%.CEL, myData1,25%.CEL, myData1,50%.CEL,
    │   │           myData1,75%.CEL, myData1,97.5%.CEL, myData1,mean%.CEL,
    │   │           myData1,sd%.CEL, myData1,rHat.CEL, myData2,2.5%.CEL etc.
    │   ├── experimentName,mu
    │   │   └── cdfName
    │   │       └── group1,2.5%.CEL, group1,25%.CEL, ... etc
    │   └── experimentName,phi
    │       └── cdfName
    │           └── group1,2.5%.CEL, group1,25%.CEL, ... etc
    └── contrastData
        ├── experimentName,logFC
        │   └── cdfName
        │       └── cont1,2.5%.CEL, cont1,25%.CEL, ... etc
        └── experimentName,phiLogFC
            └── cdfName
                └── cont1,2.5%.CEL, cont1,25%.CEL, ... etc
```

Figure B.2 – *Example directory structure as generated by BMEA after specification of the minimal structure from Figure B.1. Names shown in italics are set by the user, with group and contrast names defined within R and propagated through all directories. The experiment name, CEL file names and the CDF name are sourced from the original data structure and are also propagated through all directories.*

B.2 Algorithm Performance

The recovery of simulated parameters from Chapter 4 is given in the following sections, with assessment of the technical performance of the package, in terms of computational time and convergence statistics given in Section B.3.

B.2.1 Recovery of Simulation Parameters

All fitted values for μ_h, ϕ_{hj}, σ_μ, σ_p and σ_S were compared to the true values set during generation of the simulated data (Table B.1). Posterior means were used as point estimates for μ_h with posterior medians being used for all other parameters. Performance was assessed for both BMEA and BMEA-Z, with the latter utilising Z_g and Z_j to remove undetectable genes and exons respectively. Pearson correlations were calculated, and bias was found as defined by

$$\text{Bias}[\hat{\theta}] = E_{x|\theta}[\hat{\theta} - \theta]\,.$$

Comparison between the BMEA and BMEA-Z models revealed no specific trend of higher or lower correlations with the inclusion of the Z-score steps (Table B.1). However, with the inclusion of Z-scores the bias was slightly greater for all parameters except σ_S.

Expression Levels

Fitted Vs Simulated values for μ_h are shown in Figure B.3. Across all patterns and simulated cell-types, simulated and fitted values for expression levels showed a high degree of concordance ($0.922 < \rho < 0.944$; Table B.1). Notably, without the use of Z-scores, simulations using the splicing pattern with 40% of probes containing no true signal (Pattern 3) showed a clear downward bias in expression estimates. This was not particularly unexpected, as the expression level would have effectively been estimated as $\sim 0.6\mu$ for these data points. The alternate model incorporating Z-score filtering of exons showed none of this bias. Taking posterior means as representative of point estimates, the remainder of simulated genes showed an overall positive bias for μ_h which was most evident at the low end of the expression range.

Table B.1 – *Comparison between fitted values under BMEA and simulated values for continuous variables. Fitted values for the mean expression-level (μ_h) were taken as the posterior means, whilst fitted values for variance terms were taken as the posterior medians. Correlations between fitted and simulated values are shown (ρ), with the bias provided as $E_{x|\theta}[\hat{\theta} - \theta]$. Positive values for bias indicate that fitted values are larger on average than the true values, with the converse being true for negative values.*

Parameter	**Model**	ρ	**Bias**$[\hat{\theta}]$
μ_h	BMEA	0.922	1.121
	BMEA-Z	0.944	1.277
ϕ_{hj}	BMEA	0.719	-0.313
	BMEA-Z	0.708	-0.368
σ_μ	BMEA	0.453	0.121
	BMEA-Z	0.457	0.133
σ_p	BMEA	0.625	-0.372
	BMEA-Z	0.672	-0.417
σ_S	BMEA	0.738	-0.148
	BMEA-Z	0.765	-0.139

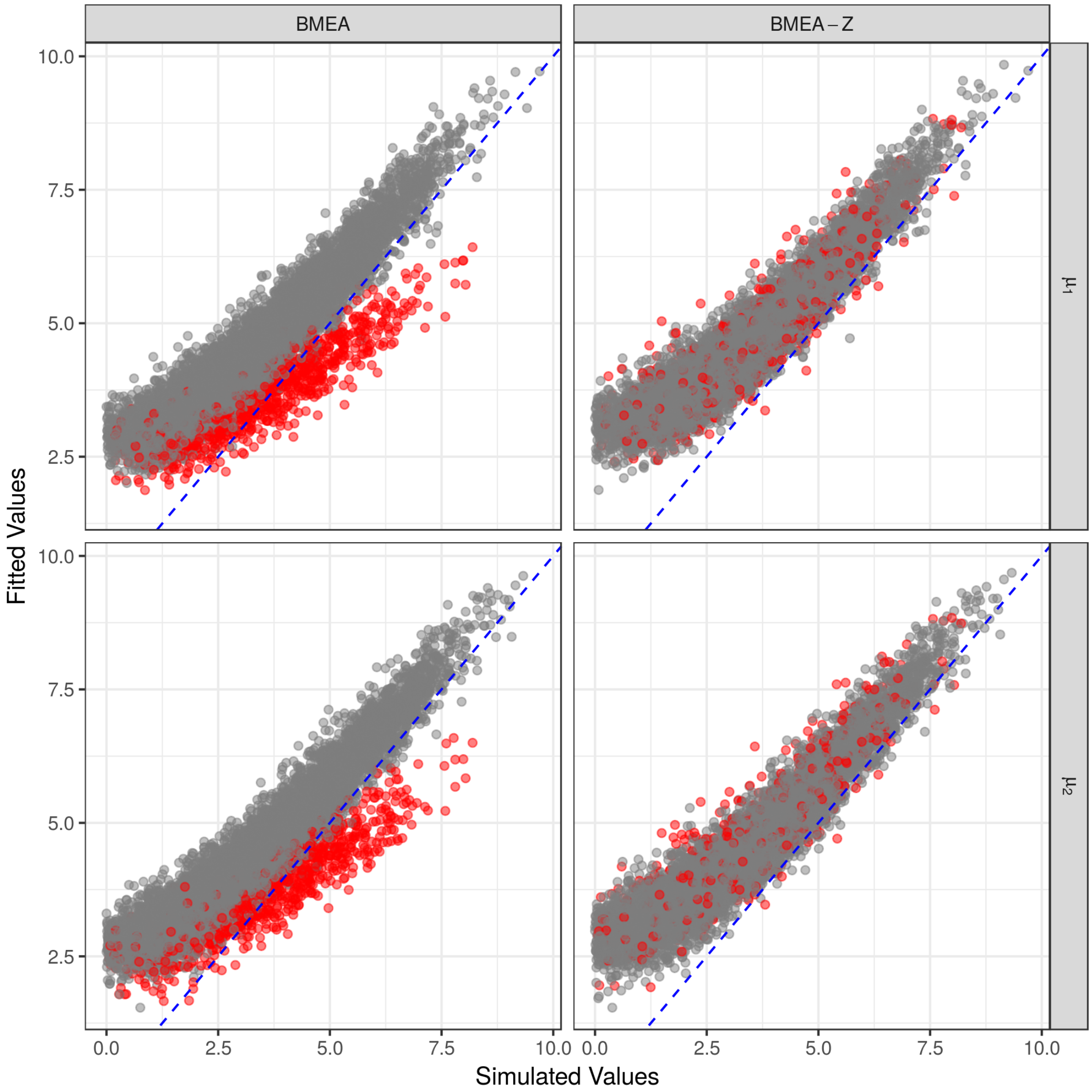


Figure B.3 – *Fitted values for μ compared to simulated values. Posterior means were used as point estimates for fitted values. Genes omitted using Z-scores prior to fitting are not shown. The dashed line represents the line $y = x$. Simulations with splicing pattern 3 are coloured red.*

Variance Components

The three variance components of the model (σ_μ, σ_p & σ_S) showed more variable results than for expression levels (Table B.1). For variance between samples within an experimental condition (σ_μ) fitted values varied widely around the simulated values by a factor of up to 5-fold, showing a generally upwards bias in the fitted values, but with a positive correlation between the two sets of values (Table B.1).

Correlations between simulated values and posterior medians were higher for probe-level variance (σ_p) than for σ_μ. A general downwards bias was also noted, with the strongest downwards bias evident amongst those data points with the lowest expression values (Figure B.4). Of the variance terms, the highest correlations between fitted and simulated values were found for the overall signal-level variance term (σ_S). Once again, an overall downwards bias was evident for the fitted values, with this being the most pronounced for data points with the lowest expression values.

Exon Proportions

Comparison of fitted values for ϕ_{hj} to simulated values was performed again using posterior median values as point estimates. For splicing pattern 1, in which all exons were consistently included, the IQR for point estimates was $0.51 < \phi_{hj} < 0.68$ for BMEA and $0.49 < \phi_{hj} < 0.68$ for the model including Z-scores (Figure B.5). These values were considerably below the simulated values of $\phi_{hj} = 1$ and likely show the influence of the prior $\phi \sim \mathcal{U}(0,1)$, especially considering only 4 probes were simulated for each exon. This also largely explains the inflated estimates of μ_h noted previously. Across the remainder of the splicing patterns, the range of values for each exon generally tracked the simulated values very closely, with the model incorporating Z_j scores showing a clear influence of removed exons by setting $\phi_{hj} = 0$ for point estimates.

A clear exception to this was Pattern 3, in which the unfiltered BMEA analysis incorrectly fitted Exons 1 to 4 with values for $\phi_{hj} > 0$. The IQR for these fitted values across all simulations was $0.25 < \phi_{hj} < 0.46$, which was considerably lower than the included exons, but still far above the true value of $\phi_{hj} = 0$, which gives a clear explanation for the

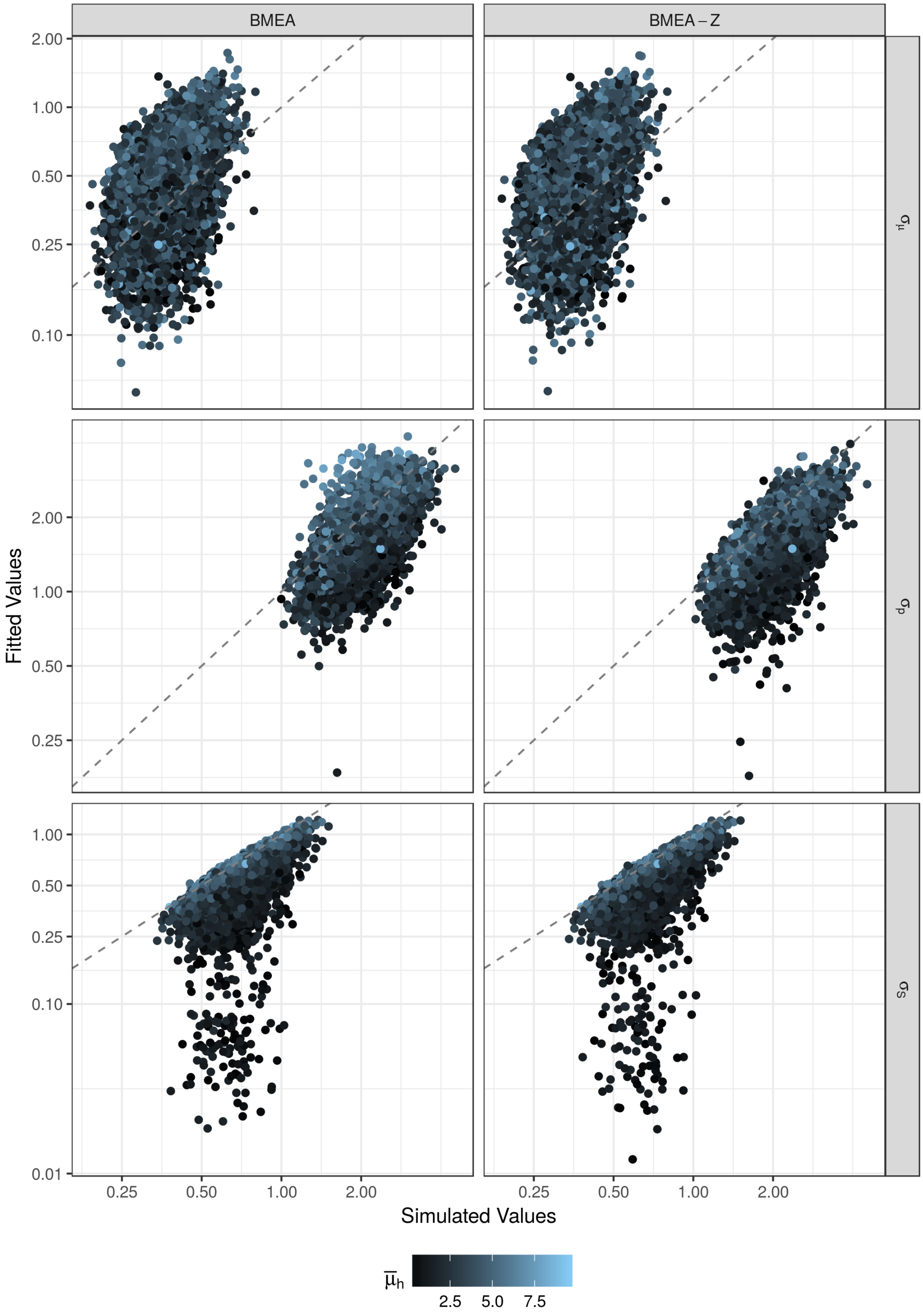


Figure B.4 – *Fitted values for σ_S, σ_p and σ_μ compared to simulated values. Posterior medians were used as point estimates for fitted values, with correlations given in Table B.1. Points are coloured based on the mean simulated expression levels ($\bar{\mu}_h$). Axes are shown on the $\log_{10}$ scale. The dashed line represents the line $y = x$. Genes omitted using Z_g-scores prior to fitting are not shown.*

behaviour seen in Figure B.3.

For patterns in which cell-type B contained a shorter transcript, patterns with a single or double exon skip (Patterns 8 and 9) showed fitted values very much in keeping with expected behaviours (Figure B.6). However a surprising artefact was noted in cell-type A for exons simulated as missing in cell-type B. In the cell-type for which these exons were included, the IQR for fitted values was higher ($0.59 < \phi_{hj} < 0.75$) using BMEA without Z_j-score filtering than for the consistently included exons. Similar, but more exaggerated behaviour was noted in Pattern 10, where higher point estimates for ϕ_{hj} were observed in cell-type A for exons simulated as missing in cell-type B. However, for exons included in both cell types for this pattern (Exons 1 to 5, 10), point estimates for ϕ_{hj} were lower in cell-type A than for cell-type B. This splicing pattern was also noted as generating considerable false positives for logFC under FIRMA (Section 4.4.1; Figure 4.6). This observation also is in keeping with the positive bias for BMEA B-statistics in Figure 4.7.

For simulations with varying proportions of Exon 3 (Patterns 6 and 7), fitted values for Exon 3 were spread across overlapping ranges, despite the different true values (Figure 4.3F and Figure 4.3G). The IQR seen in boxplots for Pattern 7 was slightly lower ($0.22 < \phi_{hj} < 0.28$) than Pattern 6 ($0.32 < \phi_{hj} < 0.40$), however the near overlap of the IQR will likely impact the accuracy of estimation of ϕ_{hj}. The impact on detection of AS events would be expected to remain minimal, however.

