The role of the 3' untranslated region in breast cancer

Andrew David Pattison
BSc (Hons)

A thesis submitted for the degree of Doctor of Philosophy at Monash University in 2018

Department of Biochemistry and Molecular Biology
Monash Biomedicine Discovery Institute

Copyright notice

Contents

Contents 1

Abstract 6

Declaration 8

Preface 9

Acknowledgements 11

Definitions and Abbreviations 13

List of figures 16

List of tables 18

Chapter 1: Literature review and introduction 19

1.1 Topics discussed in this literature review 20

1.2 Breast cancer prevalence, subtypes, impact and treatment outcomes 21

1.2.1 The incidence and impact of breast cancer 21
1.2.2 Breast cancer subtypes 22
1.2.3 Triple negative breast cancers (TNBCs) 23

1.3 Breast cancer metastasis 25

1.3.1 Metastasis and EMT are associated with negative outcome in breast cancer ... 25
1.3.2 Known cellular changes that drive breast cancer metastasis 29
1.3.3 Current methods to predict breast cancer outcome 30
1.3.4 Metastasis is driven by many complex and interacting molecular processes 31

1.4 Regulation and diversification of the human transcriptome 32

1.4.1 The canonical model of transcription initiation and regulation 32
1.4.2 Alternative splicing (AS) of exons diversifies the transcriptome and is dysregulated in cancer 34
1.4.3 Cleavage and polyadenylation (CPA) 35
1.4.4 Alternative polyadenylation (APA) further diversifies the transcriptome 36
1.4.5 The role of the poly(A) tail in translational control 39
1.4.6 MicroRNA (miRNA)-mediated regulation mRNA transcripts 40

1.5 APA in proliferation and disease 42

1.5.1 APA and viruses 42

1.5.2 APA in neurodegenerative diseases 44
1.5.3 APA in cardiac diseases 45
1.5.4 The CPA machinery and APA in proliferation 46
1.5.5 APA in cancer 47
1.5.6 APA in breast cancer 48

1.6 Themes and trends in the study of APA 49

1.6.1 Distal APA events may also increase mRNA stability 49
1.6.2 APA events are diverse and specific 51
1.6.3 Biases in methods used in previous APA research 51
1.6.4 The potential of future APA research 53

1.7 Technologies for measuring the post-transcriptional state of cells 55

1.7.1 Gene expression microarrays 55
1.7.2 RNA-Seq 57
1.7.3 PAT-seq 59
1.7.4 Alternative 3’-focused sequencing approaches for measuring APA and poly(A) tail length 62
1.7.5 APADB: A database of known APA sites in humans 63

1.8 Aims of this thesis 65

1.9 Completion of aims and summary of results 67

Chapter 2: Materials and methods 69

2.1 Assessing RNA metabolism in an increasingly metastatic mouse xenograft model of TNBC 70

2.1.1 Generation of a mouse xenograft breast cancer model 70
2.1.2 PAT-seq and TCGA gene expression processing and gene set enrichment analysis 72
2.1.3 Cell culture and RNA extraction from the MXM cell lines 73
2.1.4 Testing for miRNAs that were lost during metastatic transformation 73
2.1.5 The ePAT and TVN-PAT methods 74
2.1.6 The Extension Poly(A) Test (ePAT) 75
2.1.7 T-12 Variable Nucleotide Poly(A) Test (TVN-PAT) 76
2.1.8 PCR amplification of ePAT and TVN-PAT cDNA 77
2.1.9 ePAT and TVN-PAT gels 77
2.1.10 The mPAT and alPAT methods 77
2.1.11 The mPAT method 78
2.1.12 Primer concentrations and PCR cycles tested during mPAT optimisation 80
2.1.13 The alPAT method 81

2.2 Inferring and analysing APA from public datasets 84

2.2.1 Obtaining TCGA data 84
2.2.2 Counting reads mapping to polyadenylation sites in TCGA RNA-Seq data 84
2.2.3 Limiting the false positive error rate when inferring APA from TCGA data 85
2.2.4 Inferring APA from paired TCGA breast cancer RNA-Seq data 85
2.2.5 Panther overrepresentation analysis 87
2.2.6 Grouping samples by gene expression and APA t-SNE 87
2.2.7 Survival analysis using all TCGA tumours 87
2.2.8 Generation of a complete clinical, APA and gene expression predictive model 88
2.2.9 Testing the reproducibility of the complete model using bootstrapping 88
2.2.10 Calling APA from microarrays using the original probeset ordering method 88

2.3 Novel bioinformatics methods 90

2.3.1 Components of *3Primer* 90
2.3.2 RNA Systems Explorer App code base and implementation 91

Chapter 3: Changes to RNA metabolism in an increasingly metastatic model of TNBC 93

3.1 Introduction 94

3.2 Results 97

3.2.1 PAT-seq analysis of a mouse xenograft model of TNBC metastasis 97
3.2.2 Dynamic APA events are associated with metastasis 103
3.2.3 APA is not associated with gene expression or poly(A) tail length 107
3.2.4 The 3’ UTR binding sites and expression of miRNAs was altered with metastatic potential 107
3.2.5 RNA processing was increased and immune signalling was decreased with metastatic potential 111
3.2.6 The expression of mRNA 3’ processing factors was increased with metastatic potential 114
3.2.7 The MXM provides a controlled model for the study of TNBC metastasis 116
3.2.8 Minor transcriptome wide poly(A) tail length changes were observed with increasing metastatic potential 119
Gene specific poly(A) tail length showed no association with gene expression 121
3.2.9 Assessing the reliability of PAT-seq poly(A) tail length measurement 123
3.2.10 Investigation of poly(A) tail length changes by alternative methods 126
3.2.11 Optimisation of the mPAT method for low abundance mammalian genes 129

3.3 Discussion 135

3.3.1 PAT-seq accurately measured APA and gene expression in the MXM 135
3.3.2 APA events with increasing metastasis in TNBCs 137
3.3.3 Metastasis-associated APA selects for loss of miR-22-3p binding sites 140
3.3.4 Poly(A) tail length measurement by PAT-seq in the MXM 141
3.3.5 Gene expression changes with increasing metastasis in TNBCs 143

Chapter 4: Inferring APA events from public databases and predicting breast cancer prognosis 146

4.1 Introduction 147

4.2 Results 154

4.2.1 Inferring APA from TCGA RNA-Seq 154
4.2.2 Proximal APA is pervasive in primary human breast tumours from the TCGA 156
4.2.3 Inferring APA from Affymetrix HG-U133A microarrays without using a custom CDF file 159
4.2.4 Proximal APA is pervasive in breast cancer APA events inferred from Affymetrix HG-U133A microarrays 163
4.2.5 There is minimal agreement between competing methods when determining breast cancer associated APA events 163
4.2.6 Comparing TCGA and GEO datasets with the topconfects method yields 100 high confidence breast cancer associated APA events 166
4.2.7 TNBCs undergo more pronounced APA than ER+ breast cancers 168
was a switch to both proximal and distal sites. These results suggested that the general trend toward proximal APA does not vary with breast cancer subtype, however, the strength of APA changes present in a sample may be influenced by the state of the tumour. 173
4.2.8 An overall APA trend does not impact breast cancer outcome 173
4.2.9 Tumours from the TCGA do not always have a distinct pattern of APA when compared to normal breast tissue 176
4.2.10 APA events are not subtype-specific 178
4.2.11 Assessing miRNA expression with metastasis and subtype in the TCGA 180
4.2.12 Predicting tumour outcome with clinical data gene expression and APA 180
4.2.13 Clinical data and APA make a significantly better prognostic model than clinical data alone 188
4.2.14 Assessing the reproducibility of predictor selection using clinical data, APA and gene expression 188
4.2.15 Assessing the power of the clinical and APA model in different breast cancer subtypes 190
4.2.16 Refining the clinical and APA prognostic model for potential use in an mPAT-based prognostic test 192

4.3 Discussion 197

4.3.1 The effectiveness of the database counting (DBC) method for calling APA from RNA-Seq 197
4.3.2 The effectiveness of the original probeset ordering (OPO) method for calling APA from microarrays 198
4.3.3 Proximal APA was the predominant shift in breast cancer 200
4.3.4 The methods and datasets used to call APA can have a dramatic effect on the APA events called 201
4.3.5 Deregulation of miRNA expression in breast cancer was not subtype-specific202

4.3.6 Evaluation of prognostic modelling of breast cancer outcome in the TCGA 203
4.3.7 APA events should be considered at the gene-specific level moving forward .205

Chapter 5: Computational tools for the design and analysis of custom NGS experiments 207

5.1 Introduction........208

5.2 Results........211

5.2.1 *The RNA Systems Explorer* App (*RSER*) 211
5.2.2 *RSER* tab 1: Experiment, sample, gene selection and genomic coverage 212
5.2.3 *RSER* tab 2: Assessing raw read counts for quality control........214
5.2.4 *RSER* tab 3: The poly(A) tail reverse cumulative distribution plot........216
5.2.5 *RSER* tab 4: The read length reverse cumulative distribution plot........216
5.2.6 *RSER* tab 5: The summary statistics page 219
5.2.7 *3Primer* an automated primer design tool for 3' RACE experiments........219

5.3 Discussion 224

5.3.1 Benefits and limitations to the use of the *RSER* app 224
5.3.2 Benefits and limitations to the use of *3Primer*........225
5.3.3 The importance of accessibility of statistical analysis tools........225

Chapter 6: Conclusion 227

References........ 231

Appendix A1 - mPAT and alPAT gene specific forward primer lists 269

Appendix A2 - Paul Harrison's 3' End shift test description........ 279

Appendix A3 - Full list of APA events from paired TCGA analysis 280

Appendix A4 - High confidence APA events from TCGA and GEO microarray datasets 301

Appendix A5 - Alternative polyadenylation in the regulation and dysregulation of gene expression........ 307

Abstract

With the exception of skin cancers, breast cancer is the most common cancer affecting women. Breast cancer is commonly treated through the use of chemotherapy and hormonal therapies. Progress has been made in breast cancer detection and treatment resulting in increased survival, however, determining how likely a primary tumour will be to metastasise remains a challenge, especially in the poorly understood triple negative breast cancer (TNBC) subtype. Alternative polyadenylation (APA), has been suggested as a possible novel biomarker for the prediction of breast cancer prognosis. Poly(A) Test sequencing (PAT-seq) was used in this thesis to measure gene expression, APA and poly(A) tail length changes in primary breast tumours with increasing metastatic potential in a mouse xenograft model. Many gene expression pathways, altered with metastatic potential in this model, were associated with RNA processing, suggesting that altered RNA metabolism is key to TNBC metastasis. Relative to a lowly invasive (LNA) tumour baseline, there were 37 (31 distal, 6 proximal) metastasis-associated changes to the tumour line that targets the lung (LM2), and 47 (19 distal, 28 proximal) in the highly metastatic (HM) line. This showed that there was no clear APA trend with metastatic potential in this model. APA was also studied in primary human tumours by inferring APA events from microarrays from the Gene Expression Omnibus (GEO) and in RNA-Seq from The Cancer Genome Atlas (TCGA). Over 1,700 APA events were identified, with 100 overlapping APA events in both datasets. APA events from the TCGA were used with clinical and gene expression data to create a prognostic linear model. Interestingly, of any combination of APA, gene expression and clinical data, the model of APA + clinical data was the best predictor of breast cancer outcome and was significantly more prognostic than clinical data alone ($p < 0.05$). This model was prognostic across multiple breast cancer subtypes, including TNBCs, potentially representing a novel prognostic test for breast cancer survival. This work represents a thorough analysis of APA in breast cancer and

presents hundreds of novel breast cancer-associated APA events, some of which are prognostic of breast cancer outcome.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Andrew Patt

Andrew David Pattison

18 July 2018

Preface

-Dr Cameron Johnstone generated the mouse xenograft model including providing cell lines for subsequent *in vitro* culture and extracted the RNA that was subsequently used in the PAT-seq experiments presented here.

-PAT-seq experiments presented here were performed by Dr Traude Beilharz (Batch 1) and Melissa Curtis (Batch 2).

-The *3' end shift method*, and the t*ail-tools* and *topconfects* software packages were developed by Dr Paul Harrison.

-The RNA systems explorer app presented in Chapter 5 was also contributed to by Dr Paul Harrison, Jack Xu and Michael See.

-All other new data presented here comprises of my own original work.

Some of the work undertaken toward the completion of this degree has been published in the papers listed below:

Turner, R. E.*, **Pattison, A. D.*** & Beilharz, T. H. (2017, August). Alternative polyadenylation in the regulation and dysregulation of gene expression.

Seminars in cell & developmental biology.

* Paper was co-first authored.

Johnstone, C.N., **Pattison, A. D.**, Gorringe, K.L., Harrison, P.F., Powell, D.R., Lock, P., Baloyan, D., Ernst, M., Stewart, A.G., Beilharz, T.H. & Anderson, R.L. (2018, April). Functional and genomic characterization of a xenograft model system for the study of metastasis in triple-negative breast cancer.
Disease Models & Mechanisms.

Harrison, P.F., **Pattison, A. D.**, Powell, D.R. & Beilharz, T.H. (2018, June). Topconfects: a package for confident effect sizes in differential expression analysis provides improved usability ranking genes of interest.
BioRxiv **(preprint)**

Boag, P.R, Paul F. Harrison, P.F., Barugahare A.A., **Pattison, A. D.**, Swaminathan, A., Monk, S., Davis, G.M., Heinz, E., Powell, D.R. & Beilharz, T.H. A tail of two transcriptomes: the landscape of cytoplasmic polyadenylation in the C. elegans germline.
In preparation

A first author manuscript entitled: “Alternative polyadenylation is pervasive and prognostic of outcome across multiple breast cancer subtypes” is also currently in preparation based on the work presented in Chapter 4 of this thesis.

Acknowledgements

Firstly, I would like to thank Dr Traude Beilharz for taking me on and giving me a chance to challenge myself in her lab. She has given me the freedom to explore some very cool research in my own way. Being in her lab and taking advantage of the opportunities that she has afforded me has allowed me to grow both personally and professionally. I was also very fortunate that she allowed, and often actively encouraged me, to attend numerous conferences and training sessions throughout the course of my PhD. This experience and training has helped me to become a much more confident and competent researcher. I am hugely thankful for my time in her lab and am certain that I have received a quality education in how to do cutting edge research, the likes of which is available in few other places.

Secondly, I must thank Dr Paul Harrison for all that he has taught me about bioinformatics. Paul was always there for me to ask questions and was able to help with almost any computational problem that I might have. He always challenged me to use the best methods in my research and even wrote his own when no other satisfactory methods existed. The passion and professionalism that he demonstrates toward his work has been inspiring and I feel very privileged to have been supervised by him.

I would like to thank all the members of the Beilharz Lab, past and present, for creating a fun, supportive environment in which to do research. Lab lunches and conferences with members of the lab were always enjoyable and I value the friendships that I have built with all members of the lab. Thanks especially to Angs for always knowing where everything was, and Rachael for pointing out all my spelling mistakes.

I would like to thank to Dr Cameron Johnstone for his work in generating the mouse model studied in this thesis, as well as his support and advice throughout. I would also like to thank him for his help in setting up my cell culture experiments.

I would like to thank the external members of my advisory panel: A/Prof. Lee Wong, A/Prof. David Powell and Dr Peter Boag for all their support and advice throughout my PhD. They were always there to provide help and support whenever it was required.

I would also like to thank the Monash Bioinformatics Platform for letting me drop into their office for a chat, as well as use the eSolutions coffee machine.

A big thank you must also to Dr Belinda Goldie, Dr Paul Harrison and Dr Traude Beilharz for providing me with feedback on this thesis. I appreciate you taking the time to read it, and it has been vastly improved by your suggestions.

Finally, I would like to thank my family and friends for their support. I would especially like to thank my girlfriend Emma who has been incredibly understanding and supportive as I have completed the writing of this thesis.

This research was supported by Monash Departmental Scholarship.

Definitions and Abbreviations

3' RACE	3' Rapid amplification of cDNA ends (Oligo-dT and PCR-based)
Adjuvant chemotherapy	Chemotherapy in addition to surgery
alPAT	An adapter ligation method for quantifying RNA expression state.
APA	Alternative polyadenylation
AS	Alternative splicing
APA state	The relative APA site usage between two or more APA sites
Athymic	Lacking a thymus (athymic mice will have no T cells)
BED file	Browser extensible data file (describes a collection of locations in a genome)
bp	Base pairs
CAMERA	Correlation adjusted mean rank gene set test
CDF	Chip definition file
CSV	Comma-separated values
DBC	Database counting method
DCIS	Ductal carcinoma in situ
dTTPs	Deoxythymidine triphosphates
dUTPs	Deoxyuridine triphosphate
DSE	Downstream sequence element
EMT	Epithelial to mesenchymal transition
ECM	Extracellular matrix
ENLM	Elastic net linear modelling
ER	Estrogen receptor
FC	Fold change
FACS	fluorescence activated cell sorting

FDR	False discovery rate
GEO	Gene Expression Omnibus
GSEA	Gene set enrichment analysis
IGV	Integrative Genomics Viewer
HER2	Human epidermal growth factor receptor 2
IPSCs	Induced pluripotent stem cells
IR	Ischemia/reperfusion
Klenow polymerase	Extends RNA on a DNA template
LIMMA	Linear Models for Microarray and RNA-Seq Data
MDS	Multidimensional scaling
mPAT	Multiplex 3' RACE Illumina based sequencing method
MFP	Mammary fat pad
MM	Mismatch (microarray probe)
MSigDB	Molecular Signatures Database
MXM	Mouse xenograft model
NSG mice	NOD scid gamma mice that lack B, T, and NK cells
nt	Nucleotides
Oligo-dT	A short sequence of deoxy-thymidine nucleotides
PAL-seq	Poly(A)-tail length profiling by sequencing
PDUI	Percentage distal usage index (percentage distal APA site usage)
PM	Perfect match (microarray probe)
OPO	Original probeset ordering
PR	Progesterone receptor
Pol II	RNA polymerase II
RNA-seq	Broadly, sequencing of the transcriptome of a sample

RNAse T1	Cleaves RNA after G bases
USE	Upstream sequence element
UTR	Untranslated region
Tail-tools	Software by Paul Harrison for the analysis of PAT-seq data
TCGA	The Cancer Genome Atlas
TE	Translational efficiency
TNBC	Triple negative breast cancer
TPM	Transcripts per million
Urea PAGE	Denaturing Urea Polyacrylamide Gel Electrophoresis

List of figures

Figure 1.1. The canonical process of metastasis in breast cancer	26
Figure 1.2. The changes tumour cells undergo during EMT	28
Figure 1.3. The central dogma of molecular biology as it was understood in 1970	33
Figure 1.4. Transcription and alternative polyadenylation	37
Figure 1.5. APA genes in varying cancer and disease states	52
Figure 1.6. Schematic describing a single-channel gene expression microarray	56
Figure 1.7. Schematic describing short-read RNA-Seq	58
Figure 1.8. PAT-seq	60
Figure 2.1. The mPAT method	79
Figure 2.2. The alPAT method	82
Figure 2.3. Steps in defining and analysing APA from TCGA data	86
Figure 2.4. Schematic of the functions of 3Primer	89
Figure 2.5. The structure of a shiny app	92
Figure 3.1. TNBC xenograft tumour model	98
Figure 3.2. Detection of batch effects in PAT-seq data	100
Figure 3.3. PAT-seq identification of APA sites	102
Figure 3.4. The general trend of APA events in PAT-seq run 2	104
Figure 3.5. APA from both MXM PAT-seq runs shows variable APA trends with increasing metastatic potential	106
Figure 3.6. APA is not correlated with TL or DGE in the MXM	108
Figure 3.7. The miRNAs altered between the NI and HM lines in vitro	110
Figure 3.8. Barcode plots of the top 6 canonical pathways enriched in run 2 of the MXM	115
Figure 3.9. AS and APA factor expression is associated with metastatic potential in the MXM and TNBCs	117
Figure 3.10. PAT-seq and TCGA gene expression are consistent overall but not with metastatic potential	118
Figure 3.11. PAT-seq and TCGA gene set enrichment comparison shows broadly consistent changes with metastatic potential	120
Figure 3.12. PAT-seq gene expression and poly(A) tail length comparison	122
Figure 3.13. Variance in poly(A) tail length measurement	124
Figure 3.14. PAT-seq poly(A) tail length normalisation A	125
Figure 3.15. Tail lengths measured by mPAT	128
Figure 3.16. Measurement of MXM tail lengths and gene expression by mPAT	130
Figure 3.17. Attempting to replicate a GLUL poly(A) tail length shift	131
Figure 3.18. Optimisation of the mPAT method for lower abundance genes	133
Figure 4.1. The process of inferring APA from TCGA breast cancers	155
Figure 4.2. Significant alternative polyadenylation events from the DBC-matched tumour/normal analysis	157
Figure 4.3. Example of grouped HG-U133A probesets that bind to a single gene	160
Figure 4.4. Inferring APA from HG-U133A by ordering standard probesets	162
Figure 4.5. Breast cancer associated APA events as called from HG-U133A microarrays using the OPO method	164

Figure 4.6. Comparisons of the DBC-matched method with APA from the OPO method and two papers that inferred APA from the same datasets	165
Figure 4.7. Comparison of the unmatched APA events from the GEO and matched APA events inferred from the TCGA	167
Figure 4.8. Example coverage of high confidence APA events	169
Figure 4.9. Alternative polyadenylation in breast cancer as determined by TCGA RNA-Seq reanalysis	171
Figure 4.10. Alternative polyadenylation in breast cancer as determined by microarray reanalysis	172
Figure 4.11. Analysis of an overall tumour APA effect	175
Figure 4.12. T-SNE distributions of both tumour and normal samples generated using both gene expression and APA	177
Figure 4.13. T-SNE distributions of tumour samples generated using both gene expression and APA	179
Figure 4.14. Expression of miRNAs in the TCGA	181
Figure 4.15. Inferring good or poor prognosis using APA and clinical variables	183
Figure 4.16. Elastic net linear modelling prognostic scores generated using gene expression, APA and clinical data	187
Figure 4.17. Bootstrapping of clinical, APA and gene expression elastic net linear modelling	189
Figure 4.18. APA + clinical prognostic score within 5 intrinsic breast cancer subtypes	191
Figure 4.19. APA + clinical prognostic score within histological subtypes	193
Figure 4.20. The coefficients to be used in a targeted APA based test	194
Figure 5.1. RSER genome coverage and options tab	213
Figure 5.2. RSER raw reads count tab	215
Figure 5.3. RSER poly(A) tail inverse cumulative distribution tab	217
Figure 5.4. RSER read length cumulative distribution tab	218
Figure 5.5. RSER read length cumulative distribution tab	220
Figure 5.6. Specific primer design for 3’ RACE experiments	222
Figure 5.7. The output of the 3Primer primer design program compared to primers designed by a human	223

List of tables

Table 1.1. Differences between single and multi-3' UTR genes	39
Table 1.2. Known, disease associated APA events from the literature	43
Table 2.1. Tags added to MDA-MB-231 derived cell lines used in the MXM	71
Table 2.2. Optimising mPAT primer concentrations	81
Table 2.3. Optimising mPAT PCR cycle number	81
Table 2.4. Sequences of the primers used in the 'PAT' type methods	83
Table 3.1. The miRNA binding sites lost during metastatic transformation	111
Table 3.2. Enriched metastasis-associated GO terms	112
Table 3.3. Enriched metastasis-associated curated pathways	113
Table 4.1. The five methods used to infer APA from RNA-Seq and microarray datasets discussed in this chapter	154
Table 4.2. Top TCGA APA genes	174
Table 4.3. The contribution of clinical predictors to ENLM	185
Table 4.4. Predictor variables and coefficients used in the mPAT model	195

Chapter 1: Literature review and introduction

1.1 Topics discussed in this literature review

This literature review will discuss breast cancer, the prediction of breast cancer outcome and the regulation of gene expression at both the transcriptional and translational levels. The chapter begins by reviewing the literature surrounding breast cancer including prevalence, impacts and outcome. It features an in-depth focus on metastasis, as this is the key process associated with poor breast cancer outcome. Strengths and weaknesses of current methods to predict breast cancer metastatic potential are discussed, as is the need for better prognostic tests that potentially utilise novel markers. The processes of transcription and translation are then explained in depth to give the necessary background on where both existing and novel breast cancer markers are derived from. There is a heavy focus in this thesis on the role alternative polyadenylation (APA) in breast cancer. APA is suggested in this chapter as an important factor in the regulation of translation, however, the study of APA in disease has only recently come into prominence. To provide a more complete understanding of the dysregulation of APA, it is discussed in the context of disease more broadly including in cancer and proliferation. Themes and trends in the field of APA research are then explained in detail. Next, the methods utilised to measure the transcriptional state of cells in this thesis are introduced. Finally, the aims of this thesis and in the chapters in which they are addressed is outlined.

1.2 Breast cancer prevalence, subtypes, impact and treatment outcomes

Breast cancer is a common, highly heterogeneous disease, with each cancer largely unique to the patient from which it was derived [1]. Due to this fact, breast cancer is often better considered as a related group of diseases, rather than a single disease. Breast cancers are, therefore, often grouped into subtypes based on their treatable characteristics or their gene expression profile. This grouping may be used for prognostic purposes and assists in the determination of the best course of treatment. Breast cancer prognosis has improved over time and is generally favourable, especially when detected early [2]. The challenge for clinicians now is to differentiate patients with tumours that do not require treatment from those that do and, in the process, also identify high-risk patients that are unlikely to respond to current therapies.

1.2.1 The incidence and impact of breast cancer

Breast cancer affects 1.67 million people worldwide each year, the majority of whom are women. It is the second most common cancer type and accounts for 6.4% of cancer-related deaths [3]. Depending on ethnicity, breast cancer occurs in as many as 128 people per 100,000 and has a mortality rate ranging from 11.3-29.5% [2]. Increased mortality rates are observed in those with lower socio-economic status [4], which also explains a large portion of ethnicity-related differences [5]. In recent times, breast cancer mortality rates have declined due to the increased prevalence and application of screening, better practices of tumour characterisation and better treatments [6]. The most pressing challenges for clinicians now, are to identify patients with tumours that do not require treatment and those that are unlikely to respond to current therapies. Also of great benefit to the patient would be additional information on whether more aggressive treatments (such as chemotherapy) are necessary as well as an indication of expected survival if no effective treatments exist.

1.2.2 Breast cancer subtypes

Many breast cancers have similar histological markers or similar patterns of gene expression which can be utilised to group them into subtypes. Breast cancer subtypes tend to have similar clinical behaviours and predictable responses to therapy [7]. Breast cancer subtypes are commonly defined based on histological markers, namely the presence of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) [8, 9], with subtypes assigned based on a combination of these markers. Approximately 80% of breast cancers overexpress the ER [10] and 20-30% have an overexpression of HER2 [11]. Information about breast cancer subtype has proven invaluable in guiding treatment decisions as therapies can be targeted toward the drivers of the tumour [12]. ER+ breast cancers have a vastly increased number of ERs on their cell surface that, when bound by estrogen, activate a signalling cascade that ultimately leads to the activation of transcription factors, such as *FOXA1*, that are associated with cell proliferation and survival [13, 14]. The HER2 (*ERBB2*) gene also encodes a cell surface receptor that interacts with the Ras, Rac and PI3K pathways to decrease the expression of pro-apoptotic transcription factors and promote survival [15]. When tumours are positive for the ER and HER2 receptors, targeted endocrine therapies can be given either as the primary method of treatment or alongside chemotherapy [12]. The two prominent examples of these treatments include tamoxifen, an ER antagonist for ER+ breast cancer [16] and trastuzumab [17], a monoclonal antibody that targets tumours overexpressing the HER2 receptor. The expression of the PR receptor is often also assayed, as the presence of the receptor is indicative of a functionally intact estrogen-mediated signalling pathway [18] and is associated with better response to endocrine therapy in ER+ patients [19].

Breast cancer subtypes have also been defined at the molecular level, with gene expression microarray studies suggesting 5 ‘intrinsic subtypes’, each with its own clinical profile and pathogenicity [20]. These subtypes have been helpful in understanding the exact site of breast cancer origin (basal, luminal etc.) and tend to be associated with specific driver mutations [1]. Unfortunately, these groupings do not represent the full spectrum of breast cancer heterogeneity as these subtypes alone are not completely prognostic of breast cancer outcome and multiple sub-subtypes have also been defined [21, 22, 23, 24]. Patient ethnicity has also been shown to influence the breast cancer subtype a patient is likely to have [2]. This may be due to slightly different genetics, but may also be due to socio-economic factors such as access to health care and obesity [25]. In future, tumour classification at the patient-specific level will likely be required before breast cancer outcome can be completely predicted and the most appropriate treatment chosen with certainty.

1.2.3 Triple negative breast cancers (TNBCs)

TNBCs are named as such because they lack the expression of ER, PR and HER2 markers at levels detectable by immunohistochemical staining [26]. TNBC accounts for as low as 16% of breast cancer cases in Caucasian women under 50 and up to as many as 39% of cases in African-American women in the same age group [27]. TNBC is a highly heterogeneous subtype with at least 6 TNBC sub-subtypes previously identified [22]. Earlier detection of breast cancers, including TNBCs, through the uptake of screening has caused a decrease in disease mortality, however, it has also caused an increase in the detection of tumours that would never have gone on to pose a threat to the health of the patient [4]. This early detection is of particular issue for TNBC patients because chemotherapy is given as the standard and only treatment [28]. While TNBC prognosis with chemotherapy is better than that of other breast cancers [12], there is no way to tell if this cytotoxic treatment was ever required.

Conversely, in TNBC patients where a complete response to treatment is not obtained, higher rates of relapse and lower rates of survival are also observed [29]. Therefore, the most pressing challenges for the treatment of TNBCs are predicting the likely outcome of a primary tumour and developing less cytotoxic treatments that effectively target the heterogeneous group of tumours that make up this subtype.

1.3 Breast cancer metastasis

Metastasis refers to the process whereby tumour cells migrate from the primary tumour site to a different tissue or organ (Figure 1.1). The first step in the classical model of metastasis is the growth of the tumour and invasion of local breast tissue, before intravasation to the blood vessels or lymphatic system [30]. To do this, tumour cells must undergo changes that cause them to lose their static, interconnected form and split off into single cells. These single cells must be able to survive in the vasculature or lymphatic system, including evading the host immune system and migrate to distant tissues or organs. It has historically been thought that these changes occur through epithelial-mesenchymal transition (EMT), whereby a cell loses epithelial markers and takes on a more mesenchymal state. Under this model, once migration is complete, a tumour cell must reacquire some of its previous adherent characteristics in order to form a new tumour at this distant site. This is the reverse process, known as mesenchymal-epithelial transition (MET) [31]. More recently, it has been suggested that EMT is not always required for metastasis and other models of metastatic progression have also been suggested. Once such example is tumour cells taking on a more ameboid form and squeezing their way through the vasculature [32].

1.3.1 Metastasis and EMT are associated with negative outcome in breast cancer

Breast cancer metastasis (and metastasis in cancer more generally) has long been known as a clear indicator of poor prognosis [33]. The median survival time following a metastatic event in breast cancer is 20 months, with < 20% of patients surviving longer than 5 years following an eventual relapse and a 15% survival rate overall [34, 35]. Despite advances in chemotherapy in the past 30 years, outcomes have not greatly improved for patients with metastatic disease [34]. Associated with metastasis is breast cancer recurrence (following removal of the original tumour), which can refer to the appearance of a distant metastatic

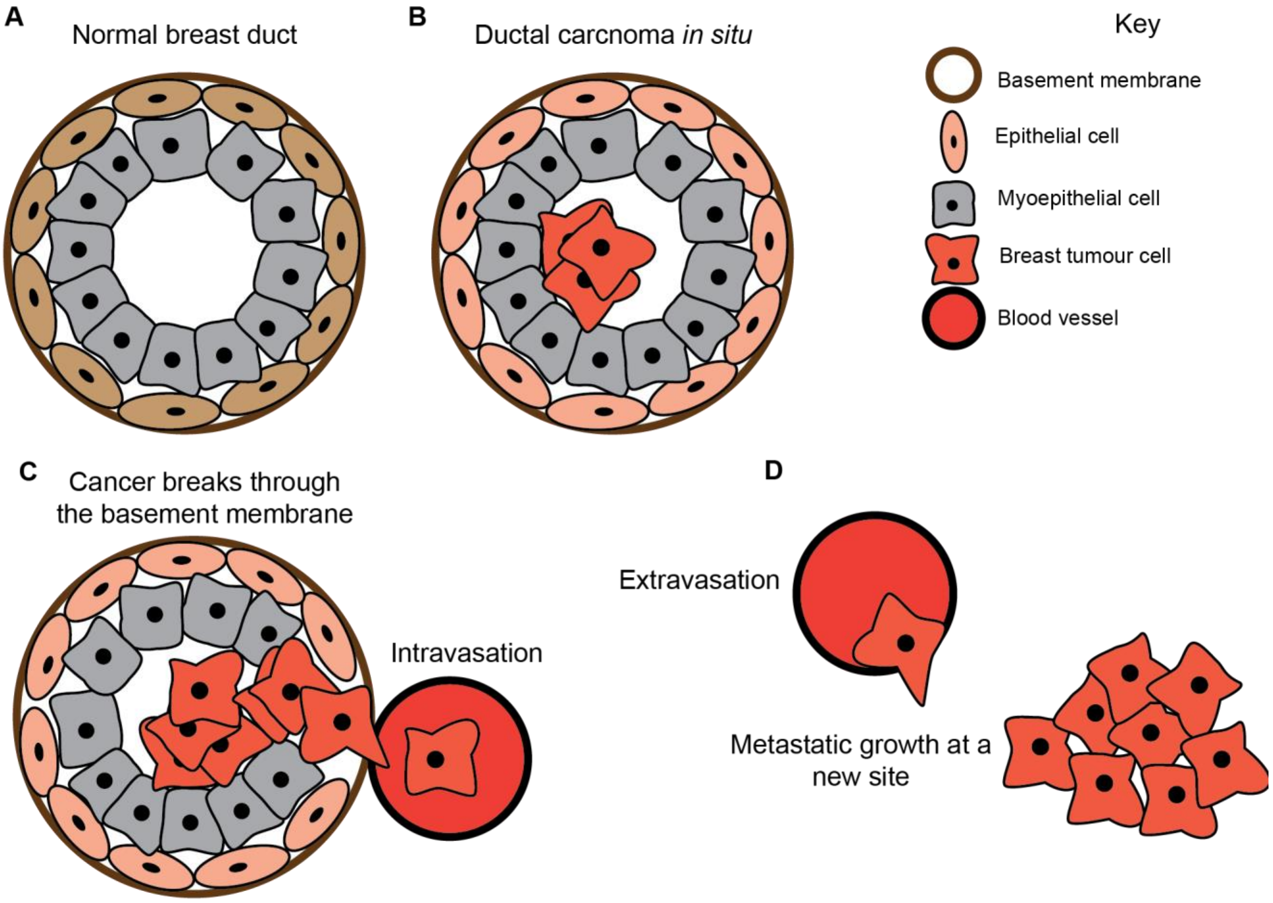


Figure 1.1. The canonical process of metastasis in breast cancer. A. A normal healthy breast duct. **B.** Ductal carcinoma *in situ* which refers to a tumour that has not yet left the breast. **C.** The cancer breaks through the basement membrane and intravasates to the blood stream or lymphatic system. **D.** The tumour extravasates from the blood stream or lymphatic system and initiates growth of a secondary tumour at a distant site.

lesion in locations such as the brain, bone and lungs or a locoregional relapse which occurs in the breast, lymph nodes in the breast region or the chest wall. Locoregional relapse occurs in ~10% of breast cancer patients and is associated with concurrent metastasis in 30% of cases, with distant metastasis occurring in the future in 60% of cases [36].

The first step in the classical metastasis pathway is the invasion of surrounding host tissue by the tumour [37]. To achieve this, tumour cells start to downregulate E-cadherin (*CDH1*), which is associated with the maintenance of tight cell-cell junctions [38]. The downregulation of E-cadherin is also the first step in EMT (Figure 1.2). EMT describes a process whereby cells lose their epithelial state, including loss of apical-basal cell polarity, loss of cell-cell adhesion and reduced E-cadherin expression [31]. These cells then adopt more mesenchymal markers including undergoing cytoskeletal changes, upregulation of N-cadherin (*CDH2*) expression, changes to miRNA expression, increased metabolic rate and the production of enzymes that degrade the extracellular matrix, making them more invasive and resistant to apoptosis [31, 39, 40]. Once cells have metastasised to an alternative location, they are able to adhere to the new metastatic site through the increased expression of N-cadherin [41] and potentially other members of the 110 genes in the cadherin family [41]. Once tumours are fixed at a new location, they may then undergo the reverse process of MET to begin developing into a new tumour [31, 40]. It should be noted that while aspects of the EMT model are required for tumour metastasis [43], the metastasis of epithelial cells has been observed [39], and there is yet to be conclusive evidence of this full process in patient tumours [35]. Due to the variable and complicated nature of the additional changes that occur during breast cancer invasion and metastasis, it may be preferable to target the underlying genetic mutations that drive metastasis rather than trying to stop metastasis itself. This approach is promising provided a tumour has not advanced sufficiently beyond reliance on

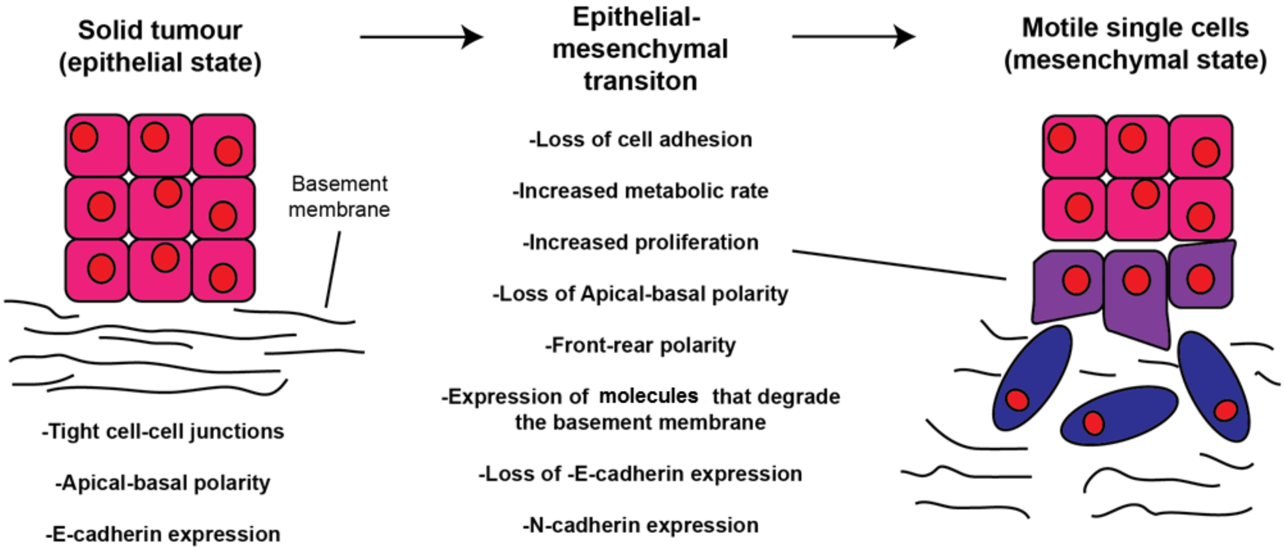


Figure 1.2. The changes tumour cells undergo during EMT. Depicted on the left panel is a more differentiated primary tumour *in situ.* Depicted on the right is the same tumour some cells undergoing EMT, losing their polarity and beginning to invade through the extracellular matrix (ECM) and the basement membrane.

these pathways.

1.3.2 Known cellular changes that drive breast cancer metastasis

Due to the high variability of breast cancers between different patients, which is often driven by different mutations, it is often better to consider breast cancer at the patient-specific level. For example, when the top mutational drivers of breast cancer were analysed by the TCGA consortium, only *GATA3*, *PIK3CA* and *TP53* were mutated in > 10% of cancers. Of these, the highest rate of mutation was in the known tumour suppressor *TP53* [44], which was significantly mutated in only 37% of tumours. Expression profiling by mRNA subtype [45] did, however, reveal that certain mutations were more likely to be associated with a certain subtype. For example, mutations in *PIK3CA* were observed in 47% of Luminal A type tumours and mutations in *TP53* were observed in 80% of basal-like (TNBC) tumours [1]. It has been shown that a metastatic tumour will generally have a similar mutational profile to the primary tumour, but will have also acquired additional mutations during the metastatic process [46, 36]. As a tumour cell disseminates it can acquire new mutations that increase metastatic ability, largely through the same mechanisms as the primary tumour, but potentially acquiring additional driver mutations such as disruptions to the JAK-STAT signalling pathway and the SWI/SNF chromatin remodelling complexes [36].

Even at the gene expression level, the drivers of metastasis are likely different depending on a given breast tumour. For example, EMT in breast cancer has some general drivers, such as the loss of E-cadherin expression, however, this driver alone can be induced by altering the expression of a host of different transcription factors including *SNAI2*, *SIP1*, *SNAI1* and *TWIST1* [47, 48, 49, 50]. There are also some general gene expression changes that may be commonly exploited by breast cancer in order to establish a secondary tumour. One prominent example is heparanase (*HPSE*) which is a key factor in the degradation of heparan sulphate and also plays a role in the remodelling of the ECM during breast cancer metastasis by cells with a more mesenchymal phenotype [51]. The increased expression of *HPSE* in breast cancer has been associated with lymph node metastasis, larger tumours, higher histological grade and poor prognosis [52]. As an alternative to the degradation of the ECM, tumour cells may also take on an ameboid formation, meaning that cells move through the ECM via the path of least resistance. This is achieved by using active myosin/actin contractions in a protease-independent manner [32]. This transition is known as mesenchymal-amoeboid transition (MAT) and has been suggested to be under the control of the Rho/ROCK signalling pathway [53], with expression of pathway members required for MAT to occur [54]. Breast cancers can, therefore, exploit a wide variety of known and unknown gene expression changes to successfully metastasise. This gene expression information has formed the basis of current breast cancer prognostic testing.

1.3.3 Current methods to predict breast cancer outcome

There are currently multiple prominent microarray-based genomic tests designed to add prognostic information about breast cancer outcome. These include PAM50 (a 50 gene signature), Oncotype DX (a 19 gene signature) and MammaPrint (a 70 gene signature) [44, 55, 56]. While these published signatures have all provided useful prognostic information to clinicians, there is generally little agreement between the genes used in these and other microarray-based signatures at large [57]. These discrepancies can be further exacerbated by technology specific biases and the heterogeneity of cancers [58]. It was suggested by Ein-Dor *et al.* that thousands of samples would be required before even a 50% overlap between gene signatures would be observed [59, 60]. It has, however, been noted that there is a broader agreement of the gene expression pathways selected in these signatures, including the estrogen signalling pathway and *BRCA1* associated pathways [61]. The concordance of these pathways has, however, been suggested to be limited to pathways associated with cellular proliferation [62, 63]. Novel prediction from other forms of translational regulators such as miRNAs [64] and APA [65] are therefore now being explored to add additional prognostic information for breast cancer prognosis.

1.3.4 Metastasis is driven by many complex and interacting molecular processes

In addition to gene expression, there are many known cellular alterations that may occur in a breast cancer primary tumour to drive metastasis including altered miRNA expression, APA and alternative splicing (AS) [64, 66, 60]. Often these processes represent the reactivation of developmental processes as a tumour cell de-differentiates. EMT, for example, is not unique to tumours and was first observed in the developing embryo [67]. Even poly(A) tail length has been suggested to be altered in the context of embryonic development [68]. These processes almost never act in isolation and will need to be understood as such before a complete model of breast cancer metastasis can be generated. The classical research into the mutational and gene expression changes that occur with breast cancer should, therefore, continue as should research into emerging mechanisms of cellular regulation to target breast cancer growth and metastasis on multiple levels.

1.4 Regulation and diversification of the human transcriptome

The central dogma of molecular biology states that a gene is transcribed from DNA into a messenger RNA (mRNA) molecule and then translated into a protein [69]. Even as these processes were beginning to be described, it was clear that this was not always a completely linear system, with non-standard information transfer occurring external to the central dogma (Figure 1.3) [69]. More recently, high throughput technologies have shown that for any given gene, there is a low correlation (usually ~0.25) between the relative amounts of mRNA and of protein present in a cell [70]. This section introduces mammalian transcriptional and post-transcriptional regulation and focuses mainly on the alterations to mRNA processing that are responsible for this discrepancy in humans.

1.4.1 The canonical model of transcription initiation and regulation

In eukaryotes, all mRNAs are synthesised from DNA templates through the process known as transcription. This process is performed by the transcriptional machinery, the core of which is RNA polymerase II (Pol II) [71]. To transcribe the mRNA, the transcriptional machinery must first gain access to the DNA, which is stored in the nucleus of the cell. When stored in the nucleus, DNA is compacted into chromatin, a complex comprised of an equal mix of protein and DNA, which is in turn comprised of repeating nucleosome units. A nucleosome is comprised of ~145–147 base pairs (bp) of DNA, wrapped around a histone octamer core [72]. For transcription to begin, Pol II-associated nucleosome modifiers, remodelers and molecular chaperones must destabilise the nucleosome and bind to the promoter sequence [73]. The promoter usually contains the ‘TATA box’, a repeated TATA sequence motif that is canonically used to signal the start of transcription [74]. The promoter site is also maintained in a more accessible state by the AT-rich nature of the promoter sequence itself, which is less able to wind around the histone octamer [75, 76], and nucleosome depletion complexes

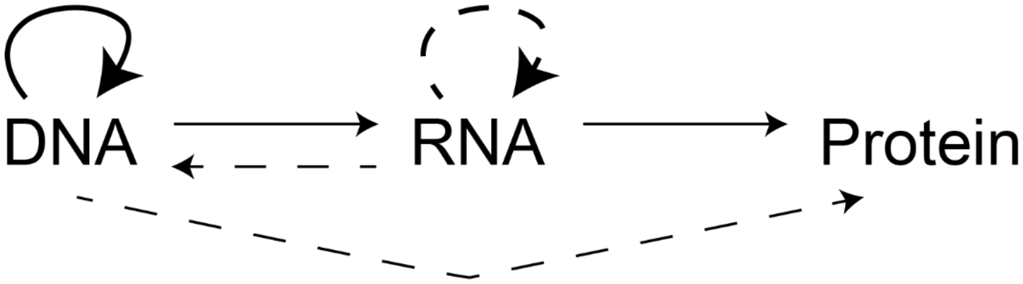


Figure 1.3. The central dogma of molecular biology as it was understood in 1970. Solid lines represent the normal transfer of information as part of the central dogma. Dotted lines represent non-standard transfers of information (systems that were known to interact outside the central dogma). Figure based on the paper by Crick [69].

such as the chromatin structure remodelling (RSC) complex [77]. During transcription initiation, Pol II forms a complex with the general transcription factors TFIIB, TFIID, TFIIE, TFIIF and TFIIH on double-stranded DNA to form the closed complex [78]. TFIID contains the TATA box-binding protein and associated factors [79] that recognise the TATA sequence motif, signalling the initiation of transcription and forming a transcriptional bubble 20-30 bp downstream of the TATA motif [80]. This is known as the open complex and allows Pol II to transcribe the single-stranded DNA. Changes in chromatin conformation have been associated with the regulation of transcription by restricting the access of Pol II to the DNA [81, 82, 83], and mechanistic links that describe these processes are beginning to be uncovered [84].

1.4.2 Alternative splicing (AS) of exons diversifies the transcriptome and is dysregulated in cancer

During the process of transcription, internal RNA sequences known as introns are spliced out of the nascent mRNA and a 5’ methylguanosine cap is added [71]. This leaves only the parts of the sequence that flank the introns, known as exons, which encode the functional protein and the 5’ and 3’ untranslated regions (UTRs) (discussed more in the next section) [85]. The process of splicing is carried out by the spliceosome, which is comprised of over 200 proteins, as well as five small nuclear RNAs (snRNAs) which form complexes with proteins to become small nuclear ribonucleo-proteins (snRNPs) [86]. The first step in splicing involves the binding of the U1 snRNP to the 5’ splice site of a pre-mRNA and recognition of the branch point by the U2 snRNP. The U5/U4/U6 tri-snRNP then joins the complex. The conformation of the spliceosome is altered, involving large rearrangements which cause the destabilisation or release of the U1 and U4 snRNPs. The spliceosome is then further structurally rearranged to form the catalytically active spliceosome, triggered by the DEAH-box protein DHX16 in an ATP dependent reaction [87]. Once the catalytically active spliceosome is formed, the pre-mRNA is cleaved at the 5’ splice site and a lariat-like structure is formed containing the 3’ exon and an intron. The second catalytic step is also ATP dependent and is promoted by another DEAH-box protein, DHX38 [88], which causes excision of the intron and the ligation of the 5’ and 3’ exons [89, 86].

It is also possible for different combinations of exons to be included in an mRNA molecule through AS [90]. AS has been suggested as the primary explanation for why the ~20,000 known human genes can encode ~100,000 different proteins [91]. Over 90% of human genes may be alternatively spliced, providing huge scope for regulation in all cellular contexts including cellular development and proliferation [92, 93]. Altered patterns of AS have also been demonstrated in tumours, with spliceosomal activity suggested to be essential for the function of the oncogenic transcription factor *MYC*, due to increased transcriptional rate and associated spliceosomal burden in *MYC* activated tumours [94]. More specifically, the expression of the splicing factor *SRSF1* has been shown to be upregulated in breast cancer [95]. The overexpression of *SRSF1* in normal-like MCF-10A cells was additionally shown to confer a tumour-like pattern of alternative splicing and was associated with delayed apoptosis and the acquisition of invasive properties [96].

1.4.3 Cleavage and polyadenylation (CPA)

To complete transcription and become a mature mRNA, a capped and spliced mRNA must undergo endonucleolytic cleavage, followed by the addition of multiple untemplated adenosine residues at the 3’ end, termed polyadenylation [97]. The process of cleavage and polyadenylation (CPA) is a crucial final step in the maturation of a transcript as it confers export competence, marking the mRNA to be transported from the nucleus to the cytoplasm for translation [98]. The polyadenylation site is co-transcriptionally selected based on the polyadenylation signal (PAS), which has the sequence AAUAAA or a close variant, and cis-regulatory elements (CREs) in the nascent mRNA that are centred around the PAS (Figure 1.4 A) [99]. These CREs include U-rich upstream sequence elements (USEs), often comprised of UGUA motifs, which are positioned 40-100 nucleotides upstream from the polyadenylation site [100]. Located within the 100 nucleotides immediately downstream of the cleavage site are U/GU-rich elements known as downstream sequence elements (DSEs) [101]. The PAS is recognised by a multimeric protein complex comprised of over 20 proteins known as the cleavage and polyadenylation specificity factor complex (CPSF), at the core of which is the CSTF64 protein and its homologue CSTF64T, which are key to PAS recognition in humans [102]. Other CPA factors that bind CREs include cleavage stimulation factor (CSTF) which binds to the DSEs and cleavage factor I (CFIM) which binds to the USEs [103, 104]. These factors then cause the binding of cleavage factor II (CFIIM), poly(A) polymerase (PAP), poly(A) binding protein nuclear 1 (PABPN1) and symplekin (SYMPK) to form the full complement of 3’ end processing machinery and cause CPA [105]. It is possible for multiple CPA sites to exist within the same mRNA molecule, with the process of differential PAS selection termed alternative polyadenylation (APA).

1.4.4 Alternative polyadenylation (APA) further diversifies the transcriptome

It has been determined that ~70% of all human mRNAs have at least two APA sites, with ~50% having three or more [106]. The majority of post-transcriptional regulation occurs at CREs encoded in mature mRNAs, which are acted upon by the *trans*-acting factors (TAFs) that bind to them. TAFs such as RNA binding proteins can affect mRNA stability, but can also have more complex effects such as moderating translational efficiency or protein localisation [107, 108, 109, 110]. Regulatory elements may bind anywhere in an mRNA molecule, however, the majority of CREs are located in the 3’ UTR. This has been suggested to be the

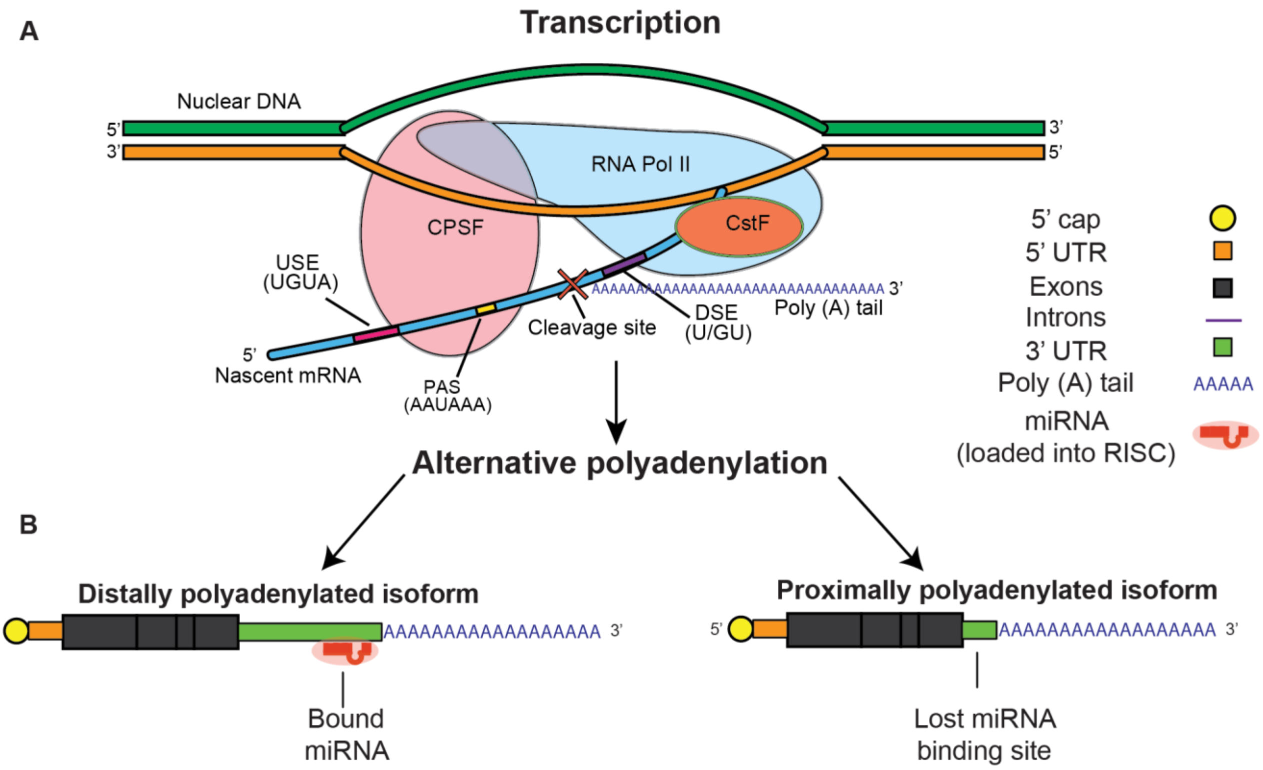


Figure 1.4. Transcription and alternative polyadenylation. **A.** The core factors in cleavage and polyadenylation. Cleavage and polyadenylation require the interaction of multiple *cis*-acting pre-mRNA sequence elements with multiple *trans*-acting multi-subunit protein complexes. **B.** Alternatively polyadenylated isoforms of the same mRNA with a bound and lost regulatory element (miRNA in this example).

case as alterations can be made to UTRs to increase the diversity of protein localisation and function through the binding of molecular chaperones without affecting protein sequence [111]. The selection of a poly(A) site proximal to the coding sequence in the 3' UTR of mRNA over one located more distally can result in the loss of the CREs, such as miRNA binding sites as depicted in Figure 1.4 B. While reduced mRNA regulation through the loss of miRNA binding sites has been the predominant area of study in the APA field [112, 109, 66, 113], there are many binding sites for mRNA stabilising proteins, such as AU-rich elements [114], which may also be lost during a shift toward proximal APA.

It has been suggested that APA disproportionately affects some genes over others [107]. Genes that do not undergo APA have a median length of ~600 nucleotides (nt), whereas multi-UTR genes have a median length of 2,300 nt [111] (Table 1.1). Also shown in Table 1.1 is that multi-UTR genes tend to have longer transcriptional units (encompassing a gene and sequence elements necessary for transcription). One study has suggested that there are also key differences in the patterns of tissue-specific expression of single and multi-UTR genes, with single-UTR genes more often expressed in a single tissue type and multi-UTR genes more often ubiquitously expressed [107]. The same study additionally reported higher evolutionary sequence conservation in multi-UTR genes that had longer UTRs. It has therefore been suggested that ubiquitously expressed genes are especially prone to regulation and diversification of protein function (often without altering the coding sequence) through their 3' UTRs [111]. This highlights the importance of APA in the regulation of gene function under altered cellular conditions and underscores the need for a more complete understanding of the factors that cause the selection of specific APA sites.

Table 1.1. Differences between single and multi-3’ UTR genes. Table adapted from the review by Mayr [111].

Descriptor	Single-UTR	Multi-UTR
3’ UTR length (nt)	625	2323
Transcription unit length (nt)	20629	40519
N expressed in 1 tissue	1,637 (23.7%)	630 (11.1%)
N expressed in 25 tissues	2,414 (34.9%)	1,854 (32.6%)
N expressed in 67 tissues	2,861 (41.4%)	3189 (56.2%)

1.4.5 The role of the poly(A) tail in translational control

One often overlooked mRNA control system is the modulation of mRNA expression by changing the length of the poly(A) tail [115]. Almost all mRNAs are polyadenylated co-transcriptionally with a nontemplated tail historically reported to be ~250-300 adenosine residues in length, (reviewed in ref [116]). This number has, however, been disputed by more recent high throughput studies, with one suggesting a median range of 50–100 nt across HeLa and NIH/3T3 cell lines [117], and another similarly suggesting a median range of 67-96 nt in HeLa, HEK293T and NIH/3T3 cell lines [68]. Typically, the poly(A) tail will be slowly de-adenylated throughout the life of the mRNA, however in some cases, such as activated or proliferating cells, cytoplasmic polyadenylation can occur, causing the poly(A) tail to be re-extended. Re-adenylation has been implicated in synaptic plasticity, including learning and memory in the brain, through the reactivation of mRNAs upon synaptic stimulation in dendrites [118]. Control of poly(A) tail length by the cytoplasmic polyadenylation element binding protein (CPEB) family members CPEB1 and CPEB4 has also been implicated as a checkpoint for proliferation and cell division in mitosis [119]. In the cancer setting, re-adenylation has also been suggested to be controlled by CPEB4 to reactivate silenced oncogenes [120].

While it is generally accepted that an mRNA must have a poly(A) tail > ~20 nt to be stably translated [121], the effect that the length of this tail has on translational efficiency (TE) is still unclear. The loss of nuclear poly(A) binding protein 1 (*PABPN1*) has been shown to lead to a shorter poly(A) tail and a decrease in cellular proliferation, suggesting that a longer tail improves TE [122]. Conversely, hyperadenylation in a nuclear context has been shown to cause the accumulation of mRNA transcripts in the nucleus, a lowering of protein expression [123] and targeting to exosomes for decay [124]. The length of the poly(A) tail has also been positively correlated with gene expression in general [109, 125, 126], although it has been suggested that this may only be the case strictly during embryonic development [68]. Challenging these findings, a recent study has suggested that short poly(A) tails are associated with highly expressed genes [127]. Adding additional uncertainty to studies of poly(A) tail length is the process whereby a short poly(A)-tail is often transiently added to mRNA decay intermediates prior to rapid exonucleolytic degradation [128], making measurements of tail-length more challenging to interpret. The exact role that poly(A) tail length plays in cellular regulation is therefore still unclear and requires further investigation.

1.4.6 MicroRNA (miRNA)-mediated regulation mRNA transcripts

MicroRNAs (miRNAs) are single-stranded RNAs ~23 nucleotides in length that post-transcriptionally regulate mRNA by translational repression through binding to complementary regions, usually in the 3’ UTR [129, 130]. A single miRNA may regulate hundreds of mRNAs by binding to them with varying degrees of specificity [131]. The miRNAs are transcribed by Pol II as a longer (typically > 1kb) primary transcript (pri-miRNA) that contains a hairpin structure where the mature miRNA sequences are located. The mature miRNAs are cleaved from the pri-miRNA by the sequential action of the RNase III type proteins Drosha and Dicer [132, 129]. Greater than 60% of human protein-coding genes have at least one evolutionarily conserved miRNA binding site; with the addition of non-conserved binding sites, it is reasonable to assume that most protein-coding genes may be regulated by miRNAs [133]. There is no set classification rule for miRNAs, but it is generally accepted that miRNAs with an identical sequence from nucleotides 2-8 (known as a ‘seed region’) of the mature miRNA form part of the same miRNA ‘family’ and have similar binding affinities for target genes [134, 135].

Since miRNAs primarily target the 3’ UTR, it is reasonable to expect an interaction between this form of post-transcriptional regulation and APA. It has been suggested that ~67% of genes that undergo 3’ UTR shortening in tumours undergo the loss of at least one potential miRNA binding site [66]. Indeed, it has been suggested that the binding of miRNAs to the 3’ UTR of mRNAs is often mitigated by the selection of a more proximal APA site [112]. This type of interaction has been suggested in aggressive breast and lung tumours, with the APA-associated loss of binding sites for the tumour suppressor miRNAs; miR-7, miR-299-5p, miR-200bc/429 and miR-496 [136]. These interactions suggested that, in addition to downregulating the expression of these miRNAs, cancer cells could overcome miRNA-based regulation through the selection of proximal APA sites.

1.5 APA in proliferation and disease

As the majority of human genes undergo some form of APA, it is not surprising that APA dysregulation has been associated with many diseases. Just as AS has received increased attention for its role in disease [137], in recent times, APA is receiving more attention as an additional form of post-transcriptional regulation which can become altered in disease. A summary of some known APA events in a variety of disease settings can be seen in Table 1.2. Discussed here are some key examples of modified APA in a variety of disease settings, highlighting the importance of APA as a target for disease prognosis and treatment in a range of cellular contexts.

1.5.1 APA and viruses

A recently emerging area of research is the role of APA in viruses and the effect both host and viral APA machinery have on viral replication. APA has been demonstrated in many viral types including retroviruses such as HIV [138, 139], and DNA viruses such as the adeno-associated virus type 5, where APA is key to the generation of viral proteins in the required ratio [148]. APA is also a key factor in the host response to viral infection. A study of the innate immune response of macrophages following infection with vesicular stomatitis virus showed a tendency toward 3’ UTR shortening as the host cells responded to the infection [149]. Furthermore, knocking down the 3’ processing factors PABPN1 and CPSF4 increased the speed of viral replication, whereas knocking down NUDT21 and CPSF6 had the inverse effect. There was a poor correlation between changes in mRNA abundance and APA, with only a quarter of the genes that exhibited a change in abundance also exhibiting APA site switching. This suggests that APA is an integral part of the viral response at more levels than simply the moderation of gene expression.

Table 1.2. Known, disease associated APA events from the literature.

Disease	Summary	Paper
Alzheimer's disease (AD)	The short Tau (MAPT) 3' UTR isoform is expressed in AD patients, escaping miR-34a repression. This generates more Tau, known to aggregate in AD.	Dickson *et al.* [140].
Heart failure	A trend toward genome wide 3' UTR shortening is seen in the failing human heart.	Creemers *et al.* [141].
Human immunodeficiency virus (HIV)	*CPSF6* assists viral integration into transcriptionally active chromatin in the host genome	Sowd *et al.* [139].
Herpes simplex virus (HSV)	Infected cell culture polypeptide 27 (*ICP27*) promotes cryptic APA in host cells	Tang *et al*. [142].
Glioblastoma	CFIm25 (*NUDT21*) depletion causes the usage of proximal 3' UTRs in glioblastoma tumours	Masamha *et al*. [143].
Myotonic dystrophy (DM)	The MBNL family of proteins regulate APA. Inhibition of MBNL proteins is a major contributing factor to misregulated APA in DM.	Batra *et al.* [144].
Parkinson's disease (PD)	APA of α-Synuclein (*SNCA*) assists in protein expression and localisation forming Lewy bodies, a characteristic of PD	Rhinn *et al.* [145].
Prostate cancer	APA in prostate cancer changed the availability of miRNA binding sites, modulating competing endogenous RNA (ceRNA) networks	Li *et al.* [146].
Sindbis virus	The Sindbis virus localises the HuR protein to the cytoplasm (as opposed to the nucleus) upon infection	Barnhart *et al.* [147].
Triple negative breast cancer	Pumilio RNA binding protein complex sites are lost in TNBC through 3' UTR shortening	Miles *et al.* [110].
7 cancer types	3' UTR shortening was associated with tumourigenesis and CstF64 (*CSTF2*) is a master regulator of APA in tumours	Xia *et al.* [66].

Viral proteins have also been found to interact with host APA processes directly. Alternative polyadenylation machinery encoded in the herpes simplex virus (HSV) has been shown to activate cryptic splicing and APA sites in about 1% of genes [142]. This splicing and APA is brought about by infected cell culture polypeptide 27 (*ICP27*), which is encoded by the virus. The endogenous splicing factor *CPSF6* has also been implicated in assisting HIV integration to more transcriptionally active regions of chromatin, thereby assisting in replication [139]. In a more extreme example, the Sindbis RNA virus has been found to cause the sequestration of *ELAVL1* RNA binding protein to the cytoplasm of cells, acting as a molecular sponge. This effect causes the loss of this protein from the nucleus and changes the APA and AS of the host [147].

1.5.2 APA in neurodegenerative diseases

Over a decade ago APA was recognised in the EAAT2 (*SLC1A2*) glutamate transporter at different sites in the human brain [150]. Defects in glutamate transporters have been associated with multiple diseases including Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS) and Huntington's disease, (reviewed in ref [151]). Despite these early findings APA was not widely considered as a contributing factor in neurodegenerative diseases until it was more recently associated with proliferation and cancer [112, 109]. It is also interesting to note that the longest 3' UTR transcripts are expressed in the brain [152], suggesting a diverse regulatory environment that is potentially heavily regulated by APA.

A recent study found APA to be associated with the aggregation of the Tau (*MAPT*) protein in AD. The Tau protein forms neurofibrillary tangles (NFTs) as part of the pathology of AD [153]. Tau expression in human neuroblastoma cell lines was shown to be altered by APA, with the longer 3' UTR isoform subject to regulation by the miR-34 family [140]. Expression of the proximal isoform may, therefore, contribute to the overexpression and build-up of the Tau protein by escaping miR-34 regulation. *MECP2* is another key gene in multiple neurodegenerative disorders and has been associated with the expression of APA processing factors such as *NUDT21*. Copy number variations in *NUDT21* were identified in 11 patients with psychiatric disorders. The resulting increase in *NUDT21* expression was shown to regulate the *MECP2* 3' UTR by causing the preferential expression of the long isoform, which in turn was associated with the production of less protein [154]. An analogous study of APA of the *α*-Synuclein (*SNCA*) gene in Parkinson's disease (PD) found 3' UTR lengthening to be associated with disease, appearing to redirect protein localisation to the mitochondria instead of the synaptic terminals [145]. In this case, accumulation of a long 3' UTR isoform was associated with increased protein production. These changes in expression and localisation are thought to assist *α*-Synuclein in forming Lewy bodies, which are a defining pathological characteristic of PD [155]. In a novel, related observation, a study of APA in ALS revealed cryptic poly(A) sites present within intron 7 of the *EAAT2* gene that were shown to only be activated upon RNA editing [156]. If this observation is reproduced, RNA editing may also need to be considered in the regulation of APA site selection not just in neurodegenerative diseases, but under all known APA-associated processes.

1.5.3 APA in cardiac diseases

An emerging area of research is the role of APA in heart conditions. A general trend toward 3' UTR shortening was seen in mouse hearts where cardiac hypertrophy (CH) had been induced [157]. It was suggested by the authors that this increase in proximal 3' UTR isoforms may result in additional protein production as part of a rapid response to CH. Interestingly, this pattern of APA was suggested to be the reverse of the pattern seen in cardiac development. The authors also note that there were no significantly enriched GO terms associated with genes from these shortening events, suggesting a more general program of APA. These findings were supported by a similar study, in which a trend toward shorter 3' UTRs and altered miRNA expression with CH was also seen [158]. It would also appear that a pattern of 3' UTR shortening persists from chronic heart conditions such as CH and is also present in failing hearts [141], however, it is not clear if these APA events occur in similar genes.

Consistent with the findings in CH is an example of APA in the heat shock protein 70 (Hsp70) complex. Hsp70 has been shown to protect cells from damage under stressful conditions, such as the oxidative stresses that are produced by Ischemia/reperfusion (IR) related insults. A key component of this complex is Hsp70.3 (*HSPA1B*), which undergoes APA to preferentially express the shorter 3' UTR isoform, in addition to an increase in gene expression, in response to IR insults [159]. The shorter Hsp70.3 isoform lacks a miR-378 binding site, which was reported to regulate the long isoform. Beyond these studies, APA has not been well studied in the cardiac setting, however, RNA binding proteins are known to play a substantial role in cardiac development [160] and as previously mentioned, transition toward a developmental pattern of APA can be seen in cardiac hypertrophy [157]. This suggests that APA is likely important in cardiac disease and will become an expanding area of study in future.

1.5.4 The CPA machinery and APA in proliferation

Preferential usage of proximal APA sites has previously been observed in proliferating cells, while mature, differentiated cells predominantly use more distal sites [112]. The exact reasons behind this phenomenon are still unclear, however, it seems likely that this is at least in part due to the greater regulatory opportunities afforded by 3' UTR-focused mechanisms such as miRNA. It follows that when cells revert to a more proliferative phenotype, such as in cancer, the change to proximal polyadenylation sites must be promoted, for example by the recruitment of specific factors to proximal APA sites. Alternatively, there may be a reduction in the APA site specificity of the core factors of the CPSF complex causing proximal binding. Both of these models could also be explained by an increase in the expression of the core factors of the CPSF, as this could cause an overall increase in binding at both APA sites.

In support of a model of increased CPA factor expression with proliferation, a review of pooled data from The Human Gene Expression Atlas [161] has shown a positive correlation between the expression of many CPSF complex members and the proliferative state of cells [162]. However, increased use of proximal APA sites has also been observed when the expression of some CPSF complex members was knocked down by siRNA [163, 164]. Thus, a straightforward hypothesis on the biochemical processes behind APA is difficult to formulate, unless proximal APA is decoupled from the proliferative state of cells. Despite the lack of specific understanding of the factors that determine APA site choice, the effects of APA on the transcriptional state of oncogenes are nonetheless an important consideration when attempting to understand the proliferative state of cells, including cancer cells. Indeed, dysregulation of APA, specifically by shortening 3’ UTRs, has been suggested to be a major driver of tumour progression [66, 136].

1.5.5 APA in cancer

Since the discovery that proliferating cells have a tendency to express shorter 3’ UTRs was extended to cancer [109], the role of APA in tumours has become the most well studied of any disease. Perhaps the most comprehensive study of APA in cancer to date was that performed by Xia *et al*. [66], which inferred APA from RNA-Seq data across 7 tumour types from the cancer genome atlas (TCGA) and found APA to be more prognostic of tumour outcome than gene expression. Similar results were obtained by Li *et al*., [146] who found that 3’ UTR shortening in prostate cancer changed the availability of miRNA binding sites, modulating competing endogenous RNA (ceRNA) networks, especially in higher-risk cancers. The results of previous studies of APA in tumours are certainly novel and highly interesting, but often did not interpret their findings in the context of what is already known about breast cancer biology such as commonly analysed gene expression [45] or clinical factors [165]. Therefore, it is still not known if APA is providing new information about the state of a tumour, or if it is simply a passenger, driven by previously characterised tumour associated processes.

1.5.6 APA in breast cancer

Cancer is a complex disease that varies greatly at the patient-specific level [166]. Despite this fact, cancers are often grouped into clinically useful groups (subtypes) that may share common molecular targets. Breast cancer is no exception, with multiple known subtypes, each with varying properties that are determined by both clinical markers [167] and associated gene expression signatures [45]. It remains to be seen whether APA could be used to subtype breast tumours and indeed whether it is still more prognostic than gene expression signatures [66] when cancer subtype is considered. The loss of Pumilio RNA binding protein complex sites in TNBC through 3' UTR shortening was associated with an increase in protein levels of target genes [110]. However, as previously mentioned, RNA binding proteins may also have a stabilising effect and 3' UTR shortening could also increase protein abundance. Indeed, not all studies of APA in cancer have found shorter 3' UTRs to be associated with tumour-derived cells. One group that looked at APA in TNBCs using a 3'-focused sequencing method found that 3' UTRs in MDA-MB-231 cell lines were longer on average when compared with the epithelial-derived MCF10A cell lines. In contrast, the luminal derived MCF7 cell line exhibited greater 3' UTR shortening [168]. While it would be convenient if all cancers expressed outcome-associated shorter 3' UTRs in similar patterns, the conflicting evidence in the literature suggests that this is likely not the case and that the deregulation of APA in cancer is substantially more nuanced than previously suggested.

It has previously been difficult to determine if observed changes to 3' UTR dynamics in tumours are isolated events or representative of more widespread cancer-associated processes. The availability of novel RNA-Seq based methods have given researchers new ways to assess these changes on a transcriptome-wide scale. It has also been discovered that it is possible to reinterpret the results of older genome-wide assays, that were designed primarily to measure gene expression, to gain additional information about APA state. These new methods unlock large cohorts of breast cancer samples for novel APA analysis that may also include associated clinical and survival data.

1.6 Themes and trends in the study of APA

The importance of altered global APA in changing cellular state has only been known for a decade [112]. Despite this, the methods discussed above for measuring global APA directly have only been under development recently [169, 117, 68]. As such, there is still much that is not known about the consequences of APA, and some conflicts have arisen over the inferences made by the seminal papers in which pervasive APA was first described. It has been suggested that the assertions made by these papers that shorter 3' UTRs lead to both increased mRNA stability and TE were not reproducible and potentially based on biased data [112, 109]. The validity of these assertions will be discussed further in this section.

1.6.1 Distal APA events may also increase mRNA stability

The majority of research presented thus far has supported what has become the widely accepted view, that proximal APA leads to loss of miRNA binding sites and therefore greater transcript stability, TE and protein production [112, 109]. While this effect has been shown to be biologically significant in many single gene examples, there is far less evidence for the global effects that were claimed by these studies. A dissenting paper by Gruber *et al*. [162]

challenges the seminal finding by Sandberg *et al*. [112], that the shorter 3' UTRs produced by T cell activation result in increased protein production. Downregulated genes in this study were actually found to have more 3' UTR shortening than their upregulated counterparts.

The proposed increased stability of shorter 3' UTR isoforms is also currently being challenged, with some studies suggesting that mRNA stabilising elements could become known as the predominant form of regulation once all interactions have been defined [183]. A recent *in vitro* study attempted to define the role of functional regulatory 3' elements [184]. The authors cloned a custom library of conserved 3' sequences into an expression vector, expressed the vector in the Flp-In 293 cell line, fluorescence activated cell sorting (FACS) binned the cells by reporter expression and performed high throughput sequencing on the resulting bins. The study found that there may be as many activating elements within 3' UTRs as repressive elements. An example of mRNA stabilisation by an RNA binding protein is the like-Sm (LSm) protein family member Ataxin-2 (*ATXN2*) which binds both U-rich and AU-rich elements in the 3' UTR, stabilising associated mRNAs and increasing the output of corresponding proteins [185]. This is consistent with many examples of older research that suggested that distal APA is important in key cellular processes such as the stabilisation of *IL2* in T-cell activation, the mitogen-activated protein kinase pathway, and Tau protein stabilisation in Alzheimer's [186, 114, 187].

There are also many recent examples of increased mRNA stability in association with the use of a distal 3' UTR isoform. In a mutant *D. melanogaster* strain with a lower transcriptional elongation rate, the forced use of the proximal APA site of the *polo* cell cycle gene was shown to cause death of the flies at the pupa stage [188]. Another study suggested that the distal isoform of the anxiety-associated serotonin transporter *SLC6A4* is bound by heterogeneous nuclear ribonucleoprotein K (HNRNPK) increasing translation [189]. The human antigen R (ELAVL1) RNA binding protein has also been shown to stabilise the key cancer vascularisation protein VEGFA by binding to AU-rich elements in the 3' UTR [190]. Studies of APA and proliferation should, therefore, be mindful of the potential inclusion of RNA stabilising CREs in longer 3' UTR isoforms as well as the loss of destabilising CREs in proximal isoforms.

1.6.2 APA events are diverse and specific

If the recent findings that APA does not have a general impact on protein production hold true, it follows that APA events may be more specific in various cellular states, rather than generalised proliferation-associated reprogramming. This point is emphasized in Figure 1.5 A, which shows that the majority of APA events are unique across 4 different disease settings where APA is known to play a role. In 5 different cancer types from the TCGA, which were determined by Xia *el al*. [66] (Figure 1.5 B), a third of APA events were cancer type specific when compared with one another [191] suggesting that the majority of cancer-associated APA events are not tumour-specific. This lack of specificity further highlights the fact that APA events may not all be part of a proliferation associated switch. Each gene may instead be individually regulated as part of a broader array of mRNA regulation that changes depending on cellular conditions. In support of this, it has been suggested that differing APA profiles exist, even within breast cancer subtypes [168]. This idea will be further explored throughout this thesis.

1.6.3 Biases in methods used in previous APA research

The study of 3' UTR regulation has highlighted the importance of APA in translational regulation and diversification, however, there are some general methodological issues that

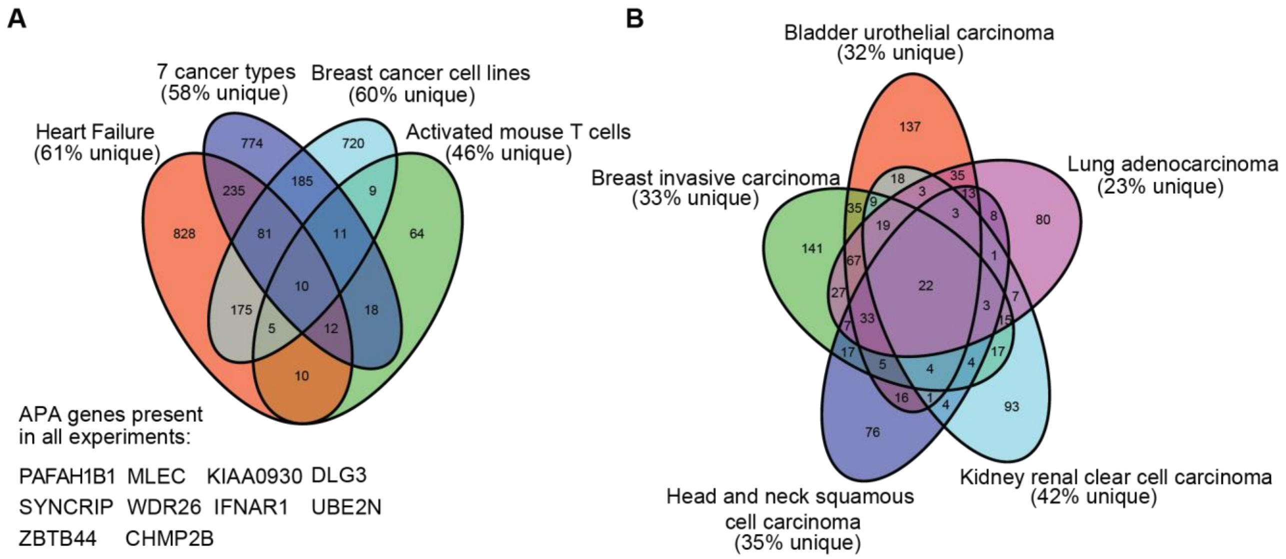


Figure 1.5. APA genes in varying cancer and disease states. A. Venn diagram of overlapping gene sets from 4 different experimental conditions: APA events taken by comparing test and control samples from the failing human heart [141], 7 cancer types [66], 2 breast cancer cell lines (HER2 positive and negative) [168] and activated mouse T cells [162]. The 10 genes that underwent differential APA in all experiments are listed in the bottom left. **B.** Common APA events from 5 of the 7 different tumour types observed by Xia *et al.* [66] using data from the TCGA (http://cancergenome.nih.gov/). The percentage of APA genes unique to a given experiment (A) or cancer type (B) are listed below each experiment label. This figure was reproduced from my co-first authored paper (Turner *at al*., Appendix A5).

should be considered in future research. Much of what is currently known about APA was discovered as part of cell line studies [112, 109, 168]. Studies of APA in cell lines can be complicated by the fact that cell lines, by nature, have been induced to continue to proliferate as part of the immortalisation process [192]. It may therefore not be appropriate to use tumour and non-tumour cell lines as models of APA in diseases with increased cellular proliferation, such as cancer [168]. Many studies of APA have also been performed with data from agglomerated publicly available microarray datasets. These methods use custom probe reanalysis methods and a database of known APA sites to determine APA, such as the work in breast and lung cancer by Lembo *et al.* and breast cancer by Akman *et al.* [136, 113]. This approach requires correct and complete databases, analysis methods and correction for batch effects where appropriate. These confounding factors are impossible to completely control for and should be kept in mind when interpreting the findings of these papers. Furthermore, not all APA events have associated probes that can be exploited in this type of analysis. Findings from these studies will need to be validated in genome-wide APA profiling experiments such as PAT-seq before genome-wide inferences can be confirmed.

