Changes in alternative polyadenylation throughout the yeast metabolic cycle and during glucose starvation in *Saccharomyces cerevisiae*

Harriet Lester



September 20, 2023

Abstract

Alternative polyadenylation (APA) is a dynamic and powerful tool to alter gene expression and protein output to rapidly adapt to changes in cellular state. The Yeast Metabolic Cycle (YMC) and a shift from glucose to galactose (the Glu-Gal shift) are two highly metabolically active and dynamic process that are good candidates to explore the role of APA in complex biological systems. Using a combinatorial approach of 3'-RNA-Seq data, in house pipelines to acess APA and gene expression changes and bio-informatics tools to examine similarities and differences, the APA landscape of these processes was profiled and the functions and mechanisms behind its use; explored. The results showed that APA was highly prevalent in both the YMC and Glu-Gal shift and follow distinct temporal profiles. APA events for these processes predominantly occurred at the peaks of metabolic activity. In the YMC, global APA lengthening occurred at the peak of oxygen consumption and during the transition into the R/B phase with the genes involved showing GO enrichment for functional groups related to cell cycle progression and DNA repair. During the Glu-Gal shift, global shortening of mRNA transcripts at genes showing GO enrichment for functional groups related to metabolic processes was observed. There was little overlap of genes that

underwent APA events in the YMC and during the Glu-Gal shift, indicating that APA events are mostly specific to a biological process and furthermore, the genes that did go APA in both, showed more generic functions such as response to metabolic stress. In conclusion, APA is a potential regulator to adapt the transcriptome to the metabolic conditions in a cell. Moreover, by examining the context of APA events using Native Elongating Transcript Sequencing (NET-Seq), it become clear that a pas site transition from 3'UTR to the prematurely terminated transcription in the coding region (CR-APA) using transcripts starting at different promoters leads to transcriptional interference and repression of the main transcript. This may represent a novel mechanism to adapt transcriptional output from different promoters to the differing metabolic states in these cells. However, more granular assignment of PAS coupled to techniques such as Transcript IsoForm sequencing (TIF-Seq) to assess the integrity of individual transcripts is now required to confirm these observations on the diversity of APA in metabolic adaptations.

Acknowledgements

I am deeply grateful to my supervisors, Jane and Andre, for their dedicated guidance and support throughout my PhD journey. Their expertise and mentorship have been invaluable in shaping my research and helping me grow as a person. I would also like to express my heartfelt thanks to my fellow lab mates, Mark and Cillian, for their friendship and support. I must also extend my sincere gratitude to my family, I am forever thankful for their encouragement and belief in me, especially my mother who passed away during my studies. Despite her absence, I know that she would be proud of me for completing my PhD.

Contents

1 Introduction **33**

1.1 Pre-mRNA processing and structure 34

1.2 5'end capping 37

1.3 Splicing and the spliceosome complex 39

1.4 Splicing in *Saccharomyces cerevisiae* 42

1.5 Cleavage and polyadenylation in *Saccharomyces cerevisiae* . . 43

1.5.1 Conserved elements between yeast and mammalian 3'-mRNA end processing 46

1.6 Alternative Cleavage and polyadenylation 49

1.7 Regulation of APA 51

1.7.1 Active co-transcriptional trans regulators of APA . . . 53

1.7.2 Active co-transcriptional cis regulators of APA 55

1.7.3 Passive post transcriptional trans regulators of APA . 56

1.7.4 Passive post transcriptional cis regulators of APA . . . 59

1.7.5 Functional role of global regulation of APA 62

1.8 APA in *Saccharomyces cerevisiae* 65

1.9 Metabolism of carbon sources by *Saccharomyces cerevisiae* . . 69

1.9.1 Physiology and growth conditions of *Saccharomyces cerevisiae* in fermentation 71
1.9.2 Yeast fermentation and the Crabtree effect in *Saccharomyces cerevisiae* 73
1.9.3 The diauxic shift in *Saccharomyces cerevisiae* 75
1.10 The yeast metabolic cycle 77
1.11 Oscillations in transcript output during the yeast metabolic cycle 84
1.12 Glucose starvation in *Saccharomyces cerevisiae* 86
1.13 Thesis Objective 88

2 Materials and methods 89

2.1 Materials and Methods 89
2.1.1 Wet laboratory materials and methods 91
2.1.2 Yeast culture chemostat fermenter conditions for the YMC 91
2.1.3 Yeast strains for YMC and Glu-Gal shift RNA-Seq data 92
2.1.4 Yeast culture conditions for Glu-Gal shift 93
2.1.5 RNA extraction protocol for 3'RNASeq 93
2.1.6 QuantSeq 3'mRNA-Seq cDNA library preparation . . . 95
2.1.7 hsDNA Bioanalyzer assay QC 95
2.1.8 3' RNA-Seq Ion Proton chef 95
2.1.9 Native Elongating Transcript-seq (NET-Seq) data for RNA samples taken during the YMC 95

2.1.10 Native Elongating Transcript-seq (NET-Seq) data for RNA samples taken during the Glu-Gal shift 96

2.2 Bioinformatics methods for exploring APA profiles in saccharomyces cerevisiae 98

2.2.1 Characterisation of pas sites in saccharomyces cerevisiae 98

2.2.2 Selection of RNA-Seq data for APA and gene expression analysis 101

2.2.3 Pipeline for characterising the APA profile for the YMC 104

2.2.4 Bootstrapping data sets for statistical analysis 107

2.2.5 Cluster analysis of YMC APA events 108

2.2.6 Identifying Δpas lengths 109

2.2.7 Cross comparison of APA changes and gene expression clusters 110

2.2.8 IGV visualisation of APA changes in the YMC 111

2.2.9 Bioinformatics Pipeline 111

2.2.10 Integrative Genomics Viewer 113

2.2.11 Yeast Mine Gene Ontology Analysis 113

2.2.12 Statistical significance tests for GE and APA values . . 114

2.2.13 NET-seq/TEF-seq sequencing data processing 114

3 The APA profile of the Yeast metabolic cycle 116

3.1 Chapter aims 116

3.2 Gene expression data shows many significant changes in gene expression throughout the YMC 117

3.3 APA changes observed in the YMC peak at time points 3a and 3b and show preference for lengthening events. 123

3.4 The Coding APA profile of the YMC shows significant changes in APA at time points 3a and 3b skewed towards APA lengthening events 127

3.5 3'UTR events occur less frequently than CR-APA in the YMC 132

3.5.1 ALL APA data set shows global lengthening events dominate the APA profile of the YMC 136

3.5.2 Shifts in specific APA events correlate to changes in gene expression 142

3.6 Gene Ontology analysis shows several significant functional characteristics for genes that undergo APA changes in the YMC 149

3.6.1 Gene Ontology analysis for genes that undergo Coding APA changes in the YMC 149

3.6.2 Gene Ontology analysis for genes that undergo 3'UTR APA changes in the YMC 152

3.6.3 Gene Ontology analysis for genes that undergo ALL APA changes in the YMC 152

3.7 Cluster analysis of genes that undergo APA changes in the YMC show have few unique gene expression clusters 154

3.8 YMC gene expression clusters overlap with YMC APA data sets in GO enrichment at each super cluster 171

3.8.1 GO analysis of genes that are present in the gene expression clusters and APA data sets are enriched for terms that include fatty acid metabolism, DNA repair and translation 172

3.8.2 Δpas lengths significantly differ in genes with APA changes in the YMC, implying targeted pas site selection shifts in the YMC APA profile 174

3.9 NET-Seq and RNA-Seq data for YMC transcripts show similar gene expression profiles. 174

3.9.1 Net-Seq profiles for genes that undergo APA changes follow similar trends to the RNA-Seq data sets 178

3.10 TIF-Seq IGV tracks unveil several mechanistic differences between changes in pas site usage 180

3.10.1 *MIS1/YBR084W* undergoes 3'UTR shortening APA at time points 3a and 3b confirmed by TIF-Seq data . 182

3.10.2 IGV tracks of 3'RNA-Seq data show *YPR151C/SUE1* undergoes ALL APA shortening at time point 3a and 3b187

3.10.3 NETSeq IGV tracks of *YPR151C/SUE1* show a complex transcriptional environment that shows potential read through from neighbouring genes 187

3.10.4 TIF-Seq data supports 3'RNASeq data for *YPR151C/SUE1* using two pas sites 188

3.10.5 IGV Tracks for *YBR158W/AMN1* show ALL lengthening APA thoughout the YMC 193

3.10.6 NETSeq IGV tracks indicate transcriptional interference influences pas site selection of *YBR158W/AMN1* in the YMC 193

3.10.7 TIF-Seq IGV tracks of *YBR158W/AMN1* confirms NET-Seq data and suggests *SUT020* read through to *AMN1* 194

3.11 Discussion 199

3.11.1 3a and 3b most active time points for gene expression changes and shifts in APA 200

3.11.2 Skew analysis of APA events are not an artifact of large shifts in gene expression change 201

3.11.3 The YMC is hallmarked by global ALL APA lengthening events 202

3.11.4 APA lengthening and gene expression increases and APA shortening and gene expression decreases are correlated 203

3.11.5 GO analysis showed genes undergoing APA were enriched in functional groups that undergo key biological mechanisms in the YMC 205

3.11.6 Cluster analysis showed two to three distinct clusters of genes that undergo APA changes and overlap with pre-established gene expression clusters 206

3.11.7 Δpas lengths overall shorter in genes that undergo APA changes in the YMC 206

3.11.8 NET-seq profiles of nascent transcripts indicate that APA is active and due to changes in pas site selection in the trancriptome 207

3.11.9 IGV reveals a range of potential mechanisms of pas site selection occurring in the YMC 208

3.11.10 Further investigations 210

4 The APA profile of the Glu-Gal shift 214

4.1 Chapter Aims 214

4.2 Large changes in the transcriptome occur during the Glu-Gal shift 215

4.3 APA changes observed in the Glu-Gal shift show skew for shortening events and peak at time point Gal15 222

4.4 The Coding APA profile of the Glu-Gal Shift shows APA events peak at time point Gal15 226

4.5 The 3'UTR APA profile of the Glu-Gal Shift shows few APA events occur in the 3'UTR 230

4.5.1 The ALL APA profile of the Glu-Gal Shift shows a preference for APA lengthening events 234

4.5.2 APA Vs Gene expression changes for Glu-Gal shift . . 239

4.6 Gene Ontology analysis for genes that undergo APA changes during Glu-Gal shift 246

4.6.1 Gene Ontology analysis for genes that undergo Coding APA changes for Glu-Gal shift 246

4.6.2 Gene Ontology analysis for genes that undergo 3'UTR APA changes for Glu-Gal shift 247
4.6.3 Gene Ontology analysis for genes that undergo ALL APA changes throughout the Glu-Gal shift 247
4.6.4 Genes undergoing shortening APA events throughout Glu-Gal shift have increased Δpas lengths 249
4.7 Discussion 251
4.7.1 Gene expression changes in the Glu-Gal shift peaks at time point Gal15 251
4.7.2 APA occurred most often at time point Gal15 in the ALL data set 252
4.7.3 APA events during Glu-Gal shift have a small preference for lengthening 253
4.7.4 Small correlation observed between Glu-Gal shift APA genes and changes in gene expression 254
4.7.5 Gene Ontology analysis shows enrichment for functional groups involved in cell cycle progression and glucose metabolism 255
4.7.6 The Δpas of genes that undergo APA shortening are significantly longer than non APA genes with multiple pas sites 257
4.7.7 Further investigations 257

5 The Comparison of APA events between the YMC and Glu-Gal shift 259

5.1 Chapter aims 259
5.2 Few APA events occur in both the YMC and Glu-Gal shift with few genes showing GO enrichment 260
5.2.1 Few genes undergo Coding APA in both the YMC and Glu-Gal shift 261
5.2.2 Few genes undergo 3'UTR APA in both the YMC and Glu-Gal shift 263
5.2.3 Several genes undergo ALL APA in both the YMC and Glu-Gal shift 265
5.2.4 NETSeq and 3'RNA-Seq IGVs for *SUE1* for Glu-Gal and YMC show similar profiles at glucose starvation . . 269
5.2.5 NET-Seq IGV shows *MIS1* is expressed in high glucose and repressed in galactose 271
5.3 Discussion 273
5.3.1 Further investigations 276
5.4 Concluding remarks 276

6 Appendix **279**

6.1 Table of descriptive statistics of the gene expression changes in the YMC 280
6.1.1 Summary table of GO analysis results for genes that undergo 3'UTR APA changes in the YMC 282
6.2 Table of APA genes that overlap with gene expression clusters 284
6.3 Table of APA genes that overlap with gene expression clusters (percentages) 286

6.3.1 Heatmap of NET-Seq log2fold gene expression changes for genes that undergo Coding APA in YMC 287

6.3.2 GO analysis of Coding APA clusters 288

6.3.3 GO analysis of 3'UTR APA clusters 288

6.3.4 Heatmap of NET-Seq log2fold gene expression changes for genes that undergo 3'UTR APA in YMC 288

6.3.5 Heatmap of NET-Seq log2fold gene expression changes for genes that undergo ALL APA in YMC 291

6.4 Spearman's rank correlation coefficent matrix for coding shortening events during the YMC 292

6.5 Spearman's rank correlation coefficent matrix for coding lengthening events during the YMC 293

6.6 Spearman's rank correlation coefficent matrix for 3'UTR shortening events during the YMC 294

6.7 Spearman's rank correlation coefficent matrix for 3'UTR lengthening events during the YMC 295

6.8 Spearman's rank correlation coefficent matrix for ALL shortening events during the YMC 296

6.9 Spearman's rank correlation coefficent matrix for ALL lengthening events during the YMC 297

6.10 Summary table of GO analysis results for genes that undergo coding APA changes during Glu-Gal shift 298

6.11 Summary table of GO analysis results for genes that undergo 3'UTR APA changes during Glu-Gal shift 300

6.11.1 Table of GO analysis for genes that undergo Coding APA events in both the YMC and the Glu-Gal shift . . 302

List of Tables

1 List of Abbreviations (continued) 30

2.1 Table breaking down PAS site location and mean distance between PAS sites (Δpas) in saccharomyces cerevisiae. 101

2.2 RNA-Seq time points along the YMC that were chosen for deseq/dexseq pipeline analysis. Phases in the YMC are defined by the oxygen consumption graph, as described in the introduction. 104

3.1 Table showing gene expression change events throughout the YMC for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC. A gene expression increase was categorised as an event of change of normalised reads per million that was less than 0.05 padj in statistical significance. 118

3.2 Table of Coding APA events throughout the YMC 129

3.3 Table of 3'UTR APA events summary 133

3.4 Table of ALL APA events throughout the YMC 138

3.5 Table showing a small preference of genes undergoing both APA and Differential gene expression events. % of APA events occur in genes that undergo differential expression and those that do not undergo differential expression. Table of the number of APA events that occur in both genes with differential expression changes throughout the YMC and genes that do not change expression throughout the YMC. Table shows DE = Differential expression, % DE is the % of genes that undergo APA events vs the total number of genes in that data set, total is the total number of APA events. 144

3.6 p-values for log fold changes in gene expression for genes that undergo APA changes vs genes that do not undergo any APA shifts. p-values highlighted in red have significance under 0.005. 145

4.1 Summary table for gene expression events throughout the Glu-Gal shift 222

4.2 Summary table of number of genes that undergo Coding APA 227

4.3 Summary table of number of genes that undergo 3'UTR APA duing the Glu-Gal shift 231

4.4 Summary table of number of genes that undergo ALL APA during the Glu-Gal shift 235

4.5 Table of APA genes throughout the Glu-Gal shift showing a small preference of genes undergoing both APA and Differential gene expression events. % of APA events occur in genes that undergo differential expression and those that do not undergo differential expression. Table of the number of APA events that occur in both genes with differential expression changes throughout the Glu-Gal shift and genes that do not change expression. The table shows DE = Differential expression, % DE is the % of genes that undergo APA events vs the total number of genes in that data set, total is the total number of APA events. 240

4.6 p-values of t-tests of genes that undergo different types of APA and changes in gene expression at different time points throughout the Glu-Gal shift. Numbers highlighted in red are statistically significant with a p-value lower than 0.05 240

4.7 Table of mean Δpas lengths of genes that undergo APA during the Glu-Gal shift 249

5.1 Table of GO analysis for genes that undergo ALL APA events in both the YMC and the Glu-Gal shift. 268

6.1 Descriptive statistics of the gene expression changes in the YMC281

6.2 table of APA genes that occur in each of the 16 gene expression YMC clusters. Clusters made by Shidong Xi, in the Mellor laboratory, Oxford university Biochemistry department. 285

6.3 Table showing the percentage of genes that are found in gene expression clusters that also undergo different types of APA changes. Data sets with an overlap greater than 10% are shown in red. 286

List of Figures

1.1 Co-transcriptional nascent mRNA processing events in yeast . 36

1.2 Diagram of basic eukaryotic mRNA 37

1.3 Diagram of 5'end cap 38

1.4 Alternative splicing 40

1.5 Schematic of sequences in pre-mRNA involved in splicing . . . 42

1.6 Map of core signal sequences on nascent yeast transcripts . . . 46

1.7 Diagram of yeast cleavage and polyadenylation machinery . . 48

1.8 Diagram of CR-APA and 3'UTR-APA types 50

1.9 Types of APA regulation 52

1.10 APA regulation via RBPs 58

1.11 APA silencing of *CSTF77* 60

1.12 Aerobic and Anaerobic metabolic pathways 70

1.13 Schematic of chemostat fermenter 72

1.14 Yeast fermentation metabolic mechanism 74

1.15 Glucose control of the GAL network in *saccharomyces cerevisiae* 76

1.16 Model for the molecular mechanism behind the yeast metabolic cycle 81

1.17 Oxgygen trace and gene expression changes in superclusters along the YMC. 82

1.18 Summary of 3 main super clusters of the YMC and physiological events that co-occur at these time points 83

2.1 experimental design flowchart 90

2.2 Histogram of the number of pas sites in the saccharomyces cerevisiae genome per gene 100

2.3 Quantseq 3-end read generation protocol 102

2.4 RNA-Seq time points along the YMC 103

2.5 PAS site bins for APA pipeline visualized on IGV track 106

3.1 Scatter plot of YMC time point log fold2 change values of gene expression changes. The plot shows logfold change in normalised reads per million of each gene vs the number of normalized reads for expression change events throughout the YMC for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC.. Significant values of gene expression change were identified as differences in normalised reads per million compared to control that had a padj of less than 0.05. Significant changes are shown in red whilst nonsignificant values are shown in black. 120

3.2 Violin plot of YMC time point logfold2 change values of gene expression changes 122

3.3 Bar chart comparing GE increases and decreases throughout the YMC 123

3.4 Heat map of logfold2 changes in normalized reads per million of genes throughout the YMC. 124

3.5 Plot of summary data of number of significant events for Coding region, 3'UTR and ALL APA that occurs during the YMC. 126

3.6 Bar chart of APA coding changes throughout the YMC 129

3.7 Heat map of the padj values for APA coding changes throughout the YMC 130

3.8 Scatter plot of gene expression vs APA events for 3b time point of the Coding data set 131

3.9 Bar chart for 3'UTR APA changes throughout the YMC showing most events occur at time points 3a and 3b with a preference for APA lengthening events. Light green bars show APA shortening and dark green bars show APA lengthening 133

3.10 Heat map of the padj values for APA 3'UTR changes throughout the YMC 134

3.11 Scatter plot of gene expression vs APA events for 3b time point of the 3'UTR data set 135

3.12 Bar chart of ALL APA events throughout the YMC 138

3.13 Heat map of the padj values for ALL APA changes throughout the YMC 139

3.14 Scatter plot of gene expression vs APA events for 3b time point of the ALL data set 140

3.15 Bar chart showing the number of APA events in the YMC that undergo differential expression 143

3.16 Venn diagram of total Coding APA changes vs gene expression changes throughout the YMC and a table which shows the % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases 146

3.17 Venn diagram of total 3'UTR APA changes vs gene expression changes throughout the YMC and a table which shows the % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases 147

3.18 Venn diagram of total ALL APA changes vs gene expression changes throughout the YMC and a table which shows the % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases 148

3.19 Summary table of GO analysis results for genes that undergo coding APA changes in the YMC 151

3.20 Summary table of GO analysis results for genes that undergo ALL APA changes in the YMC 153

3.21 Summary of all APA clusters in the YMC 156

3.23 APA clusters in the YMC 158

3.24 Cluster analysis of genes that undergo coding shortening APA changes in the YMC 160

3.25 Cluster analysis of genes that undergo coding lengthening APA changes in the YMC 162

3.26 Cluster analysis of genes that undergo 3'UTR shortening APA changes in the YMC 164

3.27 Cluster analysis of genes that undergo 3'UTR lengthening APA changes in the YMC 166
3.28 Cluster analysis of genes that undergo ALL shortening APA changes in the YMC 168
3.29 GO analysis of ALL APA Shortening clusters 168
3.30 Cluster analysis of genes that undergo ALL lengthening APA changes in the YMC 170
3.31 GO analysis of ALL APA Lengthening clusters 170
3.32 GO analysis of genes that undergo APA changes and are included in gene expression clusters 173
3.33 Scatter plot of Coding, 3'UTR and ALL APA Δpas lengths . . 175
3.34 NET-Seq and RNA-Seq gene expression profiles for YMC data sets 177
3.35 Heat maps, 3'UTR and ALL APA gene expression changes from NET-Seq data 179
3.36 IGV tracks for *MIS1/YBR84W* 183
3.37 Line graphs of distal and proximal pas shift in gene expression for *MIS1/YBR84W* 184
3.38 NETSeq IGV Tracks for *MIS1/YBR84W* 185
3.39 TIF-Seq data for *MIS1/YBR84W* 186
3.40 IGV tracks for *YPR151C/SUE1* 189
3.41 Line graphs of distal and proximal pas shift in gene expression for *YPR151C/SUE1* 190
3.42 NETSeq IGV Tracks for *YPR151C/SUE1* 191
3.43 3'RNASeq and TIF-Seq data IGV tracks for *YPR151C/SUE1* 192

3.44 IGV tracks for *YBR158W/AMN1* gene 195

3.45 Line graphs of distal and proximal pas shift in gene expression for *YBR158W/AMN1* 196

3.46 NETSeq IGV Tracks for *YBR158W/AMN1* 197

3.47 TIF-Seq data IGV tracks for *YBR158W/AMN1* 198

3.48 Regulation of *tgp1* expression by premature termination of *nc-tgp1* transcription by APA driven proximal pas site usage. 213

4.1 Heat map of logfold2 changes in normalized reads per million of genes throughout the glucose to galactose metabolic shift . 218

4.2 Bar chart showing gene expression changes throughout the Glu-Gal shift 219

4.3 Scatter plot of log fold change values of gene expression changes of genes for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the glucose to galactose metabolic shift. The plot shows logfold change in normalised reads per million of each gene vs the number of normalized reads. Significant values of gene expression change were identified as differences in normalised reads per million compared to the control that had a padj of less than 0.05. Significant changes are shown in red whilst nonsignificant values are shown in black. 220

4.4 Violin plot of log fold change values of gene expression changes for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the glucose to galactose metabolic shift. . . . 223

4.5 Plot of summary data of number of significant events for Coding region, 3'UTR and ALL APA during the Glu-Gal shift. . . 225
4.6 Bar chart of APA coding changes throughout Glu-Gal shift . . 227
4.7 Heat map of the padj values for APA coding changes throughout the Glu-Gal shift 228
4.8 logfc vs gene expression of coding APA Glu-Gal shift 229
4.9 Bar chart for 3'UTR APA changes throughout the Glu-Gal shift. Dark green is APA shortening events, light green is APA lengthening. 231
4.10 Heat map of the padj values for 3'UTR APA changes throughout the Glu-Gal shift. Darker colours represent smaller p-value and, therefore more statistically significant APA events. . . . 232
4.11 logfc vs gene expression of 3'UTR APA Glu-Gal shift 233
4.12 Bar chart values for ALL APA changes throughout the Glu-Gal shift. Light Orange indicates APA lengthening events, dark orange is APA shortening. 235
4.13 Heatmap of the PADJ values for ALL APA changes throughout the Glu-Gal shift. Darker colours represent smaller p-value, therefore more statistically significant APA events. . . . 236
4.14 logfc vs gene expression of ALL APA Glu-Gal shift 237
4.15 Venn diagram of expression data from all the points of the YMC showing Coding APA changes vs gene expression changes throughout the Glu-Gal shift and a table which shows the number and % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases. . . . 241

4.16 Venn diagram of expression data from all the points of the YMC of total UTR APA changes vs gene expression changes throughout the Glu-Gal shift and a table which shows the number and % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases. . . . 243

4.17 Venn diagram of expression data from all the points of the YMC showing total ALL APA changes vs gene expression changes throughout the Glu-Gal shift and a table which shows the number and % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases. . . 245

4.18 Summary table of GO analysis results for genes that undergo ALL APA changes during Glu-Gal shift 248

4.19 Scatter plot of coding, 3'UTR and ALL APA Δpas lengths . . 250

5.1 Venn diagram of coding APA events during Glu-Gal shift and YMC 262

5.2 Venn diagram of 3'UTR APA events during Glu-Gal shift and YMC 264

5.3 Venn diagram of ALL APA events during Glu-Gal shift and YMC 266

5.4 NETSeq and 3'RNA-Seq IGVs of *SUE1* show similar expression profiles for glucose starvation. 270

5.5 NET-Seq IGV shows *MIS1* is majoritively expressed in high glucose and repressed in galactose 272

6.1 Summary table of GO analysis results for genes that undergo 3'UTR APA changes in the YMC 283

6.2 Heat map of NET-Seq log2fold gene expression changes for genes that undergo Coding APA in YMC 287

6.3 GO analysis of coding APA Shortening clusters 288

6.4 GO analysis of coding APA Lengthening clusters 289

6.5 GO analysis of 3'UTR APA Lengthening clusters 289

6.6 GO analysis of 3'UTR APA Shortening clusters 290

6.7 Heat map of NET-Seq log2fold gene expression changes for genes that undergo 3'UTR APA in YMC 290

6.8 Heat map of NET-Seq log2fold gene expression changes for genes that undergo ALL APA in YMC 291

6.9 Spearmans correlation coefficent matrix of genes undergoing coding shortening APA changes in the YMC 292

6.10 Spearmans correlation coefficent matrix of genes undergoing coding lengthening APA changes in the YMC 293

6.11 Spearmans correlation coefficent matrix of genes undergoing 3'UTR shortening APA changes in the YMC 294

6.12 Spearmans correlation coefficent matrix of genes undergoing 3'UTR lengthening APA changes in the YMC 295

6.13 Spearmans correlation coefficent matrix of genes undergoing ALL shortening APA changes in the YMC 296

6.14 Spearmans correlation coefficent matrix of genes undergoing ALL lengthening APA changes in the YMC 297

6.15 Summary table of GO analysis results for genes that undergo coding APA changes during Glu-Gal shift 299
6.16 Summary table of GO analysis results for genes that undergo 3'UTR APA changes during Glu-Gal shift 301
6.17 Table of GO analysis for genes that undergo Coding APA events in both the YMC and the Glu-Gal shift. 302

Table 1: List of Abbreviations (continued)

Abbreviation	**Definition**
AG	Adenine-Guanine
APA	Alternative Polyadenylation
APT	Associated with Pta1
AREs	Adenylate Uridylate Elements
C-APA	Coding APA
CPE	Cytoplasmic Polyadenylation Element
CFIA	Cleavage Factor IA
CFIB	Cleavage Factor IB
Clp1	Clp1
CPSF	Cleavage and Polyadenylation Specificity Factor
CPSF73	Cleavage and Polyadenylation Specificity Factor 73
CR-APA	Coding Region-Alternative Polyadenylation
CstF	Cleavage stimulation factor
CTD	C-Terminal Domain
CT-DNA	Cisplatin-DNA
DE	Differential gene expression
dO2	Dissolved Oxygen Concentration
DNA	Deoxyribonucleic Acid
EE	Efficiency Element
Glu-Gal	Glucose to Galactose shift
GO	Gene Ontology
GREs	GU-rich Elements

Table 1: List of Abbreviations (continued)

Abbreviation	**Full Description**
GU	Guanosine-Uracil (a dinucleotide)
hnRNPs	Heterogeneous Nuclear Ribonucleoproteins
HOC	High Oxygen Consumption Phase
IRES	Internal Ribosome Entry Site
LOC	Low Oxygen Consumption Phase
mRNA	Messenger RNA
miRNAs	MicroRNAs
m7G	7-Methyl-Guanosine
NADH	Nicotinamide Adenine Dinucleotide, Reduced Form
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
ORF	Open Reading Frame
OX phase	high oxygen phase
PAP1	Polymerase PAP1
pas site	Polyadenylation Site
PE	A-rich positioning element (PE) - PE signal
PE	Positioning Element
RBPs	RNA-binding proteins
POL II	RNA polymerase II
R/B phase	Reductive building phase
R/C phase	Reductive charging phase
RNAPII	RNA Polymerase II
RNA	Ribonucleic Acid

Table 1: List of Abbreviations (continued)

Abbreviation	**Full Description**
RBBP6	Retinoblastoma Binding Protein 6
Rna14	RNA14
Rna15	RNA15
Set1	Set Domain Containing 1
Set2	Set Domain Containing 2
TNF	Tumor Necrosis Factor
UAS G	Upstream activating sites
UTR	Untranslated Region
URS	Upstream repression sites
WDR33	WD Repeat Domain 33
YMC	Yeast Metabolic Cycle
YPD	Yeast Extract-Peptone-Dextrose
3'UTR APA	3' Untranslated region APA

Chapter 1

Introduction

The expression of genes is a highly dynamic and tightly regulated process. Organisms must be able to quickly adapt their transcriptome in order to respond to changes in cellular states such as stress, growth and differentiation. One such method is the variable polyadenylation site (pas site) usage, called alternative polyadenylation (APA). Pas sites are sequence motifs, usually located in the 3'UTR of a nascent mRNA transcript and are the site at which cleavage and polyadenylation occur, therefore their usage dictates the mRNA transcript length. Proximal pas site usage creates truncated mRNA transcripts while preference for distal pas sites will produce longer isoforms of the mRNA transcripts. Altering the length of a transcript can facilitate a multitude of changes including transcript stability, localisation and the resulting protein structure. Roughly 70% of coding genes of the yeast *Saccharomyces cerevisiae* contain multiple polyA sites, that can be subjected to APA (Pelechano, W. Wei, and Steinmetz 2013). The ubiquitous nature of these polyA sites make it a potentially powerful tool of gene expression

control. The possible alternative usages of pas sites have been observed in many different contexts, however it is still poorly understood. In order to explore the capacity of APA to induce changes to transcriptome, I chose a biological process that is both exceptionally transcriptionally active and also requires precise temporal control, the yeast metabolic cycle (YMC). For this, I investigated the use of different pas sites throughout the YMC, because it is a highly dynamic process, throughout which the expression of over half of the genome is effected. This process is likely to be tightly regulated as temporal gene expression is evident as yeast progress throughout the cycle. The YMC is characterized by the spontaneous and robust oscillation of metabolism. This is observed by measures of oxygen consumption under nutrient limited conditions. The cells go through three major metabolic shifts which are tightly coupled to cell cycle progression and proliferation. As APA facilitates an additional level of gene expression control beyond basal transcription; it is plausible it plays an important part in controlling the expression of key genes. If APA changes in response to different cellular states, it could help us to unravel the function of APA.

1.1 Pre-mRNA processing and structure

Before mRNA can be translated into protein, it must be processed from its nascent (pre-mRNA) form via modifications that occur co-and post transcriptionally. The co-transcriptional processing events which include 5'capping, splicing and cleavage and polyadenyaltion (Mischo and Proudfoot 2013). These processes are distinct from transcription termination (Proudfoot, Furger, and Dye 2002), however processes are tightly coupled to transcription, as

many of these modifications to the transcript require interactions and recruitment of proteins via the RNA polymerase II C-terminal domain (CTD) (Custódio and Carmo-Fonseca 2016). As soon as the RNA polymerase begins transcribing the nascent pre-mRNA, the capping complex is recruited to the 5' end to begin the process of adding a methylguanosine cap. This stabilises the transcript and blocks degradation from 5'-3' Exoribonucleases and helps recruitment of the ribosome from the cytoplasm (Kai and Pasquinelli 2010). As the RNA polymerase moves along the gene, the splicosome cuts out the non-coding introns from the elongating nascent transcript in a process called splicing (Green 1986). The cleavage and polyadenylation machinery is then recruited via interactions with the Poll II CTD and cis acting elements that are in close proximity to the cleavage site at the 3' end of the pre-mRNA (Phatnani and Greenleaf 2006). The transcript is then cleaved and a poly Adenosine tail is added to the 3'end of the maturing pre-mRNA. The concurrent nature of these events is demonstrated in the diagram 1.1 adapted from (Desterro, Bak-Gordon, and Carmo-Fonseca 2020) on page 36.

Protein coding genes are transcribed by RNA polymerase II. The output of mRNA processing is a mature transcript which is exported from the nucleus in order to be translated into a functional protein. The mature mRNA consists of an open reading frame (ORF), that is flanked by the untranslated regions, the 3'UTR and 5'UTR (Sieber, Platzer, and S. Schuster 2018), (Dvir et al. 2013). The ORF is the sequence that occurs between the start and stop codons and contains the information that encodes the protein sequence. The 3'UTR is the untranslated region of the mRNA that starts after the stop codon 3' encompasses the region from the stop codon to the polyA tail.

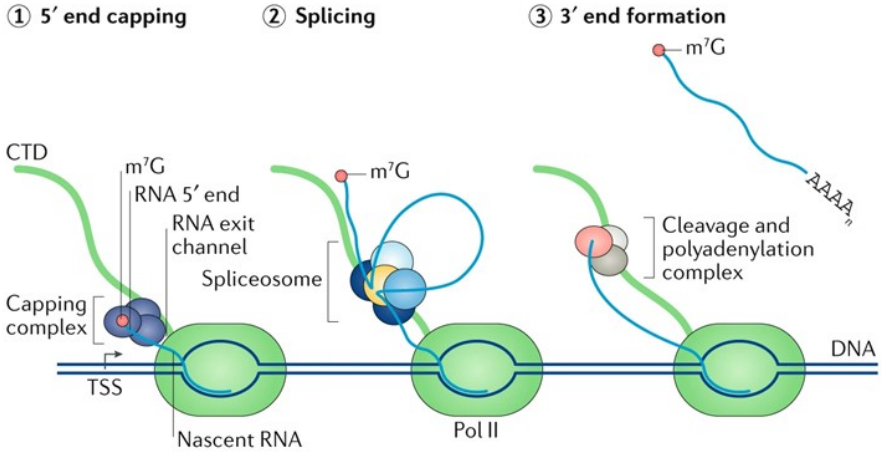


Figure 1.1: Diagram showing the co-transcriptional modification events that process nascent yeast transcripts into functional mRNA that is exported to the cytoplasm for translation into protein. 1. is the process of 5' end capping, in which a 5' methylguanosine cap is added, 2. is splicing, which is the excision of the introns from the nascent transcript. 3. is the 3' end formation, which adds the polyA tail to the end of pre-mRNA, adapted from (Desterro, Bak-Gordon, and Carmo-Fonseca 2020)

This region is crucial for the stability and localisation of many transcript as it can harbour multiple regulatory cis elements that can bind to proteins and siRNAs that can influence its fate (Szostak and Gebauer 2013). At the end of the 3'UTR, there is a polyadenosine tail which in *saccharomyces cerevisiae*, is roughly 50 nucleotides in length (Tudek et al. 2021). It contains many structural elements (see figure 1.2, including RNA G-quadruplexes, that can influence translation initiation via steric interactions and protein binding to provide specific and dynamic control over gene expression (Dvir et al. 2013). The basic structure of mRNA is summarised in the diagram 1.2 adapted from (S. L. Schuster and Hsieh 2019) on page 37.

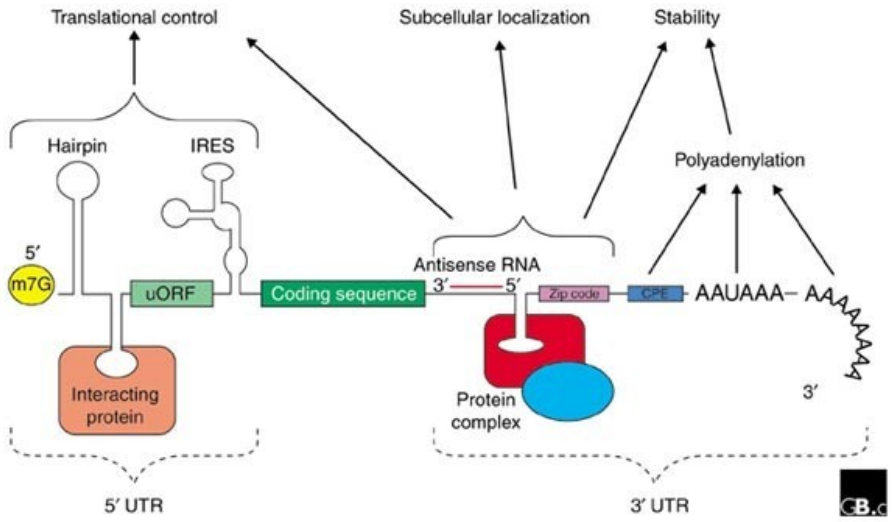


Figure 1.2: Diagram of basic eukaryotic mRNA, with cis and trans acting gene expression regulatory elements. Abbreviations : 3'UTR, untranslated region; m7G, 7-methyl-guanosine cap; hairpin, hairpin-like secondary structures; uORF, upstream open reading frame; IRES, internal ribosome entry site; CPE, cytoplasmic polyadenylation element AAUAAA, polyadenylation signal. Adapted from (S. L. Schuster and Hsieh 2019)

1.2 5'end capping

The initial mRNA modification is the addition of the 7' methyl guanosine cap to the 5' end of the pre-mRNA. This occurs co-transcriptionally, when the first 20-30 bases have been transcribed. The 5' cap stabilizes the nascent transcript and prevents its premature decay. The cap is also essential for translation initiation as it binds to the ribosomes via the translation initiation factor eIF4E that directly binds to the cap (Cowling 2010). Capping the result of three enzymatic reactions, involving the sequential action of RNA triphosphatase (Cet1) catalysed by the RNA guanylyltransferase (Ceg1) RNA triphosphatase (Cet1) and mRNA cap guanine-N7 methyltransferase, (abd1). In yeast, three enzymes form a heterodimeric complex that

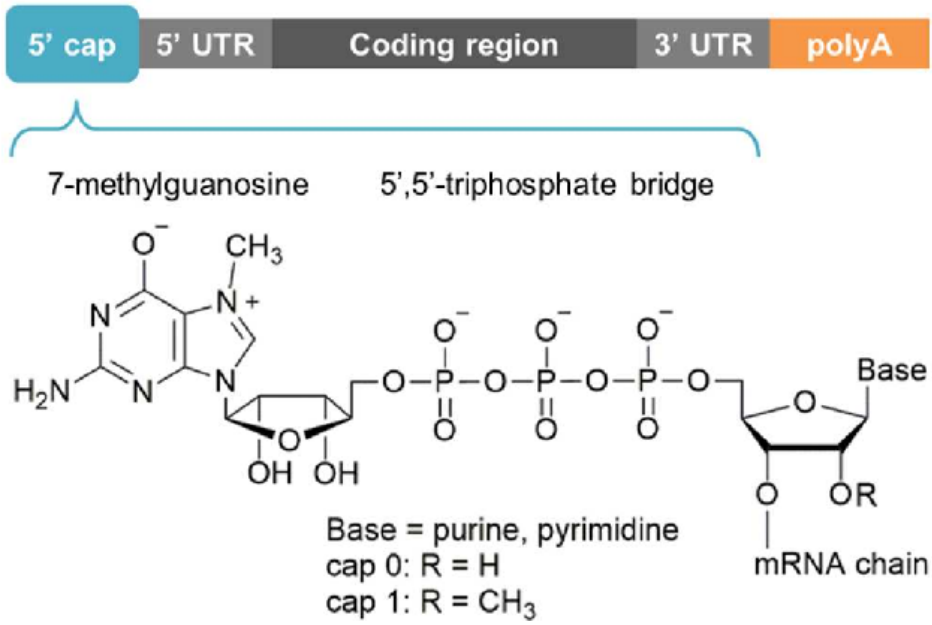


Figure 1.3: Diagram of 5'end m 7 G cap on mature mRNA. Structural diagram of all the molecular bonds between the 5' methylguanosine cap, the 5-5 triphosphate bridge and the mRNA starting base. Taken from from (Warminski et al. 2017).

associate with the serine 5' phosphorylated pol II CTD (Bharati et al. 2016). Cet1 removes the terminal phosphate on the terminal residue of the mRNA, Ceg1 fuses a GMP moiety via an unusual 5'5' phosphate linkage. This is followed by the transfer of a methyl group transfer to guanosine N7 S-adenosyl-L-methonine used as a methyl donor (Cowling 2010). Capping is completed by the association of the cap binding complex with the N7 G. Much of the translation initiation regulation that occurs via the 5'UTR occurs in a cap dependant manner (Merrick 2004). The molecular structure of the 5' cap in context to the full mRNA structure is shown in the diagram 1.3 adapted from (Warminski et al. 2017) on page 38.

1.3 Splicing and the spliceosome complex

One essential step in mRNA maturation is the removal of introns via splicing. Introns are non coding sequences present in pre-mRNA transcripts that interrupt the coding sequences (exon) and are removed during RNA processing (Cartegni, Chew, and Krainer 2002). The function of interrupted coding sequences is to greatly increase the variety of protein output from a single gene via alternative splicing (Kelemen et al. 2013) shown in diagram 1.4 taken from (Cartegni, Chew, and Krainer 2002) on page 40. By altering the combination of introns/exons being spliced out or retained, you produce a range of mRNA isoforms can be created that vastly increase the cellular proteosome. This variability of transcript output is referred to as alternative splicing, and is demonstrated in diagram 1.4 on page 42 taken from (Matera and Z. Wang 2014).

Splicing is performed via a large multisubunit ribonucleoprotien complex called the splicosome which is an assembly of roughly 80 different protein molecules and 5 small uridine rich small nucleae RNA protiens (UsnRNPS). The spliceosome is a dynamic and fluid assembly of proteins and RNA binding networks that manipulate the primary mRNA sequence and catalyse intron removal by inducing two successive transesterifcation reactions. During this process, 5 ribonucleoproteinsUsnRNPs (U1, U2, U4, U5 and U6) aided by protein splice factors play a key role in the identifications of intron exon borders and the assembly and formation of the catalytic spliceosome (Clancy et al. 2008).

Serine arginine rich proteins (SR protiensproteins) an hetero-ribonucleoproteins

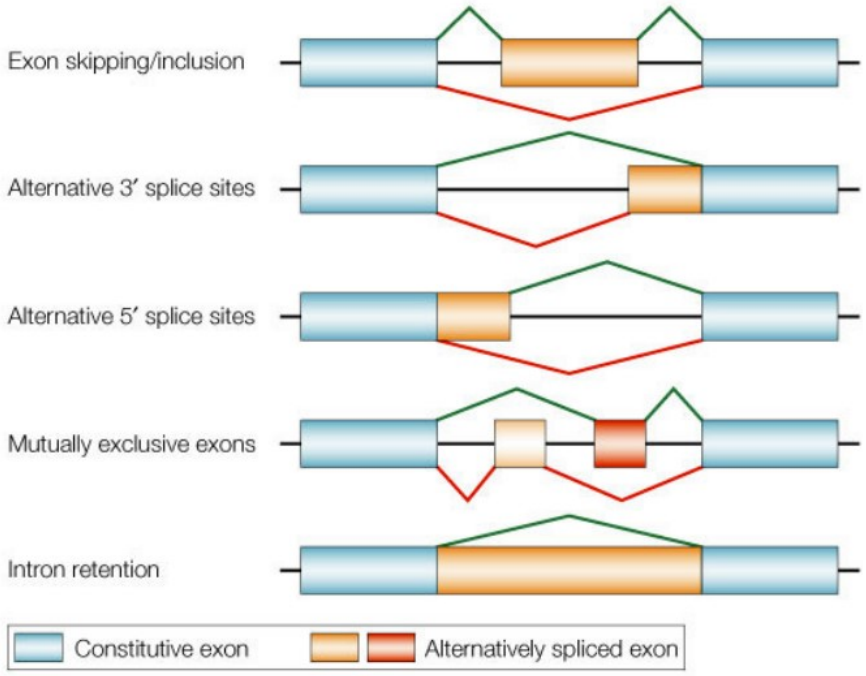


Figure 1.4: Diagram demonstrating the potential uses of alternative splicing and the variation in output produced by the inclusion or removal of intronic/exonic regions of sequence. The most common splicing is indicated at the top when consecutive exons are spliced together by removing introns. Multiple splice paths are shown in green and in red. Splicing can be used in conjunction with alternative promoters or polyadenylation sites to create complex and variable transcripts from a single gene, taken from (Matera and Z. Wang 2014).

(hnRNPs) associating with cognate binding sites on the pre-mRNA are instrumental in controlling alternative splicing pattern (Chaudhury, Chander, and P. H. Howe 2010). The components and structure of the spliceosome change throughout the splicing process. Spliceosome assembly occurs at designated sequences in the transcripts called splice sites. Splice sites demarcate the exon/intron borders thata are the 5' slice site and the 3' splice sites. Most introns start with a GU dinucleotide sequence at the 5' region and a AG at the terminal 3' region. The alteration of even a single nucleotide at these sites can result in splicing inhibition (Clancy et al. 2008). The recognition of the splice sites and the assembly of a functional spliceosome is reliant on four distinct sequence elements. The 5-splice site is defined by a 10 nucleotide motif encompassing the intron/exon border. At the 3'end of the intron three distinct features are present, the Branch point sequence which is a specific sequence containing an adenine (A) nucleotide near the 3' end of an intron. This adenine serves as the nucleophilic group that initiates the splicing process by initiating the first nucleophilic attack, followed by a pyrimidine track the 5' splice site, leading to the removal of the intron during mRNA maturation, and finally the AG dinucleotide designating the end (Konarska et al. 1985), (Ward and Cooper 2010). The outline of these splice sequences are shown in the schematic adapted from 1.5 taken from (Thapar 2015) on page 42.

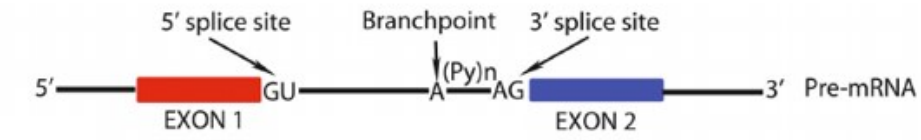


Figure 1.5: Schematic of sequences in pre-mRNA involved in splicing. These sequences include the 5-splice site, a GU dinucleotide sequence at the 5' region, the Branch point, the pyrimidine rich region Pyn on the schematic, and a AG at the terminal 3' region, from (Thapar 2015)

1.4 Splicing in *Saccharomyces cerevisiae*

The splicing mechanism in *Saccharomyces cerevisiae* is more simplistic than higher multicellular metazoans. Its simplicity makes it a good model organism for investigating this process compared to mammalian genes that often contain upwards of 8 introns per gene. *S.cerevisiae* have less than 300 genes that contain a single intron and less than 10 genes with multiple introns have been identified (Stajich, Dietrich, and Roy 2007) (Juneau, Nislow, and Davis 2009). This simplicity makes it a good model organism for investigating alternative splicing. Although, less than 5% of *Saccharomyces cerevisiae* genes contain an intron, 90% of these encode ribosomal proteins, and thus make up a large proportion of the transcriptome. As ribosomal genes are essential for growth and proliferation, splicing plays a fundamental role (Ares, Grate, and Pauling 1999).

S.cerevisiae genes that are spliced usually contain a single intron towards the 5' end of the transcript that is around 270 nucleotides in length. The four splice elements that recruit the spliceosome in mammals also exist in *S.cerevisiae*. The splice site consensus sequence, UACUAAC, that is at the

branch point is highly conserved and far less degenerate than in humans and *Schizosaccharomyces pombe*, perhaps relating to the lower complexity and usage of splicing compared to other organisms (Fair and Pleiss 2017). The splicing mechanism is similar to that in mammalian cells. However the *S.cerevisiae* genome lacks the splice factor U2AF1 that is essential for the recognition of 3' splice site in mammals. Alternatively, they contain the gene encoding MUD2 which has an equivalent function but is not essential. It is thought that *Saccharomyces cerevisiae* use a 'scanning mechanism' to find the intron borders and could also explain the heightened specificity at the other splice sites (Madhani and Guthrie 1994).

1.5 Cleavage and polyadenylation in *Saccharomyces cerevisiae*

The factors involved in the process of cleavage and polyadenylation are highly conserved in nature. The overall mechanisms of recruitment and assembly of the 3' end formation machinery onto the pre-mRNA is also conserved between humans and yeast. However there are several key differences to note in the yeast *saccharomyces cerevisiae*. The mRNA processing is catalyzed by three key factors. These are the Cleavage and polyadenylation Factor (CPF) and the multi-subunit complex Cleavage Factor IA and Cleavage Factor IB (CFIA and CFIB). The CPF is a 850kDa 14 subunit complex of three functionally distinct but associated complexes. The CPF constitutes three modules, each with a distinct enzymatic activity: the polymerase module, the nuclease module, and phosphatase module (Casañal et al. 2017).

The phosphates module is not well defined but contains a Gcl7 phosphatase Ssu72 that is thought to help promote transcription termination (Schreieck et al. 2014). The polymerase module contains the polymerase PAP1, Cft1 and Psf2. These act as a scaffold for Fip, a PAP1 binding subunit and the nuclease module via Ysh1. The nuclease module contains the nucleases Ysh1, Cft2 and Mpe1. Mpe1 appears to be instrumental in linking pre-mRNA recognition and the activity of the nuclease module (Rodrıguez-Molina et al. 2022).

The activation of the cleavage complex is further controlled by auxiliary factors CFIA and B (A. Kumar et al. 2019). The CFIA is composed of Rna14, Rna15, Clp1, and Pcf11 and CFIB which contains Hrp1. The CPF recognises the pA core signal and the more variable flanking elements via interactions with Cft1, Ysh1 and Yth1. CFIA binds to the phosphorylated CTD of PolII, which consists of 26 repeats of a heptapeptide module with the residue sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser.(Nedea, He, et al. 2003), (Nedea, Nalbant, et al. 2008), (Kyburz et al. 2003). Both the CFIA and Cstf complex bind in a comparable manner to the pre-mRNA substrate so the overall process is conserved (Baejen et al. 2014).

However, the core signal sequences that direct the cleavage polyadenylation machinery to the cleavage point in yeast differ from the mammalian counterparts (Manley and Takagaki 1996). In *saccharomyces cerevisiae*, there are four main elements that make up the core signal sequence that are involved in the assembly and positioning of the cleavage and polyadenylation machinery onto the pre-mRNA. A core signal A-rich positioning element (PE) is located 10-30 nucleotides upstream of the cleavage site (Tian and

Graber 2012). The signal is often still composed of AAUAAA, however the variation in this PE site is more common than in the mammalian PAS element equivalent. In addition, there is an efficiency element (EE), located upstream of the cleavage site and the PE, is most commonly located 20 nt upstream of where the transcript is cleaved (Dichtl and Keller 2001). This site is not involved in cleavage but is used to increase site selection efficiency and the recruitment of cleavage machinery (Guo and Sherman 1996). The PE signal has been shown to bind to the Rna15 CFIA subunit in invitro studies (Gross and C. L. Moore 2001). The EE interacts with the CFIB submit via Hrp1/Nab4 (M. M. Kessler et al. 1997). In addition, there are both upstream and downstream U-rich elements that flank the cleavage site. These sites interact with both the Ysh1/Cft2, and Yth1/Cft1 of the CPF. It has been shown experimentally that mutations at these sites can significantly reduce the cleavage activity of the CPF (Dichtl, Blank, et al. 2002). A map of of these regions, including the signal sequence/cleavage and polyadenylation sites on the nascent transcript is shown in the diagram 1.6 adapted from (Tian and Graber 2012) on page 46.

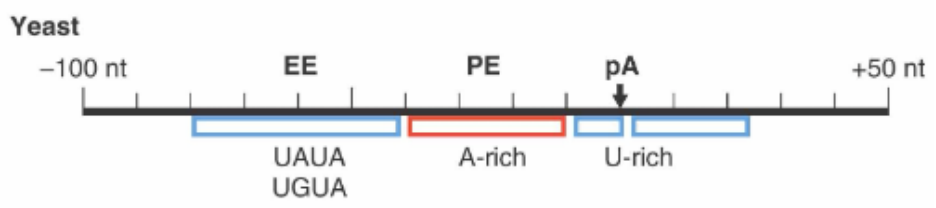


Figure 1.6: Map of core signal sequences on nascent yeast transcripts, adapted from adapted from (Tian and Graber 2012). The diagram illustrates the core signal sequences involved in directing the cleavage and polyadenylation machinery during mRNA processing in yeast. These sequences consist of four main elements: Core Signal A-rich Positioning Element (PE): Located 10-30 nucleotides upstream of the cleavage site, this element plays a crucial role in positioning the machinery. Although it often contains the AAUAAA motif, there is more variation in this element compared to its mammalian counterpart. Efficiency Element (EE): Positioned upstream of both the cleavage site and the PE, typically about 20 nucleotides upstream of the cleavage site. While not directly involved in cleavage, it enhances site selection efficiency and machinery recruitment. Upstream and Downstream U-rich Elements: These elements are found flanking the cleavage site and interact with specific components of the CPF (cleavage and polyadenylation factor) complex, including Ysh1/Cft2 and Yth1/Cft1. Mutations in these elements can significantly affect CPF's cleavage activity.
.

1.5.1 Conserved elements between yeast and mammalian 3'-mRNA end processing

Several key factors in the yeast and mammalian cleavage and polyadenylation machinery are highly similar in structure and function and many act as direct homologs to one another. In fact, all subunits in the mammalian and *Saccharomyces cerevisiae* core polyadenylation and 3énd processing factors have at least one homologue, the only exceptions being the CFI and CSFT50 (Zhao, Hyman, and C. Moore 1999). The CPSF subunits are homologous to several CPSFS. Ysh1 shares 53% identity to CPSF-73 and 23%

to CPSF-100 (which is also 24% identical to CFT2) (Jenny et al. 1996). CFT1 is homologous to CPSF-160 with 24% identity. YTH1 is highly conserved to CPSF-30 with 40% identity to one another (Barabino et al. 1997). RNA14 and RNA15 mammalian counterparts are the Cstf subunits Cstf-77 and Cstf-64 respectfully (Qu et al. 2007). Throughout the RNA-binding domain, RNA14 shares 24% identity to CstF-77, and RNA15 is 43% identical to CstF-64. The Mpe1, the essential component of the CPF required for cleavage at the pas site is homologous to the human RBBP6 (Rodrıguez-Molina et al. 2022). The PAP in yeast is highly conserved, with 47 identity to the mammalian PAP protein, however it is truncated and does not contain a AAUAAA hexamer recognition element (Barnwal et al. 2012). The *Saccharomyces cerevisiae* polyadenylation factor Hrp1p that interacts with UA-rich elements in pre-mRNAs, is missing from mammalian systems (Tian and Manley 2017). The Cleavage and polyadenylation machinery for mammals and yeast is summarised in the diagram 1.7 adapted from (Keller et al. 1991) on page 48.

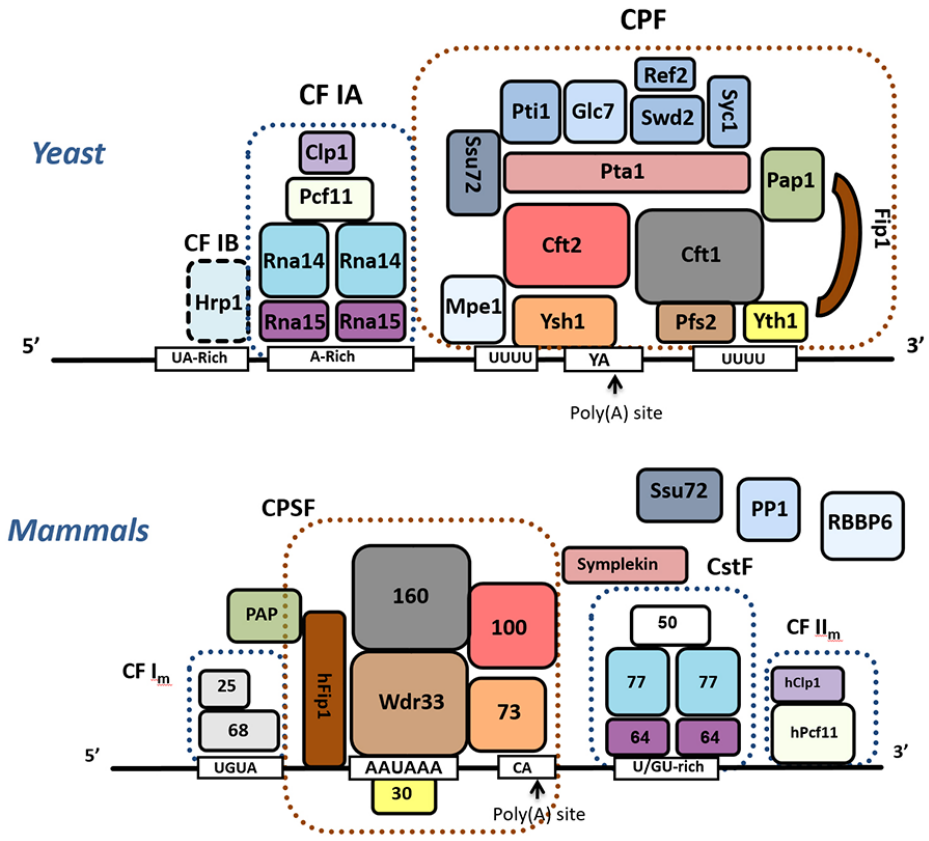


Figure 1.7: Diagram of yeast cleavage and polyadenylation machinery. Cleavage Factor IA CFIA is made up of the proteins Rna14, Rna15, Clp1 and Pcf11. CFIB which is composed of Hrp1.The CPF (holo-Cleavage and Polyadenylation Factor) consists of an APT (Associated with Pta1) subcomplex which includes Pti1, Swd2, Syc1, Ref2, two protein phosphatases, Glc7 and Ssu72, and the Pta1 factor, which acts as scaffold protein that bridges between the holo-CPF sub-complexes. The CPF subcomplex also contains Cft1/Yhh1, Brr5/Ysh1 and Cft2/Ydh. that also interact with Pfs2, adapted from adapted from (Keller et al. 1991)

.

1.6 Alternative Cleavage and polyadenylation

Alternative polyadenylation (APA) is a mechanism which controls gene expression by utilizing alternative polyadenylation sites on a gene to produce an array of distinct mRNAs isoforms (Di Giammartino, Nishida, and Manley 2011). APA is ubiquitous in the mammalian genome has roughly 80% of all protein coding genes contain multiple poly A sites (X. Liu et al. 2017). APA can occur via changes to its coding region (CR-APA). If the pas site is in the coding region it will alter the primary base sequence to alter the protein output, much like traditional splice isoform variants (Ren et al. 2020). APA also occurs in the 3' untranslated region (3'UTR-APA) (Tian and Manley 2017). Even though the 3'UTRs has no effect on the sequence of the resultant protein output, the 3'UTR often contains regions that can effect the post transcriptional regulation of gene expression. For example, this can occur by the inclusion or removal of cis regulatory elements which can alter binding sites for proteins and micro RNAs, targeting the transcript for degradation or further stabilizing it. (Tian and Manley 2017). Therefore the 3'UTR sequence often impacts greatly on gene expression in a post transcriptional manner. Global regulation of APA has been identified in controlling many fundamental biological processes, including cellular proliferation, stress response and is deregulated in many cancers (Alt et al. 1980), (Xia et al. 2014), (D. Zheng et al. 2018). The different types of APA is depicted in diagram 1.8 adapted from (Di Giammartino, Nishida, and Manley 2011) on page 50.

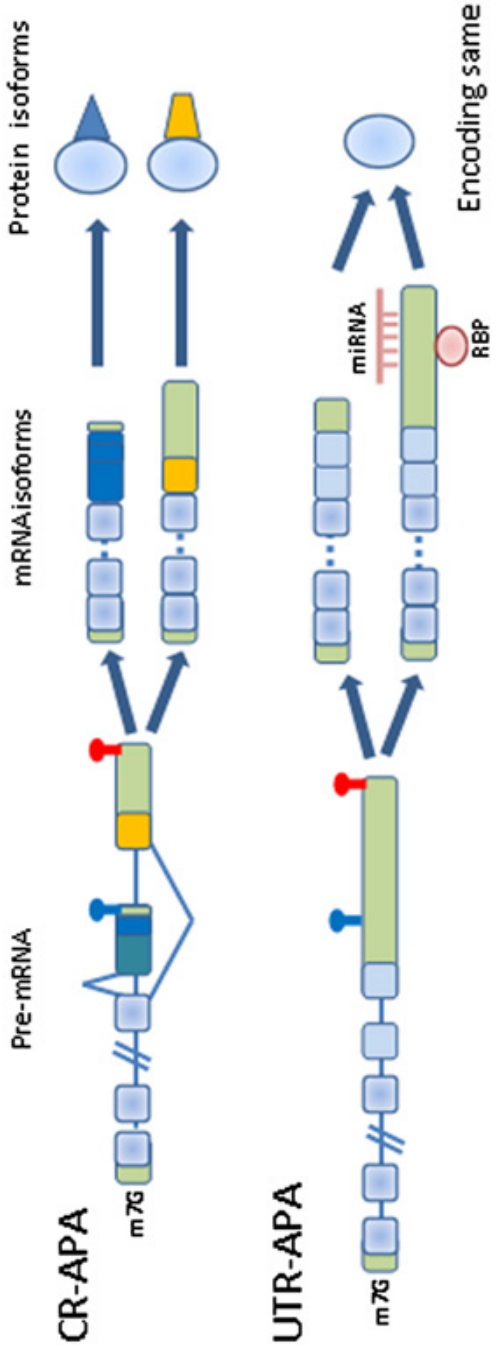


Figure 1.8: Coding region-alternative polyadenylation (CR-APA) occurs within the coding region of a transcript, which results in distinct alterations in protein outputs. 3'UTR-alternative polyadenylation (3'UTR-APA) encode the same protein coding region but have alternative lengths and sequences in the 3'UTR, which can exclude/include cis-regulatory elements that effect the mRNA stability. Blue shading is for coding regions and green are the untranslated regions, the introns are depicted as dark blue lines, adapted from (Di Giammartino, Nishida, and Manley 2011)

1.7 Regulation of APA

Regulation of APA isoform frequencies occur via passive or active mechanisms (Neve et al. 2017). Active APA is co-transcriptional and occurs via the selection of either proximal or distal pas site usage and refers to changes to selection of a pas site. This is influenced by both trans and cis acting regulators. Preference in pas site usage can be altered by many factors, including DNA methylation, cleavage and polyadenylation substrate concentrations or RNA binding proteins (Nanavaty et al. 2020), (Fox-Walsh and Fu 2010). Passive APA occurs post transcriptionally and affects the levels of mature transcript available to be translated into protein. This occurs via targeted degradation or stabilisation of the mature mRNA, by trans acting factors such as RNA binding proteins and microRNAs, or cis elements which influence the structure of the mRNA. Having multifaceted regulatory processes gives a dynamic adaptive mechanism to rapidly alter gene expression changes in response to changes in cellular states such as cell cycle progression, differentiation and metabolic stress (A. J. Gruber and Zavolan 2019), (Elkon, Ugalde, and Agami 2013). There are four different ways in which this can be considered. There are the cis acting elements in the mRNA and the trans acting factors that can interact with them. These can occur co-transcriptionally and post transcriptionally i.e. active and passive APA respectively. These categories of APA regulation are summarised in the diagram 1.9 on page 52.