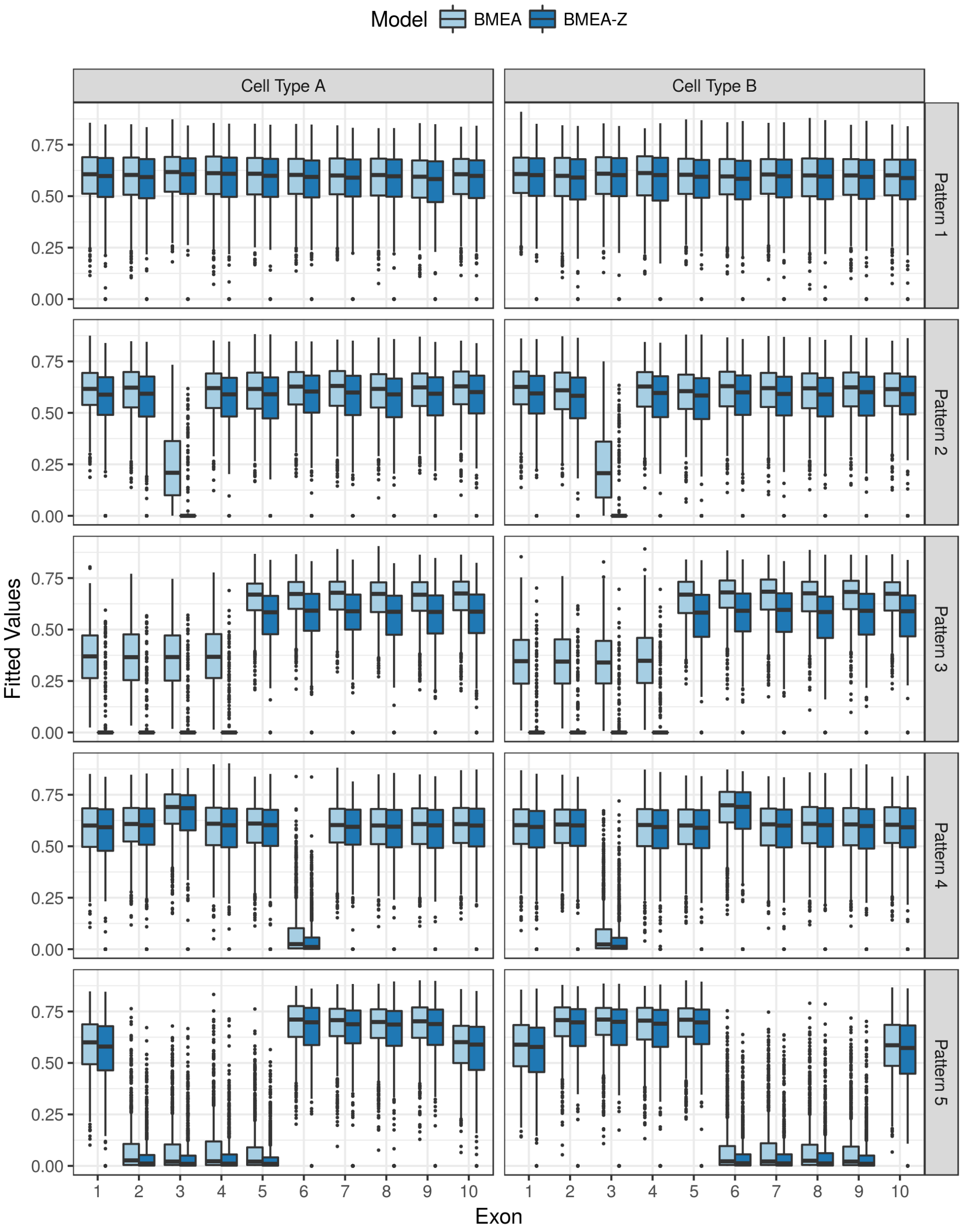


Figure B.5 – *Splicing patterns 1 to 5. Boxplot of fitted values for ϕ_{hj} compared to simulated values for each exon within each splicing pattern. Posterior medians were used as point estimates for fitted values. Exons removed during Z-score filtering are shown as being given the value $\phi_{hj} = 0$. Outlier points, based on the IQR, are shown as individual dots. All splicing patterns contain transcripts of the same length in both simulated cell-types (Figure 4.3).*

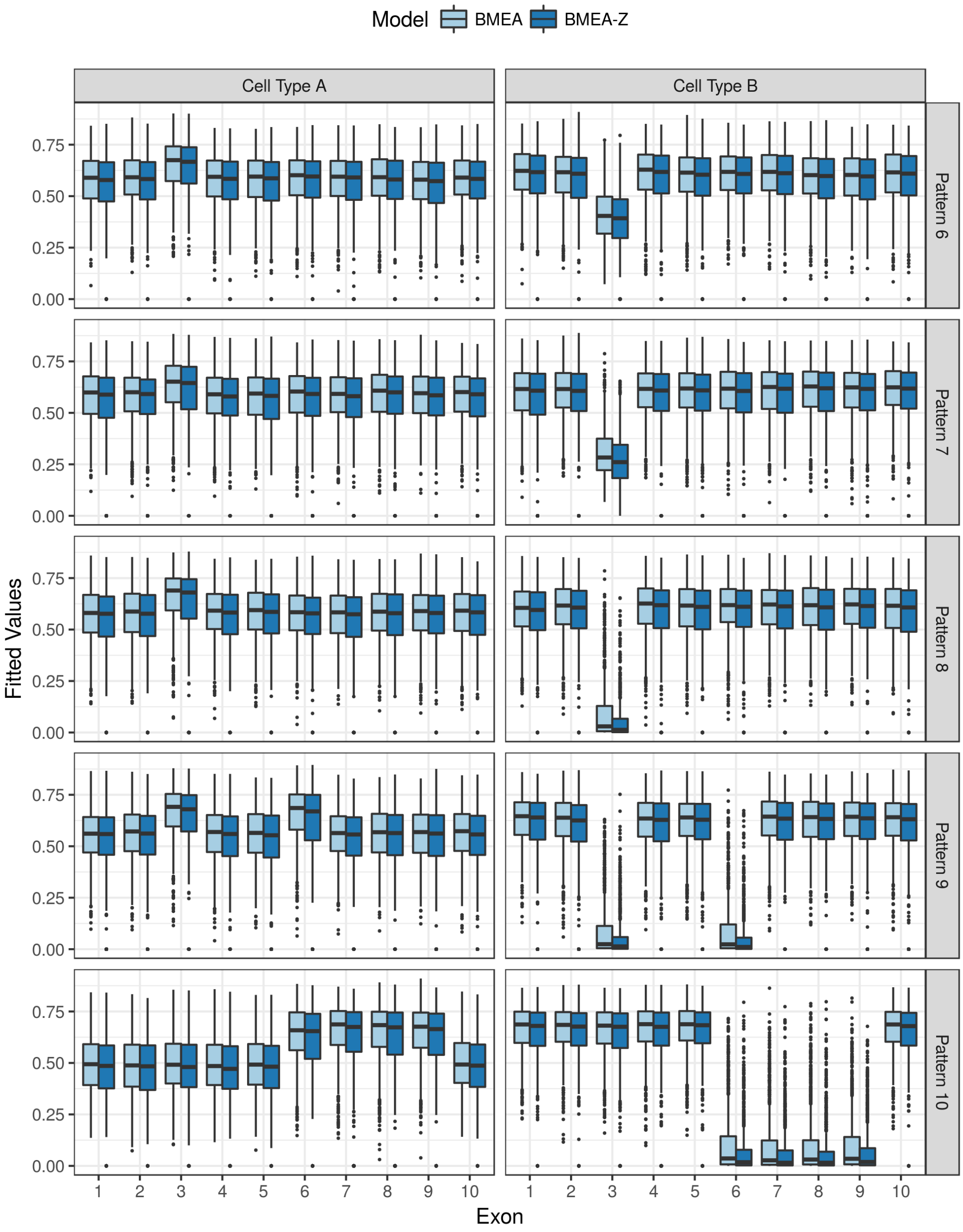


Figure B.6 – *Splicing patterns 6 to 10. Boxplot of fitted values for ϕ_{hj} compared to simulated values for each exon within each splicing pattern. Posterior medians were used as point estimates for fitted values. Exons removed during Z-score filtering are shown as being given the value $\phi_{hj} = 0$. Outlier points, based on the IQR, are shown as individual dots. Cell-type B was simulated with a shorter transcript in all splicing patterns (Figure 4.3).*

Fold-Change for Genes and Exons

In addition to the simulated model parameters, the recovery of accurate fold-change estimates was also checked both for logFC ($\Delta\mu$) and for changes in exon inclusion proportions ($\Delta \log \phi_j$). The values for the changes in expression-level are shown using posterior means as point estimates (Figure B.7), with data being transformed to the $\log_2$ scale for easier axis labelling. The difference between the BMEA approaches was minimal with correlations of $\rho_{BMEA} = 0.847$ and $\rho_{BMEA-Z} = 0.844$. For simulations with non-zero fold-change, the Bias in estimates of $|\Delta\mu|$ were -0.22 and -0.18 for the BMEA and BMEA-Z models respectively revealing a tendency to underestimate the scale of any fold change under both approaches.

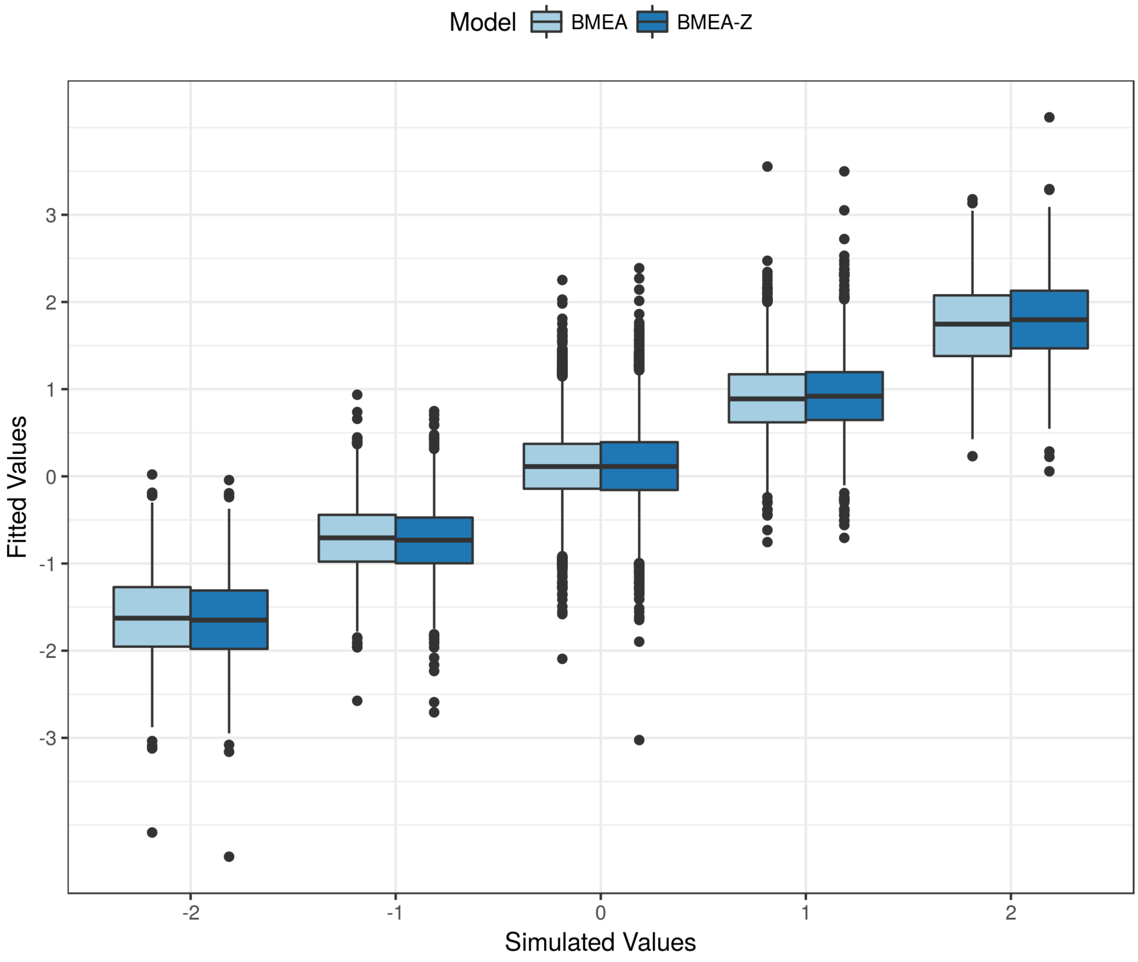


Figure B.7 – *Estimates for fold-change for each of the simulated values. Posterior means were used as point estimates for each simulated gene. Values are shown on the* $\log_2$ *scale for easier interpretation and axis-labelling.*

Assessing changes in exon proportions is a less common usage of the term logFC, and in this context the term ϕ-logFC is sometimes used interchangeably with the term $\Delta \log \phi_j$, and refers to the change in log-transformed proportions. Posterior medians were taken as point estimates for this comparison. For exons which were not simulated with any difference in proportions, the fitted values for ϕ-logFC were clearly centred around the value zero (Figure B.8).

For exon 3 in pattern 2, and the first four exons in pattern 3, which were simulated as missing in all samples, the values for ϕ-logFC were again centred around zero. Direct comparison between models is difficult for these particular exons, as the majority were removed during the Z-filtering steps for the BMEA-Z model, thus the numbers of points making up these boxplots are strongly unbalanced between the two BMEA approaches. However, for the unfiltered approach, values clearly were centred around zero, but with a more broad distribution than for those exons present in all samples, showing an increased possibility of Type I errors under this model.

For patterns 6 and 7, in which Exon 3 was simulated as being present in differing proportions, the simulated values were effectively given 2-fold and 3-fold change in proportions respectively. For the remaining exons and patterns, the exon was simulated as simply absent or present in all samples, and as such, the true numeric values for $\Delta \log \phi_j$ were either $\pm\infty$, which simply cannot be recovered under the model specification, with $\phi_{hj} \sim \mathcal{U}(0, 1)$ explicitly not permitting the distributional boundary points during the MCMC process. However, these values in general showed a great deal of variability as may be expected when true value lies outside the range of the prior, but were strongly consistent with the underlying simulated values.

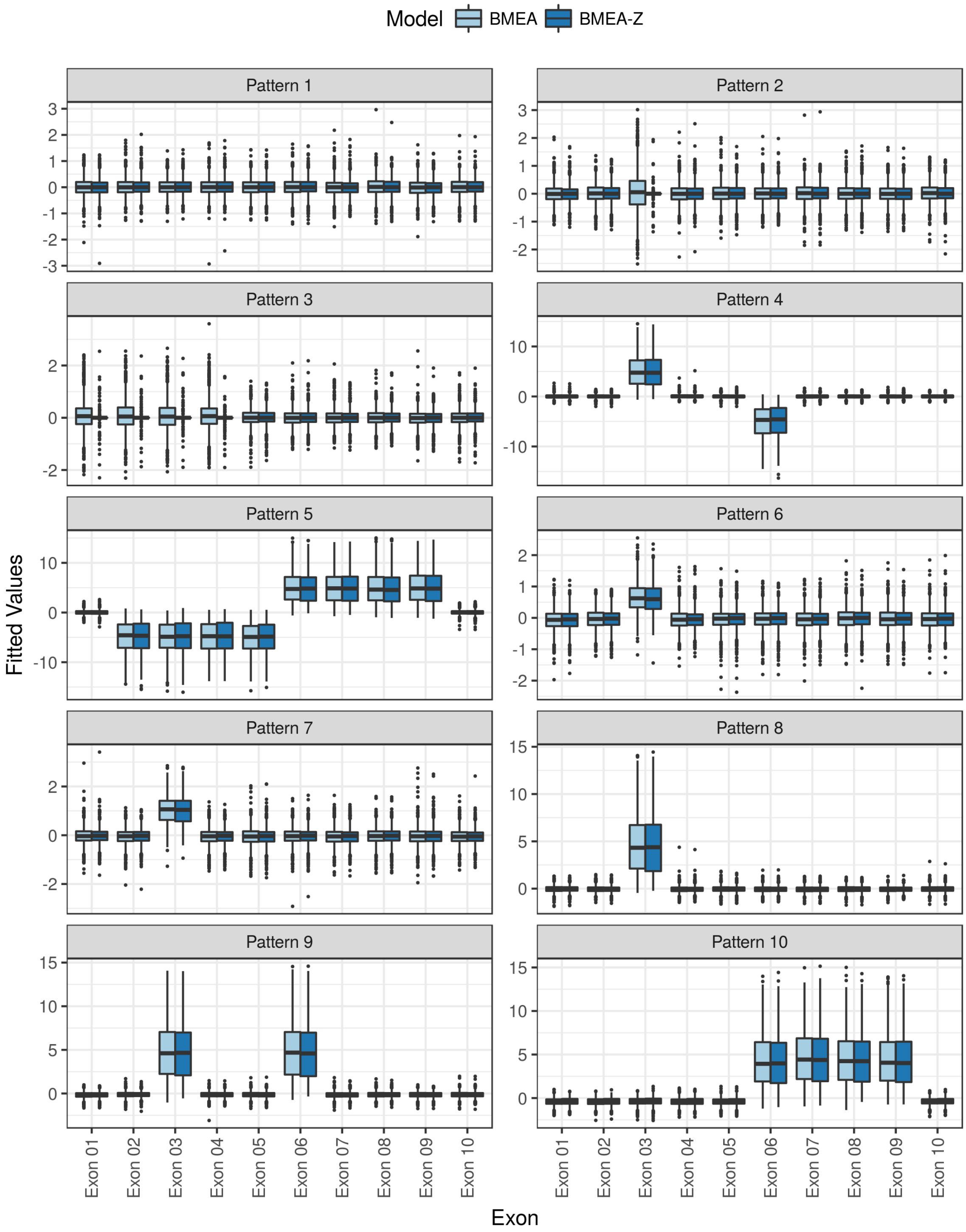


Figure B.8 – *Estimates for fold-change for each exon ($\Delta \log \phi_j$) within the simulated values. Posterior medians were used as point estimates for each exon. Values are shown on the $\log_2$ scale for easier interpretation. Outliers relative to the IQR are shown as individual points.*

B.3 Technical Performance

B.3.1 Convergence of Parameters

As well as the effectiveness of recovery for the supplied parameters, the $\hat{r}$ statistic was used to assess convergence for each parameter (Gelman et al., 2004) described in Section 4.3.2 (Figure B.9). Good rates of convergence were noted across all parameters with the majority of $\hat{r}$ values being very close to 1, as would be expected when the independent chains have converged.

The same statistics for the individual posterior distributions of the values ϕ_{hj} also showed relatively good rates of convergence(Figure B.10). However, exons which were simulated as missing were clearly unable to yield estimates which converge as strongly as for the constitutive exons, indicating longer MCMC run times may be beneficial for candidate AS exons.

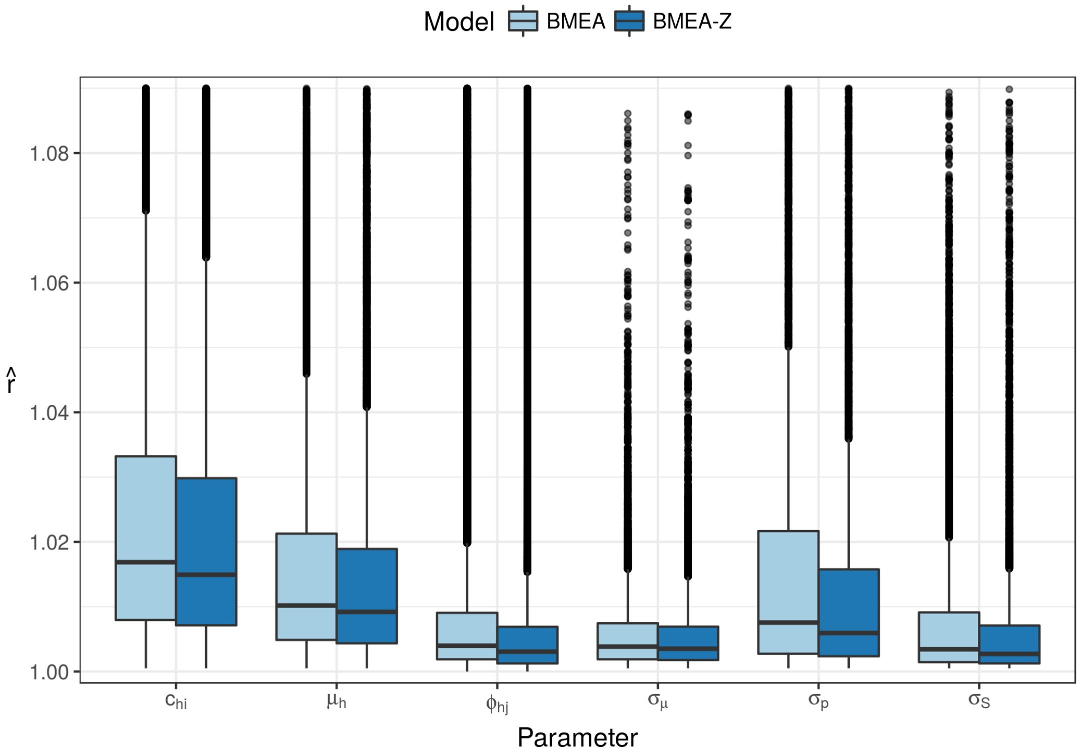


Figure B.9 – *Boxplots of the convergence statistic ($\hat{r}$) for key model parameters. Outliers beyond the limits of the y-axis have been omitted. Values approaching $\hat{r} = 1$ indicate that the model has converged for the given parameter.*

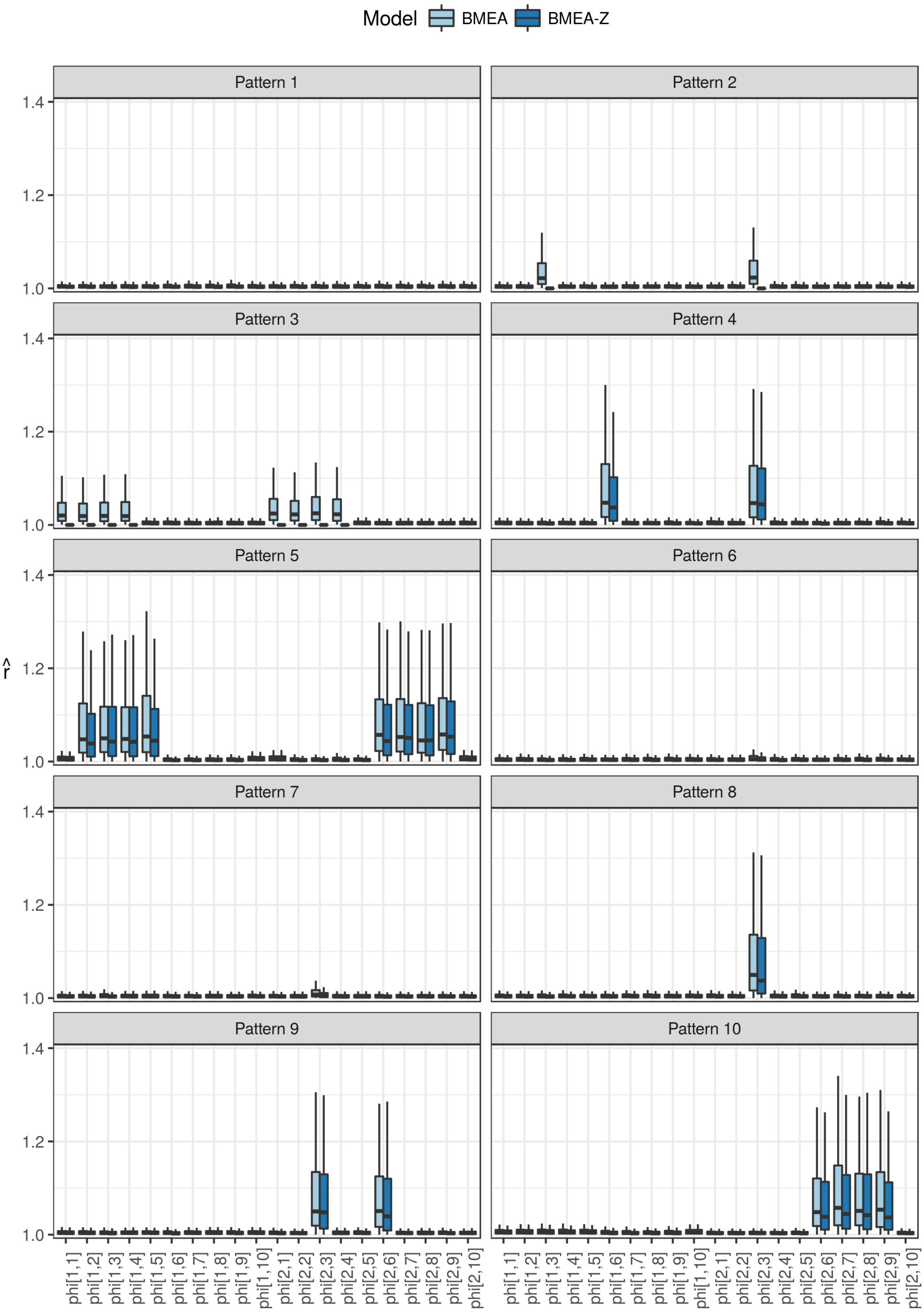


Figure B.10 – *Boxplots of the convergence statistic ($\hat{r}$) for the values ϕ_{hj}, broken down by splicing pattern. Outliers have been omitted. Values approaching $\hat{r} = 1$ indicate that the model has converged for the given parameter.*

B.3.2 Computational Time

The package is designed to run genes in batches, with each batch of genes executed on parallel “nodes” using the package *snow* (Tierney et al., 2009). As the entire process can take a number of days using a quad-core workstation, values were written to disk in chunks of 16 genes, allowing for minimal data and time loss in the event of a power failure, or other system failure. Each node writes to a separate folder during model fitting, allowing theoretical execution across multiple machines, although this is untested as it was not required for any dataset under investigation here. After process completion, the files across the separate nodes are merged into a single set of .CEL files containing all required data. No optimisation was performed with regard to memory consumption, however the process was able to run to completion on a Windows quad-core workstation within 5 days, for the set of 18 arrays in the $\mathrm{T}_{\mathrm{reg}}$ dataset. Subsequent testing of the $\mathrm{T}_{\mathrm{reg}}$ dataset using a multi-threaded Linux server running 20 nodes completed in 17 hours.