1.6.4 The potential of future APA research

It was recently discovered that APA can mediate protein-protein interactions and further diversify protein localisation and function [194]. It was discovered that when *CD47* was expressed with a short 3’ UTR it was localised to the endoplasmic reticulum, whereas when *CD47* was expressed with a longer 3’ UTR it was expressed on the cell surface. The authors found that even when localised to the same cellular compartment that the different isoforms may have distinct functions. This is another striking example of the complexity and diversity of APA regulation in humans. The dysregulation of these processes is therefore likely involved in all complex diseases and even many host-pathogen interactions. Identification and treatment of highly diversified and specific disease-associated APA events provides a massive challenge, however, once this challenge is met, there is also potential for highly diversified and specific treatments for these diseases. Consideration of APA in combination with other modes of genome regulation therefore potentially provides an opportunity for the production of novel, multi-target therapies essential in the treatment of complex diseases such as triple negative breast cancer (TNBC) [21].

1.7 Technologies for measuring the post-transcriptional state of cells

The regulation of gene expression has long been regarded as a controlling factor in almost all cellular processes, and it was widely believed that a complete understanding of the gene expression in a complex disease such as breast cancer would largely explain disease state and predict outcome [170, 56]. While gene expression explained some variation between tumours, it has become increasingly clear that there are many additional processes (such as miRNA-mediated gene expression regulation and APA) that govern the relationship between transcription and translation [109]. Early microarray-based approaches to measure genome-wide expression had some technical limitations such as cross-hybridization and poor detection of low abundance-transcripts [171]. As technology has improved, new methods such as RNA-Seq have been developed for the robust measurement of the entire transcriptome at single nucleotide resolution [172, 170], as well as for investigating post-transcriptional regulatory mechanisms such as AS. Described here are platforms for measuring gene expression and other aspects of translational regulation such as APA, poly(A) tail length and the expression of non-coding RNAs. Specific details are included in the description of these methods, because this information is key to understanding how the data that was previously generated may be reinterpreted to infer APA.

1.7.1 Gene expression microarrays

For over two decades gene expression microarrays (Figure 1.6) have been in use as a key technology to analyse the expression of thousands of genes simultaneously. An array is constructed by spotting short DNA sequences (25bp in Affymetrix HG-U133A arrays) known as probes out onto a glass slide in a pre-defined pattern. In the Affymetrix HG-U133A array, purified mRNA from the sample to be tested is first transcribed into cDNA in the presence of

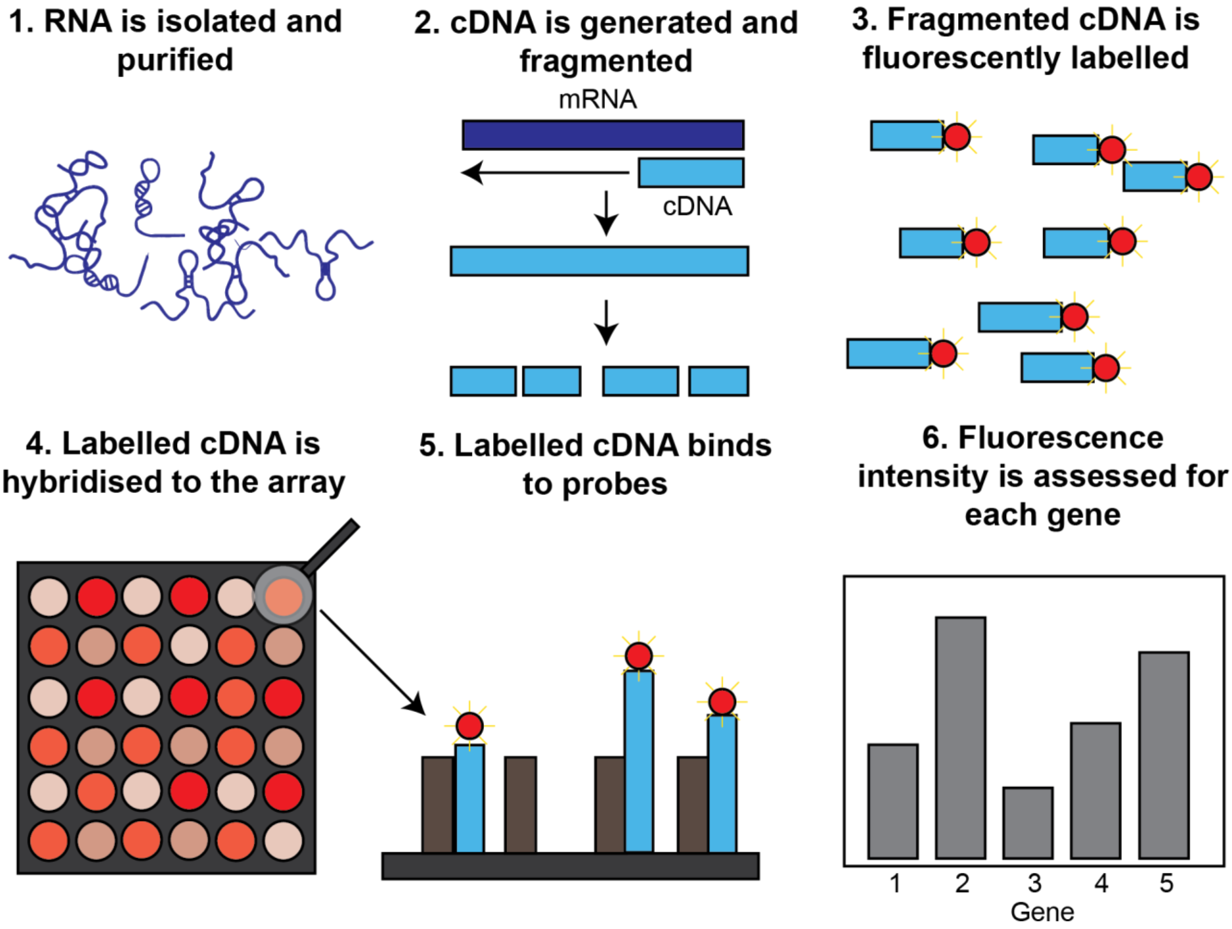

1. RNA is isolated and purified
2. cDNA is generated and fragmented
mRNA
cDNA
3. Fragmented cDNA is fluorescently labelled
4. Labelled cDNA is hybridised to the array
5. Labelled cDNA binds to probes
6. Fluorescence intensity is assessed for each gene
1
2
3
4
5
Gene

Figure 1.6. Schematic describing a single-channel gene expression microarray. Gene expression microarrays use cDNA hybridisation to measure the expression of all known genes in a sample. Known as a ‘probe’ Each spot on the array represents a cluster containing millions of copies of a short sequence from a known gene. The expression of ~11-20 probes matching the same gene is then combined to form a probeset. The expression of a single gene may often be targeted by multiple probesets.

ribonucleotide triphosphates labelled with a reporter molecule (such as biotin). The cDNA is then fragmented prior to array hybridisation, where the fragmented cDNA molecules bind to their complementary sequences. The array is then washed with a fluorescent molecule that binds to the reporter [173]. As the location of a specific sequence on the array (known as a probe) is known, the fluorescent signal for each spot can then be read by a scanner. For greater accuracy when measuring gene expression, there are normally multiple probes that bind to different sequences from the same gene. These sequences can be subsequently grouped into ‘probesets’. The expression of probes can be aggregated into the expression of probesets, which can then be interpreted and normalised by specially designed software [174] and finally used to infer gene expression [175]. Depending on the array type, there may be multiple probesets that bind to a single gene for even greater redundancy. Some gene expression microarrays can be re-analysed for the detection of APA events [136]. This is done by reorganising probes into new probesets that are defined around a known APA site and is only possible when probes or probesets were previously designed to bind to the 3’ UTR of a given gene. Microarrays are currently still in use for the measurement of gene expression, however presently, it is more common to use second-generation sequencing methods such as RNA-Seq, which reports the actual sequence of mRNA molecules, resulting in greater accuracy [176].

1.7.2 RNA-Seq

RNA-Seq (Figure 1.7) allows the estimation of the presence and abundance of every RNA molecule in a sample, enabling the discovery of novel RNA species and adding unprecedented depth to transcriptome-wide analyses. RNA-Seq and other high-throughput, small-read sequencing methods are commonly referred to as next-generation sequencing (NGS). In its simplest form, RNA-seq begins with total cellular RNA extract from which mRNA

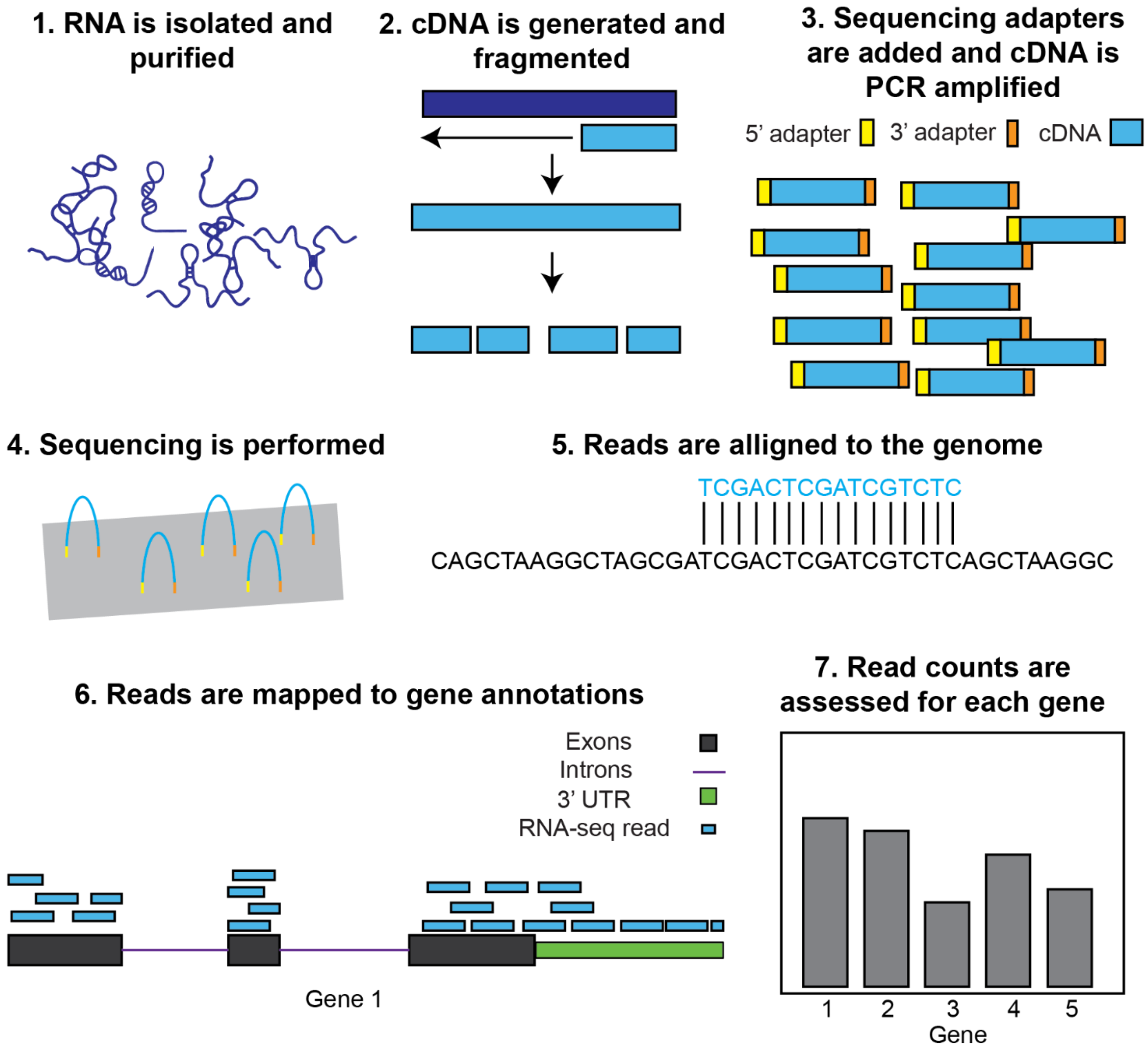
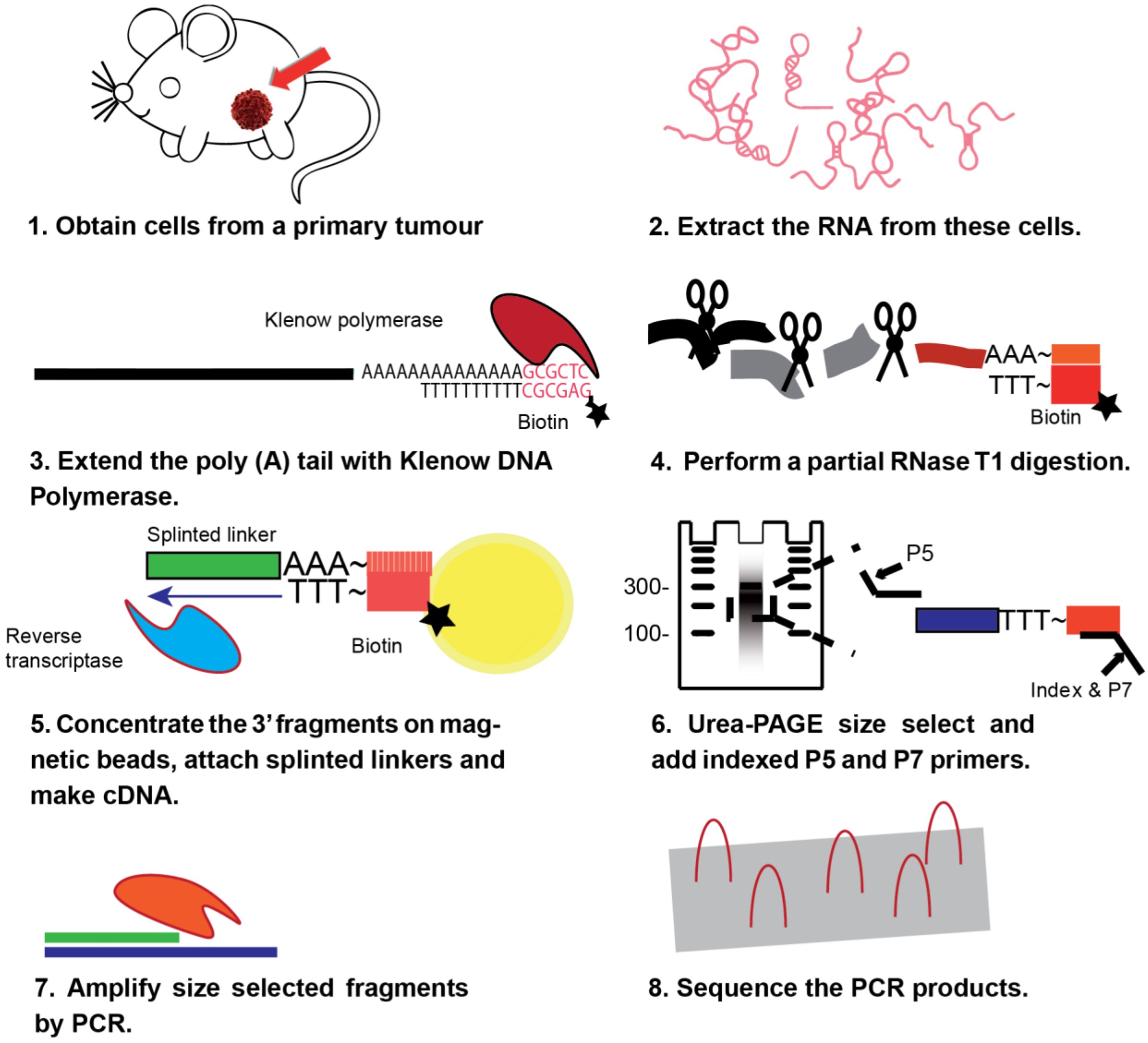

1. RNA is isolated and purified
2. cDNA is generated and fragmented
3. Sequencing adapters are added and cDNA is PCR amplified
5' adapter
3' adapter
cDNA
4. Sequencing is performed
5. Reads are alligned to the genome
TCGACTCGATCGTCTC
CAGCTAAGGCTAGCGATCGACTCGATCGTCTCAGCTAAGGC
6. Reads are mapped to gene annotations
Exons
Introns
3' UTR
RNA-seq read
Gene 1
7. Read counts are assessed for each gene
1
2
3
4
5
Gene

Figure 1.7. Schematic describing short-read RNA-Seq. RNA-Seq measures the expression of every gene in a sample at the sequence-specific level.

molecules are selected through means such as ribosome depletion (removal of the highly abundant ribosomal RNA from total RNA) [177] or poly(A) selection (the binding to any RNA molecules with a poly(A) tail using an oligo-dT sequence). Depending on the specific sequencing protocol, the RNA or resulting cDNA is then fragmented, either prior to or post reverse transcription. Sequencing adapters are then ligated to the short cDNA molecules produced to allow the reads to bind to the flow cell of the sequencer for subsequent sequencing. The short reads produced can then be aligned to a reference genome or undergo *de novo* transcriptome assembly to create a new reference genome [176]. Current ‘second-generation’ RNA-Seq methods rely on short-read sequencing (50-300bp), although long read (up to 10 kb) ‘third-generation’ sequencing methods are being developed [178]. While these new technologies hold great promise, there are currently large publicly accessible databases of short-read RNA-Seq such as the TCGA and sequence read archive (SRA) which are yet to be mined to their full potential. This potential lies in the fact that RNA-Seq can also be exploited or modified to measure other transcriptomic processes such as AS, post-transcriptional modifications and single nucleotide polymorphisms [179] among many additional novel applications. One such example of a previously untapped source of information is the differences in RNA-Seq coverage at the 3’ end of genes that can be exploited to look for APA [66].

1.7.3 PAT-seq

Not long after the development of current RNA-Seq technologies, novel methods were developed to specifically measure 3’ UTR dynamics. PAT-seq [169] (Figure 1.8) is a method developed by the RNA Systems Biology Laboratory (the laboratory in which this thesis was undertaken). In this method, all polyadenylated RNA is extended with dNTPs on an oligo-dT template using Klenow polymerase (extends mRNA on a DNA template) to 3’-end



1. Obtain cells from a primary tumour

2. Extract the RNA from these cells.

3. Extend the poly (A) tail with Klenow DNA Polymerase.

4. Perform a partial RNase T1 digestion.

5. Concentrate the 3' fragments on magnetic beads, attach splinted linkers and make cDNA.

6. Urea-PAGE size select and add indexed P5 and P7 primers.

7. Amplify size selected fragments by PCR.

8. Sequence the PCR products.

Figure 1.8. PAT-seq. This Illumina based, second generation sequencing method is relatively inexpensive to perform and can measure APA, gene expression and poly(A) tail length.

label the RNA before cDNA synthesis is carried out. The Klenow extension will only extend from the end of RNA molecules, and the high cDNA synthesis temperature restricts priming to end labelled RNA, reducing internal priming to poly(A) tracks. Included at the 5' end of the oligo-dT template is an Illumina-specific reverse primer sequence and a biotin moiety that allows for purification in subsequent steps. Similar to other 3'-focused sequencing methods such as poly(A)-tail length profiling by sequencing (PAL-seq) [68] and TAIL-Seq [117], the RNA is subjected to limited RNase T1 digestion at a low concentration. As RNase T1 cleaves only after G bases, the poly(A) tail and the DNA template are not digested, leaving a 3' UTR of variable length with an intact poly(A) tail bound to an extended DNA molecule. These fragments are then purified on streptavidin-coated magnetic beads through binding to the biotin moiety. The 5' ends of these fragments are phosphorylated and an Illumina-compatible splinted 5' adapter is ligated. Reverse transcription is then primed from the 3' end of the bead-bound mRNAs. The cDNA is eluted from the beads, size selected, and PCR amplified before directional sequencing from the 5' adapter into the poly(A) tail. The advantages of this method are that it is relatively inexpensive to perform (runs on a single lane of an Illumina Flow Cell) and can measure APA, gene expression and poly(A) tail length. The greatest practical strengths of PAT-seq are that it requires no special access to sequencing machines, is easily multiplexed (thanks to compatibility with indexed sequencing primers) and can be performed by most laboratories without any special equipment. PAT-seq is however limited by the effects of polymerase slip, a PCR artefact that has the potential to underestimate the length of the poly(A) tail, causing it to appear shorter on average. PAT-seq is also not able to measure poly(A) tails > ~120 bases due to short-read sequencing length constraints.

1.7.4 Alternative 3'-focused sequencing approaches for measuring APA and poly(A) tail length

Two alternatives to the PAT-seq method have also been recently developed, known as TAIL-seq [117] and PAL-seq [68]. TAIL-seq utilises affinity-based depletion of ribosomal RNAs and size fractionation against all remaining RNAs, selecting for mRNAs. This method of purification was chosen on the basis that the standard oligo(dT) based purification introduces a bias toward mRNAs with a longer poly(A) tail. The trade-off for this type of mRNA isolation method is the higher rate of intronic and intergenic regions that are purified [180], potentially reducing the amount of information obtained concerning gene expression. A 3' adapter that has been fused to a biotin tag is then ligated to the 3' end of the transcripts, before RNase T1 fragmentation (low concentration), which cleaves after G residues. The biotin tag then allows for the purification of 3' cleavage products. The RNA is then size selected by gel purification and a 5' adapter is ligated. The RNA is then reverse transcribed, amplified by PCR and directionally sequenced by paired-end sequencing. When using standard base calling (base calling by the sequencer), spiked in standards of known poly(A) tail length were overestimated. To address this, a base calling algorithm for poly-T stretches on the cDNA based on fluorescence signals from the sequencer was formulated. Another advantage of TAIL-seq is the ability to detect bases other than adenosine at the end of the poly(A) tail as a 3' adapter can be ligated without oligo(dT) based purification that would select against this. TAIL-seq is similarly limited to PAT-seq in that it is theoretically only able to accurately measure the poly(A) tail from 8 to 231 adenosines. The paired-end sequencing in Tail-seq may also miss some APA events depending on the length of the mRNA fragments following RNase T1 treatment.

Prior to sequencing, PAL-seq, the other PAT-seq alternative, is largely similar in design to TAIL-seq, with the major differences being a lack of purification and PCR steps in PAL-seq and in the way in which the problem of long homopolymer sequencing is addressed. After generation of Illumina sequencing clusters, but before sequencing, a primer hybridised immediately 3’ of the poly(A) tail is extended using both deoxythymidine triphosphates (dTTPs) and biotin-conjugated deoxyuridine triphosphate (dUTPs). In order to identify which gene the mRNA originated from, 36 nucleotides are then sequenced immediately 5’ of the poly(A) tail. Finally, the flow cell is incubated with fluorophore-tagged streptavidin which binds to the biotin to generate fluorescence intensity that is representative of poly(A) tail length. The advantages of this method are that it quantitatively measures the poly(A) tail with no length restrictions and does not rely on PCR amplification before sequencing. The limitations of this method are that it does not detect 3’ UTR switching (although this is possible), or the use of any uridine or guanine bases in the poly(A) tail, and requires a high input of RNA due to the lack of PCR amplification. Furthermore, PAL-Seq would not be able to measure the poly(A) tail down to single nucleotide resolution. The greatest practical limitation of this method is that it requires special access to a sequencer, which is not possible in laboratories that rely on sequencing services.

1.7.5 APADB: A database of known APA sites in humans

Often, for applications such as inferring APA from RNA-Seq data or validating sites detected by a 3’ RACE experiment, it is useful to have a reference of known polyadenylation sites. The most complete database of known human poly(A) sites generated to date is the alternative polyadenylation data base (APADB) [181]. Poly(A) sites in the APADB are identified via the massive analysis of cDNA ends (MACE) method [182]. MACE is a high throughput NGS based method that targets the 3’ end of transcripts. To generate this database for human APA, MACE was used to measure poly(A) sites from seven different cell types including three tumours. Custom bioinformatics scripts were used to identify polyadenylated reads and to perform one-dimensional clustering of these reads into poly(A) sites. On average most poly(A) sites are between 1 and 25 bp, however, there was additionally a possibility for wider sites to be included if the range of reads within a cluster was also wider.

Together, these resources and methodologies provide a suite of tools to address the emerging questions regarding the importance of APA in developing cellular diversity, regulating cell fate, and in the development and progression of disease. Although, as discussed in Section 1.5, regulation of APA in various diseases has been observed, the mechanisms governing APA site selection, as well as how these are involved in disease, remain poorly understood.

1.8 Aims of this thesis

Gene expression patterns present in a primary tumour biopsy have long been considered to be prognostic in breast cancer [56], however, gene expression in combination with clinical markers often fails to provide sufficient prognostic information to alter a treatment decision. More recent studies have associated APA with increased cellular proliferation, including in the cancer setting [112, 109]. Furthermore, it has also been recently suggested that APA is highly prognostic in breast cancer [66], even in the previously unpredictable TNBC subtype [113, 65]. These studies have, however, usually only considered breast cancer as a single disease entity, which is not the case because, as previously mentioned, breast cancer is highly heterogenous. Where subtype was considered, they have focused entirely on TNBCs, as this is the subtype for which new treatments are most required, leaving many subtype-specific questions largely unanswered. Moreover, these studies have drawn their conclusions by each using their own novel methods to re-interpret various RNA-Seq and microarray datasets [66, 113, 65]. These methods were often employed with minimal validation and without comparison to similar data sets, leading to the potential discovery of false positive APA events.

The utility of APA as a prognostic marker in all breast cancer subtypes has therefore not yet been fully explored and warrants further analysis. Poly(A) tail length has also been previously suggested to play a role in proliferation [119], however, the effect of poly(A) tail length on the metastatic potential of a primary tumour is not well studied. It was therefore hypothesised that gene expression, APA and poly(A) tail length changes would be associated with breast cancer formation and prognostic of primary tumour outcome. This hypothesis was addressed with the following aims:

Aim 1. Determine the metastasis-associated changes to RNA metabolism in a xenograft model of TNBC.

Aim 2. Infer outcome-associated APA events from microarray and RNA-Seq data in clinical patient breast tumour samples.

Aim 3. Develop a prognostic test to predict the outcome of breast cancer from RNA extracted from a primary tumour biopsy.

Aim 4. Develop novel bioinformatics tools that increase the accessibility of novel 3’ datasets to researchers lacking computational experience.

1.9 Completion of aims and summary of results

The four aims outlined in this thesis are addressed in the results chapters (3-5). Aim 1 is addressed in Chapter 3, Aims 2 and 3 are addressed in Chapter 4 and Aim 4 is addressed in Chapter 5. A brief description of the content of each results chapter is provided below:

Chapter 3 describes the identification of novel differential gene expression (DGE), APA and differential poly(A) tail length changes in an increasingly metastatic mouse xenograft primary tumour model (MXM). The MXM was generated by Cameron Johnstone in collaboration with Robin Anderson. The MXM consisted of four MDA-MB-231 derived TNBC cell lines, each with increasing metastatic potential. The APA, gene expression and poly(A) tail length state of primary breast tumours in the MXM was quantified by PAT-seq, which sequences 3' UTRs in a genome-wide fashion. Gene expression and APA changes were observed with increasing metastatic potential, however, no poly(A) tail length changes could be confirmed. Many of the top gene expression pathways altered in metastasis were also associated with RNA processing, pointing towards altered RNA metabolism being key to TNBC metastasis. It was suggested by the literature that the more metastatic primary tumours would exhibit a shift toward primarily proximal APA. While the most highly metastatic tumours behaved as expected, the tumours primarily targeting the lung exhibited predominantly distal APA. This showed that specific APA events may be associated with specific metastatic tropisms as opposed to the general metastasis-associated effect that has previously been observed.

Chapter 4 describes the identification and characterisation of novel APA events in microarrays from the Gene Expression Omnibus (GEO) and in RNA-Seq from the TCGA. More than 1700 breast cancer-associated APA events were identified overall, with 100 'high confidence' APA events shared between both the TCGA and GEO datasets. Also identified from the TCGA data was a high correlation of more pronounced overall proximal APA in some genes and more pronounced overall distal APA in others. This effect was not linked to breast cancer outcome but was associated with the expression of RNA processing genes. This points toward the possibility that the greater expression of APA processing factors further exacerbated tumour-associated APA events rather than favouring a switch to proximal or distal events. APA events from the TCGA were also used in concert with clinical and gene expression data to create a linear model for the prediction of breast cancer prognosis. Interestingly, a model of APA + clinical data was a better predictor of breast cancer outcome than gene expression + clinical data and was significantly better than clinical data alone ($p < 0.05$). This model was prognostic across multiple breast cancer subtypes including TNBCs, potentially forming the basis of a novel APA based prognostic test.

Chapter 5 describes the novel bioinformatics methods that were employed in the generation and analysis of the data presented in Chapters 3 and 4. A web application was also designed during the course of this study to visualise the cumulative distribution of PAT-seq and multiplex Poly(A) Test (mPAT) poly(A) tail length data. The mPAT method is a targeted multiplexed PCR approach that can measure the 3' UTR state of up to ~100 genes simultaneously. Also described is *3Primer*, a tool that was generated to automate the process of selecting a gene-specific forward primer for 3' Rapid Amplification of cDNA Ends (3' RACE) experiments, such as the mPAT method.

Chapter 2: Materials and methods

2.1 Assessing RNA metabolism in an increasingly metastatic mouse xenograft model of TNBC

Four cell lines with progressively increasing metastatic potential were studied in this thesis. From least to most metastatic they will be referred to as NI, LNA, LM2 and HM. These names are abbreviations of the 231_ATCC, 231_LNA, 231_LM2 and 231_HM.LNm5 lines referred to in the paper describing the functional and genomic characterisation of this model [195].

2.1.1 Generation of a mouse xenograft breast cancer model

The two least metastatic cell lines present in this study were the standard MDA-MB-231 cells (NI) and LNA, a moderately invasive line, which was derived from a late passage of the NI cell line. The MDA-MB-231 line was originally derived from a pleural effusion and was from a patient with disseminated metastasis caused by TNBC [196]. It exhibits an invasive profile *in vitro* but is poorly metastatic *in vivo* when used in mouse xenografts. Single cell populations derived from this line have exhibited differing metastatic potential despite all harbouring a similar poor prognosis gene expression signature [56].

The LM2 cell line used in this study was originally derived from the 1834 cell line, which was in turn derived from injecting MDA-MB-231 cells into the tail vein of a mouse and subculturing the resulting lung metastasis [197]. This culture was then expanded and reinoculated twice, with better lung targeting and metastatic ability each time. The LM2 line was then obtained by subculturing this second-round lung metastasis. The HM (highly metastatic) cell line was also originally derived from the MDA-MB-231 cell line [198]. Initially, 2 x 10^7 MDA-MB-231 cells were injected into the mammary fat pad (MFP) of anesthetised athymic mice. After 8-10 weeks, mice were sacrificed, and metastatic lesions were collected from lung metastases. The lesions were then minced, washed and underwent a cell culture-based process of selection for the tumour cells only. This procedure was then repeated 6 times in the same way with the HM cells subcultured after the 6th round. The HM tumours used in this work were further isolated from a lymph node metastasis generated in a previous study of glucocorticoid resistance in highly metastatic breast cancer [199].

Table 2.1. Tags added to MDA-MB-231 derived cell lines used in the MXM. Primary tumours (PTs) were FACS sorted based on the indicated marker. Table adapted from Johnstone *et al*. [195].

Tumour line	GFP vector	tdTomato vector	Luciferase Vector	PT sorted using
NI	pFB-neo_GFP	pBabe-BlaS_tdTomato	pBabe-puro_Fluc	tdTomato
LNA	pFB-neo_GFP	pBabe-BlaS_tdTomato	pBabe-puro_Fluc	GFP & tdTomato
LM2	45-TGL	pBabe-BlaS_tdTomato	45-TGL	GFP
HM	-	pBabe-BlaS_tdTomato	pBabe-puro_Fluc	tdTomato

The LNA, LM2 and HM lines were considered to be genetically identical to the parental NI line by short tandem repeat analysis (CellBank Australia). The xenograft tumours used in this study were generated by members of the Metastasis Research Laboratory (Olivia Newton-John Cancer Research Institute). To enable later *in vivo* imaging and FACS sorting of tumour cells from host cells, the cell lines were luciferase and fluorescently tagged respectively, as shown in Table 2.1. The pFB-neo_GFP, tdTomato, firefly luciferase and pBabe-BlaS_tdTomato expressing vectors have been described previously [200].

Tumour growth was initiated by inoculation of 1 x 10^6 tumour cells into the right-side inguinal mammary gland of BALB/c-SCID mice. BALB/c-SCID mice are immune suppressed, as they lack B, T, and natural killer cells, making them amenable to hosting human-derived tumours [201]. The LM2 and HM primary tumours were resected 18 days post injection, the LNA tumours were resected at day 21 and the NI tumours at day 72 (due to slow growth rate). There was no significant difference in average tumour weight among the 3 groups (ANOVA p = 0.3) [195]. To assess distant metastases, mice were maintained for an additional 3 weeks. The development of distant metastasis was evaluated by serial *in vivo* optical imaging, and by fluorescent stereo-microscopy of lung, liver, and spleen. Primary tumour cells were then FACS sorted from host cells and total RNA was isolated for PAT-seq library preparation.

2.1.2 PAT-seq and TCGA gene expression processing and gene set enrichment analysis

Sequence alignment of the PAT-seq gene expression data, peak calling (defining 3' UTR sites) and gene expression counting were performed by the 'Tail-tools' bioinformatics pipeline [169]. To obtain gene expression data from human tumours, gene expression read counts and patient information were downloaded from the TCGA using the GDC data portal. TCGA data is generated by the TCGA Research Network: http://cancergenome.nih.gov/. Triple negative tumours were selected based on clinical immunohistochemical data and separated by primary tumour stage, again based on clinical reporting. Tumours are classified into stages I-IV depending on how advanced they are at the time of observation, with a small tumour localised only to the breast classified as stage I and size and invasiveness increasing up to a tumour with distant metastasis being classified as stage IV [202]. The raw gene-wise counts from *Tail-tools* were then analysed using the *Limma-voom* approach to determine differential gene expression [203]. Gene set enrichment analysis was performed using *Camera* [204] (also implemented inside of the *Limma* software package), with gene sets taken from the molecular signatures database [205].

2.1.3 Cell culture and RNA extraction from the MXM cell lines

All tissue culture work was performed in a sterile cabinet under aseptic conditions. The 4 cell lines comprising the MXM and bottle of culture media were kindly provided by Cameron Jonhstone. For each cell line, a single vial containing 0.5 ml of cells was thawed each time a replicate was required, to ensure as much as possible that replicates were independent. Cells were cultured in Dulbecco's Modified Eagle medium (DMEM, Monash SOBS Media and Prep Services) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher Scientific), 1% Penicillin/Streptomycin (5,000 U/ml, Thermo Fisher Scientific) and 1% MEM Non-Essential Amino Acids Solution (Thermo Fisher Scientific). Cells were transported from LN_2 storage on dry ice prior to thawing. To thaw the cells, media was pre-warmed to 37°C and 5ml was added to a T25 flask. Cells were incubated at 37°C in 5% CO_2. Culture media was changed 12 hours after thawing. Cells were incubated for a further 24 hours before media was removed, then washed with 5ml sterile Phosphate-buffered saline (PBS) and trypsinised. $1.5x10^5$ cells were seeded into three T25 flasks containing 5ml of pre-warmed media. Cell counting was performed with the use of a haemocytometer. Cells were incubated for a further 48 hours prior to RNA extraction. The media was removed, cells were washed with 5ml of PBS and immediately dissolved in TRI Reagent (Sigma-Aldrich). At the time of RNA extraction cell lines were at passages 16,13,13 and 10 for NI, LNA, LM2 and HM respectively. RNA extraction was performed with the Direct-zol RNA mini kit (Zymo Corporation) as per the manufacturer's instructions.

2.1.4 Testing for miRNAs that were lost during metastatic transformation

BED files containing TargetScan [131] predictions for binding sites of miRNAs from miRNA families categorised as 'broadly conserved' (conserved across most vertebrates) were downloaded from the TargetScan website (http://www.targetscan.org/vert_71/). Both conserved (across vertebrates) and non-conserved miRNA binding sites were used in the analysis. Predicted sites were then lifted over from version hg19 of the human genome to version hg38 using the UCSC LiftOver Tool [206]. Topconfects analysis (Paul Harrison, https://github.com/pfh/topconfects) [207] of counts generated by counting reads from PAT-seq (run 2) of the NI and HM primary tumours mapping to APA sites from the APADB, resulted in 2037 genes that were suggested to have proximal APA in the HM tumour (by effect size as there was no significant APA in this comparison). Effect size as, part of the *topconfects* package, here refers to the switch to either more proximal or distal APA (-1 for a complete switch to proximal and +1 for a complete switch to distal), not moderated by a confidence interval. The names of the top 1000 of these genes were extracted from this dataset and a BED file containing regions that spanned the space between the two APA sites with the highest proportion of reads from these genes was then generated. The sites with the highest proportion of reads were used in an attempt to avoid false positive sites being used as the *topconfects* method is applied across all APA sites of a gene. The regions were then converted to GRanges objects [208] using R and overlapped (with strand considered) to determine the miRNAs that fell within the sites that underwent proximal APA. A hypergeometric test was used to detect over-represented miRNAs (as compared to a random selection of all potential miRNA binding sites), using the ‘phyper’ function of the R programming language [209] and resulting p-values were corrected using the Benjamini and Hochberg method for multiple testing correction [210].

2.1.5 The ePAT and TVN-PAT methods

The Extension Poly(A) Test (ePAT) and T-12 Variable Nucleotide Poly(A) Test (TVN-PAT) methods were developed in the RNA Systems Biology Laboratory and are low cost, PCR-based methods that offer a quick way to visualise the APA and the poly(A) tail distribution of specific genes. The ePAT method [211] uses an oligo-dT primer to bind to the end of the poly(A) tail and Klenow polymerase to extend the 3' end of the RNA on the bound DNA template to complete the adapter. TVN-PAT [212] uses variable nucleotide sequences at the end of the oligo-d(12)T primer to bind to the base of the poly(A) tail and fix poly(A) tail length to 12 bases. In both of these methods cDNA is then generated, PCR amplification is performed and the results are visualised on a gel. The poly(A) tail length distribution can then be inferred using a size ladder with the 12 (A) TVN-PAT sample as a standard reference. As the poly(A) tail appears as a smear on the gel, these methods are not limited by the size constraints or issues with the base calling of homopolymers present in short-read RNA sequencing technologies [213]. These methods are instead limited in the size specific resolution that they can provide, and it is also impossible to be certain that the RNA from the correct gene has been amplified without further testing.

2.1.6 The Extension Poly(A) Test (ePAT)

To perform an ePAT reaction [211] 200 ng-1 µg of total RNA from each sample was combined with 1 µl of the PAT-anchor primer (100 mM) in PCR strip tubes. Next, dH_2O was added as required to a total volume of 11 µl. The reaction was then heated to 80 ^oC for 5 min before being cooled to 37^oC in a PCR thermocycler (Applied Biosystems). To each tube 8 µl of a master mix was added, containing: 4 µl 5x Superscript III (SSIII) buffer (Invitrogen), 1 µl RNaseOUT (Invitrogen), 1 µl 10 mM dNTPs, 1 µl 100 mM DTT and 1 µl (5U) Klenow polymerase (New England Biolabs). The sample was then mixed by inversion 2-3 times, flash centrifuged and incubated at 37^oC for 30 min to perform the end extension reaction. The temperature was then raised to 80^oC for 10 min to inactivate the Klenow polymerase and then cooled to 55^oC for 1 min. Samples were maintained at 55^oC during the addition of 1 µl (200U) SSIII reverse transcriptase (RT, Life Technologies). To maintain sample temperature and avoid internal priming, mixing was performed by swirling the pipette tip and tapping the tubes without removing the tubes from the thermocycler. Samples were incubated at 55^{o}C for 30 min before the temperature was again raised to 80^{o}C for 5 min to inactivate the polymerase and then cooled to 12^{o}C. The resulting cDNA was then diluted by the addition of 100 µl of dH_2O. A 'no RNA' sample was included in all PAT type experiments as a negative control. A full list of all non-gene specific primer sequences used in the 'PAT' type methods can be seen in Table 2.4.

2.1.7 T-12 Variable Nucleotide Poly(A) Test (TVN-PAT)

Included alongside the ePAT as a control of known poly(A) tail length (12 A residues) is the TVN-PAT [212] sample. To generate TVN cDNA, 1 µg of pooled RNA from all samples is mixed with 1 µl TVN primer (100 µM) in a PCR tube and the total volume is made up to 12 µl with dH_2O. The structure of the TVN primer ($dT_{12}VN$; V is any nucleotide except T and N is any nucleotide) force the primer to bind only at the start of the poly(A) tail. Samples were incubated at 80^{o}C for 5 min in a PCR thermocycler (Applied Biosystems) before the temperature was reduced to 42^{o}C for 1 min. Maintaining the sample at 42^{o}C, 8 µl of a master mix was added, containing: 4 µl 5x Superscript III (SSIII) buffer (Invitrogen), 1 µl RNaseOUT (Invitrogen), 1 µl 10mM dNTPs and 1 µl 100 mM DTT. The sample was then mixed by inversion and flash spun before being incubated at 42^{o}C for another 1 min. 1 µl SSIII reverse transcriptase (200U) was then added to the sample, which was again mixed and flash spun. The sample was then incubated at 42^{o}C for 15 min and 55^{o}C for 15 min to complete the cDNA extension. Finally, the sample was incubated at 80^{o}C for 5 min to inactivate the reverse transcriptase and cooled to 12^{o}C. The resulting cDNA was then diluted by the addition of 200 µl of dH_2O.

2.1.8 PCR amplification of ePAT and TVN-PAT cDNA

The PCR reactions are the same for both ePAT and TVN-PAT cDNA. 5 µl of diluted cDNA is mixed with 15 µl of a master mix containing 0.2 µl of a 100 µM gene specific forward primer, 0.2 µl of the 100 µM PAT-anchor primer, 10 µl Amplitaq Gold 360 Master Mix (Life Technologies) and 4.8 µl dH_2O. The PCR program was as follows: 55ºC for 10 min, then cycling (95ºC for 20 sec, 60ºC for 20 sec, 72ºC for 30 sec) and finally 72ºC for 5 min. Samples were cycled between 25-30 times depending on the abundance of the target gene.

2.1.9 ePAT and TVN-PAT gels

The results of e-PAT and TVN-PAT reactions were run on 2% high-resolution agarose gels (Ultra pure 1000; Life Technologies) made with 1x Tris-borate-EDTA (TBE) Buffer (Thermo Fisher Scientific), pre-stained with SYBR safe (Life Technologies). Size quantification was performed alongside a 100 bp ladder (New England Biolabs), images were taken using an LAS 3000 imager and processed with ImageQuant TL software (IQTL ^{TM} V.7.0, Gene Expression Healthcare). Annotation of the gels was performed with Adobe Illustrator CS5.

2.1.10 The mPAT and alPAT methods

The core process of PAT-seq is quite labour intensive and is not necessary to interrogate smaller gene sets. In addition, the per-base sequencing coverage of PAT-seq may not be high enough to detect more subtle changes in APA or poly(A) tail length with certainty in genes with lower expression profiles. Furthermore, as PAT-seq is based on oligo-dT selection, it is unable to detect RNA lacking a poly(A) tail. Once desired RNAs have been identified through PAT-seq or other means (such as the literature), simplified targeted tools based on specific primers are therefore utilised to confirm these results or measure the expression of un-adenylated transcripts. These reverse transcription based, multiplexed, high throughput sequencing assays are known as the ‘multiplexed Poly(A) Test’ (mPAT, Figure 2.1) and the ‘adapter ligation Poly(A) Test’ (alPAT, Figure 2.2). The alPAT method works by the ligation of an adapter to the 3’ end of all RNA molecules present in a total RNA extraction using T4 RNA ligase 2 truncated K227Q. The adapter is pre-adenylated at the 5’ end [214] to allow for ligation to the 3’-OH group of a single-stranded sequence and is blocked with dideoxycytosine at the 3’ end to prevent circularisation. Next, cDNA is generated from the ligated RNA, and RNAs of interest are selected through a multiplex PCR approach, using gene-specific forward primers and a universal reverse primer that binds to the ligated adapter. A second round of PCR is then performed to attach Illumina adapters for sequencing. The mPAT method (Figure 2.1) [215] is similar to alPAT with the exception of the first step, in which adenylated RNAs are extended by dNTPs on an oligo-dT template (by Klenow polymerase), instead of using the adapter ligation method. The mPAT method has the advantage of being quicker to perform and more efficient in terms of retained RNA than the alPAT (due to the lower efficiency of the ligation process) but suffers from the drawback that it is limited to adenylated RNA.

2.1.11 The mPAT method

The Klenow polymerase extension and cDNA generation steps of the mPAT are almost identical to the ePAT except that the mPAT anchor oligo is used as the oligo-dT primer. As the mPAT is multiplexed, all gene-specific forward primers are pooled equally and are referred to as the ‘pooled forward primer mix’. All mPAT forward primers can be seen in Appendix A1, Tables A.1 and A.2. For the first PCR, 75 µl of a master mix, comprised of 1 µl pooled forward primer mix, 1 µl of the mPAT anchor oligo (to be used as a reverse PCR

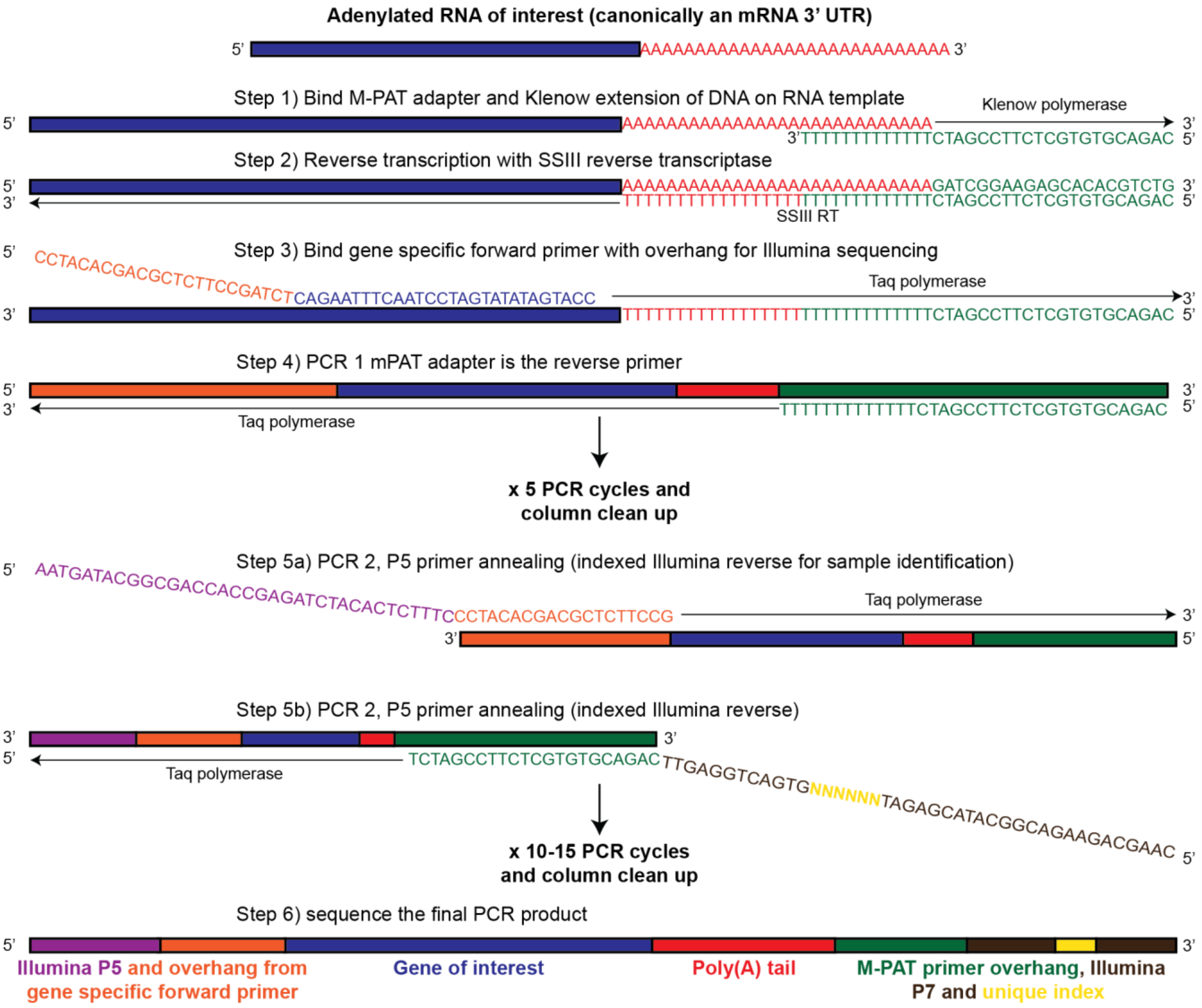

Adenylated RNA of interest (canonically an mRNA 3’ UTR)
5’ AAAAAAAAAAAAAAAAAAAAAAAAAAAA 3’
Step 1) Bind M-PAT adapter and Klenow extension of DNA on RNA template
Klenow polymerase
5’ AAAAAAAAAAAAAAAAAAAAAAAAAAAA 3’
3’TTTTTTTTTTTTTCTAGCCTTCTCGTGTGCAGAC 5’
Step 2) Reverse transcription with SSIII reverse transcriptase
5’ AAAAAAAAAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACACGTCTG 3’
3’ TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCTAGCCTTCTCGTGTGCAGAC 5’
SSIII RT
Step 3) Bind gene specific forward primer with overhang for Illumina sequencing
5’ CCTACACGACGCTCTTCCGATCTCAGAATTTCAATCCTAGTATATAGTACC
Taq polymerase
3’
3’ TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCTAGCCTTCTCGTGTGCAGAC 5’
Step 4) PCR 1 mPAT adapter is the reverse primer
5’ 3’
3’ TTTTTTTTTTTTTCTAGCCTTCTCGTGTGCAGAC 5’
Taq polymerase
x 5 PCR cycles and column clean up
Step 5a) PCR 2, P5 primer annealing (indexed Illumina reverse for sample identification)
5’ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG
Taq polymerase
3’
3’ 5’
Step 5b) PCR 2, P5 primer annealing (indexed Illumina reverse)
3’ 3’
5’ TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNTAGAGCATACGGCAGAAGACGAAC 5’
Taq polymerase
x 10-15 PCR cycles and column clean up
Step 6) sequence the final PCR product
5’ 3’
Illumina P5 and overhang from gene specific forward primer
Gene of interest
Poly(A) tail
M-PAT primer overhang, Illumina P7 and unique index

Figure 2.1. The mPAT method. The mPAT method involves the selection of polyadenylated RNAs through oligo-dT priming and Klenow polymerase-mediated 3’ end extension. Specific genes are then selected through the first round of PCR with a multiplexed combination of gene-specific forward primers. A second round of PCR is then performed to attach Illumina adapters for second-generation sequencing [215].

primer), 50 µl Amplitaq 360 master mix and 23 µl dH_2O was added to fresh PCR tubes. 25 µl of diluted cDNA was then added to each PCR tube. Samples then underwent 5 cycles of PCR using the same program as the ePAT method. Amplicons were then purified using NucleoSpin columns (Macherey-Nagel) as per manufacturer's instructions with a 50% dilution of the binding buffer with dH_2O to remove leftover primers. Samples were eluted from the columns in 49 µl of dH_2O, pre-warmed to 60^oC. The eluted cDNA was mixed with, 50 µl Amplitaq 360 master mix, 1 µl PAT-seq universal forward primer and 1 µl Illumina indexed reverse primer. The eluted cDNA was then amplified in a second round of PCR, which was run for 10-15 cycles depending on the abundance of the genes being amplified. Samples were then pooled before a final column clean up, again using 50% diluted binding buffer. Samples were eluted in the same way using 5 µl of dH_2O per included sample. To check for PCR product, 5 µl of the pooled samples was mixed with 5 µl of DNA loading buffer and then run on the same gel setup as was used for the ePAT. Samples were then handed to the Micromon sequencing facility for sequencing on the Illumina MiSeq. Standard PAT-Seq data processing was then performed using the *Tail-tools* bioinformatics pipeline.

2.1.12 Primer concentrations and PCR cycles tested during mPAT optimisation

Due to issues with the mPAT method amplifying low abundance genes in Chapter 3, some optimisation of the protocol was required. These tables describe the primer concentration and PCR cycle numbers used to optimise the mPAT method. Once the mPAT had been optimised, all further mPATs were performed as described in the previous section, with the exception that a primer concentration of 0.1 µM for each primer in the forward primer mix (i.e. more primer mix with an increasing number of primers) and 0.5 µM of mPAT anchor oligo was used in PCR 1.

Table 2.2. Optimising mPAT primer concentrations. Mixes shown correspond to lanes in Figure 3.18 D.

Mix #	Primer	Individual primer concentration in forward mix	mPAT reverse primer concentration
1	Full mix	0.1 M	0.1 M
2	Full mix	0.1 M	0.5 M
3	Full mix	0.05 M	0.1 M
4	Full mix	0.05 M	0.5 M
5	Full mix	0.0164 M	0.1 M
6	Full mix	0.00164 M	0.1 M
7	GAPDH	0.1 M	0.1 M
8	Full mix (no cDNA)	0.1 M	0.1 M

Table 2.3. Optimising mPAT PCR cycle number. Mixes correspond to lanes in Figure 3.18 E.

Program #	Primer	Individual primer concentration in forward mix	mPAT reverse primer concentration	Number of PCR cycles (PCRs 1/2)
1	Full mix	0.1 M	0.5 M	7 and 17
2	Full mix	0.1 M	0.5 M	10 and 17
3	Full mix	0.0164 M	0.1 M	10 and 17
4	Full mix	0.0164 M	0.1 M	10 and 25
5	GAPDH	0.1 M	0.1 M	5 and 15
6	Low abundance mix	0.1 M	0.1 M	5 and 17
7	Full mix (no cDNA)	0.1 M	0.5 M	10 and 25

2.1.13 The alPAT method

As the adapter ligation process is less efficient than oligo-dT based methods, 1 µg of total RNA was used as input to the alPAT and the volume was brought to a total of 11 µl with dH_2O. The mixture was heated to 80°C for 10 min before being immediately placed on ice. A master mix was added to each sample containing: 2 µl of the alPAT ligation oligo (20 µM), 2 µl 10 x T4 RNA Ligase truncated K227Q reaction buffer, 1 µl T4 RNA Ligase 2 truncated K227Q and 4 µl 50% PEG 8000. The ligation reaction was then mixed, flash spun and left to

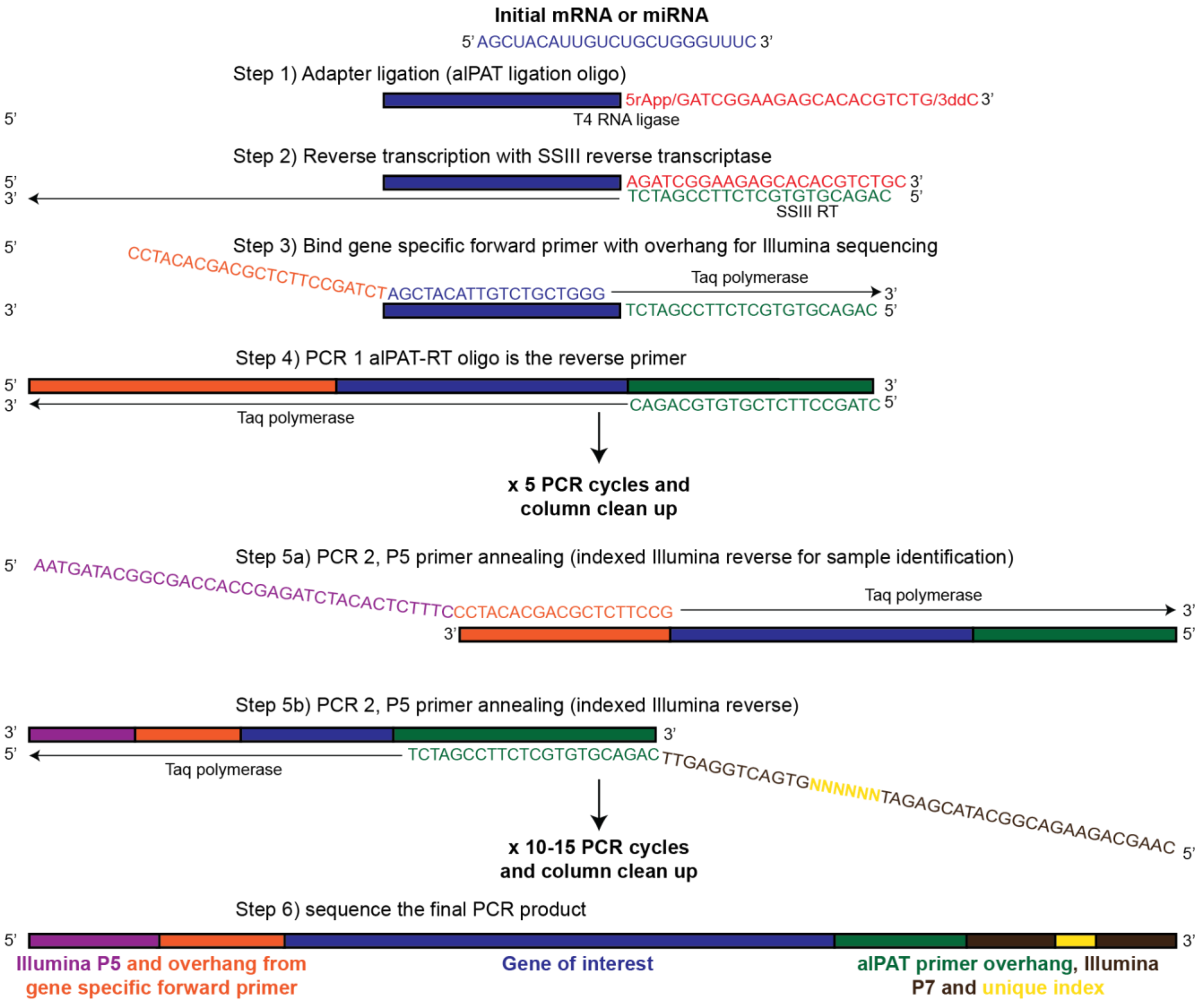

Initial mRNA or miRNA
5’ AGCUACAUUGUCUGCUGGGUUUC 3’
Step 1) Adapter ligation (alPAT ligation oligo)
5’
5rApp/GATCGGAAGAGCACACGTCTG/3ddC 3’
T4 RNA ligase
Step 2) Reverse transcription with SSIII reverse transcriptase
5’
3’
AGATCGGAAGAGCACACGTCTGC 3’
TCTAGCCTTCTCGTGTGCAGAC 5’
SSIII RT
Step 3) Bind gene specific forward primer with overhang for Illumina sequencing
5’
3’
CCTACACGACGCTCTTCCGATCTAGCTACATTGTCTGCTGGG
Taq polymerase
3’
TCTAGCCTTCTCGTGTGCAGAC 5’
Step 4) PCR 1 alPAT-RT oligo is the reverse primer
5’
3’
3’
CAGACGTGTGCTCTTCCGATC 5’
Taq polymerase
x 5 PCR cycles and column clean up
Step 5a) PCR 2, P5 primer annealing (indexed Illumina reverse for sample identification)
5’ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG
Taq polymerase
3’
3’
5’
Step 5b) PCR 2, P5 primer annealing (indexed Illumina reverse)
3’
5’
3’
TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNTAGAGCATACGGCAGAAGACGAAC 5’
Taq polymerase
x 10-15 PCR cycles and column clean up
Step 6) sequence the final PCR product
5’
3’
Illumina P5 and overhang from gene specific forward primer
Gene of interest
alPAT primer overhang, Illumina P7 and unique index

Figure 2.2. The alPAT method. An adapter is ligated to the 3’ ends of all RNA molecules. Reverse transcription is then primed from this adapter. This method, therefore, enables the sequencing of all RNA molecules in a sample, including those that do not have a poly(A) tail. Specific RNAs, including mRNAs and miRNAs, are then selected through the first round of PCR with a multiplexed combination of gene-specific forward primers. A second round of PCR is then performed to attach Illumina adapters for second-generation sequencing.

incubate at 16ºC for at least 12 hours. A column clean-up was then performed with the zymo RNA clean and concentrator -5 kit. Binding buffer was diluted 1:1 with ethanol to remove excess adaptor. Samples were eluted in 13 µl dH_2O. Reverse transcription was then performed in the same manner as was performed in the mPAT method (post Klenow extension), with the exception that the alPAT-RT oligo was used in place of the mPAT anchor oligo. Two rounds of PCR with intervening column clean ups were then performed in an identical manner to the mPAT with the exception that once again the alPAT-RT oligo was used as the reverse primer in PCR 1. Sequencing was also performed and analysed consistent with the mPAT method.

Table 2.4. Sequences of the primers used in the 'PAT' type methods.

Primer name	**Sequence**
PAT-anchor primer (ePAT)	GCGAGCTCCGCGGCCGCGTTTTTTTTTTTT
TVN primer (TVN-PAT)	GCGAGCTCCGCGGCCGCGTTTTTTTTTTTTVN; V: G, A, or C; N: any
alPAT ligation oligo	/5rApp/GATCGGAAGAGCACACGTCTG/3ddC/
alPAT-RT oligo	CAGACGTGTGCTCTTCCGATC
mPAT anchor oligo	CAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTT
mPAT gene-specific forward	CCTACACGACGCTCTTCCGATCTnnnnn-gene-specific-nnnnn
PAT-seq universal forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA-CGACGCTCTTCCG
Illumina indexed reverse	GATCGGAAGAGCACACGTCTGAACTCCAGTCACnnnnnn-ATCTCGTATGCCGTCTTCTGCTTG

2.2 Inferring and analysing APA from public datasets

2.2.1 Obtaining TCGA data

RNA-Seq bam files were downloaded from the Genomic Data Commons [216] using the *GDC Data Transfer Tool* (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool). It is important to note that TCGA data is poly(A) selected [1]. Clinical and gene expression data were downloaded through Firebrowse (http://firebrowse.org/). Gene expression data was converted to Log2 TPM prior to use. A mean untransformed TPM value > 3 across all samples was required for a gene to be used in this analysis. All data processing and plotting was performed using R [209].

2.2.2 Counting reads mapping to polyadenylation sites in TCGA RNA-Seq data

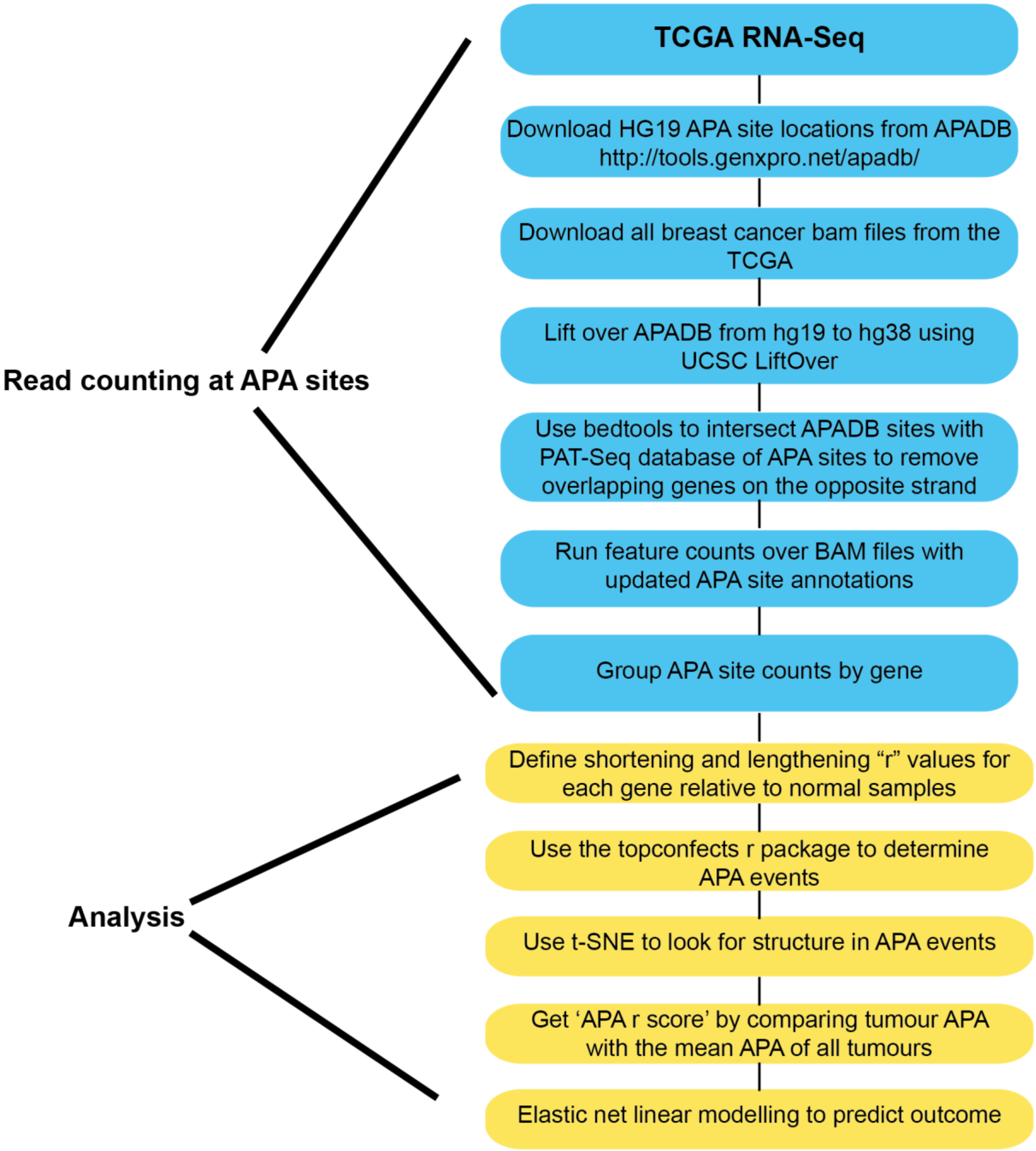
Primary tumours were used and tumours with an unknown stage were discarded from the TCGA dataset. Where the same tumour had been sequenced multiple times, only the first sequencing run was used. Tumours were classified as triple negative based on histopathological markers as lacking ER, PR and HER2 expression. To infer APA from RNA-Seq, a data base counting (DBC) method was employed. Instead of assignment to standard genomic features, reads are instead counted at known polyadenylation sites obtained from the APADB *Homo sapiens* (v2) [181]. BAM files had previously been generated by alignment to the human genome (version hg38) prior to download. APA site information was lifted over from genome version hg 19 to hg 38 using the UCSC *LiftOver Tool* [206]. Reads mapping to APADB sites were counted using *featureCounts* with unstranded and paired-end settings enabled [217]. For an APA site to be kept for further analysis, a mean read count > 20 reads was required for a particular site across all samples.

2.2.3 Limiting the false positive error rate when inferring APA from TCGA data

As TCGA data is unstranded, false positive errors can occur where the ends of two opposing genes overlap. In order to reduce these errors, a PAT-seq [169] database of directional, 3'-focused RNA sequencing data was additionally generated. PAT-seq was performed on 20 MDA-MB-231 cell line derived primary tumour xenografts that were generated in immunosuppressed mice. Based on the sequencing coverage from this data, a database of APA sites was generated. For an APA site to be called, at least 10 reads were required to map to a given 100 base span of the hg38 genome across all 20 samples. The PAT-seq APA sites had good concordance with the APADB with 50% of sites from the PAT-seq experiment overlapping sites from the APADB. APADB sites were then removed if they overlapped a PAT-seq site that had been called on the opposite strand and the counting process was repeated as described previously.

2.2.4 Inferring APA from paired TCGA breast cancer RNA-Seq data

Once APA counts were obtained, the *topconfects* R package [207] was again used in 'group effect shift' mode on paired tumour/normal samples to determine the APA events significantly implicated in tumour induction. The 'confect score' is the minimum effect size that one can have confidence in at an FDR of 0.05. The 'effect size' is defined as -1 for a complete shift to proximal APA and +1 for a complete shift to distal APA. 'APA r scores' were also calculated by comparing the mean distal usage state of all normal breast tissue samples with each individual sample in the TCGA. This method was also developed by Dr Paul Harrison and is distinct from the *topconfects* method. It works by comparing every read for a single gene in the tested sample, with every read in the same gene from the reference sample and determining the proportion that are upstream or downstream. Upstream reads are given a score of -1 and downstream reads are given a score of +1. These results are then averaged



Read counting at APA sites
TCGA RNA-Seq
Download HG19 APA site locations from APADB http://tools.genxpro.net/apadb/
Download all breast cancer bam files from the TCGA
Lift over APADB from hg19 to hg38 using UCSC LiftOver
Use bedtools to intersect APADB sites with PAT-Seq database of APA sites to remove overlapping genes on the opposite strand
Run feature counts over BAM files with updated APA site annotations
Group APA site counts by gene
Analysis
Define shortening and lengthening "r" values for each gene relative to normal samples
Use the topconfects r package to determine APA events
Use t-SNE to look for structure in APA events
Get 'APA r score' by comparing tumour APA with the mean APA of all tumours
Elastic net linear modelling to predict outcome

Figure 2.3. Steps in defining and analysing APA from TCGA data. There were two main steps in the novel methods employed in this study. The first (shown in blue) involved counting reads at known APA sites. The second (shown in yellow) was the novel *topconfects* based method for calling APA vs the mean APA state of matched normal samples. The t-SNE of 'APA r scores' defined in tumours versus the mean APA state of normal breast tissue APA and ENLM using 'APA r scores' defined in tumours versus the mean APA state of all tumours to predict outcome.

into the ‘r score’, named as such as it falls between -1 and +1. See Appendix A2 for the full description of 3’ end shifting measurement.

2.2.5 Panther overrepresentation analysis

PANTHER (**P**rotein **A**nalysis **TH**rough **E**volutionary **R**elationships, http://pantherdb.org) was used to determine over-represented gene sets with default settings [218]. In all instances a list of *Homo sapiens* input gene symbols was provided and run against PANTHER’s internal gene list. All p-values presented are Bonferroni corrected.

2.2.6 Grouping samples by gene expression and APA t-SNE

Samples were separated by the t-SNE method [219] implemented in the *Rtsne* R package [220]. Either gene expression values (in CPM) or ‘APA r scores’ (calculated by comparing the APA site use of each sample to the mean APA site usage of all normal samples in the TCGA, for the method see Appendix A2) were used as input. The theta value was set to zero, which meant that there was no trade-off of accuracy for speed by the algorithm, resulting in an exact t-SNE. The number of iterations was set to 2500 (1500 more than the default 1000) to further increase accuracy. All other parameters were left with their default settings.

2.2.7 Survival analysis using all TCGA tumours

Kaplan-Meier plots were generated using the *survival* R package [145]. Samples were grouped into good or poor prognosis groups based on the median of the prediction score (model link score, see Section 4.2.12). Samples with a higher score were considered to have a poor prognosis and vice versa. The relative predictive power of each model was measured by the D index, which is computed via a Cox model operating on the scaled rankits of the risk scores, rather than the base risk score themselves, resulting in more accurate and comparable models [146].

2.2.8 Generation of a complete clinical, APA and gene expression predictive model

To generate the complete model to use for future predictions, all 3 predictor sets were used as input for elastic net linear modelling (ENLM) using the *glmnet* R package [147] with the best α parameter (α = 0.37) determined by 10-fold CV. The best λ value was the minimum λ determined by *glmnet*. As with the 10-fold CV, clinical predictors were only penalised half as much as APA and gene expression predictors (penalty factor = 0.5).

2.2.9 Testing the reproducibility of the complete model using bootstrapping

A random sampling (with replacement) of the full 1033 TCGA samples was generated. This process was repeated 1000 times with *glmnet* ENLM [221] run on each repeat with clinical data, APA and gene expression as input. The λ and α parameters were reused and were the same as was used for generation of the complete model.

2.2.10 Calling APA from microarrays using the original probeset ordering method

The original probeset ordering (OPO) method (see Section 4.2.3) was used to call APA from Affymetrix HG-U133A microarrays, but could reasonably be applied to any array of the U133 generation or any chips that have probesets concentrated at the 3’ end of a transcript. 429 CEL files were downloaded from the GEO: Two groups of normal breast tissue samples were downloaded from GSE9574 and GSE20437 (53 total samples), TNBC samples (combined from many different studies) were downloaded from GSE31519 (259 total samples) and ER+ breast cancers (117 total samples) were downloaded from GSE2034. CEL files were read

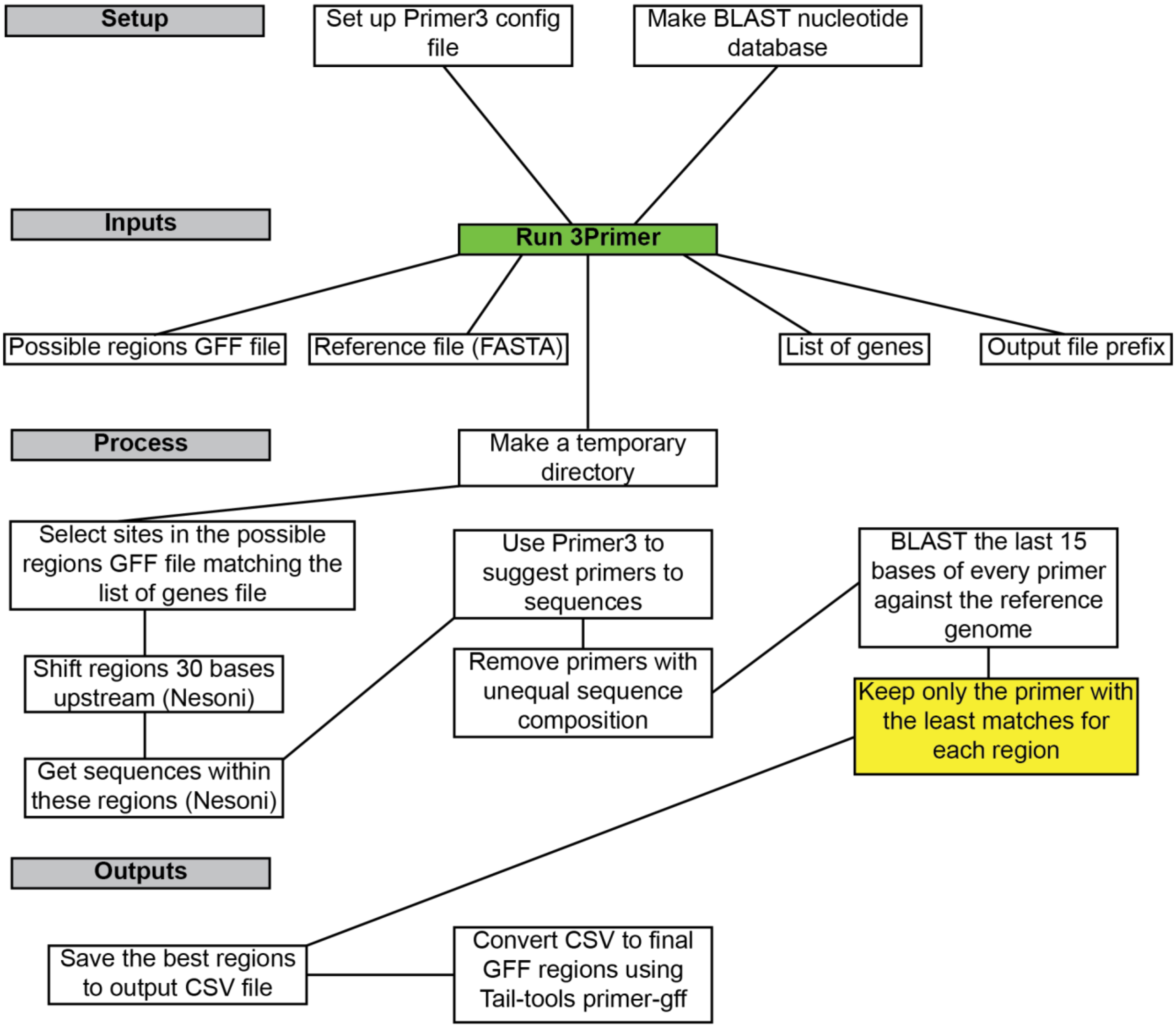

Setup
Set up Primer3 config file
Make BLAST nucleotide database
Inputs
Run 3Primer
Possible regions GFF file
Reference file (FASTA)
List of genes
Output file prefix
Process
Make a temporary directory
Select sites in the possible regions GFF file matching the list of genes file
Shift regions 30 bases upstream (Nesoni)
Get sequences within these regions (Nesoni)
Use Primer3 to suggest primers to sequences
Remove primers with unequal sequence composition
BLAST the last 15 bases of every primer against the reference genome
Keep only the primer with the least matches for each region
Outputs
Save the best regions to output CSV file
Convert CSV to final GFF regions using Tail-tools primer-gff

Figure 2.4. Schematic of the functions of *3Primer*. The steps taken by *3Primer* in order to select the most specific 3’ RACE primer. This tool was created using the Python programming language (version 2.7) and is designed for use in a Unix environment.

into R using the *affy* bioconductor package [174]. Probesets were normalised using the frozen robust multiarray analysis (fRMA) method [175].

2.3 Novel bioinformatics methods

Throughout the course of this thesis it was necessary to generate novel tools that assisted non-computational researchers in interpreting and exploring PAT-Seq datasets. It was also useful to automate some repetitive tasks, such as primer design for multiplex NGS experiments, for greater efficiency. These tools are broadly described in Chapter 5, with their more technical components outlined here.

2.3.1 Components of *3Primer*

The *3Primer* tool was created to enable easy generation of highly specific mPAT primers that also meet all thermodynamic requirements. A schematic of the steps taken by *3Primer* in specific 3' RACE primer selection can be seen in Figure 2.4. Prior to running *3Primer*, the settings for *Primer3* must be saved in 'primer config.txt'. Settings may be chosen on a user's desired PCR parameters or the default settings may be used. There are 4 main inputs to *3Primer*: A GFF file containing genomic locations that primers may be made against (must match the reference genome), a reference genome file (in FASTA format), a list of genes for which forward primers will be suggested, and the prefix for all output files. *3Primer* first makes a temporary sub-directory in the directory from which it was called. It then filters in the input GFF file keeping only the regions related to genes present in the 'list of genes' input file. Using the Nesoni bioinformatics suite [222], genomic regions are shifted 30 bases upstream and the sequences spanning these regions are obtained from the reference genome. These sequences are then passed to Primer3 [223] for the generation of possible primers using the settings in the primer config.txt file. Possible primers that have been returned with unequal sequence composition in the last 16 bases are then removed to prevent mispriming to homopolymeric sequences. The last 15 bases of every primer sequence are checked for complete matches against the reference genome using the command line version of the 'Basic Local Alignment Search Tool' (*BLAST*) [224]. The most specific primer for each region is saved to a comma-separated values (CSV) file. Finally, the CSV file is converted to a GFF that suggests the regions that the primer will amplify using the *Tail-tools* primer-gff function [169].

2.3.2 RNA Systems Explorer App code base and implementation

The RNA Systems explorer app (*RSER*) was designed for examination and quality control of poly(A) tail length changes in PAT-Seq and other 'PAT' type experiments. The full code for the *RSER* can be found at *https://github.com/AndrewPattison/PAT-seq-explorer*. The majority of the app is implemented in R, however, some JavaScript code was used to alter the browser refresh commands in the standalone version of the app used in an upcoming C.*elegans* paper (Boag *et al*., manuscript in preparation, see preface for full title). The app was used to make figures 2C, 5C and 6B in the main manuscript, and in the supplement SFig 1, SFig 3D and SFig 5C. The app follows the usual structure of a shiny app and is comprised of 3 main files; 'ui.R' contains the shiny code to manage the interface to the app including inputs and outputs, 'server.R' takes the inputs from 'ui.R' and generates the outputs to be displayed (Figure 2.5). The main functions for plotting cumulative distributions are contained in 'helper.R' and are called by 'server.R' when the app is first started.

