<table>
<thead>
<tr>
<th>Title</th>
<th>Estimation and analysis of gene expression and alternative splicing: perspectives from development and disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Korir, Paul Kibet</td>
</tr>
<tr>
<td>Publication Date</td>
<td>2014-01-30</td>
</tr>
<tr>
<td>Item record</td>
<td><a href="http://hdl.handle.net/10379/4301">http://hdl.handle.net/10379/4301</a></td>
</tr>
</tbody>
</table>
Estimation and Analysis of Gene Expression and Alternative Splicing: Perspectives from Development and Disease

by

Paul Kibet Korir

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

College of Science
School of Mathematics, Statistics and Applied Mathematics

Supervised by Professor Cathal Seoighe
Co-supervised by Dr. Tim Downing

April 2014
Declaration of Authorship

I, Paul Kibet Korir, declare that this thesis titled, ‘Estimation and Analysis of Gene Expression and Alternative Splicing: Perspectives from Development and Disease’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- 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.

Signed: [Signature]

Date: 15/04/2014
“gloria Dei celare verbum et gloria regum investigare sermonem”

Proverbia 25:2 (VULGATE)
Abstract

The development of high-throughput genomics technologies has contributed substantially to the understanding of gene expression regulation. With the growing appreciation of the importance of alternative splicing, quantitative techniques have had to keep in step with the demand for an accurate and high resolution view of the transcriptome. In this thesis, we use results and methods from quantitative genomics to explore how gene regulation may be modified in development and disease.

Precise regulation of gene expression timing can be critical for some biological processes. This is particularly the case for genes with oscillating patterns of expression. Oscillations can be brought about through negative feedback loops, with a delay between gene activation and negative autoregulation. The time required for gene transcription contributes to the delay in gene activation and, thus, the intron content of genes involved in negative autoregulatory loops can be functionally significant. An example of this occurs in *Hes7*, in which oscillation is coupled to the formation of segmental body plans during animal development. To identify further examples of genes in which the transcriptional delay introduced by introns may be functionally significant, we carried out a search for genes with conserved intron content across a diverse panel of 19 mammals and found that the set of genes with the most extreme conservation was enriched for genes involved in embryonic development. We found that these genes had both fewer insertions and deletions as well as a balance between the cumulative insertions and deletions, suggesting that selection functions to prevent indels in these introns and to balance the impact of insertions and deletions.

There has been considerable success in mapping local (cis) variants associated with phenotypes such as disease. Many cis-acting variants that cause disease disrupt splicing. However, mapping distant (trans) acting variants that affect splicing is a more formidable task. By exploiting high-density transcriptome microarrays, we show that a mutation in the splicing factor *PRPF8*, causally associated with retinitis pigmentosa, is a trans-acting splicing variant. We estimate that up to 20% of all exons are mis-spliced through higher exon inclusion in affected individuals. Characteristics of affected exons suggest that they tend to be spliced co-transcriptionally and via the exon-defined splicing pathway.

The importance of gene expression microarrays in quantitative genomics has led to the development of numerous algorithms to estimate gene expression from raw microarray intensities. But microarrays have several shortcomings relative to more recently developed sequencing-based methods for measuring gene expression. We exploited the
benefits of quantitative transcriptome sequencing (RNA-Seq) by using a statistical learning approach to obtain better expression estimates from arrays, based on a high-quality dataset for which both microarray and RNA-Seq data are available. Our analyses show that this approach compares favourably to existing algorithms for microarray analysis, with the added advantage of providing estimates of the abundance of individual transcript isoforms on an absolute scale.
Acknowledgements

I am, first and foremost, grateful to Almighty God, Maker of all things, for the strength to complete this work. I count it as non-trivial that during the study period I have retained excellent health. The words of King David echo my sentiments when he writes,

The Lord is the portion of mine inheritance and of my cup: thou maintainest my lot.
The lines are fallen unto me in pleasant places; yea, I have a goodly heritage.
I will bless the Lord, who hath given me counsel: my reins also instruct me in the night seasons.
I have set the Lord always before me: because he is at my right hand, I shall not be moved.
Therefore my heart is glad, and my glory rejoiceth: my flesh also shall rest in hope.

Psalm 16:5-9 (KJV)

A million thanks go to Prof. Cathal Seoighe, my main supervisor for being a superb advisor through this journey. I greatly admire his discipline, integrity, breadth of knowledge, and above all his patience. I have learned so much from him and he has left a deep mark on me that I hope to live up to. I also thank Dr. Tim Downing, my co-supervisor for his numerous helpful comments. I will forever remember how he patiently read through some of the early drafts even when he was a new member of staff having not seen most of this material before.

I cannot say thank you enough to Maureen Olembo, my fiancée and friend. I am especially grateful for keeping me accountable, and encouraging me and praying with me through the thick and thin of the last few years. Words are not enough, Mo!

I am also deeply indebted to the staff and fellow graduates students both in the bioinformatics research group and the wider School of Maths community. Many stimulating discussions have passed between us and it has been this atmosphere of inquisitiveness and curiosity that that I have deeply enjoyed. In particular, I would like to single out Paul Geeleher, Thong Nguyen and Peter Keane for many helpful debates and discussions—not all of which ended with agreement!

Lastly, but certainly not least, I owe an enormous debt of gratitude to my family for always being there, in particular my sister Winnie Nasurutia for her frequent phone calls and for the Galway Christian Fellowship for their prayers and support.

Paul Kibet Korir
Galway
January 2014
# Contents

Declaration of Authorship iii

Abstract vii

Acknowledgements ix

List of Figures xv

List of Tables xxiii

1 Introduction 1

2 Introns, Splicing and Alternative Splicing 5

2.1 Introns .......................................................... 5
  2.1.1 Prevalence, Designations and Structural Characterisation ... 6
  2.1.2 Origin and Evolution ........................................... 11
  2.1.3 Functions of Introns .......................................... 12

2.2 Splicing .......................................................... 14
  2.2.1 The Spliceosome ............................................... 15
  2.2.2 The Splicing Pathway ......................................... 16
  2.2.3 Transesterification Reactions in Splicing .................... 18
  2.2.4 Other Splicing Pathways .................................... 19
  2.2.5 Exon-defined Splicing ....................................... 22
  2.2.6 ‘Extreme’ Splicing ........................................... 24
  2.2.7 Timing of Splicing Relative to Transcription Termination .... 27

2.3 Alternative Splicing ........................................... 28
  2.3.1 Prevalence ................................................... 30
  2.3.2 Origin and Evolution ........................................ 31
  2.3.3 Functions of Alternative Splicing ............................ 33
  2.3.4 Modulation of Alternative Splicing ........................... 35
  2.3.5 Stochasticity of Alternative Splicing ....................... 39

2.4 Measurement of Transcript Expression ........................ 41
  2.4.1 Genes ....................................................... 41
  2.4.2 Transcript Isoforms ........................................ 43
## Contents

3 Evidence for intron length conservation in a set of mammalian genes associated with embryonic development 47

3.1 Introduction ................................................................. 48
3.2 Data and Methods ......................................................... 49

3.2.1 Randomization study of intron length evolution along the human lineage ................................................... 50
3.3 Results and Discussion ..................................................... 51

3.3.1 Intronic insertions and deletions along lineages leading to human and mouse ........................................... 54
3.3.2 Contribution of conserved sequence elements to intron length conservation .............................................. 56
3.4 Conclusion ................................................................. 57

4 A mutation in a splicing factor that causes retinitis pigmentosa has a transcriptome-wide effect on mRNA splicing 59

4.1 Introduction ................................................................. 60
4.2 Methods ................................................................. 62

4.2.1 Exon array sample preparation and data generation .......... 62
4.2.2 Expression quantification of microarrays ......................... 63
4.2.3 Characterisation of differentially spliced exons ................. 64
4.3 Results ................................................................. 64

4.3.1 Transcriptome-wide perturbation of splicing ..................... 64
4.3.2 Differential splicing primarily involves higher inclusion of exons in cases .................................................. 66
4.3.3 Characterisation of differentially included exons ............... 66
4.4 Discussion ................................................................. 70
4.5 Conclusion ................................................................. 73

5 Seq-ing improved gene expression estimates from microarrays using machine learning 75

5.1 Introduction ................................................................. 76
5.2 Methods ................................................................. 77

5.2.1 GTEx data preparation .................................................. 77
5.2.2 Selecting and tuning the best learning algorithm ............... 78
5.2.3 Effect of training set size ................................................ 79
5.2.4 Differential gene expression (DGE) between GTEx tissues .... 79
5.2.5 Application to archived samples ..................................... 80
5.2.6 Application to the Affymetrix tissue mixture dataset ........... 81
5.3 Results ................................................................. 81

5.3.1 Correlation with RNA-Seq .............................................. 82
5.3.2 Restricting to well-estimated genes ................................. 85
5.3.3 Inference of differential expression .................................. 87
5.3.4 Estimation of transcript isoform expression ...................... 87
5.3.5 Application of MaLTE trained on GTEx data to independent microarray datasets ......................................... 88
5.3.6 Evaluation of MaLTE applied to a controlled tissue mixture dataset .................................................. 90
5.4 Discussion ................................................................. 91
6 Differential expression analysis on retinitis pigmentosa samples using RNA-Seq

6.1 Introduction .......................................................... 93
6.2 Materials and Methods ............................................... 95
  6.2.1 RNA-Seq ......................................................... 95
  6.2.2 RNA-Seq data analysis ......................................... 95
  6.2.3 Preparation of MaLTE and RMA expression estimates ....... 95
6.3 Results ...................................................................... 96
  6.3.1 Quality assessment of RNA-Seq data ......................... 96
  6.3.2 Identification of retinitis pigmentosa mutations from RNA-Seq data ........................................... 103
  6.3.3 Quality control optimisation ................................... 104
  6.3.4 Comparison of RNA-Seq to MaLTE and RMA ............. 108
  6.3.5 Comparison in differential expression analysis ............ 110
6.4 Discussion .................................................................. 114

7 Conclusion .................................................................. 117

7.1 Summary ................................................................. 117
7.2 Conclusions ............................................................. 118
  7.2.1 More emphasis on isoforms rather than genes ............ 118
  7.2.2 From algorithms to better quantitative assays ............. 118
7.3 Future work ............................................................. 119
  7.3.1 Conservation of introns .......................................... 119
  7.3.2 Retinitis pigmentosa transcriptome ......................... 119
  7.3.3 Further development of the MaLTE framework .......... 120

A Supplementary results for Chapter 4 .......................... 121
A.1 Quality Assessment .................................................. 122
A.2 Sample Permutations ............................................... 123
  12.2.1 Core Probesets .................................................... 123
  12.2.2 Full Probesets ..................................................... 124
A.3 Additional Analyses ................................................... 125
  13.1 Splice Site Scores .................................................... 125
  13.2 Constitutive versus Alternative Splicing Analyses .......... 125
  13.3 U12-type Introns ...................................................... 126
A.4 Number of Differentially Included Probesets ................ 128
  14.1 Differential Inclusion by Probeset Class .................... 128
  14.2 Core Probesets ....................................................... 129
  14.3 Full Probesets ......................................................... 130
  14.4 Exonic Probesets ..................................................... 131
  14.5 Intronic Probesets .................................................... 132
  14.6 U12-type Intronic Probesets ...................................... 133
A.5 Transcript Cluster Identifiers (Metaprobesets) Associated with PRPF8 .................................................. 133
A.6 Associated Genes ...................................................... 134
A.7 Associated Exons ...................................................... 134
  17.1 Lower Inclusion in Cases .................................. 141

B Supplementary results for Chapter 5 ............................ 143
List of Figures

1.1 Electron micrograph and schematic of adenoviral hexon mRNA hybridised to single-stranded genomic DNA selected using EcoR1 restriction enzyme (region A; see [2] for full details). Loops A, B and C are genomic DNA absent from mRNA and form loops between the hybridised regions. Image reproduced from [2]. ........................................ 1

2.1 Distribution of spliceosomal introns in eukaryotes. A wide selection of organisms are indicated together with the average number of introns per gene. Anopheles gambiae; Arabidopsis thaliana; Aspergillus nidulans; Bigelowiella natans; Nucleomorph; Caenorhabditis briggsae; Caenorhabditis elegans; Candida albicans; Chlamydomonas reinhardtii; Ciona intestinalis; Cryptococcus neoformans; Cryptosporidium parvum; Cyanidioschyzon merolae; Dictyostelium discoideum; Drosophila melanogaster; Encephalitozoon cuniculi; Giardia lamblia, Guillardia theta Nucleomorph; Homo sapiens; Leishmania major; Mus musculus; Neurospora crassa; Oryza sativa; Paramecium aurelia; Phanerochaete chrysosporium; Plasmidium falciparum; Plasmidium yoelii; Saccharomyces cerevisiae; Schizosaccharomyces pombe; Takifugu rubripes; Thalassiosira pseudonana; Trichomonas vaginalis. Image reproduced from [10]. ........................................ 6

2.2 Intron density and intron length in 100 eukaryotes. Image reproduced from [9]. ........................................ 7

2.3 Structure and splicing pathway of group I introns. Image reproduced from [13]. ........................................ 8

2.4 Structure of generic group II intron secondary structure. EBS: exon-binding site; IBS: intron-binding site; MGC: Manganese ion C; J2/3: junction between domains 2 and 3. Image reproduced from [18]. ............ 9

2.5 (a) Secondary structure of tRNA with the positions of known introns marked by numbers (in parentheses). The box at the bottom indicates the anticodon. (b) Conserved ‘bulge-helix-bulge’ in tRNA (left) and rRNA (right) with cleavage sites shown by arrows. Uppercase: > 85% conservation; lowercase: 60-85% conservation; unconserved indicated by dots; Purines - R; Pyrimdines - Y. Image reproduced from [21] with permission. 10

2.6 Conserved U2-type spliceosomal motifs in five divergent species. (A) 5’ splice site (5ss). (B) branch point site (BPS) where a highly conserved adenine is located. (C) 3’ splice site (3ss). Image reproduced from [27]. 11

2.7 Distribution of intron lengths in various organisms. Image reproduced from [30]. ........................................ 11

2.8 Origin of nucleus-cytosol compartmentalisation at the onset of mitochondrial origin. Image reproduced from [34] with permission. ............ 13
2.9 Short nuclear RNA (snRNA) involved in splicing. (a) Sm-class snRNA; (b) Lsm-class snRNA. Image reproduced from [46] with permission.

2.10 Splicing cascade showing recycling of spliceosomal components. Image reproduced from [49].

2.11 Schematic depicting first and second transesterification splicing reactions. Image reproduced from [68] with permission.

2.12 tRNA splicing cascade in archaeal introns showing ligated exons (top right) and circularised intron (bottom right). Image reproduced from [21] with permission.

2.13 Splicing pathways of group II introns. Image reproduced from [72].

2.14 Various splicing reactions exhibited by group II introns. (A) Forward and reverse splicing. (B) Hydrolytic splicing. (C) Partial reverse splicing. Introns indicated in red; exons indicated in dark/light blue (E1 and E2). Image reproduced from [17] with permission.

2.15 Schematic representing (A) cis- and (B) trans-splicing. (C) Conserved motifs and elements involved in cis-splicing (ESE: exonic splicing enhancers; 5’Ss: 5’ splice site; BP: branch point; PPT: poly-pyrimidine tract; 3’Ss: 3’ splice site). (D) Schematic representing the trans-splicing between (a) androgen-binding protein (ADP) and (b) histidine decarboxylase (HDC) transcripts to yield (c) a mature chimeric transcript. Image reproduced from [80].

2.16 Schematic representing exon definition including transition to intron definition via exon justaposition. Image reproduced from [29] with permission.

2.17 Three examples of ultra-short introns together with their sequence. Images reproduced from [91] with permission.

2.18 Three mechanisms of splicing in ultra-long introns. Image reproduced from [97] with permission.

2.19 α-synuclein (SNCA) gene showing 11 transcript isoforms of various lengths and with different exon assemblies. Image from Ensembl [117].

2.20 Patterns of alternative splicing. (a) Exon skipping/cassette exons - most prevalent in higher eukaryotes but low prevalence in plants; (b & c) alternative 3’ splice sites (A3SS) and alternative 5’ splice sites (A5SS) - considered as intermediates between constitutive and alternative splicing; (d) Mutually exclusive exons; (e) Intron retention. Image reproduced from [127] with permission.

2.21 An example of extreme alternative splicing in the Drosophila Dscam gene that potentially generates 38,016 transcript isoforms through mutually exclusive splicing of four exon clusters (coloured red, blue, green, yellow). Image reproduced from [125] with permission.

2.22 Exonisation of Alu elements. (a) Mutations surrounding Alu elements that could give rise to new exons. (b) Transformation of RNA secondary structures by enzymatic editing of adenosine (A) to inosine (I), which is recognised as guanosine (G). In the presence of two adjacent Alu elements of opposite orientation can then create a functional 3’sss (AA to AG). Image reproduced from [127] with permission.
2.23 Exon transition from constitutive to alternative. (A) Degenerate splice signals that can form (Aa) A5SS of A3SS, (Ab) weak 5’ splice sites leading to skipping, (Ac) introduction of an exonic splicing suppressor, or (B) secondary RNA structure that sequester splice signals. Image reproduced from [127] with permission. 

2.24 Exon shuffling. Image reproduced from [127] with permission.

2.25 Proteomic modulation of alternative splicing. Serine-Arginine (SR) and heteronuclear ribonucleoproteins (hnRNP) classes of proteins are predominantly involved in influencing splicing decisions by binding to exonic/intronic splice enhancers (ESE/ISE) or silencers (ESS/ISS). Image reproduced from [114] with permission.

2.26 Epigenetic modulation of alternative splicing. (a) Histone modifications (H3K9ac) leads to an increase in RNA pol II elongation leading to skipping in neural cell adhesion molecule (NCAM) exon 18. (b) Histone modification (H3K36me3) at the fibroblast growth factor receptor 2 (FGFR2) recruits polypyrimidine tract binding (PTB) protein which suppresses an alternate exon. Images reproduced from [114] with permission.

2.27 Kinetic modulation of alternative splicing. (a) SR splicing factor 3 (SRSF3) is controlled by the phosphorylation state of the RNA pol II C-terminal domain (CTD), which then determines inclusion of alternate exons recruitment is affected. (b) The Mediator, which associates with transcription factors, recruits the splicing suppressor hnRNPL. (c) CCCTC-binding factor (CCTF) binds to unmethylated CG-rich DNA sequences and slows RNA pol II transcription allowing competitive splicing of a downstream exon. (d) DBIRD facilitates exclusion of A/T-rich exons by preventing RNA pol II slowing down at these exons. (e) Kinetic coupling following DNA damage, which leads to hyperphosphorylation of the RNA pol II CTD further inhibiting RNA pol II elongation. (f) SWI/SNF (a chromatin remodelling factor) recruits SAM68, which stalls poll II enhancing exon inclusion. Image reproduced from [114] with permission.

2.28 Splicing code developed for five mouse tissues. Tissue codes: CNS (C), muscle (M), embryo (E), digestive (D), and tissue-independent mixture (I). The code is read using the key on the top left of the image. For example, abundant CU-rich (nPTB) motifs (fifth row) are strongly determinative of high inclusion of the alternate exon if within 300 nt upstream. Image reproduced from [148] with permission.

2.29 An illustration of noisy splicing on the HERPUD1 locus. The top level shows average expression level at each position. Green marks exonic regions and black marks intronic regions. Middle panel shows identified exons from junction mapping RNA-Seq reads. Black bars are annotated exons; red bars are not annotated. Bottom panel shows Ensembl transcript isoforms. Image reproduced from [169].

2.30 A schematic showing various tools available for detecting and measuring splicing events using RNA-Seq. Image reproduced from [175].

3.1 Phylogeny. Phylogenetic tree illustrating relationships of the 19 mammal included in this study.
3.2 **Comparison of insertions and deletions.** The sum of the insertions and deletions that have taken place in genes in size class one along the human (a) and mouse (b) lineages since divergence from their common ancestor. Blue points, corresponding to the genes that showed as much or more intron content conservation than Hes7 are shown against a background of grey points, corresponding to all genes in the size class. All axes are truncated, but the truncated axes include all of the blue points. Grey points lying with cumulative insertions or deletions greater than 200 bp for human or 2000 bp for mouse are not shown.

3.3 phastCons scores across introns of the Hes7 gene

4.1 **Empirical distribution of p-values.** Each p-values was calculated using pairwise t-tests for a normalised probeset between cases and controls.

4.2 **Proportion of probesets indicating higher/lower inclusion in cases for binned p-value thresholds.** All bins have an equal width of $\Delta p = 0.05$ for p-values in the range [0, 1]. Red bars symbolise the proportion of probesets indicating higher inclusion in cases; similarly, blue bars refer to lower inclusion in cases. The absolute height of a bar of each colour represents the proportion of probesets with higher/lower inclusion in cases. The t-statistic was used to determine relative inclusion: $t > 0$ - higher inclusion in cases; $t < 0$ - lower inclusion in cases. Plot only shows core and exonic (full) probesets.

4.3 **Distributions of intron length.** Density plot of intron length (A) upstream and (B) downstream of differentially included exons in cases relative to controls. A similar plot for background exons (core exons) is provided for reference.

4.4 **Nucleotide content of differentially included exons.** The proportion of each nucleotide in differentially included exons (A) with exonic splicing enhancers (ESEs) present and (B) with ESEs removed.

5.1 **Estimation accuracy.** MaLTE provides an estimate of the RNA-Seq gene expression levels from microarray probe intensities. (a) The relative error (i.e. difference between MaLTE estimate and RNA-Seq, divided by the RNA-Seq value) as a function of the RNA-Seq expression level. Each point corresponds to a bin of 75 genes. The data represents all genes but with a random subset of 10 samples for each gene. Only relative errors below 2 and RNA-Seq values between 1 and 1000 are represented. Low expression genes were excluded due to high stochasticity for low read counts. A Loess regression line is shown in red, illustrating that MaLTE slightly underestimates RNA-Seq particularly for highly-expressed genes. (b) The distribution of relative error with percentage median error and median absolute error displayed with the median error indicated by the dashed red line.

5.2 **Within-sample correlation with RNA-Seq.** (a, b, c) Scatter plot of gene expression for a single exemplary sample for each method against RNA-Seq. The sample with the within-sample Pearson correlation closest to the median over all samples was chosen. Box plots are provided to show the range of within-sample (d) Pearson and (e) Spearman correlation coefficients across samples in the test dataset. Median correlations are indicated beneath in brackets.
5.3 **Cross-sample correlation with RNA-Seq.** For each gene, the cross-sample correlation was determined between the gene expression values estimated from the microarrays using MaLTE, median-polish and PLIER. Density plots show the distribution across genes of (a) Pearson and (b) Spearman correlation coefficients. Mean and median values of the correlation coefficients are provided in parentheses next to the method name in the legend. Vertical lines show mean cross-sample correlation for MaLTE (solid), median-polish (dashed) and PLIER (dotted).

5.4 **The effects of OOB filtering.** Mean cross-sample Pearson correlation as a function of OOB correlation threshold for (a) genes and (b) transcripts. Error bars correspond to two standard errors. Note that transcript-level estimates are not provided by RMA and PLIER. The black line represents the number of genes/transcripts at each level.

5.5 **Application to archived data.** MaLTE, trained using the GTEx data, was applied to predict gene expression from published microarray data based on brain samples for which RNA-Seq data was also available. Despite the fact that the two studies used different array platforms (Affymetrix Human Exon 1.0 ST arrays and Affymetrix Human Gene 1.1 ST arrays for the brain and GTEx studies, respectively), MaLTE predictions exceeded the within-sample correlations obtained using median-polish and PLIER. MaLTE predictions were based on probes shared between the two array platforms. Box plots of (a) Pearson and (b) Spearman within correlations are shown. (c) Pearson and (d) Spearman cross sample correlations with OOB filtering. The black line represents the number of genes/transcripts at each level.

6.1 Plot of GC content found in each base position.

6.2 Plot of distribution of quality scores in each base position.

6.3 Plot of mean quality score for each short read sequence.

6.4 Plot of nucleotide percentage at each base position.

6.5 Plot of k-mer enrichment profile across short read sequences.

6.6 Effect of removing duplicates on the number of mapped reads.

6.7 **First two principal components for RNA-Seq expression estimates with and without various filtering approaches.** (a) No filtering applied, (b) filtering using the subset of genes that passed OOB filtering in the GTEx data (for comparison with MaLTE), (c) combining OOB filtering and removal of noisy genes, and (d) OOB filtering, removal of noisy genes then quantile-normalisation.

6.8 **First two principal components for MaLTE and RMA expression estimates.** (a) MaLTE prior to filtering out noisy genes, (b) MaLTE after filtering out noisy genes. (c and d) Corresponding plots for RMA.

6.9 **Density of expression (a) before and (b) after filtering noisy genes.** Both plots are truncated on the x-axis because the scale extends to -700 indicating numerous genes at very low expression levels. Truncation results in the shown densities not being normalised. Expression estimates (x-axis) are log-2 transformed.
6.10 **Effect of quantile normalisation of RNA-Seq data on nominal differential expression.** (a and b) Differential expression using standard $t$-tests. (a) No quantile normalisation, (b) quantile-normalised expression estimates. (c and d) Differential expression using moderated $t$-tests (performed with limma). (c) No quantile normalisation, (d) quantile-normalised expression estimates. Red denotes the proportion of genes that are expressed higher in cases, while blue denotes those higher in controls. Under the null hypothesis of no differences in expression both proportions should be equal at all $p$-values.

6.11 **Cross-sample and within-sample correlations.** (a) Pearson cross-sample correlations, (b) Spearman cross-sample correlations, (c) Pearson within-sample correlations, (d) Spearman within-sample correlations.

6.12 **Densities of cross-sample correlations after filtering and quantile normalisation.**

6.13 **Proportion of overlap with RNA-Seq to nominally significant genes using MaLTE and RMA.** (a) Standard $t$-tests, (b) Moderated $t$-tests (performed with limma). Each bar represents a single $p$-value threshold. The proportion is computed relative to the overlap between MaLTE and RMA. Error bars indicate two standard errors.

6.14 **$\alpha$-synuclein transcript isoforms on Ensembl.**

A.1 **Quality assessment of exon array data.** 

A.2 **Sample permutations for core probesets.** Plots representing the complete set (eight) of pairwise swaps between cases ($T_i$) and controls ($C_i$). Only core probesets are considered here. Plots are ordered from the top by the value of $\pi_0$. Dashed red line represents a uniform distribution; dashed blue lines represent the value of $\pi_0$.

A.3 **Sample permutations for full probesets.** Plots representing the complete set (eight) of pairwise swaps between cases ($T_i$) and controls ($C_i$). Plots are ordered from the top by the value of $\pi_0$. Dashed red line represents a uniform distribution; dashed blue lines represent the value of $\pi_0$.

A.4 **Distribution of splice site scores.** Density plot of splice site scores of A. 5’ splice sites and B. 3’ splice sites. The dashed line indicates a similar plot for background exons (core exons).

A.5 **Proportion of probesets indicating higher/lower inclusion in cases relative to controls for binned $p$-value thresholds for U12-type introns.** Each column represents all probesets (exons) for that bin. Red represent probesets included at higher levels in cases; blue are those included at lower levels in cases.

A.6 **Illustration of differential inclusion for different classes of probesets.** A. Core, B. full, C. exonic, and D. intronic probesets. Each point represents the number of probesets presenting higher (‘up’)/lower (‘down’) inclusion in cases relative to controls within $\Delta p = 0.05$ $p$-value bins.
B.1 The effect of training size on predictions. A test set of 100 samples was used for all training sample sizes. 143
B.2 Out-of-bag (OOB) filtering. Scatter plot of cross-sample correlation coefficients from o OOB gene expression estimates and estimates obtained on test data for 1,000 randomly selected genes. 144
B.3 Comparison of differential expression results. Comparison of the performance of MaLTE and median-polish on the problem of detecting differential gene expression between five heart-muscle and five skeletal-muscle samples in the GTEx dataset. Each method results in a list of genes, ranked by the q-value from the comparison of the gene expression level in the two groups of samples. We used the (a) cumulative Jaccard index and (b) concordance correlation to compare the similarity as a function of rank and the overall similarity, respectively, between lists of genes ranked by MaLTE/median-polish and by RNA-seq. The set of genes assessed were defined by RNA-Seq: genes with RPKM of above one in both tissues. Bootstrap re-sampling (100 pseudo-replicates) was used to assess the effect of sampling error for both cumulative Jaccard index and concordance correlations. Results of bootstrapping are shown as faint lines around the observed Jaccard index and yield the density plots shown in (b). Log of fold change estimates for (c) 120 differentially expressed genes common to MaLTE and median-polish and (d) MaLTE and PLIER for six common genes. 145
B.4 Comparison of 10-sample DGE to 44-sample DGE. (a) Self-comparisons (e.g. MaLTE 10-sample DGE to MaLTE 44-sample DGE, with five and 22 samples in each tissue, respectively). Forty-four-sample DGE results are treated as putative true differential expressions at the indicated (top right or each plot) false discover rates (FDRs). FDRs were computed using the q-value method [256]. Comparisons are made using the Jaccard index. (b) Comparison of each method to 44-sample DGE using RNA-Seq. 146
B.5 Transcript isoform expression estimates. Densities of (a) Pearson and (b) Spearman cross-sample correlations for transcript isoform expression estimates obtained using MaLTE. Filtered data corresponds to transcript isoforms with \( r_{OOB} > 0 \). 147
B.6 Box plots of slopes \( \beta \) computed by linearly regressing RNA-Seq against array method. Dotted red line indicates unit gradient. Each box plot represents 12 slopes for brain samples. RNA-Seq expression is restricted to RPKM between one and 1000 because of high uncertainty at low values and few genes at high values representing 78% of genes quantified. Using all genes results in the same medians but wider variance in slopes due to outlying genes. 148
B.7 Estimation of tissue mixture proportions. (a) MaLTE, (b) median-polish (RMA), and (c) PLIER. Each plot shows comparisons of true and estimated proportions with key statistics indicated in the legends. 149
C.1 Schematic representing a MaLTE use-case. The four text files (dark green) are the main input to MaLTE. These files are obtained by processing the training data using the auxiliary scripts. 152
List of Tables

3.1 Conservation of Hes7 intron length. Intron content of orthologues of human Hes7 (OrthoDB ID EOG45TDC4). The tarsier orthologue was omitted. This gene had no annotated introns but the annotation appeared to be incomplete. ............................................................... 52

3.2 Gene set enrichment analysis. Gene set enrichment analysis of genes in size class 1 showing evidence of intron content conservation in mammals. Top twenty most significantly enriched terms are shown. ........... 54

4.1 Length of introns flanking differentially included exons. Comparison of median of intron length for exons with higher (up) and lower (down) inclusion in cases relative to controls. ........................................... 67

4.2 Length of differentially included exons. Comparison of mean and median lengths for exons with higher (up) and lower (down) inclusion in cases relative to controls. .............................................. 68

4.3 Splice site strength of differentially included exons. Mean and median splice site strength computed using MaxEntScan [260]. ............. 69

4.4 Combined splice site strength (5SS and 3SS) of differentially included exons. Similar to Table 3 but using combined splice site strength. 69

4.5 Indicators of co-transcriptional splicing. Gene length and distance from the 3' end of genes for exons differentially included. ......................... 70

6.1 Summary of quality assessment on RP RNA-Seq data from FastQC and Picard. ................................................................. 97

6.2 RP-associated mutations identified using RNA-Seq data. Mutations identified in whole blood samples.(+) present, (-) absent. ............... 104

6.3 Genes identified as differentially expressed at FDR < 10%. All genes are expressed higher in RP cases. ........................................... 114

A.1 Number of core probesets in each bin of width Δp = 0.05. Data used to produce the plots in Fig. 4.2 and Fig. A.6. ................................. 129

A.2 Number of full probesets in each bin of width Δp = 0.05 ............ 130

A.3 Number of exonic probesets in each bin of width Δp = 0.05 ....... 131

A.4 Number of intronic probesets in each bin of width Δp = 0.05 ....... 132

A.5 Number of U12-type intronic probesets in each bin of width Δp = 0.05 ................................................................. 133

A.6 Metaprobosets along PRPF8. The coordinates of transcript cluster IDs along the PRPF8 gene. ................................................. 134

C.1 Contents of MaLTE-aux and the functions they provide .................. 154
To my parents,
Joseph and Mary Keter,
for their courage, selfless sacrifice
and determination in the face of
persistent uncertainty.
Chapter 1

Introduction

By the summer of 1977, the race to delineate the structure of eukaryotic genes was in full swing [1]. Eukaryotic genes, unlike their prokaryotic counterparts, presented a comparatively greater challenge to unravel. Attention then shifted to human adenoviruses for two principal reasons: they have smaller genomes (around 35 kbp) rendering experimentation easier [1] and they are processed by the host transcriptional machinery giving insight into how human genes, and by extension eukaryotic genes, were structured and processed. Shortly thereafter, Susan Berget and Claire Moore working under Phillip A. Sharp performed an experiment based on the adenoviral hexon mRNA that encodes the major structural polypeptide of the virus [2]. Figure 1.1 shows one electron micrograph from their paper of an RNA-DNA hybrid and its schematic.