For the simulated data in Chapter 4, genes were fit using a quad-core workstation, with genes being fitted in batches of 4 using 1 gene/node. The Z-score filtered approach was generally about 10% faster for each batch of 4, with the overall completion time being somewhat comparable between the two approaches. The overall difference was primarily due to a small number of data points for which the computational time was unexpectedly long (Figure B.11). The reasons for this were not investigated, but it was assumed that the MCMC process may have ventured into areas for which parameter updating became time-consuming.

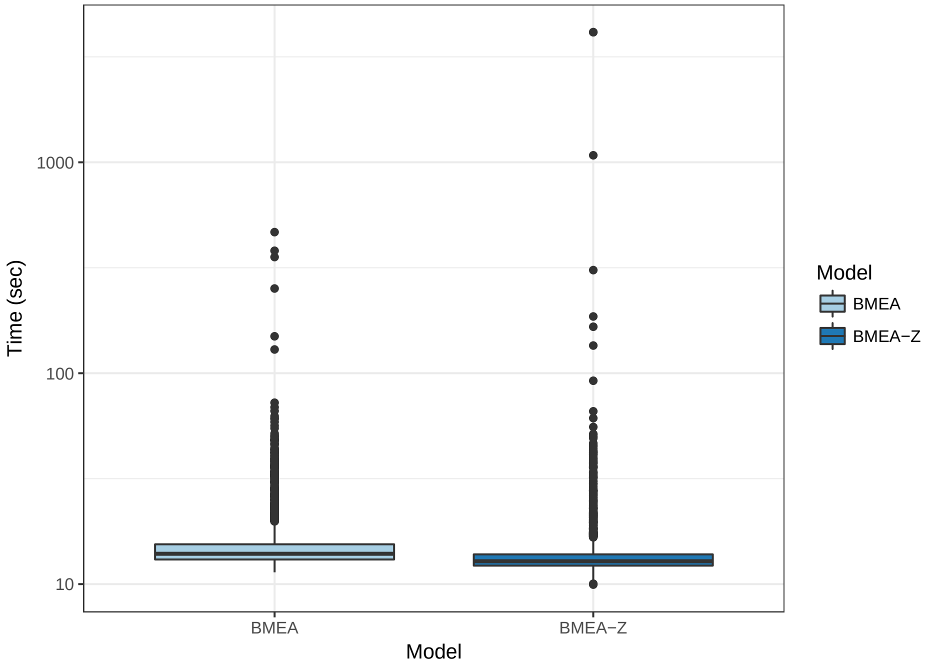


Figure B.11 – *Boxplots of time taken for each batch of 4 simulated genes for the BMEA model and for the BMEA model incorporating Z-score filtering, using the simulations from Chapter 4.*

B.4 Discussion of Package Performance

For recovery of the true values used in the simulation (Section 4.2), several noteworthy behaviours were observed. The positive bias for fitted values of μ (Figure B.3) and the negative bias for fitted values of ϕ (Figures B.5 and B.6) are likely two sides of the same coin. The small numbers of probes per exon, simulated to replicate Exon Array design, leave the posterior distributions for ϕ_{hj} as heavily influenced by the prior, with the bias towards the centre of the $\mathcal{U}(0, 1)$ prior being observed here. In order to compensate for this downwards bias in these values, posterior values for μ_h were generally forced upwards. The impact of the prior on this relatively small ($I_1 = I_2 = 4$) dataset also leaves the prior with a relatively strong impact on the posterior distributions.

The positive bias noted for σ_μ (Figure B.4) were also likely to be affected by the same phenomenon with only 8 values effectively being available for estimation of this parameter in this simulated dataset. This observation also appeared to be consistent across the range of expression levels. However, the negative bias for both σ_p and σ_S appeared to be more associated with data from the lower end of the range of expression values indicating that the model may struggle to fit parameters accurately at this end of the range, as was also noted for the parameter μ_h. For these simulated genes, the level of background signal would likely be comparable to, or greater than, the true signal, further complicating the parameter fitting process. This simply reinforces the intuitive idea that higher confidence candidates for experimental follow-up will be found when the signal component of the data strongly outweighs the background noise.

In terms of convergence between MCMC chains, both BMEA approaches were also comparable, with BMEA-Z showing slightly higher rates of convergence. The relatively poor rates of convergence for exons simulated as completely missing or absent (i.e. AS exons; Figure B.10) still showed an acceptable rate of convergence for most simulated genes with $> 60\%$ of AS exons obtaining $\hat{r}$ values < 1.1. The overall quality of the results for AS detection revealed that this observation brought minimal detrimental behaviours to the approach. Given available computational resources, more chains or a larger number of iterations may be preferable.

Appendix C

Tissue Mixture: Inspection Of Individual Exons

C.1 Manual Inspection Methods

The five most highly ranked exons from each tissue as defined in Section 5.4.5 were assessed manually in order to better confirm AS events. In order to be considered as verified, an exon-level probeset needs to:

1. Show clear separation between most probes in keeping with mixture levels of the 100% tissue samples, and the 50:50 mixture level
2. Show a clear gradient across mixture levels, with maxima/minima for ϕ_{hj} occurring in either the 90/10 or 95/05 mixture levels
3. Be predictive of a transcript combination able to be explained by known transcripts

If these were not clear AS events were either considered as either *inconclusive*, or *not confirmed*. For many of the inconclusive AS events, source biological material would be required for testing and this was not available.

Probesets were inspected based on those provided in Table C.1, however any additional probesets included as high-confidence probesets within these genes were also inspected. Other probesets which matched specific relevant transcripts were also included where appropriate.

Table C.1 – *The five most highly ranked exons from each tissue based on the defined selection criteria in Section 5.4.5. The slope of the regression line across the mixture levels is given, along with the relevant p-value for* $H_0 : \beta_1 = 0$*. Holm's method was used acros the complete set of slopes to adjust p-values, with asterisks indicating the significance of these values as per the standard conventions of R. All exons were considered as having strongly confirmed non-zero slopes with* $p_{adj} < 0.05$*.*

Exon ID	Tissue	logFC	Probes	Z_j	$\hat{\beta}_1$	p_{adj}	
ENSG00000143514_014	Brain	-4.56	4	104.1	-0.835	4.57E-02	*
ENSG00000196914_002	Brain	-3.33	8	208.9	-0.783	1.16E-03	**
ENSG00000085832_017	Brain	-3.30	8	353.9	-0.740	4.18E-02	*
ENSG00000075711_007	Brain	-3.15	4	327.4	-0.842	1.11E-02	*
ENSG00000075711_010	Brain	-2.67	4	239.9	-0.817	1.07E-02	*
ENSG00000133816_045	Heart	2.61	5	110.3	0.836	1.20E-04	***
ENSG00000149294_023	Heart	2.75	3	164.5	0.829	4.73E-02	*
ENSG00000149294_022	Heart	2.85	4	177.1	0.889	8.71E-03	**
ENSG00000133816_049	Heart	2.99	2	169.1	0.858	9.10E-05	***
ENSG00000165995_010	Heart	3.19	7	154.7	0.928	3.32E-04	***

C.2 ENSG00000143514 (TP53BP2)

The primary probeset (ENSG00000143514_014) indicates increased expression of the transcript ENST00000498843 in brain. This is supported by European Nucleotide Archive (ENA) Sequence AK316016.1, which is sourced from Brain tissue.

No clear transcripts were found supporting the high-confidence probesets 016 and 020, which were putatively heart-specific.

Table C.2 – *All exon-level probesets from ENSG00000143514 (TP53BP2) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of $\Delta \log \phi_j$, the CPI-LB and exon-level Z-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on FDR-adjusted p-values < 0.05 are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The specific exon belonging to the 5 most highly ranked exons for each tissue is ENSG00000143514_014.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000143514_012	-1.64	-0.881	19.7	-0.568	3.73E-03	1.03E-02	✓
ENSG00000143514_014*	-4.56	-2.846	104.1	-0.835	3.52E-04	2.35E-03	✓
ENSG00000143514_016*	0.80	0.416	94.3	0.456	8.17E-04	3.91E-03	✓
ENSG00000143514_017	1.31	0.875	55.8	0.469	1.18E-02	2.48E-02	✓
ENSG00000143514_019	0.72	0.211	41.1	0.432	1.51E-02	3.00E-02	✓
ENSG00000143514_020*	0.85	0.492	117.5	0.504	1.23E-03	4.93E-03	✓

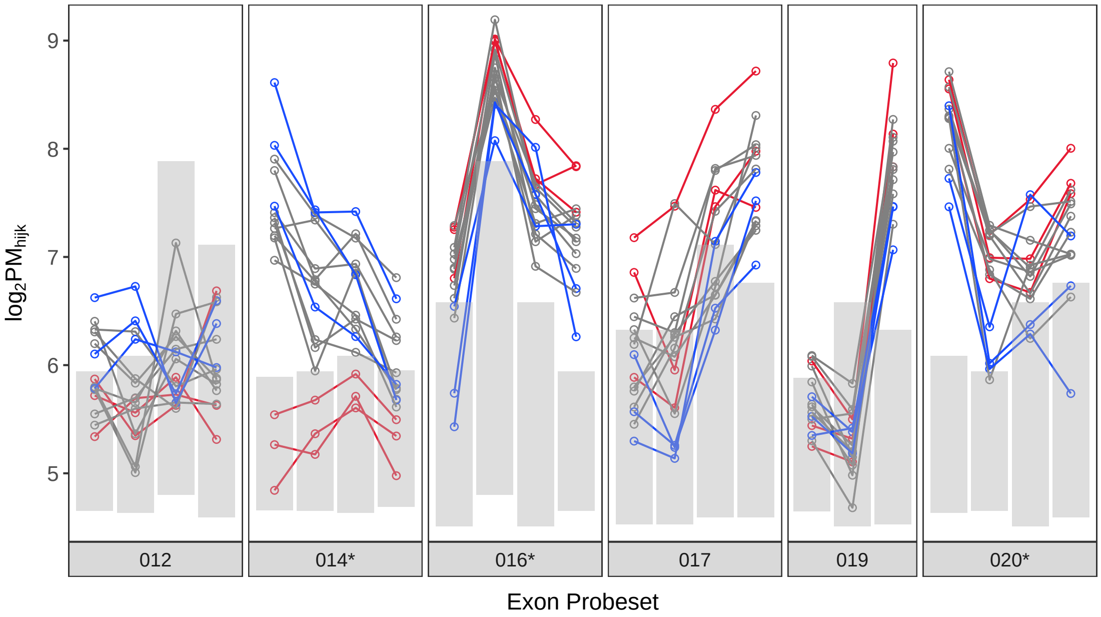


Figure C.1 – *TP53BP2 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for* $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ *as averaged across all arrays are overlaid as grey rectangles.*

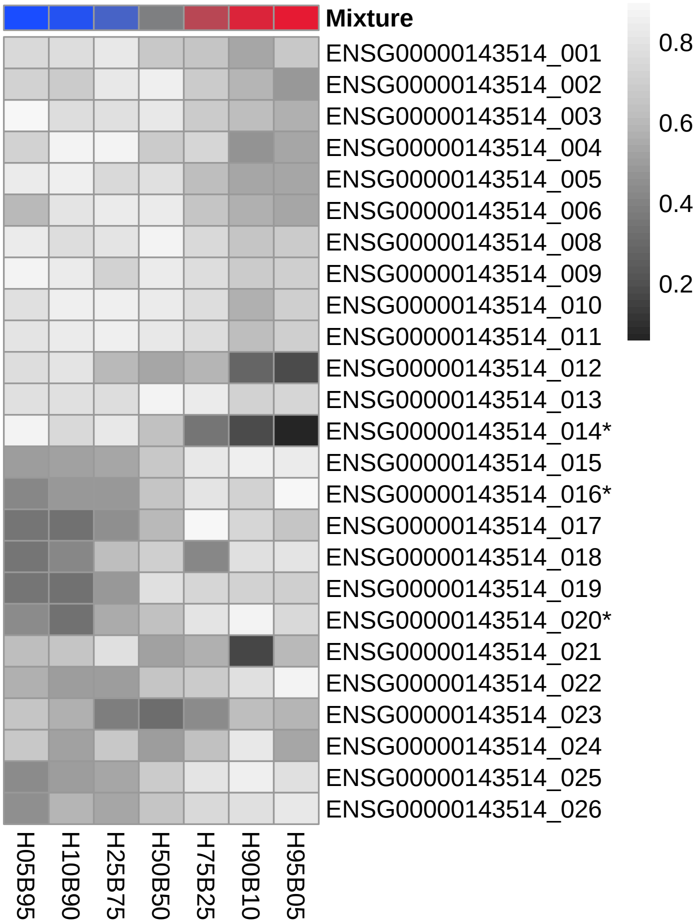


Figure C.2 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for TP53BP2. Asterisks next to the probeset number indicate a high-confidence probeset.*

C.3 ENSG00000196914 (ARHGEF12)

The next probeset specifically investigated was ENSG00000196914_002, with this probeset containing 8 probes. Whilst the first four of these showed intensity values within the range expected in the absence of true signal, the remaining probes showed a clear separation between sample groups (Figure C.3). Probeset-level estimates of ϕ_{hj} also showed a clear gradient across sample mixtures (Figure C.4). This probeset targets an alternate TSS/promoter for transcripts ENST00000532993 and ENST00000529970. Whilst no published research has provided supporting evidence for brain specific use of this promoter, a supporting cDNA for the existence of ENST00000532993 (ENA:DC350360.1) was derived from brain, as was a supporting cDNA for ENST00000529970 (ENA:AK294803.1). As such this was considered as a confirmed AS event.

The probeset ENSG00000196914_005 was also considered as a high-confidence candidate for a brain-specific AS event. The probes which were clearly detecting signal within this probeset showed an ambiguous pattern across the sample groups, however the gradient for $\hat{\phi}_{hj}$ was very consistent across mixture levels. This probeset corresponds to the second exon of the truncated, non-coding transcript ENST00000530388, which has not been detailed in any publication, but is classified as a retained intron. The supporting cDNA for this transcript (ENA:AL137456.1) was derived from a brain sample, and this evidence was considered to be a confirmed AS detection.

The probeset ENSG00000196914_044 was included in the high-confidence set as a putatively heart-specific exon, however no clear transcript was able to be determined which was targeted by the probeset. Given the unclear pattern at the probe-level, this was considered as not confirmed.

Five further probesets were included in the initial list of candidates but were excluded from the high-confidence set. ENSG00000196914_001 was the alternate promoter to ENSG00000196914_002 showing clear separation at the probe level, and an even gradient across mixture levels. Signal at this probeset was clearly detected in brain indicating the possible use of multiple promoters in brain, however this was considered as a confirmed AS event.

Signal was also clearly detected at ENSG00000196914_006 in both tissues with an increased level in heart at both the probe-level and across tissues. Whilst this appeared to indicate a potential increase in the inclusion rate of this exon, the confounding with ENSG00000196914_005 rendered this event as inconclusive.

Raw probe intensities for ENSG00000196914_015 were mainly in the region expected when $S_{hijk} = 0$ with no clear pattern for those values above this level. The gradients across the mixture levels were broadly consistent, but with greater variability than expected leaving these values as ambiguous. The probeset ENSG00000196914_015 targeted the alternate TSS for ENST00000528225.