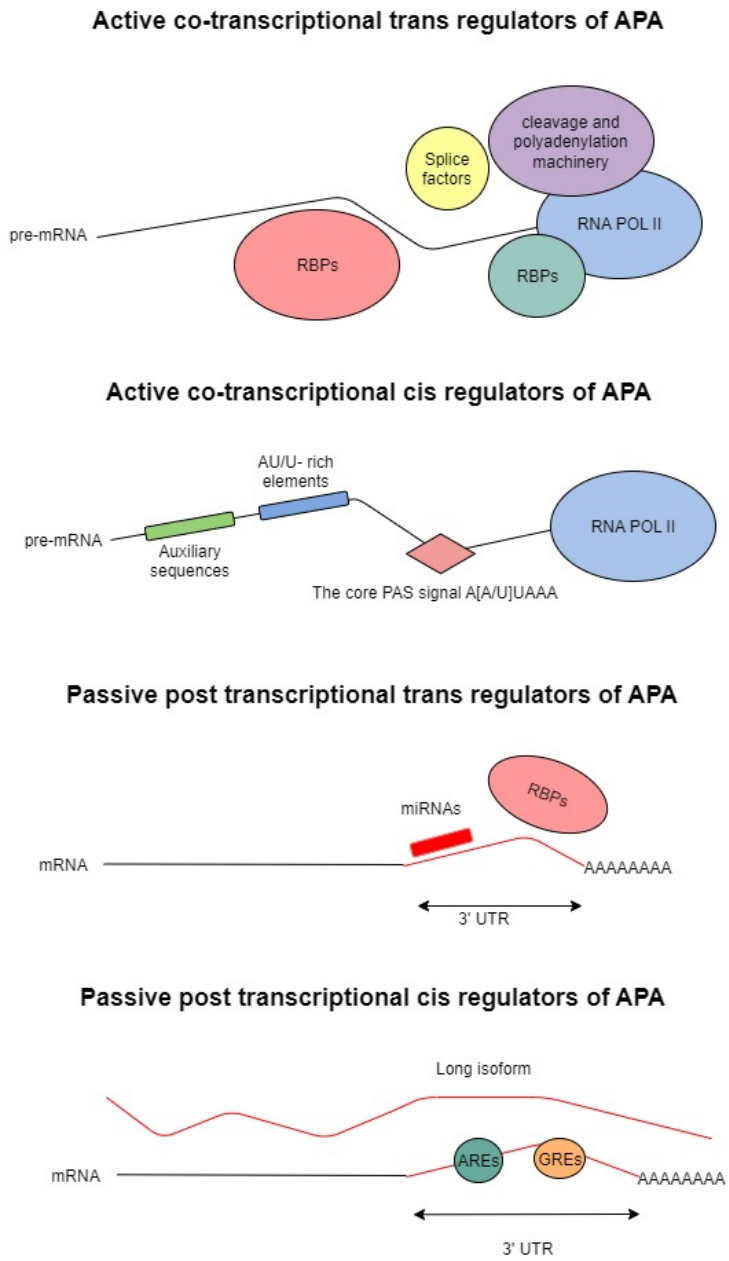


Figure 1.9: Diagram showing types of APA regulation. APA regulation can occur both co-transcriptionally (active) and post transcriptionally (passive). The cis elements in the transcript and trans elements that bind to them are also important categories of APA regulation. Active trans regulators include RBBs, Splice factors and many factors involved in the cleavage and polyadenylation machinery. active cis-regulators of APA are the elements in the pre-mRNA, including the core PAS signal, AU/U rich elements and Auxiliary sequences. Passive, post transcriptional APA regulation mostly involves trans regulatory elements interacting with the 3'UTR. Passive post transcriptional regulators include microRNAS and RBPs that bind to elements mostly in the 3'UTR of the matured mRNA. Trans elements that interact post transcriptionally include AREs and GREs. MRNAs can also regulate their own transcription via negative feebackloops of their longer or shorter isoforms.

.

1.7.1 Active co-transcriptional trans regulators of APA

Many regulators of APA are mechanisms that occur co-transcriptionally that influence the pas site selection. These include both trans regulators and cis regulators. The trans regulators include the 3' end formation, polyadenylation and cleavage machinery and binding proteins (Neve et al. 2017). The selection of a polyA site occurs alongside pre-mRNA processing and termination. Therefore, the selection of a pas site is largely dictated by the elements involved in the transcription termination process (Richard and Manley 2009). These include mRNA binding proteins, micro RNAs and most notably, the factors involved in the polyadenylation and cleavage machinery. The levels of the general polyadenylation and cleavage machinery can be enough to cause a significant shift in APA site usage. An example of this is the switch from distal to proximal pas site usage, therefore global shortening of B cells during differentiation which causes a shift from the production of the membrane-bound IgM heavy chain to its secreted form (Takagaki, Seipelt, et al. 1996). The proximal pas site binds to the CstF complex with lower affinity than the distal site. During levels of low substrate concentration, the distal site is dominantly used as it has the strongest affinity for the Csft-64 substrate. Once the levels of transcription of the CstF increase, there is less competition between the two sites for the substrate and the more proximal site is utilized first, causing APA shortening. This is believed to be the primary mechanism of global APA shortening that is seen in cell differentiation, as the distal pas site tends to be stronger compared to the proximal, however during differentiation, there is a surge in CsfT concentration, which enables

the global shift to proximal site usage. (Takagaki and Manley 1998). Splice factors can also be heavily involved in pas site selection in a similar fashion to the cleavage and polyadenylation machinery. This is due to the processes of splicing, mRNA end processing and transcription termination being tightly associated (Tian and Manley 2013). For example, the snRNP U1 can inhibit PAP and negatively influence proximal pas site usage. The inhibition of snRNP U1 causes a brief but significant shortening of transcripts during neuron activation (Berg et al. 2012). The RNA-binding proteins (RBP) in the polyadenylation machinery are main regulators of APA. RBPs can have both an enhancing or an inhibitory affect on pas site usage and the affect a RBP can have on APA is often context dependent. RBPs can interact with the cleavage and polyadenylation machinery to facilitate or obstruct their binding to cis sequences on the nascent mRNA transcript. CPSTF and CFIm are both involved in up regulation of distal pas site selection (Erson-Bensan 2016). Many RBPs facilities polyadenylation machinery binding to specific pas sites. HNRNP H (heterogeneous nuclear ribonucleoprotein H) binds to proximal pas sites, leading to the recruitment of polyadenylation machinery proteins and therefore APA shortening (Katz et al. 2010). APA shortening via the cytoplasmic polyadenylation element-binding protein 1, CPEB1 occurs via a similar mechanism. CPEB1 has a strong affinity for proximal pas sites and increases CPSF recruitment to induce 3'UTR APA shortening (Bava et al. 2013). In HeLa cells, the RBPs, Hu proteins, bind to CstF64 and CstF160 and block their ability to bind to U rich sequences to prevent cleavage at these sites (Zhu et al. 2007). The RBP MBNL that is involved in splicing also has a multitude of effects on the pas site selection.

MBNL has been shown to obstruct the core cleavage and polyadenylation machinery when bound within 100 base pairs of a specific pas site. Conversely, MBNL also acts to recruit many of the factors involved in the core polyadenylation machinery (Jenal et al. 2012), (Klerk et al. 2012). RBPs role in active APA regulation is summarised in the diagram 1.10 adapted from (Erson-Bensan 2016) on page 58.

1.7.2 Active co-transcriptional cis regulators of APA

The cis active regulators are sequences in the pre-mRNA substrate that influence the pas site usage during 3énd formation (Tian and Manley 2013). The hallmarks of the conical cis-elements of a pas site are explained earlier in the introduction. The individual strength of these elements is combinatorial in nature. The sequences surrounding the cleavage site on the pre-mRNA substrate have a variable and competitive affinity for binding to the cleavage machinery and therefore can influence pas site usage. The core PAS signal A[A/U]UAAA tends to have more variation at the proximal pas sites, therefore indicating that the proximal site is more fluid in its substrate binding affinity (Sheets, Ogg, and Wickens 1990), (Tian and Manley 2017). Pas sites often contain specific sequences to bind to specific pre-mRNA processing and termination factors. PASs that are regulated by *CSTF64* have been found to have U and GU enrichments (Tian and Manley 2017), (W. Li et al. 2015). Another specific binding site is an upstream U enrichment that specifically binds with factor interacting with PAP, *FIP1* This cis element interaction with *FIP1* is so essential to the recruitment of the CPSF, that knockdown causes global 3'UTR lengthening (Lackford et al. 2014). One of the most

common regulatory elements in cis sequences are adenylate uridylate (AU-rich) elements (AREs). These motifs are pentamer "AUUUA" which can occur with variable length repetitions. The transcripts with high frequencies of AREs are associated with low stability rates (Nourani, Robert, and Winston 2006).

1.7.3 Passive post transcriptional trans regulators of APA

Once the mature mRNA transcripts are exported out of the nucleus, they can be post transcriptionally regulated in the cytoplasm. Passive APA regulation can affect the stability and localisation of the mature mRNA transcript. One of the most common trans post transcriptional APA regulators are MicroRNAs (miRNAs). MiRNAs are a key passive, post transcriptional trans acting regulators of APA. MiRNAs are small, non coding RNAs around 20 base pairs in length. MiRNAs bind to complementary sequences on transcribed mRNAs, most commonly in the 3' untranslated region. Once bound the miRNAs target these transcripts for degradation, therefore acting as a mechanism of post transnational gene silencing (Huntzinger and Izaurralde 2011). MiRNAs can be utilized to down regulate specific APA isoforms by binding to specific sequences that only exist in the elongated transcripts that use the distal pas site, therefore causing APA shortening (Lin and Gregory 2015). As these sequences do not exist in the shorter transcripts, they are not targeted by miRNAs, so become the more dominant transcript. This has been demonstrated in colorecatal tissue from cancer patients (Yang et al. 2018). They found 16 miRNAs of significantly altered expression levels

that directly targeted 3'UTRs of 6 polyadenylation factors, all of which had altered expression levels compared to wild type cells. This gene silencing via altering expression of key genes via APA is currently being explored as major biomarkers of different cancers (Brumbaugh et al. 2018), (Mao et al. 2020). The RBPs regulate mature mRNAs by binding to specific sequences on the transcripts 3'UTRS and recruit effector proteins such as mRNA decay machinery and protein kinases to the target mRNAs. The sites at which RBPs bind to mRNA are small and often degenerate, allowing many different proteins interactions via the same site (A. R. Gruber et al. 2014). 3'UTR APA alters the availability of these binding sites and therefore affects the stability and localisation of the transcripts in a passive post transcriptional manner. One of the most significant groups of RBPs are ARE-binding proteins (ARE-BPs). One of the first AREs-BPs identified as being able to destabilise mRNAs was *AUF1*, heterogeneous nuclear ribonucleoprotein D (hnRNP D). *AUF1* targets and destabilizes key proliferation genes such as tumor necrosis factor-, *TNF*- and interleukin-1, *IL-1*, *c-fos* and *c-myc*. KO mouse models have a build up of these factors and have a severe endotoxic shock phenotype (Erson-Bensan 2016). RBPs mediated passive APA regulated is summarised in the diagram 1.10 adapted from (Erson-Bensan 2016) on page 58.

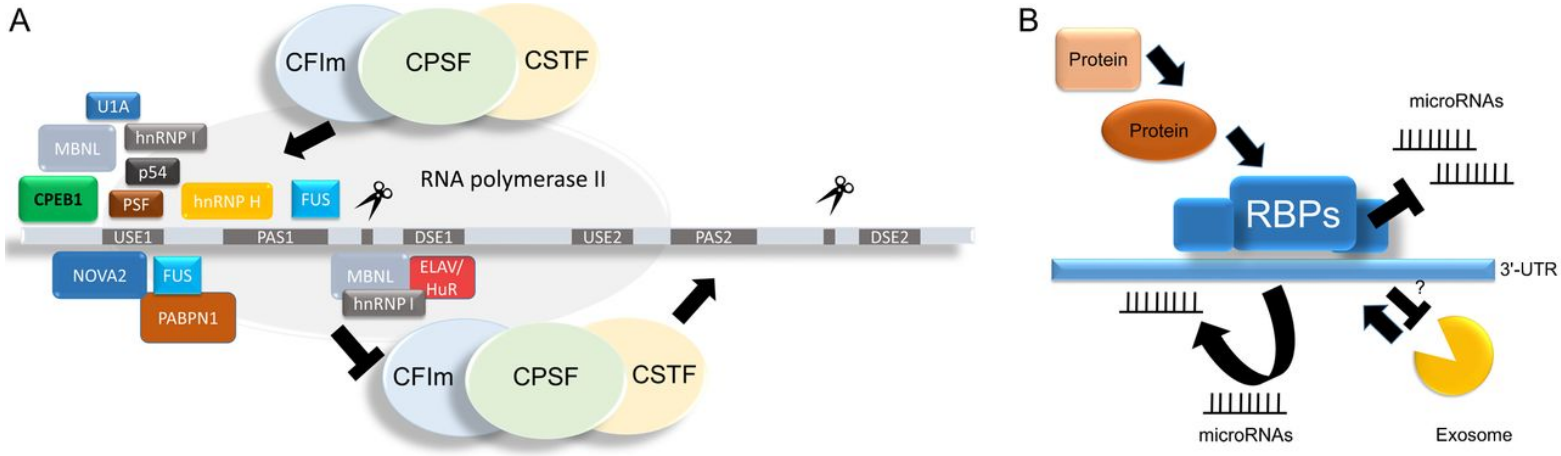


Figure 1.10: APA regulation via RBPs. (A) Co-transcriptional RBPs effecting pas site selection and therefore APA. RBPs have been shown to alter the binding and recruitment of core polyadenylation machinery complexes (CPSF, CTSF, and CFIm) to binding sites on pre-mRNA. Proximal and distal pas are marked with PAS1 and PAS2 respectfully. The U-rich/UGUA upstream elements are indicated with USE and the U-/GU-rich downstream elements are shown with DSE. The Cleavage site is indicated by a pair of scissors. (B) Post transcriptional RBP mediated APA. Variations in APA isoforms 3'UTR cis-elements affect RBP binding. RBPs can recruit effector proteins and affect stabilization of mature mRNA transcripts. Adapted from (Erson-Bensan 2016)

1.7.4 Passive post transcriptional cis regulators of APA

The cis sequences can affect the number of transcripts with specific PAS in the transcriptome via post transcriptional regulation. AU-rich (AREs) and GU-rich (GREs) elements are cis elements in matured mRNA that can bind to regulatory trans elements. AREs regions are often associated with mRNA instability and decreased translation efficiency (Bakheet et al. 2018). Cis sequences also act to repress gene expression of the transcript via self regulation in negative feedback loops. These alternative pas sites are often in the coding regions and cause the production of truncated, and therefore non-functional versions of proteins, therefore causing gene suppression. An example of pas site usage causing inhibition of gene expression is *CSTF77/CSTF3*, which encodes the mammalian polyadenylation factor cleavage stimulation factor 77 kDa subunit. *CSTF77* is one of the subunits of the CSTF and is essential for the regulation of several cell cycle progression genes (W. Luo et al. 2013). The full length, functional transcript uses the 3ḿost exon. The truncated form uses an upstream, intronic pas site that produces a shorter mRNA that lacks functionality when translated. The APA shift to the proximal pas site is controlled by the long form of the transcript. Therefore this process is self regulatory, as when the concentration of the full length isoform increases, it stimulates the usage of the proximal alternative pas site, which reduces the amount of functional *CSTF77* protein. This process of gene silencing via APA negative feedback loop is shown in in diagram 1.11 adapted from (Tian and Manley 2017) on page 60.

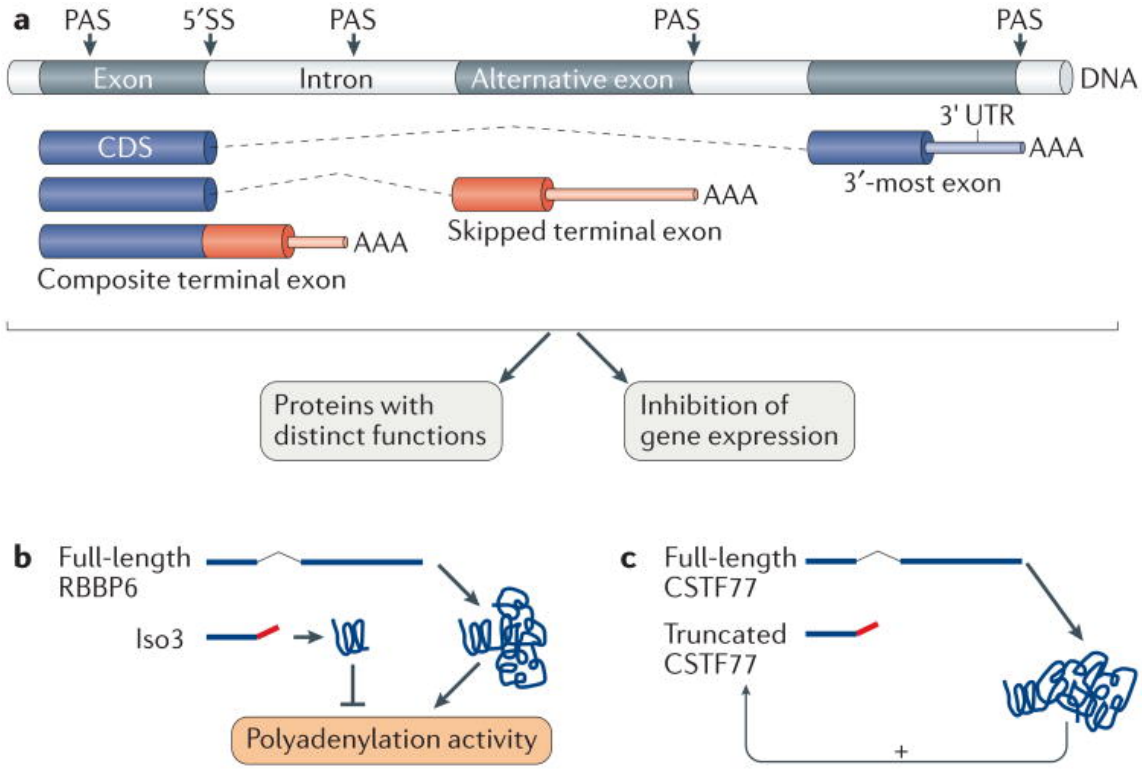


Figure 1.11: Intronic alternative pas sites in a gene enabling variation in protein outputs. Coding region pas sites cause the truncated and non functional version of transcript to effectively silence gene expression. Full length transcript self regulates its own expression by stimulating use of upstream coding pas site. Iso3 isoform of *RBBP6* is shown with its dominant negative function of inhibiting polyadenylation activity. Coding regions are shown in blue and 3' ends shown in red, adapted from (Tian and Manley 2017)

mRNA stability and localisation

As APA affects the sequence of the mature transcript via the inclusion or removal of both the coding region (CR-APA) and 3 'UTR length, APA can strongly impact on both the stability and localisation of transcripts and therefore have a significant impact on the transcriptome. MRNA stability is a key factor in determining gene expression output of a transcript as the more stable the transcript is, the longer it is around to be used as a substrate to be translated into protein. MRNA stability it determined via several features in the matured transcript. APA isoforms that have a longer half life will dominate the transcriptome over less stable ones, even if they are transcribed at similar rates (Hogg and Goff 2010). Trans-regulatory elements affecting mRNA stability such as MicroRNAs and RBPs bind to regulatory elements in the transcript to stabilise/destabilize target mRNAs. These cis-regulatory elements that are recognised by these factors are located primarily in the 3'UTR of the mRNA (Despic and Neugebauer 2018). Cis-elements that have been shown to influence the stability of a transcript include, AU-rich (AREs) and GU-rich (GREs) elements which have been associated with mRNA instability and decreased translation efficiency (Bakheet et al. 2018). The aUTR of an mRNA transcript also plays an important role in the localisation of the mRNA as it can contain localisation elements to direct the transcript to a specific compartment of the cell. An example of alternative pas sites affecting the localisation of a transcript is the the muscleblind like. *Mbnl* family of RNA-binding proteins that bind to specific alternative 3 UTRs. Several hundred genes with alternative 3UTR isoforms have different local-

isation patterns due to the cis elements found in distal regions, which were frequently bound by Mbnl1 (E. T. Wang, Cody, et al. 2012).

1.7.5 Functional role of global regulation of APA

Even though the use of APA is ubiquitous and frequent across many organisms, the function of APA is still poorly understood. However, there are some trends and specific examples of APA that have been identified. APA, especially 3'UTR APA plays a key role in gene expression regulation. This is due to the 3'UTR of mRNA having cis regulatory elements that influence the stability, localisation and translation of the transcript (Rachael E Turner and Beilharz 2021).

Several significant cellular processes have been shown to correlate with shifts in global APA. The immune system has been shown to utilise global APA shortening to increase protein output during immune cell activation, specifically T cell activation (Sandberg et al. 2008). More specific utilization of APA shortening is demonstrated in the *Hip2* gene, a ubiquitin-conjugating enzyme involved in apoptosis suppression (Mandel et al. 2006) (Song et al. 2008). The long isoform that uses the distal pas site has two seed matches for the degradation miRNAs mir-21 and mir-155. In the shorter isoform, these miRNA sites are eliminated and the levels of the ubiquitin conjugating enzyme increased significantly. It is estimated that the removal of miRNA binding sites can increase a transcripts gene expression by nearly two thirds compared to its longer APA isoform, therefore shortening can have signifcant impacts on the protein output of a gene (Mayr and Bartel 2009). Global APA shortening has also been shown to increase protein output in neurocytes with

active membrane polarization, as an increase in polarization agents have been found to correlate to an increase in proximal pas site usage and subsequently an increase in transnational potential (Tian and Manley 2017). Therefore, an APA shift to the proximal pas site often has a positive effect on the protein output of the gene (Sandberg et al. 2008). Global shifts to APA shortening have also been shown in more specific immune activation responses but have not resulted in an increase in protein output. This suggests there might be other so far uncharacterised function to APA apart from gene expression changes (A. R. Gruber et al. 2014). APA is also often used to silence genes that are essential for the processing of other genes required for a specific cellular function, and therefore acting as a master switch for these mechanism. Cell cycle progression is one of these mechanisms that uses APA to regulate multiple genes. CSTF77/CSTF3 protein acts as a homodimer that interacts with CSTF1 and CSTF2 to form the cleavage stimulation factor complex (CSTF). The CSTF is essential for polyadenylation and 3énd processing of pre-mRNAs (Grozdanov et al. 2018). This gene contains an intronic alternative pas site that can be utilized to produce a non functional, truncated form of the protein. The mechanism of this silencing is shown in diagram 1.11 adapted from (Tian and Manley 2017) on page 60 The protein CSTF77/CSTF3 has been shown via GO analysis to be a regulator of cell cycle genes. The knock down of *CSTF77* results in the suppression of several key genes involved in cell cycle progression including, *cyclin B1* and *CDK4*. This indicates that the silencing of *CSFT77/CSTF3* via APA can inhibit cell cycle progression via subsequent silencing of key genes (W. Luo et al. 2013).

Another example of APA affecting global 3' end processing is the APA

of the gene, *RBBP6*. *RBBP6* is one of the core factors involved in cleavage and polyadenylation interacts with both WDR33 and CPSF73 to activate the CPSF (Boreikaite et al. 2022). Altering the gene expression of *RBBP6* reduces global mRNA processing so has a significant impact on the cells transcriptome (Boreikaite et al. 2022). *RBBP6* is regulated via the production of a dominant negative isoform that is produced using APA. A proximal, intronic pas site is used instead of the canonical pas, to produce a truncated version of the gene called *Iso3*. *Iso3* acts as an antagonist to *RBBP6* and actively competes with the functional, full length *RBBP6* to bind with the cleavage and polyadenyaltion machinery. This reduces the cells ability to process other mRNA transcripts and therefore regulates global gene expression (Di Giammartino, W. Li, et al. 2014). *RBBP6* regulation via APA is shown in diagram 1.11 adapted from (Tian and Manley 2017) on page 60.

1.8 APA in *Saccharomyces cerevisiae*

Yeast and mammalian systems share many similarities in their biochemical functions, including the regulation of APA. Roughly 70-80% of *Saccharomyces cerevisiae* genes contain multiple 3' ends, indicating APA. This is comparable to mammalian systems that have similar numbers of genes that have the potential to perform APA (X. Liu et al. 2017). APA has been identified as a key regulator in many biological processes including growth and cellular stress in *Saccharomyces cerevisiae* (Kim Guisbert, Hao Li, and Guthrie 2007).

The cis pas sequence motifs that direct the cleavage and polyadenylation machinery to a specific pas site are more variable in *Saccharomyces cerevisiae* compared to higher organisms. Yeast have four notable pas motifs that influence pas site selection (as explained in a previous chapter). These motifs act in a similar fashion to the mammalian systems, however they show much more degeneracy and are less conserved. The canonical A[A/U]UAAA that is roughly 40 base pairs upstream of the cleavage site is far more variable in *saccharomyces cerevisiae*. There are often many A to G or U to C substitutions in the EE efficiency elements that do not affect the transcription potential of the transcripts (Guo and Sherman 1995). The regions that flank the cleavage sites are also highly enriched with AG rich motifs that might aid pas site selection. The multiple cleavage sites of yeast typically cluster together and signal sequences for different pas sites often overlap; there can be less than 10 nucleotide difference between two pas sites, which is uncommon in mammalian cells (X. Liu et al. 2017), (Schlackow et al. 2013).

Saccharomyces cerevisiae have several different binding proteins that target variants of the common core PAS sequence and also matured mRNAs. One major family of RBPs in *Saccharomyces cerevisiae* are the PUF protein family (M. Wang et al. 2018). *Saccharomyces cerevisiae* have at least 5 different PUF proteins which target 3'UTRs of mature mRNAs to inhibit translation of specific APA isoforms and enhance their turnover (Tadauchi et al. 2001). For example, Puf6p inhibits the translation of the mRNA *ASH1*, knock outs of *Puf6p* result in significant decrease in gene expression of *ASH1* (Gu et al. 2004). Another RBP that modulates APA of key transcription factors in *Saccharomyces cerevisiae* is Nab4/Hrp1. Nab4 is a heterogeneous nuclear ribonucleic acid and binding protein (hnRNP) that is a component of the cleavage factor complex and plays a role in 3' end processing but also involved in nonsense mediated decay of mRNAs (González et al. 2000). The precise mechanism of how Nab4 controls APA is not well understood, however knock out mutants show increases in cryptic pas site usage. Nab4 co-immunoprecipitates with TFIIB, which is a transcription factor that directs RNA pol II to the transcription initiation start site. TFIIB has two pas sites; under normal wild type conditions the ratio of distal to proximal pas site usage is 1.65. In nab4-7 mutants, the ratio increases to 2.37, indicating a shift towards *TFIIB* lengthening. In excess Nab4, the opposite is seen, with a preference to the upstream pas site to produce shorter transcripts. In addition to altering *TFIIB* transcript lengths, Nab4 also regulates APA of copper toxicity stress genes and its binding motif has a GO enrichment for metal ion transport. Knock out strains have a significant resistance to copper toxicity and the shortened transcripts have an 8 fold increase in transcript

levels (Kim Guisbert, Hao Li, and Guthrie 2007). This indicates that Nab4 has a destabilising effect on transcripts via its ability to effect APA.

As with humans, *Saccharomyces cerevisiae* APA regulation is heavily reliant on the subunit components of the cleavage and polyadenylation machinery. Recent studies have shown that the Mpe1, a CPF component that is homologous to the human RBBP6, is essential for 3' end processing. The Mpe1 binds to the pre-mRNA via its zinc knuckle motif and a ring finger domain and also the Ysh1 nuclease of the CPF. It is essential for facilitating the cleavage activity of CPF but also selection of the cleavage site. *Saccharomyces cerevisiae* do not have micro-RNAs so they cannot perform microRNA mediated APA regulation, however the length and structure of the transcript has been shown to influence its stability and localisation. In fact, APA isoforms that have alterations in the 3' region of only a couple of nucleotide have shown differences in the decay rates of nearly four fold (Gupta et al. 2014).

Epigenetic modifications to DNA structure such as chromatin modifications can influence the PAS selection in *saccharomyces cerevisiae*. The histone methyltransferases *Set1* and *Set2* regulate the occupancy of nucleosomes around polyadenylation sites.*Set1* and *Set2* decrease efficiency of using a single pA site and are able to regulate APA induced by nutritional stress (Kaczmarek Michaels et al. 2020).

Yeast cells undergoing undergoing changes in cellular states i.e. glucose starvation in the diauxic shift response, show global increase in proximal pas site usage (Geisberg, Moqtaderi, and Struhl 2020). This is thought to occur via changes in elongation rates in RNAPII as the patterns of global APA

profiles are similar in both. Yeast strains with RNAPII that have reduced ability to process mRNA have normal rates of transcription elongation have normal patterns of polyadenylation. This suggests that is is the elongation speed of RNAPII that determines the pas site selection. Several other studies have linked changes in elongation rates of RNAPII to changes in PAS selection.

RNAPII transcription often has variations in its elongation rate. In yeast, a mutant form of RNAPII (*S. pombe Rpb1* slow mutant, found in a genetic screen for hypersensitivity to the nucleotide-depleting drug 6-azauracil) that slows the transcription elongation rate was found to cause widespread changes in APA (Yague-Sanz et al. 2020). The mutant RNAPII had APA driven changes in gene expression via two mechanisms; shortening the 3'UTR of transcripts and prematurely terminating the transcription of upstream non-coding RNAs, i.e. transcription interference, which can cause gene silencing (Yague-Sanz et al. 2020). Furthermore, the genes affected by mutant were identified to be involved in phosphate uptake and purine synthesis, which are essential for maintaining the intracellular nucleotide pool. This suggests that RNAPII acts as a sensor of nucleotide availability and that genes important for nucleotide pool maintenance have regulatory mechanisms that respond to reduced transcription elongation rates. The elongation rate of RNAPII influencing pas site selection was also demonstrated in (Rachael Emily Turner et al. 2021). The anti-cancer drug cordycepin increased the accumulation of intracellular GTP and promoted the use of distal pas sites. This was then reversed by the drug mycophenolic acid, which reduced GTP levels and caused a slower RNAP II elongation rate, which promoted the

use of proximal sites and therefore counteracted the effects of cordycepin on alternative polyadenylation (Rachael Emily Turner et al. 2021).

1.9 Metabolism of carbon sources by *Saccharomyces cerevisiae*

In this work, the analysis of transcripts is done using data from the yeast, *saccharomyces cerevisiae*, with a focus on how the position of the 3' end of transcripts is influenced by altered metabolism and so the next section provide a brief overview of yeast metabolism. The yeast *Saccharomyces cerevisiae* can utilize a range of carbon sources for energy production. Most commonly, yeast are grown in an excess of glucose (2%) in a rich medium (YPD) in batch culture. During this phase, the yeast can grow without oxygen using primarily glycolysis to produce ATP and NADH. Glucose metabolism in *Saccharomyces cerevisiae* begins with glucose uptake, facilitated by glucose transporters on the surface of the yeast cell membrane. Once inside the cell, glucose is metabolized through glycolysis, breaking it down into pyruvate, producing ATP and NADH via both aerobic and anaerobic metabolic pathways. The utilization of both anerobic and aerobic mechanisms to produce energy makes *Saccharomyces cerevisiae* a facultative anaerobe (Rosenfeld and Beauvoit 2003). These two pathways of ATP production via glucose metabolism in emphSaccharomyces cerevisiae are shown below in the diagrams 1.12 adapted from (K. R. Reddy and DeLaune 2008) on page 70.

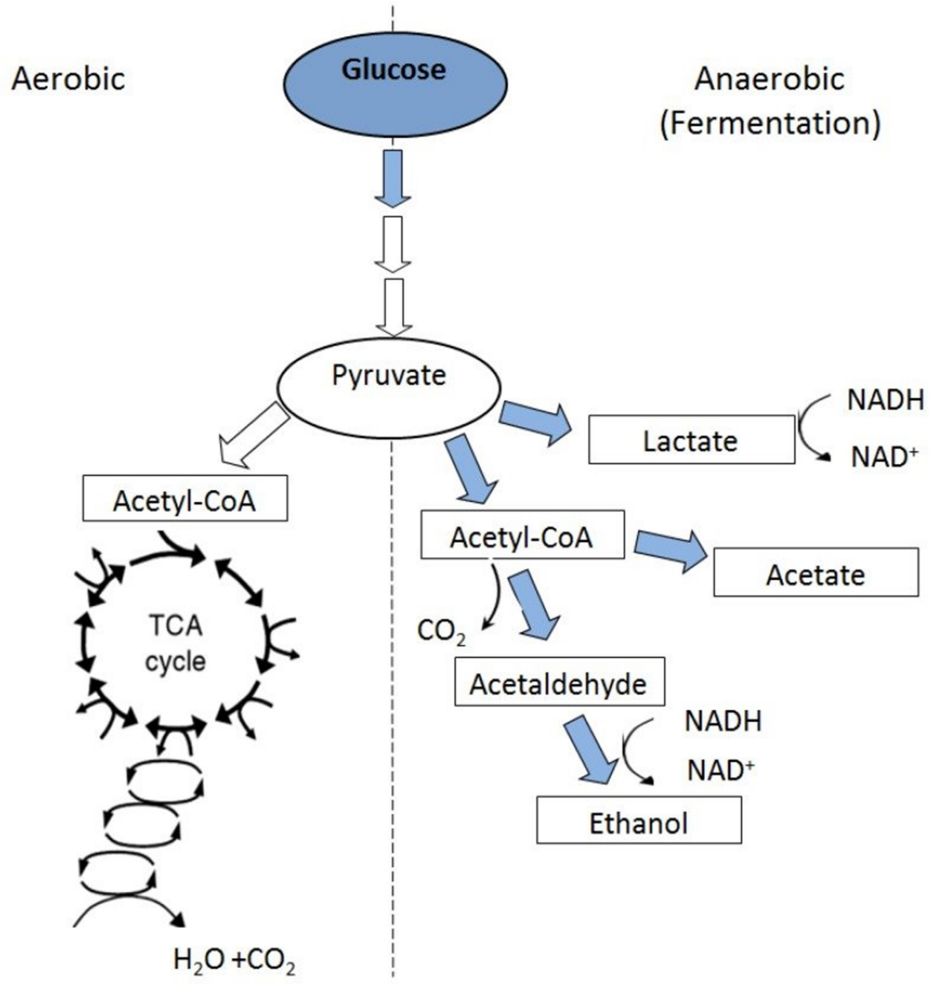


Figure 1.12: The two main metabolic pathways for carbohydrate source e.g. glucose in *saccharomyces cerevisiae*. *Saccharomyces cerevisiae* can respire aerobically to convert sugars into carbon dioxide and water. Due to the Crabtree effect, *Saccharomyces cerevisiae* even in oxygen rich conditions can convert sugars into carbon dioxide and alcohol in a process called fermentation, shown below diagram.

1.9.1 Physiology and growth conditions of *Saccharomyces cerevisiae* in fermentation

Saccharomyces cerevisiae is a unicellular fungus that requires a growth medium containing a constant supply of carbon source and oxygen to grow. These specific growth parameters require the use of a chemostat bioreactor. These bioreactors maintain a constant volume of liquid culture and a steady flow of nutrients and oxygen, pH and temperature, creating a stable environment for the yeast cell microorganisms to grow and reproduce (Ziv, Brandt, and Gresham 2013). The culture medium containing the microorganisms is constantly stirred and oxygenated, creating an optimal environment for their growth and reproduction. Optimising and controlling these parameters is essential for the systematic studying of cellular processes of these organisms as fluctuations in environmental conditions can alter their growth and increase variability in the data (Kerr and Dunham 2017). Furthermore, the chemostat provides acute control over growth parameters enabling investigations into these systems, such as the yeast metabolic cycle. Detailed explanation of the growth parameters of the chemostat bioreactor for the yeast metabolic cycle and glucose to galactose shift is shown in the methods section of this thesis. A diagram of the bioreactor used in the investigation of the yeast metabolic cycle is shown in the diagram 1.13 adapted from Shidong Xi PhD thesis Mellor lab 2020, on page 72.

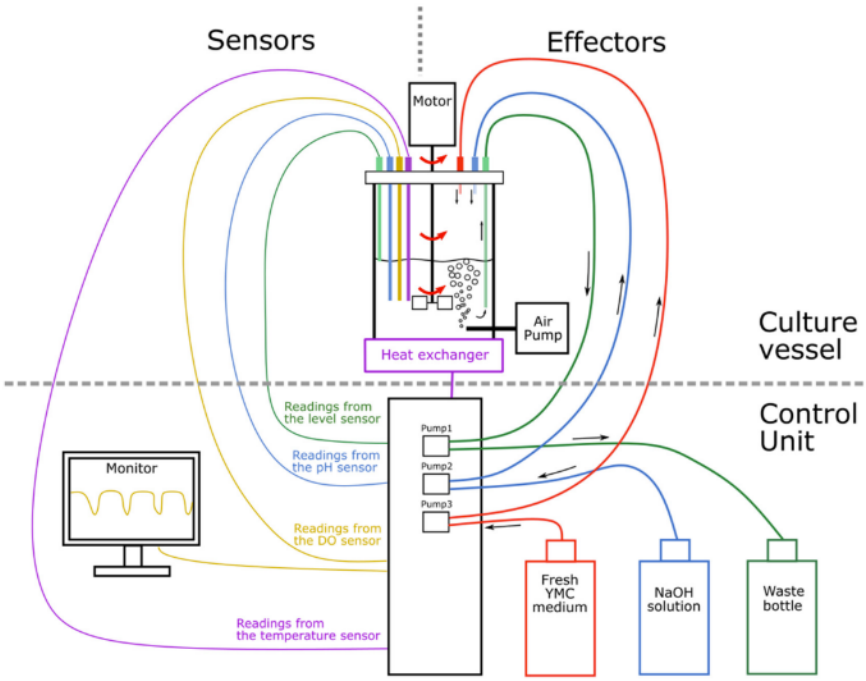


Figure 1.13: Schematic of chemostat fermenter, consisting of two main components: the culture vessel and the control unit. Within the culture vessel, various sensors (indicated by green for the level sensor, blue for the pH sensor, and purple for the temperature sensor) and effectors are installed. These sensors monitor alterations in the medium and transmit the collected data to the control unit. In response, the control unit directs the effectors to counteract these changes. The feedback loop depicted in this study ensures a consistent maintenance of pH, temperature, and culture volume. Adapted from Shidong Xi PhD thesis Mellor lab 2020

1.9.2 Yeast fermentation and the Crabtree effect in *Saccharomyces cerevisiae*

Although *Saccharomyces cerevisiae* in oxygen rich conditions is able to respire both aerobically and anaerobically, *Saccharomyces cerevisiae* preferentially consumes glucose via the fermentative pathway, even in the presence of high oxygen concentrations. This is known as the Crabtree effect (Perez-Samper et al. 2018). Initially the Crabtree effect appears counter-intuitive as ATP yield per glucose unit is nearly one order of magnitude higher via respiration than fermentation. However, it is believed that fermentation allows greater fluxes of ATP production, therefore favoring an increased ATP production rate over high ATP yield, which can lead to increased growth rates and therefore increased biomass (Van Hoek, Van Dijken, and Pronk 1998). The biochemical pathway of how *Saccharomyces cerevisiae* converts sugars such as glucose into ethanol, via the production and subsequent break down of pyruvate into ethanol and carbon dioxide is summarized in the diagram 1.14 adapted from (Maicas 2020) on page 74.

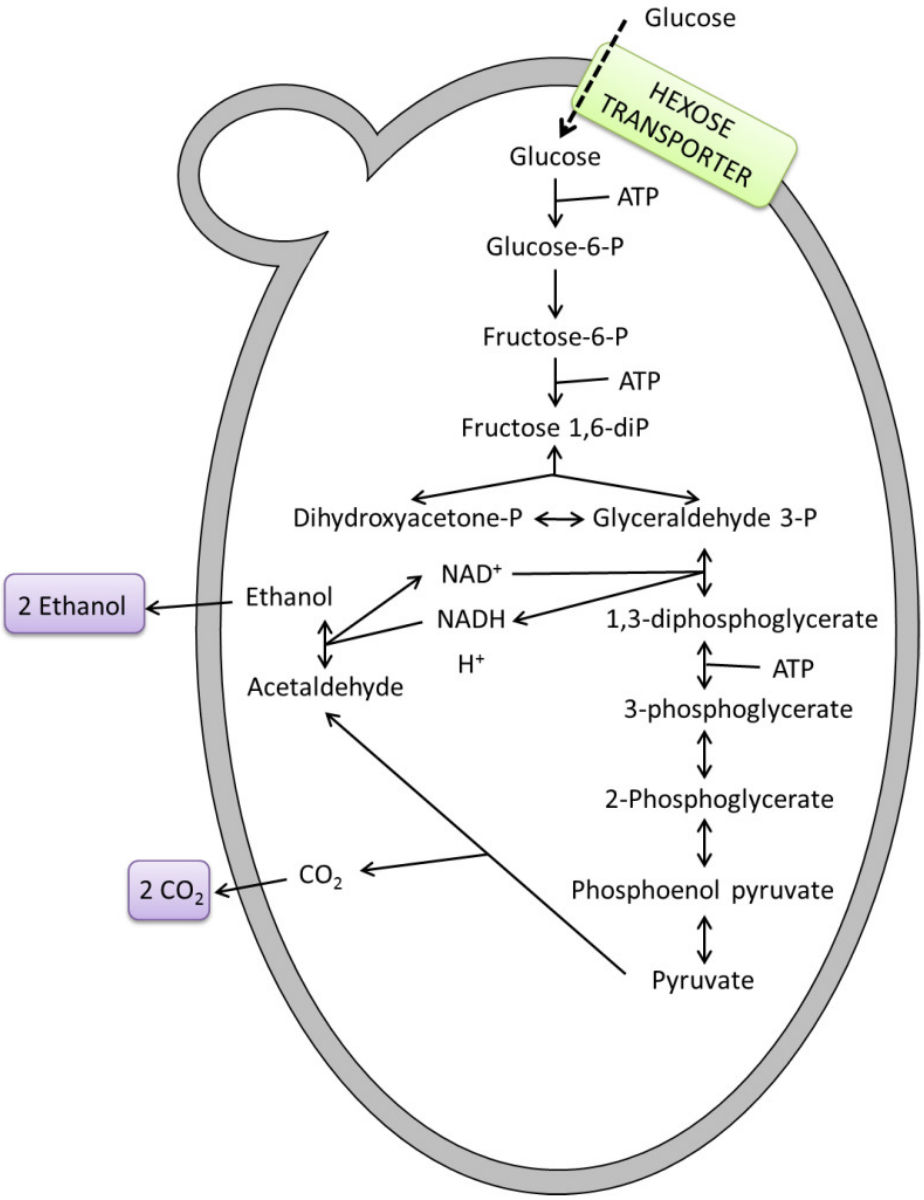


Figure 1.14: Yeast fermentation metabolic mechanism. The biochemical pathway of how yeasts convert sugars such as glucose into ethanol, via the production and subsequent break down of pyruvate into ethanol and carbon dioxide. Names of intermediate metabolites are shown, double ended arrows denotes a reversible reaction where reactant substrates and products are interchangeable, adapted from (Maicas 2020)

1.9.3 The diauxic shift in *Saccharomyces cerevisiae*

The diauxic shift occurs in when *Saccharomyces cerevisiae* is grown in the presence of two carbon sources. When glucose is present, yeast preferentially metabolizes it as its primary carbon source. Once the glucose source has been is depleted, the yeast undergoes a metabolic shift and begins to metabolize the second carbon source, such as galactose or ethanol that is produced as a by-product of glycolysis (Ostergaard, Olsson, et al. 2000). The diauxic shift is a complex process which requires the regulation and coordination of various metabolic pathways. When glucose is present, the yeast represses the expression of genes involved in the metabolism of alternative carbon sources, most importantly, the GAL gene regulatory network (Ostergaard, Walløe, et al. 2001). This repression is mediated via the protein Mig1, which binds to the promoters of these genes and prevents their transcription (Shashkova et al. 2017). The concentration of glucose in the cell drops due to its utilization as the primary carbon source, leading to the release of Mig1 from the promoters of the alternative carbon source genes such as the GAL genes. This allows the transcription of these genes to proceed enabling galactose metabolism. The diauxic shift is also accompanied by changes in the expression of other genes involved in glucose metabolism such as the expression of genes involved in the glyoxylate cycle, which enables the metabolism of acetyl-CoA to continue in the absence of glucose, is induced during the diauxic shift (Y. J. Lee et al. 2011). The induction of the GAL genes to facilitate the galactose metabolism during the diauxic shift is demonstrated in the diagram below in 1.15 adapted from (Ostergaard, Olsson, et al. 2000) on page 76.

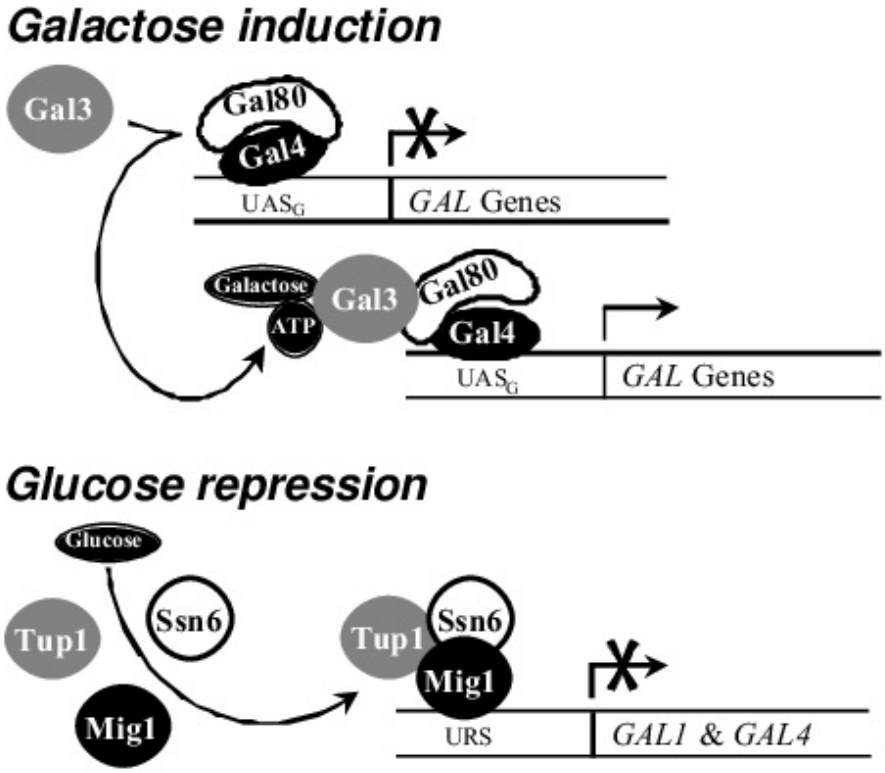


Figure 1.15: Glucose control of the GAL network in *saccharomyces cerevisiae* is regulated by intracelluar glucose levels via several mechanisms. To activate the GAL genes, *Gal4* acts as a positive transcriptional activator by binding to specific upstream activating sites (UAS G). During induction and repression of the GAL genes, *Gal80* binds to *Gal4*. However, under repressed conditions, *Gal80 binds* to *Gal4* in two regions, including the transcriptional activation domain of *Gal4*. For galactose induction, *Gal3* acts as a transducer by interacting with *Gal80*, which is bound to intracellular galactose. Mig1-mediated control prevents transcription of the *GAL1* and *GAL4* genes in the presence of glucose by participating in a protein complex with Ssn6 and Tup1 that binds to specific upstream repression sites (URS).

1.10 The yeast metabolic cycle

Yeast show alterative modes of growth depending on culture conditions. One of these is the yeast metabolic cycle (YMC), characterized by spontaneous and robust oscillations in intracellular metabolites and oxygen consumption when the yeast are cultured under constant conditions a chemostat or fermentor. In this work, yeast is cultured at a high cell density, (OD_{600} >4.0-16), constant pH (3.5), temperature (30° C), oxygen and nutrient availability (1% glucose and a dilution rate of 0.082 hr-1 (Feltham et al. 2020). Under these conditions, yeast show transient oscillations in levels of dissolved oxygen (dO_2) in the medium every 3.5 hours. This mode of growth starts after *Saccharomyces cerevisiae* is starved of nutrients for 7-14 hours and then provided with a constant flow of glucose containing medium. These oscillations are self sustaining and will continue indefinitely without interference (Klevecz and D. B. Murray 2001). An oxygen trace of *Saccharomyces cerevisiae* being induced into the indefinite YMC is shown in 1.18 on page 83, and is adapted from (Tu, Kudlicki, et al. 2005). In addition to oscillation in dissolved oxygen concentration, there are changes to intracellular and extracellular metabolites over the cycle , (Tu, Mohler, et al. 2007), Shidong's thesis. These includes spikes of ethanol in the medium and changes in the levels of intracellular storage carbohydrates (glycogen and trehalose) and are most simply explained by oscillations in intracellular glucose concentrations, despite a constant extracellular supply. These metabolite changes are coordinated with levels of dissolved oxygen. At high levels of dO_2, the intracellular glucose concentration remains constant, extracellular ethanol levels are low,

and levels of trehalose and glycogen stores start to increase over time, shown in the schematics 1.16 on page 81. Metabolism is oxidative, with the free glucose metabolized through glycolysis into the TCA to yield water, O_2 and ATP. At some point, in a proportion of cells in the population the storage carbohydrate levels become saturated and this signals (in an unknown way) their breakdown into free glucose. This excess glucose is metabolized, which is co-incident with the increase in oxygen consumption, indicating an increase in activity of the oxidative phosphorylation. As levels of dO_2 in the medium drop further, free ethanol appears in the medium, indicating that the TCA capacity becomes saturated and the excess glucose is metabolized via glycolysis into ethanol. Once all the glycogen and trehalose stores are used, oxygen consumption begins to reduce, the excess ethanol is metabolized and levels of dO_2 increase. The frequency with which the cells show this phase of increase oxygen consumption varies with the growth rate and available glucose. Coincident with these changes in intracellular metabolite concentrations and oxygen consumption rates, are oscillations to transcript levels (Tu, Kudlicki, et al. 2005) and (Feltham et al. 2020). These changes are likely to reflect increases in levels of acetylated histone proteins at the promoters of genes, although the mechanism behind this is not fully understood. These changes in levels of transcripts fall into three main superclusters known as OX, R/B and R/C that broadly reflect the period of increase oxygen consumption (OX), the period when it is decreasing (R/B) and the default state when levels remain high (R/C). These phases and the associated transcript peaks are sometime clustered into two groups known as high oxygen consumption phase/genes (HOC – the OX and R/B genes) and low oxygen consumption phase/genes

(LOC – the R/C genes). In total about half of the yeast genes show cycling in their levels over the YMC but this varies with growth conditions and the bioinformatic methods used to analyse them (Feltham et al. 2020). One of the remarkable features of the genes whose expression varies throughout the YMC is that they fall into very distinct gene ontologies (GOs). Current thinking suggests that the expression of distinct classes of genes may reflect the changes to the metabolic state of the yeast. Preliminary evidence suggests that protein levels do not oscillate in the same way as their transcripts, suggesting the oscillation in transcript do not drive the metabolic cycle per se (O'Neill et al. 2020). However, post-translational modifications to proteins do oscillate with the activity of the biological process they control suggesting that metabolites influence transcription and protein activity, but not sequentially as is often believed (Feltham et al. 2020). Nevertheless, using the GOs for transcripts whose levels peak at each phase (OX, R/B and R/C) is useful for analysis of changes to alternative polyadenylation sites and transcript expression levels throughout the YMC. These phases and the types of genes expressed are described in the next section.

There are three primary phases in the YMC; which are characterised by superclusters of gene expression changes which vary with oxygen consumption. These super clusters and the cellular functions that occur at these time points is summarised in 1.18 on page 83, and is adapted from (Tu, Mohler, et al. 2007). The first phase is the oxidative respiratory phase (OX), the second is the reductive building (RB) and the final phase is the reductive charging (RC). The OX phase is where respiration is maximal, although respiration also occurs at other phases of the YMC. At this phase the cell

prepares for cell growth and division by the expression of genes involved in protein synthesis machinery. There is a big surge in oxygen consumption and free radical production. During the RB phase the cell prepares for division. The genes involved in mitosis, such as spindle formation and histones are up regulated. The genes involved in mitochondrial production are also highly upregulated, in response to the previously intense metabolic state of the OX phase. In the final RC phase, the cell prepares for the next round of growth. This phase up regulates many genes involved in non-oxidative metabolism, such as ethanol consumption, carbohydrate catabolism and lipid metabolism. As the cell is repairing and preparing for the subsequent OX phase, the cell also produces stress and repair genes. These include, heat shock proteins, genes involved in autophagy and starvation. These super clusters of periodic gene expression are also accompanied by robust predictable fluctuations in metabolites, linking the expression of genes to specific temporal cellular events (Tu, Mohler, et al. 2007) These gene expression fluctuations along the YMC and the corresponding oxygen consumption levels are shown on the diagram 1.17 on page 82.

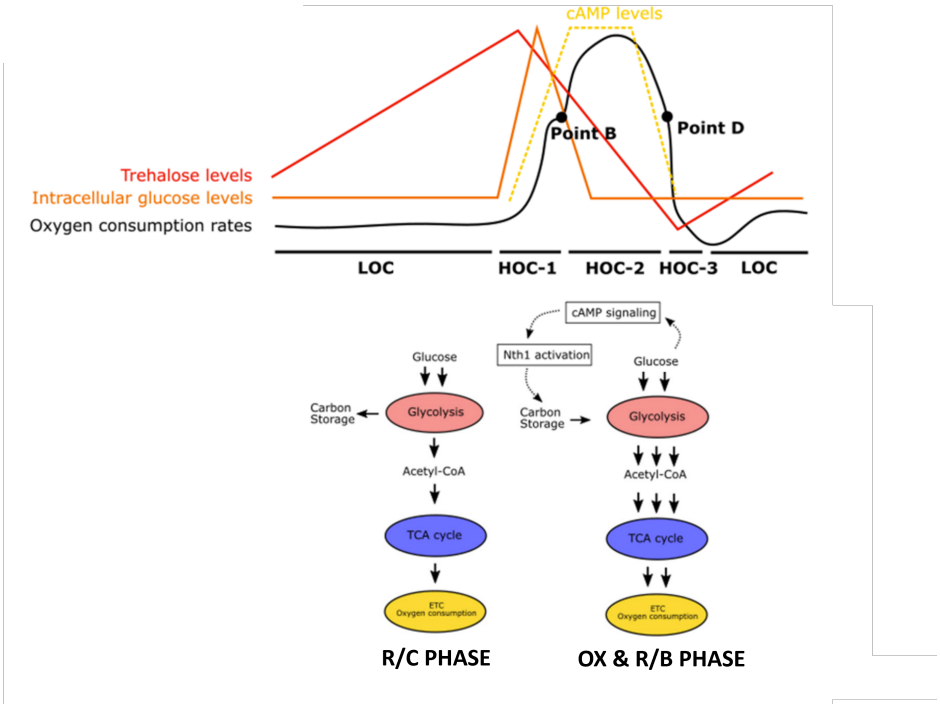


Figure 1.16: Model for the molecular mechanism behind the yeast metabolic cycle. The three major clusters are shown alongside changes in metabolites and glucose metabolism. The top panel shows the measured changes in cellular metabolites , plotted with solid lines in different colours, while hypothesised changes are plotted using dashed lines. The bottom panel shows changes in metabolism that occurs across the YMC. Adapted from Xi, Mellor Lab, 2019.

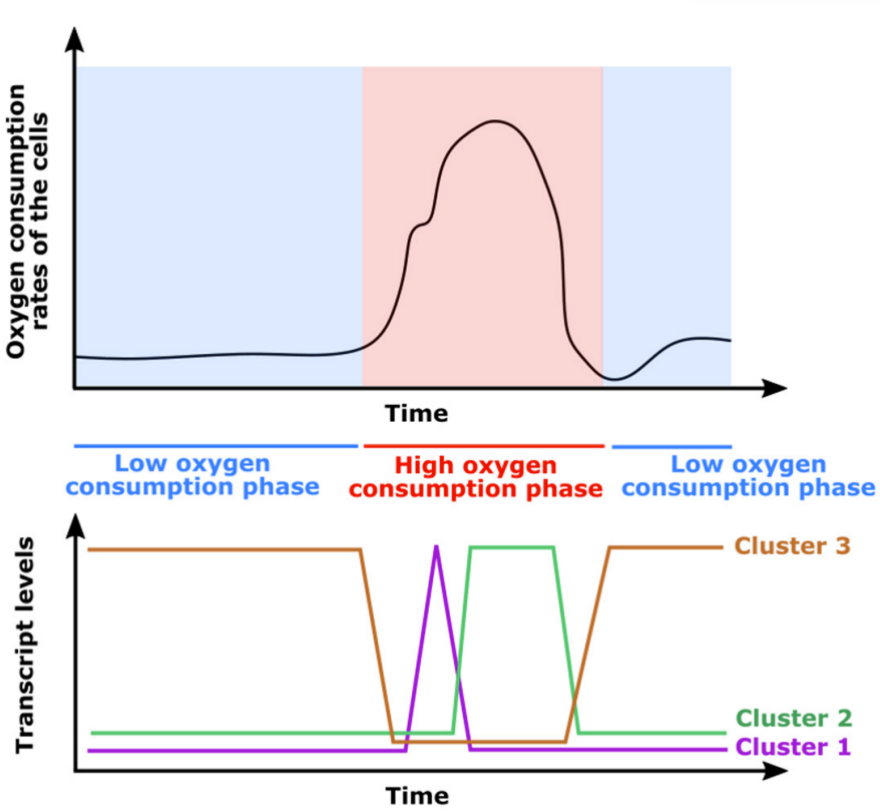
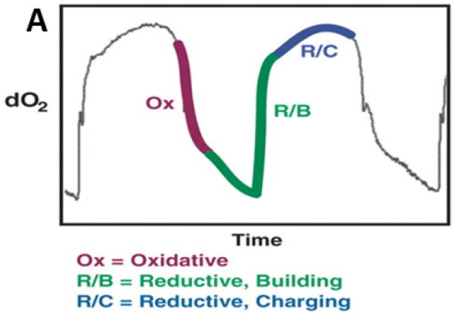


Figure 1.17: Oxgygen trace and gene expression changes in superclusters along the YMC. Schematic shows expression of genes throughout a cycle of the YMC. Red shows high oxygen consumption phase and blue is low oxygen consumption. The botom panel depicts how the levels of transcripts change throughout the three super cluster phases. Cluster 1 is the OX phase, Cluster 2 is the R/B phase and Cluster 3 is the R/C phase. Adapted from Xi, Mellor Lab, 2019.

A



B

YMC Phase	Gene expression profile
Oxidative (OX)	Ribosomal, amino acid synthesis, nuclear components (nucleolus, nuclear RNAs snRNA, tRNA, cell growth
Reductive building (RB)	Mitochondrial protein, DNA damage repair, Oxydative damage, cell division, histones spindle pole, DNA replication
Reductive charging (RC)	Peroxisomal proteins, carbohydrate metabolism, ethanol metabolism, stress/starvation genes (ubiquitination, vacuole, proteasome, autophagy, catabolism, heat shock proteins

Figure 1.18: Summary of 3 main super clusters of the YMC and physiological events that co-occur at these time points. A depicts the temporal location and order of the three super clusters. The X axis shows the dissolved oxygen levels, indicating a shift in aerobic to anaerobic respiration as the cells move through the three phases of the YMC. B i a table which characterises the Gene expression profile of each of the main phases in the YMC. Each phase has a distinct profile of genes that change in expression in specific functional groups to enact the shifts in metabolic mechanisms required to progress through the YMC. Adapted from (Tu, Mohler, et al. 2007).
.

1.11 Oscillations in transcript output during the yeast metabolic cycle

The YMC can be seen as a set of alternative metabolic phases related to growth and cell division. As mentioned, the YMC is hallmarked by roust osculations in a large proportion of its protein coding genes. These transcripts are often related to specific biological functions that are tightly affiliated with specific time points in the YMC. For instance, the cell cycle starts at the same time as high oxygen consumption phase of the YMC, which may ensure that DNA replication and cell division only occur when there is sufficient energy output of the cells. This also suggests that, despite ideas about reactive oxygen species causing DNA damage, cells in the OX phase that are competent to go through cell division (Burnetti, Aydin, and Buchler 2016). A summary of the main functional groups of genes that are expressed at different time points is shown on figure 1.18 on page 83, adapted from (Tu, Mohler, et al. 2007).

In the OX phase, there is a big increase in expression levels of genes involved in the production of growth and cell division machinery. This is identifiable as the genes that are unregulated are involved in cell cycle progression, including, translation initiation factors, ribosome synthesis, RNA processing. Genes involved in sulphur and RNA metabolism are also highly expressed at this phase. There is also a significant drop in many transcript levels as this time point, this suggests there is a period of active mRNA degradation occurring in the OX phase. This might be due to oxidative

stress, as the vigorous aerobic respiration leads to a large amount of free radical production that would cause significant damage to the transcripts (Tu, Kudlicki, et al. 2005).

In the RB phase, the cell has almost depleted its glucose stores and is repairing and rebuilding the mitochondrial machinery that was used in the previous, intense respiratory OX phase. Genes involved in mitochondrial protein synthesis are up regulated via post transcriptional stabilising events. The shift in metabolite availability of glucose depletion is a trigger for the post transcriptional regulation of mitochondrial encoding transcripts (Gancedo 1998). The cell shifts from high glycolytic metabolism to fermentable metabolites via inducing mitochondrial biogenesis to increase the ability to break down carbon sources through oxidative catabolism. The RNA-binding protein Puf3p stabilises mitochondrial transcripts during the RB phase, via binding to their 3'UTRs and prevents them from being rapidly degraded as under high glucose conditions (C.-D. Lee and Tu 2015). The RC phase repairs damage that occurred in the previous oxidative phase and NADPH stores are increased to act as a buffer for oxidative stress that will inevitably occur in the following OX phase (Cesur, Çakır, and Pir 2022). Genes that are up regulated at this time point are involved in carbohydrate and ethanol metabolism, stress genes, DNA repair and protein degradation (Rao and Pellegrini 2011).

