

Profiling the Landscape of Long Non-Coding RNAs in the Chinese Hamster Ovary Cell Transcriptome

A thesis submitted for the degree of PhD

Dublin City University



By

Krishna C. Motheramgari, BSc

The work described in this thesis was performed under the
supervision of

Dr. Paula Meleady and Dr. Colin Clarke (NIBRT)

National Institute for Cellular Biotechnology

School of Biotechnology

Dublin City University, August 2020

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ (Candidate) ID No.: 16210474

Date: _____

***This thesis is dedicated to Amma, Dad, and Butti
For being there for me, though oceans bar the way***

ACKNOWLEDGEMENTS

I'd first like to thank my advisor, Dr. Colin Clarke, for his support, guidance, and most of all, his patience. I've learned so much from him and this thesis is only possible with his mentorship. The lessons I've learned will guide me in my professional and personal life for years to come. It was an honor to be his student! I was also fortunate to have guidance from Dr. Niall Barron and Dr. Paula Meleady and I am eternally grateful for their help and support.

I'd also like to thank everyone in the Clarke lab, past and present, for their help and good humor in dealing with me over the years. This thesis is in part dedicated to Dr. Craig Monger who was always there when I needed to talk through my problems and help find solutions. I'd also like to thank Dr. Ioanna Tzani for her bemusement and help understanding biology. This thesis would not have been possible without the help and collaboration of Dr. Paul Kelly. This computational biologist would have been incredibly out of his depth without their help! Marina, Ryan, and Andrew for keeping the work lively!

The work presented in this thesis was a collaboration of the eCHO International Training Network. I'd like to honor all of our hearts (and livers) for the sacrifice they've made keeping us ESRs going! Thanks in particular to Ricardo and Prashant who talked me down from the ledge and kept my spirits up when things were at their darkest.

To Drs. Kerim Kaylan, Gargi Datta, and Mandy Kain: I get to join your august company very soon!

Contents

Abbreviations	xxi
Publications	xxiii
Abstract: Profiling the landscape of long non-coding RNAs in the Chinese hamster ovary cell transcriptome.....	1
Chapter 1 Literature Review	2
1 Introduction	3
1.1 CHO cells and biopharmaceutical production.....	3
1.2 Improving cultivation strategies in order to increase cell growth, viability, and productivity	7
1.3 Understanding CHO cell biology and the promise of rational cell engineering.....	9
1.3.1 Gene knockouts	10
1.3.2 Gene over-expression	11
1.3.3 Small non-coding RNA engineering	11
1.4 High-throughput transcriptomics analysis.....	14
1.4.1 Microarrays	14
1.4.2 Next Generation Sequencing	15
1.4.3 Illumina RNA-Seq Library Preparation	19
1.4.4 Single-cell transcriptomics	20
1.5 Long non-coding RNAs.....	21
1.5.1 Long non-coding RNA classification	22
1.5.2 Transcriptional and translational regulation by lncRNAs.....	29
1.5.2.1 Epigenetic modification of chromatin states to regulate transcription.....	31
1.5.2.2 Transcription factor and miRNA decoys and mimics	32
1.5.2.3 Translational regulation by long non-coding RNAs	33
1.5.3 State of the annotation of lncRNAs in the <i>C. griseus</i> and CHO genomes	34
1.6 Challenges in profiling long non-coding RNAs	39
1.7 The promises of lncRNA based genetic engineering.....	41
1.8 Project Aims.....	44
Chapter 2 Differential Expression of lncRNAs Following Temperature Shift	46
2 Summary	47
2.1 Introduction	48
2.2 A workflow for lncRNA discovery	50
2.2.1 Constructing gene and transcript models from NGS data.....	50
2.2.2 Coding potential calculators	52
2.2.2.1 Coding Potential Calculator 2	53
2.2.2.2 Coding Potential Assessment Tool.....	54

2.2.2.3 FIExible Extraction of lncRNAs	55
2.2.4 Feature based filtering of coding sequences	57
2.2.5 Classification of lncRNAs with respect to genome localization	60
2.3 Materials and Methods	62
2.3.1 CHO Cell culture and temperature shift	62
2.3.2 Total RNA library preparation and sequencing	62
2.3.3 lncRNA discovery pipeline	63
2.3.3.1 lncRNA discovery pipeline	65
2.3.3.2 Read alignment and transcriptome assembly	65
2.3.3.3 lncRNA discovery and annotation	66
2.3.3.4 Removal of transcripts with coding domains and homology to small ncRNAs	66
2.3.3.5 Syntenic lncRNA sequence identification	67
2.3.4 Differential expression of lncRNAs.....	67
2.3.5 Metabolite profiling between NTS and TS samples	68
2.4 Results	69
2.4.1 Temperature shift induces growth arrest in CHO cells.....	69
2.4.2 Pre-processing of RNA-Seq data and alignment to the CHOK1 genome	70
2.4.3 Reference guided assembly metrics	71
2.4.4 Assembly metrics of long non-coding RNA	71
2.4.5 Comparison of characteristics between mRNA and lncRNA loci	73
2.4.6 Isoform and exon distributions between lncRNA and mRNA genes	76
2.4.7 Identification of syntenic lncRNAs	78
2.4.8 Differential gene expression after a reduction in culture temperature.....	79
2.4.9 Enriched functional pathways of differentially expressed protein coding genes	83
2.4.10 Significantly differentially expressed evolutionarily conserved lncRNAs.....	83
2.4.11 Genome-wide differential expression of antisense transcription	84
2.4.12 Sense/antisense differential expression.....	85
2.4.13 Alterations in metabolite concentration after a reduction in culture temperature	89
2.5 Discussion	90
2.5.1 Comparison of annotation to GENCODE	90
2.5.2 Differences in isoform and exon composition to GENCODE standard	91
2.5.3 Temperature shift induced global change in gene expression	92
2.5.4 Evolutionarily conserved lncRNAs differentially expressed after TS.....	93
2.5.5 Sense/antisense differential expression.....	95
2.5.6 Limitations of analysis using lncRNA-Seq data	97
2.6 Conclusions	98

Chapter 3 Differential Expression of lncRNAs During Fed-Batch Culture	99
3 Summary	100
3.1 Introduction	101
3.2 Materials and Methods	107
3.2.1 CHO cell culture within industrial 10L stirred tank bioreactors	107
3.2.1 RNA-Seq library preparation	108
3.2.3 RNA-Seq library QC	109
3.2.4 Differential expression between cell culture phases	109
3.2.1 Correlation of expression to bioprocess variables	109
3.3 Results	111
3.3.1 CHO cell culture within industrial 10L stirred tank bioreactors	111
3.3.2 RNA-Seq data QC and lncRNA classification	112
3.3.3 Reference guided assembly metrics	113
3.3.4 Comparison of characteristics between lncRNA and mRNA loci	115
3.3.5 Differentially expressed mRNAs between exponential and stationary phases	116
3.3.6 Differentially expressed mRNAs between stationary and death phases	118
3.3.7 Differentially expressed mRNAs over time	120
3.3.8 Differentially expressed lncRNAs during exponential and stationary phases	122
3.3.9 Differentially expressed lncRNAs during death phase	126
3.3.10 Modulation of sense/antisense transcription over time	128
3.3.11 Metabolite profiling in 10L stirred tank bioreactors	130
3.3.12 Correlation of bioprocess parameters and metabolite concentration to differential expression	134
3.4 Discussion	140
3.4.1 Fed-batch culture parameters are concordant with previous CHO cell studies	140
3.4.2 Differentially expressed genes between culture phases	142
3.4.3 NEAT1 is upregulated during death phase	143
3.4.4 Sense/antisense differential expression over culture phases	144
3.4.5 Pearson's correlation between gene expression and process parameters	145
3.4.6 Correlation of metabolite concentration to gene expression	146
3.4.6.1 Glucose concentration did not significantly correlate to changes in gene expression	146
3.4.6.2 Cohort of DE lncRNAs that correlated to lactate consumption	147
3.4.6.3 Glutamine consumption and ammonia production correlations	148
3.4.7 Potential utility of identified lncRNAs	149
3.5 Conclusions	151
Chapter 4 Alternative Splicing in Fed-Batch Culture	152

4 Summary	153
4.1 Introduction	154
4.2 Materials and Methods	160
4.2.1 Transcript quantification using StringTie	160
4.2.2 Isoform switching and alternative splicing	160
4.2.3 Inferring biological significance of isoform switching	160
4.3 Results	162
4.3.1 Isoform and gene structure of lncRNA and mRNA loci	162
4.3.2 Transcript level quantification and quality control	163
4.3.3 mRNAs undergoing isoform switching between exponential and stationary phase	164
4.3.4 mRNA alternative splicing events between exponential and stationary phase	167
4.3.4 mRNA isoform switching between stationary and death phases	170
4.3.6 Alternative splicing events between stationary and death phases	174
4.3.7 Alternative splicing of mRNAs over time	176
4.3.8 lncRNAs undergoing isoform switching between exponential and stationary phase	179
4.3.9 lncRNA alternative splicing between exponential and stationary phase	181
4.3.10 lncRNA isoform switching between stationary and death phase	184
4.3.11 lncRNA alternative splicing events between stationary and death phases	186
4.3.12 lncRNA alternative splicing over the duration of fed-batch culture	189
4.4 Discussion	193
4.4.1 Isoform composition of lncRNA gene loci	193
4.4.2 Comparison of differential gene expression to alternative splicing	194
4.4.3 Isoform switching in genes undergoing differential expression	195
4.4.4 Isoform pulsing over culture period	195
4.4.5 lncRNA alternative splicing	196
4.4.6 Gene ontology enrichment of alternatively spliced mRNAs	196
4.4.7 First evidence of lncRNA isoform switching in CHO cells	197
4.4.8 Isoform switching can be used to quantify and classify different alternative splicing events	198
4.4.9 Exon skipping is the most common alternative splicing event in both mRNA and lncRNA genes	199
4.4.10 Isoform switching independent of gene level differential expression	200
4.4.11 Isoform switching over the duration of fed-batch culture	201
4.4.12 Alternative splicing induces changes to lncRNA secondary structure	201
4.5 Conclusions	204
Chapter 5 Change in Chromatin Accessibility Following Temperature Shift	205

5 Summary	206
5 Introduction.....	207
5.1 High-throughput sequencing methods to profile chromatin accessibility	213
5.2 Computational methods to profile chromatin accessibility with ATAC-Seq	216
5.2.1 Software packages for peak calling	216
5.2.2 Differential chromatin accessibility	220
5.2.3 Motif enrichment analysis.....	220
5.3 Profiling chromatin states for CHO cell line engineering	224
5.4 Materials and Methods	226
5.4.1 CHO cell culture and temperature shift	226
5.4.2 ATAC-Seq library preparation and sequencing.....	226
5.4.3 Bioinformatics data pre-processing and alignment	226
5.4.4 Evaluating peak calling algorithms	227
5.4.5 Accessibility analysis of chromatin using annotated CHIP-Seq data	228
5.4.6 Identification of active transcription with ATAC-Seq and RNA-Seq	228
5.4.7 Identification of stable transgene integration sites	229
5.4.8 Calculating differential accessibility.....	229
5.4.9 Identifying enriched TF binding sites motifs	230
5.5 Results	231
5.5.1 CHO cell culture and library preparation	231
5.5.2 NGS data pre-processing.....	232
5.5.3 Evaluation of peak calling algorithms	233
5.5.4 Classification of peaks	234
5.5.5 Active promoter and enhancer histone marks.....	235
5.5.6 Integrated RNA-Seq and ATAC-Seq to find TSS with active transcription	236
5.5.7 Identification of stable integration sites using ATAC-Seq data	239
5.5.8 Global change in chromatin accessibility after temperature shift.	240
5.5.9 Comparison between differentially accessible peaks and RNA-Seq	242
5.5.10 Motif enrichment with <i>HOMER2</i>	244
5.5.11 Total deviation of transcription factor motifs with <i>chromVAR</i>	250
5.6 Discussion	252
5.6.1 ATAC-Seq quality control metrics from sequencing data.....	253
5.6.2 Evaluation of two peak calling algorithms	253
5.6.3 Regulatory context of open chromatin regions remains difficult.....	254
5.6.4 Active chromatin states within previously annotated regulatory regions	255
5.6.5 Application of ATAC-Seq to estimate active transcription at the TSS	256

5.6.6 Combined ATAC-Seq and RNA-Seq data to identify stable sites for transgene integration	256
5.6.7 Temperature shift induces global alterations in chromatin accessibility.....	257
5.6.8 Enrichment of TF binding sites in DA peaks	257
5.6.9 Global alterations in <i>AP-1</i> binding sites	258
5.6.10 <i>TEAD4</i> has increased average accessibility after temperature shift.....	259
5.6.11 Kruppel like factors are highly accessible after temperature shift	260
5.7 Conclusions	263
Chapter 6 Overall Conclusions and Future Work.....	264
6.1 An improved lncRNA reference annotation using RNA-Seq.....	265
6.2 Future work	267
6.2.1 <i>In vitro</i> validation of differentially expressed lncRNAs	267
6.2.2 Validation of lncRNAs that are correlated with metabolite concentration.....	268
6.2.3 Characterization of the impact of alternative splicing of lncRNAs on CHO cells.....	268
6.2.4 Integration of miRNA-Seq data to identify ceRNA networks	269
6.2.5 Integration of bisulfite sequencing data to identify epigenetic remodeling complexes	269
6.2.6 Validation of differential accessibility of transcription factor binding site motifs	270
6.2.7 Validating landing pads using open chromatin maps.....	270
6.2.8 Resources for future cell line engineering research	271
6.3 Overall Conclusions	272
6.3.1 lncRNA detection pipeline	272
6.3.2 Characteristics of assembled lncRNAs	272
6.3.3 lncRNAs are differentially expressed in industrially relevant conditions	273
6.3.4 Alterations in lncRNA abundance correlated with changes in metabolite concentration and bioprocess parameters.....	273
6.3.5 Alternative splicing of lncRNAs in fed-batch culture.....	274
6.3.6 Global change in chromatin accessibility in protein coding and long non-coding genes after temperature shift	274
6.3.7 Identification of active regulatory sites using multi-omics data	275
6.3.8 Integration of ATAC-Seq with RNA-Seq to discover highly accessible landing pads	276
Bibliography.....	277

Table of Figures

Chapter 1

Figure 1.1 Plot illustrating the total number of drugs approved. A total of 353 unique drugs have been approved to date (Walsh, 2018). The largest class of approved recombinant therapeutic proteins is monoclonal antibodies which are commonly produced in CHO cells. 6

Figure 1.2 Plot of average titres (g/L) over time as reported by Rader and Langer (Rader & Langer, 2014). These data were obtained by questionnaires and a survey of the literature by the authors and tabulated from 2006-2014. Rader and Langer extrapolated titres prior to 2006 from published data and after 2014 by predictions from trends. 8

Figure 1.3 A diagram of paired-end Illumina Sequencing by Synthesis protocol. (i) RNA molecules are extracted from cell lysate and purified and (ii) forward and reverse adaptor sequences are added (iii) Loaded libraries bind to complementary probes fixed to the flow cell (v) SBS is performed with the incorporation of fluorescently labelled ddNTPs (v) with amplified sequences for both forward and reverse (vi) Fluorescently labelled ddNTPs are excited with a laser after each cycle and imaged. Each incorporated ddNTP excites at a different wavelength. Image processing software then converts snapshots into ACGT nucleotide characters and reads. 17

Figure 1.4 A plot of sequencing cost per megabase from 2001 to 2015. Sequencing costs were tracked over time by NIH for its subsidiary sequencing centres. 18

Figure 1.5 Classification of lncRNAs by genomic localization. The classifications are detailed in Table 1.3. Error! Bookmark not defined.

Figure 1.6 GENCODE classification of genes into respective categories. These data were retrieved from release 32 of the human and release 32 of the mouse annotation. 26

Figure 1.7 A diagram detailing the diverse regulatory mechanisms performed by lncRNAs within the nucleus and the cytoplasm in eukaryotic organisms. 30

Figure 1.8 Modular design of antisense SINEUP transcripts. The binding domain is thought to form Watson-Crick base pairs with the target mRNA in the 5' end. The binding domain should overlap the 5' UTR and encompass the AUG translation start site of the CDS sequence. The binding domain connected the inverted SINEB2 translational effector domain by a linker sequence. 42

CHAPTER 2

Figure 2.1 StringTie assembles transcripts using a splice graph with a network flow model where dominant isoforms are first constructed using the regions of highest coverage. The graph model is further refined using regions of lower coverage which are assumed to be other isoforms within the gene. 52

Figure 2.2 A figure summarizing cultivation conditions of 8 replicate cultures. Cells were initially seeded and grown at 37°C for a period of 48 hours. Four replicate cultures were subjected to a reduction in culture temperature to 31°C while the remaining four replicates remained at 37°C. RNA was extracted 72 hours after inoculation. 62

Figure 2.3 A flowchart diagram detailing analysis steps of our lncRNA annotation pipeline from raw data QC to lncRNA classification. Reads are initially aligned to the ENSEMBL version 98

criGri1 reference genome and assembled into 58,488 transcripts. (A) Transcripts with coding domains and homology to small ncRNAs and high coding potential are removed. (B) 6,722 putative lncRNAs are compared to RFAM and GENCODE human and mouse annotations in order to find conserved lncRNAs. 64

Figure 2.4 A plot of average cell density over time. Cells were sampled over 24 hour intervals. As expected, cells that underwent a reduction in culture temperature had a lower average cell density as they proliferated at a lower rate when compared to those cells cultured at 37°C. Total RNA was harvested from all samples after 72 hours. 70

Figure 2.5 A bar plot comparing the percentage of all lncRNA genes found in sense, antisense, and intergenic classes. The human and mouse lncRNA annotations were derived from release 32 and 23 for human and mouse respectively. Intergenic lncRNAs make up the largest classes in all three species but are the dominant class in CHO accounting for ~75%. The percentage of each class of human and mouse lncRNAs are approximately concordant between the two species. 72

Figure 2.6 Class cation of lncRNAs into respective classes with respect to localization to a protein coding gene. Intergenic and antisense genes were the most abundant classes of lncRNAs. ... 73

Figure 2.7 A density plot comparing transcript lengths between coding and long non-coding RNAs. The distribution of lncRNA transcript lengths was found to be significantly shorter with a median transcript length of 649bp compared to a median mRNA transcript length of 1,931bp which is 3X longer. 74

Figure 2.8 A density plot summarizing the GC content distribution of coding and long non-coding transcripts. mRNA GC content was observed to be slightly higher with a median of 49.5%. 75

Figure 2.9 A density plot comparing the distribution of normalized TPM abundances between lncRNA and mRNA transcripts. Coding transcripts had a wider distribution of expression and were found to be more abundant than lncRNAs. 76

Figure 2.10 A bar plot summarizing the percentage of total lncRNA and mRNA genes. Single isoform genes dominated both lncRNA and coding genes with the majority of all lncRNAs composed of genes with a single transcript. 77

Figure 2.11 A bar plot summarizing the distribution of the number of exons for coding and long non-coding transcripts. The majority of lncRNA transcripts had 3 or fewer exons while the coding transcripts had an even distribution of exons in each transcript..... 78

Figure 2.12 Venn diagram of conserved lncRNAs found homologous to BLAST searches within Rfam and cross-species syntenic loci. 79

Figure 2.13 A PCA plot summarizing the variance between samples and replicates. 68.19% of the variance could be captured by the first principal component which corresponds to culture temperature. 80

Figure 2.14 A volcano plot summarizing significant-fold-change between cells cultivated at 37°C and 31°C. A gene was considered significantly differentially expressed if it had a-fold-change $\geq \pm 1.5$ and a BH adjusted P-value < 0.05 . An even distribution of genes were found to be upregulated (blue) and downregulated (red) after a reduction in culture temperature. 81

Figure 2.15 A coverage plot comparing average TPM normalized coverage between NTS (red) and TS (blue). A 2.2 fold increase in expression of the CIRBP gene was observed at 31°C. ... 82

Figure 2.16 A bar plot of significantly differentially expressed lncRNAs segregated into different classes of lncRNAs. 82

Figure 2.17 A coverage plot comparing average TPM normalized coverage between NTS (red) and TS (blue) of the HOTAIRM1 gene. A 2.08 fold increase in expression of the HOTAIRM1 gene was observed at 31°C 84

Figure 2.18 A coverage plot comparing average TPM normalized coverage between NTS (red) and TS (blue) of the AIG1-AS gene. A 12.02 fold increase in expression of the AIG1-AS gene was observed at 31°C..... 85

Figure 2.19 A coverage plot comparing average sense and antisense TPM normalized coverage between NTS (red) and TS (blue) of the Timeless locus which contains the Timeless gene on the sense strand and a mono-exonic ncRNA on the antisense. Both the sense and antisense were downregulated after a reduction in culture temperature..... 86

Figure 2.20 A coverage plot comparing average sense and antisense TPM normalized coverage between NTS (red) and TS (blue) of the Oxct1 locus which contains the Oxct1 gene on the sense strand and a multi-exonic ncRNA on the antisense. The antisense gene is strongly upregulated while the sense protein coding gene is significantly downregulated after temperature shift. 87

Figure 2.21 Concentrations of lactate, ammonia (NH_4^+ ions), and glutamine in mmol/L between NTS and TS samples. Lactate and ammonia concentrations were reduced between TNS and TS samples while glutamine concentrations were significantly increased..... 89

CHAPTER 3

Figure 3.1 A summary of the glycolysis and TCA cycle pathways. Glucose is first metabolized to pyruvate for conversion in the acetyl coa. Aerobic respiration occurs inside the mitochondria during the citric acid cycle which generates ATP. Rapid glycolysis can lead to the accumulation of lactate which can inhibit CHO cell growth. Glutamine can also be a source of energy as well as nitrogen. 102

Figure 3.2 Viable cell concentrations over a 14 day period in two 10L bioreactors. Clear, lag, exponential, stationary, and death phases were observed in each bioreactor. Variance between replicate bioreactors was low in lag and exponential phases but started to diverge during the stationary phase. Curiously, the viable cell concentrations were observed to recover slightly at day 13 with a final VCC of $\sim 13.3 \times 10^6$ cells/mL 111

Figure 3.3 Percentage of viable cells over 14 day culture period in two replicate bioreactors. The viability was highly concordant between replicates. The dashed line in the plot depicts the time point where viability started to significantly decrease with a final viability of $\sim 70\%$ in both reactors. 112

Figure 3.4 Boxplots comparing the distribution of percentage of reads aligning to the synthetic rRNA probe sequences. Bioreactor 2 had a slightly lower median alignment rate than bioreactor 1 and samples from day 13 and 14 were deemed insufficient for downstream analysis. 113

Figure 3.5 Classification of genes into categories based on proportion of total assembly. These three categories are protein coding, long non-coding, and miscellaneous small ncRNAs..... 114

Figure 3.6 Number of transcripts for each genic and intergenic class of lncRNA transcripts. Long intergenic ncRNAs are the most abundant class followed by antisense RNAs overlapping protein

coding genes. Sense and antisense loci that are found upstream and downstream as well as divergent transcripts that are transcribed in a “head to tail” manner from coding genes compose a smaller proportion of long non-coding RNAs..... 115

Figure 3.7 (A) The distribution of transcript lengths between mRNAs and lncRNAs. lncRNAs, were on average, shorter than mRNAs but also with a wider variation in transcript lengths. (B) A density plot depicting the $\log_2$ normalized TPM abundances between coding transcripts and lncRNAs with coding transcripts expressed at higher abundances with a wider distribution of TPM values compared to non-coding transcripts. (C) A density plot comparing % GC of coding and non-coding transcripts. lncRNA transcripts had an average GC content lower than mRNAs. . 116

Figure 3.8 Average normalized expression of the *BAX* gene. The gene starts to be upregulated after day 5 and becomes significantly upregulated during the stationary phase. *BAX* does not have a change in average gene abundance during the death phase. 117

Figure 3.9 Average normalized expression of the *BCL2* gene. *BCL2* gene average normalized gene abundance begins to decrease after day 5 and begins to linearly increase during the stationary phase. 119

Figure 3.10 Average normalized abundance of the *Angiotensinogen* gene. *AGT* expression was quiescent until day 7 when gene expression started to significantly increase into the death phase. 121

Figure 3.11 A leaf plot of up and downregulated lncRNAs between exponential and stationary phases. There were 32 genes that were assigned to multiple classifications and are represented within the leaf plot..... 122

Figure 3.12 Average normalized abundance of the *LINC00536* non-coding gene. *LINC00536* gene expression was observed to increase from day 5. 123

Figure 3.13 Average normalized abundance of the novel *MSTRG.11806* ncRNA gene. Gene expression was at a maximum at day 4 in the exponential phase where abundances steadily decrease until day 10. 124

Figure 3.14 Sense and antisense normalized abundances of the *SCFD1* coding gene and the ncRNA partner. 125

Figure 3.15 A leaf plot of up and downregulated lncRNAs between stationary and death phases. There were 19 genes that were assigned to multiple classifications and are represented within the leaf plot..... 126

Figure 3.16 Average normalized abundances over time for the *TMC1-AS* ncRNA. *TMC1-AS* is modestly differentially expressed between individual days but is significantly downregulated after day 9 at the onset of apoptosis 127

Figure 3.17 Average normalized abundances over time for the conserved *NEAT1* gene. *NEAT1* is expressed at very high abundances at a constant level until day 10 when the ncRNA becomes significantly upregulated from day 11 and 12..... 127

Figure 3.18 Average normalized abundance of the sense *NAV2* gene and the antisense lncRNA gene over time. Both genes are expressed at similar abundances until day 5 when the antisense ncRNA becomes significantly upregulated and becomes dominant during the stationary phase. 129

Figure 3.19 A plot of sense and antisense coverage of the <i>NAV2</i> gene composed of transcript <i>ENSCGRT00000018790</i> and the mono-exonic antisense <i>NAV2</i> gene which contains a single transcript (<i>MSTRG.7889.1</i>) on day 5 and day 10. The sense gene is dominant during day 5 but becomes depleted during the death phase on day 10	128
Figure 3.20 Glucose concentrations were quantified before and after feed daily. This plot only displays concentrations prior to feed for reasons of clarity and in order to visualize glucose consumption over time	130
Figure 3.21 Lactate concentrations were profiled daily over a 14 day periods. Concentrations were the highest at day 3 just prior to media supplementation where consumption was observed after day 5.	131
Figure 3.22 Glutamine concentrations over 14 day culture period. Glutamine was only a component of the basal media and was not supplemented during the feeding. Glutamine is completely depleted after day 4 but concentrations begin to increase during the exponential, stationary and death phases.	132
Figure 3.23 Glutamate concentrations in mmol/L over time. Glutamate was a component of the feed media and was routinely added during the fed-batch process. Glutamate is completely depleted during the stationary phase but concentrations begin to increase during the death phase.	133
Figure 3.24 Ammonia concentrations in mmol/L over time. Ammonia concentrations began to increase during the lag phase but were reduced during the exponential and stationary phases. Ammonia concentrations increased linearly during the death phase with the highest concentration on days 13 and 14.	133
Figure 3.25 Regression plot comparing the % viability to average normalized abundance of the <i>RBM3</i> gene over time. There was a strong correlation between the two as <i>RBM3</i> gene abundance decreased proportionally to decrease in % viability.	135
Figure 3.27 Average normalized gene abundances over time for the <i>MSTRG.2854</i> intergenic ncRNA. Gene expression was quiescent during the exponential phase but increased during the stationary phase and death phases as viability dropped below 90%.	137
Figure 3.28 Regression plot describing the correlation between % of viable cells and the average normalized gene abundances of the novel <i>MSTRG.2854</i> gene.	138
Figure 3.29 (A) Regression plot of correlation between lactate concentration and average normalized gene abundances of the <i>MSTRG.16951</i> ncRNA which was positively correlated. (B) Negative correlation between the novel <i>MSTRG.18364</i> ncRNA and glutamate concentration and (C) negative correlation between ammonia concentration and intergenic <i>MSTRG.19771</i> abundance.	139

CHAPTER 4

Figure 4.1 A diagram summarizing three alternative splice isoforms generated from a double-stranded gene. The entire gene with corresponding introns and exons are first transcribed into precursor RNA molecules which are processed by the spliceosome into mature transcripts without introns.	155

Figure 4.2 The different classifications of alternative splicing as defined by IsoformSwitchAnalyzeR. Skipped exons (SE) are events where an internal exon is spliced out of the precursor whereas an intron retention (IR) event is when the intron is retained instead of being removed as normal. Mutually exclusive exons (MXE) are preferential splicing of one exon over another between conditions. Alternative 5' (A5) and 3' (A3) splice sites alter the exon boundaries during expression. Alternative start (ATSS) and stop (ATTS) sites alter the boundaries of transcription. 157

Figure 4.3 Proportion of exons per transcript for lncRNAs compared to mRNAs. Dual exonic transcripts were the largest group of lncRNAs with almost 50% of all transcripts containing 2 exons. 162

Figure 4.4 A bar plot showing the number of isoform as a percentage of the total number of coding or long non-coding genes. Single isoform genes account for almost 80% of all coding and lncRNA genes. Almost 95% of both cohorts contain 3 or fewer isoforms. 163

Figure 4.5 (A) Gene abundance over time for the PIGA gene. A 1.5 fold increase in gene expression occurred between the exponential and stationary phase after declining during the start of the death phase. (B) Normalized transcript abundances of the two dominant PIGA component isoforms. The novel isoform MSTRG.8948.7 becomes the most expressed transcript during the stationary phase and its change in expression is reflected in the gene level abundance profile. Three other isoforms exist within the PIGA gene but are not expressed beyond the lag phase and have been omitted in the plot for clarity. After day 3, ENSCGRT00000009387 and MSTRG.8948.7 are the only two expressed isoforms and make up 100% of the abundance. 165

Figure 4.6 (A) Gene abundance over time for the CS gene which was seen to be expressed at a stable abundance over time. 100% of gene expression can be attributed to two novel isoforms: MSTRG.12084.1 and MSTRG.12084.2. (B) The two component isoforms were observed to be dominant during different culture phases with MSTRG.12084.1 expressed the most during exponential phase and MSTRG.12084.2 during the stationary phase. 166

Figure 4.7 A leaf plot showing the number of alternatively spliced transcripts that underwent significant isoform switching between exponential and stationary phase 168

Figure 4.8 (A) Change in isoform fraction over time for the Laccase Domain 1 (LACC1) gene which is composed of two isoforms: a previously annotated transcript (ENSCGRT00000007293) and the novel transcript MSTRG.5985.1 which becomes dominant during the exponential and death phases of culture. (B) A 22% increase in the isoform fraction for MSTRG.5893.1 occurs with a concomitant decrease of 22% for ENSCGRT00000007293 (Adj. P-value: 4.57×10^{-6}) ... 169

Figure 4.9 (A) Gene abundance over time for the CYLBL gene. A 1.54 fold increase in gene expression occurred between the stationary and death phase (B) Normalized transcript abundances of the two dominant CYLBL component isoforms. The novel isoform MSTRG.8775.1 is periodically depleted every two days as the other isoform is slightly upregulated. The novel isoform becomes dominant during the death phase after day 10. 171

Figure 4.10 Gene abundance over time for the ZNF609 gene which was seen to be expressed at a stable abundance over time and was not significantly differentially expressed at any stage. 100% of gene expression can be attributed to two novel isoforms: MSTRG.19775.1 and MSTRG.19775.2. (B) The two component isoforms were observed to have time specific fluctuations in isoform fractions as both transcripts have anti-correlated expression. MSTRG.19775.2 is expressed at 100% of the total fraction in the death phase. 173

Figure 4.11 A leaf plot showing the number of alternatively spliced transcripts that underwent significant isoform switching between exponential and stationary phase.	175
Figure 4.12 (A) Transcript structure of the PALM gene which has one novel isoform and two previously annotated transcripts. The region highlighted in blue is magnified in plot (B) which is a sashimi plot of the last three exons of the PALM locus. Only the switching genes are depicted with MSTRG.8217.1, which contains the skipped exon, increasing 23%.....	176
Figure 4.13 A sashimi plot of the normalized coverage of the UBA2CL gene which contains two isoforms: MSTRG.18663.1 and ENSCGRT00000023929. The longer isoform contains the full length CDS sequence of the UBA2CL protein and is preferentially expressed during the stationary phase but is quiescent during the death phase.	177
Figure 4.14 277 transcripts were differentially expressed between the exponential and stationary phases with the majority belonging to the intergenic class of lncRNAs.	179
Figure 4.15 A leaf plot of significant alternative splicing events between exponential and stationary phase transcripts.	182
Figure 4.16 (A) Change in isoform fraction over time for the intergenic ENSCGRG00000025089 locus which was previously annotated within ENSEMBL. The shorter, novel isoform MSTRG.5005.3 becomes the dominant isoform at the start of the stationary phase (dIF= 13.2% ,Adj. P-value: 0.026). (B) A sashimi plot of the change in expression and splicing of the two isoforms. The multi-exonic transcript MSTRG.5005.3 becomes dominant.	183
Figure 4.17 RNA structure predicted by RNAfold for the longer (A) MSTRG.5005.1 and (B) shorter MSTRG.5005.3 transcripts. The longer isoform has a triangular shape that is lost with the switch to the shorter isoform.	184
Figure 4.18. 139 transcripts were differentially expressed between the stationary and death phases with the majority belonging to the intergenic class of lncRNAs.	185
Figure 4.19 A leaf plot of significant alternative splicing events between stationary and death phase transcripts.	187
Figure 4.20 (A) The change in isoform fraction over time for transcripts MSTRG.16574.1 and MSTRG.16574.3. The isoform fractions of the other three transcripts have been omitted for clarity but remain constant over time as the proportion of these isoform fractions change over time. (B) A sashimi plot of coverage and splicing of the MSTRG.16574 non-coding RNA locus. Exon 2 is completely depleted in the death phase.....	188
Figure 4.21 RNA structure predicted by RNAfold for the longer (A) MSTRG.16574.3 and (B) shorter MSTRG.16574.1 transcripts. The longer isoform contains an additional circular structure with stem-loops that is lost in MSTRG.16574.3 which is missing the second exon.	189
Figure 4.22 (A) The change in isoform fraction of the novel MSTRG.13357 locus which is composed of two isoforms that switch during the stationary phase. (B) A sashimi plot showing an A5 alternative splicing event that leads to a truncated second exon of MSTRG.13357.2.....	191
Figure 4.23 (A) The secondary structures of the MSTRG.13357.1 isoform that is expressed during exponential and death phase but is replaced by (B)MSTRG.13357.2 during the stationary phase.....	192

CHAPTER 5

Figure 5.1 (A) The structure of a nucleosome with 4 core histones coiled around a ~147bp DNA sequence. Histone complexes contain trailing tails which can be modified. (B) Repeating complexes of nucleosomes packaged into chromatin. Nucleosome can be further condensed into transcriptionally inactive heterochromatin or unwound for gene transcription as euchromatin. **208**

Figure 5.2 (A) An illustration of a transcriptional enhancer forming stem-loop structure to bring transcriptional activators to promoter. The enhancer region binds transcription factors in regions that can be very distal to the target promoter. After the activator/promoter complex is formed, RNA transcription by RNA Pol II is increased and transcription is upregulated. (B) An illustration of gene repression by a gene silencer. DNA stem-loops form in a manner similar to enhancers. Silencer regions bind proteins that, when forming a complex with the promoter, repress transcription initiation by RNA Pol II. **211**

Figure 5.3 A diagram detailing the ATAC-Seq library preparation process. DNA from a diluted sample of cells is first extracted using a non-ionic detergent that preserves DNA bound to nucleosomes and heterochromatin. The DNA is then treated with hyperactive Tn5 transposase that preferentially fragments regions devoid of nucleosomes and around nucleosome boundaries. The fragments are then sequenced and aligned to the genome with the shorter fragments originating from NFR regions and fragments ranging from 180-220bp equating with a mononucleosome. Larger nucleosomes will have fragment distributions that are multiples of the mononucleosome distribution. Heterochromatin will not be tagged due to protection of DNA by compacted chromatin. **215**

Figure 5.4 A plot of the average cell densities of 8 cultures 72 hours after inoculation. Eight CHO cultures were grown over a period of 72 hours with four subjected to a temperature shift after 48 hours. As expected, the cell densities of CHO cells that were cultivated at 31°C had an 18% lower cell density than those cultivated at 37°C. DNA from cells was extracted after cell counting for ATAC-Seq library preparation. **231**

Figure 5.5 Fragment length distribution of ATAC-Seq reads. A density plot depicting the distribution of unique, mapped fragments to the criGri1 CHOK1 genome. We observed the majority of reads having fragment lengths less than 147 which corresponds to a nucleosome free region which is indicative of efficient fragmentation. We also observed a mono-nucleosome (1N) peak and a lesser peak corresponding to di-nucleosomes (2N). **233**

Figure 5.6 CPM normalized coverage of coverage of NFR read density compared to nucleosome peak density. A density plot of normalized peak coverage surrounding peak summits plotted as a fraction of the peak maxima. The peak centre corresponds to the nucleosome free region with a lower coverage flanking regions upstream and downstream corresponding to nucleosomes. Approximately 58.36 of all peaks were located within gene bodies (in introns or exons) or in the promoter or TTS. Intergenic peaks found in distal regions compose the largest individual class. **234**

Figure 5.7 Distribution of peaks across the CHOK1 genome with respect to a TSS. Approximately 58.36 of all peaks were located within gene bodies (in introns or exons) or in the promoter or TTS. Intergenic peaks found in distal regions compose the largest individual class. Promoter peaks were defined as those found 1 kb upstream of the TSS or downstream 100 bp. **235**

Figure 5.8 A density plot comparing gene normalized TPM of genes that contain a peak near the TSS and genes that do not. Significantly more genes without a peak near the TSS were found to

have lower average expression. Average expression of genes that contain an ATAC-Seq derived peak was higher. 237

Figure 5.9 Heatmaps comparing average normalized coverage of around the TSS of coding and lncRNA genes. Average accessibility around mRNA TSSs was found to be significantly higher than average accessibility around the TSS of lncRNAs. 238

Figure 5.10 The average normalized coverage plot of ATAC-Seq data spanning the Dag1-Nicn/Amt locus. Coverage has been CPM normalized. A 770bp region of the larger open chromatin region was found to be differentially accessible (highlighted in box). 239

Figure 5.11 The average normalized coverage plot of ATAC-Seq data upstream of the Sacs gene. Coverage has been CPM normalized. The region of the highest peak corresponds to the promoter. 240

Figure 5.12 A principal component analysis plot visualizing variation between samples. 68% of variability is captured by the first principal component which can be explained by changes in accessibility following temperature shift. 241

Figure 5.13 A volcano plot comparing log₂-fold-change with adjusted P-values. We observed genome-wide chromatin remodelling after a reduction in culture temperature with a higher proportion of open chromatin regions compared to closed. 242

Figure 5.14 A heatmap comparing average normalized coverage within 1kb of significantly DA Fosl1 binding sites. Fosl1 sites are the most enriched motif and found equally differentially accessible between open and closed chromatin. 247

Figure 5.15 A heatmap comparing average normalized coverage within 1kb of all significantly DA TEAD4 sites. TEA domain containing peaks are approximately 20% more accessible after a reduction in culture temperature. Tead4 was also found to be 1.55 upregulated within RNA-Seq data. 248

Figure 5.16 A heatmap comparing average normalized coverage within 1kb of all significantly DA Klf4 sites. Kruppel like factors are the largest group of significantly differentially accessible TFs. Klf4 was also found to be 3.38 upregulated within RNA-Seq data. 249

CHAPTER 6

Figure 6.1 Classification of lncRNAs discovered in RNA-Seq data from Chapter 2 and Chapter 3. Intergenic lncRNAs are the most abundant class of transcripts followed by antisense overlapping lncRNAs. There was only a modest increase in the number of sense overlapping transcripts while all other classes were significantly expanded. 266

Table of Tables

CHAPTER 1

Table 1.1 The top five selling drugs by cell line in 2016. Four of the five were recombinant monoclonal antibodies produced in CHO cells. 5

Table 1.2 A table of the capabilities of currently commercially available sequencers by Illumina, Inc. 17

Table 1.3 Number of long non-coding RNA genes by classification from GENCODE. Data retrieved on December 2019 with release of human (V32) and mouse (M23) annotations. The majority of annotated lncRNA genes are found antisense or in intergenic regions. The “To Be Experimentally Confirmed” class of non-coding genes are a set of EST clusters that are polyadenylated but have not yet been verified by qRT-PCR. 22

Table 1.4 A table of lncRNA genes found in the CHO-K1, CHOK1GS, PICR, and *C.griseus* genomes. All genes were classified into intergenic, antisense, and sense loci with respect to nearby protein coding genes using the GENCODE definitions. This classification was performed using the *BEDtools* suite as detailed in Chapter 2. 37

CHAPTER 2

Table 2.1. A table of raw and quality filtered read pairs with their mapping quality to the ENSEMBL criGri1 genome. Approximately 15% of reads were removed due to insufficient quality. Quality filtered reads were mapped to the criGri1 genome and an average mapping rate of 85% was observed across all samples. Libraries sequenced from cells cultured at 31°C were observed to have a slightly higher mapping rate. 71

Table 2.2 A table summarizing the top 5 most enriched GO terms for protein coding genes. Genes involved in regulation of cell cycle, mitosis, and cellular response to stress were significantly enriched. 83

Table 2.3 Log₂-fold-changes and BH adjusted p-values for sense protein coding genes and antisense lncRNAs. There are 6 sense/antisense gene pairs that exhibit correlated expression and 11 sense/antisense gene pairs that exhibit anticorrelated expression. 88

CHAPTER 3

Table 3.1 Amino acid and ammonia composition of fed-batch feed. Glucose was continuously monitored throughout the process and supplemented when concentrations declined below 5.6 g/L 107

Table 3.2 The top ten most differentially expressed genes between exponential and stationary phase. 116

Table 3.3 The top 5 most enriched gene ontology terms as identified by DAVID which include processes involved in cell growth and cell cycle progression. 118

Table 3.4 The top ten most differentially expressed genes between stationary and death phase. 118

Table 3.5 Enriched gene ontology terms as identified by DAVID which include processes involved in cell growth and cell cycle progression.	120
Table 3.6 Top ten most significantly differentially expressed genes over time between exponential, stationary, and death phases.	121
Table 3.7 The top 5 most enriched GO biological process terms identified by DAVID for genes that were differentially expressed over all three culture phases.	122
Table 3.8 A table summarizing-fold-change and BH adjusted P-values between differentially expressed sense and antisense gene pairs. There were 10 genes with correlated expression and 9 genes with anti-correlated expression.	125
Table 3.9 Five sense/antisense gene pairs that are strongly differentially expressed during the between the stationary and death phase.	128
Table 3.10 Three differentially expressed sense/antisense gene pairs that exhibit changes in expression over all phases of the cell culture. The NAV2 gene is the most abundant of this group and has anti-correlated expression during the stationary phase.	128
Table 3.11 Pearson's correlation coefficients of four genes that were significantly correlated with multiple variables. All four genes negatively correlated with % viability and positively correlated with ammonia.	136

CHAPTER 4

Table 4.1 A table of significantly differentially transcribed isoforms that are also differentially expressed at the gene level between stationary and exponential phase. The PIGA and SNX24 genes have isoforms that become dominant during the stationary phase that attribute to the overall change in gene expression.	164
Table 4.2 A table of significantly differentially transcribed isoforms that were not differentially expressed at the gene level between stationary and exponential phase.	166
Table 4.3 Gene ontology enrichment for cellular processes for genes that undergo isoform switching without a change in expression between the exponential and stationary phases. Processes involved in ncRNA post-processing and cellular localization are significantly enriched.	167
Table 4.4 A table of significantly differentially transcribed isoforms that are also differentially expressed at the gene level between stationary and death phase.	171
Table 4.5 A table of significantly differentially transcribed isoforms that were not differentially expressed at the gene level between stationary and exponential phase.	172
Table 4.6 Gene ontology enrichment for cellular processes for genes that undergo isoform switching without a change in expression between the stationary and death phases.	174
Table 4.7 Transcripts that are alternatively spliced between exponential, stationary, and death phase.	178
Table 4.9 lncRNA genes that are differentially expressed that also undergo alternative splicing at the transcript level	180

Table 4.10 lncRNA transcripts that undergo isoform switching without any change at the gene level.	181
Table 4.11 lncRNA genes that are differentially expressed that also undergo alternative splicing at the transcript level between stationary and death phases	185
Table 4.12 lncRNA transcripts that undergo isoform switching without any change at the gene level.	186
Table 4.13 Table of transcripts that switch abundances in all three stages of cell culture. The majority of these transcripts are located within intergenic regions.	189

CHAPTER 5

Table 5.1 A table of raw and quality filtered read pairs with their mapping quality to the CHOK1 genome. An average of 10% of reads were removed after filtering. Filtered reads were aligned to the mitochondrial genome with an average of 65% being identified as mitochondrial in origin and the remaining 35% aligned to the nuclear genome.	232
Table 5.2 A table of the top ten most enriched gene ontology cellular processes. Differentially accessible peaks were annotation to the nearest TSS. Genes that were simultaneous differentially expressed and differentially accessible were selected with only genic sites used for GO enrichment. The number of genes in each process is detailed along the a Benjamini-Hochberg adjusted FDR value.	243
Table 5.3 A table containing motifs that are highly enriched in differentially accessible peaks. This table summarizes the transcription factor motif sequence as well as the percent similarity to the reference motif database and the adjusted P-value.	244
Table 5.4 A table of the top ten most enriched motifs identified by chromVar. This table summarizes the difference in mean deviation between culture conditions as well as the percent of peaks that contain a specific motif. KLF sites were found to have the highest increase in aggregate accessibility while AP-1 sites containing FOS:JUN binding sites were the most enriched.	251

Abbreviations

AS	-	Antisense
AGO	-	Argonaute
ATAC	-	Assay for Transposase-Accessible Chromatin by Sequencing
BP	-	Base Pair
BAK	-	BCL2 Antagonist
BAX	-	BCL2 Associated X
BLAST	-	Basic Local Alignment Search Tool
ceRNA	-	Competing Endogenous RNA
CHO	-	Chinese Hamster Ovary
CDS	-	Coding Sequences
cDNA	-	Complementary DNA
CPM	-	Counts per Million
DNA	-	Deoxyribonucleic Acid
DA	-	Differential Accessibility
DE	-	Differential Expression
DHFR	-	Dihydrofolate Reductatse
GB	-	Gigabase
HOTAIRM1	-	HOX Transcript Antisense Intergenic RNA M1
lincRNA	-	Long Intergenic Non-Coding RNA
lncRNA	-	Long Non-Coding RNA
KB	-	Kilobase
mRNA	-	Messenger RNA
MB	-	Megabase
MALAT	-	Metastasis Associated Lung Adenocarcinoma Transcript 1
miRNA	-	Micro RNA
mAb	-	Monoclonal Antibody
NGS	-	Next Generation Sequencing
NTS	-	Non-Temperature Shift
ncRNA	-	Non-Coding RNA
PTM	-	Post-translational Modifications
qPCR	-	Quantitative Polymerase Chain Reaction
RNA	-	Ribonucleic Acid
rRNA	-	Ribosomal RNA
scRNA-Seq	-	Single-cell RNA-Seq
sgRNA	-	Short-guide RNA
snoRNA	-	Small Nucleolar RNA
snRNA	-	Small Nuclear RNA

TS	-	Temperature Shift
TPM	-	Transcripts Per Million
XIST	-	X-Inactive Specific Transcript
ZFN	-	Zinc Finger Nuclease

Publications

“A Bioinformatics Pipeline for the Identification of CHO Cell Differential Gene Expression from RNA-Seq Data” (2017) Monger, C., Motheramgari, K., McSharry, J., Barron N., Clarke C. *Methods Mol. Biol.* **1603**: pp. 169-186 doi: 10.1007/978-1-4939-6972-2_11

“Expanding the Chinese Hamster Ovary Cell Long Non-Coding RNA Transcriptome Using RNASeq” (2019) Motheramgari K., Valdés-Bango Curell R., Tzani I., Gallagher C., Castro Rivadeneyra M., Zhang L., Barron N., Clarke C., *Biotechnol. Bioeng.* doi: 10.1002/bit.27467

“Sub Physiological Temperature Induces Pervasive Alternative Splicing in Chinese Hamster Ovary Cells” Tzani I., Monger C., Motheramgari K., Gallagher C., Hagan R., Kelly P., Costello A., Meiller J., Floris P., Zhang L., Clynes M., Bones J., Barron N., Clarke C., *Biotechnol. Bioeng.* doi: 10.1002/bit.27365

“The Landscape of Alternative Splicing in CHO Cells Cultivated in Fed-Batch” – *Manuscript in preparation*

“Reduction in Culture Temperature Induces Genome-wide Alterations in Chromatin Accessibility in CHO Cells” – *Manuscript in preparation*

Abstract: Profiling the landscape of long non-coding RNAs in the Chinese hamster ovary cell transcriptome

Chinese hamster ovary (CHO) cells are the most commonly used cell line for the production of recombinant proteins in the pharmaceutical industry. Next generation sequencing (NGS) experiments have produced *Cricetulus griseus* and CHO genomic and transcriptomic references that have expanded the capability to study molecular mechanisms of CHO cells and enabled cell line engineering strategies to enhance industrially relevant phenotypes. The quality of annotation of protein coding genes is comparable to model organisms such as humans and mice. The quality of annotation of long non-coding RNAs (lncRNAs), transcripts longer than 200 nucleotides that do not code for a protein, is not comparable and needs to be expanded in order to fully characterize the *C. griseus* and CHO genomes. lncRNAs function as regulator molecules that govern the expression of diverse pathways and are ideal targets for future cell line engineering efforts.

In this thesis, the genome and transcriptome of CHO cells were profiled in order to discover and characterize lncRNAs. A computational pipeline that can detect and annotate lncRNAs from RNA-Seq data was constructed and used to relate differential expression to a reduction in culture temperature. A multi-omics experiment was performed which included proteomics and transcriptomics to profile expression and alternative splicing of lncRNAs over the duration of an industrial culture in bioreactors. Next, global chromatin accessibility was profiled in cells that were exposed to a reduction in culture temperature using the Assay for Transposase Accessible Chromatin (ATAC-Seq) in order to identify chromatin-lncRNA regulatory networks. The findings of these studies have expanded the understanding of long non-coding RNAs within CHO cells and provide a resource for future cell line development.

Chapter 1

Literature Review

1 Introduction

In 1986, the FDA approved tissue plasminogen activator which was the first recombinant protein to be produced by the Chinese hamster ovary (CHO) cell line (Kim, Kim, and Lee 2012). In the 34 years since the first product was approved, CHO cells have become the most commonly used mammalian expression host for producing complex recombinant therapeutic proteins such as monoclonal antibodies, coagulation factors, and fusion proteins (Walsh 2014). There has been a significant effort to improve the total recombinant yield of CHO cells by industrial and academic researchers over the past three decades. This chapter will provide an introduction to CHO cells (**Section 1.1**), a brief summary of methods to improve product yield and cell density (**Section 1.2**), genetic engineering strategies to maximize phenotypes relevant for bioprocessing (**Section 1.3**), advances in high-throughput transcriptomics (**Section 1.4**), an introduction to long non-coding RNAs (**Section 1.5**) and the challenges and promises of lncRNA-mediated genetic engineering for improving production phenotypes that are advantageous for industrial bioprocessing (**Section 1.6 & 1.7**). This chapter will conclude with the objectives of this Ph.D. project (**Section 1.8**).

1.1 CHO cells and biopharmaceutical production

The first recombinant protein to be approved for therapeutic use was insulin which was produced in bacterial *Escherichia coli* (Walsh 2014). Prior to this, insulin had to be harvested from animal sources, such as horse or pig pancreases, which was problematic as this yielded low quantities and could provoke a severe immune response within the diabetic patient (Baeshen et al. 2014). Bacterial cells, such as *E. coli*, are very useful for recombinant protein production due to desirable properties like rapid doubling time and the capability of achieving high cell densities (Walsh 2014; Walsh 2018). *E. coli* cells account for the production of almost 20% of all recombinant therapeutic proteins within the pharmaceutical industry (Walsh 2014). Bacterial cells lack organelles and the enzymatic machinery that are necessary for correct protein folding and post-translation

modifications (PTMS). Post-translational modifications (PTMs), such as glycosylation, are covalently bound molecules attached to the surface of proteins. Glycosylation is the biological process of attaching sugar residues (glycans) to the surface of proteins and is one of the most important post-translational modifications and can affect patient immune response to an administered therapeutic, such as a monoclonal antibody. Furthermore, an immune response can lead to clearance of the drug thereby reducing therapeutic efficacy. This immune response can be attributed to the lack of *N-glycolyneuroaminic acid (Neu5Gc)* in human proteins which is present in proteins from other eukaryotic cells (Sumit et al. 2019a) Therefore, other eukaryotic expression systems, such as yeast cells or mammalian cells lines, are commonly used in industrial pharmaceutical protein production. Commonly used mammalian cell lines in recombinant protein production include mouse NS0, SP2/0, and Chinese hamster ovary cells (CHO) (Wurm 2004). Mouse-derived cell lines (NS0,SP2/0) can produce proteins with immunogenic isoforms which are hazardous to patient health (F. Li et al. 2010), however. CHO cells, by contrast, can produce proteins with human compatible glycosylation patterns (Sumit et al. 2019a). CHO cells can also rapidly proliferate to high cell densities with a relatively short doubling time compared to other mammalian cell lines (Sitton and Srienc 2008) and can achieve high cell densities. CHO cells can also proliferate in chemically defined media which lack serum derived from animal tissue which reduces the likelihood of transmission of adventitious agents (Pan, Streefland, and Dalm 2017). Chinese hamster ovary cells are the most commonly used mammalian cells for recombinant therapeutic production due to these qualities and CHO cell lines have a proven history of producing therapeutic proteins safely and meeting strict regulatory requirements set by agencies such as the Food and Drug Administration and the European Medicines Agency (Walsh 2018). Four of the five top selling drugs of 2016 were recombinant proteins produced in CHO cell lines (**Table 1.1**).

Table 1.1 The top five selling drugs by cell line in 2016. Four of the five were recombinant monoclonal antibodies produced in CHO cells.

Drug	Molecule	Sales (billions USD)	Production cell line	Treatment
Adalimumab	Recombinant mAb	16.08	CHO	Autoimmune disorders
Etanercept	Chimeric fusion protein	8.87	CHO	Autoimmune disorders
Infliximab	Recombinant mAb	8.23	Sp2/0	Autoimmune disorders
Rituximab	Chimeric fusion protein	7.22	CHO	Autoimmune disorders
Bevacizumab	Recombinant mAb	6.71	CHO	Cancer

Treatment for diseases such as autoimmune disorders (rheumatoid arthritis, Crohn's disease), hemophilia, and diverse types of cancer are available for patients because of the invention of recombinant DNA technology - cloning the corresponding DNA sequence for a protein into an expression vector and transfecting a cell for production (Wurm 2004; Sanchez-Garcia et al. 2016). As of 2018, a total of 353 unique recombinant therapeutic proteins have been approved by the FDA and EMA for patient use (**Figure 1.1**) with monoclonal antibodies composing the largest group of approved drugs (Walsh 2018), Global pharmaceutical sales are a multi-billion dollar industry and are projected to grow significantly due to the increased *per capita* expenditure on patients by health care systems (Walsh 2014).

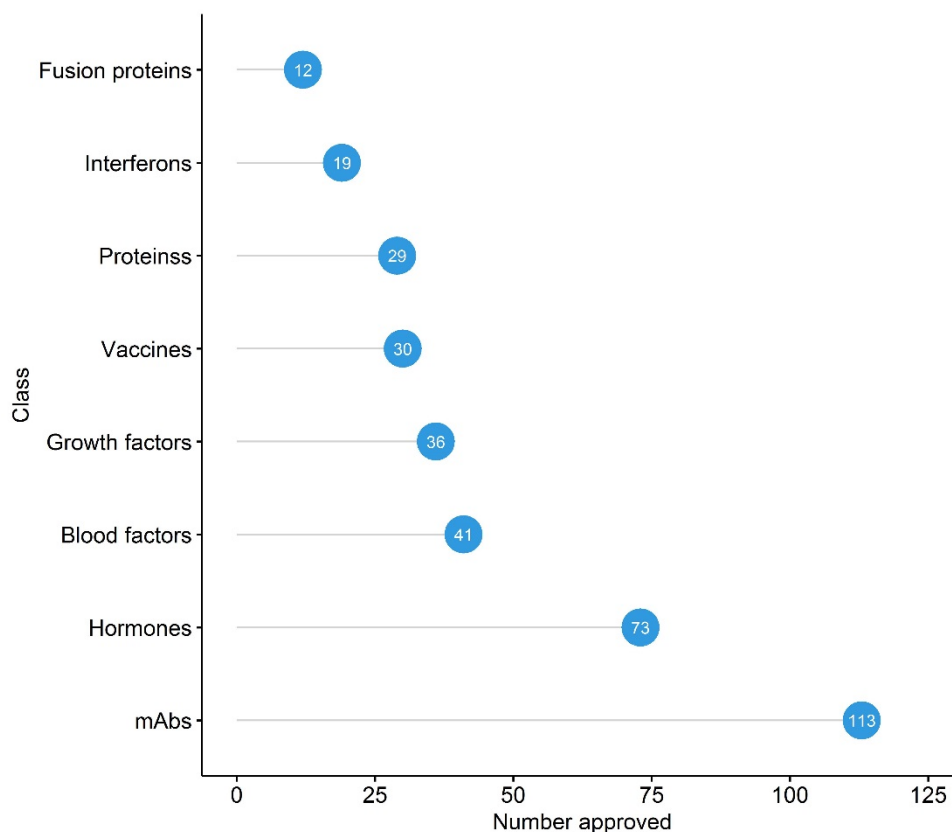


Figure 1.1 Plot illustrating the total number of drugs approved. A total of 353 unique drugs have been approved to date (Walsh, 2018). The largest class of approved recombinant therapeutic proteins is monoclonal antibodies which are commonly produced in CHO cells.

Significant resources have been dedicated by academic and industrial researchers to improve CHO cell productivity and product yields. Advances in cultivation methodology, media formulation, and vector design have all increased CHO cell density and recombinant protein yield. Optimization of phenotypes associated with increased growth, reduced apoptosis, improved protein secretion, and proper glycosylation using genetic engineering strategies is highly desirable as it would increase production capacity (Fischer, Handrick, and Otte 2015). This chapter will summarize the successes of using rational genetic engineering to improve cell line characteristics as well as present the opportunities and challenges of utilizing non-coding RNAs for designing more industrially efficient CHO cells.

1.2 Improving cultivation strategies in order to increase cell growth, viability, and productivity

Total volumetric yield of recombinant proteins has steadily increased over the years due to advances in CHO cell cultivation methods. Initially cells were cultured in “batch” mode where cells are seeded and cultivated in a closed system containing media over the duration of the culture where nutrients such as glucose are slowly depleted and toxic metabolites like ammonia and lactate steadily accumulate which would impede cell growth and viability (Y. Huang et al. 2010). “Fed-batch” cultivation methods were developed to improve cell viability by administering an infusion of nutrients (primarily glucose and amino acids) to the culture which expanded the cultivation period to up to ~21 days (Y. Huang et al. 2010). Recently, perfusion cultures have been implemented in industry to greatly extend cell viability (Hacker, Jesus, and Wurm 2009). Perfusion cultivation has significantly increased cell viability by months by continuously adding a stream of fresh media to the bioreactor while simultaneously removing spent media.

Altering media composition, in conjunction with improved cultivation methods, has also significantly increased titers. Media was originally derived from animal components, such as bovine serum but, where possible, the industry is now utilizing chemically defined media due to the risk of contamination from adventitious agents such as prions, viruses, and bacteria (F. Li et al. 2010). Media formulation can have profound effects on cell growth, productivity, and quality of recombinant proteins, however (Kim, Kim, and Lee 2012). Growth rates in chemically defined media can be cell line specific so media optimization protocols are often performed prior to the production (F. Li et al. 2010)

Titers have improved dramatically over the intervening 30 years. Industrial cell culture facilities have regularly reported average titers in the grams per liter range (**Figure 1.2**). The dramatic improvement can be attributed to improvements in CHO cell cultivation strategies. This was only possible through innovations in upstream process developments which include selecting stable cell lines, vector optimization for transfection, media optimization, and innovations in high-

throughput screening. There is still a significant need to improve CHO cell productivity which can be achieved by rational cell engineering (**Section 1.3**).

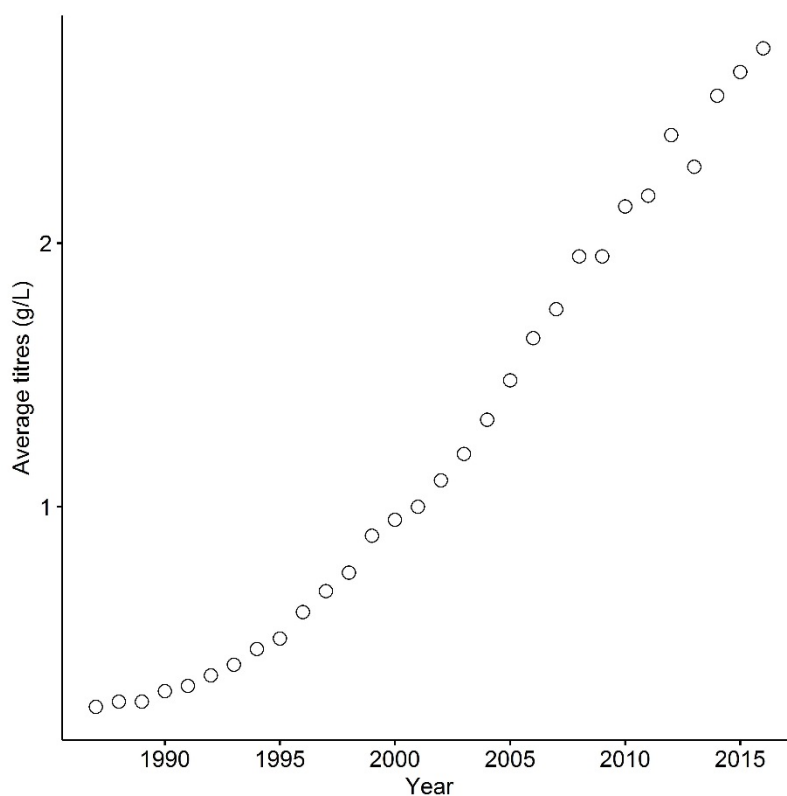


Figure 1.2 Plot of average titres (g/L) over time as reported by Rader and Langer (Rader & Langer, 2014). These data were obtained by questionnaires and a survey of the literature by the authors and tabulated from 2006-2014. Rader and Langer extrapolated titres prior to 2006 from published data and after 2014 by predictions from trends.

1.3 Understanding CHO cell biology and the promise of rational cell engineering

In addition to process development, there has been increasing research studying CHO cell biology in order to simplify CHO cell process development. Cell line development is necessary for selecting clones with optimal phenotypes (rapid growth, reduced apoptosis, high productivity) but is a laborious process (H. Zhang et al. 2010). Traditional cell line development involves random mutagenesis followed by high-throughput screening to select clones that can readily proliferate in defined media and yield a high concentration of recombinant proteins with human compatible glycosylation (H. Zhang et al. 2010). The invention of liquid handling robots, such as the Ambr 15 system (TAP Biosystems), capable of automated culture of small volume bioreactors (10-15 mL) has accelerated this stage of screening. CHO cell cultivation within the small volume bioreactors closely mimics their growth characteristics in larger industrial bioreactors and cells can be screened rapidly in parallel to identify clones for production (Delouvroy et al. 2015). Screening is still a random process, however, which can lead to highly variable timelines for selecting a stable, high producing clone that is suitable for production (Fischer, Handrick, and Otte 2015). Studying CHO cell biology, gene expression, and protein translation using high-throughput methods will enable prediction of growth rate and potential specific productivity of a CHO cell clone and would accelerate cell line selection (Farrell et al. 2014). Furthermore, increasing understanding of CHO cell biology would enable genetic modification to engineer a cell line with consistent, high productivity.

Studies of CHO cell biology and cell line engineering efforts have considerably accelerated with the publication of the CHO (X. Xu et al. 2011; Hammond et al. 2011) and Chinese hamster genomes (Brinkrolf et al. 2013; N. E. Lewis et al. 2013). The genomes were accompanied by the publication of the transcriptome (Birzele et al. 2010; Becker et al. 2011) and the first profiling of the small non-coding transcriptome (Hackl et al. 2011). These advances were fueled by the

maturation of next generation sequencing (NGS) technology which enabled high-throughput analysis of the genome and transcriptome.

The objective of rational cell engineering is a CHO cell that can attain high cell densities rapidly and remain viable for long periods without any adverse effect on recombinant protein quality. Genetic engineering strategies have focused on inhibiting regulators of apoptosis, upregulating cell proliferation genes, improving production capacity, and improving post-translational modification profiles of recombinant proteins using gene knockouts (**Section 1.3.1**), overexpression (**Section 1.3.2**), and modulating small non-coding RNA expression in order to regulate cellular phenotypes (**Section 1.3.3**).

1.3.1 Gene knockouts

Deleting genes that are deleterious for industrial production, such as those that regulate apoptosis, has shown some promise in improving recombinant protein production. Stable whole-gene deletions were traditionally performed via random mutagenesis using ultraviolet radiation or chemical treatment (Fischer, Handrick, and Otte 2015). One of the earliest successes in CHO gene engineering was the deletion of the dihydrofolate reductase (*DHFR*) gene which was accomplished by total gene knockout using ionizing radiation (Urlaub et al. 1983). Cells could then be transfected with vectors containing a functional *DHFR* copy and cultured in media lacking hypoxanthine, thymidine, and glycine and easily screened for successful transfection. The DG44 and DX11 lineages, which are *DHFR*-deficient, have been popular production cell lines for this reason (Fischer, Handrick, and Otte 2015)

Traditional methods for gene deletion relied on chemical or radiation treatments for random mutagenesis which required laborious follow-up screening to select desired mutants. Gene editing technologies, such as zinc finger nucleases (ZFN), transcription activator-like effector nucleases (TALENs), or the recently developed CRISPR/Cas9 system have the advantage of being more precise with the ability to target specific loci for deletion (Fischer, Handrick, and Otte 2015). Gene

editing has been successfully utilized to engineer apoptosis resistance by deleting the pro-apoptotic genes *BAX* and *BAK* which significantly increased mAb yield 2-5 fold as cells were viable for longer (Cost et al. 2010). Deleting genes responsible for immunogenic post-translational modifications is also an attractive option. Fucosylation, for example, of mAbs has been linked to severe cellular cytotoxicity. Zinc finger nucleases have been utilized to delete the fucosyltransferase gene, *FUT8*, to prevent the addition of fucose residues on monoclonal antibodies (Malphettes et al. 2010). Recently, Grav *et al.* used CRISPR/Cas9 to generate multi-knockout *BAX*, *BAK*, and *FUT8* cell lines which exhibited prolonged viability and double the specific productivity and produced mAbs without fucosyl residues (Grav et al. 2015). These studies demonstrate the viability of gene knockouts to improve the CHO chassis to produce cell lines with industrially beneficial phenotypes.

1.3.2 Gene over-expression

Transfection and overexpression of a gene is also a viable way to improve production capacity. It is accomplished by cloning the coding sequence (CDS) of a gene of interest into a vector containing a strong viral promoter for transient transfection. This technique can be used to extend culture viability by overexpressing anti-apoptotic or proliferative genes. Overexpression of the kinase *AKT* delayed autophagy and extended culture life span of cells producing a recombinant mAb, which increased total recoverable yield in batch culture (S. O. Hwang and Lee 2009). Cells overexpressing *AKT* were 20% more viable than wild type cells at day 12 of culture (S. O. Hwang and Lee 2009). Moreover, gene overexpression can be applied to improve production of low yield, difficult to express proteins. Pybus *et. al.* successfully increased volumetric yield of several difficult to express mAbs by co-expressing *ATF6c*, *BIP*, and *XBP1* (Pybus et al. 2014).

1.3.3 Small non-coding RNA engineering

A specific phenotype can often not be attributed solely to a singular gene but, rather, to a complex co-regulatory network of genes (Müller, Katinger, and Grillari 2008). microRNAs (miRNAs), small non-coding RNAs ~22 nt in length, have shown great promise in regulating these complex

pathways and are advantageous targets for genetic engineering as they do not add to the translational burden of a producing cell (Jadhav et al. 2013). miRNAs are the most studied class of small non-coding RNAs which include piwi-intersecting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), and small nuclear RNAs (snRNAs). They have been implicated as regulators in many diverse pathways such as cell proliferation, apoptosis, and cellular metabolism (Bartel 2004).

miRNAs function as post-transcriptional regulators of mRNA translation (Su et al. 2016). These short molecules bind to the 3'UTR of its target messenger RNA to inhibit translation. The miRNA transcript is first transcribed by RNA polymerase II and processed into ~70nt hairpin structures by *DROSHA* and *DGCR8* (Bartel 2004). The processed hairpin is then exported into the nucleus where it is cleaved by *DICER* to form miRNA duplex structures. The mature strand of the processed miRNA associates with the argonaute protein *AgoII* to form the RNA-induced silencing complex (RISC) (Bartel 2004). The miRNA-RISC complex binds antisense to the 3' untranslated region of an mRNA molecule to induce translational inhibition or mRNA degradation to control protein expression.

The publication of the Chinese hamster genome and transcriptome has improved the ease of discovering and functionally annotating miRNAs. These resources have greatly enabled miRNA based genetic engineering strategies (Jadhav et al. 2013). The first study to characterize miRNA expression used a mouse-specific microarray for cross-species comparison (Gammell et al. 2007) which identified evolutionarily conserved miRNAs but was limited in finding novel, *C. griseus* specific ones. Following studies utilized next generation small RNA sequencing to profile and annotate conserved and novel miRNAs in CHO cells (Hackl et al. 2011) which identified 235 evolutionarily conserved miRNAs as well as 11 *C. griseus* specific miRNAs. These early studies underscored the utility of non-coding RNAs, such as miRNAs, as targets to modulate CHO cell phenotypes.

Overexpression and RNA interference studies demonstrated that miRNAs play an important role in regulating bioprocessing relevant phenotypes. Sequestration of miR-7a using an engineered decoy transcript, or miRNA sponge, increased maximum cell density by 40% and prolonged cell culture viability during fed-batch cultivation thereby resulting in improved recombinant protein yield (Sanchez et al. 2014). Jadhav *et. al.* demonstrated that overexpression of miR-17 improved the average growth rate by 23% with a corresponding 3-fold increase in product titer (Jadhav et al. 2014). High-throughput sequencing has also profiled differential miRNA expression in relation to specific productivity in cell lines after hypothermic shock and identified a cohort of 89 miRNAs (among them *cgr-miR-30a-5p* and *cgr-miR-146a*) that were differentially expressed in high, low, and non-producing cell lines (Stiefel et al. 2016). These global profiling studies have demonstrated the promise miRNAs have in cell line engineering. Unfortunately, many miRNAs discovered in CHO cells have not been well characterized and there is a great need for more high-throughput screening in order to associate their differential expression with distinct, industrially relevant phenotypes.

1.4 High-throughput transcriptomics analysis

Historically, discovering and quantifying expressed RNA molecules was laborious as it relied on the conjunction of Sanger sequencing and qPCR (Dijk, Jaszczyzyn, and Thermes 2014). In recent years, our ability to profile the transcriptome has evolved considerably after the invention of microarrays and next generation sequencing as these technologies can profile thousands of transcripts rapidly. This section presents a brief introduction to cDNA microarrays (**Section 1.4.1**), the improved performance and use of Next Generation Sequencing (NGS) for transcriptomics (**Section 1.4.2**), the methodology of preparing libraries compatible for Illumina sequencing (**Section 1.4.3**), and the potential for single cell transcriptomics (**Section 1.4.4**).

1.4.1 Microarrays

Microarrays were one of the first technologies with the capacity for bulk transcriptome profiling and relied on hybridization of fluorescently-labelled, short, antisense oligonucleotides (Colombo et al. 2011a). The oligonucleotide probes are often less than 100 base pairs long with one end affixed to the glass of the array (Colombo et al. 2011b). As cDNA hybridizes to its complementary oligonucleotide, the bound probe fluoresces after excitation with a laser which is then imaged. A probe that hybridizes with its complementary cDNA will emit at a higher intensity than an unbound probe which indicates the presence of a particular transcript (C. Lee and Kikyo 2012). Increased signal intensity correlates to higher RNA expression. Microarrays have been successfully utilized in previous high-throughput studies to correlate mRNA expression to important bioprocessing relevant conditions. Vito and Smiles used a mouse microarray containing probes for annotated mRNAs and lncRNAs to identify differential expression of non-coding transcripts between batch and fed-batch culture systems over time (Vito and Smales 2018). This experiment was one of the first to specifically analyze the behavior of lncRNAs in regards to bioprocessing relevant conditions but was limited in identifying evolutionarily conserved loci (Vito and Smales 2018).

A microarray can be a cost-effective method to screen expression of thousands of target transcripts but it comes with many challenges due to their reliance on hybridization to known

probes (Bumgarner 2013). Microarrays can only profile transcripts that have sufficient complementarity to affixed oligonucleotide probes which is a limitation when attempting to analyze poorly conserved transcripts such as non-coding RNAs (Bumgarner 2013; C. Lee and Kikyo 2012). A microarray study requires *a priori* knowledge of transcript sequence composition as well as a well-annotated reference genome for probe design (C. Lee and Kikyo 2012). The resolution of a microarray is limited by its probe design; unique isoforms can only be profiled if a sequence-specific probe is include on the array (C. Lee and Kikyo 2012). This is a greater limitation in relatively poorly annotated organisms like *C. griseus* where loci might not be accurately represented in the draft assembly. Furthermore, probe design can be prohibitively expensive and labor intensive for custom arrays. While microarrays still provide a method for cost-effective, high-throughput transcriptome profiling, they have largely been superseded by next generation sequencing technologies.

1.4.2 Next generation sequencing

The invention of high-throughput, next generation sequencing has greatly accelerated the capacity to analyze the genome and transcriptome. NGS is distinguished from traditional Sanger sequencing by its ability to sequence and quantify millions of short DNA and cDNA fragments (35-1000bp) in parallel (Dijk, Jaszczyzyn, and Thermes 2014). For example, it is now possible to sequence an entire human genome in a matter of hours whereas it took an international consortium 10 years and 3 billion USD to derive the sequence of the human genome (Dijk, Jaszczyzyn, and Thermes 2014). Next generation RNA-Seq has significant advantages over microarrays in that it enables *de novo* identification and characterization of transcripts without relying on hybridization to a probe. NGS can also provide absolute quantification of transcripts instead of relying on housekeeping genes or relative fluorescence (Birzele et al. 2010)

Currently, the Illumina sequencing system is the most widely used platform which utilizes the “sequencing by synthesis” (SBS) method. SBS relies on synthetic oligonucleotide probes that

have one end affixed to the flow cell and the other free-floating. In the library preparation process, reverse transcribed RNA molecules are fragmented and synthetic sequences (referred to as adaptors) are ligated to the ends of the fragments (Bentley et al. 2009). The termini of the adapter ligated fragments contain sequences which are complementary to the oligos affixed to the flow cell (Bentley et al. 2009). The prepared library solution is injected into the flow cell and the library fragments hybridize to the flow cell affixed oligos in a process referred to as “bridge amplification”. A solution of fluorescently labelled dideoxy nucleotides (ddNTPs) is added to the flow cell at regular intervals called a cycle where the number of cycles equals the length of read length (Bentley et al. 2009; Dijk, Jaszczyszyn, and Thermes 2014). DNA polymerases incorporate ddNTPs that are antisense to the sequence fragments for each cycle and the incorporated polymers are excited with a laser and imaged. Each ddNTP (A,C,G,T) fluoresces at a different wavelength after excitation (**Figure 1.3**) so image processing software can be used to identify which base has been incorporated at each cycle (Dijk, Jaszczyszyn, and Thermes 2014).

The Ion semiconductor semi-conductor (IonTorrent) sequencing method developed by Life Technologies uses the release of hydrogen ions (H^+) during each cycle to sequence base pairs of fragmented libraries. In the library preparation process, fragmented DNA or cDNA is affixed to glass bead and a polymerase chain reaction is performed to generate multiple copies of library fragments until the surface area of the bead is covered. All the glass beads containing affixed libraries are then suspended within microscopic wells and a solution of dNTPs floods the flow cell and a new base is added to the antisense strand by DNA polymerase. Covalent bonding of the sugar-phosphate backbone of DNA releases a H^+ ion that can be detected due to the change in voltage due to the drop in pH. IonTorrent sequencing is less prevalent despite achieving slightly higher read lengths due to the higher error rates reported with the technology.

Sequencing technology has steadily improved over time. At its inception, Illumina sequencing read lengths were quite short (25 bp) but advances in microfluidics technology and kit chemistry

has increased it significantly. Currently, the longest read length available for an Illumina instrument is 2 x 300 bp whereas the IonTorrent instrument can achieve 2 x 400 bp resolution (**Table 1.2**). Advances in kit chemistry have dramatically increased the total throughput possible per sequencing run, as well. Technological advances and economies of scale have significantly dropped the cost of sequencing (**Figure 1.4**).

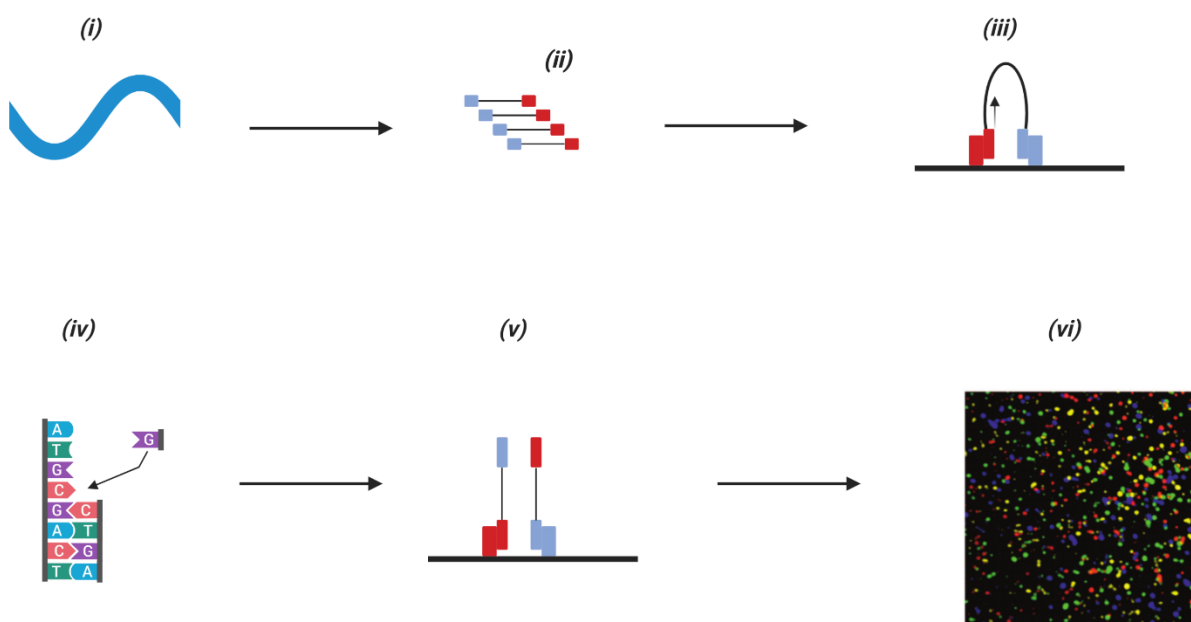


Figure 1.3 A diagram of paired-end Illumina Sequencing by Synthesis protocol. (i) RNA molecules are extracted from cell lysate and purified and (ii) forward and reverse adaptor sequences are added (iii) Loaded libraries bind to complementary probes fixed to the flow cell (v) SBS is performed with the incorporation of fluorescently labelled ddNTPs (vi) Fluorescently labelled ddNTPs are excited with a laser after each cycle and imaged. Each incorporated ddNTP excites at a different wavelength. Image processing software then converts snapshots into ACGT nucleotide characters and reads.

Table 1.2 A table of the capabilities of currently commercially available sequencers by Illumina, Inc.

Instrument	Longest Read Length (bp)	Maximum Output (Gb)	Maximum Reads per Run (Million)

MiniSeq	2 x 150	7.5	25
MiSeq	2 x 300	15	25
NextSeq	2 x 150	120	400
HiSeq 2500	2 x 125	1000	4000
HiSeq 2500-Rapid Run	2 x 250	300	600
HiSeq 3000	2 x 150	750	2500
HiSeq 4000	2 x 150	1500	5000

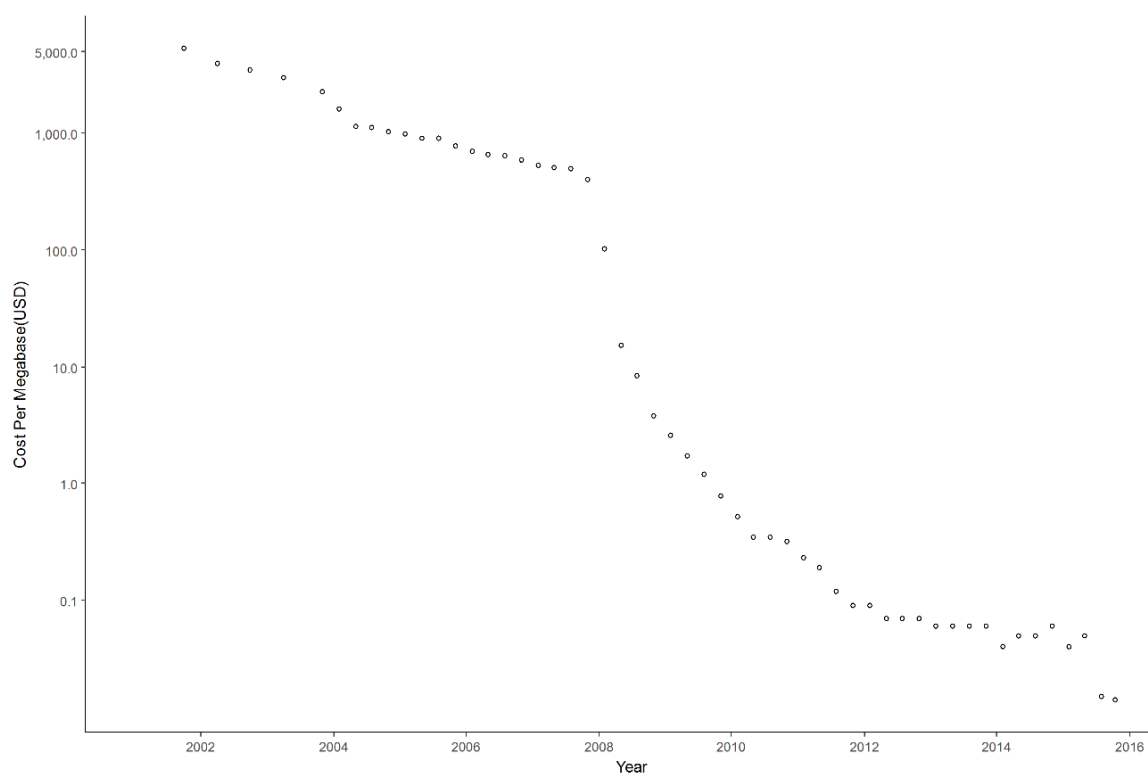


Figure 1.4 A plot of sequencing cost per megabase from 2001 to 2015. Sequencing costs were tracked over time by NIH for its subsidiary sequencing centres.