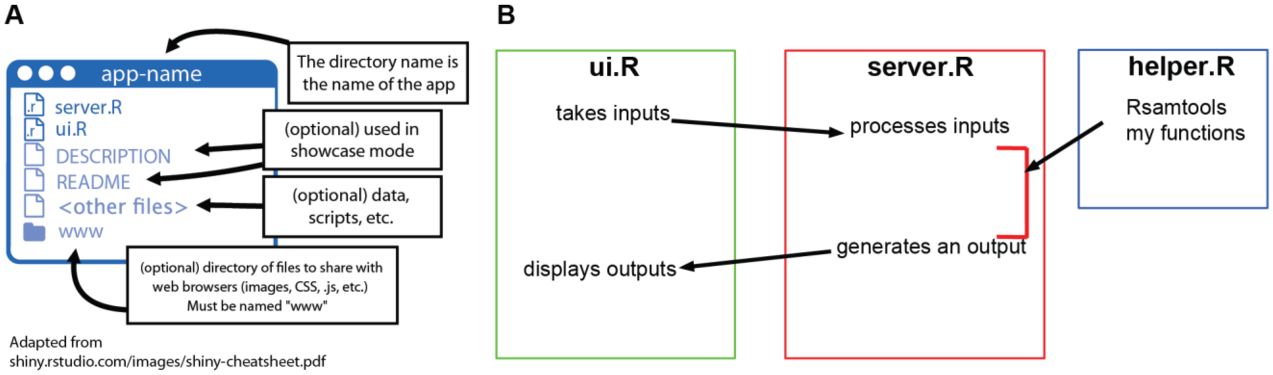

A
app-name
server.R
ui.R
DESCRIPTION
README
<other files>
www
The directory name is the name of the app
(optional) used in showcase mode
(optional) data, scripts, etc.
(optional) directory of files to share with web browsers (images, CSS, .js, etc.) Must be named "www"
Adapted from shiny.rstudio.com/images/shiny-cheatsheet.pdf
B
ui.R
takes inputs
displays outputs
server.R
processes inputs
generates an output
helper.R
Rsamtools
my functions

Figure 2.5. The structure of a shiny app. A. The common components of a shiny app. **B.** The flow of information from user inputs to displayed outputs in a shiny app.

Chapter 3: Changes to RNA metabolism in an increasingly metastatic model of TNBC

3.1 Introduction

TNBC represents the least well-understood breast cancer subtype and is complicated by limited treatment options, as these tumours lack overexpression of the ER and HER2 receptors normally targeted in endocrine therapies. Adjuvant chemotherapy is currently the standard of care (and only option) for almost all TNBC patients and, encouragingly, with this type of treatment TNBC has a better prognosis than many other forms of breast cancer [165, 12]. Early detection has also meant that more breast cancers, including TNBCs, are being detected at earlier stages, improving outcome [4]. Unfortunately, it is unclear whether a patient given cytotoxic chemotherapy for a TNBC would have ever required it had the tumour not been discovered, and although there are currently multiple gene expression-based molecular tests for the outcome of ER+ breast cancer [55, 56, 44], none of these tests are approved for use on the TNBC subtype [225]. Additional prognostic markers for the TNBC subtype are therefore required to give patients more certainty about the outcome of treatment decisions.

With genetic variants and gene expression having limited explanatory power in understanding tumour outcomes, the search for novel predictors has shifted recent focus toward mechanisms of post-transcriptional regulation, such as miRNAs and APA. Studies on the dysregulation of miRNAs in breast cancer have found altered expression of many miRNAs specific to breast cancer subtype [226, 1]. APA-mediated repression of translation through binding to *cis*-regulatory elements in the 3’ UTR of mRNAs is the best-known example of post-transcriptional mRNA regulation in proliferating and cancer cells [112, 109]. Understanding the state of miRNA expression in conjunction with APA is therefore important in understanding if APA changes are associated with changes to miRNA expression or if they moderate miRNA binding independently.

As discussed in Chapter 1, there are conflicting reports on the effect of the length of the poly(A) tail on gene expression. In the cancer setting, higher levels of polyadenylate polymerase (PAP) activity has been associated with worse prognosis in lymph node negative breast tumours [227], suggesting a possible role for poly(A) tail length in tumour progression. Other than this example, no studies could be found where poly(A) tail length was suggested to directly play a role in cancer. This is likely due to difficulties in accurately measuring homopolymeric DNA sequences and the common dogma that the poly(A) tail is unimportant for translation once it is longer than a species-specific minimum level.

As previously discussed, the addition of the poly(A) tail can occur at multiple alternative cleavage sites, altering the 3’ UTR [106], and it has been demonstrated that proliferating and tumour cells express shorter 3’ UTRs with fewer *cis*-regulatory elements [112, 109]. A recent study of APA in cancer inferred tumour-specific patterns of APA from RNA-Seq data across 7 cancer types from the TCGA, finding 1,346 cancer-associated dynamic APA events. In general, 3’ UTR shortening was found to be associated with tumourigenesis and highly prognostic of tumour outcome, with the APA factor CstF64 (*CSTF2*) suggested to be a master regulator of APA in tumours [66]. This analysis represented a good general survey of APA in the 7 cancer types but was not extended to specific subtypes. TNBC has previously been suggested to have the most APA events of any breast cancer subtype [110], suggesting that the study of APA in primary TNBCs may provide novel targets for therapy, or have some utility in the prediction of outcome.

There have previously been three other studies of APA in patient-derived TNBCs, with all three performing reanalysis of existing microarray data [113, 110, 65]. The studies by Akman *et al.* and Miles *et al.* suggested that 3’ UTR shortening events predominate in TNBCs when compared with mammary epithelial cells (normal breast tissue), and that shortening of the genes *SNX3*, *YME1L1* and *USP9X* was associated with poor outcome. Taking a more unbiased approach, the study by Wang *et al.* was not concerned with the direction of APA and simply attempted to stratify TNBC patients into high and low risk groups based on the APA profiling of 17 prognostic genes. Patients in the high-risk group were over 8 times more likely to die than the low-risk group. Although both Akman *et al.* and Wang *et al.* analysed a large number of tumour samples (520 and 327 respectively) [113, 65], the information provided by these microarray-based studies is limited, as APA can only be defined where probes are present at the 3' ends of transcripts. Moreover, this technology is subject to hybridisation artefacts such as the nonspecific binding of mRNAs to probes, as well as binding efficiency biases that may result in calling of APA that is in fact due to altered gene expression. Furthermore, the samples analysed in these studies were collated from a wide variety of sources, and completely accounting for batch effects would have been highly challenging. A more accurate method for identification and/or validation of APA events is therefore required before APA-based biomarkers are considered for clinical use.

The focus of this study was the RNA biology of primary breast tumour metastasis including gene expression, miRNA, APA and poly(A) tail length changes. This chapter describes a 3'-focused analysis across a panel of increasingly metastatic cell lines derived from the human MDA-MB-231 cell line that were used to generate a mouse xenograft model (referred to as the MXM). Previous sequencing of the MDA-MB-231 cell line by the SAPAS 3'-focused sequencing method suggested a trend toward distal APA presents in these cells when compared to an epithelial cell line [21], however, 3'-focused regulation has not previously been studied in a model of increasing metastatic potential. Data presented here were largely obtained from global 3'-focused PAT-seq experiments.

3.2 Results

3.2.1 PAT-seq analysis of a mouse xenograft model of TNBC metastasis

The aim of studying the MXM was to obtain a transcriptome-wide picture of the mRNA processing events that occur as the metastatic potential of a primary TNBC increases. In order to measure altered RNA metabolism in a controlled manner, a human in mouse xenograft model was generated by Dr Cameron Johnstone (Figure 3.1), consisting of four cell types derived from the MDA-MB-231 TNBC cell line. From least to most metastatic they were: NI (for non-invasive, the original MDA-MB-231 line), locally invasive (LNA), metastatic to the lung (LM2) and highly metastatic (HM). See methods Section 2.1.1 for full description of the generation of each cell line. Each of the increasingly metastatic TNBC cell lines were injected into the mammary fat pad of immune-suppressed NOD scid gamma (NSG) mice and a primary tumour was allowed to form. The primary tumour was then resected, and the mice were monitored for distant metastases. All cell lines were fluorescently tagged (Table 2.1) to allow the FACS separation of tumour cells from host tissue, obtaining a clean tumour sample for RNA extraction. This process was repeated in 2 batches termed run 1 and run 2. Run 1 contained 3 replicates each of the LNA, LM2 and HM lines. Run 2 was similar to run 1, but additionally included 2 replicates of tumours from the baseline NI cell line. Due to the poor growth rate of the NI line *in vivo*, only 2 biological replicates were obtained. As the cell lines were derived from the same initial line and sorted from host tissue, it was hypothesised that the majority of differences between the cell lines would be associated with the metastatic potential of the primary tumour.

PAT-seq was performed on total RNA extracted from each tumour line, yielding the gene expression, APA and poly(A) tail length of every polyadenylated transcript in a sample. The

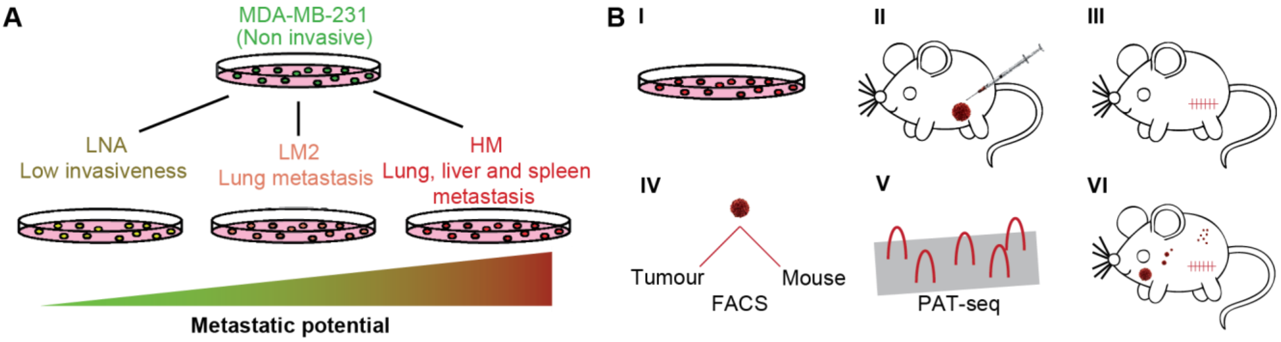


Figure 3.1*.* TNBC xenograft tumour model. A. The invasive cell lines used in this study (LNA, LM2 and HM) were derived from the MDA-MB-231 TNBC cell line (NI). **B I-III.** Each of the 4 cell lines was cultured and injected into the mammary fat pad of female NSG mice. **IV.** Primary tumours were resected after 4-6 weeks and FACS sorted from host cells to obtain a pure tumour sample. All cell lines were previously fluorescently marked to facilitate FACS sorting. **V.** RNA was extracted from primary tumours for PAT-seq RNA sequencing. **VI.** The mice were observed for tumour metastasis for a further 3 weeks before sacrifice.

resulting FASTQ files from the Illumina MiSeq were analysed using the *Tail-tools* software which was run using default settings with replicates and batches defined from a Python script. *Tail-tools* is a custom, in-house bioinformatics pipeline developed by Dr Paul Harrison that calls and visualises differential gene expression, APA and poly(A) tail length changes from PAT-Seq experiments. The final *Tail-tools* report also provides comprehensive information about gene expression, APA and poly(A) data in the form of exportable CSV files that can be further analysed. Regions where PAT-seq reads aligned to the human reference genome (version hg38) are referred to by *Tail-tools* as ‘peaks’ and represent the putative 3’ ends of a transcript. Global gene expression can be measured by PAT-seq data by aggregating the peak counts from each gene. After filtering for lowly expressed genes (defined as < 10 reads per peak averaged across all samples), PAT-seq detected the expression of 17,803 unique genes.

Gene expression was first analysed to determine the extent that cell type and sequencing batch effects had on the PAT-seq results obtained. Batch effects were clear from the patterns of expression in each run when visualised as a heat map of gene expression (Figure 3.2 A) and a multidimensional scaling (MDS) plot (Figure 3.2 B). A batch term was always included in later statistical tests where both sequencing runs were used to account for this effect. Despite the clear batch effect between runs 1 and 2, PAT-seq gene expression was relatively consistent across both sequencing runs as shown by the MDS plot (Figure 3.2 B). Samples showed separation primarily on the basis of cell type and secondarily by batch effect, clearly indicating cell type as the main driver of gene expression variation. This observation was supported by principal component analysis (PCA), which was performed in R, with a Scree plot showing that the majority of the variance between samples could be explained by the first two principal components (Figure 3.2 C, 67% and 18% respectively). Combined, these results support the robustness of the PAT-seq data in determining the gene expression state

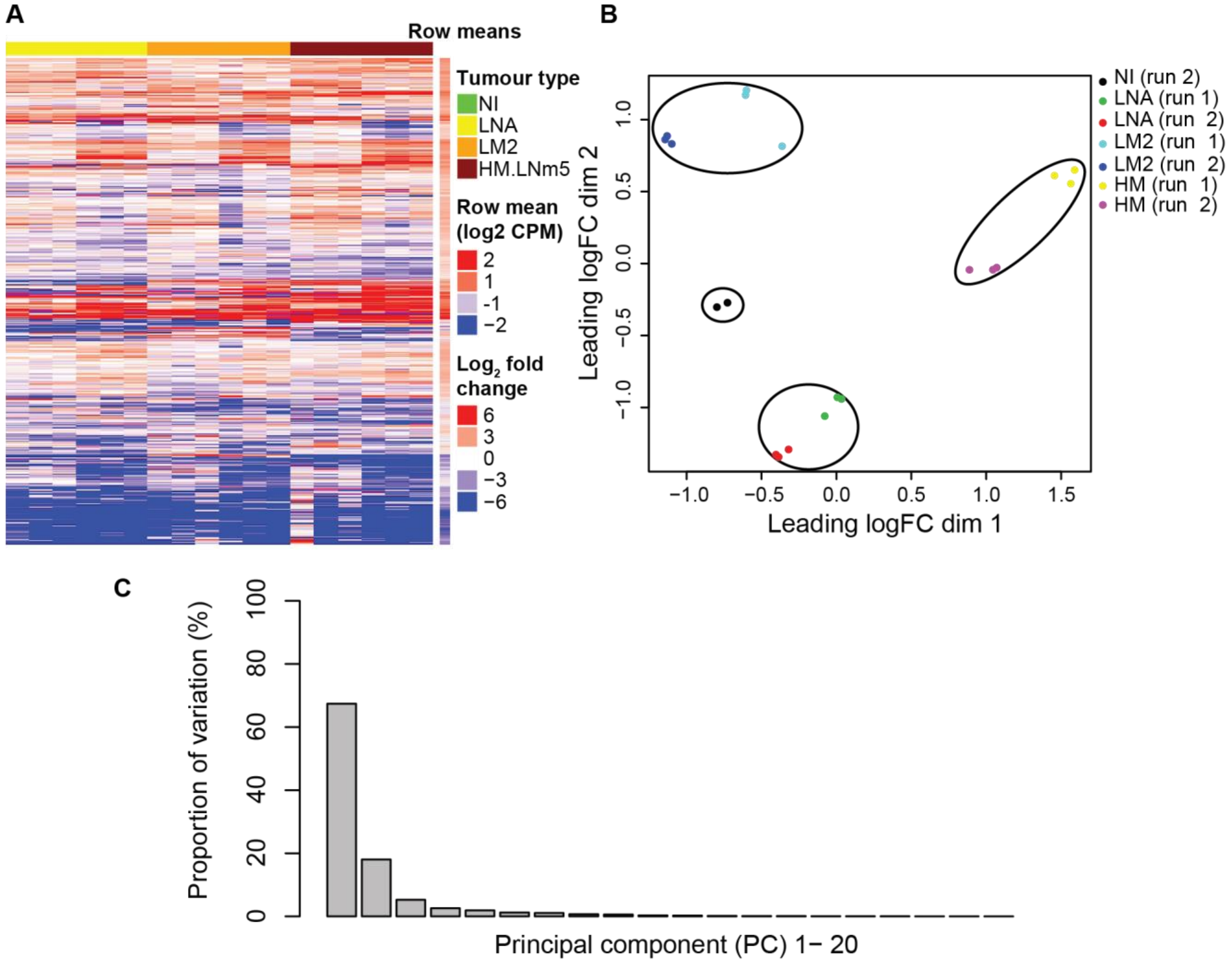


Figure 3.2. Detection of batch effects in PAT-seq data. A. Heat map showing genes with at least a 2-fold expression change from NI with a mean of ≥ 30 reads in sequencing run 2 (both runs are shown). Samples from run 2 are grouped to the left of the run 1 samples on the heatmap. **B.** Multidimensional Scaling plot (MDS) of all samples from both runs. MDS plots are implemented in the *Limma* R package [203] and approximately show the typical $\log_2$ fold changes between samples, grouping more similar samples closer together. Samples from the same cell line are circled. **C.** A Scree plot showing the proportion of variance explained by each principal component of the normalised MXM gene expression matrix (normalisation performed using the normalisation factors calculated by *Limma*).

of these cells and suggests that differences in cell type were the primary driver of observed variation in the MXM.

Before testing for metastasis-associated changes to APA, quality control of this aspect of the PAT-seq data was also performed. An example of PAT-seq coverage of the *COL1A1* gene showing two peaks as visualised in the integrative genomics viewer (*IGV*) browser [289] is presented in Figure 3.3 A. PAT-seq identified 49,425 peaks across the MXM model, of which 50% mapped to annotated 3' UTR regions of the human genome (RefSeq version hg38, Figure 3.3 B). This was similar to the 64% of APA sites that mapped to 3' UTRs in the APADB, a database of known mammalian 3' ends collected from multiple sources [181]. The APADB was chosen over the more popular PolyA_DB 2 [228], which relied on the reanalysis of expressed sequence tags, as it utilised targeted 3' sequencing that has higher coverage and better detection of novel APA sites. The location of the PAT-seq APA peaks also had good general concordance with the APADB, with 50% of PAT-Seq peaks overlapping peaks present within the APADB. Analysis of sequence motifs around PAT-seq peaks (Figure 3.3 D) revealed that the canonical polyadenylation signal (AAUAAA) was present within 100 bases upstream of a peak 38% of the time, as opposed to 10% in randomised versions of the same sequences ($p << 0.01$, Chi-squared test). Furthermore, a 1 base variant (1 mismatch allowed at any base) of the AAUAAA sequence was present in 91% of upstream regions as opposed to 68% of randomised sequences ($p << 0.01$, Chi-squared test). This was once again consistent with the APADB, in which 66% of APA sites from human whole blood samples contained a one base variant of the polyadenylation signal in the 50 bases upstream of the poly(A) site. Together these results suggest that PAT-seq effectively detected primarily true 3' ends of polyadenylated transcripts.

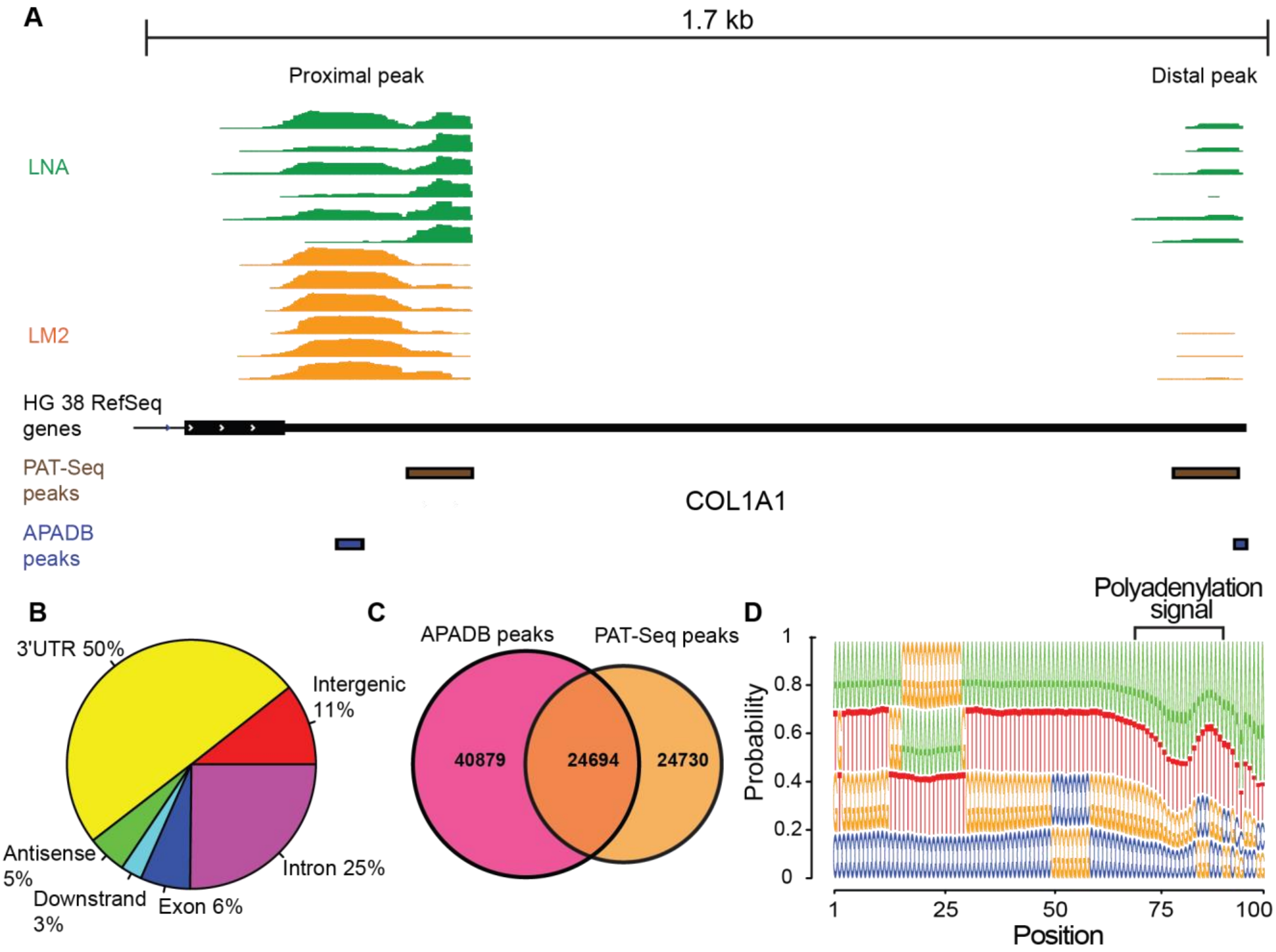


Figure 3.3. PAT-seq identification of APA sites. A. An example of PAT-seq sequencing coverage in the gene *COL1A1* from the LNA tumour line (green), vs the LM2 tumour line (orange). Shown below are the peaks (APA sites) called by *Tail-tools* (in brown) and for reference the APA sites present in the APADB (in blue). **B.** The distribution of PAT-seq reads across genomic locations. **C.** Number of overlapping APA sites between PAT-seq and the APADB. **D.** Sequence composition (by proportion of bases) of the 100 bases preceding all PAT-seq APA sites. A clear bias toward A (green) and T (red) bases can be seen in the 30 bases before the APA site, representing the polyadenylation signal (AAUAAA or a close variant).

3.2.2 Dynamic APA events are associated with metastasis

To determine metastasis-associated APA events in the MXM, the 3' UTR profile of the NI tumour line was compared with each of the other 3 lines using the *topconfects* R package [207]. In this method, developed by Dr Paul Harrison from the RNA Systems Biology Laboratory, each APA event is given an effect size score between -1 and 1, where -1 reflects a complete switch to proximal CPA, and 1 refers to a complete switch to distal CPA. Where this effect is statistically significant (FDR < 0.05 by default), an effect size that a researcher can be confident of is given, known as a 'confect' for confident effect size. See Appendix A2 for more detail on the *topconfects* method. The two NI replicates obtained from run 2 lacked sufficiently large APA switching events to overcome the variability of APA events in the MXM dataset. This meant that no statistically significant confident effects were found for the comparisons of this cell line to the others in run 2. The size of APA effects (the measure of proximal or distal APA without applying confidence bounds) was instead compared in an effort to make inferences about global APA shifts with increasing metastatic potential. The NI vs LNA comparison had an even spread of both proximal and distal APA events when comparing trends in the top 50 APA events by effect size (Figure 3.4 A). There were, however, clear trends toward distal APA in the moderately metastatic LM2 line (Figure 3.4 B) and proximal APA in the highly metastatic HM line (Figure 3.4 C). These trends suggested that APA events were present in this comparison but that the experiment may have lacked sufficient power to detect them. When APA effects from all genes from these comparisons (as opposed to the to 50) were overlaid (Figure 3.4 D) there was no clear APA trend in either direction, suggesting there were no metastasis-associated transcriptome-wide APA effects.

To determine if restricting the analysis to a single run reduced the statistical power to call APA events, a second *topconfects* analysis was performed using the cell lines analysed in

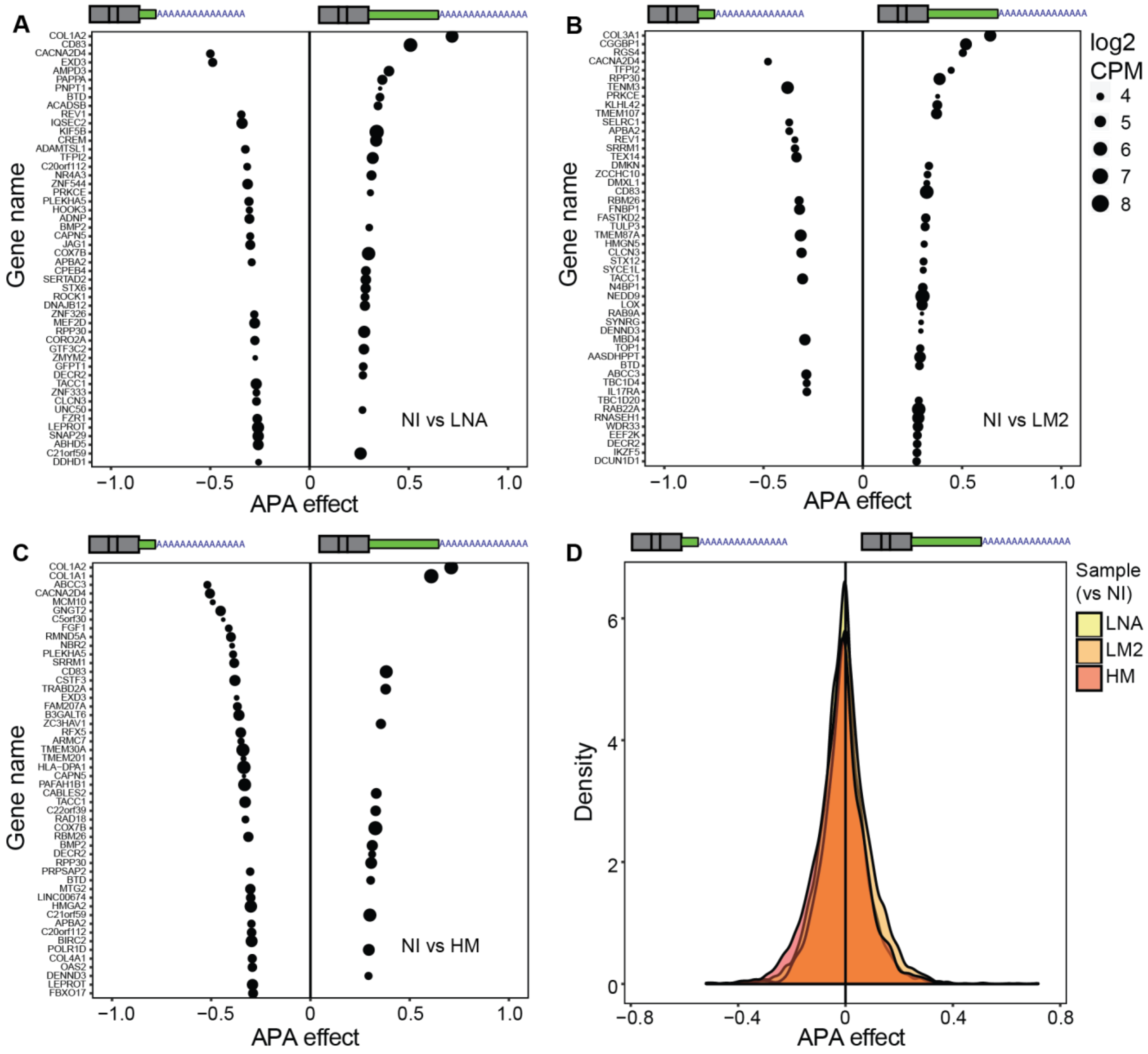


Figure 3.4. The general trend of APA events in PAT-seq run 2. A, B, C. APA comparison of the LNA (A), LM2 (B) and HM (C) cell line vs NI (N = 3, 3, 3 and 2 respectively). Values to the left of the line indicate a shift to proximal APA for the indicated gene and values to the right of the line indicate a shift to distal APA. APA effect sizes are plotted as no APA events were detected at FDR < 0.05 in any comparison. **D.** Density plot of the APA effect size values shown in A, B and C.

both runs (LNA, LM2 and HM, N=6). The analysis was repeated using LNA as a baseline as it was the least metastatic of the three. As these cell lines spanned both PAT-seq runs, a batch term was included to counter APA signals that may arise from differences between sequencing runs. When comparing the LM2 and HM tumour lines with the LNA line over both PAT-seq runs, 37 and 47 statistically significant APA effects were called respectively (Figure 3.5 A/B). There was once again a difference in the direction of these APA events, with a clear trend toward lengthening in the LM2 line (31 longer, 6 shorter) and shortening in the HM line (19 longer, 28 shorter, Figure 3.5 C). APA events largely occurred in different genes, with only 4 events common to both comparisons. These genes were *Ell2*, a Pol II elongation factor, *TPP1*, which has been suggested to be associated with the increased activity of telomerase, *PRMT2*, which has been suggested to regulate the cell cycle associated E2F group of transcription factors, and *TXNDC9*, the cell differentiation associated colorectal cancer prognostic marker [229, 230, 231, 232].

Despite the differences in APA genes with the greatest confects, there was a high overall correlation between APA effects in both the LM2 and HM lines (Pearson’s r = 0.95, Figure 3.5 D). This once again suggested that, despite the differences in significant APA changes, there may be similar overall APA changes with metastatic potential in both lines. The confect values reported were also quite low, suggesting considerable uncertainty in the underlying effect sizes. This uncertainty should be taken into account when interpreting the ranking of APA confects as it too may change if more data was available. Both lists of the 37 and 47 APA genes that were significant in the LM2 and HM comparisons with LNA were submitted for overrepresentation analysis to the PANTHER classification system [218], however, no significant GO terms were observed. Taken together, these results suggested that while APA is consistently altered in TNBC metastasis, it may not follow a defined pattern of proximal APA and does not appear to be restricted to certain gene expression pathways.

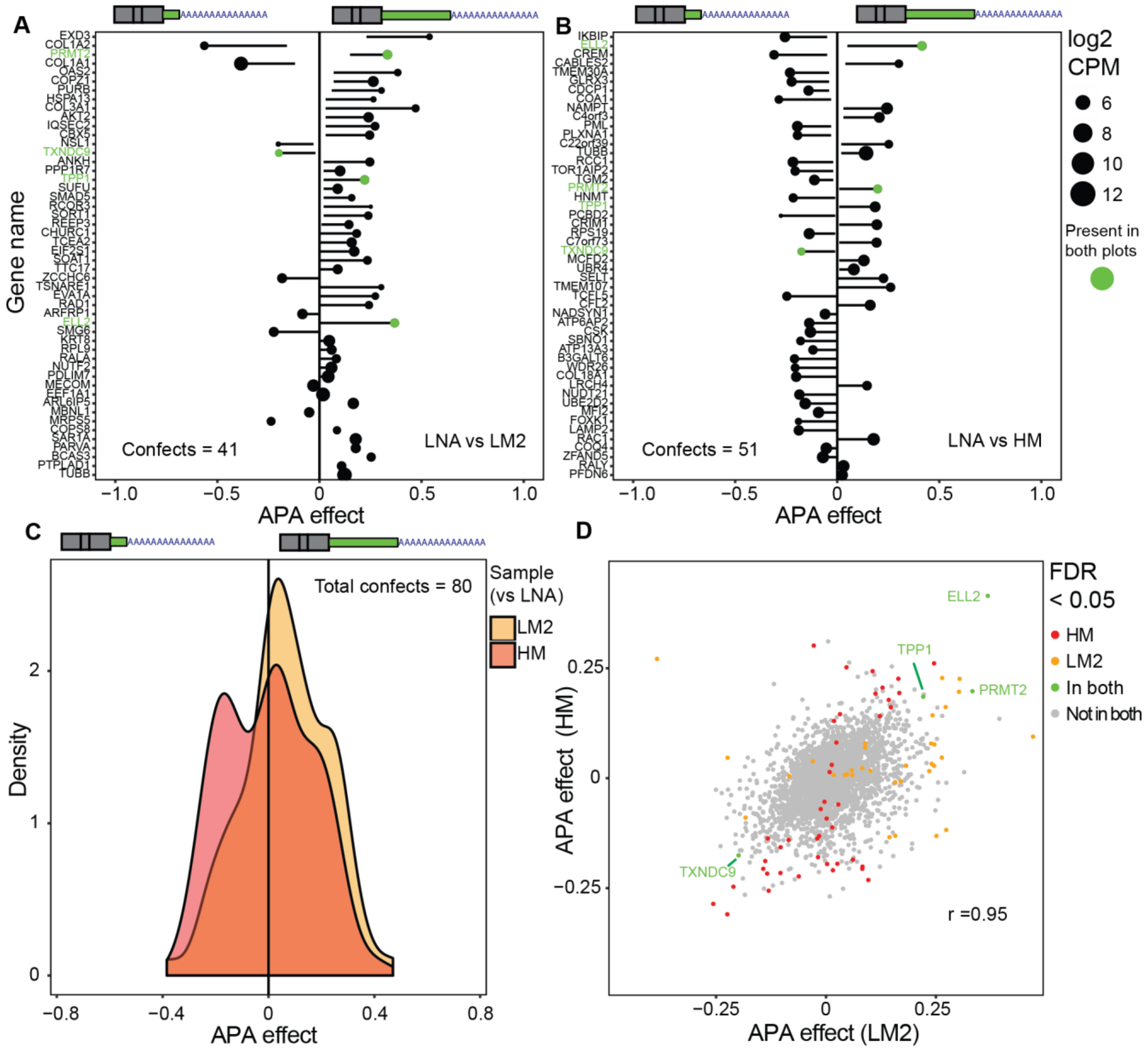


Figure 3.5. APA from both MXM PAT-seq runs shows variable APA trends with increasing metastatic potential. A. APA comparison of the LM2 cell line vs LNA (N = 6). 37 APA events were detected at FDR < 0.05. **B.** APA comparison of the HM cell line vs LNA. 47 APA events were detected at FDR < 0.05. For both A and B: APA effect (the position of the dot) and confect values (end of the line closest to 0) are displayed on the x-axis for every gene (y-axis) and are ordered by confect. Confects shared between comparisons are highlighted in green. The diameter of the effect size indicator point is proportional to the $\log_2$ CPM of the associated gene in all tested samples. **C.** Density plot of the APA effect values of the 80 confects called in A or B. A trend can be seen toward 3' UTR shortening in the LNA vs HM comparison and lengthening in the LNA vs LM2 comparison. **D.** Comparative scatter plot of all APA effects presented in A and B with significant values coloured as indicated. The two comparisons were highly correlated (Pearson's r = 0.95, p < 0.05).

3.2.3 APA is not associated with gene expression or poly(A) tail length

APA and poly(A) tail length have both been suggested to be associated with mRNA stability and translational regulation [112, 125]. To determine if there were any interactions between APA, gene expression or poly(A) tail length, APA was plotted against poly(A) tail length and gene expression in a gene-wise manner. Due to the low number of samples in the NI line, the LNA line was once again used as a baseline. Changes in gene expression were generally not associated with APA (Figure 3.6 A/B, Spearman's rho = 0.03 and 0.04 respectively), except in *COL1A1* and *COL1A2* in the LM2 lines. Both of these genes encode collagen subunits, which were also very highly expressed in the LM2 line relative to the LNA line. Poly(A) tail length change also showed no overall correlation with APA (Figure 3.6 C/D, Spearman's rho=0.09 and 0.14 respectively). These results suggested that APA generally operated independently of gene expression and poly(A) tail length but does not rule out the possibility of these systems being connected indirectly.

3.2.4 The 3' UTR binding sites and expression of miRNAs was altered with metastatic potential

As discussed in Chapter 1, it has previously been suggested that 3' UTRs are preferentially shortened in proliferating and cancer cells, causing the loss of miRNA binding sites [112, 109]. The 3' UTRs of APA genes in the MXM were therefore analysed to determine whether any miRNA binding sites were preferentially lost in the evolution of the non-invasive NI tumours to the highly metastatic HM tumours. The 3' UTRs of the top 1000 genes that underwent 3' UTR shortening in the HM cell line relative to the NI cell line were analysed for overrepresented binding sites of broadly conserved miRNAs (across vertebrates, as predicted by *TargetScan* [108]), using a Fisher's exact test and FDR p-value correction. For a full description of this method see Methods section 2.1.4. As can be seen in Table 3.1, only miR-22-3p, the well-known tumour suppressor miRNA [233], was called as significantly

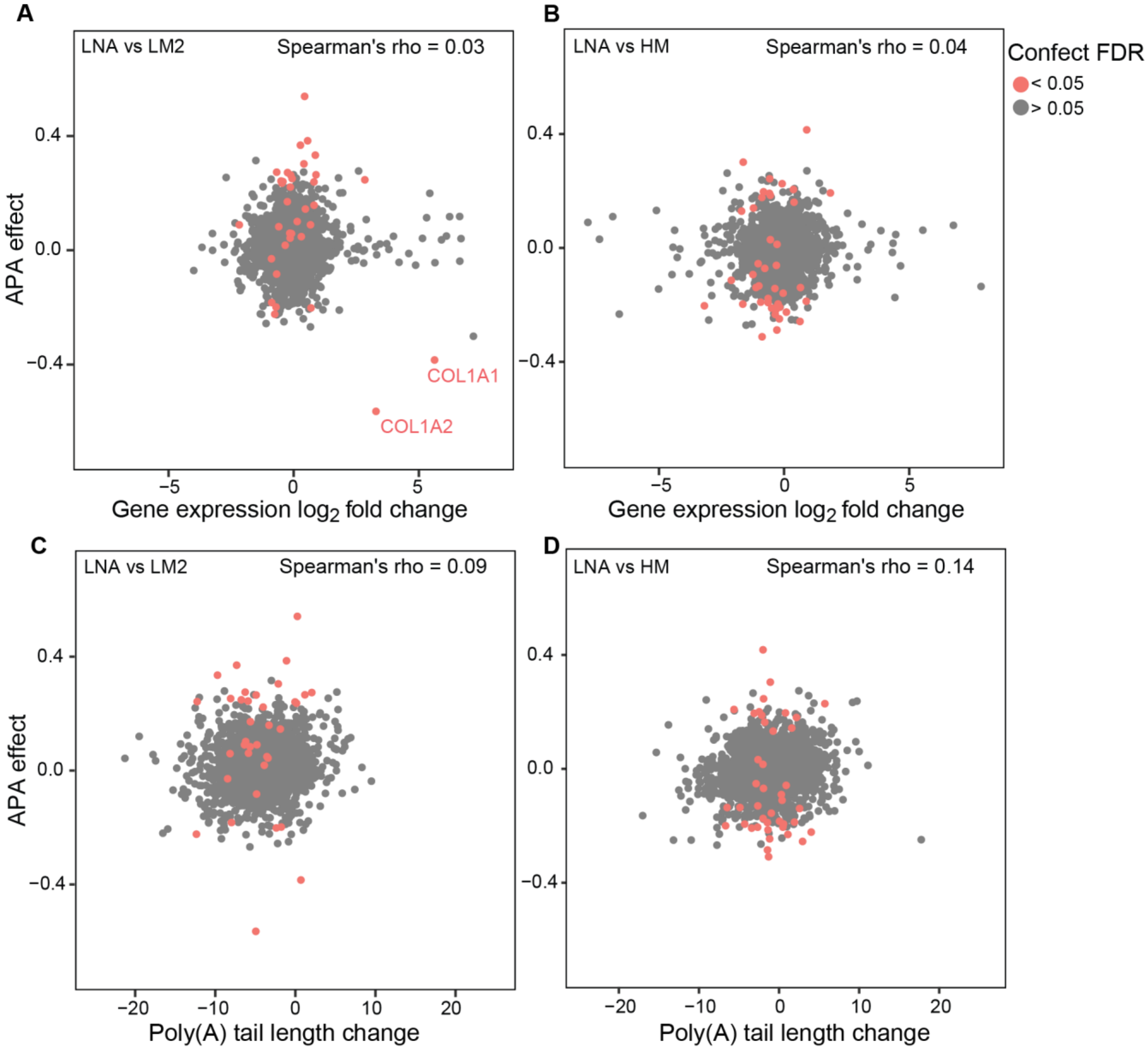


Figure 3.6. APA is not correlated with poly (A) tail length or differential gene expression in the MXM. A. Gene expression $\log_2$ fold change of the LM2 tumour line vs APA effect (LNA baseline). Positive APA effect indicates 3’ UTR lengthening and vice versa. Genes with a statistically significant confect value (FDR < 0.05) are highlighted in red. **B.** Gene expression $\log_2$ fold change of the HM tumour line vs APA effect (LNA baseline). **C.** Poly(A) tail length change of the LM2 tumour line vs APA effect (LNA baseline). **D.** Poly(A) tail length change of the HM tumour line vs APA effect (LNA baseline).

enriched in these genes at an FDR < 0.05. This suggested that APA changes from NI to HM tumours were associated with the escape of miR-22-3p repression.

To determine if APA changes were also associated with changes in miRNA expression, an alPAT was performed, targeting other known breast cancer associated miRNAs on RNA from the cell lines of the MXM [226, 1, 234]. The alPAT method was chosen as it can measure the expression of multiple non-adenylated RNAs, including miRNAs, through the ligation of a 3' adapter and provides sequence information to ensure that the correct miRNA has been amplified. Due to the low amounts of total RNA left over from the *in vivo* PAT-Seq experiments, RNA from the MXM cell lines grown *in vitro* was used in this experiment. The NI and HM lines were used in this comparison to mirror the previous miRNA overrepresentation analysis in section 3.2.4, and because this comparison had the greatest difference in metastatic potential. The full list of primers used in the alPAT can be seen in Table A.3. Of the miRNAs tested, 4 were found to be significantly altered in the HM line (Figure 3.7 B). The *Tail-tools* software, which is also used to interpret the results of alPAT experiments, was unable to assign some miRNAs to only one genomic site of origin. When this occurred, miRNAs were split evenly between possible sites. These sites were subsequently recombined for the analysis of miRNA expression. Interestingly, there was little change in the expression of miR-22-3p, suggesting that APA alone was used to escape regulation by this miRNA. These results provide evidence that miRNAs may be dysregulated in parallel to APA with increasing metastatic potential, but that the two are not always explicitly linked.

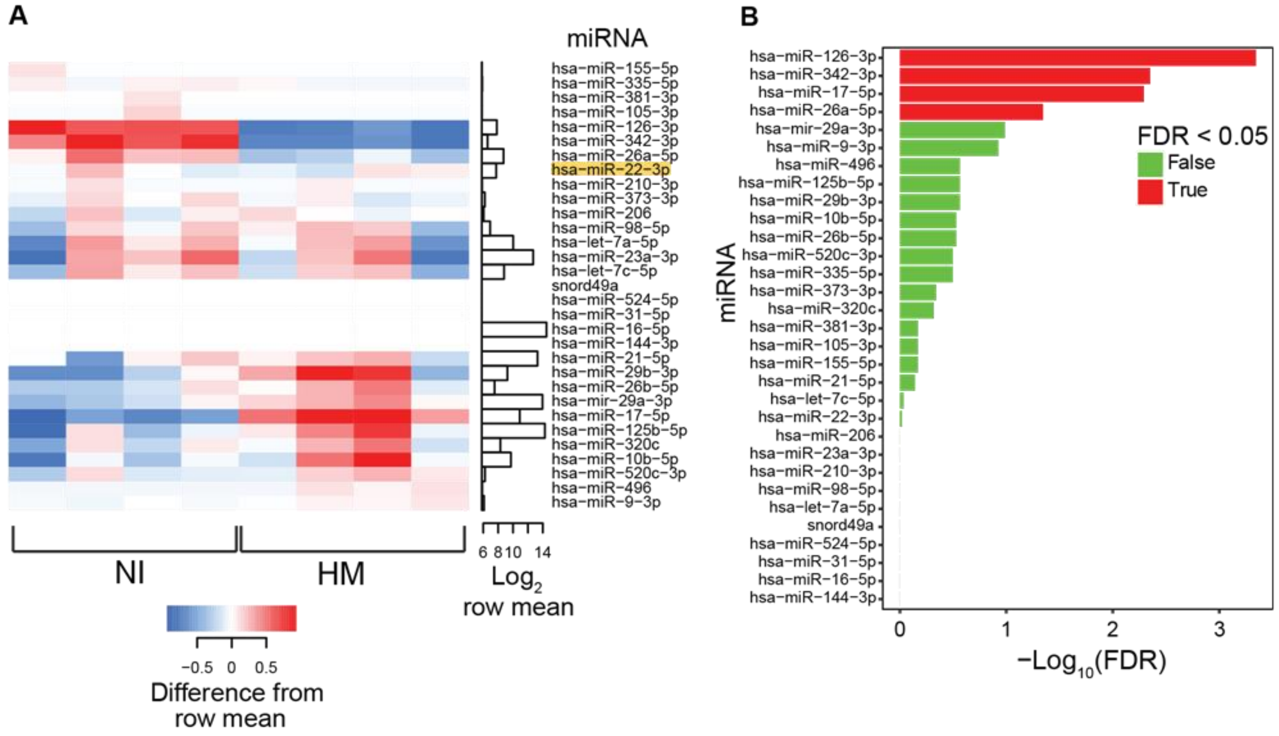


Figure 3.7. The miRNAs altered between the NI and HM lines *in vitro*. A. Heatmap of miRNA expression in the NI cell line vs the HM cell line as assayed by alPAT. Highlighted in orange is miR-22-3p, which had significantly reduced 3’ UTR binding sites in the HM tumour line. **B.** MiRNAs with a statistically significant log2 fold change (FC) from the same comparison, ordered by FDR. FDRs from 2-sided t-tests, adjusted by the Benjamini-Hochberg method for multiple testing correction [210].

Table 3.1. The miRNA binding sites lost during metastatic transformation. The top 20 (by p-value) miRNAs present in the 3' UTRs (*TargetScan* predictions) of the 1000 genes most proximally shifted in HM primary tumours (vs NI tumours).

miRNA	Count in all UTRs	p	FDR
miR-22-3p	19	4.31E-04	4.44E-02
miR-129-3p	15	2.32E-03	1.20E-01
miR-125-5p	19	1.15E-02	2.96E-01
miR-147b	4	1.15E-02	2.96E-01
miR-216-5p	27	2.25E-02	4.16E-01
miR-31-5p	15	2.42E-02	4.16E-01
miR-24-3p	26	4.50E-02	5.11E-01
miR-212-5p	17	4.68E-02	5.11E-01
miR-218-5p	14	5.21E-02	5.11E-01
miR-34-5p/449-5p	15	5.71E-02	5.11E-01
miR-138-5p	14	5.97E-02	5.11E-01
miR-7-5p	17	6.26E-02	5.11E-01
miR-191-5p	3	6.45E-02	5.11E-01
miR-103-3p/107	15	7.96E-02	5.36E-01
miR-455-3p.2	17	8.32E-02	5.36E-01
miR-208-3p	8	8.32E-02	5.36E-01
miR-133a-3p.2/133b	9	1.04E-01	5.93E-01
miR-135-5p	12	1.08E-01	5.93E-01
miR-490-3p	12	1.09E-01	5.93E-01
miR-184	3	1.25E-01	6.16E-01

3.2.5 RNA processing was increased and immune signalling was decreased with metastatic potential

Gene expression has been commonly used as a prognostic marker of breast cancer outcome [56, 44]. Metastatic potential was, therefore, analysed in relation to gene expression to determine the key metastasis-associated gene expression changes in the MXM. PAT-seq gene expression counts were exported from *Tail-tools* for *Limma* DGE analysis [203]. Of the 17,803 genes measured by PAT-seq run 2, 2,828 were differentially expressed (*Limma* FDR < 0.05) with increased metastatic potential when comparing the NI line with the mean gene expression of the other 3 lines combined. Gene set enrichment analysis of both up and down-regulated gene sets was performed in R using the *Camera* package [203] and revealed 385 dysregulated Gene Ontology (GO) gene sets (the top 20 of which, with the lowest FDR values, can be seen in Table 3.2). *Camera* also identified 131 enriched gene sets from the

Table 3.2. Enriched metastasis-associated GO terms. The top 20 (by FDR) GO terms called as enriched by *Camera* when comparing NI tumours to the mean gene expression of the other three lines (PAT-seq run 2 only).

N (Genes)	Direction	FDR	Gene set
317	Up	2.24E-06	GO RIBONUCLEOPROTEIN COMPLEX BIOGENESIS
146	Up	2.89E-06	GO RIBONUCLEOPROTEIN COMPLEX SUBUNIT ORGANIZATION
40	Down	2.69E-05	GO RESPONSE TO TYPE I INTERFERON
216	Up	2.69E-05	GO RIBOSOME BIOGENESIS
176	Up	3.18E-05	GO RRNA METABOLIC PROCESS
19	Down	4.64E-05	GO LUMENAL SIDE OF MEMBRANE
13	Down	4.64E-05	GO MHC PROTEIN COMPLEX
204	Up	5.61E-05	GO RNA SPLICING VIA TRANSESTERIFICATION REACTIONS
503	Up	6.35E-05	GO RIBONUCLEOPROTEIN COMPLEX
424	Up	9.42E-05	GO MRNA METABOLIC PROCESS
88	Up	9.80E-05	GO MULTI ORGANISM METABOLIC PROCESS
10	Down	1.07E-04	GO HEMIDESMOSOME ASSEMBLY
69	Up	1.37E-04	GO CATALYTIC STEP 2 SPLICEOSOME
591	Up	1.39E-04	GO RNA PROCESSING
43	Down	1.39E-04	GO INTERFERON GAMMA-MEDIATED SIGNALING PATHWAY
147	Up	1.39E-04	GO RIBOSOME
91	Up	1.39E-04	GO TRANSLATIONAL INITIATION
63	Up	1.49E-04	GO ESTABLISHMENT OF PROTEIN LOCALIZATION TO ENDOPLASMIC RETICULUM
36	Up	1.66E-04	GO BASE EXCISION REPAIR
378	Up	1.69E-04	GO NCRNA METABOLIC PROCESS

Table 3.3. Enriched metastasis-associated curated pathways. The top 20 (by FDR) ’curated pathways’ gene sets called as enriched by *Camera* when comparing NI tumours to the mean gene expression of the other three lines (PAT-seq run 2 only).

N Genes	Direction	FDR	Gene set
83	Up	3.10E-07	REACTOME MRNA SPLICING
106	Up	1.26E-06	REACTOME PROCESSING OF CAPPED INTRON CONTAINING PRE MRNA
122	Up	1.52E-05	REACTOME MRNA PROCESSING
38	Down	2.00E-05	REACTOME INTERFERON GAMMA SIGNALING
49	Up	4.97E-05	KEGG RIBOSOME
36	Down	8.49E-05	REACTOME INTERFERON ALPHA BETA SIGNALING
34	Up	1.87E-04	REACTOME MRNA SPLICING MINOR PATHWAY
86	Up	2.22E-04	REACTOME INFLUENZA VIRAL RNA TRANSCRIPTION AND REPLICATION
17	Up	2.22E-04	REACTOME G1 S SPECIFIC TRANSCRIPTION
28	Up	2.22E-04	PID BARD1 PATHWAY
45	Down	2.22E-04	KEGG ECM RECEPTOR INTERACTION
90	Up	2.64E-04	REACTOME 3’ UTR-MEDIATED TRANSLATIONAL REGULATION
13	Down	2.64E-04	KEGG ALLOGRAFT REJECTION
115	Down	2.75E-04	NABA ECM REGULATORS
86	Up	3.14E-04	REACTOME DNA REPAIR
13	Down	3.16E-04	KEGG AUTOIMMUNE THYROID DISEASE
7	Up	4.82E-04	BIOCARTA SET PATHWAY
115	Up	5.28E-04	REACTOME INFLUENZA LIFE CYCLE
314	Down	5.46E-04	NABA MATRISOME-ASSOCIATED
31	Up	5.46E-04	REACTOME E2F-MEDIATED REGULATION OF DNA REPLICATION

‘curated gene sets’ collection from the Molecular Signatures Database (MSigDB) [205] (Table 3.3). The top pathways and gene ontologies consistently showed the enrichment of upregulated genes associated with the ribosome and translation, indicating increased activity in these pathways with metastatic potential. There was also enrichment of downregulated genes in two pathways associated with a decrease in interferon signalling, suggesting potential immune evasion by the metastatic tumour lines. The NI tumour line was then compared individually with the three metastatic cell lines to determine the consistency of enriched gene sets. Barcode plots for the top 6 gene sets in the ‘curated gene sets’ collection for this comparison are given in Figure 3.8. The top pathways were again consistently associated with increased ribosome activity and increased mRNA processing. These results suggested that mRNA metabolism and processing pathways along with the immune response were consistently dysregulated with metastatic potential in the MXM.

3.2.6 The expression of mRNA 3’ processing factors was increased with metastatic potential

Changes in the expression of 3’ processing factors have previously been suggested to cause shifts in APA in tumours [66, 113]. Interestingly, mRNA splicing genes had the highest enrichment in the ‘curated gene sets’ collection overall (Table 3.3, Figure 3.9 A/B). Many of these splicing genes double as known APA factors and form key components of the APA machinery. Much of this splicing-associated expression change appears in the HM line (Figure 3.9 C), suggesting that higher levels of splicing factor expression are associated with the most metastatic primary tumours in this model. Also observed in the NI vs LM2 comparison was an increase in the expression of factors that mediate translation by binding to the 3’ UTR (*Camera* FDR = 3.49 x 10^{-3}). This result, paired with the primarily distal APA observed in the LM2 line, was not expected based on the prevailing view that the increased APA factor expression generally leads to proximal APA.

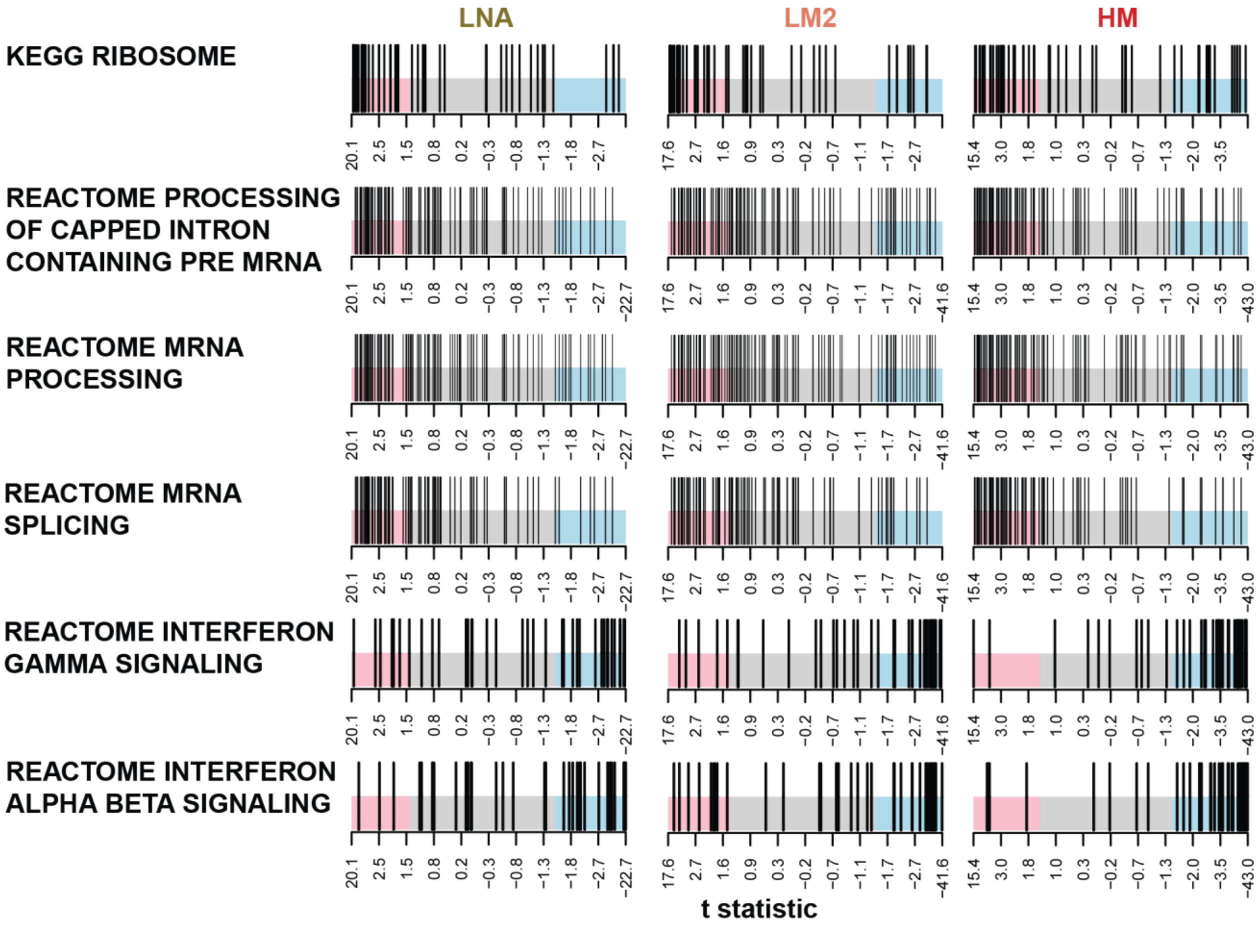
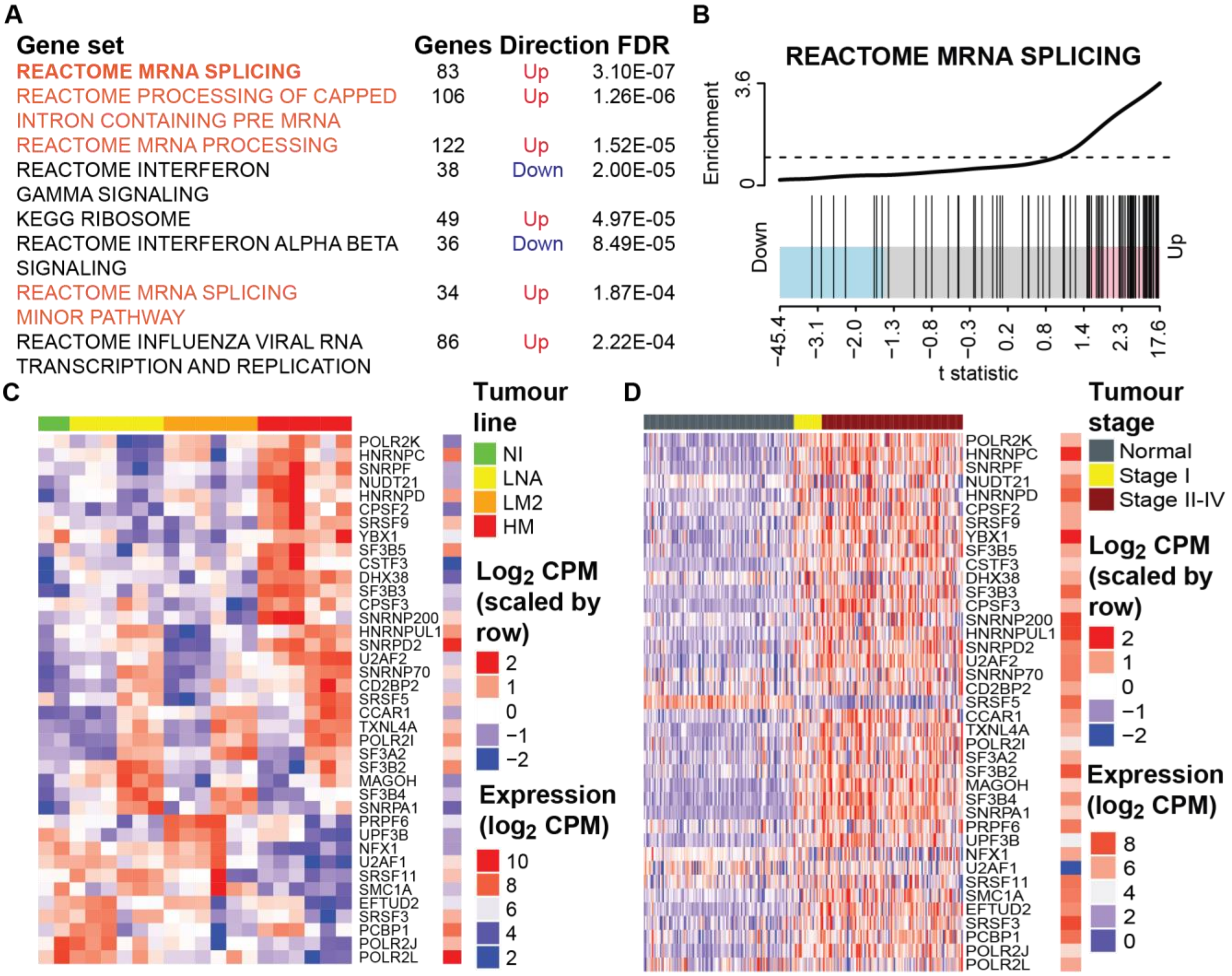


Figure 3.8. Barcode plots of the top 6 canonical pathways enriched in run 2 of the MXM. Top 6 gene sets regulated in the 3 metastatic cell lines of the MXM when compared to the NI tumours (as called by *Camera* [132]) and arranged by lowest FDR relative to the HM line. Gene sets are from the 'Canonical Pathways' collection from the MSigDB [205]. Vertical lines indicate the *Limma* t statistic value for a given gene in the set. If many genes are downregulated there will more lines at the negative end of the scale and at the positive end in the case of upregulation.

To check the MXM gene set enrichment results against real human tumours, gene expression data from primary TNBCs was downloaded from the TCGA through the GDC Data Portal (https://portal.gdc.cancer.gov/) and was also analysed using *Limma* and *Camera*. TCGA TNBC data are accompanied by patient information including tumour stage, which defines the invasive state of the tumour (higher stage = greater tumour progression). A similar pattern of deregulated gene expression was observed in primary TNBCs from the TCGA when compared with normal breast tissue (Figure 3.9 D). Expression of these genes was significantly lower in normal breast tissue than TNBC tumours (*Camera* FDR = 4.70 x 10^{-11}), however, this change may not be unique only to TNBCs (discussed further in Chapter 4). Splicing genes were also significantly upregulated in more advanced primary tumours (stage II-IV), when compared with the least advanced primary tumours (stage I; *Camera* FDR = 7.8 x 10^{-12}). Taken together, these results suggest that expression of splicing and APA genes is increased in tumour formation and is further increased with metastatic potential.

3.2.7 The MXM provides a controlled model for the study of TNBC metastasis

To determine how well the findings in the MXM would generalise to clinical TNBC samples, gene expression data from the current study was again compared with the TCGA. Following the filtering of lowly expressed genes, 23,102 genes were considered expressed in the TCGA TNBC dataset (compared with 17,803 MXM), with the expression of 13,053 genes overlapping in both datasets. To determine the gene sets that were associated with metastasis, the NI line was compared to the mean gene expression of the other three metastatic lines in the MXM (from PAT-seq run 2) and stage I TCGA primary TNBCs were compared with the mean gene expression of all later stages (Figure 3.10 A). Despite the different sample origins, there was general agreement between gene expression in both the



Gene set	Genes	Direction	FDR
REACTOME MRNA SPLICING	83	Up	3.10E-07
REACTOME PROCESSING OF CAPPED INTRON CONTAINING PRE MRNA	106	Up	1.26E-06
REACTOME MRNA PROCESSING	122	Up	1.52E-05
REACTOME INTERFERON GAMMA SIGNALING	38	Down	2.00E-05
KEGG RIBOSOME	49	Up	4.97E-05
REACTOME INTERFERON ALPHA BETA SIGNALING	36	Down	8.49E-05
REACTOME MRNA SPLICING MINOR PATHWAY	34	Up	1.87E-04
REACTOME INFLUENZA VIRAL RNA TRANSCRIPTION AND REPLICATION	86	Up	2.22E-04

Figure 3.9. AS and APA factor expression is associated with metastatic potential in the MXM and TNBCs. A. *Limma-Camera* gene set enrichment analysis of the NI cell line vs the mean expression of the LNA, LM2 and HM lines. Gene set testing was performed against the full complement of gene sets in the 'curated gene sets' collection in the MSigDB (version 5.2) [205]. The 'mRNA splicing' pathway (highlighted in red) from the 'Reactome Curated Pathway Database' [161] was the most enriched gene set. Many mRNA splicing elements also play key roles in APA. The top 3 enriched gene sets were associated with mRNA processing and all were upregulated. **B.** Barcode plot of the 'Reactome mRNA splicing' gene set enrichment by *Camera*. **C.** Heatmap of the expression of all genes in the MXM ($\log_2$ CPM) from the 'Reactome mRNA splicing' gene set that were expressed in the MXM. **D.** The gene expression ($\log_2$ CPM) of the same 'Reactome mRNA splicing'-associated genes in normal breast tissue and TNBCs from the TCGA [166]. The tumours have been split into two groups, the less advanced stage I tumours, and the more advanced stage II-IV tumours.

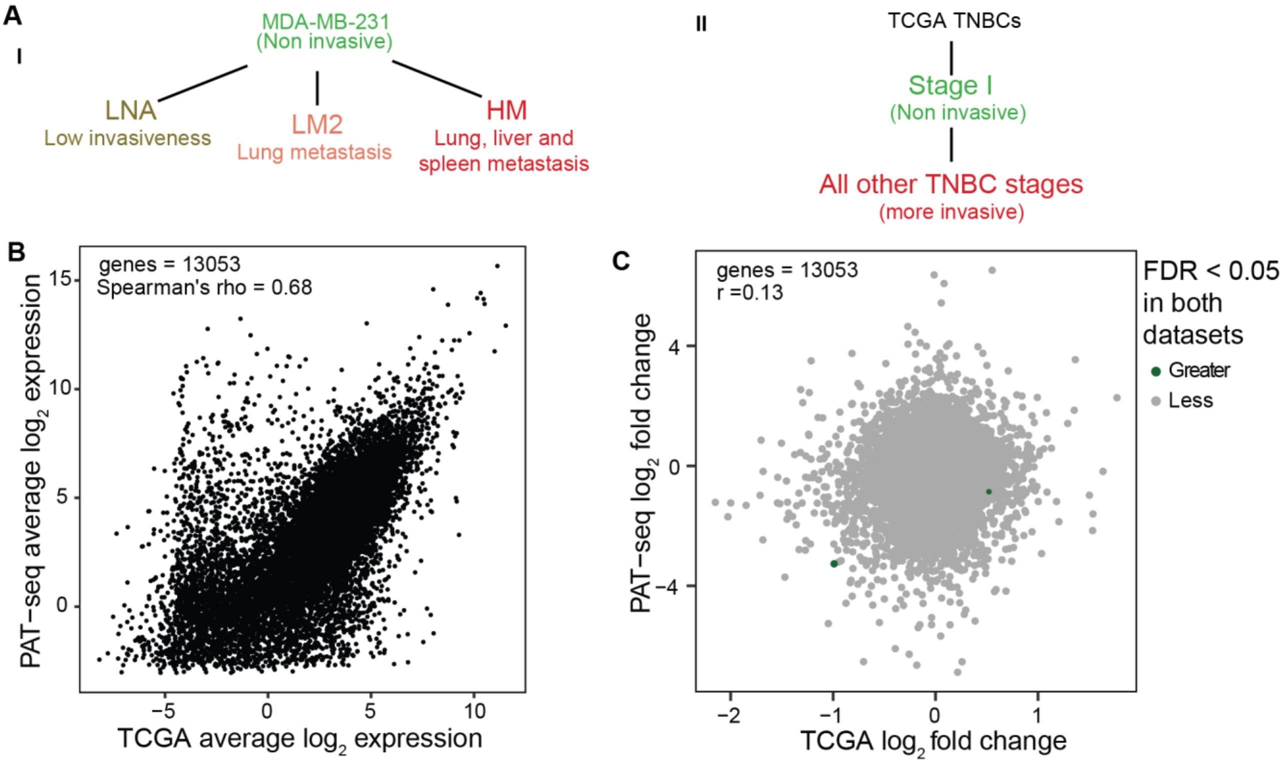


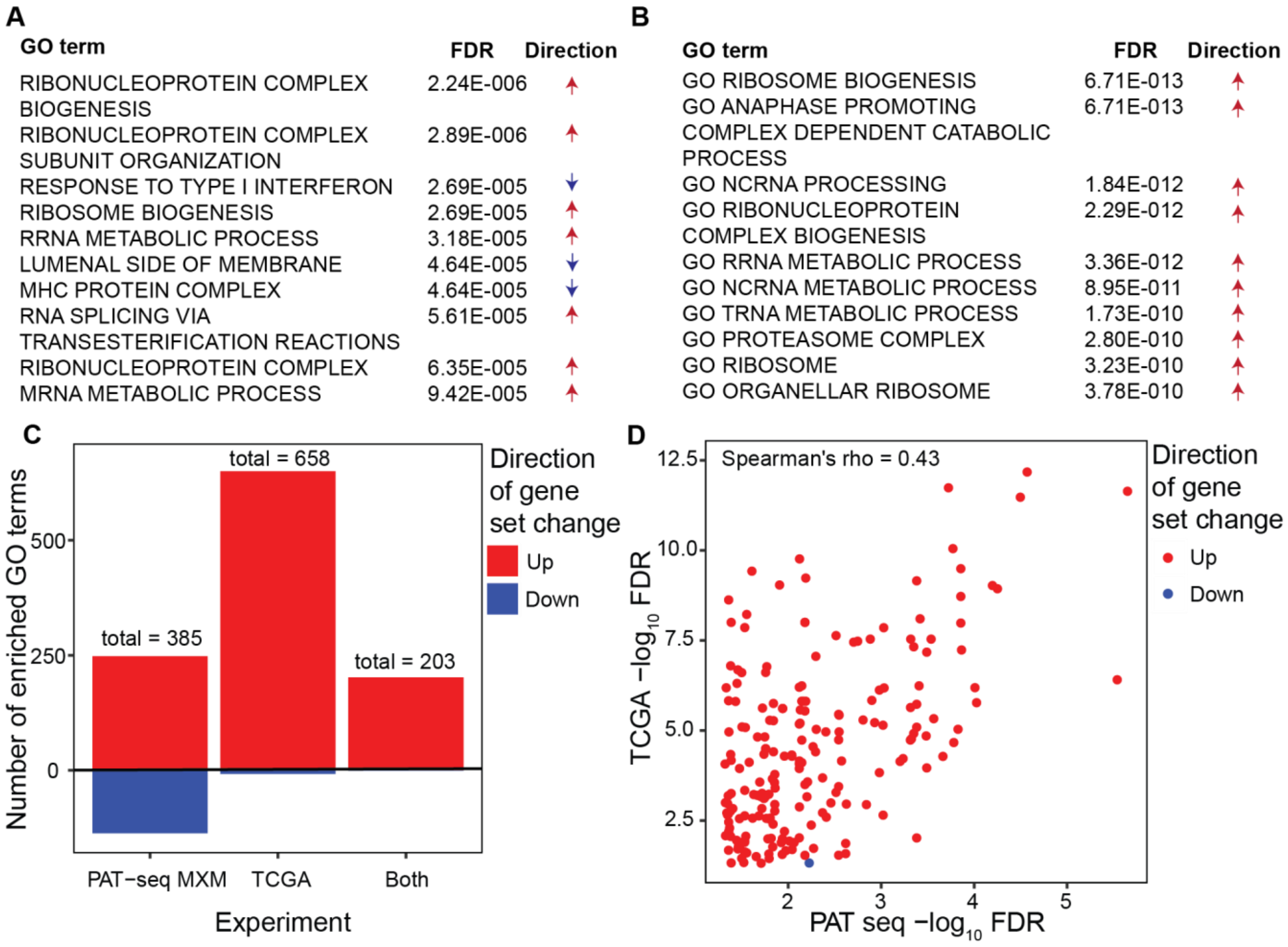
Figure 3.10. PAT-seq and TCGA gene expression are consistent overall but not with metastatic potential. A. I. The cell lines used in this study LNA, LM2 and HM were derived from and compared against the MDA-MB-231 TNBC cell line (NI). **A. II.** To compare gene expression with the MXM, TNBCs from the TCGA were compared on the basis of primary tumour stage I vs all other stages. **B.** Comparison of the $\log_2$ mean gene expression of genes expressed in both the PAT-seq of the MXM and all TNBC tumours present in the TGCA. **C.** Comparison of the $\log_2$ fold change in the mean expression of genes expressed in both the PAT-seq of the MXM (NI tumours vs the average all other tumour types) and all TNBC tumours present in the TGCA (stage I tumours vs all other tumour stages).

datasets (Figure 3.10 B, Spearman's correlation = 0.68). Interestingly, this trend was not reflected in the gene expression changes between both datasets, with the majority of changes uncorrelated (Pearson's correlation = 0.13, Figure 3.10 C). This suggested that the MXM was not altering the expression of the same genes as primary human tumours with increased metastatic potential.

While individual gene expression changes tended not to be correlated with metastasis between the two datasets, there was a clearer trend in the GO terms that were enriched in each dataset. As can be seen in Figure 3.11 A/B, gene sets associated with RNA processing and increased metabolism tended to be upregulated in both datasets, however, the downregulation of immune-associated pathways was not present in the TCGA dataset, suggesting that these changes may be specific to the MXM. In fact, there were very few downregulated pathways in the TCGA comparison overall (Figure 3.11 C), suggesting that the activity of many gene expression pathways is primarily increased with the metastatic potential of a primary breast tumour. Interestingly, despite low overall correlation of the expression of specific genes, there was modest correlation of the gene sets that were shared between both datasets (Figure 3.11 D, Spearman's correlation = 0.43). This suggested that tumour progression was occurring through different genes, but still broadly employing many of the same cellular mechanisms for proliferation and metastasis.

3.2.8 Minor transcriptome wide poly(A) tail length changes were observed with increasing metastatic potential

As mentioned in Chapter 1, cytoplasmic polyadenylation binding elements have previously been implicated in the progression of cancer [120]. To determine if poly(A) tail length was associated with metastasis in the MXM, poly(A) tail length changes across the cell lines of the MXM were compared. Overall, PAT-seq showed variable poly(A) tail lengths both within



A

GO term	FDR	Direction
RIBONUCLEOPROTEIN COMPLEX BIOGENESIS	2.24E-006	↑
RIBONUCLEOPROTEIN COMPLEX SUBUNIT ORGANIZATION	2.89E-006	↑
RESPONSE TO TYPE I INTERFERON	2.69E-005	↓
RIBOSOME BIOGENESIS	2.69E-005	↑
RRNA METABOLIC PROCESS	3.18E-005	↑
LUMENAL SIDE OF MEMBRANE	4.64E-005	↓
MHC PROTEIN COMPLEX	4.64E-005	↓
RNA SPLICING VIA TRANSESTERIFICATION REACTIONS	5.61E-005	↑
RIBONUCLEOPROTEIN COMPLEX	6.35E-005	↑
MRNA METABOLIC PROCESS	9.42E-005	↑

B

GO term	FDR	Direction
GO RIBOSOME BIOGENESIS	6.71E-013	↑
GO ANAPHASE PROMOTING COMPLEX DEPENDENT CATABOLIC PROCESS	6.71E-013	↑
GO NCRNA PROCESSING	1.84E-012	↑
GO RIBONUCLEOPROTEIN COMPLEX BIOGENESIS	2.29E-012	↑
GO RRNA METABOLIC PROCESS	3.36E-012	↑
GO NCRNA METABOLIC PROCESS	8.95E-011	↑
GO TRNA METABOLIC PROCESS	1.73E-010	↑
GO PROTEASOME COMPLEX	2.80E-010	↑
GO RIBOSOME	3.23E-010	↑
GO ORGANELLAR RIBOSOME	3.78E-010	↑

Figure 3.11. PAT-seq and TCGA gene set enrichment comparison shows broadly consistent changes with metastatic potential. A. Top 10 GO terms from the PAT-seq gene expression (NI vs more invasive tumour types) by FDR. **B.** Top 10 GO terms from the TCGA gene expression comparison (stage I vs all other stages) by FDR. **C.** Comparison of the number of enriched GO terms in the PAT-seq MXM (NI vs all other tumour types) and the TCGA (stage I vs later TNBC stages) and the direction of this expression change. **D.** Comparison of the FDRs and direction of change of gene sets enriched in both datasets.

and between tumour types. There was also variability between sequencing runs, with samples in run 2 generally having longer poly(A) tail lengths than samples in run 1 (Figure 3.12 A/B). In run 2 specifically, there was greater poly(A) tail length variation within the NI samples relative to the other lines, possibly due to a lower number of total reads in the experiment (as PAT-seq has limited sensitivity in balancing relative DNA input from each sample for sequencing), and as such, the NI samples were excluded from further analysis. When compared with the LNA line, global poly(A) tail length was significantly shorter in the LM2 and HM lines respectively (4 and 2 bases on average, paired t-tests, both sequencing runs, $p << 0.01$). This difference increased to 8 and 5 bases when comparing genes that had a statistically significant change in poly(A) tail length (*Tail-tools* ANOVA, FDR < 0.05), but may be due to biases in PAT-Seq library preparing discussed later in this chapter. These results should, therefore, not be considered as strong evidence of global poly(A) tail length changes in breast cancer.

Gene specific poly(A) tail length showed no association with gene expression

As poly(A) tail length has been suggested to play a role in mRNA transcript stability in some cellular contexts [68], the association of gene expression with poly(A) tail length was compared. Gene expression was generally uncorrelated with poly(A) tail length (Figure 3.12 C) with a weak trend toward genes with higher expression having shorter tails (Pearson's r = -0.14). Changes in gene expression also largely had no correlation with poly(A) tail length change in the LM2 and HM lines (Figure 3.12 D). Somewhat surprisingly, PAT-seq poly(A) tail length was also completely uncorrelated with TAIL-Seq sequencing of HeLa cells [117] (Spearman's rho = 0.03, Figure 3.12 E). TAIL-Seq is a comparable 3'-focused sequencing method that also relies on RNase T1 fragmentation and short-read sequencing. This lack of correlation may have been due to differences in growth conditions, prior treatment of cell lines

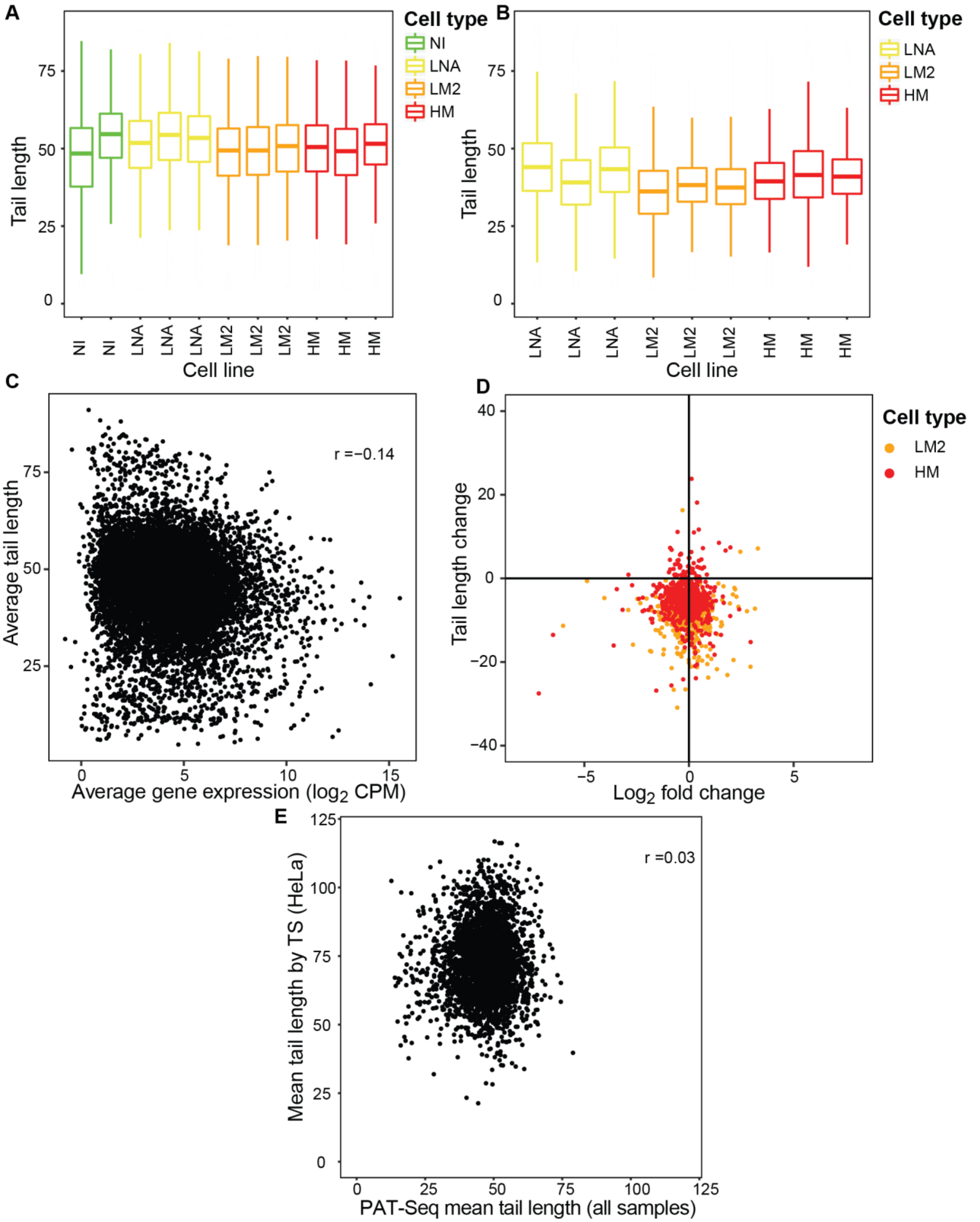


Figure 3.12. PAT-seq gene expression and poly(A) tail length comparison. A. Mean poly(A) tail length from the second PAT-seq of the MXM. **B.** Mean poly(A) tail length from the first PAT-seq of the MXM. **C.** Mean poly(A) tail length vs mean gene expression for genes expressed in the LNA, LM2 and HM primary tumours of the MXM. **D.** Comparison of poly(A) tail length change vs $\log_2$ fold change comparison for the LM2 and HM lines (vs the LNA line, both runs). **E.** Comparison of mean poly(A) tail length from all samples in the MXM PAT-seq and HeLa cell poly(A) tail length measured by the TAIL-seq method [117].

and differences in protocol. Differences in transcript processing between the two cell lines may have also resulted in different genes having different poly(A) tail lengths. It is also possible that poly(A) tail length does not vary consistently based on mRNA sequence content.