1.12 Glucose starvation in *Saccharomyces cerevisiae*

The process of glucose starvation and shift towards galactose as a food source was also investigated alongside the YMC, in order to relate possible changes to transcripts and their site of polyadenylation to these metabolic changes. *Saccharomyces cerevisiae* is able to use a range of carbon substrates as a food source, and can adapt its metabolic machinery in order to optimize its utility of a particular source. The preferable sugar for yeast to consume is glucose, and its genetic output in regards to metabolism, will favour that of genes involved in the breakdown and utilization of glucose (Schüller 2003).

Glucose starvation was a chosen as a metabolic stress to investigate and detailed analysis of yeast growing in glucose and then shifted to galactose containing medium was used to identify APA in the Glucose to Galactose (Glu-Gal) shift. Once the intracellular glucose is used up, as it takes more than 60 minutes to induce the genes required to metabolise galactose and up to four hours to produce the proteins. During the first hour after glucose is removed from the medium, the yeast undergo a stress response with widespread and highly dynamic changes in transcription and transcript levels. In response to the starvation stress, the cells undergo changes in gene expression. This includes a decrease in the expression of genes involved in growth and ribosome biogenesis (RiBi), and an increase in the expression of genes involved in stress, carbohydrate metabolism, protein degradation, autophagy, and DNA damage repair. As glucose has been depleted, the cells

also shift towards using aerobic respiration over glycolysis for ATP production, resulting in an increase in the expression of mitochondrial transcripts in order to switch to other carbon sources of energy (Fischl, Brown, et al. 2021). As with the YMC, PUF3 is involved in the change in mitochondrial proteins to facilitate changes in the transcriptome via stabilising mitochondrial transcripts via interactions with their 3'UTR region (C.-D. Lee and Tu 2015). The changes to sites of polyadenylation and possible relationships to expression levels were examined and related to changes in the YMC.

1.13 Thesis Objective

APA is a crucial mechanism of gene expression regulation which can be rapidly induced to enable tight temporal control over the function, stability and location of a transcript. This dynamic adaptive mechanism to rapidly alter gene expression changes in response to changes in cellular states is therefore a good candidate to explore transcriptionally active and complex biological systems. The YMC is a dynamic process in which metabolic activity and cell cycle progression is synchronized to optimize cell growth. The YMC progresses through three major gene expression clusters that result in alterations of expression of almost half of the yeast genome. However, the APA profile throughout the YMC has not been explored and could reveal novel biological mechanisms. Overall mechanisms of global APA gene expression control are still not fully understood. If APA changes in response to different cellular states, exploring APA in a highly metabolically and transcriptionally active system could help us to unravel the function of APA. Investigating the APA profile in both the YMC and during glucose starvation can also identify APA events related to specific metabolic growth states. If there is a significant cross over in APA trends between the YMC and Glu-Gal shift, this could indicate a more generic APA response to overall shifts in metabolic activity.

Chapter 2

Materials and methods

2.1 Materials and Methods

The YMC is categorised by a highly dynamic global gene expression change to align the proteosome with the stages in the cell cycle. Whilst the regulation at the transcriptional level has been recognised as an integral regulatory step, we currently do not know how regulation at the co-transcriptional pre-mRNA proceeding step and at the post transcriptional level coordinates gene expression with highly temporal cellular processes. In this thesis, I aimed to address this shortcoming by interrogating alternative cleavage and polyadeylation profile changes during the YMC. To achieve this goal, I set out to address the following questions summarised below, combining wet lab and bioinformatic metrics as outlined below in the flowchart 2.1 on page 90.

1. Do APA profiles change during the yeast metabolic cycle?

2. Are transitions between stages in the YMC associated with a directional

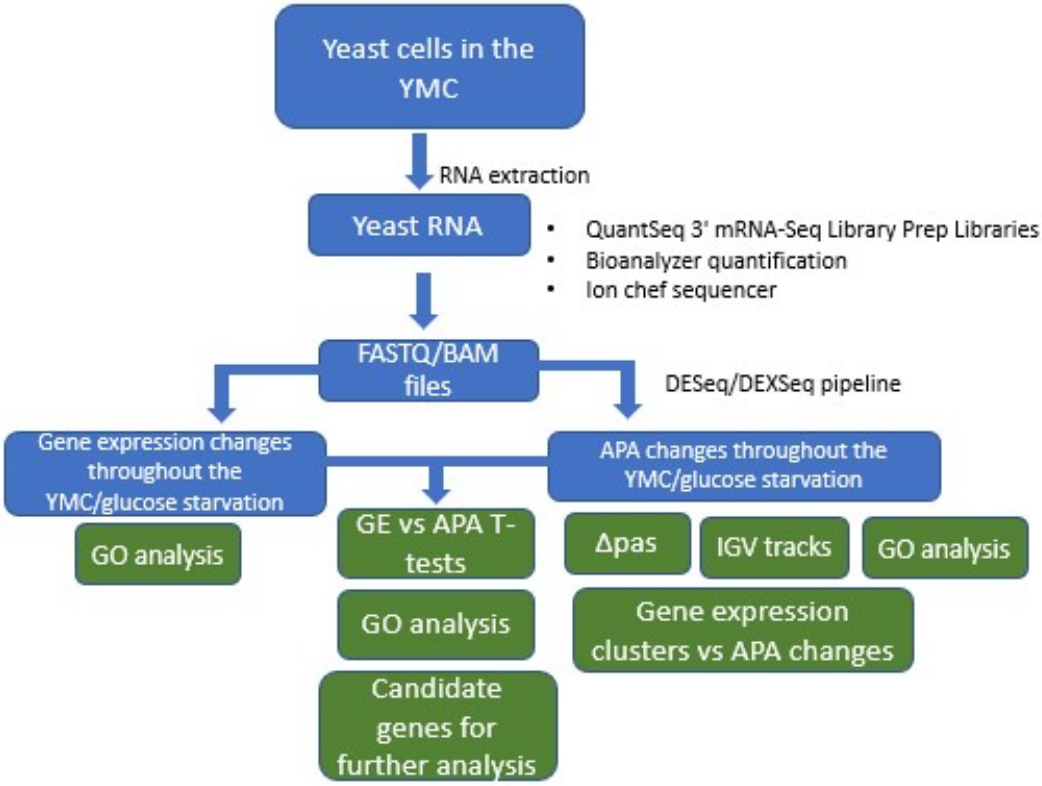


Figure 2.1: Flow chart for the experimental design to investigate APA events in *saccharomyces cerevisiae*. Experimental RNA and NET-seq data was generated by Shidong Xi feltham2020transcriptional. Blue boxes show experimental techniques used to investigate APA. Green boxes show output of these experiments and the results gathered from the data outputs of the blue boxes.

shift between long and short isoforms?

3. What is the physiological relevance of the APA profile changes during the YMC?

4. Are APA changes during the YMC a generic consequence of cellular stress or specifically regulated?

2.1.1 Wet laboratory materials and methods

2.1.2 Yeast culture chemostat fermenter conditions for the YMC

To extract RNA for 3'RNASeq at different time points in the YMC, the fermenter was set up using the following conditions. Firstly, the fermenter vessel was disassembled and thoroughly cleaned with MiliQ water and autoclaved. Next, 900ml of YMC medium was prepared in a 1L bottle and poured into the fermenter vessel. 0.5ml of antifoam 204 was also added through a syringe. The pH and dissolved oxygen probes were calibrated and installed onto the vessel, and the vessel was autoclaved again. Once the vessel cooled to room temperature, the condenser was installed in a flow hood and the probes and pipes were connected to the controller. A 10ml preculture was inoculated by transferring cells from a YPDA plate to YPD media and incubated until the following day. On the next day, the water cooler and compressor were switched on, the temperature and pH of the medium in the fermenter were adjusted, and the preculture was injected into the vessel. After the cells had used up all the glucose in the medium and entered starvation, feeding was begun and samples were collected for various purposes. When taking a small scale (1ml) sample, a 10ml lure-lock syringe was used to transfer the cells to a 15ml FALCON tube before transferring them to Eppendorf tubes. For a large scale (100ml) sample, a 100ml sampling bottle was used to transfer the cells to a FALCON tube. The large-scale sampling (100ml) typically disturbed the metabolic cycle, so only one sample was taken in a metabolic cycle. It was advised to leave a time interval

of at least 8 hours between two large-scale sampling events for the cells to recover. The large-scale sampling was intended for NET-seq, which required a large amount of starting material. To carry out this type of sampling, the glass filter unit was assembled before starting. The details of using the glass filter were referred to the NET-seq protocol for yeast published by Stirling Churchman (Churchman and Weissman 2011). A 50ml lure-lock syringe was used to extract 100ml of cell culture from the fermenter. The cell culture was immediately filtered through the nitrocellulose membrane. Cells were scraped off the membrane using a spatula. Subsequently, the cell pellet was snap-frozen using liquid nitrogen. The time of snap freezing was recorded as the sampling time. 100ml of cell culture from the fermenter typically contained 1.4g of cells, which was sufficient for one NET-seq experiment. To capture the original transcriptional status of the cells within the fermenter, the entire process had to be completed within 2.5 minutes.

2.1.3 Yeast strains for YMC and Glu-Gal shift RNA-Seq data

Wild type *Saccharomyces cerevisiae* yeast strains were used to harvest RNA for the YMC and Glu-Gal shift RNA-Seq data. These strains were procured from euroscarf, with the genotype: *CENPK113-7D MATa, URA3, HIS3, LEU2, TRP1, mal2-8c, SUC2* and had prototroph nutrient requirement, i.e. the same nutritional requirements as the origin wild type or non-mutant species.

2.1.4 Yeast culture conditions for Glu-Gal shift

Yeast were cultured under constant conditions a chemostat or fermentor. In this work, yeast were cultured at a high cell density, (OD_{600} >4.0-16), constant pH (3.5), temperature (30° C), oxygen and nutrient availability (1% glucose and a dilution rate of 0.082 hr-1 (Feltham et al. 2020). Under these conditions, yeast show transient oscillations in levels of dissolved oxygen (dO_2) in the medium every 3.5 hours. This mode of growth starts after *Saccharomyces cerevisiae* is starved of nutrients for 7-14 hours and then provided with constant flow of glucose containing medium. These oscillations are self sustaining and will continue indefinitely without interference.

2.1.5 RNA extraction protocol for 3'RNASeq

RNA was extracted for 3'RNASeq analysis using the mRNA 3'end sequencing protocol. The cells were washed twice with ice-cold PBS. 4ml of Trypsin was added to each 225cm3 flask. 10 ml of DMEM media was added to the cells, which was then centrifuged for 5 min at 1000 rpm at 4C. The supernatant was removed and the cell pellets where resuspended in 10ml of ice cold PBS and centrifuged. This step was repeated. The cell pellets were resuspended in 1ml of lysis buffer B.

50 ul of solution was removed for protein analysis. 1ul of Benzonase was added and the solution was incubated at 37degrees for 30mins. The remaining solution was centrifuged at 1,000 g at 4 degrees for 3 minutes.

500ul of the supernatant was transferred to a to new 1.7 ml tube. The fraction was centrifuged at 10,000 rpm for 1 min and the supernatant was transferred to a new tube to purify and kept on ice.

100ul of chloroform was added to the fraction. The solution was vortexed vigorously for 15 seconds. The solution was centrifuged at 12,000 g for 15 minutes at 4C. 750ul of the aqueous fraction was transferred to isopropanol the solution was left to Precipitate at room temperature for 10 minutes and were Spun at 12,000 g for 10 min at 4C. The supernatant was discarded and the pellet was Resuspended in 87 ul of RNase-free H2O. each fraction was DNase treated using the following volumes of reagent:

a. 10 ul 10 X DNase buffer b. 2 ul DNase I c. 1 ul RNase OUT

the solutions were Incubated at 32C for 30 miniutes. 100 ul of a 1:1 phenol:chloroform mix was added to each solution, which was then Vortexed and centifuged at 13000g for 5 minutes. The phenol:chloroform extraction step was repeated. The solutions were Precipitated in 250 ul of 100% ethanol and 10 ul of 3 M NaOAc for a minimum of 2 hours at -20C. the fractions were then centrifuged at 4C for 20 min and the pellet was resuspended in 32 ul of RNase free H2O.

Buffer Composition:

Lysis Buffer B For 50 mL 10 mM Tris-HCl (pH 8-8.4) 500 uL 1 M Tris-HCl 140 mM NaCl 1.4 mL 5 M NaCl 1.5 mM MgCl2 75 uL 1 M MgCl2 0.5H20 to 50ml

Detergent Stock For 10 mL 3.3% (w/v) Sodium Deoxycholate 330 mg of Powder 6.6% (v/v) Tween 40 660 uL of 100

2.1.6 QuantSeq 3'mRNA-Seq cDNA library preparation

cDNA libraries were made using the Lexogen QuantSeq 3' mRNASeq kit as per the manufacturer's instructions. 100ng of RNA was used for sequencing and was obtained via the RNA extraction experiments detailed earlier. The RNA was amplified using 14-16 PCR cycles.

2.1.7 hsDNA Bioanalyzer assay QC

CDNA libary concentrations and quality control was performed using the Agilent 2100 Bioanalyzer. The samples were placed on a hsDNA chip and the hsDNA bioanalyzer assay as performed as per the manufacture's instructions.

2.1.8 3' RNA-Seq Ion Proton chef

The cDNA 3'QuantSeq libaries were sequenced using the Ion PITM HiQ Chef Kit and ion Chef system as per the manufacturer's instructions. The libraries were diluted to 350pM before they were sequenced. The output FASTQ files were aligned to the *Saccharomyces cerevisiae* genome, sacCer3, using the Ion Torrent Server TMAP aligner with default settings.

2.1.9 Native Elongating Transcript-seq (NET-Seq) data for RNA samples taken during the YMC

The Nascent elongating transcripts sequencing (NET-Seq) was taken from Shidong Xi data sets that followed the protocol in (Churchman and Weissman 2011). The NET-seq libraries were sequenced using the illumina NEXTseq 500 sequencer, employing a high throughput flow cell, single end reads, and

50 cycles. In the original protocol, NET-seq typically required 100 million reads to achieve adequate coverage for each *S. cerevisiae* sample spiked with *S. pombe*. However, with the new protocol, 50 million reads proved to be sufficient. Initially, the adapter sequence 'ATCTCGTATGCCGTCTTCTGCTTG' was removed from the raw reads using cutadapt. Subsequently, bowtie was used to align the reads to the hybrid genome of *S. pombe* and *S. cerevisiae*, employing the following settings: '-q -p 16 -S -m 1 -n 1 -e 70 -l 28 -k 1 –best –phred33-quals'. The resulting uniquely aligned SAM files were converted to bedgraph files using samtools, bamtobed, and genomecov. For the sequencing of RNA-seq libraries, the Ion Proton sequencer with Ion PI Chip v3 was employed. Specifically, the 3' end sequencing of the *S. cerevisiae* sample spiked with *S. pombe* necessitated 6 million reads. The raw reads were aligned using STAR without employing cutadapt. STAR settings included '–runThreadN 2', '–alignIntronMax 2500', and '–outSJfilterReads Unique'. The subsequent data format conversions for RNA-seq were performed in the same manner as those for NET-seq

2.1.10 Native Elongating Transcript-seq (NET-Seq) data for RNA samples taken during the Glu-Gal shift

NET-seq data for Glu, Gal5, Gal15 and Gal60 were obtained from the Mellor lab, using ArrayExpress E-MTAB-4568 following the protocol in (Fischl, F. S. Howe, et al. 2017). The GLU-GAL shift growth phase starts after *Saccharomyces cerevisiae* is starved of nutrients for 7-14 hours and then provided with a constant flow of glucose containing medium, the first time point, GAL15 was taken 15 minutes after the introduction of galactose into the yeast

growth media. These oscillations are self sustaining and will continue indefinitely without interference. Yeast cells were cultured overnight in 25 mL of YPD medium (1% yeast extract (Difco), 1% bactopeptone, 2% glucose). The overnight culture was used to inoculate 1 L of YPD in a 2 L flask, resulting in an initial OD600 of 0.2. The cells were grown in a 2 L flask at 30°C and 170 rpm until reaching an OD600 of 0.65 (mid-log phase). To harvest the cells, they were filtered onto a 90 mm diameter nitrocellulose filter paper (Whatman) with a pore size of 0.45 m. The cells were then scraped off the filter paper using a spatula pre-cooled in liquid nitrogen and immediately flash frozen in liquid nitrogen. The frozen cells were subjected to grinding using a Retsch MM400 mixer mill with a 50 mL grinding jar (Retch) and a 25 mm stainless steel ball. The grinding was performed in six cycles of 3 minutes at a shaking frequency of 15 Hz, with the grinding chamber cooled in liquid nitrogen between cycles. 1 g of the resulting yeast grindate was stored at -80°C. To perform immunoprecipitation, 1 g of yeast grindate was resuspended in Lysis buffer A (20 mM HEPES, pH 7.4, 110 mM KOAc, 0.5% Triton X-100, 0.1% Tween 20, 10 mM MnCl2, 1x proteinase inhibitors (complete, EDTA-free (Roche)), 50 U/ml SUPERase.In (Invitrogen)). The resuspended grindate was incubated at 4°C for 20 minutes with 660 U of DNase I (Promega). Insoluble cell debris was pelleted by centrifugation at 16,000 g for 10 minutes at 4°C. The supernatants were collected and combined, with a 20 L sample taken as input and mixed with 20 L of 2× SDS loading buffer (80 mM Tris-HCl, pH 6.8, 200 mM DTT, 3.2% SDS, 0.1% Bromophenol Blue, 1.6% glycerol). The combined supernatants were added to 0.5 mL of pre-washed anti-FLAG M2 affinity agarose beads (Sigma) and

incubated at 4°C for 2.5 hours on a nutator. After incubation, the beads and supernatant were separated by spinning at 1,000 g for 2 minutes at 4°C. A 20 L unbound sample was taken and mixed with 20 L of 2× SDS loading buffer. The beads were then washed four times with 10 mL of Wash buffer A (20 mM HEPES, pH 7.4,

2.2 Bioinformatics methods for exploring APA profiles in saccharomyces cerevisiae

To explore the role of APA during the YMC, RNA was isolated at sequential time points during growth of saccharomyces cerevisiae in a chemostat or fermentor. After library preparation and sequencing, the data was subject to a range of bioinformatics pipelines to determine differential expression and potential APA during the main transition points of the YMC.

2.2.1 Characterisation of pas sites in saccharomyces cerevisiae

Before we explored APA during the YMC, we needed to characterise the annotated *saccharomyces cerevisiae* PAS sites as identified by the Tian laboratory tian2012signals. These sites were identified using 3'region extraction and deep sequencing techniques detailed in hoque2014accurate and liu2017comparative. Importantly these are PAS sites in mRNAs expressed during exponential growth in excess (2%) glucose and may not be identical to those identified in the YMC. Nevertheless, I identified the number and location of the pas sites to evaluate whether their distribution was biased and

if they clustered in ways that makes them unsuitable for further analysis. I thus characterised the annotated PAS in the context of their location and frequencies in their respective genes.

From this investigation, it was confirmed that in the saccharomyces cerevisiae genome, 56% of genes have multiple pas sites. Over the complete saccharomyces cerevisiae genome containing 7050 genes, 21934 unique pas sites were observed, giving an average of 3.11 pas sites sites per gene. The deqseq/dexseq pipeline annotated 3950 genes as having multiple pas sites, which is 56% of the 7050 genes with gene expression values. This is fitting with the literature, as 56% of genes have been recorded to have multiple pas sites tian2012signals. The average number of pas sites per gene was identified using the aggregate function in R. As the histogram in figure 2.2 on page 100 shows that almost half of the captured genes have multiple PAS. The pas sites showed no bias in regards to gene orientation and chromosomal location (data not shown). Similar numbers of PAS were located in the coding and 3-UTR regions of the genes. The average length of the distance between the different PAS (referred to as Δpas) was 155 nucleotides for the coding region and 122 nucleotides for the PAS located in the 3-UTR, as shown in the table 2.1 on page 101. In conclusion, the pas sites number and location were identified, which showed they were evenly distributed throughout the genome and occurred in similar frequency in the coding regions and 3-UTR. The average length of the Δpas was identified for both the 3-UTR and the coding region and was 155 and 122 nucleotides respectively.

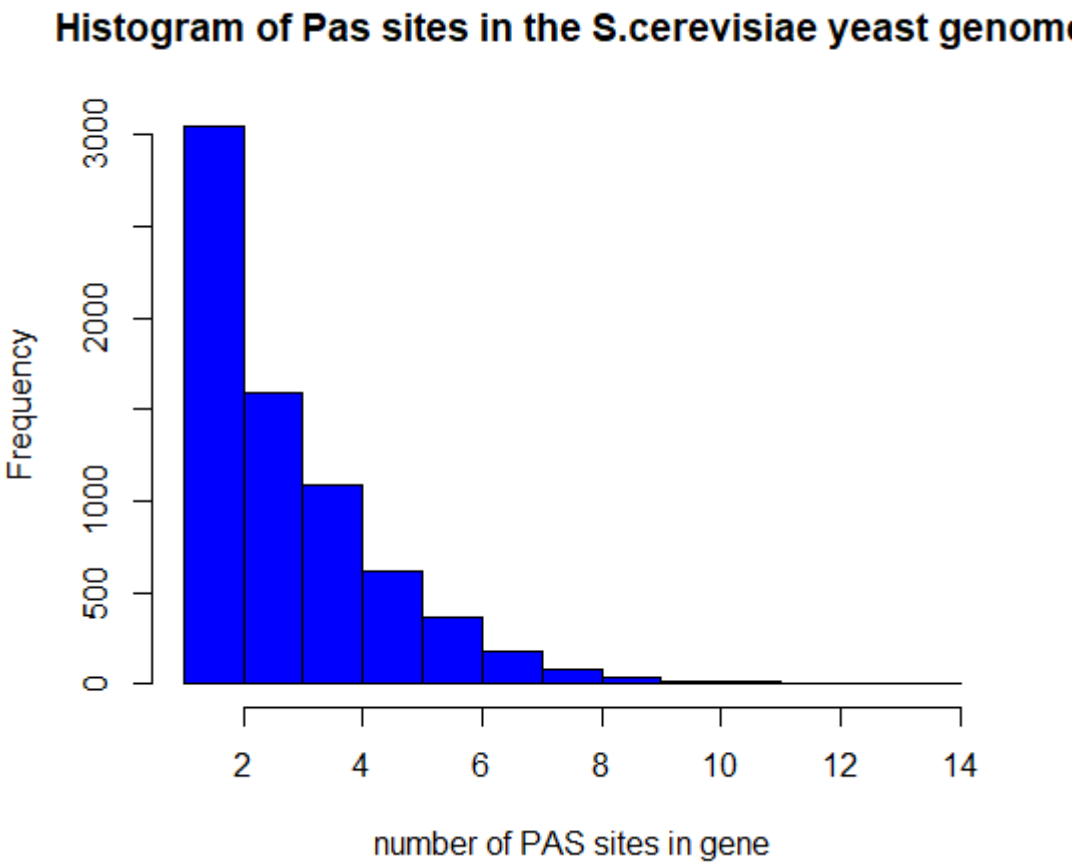


Figure 2.2: Histogram showing nearly half of the genes in saccharomyces cerevisiae contain multiple pas sites. Number of pas sites refers to number of pas in total in the genome as multiple can occur per gene. Mean Delta pas is the average mean length of the distance between two pas sites.

Type of polyA site	Number of pas sites	Mean Δpas
3'UTR	9226	122.3
Coding region (Exon)	12660	155.4
Coding region (Intronic)	29	2235.65
Total	21916	N/A

Table 2.1: Table breaking down PAS site location and mean distance between PAS sites (Δpas) in saccharomyces cerevisiae.

2.2.2 Selection of RNA-Seq data for APA and gene expression analysis

The development of RNA sequencing approaches has dramatically changed our understanding of how cells and organisms adapt their transcriptome to changing environments and in response to specific cues. I used existing data sets that were generated by Shidong Xi in the Mellor laboratory, from sequencing of RNA extracted from almost 50 time points evenly distributed during the YMC. The RNA was sequenced using the QuanrSeq 3-mRNA sequencing approach. Unlike full RNASeq, Quantseq generated data sets are ideal to interrogate APA as the reads are clustered around the polyAsites ma2019comparison. This means that the QuantSeq method pinpoints the exact 3-end of poly(A) RNA and therefore is preferable to traditional RNASeq methods. The method of generating 3-RNA end reads from Quantseq is shown in 2.3 on page 102.

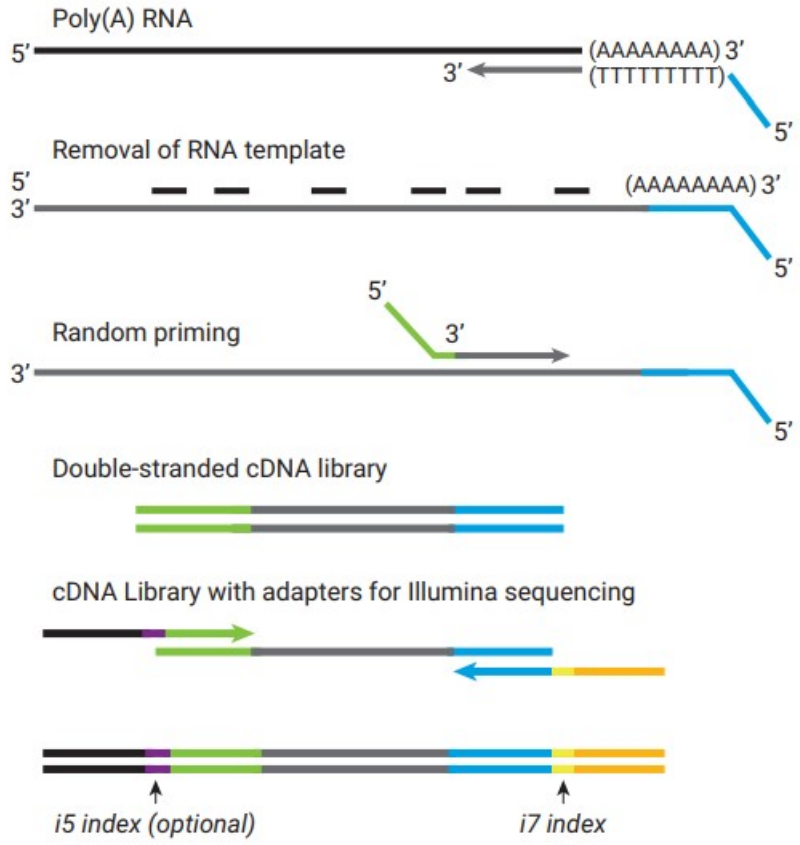


Figure 2.3: APA exploration using 3-mRNASeq enables the identification the exact pas site location. Taken from Lexogen website protocols WinNT. Sequencing using the i7 index primer will generate the unique 3'UTR sequence and report the position of the PAS. The libraries were made using the QuantSeq 3' mRNA-Seq Library Prep Kit by Lexogen, which one of the specific applications of these kits is the determination of alternative poly(A) sites. The cDNA libraries are built using a polyT oligos, which binds to the 3' UTR polyA tail, followed by random priming for second strand synthesis.WinNT.

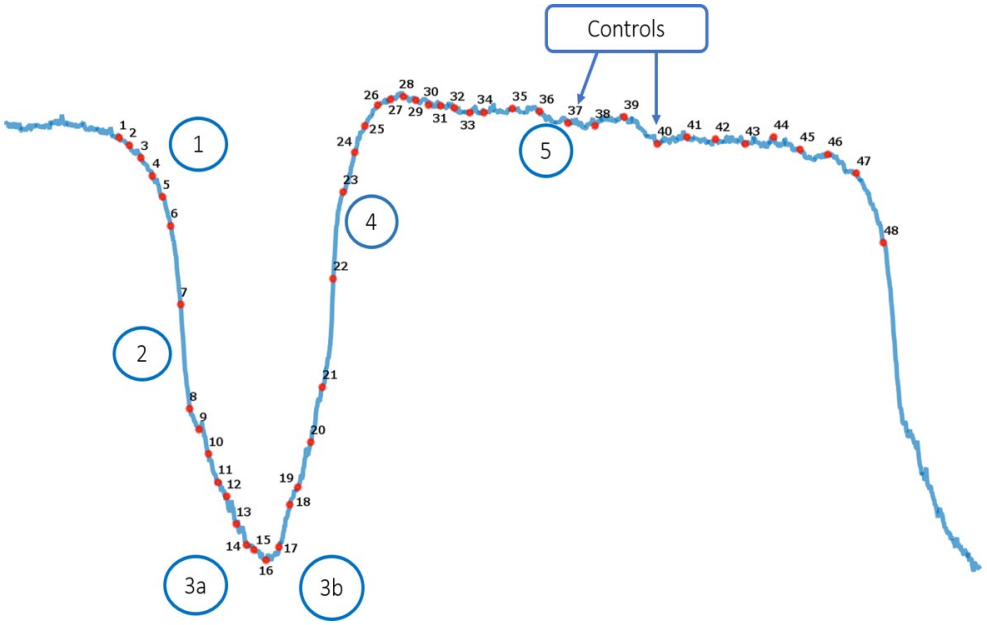


Figure 2.4: The RNA-Seq data sets available or the YMC are shown along the oxygen trace at the point in the cycle where they were extracted. The chosen time points are circled in blue and controls are identified with arrows and occur at the end of the cycle after the RC phase has ended and before the OX phase begins.

The YMC Quantseq data consists of 48 sample snapshots evenly distributed in the YMC. In order to use the dataset for APA analysis, I first divided the YMC data sets into 6 bins (1,2,3a,3b,4,5) that represented distinct stages in the YMC (as shown in table 2.2 on page 104. The bins capture the three major phases of the YMC, The OX, RB and RC. For each phase, 2 independent RNA data sets were chosen with similar read depths to use as repeats. Two RNA-Seq data sets from the middle of the RC phase were used as controls against which the gene expression and APA analysis were compared. The coverage achieved along the YMC using the time points selected is shown on the diagram 2.4 on page 103. The key for the time points used and the numbering system in showed in the table 2.2 on page 104.

YMC Time point	Repeats	YMC Phase
Controls	37&40	R/C late phase
1	1&3	OX early phase
2	8&9	OX mid phase
3a	13&16	OX late phase
3b	17&18	R/B early phase
4	22&23	R/B late phase
5	31&34	R/C early phase

Table 2.2: RNA-Seq time points along the YMC that were chosen for deseq/dexseq pipeline analysis. Phases in the YMC are defined by the oxygen consumption graph, as described in the introduction.

2.2.3 Pipeline for characterising the APA profile for the YMC

to process the RNASeq data set and interrogate APA, I employed a bioinformatics pipeline developed in the Furger lab by Mark Hickling and Cillian Brophy. The Furger DeSeq/DexSeq packages evaluates APA and gene expression changes of the RNA-Seq data. In order to process the YMC RNA data sets, the pipeline was modified for the *Saccharomyces cerevisiae* yeast genome using pas site mapping data from the Tian Laboratory (provided via personal communication).

DEXSeq is a R package that compares differential exon usage between two or more RNA-seq data sets. The script counts how many RNA reads map to each exon in a gene to give a ratio of the relative exon usage. This has been adapted to assess differential PAS usage in two or more conditions using the sum of the QuantSeq reads from the 3'end of the mRNA. First, a map of the PAS sites and their positions within the yeast genome was constructed, using data from the Bin Tian laboratory. Next, a counts table was created

relating the location of the pas sites which were then cross referenced to the mapped RNA sequencing reads. This modified DEXSeq code is able to align all of the RNAseq reads to the *Saccharomyces cerevisiae* genome and count how many reads fall within the boundaries of each annotated PAS site. The script then gives a P-adjusted (padj) value of pas site usage between two or more data sets, enabling the user to identify systematic APA changes across the data sets.

The use of these scripts comes with a few limitations. Firstly, the script is restricted to the comparison of only two different PAS sites in a gene and uses the top two most used PAS sites for this. As, the average number of PAS sites in a *Saccharomyces cerevisiae* gene is three, only changes between the top two most used PAS will be evaluated. Therefore, for each time point comparison for each gene, only the APA event with the largest change in ratio of PAS site usage will be recorded, which is the smallest padj value.

Another limitation of the investigation is the loss of resolution of pas sites that are very close together (sub 10 nucleotides). To assign and count reads, the pipeline selects a window of nucleotides surrounding each annotated exact pas site location to order for the script to allocate a pas site into categorical "bins". The bin sizes were initially set to 200 nucleotides either side of the pas map location for use with human data. These bins are too large for the *Saccharomyces cerevisiae* genome as the pas sites cluster far closer together than the human Hg19 genome that the pipeline was designed for, as the mean difference between pas sites in the 3'UTR is 122.3 bases. To fix this problem, the bins were reduced to 20 base pairs either side of the pas site for *Saccharomyces cerevisiae* so that more APA events can be analysed. This is

still a compromise however as pas sites less than 20 base pairs distance apart were not counted as APA events. 20 base pair window bins is the smallest resolution that can be used with the cDNA libraries due to repeat regions in the transcript that might cause erroneous categorisation of pas sites. This meant that pas sites with these ranges can not be distinguished between. An example of this issue is shown in 2.5 on page 106.

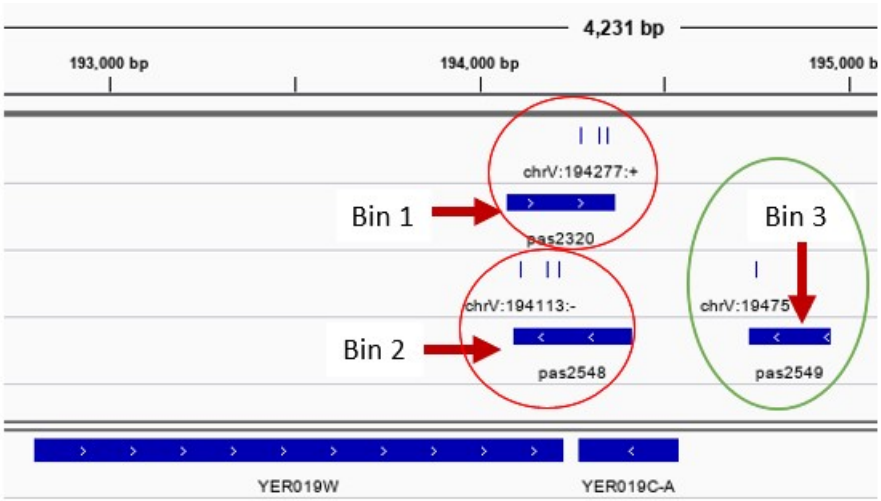


Figure 2.5: Screen shot of IGV track showing pas sites windows decreasing resolution from single nucleotide loci and to wider bin windows. Bin windows are +/- 20 nucleotides either side of the map location of the pas sites. The bins are shown as large blue bars and are indicated to using with a red arrow and labeled as 'Bin 1, Bin 2 and Bin 3'. Map locations for PAS sites from Tian map are shown as thin blue vertical dashes. The red circles show two examples of how very close together clustered PAS sites are grouped into the same PAS bin. The green circle shows the correct bin labeling for one PAS site that is further away from the other PAS sites

2.2.4 Bootstrapping data sets for statistical analysis

The 3-RNASeq data sets had to be modified in order to conduct statistical analysis. Using the D'Agostino, Pearson and Kolmogorov- Smirnov normalisation tests on the RNA-SEQ reads on time point 1 data set revealed that found the data did not follow normal distribution (data not shown). This is in line with expectations from the literature which suggest. Gene expression data should be investigated with the assumption of non normality (Mar 2019). Secondly, the RNA-Seq data sets for APA changes are limited to the number of APA events that occur, so are at risk of being skewed (e.g. ALL APA lengthening n = 12).

We fix these problems by bootstrapping. Bootstrapping is a statistical tool to create multiple simulated data sets from an original sample set. With the standard large sample statistical approach, the confidence intervals are defined for hypothesis testing from the sample size, n, of your population. The bootstrapping method is often used since the standard large sample statistical approach is limited, as it assumes the sampling distribution is normal and can be skewed by small n numbers. Bootstrapping is a useful tool for gene sequencing data as it is non-parametric method and does not ask for specific distributions.

By using the bootstrap method, we can more accurately calculate the sample variance and therefore have a more robust statistical test to investigate the null hypothesis. Bootstrapping the data means you randomly sample from the original data set with replacement to create another theoretical dataset . This is repeated 1000 times and all the statistical conclusions e.g

variance and standard distribution, are calculated using all the iterations of the theoretical data sets. 1000 iterations was chosen as this is the standard amount used for bootstrapping T-tests. All the data sets used to investigate a null hypothesis were bootstrapped to procure more robust statistical analysis.

T-tests were done on the bootstrapped data sets. The T-tests were used to test the hypothesis if the gene expression change means and/or medians of the transcripts that underwent APA changes were equal to the total data sets. The data sets were all compared to one another using unpaired two way t-test analysis in R studio. The data sets were bootstrapped 1000 replications using non parametric Bootstrapping tool in the Boot library in R studio berrar2019introduction. The T-tests were also performed in duplicate using graph pad prism 9 using the parameters of unpaired t-test assuming Gaussian distribution with welch's correction.

2.2.5 Cluster analysis of YMC APA events

The RNA-Seq data produced quantifiable gene expression and APA profiles of the YMC. This raw data is useful to see overall trends in gene expression and APA changes however more advanced data mining techniques must be applied to see more complex interactions between genes. One of the most common ways to group and categorise genes is by clustering. Cluster analysis is a powerful tool that can group large sets of genes, so that genes with similar profiles can be categorised together based on quantifiable similarity. For the YMC data sets, the most useful clustering profile would be based on the similarities between gene expression profiles of the genes that undergo

APA shifts. This is useful because we are investigating the potential impact APA changes might have on the transcriptomic profile of the YMC. Clustering based on temporal expression profiles of genes that undergo APA might uncover unknown relationships amongst genes that undergo similar types of APA and therefore suggest a functional reason for these APA shifts in the context of the YMC. The genes that underwent APA were clustered using the Maayan lab clustergrammer tool fernandez2017web. The program Clustergrammer is used to produce hierarchically clustered heat maps in highly complex, multidimensional datasets such as the ones for the YMC APA changes. Clustergrammer uses an unsupervised k-means clustering algorithm to categorise genes into groups based on the similarities of their characteristics gasch2002exploring. The tool uses Scipy library in Python to perform the clustering analysis. For the clustergrammer tool, the characteristics is the is mapped reads per million (RPM) of each gene that undergoes APA at each time point in the YMC. This was chosen in order to explore if different APA changes correlate to shifts in gene expression profiles in particular genes. The optimal number of clusters (the K number) was determined using the Elbow method, in which the number of clusters is identified when the intra-cluster variance is larger than the variance between data points inside the clusters, using the sum of squared distance to select an optimal k value based on the distance between all of the individual data points yuan2019research.

2.2.6 Identifying Δpas lengths

The Δpas is the length in base pairs between two proximal and distal pas sites on a transcript. The more upstream PAS sites is known as the

proximal site and the downstream site is the distal site. The distance between two PAS that undergo APA is an important metric to measure since changes in even a single nucleotide can affect the stability of a transcript clancy2008rna. Furthermore, investigations into correlations between transcript APA changes and aUTR lengths is important because preference for a specific aUTR length could be used identify motifs and structures associated with APA switching. The Δpas length is calculated using Bin Tians pas site map and is the output of the yeast dexseq/deseq pipline. The selection of a pas site (distal or proximal) will dictate whether the sequence of the aUTR is included in the mRNA transcript. T test of aUTR APA data sets vs aUTR of total transcripts. T test were conducted on bootstrapped datasets to compare the aUTRs of transcripts that had APA changes to the full data set of transcripts (Total). As the data sets of transcripts that undergo APA changes were smaller compared to the Total datasets, they were bootstrapped to n = 1000. This was done to reduce variance bias and to more accurately calculate the p value of the differences between aUTR lengths.

2.2.7 Cross comparison of APA changes and gene expression clusters

After the Gene Ontology and cluster analysis of the YMC APA data sets had been investigated, the APA genes were compared to YMC gene expression clusters that have been previously characterised. It was important to identify if the characteristics of the YMC functions being represented in the gene expression clusters were enriched with changes in APA. The gene expression clusters for the YMC was investigated by Shidong Xi, in

the Mellor laboratory, Oxford University Biochemistry Department. The temporal expression profiles of different genes throughout the YMC exhibit diverse shapes. The 3 super clusters were automatically separated into 16 sub-clusters using an unsupervised k-means algorithm. The genes that were in these clusters that also underwent APA changes were analysed for GO enrichment.

2.2.8 IGV visualisation of APA changes in the YMC

The APA changes that occur along the YMC can be visualised using the Integrative Genomics Viewer (IGV). IGV is a genome browser that visualises next-generation sequencing data. The browser shows all the RNA-Seq reads and aligns them to a reference genome to show depth of read coverage and how many reads are aligned to a single point in the genome. This acts as an interactive histogram of read density at each PAS, therefore showing how much each APA isoform is being expressed in each sample. IGV tracks were made for each YMC time point, therefore allowing us to visualise shifts in APA site usage. The IGV tracks were input from bigWig files. BigWigs are used to display of dense, continuous data made from BAM files. The initial RNA-Seq data uses BAM files that are converted to bigWigs in R using the GenomicAlignments package and exported using the the rtracklayer tool.

2.2.9 Bioinformatics Pipeline

The 3' RNA-seq data obtained from the wet lab sequencing experiments was analyzed for APA and gene expression changes using adapted R packages compiled into an APA analysis pipeline by Mark Hickling and Cillain Brophy in the Furger laboratory. This pipeline was adapted to the *Saccharomyces*

cerevisiae "S. cerevisiae sacCer3" genome via modifications shown below. A flow chart outlaying the steps of this pipeline and the outputs is shown in the experimental design flowchart 2.1 on page 90.

Counts Table Generation

The input counts table for the APA pipeline was generated using cleavage and polyadenylation (CPA) site annotations obtained from the Tian laboratory. This was a Pas site map with the chromosomal loci of all annotated pas sites in the *Saccharomyces cerevisiae* genome.

APA analysis pipeline

The pipeline for APA and gene expression changes was an adaptation of the DESeq and DEXSeq R packages (Love, Huber, and Anders 2014). The modifications of these pipelines so that they can be used to quantify APA changes was done by Mark Hickling and Cillain Brophy in the Furger laboratory. The counts table made in the previous pipeline was used as an input. For the *Saccharomyces cerevisiae* genome, the locations of the CPAs were put into window bin sizes of 20 nucleotides either direction (3'and 5') from the cleavage site. Pas sites with overlapping bin sizes were combined into single bins encompassing both windows. Reads were then reduced to their final 3' most nucleotide and cross referenced to the exon position in the saccer3 genome to obtain pas hits for each gene. The counts were obtained by counting how many CPAs occurred in each gene and then the differential partitioning of the CPAs was quantified using the DESeq2 R package. The output of this was a normalised read counts table in which the log2 fold change in the pas site usage between the top two most used pas sites was

identified. The significance of the log2 fold change was quantified with an adjusted p-value using the Benjamini-Hochberg method, which controls the rate of false positive hits by using a sequential modified Bonferroni correction, for multiple hypothesis tests. (Benjamini and Hochberg 1995) (Ferreira and Zwinderman 2006). The ratio of short to long isoform i.e. the direction and significance of shift in APA, was calculated by dividing the normalized reads of each APA isoform to give an RA (relative abundance) value.

2.2.10 Integrative Genomics Viewer

Integrative Genomics Viewer, IGV was used to select pas windows for the bioinformatics deseq/dexseq pipeline and also for visualisation of APA events. The bigwig files made from the bioinformatics deseq/dexseq pipeline were uploaded to IGV for visualisation.The IGV tracts from different time points were all normalized proportionality using the group auto-scale tool.

2.2.11 Yeast Mine Gene Ontology Analysis

Gene ontology analysis (GO analysis) was performed using Yeast mine online search tool. The list of genes that performed APA shifts were submitted for evaluation of Gene Ontology Enrichment under biological processes. This used default parameters without further test correction and the top 5 results with lowest p-value under 0.05 were recorded (Balakrishnan et al. 2012).

2.2.12 Statistical significance tests for GE and APA values

Statistical analysis was performed to test the hypothesis if the gene expression change means and/or medians of the transcripts that underwent APA changes were equal to the total data sets. The data sets were all compared to one another using unpaired two way t-test analysis in R studio. The data sets were bootstrapped 1000 replications using nonparametric Bootstrapping tool in the Boot library in R studio (Berrar 2019). The reliability of the T-tests was checked by redoing the tests in graph pad prism 9 using the parameters of unpaired t-test assuming Gaussian distribution with welch's correction.

2.2.13 NET-seq/TEF-seq sequencing data processing

NET-seq and TIF Seq data for Glu-Gal shift was taken and processed by the Mellor lab following the protocol in (Fischl, F. S. Howe, et al. 2017). The sequencing data in FASTQ format, with sanger-encoded quality scores (+33), were uploaded to the Galaxy platform (Blankenberg et al. 2010). Adaptor sequences were clipped from the reads, and any reads shorter than 15 nucleotides after clipping were discarded. To align the reads, Bowtie for Illumina was employed, using the S. cerevisiae genome 2011 assembly as the reference (Leek et al. 2010). Parameters were selected to discard non-uniquely aligned reads (-n 1, -e 70, -l 28, -k 1, -m 1). The resulting output was in the SAM file format, which was then converted to a BAM file and downloaded using tools from (Heng Li et al. 2009). For further analysis and visualization, the aligned reads were processed using R/Bioconductor.

Specifically, the reads were narrowed down to their 3' nucleotides, and BigWig files were created from these narrowed reads. The generated BigWig files were utilized for data visualization in the Integrative Genomics Viewer (IGV) Whenever available, data from combined biological replicates were used for visualization purposes.

Chapter 3

The APA profile of the Yeast metabolic cycle

3.1 Chapter aims

The aim of this chapter is to explore the APA profile of the YMC and the changes in gene expression that occur throughout this cycle. Interrogating the YMC APA profile via bioinformatics and statistical analysis of the APA pipeline outputs which include gene expression changes, APA padj values, GO analysis, cluster analysis, transcript lengths, and IGV tracks, we hope to gain insights into the global mechanisms of APA control and how it responds to different cellular states. This chapter investigates APA events in non-differentially expressed and differentially expressed genes independently, as APA has the capacity to impact gene expression, along with other factors such as localization and protein function, which remain unaffected by changes in gene expression. It is crucial to thoroughly investigate both datasets and

analyze the correlation between differentially expressed genes and APA.

3.2 Gene expression data shows many significant changes in gene expression throughout the YMC

Gene expression changes were identified for all 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC. This data set includes the genes that contain only one pas site and are unable to undergo APA events. Identifying changes in gene expression was an essential component of investigating the APA landscape as APA has been shown to affect the stability of transcripts and therefore the expression of genes (Cheng et al. 2020). Genes that undergo APA and gene expression changes are of special interest as the gene expression change might result from their APA shift, giving a mechanism for gene expression control in the YMC.

A gene expression increase or decrease was categorised as an event of normalised reads per million change that was <0.05 padj in statistical significance. There were significant changes in gene expression occurring throughout the YMC. Roughly 1/3 of all yeast genes underwent gene expression changes throughout the YMC. Of the 7050 genes sequenced, 1163 genes underwent GE increases (16.5%), and 962 undergo GE decreases (13.6%) and 136 genes underwent both increase and decrease in gene expression at different time points in the YMC. A significant number of gene expression increases and decreases occurred at every time point throughout the whole

YMC Time point	Gene expression increases	Gene expression decreases
1	43	73
2	350	222
3a	686	539
3b	689	684
4	339	339
5	42	23

Table 3.1: Table showing gene expression change events throughout the YMC for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC. A gene expression increase was categorised as an event of change of normalised reads per million that was less than 0.05 padj in statistical significance.

of the YMC. This is shown in the heat map 3.4 on page 124. The highest frequency of gene expression changes occurred at data points 3a and 3b. These time points marked the transition point between the point of maximal oxygen consumption and the beginning of the RC phase when oxygen consumption begins to drop off. The most transcriptionally active time point was 3b, all the significant changes in gene expression at this time point are shown in the scatter plot in 3.1 on page 120. The changes in gene expression during each time point of the YMC are summarised on table 3.3 on page 123. The full list of genes that undergo changes in gene expression throughout the YMC are shown in the table in the supplementary figures under Gene expression changes for YMC.

The lowest number of gene expression changes occurred at time point 5, which marked the end of the R/C late phase and entry to the next cycle. Overall, the YMC showed a small preference of genes increasing in expression than decreases throughout all of the recorded time points. The spread of gene

expression changes varied along the YMC. The time point with the highest logfold2 changes was at 2, that marked the middle of the OX phase. Time point 2 had a data range of 14.27 with the maximum logfold2 change in expression being 9.239 and minimum being -5.028 with a standard deviation of 0.88. The time point with the lowest standard deviation and range was time point 1, at the start of the OX phase. Time point 1 had a data range of 9.004 with the maximum log fold change in expression being 3.88 and minimum being -5.124 with a standard deviation of 0.74. The spread of the changes in gene expression were visualised on a violin plot. These plots combine box and kernel density plots to visualize the distribution of log fold changes in gene expression, as shown in 3.2 on page 122. The summary of statistics for gene expression is shown in the appendix table 6.1 on page 281. The kurtosis value is a quantitative measure of how much the data deviates from a normal Gaussian form of distribution. Any value above 3 shows a longer tail than normal distribution, which means they contain more outliers or data points with large standard deviations from the mean. The Skew kurtosis was much larger for time points 1 and 5 which were 5.253 and 6.553 respectfully. The lowest skew was at time point 3a, which was 1.34. This indicates the data sets with the least amount of extreme outliers were the ones at the center of the YMC. This skew was also demonstrated in the violin plots as shown in 3.2 on page 122. These plots show that the high kurtosis values came from extreme outliers, as the tails are very long but thin, indicating a small number of very large gene expression changes are skewing the data.

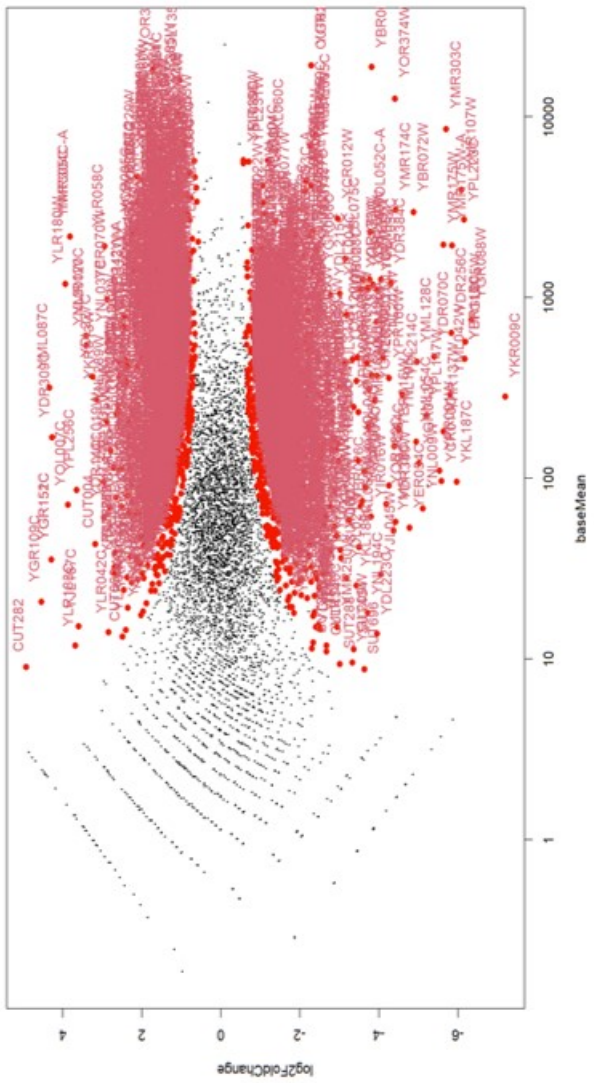


Figure 3.1: Scatter plot of YMC time point log fold2 change values of gene expression changes. The plot shows logfold change in normalised reads per million of each gene vs the number of normalized reads for expression change events throughout the YMC for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC.. Significant values of gene expression change were identified as differences in normalised reads per million compared to control that had a padj of less than 0.05. Significant changes are shown in red whilst nonsignificant values are shown in black.

The investigation into the changes in gene expression identified and quantified the gene expression changes for all genes in the YMC. The results concluded that the most transcriptionally active time points in the YMC are between time points 3a and 3b, which mark the peak of the OX phase and the transition into the early RB phase. These time point also had the least amount of outliers and had the least numbers of skew, which indicates the shifts in gene expression are due to a large number of genes all undergoing significant changes, therefore indicating this time point is very transcriptionally active for the whole genome. These results are useful for the APA analysis of the YMC as APA is tightly linked to changes in gene expression and allows us to identify genes that undergo both APA and gene expression changes throughout the YMC.

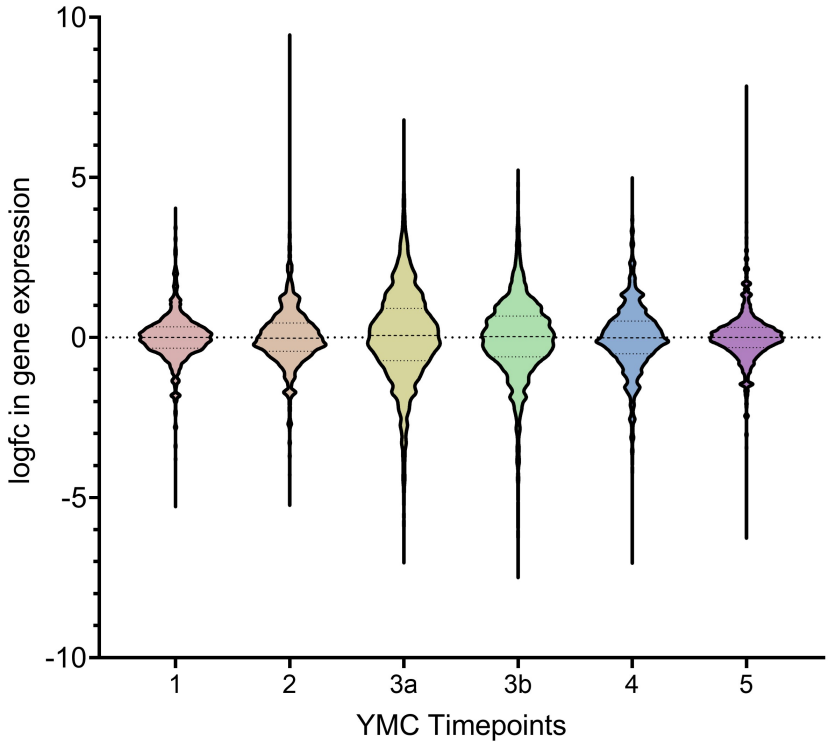


Figure 3.2: Violin plots of YMC time point logfold2 change values of gene expression change events throughout the YMC for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC. Logfc is the logfold2 change in read counts used to identify the gene expression change between the time point and the controls. Plots show the distribution in logfold2 change expression data to s visualize the distribution and skew of the data at each time points.

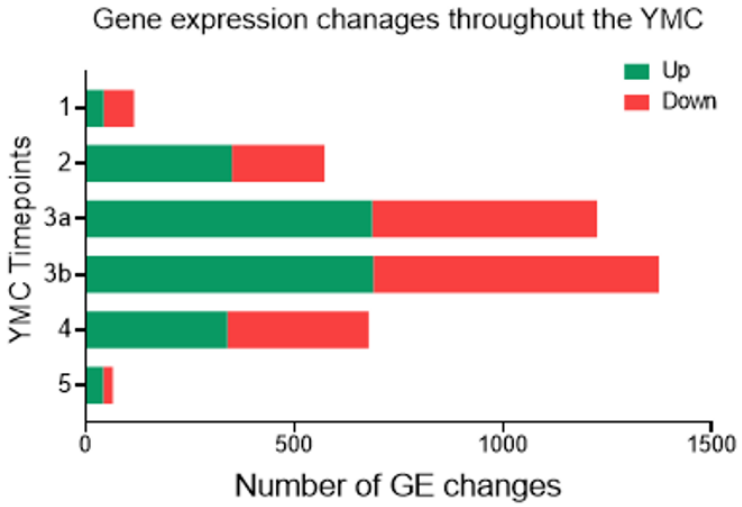


Figure 3.3: Bar chart comparing gene expression decrease and increase events throughout the YMC for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC. green bars shows for gene expression increases and red is the number of genes that undergo gene expression decreases.

3.3 APA changes observed in the YMC peak at time points 3a and 3b and show preference for lengthening events.

The gene expression data showed that a large number of genes are differently expressed in the YMC. The next step of analysis was to interrogate the pas site usage changes observed at different time points to explore the APA profile changes that occur during the YMC. We used the aforementioned deseq/dexseq pipeline to extract the 3'UTR, Coding and ALL APA (where changes of usage at two pas sites that are located in the coding region and 3'UTR). For these three APA categories there were many statistically

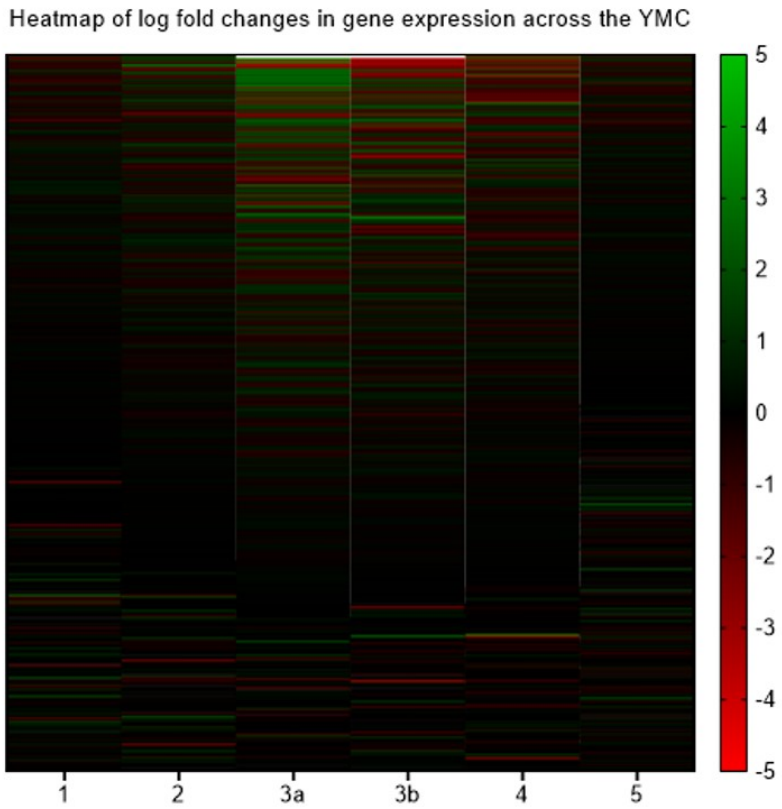


Figure 3.4: Heat map of logfold2 changes in normalized reads per million of genes for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the YMC.. Heat map shows peak of expression changes occur at time points 3a and 3b. Significant changes are identified as having a p value less than 0.05, scale shows logfold2 change values of changes in gene expression, gene expression increases are shown in green, decreases are shown in red. Black space show genes that do not significantly change in gene expression throughout the YMC.

significant events (Any value that was below 0.05 padj value was identified as significantly altering their PAS site usage compared to the control and therefore was recorded as an APA event). The APA data sets had very few overlaps in transcripts, indicating these changes were specific to a type of APA.

There were significant changes in APA occurring throughout the YMC in the coding region, 3'UTR, including changes between these two regions (named ALL). The most APA changes occurred at the peak of the YMC, at data points 3a and 3b for all data sets. The time points 3a and 3b are the time points where the YMC transitions out of the peak of the OX phase and enters the early RB phase. The 3b time point was more prevalent than the 3b time point in APA events for ALL and Coding APA. Shortening and lengthening APA events occurred in similar amounts throughout the YMC at each time point for the data sets 3'UTR and coding. For the ALL dataset, there was significantly more APA lengthening occurring than shortening at time points 2, 3a, 3b and 4. The most frequently used type of APA throughout the YMC was ALL APA and the least used was 3'UTR APA (30 vs 225 respectfully). Coding APA had over double the usage than 3'UTR APA, indicating that coding APA was the most used type of APA between the two, however, APA that changed the location of the pas site usage from either the coding region or the 3'UTR was the most prevalent overall.

This data shows that the most APA active time point of the YMC occurs between time points 3a and 3b which hallmark the end of the OX phase and start of the RB phase in the YMC. The data also showed that APA events tend to skew towards lengthening. These are significant findings as

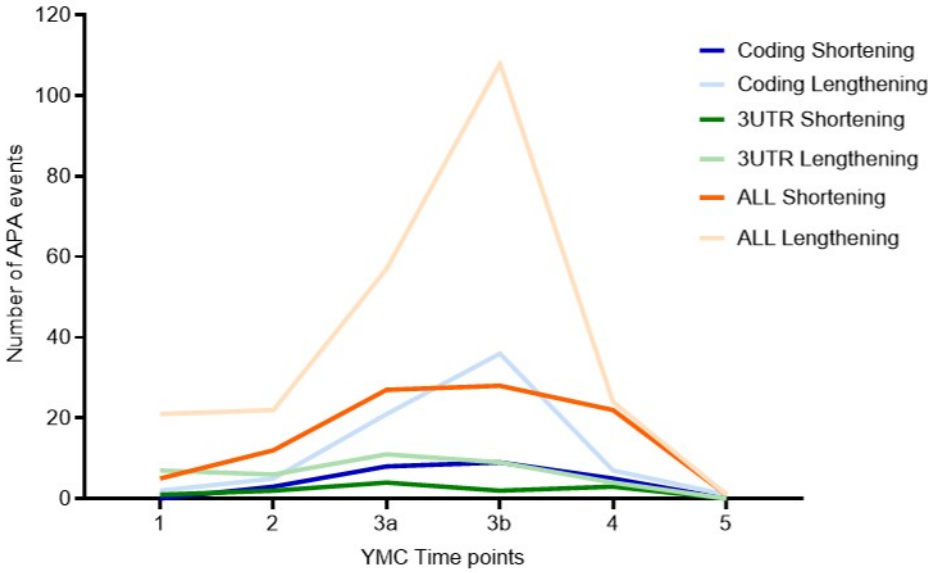


Figure 3.5: Line plot showing peak for all significant APA events occur at time points 3a and 3b which hallmark the end of the OX phase and start of the RB phase in the YMC. Plot shows peak of APA events occur at 3a and 3b with the biggest data set being ALL APA lengthening.

the most active gene expression time point matches the most prolific APA time points, indicating that APA might be influencing the gene expression profile or that changes in gene expression are correlated with changes in APA. Furthermore, time point 3a and 3b are the time point associated with cell cycle progression and metabolic shift, which are two systems that require significant transcriptional control, further implicating APA as a potential contributor to these key processes via manipulation of gene expression of a substantial number of genes.