1.4.3 Illumina RNA-Seq Library Preparation

Library preparation requires high quality RNA extracted at sufficient yield (0.1-1 µg) to ensure efficient cDNA synthesis and library quality (Borodina, Adjaye, and Sultan 2011). Low starting concentrations of RNA could result in inefficient ligation and low-quality libraries, however. The quality of sequencing can also be affected by degradation of RNA samples after extraction so great care must be taken during handling and processing (Borodina, Adjaye, and Sultan 2011) . Prior to library preparation, it is vital to reduce ribosomal RNA representation in extracted samples. Ribosomal RNA accounts for 90% of all transcribed RNA in the eukaryotic transcriptome and libraries must be enriched to enable efficient sequencing (Neil, Glowatz, and Schlumpberger 2013). One method is to enrich for poly-adenylated RNA species using poly-T probes (Sultan et al. 2014) which would select for RNA molecules that are localized to the cytoplasm. These molecules would be predominately composed of mRNAs but approximately 40% of lncRNAs can also be localized to the cytoplasm (Quinn and Chang 2015; Derrien et al. 2012). An alternative method is to deplete ribosomal RNA (rRNA) species from total RNA extracts which can be accomplished with biotinylated rRNA specific probes that can hybridize and be removed from solution, thereby enriching other RNA species (Neil, Glowatz, and Schlumpberger 2013). rRNA depletion is preferred for lncRNA transcriptomics studies as it enriches for the global population of RNA molecules instead of the small subset that is poly-adenylated (Y. Guo et al. 2015).

Following rRNA depletion/poly-A enrichment, the RNA is fragmented and reverse transcribed into cDNA followed by second strand synthesis and adapter ligation (TruSeq Stranded Total RNA Sample Prep Guide, Illumina, Inc). Strand-specific library protocols were a major innovation in RNA sequencing analysis as these protocols are especially important for long non-coding RNA analysis as it allows accurate resolution of sense/antisense and bi-directional promoter transcription (Z. Wang, Gerstein, and Snyder 2009). The dUTP second strand marking protocol is an efficient method for preparing directional libraries. During second strand synthesis, dUTPs

are added in lieu of dTTPs and the second strand is degraded, leaving only the first strand for PCR amplification (Zhao et al. 2015).

Stranded library preparation protocols are important for studying lncRNA transcription by next generation sequencing. Antisense lncRNAs are one of the largest categories of lncRNAs (Harrow et al. 2012; Derrien et al. 2012) and their differential expression can only be identified using methods that preserve strand orientation. Otherwise, expression of the sense protein coding locus would mask signal from the antisense strand.

1.4.4 Single-cell transcriptomics

Traditional total RNA-Seq averages sequencing signal across the total population of cells extracted in each sample (B. Hwang, Lee, and Bang 2018). The averaged signal would mask sub-populations that are expressing specific transcripts at very high levels. Cells within the same tissue often express transcripts heterogeneously despite possessing identical genomes. Single cell RNA-Seq (scRNA-Seq) is a sequencing method that can detect RNA-transcripts in solitary cells isolated from a population (B. Hwang, Lee, and Bang 2018).

The first stage of scRNA library preparation is isolating viable cells which can use microfluidic assisted cell sorting, fluorescently activated cell sorting (FACS), or even simple serial dilutions. The serial dilution method is not favored due to the lack of sensitivity in achieving single cell aliquots (B. Hwang, Lee, and Bang 2018). The FACS method relies on fluorescently activated monoclonal antibodies as a selection marker to identify surface markers to isolate distinct cellular sub-populations. Microfluidics assisted cell sorting is a novel method that is highly precise and can retain viable cells from a total population using the flow of nanoliters of liquid through channels on the fluidics device (J. Buenrostro et al. 2015b). Following cell isolation, RNA is extracted from each aliquot individually and strand-specific library preparation. Each aliquot of RNA will possess its own unique barcode that can be used to associate RNA expression to a specific cell type.

1.5 Long non-coding RNAs

Long non-coding RNAs, a relatively novel class of non-coding molecules, (lncRNAs) are RNAs longer than 200 nucleotides and lack the ability to code for a protein (Kung, Colognori, and Lee 2013). The Human Genome Project discovered that only 2% of the human genome encoded protein coding genes (Kung, Colognori, and Lee 2013) whereas the ENCODE consortium, a large collaboration that profiled the human transcriptome and proteome, later discovered that the human genome was pervasively transcribed (Encode and Consortium 2007). These findings were initially met with skepticism and dismissed as transcriptional noise from surrounding protein-coding genes (Struhl 2007). Genome-wide profiling of eukaryotic transcriptomes using microarrays and next generation sequencing overturned this assumption and demonstrated that these RNAs were transcribed and had regulatory functions within the cell (Iyer et al. 2015).

Xist was one of the first long non-coding RNAs to be functionally annotated (Penny et al. 1996). Following this initial finding, thousands of lncRNAs have been profiled using RNA-Seq in many different organisms such as humans, mice, zebrafish, and maize. Functional studies revealed that these transcripts are a very diverse group of molecules and exhibit similar post-transcriptional modifications to messenger RNAs such as alternative splicing and poly-adenylation (Necsulea et al. 2014). lncRNAs have been localized to both the nucleus and cytoplasm and regulate a diverse number of cellular pathways. Long non-coding RNAs have been implicated as key regulatory molecules in many diverse pathways such as X chromosome inactivation (Leung and Panning 2014), apoptosis (Su et al. 2016), cell senescence (Puvvula et al. 2014), and cell proliferation (Jiang et al. 2015). Their involvement as regulators in many key pathways make lncRNAs ideal targets for genetic engineering.

The regulatory roles of miRNAs are better understood when compared to lncRNAs as their mechanism of action is better characterized; miRNAs are complementary to the UTR region of the transcript that they target and regulate. Long non-coding RNAs are not as well understood, in

comparison, and often have multiple functions. These molecules can modulate transcription by recruiting transcription factors to promoters (Ferrè, Colantoni, and Helmer-Citterich 2015), bind and localize chromatin remodeling complexes (Mercer and Mattick 2013), titrate away miRNAs and transcription factors, and stabilize mRNAs during translation (Rashid, Shah, and Shan 2016). These diverse mechanisms have the potential for fine-grain modulation of cell phenotypes.

1.5.1 Long non-coding RNA classification

Long non-coding RNAs are a highly heterogeneous group of molecules and are classified into two groups: genic (within protein coding loci) or intergenic (outside protein coding loci), by their localization on the genome with respect to protein coding genes (Harrow et al. 2012; Derrien et al. 2012). Intergenic lncRNAs are transcribed in intervening regions (the regions between two protein coding genes) of the genomes, flanking protein coding genes, or in gene deserts (**Table 1.3**) and compose the largest proportion in mammalian genomes. Genic lncRNAs are more diverse and can be found overlapping protein coding genes in either strand or orientation (**Figure 1.5**). Intragenic lncRNAs can be classified as antisense, sense exon overlapping, sense intron overlapping, 3' overlapping, or bi-directional with respect to a protein coding gene. Antisense lncRNAs are transcribed antisense to a protein coding gene and can be found flanking the 5' or 3' end of the sense protein coding transcript or be located entirely antisense to the sense. Sense long non-coding RNA overlap protein coding genes and can be processed from introns and exons. A third, rarer group of sense non-coding RNAs are found transcribed from the 3' UTR.

Table 1.3 Number of long non-coding RNA genes by classification from GENCODE. Data retrieved on December 2019 with release of human (V32) and mouse (M23) annotations. The majority of annotated lncRNA genes are found

antisense or in intergenic regions. The “To Be Experimentally Confirmed” class of non-coding genes are a set of EST clusters that are poly-adenylated but have not yet been verified by qRT-PCR.

Figure Id	Class	GENCODE Definition
i	Antisense	Sense Overlapping: Transcribed and processed from the introns and exons on the same strand as the protein coding gene. Transcription of the exon can be complete or partial.
ii	Bi-Directional Promoter	Transcribed in the antisense strand but shares a promoter with protein coding gene
iii	Sense Intronic	Transcribed and processed from the introns on the same strand as the protein coding gene
iv	Sense Overlapping	A non-coding transcript that contains a protein coding gene within its intronic sequence on the same strand, with no overlap of exonic sequence.
iv	Intergenic	Loci transcribed from regions lacking protein coding genes

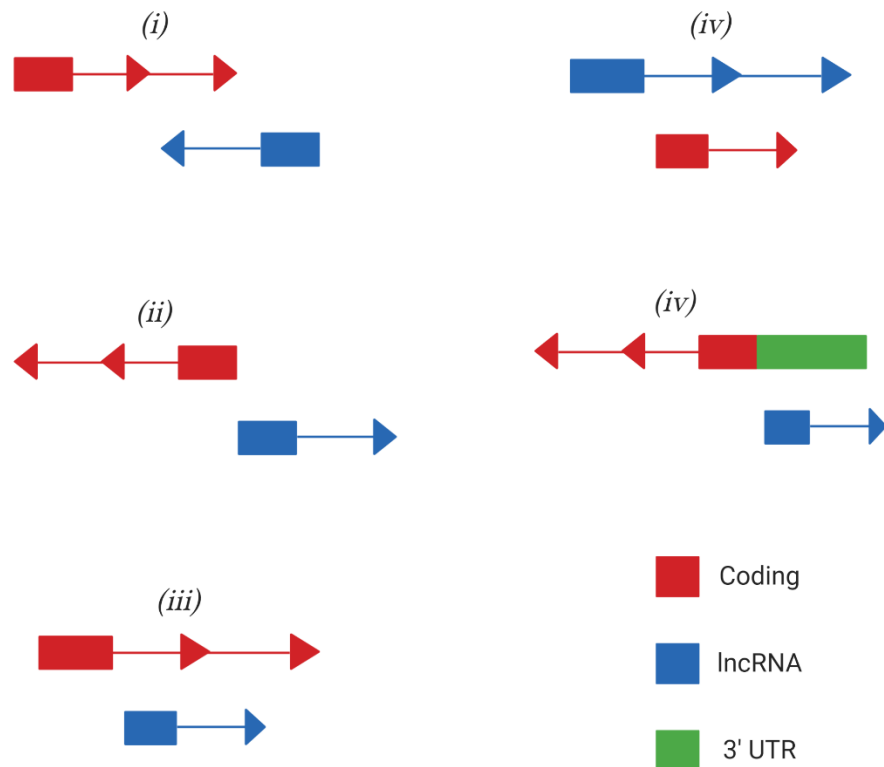


Figure 1.5 Classification of lncRNAs by genomic localization. The classifications are detailed in **Table 1.3**

1.5.1.1 GENCODE REFERENCE ANNOTATION OF HUMAN AND MOUSE GENOMES

In 2003, the ENCyclopedia Of DNA Elements (ENCODE) consortium was founded to identify all “functional elements” such as protein coding transcripts, transcription factor binding sites, CpG Islands, and histone modifications within the human genome (ENCODE Project Consortium 2012). The consortium utilized high-throughput sequencing techniques such as RNA-Seq and Cap Analysis Gene Expression (CAGE) sequencing to profile the transcriptome, of human cell lines (HeLa, H1-hESC, and GM12878) and tissue types (ENCODE Project Consortium 2012). CAGE-Seq enriches for the 5’ methyl G-capped RNAs in order to sequence the 5’ end of the molecule and can be used to identify the canonical transcriptional start site (TSS) and promoter regions of a gene (Flicek et al. 2012; Harrow et al. 2012). ENCODE also utilized other high-throughput sequencing techniques such as ChIP-Seq and DNase-Seq to profile the transcription factor binding sites, Bisulfite-Seq to profile DNA methylation sites, FAIRE-Seq, MNase-Seq, and histone ChIP-Seq to profile chromatin structure (ENCODE Project Consortium 2012).

The GENCODE consortium was co-founded to annotate the results from ENCODE and create a comprehensive human genome annotation (Harrow et al. 2012). Initially, GENCODE focused on annotating the approximately 1% of the human genome that is thought to encode a protein but was later expanded to include the full complement of genes (Harrow et al. 2012). GENCODE constructed a process pipeline that utilized computational prediction and manual curation to construct a reference set. Manual annotation was performed using analysis from the Human and Vertebrate Analysis (HAVANA) group which constructs high-confidence transcript models by aligning cDNA data and proteomic data to a genome in order to identify canonical splice sites.

A surprising discovery of the ENCODE project was that transcription occurred outside protein coding genes (ENCODE Project Consortium 2012). Some of these transcript loci were found to be non-coding RNAs and became a focus of the GENCODE annotation efforts (Derrien et al.

2012). Derrien *et al.* utilized microarrays and next generation sequencing to construct a high confidence reference annotation of the long non-coding transcriptome (Derrien et al. 2012). The GENCODE annotation project also formalized the different classification of long non-coding RNAs by their location on the human genome (Harrow et al. 2012). Gene and transcript models were first constructed from read alignments to the human genome. Transcripts longer than 200 nucleotides that overlapped known protein coding loci were removed from the dataset. The filtered transcripts were evaluated for coding potential using the PhyloCSF software package and those transcripts that possessed a low coding potential were evaluated for the presence of a protein coding domain using the hmmer software suite (Harrow et al. 2012; Derrien et al. 2012). Strand-specific RNA-Seq performed by GENCODE members and ENCODE data from different tissue types was utilized to identify an expanded cohort of antisense long non-coding RNAs which accounted for 34% of the lncRNA annotation (Derrien et al. 2012). Combined CAGE-Seq and RNA-Seq was utilized to identify canonical sense non-coding RNAs with a TSS separate from their sense protein coding gene. The authors further refined their lncRNA gene predictions using HAVANA manual annotation in order to identify canonical transcript models for each lncRNA loci. The GENCODE annotation project was designed to be a database that would be constantly revised and updated to include a larger annotation set as more comprehensive transcriptomics datasets were generated by the scientific community (Harrow et al. 2012). The GENCODE database later expanded to include similarly comprehensive annotations for the mouse database (Mudge and Harrow 2015). As of November 2019, version 32 of the human annotation and version MM 23 of the mouse reference annotation are published. The GENCODE reference annotation is a high-confidence gold standard that contains one of the most comprehensive annotations of protein coding and non-coding loci of the human and mouse genomes (**Figure 1.6**). The systematic annotation methodology can be adapted to build a similarly comprehensive database of coding and non-coding loci for non-model organisms such as *C. griseus*.

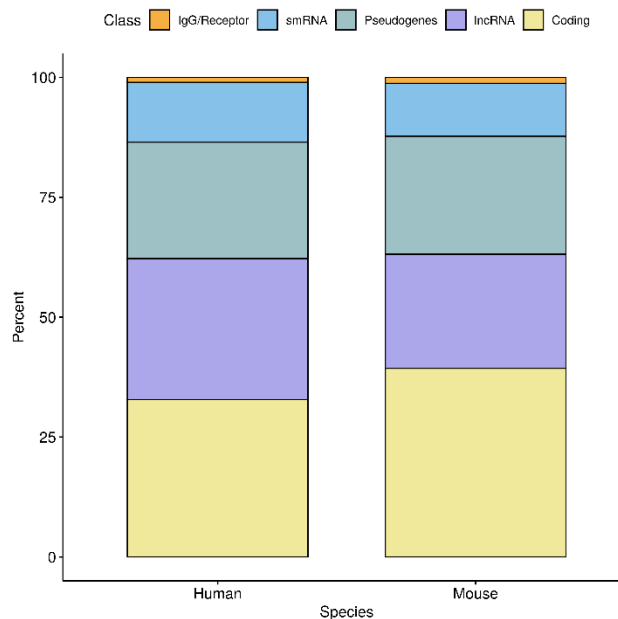


Figure 1.6 GENCODE classification of genes into respective categories. These data were retrieved from release 32 of the human and release 32 of the mouse annotation.

1.5.1.2 THE ENSEMBL GENOME ANNOTATION DATABASE

The ENSEMBL project was originally founded to computationally annotate the entirety of the human genome (Aken et al. 2017). The ENSEMBL database has since grown to include over 100 genomes animal genomes as well as plant, fungal, protist, and bacterial genomes in partner databases. ENSEMBL annotations consist of gene models for protein coding, small non-coding (microRNAs, small-nucleolar RNAs, ribosomal RNA, transfer RNAs, small nuclear RNAs), pseudogenes, and long non-coding RNAs. GENCODE creates its database of validated genes from the initial ENSEMBL annotation (Harrow et al. 2012). Similarly, to the RefSeq annotation process, ENSEMBL leverages *in silico* gene prediction and real transcriptomic datasets in order to construct models (**Section 1.5.7**). The ENSEMBL annotation process constructs gene models with unique gene (e.g. *ENSCGRG00000020358*) and transcript identifiers (e.g. *ENSCGRT00000027726*) as well as including gene symbol identifiers for evolutionarily conserved genes (e.g. *MALAT1*).

The annotation process is divided into three stages: the targeted stage, the similarity stage, followed by the merge stage. During the targeted stage, species-specific proteins are mapped to the genome using the Exonerate tool which can perform multiple pairwise alignments between protein queries and references¹. The targeted stage is followed by the similarity stage where proteins from closely related species are aligned to the assembly in order to build transcript models for evolutionarily conserved protein coding genes (Aken et al. 2016). During this stage, cDNA and EST sequences from the species of interest are mapped to the genome in order to extend transcript models and annotate UTRs. The annotation is further refined using RNA-Seq read data if available (Aken et al. 2016). During the final merge stage, overlapping transcripts are combined into a single gene model for a specific locus. Non-coding RNAs are annotated by aligning the Rfam reference database (Kalvari et al. 2017) to the genome using BLAST (Camacho et al. 2009). The Rfam reference database contains evolutionarily conserved secondary structures and sequences of non-coding RNAs which include long non-coding RNAs such as *MALAT1*, *NEAT1*, and *XIST*. The genomic *BLAST* hits are then filtered (E value $< 1 \times 10^{-5}$) to retain significant alignments in order to reduce search space and memory consumption (Aken et al. 2016). The loci with significant *BLAST* alignments are extracted from the genome and evaluated for the presence of conserved non-coding domains using *HMMER* (X. Cui et al. 2016). Please refer to **Section 2.1** for a discussion on the applications of *BLAST*, Rfam, and *HMMER* for identifying long non-coding RNAs. A long non-coding RNA is annotated as a “lincRNA” if it is intergenic and contains an ORF that covers less than 35% of the total transcript and does not contain a Pfam domain. The ENSEMBL method to annotate long non-coding RNAs often only identifies partial sequences and they are classified as “miscellaneous RNA” (Aken et al. 2016). A final reference database containing the localization on the genome is then generated for

¹ <https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate>

dissemination. There are two CHO cell line reference genomes that have been annotated by ENSEMBL and are discussed in **Section 1.5.8**.

1.5.1.3 THE REFSEQ CURATED, NON-REDUNDANT DATABASE OF GENOMES AND TRANSCRIPTOMES

The National Centre for Biotechnology Information (NCBI) administers a second comprehensive database of nucleotide and protein sequences of over 55,000 organisms of which approximately 10,000 are eukaryotes (Pruitt, Tatusova, and Maglott 2007). RefSeq entries can include manually curated data with experimental validation such as qPCR data or computationally predicted gene and transcript models from RNA-Seq or proteomics data (Pruitt, Tatusova, and Maglott 2007). RefSeq entries are constantly updated and re-evaluated based on additional data added to primary sequence repositories like the Sequence Read Archive (Pruitt, Tatusova, and Maglott 2007). Validated accessions contain the NM or NR prefix for protein coding and non-coding entries respectively whereas computationally predicted accessions contain the XM and XR prefixes.

The intensive RefSeq annotation process is performed for select eukaryotic organisms; often model organisms as well as industrially relevant species such as dogs, cows, and chickens (Pruitt, Tatusova, and Maglott 2007). The annotation process merges existing RefSeq annotations (if any), *de novo* transcriptome assemblies, and published RNA-Seq data into a unified annotation (Pruitt, Tatusova, and Maglott 2007). The RefSeq annotation also integrates user submitted annotation from the corresponding, peer-reviewed papers². All RNA-Seq data is aligned to the genome using BLAST and the splice-aware aligner, Splign (Camacho et al. 2009; Kapustin et al. 2008). Reads are aligned only to genomic contigs that are longer than 1 kb. In conjunction with RNA-Seq data, known RefSeq proteins and translated open reading frames from the non-curated GenBank repository are aligned to the genome. The genomic, transcriptomic, and proteomic gene

² The RefSeq Eukaryotic pipeline, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process

alignments are then used for constructing gene models using the Gnomon annotation pipeline (Souvorov et al. 2010). Gnomon first merges the putative alignments into transcript models then attempts to extend partial transcripts to identify start and stop codons using a Hidden Markov Model (Souvorov et al. 2010). Gnomon predictions are then aligned to the complete UniProt database of proteins (**Section 2.1.4.3**). A gene is annotated as a putative non-coding locus if there is not sufficient sequence coverage of the model in UniProt (Souvorov et al. 2010). A predicted gene model is considered validated if there is evidence of function within peer-reviewed literature (Pruitt, Tatusova, and Maglott 2007).

The RefSeq annotation is not as complete when compared to the GENCODE annotation process as it includes computationally predicted gene models. GENCODE is more stringent and integrates manually curated gene models where there is supporting evidence in high-throughput data (Harrow et al. 2012). The RefSeq pipeline infers the gene model by extending predictions to the transcription start and stop sites which might not be accurate (Frankish et al. 2015). The utility of GENCODE is severely limited for CHO transcriptome analysis as it only contains high-quality reference annotations for two very well characterized species, human and mouse, whereas NCBI's RefSeq repository contains annotations for over 10,000 eukaryotic organisms including the Chinese hamster genome. The GENCODE annotation strategy can be integrated with the existing RefSeq annotation in order to better characterize the Chinese hamster and CHO transcriptome.

1.5.2 Transcriptional and translational regulation by lncRNAs

Regulation by lncRNAs can be *cis* (on a nearby genomic locus) or *trans* where the target is on a distal genomic locus (Villegas and Zaphiropoulos 2015; Vance and Ponting 2014). Determining the regulatory target of lncRNAs is an ongoing challenge and requires significant *in vitro* validation (Baker 2011). Regulation of a specific gene can either be direct, where the ncRNA recruits effectors such as transcription factors to a locus, or indirect, where the ncRNA acts on an

upstream target to modulate expression (Nie et al. 2012). While *trans* regulators are challenging to identify, there is strong evidence of *cis* regulation by lncRNAs. Antisense transcripts, for example, have been found to regulate transcription of their neighboring sense protein coding gene (Wight and Werner 2013). lncRNAs have also been found to regulate transcription by remodeling chromatin states or by functioning as decoys by competing against regulatory targets of transcription factors and miRNAs. An important capability for lncRNAs in the context of industrial bioprocessing is their ability to regulate efficiency of protein translation in the cytoplasm. This section summarizes the functional regulatory roles of long non-coding RNAs as epigenetic modulators (**Section 1.5.2.1**), competing endogenous mimics (**Section 1.5.2.2**), and translational enhancers (**Section 1.5.2.3**). **Figure 1.7** summarizes several lncRNA-mediated regulatory mechanisms that have been identified in eukaryotic organisms.

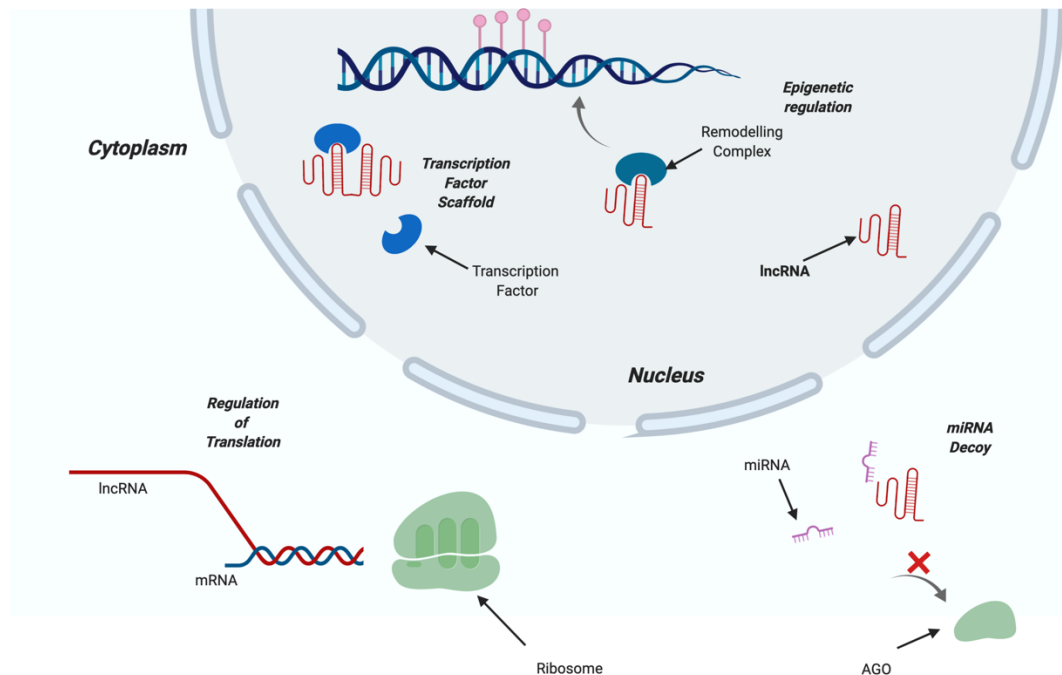


Figure 1.7 A diagram detailing the diverse regulatory mechanisms performed by lncRNAs within the nucleus and the cytoplasm in eukaryotic organisms.

1.5.2.1 Epigenetic modification of chromatin states to regulate transcription

Perhaps the most well studied regulatory function of lncRNAs is their ability to regulate chromatin states by recruiting chromatin remodeling complexes, such as polycomb repressive complexes (PRCs), to epigenetically modify DNA and histone complexes covalently (Mercer and Mattick 2013). The most studied epigenetic modification is methylation of histone complexes but other types of modifications occur such as ubiquitination, phosphorylation, and acetylation have been described. Epigenetic modifications can repress or enhance transcription at a locus (Mercer and Mattick 2013). CpG promoter methylation, for example, has been associated with transcriptional silencing (L. Liu, Jin, and Zhou 2015). While the mechanism for how lncRNAs recognize and bind to specific genomic regions is not well understood, it is known that they can discriminate between genomic regions and act as scaffolds and guides for the formation of protein complexes (K. C. Wang and Chang 2012). Secondary structures on the RNA molecule exist as binding sites for the chromatin remodeling complex. It is this ability that allows long non-coding RNAs to function as epigenetic modulators of gene transcription by activating or silencing promoters. *XIST*, one of the best characterized long non-coding RNAs, is a key regulator of X chromosome inactivation in mammals. Female mammalian embryos often have extraneous copies of this chromosome which can lead to developmental abnormalities (Penny et al. 1996). Mammals have evolved a mechanism to epigenetically silence the extra chromosome by recruiting polycomb repressive complex 2 (PRC2) to methylate histone complexes at the promoter (Leung and Panning 2014; Penny et al. 1996). *XIST* itself has an antisense *cis* lncRNA regulator, *TSIX*, which epigenetically modulates expression and demonstrates the existence of a wider, more complex lncRNA-lncRNA regulatory network (J. T. Lee, Davidow, and Warshawsky 1999). Other lncRNAs, such as *HOTAIR*, regulate transcription by scaffolding and localizing PRCs in a similar manner to *XIST* (K. Zhang et al. 2015).

Regulating epigenetic modulation of transcription is of particular importance in CHO cells as it has been implicated as one of the causes of bioprocessing heterogeneity (Veith et al. 2016). Variable volumetric yields can cause delays in production due to variable product quality or titer. Methylation of transgene viral promoters has been linked to transcriptional silencing and low production (Yang et al. 2010; Osterlehner et al. 2011). Due to promoter silencing, fewer transgene mRNAs are transcribed and translated. A methylation inhibitor can be added to the media but this is undesirable as they are often toxic (Yang et al. 2010). lncRNA-mediated inhibition of this behavior by the cell would reduce titer variability between clones.

1.5.2.2 Transcription factor and miRNA decoys and mimics

Related to their ability function as scaffolds and guides is the ability of long non-coding RNAs to function as *in vivo* mimics and decoys for miRNAs and transcription factors. Transcription factor (TF) decoys function by competitively binding and titrating away free-floating TFs from their targets as they contain RNA domains with structures similar to the TF target (Puvvula et al. 2014). Recently, it was discovered that the long intergenic ncRNA, *PANDA*, is a repressor of *P53*-mediated apoptosis by preventing the transcription factor *NF- κ B* from binding to the chromatin (Hung et al. 2011). *NF- κ B* binds to the *FAS* promoter which triggers the apoptosis pathway in the cell. Depletion of *PANDA* decreased cell survival as apoptosis was upregulated.

lncRNAs also function as competing endogenous sponges (referred to as ceRNAs) to titrate miRNAs away from targets (J.-H. Yoon, Abdelmohsen, and Gorospe 2014). They are thought to contain identical binding sites to the miRNA target and prevent translational inhibition by sequestration. It is thought that the endogenous sponge competes for the miRNA/RISC complex thereby upregulating the miRNA target transcript (Thomson and Dinger 2016). *Ptenp1*, the first identified endogenous miRNA decoy, functions as a regulator of the tumor suppressor *PTEN* by competing with its miRNA, miR-21 (Johnsson et al. 2013). Loss of the *PTENP1* locus has been linked with gastric cancer progression (Xia et al. 2014). ceRNAs can in turn be transcriptionally

regulated and degraded by miRNAs indicating the presence of a wider, more diverse ceRNA-miRNA-mRNA regulatory network.

1.5.2.3 Translational regulation by long non-coding RNAs

Recent studies suggest that long non-coding RNAs play an important role in translational regulation that is independent of their role as transcriptional regulators. lncRNAs have also been localized to the cytoplasm where they are thought to regulate protein translation which is thought to be mediated by binding and recruiting RNA binding proteins to an mRNA sequence (Kung, Colognori, and Lee 2013). Translational repression occurs at bound polysome complexes despite lncRNAs not coding for proteins (J. Yoon et al. 2012). The intergenic non-coding RNA, lincRNA-p21, inhibits translation of *CTNNB1* and *JUNB* after upregulation. lincRNA-p21 co-localizes the translational repressor *RCK* to the mRNA-polysome complex which then inhibits mRNA translation (J. Yoon et al. 2012).

lncRNAs have been demonstrated to increase endogenous protein translation which is important in the context of protein bioprocessing. Carrieri *et. al.* identified the *UCHL1* antisense locus as a translational enhancer of its sense protein coding gene, ubiquitin carboxyl-terminal hydrolase L1 (Carrieri et al. 2012). Translational upregulation was found to be independent of abundance of transcribed *UCHL1* mRNAs (Carrieri et al. 2012). The antisense transcript is transcribed in a bi-directional 5' "head to head" manner with partial overlap of the 5'UTR and the AUG start codon of the *UCHL1* gene. The overlap is important in targeting the antisense transcript to mRNA in the cytoplasm (Carrieri et al. 2012). The non-overlapping sequence contains the inverted short interspersed nuclear element (SINE), SINEB2, which function as the effector domain for increased translation. The molecular mechanism driving translational upregulation by the SINEB2 element is currently not understood but the presence of this element has been demonstrated to increase target protein production 2 fold (Zucchelli et al. 2015).

1.5.3 State of the annotation of lncRNAs in the *C. griseus* and CHO genomes

Advances in Next Generation Sequencing technology enabled high-throughput profiling of CHO cell lines used in industrial bioprocessing. The first CHO genome was published in 2011 (**Section 1.5.8.1**) with a second cell line-derived reference genome published in 2017 (**Section 1.5.8.2**) as well as a *C. griseus* genome in 2013 (**Section 1.5.8.3**). Unlike the genomes of model organisms like humans and mice which are organized into chromosome level features, the *C. griseus* and cell line genomes are composed of shorter sequences organized into scaffolds which are contiguous genomic loci which may not be complete. The GENCODE consortium was able to identify a more complete set of lncRNAs in the genome due to the availability of high quality reference genomes (Harrow et al. 2012; Mudge and Harrow 2015). A high-quality reference genome is required for long non-coding RNA analysis and any future lncRNA annotations must be informative and identify which class each non-coding RNA belongs to and reference conserved orthologues. This section will summarize the state of annotation of long non-coding RNAs of two CHO-K1 derived genomes (**Section 1.5.8.1-1.5.8.2**) and a *C. griseus* genome (**Section 1.5.8.3**) as well as a more complete genome sequence that used long read sequencing to complete long scaffolds (**Section 1.5.8.4**) This section will also summarize research in the only long non-coding RNA that has been characterized and experimentally validated by knock-out studies in CHO cells (**Section 1.5.8.5**), *GADD7*.

1.5.3.1 CHO-K1 GENOME

Xu *et al.* published the first CHO reference genome of a CHO-K1 cell line using Illumina sequencing with a range of insert sizes (200 bp to 20 kb) (X. Xu et al. 2011). The authors *de novo* assembled this genome into 109,152 scaffolds with the longest scaffold being 8,779,783 bp long (X. Xu et al. 2011). The assembled genome was approximated 2.3 Gb in length. The scaffold N50 is defined as the length of the longest scaffold that contains 50% of the genome or higher. The authors annotated the genome using homology from human, mouse, and rat protein coding genes

found in the ENSEMBL database (X. Xu et al. 2011). An ENSEMBL reference annotation was published in September 2017 (ENSEMBL Assembly ID: CriGri1, Genbank Assembly ID: GCA_000223135.1) which contains 19,617 protein coding genes and 3,332 long non-coding genes (**Table 1.4**). This lncRNA reference annotation is divided into 2,563 “lincRNAs” and 769 “miscellaneous RNAs” (**Section 1.5.6**). The lincRNAs class are all *C. griseus* specific intergenic RNAs whereas the “miscellaneous RNAs” are composed of evolutionarily conserved long non-coding RNAs such as *NEAT1*, *MALAT1*, and *XIST*.

1.5.3.2 HORIZON CHO-K1 GS GENOME

Horizon Discovery, a life science company, published a genome of a CHO-K1 *Glutamine synthetase* knockout cell line in collaboration with the Wellcome Trust Sanger Institute in 2017 (ENSEMBL Assembly ID: CHOK1GS_HDv1, Genbank Assembly ID: GCA_900186095). The sequencing strategy leveraged both Illumina short read sequencing as well as Illumina Synthetic Long Read and Dovetail library sequencing³ in order to generate longer, more complete scaffolds. 8,265 scaffolds were assembled covering greater than 90% of the genome with a scaffold N50 length of 61,810,576 which is significantly longer than either the previous CHO-K1 genome published by Xu *et al.* or the *C. griseus* genome. An ENSEMBL reference annotation was released simultaneously for the CHO-K1 GS null genome and has the smallest number of long non-coding RNAs (775) and are exclusively classified as “miscellaneous RNAs” (**Table 1.4**). These “miscellaneous RNA” annotations in both the previous CHO-K1 and the CHO-K1 GS null genomes often do not cover the full length of the long non-coding gene. For example, the *MALAT1* locus is expected to encode for a large (~7 kb) transcript but the reference annotation is significantly shorter. The shorter annotation can be attributed to ENSEMBL’s method for

³ A discussion on Illumina Synthetic Long Read and Dovetail sequencing methods is outside the scope of this thesis as they have been developed for whole genome sequencing. Please refer to Illumina, Inc for an overview of Synthetic Long Read technology (<https://www.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html>). Please refer to Dovetail Genomics, Inc for an overview of the Dovetail technology(https://dovetailgenomics.com/wp-content/uploads/2018/02/DTG_Hi-C-Kit_product-highlight.pdf).

annotating lncRNAs as this process can only identify regions that contain evolutionarily conserved motifs and would thus miss the full-length sequence. The “miscellaneous RNAs” in the CHO-K1 GS null reference correlate very closely with those annotated in the previous Xu *et al.* CHO-K1 genome with only two unique to this assembly (*Fmr1-As1* and *Smcr2*).

1.5.3.3 C. GRISEUS GENOME

Lewis *et al.* followed the publication of the CHOK1 genome with the first assembly of the *C. griseus* genome in 2013. The authors constructed Illumina Next Generation genomic libraries from Chinese hamster tissue samples (N. E. Lewis et al. 2013). The assembled genome was approximately 2.4 Gb in length which equates to 89% of the estimated *C. griseus* genome length (N. E. Lewis et al. 2013) with a scaffold N50 of 1.54 Mb. Parallel Illumina RNA-Seq libraries were constructed alongside genomic libraries and used to construct gene and transcript models using TopHat2 and Cufflinks (**Section 2.1.2**). Analysis of the transcriptomic libraries focused on protein coding genes and relied on homology to reference databases such as SwissProt with a total of 24,044 genes identified using this method (N. E. Lewis et al. 2013). The *C. griseus* genome was later fully annotated by RefSeq in 2015 (GCA_000419365.1) with protein coding, small non-coding, pseudogene, and long non-coding loci. The RefSeq annotation of the *C. griseus* genome contains 2,879 computationally predicted long non-coding RNAs (**Section 1.5.6**). Unlike the GENCODE reference, these lncRNAs are not classified based on their genomic localization. The lncRNA annotation is composed almost entirely of RefSeq locus tags which are simply numeric IDs that are uninformative. For example, the well-studied, evolutionarily conserved lncRNA *Neat1* is referenced simply as *LOC193159497*.

1.5.3.4 C. GRISEUS PACBIO SEQUENCED GENOME (PICR)

Rupp *et al.* used long read Pac Bio sequencing to generate a more complete genomic reference (ENSEMBL genome ID CriGri-PICR) derived from *C. griseus* liver tissue (Rupp et al. 2018). The authors used a hybrid assembly approach where DNA derived from liver tissue was first sequenced using long read sequencing and resulting gaps and errors were corrected using all

previous genome resequencing data (X. Xu et al. 2011; N. E. Lewis et al. 2013; Brinkrolf et al. 2013). The assembled scaffold was approximately 2.36 Gb in length which is the most concise genome published so far with only 1,830 scaffolds. The published genome was later annotated using the ENSEMBL pipeline with a total of 22,028 protein coding genes and 3,002 computationally predicted lncRNAs (2,239 lncRNAs and 763 “misc RNAs”).

Table 1.4 A table of lncRNA genes found in the CHO-K1, CHOK1GS, PICR, and *C.griseus* genomes. All genes were classified into intergenic, antisense, and sense loci with respect to nearby protein coding genes using the GENCODE definitions. This classification was performed using the *BEDtools* suite as detailed in **Chapter 2**.

Class	CHO-K1	CHOK1GS	PICR	<i>C. griseus</i>
Intergenic	2,777	537	2,716	2,374
Antisense	305	103	124	462
Sense	231	135	142	45

1.5.3.5 GROWTH ARREST AND DNA DAMAGE 7 GADD7

The *GADD7* gene, also referred to as *ADAPT15*, is identified as a “predicted” gene in the RefSeq and ENSEMBL annotation despite it being the only experimentally verified and functionally validated lncRNA in the Chinese hamster genome. The growth arrest and DNA damage 7, or *GADD7*, gene was first identified in CHO cells in 1996 (Hollander, Alamo, and Fornace 1996). Hollander *et al.* found that this lncRNA is expressed after exposure to DNA damaging elements such as UV radiation, toxic compounds such as methanosulfate (MMS), and oxidizing agents such as hydrogen peroxide (Hollander, Alamo, and Fornace 1996). *GADD7* was also found to be overexpressed after nutrient starvation in culture (Hollander, Alamo, and Fornace 1996). Overexpression of *GADD7* inhibited CHO cell densities by 85% (Hollander, Alamo, and Fornace 1996). Depletion of *GADD7* was found to delay the onset of the Endoplasmic Reticulum stress response after exposure of CHO-K1 cells to reactive oxidative species (Brookheart et al. 2009). There has not been any significant research devoted to studying long non-coding RNAs in the intervening years but the discovery of *GADD7* demonstrated the utility of studying long non-coding RNAs outside of the CHO research community. The lack of proper annotation in the reference

database for this lncRNA underscores an important challenge: correctly integrating peer-reviewed experimental results into a reference. A comprehensive reference annotation would need to accurately integrate functional information into a format that is useful for CHO cell biologists.

1.6 Challenges in profiling long non-coding RNAs

The revolution in next generation sequencing technology has enabled lncRNA discovery efforts. Thousands of lncRNAs have been detected in many diverse species such as humans, mice, zebrafish, and rice due to advances in genome and transcriptome sequencing. Detecting long non-coding RNAs can be quite challenging, however. They are often expressed at lower abundances than protein coding transcripts (Derrien et al. 2012). Furthermore, many lncRNAs are tissue and phenotype specific and can be transiently expressed which can be mitigated by increasing the depth of sequencing with a corresponding increase in project costs.

lncRNA profiling requires a high quality genome for accurate detection and annotation (Z. Wang, Gerstein, and Snyder 2009; Atkinson, Marguerat, and Bähler 2012). The majority of long non-coding RNA research has been performed in well annotated organisms. Localization on the genome is a major metric for distinguishing coding and non-coding loci which requires a well-annotated genome with accurate exon boundaries and strand information for each locus in order to differentiate sense from antisense and genic from intergenic. Currently, the *C. griseus* genome is in an early stage and is scaffolded into contigs instead of chromosomes. Gaps in the assembly would hamper accurate read mapping and transcript reconstruction (X. Chen and Tompa 2010). Furthermore, the majority of annotated transcripts in the genome are computationally predicted based on homology and many have not been validated *in vitro*. This complicates resolution of sense overlapping non-coding transcripts due to uncertainty of the accuracy of read mapping.

Accurately distinguishing between coding and non-coding transcripts can be challenging. Unlike protein coding sequences, lncRNAs show very little evolutionary conservation at a nucleotide level. Methods that rely solely on sequence homology, which are ideal for protein coding genes, are unsuitable for detecting lncRNAs (Johnsson et al. 2014b). There is evidence that exon order and secondary structure are evolutionarily conserved so comparative genomics can be used to identify syntenic regions which are genes that have conserved exon structure and location

between species but might not share conserved base pair sequences (J. Chen et al. 2016). Detecting syntenic loci requires high quality reference genomes for alignment. Furthermore, as lncRNAs exhibit a higher rate of evolution when compared to protein coding genes (Quinn et al. 2016), these strategies would miss species-specific sequences (Housman and Ulitsky 2016). Novel bioinformatics resources which can fully profile and annotate long non-coding transcripts in high-throughput sequencing data are being developed to tackle this challenge.

1.7 The promises of lncRNA based genetic engineering

Chinese hamster ovary cells have been described as the “workhorse” for recombinant therapeutic production as they can produce proteins with human compatible post-translational modifications and grow to high cell densities in chemically defined, synthetic media (Hacker, Jesus, and Wurm 2009). Because of their importance, significant time and resources have been dedicated to improving product titre. The past 30 years have seen dramatic increase in volumetric yield due to advances in bioreactor engineering, media formulation, and vector design. Genetic engineering has the potential for further improvement by improving phenotypes ideal for protein production such as reduced apoptosis and increased growth rate.

Long non-coding RNAs offer significant promise for novel cell line engineering approaches despite the challenges with discovering and functionally validating them. The primary challenge is the dearth of knowledge on the regulatory roles of many lncRNA transcripts. Advances in high throughput sequencing technology have enabled the discovery of thousands of lncRNAs. Previous work has demonstrated their roles as key regulators in pathways relevant to cell line engineering such as apoptosis, cell cycle progression, and cell growth (Geisler and Coller 2013). Further functional validation can identify other significant roles for lncRNAs.

lncRNA-mediated epigenetic silencing offers an attractive method to finely modulate cellular phenotype by selectively silencing or enhancing specific genes to control specific pathways. Production heterogeneity has been linked to silencing of transgene promoters by the cell and makes an attractive target for lncRNA guided inhibition to prevent undesirable behavior. Furthermore, the ability of lncRNAs to function as miRNA and transcription factor decoys will be potentially useful in directing cell phenotypes. Synthetic sponges, constructs containing response elements complementary to a miRNA, are widely used in CHO cell engineering. Studying endogenous sponges will provide better understanding of the CHO transcriptome and can be used to design more efficacious synthetic sponges.

lncRNAs have already been used to improve translation in CHO cells by Zucchelli *et al* (Zucchelli et al. 2015). The authors demonstrated SINEUPs, a construct containing a binding domain (BD), which partially overlaps the 5' of the target mRNA (Zucchelli et al. 2015) and provides target discrimination by base-pairing at the overlap, and an effector domain (ED) containing an inverted SINEB2 element. The molecular mechanism underlying translational upregulation by the inverted repeat is not well understood (Patrucco et al. 2015; Zucchelli et al. 2015). Transient transfection targeting translational upregulation of the endogenous *Dj-1* gene in HEK 293 cells saw 1.7-2-fold-change in protein quantity. SINEUP technology has been used to specifically upregulate a luciferase construct (Patrucco et al. 2015) in CHO-S cells as a model for recombinant protein production. A 50% increase in luciferase was observed after transfection with the SINEUP construct. Similar results were observed with a new recombinant construct expressing a single chain antibody fragment (ScFv) and a secretory leader sequence. SINEUP transfection increased ScFv secretion 2 fold (Patrucco et al. 2015).

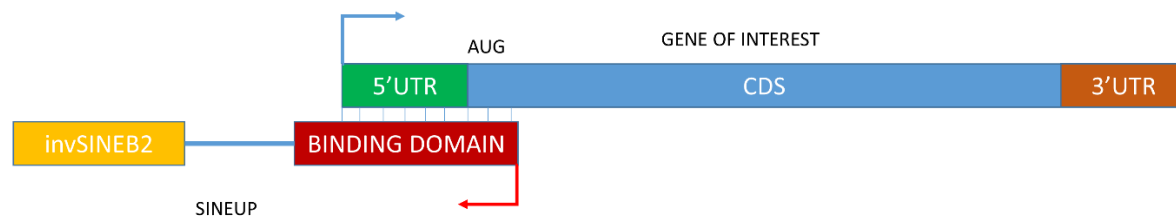


Figure 1.8 Modular design of antisense SINEUP transcripts. The binding domain is thought to form Watson-Crick base pairs with the target mRNA in the 5' end. The binding domain should overlap the 5' UTR and encompass the AUG translation start site of the CDS sequence. The binding domain connected the inverted SINEB2 translational effector domain by a linker sequence.

Long non-coding RNA-mediated translational engineering can be utilized to up or downregulate endogenous and recombinant genes. They have already been tested in CHO cells and achieved positive results. Future work will involve discovering novel targets for RNA mediated translational modulation as well as screening for stronger effectors. The success of SINEUP technology demonstrates the potential for lncRNA based cellular engineering. lncRNAs need to be better

surveyed within the CHO transcriptome for this potential to be achieved. The aim of this project is to better characterize the *C. griseus* long non-coding transcriptome and provide viable cellular engineering targets for the greater CHO community.

1.8 Project Aims

- **Chapter 2**
 - To develop bioinformatics software toolsets and pipelines that detect, characterize, and annotate long non-coding RNAs using total RNA-Seq data.
 - To profile and characterize long non-coding RNAs in Next Generation Sequencing data of CHO cells and associate their differential expression with a reduction in culture temperature.
- **Chapter 3**
 - To correlate differential expression of mRNAs and ncRNAs to metabolite concentrations and bioprocess parameters as shown over the course of a fed-batch culture.
- **Chapter 4**
- To leverage high-coverage RNA-Seq data to study alternative splicing of coding and lncRNAs over the duration of a fed-batch culture.
- **Chapter 5**
 - To profile transcriptionally active chromatin states using the novel ATAC-Seq protocol in order to refine lncRNA assembly quality and identify active regulatory regions within the CHO genome by integrating RNA-Seq
 - To utilize ATAC-Seq data to find stable, highly accessible regions within the genome that are stable loci for transgene integration.
- **Overall Aims**
 - To prioritize targets for genetic engineering to generate cells with industrially beneficial phenotypes.

- To provide a well-annotated resource of long non-coding RNAs in the CHO genome as well as publically available computational pipelines to analyse lncRNAs within RNA-Seq

Chapter 2

Differential Expression of lncRNAs After Temperature Shift

2 Summary

In this chapter, we first summarize a method to assemble transcript models using total RNA-Seq data in order to discover and annotate lncRNAs (**Section 2.1**). We present a software pipeline that can identify high confidence lncRNAs within RNA-Seq data and use this pipeline to profile lncRNAs in a transcriptomic study profiling CHO cells that underwent a reduction in culture temperature from 37°C to 31°C (**Section 2.2**). Our software pipeline was used to identify high confidence lncRNAs in CHO cells as well as identified lncRNAs that were differentially expressed after a reduction in culture temperature (**Section 2.3**). This chapter will conclude with a discussion of the utility of our software pipeline for the wider CHO community as well as the biological significance of differentially expressed lncRNAs in CHO cells (**Section 2.4-2.5**).

2.1 Introduction

lncRNAs have been implicated as regulators of diverse cellular processes such as X chromosome inactivation, cell senescence, proliferation, and apoptosis (Kung, Colognori, and Lee 2013). The molecular mechanism of these non-coding molecules is equally diverse with functions such as titrating miRNAs from the cytoplasm, functioning as scaffolds for transcription factors and other proteins, stabilizing mRNAs during translation, and recruiting chromatin remodeling complexes to regulatory loci within the genome to modify chromatin (Quinn and Chang 2015). These diverse regulatory targets and mechanisms suggest the utility of lncRNAs for rational cell engineering of CHO cells in order to enhance bioprocessing phenotypes like growth rate or productivity (Patrucco et al. 2015). A significant impediment to these engineering efforts is the lack of a high quality lncRNA reference annotation as well as the absence of robust computational pipeline for easy analysis.

Several CHO and parental Chinese hamster genomes have been published which was made possible due to advances in next generation sequencing (X. Xu et al. 2011; N. E. Lewis et al. 2013; Brinkrolf et al. 2013; Rupp et al. 2018). These reference genomes have been annotated by consortiums like *ENSEMBL* and RefSeq providing a valuable asset for the CHO community with the genomic localization of coding genes. Long non-coding loci are poorly annotated within these references, however, with the *ENSEMBL* reference annotation representing less than 20% of a theoretical total ~14-16,000 when compared to the over 14,000 genes annotated within the human and mouse GENCODE annotations. In comparison, the *ENSEMBL* CHOK1 *criGri1* reference annotation, which has the most number of annotated long non-coding transcripts, contains only 3,322 loci. The full length sequence of several well-characterized, evolutionarily conserved non-coding genes such as *MALAT1* or *PVT1*, is absent within this assembly due to the reliance on localized homology by automated pipelines (Aken et al. 2017). For example, the human and mouse reference annotation for *MALAT1* encompasses an ~22 kb gene while the version 98 annotation of the *criGri1* CHOK1 genome has a short, < 200bp gene locus. The long non-coding

transcriptome has to be significantly expanded to include more full-length sequences to enable rational cell engineering.

Another significant obstacle for lncRNA-mediated genetic engineering is the lack of information on biological function for many non-coding transcripts. Only a fraction of long non-coding genes have been characterized in model organisms such as humans and mice despite the availability of high quality reference annotation such as GENCODE (Zampetaki, Albrecht, and Steinhofel 2018). *GADD7* remains the only functionally characterized long non-coding RNA within the CHO genome despite being discovered 20 years ago. The development of strand-specific RNA-Seq enables genome-wide discovery of long non-coding RNAs which can be used to relate expression with alterations in phenotype (Ilott and Ponting 2013).

In this chapter, we describe the utilization of high-resolution RNA-Seq data to significantly expand the landscape of lncRNAs within the CHO genome. These data were acquired from a recombinant CHO cell line that underwent a temperature shift from 37°C to 31°C. Temperature shift, or a reduction in culture temperature, is another commonly used process that can be used to increase recombinant protein production (Moore et al. 1997; Wulhfard et al. 2008). CHO cells are commonly cultivated at 37°C in order to achieve high cell densities after which the culture temperature is reduced to 30-33°C in order to prolong cell culture viability (J. Xu et al. 2019). Prolonging culture duration is highly desirable as cells remain productive over a longer period of time, thereby increasing net product titers (Butler and Meneses-Acosta 2012).

2.2 A workflow for lncRNA discovery

The first step of RNA-Seq library preparation is the fragmentation of full length transcripts into shorter sequences (**Section 1.4.3**) followed by ligation of synthetic oligonucleotides adaptors to the ends of each fragment. **Section 2.1** summarizes the major stages of lncRNA analysis process from transcript assembly to classification. Reads must first be pre-processed to remove artefactual oligonucleotides after which filtered reads are aligned to a reference genome and short reads are collapsed and extended to form gene and transcript models (**Section 2.1.1**). Putative lncRNAs are then evaluated for coding potential using machine learning algorithms (**Section 2.1.2**) and genes containing protein coding domains removed (**Section 2.1.3**). The putative lncRNAs can then be classified into sense, antisense, bi-directional, or intergenic based on localization on the genome with respect to a nearby protein coding gene (**Section 2.1.**). These classified lncRNAs can then be evaluated for differential expression using tools such as DESeq2 or edgeR (Love, Huber, and Anders 2014)

2.2.1 Constructing gene and transcript models from NGS data

During the NGS library preparation process, longer transcripts are fragmented into short reads and sequenced. Information on the full-length cDNA sequence is lost so these short reads must then be computationally reconstructed in order to generate biologically relevant transcript models. The process of combining short reads is referred to as “assembly” and can use reference guided methods where reads are first aligned to a reference genome and gene and transcript models are computed from read coverage across loci (Pertea et al. 2015). Reference guided assembly methods are limited by the quality of the genome available during mapping. If there are gaps within a locus, reads would not map properly across the region and a transcript model could not be reconstructed for that loci.

Reference guided assemblers, such as *StringTie*, rely on constructing a network graph of overlapping clusters of reads (Pertea et al. 2015). The first step of the *StringTie* is to construct a splice graph where overlapping reads are clustered and abundances of putative loci are estimated

(Pertea et al. 2015). The splice graph is traversed using coverage across each cluster as a guide (Pertea et al. 2015) using a flow graph with the path which containing the most covered clusters is considered the most likely isoform. *StringTie* can also assemble different isoforms of a gene using this two-step network graph method by grouping regions of the highest coverage into the dominant isoform with regions of lower coverage divided into secondary and tertiary isoforms in a process referred to as isoform deconvolution (**Figure 2.1**). The deconvolution step performed by *StringTie* can be used to perform gene and transcript level analysis and the assembled transcripts can be assessed for coding potential.

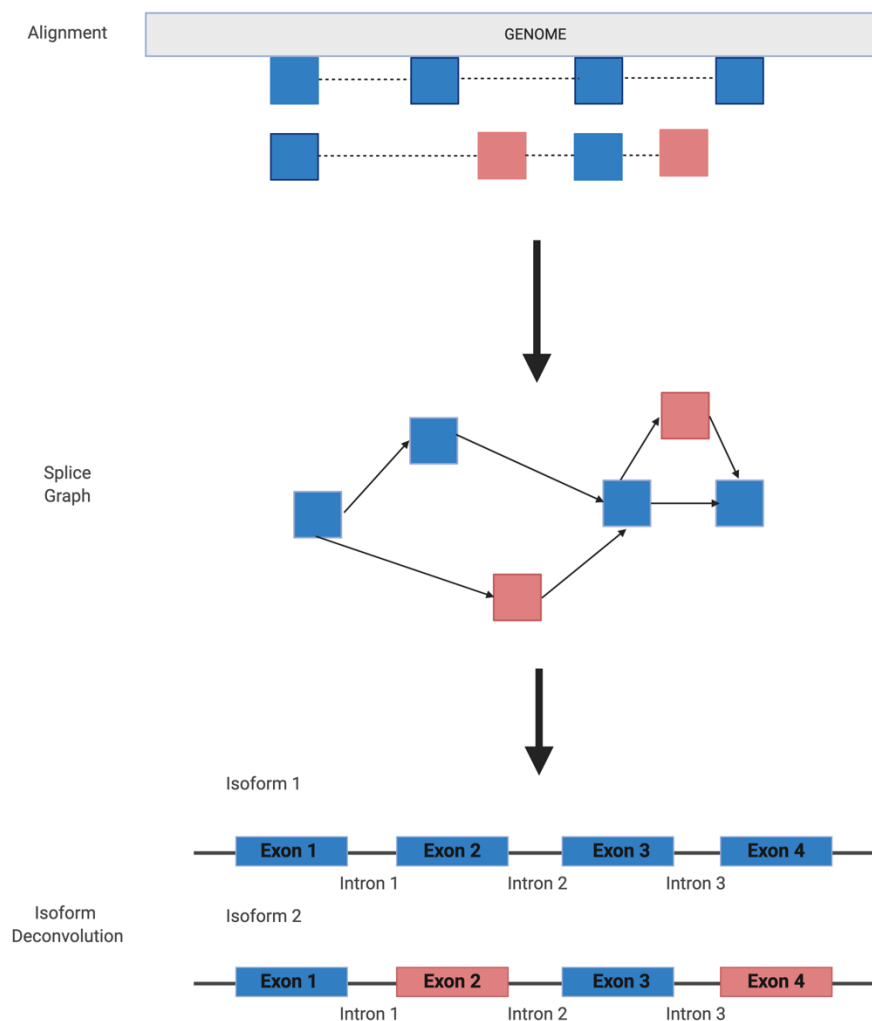


Figure 2.1 StringTie assembles transcripts using a splice graph with a network flow model where dominant isoforms are first constructed using the regions of highest coverage. The graph model is further refined using regions of lower coverage which are assumed to be other isoforms within the gene.

2.2.2 Coding potential calculators

The assembly process for detecting lncRNA and mRNA is identical as the total RNA-Seq read data is used during the reference guided or *de novo* assembly. The challenge for a lncRNA specific workflow is distinguishing long non-coding RNAs from protein coding transcripts (Kong et al. 2007; L. Wang et al. 2013). To this end, several software suites have been designed to segregate protein coding transcripts (Housman and Ulitsky 2016) by calculating the likelihood that a given transcript codes for a protein. Machine learning algorithms are used to train a model which contains known lncRNAs and protein coding genes in order to provide a reference. The trained model can then be used to evaluate novel transcripts for coding potential (Housman and Ulitsky 2016). Several metrics can be used to train these models: open reading frame length (ORF), codon composition, sequence substitution patterns, and sequence similarity to known protein coding genes (Kong et al. 2007; L. Wang et al. 2013; Lin, Jungreis, and Kellis 2011; Housman and Ulitsky 2016).

The open reading frame is the sequence of an mRNA that is likely to be translated into a protein with the assumption that lncRNAs will have shorter ORFs than protein coding transcripts (Dinger et al. 2008). ORF coverage, the percent a transcript that is composed by an ORF, and ORF length has been utilized as a metric to evaluate coding potential by tools such as *CPAT*, the Coding Potential Assessment Tool (L. Wang et al. 2013). As one of the definitions of a lncRNA is a transcript with an ORF < 100 amino acids long, these metrics can be used as a threshold to predict coding likelihood. ORF coverage and ORF length alone are not sufficient to classify a transcript as non-coding as many *bona fide* lncRNAs have open reading frames longer than 100 amino acids (Dinger et al. 2008). *XIST*, for example, has an ORF length of ~293 amino acids and is, therefore, longer than the definitive length of 100 amino acids albeit with a low ORF coverage. Nucleotide and codon frequency, in conjunction with ORF length and coverage, can also be

utilized to segregate protein coding and non-coding transcripts. Nucleotide frequency of protein coding transcripts will be “ordered” where certain codons will be more frequent. lncRNA transcripts, in contrast, exhibit a more random sequence composition as the primary structure is not essential to their function (Housman and Ulitsky 2016). Similarly, to sequence composition, codon substitution frequency can be utilized as an alternative method to evaluate coding potential. There is stronger evolutionary selection pressure to preserve nucleotide sequences in protein coding genes than non-coding and the frequency of a codon to be substituted with another can be extracted from a multiple sequence alignment to evaluate sequence conservation (Lin, Jungreis, and Kellis 2011). This section will review several commonly used machine learning algorithms, such as *CPC2*, *CPAT*, and *FEELnc*) that can be used to evaluate coding potential.

2.2.2.1 Coding Potential Calculator 2

Kang *et al.* developed Coding Potential Calculator 2 (*CPC2*) to use support vector machines to segregate coding and non-coding transcripts (Kang et al. 2017a). *CPC2* uses four intrinsic features of a transcript to calculate coding potential: ORF length, ORF integrity, calculated isoelectric point (pI), and the Fickett TESTCODE metric. *CPC2* calculates the longest ORF length for each putative lncRNA with shorter ORF lengths correlating to lack of coding potential. The ORF integrity is a binary value assigned by *CPC2* if the longest ORF begins with an AUG start codon and a canonical stop coding (UGA, UAG, or UAA). The isoelectric point is the estimated pH where the putative ORF has a net neutral charge with lncRNAs having highly variable pIs. The Fickett TESTCODE was initially proposed in 1982 as a method to calculate the compositional bias of each nucleotide at each position of a codon across the length of the transcript (Fickett 1982; L. Wang et al. 2013).

N_1 = Number of each nucleotide at position 1,4,7...

N_2 = Number of each nucleotide at position 2,5,8...

N_3 = Number of each nucleotide at position 3,6,9...

$$N_{pos} = \frac{MAX(N_1, N_2, N_3)}{MIN(N_1, N_2, N_3) + 1}$$

Where N is calculated for each nucleotide (A, C, G, T) at each codon position across the transcript. For example, A_1 is the number of adenosines in the first position of each codon whereas A_3 is the number of A's at the terminus of each codon along the total length of the transcript. A probability p is calculated for each N_{pos} and is converted to a probability score using a reference table calculated by Fickett and is a factor of the percent composition of each base across the length of the transcript (Fickett 1982). The advantage of the Fickett score is that it is independent of the ORF length or coverage and can be quite accurate in judging the coding potential of a transcript (Housman and Ulitsky 2016). All four metrics are calculated for the reference set and used to train a model to evaluate experimental data and assign coding potential (Kang et al. 2017b).

2.2.2.2 Coding Potential Assessment Tool

Like *CPC2*, the Coding Potential Assessment Tool (*CPAT*) uses an alignment free method to estimate coding potential by utilizing ORF coverage and sequence composition. An initial data set of the coding sequences of mRNAs and validated lncRNA transcripts are used to train a model and a coding potential threshold is calculated (L. Wang et al. 2013) which can then be compared to experimentally-derived transcripts. A transcript can be considered protein coding if it has a coding potential score above the calculated cut-off value. *CPAT* calculates coding potential using four metrics: the length of the longest three frame translated ORF, the ORF coverage, the Fickett score, and the hexamer usage bias score. ORF length and ORF coverage can be useful as lncRNAs are likely to have short open reading frames when compared to protein coding transcripts. *CPAT* assigns a Fickett score by calculating the sum of the products with a weight score which is the probability that each base successfully predicted coding and non-coding potential. *CPAT* also calculates a more sensitive metric known as the hexamer usage bias which can be used to calculate the occurrence of a specific hexamer across the transcript. The hexamer

usage bias is calculated using the training dataset provided when constructing the model by *CPAT* (L. Wang et al. 2013). The hexamer usage bias score is the logarithm of the ratio of the hexamers in a sequence calculated under the assumption that the sequence is protein coding versus non-coding. A positive hexamer usage bias score denotes a putative protein coding sequence whereas a negative score denotes a non-coding sequence.

When combined, all four metrics can be used to accurately segregate protein coding sequences from non-coding. A protein coding sequence would have a longer ORF, greater ORF coverage, a higher Fickett score, and a higher hexamer usage bias score as codon usage would be more ordered (i.e. less random) than in non-coding sequences. Cross-species classification can be utilized to classify lncRNAs using a well annotated genome such as the mouse transcriptome for classifying Chinese hamster transcripts. Cross-species classification can be limited by evolutionary distance between the training dataset and the species being studied as closely related species are needed to construct a training model. *CPAT* also has the advantage of a low memory requirement and runtime; we have observed that ~60,000 transcripts can be classified rapidly under a minute which is considerably shorter than most other programs.

2.2.2.3 FIEExible Extraction of lncRNAs

The FIEExible Extraction of lncRNAs (*FEELnc*) tool is a more recent set of tools that uses a Random Forest machine learning algorithm to classify transcripts (Wucher et al. 2016). Like *CPAT* and *CPC2*, *FEELnc* calculates coding potential by using an alignment-free method. It utilizes a *k*-mer frequency counting method to identify sequence composition between coding and non-coding transcripts. *K*-mers, in the context of transcriptomics and genomics, are all the possible substrings of length *k* that compose a larger sequence such as an NGS read or transcript. One of the advantages of *FEELnc* over *CPAT* is that it does not require a known lncRNA dataset for its training algorithm. Wucher *et al.* detail two distinct methods that *FEELnc* can be used to train models for classification which they term “shuffle” and “intergenic”(Wucher et al. 2016) which are

useful for performing analysis in non-model organisms which might lack annotated lncRNAs. The shuffle method randomly selects k -mer frequencies of provided mRNAs in order to simulate a set of lncRNAs as there is evidence that lncRNAs, such as *XIST*, have evolved from protein coding genes that have lost function (Johnsson et al. 2014b). The intergenic method randomly extracts intergenic regions of the provided genome and these “decoy sequences” can be used to model non-coding transcripts. *FEELnc* can also use a set of known mRNAs and lncRNAs to train a model in addition to the shuffle and intergenic methods. Wucher *et al.* specify that the shuffle method outperformed the intergenic method when classifying a set of GENCODE validated lncRNAs (Wucher et al. 2016).

During training, *FEELnc* computes a score for each 1-mer, 2-mer, 3-mer, 6-mer, 9-mer, and 12-mer in the dataset. The score can be calculated as the factor of the frequency of a specific k -mer in the mRNA dataset over the frequency of that k -mer in both coding and non-coding sequences.

$$S_K^k = \frac{F_K^m}{F_K^m + F_K^{lnc}}$$

Where K is the k -mer sequence, k is the length of the k -mer, F_K^m is the frequency of the k -mer in mRNA training data and F_K^{lnc} is the frequency of the k -mer in the lncRNA training data. A final score is calculated as the product of all occurrences of the k -mer in the sequence divided by the total number of k -mers of length k . The computed scores are then used to calculate the coding potential score for each sequence using a Random Forest (RF) machine learning algorithm. A 10-Fold cross validation is performed by *FEELnc* in order to compute a coding potential score threshold which maximizes sensitivity and specificity. Similarly, to *CPC* and *CPAT*, *FEELnc* also utilizes transcript length and ORF coverage in its classification algorithm. The output of *FEELnc* classification is the scores for all k -mers specified above (1,2,3,6,9 and 12-mers), ORF coverage, coding potential score, and a binary coding/non-coding value for each input transcript. *FEELnc* is

not as fast as *CPAT* or *CPC2*; we observed that this program could classify ~60,000 transcripts in approximately 30 minutes.

2.2.4 Feature based filtering of coding sequences

Machine learning algorithms alone are not sufficient to identify protein coding and non-coding sequences from the wealth of transcriptomic data available (Housman and Ulitsky 2016). There is the potential to introduce false positives into the analysis which could misclassify a protein coding sequence as non-coding. Partially assembled transcripts would not contain a complete open reading frame and thus can be classified as non-coding due to sequence composition and interrupted ORF (Housman and Ulitsky 2016). Alternatively, small non-coding RNAs such as small nuclear RNAs or miRNAs can be processed from longer non-coding transcripts and can be erroneously identified as *bona fide* long non-coding RNAs instead. The presence of conserved structural motifs in transcripts also indicates strong evidence for coding potential so tools that can predict these domains are also very useful. This section will summarize the sequence based homology search methodology using the Basic Local Alignment Tool (*BLAST*), identify the presence of protein coding secondary structures using the *HMMER* suite and the necessary databases for these two tools for the identification of small non-coding RNAs.

The Basic Local Alignment Search Tool (*BLAST*) is one of the most commonly used software tools and is used to identify homologous sequences using an alignment. Unlike non-coding RNAs, protein coding sequences exhibit a greater degree of sequence similarity between species (Housman and Ulitsky 2016). Significant sequence homology to a known protein coding transcript can be used to remove coding transcripts and retain non-coding sequences. The *BLAST* program has several algorithms: *BLASTN*, *BLASTP*, *BLASTX*, and *TBLASTN*. *BLASTN* is used to perform nucleotide-nucleotide alignments, whereas *BLASTP* is used to perform protein-protein alignments (Altschul et al. 1997). *BLAST* can also perform translated nucleotide queries against a protein database (*BLASTX*) or protein-protein alignments of a nucleotide query and nucleotide database

by translating ORFs using TBLASTN (Altschul et al. 1997). The *BLAST* program performs local alignments where it identifies the highest scoring alignment pair between query and reference sequence. The algorithm first breaks a larger query sequence into *k*-mers of a defined length which are referred to as “words” (Altschul et al. 1997). The default word size for a *BLASTN* search is 28 nucleotides whereas the default word size for *BLASTP* is 6 amino acids. The reference database is scanned to search for instances of each word and scored by quality of alignment. Alignments at high-scoring words are extended upstream and downstream in order to generate a more complete alignment between query and reference (Altschul et al. 1997). During extension, the alignment is scored by *BLAST* with mismatches and gaps detracting from the overall quality score. The extension occurs until a complete alignment is no longer possible or if the overall quality score drops below the threshold. *BLAST* calculates an “expect value” (E value) for each alignment which is the measure of the probability that an alignment between query and reference sequence can occur by chance with a lower E value indicating a higher quality alignment (Altschul et al. 1997). It is common to implement an E value significance threshold in order to retain only significant *BLAST* hits (Housman and Ulitsky 2016).

Protein coding genes often contain secondary structures, known as domains, which are evolutionarily conserved. Common proteins include domains for *Actin*, acetyltransferase enzymes, and tumor suppressors like p53. A statistical model can be constructed using multiple sequence alignments and built into a reference database (Finn et al. 2016). The *HMMER3* tool was constructed to search translated protein sequences for the presence of these conserved protein coding domains (Johnson, Eddy, and Portugaly 2010). *HMMER3* performs sequence alignments to a reference database to compare Hidden Markov Models which contain information on positional conservation of an amino acid or nucleotide base. *HMMER* first calculates a probability score for the null hypothesis which is the probability of the sequence being non-homologous and is used to compare the null model to *HMMER3*’s experimental model (Johnson,

Eddy, and Portugaly 2010; Eddy 2011). HMMER3 performs alignments of query sequences to the reference Markov models and selects the alignment with the best-scoring, ungapped alignment for each query which is referred to as the Multiple Segment Viterbi (MSV) model. A bias filter is performed comparing the MSV model to the null model and rescores the MSV model probability scores. The bias filtering is performed to remove poor MSV alignments. If the bias filtered score is lower than a significance threshold of 0.02, it is retained (Eddy 2011). The bias filtered query sequence is aligned to the best scoring reference model and an expect value, which is similar to the BLAST expect value, is calculated as a measure of quality of the alignment with lower expect values correlating to higher confidence domain identification by *HMMER3*. lncRNAs would not contain these protein coding domains or would exhibit poor quality alignments. It would, therefore, be strong evidence of the lack of protein coding potential if a transcript does not contain a protein coding domain.

Both BLAST and HMMER3 require a reference database for their respective analysis which need to be representative of evolutionarily conserved proteins. Reference databases, like Uniref90 and Pfam, can be aligned to putative long non-coding RNAs in order to evaluate evidence of coding potential as these are a database of conserved proteins. The UniProt database is a repository containing sequences and annotation of protein sequences (Apweiler 2004). A high confidence subset is created from the total UniProt repository by clustering sequences based on similarity referred to as the UniRef90 database (Apweiler 2004). The UniRef90 database is composed of all those protein sequences that share a sequence similarity to its longest sequence (Apweiler 2004). The UniRef90 database is updated periodically with the latest release being in September of 2017 and contains ~ 65,000,000 sequences. The UniProt database is also continuously updated and contains reference data derived from proteomics datasets.

The Pfam database is also a repository of evolutionarily conserved sequences but it contains probabilistic models of protein coding genes (Finn et al. 2016). Pfam entries are Hidden Markov

Models which are generated from multiple sequence alignments of protein coding sequences across many eukaryotic and bacterial species. The database contains the functional domains of the evolutionarily conserved protein coding genes. As Pfam is composed of evolutionarily conserved functional domains, alignment to the database provides evidence for coding potential (Finn et al. 2016). One of the steps in deriving a consensus set of non-coding RNA annotation was performing a coding domain search with Pfam and *HMMER* (Derrien et al. 2012).

Small non-coding RNAs can originate from longer transcripts that are spliced into shorter, functional RNA molecules that in turn carry out regulatory functions (Dinger et al. 2008). Pre-processed small ncRNAs differ from lncRNAs in that they are merely intermediate molecules that are spliced into their functional units. Many small non-coding RNAs exhibit sequence conservation between species and this property can be utilized to remove erroneously annotation with high confidence reference databases like MiRBase or Rfam (Kozomara and Griffiths-Jones 2014; Kalvari et al. 2017). The MiRBase version 22 database contains 38,589 hairpin sequences from 465 species that is constantly updated from data derived from validated sequencing experiments (Kozomara and Griffiths-Jones 2014). The Rfam database, by contrast, contains short sequences derived from conserved region found in multiple sequence alignments. Small nuclear, small nucleolar and tRNAs contain regions of enriched conservation which can be used to identify conserved small non-coding RNAs within novel species (Kalvari et al. 2017). The Rfam database also contains 217 motifs derived from evolutionarily conserved lncRNAs including *HOTAIRM1*, *MALAT1*, and *NEAT1*.

2.2.5 Classification of lncRNAs with respect to genome localization

As discussed in **Chapter 1**, lncRNAs are classified into genic and intergenic groups on localization to the genome with respect to protein coding genes. Genome localization may be used to infer potential regulatory targets of lncRNAs (R. Li, Zhu, and Luo 2016). The *BEDtools* software suite is an extremely versatile set of tools that can be used to quantify and manipulate coordinates of

features found on the genome such as genes, transcripts, and exons. (Quinlan and Hall 2010). The *BEDtools* suite can also be used to compare features between strands as well as identify nearby genes which is valuable for classifying lncRNAs. Alternatively, the final stage of the *FEELnc* coding potential classifier tool is identifying genes which can also be very useful in downstream analyses (Wucher et al. 2016).

2.3 Materials and Methods

In the following section, we describe the construction of a computational pipeline to detect long non-coding RNAs using the software described in **Sections 2.1**. We first describe the CHO cell culture methodology (**Section 2.2.1**) and the next generation sequencing library construction process (**Section 2.2.3**). In this section, we present our lncRNA annotation pipeline that can identify lncRNAs from raw RNA-Seq data as well as perform differential expression analysis.

2.3.1 CHO Cell culture and temperature shift

Eight CHO cultures producing a recombinant monoclonal antibody were seeded in SFM II media (Thermo Fisher Scientific) at an initial density of 2×10^5 cells/mL. Cell density was quantified using the Guava EasyCyte flow cytometer (Merck Millipore, 4000-040) at 24 hours, 48 hours, and 72 hours after inoculation (Figure 9). All cells were cultured for 72 hours. Four cultures were temperature-shifted from 37°C to 31°C after 48 hours. Total RNA was extracted at 72 hours from all culture conditions (**Figure 2.2**).

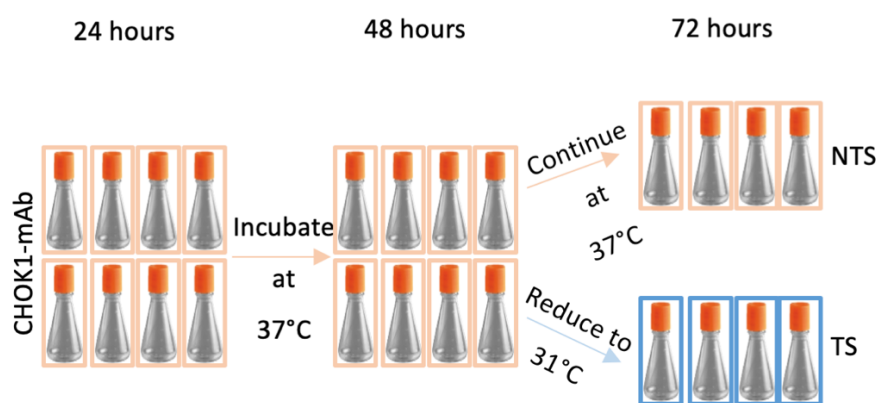


Figure 2.2 A figure summarizing cultivation conditions of 8 replicate cultures. Cells were initially seeded and grown at 37°C for a period of 48 hours. Four replicate cultures were subjected to a reduction in culture temperature to 31°C while the remaining four replicates remained at 37°C. RNA was extracted 72 hours after inoculation.

2.3.2 Total RNA library preparation and sequencing

Total RNA was extracted from each culture using the TriZol RNA isolation protocol (Invitrogen, Int). 1 µg of extracted total RNA was rRNA depleted using the Illumina RiboZero Gold Human/Mouse/Rat rRNA Removal Kit (Illumina). Stranded RNA-Seq libraries were prepared using the Illumina TruSeq Total RNA Stranded library preparation protocol (Illumina). Final

libraries were evaluated for quality using the Agilent High Sensitivity DNA assay kit and Agilent 2100 Bioanalyzer (Agilent, 5067-4626). 2 x 150 bp, paired end sequencing on a HiSeq 2500 sequencer was performed on the final libraries. Libraries were sequenced at a minimum of 50 million fragments per sample to ensure adequate profiling of transcripts with low abundance.

2.3.3 lncRNA discovery pipeline

We constructed a software pipeline that can analyze RNA-Seq data from initial data pre-processing to lncRNA identification and differential expression (**Figure 2.3**). The discovery pipeline was written with a combination of shell and PERL languages and all statistical analysis was performed using the R framework. **Section 2.2.3** will summarize each stage of the analysis pipeline and discuss relevant parameters used at each step.

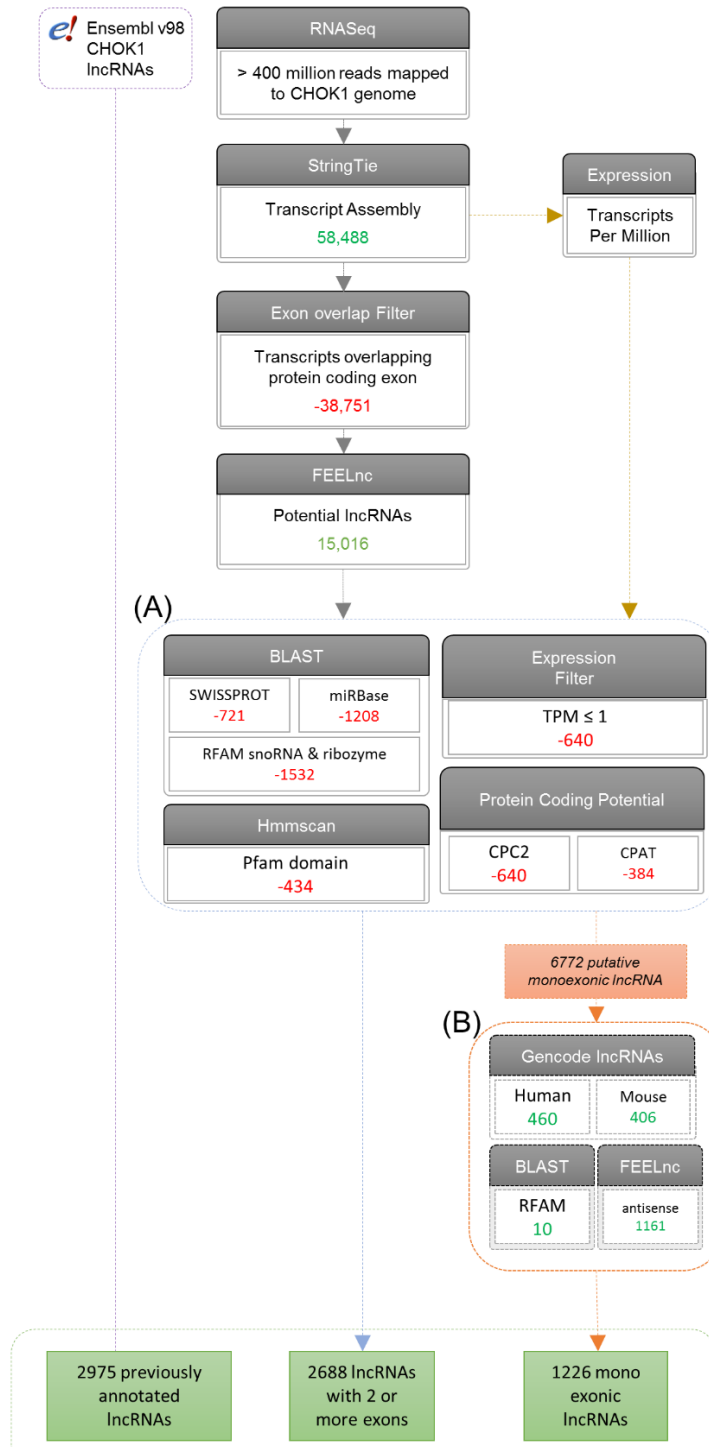


Figure 2.3 A flowchart diagram detailing analysis steps of our lncRNA annotation pipeline from raw data QC to lncRNA classification. Reads are initially aligned to the ENSEMBL version 98 criGri1 reference genome and assembled into 58,488 transcripts. **(A)** Transcripts with coding domains and homology to small ncRNAs and high coding potential are removed. **(B)** 6,722 putative lncRNAs are compared to RFAM and GENCODE human and mouse annotations in order to find conserved lncRNAs.