Figure 1.1: Electron micrograph and schematic of adenoviral hexon mRNA hybridised to single-stranded genomic DNA selected using EcoR1 restriction enzyme (region A; see [2] for full details). Loops A, B and C are genomic DNA absent from mRNA and form loops between the hybridised regions. Image reproduced from [2].
Berget and colleagues made three key discoveries. First, they observed that the length of RNA-DNA hybridisation depended on the genomic segment used (selected using different restriction enzymes). They also observed that segments of single-stranded genomic DNA formed loops between hybridised portions suggesting a process by which the mRNA was assembled from discontinuous segments with long intragenic regions somehow removed. Finally, they found that multiple mRNA fragments could hybridise to an individual genomic segment as several previous experiments had reported [3]. Based on these findings, they proposed that the multiple segments were “probably spliced to the body of this mRNA during post-transcriptional processing”. A few weeks later, Richard J. Roberts’ lab replicated these findings [4]. Sharp and Roberts were awarded the Nobel Prize in Physiology or Medicine in 1993 “for their discovery of split genes” (http://www.nobelprize.org/nobel_prizes/medicine/laureates/1993/).

These and other experiments elucidated that eukaryotic genes were subject to an additional processing step that could explain their distinction from prokaryotic genes. In one of the most cited articles on splicing [5], Walter Gilbert introduced the term intron for the ‘silent’ intragenic regions of DNA found in conjunction with expressed regions (exons). He also made several startling predictions based on the nature of the splicing process whose extent and implications have only recently been fully appreciated. First, he proposed that the presence of introns could significantly contribute to evolutionary synthesis either by allowing inclusion or exclusion of significant portions of amino acid sequence due to mutations at splicing hotspots or by tolerating splicing error as a way to introduce novel transcripts. Second, that the dogma of ‘one-gene, one-peptide’ was untenable. Third, that the presence of large introns could accelerate recombination as well as enable exon shuffling, and finally, that introns could serve regulatory roles promoting cell differentiation events.

This thesis explores the estimation and analysis of gene expression and alternative splicing in the context of development and disease. It consists of three main analyses with an additional analytical chapter that expands on work from a preceding chapter. Chapter 2 provides a broad survey of introns, splicing and alternative splicing culminating in a brief discussion of transcription isoform expression measurement.

Chapters 3, 4 and 5 constitute the main work carried out over the period of study. In Chapter 3, we present an analysis of intron content (defined as the total intron length in a gene) in nineteen mammals and show that genes with conserved intron content are preferentially associated with development. This work was based on the observation that the intron content in some genes, such as the hairy and enhancer of split 7 (Hes7), plays a crucial role in the timing and dynamics of gene induction that is important for the correct formation of segmented body parts (somites) [6].
In Chapter 4, we carried out a transcriptome-wide analysis of aberrant splicing using high-density oligonucleotide arrays on whole blood samples from individuals affected by autosomal dominant retinitis pigmentosa. One of the mutations that leads to this disease is a carboxyl-terminal domain mutation in a core component of the spliceosome, pre-mRNA processing factor 8 (PRPF8). We hypothesised that non-retinal tissue may provide evidence of non-symptomatic transcriptome-wide splicing aberration due to the ubiquity of the spliceosome.

Chapter 5 introduces a novel approach to gene- and transcript-expression quantification that uses a machine learning algorithm to provide improved estimates of gene expression. MaLTE, which stands for Machine Learning of Transcript Expression, uses statistical learning between a gold-standard (such as high-throughput transcriptome sequencing or RNA-Seq) and microarray probe fluorescence intensities to substitute conventional summarisation algorithms. In Chapter 6, we extend the analysis of Chapter 4 using RNA-Seq data and compare the results to MaLTE and RMA. Finally, we make concluding remarks in Chapter 7.
Chapter 2

Introns, Splicing and Alternative Splicing

In the preceding chapter, we introduced the seminal discovery of introns, splicing and alternative splicing. This event had a remarkable impact on our understanding of one of the most important aspects of transcriptional processing, particularly in eukaryotes. This chapter now focuses attention on these three topics and outlines some of the salient discoveries relating to each. This account provides the context for the work in the subsequent chapters.

This chapter is organised into four main sections. The first section covers key aspects of introns, from their structural diversity to their functional roles. The second section is concerned with the splicing process and the various mechanisms involved that account for the relatively high fidelity of this process. In the third section, we focus on alternative splicing and conclude the chapter with a brief introduction to approaches to measuring gene expression with an emphasis on estimating transcript isoform expression.

2.1 Introns

An intron is an intragenic region of transcribed RNA that is routinely eliminated and excluded from the final functional transcript [5]. Excision of introns predominantly proceeds co-transcriptionally [7] through splicing of flanking sequences called expressed regions (exons). A discussion of splicing (and subsequently alternative splicing) requires a description of introns because their structure largely determines the nature of splicing. Additionally, an assessment of the various intron structures and the organisms in which they occur reveals an intriguing evolutionary history that had fueled debate ever since
Introns, Splicing and Alternative Splicing

Figure 2.1: Distribution of spliceosomal introns in eukaryotes. A wide selection of organisms are indicated together with the average number of introns per gene. Anopheles gambiae; Arabidopsis thaliana; Aspergillus nidulans; Bigelowiella natans; Nucleomorph; Caenorhabditis briggsae; Caenorhabditis elegans; Candida albicans; Chlamydomonas reinhardtii; Ciona intestinalis; Cryptococcus neoformans; Cryptosporidium parvum; Cyanidioschyzon merolae; Dictyostelium discoideum; Drosophila melanogaster; Encephalitozoon cuniculi; Giardia lamblia, Guillardia theta Nucleomorph; Homo sapiens; Leishmania major; Mus musculus; Neurospora crassa; Oryza sativa; Paramecium aurelia; Phanerochaete chrysosporium; Plasmodium falciparum; Plasmodium yoelii; Saccharomyces cerevisiae; Schizosaccharomyces pombe; Takifugu rubripes; Thalassiosira pseudonana; Trichomonas vaginalis. Image reproduced from [10].

their discovery. Therefore, this section outlines key aspects of introns, commencing with their designations and structural characterisations followed by a brief discussion of their origin and evolution. We then proceed to outline functions of introns before launching into a discussion on the splicing process.

2.1.1 Prevalence, Designations and Structural Characterisation

Introns are found in all the domains of life but it is in eukaryotes that they have undergone considerable evolution exhibiting the greatest complexity in vertebrates where they constitute a large proportion of gene sequence (Fig. 2.1). The abundance and variety of introns found also varies between organisms with unicellular organisms possessing primitive introns most of which appear to be undergoing purifying selection [8, 9] (Fig. 2.2).
A survey of the literature reveals five distinct categories of introns designated as group I to group III, archaeal introns and spliceosomal (or nuclear) introns. Each category has distinct characteristics and it is the shared characteristics that indicate how introns may have originated and subsequently proliferated into eukaryotic lineages.

Group I introns were first discovered in the protozoan ciliate *Tetrahymena thermophila* [11] (Fig. 2.3). They are found in several protozoan ribosomal genes, fungal mitochondrial genes, and some phage genes [12]. In eukaryotes, group I introns are found in ribosomal RNA genes in ribosomal DNA loci. They are exclusive to functional ribosomal RNA (rRNA) genes in eukaryotes but also occur in a wide variety of protozoans. Group I introns lack conserved nucleotide residues. A substantial treatment of group I introns can be found in the review by [13] and [14].

In eukaryotes, Group II introns are principally found in mitochondrial genes. They are also found in archaea and bacteria as well as in simple eukaryotes where they ensure that housekeeping genes are accurately transcribed by preventing invasion by other harmful mobile DNA elements [15]. Group II introns are mobile DNA elements: upon excision they may be reverse transcribed back into the genome in a process termed reverse splicing. They are structurally composed of six functional domains, D1 to D6 (Fig. 2.4). They are further classified into subgroups IIA, IIB and IIC. Subgroups IIA and IIB are typically around 900 bp long and occur in bacteria, archaea and mitochondria, while subgroup IIC are about half as long (∼400 bp) and are exclusively found in prokaryotes,
representing the most primitive lineage of group II introns. More information on group II introns is available from the following reviews [15–17].

Group III introns were first characterised in the *Euglena gracilis* chloroplast genome and are mostly found in mRNA genes of chloroplasts and other euglenoids [19]. They were identified in conjunction with a set of group II introns but differed in several traits. Group III introns are much shorter and more uniform in size, ranging in length from 95-110 bp having an average length of about 102 bp (less than half the length of group II introns). They have fewer conserved splicing signals compared to group II introns. In *Euglena*, the 5’ splice site is designated by the consensus sequence 5’-NTNG as opposed to the 5’-GTGYG- of group II introns. The 3’ splice site in group III introns is ANNTNNNN-3’, while group II have ATTTTAT-3’. Their secondary structure does not resemble that of group II introns consisting of a central core with six radiating, helical domains I-VI. In *Euglena*, group III introns are mainly found in genes associated
Introns, Splicing and Alternative Splicing

Figure 2.4: Structure of generic group II intron secondary structure. EBS: exon-binding site; IBS: intron-binding site; MGC: Manganese ion C; J2/3: junction between domains 2 and 3. Image reproduced from [18].

with transcription and the transcription machinery and are highly enriched for A and T nucleotides. [19]. A substantial treatment of group III introns is provided in [20].

Archaea possess introns in their tRNA and rRNA genes. Archaeal introns share structural similarities to other intron types. The introns are typically located one nucleotide 3’ of the anticodon but may also be found in the variable arm of the pre-tRNA. In pre-rRNAs, they are found associated with important functional sites that correspond with the surface of the ribosome. Several archaeal introns have been found to posses open reading frames of about 600 nt with relatively lower GC content. Archaeal pre-tRNA and pre-rRNA introns adopt a ‘bulge-helix-bulge’ conformation at the exon-intron junction (Fig. 2.5) which is then recognised and cut at the bulges by a splicing endoribonuclease. Archaeal introns are discussed at great length in a review by Lykke-Andersen et al. [21].

Spliceosomal introns are an evolutionary success story [22]. Their diversity and the correlation of their sizes with genome complexity, particularly in metazoans, point to the central roles they play in conferring organismal complexity [23]. They fall into two main categories depending on the splicing pathway they are subject to: U2-type and U12-type introns. U2-type introns are spliced by the major spliceosome and constitute more than 99% of all spliceosomal introns in human; U12-type introns are spliced by the minor spliceosome and make up the remainder. The U* designations stem from the names of the spliceosomal short nuclear RNAs (snRNAs), which are rich in uridine (U) and are core components of the spliceosome. The initial segregation into these two classes was based on conserved dinucleotide sequences that mark the 5’ and 3’ splice sites.
Introns, Splicing and Alternative Splicing

Figure 2.5: (a) Secondary structure of tRNA with the positions of known introns marked by numbers (in parentheses). The box at the bottom indicates the anticodon. (b) Conserved ‘bulge-helix-bulge’ in tRNA (left) and rRNA (right) with cleavage sites shown by arrows. Uppercase: > 85% conservation; lowercase: 60-85% conservation; unconserved indicated by dots; Purines - R; Pyrimidines - Y. Image reproduced from [21] with permission.

(5’ss and 3’ss): U2-type introns were associated with the terminal dinucleotides GT-AG or GT-AC while U12-type introns were found in those with AT-AC dinucleotides but there are numerous exceptions to this rule [24, 25]. In addition to the terminal dinucleotides, spliceosomal introns also have a branch point site (BPS) bearing the conserved sequence YTNAY (Y = pyrimidines C and T), in which the adenine base plays a critical role as an intermediate transition site during the splicing process (Fig. 2.6). The BPS is usually found between 30 and 50 bp upstream of the 3’ splice site [26]. Between the BPS and the 3’ splice site (3’SS) is a conserved sequence of pyrimidines called the polypyrimidine tract (PPT). The PPT is a crucial docking site for auxiliary factors (U2AF heterodimer) that facilitate recognition of both the BPS and the 3’ splice site.

Spliceosomal introns vary in length both within and across eukaryotic lineages: unicellular eukaryotic organisms such as yeast (S. cerevisiae and S. pombe have intron lengths of about 270 bp and 40-75 bp, respectively [28]) have few and short introns; intron length increases from simple metazoans (C. elegans: mean of ∼500 bp) and invertebrates (D. melanogaster: ∼560 bp), and shows remarkable development in length in vertebrates (M. musculus: ∼1300 bpt; H. sapiens: ∼1700 nt) (Fig. 2.7). The length of introns plays a crucial role in influencing how splicing takes place with short introns preferring cross-intron spliceosomal assembly (intron-defined) and long introns being subject to
a transition from cross-exon assembly to cross-intron assembly (exon-defined) splicing [29].

2.1.2 Origin and Evolution

How did introns originate? Attempts to answer this question led to two main hypotheses: the *intron-early* (later renamed *introns-first*) and the *intron-late* hypotheses [9], though hybrids are more widely accepted than the extreme versions. While debate has simmered since their discovery, it is only within the last few months that decisive evidence may
finally resolve this issue. Recent empirical results that examined the function of U-rich snRNAs in the major spliceosome suggest that the evidence may point in favour of the intron-late theory [31].

The intron-early theory was first proposed by Doolittle [32]. It attempted to reconcile what then appeared as, the disparate occurrence of introns in prokaryotes and eukaryotes. It postulated that the former once had introns but had since lost them due to genome streamlining to facilitate efficient transcript processing in response to competitiveness in growth. Both domains emerged from a common ancestor having introns but introns were retained only in eukaryotes. Furthermore, it proposed that eukaryotes were an independent evolutionary line because neither prokaryotes nor archaea were known to have introns at the time [33]. According to this theory, exons are modular elements that were shuffled to generate new genes from which it followed that protein domain boundaries should coincide with exon boundaries. The absence of evidence in support of the exon-to-protein-domain prediction cast doubt on this theory [34]. Eventually, prokaryotic and archaeal introns were discovered weakening support for this theory. The dominance of introns in eukaryotes, the exon shuffling hypothesis [34], and poor conservation of introns [35] have retained some adherents. However, this theory did not take into account the diversity of introns with strong evolutionary relationships (between group II and spliceosomal introns) that support the intron-late theory.

The intron-late theory was first proposed by Thomas R. Cech [36] and takes a radically different perspective by postulating that introns survived into the eukaryotic lineage and have since undergone various phases of evolution starting with substantial accumulation in the last eukaryotic common ancestor (LECA) to substantial intron loss in most current eukaryotic lineages [9]. It proposes that introns may have been introduced into eukaryotes as group II introns in the mitochondrial endosymbiont because of their homology to spliceosomal short nuclear RNAs (snRNAs) (Fig. 2.8). The innovation of the nucleus thereafter allowed the proliferation of introns to largely go unchallenged because transcription was separated from translation [34]. Additionally, the invention of the spliceosome as an efficient splicing machine permitted unconstrained intron lengths [9]. Recent evidence has conclusively determined that the U6 spliceosomal snRNA is a ribozyme providing a strong link to group II introns [37].

### 2.1.3 Functions of Introns

Introns harbour regulatory sequences that play roles in various transcript processing pathways. As we highlighted above, introns bear the conserved splicing signals necessary for the assembly of the spliceosome. These sequences may also regulate the efficiency of
Figure 2.8: Origin of nucleus-cytosol compartmentalisation at the onset of mitochondrial origin. Image reproduced from [34] with permission.
Introns, Splicing and Alternative Splicing

splicing by the strength to which spliceosome components bind to them. For example, several proteins such as poly(pyrimidine tract binding) (PTB) protein insulate the PPT from being accessed by U2AF. In fact, similar sequences in the downstream intron allow PTB to sequester the central exon ensuring that it is skipped [38, 39]. Introns may also host sequence elements that alter the speed of RNA polymerase II (RNA pol II) transcription. An example of this is the MAZ4 elements that slow down RNA pol II in order to ensure competitive splicing of an upstream exon, which would otherwise be skipped [40].

The rate of gene induction is influenced by the total transcribing gene length [41], which is dominated by the length of introns (intron content) [42]. In higher eukaryotes, introns contribute substantially to intron length resulting in potentially very long induction times. In particular, developmental processes have been shown to be sensitive to the impact of introns on the timing of gene induction. One example of this occurs during somitogenesis when the rate of gene induction dynamically determines the formation of somites. Takashima et al. demonstrated that exclusion of the introns of the Hes7 gene resulted in abolition of oscillation in the levels of Hes7 protein and led to severe segmental defects [6].

In eukaryotes, parental genomes undergo sexual recombination to create allelic combinations. The presence of introns reduces the probability of recombination events occurring on exonic sequences particularly when long (as in most eukaryotes). Exon shuffling is regarded as one of the principal means by which new exons can be incorporated as well as propagated between genes [43].

2.2 Splicing

Splicing is the process through which intronic sequence is systematically excised and flanking exons are ligated. It is one of several transcriptional processing steps. Splicing diversifies an organism’s transcriptome by availing linear sequence absent from the genome [44]. For splicing to take place introns must first be identified before a transition is made into the splicing conformation. The actual splicing reaction then occurs, releasing the excised intron as a lariat.

This section addresses several aspects of splicing paying greater attention to spliceosomal introns. It begins with an overview of the spliceosome then proceeds to the major spliceosome (associated with U2-type introns) because it splices out more than 99% of
Introns, Splicing and Alternative Splicing

eukaryotic introns. Next, we outline the splicing pathway including the transesterification reactions involved. We omit the minor splicing pathway because it largely resembles the major splicing pathway. Thereafter, we briefly sketch other splicing pathways: self-splicing, tRNA splicing and splicing of archaeal introns. Lastly, we address several important topics in splicing such as exon-defined splicing, ‘extreme’ splicing (ultra-short, ultra-long and micro-exons), and the timing of splicing relative to transcription.

2.2.1 The Spliceosome

The spliceosome is a dynamic, macromolecular complex that is systematically assembled at splice sites to catalyse the splicing reaction. It is composed of several uridine-rich short nuclear RNA (snRNAs) in conjunction with an elaborate proteome. The snRNAs strongly associate with some of the proteins forming short nuclear ribonucleoproteins (snRNPs, pronounced ‘snurps’). In addition to core snRNPs are a host of protein factors that temporarily aid in guiding snRNPs to their destinations.

Nucleoplasmic snRNAs can further be broken down into two classes: Sm\(^1\) or Lsm [46] (Fig. 2.9). Sm snRNAs are synthesised by a specialised form of RNA polymerase II, have, at the 5’ end, a trimethylated guanosine cap and form a stem loop at the 3’ end. Additionally, they have binding sites for the proteins they associate with. They translocate into the cytoplasm where the 3’ stem loop is formed then return to the nucleus. The major spliceosome is populated by Sm-snRNAs: U1, U2, and U4; the minor spliceosome has U11, U12 and U4atac. Both spliceosomes have U5 [46].

Lsm-snRNA are synthesised by RNA polymerase III. The ‘L’ stands for ‘like’ because they are homologous to the Sm proteins. They are structurally composed of a 5’ monomethylguanosine cap and a 3’ stem loop structure. Unlike the Sm-class snRNA, Lsm-snRNAs do not leave the nucleus. The only known members are U6 and its minor-spliceosome homologue U6atac. The U6 snRNA was recently shown to be a ribozyme forming the catalytic center of the spliceosome [37].

Each of the snRNAs is found to interact, at various points in the splicing pathway, with the precursor RNA substrate and multiple proteins [47]. U1 binds at the 5’ splice site (5’ss) and is associated with the proteins U1-70K, U1-A, U1-C and SRPK1. U2 is recruited by the U2 auxiliary factor (U2AF) protein and binds to the BPS. It is associated with the proteins SPF45, SPF31, CHERP as well as many others. U4 and U6 are found base-paired to each other and weakly bound to U5 as a tri-snRNP, which

\(^1\)Named in honour of Stephanie Smith, in whom they were first identified after her diagnosis with systemic lupus erythematosus [45].
associates with the proteins *PRPF8*, *BRR2*, *SNU114*, *PRPF28* and several others. The U4/U6 pair associate with *PRPF4*, *CYPH*, *PRPF31*, *SNU13* and *PRPF3*.

### 2.2.2 The Splicing Pathway

The spliceosome assembles in stages but in some organisms such as yeast it may also be found as a complete assembly ready to be activated [48]. Stepwise assembly involves the formation of complexes that in human have been characterised as *early* (*E*), *activated* (*A*), *B*, *Bact*, *B*\(^*\), and *catalytically* (*C*) active complexes [49] (Fig. 2.10). Each stage requires the presence of various factors that facilitate recruitment of the various components of the spliceosome. Also, some of these stages are ATP-driven with others being ATP-independent. It culminates in two transesterification reactions that ligate the 5’ and 3’ exons releasing the intron as a lariat. The splicing pathway is determined by the intron to be spliced. Here we concentrate on U2-type (and by extension, U12-type due to homology) spliceosomal introns only. Splicing of other introns is discussed in Section 2.2.4.

Some of the first detailed studies of splicing were carried out in yeast [48]. It was observed that the association of U1 to the 5’ splice site committed a site to splicing. This was therefore called the *commitment complex*. In mammalian splicing systems this is referred to as the *E complex* [50, 51]. The E complex forms due to substantial complementarity between the 5’ part of U1 and the 5’ splice site [52]. However, it has also been shown that the absence of U1 does not prevent complete splicing but rather
U1 enhances splicing efficiency. The interaction that U1 makes with the 3’ splice site also plays an active role in complex formation [53]. U1 is also associated with the 3’ splice site of the upstream intron and has been shown to facilitate binding of U2AF at the upstream 3’ splice site [54]. This observation led to the discovery of a mechanism by which exons are recognised called exon definition [29] discussed in more detail in Section 2.2.5.

The presence of greater sequence complexity at the 3’ end of eukaryotic introns (BPS, PPT and 3’ss) delayed a detailed understanding of the assembly of spliceosomal components. Indeed, many more factors are involved at the 3’ end [47]. In an ATP-dependent reaction and in the presence of a host of splicing factors (SF3a, SF3b and SF1), the U2AF heterodimer binds to the polypyrimidine tract (PPT) [50]. U2AF is made up of U2AF65 and U2AF35 and it is the latter that specifically binds to the 3’ splice site [55, 56]. This binding has been shown to be influenced by enhancers (discussed in Section 2.3.4.2) and may modulate the inclusion of the downstream exon [56]. U2 is then recruited to the BPS immediately upstream of the PPT forming the activated (A) complex.

U4 and U6 transiently exist as a duplex and are joined by U5 to form the U5-U4/U6 tri-snRNP. U5 is associated with PRPF8, which interacts with the 5′ splice site [49]. The recruitment of the tri-snRNP to the U1-U2 assembly constitutes the B complex. Several helicases are now involved in destabilising U1 and U4. The U5 helicases PRP28...
and BRR2 initiate U4 unwinding from U6. BRR2 is activated by SNU14 [57]. p68 then unwinds U1 from the 5’ss [58, 59] leading to the Bact complex. PRPF2 (an ATP-dependent DEAH-box protein) triggers further rearrangements that result in the $B^*$ complex.

The spliceosome then undergoes a series of further rearrangements facilitated by the following factors and helicases: BRR2, PRPF2, PRPF5, PRPF16, PRPF22, PRPF28, PRPF43, and SUB2 [60]. The spliceosome is now in the catalytically active conformation. PRPF2 briefly interacts with the spliceosome leading to further rearrangements to prepare for the first of two transesterification reactions. This reaction joins the 5’ss to the BPS. After the first step is complete, PRPF16 is recruited to enable structural rearrangements for the second transesterification reaction specifically associating with the 3’ss [61, 62]. In order for the second step to successfully complete, PRPF8, SLU7, PRPF17, and PRPF18 are needed [63–65]. Once the splicing of adjacent exons is complete, PRPF22 and PRPF43, two other helicases, enable the spliced mRNA and lariat intron to be released [66, 67].

2.2.3 Transesterification Reactions in Splicing

Spliceosome mediated splicing takes place through two transesterification reactions. Transesterification (TE) reactions are transfer reactions that occur between two specific chemical groups: esters and alcohols. A transfer reaction occurs whenever substrates exchange functional groups. Alcohols are identified by the hydroxyl (OH) functional group while esters are characterised by the generic formula $\text{RCO}_2\text{R}^\prime$ or $\text{RCOOR}^\prime$.2

The linear structure of the RNA substrate consists of the nucleic acid bases attached to a sugar-phosphate backbone. The first TE reaction takes place between the 2’-OH of the BPS adenosine and the phosphodiester bond at the 5’ splice site (Fig. 2.11). This reaction swaps the 2’ of the BPS adenosine with the 5’ss phosphodiester forming a closed intronic loop. The upstream exon is left free but scaffolded to PRPF8 [49]. The 3’-OH at the free exon now attacks the phosphodiester bond at the 3’ss in the second TE reaction swapping out the hydroxyl for the phosphate, completing the ligation. The resulting exon junction now conforms to a regular phosphodiester bond. Because there is no resulting change in the number of phosphates, TE reactions proceed without an external energy source such as ATP or GTP [69].

---

2R (and $\text{R}^\prime$) refers to any alkyl functional group. They are formed from alkanes, which are saturated hydrocarbons. Alkyl groups are formed from the removal of one hydrogen from an alkanes e.g. the methyl group, $\text{CH}_3$. From methane, $\text{CH}_3$. 

---
2.2.4 Other Splicing Pathways

2.2.4.1 tRNA splicing

The splicing of tRNA genes in archaea takes place in the presence of three enzymes: an endonuclease, a ligase and a phosphotransferase, and requires ATP for hydrolysis. The first step of tRNA splicing is the association of an endonuclease with the pre-tRNA, which cleaves both splice sites resulting in the two loose exons and the intron (Fig. 2.12). The intron is cleaved at the 5'-hydroxyl and 3'-cyclic phosphate ends. A ligase reaction ensues, catalysed by a 90 kDa tRNA ligase in a multistep process that terminates with both exons fused but still retaining an extra 2'-phosphate. This phosphate is then removed by the last enzyme, a nicotinamide adenine dinucleotide phosphotransferase. A more detailed analysis can be found in a review by Abelson et al. [70].
2.2.4.2 Self-splicing

Self-splicing affects group I-III introns. It involves intricate folding mechanisms that lead to similar transesterification reactions as described in Section 2.2.3.

Group I introns are cleaved in two transesterification reactions but these involve phosphorylated guanosine in one of its forms (GMP, GDP or GTP) (Fig. 2.3B). Alternatively, it may proceed with unphosphorylated guanosine. In the first transesterification reaction, the guanosine attacks the phosphodiester bond at 5’ss freeing the 5’ exon now terminated by a 3’-hydroxyl. The second transesterification reaction involves the free 5’ exon attacking the phosphodiester bond at the 3’ss ligating both exons. More details of group I self-splicing are provided in a review by Cech [71].

Self-splicing with group II introns resembles spliceosomal splicing and involves the formation of a lariat (Fig. 2.14). It proceeds in a strikingly similar fashion to spliceosome-mediated splicing with the main difference being the absence of an external catalyst. An $S_N2$ reaction is involved, which entails synchronous breaking and formation of the bond. The splicing reaction is facilitated by the folding of the group II intron into the active structure and this places the active site (a 2’-hydroxyl) in contact with the 5’ splice site. The first reaction can either take place due to hydrolysis or transesterification. Thereafter, the intron is rearranged so that the 3’-hydroxyl at the loose 5’ exon attacks the 3’ splice site, ligating the exons and releasing the intron as a lariat. The released lariat intron is still able to perform its catalytic activities to facilitate its re-entry into the genome or other RNA by reverse transcription or retrotransposition.
The scarcity of group III introns coupled with their low sequence conservation means that little is understood on how they mediate self-splicing [20]. They are spliced out in a lariat after undergoing two transesterification reactions and their self-splicing shares other similarities to group II self-splicing.

Nested splicing involves self-splicing of *twintrons*, group II introns that contain within them group III introns. They undergo the self-splicing pathway by first excising the group III intron followed by the group II intron [73].

### 2.2.4.3 Trans-splicing

*Trans*-splicing involves the ligation of exons from precursor-mRNA transcribed from different loci and has only been observed in eukaryotes (Fig. 2.15). Early characterisation introduced two primary categories of *trans*-splicing: those due to a ‘spliced leader’ (SL) found in protozoans in which a short RNA sequence (the ‘splice leader’) is added to the 5’ of the precursor mRNA, and the ‘discontinuous group II introns’ found in chloroplasts (in plants and algae) and plant mitochondria, which involve joining independently transcribed sequences [74]. SL RNAs are short (∼45-140 bp) non-mRNAs transcribed by RNA pol II from intronless genes and feature a hypermodified 5’ cap structure and a
splice donor site [75]. However, subsequent results have identified non-SL type trans-
splicing (also called genic trans-splicing) in which exons from different pre-mRNA tran-
scripts are ligated [76]. There is evidence for trans-splicing in Drosophila [77, 78] and 
human [79].

Trans-splicing may be spliceosome-mediated or self-spliced. In trypanosomes, the spliceo-
osomal snRNPs have homologues for U2, U4 and U6 and it is thought that SL RNA plays 
the role of U1 because it shares partial complementarity to the 5’ splice site similar to 
U1. Splicing occurs between the splice donor on the SL RNA and the acceptor site in 
the resident RNA. Unlike in conventional splicing, where the intron is excised as a lariat, 
trans-splicing releases the intron as a Y-shaped structure with the junction at the BPS: 
one arm of the Y is the 5’ end of the SL RNA bound to the BPS by a 2’-5’ bond from 
to the first transesterification reaction.

2.2.5 Exon-defined Splicing

The intron may be regarded as the principal locus for splicing because it harbours the 
majority of the conserved splice signals. Spliceosomal components assemble after the 
U1 snRNP base pairs with the 5’ splice site. U1 has also been found to be a necessary
Introns, Splicing and Alternative Splicing

Figure 2.16: Schematic representing exon definition including transition to intron definition via exon juxtaposition. Image reproduced from [29] with permission.

recruitment factor for U2AF, which in turn recruits U2 for the upstream intron of internal exons [81].

This observation prompted Robberson and colleagues to investigate the effect of exon length on its splicing to adjacent exons for internal exons. They found that exon skipping occurred when exon length was extended beyond a length of about 300 bp. This exon-length dependent pathway is called exon definition or exon recognition [53] (Fig. 2.16). In addition to U1 and U2 snRNPs, several other components of the spliceosome including the tri-snRNP were found in an exon defined complex [82]. Schneider and colleagues observed that this complex could easily be converted into a pre-catalytic B complex conformation giving an indication of the formation of the spliceosome. However, in order for splicing to proceed to completion, there needs to be a transition from exon definition to intron definition. This transformation is called exon juxtaposition (Fig. 2.16) and results in two intron defined complexes across the flanking introns [29]. Thereafter, splicing occurs as outlined in Section 2.2.2.

A comparison between human and fruitfly showed that the transition from intron-defined to exon-defined splicing may occur when introns are about 250 bp: intron-definition takes place below this length [83]. The average length of internal exons in humans is roughly 140 bp [84] and the average length of introns is approximately 1700 bp [83], which means that majority of splicing in humans takes place via the exon definition pathway [85]. The distribution of intron lengths clearly shows a bimodal characteristic with a trough at introns of about this length (Fig. 2.7) [27, 86]. The transition from a predominantly intron-defined to a predominantly exon-defined splicing genome can also be observed
Introns, Splicing and Alternative Splicing

when one looks at the exon-intron structure in nematode, fruitfly, zebrafish, mouse and human suggesting a preference for exons definition in higher organisms [23, 86]. As we will present in Section 2.3.2, the exon-intron architecture is important in alternative splicing.

Exon-defined splicing has two important implications. First, it implies an upper bound on exon length for efficient splicing [84]. This may vary from organism to organism and in human appears to range from between 300 and 500 bp. Splicing of exons that are longer than these lengths with long flanking introns therefore becomes problematic. Bolisetty and colleagues showed that these exons contain novel exon splicing enhancers (ESE) that help overcome this limitation [84]. Second, it implies that there must exist a minimum exon length. If exon lengths get too small then the components of the spliceosome will get in the way of the splicing process (steric hindrance) [29]. An intriguing study on minimum exon length showed that skipping occurred in human beta-globin gene when exon length was shortened below 50 bp [87] but was recovered when the splice sites were mutated towards the conserved splice site sequence (strong splice sites) [88]. Finally, exon definition implies a direct mechanism that the cell may use to modulate the inclusion of exons. By blocking communication across the exon, the exon can effectively be spliced out together with the flanking introns. Specifically, many heteronuclear RNP proteins (such as hnRNP A1, F, H, L [89, 90]) have been found to play a critical role in modulating the inclusion of an exon by interfering with this activity. We briefly cover such splicing modulation in the section on alternative splicing.