3.4 The Coding APA profile of the YMC shows significant changes in APA at time points 3a and 3b skewed towards APA lengthening events

The next step of the analysis was the identification of the coding APA events. These Coding APA events were categorised as changes in pas site usage that both occur in the coding region of a gene. Coding region APA can alter the coding potential of the affected mRNA, which can therefore change the proteosome. The pipe line output showed multiple Coding APA events occurred at each time point along the YMC. Out of 1481 transcripts that were identified to have multiple pas sites in the coding region, 21 underwent shortening, 51 underwent lengthening events throughout the YMC. There was a significant increase in APA lengthening events at time point 3b, as shown in the table 3.4 on page 138. There was a preference for APA

lengthening at each time point throughout the YMC. The time point with the most APA events occurring was at 3b, in which 9 transcripts underwent APA shortening and 36 lengthened, as shown in the bar chart in 3.6 on page 129. The preference for APA lengthening changes occurring at time points 3a and 3b can be clearly seen in the heat map 3.7 on page 130. The lowest APA activity was at time point 5, with 3 transcripts shortening and 1 lengthening. The APA events were not skewed towards the most highly expressed genes. This is shown in the scatter plot 3.14 of the APA logfold changes vs the mean expression of the genes, on page 140. The tables for the 3'UTR APA events are shown in the supplementary figures under 3'UTR APA table for YMC. These results show that the coding APA is most prevalent at time points 3a and 3b, as with the gene expression changes and are skewed towards lengthening events. There is no bias for APA events occurring in genes with larger gene expression changes. This indicates that coding APA events are not just an artifact of highly expressed genes undergoing more variable use of pas sites as there is an equal distribution of APA events occurring at genes with variable amounts of gene expression changes. This adds further evidence to the fact that APA is not just a random occurrence at highly expressed genes but an intentional phenomenon that may reflect changes in the metabolic state and polyadenylation during the YMC.

YMC Time point	APA Shortening	APA Lengthening
1	0	2
2	3	5
3a	8	21
3b	9	36
4	5	7
5	0	1

Table 3.2: Table showing number and distribution of the Coding APA events occurring at each time point throughout the YMC showing peak of Coding APA events occurring at time points 3a and 3b with more lengthening events occurring at each time point compared to shortening

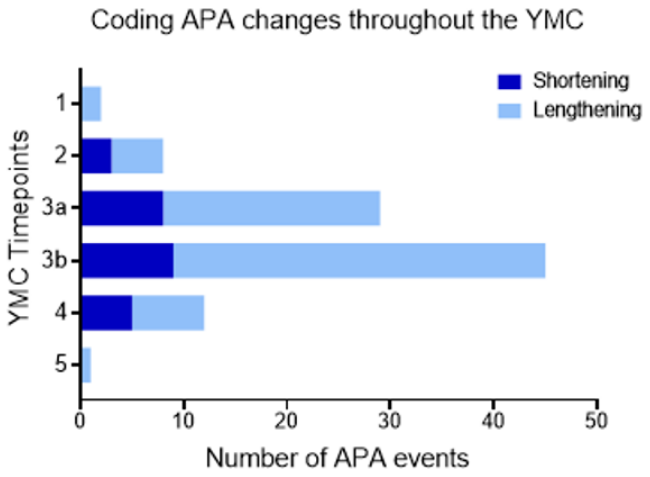


Figure 3.6: Bar chart showing the distribution of APA coding changes throughout the YMC. Chat shows peak of Coding APA events occurring at time points 3a and 3b with a preference for APA lengthening events at all time points.

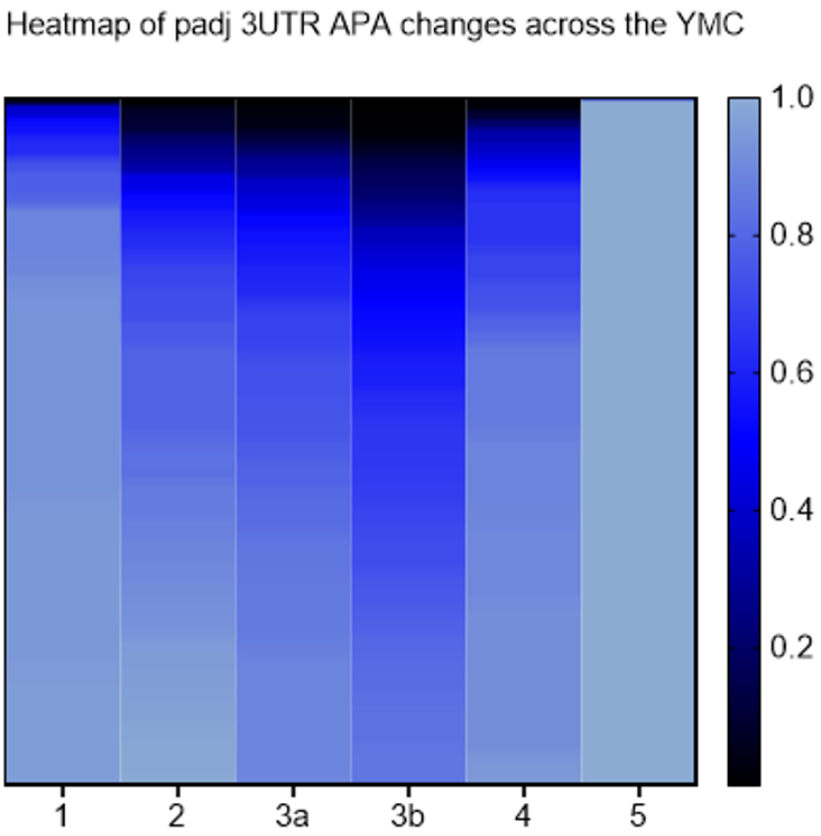


Figure 3.7: Heat map of logfold2 changes in normalized reads per million of genes that undergo coding APA throughout the YMC, significant APA identified as genes with less than 0.05 padj values. Heat map shows peak of coding APA changes occur at time points 3a and 3b. Only genes that undergo significant APA events were included in the heat map.

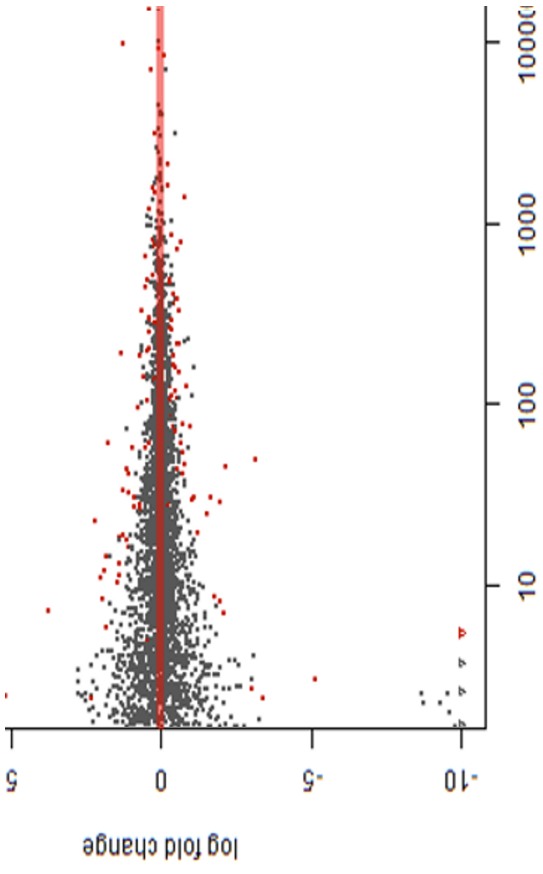


Figure 3.8: Scatter plot of gene expression vs APA events for 3b time point of the Coding data set. APA events are not skewed towards the most transcriptionally active genes as shown in this scatter plot of the 3b data set. The red dots indicate APA events that are significant with a padj value of less than 0.005, the grey dots are events that are not significant changes in APA.

3.5 3'UTR events occur less frequently than CR-APA in the YMC

In this section, APA events were categorised as changes in pas site usage that both occur in the non-coding 3'UTR region of a gene. Changes in the 3'UTR region of an mRNA transcript have been shown to alter the stability and localisation of the transcript (Yuan et al. 2021). The pipe line output showed several 3'UTR APA events occurred at each time point along the YMC. Out of 1976 genes, in total, 9 transcripts underwent shortening and 21 underwent lengthening events throughout the YMC, as summarised on table 3.3 on page 133. There was significantly more lengthening events at each time point throughout the YMC.The time point with the most 3'UTR APA events occurring was at 3a, in which 4 transcripts underwent APA shortening and 11 lengthened as shown in the bar chart in 3.9 on page 133. There was a preference for APA changes occurring at time points 3a and 3b compared to other time points , as evident in the heatmap 3.10 on page 134. The lowest APA activity was at time point 5, with 10 transcripts shortening and 2 lengthening. As with the coding data, The 3'UTR APA events were not skewed towards the most highly expressed genes. This is shown in the scatter plot 3.11 of the APA logfold changes vs the mean expression of the genes, on page 135. The tables for the Coding APA events are shown in the supplementary figures under Coding APA table for YMC.

YMC Time point	APA Shortening	APA Lengthening
1	1	7
2	2	6
3a	4	11
3b	2	9
4	3	4
5	0	0

Table 3.3: Table showing number and distribution of the 3'UTR APA events occurring at each time point throughout the YMC showing peak of 3'UTR APA events occurring at time points 3a and 3b with more lengthening events occurring at each time point compared to shortening.

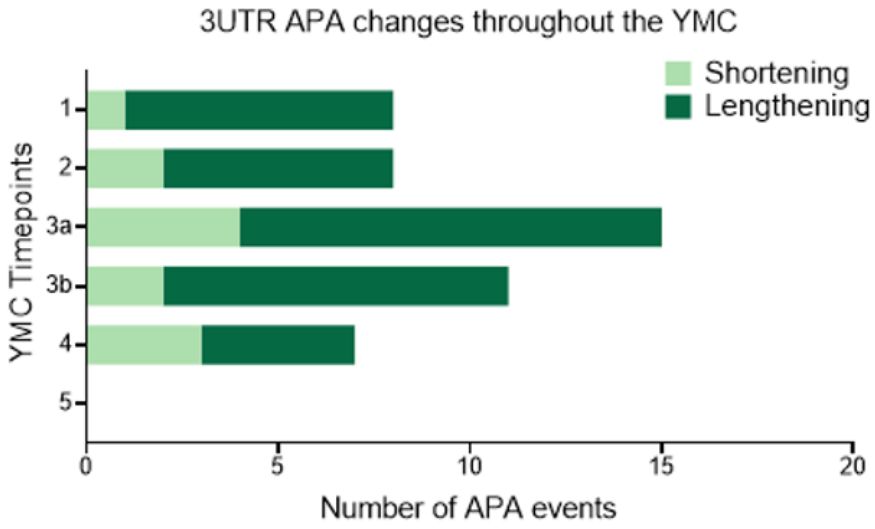


Figure 3.9: Bar chart for 3'UTR APA changes throughout the YMC showing most events occur at time points 3a and 3b with a preference for APA lengthening events. Light green bars show APA shortening and dark green bars show APA lengthening

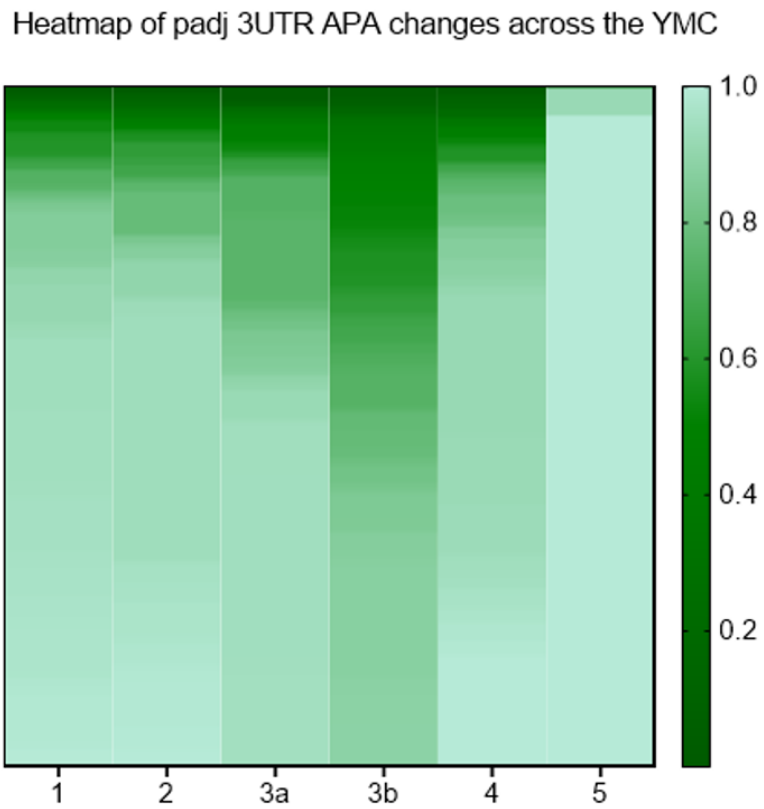


Figure 3.10: Heat map of logfold2 changes in normalized reads per million of genes that undergo 3'UTR APA throughout the YMC, significant APA identified as genes with less than 0.05 padj values. Heat map shows peak of 3'UTR APA changes occur at time points 3a and 3b. Only genes that undergo significant APA events were included in the heat map.

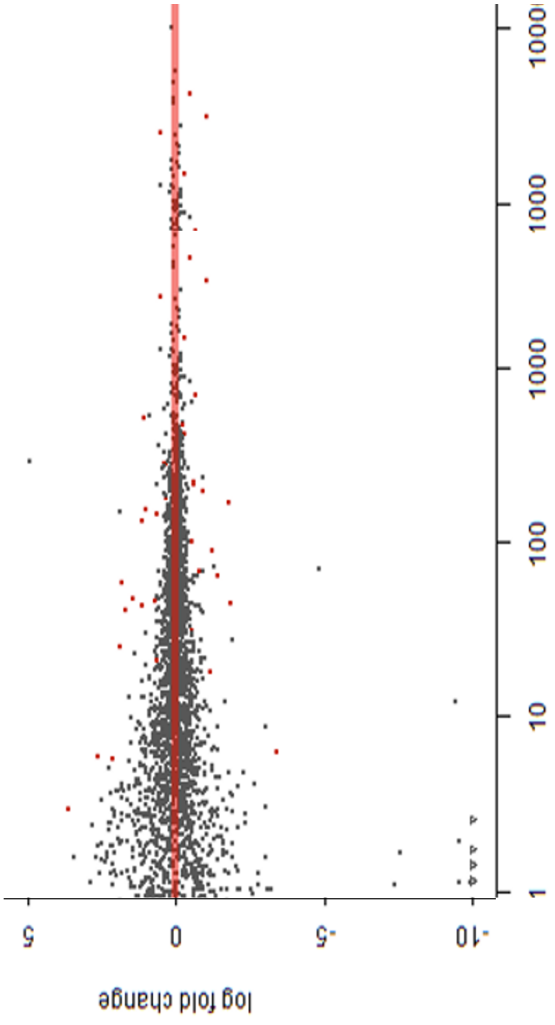


Figure 3.11: Scatter plot of gene expression vs APA events for 3b time point of the 3'UTR data set. APA events are not skewed towards the most transcriptionally active genes as shown in this scatter plot of the 3b data set. The red dots indicate APA events that are significant with a padj value of less than 0.05, the grey dots are events that are not significant changes in APA.

In summary, these results show that few 3'UTR APA event occur along the YMC. Of the few events that do occur, the most prevalent time points are again at 3a and 3b. The events are also skewed towards towards APA lengthening. As with the coding data, there is no bias for APA events occurring in genes with larger gene expression changes. This indicates that 3'UTR APA events are also not just an artifact of highly expressed genes and provides additional evidence to the fact that APA is a phenomenon linked to changes in the metabolic state during the YMC.

3.5.1 ALL APA data set shows global lengthening events dominate the APA profile of the YMC

ALL APA events were categorised as changes in pas site usage from the 3UTR to the coding region or vice versa between the two most commonly used pas for that gene that had a padj value below 0.05. The pipe line output showed multiple ALL APA events occurred at each time point along the YMC. Out of 4869 genes, 73 underwent shortening, 152 underwent lengthening events throughout the YMC. There was significantly more lengthening than shortening events occurring at time point 3a and 3b. The time point with the most ALL APA events occurring was at 3b, in which 28 transcripts underwent APA shortening and 108 lengthened as shown in the bar chart in 3.12 on page 138. The preference for ALL APA changes occurring at time points 3a and 3b can be seen in the heatmap 3.13 on page 139.the lowest APA activity was at 5. As with coding and 3'UTR, the ALL APA events were not skewed towards the most highly expressed genes. This is shown in the scatter plot 3.14 of the APA logfold changes vs the mean expression of

the genes, on page 140. The tables for the ALL APA events are shown in the supplementary figures under ALL APA table for YMC.

YMC Time point	APA Shortening	APA Lengthening
1	5	21
2	12	22
3a	27	57
3b	28	108
4	22	24
5	1	1

Table 3.4: Table of ALL APA events throughout the YMC showing most events occur at time points 3a and 3b with a preference for APA lengthening events.

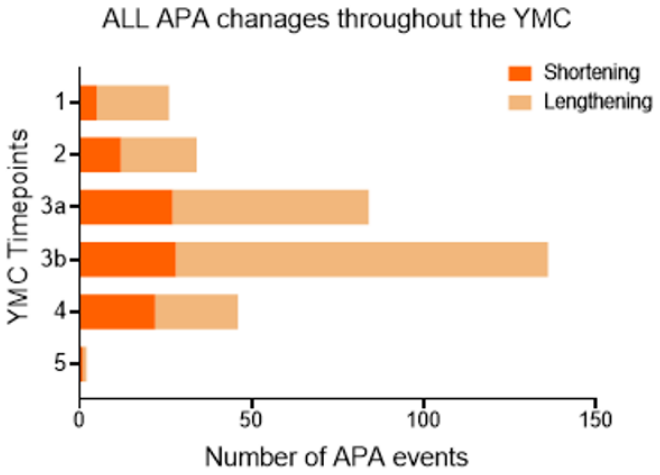


Figure 3.12: Bar chart of ALL APA events throughout the YMC showing most events occur at time points 3a and 3b with a preference for APA lengthening events. Light orange bars show APA shortening and orange green bars show APA lengthening

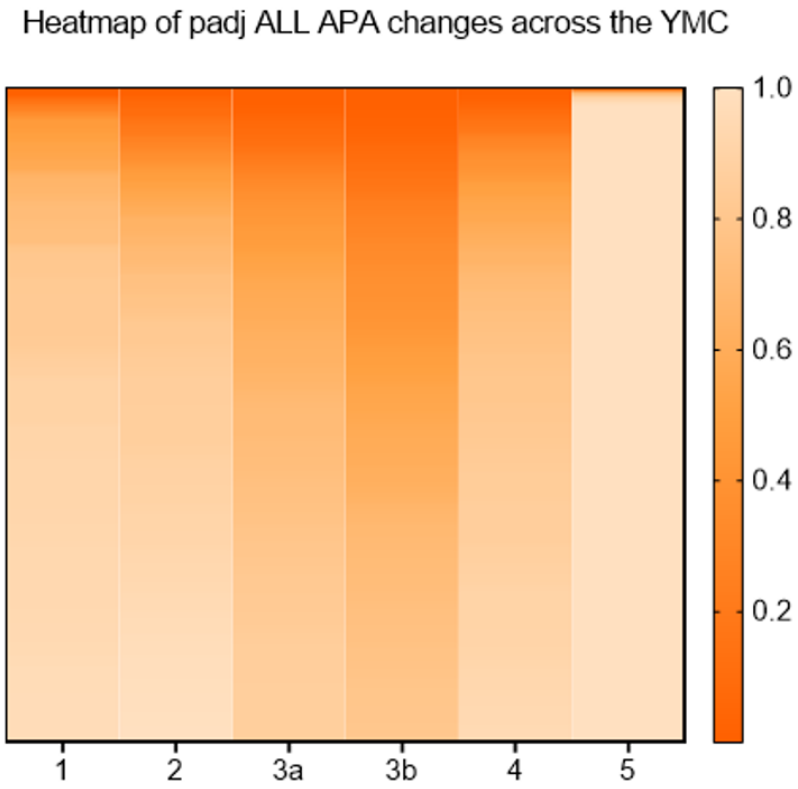


Figure 3.13: Heat map of logfold2 changes in normalized reads per million of genes that undergo ALL APA throughout the YMC, significant APA identified as genes with less than 0.05 padj values. Heat map shows peak of ALL APA changes occur at time points 3a and 3b. Only genes that undergo significant APA events were included in the heat map.

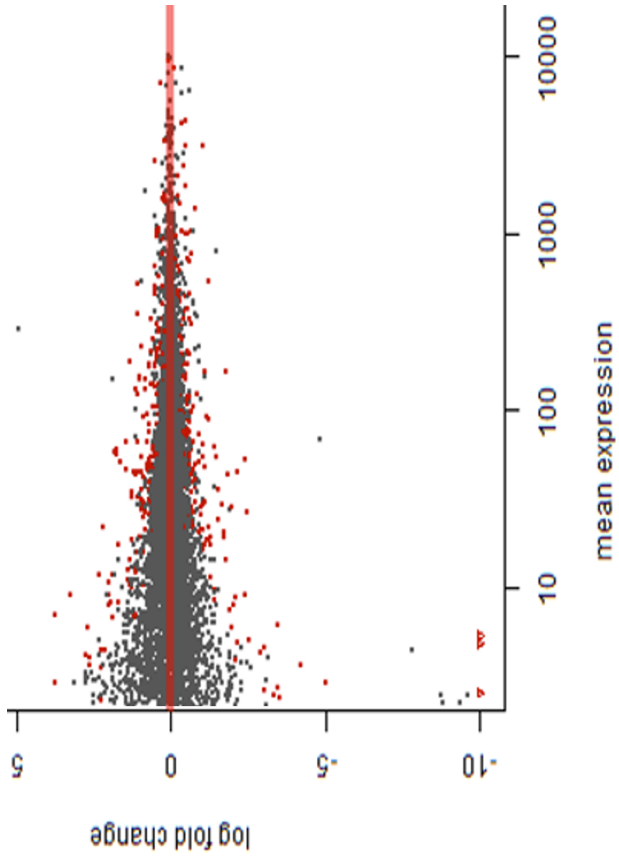


Figure 3.14: Scatter plot of gene expression vs APA events for 3b time point of the ALL data set. APA events are not skewed towards the most transcriptionally active genes as shown in this scatter plot of the 3b data set. The red dots indicate APA events that are significant with a padj value of less than 0.05, the grey dots are events that are not significant changes in APA.

The conclusion for these results is that the most APA events occur in the ALL data set during the YMC. The most prevalent time points for APA were 3a and 3b. The APA profile also shows a dominance for lengthening events at each time point. As with the other data sets the lack of skew towards the genes with larger expression changes implies that APA is not a random event that is an artifact of heavily transcribed genes, but a deliberate mechanism of gene expression control throughout the YMC.

3.5.2 Shifts in specific APA events correlate to changes in gene expression

In the previous sections, a number of genes that undergo 3'UTR, CR-APA or ALL APA events during the YMC were identified. A physiological consequence of APA can be the change in transcript stability, therefore APA can directly influence the gene expression of transcripts (Yuan et al. 2021). The data reveals a subtle preference for genes undergoing both APA events and differential Gene Expression (DE). This preference is observed across all categories of APA events. In most APA event categories, the percentage of genes with differential expression and APA events is higher than the percentage of genes without differential expression. This suggests a slight preference for genes to exhibit both APA and DE, indicating a potential association between these two processes. These trends in DE and APA are shown in the bar chart 3.15 on page 143. It's essential to note that while this preference is consistent, it is relatively small, with the percentages generally falling below 3% in each category. Nevertheless, these findings offer valuable insights into the relationship between APA events and DE in gene expression, hinting at a nuanced connection that may warrant further investigation. The summary table of the APA events for genes that undergo differential expression changes and those that do not undergo any expression changes is shown below on table 3.5 on page 144. In conclusion, the data suggests a slight preference for genes to exhibit both APA events and differential gene expression during the YMC. While this preference is consistent across all categories of APA events, the differences are small. However, these findings offer valuable insights into

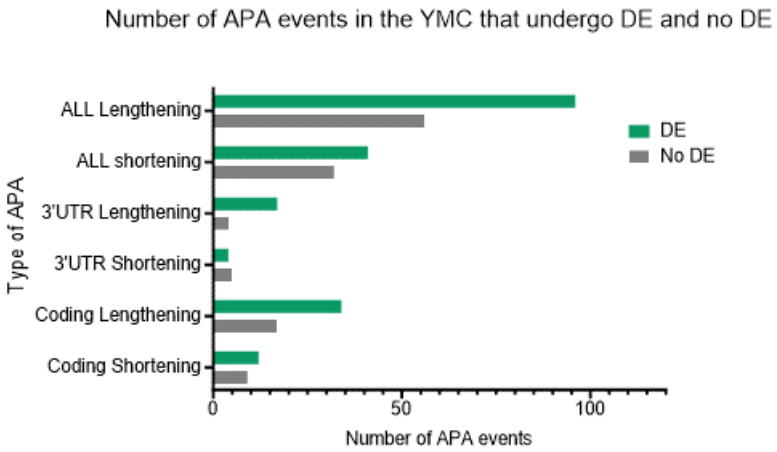


Figure 3.15: Figure shows a bar chart depicting the number of APA events in the YMC with differential expression. The genes that undergo differential expression are highlighted in green, while the genes without any differential expression are shown in grey.

the potential association between APA events and DE in gene expression, indicating a nuanced connection between these two processes.

There was only a small correlation between the prevalence of gene expression changes and APA events. However, the genes that do undergo both APA and gene expression changes showed a preference for either lengthening at genes that underwent increases in expression or shortening if gene expression decreases. For transcripts that undergo coding shortening APA, 9.52% of the transcripts also underwent gene expression increases and 47.6% undergo gene expression decreases. For transcripts that undergo coding lengthening APA, 46% of the transcripts also underwent gene expression increases and 19.2% undergo gene expression decreases. The APA and GE changes are summarised in the diagram 3.16 on page 146. For transcripts that undergo 3'UTR shortening APA, 0% of the transcripts also underwent gene expression

APA event	DE	% DE	No DE	% No DE
Coding Shortening	12	0.81	9	0.60
Coding Lengthening	34	2.29	17	1.14
3'UTR Shortening	4	0.20	5	0.25
3'UTR Lengthening	17	0.86	4	0.20
ALL shortening	41	0.84	32	0.65
ALL Lengthening	96	1.97	56	1.15

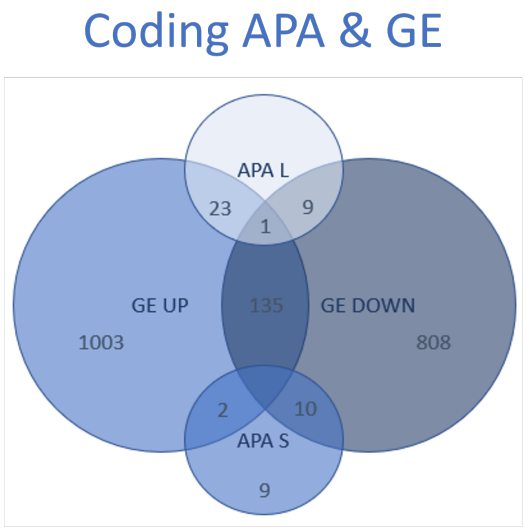
Table 3.5: Table showing a small preference of genes undergoing both APA and Differential gene expression events. % of APA events occur in genes that undergo differential expression and those that do not undergo differential expression. Table of the number of APA events that occur in both genes with differential expression changes throughout the YMC and genes that do not change expression throughout the YMC. Table shows DE = Differential expression, % DE is the % of genes that undergo APA events vs the total number of genes in that data set, total is the total number of APA events.

increases and 33.3% undergo gene expression decreases. For transcripts that undergo coding lengthening APA, 47.6% of the transcripts also underwent gene expression increases and 33.3% undergo gene expression decreases. The APA and GE changes are summarised in the diagram 3.17 on page 147.For transcripts that undergo ALL shortening APA, 13.7% of the transcripts also underwent gene expression increases and 42.5% undergo gene expression decreases. For transcripts that undergo coding lengthening APA, 40.8% of the transcripts also underwent gene expression increases and 22.3% undergo gene expression decreases. The APA and GE changes are summarised in the diagram 3.18 on page 148. T-Test between bootstrapped data sets showed some statistical correlations between APA changes and GE changes, indicating there might be a correlation between genes that undergo APA shifts in the YMC and genes that have significant changes in gene expression. The

table of results for these T-tests is shown on table 3.6 on page 145. This analysis of gene expression and APA showed that there was not an overall bias for genes that undergo APA to undergo gene expression changes or vice versa. However, of the genes that do undergo gene expression changes, there was a significant prevalence to undergo gene expression increases if the gene was also undergoing APA lengthening. Also the reverse was true that genes undergoing APA shortening tend to also have a decrease in gene expression. This is a significant finding as it identifies a clear trend linking APA events to gene expression changes. These results indicate that global APA lengthening events contribute to gene expression increases along the YMC. The tables for genes that undergo APA and gene expression changes is shown in the supplementary figures under Gene expression VS APA tables for YMC.

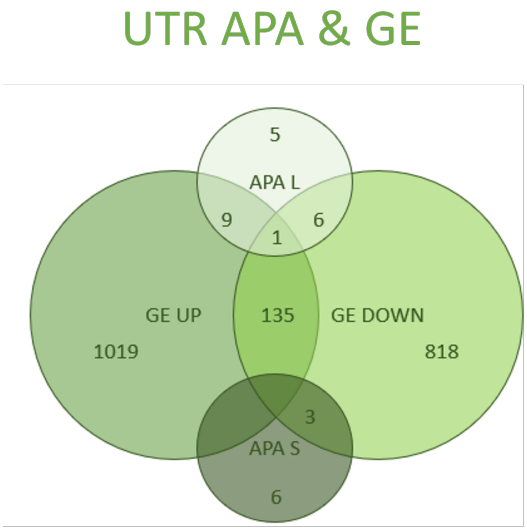
p-value at Time points	Coding S	Coding L	3UTR S	3UTR L	ALL S	ALL L
1	NA	NA	NA	0.521	0.36	0.373
2	NA	0.115	NA	0.889	0.4458	0.154
3a	0.125	0.428	0.145	0.921	0.0353	0.707
3b	0.046	0.087	NA	0.438	0.0193	0.147
4	0.085	0.114	NA	NA	0.378	0.215
5	NA	NA	NA	NA	NA	NA

Table 3.6: p-values for log fold changes in gene expression for genes that undergo APA changes vs genes that do not undergo any APA shifts. p-values highlighted in red have significance under 0.005.



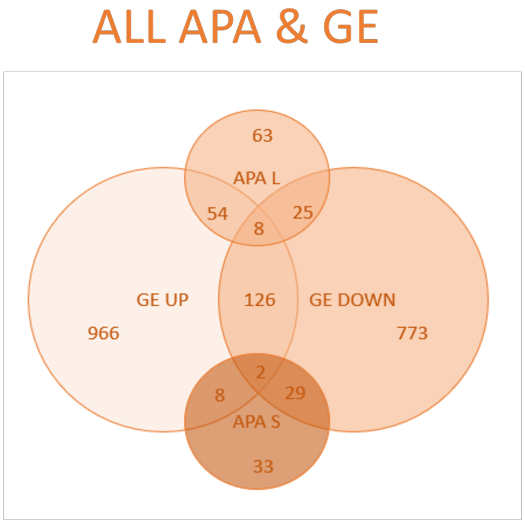
APA TYPE	Coding Up		Coding Down	
Shortening	2	9.52%	10	47.6%
Lengthening	24	46%	10	19.2%

Figure 3.16: Venn diagram of total Coding APA changes vs gene expression changes throughout the YMC and a table which shows the % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases



APA TYPE	UTR Up		UTR Down	
Shortening	0	0%	3	33.3%
Lengthening	10	47.6%	7	33.3%

Figure 3.17: Venn diagram of total 3'UTR APA changes vs gene expression changes throughout the YMC and a table which shows the % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases



APA TYPE	ALL Up		ALL Down	
Shortening	10	13.7%	31	42.5%
Lengthening	62	40.8%	34	22.3%

Figure 3.18: Venn diagram of total ALL APA changes vs gene expression changes throughout the YMC and a table which shows the % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases

3.6 Gene Ontology analysis shows several significant functional characteristics for genes that undergo APA changes in the YMC

Gene Ontology (GO) is a technique for interpreting the overall function of a group of genes by assigning genes to a set of predefined bins depending on their functional characteristics and seeing what terms are "enriched" in a data set. GO analysis shows genes undergoing APA in the YMC are enriched in GO terms for metabolism, cell cycle progression and transcription. GO enrichment was mostly captured for 3a/3b which is the peak of oxygen consumption and the transition into the reductive building early phase, in which DNA replication and cell division occur in about 10% of cells in the population (Feltham et al. 2020). When there were not enough APA events for statistically significant GO analysis, the entire data set was used instead of just the time points.

3.6.1 Gene Ontology analysis for genes that undergo Coding APA changes in the YMC

The GO analysis for coding shortening was significant at time points 3a, 3b and 4. This showed enrichment in homologous recombination, cell cycle and DNA repair. Coding APA lengthening showed GO enrichment at time points 2, 3a, 3b and 4. The enrichment functional categories were in translation elongation, nuclear division, fatty acid and glucose metabolism. The summary table for GO analysis for coding APA shortening and lengthening

data sets is shown below in 3.19 on page 151.

Coding Shortening	GO Term	P-value	Genes
1	N/A		
2	N/A		
3a	homologous recombination	0.00126	YLR320W
3b	metabolic process	0.000130	YNL247W, YBR084W
	mitotic cell cycle	0.001674	YLR353W
4	DNA repair	0.003351	YJL072C
	Nucleotide metabolic process	0.004421	YJL072C YOR359W
5	N/A		
Coding Lengthening			
1	N/A		
2	translational elongation	0.00371	YDR385W YLR249W YPL143W YPL220W
3a	nuclear division	0.000825	YBR158W YJL173C
	Translation	0.004339	YDR385W YJL050W YJL190C YLR172C YLR192C
3b	Translation	0.0000747	YPL220W YER102W YJL138C YLR441C YNL302C YOR293W
	fatty acid elongation	0.000295	YCR034W YLR372W
4	Glucose metabolism	0.0002315	YGL062W
5	N/A		

Figure 3.19: Summary table of GO enrichment analysis results for genes that undergo coding APA changes in the YMC. Significance in enrichment for a functional group was shown to be p-values less than 0.05.

3.6.2 Gene Ontology analysis for genes that undergo 3'UTR APA changes in the YMC

There was no notable enrichment in GO analysis for 3'UTR APA events in the YMC. However, the GO analysis for 3'UTR Lengthening revealed significant enrichment in glucose metabolism, cell cycle, and response to cellular stress. The summary table for the GO analysis of 3'UTR lengthening data sets can be found in the supplementary figures section, specifically 6.1 on page 283.

3.6.3 Gene Ontology analysis for genes that undergo ALL APA changes in the YMC

The GO analysis for ALL shortening was significant at time points 2, 3a, 3b and 4. The data sets showed enrichment in transport, cell cycle ad mRNA processing. ALL APA lengthening also showed GO enrichment at time points 2, 3a, 3b and 4. The enrichment functional categories were in cell cycle checkpoints, DNA damage and biosynthesis. The summary table for GO analysis for ALL APA shortening and lengthening data sets is shown below in 3.20 on page 153.

ALL Shortening	GO Term	P-value	Gene
1	transport	0.002750	YLR038C YOR254C
2	electron transport chain	0.004569	YHR001W-A YLR038C YNL169C
3a	ion transport	0.000197	YDL199C YJR077C YNR002C
	response to stress	0.04538	YBR214W YDL204W YDR258C YD R359C YML013W
3b	protein folding	0.000701	YBR169C YGR142W YJL117W YM L028W YNL209W
	mitotic cell cycle	0.009653	YBR214W YLR353W
4	mRNA poly(A) tail shortening	0.005467	YOR359W
5	N/A		
ALL Lengthening			
1	mitotic cell cycle	0.0002902	YJR072C
	Translation	0.0018542	YPL143W YPL220W
2	ribosomal ls biogenesis	0.000449	YDR083W YOR234C YPL043W YP L143W YPL220W
	cell cycle checkpoint	0.00886	YBR158W
3a	cell cycle checkpoint	0.00303	YBR158W
	mitotic cell cycle	0.000445	YBR158W YER095W YJR072C
3b	lysine metabolic process	0.00651	YBR115C YIL094C YNR050C
	cell cycle checkpoint	0.000911	YBR158W YER173W
	DNA damage checkpoint signalling	0.02964	YER173W
4	AMP biosynthetic process	0.00199	YLR359W YML022W
5	N/A		

Figure 3.20: Summary table of GO analysis results for genes that undergo ALL APA changes in the YMC. Significance in enrichment for a functional group was shown to be p-values less than 0.05.

3.7 Cluster analysis of genes that undergo APA changes in the YMC show have few unique gene expression clusters

Identifying trends in genes that undergo APA can be used to group certain genes into subcategories as to further granulate understanding of APA profiles. Furthermore, clustering genes that undergo APA by the shape of their gene expression profile might potentially reveal mechanisms of gene expression control that would not be observable when observing the full data set. One such statistical method to categorise gene groups via their gene expression profiles is cluster tool analysis. The log fold changes in gene expression for each gene undergoing APA underwent K means clustering via the clustering tool clustergrammer by the Ma'ayan Lab, the Icahn School of Medicine at Mount Sinai. The clusters were calculated based on the similarities the genes had in changes in gene expression across the YMC. The output of these clustering analysis was a heat map of the different clusters visualised to show changes in gene expression. Spearman's correlation was also calculated to show the similarity of each gene compared to each other and a matrix of these values was produced. The clusters that were produced were then put through GO analysis to identify any functional similarities between the clustered genes. The results of this analysis showed that most APA events fell into two distinct clusters, apart from all APA shortening, which had three separate clusters. The GO analysis showed that the clusters contained gene groups that were very distinct from each cluster. The func-

tional properties of many of these genes in APA clusters are important in the YMC including, regulation of carbohydrate metabolism, aerobic respiration, regulation of transcription and DNA repair. The overall trends in the clusters showed that APA shortening events correlated to gene expression decreases, whilst APA lengthening had the opposite effect and trended with gene expression increases. This trend is shown in the summary table of all APA clusters throughout the YMC, shown on 3.21 on page 156.The clusters each had a greater range gene expression changes than the mean total, indicating these data sets deviate from the normal range of gene expression. The individual genes in all of the different APA clusters are shown in figures 3.22 and 3.23 on pages 157 and 158.

The cluster analysis showed that genes that undergo APA events mostly show only two distinct clusters in gene expression changes throughout the YMC. These clusters show a correlation between gene expression decreases and APA shortening and APA lengthening and APA increases.

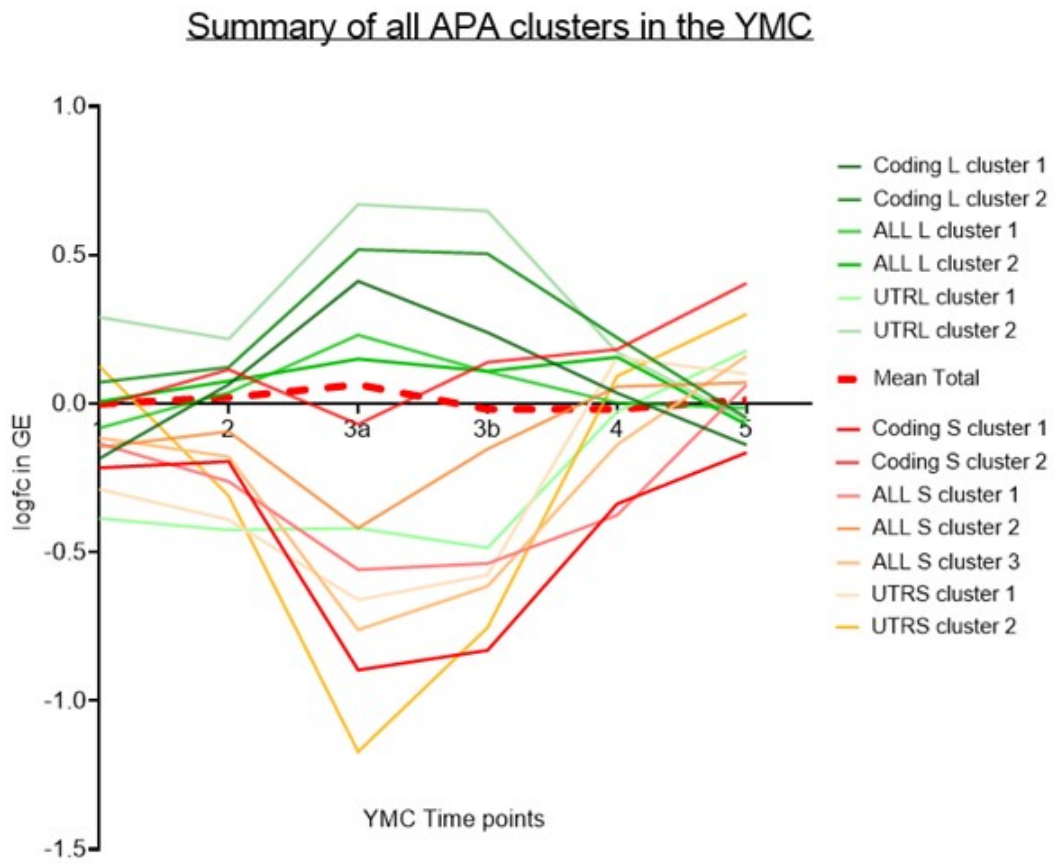


Figure 3.21: Summary of log fold changes in gene expression of all genes that undergo APA changes in the YMC. APA shortening events are coloured in red shades and lengthening are in green shades. The summary means of all log fold changes in gene expression are shown with a dashed red line

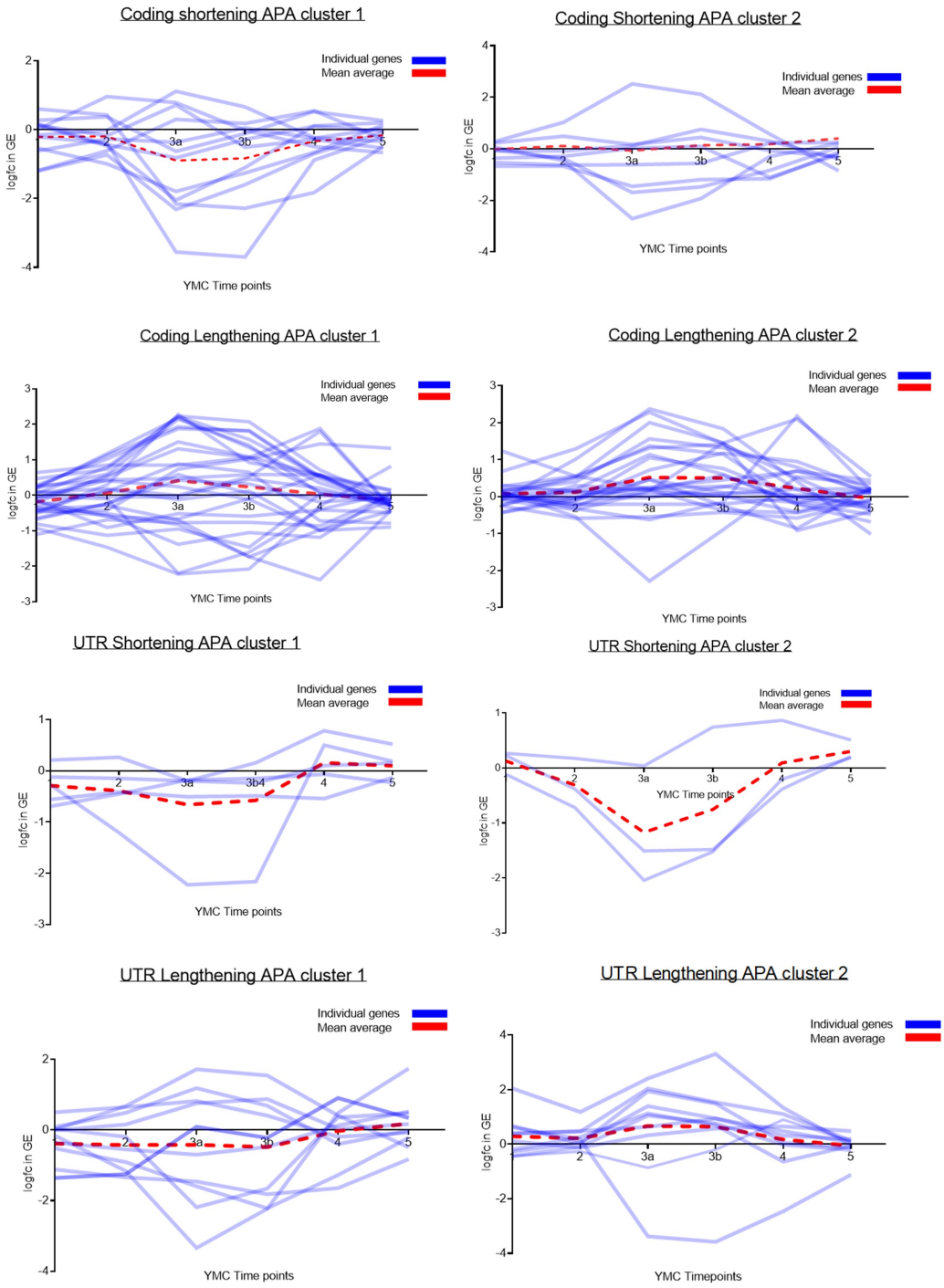

Coding shortening APA cluster 1
Individual genes
Mean average
logfc in GE
YMC Time points
Coding Shortening APA cluster 2
Individual genes
Mean average
logfc in GE
YMC Time points
Coding Lengthening APA cluster 1
Individual genes
Mean average
logfc in GE
YMC Time points
Coding Lengthening APA cluster 2
Individual genes
Mean average
logfc in GE
YMC Time points
UTR Shortening APA cluster 1
Individual genes
Mean average
logfc in GE
YMC Time points
UTR Shortening APA cluster 2
Individual genes
Mean average
logfc in GE
YMC Time points
UTR Lengthening APA cluster 1
Individual genes
Mean average
logfc in GE
YMC Time points
UTR Lengthening APA cluster 2
Individual genes
Mean average
logfc in GE
YMC Timepoints

Figure 3.22

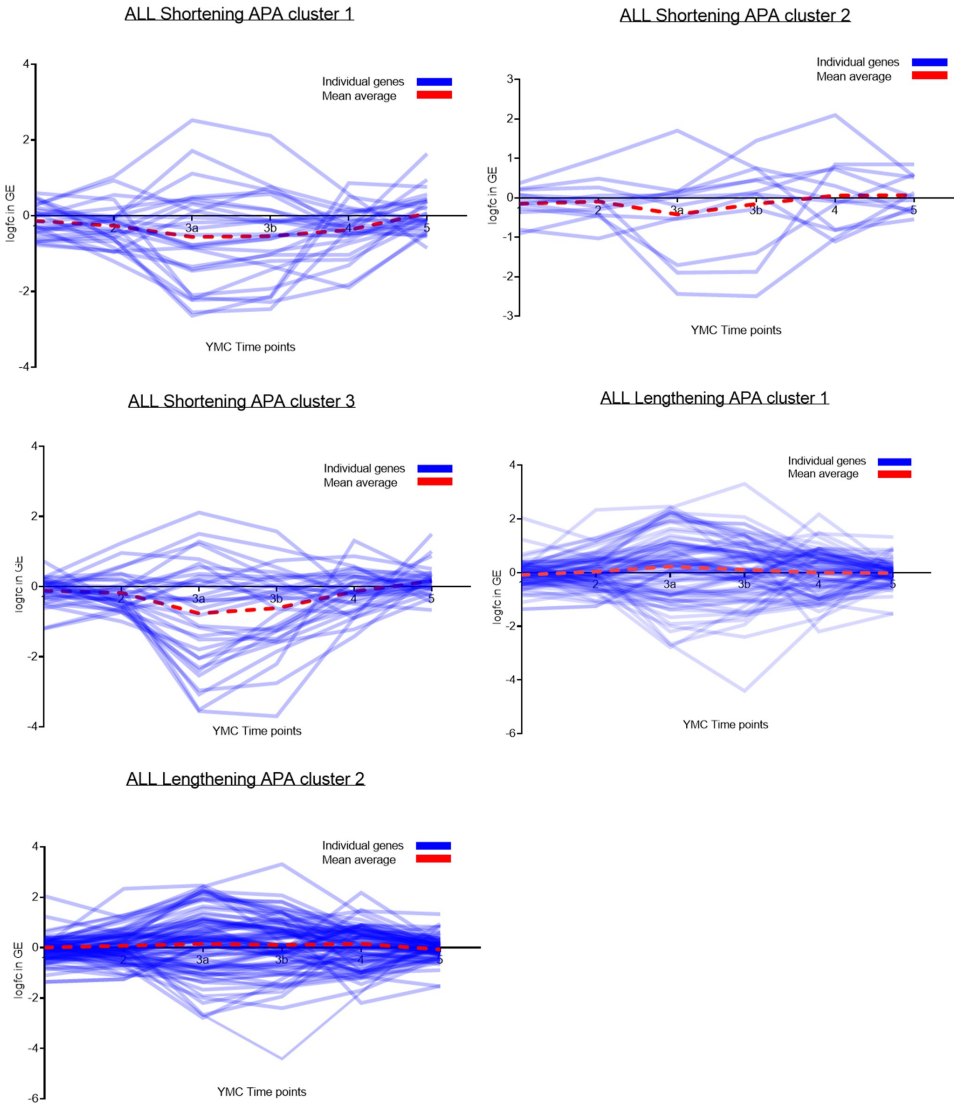


Figure 3.23: Line graphs showing the cluster data for log fold changes in gene expression along the YMC of for 3'UTR, Coding, and ALL APA data sets. The individual genes in the clusters are shown in blue and the log fold mean of the clusters is shown with a dashed red line. Each blue line as a gene that undergoes APA and has been categorised as being part of the cluster that is in the title of each graph.

Cluster Analysis of Coding Shortening APA shows two distinct clusters of GO enriched terms

For genes that undergo coding APA shortening, two gene expression clusters were identified. The hierarchically clustered heat map of these two clusters is shown below in 3.24 on page 160. Cluster 1 is composed of 12 genes. The overall trend in gene expression of this cluster is that their expression levels start to decrease as the YMC progresses with the peak at 3a, which is the end of the OX phase. The GO analysis for these genes in this cluster were involved in ion transmembrane transport. Cluster 2 contains 8 genes, whose expression steadily increases until the end of the cycle. The GO analysis of cluster two showed enrichment in heme, folic acid, and glucan metabolic processing. The GO analysis summary tables of the APA coding shortening clusters is shown on 6.3 on page 288 in the supplementary figures. The Spearman's correlation matrix of log fold change values of gene expression changes of genes that undergo coding APA shortening throughout the YMC is also shown in the heat map matrix 6.9 in the appendix on page 292.

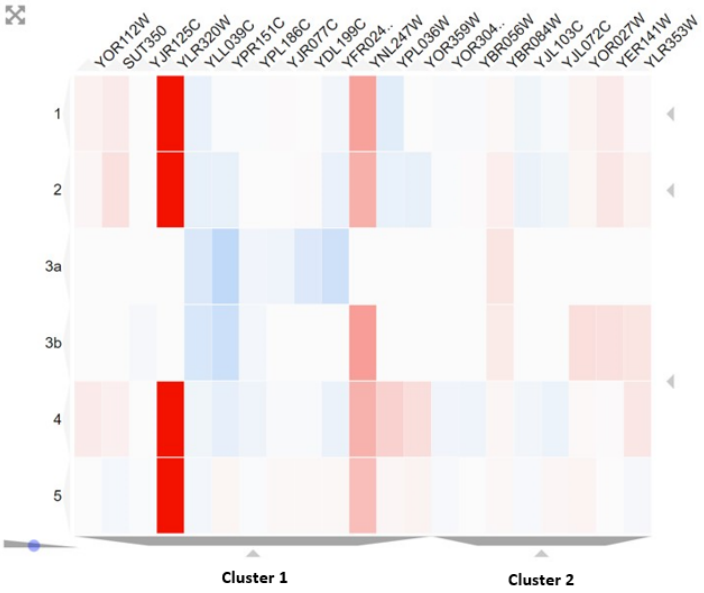


Figure 3.24: clustergram heatmap visualization of genes that undergo coding shortening APA changes in the YMC. Red/blue cells in the heatmap matrix represent positive/negative gene expression values.

Cluster Analysis of Coding Lengthening APA

For genes that undergo coding APA lengthening, two gene expression clusters were identified. The hierarchically clustered heat map of these two clusters is shown below in 3.25 on page 162. Cluster 1 is composed of 28 genes. The overall trend in gene expression of this cluster is that their expression levels slowly increase as the YMC progresses with the peak at 3a, which is the end of the OX phase and then the levels decrease to baseline again at time point 5, the end of the reductive charging phase. The GO analysis for these genes in this cluster was involved in translation. Cluster 2 contains 24 genes and, which follows a similar expression profile to cluster 1 but the peak is at 3b, the reductive building phase. The GO analysis of cluster two showed enrichment in transcription by RNA polymerase II, mitochondrial genome maintenance and regulation of TOR signaling. The GO analysis summary tables of the APA coding lengthening clusters is shown on 6.4 on page 289 in the supplementary figures. The log fold changes of individual genes in the APA coding lengthening clusters throughout the can be seen in the line graph in x. The Spearman's correlation matrix of logfold change values of gene expression changes that undergo coding APA lengthening throughout the YMC is also shown in the heat map matrix 6.10 in the appendix on page 293.

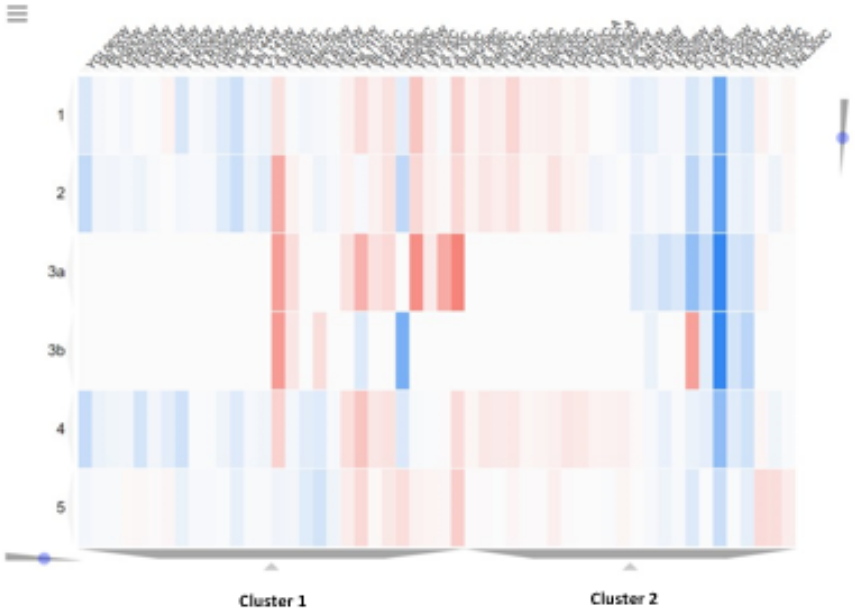


Figure 3.25: clustergram heatmap visualization of genes that undergo coding lengthening APA changes in the YMC. Red/blue cells in the heatmap matrix represent positive/negative gene expression values.

Cluster Analysis of 3'UTR Shortening APA

For genes that undergo 3'UTR APA shortening, two gene expression clusters were identified. The hierarchically clustered heat map of these two clusters is shown below in 3.26 on page 164. Cluster 1 is composed of only 5 genes. The overall trend in gene expression of this cluster is that their expression levels slowly decrease as the YMC progresses with the peak at 3a, which is the end of the OX phase and then the levels increase to baseline again at time point 5, the end of the reductive charging phase. The GO analysis for these genes in this cluster were involved in protein processing. Cluster 2 contains only 3 genes and, which follows a similar expression profile to cluster 1 but the peak is at 3b and the log fold decrease in gene expression is much larger. The GO analysis of cluster two showed enrichment in nucleotide excision repair. The GO analysis summary tables of the APA 3'UTR shortening clusters is shown on 6.6 on page 290 in the supplementary figures. Spearman's correlation matrix of logfold change values of gene expression changes of genes that undergo 3'UTR shortening throughout the YMC is shown in the heat map matrix 6.11 in the appendix on page 294.

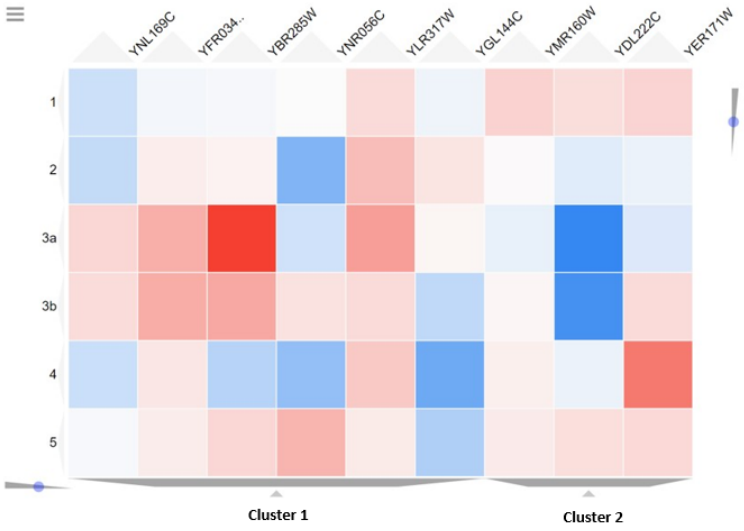


Figure 3.26: clustergram heatmap visualization of genes that undergo 3'UTR shortening APA changes in the YMC. Red/blue cells in the heatmap matrix represent positive/negative gene expression values.

Cluster Analysis of 3'UTR Lengthening APA

For genes that undergo 3'UTR APA lengthening, two gene expression clusters were identified. The hierarchically clustered heat map of these two clusters is shown below in 3.27 on page 166. Cluster 1 is composed of 11 genes. The overall trend in gene expression of this cluster is that their expression levels are low throughout the YMC and have a peak at time point 5, the end of the reductive charging phase. The GO analysis for these genes in this cluster were involved in apoptosis. Cluster 2 contains 10 genes which show a small spike in gene expression increase at 3a and 3b. The GO analysis of cluster two showed enrichment regulation in MAP kinase activity and protein adenylyation. The GO analysis summary tables of the APA 3'UTR lengthening clusters is shown on 6.5 on page 289 in the supplementary figures. Spearman's correlation matrix of logfold change values of gene expression changes of genes that undergo 3'UTR lengthening throughout the YMC is shown in the heatmap matrix 6.12 in the appendix on page 295.

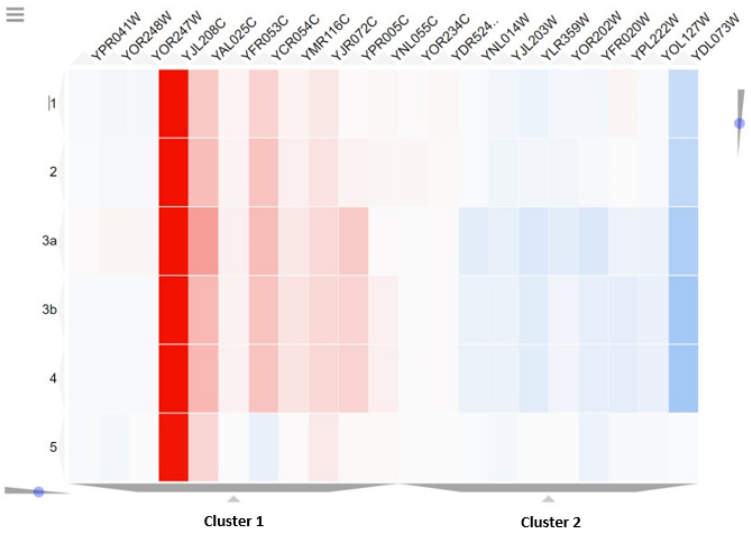


Figure 3.27: clustergram heatmap visualization of genes that undergo 3'UTR lengthening APA changes in the YMC. Red/blue cells in the heatmap matrix represent positive/negative gene expression values.

Cluster Analysis of ALL Shortening APA

For genes that undergo ALL APA shortening, three gene expression clusters were identified. The hierarchically clustered heat map of these three clusters is shown below in 3.28 on page 168. Cluster 1 is composed of 28 genes. The overall trend in gene expression of this cluster is that their expression levels decrease throughout the YMC and the peak is at the end of the OX phase in time point 3a. The GO analysis for these genes in this cluster were involved in regulation of cellular carbohydrate metabolic process, electron transport and cellular response to glucose starvation. Cluster 2 contains 12 genes and shows a small decrease in gene expression at 3a and 3b. The GO analysis of cluster two showed enrichment in aerobic respiration and sulphur metabolism. Cluster 3 is composed of 33 genes and also shows a steady decrease in gene expression that peaks at 3a, the GO analysis showed enrichment in DNA repair. The GO analysis summary tables of the ALL APA lengthening clusters is shown on 3.29 on page 168. The Spearman's correlation matrix of logfold change values of gene expression changes of genes that undergo ALL APA shortening throughout the YMC is also shown in the heat map matrix 6.13 in the appendix on page 296.