2.3.3.1 lncRNA discovery pipeline

Raw read data from all 8 libraries was inspected for low quality bases and adapter contamination using *FASTQC* (Andrews, 2010). The *cutadapt* tool was used to trim adapters found on the 3' end of the forward and reverse read with only paired end reads retained for quality trimming. Adaptor trimmed reads were subjected to second round of quality trimming with *Trimmomatic* (Bolger, Lohse, and Usadel 2014) using a sliding window size of 4bp and a Phred quality threshold of 20 (error probability 0.01). Quality filtered reads were deemed of sufficient quality for alignment after inspection with *FASTQC*.

2.3.3.2 Read alignment and transcriptome assembly

Quality filtered reads were aligned to the version 98 *ENSEMBL criGri1* CHOK1 reference genome (X. Xu et al. 2011) using *STAR* (Dobin et al. 2013) using the stringent two-pass alignment method where reads are initially mapped to the reference genome in order to infer splice junction which will be used to update the genome index to have more accurate alignment of reads. The mapping rate, or the percentage of reads that successfully aligned to the genome, was inspected for each sample to ensure analysis was of sufficient quality. Mapped reads were used to construct transcript models using the *StringTie* assembler (Pertea et al. 2015) using the version 98 annotation of the *ENSEMBL criGri1* genome as a guide. *StringTie* constructed a transcript model if a minimum of 5 reads supported a splice-junction. Transcript assembly was performed on each sample and a consensus assembly was derived using the *StringTie merge* utility. A transcript was classified as a distinct isoform if its abundance was 10% of the most abundant isoform with a minimum of 10 supporting its expression. The quality of the assembly was assessed using the *gffcompare* tool which evaluates sensitivity and specificity of assembled gene and transcript loci to a reference by comparing discovered exon coordinates on a genome to the reference (Trapnell et al. 2010). *StringTie* was used to quantify transcript and gene abundances which was used to remove low expressed genes from lncRNA identification analysis. A putative lncRNA gene had to have at least one isoform with a TPM > 1 in at least one sample to be retained.

2.3.3.3 lncRNA discovery and annotation

The high-confidence assembly was initially filtered to remove annotated protein coding transcripts and known small ncRNAs such as tRNAs, miRNAs, and snRNAs. Previously annotated lncRNAs that were identified by *ENSEMBL* were retained for downstream analysis. A second round of filtering was performed using the *BEDtools* intersect tool (Quinlan and Hall 2010); a transcript was removed if a component exon overlapped the exon of a coding gene on the same strand by a single base pair or more. The putative lncRNAs were evaluated for coding potential by *FEELnc*, *CPAT*, and *CPC2*. *FEELnc* was used to initially remove transcripts < 200bp and intergenic mono-exonic transcripts in order to remove potentially novel small ncRNAs and a Random Forest classification was performed using all annotated mRNAs as a training set (Wucher et al. 2016). The putative lncRNAs identified by *FEELnc* were further filtered by *CPC2* and *CPAT* in order to retain a high-confidence set of identified lncRNAs. *CPAT* was trained on a model of mouse mRNA and lncRNAs provided by the authors with a coding potential probability score > 0.44 suggesting a coding gene (L. Wang et al. 2013). Classification by *CPC2* used a training model provided by the authors which classified input transcripts as coding and non-coding (Kang et al. 2017b). Only transcripts that were classified as non-coding by all three calculators were retained.

2.3.3.4 Removal of transcripts with coding domains and homology to small ncRNAs

The TransDecoder tool (<https://transdecoder.github.io/>) was used to translate the longest strand-specific ORFs. These translated ORFs were compared against the high-confidence UniProt (Release version: May 12, 2015) and Pfam-A (version 32.0) reference databases (Apweiler 2004; Finn et al. 2016). *BLASTP* (Camacho et al. 2009) was utilized to identify and remove transcripts with significant homology (expect value < 1×10^{-5} , amino acid length > 100) to the UniProt database. The *HMMER* suite (Eddy 2011) was used to predict and filter out transcripts with protein coding domains (expect value < 1×10^{-5}). Transcripts that did not possess significant homology to UniProt databases and lacked a predicted protein coding domain were aligned to the miRBase (version 22.1, Kozomara & Griffiths-Jones 2014) hairpin database to remove miRNA host genes

using *BLASTN* (expect value $< 1 \times 10^{-5}$). The putative lncRNAs were also compared to a custom subset of the Rfam database (version 14.1) containing sequences for snoRNAs, snRNAs, tRNAs, ribozymes, and rRNAs using *BLASTN* (expect value $< 1 \times 10^{-5}$). A gene was removed from the assembly if one of its component isoforms had significant alignment to a protein coding or small non-coding domain. A final, high-confidence reference annotation containing all identified lncRNAs was used for synteny analysis and differential expression⁴.

2.3.3.5 Syntenic lncRNA sequence identification

A cross-species alignment between human, mouse, and Chinese hamster genomes was used to identify syntenic long non-coding loci. The University of California – Santa Cruz Genome Browser (UCSC Genome Browser) provides a repository containing cross-species alignments as well as the liftOver utility that can be used to convert coordinates from one species or genome to another (Karolchik et al. 2003; Meyer et al. 2013). A “chain file” which contains corresponding conserved regions in a multiple genome alignment is used to convert coordinates with the liftOver utility. The UCSC Genome Browser provides a cross-species alignment between the GRCh38 build of the human genome and the *criGri1* genome as well as the GRCm38 version of the mouse genome and the *criGri1* genome. Release 32 of the human and release 23 of the mouse long non-coding GENCODE annotation was converted to *criGri1* assembly coordinates using the liftOver utility. A minimum of 10% of all nucleotides of a GENCODE lncRNA transcript had to re-map to the *criGri1* genome to be retained. The BEDtools suite was used to intersect the lncRNA assembly with the GENCODE converted coordinates.

2.3.4 Differential expression of lncRNAs

A count based gene level differential expression analysis was performed using the *DESeq2* software package within the R framework using raw counts generated by *HTSeq-Count*. All

⁴ Synteny is defined as conserved gene structures between species. For example, a lncRNA gene composed of three exons that is found in humans is considered syntenic in CHO cells if a non-coding gene is similarly found in the same location in the CHO genome.

assembled genes (protein coding, long non-coding and small non-coding) were analyzed for differential expression. Count data was normalized using *DESeq2* and evaluated for differential expression. A gene was considered differentially expressed if it had a fold-change $> \pm 1.5$ and a Benjamini-Hochberg adjusted p. value < 0.05 . A gene had to have an average normalized abundance greater than 100 reads across all samples which was used to filter out low abundance genes. Significantly expressed predicted long non-coding RNAs were classified into intergenic, antisense, and sense loci as specified in **Chapter 1**.

Biological pathway enrichment of differentially expressed protein coding genes was performed using DAVID (D. W. Huang, Sherman, and Lempicki 2009). Differentially expressed antisense non-coding RNAs were compared with their sense protein coding partners to identify sense/antisense co-expression. These loci were identified as putatively co-regulated, where both sense and antisense genes are similarly up or downregulated, or anti-correlated, where the antisense gene had an opposite-fold-change than its protein coding partner.

2.3.5 Metabolite profiling between NTS and TS samples

A Nova Flex instrument (Nova Biomedical, Inc) was used to measure lactate, ammonia, and glutamine concentration in extracellular media between non-temperature-shifted and temperature-shifted samples 72 hours after initial inoculation. A two-tailed t-test was used to evaluate the average difference in extracellular metabolite concentration. Metabolites were considered significantly different between the two culture conditions if the t-test P-value was < 0.05 .

2.4 Results

In order to improve the annotation of the CHO cell transcriptome, we have generated a high-resolution RNA-Seq data that profiles the change in gene expression of cells undergoing a reduction in culture temperature which is commonly performed in industrial bioprocessing. The data is of high coverage (average of > 50 million read pairs / sample) and strand-specific which is a necessary configuration for profiling lncRNAs (**Section 1.4**). We developed a software pipeline that detects novel non-coding loci and provides further evidence for predicted loci. Our software pipeline produces a Gene Transfer Format (GTF) tab-delimited file that contains the genomic localization of transcripts and exons for each locus within the *CHOK1* genome. It leverages reference guided methods to profile the landscape of long non-coding RNAs in the Chinese hamster transcriptome in order to improve our understanding of the long non-coding transcriptome. In this section, we summarize the results of our assembly (**Section 2.3.1-2.3.3**), describe differences in transcript features between protein coding and long non-coding RNAs (**Section 2.3.4**), present evolutionarily conserved orthologs between human, mouse and Chinese hamster lncRNAs (**Section 2.3.5**), and demonstrate, for the first time correlated and anti-correlated expression of sense/antisense pairs of mRNAs and lncRNAs (**2.3.6**).

2.4.1 Temperature shift induces growth arrest in CHO cells

Eight samples were cultured over a period of 72 hours in SFM II media and viable cell density was measured every 24 hours using the Guava EasyCyte flow cytometer. Four cultures experienced a reduction in culture temperature from 37°C to 31°C 48 hours after inoculation. Average cell densities between replicates were observed to be concordant until 48 hours after initial inoculation but were observed to diverge after 72 hours (**Figure 2.4**). Growth arrest was observed in CHO cells cultivated at 31°C with a 24% reduced cell density compared to cells cultivated at 37°C after 72 hours.

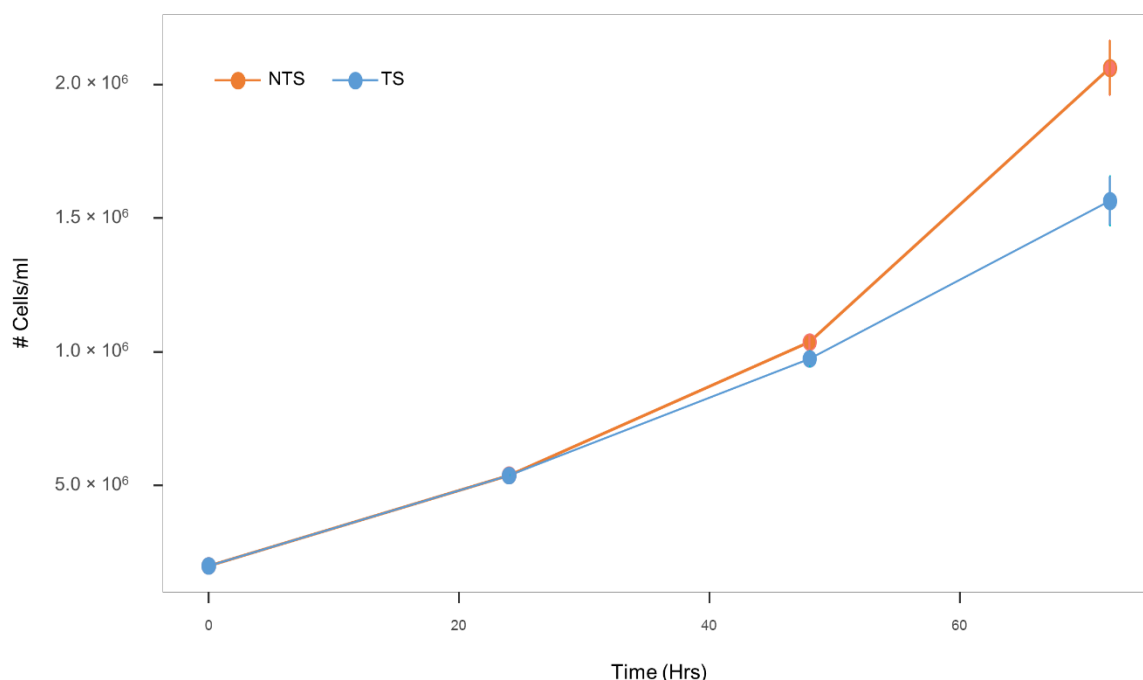


Figure 2.4 A plot of average cell density over time. Cells were sampled over 24 hour intervals. As expected, cells that underwent a reduction in culture temperature had a lower average cell density as they proliferated at a lower rate when compared to those cells cultured at 37°C. Total RNA was harvested from all samples after 72 hours.

2.4.2 Pre-processing of RNA-Seq data and alignment to the CHOK1 genome

Illumina strand-specific libraries were prepared using RNA extracted from a total of 8 CHO cultures, four cultured at 37°C and four cultured at 31°C, and sequenced on an Illumina HiSeq 2500. High coverage RNA-Seq data was obtained with an average of approximately 69 million read pairs per sample. The majority of samples (7 out of 8) were of high coverage and contained > 65 million read pairs whereas only one sample (31°C Replicate 2) had approximately 50 million read pairs (**Table 2.1**). The *cutadapt* and *Trimmomatic* tools were used to trim Illumina adapter sequences and low-quality bases (Phred quality scores < 20). An average of ~8 5% of reads were reads were retained after quality filtering. Quality filtered reads were mapped to the *C. griseus* genome using the splice-aware aligner *STAR* with an average of 85% read mapping rate of all samples (**Table 2.1**).

Table 2.1. A table of raw and quality-filtered read pairs with their mapping quality to the ENSEMBL criGri1 genome. Approximately 15% of reads were removed due to insufficient quality. Quality filtered reads were mapped to the criGri1 genome and an average mapping rate of 85% was observed across all samples. Libraries sequenced from cells cultured at 31°C were observed to have a slightly higher mapping rate.

Sample	Raw Read Pairs	Quality Filtered Read Pairs	% Aligned
NTS 1	70,447,918	60,845,892	84.95
NTS 2	76,600,082	64,504,154	86.33
NTS 3	69,654,137	58,686,630	84.59
NTS 4	70,644,823	60,743,068	84.39
TS 1	67,864,940	58,388,149	87.00
TS 2	50,979,659	42,506,156	86.87
TS 3	73,203,827	62,854,615	87.67
TS 4	77,979,243	66,647,367	87.48

2.4.3 Reference guided assembly metrics

We assembled 37,104 genes (58,486 transcripts) using the *ENSEMBL criGri1* genome as a reference with 11,313 (30.4%) genes identified as novel as they lacked a corresponding annotation within the *ENSEMBL* transcriptome (X. Xu et al. 2011). The *gffcompare* tool (Trapnell et al. 2010) was used to calculate the quality of the assembly and it was found to have a very high sensitivity (100.0%) but lower precision (62.8%). We identified 5,351 lncRNA genes (6,488 transcripts) within our assembly which accounted for 14.4% of the total assembly. Protein coding genes were significantly more represented in our assembly with 51.6% of all genes identified as coding. miRNAs, tRNAs, rRNAs, and miscellaneous mono-exonic transcripts accounted for the remaining 34% of the assembly (**Figure 2.5**).

2.4.4 Assembly metrics of long non-coding RNA

We assembled 5,351 long non-coding RNA genes with 2,862 novel (53% of all lncRNAs) loci identified as unannotated as they lacked a corresponding gene within the *ENSEMBL criGri1* version 96 annotation. Our assembly doubled the number of annotated lncRNAs compared to the *ENSEMBL* reference. The *BEDtools* suite was used to classify lncRNA genes based on their localization to protein coding genes into sense, antisense, and intergenic classes. **Figure 2.5** summarizes the proportion of each class of lncRNA and compares the updated CHO assembly to the GENCODE human and mouse assemblies. Intergenic lncRNAs are the most abundant type

of lncRNA class in all three genomes but are significantly more represented in the Chinese hamster assembly making up 75% of all identified loci. The proportion of genes in each class are approximately concordant between human and mouse lncRNAs, however. A significantly fewer number of sense lncRNAs were identified within the *StringTie* assembly.

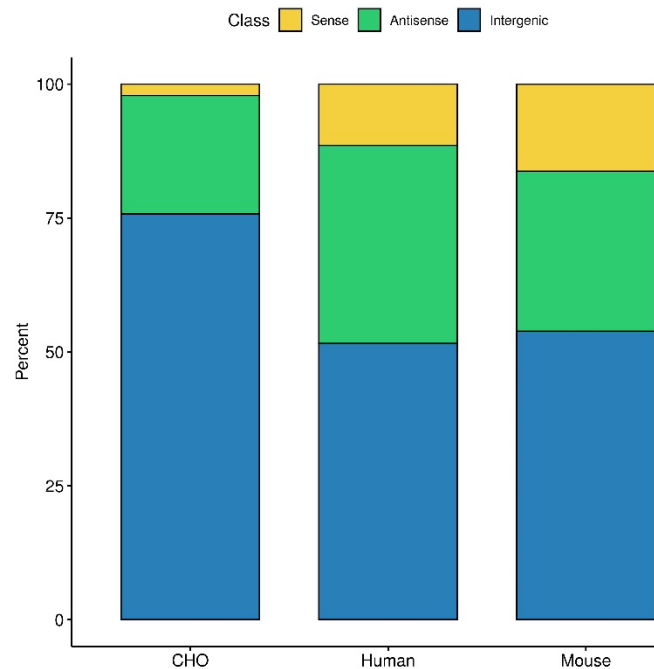


Figure 2.5 A bar plot comparing the percentage of all lncRNA genes found in sense, antisense, and intergenic classes. The human and mouse lncRNA annotations were derived from release 32 and 23 for human and mouse respectively. Intergenic lncRNAs make up the largest classes in all three species but are the dominant class in CHO accounting for ~75%. The percentage of each class of human and mouse lncRNAs are approximately concordant between the two species.

Sense, antisense, and intergenic lncRNAs were further classified with respect to overlap of protein coding genes (**Figure 2.6**). Intergenic lncRNAs that were located upstream or downstream of a coding gene were segregated by strand. A gene was considered intergenic if it there was no coding gene within 10kb. Antisense genes that were found within 2kb upstream of a protein coding gene were classified as divergent. The two most abundant classes are intergenic lncRNAs and those genes that overlap a coding gene on the antisense strand.

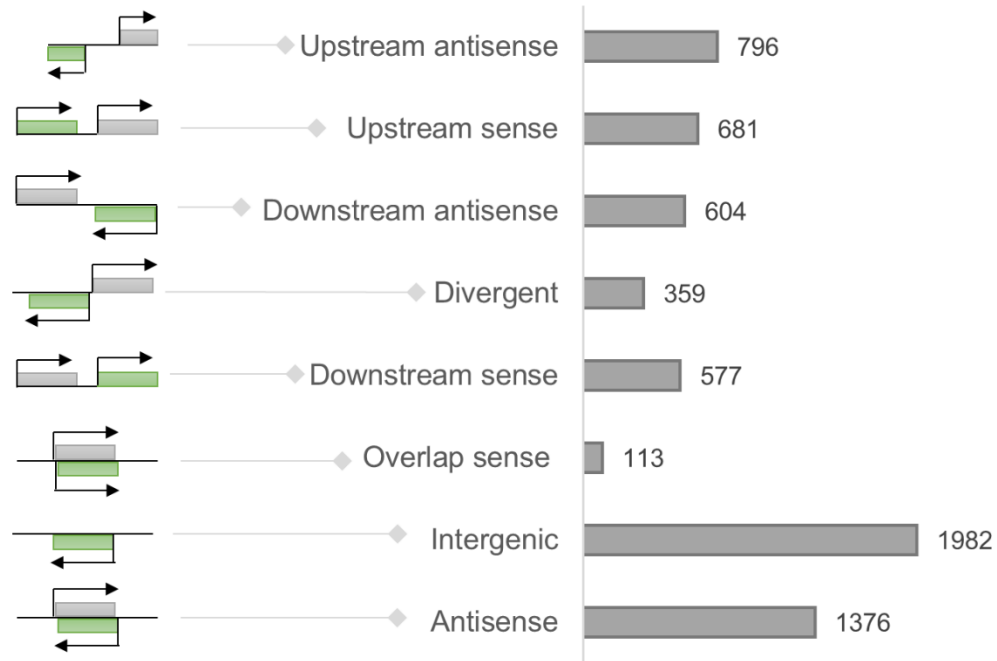


Figure 2.6 Classification of lncRNAs into respective classes with respect to localization to a protein coding gene. Intergenic and antisense genes were the most abundant classes of lncRNAs.

2.4.5 Comparison of characteristics between mRNA and lncRNA loci

lncRNAs are significantly shorter than their mRNA counterparts with a median length of 649 bp compared to a median length of 1,931 bp for protein coding genes (**Figure 2.7**). A Kolmogorov-Smirnov test was used to compare the difference between the distributions of transcript lengths between mRNA and lncRNA. The shorter transcript length distribution found in lncRNAs was found to be significantly different from mRNA transcripts (KS p-value $< 2.2 \times 10^{-16}$).

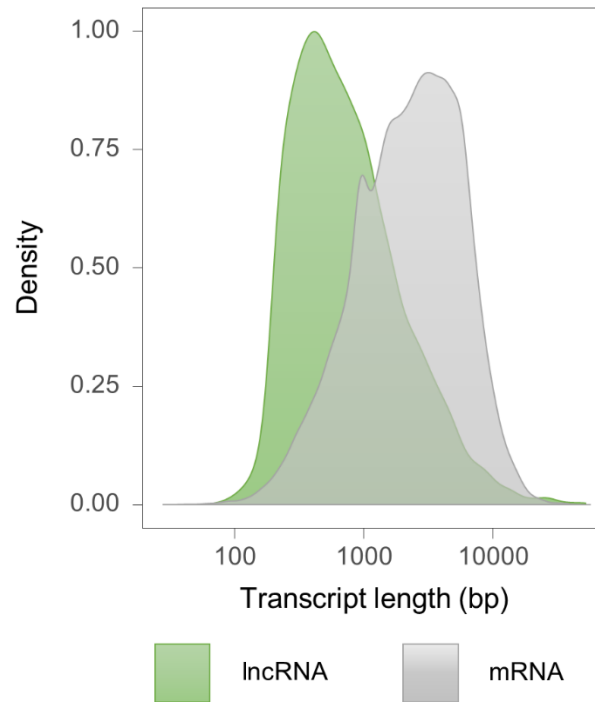


Figure 2.7 A density plot comparing transcript lengths between coding and long non-coding RNAs. The distribution of lncRNA transcript lengths was found to be significantly shorter with a median transcript length of 649bp compared to a median mRNA transcript length of 1,931bp which is 3X longer.

lncRNAs were also observed to have a slightly lower median GC content compared to mRNAs with a median GC content of 45.35 compared to a median of 45.5% for coding transcripts (**Figure 2.8**). The distribution of lncRNA GC content was also observed to be slightly narrower with a sharper peak near the median (KS p-value < 2.2×10^{-16})

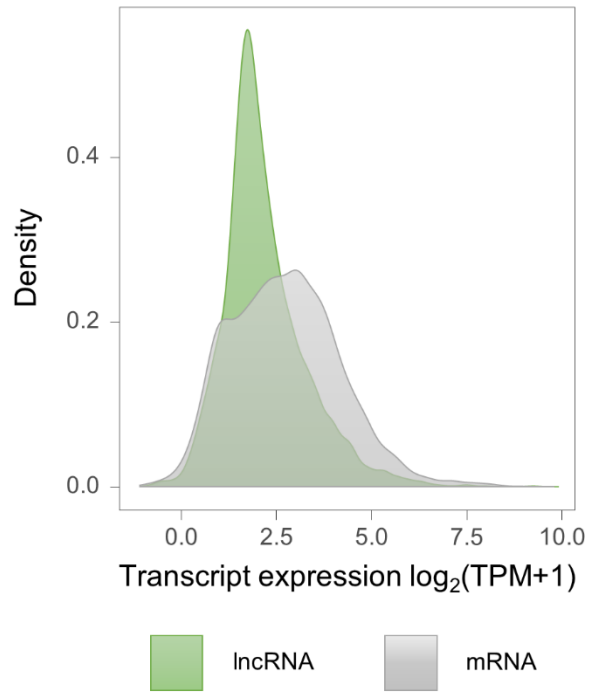


Figure 2.8 A density plot summarizing the GC content distribution of coding and long non-coding transcripts. mRNA GC content was observed to be slightly higher with a median of 49.5%.

Next, we compared the distribution of abundances between protein coding and long non-coding transcripts. Abundances were estimated for each transcript by *StringTie* across all samples and replicates. Only expressed transcripts with a minimum TPM of 1 in at least one sample were retained for equitable comparisons (**Figure 2.9**). Coding transcripts were found to be expressed at higher abundances than lncRNA transcripts. The evolutionarily conserved lncRNAs *Malat1* and *Neat1* were observed to be two of the most abundant transcripts and were expressed at two orders of magnitude higher than the average protein coding gene, however.

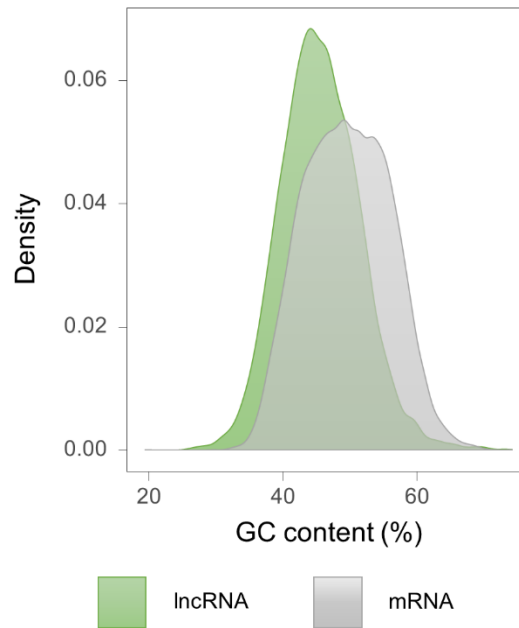


Figure 2.9 A density plot comparing the distribution of normalized TPM abundances between lncRNA and mRNA transcripts. Coding transcripts had a wider distribution of expression and were found to be more abundant than lncRNAs.

2.4.6 Isoform and exon distributions between lncRNA and mRNA genes

Single isoforms dominated both mRNA and lncRNA assemblies with approximately 50% of all coding genes being composed of a single isoform (**Figure 2.10**). The majority of all lncRNA genes (~85%) were observed to be genes containing a single isoform. Both mRNA and lncRNA genes were observed to have a low isoform count per gene with 98% of all non-coding genes and 88% of all coding genes having 3 or fewer isoforms.

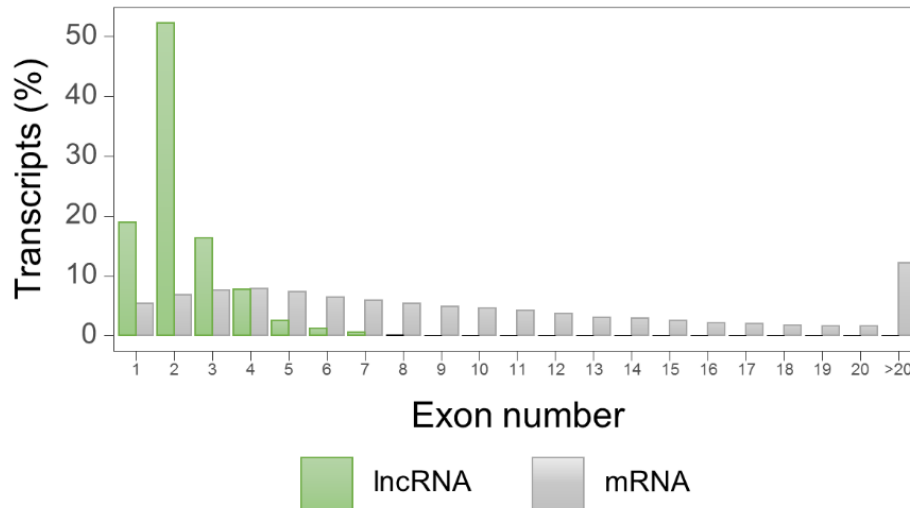


Figure 2.10 A bar plot summarizing the percentage of total lncRNA and mRNA genes. Single isoform genes dominated both lncRNA and coding genes with the majority of all lncRNAs composed of genes with a single transcript.

lncRNAs were observed to have a fewer number of exons per transcript compared to their mRNA counterparts. Over 50% of all lncRNA transcripts were found to be bi-exonic (transcripts containing only two exons) with a fewer proportion of mono-exonic transcripts (transcripts composed of only one exon). These mono-exonic transcripts were found predominantly antisense to protein coding genes. 99% of all lncRNA transcripts had fewer than 6 exons whereas the distribution of exons per transcript for protein coding genes was found to be equivalently divided for each exon number (**Figure 2.11**) with transcripts containing up to 21 exons being the largest group.

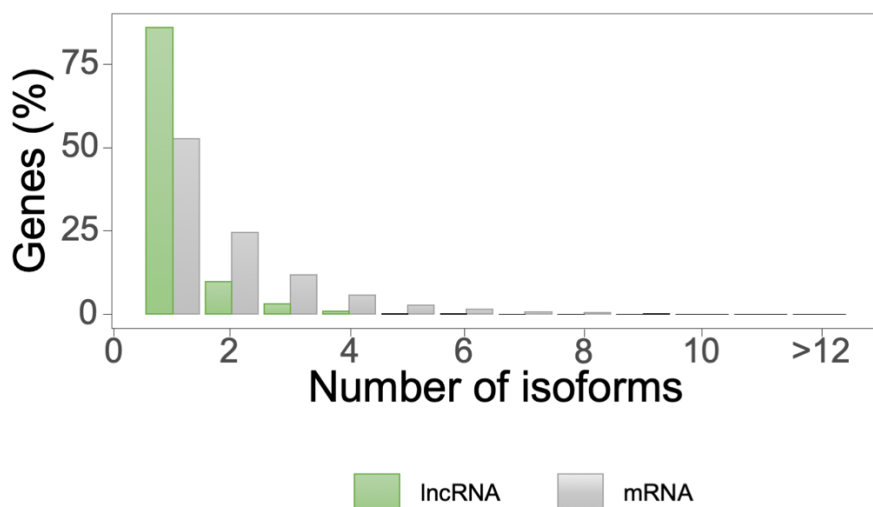


Figure 2.11 A bar plot summarizing the distribution of the number of exons for coding and long non-coding transcripts. The majority of lncRNA transcripts had 3 or fewer exons while the coding transcripts had an even distribution of exons in each transcript.

2.4.7 Identification of syntenic lncRNAs

Syntenic lncRNAs were identified using a multiple sequence alignment and the liftOver utility provided by UCSC genome browser (Karolchik et al. 2003). Releases 32 and 23 of the human and mouse GENCODE lncRNA annotation database were converted to *criGri1* coordinates and an overlap was computed between assembled CHOK1 lncRNAs. A total of 1,041 lncRNA loci were found to be conserved between human and CHO while a total of 1,035 lncRNA loci were conserved between mouse and CHO (**Figure 2.12**). Evolutionarily conserved lncRNAs among this cohort were *NEAT1*, *HOTAIRM1*, *PVT11*, and *LINC00511* which have previously been functionally validated *in vitro* (R. Li, Zhu, and Luo 2016; Vance and Ponting 2014). The version 98 reference annotation of the *ENSEMBL criGri1* genome has a total of 769 previously annotated, evolutionarily conserved lncRNAs derived from similarity to the Rfam database. Our assembly expanded this repertoire to over 1,000 genes and extended the length of *ENSEMBL* annotated transcript 15-fold including deriving the full-length sequence for *MALAT1* and *NEAT1*.

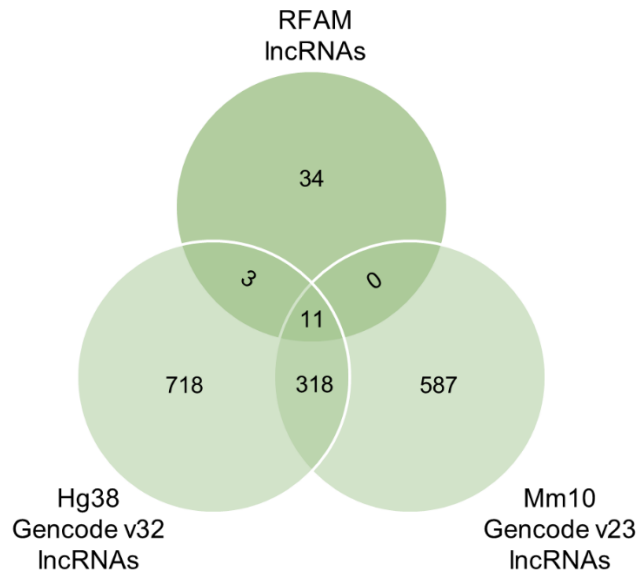


Figure 2.12 Venn diagram of conserved lncRNAs found homologous to BLAST searches within Rfam and cross-species syntenic loci.

2.4.8 Differential gene expression after a reduction in culture temperature

Aligned reads were quantified using the *HTSeq-Count* utility and differential expression was evaluated using the *DESeq2* software package within the R statistical environment. Principal component analysis of the normalized read counts found that samples can be clustered into two distinct groups that correspond to cultivation temperature (**Figure 2.13**).

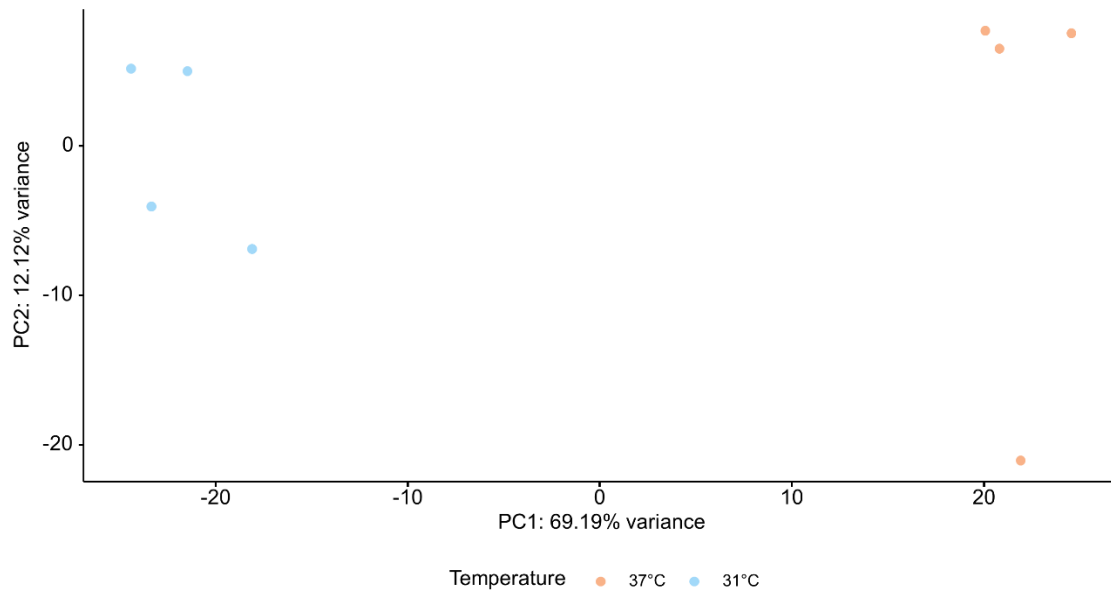


Figure 2.13 A PCA plot summarizing the variance between samples and replicates. 68.19% of the variance could be captured by the first principal component which corresponds to culture temperature.

A gene was considered significantly differentially expressed if it possessed a-fold-change $\geq \pm 1.5$ and a Benjamini-Hochberg adjusted p-value < 0.05 . A gene was retained if it had an average normalized read abundance > 100 in both conditions. This was performed to remove low expression genes that are not transcribed at biologically relevant abundances. We observed global changes in gene transcription between cells cultivated at 37°C and 31°C (**Figure 2.14**) with a total of 1,317 coding genes and 400 lncRNA genes.

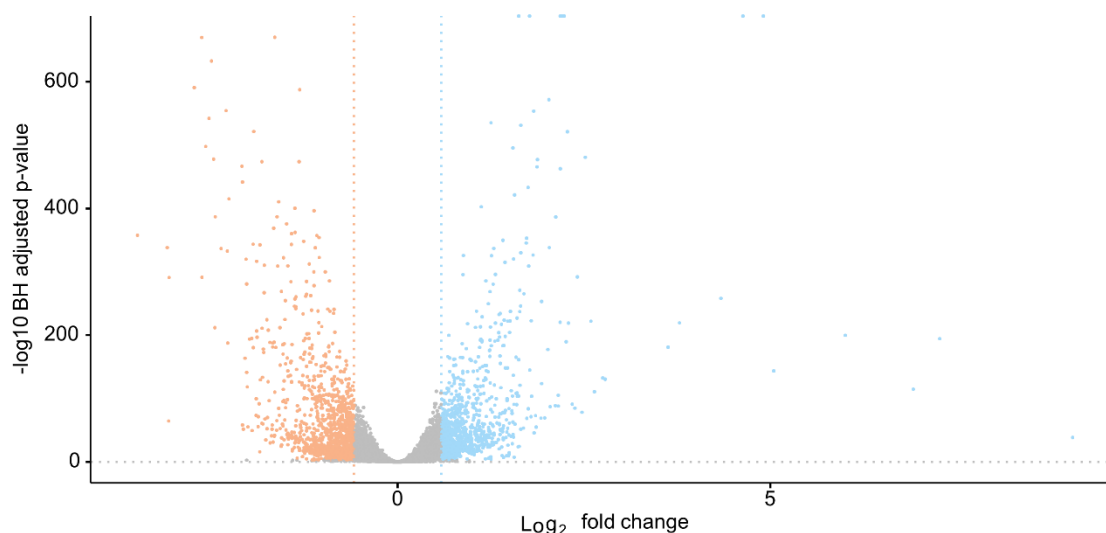


Figure 2.14 A volcano plot summarizing significant-fold-change between cells cultivated at 37°C and 31°C. A gene was considered significantly differentially expressed if it had a fold-change $\geq \pm 1.5$ and a BH adjusted P-value < 0.05 . An even distribution of genes were found to be upregulated (blue) and downregulated (red) after a reduction in culture temperature.

All significantly differentially expressed coding genes are found in **Supplementary Table 2.1**. We observed significant upregulation of *CIRBP* and *RBM3* which are two well-characterized coding genes that have been previously demonstrated to be differentially expressed after temperature shift in CHO cells (X. Zhu, Buhrer, and Wellmann 2016). *CIRBP* was found to be 2.2 fold upregulated (BH adjusted P-value: 2.80×10^{-100} , **Figure 2.14**) while *RBM3* was found to be 3.11 fold upregulated (BH adjusted P-value 2.01×10^{-119}). The *FOS* transcription factor was one of the most significantly differentially expressed and was observed to be very highly downregulated with a fold-change of -11 (BH adjusted P-value 4.36×10^{-156}). *Lactate dehydrogenase C (LDHC)* was one of the most significantly upregulated genes with a fold-change of 30 and very minuscule Benjamin-Hochberg adjusted P. value.

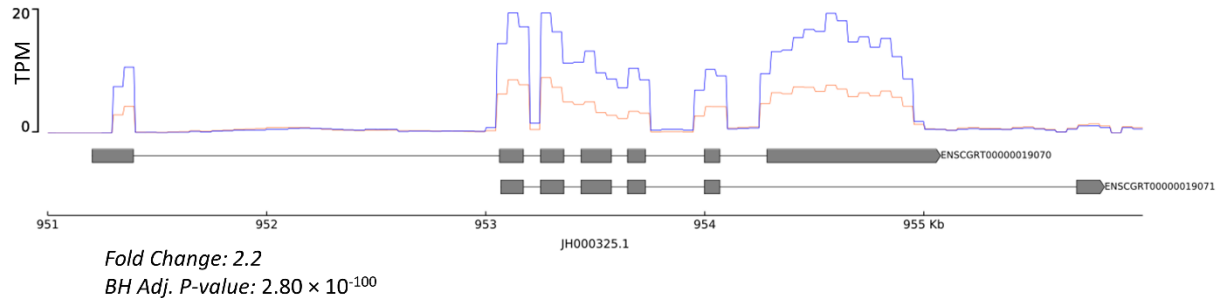


Figure 2.15 A coverage plot comparing average TPM normalized coverage between NTS (red) and TS (blue). A 2.2 fold increase in expression of the CIRBP gene was observed at 31°C.

We segregated differentially expressed lncRNAs into genic and intergenic classes with respect to localization on the genome as described in **Chapter 1**. The results of differential expression analysis identified 400 lncRNA genes that exhibited a change in expression after a reduction in culture temperature (**Figure 2.15**). Intergenic lncRNAs composed the largest group of differentially expressed non-coding genes with approximately 75% of all DE genes classified as intergenic. Differentially expressed antisense RNAs comprised 24% of the remained with the last 1% classified as sense overlapping genes. The distribution of differentially expressed lncRNA genes closely equates with the total distribution of lncRNAs summarized in **Figure 2.16**.

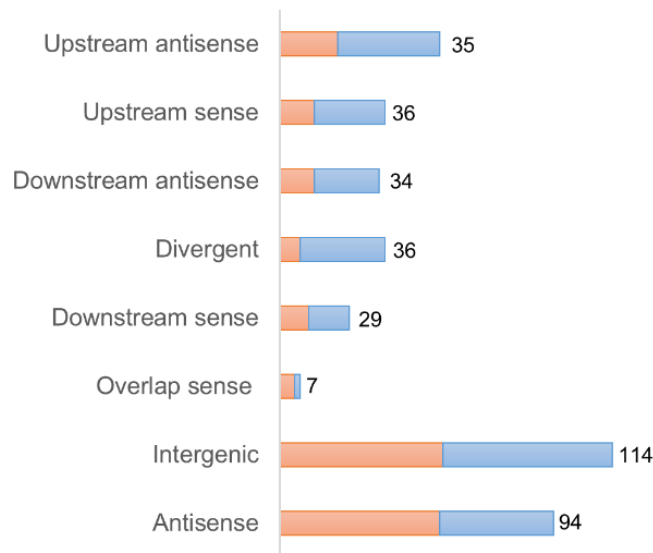


Figure 2.16 A bar plot of significantly differentially expressed lncRNAs segregated into different classes of lncRNAs.

2.4.9 Enriched functional pathways of differentially expressed protein coding genes

Gene Ontology (GO) enrichment analysis was performed using DAVID on all significantly differentially expressed protein coding genes (D. W. Huang, Sherman, and Lempicki 2009). Gene ontology analysis is used to identify significantly enriched pathways and cellular functions within a list of protein coding and small non-coding genes. A GO term was considered significantly enriched if it possessed a Benjamini-Hochberg adjusted P-value less than 0.05. **Table 2.2** summarizes the top five most enriched GO terms. Genes involved in DNA replication, response to cellular stress, and regulation of cell cycle response were observed to be significantly enriched. The complete list of enriched gene ontologies is found within **Supplementary Table 2.3**. It is currently not possible to perform GO enrichment analysis for lncRNAs as there is not sufficient, verified evidence equating expression with cellular pathways or phenotypes.

Table 2.2 A table summarizing the top 5 most enriched GO terms for protein coding genes. Genes involved in regulation of cell cycle, mitosis, and cellular response to stress were significantly enriched.

GO ID	GO Term	# of Genes	BH Adj. p-value
GO:0007049	Cell cycle	197	6.30×10^{-37}
GO:0006259	DNA metabolic process	137	1.20×10^{-29}
GO:0006974	Cellular response to DNA damage	114	7.80×10^{-28}
GO:0000278	Mitotic cell cycle	124	1.10×10^{-27}
GO:0033554	Cellular response to stress	191	5.30×10^{-27}

2.4.10 Significantly differentially expressed evolutionarily conserved lncRNAs

We identified significantly differentially expressed lncRNAs that are evolutionarily conserved between human, mice, and CHOK1 genomes. **Supplementary Table 2.2** summarizes all differentially expressed conserved lncRNAs. *PVT1* which has been previously validated by Vito & Smales (Vito and Smales 2018), was among this cohort and is significantly upregulated (fold-change 2.27, BH adjusted P-value 2.00×10^{-57}) after temperature shift but the full length gene was not assembled due to break points within the reference genome. Syntenic *LINC00511* was

discovered using cross-species alignment with human v32 GENCODE lncRNAs which was not previously annotated within the *ENSEMBL* v98 reference for the CHO genome. *LINC00511* is an abundant multi-exonic intergenic lncRNA that is upregulated after temperature shift (fold-change: 1.59, BH Adj. p-value: 1.18×10^{-11}). We also identified the conserved *HOTAIRM1* lncRNA which is a divergent antisense gene to the *HOXA1* coding gene and is one of the most significantly upregulated lncRNAs after a reduction in culture temperature (fold-change: 2.08, BH Adj. p-value: 1.73×10^{-43} , **Figure 2.17**). *HOTAIRM1* is annotated as two separate genes within the *ENSEMBL* reference that corresponds to the first and second exon of the conserved *HOTAIRM1* locus. *StringTie* assembly extended the *HOTAIRM1* gene sequence to integrate both exons into one transcript.

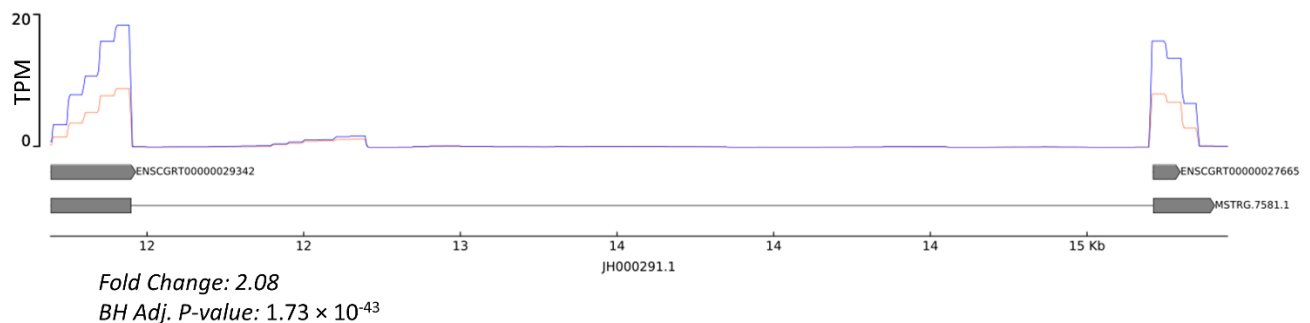


Figure 2.17 A coverage plot comparing average TPM normalized coverage between NTS (red) and TS (blue) of the *HOTAIRM1* gene. A 2.08 fold increase in expression of the *HOTAIRM1* gene was observed at 31°C

2.4.11 Genome-wide differential expression of antisense transcription

219 lncRNA genes are differentially transcribed on the antisense strand, either directly within the gene body or immediately upstream or downstream of coding gene. Antisense genes that are found overlapping protein coding genes constitute ~43% of this group (94 DE genes) with the remaining genes divided amongst upstream, downstream, and divergent classes (**Supplementary Table 2.2**). Approximately 80% of differentially expressed antisense genes are novel and were previously not present within the *ENSEMBL* v98 reference CHO annotation. Almost all antisense genes exhibit differential expression independent of the sense protein coding

gene. For example, the antisense *AIG1* is the most significantly differentially expressed gene where we observed on/off transcription (fold-change: 12.02, BH Adj. p-value: 1.68×10^{-57}) while there is no observed differential expression of the sense coding gene (**Figure 2.18**).

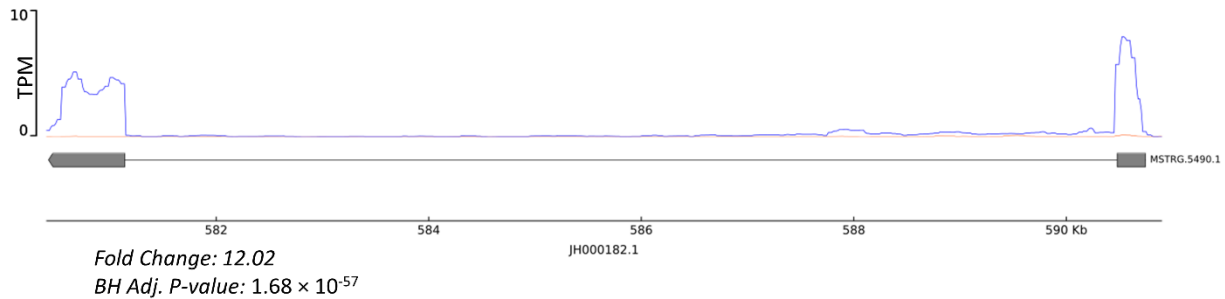


Figure 2.18 A coverage plot comparing average TPM normalized coverage between NTS (red) and TS (blue) of the AIG1-AS gene. A 12.02 fold increase in expression of the AIG1-AS gene was observed at 31°C

2.4.12 Sense/antisense differential expression

Differentially expressed sense genes were compared to differentially expressed antisense genes and we have identified a cohort of 21 mRNA encoding genes that exhibit differential expression on the overlapping antisense strand or within 2kb upstream of the mRNA (**Table 2.3**). Only two of these antisense lncRNAs are expressed above modestly abundant amounts (baseMean > 500). 10 genes were observed to have correlated expression where the sense gene and its antisense non-coding partner are similarly up or downregulated. All divergent antisense genes exhibited correlated expression. One of these loci is the circadian rhythm regulator *Timeless* that contains a mono-exonic transcript on the antisense strand. Both the sense protein coding (fold-change: -2.08, BH Adj. p-value: 1.62×10^{-47}) gene and the antisense non-coding gene (fold-change: -2.86, BH Adj. p-value: 2.15×10^{-29}) are similarly downregulated after temperature shift (**Figure 2.19**).

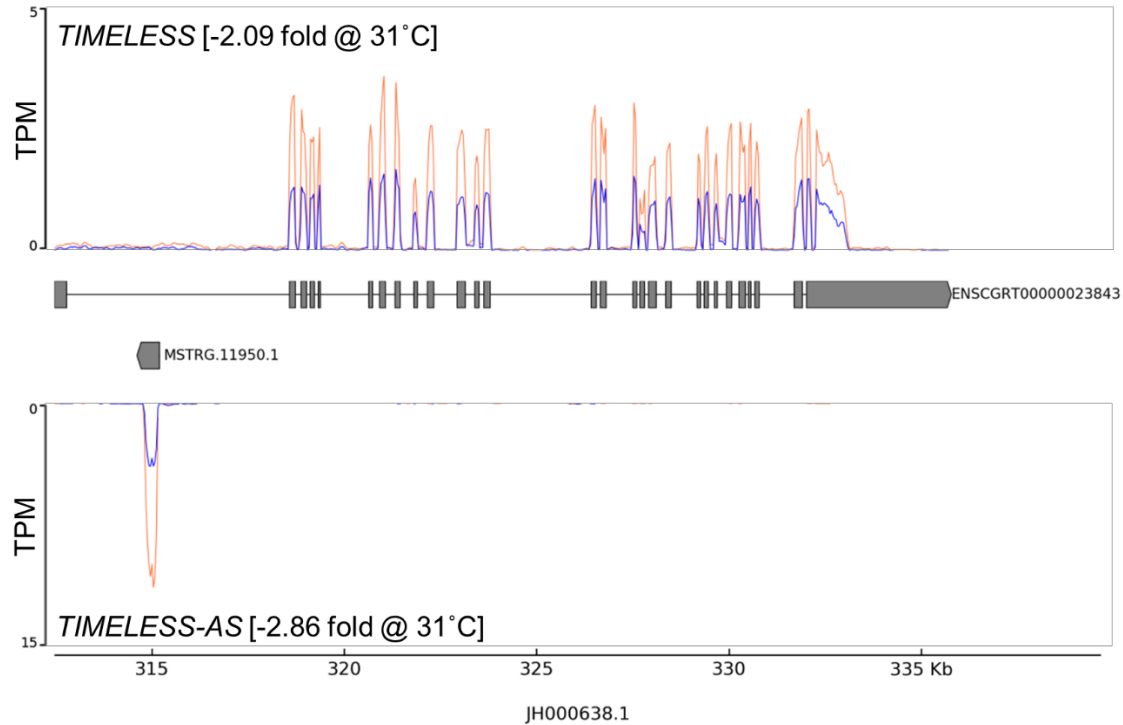


Figure 2.19 A coverage plot comparing average sense and antisense TPM normalized coverage between NTS (red) and TS (blue) of the Timeless locus which contains the Timeless gene on the sense strand and a mono-exonic ncRNA on the antisense. Both the sense and antisense were downregulated after a reduction in culture temperature.

The remaining 11 genes were found to have anti-correlated expression where the sense protein coding gene is often downregulated with an upregulated antisense ncRNA. There are two loci where the antisense ncRNA is downregulated while its partner coding gene is upregulated (**Table 2.3**): a coiled-coiled domain encoding protein and a novel CHO specific protein coding gene. The two most abundant antisense genes are anti-correlated with their protein coding gene: *Cbx3* and *Lrp8*. A novel multi-exonic gene antisense to the *3-oxoacid CoA transferase 1 (OXCT1)* protein coding gene was the most strongly differentially expressed gene within this cohort (**Figure 2.20**). *OXCT1* is downregulated -1.58 fold (BH Adj. p-value: 2.77×10^{-29}) while the antisense non-coding gene is strongly upregulated 4.78 fold (BH Adj. p-value: 4.07×10^{-83})

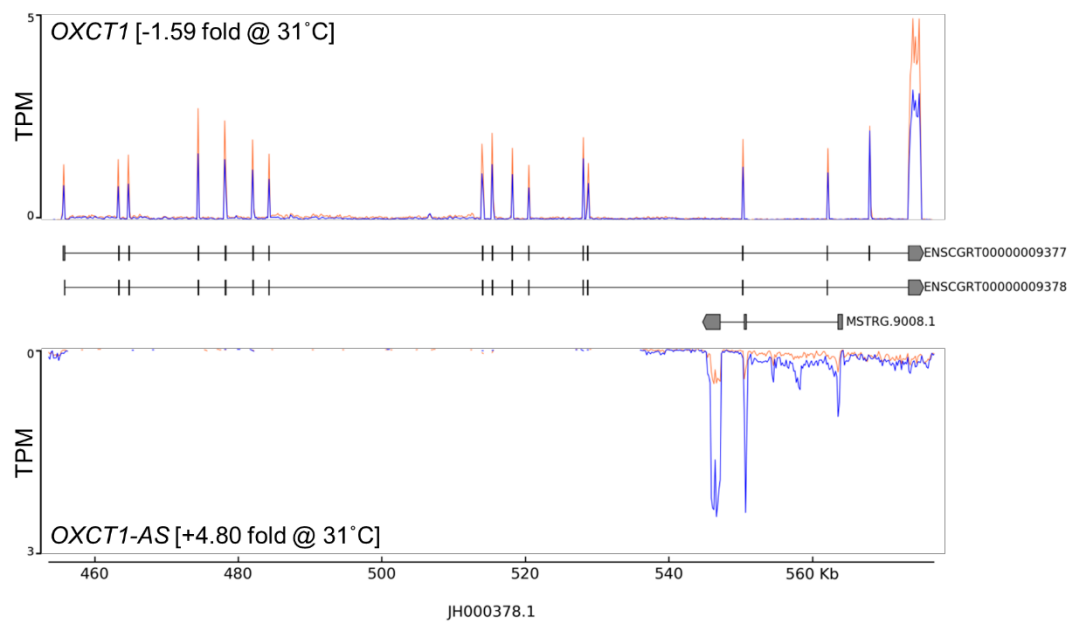


Figure 2.20 A coverage plot comparing average sense and antisense TPM normalized coverage between NTS (red) and TS (blue) of the *Oxct1* locus which contains the *Oxct1* gene on the sense strand and a multi-exonic ncRNA on the antisense. The antisense gene is strongly upregulated while the sense protein coding gene is significantly downregulated after temperature shift.

Table 2.3 Log₂-fold-changes and BH adjusted p-values for sense protein coding genes and antisense lncRNAs. There are 6 sense/antisense gene pairs that exhibit correlated expression and 11 sense/antisense gene pairs that exhibit anti-correlated expression.

Coding Gene Symbol	Sense Coding Gene	lncRNA baseMean	lncRNA log2FC	lncRNA BH Adj. p-value	lncRNA type	mRNA log2FC	mRNA BH Adj. p-value	Correlation
<i>OXCT1</i>	<i>3-oxoacid CoA transferase 1</i>	368.17	2.26	4.07×10^{-83}	antisense	-0.66	2.77×10^{-29}	Anti-correlated
<i>PRPS2</i>	<i>Phosphoribosyl pyrophosphate synthetase 2</i>	157.07	1.85	1.78×10^{-32}	antisense	-1.00	4.19×10^{-114}	Anti-correlated
<i>KIF23</i>	<i>Kinesin family member 23</i>	160.73	1.75	3.88×10^{-37}	antisense	-0.84	8.62×10^{-103}	Anti-correlated
<i>LMNA</i>	<i>Lamin A</i>	417.10	1.64	2.76×10^{-53}	antisense	0.70	2.02×10^{-58}	Correlated
<i>OSTM1</i>	<i>Osteoporosis associated transmembrane protein 1</i>	200.18	1.30	2.48×10^{-19}	divergent	0.70	6.33×10^{-29}	Correlated
<i>NFKB2</i>	<i>Nuclear factor of kappa light B2</i>	203.54	1.17	5.11×10^{-19}	divergent	0.66	5.23×10^{-39}	Correlated
<i>S100A3</i>	<i>S100 calcium binding protein A3</i>	144.82	1.13	4.95×10^{-15}	divergent	1.15	3.79×10^{-19}	Correlated
<i>BMERB1</i>	<i>bMERB domain containing 1</i>	189.64	1.08	6.29×10^{-17}	antisense	-1.38	3.22×10^{-193}	Anti-correlated
<i>CBX3</i>	<i>Chromobox 3</i>	659.00	1.08	3.02×10^{-44}	antisense	-0.75	4.00×10^{-44}	Anti-correlated
<i>GPRC5A</i>	<i>G protein-coupled receptor</i>	330.08	1.07	8.41×10^{-26}	antisense	0.60	3.07×10^{-15}	Correlated
<i>CAPN8</i>	<i>Calpain 8</i>	104.91	1.02	1.05×10^{-9}	antisense	0.68	9.69×10^{-9}	Correlated
<i>CENPF</i>	<i>Centromere protein F</i>	220.97	1.00	7.90×10^{-17}	divergent	-0.71	9.94×10^{-66}	Anti-correlated
<i>GINS4</i>	<i>GINS complex subunit 4 (Sld5 homolog)</i>	181.20	0.76	1.85×10^{-8}	antisense	-0.93	9.42×10^{-57}	Anti-correlated
<i>LRP8</i>	<i>Low density lipoprotein receptor-related protein 8</i>	705.58	0.75	4.64×10^{-19}	antisense	-0.74	2.90×10^{-33}	Anti-correlated
<i>LOC100752852</i>	<i>Schlafen family member 12</i>	124.37	0.59	1.31×10^{-3}	antisense	-0.82	4.29×10^{-21}	Anti-correlated
<i>MROH8</i>	<i>Maestro heat-like repeat family member 8</i>	208.81	-0.61	5.89×10^{-3}	antisense	-1.13	3.99×10^{-15}	Correlated
<i>LOC100752605</i>	<i>Histone H4</i>	449.32	-0.66	9.47×10^{-13}	antisense	-2.10	6.99×10^{-183}	Correlated
<i>CCDC162</i>	<i>Coiled-coil domain containing 162</i>	293.62	-0.79	3.21×10^{-14}	antisense	0.64	1.29×10^{-5}	Anti-correlated
<i>RAD51B</i>	<i>RAD51 paralog B</i>	366.34	-1.29	2.68×10^{-28}	antisense	-1.33	3.53×10^{-53}	Correlated
<i>NOVEL</i>	<i>Novel CHO gene</i>	248.32	-1.43	8.37×10^{-19}	antisense	0.78	3.26×10^{-14}	Anti-correlated
<i>TIMELESS</i>	<i>Timeless circadian regulator</i>	305.70	-1.52	2.15×10^{-29}	antisense	-1.06	1.62×10^{-47}	Correlated

2.4.13 Alterations in metabolite concentration after a reduction in culture temperature

Lactate, glutamine, and ammonium concentrations were profiled 72 hours after initial seeding between non-temperature-shifted and temperature-shifted samples. The average lactate concentration was reduced by 19% between samples cultivated at 37°C and 31°C. Glutamine concentration significant increased between the two conditions with a 97% increase in concentration while the extracellular ammonia concentration decreased by 8% between 37°C and 31°C (**Figure 2.21**).

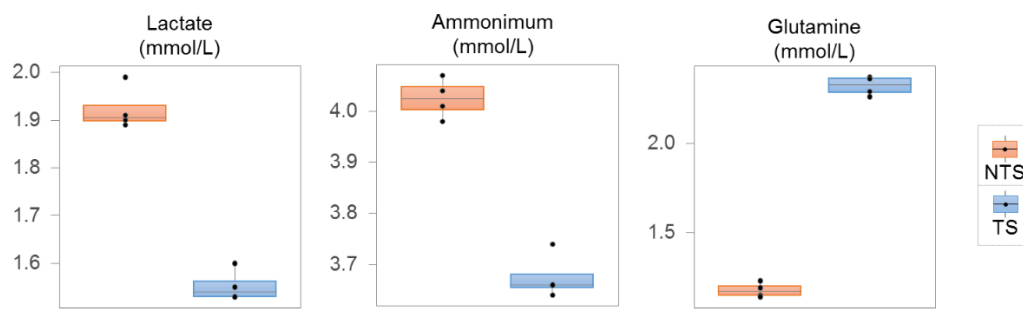


Figure 2.21 Concentrations of lactate, ammonia (NH_4^+ ions), and glutamine in mmol/L between NTS and TS samples. Lactate and ammonia concentrations were reduced between TNS and TS samples while glutamine concentrations were shown to significantly increase after TS.

2.5 Discussion

The publication of the Chinese hamster and CHOK1 genomes has greatly enhanced the capability for rational cell engineering (Rupp et al. 2014). Annotation of coding gene structures, such as exon and transcript composition, and genomic localization has enabled targeted gene knock-in and knock-out techniques to modulate the expression of specific pathways (Freeman et al. 2010; Fischer, Handrick, and Otte 2015). Prior to this, researchers were heavily reliant on mouse annotation to perform high-throughput experiments such as microarrays or proteomics assays. The generation of genomic references was only possible with the development of next generation sequencing techniques which can sequence hundreds of individual DNA molecules in high-throughput (N. E. Lewis et al. 2013; Rupp et al. 2018). Recent advances in RNA-Seq library prep combined with a reduction in price of sequencing can enable identification of lncRNA transcripts which are often expressed at lower abundances and transcribed in genic regions such as within the intron or antisense to a protein coding gene. In this study, we utilize high-coverage, strand-specific RNA-Seq to profile long non-coding expression in relation to an industrially relevant condition: a reduction in culture temperature. We have constructed a software pipeline that can process raw sequencing data in order to assemble and annotate high-confidence lncRNAs.

2.5.1 Comparison of annotation to GENCODE

The results of our annotation pipeline correlate with findings of the initial GENCODE assembly of the human and mouse lncRNA transcriptomes. We observed that lncRNAs have a lower GC content and were found to be of shorter length than mRNAs which was similar to the assemblies generated by Derrien *et al.* (Derrien et al. 2012). Protein coding genes have multiple features that are necessary for translation such as an open reading frame containing a canonical start and stop codons with the ORF body flanked by untranslated regions (Harrow et al. 2012). lncRNAs are hypothesized to be shorter as features that are necessary for translation are truncated or not present (Azali et al. 2019). The longer transcript length of mRNAs also correlates to a larger number of exons compared to lncRNAs which we observed within our *StringTie* assembly as bi

and tri-exonic lncRNA transcripts are significantly represented within the lncRNA transcriptome but an even distribution of exons per transcript was observed for mRNAs. Mono-exonic transcripts are often omitted from lncRNA transcriptomic analysis as they are likely to be sequencing noise or transcribed host genes for small non-coding RNAs (Kern et al. 2018). The initial step of the filtration stage of our pipeline removes mono-exonic intergenic lncRNAs which could introduce *in silico* bias to the final assembly, however. The filtration step is likely removing *bona fide* intergenic mono-exonic lncRNAs but this stage is necessary for generating high-confidence reference transcripts. A total of 8% of the total GENCODE human and mouse long non-coding transcriptome are mono-exonic whereas our assembly identified approximately 20% of the total lncRNA transcriptome as consisting of a single exon. Further sequencing experiments will likely expand the repertoire of annotated lncRNAs which will likely reduce the overall proportion of mono-exonic transcripts.

2.5.2 Differences in isoform and exon composition to GENCODE standard

Single isoform genes dominated both coding and lncRNA classes. Many transcript isoforms of a parent gene exhibit tissue-specific expression and may not be present in CHO cells. The high percentage of genes with single isoform genes can be attributed to profiling only those transcripts that are expressed within specialized cell lines. RNA-Seq experiments that studied tissue specific expression of isoforms found significant differences between cell types and different isoforms are found to have different functions across cell types (Kandoi and Dickerson 2019). For example, aberrant splice isoforms that result in loss of protein coding domains are frequently associated with malignant cancers (Vitting-Seerup and Sandelin 2017). Heffner *et al* used proteomics data derived from parental Chinese hamster liver and ovary tissues to identify tissue specific isoforms (Heffner et al. 2017). This study is an important first step in expanding gene and isoform numbers for CHO cells but is limited due to profiling translated proteins. Further sequencing of the transcriptome of parental Chinese hamster tissues will expand the number of isoforms.

2.5.3 Temperature shift induced global change in gene expression

Temperature shift is commonly used technique in recombinant protein production as this method improves cell viability which increases total yield. CHO cells subjected to a temperature shift undergo growth arrest where there are significantly higher proportions of cells stalled at the G₀/G₁ which would result in a reduced cell doubling (J. Xu et al. 2019). A 24% difference in cell density was observed between cells cultivated at 37°C compared to 31°C which is consistent with previous studies (Altamirano et al. 2006). A reduction in culture temperature has been previously described to induce metabolic shifts in CHO cells (Bedoya-López et al. 2016). A common change is the transition from lactate production to lactate consumption which is highly desirable as lactate is a toxic metabolite that can affect cell growth and cause culture acidification (Hartley et al. 2018). Temperature shift also reduces the accumulation of toxic metabolites such as ammonia which is highly advantageous as this metabolite can induce apoptosis (Wahrheit, Nicolae, and Heinzle 2014). Metabolite profiling of our cell culture found a 19% decrease in average lactate concentration and an 8% decrease in ammonia concentration in temperature-shifted cells which is consistent with previous studies (J. Xu et al. 2019). We also observed a dramatic increase in glutamine concentration which can be attributed to cells undergoing growth arrest. Glutamine is essential for proliferating CHO cells and the accumulation of this nutrient within the extracellular media may indicate reduced consumption of glutamine (Wahrheit, Nicolae, and Heinzle 2014). Ammonia is produced as a result of glutamine consumption so would not accumulate as readily in cells that are not using glutamine as a substrate (Wahrheit, Nicolae, and Heinzle 2014). The phenotypes associated with CHO cells subjected to temperature shift occur as a result of changes in gene expression so studying gene differential expression will provide more information on regulatory alterations within the CHO transcriptome.

Differential expression analysis of coding genes was equivalent to previous studies that profiled mRNAs with respect to a reduction in culture temperature (Bedoya-López et al. 2016). The cold induced genes *CIRBP* and *RBM3* were strongly upregulated after temperature (X. Zhu, Buhrer,

and Wellmann 2016). *CIRBP* is a chaperone protein that is thought to stabilize mRNA structure by migrating to the cytoplasm during cellular stress such as hypothermia (Hong et al. 2007) where it binds to the 3' UTR. Upregulation of *CIRBP* has previously been shown to increase total volumetric yield of recombinant proteins 2-5 fold (Tan et al. 2008). *RBM3* is a well-characterized gene that encodes an RNA binding protein that can associate with the 60S subunit to increase global translation after mild hypothermia (X. Zhu, Buhrer, and Wellmann 2016). Both *CIRBP* and *RBM3* are activated by temperature sensitive triggers with *CIRBP* often expressed at higher levels than *RBM3* (X. Zhu, Buhrer, and Wellmann 2016). In addition to previously identified cold-inducible genes, we found strong upregulation of the *Lactate dehydrogenase A (Ldha)* gene and strong downregulation of the *FOS* transcription factor. *LDHA* is one of the most upregulated genes after a reduction in culture temperature and preferentially catalyzes pyruvate into lactate (Ratter, Rooijackers, and Hooiveld 2018) which is then consumed by the cell to regenerate NADH. The *FOS* transcription factor regulated transcription of a wide array of genes and cellular pathways. Please refer to **Chapter 5** for a more thorough discussion on the *FOS* transcription factor complex. Modulation of expression of cold inducible coding genes has previously been demonstrated to increase recombinant protein production. lncRNAs have a similar potential as cell line engineering molecules as these regulatory transcripts can govern entire pathways.

2.5.4 Evolutionarily conserved lncRNAs differentially expressed after TS

A cohort of evolutionarily conserved lncRNAs, such as *LINC00511*, *PVT1*, and *HOTAIRM1*, were found to be abundant and differentially expressed after a reduction in culture temperature. These conserved loci are relatively well-characterized lncRNAs which can be used to infer probable regulatory function within Chinese hamster ovary cells. A recent study by Zhang *et al.* demonstrated the importance of *LINC00511* in cell proliferation and viability of breast cancer cell lines (Jian Zhang et al. 2019a). Overexpression of *LINC00511* was demonstrated to increase cell viability by ~25% by recruiting chromatin modifying complexes to methylate cyclin dependent kinases (Jian Zhang et al. 2019b). This study suggests that *LINC00511* is involved in increased

cell viability after temperature shift in CHO cells. We have also identified significant differential expression of *Plasmacytoma variant translocation 1 (PVT1)* lncRNA gene which is a known oncogene that drives tumorigenesis in diverse cancer cells (Y. Yu et al. 2018). In humans, the *PVT1* gene is composed of 182 individual isoforms and has diverse regulatory functions depending on which isoform is expressed such as forming fusion transcripts during DNA rearrangement, titrating endogenous miRNAs away from target sequences, and stabilizing *cMYC* mRNAs (M. Cui et al. 2016). Vito and Smales previously demonstrated differential *PVT1* expression between batch and fed-batch cultures using a mouse microarray (Vito and Smales 2018) and we have expanded its functional role to include differential expression after temperature shift. We *HOXA1* antisense transcript *HOTAIRM1* was found to be one of the most abundant, most differentially expressed lncRNAs after a reduction in culture temperature. Previous studies that profiled *HOTAIRM1* expression in colorectal cancer cell lines identified this gene as a tumor suppressor that regulated cell proliferation (Wan et al. 2016) as well as a key initiator of autophagy within eukaryotic cells (Z. Chen et al. 2016). Autophagy is the process where eukaryotic organisms catabolize macromolecules and organelles and often occurs in response to cell stress which can also regulate cell growth and cell cycle progression (Gump and Thorburn 2011; Jung et al. 2010). The regulatory relationship between cell cycle progression and autophagy can be complex but previous studies have shown an increase in autophagy during G_0/G_1 phases (Mathiassen, De Zio, and Cecconi 2017). Cell cycle arrest in G_0/G_1 phases is often observed within CHO cells that were subjected to a reduction in culture temperature (Bedoya-López et al. 2016) and we hypothesize that *HOTAIRM1* is a key regulatory molecule that induces autophagy after temperature shift. The cohort of annotated, evolutionarily conserved non-coding RNAs are useful for future cell line engineering efforts as they are relatively understood in model organisms like humans and mice and conservation suggests similar regulatory functions. Further *in vitro* validation is required to evaluate conserved function, however.

2.5.5 Sense/antisense differential expression

Analysis of antisense transcription is only possible with library preparation methods that preserve strand orientation (Wight and Werner 2013). The putative *cis* regulatory function of this cohort of antisense RNAs could be time dependent. lncRNAs are known to be actively transcribed after cellular stress response. Antisense lncRNAs can regulate in *trans* at distal genomic loci and therefore the expression of the sense protein coding gene would be unaffected. Computational methods to predict potential regulatory functions from RNA-Seq data are still quite immature and are of limited utility even within model organisms. These computational tools rely on novel high-throughput sequencing methods such as CLIP-Seq, which uses chromatin immunoprecipitation to preferentially enrich RNA bound to specific proteins, to infer potential function such as binding epigenetic remodeling complexes. Computational tools that have been designed to take advantage of this enriched sequencing data are dependent on high quality reference annotations to predict potential function. The high-quality annotation of novel antisense ncRNAs will be useful for future high-throughput studies that can probe regulatory targets.

There are 21 gene loci that are differentially expressed on both the sense and antisense strand with 10 of these pairs found to have correlated expression. There is debate whether correlated expression of antisense transcription is not simply transcriptional noise due to increased transcription on the sense strand (Pelechano and M. 2013; Wight and Werner 2013). There is evidence of a functional role for antisense lncRNAs co-expressed with their sense protein coding gene, however. Balbin *et al.* discovered that siRNA-mediated knockdown of the sense *NKX2* oncogene downregulated its antisense non-coding gene, *NKX2-AS* (Balbin et al. 2015) in a lung cancer cell line model. Conversely, knock down of the antisense *NKX2-AS* reduced expression of the sense *NKX2* gene as well as inhibited cell proliferation (Balbin et al. 2015) which suggests the existence of tightly modulated sense/antisense co-regulatory networks. One proposed mechanism of *cis* regulation by antisense RNAs is to function as enhancers to increase transcription of the sense strand by recruiting transcription factors to genomic loci (Ilott et al.

2014). Downregulation of the antisense enhancer would result in a similar downregulation of the protein coding gene. The circadian rhythm regulator *TIMELESS* locus contains a strongly downregulated mono-exonic gene that could function as a potential enhancer RNA. Circadian rhythms govern checkpoints between cell cycle states with *TIMELESS* necessary for the transition to the G₂/M phases (Feillet et al. 2015). Downregulation of this circadian rhythm regulator would result in increased G₀/G₁ phase cells which is a common phenotype of cells cultured at sub-physiological temperatures. The mono-exonic antisense gene could potentially upregulate *TIMELESS* expression at 37°C while a temperature shift would result in significant downregulation of the entire locus. Further analysis is required to evaluate whether the sense and antisense genes share expression or are independent.

We present the first evidence of anti-correlated expression of sense/antisense pairs within Chinese hamster ovary cells. The majority of these pairs exhibited increased expression on the antisense strand and downregulation of the sense strand which suggests that these transcripts function as transcriptional inhibitors. The mechanism of inhibition can be diverse and includes epigenetic silencing of the sense promoter or sequestering transcription factors from their target binding sites (Wight and Werner 2013). There are two genes where the converse is true where the antisense is downregulated and the sense, upregulated so not all antisense non-coding genes can be classified as repressors. The mitochondrial matrix protein has one of the most strongly anti-correlated antisense lncRNAs after a reduction in culture temperature. *Oxct1* is significantly downregulated while its multi-exonic antisense counterpart is strongly upregulated. *OXCT1* catalyzes the conversion of succinyl-CoA ketone body into acetate which can be converted into acetyl-CoA for entry into the citric acid cycle (S. Zhang and Xie 2017). Downregulation of *OXCT1* would result in a reduced metabolic rate which is an important transition in cells that undergo a reduction in culture temperature. We propose that upregulation of the antisense *OXCT1* ncRNA gene is the mechanism of reduced transcription of the sense protein coding gene. Anti-correlated

lncRNAs are expressed at very low abundances so further *in vitro* validation is required to test whether these antisense molecules are functional or whether they are transcriptional noise.

2.5.6 Limitations of analysis using lncRNA-Seq data

A limitation of this study is that we only profiled gene expression in a single phenotype, temperature shift, within a single cell line. lncRNAs are regulatory molecules that are often transcribed first due to cellular stress as these transcripts are necessary for modulation of cellular pathways (Su et al. 2016). As RNA was extracted 24 hours after the initial hypothermic shock, the immediate transcriptional regulatory response would not have been captured by library preparation. In **Chapter 3**, we profiled lncRNA expression over the duration of a 14-day fed-batch culture in order to identify temporal patterns in non-coding gene expression.

2.6 Conclusions

We have constructed a software pipeline that can assemble, detect, and classify long non-coding RNAs from strand-specific, high coverage RNA-Seq data. We utilized NGS data derived from CHO cells that underwent a reduction in culture temperature in order to detect lncRNAs. We have expanded the annotation of the long non-coding transcriptome by over 50% using this data and classify them into categories as specified by GENCODE and identify a cohort of novel intergenic, antisense, and sense loci. Many of these non-coding RNAs are not functionally annotated so it is quite challenging to extract industrially relevant information from these molecules. We have identified significantly differentially expressed long non-coding RNAs after a reduction in culture temperature from 37°C to 31°C with our software pipeline and present a set of differentially expressed, evolutionarily conserved lncRNAs. We also presented the first evidence of correlated and anti-correlated co-expression of sense protein coding genes with their antisense long non-coding RNAs.

Chapter 3

Differential Expression of

lncRNAs During Fed-Batch Culture

3 Summary

In **Chapter 3**, we first summarize important bioprocessing relevant parameters such as viable cell concentration and viability during industrial fed-batch cultivation as well as the relevance of several metabolites such as glucose and lactate that are routinely profiled in industrial culture (**Section 3.1**). A large RNA-Seq data was generated from CHO cells cultivated as fed-batch cultures within two 10L bioreactors. This RNA-Seq dataset was part of a comprehensive, multi-omics study of CHO cells that also included genomics, proteomics, and metabolomics data in order to profile alterations in cellular phenotype over an industrial culture period. The pipeline used in **Chapter 2** was used to detect and annotate lncRNAs and correlate expression with exponential, stationary, and death phases (**Section 3.2-Section 3.3**). Bioprocessing relevant parameters were monitored over the course of the culture and were correlated to mRNA and lncRNA alterations in expression. **Chapter 3** will conclude on a discussion of the biological significance of alterations lncRNA expression over time and the potential relation with metabolite concentration.

3.1 Introduction

The invention of fed-batch cultivation has significantly increased cell growth, viability, and total recombinant protein yield (Dean et al. 2013). During traditional batch cultivation, mammalian cells are seeded into a closed system containing a defined concentration of metabolites. Nutrients are depleted and toxic metabolites accumulate as over the duration of the culture which limits growth and product yield (Butler and Meneses-Acosta 2012). In fed-batch culture, a bolus of additional nutrients, such as glucose and amino acids, is periodically administered to the reactor to replenish depleted metabolites (Y. Huang et al. 2010). The addition of additional nutrients to the culture media significantly increases growth rate and viability when compared to cells cultured in batch conditions (Butler and Meneses-Acosta 2012). Fed-batch cultivation can also drastically increase cell densities with > 5g/L reported (Y. Huang et al. 2010). Periodic feeding also reduces the accumulation of toxic metabolites like lactate and ammonia in extracellular media which would reduce cell growth and viability (Y. Huang et al. 2010). As a result, fed-batch cultures are the most common cultivation method in industrial recombinant protein production within the pharmaceutical industry.

Increased proliferation and decreased apoptosis are two highly desired phenotypes for industrial bioprocessing. A high yield growth phase is essential as total recombinant protein yield is positively correlated with cell density and viability (Fischer, Handrick, and Otte 2015). Delaying the onset of the death phase and preventing apoptosis is of significant industrial interest as cells remain productive for a longer period of time. Lysis during programmed cell death can also affect the final quality due to the release of proteases into the media which could introduce unwanted protein heterogeneity (Y. K. Han et al. 2011). Fed-batch media has significantly prolonged cell viability and reduced apoptosis which can be attributed to the significant improvements in basal and feed media composition (Y. Huang et al. 2010).

The media used in CHO cell culture contain nutrients necessary for cell growth with glucose and glutamine as the main energy sources (Reinhart et al. 2015). Glucose is the major source of energy and carbon for eukaryotic cells and is a core component of culture media that is necessary for cell growth (Vergara et al. 2018). Glucose is metabolized during aerobic respiration in eukaryotic organisms by conversion into pyruvate during glycolysis which, in turn, serves as substrate for the citric acid (TCA) cycle (**Figure 3.1**).

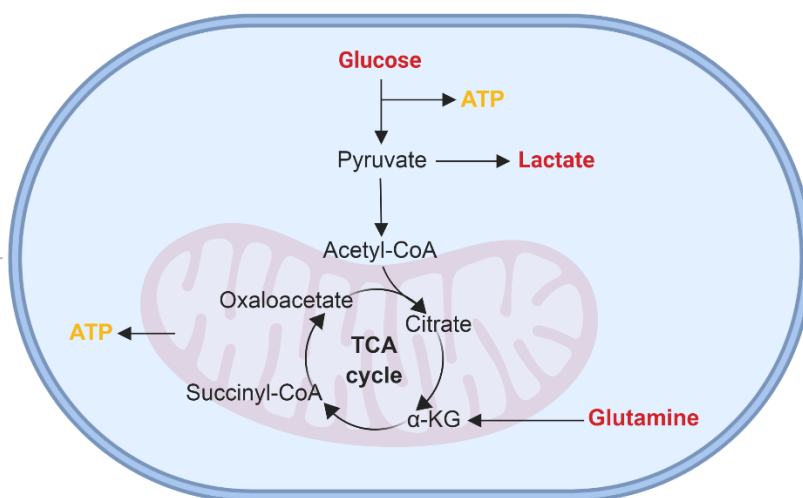


Figure 3.1 A summary of the glycolysis and TCA cycle pathways. Glucose is first metabolized to pyruvate for conversion in the Acetyl-CoA. Aerobic respiration occurs inside the mitochondria during the citric acid cycle which generates ATP. Rapid glycolysis can lead to the accumulation of lactate which can inhibit CHO cell growth. Glutamine can also be a source of energy as well as nitrogen.