2.2.6 ‘Extreme’ Splicing

‘Extreme’ here refers to extremes of intron or exon lengths in relation to spliceosomal dimensions. There are three scenarios that may be defined as ‘extremes’: ultra-short introns, ultra-long introns and micro exons. The extremity of these features introduces new challenges for the splicing machinery. Nevertheless, efficient splicing can still occur.

2.2.6.1 Ultra-Short Introns

The existence of conserved sequences in introns implies that introns need to have a minimum length for splicing to take place. The distribution of intron length in human (Fig. 2.7) clearly shows a hard limit below which splicing is impossible by the pathways described so far. Introns shorter than 65 bp have been referred to as ultra-short introns [91]. This in turn has consequences on the splicing pathway undertaken. The fully assembled major spliceosome has a footprint of approximately $26 \times 20 \times 19.5$ nm and given that 1 bp in RNA is approximately 0.23 nm this translates to between 85 and 113
bp of linearised RNA [91, 92]. Three examples of ultra-short introns are the 56 bp intron in *HNRNPH1*, the 49 bp intron in *NDOR1*, and the 43 bp intron in *ESRP2* genes [91] all of which appear to be efficiently spliced (Fig. 2.17).

Splicing of ultra-short introns is *ad hoc* because these introns lack conserved regions such as the polypyrimidine tract (PPT). However, all introns analysed in this study required the U2-associated splicing factor SF3b. The longest of these introns, the 56 bp intron of *HNRNPH1* was excised with a lariat structure *in vitro* while two of the other ultra-short introns harboured previously unidentified conserved regions consisting of G-rich domains, which have previously been shown to be associated with intron splicing enhancer (ISE) domains [93, 94] and also play a role by enforcing the location of boundaries [95]. Furthermore, these G-rich domains, when mutated, resulted in a moderate splicing defect with the retention of these introns.

### 2.2.6.2 Ultra-Long Introns

At the extreme end of the scale are the ultra-long introns with the longest known human intron being 740,920 bp in the *heparan sulfate 6-O-sulfotransferase 3 (HS6ST3)* gene [96]. This gene appears to have a single intron and only one transcript isoform therefore exon definition is inadmissible. Fedorova and Fedorov also pointed out that over 3000 human introns are over 50 kb, ~1200 over 100 kb and ~300 over 200 kb, and 9 over 500
Introns, Splicing and Alternative Splicing

Introns, Splicing and Alternative Splicing

Figure 2.18: Three mechanisms of splicing in ultra-long introns. Image reproduced from [97] with permission.

kb. Apart from the obvious challenge of locating ~200 bp exons among such introns, it is intriguing that such introns are correctly excised. Even where exon definition might explain how tandem spliceosomes may form it still does not account for how exon juxtaposition may take place.

There are two known ways in which long introns may be spliced: recursive splicing as in the *Drosophila* gene *Ubx* and nested splicing as in the human gene *DMD*. In fruitfly, splicing of an ultra-long intron proceeds recursively. This occurs by stepwise reuse of one splice site (Fig. 2.18) ultimately leading to excision of the full intronic sequence [98, 99]. In order for this to take place, these relatively long introns have splice sites referred to as *ratchetting points* (RP)-sites (Fig. 2.18) leading to new intermediate splice sites called *regenerated splice* (RS) sites. The RP-site takes on a dual role as donor and acceptor and effectively simulates a zero-length exon [100]. The splicing can then be carried out either upstream-first or downstream-first. However, a computational survey of vertebrate introns showed that RP-sites are depleted suggesting an alternative mechanism for splicing of long introns [100]. Additionally, vertebrate long introns are enriched with short interspersed elements (SINEs) and long interspersed elements (LINEs) that may form secondary structures of introns, thereby reducing apparent intron length to facilitate the transition from exon definition to intron definition [100].

Recently, it was shown that nested splicing appears to take place in the human *dystrophin* (*DMD*) gene [97]. The seventh intron of DMD has a 110,199 bp intron. Using RT-PCR, Suzuki and colleagues were unable to observe part of the intronic sequence and instead only observed an alternative BPS between the 5’ss and the authentic BPS.
Furthermore, they also observed a closed lariat missing sequence from the original long intron suggesting that the intron was modified through a nested splicing mechanism.

2.2.6.3 Micro-Exons

Splicing of extremely short exons appears to be strongly influenced by the presence of enhancers located in either of the flanking introns or nearby exons. For example, the 61 bp exon of the chicken cardiac troponin T (cTNT) gene was found to require a 134 bp sequence in the downstream intron [94]. This intronic splicing enhancer (ISE) contains a G-rich motif that serves as the active site and may interact with components of the splicing machinery. The 7 bp exon in chicken skeletal troponin I (sTNI) depends on the presence of the upstream exon (an exonic splicing enhancer, ESE) [101]. Another example is the 18 bp N1 src exon spliced in an enhancer-dependent manner [102–104].

2.2.7 Timing of Splicing Relative to Transcription Termination

The timing and intra-nuclear localisation of splicing has important consequences for induction of functional transcripts [105]. Timing may be co-transcriptional, in which it is coupled to transcription and occurs along the nascent transcript or may be post-transcriptional. However, this requires a functional interaction between the transcriptional machinery and the splicing machinery.

2.2.7.1 Coupling between Transcription and Splicing

The carboxyl-terminal domain (CTD) of RNA polymerase II is necessary for in vivo splicing and cleavage as well as other aspects of post-transcriptional processing [106]. RNA pol II lacking a CTD led to accumulation of unspliced product as well as of read-around (RA) transcripts around a full plasmid sequence. The phosphorylation state of the CTD of RNA pol II plays a vital role splicing in vitro: hyperphosphorylated RNA pol II (RNA pol IIO) activates while hypophosphorylated RNA pol II (RNA pol IIA) may inhibit splicing. This suggests that specific sites along the CTD may play a decisive role in activating splicing both in vivo and in vitro [107, 108].

RNA pol II also takes part in splicing by influencing the early stages of spliceosome formation. The CTD activates splicing through its interaction with two splicing factors: U2AF65 and PRPF19 complex (PRPF19C, also called the nineteen complex, NTC). PRPF19C is poorly understood but it is seen to play an important role through protein-protein interactions with other spliceosomal components [109]. U2AF65 directly binds to
the phosphorylated CTD, which enhances further recruitment of U2AF65 and PRPF19C [110].

There is also evidence that the exon-defined splicing pathway exhibits coupling between transcription and splicing via exon definition [111]. Mammalian internal exons were spliced in vitro only when complete exon-defined complexes were formed in the presence of RNA pol II with a recombinant CTD. Furthermore, exon-defined complexes were stably associated with the recombinant CTD implying a direct interaction between RNA pol II in vitro and splicing.

2.2.7.2 Co-transcriptional Splicing versus Post-transcriptional Splicing

The majority of splicing events take place co-transcriptionally [7] but there are multiple examples of post-transcriptional splicing. However, the relative extent of co- versus post-transcriptionality of splicing as well as its predictability is unknown. Coupling of transcription to splicing avails an avenue through which alternative splicing may be effected either through kinetic coupling or structural coupling [112, 113]. Nevertheless, co-transcriptionality does not necessarily imply that the order of intron excision is strictly a 5’-to-3’ affair - a ‘first-come, first-served’ model; rather, it appears that the commitment to splicing (by formation of the U1-5’ss bond) may be decisive - a ‘first-committed, first-served’ model [114].

There are good reasons why splicing may need to take place post-transcriptionally. For example, in anucleate platelets, splicing of interleukin-1β (IL-1β) requires external activation. Prior to splicing, IL-1β pre-mRNA is present as unspliced transcripts [115]. This allows rapid induction of IL-1β for translation without the need for time-consuming transcription. Vargas et al. also demonstrated post-transcriptional splicing of sex lethal (Sxl) in Drosophila and polypyrimidine tract-binding (PTB) protein in HeLa cells by visualising splicing in live cells [113]. They observed that constitutive splicing (in male Sxl and canonical PTB, respectively) occurred co-transcriptionally but regulated splicing (in female Sxl and paralogous neuronal PTB, nPTB) resulted in intron-containing pre-mRNA dispersed throughout the nucleoplasm indicating post-transcriptional splicing. However, it has been argued that it may still be the case that splicing factors are co-transcriptionally recruited with splicing catalysis occurring post-transcriptionally [114].

2.3 Alternative Splicing

Splicing that results in ligation of all exons is termed constitutive and occurs predominantly in simple eukaryotes such as yeast [116]. However, in higher eukaryotes, splicing
Figure 2.19: α-synuclein (SNCA) gene showing 11 transcript isoforms of various lengths and with different exon assemblies. Image from Ensembl [117].

takes place with a remarkable diversity in splicing decisions, significantly increasing the information coding capacity of the organism’s genome. This form of splicing generates a set of alternative transcript isoforms from individual genes and may employ previously uncharacterised, cryptic splice sites, a process called alternative splicing. Figure 2.19 shows a gene with multiple transcript isoforms displayed Ensembl. Multiple transcript isoforms are indicated as parallel assemblies of exons. Figure 2.20 shows the set of splicing patterns that can be generated through alternative splicing.

Current estimates suggest that nearly all multi-exon genes in humans and a large proportion in vertebrates are alternatively spliced [118, 119]. Widespread application of high-throughput assays that interrogate genes along their entire length such as exon arrays [120], exon junction arrays [121], tiling arrays [122] and high-throughput sequencing [118] provide unprecedented transcriptome-wide quantification, allowing investigators to identify aberrant and alternative splicing events. Notwithstanding these developments, transcriptome analysis is still in its infancy [123, 124].

This section covers various themes in alternative splicing. In particular, we address five broad topics: prevalence, origin and evolution, functions, modulation mechanisms, and stochasticity. To keep this chapter brief, we only highlight salient points in each section commencing with one of the most cited examples of alternative splicing.

The Down’s syndrome cell adhesion molecule (DSCAM) homologue in Drosophila exemplifies the astounding potential of alternative splicing with the capacity to generate a possible 38,016 individual transcript isoforms. Figure 2.21 shows a schematic of this
gene. Dscam has 24 exons, three of which (exons 4, 6, 9 and 17) occur in clusters of 12, 48, 33 and 2 (for a total of 95) mutually exclusive exons. This gene is expressed in neurons where it is crucial for cell recognition during axonal modeling [125]. Alternative splicing allows a choice of exons from the clusters while the remaining exons are constitutively spliced. A similar example of a high diversity gene in human is the neurexin 3 (NRXN3) gene having 1,728 transcript isoforms [126].

### 2.3.1 Prevalence

Alternative splicing has only been observed in eukaryotes even though introns occur in other domains of life. Alternative splicing in unicellular eukaryotes has been extensively studied in yeast (S. cerevisiae and S. pombe) [128]. Both species differ markedly in their transcriptomes with S. pombe showing closer resemblance to mammalian transcriptomes. Alternative splicing in yeast takes place mainly through intron retention because almost all intron-containing genes have only one intron.

Metazoans display alternative splicing complexity that correlates with genome size [128]. Similarly, the relationship between genome size and number of genes saturates at about
Introns, Splicing and Alternative Splicing

Figure 2.21: An example of extreme alternative splicing in the Drosophila Dscam gene that potentially generates 38,016 transcript isoforms through mutual-exclusive splicing of four exon clusters (coloured red, blue, green, yellow). Image reproduced from [125] with permission.

20,000 genes suggesting that increased transcriptome complexity is achieved through increased alternative splicing [129]. Vertebrates exhibit a greater diversity in alternative splicing compared to invertebrates [130]. One study examined the rate and characterisation of alternative splicing across vertebrates and invertebrates and found alternatively spliced exons exhibit longer flanking introns but only for cassette exons (as opposed to alternative 5’ss (A5SS) and alternative 3’ss (A3SS) exons) [131]. Furthermore, they found that intron retention appears to be the rarest of all observed alternative splicing patterns.

2.3.2 Origin and Evolution

To explain the origin and evolution of alternatively spliced exons, it is instructive to outline their properties. Alternatively spliced exons tend to be short, have flanking splice sites with weak conservation of splice motifs and have higher enrichment of regulatory sequences compared to constitutively spliced exons [127]. Their sequences show higher conservation and their lengths are preferentially divisible by three (suggesting that they may function as cassettes) [128]. These properties of alternatively spliced exons help shed light on how they may have emerged.

There are currently three theories that attempt to explain the emergence of alternative splicing: de novo exonisation, exon shuffling and exon transition [127]. De novo exonisation proposes that previously non-exonic sequence may have acquired exonic status when cryptic splice sites converted to weak splice sites. The weak splice sites resulted in the exon being included in a fraction of transcript isoforms. One example of such a scenario is provided by Alu elements, which are primate-specific repetitive and transposable sequence elements (Fig. 2.22). About 5% of alternatively spliced exons are known
Introns, Splicing and Alternative Splicing

**Figure 2.22:** Exonisation of Alu elements. (a) Mutations surrounding Alu elements that could give rise to new exons. (b) Transformation of RNA secondary structures by enzymatic editing of adenosine (A) to inosine (I), which is recognised as guanosine (G). In the presence of two adjacent Alu elements of opposite orientation can then create a functional 3’ss (AA to AG). Image reproduced from [127] with permission.

to have originated from Alu elements [132, 133]. Alu elements are also found in introns and explain a possible route by which these exons may have emerged. Additionally, de novo alternative exons may have emerged through tandem exon duplication [128, 134]. The duplicated exons may later be subject to different selective pressures leading to exon specialisation [128].

Exon transition theory proposes that alternatively spliced exons may have evolved from extant exons in two possible ways: relaxation of splicing motifs or evolution of novel regulatory proteins. Relaxation of splicing signals (splice sites, BPS or PPT) may have resulted from splice site mutations influencing exon inclusion. For example, 5’ss differ between yeast and metazoans with the former having fewer base pairing to U1, while the latter have base pairing that extends into the exon [28] and are alternatively spliced. Furthermore, one experiment showed stronger binding of U1 to constitutive versus alternative exons [133]. The second method may involve evolution of regulatory elements
that may associate either with exonic, or intronic sequence to negatively influence splicing. Over time these regulatory elements (mostly protein) may then render constitutive exons alternative [127] (Fig. 2.23).

Exon shuffling relies on how eukaryotic exons typically constitute a miniscule proportion of genomic sequence, which elevates the likelihood for novel recombination events to occur predominantly at intronic loci. This introduces opportunities to shuffle exons thereby introducing phenotypic diversity. The resulting shuffled exons would have been propagated as alternatively spliced exons to their new resident genes (Fig. 2.24) and appears to have been predominant in higher organisms due to their abundance of intronic sequence. Exon shuffling may therefore have appeared much later in evolution and played an important role in distributing alternatively spliced exons as opposed to creating them de novo [127].

2.3.3 Functions of Alternative Splicing

As the preceding section has highlighted, alternative splicing has made a substantial contribution to the evolution of genome complexity particularly in metazoans leading up to higher eukaryotes. In particularly, alternative splicing introduced a space-saving means by which genomic information may be regulated [129]. However, this innovation was introduced at the considerable cost of transcribing non-coding introns as well as requiring the development of the sophisticated spliceosome. The ability to explore transcript space at low evolutionary risk was a useful invention for introducing new functions [5].

Nearly 75% of all alternative splicing events occur within the coding sequence, leading to a direct impact on protein sequence [135, 136]. These can lead to changes in binding properties, intracellular localisation, enzymatic and signalling activities, protein stability, addition of domains that lead to protein conformational changes and changes in ion channel properties [137].

Alternative splicing also facilitates developmental programmes and cell differentiation programmes, which results in tissue specificity. The facility for a single genome to encode the full range of transcripts required in the course of development and cell differentiation is largely attributable to alternative splicing programmes [138]. For example, FOXP1 has an alternative splicing event that leads to pluripotency of embryonic stem cells [139] while brain is a well-studied example of tissue specificity [140]. This feature of metazoans also reflects the difference in transcriptome complexity between unicellular eukaryotes, which exhibit primitive alternative splicing (predominated by intron
Introns, Splicing and Alternative Splicing

Figure 2.23: Exon transition from constitutive to alternative. (A) Degenerate splice signals that can form (Aa) 5'SS of 3'SS, (Ab) weak 5' splice sites leading to skipping, (Ac) introduction of an exonic splicing suppressor, or (B) secondary RNA structure that sequester splice signals. Image reproduced from [127] with permission.
Introns, Splicing and Alternative Splicing

Figure 2.24: Exon shuffling. Image reproduced from [127] with permission.

retention) and metazoans (show increasing exon skipping dominance with genome complexity) [141].

Other roles of alternative splicing have been extensively discussed in [137] and [142] and include gene expression regulation by introduction of premature stop codons (PTC) leading to nonsense mediated decay (NMD) [143], modification of mRNA properties such as stability [144] as well as the role alternative splicing plays when dysregulated in diseases such as cancer [145–147].

2.3.4 Modulation of Alternative Splicing

There are four major ways through which alternative splicing is modulated: genetic (at the sequence level), proteomic, epigenetic (chromatin), and kinetic (coupling to transcription). In many instances, modulation employs several means simultaneously resulting in a splicing code [148], which can be used to predict tissue-specific splicing patterns based on the set of available modulators.

2.3.4.1 Genetic

There are two main ways in which the nucleotide sequence can serve to modulate splicing: through sequence motifs and through sequence variants [149]. Since the nucleotide content of genes is typically unchanged for the duration of an organism’s lifetime, the main genetic regulatory elements are those that lead to differential splicing without harmful phenotypes. Genetic regulators typically work in conjunction with proteomic, epigenetic and kinetic regulators for dynamic modulation of splicing.

Sequence motifs come in two forms: (i) those that influence some dynamic of transcription either through interaction with the spliceosome or an associated factor [150, 151]; (ii) those that influence binding of external elements that have a subsequent effect on the activity of the spliceosome [89, 152]. The main candidates for (i) are the conserved
sequence elements themselves. In fact, it is by selectively mutating splice sites that a lot has been learned about the various steps and factors involved in splicing and some of the sequence variants (e.g. single nucleotide polymorphisms, tandem repeats) may lead to deviation from the consensus. In these cases, regulation is unintended and often leads to unwanted consequences (defective transcripts). Also, the SINEs and LINEs mentioned in Section 2.2.6.2 in conjunction with long introns create RNA secondary structure that may influence accessibility or efficiency in splicing. Sequence motifs in (ii) are associated with proteomic regulators and are discussed below.

Of particular interest are sequence variants that are linked to changes in abundance of individual transcript isoforms. These are referred to as splicing quantitative trait loci (sQTLs). They may be found proximal to the gene (cis-sQTLs) or may operate in trans (trans-sQTLs).

### 2.3.4.2 Proteomic

Proteomic modulators work in conjunction with other components of the splicing system to positively or negatively influence splicing. Their action may directly influence the splicing complex or may involve interaction with the pre-mRNA. Their association with the transcript partitions them into those that interact with exonic or intronic sequence. Consequently, many proteomic splicing modulators have extensive RNA binding domains [89].

The most widely characterised exonic splicing enhancers (ESEs) are from the class of serine-arginine (SR) proteins (Fig. 2.25). ESEs preferentially bind to purine-rich sequences generally found in protein-coding regions bearing an A-rich consensus sequence...
ESEs have been reported to be present in almost all exons independent of whether they are constitutively or alternatively spliced [153]. However, individual enhancer proteins typically may have sharply divergent preferential binding motifs that depart from this pattern [154]. SR proteins consist of an RNA binding domain (RBD) and a repeated sequence of arginine/serine (RS) dipeptides, which appear to be interchangeable between SR ESEs [155, 156]. Some examples of proteins in this family are ASF, SC35, SR 20, SR 30c, SR 40, SR 55 and SR 70 [157]. Long exons (>300 bp) have been shown to be enriched for ESEs, likely to aid in the exon definition process [84].

In addition to exonic splicing enhancers are intronic splicing enhancers (ISE), exonic splicing silencers/suppressors (ESS), and intronic splicing silencers/suppressors (ISS). An example of an intronic splicing silencer is the polypyrimidine tract binding (PTB) protein which binds to the PPT and at the 5’ end of the downstream introns [158]. It base pairs with the transcript then folds to sequester the exon away from the spliceosome. Exonic splicing silencers are enriched in exons that are flanked by long introns likely to be spliced by exon definition. They operate by preventing U2AF and U1 from interacting across the exon leading to exon skipping.

2.3.4.3 Epigenetic

There is a lot of interest on the link between chromatin modifications and alternative splicing [114, 159]. Epigenetic changes involving histone modifications can have a major influence on splicing decisions (Fig. 2.26). This is largely due to the fact that nucleosomes are preferentially located at exons and particularly alternatively spliced exons [29]. Additionally, in relation to RNA pol II kinetics, the speed of transcription at nucleosomes is lower than between nucleosomes favouring higher inclusion because spliceosomal components have more time to aggregate at splice sites [160, 161]. It has also been demonstrated that some histone acetyltransferases such as Gcn5 in yeast and STAGA in human physically interact with spliceosomal snRNPs implying that the chromatin complex may directly influence accurate assembly of the pre-spliceosome [162, 163].

Recent work has shown a link between epigenetic modifications and splicing regulation. In particular, methylation at CTCF sites has been linked to alternative splicing of CD45. In this case, methylation at CTCF sites modifies the rate of RNA pol II (see Section 2.3.4.4), influencing the amount of time available to recognise the 5’ splice site of intron 5 [90].
2.3.4.4 Kinetic

Coupling of transcription and splicing provides kinetic modulation: faster transcription means that weak splice sites may be ignored leading to exon skipping; slower transcription implies longer time for the spliceosome to form and catalyse splicing (Fig. 2.27). In this way, kinetic modulation introduces competition between splice sites conditional on the speed of transcription. One study on co-transcriptionality of splicing in yeast found that RNA pol II appears to slow down in the last intron to increase the likelihood of its excision [164]. Similarly, pausing elements such as MAZ4 immediately downstream of an alternatively spliced exon ensure its inclusion [40].

One recent example involved elucidation of a novel complex found in proximity with RNA pol II. The DBIRD complex [165] was found to speed up transcription at A+T rich exons, increasing the likelihood of them being skipped. A+T rich regions are associated with RNA pol II pausing [166]. When present, DBIRD slowed down RNA pol II leading to higher inclusion of those exons.

2.3.4.5 Combinatorial Control: The Splicing Code

The mechanisms outlined above typically act in concert implying that they combine in various ways to exert complex control patterns. Analysis of how various factors contribute to alternative splicing events has given rise to the notion of a splicing code [167]. Barash et al. [148] studied the combinatorial effect of multiple splicing regulators
Figure 2.27: Kinetic modulation of alternative splicing. (a) SR splicing factor 3 (SRSF3) is controlled by the phosphorylation state of the RNA pol II C-terminal domain (CTD), which then determines inclusion of alternate exons recruitment is affected. (b) The Mediator, which associates with transcription factors, recruits the splicing suppressor hnRNPL. (c) CCCTC-binding factor (CCTF) binds to unmethylated CG-rich DNA sequences and slows RNA pol II transcription allowing competitive splicing of a downstream exon. (d) DBIRD facilitates exclusion of A/T-rich exons by preventing RNA pol II slowing down at these exons. (e) Kinetic coupling following DNA damage, which leads to hyperphosphorylation of the RNA pol II CTD further inhibiting RNA pol II elongation. (f) SWI/SNF (a chromatin remodelling factor) recruits SAM68, which stalls pol II enhancing exon inclusion. Image reproduced from [114] with permission.

to come up with the code table shown in Figure 2.28. Their assessment centred on exon inclusion as the main alternative splicing event. However, development of a complete splicing code will require considerable effort given the tissue- and development-specific natures of alternative splicing.

2.3.5 Stochasticity of Alternative Splicing

As we mentioned in Section 2.3.3, alternative splicing may allow the cell to explore a range of transcripts prior to committing to produce them exclusively or at abundant levels [5]. This evolutionary trial-and-error would require that splicing possesses a degree of tolerable noise in order to generate novel transcript isoforms. Therefore, splicing is an inherently stochastic process.
Introns, Splicing and Alternative Splicing

Figure 2.28: Splicing code developed for five mouse tissues. Tissue codes: CNS (C), muscle (M), embryo (E), digestive (D), and tissue-independent mixture (I). The code is read using the key on the top left of the image. For example, abundant CU-rich (nPTB) motifs (fifth row) are strongly determinative of high inclusion of the alternate exon if within 300 nt upstream. Image reproduced from [148] with permission.
Studies on the stochasticity of the splicing process have provided evidence that even under correct functioning, the spliceosome is tolerable to some degree of noise. For example, Melamud et al. used over 8,000 EST libraries to estimate noise levels using a stochastic model based on the number of introns and gene expression [168]. They estimated that between 1% and 10% of all transcripts are spliced erroneously. Interestingly, they found that the number of introns correlates with number of alternative splicing events (per transcript) but while admitting fewer splicing errors. They also observed a higher density of ESEs with an increase in the number of splicing reactions suggesting selective pressure towards less splicing noise. They argue that these observations imply that low errors are expected in the presence of many introns to ensure that useful transcripts are produced (otherwise, nothing would be produced at all). Additionally, lower errors could also ensure lower toxicity due to accumulation of aberrant transcripts at high expression.

Pickrell and colleagues used RNA-Seq data to assess splicing error rates and characterised over 150,000 previously unannotated splice sites conforming to consensus splice sites (GT-AG and GC-AG) but lacking evolutionary conservation [169] (Fig. 2.29). They estimated that the spliceosome admits about 0.7% error per human intron. They found that longer introns tend to have a higher rate of mis-splicing and that highly expressed genes have lower errors perhaps due to their shorter introns spliced via intron-defined pathway unlike those with higher noise that bore resemblance to exons spliced via exon-defined pathway and were enriched for ESEs.

In a recent analysis of yeast transcriptome using transcript isoform sequencing (TIF-Seq), Pelechano and colleagues demonstrated remarkable transcript diversity even in an organism with very little alternative splicing [170]. TIF-Seq involves high-throughput sequencing of captured transcript ends revealing the diversity in transcription start and end sites. It will be intriguing to see what results TIF-Seq (or comparable methods) produce from human transcriptomes.

### 2.4 Measurement of Transcript Expression

#### 2.4.1 Genes

The steady state abundance of products of gene transcription is an important measure that can be used to infer the endogenous state or response of a cell under various conditions. Identifying differentially expressed genes is a powerful approach to determine their functions. This has elevated gene expression measurement to a major biotechnological and bioinformatic challenge leading to remarkable innovation in developing high
resolution and accurate expression assays. Expression assays may be divided into *quantitative* and *non-quantitative* methods reflecting advances in the underlying technology. Each class of methods may be further split into *targeted* [171] or *global* [172] techniques, whereby targeted techniques probe a handful to thousands of genes while global approaches attempt to probe all genes in the genome of an organism. However, for a gene that undergoes alternative splicing, gene expression estimates may be uninformative particularly if the overall expression remains unchanged even when the transcription profile changes [173]. This is shifting attention towards measurement of transcript isoforms.
Figure 2.30: A schematic showing various tools available for detecting and measuring splicing events using RNA-Seq. Image reproduced from [175].

### 2.4.2 Transcript Isoforms

As the preceding sections have developed, alternative splicing plays an important role particularly in metazoans, giving rise to the need to identify and estimate the abundance of all transcript isoforms (Fig. 2.30). The shift towards a transcript-isoform-centric paradigm has been slow, possibly hampered by the sharp drop in the cost of microarrays for which well-developed analytical approaches abound; on the bright side, there has been a correspondingly sharp drop in the cost of more appropriate approaches like high-throughput sequencing ([http://www.genome.gov/sequencingcosts/](http://www.genome.gov/sequencingcosts/)) as the technology evolves towards greater accuracy and reproducibility [174].

Whereas measurement of gene expression is often based on well-defined gene models, transcriptome complexity routinely integrates transcript isoform discovery in addition
to estimating the abundance of each transcript isoform. The stochastic nature of alternative splicing makes identifying and measuring transcript isoforms especially challenging. Therefore, a middle-ground has been devised that entails detection of alternative splicing events [176].

This section outlines the three main aspects of transcript isoform measurement: detection of alternative splicing events, identification of transcript isoforms, and measurement of transcript isoform abundance. This discussion will interchangeably refer to microarray and high-throughput sequencing methods.

### 2.4.2.1 Detection of Alternative Splicing Events

Detection of alternative splicing events usually involves assessing part of a gene associated with the transcript isoform of interest. Typically, this will be the exon though in organisms that show a preponderance of intron-retention (such as plants) introns may be used [141]. For example, consider a simple case of a gene with a cassette exon and two transcript isoforms. The presence or absence of one or more transcript isoforms will be indicated by the relative expression of this exon. This can be assessed by a test between experimental groups for the normalised expression of that exon. The advent of high-density oligonucleotide arrays has been instrumental in availing affordable, genome-wide access to such analyses. Another method to infer splicing events is by using junction tiling arrays [177], which check the presence of predefined splice junctions.

Some widely used measures used in alternative splicing detection (some of which are used in later analytical chapters) are:

1. **Differential inclusion.** The most straightforward approach involves a direct comparison of normalised exon expression. Exon inclusion is measured as the exon signal normalised by the gene signal. Thereafter, a simple test can be used to compare exon inclusion levels between two samples (e.g. using a t-test or more advanced statistical test). Because the number of exons and hence tests is typically large, one must control the false discovery rate. Additionally, one may employ strict filtering to consider only exons that have strong evidence of expression. For Affymetrix arrays, this is done using the detection above background (DABG) metric. We apply this method to study aberrant splicing in Chapter 4.

2. **Splicing index.** The splicing index employs the idea of gene expression fold change to compare the inclusion between two groups. It is computed as,

\[
SI = \log_2 \frac{p_1}{q_1} \div \log_2 \frac{p_2}{q_2},
\]
where \( p_i \) is the probe set signal and \( g_i \) is the metaprobe set signal in groups \( i = 1, 2 \). Typically, absolute values greater than one suggest differential inclusion of the targeted exon. The splicing index is normally used in conjunction with statistical tests such as conventional \( t \)-tests or modified \( t \)-tests [178]. One limitation of the splicing index is that it only applies to two-sample data. An extension of the splicing index is microarray detection of alternative splicing (MIDAS), which performs an ANOVA on the splicing index.

3. **Percent/Proportion Spliced In.** One powerful advantage of high-throughput sequencing is that it is able to give a snapshot of actual splicing events by using splice junctions. When reads are mapped back to the organism’s genome, the presence of gaps that coincide with introns captures the set of splicing events that occurred. This is useful in also identifying novel splice variants that are absent from available gene models. The simplest method to discern splicing events is by the percent/proportion spliced in (PSI) given by

\[
\Psi = \frac{\text{no. of reads that include an exon}}{\text{no. of reads that include OR exclude the exon}}
\]

The PSI is a single-sample metric (as opposed to differential inclusion or splicing index) because it allows an assessment of alternative splicing within a single sample.

### 2.4.2.2 Identification of Transcript Isoforms

Complete mapping of a complex organism’s transcriptome is a perpetual task because of the diversity of tissues brought about by elaborate developmental programmes. Therefore, it is necessary to incorporate a search for novel transcript isoforms into the analysis. Microarrays have limited applicability to this task making room for sequencing-based techniques [179].

There are two ways in which a transcriptome may be assembled to unravel the set of isoforms. In a **guided** approach, transcriptome assembly leverages a reference annotation to infer novel transcripts. For example, **Cufflinks** uses a graph theoretic approach by which it attempts to construct the most parsimonious assembly that the read fragments support [180, 181].

Another method involves **ab initio** transcriptome assembly, where evidence for transcript isoforms is based on structural motifs found in aligned reads (e.g. transcription start sites (TSS), exon-intron boundaries etc). **Oases** accomplishes this by building an array of hash lengths (where each hash is a \( k \)-mer) followed by a filtering step in which noisy data
Introns, Splicing and Alternative Splicing

Introns, lacking structural motifs, are iteratively eliminated. It also considers various alternative splicing events (e.g. cassette exons) produced from multiple primary assemblies [182].

2.4.2.3 Measurement of Transcript Isoform Abundance

Transcript quantification is currently an active area of research and new tools continue to be developed as sequencing technology improves and sequencing costs drop. A recent survey examined the performance of 14 tools and concluded that transcriptome assembly is still a challenging task [174]. Just as in transcriptome assembly, abundance disambiguation may proceed either with a reference annotation or may be ab initio. Several tools have also been developed that can infer transcript isoform abundance from Affymetrix exon arrays, including MMBGX (based on multi-mapping probes and Bayesian hierarchical models) [183], PUMA 3.0 [184], and SPACE [185]. However, these have been superseded by sequence based approaches.