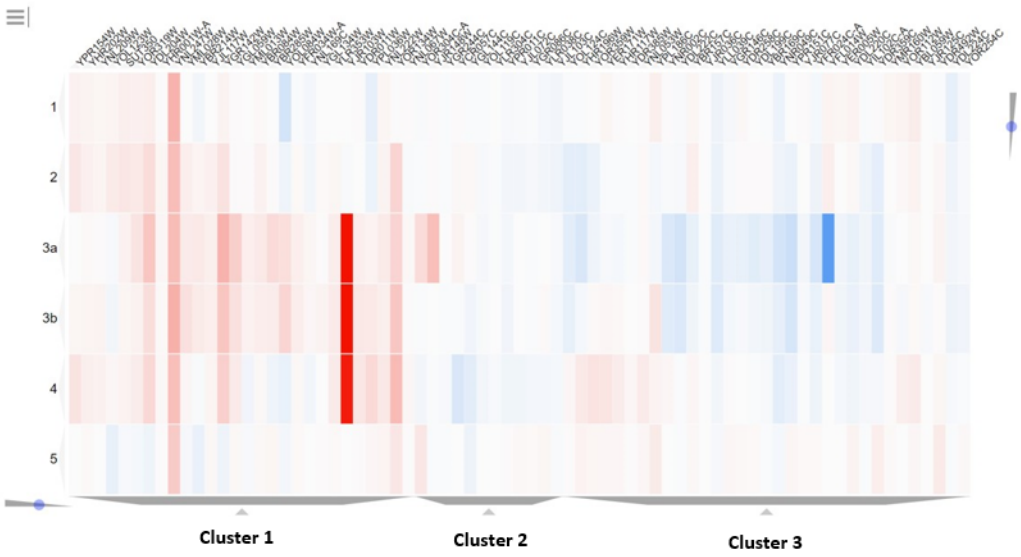


Figure 3.28: clustergram heatmap visualization of genes that undergo ALL shortening APA changes in the YMC. Red/blue cells in the heatmap matrix represent positive/negative gene expression values.

APA ALL Shortening Cluster GO analysis		
Cluster 1		
GO Term	p-Value	genes
regulation of cellular carbohydrate metabolic process [GO:0010675]	6.05E-04	YGL059W YGL134W YML028W
mitochondrial electron transport, ubiquinol to cytochrome c [GO:0006122]	0.001175	YHR001W-A YNL169C
cellular response to glucose starvation [GO:0042149]	0.001385	YBR214W YNL209W
Cluster 2		
GO Term	p-Value	genes
aerobic respiration [GO:0009060]	9.27E-04	YGR244C YLR038C YLR304C
sulfur compound metabolic process [GO:0006790]	0.037843	YGR244C YJR148W
Cluster3		
GO Term	p-Value	genes
nucleotide-excision repair, DNA duplex unwinding [GO:0000717]	0.005237	YER171W

Figure 3.29: GO analysis of the genes in the ALL APA Shortening clusters. Tables show GO terms for genes that have enrichment in GO functional groups with p-values less than 0.05.

Cluster Analysis of ALL Lengthening APA

For genes that undergo ALL APA lengthening, two gene expression clusters were identified. The hierarchically clustered heat map of these two clusters is shown below in 3.30 on page 170. Cluster 1 is composed of 88 genes. The overall trend in gene expression of this a small peak in gene expression at timepoint 3a at the end of the OX phase. The GO analysis for these genes in this cluster were involved in monocarboxylic acid biosynthetic processing, fatty acid elongation and regulation of intracellular signal transduction. Cluster 2 contains 61 genes and shows a very small spike in gene expression increase at 3a and 3b and decrease at 5, the end of the YMC cycle. The GO analysis of cluster two showed enrichment regulation in aspartate family amino acid biosynthetic process and lysine biosynthetic process via aminoadipic acid. The GO analysis summary tables of the ALL APA lengthening clusters are shown on 3.31 on page 170. The log fold changes of individual genes in the ALL APA lengthening clusters throughout the YMC can be seen in the line graph in 6.14 on page 297. Spearman's correlation matrix of log fold change values of gene expression changes of genes that undergo ALL lengthening throughout the YMC is shown in the heatmap matrix 6.14 in the appendix on page 297.

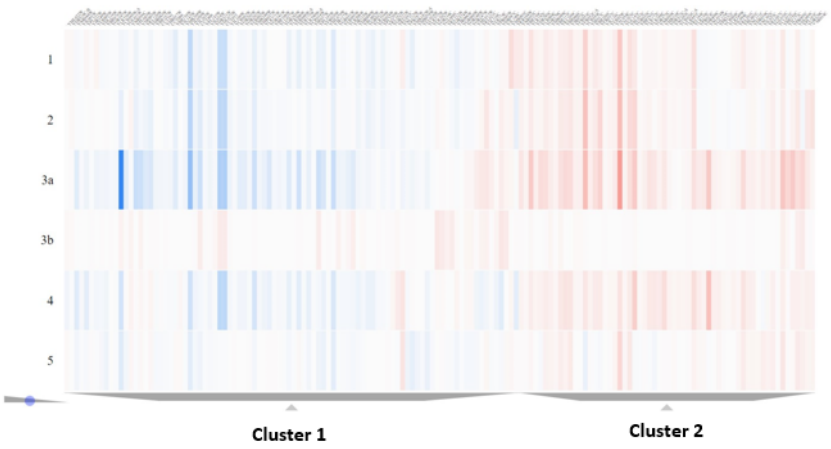


Figure 3.30: clustergram heatmap visualization of genes that undergo ALL lengthening APA changes in the YMC. Red/blue cells in the heatmap matrix represent positive/negative gene expression values.

APA ALL Lengthening Cluster GO analysis		
Cluster 1		
GO Term	p-Value	genes
monocarboxylic acid biosynthetic process [GO:0072330]	1.46E-04	YCR034W YKL182W YLL027W YLR372W YNR056C YOR374W
fatty acid elongation [GO:0019367]	5.44E-04	YCR034W YLR372W
positive regulation of intracellular signal transduction [GO:1902533]	7.57E-04	YDL073W YGL008C YLR208W YMR001C
Cluster 2		
GO Term	p-value	genes
aspartate family amino acid biosynthetic process [GO:0009067]	1.25E-04	YBR115C YCR054C YIL094C YIR017C YNR050C
lysine biosynthetic process via aminoadipic acid [GO:0019878]	2.24E-04	YBR115C YIL094C YNR050C

Figure 3.31: GO analysis of the genes in the ALL APA Lengthening clusters.Tables show GO terms for genes that have enrichment in GO functional groups with p-values less than 0.05.

3.8 YMC gene expression clusters overlap with YMC APA data sets in GO enrichment at each super cluster

After the Gene Ontology and cluster analysis of the YMC APA data sets had been investigated, the APA genes were compared to YMC gene expression clusters. It was important to identify if the characteristics of the YMC functions being represented in the gene expression clusters were enriched with changes in APA. The gene expression clusters for the YMC was investigated by Shidong Xi, in the Mellor laboratory, Oxford university Biochemistry department. The temporal expression profiles of different genes throughout the YMC exhibit diverse shapes. The 3 super clusters was automatically separated into 16 sub-clusters using an unsupervised k-means algorithm using a similar method to cluster the APA data sets. These data sets were cross compared to the APA data sets and it was found that the data sets shared many common genes and also GO terms at similar points in the YMC. The most overlap for genes that undergo APA and are included in clusters was seen cluster 10, 14, and 16, which all contained more than 10% of APA genes listed. The tables summarising of APA transcripts in clusters are shown in the appendix in table 6.3 on page 286 and 6.2 on page 285. The super clusters also showed similar GO terms to the GO terms shown in the APA data sets which included cell cycle progression, ribosome synthesis and stress response.

3.8.1 GO analysis of genes that are present in the gene expression clusters and APA data sets are enriched for terms that include fatty acid metabolism, DNA repair and translation

The genes in the clusters that had overlap with APA events underwent GO analysis and several data sets show GO enrichment for functional terms that include fatty acid metabolism, DNA repair and translation. A summary of all the enriched cluster/APA data sets is shown in table 3.32 on page 3.32.

Coding Lengthening cluster 10	Genes
translation	YLR208W YOR038C
Coding Lengthening cluster 14	
fatty acid elongation	YCR034W YLR372W
Coding Lengthening cluster 16	
translation	YJL138C YJL190C YNL302C YPL220W
ALL Shortening cluster 10	
electron transport chain	YHR001W-A YNL169C
ALL Shortening cluster 15	
vesicle transport	YDL222C YGR086C YGR142W
ALL Lengthening cluster 1	
rRNA processing	YAL025C YDR083W YDR087C YJL050W YNL112W
ion transmembrane transport	YBR187W YBR291C YGL008C YJL116C YKL120W
ALL Lengthening cluster 2	
glucose import	YFR053C
ALL Lengthening cluster 4	
nucleotide-excision repair	YER095W YJL173C
ALL Lengthening cluster 10	
protein deacetylation	YDR295C YER174C
ALL Lengthening cluster 14	
fatty acid elongation	YCR034W YLR372W
ALL Lengthening cluster 16	
translation	YJL138C YJL190C YNL302C YPL220W

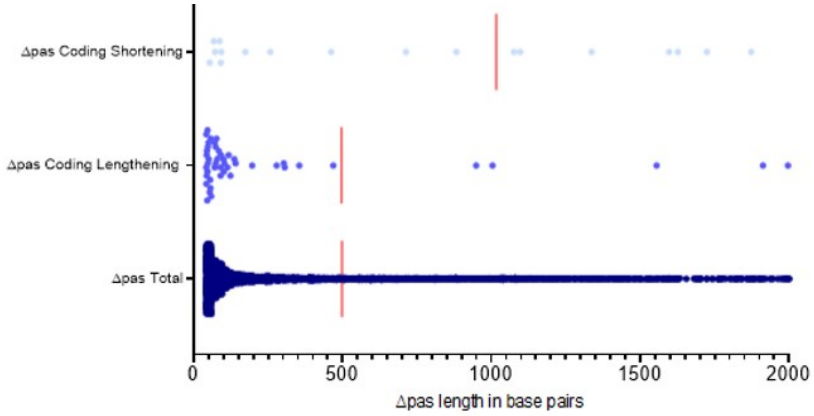
Figure 3.32: GO analysis of genes that undergo APA changes and are included in gene expression clusters. Clusters with enriched terms are listed.

3.8.2 Δpas lengths significantly differ in genes with APA changes in the YMC, implying targeted pas site selection shifts in the YMC APA profile

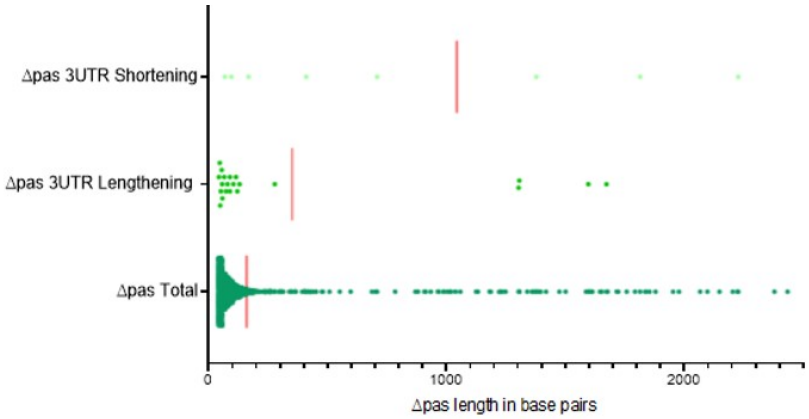
Δpas length is the length in base pairs between two alternative pas sites and therefore the stretch of sequence that is included or removed from a nascent mRNA transcript if APA occurs. For all types of APA apart from coding lengthening, genes that underwent APA all had significantly longer Δpas lengths then genes that had multiple pas sites but did not undergo APA. This suggests that the APA events are not random or due to error as the longer Δpas distances seem to be preferable and selected over the shorter, instead of a random selection of lengths being removed or included due to APA shifts. The distribution of Δpas lengths for all the APA events can be seen in 3.33 on page 175.

3.9 NET-Seq and RNA-Seq data for YMC transcripts show similar gene expression profiles.

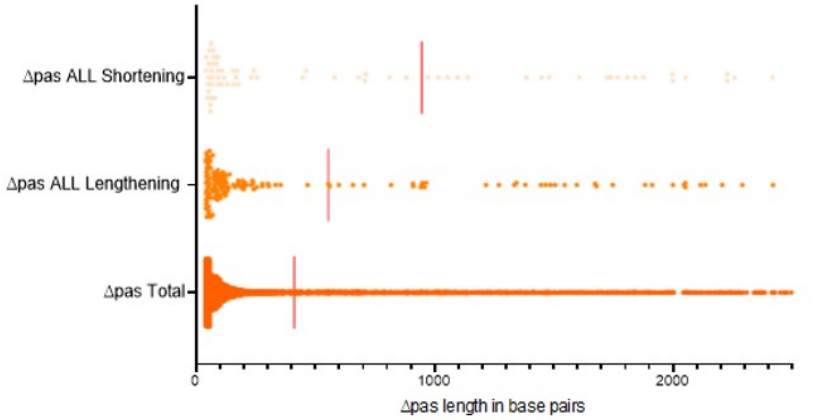
Net-seq data sets for the YMC was shared by the Mellor lab. NET-Seq detects nascent, actively transcribed RNAPII RNAs, through the capture of 3' RNA. The RNAPII elongation complex is immunoprecipitated, and RNA is extracted and reverse-transcribed to cDNA and then sequenced to output the gene expression level of nascent transcribed transcripts. This provides us with information on what genes are being actively transcribed



(a) Δpas lengths for coding APA events.



(b) Δpas lengths for 3'UTR APA events.



(c) Δpas lengths for ALL APA events.

Figure 3.33: Scatter plot of Coding, 3'UTR and ALL APA Δpas lengths. Mean length in base pairs between alternative pas sites is shown with a red line.

at different time points in the YMC. This data is valuable in the investigations into the YMC as it can show us if the APA changes are just due to degradation levels of APA transcripts or if these APA transcripts are being actively more transcribed. Identifying whether these transcripts are showing increased transcription will confirm if APA events in the YMC are due to targeted degradation or targeted increases in transcription. The Net-Seq profiles show the same overall trends in gene expression compared to the RNA-Seq data sets. The peak of both gene expression changes was at the transition between the R/B late phase and R/C early phase. The least amount of transcriptional activity was at the end of the R/C phase/ the transition time between two cycles for both the RNA-Seq and Net-Seq time points.

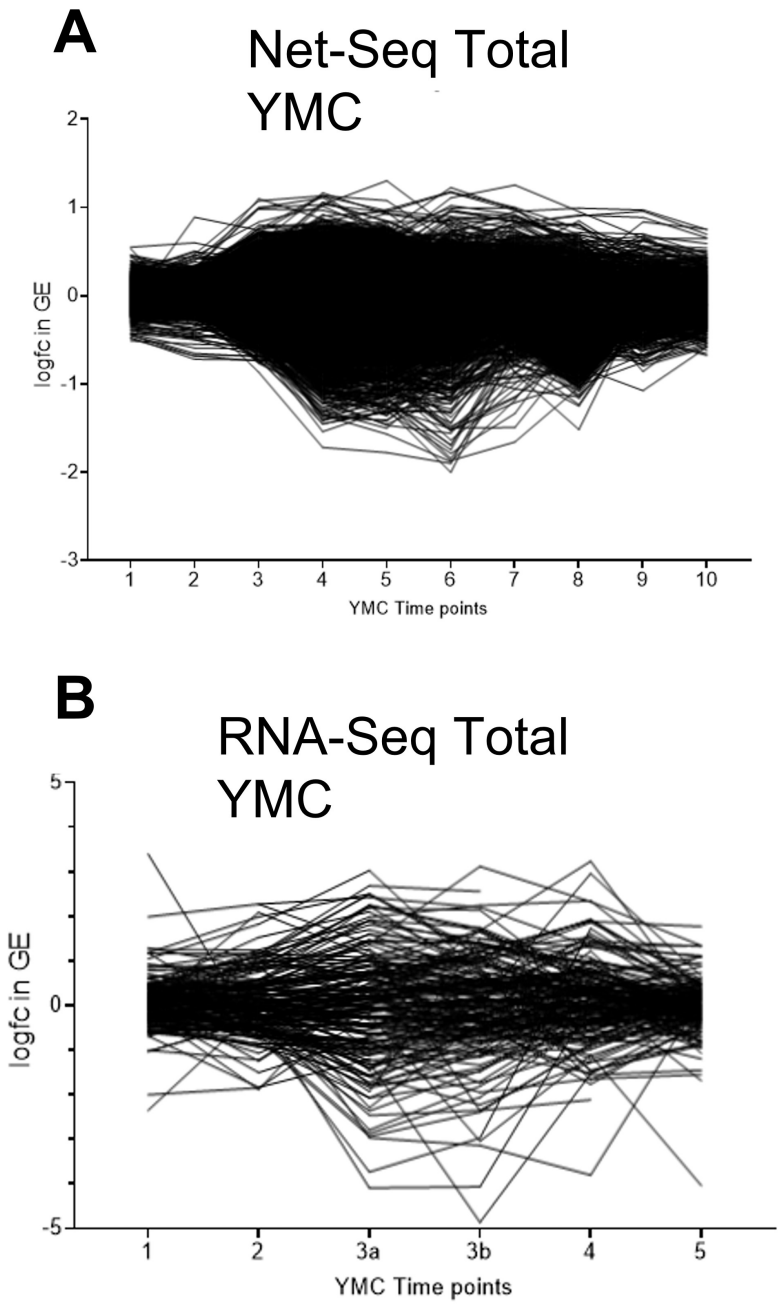
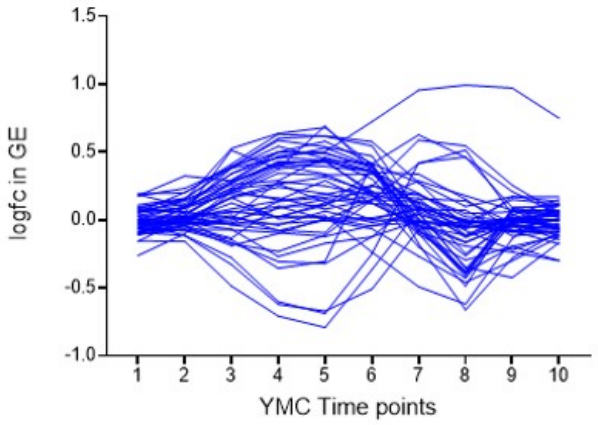


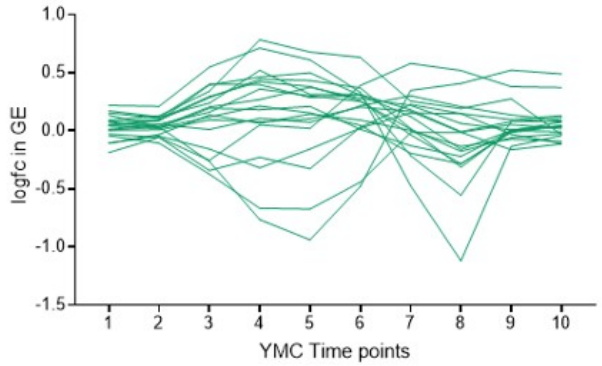
Figure 3.34: NET-Seq and RNA-Seq gene expression profiles for YMC data sets. A is a line graph of the the Net-seq data log fold change in gene expression of all genes captured in the YMC. B is a line graph of the the RNA-seq data log fold change in gene expression of all genes captured in the YMC.

3.9.1 Net-Seq profiles for genes that undergo APA changes follow similar trends to the RNA-Seq data sets

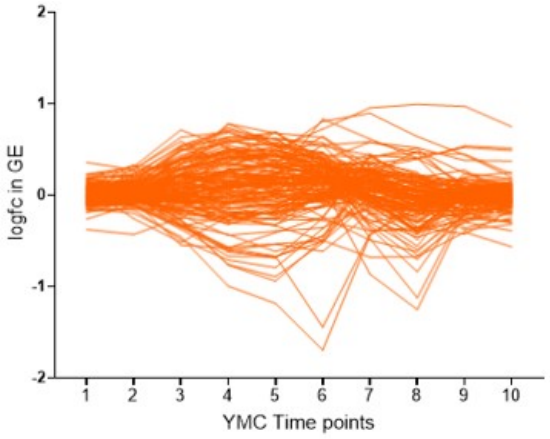
The Net-Seq reads for the genes that underwent Coding, 3'UTR and ALL were identified from the Net-seq data sets from the Mellor lab. The genes that underwent APA changes in all of the data sets showed comparable trends in gene expression shifts across the YMC in both the Net-Seq and RNA-Seq data. The amplitude of the Net-Seq reads for APA genes peak at the middle of the YMC which is comparable to time points 3a and 3b of the RNA-Seq data. The similarities in shapes of the expression profiles for all the APA types can be seen in the line graphs 3.35 on page 179. The heat maps of the logfold2 changes in gene expression for the Net-Seq data for coding, 3'UTR and ALL APA genes also demonstrate the peaks in gene expression changes occur at the end of the R/B beginning of R/C are shown in the appendix in figure 6.2 on page 287 . This comparison between the NET-Seq and RNA-Seq data sets confirms that APA events in the YMC are not just due to increased transcript stability but due to active changes in transcription of mRNAs with alternative pas site usage. This indicates that active APA is occurring that changes the pas site usage of transcripts.



(a) Net-Seq gene expression profiles for genes that undergo coding APA throughout the YMC



(b) Net-Seq gene expression profiles for genes that undergo 3'UTR APA throughout the YMC



(c) Net-Seq gene expression profiles for genes that undergo ALL APA throughout the YMC

Figure 3.35: Line graphs of coding, 3'UTR and ALL APA logfold change in gene expression along the 10 time points in the YMC Net-Seq data sets.

3.10 TIF-Seq IGV tracks unveil several mechanistic differences between changes in pas site usage

Genes that undergo APA shifts throughout the YMC were visualised in IGV. IGV tracks show the counts for each pas site usage in a gene for all the time points, giving a full overview of how the number of reads of a transcript per pas site (i.e. distal or pas site usage) changes along the YMC. These visualised data sets come from the original BAM files that were imputed into the adapted Furger APA pipeline. The data was converted from BAM files to BigWigs, which allows the count reads to be displayed as continuous data in a genome browser track, such as IGV. The IGV reads are extremely important information for investigating the YMC APA profile as they provides essential context as to how these APA shifts are occurring. The RNA-Seq pipeline outputs just show that an event has occurred and the padj significance of that event. Using just this data, it is unknown whether an APA "lengthening" event is due to increased distal pas site usage or a collapse in proximal pas site usage, i.e. changes in ratio of short to long reads. Knowing if shortening or lengthening event is due to increase or decrease usage of a proximal or distal pas site gives essential information on the mechanisms driving that APA shift. The IGV tracks for the APA events showed that APA lengthening and shortening events occur via a mix of varying count reads at both the proximal and distal PAS. This was observable only by looking at genes individually in IGV. I therefore took a selection of genes that undergo

APA that are also physiologically relevant to the YMC to get a more granular look at how these genes undergo shortening and lengthening events. Furthermore, the IGV tracks identified several potential other mechanisms of gene expression control that might not be strictly APA events, such as transcriptional interference with overlapping genes. Therefore, the discovery of these expression patterns in IGV that were identified by the pipeline as APA, might in fact be unknown novel interference mechanisms. These discoveries show how powerful the combinatorial bioinformatics approach is to exploring mechanisms of gene expression control in complex biological systems such as the YMC. These shifts in pas site usage where also confirmed in IGV using NET-Seq and TIF-Seq data from the Mellor Laboratory. The TIF seq data set was produced from RNA extracted from exponentially growing cultures in glucose at mid log show, so is a good baseline control for APA in *saccharomyces cerevisiae*. TIF-seq (Transcript isoform sequencing) identifies isoforms of transcripts by sequencing full-length mRNA molecules with both the 5-cap and poly(A) tails to show the transcription start sites (TSSs) through to the PAS at the 3-end of the transcript (Pelechano, W. Wei, and Steinmetz 2013). This sequencing tool can therefore be used to independently confirm the APA events identified by the pipeline. Using these three different data sets in IGV identified potentially non traditional APA events occurring due to transcriptional interference and changes in TSS.

3.10.1 *MIS1/YBR084W* undergoes 3'UTR shortening APA at time points 3a and 3b confirmed by TIF-Seq data

The RNA-Seq data set identified the gene *MIS1/YBR084W* as undergoing an APA shortening event at time points 3a and 3b in the YMC. The *MIS1* gene is responsible for encoding the enzyme Mitochondrial C1-tetrahydrofolate synthase which is involved in folic acid biosynthesis and plays a vital role in the regulation of the cell cycle (Shannon and Rabinowitz 1988). The IGV tracks for all time points for the gene *YBR084W* show this 3'UTR APA shortening event is shown in on page 183. In the IGV tracks you can see how there is a large increase in count reads for both pas sites, which is due to the gene expression increase shown at the 3a and 3b time point for this gene. The proximal pas site increases in count reads much more compared to the distal pas site, therefore causing more transcripts with a truncated 3'UTR to be produced, i.e. APA shortening. These shifts in pas site count reads along the YMC is shown in the graph 3.37 on page 184.

NET-Seq and TIF-Seq data confirm 3'UTR APA events in *MIS1/YBR084W*

Transcription of *MIS1* peaks in the OX phase of the YMC, which is the peak of oxygen consumption. The NET-Seq data on this gene shows the shifts in pas site usage is due to active transcription of *MIS1* transcripts with the proximal PAS. This is shown in the NET-Seq data on 3.38 on page 185. The TIF seq data shows clear evidence for APA in the 3'UTR, with two variants of transcripts occurring at both 3'UTR pas site locations, as shown in the IGV tracks 3.39 on page 186.

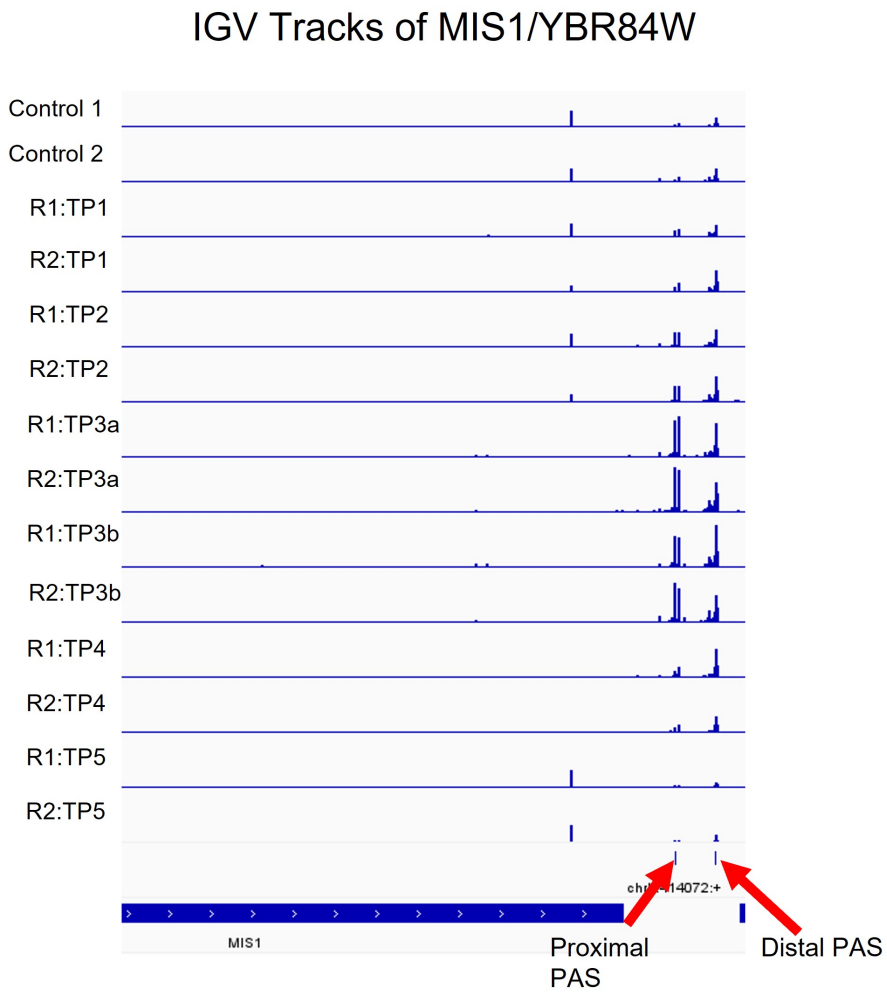


Figure 3.36: IGV tracks for shifts in pas site usage for the gene *MIS1/YBR84W*. R1 is repeat 1 and 2 of each time point respectively. Red arrows indicate the pas site location for proximal and distal pas sites.

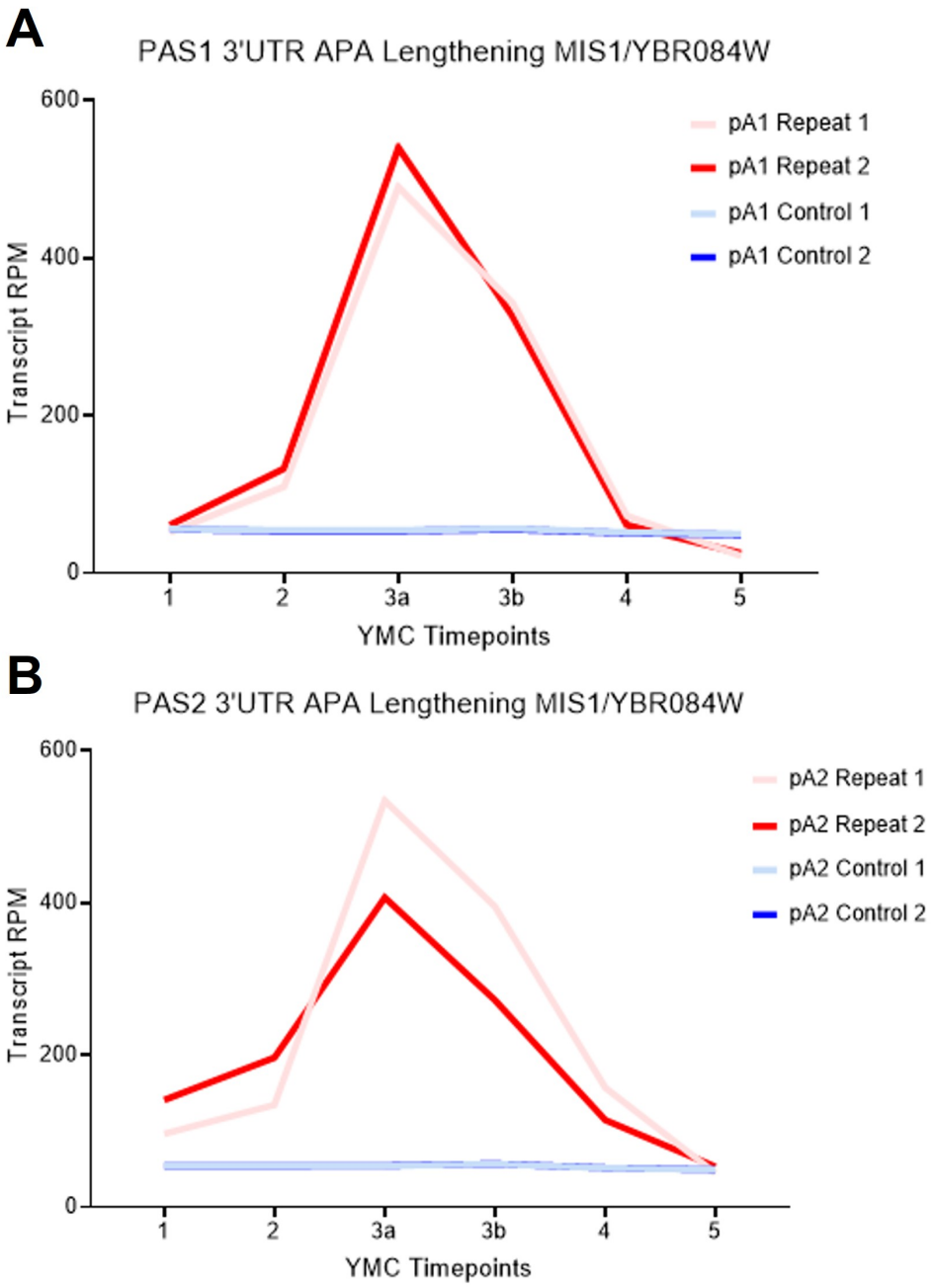


Figure 3.37: Line graphs of distal and proximal pas shift in gene expression for *MIS1/YBR84W*. A shows the changes in RPM, i.e. number of transcripts with the proximal pas end throughout the YMC. B shows the changes in RPM, i.e. number of transcripts with the distal pas end throughout the YMC.

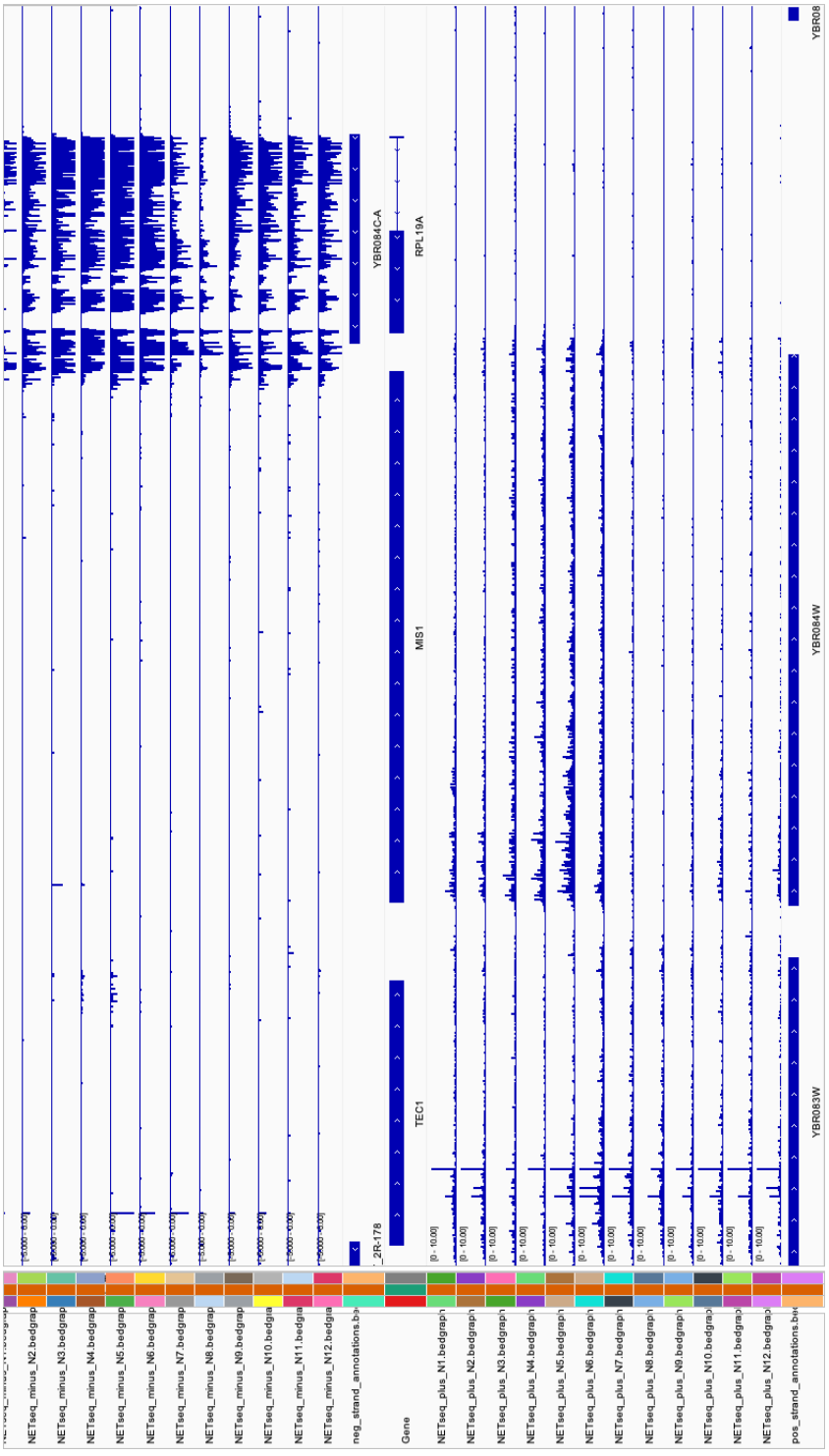


Figure 3.38: IGV tracks for Net-Seq data showing shifts in pas site usage for *MIS1/YBR84W* in the OX phase of the YMC. Tracks are shown on the top, negative strand. Time point N1 is the start of the OX phase in the YMC and N5 is the peak of oxygen consumption (see below for profile of sampling for NET-seq on oxygen use profile) and matches the increase in transcript levels as the IGV tracks for 3'- RNA-Seq data shown above.

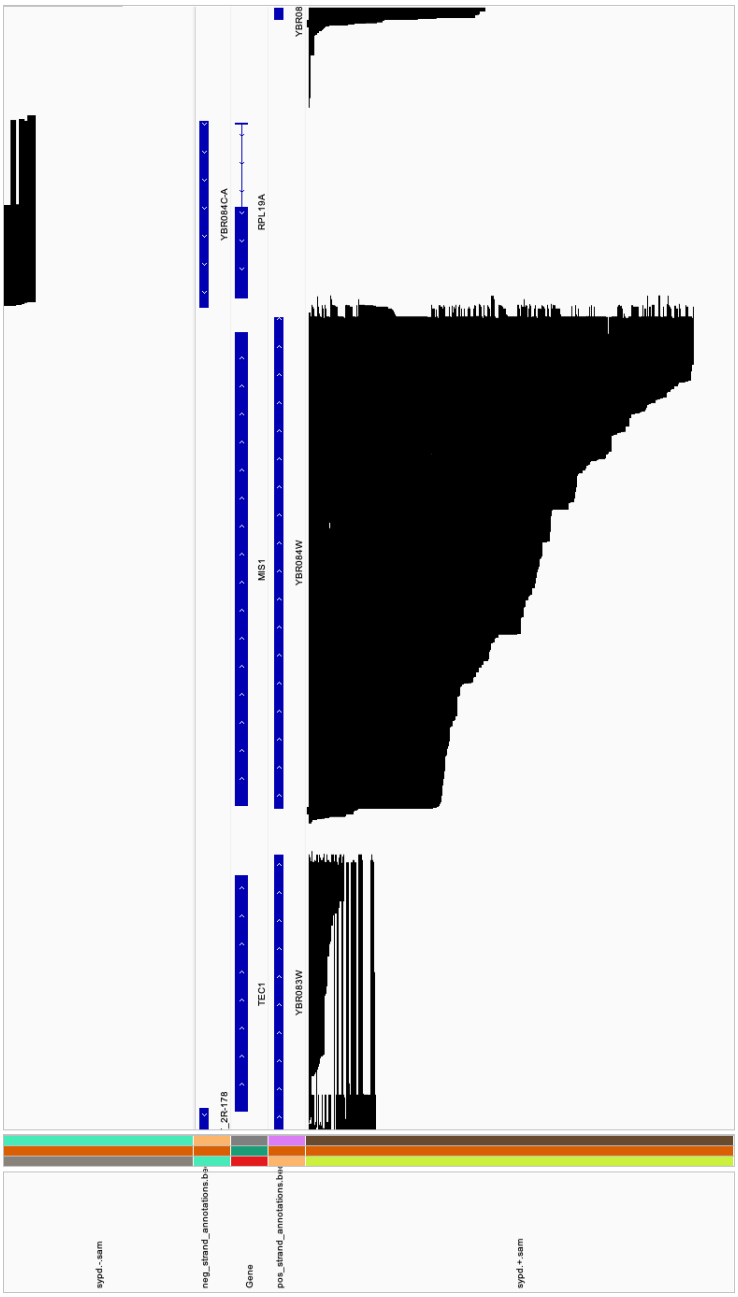


Figure 3.39: IGV tracks for TIF-Seq data showing both distal and proximal *MIS1/YBR84W* 3'UTR transcripts, therefore confirming this APA type occurs for *MIS1/YBR84W*.

3.10.2 IGV tracks of 3'RNA-Seq data show *YPR151C/SUE1* undergoes ALL APA shortening at time point 3a and 3b

SUE1 is a a nuclear gene that encodes a mitochondrial protein whose function is to help facilitate protein degradation, localized to the mitochondrial envelope (J. Wei and Sherman 2004). The RNA-Seq data set identified the gene *YPR151C/SUE1* as undergoing a Coding APA shortening event at time points 3a and 3b in the YMC. The IGV tracks for *SUE1/PR151C* show the swap in pas site usage from the distal pas site in the 3'UTR to the proximal pas site that is located in the coding region of the gene. The APA shortening event is shown in on page 189. In the IGV tracks show how the distal pas site collapses with a big decrease in count reads in time point 3a and 3b and a small increase in reads for the proximal pas site. The collapse of reads for the distal pas site causes the ratio of shorter transcripts to increase, i.e. APA shortening. These shifts in pas site count reads along the YMC is shown in the graph 3.41 on page 190. The NETseq data for SUE1 shows a more complicated pattern of transcription however.

3.10.3 NETSeq IGV tracks of *YPR151C/SUE1* show a complex transcriptional environment that shows potential read through from neighbouring genes

The NETSeq data shows the peak of transcription at the RB phase, however, there is also some convergent overlapping transcription on the opposite strand to *SUE1* in the gene *NCE102* shown in the NET-SEQ IGV tracks

in 3.42 on page 191. The NET-seq data is shows different read densities in SUE1 between the CR-APA pas sites in time points between N5 and N8, which corresponds to the 3a and 3b time points the pas site swaps are shown in the 3'RNA-SEQ data. However, the IGV tracks show that transcription may not start at the SUE1 TSS, as there is some read through from the other gene so it is not clear how much of the change in pas site usage is due to APA. Furthermore the two genes both show a decrease in reads at time points N2-N8, therefore indicating some other regulatory mechanism occurring for suppressing the expression of *SUE1* in the OX phase of the YMC.

3.10.4 TIF-Seq data supports 3'RNASeq data for *YPR151C/SUE* using two pas sites

To identify whether the changes in reads seen in the NET-Seq data were due to CR-APA shifts, the TIF-seq data was analysed. The TIF-Seq IGV reads supports the notion that SUE1 uses the two pas sites identified in the APA pipeline, as shown in 3.43 on page 192. TIF-seq data is shows both forms of the APA isoform using both pas sites. This indicates that SUE1 is able to perform CR-APA, however, TIF-Seq data on the YMC would be needed to confirm this is happening specifically in the YMC.

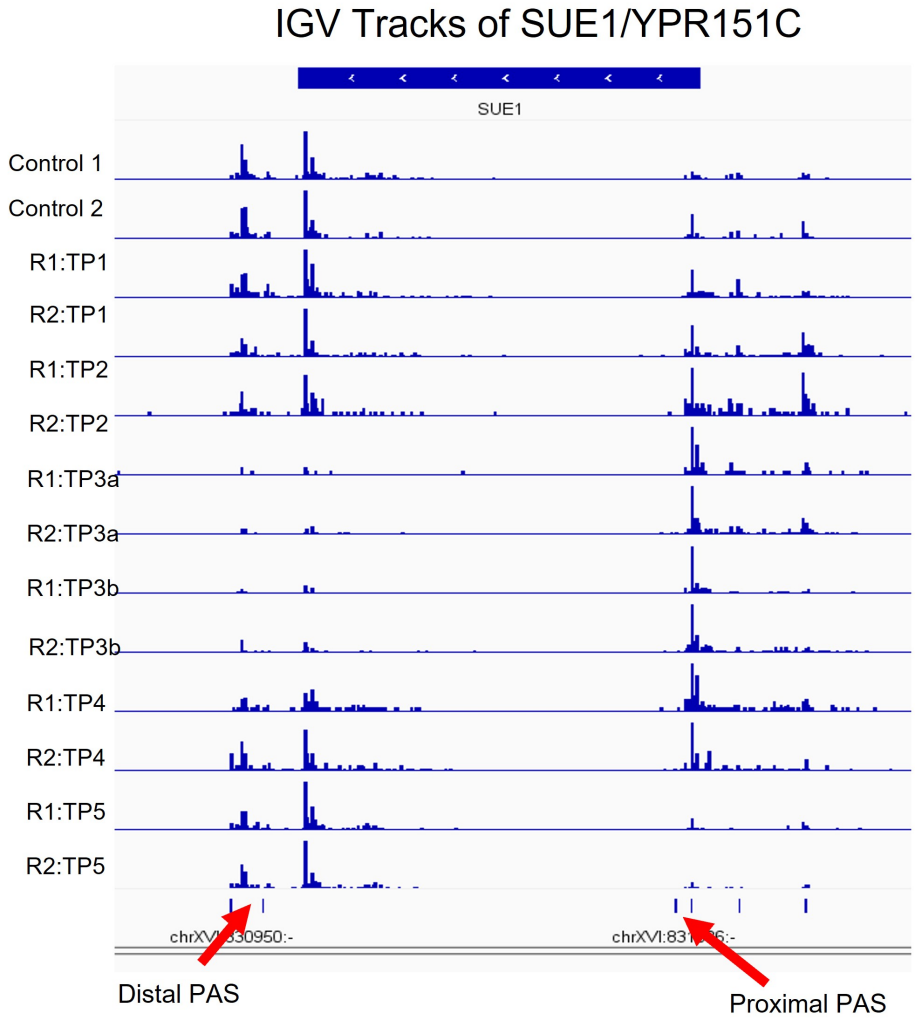


Figure 3.40: IGV tracks for shifts in pas site usage for the gene *SUE1/YPR151C*. R1 is repeat 1 and 2 of each time point respectively. Red arrows are indicating pas site location for proximal and distal pas sites.

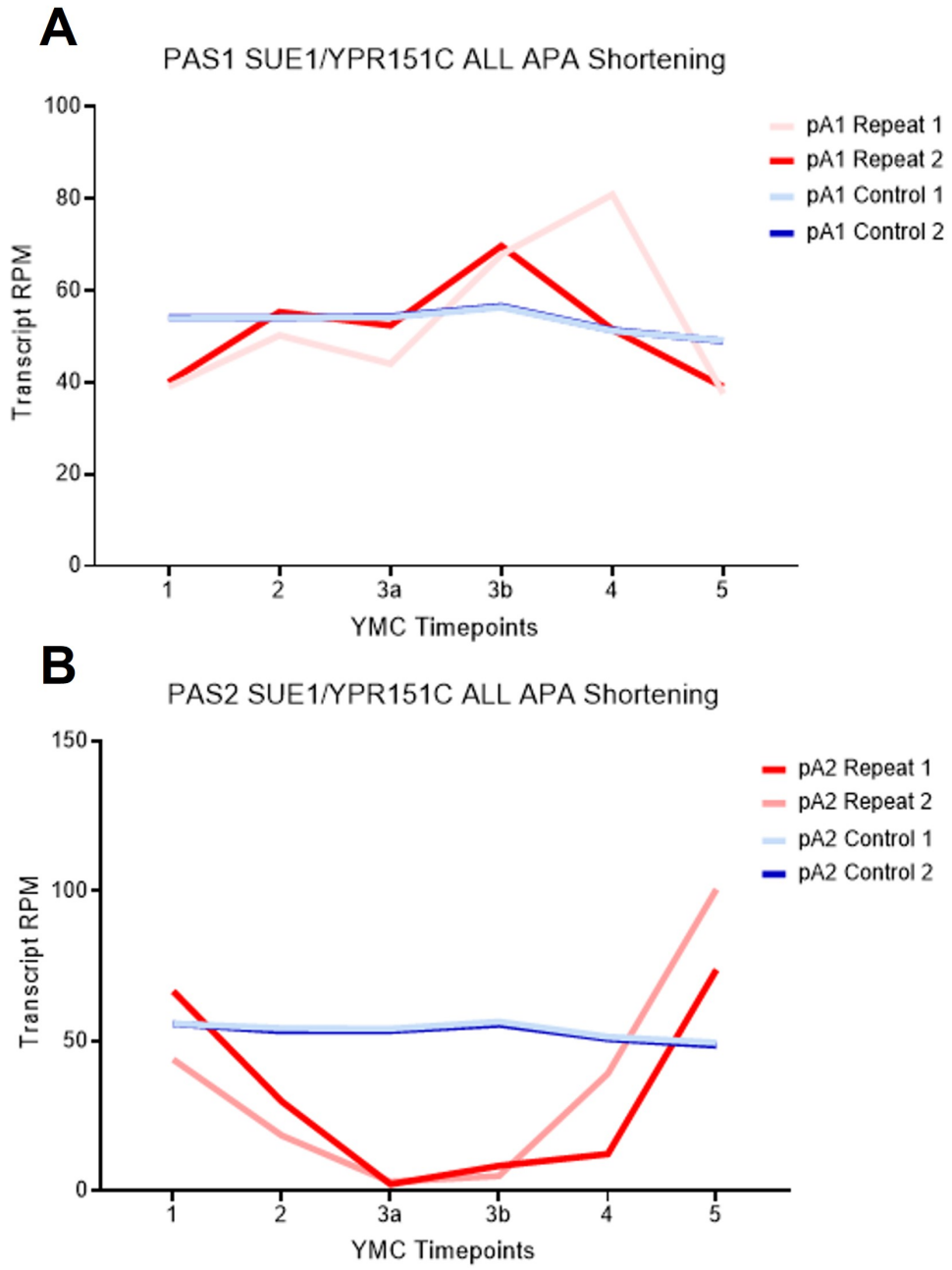


Figure 3.41: Line graphs of distal and proximal pas shift in gene expression for *YPR151C/SUE1* showing collapse of distal (pas2) usage at time points 3a and 3b. A shows the changes in RPM, i.e. number of transcripts with the proximal pas end throughout the YMC. B shows shows the changes in RPM, i.e. number of transcripts with the distal pas end throughout the YMC.

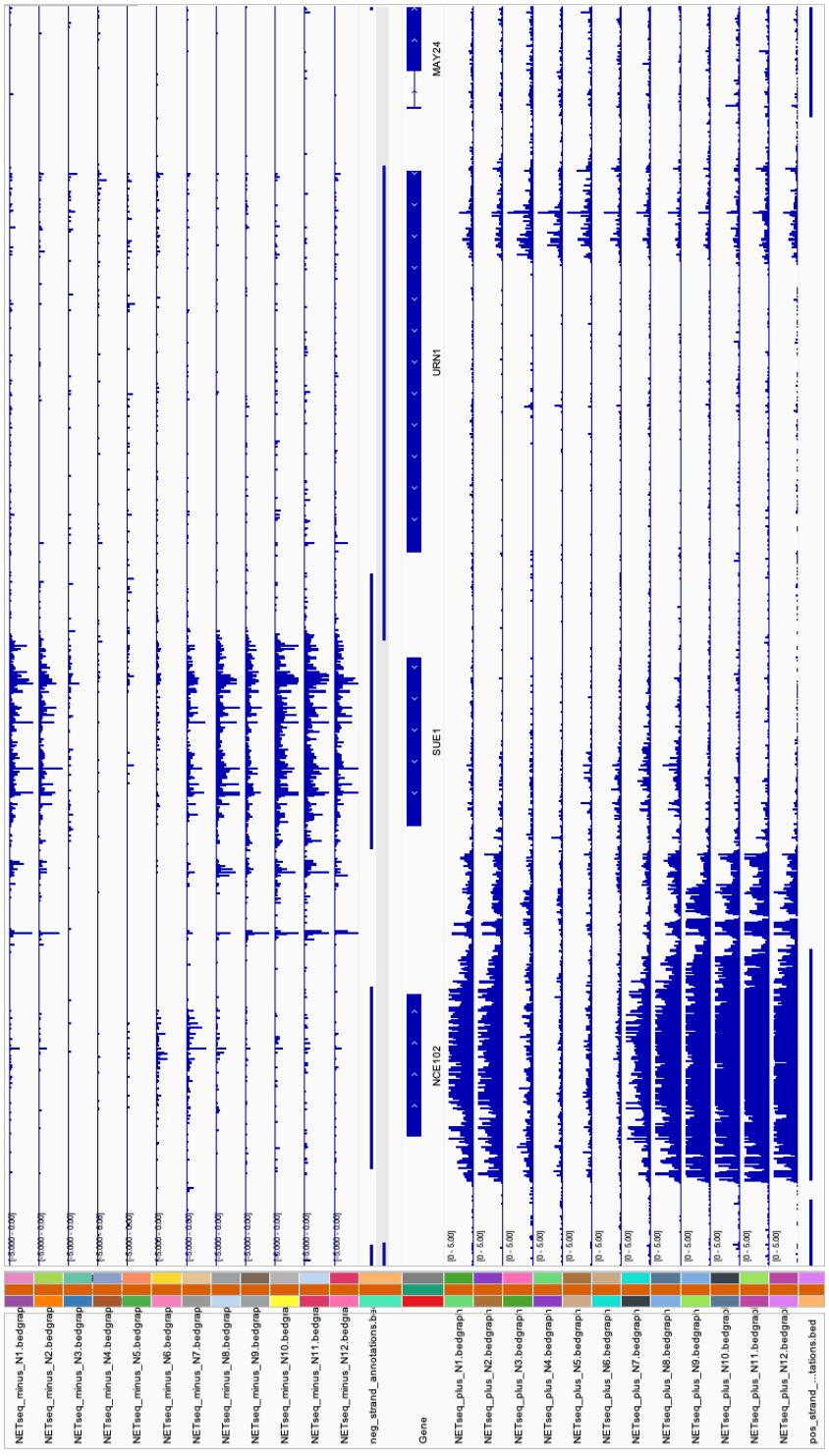


Figure 3.42: IGV tracks for Net-Seq data showing shifts in pas site usage for *YPR151C/SUE1* in the RB phase of the YMC. Tracks are shown on the top, negative strand. Time point N8 is the start of the OX phase in the YMC and matches the same pattern in PAS usage as the IGV tracks for 3-RNA-Seq data shown above.

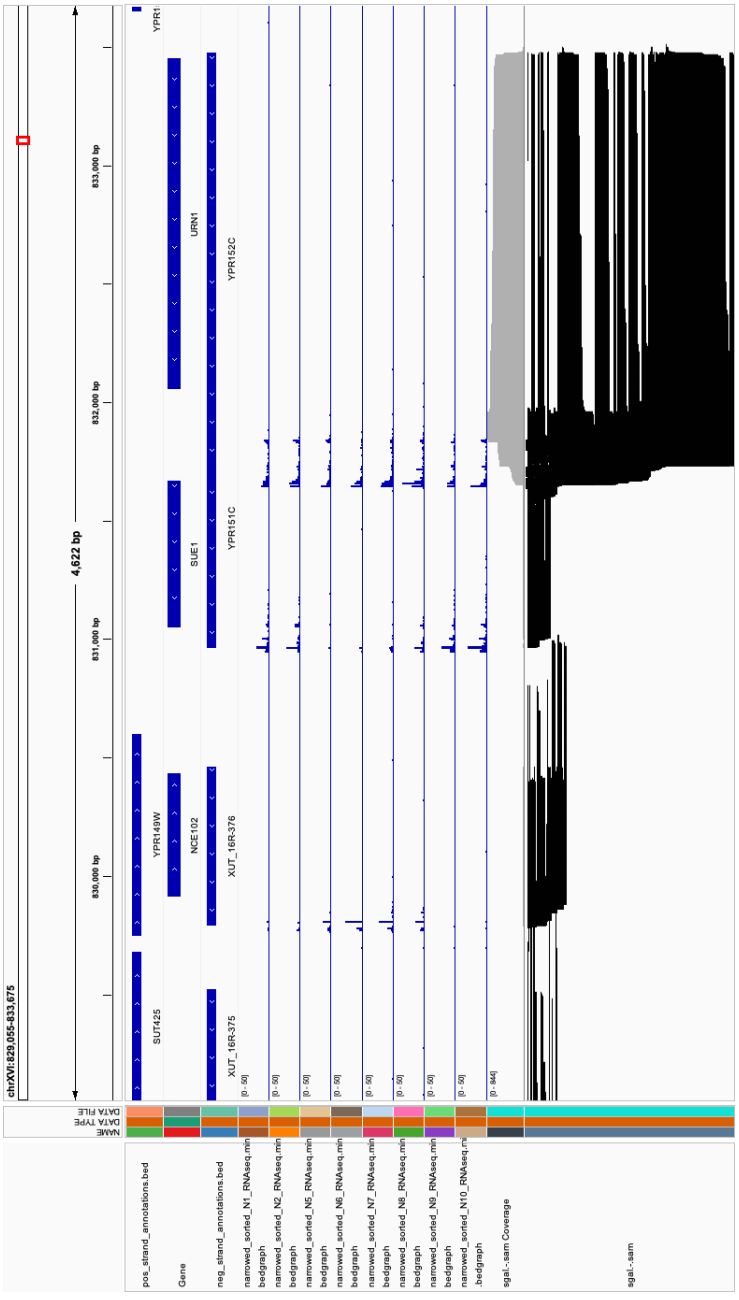


Figure 3.43: IGV tracks for TIF-Seq and 3'RNA-Seq data. RNA-Seq data is in blue and shows the collapse of distal pas site during time points 3a and 3b. TIF-seq data is shown in black and shows both forms of the APA isoform using both pas sites.

3.10.5 IGV Tracks for *YBR158W/AMN1* show ALL lengthening APA thoughout the YMC

AMN1 is a yeast gene that plays a role in regulating the exit of cells from the mitotic phase of the cell cycle. *AMN1* interacts with the Tem1p GTPase, breaking up the Tem1p-Cdc15p complex, and inhibiting cell separation through the proteolysis of Ace2p, *AMN1* is also involved in the mitotic cell cycle checkpoint (Y. Wang et al. 2003).The RNA-Seq data set identified the gene *YBR158W/AMN1* as undergoing ALL lengthening at time points 1, 2, 3a and 3b in the YMC. The pas sites that switch in usage are the proximal upstream site and the pas site in the 3'UTR. The IGV tracks for the gene *YBR158W* visualised this lengthening event as shown in on page 195. Visualising the data that is made from the count reads of the two pas sites from which the pas site shift occurs shows that there is a reduction in proximal pas site usage with the distal site having little change in count reads in 3a and 3b. This shift in proximal pas site count reads is shown in the line graph 3.45 on page 196,

3.10.6 NETSeq IGV tracks indicate transcriptional interference influences pas site selection of *YBR158W/AMN1* in the YMC

NETSeq IGV tracks show similar trends to RNA-Seq data with a proximal pas collapse with no change in reads at the distal pas site at time points N4-N7, which also occurs at RNA-Seq time points 3a and 3b. NET-seq data shows transcripts starting at the proximal pas site predominately occur in

the OX phase which are then repressed throughout the rest of the YMC. There are reads that overlap heavily with the upstream gene *SUT020*. This infers transcription interference might mediate alterations in pas site usage in *AMN1*, as the proximal pas site is in close proximity to the neighbouring gene and the distal site no change in usage throughout the YMC. The overlap of reads can be seen of the NET-Seq tracks 3.46 on page 197.

3.10.7 TIF-Seq IGV tracks of *YBR158W/AMN1* confirms NET-Seq data and suggests *SUT020* read through to *AMN1*

The TIF-seq IGV tracks support the NET-Seq data which shows read through transcripts from *SUT020* to the proximal pas site of *AMN1* 3.47 on page 198. TIF-Seq shows most of the transcripts start in the middle of SUT020 and overlap with the TSS of *AMN1* and thus are highly likely to mediate transcriptional interference. This is not traditional APA as the *AMN1* is being repressed via blocking the proximal pas and the TTS are being affected by an upstream gene due to read through. This will have to be confirmed via TIF-Seq during the YMC to see if these read through transcripts exist in the YMC and co-occur with proximal pas site collapse.

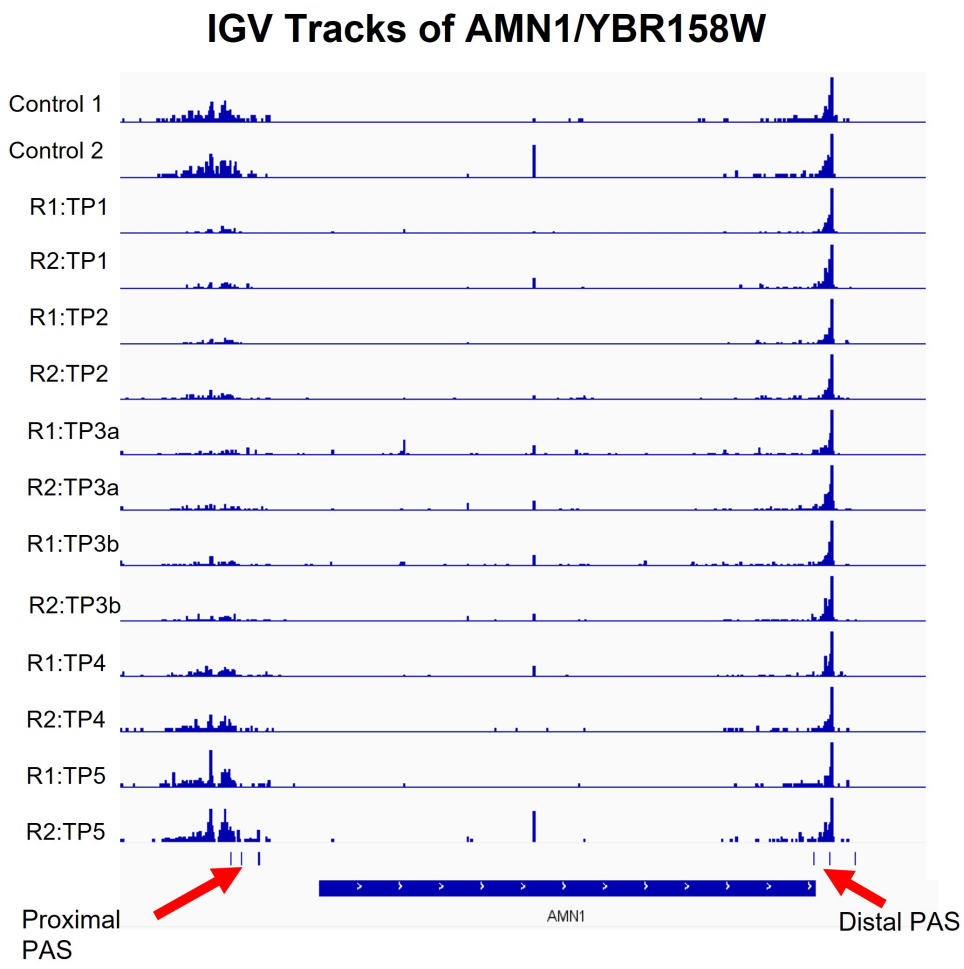


Figure 3.44: IGV tracks for shifts in pas site usage for the gene *YBR158W/AMN1*. R1 is repeat 1 and 2 of each time point respectively. Red arrows are indicating pas site location for proximal and distal pas sites.