Glycolytic activity is the highest during the exponential phase of culture whereas glucose metabolic pathways within the TCA cycle are more enriched during stationary phase. Glucose concentrations within bioreactors are routinely monitored within industrial bioreactors as sugar consumption is linearly correlated with cell growth and productivity (Tsao et al. 2005). Modern industrially available media formulations contain glucose concentrations that range between 25 to 35 mM (Vergara et al. 2018). Alternative sugars, such as galactose, have been suggested as the main energy source in media in lieu of glucose but altering this critical component leads to reduced growth rate and unwanted variability in product glycosylation (Altamirano et al. 2006).

Rapid glucose consumption can lead to the accumulation of lactate which can be a toxic metabolite as it can directly inhibit cell growth, downregulate metabolic enzymes, and alter the pH of the bioreactor (Hartley et al. 2018). Lactate is a byproduct of glycolytic pathways and is necessary for the replenishment of the NAD^+ co-factor which is required for continued glycolytic metabolism (Hartley et al. 2018). Lactate is converted from pyruvate by the lactate dehydrogenase (*LADH*) enzyme and is a common metabolite that is profiled throughout the duration of a production run (Hartley et al. 2018). This toxic metabolite is commonly produced in cells during the exponential phase, referred to as the “Warburg effect”, and a metabolic transition occurs in the stationary phase where CHO cells switch from producing lactate to consuming lactate (Hartley et al. 2018). The lactate production/consumption switch can also occur due to environmental triggers such as depleted glucose, shift in pH as well as alteration of culture temperature (Zalai et al. 2015). This metabolic transition is highly desirable as it depletes cellular lactate and is highly correlated with improved metabolic efficiency as well as increased recombinant protein production (Le et al. 2012).

Glutamine is also a major energy and carbon source as well as an important source of nitrogen during CHO cell cultivation. Glutamine is converted into energy by entering the mitochondria where it is first converted to glutamate by the *glutaminase* (*GLS*) gene followed by entry into the TCA cycle as α *ketoglutarate*. Glutamine is a component in many basal media formulations as it is readily used by rapidly dividing cells during the exponential phase of the cell cycle. Glutamine based CHO selection is commonly used in industrial bioprocessing where CHO cells deficient in the glutamine synthetase (*GS*) gene are transfected with recombinant DNA that contains a functional copy of the *GS* gene so cells expressing product can be selected (Noh, Shin, and Lee 2018). Glutamine metabolism results in the production of ammonium ions which have cytotoxic properties and can inhibit cell growth, however (Wahrheit, Nicolae, and Heinzle 2014). Increasing ammonia concentrations can also have an adverse effect on product glycosylation as well

(Wahrheit, Nicolae, and Heinzle 2014). Metabolic engineering has had some success in reducing ammonia production during CHO cell cultivation by overexpressing the GS gene in media devoid of glutamine to instead use glutamate as a substrate as a substrate for the TCA cycle (F. Zhang et al. 2006).

Metabolite concentrations have profound effects on CHO cell growth rate and viability and are therefore routinely measured and quantified in order to monitor cell growth and viability (Butler and Meneses-Acosta 2012). High-throughput omics technologies, which includes transcriptomics, proteomics, and metabolomics have previously been used to study phenotypic changes during fed-batch culture. Schaub *et al.* were one of the first to study the transcriptome of cells producing an IGG in fed-batch conditions over time (Schaub et al. 2010). The authors compared high and low producing cell lines and found that high-titer cell lines have genes involved in lipid metabolism upregulated (Schaub et al. 2010). Dean & Reddy performed a comprehensive metabolomics profiling study over an 8 culture period into lag, exponential, stationary, and death phases between high and low producing cell lines (Dean et al. 2013). A metabolic shift was observed where cells started to consume glucose over glutamine as the growth rate decreased (Dean et al. 2013). The production of lactate was also observed to dramatically decrease during the stationary phase compared to the exponential phase (Dean et al. 2013). More recently, Hsu *et al.* demonstrated an integrated method that combined matched metabolomics and transcriptomics data over time in order to identify gene pathways that regulate cell metabolism in fed-batch culture (Hsu et al. 2017). This study was the first to combine metabolome profiling with RNA-Seq and successfully demonstrated the utility of using paired data in order to identify correlation between gene expression and metabolite concentration such as lactate (Hsu et al. 2017). The RNA-Seq dataset focused exclusively on coding genes, however, and there has not yet been a study that correlated ncRNA expression with industrially relevant metabolite concentrations within batch or fed-batch cultures.

Previous “omics” experiments have successfully been used to profile changes in gene expression and the metabolome in CHO cells cultivated in fed-batch systems. There has, to date, not been a comprehensive study of the long non-coding transcriptome in fed-batch conditions. The first transcriptomic study to profile lncRNAs used a mouse microarray to study differential expression between batch and fed-batch cultures between two time points (Vito and Smales 2018). Vito and Smales found the evolutionarily conserved lncRNAs *MALAT1*, *NEAT1*, and *PVT1* were significantly upregulated (fold-change > 3) in fed-batch cultures over batch cultures. The authors’ reliance on a mouse microarray limited the number of lncRNAs that could be profiled as only 28% of mouse probes shared significant similarity to the *C. griseus* non-coding transcripts (Vito and Smales 2018). The work performed by Vito and Smales is an important first step in profiling non-coding gene expression in fed-batch cultures as it demonstrated modulation of non-coding genes over time and after feeds.

In this chapter, we present the results from an RNA-Seq experiment that profiled the transcriptome of two duplicate 10L, stirred-tank bioreactors where the seeded cells are expressing a recombinant IgG1 antibody over a 14-day culture period. Over 22 variables that are relevant to industrial bioprocessing, including viable cell concentrations, viability, and growth rate were tracked over the duration of the culture. This experiment is part of a comprehensive, multi-omics study that used diverse technologies including genomics, metabolomics, miRNA-Seq, proteomics, phosphoproteomics, and glycomics to develop a high-resolution map of CHO cell behavior and expression during the course of an industrial production run and will be the most comprehensive dataset available to date. The multi-omics experiment was conducted by Marie Curie eCHO international training network whose members include industrial and academic researchers. The software pipeline presented in **Chapter 2** was used to quality filter raw data, align reads to the *ENSEMBL* version 98 *criGri1* genome, and identify evolutionarily conserved and novel long non-coding loci. lncRNAs were then segregated into genes that were associated

with the exponential, stationary, and death phases. The RNA-Seq data was also integrated with metabolite profiling to identify differentially expressed lncRNAs that have expression correlated with metabolite concentrations.

3.2 Materials and Methods

3.2.1 CHO cell culture within industrial 10L stirred tank bioreactors

The cell cultures and metabolomics profiling presented in this chapter were performed by Cyrielle Calmels and Linas Tamošaitis with industrial partners at UCB Biopharma in Brussels, Belgium. All sequencing was performed in collaboration with the Austrian Center of Industrial Biotechnology in Vienna, Austria. A DG44 cell line producing a recombinant IgG1 antibody was cultivated with industrial partners in two replicates, glass stirred tank 10L bioreactors (Sartorius Stedim Biotech). The two replicate 10L bioreactors were initially inoculated with 3.5×10^5 in 6.5 L of Hyclone ActiPro chemically defined media (GE Healthcare) which was supplemented with glutamate. Cultivation temperature and pH were kept constant at 36.8°C and 7.0 respectively and monitored by a multi-fermenter control system (MFCS, Sartorius Stedim Biotech). Dissolved O₂ was maintained at 40% of saturation and replicate bioreactors were monitored over a 14-day cultivation period using a fed-batch process with daily feeding after 72 hours post-inoculation. A bolus of glucose was administered to the reactors if the dissolved glucose concentration fell below 5.6 g/L. **Table 3.1** summarizes the amino acid composition of the bolus administered during the fed-batch culture. Cells were sampled daily prior to glucose feeding to evaluate cell density and viability using a VI-Cell XR (Beckman-Coulter) with trypan blue staining. Cells were harvested for downstream multi-omics analysis with 1.2×10^7 cells extracted at each day prior to feeding for total RNA-Seq library preparation. Metabolite concentrations including glucose, lactate, glutamine, glutamate, and ammonia were calculated using a Cedex Bio HT Analyzer (Innovatis).

Table 3.1 Amino acid and ammonia composition of fed-batch feed. Glucose was continuously monitored throughout the process and supplemented when concentrations declined below 5.6 g/L

Feed	Concentration (mmol/L)
NH3	0
His	10.85
Asn	68.75
Tau	0.95
Ser	86.85
Gln	0
Arg	28.35
Gly	0
Asp	48.75
Glu	85.65
Thr	47.45
Ala	15.8
Pro	29.8
Cys	0
Lys	43.65
Tyr	0
Met	12.05
Val	64.35
Ile	34.25
Leu	59.15
Phe	33.6
Trp	0

3.2.1 RNA-Seq library preparation

Harvested cells were centrifuged and extracted total RNA was used as template for rRNA depletion using the RiboZero Gold kit (Illumina, Inc). A 40 μ L reaction volume containing up to 1 μ g of extracted RNA was incubated at 68°C for 10 minutes to ensure hybridization of rRNAs to their complementary probes. Magnetic bead based enrichment was used to remove hybridized ribosomal sequences and the purified RNA was used as template for RNA-Seq library preparation using the NEBNext total RNA Ultra II Directional RNA library prep kit (New England Biolabs). Samples were initially treated with DNase I to remove gDNA contaminants and strand-specific libraries were prepared using random hexamer priming. Paired-end, 2 x 125bp sequencing was performed on an Illumina HiSeq instrument in order to ensure a minimum of 50 million fragments

per sample. Library preparation and sequencing was repeated when samples were found to have a high percentage of RiboZero antisense probe contamination.

3.2.3 RNA-Seq library QC

Raw read data was pre-processed using *cutadapt* and *Trimmomatic* using the QC and assembly pipeline as previously described in **Chapter 2**. QC filtered reads were aligned to the *ENSEMBL criGri1* genome with the version 98 annotation. Transcript models were assembled for each sample using StringTie with each putative sample requiring a minimum of 5 reads supporting a junction. A consensus transcriptome model was then constructed and long non-coding RNAs were identified using parameters outlined in **Chapter 2**. Identified lncRNAs were classified into sense, antisense, and intergenic groups with respect to genome localization and orientation.

3.2.4 Differential expression between cell culture phases

The consensus transcriptome assembly was used as a reference to quantify reads with *HTSeq* in a strand-specific manner. Raw gene counts were imported into the *DESeq2* software package within the R statistical framework. Sample libraries from days 13 and 14 were excluded from the time series analysis due to insufficient read depth and poor read quality and remaining samples were classified into exponential, stationary, and death phases. Pairwise differential expression was performed between exponential vs. stationary and stationary vs. death phase samples. A gene was considered significantly differentially expressed if it had a fold-change greater than or equal to ± 1.5 and a Benjamini-Hochberg adjusted P-value < 0.05 .

3.2.1 Correlation of expression to bioprocess variables

The average measurements between bioreactor 1 and 2 for 18 process variables were computed. A Pearson's correlation coefficient was computed between all process variables over time. The Pearson's correlation matrix was visualized in order to identify strong associations between process variables to evaluate alterations in gene expression with metabolic transitions. In order

to identify correlation between lncRNA expression and bioprocess variables, a second correlation matrix was designed that included differentially expressed lncRNAs and 18 bioprocess variables. Only lncRNAs with a Pearson's correlation coefficient greater than or equal to ± 0.90 with a process variable were retained for analysis.

3.3 Results

3.3.1 CHO cell culture within industrial 10L stirred tank bioreactors

A cell line producing a recombinant IgG1 monoclonal antibody was seeded into two replicate 10L stirred tank bioreactor. Growth and cell viability was calculated every 24 hours over 14 days. The viable cell concentration (VCC) and viability of cells in replicate bioreactors was found to be highly comparable (**Figure 3.2**) between samples during all time points with a mean variance of 4.7×10^4 cells/mL the exponential phase was observed to begin at day 3 in both reactors and end at day 6 followed by an approximately 3-day stationary phase from day 7 to day 9. Cells showed the greatest variance between replicate bioreactors during the death phase which started on day 10 but were observed to recover slightly after day 13.

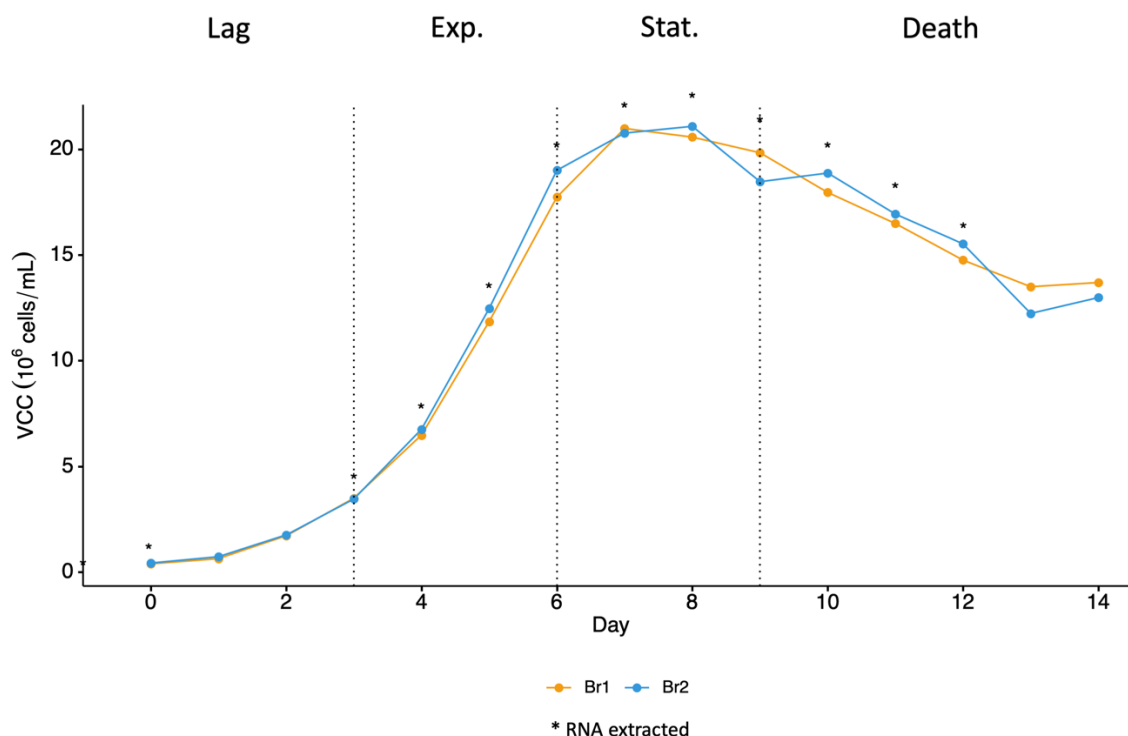


Figure 3.2 Viable cell concentrations over a 14-day period in two 10L bioreactors. Clear, lag, exponential, stationary, and death phases were observed in each bioreactor. Variance between replicate bioreactors was low in lag and exponential phases but started to diverge during the stationary phase. The viable cell concentrations were observed to recover slightly at day 13 with a final VCC of $\sim 13.3 \times 10^6$ cells/mL

Offline viability was measured using a VI-Cell XR (Beckman-Coulter) with % Viability holding constant above 95% until day 6. Viability in bioreactors was concordant with little variation between each replicate at each time point. The viability started to decrease in a linear fashion after day 6 which corresponds to the start of the stationary phase. The final viability at day 14 was approximately 70% in both bioreactors (**Figure 3.3**).

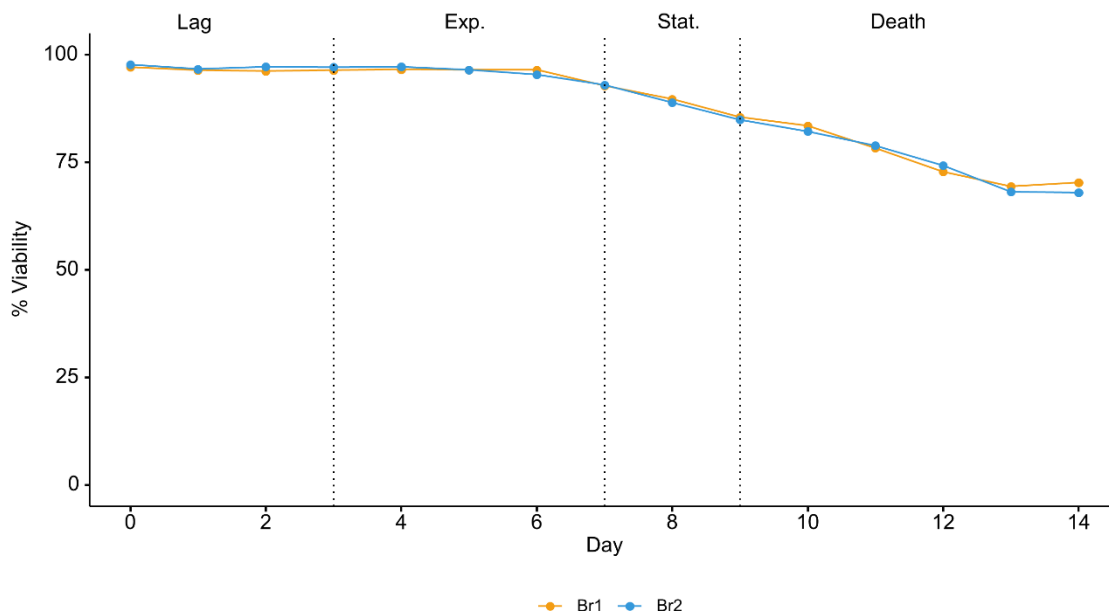


Figure 3.3 Percentage of viable cells over 14 day culture period in two replicate bioreactors. The viability was highly concordant between replicates. The dashed line in the plot depicts the time point where viability started to significantly decrease with a final viability of ~70% in both reactors.

3.3.2 RNA-Seq data QC and lncRNA classification

Reads were quality filtered and trimmed using *cutadapt* and *Trimmomatic* using the data pre-processing pipeline described in **Chapter 2**. Contamination from antisense rRNA probes was found in the raw read data and removed by aligning reads to the corresponding Illumina rRNA probe sequences. An average of 20% of reads were removed due insufficient quality or contamination (**Figure 3.4**). There was a higher degree of rRNA contamination in reads sequenced from bioreactor 1 compared to bioreactor 2. Reads derived from RNA on days 13 and 14 were deemed of insufficient quality and removed from all downstream analysis. Quality filtered

reads were aligned to the *ENSEMBL criGri1* CHOK1 genome with an average alignment rate of ~85% across all samples.

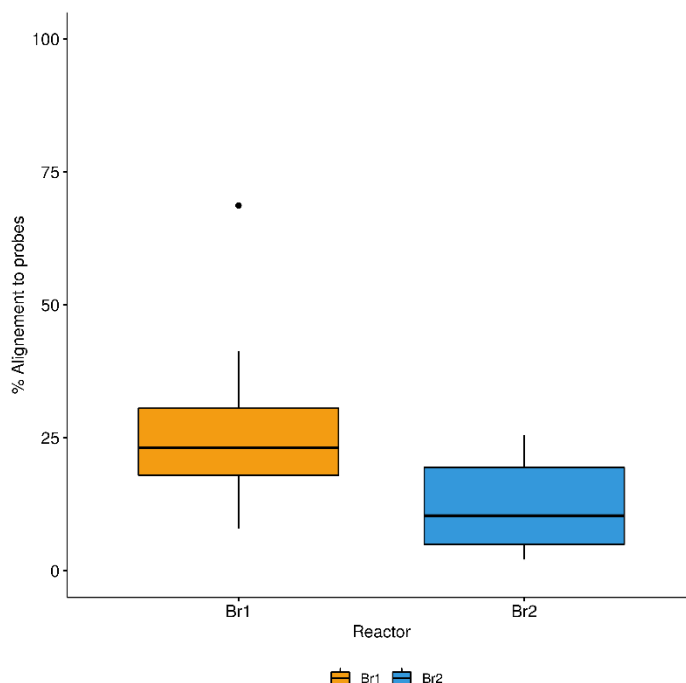


Figure 3.4 Boxplots comparing the distribution of percentage of reads aligning to the synthetic rRNA probe sequences. Bioreactor 2 had a slightly lower median alignment rate than bioreactor 1 and samples from day 13 and 14 were deemed insufficient for downstream analysis.

3.3.3 Reference guided assembly metrics

36,470 genes were assembled by *StringTie* (60,855 transcripts) using the *ENSEMBL criGri1* genome as a reference with 10,904 (29.74%) genes identified as novel because they lacked a corresponding annotation within the *ENSEMBL* transcriptome (X. Xu et al. 2011) as described in **Chapter 2**. The *gffcompare* tool was used to evaluate the quality of the assembly and found a 100% sensitivity and 81.2% precision at the exon level. Coding genes comprise 52% of all genes assembled while 14% of assembled genes belong to the lncRNA class of genes and the remaining 34% being miscellaneous small ncRNAs such as miRNAs, snoRNAs, and tRNAs (**Figure 3.5**). A total of 5,965 lncRNA genes (8,406 transcripts) were annotated. The percentage of each class of gene are concordant with the previous assembly presented in **Chapter 2**.

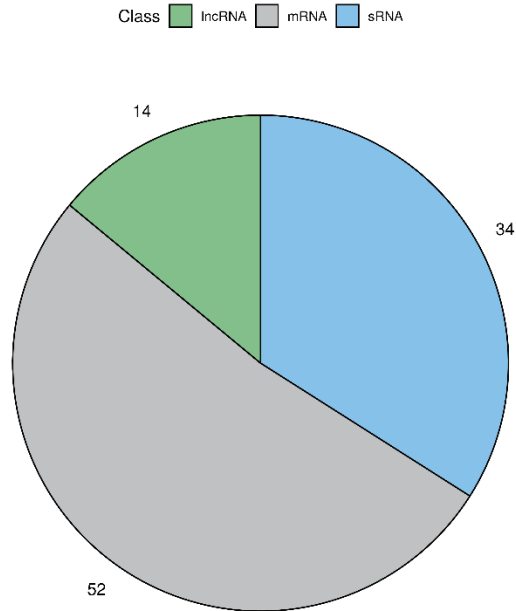


Figure 3.5 Classification of genes into categories based on proportion of total assembly. These three categories are protein coding, long non-coding, and miscellaneous small ncRNAs.

lncRNA transcripts were further classified into different groups using their proximity and orientation to nearby coding genes as described in **Chapter 2**. Intergenic transcripts composed the largest group followed by lncRNAs overlapping sense coding genes on the antisense strand. A transcript was further classified as upstream or downstream with respect to strand if the ncRNA was located within 100kb of a protein coding gene. A similar number of sense overlapping and divergent genes were identified when compared to the results of **Chapter 2**. The total number of lncRNA transcripts are summarized in **Figure 3.6**. The *gffcompare* tool was used to evaluate the difference between this assembly and the previous assembly from **Chapter 2** as well as the reference assembly.

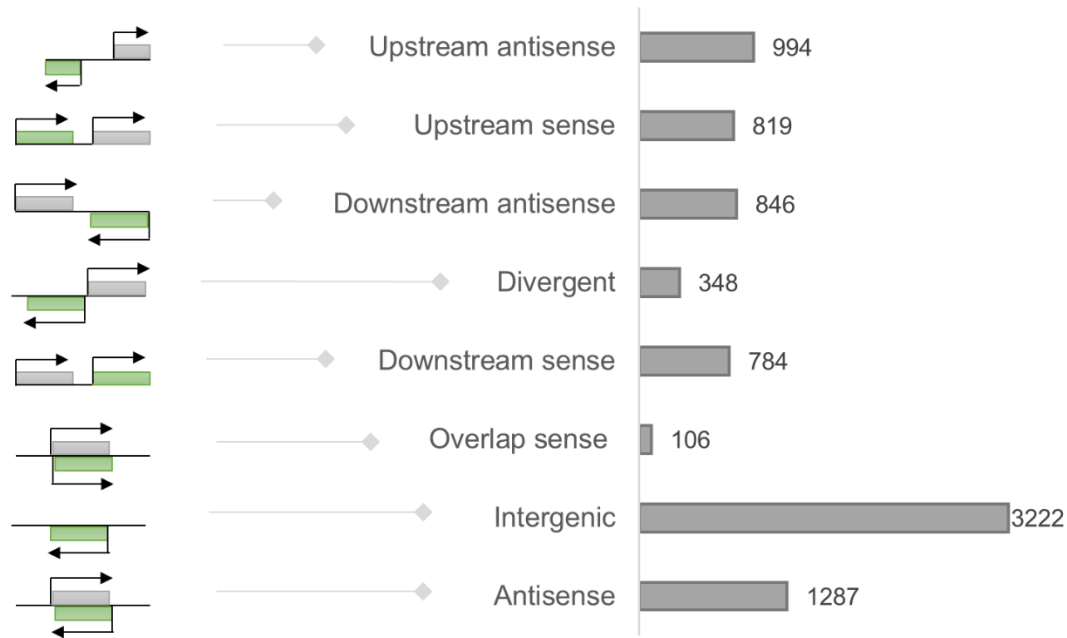


Figure 3.6 Number of transcripts for each genic and intergenic class of lncRNA transcripts. Long intergenic ncRNAs are the most abundant class followed by antisense RNAs overlapping protein coding genes. Sense and antisense loci that are found upstream and downstream as well as divergent transcripts that are transcribed in a “head to tail” manner from coding genes compose a smaller proportion of long non-coding RNAs.

3.3.4 Comparison of characteristics between lncRNA and mRNA loci

The transcript length, abundance, and % GC was compared between coding and non-coding transcripts. These characteristics agree closely with the assembly presented in **Chapter 2**. lncRNAs are shorter than mRNA transcripts (**Figure 3.7A**) and expressed at lower abundances with a much narrower range of expression (**Figure 3.7B**). lncRNAs also have a lower GC content per transcript compared to mRNA transcripts (**Figure 3.7C**). A Kolmogorov-Smirnov test found that all three distributions were significantly different ($p\text{-value} < 2.2 \times 10^{-16}$). A comparison of exon and isoform structure between coding and non-coding loci is presented in **Chapter 4**.

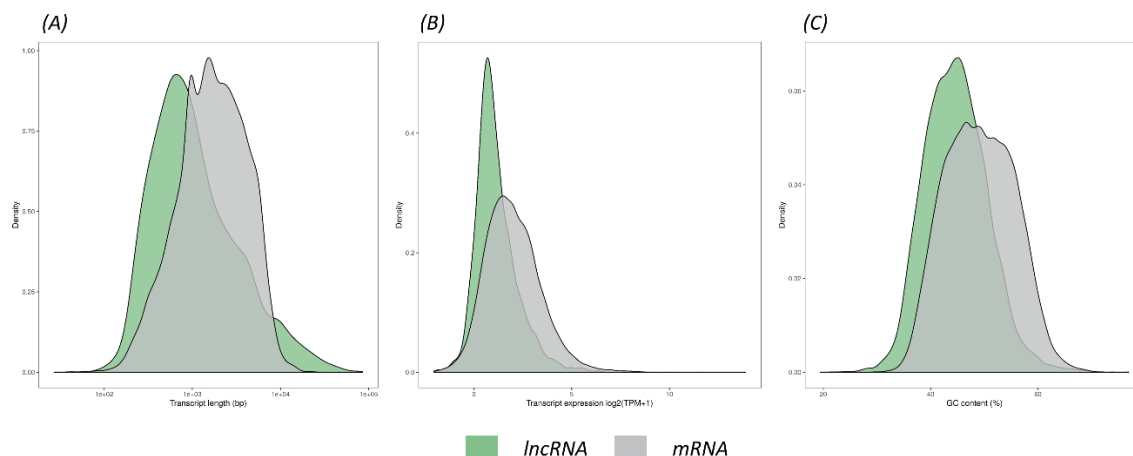


Figure 3.7 (A) The distribution of transcript lengths between mRNAs and lncRNAs. lncRNAs, were on average, shorter than mRNAs but also with a wider variation in transcript lengths. (B) A density plot depicting the $\log_2$ normalized TPM abundances between coding transcripts and lncRNAs with coding transcripts expressed at higher abundances with a wider distribution of TPM values compared to non-coding transcripts. (C) A density plot comparing % GC of coding and non-coding transcripts. lncRNA transcripts had an average GC content lower than mRNAs.

3.3.5 Differentially expressed mRNAs between exponential and stationary phases

A comparison of differentially expressed genes between exponential and stationary phases was performed using RNA-Seq data from days 3 to 9. A total 2,523 genes were found to exhibit differential expression between these two culture stages with 50% upregulated and 50% downregulated with **Table 3.2** summarizing the top ten most differentially expressed genes which were all observed to be significantly upregulated, albeit at comparatively low abundances.

Table 3.2 The top ten most differentially expressed genes between exponential and stationary phase.

Gene	baseMean	Log ₂ FC	Adj. P-value
<i>RAB38</i>	553.79	5.55	2.54×10^{-12}
<i>CDH12</i>	160.86	5.29	1.02×10^{-12}
<i>OLFR994</i>	721.24	5.05	2.06×10^{-11}
<i>FGF12</i>	471.14	5.02	2.18×10^{-11}
<i>SFTPD</i>	132.59	4.92	1.01×10^{-6}
<i>ARPP19</i>	136.54	4.83	3.72×10^{-10}
<i>KYNU</i>	869.56	4.82	6.43×10^{-6}
<i>SVEP1</i>	147.25	4.71	2.90×10^{-9}
<i>ADH1</i>	278.27	4.69	5.22×10^{-11}
<i>MIGA1</i>	346.63	4.66	1.61×10^{-17}

Supplementary Table 2.1 summarizes all significantly differentially expressed genes between exponential and stationary phases. The *BAX* gene was also significantly upregulated during the stationary phase with a fold-change of 1.55 (BH Adj. P-value: 2.32×10^{-4}). **Figure 3.8** illustrates the change in average TPM abundance over time for the *BAX* locus. *BAX* becomes upregulated after day 5 during the exponential phase and stays constant during the stationary and death phases.

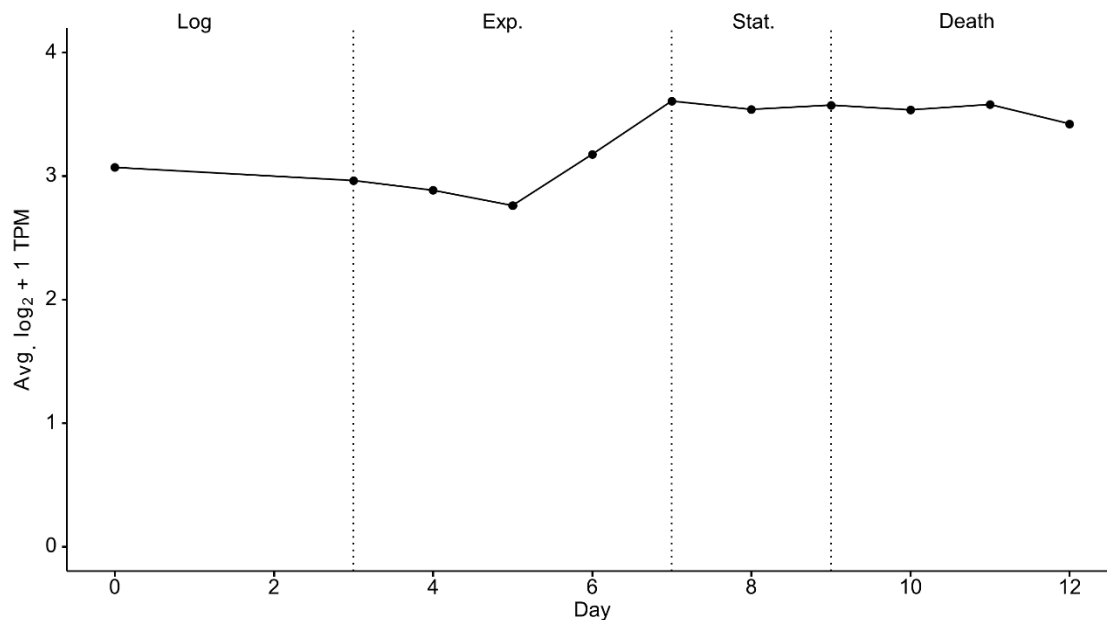


Figure 3.8 Average normalized expression of the *BAX* gene. The gene starts to be upregulated after day 5 and becomes significantly upregulated during the stationary phase. *BAX* does not have a change in average gene abundance during the death phase.

Gene ontology enrichment for biological processes was performed using DAVID on all differentially expressed mRNA encoding genes. A gene ontology was considered significantly enriched if there was a Benjamini-Hochberg adjusted P-value < 0.05 . Genes associated with cell cycle and mitosis were the most enriched GO terms in both cohorts of mRNAs. **Table 3.3** summarizes the top five most enriched terms in coding genes that are differentially expressed between exponential and stationary phases. *DAVID* analysis identified 36 significant enriched

biological processes with the full complement of significantly enriched terms found in **Supplementary Table 2.2.**

Table 3.3 The top 5 most enriched gene ontology terms as identified by *DAVID* which include processes involved in cell growth and cell cycle progression.

GO ID	GO Term	BH Adj. P-value
GO:0007049	Cell cycle	1.02×10^{-54}
GO:1903047	Mitotic cell cycle process	1.33×10^{-45}
GO:0006259	DNA metabolic process	1.45×10^{-32}
GO:0048285	Organelle fission	9.94×10^{-32}
GO:0007067	Mitotic nuclear division	1.39×10^{-31}

3.3.6 Differentially expressed mRNAs between stationary and death phases

A comparison of differentially expressed genes between stationary and death phases was performed using RNA-Seq data from days 7 to 12. A fewer number of protein coding genes were found to be differentially expressed compared to changes between the exponential and stationary phase with a total of 1,482 genes found to exhibit differential expression between these two culture stages. **Table 3.4** summarizes the top ten most differentially expressed genes which were all observed to be significantly upregulated, albeit at comparatively low abundances.

Table 3.4 The top ten most differentially expressed genes between stationary and death phase.

Gene	baseMean	Log ₂ FC	Adj. P-value
<i>PRXL2A</i>	133.44	5.37	5.47×10^{-15}
<i>AGT</i>	5250.08	5.03	1.72×10^{-8}
<i>GSTP3</i>	173.81	4.39	1.06×10^{-11}
<i>HIST1H3B</i>	1712.74	-4.27	1.13×10^{-13}
<i>HIST1H2BB</i>	550.90	-3.83	1.76×10^{-28}
<i>RANBP3L</i>	154.92	3.76	4.04×10^{-17}
<i>ENSCGRG00000019883</i>	328.27	-3.75	9.94×10^{-23}
<i>ENSCGRG00000018568</i>	336.48	-3.75	9.69×10^{-13}
<i>HP</i>	225.82	3.71	1.24×10^{-6}
<i>ENSCGRG00000025924</i>	668.33	3.63	1.98×10^{-5}

Supplementary Table 2.3 details all significantly differentially expressed genes between exponential and stationary phases. The *BCL2* gene was also significantly upregulated during the stationary phase with a fold-change of -1.83 (BH Adj. P-value: 6.53×10^{-14}). **Figure 3.9** illustrates the change in average TPM abundance over time for the *BCL2* locus. *BCL2* is downregulated after day 5 during the exponential phase and becomes upregulated after day 8 during the stationary phase where the

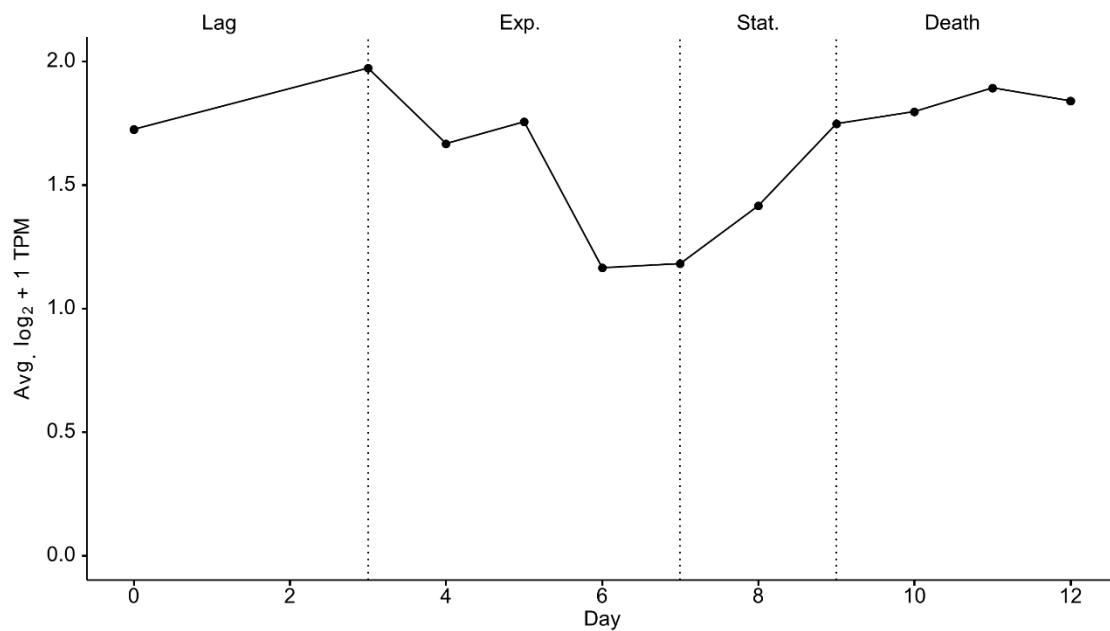


Figure 3.9 Average normalized expression of the *BCL2* gene. *BCL2* gene average normalized gene abundance begins to decrease after day 5 and begins to linearly increase during the stationary phase.

Gene ontology enrichment for biological processes was performed using DAVID on all differentially expressed mRNA encoding genes. Genes associated with cell cycle and mitosis were the most enriched GO terms in both cohorts of mRNAs. **Table 3.5** summarizes all 14 significantly enriched bioprocess terms

Table 3.5 Enriched gene ontology terms as identified by *DAVID* which include processes involved in cell growth and cell cycle progression.

GO ID	GO Term	FDR
GO:0007049	Cell cycle	3.80×10^{-28}
GO:0022403	Cell cycle phase	1.02×10^{-21}
GO:0022402	Cell cycle process	1.30×10^{-21}
GO:0000279	M phase	6.02×10^{-21}
GO:0000278	Mitotic cell cycle	4.30×10^{-16}
GO:0000087	M phase of mitotic cell cycle	3.58×10^{-15}
GO:0051301	Cell division	2.56×10^{-14}
GO:0048285	Organelle fission	3.82×10^{-14}
GO:0007067	Mitosis	4.15×10^{-14}
GO:0000280	Nuclear division	4.15×10^{-14}
GO:0006259	DNA metabolic process	6.00×10^{-13}
GO:0006260	DNA replication	2.94×10^{-11}
GO:0033554	Cellular response to stress	7.49×10^{-9}
GO:0006281	DNA repair	4.08×10^{-8}

3.3.7 Differentially expressed mRNAs over time

695 mRNA encoding genes were differentially expressed over time across exponential, stationary, and death phases. **Supplementary Table 2.4** contains all differentially expressed genes over time with **Table 3.6** summarizing the top 10 most differentially expressed genes over all three stages. The *Angiotensinogen* gene had one of the largest change in expression between stationary and death phase but becomes 32 fold more abundant during the death phase (**Figure 3.10**).

Table 3.6 Top ten most significantly differentially expressed genes over time between exponential, stationary, and death phases.

Gene	EXP-STAT baseMean	EXP-STAT Log2FC	EXP-STAT Adj. P-Value	STAT-DEATH baseMean	STAT-DEATH Log2FC	STAT-DEATH Adj. P-Value
<i>SFTPD</i>	132.59	4.92	1.01×10^{-6}	227.81	-0.71	5.06×10^{-3}
<i>ARPP19</i>	136.54	4.83	3.72×10^{-10}	205.11	-1.30	7.53×10^{-13}
<i>KYNU</i>	869.56	4.82	6.43×10^{-6}	1274.55	-1.41	9.27×10^{-5}
<i>ADH1</i>	278.27	4.69	5.22×10^{-11}	1687.04	2.25	6.87×10^{-5}
<i>TNFSF11</i>	495.26	4.10	1.25×10^{-5}	650.33	-1.90	1.27×10^{-15}
<i>ABCA13</i>	1816.75	4.03	9.32×10^{-9}	5205.63	0.84	8.08×10^{-3}
<i>SULT1D1</i>	154.14	3.85	7.67×10^{-22}	186.90	-2.45	6.52×10^{-16}
<i>EFCAB5</i>	183.15	3.69	1.36×10^{-7}	241.70	-1.74	1.92×10^{-13}
<i>AGT</i>	156.27	3.62	1.95×10^{-6}	5250.08	5.03	1.72×10^{-8}
<i>CTLA2A</i>	315.25	3.51	1.33×10^{-19}	390.65	-2.06	2.71×10^{-19}

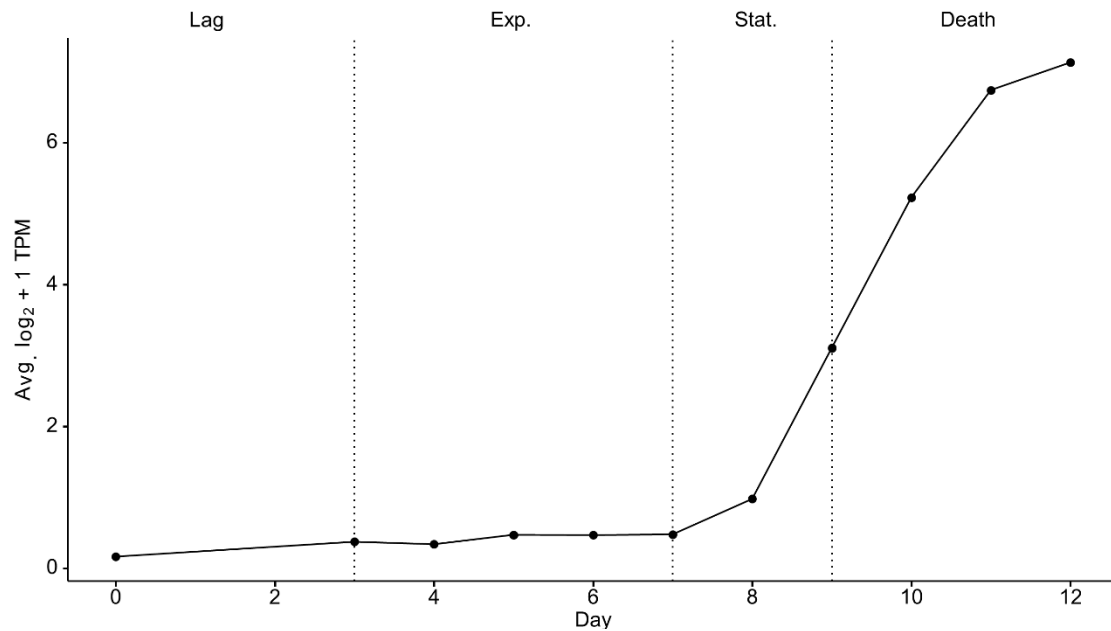


Figure 3.10 Average normalized abundance of the *Angiotensinogen* gene. *AGT* expression was quiescent until day 7 when gene expression started to significantly increase into the death phase.

Gene ontology enrichment for biological processes was performed using DAVID on all differentially expressed mRNA encoding genes. Genes associated with cell cycle and mitosis were the most enriched GO terms in both cohorts of mRNAs. **Table 3.7** summarizes the top 5 most enriched terms with the full cohort found within **Supplementary Table 2.5**.

Table 3.7 The top 5 most enriched GO biological process terms identified by DAVID for genes that were differentially expressed over all three culture phases.

GO ID	GO Term	BH Adj. P-value
GO:0007049	Cell cycle	1.99×10^{-44}
GO:0000279	M phase	1.82×10^{-33}
GO:0022403	Cell cycle phase	2.44×10^{-32}
GO:0022402	Cell cycle process	7.72×10^{-32}
GO:0000087	M phase of mitotic cell cycle	5.36×10^{-26}

3.3.8 Differentially expressed lncRNAs during exponential and stationary phases

A total of 861 lncRNA genes were significantly differentially expressed only between the exponential phase to stationary phase. There were more significantly upregulated genes (576 genes) than downregulated genes (285 genes) during this transition. 190 significantly differentially expressed CHO lncRNA genes were evolutionarily conserved with the human GENCODE annotation while only 6 were conserved with the mouse GENCODE annotation. Differentially expressed genes were classified by localization to the genome with intergenic and antisense overlapping lncRNAs representing the two most abundant classes of non-coding genes (**Figure 3.11**).

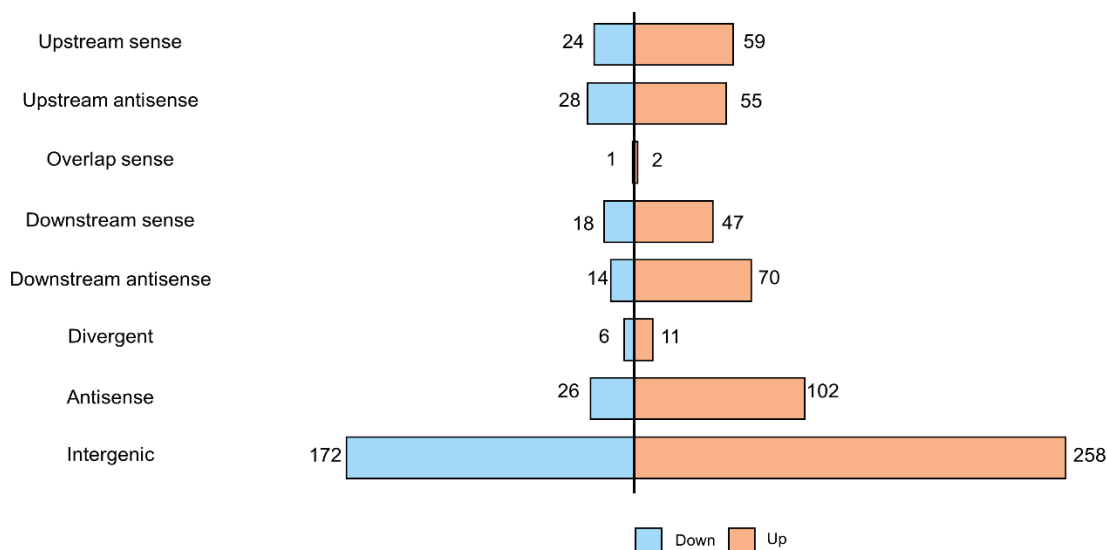


Figure 3.11 A leaf plot of up and downregulated lncRNAs between exponential and stationary phases. There were 32 genes that were assigned to multiple classifications and are represented within the leaf plot.

The evolutionarily conserved competing endogenous lncRNA *LINC00536* was one of the most abundant, differentially expressed genes in this group (**Figure 3.12**) and was significantly upregulated between exponential and stationary phases (fold-change: 23.03, BH Adj. P-value: 2.23×10^{-10}).

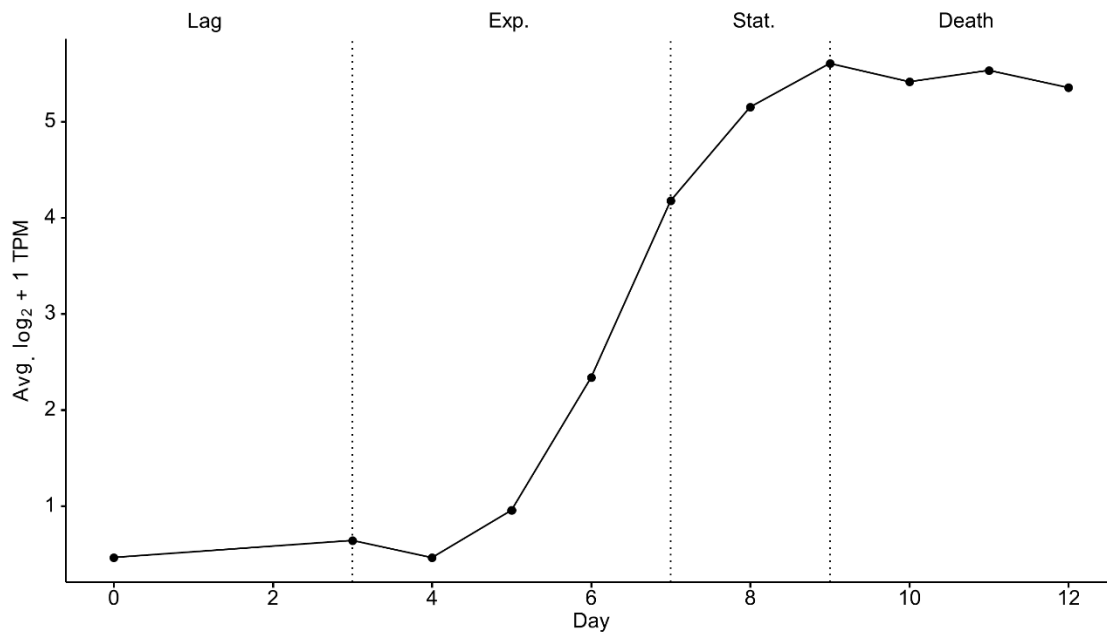


Figure 3.12 Average normalized abundance of the *LINC00536* non-coding gene. *LINC00536* gene expression was observed to increase from day 5.

Alternatively, a novel multi-exonic, *MSTRG.11806*, gene found upstream of the *IER3* coding gene is the most abundant, downregulated gene between exponential and stationary phase (fold-change: -3.30, BH Adj. P-value 6.01×10^{-14} , **Figure 3.13**). This gene was downregulated during the stationary phase but abundances started to recover on day 11. **Supplementary Table 2.6** contains the full cohort of differentially expressed lncRNAs and the respective classification.

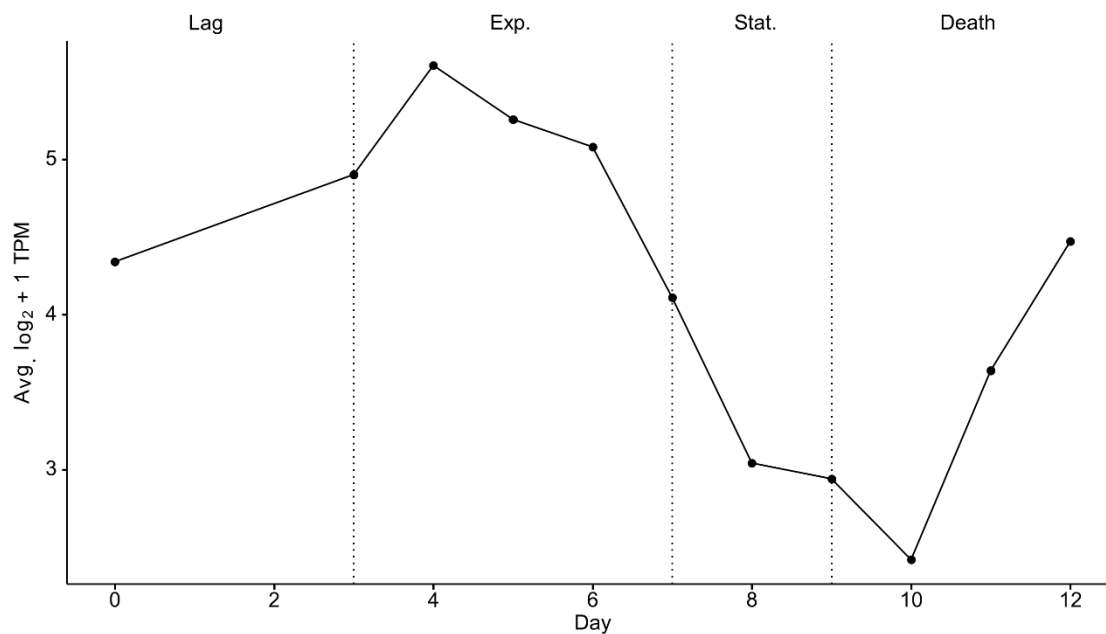


Figure 3.13 Average normalized abundance of the novel *MSTRG.11806* ncRNA gene. Gene expression was at a maximum at day 4 in the exponential phase where abundances steadily decrease until day 10.

A total of 19 sense/antisense genes exhibited differential expression on both strands between the exponential and stationary phases. 10 loci had correlated expression on the sense and antisense strand while 9 genes had anti-correlated expression. The *SCFD1* coding gene is the most abundant, anti-correlated locus with strong upregulation of the antisense gene and downregulation of the sense coding gene (**Table 3.8, Figure 3.14**).

Table 3.8 A table summarizing-fold-change and BH adjusted P-values between differentially expressed sense and antisense gene pairs. There were 10 genes with correlated expression and 9 genes with anti-correlated expression.

Gene	Sense L2FC	Sense P Adj	Antisense L2FC	Antisense P Adj	Correlation
<i>FILIP1L</i>	-1.91	1.75×10^{-19}	-1.57	3.58×10^{-9}	Correlated
<i>H2-T23</i>	0.97	2.06×10^{-9}	0.65	1.41×10^{-7}	Correlated
<i>GMPR</i>	2.34	3.48×10^{-16}	1.01	1.09×10^{-5}	Correlated
<i>ENSCGRG000000015813</i>	0.94	4.78×10^{-6}	1.29	6.40×10^{-9}	Correlated
<i>SVEP1</i>	4.71	2.90×10^{-9}	2.73	1.18×10^{-6}	Correlated
<i>CENPN</i>	-1.53	4.58×10^{-20}	1.33	1.62×10^{-10}	Anti-correlated
<i>EPHX1</i>	1.37	8.46×10^{-12}	-1.88	8.02×10^{-7}	Anti-correlated
<i>RPS6KA5</i>	-0.70	5.38×10^{-5}	1.71	5.60×10^{-21}	Anti-correlated
<i>CRACR2B</i>	1.04	8.04×10^{-10}	0.77	3.19×10^{-5}	Correlated
<i>ENSCGRG00000000226</i>	0.62	4.71×10^{-5}	-0.97	4.47×10^{-4}	Anti-correlated
<i>CDT1</i>	-1.85	1.33×10^{-13}	0.74	1.15×10^{-3}	Anti-correlated
<i>CALCR</i>	2.05	1.13×10^{-8}	-0.83	6.57×10^{-5}	Anti-correlated
<i>SCFD1</i>	-0.67	1.51×10^{-20}	2.74	2.97×10^{-8}	Anti-correlated
<i>STX11</i>	0.93	3.99×10^{-6}	0.67	7.49×10^{-3}	Correlated
<i>TFEC</i>	1.60	8.90×10^{-18}	0.81	4.40×10^{-5}	Correlated
<i>CTNNA3</i>	3.52	4.75×10^{-27}	2.71	2.31×10^{-18}	Anti-correlated
<i>GPRC5A</i>	-0.86	5.06×10^{-5}	-2.16	4.86×10^{-11}	Correlated
<i>LRP8</i>	-1.08	3.91×10^{-5}	-0.63	7.49×10^{-4}	Correlated

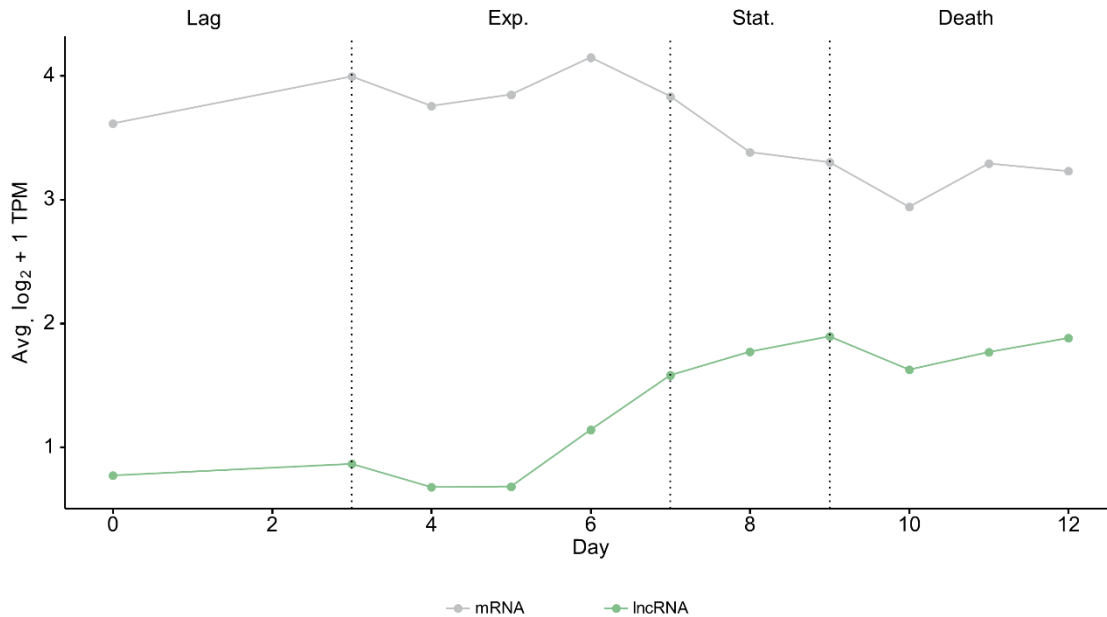


Figure 3.14 Sense and antisense normalized abundances of the *SCFD1* coding gene and the ncRNA partner.

3.3.9 Differentially expressed lncRNAs during death phase

491 lncRNAs were found to be significantly differentially expressed only during the death phase (day 10, 11, and 12). There were an equivalent number of upregulated (244 genes) and downregulated (247) genes between the two stages. 76 significantly differentially expressed genes were conserved between CHO and the human GENCODE annotation whereas only 2 were conserved between CHO and the mouse annotation. Differentially expressed genes were classified by localization to the genome with intergenic and antisense overlapping lncRNAs representing the two most abundant classes of non-coding genes (**Figure 3.15**).

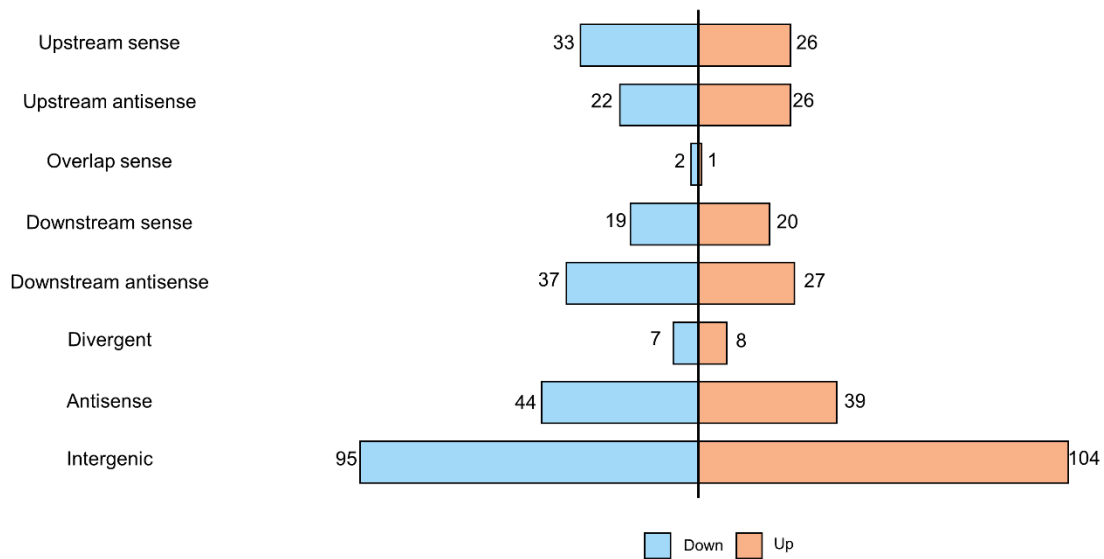


Figure 3.15 A leaf plot of up and downregulated lncRNAs between stationary and death phases. There were 19 genes that were assigned to multiple classifications and are represented within the leaf plot.

Supplementary Table 2.7 contains all differentially expressed genes between these two phases as well as the corresponding lncRNA classifications. A novel antisense gene, *MSTRG.19771*, found within the *TMC1* gene (fold-change: -4.02, BH Adj. P-value: 6.62×10^{-12} , **Figure 3.16**) as well as the evolutionarily conserved *NEAT1* lncRNA (fold-change: 2.55, BH Adj. P-value: 0.038, **Figure 3.17**) had variable expression over time during the exponential phase but were significantly differentially expressed from the start of the death phase.

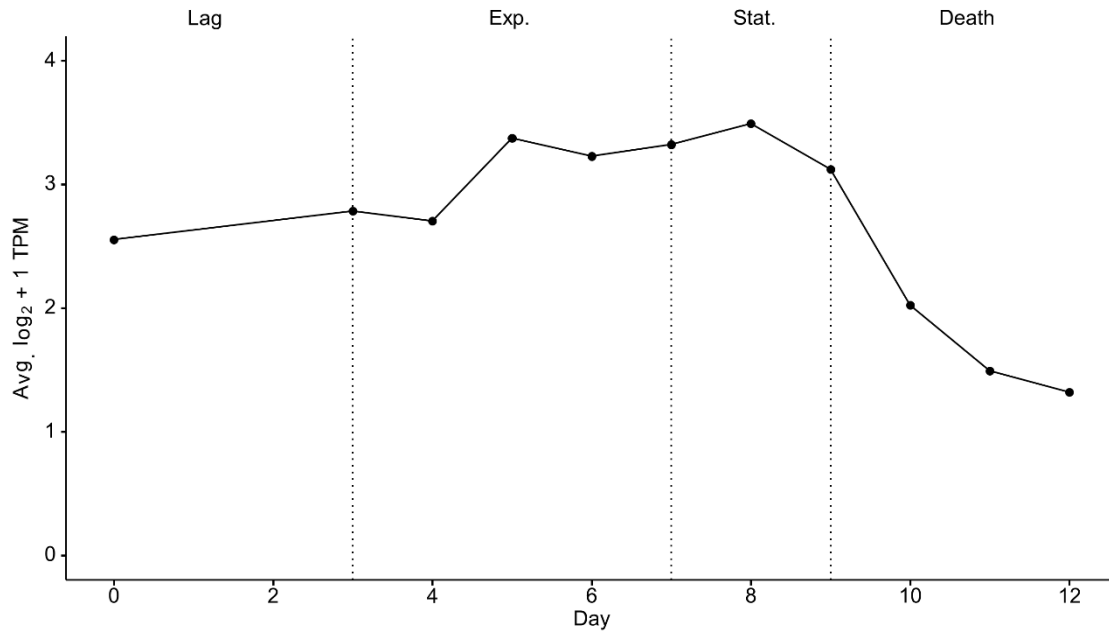


Figure 3.16 Average normalized abundances over time for the *TMC1-AS* ncRNA. *TMC1-AS* is modestly differentially expressed between individual days but is significantly downregulated after day 9 at the onset of apoptosis.

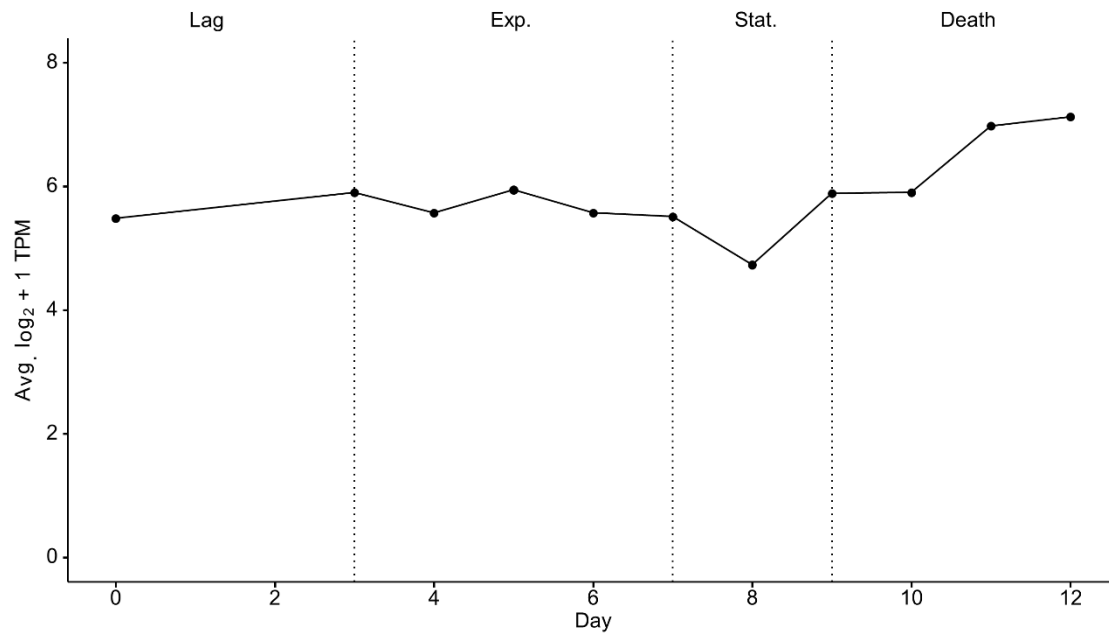


Figure 3.17 Average normalized abundances over time for the conserved *NEAT1* gene. *NEAT1* is expressed at very high abundances at a constant level until day 10 when the ncRNA becomes significantly upregulated from day 11 and 12.

Five gene loci were found to have differential expression on both the coding sense strand and the antisense non-coding strand. Three sense/antisense pairs had correlated expression and two

genes, anti-correlated expression with *SLFN12* being the most abundant sense/antisense pair (Table 3.9)

Table 3.9 Five sense/antisense gene pairs that are strongly differentially expressed during the between the stationary and death phase.

mRNA	Sense L2FC	Sense P- Adj	Antisense L2FC	Antisense P -Adj	Correlation
<i>IYD</i>	1.05	3.67×10^{-2}	0.76	4.30×10^{-2}	Correlated
<i>SLFN12</i>	1.03	1.18×10^{-10}	-1.49	2.58×10^{-13}	Anti-correlated
<i>FEN1</i>	-0.60	4.66×10^{-9}	-1.08	9.77×10^{-16}	Correlated
<i>MGAT4A</i>	-0.65	1.38×10^{-7}	-0.64	8.40×10^{-3}	Correlated
<i>GINS4</i>	1.22	5.32×10^{-10}	-1.51	3.46×10^{-22}	Anti-correlated

3.3.10 Modulation of sense/antisense transcription over time

Three genes exhibited differential expression on the sense and antisense strand between the transition from exponential phase to stationary phase and the transition from stationary phase to death phase: *PLCB4*, *NAV2*, and *FAXDC2* (Table 3.10). The *Neuron navigator 2* (*NAV2*) antisense gene is expressed at similar abundances to its coding counterpart until the 5th day when expression of the sense coding gene becomes significantly suppressed as the antisense ncRNA becomes highly upregulated, reaching its maximum abundance at day 9 (Figure 3.18). Figure 3.19 shows the average sense and antisense coverage of the *NAV2* locus between day 5 and day 10. *PLBC4* and *FAXDC2* are expressed at comparatively lower abundances on both strands across all time points.

Table 3.10 Three differentially expressed sense/antisense gene pairs that exhibit changes in expression over all phases of the cell culture. The *NAV2* gene is the most abundant of this group and has anti-correlated expression during the stationary phase.

Exponential to Stationary Phase					Stationary to Death Phase			
Gene	Sense L2FC	Sense P- Adj	Antisense L2FC	Antisense P- Adj	Sense L2FC	Sense P- Adj	Antisense L2FC	Antisense P- Adj
<i>PLCB4</i>	-1.08	8.29×10^{-5}	5.24	9.47×10^{-4}	-0.90	6.15×10^{-5}	-1.88	3.53×10^{-6}
<i>NAV2</i>	0.75	3.63×10^{-3}	2.17	1.08×10^{-20}	-0.88	6.84×10^{-4}	-0.74	6.13×10^{-3}
<i>FAXDC2</i>	-2.12	4.78×10^{-15}	-0.77	1.66×10^{-8}	0.60	5.57×10^{-4}	-0.63	8.14×10^{-6}

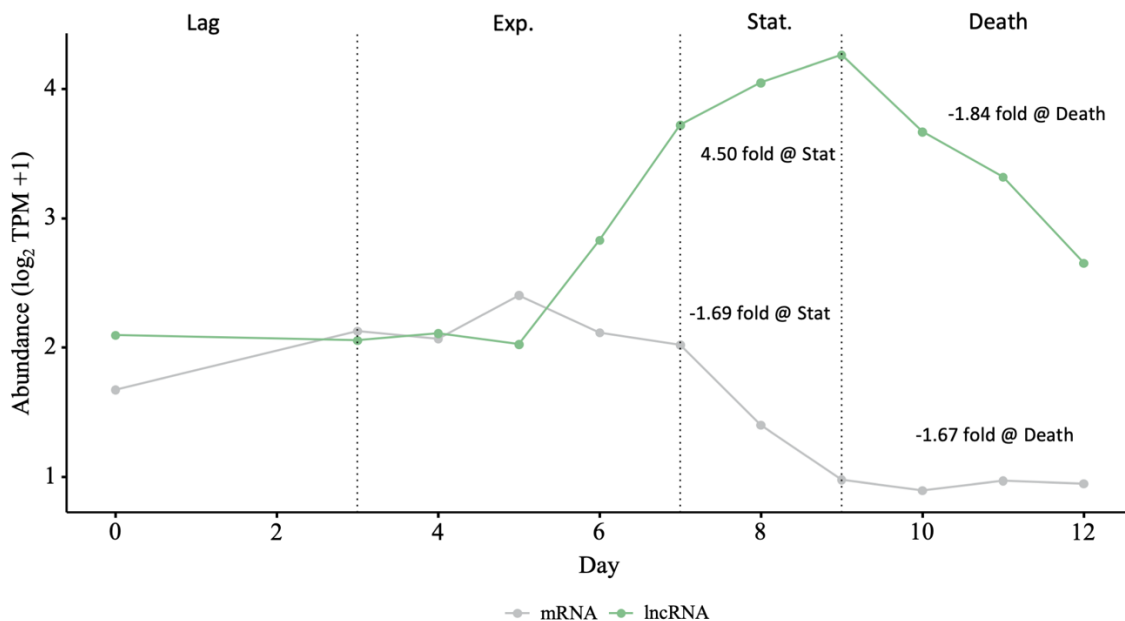


Figure 3.18 Average normalized abundance of the sense *NAV2* gene and the antisense lncRNA gene over time. Both genes are expressed at similar abundances until day 5 when the antisense ncRNA becomes significantly upregulated and becomes dominant during the stationary phase.

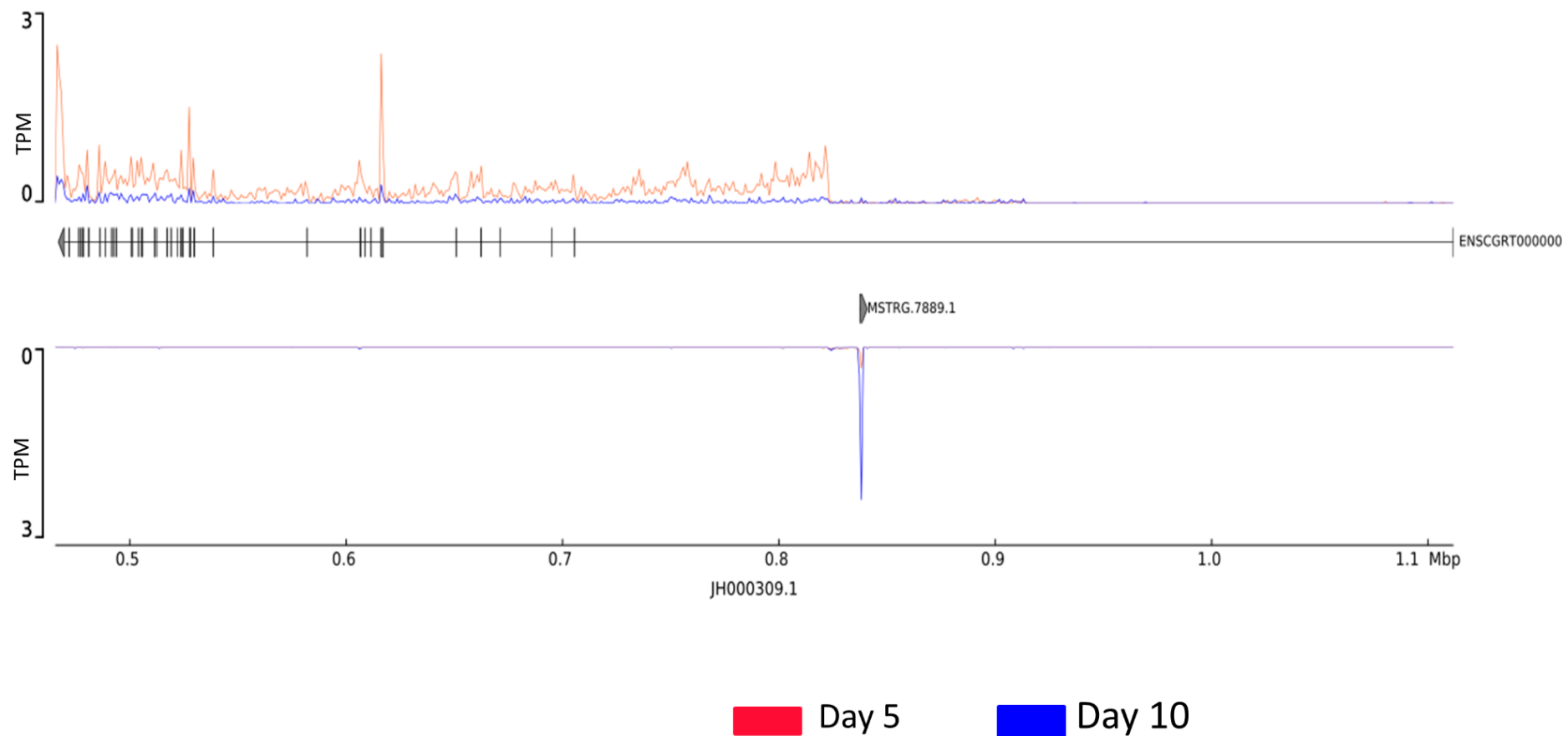


Figure 3.19 A plot of sense and antisense coverage of the *NAV2* gene composed of transcript *ENSCGRT00000018790* and the mono-exonic antisense *NAV2* gene which contains a single transcript (*MSTRG.7889.1*) on day 5 and day 10. The sense gene is dominant during day 5 but becomes depleted during the death phase on day 10.

3.3.11 Metabolite profiling in 10L stirred tank bioreactors

Metabolite concentrations in extracellular media were calculated daily using a Cedex HT Bioanalyzer. Glucose, lactate, glutamine, glutamate, and ammonia were among the metabolites profiled over time. A bolus of nutrients including glucose was supplemented into the media from day 4 to day 13. Glucose concentration steadily decreased following initial inoculation until day 4 when fed-batch supplementation began (**Figure 3.20**).

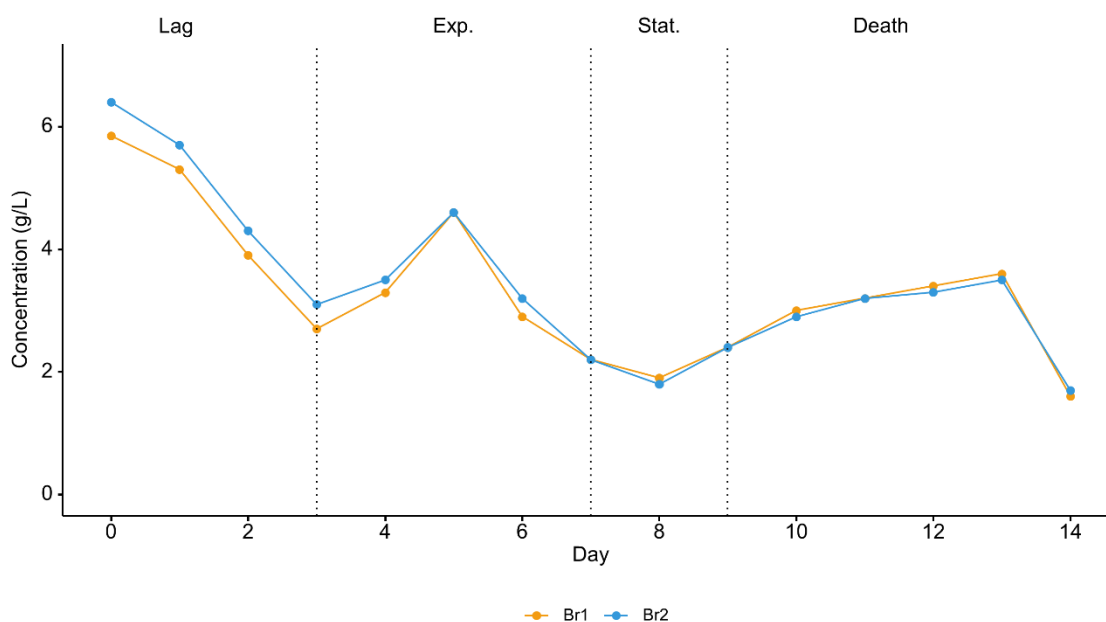


Figure 3.20 Glucose concentrations were quantified before and after feed daily. This plot only displays concentrations prior to feed for reasons of clarity and in order to visualize glucose consumption over time

Lactate concentrations increased inversely to glucose consumption from day 0 until day 5-6 where cells began to consume lactate. Lactate concentration steadily decrease from day 5 until completely depleted from the media at day 8 (**Figure 3.21**).

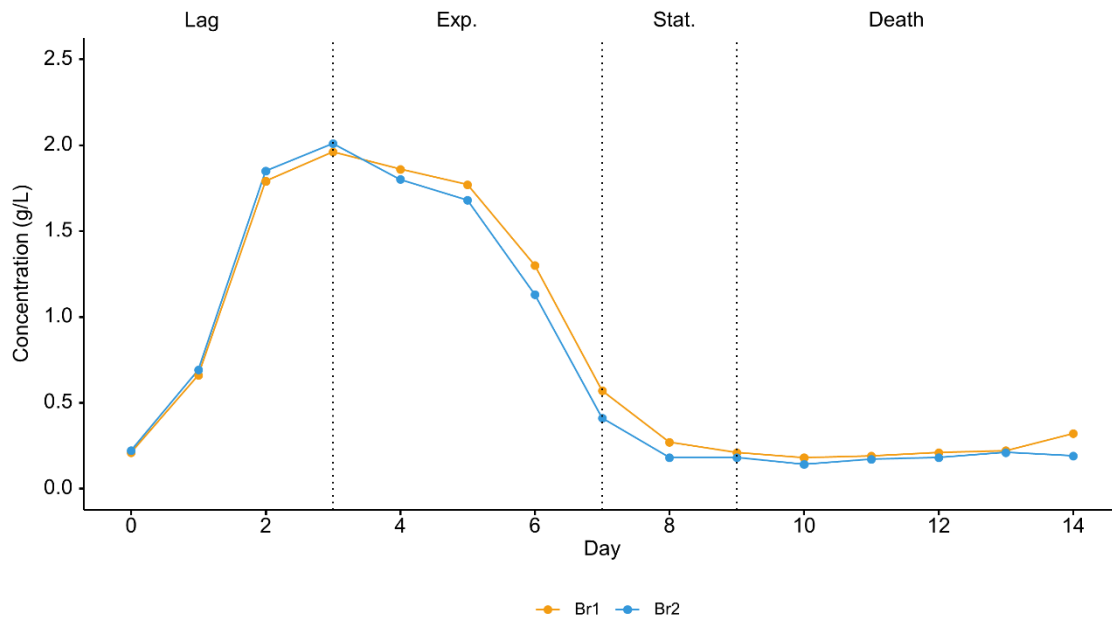


Figure 3.21 Lactate concentrations were profiled daily over a period. Concentrations were the highest at day 3 just prior to media supplementation where consumption was observed after day 5.

Glutamine was only a component of the basal media prior whereas glutamate was a component of the fed-batch supplemental media. Glutamine concentrations decreased linearly until day 4 when it was depleted from the media. Following media supplementation, glutamine concentration began to increase to approximately 1.5 mmol/L (**Figure 3.22**).

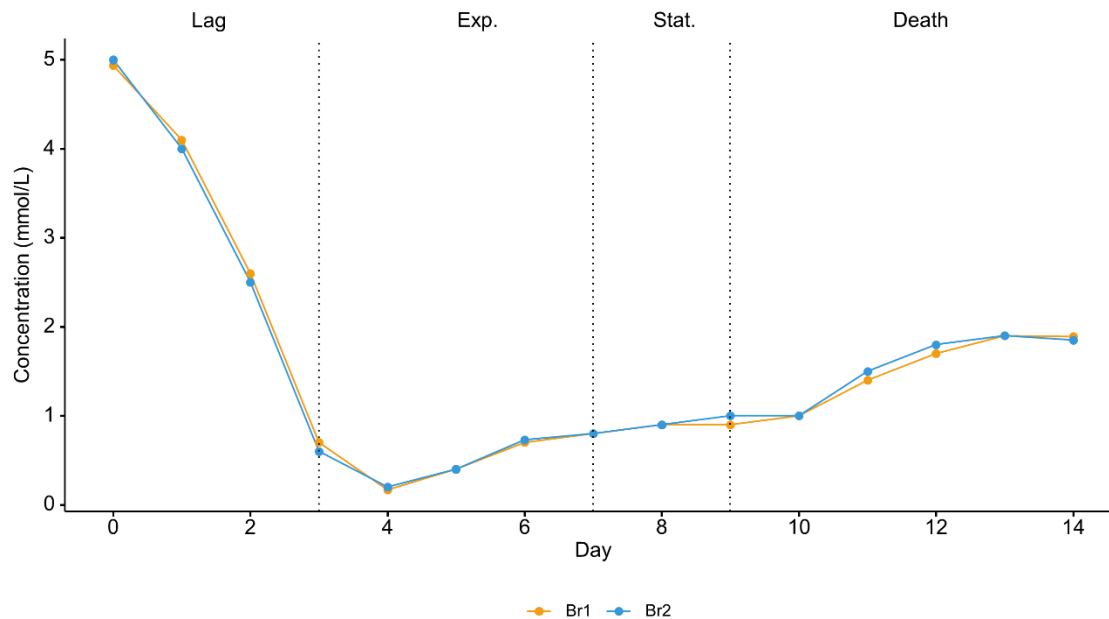


Figure 3.22 Glutamine concentrations over 14 day culture period. Glutamine was only a component of the basal media and was not supplemented during the feeding. Glutamine is completely depleted after day 4 but concentrations begin to increase during the exponential, stationary and death phases.