The majority of transcript disambiguation methods employ an expectation maximisation method on some generative model. One of the first such models was proposed by Jiang and Wong [186] upon which iReckon [187], IsoEM [188], IsoLasso [189], RSEM [190], among many others have built. Others such as mGene [191] use discriminative models based on a hidden semi-Markov SVM [192], while others like SLIDE [193] rely on sparse matrix linear models that consider transcript structure (exon boundaries) with the possibility of incorporating alternative transcriptomic data (e.g. EST) to augment the transcript discovery step prior to quantification.

By far, the most popular tool is Cufflinks, which may be guided or ab initio. In both approaches it uses a likelihood model for the alignment of fragments (determined using a gapped aligner such as TopHat [194]) to assign abundances. It then uses numerical optimisation to obtain the maximum likelihood transcript abundances.

In Chapter 5, we introduce a novel approach to gene and transcript expression estimation that uses both high-throughput sequencing data and microarray data in conjunction with a statistical learning algorithm. Using a high-quality, public data resource from the Genotype-Tissue Expression (GTEx) project, our method learns the relationship between probe intensities and gene expression as estimated by a gold-standard method (e.g. RNA-Seq) then predicts gene expression from probe intensities alone.
Chapter 3

Evidence for intron length conservation in a set of mammalian genes associated with embryonic development

The content of this chapter was published as:

PKK carried out analyses and co-wrote the manuscript.

Abstract

Background: We carried out an analysis of intron length conservation across a diverse group of nineteen mammalian species. Motivated by recent research suggesting a role for time delays associated with intron transcription in gene expression oscillations required for early embryonic patterning, we searched for examples of genes that showed the most extreme conservation of total intron content in mammals.

Results: Gene sets annotated as being involved in pattern specification in the early embryo or containing the homeobox DNA-binding domain were significantly enriched among genes with highly conserved intron content. We used ancestral sequences reconstructed with probabilistic models that account for insertion and deletion mutations to
distinguish insertion and deletion events on lineages leading to human and mouse from their last common ancestor. Using a randomization procedure, we show that genes containing the homeobox domain show less change in intron content than expected, given the number of insertion and deletion events within their introns.

Conclusions: Our results suggest selection for gene expression precision or the existence of additional development-associated genes for which transcriptional delay is functionally significant.

3.1 Introduction

One of the salient features of eukaryotic genomes is the pervasive presence of introns, comprising up to 95% of transcribed primary protein-coding sequences in mammals [195]. The functions, origins and evolutionary trajectory of introns have long been of great interest in genomics and genome evolution. Although some theories of the spread of introns postulate that they can accumulate passively as a consequence of insufficient purifying selection to remove them in organisms with relatively low effective population sizes [196, 197], introns have been shown to play a number of functional roles [198]. They give rise to the possibility of alternative splicing, contributing to the diversity of biomolecules, and they can also have a substantial impact on levels of gene expression [199–201]. Introns contain most of the sequence features required for splicing (5’ and 3’ splice sites, branch point sites (BPS), poly-pyrimidine tracts and various intronic splicing elements). Many introns also contain functional non-coding RNAs [198], which can play critical roles in fine-tuning gene expression [202]. Lastly, introns have been proposed to control the timing of gene expression by delaying transcription [203].

Negative feedback loops with a time delay can result in oscillating patterns of gene expression, which can be exploited by living organisms as biological time-keeping devices. This appears to be particularly important in development [203]. The hairy and enhancer of split 7 (Hes7) gene is involved in the control of somite formation through oscillatory patterns of gene expression [204]. Recently, Takashima et al. [205] investigated in vivo the impact of removing the introns from the mouse Hes7 gene. They found that expression of the mutant Hes7 gene occurred approximately 19 minutes earlier than for the wild-type and that this reduction in gene expression delay resulted in abolition of oscillations and segmentation defects.

Given that the delay associated with transcribing introns appears to play a crucial role in the functioning of this gene, the length of the introns may be evolving under
purifying selective pressure. Moreover, further examples of genes with highly conserved intron content across species may reveal more genes for which transcriptional delay is an important aspect of gene regulation. We investigated the conservation of the intron content of *Hes7* across a diverse set of nineteen mammals and carried out a search for other genes for which total intron length shows evidence of evolutionary conservation. In such cases introns are likely to play an important functional role and, in a subset, time delays associated with transcribing introns may be a significant aspect of gene regulation.

### 3.2 Data and Methods

Complete sets of gene models were downloaded from Ensembl release 62 [206] via BioMart for 19 mammalian species. These were *Homo sapiens*, *Bos taurus*, *Canis familiaris*, *Ornithorhynchus anatinus*, *Equus caballus*, *Erinaceus europaeus*, *Gorilla gorilla*, *Loxodonta africana*, *Monodelphis domestica*, *Myotis lucifugus*, *Microcebus murinus*, *Mus musculus*, *Oryctolagus cuniculus*, *Sus scrofa*, *Spermophilus tridecemlineatus*, *Tarsius syrichta*, *Tursiops truncatus*, *Vicugna pacos* and *Felis catus*. Species were selected to sample a broad range of mammalian evolutionary history. A phylogenetic tree of these taxa, obtained from the interactive tree of life [207, 208] is provided for illustrative purposes (Fig. 3.1).
For each Ensembl gene in each species we considered all annotated exons and calculated the total intronic content of the gene as the sum of the gaps between non-overlapping successive exons in the canonical transcript associated with the gene. Genes were divided into four classes, depending on the total intron content of the gene (1–5 kb, 5–15 kb, 20–50 kb and 50–100 kb). Orthologous groups of mammalian proteins were downloaded from OrthoDB [209]. We extracted the protein identifiers from OrthoDB corresponding to each of the mammalian species included in the study. Where more than one protein from a species was included in the same orthologous group, we selected one paralogue at random. For each orthologous group, with a representative in at least half of the mammalian species considered we determined the number of species in which the intron content, as defined above, was within 10% of the length of the intron content of the human gene. This was done separately for genes in different intron content classes. Functional analysis of genes with conserved intron content was carried out using DAVID [210, 211]. For each size class, the genes for which the intron content showed evidence of conservation was used as the foreground set and the complete set of genes in the size class was used as the background set.

Genomic multiple sequence ancestor alignments, inferred using the Ensembl Enredo-Pecan-Ortheus pipeline [212], were downloaded from the comparative genomics section of the Ensembl FTP site. Insertions and deletions were inferred for ancestral sequences included in these alignments using a branch transducer method that has been shown to achieve high accuracy [213]. This allowed insertions and deletions to be placed on branches of the mammalian phylogeny. For reconstructed insertion and deletion events we focussed on the branches leading from the common ancestor of the euarchontoglires (which includes rodents and primates) to humans and mouse. All insertion and deletion events occurring within introns (at least 20 bp from exon-intron) boundaries were identified, based on Ensembl gene models. In this case, intronic indels were defined as indels that occur within the boundaries of the gene but not within 20 bp of any annotated exon associated with the gene.

### 3.2.1 Randomization study of intron length evolution along the human lineage

For each intron size class we used the complete set of insertion and deletion events within the introns to define the null expectation of events in that intron size class. To test for evidence of purifying selection acting on the intron content of a gene or a set of genes we considered all of the insertion and deletion events in the gene(s) under consideration and for each event sampled an insertion or deletion from the total set of events in the introns of genes in that class. Given a set of insertions and deletions randomly sampled
in this way we calculated the change in intron content implied by this set of insertions and deletions (sum of insertion lengths minus the sum of the lengths of the deletions), separately along the human and mouse lineages. This was repeated 10,000 times and in each case the implied change in intron content in the randomized data was compared to the change in intron content, given the actual insertions and deletions inferred to have occurred. The number of times the absolute value of the change in intron content in the randomized data was less than or equal to the absolute value of the change in intron content in the observed data was used to calculate a p-value, with a separate p-value calculated for the human and mouse lineages. These p-values indicate the probability of observing as little or less variation in the intron content of a gene or set of genes, given the number of insertion and deletion events that have occurred and under the assumption that these indel events have the same distribution as events in other genes in the same intron size class.

3.3 Results and Discussion

Delayed expression of Hes7 resulting, at least in part, from intron transcription has been proposed to play a key role in somite formation in animal embryos by establishing an oscillating pattern of gene expression [205, 214–216]. Given the important role played by the length of the introns rather than their specific sequence content in this gene, we compared the combined length of Hes7 introns across a diverse set of mammalian species (Fig. 3.1) to determine the extent to which the intron content of this gene is conserved over evolutionary time. For each orthologue, the sum of the distances between successive exons in the canonical transcript (according to Ensembl) was calculated. Despite the fact that there may be differences in gene annotation across species the intron content of most of the orthologues was similar. Nine out of the twelve of the available orthologues of Hes7 among the mammalian species in our dataset differed in combined intron length from the human gene by less than 10% (Table 3.1).
### Table 3.1: Conservation of Hes7 intron length

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Gene</th>
<th>#introns</th>
<th>CIL(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>Human</td>
<td>ENSG00000179111</td>
<td>3</td>
<td>1839</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>Mouse</td>
<td>ENSMUSG0000023781</td>
<td>3</td>
<td>1846</td>
</tr>
<tr>
<td><em>Oryctolagus cuniculus</em></td>
<td>Rabbit</td>
<td>ENSOCUG000002606</td>
<td>3</td>
<td>1696</td>
</tr>
<tr>
<td><em>Canis familiaris</em></td>
<td>Dog</td>
<td>ENSCAFG0000016957</td>
<td>3</td>
<td>1823</td>
</tr>
<tr>
<td><em>Felis catus</em></td>
<td>Cat</td>
<td>ENSFCAG0000015190</td>
<td>3</td>
<td>1958</td>
</tr>
<tr>
<td><em>Equus caballus</em></td>
<td>Horse</td>
<td>ENSECAG0000013278</td>
<td>3</td>
<td>1804</td>
</tr>
<tr>
<td><em>Myotis lucifugus</em></td>
<td>Little brown bat</td>
<td>ENSMLUG000008014</td>
<td>3</td>
<td>1810</td>
</tr>
<tr>
<td><em>Erinaceus europaeus</em></td>
<td>Hedgehog</td>
<td>ENSEEUG000006336</td>
<td>3</td>
<td>2774</td>
</tr>
<tr>
<td><em>Sus scrofa</em></td>
<td>Pig</td>
<td>ENSSSCG0000017982</td>
<td>4</td>
<td>1550</td>
</tr>
<tr>
<td><em>Tursiops truncatus</em></td>
<td>Bottlenose dolphin</td>
<td>ENSTTRG0000010312</td>
<td>3</td>
<td>1827</td>
</tr>
<tr>
<td><em>Loxodonta africana</em></td>
<td>African elephant</td>
<td>ENSLAFG0000002331</td>
<td>3</td>
<td>1870</td>
</tr>
<tr>
<td><em>Monodelphis domestica</em></td>
<td>Grey short-tailed opossum</td>
<td>ENSMODG0000007704</td>
<td>3</td>
<td>2364</td>
</tr>
<tr>
<td><em>Ornithorhynchus anatinus</em></td>
<td>Platypus</td>
<td>ENSOANG0000022456</td>
<td>2</td>
<td>2018</td>
</tr>
</tbody>
</table>

\(^1\)Cumulative intron length

Intron content of orthologues of human *Hes7* (OrthoDB ID EOG45TDC4). The tarsier orthologue was omitted. This gene had no annotated introns but the annotation appeared to be incomplete.
To determine whether intron content, and thus potentially transcription time, of *Hes7* was exceptionally conserved compared to other genes and to discover additional examples of genes for which there is evidence of intron content conservation we compared intron content (defined as the sum of the lengths of all introns in the canonical transcript) between human gene models from Ensembl and sets of orthologous genes, obtained from OrthoDB [209]. Because the pattern of insertion and deletion may differ between very large and smaller introns we restricted to genes with introns of similar lengths to *Hes7* (specifically we considered genes for which the sum of the intron lengths in the canonical transcript was between one and five kilobases). Restricting also to orthologous groups represented in at least half of the species, only 11 other orthologous groups (from a total of 1875 satisfying these criteria) had intron content within 10% of the human gene in as high a proportion of the orthologues as the *Hes7* group. Remarkably, of the corresponding 12 human genes (including *Hes7*), five are annotated with the gene ontology (GO) term GO:0007389 (pattern specification process) from the biological process component of GO, defined as any developmental process that results in the creation of defined areas or spaces within an organism to which cells respond and eventually are instructed to differentiate. Using the DAVID functional annotation tool [210, 211], this is the most statistically enriched GO term ($p = 9.4 \times 10^{-5}$) and remains significant following correction for multiplicity of testing using the Benjamini-Hochberg method (FDR = 0.03). Six of the genes were annotated with the Swissprot and Protein Information Resource (SP-PIR) keywords term developmental protein (FDR = 0.003).

We also used a more relaxed conservation threshold and identified a larger set of 144 genes with relatively conserved intron content among the genes with between one and five kilobases of intron. Genes for which more than 50% of the orthologues have intron content within 10% of the intron content of the human member of the group were included in this group, again restricting to orthologue groups represented in at least half of the species. We again found evidence for a set of genes with conserved intron content which was very highly enriched for genes involved in development, identified using DAVID (Table 3.2). Only the top 20 enriched terms are shown in Table 2. These include terms corresponding to the presence of homeobox DNA-binding domains, frequently involved in developmental gene regulation. To look for examples of conserved intron content among genes with higher intron content, we separated the human genes into five intron content size classes (1–5 kb, 5–20 kb, 20–50 kb and 50–100 kb), to account, to some extent, for the fact that patterns of intron length evolution may differ between genes with larger versus smaller introns. The proportion of genes from the other size classes (other than size class one, discussed above) that showed evidence of conserved intron content was smaller and we found no evidence for enrichment for the default DAVID gene sets (after multiplicity correction).
3.3.1 Intronic insertions and deletions along lineages leading to human and mouse

To investigate evolution of intron length through insertion and deletion in greater detail we considered changes in intron length along the human and mouse lineages following divergence from their last common ancestor. Genomic sequences at ancestral nodes of the mammalian phylogeny were obtained from Ensembl. These sequences were inferred using a probabilistic method that has been shown to have high accuracy for the inference of insertion and deletion events. We mapped insertion and deletion events to introns. Conservation of intron content could result from the absence of insertion and deletion events or from balance between insertion and deletion. For the 11 genes that showed as much conservation as Hes7 across the mammalian panel, we found qualitative support for both effects. The total amount of insertion and deletion was lower in these genes but there were also some genes with substantial insertion and deletions in balance (Fig. 3.2). This was particularly evident in the case of indels along the lineage leading to mouse,

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Count</th>
<th>(P)-value</th>
<th>FDR (BH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP_SEQ_FEATURE</td>
<td>DNA-binding region:Homeobox</td>
<td>20</td>
<td>(2 \times 10^{-10})</td>
<td>(8 \times 10^{-8})</td>
</tr>
<tr>
<td>INTERPRO</td>
<td>Homeobox, conserved site</td>
<td>20</td>
<td>(9 \times 10^{-10})</td>
<td>(2 \times 10^{-7})</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>Homeobox</td>
<td>20</td>
<td>(2 \times 10^{-9})</td>
<td>(4 \times 10^{-7})</td>
</tr>
<tr>
<td>INTERPRO</td>
<td>Homeobox</td>
<td>19</td>
<td>(5 \times 10^{-9})</td>
<td>(6 \times 10^{-7})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Regulation of transcription, DNA-dependent</td>
<td>41</td>
<td>(7 \times 10^{-9})</td>
<td>(1 \times 10^{-6})</td>
</tr>
<tr>
<td>INTERPRO</td>
<td>Homeodomain-related</td>
<td>19</td>
<td>(1 \times 10^{-8})</td>
<td>(8 \times 10^{-7})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Regulation of RNA metabolic process</td>
<td>41</td>
<td>(1 \times 10^{-8})</td>
<td>(8 \times 10^{-6})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of transcription from RNA polymerase II promoter</td>
<td>19</td>
<td>(1 \times 10^{-8})</td>
<td>(6 \times 10^{-7})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of biosynthetic process</td>
<td>27</td>
<td>(3 \times 10^{-8})</td>
<td>(1 \times 10^{-6})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of cellular biosynthetic process</td>
<td>27</td>
<td>(3 \times 10^{-8})</td>
<td>(1 \times 10^{-6})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of transcription, DNA-dependent</td>
<td>21</td>
<td>(8 \times 10^{-8})</td>
<td>(2 \times 10^{-7})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of RNA metabolic process</td>
<td>21</td>
<td>(8 \times 10^{-8})</td>
<td>(2 \times 10^{-7})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of macromolecule biosynthetic process</td>
<td>25</td>
<td>(9 \times 10^{-8})</td>
<td>(2 \times 10^{-7})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of macromolecule metabolic process</td>
<td>27</td>
<td>(1 \times 10^{-7})</td>
<td>(2 \times 10^{-5})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of nitrogen compound metabolic process</td>
<td>25</td>
<td>(1 \times 10^{-7})</td>
<td>(2 \times 10^{-5})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Sequence-specific DNA binding</td>
<td>25</td>
<td>(1 \times 10^{-7})</td>
<td>(2 \times 10^{-5})</td>
</tr>
<tr>
<td>SMART</td>
<td>HOX</td>
<td>19</td>
<td>(2 \times 10^{-7})</td>
<td>(8 \times 10^{-6})</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>Positive regulation of transcription</td>
<td>23</td>
<td>(2 \times 10^{-7})</td>
<td>(3 \times 10^{-6})</td>
</tr>
</tbody>
</table>

1Benjamini-Hochberg FDR corrected \(P\)-values

**Table 3.2:** Gene set enrichment analysis. Gene set enrichment analysis of genes in size class 1 showing evidence of intron content conservation in mammals. Top twenty most significantly enriched terms are shown.
Intron Length Conservation

Figure 3.2: Comparison of insertions and deletions. The sum of the insertions and deletions that have taken place in genes in size class one along the human (a) and mouse (b) lineages since divergence from their common ancestor. Blue points, corresponding to the genes that showed as much or more intron content conservation than *Hes7* are shown against a background of grey points, corresponding to all genes in the size class. All axes are truncated, but the truncated axes include all of the blue points. Grey points lying with cumulative insertions or deletions greater than 200 bp for human or 2000 bp for mouse are not shown.

However, as these genes were selected on the basis of their conservation in a panel that included humans and mouse this result should be considered to be illustrative only.

We also identified the complete set of genes in size class one that were annotated with a homeobox domain by Interpro [217]. To investigate whether the change in intron content since the common ancestor of human and mouse was less than expected, given the inferred numbers of indels we used a randomization procedure (described in Data and Methods). We applied this procedure to the set of homeobox domain proteins and found that the change in intron content of these genes along both the human and mouse lineages was far less than expected, given the number of indels inferred to have occurred in each lineage and under the null assumption that the ratio of insertions and deletions and size distribution of these events did not differ for these genes compared to the rest of genes in the size class. In the case of human, the mean absolute change in intron content was 450 bp per gene, whereas the median of this value in the randomized data was 846.
Intron Length Conservation

bp per gene. The randomized values were less than or equal to the observed value in 87 of 10,000 randomizations (one-tailed \( p = 0.009 \)). In mouse the mean absolute change in intron length was just 9 bp per gene. The median of this quantity across randomizations was 398 bp per gene. The values in the random data were lower than or equal to the observed data in 33 of 10,000 randomizations (one-tailed \( p = 0.003 \)).

3.3.2 Contribution of conserved sequence elements to intron length conservation

From our results it appears that there is selection to prevent large changes in intron content in specific sets of genes, notably in genes involved in development and especially in early embryonic patterning. This selection could result from the presence of cis or trans (e.g. non-coding RNAs) regulatory elements within introns of these genes, which would be disrupted by insertions and even more so by deletions. However, the presence of regulatory elements would tend to increase the intron length [218], yet we found good evidence for classes of genes with well conserved intron content only among the genes with lower intron content. To test the contribution of conserved functional elements within intronic sequences on the conservation of intron content we obtained conservation scores, calculated using phastCons, from the UCSC genome database. phastCons scores for the intronic regions of \( Hes7 \) are shown as Figure 3.3. While there is evidence of functional elements these are relatively few and short and seem unlikely to prevent changes in intron length over evolutionary time. More quantitatively, we compared phastCons scores, averaged across all intronic positions, between the 144 genes with evidence of conserved intron content and the remainder of the genes in the smallest intron size class. The difference in mean phastCons scores was not significant (\( p = 0.33 \), Wilcoxon rank sum test). The difference in the proportion of conserved sites (with phastCons scores > 0.95) was also non-significant (\( p = 0.69 \)), suggesting that a greater proportion of conserved intronic functional elements is not the cause of the intron length conservation in these genes. However, we cannot rule out that conserved sequence elements contribute to the conservation of the intron content.

To explain our observations, conservation acting on functional elements would, perhaps, have to be balanced with selection for rapid induction, as rapidly induced genes have been shown to have short introns [219]. If the conservation of intron content is a consequence of balance between conservation of intronic functional elements and selection either for efficiency [220] or for rapid induction, this balance appears to be particularly significant for development-associated genes. For these genes, expression precision at the critical developmental junctures in which patterns are established in the developing embryo may be particularly crucial. An intriguing role for introns that has recently gained
experimental support is in the establishment of oscillating patterns of gene expression through the control of time delays in expression [205]. These authors attributed a delay of approximately 19 minutes to the presence of introns. However, the introns of Hes7 are relatively short (< 2 kb) and RNA polymerase II processes nucleotides at a rate in the order of 2 kb per minute [221, 222]. This may suggest either the existence of other functional elements within the intron that caused the delay in transcription or specific properties of the intron that result in an exceptionally slow transcription rate.

### 3.4 Conclusion

Some theories of intron evolution have focussed on the energy cost associated with introns. Reduced intron lengths in highly expressed genes were proposed to result from selection for efficiency [220]. In support of this model, greater selective efficiency associated with larger effective population sizes of unicellular versus multicellular organisms, is associated with shorter introns [196, 197]. Alternatively longer introns in highly and ubiquitously expressed genes may be associated with greater regulatory complexity and the preservation of regulatory elements in introns [218]. Selection to reduce energetic
costs of transcription and for rapid induction of some genes as well as selection to preserve regulatory elements in highly regulated genes may all be important factors for intron evolution. However, comparison of genes with very conserved intron contents carried out here, suggests that there may be a substantial number of genes for which the size of the primary transcript is an important and conserved feature. These genes are enriched for key developmental processes, particularly the establishment of early embryonic patterns. Since time delays in gene induction have been shown to be important for some such genes, we propose that selection to conserve transcription time is an important factor in the evolution of intron lengths, particularly in development-associated genes. Our analysis involved a combination of a heuristic examination of intron content across a panel of mammalian species and a statistical randomization approach, which does not take into account all of the factors that may affect the fixation probabilities of insertions and deletions in introns. A better understanding of the evolution of intron sizes through insertion and deletion requires the development of evolutionary models of intron length evolution, though this is challenging because of the diversity of insertion and deletion events and the difficulty in modelling the constraints imposed by functional elements within the introns on the occurrence and size distribution of these events. If an appropriate model of neutral evolution by insertion and deletion can be derived, such a model could be used to identify and quantify purifying selection acting on intron lengths and, perhaps, to discover examples of positive selection acting on changes in intron length over a phylogeny or specific branches of a phylogenetic tree.
Chapter 4

A mutation in a splicing factor that causes retinitis pigmentosa has a transcriptome-wide effect on mRNA splicing

The content of this chapter has been submitted as:

Paul K Korir, Lisa Roberts, Raj Ramesar and Cathal Seoighe “A mutation in a splicing factor that causes retinitis pigmentosa has a transcriptome-wide effect on mRNA splicing.” under review in BMC Research Notes.

PKK carried out the bioinformatic analyses and drafted the manuscript.

Abstract

Background: Substantial progress has been made in the identification of sequence elements that control mRNA splicing and the genetic variants in these elements that alter mRNA splicing (referred to as splicing quantitative trait loci – sQTLs). Genetic variants that affect mRNA splicing in trans are harder to identify because their effects can be more subtle and diffuse, and the variants are not co-located with their targets. We carried out a transcriptome-wide analysis of the effects of a mutation in a ubiquitous splicing factor that causes retinitis pigmentosa (RP) on mRNA splicing, using exon microarrays.

Results: Exon microarray data was generated from whole blood samples obtained from four individuals with a mutation in the splicing factor PRPF8 and four sibling controls. Although the mutation has no known phenotype in blood, there was evidence of widespread differences in splicing between cases and controls (affecting between 10% and 25% of exons). Most probesets with significantly different inclusion
Transcriptome-wide Aberrant Splicing

(defined as the expression intensity of the exon divided by the expression of the corresponding transcript) between cases and controls had higher inclusion in cases and corresponded to exons that were shorter than average, AT rich, located towards the 5’ end of the gene and flanked by long introns. Introns flanking affected probesets were particularly depleted for the shortest category of introns, associated with splicing via intron definition.

Conclusions: Our results show that a mutation in a splicing factor, with a phenotype that is restricted to retinal tissue, acts as a trans-sQTL cluster in whole blood samples. Characteristics of the affected exons suggest that they are spliced co-transcriptionally and via exon definition.

4.1 Introduction

Splicing sits at the intersection between transcription and translation and directly regulates both the abundance and diversity of transcripts. It is effected by the spliceosome, a macromolecular complex composed of uridine-rich snRNAs (U1, U2, U5, and U4/U6 duplex for the U2-type spliceosome) and numerous proteins [47], which systemically assemble at conserved sequence motifs on newly-transcribed RNA. The spliceosome undergoes a series of structural rearrangements to catalyse two transesterification reactions ligating adjacent exons [49]. The PRPF8 protein mediates most spliceosomal interactions as it interfaces with splice sites on the transcribed RNA, snRNAs, and other proteins. Over the years, mapping of PRPF8 protein domains has revealed numerous splicing-related roles. For example, the 3’ fidelity region attenuates the impact of 3’ splice site (3SS) mutations [223]. Also, the C-terminal domain (CTD) of Prp8p (yeast homologue of PRPF8) interacts with Snu114p and Brr2p (another protein that interacts with the CTD) to unwind and release the U4 snRNA, which activates the spliceosome [224–228]. Recent evidence suggests a direct role in splicing through catalysis of the transesterification reactions [229]. It is therefore unsurprising that mutations in PRPF8 have been shown to affect splicing efficiency in yeast [230], mouse [231, 232], zebrafish [233] and human [234].

Genome-wide association studies (GWAS) have revealed a large number of genetic variants that are associated with diseases and phenotypes but for many of these associations the exact causal mechanism remains unknown. A large proportion of the associations are likely to be mediated by the effects of polymorphic variants on gene expression. In support of this view, GWAS results have been found to be enriched for expression quantitative trait loci (eQTLs) [235]. Genetic variants may also affect transcript processing. The contribution of variants that affect mRNA splicing, in particular, is thought to be
substantial [236, 237]. Although many studies have assessed the impact of genetic variants acting in cis on mRNA splicing [236, 238–242] variants that act in trans have been less extensively studied [243].

Genetic variants affecting components of the spliceosome can have trans-acting effects on splicing. In particular, maturation defects, structural malformations and splicing factor mutations can lead to aberrant transcripts due to mis-splicing, resulting in genetic diseases [147, 244]. For example microencephalic osteodysplastic primordial dwarfism type I (MOPDI), also known as Taybi-Linder Syndrome (TALS), is caused by mutations in the minor spliceosome resulting in incomplete splicing of a small number of critical transcripts [245, 246]. MOPDI is characterised by gross developmental retardation and a lifespan of under one year. Mutations affecting components of the spliceosome can also have a more restricted phenotype, such as in the case of splicing factor associated retinitis pigmentosa (RP). RP is a broad spectrum of eye diseases featuring gradual degeneration of rod and cone photoreceptors, causally associated with mutations in over 100 genes and may be autosomal dominant, autosomal recessive or X-linked [247, 248]. A subset of autosomal dominant mutations are located on components of the spliceosome PRPF3, PRPF6, PRPF8, and PRPF31 [247, 249]. The progression of the disease is marked by night blindness and tunnel vision, and in later stages, complete blindness. However, no adverse phenotypes in non-retinal tissues are known [243, 247].

There are two hypotheses that could explain how mutations in splicing factors result in RP [234]. First, because the pathology of the disease is restricted to the retina, RP may be the result of aberrant splicing of a transcript isoform that is specific to retinal tissue. Recently, transcriptome analysis of retinal tissue in a mouse model of RP revealed a large number of novel and/or aberrant transcripts [232]. This hypothesis does not rule out the second possibility: that these mutations in core components of the spliceosome increase the transcriptome-wide rate of splicing error and that retinal tissue is particularly sensitive to this increased rate of mis-splicing. Intermediates between these two extremes are also possible.

Here we test the hypothesis that a mutation in a splicing factor that causes RP leads to transcriptome-wide splicing defects in a non-retinal tissue. We used exon microarrays to profile transcript expression in whole blood samples obtained from RP-PRPF8 individuals bearing the p.H2309R mutation in the splicing factor PRPF8 [248] and paired sibling controls. For a large proportion of exons we found evidence of differential inclusion of the exon in mature transcripts of cases compared to controls. Differentially spliced exons were disproportionately associated with longer flanking introns and likely to be spliced through exon, rather than intron definition.
4.2 Methods

4.2.1 Exon array sample preparation and data generation

This project was approved by the University of Cape Town Research Ethics Committee (REC REF: 180/2009), which is in compliance with the guidelines of the declaration of Helsinki. Written informed consent for participation in the study was obtained from participants.

Blood from five individuals carrying the autosomal dominant RP mutation p.H2309R (referred to as cases) together with that from unaffected sibling controls was collected and preparation coordinated to ensure equal sample incubation times. Three blood samples per subject were collected into PAXgene™ tubes (PreAnalytiX), according to the manufacturer’s instructions, and incubated at room temperature. RNA extractions were performed 16 hours (A sample), 20 hours (B sample) and 39 hours (C sample) after collection. Total RNA was extracted using the PAXgene™ Blood RNA kit, following the manufacturer’s recommended protocol. The RNA samples were not heat denatured following elution, but immediately stored at -80°C.

The A and B RNA isolations of each sample were pooled, and the Affymetrix GeneChip® Blood RNA Concentration Kit used to concentrate the RNA. Quality control checks to determine RNA concentration and integrity were performed for each sample, using the Nanodrop (Thermo Scientific) and Agilent 2100 Bioanalyser (Agilent Technologies), respectively. It was determined that one of the cases showed poor integrity of A/B and C RNA samples leading to its exclusion together with its sibling-pair. RNA yields of between 2.22 and 5.16µg were obtained for the remaining samples, which were of satisfactory integrity to proceed with microarray analysis. The 260/280 ratios for the samples ranged from 1.97 to 2.06. All samples exceeded the RNA integrity number (RIN) quality threshold ≥ 7, as samples ranged from RIN 8.3–9.10.

Ribosomal RNA reduction was performed on the eight remaining samples using the RiboMinus™ Transcriptome Isolation Kit (Human/Mouse) (Invitrogen), in accordance with a modified Affymetrix protocol. The RNA was then processed with the Whole Transcript (WT) Sense Target Labelling assay. The labelled samples were hybridised to Affymetrix GeneChip® Human Exon Array 1.0 ST chips according to the prescribed protocol. A total of 5.5µg of single stranded cDNA (generated from cRNA) was hybridised to the arrays. Following hybridization, the arrays were scanned in the Affymetrix GeneChip® Scanner 3000 7G.

Cases were labelled $T_1$, $T_2$, $T_3$, and $T_4$ and corresponding sibling controls $C_1$, $C_2$, $C_3$, and $C_4$. 
4.2.2 Expression quantification of microarrays

The Affymetrix GeneChip Human Exon Array consists of oligonucleotide probes clustered into sets of probes (probesets), which are further clustered into metaprobesets. Raw exon array intensities were summarised into probesets and metaprobesets at the core and full probeset/metaprobeset level using Affymetrix Power Tools (APT) by the robust multi-chip averaging (RMA) algorithm [250]. Quality control of array data was carried out according to a previously described procedure [120]. We applied hierarchical clustering and principal components analysis to the raw intensities to identify possible outliers. APT was also used to determine probesets and metaprobesets that were detectable above the background (DABG) [251]. We excluded from further analysis all probesets that were detectable above background in fewer than half of the samples. We also used the annotation files provided by Affymetrix to exclude all probesets with probes likely to cross-hybridise. A similar procedure was applied to metaprobesets: only those having at least half of the associated probesets expressed above background in all samples were retained [120]. Finally, we normalised each probeset intensity by subtracting its value from its parent metaprobeset intensity since these values lie on a logarithmic scale [252]. The number of probesets remaining after these filtering steps were 103,268 and 149,835 for the core and full sets, respectively.