Table C.3 – *All exon-level probesets from ENSG00000196914 (ARHGEF12) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of $\Delta \log \phi_j$, the CPI-LB and exon-level Z-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on FDR-adjusted p-values < 0.05 are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The specific exon belonging to the 5 most highly ranked exons for each tissue is ENSG00000196914_002.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000196914_001	0.53	0.297	169.8	0.461	1.42E-04	1.40E-03	✓
ENSG00000196914_002*	-3.33	-3.069	208.9	-0.783	5.28E-06	3.29E-04	✓
ENSG00000196914_005*	-1.14	-0.906	124.5	-0.330	1.06E-04	1.22E-03	✓
ENSG00000196914_006	0.43	0.164	243.3	0.219	4.56E-02	7.12E-02	×
ENSG00000196914_014	1.06	0.721	59.0	0.193	2.33E-01	2.91E-01	×
ENSG00000196914_015	-0.77	-0.426	29.9	-0.392	2.88E-03	8.63E-03	✓
ENSG00000196914_044*	0.75	0.457	73.0	0.395	1.70E-02	3.25E-02	✓
ENSG00000196914_046	-0.63	-0.330	79.2	-0.374	6.83E-03	1.65E-02	✓

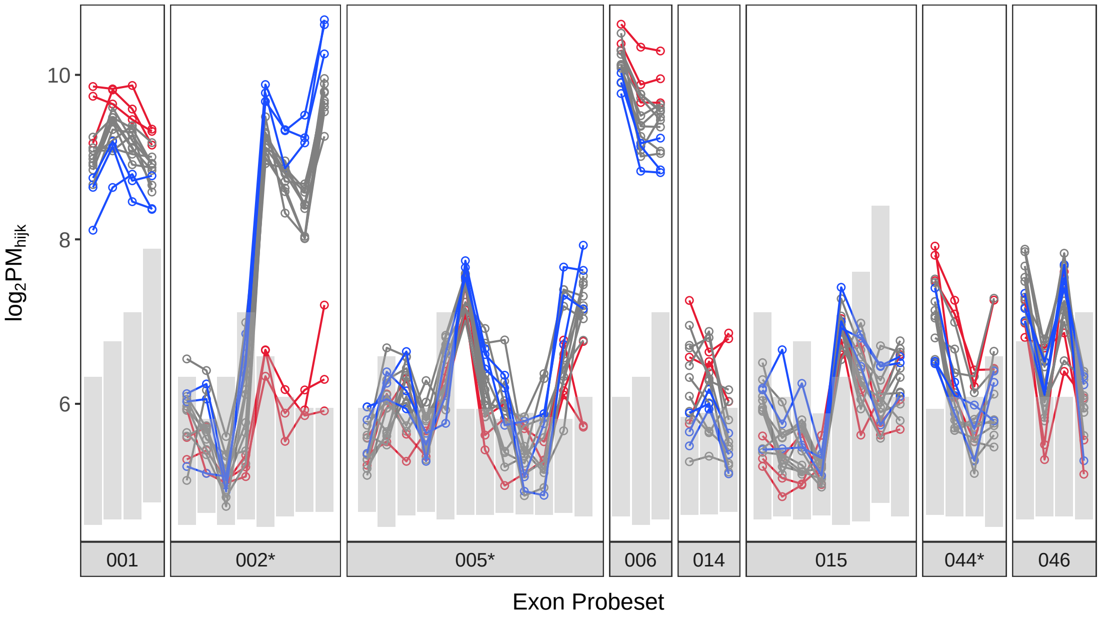


Figure C.3 – *ARHGEF12 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ as averaged across all arrays are overlaid as grey rectangles.*

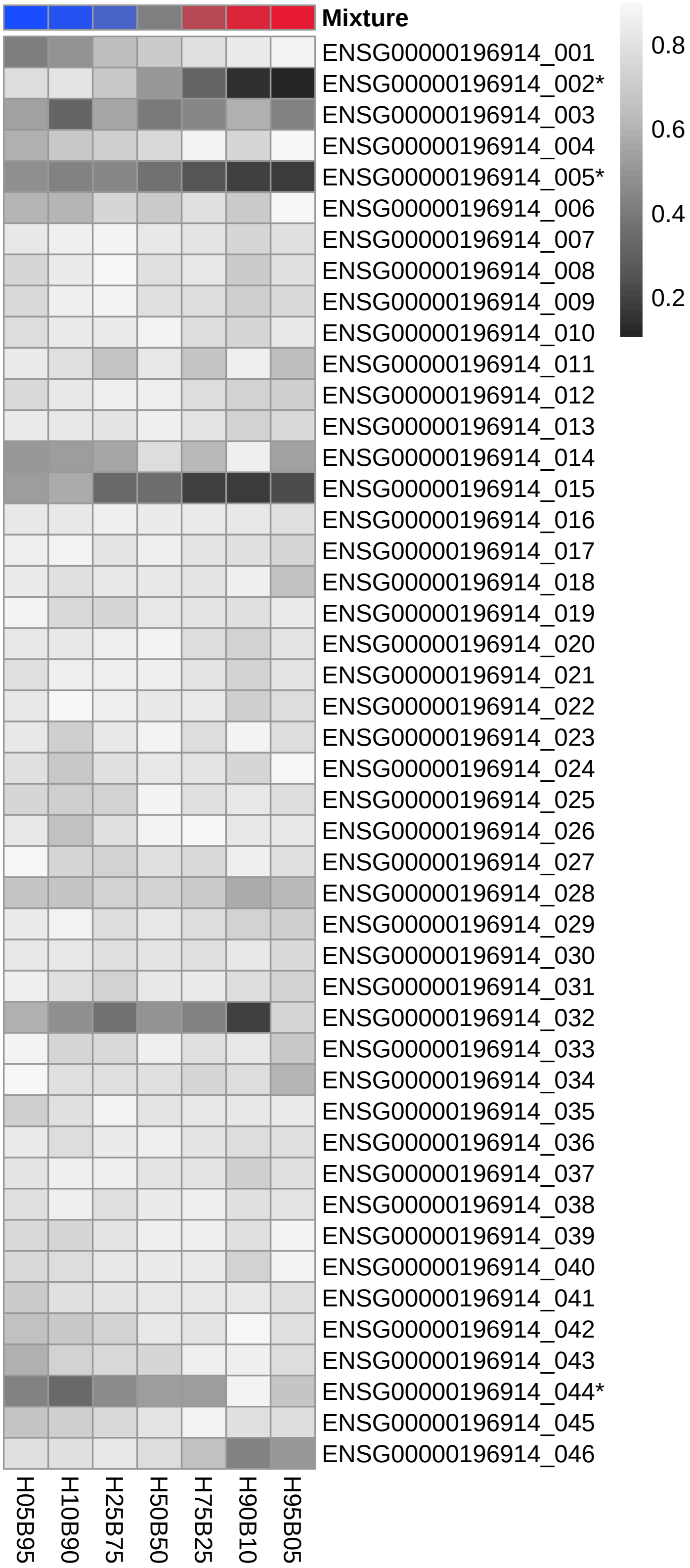


Figure C.4 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for ARHGEF12. Asterisks next to the probeset number indicate a high-confidence probeset.*

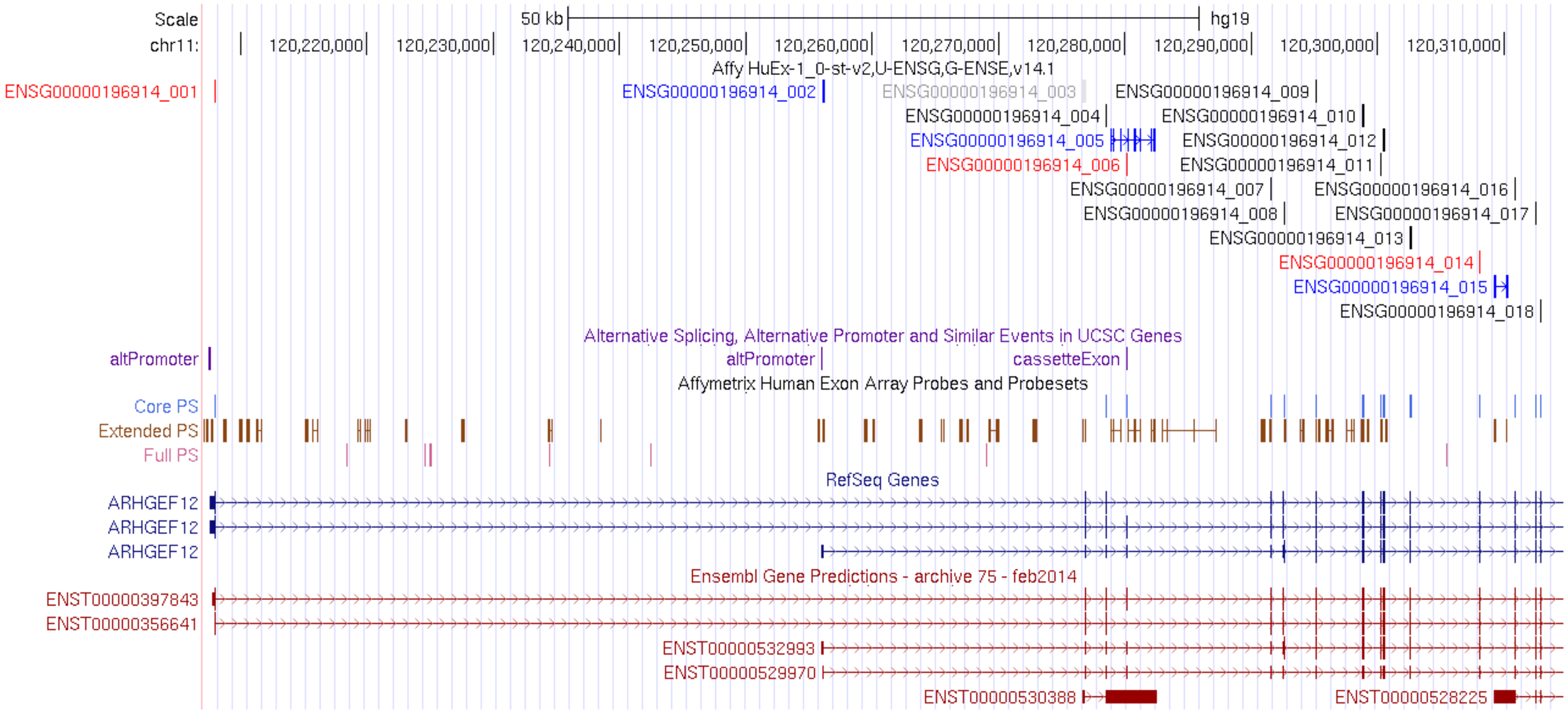


Figure C.5 – *UCSC Genome Browser plot for ARHGEF12 along with exon-level probesets as defined on the custom CDF, as well as the original Affymetrix probesets. Putative heart-specific exons are shown in red on the custom CDF track, whilst putative brain-specific exons are shown in blue. All exons with an initial B-statistic > 8 are shown in colour, whilst any undetectable exons are shown in grey.*

C.4 ENSG00000085832 (EPS15)

Another probeset for brain-specific inclusion was ENSG00000085832_017, and four of the eight probes showed a clear separation between samples, whilst the remaining four only marginally exceeded the expected values for background signal only. The gradient for ϕ_{hj} was consistent across mixture levels, and noting the clear discrepancy between RefSeq and Ensembl transcripts, this probeset was found to target the alternate TSS for NM_001159969.1. Given these observations, this was considered a confirmed AS event.

The probeset ENSG00000085832_001 was also considered as a high confidence target, with a generally consistent separation between sample groups observed at the probe level. However, this was considered to be ambiguous as some probes appeared to be highly variable within the same sample types. The gradient was also consistent across mixture levels, with extrema for $\hat{\phi_{hj}}$ in the 95% heart and 90% brain samples. The transcript being targeted by this probeset was unclear however, as the discrepancy between databases was contradictory. As such this probeset was considered to be inconclusive.

One further probeset (029) for heart-specific inclusion was included in the list of high-confidence candidates. However, raw intensity values were contradictory where PM_{hijk} values for some 100% brain samples were higher than for the 100% heart samples. Whilst extrema for $\hat{\phi}_{hj}$ were observed at the appropriate ends of the mixture levels, the gradient was relatively variable and was not considered supportive of a true AS event.

Table C.4 – *All exon-level probesets from ENSG00000085832 (EPS15) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of $\Delta \log \phi_j$, the CPI-LB and exon-level Z-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on FDR-adjusted p-values < 0.05 are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The specific exon belonging to the 5 most highly ranked exons for each tissue is ENSG00000085832_017.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000085832_001*	-0.80	-0.566	709.2	-0.511	2.17E-04	1.74E-03	✓
ENSG00000085832_002	-0.43	-0.159	504.6	-0.347	1.88E-03	6.38E-03	✓
ENSG00000085832_003	-0.42	-0.153	519.5	-0.193	5.21E-02	7.92E-02	×
ENSG00000085832_005	-0.48	-0.150	519.5	-0.226	9.58E-03	2.10E-02	✓
ENSG00000085832_017*	-3.30	-2.978	353.9	-0.740	3.10E-04	2.14E-03	✓
ENSG00000085832_018	0.65	0.323	281.2	0.259	5.85E-02	8.74E-02	×
ENSG00000085832_024	0.55	0.237	333.8	0.329	4.28E-03	1.15E-02	✓
ENSG00000085832_025	0.52	0.210	475.8	0.384	6.85E-04	3.58E-03	✓
ENSG00000085832_026	0.64	0.341	384.9	0.411	1.81E-03	6.26E-03	✓
ENSG00000085832_028	0.59	0.282	283.1	0.360	6.46E-03	1.59E-02	✓
ENSG00000085832_029*	1.28	0.907	110.1	0.435	1.84E-02	3.46E-02	✓

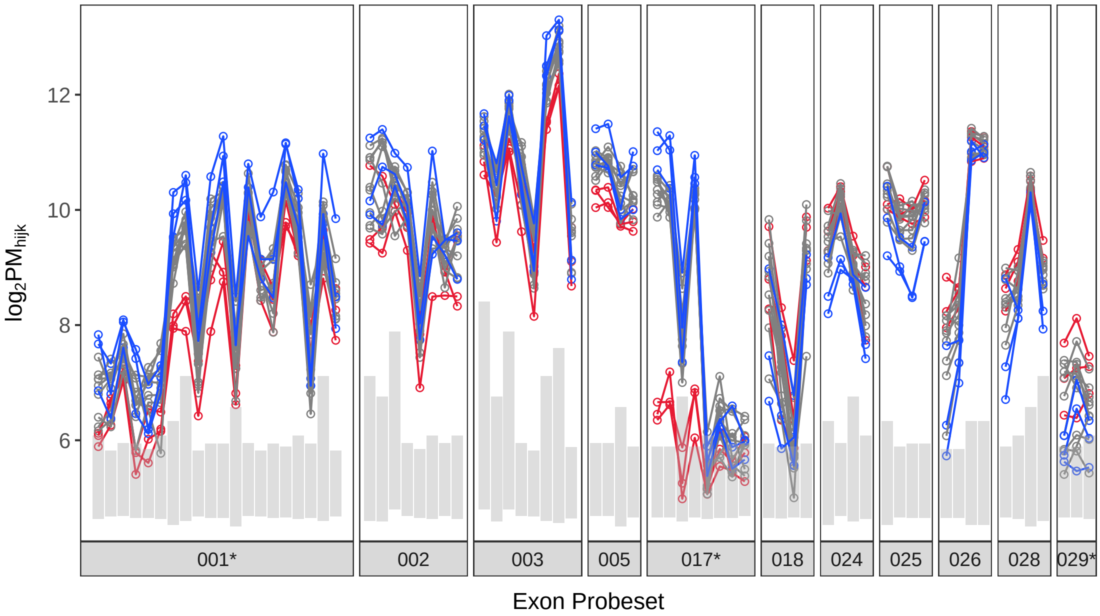


Figure C.6 – *EPS15 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ as averaged across all arrays are overlaid as grey rectangles.*

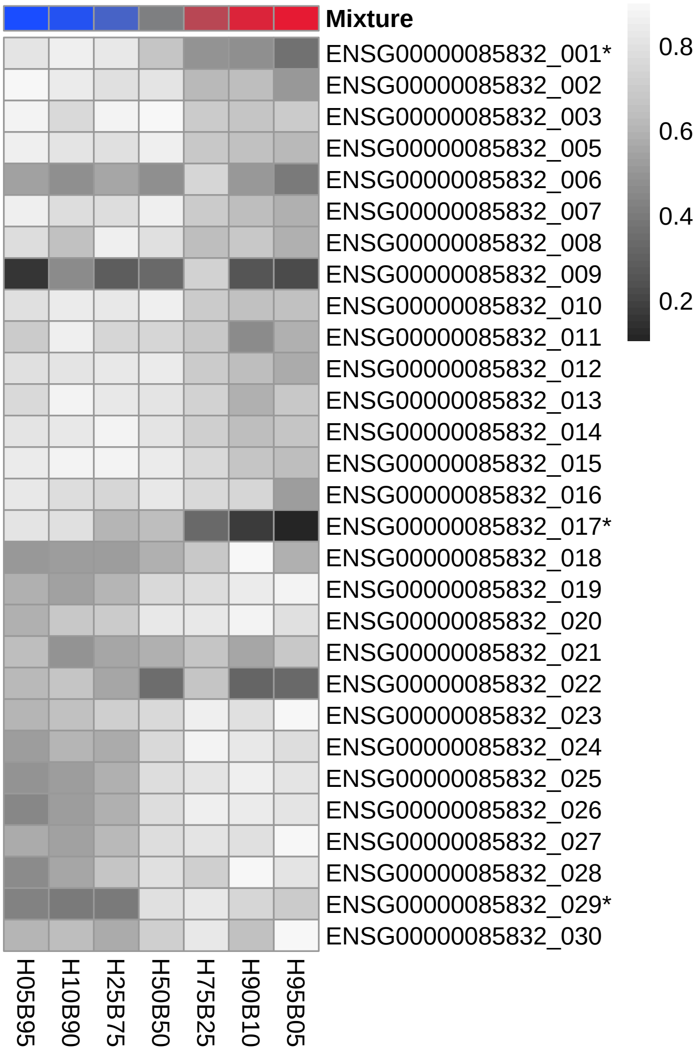


Figure C.7 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for EPS15. Asterisks next to the probeset number indicate a high-confidence probeset.*

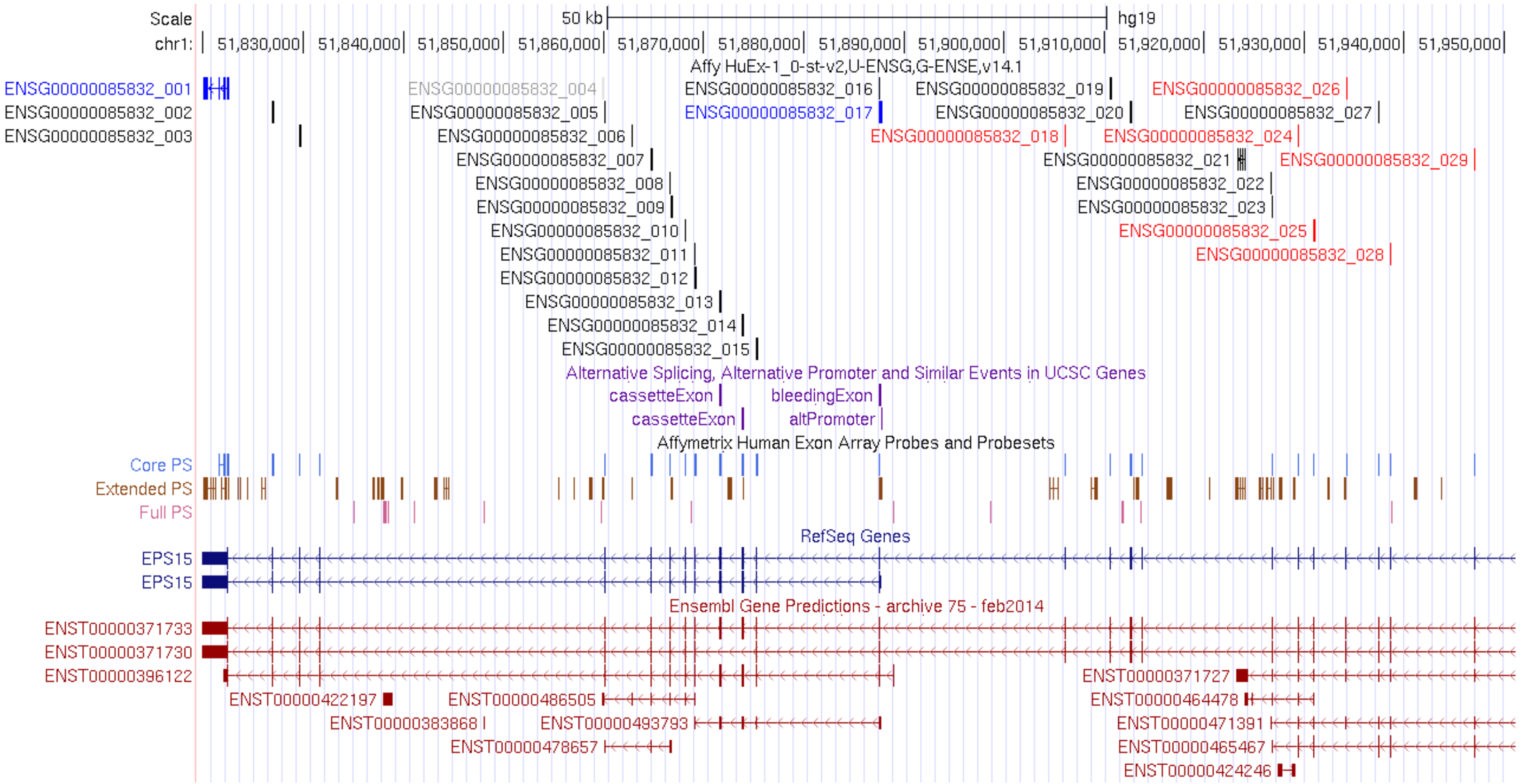


Figure C.8 – *UCSC Genome Browser plot for EPS15 along with exon-level probesets as defined on the custom CDF, as well as the original Affymetrix probesets. Putative heart-specific exons are shown in red on the custom CDF track, whilst putative brain-specific exons are shown in blue. All exons with an initial B-statistic > 8 are shown in colour, whilst any undetectable exons are shown in grey.*

C.5 ENSG00000075711 (DLG1)

Probesets ENSG00000075711_007 and ENSG00000075711_011 both target the transcript ENST00000443183 which is supported by ENA record AK294772.1, derived from brain. Probeset ENSG00000075711_027 also targets the 5' end of this transcript. All probe and ϕ_{hj} gradient evidence is strongly supportive and this is another confirmed AS event. Probeset ENSG00000075711_030 targets ENST00000453607 which is supported by ENA record DA121908.1 and was also derived from brain. All evidence appears confirmatory for this AS event.