Glutamate concentrations were at their highest during the days 3-6 until completely depleted during the stationary phase. Glutamate concentrations increased linearly after day 10 until reaching a final concentration of ~5 mmol/L (**Figure 3.23**). The ammonium ion concentrations correlated with the glutamate concentrations and followed a similar pattern of depletion during the stationary phase followed by steady increase during the death phase where high ammonium ion concentrations were found (**Figure 3.24**).

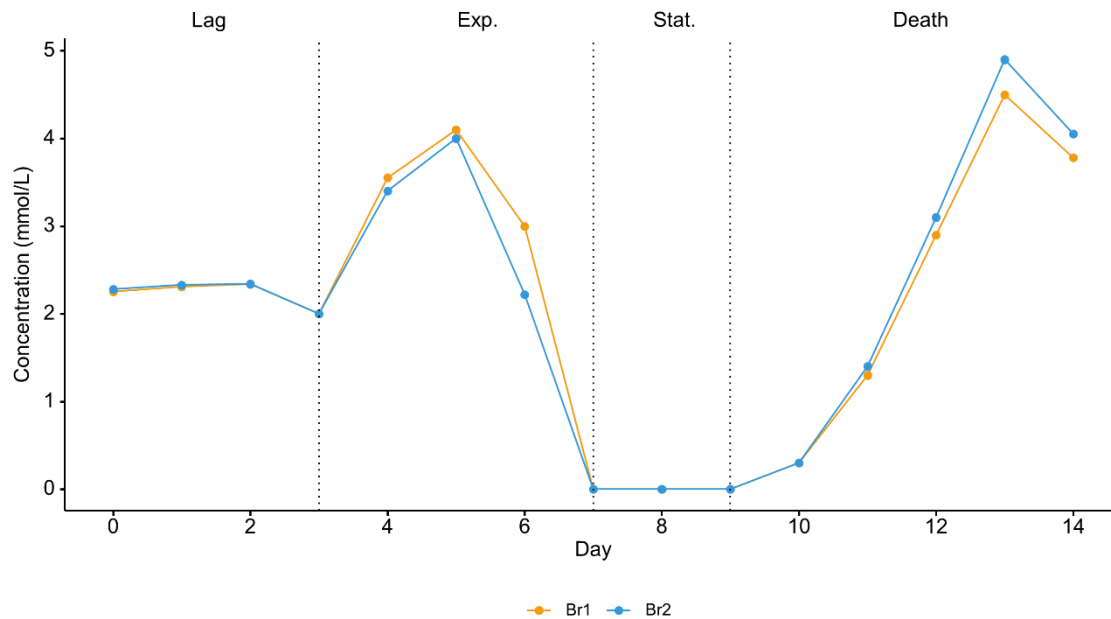


Figure 3.23 Glutamate concentrations in mmol/L over time. Glutamate was a component of the feed media and was routinely added during the fed-batch process. Glutamate is completely depleted during the stationary phase but concentrations begin to increase during the death phase.

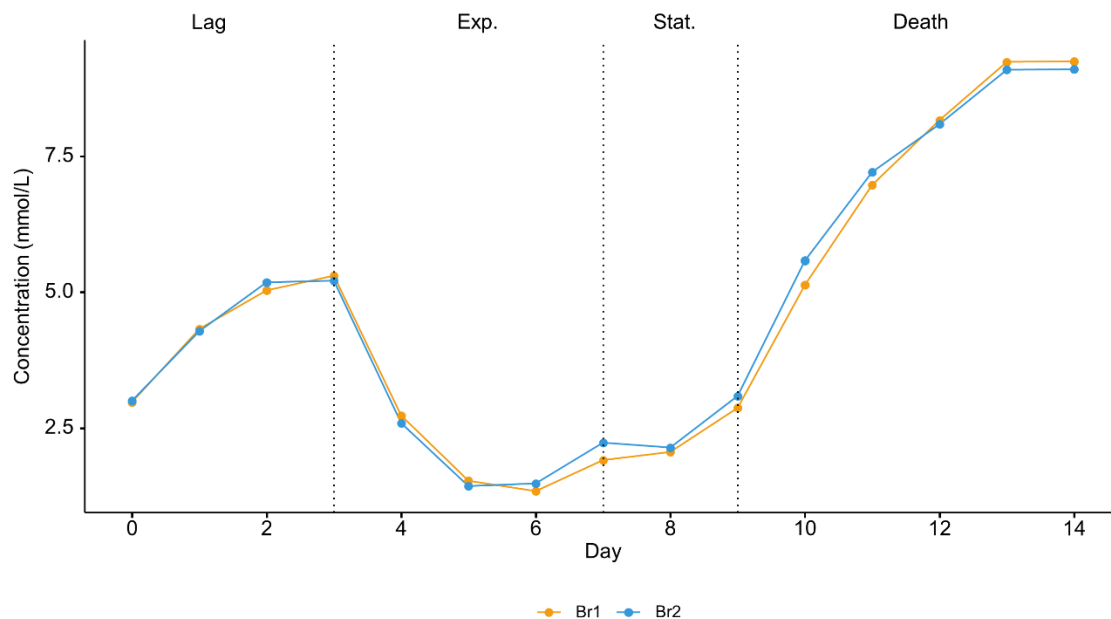


Figure 3.24 Ammonia concentrations in mmol/L over time. Ammonia concentrations began to increase during the lag phase but were reduced during the exponential and stationary phases. Ammonia concentrations increased linearly during the death phase with the highest concentration on days 13 and 14.

3.3.12 Correlation of bioprocess parameters and metabolite concentration to differential expression

Pearson's correlation coefficients were computed between differentially expressed coding and long non-coding gene normalized gene abundances and five metabolite concentrations: glucose, lactate, glutamine, glutamate, and ammonia as well as the viable cell concentration and % viability. A gene was considered highly correlated to a metabolite if the Pearson's correlation was greater than or equal to ± 0.9 . 1,259 protein coding genes were significantly correlated with the two process parameters or five metabolite concentrations whereas 337 lncRNA genes were correlated. **Supplementary Tables 2.8 and 2.9** summarizes all significantly correlated genes with metabolite concentration or change in VCC or % viability. There were no significantly differentially expressed coding or non-coding genes that correlated with glucose or glutamine. Viability had the largest number of correlations between process parameter and changes in gene abundance for mRNAs. The coding gene *RBM3* correlated strongly with % viability with a Pearson's coefficient of 0.98 (**Figure 3.25**).

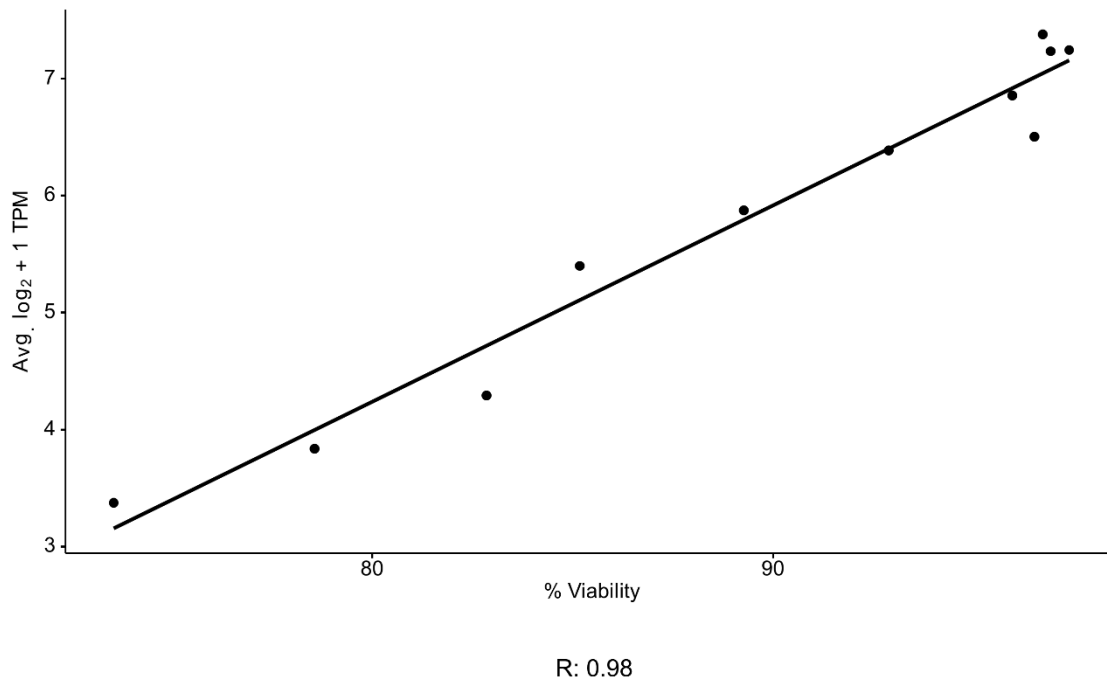


Figure 3.25 Regression plot comparing the % viability to average normalized abundance of the *RBM3* gene over time. There was a strong correlation between the two as *RBM3* gene abundance decreased proportionally to decrease in % viability.

RNA abundances were also significantly correlated with lactate abundance with 57 genes having a Pearson's correlation greater than or equal to ± 0.90 . The *Cyclooxygenase 2* (*COX2*) gene was the most negatively correlated mRNA with a Pearson's correlation coefficient of -0.92 (**Figure 3.26**). *COX2* was highly expressed with an average normalized TPM abundance of ~9,500 which is an order of magnitude higher than the average coding gene expression and is upregulated during the stationary phase with a fold-change of 1.78 (BH Adj. P-value: 1.50×10^{-7}).

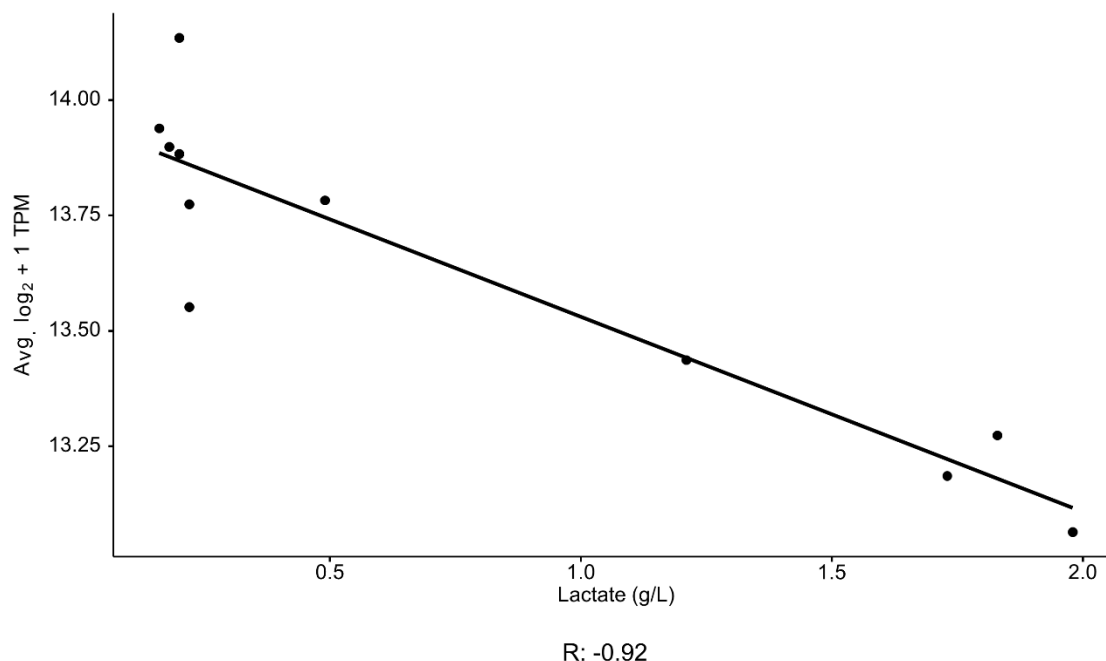


Figure 3.26 Regression plot comparing lactate concentration to average normalized abundance of the COX2 gene. Gene expression was observed to be upregulated during the stationary phase when lactate is completely depleted from the culture media.

There were four genes that correlated to more than one process with all four being negatively correlated with viability and positively correlated with ammonia. **Table 3.11** contains the Pearson's correlation coefficients between gene abundance and % viability or ammonia. *LPCAT3* was the most abundant of this cohort with an average TPM of 83 and was upregulated during the death phase with a fold-change of 2.07 (BH Adj. P-value: 1.11×10^{-11}).

Table 3.11 Pearson's correlation coefficients of four genes that were significantly correlated with multiple variables. All four genes negatively correlated with % viability and positively correlated with ammonia.

Gene	% Viability	Ammonia
<i>ASP</i>	-0.92	0.92
<i>IL13RA2</i>	-0.92	0.93
<i>LPCAT3</i>	-0.93	0.93
<i>SHISA4</i>	-0.91	0.94

Viability had the largest number of correlations between process parameter and changes in gene abundance for lncRNAs with 223 long non-coding RNA genes were highly correlated with change

in viability. The novel intergenic gene *MSTRG.2854* was one of the non-coding genes that strongly correlated with % viability ($R: -0.94$) and was highly upregulated during the death phase with a fold-change of 65.01 (BH Adj. P-value: 4.16×10^{-11}). This gene remained quiescent during exponential phase until day 7 during the stationary phase when gene abundance was observed to increase (**Figure 3.27**). **Figure 3.28** contains the correlation between % viability and average normalized gene abundances for *MSTRG.2854*.

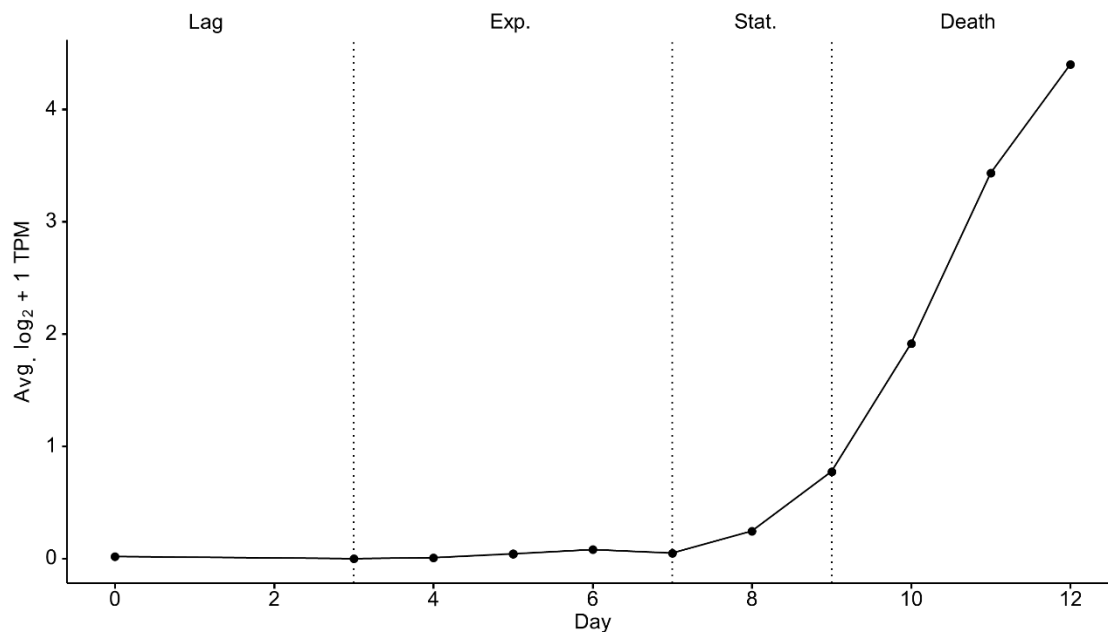


Figure 3.27 Average normalized gene abundances over time for the *MSTRG.2854* intergenic ncRNA. Gene expression was quiescent during the exponential phase but increased during the stationary phase and death phases as viability dropped below 90%.

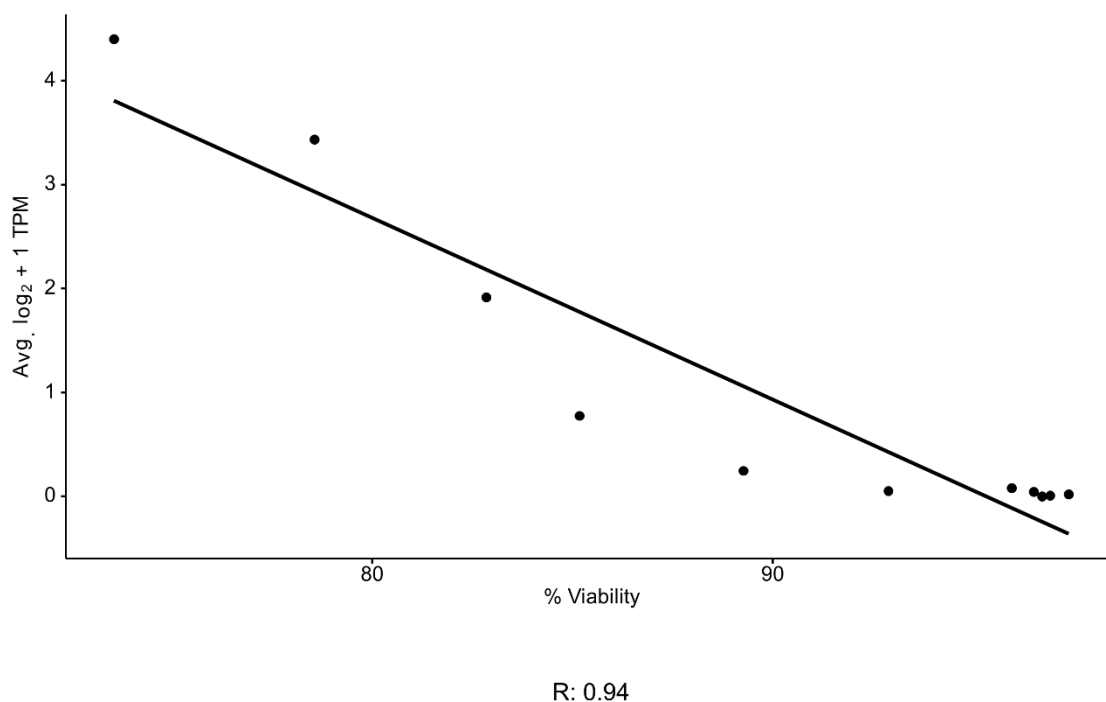


Figure 3.28 Regression plot describing the correlation between % of viable cells and the average normalized gene abundances of the novel *MSTRG.2854* gene.

Long non-coding RNAs also strongly correlated with metabolites but were not observed to correlate with glucose or glutamine concentrations. A total of 30 long non-coding RNAs strongly correlated with lactate, glutamate, and NH_4^+ (lactate:8, glutamate: 17, ammonia: 5). **Figure 3.29** contains the regression plots between metabolite concentration and normalized abundance of the top three most abundant ncRNA genes for lactate, ammonia, and glutamate. *MSTRG.16951* is a mono-exonic transcript found downstream of the *LRN3* gene on the antisense strand and was downregulated 2.27 fold during the stationary phase (BH Adj. P-value: 1.33×10^{-8}). This novel ncRNA was observed to positively correlate with lactate concentrations (R: 0.94). *MSTRG.18364* is a novel dual exonic transcript located downstream of the *HERC3* gene and was upregulated during the stationary phase (Fold-change: 2.94, BH Adj. P-value: 2.11×10^{-23}) but downregulated during the death phase (Fold-change: 1.70, BH Adj. P-value: 1.52×10^{-4}). This gene negatively correlated with glutamate concentrations over the duration of a fed-batch culture (R: -0.95). The multi-exonic antisense gene *MSTRG.19771* is composed of four component isoforms and

negatively correlated with ammonia concentrations ($R: -0.93$) and was downregulated 4 fold during the stationary phase (BH Adj. P-value: 2.33×10^{-10}).

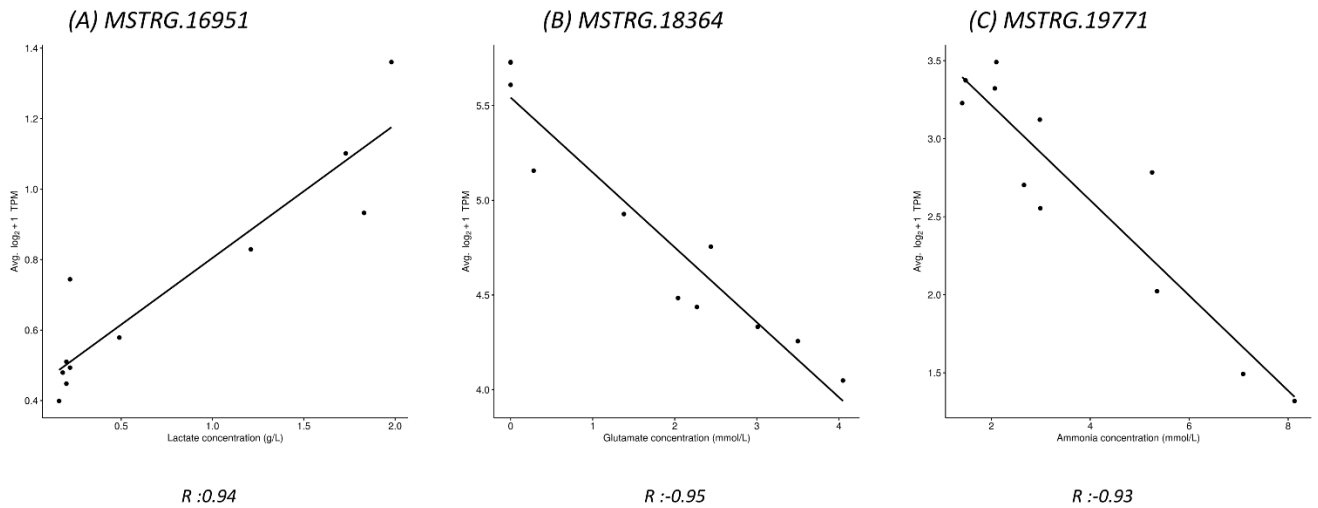


Figure 3.29 (A) Regression plot of correlation between lactate concentration and average normalized gene abundances of the *MSTRG.16951* ncRNA which was positively correlated. (B) Negative correlation between the novel *MSTRG.18364* ncRNA and glutamate concentration and (C) negative correlation between ammonia concentration and intergenic *MSTRG.19771* abundance.

3.4 Discussion

There is significant industrial and academic interest in understanding changes in phenotype, metabolism, and gene expression during fed-batch culture (Hsu et al. 2017). Previous high-throughput studies used metabolomics, proteomics, and transcriptomics in order to identify important changes in phenotype over the duration of cell culture (A. M. Lewis et al. 2016). These previous studies focused on the relationship between coding genes and culture phases and phenotypes. Currently, there is very little understood about the role long non-coding RNAs play in regulating cell cycle transitions and metabolism. In this chapter, we correlated, for the first time, lncRNA expression with the exponential, stationary and death phases over the duration of a 14-day fed-batch culture and link changes in gene abundance with proliferation and apoptosis. We also present the first evidence of correlated and anti-correlated expression of long non-coding RNAs to metabolites that are important for industrial bioprocessing.

3.4.1 Fed-batch culture parameters are concordant with previous CHO cell studies

Fed-batch culture processes can be divided into four distinct phases: lag, exponential, stationary, and death phase. The lag phase is the initial period of cell growth following inoculation which is followed by the exponential phase where cell density increases rapidly. Cell proliferation decreases during the stationary phase where specific productivity of recombinant proteins is the highest. A decline or death phase occurs near the end of the culture period when viable cell concentrations diminish due to onset of apoptosis. All of these phase transitions were clearly observed within two 10L stirred tank bioreactors over a 14-day fed-batch culture with an exponential phase that resulted in $\sim 2 \times 10^7$ viable cells/mL at the onset of the stationary phase after day 7 as viability began to decrease linearly. The difference between phases can be used to infer the correlation between lncRNA expression and phenotypes such as proliferation and apoptosis.

The growth phase began at day 3 and continued until after day 6 which coincided with the start of the stationary phase. A clear death phase was observed after day 10 where the total proportion of viable cells declined linearly until day 14. There were more differentially expressed coding genes between the exponential and stationary phase compared to the stationary and death phases. Gene ontology enrichment analysis for bioprocess terms found that genes involved in cell cycle, mitosis and DNA repair pathways were significantly differentially expressed during all stages. Genes associated with cellular division would be upregulated during the exponential phase while downregulated during the stationary phase as the specific growth rate decreases and these genes would be differentially expressed over the course of the culture (Borys et al. 2010). For example, the *Angiotensinogen (AGT)* gene was only modestly expressed during the exponential and stationary phases but becomes significantly abundant during the death phase. *Angiotensinogen* is a precursor molecule that is necessary for angiogenesis and vascular remodeling (Wu et al. 2015). Wu *et al.* found that *AGT* overexpression induced apoptosis in human epithelial cells (Wu et al. 2015). *AGT* function has not been studied in CHO cells but the results presented in **Chapter 3** suggest that this gene would be a viable target for cell engineering as *AGT* becomes significantly abundant only during the death phase. *BAX* and *BCL2* were among this cohort of genes with *BAX* differentially expressed between exponential and stationary phase and *BCL2* differentially expressed during the death phase. The *BAX* gene is a pro-apoptotic genes that is involved in the permeabilization of the mitochondrial membrane which triggers the release of caspases and the initiation of cell death (J. S. Lee et al. 2013). Previous research has shown that deletion or RNAi-mediated suppression of the *BAX* and *BAK* pro-apoptotic genes significantly increases apoptosis resistance (Grav et al. 2015). Deletion of these genes has also been shown to increase tolerance to cell stress such as DNA damage (Misaghi et al. 2013). The *BCL2* gene is an apoptosis regulator that is located on the mitochondrial membrane and plays an important role in cellular resistance to apoptosis (J. S. Lee et al. 2013). These two genes have previously been demonstrated as viable genetic engineering targets with deletion of the *BAX* gene and

overexpression of the *BCL2* genes shown to significantly delay apoptosis and increase viability (Grav et al. 2015; Misaghi et al. 2013). Previous successes with modulating the expression of genes involved in apoptosis informs the viability of genetic engineering strategies to improve total recombinant protein yield.

3.4.2 Differentially expressed genes between culture phases

We identified 861 genes that were that were significantly differentially expressed only during the transition from exponential to stationary phase and 491 between stationary and death phase. A combined 262 genes in these groups found to be evolutionarily conserved between Chinese hamster and humans. Very few genes were found to be conserved between the *criGri1* CHO annotation and the mouse annotation, however. This could simply be an artificial inflation of human lncRNAs as many non-coding RNAs that have been implicated in proliferative or apoptotic pathways have been studied in human cancer cell lines and would be overrepresented in the cross-species analysis (Nie et al. 2012; Su et al. 2016). Evolutionarily conserved intergenic lncRNA *NEAT1* was one of these genes that were significantly differentially expressed. Other conserved genes such as *HOTAIRM1* which were found in **Chapter 2** were not expressed in the fed-batch RNA-Seq data suggesting that these lncRNAs have specific regulatory mechanisms that are only present during temperature shift but not altered over time.

A significant barrier for lncRNA-mediated cell line engineering is the lack of information on regulatory function for the majority of transcripts. We used a “guilt by association” strategy where expression during a specific phase is linked to an observed phenotype. For example, alterations in abundance between exponential and stationary phases suggest involvement in cell division and growth processes. The *SCFD1* gene, which is involved in regulating the secretory pathway within the Golgi body, contains an abundant antisense gene that is upregulated while the sense coding gene becomes downregulated in the stationary phase. *SCFD1* is involved the secretion of extracellular matrix (ECM) components which are necessary for cellular proliferation (Hou et al.

2017). As cell growth decreases during the stationary phase, secretion of ECM proteins also decreases and this would result in a decrease in expression of the *SCFD1* gene (Hou et al. 2017). Downregulation of the *SCFD1* gene could occur due to transcriptional inhibition by the antisense gene. The change in abundance during the death phase can be used to infer the regulatory mechanism of these genes: genes with a positive increase are potentially active regulators of apoptotic pathways while downregulated genes could represent non-coding RNAs that are necessary for cell viability. For example, a novel antisense gene to the transmembrane protein *TMC1* becomes significantly downregulated during the death phase. The sense protein coding gene is not differentially expressed which suggests a *trans* regulatory function for the antisense gene. lncRNA differential expression could be independent of culture phase but this “guilt by association” strategy can be used to contextualize gene expression for future *in vitro* validation studies. These unique genes can be useful biomarkers that can be used to predict genomic changes during important metabolic transitions (Fatima et al. 2015).

3.4.3 NEAT1 is upregulated during death phase

One of the well-studied functions of lncRNAs is *NEAT1* mediated paraspeckle formation within the nucleus. Paraspeckles are aggregates of proteins found within the nucleus and *NEAT1* functions as a scaffold for proteins to bind and organize (West et al. 2014). These sub-nuclear RNA-protein complexes govern eukaryotic alternative splicing machinery (Bond and Fox 2009). *NEAT1*, along with its lncRNA partner *MALAT1*, are commonly found to be significantly enriched within highly proliferative cancer tissues such as breast and colorectal cancer (West et al. 2014). Curiously, we found that *NEAT1* was 1.25 fold upregulated during the death phase of the culture despite this lncRNA’s canonical role in cell growth and proliferation. Ding *et al.* proposed an alternative regulatory mechanism where *NEAT1* functions as an endogenous sponge for miRNA 34a-5p in ovarian cell lines which targets the 3’UTR of the *BCL2* gene (Ding et al. 2017). The authors found that *BCL2* expression increased 50% as *NEAT1* sequesters miR-34a-5p which prevents degradation of the anti-apoptotic gene (Ding et al. 2017). This novel result for alternative

regulatory control by *NEAT1* can be validated using the matched miRNA-Seq data generated for this study.

3.4.4 Sense/antisense differential expression over culture phases

Sense/antisense expression was observed throughout the fed-batch culture process with 27 loci exhibiting differential expression. Among this cohort was the heterodimeric DNA replication complex *GINS4* which exhibits anti-correlated expression and the Schlafen family member *SLFN12* which has both sense and antisense genes upregulated during the death phase. The *GINS4* locus was previously found to have anti-correlated expression on the sense and antisense strand after a reduction in culture temperature as reported in **Chapter 2**. The *GINS4* gene encodes a protein that is necessary for transcription initiation with upregulation correlated with increased proliferation and metastasis with downregulated apoptosis in gastric cancer cells (Z. Zhu et al. 2019). Gene abundances for the sense and antisense strands for the *GINS4* locus are comparatively low in both RNA-Seq datasets so it remains to be seen whether the sense coding gene or its antisense partner are expressed at physiologically relevant levels. The other gene that exhibited differential expression on both strands is a member of the *Schlafen* family of proteins which encode diverse function within eukaryotic cells. Elevated levels of the *SLFN12* has recently been discovered to associate with the ribosome in the endoplasmic reticulum which results in reduced translation of the *BCL2* anti-apoptotic gene which has increased rate of cell death (Turnover et al. 2019). Only three loci had differential expression on the sense/antisense strand across all three stages with only the *NAV2* locus expressed at physiologically relevant abundances. The *Neuron navigator 2* gene has been identified as an oncogene in human tumors but the regulatory function is not currently understood. Overexpression of *NAV2* has recently been linked to increased colorectal cancer cell proliferation which suggests that this locus might regulate cell division (Hu et al. 2019). After day 5, the antisense ncRNA becomes extremely abundant while the sense gene becomes depleted which suggests that *NAV2* is involved in cellular proliferation in CHO cells as the sense coding gene becomes downregulated as growth

rate and viability decrease. The antisense gene could function as a *cis* repressor of the *NAV2* gene and involved in the regulation of cellular proliferation over the course of a fed-batch culture.

3.4.5 Pearson's correlation between gene expression and process parameters

Differentially expressed coding and non-coding genes were correlated to process parameters including viable cell concentrations and % viability. The majority of correlated genes had a strong association with viability including the RNA binding protein *RBM3* which was previously found to be strongly upregulated after a reduction in culture temperature in **Chapter 2**. *RBM3* is thought to stabilize the 3'UTR of mRNAs in response to stress such as mild hypothermia (Dresios et al. 2005). *RBM3* has been linked to cellular proliferation as deletion of this locus resulted in reduced cell viability and increased apoptosis (Wellmann et al. 2010). *RBM3* expression within this dataset agrees with previous findings as we observed that gene expression was constant until day 7, at the start of the stationary phase, followed by a linear decrease in gene abundance that strongly correlated with the decrease in viable cells. Viability was constant at ~95% until day 7 when the percent of viable cells began to decrease linearly at the same time point *RBM3* expression decreased which further supports the relation between this gene and apoptosis regulation. A similar correlation analysis was performed for long non-coding RNAs and a cohort of 223 lncRNAs were found to be strongly associated with changes in % viability. An intergenic lncRNA, *MSTRG.2854*, was inversely correlated with viability (Pearson's correlation: -0.94). *MSTRG.2854* is a gene composed of 6 multi-exonic isoforms that was previously not annotated within the *ENSEMBL* version 98 *criGri1* CHO reference. Cross-species comparison to the human v32 GENCODE annotation found that this gene was located within a conserved locus that contains the *LINC02533* long intergenic ncRNA albeit in the antisense orientation to the human gene which suggests that *MSTRG.2854* is unique to *C. griseus*. *MSTRG.2854* is expressed at low abundances until day 6 when gene expression increases exponentially which coincides with a decrease in proportion of viable cells. The inverse correlation between cell viability and gene expression suggests that *MSTRG.2854* is a potential biomarker that can be used to evaluate

changes in genomic transcription caused by entry into the death phase. Future work will require *in vitro* detection of regulatory targets for this intergenic lncRNA.

3.4.6 Correlation of metabolite concentration to gene expression

We also correlated metabolite concentrations with mRNA and lncRNA expression over time.

Extracellular metabolite concentrations, including glucose, lactate, glutamine, glutamate and ammonia were tracked daily over the 14-day duration of the culture in duplicate bioreactors. Glucose is the major source of energy within media and is constantly supplemented during fed-batch culture. Glucose is catabolized into pyruvate during glycolysis for entry into the citric acid cycle. Rapid glycolysis can lead to accumulation of the toxic metabolite lactate which is converted from pyruvate. Glutamine is another major energy source that can be converted for use in the citric acid cycle but also functions as a major source of nitrogen which is necessary for amino acid and nucleic acid catabolism. Conversion of glutamine to glutamate produces ammonia which has cytotoxic properties during cell culture. These metabolites are commonly monitored during industrial recombinant protein production within bioreactors. Growth rate and viability are intrinsically linked to the availability of nutrients and metabolites such as glucose and glutamine (Luo et al. 2012; Noh, Shin, and Lee 2018). The work presented in this chapter is the first to correlate non-coding RNA expression to metabolite accumulation and depletion in industrially relevant conditions.

3.4.6.1 Glucose concentration did not significantly correlate to changes in gene expression

The rate of glucose consumption was observed to be at the highest during the lag and exponential phases where specific cell growth was the highest. There was no expression correlation between lncRNAs and glucose concentrations. The lack of correlation was surprising as Vito and Smales previously reported differential expression of mRNAs and lncRNAs before and after feeding in

fed-batch cultures (Vito and Smales 2018). This lack of correlation could be attributed to the daily feeding regime as glucose was added to the media when the concentration fell below 5 g/L which would result in a consistent concentration of the sugar during the period when cells were sampled for downstream RNA-Seq library preparation. There would, therefore, be no correlation between glucose concentration and lncRNA abundances. RNA-Seq experiments that studied lncRNA expression in relation to insulin deficiency and diabetes in human patients demonstrated that several evolutionarily conserved lncRNAs, including the *H19* non-coding gene, regulate glucose homeostasis (Sun and Wong 2016). *H19* regulates the expression of the *Insulin receptor (INSR)* gene via titrating *miR-let-7* away from its target UTR thereby increasing expression which regulated glucose bioavailability (Sun and Wong 2016). These previous studies directly implicate lncRNA expression with glucose consumption and homeostasis in humans and it is extremely likely that these molecules have similar regulatory effects on glucose availability within CHO cells. Future RNA-Seq experiments that study genome regulation directly before, during, and after feeding in industrial setting will doubtless identify glucose-mediated lncRNA expression networks.

3.4.6.2 Cohort of DE lncRNAs that correlated to lactate consumption

A transition from lactate production to lactate consumption within four days of the initial seeding with lactate completely depleted from the media after day 7. Lactate depletion coincides with the transition from exponential to stationary phase. Significant effort has been devoted to preventing the accumulation of lactate within the media while maintaining high growth rates (Luo et al. 2012). For example, a combinatorial strategy where *Lactate dehydrogenase A* was downregulated while *BCL2* was upregulated saw a significant decrease in lactate production (Jeon, Yu, and Lee 2011). This successful genetic engineering strategy suggests the utility of targeted modulation of expression as a viable method of improving bioprocess conditions. mRNA and lncRNA expression was correlated to lactate concentration with 57 coding genes and 8 non-coding genes having a Pearson's correlation coefficient greater than or equal to ± 0.90 . The *COX2* gene was the most anti-correlated mRNA to metabolite concentrations. This gene catabolizes the formation of

prostaglandins in mammalian cells and transcript stability has been shown to increase after overexpression of *RBM3* which suggests that *COX2* has a significant role in cell survival (Wellmann et al. 2010). While *COX2* is differentially expressed between the exponential and stationary phases, this gene remains highly abundant so the correlation between gene expression, lactate concentration, and *RBM3* expression may be spurious. lncRNAs offer a novel avenue for metabolic engineering to reduce lactate production. In this chapter, we present eight differentially expressed lncRNAs whose abundances correlated strongly with extracellular lactate concentration. Of these 8 genes, only two were found to have physiologically likely levels of abundance. The most abundant gene, *MSTRG.16951*, was found on the antisense strand downstream of the *LRRN3* gene with a strong positive correlation (Pearson's correlation: 0.94). The highest abundance of this gene was observed to occur on days 3-5 after which the gene became quiescent. *MSTRG.16951* was expressed at low abundances on day 0 at the start of the fed-batch culture so future sequencing experiments that more closely follow the lactate metabolism over the lag phase would be necessary to understand the relationship between lactate concentration and *MSTRG.16951*.

3.4.6.3 Glutamine consumption and ammonia production correlations

Glutamine was a component of the basal media at the time of initial feeding at ~3 mmol/L concentrations. Glutamine concentrations decreased linearly during the lag and exponential phase and was completely depleted by day 4. Glutamine was not a component of the feed during the fed-batch stages of the culture yet was observed to steadily increase in concentration. Glutamate was completely depleted from the extracellular media during the stationary phase when glutamine levels started to replenish. Glutamate, which is a component of the fed-batch bolus, can be converted into glutamine which can in turn be a substrate and lead to ammonia accumulation (F. Zhang et al. 2006). A surprising result of the correlation analysis was the lack of significant correlation of glutamine concentration to coding and non-coding normalized gene abundances. This can be attributed to the lack of RNA-Seq data from days 1 and 2 where the

rate of glutamine consumption was the highest. RNA-Seq profiling over a shorter interval during the lag phase could potentially find genes that more closely correlate with glutamine concentrations. 17 lncRNA genes were observed to be strongly correlated with glutamate concentrations, however. Among this cohort was the novel gene, *MSTRG.18364*, which is located downstream of the *HERC3* gene. This gene exhibited anti-correlation to glutamate concentrations (R: -0.95) with the highest abundance occurring on days 7-9 when glutamate was completely depleted from the culture media. This gene is unique to the CHO genome so it is difficult to evaluate the potential function role of *MSTRG.18364* but the correlation analysis is useful in contextualizing expression with potential regulatory role. We also found four mRNA coding genes that inversely correlated with viability and positively correlated with ammonia concentrations. Ammonia is a toxic metabolite that results in increased apoptosis which supports the inverse relationship between these genes and viability (Wahrheit, Nicolae, and Heinzle 2014). *LPCAT3*, which is the most abundant gene in this small cohort, is involved in acylation of fatty acids during synthesis (Jain et al. 2009). There has not been a significant study of *LPCAT3* and its relation to ammonia excretion, however, but deficiency in long chain fatty acids has been linked to increased NH_4^+ accumulation in clinical studies (Gauthier et al. 2013). We performed a similar analysis of lncRNA abundance compared to ammonia concentration and found 4 long non-coding genes that positively correlate with NH_4^+ accumulation and one negatively correlated gene, the antisense *TMC1-AS* gene which was significantly differentially expressed during the death phase and very abundant. The cohort of lncRNAs that correlate with glutamate and ammonia concentration provide an alternative avenue to reduce toxic metabolite accumulation.

3.4.7 Potential utility of identified lncRNAs

We present the first evidence of correlated and anti-correlated long non-coding RNA expression to metabolites that are relevant for bioprocessing such as lactate, glutamate, and ammonia. Correlation of gene expression to metabolite concentration does not necessarily equate to direct regulation by lncRNAs, however, and further validation is required to determine whether these

genes play a deterministic role in cellular metabolism (Hsu et al. 2017). We focused on the potential relationship between non-coding genes to glucose and glutamine metabolism as these pathways are of significant interest in industrial contexts but the analysis presented here can be adapted to profile other metabolite-lncRNA interactions such as pyruvate and acetate. lncRNAs are known to play an active regulatory role in eukaryotic cell metabolism and this property can be leveraged to engineer CHO cells with more robust metabolism and reduced toxic metabolite accumulation.

3.5 Conclusions

The long non-coding RNA annotation was significantly expanded to include 8,406 transcripts which will be an extremely useful resource for the CHO community. In this chapter, we present the first evidence of the alterations in lncRNA expression within a fed-batch culture that correlate to important transitions. We have also correlated long non-coding RNA gene expression with industrially important metabolite concentrations: lactate, glutamine, and ammonia. These genes will be of great significance for future cell engineering efforts as their expression suggests a regulatory role in industrially relevant metabolic pathways.

Chapter 4

Alternative Splicing in

Fed-Batch Culture

4 Summary

In **Chapter 4**, we first summarize alternative splicing and a computation method to detect this phenomenon using alterations in isoform abundance using RNA-Seq data derived from fed-batch culture presented in the previous chapter. A transcript level analysis was performed using high-coverage RNA-Seq data with the *IsoformAnalyzeR* tool (**Section 4.2**). In **Section 4.3** we present an analysis of global alternative splicing of mRNAs between the exponential, stationary, and death phases of a fed-batch culture. We also present the first evidence of global alternative splicing of long non-coding RNAs within CHO cells and correlated the splicing events to culture phases. **Chapter 4** will conclude on a discussion of the alterations in gene function mediated by alternative splicing and the significance to CHO cell biology.

4.1 Introduction

In **Chapters 2** and **3**, gene level quantification was performed in order to correlate changes in lncRNA abundance with industrially relevant conditions: temperature shift and fed-batch cultivation respectively. Gene level analysis of long non-coding RNAs is extremely useful for inferring the relationship between phenotype and global alterations in expression. One of the significant results of the GENCODE human and mouse annotation project was that lncRNA genes can contain multiple isoforms and are alternatively spliced in a manner similar to mRNAs (Derrien et al. 2012). Very little is currently understood about lncRNA splicing despite the knowledge of widespread alternative splicing within model organisms but different isoforms are thought to have altered regulatory function within the eukaryotic cell (Gonzalez et al. 2015a). Our group was the first to demonstrate genome-wide alternative splicing of mRNAs within Chinese hamster ovary cells that were subjected to a reduction in culture temperature from 37°C to 31°C and found that alternative splicing can occur independently of changes in gene expression (Tzani et al. 2019). In this chapter, we present the first study of lncRNA alternative splicing and correlated transcript expression to culture stages within duplicate bioreactors over a 14-day fed-batch cultivation period.

Alternative splicing (AS) is a cellular process that generates multiple, unique transcripts from the same parent gene and underlies transcriptome diversity (Park et al. 2018). DNA is initially transcribed into a precursor RNA that contains all introns and exons within a gene (Reyes and Huber 2018). The intronic sequences are removed with the exon sequences arranged in different configurations by the spliceosome: a complex of small ncRNAs and splicing factor proteins. The spliceosome complex recognizes binding sites within precursor RNAs that govern the production of alternative splice isoforms (**Figure 4.1**). The mechanism that generates alternative splice isoforms within long non-coding RNAs is not understood but is thought to be similar to the process that generates mRNA transcripts (Melé et al. 2017).

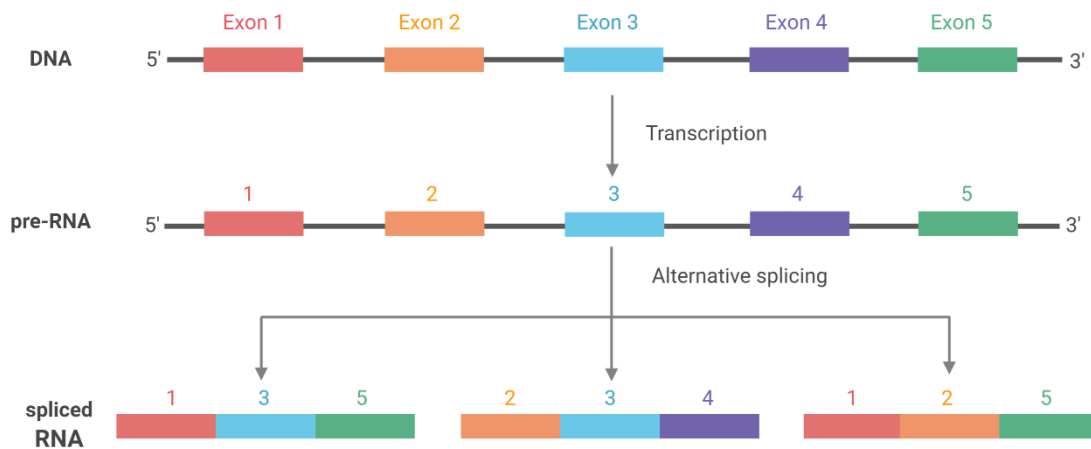


Figure 4.1 A diagram summarizing three alternative splice isoforms generated from a double-stranded gene. The entire gene with corresponding introns and exons are first transcribed into precursor RNA molecules which are processed by the spliceosome into mature transcripts without introns.

The efficiency of splicing was found to be lower than mRNAs due to a comparative lack of splicing factor binding sites (Krchňáková et al. 2019). The distribution of splice sites are thought to play a larger role in driving alternative splicing than in mRNAs which interact with splicing regulatory proteins (Krchňáková et al. 2019). Several lncRNAs have been implicated in regulating mRNA alternative splicing in model organisms by controlling post-translational modification of splicing factors, forming RNA-RNA duplexes, and epigenetically controlling transcription by guiding chromatin remodeling complexes to the promoter. *MALAT1*, which is enriched within the nucleus in bodies known as paraspeckles which are complexes of splicing factors, is thought to regulate alternative splicing by controlling phosphorylation of splicing factors which in turn are responsible for splicing pre-mRNAs into mature transcripts (Tripathi et al. 2010). Regulating phosphorylation of splicing factors controls the distribution of these proteins within nuclear paraspeckles which in turn drive alternative splicing (Gutschner et al. 2013). Regulation of alternatively splicing by lncRNA-mRNA duplexes is not currently well-understood but one proposed mechanism is masking of mRNA splice sites on the pre-mRNA (Ciarlo et al. 2013). Masking the splice site prevents splicing factor binding which would drive splicing at other intron-exon junctions and generate other isoforms (Ciarlo et al. 2013).

Previous studies have predominantly focused on how lncRNAs regulate alternative splicing of coding transcripts but little is currently understood about what regulates lncRNA transcription and splicing. Alternative splicing can also be regulated at the genomic level by guiding chromatin remodeling complexes to alter chromatin marks to prevent splicing factors from binding (Gonzalez et al. 2015a). One theory is that alternative splicing is governed by methylation of H3K4me3 histone proteins within open chromatin (Niemczyk et al. 2013a). This mechanism was first identified in the *GNG12-AS* locus that has isoforms that are co-transcribed with the sense protein coding gene. Loss of methylation within the gene body suppressed the sense protein coding gene and altered splicing in the *GNG12-AS* gene in breast cancer cell lines (Niemczyk et al. 2013b). It is currently unknown whether other mechanisms exist that regulate alternative lncRNA splicing (Ziegler and Kretz 2017).

AS can be broadly classified into six events (**Figure 4.2**): intron retention, exon skipping, mutually exclusive exons, alternative 5' splice sites, alternative 3' splice sites, and alternative transcript start and termination sites (Park et al. 2018). Intron retention occurs when an intronic sequence that is normally removed is spliced into the mature transcript. Exon skipping and mutually exclusive exons are splicing events where different exons are included within the mature transcript (Matlin, Clark, and Smith 2005). Alternative 5' and 3' splice sites result in longer or shorter exon sequences due to internal splice sites found within exonic sequences while alternative start and termination sites alter the length of the untranslated regions (Matlin, Clark, and Smith 2005).

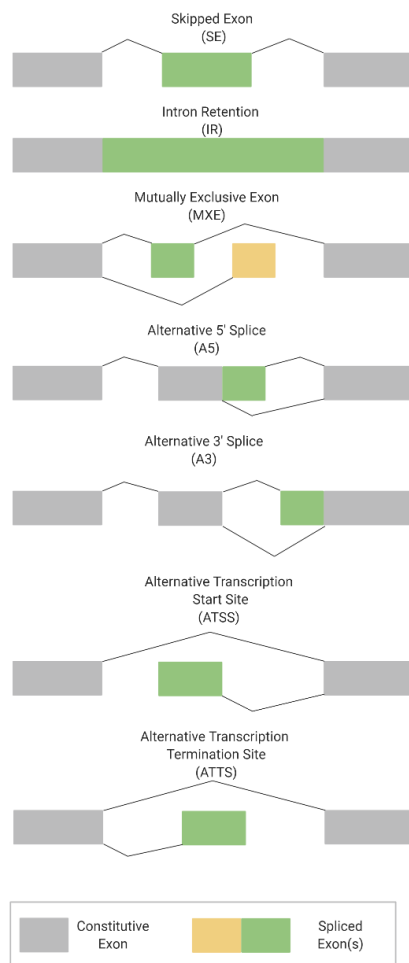


Figure 4.2 The different classifications of alternative splicing as defined by IsoformSwitchAnalyzeR. Skipped exons (SE) are events where an internal exon is spliced out of the precursor whereas an intron retention (IR) event is when the intron is retained instead of being removed as normal. Mutually exclusive exons (MXE) are preferential splicing of one exon over another between conditions. Alternative 5' (A5) and 3' (A3) splice sites alter the exon boundaries during expression. Alternative start (ATSS) and stop (ATTS) sites alter the boundaries of transcription.

High-throughput RNA sequencing has enabled genome-wide transcript level quantification and assembly of individual isoforms. RNA-Seq is advantageous for alternative splicing analysis as this method does not require hybridization to known probes so novel isoforms can be detected. RNA-Seq also provides absolute transcript abundances so it is possible to evaluate absolute changes in isoform expression independent of gene expression (Soneson, Love, and Robinson 2018). RNA-Seq identified pervasive alternative splicing events within the human and mouse genome in both coding and non-coding loci.

Total gene abundance can be quantified as the fraction of each component isoform which can be expressed at different proportions of the total abundance (Vitting-Seerup and Sandelin 2019). The change in isoform fraction between conditions can be used to quantify alternative splicing events. Computational pipelines, such as *IsoformSwitchAnalyzeR* (Vitting-Seerup and Sandelin 2019), have been developed that can quantify differential alternative splicing by calculating the changing proportions of isoforms with respect to the aggregate gene abundance in a method referred to as isoform switching (Frazee et al. 2015). *IsoformSwitchAnalyzeR* was first used to study the global changes in isoform expression and alternative splicing in 1 previously sequenced RNA-Seq data within the Cancer Genome Atlas (TCGA) from 12 different cancer types including breast and colorectal cancer (Vitting-Seerup and Sandelin 2017). Vitting-Seerup & Sandelin found that 19% of all protein coding genes underwent statistically significant isoform switching (Vitting-Seerup and Sandelin 2017). The authors correlated these switches to patient survival and found a set of 31 transcripts that are common to all analyzed cancers that have potential functional consequences for the translated protein (Vitting-Seerup and Sandelin 2017). These results suggest that aberrant alternative splicing is important in cancer progression and important for clinically relevant cellular processes such as reduced apoptosis and increased cellular proliferation.

This tool uses the normalized gene and transcript abundances as well as the transcript models created by *StringTie*. Isoform TPM normalized gene abundances are then converted to fractions of the total normalized gene abundance evaluated for transcript level differential expression (Love et al. 2019). The change in isoform fraction between experimental groups is evaluated using *DEXseq* which calculates a negative binomial distribution of isoform fractions (Love et al. 2019). The difference in isoform fraction (dIF) is calculated between two groups using a generalized linear model which also calculates a Benjamini-Hochberg adjusted P-value that is a measure of the false discovery rate (FDR) (Vitting-seerup and Sandelin 2019; Love et al. 2019). The final

stage of the *IsoformSwitchAnalyzeR* workflow is the classification of significant switching isoforms into alternative splicing events by first constructing a theoretical precursor RNA which is a theoretical transcript model containing all exons of each component isoform. An AS event is classified by comparing the assembled isoform to the theoretical precursor transcript (Vitting-Seerup and Sandelin 2019). These splicing events can then be evaluated for functional consequence such as the loss of coding domains.

Previous studies in model organisms has demonstrated genome-wide alternative splicing events in coding genes but there have been very few high-throughput studies that profiled global AS for lncRNAs. In this chapter, we use high-coverage RNA-Seq data presented in **Chapter 2** of CHO cells cultivated as fed-batch cultures in profile isoform switching and alternative splicing in mRNAs and lncRNAs between the exponential, stationary, and death phases. We present the first evidence of alternative splicing of lncRNAs within CHO cells and relate the expression to cell culture phase.

4.2 Materials and Methods

4.2.1 Transcript quantification using StringTie

Aligned reads from **Chapter 3** were assembled by *StringTie* into transcript loci using the version 98 *criGri1* ENSEMBL reference annotation. The reference guided assembly was performed on each individual alignment prior to being merged with the *StringTie* *-merge* tool. A putative novel isoform was included within the assembly if the transcript was at least 10% of the total isoform fraction, had a minimum of average coverage of 10 reads with at least 5 reads aligning to novel splice junctions. The assembly was used to quantify transcript abundance using the *-eB* parameters which are used to generate estimates of transcript expression.

4.2.2 Isoform switching and alternative splicing

Transcript Per Million (TPM) normalized abundances were imported into *IsoformSwitchAnalyzeR* for isoform switching and alternative splicing analysis. The full transcript sequence and annotation file were also imported into the R environment. Single isoform genes and transcripts that did not have a TPM of 1 in at least sample were removed from the analysis. Isoform switching analysis was performed to determine the difference in transcript abundance between the exponential, stationary, and death culture phases over 11 days as described in **Chapter 2**. A transcript was considered to be significant if the difference in isoform fraction (dIF) was greater than 10% with a Benjamini-Hochberg adjusted P-value < 0.05.

4.2.3 Inferring biological significance of isoform switching

Isoform switching transcripts were segregated into coding and non-coding genes using the results of the lncRNA classification performed in **Chapter 3**. Isoforms that were found to be significantly differentially expressed were classified into skipped exon, alternative 5' and 3' splice site, retained introns, and mutually exclusive exon alternative splicing categories using the *analyzeAlternativeSplicing* tool within *IsoformSwitchAnalyzeR*. The significance of an mRNA alternative splicing event was evaluated by searching domains against the *PFAM* database. Sashimi plots of alternative splicing and differential transcript utilization were generated using the

rmats2sashimi utility. The *RNAfold* webserver was used to visualize secondary structure of lncRNAs of interest.

4.3 Results

4.3.1 Isoform and gene structure of lncRNA and mRNA loci

StringTie was used to assemble aligned reads from **Chapter 3** into transcript models which were used as input for the lncRNA detection pipeline as described in **Chapter 2**. A similar distribution of the number of exons per transcripts was found with the exon distribution in **Chapter 2**. Dual exonic lncRNA transcripts were the most frequently occurring class accounting for almost 50% of all lncRNA transcripts (**Figure 4.3**). The distribution of the number of exons per mRNA was found to be constant with an approximately equal proportion.

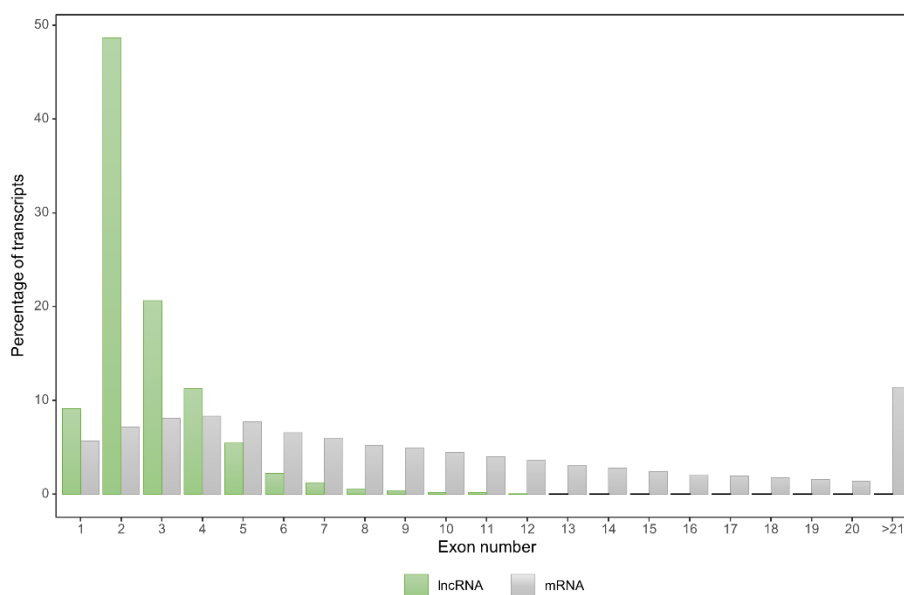


Figure 4.3 Proportion of exons per transcript for lncRNAs compared to mRNAs. Dual exonic transcripts were the largest group of lncRNAs with almost 50% of all transcripts containing 2 exons.

This assembly expanded the total number of coding transcripts by 30% as these mRNAs compared to the *ENSEMBL* reference sequence as these sequences were not found within the *ENSEMBL criGri1* version 98 reference annotation. The results of the lncRNA detection pipeline annotated 8,406 transcripts with over 85% novel non-coding transcripts. Approximately 75-80% of mRNAs and lncRNAs genes were single isoform loci in the merged *StringTie* assembly. There was a similar distribution of isoforms per gene for coding and non-coding genes (**Figure 4.5**) with < 5% of both gene sets composed of 3 or more isoforms. *GFFCOMPARE* was used to compare

the previous assembly derived from RNA-Seq data from a temperature shift experiment presented in **Chapter 2**. 65% of lncRNA exons (60% transcripts) were represented in the assembly from **Chapter 2** and the assembly using fed-batch RNA-Seq data from this chapter.

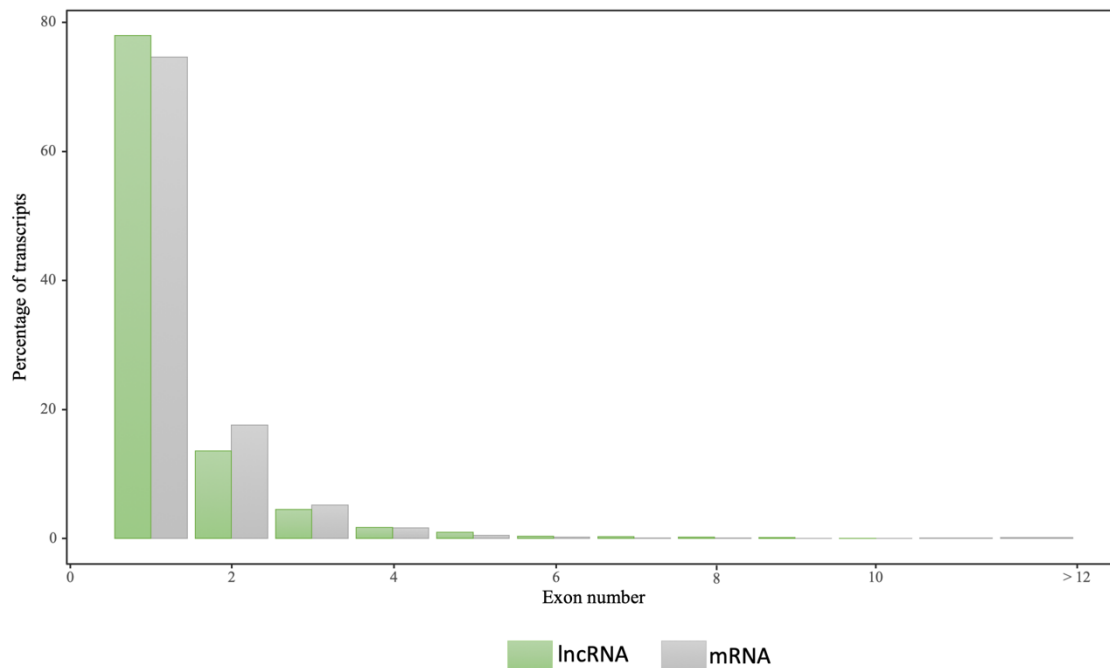


Figure 4.4 A bar plot showing the number of isoform as a percentage of the total number of coding or long non-coding genes. Single isoform genes account for almost 80% of all coding and lncRNA genes. Almost 95% of both cohorts contain 3 or fewer isoforms.

4.3.2 Transcript level quantification and quality control

A transcript level quantification was performed using *StringTie* and the consensus assembly presented in **Chapter 3**. The transcript abundances were imported into the *IsoformSwitchAnalyzeR* package within the *R* statistical framework. Single isoform genes and low expressed isoforms were removed from the dataset prior to isoform switching analysis. An isoform had to have a minimum of 1 TPM in at least one time point to be retained for further analysis. A combined total of 74% of all transcripts were removed due to insufficient abundance or lack of multiple isoforms. 12,130 mRNA transcripts and 1,280 lncRNA transcripts were retained after filtering which were evaluated for isoform switching.

4.3.3 mRNAs undergoing isoform switching between exponential and stationary phase

A total of 779 mRNA isoforms originating from 645 genes exhibited difference in isoform fraction between the exponential and stationary phases within the RNA-Seq data. There were an equivalent number of upregulated transcripts (401) and downregulated transcripts (378) with 277 transcripts also observed to be differentially expressed at the gene level. **Table 4.1** summarizes the top 10 differentially expressed transcripts with the corresponding-fold-change in gene expression between exponential and stationary phases. There were two differentially expressed genes, *PIGA* and *SNX24*, where the dominant isoforms change between the exponential and stationary phases. *PIGA* contains 7 isoforms but only two (*ENSCGRT00000009387* and *MSTRG.8948.7*) were found to be expressed the RNA-Seq data. **Figure 4.5** shows that the change in average gene abundance over time can be attributed to a ~70% increase in isoform fraction of the novel *MSTRG.8948.7* transcript.

Table 4.1 A table of significantly differentially transcribed isoforms that are also differentially expressed at the gene level between stationary and exponential phase. The *PIGA* and *SNX24* genes have isoforms that become dominant during the stationary phase that attribute to the overall change in gene expression.

Gene	Gene l2FC	Gene Adj. P-value	Isoform ID	dIF	Isoform Adj. P-value
<i>PIGA</i>	0.60	4.51×10^{-4}	<i>MSTRG.8948.7</i>	0.75	2.14×10^{-7}
<i>PIGA</i>	0.60	4.51×10^{-4}	<i>ENSCGRT00000009387</i>	-0.75	3.78×10^{-11}
<i>CD276</i>	0.65	4.33×10^{-5}	<i>ENSCGRT00000018557</i>	-0.64	7.11×10^{-7}
<i>SNX24</i>	0.62	1.46×10^{-7}	<i>MSTRG.9907.1</i>	-0.63	4.11×10^{-7}
<i>SNX24</i>	0.62	1.46×10^{-7}	<i>MSTRG.9907.3</i>	0.63	4.11×10^{-7}
<i>GABRG3</i>	3.48	3.22×10^{-11}	<i>ENSCGRT00000019489</i>	-0.61	1.45×10^{-3}
<i>USP49</i>	-0.89	7.28×10^{-7}	<i>MSTRG.15182.2</i>	-0.57	2.89×10^{-7}
<i>NHSL2</i>	0.64	2.95×10^{-8}	<i>MSTRG.14977.3</i>	0.55	1.62×10^{-4}
<i>TEX10</i>	0.68	1.92×10^{-11}	<i>MSTRG.12037.2</i>	0.55	6.88×10^{-6}
<i>COBL</i>	2.21	8.78×10^{-6}	<i>MSTRG.8659.4</i>	0.54	3.44×10^{-3}

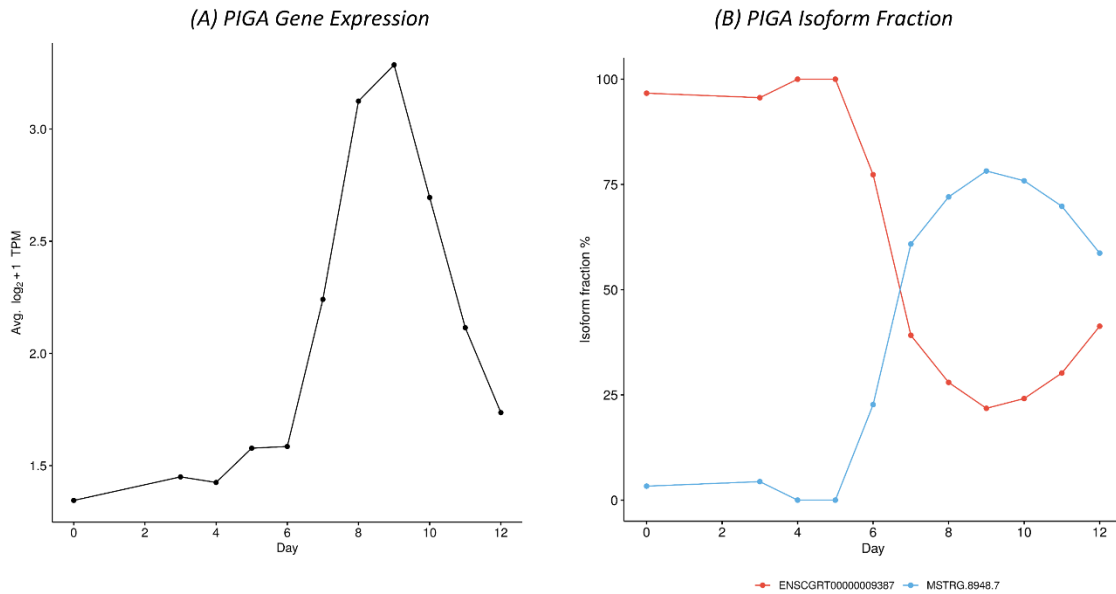


Figure 4.5 (A) Gene abundance over time for the PIGA gene. A 1.5 fold increase in gene expression occurred between the exponential and stationary phase after declining during the start of the death phase. **(B)** Normalized transcript abundances of the two dominant PIGA component isoforms. The novel isoform MSTRG.8948.7 becomes the most expressed transcript during the stationary phase and its change in expression is reflected in the gene level abundance profile. Three other isoforms exist within the PIGA gene but are not expressed beyond the lag phase and have been omitted in the plot for clarity. After day 3, ENSCGRT00000009387 and MSTRG.8948.7 are the only two expressed isoforms and make up 100% of the abundance.

484 mRNA transcripts exhibited isoform switching independently of gene differential expression with an equivalent number of upregulated (250) and downregulated (254) transcripts. **Table 4.2** summarizes the top ten transcripts that have a significant change in isoform fraction between the exponential and stationary phase. *Citrate synthase* gene contains 5 isoforms but only two transcripts were expressed: *MSTRG.12084.1* and *MSTRG.12084.2*. The first novel isoform, *MSTRG.12084.1* accounted for almost 100% of *CS* gene expression during the exponential phase while the second isoform, *MSTRG.12084.2*, accounted for the majority of expression during the stationary phase (**Figure 4.6**).

Table 4.2 A table of significantly differentially transcribed isoforms that were not differentially expressed at the gene level between stationary and exponential phase.

Gene	Isoform ID	dIF	Isoform Adj. P-value
ZFP68	ENSCGRT00000012749	-0.62	5.31×10^{-3}
ZFP68	ENSCGRT00000012750	0.62	5.27×10^{-3}
CS	MSTRG.12084.1	-0.73	1.14×10^{-4}
CS	MSTRG.12084.2	0.57	1.60×10^{-4}
PCMTD1	MSTRG.13434.2	0.71	7.27×10^{-8}
CTTNBP2NL	MSTRG.21314.3	-0.56	1.90×10^{-2}
FLNC	MSTRG.2901.4	-0.71	4.19×10^{-5}
PGS1	MSTRG.4371.1	-0.72	1.21×10^{-4}
SS18	MSTRG.7828.3	-0.66	1.36×10^{-4}
NEO1	MSTRG.8021.3	0.59	6.82×10^{-3}

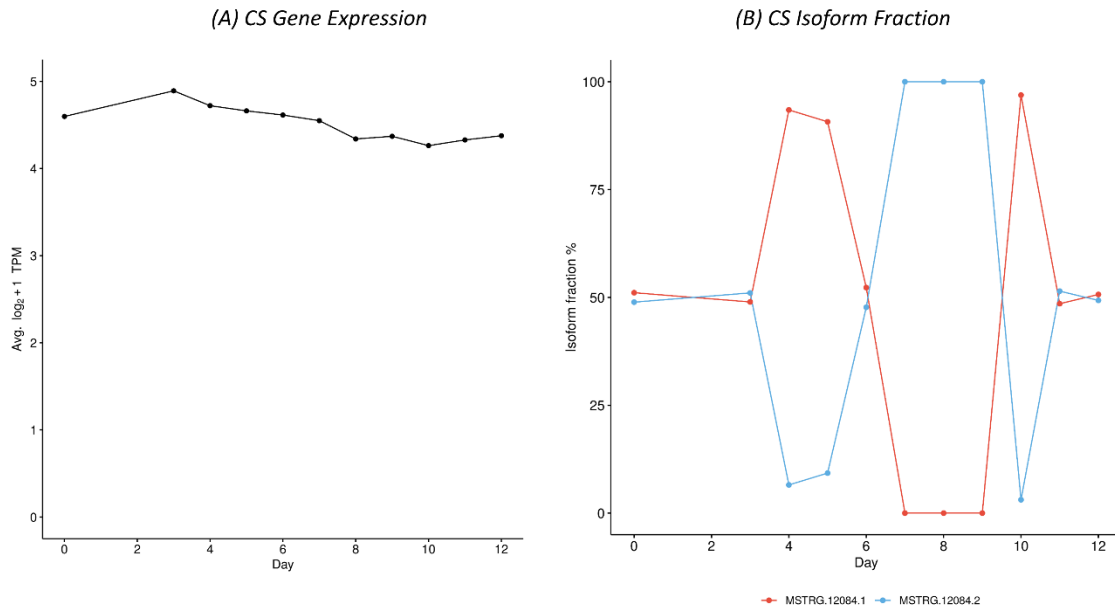


Figure 4.6 (A) Gene abundance over time for the CS gene which was seen to be expressed at a stable abundance over time. 100% of gene expression can be attributed to two novel isoforms: MSTRG.12084.1 and MSTRG.12084.2. (B) The two component isoforms were observed to be dominant during different culture phases with MSTRG.12084.1 expressed the most during exponential phase and MSTRG.12084.2 during the stationary phase.

Supplementary Table 5.1 summarizes all transcripts that had significant isoform switching events. Gene ontology enrichment of biological processes was performed using *DAVID* functional annotation suite with 11 processes significantly enriched below a FDR threshold of 0.05. **Table 4.3** details all enriched of transcripts that undergo isoform switching between the two cell culture

phases. Protein coding genes that play a role in the production and metabolism of non-coding RNAs were significantly enriched as well as those processes that are necessary for ncRNA genomic localization.

Table 4.3 Gene ontology enrichment for cellular processes for genes that undergo isoform switching without a change in expression between the exponential and stationary phases. Processes involved in ncRNA post-processing and cellular localization are significantly enriched.

GO ID	GO Term	Count	FDR
GO:0034470	ncRNA processing	27	3.15×10^{-4}
GO:0006396	RNA processing	46	3.36×10^{-4}
GO:0044085	Cellular component biogenesis	110	3.52×10^{-4}
GO:0034660	ncRNA metabolic process	32	5.58×10^{-4}
GO:0033036	Macromolecule localization	110	9.28×10^{-4}
GO:0015931	Nucleobase-containing compound transport	18	1.20×10^{-3}
GO:0051236	Establishment of RNA localization	15	2.13×10^{-2}
GO:0022607	Cellular component assembly	95	3.13×10^{-2}
GO:0033554	Cellular response to stress	72	3.71×10^{-2}
GO:0006403	RNA localization	16	4.94×10^{-2}

4.3.4 mRNA alternative splicing events between exponential and stationary phase

Significantly switching transcripts had multiple alternative splicing events with an average of 2 events per transcript. **Figure 4.7** summarizes the number of alternative splicing events in transcripts that underwent isoform switching. There were no upregulated mutually exclusive exon events but there were two downregulated switches during the stationary phase.

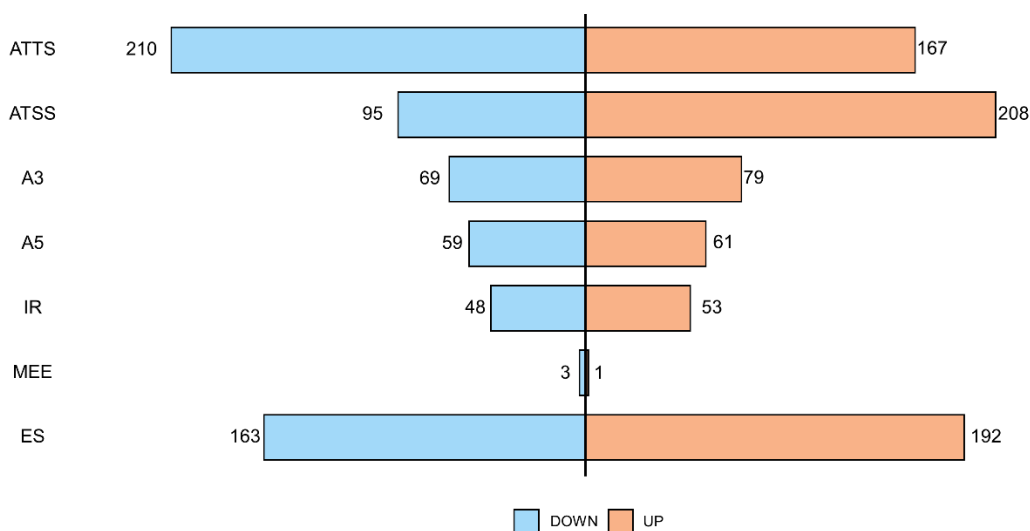


Figure 4.7 A leaf plot showing the number of alternatively spliced transcripts that underwent significant isoform switching between exponential and stationary phase

Figure 4.8 shows an example of alternative transcript termination site splicing between component transcripts of the dual isoform *Laccase domain containing 1 (LACC1)* gene where a novel isoform, *MSTRG.5965.1*, becomes dominant during the stationary and death phases. The *LACC1* is not differentially expressed at the gene level between exponential and stationary phases. The novel isoform contains a putative alternative 3' UTR that becomes 22% more abundant during the stationary phase (Adj. P-value: 4.57×10^{-5}). The CDS sequence of each isoform remained unchanged but contained an alternatively spliced 3'UTR.

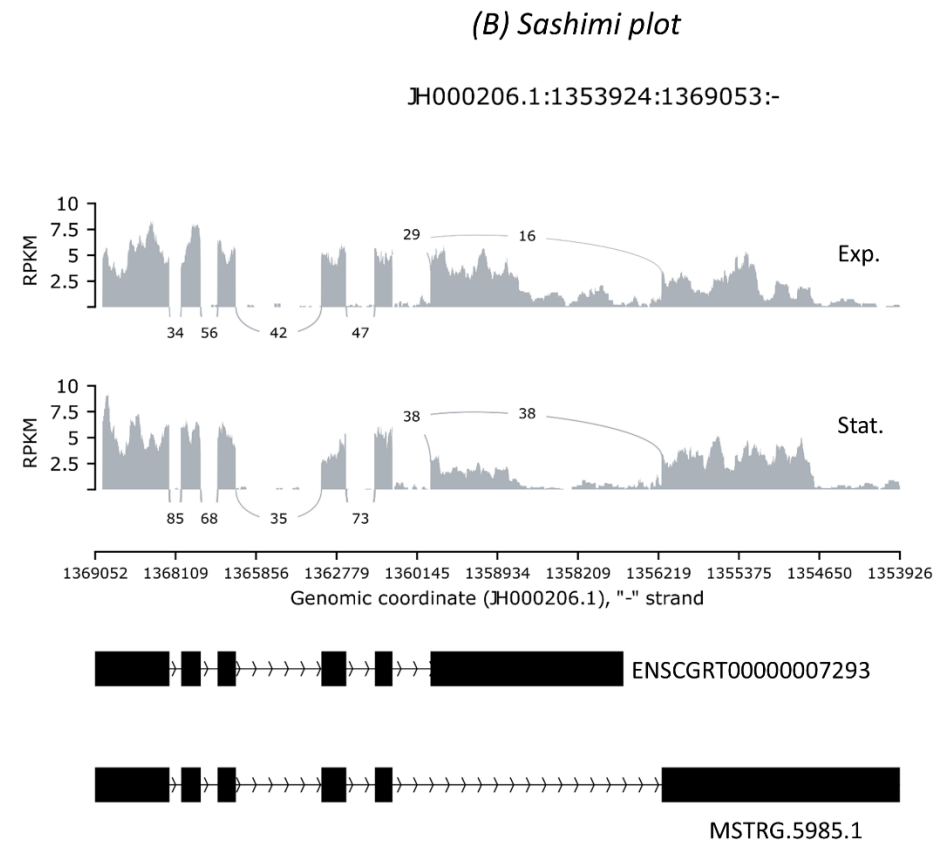
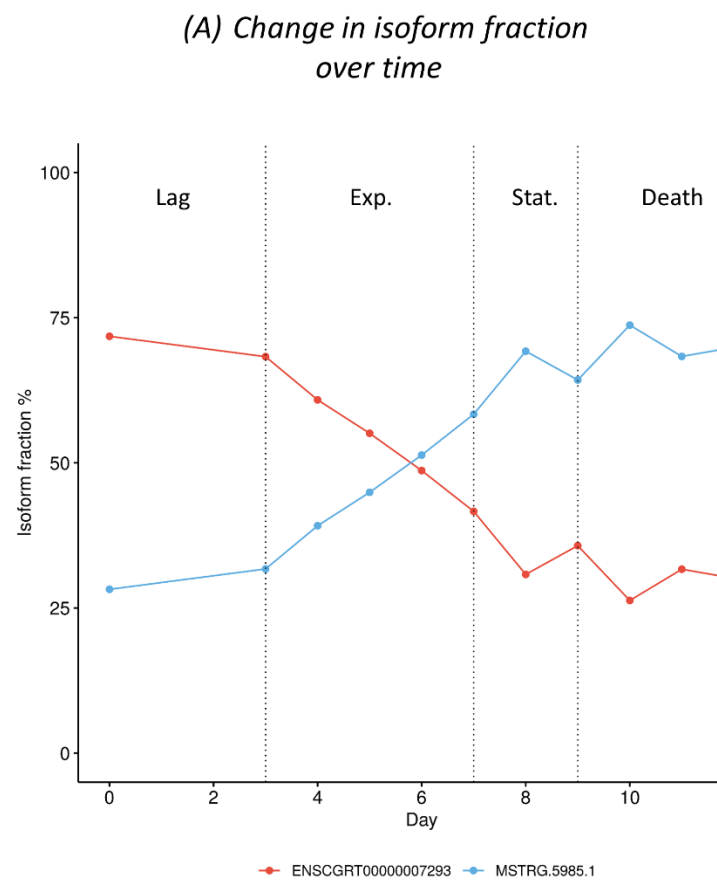


Figure 4.8 (A) Change in isoform fraction over time for the Laccase Domain 1 (LACC1) gene which is composed of two isoforms: a previously annotated transcript (ENSCGRT00000007293) and the novel transcript MSTRG.5985.1 which becomes dominant during the exponential and death phases of culture. **(B)** A 22% increase in the isoform fraction for MSTRG.5985.1 occurs with a concomitant decrease of 22% for ENSCGRT00000007293 (Adj. P-value: 4.57×10^{-10})

4.3.4 mRNA isoform switching between stationary and death phases

A total of 471 mRNA isoforms from 289 genes exhibited isoform switching between the stationary and death phases within the RNA-Seq data. There were an equivalent number of upregulated transcripts (231) and downregulated transcripts (240). 134 switching transcripts were also observed to be differentially expressed at the gene level. **Table 4.4** summarizes the top 10 differentially expressed transcripts with the corresponding-fold-change in gene expression between stationary and death phases. There were two differentially expressed genes, *GEN1* and *CLYBL*, where the dominant isoforms change between the exponential and stationary phases. Both of these coding genes are dual isoform genes where the novel transcript is expressed reciprocally to an *ENSEMBL* annotated transcript. **Figure 4.9** shows that the change in average gene and transcript abundance over time for the *CLYBL* gene. The gene abundance increased 1.54 fold during the death phase while the abundance of the novel isoform *MSTRG.8775.1* increases by 69%. During the exponential and stationary phases, the novel isoform is seen to have a “pulsing” expression where the abundance is periodically depleted every two days until reaching the death phase when this transcript accounts for ~80% of all expression within the *CLYBL* gene.

Table 4.4 A table of significantly differentially transcribed isoforms that are also differentially expressed at the gene level between stationary and death phase.