All analyses were performed on the hg19 build of the human genome. Exon boundaries and splice sites were defined based on build 66 of the Ensembl database of gene models [253]. Exonic and intronic probesets were isolated using the R tool xmapcore, using the database based on build 66 of the Ensembl gene model [117]. U12-type introns were obtained from the U12DB [25] with coordinates modified to hg19 using liftOver [254].

For each probeset, log-transformed expression intensities of probesets were compared between cases and controls. The log-transformed expression intensities were normalized by subtracting the log expression intensities of corresponding metaprobesets from that of probesets. We compared the expression intensity for each probeset using paired tests first using Welch $t$-tests from the standard R library then using moderated paired $t$-tests implemented in the limma [255] package. Correction for multiple testing was based on the $q$-value method as implemented in the R package qvalue [256]. The qvalue package was also used to estimate $\pi_0$, the expected proportion of tests consistent with the null hypothesis.
4.2.3 Characterisation of differentially spliced exons

To avoid the potential for bias resulting from exons to which multiple probesets mapped or individual probesets that mapped to multiple overlapping exons, we constructed a one-to-one mapping of probesets to exons by randomly sampling at most one exon for each probeset and one probeset for each exon. We compared the characteristics of exons for which the corresponding probeset was differentially included between cases and controls using a significance threshold of 0.01. Exons with significantly higher and lower inclusion levels in cases relative to controls were considered separately and, in each category, eight features of the exon (5’ and 3’ splice site scores, length of up and downstream introns, exon length, nucleotide content, distance from the 3’ end and associated gene length) were compared between the significantly included/excluded exons and the remainder of the exons tested. For each feature, we compared groups using a non-parametric test (Wilcoxon rank-sum test), followed by correction for multiple testing either using the Benjamini-Hochberg method [257] or the \( q \)-value method [256].

4.3 Results

4.3.1 Transcriptome-wide perturbation of splicing

We generated exon-level microarray expression data from four individuals with a mutation in \( PRPF8 \), a core component of the spliceosome, and sibling controls. Preprocessing and quality control steps were carried out as described in the Methods section. Following quality control steps and removal of probesets that were not expressed in at least 50% of the samples, a total of 103,268 core and 149,835 full probesets remained. To test for evidence of differential splicing we compared probeset inclusion (defined as the log of the ratio of probeset expression intensity to the expression intensity of the corresponding metaproboset/transcript) between cases and controls. Comparisons were carried out separately for probesets in four overlapping categories, defined by whether the probeset mapped to an annotated intron or exon (referred to as the exonic and intronic categories, respectively) and by whether the probeset belonged to the core probesets or to the full probesets. The latter two categories are an aspect of the microarray design and reflect the confidence of the exon annotations on which the probesets were based. The core probesets correspond to high-confidence exon annotations, while the full probesets also target exons in computationally predicted transcripts. Intronic probesets were further partitioned based on splice type (major and minor).

Probeset inclusion was compared between cases and sibling controls for each probeset that passed the detection and quality control filters described above. Instead of using
Figure 4.1: Empirical distribution of p-values. Each p-values was calculated using pairwise t-tests for a normalised probeset between cases and controls.

limma we used standard t-tests so that we could estimate \( \pi_0 \), the proportion of true null hypotheses. The limma approach has higher statistical power to detect differences at the individual probeset level, because it shares information across probesets but this violates the independence assumption when estimating \( \pi_0 \). Consequently, the expected distribution of p-values is no longer uniform under the null hypothesis [258]. The histogram of p-values obtained from these comparisons (Fig. 4.1) shows a large excess of small values compared to the uniform distribution, expected under a null hypothesis of no difference in splicing between cases and controls. Statistical methods exist to estimate \( \pi_0 \) from a distribution of p-values. We used the qvalue package from BioConductor [256, 259] for this purpose. For core probesets \( \pi_0 \) was 0.76 while full probesets had a slightly higher value of 0.79. This suggests that for approximately one fifth of the probesets the null hypothesis of no difference in inclusion between cases and controls does not hold. For exonic probesets \( \pi_0 \) was 0.77 while intronic probesets had \( \pi_0 = 0.81 \).

It is possible that this high proportion of affected probesets inferred using the qvalue package could be the result of a failure of the assumptions underlying the estimation of \( \pi_0 \) and not a consequence of differences in splicing between the case and control groups. To investigate this possibility, we compared the estimates of \( \pi_0 \) obtained from
the true case-control groups to estimates obtained when sample labels were permuted exhaustively within sibling pairs (so that comparable paired $t$-tests could still be carried out). The estimate of $\pi_0$ for the correctly labelled samples was lower than for any of the eight alternative configurations that can be obtained in this way (Supplementary Figs. A.2 and A.3). The value of $\pi_0$ was strongly anti-correlated with the number of properly paired sibling pairs (Pearson $r = -0.84$, $p = 0.008$; Spearman $\rho = -0.92$, $p = 0.001$), suggesting that the PRPF8 mutation does have a transcriptome-wide effect on mRNA splicing in whole blood samples.

### 4.3.2 Differential splicing primarily involves higher inclusion of exons in cases

We used limma to compute $p$-values and $t$-statistics for differential inclusion of individual probesets because this approach has been shown to have higher statistical power than standard $t$-tests for small sample sizes [255]. The majority of probesets with significantly different inclusion between cases and controls had higher inclusion in cases (Fig. 4.2). Below a $p$-value threshold of 0.2, the enrichment for probesets with higher inclusion in cases was highly statistically significant (core probesets: $p = 4.8 \times 10^{-4}$, full: $p = 1.3 \times 10^{-4}$, exonic: $4.6 \times 10^{-10}$; Supplementary Tables A.1–A.3, Supplementary Fig. A.6). However, intronic probesets showed no significant excess of positive $t$-statistics among probesets with low $p$-values ($p = 0.24$ at $p$-value threshold of 0.2; Supplementary Tables A.4 and A.5). This suggests that most of the differential splicing involves higher exon inclusion in cases. The affected probesets tended to be expressed at a lower level than the remainder of the probesets in the meta-probeset (gene) to which they mapped. Probesets with higher inclusion in cases had a mean expression intensity (across cases and controls) that was, on average, less than half the mean expression intensities of unaffected probesets ($p = 9.6 \times 10^{-112}$). The difference was much smaller for probesets with lower inclusion in cases; these probesets had raw expression intensities that were on average 14% lower than the average of the unaffected probesets ($p = 0.004$). Thus, the affected probesets are from low inclusion exons that are typically included at higher levels in the mutant samples compared to controls.

### 4.3.3 Characterisation of differentially included exons

The introns upstream of exons containing probesets with evidence of differential inclusion between cases and controls were significantly longer than average (Table 4.1). Short introns (< 250 bp) are thought to be excised primarily through intron definition, and longer introns through exon definition [29, 51, 83]. These two classes of introns can be
**Figure 4.2:** Proportion of probesets indicating higher/lower inclusion in cases for binned $p$-value thresholds. All bins have an equal width of $\Delta p = 0.05$ for $p$-values in the range [0, 1]. Red bars symbolise the proportion of probesets indicating higher inclusion in cases; similarly, blue bars refer to lower inclusion in cases. The absolute height of a bar of each colour represents the proportion of probesets with higher/lower inclusion in cases. The $t$-statistic was used to determine relative inclusion: $t > 0$ - higher inclusion in cases; $t < 0$ - lower inclusion in cases. Plot only shows core and exonic (full) probesets.

<table>
<thead>
<tr>
<th></th>
<th>Upstream intron length (5')</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median $p$-value $q$-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>up</td>
<td>1966</td>
<td>0.002</td>
<td>0.0064</td>
</tr>
<tr>
<td>down</td>
<td>2053.5</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>all</td>
<td>1625</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Downstream intron length (3')</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median $p$-value $q$-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>up</td>
<td>1662</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>down</td>
<td>1843.5</td>
<td>0.14</td>
<td>0.18</td>
</tr>
<tr>
<td>all</td>
<td>1625</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 4.1:** Length of introns flanking differentially included exons. Comparison of median of intron length for exons with higher (up) and lower (down) inclusion in cases relative to controls.

seen on the density plot of intron lengths (Fig. 4.3). The density plot of introns bordering differentially included exons had a diminished peak at short lengths (Fig. 4.3), suggesting that aberrant splicing in cases does not affect splicing via intron definition. Instead, longer introns were particularly abundant upstream of preferentially included exons (Fig. 4.3). Thus, differential splicing between cases and controls appears to primarily affect the exon definition pathway.

Exon definition occurs for short to moderate length (< 500 bp) exons since the components of the spliceosome need to associate across them [84]. Long exons tend to have additional splicing enhancer motifs, perhaps to aid the binding of spliceosome components since they cannot associate across the length of the exon [84]. We found that
Figure 4.3: Distributions of intron length. Density plot of intron length (A) upstream and (B) downstream of differentially included exons in cases relative to controls. A similar plot for background exons (core exons) is provided for reference.

Table 4.2: Length of differentially included exons. Comparison of mean and median lengths for exons with higher (up) and lower (down) inclusion in cases relative to controls.

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Mean</th>
<th>p-value</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>up</td>
<td>368.88</td>
<td>127</td>
<td>7.52 × 10^{-18}</td>
<td>1.50 × 10^{-17}</td>
</tr>
<tr>
<td>down</td>
<td>562.81</td>
<td>166</td>
<td>2.44 × 10^{-3}</td>
<td>2.44 × 10^{-3}</td>
</tr>
<tr>
<td>all</td>
<td>436.5</td>
<td>153</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

exons containing probesets with evidence of higher inclusion in cases than in controls were significantly shorter than average (Table 4.2). We also found that exons with lower inclusion in cases were somewhat longer than other exons (Table 4.2). Exons that showed evidence of higher inclusion in cases relative to controls were depleted of long exons (> 500 bp), suggesting that the majority of these exons are spliced via exon definition [53].

We used MaxEntScan [260] to calculate splicing site scores at intron-exon junctions and compared these between affected and unaffected exons. For exons with higher inclusion in cases, the median score of 5’ splice sites (5SS) was significantly higher than background (Tables 4.3 and 4.4) but the 3’ splice sites (3SS) showed no difference in score.
Transcriptome-wide Aberrant Splicing

<table>
<thead>
<tr>
<th></th>
<th>Splice donors (5SS)</th>
<th></th>
<th>Splice acceptors (3SS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>p-value</td>
</tr>
<tr>
<td>up</td>
<td>8.33</td>
<td>8.76</td>
<td>8.86 × 10⁻³</td>
</tr>
<tr>
<td>down</td>
<td>8.12</td>
<td>8.68</td>
<td>0.75</td>
</tr>
<tr>
<td>all</td>
<td>8.08</td>
<td>8.68</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.3: Splice site strength of differentially included exons. Mean and median splice site strength computed using MaxEntScan [260].

<table>
<thead>
<tr>
<th></th>
<th>Sum of 5SS and 3SS scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>up</td>
<td>16.76</td>
</tr>
<tr>
<td>down</td>
<td>16.31</td>
</tr>
<tr>
<td>all</td>
<td>16.38</td>
</tr>
</tbody>
</table>

Table 4.4: Combined splice site strength (5SS and 3SS) of differentially included exons. Similar to Table 3 but using combined splice site strength.

In contrast, for exons with lower inclusion in cases, neither the 5SS nor 3SS showed differences in scores (Table 4.3 and 4.4). The distributions of splice site scores are shown in supplementary figure A.4.

We found that the exons with higher inclusion in cases had much higher proportions of A and T nucleotides than background (core) exons and, correspondingly, lower proportions of G and C nucleotides (Fig. 4.4). Exons with lower inclusion in cases showed no significant difference from background. We also obtained a list of 238 candidate hexameric exonic splicing enhancers (ESEs) sequences [261]. For each ESE sequence, we counted the number of times it occurred in each exon category and normalised this by the sum of exon lengths in that category. Unsurprisingly, considering that ESEs are A-rich (nearly 50% of bases of the 238 ESEs were A), we found that highly included exons did indeed show evidence of ESE-enrichment (mean ESE prevalence of 4.14 × 10⁻⁴ against 3.70 × 10⁻⁴ for core exons, \( p = 0.02 \)). We did not observe any ESE enrichment in exons with lower inclusion in cases (mean ESE prevalence of 3.76 × 10⁻⁴, \( p = 0.72 \)). The enrichment of ESEs in the exons with increased inclusion in cases does not explain the A+T richness because the excess of A and T nucleotides persisted even when all instances of the 238 ESEs were removed. Indeed, the excess of ESEs may be a consequence of the A-richness of the affected exons.

Splicing can take place co-transcriptionally (while the polymerase is still associated with the DNA template) or post-transcriptionally [40, 112, 113, 164, 262–264]. Exons belonging to longer genes or located far from the 3’ ends of genes are more likely than other exons to be spliced co-transcriptionally [265]. To investigate whether the mutation may have a greater impact on co- or post-transcriptional splicing we compared gene length and distance from 3’ gene ends between affected exons (differentially included
Table 4.5: Indicators of co-transcriptional splicing. Gene length and distance from the 3' end of genes for exons differentially included.

<table>
<thead>
<tr>
<th></th>
<th>Gene length</th>
<th>Distance from 3' end</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>up</td>
<td>100,954</td>
<td>53,464</td>
</tr>
<tr>
<td>down</td>
<td>103,761</td>
<td>53,741</td>
</tr>
<tr>
<td>all</td>
<td>91,811</td>
<td>42,729</td>
</tr>
</tbody>
</table>

4.4 Discussion

Cis-acting factors have been shown to be important regulators of alternative splicing and several previous studies have reported cis-acting genetic variants that affect mRNA splicing [236, 238–241, 266–268]. Such cis-acting splicing quantitative trait loci (sQTLs)
Transcriptome-wide Aberrant Splicing

typically affect the splicing of a single gene. *Trans*-acting variants, however, can have widespread impact and when they affect ubiquitous components of the transcript processing machinery the effects may be transcriptome-wide. To date, there have been fewer studies on *trans*-sQTLs than on *cis*-acting variants. Given that almost all human genes are spliced and a large majority of multi-exon genes are alternatively spliced, often in a tissue-specific manner [119], *cis* - and *trans*-acting factors that cause splicing errors or affect the regulation of alternative splicing have the potential to have a substantial impact on the transcriptome. Indeed, a large proportion of human genetic diseases are likely to result from mutations that affect splicing [119, 244, 269, 270].

In this study, we applied statistical analyses that interrogated exon inclusion events across the human transcriptome. We estimated the transcriptome-wide effect of the *PRPF8* mutation by using the $\pi_0$ statistic, which in the context of multiple hypothesis testing, is an estimate of the proportion of tests that conform to the null hypothesis. This approach has previously been applied in transcriptome-wide studies of population differential gene expression between human populations [271] and, more recently, in the study of *cis* - and *trans*-expression quantitative loci [242]. The observed value of $\pi_0$ was the lowest among all permutations of the case-control labels. Nevertheless, we acknowledge that, given the relatively small sample size in this study, this finding requires independent validation particularly given the often noisy nature of microarray data.

Our results provide evidence that a non-synonymous mutation in the splicing factor *PRPF8* that leads to retinitis pigmentosa affects the inclusion of approximately 20% of the exons in genes expressed in whole blood samples. Affected exons were most often included at a higher level in the transcripts of cases than of controls. Averaging across all samples (cases and controls), these exons had lower inclusion levels than unaffected exons, suggesting that they are absent from some of the transcripts of the corresponding gene. The fact that the mutated form of *PRPF8* was associated with higher inclusion of exons that appear to be skipped in some transcripts was an unexpected result. It suggests that the mutation may affect alternative splicing (e.g. tissue-specific regulation of alternative splice isoforms) rather than giving rise to a high rate of splicing errors involving skipping of constitutive exons or intron retention. Exon skipping is the most common type of alternative splicing event [272]. Because the mutation is associated with higher inclusion levels of exons that are expressed at relatively low levels, we propose that the *PRPF8* mutation may reduce the likelihood of exon skipping. Under this model the adverse phenotype could result from failure of a regulated exon skipping event which is required in retina. The exon microarray platform we used also includes some probesets that map to intronic regions. Although the number of expressed probesets mapping to introns was far lower than for exons, we found some evidence of differential intron retention between cases and controls. However, the proportion of affected probesets was
lower for intronic compared to exonic probesets and the overall effect was approximately equally divided between increased and decreased intron inclusion.

The introns upstream of exons with higher inclusion in cases were significantly longer than average. In fact, short introns were almost entirely absent upstream of affected exons (Fig. 4.3). Short introns (< 250 bp) are more efficiently excised through the assembly of spliceosomal components across the intron before pairing up (intron definition) and they generally have weaker splice sites [273]. Long introns rely on spliceosome components first assembling across exons before juxtaposing across the target introns.

One of the most striking features of probesets with significantly different inclusion in cases compared to controls was that they were found predominantly on shorter exons. The majority of affected exons had increased inclusion in cases and we found that shorter exon length applied only to these exons. In fact, probesets with lower inclusion in cases compared to controls mapped to exons that were slightly longer than the background unaffected exons. Exon defined splicing is most efficient when exon length is between 50 and 500 bp [29, 53, 274]. Splicing via intron definition requires short introns (< 250 bp) and, as discussed above, these were depleted among introns flanking affected exons, particularly for the exons with higher inclusion in cases. Consequently, we propose that the \textit{PRPF8} mutation affects splicing via exon definition, but may have no effect on splicing via intron definition. We also observed that exons with higher inclusion in cases tended to have stronger 5' splice site (5SS). Interestingly, \textit{PRPF8} is known to interact directly with the 5SS [49] (and references therein). It contacts the 5SS dinucleotide at residues \textit{QACLK} (positions 1894-1898) as part of the U5 snRNP (a constituent of the U5-U4/U6 tri-snRNP) [275–278]. However, the RP mutation is a histidine to arginine change at residue 2309 [248] suggesting that the mutation does not directly affect contact with the 5SS and may, instead, affect interaction with another protein or snRNA.

On average, exons that were differentially included between cases and controls were much further from 3' gene ends and belonged to significantly longer genes than unaffected exons. Both distance from the 3' end and gene length are associated with the efficiency of co-transcriptional splicing [265]. The majority of affected exons had higher inclusion in cases and exons with higher inclusion in cases (but not exons with lower inclusion) had significantly greater proportions of A and T nucleotides than unaffected exons. A-rich tracts have been observed to lead to pol II pausing [166, 279], potentially increasing the time available for mRNA splicing to occur prior to the completion of transcription. Taking these observations together we propose that the \textit{PRPF8} mutation may primarily or, at least, disproportionately affect splicing that takes place co-transcriptionally.

Two previous studies of the effects of mutations in three splicing factors (\textit{PRPF3}, \textit{PRPF8} and \textit{PRPF31}) linked to retinitis pigmentosa on mRNA splicing in lymphoblast cells
reported evidence of retained introns [231, 234]. Both of these studies investigated splicing for a small number of introns. Only one intron showed evidence of retention associated with the mutation in PRPF8 [234]. This was a U12-type intron (the third intron of STK11). However, the corresponding probesets did not show evidence of expression above background, thus we did not find evidence to support retention of this intron in our dataset. Furthermore, we did not find any evidence of retention of U12-type introns in cases (Supplementary Fig. A.5).

We observed that exons with specific structural characteristics were more likely to be differentially-included: short exons flanked by long introns, high A+T-content, located towards the 5 end of the gene. It is possible that these exons are more prone to microarray hybridisation noise. For example, it can be more difficult to design microarray probesets targeting short exons (short sequences limit the choices of array probes), resulting in lower hybridisation affinity. Indeed, we observed higher variability of expression signals in affected exons compared to unaffected exons even when restricting to controls alone (data not shown). Nevertheless, given that sample preparation and processing was carried out independent of treatment labels, technical noise should not be biased towards increased exon inclusion in cases. Our results suggest that a mutation in PRPF8, implicated in RP, has a subtle effect on the inclusion of a large number of human exons. However, further investigation using newer sequencing based technologies (RNA-Seq) and an independent set of samples will enable the impact of the mutation on splicing error to be confirmed and dissected in greater detail.

4.5 Conclusion

Overall, our results support the hypothesis that a mutation in the splicing factor PRPF8 that leads to retinitis pigmentosa, has a widespread impact on mRNA splicing across the transcriptome. Given that splicing of such a large proportion of exons is effected by the mutation in blood, it is surprising that the phenotype associated with this mutation is restricted to the retina. However, because the differentially included probesets were from exons with low inclusion overall and had higher inclusion in cases compared to controls, we propose that the mutation does not lead to an increase in the rate of mis-splicing of constitutive exons. Instead, the mutation may influence the inclusion of alternatively spliced exons. Consequently, we suggest that the disease phenotype in retinal tissue could result from failure to produce one or several retina-specific isoforms that require exon skipping.
Availability of supporting data

Affymetrix exon array CEL files are available on GEO website under accession GSE43134 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43134). This manuscript is accompanied by supplementary information.
Chapter 5

Seq-ing improved gene expression estimates from microarrays using machine learning

The content of this chapter has been submitted as:


PKK implemented the algorithm, analysed the data and drafted the initial manuscript. Paul Geeleher and Cathal Seoighe were involved in improving the manuscript.

Abstract

The quantification of gene expression by high-throughput sequencing (RNA-Seq) has several advantages over microarrays, including better reproducibility, greater dynamic range and gene expression estimates on an absolute, rather than a relative scale. We propose a novel approach to microarray analysis that attains some of the advantages of RNA-Seq. The method, called Machine Learning of Transcript Expression (MaLTE), leverages samples for which both microarray and RNA-Seq to learn the relationship between the fluorescence intensity of sets of microarray probes and RNA-Seq transcript expression estimates. We trained MaLTE on data released by the Genotype-Tissue Expression (GTEx) project, consisting of Affymetrix gene arrays and RNA-Seq from over 700 samples across a broad range of human tissues. We show that regression models learned from a subset of GTEx samples can accurately estimate RNA-Seq values for the remainder and can be used to estimate absolute gene expression levels from archived microarray data.
5.1 Introduction

Initially designed for DNA mapping and sequencing-by-hybridization [280–282], microarrays were first used to quantify gene expression in the 1990s [283]. Since then, the technology has been adapted for diverse biological assays such as genotyping [284], detection of alternative splicing [121], epigenetic modifications [285], protein-DNA interactions [286] (and references therein), among many others [287]. Because of the importance of gene expression analysis, much effort has been invested in developing accurate methods to infer gene expression levels from the fluorescence intensities of microarray probes. Typical analysis pipelines for oligonucleotide microarrays include subtraction of background signals, normalization of the signal intensities across samples and summarization of the fluorescence intensities of probes that map to the same genomic feature (gene or transcript) [121, 288]. Numerous approaches have been applied to obtain gene-level expression estimates from probe intensities [289] with robust linear models (RMA, parameter estimation by median-polish) and multiplicative models (PLIER, MAS5.0, and dChip) suggested to give the best results [290].

Although they remain in widespread use, microarrays have a number of limitations in comparison to sequencing-based methods for the quantification of gene expression (generally referred to as RNA-Seq). Documented advantages of RNA-Seq include improved reproducibility and dynamic range and gene expression estimates on an absolute scale [291]. The latter attribute facilitates comparison of gene expression levels between different experiments and also enables the expression of different genes within the same sample to be compared. Data on the expression levels of different genes within a sample can be useful, for instance in modeling regulatory networks [291]. The summarized probe intensity values obtained from microarrays, by contrast, are on an arbitrary scale, cannot easily be compared between experiments without renormalization and are not well suited to the comparison of expression levels of different genes in the same sample [292–294]. In addition, the chips are designed according to specific genome annotations and need to be adapted when these annotations change. Finally, the dynamic range of microarrays (a few hundred fold) is an order of magnitude lower than that of RNA-Seq (approx. 9,000 fold depending on the coverage) making them less sensitive [295, 296]. However, because of their relatively low cost, microarrays remain in widespread use and are particularly important for large-scale studies [297–299].

Here we investigate an alternative approach to estimating gene or transcript expression levels from microarray probe level fluorescence intensities, which addresses several of the limitations discussed above. We refer to this method as machine learning of transcript expression (MaLTER). Given a set of samples for which both microarray and RNA-Seq data are available, MaLTER uses statistical regression techniques to learn the relationship.
between the intensities of a set of probes associated with the gene and the expression level of the feature as estimated by RNA-Seq. The regression models can then be applied to estimate gene expression from microarray data for which RNA-Seq data is not available. We applied MaLTE to 716 samples from a broad range of tissues profiled by the Genotype-Tissue Expression (GTEx) project, training the regression models on a random subset of the samples and testing on the remainder. Within individual test samples the gene expression estimates from MaLTE approximate closely the values estimated by RNA-Seq. Cross-sample correlation between RNA-Seq and microarray expression estimates for individual genes was also significantly higher using MaLTE compared to existing methods to estimate gene expression from microarrays. In addition, MaLTE can leverage transcript isoform expression estimates produced from RNA-Seq to provide a means to estimate the expression levels of specific isoforms from the microarray data.

5.2 Methods

5.2.1 GTEx data preparation

Affymetrix Human Gene 1.1 ST microarray and RNA-Seq expression data (derived from 837 samples across 29 tissue types and three cell lines) were downloaded from the genotype-tissue expression (GTEx) project website (http://www.broadinstitute.org/gtex) on 30th July 2013. For 91 of the samples, only microarray data was available and multiple biological replicates were included for several of the cell lines. A total of 716 of the samples were used in this study.

Raw and RMA (median-polish) processed microarray data were downloaded from GEO (accession GSE45878). Library files for Affymetrix Human Gene 1.1 ST arrays were downloaded from the Affymetrix website (http://www.affymetrix.com). Custom CDF files for Human Gene (HuGene11stv1_Hs_ENSG.cdf) that were used in the GTEx project [300] were also downloaded from the Brainarray resource (http://brainarray.mbl.org). Probes were extracted using Affymetrix Power Tools (APT v.1.12.0) with and without quantile-normalization and with background correction (http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx). Expression estimates obtained using RMA were downloaded from GEO. Gene expression was also estimated using PLIER [301] as implemented in APT using the custom CDF (HuGene11stv1_Hs_ENSG.cdf) by passing the following analysis string to APT:

-a quant-norm.sketch=-1.usepm=true.bioc=true,pm-only,plier
We mapped genes to probes using two datasets: (i) the mapping between probes and probe sets implied within the probe intensities file and (ii) the mapping between probe sets and genes. The data for (ii) were constructed using the BEDTools \texttt{intersect} utility. This utility takes two BED files as input: one for gene and another for probe set genomic coordinates. Gene coordinates were determined from the Ensembl v.72 gene model [253] while probe set coordinates were downloaded from the Affymetrix website (http://www.affymetrix.com).

Altogether, we assessed expression estimates for 26,215 genes. These corresponded to 140,212 transcript isoforms. Quantile normalization of gene RPKM values for the 26,215 genes was performed prior to training and testing. MaLTE as well as RMA and PLIER expression estimates were also quantile-normalized prior to downstream analyses.

### 5.2.2 Selecting and tuning the best learning algorithm

We used an independent dataset to select and tune the optimal learning algorithm. We obtained RNA-Seq data [236, 302] and Affymetrix Human Exon 1.0 ST array CEL files [303] for 53 Yoruba (YRI) and 44 European (CEU) HapMap cell lines [304]. All individuals were unrelated. Data was downloaded from the GEO database [305] under accessions GSE25030 (CEU RNA-Seq), GSE19480 (YRI RNA-Seq) and GSE7851 (CEU and YRI exon array data).

We compared the performance of several learning algorithms: \textit{classification and regression trees (CART), multivariate adaptive regression splines (MARS), boosted regression trees (BRT), random forest (RF), conditional random forest (CRF) and quantile regression random forest (QRF)} [306–310]. All learning algorithms were compared to median-polish and PLIER. All methods were applied in the R statistical computing environment [311]. Performance was evaluated based on the mean cross-sample correlation over all genes (see Section 5.3.1 for a definition). The R \texttt{multicore} package (http://cran.r-project.org/web/packages/multicore/) was used to speed up computations.

We identified CRF as the best performing algorithm and then optimized its performance by identifying the best parameter settings. To do this, we created ten datasets, each with 1,000 randomly selected genes. We examined the following parameters: \textit{minimum number of features (probes) selected (FS), number of randomly sampled predictor probes (mtry) and number of trees (ntree)}. We chose the optimal value of each parameter based on the mean cross-sample Pearson correlation coefficient of out-of-bag (OOB) estimates for positively correlated genes ($r_{\text{OOB}} > 0$). We identified the optimal $FS$ by restricting the $n$ probes that were most highly correlated with the response in the training
set. We also identified quantile regression random forest (QRF) as having equivalent performance with the added benefit of producing prediction intervals (upper and lower quantiles). Tuned parameters for the best algorithm were applied without modification on the GTEx dataset.

5.2.3 Effect of training set size

We tested the effect of training set size on prediction accuracy. To do this, we generated training sets having between 20 and 600 randomly selected samples (20, 30, 40, 50, 75, 100, 150, 200, 300, 400, 500, and 600) and evaluated the performance of each training set on a random set of 100 test samples (Fig. B.1). Additionally, we counted the number of genes that passed a nominal OOB filter threshold of zero (Fig. B.1c). From these results, we decided on a training set size of approximately one-fifth of all samples for the training set with the remainder used to test performance.

5.2.4 Differential gene expression (DGE) between GTEx tissues

We selected two GTEx tissues for differential expression analysis: heart muscle (left ventricle) and skeletal muscle. Five samples [312] for each tissue that were among the 570 samples in the test set were randomly selected (designated 10-sample DGE). Only genes for which expression could be estimated using all of the methods tested and with RPKM (in the RNA-Seq data) greater than one in all samples were considered in this assessment. Differential expression analysis was performed using standard $t$-tests [290]. While standard $t$-tests are sub-optimal in practice [313], they enable a fair comparison of quantification methods without platform-specific optimizations. For example, the popular limma [255] method uses moderated $t$-tests that make distributional assumptions that do not necessarily hold for RNA-Seq data [314]. Performance on differential expression was compared using the cumulative Jaccard index (Eqn. 5.1) [315] traversing the list of genes sorted by $q$-value in steps of 10 genes (for computational feasibility) and the concordance correlation coefficient ($\rho_c$) (Eqn. 5.2) [316]. Here, as in the within-sample and cross-sample tests of correlation, we treat RNA-Seq as the gold standard and the microarray methods (MaLTE, RMA and PLIER) are evaluated based on similarity of their DGE results to the RNA-Seq results. We repeated this computation with 100 bootstrap replicates to estimate variability (Fig. B.3).

The Jaccard index $J$ is defined as the ratio of the intersection to the union of two sets,

\begin{equation}
J(A, B) = \frac{|A \cap B|}{|A \cup B|}.
\end{equation}  (5.1)
The concordance correlation coefficient gives the degree of departure from the unit gradient line. Two random variables $Y_1$ and $Y_2$ have a concordance correlation $r_c$ given by

$$r_c = \frac{2\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 + \mu_2)^2}r_{12} = C_b r_{12},$$ (5.2)

where $\mu_i$ and $\sigma_i^2$ are the mean and variance for $Y_i$, and $r_{12}$ is the Pearson correlation coefficient for $Y_1$ with $Y_2$ [316]. The value of $r_c$ was determined using gene lists from differential expression results for each method ranked by the $q$-value. The $q$-values were computed using the Bioconductor `qvalue` package [256].

We also compared the log of fold change between MaLTE and each of median-polish and PLIER separately. First, we identified the common set of genes differentially expressed at FDR < 1% then compared to the respective RNA-Seq log of fold change values. There were 120 genes found in common between MaLTE and median-polish and only six genes common between MaLTE and PLIER.

To determine the consistency of differential expression results, we repeated differential expression analysis using 22 samples for each tissue (44-sample DGE). We then compared the 10-sample DGE results to the 44-sample DGE results treating the latter as true differentially expressed genes using the Jaccard index (Eqn. 5.1). We refer to these as self-comparisons because the 10-sample DGE results were only compared to the 44-sample DGE results from the same method (e.g. MaLTE 10-sample DGE was compared to MaLTE 44-sample DGE; Fig. B.4a). A similar assessment of all methods to the 44-sample RNA-Seq DGE only was also carried out (Fig. B.4b).