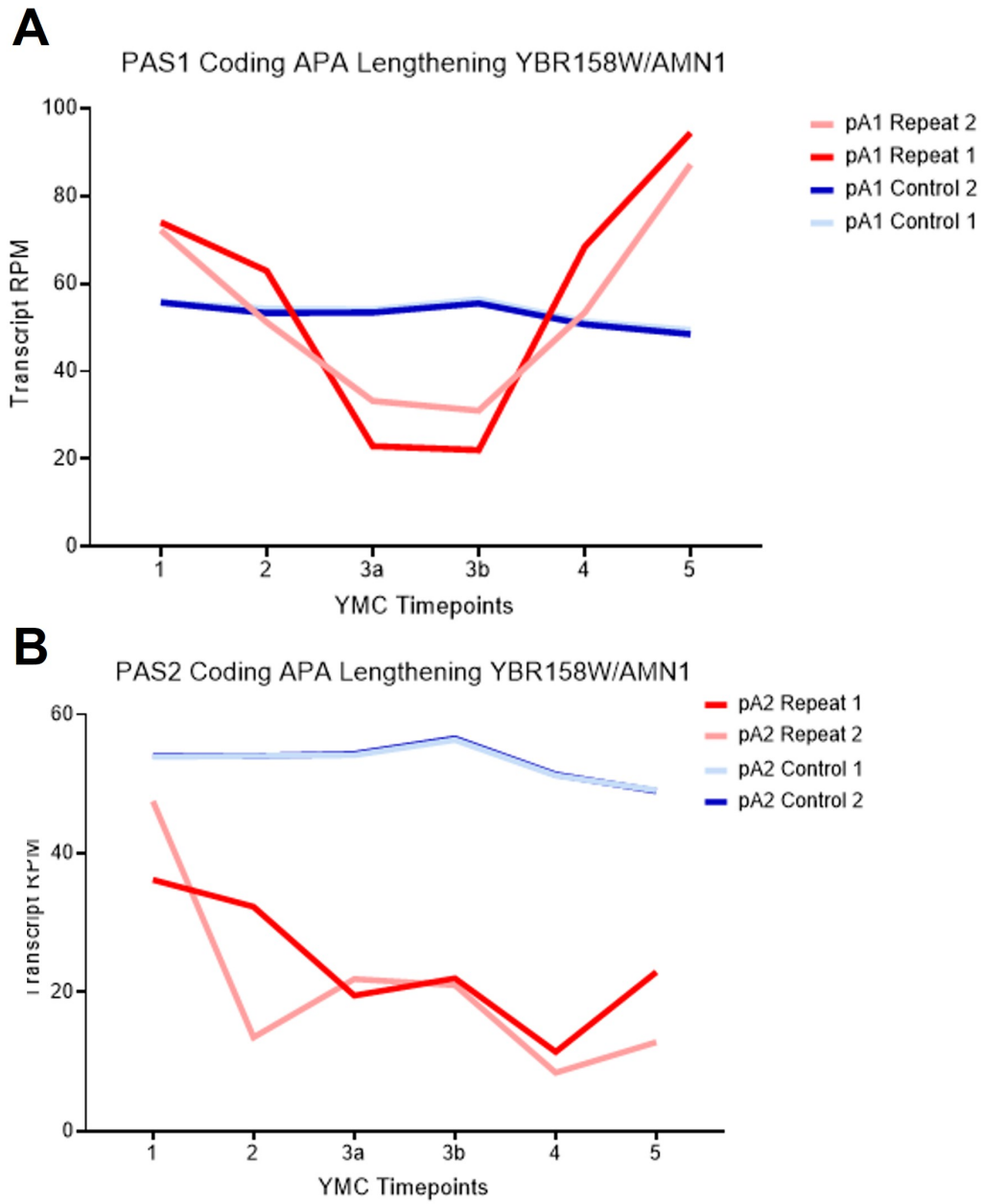


Figure 3.45: IGV tracks for shifts in pas site usage for *YBR158W/AMN1* of proximal (pas1) shown in red and distal (pas2) shown in blue, usage at time points throughout the YMC. A shows the changes in RPM, i.e. number of transcripts with the proximal pas end throughout the YMC. B shows the changes in RPM, i.e. number of transcripts with the distal pas end throughout the YMC.

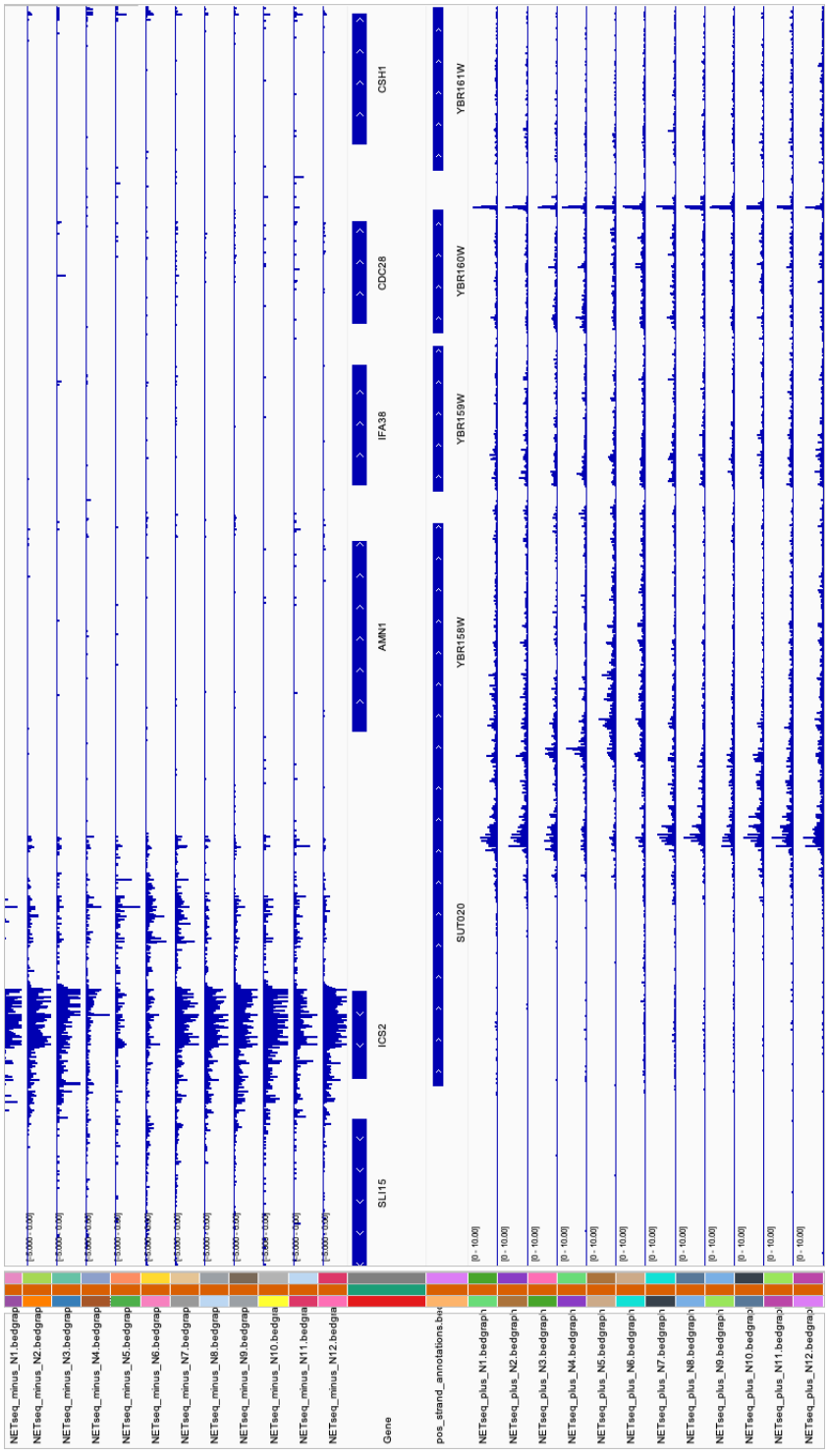


Figure 3.46: IGV tracks for Net-Seq data showing shifts in pas site usage for *YBR158W/AMN1*. Tracks show *YBR158W/AMN1* runs on the opposite strand to *SUT020* and transcripts show potential transcriptional interference.

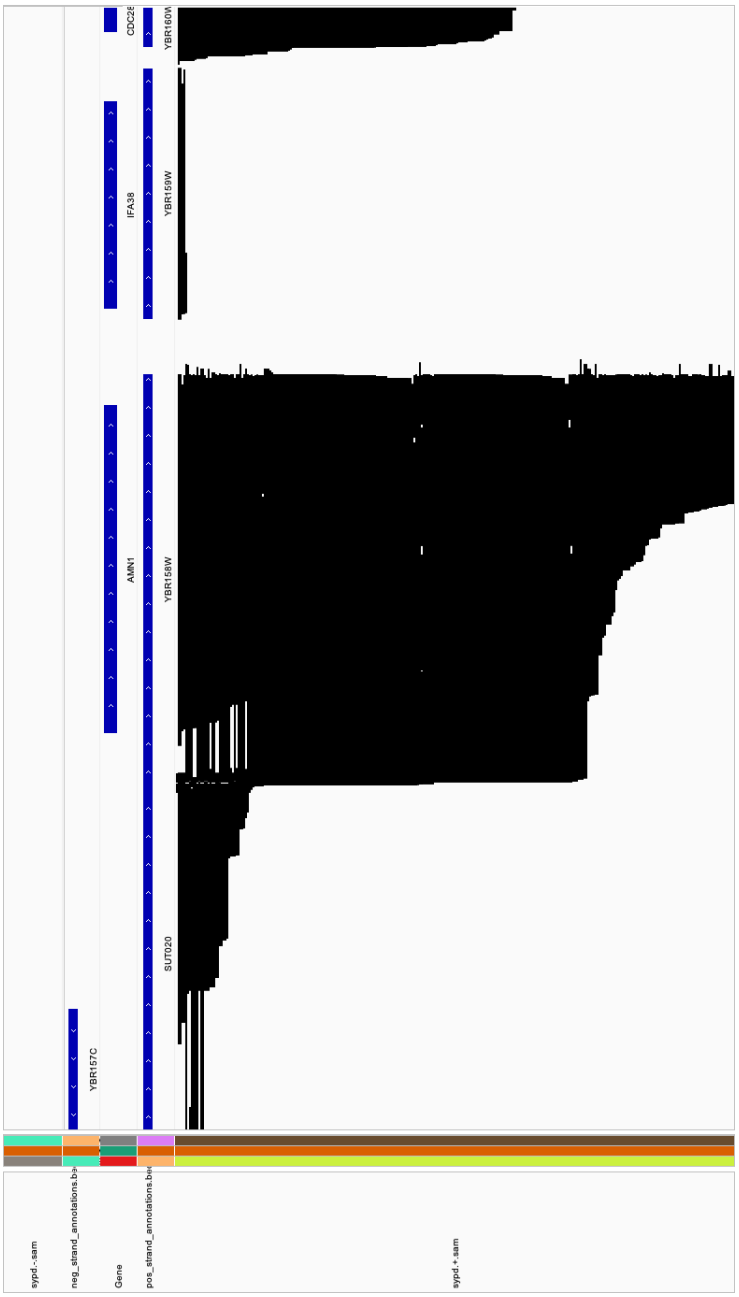


Figure 3.47: IGV tracks for *YBR158W/AMN1* TIF-Seq. TIF-Seq tracks shown in black show many transcripts are read through from the upstream gene *SUT020*, indicating the reduction in proximal pas usage is due to transcriptional interference and suppression of these upstream transcripts and not a traditional APA event.

3.11 Discussion

The key findings of this chapter are that APA and gene expression changes are highly prevalent at all time points during the YMC. The results indicate that the most transcriptionally active period of the YMC occurs at the end of the reductive building phase and the start of the reductive charging phase. The APA events peak at the same time points as gene expression, which suggests a functional purpose for APA in gene expression control. The functional nature of APA in the YMC is supported by the lack of skew towards more transcriptionally active genes, the difference in Δpas lengths of the genes that underwent APA and the correlation between specific gene expression changes and APA events (gene expression decreases and shortening and gene expression increases and APA lengthening). The most prevalent APA events were APA lengthening that occurred in the ALL data set. The GO analysis of the genes that undergo APA changes were enriched in several functional groups related to the YMC, such as cell cycle progression and DNA damage. Cluster analysis identified two to three distinct clusters of these genes, which also overlapped with previously established gene expression clusters. The comparison between the NET-Seq and 3'RNA-Seq data sets showed that APA events are not just due to passive APA via increased transcript stability, but are the result of active APA due to changes in the selection of alternative polyadenylation sites. Further analysis of NET-seq, RNA-seq and TIF-Seq IGV track data showed similar patterns of gene expression in the YMC, indicating that the changes in APA genes are due to active changes in transcription. Furthermore, the IGV tracks revealed potential mechanisms of

polyadenylation site selection that may involve transcriptional interference of upstream genes. Overall, the results suggest that APA plays a significant role in regulating gene expression in the YMC and may have functional importance in this process.

3.11.1 3a and 3b most active time points for gene expression changes and shifts in APA

The results show that APA and gene expression changes are highly prevalent at all time points along the YMC. The overall number of genes undergoing gene expression changes was in line with the literature also, as 30.1% of genes underwent changes, as other work has shown between one third and half of all genes undergo these changes (Tu, Mohler, et al. 2007). The results indicate that APA events and changes in gene expression mostly occurred at time points 3a and 3b, which was the end of the reductive building phase and the start of the reductive charging phase. This indicates that the most transciptomically active period of the YMC occurs between these two time points. This is consistent with the literature regarding gene expression profiles of the YMC, that indicates that reductive building/reductive charging are extremely active periods of gene expression (Tu, Kudlicki, et al. 2005). The RB/RC phase is one that follows the shift from aerobic to anaerobic glycolysis and also is where DNA replication and Cell cycle progresses. These processes all require the expression of a plethora of genes and subsequent proteins to enact these key cellular activities, therefore it is expected to be a high point in the cells gene expression output. The results found are also in concordance with the NET-Seq data from the Mellor laboratory, which

showed increase number and amplitude of gene expression changes at these time points. The changes in APA also follow the same trends as gene expression and also peak at time points 3a and 3b. APA profiles along the YMC have not been well established previously, however, APA has been found to affect the expression, localisation, and stability of transcripts. The APA events peaking at the same time points as gene expression could indicate a functional purpose of APA, as it has been shown to be a powerful mechanism of gene expression control.

There are a few limitations of these findings. Firstly, it is not possible to say if the APA shifts are having an effect on the gene expression profiles or vice versa, therefore we can only say they both seem to peak at time points 3a and 3b and using what has already been established by previous studies on APA , we can infer, but cannot prove that APA might be directly causing gene expression shifts. However, there wasn't a strong correlation between increased changes in gene expression and APA changes of genes, which we would expect to see if it the increase in APA was just a characteristic of highly expressed genes.

3.11.2 Skew analysis of APA events are not an artifact of large shifts in gene expression change

The scatter plots comparing APA events to logfold2 changes in gene expression showed there was no skew of the data towards highly transcribed genes. This means that the genes that were highly active in regards to gene expression changes were not more or less likely to undergo APA as less active ones. This indicates that shifts in PAS usage is not just a consequence of

active genes and indicates a functional use for APA. The functional nature of APA in the YMC is further suggested due to the fact that APA genes is correlated with gene expression expression changes. This implys that these gene expression changes are due to APA and not the other way around, since if this were the case, we would see genes with larger gene expression shifts undergoing more APA than ones with smaller changes. As APA is known to effect gene expression, this is strong evidence that APA is being used to influence the transcriptome of the YMC (Yuan et al. 2021).

3.11.3 The YMC is hallmarked by global ALL APA lengthening events

The most prevalent APA events are those that occur in the ALL data set, i.e. shift in pas usage from the 3'UTR to the coding region and vice versa. In all types of APA events i.e. coding, 3'UTR and ALL, APA lengthening was significantly more common than APA shortening. The ALL data set being the most common type of APA is not surprising as it contains more PAS sites than both the 3'UTR and Coding data sets, therefore more APA events are likely to be observed. Previous studies have shown that the 3' 3'UTR is the most prevalent location for PAS sites in the yeast genome, however, the 3'UTR data set showed less APA events then the coding region (X. Liu et al. 2017). However, this could be due to the pipeline only showing the two most used pas sites per gene and as there are often many more than two pas sites capable of performing APA, many APA shifts will not be observed. APA lengthening was significantly more common than shortening for all APA types and at all time points along the YMC. Global APA length-

ening patterns has been shown to be associated with cell differentiation and development (Shepard et al. 2011)(Cheng et al. 2020). The YMC doesn't involve cellular differentiation or development so it is not possible to credit these processes in causing global APA lengthening, however it is important to acknowledge that these are essential and complex cellular functions which require strict temporal control over gene expression, which could be the underlying reason why we see similar trends in the YMC that also shares these features. These results are somewhat surprising as the cell is experiencing glucose starvation and therefore stress once the OX phase has used up intercellular glucose stores, which is usually hallmarked with global shortening (Badjatia et al. 2021). However, this shortening is usually accredited to the stalling of RNAPII elongation, which is an unknown factor in this analysis. In order to investigate the underlying mechanisms that the APA is having in these key functional genes, further studies need to be conducted, such as RNAPII elongation rates, knocking out the pas site that is switched to during APA and seeing the effects on the expression levels of the genes.

3.11.4 APA lengthening and gene expression increases and APA shortening and gene expression decreases are correlated

The analysis of genes that undergo APA and differential expression changes identified a small preference towards genes that display both APA events and differential gene expression, indicating a possible connection between the two processes that merits further exploration. It's worth noting, however, that the proportion of genes that exhibit both APA and DE is relatively low, typ-

ically amounting to less than 3% in each data set. The genes that undergo APA and DE were more prevalent but the proportion of the type of APA events (i.e. 3'UTR, Coding) did not appear to be influenced by differential expression occurring. This was a surprising result as the literature suggests that genes that undergo APA but do not change differential expression are being influenced to change stability or localiztion, as these events do not require a change in gene expression. As both localization and stabilization have usually been shown to occur in the 3'UTR of the transcript, you would expect genes undergoing APA without showing differential expression would show an increased proportion of these 3'UTR APA events. This suggests that there might be other reasons for APA to occur beyond what the literature suggests or that APA events might not all cause functional changes to the transcript. More discernable observations occurred after conducting further analysis after separating the lengthening and shortening APA events and changes in gene expression.Of the genes that underwent both APA and gene expression changes, there was a strong preference for lengthening and gene expression increases and shortening and gene expression decreases. These results are somewhat surprising as previous investigations into pas lengths and gene expression changes have indicated APA shortening causes gene expression to increase. This is due to the shorter transcripts having shorter 3'UTR lengths and therefore lack signals for negative regulation such as binding proteins that might destabilise the transcripts (Mayr and Bartel 2009), (E. T. Wang, Sandberg, et al. 2008). However, most of these studies were done on mammalian cells that rely on micro-RNAs to target 3'UTRs of transcripts for degradation. As yeast do not have miRNAs this method is not used

and therefore APA lengthening might have another role such as localisation. Furthermore, there is no one rule for APA length and effect on a transcript and there are many conflicting examples of lengthening and shortening having a range of effects on gene expression. The longer transcripts produced by preferable distal pas site usage could be increasing the stability and therefore increasing the protein output of the transcripts that undergo APA.

3.11.5 GO analysis showed genes undergoing APA were enriched in functional groups that undergo key biological mechanisms in the YMC

The GO analysis of genes that undergo APA changes were enriched in several functional groups that are tightly associated with mechanisms that occur in the YMC such as cell cycle progression and DNA damage. As APA has been shown to to be a key mechanism of gene expression control, it is possible that the over representation of of these genes undergoing APA is functional and not by chance. The genes that were enriched for GO terms related to the YMC mostly underwent lengthening events, which shows that YMC genes undergo a general trend of APA lengthening throughout the YMC.

3.11.6 Cluster analysis showed two to three distinct clusters of genes that undergo APA changes and overlap with pre-established gene expression clusters

Cluster analysis of genes that undergo shift in APA throughout the YMC show 2 to 3 distinct clusters for each data set. The K-means clustering algorithm used to identify the shape and numbers of these clusters showed most data sets had just two clusters apart from ALL apa shortening that had three. The cluster analysis of the APA data sets demonstrated the overall trend of genes undergoing APA lengthening also undergoing gene expression increases and shortening undergoing decreases in gene expression. Furthermore, the gene expression clusters established by the Mellor laboratory group had significant overlap with the genes that underwent APA changes, especially in the 3rd group of the super clusters. This indicates that the genes that underwent APA changes are significant in the temporal gene expression clusters of the YMC, which could indicate a functional purpose related to controlling the gene expression profiles of these clusters.

3.11.7 Δpas lengths overall shorter in genes that undergo APA changes in the YMC

The Δpas lengths of the genes that undergo APA changes in the were significantly shorter than genes that had multiple pas sites but did not undergo shifts in pas site usage, apart from the coding shortening data set. This indicates that the APA shifts observed are not randomly selected but

favour shorter Δpas lengths. The Δpas lengths being shorter for APA genes compared to the genes that do not undergo APA could indicate specific sequences are being targeted for inclusion/removal via APA changes. Previous studies have shown that changes in a single base pair in the 3'UTR mRNA sequence can affect the stability rates of transcripts (Schwerk and Savan 2015). Therefore APA genes having short distances between these alternative pas sites would still contribute significant changes in the stability, transcription and localisation of transcripts.

3.11.8 NET-seq profiles of nascent transcripts indicate that APA is active and due to changes in pas site selection in the trancriptome

The Net seq profiles followed the same trends that were observed in the RNA-Seq data. As NET-Seq detects nascent, actively transcribed mRNA, this indicates that the changes seen in the genes that undergo APA shifts are due to active changes in transcription occurring throughout the YMC.

There are limitations in the comparisons we can draw from the RNA-Seq and NET-Seq data sets. The time points are different for both data sets even though they both cover one oscillation of the YMC. The NET-Seq data has 10 time points whilst the RNA-Seq data has only 6 time points. This means that the comparisons made between these different data sets are not going to be at the exact same time points in the YMC as they are not recorded at the same times. This is negated by looking at the overall trends in the shifts between the three super clusters OX, RB and RC. The oxygen traces for both data sets are available and the closest time points to cover these

three regions were chosen so that the data matches up as close as possible.

3.11.9 IGV reveals a range of potential mechanisms of pas site selection occurring in the YMC

The IGV tracks for the APA events identified in the pipeline showed that pas site usage changes throughout the YMC occur via a range of different mechanisms, furthermore comparing IGV tracks for the 3'RNA-Seq, NET-Seq and TIF-Seq revealed potentially non traditional APA events occurring due to transcriptional interference. There was a wide range of variations of absolute changes in gene expression that led to APA shifts. Looking at the APA events on the igv tracks showed that there was not one type of gene expression change that resulted in an APA event. For example, two APA events might both be categorized as shortening but one is due to a large increase in proximal reads and the other due to a big decline in distal reads. These two APA shortening events must occur via different mechanisms of action. An APA shortening event that occurs via the distal pas site having a decreased number of reads with no change in the absolute number of transcripts with proximal pas site usage, is occurring due to targeted degradation of the distal isoform or repression of that pas site being used during 3 end formation. An APA shortening event that occurs due to an increase in reads for the shorter isoform is due to an increase in proximal pas site usage or a stabilisation of the shorter transcript. Seeing that the APA events are due to a mix of gene expression changes via the proximal and distal isoforms indicates that APA in the YMC is not a result of a one specific mechanism. To further explore the mechanisms of APA, it would be useful to separate

out the APA events that occur via the same alterations in gene expression i.e. shortening via increased proximal pas site usage/increasing transcript stability and shortening via distal isoform degradation/reduction in as site usage. Identifying these subcategories of APA events would bring us closer to understanding the mechanisms behind these global shifts in APA.

The integration of IGV tracks with the APA pipeline has led to the identification of several novel insights into non-conventional APA events and transcriptionally interesting loci within the YMC, which could be further studied to enhance our understanding of gene expression control mechanisms. The *SUE1* gene was identified as undergoing CR-APA during the YMC. The IGV tracks confirmed this event but also showed major suppression of gene expression of the neighbouring gene *NCE102* on the opposite strand. This indicates a potentially novel mechanism of transcriptional repression at the OX phase of the YMC.

A further example of how this combinatorial analysis approach to investigating APA has revealed a wider range of gene expression control mechanisms is the pas site changes observed in *AMN1* during the YMC. The APA pipeline identified APA occurring in *AMN1*, however, the IGV TIF-Seq tracks showed the existence of a transcript read-through from the genes *SUT20* to *AMN1*. This transcript could block the TSS of *AMN1* and transcriptional interference. The Proximal pas usage collapsing might be due to overall gene suppression of the transcripts as the gene expression peaks in the OX phase and drops throughout the YMC. This mechanism of gene expression control in yeast has been identified in previous studies (Yague-Sanz et al. 2020). Furthermore, studies of transcriptional interference identify APA

changes in upstream interfering noncoding RNAs as being able to cause gene derepression in neighbouring genes, acting as a powerful mechanism of gene expression control.

3.11.10 Further investigations

Further investigations can be done to develop these initial findings. Firstly, the APA events seen in these data sets can be confirmed via qPCR. This would confirm the results from the 3-RNA sequencing and quantify the changes in gene expression that occurs with the shifts in APA. To investigate if the APA changes observed had an effect on the protein output of the genes, western blott analysis could also be performed. As the literature shows that APA can alter the protein output of transcripts, it would be useful in understanding the function of specific APA events to see if the protein output is altered when they undergo APA. Further investigations using the data that has been generated in this thesis can be done to identify specific APA events as a mechanism of gene expression control of key genes in the YMC. Due to time constraints, it was not possible to explore all of the individual genes and identify mechanistic functions relating to APA. The key APA genes identified to be GO enriched for essential YMC functions such as cell cycle progression should be further investigated. This could be achieved via exploring common ratio shifts using the IGV tracks, if a group of these functionally related genes show similar IGV patterns, this would imply a common APA mechanism. This could then be further proven via motif analysis of the Δ pas sequence region. If these genes show APA is controlling the inclusion or exclusion of a regulatory sequence that they all share e.g. a bind-

ing site for a protein that targets that transcript for degradation, this would be strong evidence of APA directly being used to alter the transcriptome of key genes in the YMC.

Transcriptional interference was seen in the genes that were identified as undergoing APA throughout the YMC. This is an interesting gene expression observation in of itself, however it also opens up another area of exploration which can be done with the data sets already acquired. Shift in pas site usage from proximal to distal in upstream interfering noncoding RNAs can be used to cause gene derepression in downstream genes. This has recently been observed in (Yague-Sanz et al. 2020), in which the gene *tgp1*, a gene essential for responding to nutrient stress, is deprepressed via premature termination of *nc-tgp1* transcription, by switching from distal to proximal pas site usage, i.e. APA. Under normal conditions, the *nc-tgp1* used the distal pas site, and blocks the TSS of the downstream *tgp1* gene, causing silencing. During phosphatase starvation, RNAPII elongation is slowed, causing proximal pas site preference in *nc-tgp1*. This frees up the TSS of *tgp1*, allowing its expression. This mechanism is demonstrated in the schematic 3.48 on page 213 adapted from (Yague-Sanz et al. 2020). As this mechanism of APA driven gene expression control has been shown in nutrient starvation to control key genes, it is possible it is occurring in the YMC too. This can be evaluated by identifying genes in the APA data sets that overlap with the TSS of neighbouring genes. Using the gene expression data, you can identify what downstream genes undergo changes in gene expression. The TIF-Seq data will be able to show if read through transcripts exist at this loci. This will give a list of good candidate genes for APA mediated transcriptional interference. GO analysis

of these genes will then show if these candidate genes have functional roles in the YMC/stress response or have other important biological functions. The APA of these genes can then be confirmed via qPCR and knock out distal pas site mutants can also confirm the upstream APA is causing the changes in gene expression.

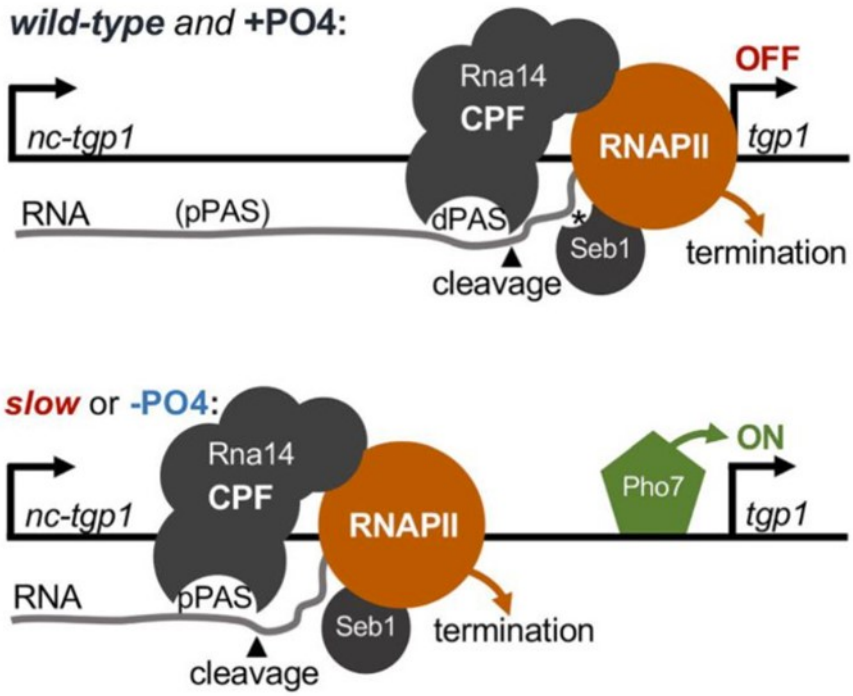


Figure 3.48: Regulation of *tgp1* expression by premature termination of *nc-tgp1* transcription by APA driven proximal pas site usage. The top diagram shows *tgp1* repression under normal cellular conditions. Distal pas site usage is occurring in upstream *nc-tgp1* and blocks TSS of down stream *tgp1*. The bottom diagram shows nutrient starvation/slow elongation of RNAPII. Slow elongation of RNAPII causes a shift from distal to proximal pas site usage, i.e. shortening APA. This frees up the TSS of *tgp1* and therefore enables expression and switching on of the gene.

Chapter 4

The APA profile of the Glu-Gal shift

4.1 Chapter Aims

The analysis of the APA profile of the YMC, investigated in previous chapters, identified several key genes that undergo APA changes that might reveal insights into gene expression regulation via APA in the YMC. In order to fully explore and understand the underlying mechanisms of these APA events, is important to identify if these events are specific to the YMC, or just a universal response to stress. Therefore, the glucose to galactose (Glu-Gal) shift, another process that induces metabolic stress in *Saccharomyces cerevisiae* was also scrutinised for APA events. The data sets analysed in this chapter are from yeast cells undergoing Glu-Gal shift. The Glu-Gal shift growth phase starts after *Saccharomyces cerevisiae* is starved of nutrients for 7-14 hours and then provided with a constant flow of glucose containing

medium, the first time point, GAL15 was taken 15 minutes after the introduction of galactose into the yeast growth medium. The Glu-Gal shift is a type of metabolic stress as it requires yeast cells to switch from glucose, the preferred carbon source to a non-preferred carbon source, galactose. This causes significant metabolic stress due to glucose starvation and requires the induction of the GAL pathway and an array of metabolic shifts via changes in gene expression of key galactose catabolic genes. Comparing the APA profiles of the YMC and Glu-Gal shift will therefore show if APA is just due to glucose starvation stress or is a targeted response mechanism. As with the analysis of the YMC, this chapter examines APA events in genes that are differentially expressed and those that are not. APA can affect gene expression, as well as other factors like localization and protein function that are not influenced by changes in gene expression. It's essential to carefully examine both datasets and analyze the relationship between APA and differentially expressed genes.

4.2 Large changes in the transcriptome occur during the Glu-Gal shift

As with the YMC data, identifying changes in gene expression was an essential component of investigating the APA landscape of the Glu-Gal shift, since APA can affect the stability of transcripts and therefore the expression of genes (Cheng et al. 2020). Genes that undergo APA and gene expression changes are of special interest as the gene expression change might result from their APA shift, giving a mechanism for gene expression control. Comparing

both the APA and gene expression changes of genes in the YMC and Glu-Gal shift can give us further insights into the mechanisms behind these shifts in pas site usages.

Significant changes in gene expression were observed during the glucose to galactose metabolic shift. To fully investigate the gene expression profile of the Glu-Gal shift, the data set includes the genes that contain only one pas site and are unable to undergo APA events. Similar to the YMC, roughly 1/3 of all yeast genes underwent gene expression changes. Of the 7050 genes sequenced, 1240 genes underwent GE increases (17.8%), and 950 underwent GE decreases (24.2%) and only 8 genes underwent both increases and decreases in gene expression. A significant number of gene expression increases and decreases occurred at each of the time points. The number of genes that undergo gene expression increases and decreases are shown in the bar chart 4.2 on page 219. The overall trends of changes in gene expression throughout the Glu-Gal shift is shown in the heat map 4.1 on page 218. The highest frequency of gene expression changes occurred at data point gal15. This time point marked 15 minutes post glucose starvation and the beginning of the cells shift to galactose metabolism. The most transcriptionally active time point was gal15. The Glu-Gal APA events were not skewed towards the most highly expressed genes. This is shown in the scatter plot 3.14 of the APA logfold changes vs the mean expression of the genes, on page 140. all the significant changes in gene expression at this time point are shown in the scatter plot in 4.3 on page 220. The changes in gene expression during each time point are summarised on table 4.1 on page 222. The data tables for gene expression events in the Glu-Gal are shown in the supplementary

figures under Gene expression changes for Glu-Gal.

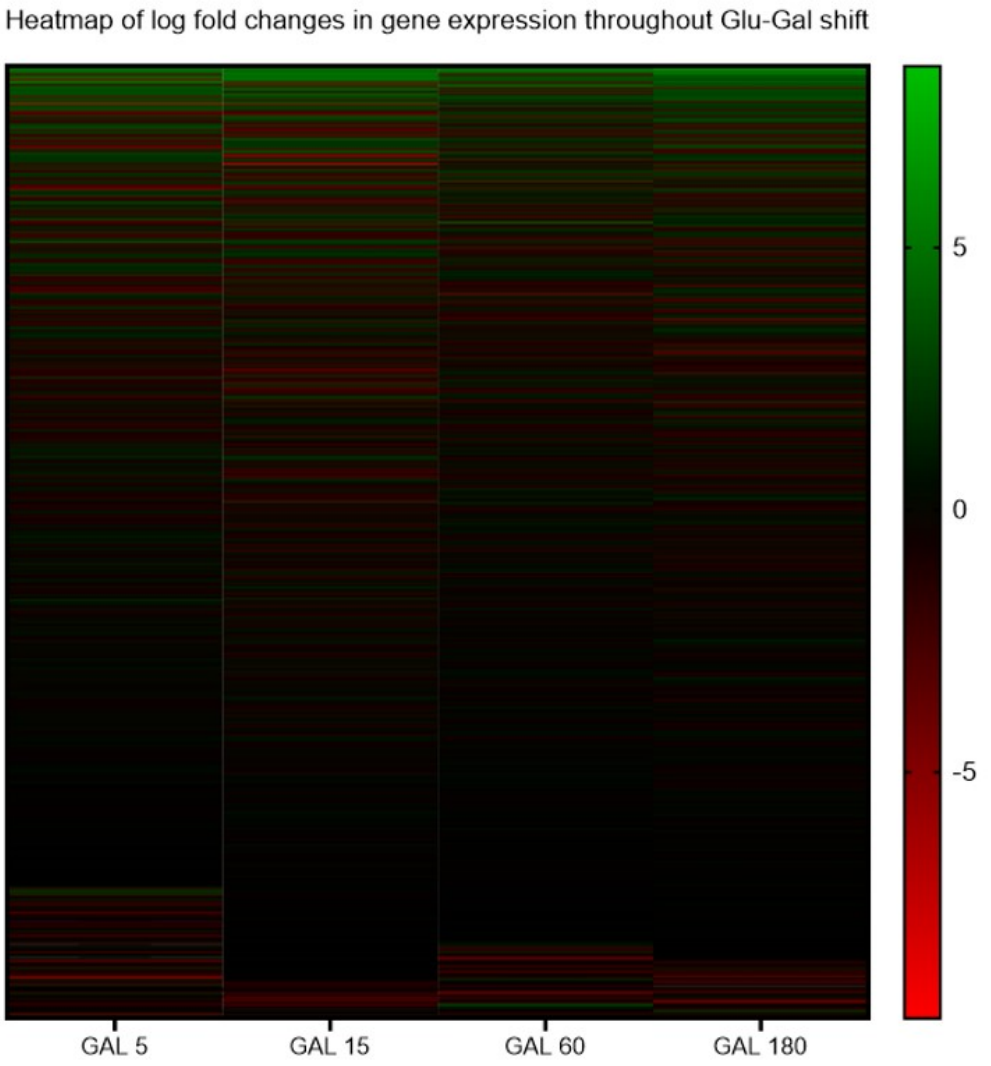


Figure 4.1: Heat map of logfold2 changes in normalized reads per million of genes for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the glucose to galactose metabolic shift. Scale shows gene expression increases are shown in green, decreases are shown in red. Black space show genes that do not significantly change in gene expression throughout the YMC.

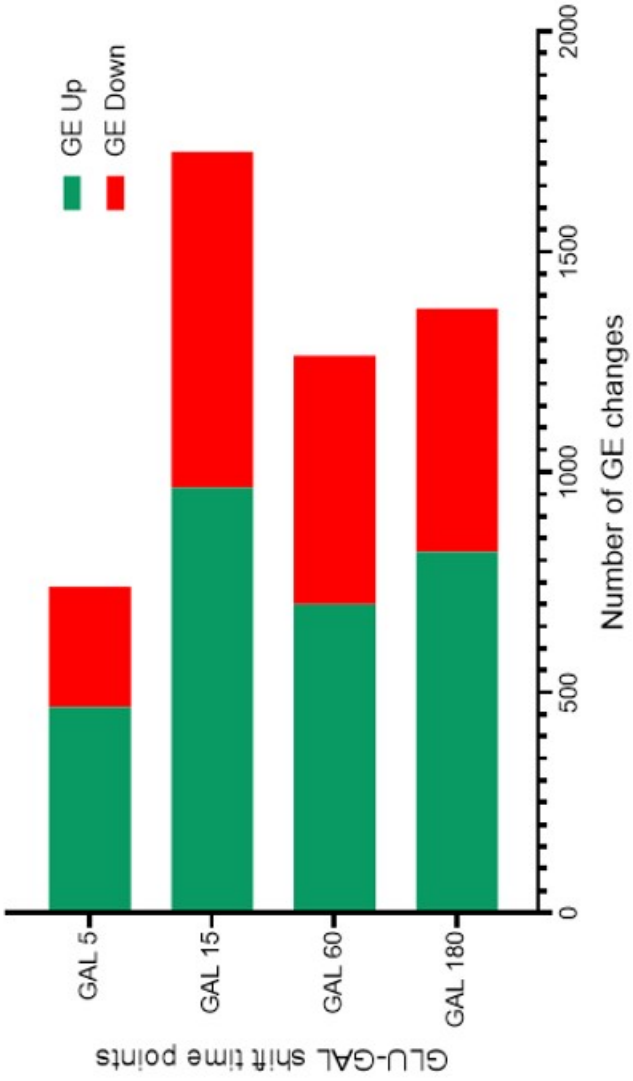


Figure 4.2: Bar chart comparing gene expression increases and decreases for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the glucose to galactose metabolic shift, green bars show for gene expression increases and red is the number of genes that undergo gene expression decreases.

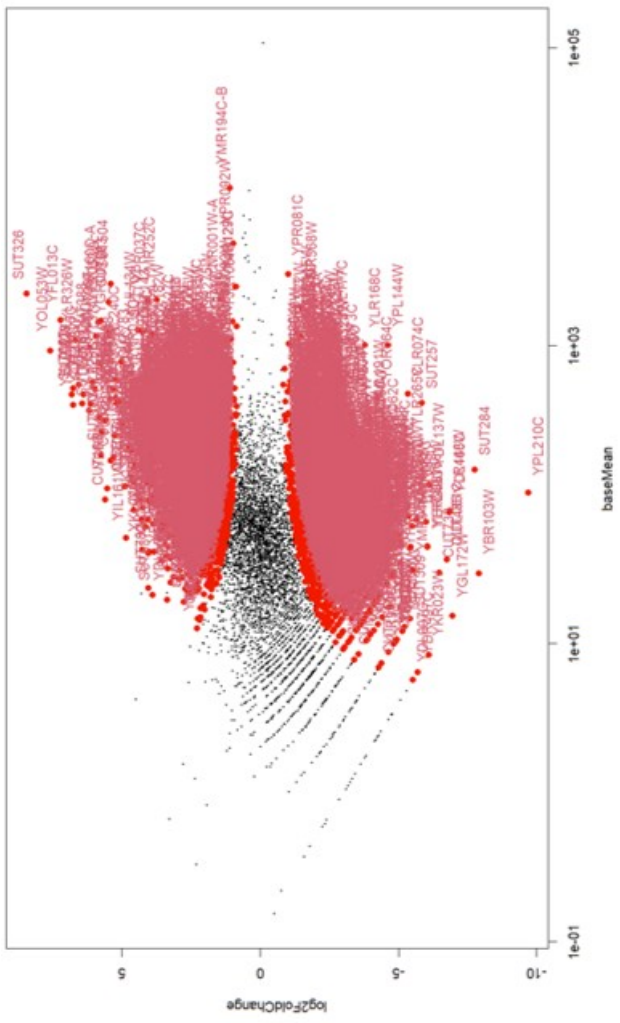


Figure 4.3: Scatter plot of log fold change values of gene expression changes of genes for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the glucose to galactose metabolic shift. The plot shows logfold change in normalised reads per million of each gene vs the number of normalized reads. Significant values of gene expression change were identified as differences in normalised reads per million compared to the control that had a padj of less than 0.05. Significant changes are shown in red whilst nonsignificant values are shown in black.

The lowest amount of gene expression changes occurred at time point Gal5, which was the first reading taken after glucose starvation. Overall, there was a modest but significant of genes increasing in expression than decreases throughout all of the recorded time points. The spread of gene expression changes varied along the YMC. The time point with the highest range of changes was at Gal15, which was also the most transcriptionically active. Timepoint Gal15 had a data range of 15.64 with the maximum log fold change in the expression being 8.083 and the minimum being -7.56 with a standard deviation of 1.62. The time point with the lowest range of gene expression changes and standard deviation was time point Gal60, which had a data range of 14.03 with the maximum log fold change in expression being 7.681 and minimum being -6.35 with a standard deviation of 1.35. The spread of the changes in gene expression was visualised on a violin plot. These plots combine box and kernel density plots to visualize the distribution of log fold changes in gene expression, as shown in 3.2 on page 122. The summary of statistics for gene expression is shown in table 6.1 on page 281. The Skew kurtosis was highest for time point GAL60 which was 2.086 The lowest skew was at time point GAL5, which was 1.285. This skew was also demonstrated in the violin plots as shown in 4.4 on page 223. The tables for the gene expression events are shown in the supplementary figures under Gene expression tables for Glu-Gal.

These results conclude that the most transcriptionally active time point during the Glu-Gal shift is at Gal15. Gal15 also had the least amount of outliers and skew, which indicates the shifts in gene expression are due to a large number of genes all undergoing significant changes, therefore indicating

this time point is very transcriptionally active for the whole genome. These results are useful for the APA analysis as APA is tightly linked to changes in gene expression and allows us to identify genes that undergo both APA and gene expression changes along the Glu-Gal shift.

Time point	Gene expression increase	Gene expression decrease
GAL 5	466	273
GAL 15	964	762
GAL 60	699	564
GAL 180	818	552

Table 4.1: Summary table of number of genes that have changes in gene expression at all time points throughout the glucose to galactose metabolic shift.

4.3 APA changes observed in the Glu-Gal shift show skew for shortening events and peak at time point Gal15

3'UTR, Coding and ALL APA events for the Glu-Gal shift time points were analyzed using the deseq/dexseq pipeline using the same parameters for the YMC. This was done to compare the APA events between Glu-Gal and the YMC to see the depth of overlap and therefore how common or specific these APA events are.

The APA data sets had very few overlaps in transcripts, indicating these changes were specific to a type of APA. There were significant changes in APA occurring at each time point throughout the Glu-Gal shift in the coding region, 3'UTR and ALL. The most APA changes occurred at time point

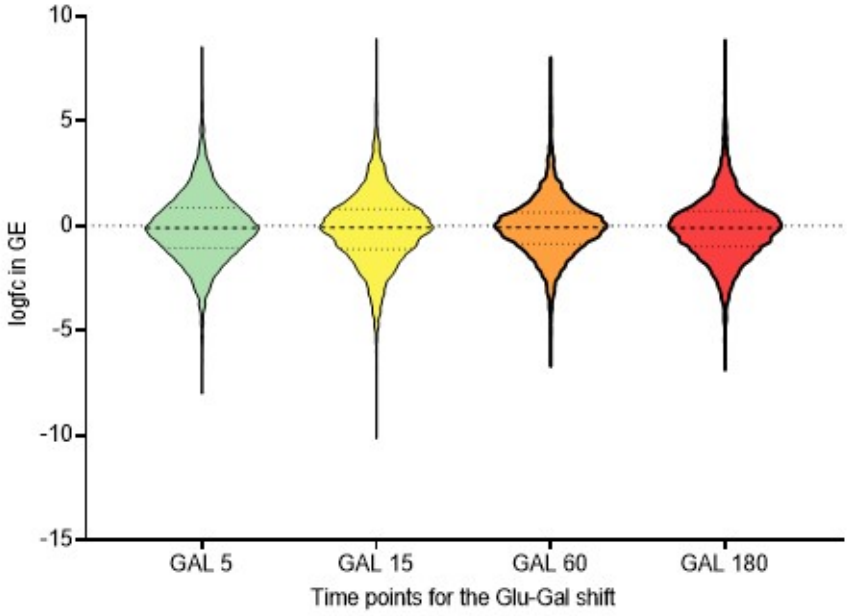


Figure 4.4: Violin plot of log fold change values of gene expression for all the 7050 genes in *Saccharomyces cerevisiae* at each time point throughout the glucose to galactose metabolic shift. Logfc is the logfold2 change in read counts used to identify the gene expression change between the time point and the controls. Plots show the distribution in logfold2 change expression data to visualize the distribution and skew of the data at each time point. Expression changes show skew of gene expression values of each time point throughout the Glu-Gal shift

GAL60, which was taken 60 minutes after the start of glucose starvation, for every type of APA. Shortening and lengthening APA events occurred more frequently for the data sets 3'UTR and coding. For the ALL data set, there was much more APA lengthening occurring than shortening at all time points. The most frequently used type of APA throughout the Glu-Gal shift was ALL APA lengthening, which had a total of 295 different genes undergoing lengthening. The least used type of APA was 3'UTR APA shortening. There was only 22 genes undergoing 3'UTR shortening throughout the Glu-Gal shift. The summary of all APA data sets throughout the Glu-Gal shift is shown in 4.5 on page 225.

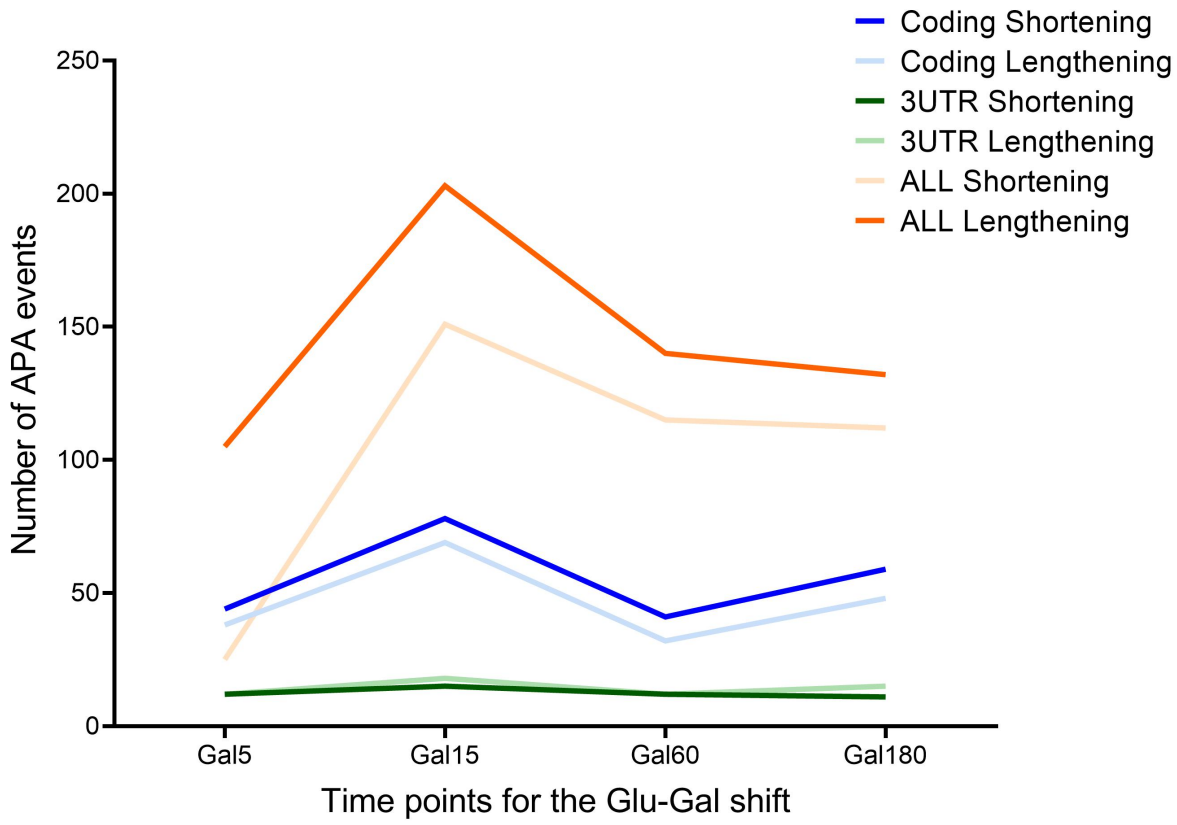


Figure 4.5: Line plot showing the number of Coding, 3'UTR and ALL APA events at each time point along the Glu-Gal shift. Shortening and Lengthening APA events have been separated to identify trends in both individually. The graph shows the peak for all significant APA events occur at time point Gal15 which is 15 minutes after the start of glucose starvation, with the biggest data set being ALL APA lengthening.

4.4 The Coding APA profile of the Glu-Gal Shift shows APA events peak at time point Gal15

Coding APA events were categorised as changes in pas site usage that both occur in the coding region of a gene. The pipe line output showed several Coding APA events occurred at each time point throughout the Glu-Gal shift. Out of 1481 transcripts that were identified to have multiple pas sites in the coding region, 118 underwent shortening, 101 underwent lengthening events, as shown in the table 4.2 on page 227. There was a small preference for APA shortening at each time point. The time point with the most APA events occurring was at GAL15, in which 78 transcripts underwent APA shortening and 69 lengthened as shown in the bar chart in 4.6 on page 227. The preference for APA changes occurring for all genes along the Glu-Gal shift is shown in the heat map 4.7 on page 228. The lowest APA activity was APA lengthening at time point GAL5, with 44 transcripts undergoing coding APA lengthening. The significant changes in gene expression at GAL15, the time point with the most APA activity, are shown in the scatter plot in 4.8 on page 229. The coding APA events for Glu-Gal were not skewed towards the most highly expressed genes.

Coding APA changes		
Time point	APA Shortening	APA Lengthening
GAL 5	44	38
GAL 15	78	69
GAL 60	41	32
GAL 180	59	48

Table 4.2: Summary table of number of genes that undergo Coding APA shortening and Lengthening at different time points throughout the glucose to galactose metabolic shift.

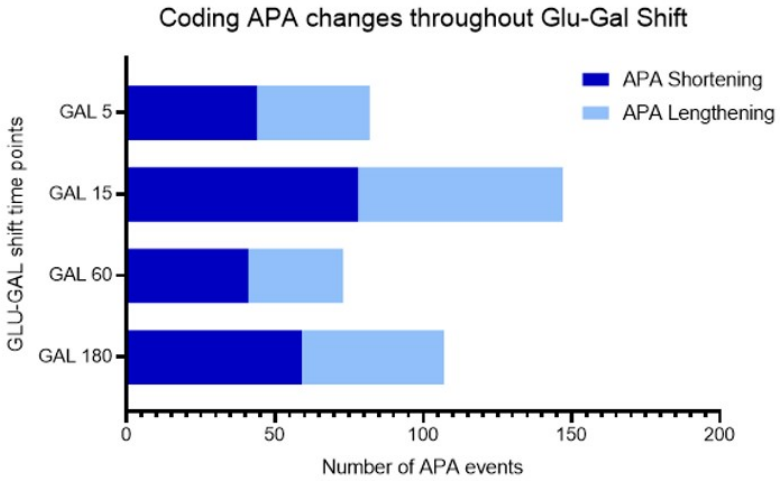


Figure 4.6: Bar chart showing the distribution of APA coding changes throughout the Glu-Gal shift. Chart shows peak of Coding APA events occurring at time GAL15 with a small preference for APA lengthening events at all time points.

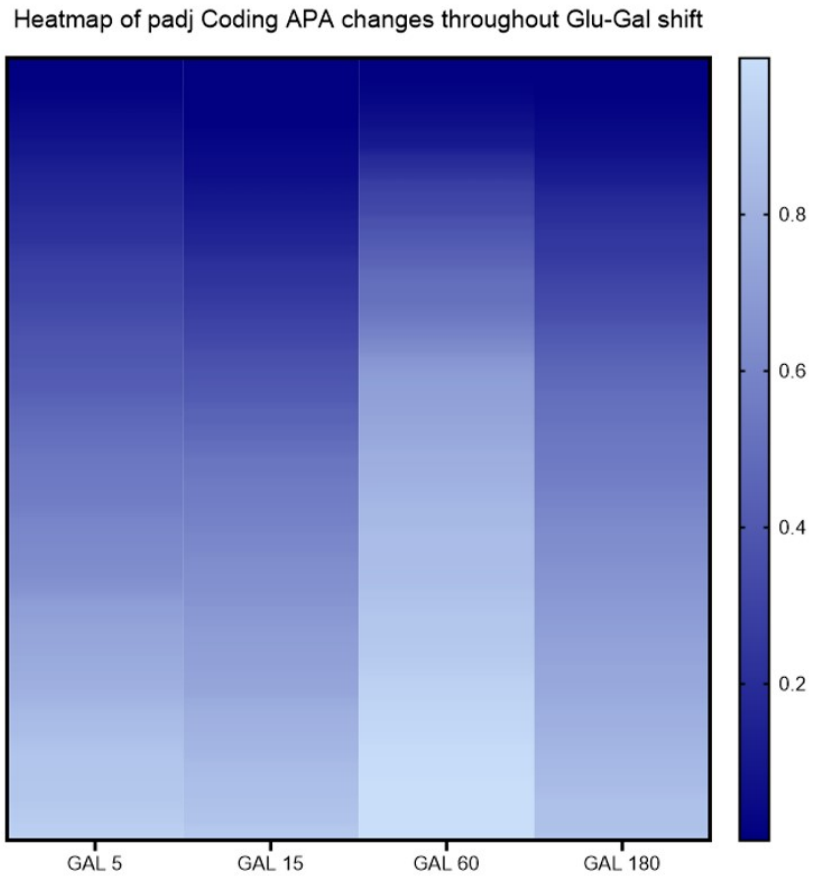


Figure 4.7: Heat map of logfold2 changes in normalized reads per million of genes that undergo coding APA throughout the Glu-Gal shift, significant APA identified as genes with less than 0.05 padj values. Heatmap shows peak of coding APA changes occur at time point GAL15. Only genes that undergo significant APA events were included in the heatmap.

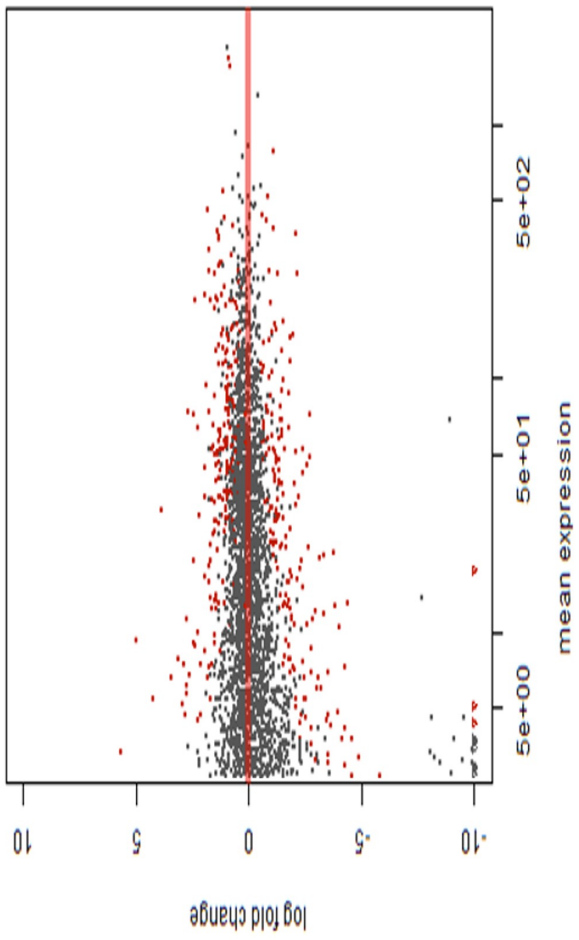


Figure 4.8: Log fold change in gene expression across Coding APA events vs the overall mean expression of the transcript, taken from time point Gal15 in the Glu-Gal shift.

4.5 The 3'UTR APA profile of the Glu-Gal Shift shows few APA events occur in the 3'UTR

The 3'UTR data shows that APA was not very prevalent in the 3'UTR for the Glu-Gal shift but there is still a clear peak at GAL15, indicating this time point is the most active for APA globally. 3'UTR APA events were categorised as changes in pas site usage that occur in the non-coding 3'UTR region of a gene. The pipeline output showed several 3'UTR APA events occurred at each time point recorded for Glu-Gal shift. Out of 1976 genes, in total, 22 transcripts underwent shortening and 27 underwent lengthening events throughout the YMC, as summarised on table 4.3 on page 231. There were more lengthening APA events at each time point throughout the Glu-Gal shift bar GAL5 and GAL60 which had equal amounts. The time point with the most APA events occurring was at GAL15, in which 18 transcripts underwent APA shortening and 15 lengthened as shown in the bar chart in 4.9 on page 231. The heat map 4.10 on page 232 shows the p-values of the changes in APA throughout the Glu-Gal shift. The lowest time point of APA activity was at time point GAL180, with 11 transcripts shortening and 15 lengthening. The significant changes in gene expression at GAL15, the time point with the most APA activity, are shown in the scatter plot in 4.14 on page 237. The tables for the 3'UTR APA events are shown in the supplementary figures under 3UTR APA tables for Glu-Gal.

UTR APA changes		
Time point	APA Shortening	APA Lengthening
GAL 5	12	12
GAL 15	15	18
GAL 60	12	12
GAL 180	11	15

Table 4.3: Summary table of number of genes that undergo 3'UTR APA shortening and Lengthening at different time points throughout the glucose to galactose metabolic shift.

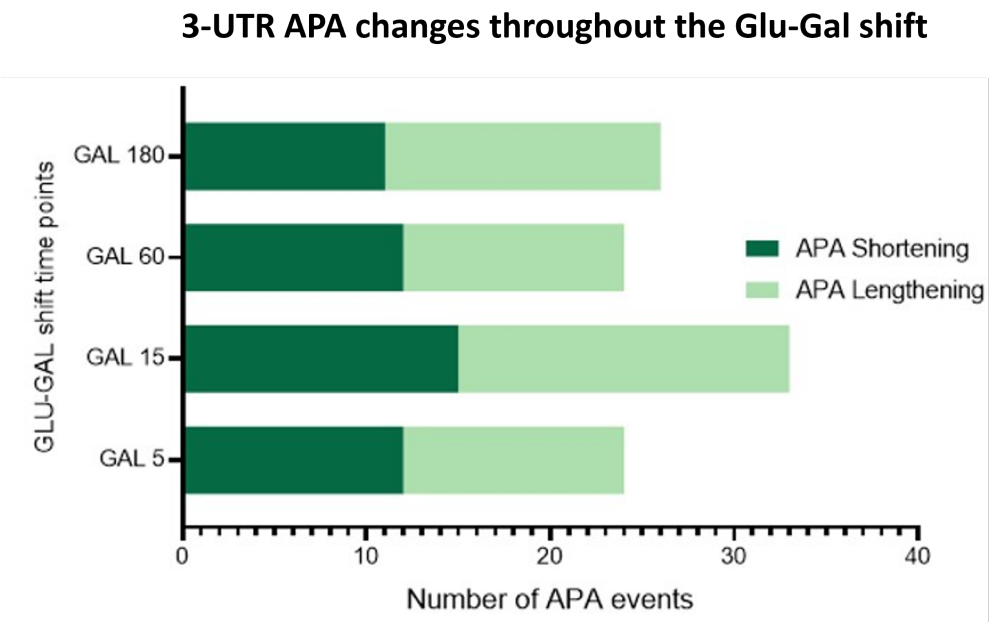


Figure 4.9: Bar chart for 3'UTR APA changes throughout the Glu-Gal shift. Dark green is APA shortening events, light green is APA lengthening.

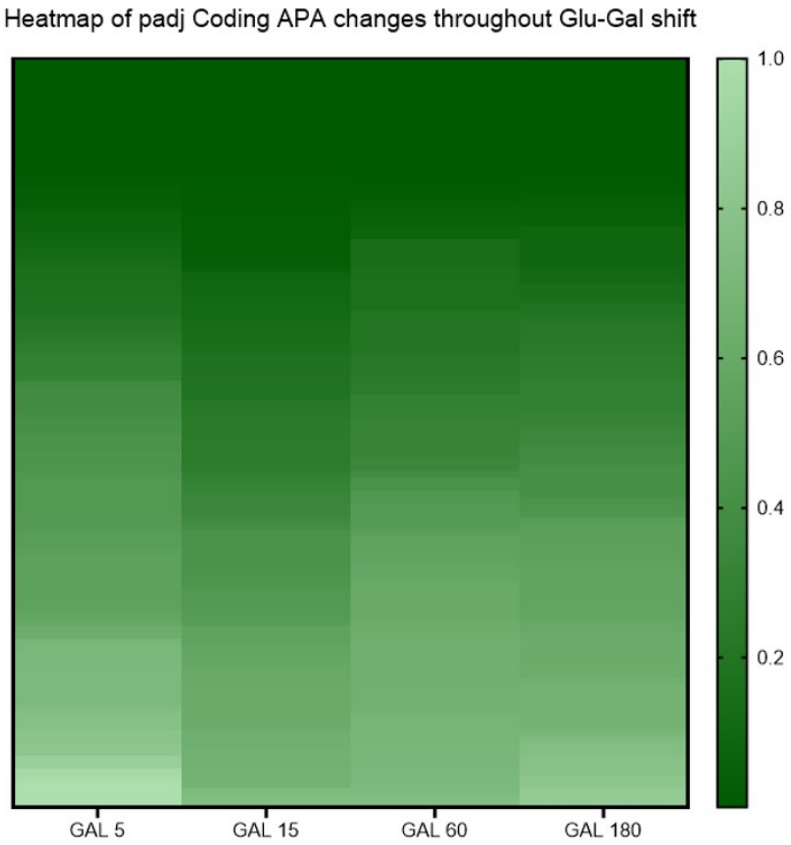


Figure 4.10: Heat map of the padj values for 3'UTR APA changes throughout the Glu-Gal shift. Darker colours represent smaller p-value and, therefore more statistically significant APA events.