Gene	Gene l2FC	Gene Adj. P-value	Isoform ID	dIF	Isoform Adj. P-value
<i>BARD1</i>	-2.15	1.51×10^{-17}	<i>MSTRG.13401.1</i>	0.70	1.55×10^{-2}
<i>ENSCGRG00000003859</i>	-2.28	2.99×10^{-30}	<i>ENSCGRT00000005371</i>	0.70	7.73×10^{-6}
<i>ENSCGRG00000016618</i>	-2.45	4.14×10^{-25}	<i>MSTRG.14422.2</i>	-0.70	2.98×10^{-3}
<i>ENSCGRG00000017799</i>	-2.45	4.14×10^{-25}	<i>MSTRG.14422.2</i>	-0.70	2.98×10^{-3}
<i>ENSCGRG00000016618</i>	-2.45	4.14×10^{-25}	<i>ENSCGRT00000024890</i>	0.70	3.09×10^{-3}
<i>ENSCGRG00000017799</i>	-2.45	4.14×10^{-25}	<i>ENSCGRT00000024890</i>	0.70	3.09×10^{-3}
<i>ORC1</i>	-2.40	4.53×10^{-23}	<i>ENSCGRT00000020748</i>	-0.67	7.52×10^{-5}
<i>KIF14</i>	-2.47	1.77×10^{-20}	<i>ENSCGRT00000025123</i>	0.70	5.29×10^{-3}
<i>GEN1</i>	-2.66	2.05×10^{-26}	<i>ENSCGRT00000011868</i>	-0.69	1.37×10^{-5}
<i>GEN1</i>	-2.66	2.05×10^{-26}	<i>MSTRG.8732.1</i>	0.69	1.42×10^{-5}
<i>CLYBL</i>	0.63	6.86×10^{-6}	<i>ENSCGRT00000022202</i>	-0.69	5.44×10^{-5}
<i>CLYBL</i>	0.63	6.86×10^{-6}	<i>MSTRG.8775.1</i>	0.69	5.44×10^{-5}

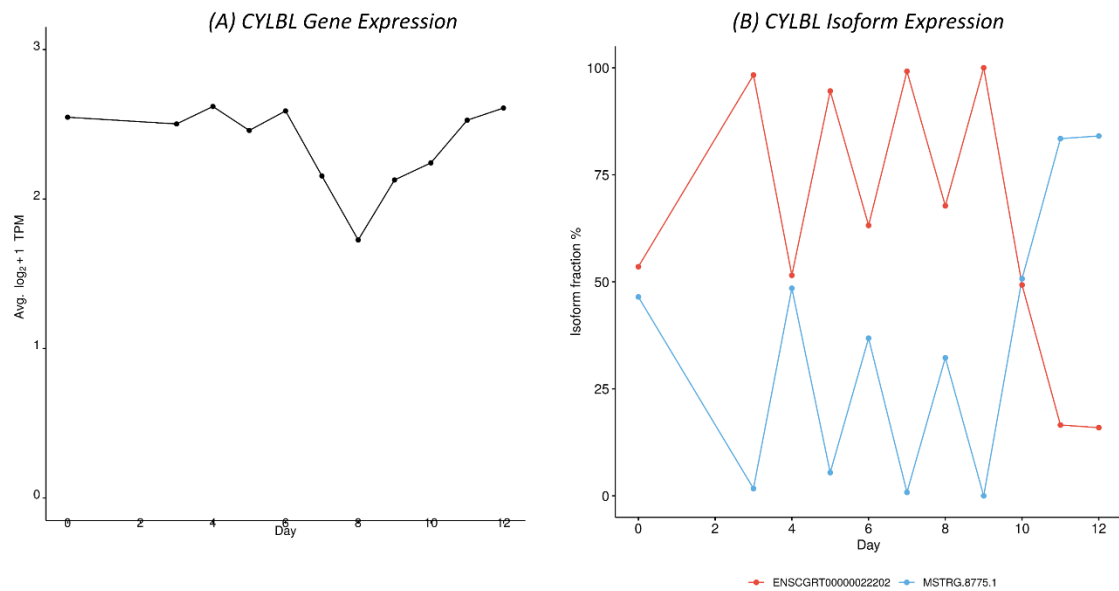


Figure 4.9 (A) Gene abundance over time for the CYLBL gene. A 1.54 fold increase in gene expression occurred between the stationary and death phase (B) Normalized transcript abundances of the two dominant CYLBL component isoforms. The novel isoform MSTRG.8775.1 is periodically depleted every two days as the other isoform is slightly upregulated. The novel isoform becomes dominant during the death phase after day 10.

337 mRNA transcripts exhibited isoform switching independently of gene differential expression with an equivalent number of upregulated (168) and downregulated (169) transcripts. **Table 4.5** summarizes the top ten transcripts that have a significant change in isoform fraction between the stationary and death phases.

Table 4.5 A table of significantly differentially transcribed isoforms that were not differentially expressed at the gene level between stationary and exponential phase.

Gene	Isoform ID	dIF	Isoform Adj. P-value
<i>SYNE3</i>	<i>MSTRG.14567.1</i>	-0.62	4.03×10^{-3}
<i>SYNE3</i>	<i>MSTRG.14567.2</i>	-0.62	4.03×10^{-3}
<i>RPL10</i>	<i>ENSCGRT00000003266</i>	0.64	5.28×10^{-5}
<i>RPL10</i>	<i>MSTRG.18569.1</i>	-0.64	5.28×10^{-5}
<i>ZNF609</i>	<i>MSTRG.19775.1</i>	-0.78	1.97×10^{-4}
<i>ZNF609</i>	<i>MSTRG.19775.2</i>	0.78	1.94×10^{-4}
<i>TRMT6</i>	<i>MSTRG.386.2</i>	0.65	2.50×10^{-5}
<i>EXO1</i>	<i>ENSCGRT00000002852</i>	0.75	3.51×10^{-6}
<i>EXO1</i>	<i>MSTRG.5493.2</i>	-0.75	1.11×10^{-6}
<i>WWC2</i>	<i>MSTRG.608.1</i>	0.68	1.52×10^{-4}

The *ZNF609* gene has the greatest difference in isoform fractions between the transcripts. *ZNF609* had two novel transcripts, *MSTRG.19775.1* and *MSTRG.19775.2*, and a previously annotated transcript *ENSCGRT00000002521* which was not expressed. The two novel isoforms exhibited daily switching (**Figure 4.10**). Two novel isoforms for the *Citrate synthase* gene were the only transcripts expressed. The first novel isoform, *MSTRG.12084.1* accounted for almost 100% of CS gene expression during the exponential phase while the second isoform, *MSTRG.12084.2*, accounted for the majority of expression during the stationary phase (**Figure 4.7**).

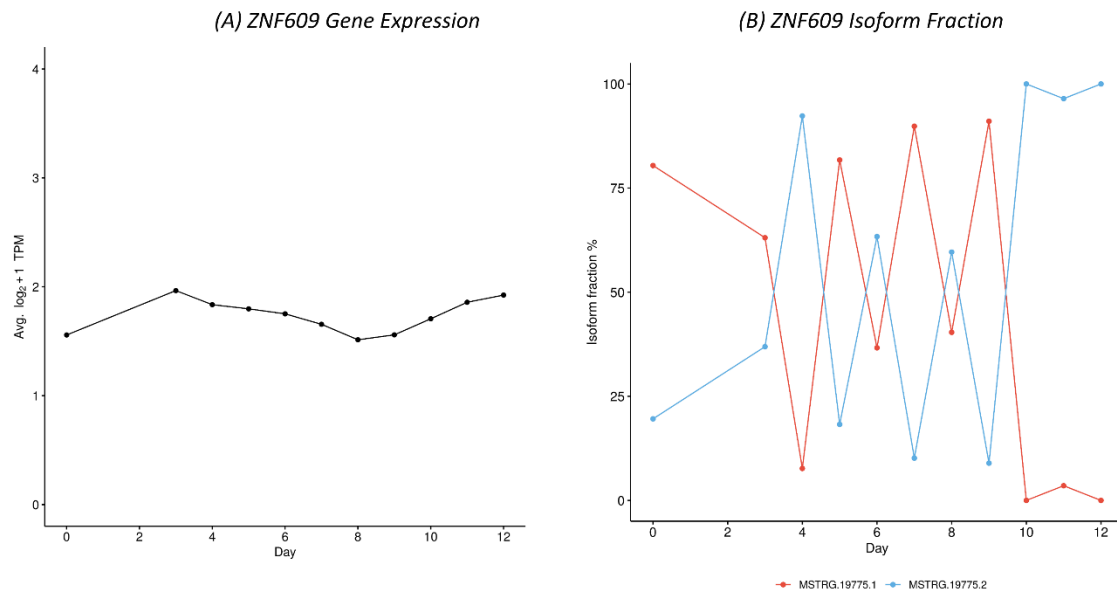


Figure 4.10 Gene abundance over time for the ZNF609 gene which was seen to be expressed at a stable abundance over time and was not significantly differentially expressed at any stage. 100% of gene expression can be attributed to two novel isoforms: MSTRG.19775.1 and MSTRG.19775.2. **(B)** The two component isoforms were observed to have time specific fluctuations in isoform fractions as both transcripts have anti-correlated expression. MSTRG.19775.2 is expressed at 100% of the total fraction in the death phase.

Supplementary Table 5.2 summarizes all transcripts that had significant isoform switching events. Gene ontology enrichment of biological processes was performed using *DAVID* functional annotation suite with 7 processes significantly enriched below a FDR threshold of 0.05. **Table 4.6** details all enriched transcripts that undergo isoform switching between the stationary and death phases. Protein coding genes that are involved in producing and metabolizing post-translational modification processes and ncRNA synthesis were significantly enriched during the death phase.

Table 4.6 Gene ontology enrichment for cellular processes for genes that undergo isoform switching without a change in expression between the stationary and death phases.

GO ID	GO Term	Count	FDR
GO:0006464	Cellular protein modification process	187	3.93×10^{-4}
GO:0036211	Protein modification process	187	3.93×10^{-4}
GO:0010467	Gene expression	256	2.47×10^{-3}
GO:0034660	ncRNA metabolic process	38	8.5×10^{-3}
GO:0006396	RNA processing	55	0.017
GO:0016070	RNA metabolic process	213	0.029
GO:0044085	Cellular component biogenesis	141	0.045

4.3.6 Alternative splicing events between stationary and death phases

All transcripts that had significant changes in isoform fraction during the death phase had respective alternative splicing events classified. Significantly switching transcripts had multiple alternative splicing events with an average of 2 events per transcript similar to the difference between exponential and stationary phases. **Figure 4.11** summarizes the number of alternative splicing events in transcripts that underwent isoform switching. The three most common events were alternative transcription and termination sites as well as exon skipping.

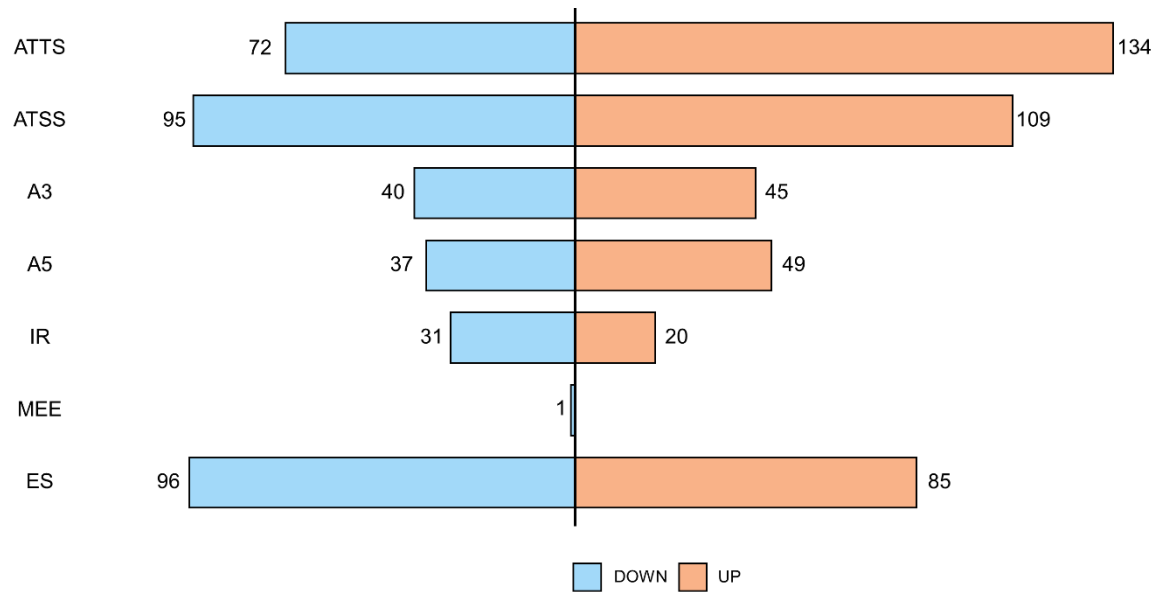


Figure 4.11 A leaf plot showing the number of alternatively spliced transcripts that underwent significant isoform switching between exponential and stationary phase.

Figure 4.12 shows an example of an exon skipping event between component transcripts of the tri isoform *Paralemmmin* (*PALM*) gene. *ENSCGRT00000014191*, which lacks exon 8, is skipped in the stationary phase until the death phase where a novel isoform containing this exon, *MSTRG.8217.1*, becomes dominant (dIF: 0.23, 8.67×10^{-3}). The inclusion of exon 8 in the translated *PALM* gene adds a 13 amino acid residue to the *PALM* protein during the death phase.

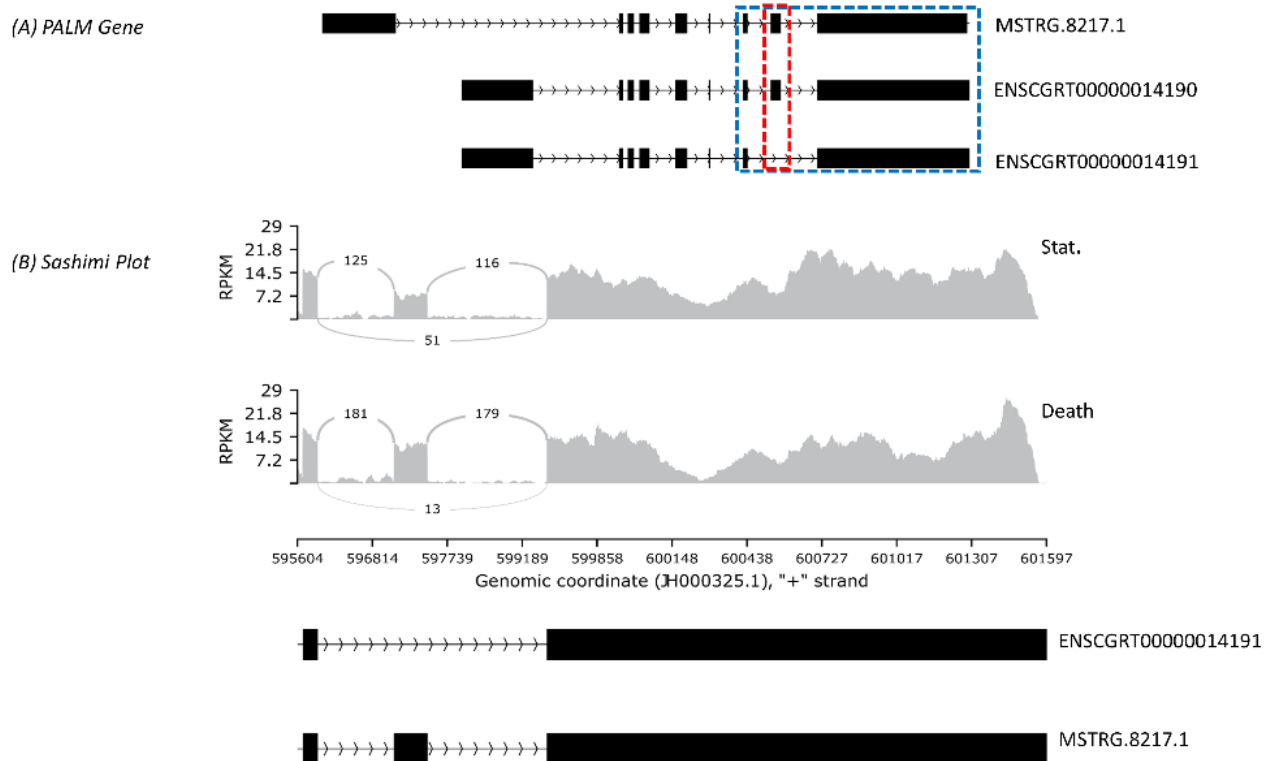


Figure 4.12 (A) Transcript structure of the PALM gene which has one novel isoform and two previously annotated transcripts. The region highlighted in blue is magnified in plot **(B)** which is a sashimi plot of the last three exons of the PALM locus. Only the switching genes are depicted with MSTRG.8217.1, which contains the skipped exon, increasing 23%.

4.3.7 Alternative splicing of mRNAs over time

Transcripts that were only differentially expressed at the transcript level were compared to find a cohort of mRNAs that were alternatively spliced during all three phases of cell culture. 35 transcripts were alternatively spliced between exponential, stationary, and death phases in fed-batch culture over time (**Table 4.7**). 10 genes contained transcripts that were reciprocally alternatively spliced with 6 pairs undergoing anti-correlated switching where the one isoform was upregulated during the exponential phase and downregulated during the death phase.

Preferential transcription of an intron retaining novel isoform of the *ubiquitin associated protein 2-like* gene (*UBA2L*) gene was observed during the exponential and stationary phase. The longer, mRNA encoding isoform containing multiple exons is transcribed reciprocally to the mono-exonic, shorter sequence during the stationary phase (**Figure 4.13**). The longer isoform,

ENSCGRT00000023929, contains the full-length coding sequence for the *UBA2CL* protein while the shorter, mono-exonic variant contains a partial 37 amino acid sequences. The mono-exonic isoform has a 49% decrease in isoform fraction during the stationary phase (Adj. P-value: 2.07×10^{-3}) but a 43% increase during the death phase (2.80×10^{-2}).

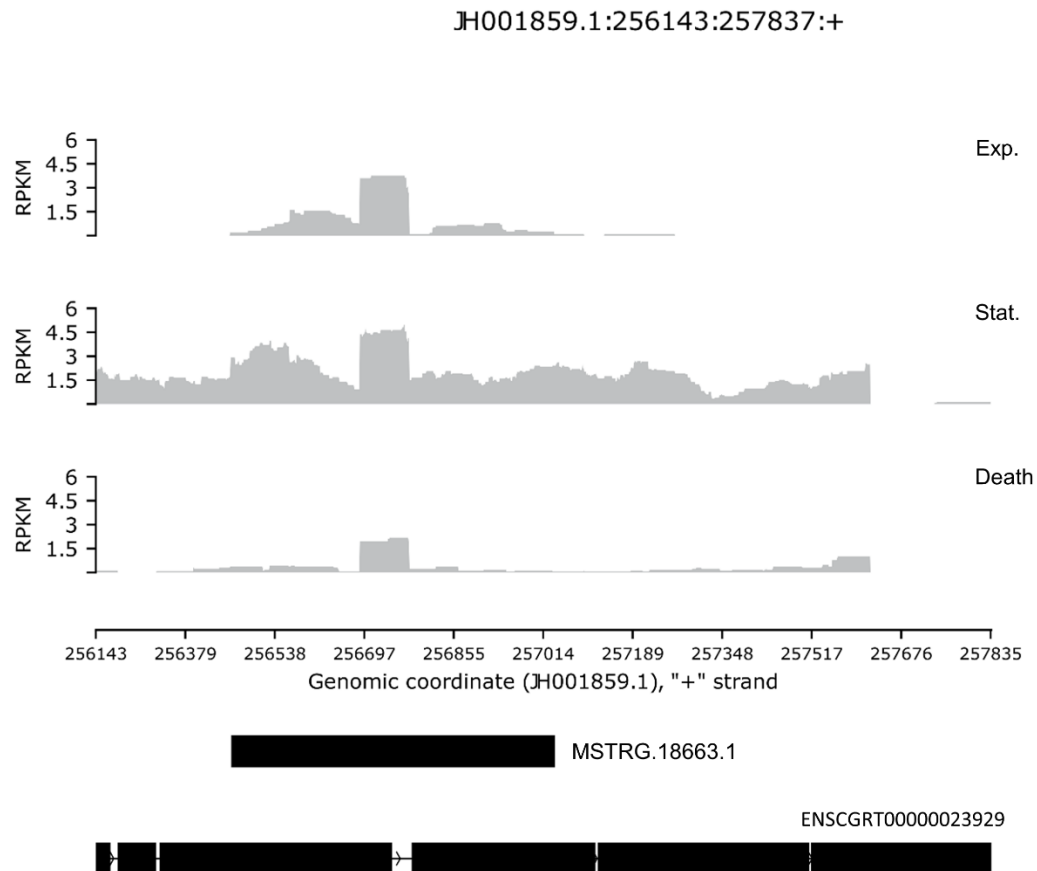


Figure 4.13 A sashimi plot of the normalized coverage of the *UBA2CL* gene which contains two isoforms: MSTRG.18663.1 and ENSCGRT00000023929. The longer isoform contains the full length CDS sequence of the *UBA2CL* protein and is preferentially expressed during the stationary phase but is quiescent during the death phase.

Table 4.7 Transcripts that are alternatively spliced between exponential, stationary, and death phase.

Gene	Isoform ID	Exp. Vs Stat dIF	Exp. Vs. Stat Adj. P-value	Stat. Vs. Death dIF	Stat. Vs. Death Adj. P-value
<i>MAT2A</i>	<i>MSTRG.11551.3</i>	0.29	9.80×10^{-5}	-0.29	2.32×10^{-3}
<i>MAT2A</i>	<i>MSTRG.11551.2</i>	-0.38	2.58×10^{-4}	0.29	3.15×10^{-2}
<i>ENSCGRG00000013966</i>	<i>MSTRG.12084.1</i>	-0.73	1.14×10^{-4}	0.60	3.77×10^{-2}
<i>UNK</i>	<i>ENSCGRT00000012042</i>	0.15	5.64×10^{-3}	0.19	6.86×10^{-3}
<i>UNK</i>	<i>MSTRG.12805.1</i>	-0.15	5.61×10^{-3}	-0.19	6.90×10^{-3}
<i>ENSCGRG00000007393</i>	<i>ENSCGRT00000010302</i>	0.11	3.88×10^{-2}	-0.11	2.40×10^{-2}
<i>ENSCGRG00000007393</i>	<i>MSTRG.13875.1</i>	-0.11	3.54×10^{-2}	0.11	2.19×10^{-2}
<i>EHMT2</i>	<i>MSTRG.14619.7</i>	0.44	6.87×10^{-3}	-0.51	2.98×10^{-3}
<i>PMPCA</i>	<i>MSTRG.17031.1</i>	0.14	8.27×10^{-6}	0.12	2.12×10^{-2}
<i>UBA2CL</i>	<i>ENSCGRT00000023929</i>	0.49	2.07×10^{-3}	-0.43	2.80×10^{-2}
<i>UBA2CL</i>	<i>MSTRG.18663.2</i>	-0.49	2.05×10^{-3}	0.43	2.80×10^{-2}
<i>PTGES</i>	<i>MSTRG.20158.3</i>	-0.21	6.66×10^{-3}	0.21	4.13×10^{-3}
<i>PTGES</i>	<i>MSTRG.20158.2</i>	0.23	9.87×10^{-6}	-0.26	1.27×10^{-3}
<i>NXF1</i>	<i>ENSCGRT00000023495</i>	0.18	3.10×10^{-3}	-0.12	7.30×10^{-3}
<i>TMEM223</i>	<i>ENSCGRT00000023495</i>	0.18	3.10×10^{-3}	-0.12	7.30×10^{-3}
<i>NXF1</i>	<i>MSTRG.20500.1</i>	-0.15	3.90×10^{-3}	0.11	5.29×10^{-3}
<i>TMEM223</i>	<i>MSTRG.20500.1</i>	-0.15	3.90×10^{-3}	0.11	5.29×10^{-3}
<i>PLA2G4A</i>	<i>MSTRG.2054.2</i>	0.18	4.22×10^{-2}	-0.20	1.33×10^{-3}
<i>PLA2G4A</i>	<i>ENSCGRT00000020344</i>	-0.31	1.06×10^{-4}	0.29	5.29×10^{-5}
<i>PEX16</i>	<i>ENSCGRT00000009061</i>	0.19	4.07×10^{-2}	0.16	1.02×10^{-2}
<i>1700029I15RIK</i>	<i>ENSCGRT00000009061</i>	0.19	4.07×10^{-2}	0.16	1.02×10^{-2}
<i>PEX16</i>	<i>ENSCGRT00000009240</i>	-0.19	4.02×10^{-2}	-0.15	1.01×10^{-2}
<i>1700029I15RIK</i>	<i>ENSCGRT00000009240</i>	-0.19	4.02×10^{-2}	-0.15	1.01×10^{-2}
<i>RAD51B</i>	<i>MSTRG.3748.6</i>	-0.27	9.68×10^{-3}	-0.16	2.80×10^{-2}
<i>CLCN3</i>	<i>MSTRG.4341.5</i>	0.28	3.22×10^{-2}	-0.36	3.14×10^{-2}
<i>PPP1R36</i>	<i>MSTRG.4951.2</i>	-0.16	1.34×10^{-2}	0.16	1.00×10^{-2}
<i>PLEKHG3</i>	<i>MSTRG.4952.2</i>	-0.45	5.99×10^{-3}	0.45	2.62×10^{-2}
<i>CWC25</i>	<i>ENSCGRT00000007104</i>	0.17	4.24×10^{-4}	0.19	4.62×10^{-5}
<i>CWC25</i>	<i>MSTRG.5218.2</i>	-0.17	4.24×10^{-4}	-0.19	4.66×10^{-5}
<i>ENSCGRG00000009770</i>	<i>MSTRG.6547.2</i>	0.22	8.54×10^{-4}	-0.21	1.93×10^{-2}
<i>ENSCGRG00000010536</i>	<i>MSTRG.6644.3</i>	0.35	2.02×10^{-2}	-0.21	2.76×10^{-3}
<i>ENSCGRG00000012304</i>	<i>ENSCGRT00000017163</i>	-0.10	2.83×10^{-2}	-0.14	4.03×10^{-4}
<i>TBXAS1</i>	<i>ENSCGRT00000024920</i>	0.17	1.98×10^{-6}	0.35	8.05×10^{-4}
<i>TBXAS1</i>	<i>MSTRG.8693.2</i>	-0.17	1.32×10^{-6}	-0.35	7.82×10^{-4}

4.3.8 lncRNAs undergoing isoform switching between exponential and stationary phase

277 ncRNA transcripts were observed to undergo isoform switching between exponential and stationary phases with 143 isoforms transcribed from genes that are differentially expressed and 134 isoforms differentially expressed at the transcript level but not at the gene level between exponential and stationary phase. 125 lncRNAs found within or near a gene body and the remaining found in intergenic loci (**Figure 4.15**). Intergenic loci had the greatest number of isoforms switching between phases with the downstream sense and antisense as well as antisense overlapping transcripts making up the remainder. 142 transcripts were upregulated and 135 transcripts downregulated between the two cell culture stages.

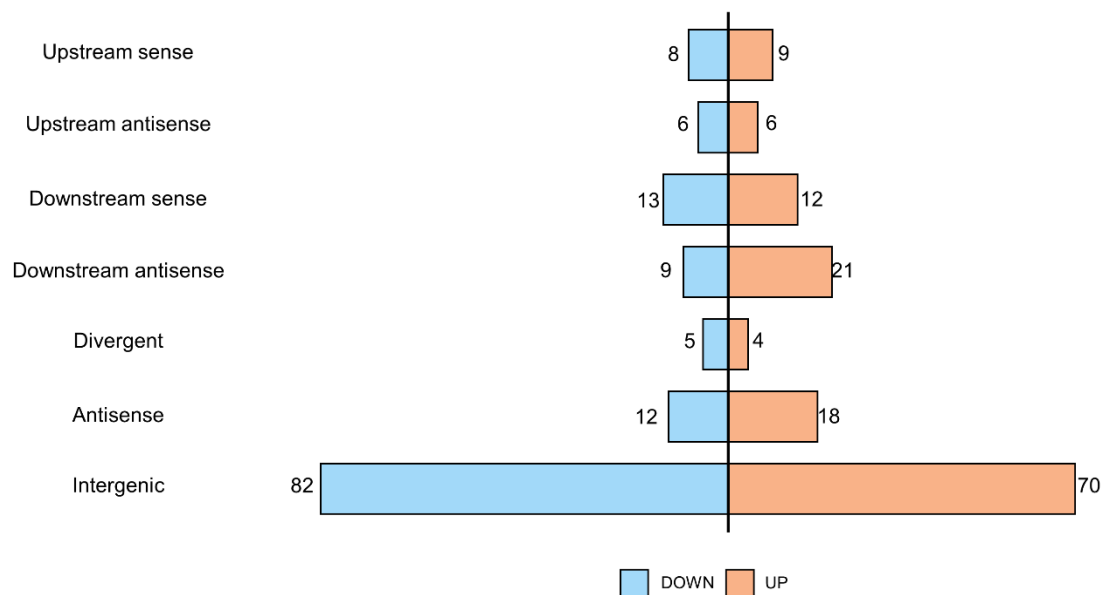


Figure 4.14 277 transcripts were differentially expressed between the exponential and stationary phases with the majority belonging to the intergenic class of lncRNAs.

Table 4.9 summarizes the top ten lncRNA genes that are simultaneously differentially expressed at the isoform level as well as the gene level with 5 gene pairs having reciprocal isoform switching. *MSTRG.13823*, a novel antisense lncRNA found within the *GABPRP1* coding gene, is the most differentially expressed gene with its two component isoforms switching during the stationary

phase. All of these loci are expressed at significantly lower abundances compared to mRNA isoforms, however.

Table 4.8 lncRNA genes that are differentially expressed that also undergo alternative splicing at the transcript level

Gene	Gene log2FC	Gene Adj. P-value	Isoform ID	dIF	Isoform Adj. P-value	Class
<i>MSTRG.14216</i>	4.11	2.68×10^{-9}	<i>MSTRG.14216.2</i>	-0.75	1.24×10^{-22}	Downstream sense
<i>MSTRG.13823</i>	3.80	7.23×10^{-8}	<i>MSTRG.13823.1</i>	0.72	2.71×10^{-2}	Antisense
<i>MSTRG.13823</i>	3.80	7.23×10^{-8}	<i>MSTRG.13823.2</i>	-0.72	3.01×10^{-2}	Antisense
<i>MSTRG.13512</i>	2.74	2.43×10^{-11}	<i>MSTRG.13512.2</i>	0.63	6.05×10^{-4}	Intergenic
<i>MSTRG.13512</i>	2.74	2.43×10^{-11}	<i>MSTRG.13512.1</i>	-0.63	6.10×10^{-4}	Intergenic
<i>MSTRG.13821</i>	5.54	1.70×10^{-10}	<i>MSTRG.13821.2</i>	-0.63	3.72×10^{-4}	Upstream antisense
<i>MSTRG.15589</i>	3.61	5.93×10^{-23}	<i>MSTRG.15589.1</i>	-0.58	7.12×10^{-5}	Intergenic
<i>MSTRG.15589</i>	3.61	5.93×10^{-23}	<i>MSTRG.15589.2</i>	0.58	6.87×10^{-5}	Downstream antisense
<i>MSTRG.13438</i>	5.22	8.15×10^{-7}	<i>ENSCGRT00000034148</i>	-0.54	2.80×10^{-2}	Upstream sense
<i>MSTRG.13438</i>	5.22	8.15×10^{-7}	<i>MSTRG.13438.1</i>	0.54	2.52×10^{-2}	Intergenic

48% of lncRNA transcripts were differentially expressed at the transcript level independently while the expression at the gene level remained stable between exponential and stationary phase.

Table 4.10 displays top ten transcripts that have a significant change in isoform fraction between the exponential and stationary phase with four genes having reciprocal switching between phases.

Table 4.9 lncRNA transcripts that undergo isoform switching without any change at the gene level.

Isoform ID	dIF	Isoform Adj. P-value	Class
<i>ENSCGRT00000033861</i>	-0.70	3.60×10^{-2}	Intergenic
<i>MSTRG.10133.1</i>	0.70	2.68×10^{-2}	Intergenic
<i>MSTRG.6466.3</i>	-0.65	2.05×10^{-4}	Intergenic
<i>MSTRG.13625.7</i>	0.52	3.95×10^{-13}	Intergenic
<i>MSTRG.22547.1</i>	-0.50	1.78×10^{-2}	Intergenic
<i>MSTRG.22547.2</i>	0.50	1.76×10^{-2}	Intergenic
<i>MSTRG.7929.1</i>	0.50	3.10×10^{-2}	Upstream antisense
<i>MSTRG.7929.2</i>	-0.50	3.10×10^{-2}	Intergenic
<i>MSTRG.4956.3</i>	0.41	3.75×10^{-4}	Intergenic
<i>MSTRG.1924.2</i>	0.41	1.74×10^{-9}	Intergenic

4.3.9 lncRNA alternative splicing between exponential and stationary phase

579 alternative splicing events were identified in all lncRNAs isoforms that had a change in expression between exponential and stationary phase (**Figure 4.16**). There was only a single mutually exclusive exon event which was upregulated in the stationary phase. Exon skipping and alternative transcription start and termination sites were the most abundant. The distribution of alternative splicing events was similar between lncRNAs and mRNAs presented in **Section 4.3.6**. The complete summary of all alternative splicing events for each significant isoform switch is found in **Supplementary Table 4.2**.

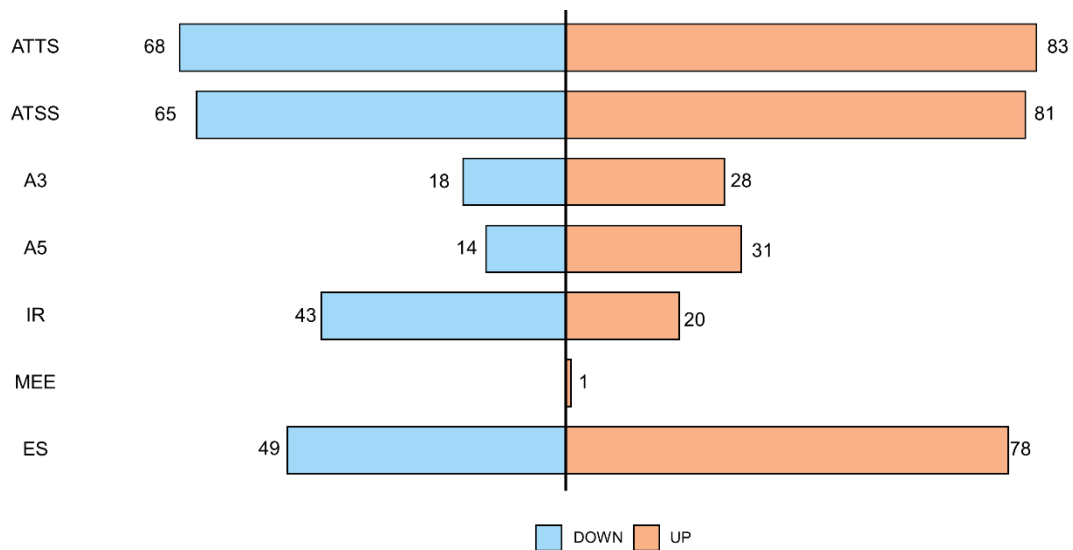


Figure 4.15 A leaf plot of significant alternative splicing events between exponential and stationary phase transcripts.

Figure 4.17 illustrates an example of an intergenic lncRNA, *ENSCGRG00000025089*, that exhibits an exon skipping event between exponential and stationary phases. This lncRNA loci is composed of 4 isoforms with only the novel *MSTRG.5005.1* and *MSTRG.5005.2* isoforms expressed. *ENSCGRT00000032613*, an isoform that was previously annotated within *ENSEMBL* is not expressed. The novel isoforms are both expressed at equal abundances during the lag and exponential phases but *MSTRG.5005.3* becomes the dominant isoform during the stationary phase.

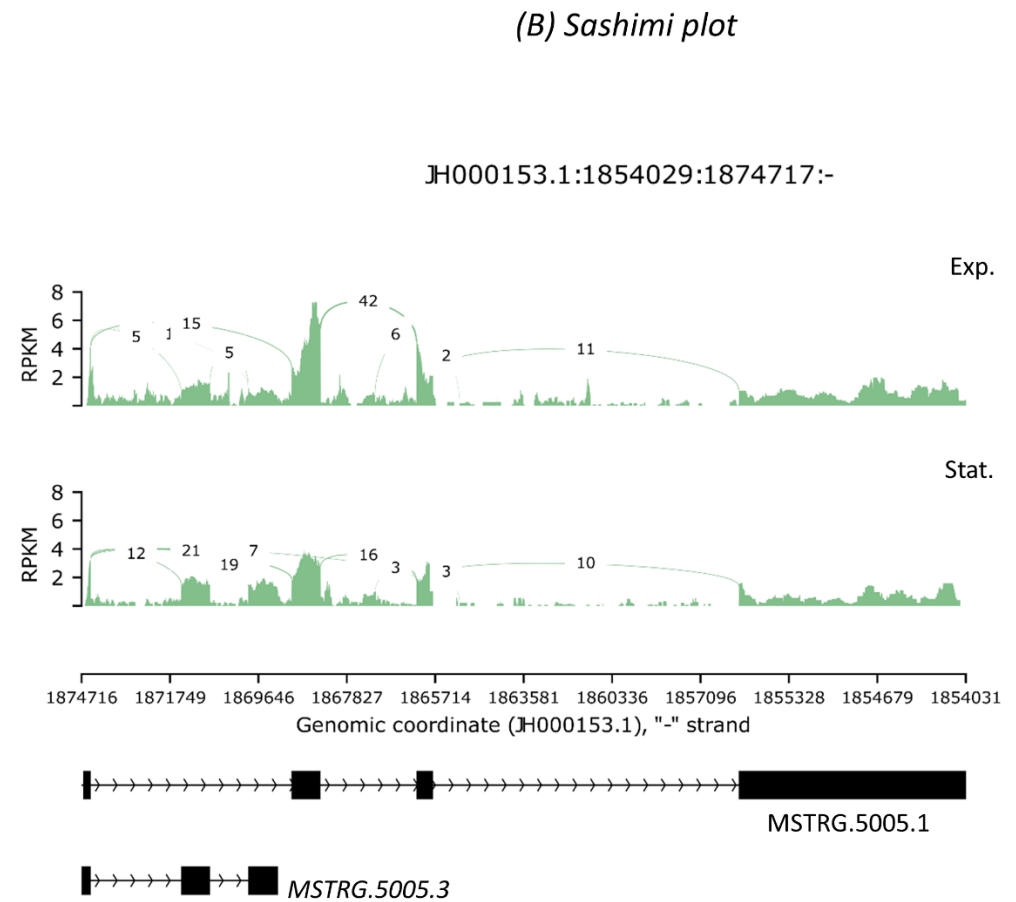
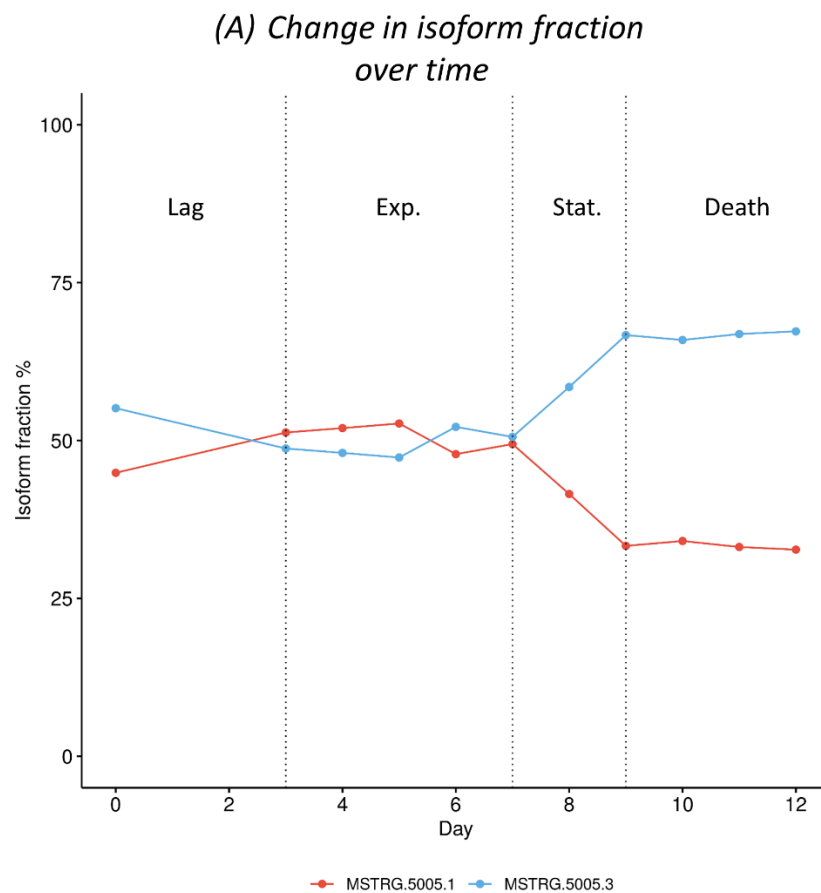


Figure 4.16 (A) Change in isoform fraction over time for the intergenic ENSCGRG00000025089 locus which was previously annotated within ENSEMBL. The shorter, novel isoform MSTRG.5005.3 becomes the dominant isoform at the start of the stationary phase (dIF= 13.2% ,Adj. P-value: 0.026). (B) A sashimi plot of the change in expression and splicing of the two isoforms. The multi-exonic transcript MSTRG.5005.3 becomes dominant.

Secondary structures of *MSTRG.5005.1* and *MSTRG.5005.3* were predicted using the *RNAfold* web server utility. **Figure 4.17** shows the secondary structures of the two isoforms. *MSTRG.5005.1*, which is expressed at a higher abundance during the exponential phase contains stem-loop motifs that are absent in *MSTRG.5005.3*.



Figure 4.17 RNA structure predicted by RNAfold for the longer (A) *MSTRG.5005.1* and (B) shorter *MSTRG.5005.3* transcripts. The longer isoform has a triangular shape that is lost with the switch to the shorter isoform.

4.3.10 lncRNA isoform switching between stationary and death phase

139 ncRNA transcripts were observed to perform isoform switching between stationary and death phases with 37 isoforms transcribed from genes that are differentially expressed and 102 isoforms differentially expressed at the transcript level but not at the gene level between exponential and stationary phase. 75 lncRNAs were found within or near a gene body and the remaining found in intergenic loci (**Figure 4.18**). Intergenic loci had the greatest number of isoforms switching between phases with the downstream sense and antisense as well as antisense overlapping transcripts making up the remainder. **Table 4.11** details the top 10 differentially expressed genes with transcripts that also have altered expression between stationary and death phases.

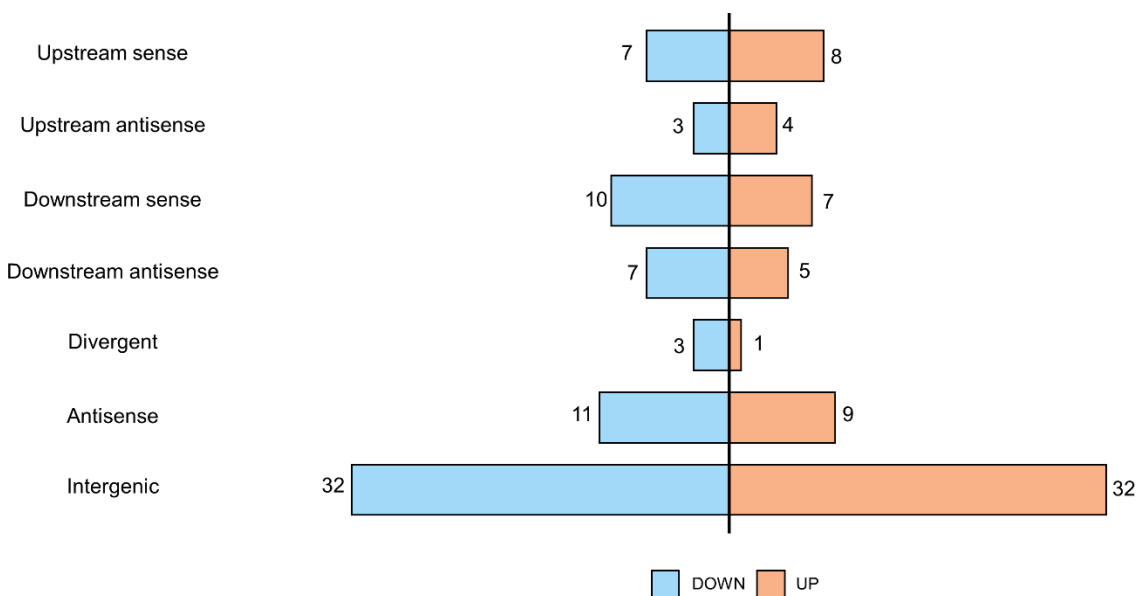


Figure 4.18. 139 transcripts were differentially expressed between the stationary and death phases with the majority belonging to the intergenic class of lncRNAs.

There are three genes that have reciprocal switching between the component isoforms. All three loci are dual isoform genes and are expressed at lower abundances than protein coding genes.

Table 4.10 lncRNA genes that are differentially expressed that also undergo alternative splicing at the transcript level between stationary and death phases

Gene	Gene log2FC	Gene Adj. P-value	Isoform ID	dIF	Isoform Adj. P-value	Class
<i>MSTRG.4831</i>	-3.24	4.86×10^{-24}	<i>MSTRG.4831.1</i>	0.43	4.30×10^{-2}	Downstream antisense
<i>MSTRG.9034</i>	1.22	5.32×10^{-10}	<i>MSTRG.9034.3</i>	-0.34	2.44×10^{-6}	Antisense
<i>MSTRG.13503</i>	1.75	9.43×10^{-8}	<i>MSTRG.13503.3</i>	0.34	3.20×10^{-2}	Upstream sense
<i>MSTRG.20473</i>	-0.59	1.49×10^{-3}	<i>MSTRG.20473.1</i>	0.28	1.19×10^{-2}	Downstream antisense
<i>MSTRG.20473</i>	-0.59	1.49×10^{-3}	<i>MSTRG.20473.2</i>	-0.28	1.22×10^{-2}	Downstream antisense
<i>MSTRG.13074</i>	-0.76	4.85×10^{-8}	<i>MSTRG.13074.3</i>	-0.26	3.87×10^{-2}	Downstream antisense
<i>MSTRG.5502</i>	-1.00	3.62×10^{-4}	<i>MSTRG.5502.3</i>	0.25	2.74×10^{-2}	Intergenic
<i>MSTRG.3020</i>	-0.91	1.37×10^{-7}	<i>MSTRG.3020.1</i>	0.23	3.38×10^{-2}	Upstream antisense
<i>MSTRG.3020</i>	-0.91	1.37×10^{-7}	<i>MSTRG.3020.2</i>	-0.23	3.42×10^{-2}	Intergenic
<i>MSTRG.9839</i>	-0.65	1.01×10^{-3}	<i>MSTRG.9839.1</i>	-0.23	3.05×10^{-2}	Intergenic

73% of lncRNA transcripts were differentially expressed at the transcript level independently while the expression at the gene level remained stable between stationary and death phase. **Table 4.12** displays top ten transcripts that have a significant change in isoform fraction with five genes having reciprocal switching between phases. *ENSCGRT00000033332* and *ENSCGRT00000033333* dual isoforms of the same locus and alter expression reciprocally.

Table 4.11 lncRNA transcripts that undergo isoform switching without any change at the gene level.

Isoform ID	dIF	Isoform Adj. P-value	Class
<i>ENSCGRT00000033332</i>	0.78	6.92×10^{-3}	Downstream sense
<i>ENSCGRT00000033333</i>	-0.78	6.92×10^{-3}	Downstream
<i>MSTRG.3418.1</i>	0.58	4.42×10^{-2}	Intergenic
<i>MSTRG.3418.2</i>	-0.58	4.42×10^{-2}	Downstream antisense
<i>MSTRG.16737.1</i>	-0.55	2.05×10^{-2}	Intergenic
<i>MSTRG.16737.2</i>	0.48	2.82×10^{-2}	Downstream antisense
<i>MSTRG.3563.3</i>	0.46	3.97×10^{-3}	Upstream sense
<i>MSTRG.9594.2</i>	0.43	3.95×10^{-2}	Intergenic
<i>MSTRG.23471.1</i>	0.41	2.26×10^{-3}	Intergenic
<i>MSTRG.23471.2</i>	-0.41	2.29×10^{-3}	Intergenic

4.3.11 lncRNA alternative splicing events between stationary and death phases

265 alternative splicing events were identified in all lncRNAs isoforms that had a change in expression between stationary and death phases (**Figure 4.20**). There were no mutually exclusive exon events in death phase enriched transcripts. Exon skipping and alternative transcription start and termination sites were the most abundant. The complete summary of all alternative splicing events for each significant isoform switch is found in **Supplementary Table 4.2**.

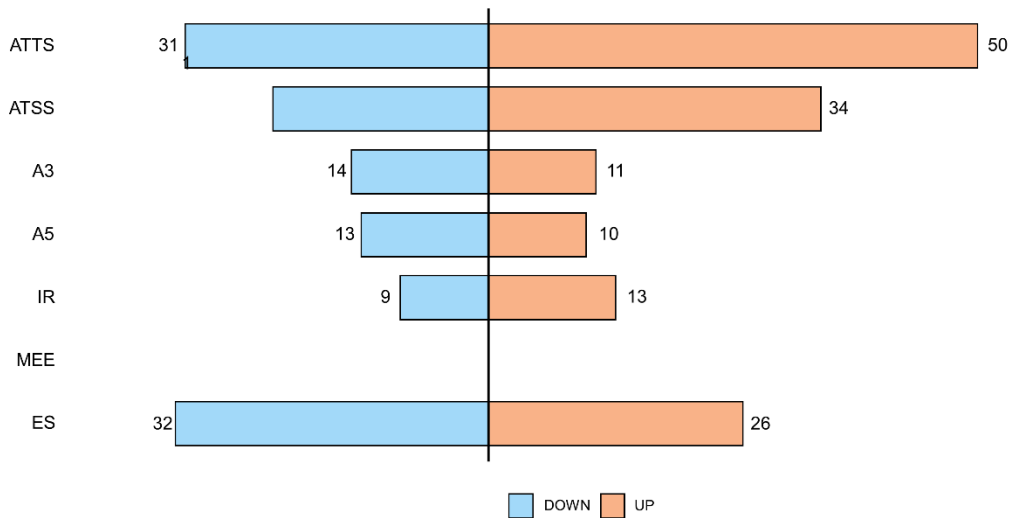


Figure 4.19 A leaf plot of significant alternative splicing events between stationary and death phase transcripts.

Figure 4.20 illustrates an example of an exon skipping event in an antisense gene locus, *MSTRG.16574*, to the *Glutathione S-Transferase Alpha 2* gene (*GSTA2*). *MSTRG.16574* is composed of five isoforms where only two isoforms, *MSTRG.16574.1* and *MSTRG.16574.3*, alternative spliced while there was no observed change in the remaining three. *MSTRG.16574.1*, which is missing exon 2, became the dominant isoform accounting for ~ 65% of all isoforms. The gene abundance remained constant between stationary and death phases. **Figure 4.21** displays the predicted secondary structure of the RNA molecule that contains exon 2 and lacks exon 2.

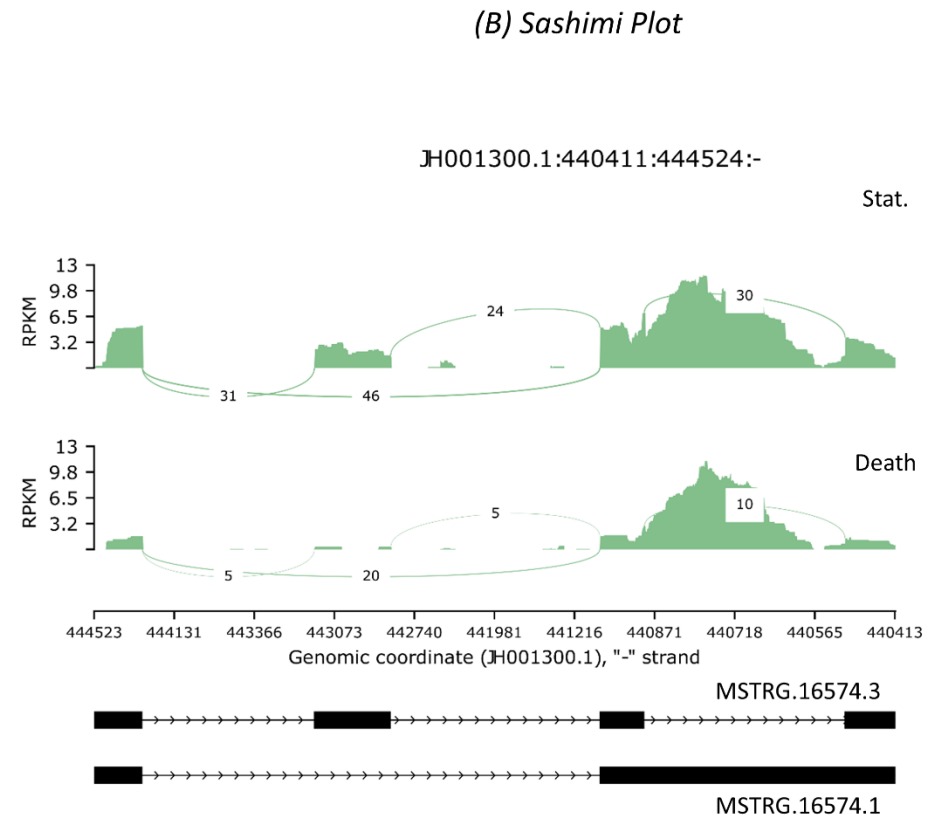
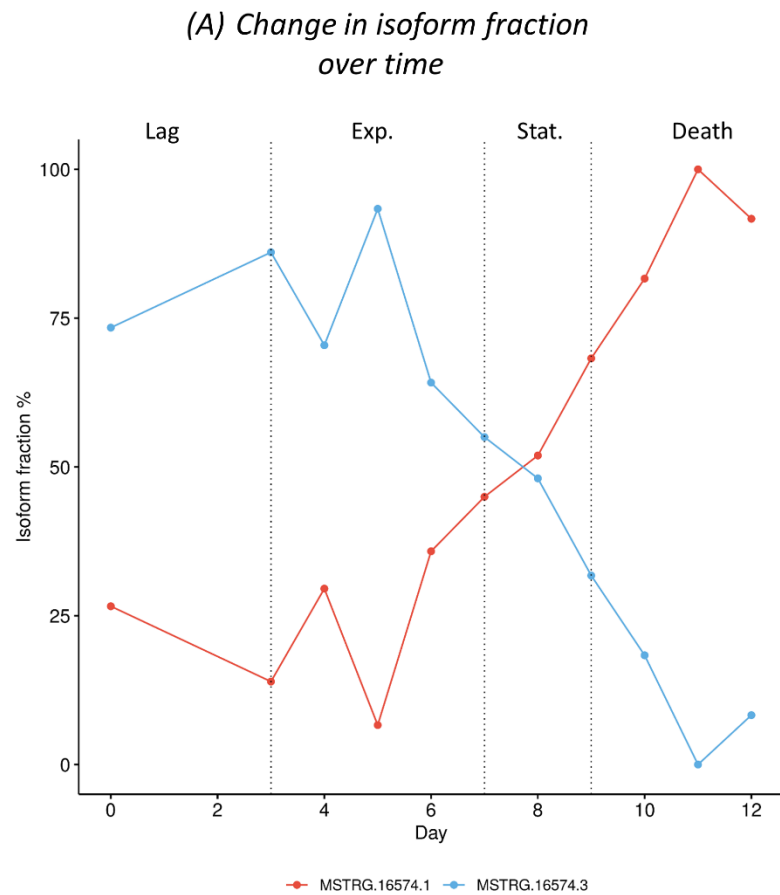


Figure 4.20 (A) The change in isoform fraction over time for transcripts MSTRG.16574.1 and MSTRG.16574.3. The isoform fractions of the other three transcripts have been omitted for clarity but remain constant over time as the proportion of these isoform fractions change over time. (B) A sashimi plot of coverage and splicing of the MSTRG.16574 non-coding RNA locus. Exon 2 is completely depleted in the death phase

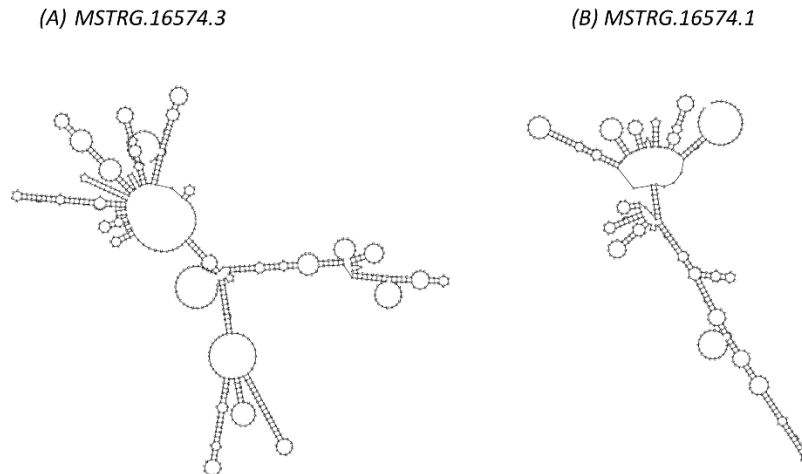


Figure 4.21 RNA structure predicted by RNAfold for the longer **(A)** *MSTRG.16574.3* and **(B)** shorter *MSTRG.16574.1* transcripts. The longer isoform contains a circular structure with stem-loops that is lost in *MSTRG.16574.3* which is missing the second exon.

4.3.12 lncRNA alternative splicing over the duration of fed-batch culture

18 long non-coding RNA transcripts underwent alternative splicing during all three phases of the culture between exponential, stationary, and death phase. **Table 4.13** contains all isoforms that have altered expression during all three cell culture stages. There are five pairs of transcripts within this group that have reciprocal differences in isoform fraction. All of these transcripts are found inside genes containing two isoforms. The novel, intergenic gene *MSTRG.13357*, is a member of this cohort and was observed to have an A5 splicing event where an alternative, shorter second exon becomes dominant during the stationary phase (dIF: 19%, Adj. P-value: 3.52×10^{-5}). The initial, longer isoform becomes dominant again at the start of the death phase (dIF: 14%, 3.90×10^{-2}). The change in isoform fraction for the switching isoforms and a sashimi plot is shown in **Figure 4.22**.

Table 4.12 Table of transcripts that switch abundances in all three stages of cell culture. The majority of these transcripts are located within intergenic regions.

Isoform ID	Exp vs. Stat dIF	Exp vs. Stat Adj. P-value	Stat vs. Death dIF	Stat vs. Death Adj. P-value	Class
<i>ENSCGRT00000032776</i>	-0.14	8.41×10^{-3}	0.40	4.62×10^{-2}	Intergenic
<i>ENSCGRT00000033732</i>	-0.15	2.57×10^{-3}	0.16	4.41×10^{-9}	Intergenic
<i>MSTRG.10407.6</i>	0.21	7.66×10^{-4}	-0.13	5.01×10^{-3}	Downstream sense
<i>MSTRG.12785.2</i>	0.15	2.15×10^{-3}	-0.16	3.69×10^{-9}	Intergenic
<i>MSTRG.13357.1</i>	-0.19	3.51×10^{-5}	0.14	3.90×10^{-2}	Intergenic
<i>MSTRG.13357.2</i>	0.19	3.52×10^{-5}	-0.14	3.90×10^{-2}	Intergenic
<i>MSTRG.13859.1</i>	-0.29	4.30×10^{-3}	-0.26	9.68×10^{-4}	Upstream sense
<i>MSTRG.13859.2</i>	0.29	4.59×10^{-3}	0.26	9.68×10^{-4}	Upstream sense
<i>MSTRG.19195.1</i>	0.14	1.16×10^{-2}	-0.27	1.37×10^{-5}	Intergenic
<i>MSTRG.19195.2</i>	-0.14	1.21×10^{-2}	0.27	1.42×10^{-5}	Upstream sense
<i>MSTRG.21040.2</i>	0.24	4.58×10^{-2}	-0.27	4.30×10^{-2}	Intergenic
<i>MSTRG.21201.1</i>	-0.25	5.57×10^{-4}	0.20	1.83×10^{-2}	Upstream antisense
<i>MSTRG.21201.2</i>	0.25	5.58×10^{-4}	-0.20	1.83×10^{-2}	Upstream antisense
<i>MSTRG.2926.1</i>	0.26	1.45×10^{-2}	-0.32	1.34×10^{-7}	Upstream sense
<i>MSTRG.2926.2</i>	-0.26	1.48×10^{-2}	0.32	1.67×10^{-7}	Upstream sense
<i>MSTRG.4911.3</i>	-0.15	2.15×10^{-2}	-0.11	3.72×10^{-3}	Downstream sense
<i>MSTRG.5714.3</i>	-0.16	4.99×10^{-4}	0.13	7.60×10^{-3}	Intergenic
<i>MSTRG.7482.3</i>	0.15	1.72×10^{-3}	-0.16	9.63×10^{-8}	Antisense

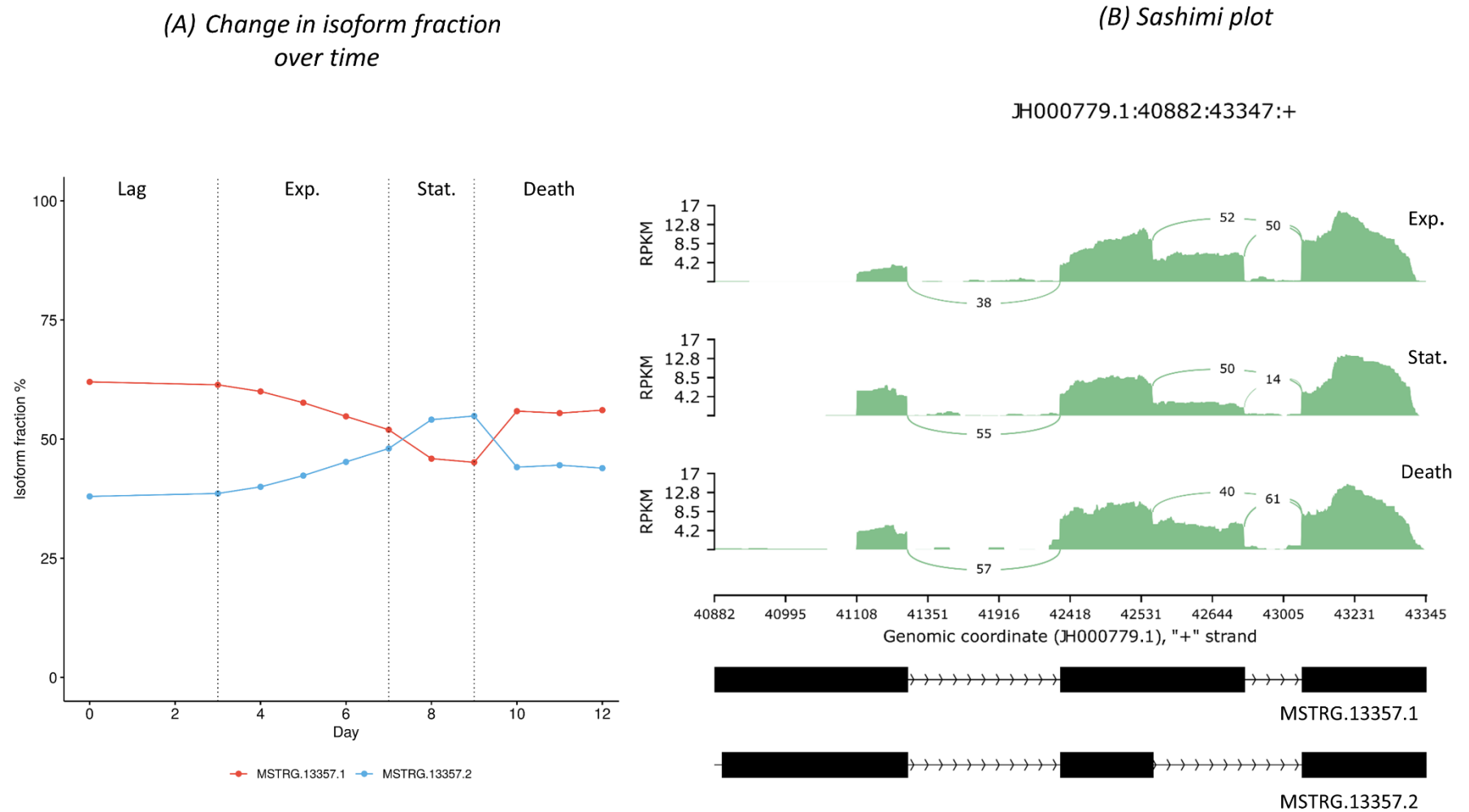


Figure 4.22 (A) The change in isoform fraction of the novel MSTRG.13357 locus which is composed of two isoforms that switch during the stationary phase. (B) A sashimi plot showing an A5 alternative splicing event that leads to a truncate second exon of MSTRG.13357.2.

Figure 4.23 illustrates the secondary structures of the two component isoforms: the longer MSTRG.13357.1 which is dominant during the exponential and stationary phases of culture and the shorter isoform of MSTRG.13357.1 which contains an A5 alternative splicing event that has a truncated second exon. The truncated exon leads to an altered secondary structure for MSTRG.13357.2 which is missing a circular secondary structure.

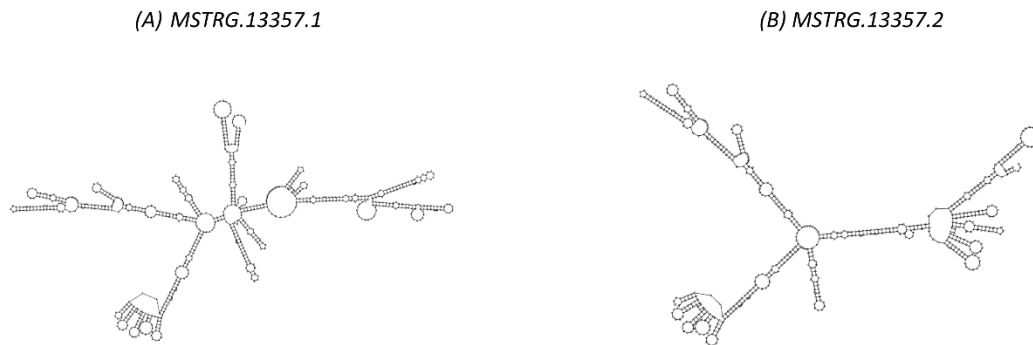


Figure 4.23 (A) The secondary structures of the MSTRG.13357.1 isoform that is expressed during exponential and death phase but is replaced by (B)MSTRG.13357.2 during the stationary phase.

4.4 Discussion

In this chapter, we present the first evidence of global alternative splicing and isoform switching of long non-coding RNAs within Chinese hamster ovary cells. This analysis used high-coverage RNA-Seq data in order to quantify transcripts and associate their expression to cell culture phases in fed-batch culture. A minimum isoform fraction of 10% was used during the assembly stage to retain high quality isoforms for isoform switching and alternative splicing analysis. A comparatively high isoform fraction threshold is necessary in order to reduce false positives as putative, assembled transcripts could, in fact, be originating from spurious read alignment (Frazee et al. 2015). *StringTie* was used to expand the Chinese hamster ovary transcriptome by 80% compared to the version 98 *ENSEMBL* reference annotation. This assembly was highly similar to the assembly presented in **Chapter 2** with over 90% similarity at the exon level which further validates both assemblies. The assembly conducted during the work for this chapter resulted in a 30% increase in the number of annotated lncRNAs compared to the previous assembly from **Chapter 2** with a more complete representation of the non-coding transcriptome.

4.4.1 Isoform composition of lncRNA gene loci

Over 80% of the coding and long non-coding transcriptome was represented by single isoform genes. A similar distribution of single isoform genes was observed in the assembly derived from temperature shift RNA-Seq data. The initial results of the *GENCODE* annotation had more multi-isoform coding genes (Derrien et al. 2012), however, which suggests a possible limitation of sequencing the Chinese hamster ovary cell transcriptome. The *GENCODE* consortium constructed the version 7 assembly by combining RNA-Seq data from 11 tissues and cell lines thereby expanding the number of isoforms detected (Harrow et al. 2012). Many genes exhibit tissue specific isoform expression which could only be detected by increasing the breadth of sample types (Harrow et al. 2012). The reported 80% of single isoform mRNA and lncRNA loci could represent a theoretical lower limit of isoforms expressed within *C. griseus* ovary tissues. Sequencing tissues from parental hamster would significantly expand the number of annotated

lncRNA and mRNA isoforms and provide a more complete assembly (Heffner et al. 2017). Furthermore, the *GENCODE* annotation is a living repository that is continuously updated from published RNA-Seq data that is used to refine the reference human and mouse annotation (Aken et al. 2016).

4.4.2 Comparison of differential gene expression to alternative splicing

We recently presented data from analyzing alternative splice variants in CHO cells after temperature shift where there are cell line specific variability of expression of some mRNA isoforms (Tzani et al. 2019). This variability could also be present in lncRNA loci as these regulatory molecules have been previously demonstrated to be highly tissue and cell type-specific (Kern et al. 2018). lncRNA expression has also been found to be highly tissue and condition specific which would suggest an even higher degree of alternative splicing variation than coding genes (Kern et al. 2018). These variations could attribute to the variability seen in different CHO cell lineages and clones. Further sequencing of parental CHO cell lines from parental CHOK1 as well as DG44 and GS lineages could potentially identify differences in expression of non-coding RNAs.

Less than 1500 transcripts were differentially expressed at the transcript level compared to the 5,561 differentially expressed genes presented in **Chapter 3**. This discrepancy can be explained by the significantly large proportion of single isoform genes which would not be alternatively spliced. Isoform switching measures the change in the proportion of each transcript relative to the total gene between each condition (Vitting-Seerup and Sandelin 2019). Genes composed of only two genes would, by definition, have to occur reciprocally as changes in expression of one isoform would necessitate an equal and opposite change in the proportion of the other. In addition, we removed low expressed transcripts (> 1 TPM in all samples) which would further reduce the number of analyzed loci. Alternatively, comparing between phases would mask transient changes in expression between time points. For example, an individual isoform could be the dominant

transcript for only one day but the aggregate expression between stages would be stable and thus no isoform switching would be detected (W. Guo et al. 2017). Transient changes in transcript expression could be indicative of important fluctuations in cellular metabolism (Hsu et al. 2017). Focusing on aggregate changes over time between cell culture phases will be of more utility, however, as the change in transcript proportion can be better correlated with phenotype.

4.4.3 Isoform switching in genes undergoing differential expression

Isoform switching can be used to identify transcript level alterations in expression that would be masked by gene expression. For example, we observed phase-specific switching of the *Citrate synthase* gene where there are only modest changes in gene abundance but alterations were observed in the transcript abundance where one isoform accounted for almost 100% of expression at a particular stage of cell culture. *Citrate synthase* is an evolutionarily conserved enzyme that converts Acetyl-CoA into citrate during the TCA cycle (Cheng et al. 2009) The human CS gene has been found to have two splice variants, *CSa* and *CSb* with the *CSa* domain containing an additional N-terminal mitochondrial targeting sequence that directs the translated protein to the mitochondria (Cheng et al. 2009). The *CSb* gene accumulates outside the mitochondria as this splice variant lacks the mitochondrial targeting sequence (Cheng et al. 2009). This gene is essential for regulation of the TCA cycle and alterations in splicing could reflect the metabolic transitions that occur when a cell switches from proliferation to a stationary phase.

4.4.4 Isoform pulsing over culture period

Curiously, we also observed temporal variations in isoform fractions which we defined as “isoform pulsing” where component isoforms become dominant during different days despite little variation in gene abundance. One of the most dramatic examples of isoform pulsing was observed in the *ZNF609* gene which is a zinc-finger nuclease and transcription factor (Legnini et al. 2017). Alternative splicing of this gene can result in a circular variant of the transcript where the 5' and 3' ends of the mRNA are joined to make a circular structure which, in turn, can be translated into a protein that regulates polysome size (Legnini et al. 2017). Temporal isoform switching of this

gene could indicate a time-specific regulatory mechanism that is occurring in response to feed which were not identified at the gene level.

4.4.5 lncRNA alternative splicing

398 lncRNAs exhibited changes in isoform fraction between the three culture phases over the 11 days profiled during day fed-batch cultivation period. Long intergenic ncRNAs comprised the largest group of isoform switching non-coding transcripts which reflects the distribution of lncRNA genes described in **Chapter 3**. A stringent filter was used to remove single isoform genes as these genes, by definition, would not undergo alternative splicing. The majority of lncRNA genes that are composed of 2 or more isoforms were found to be long intergenic ncRNAs. lincRNA transcripts are thought to be the more similar to mRNA encoding genes including a similar mechanism of alternative splicing involving the spliceosome although the exact functional mechanism is not currently understood (Krchňáková et al. 2019). This similarity could explain the higher abundance of intergenic transcripts as lincRNAs are more likely to be alternatively spliced (Krchňáková et al. 2019).

4.4.6 Gene ontology enrichment of alternatively spliced mRNAs

Biological process enrichment was performed using *DAVID* webserver to identify significant pathways and processes that are a result of mRNA isoform switching. Functional annotation of lncRNAs with biological processes is currently not possible so this analysis was not performed on non-coding RNAs. There were significantly fewer enriched processes due to differential expression at the transcript level compared to the gene level. A similar, gene level analysis using *DAVID* was presented in **Chapter 3** where genes involved in cell cycle maintenance were some of the most significant whereas functional annotation of isoform switching mRNAs identified processes that are involved in ncRNA post-transcriptional processing and modifications. Intriguingly, these processes interact with non-coding RNAs in order to regulate metabolic pathways which suggests isoform switching of mRNAs is co-regulated alongside lncRNA switching (Kornfeld and Brüning 2014). lncRNAs, like *MALAT1*, are known to regulate alternative

splicing by functioning as scaffolds for splicing factors and the enrichment of processes involved in ncRNA processing and protein modification could reflect the genomic (Gonzalez et al. 2015b)

We discovered significantly fewer isoform switches but this can be attributed to sequencing only one cell line compared to morphologically diverse tissue types. Transcript level analysis can be used to identify specific isoforms driving changes in gene expression. The expression profile of the *PIGA* gene, which is necessary for the anchoring of the GlcNAc-PI glycoprotein, is an example of this pattern (X. Yuan et al. 2017b). The *PIGA* locus contains 7 isoforms but only two were found to be expressed with one isoform accounting for almost 100% of expression during the exponential phase before significantly decreasing while a novel transcript becomes dominant during the stationary phase before declining during the death phase (X. Yuan et al. 2017a). Glycosylation within CHO suspension cultures are often heterogeneous and often change during the duration of culture which could be attributed to preferential production of specific glycoforms (Sumit et al. 2019b).

4.4.7 First evidence of lncRNA isoform switching in CHO cells

Similar patterns of expression were found in lncRNAs between differentially expressed non-coding genes during the exponential, stationary, and death phases of culture. Very little is currently understood about lncRNA transcript expression which severely limits the capacity for functional characterization of these isoforms. The correlation of differential isoform expression between culture phases can be used to infer potential regulatory targets by evaluating putative function as expression changes between stages. *MSTRG.13823*, which is antisense to the *GAPRP1P* coding gene, for example saw a 72% decrease in the dominant isoform, while there was a 13 fold increase in gene expression which can be attributed to a 72% increase in expression of the second isoform. These results are the first evidence of long non-coding RNA transcript level differential expression in Chinese hamster ovary cells.

4.4.8 Isoform switching can be used to quantify and classify different alternative splicing events

IsoformSwitchAnalyzeR uses isoform switching to infer alternative splicing within a locus as differential splicing would be reflected in a change in isoform fraction (Vitting-Seerup and Sandelin 2019). The profile of alternatively spliced mRNAs and lncRNAs in both conditions were similar with the majority of switching transcripts found to undergo alternative transcription start site and alternative transcription termination site splicing. Different transcription start sites are indicative of alternative promoters that are found within the gene body (Reyes and Huber 2018). Studying alternative start sites that produce non-coding isoforms can be a way to differentiate sense long non-coding RNAs that have been assembled into the sense coding loci. Combining profiles of alternative 5' splice sites alongside chromatin accessibility data such as ATAC-Seq (**Chapter 5**) can be useful in identifying canonical sense long non-coding RNAs. Alterations in the 3' termination site govern poly-Adenylation of the transcript with alternative sites leading to a change in the length of poly-A tails (Reyes and Huber 2018). Alternative poly-Adenylation can change miRNA binding sites within the 3'UTR which canonically degrade transcripts by forming duplexes with recognition sites at the end of transcripts and recruiting the RISC complex to fragment the mRNA (Pilotte, Dupont-Versteegden, and Vanderklish 2011; Reyes and Huber 2018). We observed an alternative 3' transcript splice isoforms within the *LACC1* gene where the novel isoform becomes dominant during the stationary and death phases of the culture. The change in isoform abundance did not affect the coding sequence of either isoform as the two transcripts differ only by the location of the last exon. *LACC1* encodes a protein that is necessary for peroxisome formation and regulates mitochondrial efficiency and the rate of glycolysis (Cader et al. 2016). As discussed in **Chapter 3**, the rate of glycolysis is reduced as cells enter the stationary phase. The alternative 3' splicing of the *LACC1* gene could have a potential role in regulating the metabolic transition from the exponential phase to the stationary phase. Alternative 3' splicing was also the most common event observed for lncRNAs with 400 events between exponential, stationary, and death phases. Differences in transcription termination sites of lncRNAs may

govern cellular localization of a transcript to the nucleus, organelles, or cytoplasm (Hoque et al. 2013). Processing of lncRNAs at the 3' end can alter regulatory function and a study performed by Hoque *et al.* that mapped alternative poly-Adenylation sites in mouse cell lines found that 66% of all assembled lncRNAs have alternative 3' splicing (Hoque et al. 2013). Differential 3' splicing of non-coding transcripts has the potential for identifying specific individual regulatory functions.

4.4.9 Exon skipping is the most common alternative splicing event in both mRNA and lncRNA genes

Exon skipping was also a prevalent alternative splicing event over the duration of the culture period and was found in both mRNA and lncRNA transcripts. After transcription of the precursor mRNA, splicing by exon skipping can occur where an exon may be removed by the spliceosome in addition to introns. Exon skipping can result in changes to the coding sequence which can modify function and protein localization for mRNAs. An exon skipping event was discovered within the 8th exon of the *Paralemmin* gene (*PALM*) which is an evolutionarily conserved phosphoprotein that is localized to cytoplasmic plasma membranes (Turk et al. 2012). *PALM* was initially characterized within neuronal cells where this gene was found to regulate plasma membrane dynamics, cell motility, and cell shape (Hultqvist et al. 2012). The 8th exon of this conserved gene was previously found to be alternatively spliced within breast cancer tissues (Turk et al. 2012). Both isoforms were found to be expressed in different proportions within carcinogenic cells derived from patient biopsies but the transcript variant lacking the 8th exon was more dominant within healthy tissues (Turk et al. 2012). The longer, alternatively spliced transcript containing the exon is necessary for recruiting AMPA glutamate transmembrane receptors to the plasma membrane and upregulation of these receptors has previously been correlated with increased apoptosis (Arstikaitis et al. 2008). We propose that this exon skipping event is mediated by the metabolic shift of increased glutamate production from glutamine during the death phase of culture. As discussed in **Chapter 3**, glutamate is completely depleted during the stationary phase but concentrations increase at the onset of death phase which coincides with the alternative

splicing of a longer *PALM* isoform that recruits glutamate receptors that drive the progression to apoptosis. This initial finding is promising but further work is needed to identify the precise interaction between the isoform and glutamate production. Exon skipping was also prevalent for lncRNAs in both conditions tested. Exon skipping has not yet been widely profiled in lncRNA transcripts as the relationship between primary sequence and secondary structure is not yet understood (Ziegler and Kretz 2017). Differential splicing of exons can alter the secondary structure of lncRNA transcripts which can potentially govern regulatory function of each isoform (Johnsson et al. 2014b). We identified two isoform skipping events in the RNA-Seq data: the *ENSCGRG00000025089* gene and the novel *MSTRG.16574* locus. The *ENSCGRG00000025089* gene generates two isoforms that only have the first exon in common while alternative splicing of the *MSTRG.16574* results in complete depletion of the second exon but retention of the 5' and 3' structure. Exon skipping events in both genes resulted in more linear secondary structures with the loss of stem-loop and secondary structures after the event. Exon skipping analysis of lncRNAs alternative splicing can be of significant use for the identification of specific domains that are necessary for ncRNA mediated regulatory functions such as the domains that can recruit epigenetic remodeling complexes.

4.4.10 Isoform switching independent of gene level differential expression

We also observed isoform switching occurring independently of gene expression which has been previously reported in transcriptome-wide studies of coding transcripts (Vitting-Seerup and Sandelin 2017). These results agree with previous work where we have demonstrated alternative splicing can occur at the transcript level without any significant alterations in expression at the gene level (Tzani et al. 2019). Traditional gene level analysis is of considerable utility in studying alteration of gene expression in different phenotypes but would miss transcript level alterations that do not result in changes in the total gene abundance (Love et al. 2019). We present, for the first time, lncRNAs that are differentially expressed at the transcript level due to changes in transcript isoform as lncRNA gene expression remains stable. These findings suggest that many

lncRNAs are differentially spliced due to changes in cell culture physiology with those splice variants expressed during the exponential phase conducive to cell proliferation and those expressed during the death phase associated with viability.

4.4.11 Isoform switching over the duration of fed-batch culture

Significant isoform switching and alternative splicing events over the duration of fed-batch cell culture represent only a small cohort of transcripts. Only 35 mRNA transcripts and 18 lncRNA transcripts had significant differences in isoform fraction between exponential, stationary, and death phases. The novel dual isoform gene *MSTRG.13357* is one such gene where one of its component isoforms, *MSTRG.13357.2*, becomes dominant only during the stationary phase. An alternative A5 splicing event occurs between *MSTRG.13357.1* and *MSTRG.13357.2* where the second exon becomes truncated. This shortening of the exonic sequence during the stationary phase results in the loss of hairpin and stem-loops within the secondary structure. Preferential expression of the *MSTRG.13357.2* is particularly intriguing as it suggests a change in the non-coding transcriptome that is unique only to the stationary phase. This event could potentially be the result of an error within the assembly, however, and further *in vitro* validation is necessary in order to evaluate putative biological function.