### 5.2.5 Application to archived samples

A dataset consisting of Affymetrix Human Exon 1.0 ST array and RNA-Seq expression estimates from a set of human brain samples was downloaded from GEO (accession GSE26586) [317]. To apply MaLTE, trained on Affymetrix Human Gene 1.1 ST arrays, to data from the Affymetrix Human Exon 1.0 ST arrays, we needed to map probes shared between the two platforms using comparison spreadsheets. Spreadsheets relating exon array to gene array meta-probe sets were downloaded from the Affymetrix website (http://www.affymetrix.com). We used meta-probe sets classified as “Best Match” to map exon array probes to gene array probes. A probe was converted if it was part of a probe set that was in a meta-probe set in the “Best Match” spreadsheet and only if it shared 100% sequence similarity between arrays. In total 425,268 of the 861,493 probes on the gene array could be mapped to exon array probes in this way. GTEx and brain RNA-Seq and probe intensities were then quantile-normalized together but we excluded
background correction because the process of transforming the exon arrays excluded several background probes, which would hamper comparison to non-background-corrected median-polish and PLIER.

To compare performance between median-polish, PLIER and MaLTE, we redefined the exon array meta-probe set mappings. (For a detailed description of the exon array design please refer to [318],). We did this to ensure that the same gene-to-probe set definitions were applied to all methods. Meta-probe sets represent the gene/transcript level and are quantified by summarizing the estimates from the individual probe sets. For each Ensembl gene identifier, we identified all exonic probe sets using xmapcore (now annmap) [319]. An exonic probe set is defined as having: (i) all probes mapping to the genome only once, and (ii) all probes falling within an exon boundary. We then used the Ensembl gene identifiers as meta-probe set identifiers. We also constructed a text file of ‘kill list’ probes, which consisted of probes that were excluded from the modified exon array. Gene expression estimates were then computed using APT for custom and core meta-probe sets with the following analysis strings:

```
-a quant-norm.sketch=0.usepm=true.bioc=true,pm-only,med-polish \
-a quant-norm.sketch=0.usepm=true.bioc=true,pm-only,plier \
--kill-list <kill list file>
```

### 5.2.6 Application to the Affymetrix tissue mixture dataset

The Affymetrix tissue mixture dataset was downloaded from the Affymetrix website (http://www.affymetrix.com). This consisted of nine mixtures of brain and heart in varying proportions (1:0 (heart to brain), 0.95:0.5, 0.9:0.1, 0.75:0.25, 0.5:0.5 (three biological replicates), 0.25:0.75, 0.1:0.9, 0.05:0.95, and 0:1) with three technical replicates of each. Because this dataset consisted of Affymetrix Human Gene 1.0 ST arrays, we mapped probes to Human Gene 1.1 ST probe identifiers. The same custom library files (described in Section 5.2.5) were used to evaluate median-polish and PLIER summarization after background correction and quantile normalization. Tissue mixture proportions were then estimated using the R package CellMix [320] for common genes following OOB filtering.

### 5.3 Results

We hypothesized that learning the relationship between single-channel microarray probe intensities and RNA-Seq expression estimates from samples for which microarray and
RNA-Seq data are available could provide a means to obtain improved estimates of gene expression from microarrays. We refer to this general approach as MaLTE. To investigate the performance of the MaLTE approach we applied it to 716 samples from a broad range of human tissues for which high quality Affymetrix Human Gene 1.1 ST microarray and RNA-Seq data have been generated through the pilot phase of the GTEx project [300]. We selected approximately one-fifth (146) of the 716 unique samples at random for use as the training set and tested the performance of MaLTE on the remaining samples. Training on a larger number of samples did not lead to substantial improvement in prediction accuracy (Supplementary Fig. B.1). For each feature of interest (gene or transcript), we identified a set of probes with genomic co-ordinates that overlap the coordinates of the feature and used quantile regression random forest, including a feature selection step, to learn the relationship between the RNA-Seq expression estimates and the probe intensities in the training data (see Section 5.2.2 for details of the choice of regression algorithm). MaLTE is available as an R software package from http://bioinf.nuigalway.ie/MaLTE/malte.html.

5.3.1 Correlation with RNA-Seq

Given a putatively accurate measure of gene expression (here RNA-Seq), our goal is to approximate this measure from the array probe intensities. For most genes we could estimate the RNA-Seq expression level, given as reads per kilobase of transcript per million mapped reads (RPKM) relatively accurately (Fig. 5.1). For the lowest expressed genes, accuracy is limited by stochastic fluctuation in the number of reads from a given transcript and for the highest expressed genes we underestimate expression level due to saturation of microarray probe intensities. We also calculated the correlation between MaLTE and RNA-Seq and compared it to the correlations with RNA-Seq obtained from two existing widely-used methods to estimate expression from sets of microarray oligonucleotide probe intensities (median polish [288] and PLIER [301]). Comparison was carried out both for correlation within samples and across samples. The former provides an indication of the agreement between the methods on the relative expression of different genes within the same sample, while the latter is an indication of how well the variation across samples detected by RNA-Seq is captured by the array estimates. Gene expression levels estimated by MaLTE on the test samples showed much higher within-sample Pearson correlation with the RNA-Seq estimates than median-polish or PLIER (Fig. 5.2a–5.2d). Importantly, the improved within-sample Pearson correlation is not simply a result of MaLTE rescaling the microarray expression estimates to match the RNA-Seq data, as MaLTE also results in substantially higher within-sample rank correlation (Fig. 5.2e). The slopes of the within-sample regression lines were close to
unity for MaLTE (e.g. Fig. 5.2c). In contrast, the expression estimates from the other methods are not on the same scale as the RNA-Seq values (Fig. 5.2a and 5.2b). By placing gene expression estimated from the arrays on the absolute scale defined by RNA-Seq, MaLTE allows comparison of gene expression between genes on the array. This is not possible with standard summarization techniques, such as median-polish and PLIER [291, 293, 294].

**Figure 5.1: Estimation accuracy.** MaLTE provides an estimate of the RNA-Seq gene expression levels from microarray probe intensities. (a) The relative error (i.e. difference between MaLTE estimate and RNA-Seq, divided by the RNA-Seq value) as a function of the RNA-Seq expression level. Each point corresponds to a bin of 75 genes. The data represents all genes but with a random subset of 10 samples for each gene. Only relative errors below 2 and RNA-Seq values between 1 and 1000 are represented. Low expression genes were excluded due to high stochasticity for low read counts. A Loess regression line is shown in red, illustrating that MaLTE slightly underestimates RNA-Seq particularly for highly-expressed genes. (b) The distribution of relative error with percentage median error and median absolute error displayed with the median error indicated by the dashed red line.
The correlation of microarray and RNA-Seq estimates of gene expression has been investigated previously by several studies [236, 321, 322]. Because not all genes vary substantially across samples, while within individual samples mRNA abundance ranges over several orders of magnitude [323], cross-sample correlations tend to be lower than within-sample correlations. MaLTE significantly outperformed median-polish and PLIER in cross-sample correlation (Fig. 5.3). For example, mean cross-sample Pearson correlation (\( \bar{r} \)), in test data was 0.76 for MaLTE compared to 0.72 for median-polish (\( p < 1 \times 10^{-322} \), from a Wilcoxon rank sum test of cross-sample correlations) and 0.68 for PLIER (Fig. 5.3a). Mean Spearman cross-sample correlations (\( \bar{\rho} \)) obtained from MaLTE were also much higher (0.69 compared to 0.64 for median-polish and 0.61 for PLIER; Fig. 5.3b).

**Figure 5.2:** Within-sample correlation with RNA-Seq. (a, b, c) Scatter plot of gene expression for a single exemplary sample for each method against RNA-Seq. The sample with the within-sample Pearson correlation closest to the median over all samples was chosen. Box plots are provided to show the range of within-sample (d) Pearson and (e) Spearman correlation coefficients across samples in the test dataset. Median correlations are indicated beneath in brackets.
Figure 5.3: Cross-sample correlation with RNA-Seq. For each gene, the cross-sample correlation was determined between the gene expression values estimated from the microarrays using MaLTE, median-polish and PLIER. Density plots show the distribution across genes of (a) Pearson and (b) Spearman correlation coefficients. Mean and median values of the correlation coefficients are provided in parentheses next to the method name in the legend. Vertical lines show mean cross-sample correlation for MaLTE (solid), median-polish (dashed) and PLIER (dotted).

5.3.2 Restricting to well-estimated genes

MaLTE does not perform equally well for all genes. For example, low values of cross-sample correlation between MaLTE and RNA-Seq can be obtained for genes with low variation in expression across samples. Such genes will typically also show poor cross-sample correlation when their expression is estimated using median-polish and PLIER. However, MaLTE has the advantage that it provides an estimate of the accuracy with which the expression level of a given gene can be predicted. This is provided by the
cross-validation carried out by random forest when the gene-specific regression model is learned from the training data [308]. Each regression tree in the forest is constructed from a subset of the samples. The expression level of the gene in a given sample can be estimated from the regression trees from which that sample was omitted. This is called the out-of-bag (OOB) estimate. For example, to estimate how well MaLTE will perform for a given gene as assessed by cross-sample correlation with RNA-Seq, we calculate the cross-sample correlation between the OOB estimates and the RNA-Seq data from the training samples. This provides an accurate estimate of the cross-sample correlation in test data (Supplementary Fig. B.2). The OOB estimate can be used as a filter, so that MaLTE returns expression estimates only for genes with a desired property, such as high cross-sample correlation with hypothetical RNA-Seq (note the RNA-Seq data is hypothetical here as MaLTE will typically be applied to samples for which RNA-Seq data is not available). By thresholding on the OOB cross-sample correlation, we found that very high values of cross-sample correlation can be achieved for a subset of genes (Fig. 5.4a). Because genes that pass the OOB cross-sample correlation threshold are likely to have high cross-sample variation, median-polish and PLIER also achieve higher cross-sample correlation for these genes. However, MaLTE maintains a performance advantage over the other methods with increasing threshold values (Fig. 5.4a).

**Figure 5.4: The effects of OOB filtering.** Mean cross-sample Pearson correlation as a function of OOB correlation threshold for (a) genes and (b) transcripts. Error bars correspond to two standard errors. Note that transcript-level estimates are not provided by RMA and PLIER. The black line represents the number of genes/transcripts at each level.
5.3.3 Inference of differential expression

For many gene expression studies, a key objective is to identify the set of genes that are differentially expressed between groups of samples (e.g. disease versus non-disease or treatment versus control). To assess the performance of MaLTE in differential expression analysis we compared gene expression between two different GTEx tissues: heart muscle and skeletal muscle. RNA-Seq has been shown to have higher statistical power than microarrays for detecting differentially expressed genes [321]; therefore, we used similarity to the set of differentially expressed genes identified by RNA-Seq as the metric to evaluate the performance of MaLTE compared to median-polish and PLIER (Supplementary Fig. B.3 and Fig. B.4). To limit the influence of differences in platform-specific techniques for identifying differentially expressed genes (e.g. Cuffdiff for RNA-Seq, limma for microarrays) we used standard t-tests, and ranked the 21,367 common genes by the q-value. We measured the agreement between the ranked lists produced from the microarrays by alternative methods and from RNA-Seq using the concordance correlation coefficient ($\rho_c$) and cumulative Jaccard index (see Section 5.2.4; Supplementary Fig. B.3a). MaLTE showed significantly higher concordance with RNA-Seq than median-polish or PLIER ($\rho_c = 0.34, 0.16$ and $0.12$, respectively; Supplementary Fig. B.3b).

5.3.4 Estimation of transcript isoform expression

A major advantage of RNA-Seq over microarrays is that the former can be used to discover and quantify the expression of novel transcript isoforms, resulting, for example, from alternative splicing. Microarray platforms have been developed to assess alternative splicing by incorporating probes designed to target individual exons or splice junctions. Differential mRNA splicing can be detected by comparing exon inclusion rates or using exon-level parametrized models [324–327]. It is difficult to obtain reliable estimates of expression at the level of transcript isoforms from microarrays, although some methods have been developed for this purpose [183, 185, 267]. MaLTE extends naturally to the estimation of the abundance of specific isoforms by replacing gene expression as the fitted variable with transcript expression. Although it contains a lower density of probes than Affymetrix exon microarrays, the Affymetrix Human Gene 1.1 ST array contains multiple probe sets along human genes and can be used to measure alternative splicing [328]. Using multiple response regression, we learned the relationship between the expression level of multiple transcript isoforms estimated from the RNA-Seq data and the fluorescence intensity of all probes mapping to the gene. We predicted the expression of 140,212 transcript isoforms, achieving mean cross-sample correlation with RNA-Seq
estimates of 0.29. Again, the OOB expression estimates enabled us to identify smaller sets of transcript isoforms whose variation across samples can be predicted more accurately (Fig 5.4b). For example, about 25,000 transcript isoforms have a mean Pearson cross-sample correlation of over 0.65 (Fig. 5.4b and Supplementary Fig. B.5).

5.3.5 Application of MaLTE trained on GTEx data to independent microarray datasets

To determine whether MaLTE regression models, trained on a diverse panel of GTEx tissues, can be applied to estimate expression from microarrays generated independently, we downloaded a dataset, consisting of Affymetrix Human Exon 1.0 ST microarrays and RNA-Seq data from a set of brain samples from a recent study [317] (we refer to this as the Mazin dataset). The fact that these data were from a different array platform (an exon rather than gene array) posed a particular challenge, requiring that we restrict MaLTE to the subset of probes that are shared between the platforms (425,268 of 5,432,523 of the exon array probes are on the gene array). In spite of this, MaLTE again provided considerable improvements in within-sample correlations compared to median-polish and PLIER and similar performance in cross-sample correlation (Figs. 5.5 and Supplementary Fig. B.6). For this comparison, all of the methods used only the set of probes shared between the platforms because these are the only probes available to MaLTE. Without this restriction, the cross-sample correlations obtained using median-polish and PLIER applied to all core probe sets were, in fact, lower than for MaLTE. This is likely to be the result of noise resulting from lower quality probe sets that are not shared between the two platforms. Indeed, the majority of exon array probes have been shown to contribute little to expression signals [328].
Figure 5.5: Application to archived data. MaLTE, trained using the GTEx data, was applied to predict gene expression from published microarray data based on brain samples for which RNA-Seq data was also available. Despite the fact that the two studies used different array platforms (Affymetrix Human Exon 1.0 ST arrays and Affymetrix Human Gene 1.1 ST arrays for the brain and GTEx studies, respectively), MaLTE predictions exceeded the within-sample correlations obtained using median-polish and PLIER. MaLTE predictions were based on probes shared between the two array platforms. Box plots of (a) Pearson and (b) Spearman within correlations are shown. (c) Pearson and (d) Spearman cross sample correlations with OOB filtering. The black line represents the number of genes/transcripts at each level.
5.3.6 Evaluation of MaLTE applied to a controlled tissue mixture dataset

Both of the evaluations of MaLTE discussed thus far involve comparison of expression estimates learned from microarrays with RNA-Seq estimates. This is the primary evaluation because the objective of MaLTE is to use arrays to approximate RNA-Seq (or any gold-standard measure that replaces it and for which a suitable training dataset is available). However, we also applied MaLTE to a tissue mixture dataset, provided by Affymetrix (http://www.affymetrix.com), for which only microarray data are available. The dataset consists of expression estimates from nine mixtures of commercially available heart and brain total RNA, with proportions 1:0 (pure heart), 0.95:0.05, 0.9:0.1, 0.5:0.5 (three biological replicates), 0.1:0.9, 0.05:0.95 and 0:1 (pure brain). At each mixture ratio three technical replicates were conducted, for a total of 33 arrays. Tissues were assayed using Affymetrix Human Gene 1.0 ST arrays. There were 9,455 genes that were called differentially expressed (FDR < 0.05) between the pure heart and pure brain samples with each of the three methods. In the absence of biological noise and measurement error we expect to find a perfect linear correlation between the tissue proportion (i.e. brain or heart proportion) and the expression level for these genes. Because the biological noise is shared between the methods, the Pearson correlation coefficient gives an estimate of how consistent the gene expression measurements are across different sample mixtures. In this test, we would expect PLIER and RMA to outperform MaLTE because PLIER and RMA directly summarize the probe intensities, whereas MaLTE allows more complex relationships between expression levels of different probes and the gene expression estimate. The key advantage of MaLTE is that it provides an estimate of the absolute expression level, whereas, although the RMA and PLIER estimates are consistent across tissue mixtures, their relationship with actual gene expression level can be unclear. Nonetheless, the mean absolute value of the Pearson correlation coefficient from MaLTE was similar to that of RMA and PLIER (0.839 versus 0.874 and 0.896 for MaLTE, RMA and PLIER, respectively).

Because MaLTE provides gene expression estimates on the absolute scale, characteristic of RNA-Seq it has significant advantages over the other methods in certain settings. For example, MaLTE outperformed the other methods when we applied CellMix [320], to estimate the tissue mixture proportions from the expression data, using gene expression deconvolution techniques [329]. The correlation between true and estimated proportions was high for all methods (Supplementary Fig. B.7), but values estimated using MaLTE were closest to the true proportions (the slope of the regression line was 1.02 for MaLTE, compared to 0.86 and 0.96 for RMA and PLIER, respectively; Supplementary Fig. B.7). Both Spearman and Pearson correlation coefficients between the true and estimated
proportions were also highest for the estimates from MaLTE. In all cases filtering was applied (using OOB and removal of noisy genes; see Section 5.2.6) at the training stage to determine which genes could be estimated accurately by MaLTE, but expression estimates for the same set of genes were provided to CellMix for each of the array estimation techniques.

### 5.4 Discussion

Oligonucleotide expression microarrays frequently include multiple probes targeting genes or parts of genes. For these arrays, a key consideration is how to summarize fluorescence intensity signals from multiple probes in order to arrive at an estimate of the feature (e.g. gene or exon) of interest. A wide range of strategies to evaluate the performance of different microarray analysis tools and pipelines have been developed [289, 290]. We propose an alternative method to estimate gene and transcript expression from microarrays as well as a very different approach to the evaluation of performance. Given a gold standard measure of gene expression our goal is to obtain expression estimates from the microarray data that approximate it as closely as possible. The evaluation of performance then becomes a matter of determining how closely the expression estimates derived from the microarrays match the gold standard estimates. In principal, any high-throughput measure of gene expression can play the role of gold standard, provided a sufficiently large and diverse set of training samples with both microarray and gold standard expression data is available. Here we use RNA-Seq because it has been shown to have several important advantages over microarrays [295, 330]. Our goal is to estimate gene expression from microarrays in a way that attains some of these advantages. To achieve this, we learned the relationship between RNA-Seq transcript expression levels and microarray probe intensities from a subset of the GTEx [300] samples, evaluating performance on the remainder, as well as on data from a second study [317] that also generated RNA-Seq and microarrays from the same set of samples.

Compared to two widely-used methods to estimate expression levels from microarrays, MaLTE obtained much greater within-sample correlation with RNA-Seq estimates on both the test GTEx data [300] and on the brain dataset [317] and comparable (for the brain dataset) or significantly better (for the GTEx test data) cross-sample correlation. The performance on the brain dataset was achieved in spite of the use of very different array platforms for the training and test samples in this case that shared only a small proportion of probes (an exon microarray for the brain dataset and a gene microarray for GTEx). In addition to high within-sample correlation coefficients, the slopes of the
regression lines of MaLTE against RNA-Seq for each sample were close to one (Supplementary Fig. B.6). Taken together, this indicates that MaLTE provides an estimate of the RNA-Seq data on the same scale as RNA-Seq. MaLTE can be applied in the context of large studies where RNA-Seq and arrays are applied to a subset of the samples and arrays only to the remainder. The performance on the Mazin dataset suggests that MaLTE, trained on the GTEx data, can also be applied to archived samples, despite batch effects and, in this case, differences in array platforms. Further improvements in performance are likely to be possible if batch effects are modeled and appropriate methods exist for this purpose.

Tissue-specific alternative splicing can complicate the relationship between the signal from a collection of probes and gene expression level because different transcript isoforms may dominate in different tissues. We have found (data not shown) that the performance of MaLTE can be further improved by first carrying out principal component analysis on the probe intensity matrix and including a subset of the principal components as features that MaLTE uses to predict gene expression. In this case the feature set for a given gene includes shared (experiment-wide) principal components in addition to the gene-specific probe intensities.

Our results show that MaLTE, trained on the GTEx dataset, can be applied to estimate gene expression accurately from microarray data generated in other studies. There are currently over 24,000 expression microarray datasets in the GEO database [331] including more than 9,000 from humans. Affymetrix GeneChip gene and exon array platforms together account for over 1,100 expression array experiments, involving over 19,000 samples. MaLTE offers an alternative approach to the analysis of these data, which will allow the estimated gene and transcript expression levels to become more comparable with expression estimates from RNA-Seq. Archived microarray datasets represent a rich resource of data and, by learning the relationship between probe intensities and RNA-Seq expression estimates, MaLTE offers the possibility of joint analysis of data generated using RNA-Seq and microarrays. In general, training datasets that have been assayed using different platforms represent a Rosetta Stone for gene expression measurement, allowing measurements from one platform to be translated to another. Due to the study size and the breadth of sample types, GTEx serves this role for Affymetrix gene oligonucleotide arrays and RNA-Seq.
Chapter 6

Differential expression analysis on retinitis pigmentosa samples using RNA-Seq

*RNA-Seq data was sequenced at TriSeq at Trinity College Dublin.*
*Analysis and authorship of this chapter is by PKK.*

Abstract

RNA-Seq data has become a standard quantification tool in the biologist’s toolbox and is taking over from microarrays. Often, as with other technologies, data may be generated presenting systematic biases introduced either during sample preparation or sequencing placing a high demand on normalisation. Rigorous quality control can help identify and guide in salvaging affected data. We present a thorough analysis of the RNA-Seq data that was generated from similar samples to those analysed in Chapter 4, which had several technical artefacts. From these preliminary analyses we found that filtering out noisy genes prior to simple quantile normalisation reduced the impact of poor coverage and high duplication. We then performed differential gene expression analysis using RNA-Seq, MaLTE and microarray (RMA) expression measures to identify potential aberrantly spliced genes and compared performance and results from each.

6.1 Introduction

As we have presented in the preceding chapter, high-throughput sequencing exhibits several desirable advantages over alternative technologies and we took advantage of these to develop the MaLTE framework. In this chapter, we outline differential gene
expression analysis using RNA-Seq data that was generated as an extension to Chapter
4. This chapter, therefore, extends the analysis of Chapter 4 and 5 and compares
differential expression results obtained using both MaLTE and RMA (exon array) to
RNA-Seq results.

The output of an RNA-Seq experiment are short cDNA sequences called reads (usu-
ally sequenced in pairs from the end of a cDNA fragment) accompanied with per-base
quality scores. Processing RNA-Seq data involves three main steps: mapping reads
to a reference (genome, transcriptome or exome); summarisation of reads into a sin-
gle measure, and normalisation to correct for library-specific differences [332]. Ex-
amples of publicly available mapping tools include BWA [333], Bowtie [334], MAQ
(http://maq.sourceforge.net/index.shtml), SOAP [335] and many others (http://
en.wikipedia.org/wiki/List_of_sequence_alignment_software). A fraction of
the reads will have exon-exon junctions because the cDNA is derived from spliced
mRNA therefore gapped aligners such as TopHat [194] are frequently used. Summarisa-
tion and normalisation are typically performed simultaneously and can be carried out
either arithmetically (sums of reads over specified regions) such as with RPKM [322]
and counts-per-million [336] or using statistical models [337]. Despite the advantages
gained using sequencing-based quantification, there are a number of biases which may
have a severe impact on differential expression analysis. Some of the biases observed
are due to sequencing coverage [338], GC content [339], feature length [340, 341], and
read distribution biases [342, 343]. Correction for biases often entails elaborate statis-
tical normalisation that aims to model and eliminate the bias. Normalisation can be
affected to minimise either between-library biases (such as coverage) or within-library
biases (such as transcript length).

Here we analyse RNA-Seq data that exhibited considerable technical biases in the form
of coverage and duplication. We first performed quality analysis of the RNA-Seq data
highlighting the two major technical biases. We show that excluding poorly covered
genes prior to quantile normalisation can mitigate the effects of poor coverage. We
then perform differential expression analysis using the normalised data and compare
the results to those obtained using MaLTE and RMA. From these analyses, we found
that the Parkinson’s-associated gene α-synuclein (SNCA) showed evidence of differential
expression using all three expression measures and, in conjunction with results from our
previous analysis, may be differentially spliced between cases and controls. We conclude
this chapter with a discussion on SNCA based on a survey of its role in pathology.
6.2 Materials and Methods

6.2.1 RNA-Seq

RNA was extracted from whole blood samples of the four sibling-control pairs, described in Chapter 4, using the QuantiTect® Reverse Transcription (RT) kit (Qiagen) and cDNA libraries were prepared according to the Illumina RNA sequencing protocol. Poly-A containing molecules were purified, fragmented and cDNA was synthesized, followed by end repair and adaptor ligation and purification. Paired-end sequencing was carried out on an Illumina Genome Analyzer II producing reads of 50bp (74bp for sample T3). All lanes were processed using the Illumina GA Pipeline version 1.4.0. We used one lane to sequence the ΦX174 DNA control to calibrate the run quality scores.

6.2.2 RNA-Seq data analysis

Paired-end Illumina reads were mapped to the hg19 build of the human genome using TopHat v.1.1.4 [194] enforcing unique read mapping (~g flag). We evaluated RP mutations based on data provided from the Online Mendelian Inheritance of Man (OMIM) resource (http://omim.org/) and used the samtools mpileup utility [344] to assess variation present in all samples. The quality of aligned reads was evaluated using FastQC v.0.9 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Picard v.1.40 (http://picard.sourceforge.net/). Picard was used to estimate the level of optical duplicates and to mark PCR duplicates. Quantification of gene expression was achieved using Cufflinks v.2.1.1 [181] using the following non-default parameters: -u (multiread correct) and -b ref.fa (fragment-bias correct). Genes were quantified against the Ensembl v.66 gene model [206]. We filtered out all genes having at least four samples expressed below an FPKM of one as well as those having a mean expression in the top percentile. The remaining genes were quantile normalised using the normalizeQuantiles() function from the limma R package [255]. Differential expression was carried out using standard paired t-tests in R [345] and moderated paired t-tests provided in the limma R package. All intermediate analyses (expression asymmetry (Fig. 6.10), overlapping gene lists (Fig. 6.13), were performed using custom R scripts.

6.2.3 Preparation of MaLTE and RMA expression estimates

A detailed outline on how MaLTE expression estimates are obtained is described in Appendix C. Briefly, gene-specific regression models were trained based on the GTEx data described in Chapter 5, using the quantile regression forest algorithm. In order
RNA-Seq Differential Expression Analysis

to make predictions based on the RP exon array data, probe sets from the exon array were mapped to gene array probe sets. The exon array comprises a much larger number of probe sets and exon array probe sets that do not occur on the gene array were discarded. Predictions were filtered at an OOB correlation threshold of 0.5 and quantile normalised. RMA expression estimates were also generated from exon array data based on a modified library consisting only of probes used by MaLTE to ensure they were comparable. RMA expression estimates were computed using Affymetrix Power Tools v.1.12.0 (http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx).

As stated in Chapter 4, RP-PRPF8 samples were designated $T_{1-4}$ and corresponding sibling controls $C_{1-4}$.

6.3 Results

6.3.1 Quality assessment of RNA-Seq data

We generated the RNA-Seq data to validate the microarray results presented in Chapter 4. However, we encountered several technical challenges with this data. This section highlights some of these concerns.

We used FastQC and Picard to assess the quality of the RNA-Seq data. FastQC performs several diagnostic tests on the data to evaluate read quality based on quality scores, GC content, presence of over-represented sequences, and the overall duplication rate. We used Picard to estimate the levels of both PCR and optical duplication. We illustrate each quality assessment result in turn using plots for sample $C_{1}$.

Coverage is defined as the number of times a region (entire genome, transcriptome or targeted locus) is sequenced and gives the expected number of reads covering all genomic bases. It is approximated by the expression

$$C = \frac{LN}{G},$$

where $L$ is the read length, $N$ is the number of reads and $G$ is the genomic length [346] (http://res.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf). Columns two and four in Table 6.1 show the number of mapped reads in millions and the estimated coverage, respectively. It is evident that sample $T_{3}$ has about one-quarter of the reads present in each of the other samples and less than half their coverage and is likely to have a negative impact on the accuracy of differential gene expression analyses. It is important to note that sample $T_{3}$ also differed in read length which

96
Table 6.1: Summary of quality assessment on RP RNA-Seq data from FastQC and Picard.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mapped reads millions</th>
<th>Read length bp</th>
<th>Coverage</th>
<th>% GC</th>
<th>% dupl.</th>
<th>% optical dupl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>29.05</td>
<td>50</td>
<td>0.48</td>
<td>50</td>
<td>77.98</td>
<td>0.011</td>
</tr>
<tr>
<td>$T_2$</td>
<td>34.72</td>
<td>50</td>
<td>0.58</td>
<td>48</td>
<td>68.37</td>
<td>0.0091</td>
</tr>
<tr>
<td>$T_3$</td>
<td>7.97</td>
<td>74</td>
<td>0.20</td>
<td>54</td>
<td>43.61</td>
<td>0.24</td>
</tr>
<tr>
<td>$T_4$</td>
<td>27.65</td>
<td>50</td>
<td>0.46</td>
<td>52</td>
<td>65.08</td>
<td>0.012</td>
</tr>
<tr>
<td>$C_1$</td>
<td>27.77</td>
<td>50</td>
<td>0.46</td>
<td>50</td>
<td>68.16</td>
<td>0.016</td>
</tr>
<tr>
<td>$C_2$</td>
<td>33.45</td>
<td>50</td>
<td>0.56</td>
<td>51</td>
<td>66.68</td>
<td>0.014</td>
</tr>
<tr>
<td>$C_3$</td>
<td>31.40</td>
<td>50</td>
<td>0.52</td>
<td>50</td>
<td>60.62</td>
<td>0.079</td>
</tr>
<tr>
<td>$C_4$</td>
<td>30.56</td>
<td>50</td>
<td>0.51</td>
<td>50</td>
<td>65.08</td>
<td>0.0071</td>
</tr>
</tbody>
</table>

had a positive effect on coverage but not enough to salvage the poor coverage. This single-sample low coverage presented one of the main technical setbacks to using this dataset.

Overall GC content varied from a low of 48% to a high of 54%. However, the first 14 or so bases appeared to exhibit higher GC content than the rest (Fig. 6.1), which arise due to biases introduced by random hexamer primers [343]. The pattern observed in the first 13 bases was identical for all samples characteristic of the primer bias. Moreover, the read quality for the first 14 bases was the highest of all bases confirming that they were not the result of sequencing error (Fig. 6.2). In effect, using the random hexamers non-uniformly samples cDNA fragments from the transcriptome. Therefore, trimming the first 14 bp does not resolve this problem [343] because the underlying sampling bias would persist in the remaining sequences. We therefore used the full read lengths because we could not circumvent this bias. The effect of random hexamer bias was also noticeable in the sequence content (Fig. 6.4) and the over-represented sequences (Fig. 6.5), which were CCAGG, TGTCG, CTGGG, CAGCA, AAAAA (possible poly-A remnants) and CCAGC (60% GC in these pentamers). Base quality had overall means Phred scores per sample of over 36 (Fig. 6.3) reflecting base calling accuracy of over 99.9% [347].
Unfortunately, the data exhibited a high variance in duplication levels ranging from a low of about 44% ($T_3$) to over 77% ($T_1$). There are two forms of undesirable duplication: *optical* and *PCR duplication*. Optical duplicates result during the base reading phase of sequencing in which two or more fragment clusters are located very close to one another on the flow cell resulting in identical coordinates. Therefore, the sequence of nucleotides could be from either cluster introducing uncertainty in the inferred sequence. Using Picard ([http://picard.sourceforge.net](http://picard.sourceforge.net)), we found very low levels of optical duplicates (well under 1%; column 7 of Table 6.1) and therefore ignored them. PCR duplicates, on the other hand, are harder to identify. Defining PCR duplicates as reads having the same sequence content and sharing 5’ coordinates is ambiguous. It ignores the possibility that similar fragments may be sequenced independently. Nevertheless, conditional on the gene expression level, one would expect that the proportion of duplicates arising because of PCR amplification would constitute a substantial proportion of observed duplicates: low for highly expressed genes but may be sizeable at lower levels. Excluding duplicates would have presented a serious challenge in further analysis (Fig. 6.6) therefore we opted to retain them and attempted to correct this through alternative normalisation.
In summary, we observed two main biases that had the potential to severely affect expression analyses: coverage differences (low in $T_3$) and high duplication (most samples but particularly in $T_1$). Overall read quality was good indicated by the high quality scores.
Figure 6.3: Plot of mean quality score for each short read sequence.
Figure 6.4: Plot of nucleotide percentage at each base position.
Figure 6.5: Plot of k-mer enrichment profile across short read sequences.
6.3.2 Identification of retinitis pigmentosa mutations from RNA-Seq data

Sequencing-based transcriptome assays have the advantage of simultaneously providing quantitative and descriptive information. We exploited this feature of RNA-Seq to assess 20 known retinitis pigmentosa (RP) associated genes at 120 loci. To do this, we used the `samtools mpileup` v.0.12.7 utility by compiling bases of reads aligned at individual genomic coordinates [344]. We used the coordinates as specified in OMIM diseases database for the RP mutations (http://www.omim.org). Of the 120 loci examined only 20 were covered in at least one sample and eight loci in all eight samples. The eight loci were all found on PRPF8 and CA4, which showed that, in addition to the PRPF8 mutation (RP13) designating the samples as case-control, there was an additional disease-associated mutation on CA4 (RP17) in samples T2, C2 and C4 (Table 6.2). However, C2 and C4 (both control samples) have not presented with any RP symptoms (Lisa Roberts, personal communication, 27\textsuperscript{th} February 2012).
Table 6.2: **RP-associated mutations identified using RNA-Seq data.** Mutations identified in whole blood samples. (+) present, (-) absent.