3.2.9 Assessing the reliability of PAT-seq poly(A) tail length measurement

Due to the variable nature of the transcriptome-wide poly(A) tail length changes that were previously observed, the ability of PAT-seq to effectively measure poly(A) tail length in this model was evaluated and validated. Mean-variance analysis was used to compare poly(A) tail length variability with gene expression. It is expected that as the number of PAT-Seq reads increase, the reliability of poly(A) tail length measurement will increase, and the variance of this measurement will decrease. The length of the poly(A) itself was also compared with poly(A) tail length variability to determine if there was any variability in the measurement of tails of a particular length. In both sequencing runs the variability of poly(A) tail length measurement was reduced after ~25 bp (Figure 3.13 A/B). This is perhaps due to the decreased incidence of genomic A-rich regions (included by mispriming of the oligo-dT primer) beyond this length. Poly(A) tail length measurement was less variable with increasing gene expression (Figure 3.13 C/D), adding confidence that PAT-seq is sampling non-random poly(A) tail lengths. When changes in global poly(A) tail length were accounted for by standardising poly(A) tail length, no changes in poly(A) tail length for any single gene could be observed by *Tail-tools* (Figure 3.14 B/D). Standardisation was performed by multiplying all poly(A) tail length values by the mean of the mean poly(A) tail lengths for each sample and then dividing the poly(A) tail lengths in each sample by the sample mean. This suggested that there were minor, genome-wide poly(A) tail length changes in the tumour lines that this experiment lacked the resolution to reliably detect, or that poly(A) tail length did not change with metastatic potential in this model.

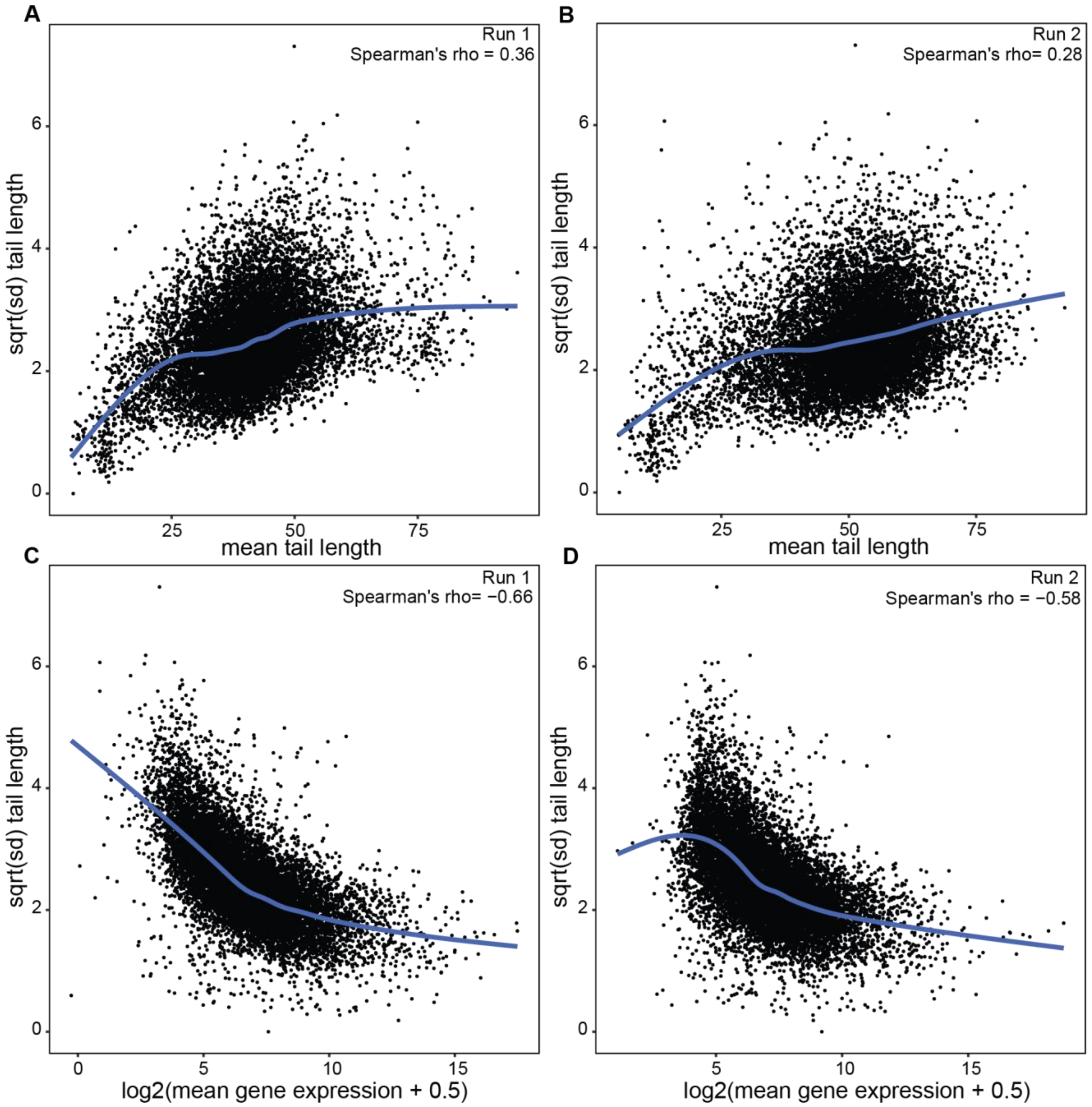


Figure 3.13. Variance in poly(A) tail length measurement. A/B. The mean-variance trend plot of poly(A) tail length for each run of the MXM PAT-seq experiments. The variability of poly(A) tail length measurement increases roughly linearly with poly(A) tail length until ~25 bp where it begins to stabilise. **C/D.** The mean-variance trend plot of gene expression (x-axis) and poly(A) tail length (y-axis) for each run of the MXM PAT-seq experiments. The plots show that higher gene expression tends to be associated with more reliable poly(A) tail length measurement.

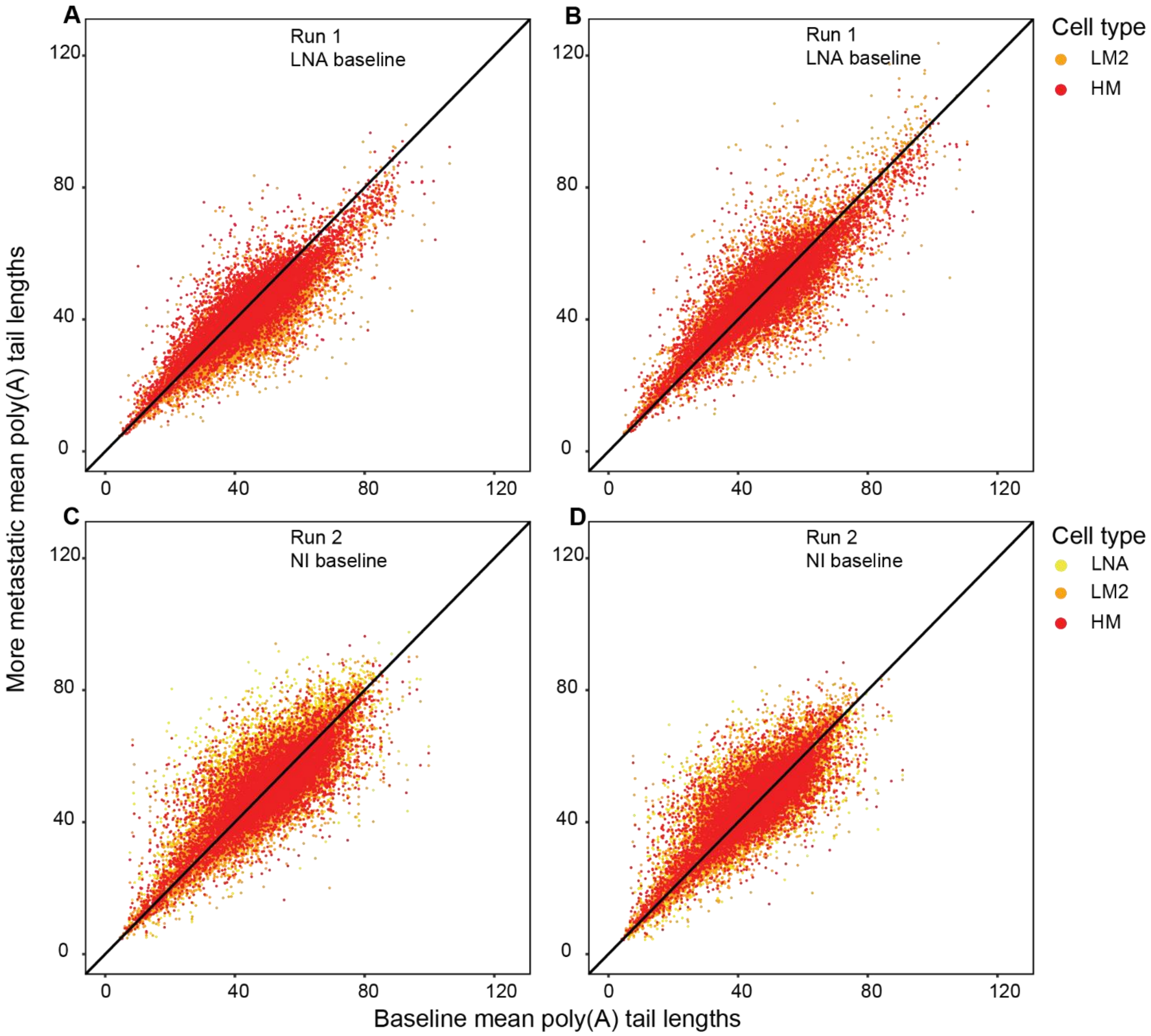


Figure 3.14. PAT-seq poly(A) tail length normalisation A. Mean poly(A) tail lengths of all genes expressed in the first PAT-seq of the MXM. Both the LM2 and HM tumour lines are plotted against the LNA baseline. **B.** The same comparison as in A scaled to have the same mean by multiplicative normalisation. **C.** Mean poly(A) tail lengths of all genes expressed in the first PAT-seq of the MXM. The LNA, LM2 and HM tumour lines are plotted against the LNA baseline. **D.** The same comparison as in C scaled to have the same mean by multiplicative normalisation.

3.2.10 Investigation of poly(A) tail length changes by alternative methods

Poly(A) tail length is notoriously difficult to accurately measure [68], possibly explaining the absence of significant changes observed in the MXM. The mPAT [215] multi-gene, NGS-based method and the simpler ePAT [211] agarose gel-based method were utilised to determine if the most promising candidate genes with a suggested poly(A) tail length change from the MXM PAT-seq could be validated by more targeted approaches. The mPAT method is a 3'-focused targeted re-sequencing method that gives similar results to PAT-seq, however instead of operating on a transcriptome-wide scale, it requires targeted primers to be designed to each CPA site tested. The method is capable of measuring polyadenylated transcripts up to 300bp in total length (as opposed to 150 in PAT-seq) and gives a more specific and in-depth readout due to the vast reduction in the number of measured genes. To validate the single-gene poly(A) tail length changes observed (prior to normalisation) in PAT-Seq run 1, an mPAT was performed using total RNA from MXM cell lines cultured *in vitro*, referred to as the *in vitro* mPAT. Three vials each of the NI, LNA, LM2 and HM cell lines were used as pseudo-replicates (see methods Section 2.1.3 for a full description of the cell culture protocol). An mPAT was also performed on the remaining xenograft tumour total RNA that was used for the PAT-seq experiments, referred to as the *in vivo* mPAT. Genes were selected based on poly(A) tail length changes suggested by the PAT-seq experiment, or as validation for observed gene expression changes. The primers used in these experiments can be seen in Tables A.1 and A.2 respectively and were designed using the *3Primer* primer design tool, discussed in section 5.2.7.

The correlation of poly(A) tail lengths from the *in vitro* mPAT and the *in vivo* PAT-seq was low (Pearson's r = 0.35, Figure 3.15 A), once again suggesting variable poly(A) tail length between conditions. Poly(A) tail length showed a slight decrease with increasing metastatic potential across the *in vitro* mPAT (Figure 3.15 B), however, this difference was not significant (p >0.05, t-test). Inspection of poly(A) tail length at the single gene level showed that poly(A) tail length tended to vary more by sample than by cell line in this experiment (Figure 3.16 A, left heatmap). This was not the case for gene expression which tended to vary with cell line (Figure 3.16 A, right heatmap). This was also the case in the PAT-seq experiments, in which gene expression changes could be much more clearly determined than poly(A) tail length changes. The poly(A) tail lengths measured by mPAT were also generally shorter than those measured by PAT-seq, possibly owing to PCR slippage [235] caused by the 5-10 additional PCR cycles in the mPAT method. The shortening could also have been caused by sequencing errors associated with the lower sequence heterogeneity in the mPAT library. Short, highly variable tails were also observed in the *in vivo* mPAT (Figure 3.16 B, left heatmap), in which gene expression was once again consistent (Figure 3.16 B, right heatmap).

The most promising poly(A) tail length change from PAT-seq experiments was detected in run 1 and was a decrease in poly(A) tail length in the HM line (vs LNA baseline) in the glutamate-ammonia ligase (*GLUL*) gene (Figure 3.17 A). This poly(A) tail length change could not be recapitulated in the second PAT-seq experiment (Figure 3.17 B), the *in vitro* mPAT (Figure 3.17 C) or the *in vivo* mPAT (Figure 3.17 D). An ePAT was also performed from the RNA that was used in the *in vitro* mPAT (Figure 3.17 F) to visualise the poly(A) tail as a smear on an agarose gel and, once again, the poly (A) tail length change could not be replicated. The darker bands in the HM samples of this ePAT are likely due to increased *GLUL* gene expression, as was shown in the *in vitro* mPAT (Figure 3.17 E). These results suggested that even when utilising alternative methods, poly(A) tail length changes to *GLUL*

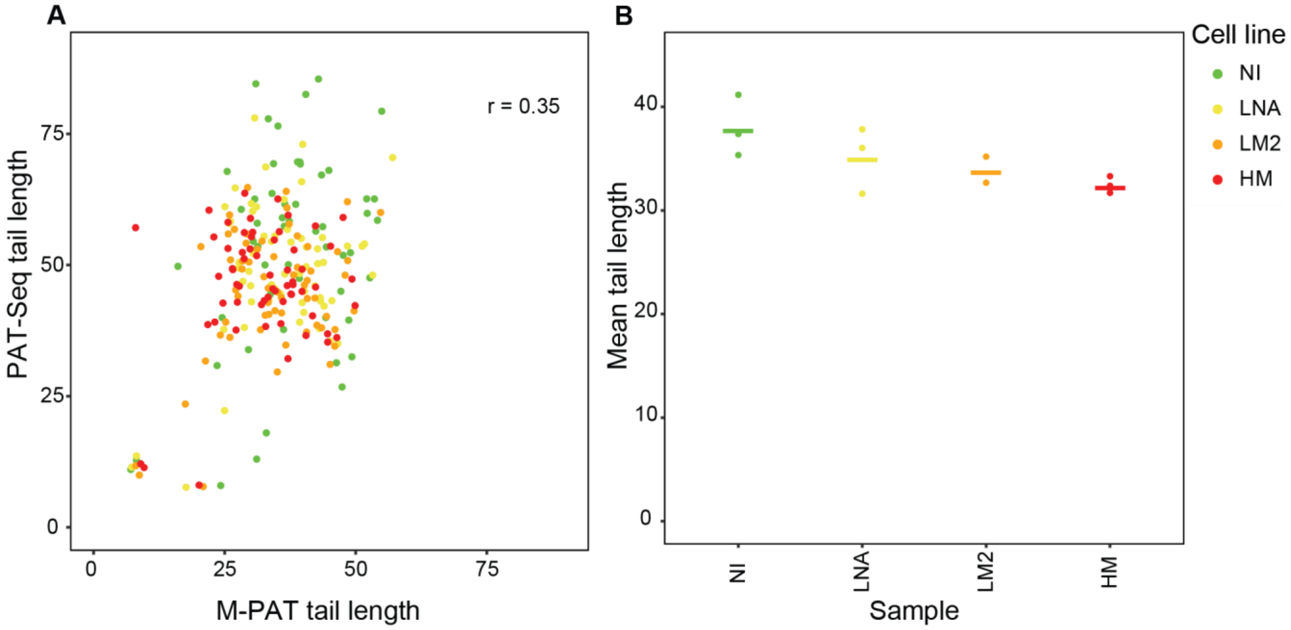


Figure 3.15. Tail lengths measured by mPAT. A. Correlation of mean tail lengths of the genes tested by mPAT with their counterparts in the MXM PAT-seqs. The same MXM in *in vivo* RNA was used for both experiments. **B.** The mean tail lengths from the *in vitro* mPAT.

identified in run 1 could not be recapitulated in these samples. Taken together, the high variability of poly(A) tail length measurement in the MXM meant that no single-gene poly(A) tail length changes could be confirmed.

3.2.11 Optimisation of the mPAT method for low abundance mammalian genes

The mPAT method is a novel protocol that has been developed in the RNA Systems Biology Laboratory and still requires some optimisation. Prior to the mPAT experiments described above, two unsuccessful attempts were made to obtain a product for sequencing as part of the mPAT from the RNA of the *in vitro* MXM cell lines. A substantial amount of time was subsequently spent testing the mPAT protocol in order to trouble-shoot the method.

To determine if the failure to obtain a PCR product was due to the non-annealing of gene-specific forward primers that I had designed, the annealing step of the first mPAT PCR was performed on single genes at both 60°C (Figure 3.18 A) and 55°C (Figure 3.18 B). In both cases the amplification of genes from single, gene-specific forward primers was successful, producing clear bands at the expected size. This result suggested that individual primers in the original pooled forward primer mix were not the cause of the issue and had indeed been designed appropriately. Even at the lower annealing temperature with as many as 35 PCR cycles (as compared to the standard 15-20), no products of the expected size could be observed in the full primer mix, suggesting annealing temperature was also not the reason for the failure. The design process of the PCR primers was reviewed, and it was determined that a failed PCR reaction could not be easily attributed to an obvious flaw in primer design. It was next hypothesised that one 'bad' primer may be somehow interfering with the reaction. Primers were split into 6 groups of ~10 primers and the first m-PAT PCR was repeated with each group (Figure 3.18 C). Products that were in the expected size range were obtained in

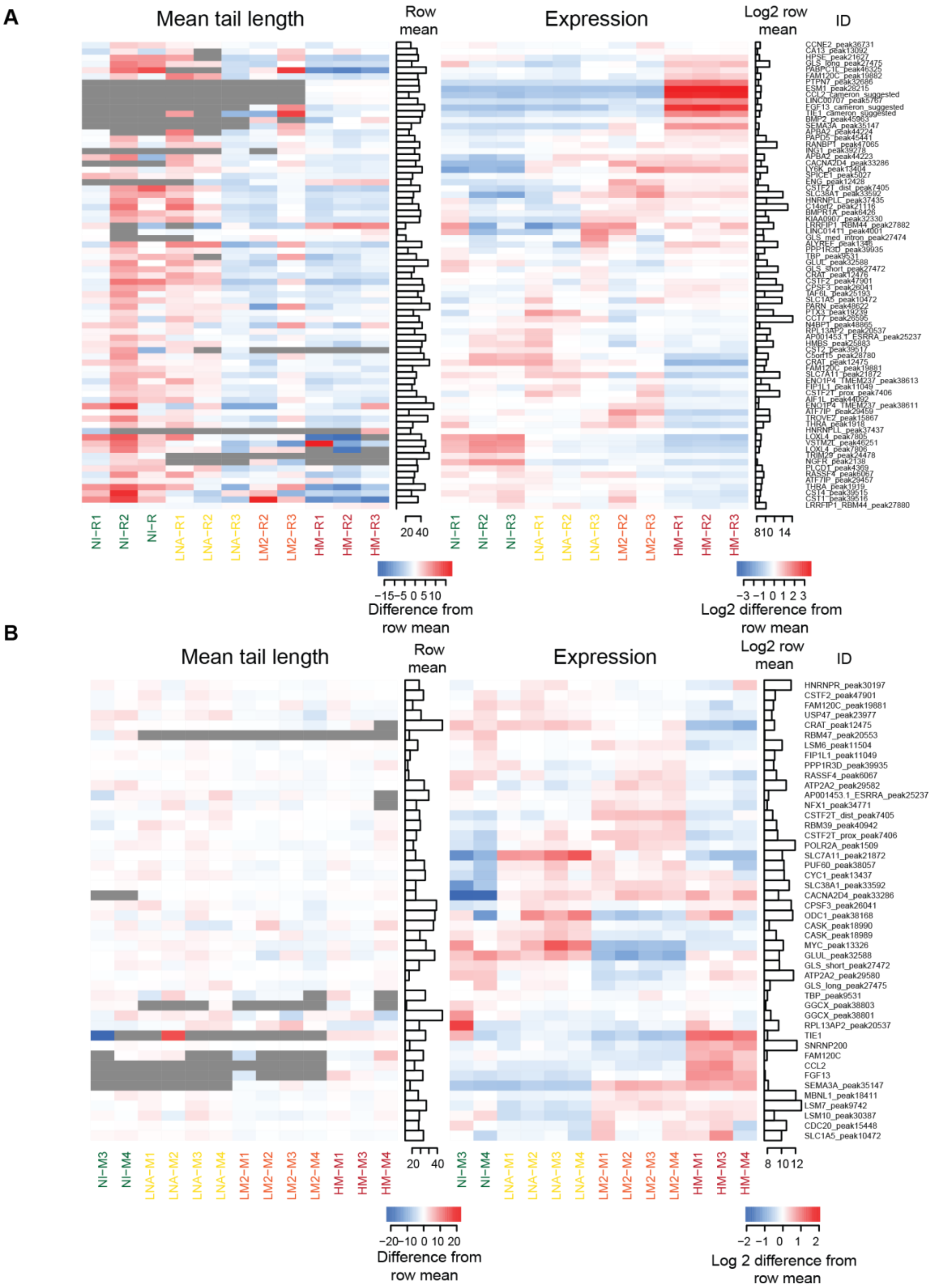


Figure 3.16. Measurement of MXM tail lengths and gene expression by mPAT. A. The heatmap on the left represents poly(A) tail length from an mPAT [125] of the cell lines used in the xenograft model grown *in vitro*. The heatmap on the right represents the expression of the same genes. **B.** The same as A, except the mPAT was performed on samples from run 2 of the mouse xenograft model. Clear gene expression patterns can be seen in both A and B, however, poly(A) tail length is far less consistent. A mean expression count of 20 reads was required for a primed region to be shown and a mean count of 20 polyadenylated reads was required for poly(A) tail length not to be greyed out.

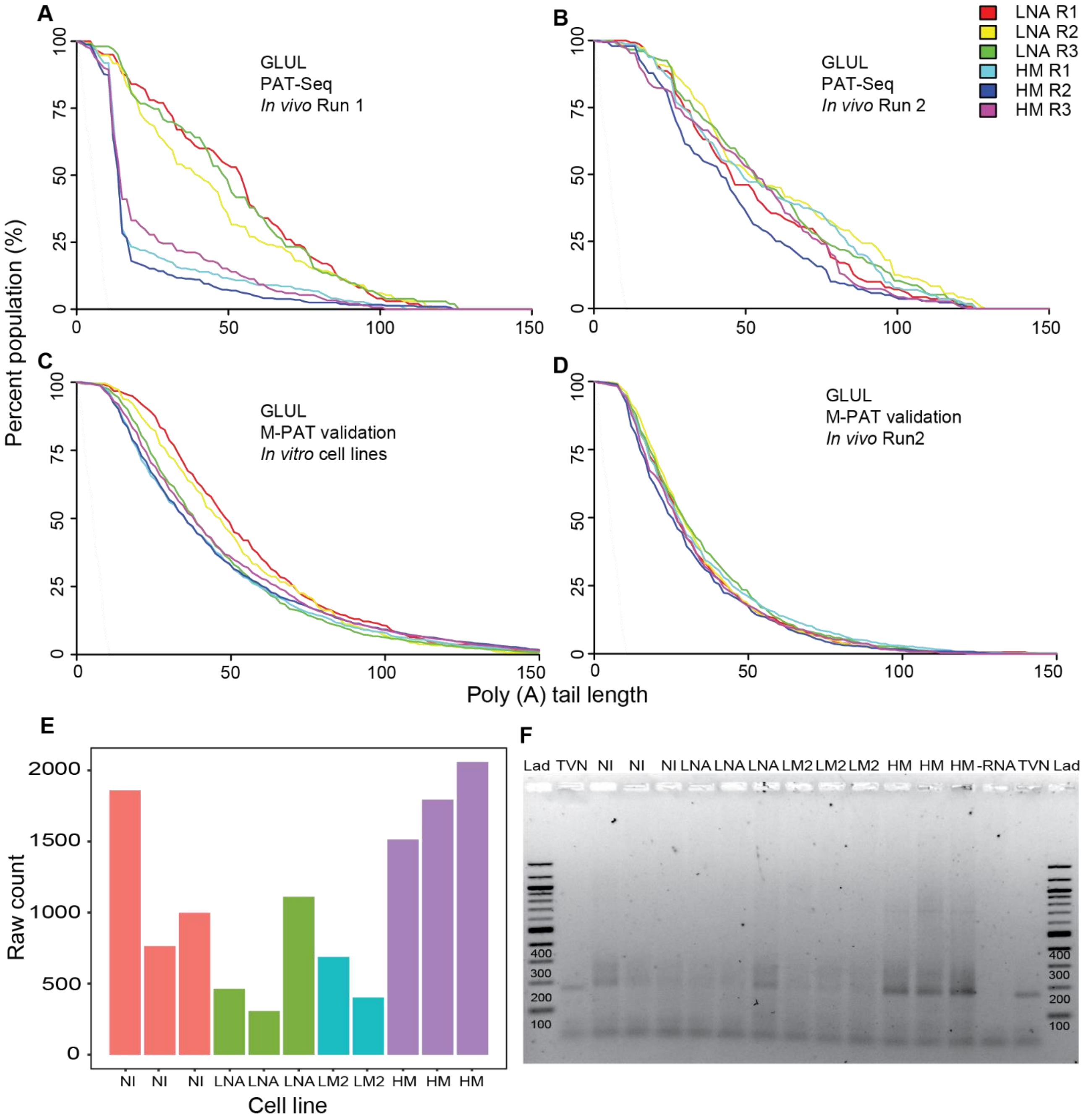


Figure 3.17. Attempting to replicate a *GLUL* poly(A) tail length shift. A. Cumulative distribution of poly(A) tail lengths of the LNA and HM xenograft tumour lines measured from reads mapping to the GLUL gene as measured by PAT-seq (run 1). A clear difference in poly(A) tail length can be observed between the LNA and HM lines **B.** The same plot as A for the second run of the xenograft tumour lines. In this case, no difference in poly(A) tail length is apparent. **C.** The same plot as A and B, this time generated from a targeted mPAT [125] experiment based on the LNA and HM cell lines grown *in vitro*. Once again, no clear difference in poly(A) tail length can be seen. **D.** An mPAT of the same samples as B. Again, no poly(A) tail length change can be observed. **E.** gene expression of the *GLUL* gene from the second *in vitro* mPAT. **F.** An ePAT [121] of all 4 cell lines (n = 3) grown *in vitro*. A change in expression can be observed, but no shift in the distribution of the bands suggests that there is no poly(A) tail length change.

combinations 1, 4 and 6 but not in combinations 2, 3, and 5 suggesting that a single primer was not the cause of the problem. It was unlikely that there were multiple ‘bad’ primers in the mix, as there were never usually any issues with primers ordered from Integrated DNA Technologies, Inc. (IDT). The fact that some mixes had now begun to produce products in the expected size range suggested that a lower total number of primers in the mix favoured amplification.

One aspect of the mPAT experiment that is not normally controlled for is the concentration of each individual primer in the forward primer mix. While the same total volume (1µl) of forward primer or forward primer mix was added in Figure 3.18 A-C, the concentration of each individual primer varies depending on the number of primers in the mix (as the same volume of primer mix is always used). Closer inspection of the protocol for the AmpliTaq Gold 360 master mix showed a suggested concentration of 0.2-1 µM of each primer. According to the

mPAT protocol, the primer (staring at a concentration of 100 µM) is diluted 1 in 10 prior to use and then 1 in 100 in the reaction itself. Each individual primer is also diluted by the total number of primers make up the pooled forward primer mix. The final concentration of each of the 61 primers in this experiment is outlined below:

$$\frac{100\ \mu\mathrm{M}}{61 \text{ x } 10 \text{ x } 100} = 0.00164\ \mu\mathrm{M}$$

When the full primer mix was used, the final concentration of each individual primer in the reaction was 0.00164 µM. This value is 122 - 610 times lower than the recommended primer concentration range for the AmpliTaq Gold 360 master mix, and may have been insufficient

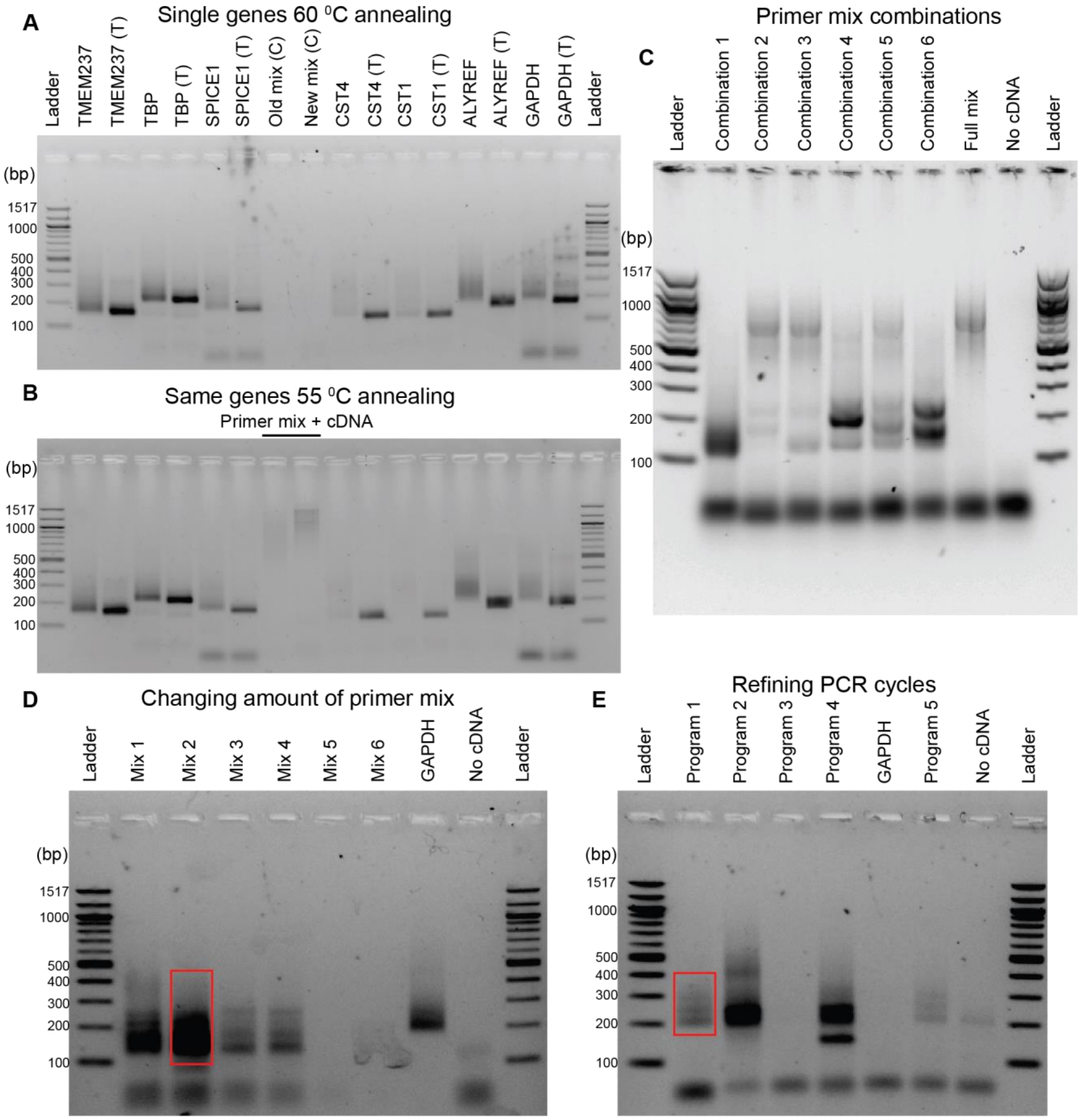


Figure 3.18. Optimisation of the mPAT method for lower abundance genes. Various steps of the mPAT protocol [215] were modified in an attempt to obtain a PCR product from lower abundance genes. **A.** PCRs targeting the 3' UTRs of single genes present in the failed primer mixes with a 60°C annealing temperature. **B.** The same as A except that cDNA is present within the PCR reactions containing the full primer mix and the annealing temperature of the PCR reaction was reduced to 55°C. All PCRs were performed using cDNA generated with 1 µg of RNA from the NI cell line as part of the mPAT protocol using the mPAT primer or the T12VN primer [211]. **C.** PCRs using smaller subsets of the full mPAT forward primer mix. **D.** PCRs of the full mPAT forward primer mix at varying primer concentrations (both forward and reverse). **E.** PCRs of the full mPAT forward primer mix at the optimal concentration with varying PCR cycle numbers.

for efficient product amplification, especially for low abundance genes. The full primer mix was therefore tested at varying concentrations of the forward and reverse primer at 35 cycles (Figure 3.18 D). The exact primer concentrations that were used are shown in Table 2.2. The most product was clearly obtained in mix 2 (Figure 3.18 D red box) and thus, a concentration of 0.1 µM of each individual forward primer and 0.5 µM for the reverse primer mix was selected as optimal for mPAT reactions primed from MXM cell line cDNA.

The last step in the optimisation process was to determine an appropriate number of PCR cycles to obtain enough mPAT product without introducing PCR artefacts through over-cycling. Ideally, an mPAT would be performed from cDNA in the exponential phase of PCR amplification. This can be visualised as a faint smear on an ultrapure agarose gel when the product of PCR cycle 2 is run. The number of PCR cycles required to produce this band was evaluated in both PCR steps (Table 2.3). As shown in Figure 3.18 E, 7 and 17 cycles in mPAT PCRs 1 and 2 respectively (program 1, red box) produced an appropriately sized smear at the correct intensity. When applied as part of the mPAT protocol, effectiveness was similar at 5 and 19 cycles and so the protocol was adjusted only by increasing primer concentrations in the first PCR to 0.1 µM (per primer) forward primer mix and 0.5 µM mPAT anchor oligo. This change produced products that were then successfully measured by sequencing.

3.3 Discussion

In the breast cancer subtypes for which gene expression-based prognostic tests are available, there is often still insufficient information provided that would alter a treatment decision for the patient [165]. In particular, there are currently no prognostic markers for the TNBC subtype approved for clinical use, a problem compounded by the heterogeneity of tumours within this subtype, which yields inconclusive results in gene expression-focussed studies. The 3’ end of mRNAs has been suggested as a novel prognostic biomarker that may potentially address these challenges [66, 113, 120]. To study mRNA 3’ dynamics in a controlled and reproducible environment, a xenograft model was generated from increasingly metastatic TNBC cell lines in immune-compromised mice. Alterations in primary tumours were studied in this model using custom RNA systems biology methods, developed in-house, to determine consistent metastasis-associated changes in 3’ UTR regulation.

3.3.1 PAT-seq accurately measured APA and gene expression in the MXM

The tumours used in the PAT-seq study of the MXM were all derived from an identical MDA-MB-231 TNBC parental cell line and were FACS sorted from host mouse cells prior to RNA extraction. This allowed the study of a clean model of the factors that were directly contributing to the metastatic potential of these tumours. Changes to mRNA metabolism were primarily observed, reflected in mRNA and APA changes. However, possibly due to methodological limitations, no consistent changes to poly(A) tail length were observed. To compare these results with human TNBCs, the gene expression of APA factors taken from a panel of TNBCs from the TCGA was also studied. APA effects were present in samples with differing metastatic potential, however, proximal APA was not always the major trend, in contrast with previous suggestions [112, 109, 113]. The results of GO analysis additionally suggested that the most metastatic TNBCs in this model and the TCGA have the highest expression of splicing and APA factors (Figure 3.9 C and D). While these effects were generally similar in terms of enriched gene sets, metastasis was achieved using different genes in each dataset.

PAT-seq appears to have measured APA sites quite accurately in this study. There was a large amount of overlap of PAT-seq APA sites (~50%) with the APADB (Figure 3.3 C). This overlap, and the presence of the canonical 'AAUAAA' polyadenylation signal, or a close variant, ~30 bases upstream from each poly(A) site suggested that PAT-seq was able to accurately detect mRNA 3' ends. The 24,730 non-overlapping APA sites between PAT-seq and the APADB may represent novel, previously unannotated APA sites or other polyadenylated transcripts such as some long non-coding RNAs [236]. It is also possible that some of these sites may represent internal priming to A rich stretches of the genome not filtered out by *Tail-tools* or the APADB. This possibility was also highlighted in the PAT-Seq poly(A) tail length measurements, with more consistent tail lengths observed after ~12 bases (roughly the length that would be expected from internal priming). It should be noted that *Tail-tools* removes genomic poly(A) stretches in PAT-seq reads, and so internal priming must either extend these regions with nontemplated A residues from the Illumina adapter sequence or, in some cases, the adapter may bind imperfectly.

The rationale for using PAT-seq in this study was that it can measure genome-wide 3' UTR switching, gene expression, poly(A) tail length distribution change and is relatively inexpensive to perform (runs on a single lane of an Illumina Flow Cell). The greatest practical strengths of PAT-seq are that it requires no special access to sequencing machines (as is the case for similar methods such as PAL-Seq [68]), is easily multiplexed (thanks to compatibility with indexed sequencing primers) and can be performed by most laboratories without any special equipment. In terms of measuring gene expression counts and considering the differences in the collection methods of measuring gene expression in both the TCGA and PAT-seq datasets (frozen human tumours vs a xenograft model), a Spearman's correlation of 0.68 (Figure 3.10 B) suggests that the PAT-seq and RNA-Seq methods are quite similar. To put these results in perspective, microarray and RNA-seq technologies can have Spearman's correlations of ~0.75 for overall gene expression when measuring the same sample [237]. It would be expected that both methods would have better concordance when measuring DGE, as opposed to absolute gene expression, as platform-specific biases do not need to be accounted for. This was not the case for MXM and TCGA data, however, this was likely influenced by the vastly different growth conditions and survival pressures between primary human tumours and a xenograft model generated using immortalised cell lines.

3.3.2 APA events with increasing metastasis in TNBCs

Previous studies of APA in breast cancer have found 3' UTR shortening to be associated with poor prognosis [136, 66]. These reports have identified an important trend, but likely represent an oversimplification of the role of APA in breast cancer. As previously discussed, breast cancer is a complex, highly heterogeneous disease, and APA that may vary with breast cancer subtype or even from patient to patient was not studied. Currently, breast cancer subtype is determined from both clinical markers [167] and gene expression signatures [45], and it therefore, stands to reason that the different subtypes would also likely have different APA profiles. Indeed, the TNBC subtype has been suggested to have the greatest levels of APA [110] and is the least well-characterised in terms of biomarkers, indicating that the study of APA in TNBC could yield potentially useful findings. Metastasis-associated APA was found in the two most metastatic lines in this study, suggesting APA is present in the progression of primary TNBCs.

No APA events could be found when comparing the NI samples to the other lines from PAT-seq run 2. This lack of statistical significance was likely due to the low number of biological replicates (N = 2) in the NI samples. This low sample number was due to difficulties in inducing primary tumour formation in the NI samples, owing to their low invasiveness and proliferation in mice. Nonetheless, the trends in overall APA effect values suggested a dynamic pattern of APA events in the metastasis of a primary tumour. Stronger evidence of APA was obtained when comparing the LM2 and HM tumour lines with the LNA line over both PAT-seq runs. For these comparisons, 88 statistically significant APA events were found. This is likely an underestimation of the true amount of APA, largely due to lower confidence in the variability of APA sites with lower sequencing depth. This could potentially be remedied in future by utilising RNA inputs > 2 µg in the PAT-seq experiments and fewer PCR cycles, although it would be more difficult to obtain greater volumes of RNA from tumour samples that may often be relatively small (< 2cm^3). An increased number of replicates would also enable greater power in distinguishing metastasis-associated signals from background variation. It is encouraging, however, that this APA signal was not lost over two distinct PAT-seq batches.

Interestingly, while proximal APA was the trend in the HM line as expected, distal APA was more common in the LM2 line (Figure 3.5 C). This suggested that there may be some key APA events associated with different metastatic tropisms within this model (HM is highly metastatic to the lung, liver and spleen and LM2 tends to metastasise to the lung). If the LM2 and HM lines were once again derived in the same way and if the experiment was repeated, the same metastatic profiles may not be selected for. This is because there may be an element of chance to the selection of a successful phenotype, that may achieve metastasis through alternative means, and then be propagated throughout the selection process.

Overall, however, the APA trend was similar between the two metastatic samples (Figure 3.5 D). A general signature of proliferation may, therefore, be governed by a more general APA trend as has been suggested in the literature [112], with greater APA trends governing specific metastatic tropisms. It is also possible that the site of metastasis is independent of APA and that APA events are simply assisting the tumour through increased proliferative capacity.

The results reported in the LM2 xenograft tumour line analysed here are not the first case of proliferative cells exhibiting a shift toward primarily distal APA. An *in vitro* study by Fu *et al.* found predominantly 3' UTR lengthening events in the MDA-MB-231 TNBC cell line and shortening events in the ER+ MCF7 cell line when compared with the MCF10 mammary epithelial-derived cell line [168]. Much like the present study, Fu *et al.* had the advantage of being able to measure APA in all genes in a targeted manner using a 3'-focused sequencing method, providing additional evidence of dynamic APA in cell lines. It should be noted that as APA is associated with proliferation [112], immortalised cell lines likely exhibit artificial APA profiles. By examining proliferation-associated APA events in derivatives of the same cell line, using the 3'-focused, genome-wide PAT-seq methodology, this study retained the power to detect metastasis-associated events, above and beyond the general proliferation-associated APA events that may be expected in cell lines.

The loss of regulatory elements by 3' UTR shortening has been the primary focus of the APA field recently. Collagen genes *COL1A1* and *COL1A2* both had significant 3' UTR shortening in the LM2 line in this study. This effect may be indicative of alterations to the cytoskeleton and extracellular matrix remodelling [238] prior to EMT and metastasis. Another example of a shift to proximal APA in the literature is the loss of Pumilio RNA binding protein complex sites in TNBC through 3’ UTR shortening, which was shown to increase protein levels of target genes [110]. In contrast to these findings, however, it is becoming more apparent that it is just as likely that longer 3’ UTRs may include mRNA stabilising elements [239, 188, 189]. The loss of AU-rich elements during proximal APA is one such example. AU-rich elements are present in as many as ~22% of 3’ UTRs of mRNAs [240], have been previously associated with invasive breast cancer [241], and represent only one contributor to mRNA stabilisation. When testing genes that underwent APA in this study no enriched gene sets could be found. This lack of direction, along with the variable APA events seen here and elsewhere [168] and the confounding effects of proliferation-associated APA from the immortalisation of cell lines [112], makes it difficult to untangle the purpose of genes with altered APA profiles in cell line studies. It is, however, clear that there is no consistent pattern of metastasis-associated APA events in any one pathway of this model system.

3.3.3 Metastasis-associated APA selects for loss of miR-22-3p binding sites

The miRNA binding sites lost with proximal APA were also analysed in this study. There was a significant reduction in the number of miR-22-3p binding sites associated with 3’ UTR shortening between the NI and HM tumour lines (Table 3.1). A known tumour suppressor, miR-22-3p represses cancer progression through the induction of cellular senescence [233]. Interestingly, when the expression of common breast cancer-associated miRNAs was evaluated by the alPAT method, the expression of 4 miRNAs was significantly altered, however, miR-22-3p was not among them (Figure 3.7, highlighted in orange). The miRNAs that did undergo a change in expression include the downregulation of miR-126-3p, which is known to suppress breast cancer metastasis [242], miR-342-3p, which has been associated with tamoxifen-resistant breast cancer [243] and miR-26a-5p, which has been suggested to inhibit breast cancer proliferation and invasion [244]. The only miRNA that was upregulated was miR-17-5p, which is a known oncomiR that has been associated with breast cancer invasion through the repression of the *HBP1* cell cycle gene [245]. It is therefore suggested that in the MXM, APA alone was used to escape miR-22-3p-mediated regulation of cellular senescence. This is an example of a tumour escaping regulation through altered mRNA processing rather than the differential decay of long transcripts by altering miRNA expression. It is also likely that miR-22-3p binding sites are not the only miRNA binding sites that are preferentially lost with increasing metastatic potential in this model. Were this experiment to be repeated with additional tumours, other miRNA binding sites may be shown to be preferentially lost. Unfortunately, time and cost constraints associated with the generation of the MXM and PAT-seq analysis of tumours limited the power of this analysis but nonetheless revealed a novel mechanism for the escape of miR-22-3p-mediated repression of metastasis.

3.3.4 Poly(A) tail length measurement by PAT-seq in the MXM

Poly(A) tail length has been shown to vary dynamically depending on cellular conditions [125, 246, 68]. It was therefore surprising that poly(A) tail length could not be associated with metastasis in this study. Previous studies of poly(A) tail length have also suggested a positive association with mRNA half-life [117, 68]. Here, however, poly(A) tail length was found to be only weakly negatively associated with mRNA abundance and was not associated with APA. While it cannot be inferred from these findings that mRNA half-life is not associated with poly(A) tail length, it can be inferred that, as a general rule, more abundant transcripts do not have longer poly(A) tails under normal cellular conditions. Differences that were found in poly(A) tail length for specific genes were also not reproducible between sequencing runs of PAT-seq experiments or by the targeted mPAT and ePAT methods. There was also a poor correlation between PAT-seq and mPAT tail lengths when measuring *in vitro* cell line samples vs the MXM tumours. Combined, these data suggested that the study of poly(A) tail length in tumours will require greater sample numbers and more controlled conditions before any differences may be found.

The PAT-seq method is designed for robust and powerful measurement of differential poly(A) tail length on a genome-wide scale. The cDNA reads obtained from the library preparation step of PAT-seq will be on average ~150 bp long and each gene will have its own distribution of poly(A) tail lengths. This distribution will vary depending on the age of transcripts and the stabilisation of mRNAs in the cytoplasm by cytoplasmic polyadenylation proteins [247]. As the RNase T1 used in PAT-seq cleaves mRNA after G bases, the sequence composition of the 3' UTR of a gene will also play a role in how long a sequenced cDNA transcript is. The proportion of poly(A) tail lengths that can be captured in a PAT-seq experiment will, therefore, be altered by the sequence composition of the cleaved 3' UTR. PAT-seq is also limited by the effects of polymerase slip [235], a PCR artefact that occurs when amplifying repetitive sequences that has the potential to underestimate the length of the poly(A) tail and cause it to appear shorter on average. A further limitation of PAT-seq is that it is not able to measure poly(A) tails > ~120 bases due to the 150 base sequence constraints of the method. This is especially problematic when measuring the mammalian poly(A) tail, which has been suggested to be as long as 250 bases in length [248, 249]. As size selection is roughly equal across a given experiment, PAT-seq should be able to detect a change in poly(A) tail length distribution but should not be relied upon to accurately detect absolute poly(A) tail length in mammalian samples.

In addition to the mPAT results, MXM PAT-seq poly(A) tail lengths did not correlate at all with the poly(A) tail lengths measured by TAIL-seq sequencing of HeLa cells [117] (Figure 3.14 D). TAIL-seq is similar to PAT-seq in methodology (except that is uses pair-end as opposed to single-end sequencing) and it would be expected that if specific genes tended to have comparable tail lengths in mammalian systems, just as they had similar gene expression levels in the TCGA comparison, that this would be reflected in the results. This variation between methods has also been observed in laboratory yeast strains (*Saccharomyces cerevisiae*, W303 vs BY4741). When PAT-seq was compared by Harrison *et al.* with PAL-seq [68], another poly(A) tail length measurement method, a correlation of only 0.3 was obtained. The aforementioned difficulties with measuring poly(A) tail length using short-read sequencing may go some way to explaining these results, however it is likely that there are more factors contributing to poly(A) tail length such as the circadian rhythm of cells, which has been shown to induce global tail length changes [246].

3.3.5 Gene expression changes with increasing metastasis in TNBCs

Gene expression was by far the clearest variable measured by PAT-seq that was altered with metastasis in the MXM. Almost 3000 genes had significant metastasis-associated changes in the MXM (run 2). Despite batch effects, samples grouped by cell line rather than by batch on a *Limma* MDS plot (Figure 3.2 B). *Camera* gene set enrichment analysis (GSEA) revealed that many of these gene expression changes were associated with the upregulation of mRNA processing genes or the suppression of immune signalling genes. These changes were compared with increasingly metastatic TNBCs from the TCGA to determine if these effects were consistent with the metastatic state of human tumours. In general, there was poor agreement between the genes that were differentially regulated between the MXM and the TCGA (Figure 3.10 C). There were, however, the same types of changes occurring in both systems (Figure 3.11 D) suggesting that the MXM had somewhat effectively recapitulated an increasingly metastatic panel of primary patient TNBCs.

Overall there were 273 more enriched gene sets in the TCGA compared with the MXM (Figure 3.11 C). The larger numbers of enriched GO term gene sets with metastasis are likely for two reasons: first, the MXM lines are all derived from the MDA-MB-231 line and therefore likely have less underlying biological variation to begin with, and second, there are many more samples in the TCGA, which may result in lower p-values through added statistical power. Furthermore, there was suppression of immune related gene sets such as interferon signalling present in the MXM but not in the TCGA (Figure 3.11 A/B). These changes may reflect better adaptation of the human tumours to surviving in the murine host. The BALB/c-SCID mice are largely immune-suppressed, however, these mice still likely have some remaining immunity as they do not spontaneously develop tumours [201]. Human tumours would have already had to escape the host immune system to avoid detection during development and, as such, these gene sets were not required for increased metastatic potential. It is, therefore, possible that metastasis in the MXM was governed by both adaptation to survival in a host environment as well as metastatic gene expression changes that are more broadly conserved in patient primary tumours. The increase in the expression of RNA processing factors combined with APA in TNBC metastasis in this model suggested that altered mRNA processing and resulting APA is a key driver in metastasis and should be further studied in human tumours free from the confounding proliferation-associated effects present in cell line models. It is also interesting that the top APA changes (distal in LM2 and proximal in HM) had different trends in tumours that both had a consistent increase in the expression of RNA processing factors, which was expected generally to be associated with only proximal APA [66, 113].

The data presented in this chapter clearly shows that APA is dysregulated in primary TNBC metastasis, although there is still more to uncover regarding the role of APA in TNBCs with increasing metastatic potential. It remains to be seen whether APA could be used as a novel classifier in the TNBC subtype and whether it could add prognostic power in addition to gene expression signatures. The clearest way to determine APA at the patient-specific level would be to apply a genome-wide 3’ sequencing method such as PAT-seq to a large breast cancer patient cohort. This could potentially be done with the addition of small RNA-Seq or an additional sequencing method able to more accurately capture the expression of mRNA regulatory elements. In the interim, confirming APA events seen in the reanalysis of TNBC microarray data [113, 110] and in RNA-seq data with novel methods such as those employed by Xia *et al.* [66] (discussed in Chapter 4) may give additional confidence to observed APA trends at the transcriptome-wide scale, and enable the generation of effective prognostic models.

Chapter 4: Inferring APA events from public databases and predicting breast cancer prognosis

4.1 Introduction

Gene expression is well known to be altered in breast cancer [56] and, therefore, thousands of experiments have been performed in order to better understand these alterations. There are now large collections of breast cancer gene expression data from RNA-Seq and microarray technologies, which are present in repositories such as the TCGA [166] and the GEO [250]. If reanalysed correctly, however, these datasets can also be exploited to infer APA [136, 113, 66, 65]. As discussed in Chapter 3, it has been a common claim of these studies that APA is prognostic of breast cancer outcome, in some cases, even more so than gene expression [66], which is currently the gold standard for breast cancer prognostic testing [251]. While there are accepted standards for the analysis of gene expression data, the study of APA is comparatively new. This has meant that there is currently no accepted standard to infer APA from gene expression data and recent studies of APA in breast cancer have all employed their own novel methods to detect APA. When compared with one another, they often show little agreement as to the specific APA events that are associated with breast cancer. To determine breast cancer-associated APA events with greater certainty, the present study employed novel APA calling pipelines. These pipelines attempted to utilise consistent statistical methods and minimise false positive errors when calling APA from both RNA-seq and microarray datasets.

The idea behind detecting APA from RNA-Seq is relatively simple and relies on looking at variations in genomic coverage (the number of reads that overlap a given portion of the genome, Figure 4.1 A). The RNA-Seq coverage profile of short-read sequencing methods is known to be non-uniform across the length of a gene [176]. This variability of RNA-Seq coverage also extends to the 3’ ends of genes, especially in poly(A)-selected libraries such as the TCGA [1]. This difference in coverage can be due to technical bias, associated with the transcript sequences themselves, but may also be caused by the differing expression of alternative isoforms of the same gene, which are generated through processes such as AS or APA. This variability can be detected by software when looking at differences in coverage across the same gene when comparing two or more conditions [203]. Methods that call APA from RNA-Seq exploit these changes in coverage by looking for changes at the 3' end of genes (between groups of samples) at potential APA sites. Potential APA sites are defined either through *de novo* methods or using databases of known APA sites [252, 181] that were previously determined from 3'-focused sequencing methods similar to PAT-seq [169].

Xia *et al*. attempted to infer APA events with all paired tumour-normal samples in the TCGA using their novel dynamic analysis of alternative polyadenylation from RNA-seq (DaPars) method [66]. The DaPars method involves inferring a proximal APA site from RNA-Seq coverage (with the distal peak taken from a modified database of APA sites) and then computing a percentage distal usage index (PDUI). The PDUI is computed by comparing the ratio of RNA-Seq read counts, at two APA sites, for a single gene, in a given sample. The change in this ratio between conditions is termed the 'change in percentage distal usage index' (ΔPDUI). Significant APA events are then called from reproducible changes in this ΔPDUI between samples. This was a landmark study, but did not attempt to focus in detail on any one cancer type and ignored many of the unmatched tumour samples that are also available for analysis in the TCGA.

Reanalysis of microarrays at the probe level has also been utilised to call APA from existing microarray data in breast cancer and involves similar, but slightly more complex methods than are used to call APA from RNA-Seq [136, 113, 65]. These studies were based around Affymetrix arrays that are comprised of multiple 25 bp probes that bind to fragmented cDNA.

As discussed in Chapter 1, more binding of a transcript to a probe will cause more fluorescence on the array and more signal to be reported for that probe. Probes are organised into collections (usually all binding to sequences in a ~200 bp genomic region) known as probesets, with one or more sets used to measure the expression of a given gene. Similar to the analysis of APA from RNA-Seq, the changes in the amount of RNA present at the 3’ end of a transcript will cause a difference in signal between samples [136]. Methods that measure APA from microarrays have previously relied on reorganising probes into custom probesets, defined around known 3’ ends of transcripts [253]. These methods are, of course, limited to array types that have probes designed to the 3’ end of mRNA transcripts. They also suffer from the drawback that they are not able to utilise historical methods developed to more accurately normalise and measure the standard gene expression from the original probesets. As only a small number of studies have examined APA in breast cancer, even fewer studies have examined APA in breast cancer subtypes. Only the microarray studies by Akman *et al.* [113] and Wang *et al.* [176] have studied APA in the TNBC subtype and both are limited in the genes that they can test, due to the aforementioned requirement for probesets around a given APA site. Reanalysing breast cancer gene expression data, with a focus on subtype, may yet provide novel breast cancer-associated APA events or prognostic markers.

An appropriate modelling technique was required to potentially derive new prognostic models from the APA events determined in this study and test them against previous claims that APA is prognostic of breast cancer outcome [66, 113]. In the biological context, where there could potentially be thousands of predictors (such as the expression of ~20,000 genes), it is usually desirable for a statistical model to be sparse, meaning most of the values in the model are zero. Sparse models have the advantage of being more easily interpretable by researchers, owing to the smaller number of predictors to consider. A reduction in predictors is also necessary in the p >> n scenario where there are many more predictors (p) than samples (n), another common problem in biological datasets that measure changes associated with every gene or protein in a sample [254].

An optimal modelling approach, in the context of this thesis, would only select predictors, such as clinical data, gene expression changes or APA events, that have good prognostic power, while ignoring predictors that do not. Two prominent examples of methods that moderate the impact of predictors in a linear model include ridge regression [255] and the lasso [256]. Both methods apply regularisation (application of penalty values to predictors in a model) to a list of predictors, however, they differ in their implementation. The key difference between both methods is that the lasso penalty forces at least some predictors to be zero (to have no contribution to the model), whereas ridge regression utilises all predictors, but shrinks the contribution of some versus others. This makes the lasso the clear choice for obtaining a sparse model with a shrunken list of interpretable predictors. In experiments involving complex biological data, the best linear models will be formed using predictors from a range of biological pathways. This methodology is superior to those that simply identify the predictors with the highest correlation with the response variable, as many of these predictors may be members of the most altered biological pathway. Methods such as the lasso allow for the capture of more of the biological variation driving the response variable (patient survival in this case), which is especially important in complex diseases such as cancer. The lasso method achieves this by including only one of a highly correlated set of predictors in the final model [256]. Unfortunately, when predictors are highly correlated, with similar predictive power, the lasso will select one at random, possibly removing some predictive power from the final model [257].

The best way to account for both feature selection, weighting and the production of an interpretable model, has been suggested to be elastic net linear modelling (ENLM) [257]. The ENLM approach uses a linear combination of both ridge regression and the lasso penalty terms. ENLM performs feature selection using the lasso penalty, while the ridge penalty averages out the effect from highly correlated genes, encouraging a grouping effect. The result is that when a gene from a gene family or distinct biological pathway is included in the model as a representative, other highly correlated genes will not be selected due to the lasso component. This selection can now also occur without the loss of statistical power that may occur in the lasso method by picking a predictor at random thanks to the ridge regression component.

In order to determine the best prognostic markers for disease outcome, including breast cancer survival rates, it is important for prognostic models that were generated using different datasets or methodologies to be comparable. Typically, model performance is measured by Cox proportional hazards modelling [258]. Cox proportional hazards modelling gives a coefficient indicating the overall usefulness of a predictive model in predicting survival and an associated p-value. One issue with the Cox proportional hazards model that may complicate the comparison of two models generated using different data, is that the coefficient is dependent on the scaling of the prognostic scores and it may be overly influenced by outliers (in terms of patient survival time). These problems are often solved by splitting the patients into two equally sized groups based on the median of the prognostic scores generated for each sample by the model. This results in a hazard ratio (HR) which is comparable across models generated using different prognostic scoring methods. The HR is a measure of the risk of a patient dying at any short interval of time across the indicated time frame. The results of this type of median split are often visualised using Kaplan-Meier plots.

The generation of a hazard ratio does, however, discard a lot of information contained in the prognostic scores by reducing them to a binary value (greater or less than the median).

Due to the information loss associated with the hazard ratio that is often generated as part of Cox modelling, the D index was chosen as the measure of the prognostic power of a prognostic model in this chapter [259]. The D index is also calculated via Cox proportional hazards modelling, however, it is computed using the scaled rankits (expected values from order statistics) of the risk scores, rather than the base risk scores themselves. The idea behind this approach is to normalise the predictor by rescaling it and forcing any outliers to fall into a normal distribution. This allows the coefficient generated by the Cox method to be used as an effect size, resulting in more accurate and comparable models. The D index is also slightly scaled to aid in its interpretation as a log HR. Furthermore, the D index has an intuitive approach to approximating the HR between two groups split by the median of a predictor (referred to as the D index HR, which is calculated by taking the exponential function of the D index). It can be viewed as similar to the HR obtained by performing a median-split. When compared with the Cox model generated from this median split, the D index is smoother, using more information from the predictor.

Presented in this chapter are the results of defining APA from RNA-Seq and microarray datasets using new methodologies. Importantly, the APA results presented here were also compared both within this study, as well as with previously published breast cancer APA datasets [66, 113], finding the most, and best supported, breast cancer APA events discovered to date. As it was the largest dataset with the best annotation, the TCGA RNA-seq APA data was further utilised, in combination with gene expression and clinical data, to enhance the prediction of breast cancer outcome. The results of this work suggest that APA can be used to form the basis of a prognostic test for breast cancer outcome.

4.2 Results

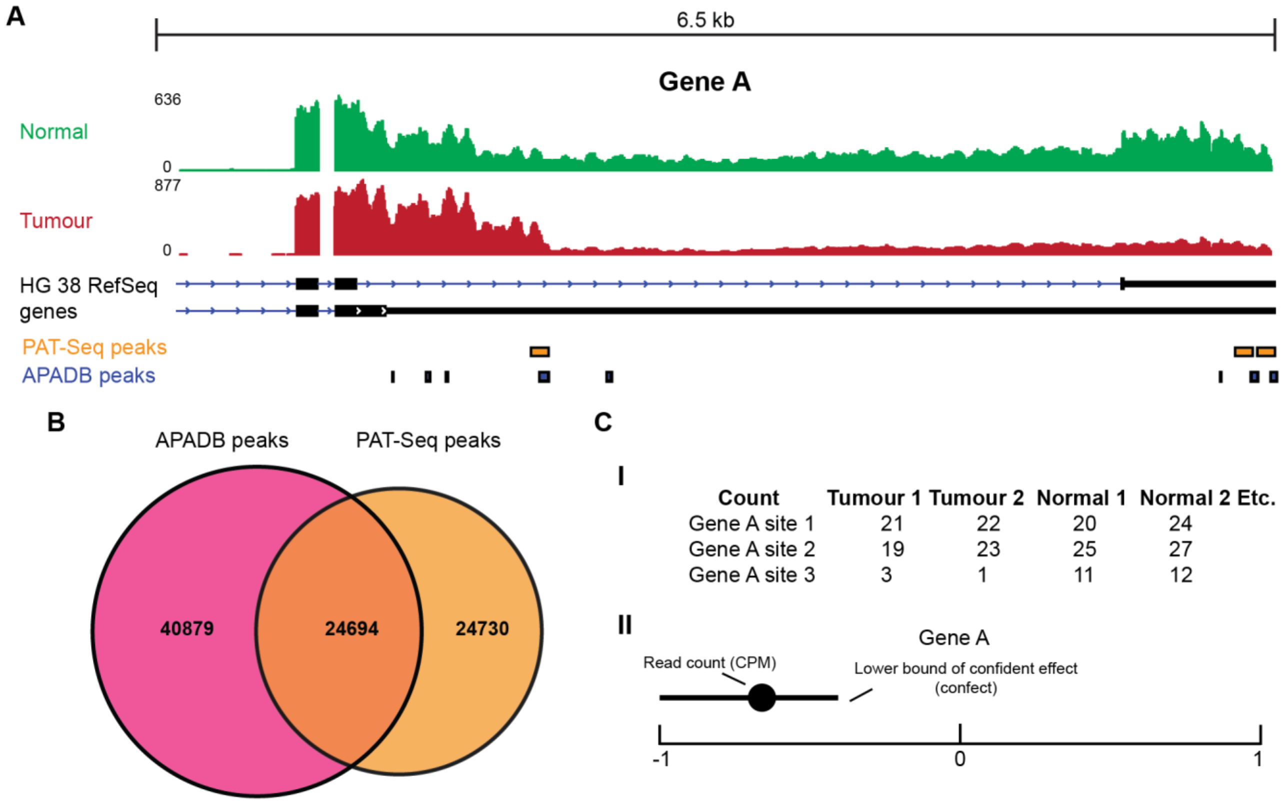
Multiple datasets that infer APA from novel RNA-Seq and microarray data were analysed in this chapter. Table 4.1 is provided as a reference to easily determine the APA calling methods and results that are described.

Table 4.1. The five methods used to infer APA from RNA-Seq and microarray datasets discussed in this chapter. This table is to assist with reference to studies and methods that were used to infer APA from the TCGA and GEO.

APA calling method	Dataset analysed	Data source	Description	Referred to as
Log short/long ratio	Unmatched tumour and normal breast tissue samples agglomerated from the GEO	Akman *et al*. [113]	Method for calling APA from Affymetrix microarrays	SLR
Dynamic analysis of alternative polyadenylation from RNA-seq	Matched tumour and normal breast tissue samples from the TCGA	Xia *et al*. [66]	Method for calling APA from RNA-Seq	DaPars
Original probeset ordering	Unmatched tumour and normal breast tissue samples agglomerated from the GEO	This thesis	Method for calling APA from Affymetrix microarrays	OPO
Database counting (matched)	Matched tumour and normal breast tissue samples from the TCGA	This thesis	Method for calling APA from RNA-Seq	DBC-matched
Database counting (unmatched)	All tumour and normal breast tissue samples from the TCGA	This thesis	Method for calling APA from RNA-Seq	DBC-unmatched

4.2.1 Inferring APA from TCGA RNA-Seq

To obtain a detailed picture of 3’ end usage in breast cancer, APA was first studied in the TCGA, the largest available breast cancer dataset. TCGA primary tumours are not permitted to have undergone neoadjuvant therapy prior to collection, giving a consistently untreated starting point for analysis. APA was inferred from TCGA breast cancer data the ‘Homo sapiens (v2)’ database of annotated APA sites, downloaded from the APADB [181]. Read



Count	Tumour 1	Tumour 2	Normal 1	Normal 2	Etc.
Gene A site 1	21	22	20	24	
Gene A site 2	19	23	25	27	
Gene A site 3	3	1	11	12	

Figure 4.1. The process of inferring APA from TCGA breast cancers. A. A hypothetical example of RNA-Seq coverage from a primary tumour sample and a normal breast tissue sample in the TCGA. Displayed below are the APADB peaks used for inference of APA and the MXM PAT-seq peaks for comparison. **B.** Venn diagram depicting the overlap of APADB peaks and PAT-seq peaks (peaks must be on the same strand). **C.** The process of inferring APA from counts obtained in A. **I.** The number of reads that mapped to the genome were counted at each APADB site, for every sample. This process was repeated for all primary breast tumours and normal breast tissue samples. **II.** A 'confects' object was computed for each gene using the *topconfects* R package. The topconfects method computes the magnitude, direction and confidence bounds of detected APA events for each gene when comparing between two conditions.

counting was performed simply by counting the number of reads present at each APA site in each sample (Figure 4.1 A). This method was named the database counting (DBC) method.

As TCGA RNA-Seq data is unstranded, calling APA from overlapping genes can be challenging and potentially lead to erroneous results. To address this, the APADB peaks were intersected with the stranded, 3'-focused, PAT-seq MXM APA sites described in Chapter 3. In order to remove infrequent, but potentially misleading APA sites that may measure coverage from the wrong gene, sites from the APADB that overlapped the PAT-seq APA sites on opposite strands were removed. This method worked well, largely eliminating calls of statistically significant APA events from overlapping 3' ends, likely due to the 50% overlap between PAT-Seq APA sites and the APADB (Figure 4.1 B). Next, any APADB sites that overlapped the 5' UTR of the next downstream gene were also removed, to prevent APA events being called from reads corresponding to downstream genes on the same strand. This process resulted in a filtered database of APA sites, that could be utilised to vastly simplify the process of quantifying APA from unstranded RNA-Seq data. The *topconfects* method (Figure 4.1 C), was then used to compare both matched tumour vs normal samples (DBC-matched) and all tumour samples vs the mean APA state of all normal tumours (DBC-unmatched), to determine the shifted APA events with the greatest confident effect size. For a full description of the implementation of the topconfects method see Chapter 2, Section 2.2.4.

4.2.2 Proximal APA is pervasive in primary human breast tumours from the TCGA

The first objective in studying APA in breast cancer was to determine the difference between normal breast tissue and a primary breast tumour. Despite being ~10% of the size of the full TCGA breast cancer dataset (discussed in section 4.2.8), the DBC-matched method had the most statistical power when calling breast cancer APA (Figure 4.2 A), finding 914 significant

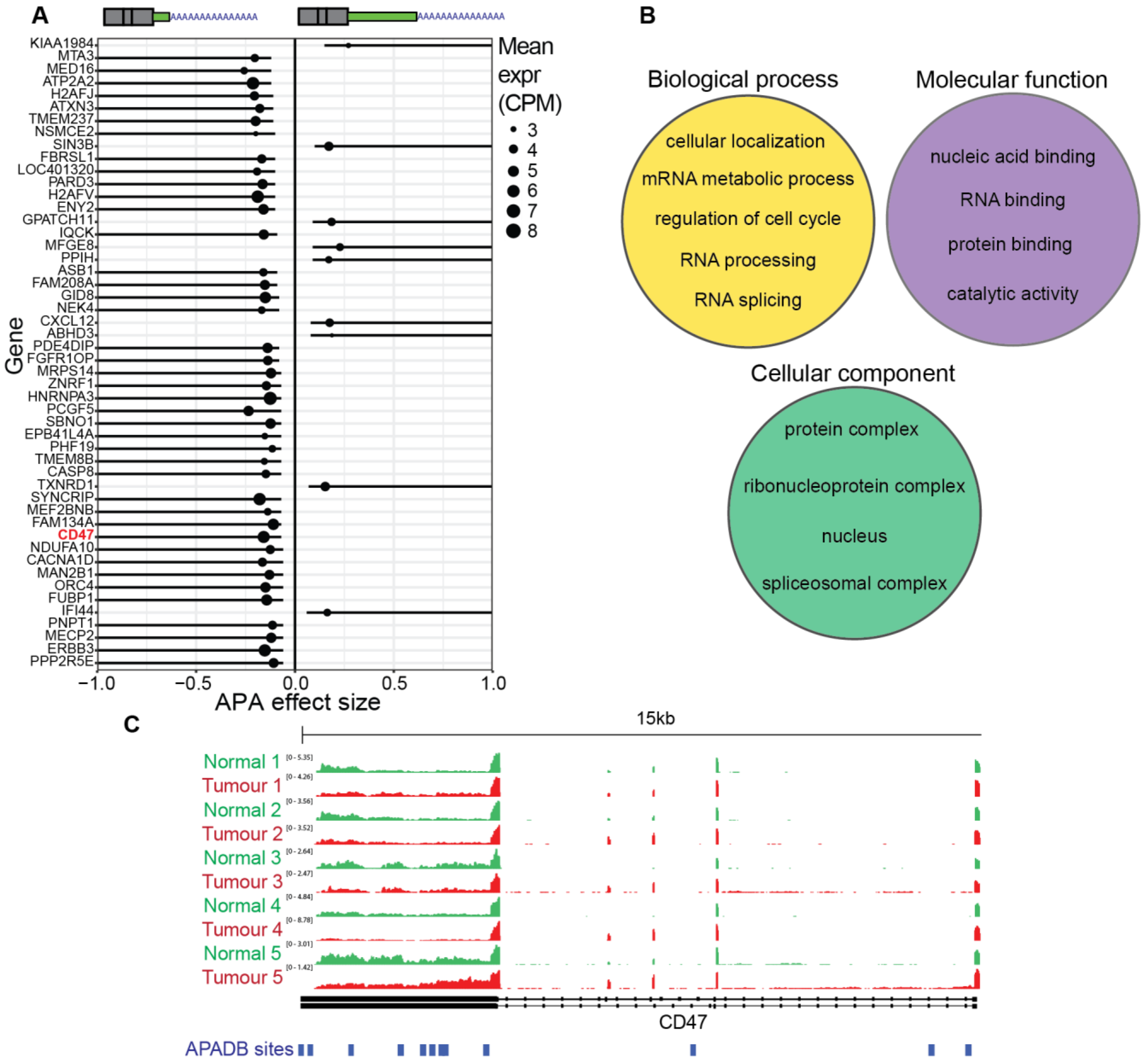


Figure 4.2. Significant alternative polyadenylation events from the DBC-matched tumour/normal analysis. A. A confects plot of the top 50 APA events from DBC-matched analysis. 914 genes were called in total. **B.** Selected GO terms called as significant (FDR < 0.05) by PANTHER overrepresentation analysis, comparing all 914 significant APA genes against all known human genes. **C.** Representative 3' read coverage (in CPM) of *CD47* from 5 randomly selected matched tumour/normal samples. Proximal APA is favoured in tumours for this gene.

APA events compared with < 50 in the full DBC dataset (FDR < 0.05, absolute confect ≥ 0). The full list of events and associated confect values can be found in Appendix A4. Some of the top APA genes included *MTA3*, which has previously been suggested as a key component of growth and proliferation through the estrogen receptor pathway [260]. Also called was the well-known cancer-associated cell surface receptor *CD47* [261] (Figure 4.2 C), the overexpression of which has been suggested to assist breast tumours in evading the immune system [262]. Interestingly, CD47 appears to have switched to preferential use of the proximal isoform in breast tumours, which as discussed in Chapter 1, leads to localisation away from the cell surface [194]. Consistent with previous reports [109, 66, 113], the overall APA trend in breast cancer was toward 3' UTR shortening, with distal APA present at 174 sites and proximal APA at 740.

To determine if APA genes were enriched in any biological pathways, PANTHER [218] overrepresentation analysis was performed using the 914 genes that had significant APA (at an FDR of 0.05). When compared to the default database provided by PANTHER (all human genes) there was a strong enrichment for pathways related to RNA processing (Figure 4.2 B). However, when the full list of genes that could possibly have had an APA event (genes from the APADB that passed expression filters) was used as the background, there were no enriched GO terms. This lack of enrichment with an APA biased background suggests a broadly auto-regulatory role for alternatively polyadenylated genes and that cancer-associated APA events may not be localised to specific pathways. This result is consistent with APA events in the MXM analysed in Chapter 3, which were also not overrepresented in any gene sets.

4.2.3 Inferring APA from Affymetrix HG-U133A microarrays without using a custom CDF file

In order to validate the breast cancer-associated APA events from the TCGA a new dataset was sought. APA was, therefore, also inferred from breast cancer and normal breast tissue samples from microarray studies that were previously uploaded to GEO. Samples used were a subset of those compiled by Akman *et al.* [113] (see Methods chapter Section 2.2.10 for a complete list of sample sources). A standard Affymetrix HG-U133A microarray is comprised of multiple probesets (which are, in turn, usually comprised of 11, 25 bp probes) designed to measure the expression of a target gene. An example of the structure of multiple probesets targeting a single gene can be seen in Figure 4.3. Standard practice for calling APA from Affymetrix microarrays has been to reorder probes into new probesets (post-experimentally), using a custom chip definition file (CDF). A database of sequences representing alternative forms of 3' UTRs is generated, probe sequences are mapped to this database, and a custom CDF file is generated with new probesets defined on either side of an APA site. This method has been used effectively in the past to study APA in breast and lung cancer, and pluripotent stem cell generation [136, 253]. The issue with this method, however, is that it may, in some cases, rely only on one probe to call APA, and does not take into account information that has been gained about the biases of the standard manufacturer-annotated probesets over time.

To address the aforementioned issues with defining custom probesets, a new method for calling APA form microarrays, known as the original probeset ordering (OPO) method, was developed in this study. The GEO dataset analysed by this method will be referred to as the OPO dataset. The OPO utilised gene expression information from the standard Affymetrix probesets, and ordered them based on their position within the genome (proximal-distal). Affymetrix arrays have both perfect match (PM) and mismatch (MM) probes. Probeset

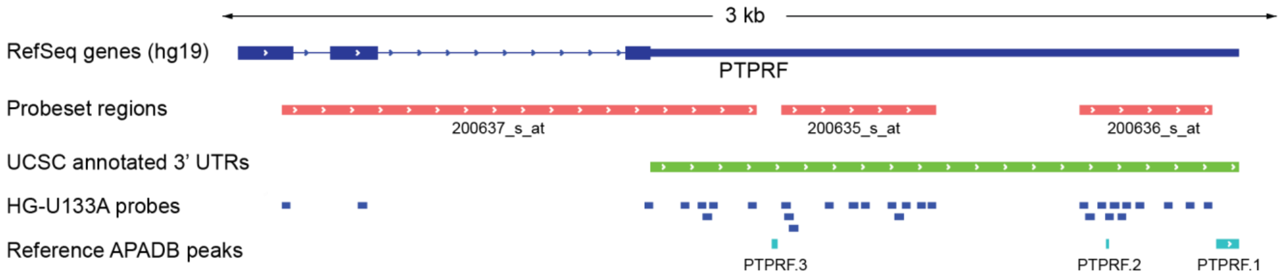


Figure 4.3. Example of grouped HG-U133A probesets that bind to a single gene. From top to bottom: The human genome (hg19), regions that cover a full probeset, UCSC known 3' UTRs, HG-U133A probe binding locations and known APA sites from the APADB.

normalisation was performed using the fRMA method from the *fRMA* Bioconductor package [175], which utilises only PM probes, as these are generally more accurate than older methods that also utilise mismatched probes [263]. The reason for using the fRMA method over the standard RMA method, the most popular method for PM probeset analysis, is that it utilises a large database of publicly available arrays. Probe specific effects and variances can be precomputed from this large database and then 'frozen' for use with new datasets as they arise. The fRMA method also outperforms the standard RMA method when applied to samples arising from multiple batches, as was the case in the datasets used here. Once the expression of each probeset had been established, the 'panp' function from the *panp* Bioconductor package [264] was then used, with default settings, to determine the likelihood that a gene was actually expressed (expression signal above background levels) [264]. Probesets were removed if they were not assigned a call of 'present' ($p < 0.01$) in at least 53 samples. This represented the lowest number of samples in a single sample type, although samples were not specifically required to be of the same type to pass this filter.

As a result of this processing, probesets could then be ordered by 3' end for further processing. A bed file containing the locations of each individual probe (human genome version hg19) was downloaded from the Affymetrix website. Probe locations were annotated from the start of the most proximal probe to the end of the most distal probe, so that every probeset was covered by a single bed file entry. The *Bedtools* 'overlap' function was then performed against the UCSC human genome (version hg19), keeping only probesets that at least partially overlapped a 3' UTR. Probesets were then ordered from proximal to distal based on this positional information (Figure 4.4). Similar to the TCGA RNA-Seq analysis, the topconfects $\log_2$ group effect shift function was then used to call significant 3' end shifting, with the APA state of all normal breast tissue as the baseline, as matched samples were not available.

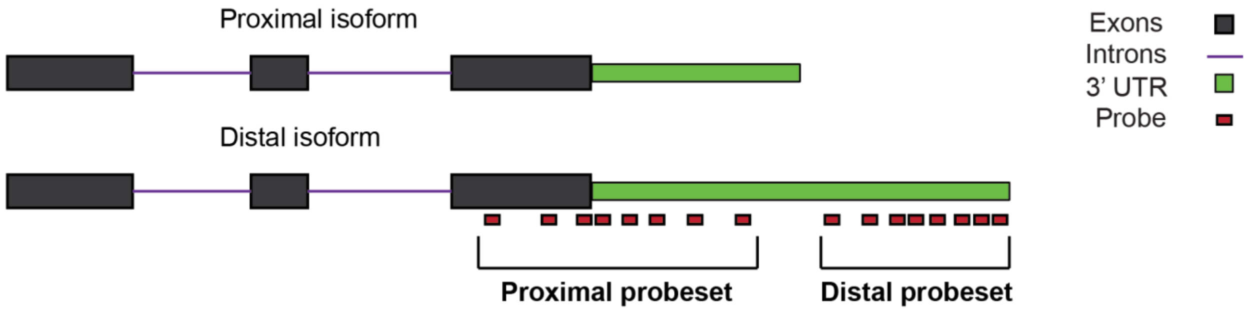


Figure 4.4. Inferring APA from HG-U133A by ordering standard probesets. Standard microarray probesets are ordered from proximal to distal in order to infer APA events, while still retaining the ability to utilise probeset expression information obtained about these probesets from previous experiments.