Putatively heart specific AS events targeted by ENSG00000075711_032/033 could not be ascribed to any transcripts, and probe and ϕ_{hj} gradients were inconclusive. These were not considered as confirmed.

Probeset ENSG00000075711_026 targets a common cassette exon and offered broadly supportive evidence at the probe and gradient level. As such, this was considered a confirmed heart-specific AS event.

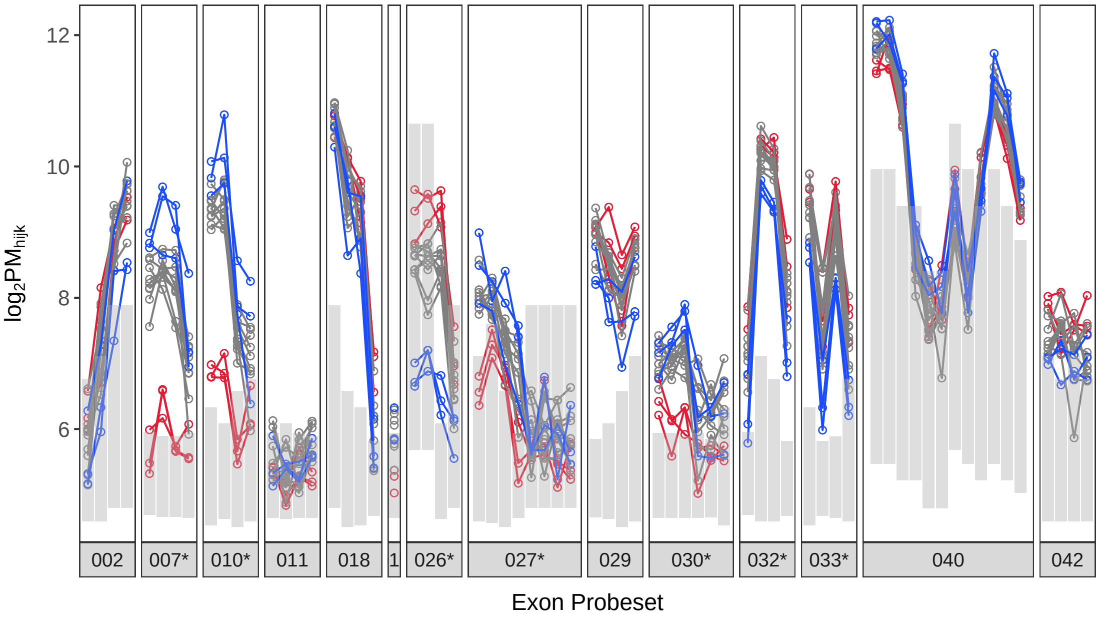


Figure C.9 – *DLG1 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for* $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ *as averaged across all arrays are overlaid as grey rectangles.*

Table C.5 – *All exon-level probesets from ENSG00000075711 (DLG1) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of* $\Delta \log \phi_j$*, the CPI-LB and exon-level* Z*-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on* FDR*-adjusted p-values* < 0.05 *are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The exons belonging to the 5 most highly ranked exons for each tissue are ENSG00000075711_007 and ENSG00000075711_010.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000075711_002	0.55	0.207	94.2	0.027	7.33E-01	7.69E-01	×
ENSG00000075711_007*	-3.15	-2.791	327.4	-0.842	6.03E-05	9.56E-04	✓
ENSG00000075711_010*	-2.67	-2.299	239.9	-0.817	5.69E-05	9.51E-04	✓
ENSG00000075711_011	-1.47	-0.466	22.1	-0.407	1.05E-01	1.44E-01	×
ENSG00000075711_018	0.47	0.173	244.6	0.185	4.94E-03	1.28E-02	✓
ENSG00000075711_019	-2.44	-1.175	31.4	-0.639	1.64E-02	3.17E-02	✓
ENSG00000075711_026*	2.50	2.050	92.0	0.823	4.59E-07	1.11E-04	✓
ENSG00000075711_027*	-1.63	-1.262	74.6	-0.593	1.16E-05	4.91E-04	✓
ENSG00000075711_029	0.49	0.205	308.1	0.294	1.32E-03	5.11E-03	✓
ENSG00000075711_030*	-1.40	-1.100	197.3	-0.665	7.73E-05	1.05E-03	✓
ENSG00000075711_032*	0.72	0.423	286.9	0.379	3.85E-03	1.06E-02	✓
ENSG00000075711_033*	0.92	0.610	316.8	0.438	5.69E-03	1.45E-02	✓
ENSG00000075711_040	-0.38	-0.146	155.3	-0.075	3.48E-01	4.11E-01	×
ENSG00000075711_042	0.52	0.181	101.5	0.401	7.02E-03	1.67E-02	✓

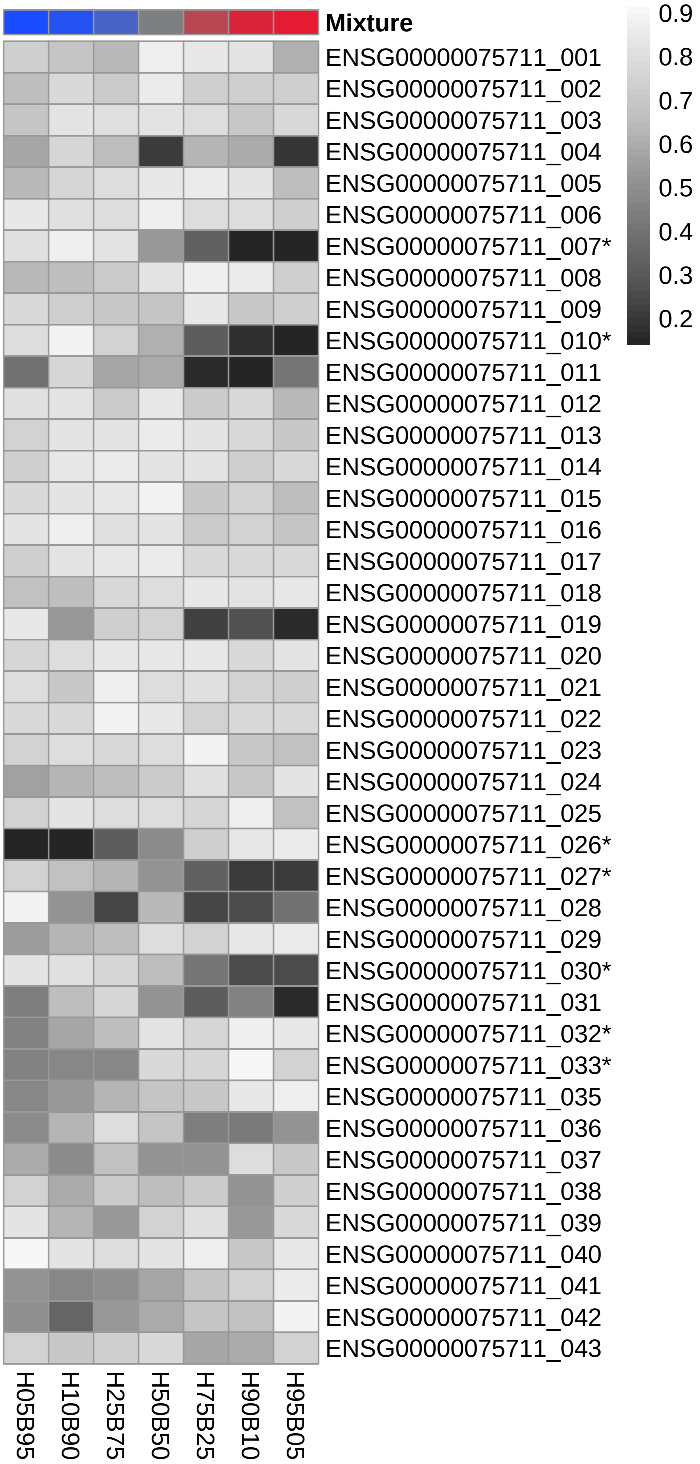


Figure C.10 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for DLG1. Asterisks next to the probeset number indicate a high-confidence probeset.*

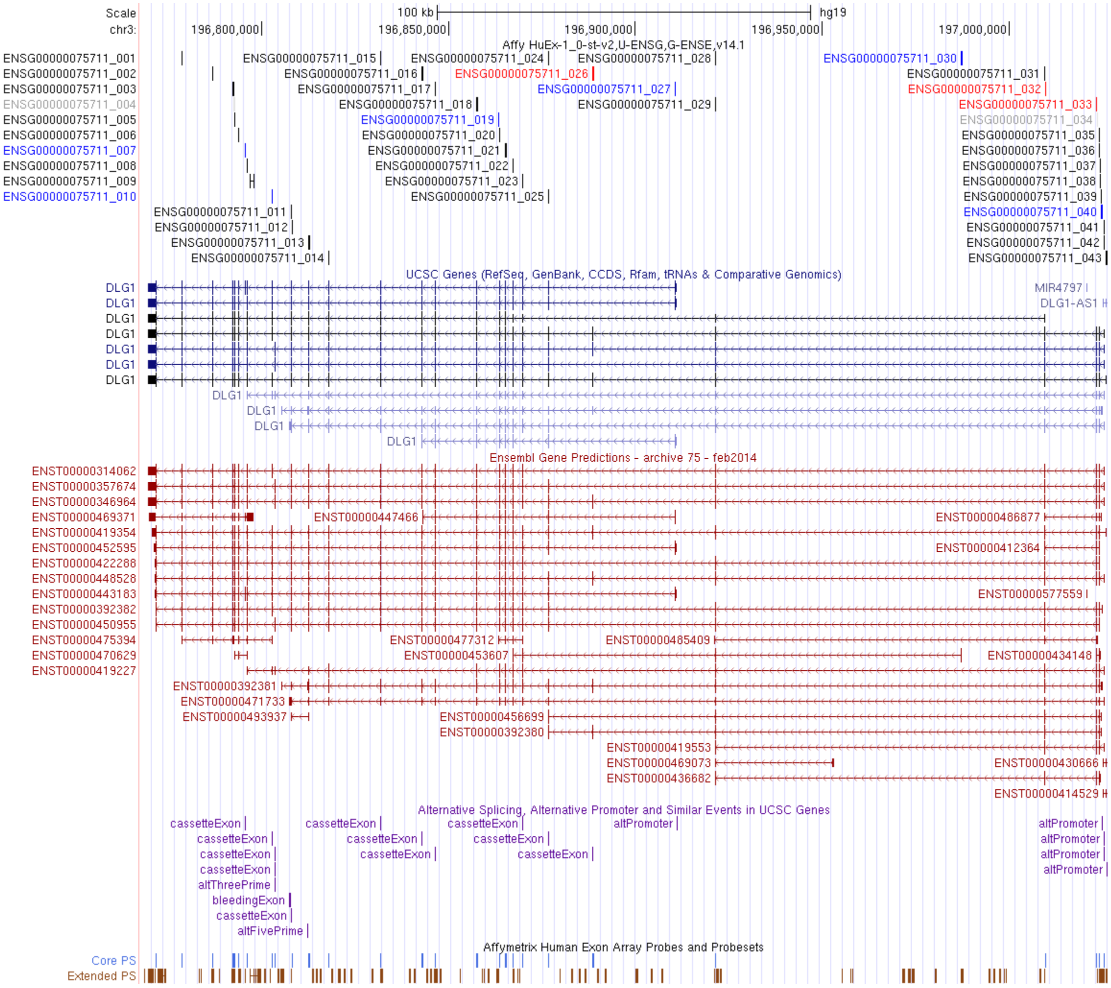


Figure C.11 – *UCSC Genome Browser plot for DLG1 along with exon-level probesets as defined on the custom CDF, as well as the original Affymetrix probesets. Putative heart-specific exons are shown in red on the custom CDF track, whilst putative brain-specific exons are shown in blue. All exons with an initial B-statistic > 8 are shown in colour, whilst any undetectable exons are shown in grey.*

C.6 ENSG00000165995 (CACNB2)

The highly ranked heart-specific probesets were ENSG00000165995_009/010 and raw probe intensities (Figure C.12) showed clear separation as expected. Estimates of ϕ_{hj} also showed a consistent gradient across mixture levels (Figure C.13) for both probesets. Both probesets target an alternate TSS for the coding transcript ENST00000377329, and non-coding transcript ENST00000498816, and the supporting cDNA (ENA:CR858773.1) for the coding transcript was derived from heart tissue. This was then considered to be a confirmed AS event.

In addition to the two heart-specific probesets, ENSG00000165995_011 was also included in the list of high-confidence candidates, and was putatively targeting a brain-specific exon. Raw intensities at the probe level (Figure C.12) show patterns supporting this as a brain-specific exon. However, values for $\hat{\phi}_{hj}$ across mixture levels had an unexpected peak in the 75% brain sample, with the remainder of samples following the expected continuous gradient, leaving this as an ambiguous pattern. This probeset corresponds to an alternate TSS which produces the coding transcript ENST00000377315, with supporting cDNAs for this transcript have been found in multiple tissues, of which ENA:L20343.1 was derived from brain. The alternate transcripts for this gene have also been shown to play a key role in both cardiac and mental disorders (Soldatov, 2015) and this was also considered to be a confirmed detection of an AS event.

The additional probeset ENSG00000165995_007 was included in the high-confidence list of probesets. Some PM_{hijk} values in the heart samples exceed the brain samples, and these were not considered supportive. The gradient for $\hat{\phi}_{hj}$ was also somewhat inconsistent, with the second-highest value being obtained in the 50:50 mixture level. Whilst this probeset targets the alternate TSS for transcript ENST00000396576, this was not considered to be a confirmed AS event

Table C.6 – *All exon-level probesets from ENSG00000165995 (CACNB2) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of $\Delta \log \phi_j$, the CPI-LB and exon-level Z-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on FDR-adjusted p-values < 0.05 are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The specific exon belonging to the 5 most highly ranked exons for each tissue is ENSG00000165995_010.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000165995_007*	-0.91	-0.545	99.9	-0.385	1.77E-02	3.37E-02	✓
ENSG00000165995_008	-0.61	-0.270	71.5	-0.331	1.09E-02	2.33E-02	✓
ENSG00000165995_009	3.85	2.707	57.4	0.818	2.88E-06	2.69E-04	✓
ENSG00000165995_010*	3.19	2.795	154.7	0.928	1.45E-06	2.06E-04	✓
ENSG00000165995_011*	-2.10	-1.719	112.5	-0.722	6.87E-04	3.58E-03	✓
ENSG00000165995_025	0.64	0.270	90.4	0.159	6.55E-03	1.60E-02	✓

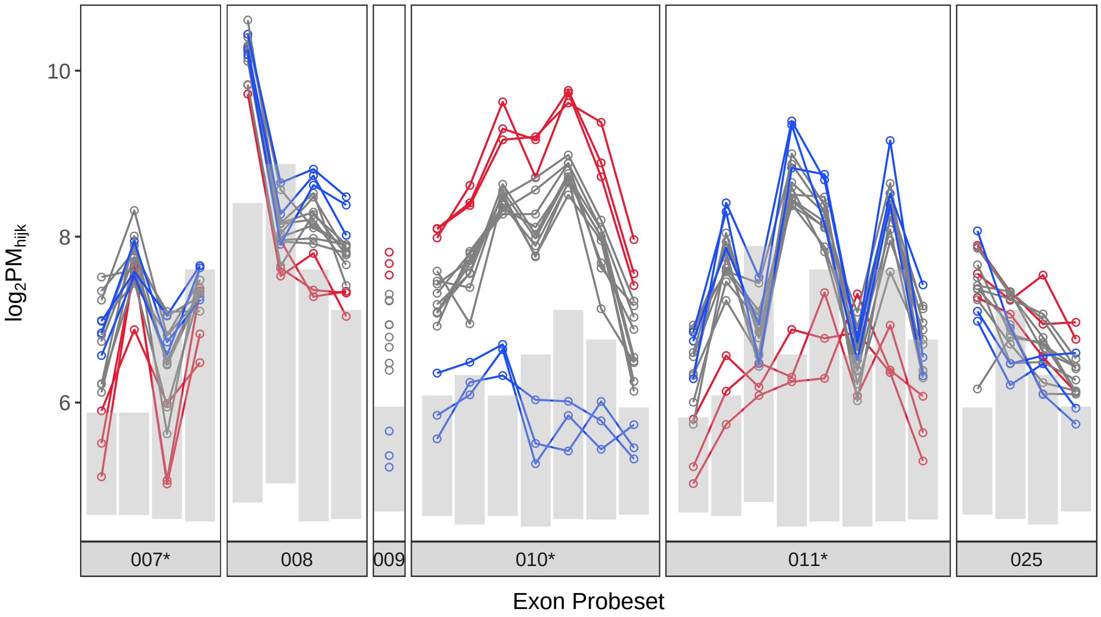


Figure C.12 – *CACNB2 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ as averaged across all arrays are overlaid as grey rectangles.*

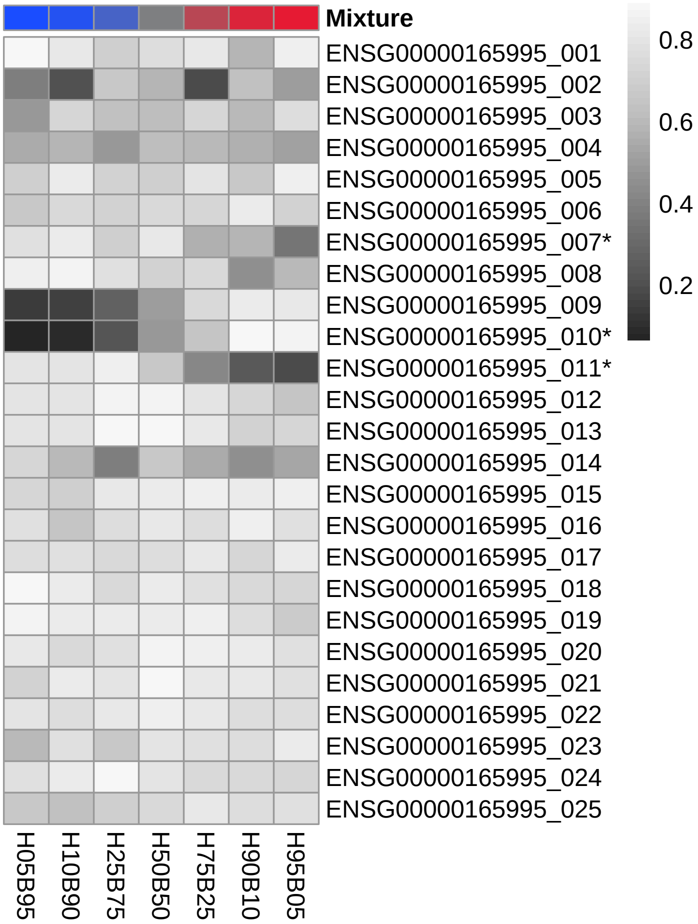


Figure C.13 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for CACNB2. Asterisks next to the probeset number indicate a high-confidence probeset.*

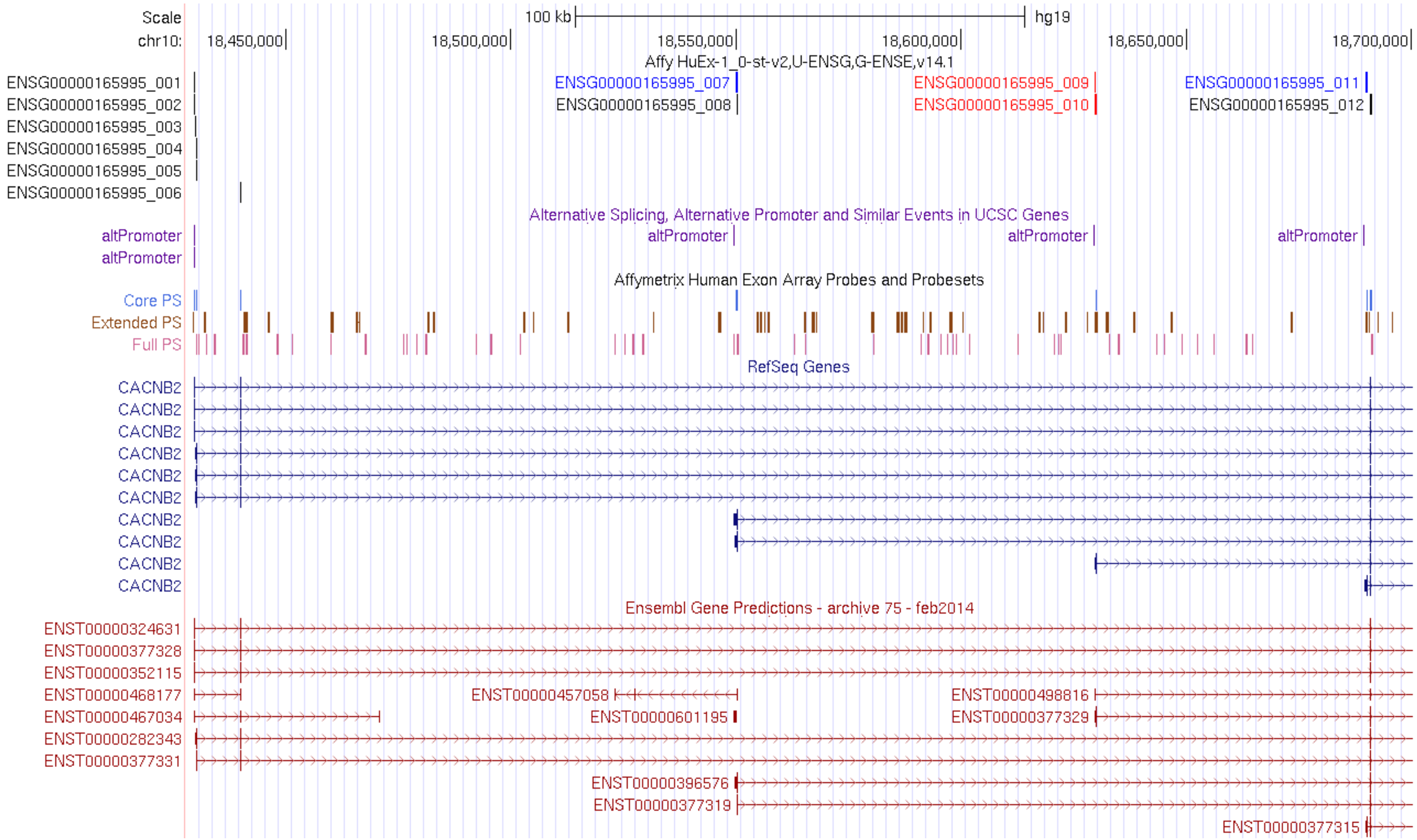


Figure C.14 – *UCSC Genome Browser plot for CACNB2 along with exon-level probesets as defined on the custom CDF, as well as the original Affymetrix probesets. Putative heart-specific exons are shown in red on the custom CDF track, whilst putative brain-specific exons are shown in blue. All exons with an initial B-statistic > 8 are shown in colour, whilst any undetectable exons are shown in grey.*

C.7 ENSG00000133816 (MICAL2)

Probesets from the gene MICAL2 which were in the top 5 candidates for a heart-specific inclusion event were ENSG00000133816_045/046/049 . Raw probe intensities show extremely clear separation between sample types for all three probesets (Figure C.15), with a near constant gradient also being observed across all mixtures at the probeset level (Figure C.16). An additional probeset in the high-confidence list of candidate probesets (ENSG00000133816_050) also showed this clear separation and even gradient.