<table>
<thead>
<tr>
<th>Design</th>
<th>Gene</th>
<th>Mutation</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP13</td>
<td>PRPF8</td>
<td>p.H2309R</td>
<td>T1 +</td>
</tr>
<tr>
<td>RP17</td>
<td>CA4</td>
<td>p.R14W</td>
<td>T2 +</td>
</tr>
</tbody>
</table>

### 6.3.3 Quality control optimisation

We assessed how experimental groups clustered using principal components and observed the effect of filtering and quantile normalisation on the densities of expression. Principal components can give a global view on the relationship between experimental groups. They can also identify outliers and guide in quality control steps. To the best of our knowledge, applying quantile normalisation after filtering has not been done before. We demonstrate that it may be used to alleviate low coverage bias. First, we quantified expression using **Cufflinks** as described in *Materials and Methods* then applied a log-2 transformation on expression estimates. Clustering was evaluated following different filtering strategies (no filtering, using well-predicted genes using MaLTE OOB filtering, and filtering out noisy genes). The OOB filtering step was done because the final comparison with MaLTE would be based only on common genes between RNA-Seq, MaLTE and RMA. Because the MaLTE genes were OOB genes (see *Materials and Methods*), RNA-Seq would effectively undergo this type of filtering.

Figure 6.7a shows a plot of the first two principal components (PCs) which account for 46.49% (PC1) and 14.9% (PC2) of the variation, respectively. Sample $T_3$ failed to cluster with the other samples and there was no clear partitioning of samples into cases and controls. The truncated densities of expression were also different across samples (Figure 6.9a).
Using OOB-filtered genes (Figure 6.7b), the proportion of variation of the first PCs increased (to 52.47%), though $T_3$ was still an outlier. Next, we removed noisy genes to assess their effect. Typically, low expression (having an expected expression of fewer than one fragment per kilobase of million-mapped reads, one FPKM) and high expression (top 1%) genes may be excluded to eliminate genes dominated by noise [348, 349]. Therefore, we excluded all genes expressed at less than one FPKM in at least four samples and those in the top percentile of mean expression across samples. Sample $T_3$ exhibited much higher densities (Fig. 6.9b) and was even further from the other samples. However, the proportion of variation captured by $PC_1$ markedly increased to 82.5% (Fig. 6.7c). The densities of filtered genes were symmetrical about the same median and mode for all
samples (including $T_3$) though $T_3$ had much higher densities. Because of this, we applied quantile normalisation to eliminate the coverage bias [349]. While quantile-normalisation did not entirely resolve the outlier effect, it imposed equal distributions on all samples. Figure 6.7d shows that $T_1$ and $T_3$ (samples with high duplication and low coverage, respectively) were isolated relatively to the other samples. Additionally, the proportion of variance covered by the $PC_1$ was as high (96.03%) as exhibited by MaLTE (Figs. 6.8a and b; over 97%) and RMA (Figs. 6.8c and d; over 93%). In summary, the combined effect of OOB filtering, filtering out noisy genes and quantile normalisation were able to substantially reduce the coverage and duplication bias.

To assess the effect of quantile normalisation, we performed nominal differential expression analysis using standard $t$-tests. Figure 6.10a shows the proportion of genes

---

**Figure 6.8:** First two principal components for MaLTE and RMA expression estimates. (a) MaLTE prior to filtering out noisy genes, (b) MaLTE after filtering out noisy genes. (c and d) Corresponding plots for RMA.
expressed higher in controls across the full range of $p$-values in 0.05 width bins before quantile normalisation. Figure 6.10b shows a similar plot after quantile normalisation. The asymmetrical attribution to lower expression in cases is almost entirely abolished. Instead, there is a modest abundance of higher expression in cases for a subset of genes reminiscent of observations in Chapter 4 for exons (Fig. 4.2). Figures 6.10c and d show results using moderated $t$-tests (limma) with a similar outcome.

**Figure 6.9:** Density of expression (a) before and (b) after filtering noisy genes. Both plots are truncated on the $x$-axis because the scale extends to -700 indicating numerous genes at very low expression levels. Truncation results in the shown densities not being normalised. Expression estimates ($x$-axis) are log-2 transformed.
Figure 6.10: Effect of quantile normalisation of RNA-Seq data on nominal differential expression. (a and b) Differential expression using standard \( t \)-tests. (a) No quantile normalisation, (b) quantile-normalised expression estimates. (c and d) Differential expression using moderated \( t \)-tests (performed with limma). (c) No quantile normalisation, (d) quantile-normalised expression estimates. Red denotes the proportion of genes that are expressed higher in cases, while blue denotes those higher in controls. Under the null hypothesis of no differences in expression both proportions should be equal at all \( p \)-values.

6.3.4 Comparison of RNA-Seq to MaLTE and RMA

We then assessed how MaLTE and RMA expression estimates derived from the exon microarray experiment described in Chapter 4 compared to the normalised RNA-Seq expression estimates. We first measured the degree of agreement using Pearson and Spearman correlation coefficients as described in Chapter 5 (Figs. 5.2 and 5.3). We used
MaLTE genes filtered at an OOB Pearson correlation of 0.5 to exclude poorly-predicted genes. The comparison was then restricted to 7,455 genes common to all methods.

MaLTE and RMA were almost identical in median cross-sample correlations (Figs. 6.11a, b and Fig. 6.12) but MaLTE had significantly higher within-sample Pearson and Spearman correlations ($p = 1.55 \times 10^{-5}$ for both; Figs. 6.11c and d) indicative of its ability to better estimate absolute expression estimates. Thus, even on this cross-array application, MaLTE exhibited good performance.

**Figure 6.11:** Cross-sample and within-sample correlations. (a) Pearson cross-sample correlations, (b) Spearman cross-sample correlations, (c) Pearson within-sample correlations, (d) Spearman within-sample correlations.
6.3.5 Comparison in differential expression analysis

Finally, we assessed how all expression measures performed in differential expression using both standard and moderated (limma) t-tests. Standard t-tests not only assume that all genes are independent, which is unrealistic, but have lower power, particularly when the number of samples is small (as is often the case) [178]. The moderated t-tests in limma can be applied to RNA-Seq (and by extension MaLTE estimates) but this utilises a different pipeline commencing with raw read counts. To facilitate comparison with array we therefore used limma on filtered RNA-Seq estimates having excluded the genes dominated by noise.
We first estimated $\pi_0$, the expected proportion of true nulls based on the standard (paired) $t$-tests. We estimated these to be 0.84, 0.57 and 0.69 for RNA-Seq, MaLTE and RMA, respectively, all of which provide varying degrees of evidence in favour of the existence of a set of differentially expressed genes. The RNA-Seq data had the weakest evidence, potentially due to lower data quality.

Next, we measured the proportion of overlap with RNA-Seq for top nominally significant genes at various $p$-value thresholds for MaLTE and RMA (Fig. 6.13). At each $p$-value threshold, we obtained the proportion of overlap between the genes identified using each method then compared it to the size of overlap between MaLTE and RMA. We used this to infer the relative ability to identify differentially expressed genes in comparison to RNA-Seq. RMA exhibited consistently higher proportions of overlap using limma (Fig. 6.13b). The results were nearly identical for both methods with standard (Fig. 6.13a) $t$-tests. We also compared how log of fold change value compared to those from RNA-Seq. We first found the common set of genes at FDR < 10% between MaLTE and RMA then computed the corresponding correlations. We found seven genes in common whose log of fold change values with RMA had marginally better Pearson correlations ($r = 0.82, p = 0.024$) compared to MaLTE ($r = 0.76, p = 0.046$). Spearman correlations were not significant.
Figure 6.13: Proportion of overlap with RNA-Seq to nominally significant genes using MaLTE and RMA. (a) Standard $t$-tests, (b) Moderated $t$-tests (performed with limma). Each bar represents a single $p$-value threshold. The proportion is computed relative to the overlap between MaLTE and RMA. Error bars indicate two standard errors.
**Figure 6.14:** α-synuclein transcript isoforms on Ensembl.
Lastly, from the genes with the highest significance, only one was significant using all three measures at an FDR < 10%: α-synuclein (SNCA) (Table 6.3). The log-2 of fold was estimated as 1.14 (RMA), 2.72 (MaLTE) and 2.31 (RNA-Seq). The RNA-Seq estimates indicate that cases expressed SNCA roughly five times higher than controls: the mean (standard deviation) in cases was 59.48 (6.01) FPKM and 12.44 (4.20) FPKM in controls. There were three additional genes at the same FDR shared between RNA-Seq and RMA: solute carrier family 4 (anion exchanger) member 1 (SLC4A1); fatty acid hydroxylase domain-containing 2 (FAXDC2); and oxysterol binding protein 2 (OSBP2), all of which were expressed at higher levels in cases (Table 6.3).

### 6.4 Discussion

The utility of RNA-Seq can easily be offset by technical biases introduced either during library preparation or sequencing. We have presented an analysis of RNA-Seq data based on whole blood samples from individuals with RP-PRPF8 and sibling controls. We were able to confirm the PRPF8 variants and found another RP mutation, although no symptoms were observed in the affected controls. Two technical problems affected the RNA-Seq data we generated: one sample was sequenced at less than half the coverage than the rest and another was sequenced with considerably higher duplication than the others. While there are no standard duplication thresholds, nearly all samples had over 60% duplication. Despite these technical biases, the filtering and normalisation we employed appeared to reduce their effects. Our analysis also presented the potential utility of MaLTE in an analysis similar to the brain microarray data generated by Mazin et al. [317] in which related arrays were employed. As before, our results highlight how MaLTE provides higher within-sample correlations but did not perform as well as RMA on differential expression analysis.

All the genes we identified as differentially expressed were at higher levels in cases. Only one gene, α-synuclein (SNCA) was differentially expressed at a maximum FDR of 10% based on all measures. In Chapter 4, we argued that missplicing appeared to target exons spliced through the exon-defined splicing pathway (Section 2.2.5) in which shorter
than average exons were flanked by longer than average introns. Interestingly, exon five of SNCA was one of those identified (see Section A.7 for accession ENST00000508895:5). This phase 0-0 exon is 84 bp long and flanked by a 92 kbp upstream intron and a 2,533 bp downstream intron making it especially vulnerable to missplicing or being processed as a cassette. It is found in seven (of 11 known) splice isoforms, five of which are translated into 140 bp amino acid sequences (Fig. 6.14). The five transcripts encode the main protein product of SNCA, designated SNCA140 [350]. The other two transcript isoforms containing this exon are translated into 126 bp amino acids (SNCA126) involving a frameshift deletion of exon five [350]. Of the other four protein-coding isoforms, three exclude exon five (all designated SNCA112 because of they encode of 112-115 bp amino acids) [351] and one has an incomplete CDS (see Ensembl transcript table for SNCA). Aberrant splicing in this gene could therefore be the result of higher inclusion of exon five through one of two ways. Higher inclusion could occur if more than the required SNCA140 or SNCA126 isoforms (having exon five) are spliced. On the other hand, higher inclusion could be the result of failure to skip exon five in a SNCA112 isoform. Both scenarios may also occur simultaneously.

The SNCA gene has been intensively studied in relation to its role in neurodegenerative diseases. It is predominantly associated with familial autosomal dominant, juvenile forms of Parkinsons disease (PD) [352, 353]. The SNCA mRNA and protein are expressed in the retinal pigment epithelium (RPE) and neural retinal cells, the latter being consistent with its association with the central nervous system specifically with presynaptic terminals [354]. The latter study was prompted by previous reports citing instances of visual and morphological impairments in retina of PD patients [355]. It had previously been shown that transgenic fruit flies harbouring both normal and mutant (A30P and A53T) SNCA showed a development-associated retinal tolerance of human SNCA only resulting in retinal degeneration in adult flies [356]. However, at present, our results are only speculative and splicing sensitive expression profiling on RP-affected retinal tissue will need to be done to expose any link between SNCA missplicing and splicing factor RP.
Chapter 7

Conclusion

In this thesis, we have examined several questions relating to estimation and analysis of gene expression and alternative splicing focusing on development and disease. We began with a survey of the literature commencing with the seminal work by Sharp [2] and Roberts [4] and ended that discussion on the pervasiveness of alternative splicing in higher order organisms, highlighting some approaches to measurement of alternative splicing. This final chapter briefly summarises the key findings, draws concluding remarks then anticipates future directions.

7.1 Summary

In our first analysis (Chapter 3), we presented evidence that the intron content in a subset of genes associated with embryonic development is conserved in mammals. We showed that homeodomain genes are enriched among the set of genes with conserved intron content. In particular, a far higher proportion of genes that play a role in patterning processes such as somitogenesis have conserved intron content. Next, we performed a transcriptome wide analysis of splicing using exon arrays (Chapter 4). We found that a subset of exons exhibited higher inclusion in individuals affected with an autosomal dominant mutation on PRPF8 that leads to retinitis pigmentosa (RP). PRPF8 is a key splicing factor and our results support the hypothesis that this mutation may act as a trans-sQTL. We then outlined a new approach to gene and transcript isoform expression estimation that provides RNA-Seq-like expression estimates (Chapter 5). Our method, named MaLTE for machine learning of transcript expression, improves microarray expression estimates by learning regression forest models of gene expression between microarray probes and RNA-Seq. From our analysis based on additional expression estimates from RNA-Seq and MaLTE, we were able to provide further evidence for our
Conclusion

results in Chapter 4 that showed higher expression of a subset of genes consistent with higher exon inclusion in RP cases (Chapter 6).

7.2 Conclusions

7.2.1 More emphasis on isoforms rather than genes

It is now widely accepted that almost all multi-exon genes in humans are alternatively spliced with most spliced in a tissue-specific manner [119]. To say that a gene is highly expressed is, therefore, not informative as this provides little indication of the actual transcripts involved. For example, the results presented in Chapter 6 implicated at least one transcript isoform in SNCA of being aberrantly spliced, all of which, when viewed alongside each other, are virtually identical (Fig. 6.14). Given the recent transcript isoform sequencing (TIF-Seq) results from yeast that showed wide diversity in transcript isoforms from an organism that undergoes relatively little alternative splicing [170], it is likely that the resolution of current transcriptome sequencing merely scratches the surface of transcriptome complexity. Isoform expression estimation will undoubtedly take over from gene expression. This will demand advances in quantitative assays and computational approaches that paint a better picture of alternative splicing.

In light of this, there is a noticeable shift towards new approaches that combine the best of various technologies or incorporate new information in isoform analysis. As we outlined in Chapter 2, some of the tools available for isoform quantification incorporate both identification and quantification. However, one group has come up with a novel approach using hybrid sequencing: combining the long reads from PacBios pyrosequencing for isoform discovery with conventional RNA-Seq involving the newly identified isoforms in quantification [357]. It has also been proposed that integrating information on chromatin 3D conformation could enable a better view of transcriptional complexity [358]. The advent of single cell sequencing is also likely to introduce cell-specific features which are typically lost when analysing cell populations [359].

7.2.2 From algorithms to better quantitative assays

One of the observation made during in the course of developing the MaLTE framework was that only a subset of predictors (identified using simple feature selection) were required to predict expression with equivalent accuracy to widely used microarray summarisation. Microarray design may be redundant and could benefit from a design
strategy that works in reverse: identifying optimal probes starting from known expression estimates. High density arrays, such as the exon array, currently rely on the coverage of probes across the entire length of a gene with higher coverage being preferred [318]. However, as has been articulated by others [328], this increases the number of probes expressed at background levels (approx. 80% in the exon array). Understandably, competition between targets as the number of probes increases is likely to be the main source of unreliability. Therefore, higher coverage might not imply better expression resolution but may only be beneficial in detection of splicing events. Alternatively, one study has already demonstrated the clinical potential of isoform-specific microarrays in which a subset of pre-mRNA splicing events of interest may be captured [360]. A MaLTE-like framework could be useful in identifying the most informative probes for the array.

7.3 Future work

7.3.1 Conservation of introns

Our results highlight an important functional role played by intron content in mammals by introducing delays. However, using conservative estimates for RNA pol II elongation rate (e.g. 1.9 kbp min$^{-1}$ [222]) does not full account for the 19 minute delay observed by Takashima et al. [6]. Other avenues we did not investigate are whether the speed of RNA pol II is varied during transcription due to other factors such as RNA pol II queues [166], spliceosomal influences on transcription [361–363], or even epigenetic factors [364, 365].

Future work on intron conservation will require models that explicitly capture insertion and deletion events across several lineages. The work presented here highlighted the potential of using the Enredo-Pecan-Ortheus (EPO) pipeline [366, 367] to infer insertions and deletions from ancestral sequences from which these can be discerned.

7.3.2 Retinitis pigmentosa transcriptome

The retinal transcriptome is complex [232]. Identifying the underlying transcriptome aberrations involved in RP require far higher coverage than our experiment was able to accomplish. Unfortunately, obtaining retinal tissue samples is a challenge. Nevertheless, more work needs to be done on closely related species to study the etiology of retinal degeneration in light of functional transcriptomics. Another alternative would be to use induced pluripotent stem cells using non-retinal samples from individuals affected by RP-PRPF8 (Raj Ramesar, personal communication).
7.3.3 Further development of the MaLTE framework

The MaLTE framework is a starting point for incorporating statistical learning techniques to bridge the gap between legacy assays such as microarrays and newer sequencing-based technologies. While forest-based methods display attractive properties, they have a few limitations. As we pointed out in Appendix B, boosted regression trees (BRT) performed just as well as the random forest variants we used with the exception that BRT was more computationally intensive. In general, forests can be computationally intensive and, depending on the number of trees used, storing the forests consumes a large amount of memory. We were able to bypass the large storage requirements by treating each gene independently then carrying out test data predictions immediately after training and easily lends itself to parallel computation. Improvements can also be made on how forest-based learners estimate the variance without sacrificing accuracy. Resolving these limitations will make MaLTE a powerful tool in extending the benefits of high-throughput sequencing quantification to the more affordable array-based platforms and unlock the vast collection of archived array datasets.

There is no doubt that deeper insight into gene expression and alternative splicing will be instrumental in furthering biological understanding. It is seminal discoveries such as those made by Sharp and Roberts, which have the ability to radically revise on our perspectives, that will advance this goal. The staggering diversity of tissues promises to makes this a long-standing challenge. Ironically, it is often in the context of the transitory effects of development or disease-causing aberrations that we learn much on the intricacies of cellular mechanisms.
Appendix A

Supplementary results for Chapter 4

A.1 Quality Assessment

Figure A.1: Quality assessment of exon array data. A. Hierarchical cluster of samples. $T_i$ are cases, $C_i$ are controls. B. Plot of first two principal components (PCs) for full probesets. C.,D. First two PCs for exonic and intronic probesets, respectively.
A.2 Sample Permutations

A.2.1 Core Probesets

Figure A.2: Sample permutations for core probesets. Plots representing the complete set (eight) of pairwise swaps between cases ($T_i$) and controls ($C_i$). Only core probesets are considered here. Plots are ordered from the top by the value of $\pi_0$. Dashed red line represents a uniform distribution; dashed blue lines represent the value of $\pi_0$. 

123
A.2.2 Full Probesets

Figure A.3: Sample permutations for full probesets. Plots representing the complete set (eight) of pairwise swaps between cases \((T_i)\) and controls \((C_i)\). Plots are ordered from the top by the value of \(\pi_0\). Dashed red line represents a uniform distribution; dashed blue lines represent the value of \(\pi_0\).
A.3 Additional Analyses

A.3.1 Splice Site Scores

Figure A.4: Distribution of splice site scores. Density plot of splice site scores of A. 5’ splice sites and B. 3’ splice sites. The dashed line indicates a similar plot for background exons (core exons).

A.3.2 Constitutive versus Alternative Splicing Analyses

We reasoned that exons differentially included would also be enriched for alternatively spliced exons. However, we found no evidence of splicing bias for exons prone to alternative splicing. We tested this by constructing reference lists for constitutive and alternative exons in lymphoblastoid (similar to the whole blood samples we used) tissue [236] using the following procedure.

First, we aligned RNA-Seq data for 53 YRI samples to the hg19 build of the genome and isolated all junction spanning reads which numbered about 62 million reads. We then constructed a reference set of exons from the Ensembl gene model. Exons were defined at the transcript level by associating each transcript ID with the exon number (e.g. ENST*:1 for the first exon in ENST*). This method produced a redundant list because many exons overlap. Overlapping exons were eliminated using a randomisation procedure: from
every set of exons that shared at least one terminal coordinate (start/stop) one exon
was chosen at random.

For each internal exon we counted the number of reads that mapped to the upstream
\((C_1 : A)\) and downstream \((A : C_2)\) junction as well as those that mapped only to flanking
exons (skipping) \((C_1 : C_2)\) [368]. We computed the percent-/proportion-spliced-in (PSI
or \(\Psi\)) using the expression

\[
\Psi = \frac{\text{avg}(C_1 : A, A : C_2)}{\text{avg}(C_1 : A, A : C_2) + C_1 : C_2}.
\]

An exon was considered constitutive if \(\Psi - 2 \cdot \text{SE}(\Psi) > 0.5\) and alternative if \(\Psi + 2 \cdot \text{SE}(\Psi) < 0.5\). In this way we identified 19,685 constitutive and 3,078 alternative
exons. From this list of constitutive and alternative exons we tallied those present in
higher/lower inclusion in cases and compared them to similar counts in the background
set of exons using Fisher’s exact test. For exons included at higher levels in cases the
odds ratio was 1.57 \((p = 0.45)\) while those included at lower levels had \(OR = 3.59\)
\((p = 0.27)\) suggesting no splicing-specific influence.

### A.3.3 U12-type Introns

We did not find any distinction in probeset intensities based on intron type. U2-type
introns form the majority of introns and Fig. A.6D adequately represents them. U12-
type introns had a less even distribution (Supplementary Fig.A.5) likely due to few
probesets associated with them. No significant bias in inclusion levels was observed. We
also compared U2-type and U12-type introns to each other based on their \(t\)-statistics.
We found no significant difference in the \(t\)-statistics \((p = 0.6268, 95\% \text{ CI of difference of}
means: [-0.27, 0.17])\) suggesting similar distributions for both.
Figure A.5: Proportion of probesets indicating higher/lower inclusion in cases relative to controls for binned p-value thresholds for U12-type introns. Each column represents all probesets (exons) for that bin. Red represent probesets included at higher levels in cases; blue are those included at lower levels in cases.
A.4 Number of Differentially Included Probesets

A.4.1 Differential Inclusion by Probeset Class

**Figure A.6:** Illustration of differential inclusion for different classes of probesets. A. Core, B. full, C. exonic, and D. intronic probesets. Each point represents the number of probesets presenting higher ('up')/lower ('down') inclusion in cases relative to controls within $\Delta p = 0.05$ $p$-value bins.
### A.4.2 Core Probesets

<table>
<thead>
<tr>
<th>$p$ threshold</th>
<th>down</th>
<th>up</th>
<th>equal</th>
<th>$p$-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>473</td>
<td>1287</td>
<td>0</td>
<td>$7.64 \times 10^{-87}$</td>
<td>$1.61 \times 10^{-85}$</td>
</tr>
<tr>
<td>0.05</td>
<td>2417</td>
<td>3963</td>
<td>0</td>
<td>$3.61 \times 10^{-84}$</td>
<td>$3.79 \times 10^{-83}$</td>
</tr>
<tr>
<td>0.1</td>
<td>3008</td>
<td>3760</td>
<td>0</td>
<td>$6.36 \times 10^{-20}$</td>
<td>$4.45 \times 10^{-19}$</td>
</tr>
<tr>
<td>0.15</td>
<td>2801</td>
<td>3230</td>
<td>0</td>
<td>$3.52 \times 10^{-8}$</td>
<td>$1.85 \times 10^{-7}$</td>
</tr>
<tr>
<td>0.2</td>
<td>2625</td>
<td>2913</td>
<td>0</td>
<td>$1.2 \times 10^{-4}$</td>
<td>$4.8 \times 10^{-4}$</td>
</tr>
<tr>
<td>0.25</td>
<td>2446</td>
<td>2557</td>
<td>0</td>
<td>0.12</td>
<td>0.36</td>
</tr>
<tr>
<td>0.3</td>
<td>2253</td>
<td>2362</td>
<td>0</td>
<td>0.11</td>
<td>0.36</td>
</tr>
<tr>
<td>0.35</td>
<td>2190</td>
<td>2271</td>
<td>0</td>
<td>0.23</td>
<td>0.58</td>
</tr>
<tr>
<td>0.4</td>
<td>2073</td>
<td>2122</td>
<td>0</td>
<td>0.46</td>
<td>0.81</td>
</tr>
<tr>
<td>0.45</td>
<td>1975</td>
<td>1959</td>
<td>0</td>
<td>0.81</td>
<td>0.99</td>
</tr>
<tr>
<td>0.5</td>
<td>2003</td>
<td>1956</td>
<td>0</td>
<td>0.46</td>
<td>0.81</td>
</tr>
<tr>
<td>0.55</td>
<td>1881</td>
<td>1856</td>
<td>0</td>
<td>0.69</td>
<td>0.99</td>
</tr>
<tr>
<td>0.6</td>
<td>1857</td>
<td>1848</td>
<td>0</td>
<td>0.90</td>
<td>0.99</td>
</tr>
<tr>
<td>0.65</td>
<td>1784</td>
<td>1761</td>
<td>0</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>0.7</td>
<td>1774</td>
<td>1779</td>
<td>0</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>0.75</td>
<td>1764</td>
<td>1713</td>
<td>0</td>
<td>0.40</td>
<td>0.81</td>
</tr>
<tr>
<td>0.8</td>
<td>1710</td>
<td>1642</td>
<td>0</td>
<td>0.25</td>
<td>0.58</td>
</tr>
<tr>
<td>0.85</td>
<td>1646</td>
<td>1637</td>
<td>0</td>
<td>0.89</td>
<td>0.99</td>
</tr>
<tr>
<td>0.9</td>
<td>1643</td>
<td>1617</td>
<td>0</td>
<td>0.66</td>
<td>0.99</td>
</tr>
<tr>
<td>0.95</td>
<td>1688</td>
<td>1685</td>
<td>0</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>1</td>
<td>1631</td>
<td>1629</td>
<td>207</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table A.1: Number of core probesets in each bin of width $\Delta p = 0.05$. Data used to produce the plots in Fig. 4.2 and Fig. A.6.
### A.4.3 Full Probesets

<table>
<thead>
<tr>
<th>$p$ threshold</th>
<th>down</th>
<th>up</th>
<th>equal</th>
<th>$p$-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1147</td>
<td>1400</td>
<td>0</td>
<td>$5.82 \times 10^{-7}$</td>
<td>$3.054 \times 10^{-6}$</td>
</tr>
<tr>
<td>0.05</td>
<td>3825</td>
<td>4370</td>
<td>0</td>
<td>$1.84 \times 10^{-9}$</td>
<td>$3.86 \times 10^{-8}$</td>
</tr>
<tr>
<td>0.1</td>
<td>4028</td>
<td>4524</td>
<td>0</td>
<td>$8.60 \times 10^{-8}$</td>
<td>$6.02 \times 10^{-7}$</td>
</tr>
<tr>
<td>0.15</td>
<td>3534</td>
<td>4048</td>
<td>0</td>
<td>$3.78 \times 10^{-9}$</td>
<td>$3.40 \times 10^{-8}$</td>
</tr>
<tr>
<td>0.2</td>
<td>3275</td>
<td>3622</td>
<td>0</td>
<td>$3.09 \times 10^{-5}$</td>
<td>0.00013</td>
</tr>
<tr>
<td>0.25</td>
<td>3040</td>
<td>3366</td>
<td>0</td>
<td>$4.88 \times 10^{-5}$</td>
<td>0.00017</td>
</tr>
<tr>
<td>0.3</td>
<td>2991</td>
<td>3129</td>
<td>0</td>
<td>0.080</td>
<td>0.1291</td>
</tr>
<tr>
<td>0.35</td>
<td>2685</td>
<td>2913</td>
<td>0</td>
<td>0.0024</td>
<td>0.0063</td>
</tr>
<tr>
<td>0.4</td>
<td>2642</td>
<td>2887</td>
<td>0</td>
<td>0.0010</td>
<td>0.0031</td>
</tr>
<tr>
<td>0.45</td>
<td>2470</td>
<td>2650</td>
<td>0</td>
<td>0.012</td>
<td>0.026</td>
</tr>
<tr>
<td>0.5</td>
<td>2399</td>
<td>2548</td>
<td>0</td>
<td>0.035</td>
<td>0.062</td>
</tr>
<tr>
<td>0.55</td>
<td>2391</td>
<td>2590</td>
<td>0</td>
<td>0.0050</td>
<td>0.012</td>
</tr>
<tr>
<td>0.6</td>
<td>2408</td>
<td>2378</td>
<td>0</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>0.65</td>
<td>2230</td>
<td>2306</td>
<td>0</td>
<td>0.27</td>
<td>0.35</td>
</tr>
<tr>
<td>0.7</td>
<td>2237</td>
<td>2352</td>
<td>0</td>
<td>0.093</td>
<td>0.14</td>
</tr>
<tr>
<td>0.75</td>
<td>2263</td>
<td>2337</td>
<td>0</td>
<td>0.28</td>
<td>0.35</td>
</tr>
<tr>
<td>0.8</td>
<td>2123</td>
<td>2155</td>
<td>0</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>0.85</td>
<td>2058</td>
<td>2150</td>
<td>0</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>0.9</td>
<td>2089</td>
<td>2143</td>
<td>0</td>
<td>0.42</td>
<td>0.48</td>
</tr>
<tr>
<td>0.95</td>
<td>2059</td>
<td>2032</td>
<td>0</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>1</td>
<td>2194</td>
<td>2052</td>
<td>25057</td>
<td>0.030</td>
<td>0.058</td>
</tr>
</tbody>
</table>

**Table A.2:** Number of full probesets in each bin of width $\Delta p = 0.05$
### A.4.4 Exonic Probesets

<table>
<thead>
<tr>
<th>$p$ threshold</th>
<th>down</th>
<th>up</th>
<th>equal</th>
<th>$p$-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>347</td>
<td>790</td>
<td>0</td>
<td>$2.95 \times 10^{-40}$</td>
<td>$4.59 \times 10^{-30}$</td>
</tr>
<tr>
<td>0.05</td>
<td>1809</td>
<td>2862</td>
<td>0</td>
<td>$6.61 \times 10^{-54}$</td>
<td>$2.059 \times 10^{-53}$</td>
</tr>
<tr>
<td>0.1</td>
<td>2139</td>
<td>2769</td>
<td>0</td>
<td>$2.46 \times 10^{-19}$</td>
<td>$2.55 \times 10^{-19}$</td>
</tr>
<tr>
<td>0.15</td>
<td>1985</td>
<td>2338</td>
<td>0</td>
<td>$8.48 \times 10^{-8}$</td>
<td>$5.29 \times 10^{-8}$</td>
</tr>
<tr>
<td>0.2</td>
<td>1771</td>
<td>2160</td>
<td>0</td>
<td>$5.89 \times 10^{-10}$</td>
<td>$4.59 \times 10^{-10}$</td>
</tr>
<tr>
<td>0.25</td>
<td>1672</td>
<td>1872</td>
<td>0</td>
<td>$0.00083$</td>
<td>$0.0037$</td>
</tr>
<tr>
<td>0.3</td>
<td>1590</td>
<td>1827</td>
<td>0</td>
<td>$5.37 \times 10^{-5}$</td>
<td>$2.79 \times 10^{-5}$</td>
</tr>
<tr>
<td>0.35</td>
<td>1427</td>
<td>1607</td>
<td>0</td>
<td>$0.0012$</td>
<td>$0.00045$</td>
</tr>
<tr>
<td>0.4</td>
<td>1392</td>
<td>1558</td>
<td>0</td>
<td>$0.0024$</td>
<td>$0.00082$</td>
</tr>
<tr>
<td>0.45</td>
<td>1323</td>
<td>1432</td>
<td>0</td>
<td>$0.040$</td>
<td>$0.010$</td>
</tr>
<tr>
<td>0.5</td>
<td>1340</td>
<td>1379</td>
<td>0</td>
<td>$0.47$</td>
<td>$0.085$</td>
</tr>
<tr>
<td>0.55</td>
<td>1212</td>
<td>1342</td>
<td>0</td>
<td>$0.011$</td>
<td>$0.0033$</td>
</tr>
<tr>
<td>0.6</td>
<td>1243</td>
<td>1311</td>
<td>0</td>
<td>$0.19$</td>
<td>$0.038$</td>
</tr>
<tr>
<td>0.65</td>
<td>1190</td>
<td>1308</td>
<td>0</td>
<td>$0.019$</td>
<td>$0.0055$</td>
</tr>
<tr>
<td>0.7</td>
<td>1229</td>
<td>1266</td>
<td>0</td>
<td>$0.47$</td>
<td>$0.085$</td>
</tr>
<tr>
<td>0.75</td>
<td>1176</td>
<td>1186</td>
<td>0</td>
<td>$0.85$</td>
<td>$0.13$</td>
</tr>
<tr>
<td>0.8</td>
<td>1072</td>
<td>1136</td>
<td>0</td>
<td>$0.18$</td>
<td>$0.038$</td>
</tr>
<tr>
<td>0.85</td>
<td>1138</td>
<td>1157</td>
<td>0</td>
<td>$0.71$</td>
<td>$0.11$</td>
</tr>
<tr>
<td>0.9</td>
<td>1170</td>
<td>1136</td>
<td>0</td>
<td>$0.49$</td>
<td>$0.085$</td>
</tr>
<tr>
<td>0.95</td>
<td>1072</td>
<td>1098</td>
<td>0</td>
<td>$0.59$</td>
<td>$0.097$</td>
</tr>
<tr>
<td>1</td>
<td>1202</td>
<td>1110</td>
<td>649</td>
<td>$0.058$</td>
<td>$0.014$</td>
</tr>
</tbody>
</table>