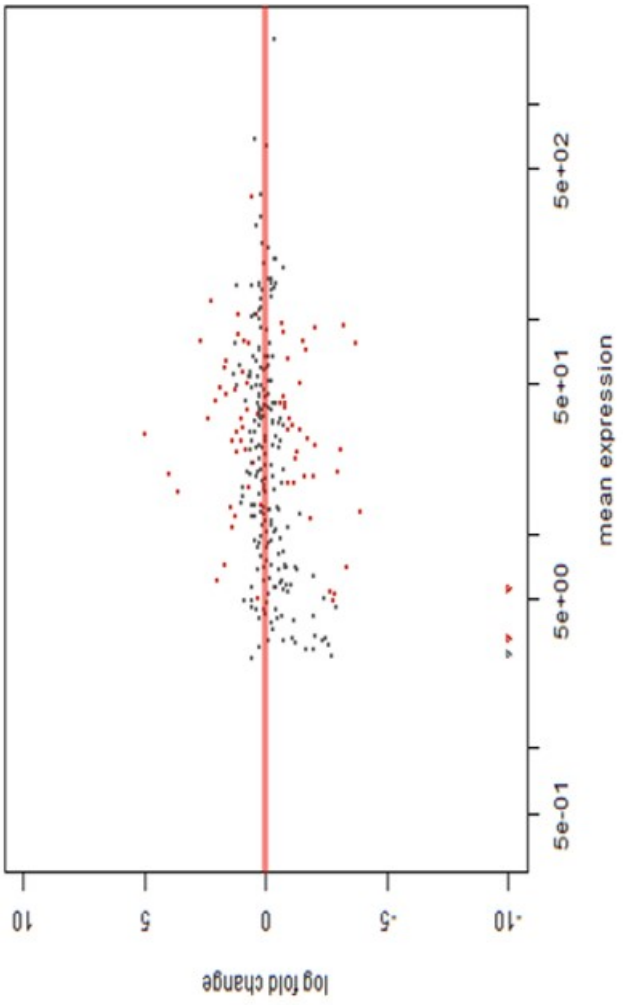


Figure 4.11: log fold change in gene expression across 3'UTR APA events vs the overall mean expression of the transcript, taken from time point Gal15 in the Glu-Gal shift.

4.5.1 The ALL APA profile of the Glu-Gal Shift shows a preference for APA lengthening events

ALL APA events were categorised as changes in pas site usage from the 3'UTR to the coding region or vice versa or any APA switch event that occurred between the two most transcriptionally active pas for that gene that had a padj value below 0.05. The pipeline output showed multiple APA events occurred at each time point along the Glu-Gal shift. Out of 4869 genes that had multiple pas sites in the coding and 3'UTR regions and therefore were capable of undergoing ALL APA, 221 underwent shortening, 294 underwent lengthening events throughout the Glu-Gal shift, as summarised in table 4.4 on page 235. There were significantly more lengthening than shortening events occurring at all time points, indicating a large preference for APA lengthening. The time point with the most APA events occurring was at Glu15, in which 151 transcripts underwent APA shortening and 203 lengthened as shown in the bar chart in 4.12 on page 235. The preference for APA changes occurring at time points 3a and 3b can be seen in the heat map 3.13 on page 139.the lowest APA activity was at 5. The tables for the ALL APA events are shown in the supplementary figures under ALL APA tables for Glu-Gal.

ALL APA changes		
Time point	APA Shortening	APA Lengthening
GAL 5	25	105
GAL 15	151	203
GAL 60	115	140
GAL 180	112	132

Table 4.4: Summary table of number of genes that undergo ALL APA shortening and Lengthening at different time points throughout the glucose to galactose metabolic shift.

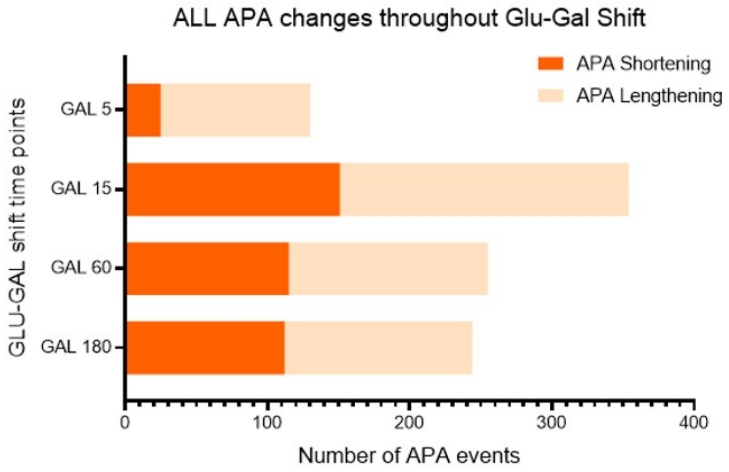


Figure 4.12: Bar chart values for ALL APA changes throughout the Glu-Gal shift. Light Orange indicates APA lengthening events, dark orange is APA shortening.

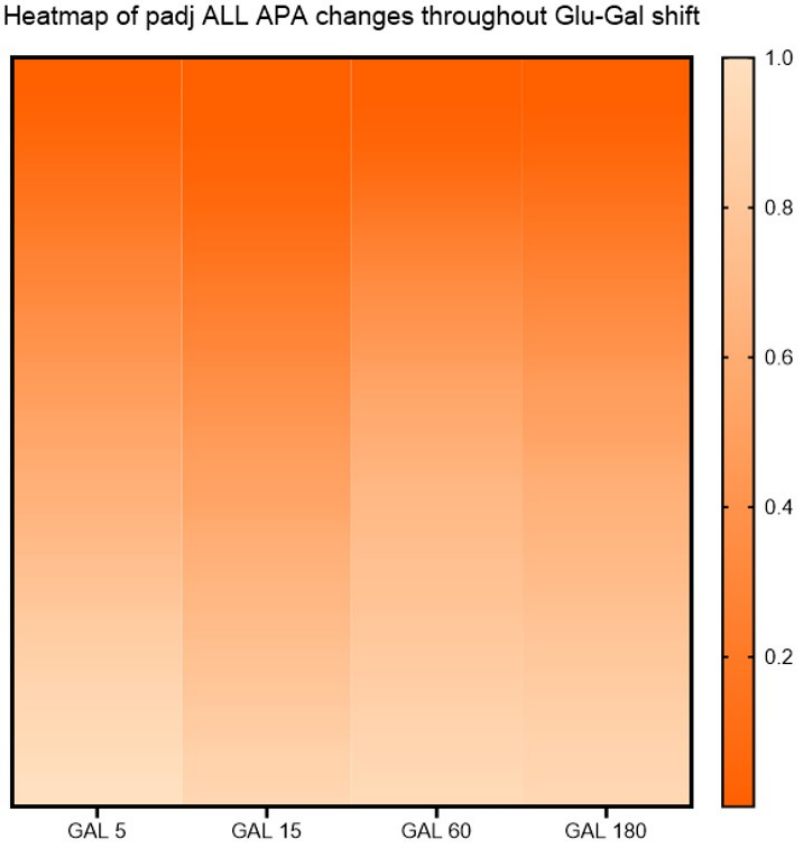


Figure 4.13: Heatmap of the PADJ values for ALL APA changes throughout the Glu-Gal shift. Darker colours represent smaller p-value, therefore more statistically significant APA events.

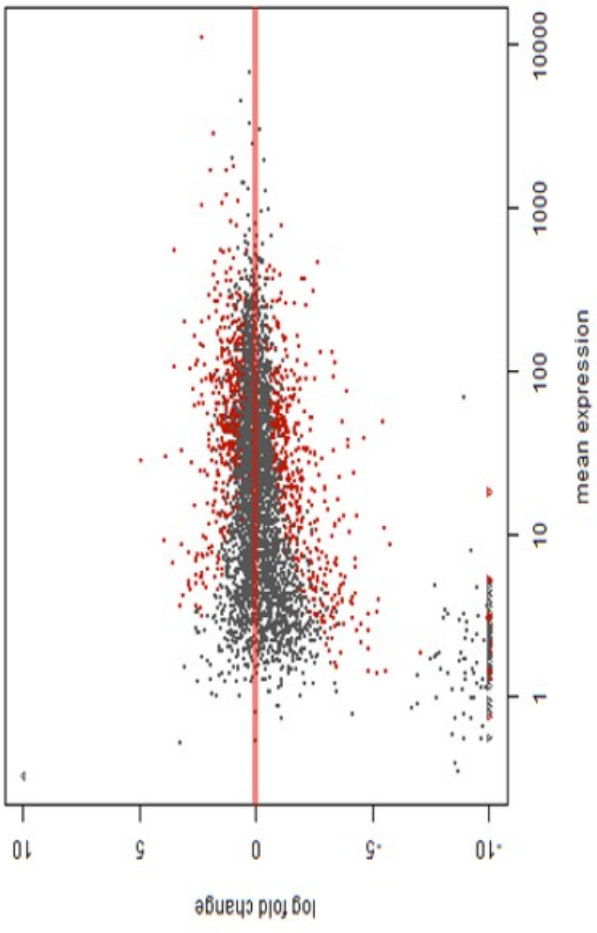


Figure 4.14: Log fold change in gene expression across ALL APA events vs the overall mean expression of the transcript, taken from time point Gal15 in the Glu-Gal shift.

To conclude, the APA profiling of the Glu-Gal shift, several key findings were discovered in the exploration of these data sets. These data shows that the most APA active time point of the Glu-Gal shift was Gal15. The data also showed that APA events skewed towards shortening in the coding region but lengthening in the ALL data set. As with the YMC, the scatter plots show no is no skew towards APA events occurring in genes with larger gene expression changes. This indicates that these APA events are not just an artifact of highly expressed genes undergoing more variable use of pas sites. The APA data sets also showed that the most active gene expression time point matches the most prolific APA time points, indicating that APA might be influencing the gene expression profile in the Glu-Gal shift as with the YMC. Furthermore, Gal15 is a time point associated with the induction of the GAL genes, which are key regulators of the Glu-Gal shift further implicating APA as a potential contributor to these key processes via manipulation of gene expression of a substantial number of genes.

4.5.2 APA Vs Gene expression changes for Glu-Gal shift

Many genes undergo both APA and GE events throughout the Glu-Gal shift. The intersection of these two events was identified. Overall, genes that undergo APA events are more likely to undergo gene expression changes, compared to genes that do not undergo any APA changes throughout the Glu-Gal shift. The data sets of genes undergoing 3'UTR, CR-APA, or ALL APA events along the Glu-Gal shift were analysed for changes in gene expression. It is noteworthy that APA has been associated with alterations in transcript stability, which can exert a direct influence on gene expression dynamics, as established by prior research (Yuan et al. 2021). The analysis of the data showed a small significant predilection for genes experiencing both APA events and exhibiting differential Gene Expression. This inclination persists across all categories of APA events under investigation. The summary table for these data sets is shown in 4.5 on page **??**. The T-tests conducted on the bootstrapped data sets showed some statistical correlations between APA changes and GE changes, indicating there might be a small correlation between genes that undergo APA shifts in the Glu-Gal shift and genes that have significant changes in gene expression. The table of results for these T-tests is shown on table 4.6 on page 240.

Coding APA Vs Gene expression changes for Glu-Gal shift

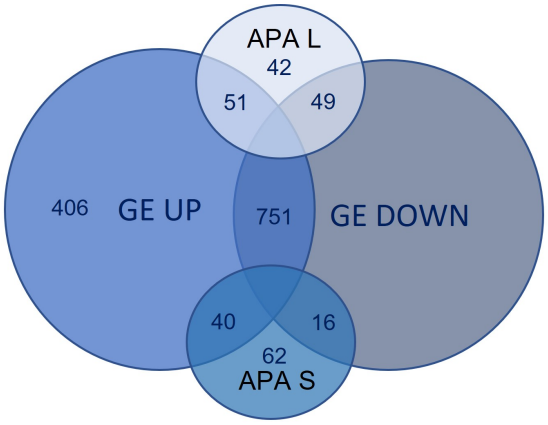
For the Coding APA data sets, there was no statistical correlation between gene expression changes and APA lengthening and APA shortening. Genes that underwent coding APA lengthening had higher percentages of

APA event	DE	% DE	No DE	% No DE
Coding Shortening	85	5.74	78	5.27
Coding Lengthening	100	6.75	69	4.66
3'UTR Shortening	15	1.11	22	1.11
3'UTR Lengthening	17	1.36	27	1.37
ALL shortening	165	3.39	151	3.11
ALL Lengthening	244	5.01	203	4.17

Table 4.5: Table of APA genes throughout the Glu-Gal shift showing a small preference of genes undergoing both APA and Differential gene expression events. % of APA events occur in genes that undergo differential expression and those that do not undergo differential expression. Table of the number of APA events that occur in both genes with differential expression changes throughout the Glu-Gal shift and genes that do not change expression. The table shows DE = Differential expression, % DE is the % of genes that undergo APA events vs the total number of genes in that data set, total is the total number of APA events.

Time points	Coding S	Coding L	3'UTR S	3'UTR L	ALL S	ALL L
GAL5	0.1278	0.6527	0.4069	0.8327	0.2189	0.2668
GAL15	0.999	0.4897	0.2512	0.8769	0.9897	0.0005
GAL60	0.4507	0.5348	0.9268	0.7864	0.6802	0.0003
GAL180	0.6347	0.603	0.5962	0.2824	0.107	0.0249

Table 4.6: p-values of t-tests of genes that undergo different types of APA and changes in gene expression at different time points throughout the Glu-Gal shift. Numbers highlighted in red are statistically significant with a p-value lower than 0.05



APA TYPE	Coding Up		Coding Down	
Shortening	40	39%	45	38.1%
Lengthening	51	49.5%	49	48.5%

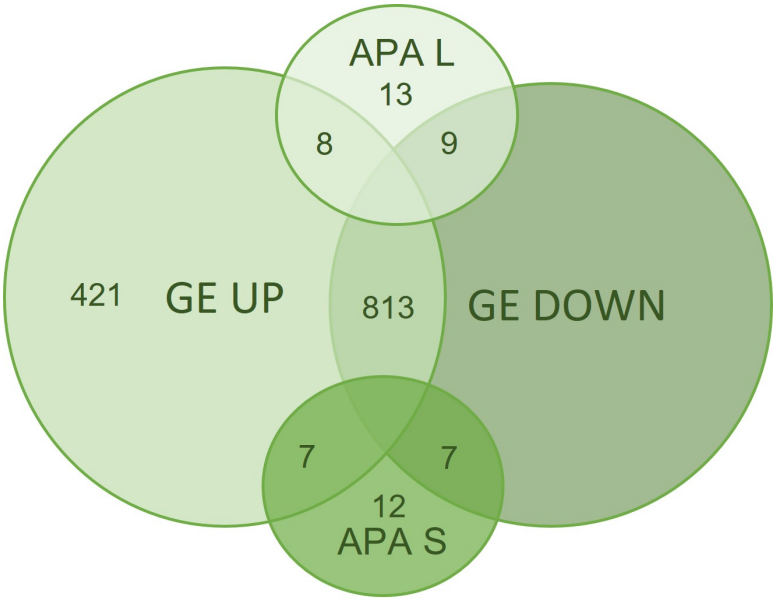
Figure 4.15: Venn diagram of expression data from all the points of the YMC showing Coding APA changes vs gene expression changes throughout the Glu-Gal shift and a table which shows the number and % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases.

gene expression increases and decreases, with gene expression increases and APA lengthening being the biggest overlapping group with 49.5% of genes undergoing APA lengthening also undergoing increases in gene expression. The Coding APA and GE changes are summarised in the diagram 4.15 on page 241.

3'UTR APA Vs Gene expression changes for Glu-Gal shift

The 3'UTR APA data sets had almost equal amounts of genes undergoing shortening and lengthening events and the genes that underwent gene expression increases/decreases. No observable preference or statistical correlation was found between APA and gene expression events. The APA and

gene expression changes are summarised in the diagram 4.16 on page 243.

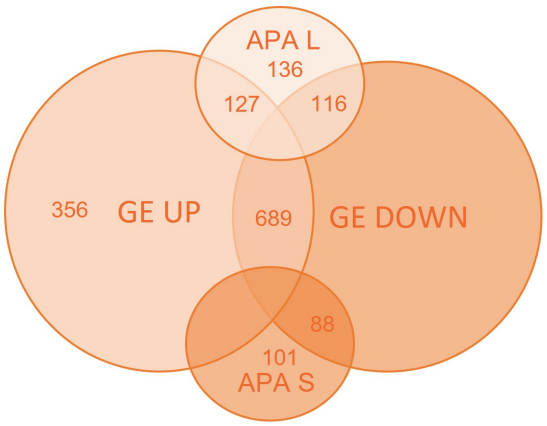


APA TYPE	3UTR Up		3UTR Down	
Shortening	7	31.8%	8	36.4%
Lengthening	8	29.6%	9	33.3%

Figure 4.16: Venn diagram of expression data from all the points of the YMC of total UTR APA changes vs gene expression changes throughout the Glu-Gal shift and a table which shows the number and % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases.

ALL APA Vs Gene expression changes for Glu-Gal shift

For the ALL APA data set, there was a small significant correlation between APA changes and gene expression changes. For the coding and ALL APA data sets there was a trend for the genes that underwent lengthening to have increases in gene expression versus genes that didn't undergo APA or underwent APA shortening events. This indicated there was a small but statistically significant correlation between APA lengthening and gene expression increases. For transcripts that undergo ALL shortening APA, 13.7% of the transcripts also underwent gene expression increases and 42.5% undergo gene expression decreases. For transcripts that undergo coding lengthening APA, 40.8% of the transcripts also underwent gene expression increases and 22.3% undergo gene expression decreases. The APA and GE changes are summarised in the diagram 4.17 on page 245.



APA TYPE	ALL Up		ALL Down	
Shortening	77	34.8%	88	39.8%
Lengthening	128	43.5%	116	39.5%

Figure 4.17: Venn diagram of expression data from all the points of the YMC showing total ALL APA changes vs gene expression changes throughout the Glu-Gal shift and a table which shows the number and % of transcripts that undergo APA shortening or lengthening and gene expression increases and decreases.

4.6 Gene Ontology analysis for genes that undergo APA changes during Glu-Gal shift

GO analysis was performed on the Glu-Gal shift data sets using the same tools and parameters as done for the YMC data. The GO analysis showed that genes that undergo APA in the Glu-Gal shift are enriched in GO terms for cell cycle and DNA maintenance and repair. GO enrichment was found for all time points in the Glu-Gal shift data and for all of the types of APA events.

4.6.1 Gene Ontology analysis for genes that undergo Coding APA changes for Glu-Gal shift

The data sets for each individual time point for GO analysis for APA coding were too small to gain meaningful and statistically significant results. Therefore, the results were identified for the total genes in the data set and not for specific time points. The GO analysis for genes that underwent coding APA shortening showed enrichment in GO terms that included cell development and organophosphate metabolic process. Coding APA lengthening also showed GO enrichment. The functional GO categories included cellular chemical homeostasis and DNA genomic change. The summary table showing all GO analysis relating to coding APA shortening and lengthening is shown in the supplementary data table 6.15 on page 299.

4.6.2 Gene Ontology analysis for genes that undergo 3'UTR APA changes for Glu-Gal shift

Due to small data sets for each individual time point in GO analysis for 3'UTR APA, meaningful and statistically significant results couldn't be obtained. Thus, the results were identified for the total genes in the data set rather than specific time points. The GO analysis for 3'UTR shortening showed enrichment in GO terms that included cell cycle, malate metabolic process, and protein deglutathionylation. 3'UTR APA lengthening also showed GO enrichment in functional categories that included mRNA processing and proline methylation. The summary table showing all GO analysis for all time points relating to 3'UTR APA shortening and lengthening is shown in the supplementary figure table 6.16 on page 301.

4.6.3 Gene Ontology analysis for genes that undergo ALL APA changes throughout the Glu-Gal shift

The GO analysis for ALL shortening was significant at all the time points recorded. This showed enrichment in GO terms that included DNA repair and trans-membrane transport. The enrichment functional categories included regulation of cell cycle phase transition, negative regulation of nuclear division, and negative regulation of cell growth. The summary table showing all GO analysis for all time points relating to ALL APA shortening and lengthening is shown in 4.18 on page 248.

ALL Shortening	GO Term	Genes	P value
5	double-strand break repair via synthesis-dependent strand annealing [GO:0045003]	YIL132C,YMR224C	0.001348
	inosine salvage [GO:0006190]	YOR155C	0.00349
15	membrane lipid biosynthetic process [GO:0046467]	YCR017C YDR072C YJL062W YKR053C YNL183C	6.812754e-4
	protein refolding [GO:0042026]	YDR258C YLR090W YLR216C YNL077W	0.003384
60	transmembrane transport [GO:0055085]	YBR008C,YDL199C,YDR011W,YDR322C-A,YDR414C,YGL080W,YIL111W,YKR093W,YLR090W,YLR220W,YNL142W,YNL183C,YNR002C,YOR071C,YPL189W,YPR020W	0.001595
	protein modification process [GO:0036211]	0.004056	
180	transmembrane transport [GO:0055085]		1.97E-04
ALL Lengthening	GO Term	Genes	P value
5	peptidyl-amino acid modification [GO:0018193]	YDR448W,YGL249W,YHR109W,YJL127C,YLL002W,YLR137W,YLR153C,YOR156C,YOR213C,YOR355W	8.99E-05
	protein acetylation [GO:0006473]	YDR448W,YJL127C,YLL002W,YLR153C,YOR213C,YOR355W	6.12E-04
15	cellular hyperosmotic salinity response [GO:0071475]		7.84E-04
	negative regulation of cell growth [GO:0030308]	YBR066C,YDR043C,YNL068C	0.001614
60	regulation of translational termination [GO:0006449]	YDL207W,YER176W,YPL226W	0.001245
	U4 snRNA 3'-end processing [GO:0034475]	YGR158C,YJL050W,YNL299W	
180	positive regulation of cell cycle phase transition [GO:1901989]	YGL116W,YGL190C,YMR273C,YNL201C,YOR373W	0.0007741
	negative regulation of nuclear division [GO:0051784]	YGL116W,YGL190C,YML074C,YNL201C,YPL018W	0.001792

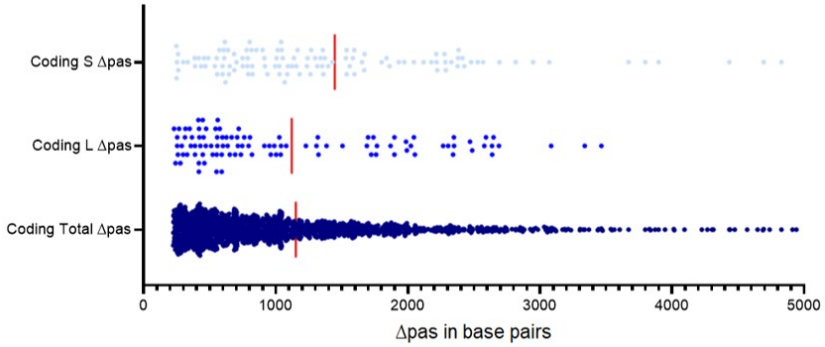
Figure 4.18: Summary table of GO analysis results for genes that undergo ALL APA changes during Glu-Gal shift. Significance in enrichment for a functional group was shown to be p-values less than 0.05.

4.6.4 Genes undergoing shortening APA events throughout Glu-Gal shift have increased Δpas lengths

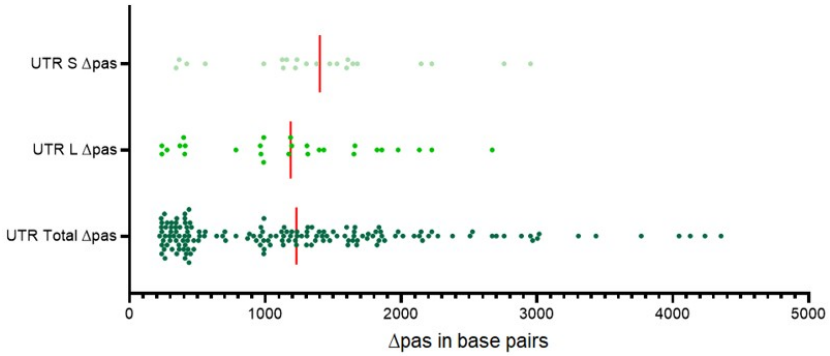
Δpas length is the length in base pairs between two alternative pas sites and therefore the stretch of sequence that is included or removed from a nascent mRNA transcript if APA occurs. All APA shortening types had a Δpas length longer than genes that had multiple pas sites but did not undergo APA. This suggests that the APA events are not random or due to error as the longer Δpas distances seem to be preferable and selected over the shorter, instead of a random selection of lengths being removed or included due to APA shifts. The distribution of Δpas lengths for all the APA events can be seen in 4.19 on page 250.

APA Type	p-value	Mean Δpas APA	Mean Δpas Total
Coding Shortening	0.001	1446	1151
Coding Lengthening	0.75	1121	1151
3'UTR Shortening	0.404	1402	1228
3'UTR Lengthening	0.823	1186	1228
ALL Shortening	0.0015	1513	1268
ALL Lengthening	0.893	1259	1268

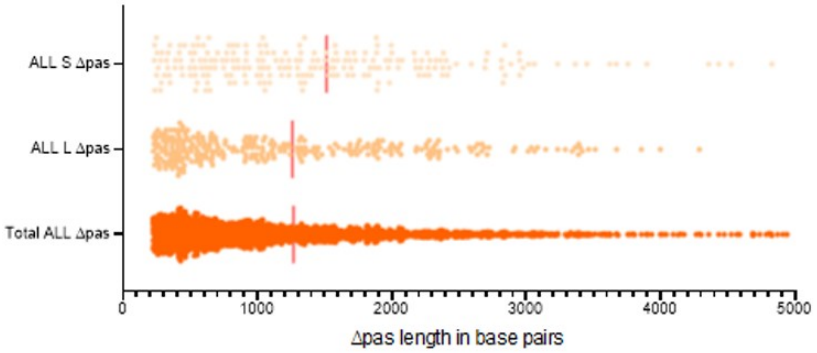
Table 4.7: Table of mean Δpas lengths of genes that undergo APA during the Glu-Gal shift. Table shows p-values of mean Δpas lengths for Coding, 3'UTR and ALL APA datasets compared to genes that do not undergo APA in the Glu-Gal shift but contain multiple pas sites. Coding and ALL shortening have p values less than 0.05, which is statistically significant shorter mean lengths than the control genes.



(a) Δpas lengths for coding APA events.



(b) Δpas lengths for 3'UTR APA events.



(c) Δpas lengths for ALL APA events.

Figure 4.19: Scatter plot of coding, 3'UTR and ALL APA Δpas lengths for genes that undergo APA in the Glu-Gal shift. Mean length in base pairs between alternative pas sites is shown with a red line.

4.7 Discussion

The main conclusions of this chapter showed gene expression and APA changes observed during the Glu-Gal shift in *saccharomyces cerevisiae* showed that the largest number of changes occurred at the 15 minute time point, Gal15, which is when glucose metabolism is induced. This time point appears to be a stress response to the lack of glucose, as the expression of genes involved in metabolizing galactose, known as the GAL network, does not begin until 60 minutes after induction and protein translation does not occur until 4 hours after induction. The APA events during the Glu-Gal shift showed a small preference for lengthening, which is a surprising result as similar situations of metabolic stress in yeast have shown a preference for global shortening. There was also a trend for genes undergoing lengthening APA to have increases in gene expression, while genes undergoing no APA or shortening APA showed decreases in expression. Further investigation of potential binding sequences and RNAPII elongation rates during the Glu-Gal shift may provide insight into the mechanisms behind these gene expression and APA changes.

4.7.1 Gene expression changes in the Glu-Gal shift peaks at time point Gal15

The gene expression changes seen in the Glu-Gal shift showed the largest number of gene expression changes occurred at time point glu15, which is 15 minutes after galactose metabolism is induced. This seems to fit with the literature as the removal of glucose from the environment is the trigger

for metabolic stress which can result in changes in both transcription and transnational output in yeast (Janapala, Preiss, and Shirokikh 2019). Depletion of glucose does also initiate the expression of genes that can metabolise galactose, called the GAL network, (Nguyen-Huu et al. 2015). Glucose also inhibits the expression of enzymes needed to metabolize the less preferred sugar, therefore its removal will result in an increase in expression of the genes encoding these proteins, (Venturelli et al. 2015). The GAL network is the mechanism that governs how yeast decides what to metabolise between available carbon sources. High glucose levels suppress genes involved in GAL metabolism via the protein Mig1 (Mahilkar, Saini, et al. 2022). When glucose is removed, the GAL4 protein activates a cascade of downstream GAL genes. Activation of these genes result in a net increase in gene expression levels. However, the GAL genes are only switched on after 60 minutes, with the translation of the proteins occurring after 4 hours (Brown et al. 2018). Therefore, this initial gene expression changes at gal15 are most likely to be a stress response due to lack of carbon response and not the GAL genes being expressed, but later time points might reflect this change.

4.7.2 APA occurred most often at time point Gal15 in the ALL data set

APA events for the Glu-Gal shift occurred most frequently at time point Gal15, which is 15 minutes after the induction of the Glu-Gal shift. The peak of APA activity is 15 minutes after glucose starvation. As with the gene expression data, this initially looks like it could be due to the increase demand for gene expression of genes required for galatose metabolism. However, as

the time point is too early for these genes to be switched on, it is also likely to be a stress response to glucose starvation (Brown et al. 2018).

4.7.3 APA events during Glu-Gal shift have a small preference for lengthening

There was a small preference for APA lengthening over shortening during the Glu-Gal shift, therefore showing an increase in transcripts with distal pas site usage. This was an unexpected outcome as similar situations of metabolic stress in yeast have shown global shortening is the preferred type of APA (Badjatia et al. 2021). 3' READS data has shown that under diauxic conditions, yeast favour shorter 3' isoforms and Poly(A) sites are shifted upstream (Geisberg, Moqtaderi, and Struhl 2020). The global lengthening APA events may be due in part by the retention of stabilizing elements in the longer transcript (Tian and Manley 2017), however analysis of potential binding sequences in the needs to be done to confirm this theory. Additionally, the unexpected APA shift could be a different mechanism of gene expression control being used for Glu-Gal and diauxic shift. A theory for the global APA shortening for diauxic shift is a slower elongation rate of RNAPII, which has an extremely similar APA profile to diaxuic shift. Slower elongation rates lead to preferential usage of upstream proximal pas sites, which isn't seen in the Glu-Gal shift. Therefore if the Glu-Gal shift is not causing this change in RNAPII elongation, it might not affect APA in this way. Investigating the elongation rate of RNAPII during the Glu-Gal shift would be a good next step in investigating the mechanism behind the APA profile.

4.7.4 Small correlation observed between Glu-Gal shift APA genes and changes in gene expression

The data reveals a small preference for genes that undergo APA to also undergo changes in gene expression during the Glu-Gal shift. These findings are consistent across all categories of APA events, where the percentage of genes concurrently experiencing both APA events and DE surpasses the percentage of genes with APA events but without DE. This trend suggests a discernible preference for genes to undergo both APA and DE simultaneously during the Glu-Gal shift, implying a potential interplay between these two molecular processes in the regulation of gene expression dynamics under these conditions. These findings underscore the significance of understanding the intricate relationship between APA and DE in the context of gene expression regulation during dynamic cellular transitions such as the Glu-Gal shift. For the coding and ALL APA data sets there was a trend for the genes that underwent lengthening to have increases in gene expression versus genes that did not undergo APA or underwent APA shortening events. These trends could be due to stabilising elements being retained in the longer transcript. This is a common method of increasing transcript stability. Previous studies have found yeast mRNAs contain hundreds of varying sequences affecting the stabilization or destabilization of mRNAs terminating downstream (Geisberg, Moqtaderi, Fan, et al. 2014). Further studies could be done to clarify if transcripts are being stabilised by the retention of these elements. This would entail identifying the sequences of the Δpas and cross referencing them to see if they contain shared motifs that could be targeted for stabilization. These

sequences could be investigated using online databases of stability sequences that have already been confirmed. As with the YMC data, the prevalence of genes that underwent APA and DE was higher, but the type of APA events (3'UTR, Coding) did not seem to be affected by differential expression. The outcome of this analysis was unexpected as previous research into genes that undergo APA but do not exhibit differential expression is affected in terms of their stability or localization Juneau, Nislow, and Davis 2009 as such changes do not require a variation in gene expression. Typically, localization and stabilization are known to happen in the 3'UTR of the transcript. Therefore, it is reasonable to anticipate that genes undergoing APA without differential expression would show a higher proportion of 3'UTR APA events. However, the results suggest that there might be additional reasons for APA to occur beyond what the literature suggests. Alternatively, it is possible that not all APA events cause functional changes to the transcript.

4.7.5 Gene Ontology analysis shows enrichment for functional groups involved in cell cycle progression and glucose metabolism

The GO analysis for APA genes in the Glu-Gal shift showed enrichment for functional groups involved in cell cycle progression and glucose metabolsim. These biological processes are essential for cellular function. Cell cycle progression is tightly tied to changes in environment in yeast, especially a shift in metabolism such as glucose starvation. In s.pombe yeast, transient cell cycle arrest in G2 phase is induced via limiting glucose availability (Masuda et al. 2016). During the Glu-Gal shift, the cell cycle temporally length-

ens. The lengthening of the cell cycle is grouped into three phases. The first phase is the shortest, this is when the glucose is still in the growth media. As glucose is depleted, the second phase starts, which is a more unpredictable but generally longer cell cycle phase. The final phase occurs once glucose has been fully depleted and galactose is the primary energy source, this is the longest cell cycle phase (Escalante-Chong et al. 2015). The genes that were in the GO group for cell cycle are all involved in cell cycle procession inhibition, many in response to stress. It is possible that these genes are being targeted for APA events to alter their gene expression and therefore protein output to contribute to the induction of the alternative galactose induced cell cycle. This mechanism of cell cycle control could be investigated by doing western blots of the genes that induce cell cycle inhibition that also undergo APA shifts and see if APA is causing a change in protein output. Genes involved in the metabolism of glucose were also enriched in the GO analysis. *YGL059W/Pkp2* which underwent shortening is a negative regulator of the Yeast pyruvate dehydrogenase complex (PDC) (Gey et al. 2008). The PDC mediates a major regulatory step between glycolysis and the tricarboxy acid cycle via the oxidative decarboxylation of pyruvate, committing it to the TCA cycle (Jeoung 2015). In the related yeast strain, D. bruxellensis, it was found that the PDC activity was lower in yeast that had a galactose food source than those with glucose. The APA of this regulatory factor of PDC could be facilitating this decrease in activity shown in the Glu-Gal shift. This mechanism of PDC regulation via APA of PKP2 could be explored further via knock outs of the proximal pas site for PKP2 and then measuring PDC activity levels in the mutant strains.

4.7.6 The Δpas of genes that undergo APA shortening are significantly longer than non APA genes with multiple pas sites

The Δpas lengths of the genes that undergo APA shortening changes throughout the Glu-Gal shift were significantly longer than genes that had multiple pas sites but did not undergo shifts in pas site usage. This indicates that the APA shifts observed might be influenced by having longer Δpas lengths. The Δpas lengths being significantly different to that of genes that aren't undergoing APA might be due to specific sequences are being targeted for inclusion/removal via APA changes. Cross comparing the different Δpas sequences for similarities could help identify specific sequences that are being included/excluded in the APA for Glu-Gal events.

4.7.7 Further investigations

As with the YMC APA data sets, further investigations can be done to develop these initial findings. First, qPCR can be used to confirm the APA events observed in the genes that undergo APA that are involved in essential cellular functions. This will not only confirm the results from the 3-RNA sequencing but also allow for the quantification of changes in gene expression that occur with shifts in APA. Additionally, western blot analysis can be performed to determine if the observed APA changes have an effect on the protein output of the genes. Previous research has suggested that APA can alter the protein output of transcripts, so examining this connection in the present study would be useful for understanding the function of specific APA

events. Additional analysis of the Δ pas sequence region such as motif analysis could be used to identify potential functions of APA. As with the YMC data, identifying shifts in pas site usage from proximal to distal in upstream interfering noncoding RNAs could potentially be used in the Glu-Gal shift to cause gene derepression in downstream genes. This can be investigated with the data sets already produced by this investigation and then confirmation of these APA events is confirmed using qPCR and knock out distal pas site mutants.

Chapter 5

The Comparison of APA events between the YMC and Glu-Gal shift

5.1 Chapter aims

This chapter explores the intersection of APA events in the YMC and Glu-Gal shift. Investigating how APA affects the YMC requires us to identify what APA events are specific to the YMC and what events are more generic to metabolic stress and glucose starvation. The Glu-Gal shift is a process that causes major changes in the metabolism of yeast and acute metabolic stress due to glucose starvation. By investigating and comparing the APA landscape of both the YMC and Glu-Gal shift, we can identify if the APA changes observed in these processes are specific to these biological processes or something that occurs due to generic metabolic stress responses. The

comparison of APA events between the YMC and Glu-Gal shift showed that there is very little overlap of genes that undergo APA changes in these two events. Therefore this indicates that APA is specific to both YMC and Glu-Gal shift and is unlikely to be controlled mechanistically by glucose starvation or other metabolic stresses shared by both of these cellular states.

5.2 Few APA events occur in both the YMC and Glu-Gal shift with few genes showing GO enrichment

There is limited overlap in the genes that undergo APA events in the YMC and Glu-Gal shift. The ALL APA Shortening data set had the largest overlap with 11 genes undergoing APA in both sets. The GO analysis revealed few significant functional hits for the ALL data set and Coding APA, but none for 3'UTR. The GO analysis for APA coding for each individual time point had data sets that were too small to provide meaningful and statistically significant results. Therefore, the overall results were identified for all genes in the data set, rather than for specific time points. The supplementary figures contain data tables for APA events for genes in both the Glu-Gal shift and APA, which can be found in the supplementary table, Glu-Gal VS YMC APA events.

5.2.1 Few genes undergo Coding APA in both the YMC and Glu-Gal shift

Of the genes that undergo Coding APA changes in the YMC and Glu-Gal shift, only one was found to be common to both of these data sets. Out of the 170 genes that undergo coding APA in the Glu-Gal shift and 122 in the YMC only one gene was found to occur in both. The only gene in the overlap was YPL186C, that undergoes APA shortening. The intersection of these data sets are shown in the venn diagram in 5.3 on page 266.

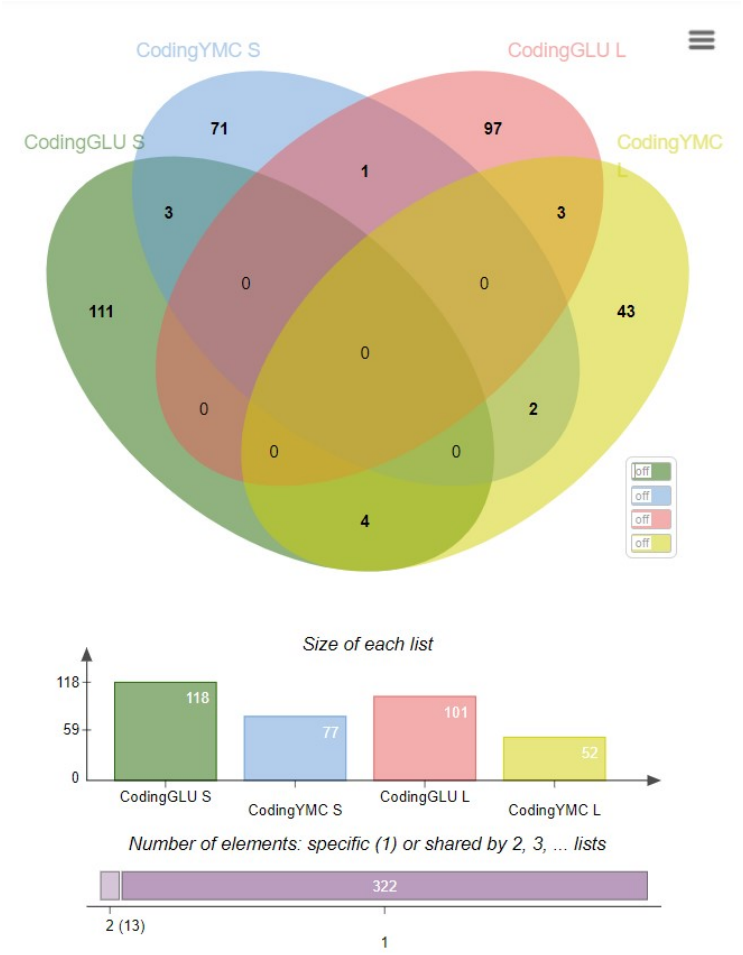


Figure 5.1: Venn diagram and bar charts visualise the APA event overlaps in the Glu-Gal shift and YMC data sets. These charts show all the genes that undergo coding shortening and lengthening APA events in both the Glu-Gal shift and YMC. The first bar chart underneath shows the number of genes in each individual data set and the purple bar chart at the bottom of the image is the total overlaps that occur.

5.2.2 Few genes undergo 3'UTR APA in both the YMC and Glu-Gal shift

Of the genes that undergo 3'UTR APA changes in the YMC and Glu-Gal shift, a few genes were common to both of these data sets. Out of the 49 genes that undergo 3'UTR APA in the Glu-Gal shift and 30 in the YMC 5 genes appear in both. The intersection of these data sets are shown in the venn diagram in 5.2 on page 264.

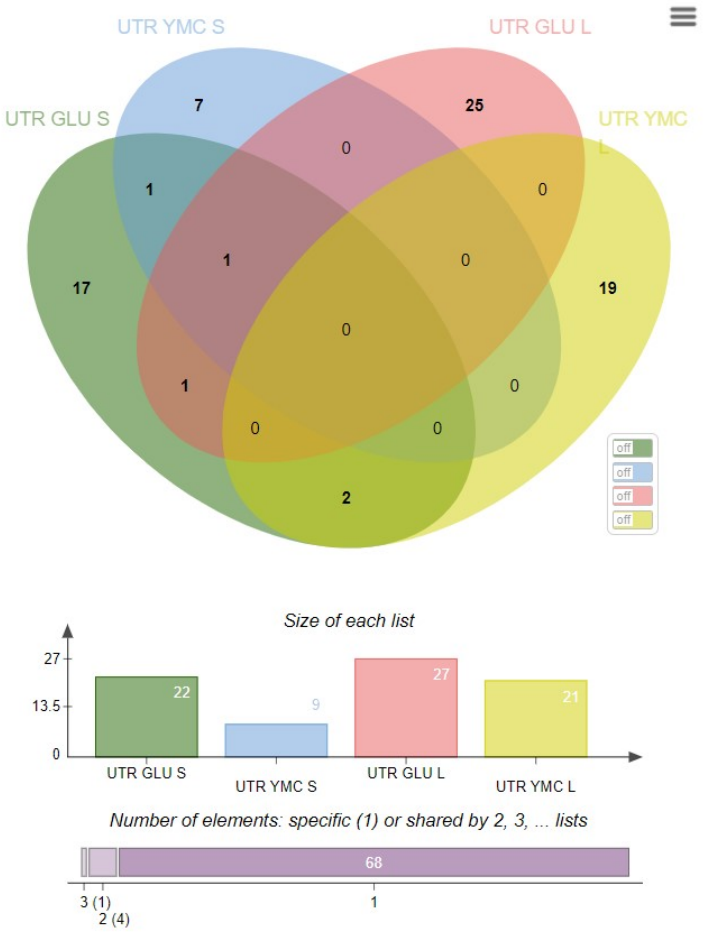


Figure 5.2: Venn diagram and bar charts visualise the APA event overlaps in the Glu-Gal shift and YMC data sets. These charts show all the genes that undergo 3'UTR shortening and lengthening APA events in both the Glu-Gal shift and YMC. The first bar chart underneath shows the number of genes in each individual data set and the purple bar chart at the bottom of the image is the total overlaps that occur.

5.2.3 Several genes undergo ALL APA in both the YMC and Glu-Gal shift

Of the genes that undergo ALL APA changes in the YMC and Glu-Gal shift, a small number of genes were common to both of these data sets. Out of the 513 genes that undergo UTR APA in the Glu-Gal shift and 225 in the YMC 5 genes appear in both. The intersection of these data sets are shown in the Venn diagram in 5.2 on page 264.

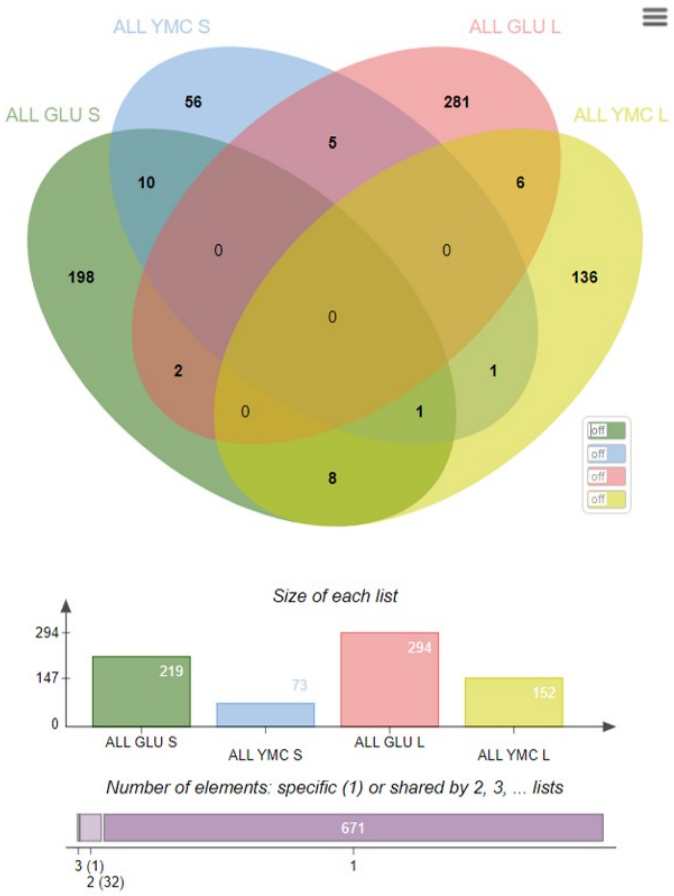


Figure 5.3: Venn diagram and bar charts visualise the APA event overlaps in the Glu-Gal shift and YMC data sets. These charts show all the genes that undergo ALL shortening and lengthening APA events in both the Glu-Gal shift and YMC. The first bar chart below shows the number of genes in each individual data set and the purple bar chart at the bottom of the image is the total overlaps that occur.

GO analysis for genes that undergo APA in the YMC and Glu-Gal shift shows enrichment in few functional categories

Based on the GO analysis, only a few genes exhibited enrichment in functional categories when considering the intersection of YMC and Glu-Gal coding APA data sets. The limited size of the individual time point data sets for APA coding prevented the attainment of statistically significant results via GO analysis. Consequently, an aggregate analysis was conducted across all genes within the data sets rather than specific time points. Notably, the sole category displaying enrichment was the intersection of YMC Lengthening and Glu-Gal Lengthening, with the gene YJL050W being enriched, primarily associated with nuclear mRNA surveillance during mRNA 3'-end processing. The summary table for the GO analysis is shown in the supplementary table 6.17 on page 302. In contrast, the GO analysis of the 3'UTR data sets revealed no enriched genes in functional categories for the intersection of YMC and Glu-Gal coding APA data sets. However, the intersection of the YMC and Glu-Gal ALL APA data sets, several genes exhibited enrichment in functional categories. Specifically, YMC and Glu-Gal shortening demonstrated enrichment in the negative regulation of pyruvate dehydrogenase activity, while various genes exhibited enrichment related to responses to nutrient levels within this intersection. The Venn diagram illustrating the intersection of these data sets is shown below in 5.1 on page 268.

ALL APA YMC Vs Glu-Gal			
YMC Shortening Glu-Gal shortening	GO Term	p-Value	genes
	negative regulation of pyruvate dehydrogenase activity [GO:1904183]	0.001742	YGL059W
	response to nutrient levels [GO:0031667]	0.001763	YBR214W YGL059W YNR002C
YMC Lengthening Glu-Gal Lengthening	negative regulation of septum digestion after cytokinesis [GO:2001042]	6.97E-04	YBR158W
	nuclear mRNA surveillance of mRNA 3'-end processing [GO:0071031]	0.003479	YJL050W
YMC Shortening Glu-Gal Lengthening			
	ATP biosynthetic process [GO:0006754]	2.40E-04	YDR322C-A YJL116C
YMC Lengthening Glu-Gal Shortening			
	N/A	N/A	N/A

Table 5.1: Table showing results of GO analysis for genes that undergo ALL APA events in both the YMC and the Glu-Gal shift. N/A is for data sets that did not have enough genes to have statistical significance or did not have any statistically significant GO enrichments

5.2.4 NETSeq and 3'RNA-Seq IGVs for *SUE1* for Glu-Gal and YMC show similar profiles at glucose starvation

Comparing expression profiles of genes that undergo similar changes during glucose starvation in the Glu-Gal shift and the RB phase of the YMC shows us potential mechanisms of gene expression control. *SUE1* undergoes APA in the YMC but does not in the Glu-Gal shift. The IGV tracks for the 3'RNA-Seq and NET-Seq data for the Glu-Gal shift shows similar gene expression profiles to the YMC at the RB phase which is during glucose starvation in the YMC and during the first stages of glucose starvation. *SUE1* is expressed more heavily in galactose, and at the RB phase, therefore indicating they are both repressed by glucose. The NETSeq and 3'RNA-Seq IGVs tracks are laid out side by side for comparison on 5.4 on page 270. This shows that APA is not controlling the expression profiles of *SUE1* but changes in its expression are most likely due to changes in glucose levels.

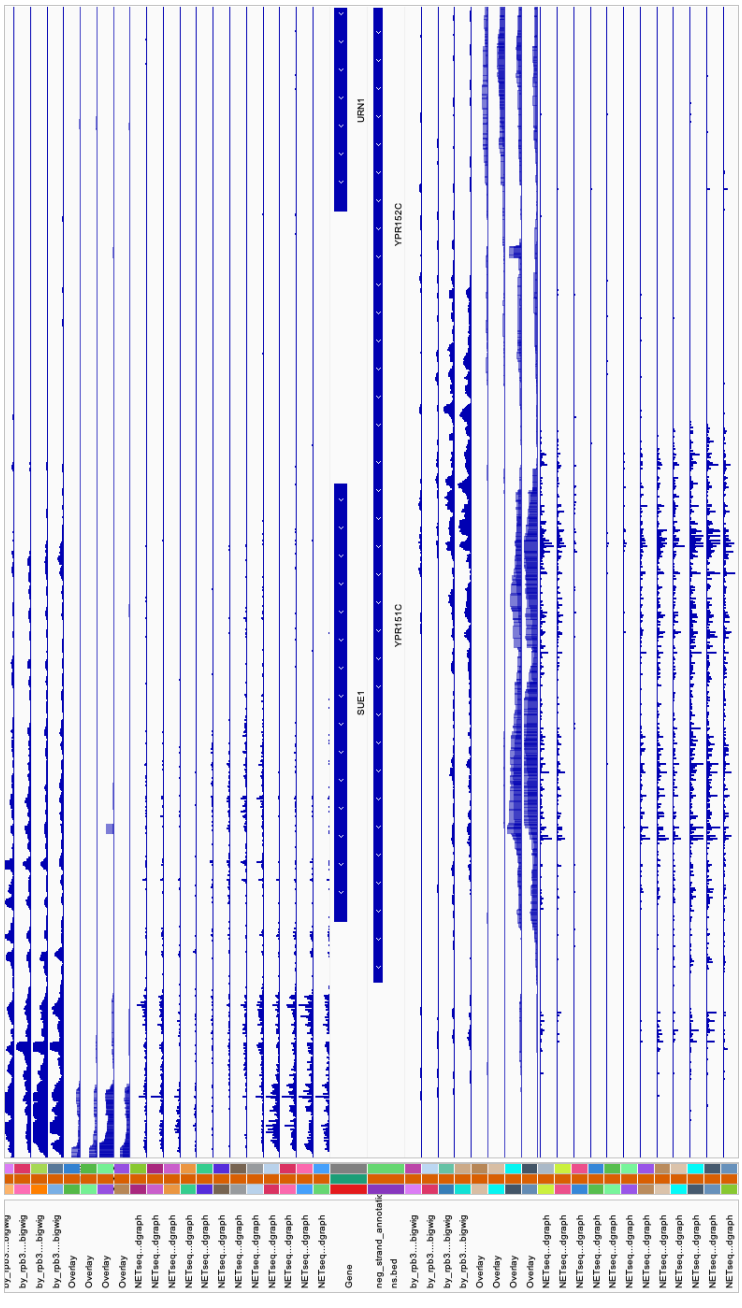


Figure 5.4: NETSeq and 3'RNA-Seq IGVs of *SUE1* show similar expression profiles for glucose starvation. NET-Seq and 3'RNA-Seq data are laid out side by side to show similar expression levels. Tracks N5 to N8 shows *SUE1* is expressed more heavily in galactose, and at the RB phase of the YMC, therefore indicating they are both repressed by glucose.

5.2.5 NET-Seq IGV shows *MIS1* is expressed in high glucose and repressed in galactose

As with *SUE1, MIS1* undergoes APA in the YMC but not during Glu-Gal shift. *MIS1* profiles for 3'RNA and NET-Seq show the gene expression changes in glucose starvation, i.e. gal 60 in the Glu-Gal shift and RC phase of the YMC 5.5 on page 272. As with *SUE1* this indicates that APA is not controlling the expression profiles of *MIS1* but changes in its expression are most likely due to changes in glucose levels. These comparisons of transcript level changes and APA in the Glu-Gal shift and the YMC are essential at revealing mechanisms of gene expression control in metabolic stress.



Figure 5.5: NET-Seq IGV shows *MIS1* is majoritively expressed in high glucose and repressed in galactose. NET-Seq and 3'RNA-Seq data are laid out side by side to show similar expression levels at comparable glucose starvation time points.

5.3 Discussion

The YMC and Glu-Gal shift are two biological events in *Saccharomyces cerevisiae* that require drastic shifts in metabolic activity and subsequently require acute temporal control of gene expression over a plethora of key genes. Comparing the two different biological processes was necessary to identify whether or not the APA events that occur are part of specific control pathways or just generic responses to metabolic stress. Identifying if these APA events are non-specific or targeted helps unravel the mechanisms and functions behind APA. The overlap of the genes that undergo APA in both the YMC and the Glu-Gal shift was very small. This indicates that most APA events observed were not generic stress responses but targeted discrete for these processes. Of the genes that underwent APA in both, many had GO analysis enrichment for more generic functions such as stress and 3-end processing. The events that undergo APA in both the YMC and Glu-Gal shift included the gene *YGL059W/PKP2* undergoing APA shortening in the 3'UTR for both. The gene *PKP2* is a negative regulator of the PDC an therefore moderates the step between glycolysis and the TCA cycle (Jeoung 2015). As this process is essential to both YMC and Glu-Gal shift, having the same gene that's key to PDC inhibition undergoing the same APA event in both would indicate APA might be used as a regulator of this pathway.

Both the YMC and Glu-Gal shift had a significant number of genes undergoing APA at all time points, with a peak in APA activity at the mid point of both processes. This indicates APA and biological activity are intrinsically linked and furthermore suggests that APA might be a key regulatory in

gene expression changes that are essential to facilitate these processes. There was APA identified between all time points, which peak at 15-60 minutes for Glu-Gal and 40-100 minutes for the 3a and 3b time points in the YMC. This confirms that APA is a rapid response to changes in the environment and alters the transcriptome within minutes. The YMC and Glu-Gal shift data suggest a higher prevalence of genes undergoing APA and DE, but there appears to be no effect on the type of APA events (i.e. the proportion of genes undergoing Codoing or 3'UTR APA). These unexpected results challenge prior research which suggested changes in stability or localization for genes that undergo APA but do not exhibit differential expression, as localization and stabilization typically occur in the 3'UTR of the transcript, so it was expected that genes undergoing APA without differential expression would show a higher proportion of 3'UTR APA events. However, these results suggest that there may be additional reasons for APA to occur beyond what the literature suggests. It is also possible that not all APA events cause functional changes to the transcript. These findings highlight the need for further investigation into the mechanisms underlying APA and its potential functional consequences. YMC APA profile showed a large preference for APA lengthening for all APA types and at all time points. This trend was not seen in the Glu-Gal shift APA and there was a small but statistically significant preference for shortening in the ALL APA data set. These differences in APA are not unexpected as the overlap of genes is very small so it is unlikely to be many overlapping events occurring of the same APA type. Global APA shortening has been attributed to shifts in sugar source availability, specifically yeast cells undergoing the diauxic response (Geis-

berg, Moqtaderi, Fan, et al. 2014). APA lengthening has been attributed to proliferation and APA lengthening reverses cellular quiescence allowing them to re-enter the cell cycle (Mitra et al. 2018). The two different global APA profiles seen in the YMC and Glu-Gal shift could be due to these two linked systems that occur in each respectively.

Both Glu-Gal and YMC had the most APA events in the ALL data set and the least in the 3'UTR data set. This is predictable as this data set has the largest number of genes so would therefore capture the largest number of APA events. The 3'UTR data sets being the smallest for both is more surprising as most of the literature refers to APA changes affecting the 3'UTR. This does not necessarily suggest that 3'UTR APA is more prevalent than pas site use changes that remove or include protein coding regions but just has more studies done on how it affects gene expression.

The IGV tracks for the genes that undergo APA in the YMC but not in the Glu-Gal shift show similar expression levels for these genes during glucose starvation. The RC in the YMC corresponds to the point as which the cells have depleted their glucose levels and are therefore undergoing glucose starvation, which is comparable to the Glu-Gal shift in which cells are starved of glucose. These genes undergoing APA in the YMC but not in the Glu-Gal shift indicate that the APA events are not being influenced by glucose levels but by another mechanism. It also indicates that gene expression is linked to glucose levels and not APA. Using this approach to cross compare genes in the YMC and Glu-Gal shift helps us unpick the complex and interconnected mechanisms of gene expression control during metabolic stress.

5.3.1 Further investigations

As with the previous assessments of further investigation in the Glu-Gal and YMC data sets, confirmation of the APA events via qPCR would be an essential first step. As many of the identified pas site changes during the YMC and Glu-Gal shift occur between one or both pas sites occurring in the coding region, it would be useful to conduct further studies on how altering the coding regions of mRNA via APA shifts alter the function of the protein. This would help us identify the full scale of effect APA has on both the YMC and Glu-Gal shift. There was little overlap in APA events that occur in the YMC and in the Glu-Gal shift. The genes that do overlap in both should be investigated for APA events that occur in upstream potentially interfering genes. This can be done via analysis explored in the previous chapters. Identifying the genes that undergo APA that might be causing downstream interference that occur in both the Glu-Gal shift and the YMC would be a strong indication of functional APA as a direct response to glucose starvation.

5.4 Concluding remarks

The combinatorial approach of RNA sequencing data and bioinformatics tools to investigate the APA profile of the YMC and Glu-Gal shift revealed many insights into mechanisms of gene expression control that involve shifts in pas site usage. Some of they key findings are as follows. It was found that APA was highly prevalent in both the YMC and Glu-Gal shift, and that APA events occurred predominantly at the peaks of metabolic activ-

ity, the RC phase and 15 minutes post glucose starvation respectively. The statistical analysis of the skew of the APA events, their Δpas lengths and their correlation to gene expression changes indicated that they were not random occurrences in heavily transcriptionally active genes and had a functional significance. The APA events for both the YMC and the Glu-Gal shift also mostly lead to an increase use of the distal pas site, i.e. global APA lengthening, which is surprising as the literature assumes shortening events are more favourable during stress events. Many genes that underwent APA events were enriched for GO functional categories that are tightly associated with mechanisms that occur in the YMC and Glu-Gal shift such as cell cycle progression, DNA damage and metabolic stress, which suggests they are functional or influenced by these events. The YMC and the Glu-Gal shift had few APA events that occur in both the YMC and Glu-Gal shift, which suggests that APA is not mainly driven by shared metabolic stresses, i.e. glucose starvation. Additionally, we found that many genes that the APA pipeline identified as undergoing APA events could potentially be undergoing transcriptional interference by neighbouring genes. Furthermore, the shifts in pas site usage observed in upstream noncoding RNAs could potentially lead to gene derepression downstream in key functional genes. This can be explored using the current data sets, and then confirmed with qPCR and knock out mutants of distal PAS sites. Understanding how these changes in APA are regulated can provide insights into how cells adapt to and survive under metabolic stress conditions. This is particularly relevant in the context of diseases and disorders that involve metabolic dysregulation, such as diabetes and obesity. By understanding the mechanisms of gene expres-

sion control during metabolic stress, it may be possible to develop therapies or interventions that can modulate gene expression in a way that protects against disease. In conclusion, this study of the APA profiles of the YMC and Glu-Gal shift provided insight into gene expression control mechanisms involving shifts in PAS site usage and can be used to identify novel mechanisms for adapting transcriptional outputs in response to changes in cellular states.

Chapter 6

Appendix

6.1 Table of descriptive statistics of the gene expression changes in the YMC

YMC Time point	1	2	3a	3b	4	5
Number of values	6633	6654	6694	6694	6649	6659
Minimum	-5.124	-5.028	-6.647	-7.195	-6.806	-6.115
Maximum	3.88	9.239	6.408	4.923	4.743	7.706
Range	9.004	14.27	13.05	12.12	11.55	13.82
Mean	-0.00106	0.02041	0.06407	-0.01932	-0.01852	0.01267
Std. Deviation	0.7378	0.8823	1.404	1.214	1.03	0.7712
Std. Error of Mean	0.009059	0.01082	0.01716	0.01484	0.01263	0.009451
Skewness	-0.203	0.1193	-0.2342	-0.5871	-0.2615	0.3212
Kurtosis	5.254	4.516	1.342	3.021	2.698	6.553

Table 6.1: Descriptive statistics of the gene expression changes in the YMC. The statistics were done on the logfold change in normalised reads per million of each gene.

6.1.1 Summary table of GO analysis results for genes that undergo 3'UTR APA changes in the YMC

UTR Lengthening	GO Term	P-value	Genes
1	glucose metabolic process	0.003529	YFR053C
	response to stress	0.00492	YDL073W
2	mitotic cell cycle	0.00091	YJR072C
3a	mitotic cell cycle	0.00536	YJR072C
3b	response to glucose limitation	0.00886	YMR116C
4	response to glucose limitation	0.008943	YMR116C
5	N/A		

Figure 6.1: Summary table of GO analysis results for genes that undergo 3'UTR APA changes in the YMC. Significance in enrichment for a functional group was shown to be p-values less than 0.05.