All four of these probesets target a retained intron from the non-coding transcript ENST00000 530691, with two further RefSeq coding transcripts being targeted. This makes a fairly strong case for this being the correct detection of a true AS event, however no published evidence for this retained intron could be found in the literature with regard to heart expression, and supporting cDNAs from the European Nucleotide Archive were all derived from non-cardiac tissues. Despite this, the intensity patterns at the probe-level were so striking that this was considered as a confirmed AS event.

One further putative heart-specific inclusion event was suggested in the original list by probeset ENSG00000133816_008, however this probeset was excluded from the high-confidence list due to a low Z score. Only one probe returned PM_{hijk} values beyond the levels expected from background signal alone and this was broadly supportive of the sample groupings. Estimates of ϕ_{hj} across mixture levels were broadly supportive, however an unexpected drop was noted in the 95% heart sample. This probeset targeted the alternate TSS for ENST00000524685 and this was considered to be inconclusive.

Whilst not included in the list of high-confidence candidates, the probesets ENSG0000013 3816_031/032/033/034/037 were broadly supportive of increased inclusion rates in brain, and as these all targeted cassette exons, these were considered as probable AS events. Probes for ENSG00000133816_041 were mainly in the range of background signal and were considered as inconclusive.

Table C.7 – *All exon-level probesets from ENSG00000133816 (MICAL2) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of $\Delta \log \phi_j$, the CPI-LB and exon-level Z-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on FDR-adjusted p-values < 0.05 are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The exons belonging to the 5 most highly ranked exons for each tissue are ENSG00000133816_049 and ENSG00000133816_045.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000133816_001	0.52	0.159	131.5	0.291	9.21E-03	2.03E-02	✓
ENSG00000133816_008	2.62	1.581	19.8	0.442	1.97E-02	3.68E-02	✓
ENSG00000133816_014	-0.93	-0.307	33.3	0.020	9.10E-01	9.27E-01	✗
ENSG00000133816_031	-0.42	-0.207	269.6	-0.292	2.03E-02	3.77E-02	✓
ENSG00000133816_032	-0.34	-0.109	147.2	-0.148	8.70E-02	1.23E-01	✗
ENSG00000133816_033	-0.32	-0.098	328.7	-0.193	4.41E-02	6.94E-02	✗
ENSG00000133816_034	-0.39	-0.157	150.0	-0.325	2.93E-03	8.72E-03	✓
ENSG00000133816_037	-0.40	-0.182	357.9	-0.253	1.82E-02	3.45E-02	✓
ENSG00000133816_041	-0.73	-0.383	43.3	-0.225	3.98E-02	6.40E-02	✗
ENSG00000133816_045*	2.61	2.220	110.3	0.836	5.21E-07	1.11E-04	✓
ENSG00000133816_046*	2.57	2.369	358.5	0.866	2.00E-08	2.54E-05	✓
ENSG00000133816_049*	2.99	2.611	169.1	0.858	3.91E-07	1.11E-04	✓
ENSG00000133816_050*	1.86	1.516	98.8	0.784	1.69E-06	2.15E-04	✓

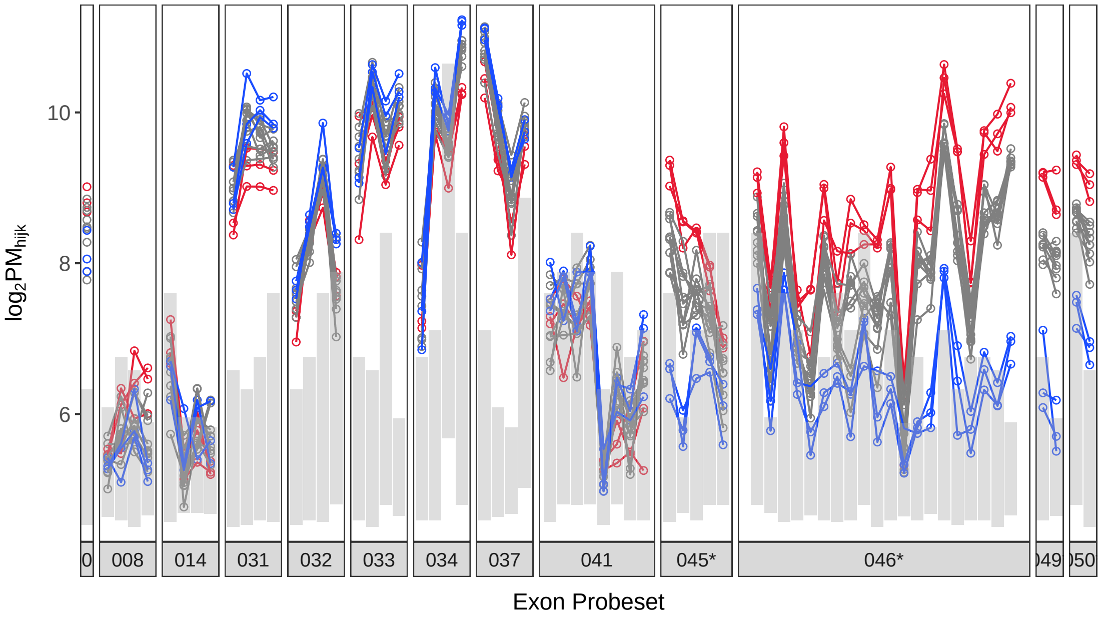


Figure C.15 – *MICAL2 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ as averaged across all arrays are overlaid as grey rectangles.*

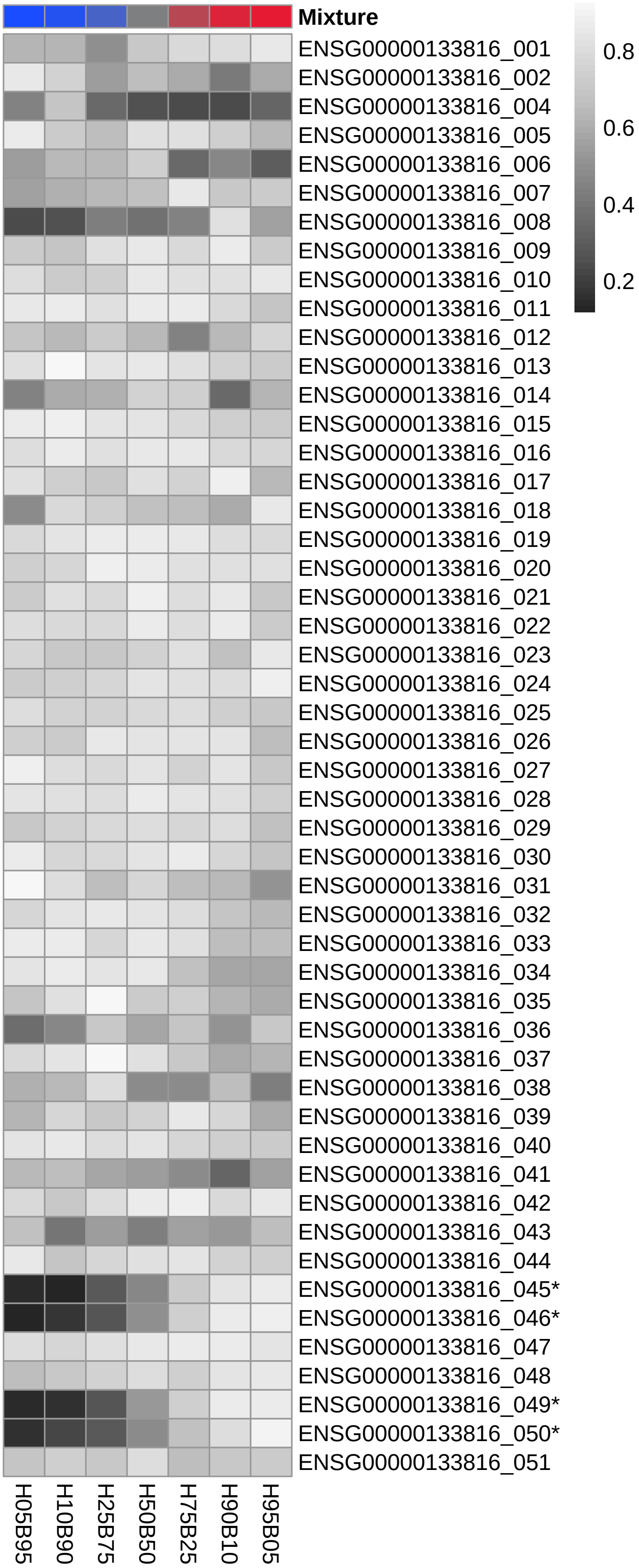


Figure C.16 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for MICAL2. Asterisks next to the probeset number indicate a high-confidence probeset.*

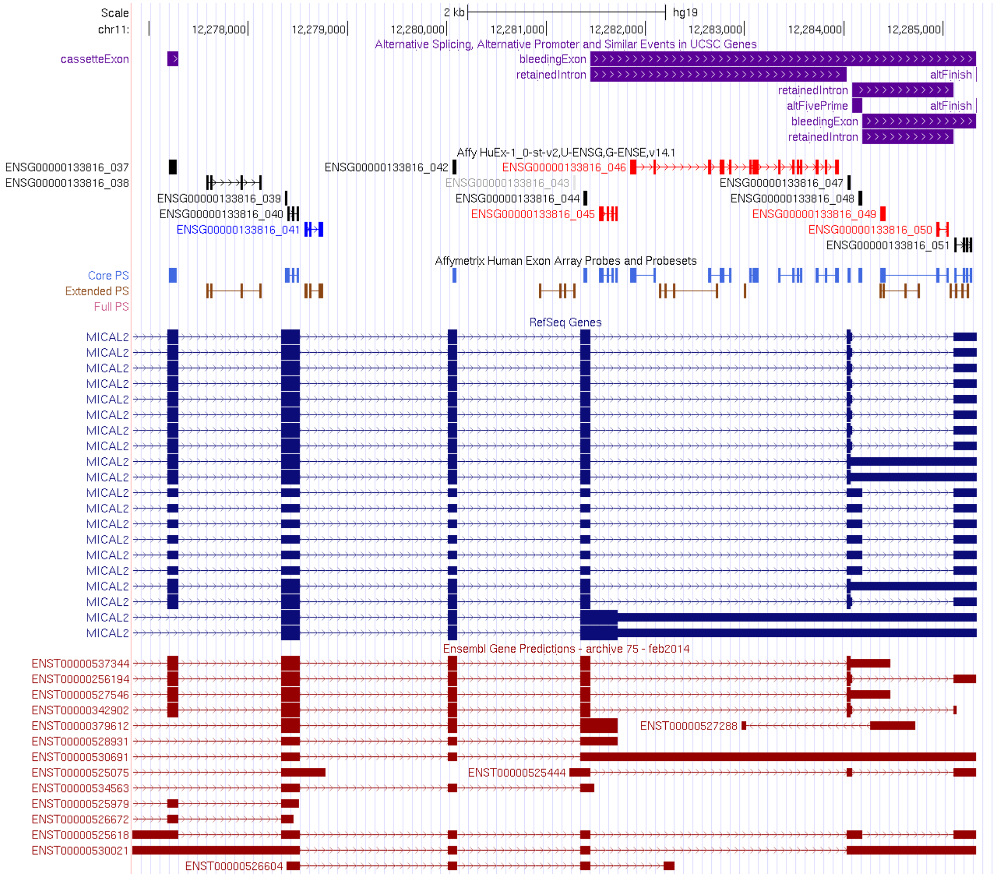


Figure C.17 – *UCSC Genome Browser plot for MICAL2 along with exon-level probesets as defined on the custom CDF, as well as the original Affymetrix probesets. Putative heart-specific exons are shown in red on the custom CDF track, whilst putative brain-specific exons are shown in blue. All exons with an initial B-statistic > 8 are shown in colour, whilst any undetectable exons are shown in grey.*

C.8 ENSG00000149294 (NCAM1)

Putatively brain-specific probeset ENSG00000149294_015 targets ENST00000529420, and whilst the probe-level plot was inconclusive, the gradient across mixture levels was supportive. This transcript is supported by ENA record BC029119.1 which was derived from brain & this was considered as a confirmed AS event.

Putatively heart-specific probesets ENSG00000149294_022 and ENSG00000149294_023 showed very supportive patterns in all plots and targeted known cassette exons for multiple transcripts. This were considered as confirmed AS events.

Probeset ENSG00000149294_024 was inconclusive at the probe level, but supported across the mixture levels. This targeted two possible transcripts and was considered as inconclusive.

Probes and gradients were supportive for ENSG00000149294_028, which targets the 3'UTR for ENST00000533760. However, the supporting evidence from ENA record AB209443.1 was derived from brain. This also contradicted the results from probesets 022 and 023, and as such this was considered as inconclusive.

Two possible transcripts (ENST00000531927 and ENST00000531044) were targeted by probesets ENSG00000149294_030 to 032, whilst probesets 034 to 037 targeted ENST00000528158. All supporting records at ENA were brain or neuron derived and this were considered as confirmed, given the strongly supportive probe and gradient plots. No direct transcript was found for ENSG00000149294_038, however all other plots were supportive and this also considered as confirmed.