Table A.3: Number of exonic probesets in each bin of width $\Delta p = 0.05$
### A.4.5 Intronic Probesets

<table>
<thead>
<tr>
<th>$p$ threshold</th>
<th>down</th>
<th>up</th>
<th>equal</th>
<th>$p$-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>280</td>
<td>304</td>
<td>0</td>
<td>0.34</td>
<td>0.52</td>
</tr>
<tr>
<td>0.05</td>
<td>914</td>
<td>879</td>
<td>0</td>
<td>0.42</td>
<td>0.52</td>
</tr>
<tr>
<td>0.1</td>
<td>1014</td>
<td>919</td>
<td>0</td>
<td>0.033</td>
<td>0.17</td>
</tr>
<tr>
<td>0.15</td>
<td>800</td>
<td>887</td>
<td>0</td>
<td>0.036</td>
<td>0.17</td>
</tr>
<tr>
<td>0.2</td>
<td>743</td>
<td>673</td>
<td>0</td>
<td>0.067</td>
<td>0.24</td>
</tr>
<tr>
<td>0.25</td>
<td>731</td>
<td>732</td>
<td>0</td>
<td>1</td>
<td>0.85</td>
</tr>
<tr>
<td>0.3</td>
<td>710</td>
<td>665</td>
<td>0</td>
<td>0.24</td>
<td>0.52</td>
</tr>
<tr>
<td>0.35</td>
<td>602</td>
<td>639</td>
<td>0</td>
<td>0.31</td>
<td>0.52</td>
</tr>
<tr>
<td>0.4</td>
<td>628</td>
<td>624</td>
<td>0</td>
<td>0.93</td>
<td>0.83</td>
</tr>
<tr>
<td>0.45</td>
<td>529</td>
<td>617</td>
<td>0</td>
<td>0.010</td>
<td>0.17</td>
</tr>
<tr>
<td>0.5</td>
<td>564</td>
<td>574</td>
<td>0</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>0.55</td>
<td>532</td>
<td>580</td>
<td>0</td>
<td>0.16</td>
<td>0.40</td>
</tr>
<tr>
<td>0.6</td>
<td>536</td>
<td>561</td>
<td>0</td>
<td>0.47</td>
<td>0.52</td>
</tr>
<tr>
<td>0.65</td>
<td>567</td>
<td>538</td>
<td>0</td>
<td>0.40</td>
<td>0.52</td>
</tr>
<tr>
<td>0.7</td>
<td>541</td>
<td>474</td>
<td>0</td>
<td>0.038</td>
<td>0.17</td>
</tr>
<tr>
<td>0.75</td>
<td>517</td>
<td>543</td>
<td>0</td>
<td>0.44</td>
<td>0.52</td>
</tr>
<tr>
<td>0.8</td>
<td>483</td>
<td>519</td>
<td>0</td>
<td>0.27</td>
<td>0.52</td>
</tr>
<tr>
<td>0.85</td>
<td>482</td>
<td>468</td>
<td>0</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>0.9</td>
<td>469</td>
<td>496</td>
<td>0</td>
<td>0.40</td>
<td>0.52</td>
</tr>
<tr>
<td>0.95</td>
<td>458</td>
<td>453</td>
<td>0</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>1</td>
<td>513</td>
<td>459</td>
<td>4235</td>
<td>0.089</td>
<td>0.26</td>
</tr>
</tbody>
</table>

**Table A.4:** Number of intronic probesets in each bin of width $\Delta p = 0.05$
A.4.6 U12-type Intronic Probesets

<table>
<thead>
<tr>
<th>$p$ threshold</th>
<th>down</th>
<th>up</th>
<th>equal</th>
<th>$p$-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.05</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.1</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.15</td>
<td>9</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>0.25</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0.51</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0.51</td>
<td>1</td>
</tr>
<tr>
<td>0.35</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0.18</td>
<td>1</td>
</tr>
<tr>
<td>0.4</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.45</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>0.79</td>
<td>1</td>
</tr>
<tr>
<td>0.5</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0.55</td>
<td>1</td>
</tr>
<tr>
<td>0.55</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>0.61</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0.51</td>
<td>1</td>
</tr>
<tr>
<td>0.65</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.75</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0.51</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.063</td>
<td>1</td>
</tr>
<tr>
<td>0.85</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0.73</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
<td>40</td>
<td>0.18</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.5: Number of U12-type intronic probesets in each bin of width $\Delta p = 0.05$

A.5 Transcript Cluster Identifiers (Metaprobesets) Associated with $PRPF8$

$PRPF8$ is located on the reverse strand of chromosome 17 between base 1553923 and 1588154 (hg19 and EnsEMBL v.66). The following table shows corresponding metaprobesets.
Table A.6: **Metaprobesets along PRPF8**. The coordinates of transcript cluster IDs along the PRPF8 gene.

Only 3740589 and 3740625 passed quality control filtering but were not differentially expressed between cases and controls.

### A.6 Associated Genes

Lists of genes with higher/lower inclusion are available upon request.

### A.7 Associated Exons

The list of exons analysed represent a one-to-one mapping of probesets to exons (see Methods in the main text). The set of exons represent a random choice of all possible exons that overlap with the exon array probeset. The number after the colon (‘:’) represents the exon number with the transcript isoform.

<table>
<thead>
<tr>
<th>ID</th>
<th>chr</th>
<th>strand</th>
<th>start</th>
<th>end</th>
</tr>
</thead>
<tbody>
<tr>
<td>3705896</td>
<td>chr17</td>
<td>+</td>
<td>1577504</td>
<td>1577972</td>
</tr>
<tr>
<td>3705894</td>
<td>chr17</td>
<td>+</td>
<td>1574104</td>
<td>1574222</td>
</tr>
<tr>
<td>3705898</td>
<td>chr17</td>
<td>+</td>
<td>1581163</td>
<td>1581608</td>
</tr>
<tr>
<td>3740589</td>
<td>chr17</td>
<td>-</td>
<td>1574051</td>
<td>1574859</td>
</tr>
<tr>
<td>3740625</td>
<td>chr17</td>
<td>-</td>
<td>1583120</td>
<td>1583533</td>
</tr>
<tr>
<td>3740649</td>
<td>chr17</td>
<td>-</td>
<td>1587369</td>
<td>1587620</td>
</tr>
</tbody>
</table>

ENST00000005340:4 ENST00000011700:2 ENST000000219070:4 ENST000000221973:2
ENST00000013125:29 ENST00000025301:2 ENST000000223642:40 ENST000000225308:3
ENST00000040877:23 ENST00000075503:8 ENST000000229214:5 ENST000000229729:17
ENST00000082259:9 ENST00000083182:7 ENST000000230568:2 ENST000000230901:15
ENST00000158149:10 ENST00000158771:2 ENST00000231368:11 ENST000000234420:4
ENST00000164133:13 ENST00000164133:5 ENST00000236959:12 ENST000000237186:4
ENST00000198765:9 ENST00000200135:14 ENST00000237612:5 ENST00000237612:8
ENST00000200135:5 ENST00000202017:4 ENST00000238831:3 ENST00000242057:7
ENST00000202816:5 ENST00000207549:16 ENST00000242210:4 ENST00000242351:9
ENST00000210060:8 ENST00000215862:19 ENST000000243346:3 ENST00000243346:7
ENST00000216115:4 ENST00000216492:6 ENST00000243583:13 ENST00000243662:2
Supplements for Chapter 4
Supplements for Chapter 4
Supplements for Chapter 4

ENST00000527141:2 ENST00000527587:10 ENST00000542526:15 ENST00000542979:6
ENST00000527591:9 ENST00000528744:2 ENST00000543127:5 ENST00000543192:11
ENST00000529244:5 ENST00000529588:4 ENST00000543523:16 ENST00000543645:9
ENST00000529744:3 ENST00000529970:20 ENST00000543648:2 ENST00000543662:9
ENST00000530044:5 ENST00000531107:19 ENST00000543949:10 ENST000005444784:7
ENST00000530879:2 ENST00000533349:4 ENST00000544853:4 ENST000005444933:3
ENST00000532259:5 ENST00000532502:2 ENST00000545132:6 ENST00000546208:10
ENST00000532742:2 ENST00000533349:4 ENST00000546893:8 ENST00000547043:7
ENST00000533801:14 ENST00000534548:12 ENST00000547101:6 ENST00000547372:15
ENST00000534548:9 ENST00000534738:5 ENST00000547540:2 ENST00000548465:2
ENST00000534881:11 ENST00000535095:9 ENST00000549049:5 ENST00000549379:5
ENST00000535106:7 ENST00000535117:4 ENST00000549851:7 ENST00000549944:7
ENST00000535256:13 ENST00000535346:2 ENST00000550672:5 ENST00000551079:2
ENST00000535650:11 ENST00000536121:2 ENST00000551673:2 ENST00000552351:7
ENST00000536256:11 ENST00000536291:4 ENST00000552618:13 ENST00000552641:5
ENST00000536649:5 ENST00000536681:8 ENST00000552738:23 ENST00000552994:23
ENST00000536688:10 ENST00000536701:6 ENST00000553217:2 ENST00000553369:5
ENST00000536756:4 ENST00000536796:17 ENST00000553412:15 ENST00000554164:2
ENST00000536845:17 ENST00000537036:8 ENST00000554439:2 ENST00000554465:4
ENST00000537081:8 ENST00000537084:21 ENST00000554715:5 ENST00000554752:9
ENST00000537174:11 ENST00000537599:11 ENST00000555223:6 ENST00000555358:13
ENST00000538066:16 ENST00000538611:3 ENST000005556626:19 ENST000005556626:5
ENST00000539368:8 ENST00000539476:8 ENST00000557018:4 ENST00000557332:9
ENST00000539513:6 ENST00000539524:29 ENST00000557381:27 ENST00000558157:14
ENST00000539592:11 ENST00000539625:10 ENST00000558768:20 ENST00000559236:2
ENST00000539933:3 ENST00000540043:8 ENST00000560792:5 ENST00000561552:4
ENST00000540216:7 ENST00000540246:3 ENST00000562676:8 ENST00000563197:2
ENST00000540418:6 ENST00000540444:42 ENST00000563264:3 ENST00000564674:6
ENST00000540823:7 ENST00000540905:5 ENST00000565267:11 ENST00000565435:3
ENST00000541133:5 ENST00000541188:8 ENST00000565488:4 ENST00000565502:3
ENST00000541296:8 ENST00000541377:29 ENST00000565644:3 ENST00000565728:10
ENST00000541394:10 ENST00000541549:8 ENST00000566389:3 ENST00000566621:3
ENST00000542094:5 ENST00000542289:2 ENST00000569057:4
ENST00000542339:4 ENST00000542445:9
A.7.1 Lower Inclusion in Cases

ENST00000040663:4 ENST000000199708:2 ENST000000342111:2 ENST000000342711:11
ENST000000218436:8 ENST000000221856:3 ENST000000343516:15 ENST000000344799:7
ENST000000225275:10 ENST000000225805:8 ENST000000345125:2 ENST000000348962:6
ENST000000226432:6 ENST000000228916:7 ENST000000349874:9 ENST000000350221:8
ENST000000231061:9 ENST000000237163:30 ENST000000352115:7 ENST000000352896:14
ENST000000237201:3 ENST000000243300:26 ENST000000354381:3 ENST000000355697:13
ENST000000243914:10 ENST000000246199:2 ENST000000355758:2 ENST000000356865:9
ENST000000251507:10 ENST000000252729:3 ENST000000357041:15 ENST000000357943:9
ENST000000253110:5 ENST000000255174:3 ENST000000358552:3 ENST000000358866:2
ENST000000256379:5 ENST000000257100:11 ENST000000359169:2 ENST000000359495:2
ENST000000257198:4 ENST000000257290:9 ENST000000360003:15 ENST000000360064:20
ENST000000258341:21 ENST000000259351:2 ENST000000360790:17 ENST000000360825:3
ENST000000261292:9 ENST000000261707:9 ENST000000361731:7 ENST000000361924:17
ENST000000261722:18 ENST000000263063:7 ENST000000366652:2 ENST000000366829:5
ENST000000264345:2 ENST000000264705:2 ENST000000367492:100 ENST000000367492:49
ENST000000264741:24 ENST000000268695:10 ENST000000367714:2 ENST000000368125:2
ENST000000269554:8 ENST000000270225:5 ENST000000368169:4 ENST000000368416:3
ENST000000271854:15 ENST000000275560:5 ENST000000368761:4 ENST000000369085:3
ENST000000276440:18 ENST00000281017:3 ENST000000369538:7 ENST000000370296:7
ENST000000281172:6 ENST000000286332:5 ENST000000370336:11 ENST000000370449:10
ENST000000291547:6 ENST000000292246:5 ENST000000371494:6 ENST000000371923:6
ENST000000293590:10 ENST000000294618:4 ENST000000372044:9 ENST000000374324:8
ENST000000295294:2 ENST000000295408:5 ENST000000375180:3 ENST000000375755:5
ENST000000296046:2 ENST000000296794:34 ENST000000377314:8 ENST000000377882:6
ENST000000296932:4 ENST000000297404:2 ENST000000379922:10 ENST000000380550:20
ENST000000298527:2 ENST000000301336:4 ENST000000380968:4 ENST000000380993:22
ENST000000302174:7 ENST000000306317:7 ENST000000381480:9 ENST000000381913:14
ENST000000306821:6 ENST000000307092:4 ENST000000382040:4 ENST000000382233:8
ENST000000309930:7 ENST000000310018:2 ENST000000383702:12 ENST000000383807:9
ENST000000312584:2 ENST000000312655:2 ENST000000388781:13 ENST000000389374:10
ENST000000314607:11 ENST000000322349:22 ENST000000389501:10 ENST000000389934:8
ENST000000322893:6 ENST000000333479:34 ENST000000390015:20 ENST000000391986:5
ENST000000333850:8 ENST000000335968:15 ENST000000392200:16 ENST000000393853:6
ENST000000335968:17 ENST000000336148:14 ENST000000394809:9 ENST000000395314:13
Appendix B

Supplementary results for Chapter 5

B.1 Effect of training set size

Figure B.1: The effect of training size on predictions. A test set of 100 samples was used for all training sample sizes.
B.2 OOB filtering

Figure B.2: Out-of-bag (OOB) filtering. Scatter plot of cross-sample correlation coefficients from on OOB gene expression estimates and estimates obtained on test data for 1,000 randomly selected genes.
B.3 Comparison of DGE results

**Figure B.3:** Comparison of differential expression results. Comparison of the performance of MaLTE and median-polish on the problem of detecting differential gene expression between five heart-muscle and five skeletal-muscle samples in the GTEx dataset. Each method results in a list of genes, ranked by the $q$-value from the comparison of the gene expression level in the two groups of samples. We used the (a) cumulative Jaccard index and (b) concordance correlation to compare the similarity as a function of rank and the overall similarity, respectively, between lists of genes ranked by MaLTE/median-polish and by RNA-seq. The set of genes assessed were defined by RNA-Seq: genes with RPKM of above one in both tissues. Bootstrap re-sampling (100 pseudo-replicates) was used to assess the effect of sampling error for both cumulative Jaccard index and concordance correlations. Results of bootstrapping are shown as faint lines around the observed Jaccard index and yield the density plots shown in (b). Log of fold change estimates for (c) 120 differentially expressed genes common to MaLTE and median-polish and (d) MaLTE and PLIER for six common genes.
Figure B.4: Comparison of 10-sample DGE to 44-sample DGE. (a) Self-comparisons (e.g. MaLTE 10-sample DGE to MaLTE 44-sample DGE, with five and 22 samples in each tissue, respectively). Forty-four-sample DGE results are treated as putative true differential expressions at the indicated (top right or each plot) false discover rates (FDRs). FDRs were computed using the q-value method [256]. Comparisons are made using the Jaccard index. (b) Comparison of each method to 44-sample DGE using RNA-Seq.
B.4 Cross-sample correlations for transcript isoform expression estimates

**Figure B.5:** Transcript isoform expression estimates. Densities of (a) Pearson and (b) Spearman cross-sample correlations for transcript isoform expression estimates obtained using MaLTE. Filtered data corresponds to transcript isoforms with $r_{OON} > 0$. 
B.5 Slopes of regression between RNA-Seq and array methods assessed for Brain data

Figure B.6: Box plots of slopes $\beta$ computed by linearly regressing RNA-Seq against array method. Dotted red line indicates unit gradient. Each box plot represents 12 slopes for brain samples. RNA-Seq expression is restricted to RPKM between one and 1000 because of high uncertainty at low values and few genes at high values representing 78% of genes quantified. Using all genes results in the same medians but wider variance in slopes due to outlying genes.
### B.6 Prediction of tissue-mixture proportions

**Figure B.7: Estimation of tissue mixture proportions.** (a) MaLTE, (b) median-polish (RMA), and (c) PLIER. Each plot shows comparisons of true and estimated proportions with key statistics indicated in the legends.
Appendix C

Use Case: MaLTE Trained with GTEx Data Applied on Exon Array

C.1 Introduction

Using MaLTE with the GTEx training dataset consists of three main steps illustrated in Figure C.1.

- Obtaining the package and training data
- Preparing the data for training-and-testing
- Predicting, filtering and collating expression estimates
Figure C.1: Schematic representing a MaLTE use-case. The four text files (dark green) are the main input to MaLTE. These files are obtained by processing the training data using the auxiliary scripts.
In the first step, the user needs to download a set of auxiliary data and scripts before installing the MaLTE R package. The training (and test) data may then be downloaded after which several preprocessing steps may be carried out depending on the type of array involved. For example, we will need to transform the exon array to a gene array. The transformed array data is then quantile normalized to reduce batch effects.

The second and third steps are relatively straightforward and employ only MaLTE functions as outlined from Sections C.6.

## C.2 Obtaining Auxiliary Data and Scripts

The use-case presented here requires several auxiliary data and scripts to convert the exon array to a gene array. These are all contained in the MaLTE-aux Github repository which may be directly downloaded from https://github.com/polarise/MaLTE-aux. Click the ‘Download’ button to get the whole directory as a zip file.

```bash
unzip MaLTE-aux-master.zip
cd MaLTE-aux
```

Alternatively, it is much faster to use git as follows:

```bash
git clone https://github.com/polarise/MaLTE-aux.git
cd MaLTE-aux
```

This directory contains pre-compiled resource files described in Table C.1. We strongly recommend that all analysis be carried out in this directory.

The script `create_samples_template.py` compiles the resource file `samples.txt` by choosing a random subset of samples (the user specifies the number) to be used for training specified in `training_samples.txt`. More information on how it works can be obtained by running

```bash
./create_samples_template.py --help
```

The script `transform_microarrays.py` converts one array to another using several resource files provided in MaLTE-aux. A detailed specification of how to use this script is available in the file `transform_microarrays.help`.
MaLTE Use-Case

<table>
<thead>
<tr>
<th>File</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>create_samples_template.py</td>
<td>These script and input data files are used to generate samples.txt.</td>
</tr>
<tr>
<td>training_samples.txt</td>
<td></td>
</tr>
<tr>
<td>transform_microarrays.py</td>
<td>The Python script is used to transform the exon array to a gene array. Some probes are omitted in the process. The .help file gives the explicit use of the data files below.</td>
</tr>
<tr>
<td>transform_microarrays.help</td>
<td></td>
</tr>
<tr>
<td>HuEx-1.0-st-v2.dt1.hg18.full.mps.gz</td>
<td>Data files required by transform_microarrays.py</td>
</tr>
<tr>
<td>HuEx-1.0-st-v2.r2.pgf.txt.gz</td>
<td></td>
</tr>
<tr>
<td>HuExVsHuGene_BestMatch.txt.gz</td>
<td></td>
</tr>
<tr>
<td>HuGene-1.1-st-v1.r4.mps.gz</td>
<td></td>
</tr>
<tr>
<td>HuGene-1.1-st-v1.r4.pgf.txt.gz</td>
<td></td>
</tr>
<tr>
<td>gene_probesets_HuGene_Ens72.txt.gz</td>
<td>Input data files required to run the MaLTE function prepare.data()</td>
</tr>
<tr>
<td>transcript_probesets_HuGene_Ens72.txt.gz</td>
<td></td>
</tr>
</tbody>
</table>

Table C.1: Contents of MaLTE-aux and the functions they provide

C.3 Working Directory Structure

As all analysis will be carried out in MaLTE-aux directory, we need to create two directories for the training and test array CEL files.

```bash
mkdir GTEx_CELL
mkdir RP_CELL
```

C.4 Obtaining MaLTE

MaLTE may be downloaded as a pre-built or source package. Pre-built versions can be found at [https://github.com/polarise/MaLTE-packages](https://github.com/polarise/MaLTE-packages). The latest version should be downloaded (files are not sorted by version), ensuring that all prerequisites (see the manual available [http://bioinf.nuigalway.ie/MaLTE/malte.html](http://bioinf.nuigalway.ie/MaLTE/malte.html)) are installed before installing MaLTE.

```R
R CMD INSTALL MaLTE_<version>.tar.gz
```

The source package requires an installation of git and is obtained as followed:

```bash
git clone https://github.com/polarise/MaLTE.git
```
then built with

R CMD build MaLTE

and installed using R CMD INSTALL as shown above.

C.4.1 Creating the file samples.txt

The samples.txt file is created by running

./create_samples_template.py

The names of the test samples should be appended at the bottom of this file using a text editor. For example, to add a test sample called Test01.CEL add the following line to samples.txt:

*NA<tab>Test01.CEL

The asterisk (‘*’) marks this sample as a test sample and ‘NA’ indicates that RNA-Seq data is not available. Repeat this for as many samples as are present in the test data.

C.4.2 Obtaining the training data

**GTEx RNA-Seq data.** Gene expression levels estimated from RNA-Seq data may be downloaded from [http://www.broadinstitute.org/gtex](http://www.broadinstitute.org/gtex). The downloaded data needs to be processed to arrive at the training state by excluding extraneous information and converting gene annotation to Ensembl (from GENCODE) then quantile-normalizing only those genes with probe sets. Pre-processed data may be downloaded from [https://drive.google.com/file/d/0BxLgaMV5aZahVDNjM29WeGh2NHM/edit?usp=sharing](https://drive.google.com/file/d/0BxLgaMV5aZahVDNjM29WeGh2NHM/edit?usp=sharing).

**GTEx gene array data.** CEL files for the GTEx microarray data are available as a single zipped archive which may be downloaded from GEO archive GSE45878 into the directory GTEx_CEL. Background-corrected probes should be extracted as follows using the Affymetrix Power Tools function apt-cel-extract, which depends on the Affymetrix GeneChip® Human Gene 1.1 ST Array library files available at [http://www.affymetrix.com/estore/catalog/prod350003/AFFY/Human-Gene-ST-Array-Strips#1_3](http://www.affymetrix.com/estore/catalog/prod350003/AFFY/Human-Gene-ST-Array-Strips#1_3). Select the ‘Technical Documentation’ tab to get a link to the library files.
MaLTE Use-Case

apt-cel-extract -o GTEx_probe_intensities.txt \
-c /path/to/HuGene-1_1-st-v1.r4.clf \
-p /path/to/HuGene-1_1-st-v1.r4.pgf \
-b /path/to/HuGene-1_1-st-v1.r4.bgp \
-a pm-gcbg GTEx_CEL/*.CEL

It may be necessary to extract background probes file using apt-dump-pgf.

apt-dump-pgf \
-p /path/to/HuGene-1_1-st-v1.r4.pgf \
-c /path/to/HuGene-1_1-st-v1.r4.clf \
--probeset-type antigenomic \n-o HuGene-1_1-st-v1.r4.bgp

Exon array data. We outline this procedure using a previously uploaded dataset available under GEO accession GSE43134 into the directory RP_CEL. The Affymetrix GeneChip® Human Exon 1.0 ST Array library files are available from http://www.affymetrix.com/estore/catalog/131452/AFFY/Human-Exon-ST-Array. Similarly, background-corrected probe fluorescence intensities should be extracted as follows:

apt-cel-extract \n-o RP_probe_intensities.txt \n-c /path/to/HuEx-1_0-st-v2.r2.clf \n-p /path/to/HuEx-1_0-st-v2.r2.pgf \n-b /path/to/HuEx-1_0-st-v2.r2.antigenomic.bgp \n-a pm-gcbg RP_CEL/*.CEL

C.5 Transforming exon array probes to gene array probes

We now run transform_microarrays.py using the following template

./transform_microarrays.py \
-e RP_probe_intensities.txt \
-g GTEx_probe_intensities.txt \
-c HuExVsHuGene_BestMatch.txt.gz \
-i HuGene-1_1-st-v1.r4.pgf.txt.gz
MaLTE Use-Case

-`f HuEx-1_0-st-v2.r2.pgf.txt.gz`
-`p HuEx-1_0-st-v2.r2.dt1.hg18.full.mps.gz`
-`q HuGene-1_1-st-v1.r4.mps.gz`
--`huex-out BestMatch_HuEx_probe_intensities.txt.gz`
--`huge-out BestMatch_HuGe_probe_intensities.txt.gz`

Once the modified probe intensity files have been created they need to be quantile-normalized together with the training data as follows:

```r
# BESTMATCH DATA
huex.best <- read.table("BestMatch_HuEx_probe_intensities.txt.gz", header=TRUE, stringsAsFactors=FALSE, check.names=FALSE)

huge.best <- read.table("BestMatch_HuGe_probe_intensities.txt.gz", header=TRUE, stringsAsFactors=FALSE, check.names=FALSE)

# all raw data
data <- cbind( huge.best[,8:ncol( huge.best )], huex.best[,8:ncol( huex.best )] )

# quantile normalisation only
qnorm.data <- normalizeQuantiles( data, ties=FALSE )

# save the quantile-normalised data for later
# (extracting principal components)
save( qnorm.data, file="qnorm.data.Rdata" )

# re-insert probe metadata
hugeex.qnorm.data <- cbind( huge.best[,1:7], qnorm.data )

# save to file
write.table( hugeex.qnorm.data, file="BestMatch_GTEX_RP_probe_intensities_QN.txt", col.names=TRUE, row.names=FALSE, quote=FALSE, sep="\t" )
```

All training data needed is now available to use MaLTE.
C.6 Preparing the data for training-and-testing

We can now use MaLTE to prepare the data for training and testing. First, we load the package after launching R.

> library( MaLTE )

We use the prepare.data() function which takes as arguments the samples.txt, GTEx_subset_gene_rpkm_QN.txt, BestMatch_GTEx_RP_probe_intensities_QN.txt.gz, and gene_probesets_HuGene_Ens72.txt.gz (available in MaLTE-aux directory).

> prepare.data( samples.fn='samples.txt',
hts.fn='GTEx_subset_gene_rpkm_QN.txt.gz,'\nma.fn='BestMatch_GTEx_RP_probe_intensities_QN.txt.gz,'\ng2p.fn='gene_probesets_HuGene_Ens72.txt.gz', raw=TRUE )

We then read in the training and test data.

> tt.ready = read.data( train.fn='train_data.tar.gz',
test.fn='test_data.tar.gz' )

C.7 Prediction

Prediction can either be performed on the test data training data (out-of-bag, OOB). We outline OOB shortly. First, we need to define training-and-test parameters.

> tt.params = TT.Params() # default parameters

This sets the following values:

> tt.params
mtry = 2
ntree = 1000
feature.select = TRUE
min.probes = 15
cor.thresh = 0
OOB = FALSE
QuantReg = FALSE
Tune (OOB cor.P) = NA
Predictions are performed using the `array2seq()` function.

```r
> tt.seq = array2seq( tt.ready, tt.params )
```

OOB predictions are performed using the `array2seq.oob()` function.

```r
> tt.seq.oob = array2seq.oob( tt.ready, tt.params )
```

### C.8 Filtering by OOB

OOB filtering is performed using the `oob.filter()` function, which takes a `tt.seq` pair (`tt.seq` and `tt.seq.oob`) and a filtering threshold together with the correlation to use for filtering (Pearson (default) or Spearman),

```r
> tt.filtered = oob.filter( tt.seq, tt.seq.oob, thresh=0.5, method='spearman' )
```

### C.9 Collating predicted gene expression values

Predicted values should be collated as a data frame using the `get.predictions()` function. First, we get the names of test samples using the `get.names()` or `get.test()` functions.

```r
> test.names = get.names( samples.fn='samples.txt', test=TRUE )
> test.names = get.test( samples.fn='samples.txt' )
```

The predictions can be quantile-normalized but this may be omitted.

```r
> df.malte.qn = get.predictions( tt.filtered, sample.names=test.names, qnorm=TRUE )
```

then saved to file

```r
> write.table( df.malte, file='RP_malte_OOB_0p5_QN.txt',
              col.name=TRUE, row.names=FALSE, quote=FALSE, sep='	' )
```
MaLTE Use-Case

C.10 Experimental features

C.10.1 Per-gene/transcript tuning

Set \texttt{tune.cor.P=TRUE} in \texttt{TT.Params()}

\begin{verbatim}
> tt.params = TT.Params( tune.cor.P=TRUE )
\end{verbatim}

C.10.2 Incorporating principal components

Principal components of probe intensities may be incorporated into the prediction. The
same principal component will be used for all genes/transcript therefore they must be
incorporated into the \texttt{samples.txt} file to be dispatched to all genes/transcripts. We have
experimented with the first 10 principal components.

\begin{verbatim}
> load( "dnorm.data.Rdata" )  \# load qnorm.data from before
> pc.data = princomp( qnorm.data )

\# save the first 10 principal components
> pcs = pc.data$loadings[,1:10]
> samples = read.table( samples.txt', header=TRUE )

\# create new samples data frame augmented with principal components
> samples.aug = cbind( samples, pcs )
> colnames( samples.aug ) = paste( P', 1:10, sep='' )

\# create a new samples file
> write.table( samples.aug, file='samples_aug.txt',
\quad col.names=TRUE, row.names=FALSE, quote=FALSE, sep='\t' )
\end{verbatim}

We then run \texttt{prepare.data()} using \texttt{samples_aug.txt} (instead of \texttt{samples.txt} as before)
setting the variable \texttt{PCs=TRUE}. This creates two files: \texttt{train_PCs.txt} having the training
principal components and a corresponding \texttt{test_PCs.txt}.

\begin{verbatim}
> prepare.data( samples.fn=samples.fn, hts.fn=hts.fn, ma.fn=ma.fn,
\quad g2p.fn=g2p.fn, PCs=TRUE )
\end{verbatim}

Also, we need to run \texttt{read.data()} setting \texttt{PCs.present=TRUE},
\texttt{train.PCs.fn='train_PCs.txt'} and \texttt{test.PCs.fn='test_PCs.txt'}.
> tt.ready.pc <- read.data( train.fn="train_data.txt.gz",
    test.fn="test_data.txt.gz", PCs.present=TRUE,
    train.PCs.fn="train_PCs.txt", test.PCs.fn="test_PCs.txt" )

The principal components are now embedded in the data for each gene.
Bibliography


Bibliography


Bibliography


Bibliography