4.2.4 Proximal APA is pervasive in breast cancer APA events inferred from Affymetrix HG-U133A microarrays

To determine if measuring APA from microarrays resulted in a similar trend to the TCGA (Figure 4.2 A), the APA state of 376 breast tumour samples downloaded from GEO was compared with 53 normal breast tissue samples using the OPO method. Overall there were 994 APA events called between tumour and normal conditions. Once again, proximal APA predominated (Figure 4.5 A), with 836 proximal and 158 distal APA events observed. The gene with the greatest APA shift overall was *ATP2A2* (Figure 4.5 B), which was also shown to have a pronounced switch toward proximal APA when called from the TCGA using the DBC-matched method. Interestingly, in both datasets *ATP2A2* APA was consistent regardless of breast cancer subtype. *ATP2A2* has not previously been associated with breast cancer, however, mutations in this gene have been linked to a loss of adhesion between epithelial skin cells, suggesting a possible role in maintaining cell-cell contact [265]. The second greatest shift toward proximal APA was observed in the *FUBP1* gene (Figure 4.5 C), which acts as a splicing factor for the *MDM2* proto-oncogene, and is well known to be involved in the transcriptional upregulation of the *MYC* oncogene [266, 267]. The greatest shift to a distal APA event was observed in the *RUFY3* gene (Figure 4.5 D), which plays a role in neuronal polarity by suppressing the formation of excess axons [268]. *RUFY3* overexpression has also been associated with cell migration in invasive gastric cancer [269]. Combined, these results present a multitude of breast cancer-associated APA events that could be investigated further.

4.2.5 There is minimal agreement between competing methods when determining breast cancer associated APA events

To assess the reproducibility of the APA events called in the TCGA and microarray datasets, APA events were compared with breast cancer APA events from published studies. The first

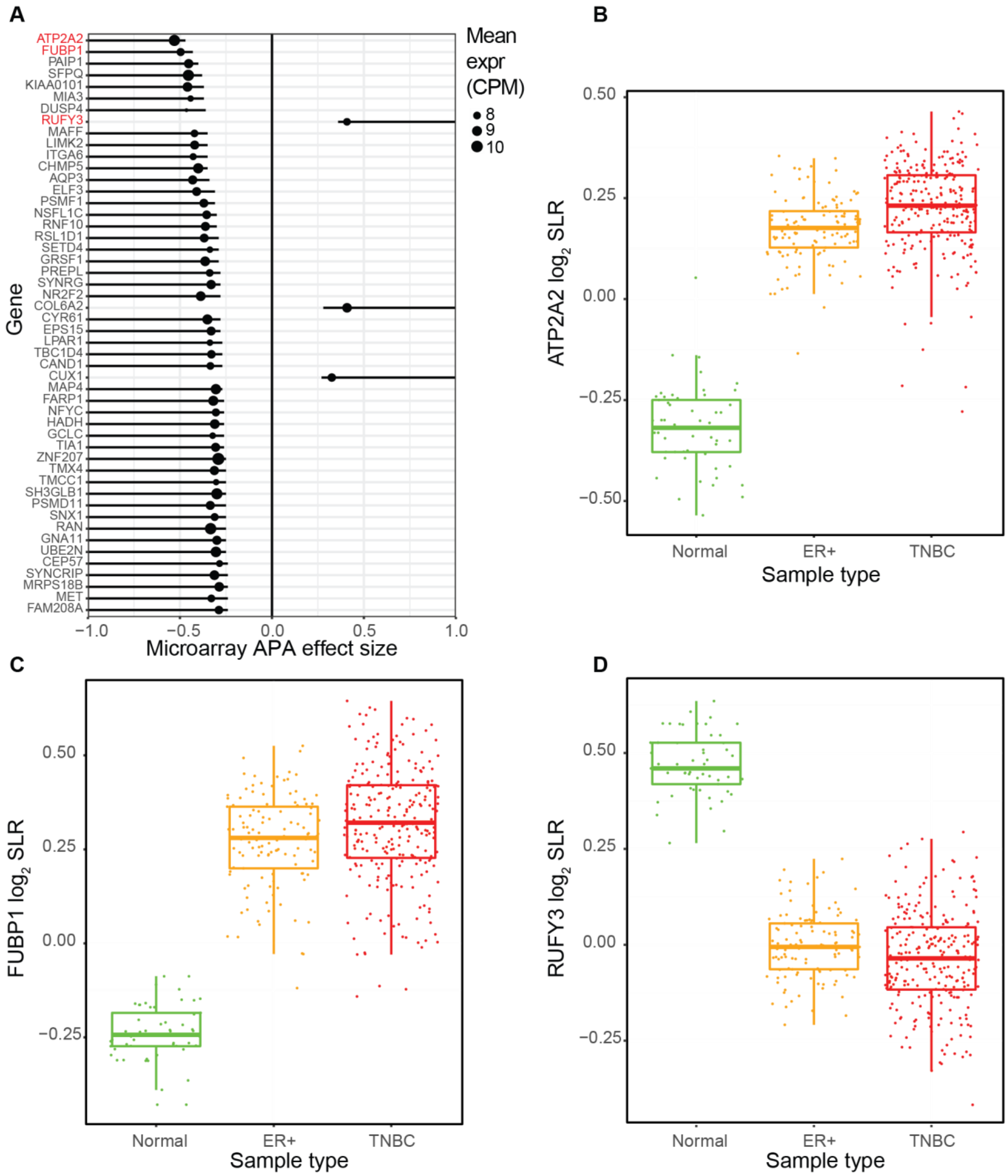


Figure 4.5. Breast cancer associated APA events as called from HG-U133A microarrays using the OPO method. A. The top 50 confect values for all tumours (ER+ samples from GSE2034 and TNBC samples from GSE31519) analysed using the OPO method as compared to all normal breast tissue samples (pooled from GSE9574 and GSE20437). Highlighted in red are the genes plotted in B-D. **B.** The $\log_2$ short/long ratio ($\log_2$ SLR) for the gene *ATP2A2*, showing a shift to proximal APA in tumour samples (both ER+ and TNBC tumours). **C.** The $\log_2$ SLR for the gene *FUBP1*, showing a shift to proximal APA in tumour samples. **D.** The $\log_2$ SLR for the gene *RUFY3*, showing a shift to distal APA in tumour samples. In plots B-D, the SLR was calculated similarly to that calculated by Akman *et al*., with the key difference being that, as opposed to custom probesets, manufacturer designed probesets (with expression determined by the fRMA method) were used here instead, resulting in the calling of clearer APA shifts [120].

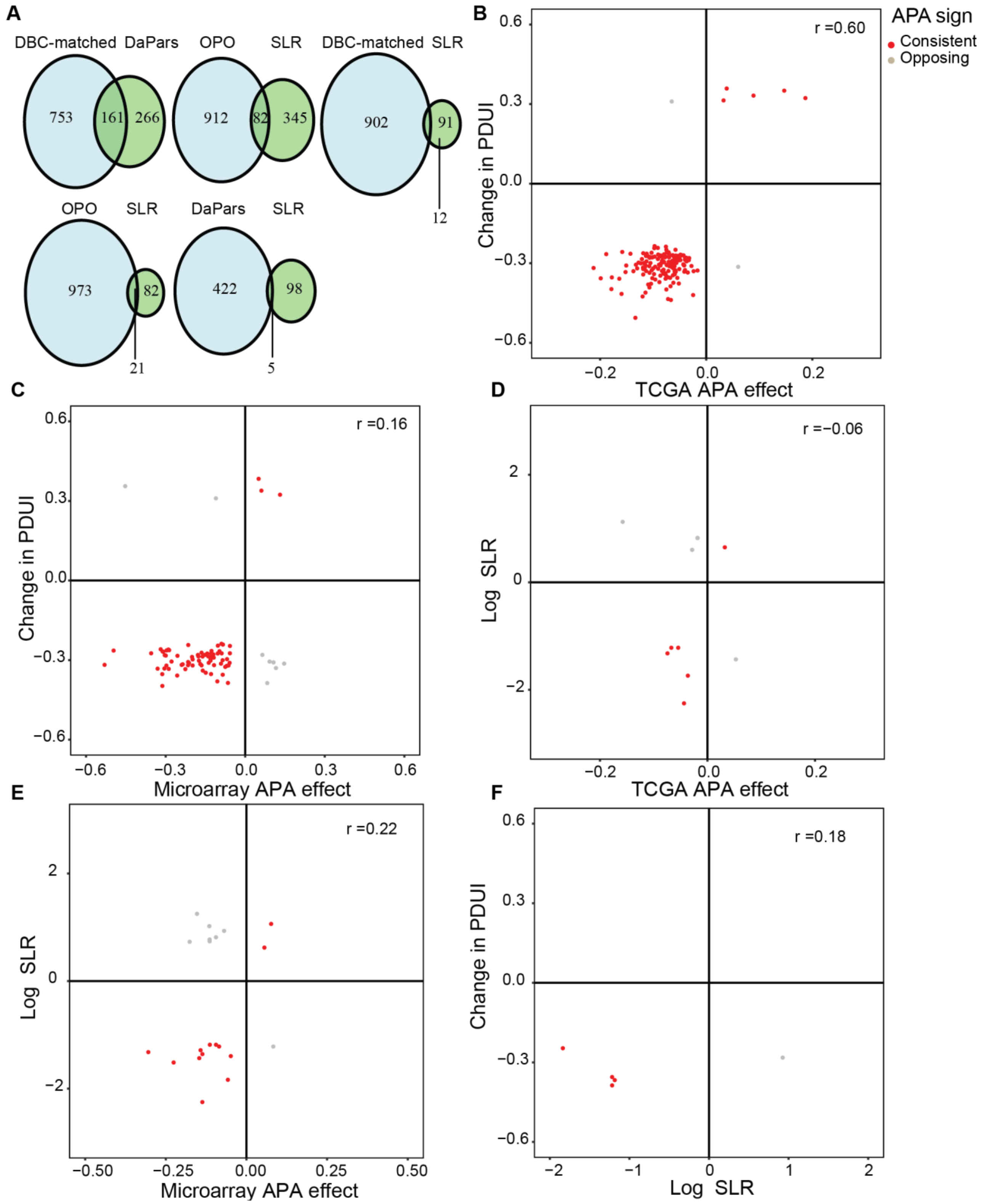


Figure 4.6. Comparisons of the DBC-matched method with APA from the OPO method and two papers that inferred APA from the same datasets. A. Venn diagrams of the overlap of significant APA events from each experiment analysed in this thesis. **B.** Is the comparison between the DBC-matched and DaPars datasets. **C.** Is the comparison between the OPO and DaPars datasets. **D.** Is the comparison between the DBC-matched and SLR datasets. **E.** The comparison between the OPO and SLR datasets. **F.** Is the comparison between the SLR and DaPars datasets. Where confect values are used, the APA effect is plotted for every gene that was assigned a confect value. All r values are Pearson's r. ΔPDUI values are from Xia *et al.* [66] and log SLR values are from Akman *et al.* [113].

study that was compared with the DBC-matched dataset was the DaPars method for calling APA in paired tumour-normal samples from the TCGA [66]. Despite being applied to the same Datasets, DaPars called less APA events than the DBC-matched method (427 vs 914) as shown in Figure 4.6 A (first Venn diagram), with an overlap of 161 APA events. The overlapping APA events were highly correlated, with all but 2 occurring in the same direction (Spearman's rho = 0.6, $p < 0.05$). Next, the unmatched microarray confects were compared with the $\log_{10}$ short/long ratio (SLR) values generated by Akman *et al.* using agglomerated GEO microarrays [113]. Once again, there were fewer significant SLR APA events when compared with the OPO (103 vs 994) as shown in Figure 4.6 A (fourth Venn diagram), with an overlap of 21 APA events. The correlation of APA events between these datasets was also low and not significant (Spearman's rho = 0.22, $p > 0.05$). This was despite the fact that the OPO dataset was actually derived from a slightly smaller subset of the Akman *et al.* dataset. This lack of correlation was surprising considering that both methods similarly relied on the relative expression of proximal and distal probes for a given gene, and only the events present in both datasets were compared. There were also no significant correlations observed when additional pairwise comparisons between the DBC-matched or OPO methods, were performed with the SLR or DaPars datasets respectively (Figure 4.6 C/D/F). This suggested that there is a high level of discordance between APA events that have previously been suggested to be associated with breast cancer and that the chosen method for inferring APA has a substantial effect on the results obtained.

4.2.6 Comparing TCGA and GEO datasets with the topconfects method yields 100 high confidence breast cancer associated APA events

To determine if APA was being consistently called in both the OPO and DBC-matched datasets, the *topconfects* APA values from both datasets were compared. The DBC-matched method was chosen for comparison with the OPO over the DBC unmatched method, as the

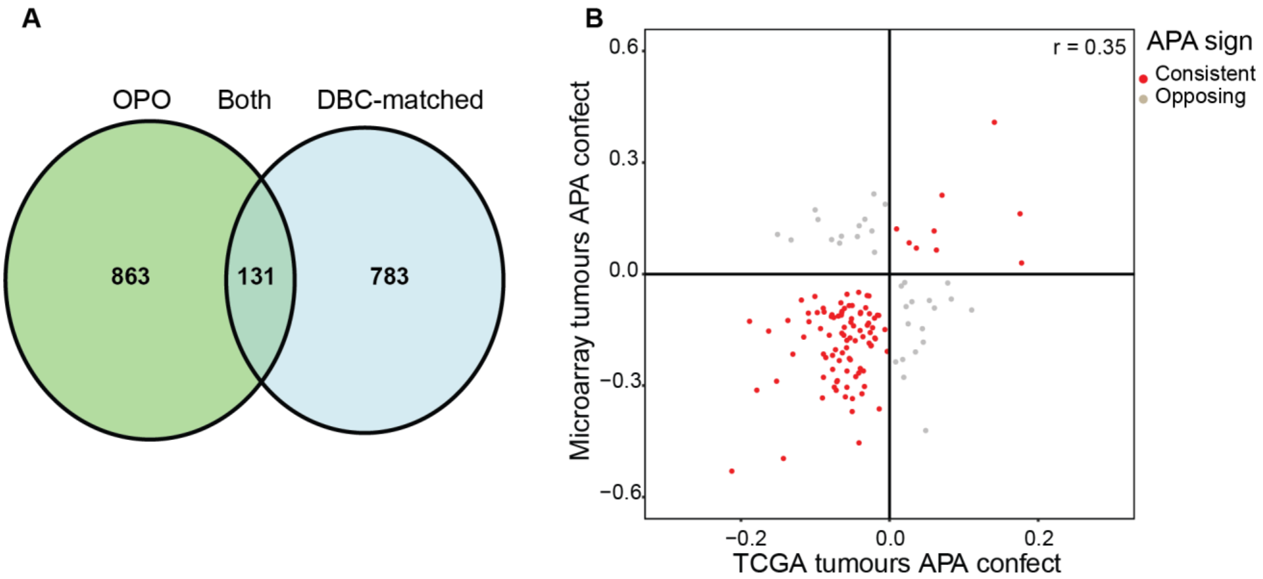


Figure 4.7. Comparison of the unmatched APA events from the GEO and matched APA events inferred from the TCGA. A. Venn diagram of overlapping APA events determined from unmatched tumour-normal comparisons in the OPO microarray dataset (left) and matched tumour-normal comparisons from the DBC-matched dataset (right). **B.** A scatter plot of the gene-wise APA events from the same comparison. Values in red indicate the 100 high confidence APA events that were assigned a confect value with the same sign.

matched tumour vs normal comparison was more powerful for calling APA than the unmatched version. Overall, there were 1,421 possible APA events (genes with > 1 associated probeset) present in the GEO dataset and 6,282 genes present in the TCGA dataset (with > 1 APA site and counts that passed expression cut-offs), with 1,020 common to both. Of these 1,020 potentially overlapping APA events, there were 131 that were assigned a confect value in both datasets (Figure 4.7 A). Within these 131 genes, 100 (76%) had a confect value with the same direction of APA change (Figure 4.7 B), with an overall Pearson's correlation coefficient of 0.35. These 100 'high confidence' APA events can be seen in Table A.1, Appendix A4. This concordance is relatively low, and may have partially been driven by variation in the power to call APA events in both methods across different platforms. Nonetheless, it was highly unlikely that the overlap of 100 genes occurred by chance (p << 0.01, Chi-squared test). Other than the overlap of the DaPars method with the DBC-matched dataset (Spearman's rho = 0.6), this was also the highest correlation that could be observed between any two APA calling methods examined in this thesis. Clear examples of a difference in RNA-Seq coverage of 4 randomly selected tumours and 4 randomly selected normal samples from the TCGA, for the top 6 high confidence APA events (ordered by TCGA confect), are shown as coverage plots in Figure 4.8. These plots clearly highlight striking patterns of APA and act as a valuable qualitative validation of the predicted events. The high confidence APA events discovered here, therefore, present compelling, novel APA events that are clear candidates for validation, using more targeted and precise methods, as well as functional studies.

4.2.7 TNBCs undergo more pronounced APA than ER+ breast cancers

It has been suggested previously that TNBCs undergo more APA events than their ER+ counterparts [168, 113]. To see if this theory holds, using the methods of APA prediction

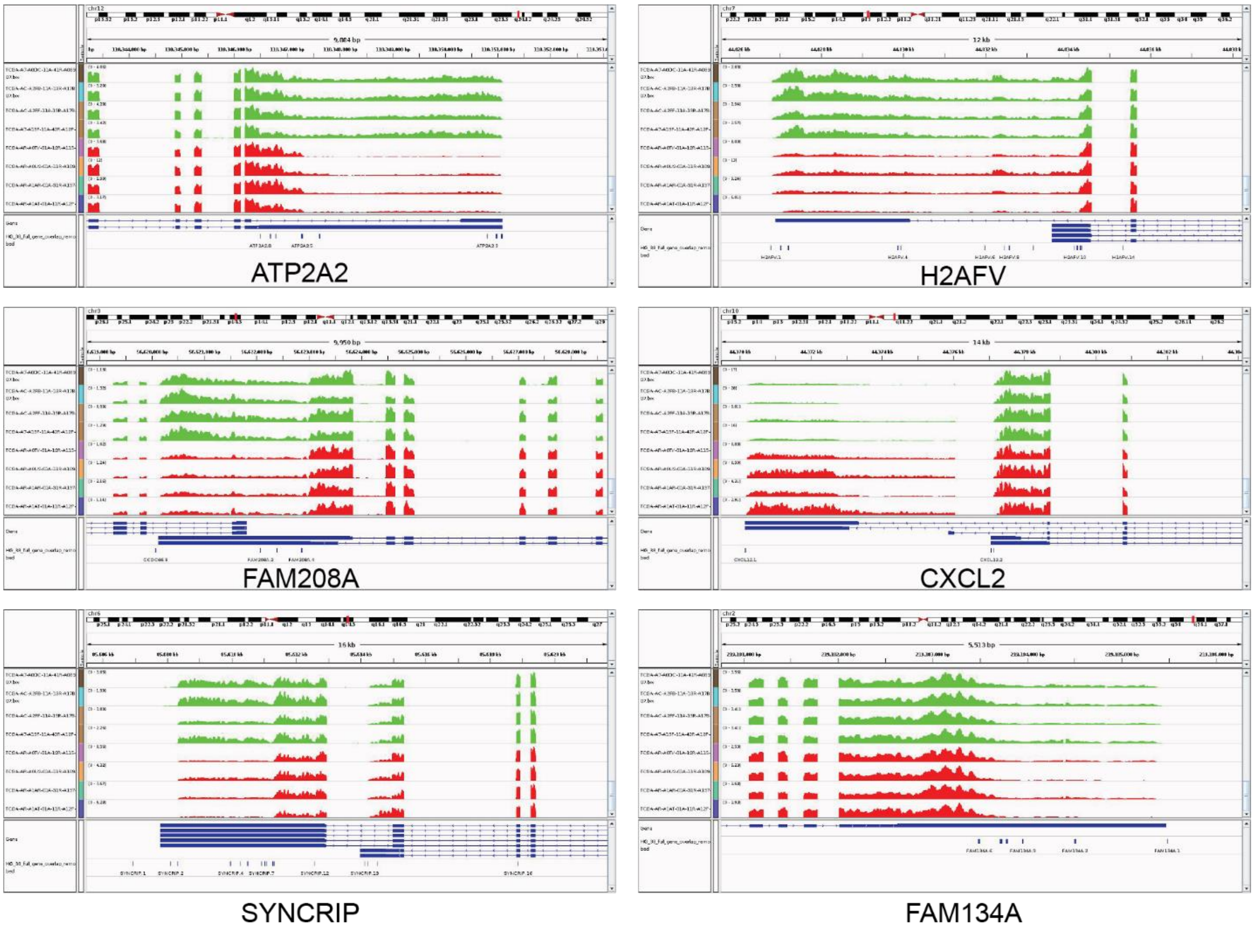


Figure 4.8. Example coverage of high confidence APA events. *IGV* screen shots of RNA-Seq coverage from the top 6 APA events present in both the TCGA and microarray datasets. Genes ranked by TCGA confect score. Tumour samples are shown in red and normal samples are shown in green.

outlined here, the APA r scores (named as such because they are between -1 and 1, Appendix A2) of ER+ cancers were compared with the r scores of TNBCs in both the TCGA (Figure 4.9) and GEO (Figure 4.10) datasets. APA r scores were used as a score can be assigned to every sample, rather than the condition-based effects of the topconfects method. APA effect was calculated using Paul Harrison's 3' end shifting method (Appendix A2), comparing tumour samples to the mean of all normal samples in both comparisons. Heatmaps of APA r scores can be seen in Figure 4.9 A for the TCGA dataset and Figure 4.10 A for the microarray dataset. Samples were grouped by subtype (either TNBC or ER+) to test the hypothesis that APA varies with tumour subtype. Genes were then ordered vertically by their mean APA r score from most proximal to most distal. In both plots there is a clear distinction between tumour and normal APA, however, beyond this striking shift, APA appears to vary more with sample than with tumour subtype. This effect can be seen in the red and blue streaks moving along the vertical axis of the heatmaps (all from the same sample), with no clear pattern along the horizontal axis except for the lower overall APA in normal samples.

Unmatched confect values (DBC-unmatched method) were also calculated for TNBC and ER+ tumours in both datasets and were highly correlated when plotting TNBC APA against ER+ breast cancer APA (Figure 4.9 B and Figure 4.10 B). These correlations were strikingly consistent in both comparisons (Pearson's r > 0.9) even despite the fact that variability in APA effects was not considered. The top 50 confect values from the ER+ vs normal and TNBC vs normal comparisons for both datasets can be seen in Figure 4.9 C/D and Figure 4.10 C/D respectively. In both the TCGA-unmatched and OPO datasets, 3' UTR shortening was the primary form of APA seen in both tumour subtypes. Interestingly, and despite the high correlation of APA effects, there tended to be more pronounced APA in TNBCs than ER+ tumours in both datasets. This means that APA effects tended to be larger when there

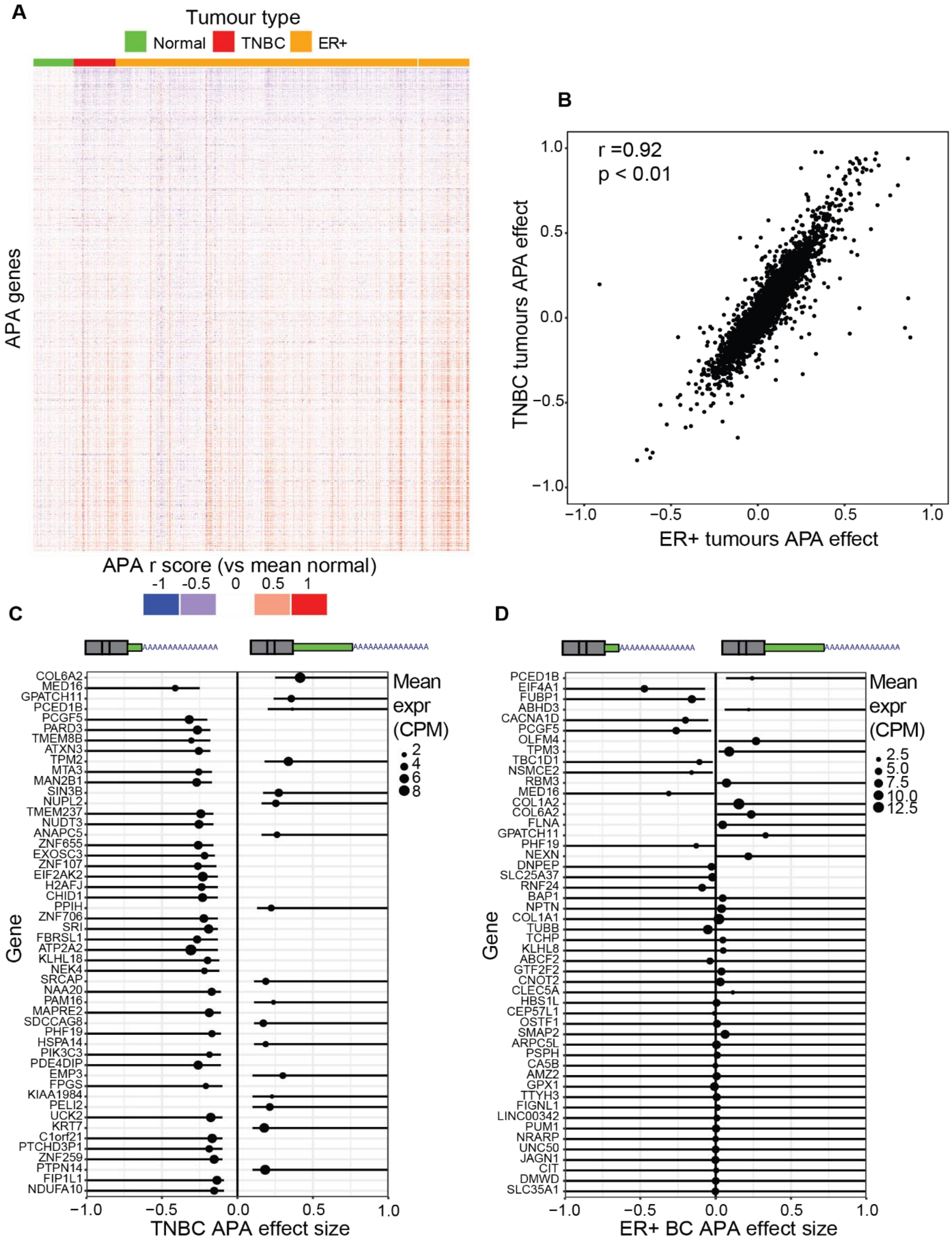


Figure 4.9. Alternative polyadenylation in breast cancer as determined by TCGA RNA-Seq reanalysis. A. A heatmap of APA effects as compared to the mean APA effect of all normal breast tissue samples. The heatmap is ordered vertically, by mean APA effect across all samples (proximal-distal). **B.** The correlation of the mean APA effect (effect size for each gene from *topconfects* without taking into account confidence) between TNBCs and ER+ cancers (Pearson's r = 0.92, p << 0.01). **C.** The top 50 confect values for TNBC tumours, determined using the DBC-unmatched method. **D.** The top 50 confect values for ER+ tumours, determined using the DBC-unmatched method.

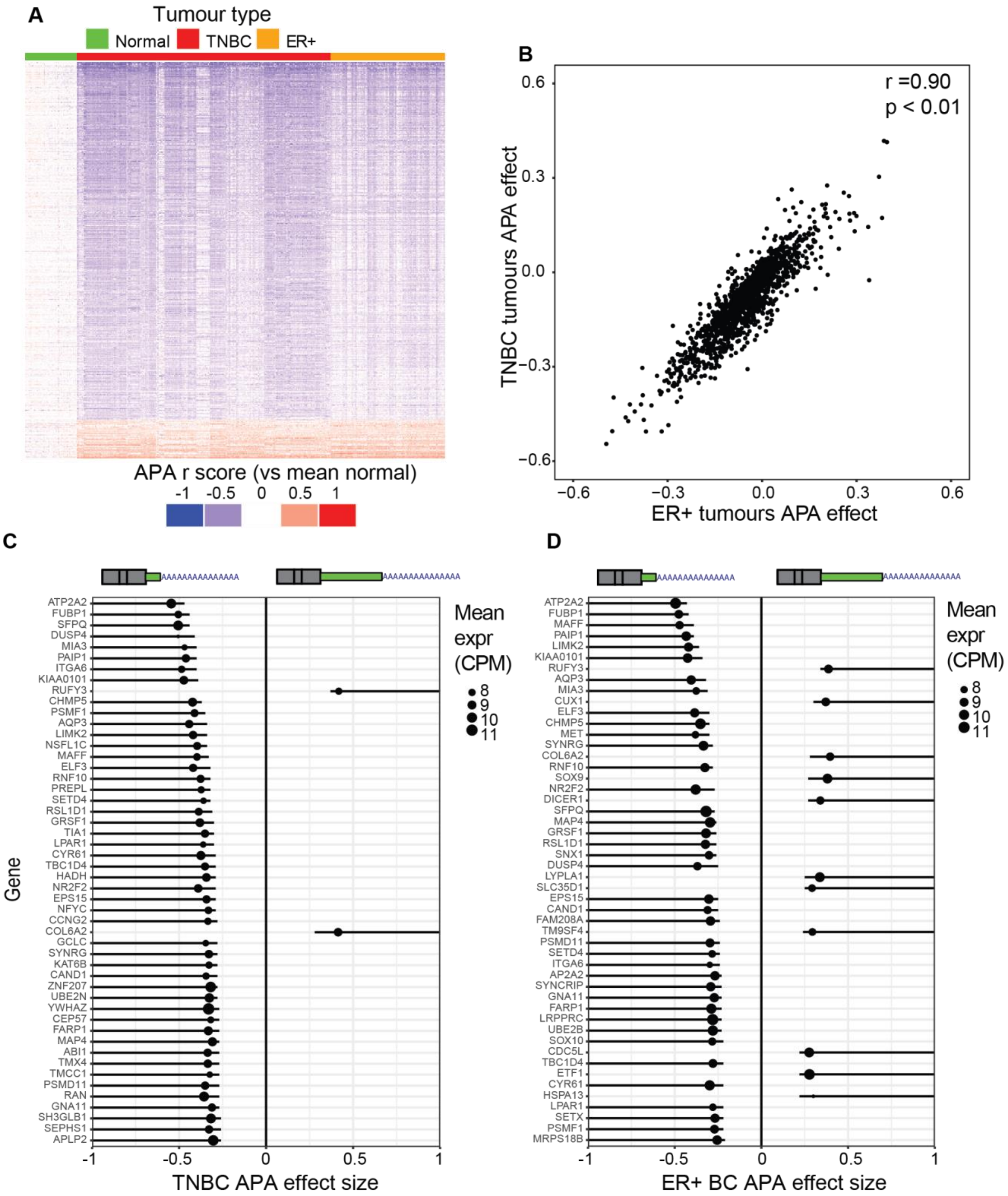


Figure 4.10. Alternative polyadenylation in breast cancer as determined by microarray reanalysis. A. A heatmap of APA effects as compared to the mean APA effect of all normal breast tissue samples. The heatmap is ordered vertically, by mean APA effect across all samples (proximal-distal). **B.** The correlation of the mean APA effect (effect size for each gene from topconfects without taking into account confidence) between TNBCs and ER+ breast cancers (Pearson's r = 0.90, p << 0.01). **C.** The top 50 confect values for TNBC tumours, determined using the OPO method. **D.** The top 50 confect values for ER+ tumours, determined using the OPO method.

was a switch to both proximal and distal sites. These results suggested that the general trend toward proximal APA does not vary with breast cancer subtype, however, the strength of APA changes present in a sample may be influenced by the state of the tumour.

4.2.8 An overall APA trend does not impact breast cancer outcome

From the heatmap of APA r values from the TCGA shown in Figure 4.9 A, it appeared that some samples undergo more pronounced APA than other samples and that this phenomenon may not be limited to tumour subtype. This APA trend is plotted in Figure 4.11 A, which shows that samples with stronger proximal APA, almost always also had stronger distal APA effects as well (Pearson’s r = -0.83, p << 0.01). To attempt to understand this phenomenon, each sample was assigned an ‘overall APA effect’ by taking the mean of the absolute APA r score (the absolute mean of the 3’ end shift r score, for every APA gene in a sample). Since this effect varied by sample, it was hypothesised that this effect may have some bearing on patient survival. Survival analysis was therefore performed, splitting the TCGA cohort on the median ’overall APA effect’. This score was not found to be significantly predictive of patient survival (Figure 4.11 E, p > 0.05 by Cox proportional hazards modelling). This suggested that greater APA overall was not contributing to negative breast cancer outcome, and may instead be associated with other cellular phenomena.

Table 4.2. Top TCGA APA genes. The top 20 (ordered by FDR) genes correlated with an overall APA effect present in both tumour and normal samples from the TCGA.

Gene	**Spearman's rho**	**p**	**FDR**
LOC100271831	0.74	1.80E-205	2.36E-201
GADD45GIP1	0.69	3.12E-165	2.04E-161
ZNHIT1	0.69	1.77E-163	7.73E-160
ATP5D	0.68	3.79E-156	1.24E-152
BCL7C	0.66	4.00E-148	1.05E-144
TGFBRAP1	-0.65	7.23E-142	1.58E-138
C16orf13	0.65	2.75E-141	5.15E-138
MRPL23	0.64	8.95E-135	1.46E-131
ZNF593	0.64	2.92E-133	4.24E-130
TSTD2	-0.63	3.13E-132	4.09E-129
TCEB2	0.63	8.50E-132	1.01E-128
YLPM1	-0.63	5.86E-131	6.39E-128
SF3A1	-0.63	8.51E-130	8.57E-127
NDUFA11	0.63	7.22E-128	6.75E-125
DDX3X	-0.62	5.83E-126	5.08E-123
EDF1	0.62	4.26E-125	3.48E-122
WDR82	-0.62	3.98E-124	3.06E-121
C9orf142	0.62	1.79E-122	1.30E-119
NDUFA13	0.61	6.06E-122	4.17E-119
EPC1	-0.61	8.59E-122	5.62E-119

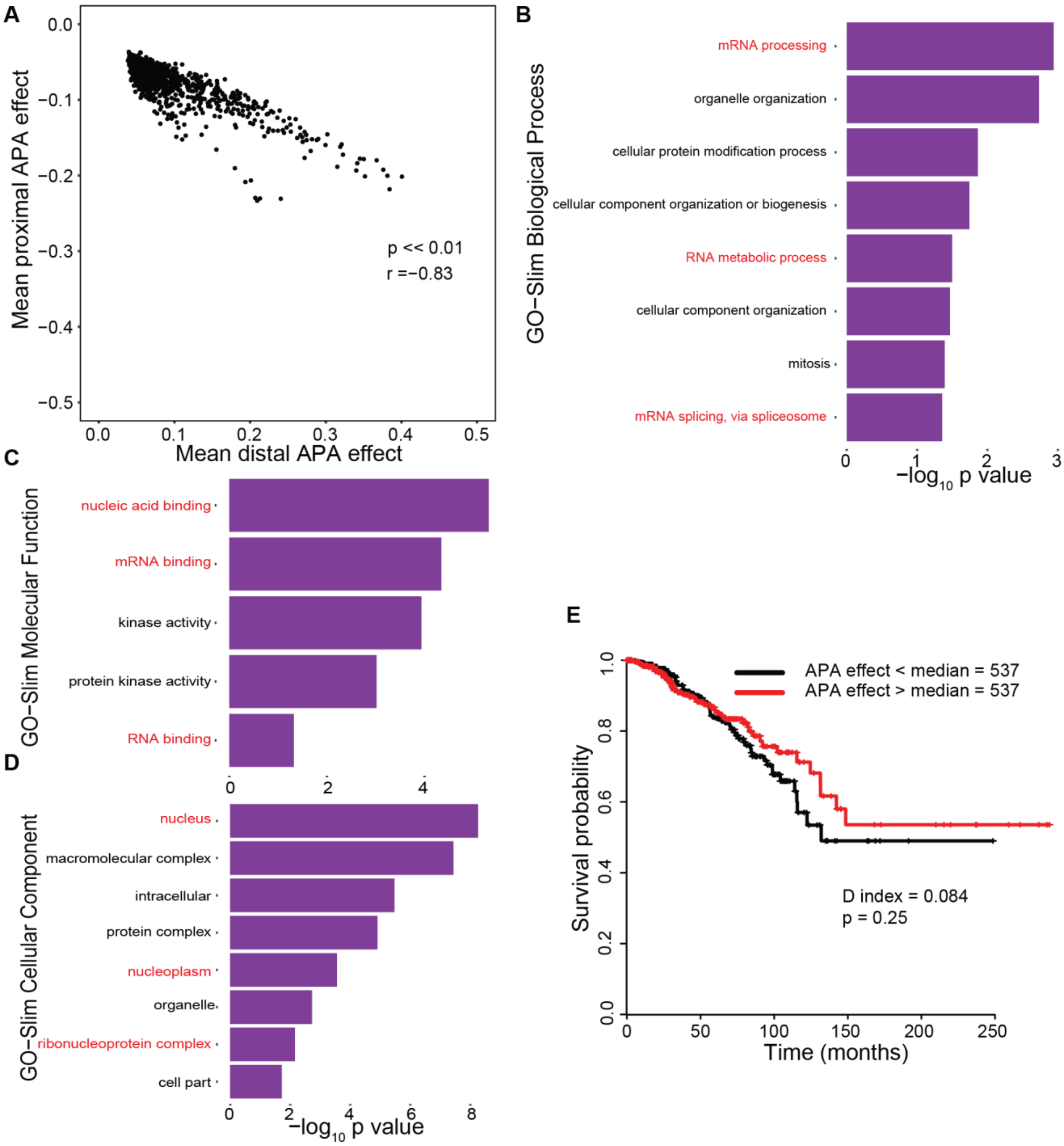


Figure 4.11. Analysis of an overall tumour APA effect. A. The correlation of the mean gene-wise APA effects for each sample. Mean distal APA is plotted on the x axis vs mean proximal APA on the y axis. **B-D.** Significantly Enriched GO-Slim biological process terms (B), molecular function terms (C) and cellular component terms (D). All GO p-values were Bonferroni corrected for multiple testing by PANTHER. **E.** Kaplan-Meier plot of patient survival separated by mean APA effect.

In order to understand the principal drivers of the APA effect, every gene expressed above background levels in a sample was correlated with that sample's APA effect. The p-values were FDR corrected, and the list of significant genes was given to the PANTHER [218] online tool for GO term overrepresentation analysis. The top 20 genes (by FDR) can be seen in Table 4.2. The genes most correlated with the APA effect were mostly related to nucleic acid binding and RNA processing (Figure 4.11 B-D). This suggested once again that a large portion of APA regulation may operate as part of a self-regulatory feedback loop, with an increased expression of CPA factors associated with both more extreme proximal and distal APA.

4.2.9 Tumours from the TCGA do not always have a distinct pattern of APA when compared to normal breast tissue

It is well known that APA state changes with cell type [111], however, it is not known if these alterations are consistent with changes to gene expression in breast cancer, or if they occur through alternative mechanisms. To test the grouping of samples generated by APA compared with gene expression, the similarity of all tumour and normal cells was analysed using t-Distributed Stochastic Neighbour Embedding (t-SNE). The t-SNE method is a tool for dimensionality reduction and visualisation of high dimensional datasets in two-three dimensions. In this thesis t-SNE was used to group samples based on their similarly in high dimensional space (such as the APA effect for each gene in a sample). Exactly as for the heatmaps in Figure 4.9 A, the matrix of APA r scores for input into the t-SNE was calculated by comparing each tumour or normal sample to the mean APA state of all normal samples (DBC-unmatched method). A t-SNE was also performed using gene expression for all samples in the TCGA, which was expressed as transcripts per million (TPM). The t-SNE plots showed a distinct grouping of normal samples when using both gene expression (Figure 4.12 A) and APA (Figure 4.12 B) as inputs. More distinct t-SNE groupings were, however, formed

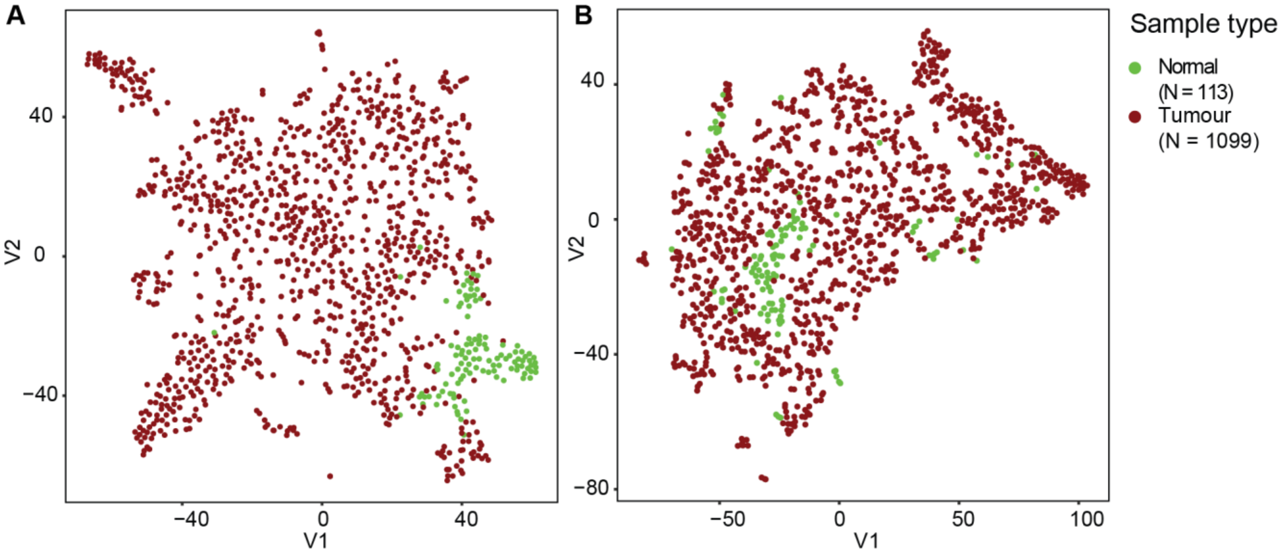


Figure 4.12. T-SNE distributions of both tumour and normal samples generated using both gene expression and APA. A. A t-SNE plot of TCGA breast tumour and normal samples, separated based on gene expression (in TPM). Normal samples tend to be shown together. **B.** The same as A, except that samples were separated based on APA r values, calculated against the mean normal APA state (DBC-unmatched method).

based on gene expression, suggesting that APA may not change with subtype in the same way that gene expression does. Interestingly, many more tumour samples exhibited patterns of APA more similar to normal samples than the other tumour samples in the analysis than was the case for gene expression. This effect can also be seen in the heatmap of the same APA r scores shown in Figure 4.9 A, with some tumour samples having less pronounced APA r scores in general than some normal samples. This result suggested that, in some cases, it is possible for tumours to have a similar pattern of APA to normal samples.

4.2.10 APA events are not subtype-specific

As breast cancer can be broadly classified into histological and gene expression subtypes [9, 23], it was hypothesised that APA may form distinctive patterns that also segregate with these subtypes. However, while ‘overall APA effects’ were clear in Figure 4.9, there also seemed to be few distinguishing differences between the TNBC subtype and ER+ breast cancers. To determine if APA patterns broadly followed defined gene expression subtypes, another t-SNE comparison was performed in exactly the same manner as was used to compare tumour and normal samples in Section 4.2.9, except that normal samples were excluded from the analysis. The PAM50 subtypes for a number of tumour samples in the TCGA have been determined previously [1], and these were overlaid onto the results as different colours. As expected, basal-like tumours were grouped together by gene expression (Figure 4.13 A), however, it was more difficult to distinguish the other subtypes. In contrast, there was no clear pattern in the plot generated from the APA r scores (Figure 4.13 B). This result suggested, once again, that the APA state of a tumour may vary independently to gene expression state.

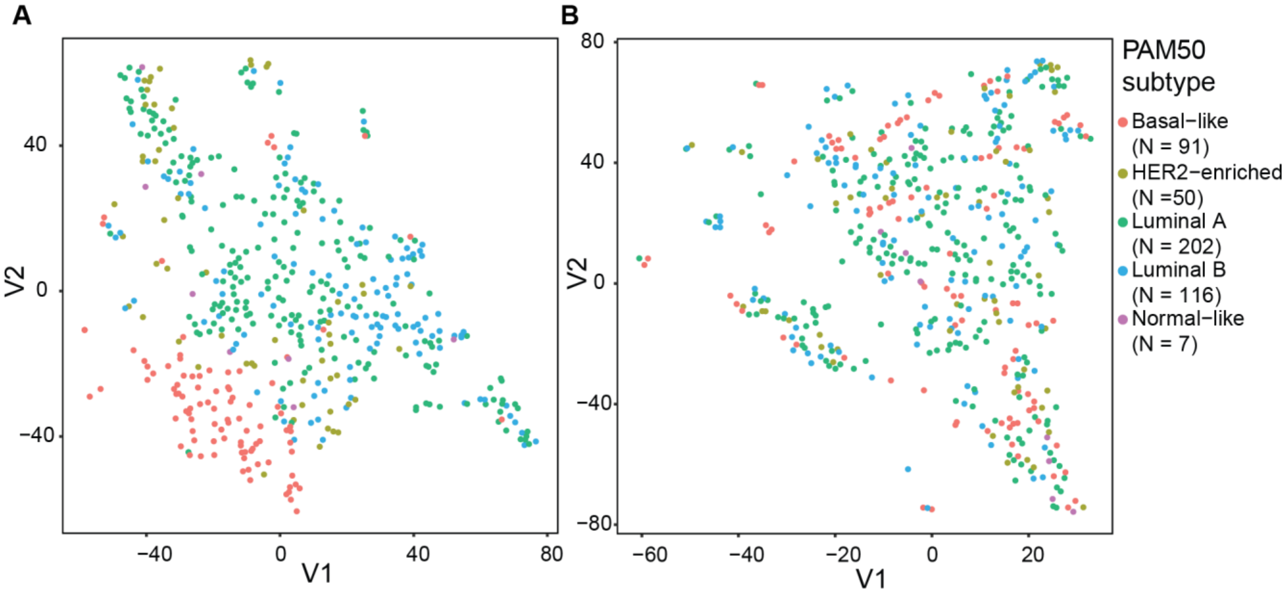


Figure 4.13. T-SNE distributions of tumour samples generated using both gene expression and APA. A. A t-SNE plot of TCGA breast tumour samples, separated based on gene expression (in TPM) and coloured by their PAM50 subtype. Some discrimination between subtypes can be seen. **B.** The same as A, except that samples were separated based on APA r values calculated against the mean normal APA state.

4.2.11 Assessing miRNA expression with metastasis and subtype in the TCGA

miRNAs are the elements best known to interact with the 3' UTRs of mRNAs [112]. To determine the degree to which miRNAs were dysregulated in the TCGA, miRNA expression data was downloaded from GDC data portal. The expression of the 368 miRNAs measured by the TCGA in both tumour and normal samples can be seen in Figure 4.14 A. There did not appear to be any trend associated with tumour stage, however, as with gene expression and APA (Figure 4.12 A/B), there was a clear difference between tumour and normal samples. This trend is further highlighted in Figure 4.14 B where 77% of miRNAs were significantly changed (FDR < 0.05) between tumour and normal samples. There were only small differences in miRNA expression between TNBC and non-TNBC tumours (Figure 4.14 C) with only 10 miRNAs not having a similar direction of gene expression change (vs the mean of the normal breast tissue samples). This suggested that the majority of miRNA changes are less associated with gene expression subtype than has been previously suggested [270]. This is also consistent with the APA data, which did not show any subtype specific patterning, leaving open the possibility that these two processes may be linked.

4.2.12 Predicting tumour outcome with clinical data gene expression and APA

To study the effect of clinical data, as well as gene expression and APA on the survival outcome of breast cancer, ENLM was employed using the *glmnet* R package [221]. The *glmnet* package allows the elastic-net method to be used to fit both generalised linear models and Cox survival models. In this chapter, *glmnet* was used to fit regularised Cox survival models. Clinical variables such as tumour stage, weight and receptor status were also studied to determine if gene expression or APA added additional prognostic power, in addition to the standard information that is currently obtained about a primary breast tumour. In the context of breast cancer prognosis, ENLM allowed for the selection of the coefficients with the greatest effect on survival, while minimising the overrepresentation of correlated coefficients.

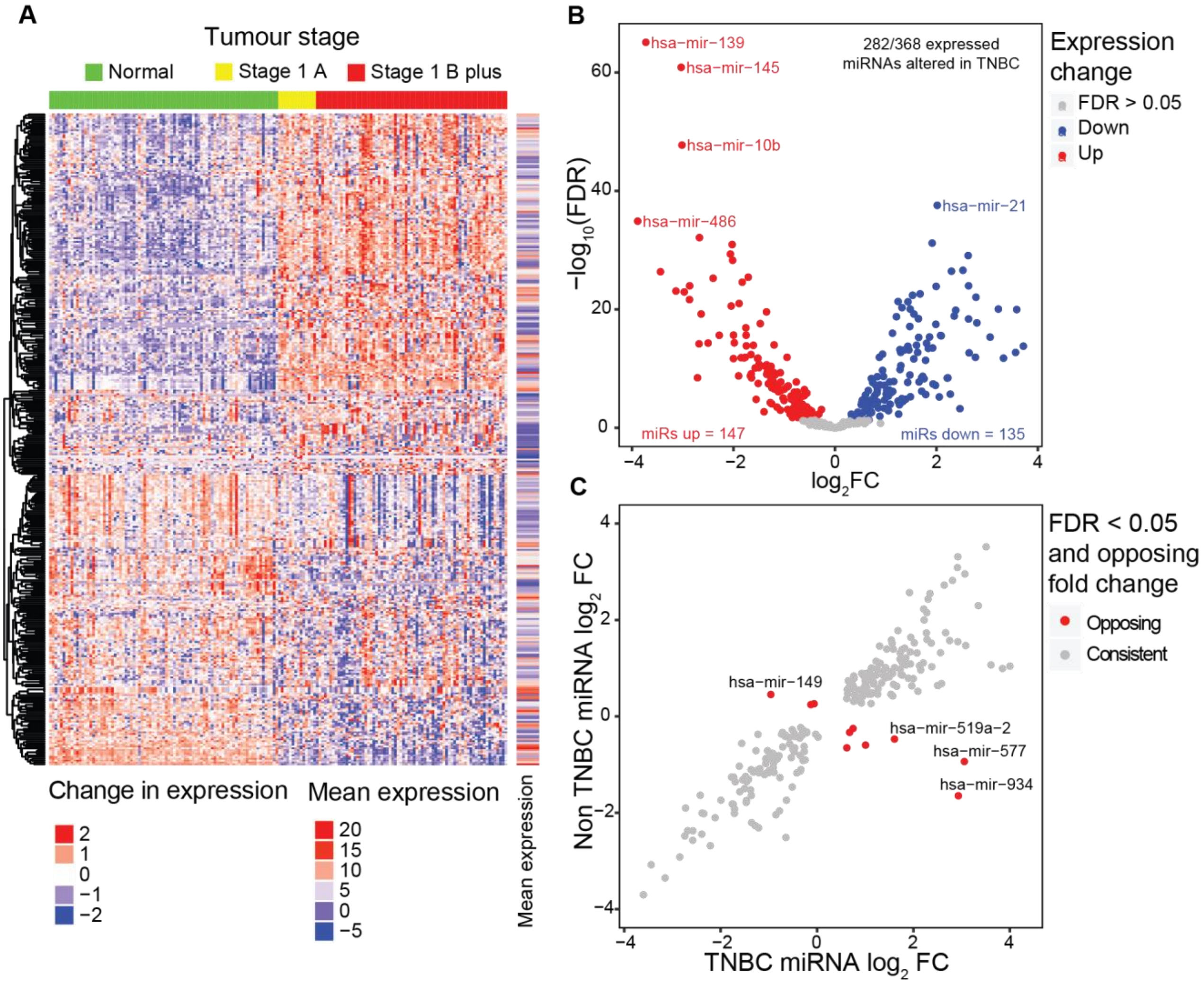
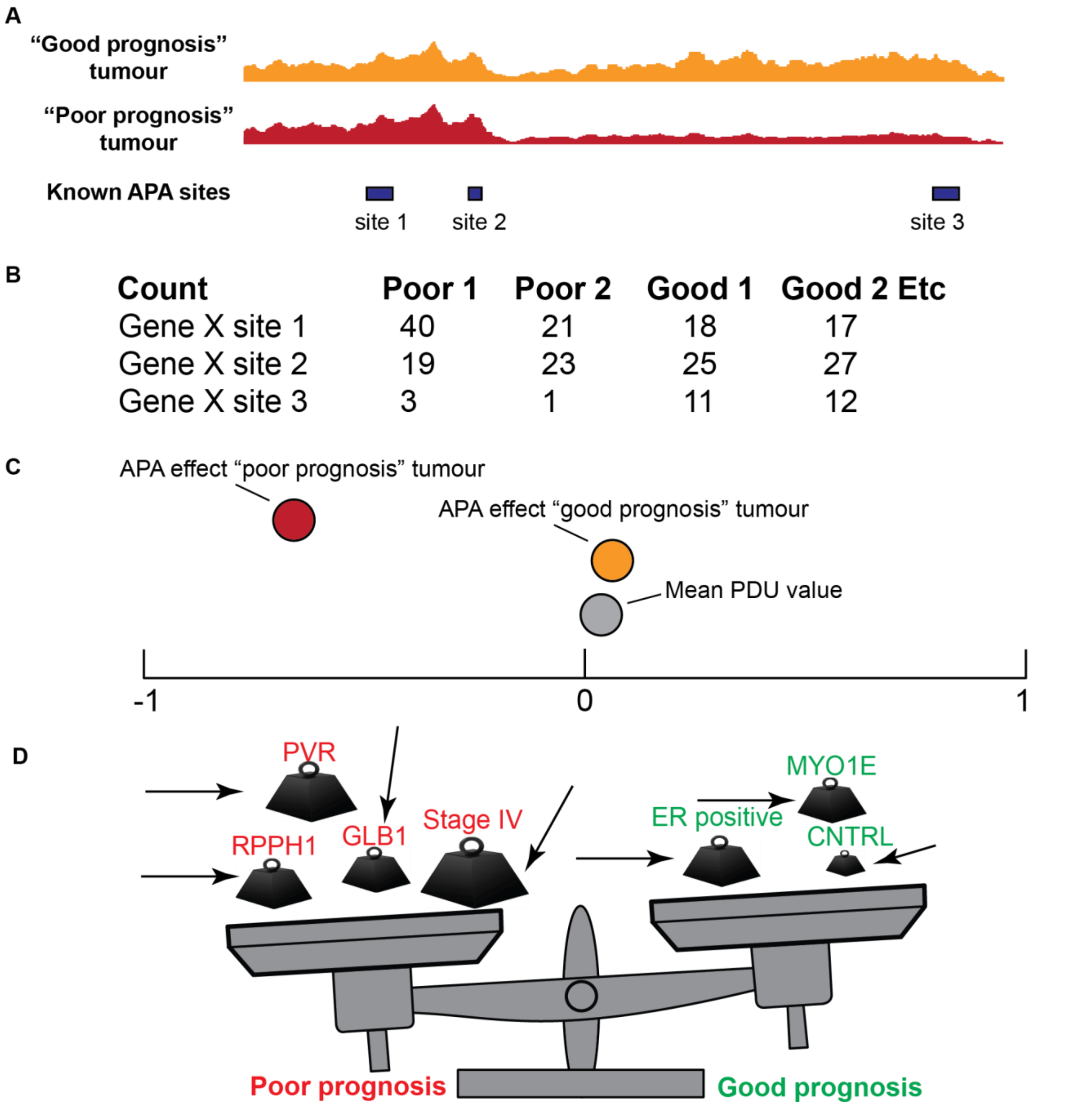


Figure 4.14. Expression of miRNAs in the TCGA A. Heatmap of miRNA expression (in $\log_2$ CPM) of TNBCs and normal breast tissue in the TCGA. **B.** Volcano plot of the miRNA expression changes and the FDRs of the TCGA TNBCs, as compared to normal breast tissue. **C.** Scatter plot of fold changes of TNBCs compared with other breast cancers in the TCGA. Highlighted miRNAs have a different direction of $\log_2$ fold change.

Glmnet allows for penalty values to be altered prior to the start of the algorithm, allowing predictors that are already known to be effective, such as common clinical variables, to be preferentially utilised. Reproducibility of the models generated in this study was measured by 10-fold cross-validation (10-fold CV). Datasets were split into a 90% training and 10% validation cohort, with a different 10% of samples used for validation in each round of CV. Once each sample had been assigned a prognostic score (model link score for each sample) by ENLM as part of a 10% validation cohort, the prognostic scores were then used in combination with survival data to calculate the D index and D index HR.

The regularisation in *glmnet* is controlled by the two parameters, namely the lambda (λ) and alpha (α) parameters. The λ value represents the penalty value applied to every predictor for inclusion in the model. The optimal λ values were determined by glmnet using the ‘cv.glmnet’ function (which uses 10-fold cv to suggest possible λ values) on the full dataset including all samples but not always all predictors depending on the combination of predictors being tested. The same lambda value was then used for each round of CV. The α score represents the balance between ridge regression and the lasso (α = 1 is lasso and, α = 0 is ridge regression). A higher α value will therefore generally result in fewer predictors being used in the model and vice versa. After testing multiple models using APA, gene expression and clinical variables, an α score of 0.37 was settled upon to control for optimal predictive capability and a manageable number of coefficients.

A diagram of ENLM for a model comprised of APA and clinical data can be seen in Figure 4.15. Once a model was generated, it was assessed by D index [259], p-value and D index HR (as explained in the introduction to this chapter). D indexes were compared between



Count	Poor 1	Poor 2	Good 1	Good 2 Etc
Gene X site 1	40	21	18	17
Gene X site 2	19	23	25	27
Gene X site 3	3	1	11	12

Figure 4.15. Inferring good or poor prognosis using APA and clinical variables. A. Read counting was performed at known APA sites. **B.** Sites were organised by gene and an APA effect value was determined. **C.** The differences from mean APA state were calculated for every gene in every tumour (DBC-unmatched method). **D.** ENLM was used to select and weight the changes in APA state (in combination with clinical variables) that best predict survival. Arrows refer to the selection of a predictor in the model.

competing models to determine if one was significantly more predictive than another. Clinical variables are most commonly used to predict likely breast cancer outcome and determine the optimal course of treatment [165]. Breast cancer outcome was defined in this thesis as overall survival (OS) over time. Before moving on to gene expression and APA, an elastic net model was generated to assess the utility of clinical variables alone in the prediction of survival from tumour samples in the DBC-unmatched dataset, giving a baseline level of prognostic power. Only primary tumours for which gene expression, APA and clinical data could be obtained were therefore used in this study. Distant metastases and duplicate sequencing runs of the same primary tumour sample were also not included, yielding 1,033 samples in total.

Initial analysis showed that tumour stage III did not contribute to the model as expected, likely due to the low number of stage III samples (< 20) present in the TCGA, and so it was merged with stage II in all models utilising clinical data. Results were analysed in all models as part of 10-fold CV analysis, with survival predictions generated by splitting samples by median link score (prognostic score), unless otherwise stated. As expected, clinical variables alone were able to significantly predict patient outcome (Figure 4.16 A, $p < 0.05$, D index HR = 3.01). This showed that ENLM performed as expected when predicting patient survival. The 15 clinical variables included in the model, and the standardised contribution of each to the prediction of tumour outcome can be seen in Table 4.3. Standardisation was performed by multiplying each coefficient by the standard deviation of the values that made up the predictor. This made the standard deviation 1 in all predictors and enabled the comparison of the absolute contribution of a predictor to the model. Many clinical variables were chosen for inclusion in the model, as the consideration of many clinical variables may circumvent the recapitulation of these signals from gene expression or APA based predictors. For comparison with numeric variables (such as tumour weight), binary variables (such as stage or race) were encoded numerically as having a value of 0 or 1. In order to be able to merge all scores obtained by 10-fold CV into one dataset, all predictors were normalised to have a mean of zero (subtracting the mean of a predictor from each predictor value) prior to model generation in this model and all future models, including those containing gene expression and APA predictors. As expected, clinical data alone established a strong baseline of prognostic power for comparison with more complicated models including APA and gene expression.

Table 4.3. The contribution of clinical predictors to ENLM. Data presented is mean clinical coefficient contribution following 10-fold CV. Also provided is a description of each clinical variable and the units in which it is measured. Coefficients were standardised to enable the comparison of the absolute contribution of a predictor to the model (standard deviation = 1 in all predictors).

Clinical variable	Description	Units	Mean coefficient
Age	Age of the patient	Years	0.44
Stage IV	Tumour stage, tumour has metastasised	Binary	0.23
Lymph node met	The tumour has metastasised to the lymph nodes	Binary	0.20
HER2 pos	The tumour overexpresses HER2 by immunohistochemical analysis	Binary	0.19
Mono int pct	The percentage of the tumour that has been infiltrated by monocytes	Percentage	0.09
Stage II	Tumour stage, tumour is usually larger or more invasive than Stage I	Binary	0.07
Neutro int pct	The percentage of the tumour that has been infiltrated by neutrophils	Percentage	0.04
Tumour weight	The weight of the extracted tumour	Grams	0.00
Race Black	The race of the patient	Binary	0.00
Race Asian	The race of the patient	Binary	-0.01
Tumour necrosis pct	The percentage of the tumour that is necrotic in appearance	Percentage	-0.01
Lmpho int pct	The percentage of the tumour that has been infiltrated by Lymphocytes	Percentage	-0.08
PR pos	The tumour overexpresses PR by immunohistochemical analysis	Binary	-0.10
ER pos	The tumour overexpresses ER by immunohistochemical analysis	Binary	-0.28

Gene expression has long been known to be predictive of tumour outcome [56]. The predictive power of gene expression was therefore evaluated in the TCGA dataset. Gene expression (in TPM) was calculated in the same 1,033 samples from the TCGA that were analysed using clinical data. ENLM, assessed once again by 10-fold CV, was then used to determine the subset of genes most predictive of tumour outcome, with the same parameters as the clinical only model. As expected, gene expression was a strong predictor of survival (D index HR = 2.35, $p << 0.01$, Figure 4.16 B). Gene expression alone, however, had less prognostic power than clinical variables alone, with a D index HR of 3.01 for the clinical model vs 2.35 in the gene expression model. This result served as evidence that ENLM was also able to build useful predictive models in the gene expression context using this dataset, but that this information alone was not more useful than clinical predictors (Figure 4.16 A). Interestingly, even a combination of gene expression and clinical predictors was not significantly better than clinical data alone ($p > 0.05$, D index comparison, Figure 4.16 D).

Following previous suggestions that APA was associated with tumour outcome [66], the effect of the deviation of APA from the mean of all tumours on patient survival was investigated. An APA r score was generated for every gene that had > 1 poly(A) site (DBC-unmatched). This method allowed multiple APA sites to be utilised, and matched normal samples were not required, allowing all tumours in the TCGA to be analysed. All APA sites were considered as candidates, regardless of whether they were previously found to be significantly associated with breast cancer (Section 4.2.6). ENLM was performed using the APA r scores in the same way as was previously used for gene expression. APA was significantly able to predict survival ($p < 0.01$, D index p-value, D index HR = 2.21, Figure 4.16 C), but once again was not as prognostic as clinical variables alone (D index HR = 3.01) or gene expression alone (D index HR = 2.35).

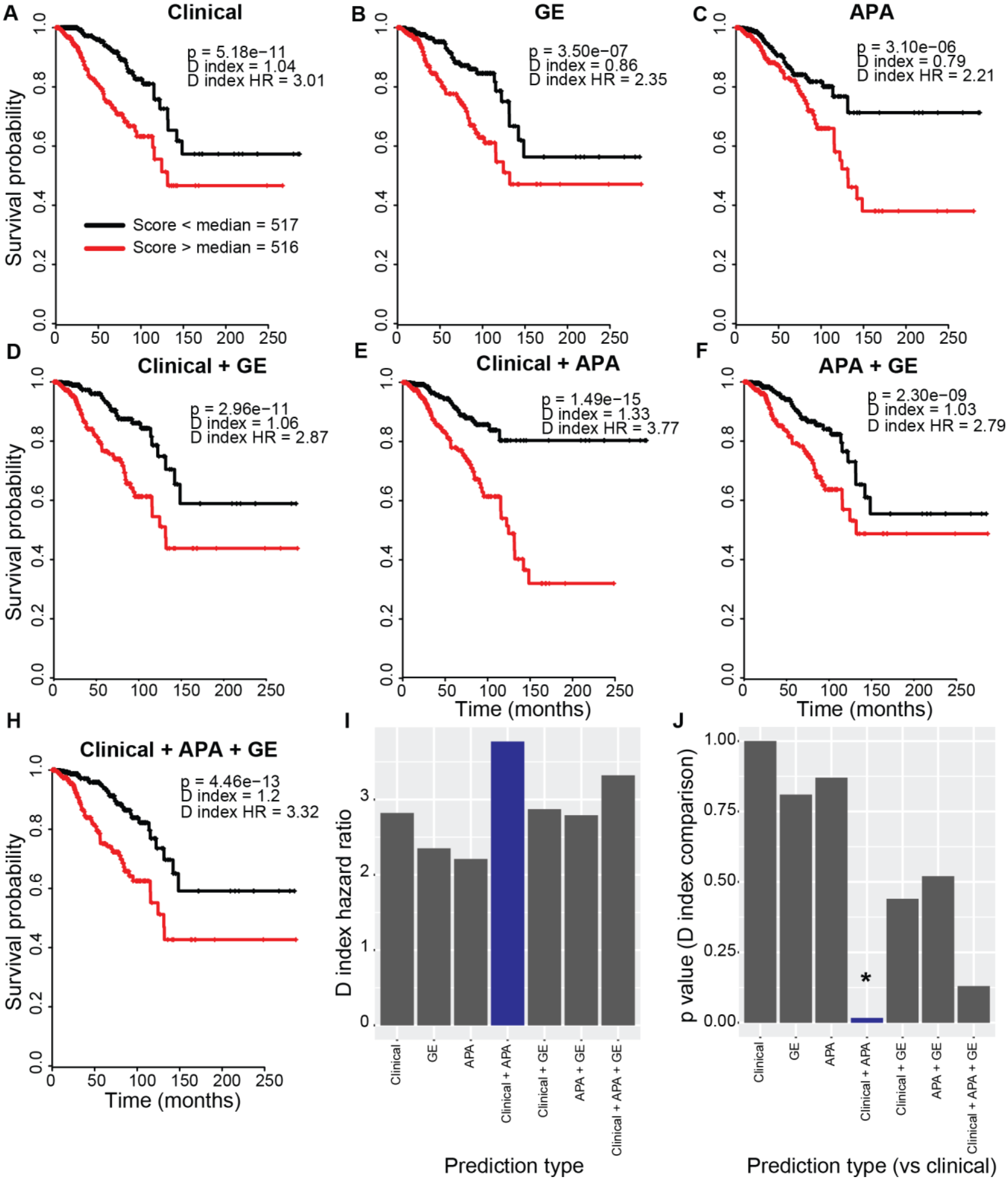


Figure 4.16. Elastic net linear modelling prognostic scores generated using gene expression, APA and clinical data. A-H. Relapse-free survival of 1,033 patients when tumour outcome was predicted using combinations of APA, gene expression and clinical data as input to ENLM. Good and poor prognosis were split on median link score (prognosis score). Scores were calculated from 10-fold CV. Clinical + APA had the greatest prognostic power of any model (D index HR = 3.77, p << 0.01). **I.** D indexes for each prediction type combination. **J.** The p-values for D index comparisons (each model type vs clinical alone).

4.2.13 Clinical data and APA make a significantly better prognostic model than clinical data alone

All combinations of predictor sets (gene expression, APA and clinical data) were tested by ENLM and 10-fold CV in order to determine the combination of predictors that results in the most powerful prognostic model. To be certain that APA and gene expression were adding prognostic power beyond clinical variables, clinical predictors were penalised half as much (penalty factor = 0.5), when combined with APA and gene expression predictor sets, making clinical variables more likely to be included in the model. The clinical + gene expression model did not have significantly greater predictive power than clinical data alone ($p > 0.05$, D index comparison, Figure 4.16 A and D). All D index comparisons were performed using the *survcomp* R package. This suggested that much of the prognostic capacity of gene expression may already be reflected in common clinical variables that already measure the proliferative state of a primary tumour. Interestingly, clinical + APA (Figure 4.16 A) was the most prognostic combination of predictor sets of any combination, including the full clinical + APA + gene expression set. Clinical + APA was also the only predictor set to have a significantly better D index than clinical data alone ($p = 0.02$, D index comparison, Figure 4.16 J). D index values for each comparison are summarised in Figure 4.16 I. It was not expected that the Clinical + APA would have the most prognostic power, as it would be expected that more predictor sets would each bring additional prognostic power. This suggested that gene expression and clinical data may be comprised of a similar poor prognosis signal, while APA may add additional prognostic information.

4.2.14 Assessing the reproducibility of predictor selection using clinical data, APA and gene expression

In order to infer how generalisable the models generated here may be when applied to new breast cancer datasets, the reproducibility of the selection of predictors was tested by bootstrapping. For the simulation of predictor selection from alternative populations, the

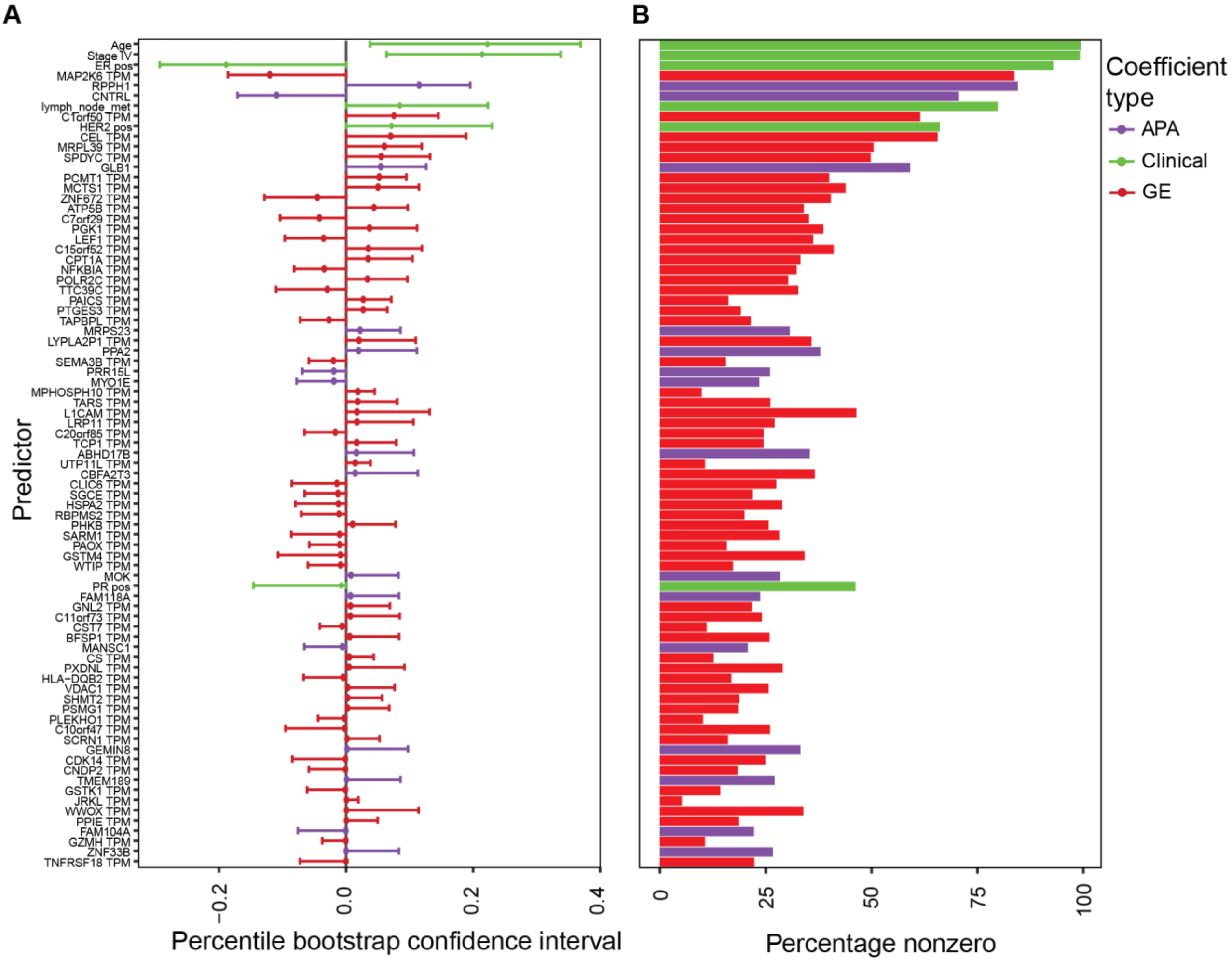


Figure 4.17. Bootstrapping of clinical, APA and gene expression elastic net linear modelling. A. A 95% percentile bootstrap confidence interval of each coefficient in the clinical + APA + gene expression model, trained on the full set of 1033 samples. The dot is the coefficient value from the original model and the 95 % confidence interval is generated from coefficient values taken from 1,000x bootstrapping. Coefficients were standardised to enable the comparison of the absolute contribution of a predictor to the model (standard deviation = 1 in all predictors). **B.** The percentage of the time that each predictor variable was selected by ENLM in the 1,000x bootstrapping. Clinical predictors were penalised half as much as other gene expression or APA predictors and were, therefore, more likely to be included in the model.

complete predictive model (all 1,033 samples) was constructed using the full clinical + APA + gene expression predictor sets, and 1,000x bootstrapping (random sampling of the data with replacement) was performed. The coefficient values of predictors from the original full model and 95% bootstrap confidence intervals (CIs) are shown in Figure 4.17 A. While APA predictors were selected at a lower frequency to gene expression predictors (Figure 4.17 B), the APA genes *RPPH1*, *CNTRL* and *GLB1* featured near the top of the list of coefficients. This again suggested that APA events may have additional impact on survival prediction that is not measured by gene expression in combination with clinical data and that this effect is reproducible by bootstrapping.

4.2.15 Assessing the power of the clinical and APA model in different breast cancer subtypes

It has been suggested that breast cancer outcome varies with subtype [45]. As it had the greatest prognostic power, the APA + clinical model was further evaluated across both histological and gene expression subtypes. Models were first re-generated using 20-fold CV to obtain slightly more accurate scores, increasing the discriminatory power of further analyses that may rely on less samples. The APA prognostic signal was still useful, even when samples were separated by PAM50 ‘intrinsic subtypes’ [20] (Figure 4.18) that were previously annotated by the Cancer Genome Atlas Network [1]. Similarly, the APA signal was able to add prognostic power when samples were separated based on histological receptor status (Figure 4.19), even adding significant prognostic power in the heterogeneous TNBC subtype (Figure 4.19 I, D index HR = 2.2, p = 0.018). The increased prognostic power of APA, therefore, did not appear to discriminate by breast cancer subtype, reflecting the same lack of separation seen in tumour vs normal comparisons. If successfully validated, this model could be potentially used to predict outcome in breast cancer subtypes, such as the TNBC subtype, where no clinically approved prognostic test yet exists.

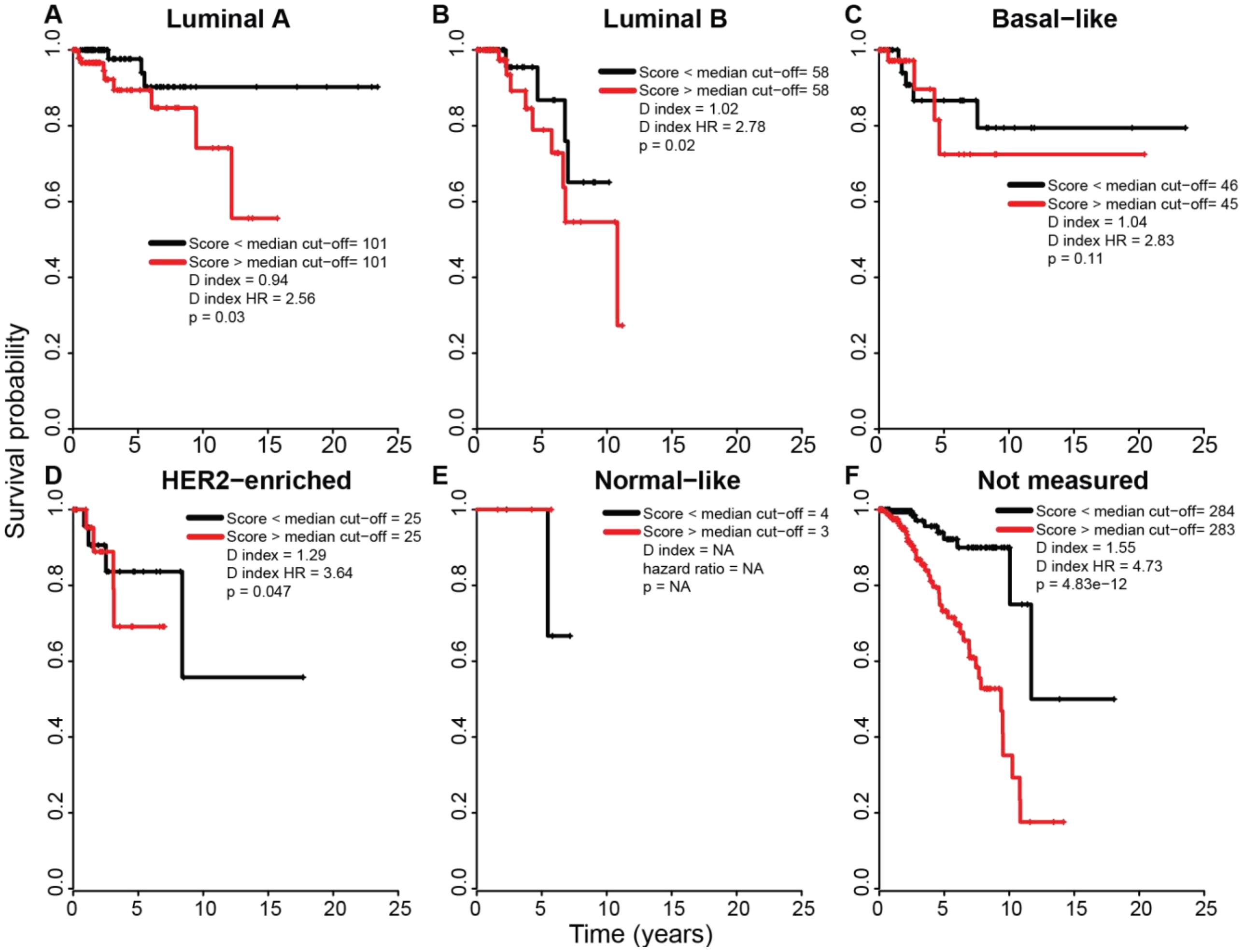


Figure 4.18. APA + clinical prognostic score within 5 intrinsic breast cancer subtypes. Survival split on the median prognostic score, generated using clinical and APA predictors, for each of the 5 intrinsic breast cancer subtypes [93] (**A-E**) and unclassified tumours (**F**).

4.2.16 Refining the clinical and APA prognostic model for potential use in an mPAT-based prognostic test

If the best predictive models presented thus far were to be implemented as a multiplexed 3’ RACE-based clinical test, then using a large number of APA sites would likely make the design of such a test more difficult. The α score of the most predictive APA model was increased to 1 (complete lasso), in order to generate a model with the least included predictors, that was still manageable from the standpoint of developing a clinical test. Despite the fixed α score, this prognostic score (Figure 4.20 B, D index HR = 3.81) was comparable to that of the best APA + clinical model from the previous set of comparisons.

The final proposed mPAT model consisted of 45 predictors (Table 4.4), comprised of 6 clinical variables and 39 APA events. A penalty factor of 0.5 was once again given to clinical variables, selecting for their preferential inclusion in the model (Figure 4.20 A). The APA + clinical model represented a set of APA events that, once validated, could potentially be used as a diagnostic test for primary tumour biopsies from any breast cancer subtype. In future, primers may be designed to test this model using the primer design tool described in Chapter 5.