Table C.8 – *All exon-level probesets from ENSG00000149294 (NCAM1) considered as candidates for AS events based on the initial selection criteria in Section 5.4.4. Initial point estimates of $\Delta \log \phi_j$, the CPI-LB and exon-level Z-scores are shown, as these were utilised in selection of the high-confidence candidates. High-confidence probesets are indicated with an asterisk. Exons with a confirmed non-zero slope based on FDR-adjusted p-values < 0.05 are indicated with a tick. A negative slope indicates a putative brain-specific exon, whilst a positive slope indicates a putative heart-specific exon. The exons belonging to the 5 most highly ranked exons for each tissue are ENSG00000149294_022 and ENSG00000149294_023.*

Exon ID	$\widehat{\Delta \log \phi_j}$	CPI-LB	Z_j	$\hat{\beta}_1$	p	FDR	
ENSG00000149294_001	-0.63	-0.362	161.0	-0.403	1.50E-03	5.57E-03	✓
ENSG00000149294_006	-0.70	-0.320	107.9	-0.264	2.55E-02	4.52E-02	✓
ENSG00000149294_011	0.43	0.178	445.0	0.235	7.32E-03	1.72E-02	✓
ENSG00000149294_012	0.45	0.201	484.7	0.177	9.44E-02	1.32E-01	×
ENSG00000149294_013	0.36	0.105	542.5	0.106	2.27E-01	2.85E-01	×
ENSG00000149294_015*	-0.86	-0.586	199.9	-0.474	3.93E-06	3.01E-04	✓
ENSG00000149294_016	0.42	0.168	376.7	0.204	6.31E-02	9.33E-02	×
ENSG00000149294_019	0.45	0.204	718.8	0.156	2.17E-02	3.96E-02	✓
ENSG00000149294_020	0.51	0.262	541.3	0.245	2.75E-02	4.76E-02	✓
ENSG00000149294_021	0.60	0.345	236.3	0.337	8.68E-03	1.97E-02	✓
ENSG00000149294_022*	2.85	2.519	177.1	0.889	4.52E-05	8.45E-04	✓
ENSG00000149294_023*	2.75	2.432	164.5	0.829	3.66E-04	2.40E-03	✓
ENSG00000149294_024*	-0.74	-0.428	152.1	-0.372	2.33E-03	7.51E-03	✓
ENSG00000149294_025	-0.39	-0.116	162.7	-0.533	1.59E-03	5.80E-03	✓
ENSG00000149294_026	0.51	0.264	616.6	0.239	1.44E-02	2.90E-02	✓
ENSG00000149294_027	0.50	0.253	487.2	0.275	1.49E-03	5.56E-03	✓
ENSG00000149294_028*	0.80	0.577	642.2	0.387	2.10E-05	5.68E-04	✓
ENSG00000149294_030*	-1.14	-0.895	639.6	-0.626	1.47E-05	5.52E-04	✓
ENSG00000149294_031*	-0.95	-0.676	153.3	-0.518	1.20E-04	1.30E-03	✓
ENSG00000149294_032*	-1.81	-1.528	203.5	-0.758	6.51E-05	9.72E-04	✓
ENSG00000149294_034*	-1.10	-0.813	154.6	-0.543	6.82E-05	9.72E-04	✓
ENSG00000149294_035	-0.54	-0.285	175.6	-0.369	5.55E-04	3.14E-03	✓
ENSG00000149294_036*	-0.88	-0.621	459.4	-0.598	2.06E-05	5.68E-04	✓
ENSG00000149294_037*	-0.90	-0.685	891.2	-0.662	1.45E-06	2.06E-04	✓
ENSG00000149294_038*	-0.95	-0.673	334.2	-0.569	1.76E-03	6.14E-03	✓

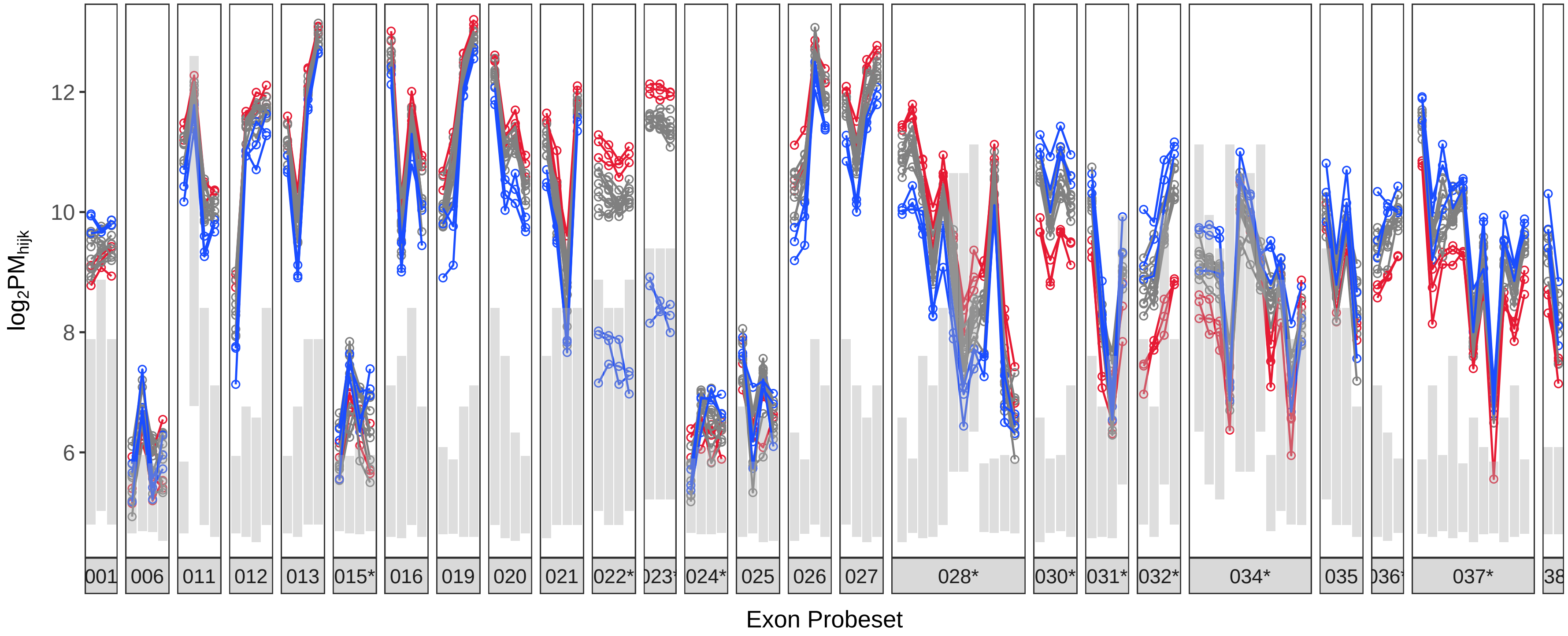


Figure C.18 – *NCAM1 raw probe intensities for exons in the initial list of candidates for fitting across all arrays. Only the 100% Heart (red), 50:50 (grey) and 100% Brain (blue) samples are shown. Asterisks next to the probeset number indicate a high-confidence probeset. The values for* $\hat{\lambda}_{..jk} \pm 2\hat{\delta}_{..jk}$ *as averaged across all arrays are overlaid as grey rectangles.*

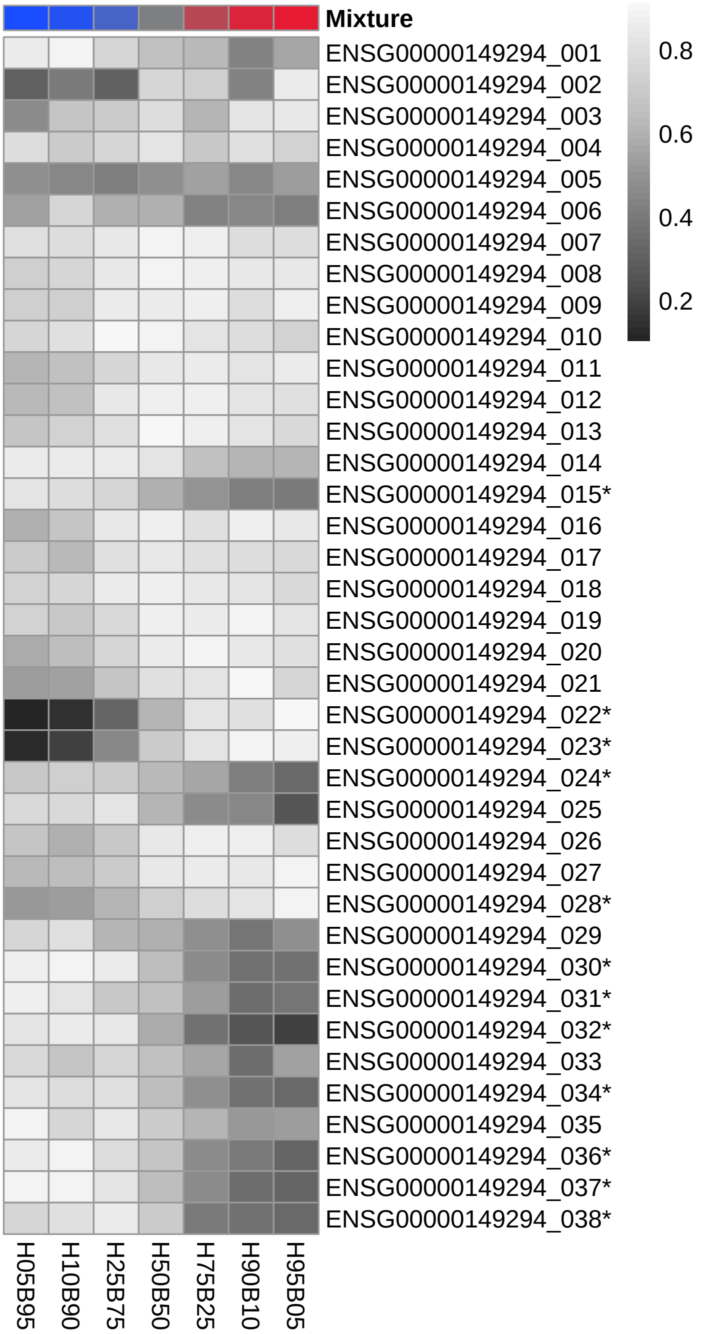


Figure C.19 – *Posterior medians as point estimates for ϕ_{hj} across all exon-level probesets and tissue mixtures for NCAM1. Asterisks next to the probeset number indicate a high-confidence probeset.*

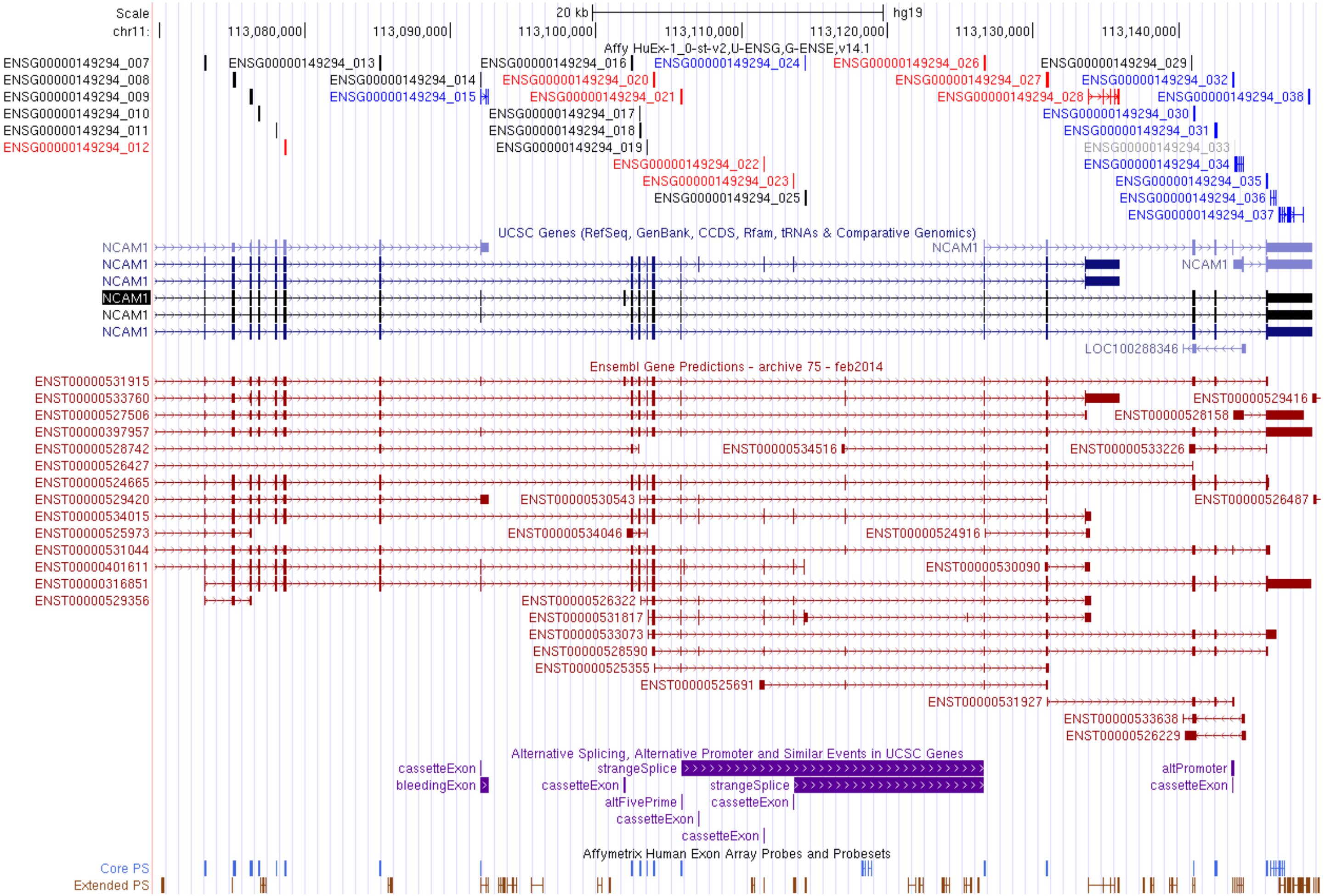


Figure C.20 – *UCSC Genome Browser plot for NCAM1 along with exon-level probesets as defined on the custom CDF, as well as the original Affymetrix probesets. Putative heart-specific exons are shown in red on the custom CDF track, whilst putative brain-specific exons are shown in blue. All exons with an initial B-statistic > 8 are shown in colour, whilst any undetectable exons are shown in grey.*

Appendix D

Primer Design for qPCR

Table D.1 – *Primer design for qPCR in Section 6.1.3*

Probeset	Gene	Target Exon	Forward	Reverse
ENSG00000119314_018	PTBP3	chr9:115092722-115092754 (-)	>chr9:115092735→115060191 ACAGTCGGTTTAAAGCGGGG	>chr9:115060161→115060172 AGGTCCGTTAATGATGCCAGA
ENSG00000021776_026	AQR	chr15:35226773-35226820 (-)	>chr15:35231015→35230992 GGGCATATCAAGAGAGGAGATTT	>chr15:35226814→35230944 CATAGTGACAGGTTCCAGAAAGT
ENSG00000077147_014	TM9SF3	chr10:98325062-98325168 (-)	>chr10:98336503→98336482 ACTTCCATTCTGTGTGGGGTC	>chr10:98325169→98336394 TGGCATCACATCATCACCTTTA
ENSG00000196367_003	TRRAP	chr7:98479599-98479643 (+)	>chr7:98478804→98478824 ACTTCCATTCTGTGTGGGGTC	>chr7:98479636→98479612 TGGCATCACATCATCACCTTTA
ENSG00000196367_008	TRRAP	chr7:98493400-98493443 (+)	>chr7:98491461→98491486 TATTGAGCTACACAAACAGTTCAGG	>chr7:98493443→98493420 CACTACTTTTGGAAGCTCCTTGT
ENSG00000118816_011	CCNI	chr4:77996323-77996416 (-)	>chr4:77996379→77996359 GGCCAGTTTTTCTGTCAGCG	>chr4:77987533→77987553 CTGCTACCCAGCTTGCTGTA
ENSG00000172890_047	NADSYN1	chr11:71210454-71210599 (+)	>chr11:71210447→71210467 GACAGGTCTGGCATGCACTC	>chr11:71210592→71210572 GCCATGACGCCATCTGGATA
ENSG00000143119_007	CD53	chr1:111435158-111435338 (+)	>chr1:111434081→111435164 TTGCTCTTTTGGGGAGTCCTT	>chr1:111435296→111435276 GCAGACTTGGCTGATCTGAGG

References

Affymetrix. (2005a, 11-10-2005). *Alternative transcript analysis methods for exon arrays* (Tech. Rep.). Affymetrix Inc.

Affymetrix. (2005b, 27-09-2005). *Exon array background correction* (Tech. Rep.). Affymetrix Inc.

Affymetrix. (2005c, 27-09-2005). *Exon probeset annotations and transcript cluster groupings* (Tech. Rep.). Affymetrix Inc.

Affymetrix. (2005d). *Guide to probe logarithmic intensity error (PLIER) estimation* (Tech. Rep.). Affymetrix Inc.

Affymetrix. (2005e). *GeneChip® exon array design* (Tech. Rep.). Affymetrix Inc.

Alarcón, B., & Martínez-Martín, N. (2012). Rras2, rhog and t-cell phagocytosis. *Small GTPases*, *3*(2), 97-101. Retrieved from `https://doi.org/10.4161/sgtp.19138` (PMID: 22790196) doi: 10.4161/sgtp.19138

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... Soboleva, A. (2013). Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991-D995. Retrieved from `+http://dx.doi.org/10.1093/nar/gks1193` doi: 10.1093/nar/gks1193

Belkaid, Y. (2007). Regulatory T cells and infection: a dangerous necessity. *Nat Rev Immunol*, *7*(11), 875-888. (10.1038/nri2189)

Bengtsson, H., Simpson, K., Bullard, J., & Hansen, K. (2008). *aroma.affymetrix: A generic framework in r for analyzing small to very large Affymetrix data sets in bounded memory* (Tech. Rep.). Department of Statistics, University of California.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. Retrieved from http://www.jstor.org/stable/2674075

Blat, Y., & Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell*, *98*(2), 249-259.

Bluestone, J. A., & Abbas, A. K. (2003). Natural versus adaptive regulatory T cells. *Nat Rev Immunol*, *3*(3), 253-257. (10.1038/nri1032)

Bolstad, B., Collin, F., Simpson, K., Irizarry, R., & Speed, T. (2004). Experimental design and low-level analysis of microarray data. In *Dna arrays in neurobiology* (Vol. 60, p. 25 - 58). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S007477420460002X doi: https://doi.org/10.1016/S0074-7742(04)60002-X

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016, May). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, *34*(5), 525–527.

Bresatz, S., Sadlon, T., Millard, D., Zola, H., & Barry, S. C. (2007). Isolation, propagation and characterization of cord blood derived CD4+ CD25+ regulatory T cells. *Journal of Immunological Methods*, *327*(1–2), 53-62.

Brunkow, M. E., Jeffery, E. W., Hjerrild, K. A., Paeper, B., Clark, L. B., Yasayko, S. A., ... Ramsdell, F. (2001). Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. *Nat Genet*, *27*(1), 68-73. (Brunkow, M E Jeffery, E W Hjerrild, K A Paeper, B Clark, L B Yasayko, S A Wilkinson, J E Galas, D Ziegler, S F Ramsdell, F United states Nature genetics Nat Genet. 2001 Jan;27(1):68-73.)

Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., ... Brown, M. (2006, Nov). Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.*, *38*(11), 1289–1297.

Chang, C. H., & Pearce, E. L. (2016, Apr). Emerging concepts of T cell metabolism as a target of immunotherapy. *Nat. Immunol.*, *17*(4), 364–368.

Chen, W., Jin, W., Hardegen, N., Lei, K.-j., Li, L., Marinos, N., ... Wahl, S. M. (2003). Conversion of peripheral CD4+CD25- naive t cells to CD4+CD25+ regulatory t cells by TGF-βinduction of transcription factor foxp3. *The Journal of Experimental Medicine*, *198*(12), 1875-1886.

Choudhuri, K., Llodra, J., Roth, E. W., Tsai, J., Gordo, S., Wucherpfennig, K. W., ... Dustin, M. L. (2014, Mar). Polarized release of T-cell-receptor-enriched microvesicles at the immunological synapse. *Nature*, *507*(7490), 118–123.

Clark, T. A., Schweitzer, A. C., Chen, T. X., Staples, M. K., Lu, G., Wang, H., … Blume, J. E. (2007). Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol*, *8*(4), R64. (Clark, Tyson A Schweitzer, Anthony C Chen, Tina X Staples, Michelle K Lu, Gang Wang, Hui Williams, Alan Blume, John E England Genome biology Genome Biol. 2007;8(4):R64.)

Clark, T. A., Sugnet, C. W., & Ares, J., M. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, *296*(5569), 907-10. (Clark, Tyson A Sugnet, Charles W Ares, Manuel Jr CA77813/CA/NCI NIH HHS/United States GM40478/GM/NIGMS NIH HHS/United States Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United States Science (New York, N.Y.) Science. 2002 May 3;296(5569):907-10.)

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., … Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, *33*(20), e175.