4.4.12 Alternative splicing induces changes to lncRNA secondary structure

The secondary structures of lncRNAs have been found to be conserved despite the lack of conservation of the nucleotide sequences (Johnsson et al. 2014a). Regulatory functions of lncRNAs are governed by secondary structures that are formed *in vivo* by folding RNA molecules into complex domains such as pseudoknots, hairpins, stem-loops and cruciform shapes. These secondary structures are thought to function as binding sites for proteins such as chromatin remodeling complexes and transcription factors (H. Yuan et al. 2016). One of the important regulatory functions of these molecules is to form complexes with proteins in order to guide transcription factors and remodeling complexes to target genomic loci in order to modulate gene expression or alternative splicing (R. Li, Zhu, and Luo 2016). The potential regulatory function of

a transcript can be inferred by what three dimensional shapes form after transcription (R. Li, Zhu, and Luo 2016). Very little is currently understood about the formation of secondary structures and their function *in vivo*, however, but it is known that alternative splicing induces alterations to the secondary structure in both mRNAs and lncRNAs (Yan et al. 2016)

A significant limitation in predicting lncRNA function from secondary structures is the difficulty in accurately modeling RNA folding *in silico* (Delli Ponti et al. 2018). Computational methods, such as *RNAfold* (Lorenz et al. 2011), predict structure by estimating the minimum free energy necessary to form optimal hydrogen bonds within an RNA molecule. Transcripts do not necessarily adopt the lowest energy state necessary to form these structures *in vivo* and often adopt a dynamic range of folding states (Yan et al. 2016). Furthermore, lncRNAs can be enzymatically folded and modified to adopt conformations that are not achievable by simple Watson-Crick base pairing which significantly increases the complexity of secondary structures. These complex secondary structures may also form transiently *in vivo* so isolating these transcripts may be extremely difficult using traditional methods (Yan et al. 2016). The computational methods necessary for accurate prediction are not yet mature so we are reliant upon the free energy state prediction methods in order to model any putative regulatory functions of lncRNAs. These computational methods are improving as more *bona fide* non-coding molecules are being experimentally characterized which are then used to further refine algorithms. The computational prediction of secondary structures we presented in this chapter are an important first step in studying the regulatory function of *C. griseus* specific motifs within non-coding transcripts.

4.5 Conclusions

In this chapter, we present the first evidence of lncRNA isoform switching and alternative splicing in Chinese hamster ovary cells. Alternative splicing is a cellular mechanism that can introduce significant diversity to the genome with multiple transcripts transcribed from the same gene. We used high-coverage RNA-Seq data from CHO cells cultivated in fed-batch mode over 14 days in a 10L bioreactor in order to generate the first genome wide map of lncRNA isoform switching during the exponential, stationary, and death phases of culture. We identified 398 lncRNA transcripts that are alternatively spliced during different stages of the culture including 18 transcripts that are alternatively spliced across all three stages.

Chapter 5

Change in Chromatin Accessibility

Following Temperature Shift

5 Summary

In this chapter, we provide a brief introduction to chromatin accessibility of regulatory features within the genome such as promoters, enhancers, and repressors (**Section 5.1**). We also introduce the Assay for Transposase Accessible Chromatin (ATAC-Seq) method that can be applied to profile genome-wide alterations in the accessibility of regulatory regions (**Section 5.2**) and discuss computational tools that are currently available to analyse ATAC-Seq data (**Section 5.3**). We present a computation pipeline that can integrate RNA-Seq and ATAC-Seq data in order to profile transcriptional regulatory changes within the CHO genome after a reduction in temperature shift (**Section 5.4**). ATAC-Seq data was integrated with the assembly presented in **Chapter 2** in order to validate active transcription of lncRNAs and identify transcription factor regulatory networks (**Section 5.5**). We also used ATAC-Seq to stable genomic regions of high accessibility that can function as “landing pads” for efficient transgene integration. This chapter will conclude with a discussion on the applicability of ATAC-Seq for CHO cell engineering from profiling transcription start sites to transcription factor binding sites (**Section 5.6**).

5 Introduction

In previous chapters, we demonstrated the utility of RNA-Seq for identifying long non-coding RNAs and identifying potential *cis* regulation between protein coding genes and their non-coding counterparts. We have also demonstrated the application of strand-specific RNA-Seq to relate lncRNA expression with industrially relevant conditions. RNA-Seq is a common method to profile the total population of RNA molecules within a cell but cannot capture genomic regulation which governs transcription (Klemm, Shipony, and Greenleaf 2019a) and will only present a partial understanding of complex regulatory interactions. RNA transcription is regulated at the genomic level by local genome architecture and transcription factor binding affinity (Mercer and Mattick 2013; Tsompana and Buck 2014). Genomic regulation of coding genes is relatively well characterized but regulation of lncRNA transcription is not as well characterized. RNA-Seq, when combined with methods that can probe chromatin architecture (e.g. DNase-Seq, ATAC-Seq) can be used to probe transcription regulatory networks (Gendron et al. 2019).

DNA within mammalian cells is estimated to be 2 meters in length but is packaged into a 6 μ m nucleus by being coiled around a protein complex referred to as a nucleosome (Song et al. 2011a). Nucleosomes are composed of an octamer of 4 histone proteins (histone 2a, histone 2b, histone 3 and histone 4) arranged into duplexes (Zhou, Gaullier, and Luger 2019a) which forms a cylindrical shape where ~147bp of DNA is coiled around (**Figure 5.1**). Histone protein duplexes arrange to form stable nucleosomes which in turn protect and package DNA (Klemm, Shipony, and Greenleaf 2019b). Nucleosomes form repeating units ~200bp apart that were shown to have a characteristic “bead on a string” shape using microscopy (Lawrence, Daujat, and Schneider 2016a). Regions of tightly condensed nucleosomes are referred to as heterochromatin where RNA transcription is quiescent. Regions of actively transcribed DNA is referred to as euchromatin where DNA is not as protected by nucleosomes (Zhou, Gaullier, and Luger 2019b).

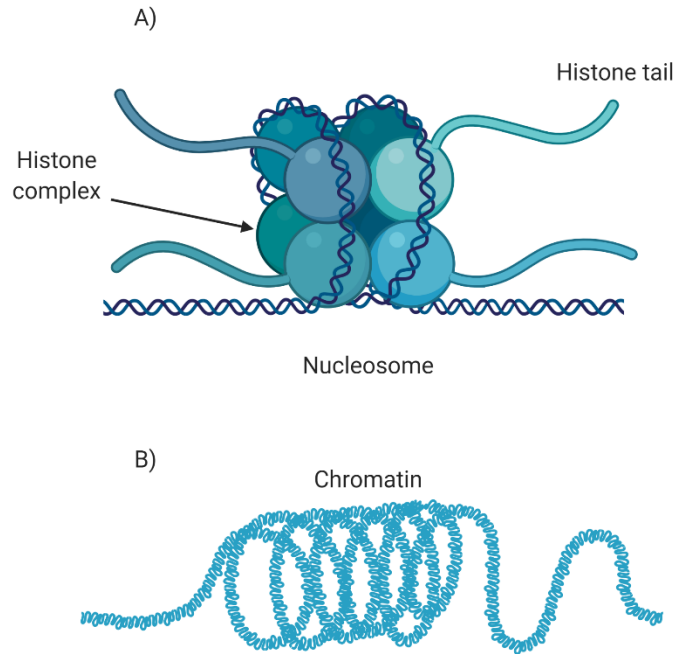


Figure 5.1 (A) The structure of a nucleosome with 4 core histones coiled around a ~147bp DNA sequence. Histone complexes contain trailing tails which can be modified. (B) Repeating complexes of nucleosomes packaged into chromatin. Nucleosome can be further condensed into transcriptionally inactive heterochromatin or unwound for gene transcription as euchromatin.

Histone proteins contain a core that coils around DNA and a N-terminal “tail” or chain of proteins that is commonly altered by post-translational modifications (Lawrence, Daujat, and Schneider 2016b). PTMs such as methylation, acetylation, phosphorylation, and ubiquitination regulate chromatin structure and accessibility. Methylation and acetylation, or the addition and removal of $-CH_3$ and $-OC_2H_3$, of peptides occurring at histone tails are some of the best characterized post-translational modifications in eukaryotes. For example, methylation of the fourth lysine residue of histone 3 (H3K4me3) has been found in regions of DNA that is being actively transcribed and used as a marker for gene promoters (Feichtinger et al. 2016). The addition of a ubiquitin protein at histone 2a tails has been correlated with transcriptional silencing whereas the addition at a histone 2b complex has been identified as marker for transcriptional activation (Cao and Yan 2012). Phosphorylation, or the addition of $-O_3P$ groups, at histone 3 tails has been implicated in DNA damage repair pathways (Lawrence, Daujat, and Schneider 2016b). Long non-coding RNAs

play an important role in regulating PTMs of histone tails by functioning as scaffolds for epigenetic remodeling enzymes that, in turn, recruit complexes to specific genomic loci (Nie et al. 2012). *Xist*, for example, recruits the polycomb repressive complex 2 enzyme to tri-methylate the fourth lysine residue of histone 3 (Leung and Panning 2014) during X chromosome inactivation. *Xist* recruits PRC2 complexes across all extant promoters in extraneous X chromosomes in a process known as “coating” where the H3K4 residues are tri-methylated and rendered transcriptionally inert (Dimond and Fraser 2013).

The three-dimensional structure of DNA also plays a significant role in regulation of gene transcription with the presence of transcription factor binding sites in *cis* regulatory elements (CREs), such as promoters, enhancers, and silencers. CREs contain short DNA sequences ~10 nucleotides long (Stewart and Plotkin 2012) that function as binding sites for transcription factors (TFs) which, in turn, modulate RNA transcription (Wittkopp and Kalay 2012). The regulatory activity of these transcription factor binding sites with CREs is governed by physical access to the DNA sequence; if the transcription factor binding site is occupied or protected by a nucleosome, the TF cannot bind to its constituent site and activate or repress transcription (A. Chen, Chen, and Chen 2018)

Promoters are genomic loci that range from approximately 100 bp to 1 kb (Klemm, Shipony, and Greenleaf 2019b) located adjacent to a gene’s transcription start site at the 5’ end of a gene (Tsompana and Buck 2014). RNA polymerases, including RNA Polymerase II which transcribes mRNA and many lncRNAs (Böhmdorfer and Wierzbicki 2015), are recruited to the promoter region and form the pre-initiation complex (PIC) which bind to the promoter and begin transcription of DNA into RNA. Next generation sequencing using chromatin immunoprecipitation (ChIP-Seq) enabled identification of genome-wide promoter loci by using antibodies that selectively bind histone proteins that contain the methylated H3K4me3 marker which are often enriched in regions near the TSS (L. Han and Zhao 2009) and denote transcriptionally active coding genes. Long

non-coding RNAs are also similarly transcribed by RNA polymerase II and have enriched H3K4me3 histone modifications near the transcription start site (Guttman et al. 2009). However, Melee *et al.* found that the promoters of intergenic long non-coding RNAs are enriched for the H3K9me3 indicating the possibility of alternative transcriptional regulation machinery for some intergenic lncRNAs (Melé et al. 2017).

Enhancers are regulatory elements that increase transcription of target genes and are interspersed throughout the eukaryotic upstream or downstream of the target gene or within introns (Klemm, Shipony, and Greenleaf 2019b). Enhancer elements found within the genome can increase gene expression compared to basal expression of genes initiating from just the promoter (**Figure 5.2**). Distal enhancer elements increase transcription by binding transcription factors and forming DNA stem-loops to bring DNA-protein complexes to the promoters of the target gene (Halfon 2019a). Enhancers often contain H3K4me1 post-translational modifications which can be used as identifying markers for detection similar to promoters (Zabidi and Stark 2016). Furthermore, certain short lncRNAs, referred to as enhancer RNAs (eRNAs), are transcribed bi-directionally from genomic regions containing enhancer sites and are thought to be involved in enhancer/promoter stem loop formation (Ilott et al. 2014).

Gene silencers, by contrast, function as the converse of enhancers by preventing binding of RNA polymerase II which in turn prevents transcription initiation (Tsompana and Buck 2014). Transcriptional silencers are not as well-characterized in the genome compared to enhancers as the genomic mechanism is poorly understood (D. Huang et al. 2019). The presence of gene H3K27me3 marks have been correlated to the presence of silencers, however (Jayavelu et al. 2018; D. Huang et al. 2019) but is not as specific of a histone modification when compared to H3K4me1 for enhancers or H3K4me3 for promoters (D. Huang et al. 2019).

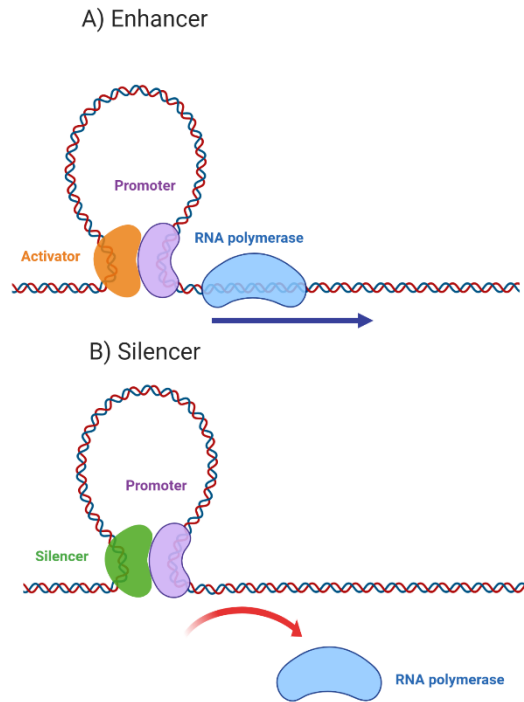


Figure 5.2 (A) An illustration of a transcriptional enhancer forming stem-loop structure to bring transcriptional activators to promoter. The enhancer region binds transcription factors in regions that can be very distal to the target promoter. After the activator/promoter complex is formed, RNA transcription by RNA Pol II is increased and transcription is upregulated. **(B)** An illustration of gene repression by a gene silencer. DNA stem-loops form in a manner similar to enhancers. Silencer regions bind proteins that, when forming a complex with the promoter, repress transcription initiation by RNA Pol II.

Genome wide histone modification and chromatin accessibility analysis has been enhanced by the development of high-throughput sequencing methods such as CHIP-Seq, DNase-Seq, and ATAC-Seq (Tsompana and Buck 2014). The Chromatin Immunoprecipitation followed by sequencing (CHIP-Seq) is a method that leverages next generation sequencing to identify protein bound to DNA (Furey 2012). CHIP-Seq is a technique that can be applied to probe specific histone modifications or transcription factors bound to chromatin. Protein-bound DNA is first treated with form covalent bonds between the complexes which is referred to as “cross-linking” (Landt et al. 2012). The cross-linked DNA is then treated to an endonuclease digestion or preferential DNA fragmentation by sonication to remove unlinked DNA (Furey 2012). An antibody that is specific to the DNA binding protein of interest is used to precipitate selective protein-DNA complexes prior to sequencing (Furey 2012). CHIP-Seq has been applied to generate maps of histone modifications and transcription factor binding by the ENCODE consortium (Landt et al. 2012). The

data derived from multiple sequencing experiments within ENCODE was used to refine lncRNA annotation by the GENCODE project as lncRNAs canonically exhibit similar histone modifications to mRNAs (Derrien et al. 2012).

The first significant study of histone modifications within the CHO genome was performed by Hechinger *et al.* where the authors profiled changes in histone modifications over the course of an 8 day batch culture (Feichtinger et al. 2016). DNA was extracted over 12 hour increments over the 8 day culture period and ChIP-Seq libraries were prepared for H3K4me1, H3K4me3, H3K6me3, H3K9me3, H3K27me3, and H3k27ac histone modifications (Feichtinger et al. 2016). In addition, genome-wide methylome profiling (bisulfite sequencing) was performed to identify methylation sites across the CHO genome and the methylome was found to be very stable over the period of batch cultivation (Feichtinger et al. 2016). Feichtinger *et al.* found that histone modifications continuously change over the culture period as cells adapt to nutrient stress by the methylome is comparatively stable (Feichtinger et al. 2016). The authors further found that methylation at the promoter can be used to identify ON/OFF gene transcription as well as found that genic histone modifications are positively correlated with gene expression with RNA-Seq (Hernandez et al. 2019). The data derived from these sequencing experiments have been provided as publically available chromatin state maps at <http://cho-epigenome.boku.ac.at/index.html> which is a significant resource for the CHO research community.

Chromatin accessibility and epigenetic regulation is of particular importance for industrial production of recombinant proteins within CHO cells as changes in chromatin states can lead to variable product quality or loss of transgene expression (Veith et al. 2016). High nucleosome occupancy near integration sites for transgenes have been shown to correlate with low expression of the gene of interest (Veith et al. 2016). Chromatin state models would be a significant resource for the CHO research community yet it has been poorly studied. There is significant need of high-

resolution maps of chromatin accessibility that can be used for identifying stable transgene integration sites as well as identifying *cis* regulatory regions for rational cell engineering. In this chapter, we summarize high-throughput sequencing methods to profile chromatin accessibility (**Section 5.2**) including a discussion of ATAC-Seq (**Section 5.2**), computational methods to analyse ATAC-Seq data (**Section 5.3**), describe the utility of chromatin accessibility analysis for transgene integration (**Section 5.4**), and will describe an experiment that profiled chromatin accessibility following a reduction in culture temperature from 37°C to 31°C (**5.5-5.8**)

5.1 High-throughput sequencing methods to profile chromatin accessibility

Methods to profile chromatin accessibility all use steric hindrance of enzymes to DNA where closed chromatin would not be readily digested due to protection by nucleosomes or transcription factors whereas open chromatin would be more readily digested. Initially, chromatin accessibility assays relied on DNA digestion by DNase I followed by laborious Southern blotting or qPCR which severely limited the number of open chromatin sites that could be analysed. High-throughput methods such as DNase-Seq, FAIRE-Seq, MNase-Seq, and ATAC-Seq use sequencing after enzymatic cleavage or formaldehyde cross-linking to identify accessibility as there will be a higher intensity of signal from open chromatin compared to closed. In DNase-Seq, extracted nuclei from > 50 million cells is treated with DNase I which selectively cleaves open chromatin followed by sequencing (Boyle et al. 2008b; Hesselberth et al. 2009). FAIRE-Seq and MNase-Seq both rely on formaldehyde cross-linking followed by enzymatic cleavage to find open chromatin with MNase-Seq selectively enriching for regions bound to nucleosomes while FAIRE-Seq identifies large regions of open chromatin (Giresi et al. 2007; Song et al. 2011b).

Prior sequencing methods such as DNase, FAIRE, and MNase sequencing all required a high number of cells in order to accurately profile chromatin states or nucleosome positioning. Buenrostro *et al.* developed the Assay for Transposase Accessible Chromatin by sequencing (ATAC-Seq) that can quantify chromatin accessibility using significantly fewer cells (J. D.

Buenrostro et al. 2013). The authors demonstrated the efficacy of ATAC-Seq to identify accessible chromatin with as few as 5,000 cells with a suggested cell number of 50,000 which is significantly lower than previous methods which require hundreds of thousands to millions of cells (J. D. Buenrostro et al. 2013; J. D. Buenrostro et al. 2015b). ATAC-Seq functions in a manner similar to DNase-Seq by using a hyperactive Tn5 transposase to selectively fragment DNA; DNA originating from open chromatin is preferentially fragmented over DNA that is coiled around nucleosomes or bound to proteins (J. D. Buenrostro et al. 2013). (J. D. Buenrostro et al. 2015a). The number of cells used as input for the transposition reaction is a crucial parameter, as it determines the library complexity or the number of unique sequenced fragments. A lower cell number can be correlated with over-fragmented DNA that produces a false signal in closed chromatin; higher cell number can lead to under-fragmentation and poor signal-noise ratio (J. Buenrostro et al. 2015b). ATAC-Seq has been successfully used for very low cell numbers, however, and has achieved single-cell resolution (J. Buenrostro et al. 2015a). The library preparation protocol is similar to the DNase-Seq method where crude nuclei are harvested by centrifugation and a lysis buffer containing a non-ionic detergent. Unlike FAIRE and MNase-Seq, addition of formaldehyde is not recommended by the authors as the fixative can significantly impede transposition efficiency (J. D. Buenrostro et al. 2015a). The Nextera hyperactive transposase enzymatically cleaves double-stranded DNA while simultaneously ligating Illumina specific 12bp oligonucleotides to both ends of the fragment in a process referred to as “tagmentation” (Lamble et al. 2013). Nucleosome-bound DNA would hinder Tn5 binding due to steric hindrance while open chromatin would be readily tagmented (J. D. Buenrostro et al. 2015a). A small aliquot of transposed DNA is used as template for a qPCR reaction to estimate the number of total PCR cycles required for efficient amplification and is an essential step in ATAC library preparation. Over-amplification of fragments would lead to increased GC bias and saturation of signal from closed chromatin regions (J. D. Buenrostro et al. 2013). The remaining reaction of transposed DNA is used as template for a PCR reaction in order to add Illumina specific adaptors

and barcodes to the fragments using the calculated number of cycles estimated by the side reaction prior to sequencing. Mitochondrial DNA is a common source of contamination in ATAC-Seq which might compose up to 80% of sequenced reads depending on cell type (Montefiori et al. 2017). Montefiori *et al.* used CRISPR/Cas9 strategy to deplete mitochondrial contamination and enrich nuclear DNA 1.7 fold (Montefiori et al. 2017). An alternative method is to add multiple detergents, such as Tween-20 and digitonin, to the lysis buffer to improve selective permeabilization of the nuclear genome (Corces et al. 2017) and enriching for chromatin accessible regions 3 fold.

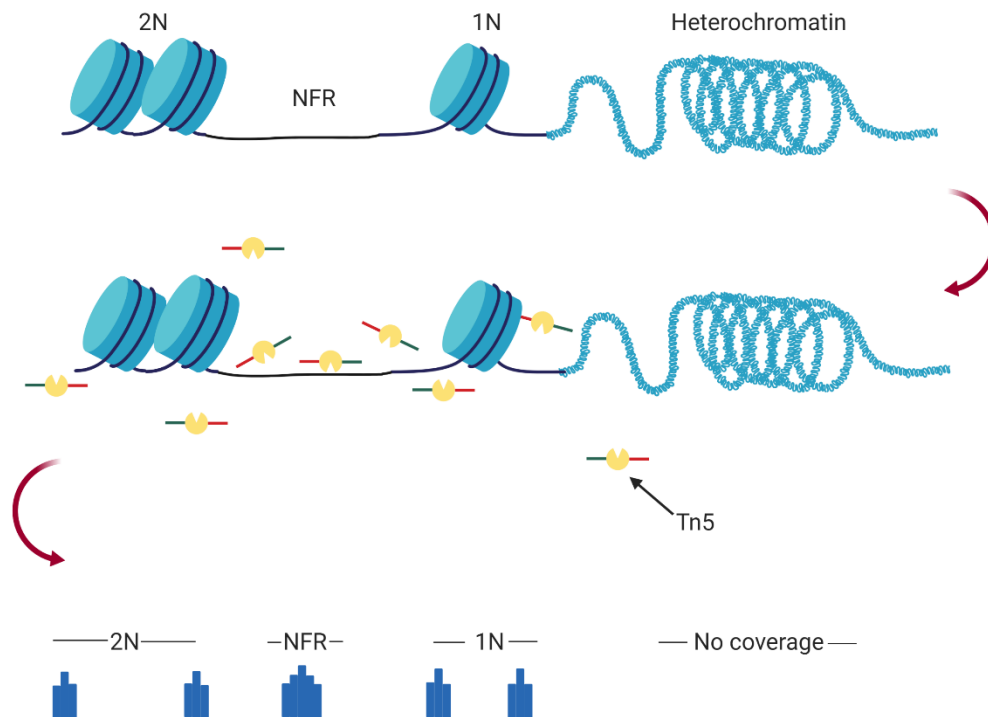


Figure 5.3 A diagram detailing the ATAC-Seq library preparation process. DNA from a diluted sample of cells is first extracted using a non-ionic detergent that preserves DNA bound to nucleosomes and heterochromatin. The DNA is then treated with hyperactive Tn5 transposase that preferentially fragments regions devoid of nucleosomes and around nucleosome boundaries. The fragments are then sequenced and aligned to the genome with the shorter fragments originating from NFR regions and fragments ranging from 180-220bp equating with a mononucleosome. Larger nucleosomes will have fragment distributions that are multiples of the mononucleosome distribution. Heterochromatin will not be tagged due to protection of DNA by compacted chromatin.

5.2 Computational methods to profile chromatin accessibility with ATAC-Seq

Sequencing methods to assess chromatin accessibility all follow four basic stages of analysis:

first raw reads are trimmed to remove adaptors and low-quality bases, second the quality filtered data is aligned to the genome. The Nextera library preparation protocol for ATAC-Seq might introduce redundant, duplicate reads due to amplification of the same fragments and must be removed in order to estimate accessibility accurately. After de-duplication, forward and reverse reads need to be offset by ~9 bp (+4 in the forward, -5 in the reverse) in order to account for the fragmentation of the Tn5 transposase (J. D. Buenrostro et al. 2015a). Regions of enriched read coverage which correspond to accessible regions are calculated using peak calling algorithms like *MACS2*, *F-Seq*, and *HMMRATAC* and annotated to features on the genome. After the peak calling stage, differential accessibility is evaluated in a manner that is similar to RNA-Seq data (**Section 1.6**) differentially accessible peaks are examined for enriched transcription factor binding sites using tools such as *HOMER2* or *MEME*. **Section 5.3** will summarize peak calling algorithms (**Section 5.3.1**), describe methods to identify transcription factor binding sites (**Section 5.3.2**), and will conclude with a discussion on software packaged developed specifically for ATAC-Seq analysis such as *chromVAR* (**Section 5.3.3**)

5.2.1 Software packages for peak calling

One of the most important steps in ATAC-Seq data analysis is the peak call stage where accessible loci are identified. Peak callers, such as *MACS2* and *F-Seq*, were originally developed to analyse CHIP-Seq data but have been adapted for ATAC-Seq albeit by only utilizing a fraction of the sequencing data in order to model nucleosome free regions (Tarbell and Liu 2019). *HMMRATAC*, a recently released peak caller, was the first software package developed specifically for analyzing paired-end ATAC-Seq data. **Section 3.3.1** will summarize parameters and utility of these algorithms.

The *Model-based Analysis for CHIP-Seq Version 2* (*MACS2*) is the most commonly used peak caller for DNA enrichment sequencing studies like CHIP-Seq, ATAC-Seq, DNase-Seq, and

MNase-Seq with over 6,000 citations since its initial publication (Feng et al. 2012a). *MACS2* computes localized fold enrichment for sequenced reads compared to a background model in order to identify peaks. *MACS2* can compute fold enrichment over relatively narrow regions (< 1kb) and longer broad regions across the genome (Feng et al. 2012b).

The first stage of the analysis is to compute a fragment length distribution for read position adjustment: reads in CHIP-Seq data need to be shifted by the average fragment length in order to account for enrichment of signal at 5' and 3' ends of the DNA molecule due to protection by a TF (Thomas et al. 2017). Read re-positioning is not desirable for chromatin accessibility data such as DNase-Seq or ATAC-Seq as the likelihood of enzymatic cleavage directly measure accessibility. Instead, the authors recommend taking the 5' end of the first read and extending the peak length uniformly in the 3' end for 200bp to enrich for nucleosome free regions and mononucleosomes (Feng et al. 2012b) which enriches for signal found in the cutting sites. *MACS2* will only use the first read when centering on cutting sites so half the data found in the second read is discarded. An alternative is for *MACS2* to utilize the total fragment length (--format BAMPE) to calculate variable peak lengths that can span multiple nucleosomes (Feng et al. 2012b).

MACS2 calculates peak enrichment compared to the background signal by using a sliding window that is twice the estimated fragment size with overlapping windows being merged for downstream analysis. The peak caller first computes a Poisson distribution of background signal, λ_{BG} , that is compared to the local distribution, λ_x , of the candidate region, P .

$$P = \max(\lambda_{BG}, \lambda_x)$$

MACS2 assigns an enrichment P-value for each candidate region with those regions that met a minimum threshold of 1×10^{-5} assigned as final, high-confidence peaks (Feng et al. 2012b). The

output of *MACS2* is a table of genomic coordinates in BED format with the P-value indicating the confidence of each peak.

The majority of published ATAC-Seq studies used *MACS2* in the default, single-end mode that enriched for cutting sites despite discarding half the data and ignoring the fragment length distribution for nucleosomes (Gaspar 2018) despite the recommendations by Buenrostro *et al.* to perform paired-end sequencing for ATAC-Seq (J. D. Buenrostro et al. 2015a). As *MACS2* was explicitly designed for CHIP-Seq data analysis, it is not wholly suited for analyzing variable fragment length data (Tarbell and Liu 2019).

An alternative algorithm for chromatin accessibility analysis is *F-seq* which has successfully been adapted for CHIP-Seq and DNase-Seq experiments. *F-seq* calculates continuous probability density across the genome with a greater probability directly proportional to the likelihood of a read aligning to the locus (Boyle et al. 2008a). The density is calculated across a user defined window along the genome and is estimated using the Gaussian kernel function K below with b is the window size (described as “bandwidth” by the authors), x is a point along the genome, and n the number of points that are sampled.

$$P = \frac{1}{nb} \sum_{i=1}^n K\left(\frac{x - x_i}{b}\right)$$

F-seq requires the pre-processing of alignment data into a BED format that only contains the first base pair of the 5' end of the first read thereby discarding the rest of the alignment in order to enrich for enzymatic cleavage sites within the DNA. *F-seq* was originally developed to analyse DNase-seq data by identifying the start of *DNase* hypersensitivity sites but can be adapted to use for ATAC-Seq due to the similarity of *Tn5* transposition (Chang et al. 2018). The output of *F-seq*

is an alignment track (WIG format) containing computed coverage across the genome detailing continuous probability distributions for aligned reads.

MACS2 and *F-Seq* parameters often need to be modified in order to accurately model ATAC-Seq data. *HMMRATAC*, which was developed by the authors of *MACS2*, is the first dedicated ATAC-Seq peak caller and functions by creating Hidden Markov Models (HMMs) of nucleosome free and nucleosome occupied signals in order to accurately identify regions of open chromatin as well as nucleosome bound fragments (Tarbell and Liu 2019). The first stage of peak calling is segregation of fragment length distribution into nucleosome free region and nucleosome bound signal with NFR data modelled with an exponential distribution (Tarbell and Liu 2019). The distribution, P , of NFR fragments is defined below where β is the mean fragment length and x represents the observed insert size:

$$P_{NFR} = \frac{1}{\beta} e^{-x/\beta}$$

The fragment length distribution of each mono-, di-, and tri-nucleosome signal is calculated in a similar manner with μ representing the average fragment length and σ representing the standard deviation of fragment lengths.

$$P_{NUC} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

HMMRATAC then generates a training region by calculating localized, enriched-fold-change compared to 5kb background flanking sequence. The refined HMM is used to classify each putative peak into NFR, nucleosome, or background categories (Tarbell and Liu 2019).

HMMRATAC has the ability to distinguish between nucleosome free regions and flanking nucleosome bound fragments which is useful for inferring nucleosome positioning and accessible

chromatin around transcription start sites. One significant advantage of *HMMRATAC* is that it can de-convolute signal in peaks found in distal regions devoid of protein coding genes and potentially identify active chromatin features (Tarbell and Liu 2019). *HMMRATAC* demonstrates the utility of dedicated peak callers for ATAC-Seq over callers developed for other data such as CHIP-Seq (**Section 3.4.1**).

5.2.2 Differential chromatin accessibility

The next stage of ATAC-Seq calling is determining changes in chromatin accessibility with respect to experimental conditions. The process is very similar to differential expression analysis where reads are counted and normalized using R packages such as *edgeR* (Nikolayeva and Robinson 2014) and *DESeq2* (Love, Huber, and Anders 2014) and the difference of means between groups is used to calculate a-fold-change and a Benjamini-Hochberg adjusted P-value. The *DiffBind* R package was developed for identifying differentially bound DNA in CHIP-Seq data but can be used for assessing differential accessibility (Ross-Innes et al. 2012b). *DiffBind* performs differential affinity analysis using *DESeq2* for normalization and identification of mean variance of peaks between conditions (Ross-Innes et al. 2012a). Significantly differentially accessible peaks can then be inspected to identify enriched transcription factor binding sites that correlate with changes in chromatin states (**Section 5.3.3**).

5.2.3 Motif enrichment analysis

Transcription factors regulate gene transcription by binding to recognition sites in open chromatin in promoters, enhancers, and repressors. Transcription factors preferentially bind to protein specific DNA sequences, referred to as binding sites, with physical access by the TF governing regulatory effect. It is possible to infer the genomic location of TFBSs by identifying short (~12bp) sequences or motifs that are likely TF targets. It is possible to generate high confidence maps TFs using high-throughput genomic enrichment methods such as CHIP-Seq, DNase-Seq, and ATAC-Seq.

The Multiple Expectation Maximization for Motif Elicitation (MEME) suite of tools was one of the first methods developed to detect transcription factor motifs in enriched DNA data with tools for *de novo* discovery as well as annotation of known motifs (Grant, Bailey, and Noble 2011b). The Find Individual Motif Occurrence (FIMO) tool within the MEME suite is used to identify the localization of specific motifs from a set of sequences derived from DNA enrichment data (Grant, Bailey, and Noble 2011a). PWMs have been experimentally derived from CHIP-Seq data for a wide variety of organisms include humans, mice, zebrafish, and yeast. FIMO takes a user-specific position weight matrix for motif discovery and enrichment by first computing a log-likelihood ratio score for a motif across each individual base along a set of given sequences. The likelihood-ratio score is compared to randomly generated background sequenced to evaluate probability of occurrence for a motif.

HOMER2 is a commonly used software package that can analyse and annotate enriched motifs in CHIP-Seq, DNase-Seq, GRO-Seq, and ATAC-Seq from peak calling to motif discovery (Heinz et al. 2010). During the peak calling stage, *HOMER2* will compute localized fold enrichment of all reads and compare the signal to a surrounding 10kb window (Benner 2018a). A peak is called if there is a 4 fold or more increase in localized enrichment compared to the background (Benner 2018b). The coordinates of each peak can be interrogated for the presence of enriched transcription factor motifs. A background set of peaks is calculated from random genomic intervals and used for %GC normalization of identified peaks. Peaks that contain high percentage of homopolymers such as AT rich motifs are removed and a k-mer table is created of all peaks (Benner 2018b).

The k-mers are searched for similarity to a known motif and an enrichment score is calculated between experimental peaks and background peaks. In *de novo* motif discovery, *HOMER2* computes the distribution of k-mers to find over-represented k-mers in the oligo table compared to the background and retains those that are significantly enriched. The *de novo* predicted motifs

are compared to the database of known motifs using Pearson's correlation coefficient with a higher correlation indicating similarity to known motifs. Unless specified, *HOMER2* will search both strands of the *de novo* predicted motif for similarity to a known motif. A p. value describing the confidence of the motif identification is also provided along with a matrix containing the probability of each base at that position (Benner 2018b). *HOMER2* is often combined with other peak callers such as *MACS2* and the motif enrichment stage is often performed after differential accessibility analysis (Tsompana and Buck 2014).

Schepp *et al.* developed *chromVAR* (Schep et al. 2017), one of the first software packages that is dedicated to analyzing chromatin accessibility in single-cell ATAC-Seq (scATAC-Seq). In bulk or total ATAC-Seq, reads originating from thousands of cells are aggregated to get an average measure of chromatin accessibility which is not possible in scATAC-Seq due to the low copy number of nuclear DNA per cell (Schep et al. 2017). *chromVAR* instead measures the changes in chromatin accessibility in a set of pre-defined or *de novo* predicted TF motifs across a population of cells by calculating the deviation of each peak containing a motif with a set of background peaks with similar coverage and GC content. *chromVAR* can also be used to analyse the net accessibility of TF motifs in bulk ATAC-Seq data and takes as input a coordinate file containing consensus peak coordinates and alignment files. The deviation of accessibility, D , is calculated for all motifs in a position weight matrix M , against the difference in observed fragments, X , against expected fragments E :

$$D = M \left(\frac{X - E}{E} \right)$$

The deviations for a background set of peaks is similarly calculated and a bias adjusted z score is calculated for each motif as well as a false discovery rate p. value. These bias adjusted z scores can be used to infer changes in net accessibility between experimental conditions as well as identify sub-populations of variable accessibility within single cell data (Jia et al. 2018).

5.3 Profiling chromatin states for CHO cell line engineering

Sequencing mediated analysis of chromatin accessibility has significant potential utility in CHO cell engineering strategies such as identifying transgene integration sites and *cis* regulatory elements for expression engineering (O'Brien et al. 2018). The first stage in generating stably expressing cell lines (**Section 1.3**) traditionally involves random integration followed by a lengthy screening process to retain a pool of cells that have the transgene inserted within an open chromatin region and expressing at high yields. The random nature of this method means that a transgene could be integrated within heterochromatin or in an unstable genomic loci rendering the gene of interest inoperative or unstable (Veith et al. 2016). Random integration also has the risk of introducing a transgene into an open chromatin region containing important endogenous genes which would disrupt vital cellular pathways and affecting growth and protein production. Furthermore, heterogeneity is also introduced into the pool of producing cells due to the random distribution of inserted transgenes within the population. Targeted transgene insertion and integration using TALENS or CRISPR technologies offers an alternative to random integration as the gene of interest can be localized to specific genomic regions. CRISPR/Cas9 targeting efficiency, for example, has been previously demonstrated to correlate with local chromatin architecture (Zebrafish) due to steric hindrance by chromatin. Next generation sequencing methods that can profile chromatin accessibility and architecture such as ATAC-Seq be used to identify “landing pads” which are regions of stable accessibility that would readily integrate a gene of interest.

O'Brien *et al.* were the first to use ATAC-Seq to profile chromatin accessibility with respect to transgene localization sites in CHO cells (O'Brien et al. 2018). The authors used low copy number transgene integration and a titrated lentiviral vector to ensure low ratio of viral vector to CHO cells followed by RNA-Seq and ATAC-Seq of low producing pools to high producing. High IgG producing pools of cells were found to have transgenes incorporated in open chromatin with nearby, highly expressing genes (O'Brien et al. 2018). O'Brien also found that a low copy number

transgene integrated into highly active chromatin will stably produce a recombinant IgG after transfection (O'Brien et al. 2018). The pioneering work performed by O'Brien *et al.* is an important first step in identifying stable integration sites but further work is required to profile chromatin accessibility across more bioprocessing relevant conditions.

ATAC-Seq also has the potential to assist in CRISPR/Cas9 mediated cell engineering efforts as accessibility of genomic loci is positively correlated with Cas9 efficiency (Chari, Moosburner, and Church 2015a). Steric hindrance by nucleosomes protect the underlying DNA from being cut by the enzyme thereby reducing CRISPR efficiency. Furthermore, Cas9 has higher off target effects at distal loci that contain highly accessible loci (Chari, Moosburner, and Church 2015b). Chromatin remodeling due to cellular stress, such as temperature shift, can alter chromatin accessibility thereby changing the initial PAM site used for CRISPR/Cas9 engineering (Horlbeck et al. 2016). In this chapter, we present a map of chromatin accessibility where we have identified stable genomic loci. We have also generated a map of enhancer and promoter loci that will aid in future rational cell line engineering efforts.

5.4 Materials and Methods

5.4.1 CHO cell culture and temperature shift

Eight CHO cultures were seeded in SFM II media at an initial density of 2×10^5 cells/mL with four subjected to a reduction in culture temperature as described in **Section 2.4.1**. All cells were cultured for a period of 72 hours with four cultures temperature-shifted from 37°C to 31°C after 48 hours. Total DNA was extracted from all cells using non-ionic detergents using the protocol described by Corces *et al.* (Corces *et al.* 2017) where cells were centrifuged at 4°C and re-suspended in a lysis buffer containing 0.1% Tween20 in order to preferentially extract nuclei. Extracted nuclei were used as input for ATAC-Seq library preparation.

5.4.2 ATAC-Seq library preparation and sequencing

Extracted nuclei were treated with hyperactive Illumina Tn5 transposase at 37°C for 30 minutes to ensure preferential tagmentation of open chromatin as described by Buenrostro *et al.* (J. D. Buenrostro *et al.* 2013; J. D. Buenrostro *et al.* 2015a). A limited cycle PCR reaction was performed on transposase treated DNA in order to amplify “tagmented” fragments and add Illumina specific barcodes and primers for sequencing using the conditions specified in **Table 3.1**. The amplified libraries were pooled and run on an Agilent 2100 Bioanalyzer (Agilent, 5067-4626) to verify nucleosome periodicity. 2×75 bp paired end sequencing on a Illumina NextSeq instrument was performed with a minimum depth of 60 million fragments per sample to ensure adequate depth for chromatin accessibility analysis (J. D. Buenrostro *et al.* 2015a).

5.4.3 Bioinformatics data pre-processing and alignment

Raw read data was processed using *cutadapt* and *Trimmomatic* using the parameters outlined in **Section 2.4.2** in order to remove adapter contamination from fragment ends and bases with Phred quality scores less than 20. Only paired end reads with both mates passing these quality filters were retained for downstream analysis. Quality filtered reads were aligned to the RefSeq mitochondrial genome (GenBank ID: KX576660.1) using *Bowtie2* in order to remove reads originating from mtDNA using the `–very-sensitive` command and max fragment size (`-X`) set to 2000bp. Only unaligned reads where both forward and reverse reads failed to map to the

mitochondrial sequence were retained for analysis. mtDNA removed reads were aligned to the *C. griseus* *CHOK1* genome (X. Xu et al. 2011) using *Bowtie2* using the `–very-sensitive` and max fragment size parameters described above to retain high quality genomic alignments. The aligned reads were filtered to remove optical duplicates that are artefacts of sequencing as well as retained read pairs with a Phred alignment quality score greater than or equal to 30 using the *Picard* tools dataset.

5.4.4 Evaluating peak calling algorithms

De-duplicated alignment files were used as input for *MACS2* and *HMMRATAC* algorithms. *F-seq* was omitted from the analysis as it is not able to perform paired-end analysis which is necessary to properly evaluate ATAC-Seq data. Peak calling with *MACS2* was performed using single-end and paired-end modes. Single-end peak calling was performed using the `–format BAM –shift-100 –extsize 200 –nomodel –nolambda –q 0.05` options specified by the authors specifically for cutting site enrichment analysis for DNase-Seq data. Paired-end peak calling was performed using the `–format BAMPE –nomodel –nolambda –q 0.05` options to compute peak coverage for peaks using both forward and reverse reads to infer the fragment length distribution. Condition-specific peaks were identified using the *BEDtools* suite (Quinlan and Hall 2010); a peak had to be found at the same region in all four replicates of each conditions to be retained. Overlapping peaks were merged using *BEDtools* in order to retain a consensus peak set for single-end and paired-end *MACS2* output. Peak calling with *HMMRATAC* was performed after combining replicates for each condition using author recommended parameters (`–q 30 –removeDuplicates TRUE –printExclude T`). A consensus peak file was generated from *HMMRATAC* outputs by merging overlapping peaks using *BEDtools*. The consensus peak files for paired-end *MACS2* output and *HMMRATAC* output was compared to the single-end output of *MACS2* using *BEDtools intersect* in order to evaluate parameters for peak calling.

The consensus peak set generated by *HMMRATAC* was selected for differential chromatin accessibility analysis. Peaks were annotated to genic sites such as promoters, exons, introns, and transcription termination sites by the *ChIPseeker* R statistical package (G. Yu, Wang, and He 2015). The promoter region was specified as a region 1000bp upstream and 100bp downstream of the transcription start site. Peaks were classified as intergenic by *ChIPseeker* if they were localized to regions outside of gene bodies or promoters.

5.4.5 Accessibility analysis of chromatin using annotated CHIP-Seq data

CHIP-Seq data generated from a sequencing experiment by Feichtinger *et al.* that profiled the change in epigenetic states over a period of 8 days (Feichtinger et al. 2016) was downloaded from the CHO-EPIGENOME database (<http://cho-epigenome.boku.ac.at/download.html>). Data for histone marks for promoters (H3K4me3), enhancers (H3K4me1), and repressors (H327me3) for days 2, 4, 6, and 8 were downloaded along with data for an untreated control. Matched control/immunoprecipitated samples were only available for these four time points. Reads were aligned to the ENSEMBL *criGri1* CHOK1 genome using *Bowtie2* and peaks were called using *MACS2* with the parameters (-c control.bam -t experimental.bam -g 2.3e9). The *BEDtools* suite was used to generate a consensus peak set of stable promoters, enhancers, or repressors. A peak was considered as stable if it was represented in all four time points. The stable peak set for each regulatory element was overlapped with the consensus peak set derived from ATAC-Seq data.

5.4.6 Identification of active transcription with ATAC-Seq and RNA-Seq

Active transcription was estimated using ATAC-Seq accessibility data and RNA-Seq data presented in **Chapter 2**. Regions within 1kb upstream and downstream of a transcription start were inspected for the presence of an ATAC-Seq peak and segregated into “active” and “inactive” genes. A gene was classified as actively transcribed if it possessed a peak within 1kb of the TSS and inactivate if it did not. The TPM normalized gene abundances of active and inactive genes were compared to evaluate transcription patterns. Actively transcribed coding and long non-

coding genes were identified and the accessibility between the two gene sets was plotted using the *deepTools* software package (Ramírez et al. 2016).

5.4.7 Identification of stable transgene integration sites

Consensus peaks were extended in 1kb in both directions and overlapping peaks were merged to create a map of active chromatin. Broad, active chromatin sites were removed if they were located within gene bodies or in short contigs (< 1kb) and the filtered sites were quantified using *featureCounts* (Liao, Smyth, and Shi 2014) and normalized using *DESeq2* (Love, Huber, and Anders 2014). Intergenic open chromatin sites that had an average normalized accessibility > 1,000 reads were annotated as highly accessible and annotated to their nearest transcription start site. Highly accessible loci that were located within 25kb (O'Brien et al. 2018) of a gene were retained and accessibility was compared to gene expression with highly accessible loci that were located near highly expressed (average normalized expression $\geq 10,000$ reads) selected as potential stable transgene integration sites.

5.4.8 Calculating differential accessibility

The consensus peak set was quantified using *featureCounts* as previously described in **Section 3.6** in paired-end mode. Quantified peaks were normalized using the *DESeq2* package within the R statistical framework and assessed for differential accessibility. A peak was considered significantly differentially accessible if it had a fold-change $\geq \pm 1.5$ and a Benjamini-Hochberg adjusted p. value < 0.05. Significantly differentially accessible peaks were classified with respect to localization on the genome using *ChIPseeker*. Functional, biological pathway enrichment was performed for genes that contained a genic, differentially accessible peak using DAVID (D. W. Huang, Sherman, and Lempicki 2009). Differential accessibility was compared to differential expression analysis and functional pathway enrichment was repeated for genes that exhibited differential accessibility and differential expression. Differentially accessible peak coordinates were intersected with coordinates for stable promoter, enhancer, and repressor loci in order to classify function.

5.4.9 Identifying enriched TF binding sites motifs

The summits, or the region of a peak with the highest coverage, of differentially accessible loci were evaluated for the presence of transcription factor binding sites with *HOMER2* (Benner 2018b) using *-size* given parameter to evaluate the entire accessible region. *HOMER2* uses a position weight model against a compiled database of motif sequences in order to identify motifs that are enriched within genomic intervals and provides a calculated adjusted P-value as well as percent identity score that describes the similarity between predicted motif and reference motif. A motif was identified as likely significant if it possessed an adjusted p. value $< 1 \times 10^{-5}$, a percent identity ≥ 0.9 , and represented in more than 5% of target sequences.

Global changes in transcription factor variability between cells cultivated at 37°C and 31°C was assessed using the *chromVAR* statistical package within the R framework. *HMMRATAC* consensus peaks were quantified within *chromVAR* and peaks containing an average normalized accessibility $\geq 1,000$ reads were retained for motif searching. *chromVAR* uses position weight matrices provided by the high-confidence JASPAR database of motifs to search for the presence of transcription factor binding sites (Schep et al. 2017). An average deviation for each TFBS was calculated with an adjusted P-value that measures the confidence of aggregate annotation. A motif was considered differentially accessible if it was present in more than 5% of all consensus peaks and had an adjusted p. value < 0.05 .

5.5 Results

5.5.1 CHO cell culture and library preparation

Eight CHO cultures were cultivated for a period of 72 hours with four undergoing a reduction in culture temperature as described in (**Section 2.5**). We observed a reduction in growth rate with ~18% fewer cells in cultures grown at 31°C compared to those grown at 37°C (**Figure 5.4**) which was comparable to the results from cells used for RNA-Seq analysis. ATAC-Seq libraries were prepared using the Omni-ATAC sequencing protocol as described by Corces *et al.* with 50,000 cells of each culture centrifuged with a mild, non-ionic detergent to extract crude nuclei followed by treatment with Nextera Tn5 hyperactive transposase. The fragment distribution of final pooled libraries was analysed using an Agilent 2100 Bioanalyzer and we observed a non-uniform distribution with a significant number of reads found at 575bp (**Supplementary Figure 5.1**). The final pooled libraries were diluted to 8 pmol concentration and sequenced on an Illumina NextSeq instrument. All libraries were sequenced in order to attain a minimum of 60 million 2 x 75 paired-end fragments per sample.

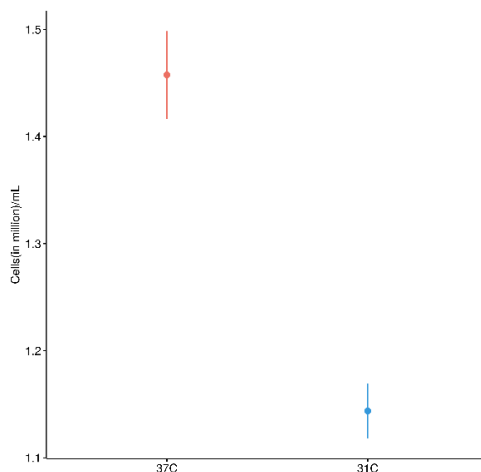


Figure 5.4 A plot of the average cell densities of 8 cultures 72 hours after inoculation. Eight CHO cultures were grown over a period of 72 hours with four subjected to a temperature shift after 48 hours. As expected, the cell densities of CHO cells that were cultivated at 31°C had an 18% lower cell density than those cultivated at 37°C. DNA from cells was extracted after cell counting for ATAC-Seq library preparation.

5.5.2 NGS data pre-processing

Raw reads were trimmed in order to remove synthetic adaptor sequences and low-quality bases with an average of ~10% of reads removed due to insufficient quality (**Table 5.1**). The quality-filtered reads were aligned to the mitochondrial genome in order to filter out reads originating from mtDNA. We observed an average of ~65% of reads originating from mitochondrial origin. The mitochondrial DNA filtered reads were aligned to the ENSEMBL *criGri1* CHOK1 reference genome and we observed over 90% read alignment in all samples.

Table 5.1 A table of raw and quality-filtered read pairs with their mapping quality to the CHOK1 genome. An average of 10% of reads were removed after filtering. Filtered reads were aligned to the mitochondrial genome with an average of 65% being identified as mitochondrial in origin and the remaining 35% aligned to the nuclear genome.

SampleID	Raw Fragments	Quality Filtered Fragments	% mtDNA
37°C-Rep1	64,793,116	58,095,428	64.16
37°C-Rep2	71,474,157	64,263,695	65.40
37°C-Rep3	65,588,299	58,801,065	64.48
37°C-Rep4	67,506,411	60,934,351	65.61
31°C-Rep1	84,523,801	76,374,506	65.96
31°C-Rep2	89,025,241	79,836,651	67.90
31°C-Rep3	69,362,462	62,901,042	65.71
31°C-Rep4	73,618,312	66,813,289	67.27

The fragment size distributions were quantified and visualized within the R statistical environment (**Figure 5.5**). The majority of fragments had insert sizes less than 147bp which corresponds to the nucleosome-free region indicative of efficient tagmentation (J. D. Buenrostro et al. 2015a). We also observed a mono-nucleosome peak with insert sizes ranging from 180-250bp but only observed a slight peak corresponding to di-nucleosome fragments (310-440bp).

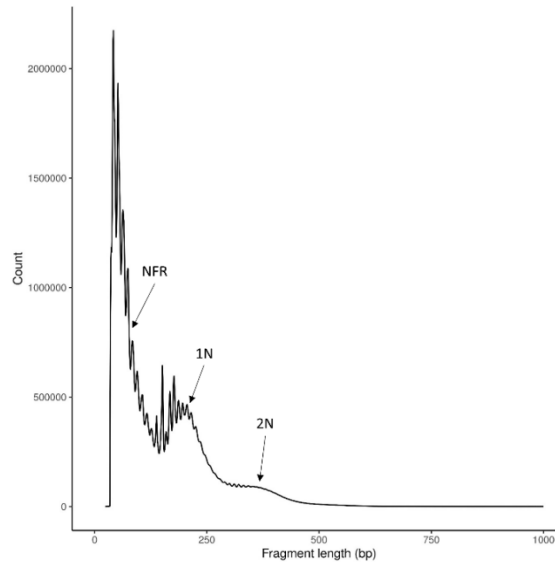


Figure 5.5 Fragment length distribution of ATAC-Seq reads. A density plot depicting the distribution of unique, mapped fragments to the *criGri1* CHOK1 genome. We observed the majority of reads having fragment lengths less than 147 which corresponds to a nucleosome free region which is indicative of efficient tagmentation. We also observed a mono-nucleosome (1N) peak and a lesser peak corresponding to di-nucleosomes (2N).

5.5.3 Evaluation of peak calling algorithms

Peaks corresponding to regions of enriched localized coverage were called using *MACS2* in single-end and paired-end mode and *HMMRATAC* using parameters recommended by the authors. Only reads with an alignment Phred score greater than or equal to 30 were used for peak calling analysis. The consensus peak sets for all peak calling algorithms was compared. *MACS2* in single-end mode with reads centered on Tn5 cutting sites had the greatest degree of concordance with both *MACS2* in paired-end mode and *HMMRATAC* with over 95% of peaks represented in all consensus sets. We repeated single-end peak calling with *MACS2* using parameters optimal for calling nucleosome positions and identified over 170,000 peaks that correspond to ostensible nucleosome positions but observed a lower degree of concordance between these peaks and those centered on cutting sites. Peaks called with *MACS2* in paired-end mode saw a variable distribution of peak lengths with a minority of peaks ~10kb in length but still over 90% concordance between single-end calls and *HMMRATAC*. We visualized peak coverage for the regions of highest coverage made with *HMMRATAC* and observed the highest

degree of coverage at the nucleosome free region with regions of enriched coverage flanking the NFR corresponding to the +1N and -1N nucleosomes (**Figure 5.6**).

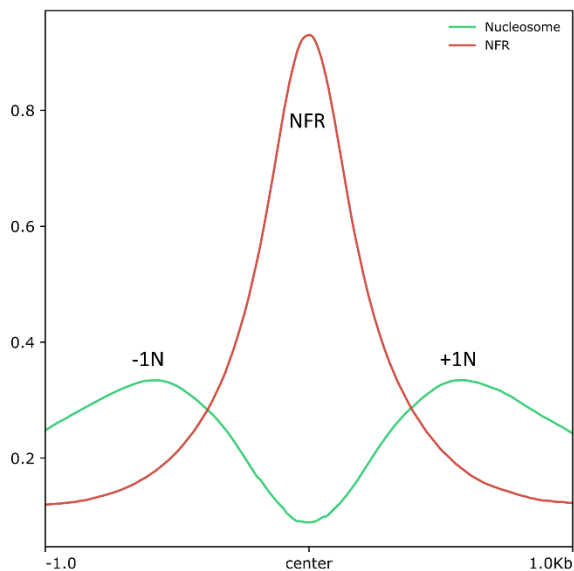


Figure 5.6 Normalized coverage, in Counts Per Million (CPM) of coverage of NFR read density compared to nucleosome peak density. CPM is defined as a factor of the number of reads per million for each gene or locus. A density plot of normalized peak coverage surrounding peak summits plotted as a fraction of the peak maxima. The peak centre corresponds to the nucleosome free region with a lower coverage flanking regions upstream and downstream corresponding to nucleosomes. Approximately 58.36 of all peaks were located within gene bodies (in introns or exons) or in the promoter or TTS. Intergenic peaks found in distal regions compose the largest individual class.

5.5.4 Classification of peaks

Peaks identified by *HMMRATAC* were annotated to the closest TSS using the *ChIPseeker* R statistical package. A peak was classified as genic falling within the gene body or within 1000 bp of a gene and was classified as intergenic if located in distal genomic location. Intergenic and intronic peaks were the two largest classes of peaks with the remainder located within the promoter, which was defined as the region 1kb upstream and 100 bp downstream of the TSS, and exons (**Figure 5.7**). Only a small minority of peaks were located in regions immediately downstream (< 300bp) of the transcription termination site. We observed enrichment of peaks around transcription start sites with 43% of all peaks located within 10kb of a TSS.

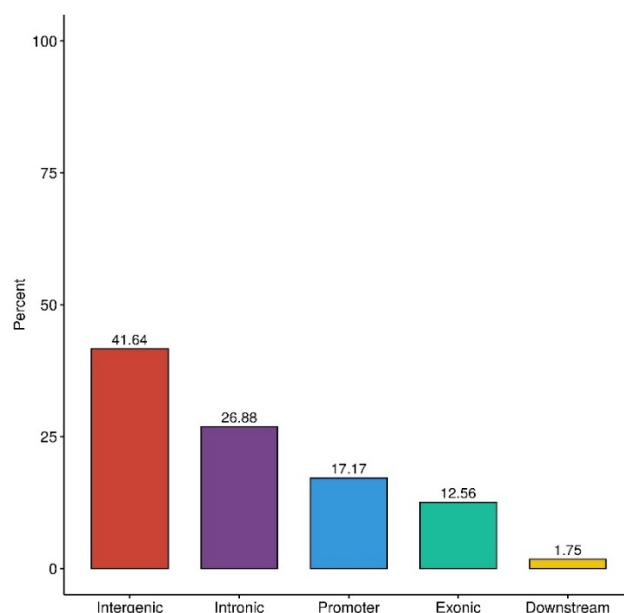


Figure 5.7 Distribution of peaks across the CHOK1 genome with respect to a TSS. Approximately 58.36 of all peaks were located within gene bodies (in introns or exons) or in the promoter or TTS. Intergenic peaks found in distal regions compose the largest individual class. Promoter peaks were defined as those found 1 kb upstream of the TSS or downstream 100 bp.

5.5.5 Active promoter and enhancer histone marks

Pre-processed alignments for matched control and CHIP-Seq data from four time points: 5, 9, 13, and 17 were downloaded from the CHO-epigenome database (<http://cho-epigenome.boku.ac.at/download.html>) for H3K4me3 (promoters), H3K4me1 (enhancers), and H3k27m3 (repressors). The other time points were omitted due to a lack of matched control samples containing untreated DNA. Peaks for all three histone marks were called for all time points and a consensus peak set of stable promoters and enhancers was generated. A total of 49,609 peaks containing H3K4me3 stable promoter histone marks and 120,740 H3K4me1 stable enhancer marks were identified with, 76% of all putative stable promoters overlapping with putative enhancer sites. Coordinates for ATAC-Seq consensus peaks were compared to promoter and enhancer loci with 14% of promoter and 30% of enhancer histone modification sites not represented by a consensus peak. We identified a total of 44,948 active enhancer sites that do not overlap H3K4me3 promoter sites and contained a peak within the ATAC-Seq data. There was

poor overlap between H3K27me3 silencer peaks and ATAC-Seq with only 126 of ~44,000 putative silencer peaks found to overlap with ATAC-Seq peaks.

5.5.6 Integrated RNA-Seq and ATAC-Seq to find TSS with active transcription

Previous experiments have demonstrated correspondence between accessibility near the transcription start site and gene expression (Gendron et al. 2019). Gene expression data from **Chapter 2** was compared to ATAC-Seq chromatin accessibility data. A total of 14,077 genes contained a peak within the TSS while 23,839 genes did not contain ATAC-Seq signal at the TSS. Genes that contained an ATAC-Seq peak at the TSS were defined as actively transcribed and we observed a higher proportion of genes expressed above a TPM threshold of 1. The converse was observed with genes that lacked a peak at the TSS with the majority exhibiting very little expression (**Figure 5.8**). We repeated this analysis by removing peaks located within 1kb of the transcription start site in order to remove potential distal promoters and found the same density distribution of RNA-Seq abundances between the two groups.

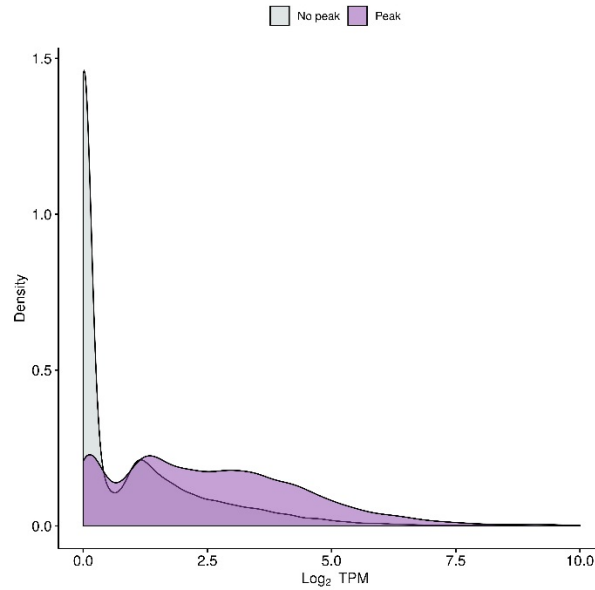


Figure 5.8 A density plot comparing gene normalized TPM of genes that contain a peak near the TSS and genes that do not. Significantly more genes without a peak near the TSS were found to have lower average expression. Average expression of genes that contain an ATAC-Seq derived peak was higher.

We next compared TSS accessibility between coding and long non-coding RNA genes. Coding genes have 12,068 genes (~61%) that overlap the TSS which is a higher proportion compared to lncRNAs that have 1,266 (~51%) genes. We examined the accessibility at the TSS between mRNAs and lncRNAs and found that coding genes had greater accessibility within 1kb of the gene start compared to lncRNAs (**Figure 5.9**) which is comparable to gene expression distributions as lncRNAs were observed to be expressed at lower levels than protein coding genes.

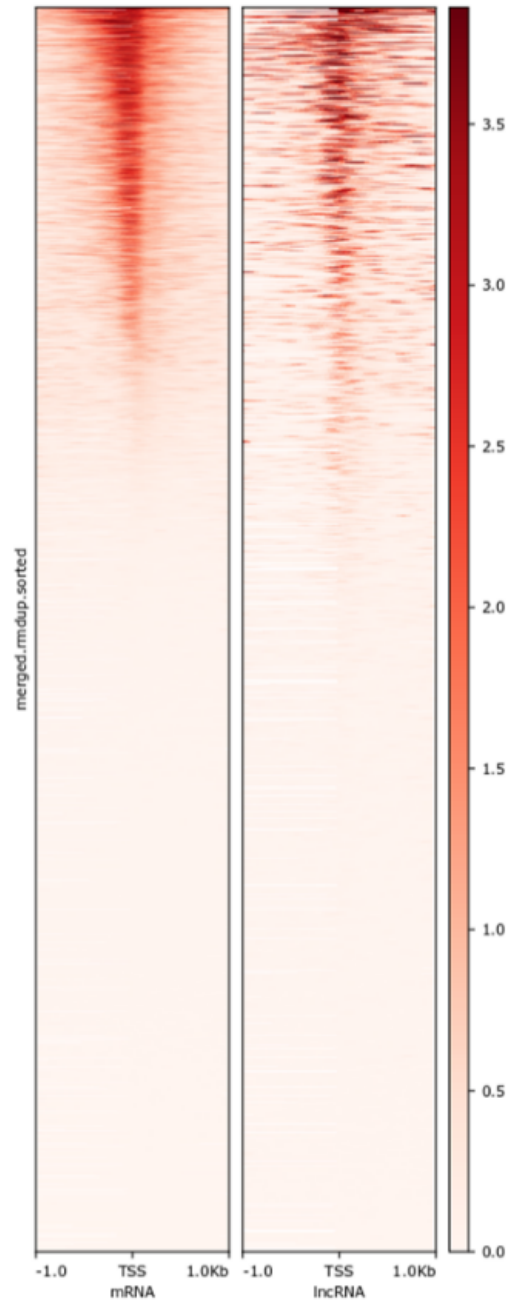


Figure 5.9 Heat maps comparing average normalized coverage of coding and lncRNA genes around the TSS. Average accessibility around mRNA TSSs was found to be significantly higher than average accessibility around the TSS of lncRNAs.

We observed evidence of active transcription 398 evolutionarily conserved lncRNAs including *MALAT1*, *HOTAIRM1*, *LINC00511* which all contain a peak at the transcription start site. Only 21 sense lncRNAs exhibit evidence of active transcription compared to 333 antisense genes whereas a total of 1,116 intergenic genes exhibited a peak at the transcription start site.

5.5.7 Identification of stable integration sites using ATAC-Seq data

In order to identify stable integration sites, peaks were extended by 1kb in both directions and overlapping peaks merged in both temperature conditions to identify open chromatin states. We compared the genomic coordinates for open chromatin with transcript coordinates using the assembly generated in **Section 2.6** and retained a total of 24,843 intervals that were common in both temperature conditions that did not overlap a gene.

Intergenic intervals were quantified and normalized and we identified 161 highly accessible, broad genomic regions with an average accessibility $\geq 1,000$ reads which were annotated to the nearest TSS. We identified two broad, open chromatin regions that are within 25kb of highly expressed (*DESeq2* normalized average expression $> 10,000$) genes. The first site is a ~21kb locus that is located between the *DAG1* and *NICN1* genes. This interval contains a differentially accessible promoter peak (fold-change: -1.55, Adj. P-value: 3.95×10^{-17}) but only composes 3% of the total interval with the remainder exhibiting stable accessibility between temperature conditions. This interval contains promoter and enhancer histone modifications with the summit of the highest peak contained within an enhancer region.

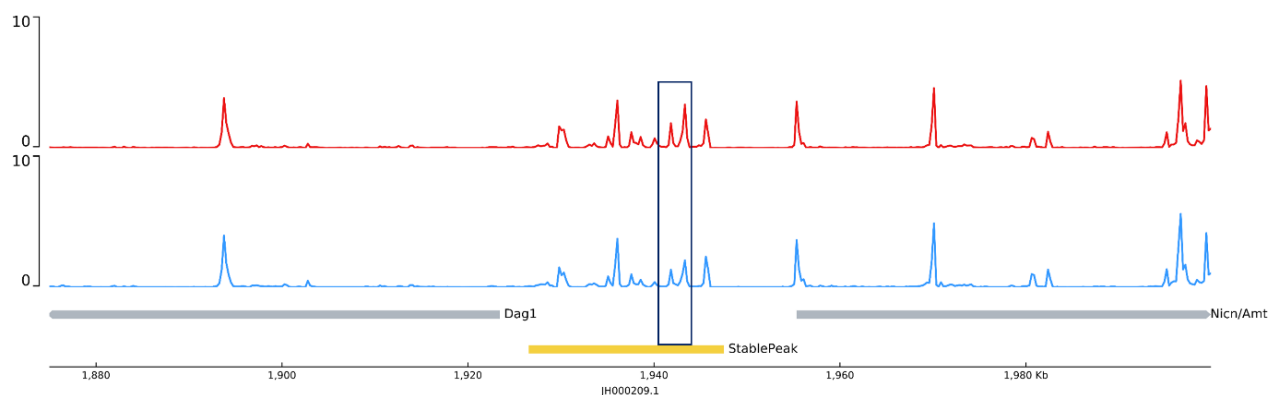


Figure 5.10 The average normalized coverage plot of ATAC-Seq data spanning the Dag1-Nicn/Amt locus. Coverage has been CPM normalized. A 770bp region of the larger open chromatin region was found to be differentially accessible (highlighted in box).

The second site is a ~6.6kb locus that is found immediately upstream of the *Sacs* gene. There are no differentially accessible sites within this interval which similarly contains enhancer and promoter histone modifications.

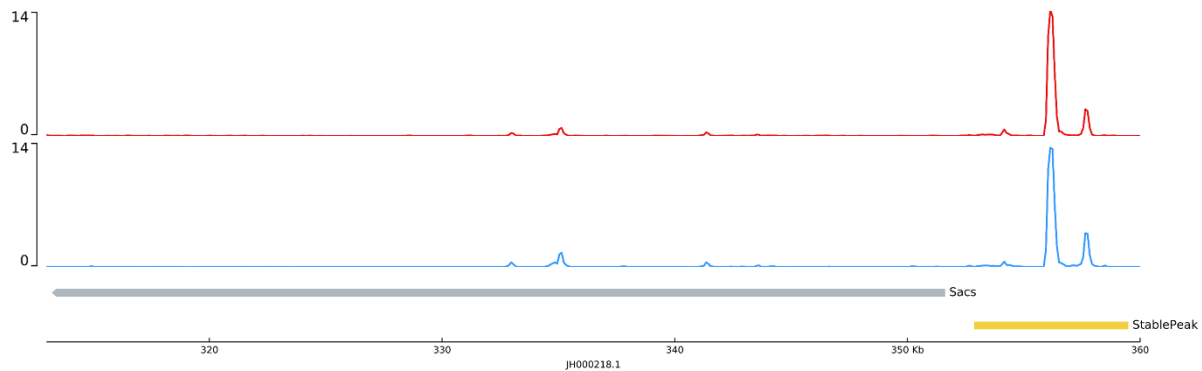


Figure 5.11 The average normalized coverage plot of ATAC-Seq data upstream of the *Sacs* gene. Coverage has been CPM normalized. The region of the highest peak corresponds to the promoter.

5.5.8 Global change in chromatin accessibility after temperature shift.

We observed a global change in peak accessibility after a reduction in culture temperature from 37°C to 31°C. The consensus peak file derived from *HMMRATAC* peak calling was used to quantify reads within peaks. We observed a comparatively high FRiP score with an average of 67% reads assigned to a called peak for each sample. We initially observed that peak accessibility is significantly associated with culture condition (**Figure 5.12**) with 68% of variance corresponding to the first principal component.

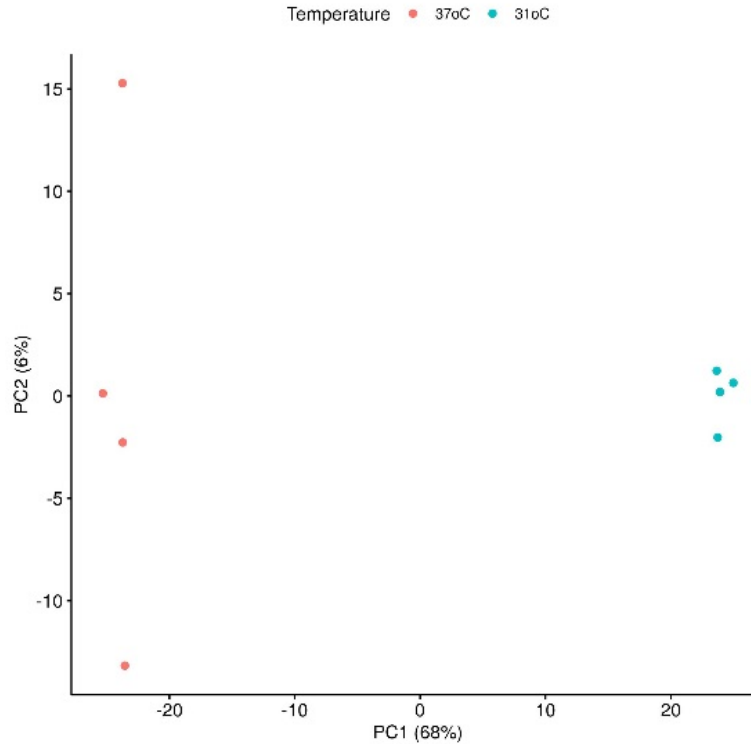


Figure 5.12 A principal component analysis plot visualizing variation between samples. 68% of variability is captured by the first principal component which can be explained by changes in accessibility following temperature shift.

Differential peak accessibility was evaluated with *DESeq2* with a fold-change threshold $\geq \pm 1.5$ and a Benjamini-Hochberg adjusted P-value < 0.05 (Love, Huber, and Anders 2014). A total of 14,187 peaks (~15% of all peaks) were found to be significantly differentially accessible after temperature shift with 8,599 peaks exhibiting increased accessibility and 5,589 peaks exhibiting decreased accessibility (**Figure 5.13**).

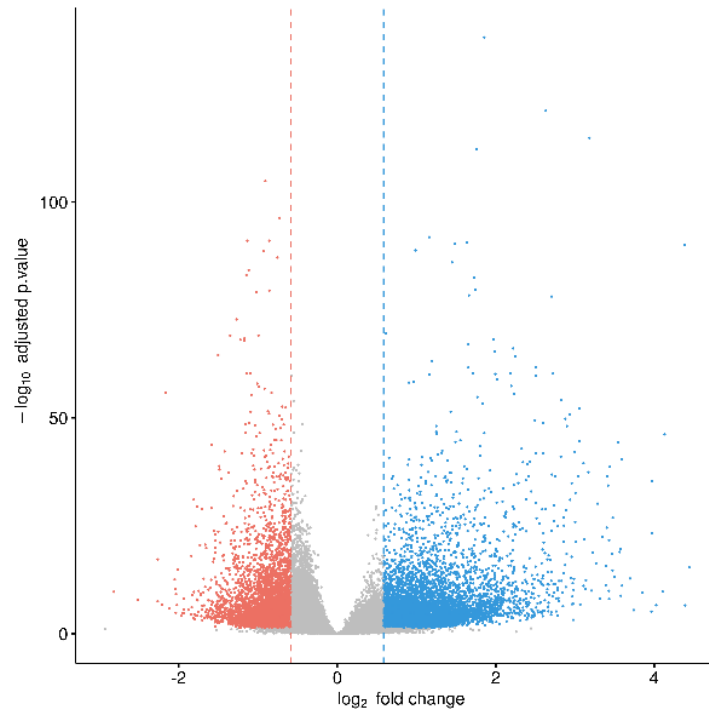


Figure 5.13 A volcano plot comparing log₂-fold-change with adjusted P-values. We observed genome-wide chromatin remodelling after a reduction in culture temperature with a higher proportion of open chromatin regions compared to closed.

The largest individual group of differentially accessible peaks were located within intergenic regions composing approximately 50% of all differentially accessible (DA) peaks. A smaller percentage (7.7%) of peaks located within the promoter were found to be differentially accessible whereas intronic peaks were the largest genic group that were significantly differentially accessible (~30%) with the remaining 13% being composed of exonic (~12%) and TES overlapping (~1.5%) peaks. *Cirbp* and *Rbm3* contained highly accessible peaks at the TSS but were not found to be differentially accessible in the ATAC-Seq data. **Supplementary Table 4.1** summarizes all differentially accessible peaks with the respective gene classification.

5.5.9 Comparison between differentially accessible peaks and RNA-Seq

Significantly differentially accessible peaks annotated with *ChIPseeker* were compared to differentially expressed protein coding genes identified in **Section 2.6** using the *DAVID* gene ontology enrichment suite (D. W. Huang, Sherman, and Lempicki 2009). Only differentially accessible peaks located within genic regions (promoters, introns, exons, and TES) were selected

for gene ontology enrichment. All gene set enrichments with a Benjamini-Hochberg adjusted P-value < 0.05 were considered significantly enriched. **Table 5.2** summarizes the ten most enriched biological processes as well as provides the number of genes that are simultaneously differentially. Cellular pathways involved in cellular division and cellular proliferation were significantly enriched. **Supplementary Table 4.2** contains all enriched gene ontology bioprocess terms.

Table 5.2 A table of the top ten most enriched gene ontology cellular processes. Differentially accessible peaks were annotation to the nearest TSS. Genes that were simultaneous differentially expressed and differentially accessible were selected with only genic sites used for GO enrichment. The number of genes in each process is detailed along the a Benjamini-Hochberg adjusted False Discovery Rate (FDR) value.

Term	Process	Count	FDR
GO:0007059	Chromosome segregation	20	4.18×10^{-4}
GO:0022402	Cell cycle process	40	2.42×10^{-3}
GO:1902589	Single-organism organelle organization	53	3.11×10^{-3}
GO:0006928	Movement of cell or subcellular component	53	3.69×10^{-3}
GO:0035556	Intracellular signal transduction	66	4.67×10^{-3}
GO:0000280	Nuclear division	26	6.76×10^{-3}
GO:0007049	Cell cycle	46	8.25×10^{-3}
GO:0000278	Mitotic cell cycle	31	9.60×10^{-3}
GO:0007265	Ras protein signal transduction	18	9.92×10^{-3}
GO:0040011	Locomotion	47	1.72×10^{-2}

We have also identified a cohort of 101 significantly differentially expressed lncRNA genes that was annotated to a differentially accessible peak with 66 identified in genic regions. The number of differentially expressed/differentially accessible pairs is lower than protein coding genes. There are 16 differentially expressed lncRNA genes that contain differentially expressed peaks at the promoter with 6 genes exhibiting low average expression (average normalized read count < 200). The most abundant lncRNA is a novel intergenic single-exon gene that is 1.7 fold downregulated (Adjusted P-value 1.67×10^{-09}) that contains a downregulated peak that completely encompasses this gene. No stable enhancer, repressor, or promoter peaks from CHIP-Seq data derived from Feichtinger *et al.* were found overlapping this gene (Feichtinger *et al.* 2016). Two genes within

this subset have been previously annotated within the ENSEMBL *criGri1* CHOK1 V98 database with ENSCGRG00000026258 found on a short contig (2,685bp) and ENSCGRG00000026238 found antisense to the *Mroh2b* protein coding gene. Both of these genes are expressed at comparatively lower abundances.

5.5.10 Motif enrichment with HOMER2

Significantly differentially accessible peaks were searched using the *HOMER2* algorithm in order to identify commonly enriched transcription factor binding site motifs (Benner 2018b). The regions of a peak corresponding to the highest coverage, referred to as a summit, were selected for motif enrichment analysis as this region equates with a nucleosome-free region (NFR). A motif had to be present in $\geq 5\%$ of peaks, have a percent similarity $\geq 90\%$, and an adjusted P-value $< 10^{-5}$. Motif alignments were visually inspected and low-quality alignments were removed and a high confidence set of transcription factor binding site (TFBS) motifs were selected for analysis (**Table 3**).

Table 5.3 A table containing motifs that are highly enriched in differentially accessible peaks. This table summarizes the transcription factor motif sequence as well as the percent similarity to the reference motif database and the adjusted P-value.

TF	Motif	% Similarity	Log P-value
AP-1		99.0	1×10^{-1819}
TEAD		91.6	1×10^{-245}
KLF		98.6	1×10^{-134}

We observed that the *AP-1* transcription factor binding site is enriched in greater than 40% (compared to a 6% enrichment in background regions) of all significantly differentially accessible. *AP-1* sites are equally represented within peaks that are openly accessible as well as having reduced accessibility after a reduction in culture temperature. We identified a total of 3,673 *AP-1* TFBS that contain a differentially accessible peak. *AP-1* is a large class of transcription factors that are composed of complexes between *FOS*, *JUN*, and *ATF* proteins and bind to the DNA motif containing the sequence TGAVTCA. *FOS* and *JUN* are significantly downregulated after a reduction in culture temperature whereas *Fos/1* is modestly upregulated with a fold-change of 1.45 and Benjamini-Hochberg adjusted P-value of 7.06×10^{-17} suggesting that this TF is actively binding *AP-1* motifs. *JUND* is not annotated within the ENSEMBL *criGri1* CHOK1 genome so we could not observe any expression. The majority (53%) of all significantly differentially accessible *AP-1* containing peaks were located in distal intergenic regions with the second largest group (30%) located within introns. Differentially accessible putative *AP-1* transcription factor binding sites were distributed in genic regions within lncRNAs and mRNAs. Genic sites containing *AP-1* binding motifs that exhibit increased accessibility saw an upregulation of partner genes whereas decreased accessibility equated with decreased expression. Four upregulated antisense

transcripts containing *AP-1* transcription factor binding sites were found to be upregulated with increased accessibility at the promoter including the ON/OFF antisense transcript *AIG1-AS* and the previously annotated *MROH2B-AS* gene (ENSCGRG00000026238). The converse was not observed for differentially accessible TSSs containing an *AP-1* motif. Five differentially expressed genes were identified with three being mono-exonic and two containing two exons or more. The mono-exonic gene expression correlated positively with TSS accessibility whereas the multi-exonic genes correlated inversely.

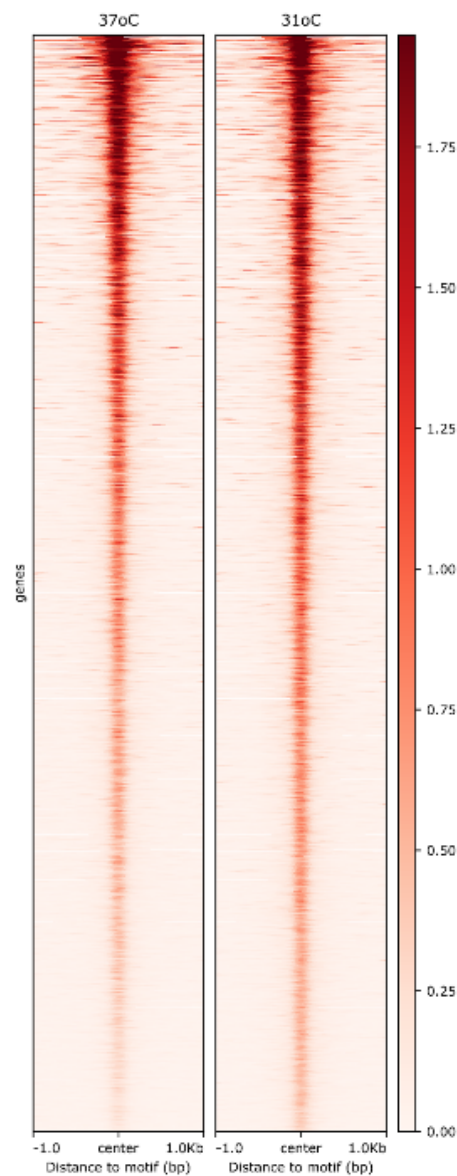


Figure 5.14 A heat map comparing average normalized coverage within 1kb of significantly DA FosI1 binding sites. FosI1 sites are the most enriched motif and found equally differentially accessible between open and closed chromatin.