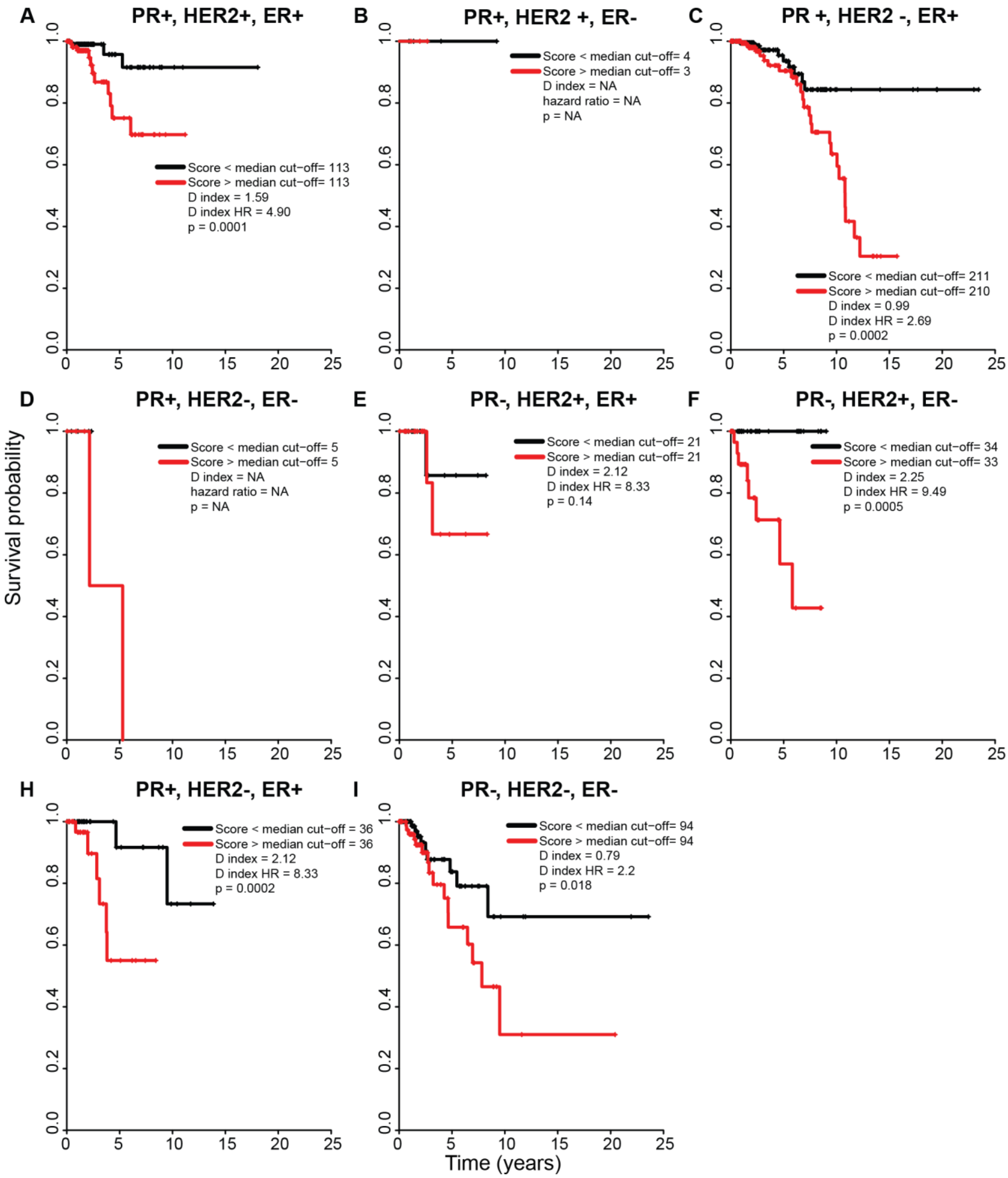


Figure 4.19. APA + clinical prognostic score within histological subtypes. Survival split on the median prognostic score for all possible histological combinations for ER, PR and HER2 receptor overexpression.

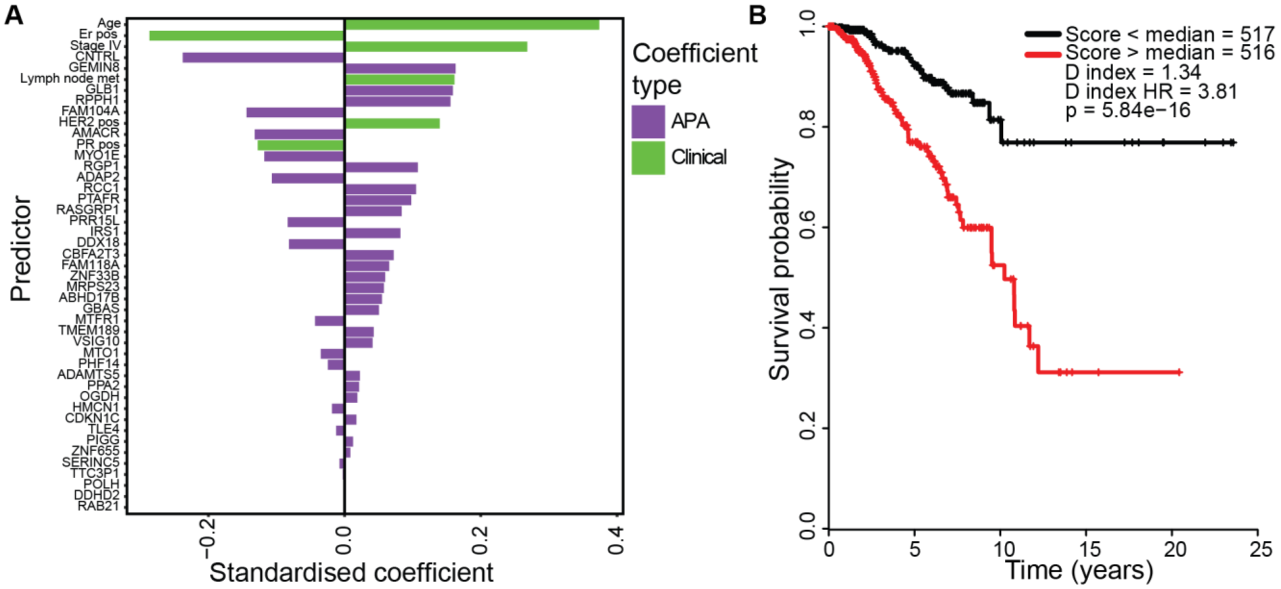


Figure 4.20. The coefficients to be used in a targeted APA based test. A. The coefficient of each predictor of the prognostic score. A higher coefficient indicates a greater impact on the model. Coefficients were standardised to enable the comparison of the absolute contribution of a predictor to the model (standard deviation = 1 in all predictors). B. Survival curve based on 10-fold CV of the model generated using the targeted test parameters. Only APA and gene expression were used as input, α was set to 1.

Table 4.4. Predictor variables and coefficients used in the mPAT model. The table is ordered by contribution of each predictor to the model. Coefficients were standardised to enable the comparison of the absolute contribution of a predictor to the model (standard deviation = 1 in all predictors).

Predictor	Coefficient
Age	0.37
ER pos	-0.29
Stage IV	0.27
CNTRL	-0.24
GEMIN8	0.16
Lymph node met	0.16
GLB1	0.16
RPPH1	0.16
FAM104A	-0.14
HER2 pos	0.14
AMACR	-0.13
PR pos	-0.13
MYO1E	-0.12
RGP1	0.11
ADAP2	-0.11
RCC1	0.10
PTAFR	0.10
RASGRP1	0.08
PRR15L	-0.08
IRS1	0.08
DDX18	-0.08
CBFA2T3	0.07
FAM118A	0.07
ZNF33B	0.06
MRPS23	0.06
ABHD17B	0.05
GBAS	0.05
MTFR1	-0.04
TMEM189	0.04
VSIG10	0.04
MTO1	-0.03
PHF14	-0.02
ADAMTS5	0.02
PPA2	0.02
OGDH	0.02
HMCN1	-0.02
CDKN1C	0.02
TLE4	-0.01
PIGG	0.01
ZNF655	0.01
SERINC5	-0.01

TTC3P1	0.00
POLH	0.00
DDHD2	0.00
RAB21	0.00

4.3 Discussion

There are gene expression-based tests for breast cancer prognosis that are currently in clinical use, such as MammaPrint and Oncotype DX [55, 56]. Unfortunately, these tests are not completely prognostic of breast cancer outcome and are not recommended for all breast cancer subtypes. APA has been suggested as a novel prognostic marker that may help fill these gaps [113, 66, 65]. In this chapter, a database of known APA sites was used to infer APA from TCGA RNA-Seq data. GEO microarray data was also reanalysed, using information about probeset location to determine APA. APA was found to be pervasive in breast cancer, regulated differently to gene expression, and added significant additional prognostic power when determining breast cancer survival.

4.3.1 The effectiveness of the database counting (DBC) method for calling APA from RNA-Seq

As discussed in the introduction to this chapter, a previous attempt has been made at inferring APA from the TCGA by Xia *et al.* [66]. The advantages of using the DBC method presented here over the DaPars method, is that it covers all known APA sites in the human genome and does not rely on inferring APA sites from RNA-Seq coverage profiles. Due to the variable nature of short-read RNA-Seq coverage, it is possible for DaPars to incorrectly call APA sites, resulting in potential false positive discoveries, especially in areas of lower RNA-Seq coverage [271]. The DBC avoids these potentially erroneous results by using a database of known APA sites and does not require the use of any arbitrary cut-offs that are suggested by the DaPars software (e.g. ΔPDUI > 20%), that may have been implemented in an attempt to safeguard against these errors. One potential downside to using the counts from all potential poly(A) sites is that when compared with DaPars, the DBC method is likely to underestimate the magnitude of an APA event if it does not occur across all APA sites. Another downside to using all APA sites from a database in the DBC method is that any incorrectly annotated sites can still be used to call APA. Both methods would require modification before calling more than a single APA event for a given gene, although the DBC method is more likely to capture more subtle changes outside of the two main APA sites, as it incorporates all known sites.

While the TCGA APA results obtained by the DBC method broadly agreed with Xia *et al*. (Figure 4.6 B), many more significant APA events were discovered in this study (427 vs 914). This may have been because the DBC method did not require the use of the arbitrary cut-offs that are utilised by the DaPars method (such as a minimum amount of APA), that may cause false negatives. The DBC method is, however, limited to known APA sites, and would not be appropriate for the detection of novel APA sites or in organisms where an APA database is not available. Furthermore, removing overlapping APA sites to guard against errors due to the unstranded nature of TCGA data may have caused the loss of the detection of some potentially significant APA events. This limitation, however, would not apply if this method were applied to stranded data and is likely more targeted than gradually separating overlapping 3' UTRs, as is performed by the DaPars method.

4.3.2 The effectiveness of the original probeset ordering (OPO) method for calling APA from microarrays

As there are large databases of breast cancer microarray data available from sources such as the GEO [250], APA was also inferred from microarrays to validate the APA events called by the DBC-matched method. Previous methods used to infer APA from microarrays vary slightly but are all based around comparing the ratio of specially defined proximal and distal probesets [136, 113] (Figure 4.4). While these methods unlock APA information from a large cohort of publicly available datasets, microarrays were not designed with APA analysis in mind, and not all platforms are appropriate for this type of analysis. Even when using platforms where this type of analysis is possible, not all genes are covered by probes that were designed to cover multiple APA sites. It should be noted here, as a general point on the analysis of older microarray data, that 3’ probes are a possible confounding factor when interpreting the standard gene expression information (which were designed before the widespread prevalence of APA was known), as a change in 3’ probe expression may indeed be caused by APA.

Analysis of APA from microarrays is also affected by many of the known issues with microarrays, such as cross-hybridization [272] and poor measurement of low abundance transcripts [171]. These issues have been previously addressed for the analysis of Affymetrix HG-U133A arrays by using the fRMA method [175] for probeset expression and normalisation, and the PANP [264] method for determining how likely a gene was to have been expressed above background levels. The OPO method was designed in this study because it is a more conservative pipeline for calling APA from microarray data than previous methods [136, 113]. The OPO suffered from the drawback that it could not distinguish APA when a single probeset fell on both sides of an APA site, which is possible using previous methods that reannotate probes into novel probesets. At the same time, however, the OPO method did not rely on the expression of a few probes partitioned on each side of an APA site, making APA calls more reliable. The OPO method, therefore, increased false negative errors, in order to decrease false positive errors, likely a suitable approach considering the variability in APA calls also highlighted here (Figure 4.6, discussed further in Section 4.3.4). Furthermore, the analysis of APA using the *topconfects* R package (Appendix A2) unified the APA calling statistics and outputs of both the DBC and OPO methods, maintaining consistency between the methods. Methods of calling APA based on newer, higher nucleotide resolution RNA-Seq data will be free from the issues outlined here, but also have their own limitations, as previously discussed, highlighting the importance of utilising multiple datasets to infer APA.

4.3.3 Proximal APA was the predominant shift in breast cancer

Consistent with Xia *et al.* and others [113, 109], it was found in this work that 3' UTR shortening was the predominant APA change in breast cancer in both the TCGA and GEO datasets (Figure 4.2 A). APA was found to be largely independent of tumour subtype, disagreeing with the suggestion from a previous cell line study that different breast cancer subtypes have different APA patterns [168]. Cell lines were likely used in this previous study of APA in breast cancer because they are able to provide a largely homogenous system for the study of the disease. This is not the case for primary tumour biopsies, as they may have varying proportions of stromal and immune-cell infiltrate. Primary tumours may also have intra-tumour heterogeneity, with different sections of a tumour showing distinct gene expression profiles. In this previous cell line study and in the MXM presented in Chapter 3, APA events were highly variable within cell lines, even in cell lines derived from the same parental strain. This is likely because all immortalised cell lines are, by nature, proliferative and would likely undergo many APA alterations during the immortalisation process. Additional APA effects brought about after this process, and subsequent rounds of cell culture are likely to be reflective of altered cellular states, and potentially metastatic ability, but would likely not follow the general cancer-associated trend observed by the DBC-matched method used here. Therefore, it is suggested that future efforts should focus in APA events from primary tumours, ideally using multiple sections of primary tumour biopsies or even higher resolution methods such as single-cell sequencing.

4.3.4 The methods and datasets used to call APA can have a dramatic effect on the APA events called

The largest number of APA events called to date in breast cancer were inferred in this thesis using the DBC-matched and OPO methods (914 and 994 respectively). It was important to compare these results with similar published research, and with one other, to determine the reproducibility of these APA events. The most highly correlated datasets were the DBC-matched dataset with the DaPars dataset (Pearson's r = 0.6) and the OPO dataset with the DBC-matched dataset (Pearson's r = 0.35). The DBC-matched overlap with the DaPars dataset is not surprising considering that the APA events were generated from largely the same samples, and, although it was still the second best-correlated comparison, the low overlap between the DBC-matched and OPO datasets was somewhat unexpected. This suggested that the chosen platform and the method of inferring APA could have a dramatic effect on the genes called. This does not mean that the non-overlapping APA events are not correct, however, as this lack of correlation may be due to the respective restrictions on the number of APA sites that can be called by each method. AS may have also been called as APA in the DBC-matched method where APADB sites were called within internal exons (which the occasionally are). Conversely, it is also possible that given enough samples many more breast cancer APA events may be called by both methods, greatly increasing concordance between datasets. Were this to be the case, it should also be noted that, due to the different APA calling methods used, the ordering of the measured effect size would still likely be different.

The lack of concordance between published methods further highlights the importance of the 100 high confidence APA events that overlapped between the DBC-matched and OPO datasets (Appendix A4). This result was likely only possible due to the conservative approach taken when calling APA that was applied to both datasets. The significance of these results

($p < 0.05$, Chi-squared test) was further highlighted by the fact that a 100% overlap was impossible, as HG-U133A microarrays can only be used to infer APA from genes covered by multiple probesets. RNA-Seq coverage plots also appeared to strongly support these 100 events (Figure 4.8), providing more evidence that the APA events presented in this thesis are the most reproducible of any of the previous studies of APA in breast cancer performed to date. These 100 high confidence events could potentially serve as novel biomarkers or therapeutic targets for breast cancer treatment in future.

4.3.5 Deregulation of miRNA expression in breast cancer was not subtype-specific

As miRNAs are known to bind primarily to the 3' UTRs of mRNAs, miRNA expression from the TCGA was also evaluated in this work. There were broad changes in miRNA expression in breast cancer across the TCGA cohort, with 77% of miRNAs changing in expression between tumour and normal samples (Figure 4.14 A/B). Interestingly, the vast majority of miRNA expression changes in the TCGA (Figure 4.14 C) were not subtype-specific. This is contrary to previous suggestions in the literature that APA in TNBCs is subtype specific [273, 270]. This may be due to the time and cost associated with analysing miRNA expression in primary human tumours in these early studies, resulting in the analysis of a relatively low number of tumours (93 and 15 samples respectively). Studies of miRNA expression in breast cancer have also often only considered miRNA expression in a single subtype, such as TNBCs [274, 275], resulting in the identification of 'subtype-specific' miRNA markers that may also be dysregulated in many other breast cancer subtypes. The results presented here are also consistent with the broader deregulation of APA events, which were also observed not to be subtype-specific in this research. This suggests that there are broad patterns of APA and miRNA dysregulation in breast cancer and that tumour subtype is not a major driver of these changes.

4.3.6 Evaluation of prognostic modelling of breast cancer outcome in the TCGA

APA has previously been used to determine breast cancer prognosis using TCGA RNA-Seq data [66], which was performed on matched tumour vs normal pairs of samples. There has, however, not yet been a predictive model generated using APA data, that was trained on all breast cancers in the TCGA. APA, gene expression and clinical data were, therefore, examined in this thesis for the prediction of patient survival. APA, in addition to clinical data, was found to be significantly more prognostic than clinical data alone ($p < 0.05$, D index comparison). It is especially encouraging that outcome could be predicted using APA in this study, even in the previously unpredictable TNBC subtype (Figure 4.19 I). This result is consistent with predictions based on inferring APA in TNBCs from microarrays [65], which also found APA to be highly prognostic of breast cancer outcome. Strangely, this microarray-based study has been cited very few times, possibly due to the technical challenges associated with inferring APA from microarrays previously discussed. When the final prognostic model was tested by bootstrapping, using all APA, gene expression and clinical predictors (penalty factor = 0.5 for clinical predictors, Figure 4.17 A), 95% CI's all touched zero, except for the stage IV predictor, meaning APA events were not always included in 95% of bootstrapped models. It would be preferable if these coefficients were always included, however, it is encouraging to note that when utilised, they have a consistent sign (direction of APA) and often a large contribution to the model.

While Xia *et al.* [66] have suggested that APA + clinical data is more prognostic of breast cancer outcome than gene expression + clinical data, their comparison only included the expression of genes that they had selected for APA prediction, making the gene expression predictions from these genes almost certainly an underestimate of the predictive power of gene expression. Despite this, it was also found in this study that APA + clinical data was more predictive of survival than gene expression + clinical data. It is, therefore, suggested that clinical and gene expression data may be measuring largely the same proliferation-associated pathways, while APA may be adding some new prognostic information. The gene expression-based Oncotype DX recurrence score, for example, can be inferred using only clinical variables [55], suggesting that there is indeed an overlap between prognostic clinical and gene expression markers.

Most previous genomic studies that attempt to predict breast cancer outcome [44, 56, 276] did not take into account as many clinical variables as are currently available. For example, the PAM50 study [45] assessed tumour prognosis after separating tumours into 5 ‘intrinsic subtypes’ based on the expression of 50 genes. While it is convenient to classify breast cancer into 5 groups, the subtypes do not account for the full spectrum of breast cancer variability, and it is still often necessary to consider tumour heterogeneity within these subtypes when determining treatment decisions [165]. It has been demonstrated that the clinical + APA model generated here adds additional prognostic power, even within these gene expression subtypes (Figure 4.18). This provides further evidence that APA may be operating as a somewhat separate system to gene expression, which would explain the increase in prognostic power that has consistently been observed. It follows that additional known systems of mRNA regulation, such as miRNA expression and AS, would also add additional prognostic power, were they to be included in future prognostic models. Based on the separation of patients on the Kaplan-Meier plots, APA seems to be more prognostic of survival around 7 years after diagnosis, whereas clinical and gene expression data seems to have stronger predictive capacity in the first 7 years (Figure 4.16 D and E). The loss of prognostic power after more than 5 years post-diagnosis has been identified as a known issue with current multigene breast cancer prognostic tests [225]. The prognostic findings of this research, and the tumour-like patterns of APA that are present in some normal breast tissue samples, hints at the hypothesis that APA may be reflective of a patient’s inherent potential to retain dormant tumour cells following treatment, however, investigating this mechanism further is beyond the scope of this thesis.

The final step in analysing the prognostic capacity of APA was to validate these events for use in a diagnostic test. With this in mind, a model was generated that had a smaller number of prognostic APA events, suitable for eventual validation by the mPAT method [215]. This model was still highly prognostic of breast cancer outcome, while only requiring the measurement of 39 APA events and six clinical variables (Figure 4.20). It is, unfortunately, unlikely that the final mPAT based test will exactly replicate a model designed around the TCGA, due to different primer efficiencies and different biases of both methods. APA events that are recapitulated, however, may prove a strong source of additional prognostic information available to a breast cancer patient, regardless of subtype.

4.3.7 APA events should be considered at the gene-specific level moving forward

It has been suggested by Lembo *et al*. [136] that shorter 3' UTRs are in general associated with poorer breast cancer outcome. This assertion is also consistent with reports of predominantly proximal proliferation- and cancer-associated APA events [112, 109]. While many APA events tend to be proximal in tumour formation [109], APA regulation more broadly, has been suggested to be far more complex at individual sites, rather than generally increasing gene expression and protein output [162]. It has been shown that, in many cases, it is actually the gain of RNA stabilising elements through distal APA that causes an increase in translational efficiency [239, 188, 189]. This is consistent with the prognostic test developed here in which both proximal and distal APA events were included in the final model. Proximal APA was, however, the dominant change between breast tumours and normal breast tissue in this study, possibly providing an explanation for why these events are also overrepresented in previous studies of APA in breast cancer prognosis.

Despite the identification of many novel and prognostic APA events in this chapter, the exact mechanisms behind APA in breast cancers and in breast cancer metastasis are still unclear. It has been suggested previously that an increase in the expression of mRNA processing factors is associated with an increase in the amount of proximal APA [66, 113]. Surprisingly, in this thesis, this trend was also associated with an increase in distal APA (Figure 4.11 A). It is therefore suggested that the expression of additional CPA factors, in combination with more gene-specific regulation of APA, is required to drive tumour progression and that APA events should be considered and targeted individually in future. APA will likely need to be studied at the single gene level, on a genome-wide scale, in combination with the host of associated CREs (such as AU rich elements) and TAFs (such as miRNAs), before a complete understanding of these events can be obtained. In future, there is the potential for the 100 high confidence breast cancer-associated APA events determined here, such as the cell adhesion-related *ATP2A2* [265], to be investigated as possible biomarkers or evaluated as targets for novel breast cancer treatment. There is also the potential for the APA based prognostic test suggested here to be validated and implemented. Some recalibration of this model and optimisation of APA sites may be required when transitioning from RNA-Seq to mPAT. The model may require recalibration as the APA site counts at proximal sites obtained from RNA-Seq will be a mix of reads from proximal and distal transcripts. Optimisation of APA sites may also be required due to the specific primer requirements of the mPAT method and other method-specific biases.

Chapter 5: Computational tools for the design and analysis of custom NGS experiments

5.1 Introduction

It has been suggested for nearly two decades that biologists now require computational skills to interpret the ever-growing amounts of data that is generated from high throughput genomic methods [277]. It is also now well known that biology has evolved into a far more quantitative discipline than it previously has been, and is now based largely around data science [278, 279, 280, 281]. Although this shift is slowly starting to filter through to both undergraduate and postgraduate biology courses [282, 283, 284], the widespread adoption of computational methods by biologists is still in its infancy. Furthermore, education of students alone does not address the large knowledge gap that currently exists among established researchers. This challenge must be managed going forward, in order to effectively leverage the wealth of existing scientific expertise against exciting new statistical methods.

While a wet-lab researcher will typically understand the biological questions they are trying to answer, when NGS methods are involved they may be unsure how to go about programmatically answering these questions. Incorrect and misleading conclusions may be drawn when inexperienced researchers make attempts to interpret NGS data without being fully aware of how to spot potential biases that exist in almost all NGS libraries [285]. Successfully connecting biologists to robust analysis of their own NGS data, therefore, often requires an intermediate step. Examples such as *Degust* (http://degust.erc.monash.edu/) and *Glimma* [286], represent attempts to abstract away the need for programming knowledge in the analysis of RNA-Seq data. These tools provide a graphical interface to a biologist for the analysis of RNA-Seq experiments and do not require any programming skills to use. While an understanding of the core principles of an experiment is still inevitably required for a proper statistical analysis, this abstraction removes much of the burden of understanding the statistical methods and programming from the biologist, while remaining statistically rigorous.

This is not the optimal solution, as ideally, all researchers would have a complete understanding of every stage of an experiment from start to finish. Computational skills can also prove useful in the design of an experiment, as the later analysis steps will be better considered before beginning the experiment. However, these tools and many others provide a lower barrier of entry to bioinformatic data analysis and may represent the first step for a wet-lab researcher learning bioinformatics.

Computational power is often also leveraged in the design phase of an experiment, including in the design of PCR-based experiments. Predicting the melting temperature of primers in a standard PCR reaction, for example, allows for more targeted and efficient reactions [287]. Similar in design to standard PCR-based methods are the 3' rapid amplification of cDNA ends (3' RACE) methods. These methods generally utilise an oligo-dT universal reverse primer, to bind to all polyadenylated molecules in a total RNA extraction and a gene-specific forward primer to select for a gene of interest. There are a number of considerations that need to be taken into account when designing primers for 3' RACE experiments, such as the mPAT method, that are not required to be accounted for when designing primers for a standard PCR reaction [215]. In conventional PCR reactions, specific sequences can be selected for amplification through the use of both forward and reverse primers, resulting in two layers of selection that, in turn, results in a specific product. As the reverse primer in an mPAT experiment takes advantage of the poly(A) tail of every mRNA transcript, selection of a specific gene sequence to amplify is reliant on the forward primer alone, greatly increasing the requirement for specificity. Maintaining this specificity, while accounting for other factors such as melting temperature (Tm), GC base content, appropriate length and avoiding self-duplexing, can also be quite laborious when designing multiple primers for a multiplex experiment. Automation of this process ultimately saves time and reduces the impact of human error in these experiments.

Described in this chapter are methods to allow biologists without computational skills to perform point and click data analysis of custom datasets generated in the RNA Systems Biology Laboratory, and to design effective primers for 3’ RACE experiments. The first tool described in this chapter was primarily designed to visualise the cumulative distribution of poly(A) tail lengths for any gene, in any sample, from any PAT-Seq experiment undertaken in the RNA Systems Biology Laboratory. The second tool described in this chapter is a method for automated selection of primers that improves binding specificity in the targeted mPAT [215] and alPAT methods. This method considers all parameters that a researcher would normally be manually required to account for when designing primers for a targeted ‘PAT’ type experiment. Together, these methods allow for more streamlined experimental design and analysis of data generated from ‘PAT’ type experiments (or similar methods), with increased accuracy and reduced potential for human error.

5.2 Results

5.2.1 *The RNA Systems Explorer* App (*RSER*)

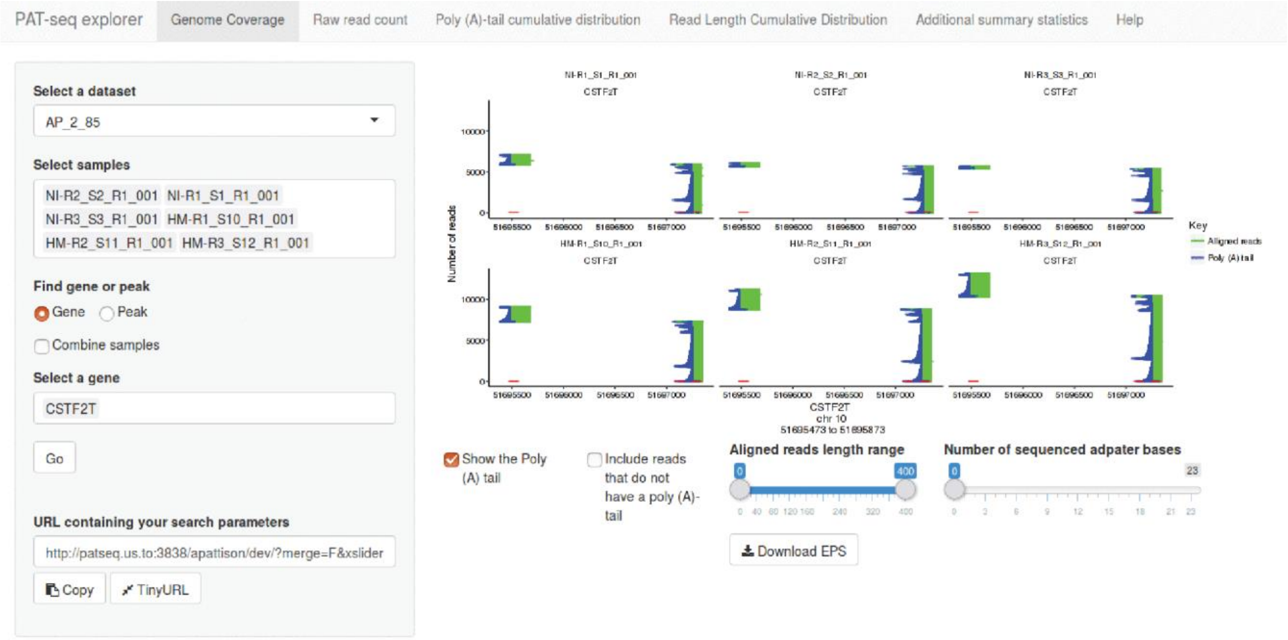
As described in Chapter 1, the distribution of poly(A) tail lengths observed, for a given gene in a PAT-seq library, gives information about the actual distribution of poly(A) tail lengths for that gene. Some distortions in absolute poly(A) tail length are, however, expected due to the limited length of PAT-seq reads and issues with sequencing homopolymers present in current sequencing technologies [213]. Tails beyond ~130 nt must also either be truncated or not seen at all, due to the sequence length constraints of the Illumina HiSeq (150 nt), when used as part of the PAT-seq protocol. Differences in the distribution of poly(A) tail lengths observed in reads between samples are therefore of particular interest, as these changes are still measurable despite the underlying uncertainty in poly(A) tail length measurement. As all samples are presumed to be subject to the same distortions of observed poly(A) tail lengths, any difference in observed distributions indicates some difference in the true distributions of the poly(A) tail lengths for that gene.

Poly(A) tail length and its role in the regulation of cellular processes has historically received less attention than other known forms of transcriptional regulation, especially on a genome-wide scale. As such, there have been few methods previously developed for the specific purpose of visualising poly(A) tail lengths. Statistics, such as the mean and median are helpful in determining a general shift in poly(A) tail length between conditions, but may not accurately reflect changes to sub-populations within a particular sample or group of samples. A clearer way of visualising poly(A) tail data is the cumulative distribution, as it can distinguish changes in poly(A) tail length sub-populations that may not be clear from other descriptive statistics. Plots of poly(A) tail length distribution would therefore be useful in the visualisation of the output of PAT-seq experiments. The genome wide nature of PAT-seq makes generating these plots for every gene impractical, time consuming and a waste of hard disk storage space.

A web application (web app) was built to generate poly(A) tail cumulative distribution plots and associated information as required. Using the R programming language [209] and the Shiny R package [288], the *RNA Systems Explorer* (*RSER*) app was created (rnasystems.erc.monash.edu:3838/apattison/dev/), which was primarily designed to plot a cumulative distribution of poly(A) tail lengths for given gene(s), in given sample(s) (or specific grouping of samples), in a PAT-seq (or 'PAT' type) experiment, and was also expanded to include additional information. The app can be easily applied to any *Tail-tools* [169] output and allows a user to look in detail at the poly(A) tail length and associated 3' UTR dynamics of any gene(s) or adenylation site(s) present in a dataset. Examples of plots downloaded from the *RSER* can be seen in Figure 3.17.

5.2.2 *RSER* tab 1: Experiment, sample, gene selection and genomic coverage

The first tab of the web app (Figure 5.1) enables a user to select from any PAT-seq or mPAT experiment that was previously performed in the RNA Systems Biology Laboratory. The user can then select samples as required, or arrange the samples into any grouping they choose. The user then selects genes or peaks of interest and starts the web app by pressing the "Go" button. The initial output is similar to that seen when examining BAM files in *IGV*, with peaks identified by *Tail-tools* (or manually depending on the type of experiment) shown in red. The advantage of this plot over *IGV* is that it also shows poly(A) tail length, which is not a standard feature of *IGV*. The reads are sorted by 3' position, then by poly(A) tail length, and are piled up to give the user a visual representation of the reads that contributed to the poly(A) tail length distribution. The user can also choose to see just the genomic part of the reads, or



Select a dataset

AP_2_85

Select samples

NI-R2_S2_R1_001 NI-R1_S1_R1_001 NI-R3_S3_R1_001 HM-R1_S10_R1_001 HM-R2_S11_R1_001 HM-R3_S12_R1_001

Find gene or peak

● Gene ○ Peak

□ Combine samples

Select a gene

CSTF2T

URL containing your search parameters

http://patseq.us.to:3838/apattison/dev/?merge=F&xslider

☑ Show the Poly (A) tail

□ Include reads that do not have a poly (A)-tail

Aligned reads length range: 0 – 400

Number of sequenced adpater bases: 0 – 23

Figure 5.1. *RSER* genome coverage and options tab. This is the main page of the RSER. It is where a dataset that was generated in the RNA Systems Biology Laboratory can be selected. Specific samples, or a custom sample grouping, can also be selected from this tab. A user can then select a gene or peak and view any mapped reads from all selected samples or groupings.

include reads that did not end in a poly(A) tail. To modify reads that are plotted by alignment to the genome, there is a slider that allows for the filtering of reads in a certain genomic length range. Finally, to obtain reads that have been sequenced to completion, there is a slider to filter out reads that do not contain a required number of 3’ adapter bases. The 3’ adapter is the sequence added to all mRNA molecules in PAT type experiments for the subsequent binding of Illumina adapters prior to sequencing. If a user wishes to save their currently selected parameters, they can copy the URL from the box provided. Entering the URL into a web browser will then bring up the web app with the saved search parameters loaded. All plots generated by the web app can additionally be downloaded in EPS format to be saved locally.

5.2.3 *RSER* tab 2: Assessing raw read counts for quality control

As was discussed in Chapter 3, a poly(A) tail distribution may be unreliable if it is based off a small number of reads. The raw read count tab (Figure 5.2) shows the raw read counts from all samples/groups selected by the user on the previous page. It is useful to see the number of reads that go into the poly(A) cumulative distribution plot to prevent false conclusions based on a small number of poly(A) reads. It is also useful to check that negative control samples in mPAT experiments have not produced any reads, as reads in a control sample may indicate PCR product contamination. Finally, discrepancies in the overall number of reads present in samples from the same condition may indicate an issue in the reproducibility of the experiment, or pooling of uneven amounts of PCR product when combining multiplex PAT type experiments for sequencing.

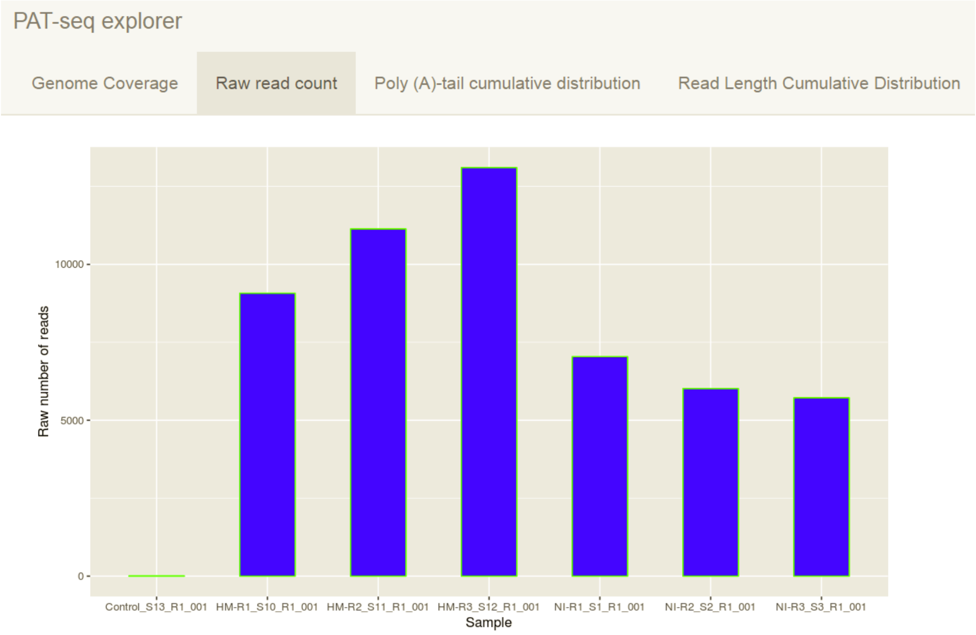


Figure 5.2. *RSER* raw reads count tab. This tab shows the raw read count taken from each BAM file in a PAT type experiment. This provides an indication of the reliability of poly(A) tail cumulative distributions in the next panel.

5.2.4 *RSER* tab 3: The poly(A) tail reverse cumulative distribution plot

The poly(A) tail reverse cumulative distribution tab (Figure 5.3) shows the distribution of poly(A) tail lengths for the samples/groups that were selected in the first tab. The x-axis represents poly(A) tail length and the y-axis of the plot is an reverse cumulative distribution, converted to a percentage. The reverse cumulative distribution was used as opposed to a standard cumulative distribution, as poly(A) tails will slowly be decayed over the life of an mRNA molecule. As the longest tail length may vary between experiments depending on the organism, there is a slider that allows the user to resize the x axis. The selected genes are displayed both at the top of the plot and in the legend, which can be removed by un-checking the check-box. The distribution can then be used to assess poly(A) tail length changes between conditions for any gene measured in a PAT-seq experiment. Multiple genes may also be plotted simultaneously.

5.2.5 *RSER* tab 4: The read length reverse cumulative distribution plot

The last plot tab is the read length cumulative distribution tab (Figure 5.4). This tab is similar to the poly(A) tail cumulative distribution plot, except that it shows the length distribution of the genomic portion of the reads, rather than the poly(A) tail. The reads presented in this tab are expected to group into tighter populations around APA sites, rather than forming a smooth distribution. This is because 3’ ends are relatively fixed by sequence motifs such as the polyadenylation signal [200]. As PAT-seq and mPAT can only measure reads of a fixed length, the length of the genomic portion of the read will also influence the maximum poly(A) tail length that can be measured. Showing a read length distribution allows a user to check if their tail length changes may be due to differences in poly(A) tail length or slight variations in

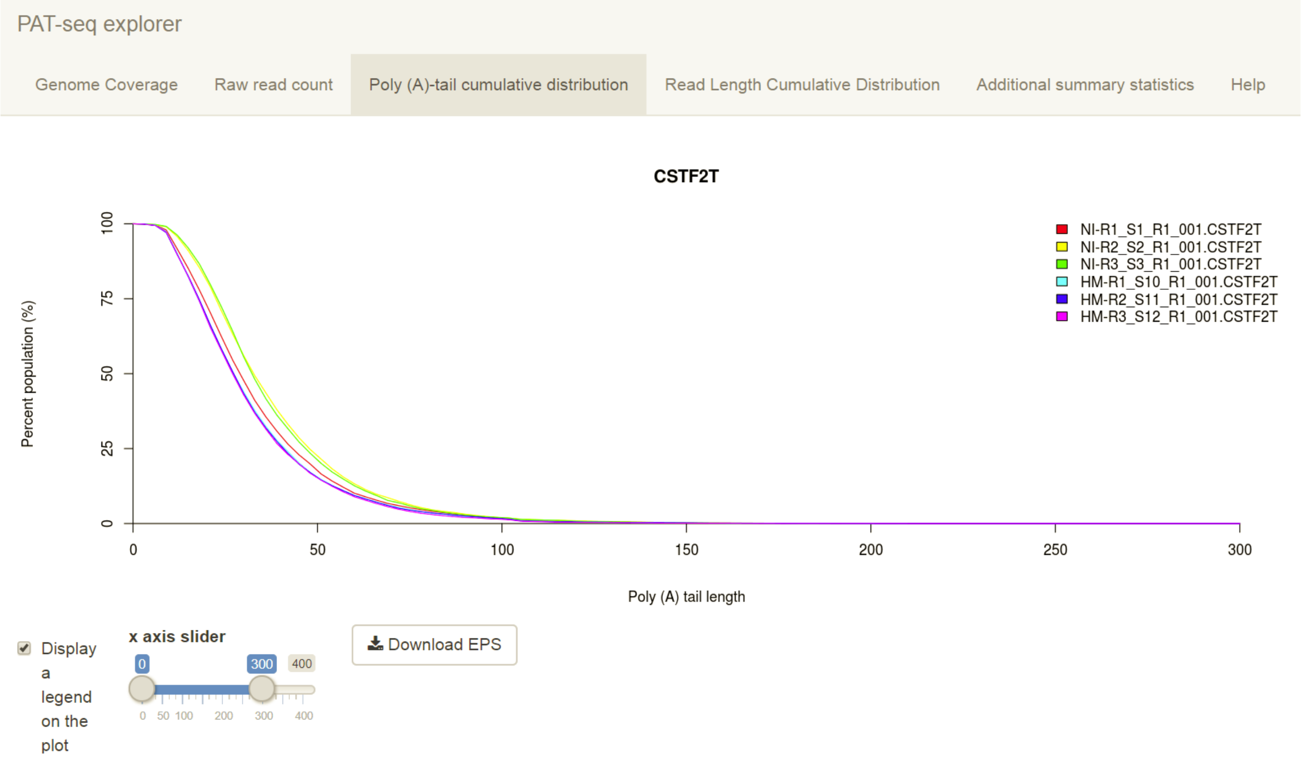
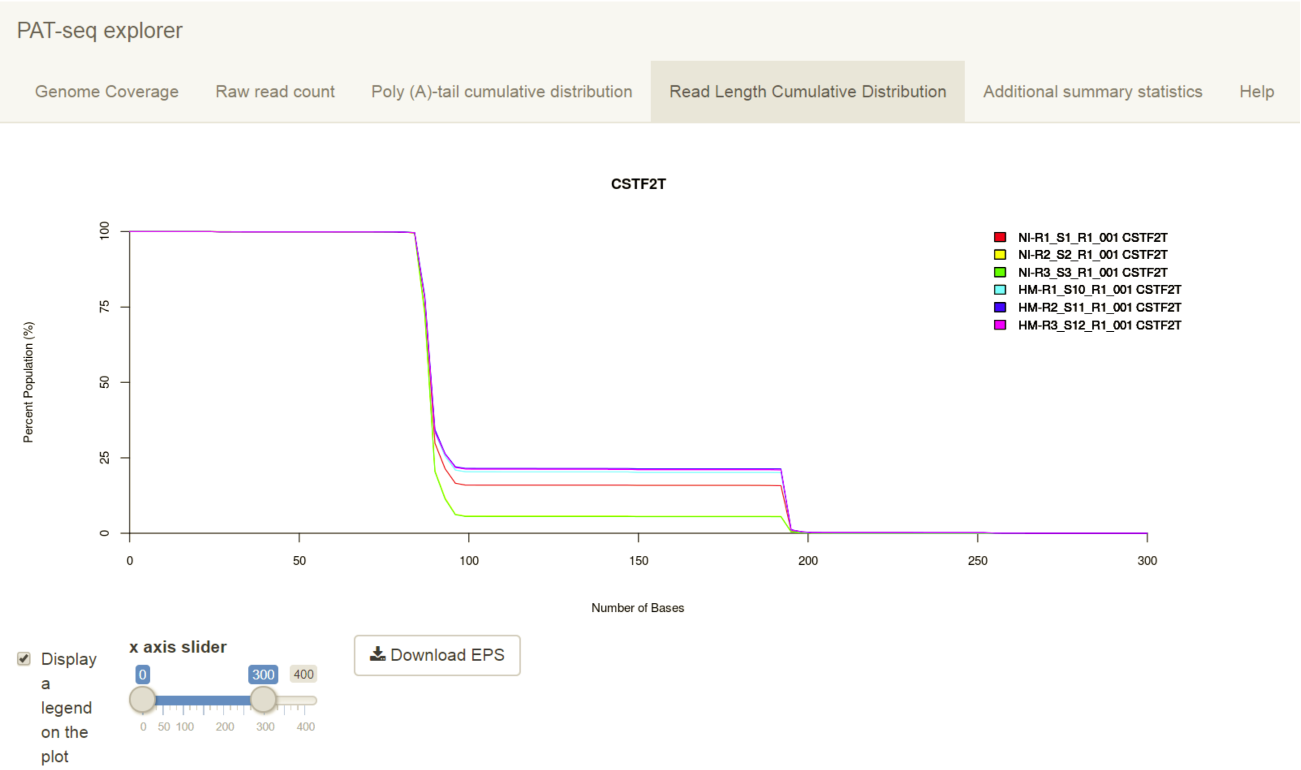


Figure 5.3. *RSER* poly(A) tail inverse cumulative distribution tab. This tab shows a cumulative distribution of poly(A) tail lengths of the selected gene/peak, for each sample/group in a selected experiment.



☑ Display a legend on the plot

x axis slider: 0 – 300 (max 400)

Figure 5.4. *RSER* read length cumulative distribution tab. This tab shows the inverse cumulative distribution of the genomic portion of the read lengths of a PAT type experiment. This panel can be compared with the poly(A) tail cumulative distributions panel to determine if tail length changes are real or due to a change in read lengths.

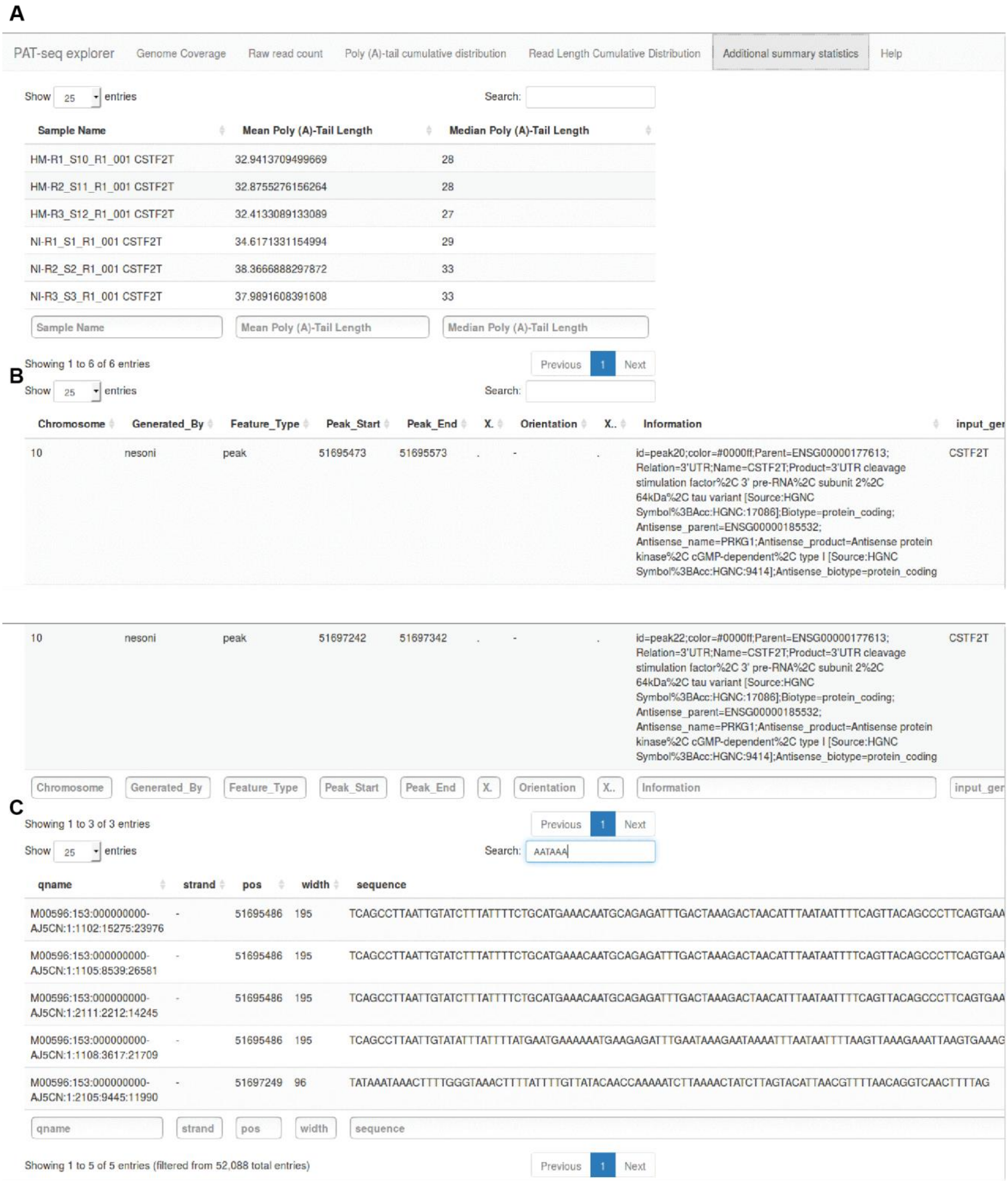
3’ end usage. Visualising 3’ end usage may also show micro heterogeneity around the 3’ end of a single peak.

5.2.6 *RSER* tab 5: The summary statistics page

While the reverse cumulative distribution is the optimal way to assess poly(A) tail length changes with the *RSER*, summary statistics are also provided, as these may be useful in the results text of a paper in circumstances where clear differences can already be observed from the plot. A summary statistics tab was, therefore, also included to show some additional information about the raw reads that are plotted by the web-app, as well as some other useful summary statistics. The top panel of the summary statistics tab (Figure 5.4 A) shows the mean and median of the poly(A) tail cumulative distributions currently displayed by the web app. More information about the genomic location of the regions used to select adenylated reads is also provided (Figure 5.4 B). The user can also look at the raw reads that contributed to the output of the web app (Figure 5.4 C). The reads displayed in this section can be interactively searched if a researcher wishes to see more specific detail about the raw BAM file entries. Also included is a ‘help’ tab with information about the PAT-Seq experiment and the authors of the app.

5.2.7 *3Primer* an automated primer design tool for 3’ RACE experiments

As discussed in the introduction to this chapter, in many 3’ RACE experiments, there is only one site of specificity for primer binding as a universal reverse primer is often used. Choosing a specific binding site, within a limited sequence range, while at the same time maintaining optimal PCR parameters can be a challenging task to perform manually. Based on Primer3 [223], the custom primer design tool, *3Primer* was designed, using the Python programming


A
PAT-seq explorer
Genome Coverage
Raw read count
Poly (A)-tail cumulative distribution
Read Length Cumulative Distribution
Additional summary statistics
Help

Show 25 entries Search:

Sample Name	Mean Poly (A)-Tail Length	Median Poly (A)-Tail Length
HM-R1_S10_R1_001 CSTF2T	32.9413709499669	28
HM-R2_S11_R1_001 CSTF2T	32.8755276156264	28
HM-R3_S12_R1_001 CSTF2T	32.4133089133089	27
NI-R1_S1_R1_001 CSTF2T	34.6171331154994	29
NI-R2_S2_R1_001 CSTF2T	38.3666888297872	33
NI-R3_S3_R1_001 CSTF2T	37.9891608391608	33

Showing 1 to 6 of 6 entries

Show 25 entries Search:

Chromosome	Generated_By	Feature_Type	Peak_Start	Peak_End	X.	Orientation	X..	Information	input_ger
10	nesoni	peak	51695473	51695573	.	-	.	id=peak20;color=#0000ff;Parent=ENSG00000177613; Relation=3'UTR;Name=CSTF2T;Product=3'UTR cleavage stimulation factor%2C 3' pre-RNA%2C subunit 2%2C 64kDa%2C tau variant [Source:HGNC Symbol%3BAcc:HGNC:17086];Biotype=protein_coding; Antisense_parent=ENSG00000185532; Antisense_name=PRKG1;Antisense_product=Antisense protein kinase%2C cGMP-dependent%2C type I [Source:HGNC Symbol%3BAcc:HGNC:9414];Antisense_biotype=protein_coding	CSTF2T
10	nesoni	peak	51697242	51697342	.	-	.	id=peak22;color=#0000ff;Parent=ENSG00000177613; Relation=3'UTR;Name=CSTF2T;Product=3'UTR cleavage stimulation factor%2C 3' pre-RNA%2C subunit 2%2C 64kDa%2C tau variant [Source:HGNC Symbol%3BAcc:HGNC:17086];Biotype=protein_coding; Antisense_parent=ENSG00000185532; Antisense_name=PRKG1;Antisense_product=Antisense protein kinase%2C cGMP-dependent%2C type I [Source:HGNC Symbol%3BAcc:HGNC:9414];Antisense_biotype=protein_coding	CSTF2T

Showing 1 to 3 of 3 entries

Show 25 entries Search: AATAAA

qname	strand	pos	width	sequence
M00596:153:000000000-AJ5CN:1:1102:15275:23976	-	51695486	195	TCAGCCTTAATTGTATCTTTATTTTCTGCATGAAACAATGCAGAGATTTGACTAAAGACTAACATTTAATAATTTTCAGTTACAGCCCTTCAGTGAA
M00596:153:000000000-AJ5CN:1:1105:8539:26581	-	51695486	195	TCAGCCTTAATTGTATCTTTATTTTCTGCATGAAACAATGCAGAGATTTGACTAAAGACTAACATTTAATAATTTTCAGTTACAGCCCTTCAGTGAA
M00596:153:000000000-AJ5CN:1:2111:2212:14245	-	51695486	195	TCAGCCTTAATTGTATCTTTATTTTCTGCATGAAACAATGCAGAGATTTGACTAAAGACTAACATTTAATAATTTTCAGTTACAGCCCTTCAGTGAA
M00596:153:000000000-AJ5CN:1:1108:3617:21709	-	51695486	195	TCAGCCTTAATTGTATATTTATTTTATGAATGAAAAAATGAAGAGATTTGAATAAAGAATAAAATTTAATAATTTTAAGTTAAAGAAATTAAGTGAAAG
M00596:153:000000000-AJ5CN:1:2105:9445:11990	-	51697249	96	TATAAATAAACTTTTGGGTAAACTTTTATTTTGTTATACAACCAAAAATCTTAAAACTATCTTAGTACATTAACGTTTTAACAGGTCAACTTTTAG

Showing 1 to 5 of 5 entries (filtered from 52,088 total entries)

Figure 5.5. *RSER* read length cumulative distribution tab. This tab shows the inverse cumulative distribution of the genomic portion of the read lengths of a PAT type experiment. This panel can be compared with the poly(A) tail cumulative distributions panel to determine if tail length changes are real or due to a change in read lengths.

language, to address these issues. *3Primer* takes Tail-tools peak calling (or any correctly formatted APA site information) from any PAT-seq experiment [169] or any database of possible APA sites in GFF format as input. The tool then designs specific primers for any genes required by the user. The primers that are generated are optimised to the parameters that are specific to the mPAT experiment. A schematic of the basic functionality of *3Primer* can be seen in Figure 5.6. The parameters of *3Primer* are not limited to the mPAT method, however, and could be altered to suit most 3'-focused PCR-based experiments by changing the program's settings. The tool also allows users to set their own melting temperatures, GC bias, GC clamps and optimal primer length as they would normally when using Primer3 from the command line. Importantly, the pipeline also has the added benefit of selecting the most specific primer from the pool of possible primers that are returned.

The *3Primer* method selects for primers with the highest sequence specificity using *BLAST* [224], making off-target binding of primers far less likely. Specificity is defined as the lowest number of perfect 10+ base matches to the reference genome in the last 15 bases of the primer sequence. An example of this increased specificity can be seen in Figure 5.7, where *3Primer* designed primers were 3-7 times more specific than human designed primers, while still binding within the required range, and meeting all thermodynamic specifications. For greater computational speed, the *BLAST* portion of the tool may also be multi-threaded. The full process undertaken by *3Primer* to suggest primer sequences can be seen in the methods section and Figure 2.4. If used correctly, the *3Primer* pipeline greatly reduces human error, saves a substantial amount of time in PCR primer design and results in the most specific primers possible (given experimental constraints) for 3' RACE experiments. The code that comprises the tool is available on GitHub at the following address: https://github.com/AndrewPattison/3Primer/blob/master/Software/3PrimeR.py

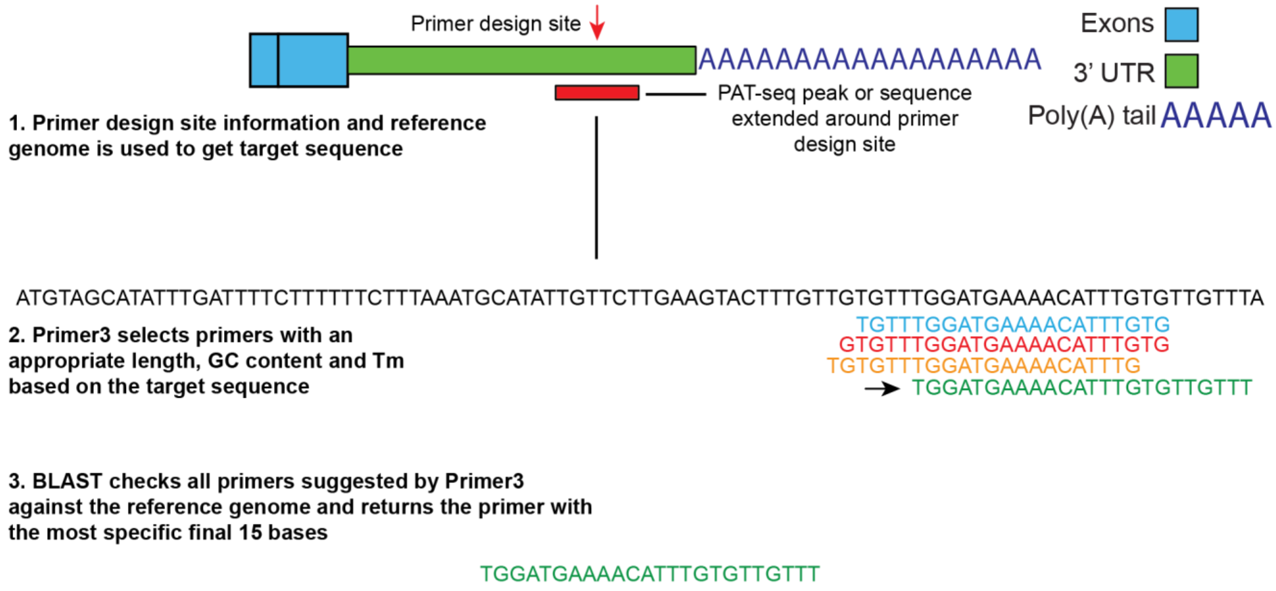

Primer design site
AAAAAAAAAAAAAAAAAA
Exons
3' UTR
Poly(A) tail AAAAA
PAT-seq peak or sequence extended around primer design site
1. Primer design site information and reference genome is used to get target sequence
ATGTAGCATATTTGATTTTCTTTTTTCTTTAAATGCATATTGTTCTTGAAGTACTTTGTTGTGTTTGGATGAAAACATTTGTGTTGTTTA
2. Primer3 selects primers with an appropriate length, GC content and Tm based on the target sequence
TGTTTGGATGAAAACATTTGTG
GTGTTTGGATGAAAACATTTGTG
TGTGTTTGGATGAAAACATTTG
TGGATGAAAACATTTGTGTTGTTT
3. BLAST checks all primers suggested by Primer3 against the reference genome and returns the primer with the most specific final 15 bases
TGGATGAAAACATTTGTGTTGTTT

Figure 5.6. Specific primer design for 3' RACE experiments. A schematic of the major steps that are taken by the 3Primer pipeline to generate specific PCR primers.

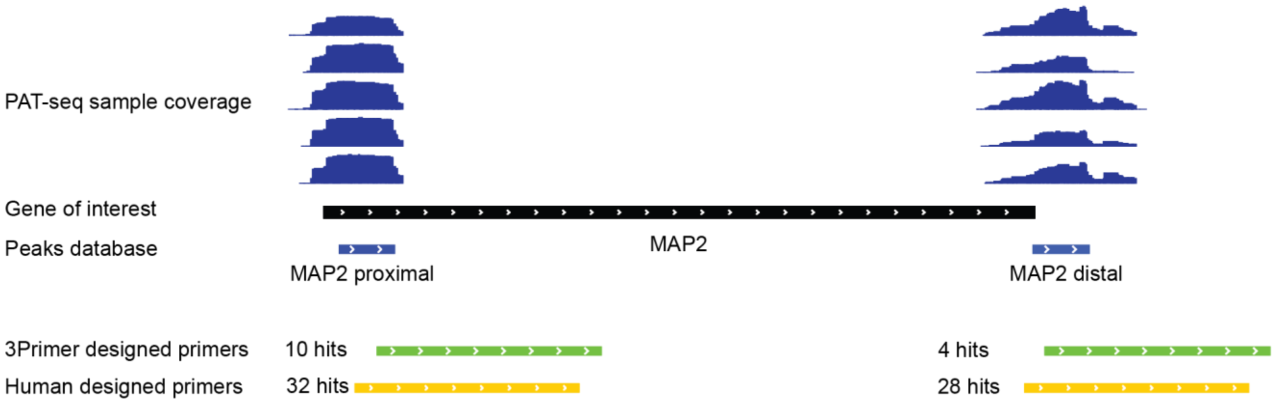


Figure 5.7. The output of the *3Primer* primer design program compared to primers designed by a human. This figure shows two sets of primers designed to test, using the mPAT method, APA that was initially measured in the *MAP2* gene of Saccharomyces cerevisiae using the PAT-seq method. The regions covered by the *3Primer*-designed primers (green bars) were positioned at a similar location to those designed by a human (yellow bars), with the key difference being primer specificity. For the proximal APA site, there were 32 perfect 10+ base matches to the reference genome in the last 15 bases of the human-designed primers, while there were only 10 in the *3Primer*-designed primer. At the distal APA site, there was once again 28 matches for the human-designed primer, vs only 4 for the 3Primer-designed primer.

5.3 Discussion

Currently, there is a knowledge gap that wet-lab biologists must bridge in order to design and reliably interpret the results of their NGS experiments [281]. It may be beneficial to bridge this gap with tools that provide a graphical interface to simplify a complex analysis, such as the interpretation of RNA-Seq experiments [286], or a complicated process (such as primer design [223]) that would normally require programming skills. The two tools described in this chapter (*RSER* and *3Primer*) were created with both better experimental design and simplified interpretation of complex datasets in mind. These tools minimise the requirement for advanced technical knowledge or computing skills that would otherwise be essential to perform these tasks.

5.3.1 Benefits and limitations to the use of the *RSER* app

The *RSER* app was designed to allow researchers lacking in computational experience to explore complex PAT-seq (or 'PAT' type) data, with the point and click functionality that they are generally more familiar with. A simplified version of the RSER is being utilised in an upcoming paper on poly(A) tail length in *C. elegans Gld2* mutants. One key advantage of the *RSER* and other R/Shiny apps for the communication of complex genomic datasets is that they enable quick, reproducible analyses to be performed by the researcher that designed the experiment. This allows the researcher to spend less time seeking help from bioinformaticians, and spend more time analysing their own data, saving time for both parties. These apps also remove the need for a bioinformatician to consistently reproduce the results of their original analyses with only slight variations each time. The online availability of these tools also makes published data more easily accessible to members of the wider research community, that may lack the technical expertise or time to reproduce the results of a paper from the original data and draw their own conclusions. However, there can also be drawbacks to the use of these apps, especially if they are not properly implemented, as a mistake in the generation of results could also be consistently repeated. The reliance on these apps may also stymie the motivation of researchers to study statistical methods, potentially resulting in a reduced statistical knowledge base in the biological sciences, which is not desirable.

5.3.2 Benefits and limitations to the use of *3Primer*

The *3Primer* tool is less about engagement with biologists than the *RSER*, and more about increasing the efficiency of a targeted 3' sequencing experiment. Ideally, however, it will eventually be converted to a web app, with similar point and click functionality as is present in the *RSER*. By utilising the power of computers to perform a repetitive task, primer design is achieved much more rapidly and the results are as specific as they can possibly be, making the chances of a highly non-specific primer overwhelming a sequencing run far less likely. The *3Primer* method could be improved in future to better handle APA sites in close proximity. Currently, the method selects primers from the database of APA sites that it is given. When these sites overlap, the same primer may be chosen for both sites, which may not be desirable when both sites require analysis. The *3Primer* method could also be updated to determine primer specificity based on thermodynamics, rather than sequence, as this is a slightly more accurate way to determine primer specificity [287], due to different binding affinities of individual nucleotide pairs [290].

5.3.3 The importance of accessibility of statistical analysis tools

It is important for cutting edge bioinformatics to be accessible to biologists, as even the best tools will not have an impact if they are not adopted. The *DAVID* Bioinformatics Resources [291], for example, are often used for GSEA from RNA-Seq data, despite the existence of superior programmatical tools, such as *Camera* and *ROAST* [204, 292], that may be used instead. *DAVID* is also not often updated, however, it remains a popular tool due to the interactivity of the interface and the fact that it is simple to use. It is therefore clearly essential that basic statistical techniques and programming must be taught to all new researchers, to improve the general quality of statistical analysis in biology. It is also evident that interactive tools may be useful instruments when filling the current gap that exists between biologists and the use of best practice bioinformatics methods, but should not replace adequate training.

Chapter 6: Conclusion

This thesis primarily investigated APA in breast cancer and used APA events to add additional prognostic power for the prediction of breast cancer survival. Also described were computational tools for the exploration of PAT-Seq datasets and the design of primers for 3' RACE experiments. The results of this thesis support the current dogma that proximal APA events predominate in proliferative cell types, but do not support the idea that it is only further proximal APA that contributes to the metastatic potential of a primary tumour. Rather, it is the specific APA profile of an individual tumour, including both proximal and distal APA events, that contributes to breast cancer metastatic ability.

In a controlled, mouse xenograft model of increasing breast cancer metastatic potential, depending on the metastatic cell line under investigation, there was a switching to both predominantly proximal or predominantly distal APA. Increases in the gene expression of RNA processing genes were also observed alongside APA. Poly(A) tail length could not be associated with metastasis, however the methods currently available for poly(A) tail length measurement may have been insufficient to determine these changes. It was also concluded that cell lines may be inappropriate for the study of proliferation-associated APA events, but may still prove useful in the of study APA for specific metastatic tropisms.

APA was also studied in primary human tumours after downloading, and inferring APA from, data from the TCGA and GEO. This resulted in the largest known collection of breast cancer-associated APA events produced to date, including 100 high confidence events that present highly attractive targets for further study as breast cancer biomarkers, or even potentially targetable genes. Through the reanalysis of these databases of publicly available primary tumour data, in the form of RNA-Seq and microarray datasets, it was once again concluded that APA is associated with breast cancer metastasis and occurs in combination with changes to mRNA processing genes, including splicing and APA factors. Furthermore, rather than a general trend, individual APA events were far more likely to be associated with breast cancer outcome. It has also been demonstrated that APA and clinical data are a better predictor of breast cancer outcome than the combination of gene expression and clinical data. A model was suggested to predict the outcome of breast cancer, using a combination of both clinical and APA data. In future, the RNA-Systems biology laboratory aims to design a targeted APA + clinical data based prognostic test that will add significant novel information about likely breast cancer outcome. This test will primarily be targeted toward the TNBC subtype, for which no clinical test currently exists.

When considering the 'rules' governing APA in breast cancer, it was shown that there can be more pronounced proximal and distal APA in both tumour and normal cell types (relative to an average normal breast tissue sample) at the same time and that this 'APA effect' increases with the expression of APA processing genes. It was shown that there is more APA in TNBC tumours generally but that these events did not differ significantly from other tumour subtypes. An APA-based predictor was shown to be able to significantly predict breast cancer outcome regardless of subtype. The lack of alignment of APA regulation with gene expression subtypes observed in this study shows that APA regulation may not be as closely related to gene expression as has been previously suggested and should be studied further as its own entity. This result was also supported by the expression of miRNA molecules, known to bind primarily within the 3' UTR, which also did not follow a subtype-specific pattern of expression.

Throughout the course of this study, two new bioinformatics tools were generated. The first was the *RSER* which enables novel analysis of PAT-seq data. The *RSER* web app also includes novel quality control and poly(A) tail length visualisation tools, that greatly enhance the ability of a researcher to explore 'PAT' type data. Also presented in this work is *3Primer*, a primer design tool for designing highly specific primers for 3’ RACE experiments. This tool has already proven invaluable in generating specific primers for targeted 3’ sequencing experiments presented here.

References

[1] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[2] Carol E DeSantis, Jiemin Ma, Ann Goding Sauer, Lisa A Newman, and Ahmedin Jemal. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: a cancer journal for clinicians*, 67(6):439–448, 2017.

[3] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5), 2015.

[4] Carol DeSantis, Jiemin Ma, Leah Bryan, and Ahmedin Jemal. Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, 64(1):52–62, 2014.

[5] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.

[6] Adetunji T Toriola and Graham A Colditz. Trends in breast cancer incidence and mortality in the United States: implications for prevention. *Breast cancer research and treatment*, 138(3):665–673, 2013.

[7] Adedayo A Onitilo, Jessica M Engel, Robert T Greenlee, and Bickol N Mukesh. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. *Clinical medicine & research*, 7(1-2):4–13, 2009.

[8] Paul L Nguyen, Alphonse G Taghian, Matthew S Katz, Andrzej Niemierko, Rita F Abi Raad, Whitney L Boon, Jennifer R Bellon, Julia S Wong, Barbara L Smith, and Jay R Harris. Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy. *Journal of clinical oncology*, 26(14):2373–2378, 2008.

[9] Carol A Parise, Katrina R Bauer, Monica M Brown, and Vincent Caggiano. Breast cancer subtypes as defined by the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) among women with invasive breast cancer in California, 1999–2004. *The breast journal*, 15(6):593–602, 2009.

[10] Nadia Howlader, Sean F Altekruse, Christopher I Li, Vivien W Chen, Christina A Clarke, Lynn AG Ries, and Kathleen A Cronin. US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *JNCI: Journal of the National Cancer Institute*, 106(5), 2014.

[11] Dennis J Slamon, William Godolphin, Lovell A Jones, John A Holt, Steven G Wong, Duane E Keith, Wendy J Levin, Susan G Stuart, Judy Udove, Axel Ullrich, et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, 244(4905):707–712, 1989.

[12] Giuseppe Curigliano, Harold J Burstein, E P Winer, Michael Gnant, Peter Dubsky, Sibylle Loibl, Marco Colleoni, Meredith M Regan, M Piccart-Gebhart, H-J Senn, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. *Annals of Oncology*, 28(8):1700–1712, 2017.

[13] Antoni Hurtado, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt, and Jason S Carroll. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature genetics*, 43(1):27, 2011.

[14] William A Knight, Robert B Livingston, Ernest J Gregory, and William L McGuire. Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. *Cancer research*, 37(12):4669–4671, 1977.

[15] G Valabrega, F Montemurro, and M Aglietta. Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. *Annals of oncology*, 18(6):977–984, 2007.

[16] Early Breast Cancer Trialists' Collaborative Group et al. Tamoxifen for early breast cancer: an overview of the randomised trials. *The Lancet*, 351(9114):1451–1467, 1998.

[17] Charles L Vogel, Melody A Cobleigh, Debu Tripathy, John C Gutheil, Lyndsay N Harris, Louis Fehrenbacher, Dennis J Slamon, Maureen Murphy, William F Novotny, Michael Burchmore, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *Journal of clinical oncology*, 20(3):719–726, 2002.

[18] KB Horwitz and WL McGuire. Specific progesterone receptors in human breast cancer. *Steroids*, 25(4):497–505, 1975.

[19] Xiaojiang Cui, Rachel Schiff, Grazia Arpino, C Kent Osborne, and Adrian V Lee. Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy. *Journal of clinical oncology*, 23(30):7721–7735, 2005.

[20] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[21] Hiroko Masuda, Keith A Baggerly, Ying Wang, Ya Zhang, Ana Maria Gonzalez-Angulo, Funda Meric-Bernstam, Vicente Valero, Brian D Lehmann, Jennifer A Pietenpol, Gabriel N Hortobagyi, et al. Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clinical cancer research*, 19(19):5533–5540, 2013.

[22] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750–2767, 2011.

[23] Anna V Ivshina, Joshy George, Oleg Senko, Benjamin Mow, Thomas C Putti, Johanna Smeds, Thomas Lindahl, Yudi Pawitan, Per Hall, Hans Nordgren, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, 66(21):10292–10301, 2006.

[24] Maggie CU Cheang, David Voduc, Chris Bajdik, Samuel Leung, Steven McKinney, Stephen K Chia, Charles M Perou, and Torsten O Nielsen. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clinical cancer research*, 14(5):1368–1376, 2008.

[25] Rebecca Siegel, Elizabeth Ward, Otis Brawley, and Ahmedin Jemal. Cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(4):212–236, 2011.

[26] Lisa A Carey, E Claire Dees, Lynda Sawyer, Lisa Gatti, Dominic T Moore, Frances Collichio, David W Ollila, Carolyn I Sartor, Mark L Graham, and Charles M Perou. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clinical cancer research*, 13(8):2329–2334, 2007.

[27] Idil Cetin and Mehmet Topcul. Triple negative breast cancer. *Asian Pac J Cancer Prev*, 15(6):2427–2431, 2014.

[28] Susan Cleator, Wolfgang Heller, and R Charles Coombes. Triple-negative breast cancer: therapeutic options. *The lancet oncology*, 8(3):235–244, 2007.

[29] Cornelia Liedtke, Chafika Mazouni, Kenneth R Hess, Fabrice André, Attila Tordai, Jaime A Mejia, W Fraser Symmans, Ana M Gonzalez-Angulo, Bryan Hennessy, Marjorie

Green, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of clinical oncology*, 26(8):1275–1281, 2008.

[30] Kent W Hunter, Nigel PS Crawford, and Jude Alsarraj. Mechanisms of metastasis. *Breast Cancer Research*, 10(1):S2, 2008.

[31] Katarina Wolf, Irina Mazo, Harry Leung, Katharina Engelke, Ulrich H Von Andrian, Elena I Deryugina, Alex Y Strongin, Eva-B Bröcker, and Peter Friedl. Compensation mechanism in tumor cell migration: mesenchymal–amoeboid transition after blocking of pericellular proteolysis. *The Journal of cell biology*, 160(2):267–277, 2003.

[32] Raghu Kalluri and Robert A Weinberg. The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation*, 119(6):1420–1428, 2009.

[33] Isaiah J Fidler. The pathogenesis of cancer metastasis: the seed and soil hypothesis revisited. *Nature Reviews Cancer*, 3(6):453–458, 2003.

[34] Amye J Tevaarwerk, Robert J Gray, Bryan P Schneider, Mary Lou Smith, Lynne I Wagner, John H Fetting, Nancy Davidson, Lori J Goldstein, Kathy D Miller, and Joseph A Sparano. Survival in patients with metastatic recurrent breast cancer after adjuvant chemotherapy. *Cancer*, 119(6):1140–1148, 2013.

[35] Jeremy Bastid. EMT in carcinoma progression and dissemination: facts, unanswered questions, and clinical considerations. *Cancer and Metastasis Reviews*, 31(1-2):277–283, 2012.

[36] Lucy R Yates, Stian Knappskog, David Wedge, James HR Farmery, Santiago Gonzalez, Inigo Martincorena, Ludmil B Alexandrov, Peter Van Loo, Hans Kristian Haugland, Peer Kaare Lilleng, et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer cell*, 32(2):169–184, 2017.

[37] Olivia Jane Scully, Boon-Huat Bay, George Yip, and Yingnan Yu. Breast cancer metastasis. *Cancer Genomics-Proteomics*, 9(5):311–320, 2012.

[38] Dong-Mei Li and Yu-Mei Feng. Signaling mechanism of cell adhesion molecules in breast cancer metastasis: potential therapeutic targets. *Breast cancer research and treatment*, 128(1):7, 2011.

[39] Kari R Fischer, Anna Durrans, Sharrell Lee, Jianting Sheng, Fuhai Li, Stephen TC Wong, Hyejin Choi, Tina El Rayes, Seongho Ryu, Juliane Troeger, et al. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature*, 527(7579):472, 2015.

[40] Eva Tomaskovic-Crook, Erik W Thompson, and Jean Paul Thiery. Epithelial to mesenchymal transition and breast cancer. *Breast cancer research*, 11(6):213, 2009.

[41] Anna Labernadie, Takuya Kato, Agust Brugués, Xavier Serra-Picamal, Stefanie Derzsi, Esther Arwert, Anne Weston, Victor González-Tarragó, Alberto Elosegui-Artola, Lorenzo Albertazzi, et al. A mechanically active heterotypic E-cadherin/N-cadherin adhesion enables fibroblasts to drive cancer cell invasion. *Nature cell biology*, 19(3):224, 2017.

[42] Frans Van Roy. Beyond E-cadherin: roles of other cadherin superfamily members in cancer. *Nature Reviews Cancer*, 14(2):121, 2014.

[43] Xin Ye, Thomas Brabletz, Yibin Kang, Gregory D Longmore, M Angela Nieto, Ben Z Stanger, Jing Yang, and Robert A Weinberg. Upholding a role for EMT in breast cancer metastasis. *Nature*, 547(7661):E1, 2017.

[44] Arnold J Levine, Jamil Momand, and Cathy A Finlay. The p53 tumour suppressor gene. *Nature*, 351(6326):453, 1991.

[45] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167, 2009.

[46] Peter Savas, Zhi Ling Teo, Christophe Lefevre, Christoffer Flensburg, Franco Caramia, Kathryn Alsop, Mariam Mansour, Prudence A Francis, Heather A Thorne, Maria Joao Silva, et al. The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program CASCADE. *PLoS medicine*, 13(12):e1002204, 2016.

[47] Amparo Cano, Mirna A Pérez-Moreno, Isabel Rodrigo, Annamaria Locascio, Mará J Blanco, Marta G del Barrio, Francisco Portillo, and M Angela Nieto. The transcription factor snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. *Nature cell biology*, 2(2):76–83, 2000.

[48] Joke Comijn, Geert Berx, Petra Vermassen, Kristin Verschueren, Leo van Grunsven, Erik Bruyneel, Marc Mareel, Danny Huylebroeck, and Frans Van Roy. The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion. *Molecular cell*, 7(6):1267–1278, 2001.

[49] Victoria Bolós, Hector Peinado, Mirna A Pérez-Moreno, Mario F Fraga, Manel Esteller, and Amparo Cano. The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with Snail and E47 repressors. *Journal of cell science*, 116(3):499–511, 2003.

[50] Jing Yang, Sendurai A Mani, Joana Liu Donaher, Sridhar Ramaswamy, Raphael A Itzykson, Christophe Come, Pierre Savagner, Inna Gitelman, Andrea Richardson, and Robert A Weinberg. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *cell*, 117(7):927–939, 2004.

[51] I Vlodavsky, A Eldor, A Haimovitz-Friedman, Y Matzner, R Ishai-Michaeli, O Lider, Y Naparstek, IR Cohen, and Z Fuks. Expression of heparanase by platelets and circulating cells of the immune system: possible involvement in diapedesis and extravasation. *Invasion & metastasis*, 12(2):112–127, 1992.

[52] Xu Sun, Ganlin Zhang, Jiayun Nian, Mingwei Yu, Shijian Chen, Yi Zhang, Guowang Yang, Lin Yang, Peiyu Cheng, Chen Yan, et al. Elevated heparanase expression is associated with poor prognosis in breast cancer: a study based on systematic review and TCGA data. *Oncotarget*, 8(26):43521, 2017.

[53] Simon Wilkinson, Hugh F Paterson, and Christopher J Marshall. Cdc42–MRCK and Rho–ROCK signalling cooperate in myosin phosphorylation and cell invasion. *Nature cell biology*, 7(3):255, 2005.

[54] Erik Sahai and Christopher J Marshall. Differing modes of tumour cell invasion have distinct requirements for Rho/ROCK signalling and extracellular proteolysis. *Nature cell biology*, 5(8):711, 2003.

[55] Molly E Klein, David J Dabbs, Yongli Shuai, Adam M Brufsky, Rachel Jankowitz, Shannon L Puhalla, and Rohit Bhargava. Prediction of the Oncotype DX recurrence score: use of pathology-generated equations derived by linear regression analysis. *Modern Pathology*, 2013.

[56] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[57] John PA Ioannidis. Microarrays and molecular research: noise discovery? *The Lancet*, 365(9458):454–455, 2005.

[58] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210, 2011.

[59] Liat Ein-Dor, Or Zuk, and Eytan Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923–5928, 2006.

[60] Fabien Reyal, Martin H Van Vliet, Nicola J Armstrong, Hugo M Horlings, Karin E de Visser, Marlen Kok, Andrew E Teschendorff, Stella Mook, Laura van't Veer, Carlos Caldas, et al. A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Research*, 10(6):R93, 2008.

[61] Ronglai Shen, Arul M Chinnaiyan, and Debashis Ghosh. Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC medical genomics*, 1(1):28, 2008.

[62] Yotam Drier and Eytan Domany. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PloS one*, 6(3):e17795, 2011.

[63] Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frédéric Schütz, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):R65, 2008.

[64] Li-Xu Yan, Xiu-Fang Huang, Qiong Shao, MA-Yan Huang, Ling Deng, Qiu-Liang Wu, Yi-Xin Zeng, and Jian-Yong Shao. MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *Rna*, 14(11):2348–2360, 2008.

[65] Lei Wang, Xin Hu, Peng Wang, and Zhi-Ming Shao. The 3′ UTR signature defines a highly metastatic subgroup of triple-negative breast cancer. *Oncotarget*, 7(37):59834, 2016.