Davidson, T. S., DiPaolo, R. J., Andersson, J., & Shevach, E. M. (2007). Cutting edge: IL-2 is essential for TGF-β-mediated induction of foxp3+ t regulatory cells. *The Journal of Immunology*, *178*(7), 4022-4026.

Demerath, N. J. (1949). The american soldier: Volume i, adjustment during army life. by s. a. stouffer, e. a. suchman, l. c. devinney, s. a. star, r. m. williams, jr. volume ii, combat and its aftermath. by s. a. stouffer, a. a. lumsdaine, m. h. lumsdaine, r. m. williams, jr., m. b. smith, i. l. janis, s. a. star, l. s. cottrell, jr. princeton, new jersey: Princeton university press, 1949. vol. i, 599 pp., vol. ii, 675 pp. 7.50*each*;13.50 together. *Social Forces*, *28*(1), 87-90. Retrieved from `+http://dx.doi.org/10.2307/2572105` doi: 10.2307/2572105

Draber, P., Vonkova, I., Stepanek, O., Hrdinka, M., Kucova, M., Skopcova, T., … Brdicka, T. (2011). Scimp, a transmembrane adaptor protein involved in major histocompatibility complex class ii signaling. *Molecular and Cellular Biology*, *31*(22), 4550-4562. Retrieved from `http://mcb.asm.org/content/31/22/4550.abstract` doi: 10.1128/MCB.05817-11

Du, J., Huang, C., Zhou, B., & Ziegler, S. F. (2008). Isoform-specific inhibition of RORα-mediated transcriptional activation by human FOXP3. *The Journal of Immunology*, *180*(7), 4785-4792.

Dudoit, S., Yang, Y. H., Callow, M., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.

Statistica Sinica, *12*(1), 111-139.

Efron, B., & Morris, C. (1972). Limiting the risk of bayes and empirical bayes estimators–part ii: The empirical bayes case. *Journal of the American Statistical Association*, *67*(337), 130–139. Retrieved from `http://www.jstor.org/stable/2284711`

Fontenot, J. D., Rasmussen, J. P., Williams, L. M., Dooley, J. L., Farr, A. G., & Rudensky, A. Y. (2005). Regulatory t cell lineage specification by the forkhead transcription factor foxp3. *Immunity*, *22*(3), 329-341.

Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, *13*(3), 539-552. Retrieved from `+http://dx.doi.org/10.1093/biostatistics/kxr034` doi: 10.1093/biostatistics/kxr034

Gaidatzis, D., Jacobeit, K., Oakeley, E. J., & Stadler, M. B. (2009). Overestimation of alternative splicing caused by variable probe characteristics in exon arrays. *Nucleic Acids Research*, *37*(16), e107.

Gardina, P., Clark, T., Shimada, B., Staples, M., Yang, Q., Veitch, J., … Turpaz, Y. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, *7*(1), 325.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall / CRC.

Geman, S., & Geman, D. (1984, Nov). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *PAMI-6*(6), 721-741. doi: 10.1109/TPAMI.1984.4767596

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Gerisch, G. (2010, Mar 18). Self-organizing actin waves that simulate phagocytic cup structures. *PMC Biophysics*, *3*(1), 7. Retrieved from `https://doi.org/10.1186/1757-5036-3-7` doi: 10.1186/1757-5036-3-7

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex bayesian modelling. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *43*(1), 169-177.

Gilmour, D. S., & Lis, J. T. (1985). In vivo interactions of RNA polymerase II with genes of drosophila melanogaster. *Molecular and Cellular Biology*, *5*(8), 2009-2018.

Godfrey, W. R., Spoden, D. J., Ge, Y. G., Baker, S. R., Liu, B., Levine, B. L., … Porter, S. B. (2005). Cord blood CD4+CD25+-derived T regulatory cell lines express FOXP3 protein and manifest potent suppressor function. *Blood*, *105*(2), 750-758.

Gomez, T. S., & Billadeau, D. D. (2008). T cell activation and the cytoskeleton: you can't have one without the other. *Adv. Immunol.*, *97*, 1–64.

Goodson, M. L., Jonas, B. A., & Privalsky, M. L. (2005). Alternative mRNA splicing of SMRT creates functional diversity by generating corepressor isoforms with different affinities for different nuclear receptors. *Journal of Biological Chemistry*, *280*(9), 7493-7503.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97-109. Retrieved from `http://biomet.oxfordjournals.org/content/57/1/97.abstract` doi: 10.1093/biomet/57.1.97

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York: Wiley. (82008528 US edited by David C. Hoaglin, Frederick Mosteller, John W. Tukey ill ; 24 cm. Includes bibliographical references (p. 427-4290 and index)

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, *6*(9), 813-827.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346-363.

Hubbell, E., Liu, W.-M., & Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, *18*(12), 1585-1592.

Huber, P. J. (2005). The basic types of estimates. In *Robust statistics* (p. 43-72). John Wiley & Sons, Inc.

Huse, M., Quann, E. J., & Davis, M. M. (2008, Oct). Shouts, whispers and the kiss of death: directional secretion in T cells. *Nat. Immunol.*, *9*(10), 1105–1111.

Hyatt, G., Melamed, R., Park, R., Seguritan, R., Laplace, C., Poirot, L., . . . Benoist, C. (2006). Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol*, *7*(7), 686-691. (10.1038/ni0706-686)

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249-264.

Irizarry, R. A., Wu, Z., & Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, *22*(7), 789-794.

Janeway, J., Charles A, Travers, P., Walport, M., & Shlomchik, M. J. (2001a). Antigen recognition by T cells. In *Immunobiology: The immune system in health and disease* (5th ed.). New York: Garland Science.

Janeway, J., Charles A, Travers, P., Walport, M., & Shlomchik, M. J. (2001b). Generation of lymphocytes in bone marrow and thymus. In *Immunobiology: The immune system in health and disease* (5th ed.). New York: Garland Science.

Janeway, J., Charles A, Travers, P., Walport, M., & Shlomchik, M. J. (2001c). The production of armed effector T cells. In *Immunobiology: The immune system in health and disease* (5th ed.). New York: Garland Science.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *186*(1007), 453–461. Retrieved from `http://rspa.royalsocietypublishing.org/content/186/1007/453` doi: 10.1098/rspa.1946.0056

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, *316*(5830), 1497–1502. Retrieved from `http://science.sciencemag.org/content/316/5830/1497` doi: 10.1126/science.1141319

Johnson, W. E., Li, W., Meyer, C. A., Gottardo, R., Carroll, J. S., Brown, M., & Liu, X. S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences*, *103*(33), 12457-12462.

Julius, M. H., Masuda, T., & Herzenberg, L. A. (1972). Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proc Natl Acad Sci U S A*, *69*(7), 1934-8. (Julius, M H Masuda, T Herzenberg, L A United states Proceedings of the National Academy of Sciences of the United States of America Proc Natl Acad Sci U S A. 1972 Jul;69(7):1934-8.)

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. A., & Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, *296*(5569), 916-919.

Kapur, K., Xing, Y., Ouyang, Z., & Wong, W. (2007). Exon arrays provide accurate assessments of gene expression. *Genome Biology*, *8*(5), R82.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., ... Brazma, A. (2015). Arrayexpress update—simplifying data submissions. *Nucleic Acids Research*, *43*(D1), D1113-D1116. Retrieved from `+http://dx.doi.org/10.1093/nar/gku1057` doi: 10.1093/nar/gku1057

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014, Feb 03). voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, *15*(2), R29. Retrieved from `https://doi.org/10.1186/gb-2014-15-2-r29` doi: 10.1186/gb-2014-15-2-r29

Lee, M.-L. T., & Whitmore, G. A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, *21*(23), 3543-3570.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014, Dec). Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983-1992. doi: 10.1109/TVCG.2014.2346248

Li, W., Meyer, C. A., & Liu, X. S. (2005). A hidden markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, *21*(suppl 1), i274-i282.

Liu, W., Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., ... Smeekens, S. P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, *18*(12), 1593-1599.

Liu, W., Putnam, A. L., Xu-yu, Z., Szot, G. L., Lee, M. R., Zhu, S., ... Bluestone, J. A. (2006). CD127 expression inversely correlates with foxp3 and suppressive function of human CD4+ t reg cells. *The Journal of Experimental Medicine*, *203*(7), 1701-1711.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., ... Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech*, *14*(13), 1675-1680. (10.1038/nbt1296-1675)

Love, M. I., Hogenesch, J. B., & Irizarry, R. A. (2016, Dec). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.*, *34*(12), 1287–1291.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000a). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325-337. Retrieved from `http://dx.doi.org/10.1023/A%3A1008929526011` doi: 10.1023/A:1008929526011

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000b). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325-337. Retrieved from `http://dx.doi.org/10.1023/A%3A1008929526011` doi: 10.1023/A:1008929526011

Mestas, J., & Hughes, C. C. W. (2004). Of mice and not men: Differences between mouse and human immunology. *The Journal of Immunology*, *172*(5), 2731-2738.

Mittelbrunn, M., Gutierrez-Vazquez, C., Villarroya-Beltri, C., Gonzalez, S., Sanchez-Cabo, F., Gonzalez, M. A., … Sanchez-Madrid, F. (2011). Unidirectional transfer of microRNA-loaded exosomes from T cells to antigen-presenting cells. *Nat Commun*, *2*, 282.

Mohandas, A., Zola, H., Barry, S., & Krumbiegel, D. (2014). Peptidase inhibitor 16 identifies a unique subset of memory t helper cells with hyperproliferative and proinflammatory properties (irc8p.477). *The Journal of Immunology*, *192*(1 Supplement), 190.5–190.5. Retrieved from `http://www.jimmunol.org/content/192/1_Supplement/190.5`

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008, Jul). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, *5*(7), 621–628.

Nicholson, I. C., Mavrangelos, C., Bird, D. R. G., Bresatz-Atkins, S., Eastaff-Leung, N. G., Grose, R. H., … Krumbiegel, D. (2012). PI16 is expressed by a subset of human memory treg with enhanced migration to CCL17 and CCL20. *Cellular Immunology*, *275*(1–2), 12-18.

Ocklenburg, F., Moharregh-Khiabani, D., Geffers, R., Janke, V., Pfoertner, S., Garritsen, H., … Probst-Kepper, M. (2006). UBD, a downstream element of FOXP3, allows the identification of LGALS3, a new marker of human regulatory T cells. *Lab Invest*, *86*(7), 724-737.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, *40*(12), 1413-5. (Pan, Qun Shai, Ofer Lee, Leo J Frey, Brendan J Blencowe, Benjamin J Research Support, Non-U.S. Gov't United States Nature genetics Nat Genet. 2008 Dec;40(12):1413-5. Epub 2008 Nov 2.)

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015, Mar). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, *33*(3), 290–295.

Pfoertner, S., Jeron, A., Probst-Kepper, M., Guzman, C., Hansen, W., Westendorf, A., … Geffers, R. (2006). Signatures of human regulatory T cells: an encounter with old friends and new players. *Genome Biology*, *7*(7), R54.

Purdom, E., Simpson, K. M., Robinson, M. D., Conboy, J. G., Lapuk, A. V., & Speed, T. (2008). FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, *24*(15), 1707-1714.

Qiu, W., Lee, M.-L. T., & Whitmore, G. A. (2006). *sizepower: Sample size and power calculation in micorarray studies.* (R package version 1.26.0)

R Development Core Team. (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Ritchie, M., Diyagama, D., Neilson, J., van Laar, R., Dobrovic, A., Holloway, A., & Smyth, G. (2006). Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics*, *7*(1), 261.

Sadlon, T. J., Wilkinson, B. G., Pederson, S., Brown, C. Y., Bresatz, S., Gargett, T., ... Barry, S. C. (2010). Genome-wide identification of human FOXP3 target genes in natural regulatory T cells. *The Journal of Immunology*, *185*(2), 1071-1081.

Sakaguchi, S. (2004). Naturally arising CD4+ regulatory T cells for immunologic self-tolerance and negative control of immune responses. *Annu Rev Immunol*, *22*, 531-62. (Sakaguchi, Shimon Research Support, Non-U.S. Gov't Review United States Annual review of immunology Annu Rev Immunol. 2004;22:531-62.)

Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M., & Toda, M. (1995). Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *The Journal of Immunology*, *155*(3), 1151-64.

Sanchez-Graillet, O., Rowsell, J., Langdon, W. B., Stalteri, M., Arteaga-Salas, J. M., Upton, G. J., & Harrison, A. P. (2008). Widespread existence of uncorrelated probe intensities from within the same probeset on affymetrix GeneChips. *J Integr Bioinform*, *5*(2). (Sanchez-Graillet, Olivia Rowsell, Joanna Langdon, William B Stalteri, Maria Arteaga-Salas, Jose M Upton, Graham J G Harrison, Andrew P BB/E001742/1/Biotechnology and Biological Sciences Research Council/United Kingdom BBS/S/H/2005/11996A/Biotechnology and Biological Sciences Research Council/United Kingdom Research Support, Non-U.S. Gov't Germany Journal of integrative bioinformatics J Integr Bioinform. 2008 Aug 25;5(2). doi: 10.2390/biecoll-jib-2008-98.)

Sandberg, R., & Larsson, O. (2007). Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, *8*(1), 48.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467-470.

Schmitz, U., Pinello, N., Jia, F., Alasmari, S., Ritchie, W., Keightley, M.-C., ... Rasko, J. E. J. (2017, Nov 16). Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biology*, *18*(1), 216. Retrieved from `https://doi.org/10.1186/`

s13059-017-1339-3 doi: 10.1186/s13059-017-1339-3

Setoguchi, R., Hori, S., Takahashi, T., & Sakaguchi, S. (2005). Homeostatic maintenance of natural foxp3+ CD25+ CD4+ regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *The Journal of Experimental Medicine*, *201*(5), 723-735.

Silkworth, J. B., Carlson, E. A., McCulloch, C., Illouz, K., Goodwin, S., & Sutter, T. R. (2008). Toxicogenomic analysis of gender, chemical, and dose effects in livers of TCDD- or aroclor 1254–exposed rats using a multifactor linear model. *Toxicological Sciences*, *102*(2), 291-309.

Smyk-Pearson, S. K., Bakke, A. C., Held, P. K., & Wildin, R. S. (2003). Rescue of the autoimmune scurfy mouse by partial bone marrow transplantation or by injection with T-enriched splenocytes. *Clinical & Experimental Immunology*, *133*(2), 193-199.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, *3*, Article3. (Smyth, Gordon K United States Statistical applications in genetics and molecular biology Stat Appl Genet Mol Biol. 2004;3:Article3. Epub 2004 Feb 12.)

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, & W. Huber (Eds.), *Bioinformatics and computational biology solutions using r and bioconductor* (pp. 397–420). New York: Springer.

Smyth, G. K., Michaud, J., & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, *21*(9), 2067. Retrieved from +http://dx.doi.org/10.1093/bioinformatics/bti270 doi: 10.1093/bioinformatics/bti270

Soldatov, N. M. (2015). Cacnb2: An emerging pharmacological target for hypertension, heart failure, arrhythmia and mental disorders. *Current Molecular Pharmacology*, *8*(1), 32-42. Retrieved from http://www.eurekaselect.com/node/131114/article doi: 10.2174/1874467208666150507093258

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*(3), 479–498. Retrieved from http://www.jstor.org/stable/3088784

Sturtz, S., Ligges, U., & Gelman, A. (2005a). R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, *12*(3), 1–16. Retrieved from http://www.jstatsoft.org

Sturtz, S., Ligges, U., & Gelman, A. (2005b). R2winBUGS: A package for running winBUGS from r. *Journal of Statistical Software*, *12*(3), 1-16.

The Ensembl Project. (2012, 06-Aug-2012). *Ensembl release 68.*

Tierney, L., Rossini, A., & Li, N. (2009). Snow: A parallel computing framework for the r system. *International Journal of Parallel Programming*, *37*(1), 78-90. Retrieved from `http://dx.doi.org/10.1007/s10766-008-0077-2` doi: 10.1007/s10766-008-0077-2

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012, Mar). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, *7*(3), 562–578.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, Mass: Addison-Wesley Pub. Co. (76005080 John W. Tukey ill ; 25 cm. On spine: EDA Includes index)

Turro, E., Lewin, A., Rose, A., Dallman, M. J., & Richardson, S. (2010). MMBGX: a method for estimating expression at the isoform level and detecting differential splicing using whole-transcript Affymetrix arrays. *Nucleic Acids Research*, *38*(1), e4.

Vukmanovic-Stejic, M., Zhang, Y., Cook, J. E., Fletcher, J. M., McQuaid, A., Masters, J. E., . . . Akbar, A. N. (2006). Human CD4+ CD25hi foxp3+ regulatory T cells are derived by rapid turnover of memory populations in vivo. *The Journal of Clinical Investigation*, *116*(9), 2423-2433. (10.1172/JCI28941)

Walker, M. R., Kasprowicz, D. J., Gersuk, V. H., Bènard, A., Van Landeghen, M., Buckner, J. H., & Ziegler, S. F. (2003). Induction of foxp3 and acquisition of t regulatory activity by stimulated human CD4+CD25– T cells. *The Journal of Clinical Investigation*, *112*(9), 1437-1443. (10.1172/JCI19441)

Wildin, R. S., Ramsdell, F., Peake, J., Faravelli, F., Casanova, J.-L., Buist, N., . . . Brunkow, M. E. (2001). X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. *Nat Genet*, *27*(1), 18-20. (10.1038/83707)

Wildin, R. S., Smyk-Pearson, S., & Filipovich, A. H. (2002). Clinical and molecular features of the immunodysregulation, polyendocrinopathy, enteropathy, x linked (IPEX) syndrome. *Journal of Medical Genetics*, *39*(8), 537-545.

Williams, L. M., & Rudensky, A. Y. (2007). Maintenance of the foxp3-dependent developmental program in mature regulatory t cells requires continued expression of foxp3. *Nat Immunol*, *8*(3), 277-284. (10.1038/ni1437)

Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., . . . Svejstrup, J. Q. (2017, Feb). UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene. *Cell*, *168*(5), 843–855.

Wood, S. N., & Augustin, N. H. (2002). {GAMs} with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, *157*(2–3), 157 - 177. Retrieved from `http://www.sciencedirect.com/science/article/pii/S030438000200193X` doi: http://dx.doi.org/10.1016/S0304-3800(02)00193-X

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., & Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, *99*(468), 909-917.

Wu, Z., Irizarry, R. A., MacDonald, J., & Gentry, J. (2011). *gcrma: Background adjustment using sequence information.*

Yagi, H., Nomura, T., Nakamura, K., Yamazaki, S., Kitawaki, T., Hori, S., … Sakaguchi, S. (2004). Crucial role of FOXP3 in the development and function of human CD25+CD4+ regulatory T cells. *International Immunology*, *16*(11), 1643-1656.

Zakharkin, S., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K., … Page, G. (2005). Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*, *6*(1), 214.

Zhang, C., Zhang, B., Lin, L.-L., & Zhao, S. (2017, Aug 07). Evaluation and comparison of computational tools for rna-seq isoform quantification. *BMC Genomics*, *18*(1), 583. Retrieved from `https://doi.org/10.1186/s12864-017-4002-1` doi: 10.1186/s12864-017-4002-1

Zhou, L., Lopes, J. E., Chong, M. M., Ivanov, I. I., Min, R., Victora, G. D., … Littman, D. R. (2008, May). TGF-beta-induced Foxp3 inhibits T(H)17 cell differentiation by antagonizing RORgammat function. *Nature*, *453*(7192), 236–240.

Zola, H., Swart, B., Banham, A., Barry, S., Beare, A., Bensussan, A., … Yang, X. (2007). CD molecules 2006 — human cell differentiation molecules. *Journal of Immunological Methods*, *319*(1–2), 1-5.