6.2 Table of APA genes that overlap with gene expression clusters

	Number of APA transcripts in clusters					
Cluster	Coding S	Coding L	3UTR S	3UTR L	ALL S	ALL L
1	1	4	0	2	5	14
2	5	3	0	2	7	7
3	0	2	1	0	2	5
4	0	3	0	1	5	7
5	0	2	0	0	1	8
6	1	2	0	1	2	6
7	1	2	0	0	2	5
8	0	0	0	0	1	3
9	0	1	1	1	2	5
10	4	6	1	0	7	11
11	0	1	0	0	4	3
12	0	2	0	0	0	4
13	1	0	0	1	4	5
14	3	8	0	6	3	19
15	1	0	1	0	7	4
16	0	7	0	3	2	13

Table 6.2: table of APA genes that occur in each of the 16 gene expression YMC clusters. Clusters made by Shidong Xi, in the Mellor laboratory, Oxford university Biochemistry department.

6.3 Table of APA genes that overlap with gene expression clusters (percentages)

	% of APA transcripts in clusters					
Cluster	Coding S	Coding L	3UTR S	3UTR L	ALL S	ALL L
1	4.76	7.69	0.00	9.52	6.85	9.21
2	23.81	5.77	0.00	9.52	9.59	4.61
3	0.00	3.85	11.11	0.00	2.74	3.29
4	0.00	5.77	0.00	4.76	6.85	4.61
5	0.00	3.85	0.00	0.00	1.37	5.26
6	4.76	3.85	0.00	4.76	2.74	3.95
7	4.76	3.85	0.00	0.00	2.74	3.29
8	0.00	0.00	0.00	0.00	1.37	1.97
9	0.00	1.92	11.11	4.76	2.74	3.29
10	19.05	11.54	11.11	0.00	9.59	7.24
11	0.00	1.92	0.00	0.00	5.48	1.97
12	0.00	3.85	0.00	0.00	0.00	2.63
13	4.76	0.00	0.00	4.76	5.48	3.29
14	14.29	15.38	0.00	28.57	4.11	12.50
15	4.76	0.00	11.11	0.00	9.59	2.63
16	0.00	13.46	0.00	14.29	2.74	8.55

Table 6.3: Table showing the percentage of genes that are found in gene expression clusters that also undergo different types of APA changes. Data sets with an overlap greater than 10% are shown in red.

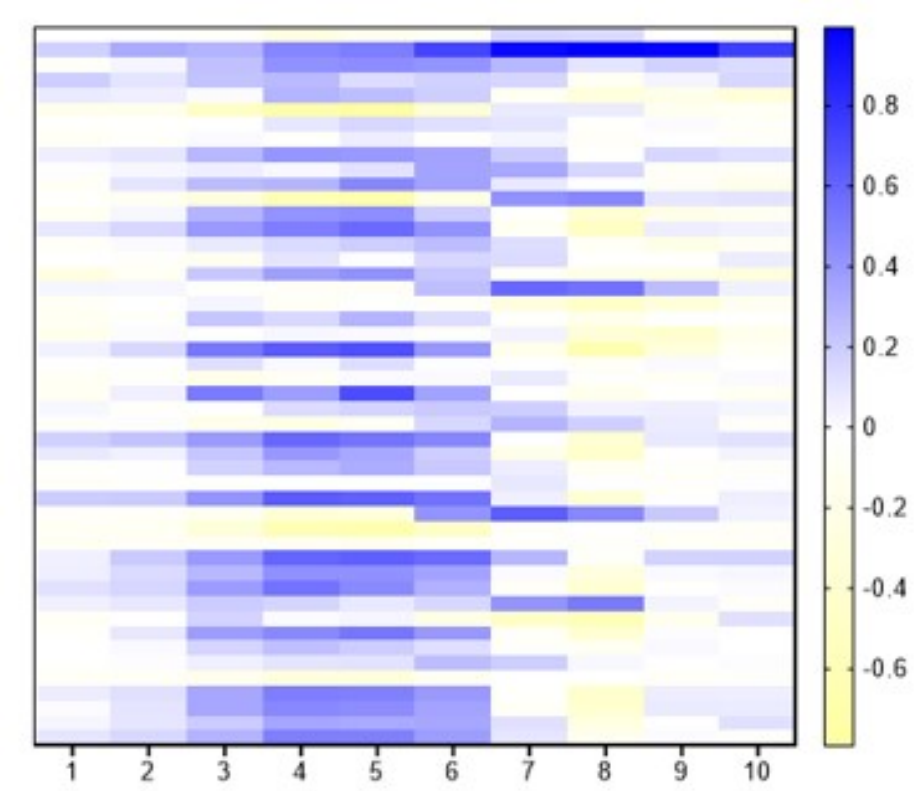


Figure 6.2: Heat map of NET-Seq log2fold gene expression changes for genes that undergo Coding APA in YMC showing peaks of gene expression occurs at the OX/RB phase, time points 3-6. Logfold2 changes that are above 0 i.e. increases in gene expression shown in green, logfold2 changes below 0 i.e. gene expression decreases shown in red

6.3.1 Heatmap of NET-Seq log2fold gene expression changes for genes that undergo Coding APA in YMC

APA Coding Shortening Cluster GO analysis		
Cluster 1		
Go Term	p-Value	genes
cysteinyl-tRNA aminoacylation [GO:0006423]	0.002095	YNL247W
inorganic ion transmembrane transport [GO:0098660]	0.006424	YDL199C YJR077C YPL036W
Cluster 2		
Go Term	p-Value	genes
heme A biosynthetic process [GO:0006784]	0.004184	YER141W
folic acid biosynthetic process [GO:0046656]	0.008353	YBR084W
glucan catabolic process [GO:0009251]	0.015267	YBR056W

Figure 6.3: GO analysis of the genes in the coding APA Shortening clusters showing enrichment in functional categories.Tables show GO terms for genes that have enrichment in GO functional groups with p-values less than 0.05.

6.3.2 GO analysis of Coding APA clusters

6.3.3 GO analysis of 3'UTR APA clusters

6.3.4 Heatmap of NET-Seq log2fold gene expression changes for genes that undergo 3'UTR APA in YMC

APA Coding Lengthening Cluster GO analysis		
Cluster 1		
Go Term	p-Value	genes
cytoplasmic translation [GO:0002181]	2.34E-04	YER102W YJL190C YLR192C YOR293W YPL143W, YPL220W
Cluster 2		
Go Term	p-Value	genes
mitochondrial genome maintenance [GO:0000002]	6.21E-04	YCR028C-A YJL143W YLR304C
regulation of transcription by RNA polymerase III [GO:0006359]	8.26E-04	YIL035C YOR038C
positive regulation of TOR signaling [GO:0032008]	9.74E-04	YGL008C YLR208W

Figure 6.4: GO analysis of the genes in the coding APA Lengthening clusters showing enrichment in functional categories.Tables show GO terms for genes that have enrichment in GO functional groups with p-values less than 0.05.

APA UTR Lengthening Cluster GO analysis		
Cluster 1		
Go Term	p-Value	genes
positive regulation of protein phosphorylation [GO:0001934]	0.001674	YMR116C YNL055C
apoptotic process [GO:0006915]	0.002065	YJL208C Y NL055C
Cluster 2		
GO Term	p-Value	genes
regulation of MAP kinase activity [GO:0043405]	0.001396	YDL073W
protein adenylylation [GO:0018117]	0.002791	YPL222W

Figure 6.5: GO analysis of the genes in the 3'UTR APA Lengthening clusters.Tables show GO terms for genes that have enrichment in GO functional groups with p-values less than 0.05.

APA UTR Shortening Cluster GO analysis		
Cluster 1		
Go Term	p-Value	genes
positive regulation of protein processing [GO:0010954]	5.24E-04	YNL169C
Cluster 2		
Go Term	p-Value	genes
nucleotide-excision repair, DNA duplex unwinding [GO:0000717]	5.24E-04	YER171W

Figure 6.6: GO analysis of the genes in the 3'UTR APA Shortening clusters that show enriched functional categories for APA genes in each cluster.Tables show GO terms for genes that have enrichment in GO functional groups with p-values less than 0.05.

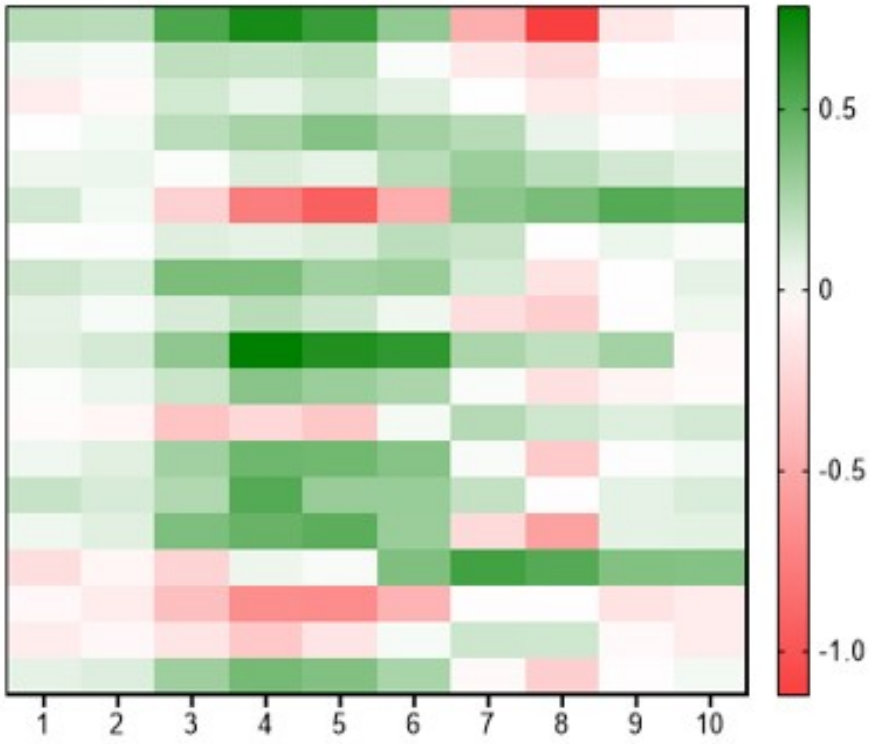


Figure 6.7: Heat map of NET-Seq log2fold gene expression changes for genes that undergo 3'UTR APA in YMC showing peaks of gene expression occurs at the OX/RB phase, time points 3-6. Logfold2 changes that are above 0 i.e. increases in gene expression shown in blue, logfold2 changes below 0 i.e. gene expression decreases shown in yellow

6.3.5 Heatmap of NET-Seq log2fold gene expression changes for genes that undergo ALL APA in YMC

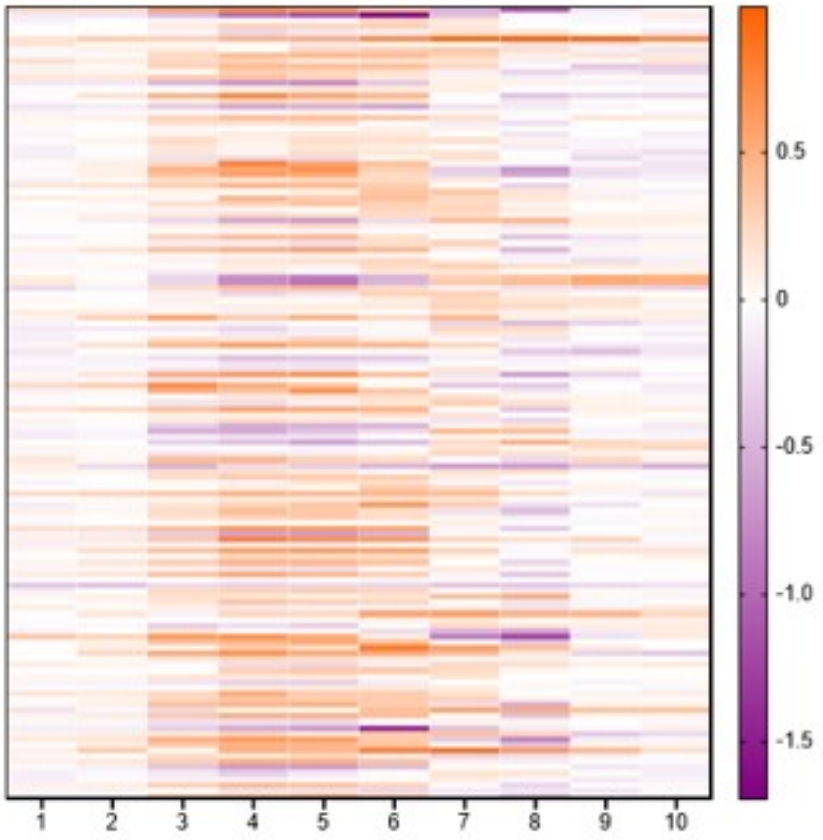


Figure 6.8: Heat map of NET-Seq log2fold gene expression changes for genes that undergo ALL APA in YMC showing peaks of gene expression occurs at the OX/RB phase, time points 3-6. Logfold2 changes that are above 0 i.e. increases in gene expression shown in orange, logfold2 changes below 0 i.e. gene expression decreases shown in purple

6.4 Spearman's rank correlation coefficent matrix for coding shortening events during the YMC

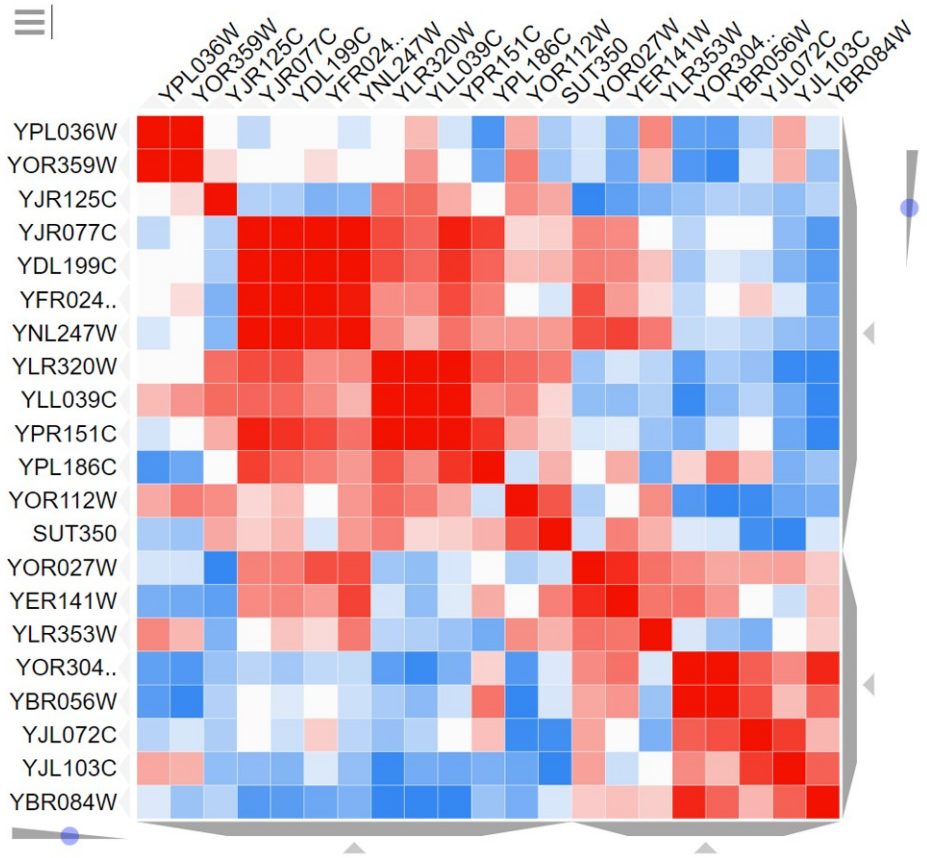


Figure 6.9: Spearman's rank correlation coefficent matrix of log fold change values of gene expression changes in genes that undergo APA coding shortening at each time point in the YMC. Red/blue cells in the heatmap matrix represent positive/negative coefficient values, darker blue is closest to -1, total negative correlation and red 1, positive correlation.

6.5 Spearman's rank correlation coefficent matrix for coding lengthening events during the YMC

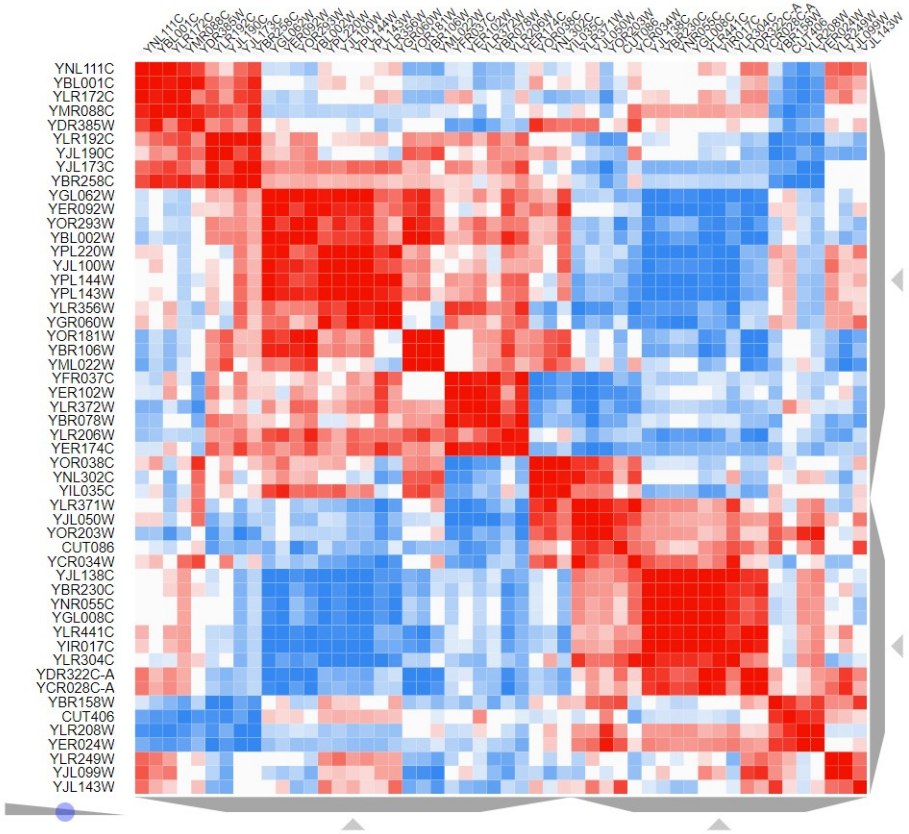


Figure 6.10: Spearman's rank correlation coefficent matrix of log fold change values of gene expression changes in genes that undergo APA coding lengthening at each time point in the YMC. Red/blue cells in the heatmap matrix represent positive/negative coefficient values, darker blue is closest to -1, total negative correlation and red 1, positive correlation.

6.6 Spearman's rank correlation coefficent matrix for 3'UTR shortening events during the YMC

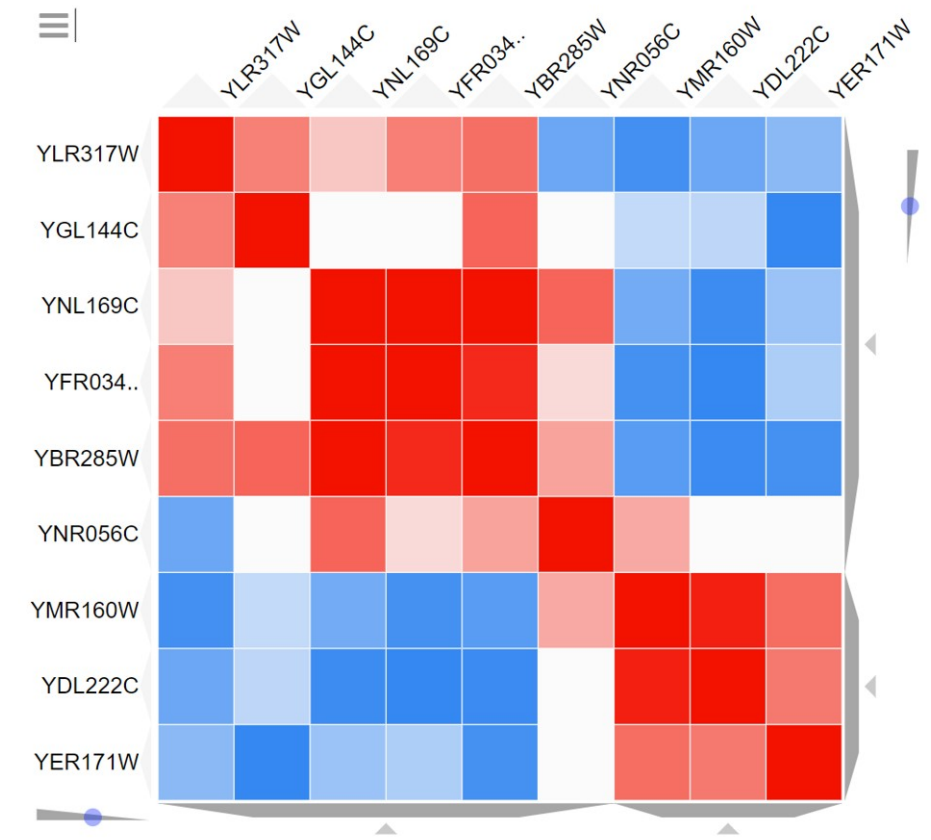


Figure 6.11: Spearman's rank correlation coefficent matrix of log fold change values of gene expression changes in genes that undergo APA 3'UTR shortening at each time point in the YMC. Red/blue cells in the heatmap matrix represent positive/negative coefficient values, darker blue is closest to -1, total negative correlation and red 1, positive correlation.

6.7 Spearman's rank correlation coefficent matrix for 3'UTR lengthening events during the YMC

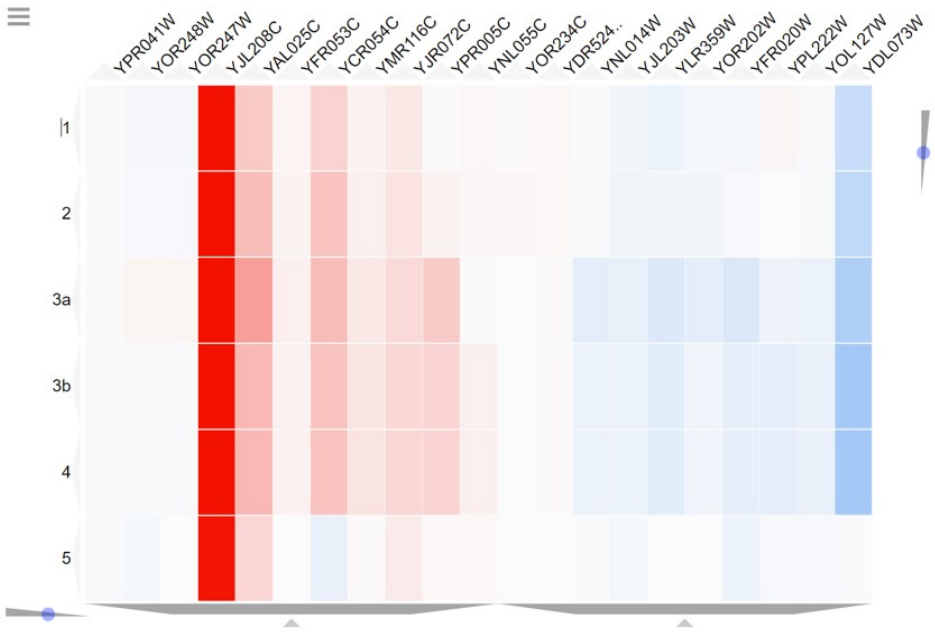


Figure 6.12: Spearman's rank correlation coefficent matrix of log fold change values of gene expression changes in genes that undergo APA 3'UTR lengthening at each time point in the YMC. Red/blue cells in the heatmap matrix represent positive/negative coefficient values, darker blue is closest to -1, total negative correlation and red 1, positive correlation.

6.8 Spearman's rank correlation coefficent matrix for ALL shortening events during the YMC

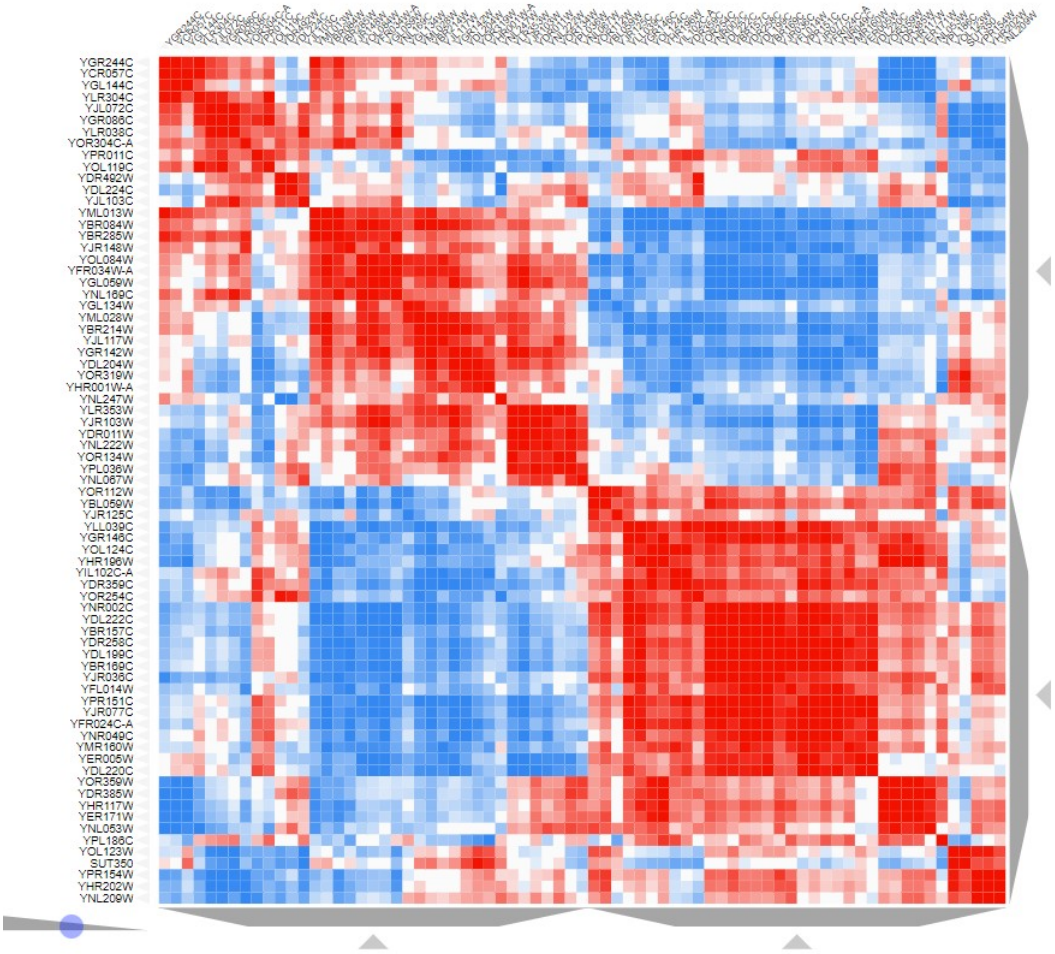


Figure 6.13: Spearman's rank correlation coefficent matrix of log fold change values of gene expression changes in genes that undergo APA ALL shortening at each time point in the YMC. Red/blue cells in the heatmap matrix represent positive/negative coefficient values, darker blue is closest to -1, total negative correlation and red 1, positive correlation.

6.9 Spearman's rank correlation coefficent matrix for ALL lengthening events during the YMC

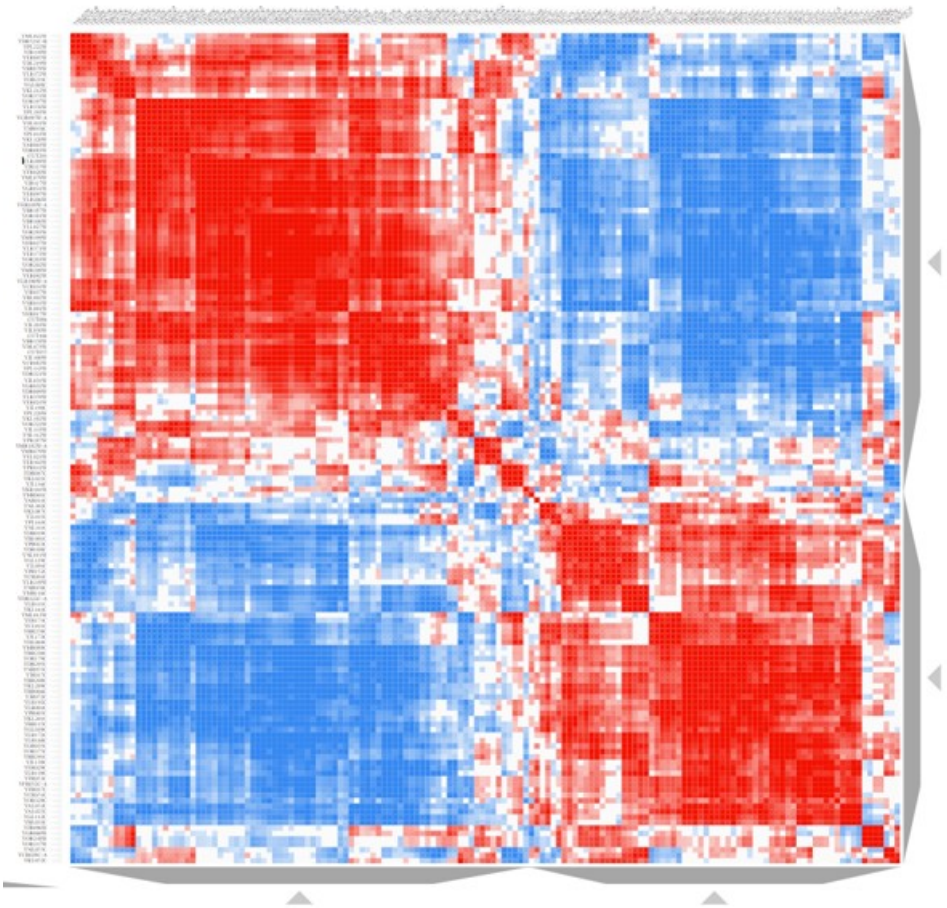


Figure 6.14: Spearman's rank correlation coefficent matrix of log fold change values of gene expression changes in genes that undergo APA ALL lengthening at each time point in the YMC. Red/blue cells in the heatmap matrix represent positive/negative coefficient values, darker blue is closest to -1, total negative correlation and red 1, positive correlation.

6.10 Summary table of GO analysis results for genes that undergo coding APA changes during Glu-Gal shift

Coding Shortening	GO Term	Genes	P value
5	organophosphate metabolic process [GO:0019637]	YDR322C-A,YJL062W,YKL004W,YLR345W,YML035C,YOR155C,YPL189W,YPR020W	9.14E-04
	protein localization to extracellular region [GO:0071692]	YGR127W,YOR329C	0.005064
15	export from cell [GO:0140352]	YBR008C,YDR011W,YDR084C,YER008C,YGR127W	0.001571
	cell development [GO:0048468]	YBR107C,YDL222C,YGL197W,YIL017C,YOR313C	0.005825
60	cristae formation [GO:0042407]	YDR322C-A,YPR020W	0.001062
	fluconazole transport [GO:0015903]	YBR008C	0.004537
180	xenobiotic export from cell [GO:0046618]	YBR008C YDR011W	0.001269
	cristae formation [GO:0042407]	YDR322C-A YPR020W	0.001924
Coding Lengthening	GO Term	Genes	P value
5	signal transduction involved in filamentous growth [GO:0001402]	YDR144C YGR040W	0.001274
	DNA geometric change [GO:0032392]	YER176W YHR031C YMR106C	0.001710
15	cellular cation homeostasis [GO:0030003]	YBR037C,YDR456W,YDR492W,YGR092W,YIL165C,YJR104C	0.001705
	regulation of DNA-binding transcription factor activity [GO:0051090]	YGR040W,YJR104C	0.002382
60	tRNA N2-guanine methylation [GO:0002940]	YDR120C	0.004188
	carbohydrate transmembrane transport [GO:0034219]	YFL054C YHR094C	0.012469
180	cellular chemical homeostasis [GO:0055082]	YBR037C,YDR492W,YJL165C,YJR049C,YLR446W,YMR244C-A	0.000661
	chaperone-mediated protein folding [GO:0061077]	YJL162C,YKL073W	0.0103157

Figure 6.15: Summary table of GO analysis results for genes that undergo coding APA changes during Glu-Gal shift.

6.11 Summary table of GO analysis results for genes that undergo 3'UTR APA changes during Glu-Gal shift

UTR Shortening	GO Term	Genes	P value
5	protein deglutathionylation [GO:0080058]	YGR209C	0.003139
	response to osmotic stress [GO:0006970]	YLR113W,YPR005C	0.00487
15	cell cycle [GO:0007049]	DL222C,YHR071W,YJR072C,YLR113W,YNL231C,YOR026W	0.008027
	cellular protein modification process [GO:0006464]	YDL091C,YGR209C,YHR071W,YIL062W,YLR113W,YOR026W	0.008693
60	protein deglutathionylation [GO:0080058]	YGR209C	0.003139
	malate metabolic process [GO:0006108]	YKL085W	0.00783
180	protein deglutathionylation [GO:0080058]	YGR209C	0.002442
	malate metabolic process [GO:0006108]	YKL085W	0.006094
UTR Lengthening	GO Term	Genes	P value
5	N-terminal peptidyl-proline methylation [GO:0035568]	YBR261C	0.001396
	mRNA transport [GO:0051028]	YDL207W,YLL023C	0.007729
15	folic acid-containing compound biosynthetic process [GO:0009396]	YGR267C,YKL132C	4.16E-04
	microtubule anchoring at spindle pole body [GO:0034631]	YAL047C	0.002617
60	folic acid-containing compound biosynthetic process [GO:0009396]	YGR267C,YKL132C	1.44E-04
	nuclear mRNA surveillance of mRNP export [GO:0071032]	YOR166C	0.00157
180	reproductive process in single-celled organism [GO:0022413]	YBR045C,YDR273W,YMR144W	3.50E-03
	dihydrofolate biosynthetic process [GO:0006761]	YKL132C	4.18E-03

Figure 6.16: Summary table of GO analysis results for genes that undergo 3'UTR APA changes during Glu-Gal shift.

6.11.1 Table of GO analysis for genes that undergo Coding APA events in both the YMC and the Glu-Gal shift

Coding APA YMC Vs Glu-Gal			
YMC Shortening Glu-Gal shortening	GO Term	p-Value	genes
	N/A	N/A	N/A
YMC Lengthening Glu-Gal Lengthening			
	nuclear mRNA surveillance of mRNA 3'-end processing [GO:0071031]	0.001741	YJL050W
YMC Shortening Glu-Gal Lengthening			
	N/A	N/A	N/A
YMC Lengthening Glu-Gal Shortening			
	N/A	N/A	N/A

Figure 6.17: Table showing results of GO analysis for genes that undergo Coding APA events in both the YMC and the Glu-Gal shift. N/A is for data sets that did not have enough genes to have statistical significance or did not have any statistically significant GO enrichments.

Bibliography

Alt, Frederick W et al. (1980). "Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3 ends". In: *Cell* 20.2, pp. 293–301.

Ares, Manuel, LESLIE Grate, and MICHELLE H Pauling (1999). "A handful of intron-containing genes produces the lion's share of yeast mRNA". In: *Rna* 5.9, pp. 1138–1139.

Badjatia, Nitika et al. (2021). "Acute stress drives global repression through two independent RNA polymerase II stalling events in Saccharomyces". In: *Cell reports* 34.3, p. 108640.

Baejen, Carlo et al. (2014). "Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition". In: *Molecular cell* 55.5, pp. 745–757.

Bakheet, Tala et al. (2018). "The AU-rich element landscape across human transcriptome reveals a large proportion in introns and regulation by ELAVL1/HuR". In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1861.2, pp. 167–177.

Balakrishnan, Rama et al. (2012). "YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit". In: *Database* 2012.

Barabino, SM et al. (1997). "The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins." In: *Genes & development* 11.13, pp. 1703–1716.

Barnwal, Ravi Pratap et al. (2012). "Structural and biochemical analysis of the assembly and function of the yeast pre-mRNA 3 end processing complex CF I". In: *Proceedings of the National Academy of Sciences* 109.52, pp. 21342–21347.

Bava, Felice-Alessio et al. (2013). "CPEB1 coordinates alternative 3-UTR formation with translational regulation". In: *Nature* 495.7439, pp. 121–125.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Berg, Michael G et al. (2012). "U1 snRNP determines mRNA length and regulates isoform expression". In: *Cell* 150.1, pp. 53–64.

Berrar, Daniel (2019). *Introduction to the Non-Parametric Bootstrap.*

Bharati, Akhilendra Pratap et al. (2016). "The mRNA capping enzyme of Saccharomyces cerevisiae has dual specificity to interact with CTD of RNA Polymerase II". In: *Scientific reports* 6.1, pp. 1–12.

Blankenberg, Daniel et al. (2010). "Galaxy: a web-based genome analysis tool for experimentalists". In: *Current protocols in molecular biology* 89.1, pp. 19–10.

Boreikaite, Vytaute et al. (2022). "RBBP6 activates the pre-mRNA 3 end processing machinery in humans". In: *Genes & development* 36.3-4, pp. 210–224.

Brown, Thomas et al. (2018). "Antisense transcription-dependent chromatin signature modulates sense transcript dynamics". In: *Molecular systems biology* 14.2, e8007.

Brumbaugh, Justin et al. (2018). "Nudt21 controls cell fate by connecting alternative polyadenylation to chromatin signaling". In: *Cell* 172.1-2, pp. 106–120.

Burnetti, Anthony J, Mert Aydin, and Nicolas E Buchler (2016). "Cell cycle Start is coupled to entry into the yeast metabolic cycle across diverse strains and growth rates". In: *Molecular biology of the cell* 27.1, pp. 64–74.

Cartegni, Luca, Shern L Chew, and Adrian R Krainer (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing". In: *Nature reviews genetics* 3.4, pp. 285–298.

Casañal, Ana et al. (2017). "Architecture of eukaryotic mRNA 3-end processing machinery". In: *Science* 358.6366, pp. 1056–1059.

Cesur, Müberra Fatma, Tunahan Çakır, and Pınar Pir (2022). "Genome-Wide Analysis of Yeast Metabolic Cycle through Metabolic Network Models Reveals Superiority of Integrated ATAC-seq Data over RNA-seq Data". In: *Msystems*, e01347–21.

Chaudhury, Arindam, Praveen Chander, and Philip H Howe (2010). "Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus

on hnRNP E1's multifunctional regulatory roles". In: *Rna* 16.8, pp. 1449–1462.

Cheng, Larry C et al. (2020). "Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation". In: *Nature communications* 11.1, pp. 1–14.

Churchman, L Stirling and Jonathan S Weissman (2011). "Nascent transcript sequencing visualizes transcription at nucleotide resolution". In: *Nature* 469.7330, pp. 368–373.

Clancy, Suzanne et al. (2008). "RNA splicing: introns, exons and spliceosome". In: *Nature Education* 1.1, p. 31.

Cowling, Victoria H (2010). "Regulation of mRNA cap methylation". In: *Biochemical Journal* 425.2, pp. 295–302.

Custódio, Noélia and Maria Carmo-Fonseca (2016). "Co-transcriptional splicing and the CTD code". In: *Critical reviews in biochemistry and molecular biology* 51.5, pp. 395–411.

Despic, Vladimir and Karla M Neugebauer (2018). "RNA tales–how embryos read and discard messages from mom". In: *Journal of Cell Science* 131.5, jcs201996.

Desterro, Joana, Pedro Bak-Gordon, and Maria Carmo-Fonseca (2020). "Targeting mRNA processing as an anticancer strategy". In: *Nature Reviews Drug Discovery* 19.2, pp. 112–129.

Di Giammartino, Dafne Campigli, Wencheng Li, et al. (2014). "RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3 UTRs". In: *Genes & Development* 28.20, pp. 2248–2260.

Di Giammartino, Dafne Campigli, Kensei Nishida, and James L Manley (2011). "Mechanisms and consequences of alternative polyadenylation". In: *Molecular cell* 43.6, pp. 853–866.

Dichtl, Bernhard, Diana Blank, et al. (2002). "Yhh1p/Cft1p directly links poly (A) site recognition and RNA polymerase II transcription termination". In: *The EMBO journal* 21.15, pp. 4125–4135.

Dichtl, Bernhard and Walter Keller (2001). "Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor". In: *The EMBO Journal* 20.12, pp. 3197–3209.

Dvir, Shlomi et al. (2013). "Deciphering the rules by which 5-UTR sequences affect protein expression in yeast". In: *Proceedings of the National Academy of Sciences* 110.30, E2792–E2801.

Elkon, Ran, Alejandro P Ugalde, and Reuven Agami (2013). "Alternative cleavage and polyadenylation: extent, regulation and function". In: *Nature Reviews Genetics* 14.7, pp. 496–506.

Erson-Bensan, Ayse Elif (2016). "Alternative polyadenylation and RNA-binding proteins". In: *Journal of molecular endocrinology* 57.2, F29–F34.

Escalante-Chong, Renan et al. (2015). "Galactose metabolic genes in yeast respond to a ratio of galactose and glucose". In: *Proceedings of the National Academy of Sciences* 112.5, pp. 1636–1641.

Fair, Benjamin Jung and Jeffrey A Pleiss (2017). "The power of fission: yeast as a tool for understanding complex splicing". In: *Current genetics* 63.3, pp. 375–380.

Feltham, Jack E et al. (2020). "Transcriptional changes are regulated by metabolic pathway dynamics but decoupled from protein levels". In: *BioRxiv*, p. 833921.

Ferreira, JA and AH Zwinderman (2006). "On the benjamini–hochberg method". In: *The Annals of Statistics* 34.4, pp. 1827–1849.

Fischl, Harry, Thomas Brown, et al. (2021). "Global and Gene-specific Transcriptional Responses to Acute Stress". In: *bioRxiv*.

Fischl, Harry, Françoise S Howe, et al. (2017). "Paf1 has distinct roles in transcription elongation and differential transcript fate". In: *Molecular cell* 65.4, pp. 685–698.

Fox-Walsh, Kristi and Xiang-Dong Fu (2010). "Chromatin: the final frontier in splicing regulation?" In: *Developmental cell* 18.3, pp. 336–338.

Gancedo, Juana M (1998). "Yeast carbon catabolite repression". In: *Microbiology and molecular biology reviews* 62.2, pp. 334–361.

Geisberg, Joseph V, Zarmik Moqtaderi, Xiaochun Fan, et al. (2014). "Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast". In: *Cell* 156.4, pp. 812–824.

Geisberg, Joseph V, Zarmik Moqtaderi, and Kevin Struhl (2020). "The transcriptional elongation rate regulates alternative polyadenylation in yeast". In: *Elife* 9, e59810.

Gey, Uta et al. (2008). "Yeast pyruvate dehydrogenase complex is regulated by a concerted activity of two kinases and two phosphatases". In: *Journal of Biological Chemistry* 283.15, pp. 9759–9767.

González, Carlos I et al. (2000). "The yeast hnRNP-like protein Hrp1/Nab4 marks a transcript for nonsense-mediated mRNA decay". In: *Molecular cell* 5.3, pp. 489–499.

Green, Michael R (1986). "Pre-mRNA splicing". In: *Annual review of genetics* 20.1, pp. 671–708.

Gross, Stefan and Claire L Moore (2001). "Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3-end formation". In: *Molecular and cellular biology* 21.23, pp. 8045–8055.

Grozdanov, Petar N et al. (2018). "The structural basis of CstF-77 modulation of cleavage and polyadenylation through stimulation of CstF-64 activity". In: *Nucleic acids research* 46.22, pp. 12022–12039.

Gruber, Andreas J and Mihaela Zavolan (2019). "Alternative cleavage and polyadenylation in health and disease". In: *Nature Reviews Genetics* 20.10, pp. 599–614.

Gruber, Andreas R et al. (2014). "Global 3 UTR shortening has a limited effect on protein abundance in proliferating T cells". In: *Nature communications* 5.1, pp. 1–10.

Gu, Wei et al. (2004). "A new yeast PUF family protein, Puf6p, represses ASH1 mRNA translation and is required for its localization". In: *Genes & development* 18.12, pp. 1452–1465.

Guo, Zijian and Fred Sherman (1995). "3'-end-forming signals of yeast mRNA". In: *Molecular and cellular biology* 15.11, pp. 5983–5990.

— (1996). "3-end-forming signals of yeast mRNA". In: *Trends in biochemical sciences* 21.12, pp. 477–481.

Gupta, Ishaan et al. (2014). "Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA–protein interactions". In: *Molecular systems biology* 10.2, p. 719.

Hogg, J Robert and Stephen P Goff (2010). "Upf1 senses 3 UTR length to potentiate mRNA decay". In: *Cell* 143.3, pp. 379–389.

Huntzinger, Eric and Elisa Izaurralde (2011). "Gene silencing by microRNAs: contributions of translational repression and mRNA decay". In: *Nature Reviews Genetics* 12.2, pp. 99–110.

Janapala, Yoshika, Thomas Preiss, and Nikolay E Shirokikh (2019). "Control of translation at the initiation phase during glucose starvation in yeast". In: *International Journal of Molecular Sciences* 20.16, p. 4043.

Jenal, Mathias et al. (2012). "The poly (A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites". In: *Cell* 149.3, pp. 538–553.

Jenny, Andreas et al. (1996). "Sequence similarity between the 73-kilodalton protein of mammalian CPSF and a subunit of yeast polyadenylation factor I". In: *Science* 274.5292, pp. 1514–1517.

Jeoung, Nam Ho (2015). "Pyruvate dehydrogenase kinases: therapeutic targets for diabetes and cancers". In: *Diabetes & metabolism journal* 39.3, pp. 188–197.

Juneau, Kara, Corey Nislow, and Ronald W Davis (2009). "Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization". In: *Genetics* 183.1, pp. 185–194.

Kaczmarek Michaels, Katarzyna et al. (2020). "Regulation of alternative polyadenylation in the yeast Saccharomyces cerevisiae by histone H3K4

and H3K36 methyltransferases". In: *Nucleic acids research* 48.10, pp. 5407–5425.

Kai, Zoya S and Amy E Pasquinelli (2010). "MicroRNA assassins: factors that regulate the disappearance of miRNAs". In: *Nature structural & molecular biology* 17.1, pp. 5–10.

Katz, Yarden et al. (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation". In: *Nature methods* 7.12, pp. 1009–1015.

Kelemen, Olga et al. (2013). "Function of alternative splicing". In: *Gene* 514.1, pp. 1–30.

Keller, Walter et al. (1991). "Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3 processing signal AAUAAA." In: *The EMBO journal* 10.13, pp. 4241–4249.

Kerr, Emily O and Maitreya J Dunham (2017). "Chemostat culture for yeast physiology". In: *Cold Spring Harbor Protocols* 2017.7, pdb–prot089003.

Kessler, Marco M et al. (1997). "Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3-end formation in yeast". In: *Genes & development* 11.19, pp. 2545–2556.

Kim Guisbert, Karen S, Hao Li, and Christine Guthrie (2007). "Alternative 3 pre-mRNA processing in Saccharomyces cerevisiae is modulated by Nab4/Hrp1 in vivo". In: *PLoS biology* 5.1, e6.

Klerk, Eleonora de et al. (2012). "Poly (A) binding protein nuclear 1 levels affect alternative polyadenylation". In: *Nucleic acids research* 40.18, pp. 9089–9101.

Klevecz, Robert R and Douglas B Murray (2001). "Genome wide oscillations in expression–Wavelet analysis of time series data from yeast expression arrays uncovers the dynamic architecture of phenotype". In: *Molecular Biology Reports* 28.2, pp. 73–82.

Konarska, Maria M et al. (1985). "Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors". In: *Nature* 313.6003, pp. 552–557.

Kumar, Ananthanarayanan et al. (2019). "Mechanistic insights into mRNA 3-end processing". In: *Current Opinion in Structural Biology* 59, pp. 143–150.

Kyburz, Andrea et al. (2003). "The role of the yeast cleavage and polyadenylation factor subunit Ydh1p/Cft2p in pre-mRNA 3-end formation". In: *Nucleic acids research* 31.14, pp. 3936–3945.

Lackford, Brad et al. (2014). "Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal". In: *The EMBO journal* 33.8, pp. 878–889.

Lee, Chien-Der and Benjamin P Tu (2015). "Glucose-regulated phosphorylation of the PUF protein Puf3 regulates the translational fate of its bound mRNAs and association with RNA granules". In: *Cell reports* 11.10, pp. 1638–1650.

Lee, Yong Joo et al. (2011). "TCA cycle-independent acetate metabolism via the glyoxylate cycle in Saccharomyces cerevisiae". In: *Yeast* 28.2, pp. 153–166.

Leek, Jeffrey T et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10, pp. 733–739.

Li, Heng et al. (2009). "The sequence alignment/map format and SAMtools". In: *bioinformatics* 25.16, pp. 2078–2079.

Li, Wencheng et al. (2015). "Systematic profiling of poly (A)+ transcripts modulated by core 3'end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation". In: *PLoS genetics* 11.4, e1005166.

Lin, Shuibin and Richard I Gregory (2015). "MicroRNA biogenesis pathways in cancer". In: *Nature reviews cancer* 15.6, pp. 321–333.

Liu, Xiaochuan et al. (2017). "Comparative analysis of alternative polyadenylation in S. cerevisiae and S. pombe". In: *Genome research* 27.10, pp. 1685–1695.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12, pp. 1–21.

Luo, Wenting et al. (2013). "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". In: *PLoS genetics* 9.7, e1003613.

Madhani, Hiten D and Christine Guthrie (1994). "Dynamic RNA-RNA interactions in the splicesome". In: *Annual review of genetics* 28, pp. 1–27.

Mahilkar, Anjali, Supreet Saini, et al. (2022). "GAL Regulon in the Yeast S. cerevisiae is Highly Evolvable via Acquisition in the Coding Regions of the Regulatory Elements of the Network". In: *Frontiers in molecular biosciences* 9.

Maicas, Sergi (2020). *The role of yeasts in fermentation processes*.

Mandel, Corey R et al. (2006). "Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease". In: *Nature* 444.7121, pp. 953–956.

Manley, James L and Yoshio Takagaki (1996). "The end of the message—another link between yeast and mammals". In: *Science* 274.5292, pp. 1481–1482.

Mao, Zhanrui et al. (2020). "Post-transcriptional dysregulation of microRNA and alternative polyadenylation in colorectal cancer". In: *Frontiers in genetics* 11, p. 64.

Mar, Jessica C (2019). "The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond". In: *Biophysical reviews* 11.1, pp. 89–94.

Masuda, Fumie et al. (2016). "Glucose restriction induces transient G2 cell cycle arrest extending cellular chronological lifespan". In: *Scientific reports* 6.1, pp. 1–9.

Matera, A Gregory and Zefeng Wang (2014). "A day in the life of the spliceosome". In: *Nature reviews Molecular cell biology* 15.2, pp. 108–121.

Mayr, Christine and David P Bartel (2009). "Widespread shortening of 3 UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells". In: *Cell* 138.4, pp. 673–684.

Merrick, William C (2004). "Cap-dependent and cap-independent translation in eukaryotic systems". In: *Gene* 332, pp. 1–11.

Mischo, Hannah E and Nick J Proudfoot (2013). "Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast". In: *BBA-Gene Regulatory Mechanisms* 1.1829, pp. 174–185.

Mitra, Mithun et al. (2018). "Alternative polyadenylation factors link cell cycle to migration". In: *Genome biology* 19.1, pp. 1–24.

Nanavaty, Vishal et al. (2020). "DNA methylation regulates alternative polyadenylation via CTCF and the cohesin complex". In: *Molecular cell* 78.4, pp. 752–764.

Nedea, Eduard, Xiaoyuan He, et al. (2003). "Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3-ends". In: *Journal of Biological Chemistry* 278.35, pp. 33000–33010.

Nedea, Eduard, Demet Nalbant, et al. (2008). "The Glc7 phosphatase subunit of the cleavage and polyadenylation factor is essential for transcription termination on snoRNA genes". In: *Molecular cell* 29.5, pp. 577–587.

Neve, Jonathan et al. (2017). "Cleavage and polyadenylation: Ending the message expands gene regulation". In: *RNA biology* 14.7, pp. 865–890.

Nguyen-Huu, Truong D et al. (2015). "Timing and variability of galactose metabolic gene activation depend on the rate of environmental change". In: *PLoS computational biology* 11.7, e1004399.

Nourani, Amine, Francois Robert, and Fred Winston (2006). "Evidence that Spt2/Sin1, an HMG-like factor, plays roles in transcription elongation,

chromatin structure, and genome stability in Saccharomyces cerevisiae". In: *Molecular and Cellular Biology* 26.4, pp. 1496–1509.

O'Neill, John S et al. (2020). "Eukaryotic cell biology is temporally coordinated to support the energetic demands of protein homeostasis". In: *Nature communications* 11.1, pp. 1–11.

Ostergaard, Simon, Lisbeth Olsson, et al. (2000). "Increasing galactose consumption by Saccharomyces cerevisiae through metabolic engineering of the GAL gene regulatory network". In: *Nature biotechnology* 18.12, pp. 1283–1286.

Ostergaard, Simon, Kristian O Walløe, et al. (2001). "The impact of GAL6, GAL80, and MIG1 on glucose control of the GAL system in Saccharomyces cerevisiae". In: *FEMS yeast research* 1.1, pp. 47–55.

Pelechano, Vicent, Wu Wei, and Lars M Steinmetz (2013). "Extensive transcriptional heterogeneity revealed by isoform profiling". In: *Nature* 497.7447, pp. 127–131.

Perez-Samper, Gemma et al. (2018). "The crabtree effect shapes the Saccharomyces cerevisiae lag phase during the switch between different carbon sources". In: *MBio* 9.5, e01331–18.

Phatnani, Hemali P and Arno L Greenleaf (2006). "Phosphorylation and functions of the RNA polymerase II CTD". In: *Genes & development* 20.21, pp. 2922–2936.

Proudfoot, Nick J, Andre Furger, and Michael J Dye (2002). "Integrating mRNA processing with transcription". In: *Cell* 108.4, pp. 501–512.

Qu, Xiangping et al. (2007). "The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA

3-end processing". In: *Journal of Biological Chemistry* 282.3, pp. 2101–2115.

Rao, Aliz R and Matteo Pellegrini (2011). "Regulation of the yeast metabolic cycle by transcription factors with periodic activities". In: *BMC systems biology* 5.1, pp. 1–15.

Reddy, K Ramesh and Ronald D DeLaune (2008). *Biogeochemistry of wetlands: science and applications*. CRC press.

Ren, Fanggang et al. (2020). "Alternative Polyadenylation: A new frontier in post transcriptional regulation". In: *Biomarker Research* 8.1, pp. 1–10.

Richard, Patricia and James L Manley (2009). "Transcription termination by nuclear RNA polymerases". In: *Genes & development* 23.11, pp. 1247–1269.

Rodrıguez-Molina, Juan B et al. (2022). "Mpe1 senses the binding of pre-mRNA and controls 3 end processing by CPF". In: *Molecular Cell*.

Rosenfeld, Eric and Bertrand Beauvoit (2003). "Role of the non-respiratory pathways in the utilization of molecular oxygen by Saccharomyces cerevisiae". In: *Yeast* 20.13, pp. 1115–1144.

Sandberg, Rickard et al. (2008). "Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites". In: *Science* 320.5883, pp. 1643–1647.

Schlackow, Margarita et al. (2013). "Genome-wide analysis of poly (A) site selection in Schizosaccharomyces pombe". In: *Rna* 19.12, pp. 1617–1631.

Schreieck, Amelie et al. (2014). "RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7". In: *Nature structural & molecular biology* 21.2, pp. 175–179.

Schüller, Hans-Joachim (2003). "Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae". In: *Current genetics* 43.3, pp. 139–160.

Schuster, Samantha L and Andrew C Hsieh (2019). "The untranslated regions of mRNAs in cancer". In: *Trends in cancer* 5.4, pp. 245–262.

Schwerk, Johannes and Ram Savan (2015). "Translating the untranslated region". In: *The Journal of Immunology* 195.7, pp. 2963–2971.

Shannon, Karen W and Jesse C Rabinowitz (1988). "Isolation and characterization of the Saccharomyces cerevisiae MIS1 gene encoding mitochondrial C1-tetrahydrofolate synthase." In: *Journal of Biological Chemistry* 263.16, pp. 7717–7725.

Shashkova, Sviatlana et al. (2017). "The yeast Mig1 transcriptional repressor is dephosphorylated by glucose-dependent and-independent mechanisms". In: *FEMS microbiology letters* 364.14.

Sheets, Michael D, Stephen C Ogg, and Marvin P Wickens (1990). "Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro". In: *Nucleic acids research* 18.19, pp. 5799–5805.

Shepard, Peter J et al. (2011). "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq". In: *Rna* 17.4, pp. 761–772.

Sieber, Patricia, Matthias Platzer, and Stefan Schuster (2018). "The definition of open reading frame revisited". In: *Trends in Genetics* 34.3, pp. 167–170.

Song, Sungmin et al. (2008). "E2-25K/Hip-2 regulates caspase-12 in ER stress–mediated Aβ neurotoxicity". In: *The Journal of cell biology* 182.4, pp. 675–684.

Stajich, Jason E, Fred S Dietrich, and Scott W Roy (2007). "Comparative genomic analysis of fungal genomes reveals intron-rich ancestors". In: *Genome biology* 8.10, pp. 1–13.

Szostak, Emilia and Fátima Gebauer (2013). "Translational control by 3-UTR-binding proteins". In: *Briefings in functional genomics* 12.1, pp. 58–65.

Tadauchi, Tomofumi et al. (2001). "Post-transcriptional regulation through the HO 3-UTR by Mpt5, a yeast homolog of Pumilio and FBF". In: *The EMBO Journal* 20.3, pp. 552–561.

Takagaki, Yoshio and James L Manley (1998). "Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation". In: *Molecular cell* 2.6, pp. 761–771.

Takagaki, Yoshio, Rebecca L Seipelt, et al. (1996). "The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation". In: *Cell* 87.5, pp. 941–952.

Thapar, Roopa (2015). "Roles of prolyl isomerases in RNA-mediated gene expression". In: *Biomolecules* 5.2, pp. 974–999.

Tian, Bin and Joel H Graber (2012). "Signals for pre-mRNA cleavage and polyadenylation". In: *Wiley Interdisciplinary Reviews: RNA* 3.3, pp. 385–396.

Tian, Bin and James L Manley (2013). "Alternative cleavage and polyadenylation: the long and short of it". In: *Trends in biochemical sciences* 38.6, pp. 312–320.

— (2017). "Alternative polyadenylation of mRNA precursors". In: *Nature reviews Molecular cell biology* 18.1, pp. 18–30.

Tu, Benjamin P, Andrzej Kudlicki, et al. (2005). "Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes". In: *Science* 310.5751, pp. 1152–1158.

Tu, Benjamin P, Rachel E Mohler, et al. (2007). "Cyclic changes in metabolic state during the life of a yeast cell". In: *Proceedings of the National Academy of Sciences* 104.43, pp. 16886–16891.

Tudek, Agnieszka et al. (2021). "Global view on the metabolism of RNA poly (A) tails in yeast Saccharomyces cerevisiae". In: *Nature communications* 12.1, pp. 1–14.

Turner, Rachael E and Traude H Beilharz (2021). "Seeking a Role for Translational Control by Alternative Polyadenylation in Saccharomyces cerevisiae". In: *Microorganisms* 9.9, p. 1885.

Turner, Rachael Emily et al. (2021). "Genetic and pharmacological evidence for kinetic competition between alternative poly (A) sites in yeast". In: *Elife* 10.

Van Hoek, PIM, Johannes P Van Dijken, and Jack T Pronk (1998). "Effect of specific growth rate on fermentative capacity of baker's yeast". In: *Applied and environmental microbiology* 64.11, pp. 4226–4233.

Venturelli, Ophelia S et al. (2015). "Population diversification in a yeast metabolic program promotes anticipation of environmental shifts". In: *PLoS biology* 13.1, e1002042.

Wang, Eric T, Neal AL Cody, et al. (2012). "Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins". In: *Cell* 150.4, pp. 710–724.

Wang, Eric T, Rickard Sandberg, et al. (2008). "Alternative isoform regulation in human tissue transcriptomes". In: *Nature* 456.7221, pp. 470–476.

Wang, Ming et al. (2018). "The PUF protein family: overview on PUF RNA targets, biological functions, and post transcriptional regulation". In: *International journal of molecular sciences* 19.2, p. 410.

Wang, Yanchang et al. (2003). "Exit from exit: resetting the cell cycle through Amn1 inhibition of G protein signaling". In: *Cell* 112.5, pp. 697–709.

Ward, Amanda J and Thomas A Cooper (2010). "The pathobiology of splicing". In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 220.2, pp. 152–163.

Warminski, Marcin et al. (2017). "Amino-functionalized 5 cap analogs as tools for site-specific sequence-independent labeling of mRNA". In: *Bioconjugate Chemistry* 28.7, pp. 1978–1992.

Wei, Jun and Fred Sherman (2004). "Sue1p is required for degradation of labile forms of altered cytochromes C in yeast mitochondria". In: *Journal of Biological Chemistry* 279.29, pp. 30449–30458.

Xia, Zheng et al. (2014). "Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3-UTR landscape across seven tumour types". In: *Nature communications* 5.1, pp. 1–13.

Yague-Sanz, Carlo et al. (2020). "Nutrient-dependent control of RNA polymerase II elongation rate regulates specific gene expression programs by alternative polyadenylation". In: *Genes & development* 34.13-14, pp. 883–897.

Yang, Xiaochen et al. (2018). "Genome-wide profiling reveals cancer-related genes with switched alternative polyadenylation sites in colorectal cancer". In: *OncoTargets and therapy* 11, p. 5349.

Yuan, Fuwen et al. (2021). "Alternative polyadenylation of mRNA and its role in cancer". In: *Genes & Diseases* 8.1, pp. 61–72.

Zhao, Jing, Linda Hyman, and Claire Moore (1999). "Formation of mRNA 3 ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis". In: *Microbiology and molecular biology reviews* 63.2, pp. 405–445.

Zheng, Dinghai et al. (2018). "Cellular stress alters 3 UTR landscape through alternative polyadenylation and isoform-specific degradation". In: *Nature communications* 9.1, pp. 1–14.

Zhu, Hui et al. (2007). "Hu proteins regulate polyadenylation by blocking sites containing U-rich sequences". In: *Journal of Biological Chemistry* 282.4, pp. 2203–2210.

Ziv, Naomi, Nathan J Brandt, and David Gresham (2013). "The use of chemostats in microbial systems biology". In: *JoVE (Journal of Visualized Experiments)* 80, e50168.