[66] Zheng Xia, Lawrence A Donehower, Thomas A Cooper, Joel R Neilson, David A Wheeler, Eric J Wagner, and Wei Li. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nature communications*, 5, 2014.

[67] Elizabeth D Hay. An overview of epithelio-mesenchymal transformation. *Cells Tissues Organs*, 154(1):8–20, 1995.

[68] Alexander O Subtelny, Stephen W Eichhorn, Grace R Chen, Hazel Sive, and David P Bartel. Poly (A)-tail profiling reveals an embryonic switch in translational control. *Nature*, 508(7494):66–71, 2014.

[69] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

[70] Marcus Gry, Rebecca Rimini, Sara Strömberg, Anna Asplund, Fredrik Pontén, Mathias Uhlén, and Peter Nilsson. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC genomics*, 10(1):365, 2009.

[71] Jing-Ping Hsin and James L Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & development*, 26(19):2119–2137, 2012.

[72] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251, 1997.

[73] Swaminathan Venkatesh and Jerry L Workman. Histone exchange, chromatin structure and the regulation of transcription. *Nature reviews Molecular cell biology*, 16(3):178, 2015.

[74] Richard Breathnach and Pierre Chambon. Organization and expression of eucaryotic split genes coding for proteins. *Annual review of biochemistry*, 50(1):349–383, 1981.

[75] Horace R Drew and Andrew A Travers. DNA bending and its relation to nucleosome positioning. *Journal of molecular biology*, 186(4):773–790, 1985.

[76] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K Moore, Ji-Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772, 2006.

[77] Paul D Hartley and Hiten D Madhani. Mechanisms that specify promoter nucleosome location and identity. *Cell*, 137(3):445–458, 2009.

[78] Sebastian Grünberg and Steven Hahn. Structural insights into transcription initiation by RNA polymerase II. *Trends in biochemical sciences*, 38(12):603–611, 2013.

[79] Sarah Sainsbury, Carrie Bernecky, and Patrick Cramer. Structural basis of transcription initiation by RNA polymerase II. *Nature reviews Molecular cell biology*, 16(3):129, 2015.

[80] Charles Giardina and John T Lis. DNA melting on yeast RNA polymerase II promoters. *Science*, 261(5122):759–762, 1993.

[81] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[82] Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290, 2013.

[83] Hyejung Won, Luis de La Torre-Ubieta, Jason L Stein, Neelroop N Parikshak, Jerry Huang, Carli K Opland, Michael J Gandal, Gavin J Sutton, Farhad Hormozdiari, Daning Lu, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, 538(7626):523, 2016.

[84] Tom Sexton and Giacomo Cavalli. The role of chromosome domains in shaping the functional genome. *Cell*, 160(6):1049–1059, 2015.

[85] Markus C Wahl, Cindy L Will, and Reinhard Lührmann. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, 2009.

[86] Anna Hegele, Atanas Kamburov, Arndt Grossmann, Chrysovalantis Sourlis, Sylvia Wowro, Mareike Weimann, Cindy L Will, Vlad Pena, Reinhard Lührmann, and Ulrich Stelzl. Dynamic protein-protein interaction wiring of the human spliceosome. *Molecular cell*, 45(4):567–580, 2012.

[87] Marieta Gencheva, Mitsuo Kato, Alain NS Newo, and Ren-Jang Lin. Contribution of DEAH-box protein DHX16 in human pre-mRNA splicing. *Biochemical Journal*, 429(1):25–32, 2010.

[88] Zhaolan Zhou and Robin Reed. Human homologs of yeast Prp16 and Prp17 reveal conservation of the mechanism for catalytic step II of pre-mRNA splicing. *The EMBO Journal*, 17(7):2095–2106, 1998.

[89] Zbigniew Warkocki, Peter Odenwälder, Jana Schmitzová, Florian Platzmann, Holger Stark, Henning Urlaub, Ralf Ficner, Patrizia Fabrizio, and Reinhard Lührmann. Reconstitution of both steps of Saccharomyces cerevisiae splicing with purified spliceosomal components. *Nature Structural and Molecular Biology*, 16(12):1237, 2009.

[90] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345, 2010.

[91] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nature genetics*, 30(1):13, 2002.

[92] John A Calarco, Simone Superina, Dave O'Hanlon, Mathieu Gabut, Bushra Raj, Qun Pan, Ursula Skalska, Laura Clarke, Danielle Gelinas, Derek van der Kooy, et al. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell*, 138(5):898–910, 2009.

[93] Elias G Bechara, Endre Sebestyén, Isabella Bernardis, Eduardo Eyras, and Juan Valcárcel. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Molecular cell*, 52(5):720–733, 2013.

[94] Tiffany Y-T Hsu, Lukas M Simon, Nicholas J Neill, Richard Marcotte, Azin Sayad, Christopher S Bland, Gloria V Echeverria, Tingting Sun, Sarah J Kurley, Siddhartha Tyagi, et al. The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature*, 525(7569):384, 2015.

[95] Rotem Karni, Elisa de Stanchina, Scott W Lowe, Rahul Sinha, David Mu, and Adrian R Krainer. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural and Molecular Biology*, 14(3):185, 2007.

[96] Olga Anczuków, Martin Akerman, Antoine Cléry, Jie Wu, Chen Shen, Nitin H Shirole, Amanda Raimer, Shuying Sun, Mads A Jensen, Yimin Hua, et al. SRSF1-regulated alternative splicing in breast cancer. *Molecular cell*, 60(1):105–117, 2015.

[97] Tom Maniatis and Robin Reed. An extensive network of coupling among gene expression machines. *Nature*, 416(6880):499–506, 2002.

[98] Hiroyuki Fuke and Mutsuhito Ohno. Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic acids research*, 36(3):1037–1049, 2007.

[99] Emmanuel Beaudoing, Susan Freier, Jacqueline R Wyatt, Jean-Michel Claverie, and Daniel Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome research*, 10(7):1001–1010, 2000.

[100] JUN Hu, Carol S Lutz, Jeffrey Wilusz, and BIN Tian. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *Rna*, 11(10):1485–1493, 2005.

[101] Anna Gil and Nicholas J Proudfoot. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β-globin mRNA 3′ end formation. *Cell*, 49(3):399–406, 1987.

[102] Chengguo Yao, Jacob Biesinger, Ji Wan, Lingjie Weng, Yi Xing, Xiaohui Xie, and Yongsheng Shi. Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation. *Proceedings of the National Academy of Sciences*, 109(46):18773–18778, 2012.

[103] Yoshio Takagaki, Lisa C Ryner, and James L Manley. Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes & development*, 3(11):1711–1724, 1989.

[104] Corey R Mandel, Yun Bai, and Liang Tong. Protein factors in pre-mRNA 3′-end processing. *Cellular and Molecular Life Sciences*, 65(7-8):1099–1122, 2008.

[105] Diana F Colgan and James L Manley. Mechanism and regulation of mRNA polyadenylation. *Genes & development*, 11(21):2755–2766, 1997.

[106] Adnan Derti, Philip Garrett-Engele, Kenzie D MacIsaac, Richard C Stevens, Shreedharan Sriram, Ronghua Chen, Carol A Rohl, Jason M Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome research*, 22(6):1173–1183, 2012.

[107] Steve Lianoglou, Vidur Garg, Julie L Yang, Christina S Leslie, and Christine Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development*, 27(21):2380–2396, 2013.

[108] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105, 2007.

[109] Christine Mayr and David P Bartel. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, 2009.

[110] Wayne O Miles, Antonio Lembo, Angela Volorio, Elena Brachtel, Bin Tian, Dennis Sgroi, Paolo Provero, and Nicholas Dyson. Alternative Polyadenylation in Triple-Negative Breast Tumors Allows NRAS and c-JUN to Bypass PUMILIO Posttranscriptional Regulation. *Cancer Research*, 76(24):7231–7241, 2016.

[111] Christine Mayr. Evolution and biological roles of alternative 3′ UTRs. *Trends in cell biology*, 26(3):227–237, 2016.

[112] Rickard Sandberg, Joel R Neilson, Arup Sarma, Phillip A Sharp, and Christopher B Burge. Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–1647, 2008.

[113] Hesna Begum Akman, Merve Oyken, Taner Tuncer, Tolga Can, and Ayse Elif Erson-Bensan. 3'UTR Shortening and EGF signaling: Implications for breast cancer. *Human molecular genetics*, page ddv391, 2015.

[114] Jonathan LE Dean, Gareth Sully, Andrew R Clark, and Jeremy Saklatvala. The involvement of AU-rich element-binding proteins in p38 mitogen-activated protein kinase pathway-mediated mRNA stabilisation. *Cellular signalling*, 16(10):1113–1121, 2004.

[115] Laure Weill, Eulàlia Belloc, Felice-Alessio Bava, and Raúl Méndez. Translational control by changes in poly (A) tail length: recycling mRNAs. *Nature structural & molecular biology*, 19(6):577–585, 2012.

[116] Dafne Campigli Di Giammartino, Kensei Nishida, and James L Manley. Mechanisms and consequences of alternative polyadenylation. *Molecular cell*, 43(6):853–866, 2011.

[117] Hyeshik Chang, Jaechul Lim, Minju Ha, and V Narry Kim. TAIL-seq: genome-wide determination of poly (A) tail length and 3′ end modifications. *Molecular cell*, 53(6):1044–1052, 2014.

[118] Tsuyoshi Udagawa, Sharon A Swanger, Koichi Takeuchi, Jong Heon Kim, Vijayalaxmi Nalavadi, Jihae Shin, Lori J Lorenz, R Suzanne Zukin, Gary J Bassell, and Joel D Richter. Bidirectional control of mRNA translation and synaptic plasticity by the cytoplasmic polyadenylation complex. *Molecular cell*, 47(2):253–266, 2012.

[119] Isabel Novoa, Javier Gallego, Pedro G Ferreira, and Raul Mendez. Mitotic cell-cycle progression is regulated by CPEB1 and CPEB4-dependent translational control. *Nature cell biology*, 12(5):447–456, 2010.

[120] Elena Ortiz-Zapater, David Pineda, Neus Martnez-Bosch, Gonzalo Fernández-Miranda, Mar Iglesias, Francesc Alameda, Mireia Moreno, Carolina Eliscovich, Eduardo Eyras, Francisco X Real, et al. Key contribution of CPEB4-mediated translational control to cancer progression. *Nature medicine*, 18(1):83–90, 2012.

[121] Aimee L Jalkanen, Stephen J Coleman, and Jeffrey Wilusz. Determinants and implications of mRNA poly (A) tail size–Does this protein make my tail look big? In *Seminars in cell & developmental biology*, volume 34, pages 24–32. Elsevier, 2014.

[122] Luciano H Apponi, Sara W Leung, Kathryn R Williams, Sandro R Valentini, Anita H Corbett, and Grace K Pavlath. Loss of nuclear poly (A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. *Human molecular genetics*, 19(6):1058–1065, 2009.

[123] G Renuka Kumar and Britt A Glaunsinger. Nuclear import of cytoplasmic poly (A) binding protein restricts gene expression via hyperadenylation and nuclear retention of mRNA. *Molecular and cellular biology*, 30(21):4996–5008, 2010.

[124] Stefan M Bresson and Nicholas K Conrad. The human nuclear poly (a)-binding protein promotes RNA hyperadenylation and decay. *PLoS genetics*, 9(10):e1003893, 2013.

[125] Traude H Beilharz and Thomas Preiss. Widespread use of poly (A) tail length control to accentuate expression of the yeast transcriptome. *Rna*, 13(7):982–997, 2007.

[126] Daniel H Lackner, Traude H Beilharz, Samuel Marguerat, Juan Mata, Stephen Watt, Falk Schubert, Thomas Preiss, and Jürg Bähler. A network of multiple regulatory layers shapes gene expression in fission yeast. *Molecular cell*, 26(1):145–155, 2007.

[127] Sarah Azoubel Lima, Laura B Chipman, Angela L Nicholson, Ying-Hsin Chen, Brian A Yee, Gene W Yeo, Jeff Coller, and Amy E Pasquinelli. Short poly (A) tails are a conserved feature of highly expressed genes. *Nature structural & molecular biology*, 2017.

[128] Shimyn Slomovic, Ella Fremder, Raymond HG Staals, Ger JM Pruijn, and Gadi Schuster. Addition of poly (A) and poly (A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proceedings of the National Academy of Sciences*, 107(16):7407–7412, 2010.

[129] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

[130] Bruce Wightman, Ilho Ha, and Gary Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862, 1993.

[131] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell*, 120(1):15–20, 2005.

[132] György Hutvágner, Juanita McLachlan, Amy E Pasquinelli, Éva Bálint, Thomas Tuschl, and Phillip D Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838, 2001.

[133] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105, 2009.

[134] David P Bartel. MicroRNAs: target recognition and regulatory functions. *cell*, 136(2):215–233, 2009.

[135] Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology*, 15(8):509–524, 2014.

[136] Antonio Lembo, Ferdinando Di Cunto, and Paolo Provero. Shortening of 3′ UTRs correlates with poor prognosis in breast and lung cancer. *PloS one*, 7(2):e31129, 2012.

[137] Maria Paola Paronetto, Ilaria Passacantilli, and Claudio Sette. Alternative splicing and cell survival: from tissue homeostasis to disease. *Cell Death & Differentiation*, 2016.

[138] Mark P Ashe, Philip Griffin, William James, and Nick J Proudfoot. Poly (A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site. *Genes & development*, 9(23):3008–3025, 1995.

[139] Gregory A Sowd, Erik Serrao, Hao Wang, Weifeng Wang, Hind J Fadel, Eric M Poeschla, and Alan N Engelman. A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proceedings of the National Academy of Sciences*, 113(8):E1054–E1063, 2016.

[140] John R Dickson, Carla Kruse, Daniel R Montagna, Bente Finsen, and Michael S Wolfe. Alternative polyadenylation and miR-34 family members regulate tau expression. *Journal of neurochemistry*, 127(6):739–749, 2013.

[141] Esther E Creemers, Amira Cholid Bawazeer, Alejandro P Ugalde, Hanneke WM van Deutekom, Ingeborg van der Made, Nina E de Groot, Michiel E Adriaens, Stuart Cook, Connie Bezzina, Norbert Hübner, et al. Genome-wide polyadenylation maps reveal dynamic

mRNA 3’-end formation in the failing human heart. *Circulation research*, pages CIRCRESAHA–115, 2015.

[142] Shuang Tang, Amita Patel, and Philip R Krause. Herpes simplex virus ICP27 regulates alternative pre-mRNA polyadenylation and splicing in a sequence-dependent manner. *Proceedings of the National Academy of Sciences*, page 201609695, 2016.

[143] Chioniso P Masamha, Zheng Xia, Jingxuan Yang, Todd R Albrecht, Min Li, Ann-Bin Shyu, Wei Li, and Eric J Wagner. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, 510(7505):412, 2014.

[144] Ranjan Batra, Konstantinos Charizanis, Mini Manchanda, Apoorva Mohan, Moyi Li, Dustin J Finn, Marianne Goodwin, Chaolin Zhang, Krzysztof Sobczak, Charles A Thornton, et al. Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Molecular cell*, 56(2):311–322, 2014.

[145] Herve Rhinn, Liang Qiang, Toru Yamashita, David Rhee, Ari Zolin, William Vanti, and Asa Abeliovich. α-Synuclein transcript alternative 3′ UTR usage as a convergent mechanism in Parkinsons disease pathology. *Nature communications*, 3:1084, 2012.

[146] Li Li, Duolin Wang, Mengzhu Xue, Xianqiang Mi, Yanchun Liang, and Peng Wang. 3′ UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network. *Scientific reports*, 4:5406, 2014.

[147] Michael D Barnhart, Stephanie L Moon, Alexander W Emch, Carol J Wilusz, and Jeffrey Wilusz. Changes in cellular mRNA stability, splicing, and polyadenylation through HuR protein sequestration by a cytoplasmic RNA virus. *Cell reports*, 5(4):909–917, 2013.

[148] Jianming Qiu, Ramnath Nayak, and David J Pintel. Alternative polyadenylation of adeno-associated virus type 5 RNA within an internal intron is governed by both a downstream element within the intron 3′ splice acceptor and an element upstream of the P41 initiation site. *Journal of virology*, 78(1):83–93, 2004.

[149] Xin Jia, Shaochun Yuan, Yao Wang, Yonggui Fu, Yong Ge, Yutong Ge, Xihong Lan, Yuchao Feng, Feifei Qiu, Peiyi Li, et al. The role of alternative polyadenylation in the antiviral innate immune response. *Nature Communications*, 8, 2017.

[150] Christoph Münch, Birgit Schwalenstöcker, Christine Hermann, Stanko Cirovic, Stefan Stamm, Albert Ludolph, and Thomas Meyer. Differential RNA cleavage and polyadenylation of the glutamate transporter EAAT2 in the human brain. *Molecular brain research*, 80(2):244–251, 2000.

[151] Jan Lewerenz and Pamela Maher. Chronic glutamate toxicity in neurodegenerative diseases—what is the evidence? *Frontiers in neuroscience*, 9, 2015.

[152] Pedro Miura, Sol Shenker, Celia Andreu-Agullo, Jakub O Westholm, and Eric C Lai. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome research*, 23(5):812–825, 2013.

[153] I Alafuzoff, K Iqbal, H Friden, Rolf Adolfsson, and B Winblad. Histopathological criteria for progressive dementia disorders: clinical-pathological correlation and classification by multivariate data analysis. *Acta neuropathologica*, 74(3):209–225, 1987.

[154] Vincenzo A Gennarino, Callison E Alcott, Chun-An Chen, Arindam Chaudhury, Madelyn A Gillentine, Jill A Rosenfeld, Sumit Parikh, James W Wheless, Elizabeth R Roeder, Dafne DG Horovitz, et al. NUDT21-spanning CNVs lead to neuropsychiatric disease and altered MeCP2 abundance via alternative polyadenylation. *Elife*, 4:e10782, 2015.

[155] Maria Grazia Spillantini, Marie Luise Schmidt, Virginia M-Y Lee, John Q Trojanowski, Ross Jakes, and Michel Goedert. α-Synuclein in Lewy bodies. *Nature*, 388(6645):839–840, 1997.

[156] Rachel Flomen and Andrew Makoff. Increased RNA editing in EAAT2 pre-mRNA from amyotrophic lateral sclerosis patients: involvement of a cryptic polyadenylation site. *Neuroscience letters*, 497(2):139–143, 2011.

[157] Ji Yeon Park, Wencheng Li, Dinghai Zheng, Peiyong Zhai, Yun Zhao, Takahisa Matsuda, Stephen F Vatner, Junichi Sadoshima, and Bin Tian. Comparative analysis of mRNA isoform expression in cardiac hypertrophy and development reveals multiple post-transcriptional regulatory modules. *PLoS One*, 6(7):e22391, 2011.

[158] Rina Soetanto, Carly J Hynes, Hardip R Patel, David T Humphreys, Maurits Evers, Guowen Duan, Brian J Parker, Stuart K Archer, Jennifer L Clancy, Robert M Graham, et al. Role of miRNAs and alternative mRNA 3′-end cleavage and polyadenylation of their mRNA targets in cardiomyocyte hypertrophy. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(5):744–756, 2016.

[159] Michael Tranter, Robert N Helsley, Waltke R Paulding, Michael McGuinness, Cole Brokamp, Lauren Haar, Yong Liu, Xiaoping Ren, and W Keith Jones. Coordinated post-transcriptional regulation of Hsp70. 3 gene expression by microRNA and alternative polyadenylation. *Journal of Biological Chemistry*, 286(34):29828–29837, 2011.

[160] Andrea N Ladd. New Insights Into the Role of RNA-Binding Proteins in the Regulation of Heart Development. *International review of cell and molecular biology*, 324:125–185, 2016.

[161] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[162] Andreas R Gruber, Georges Martin, Philipp Müller, Alexander Schmidt, Andreas J Gruber, Rafal Gumienny, Nitish Mittal, Rajesh Jayachandran, Jean Pieters, Walter Keller, et al. Global 3′ UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nature communications*, 5, 2014.

[163] Mathias Jenal, Ran Elkon, Fabricio Loayza-Puch, Gijs van Haaften, Uwe Kühn, Fiona M Menzies, Joachim AF Oude Vrielink, Arnold J Bos, Jarno Drost, Koos Rooijers, et al.

The poly (A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149(3):538–553, 2012.

[164] Georges Martin, Andreas R Gruber, Walter Keller, and Mihaela Zavolan. Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell reports*, 1(6):753–763, 2012.

[165] Alan S Coates, Eric P Winer, Aron Goldhirsch, Richard D Gelber, Michael Gnant, M Piccart-Gebhart, Beat Thürlimann, H-J Senn, Panel Members, Fabrice André, et al. Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Annals of oncology*, 26(8):1533–1546, 2015.

[166] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[167] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart, and Christos Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical cancer research*, 14(16):5158–5165, 2008.

[168] Yonggui Fu, Yu Sun, Yuxin Li, Jie Li, Xingqiang Rao, Chong Chen, and Anlong Xu. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome research*, 21(5):741–747, 2011.

[169] Paul F Harrison, David R Powell, Jennifer L Clancy, Thomas Preiss, Peter R Boag, Ana Traven, Torsten Seemann, and Traude H Beilharz. PAT-seq: a method to study the integration of 3′-UTR dynamics with gene expression in the eukaryotic transcriptome. *RNA*, 21(8):1502–1510, 2015.

[170] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621, 2008.

[171] Sorin Draghici, Purvesh Khatri, Aron C Eklund, and Zoltan Szallasi. Reliability and reproducibility issues in DNA microarray measurements. *TRENDS in Genetics*, 22(2):101–109, 2006.

[172] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[173] David J Lockhart, Helin Dong, Michael C Byrne, Maximillian T Follettie, Michael V Gallo, Mark S Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Norton, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675, 1996.

[174] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[175] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, 2010.

[176] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[177] Dominic O’Neil, Heike Glowatz, and Martin Schlumpberger. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. *Current protocols in molecular biology*, pages 4–19, 2013.

[178] Christopher E Mason and Olivier Elemento. Faster sequencers, larger datasets, new challenges. *Genome biology*, 13(3):314, 2012.

[179] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[180] Peng Cui, Qiang Lin, Feng Ding, Chengqi Xin, Wei Gong, Lingfang Zhang, Jianing Geng, Bing Zhang, Xiaomin Yu, Jin Yang, et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, 96(5):259–265, 2010.

[181] Sören Müller, Lukas Rycak, Fabian Afonso-Grunz, Peter Winter, Adam M Zawada, Ewa Damrath, Jessica Scheider, Juliane Schmäh, Ina Koch, Günter Kahl, et al. APADB: a database for alternative polyadenylation and microRNA regulation events. *Database*, 2014:bau076, 2014.

[182] Adam M Zawada, Kyrill S Rogacev, Sören Müller, Björn Rotter, Peter Winter, Danilo Fliser, and Gunnar H Heine. Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease. *Epigenetics*, 9(1):161–172, 2014.

[183] Hani Goodarzi, Hamed S Najafabadi, Panos Oikonomou, Todd M Greco, Lisa Fish, Reza Salavati, Ileana M Cristea, and Saeed Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485(7397):264, 2012.

[184] Panos Oikonomou, Hani Goodarzi, and Saeed Tavazoie. Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. *Cell reports*, 7(1):281–292, 2014.

[185] Moe Yokoshi, Quan Li, Munetaka Yamamoto, Hitomi Okada, Yutaka Suzuki, and Yukio Kawahara. Direct binding of Ataxin-2 to distinct elements in 3′ UTRs promotes mRNA stability and protein expression. *Molecular cell*, 55(2):186–198, 2014.

[186] Ching-Yi Chen, Roberto Gherzi, Jens S Andersen, Guido Gaietta, Karsten Jürchott, Hans-Dieter Royer, Matthias Mann, and Michael Karin. Nucleolin and YB-1 are required for

JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. *Genes & development*, 14(10):1236–1248, 2000.

[187] Stella Aronov, Ruth Marx, and Irith Ginzburg. Identification of 3′ UTR region implicated in tau mRNA stabilization in neuronal cells. *Journal of Molecular Neuroscience*, 12(2):131–145, 1999.

[188] Pedro AB Pinto, Telmo Henriques, Marta O Freitas, Torcato Martins, Rita G Domingues, Paulina S Wyrzykowska, Paula A Coelho, Alexandre M Carmo, Claudio E Sunkel, Nicholas J Proudfoot, et al. RNA polymerase II kinetics in polo polyadenylation signal selection. *The EMBO journal*, 30(12):2431–2444, 2011.

[189] YoneJung Yoon, Morgan C McKenna, David A Rollins, Minseok Song, Tal Nuriel, Steven S Gross, Guoqiang Xu, and Charles E Glatt. Anxiety-associated alternative polyadenylation of the serotonin transporter mRNA confers translational regulation by hnRNPK. *Proceedings of the National Academy of Sciences*, 110(28):11624–11629, 2013.

[190] Nina S Levy, Sangmi Chung, Henry Furneaux, and Andrew P Levy. Hypoxic stabilization of vascular endothelial growth factor mRNA by the RNA-binding protein HuR. *Journal of Biological Chemistry*, 273(11):6417–6423, 1998.

[191] Rachael Emily Turner, Andrew David Pattison, and Traude Helene Beilharz. Alternative polyadenylation in the regulation and dysregulation of gene expression. In *Seminars in cell & developmental biology*, 2017.

[192] Carmela P Morales, Shawn E Holt, Michel Ouellette, Kiran J Kaur, Ying Yan, Kathleen S Wilson, Michael A White, Woodring E Wright, and Jerry W Shay. Absence of cancer–associated changes in human fibroblasts immortalized with telomerase. *Nature genetics*, 21(1):115, 1999.

[193] Mainul Hoque, Zhe Ji, Dinghai Zheng, Wenting Luo, Wencheng Li, Bei You, Ji Yeon Park, Ghassan Yehia, and Bin Tian. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nature methods*, 10(2):133, 2013.

[194] Binyamin D Berkovits and Christine Mayr. Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization. *Nature*, 522(7556):363–367, 2015.

[195] Cameron N Johnstone, Andrew D Pattison, Kylie L Gorringe, Paul F Harrison, David R Powell, Peter Lock, David Baloyan, M Ernst, Alastair G Stewart, Traude H Beilharz, et al. Functional and genomic characterization of a xenograft model system for the study of metastasis in triple-negative breast cancer. *Disease models & mechanisms*, pages dmm–032250, 2018.

[196] R Cailleau, R Young, M Olive, and WJ Reeves Jr. Breast Tumor Cell Lines From Pleural Effusions 2. *Journal of the National Cancer Institute*, 53(3):661–674, 1974.

[197] Andy J Minn, Gaorav P Gupta, Peter M Siegel, Paula D Bos, Weiping Shu, Dilip D Giri, Agnes Viale, Adam B Olshen, William L Gerald, and Joan Massagué. Genes that mediate breast cancer metastasis to lung. *Nature*, 436(7050):518–524, 2005.

[198] Xin-Zhong Chang, Da-Qiang Li, Yi-Feng Hou, Jiong Wu, Jin-Song Lu, Gen-Hong Di, Wei Jin, Zhou-Luo Ou, Zhen-Zhou Shen, and Zhi-Ming Shao. Identification of the functional role of peroxiredoxin 6 in the progression of breast cancer. *Breast Cancer Research*, 9(6):R76, 2007.

[199] Ebony R Fietz, Christine R Keenan, Guillermo López-Campos, Yan Tu, Cameron N Johnstone, Trudi Harris, and Alastair G Stewart. Glucocorticoid resistance of migration and gene expression in a daughter MDA-MB-231 breast tumour cell line selected for high metastatic potential. *Scientific reports*, 7:43774, 2017.

[200] Cameron N Johnstone, Perry S Mongroo, A Sophie Rich, Michael Schupp, Mark J Bowser, John W Tobias, Yingqiu Liu, Gregory E Hannigan, Anil K Rustgi, et al. Parvin-β

inhibits breast cancer tumorigenicity and promotes CDK9-mediated peroxisome proliferator-activated receptor gamma 1 phosphorylation. *Molecular and cellular biology*, 28(2):687–704, 2008.

[201] Leonard D Shultz, Bonnie L Lyons, Lisa M Burzenski, Bruce Gott, Xiaohua Chen, Stanley Chaleff, Malak Kotb, Stephen D Gillies, Marie King, Julie Mangada, et al. Human lymphoid and myeloid cell development in NOD/LtSz-scid IL2Rγnull mice engrafted with mobilized human hemopoietic stem cells. *The Journal of Immunology*, 174(10):6477–6489, 2005.

[202] S Eva Singletary and James L Connolly. Breast cancer staging: working with the sixth edition of the AJCC Cancer Staging Manual. *CA: a cancer journal for clinicians*, 56(1):37–47, 2006.

[203] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, page gkv007, 2015.

[204] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133, 2012.

[205] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[206] Laurence R Meyer, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Robert M Kuhn, Matthew Wong, Cricket A Sloan, Kate R Rosenbloom, Greg Roe, Brooke Rhead, et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research*, 41(D1):D64–D69, 2013.

[207] Paul F Harrison, Andrew D Pattison, David R Powell, and Traude H Beilharz. Topconfects: a package for confident effect sizes in differential expression analysis provides improved usability ranking genes of interest. *bioRxiv*, DOI: https://doi.org/10.1101/343145, 2018.

[208] Michael Lawrence, Wolfgang Huber, Hervé Pages, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

[209] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[210] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[211] Amrei Jänicke, John Vancuylenberg, Peter R Boag, Ana Traven, and Traude H Beilharz. ePAT: a simple method to tag adenylated RNA to measure poly (A)-tail length and other 3′ RACE applications. *RNA*, 18(6):1289–1295, 2012.

[212] Traude H Beilharz and Thomas Preiss. Transcriptome-wide measurement of mRNA polyadenylation state. *Methods*, 48(3):294–300, 2009.

[213] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012, 2012.

[214] Peter J Unrau and David P Bartel. RNA-catalysed nucleotide synthesis. *Nature*, 395(6699):260, 1998.

[215] Traude H Beilharz, Paul F Harrison, Douglas Maya Miles, Michael Ming See, Uyen Minh Merry Le, Ming Kalanon, Melissa Jane Curtis, Qambar Hasan, Julie Saksouk, Thanasis Margaritis, et al. Coordination of Cell Cycle Progression and Mitotic Spindle Assembly

Involves Histone H3 Lysine 4 Methylation by Set1/COMPASS. *Genetics*, 205(1):185–199, 2017.

[216] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[217] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[218] Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, 8(8):1551, 2013.

[219] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[220] JH Krijthe. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R package version 0.13, URL https://github. com/jkrijthe/Rtsne*, 2015.

[221] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[222] Paul Harrison and Torsten Seemann. From high-throughput sequencing read alignments to confident, biologically relevant conclusions with Nesoni. 2009.

[223] Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C Faircloth, Maido Remm, and Steven G Rozen. Primer3—new capabilities and interfaces. *Nucleic acids research*, 40(15):e115–e115, 2012.

[224] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[225] Balázs Győrffy, Christos Hatzis, Tara Sanft, Erin Hofstatter, Bilge Aktas, and Lajos Pusztai. Multigene prognostic tests in breast cancer: past, present, future. *Breast cancer research*, 17(1):11, 2015.

[226] Jayanth Kumar Palanichamy and Dinesh S Rao. miRNA dysregulation in cancer: towards a mechanistic understanding. *Frontiers in genetics*, 5, 2014.

[227] Andreas Scorilas. Polyadenylate polymerase (PAP) and 3'end pre-mRNA processing: function, assays, and association with disease. *Critical reviews in clinical laboratory sciences*, 39(3):193–224, 2002.

[228] Ju Youn Lee, Ijen Yeh, Ji Yeon Park, and Bin Tian. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic acids research*, 35(suppl_1):D165–D168, 2007.

[229] Ali Shilatifard, D Roxanne Duan, Dewan Haque, Charles Florence, William H Schubach, Joan Weliky Conaway, and Ronald C Conaway. ELL2, a new member of an ELL family of RNA polymerase II elongation factors. *Proceedings of the National Academy of Sciences*, 94(8):3639–3643, 1997.

[230] Feng Wang, Elaine R Podell, Arthur J Zaug, Yuting Yang, Paul Baciu, Thomas R Cech, and Ming Lei. The POT1–TPP1 telomere complex is a telomerase processivity factor. *Nature*, 445(7127):506, 2007.

[231] Takanobu Yoshimoto, Manfred Boehm, Michelle Olive, Martin F Crook, Hong San, Thomas Langenickel, and Elizabeth G Nabel. The arginine methyltransferase PRMT2 binds RB and regulates E2F function. *Experimental cell research*, 312(11):2040–2053, 2006.

[232] Aiguo Lu, Xiongzhi Wangpu, Dingpei Han, Hao Feng, Jingkun Zhao, Junjun Ma, Shun Qu, Xuehua Chen, Bingya Liu, and Minhua Zheng. TXNDC9 expression in colorectal cancer cells and its influence on colorectal cancer prognosis. *Cancer investigation*, 30(10):721–726, 2012.

[233] Dan Xu, Fumitaka Takeshita, Yumiko Hino, Saori Fukunaga, Yasusei Kudo, Aya Tamaki, Junko Matsunaga, Ryou-u Takahashi, Takashi Takata, Akira Shimamoto, et al. miR-22 represses cancer progression by inducing cellular senescence. *The Journal of cell biology*, 193(2):409–424, 2011.

[234] Ramiro Garzon, George A Calin, and Carlo M Croce. MicroRNAs in cancer. *Annual review of medicine*, 60:167–179, 2009.

[235] Deepali Shinde, Yinglei Lai, Fengzhu Sun, and Norman Arnheim. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis:(CA/GT) n and (A/T) n microsatellites. *Nucleic acids research*, 31(3):974–980, 2003.

[236] Jeffrey J Quinn and Howard Y Chang. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, 17(1):47, 2016.

[237] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[238] Mikala Egeblad, H-C Jennifer Shen, Danielle J Behonick, Lisa Wilmes, Alexandra Eichten, Lidiya V Korets, Farrah Kheradmand, Zena Werb, and Lisa M Coussens. Type I collagen is a genetic modifier of matrix metalloproteinase 2 in murine skeletal development. *Developmental Dynamics*, 236(6):1683–1693, 2007.

[239] Noah Spies, Christopher B Burge, and David P Bartel. 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome research*, 23(12):2078–2090, 2013.

[240] Tala Bakheet, Edward Hitti, and Khalid S A Khabar. ARED-Plus: an updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs. *Nucleic acids research*, 46(D1):D218–D220, 2017.

[241] Wijdan Al-Ahmadi, Maha Al-Ghamdi, Norah Al-Souhibani, and Khalid SA Khabar. miR-29a inhibition normalizes HuR over-expression and aberrant AU-rich mRNA stability in invasive cancer. *The Journal of pathology*, 230(1):28–38, 2013.

[242] Sohail F Tavazoie, Claudio Alarcón, Thordur Oskarsson, David Padua, Qiongqing Wang, Paula D Bos, William L Gerald, and Joan Massagué. Endogenous human microRNAs that suppress breast cancer metastasis. *Nature*, 451(7175):147, 2008.

[243] Diana M Cittelly, Partha M Das, Nicole S Spoelstra, Susan M Edgerton, Jennifer K Richer, Ann D Thor, and Frank E Jones. Downregulation of miR-342 is associated with tamoxifen resistant breast tumors. *Molecular cancer*, 9(1):317, 2010.

[244] Jie Gao, Laisheng Li, Minqing Wu, Min Liu, Xinhua Xie, Jiaoli Guo, Hailin Tang, and Xiaoming Xie. MiR-26a inhibits proliferation and migration of breast cancer through repression of MCL-1. *PloS one*, 8(6):e65138, 2013.

[245] Hongling Li, Chunjing Bian, Lianming Liao, Jing Li, and Robert Chunhua Zhao. miR-17-5p promotes human breast cancer cell migration and invasion through suppression of HBP1. *Breast cancer research and treatment*, 126(3):565–575, 2011.

[246] Shihoko Kojima, Elaine L Sher-Chen, and Carla B Green. Circadian control of mRNA polyadenylation dynamics regulates rhythmic protein expression. *Genes & development*, 26(24):2724–2736, 2012.

[247] Maria Piqué, José Manuel López, Sylvain Foissac, Roderic Guigó, and Raúl Méndez. A combinatorial code for CPE-mediated translational control. *Cell*, 132(3):434–448, 2008.

[248] Wei Chen, Qi Jia, Yifan Song, Haihui Fu, Gang Wei, and Ting Ni. Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics, proteomics & bioinformatics*, 2017.

[249] Elmar Wahle. Poly (A) tail length control is caused by termination of processive synthesis. *Journal of Biological Chemistry*, 270(6):2800–2808, 1995.

[250] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[251] Elzbieta A Slodkowska and Jeffrey S Ross. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert review of molecular diagnostics*, 9(5):417–422, 2009.

[252] Ruijia Wang, Ram Nambiar, Dinghai Zheng, and Bin Tian. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic acids research*, 46(D1):D315–D319, 2017.

[253] Zhe Ji and Bin Tian. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PloS one*, 4(12):e8419, 2009.

[254] Ben J Marafino, W John Boscardin, and R Adams Dudley. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of biomedical informatics*, 54:114–120, 2015.

[255] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[256] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[257] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[258] David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

[259] Patrick Royston and Willi Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in medicine*, 23(5):723–748, 2004.

[260] Naoyuki Fujita, David L Jaye, Masahiro Kajita, Cissy Geigerman, Carlos S Moreno, and Paul A Wade. MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell*, 113(2):207–219, 2003.

[261] Mark P Chao, Irving L Weissman, and Ravindra Majeti. The CD47–SIRPα pathway in cancer immune evasion and potential therapeutic implications. *Current opinion in immunology*, 24(2):225–232, 2012.

[262] Paola A Betancur, Brian J Abraham, Ying Y Yiu, Stephen B Willingham, Farnaz Khameneh, Mark Zarnegar, Angera H Kuo, Kelly McKenna, Yoko Kojima, Nicholas J Leeper, et al. A CD47-associated super-enhancer links pro-inflammatory signalling to CD47 upregulation in breast cancer. *Nature Communications*, 8, 2017.

[263] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[264] Peter Warren. panp: Presence-Absence Calls from Negative Strand Matching Probesets. *R package version 1.26*, 2016.

[265] Anavaj Sakuntabhai, Susan Burge, Sarah Monk, and Alain Hovnanian. Spectrum of novel ATP2A2 mutations in patients with Darier's disease. *Human molecular genetics*, 8(9):1611–1619, 1999.

[266] Aishwarya G Jacob, Ravi K Singh, Fuad Mohammad, Thomas W Bebee, and Dawn S Chandler. The splicing factor FUBP1 is required for the efficient splicing of oncogene MDM2 pre-mRNA. *Journal of Biological Chemistry*, 289(25):17350–17364, 2014.

[267] Robert Duncan, Leonard Bazar, Greg Michelotti, Takeshi Tomonaga, Henry Krutzsch, Mark Avigan, and David Levens. A sequence-specific, single-strand binding protein activates

the far upstream element of c-myc and defines a new DNA-binding motif. *Genes & Development*, 8(4):465–480, 1994.

[268] Tatsuya Mori, Tomoe Wada, Takahiro Suzuki, Yoshitsugu Kubota, and Naoyuki Inagaki. Singar1, a novel RUN domain-containing protein, suppresses formation of surplus axons for neuronal polarity. *Journal of Biological Chemistry*, 282(27):19884–19893, 2007.

[269] G Wang, Q Zhang, Y Song, X Wang, Q Guo, J Zhang, J Li, Y Han, Z Miao, and F Li. PAK1 regulates RUFY3-mediated gastric cancer cell migration and invasion. *Cell death & disease*, 6(3):e1682, 2016.

[270] VY Shin, JM Siu, I Cheuk, EKO Ng, and A Kwong. Circulating cell-free miRNAs as biomarker for triple-negative breast cancer. *British journal of cancer*, 112(11):1751–1759, 2015.

[271] Brian J Haas and Michael C Zody. Advancing RNA-seq analysis. *Nature biotechnology*, 28(5):421–423, 2010.

[272] Tineke Casneuf, Yves Van de Peer, and Wolfgang Huber. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC bioinformatics*, 8(1):461, 2007.

[273] Cherie Blenkiron, Leonard D Goldstein, Natalie P Thorne, Inmaculada Spiteri, Suet-Feung Chin, Mark J Dunning, Nuno L Barbosa-Morais, Andrew E Teschendorff, Andrew R Green, Ian O Ellis, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome biology*, 8(10):R214, 2007.

[274] Jelena Radojicic, Apostolos Zaravinos, Thomas Vrekoussis, Maria Kafousi, Demetrios A Spandidos, and Efstathios N Stathopoulos. MicroRNA expression analysis in triple-negative (ER, PR and Her2/neu) breast cancer. *Cell cycle*, 10(3):507–517, 2011.

[275] Luciano Cascione, Pierluigi Gasparini, Francesca Lovat, Stefania Carasi, Alfredo Pulvirenti, Alfredo Ferro, Hansjuerg Alder, Gang He, Andrea Vecchione, Carlo M Croce, et al.

Integrated microRNA and mRNA signatures associated with survival in triple negative breast cancer. *PloS one*, 8(2):e55910, 2013.

[276] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

[277] Potter Wickware. Next-generation biologists must straddle computation and biology. *Nature*, 404(6778):683–684, 2000.

[278] Florian Markowetz. All biology is computational biology. *PLoS biology*, 15(3):e2002050, 2017.

[279] Andrew J Severin. Dealing with data: training new scientists. *Science*, 331(6024):1516–1516, 2011.

[280] Roberto Spreafico, Simon Mitchell, and Alexander Hoffmann. Training the 21 st Century Immunologist. *Trends in immunology*, 36(5):283–285, 2015.

[281] Pavel Pevzner and Ron Shamir. Computing has changed biology—biology education must catch up. *Science*, 325(5940):541–542, 2009.

[282] Tin Wee Tan, Shen Jean Lim, Asif M Khan, and Shoba Ranganathan. A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the"-omics" era. *BMC genomics*, 10(3):S36, 2009.

[283] Ina Koch and Georg Fuellen. A review of bioinformatics education in Germany. *Briefings in bioinformatics*, 9(3):232–242, 2008.

[284] Russ B Altman and Teri E Klein. Biomedical informatics training at Stanford in the 21st century. *Journal of biomedical informatics*, 40(1):55–58, 2007.

[285] Erwin L van Dijk, Yan Jaszczyszyn, and Claude Thermes. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, 322(1):12–20, 2014.

[286] Charity W Law, Monther Alhamdoosh, Shian Su, Gordon K Smyth, and Matthew E Ritchie. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*, 5, 2016.

[287] Tobias Mann, Richard Humbert, Michael Dorschner, John Stamatoyannopoulos, and William Stafford Noble. A thermodynamic approach to PCR primer design. *Nucleic acids research*, 37(13):e95–e95, 2009.

[288] Winston Chang, Joe Cheng, J Allaire, Yihui Xie, and Jonathan McPherson. Shiny: web application framework for R. *R package version 0.11*, 1, 2015.

[289] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[290] Kenneth J Breslauer, Ronald Frank, Helmut Blöcker, and Luis A Marky. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences*, 83(11):3746–3750, 1986.

[291] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44, 2008.

[292] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010.

[293] Steven P Lund, Dan Nettleton, Davis J McCarthy, and Gordon K Smyth. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology*, 11(5), 2012.

[294] Aaron TL Lun, Yunshun Chen, and Gordon K Smyth. Its DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Statistical Genomics: Methods and Protocols*, pages 391–416, 2016.

[295] Davis J McCarthy and Gordon K Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, 2009.

Appendix A1 - mPAT and alPAT gene specific forward primer lists

Table A.1. Primers used for *in vitro* mPAT of MXM cell lines.

Chr	Start	End	Strand	Gene	Sequence
1	43322999	43323398	+	*TIE1*	ACTCCAGCTCCTTCG CTTAA
1	155912813	155913212	-	*KIAA0907*	AGTGAAGGAGCTAGT CTGCTGAAC
1	182382614	182383013	-	*GLUL*	TCTTTTGCCATGACAA CTTCTAA
1	186671537	186671936	-	*PTGS2*	GTTGTGTATGCGAAT GTTTCAG
1	186671537	186671936	-	*ENG*	GTTGTGTATGCGAAT GTTTCAG
1	193091647	193092046	+	*TROVE2*	CTAGCCTTGCAAAAT ACCCTACTA
1	202146830	202147229	-	*PTPN7*	GTGAGAGGCACGACT GTCTATG
10	6842730	6843129	+	*LINC00707*	CAAACCCGCAGTAAA CTTTC
10	44991149	44991548	+	*RASSF4*	CCAGAGCCCTGTCAG TTGAT
10	44994634	44995033	+	*RASSF4*	CCCACTGTTTTGTCG ACTCT
10	51695281	51695680	-	*CSTF2T*	AGGATAACTCGCGTT TACATATG
10	51696944	51697343	-	*CSTF2T*	CTAAAAGTTGACCTG TTAAAACGTT
10	86925020	86925419	+	*BMPR1A*	GTCAAATATGTTCTG GACAGCTAA
10	98247690	98247790	-	*LOXL4*	GTCAGTTTAGTTAAG GATGGAACC
10	98247794	98247894	-	*LOXL4*	GTCAGTTTAGTTAAG GATGGAACC
11	62787264	62787663	+	*TAF6L*	CTCTCGGACTACTCG CTGTACTT
11	64316554	64316953	+	*AP001453.1*	CCCTGCAGAGCAATA ACACTAT
11	119093372	119093771	+	*HMBS*	AGTGATTACCCCGGG AGAC

11	120111071	120111470	-	*TRIM29*	GTTGCATATCAGGGTGCTCA
12	1791635	1792034	-	*CACNA2D4*	GATGTATTCACGTAACATGCTTGA
12	14498127	14498526	+	*ATF7IP*	TTTTAATCTTGTGCATGATACCC
12	14502860	14503259	+	*ATF7IP*	GTTACCTGTAGTGGGGTTTTGTC
12	46182839	46183238	-	*SLC38A1*	CAGATGGGTGATTTAAGTGAGTCA
13	110722236	110722635	+	*ING1*	TTTAAGAGTCTCGGGTGTTTAAAT
14	47200212	47200611	-	*RPL13AP2*	TACAGAGGTCTTCAAGACTCACG
14	103911983	103912382	-	*C14orf2*	TACCTCAGTTGTACAGGACTTGG
14	103912221	103912620	-	*C14orf2*	GTAAATTTCAGCAAGCCGTGTTA
15	29117633	29118032	+	*APBA2*	CCAGGACACGACTTGTAATGA
15	29118129	29118528	+	*APBA2*	GTCATCTTGTCTCGAAGTCTCTTT
16	14435468	14435867	-	*PARN*	CCCATTCTCCGTTGAAACA
16	48538519	48538918	-	*N4BP1*	CACTTTGTGTCTCGCTTTGAG
16	48542172	48542571	-	*N4BP1*	CAAATGGATATGGTTAGCCTTTAT
16	50231732	50232131	+	*PAPD5*	ATGACCAGCATTGTATTCGTG
17	34257060	34257459	+	*CCL2*	CTACCCCTGGGATGTTTTGA
17	40092494	40092893	+	*THRA*	TCAGTATGTCTGTTATGTGCGATT
17	40093742	40094141	+	*THRA*	TTATAATTAGTCGGGCATGAGTCT
17	49514842	49515241	+	*NGFR*	GAGATGGAACCCTTTTGGC

17	81887548	81887947	-	*ALYREF*	AGACCTGTTTTGTACCGAGTTATT
19	46774644	46775043	-	*SLC1A5*	TCCTGTCCCCATGGTACGT
2	9472976	9473375	+	*CPSF3*	GCTGCACAGAGACTGTACGAG
2	38563239	38563638	-	*HNRNPLL*	TGGATACACGTGTACAGTATGCA
2	38563844	38564243	-	*HNRNPLL*	TTACAGATGGTTCCAATCCCTA
2	73252946	73253345	+	*CCT7*	GGTAGTAATTGGCCCACTCTC
2	190935157	190935556	+	*GLS*	GCATTTTATCTTAGTGTTTGGATG
2	190945512	190945911	+	*GLS*	GAAATCATCCAGGTTTGGTG
2	190965372	190965771	+	*GLS*	ATTTAGATTCTTCCTGGGGATTAC
2	201619893	201620292	-	*ENO1P4*	AGGAATAAATAATGATCCTGAGCT
2	201623506	201623905	-	*ENO1P4*	CAGGTTTAAACATAAATTTAGCG
2	237763345	237763744	+	*LRRFIP1*	CTTAGATATGAAAGAGCCCGAT
2	237765287	237765686	+	*LRRFIP1*	ACACACAACACAATGTTTTCACG
20	6779805	6780204	+	*BMP2*	AAAGAATAAAGCAGGATCCATAGA
20	23685311	23685710	-	*CST4*	TTCTTGCTTCTAATAGACCTGGTA
20	23747231	23747630	-	*CST1*	CTTGCTTCTAATAGCCCTGGTA
20	23823587	23823986	-	*CST2*	CTACTCCCACCCCTTGTAGTGCTC
20	37945191	37945590	+	*VSTM2L*	AATTCTCCTCGGAATTGGC
20	44939198	44939597	+	*PABPC1L*	CTCTAACTTATTTCCCAATTAGTCTGTA

20	44958967	44959366	+	*PABPC1L*	GAGGAGAGGTCTAGCAAAATGA
20	59936546	59936945	-	*PPP1R3D*	TAAAATTTTCTAGTGGGCTGTGC
22	20127260	20127659	+	*RANBP1*	CATGAAAATGTACTGTGCTAACTTTC
3	38007259	38007658	-	*PLCD1*	GCCTTCAGCCCTAACATAGTGT
3	113444351	113444750	-	*SPICE1*	GCTTAAGCAGGTAAGAGTGGT
3	157443462	157443861	+	*PTX3*	ATGTGTTATAATCGAATGTCACGT
4	53459459	53459858	+	*AC098587.1*	TGGCCTTTTGTGTATATTAGTACCA
4	83295104	83295503	-	*HPSE*	ATCTGTCCAACTCAATGGTCTAAC
4	138163809	138164208	-	*SLC7A11*	TATACCTGTCACGCTTCTAGTTGC
5	54977553	54977952	-	*ESM1*	CCTTTGAATGTAAAGCTGCATAAG
5	133955216	133955615	-	*C5orf15*	AAATACCCCTGAACCGTTTTA
5	174418680	174419079	+	*LINC01411*	GAGCAAAACTCTGTAAGAAAGAAAG
6	170572698	170573097	+	*TBP*	TTTCTAATTTATAACTCCTAGGGGTT
7	83958124	83958523	-	*SEMA3A*	GTGCGTGCCCGTTCAATAA
8	85283905	85284304	+	*CA13*	TGTGTTAAAATGGTTATATTGCC
8	94880017	94880416	-	*CCNE2*	TTTATTGTTACGGTATGAAGTCTTC
8	142703384	142703783	+	*LY6K*	CCACAGACTGAGCCTTCCG
9	129094512	129094911	-	*CRAT*	CTGTGGATAACATTGCTAGCG
9	129107140	129107539	-	*CRAT*	ACGGCAGGTACAACCAGATA

9	131122974	131123373	+	*AIF1L*	TCTAATGTAACCAGTAACGTGAGG
X	54068026	54068425	-	*FAM120C*	TCTGTGATGGAAATTGGTCTG
X	54173618	54174017	-	*FAM120C*	CTGCCAACAGCCACGTACGT
X	100840786	100841185	+	*CSTF2*	CATCCTAACCCTTGAATGACTC
X	138631363	138631762	-	*FGF13*	TGGCATAGAGTTGCATGATATGTA

Table A.2. Primers used for *in vivo* mPAT validation of MXM primary tumours (run 2).

Chr	Start	End	Strand	Gene	Sequence
1	23309482	23309881	-	*HNRNPR*	GTCAAAAGCCGTGACAATT
1	36393111	36393510	-	*LSM10*	TGACTTATTGATTATGGAACCTGT
1	43322999	43323398	+	*TIE1*	ACTCCAGCTCCTTCGCTTAA
1	43363030	43363429	+	*CDC20*	GAGGCTATGGCGCTGTTTT
1	182382614	182383013	-	*GLUL*	TCTTTTGCCATGACAACTTCTAA
10	44991149	44991548	+	*RASSF4*	CCAGAGCCCTGTCAGTTGAT
10	44994634	44995033	+	*RASSF4*	CCCACTGTTTTGTCGACTCT
10	51695281	51695680	-	*CSTF2T*	AGGATAACTCGCGTTTACATATG
10	51696944	51697343	-	*CSTF2T*	CTAAAAGTTGACCTGTTAAAACGTT
11	11956210	11956609	+	*USP47*	TGGTTTTAATTAGATGGTTCACTAC
11	11956796	11957195	+	*USP47*	GTTTCCCTTAGACCGATCC
11	64316554	64316953	+	*AP001453.1*	CCCTGCAGAGCAATAACACTAT

11	66016412	66016811	-	*CATSPER1*	TGCTGGAATGATTGTCCGG
12	1791635	1792034	-	*CACNA2D4*	GATGTATTCACGTAACATGCTTGA
12	46182839	46183238	-	*SLC38A1*	CAGATGGGTGATTTAAGTGAGTCA
12	95865973	95866372	+	*SNRPF*	GGAACAACAAAATCGACTTTT
12	95866255	95866654	+	*SNRPF*	CTTTTCTTTTGTAAGCCCAATAT
12	110347157	110347556	+	*ATP2A2*	AGATTCAATCGACTGGGTTTAT
12	110350961	110351360	+	*ATP2A2*	CTAAATGTCAATTTATCACTGCGC
14	47200212	47200611	-	*RPL13AP2*	TACAGAGGTCTTCAAGACTCACG
17	7514511	7514910	+	*POLR2A*	GTGAGTGGTTACAGCTGATCC
17	34257060	34257459	+	*CCL2*	CTACCCCTGGGATGTTTTGA
19	2321203	2321602	-	*LSM7*	CAGTACCGCCTCCTGGAAC
19	46774644	46775043	-	*SLC1A5*	TCCTGTCCCCATGGTACGT
2	9472976	9473375	+	*CPSF3*	GCTGCACAGAGACTGTACGAG
2	10440122	10440521	-	*ODC1*	TTATTCACTCTTCAGACACGCTAC
2	85549108	85549507	-	*GGCX*	GTGCCTGTAATCCAACTACCC
2	85549662	85550061	-	*GGCX*	CCCAGGAGGTGACTTATGC
2	96274435	96274834	-	*SNRNP200*	AGCAGGTGTCATGGGTCAA
2	190935157	190935556	+	*GLS*	GCATTTTATCTTAGTGTTTGGATG
2	190965372	190965771	+	*GLS*	ATTTAGATTCTTCCTGGGGATTAC

20	35703277	35703676	-	*RBM39*	ACAAATGACTTTCATATTGCAAC
20	59936546	59936945	-	*PPP1R3D*	TAAAATTTTCTAGTGGGCTGTGC
21	6484697	6485096	+	*U2AF1*	TTCCCCTTATGAACTGGTTTG
21	43093030	43093429	+	*U2AF1*	TTCCCCTTATGAACTGGTTTG
22	19520435	19520834	+	*CDC45*	ACATCAACATCGTTTGAAACTTG
22	36560991	36561390	+	*CACNG2*	GGAGGTTAGTTTCTTGAACTGGT
3	152465705	152466104	+	*MBNL1*	ATTACTGCAGTAGTTGACTTTGCT
4	40423007	40423406	-	*RBM47*	TTCAAACATTGCTAGTGGTTTAGT
4	53459459	53459858	+	*FIP1L1*	TGGCCTTTTGTGTATATTAGTACCA
4	138163809	138164208	-	*SLC7A11*	TATACCTGTCACGCTTCTAGTTGC
4	146189847	146190246	+	*LSM6*	AGTCATTTTCTTTTACCTCGTTGT
6	170572698	170573097	+	*TBP*	TTTCTAATTTATAACTCCTAGGGGTT
7	83958124	83958523	-	*SEMA3A*	GTGCGTGCCCGTTCAATAA
8	127741277	127741676	+	*MYC*	CAGAATTTCAATCCTAGTATATAGTACC
8	143816171	143816570	-	*PUF60*	GCTGAAGTGTACGACCAGG
8	144097329	144097728	+	*CYC1*	CCAAGTGACCCTGTCCAGT
9	33370977	33371376	+	*NFX1*	AGGTGCATTGATAGTTCCATTAGT
9	129094512	129094911	-	*CRAT*	CTGTGGATAACATTGCTAGCG
9	129107140	129107539	-	*CRAT*	ACGGCAGGTACAACCAGATA

X	41514644	41515043	-	*CASK*	TGTAGAATATACATACCTGTAGGATGC
X	41518881	41519280	-	*CASK*	CTTTAAAATGATAACTAACAGGACAG
X	54068026	54068425	-	*FAM120C*	TCTGTGATGGAAATTGGTCTG
X	54173618	54174017	-	*FAM120C*	CTGCCAACAGCCACGTACGT
X	100840786	100841185	+	*CSTF2*	CATCCTAACCCTTGAATGACTC
X	138631363	138631762	-	*FGF13*	TGGCATAGAGTTGCATGATATGTA

Table A.3. Primers used for *in vitro* alPAT of MXM cell lines.

miRNA	Primer
hsa-miR-9-3p	ATAAAGCTAGATAACCGAAAGT
hsa-miR-29b-3p	TAGCACCATTTGAAATCAGTGTT
hsa-miR-210-3p	TGCGTGTGACAGCGGCTGA
hsa-miR-125b-5p	TCCCTGAGACCCTAACTTGTGA
hsa-let-7a-5p-	TGAGGTAGTAGGTTGTATAGTT
hsa-miR-26a-5p	TTCAAGTAATCCAGGATAGGCT
hsa-miR-16-5p	TAGCAGCACGTAAATATTGGCG
hsa-miR-17-5p	CAAAGTGCTTACAGTGCAGGTAG
hsa-miR-342-3p	TCTCACACAGAAATCGCACC
hsa-miR-381-3p	TATACAAGGGCAAGCTCTCTGT
hsa-miR-496	TGAGTATTACATGGCCAATCTC
hsa-miR-22-3p	AAGCTGCCAGTTGAAGAACT
snord49a	GACGAAGACTACTCCTGTCTGATT
hsa-miR-144-3p	TACAGTATAGATGATGTACT
hsa-miR-21-5p	TAGCTTATCAGACTGATGTTGA
hsa-miR-320c	AAAAGCTGGGTTGAGAGGGT
hsa-miR-23a-3p	ATCACATTGCCAGGGATTTCC
hsa-miR-520c-3p	AAAGTGCTTCCTTTTAGAGGGT
hsa-miR-524-5p	CTACAAAGGGAAGCACTTTCTC
hsa-miR-373-3p	GAAGTGCTTCGATTTTGGGGTGT
hsa-miR-10b-5p	TACCCTGTAGAACCGAATTTGTG
hsa-miR-26b-5p	TTCAAGTAATTCAGGATAGGT
hsa-let-7c-5p	TGAGGTAGTAGGTTGTATGGTT
hsa-miR-155-5p	TTAATGCTAATCGTGATAGGGGT
hsa-miR-206	TGGAATGTAAGGAAGTGTGTGG

hsa-miR-335-5p	TCAAGAGCAATAACGAAAAATGT
hsa-mir-29a-3p	TAGCACCATCTGAAATCGGTTA
hsa-miR-31-5p	AGGCAAGATGCTGGCATAGCT
hsa-miR-126-3p	TCGTACCGTGAGTAATAATGCG
hsa-miR-98-5p	TGAGGTAGTAAGTTGTATTGTT
hsa-miR-105-3p	ACGGATGTTTGAGCATGTGCTA

Appendix A2 - Paul Harrison’s 3’ End shift test description

$3'$ end shift description

An effect size for $3'$ end shifting

The $3'$ end shift test builds on generalized linear modelling, in particular building on the QLSpline method (Lund et al. 2012) implemented in the edgeR package (Lun, Chen, and Smyth 2016). This is a negative binomial model with $\log_2$ link function and quasi-likelihood testing with shrunken overdispersion estimates.

We have n^{gene} genes, each gene i having n_i^{peak} peaks. We have data from n^{samp} biological samples.

The first step, performed by edgeR, is to fit a generalized linear model with $\log_2$ link function and negative binomial distribution for each peak.

We obtain estimated coefficients $\beta_{i,j,k}$ where i is gene, j is peak within the gene, and k is experimental group (1 or 2). The linear model may contain further terms, for example to account for a batch effect, so long as $\beta_{i,j,1}$ represents the $\log_2$ abundance in group 1 and $\beta_{i,j,2}$ represents the $\log_2$ abundance in group 2. Call the number of terms for each peak n^{term}.

From these coefficients we calculate an effect size r_i, which we refer to as the $3'$ end shift. First define proportions within the two groups

$$p_{i,j,k} = \frac{2^{\beta_{i,j,k}}}{\sum_{l=1}^{n_i^{\text{peak}}} 2^{\beta_{i,l,k}}}$$

The $3'$ end shift is

$$r_i = \sum_{j=1}^{n_i^{\text{peak}}} \sum_{k=1}^{n_i^{\text{peak}}} \text{sign}(k-j) p_{i,j,1} p_{i,k,2}$$

r_i is bounded between -1 and +1, with -1 representing a complete shift to a proximal peak and +1 representing a complete shift to a distal peak.

Testing whether the effect size exceeds a given magnitude

Testing may be performed using quasi-likelihood, with the quasi-likelihood overdispersion estimate moderated as in edgeR. edgeR's estimates of the negative binomial dispersion parameter, prior quasi-likelihood overdispersion, and prior degrees of freedom are used.

A deviance D_1 is obtained for an unconstrained model, and D_0 for a model with coefficients constrained to have effect size smaller than a specified amount. In normal edgeR usage, this constraint would be to a subspace of the model coefficients (using a "contrast"), however here the constraint is that the effect size is smaller than a specified magnitude $|r| \leq e$. We have therefore implemented our own constrained maximum-likelihood fitting procedure.

Since multiple peaks for a gene from a single sample will not have independent biological variation, multiple peaks within a sample are counted as a single degree of freedom when accounting degrees of freedom.

edgeR supplies a prior deviance D^{prior} with d^{prior} degrees of freedom (more precisely, it supplies d^{prior} and $D^{\text{prior}}/d^{\text{prior}}$). These moderate the estimate of the overdispersion. The posterior deviance and degrees of freedom for the alternative hypothesis are

$$D^{\text{post}} = D^{\text{prior}} + \frac{D_1}{n^{\text{peak}}}$$

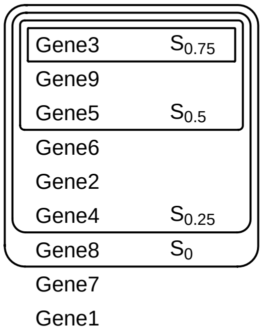


Figure 1: Example of nesting of sets of genes, and a resulting ordering of genes.

$$d^{\text{post}} = d^{\text{prior}} + n^{\text{samp}} - n^{\text{term}}$$

The moderated quasi-likelihood test statistics is then

$$F = \frac{D_1 - D_0}{D^{\text{post}}/d^{\text{post}}}$$

If we were restricting the null hypothesis to a single effect size $r = e$, we would expect this to follow an $\mathcal{F}(1, d^{\text{post}})$ distribution and calculate a p-value accordingly. Since the null hypothesis is instead $|r| \leq e$, this is a slightly conservative method of calculating a p-value—p-values will not follow a uniform distribution in the case of the null hypothesis, but rather be skewed to the right. The treatment of each sample rather than each peak as independent is also conservative—the technical variation component from each peak in each sample might also be considered independent. Using the $\mathcal{F}(1, d^{\text{post}})$ distribution is therefore a valid but conservative way calculate p-values.

A small improvement on this method of calculating p-values is possible, which can reduce the p-value by up to a factor of 2. This will be described in the final section.

Top confident effect sizes

For a given effect size e and for each gene i, calculate the p-value $p_{e,i}$ as described above. A set of genes with effect size at least e at a given False Discovery Rate (FDR) q may be obtained using the procedure of Benjamini and Hochberg (1995). This set S_e is the *largest* set such such that

$$S_e = \left\{ i : p_{e,i} \leq \frac{|S_e|}{n} q \right\}$$

Sets for different effect sizes nest. If $e > e'$ then $S_e \subseteq S_{e'}$. Let $|c_i|$ be the largest e such that $i \in S_e$, and for convenience let this quantity have have the actual sign of the effect, sign $c_i =$ sign r_i. We call this quantitity the "confect", for *con*fident ef*fect* size.

By presenting genes in order from largest to smallest $|c_i|$, the reader may easily choose an effect size resulting in a set of genes S_e of a size suitable for their purpose. Some genes are not a member of any set, and are not given a confect. These are listed last. An illustration of this idea is shown in Figure 1.

There is some similarity in this procedure to the assigning of q-values to genes, for example as produced by the TREAT procedure in the limma R package. q-values allow the reader to select a desired FDR for a fixed effect size. Here instead we have fixed the FDR and allowed the reader to select an effect size.

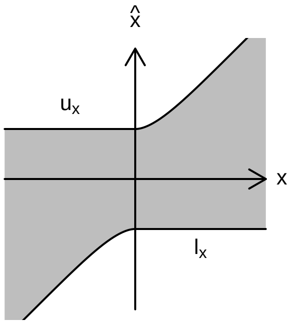


Figure 2: General shape of the upper and lower bounds of the non-rejection region.

A small improvement over the quasi-likelihood F test

Finally we note that a small improvement is possible over the quasi-likelihood test, which may result in up to halved p-values. The idea is that for small effect sizes we must use a two sided test, but for larger effect sizes a one sided test becomes valid. Our method is inspired by the TREAT procedure (McCarthy and Smyth 2009), but our derivation uses confidence sets rather than a test statistic. This is to ensure that we not only determine that the absolute effect size is larger than some value, but also determine that it has a particular sign.

Consider first the estimation of a parameter x with estimate $\hat{x} \sim \mathcal{F}(x)$, with $\mathcal{F}(x)$ having cumulative distribution $F(\hat{x}|x)$. As a typical example, $\hat{x}$ might be normally distributed about x, $\hat{x} \sim \mathcal{N}(x, 1)$.

For each x we will specify a valid non-rejection interval, and use this to construct a confidence-bound procedure. For false rejection probability α, any value of x we must have upper bound u_x and lower bound l_x on $\hat{x}$ satisfying

$$F(u_x|x) - F(l_x|x) \geq 1 - \alpha$$
$$\hat{x} \in [l_x, u_x]$$

For $x = 0$ a valid non-rejection interval is $l_0 = F^{-1}(\alpha/2|0)$, $u_0 = F^{-1}(1 - \alpha/2|0)$. We then choose non-rejection intervals for other values of x that encompass the non-rejection interval for $x = 0$. We will describe the case of $x > 0$, the case for $x < 0$ is symmetric. The lower bound on $\hat{x}$ is $l_x = l_0$. An upper bound u_x on for a valid interval is then

$$F(u_x|x) - F(l_0|x) = 1 - \alpha$$
$$u_x = F^{-1}(1 - \alpha + F(l_0|x)|x)$$

Now consider the transpose of these intervals, the set of non-rejected values of x for a given $\hat{x}$. These are horizontal slices through the graph in Figure 2. We see that either no values of x are rejected, or that a confidence bound is given on x.

If instead of fixing α we wish to obtain a p-value for given x and $\hat{x}$, the p-value is the value of α where $\hat{x}$ lies at the very edge of the interval. This requires solving

$$F(\hat{x}|x) - F(F^{-1}(p/2|0)|x) = 1 - p$$

The solution may be found numerically using Newton's method.

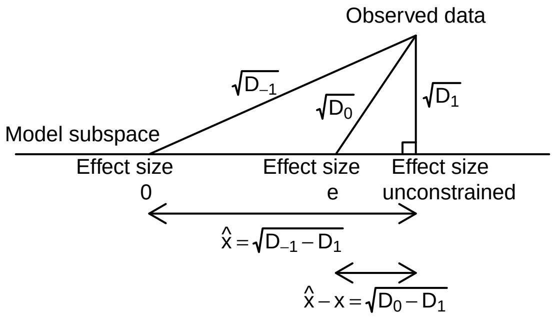


Figure 3: Deviances as measuring-sticks in data-space, in the style of McCullagh and Nelder (1983) pp. 33.

Approximation in a maximum-likelihood context

We now seek to approximate this within a maximum-likelihood framework using z-scores calculated from deviances as measuring-sticks. We fit a third model with the constraint $r = 0$, and call the resulting deviance D_{-1}. Our $\hat{x} = \sqrt{D_{-1} - D_1}$, and $\hat{x} - x = \sqrt{D_0 - D_1}$, and $\hat{x} \sim \mathcal{N}(x, 1)$.

A diagram of the geometry at work is shown in Figure 3. For a linear model with normally distributed residuals having unit standard deviation, identity link function, and effect size based on a linear contrast of model coefficients, the lines and lengths are exactly correct, as is the orthogonality of the line from the unconstrained model to the observed data with the model subspace. If the residuals do not have unit standard deviation, the diagram is correct in a data space of observed values scaled by these standard deviations. With the further generalization to generalized linear models and non-linear effect sizes, the diagram and this method of obtaining x and $\hat{x}$ are only an approximation.

Approximation in a quasi-likelihood context

The final step is to the quasi-likelihood framework. To take into account overdispersion and uncertainty about the overdispersion, we say that our uncertainty of estimates $\hat{x}$ varies about the true value x according to a scaled t distribution.

$$\hat{x} \sim x + \sqrt{\frac{D^{\text{post}}}{d^{\text{post}}}} t(d^{\text{post}})$$

This is the final form of the test used to find top confident $3'$ end shifts.

References

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Lun, Aaron T., Yunshun Chen, and Gordon K. Smyth. 2016. “It’s DE-Licious: A Recipe for Differential Expression Analyses of RNA-Seq Experiments Using Quasi-Likelihood Methods in edgeR.” *Methods in Molecular Biology* 1418: 391–416.

Lund, Steven P., Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth. 2012. “Detecting Differential Expression in RNA-Sequence Data Using Quasi-Likelihood with Shrunken Dispersion Estimates.” *Statistical*

Applications in Genetics and Molecular Biology 11 (5). doi:10.1515/1544-6115.1826.

McCarthy, Davis J., and Gordon K. Smyth. 2009. "Testing Significance Relative to a Fold-Change Threshold Is a TREAT." *Bioinformatics* 25 (6): 765–71. doi:10.1093/bioinformatics/btp053.

McCullagh, P., and J. A. Nelder. 1983. *Generalized Linear Models*. Chapman and Hall.

Appendix A3 - Full list of APA events from paired TCGA analysis

Table A.4. APA events that were assigned a confect value (FDR = 0.05) in matched tumour-normal analysis of TCGA datasets.

Gene	Effect	Confect	Log_2 CPM
KIAA1984	0.27	0.15	1.99
MTA3	-0.20	-0.12	3.54
MED16	-0.26	-0.12	2.94
ATP2A2	-0.21	-0.12	8.39
H2AFJ	-0.20	-0.11	4.37
ATXN3	-0.18	-0.11	4.33
TMEM237	-0.20	-0.11	5.27
NSMCE2	-0.20	-0.1	1.95
SIN3B	0.17	0.1	4.37
FBRSL1	-0.17	-0.1	4.34
LOC401320	-0.19	-0.1	3.40
PARD3	-0.16	-0.1	5.40
H2AFV	-0.19	-0.1	8.31
ENY2	-0.16	-0.1	5.78
GPATCH11	0.19	0.09	3.62
IQCK	-0.16	-0.09	5.29
MFGE8	0.23	0.09	3.23
PPIH	0.17	0.09	2.81
ASB1	-0.16	-0.09	3.57
FAM208A	-0.15	-0.09	5.18
GID8	-0.15	-0.08	6.90
NEK4	-0.17	-0.08	3.09
CXCL12	0.18	0.08	3.70

ABHD3	0.19	0.08	1.85
PDE4DIP	-0.14	-0.08	5.49
FGFR1OP	-0.14	-0.08	4.68
MRPS14	-0.12	-0.07	5.74
ZNRF1	-0.14	-0.07	4.32
HNRNPA3	-0.13	-0.07	9.22
PCGF5	-0.23	-0.07	5.53
SBNO1	-0.12	-0.07	5.65
EPB41L4A	-0.15	-0.07	2.45
PHF19	-0.11	-0.07	3.02
TMEM8B	-0.15	-0.07	2.64
CASP8	-0.15	-0.07	3.79
TXNRD1	0.15	0.07	4.54
SYNCRIP	-0.18	-0.07	8.16
MEF2BNB	-0.14	-0.07	3.01
FAM134A	-0.11	-0.07	6.70
CD47	-0.16	-0.07	7.74
NDUFA10	-0.12	-0.06	4.25
CACNA1D	-0.16	-0.06	4.25
MAN2B1	-0.13	-0.06	5.02
ORC4	-0.15	-0.06	5.63
FUBP1	-0.14	-0.06	6.40
IFI44	0.16	0.06	3.08
PNPT1	-0.11	-0.06	4.10
MECP2	-0.12	-0.06	5.43

ERBB3	-0.15	-0.06	7.96
PPP2R5E	-0.11	-0.06	4.98
UCK2	-0.12	-0.06	5.42
NUDT3,RPS10-NUDT3	-0.12	-0.06	5.18
KLC1	-0.14	-0.06	5.85
GATAD2B	-0.11	-0.06	5.42
CTNNBIP1	-0.11	-0.06	6.39
N6AMT1	-0.12	-0.06	4.22
ZNF259	-0.10	-0.05	5.19
MICALL1	-0.12	-0.05	5.51
ARSA	-0.10	-0.05	4.12
PCED1B	0.18	0.05	1.99
ZNF655	-0.13	-0.05	5.17
FAM208B	-0.11	-0.05	5.21
SLMAP	-0.10	-0.05	6.53
KRT7	0.13	0.05	5.80
ABCG1	-0.10	-0.05	5.05
MAPK13	-0.11	-0.05	6.12
TFDP2	-0.11	-0.05	4.61
MRPS25	-0.11	-0.05	5.00
GGCX	-0.13	-0.05	6.73
SLC25A30	-0.12	-0.05	2.97
SAR1B	-0.11	-0.05	5.68
COL6A2	0.14	0.05	6.93
FAM92A1	0.10	0.05	5.23

LINC00667	-0.13	-0.05	5.59
ODF2	0.11	0.05	3.21
GOSR1	-0.10	-0.05	4.67
BID	-0.11	-0.05	5.23
ARIH2	-0.10	-0.05	6.29
TPM2	0.10	0.05	5.65
HGSNAT	-0.10	-0.05	5.75
SNRNP200	-0.10	-0.05	4.23
PRPF38A	-0.10	-0.05	5.85
KIAA1217	-0.09	-0.05	6.74
EEF2K	-0.09	-0.05	5.65
UBE2N	-0.07	-0.05	7.49
TM9SF3	-0.09	-0.05	10.01
C6orf89	-0.10	-0.05	7.75
RAN	-0.09	-0.05	7.63
RFT1	-0.11	-0.04	2.29
ZNF706	-0.10	-0.04	4.95
BMS1P20	-0.10	-0.04	3.76
JPX	-0.09	-0.04	4.15
GSE1	-0.09	-0.04	5.59
C4orf3	-0.09	-0.04	8.10
CDS2	-0.12	-0.04	6.90
C6orf48	-0.10	-0.04	4.19
ANXA9	-0.11	-0.04	5.16
GALNT16	-0.14	-0.04	3.07

WDR20	-0.09	-0.04	4.54
C1orf21	-0.10	-0.04	6.42
NICN1	-0.10	-0.04	3.33
TM7SF3	-0.08	-0.04	4.76
AP1S2	-0.09	-0.04	5.54
FNTA	-0.09	-0.04	4.94
TMEM64	-0.13	-0.04	5.65
DGKZ	0.11	0.04	3.08
ALDH7A1	-0.08	-0.04	6.08
YRDC	0.08	0.04	5.31
USP31	-0.11	-0.04	3.85
PTCHD3P1	-0.10	-0.04	3.81
GNL1	-0.11	-0.04	4.14
KRT8	0.07	0.04	6.23
EIF4E2	-0.12	-0.04	6.58
RERE	-0.15	-0.04	7.13
CYLD	-0.11	-0.04	6.15
C21orf2	0.08	0.04	3.81
TXNDC17	-0.08	-0.04	4.41
FAF1	-0.09	-0.04	3.67
MUT	-0.07	-0.04	5.96
CCDC50	-0.12	-0.04	6.92
GPATCH2L	-0.10	-0.04	4.56
ZBTB44	-0.09	-0.04	6.60
PTP4A1	-0.06	-0.04	8.74

FBXO22	-0.09	-0.03	3.95
SLIT3	-0.12	-0.03	5.95
HEXIM1	-0.08	-0.03	6.52
DNAJC25	-0.09	-0.03	3.64
SEC22A	-0.08	-0.03	4.09
SPRED1	-0.09	-0.03	5.09
KIF20A	-0.12	-0.03	2.26
PPAPDC1B	-0.12	-0.03	2.34
RNF220	-0.10	-0.03	3.09
OBSL1	-0.09	-0.03	4.18
PDF,COG8	-0.11	-0.03	5.43
NUPL2	0.10	0.03	3.55
PDCD6IP	-0.12	-0.03	7.44
CCDC57	0.13	0.03	2.26
KHSRP	-0.07	-0.03	7.80
BCAS4	-0.09	-0.03	4.31
DNMT3A	-0.08	-0.03	3.84
ANAPC5	0.12	0.03	2.77
PA2G4	-0.08	-0.03	7.31
EPSTI1	-0.11	-0.03	4.18
PCDH7	-0.11	-0.03	3.73
SLC19A1	-0.08	-0.03	4.21
ARL3	-0.10	-0.03	2.53
KIF18B	-0.15	-0.03	1.79
NFIA	-0.10	-0.03	4.39

EXOSC3	-0.09	-0.03	3.96
SRI	-0.07	-0.03	6.04
MAP3K2	-0.09	-0.03	7.00
ANXA11	-0.16	-0.03	6.80
ABI2	-0.10	-0.03	6.89
SUPT7L	-0.07	-0.03	5.89
IDH2	-0.08	-0.03	4.79
AK2	-0.07	-0.03	6.61
SRSF6	-0.10	-0.03	8.03
TBC1D1	-0.08	-0.03	3.57
DDA1	-0.07	-0.03	4.82
TMEM192	-0.07	-0.03	5.58
USP14	-0.10	-0.03	6.71
IP6K2	-0.08	-0.03	5.36
COMMD2	-0.07	-0.03	5.33
RBCK1	-0.08	-0.03	3.06
YIPF6	-0.10	-0.03	7.46
CRHR1	-0.08	-0.03	3.67
RBM17	-0.07	-0.03	7.30
TMEM248	-0.09	-0.03	7.88
ETNK1	-0.09	-0.03	6.98
MTF2	-0.08	-0.03	5.00
CCDC127	-0.07	-0.03	3.99
GLB1	0.08	0.03	3.48
RAB22A	-0.07	-0.03	4.56

NCBP1	-0.09	-0.03	5.86
MAPRE2	-0.08	-0.03	5.12
VAMP4	-0.07	-0.03	5.16
UFD1L	0.07	0.03	3.04
STRIP1	-0.07	-0.03	4.17
KDM5C	-0.07	-0.03	5.50
ZNF506	-0.08	-0.03	4.29
TMEM261	-0.07	-0.03	3.44
MAGOHB	-0.06	-0.03	4.42
VTA1	-0.07	-0.03	5.42
MAN2A2	-0.06	-0.03	5.59
ZMIZ2	0.08	0.03	6.33
GM2A	-0.08	-0.03	7.74
ESYT2	-0.08	-0.03	8.66
EPS15	-0.06	-0.03	6.58
ANKH	-0.07	-0.03	6.06
COA4	-0.05	-0.03	4.35
HS2ST1	-0.07	-0.03	6.44
AGO2	-0.09	-0.03	5.10
CNIH4	-0.07	-0.02	5.09
TMEM110	-0.08	-0.02	5.04
PMM2	-0.06	-0.02	4.52
PDXK	-0.13	-0.02	7.41
GSPT1	-0.14	-0.02	8.04
VPS53	0.09	0.02	5.20

TRAK1	-0.06	-0.02	5.19
USP47	-0.11	-0.02	7.29
ORC2	-0.06	-0.02	5.32
SPIDR	-0.08	-0.02	5.57
PSMG3-AS1	-0.09	-0.02	2.57
AP3S2,C15orf38-AP3S2	-0.07	-0.02	5.05
RCCD1	-0.08	-0.02	3.01
MRP63	-0.07	-0.02	6.45
CNOT6	-0.06	-0.02	5.73
TRIP12	-0.06	-0.02	6.45
RASAL2	-0.07	-0.02	4.50
HIF1AN	-0.07	-0.02	6.11
DPH3	-0.07	-0.02	4.93
HNMT	-0.08	-0.02	5.14
C19orf52	-0.07	-0.02	3.78
DIDO1	0.18	0.02	6.04
ZFYVE16	0.08	0.02	6.76
CAAP1	-0.06	-0.02	3.68
PCCB	-0.08	-0.02	4.32
GNG7	-0.08	-0.02	5.15
ITGB1BP1	-0.06	-0.02	5.79
SHANK2	-0.09	-0.02	3.58
COX19	-0.07	-0.02	4.72
ADCY7	-0.08	-0.02	4.74
PDK3	-0.11	-0.02	5.28

AP4S1	-0.08	-0.02	2.67
RDH13	-0.10	-0.02	2.29
CSTB	-0.05	-0.02	5.80
GSTM3	-0.10	-0.02	5.94
TMEM184C	0.06	0.02	4.60
EIF2AK2	-0.13	-0.02	6.62
VPS13D	-0.06	-0.02	4.87
PPP6C	-0.09	-0.02	7.08
DHX15	0.06	0.02	6.46
TCOF1	-0.07	-0.02	4.21
RAD1	-0.06	-0.02	5.18
PIK3C3	-0.08	-0.02	2.96
NEDD4L	-0.10	-0.02	6.13
LETMD1	-0.08	-0.02	2.81
PPM1A	-0.10	-0.02	6.15
CPSF6	-0.10	-0.02	6.95
ATP2C2	-0.11	-0.02	2.38
C2orf49	-0.06	-0.02	4.55
YY1	-0.08	-0.02	8.67
CRELD1	0.07	0.02	3.39
PREP	-0.07	-0.02	3.64
ELL2	-0.08	-0.02	5.05
ANXA1	0.05	0.02	6.96
SNRPD1	-0.06	-0.02	5.32
HCP5	-0.07	-0.02	3.54

FASTKD2	-0.06	-0.02	5.20
UQCC2	-0.06	-0.02	3.36
PGM2L1	-0.08	-0.02	3.42
HPS1	0.07	0.02	4.67
ZNF586	0.07	0.02	3.15
WDR12	-0.06	-0.02	4.18
SRA1	-0.06	-0.02	5.25
NAA20	-0.06	-0.02	4.55
BMP1	0.06	0.02	3.92
SRF	0.06	0.02	4.42
TMEM19	-0.08	-0.02	5.75
RPS23	-0.06	-0.02	8.16
PLAA	-0.05	-0.02	4.83
CYP20A1	-0.06	-0.02	5.09
SETD5-AS1	-0.07	-0.02	4.22
ARID2	-0.07	-0.02	4.40
ZHX3	-0.07	-0.02	3.97
TRIQK	-0.07	-0.02	5.36
TPST2	-0.07	-0.02	4.94
MRPS21	-0.05	-0.02	5.92
DYM	-0.06	-0.02	4.50
ATF6	-0.09	-0.02	7.59
DDRGK1	-0.06	-0.02	4.07
KIF13A	0.06	0.02	6.30
MSL1	0.10	0.02	7.94

KCTD1	-0.05	-0.02	4.52
DNAJC30	-0.05	-0.02	3.78
KIF1B	0.08	0.02	4.57
TBCEL	-0.05	-0.02	4.57
CDK9	-0.05	-0.02	5.95
NUDT21	-0.06	-0.02	6.34
PHF21A	-0.06	-0.02	5.07
LONP2	-0.06	-0.02	6.83
TMX4	-0.07	-0.02	7.68
GTF2H5	-0.04	-0.02	3.85
USP13	-0.06	-0.02	4.70
WDFY3	-0.06	-0.02	5.85
EPT1	-0.05	-0.02	6.29
CCNL2	-0.10	-0.02	6.62
IMPAD1	-0.06	-0.02	9.03
SLC25A51	0.06	0.02	5.27
LUC7L3	-0.05	-0.02	9.10
DDX18	-0.04	-0.02	7.46
STRN	-0.04	-0.02	6.33
INPP4A	-0.04	-0.02	5.39
PSMB2	-0.05	-0.02	7.25
KIAA1715	-0.03	-0.02	6.56
WASF2	0.03	0.02	9.27
SNRPD3	-0.08	-0.01	7.48
MAFF	0.05	0.01	5.27

PDHA1	-0.06	-0.01	4.35
ZNF107	-0.10	-0.01	3.95
CSNK2A1	-0.06	-0.01	7.15
RBM33	-0.07	-0.01	4.17
TPCN2	-0.07	-0.01	4.36
CLN8	-0.08	-0.01	4.77
NIT2	-0.06	-0.01	2.93
PDE12	-0.07	-0.01	5.27
RBPMS	-0.15	-0.01	6.23
ATP6V1G2-DDX39B,DDX39B	0.07	0.01	4.34
SRSF8	-0.06	-0.01	4.06
RGS5	-0.09	-0.01	6.37
PSEN1	-0.05	-0.01	6.11
MFSD12	-0.06	-0.01	3.14
NEXN	0.10	0.01	6.54
PAM16,CORO7-PAM16	0.10	0.01	1.82
CEP104	-0.06	-0.01	4.79
USP24	-0.05	-0.01	5.75
MOB1A	-0.07	-0.01	8.51
CHID1	-0.09	-0.01	4.73
ZNF606	0.08	0.01	3.89
AKT3	-0.08	-0.01	5.24
MBOAT2	-0.09	-0.01	5.71
PCNX	-0.06	-0.01	6.33
CD109	-0.12	-0.01	5.39

EIF2AK4	-0.07	-0.01	4.45
ZBTB8A	-0.07	-0.01	3.22
METTL2B	-0.05	-0.01	5.47
DENND4C	-0.05	-0.01	5.69
AKT2	-0.04	-0.01	5.97
RASA1	-0.06	-0.01	5.38
GPCPD1	-0.07	-0.01	4.32
RBMS2	-0.09	-0.01	6.41
AGAP1	-0.07	-0.01	4.28
PYGL	-0.06	-0.01	5.36
PGGT1B	-0.08	-0.01	5.45
SLC27A4	0.11	0.01	2.60
PPP6R3	-0.06	-0.01	6.33
PSMF1	-0.05	-0.01	7.55
TRAF3	-0.08	-0.01	4.72
VAPB	-0.05	-0.01	6.04
STX12	-0.05	-0.01	6.43
PSD3	-0.10	-0.01	4.83
SPPL2A	-0.08	-0.01	6.42
C3orf17	-0.06	-0.01	5.85
TNKS	-0.07	-0.01	5.88
ITGBL1	-0.16	-0.01	6.55
CEP68	-0.06	-0.01	4.83
SLC25A53	-0.07	-0.01	2.95
COL1A2	0.15	0.01	12.67

CCNH	-0.07	-0.01	2.57
RNF115	-0.05	-0.01	6.48
CYTH2	0.06	0.01	5.58
POMT2	-0.07	-0.01	2.73
ELOVL6	-0.08	-0.01	4.37
EFNA5	-0.10	-0.01	3.05
NUTF2	-0.05	-0.01	6.02
ARL1	-0.06	-0.01	6.70
VPS41	-0.08	-0.01	5.63
AQR	-0.06	-0.01	4.64
HNRNPUL2	-0.05	-0.01	6.32
TPGS2	-0.07	-0.01	5.30
CSGALNACT1	0.07	0.01	3.06
MTL5	-0.10	-0.01	3.46
UBE2L3	-0.05	-0.01	6.49
TMED10	-0.06	-0.01	8.60
ELP5	-0.08	-0.01	2.70
RBM15B	-0.05	-0.01	6.26
IPP	-0.06	-0.01	3.42
QSOX1	-0.07	-0.01	5.64
HLA-DRB1	-0.04	-0.01	7.01
ATL2	-0.05	-0.01	5.61
C15orf39	0.06	0.01	4.09
GINS2	-0.05	-0.01	3.93
ABHD13	-0.07	-0.01	5.23

ATG5	-0.07	-0.01	6.07
FBXO21	-0.05	-0.01	5.03
NEK6	-0.07	-0.01	6.79
HSPA14	0.06	0.01	2.43
BTD	-0.07	-0.01	3.80
LYRM7	-0.06	-0.01	4.06
AGK	-0.06	-0.01	4.09
STAMBP	-0.05	-0.01	4.91
BTN3A1	-0.04	-0.01	4.53
CHMP3,RNF103-CHMP3	-0.06	-0.01	8.21
TIMM50	-0.08	-0.01	5.51
RPP30	-0.05	-0.01	4.08
NUP98	-0.04	-0.01	5.49
RAI1	-0.06	-0.01	3.10
SAMD4A	-0.08	-0.01	4.22
MAP4	-0.06	-0.01	7.51
COQ7	-0.07	-0.01	5.24
PIGM	-0.07	-0.01	6.35
PLEKHA6	-0.08	-0.01	5.72
FKBP7	-0.06	-0.01	3.72
PIGL	-0.07	-0.01	2.86
RHEB	-0.07	-0.01	3.00
RAB1A	-0.04	-0.01	8.16
FAM104A	-0.04	-0.01	5.35
WWC1	-0.06	-0.01	4.90

PGPEP1	-0.06	-0.01	4.64
NUCKS1	-0.06	-0.01	10.02
RPAP3	-0.06	-0.01	5.08
SNX13	-0.08	-0.01	5.70
PSMD11	-0.05	-0.01	6.68
KMT2D	-0.05	-0.01	5.20
KLF7	-0.07	-0.01	4.52
ELMOD2	-0.06	-0.01	5.43
SIPA1L1	-0.06	-0.01	3.70
METTL7A	0.05	0.01	7.27
PFDN6	-0.06	-0.01	2.63
BTBD9	-0.05	-0.01	3.91
MAFG	-0.06	-0.01	5.01
PHLDA3	-0.05	-0.01	3.77
SMIM7	-0.04	-0.01	7.18
PPID	-0.04	-0.01	5.61
RPL29	-0.06	-0.01	3.22
PQLC2	0.04	0.01	4.09
NAA35	-0.05	-0.01	4.56
NHLRC2	-0.07	-0.01	5.64
POLR3E	-0.05	-0.01	5.02
RBM7	0.05	0.01	4.99
MRPS22	0.04	0.01	5.20
NT5DC1	-0.05	-0.01	5.02
DAB2	-0.05	-0.01	6.65

CUL3	-0.05	-0.01	5.04
COL14A1	-0.09	-0.01	5.65
RNF24	-0.09	-0.01	6.01
TP53BP1	-0.04	-0.01	4.86
TBC1D20	-0.05	-0.01	4.91
MRPL3	-0.04	-0.01	7.05
FBXW2	-0.05	-0.01	5.08
DUSP22	0.04	0.01	5.95
OBFC1	-0.05	-0.01	4.41
TPBG	-0.06	-0.01	5.47
RAP2B	-0.04	-0.01	5.46
ITGB5	-0.05	-0.01	6.78
NAPG	-0.05	-0.01	5.29
PPP4R2	-0.05	-0.01	6.34
REV1	0.04	0.01	4.96
TIPRL	-0.05	-0.01	6.12
KIAA1598	-0.04	-0.01	4.40
PVRL2	0.03	0.01	4.70
AGGF1	-0.04	-0.01	5.55
PKNOX1	-0.04	-0.01	4.72
MTERFD3	0.04	0.01	4.20
WBP4	-0.04	-0.01	4.56
ERP44	-0.04	-0.01	5.26
CHCHD3	0.03	0.01	5.53
SYAP1	-0.03	-0.01	8.87

NEMF	-0.02	-0.01	6.98
PSMD12	-0.05	-0.01	7.31
EXD2	-0.04	-0.01	3.94
TOX4	-0.04	-0.01	6.53
MTR	-0.04	-0.01	5.22
UGCG	-0.03	-0.01	4.55
CSNK1G1	0.06	0.01	4.61
CC2D1B	-0.04	-0.01	4.21
SMAP1	-0.04	-0.01	6.52
HIPK2	0.06	0.01	5.82
FIP1L1	-0.04	-0.01	5.35
GGNBP2	0.04	0.01	6.50
FUCA2	-0.04	-0.01	5.14
ZNF592	-0.04	-0.01	4.37
FAM162A	-0.03	-0.01	5.20
SERBP1	-0.05	-0.01	10.05
TRIP11	-0.07	-0.01	6.14
DNAJB12	0.03	0.01	6.51
NUS1	-0.03	-0.01	4.80
MTHFD2L	-0.05	-0.01	4.68
GATC	-0.03	-0.01	6.28
SLC35F5	-0.07	-0.01	6.49
NOL7	-0.05	-0.01	7.44
COPB1	0.05	0.01	7.84
PAK2	-0.05	-0.01	7.51

SCO1	-0.03	-0.01	4.82
C17orf85	-0.05	-0.01	5.77
NOLC1	-0.03	-0.01	7.72
ATXN10	-0.03	-0.01	6.16
TARDBP	-0.05	-0.01	7.30
WNK1	-0.05	-0.01	8.66
TFAM	-0.03	-0.01	5.93
COX20	-0.04	-0.01	7.76
ANAPC16	-0.05	-0.01	8.63
PAPOLA	0.05	0.01	7.74
CPPED1	-0.05	-0.01	6.84
HAUS2	-0.03	-0.01	5.32
PCBP2	0.04	0.01	10.68
ZRANB1	-0.03	-0.01	7.17
UBAP2L	0.03	0.01	7.22
ARL5A	-0.04	-0.01	7.17
CSNK2A2	-0.02	-0.01	5.90
OSBPL2	-0.02	-0.01	6.85
ABI1	-0.03	-0.01	7.26
OSGIN2	0.04	0.01	7.28
TUSC3	-0.03	-0.01	6.16
ANKRD12	0.02	0.01	7.45
SENP6	-0.02	-0.01	6.87

Appendix A4 - High confidence APA events from TCGA and GEO microarray datasets

Table A.5. High confidence APA events from microarrays and the TCGA. Samples were required to have a confect value and the same direction of APA change. Genes ranked by TCGA confect score.

Gene	TCGA confect	TCGA effect	Microarray confect	Microarray effect
ATP2A2	-0.12	-0.21	-0.47	-0.53
H2AFV	-0.10	-0.19	-0.09	-0.13
FAM208A	-0.09	-0.15	-0.24	-0.29
CXCL12	0.08	0.18	0.08	0.16
SYNCRIP	-0.07	-0.18	-0.24	-0.31
FAM134A	-0.07	-0.11	-0.05	-0.11
FUBP1	-0.06	-0.14	-0.43	-0.50
MICALL1	-0.05	-0.12	-0.02	-0.07
COL6A2	0.05	0.14	0.28	0.41
UBE2N	-0.05	-0.07	-0.25	-0.30
RAN	-0.05	-0.09	-0.25	-0.33
GSE1	-0.04	-0.09	-0.09	-0.15
CDS2	-0.04	-0.12	-0.14	-0.17
FNTA	-0.04	-0.09	-0.08	-0.10
CYLD	-0.04	-0.11	-0.09	-0.13
PTP4A1	-0.04	-0.06	-0.17	-0.21
HEXIM1	-0.03	-0.08	-0.07	-0.11
ANXA11	-0.03	-0.16	-0.08	-0.15
USP14	-0.03	-0.10	-0.04	-0.10
MTF2	-0.03	-0.08	-0.06	-0.11
RAB22A	-0.03	-0.07	-0.16	-0.20

MAPRE2	-0.03	-0.08	-0.16	-0.22
EPS15	-0.03	-0.06	-0.28	-0.33
PDXK	-0.02	-0.13	-0.16	-0.22
GSPT1	-0.02	-0.14	-0.08	-0.12
TRAK1	-0.02	-0.06	-0.06	-0.10
DIDO1	0.02	0.18	0.00	0.03
SHANK2	-0.02	-0.09	-0.15	-0.22
VPS13D	-0.02	-0.06	-0.12	-0.18
PPP6C	-0.02	-0.09	-0.19	-0.22
RAD1	-0.02	-0.06	-0.05	-0.09
TMX4	-0.02	-0.07	-0.25	-0.31
WDFY3	-0.02	-0.06	-0.14	-0.20
DDX18	-0.02	-0.04	-0.12	-0.17
CSNK2A1	-0.01	-0.06	-0.12	-0.16
RGS5	-0.01	-0.09	-0.19	-0.28
AKT3	-0.01	-0.08	-0.05	-0.12
RASA1	-0.01	-0.06	-0.08	-0.11
PSMF1	-0.01	-0.05	-0.31	-0.37
STX12	-0.01	-0.05	-0.07	-0.13
CEP68	-0.01	-0.06	0.00	-0.05
ARL1	-0.01	-0.06	-0.21	-0.26
TPGS2	-0.01	-0.07	-0.03	-0.08
UBE2L3	-0.01	-0.05	-0.12	-0.17
TMED10	-0.01	-0.06	-0.11	-0.14

C15orf39	0.01	0.06	0.04	0.06
ATG5	-0.01	-0.07	-0.22	-0.29
MAP4	-0.01	-0.06	-0.27	-0.30
PSMD11	-0.01	-0.05	-0.25	-0.34
DAB2	-0.01	-0.05	-0.17	-0.23
RNF24	-0.01	-0.09	-0.05	-0.09
FBXW2	-0.01	-0.05	-0.10	-0.14
SERBP1	-0.01	-0.05	-0.14	-0.18
PAK2	-0.01	-0.05	-0.08	-0.12
NOLC1	-0.01	-0.03	-0.11	-0.13
WNK1	-0.01	-0.05	-0.22	-0.28
UBAP2L	0.01	0.03	0.05	0.08
ABI1	-0.01	-0.03	-0.24	-0.30
HSBP1	0.00	-0.07	-0.23	-0.29
RAB11A	0.00	-0.10	-0.02	-0.06
NFATC2IP	0.00	-0.08	-0.21	-0.26
HNRNPK	0.00	-0.04	-0.09	-0.15
MARCH6	0.00	-0.08	-0.12	-0.16
RCHY1	0.00	-0.07	-0.18	-0.23
NCK1	0.00	-0.05	-0.16	-0.23
PCMT1	0.00	-0.03	-0.15	-0.19
TNPO1	0.00	-0.05	-0.06	-0.08
ARHGEF40	0.00	-0.07	-0.06	-0.11
RAD23B	0.00	-0.05	-0.03	-0.08

IGF2R	0.00	0.06	0.06	0.12
PSMA2	0.00	-0.03	-0.09	-0.16
GCLC	0.00	-0.04	-0.26	-0.32
RAB7A	0.00	-0.03	-0.03	-0.06
AGFG1	0.00	-0.06	-0.11	-0.16
GFPT1	0.00	-0.06	-0.06	-0.10
USP7	0.00	-0.04	-0.22	-0.27
CELF1	0.00	0.07	0.17	0.21
SMARCA5	0.00	-0.04	-0.21	-0.26
C11orf24	0.00	0.04	0.02	0.07
TLK2	0.00	-0.03	-0.09	-0.14
GAPVD1	0.00	-0.03	-0.06	-0.09
CSDE1	0.00	-0.04	-0.04	-0.11
METTL13	0.00	-0.02	-0.09	-0.14
KDELR2	0.00	-0.04	-0.19	-0.25
SWAP70	0.00	-0.03	-0.01	-0.06
SFPQ	0.00	-0.04	-0.38	-0.45
CHMP1B	0.00	-0.04	-0.05	-0.10
RAB2A	0.00	-0.02	-0.12	-0.19
CTBP2	0.00	-0.03	-0.07	-0.11
TROVE2	0.00	-0.04	-0.01	-0.05
CD44	0.00	-0.02	-0.13	-0.17
TMCO1	0.00	-0.02	-0.10	-0.17
TMEM41B	0.00	-0.03	-0.15	-0.19

LRRFIP2	0.00	-0.02	-0.08	-0.12
GRSF1	0.00	-0.01	-0.29	-0.36
SMC4	0.00	-0.02	-0.05	-0.11
IPO7	0.00	-0.01	-0.05	-0.11
RAB14	0.00	0.01	0.07	0.12
SUMO1	0.00	-0.01	-0.11	-0.15
NCOA1	0.00	0.00	-0.15	-0.21

Appendix A5 - Alternative polyadenylation in the regulation and dysregulation of gene expression.

Seminars in Cell & Developmental Biology 75 (2018) 61–69

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Alternative polyadenylation in the regulation and dysregulation of gene expression

Rachael Emily Turner[1], Andrew David Pattison[1], Traude Helene Beilharz*

Development and stem cells Program, Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria, 3800, Australia

ARTICLE INFO

Article history:
Received 14 June 2017
Received in revised form 30 August 2017
Accepted 30 August 2017
Available online 1 September 2017

Keywords:
Alternative polyadenylation
3'UTR
Poly(A)-tail
3'-End formation

ABSTRACT

Transcriptional control shapes a cell's transcriptome composition, but it is RNA processing that refines its expression. The untranslated regions (UTRs) of mRNA are hotspots for regulatory control. Features in these can impact mRNA stability, localisation and translation. Here we describe how alternative cleavage and polyadenylation can change mRNA fate by changing the length of its 3'UTR.

Contents

1. Introduction 61
2. RNA processing for alternative polyadenylation 61
3. Mechanisms of APA choice: first come, first served or survival of the fittest? 63
4. The influence of cleavage factor concentration on poly(A)-site choice 64
5. Coupling between transcriptional and 3′-end processing machineries 64
6. Coupling between splicing and 3′-end processing machineries 65
7. Regulation of alternative polyadenylation by PAS occlusion 65
8. Dysregulation of alternative polyadenylation in disease 66
Acknowledgements 67
References 67

1. Introduction

Advances in RNA analysis technologies have led to a new level of appreciation for the complexity of mRNA metabolism. In the area of 3′-end processing, increasing sophistication of first tiling arrays [1–3] and then in next generation sequencing approaches [4–11] has revealed that alternative polyadenylation (APA) is common in eukaryotic mRNA. APA refers to situations where more than one potential polyadenylation site exists within the 3′UTR of a mRNA molecule. The result of this 3′-end diversity, is a highly expanded regulatory and protein coding repertoire over what was previously thought. This provides scope for sophisticated regulatory paradigms in normal growth and development but also means relatively minor changes to mRNA processing can have major implications in disease. Here we will discuss the mechanisms of alternative polyadenylation and how its dysregulation can lead to disease.

2. RNA processing for alternative polyadenylation

Where a nascent mRNA is associated with a single encoded cleavage and polyadenylation site it is termed constitutive polyadenylation (Fig. 1a). Such mRNA polyadenylation can be further classified into four distinct groups based on the position of the poly(A) sites (Fig. 1b–e). The most frequent form of APA is tandem 3′UTR APA, where multiple mRNA isoforms are produced that vary

* Corresponding author.
E-mail address: traude.beilharz@monash.edu (T.H. Beilharz).

[1] Denotes equal author contribution.

https://doi.org/10.1016/j.semcdb.2017.08.056

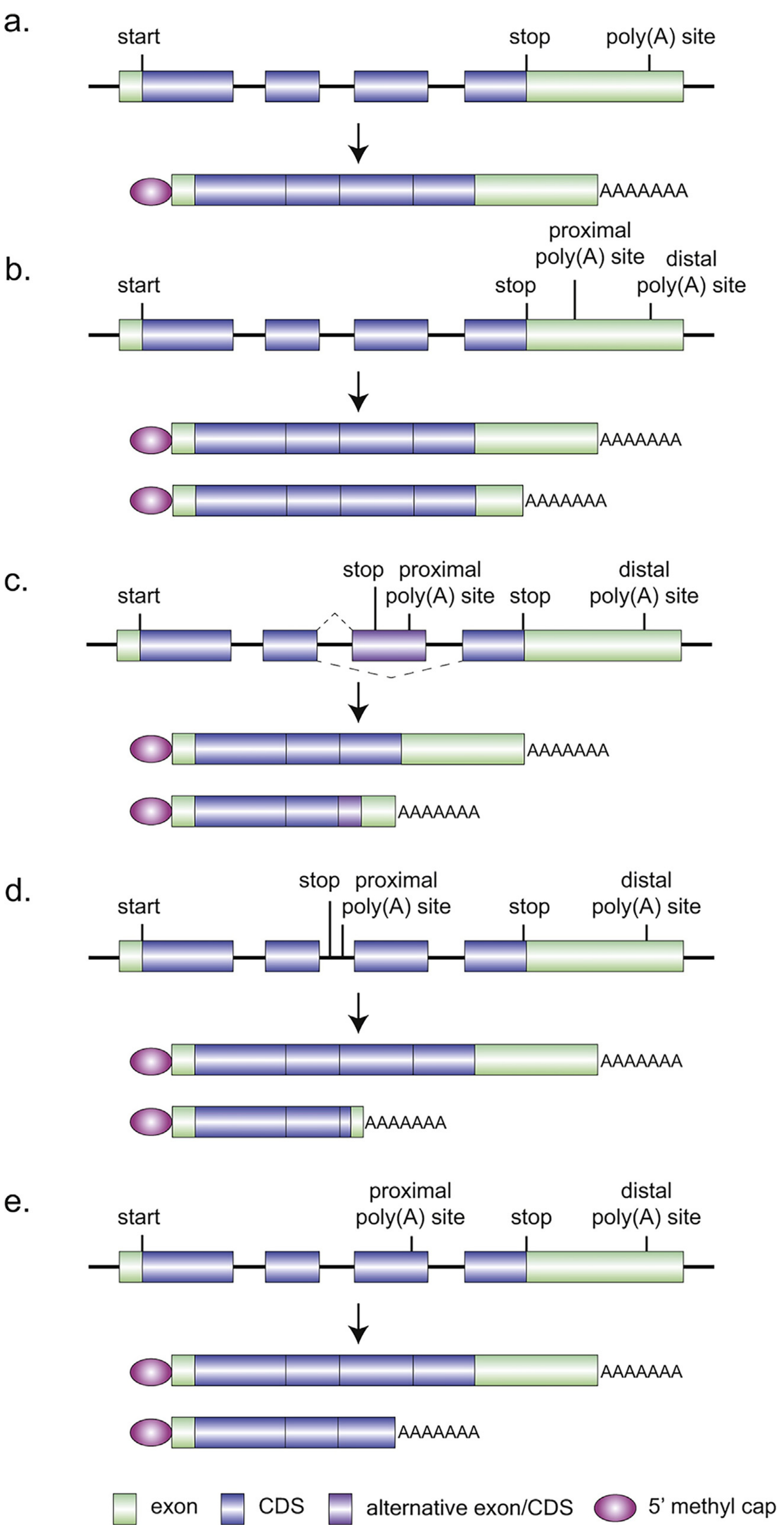


Fig. 1. Mechanisms of Alternative Polyadenylation.

Polyadenylation events can be split into five different categories, four of which involve alternative polyadenylation. **A)** Constitutive polyadenylation involves the occurrence of a single potential poly(A) site within the 3′ UTR of the transcript. **B)** Tandem 3′UTR APA genes possess two or more cleavage poly(A) sites within their 3′UTR. Resulting transcripts only differ by the length of their untranslated region. **C)** Alternative terminal exon APA requires alternative splicing to occur that changes the last exon and therefore the available poly(A) site. **D)** Intronic APA involves the use of cryptic alternative poly(A) sites found within introns. **E)** Internal exon APA uses poly(A) sites within upstream exons and results in a transcript lacking a stop codon or a 3′UTR.

only in the length of their 3′UTR. This version is generally thought to be associated with altered stability, translational efficiency and/or localisation between 3′UTR mRNA isoforms. Unlike tandem 3′UTR APA, the three other forms involve changes to the protein-coding potential and therefore exist under the umbrella term, coding site (CDS) APA. In this case, APA events are linked to alternative splicing and result in distinct protein products between isoforms [12,13]. Firstly, terminal exon APA involves the incorporation of alternative terminal exons into the nascent transcript through the use of different poly(A) sites between isoforms. Secondly, alternative intronic APA results from the recognition of cryptic cleavage sites present within introns. This transforms the normally non-coding region upstream of this site into a composite exon [12]. Lastly, exons other than the terminal exon can also contain cryptic poly(A) sites and this is referred to as internal exon APA. Unlike alternative terminal exon APA, mRNA transcripts produced from internal exon APA will not possess a stop codon unless the site of cleavage directly follows a T or TG residue that is converted into a stop-codon by the addition of a poly(A)-tract. In the absence of a stop, no 3′UTR is generated, and since A-tracts generate C-terminal poly-lysine tags, any resulting protein products are likely targeted for decay [14]. Intronic APA can also suffer this fate, in the absence of an alternative stop codon within the retained part of the intron, upstream of the poly(A) site. It is possible for more than one type of APA to exist for a single transcript.

3. Mechanisms of APA choice: first come, first served or survival of the fittest?

Despite the pervasive nature of APA, how any particular site is specified for cleavage under a given circumstance is still far from clear. Albeit, several models have emerged. Firstly, it has been proposed that the proximal cleavage site has an intrinsic temporal advantage over the distal site [15]. Early work by Denome and Cole [16] demonstrated that when two identical polyadenylation signals were present in the 3′UTR of a reporter construct, increasing the distance between these sites biased poly(A)-site usage further toward proximal by increasing its temporal advantage. Further, in the 'enhancer of rudimentary' e(r) transcript in Drosophila, the proximal cleavage site is preferentially used in males, whereas the distal site is utilised in females. Switching the sequence of these sites did not alter processing [17]. This suggests that position can trump sequence, perhaps simply because a proximal site is transcribed first and therefore has more time to be recognised by the cleavage and polyadenylation machinery. This idea has become the “first come, first served” model [18].

This proximal advantage can be influenced by the transcriptional elongation rate. When APA was monitored in an RNA polymerase II mutant with a slower elongation rate, an aberrant and functionally deleterious proximal cleavage site in the Polo transcript of *D. melanogaster* was preferentially used [19]. This increased proximal usage was interpreted as allowing more time for the proximal site to be recognised by the 3′end processing machinery. If poly(A)-site choice then depends on transcriptional elongation rate, it does not appear to generally correspond to a gene's transcriptional frequency. In both the human and mouse transcriptome, more abundant genes tend to have shorter 3′UTRs than their more lowly expressed counterparts [20], a finding that was recapitulated in reporter assays where expression from stronger promoters caused a preference for proximal site choice [20]. An exception to this general rule is seen in yeast where strong transcriptional up-regulation can shift cleavage to a distal site [21]. This might be explained by the combined effect of very short 3′UTRs in yeast and a high transcriptional rate, such that the 3′-end machinery 'misses the moment' for cleavage at the proximal site, cleaving at a distal site once it gets the chance. Given that the cleavage and adenylation machinery is thought to travel with the polymerase, perhaps it is only once transcription slows, that end formation can occur.

RNA polymerase II pausing has been correlated with switches in cleavage-site choice. Pausing at the poly(A) signal was shown to be heightened at more highly expressed genes indicating that even a local decrease in elongation rate increases proximal site usage [20]. In a further example of this kinetic model for poly(A)-choice, deletion of a pause site downstream of the proximal cleavage site in the IgM gene causes increased use of the more distal site [22]. Such stalls can come from structured DNA or RNA elements. Moreover, local chromatin structure and epigenetic marks impact alternative polyadenylation. Poly(A)-sites are associated with a strong depletion of nucleosomes whereas downstream regions are enriched for nucleosomes [23–25]. For genes with multiple sites, the more highly used site was associated with a lower nucleosome occupancy directly surrounding the poly(A) site and more pronounced nucleosome enrichment downstream [25]. This may be due to altered RNA polymerase II kinetics or to the presence of a more favourable environment for efficient assembly of the cleavage and polyadenylation machinery [26]. Finally, the fact that epigenetics can play a role is demonstrated in findings that cleavage site choice can occur in an allele-specific manner. For the mouse imprinted gene H13, the alternative poly(A)-sites are separated by a CpG island. However, this is only methylated on the maternal copy. This methylation causes stalling of RNA polymerase II at this site only on the maternal allele, biasing cleavage toward the proximal site, whereas distal sites are selected in the absence of methylation at the paternal allele [27]. Such findings have prompted an alternative model, being that of 'survival of the fittest'.

The “survival of the fittest” model explains changes in cleavage site choice that favour the use of more distal poly(A) sites. This model focuses on findings showing that the positioning and efficiency of 3′-end cleavage and polyadenylation are largely determined by the interaction of cleavage and polyadenylation factors with *cis* elements surrounding the potential poly(A)-site. Several core *cis* elements within the pre-mRNA appear to be important for 3′-end processing. Proudfoot and Brownlee [28] discovered a conserved AAUAAA hexamer in the region upstream of the cleavage site in metazoans (Fig. 2). This element has been termed the polyadenylation signal (PAS) and typically occurs within 40 nucleotides of the cleavage site [29]. Although AAUAAA is the canonical PAS sequence, variants of this hexamer are also observed in metazoans [5,30,31]. In general, deviations from the canonical sequence are associated with weaker PAS [reviewed: [32]]. Such variant PAS are more frequent in genes with multiple polyadenylation sites [33,34] with the canonical PAS tending to occur at distal sites, and proximal sites often being variant signals [5,35].

In addition to the PAS element, there are distinct U-rich and GU-rich elements located within 100 nucleotides downstream (DSE) of the cleavage site [36]. An upstream U-rich element (USE) containing UGUA motifs also tends to be positioned within 40–100 nucleotides of the cleavage site [37]. This *cis* element pattern is generally conserved, however, higher variation in signals are seen for plants and the budding yeast *Saccharomyces cerevisiae* [38]. Together, these core motifs along with auxiliary elements help to define the cleavage site of the mRNA and impact the efficiency of 3′end processing. It has been suggested that these elements act cooperatively so that the absence of, or weak sequences for one element may be compensated for by stronger elements also present at that cleavage site [38]. To this end, sites lacking a PAS element tend to rely on strong USEs, DSEs or auxiliary elements to bind the 3′end processing machinery [33,39].

In general, the more distal poly(A) sites utilise stronger *cis* elements than proximal sites and are also more likely to be conserved

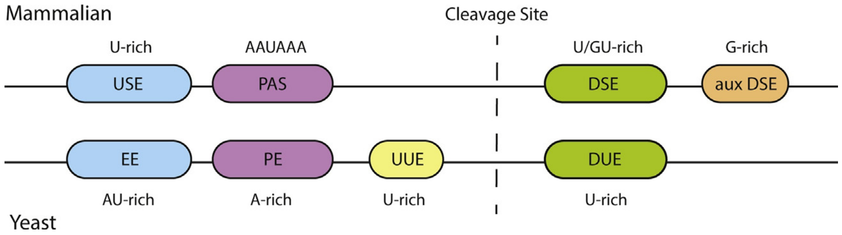


Fig. 2. Core Cis Elements Involved in Cleavage and Polyadenylation Site Recognition.
The cleavage and polyadenylation machinery is guided to target RNA by a series of *cis* elements as indicated. These are generalised here in schematic form, but it is important to note that few given genes will encode all elements. The positioning elements in simple eukaryotes such as *S. cerevisiae* tend to be more divergent than their mammalian counterparts.

[35]. Thus, proximal sites tend to be weaker than distal sites and are less likely to be recognised by the 3′end processing machinery [40]. However favourable the sequence configuration in *cis*, the actual site that is chosen for cleavage and polyadenylation is strongly impacted by the concentration of the core cleavage and polyadenylation factors as well as other proteins.

4. The influence of cleavage factor concentration on poly(A)-site choice

A complex machinery of protein factors is involved in the 3′-end processing of eukaryotic pre-mRNA. This consists of 20 characterised proteins in yeast [41] and perhaps more than 80 proteins in human cells [42]. The core factors involved in cleavage and polyadenylation in these two organisms are generally conserved, although there are some differences in the consensus sequence for PAS recognition and sub-complex organisation [43] (See also Table 1). The mammalian 3′-end processing machinery is comprised of four core sub-complexes. These include cleavage and polyadenylation specification factor (CSPF), cleavage stimulation factor (CstF), cleavage factor I (CFIm) and cleavage factor II (CFIIm) [44]. These factors bind selectively to the cleavage and polyadenylation site prior to any reaction taking place [43]. The CPSF, CstF and CFIm complexes recognise and bind to the PAS, DSE and USE respectively [45]. These three factors then recruit CFIIm and other proteins such as the polyadenylation polymerase (Pap), symplekin, and polyadenylated-binding nuclear protein 1 (PABPN1) to form the 3′-end processing machinery and allow cleavage and polyadenylation to take place [46]. In contrast to the mammalian system, the yeast machinery is comprised of three main complexes including cleavage and polyadenylation factor (CPF), cleavage factor IA (CFIA) and cleavage factor IB (CFIB) [47]. CFIA contains homologous subunits to those in mammalian CFIIm and CstF, except for the apparent absence of a CstF-50 equivalent [47]. Similarly, CPF contains subunits that are homologous to those in mammalian CPSF. However, these are distributed into the two different sub-complexes, CFII and PFI, which make up CPF [47]. Some factors also appear to be unique to either mammals or yeast. The yeast CFIB factor, for example, appears to lack a homologue in mammals whereas the mammalian factor CFIm appears to be absent from yeast [48]. The core cleavage factors in both mammalian and yeast systems are indicated in Table 1.

The effect of this machinery's concentration on the regulation of APA was first reported by Takagaki et al. [49,50]. They demonstrated that in resting B cells, the low levels of CstF-64 were associated with the preferential use of the distal cleavage site in IgM mRNAs, which, contained a strong CstF-64 binding site. However, during B cell activation, elevated levels of CstF-64 correlated with a switch to the proximal site with weak CstF-64 binding capability. This suggests that higher levels of this 3′-end processing factor promote the recognition of weaker cleavage sites whilst limiting concentrations cause the preferential use of the stronger sites. This has since been shown to be a global trend for CstF-64 [40,51]. However, the overall effect of CstF-64 knockdown is relatively small compared to the co-depletion of CstF-64 and its variant form CstF-64τ [40]. As CstF-64τ depletion alone also has a comparatively mild impact on cleavage site switching, this suggests that CstF-64 and CstF-64τ have at least partially redundant functions [52]. Similarly, reduced levels of the CPSF factor Fip1 or yeast Hrp1 have been associated with increased distal site usage [53,54]. In striking contrast to this, depletion of the CFIm subunits CFIm-25 or CFIm-68 causes an increased use of more proximal cleavage sites [29,55–58]. This implies that unlike the other core 3′-end processing complexes, CFIm normally represses cleavage at the proximal site. The cellular balance of 3′-end processing sub-machineries is therefore of critical importance to APA choice.

Together these data indicate that the strength of mRNA *cis* elements and the concentration of 3′-end processing factors together define the pattern of APA for many genes. However, it is of note that when target mRNA affected by changes in cleavage machinery concentrations were compared, there was little overlap, suggesting that 3′-end processing factors each influence the alternative polyadenylation of a specific subset of genes [53] (see also Fig. 2). Another possibility is that each different experimental system that has identified shifts in 3′UTR choice [29,40,59] is prone to unique gene regulatory paradigms. A clearer understanding will require systematic depletion of the machinery within a single experimental system [13], and to understand the native 3′-end profile differences between systems.

5. Coupling between transcriptional and 3′-end processing machineries

Functional coupling between transcription and 3′-end processing extends beyond the role of the transcriptional elongation rate in cleavage site choice. RNA polymerase II is an essential mRNA polyadenylation factor in its own right, both *in vivo* and *in vitro* [60,61]. The RNA polymerase II carboxyl-terminal domain (CTD) interacts with many of the 3′-end processing factors [62–65]. Interaction with the CTD is thought to recruit the 3′-end processing sub-complexes to the pre-mRNA, positioning them for cleavage and polyadenylation, with RNA polymerase II acting as a platform for 3′-end processing. Furthermore, various transcription factors have also been shown to play a role in cleavage site selection. PAF1C is a transcriptional elongation factor that has been associated with enhanced 3′-end processing [66]. Depletion of the PAF1C subunits Cdc73, Paf1 or Ski8 resulted in a global increase in proximal site usage in murine myocytes [67]. Paf1 ablation was also linked to the accumulation of RNA polymerase II along gene bodies suggesting that this is due to a reduced transcriptional elongation rate [67]. Similarly, decreased expression of the murine transcription elongation factor Ell2 also caused enhanced proximal site usage and a concomitant switch from membrane-tethered immunoglobulins to their secreted form [68]. In this case, Ell2 increased CstF-64

Table 1
The core cleavage and polyadenylation factors in yeast and mammalian systems.

Mammalian Factor	Mammalian Complex	Yeast Homologue Factor	Yeast Complex
Symplekin		Pta1	CPF(CFII)
CPSF-160 (CPSF1)	CPSF	Cft1 (Yhh1)	CPF(CFII)
CPSF-100 (CPSF2)	CPSF	Cft2 (Ydh1)	CPF(CFII)
CPSF-73 (CPSF3)	CPSF	Ysh1 (Brr5)	CPF(CFII)
CPSF-30 (CPSF4)	CPSF	Yth1	CPF(PFI)
Fip1 (FIP1L1)	CPSF	Fip1	CPF(PFI)
Wdr33	CPSF	Pfs2	CPF(PFI)
PP1		Glc7	CPF
Pap (PAPOLA)		Pap1	CPF
Rbbp6		Mpe1	CPF
		Pti1	
		Ref2	CPF
Ssu72		Ssu72	CPF
Wdr82		Swd2	
		Syc1	
CstF-50 (CSTF1)	CstF		
CstF-64 (CSTF2)	CstF	Rna15	CFIA
CstF-77 (CSTF3)	CstF	Rna14	CFIA
CFIm-25 (CPSF5/NUDT21)	CFIm		
CFIm-68 (CPSF6)	CFIm		
CFIm-59 (CPSF7)	CFIm		
Pcf11	CFIIm	Pcf11	CFIA
Clp1	CFIIm	Clp1	CFIA
		Hrp1 (Nab4)	CFIB

association with RNA polymerase II and thus caused a high local concentration of CstF-64 at more proximal cleavage sites.

6. Coupling between splicing and 3′-end processing machineries

A link between alternative polyadenylation and splicing is becoming increasingly clear (Fig. 1c–e). For example, the U1 snRNP affects cleavage and polyadenylation independently of its role in splicing. Knockdown of U1 snRNP promotes the use of cryptic polyadenylation sites within introns close to the 5′-end of the transcript [69]. It was suggested that binding of U1 snRNP to these regions blocks their recognition. Surprisingly, more moderate decreases in U1 snRNP levels elicited a shift to more proximal 3′UTR cleavage sites rather than the use of upstream intronic cleavage sites indicating that the impact of U1 snRNP is dose-dependent [70]. Furthermore, U2 snRNP mediates CPSF loading to the pre-mRNA and the U2 snRNP auxiliary factor (U2AF) interacts with CFIm-59, both of which stimulate cleavage and polyadenylation [71,72]. A confounding factor to the interpretation of links between splicing and polyadenylation is the degree of cross-talk between the two systems [73]. Movassat et al. [74], suggest for example, that extensive terminal exon splicing induced by CstF-64 knockdown can be explained by changes to 3′UTR choice in core splicing factors, the altered expression of which then indirectly alter transcriptome-wide splicing choices.

7. Regulation of alternative polyadenylation by PAS occlusion

Polyadenylation sites can be blocked by protein and/or RNA elements that compete with the binding of 3′-end processing factors. Multiple proteins have been identified that interfere with CstF-64 binding and cause a shift in polyadenylation site choice [73,75,76]. For example, the poly(A) binding protein nuclear 1 (PABPN1) plays an important role in 3′end processing [77]. In addition to its function in poly(A)-length control, PABPN1 has been shown to associate with weaker proximal APA sites suppressing their cleavage [59]. As a result, PABPN1 knockdown facilitates a shift to proximal site usage. Depletion of the major cytoplasmic poly(A) binding protein, PABPC1, appears to cause 3′UTR shortening to a similar extent as PABPN1 indicating that this may be a general function of PABPs [13]. Conversely, some RNA binding proteins aid the 3′-end processing machinery in site recognition, promoting the use of weaker *cis* elements. The human TREX subunit Thoc5 co-transcriptionally loads CFIm-68 onto target genes [78]. Therefore, like CFIm-68 depletion, a decrease in Thoc5 results in a shift towards proximal site usage [78]. Furthermore, the cytoplasmic polyadenylation element binding protein 1 (CPEB1) can shuttle to the nucleus and bind to cytoplasmic polyadenylation elements (CPE) within the pre-mRNA sequence [79]. This aids in the recruitment of CPSF to weaker DSEs and therefore proximal cleavage site use is increased. Importantly, occluded cleavage sites can be responsive to, and revealed by, cellular signalling. Danckwardt et al. [80], identified the RNA binding proteins FBP2 and FBP3, at upstream elements within the F2 (thrombin) 3′UTR. External stresses such as inflammatory cytokines that activate p38 MAPK, result in phosphorylation of the FBP RNA binding proteins, causing their dissociation from target mRNA, and activation of 3′-end processing [80].

An emerging theme in modulators of APA is to control their own cleavage and polyadenylation in an auto-regulatory fashion. For example, the Drosophila embryonic lethal abnormal visual system protein (ELAV) and its homologue in mammals, HuR, directly bind to proximal cleavage sites suppressing 3′end processing through steric hindrance [81,82]. Both ELAV and HuR auto-regulate their abundance through binding to proximal polyadenylation sites within their own pre-mRNAs causing longer 3′UTR isoforms [83,84]. The cleavage and polyadenylation machinery itself also appears to be self-equilibrating. The CstF-77 primary transcript for example, harbours a cryptic truncating adenylation site that buffers expression of the full-length, functional protein [85]. Thus an excess of cleavage and polyadenylation activity feeds back to limit its own expression.

Pre-mRNA secondary structures have also been shown to impact site selection [86,87]. IEAAT2 possesses a stem-loop structure within its 7th intron. Adenosine/Inosine RNA editing results in increased use of a cryptic alternative polyadenylation site within the stem-loop region [88]. Finally, non-coding RNAs may also play a role in polyadenylation site switching. The long non-coding RNA colon cancer-associated transcript 2 (CCAT2), has been implicated in cleavage site selection for the gene glutaminase (GLS) [89]. CCAT2 interacts with the CFIm complex, causing CFIm-25 recruitment,

Table 2
Recent studies with disease associated APA events.

Disease	Summary
Alzheimer's disease (AD)	Switch to short Tau 3′ UTR isoform in AD patients. Loss of mi*R*-34a repression leads to more aggregates [105]
Parkinson's disease (PD)	Switch to long α-Synuclein 3′UTR isoforms increases protein expression and formation of Lewy bodies in PD [106]
Myotonic dystrophy (DM)	Misregulated APA in DM by inhibition of MBNL proteins [76]
Neuropsychiatric disease	Increased CFIm-25 causes switch to distal cleavage site in the MECP2 3′UTR and neuropsychiatric disease [107]
Heart failure	Altered APA of a subset of genes in the failing human heart [91]
Cardiac hypertrophy	Global trend toward 3′ UTR shortening in cardiac hypertrophy [3,108]
Sindbis virus	Infection causes redistribution of HuR protein to the cytoplasm and altered HuR-regulated APA events [109]
Herpes simplex virus (HSV)	Infected cell culture polypeptide 27 (ICP27) promotes cryptic APA in host cells [110]
Glioblastoma	CFIm-25 depletion causes switch to proximal 3′ UTRs in glioblastoma tumours [58]
Pan cancer APA	Global 3′ UTR shortening associated with tumorigenesis and CstF-64 proposed as master regulator of APA [51]
Prostate cancer	APA in prostate cancer changed the availability of miRNA binding sites, modulating competing endogenous RNA (ceRNA) networks [95]
Triple negative breast cancer (TNBC)	RNA binding protein sites are lost in TNBC through 3′ UTR shortening [111]

which, in this case, promotes cleavage at a cryptic PAS within intron 14 of GLS. This results in the synthesis of the short GAC protein isoform that triggers glutamine metabolism and is implicated in metastasis.

Rather than any of these models acting alone, it is likely that alternative polyadenylation site selection is under the combinatorial control of various mechanisms and that site choice can be regulated in a multitude of ways including those yet to be identified.

8. Dysregulation of alternative polyadenylation in disease

As the majority of human genes undergo alternative 3′-end processing at some time during development [7], it is not surprising that APA dysregulation has been associated with multiple disease states (See Table 2). In general, genome-wide research in disease tissue and in models of disease have suggested a trend toward global shortening of 3′UTRs. Although this trend is being challenged [90]. Furthermore, it is important to note that comparison of the specific transcripts subject to change in 3′-end usage in these studies shows surprisingly little overlap. Of the 3137 different transcripts identified subject to APA across 4 different studies [51,90–92] compared here, only 10 are shared, and of these only 7 change in the same direction, toward shortening (Fig. 3a). Since the discovery that proliferating and cancer cells have a tendency to express shorter 3′UTRs [93,94], the role of APA in cancer has become the most well-studied of any disease state [1,58,95]. In a comprehensive recent review, Gruber et al. [96] provide evidence that the expression level of 3′-end processing factors is highly correlated to the proliferative index of human cells. Given that high levels of such factors tend to alter APA site selection, it is not surprising that APA in cancer has been a major topic of current research.

Perhaps the most comprehensive study of APA in cancer to date was performed by Xia et al. [51], who inferred APA events from the RNA-Seq read coverage of tumour/normal pairs in the cancer genome atlas (TCGA). Their bioinformatic analysis identified 1346 dynamic APA events across 7 tumour types. In general, 3′UTR shortening was found to be associated with tumourigenesis and CstF-64 was suggested to be a master regulator of APA in tumours. The authors also found APA to be more predictive of tumour outcome than gene expression. The possibility of the prognostic value of APA events was independently shown also in prostate cancer by Li et al. [95], albeit, these authors also note a significant sub-population of transcripts that switch to longer 3′UTR isoforms. The heterogeneity of APA events is underpinned by comparison of the overlap of APA events in different cancer types from the study by Xia et al. [51] (Fig. 3b). Of the 920 transcripts subject to APA, only 22 are shared between the five cancers shown. Of these 22, 18 are shortened in all cancer types.

Cancer can be a heterogeneous disease. In breast cancer, for example, multiple subtypes are known, with varying properties determined from both clinical markers [97] and associated gene expression signatures [98]. It remains to be seen whether APA could provide additional prognostic value over current clinical parameters in breast cancer. To date, the research has been mixed. By contrasting two different breast cancer cell lines to a cultured mammary epithelial cell line MCF10A, Fu et al. [92], showed a polar opposite trend in 3′UTR choices. The luminal derived MCF7 cell line exhibited the expected 3′UTR shortening. In contrast, the MDA-MB-231 cells showed lengthening relative to the epithelial control. Such data highlight the complexity of the underlying biology in each individual tumour type and suggest that definitive global statements are no longer useful as descriptors in disease. Instead, new research should focus on identification of the specific APA events that are prognostic for disease outcome.

Importantly, as with alternative splicing, alternative 3′-end processing can have major regulatory effects with very little change in overall mRNA expression level [99]. The addition/removal, of stability, translation and localisation elements in just a few key regulatory genes could send gene expression programs down divergent developmental trajectories. If this were not already complicated enough, new research shows that 3′UTRs can act as protein complex assembly scaffolds [100,101]. In this case, a change in 3′-end usage can impact the co-translational assembly of cellular machines. It is therefore important not to be drawn into generalisations. Although some 3′UTR shortening events clearly result in increases in protein expression (See Table 2), multiple studies now show that shortened 3′UTRs do not necessarily result in more protein synthesis [90]. Moreover, some research suggests that distal isoforms can be more efficiently translated [19,102,103]. Indeed there may be as many activating elements in 3′UTRs as there are repressive elements [104].

While the number of protein-coding genes are maintained, APA events appear to scale with complexity, [Reviewed: [100]], so it follows that APA is likely a means of increasing functional and regulatory diversity. The dysregulation of these processes is therefore likely common to all complex diseases. Systematic identification of the highly diversified and specific disease-associated APA events provides a massive challenge. However, if the early suggestions of the prognostic power of specific APA events in cancer prove correct, there is also a considerable opportunity. This might be particularly the case when APA events are considered in combination with other modes of gene expression control such as alternative splicing and post transcriptional regulatory mechanisms. Finally, in analogy to the breakthroughs in exon-skipping for muscular dystrophy, perhaps future therapeutics that selectively switch 3′UTR choice, will provide an opportunity to produce among the most targeted, medicines ever created.

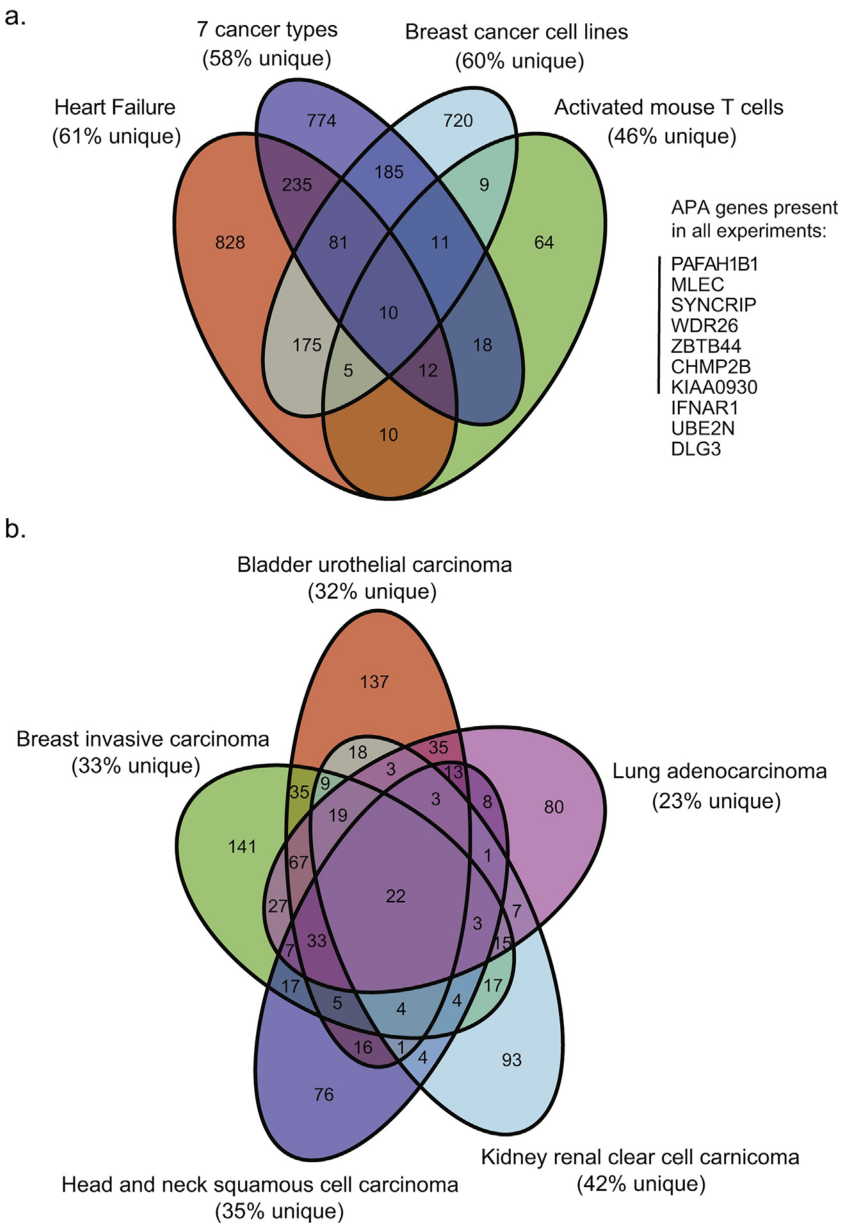


Fig. 3. APA is disease and condition specific.
Genome-wide APA genes are more likely to be unique than shared among disease states. **A)** Venn-diagram of overlapping gene sets from 4 different experimental datasets. The APA events determined by comparing test versus control samples from the failing human heart [91], 7 cancer types [51], 2 breast cancer cell lines (HER2 positive and negative) [92] and activated mouse T cells [90]. Only 10 genes were shared in APA genes in all experiments and these are listed to the right of the diagram. **B)** Venn-diagram of overlapping APA events from 5 of the 7 tumour types inferred from RNA-seq coverage [51]. The percentage of APA genes unique to each condition are indicated.

Acknowledgements

We thank members of the Beilharz Lab for constructive comment on the manuscript. RET was supported by an Australian Postgraduate scholarship, ADP was supported by a Monash Postgraduate scholarship and THB was supported by a Monash Biodiscovery Fellowship. Research in the Beilharz Lab was funded by grants from the NHMRC (APP1128250) and ARC (DP170100569).

References

[1] A. Lembo, F. Di Cunto, P. Provero, Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer, PLoS One 7 (2) (2012) e31129.

[2] Z. Ji, B. Tian, Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types, PLoS One 4 (12) (2009) e8419.

[3] J.Y. Park, W. Li, D. Zheng, P. Zhai, Y. Zhao, T. Matsuda, S.F. Vatner, J. Sadoshima, B. Tian, Comparative analysis of mRNA isoform expression in cardiac hypertrophy and development reveals multiple post-transcriptional regulatory modules, PLoS One 6 (7) (2011) e22391.

[4] O.K. Yoon, R.B. Brem, Noncanonical transcript forms in yeast and their regulation during environmental stress, RNA 16 (6) (2010) 1256–1267.

[5] P.J. Shepard, E.A. Choi, J. Lu, L.A. Flanagan, K.J. Hertel, Y. Shi, Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq, RNA 17 (4) (2011) 761–772.

[6] A.H. Beck, Z. Weng, D.M. Witten, S. Zhu, J.W. Foley, P. Lacroute, C.L. Smith, R. Tibshirani, M. van de Rijn, A. Sidow, R.B. West, 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples, PLoS One 5 (1) (2010) e8768.

[7] A. Derti, P. Garrett-Engele, K.D. Macisaac, R.C. Stevens, S. Sriram, R. Chen, C.A. Rohl, J.M. Johnson, T. Babak, A quantitative atlas of polyadenylation in five mammals, Genome Res. 22 (6) (2012) 1173–1183.

[8] F. Ozsolak, P. Kapranov, S. Foissac, S.W. Kim, E. Fishilevich, A.P. Monaghan, B. John, P.M. Milos, Comprehensive polyadenylation site maps in yeast and

human reveal pervasive alternative polyadenylation, Cell 143 (6) (2010) 1018–1029.

[9] S. Wilkening, V. Pelechano, A.I. Jarvelin, M.M. Tekkedil, S. Anders, V. Benes, L.M. Steinmetz, An efficient method for genome-wide polyadenylation site mapping and RNA quantification, Nucleic Acids Res. 41 (5) (2013) e65.

[10] M. Hoque, Z. Ji, D. Zheng, W. Luo, W. Li, B. You, J.Y. Park, G. Yehia, B. Tian, Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing, Nat. Methods 10 (2) (2013) 133–139.

[11] P.F. Harrison, D.R. Powell, J.L. Clancy, T. Preiss, P.R. Boag, A. Traven, T. Seemann, T.H. Beilharz, PAT-seq: a method to study the integration of 3'-UTR dynamics with gene expression in the eukaryotic transcriptome, RNA 21 (8) (2015) 1502–1510.

[12] B. Tian, Z. Pan, J.Y. Lee, Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing, Genome Res. 17 (2) (2007) 156–165.

[13] W. Li, B. You, M. Hoque, D. Zheng, W. Luo, Z. Ji, J.Y. Park, S.I. Gunderson, A. Kalsotra, J.L. Manley, B. Tian, Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation, PLoS Genet. 11 (4) (2015) e1005166.

[14] M. Chiabudini, C. Conz, F. Reckmann, S. Rospert, Ribosome-associated complex and Ssb are required for translational repression induced by polylysine segments within nascent chains, Mol. Cell. Biol. 32 (23) (2012) 4769–4779.

[15] S. Danckwardt, M.W. Hentze, A.E. Kulozik, 3' end mRNA processing: molecular mechanisms and implications for health and disease, EMBO J. 27 (3) (2008) 482–498.

[16] R.M. Denome, C.N. Cole, Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals, Mol. Cell. Biol. 8 (11) (1988) 4829–4839.

[17] B. Gawande, M.D. Robida, A. Rahn, R. Singh, Drosophila Sex-lethal protein mediates polyadenylation switching in the female germline, EMBO J. 25 (6) (2006) 1263–1272.

[18] Y. Shi, Alternative polyadenylation: new insights from global analyses, RNA 18 (12) (2012) 2105–2117.

[19] P.A. Pinto, T. Henriques, M.O. Freitas, T. Martins, R.G. Domingues, P.S. Wyrzykowska, P.A. Coelho, A.M. Carmo, C.E. Sunkel, N.J. Proudfoot, A. Moreira, RNA polymerase II kinetics in polo polyadenylation signal selection, EMBO J. 30 (12) (2011) 2431–2444.

[20] Z. Ji, W. Luo, W. Li, M. Hoque, Z. Pan, Y. Zhao, B. Tian, Transcriptional activity regulates alternative cleavage and polyadenylation, Mol. Syst. Biol. 7 (2011) 534.

[21] A. Swaminathan, T.H. Beilharz, Epitope-tagged yeast strains reveal promoter driven changes to 3'-end formation and convergent antisense-transcription from common 3' UTRs, Nucleic Acids Res. 44 (1) (2016) 377–386.

[22] M.L. Peterson, S. Bertolino, F. Davis, An RNA polymerase pause site is associated with the immunoglobulin s Poly(A) site, Mol. Cell. Biol. 22 (15) (2002) 5606–5615.

[23] T.N. Mavrich, I.P. Ioshikhes, B.J. Venters, C. Jiang, L.P. Tomsho, J. Qi, S.C. Schuster, I. Albert, B.F. Pugh, A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome, Genome Res. 18 (7) (2008) 1073–1083.

[24] S. Shivaswamy, A. Bhinge, Y. Zhao, S. Jones, M. Hirst, V.R. Iyer, Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, PLoS Biol. 6 (3) (2008) e65.

[25] N. Spies, C.B. Nielsen, R.A. Padgett, C.B. Burge, Biased chromatin signatures around polyadenylation sites and exons, Mol. Cell 36 (2) (2009) 245–254.

[26] W. Li, J.Y. Park, D. Zheng, M. Hoque, G. Yehia, B. Tian, Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control, BMC Biol. 14 (2016) 6.

[27] A.J. Wood, R. Schulz, K. Woodfine, K. Koltowska, C.V. Beechey, J. Peters, D. Bourc'his, R.J. Oakey, Regulation of alternative polyadenylation by genomic imprinting, Genes Dev. 22 (9) (2008) 1141–1146.

[28] N.J. Proudfoot, G.G. Brownlee, 3' Non-coding region seqeunces in eukaryotic messenger RNA, Nature 263 (1976) 211–214.

[29] G. Martin, A.R. Gruber, W. Keller, M. Zavolan, Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length, Cell Rep. 1 (6) (2012) 753–763.

[30] M.I. Zarudnaya, I.M. Kolomiets, A.L. Potyahaylo, D.M. Hovorun, Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures, Nucleic Acids Res. 31 (5) (2003) 1375–1386.

[31] M. Mangone, A.P. Manoharan, D. Thierry-Mieg, J. Thierry-Mieg, T. Han, S.D. Mackowiak, E. Mis, C. Zegar, M.R. Gutwein, V. Khivansara, O. Attie, K. Chen, K. Salehi-Ashtiani, M. Vidal, T.T. Harkins, P. Bouffard, Y. Suzuki, S. Sugano, Y. Kohara, N. Rajewsky, F. Piano, K.C. Gunsalus, J.K. Kim, The landscape of C. elegans 3'UTRs, Science 329 (5990) (2010) 432–435.

[32] N.J. Proudfoot, Ending the message: poly(A) signals then and now, Genes Dev. 25 (17) (2011) 1770–1782.

[33] N.M. Nunes, W. Li, B. Tian, A. Furger, A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence, EMBO J. 29 (9) (2010) 1523–1536.

[34] B. Tian, J. Hu, H. Zhang, C.S. Lutz, A large-scale analysis of mRNA polyadenylation of human and mouse genes, Nucleic Acids Res. 33 (1) (2005) 201–212.

[35] E. Beaudoing, S. Freier, J.R. Wyatt, J.M. Claverie, D. Gautheret, Patterns of variant polyadenylation signal usage in human genes, Genome Res. 10 (7) (2000) 1001–1010.

[36] A. Gil, N.J. Proudfoot, Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit ß-globin mRNA 3'end formation, Cell 49 (1987) 399–406.

[37] J. Hu, C.S. Lutz, J. Wilusz, B. Tian, Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation, RNA 11 (10) (2005) 1485–1493.

[38] J.H. Graber, C.R. Cantor, S.C. Mohr, T.F. Smith, In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species, Proc. Natl. Acad. Sci. U. S. A. 96 (24) (1999) 14055–14060.

[39] K. Venkataraman, K.M. Brown, G.M. Gilmartin, Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition, Genes Dev. 19 (11) (2005) 1315–1327.

[40] C. Yao, J. Biesinger, J. Wan, L. Weng, Y. Xing, X. Xie, Y. Shi, Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation, Proc. Natl. Acad. Sci. U. S. A. 109 (46) (2012) 18773–18778.

[41] J. Zhao, L. Hyman, C. Moore, Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis, Microbiol. Mol. Biol. Rev. 63 (2) (1999) 405–445.

[42] Y. Shi, D.C. Di Giammartino, D. Taylor, A. Sarkeshik, W.J. Rice, J.R. Yates, 3rd, J., frank, J.L., manley, molecular architecture of the human pre-mRNA 3' processing complex, Mol. Cell 33 (3) (2009) 365–376.

[43] C.R. Mandel, Y. Bai, L. Tong, Protein factors in pre-mRNA 3'-end processing, Cell. Mol. Life Sci. 65 (7-8) (2008) 1099–1122.

[44] Y. Takagaki, L.C. Ryner, J.L. Manley, Four factors are require for 3'-end cleavage of pre-mRNAs, Genes Dev. 3 (1989) 1711–1724.

[45] S. Chan, E.A. Choi, Y. Shi, Pre-mRNA 3'-end processing complex assembly and function, Wiley Interdiscip. Rev. RNA 2 (3) (2011) 321–335.

[46] D.F. Colgan, J.L. Manley, Mechanism and regulation of mRNA polyadenylation, Genes Dev. 11 (21) (1997) 2755–2766.

[47] S. Millevoi, S. Vagner, Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation, Nucleic Acids Res. 38 (9) (2010) 2757–2774.

[48] K. Xiang, L. Tong, J.L. Manley, Delineating the structural blueprint of the pre-mRNA 3'-end processing machinery, Mol. Cell. Biol. 34 (11) (2014) 1894–1910.

[49] Y. Takagaki, R.L. Seipelt, M.L. Peterson, J.L. Manley, The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation, Cell 87 (5) (1996) 941–952.

[50] Y. Takagaki, J.L. Manley, Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation, Mol. Cell 2 (6) (1998) 761–771.

[51] Z. Xia, L.A. Donehower, T.A. Cooper, J.R. Neilson, D.A. Wheeler, E.J. Wagner, W. Li, Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types, Nat. Commun. 5 (2014) 5274.

[52] C. Yao, E.A. Choi, L. Weng, X. Xie, J. Wan, Y. Xing, J.J. Moresco, P.G. Tu, J.R. Yates 3rd, Y. Shi, Overlapping and distinct functions of CstF64 and CstF64tau in mammalian mRNA 3' processing, RNA 19 (12) (2013) 1781–1790.

[53] B. Lackford, C. Yao, G.M. Charles, L. Weng, X. Zheng, E.A. Choi, X. Xie, J. Wan, Y. Xing, J.M. Freudenberg, P. Yang, R. Jothi, G. Hu, Y. Shi, Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal, EMBO J. 33 (8) (2014) 878–889.

[54] K.S.K. Guisbert, H. Li, C. Guthrie, Alternative 3' pre-mRNA processing in Saccharomyces cerevisiae is modulated by Nab4/Hrp1 in vivo, PLoS Biol. 5 (1) (2007) e6.

[55] T. Kubo, T. Wada, Y. Yamaguchi, A. Shimizu, H. Handa, Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3'-UTRs, Nucleic Acids Res. 34 (21) (2006) 6264–6271.

[56] S. Kim, J. Yamamoto, Y. Chen, M. Aida, T. Wada, H. Handa, Y. Yamaguchi, Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation, Genes Cells 15 (9) (2010) 1003–1013.

[57] A.R. Gruber, G. Martin, W. Keller, M. Zavolan, Cleavage factor Im is a key regulator of 3' UTR length, RNA Biol. 9 (12) (2012) 1405–1412.

[58] C.P. Masamha, Z. Xia, J. Yang, T.R. Albrecht, M. Li, A.B. Shyu, W. Li, E.J. Wagner, CFIm25 links alternative polyadenylation to glioblastoma tumour suppression, Nature 510 (7505) (2014) 412–416.

[59] M. Jenal, R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kuhn, F.M. Menzies, J.A. Oude Vrielink, A.J. Bos, J. Drost, K. Rooijers, D.C. Rubinsztein, R. Agami, The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites, Cell 149 (3) (2012) 538–553.

[60] S. McCracken, N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S.D. Patterson, M. Wickens, D.L. Bentley, The C-terminal domain of RNA polymerase II couples mRNA processing to transcription, Nature 385 (6614) (1997) 357–361.

[61] Y. Hirose, J.L. Manley, RNA polymerase II is an essential mRNA polyadenylation factor, Nature 395 (1998) 93–96.

[62] K. Glover-Cutter, RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes, Nat. Struct. Mol. Biol. 15 (1) (2008) 71–78.

[63] D.D. Licatalosi, G. Geiger, M. Minet, S. Schroeder, K. Cilli, J.B. McNeil, D.L. Bentley, Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II, Mol. Cell 9 (5) (2002) 1101–1111.

[64] B. Dichtl, D. Blank, M. Sadowski, W. Hubner, S. Weiser, W. Keller, Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination, EMBO J. 21 (15) (2002) 4125–4135.
[65] M. Sadowski, B. Dichtl, W. Hubner, W. Keller, Independent functions of yeast Pcf11p in pre-mRNA 3'end processing and in transcription termination, EMBO J. 22 (9) (2003) 2167–2177.
[66] T. Nagaike, C. Logan, I. Hotta, O. Rozenblatt-Rosen, M. Meyerson, J.L. Manley, Transcriptional activators enhance polyadenylation of mRNA precursors, Mol. Cell 41 (4) (2011) 409–418.
[67] Y. Yang, W. Li, M. Hoque, L. Hou, S. Shen, B. Tian, B.D. Dynlacht, PAF complex plays novel subunit-specific roles in alternative cleavage and polyadenylation, PLoS Genet. 12 (1) (2016) e1005794.
[68] K. Martincic, S.A. Alkan, A. Cheatle, L. Borghesi, C. Milcarek, Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing, Nat. Immunol. 10 (10) (2009) 1102–1109.
[69] D. Kaida, M.G. Berg, I. Younis, M. Kasim, L.N. Singh, L. Wan, G. Dreyfuss, U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation, Nature 468 (7324) (2010) 664–668.
[70] M.G. Berg, L.N. Singh, I. Younis, Q. Liu, A.M. Pinto, D. Kaida, Z. Zhang, S. Cho, S. Sherrill-Mix, L. Wan, G. Dreyfuss, U1 snRNP determines mRNA length and regulates isoform expression, Cell 150 (1) (2012) 53–64.
[71] A. Kyburz, A. Friedlein, L.H.W. Keller, Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing, Mol. Cell 23 (2006) 195–205.
[72] S. Millevoi, C. Loulergue, S. Dettwiler, S.Z. Karaa, W. Keller, M. Antoniou, S. Vagner, An interaction between U2AF 65 and CFIm links the splicing and 3' end processing machineries, EMBO J. 25 (2006) 4854–4864.
[73] M. Nazim, A. Masuda, M.A. Rahman, F. Nasrin, J.I. Takeda, K. Ohe, B. Ohkawara, M. Ito, K. Ohno, Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms, Nucleic Acids Res. 45 (3) (2017) 1455–1468.
[74] M. Movassat, T.L. Crabb, A. Busch, C. Yao, D.J. Reynolds, Y. Shi, K.J. Hertel, Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns, RNA Biol. 13 (7) (2016) 646–655.
[75] P. Castelo-Branco, A. Furger, M. Wollerton, C. Smith, A. Moreira, N. Proudfoot, Polypyrimidine tract binding protein modulates efficiency of polyadenylation, Mol. Cell. Biol. 24 (10) (2004) 4174–4183.
[76] R. Batra, K. Charizanis, M. Manchanda, A. Mohan, M. Li, D.J. Finn, M. Goodwin, C. Zhang, K. Sobczak, C.A. Thornton, M.S. Swanson, Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease, Mol. Cell 56 (2) (2014) 311–322.
[77] U. Kuhn, M. Gundel, A. Knoth, Y. Kerwitz, S. Rudel, E. Wahle, Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor, J. Biol. Chem. 284 (34) (2009) 22803–22814.
[78] J. Katahira, D. Okuzaki, H. Inoue, Y. Yoneda, K. Maehara, Y. Ohkawa, Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I, Nucleic Acids Res. 41 (14) (2013) 7060–7072.
[79] F.A. Bava, C. Eliscovich, P.G. Ferreira, B. Minana, C. Ben-Dov, R. Guigo, J. Valcarcel, R. Mendez, CPEB1 coordinates alternative 3'-UTR formation with translational regulation, Nature 495 (7439) (2013) 121–125.
[80] S. Danckwardt, A.S. Gantzert, S. Macher-Goeppinger, H.C. Probst, M. Gentzel, M. Wilm, H.J. Grone, P. Schirmacher, M.W. Hentze, A.E. Kulozik, p38 MAPK controls prothrombin expression by regulated RNA 3' end processing, Mol. Cell 41 (3) (2011) 298–310.
[81] V. Hilgers, S.B. Lemke, M. Levine, ELAV mediates 3' UTR extension in the Drosophila nervous system, Genes Dev. 26 (20) (2012) 2259–2264.
[82] K.D. Mansfield, J.D. Keene, Neuron-specific ELAV/Hu proteins suppress HuR mRNA during neuronal differentiation by alternative polyadenylation, Nucleic Acids Res. 40 (6) (2012) 2734–2746.
[83] W. Dai, G. Zhang, E.V. Makeyev, RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage, Nucleic Acids Res. 40 (2) (2012) 787–800.
[84] M. Samson, Evidence for 3'untranslated region-dependent autoregulation of the Drosophila gene encoding the neuronal nuclear RNA-binding protein ELAV, Genetics 150 (1998) 723–733.
[85] W. Luo, Z. Ji, Z. Pan, B. You, M. Hoque, W. Li, S.I. Gunderson, B. Tian, The conserved intronic cleavage and polyadenylation site of CstF-77 gene imparts control of 3' end processing activity through feedback autoregulation and by U1 snRNP, PLoS Genet. 9 (7) (2013) e1003613.
[86] B.R. Graverley, E.S. Fleming, G.M. Gilmartin, RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor, Mol. Cell. Biol. 16 (9) (1996) 4942–4951.
[87] M.J. Munoz, R.R. Daga, A. Garzon, G. Thode, J. Jimenez, Poly(A) site choice during mRNA 3'-end formation in the Schizosaccharomyces pombe wos2 gene, Mol. Genet. Genomics 267 (6) (2002) 792–796.
[88] A. Flomen, Increased RNA. editing in EAAT2 pre-mRNA from amyotrophic lateral sclerosis patients: involvement of a cryptic polyadenylation site, Neurosci. Lett. 497 (2) (2011) 139–143.
[89] R.S. Redis, L.E. Vela, W. Lu, J. Ferreira de Oliveira, C. Ivan, C. Rodriguez-Aguayo, D. Adamoski, B. Pasculli, A. Taguchi, Y. Chen, A.F. Fernandez, L. Valledor, K. Van Roosbroeck, S. Chang, M. Shah, G. Kinnebrew, L. Han, Y. Atlasi, L.H. Cheung, G.Y. Huang, P. Monroig, M.S. Ramirez, T. Catela Ivkovic, L. Van, H. Ling, R. Gafa, S. Kapitanovic, G. Lanza, J.A. Bankson, P. Huang, S.Y. Lai, R.C. Bast, M.G. Rosenblum, M. Radovich, M. Ivan, G. Bartholomeusz, H. Liang, M.F. Fraga, W.R. Widger, S. Hanash, I. Berindan-Neagoe, G. Lopez-Berestein, A.L. Ambrosio, S.M. Gomes Dias, G.A. Calin, Allele-specific reprogramming of cancer metabolism by the long non-coding RNA CCAT2, Mol. Cell 61 (4) (2016) 520–534.
[90] A.R. Gruber, G. Martin, P. Muller, A. Schmidt, A.J. Gruber, R. Gumienny, N. Mittal, R. Jayachandran, J. Pieters, W. Keller, E. van Nimwegen, M. Zavolan, Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells, Nat. Commun. 5 (2014) 5465.
[91] E.E. Creemers, A. Bawazeer, A.P. Ugalde, H.W. van Deutekom, I. van der Made, N.E. de Groot, M.E. Adriaens, S.A. Cook, C.R. Bezzina, N. Hubner, J. van der Velden, R. Elkon, R. Agami, Y.M. Pinto, Genome-wide polyadenylation maps reveal dynamic mRNA 3'-End formation in the failing human heart, Circ. Res. 118 (3) (2016) 433–438.
[92] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, A. Xu, Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing, Genome Res. 21 (5) (2011) 741–747.
[93] R. Sandberg, J.R. Neilson, A. Sarma, P.A. Sharp, C.B. Burge, Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites, Science 320 (5883) (2008) 1643–1647.
[94] C. Mayr, D.P. Bartel, Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells, Cell 138 (4) (2009) 673–684.
[95] L. Li, D. Wang, M. Xue, X. Mi, Y. Liang, P. Wang, 3'UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network, Sci. Rep. 4 (2014) 5406.
[96] A.R. Gruber, G. Martin, W. Keller, M. Zavolan, Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors, Wiley Interdiscip. Rev. RNA 5 (2) (2014) 183–196.
[97] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, C. Sotiriou, Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes, Clin. Cancer Res. 14 (16) (2008) 5158–5165.
[98] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J.F. Quackenbush, I.J. Stijleman, J. Palazzo, J.S. Marron, A.B. Nobel, E. Mardis, T.O. Nielsen, M.J. Ellis, C.M. Perou, P.S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (2009) 1160–1167.
[99] S. Lianoglou, V. Garg, J.L. Yang, C.S. Leslie, C. Mayr, Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression, Genes Dev. 27 (21) (2013) 2380–2396.
[100] C. Mayr, Evolution and biological roles of alternative 3'UTRs, Trends Cell Biol. 26 (3) (2016) 227–237.
[101] B.D. Berkovits, C. Mayr, Alternative 3' UTRs act as scaffolds to regulate membrane protein localization, Nature 522 (7556) (2015) 363–367.
[102] N. Spies, C.B. Burge, D.P. Bartel, 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts, Genome Res. 23 (12) (2013) 2078–2090.
[103] Y. Yoon, M.C. McKenna, D.A. Rollins, M. Song, T. Nuriel, S.S. Gross, G. Xu, C.E. Glatt, Anxiety-associated alternative polyadenylation of the serotonin transporter mRNA confers translational regulation by hnRNPK, Proc. Natl. Acad. Sci. U. S. A. 110 (28) (2013) 11624–11629.
[104] P. Oikonomou, H. Goodarzi, S. Tavazoie, Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts, Cell Rep. 7 (1) (2014) 281–292.
[105] J.R. Dickson, C. Kruse, D.R. Montagna, B. Finsen, M.S. Wolfe, Alternative polyadenylation and miR-34 family members regulate tau expression, J. Neurochem. 127 (6) (2013) 739–749.
[106] H. Rhinn, L. Qiang, T. Yamashita, D. Rhee, A. Zolin, W. Vanti, A. Abeliovich, Alternative alpha-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology, Nat. Commun. 3 (2012) 1084.
[107] V.A. Gennarino, C.E. Alcott, C.A. Chen, A. Chaudhury, M.A. Gillentine, J.A. Rosenfeld, S. Parikh, J.W. Wheless, E.R. Roeder, D.D. Horovitz, E.K. Roney, J.L. Smith, S.W. Cheung, W. Li, J.R. Neilson, C.P. Schaaf, H.Y. Zoghbi, NUDT21-spanning CNVs lead to neuropsychiatric disease and altered MeCP2 abundance via alternative polyadenylation, Elife 4 (2015).
[108] R. Soetanto, C.J. Hynes, H.R. Patel, D.T. Humphreys, M. Evers, G. Duan, B.J. Parker, S.K. Archer, J.L. Clancy, R.M. Graham, T.H. Beilharz, N.J. Smith, T. Preiss, Role of miRNAs and alternative mRNA 3'-end cleavage and polyadenylation of their mRNA targets in cardiomyocyte hypertrophy, Biochim. Biophys. Acta 1859 (5) (2016) 744–756.
[109] M.D. Barnhart, S.L. Moon, A.W. Emch, C.J. Wilusz, J. Wilusz, Changes in cellular mRNA stability, splicing, and polyadenylation through HuR protein sequestration by a cytoplasmic RNA virus, Cell Rep. 5 (4) (2013) 909–917.
[110] S. Tang, A. Patel, P.R. Krause, Herpes simplex virus ICP27 regulates alternative pre-mRNA polyadenylation and splicing in a sequence-dependent manner, Proc. Natl. Acad. Sci. U. S. A. 113 (43) (2016) 12256–12261.
[111] W.O. Miles, A. Lembo, A. Volorio, E. Brachtel, B. Tian, D. Sgroi, P. Provero, N. Dyson, Alternative polyadenylation in triple-negative Breast tumors allows NRAS and c-JUN to bypass PUMILIO posttranscriptional regulation, Cancer Res. 76 (24) (2016) 7231–7241.