We observed that *TEA domain* binding sites were significantly upregulated after a reduction in culture temperature 37°C to 31°C. The *TEAD* family are composed of four genes: *Tead1*, *Tead2*, *Tead3*, and *Tead4* and all the *TEA* domains share a core NAGGAATGY motif. *Tead4* is the only significantly differentially expressed member of this family with a fold-change of 1.55 and a Benjamini-Hochberg adjusted P-value of 1.85×10^{-175}

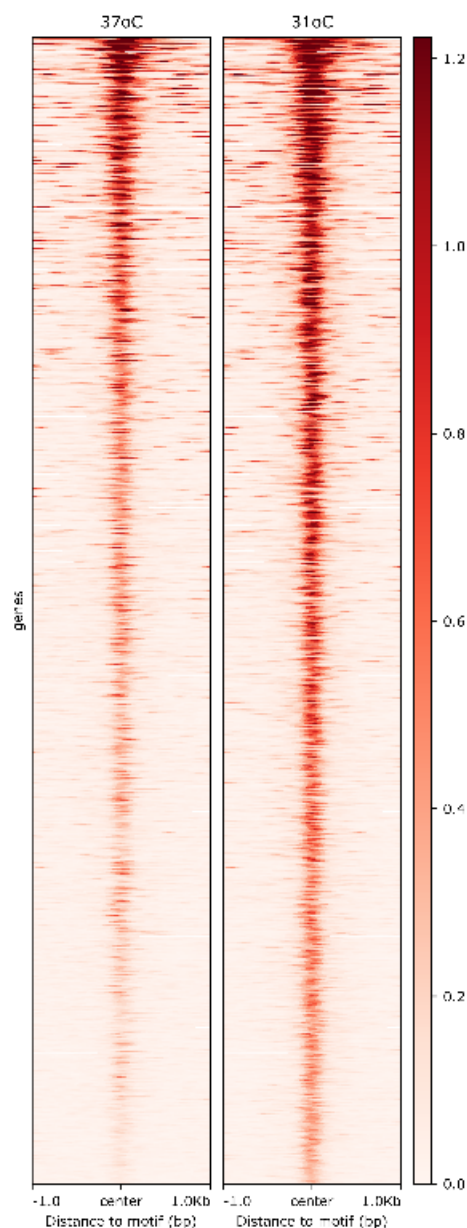


Figure 5.15 A heat map comparing average normalized coverage within 1kb of all significantly DA TEAD4 sites. TEAD4 domain containing peaks are approximately 20% more accessible after a reduction in culture temperature. TEAD4 was also found to be 1.55 upregulated within RNA-Seq data.

We observed that the KLF4 binding sites were significantly upregulated after a reduction in culture temperature from 37°C to 31°C. The Kruppel Like Factor family are evolutionarily conserved and are composed of 17 KLF transcription factors within humans with 10 homologous genes annotated within the ENSEMBL *criGri1* CHOK1 genome. HOMER2 identified the core KLF binding site is shared with *Klf1*, *Klf12*, and *Klf4*. Kruppel Like Factor 4 is the only one of this set

that is expressed with *Klf1* and *Klf12* having zero expression. *KLF4* was also identified as highly upregulated after a reduction in culture temperature with a fold-change of 3.38 and a Benjamini-Hochberg adjusted p. value of 4.92×10^{-1135} .

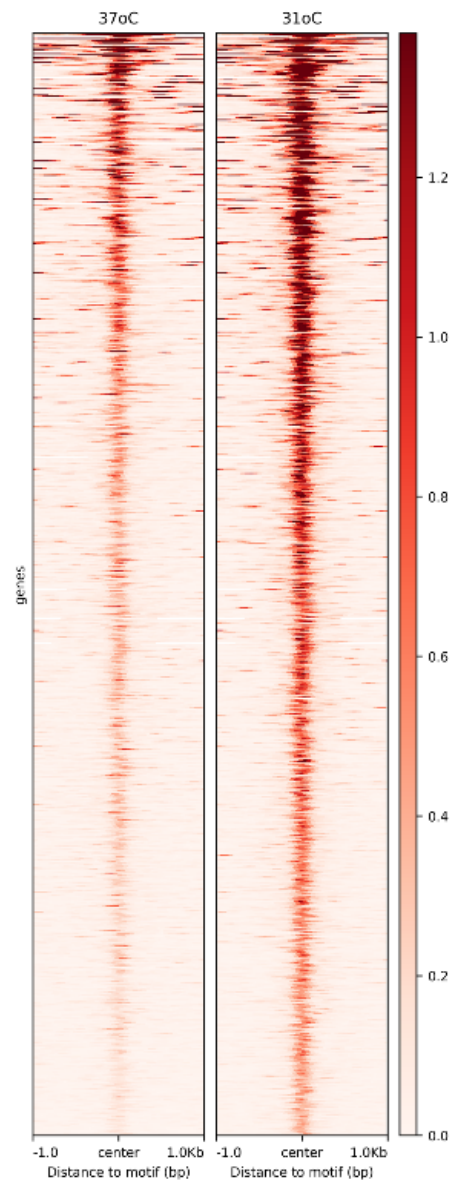


Figure 5.16 A heat map comparing average normalized coverage within 1kb of all significantly DA Klf4 sites. Kruppel like factors are the largest group of significantly differentially accessible TFs. Klf4 was also found to be 3.38 upregulated within RNA-Seq data.

5.5.11 Total deviation of transcription factor motifs with *chromVAR*

The consensus peak set generated by *HMMRATAC* was quantified with aligned read data using *chromVAR*. 128 mouse position weight matrices from the *JASPAR* database were used to search the *criGri1* CHOK1 genome and a deviation score between cells cultivated at 37°C and 31°C was calculated for each aggregate motif with 70 motifs having a significant Adj. P-value < 0.05. Significant peaks represented in a low fraction of peaks (< 5%) were removed in order to identify significantly differentially accessible motifs. *Kruppel like factors* 1, 4, and 12 were some of the most significantly upregulated transcription factor binding sites which correlated with the results of *HOMER2* where KLFs are significantly enriched. *chromVAR* analysis also found that the *AP-1* transcription factor family that encompasses *FOS:JUN*, *Nfe2l2*, and *Bach1:Mafk* transcription factors was the most significantly enriched motif within the consensus peak set but classified the family as downregulated after temperature shift. The *Gabpa* transcription factor motif is similarly highly differentially accessible and found in 5.73% of all peaks but identified in only 1.73% of all motifs with *HOMER2* and was therefore eliminated. The *Myog*, *Tcf12*, and *Myod1* are all members of BHLH family of transcription factors. The *TEAD* transcription factor family is not present within the *chromVAR* analysis as there is no *Mus musculus* *TEAD* position weight matrices present in the reference database

Table 5.4 A table of the top ten most enriched motifs identified by chromVAR. This table summarizes the difference in mean deviation between culture conditions as well as the percent of peaks that contain a specific motif. KLF sites were found to have the highest increase in aggregate accessibility while AP-1 sites containing FOS:JUN binding sites were the most enriched.

TF motif	Δ mean deviation	% of peaks with motif	Adj. P-value
<i>Klf1</i>	0.04	11.04	6.39×10^{-4}
<i>Gabpa</i>	0.03	5.73	2.82×10^{-3}
<i>Klf12</i>	0.03	9.29	9.74×10^{-4}
<i>Klf4</i>	0.03	14.35	8.84×10^{-4}
<i>Bach1::Mafk</i>	-0.02	12.90	8.46×10^{-3}
<i>FOS::JUN</i>	-0.02	37.91	6.44×10^{-3}
<i>Nfe2l2</i>	-0.02	13.70	1.56×10^{-2}
<i>Myog</i>	-0.02	6.61	3.79×10^{-3}
<i>Tcf12</i>	-0.01	6.48	6.44×10^{-3}
<i>Myod1</i>	-0.01	5.71	1.40×10^{-2}

5.6 Discussion

In this Chapter, we present the first map of global chromatin accessibility of the Chinese hamster ovary cell genome. We utilized the Assay for Transposase Accessible Chromatin in order to probe local chromatin organization to compare cells cultivated at 37°C and 31°C. ATAC-Seq is a very useful method to probe chromatin architecture in CHO cells due to ease of library production and tolerance for low cell numbers (J. D. Buenrostro et al. 2013). One significant *caveat* in library preparation is the significant amount of mtDNA that is extracted during the initial differential centrifugation step. We observed an average of 60% of mtDNA read contamination across all conditions and replicates which is within the parameters reported by Buenrostro *et al.* (J. D. Buenrostro et al. 2015c). mtDNA contamination is highly dependent on cell and tissue type and mtDNA contamination as high as 80% of total data has previously been reported (Montefiori et al. 2017). Montefiori *et al.* presented a protocol that adds CRISPR-mediated cleavage of mtDNA in addition to differential centrifugation with a non-ionic detergent in order to efficiently enrich for nuclear DNA (Montefiori et al. 2017). The addition of a CRISPR-mediated mtDNA reduction step, while useful, would obviate one of ATAC-Seq's greatest virtues: its simplicity and quick preparation. ATAC-Seq library preparation from DNA extraction to final library synthesis requires can be completed within four hours which contrasts very favorably with more laborious methods like MNase-Seq (J. D. Buenrostro et al. 2013). Corces *et al.* developed a modified protocol that added digitonin and Tween 20% to the non-ionic detergent to reduce mtDNA contamination (Corces et al. 2017). We still observed a high degree of reads originating from mtDNA with this method, however, and future ATAC-Seq experiments will need to evaluate different ratios of detergents in order to identify the optimal mixture for probing chromatin accessibility in CHO cells.

5.6.1 ATAC-Seq quality control metrics from sequencing data

Despite having to discard 60% of the sequencing data, we were still able to achieve a high-quality fragment length distribution. We observed a single, large peak containing the majority of sequenced reads which correspond to the nucleosome free region. We were also able to identify a mononucleosome peak that spans 180-220bp region. The initial protocol developed by Buenrostro *et al.* demonstrated multi-nucleosome peaks that spanned up to 6 nucleosomes (J. D. Buenrostro et al. 2013). Improving DNA extraction methods will undoubtedly provide clearer signal for higher order nucleosomes in CHO data. Nucleosome positioning data can be useful in identifying heterochromatin (Zhou, Gaullier, and Luger 2019a). The NFR-enriched fragments are extremely useful in extrapolating promoters and enhancer regions.

5.6.2 Evaluation of two peak calling algorithms

We compared peak calling algorithms using *MACS2* and *HMMRATAC* and found a great deal of concordance between *MACS2* in single end mode and *HMMRATAC* (Thomas et al. 2017; Tarbell and Liu 2019). The majority of published studies use *MACS2* in single-end mode which discards half the read data to selectively enrich for cutting sites (Gaspar 2018). The remaining data is extended uniformly from the start of the cutting site in order to theoretically enrich for nucleosome free regions. This assumption of uniform fragment length distribution is not viable for ATAC-Seq as fragments of variable length are desired in order to capture NFR and reads flanking nucleosomes. Peaks called using *HMMRATAC* had high percentage of overlap with single-end *MACS2* peak calls. *MACS2* using paired-end options had the greatest degree of divergence between the peak callers but > 90% similarity was still observed between consensus peak sets. *HMMRATAC* was able to recapitulate nucleosome architecture with a single high coverage NFR region flanked by two bimodal nucleosome peaks. More ATAC-Seq specific data analysis tools are necessary for processing datasets.

Peak distribution was found to be dispersed around the genome with all peaks located within 100kb of a transcription start site. The genomic distribution of peaks is concordant with DNase-

Seq datasets where the majority of hypersensitivity sites are found in intergenic regions within 200kb of a transcription start site (A. Chen, Chen, and Chen 2018). Extrapolating function for intergenic peaks is difficult due to the three dimensional nature of chromatin complexes; an enhancer region that is located 50kb from a transcription start site might affect gene transcription due to the formation of long DNA loops that bring distal elements near promoters (Klemm, Shipony, and Greenleaf 2019a). Extrapolating regulatory function from genic loci is comparatively easier as RNA-Seq data can be combined in order to find co-expression with accessibility. The largest group of genic *cis* regulatory elements with some of these loci functioning as internal promoters during alternative splicing (Vacik 2017a). Alternative splicing has been shown to be tissue and condition specific and an intronic promoter could be the mechanism that drives differential isoform transcription (Vacik 2017b). These intronic CREs could also function as promoters, enhancers, and silencers for antisense and sense lncRNAs that are contained within the gene body. Ritter *et al.* found that an equivalent number of coding and non-coding genes have internal exonic enhancer regions that drive host gene transcription amplification (Ritter *et al.* 2012).

5.6.3 Regulatory context of open chromatin regions remains difficult

De-convoluting peak localization with regulatory function as peak annotation often identifies regulatory targets to its nearest TSS. This assumption is comparatively simple to evaluate for solitary genes where the nearest *cis* regulatory element is highly likely to modulate expression but is not necessarily valid for gene dense genomic regions (Kondili *et al.* 2017). Complex loci containing many sense/antisense pairs would contain numerous internal enhancers and silencers that could regulate multiple target genes within the genomic interval (Karnuta and Scacheri 2018). Tools such as *ChIPseeker* were developed to analyse CHIP-Seq data derived from histone modifications where the nearest TSS is a highly likely candidate as regulatory targets (G. Yu, Wang, and He 2015). New software tools need to be developed that can reliably integrate chromatin accessibility data.

We used existing CHIP-Seq data derived from a study by Feichtinger *et al.* to annotate CREs within the genome (Feichtinger et al. 2016; Hernandez et al. 2019). One significant computational challenge is the lack of a unified, well-annotated genome that utilizes a single reference annotation. Sequencing data was originally aligned to a chromosome sorted genome (Brinkrolf et al. 2013) and later to a more complete genome derived from long read DNA sequencing (Rupp et al. 2018) with each genome containing different levels of annotation. We had to re-process data by aligning CHIP-Seq reads to the ENSEMBL *criGri1* CHOK1 genome (X. Xu et al. 2011) in order to combine RNA-Seq data and gene annotation information. A unified reference sequence with a single, high-quality reference annotation from repositories such as ENSEMBL would simplify peak calling and feature annotation.

5.6.4 Active chromatin states within previously annotated regulatory regions

We observed significant overlap between H3K4me3 promoter histone modifications and ATAC-Seq peaks with only ~14% of promoter histone modifications not containing an accessible peak. H3K4me3 is a well-characterized epigenetic mark that is enriched near transcription start sites and is correlated with active transcription (Lawrence, Daujat, and Schneider 2016a) which demonstrates that ATAC-Seq can be reliably utilized to identify active promoter regions. There are significantly more enhancer and repressor modifications than promoter modifications and are interspersed throughout the genome. ATAC-Seq had a 70% concordance with enhancer histone modifications and can be used to identify active enhancer loci. We have combined ATAC-Seq with previously sequenced CHIP-Seq data to generate chromatin maps of highly accessible enhancer regions in distal genomic intervals. ATAC-Seq had poor concordance with H3K27me3 silencer data, however. Silencers are poorly characterized compared to enhancers due to the lack of a consistent histone modification; the H3K27me3 histone modification does not necessarily correlate to transcriptional repression in the same way that H3K4me1 correlates to transcriptional activation (Jayavelu et al. 2018). The initial experiments in ATAC-Seq were greatly enhanced by detailed histone and transcription factor binding states derived from multiple tissues and

conditions performed in model organisms such as humans and mice (Chang et al. 2018). There are significantly fewer epigenetic datasets available for CHO cells. The few datasets that are available are more limited in that they profiled single bioprocessing conditions. More sequencing experiments are needed in order to generate high-resolution chromatin accessibility and modification maps.

5.6.5 Application of ATAC-Seq to estimate active transcription at the TSS

Chromatin accessibility has been demonstrated to be correlated with the level of gene transcription by Dendron *et al.* (Gendron et al. 2019). We utilized this property to estimate active transcription after integrating ATAC-Seq and RNA-Seq data. We observed that the presence of a peak within 1kb of a TSS is correlated with higher average expression compared to genes that lack peaks. The lack of a peak does not necessarily equate with transcriptional quiescence, however, as we still observed a smaller proportion of genes that are expressed above a TPM threshold of 1. This observation can potentially be attributed to distal promoters and enhancers that act to drive transcription of a gene (Halfon 2019a). We compared accessibility at the TSS between coding genes and lncRNA genes and found that mRNAs exhibited a higher average accessibility than non-coding genes. RNA-Seq data previously demonstrated that lncRNAs are expressed at lower abundances than their protein coding counterparts so the accessibility around the TSS is correspondingly lower.

5.6.6 Combined ATAC-Seq and RNA-Seq data to identify stable sites for transgene integration

O'Brien *et al.* previously demonstrated the utility of combined ATAC-Seq and RNA-Seq to identify the characteristics of lentiviral integration sites for low copy number, high expression transgene expression (O'Brien et al. 2018). In this study, we similarly integrated RNA-Seq and ATAC-Seq data to identify highly accessible and stable loci within the CHO genome. We identified two stable

genomic regions that are located within 25 kb of highly expressed genes. We selected intergenic sites over those found within gene bodies to minimize potential disruption of transcription (Gaidukov et al. 2018). Open chromatin is enriched in gene bodies with a higher average accessibility compared to intergenic regions devoid of genes. Further *in vitro* validation of the two identified sites is necessary to determine whether transgene integration at putative regulatory elements will fundamentally disrupt cell growth or recombinant protein productivity. Furthermore, it is as yet unclear whether these sites are stable across different cell lines and future ATAC-Seq experiments are essential in identifying potential common integration sites.

5.6.7 Temperature shift induces global alterations in chromatin accessibility

A global change in chromatin accessibility was observed after a reduction in culture temperature from 37°C to 31°C and average peak accessibility was found to be condition-specific. Approximately 15% of all consensus peaks were found to be significantly differentially accessible which suggest a genome-wide regulatory response after temperature shift. This observation is supported by RNA-Seq data which identified 2,000 genes as highly expressed after temperature shift. Gene ontology annotation with *DAVID* of differentially expressed genes that contained differentially accessible genic peaks found that pathways involved in DNA replication and cell cycle regulation were the most enriched. The inclusion of significantly DA intergenic peaks added spurious enrichment for cell differentiation and neuronal development. Cell cycle arrest and reduction in cell division commonly occur in CHO cells that are subjected to mild hypothermia as the cells convert metabolic pathways to consume lactate for cell survivability (Bedoya-López et al. 2016).

5.6.8 Enrichment of TF binding sites in DA peaks

All significantly differentially accessible peaks were evaluated for motif enrichment using *HOMER2* which identified *AP-1*, *KLF*, and *AP-2* transcription factor families as differentially accessible. This analysis was repeated using consensus peak set with *chromVAR*. There was significant agreement between *HOMER2* and *chromVAR* analysis as both methods identified the

core *AP-1* motif as significantly enriched. *HOMER2* provides a list of similar motifs that share a percent identity with the *de novo* predicted motif which correspond to transcription factor families (Benner 2018a). *chromVAR*, by contrast, computes a similarity score to each individual motif within a reference database (Schep et al. 2017). We leveraged RNA-Seq in order to identify likely members of transcription factor families that are differentially expressed after temperature shift. *Klf4* and *Tead4* were identified as likely active transcription factors as they exhibit correlated upregulation with their constituent binding sites. Other members of their respective transcription factor families were not expressed. ATAC-Seq alone would not have been able to identify transcriptional activity and integration of RNA-Seq data enables identification of high confidence TF regulatory networks.

5.6.9 Global alterations in *AP-1* binding sites

Global chromatin remodeling of *AP-1* binding sites has been observed after a reduction in culture temperature in CHO cells. *AP-1* is a conserved family of transcription factors that include *cFOS* which has been previously demonstrated to be one of the most downregulated genes in CHO cells following temperature shift (Bedoya-López et al. 2016). *AP-1* transcription factor are composed of dimers between the *FOS* subunits (*cFOS*, *FOSL1*, *FOSL2*) and *JUN* subunits (*JUN*, *JUNB*, *JUND*) as well as dimers between *MAF* and *CREB* families (Madrigal and Alasoo 2018). The core *AP-1* motif was represented in both differentially accessible promoters and enhancers indicating this transcription factor family has a global regulatory role. We observed a significant downregulation of *cFos* but a modest upregulation of the *Fos* like antigen 1 protein which suggests that temperature shift drives subunit specificity in CHO cells. Different arrangements of constituent *AP-1* subunits mediate cell-specific control of genome wide transcription (B.-K. Lee et al. 2018). For example, *Fosl1:JunB* complexes regulate migration and differentiation of trophoblasts (Renaud et al. 2014). Gene abundances for the corresponding *Jun* subunits were found to be high and were not found to be differentially expressed. Dimerization with *Jun* subunits is essential for *Fos* subunit DNA binding (Hess, Angel, and Schorpp-Kistner 2004). *AP-1* transcription factors

are key regulators of cell proliferation, apoptosis, and differentiation (Madrigal and Alasoo 2018) with an increase in *Fos/1* expression correlated with increased cell proliferation and cell survivability in lung cancer samples in a mouse model (Elangovan et al. 2018). Elangovan *et al.* also demonstrated that *Fos/1* increased cell survival and reduced apoptosis by upregulating the epidermal growth factor amphiregulin (*Areg*) (Elangovan et al. 2018) which was expressed at a low abundance in our RNA-Seq data despite containing a distal intergenic enhancer with a *Fos/1* binding site. *Fos/1* is an immediate early response gene that reacts to cellular stress triggers and is activated as part of the *Ras* signaling cascade (D. H. Khan and Davie 2013). This signaling cascade was found to be one of the most enriched gene ontology enriched pathways of differentially expressed genes that contain differentially accessible peaks. We propose that *Fos/1-Jun* binding sites become significantly open after a reduction in culture temperature in order to prolong cell survival due to cellular stress.

5.6.10 *TEAD4* has increased average accessibility after temperature shift

The *TEAD4* transcription factor was observed to have increased accessibility at 31°C and localized to enhancer regions within the genome. *TEAD4* was found to be enriched by the *HOMER2* software suite but not by the *chromVAR* tool. *chromVAR* relies on a position weight matrix from the JASPAR database of experimentally validated motifs (A. Khan et al. 2018) which does not contain *TEA* domains within the mouse-specific reference database. *TEA* domain factors have been found to be tissue-specific regulators of transcription and function as oncogenes that modulate proliferation (Jie Zhang et al. 2019). Downregulation of *TEAD4* was found to consistently result in reduced proliferation (Jie Zhang et al. 2019). *TEAD4* is also known to co-regulate gene expression with the *AP-1* transcription factor family in order to regulate cell motility (X. Liu et al. 2016). We identified a total of 72 differentially expressed genes that contained differentially accessible *TEAD4* and *AP-1* transcription factor binding sites including within an intron of the ENSEMBL annotated intergenic lncRNA *ENSCGRG00000024491* and the last exon of an evolutionarily conserved antisense transcript *Gm32980*. Cellular processes involved in

organelle and cell motility were highly represented within the gene ontology enrichment analysis which adds support to the *AP-1:TEAD4* transcriptional co-regulation.

5.6.11 Kruppel like factors are highly accessible after temperature shift

Transcription factor binding sites containing *Kruppel like factor* (KLF) motifs were the most significantly upregulated accessible sites. KLFs are a family of evolutionarily conserved, zinc finger containing transcription factors that have been demonstrated to control cellular metabolism and proliferation and share a common motif. Analysis with *chromVAR* identified *KLF1*, *KLF4*, and *KLF12* sites which all shares a core motif with peaks containing a *KLF* motif being 4-fold more accessible at 31°C. Comparison with RNA-Seq data found that *KLF1* and *KLF12* were not expressed while *KLF4* was significantly upregulated after temperature shift and abundant which suggests that this transcription factor has a role in regulating cellular pathways after TDS in CHO cells.

KLF4 has a well-characterized role in regulating growth arrest after a response to cellular stress and DNA damage. Overexpression of *Kruppel like factor 4* has been correlated with cell cycle arrest in gastric cancer cell lines *KLF4* mediated growth arrest has been linked to inhibited cell cycle progression via induced expression of the *Cyclin dependent kinase inhibitor 1A* (*CDKN1A*) which in turn inhibits cell cycle progression and stalling cells in the G0/G1 phase. Arrest at G0/G1 phase of the cells cycle is a common phenotype of CHO cells that have been exposed to mild hypothermic shock and our observations suggest that *KLF4* is a strong regulatory component of this pathway. *CDKN1A* expression is one of the most significantly upregulated genes after a reduction in culture temperature in RNA-Seq data which further supports the role of *KLF4* as an important regulator. We identified a temperature specific *KLF4* binding site that is found upstream of the *CDKN1A* promoter which further supports induced regulation by *KLF4*. *KLF4* has also been demonstrated to have an antagonistic effect on the *Lactate dehydrogenase A* gene (*LDHA*) in pancreatic cancer although we found *LDHA* as significantly upregulated which is necessary for

the shift from Warburg metabolism to OXPHOS metabolism that occurs after temperature shift. We did not identify a peak containing a *KLF4* site within the CHO LDHA gene suggesting that this gene is regulated by an independent mechanism.

We have also identified significantly differentially accessible *KLF4* sites within gene bodies and promoters of differentially expressed lncRNAs. The existence of *KLF4* transcription factor binding sites suggest the importance of the target lncRNAs in metabolic and growth arrest pathways like their coding counterparts. The presence of *KLF4* motifs can, therefore, be used to infer the regulatory function of a particular lncRNA as well as identify potential biological pathways that are regulatory targets. Similar to coding genes, differentially accessible sites within lncRNAs do not necessarily equate with differential lncRNA expression so further *in vitro* validation is required to elucidate specific molecular mechanisms of actions. *KLF4* has been demonstrated to be downregulated by the expression of the conserved lncRNA *TUG1* which was found to be quite abundant (average normalized abundance 5442.2) but not differentially expressed within RNA-Seq data indicating an alternative induction of *KLF4*. Evidence of *KLF4-lncRNA* regulatory networks suggests that Kruppel Like Factors have a role in lncRNA expression following cellular stress.

Motifs for *FOSL1:JUN*, *TEAD4*, and *KLF4* transcription factors were found to be significantly enriched in differentially accessible chromatin. Chromatin accessibility of a transcription factor binding site is correlated with active regulation by the TF but *in vitro* validation by binding assays are necessary in order to identify a determinative role. DNA foot printing assays using ATAC-Seq would be useful in identifying actively TF motifs (J. D. Buenrostro et al. 2015b). High coverage (>200 million fragments) can be aligned to the genome in order to identify localized decreases in coverage that correspond to protein bound DNA. Alternatively, CHIP-Seq libraries with TF bound DNA can be prepared to map actively bound motifs.

5.7 Conclusions

In this chapter, we performed the first ATAC-Seq experiment that profiled genome wide chromatin alterations in order to study a bioprocessing relevant phenotype: a reduction in culture temperature from 37°C to 31°C. We combined existing CHIP-Seq datasets and a previously sequenced RNA-Seq dataset presented in Chapter 2 in order to generate maps of open chromatin. Integration of RNA-Seq and ATAC-Seq was useful in locating two stable, highly accessible intergenic loci that have potential for site-directed targeted integration of transgenes. The ATAC-Seq and RNA-Seq data was also integrated to identify putative transcription factor regulatory networks.

Chapter 6

Future Work and

Overall Conclusions

6.1 An improved lncRNA reference annotation using RNA-Seq

RNA-Seq data from **Chapter 2** and **Chapter 3** were used as input for the lncRNA discovery pipeline in order to generate a merged annotation file that integrates the *Ensembl* version 98 reference assembly with high-coverage RNA-Seq data. The high-confidence, merged assembly had 7,312 lncRNA genes (10,312 lncRNA transcripts) which represents more than 2 fold increase in the number of annotated non-coding transcripts within the *criGri1* genome. 6,435 transcripts (64%) are novel and not previously found within the *ENSEMBL* version 98 reference annotation. The high-confidence transcripts were classified with respect to localization to the genome to nearby protein coding genes (**Figure 6.1**). The consensus assembly doubled the number of antisense overlapping and intergenic transcripts compared to the initial assembly presented in **Chapter 2** while there was no appreciable increase in sense overlapping transcripts and only a modest increase in divergent transcripts. These classes of lncRNAs represent only a small fraction of the total abundance in the human and mouse GENCODE reference annotation so the small number of sense and divergent transcripts may represent those loci that are expressed in CHO cells. Sense and antisense transcripts localized upstream and downstream to protein coding genes were expanded by approximately 33% compared to the *StringTie* assembly from **Chapter 2**. Version 23 of the mouse GENCODE reference annotation contains 18,800 long non-coding transcripts while version 32 of the human reference annotation contains 48,351 transcripts. The improved CHO annotation represents 53% of the mouse and 20% of the human annotation lncRNAs which is a significant improvement from the *Ensembl* version 98 reference. The work conducted in this Ph.D. thesis has resulted in the most comprehensive annotation of lncRNAs to date and will be available to the CHO community as a gene transfer format (GTF) file will be a very useful reference for future cell line engineering efforts by providing more complete coordinates for long non-coding genes.

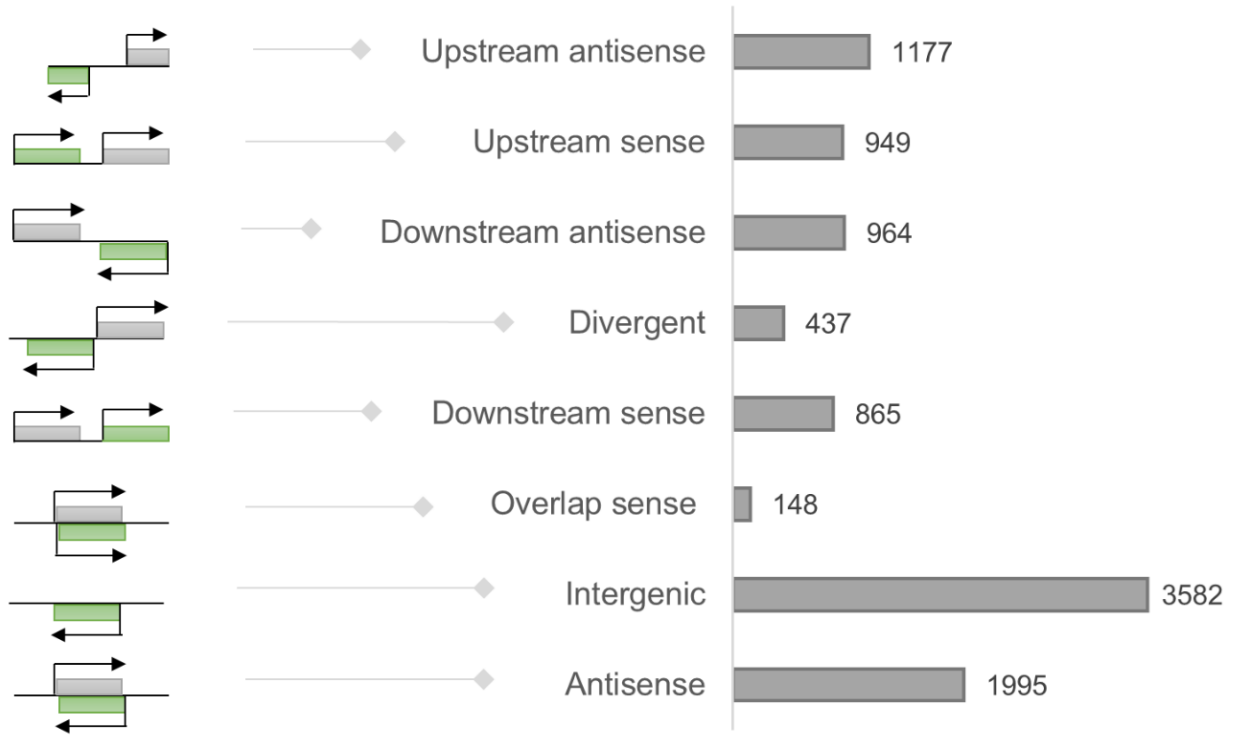


Figure 6.1 Classification of lncRNAs discovered in RNA-Seq data from **Chapter 2** and **Chapter 3**. Intergenic lncRNAs are the most abundant class of transcripts followed by antisense overlapping lncRNAs. There was only a modest increase in the number of sense overlapping transcripts while all other classes were significantly expanded.

6.2 Future work

6.2.1 *In vitro* validation of differentially expressed lncRNAs

We have presented global differential expression of long non-coding RNA genes after a reduction in culture temperature of over the 11-day span of fed-batch culture. Many of these lncRNAs are expressed at lower abundances than protein coding genes so it is necessary to perform further validation to evaluate whether these molecules are expressed at physiologically relevant levels. In **Chapter 2**, we presented a cohort of 400 genes that were significantly differentially expressed after temperature shift. Some of the most abundant, differentially expressed non-coding genes in RNA-Seq data from temperature-shifted cells such as the conserved *HOTAIRM1* and *LINC00511* lncRNAs. In the fed-batch culture (as presented in **Chapter 3**), *NEAT1* was highly upregulated during the death phase despite being previous research showing overexpression of this conserved gene leads to cellular proliferation in model organisms. The regulatory mechanism of these conserved genes is not thoroughly understood, however. Further *in vitro* and *in vivo* validation studies will use overexpression and gene knockouts to probe the function of these conserved genes. We hypothesize that overexpression of *HOTAIRM1* results in increased autophagy within CHO cells while knockout of *LINC00511* will result in cell cycle arrest.

Differential expression of sense coding genes and the antisense non-coding gene was also observed after temperature shift and during fed-batch culture. Differential expression on the antisense strand does not necessarily equate to *cis* regulation of the sense protein coding gene, however. Further validation is necessary to test whether antisense molecules are altering the expression of the sense coding partner. Among this cohort, we identified a set of antisense lncRNAs that have anti-correlated expression with the sense protein coding gene where the antisense gene is upregulated and the sense gene is downregulated. We initially hypothesized that these genes function as transcriptional repressors but this needs to be experimentally verified. A validation experiment that overexpresses antisense lncRNA, such as for the *CBX3-AS*

or *OXCT1-AS* genes, will provide evidence on whether these non-coding genes repress sense expression. Future work beyond the initial validation experiment would focus on identifying functional domains within the exons of this cohort that are responsible for the regulatory function which can, in turn, be adapted to future cell line engineering efforts.

6.2.2 Validation of lncRNAs that are correlated with metabolite concentration

In **Chapter 3**, normalized gene abundances were correlated to concentrations of metabolites lactate, glutamine, and ammonia as well as process parameters like viable cell concentration and cell viability. It is not currently known whether the relationship between metabolite concentration and non-coding gene expression is deterministic where alterations in concentration directly cause differential gene expression. Future experiments would involve adding incremental doses of a metabolite like lactate to culture media and evaluate the change in target gene expression. Genes that are strongly negatively correlated with % viability could be involved in governing apoptosis and an experiment that uses RNA interference or *CRISPR*-mediated gene deletion of these genes would be necessary to evaluate potential regulatory function.

6.2.3 Characterization of the impact of alternative splicing of lncRNAs on CHO cells

In **Chapter 4**, we presented 398 lncRNAs that underwent alternative splicing and isoform switching over the duration of a fed-batch culture. An exon skipping event was observed for the novel *MSTRG.16574* gene where that isoform became the predominant transcript during the death phase while the previously predicted ENSCGRG00000025089 gene has an alternative 3' splice site which becomes dominant during the stationary phase. Overexpression of exponential stage-specific and death phase-specific transcripts could be used to investigate the functional role of these isoforms. We also reported a cohort of 18 lncRNAs that underwent isoform switching over all three stages such as the *MSTRG.13357* gene which is composed of two isoforms. A longer isoform, *MSTRG.13357.1*, is dominant during the exponential and death phases while the second, shorter isoform, *MSRTRG.13357.2*, is dominant only during the stationary phase. Investigation of the impact of this alternative 5' splice event would require overexpression of the

shorter isoform to evaluate any effect on cell physiology. Differentially spliced lncRNA isoforms were shown to have differences in secondary structure. These altered secondary structures may represent functional domains that are necessary for different regulatory function so future validation experiments could probe the regulatory function of individual isoforms.

6.2.4 Integration of miRNA-Seq data to identify ceRNA networks

One of the roles of lncRNAs in the cytoplasm is to function as ceRNA sponges for miRNAs (Du et al. 2016). These ceRNAs sequester miRNAs to prevent binding of the small ncRNA to target 3'UTR. Several putative ceRNAs, such as *LINC002553*, were discovered by cross-species comparison of CHO lncRNAs to *GENCODE* annotated human and mouse transcripts. High depth miRNAs were profiled over the course of a fed-batch culture alongside total RNA. Integrating these two datasets will enable identification of lncRNAs that function as competing endogenous RNAs. Correlation analysis will be applied to discover long non-coding RNAs that are anti-correlated with miRNA expression and use this information to infer mRNA-lncRNA-miRNA regulatory networks in CHO cells. Identifying competing endogenous RNAs will assist in identifying miRNA targets that downregulate mRNAs associated with bioprocessing relevant traits. Synthetic ceRNAs, referred to as miRNA sponges, are a common tool in molecular biology and have been previously demonstrated to alter CHO cell physiology after transient transfection. Binding sites predicted within ceRNAs can be used to refine sponge design to for future cell line engineering. Sequencing data is currently being analyzed in collaboration with Ricardo Valdés-Bango Currell at the National Institute for Cellular Biotechnology in Dublin, Ireland.

6.2.5 Integration of bisulfite sequencing data to identify epigenetic remodeling complexes

High-throughput bisulfite sequencing was performed to profile alterations to genome-wide methylation over the course of the fed-batch culture. This is a high-throughput methylome sequencing method that first treats extracted DNA with bisulfite that converts unmethylated cytosine nucleic acids into uracil while not affecting methylated cytosine residues. Alterations in

methylation at promoters, enhancers, and repressors modulates gene expression (Deaton and Bird 2011). Many lncRNAs, like *XIST* and *PVT1*, regulate gene transcription by guiding methylases to regulatory regions which are then modified the addition or removal of methyl groups (Leung and Panning 2014). Differentially expressed lncRNAs between culture phases can be correlated with changes in methylation states of nearby protein coding genes to identify *cis* epigenetic regulation. The chromatin state map developed in **Chapter 5** will also be extremely useful in inferring functional consequences to alterations in methylation at a regulatory site such as a promoter, enhancer, or repressor. Analysis of bisulfite sequencing data is currently ongoing and is being performed by Dr. Heena Dhiman Pal at the Austrian Center for Industrial Biotechnology in Vienna, Austria.

6.2.6 Validation of differential accessibility of transcription factor binding site motifs

Transcription factors function as global regulators of gene expression by binding to specific motifs.

Cell line engineering strategies that target these binding sites have been proposed as a method to modulate transcription as short motif sequences can be more readily deleted or inserted using targeted engineering strategies compared to whole genes (Gutiérrez-González et al. 2019). Some transcription factors can function both as transcriptional repressors and enhancers so further *in vitro* work is needed to evaluate TF binding affinity to nearby gene regulation (Hess, Angel, and Schorpp-Kistner 2004). Transcription factor binding sites that contain *KLF4* motifs were some of the most differentially accessible after a reduction in culture temperature which suggests that these sites are ideal for future, more complex *in vitro* and *in vivo* studies. Targeted deletion of *KLF4* sites in enhancer elements upstream of genes, such as *ANXA8*, would evaluate the relationship between motif and gene expression.

6.2.7 Validating landing pads using open chromatin maps

In addition to identifying transcription factor binding sites and potential regulatory regions, we used ATAC-Seq to discover two “landing pads” which are stable genomic regions are highly accessible that do not change in conditions. Identification of stable genomic sites are of significant

interest in industrial CHO cell bioprocessing as these sites reduce the reliance on random integration and laborious screening during initial cell line development. A transgene can be reliably integrated into a landing pad loci using *CRISPR* to induce double stranded breaks with single guide RNAs (sgRNAs). Altering the nucleotide composition of accessible regions could have unknown functional consequences for the cell as this could disrupt regulatory regions such as enhancers and repressors that function as transcription factor binding sites. *In vivo* validation would involve performing site directed transgene integration into these two sites to evaluate whether cell growth or recombinant protein titer are affected compared to random integration.

6.2.8 Resources for future cell line engineering research

The differential expression and chromatin accessibility data presented in this thesis will be integrated into a publically available database that will be continuously updated with results from future experiments. Gene expression and alternative splicing data that was correlated to metabolite concentration and culture conditions will provide a resource of targets that researchers in the community can search. Future ATAC-Seq experiments will also be integrated into this repository in order to generate more complete chromatin state maps of CHO cells in industrially relevant conditions.

6.3 Overall Conclusions

6.3.1 lncRNA detection pipeline

We constructed a software pipeline that can detect conserved lncRNAs as well as identify novel loci within raw RNA-Seq data. This software pipeline requires strand-specific transcriptome data in order to identify sense and antisense genic loci. The pipeline first performs raw data QC to trim low-quality reads and adaptors. Quality filtered reads are then aligned to the CHOK1 reference genome and mapped reads are assembled into transcript loci. These loci are then evaluated for coding potential and the presence of protein coding domains or similarity to small non-coding RNAs. This pipeline uses highly stringent parameters in order to retain high-quality long non-coding transcripts which are then classified into genic and intergenic groups using localization to nearby protein coding genes. All code for this analysis pipeline is available as a *GITLAB* repository and can be easily adapted for use in future experiments and will be a very useful resource to the CHO community in future RNA-Seq analysis of lncRNAs.

6.3.2 Characteristics of assembled lncRNAs

lncRNAs were found to be expressed at lower average abundances than protein coding genes in both assemblies presented in **Chapter 2** and **Chapter 3**. lncRNAs are also significantly shorter than the mRNA transcripts with an average length of ~750bp. lncRNA loci had fewer isoforms per gene and the non-coding transcripts had fewer exons compared to mRNAs. Single isoform genes are the majority of lncRNA and mRNA loci, however. These results are consistent with those results presented by *GENCODE* which suggests that our annotation workflow is rigorous and can reproduce results from more complex annotations (Derrien et al. 2012). These results further support the accuracy and consistency of our pipeline and our annotation is comparable to reference annotations of model organisms.

6.3.3 lncRNAs are differentially expressed in industrially relevant conditions

We observed global differential expression of long non-coding RNAs in CHO cells in industrially relevant conditions. 400 non-coding genes were differentially expressed after a reduction in temperature from 37°C to 31°C while almost 3,000 genes were differentially expressed over the duration of a fed-batch culture. Intergenic genes were the most abundant class of long non-coding RNAs and composed the largest class within differentially expressed non-coding genes. Differential expression of genes found antisense to protein coding genes was prevalent in both datasets. Several evolutionarily conserved loci, including *HOTAIRM1* and *PVT1*, were strongly differentially expressed after temperature shift but not differentially expressed during fed-batch culture while the nuclear enriched *NEAT1* gene was one of the most abundant, differentially expressed non-coding genes during the death phase of the fed-batch culture.

We presented the first evidence of differential expression of both sense protein coding genes with the antisense non-coding gene after a temperature shift as well as over the duration of a fed-batch culture. A total of 47 sense/antisense pairs were discovered in both datasets with the *GINS4* locus being differentially expressed in both assemblies. Differential expression could be correlated, with the sense and antisense gene similarly up or downregulated, or anti-correlated, where, for example, the sense coding gene is downregulated while the partner non-coding gene is upregulated. Downregulation of the sense protein coding gene, such as with the *Oxct1* mRNA with the *Oxct1*-AS non-coding partner, was more common when the antisense non-coding gene was upregulated suggesting that many of the antisense lncRNAs function as transcriptional repressors. These alterations in sense/antisense expression could occur independently and the lncRNA could be regulating a distal target instead (Goyal et al. 2017)

6.3.4 Alterations in lncRNA abundance correlated with changes in metabolite concentration and bioprocess parameters

We correlated normalized gene abundances with lactate, glutamate, and ammonia concentrations during a fed-batch culture in **Chapter 3**. 30 long non-coding RNA genes strongly

correlated with these three metabolite concentrations. There were no genes that significantly correlated with glucose or glutamate concentrations, however. Pearson's correlation coefficients were also computed for bioprocessing relevant parameters including cell viability and viable cell concentrations. lncRNAs that strongly correlate with a decrease in cell viability could be involved in regulating apoptosis while those genes that are positively correlated with viable cell concentrations could be involved in proliferation mechanisms.

6.3.5 Alternative splicing of lncRNAs in fed-batch culture

Chapter 4 presented the first genome-wide analysis of lncRNA alternative splicing in CHO cells with 398 lncRNA transcripts undergoing significant switching events between exponential, stationary, and death phases. 18 of these genes had changes in isoform fraction throughout the duration of the culture where different component transcripts become dominant during different culture phases. The novel *MSTRG.13357* lncRNA gene contains an isoform that was dominant only during the stationary phase which had an Alternative 5' splice site in the second exon. The expression of different transcripts during different culture stages may be useful in inferring likely regulatory roles of lncRNA. The predicted secondary structure was observed to be significantly different between isoforms.

6.3.6 Global change in chromatin accessibility in protein coding and long non-coding genes after temperature shift

We used the ATAC-Seq method to evaluate differences in chromatin accessibility after a reduction in culture temperature. The presence of ATAC-Seq signal near the transcription start site or within promoters is correlated to the active transcription of a gene (Ackermann et al. 2016). This analysis was performed to further refine the annotation of lncRNAs by identifying a cohort of high-confidence loci that have an accessible region near the transcription start site. 51% of all lncRNA genes had an accessible peak within 1kb of the transcription start site compared to 61% of all mRNA genes. The intensity of signal surrounding the TSS of lncRNAs is lower than the intensity of signal surrounding mRNA genes which agrees with previous findings from **Chapter 2**

and **Chapter 3** that show that lncRNAs are expressed at lower abundances than coding genes. While the average accessibility is lower compared between these two classes of genes, alterations in nucleosome positioning and accessibility were found to be similar between mRNA genes and long non-coding genes which supports previous findings (Gendron et al. 2019).

6.3.7 Identification of active regulatory sites using multi-omics data

The majority (42%) of ATAC-Seq peaks occurred in intergenic regions outside of gene bodies or promoters with the remainder found within introns. Accessible sites are often found to overlap regulatory regions such as promoters, enhancers, and repressors (Halfon 2019b). The ATAC-Seq data was integrated with previously published CHIP-Seq data (Feichtinger et al. 2016) to find accessible promoters, enhancers, and repressors. The majority of high-confidence promoter CHIP-Seq peaks overlapped with ATAC-Seq signal while comparatively fewer enhancer and repressor peaks overlapped with ATAC-Seq data. These active regulatory sites contain binding sites for transcription factors that in turn modulate gene expression. The high-coverage RNA-Seq data presented in **Chapter 2** was used to compare differential expression of transcription factor genes with changes in chromatin accessibility which is the first combined analysis of TF expression and accessibility in CHO cells. Global rearrangement of *FOSL1* sites was observed after temperature shift with 40% of all differentially accessible peaks containing motifs that are unique to this TF complex. *FOSL1* was found to be moderately upregulated in the RNA-Seq data which suggests gene expression is regulated by transcription factor binding site accessibility. In addition to *FOSL1*, the *TEAD4* and *KLF4* transcription factors were significantly upregulated after a reduction in culture temperature with motifs containing *KLF* domains becoming significantly more accessible after temperature shift. *KLF4* has previously been shown to play a role in cell cycle arrest which is a common phenotype in cells exposed to mild hypothermia (Kumar et al. 2008). We propose that increased expression of transcription factor *KLF4*, combined with increased availability of binding sites, is involved in cell cycle arrest in the G₀/ G₁ phase of the cell cycle.

6.3.8 Integration of ATAC-Seq with RNA-Seq to discover highly accessible landing pads

Chromatin state maps over large genomic (>1kb) intervals were generated using ATAC-Seq in order to find sites that were continuously accessible. The chromatin state maps are a useful utility for discovering highly accessible, stable genomic regions that are likely sites for targeted transgene integration. We selected two large intervals that are located within intergenic regions flanking highly expressed genes: a large 21 kb interval between the *DAG1/NICN1* locus and a 6.6kb region immediately upstream of the *SACS* locus. These genes are among the top 5% most expressed loci and were not found to be significantly differentially expressed using RNA-Seq data. Transgenes will more readily integrate in regions of high accessibility due to reduced protection by the nucleosomes where there are highly transcribed genes (O'Brien et al. 2018).

Bibliography

- Ackermann, Amanda M., Zhiping Wang, Jonathan Schug, Ali Naji, and Klaus H. Kaestner. 2016. "Integration of ATAC-Seq and RNA-Seq Identifies Human Alpha Cell and Beta Cell Signature Genes." *Molecular Metabolism* 5 (3). Elsevier GmbH: 233–44. doi:10.1016/j.molmet.2016.01.002.
- Aken, Bronwen L., Premanand Achuthan, Wasiru Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, et al. 2017. "Ensembl 2017." *Nucleic Acids Research* 45 (D1): D635–42. doi:10.1093/nar/gkw1104.
- Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, et al. 2016. "The Ensembl Gene Annotation System." *Database : The Journal of Biological Databases and Curation* 2016: 1–19. doi:10.1093/database/baw093.
- Altamirano, Claudia, Andres Illanes, Silvana Becerra, Jordi Joan Cairó, and Francesc Gòdia. 2006. "Considerations on the Lactate Consumption by CHO Cells in the Presence of Galactose." *Journal of Biotechnology* 125 (4): 547–56. doi:10.1016/j.jbiotec.2006.03.023.
- Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. doi:10.1093/nar/25.17.3389.
- Apweiler, R. 2004. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 32 (90001): 115D–119. doi:10.1093/nar/gkh131.
- Arstikaitis, Pamela, Catherine Gauthier-Campbell, Rosario Carolina Gutierrez Herrera, Kun Huang, Joshua N Levinson, Timothy H Murphy, Manfred W Kilimann, Carlo Sala, Michael A Colicos, and Alaa El-Husseini. 2008. "Paralemm-1, a Modulator of Filopodia Induction Is Required for Spine Maturation." *Molecular Biology of the Cell* 19 (5): 2026–38. doi:10.1091/mbc.e07-08-0802.
- Atkinson, Sophie R., Samuel Marguerat, and Jürg Bähler. 2012. "Exploring Long Non-Coding RNAs through Sequencing." *Seminars in Cell and Developmental Biology* 23 (2). Elsevier Ltd: 200–205. doi:10.1016/j.semcdb.2011.12.003.
- Azali, A, Sattam M Obeidat, Muhammad Amir Yunus, and A Ghows. 2019. "Systematic Identification and Characterization of Aedes Aegypti Long Noncoding RNAs (lncRNAs)," no. January: 1–9. doi:10.1038/s41598-019-47506-9.
- Baeshen, Nabih A, Mohammed N Baeshen, Abdullah Sheikh, Roop S Bora, Mohamed Morsi M Ahmed, Hassan A I Ramadan, Kulvinder Singh Saini, and Elrashdy M Redwan. 2014. "Cell Factories for Insulin Production," 1–9.
- Baker, Monya. 2011. "Long Noncoding RNAs: The Search for Function." *Nature Methods* 8 (5). Nature Publishing Group: 379–83. doi:10.1038/nmeth0511-379.
- Balbin, O Alejandro, Rohit Malik, Saravana M Dhanasekaran, John R Prensner, Xuhong Cao, Yi-mi Wu, Dan Robinson, et al. 2015. "The Landscape of Antisense Gene Expression in Human Cancers." *Genome Research* 25 (7): 1068–79. doi:10.1101/gr.180596.114.
- Bartel, David P. 2004. "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." *Cell* 116 (2): 281–97. doi:10.1016/S0092-8674(04)00045-5.

- Becker, Jennifer, Christina Timmermann, Tobias Jakobi, Oliver Rupp, Rafael Szczepanowski, Matthias Hackl, Alexander Goesmann, et al. 2011. "Next-Generation Sequencing of the CHO Cell Transcriptome" 5 (Suppl 8): 5–6. doi:10.1186/1753-6561-5-S8-P6.
- Bedoya-López, Andrea, Karel Estrada, Alejandro Sanchez-Flores, Octavio T. Ramírez, Claudia Altamirano, Lorenzo Segovia, Juan Miranda-Ríos, Mauricio A. Trujillo-Roldán, and Norma A. Valdez-Cruz. 2016. "Effect of Temperature Downshift on the Transcriptomic Responses of Chinese Hamster Ovary Cells Using Recombinant Human Tissue Plasminogen Activator Production Culture." *PLoS ONE* 11 (3): 1–26. doi:10.1371/journal.pone.0151529.
- Benner, Christopher. 2018a. "HOMER2." <http://homer.ucsd.edu/homer/index.html>.
- Bentley, David R, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, et al. 2009. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59. doi:10.1038/nature07517.Accurate.
- Birzele, Fabian, Jochen Schaub, Werner Rust, Christoph Clemens, Patrick Baum, Hitto Kaufmann, Andreas Weith, Torsten W. Schulz, and Tobias Hildebrandt. 2010. "Into the Unknown: Expression Profiling without Genome Sequence Information in CHO by next Generation Sequencing." *Nucleic Acids Research* 38 (12): 3999–4010. doi:10.1093/nar/gkq116.
- Böhmdorfer, Gudrun, and Andrzej T. Wierzbicki. 2015. "Control of Chromatin Structure by Long Noncoding RNA." *Trends in Cell Biology* 25 (10). Elsevier Ltd: 623–32. doi:10.1016/j.tcb.2015.07.002.
- Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. doi:10.1093/bioinformatics/btu170.
- Bond, Charles S., and Archa H. Fox. 2009. "Paraspeckles: Nuclear Bodies Built on Long Noncoding RNA." *Journal of Cell Biology* 186 (5): 637–44. doi:10.1083/jcb.200906113.
- Borodina, Tatiana, James Adjaye, and Marc Sultan. 2011. *A Strand-Specific Library Preparation Protocol for RNA Sequencing. Methods in Enzymology*. 1sted. Vol. 500. Elsevier Inc. doi:10.1016/B978-0-12-385118-5.00005-0.
- Borys, Michael C., Nimish G. Dalal, Nicholas R. Abu-Absi, Sarwat F. Khattak, Ying Jing, Zizhuo Xing, and Zheng Jian Li. 2010. "Effects of Culture Conditions on N-Glycolylneuraminic Acid (Neu5Gc) Content of a Recombinant Fusion Protein Produced in CHO Cells." *Biotechnology and Bioengineering* 105 (6): 1048–57. doi:10.1002/bit.22644.
- Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. 2008a. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." *Cell* 132 (2): 311–22. doi:10.1016/j.cell.2007.12.014.
- Brinkrolf, Karina, Oliver Rupp, Holger Laux, Florian Kollin, Wolfgang Ernst, Burkhard Linke, Rudolf Kofler, et al. 2013. "Chinese Hamster Genome Sequenced from Sorted Chromosomes." *Nature Biotechnology* 31 (8): 694–95. doi:10.1038/nbt.2645.
- Brookheart, Rita T., Carlos I. Michel, Laura L. Listenberger, Daniel S. Ory, and Jean E. Schaffer. 2009. "The Non-Coding RNA gadd7 Is a Regulator of Lipid-Induced Oxidative and Endoplasmic Reticulum Stress." *Journal of Biological Chemistry* 284 (12): 7446–54.

doi:10.1074/jbc.M806209200.

- Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18. doi:10.1038/nmeth.2688.
- Buenrostro, Jason D, Beijing Wu, Howard Y Chang, and William J Greenleaf. 2015a. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Current Protocols in Molecular Biology* 109 (January): 21.29.1-9. doi:10.1002/0471142727.mb2129s109.
- Buenrostro, Jason, Beijing Wu, Ulrike M. Litzénburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015a. "Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation." *Nature* 523 (7561): 486–90. doi:10.1038/nature14590.
- Bumgarner, Roger. 2013. "DNA Microarrays: Types, Applications and Their Future." *Curr Protoc Mol Biol.* 6137 (206): 1–17. doi:10.1002/0471142727.mb2201s101.DNA.
- Butler, M., and A. Meneses-Acosta. 2012. "Recent Advances in Technology Supporting Biopharmaceutical Production from Mammalian Cells." *Applied Microbiology and Biotechnology* 96 (4): 885–94. doi:10.1007/s00253-012-4451-z.
- Cader, M Zaeem, Katharina Boroviak, Qifeng Zhang, Ghazaleh Assadi, Sarah L Kempster, Gavin W Sewell, Svetlana Saveljeva, et al. 2016. "C13orf31 (FAMIN) Is a Central Regulator of Immunometabolic Function." *Nature Immunology* 17 (9): 1046–56. doi:10.1038/ni.3532.
- Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. doi:10.1186/1471-2105-10-421.
- Cao, Jian, and Qin Yan. 2012. "Histone Ubiquitination and Deubiquitination in Transcription, DNA Damage Response, and Cancer." *Frontiers in Oncology* 2: 26. doi:10.3389/fonc.2012.00026.
- Carrieri, Claudia, Laura Cimatti, Marta Biagioli, Anne Beugnet, Silvia Zucchelli, Stefania Fedele, Elisa Pesce, et al. 2012. "Long Non-Coding Antisense RNA Controls Uchl1 Translation through an Embedded SINEB2 Repeat." *Nature* 491 (7424). Nature Publishing Group: 454–57. doi:10.1038/nature11508.
- Chang, Pearl, Moloya Gohain, Ming Ren Yen, and Pao Yang Chen. 2018. "Computational Methods for Assessing Chromatin Hierarchy." *Computational and Structural Biotechnology Journal* 16. The Authors: 43–53. doi:10.1016/j.csbj.2018.02.003.
- Chari, Raj, Mark Moosburner, and George M Church. 2015a. "Unraveling CRISPR-Cas9 Genome Engineering Parameters via a Library- on-Library Approach." *Nature Methods*, no. July. Nature Publishing Group: 1–7. doi:10.1038/nmeth.3473.
- Chen, Ailing, Daozhen Chen, and Ying Chen. 2018. "Advances of DNase-Seq for Mapping Active Gene Regulatory Elements across the Genome in Animals." *Gene* 667 (December 2017). Elsevier: 83–94. doi:10.1016/j.gene.2018.05.033.
- Chen, Jenny, Alexander A. Shishkin, Xiaopeng Zhu, Sabah Kadri, Itay Maza, Mitchell Guttman, Jacob H. Hanna, Aviv Regev, and Manuel Garber. 2016. "Evolutionary Analysis across Mammals Reveals Distinct Classes of Long Non-Coding RNAs." *Genome Biology* 17 (1).

- Genome Biology: 19. doi:10.1186/s13059-016-0880-9.
- Chen, Xiaoyu, and Martin Tompa. 2010. "Comparative Assessment of Methods for Aligning Multiple Genome Sequences." *Nature Biotechnology* 28 (6). Nature Publishing Group: 567–72. doi:10.1038/nbt.1637.
- Chen, Zhen-hua, Wen-tao Wang, Wei Huang, Ke Fang, Yu-meng Sun, Shu-rong Liu, Xue-qun Luo, and Yue-qin Chen. 2016. "The lncRNA HOTAIRM1 Regulates the Degradation of PML-RARA Oncoprotein and Myeloid Cell Differentiation by Enhancing the Autophagy Pathway" 24 (2). Nature Publishing Group: 212–24. doi:10.1038/cdd.2016.111.
- Cheng, Tsung-Lin, Ching-Chun Liao, Wen-Hui Tsai, Chin-Chih Lin, Chin-Wei Yeh, Chiao-Fang Teng, and Wen-Tsan Chang. 2009. "Identification and Characterization of the Mitochondrial Targeting Sequence and Mechanism in Human Citrate Synthase." *Journal of Cellular Biochemistry* 107 (5): 1002–15. doi:10.1002/jcb.22200.
- Ciarlo, Eleonora, Sara Massone, Ilaria Penna, Mario Nizzari, Arianna Gigoni, Giorgio Dieci, Claudio Russo, and Tullio Florio. 2013. "An Intronic ncRNA-Dependent Regulation of SORL1 Expression Affecting A β Formation Is Upregulated in Post-Mortem Alzheimer ' S Disease Brain Samples," no. May 2014. doi:10.1242/dmm.009761.
- Colombo, Pierre-Emmanuel, Fernanda Milanezi, Britta Weigelt, and Jorge S Reis-Filho. 2011. "Microarrays in the 2010s: The Contribution of Microarray-Based Gene Expression Profiling to Breast Cancer Classification, Prognostication and Prediction." *Breast Cancer Research* 13 (3): 212. doi:10.1186/bcr2890.
- Corces, M. Ryan, Alexandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, et al. 2017. "An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues." *Nature Methods* 14 (10): 959–62. doi:10.1038/nmeth.4396.
- Cost, Gregory J, Yevgeniy Freyvert, Annamaria Vafiadis, Yolanda Santiago, Jeffrey C Miller, Edward Rebar, Trevor N Collingwood, et al. 2010. "A RTICLE BAK and BAX Deletion Using Zinc-Finger Nucleases Yields Apoptosis-Resistant CHO Cells" 105 (2): 330–40. doi:10.1002/bit.22541.
- Cui, Ming, Lei You, Xiaoxia Ren, Wenjing Zhao, Quan Liao, and Yupei Zhao. 2016. "Long Non-Coding RNA PVT1 and Cancer." *Biochemical and Biophysical Research Communications* 471 (1). Elsevier Ltd: 10–14. doi:10.1016/j.bbrc.2015.12.101.
- Cui, Xuefeng, Zhiwu Lu, Sheng Wang, Jim Jing-Yan Wang, and Xin Gao. 2016. "CMsearch: Simultaneous Exploration of Protein Sequence Space and Structure Space Improves Not Only Protein Homology Detection but Also Protein Structure Prediction." *Bioinformatics* 32 (12): i332–40. doi:10.1093/bioinformatics/btw271.
- Dean, Jason, Pranhitha Reddy, Amgen Cell Sciences, and Amgen Court West. 2013. "Metabolic Analysis of Antibody Producing CHO Cells in Fed-Batch Production" 110 (6): 1735–47. doi:10.1002/bit.24826.
- Deaton, Am, and Adrian Bird. 2011. "CpG Islands and the Regulation of Transcription." *Genes & Development* 25 (10): 1010–22. doi:10.1101/gad.2037511.1010.
- Delli Ponti, Riccardo, Alexandros Armaos, Stefanie Marti, and Gian Gaetano Tartaglia. 2018. "A Method for RNA Structure Prediction Shows Evidence for Structure in lncRNAs." *Frontiers in Molecular Biosciences* 5: 111. doi:10.3389/fmolb.2018.00111.

- Delouvroy, Frédéric, Gaetan Siriez, An-Vy Tran, Larissa Mukankurayija, Nadine Kochanowski, and Laetitia Malphettes. 2015. "AmbrTM Mini-Bioreactor as a High-Throughput Tool for Culture Process Development to Accelerate Transfer to Stainless Steel Manufacturing Scale: Comparability Study from Process Performance to Product Quality Attributes." *BMC Proceedings* 9 (Suppl 9). BioMed Central Ltd: P78. doi:10.1186/1753-6561-9-S9-P78.
- Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. "The GENCODE v7 Catalog of Human Long Noncoding RNAs : Analysis of Their Gene Structure , Evolution , and Expression," 1775–89. doi:10.1101/gr.132159.111.
- Dijk, Erwin L Van, Yan Jaszczyszyn, and Claude Thermes. 2014. "Ten Years of next-Generation Sequencing Technology" 30 (9). doi:10.1016/j.tig.2014.07.001.
- Dimond, Andrew, and Peter Fraser. 2013. "Long Noncoding RNAs Xist in Three Dimensions." *Science* 341 (6147): 720–21. doi:10.1126/science.1243257.
- Ding, Nan, Haiying Wu, Tao Tao, and Erxuan Peng. 2017. "NEAT1 Regulates Cell Proliferation and Apoptosis of Ovarian Cancer by miR-34a-5p/BCL2." *OncoTargets and Therapy* 10: 4905–15. doi:10.2147/OTT.S142446.
- Dinger, Marcel E., Ken C. Pang, Tim R. Mercer, and John S. Mattick. 2008. "Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities." *PLoS Computational Biology* 4 (11). doi:10.1371/journal.pcbi.1000176.
- Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. doi:10.1093/bioinformatics/bts635.
- Dresios, J., A. Aschrafi, G. C. Owens, P. W. Vanderklish, G. M. Edelman, and V. P. Mauro. 2005. "Cold Stress-Induced Protein Rbm3 Binds 60S Ribosomal Subunits, Alters microRNA Levels, and Enhances Global Protein Synthesis." *Proceedings of the National Academy of Sciences* 102 (6): 1865–70. doi:10.1073/pnas.0409764102.
- Du, Zhou, Tong Sun, Ezgi Haciosuleyman, Teng Fei, Xiaodong Wang, Myles Brown, John L. Rinn, et al. 2016. "Integrative Analyses Reveal a Long Noncoding RNA-Mediated Sponge Regulatory Network in Prostate Cancer." *Nature Communications* 7 (March). Nature Publishing Group: 10982. doi:10.1038/ncomms10982.
- Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10). doi:10.1371/journal.pcbi.1002195.
- Elangovan, Indira M, Michelle Vaz, Chandramohan R Tamatam, Haranatha R Potteti, Narsa M Reddy, and Sekhar P Reddy. 2018. "FOSL1 Promotes Kras-Induced Lung Cancer through Amphiregulin and Cell Survival Gene Regulation." *American Journal of Respiratory Cell and Molecular Biology* 58 (5): 625–35. doi:10.1165/rcmb.2017-0164OC.
- Encode, The, and Project Consortium. 2007. "Identification and Analysis of Functional Elements in 1 % of the Human Genome by the ENCODE Pilot Project" 447 (June). doi:10.1038/nature05874.
- ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74. doi:10.1038/nature11247.
- Farrell, Amy, Niaobh McLoughlin, John J. Milne, Ian W. Marison, and Jonathan Bones. 2014. "Application of Multi-Omics Techniques for Bioprocess Design and Optimization in Chinese

- Hamster Ovary Cells." *Journal of Proteome Research* 13 (7): 3144–59. doi:10.1021/pr500219b.
- Fatima, Roshan, Vijay Suresh Akhade, Debosree Pal, and Satyanarayana Mr Rao. 2015. "Long Noncoding RNAs in Development and Cancer: Potential Biomarkers and Therapeutic Targets." *Molecular and Cellular Therapies* 3. Molecular and Cellular Therapies: 5. doi:10.1186/s40591-015-0042-6.
- Feichtinger, Julia, Inmaculada Hernández, Christoph Fischer, Michael Hanscho, Norbert Auer, Matthias Hackl, Vaibhav Jadhav, et al. 2016. "Comprehensive Genome and Epigenome Characterization of CHO Cells in Response to Evolutionary Pressures and over Time." *Biotechnology and Bioengineering* 113 (10): 2241–53. doi:10.1002/bit.25990.
- Feillet, Celine, Gijsbertus T.J. van der Horst, Francis Levi, David A. Rand, and Franck Delaunay. 2015. "Coupling between the Circadian Clock and Cell Cycle Oscillators: Implication for Healthy Cells and Malignant Growth." *Frontiers in Neurology* 6 (MAY): 1–7. doi:10.3389/fneur.2015.00096.
- Feng, Jianxing, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. 2012. "Identifying ChIP-Seq Enrichment Using MACS." *Nature Protocols* 7 (9): 1728–40. doi:10.1038/nprot.2012.101.
- Ferrè, Fabrizio, Alessio Colantoni, and Manuela Helmer-Citterich. 2015. "Revealing Protein – lncRNA Interaction," no. April: 1–11. doi:10.1093/bib/bbv031.
- Fickett, James W. 1982. "Recognition of Protein Coding Regions in DNA Sequences." *Nucleic Acids Research* 10 (17): 5303–18. doi:10.1093/nar/10.17.5303.
- Finn, Robert D., Penelope Coghill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, et al. 2016. "The Pfam Protein Families Database: Towards a More Sustainable Future." *Nucleic Acids Research* 44 (D1): D279–85. doi:10.1093/nar/gkv1344.
- Fischer, Simon, René Handrick, and Kerstin Otte. 2015. "The Art of CHO Cell Engineering: A Comprehensive Retrospect and Future Perspectives." *Biotechnology Advances* 33 (8). Elsevier Inc.: 1878–96. doi:10.1016/j.biotechadv.2015.10.015.
- Flicek, Paul, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalhosa, Peter Clapham, et al. 2012. "Ensembl 2012" 40 (November 2011): 84–90. doi:10.1093/nar/gkr991.
- Frankish, Adam, Barbara Uszczyńska, Graham RS Ritchie, Jose M Gonzalez, Dmitri Pervouchine, Robert Petryszak, Jonathan M Mudge, et al. 2015. "Comparison of GENCODE and RefSeq Gene Annotation and the Impact of Reference Geneset on Variant Effect Prediction." *BMC Genomics* 16 (Suppl 8): S2. doi:10.1186/1471-2164-16-S8-S2.
- Frazee, Alyssa C, Geo Pertea, Andrew E Jaffe, Ben Langmead, Steven L Salzberg, and Jeffrey T Leek. 2015. "Ballgown Bridges the Gap between Transcriptome Assembly and Expression Analysis." *Nature Biotechnology* 33 (3): 243–46. doi:10.1038/nbt.3172.
- Freeman, Jamie, Brian Burke, Jane Elliott, Chris Thorne, Fan Chen, Hua Tu, Fai Poon, et al. 2010. "Building on GS : Engineering the next Generation CHO Cell Platform." *Horizon Discover* 106 (97): 655581.
- Furey, Terrence S. 2012. "ChIP-Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions." *Nature Reviews Genetics* 13 (12). Nature

- Publishing Group: 840–52. doi:10.1038/nrg3306.
- Gaidukov, Leonid, Liliana Wroblewska, Brian Teague, Tom Nelson, Xin Zhang, Yan Liu, Kalpana Jagtap, et al. 2018. “A Multi-Landing Pad DNA Integration Platform for Mammalian Cell Engineering” 46 (8). Oxford University Press: 4072–86. doi:10.1093/nar/gky216.
- Gammell, Patrick, Niall Barron, Niraj Kumar, and Martin Clynes. 2007. “Initial Identification of Low Temperature and Culture Stage Induction of miRNA Expression in Suspension CHO-K1 Cells.” *Journal of Biotechnology* 130 (3): 213–18. doi:10.1016/j.jbiotec.2007.04.020.
- Gaspar, John M. 2018. “Improved Peak-Calling with MACS2.” *bioRxiv*, 496521. doi:10.1101/496521.
- Gauthier, Nicolas, Jiang Wei Wu, Shu Pei Wang, Pierre Allard, Orval A. Mamer, Lawrence Sweetman, Ann B. Moser, et al. 2013. “A Liver-Specific Defect of Acyl-CoA Degradation Produces Hyperammonemia, Hypoglycemia and a Distinct Hepatic Acyl-CoA Pattern.” Edited by François Blachier. *PLoS ONE* 8 (7): e60581. doi:10.1371/journal.pone.0060581.
- Geisler, Sarah, and Jeff Collier. 2013. “RNA in Unexpected Places: Long Non-Coding RNA Functions in Diverse Cellular Contexts.” *Nature Publishing Group* 14 (11). Nature Publishing Group: 699–712. doi:10.1038/nrm3679.
- Gendron, J., C. Colace-Sauty, N. Beaume, H. Cartonnet, J. Guegan, D. Ulveling, C. Pardanaud-Glavieux, I. Moszer, H. Cheval, and P. Ravassard. 2019. “Long Non-Coding RNA Repertoire and Open Chromatin Regions Constitute Midbrain Dopaminergic Neuron - Specific Molecular Signatures.” *Scientific Reports* 9 (1): 1–16. doi:10.1038/s41598-018-37872-1.
- Giresi, Paul G., Jonghwan Kim, Ryan M. McDaniel, Vishwanath R. Iyer, and Jason D. Lieb. 2007. “FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Isolates Active Regulatory Elements from Human Chromatin.” *Genome Research* 17 (6): 877–85. doi:10.1101/gr.5533506.
- Gonzalez, Inma, Roberto Munita, Eneritz Agirre, Travis A Dittmer, Katia Gysling, Tom Misteli, and Reini F Luco. 2015. “A lncRNA Regulates Alternative Splicing via Establishment of a Splicing-Specific Chromatin Signature.” *Nature Structural & Molecular Biology* 22 (5): 370–76. doi:10.1038/nsmb.3005.
- Goyal, Ashish, Evgenij Fiškin, Tony Gutschner, Maria Polycarpou-Schwarz, Matthias Groß, Julia Neugebauer, Minakshi Gandhi, Maiwen Caudron-Herger, Vladimir Benes, and Sven Diederichs. 2017. “A Cautionary Tale of Sense-Antisense Gene Pairs: Independent Regulation despite Inverse Correlation of Expression.” *Nucleic Acids Research* 45 (21): 12496–508. doi:10.1093/nar/gkx952.
- Grant, Charles E, Timothy L Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” *Bioinformatics (Oxford, England)* 27 (7): 1017–18. doi:10.1093/bioinformatics/btr064.
- Grav, Lise Marie, Jae Seong Lee, Signe Gerling, Thomas Beuchert Kallehauge, Anders Holmgaard Hansen, Stefan Kol, Gyun Min Lee, Lasse Ebdrup Pedersen, and Helene Fastrup Kildegaard. 2015. “One-Step Generation of Triple Knockout CHO Cell Lines Using CRISPR Cas9 and Fluorescent Enrichment.” *Biotechnology Journal*, n/a-n/a. doi:10.1002/biot.201500027.
- Gump, Jacob M, and Andrew Thorburn. 2011. “Autophagy and Apoptosis: What Is the

- Connection?" *Trends in Cell Biology* 21 (7): 387–92. doi:10.1016/j.tcb.2011.03.007.
- Guo, Wenbin, Cristiane P G Calixto, John W S Brown, and Runxuan Zhang. 2017. "TSIS: An R Package to Infer Alternative Splicing Isoform Switches for Time-Series Data." *Bioinformatics (Oxford, England)* 33 (20): 3308–10. doi:10.1093/bioinformatics/btx411.
- Guo, Yan, Shilin Zhao, Quanhui Sheng, Mingsheng Guo, Brian Lehmann, Jennifer Pietenpol, David C Samuels, and Yu Shyr. 2015. "RNAseq by Total RNA Library Identifies Additional RNAs Compared to Poly (A) RNA Library" 2015.
- Gutiérrez-González, Matías, Yesenia Latorre, Roberto Zúñiga, Juan Carlos Aguillón, María Carmen Molina, and Claudia Altamirano. 2019. "Transcription Factor Engineering in CHO Cells for Recombinant Protein Production." *Critical Reviews in Biotechnology* 39 (5): 665–79. doi:10.1080/07388551.2019.1605496.
- Gutschner, Tony, Monika Hämmerle, Moritz Eissmann, Jeff Hsu, Youngsoo Kim, Gene Hung, Alexey Revenko, et al. 2013. "The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells." *Cancer Research* 73 (3): 1180–89. doi:10.1158/0008-5472.CAN-12-2850.
- Guttman, Mitchell, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, et al. 2009. "Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals." *Nature* 457 (7235). Nature Publishing Group: 223–27. doi:10.1038/nature07672.
- Hacker, David L, Maria De Jesus, and Florian M Wurm. 2009. "25 Years of Recombinant Proteins from Reactor-Grown Cells — Where Do We Go from Here ?" *Biotechnology Advances* 27 (6). Elsevier Inc.: 1023–27. doi:10.1016/j.biotechadv.2009.05.008.
- Hackl, Matthias, Tobias Jakobi, Jochen Blom, Daniel Doppmeier, Karina Brinkrolf, Rafael Szczepanowski, Stephan H Bernhart, et al. 2011. "Next-Generation Sequencing of the Chinese Hamster Ovary microRNA Transcriptome : Identification , Annotation and Profiling of microRNAs as Targets for Cellular Engineering □." *Journal of Biotechnology* 153 (1–2). Elsevier B.V.: 62–75. doi:10.1016/j.jbiotec.2011.02.011.
- Halfon, Marc S. 2019a. "Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases." *Trends in Genetics* 35 (2). Elsevier Ltd: 93–103. doi:10.1016/j.tig.2018.11.004.
- Hammond, Stephanie, Jeffrey C Swanberg, Mihailo Kaplarevic, and Kelvin H Lee. 2011. "Genomic Sequencing and Analysis of a Chinese Hamster Ovary Cell Line Using Illumina Sequencing Technology."
- Han, Leng, and Zhongming Zhao. 2009. "CpG Islands or CpG Clusters: How to Identify Functional GC-Rich Regions in a Genome?" *BMC Bioinformatics* 10 (1): 65. doi:10.1186/1471-2105-10-65.
- Han, Young Kue, Tae Kwang Ha, So Jeong Lee, Jae Seong Lee, and Gyun Min Lee. 2011. "Autophagy and Apoptosis of Recombinant Chinese Hamster Ovary Cells during Fed-Batch Culture: Effect of Nutrient Supplementation." *Biotechnology and Bioengineering* 108 (9): 2182–92. doi:10.1002/bit.23165.
- Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, et al. 2012. "GENCODE : The Reference Human Genome Annotation for The ENCODE Project," 1760–74. doi:10.1101/gr.135350.111.

- Hartley, Fiona, Tracy Walker, Vicky Chung, and Karl Morten. 2018. "Mechanisms Driving the Lactate Switch in Chinese Hamster Ovary Cells." *Biotechnology and Bioengineering* 115 (8): 1890–1903. doi:10.1002/bit.26603.
- Heffner, Kelley M., Deniz Baycin Hizal, George S. Yerganian, Amit Kumar, Özge Can, Robert O'meally, Robert Cole, et al. 2017. "Lessons from the Hamster: Cricetulus Griseus Tissue and CHO Cell Line Proteome Comparison." *Journal of Proteome Research* 16 (10): 3672–87. doi:10.1021/acs.jproteome.7b00382.
- Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4). Elsevier Inc.: 576–89. doi:10.1016/j.molcel.2010.05.004.
- Hernandez, Inmaculada, Heena Dhiman, Gerald Klanert, Vaibhav Jadhav, Norbert Auer, Michael Hanscho, Martina Baumann, et al. 2019. "Epigenetic Regulation of Gene Expression in Chinese Hamster Ovary Cells in Response to the Changing Environment of a Batch Culture." *Biotechnology and Bioengineering* 116 (3): 677–92. doi:10.1002/bit.26891.
- Hess, Jochen, Peter Angel, and Marina Schorpp-Kistner. 2004. "AP-1 Subunits: Quarrel and Harmony among Siblings." *Journal of Cell Science* 117 (25): 5965–73. doi:10.1242/jcs.01589.
- Hesselberth, Jay R, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, et al. 2009. "Global Mapping of Protein-DNA Interactions in Vivo by Digital Genomic Footprinting" 6 (4): 283–89. doi:10.1038/NMETH.1313.
- Hollander, M. Christine, Isaac Alamo, and Albert J. Fornace. 1996. "A Novel DNA Damage-Inducible Transcript, gadd7, Inhibits Cell Growth, but Lacks a Protein Product." *Nucleic Acids Research* 24 (9): 1589–93. doi:10.1093/nar/24.9.1589.
- Hong, Jong Kwang, Yeon G. Kim, Sung Kwan Yoon, and Gyun Min Lee. 2007. "Down-Regulation of Cold-Inducible RNA-Binding Protein Does Not Improve Hypothermic Growth of Chinese Hamster Ovary Cells Producing Erythropoietin." *Metabolic Engineering* 9 (2): 208–16. doi:10.1016/j.ymben.2006.11.003.
- Hoque, Mainul, Zhe Ji, Dinghai Zheng, Wenting Luo, Wencheng Li, Bei You, Ji Yeon Park, Ghassan Yehia, and Bin Tian. 2013. "Analysis of Alternative Cleavage and Polyadenylation by 3' Region Extraction and Deep Sequencing." *Nature Methods* 10 (2): 133–39. doi:10.1038/nmeth.2288.
- Horlbeck, Max A, Lea B Witkowsky, Benjamin Guglielmi, Joseph M Replogle, Luke A Gilbert, Jacqueline E Villalta, Sharon E Torigoe, Robert Tjian, and Jonathan S Weissman. 2016. "Nucleosomes Impede Cas9 Access to DNA in Vivo and in Vitro," 1–21. doi:10.7554/eLife.12677.
- Hou, Ningning, Yuxi Yang, Ian C. Scott, and Xin Lou. 2017. "The Sec Domain Protein Scfd1 Facilitates Trafficking of ECM Components during Chondrogenesis." *Developmental Biology* 421 (1): 8–15. doi:10.1016/j.ydbio.2016.11.010.
- Housman, Gali, and Igor Ulitsky. 2016. "Methods for Distinguishing between Protein-Coding and Long Noncoding RNAs and the Elusive Biological Purpose of Translation of Long Noncoding RNAs." *Biochimica et Biophysica Acta* 1859 (1). Elsevier B.V.: 31–40.

doi:10.1016/j.bbagrm.2015.07.017.

- Hsu, Han-hsiu, Michihiro Araki, Masao Moc, Yoshimi Hori, and Masahiro Mura. 2017. "A Systematic Approach to Time- Series Metabolite Profiling and RNA-Seq Analysis of Chinese Hamster Ovary Cell Culture," no. July 2016: 1–13. doi:10.1038/srep43518.
- Hu, Wei, Xiaoqing Li, Ruimin Cheng, Jingru Ke, Yamei Liu, Menglan Ma, Yuchun Cao, and Dongxian Liu. 2019. "NAV2 Facilitates Invasion of Cutaneous Melanoma Cells by Targeting SNAI2 through the GSK-3 β / β -Catenin Pathway." *Archives of Dermatological Research* 311 (5): 399–410. doi:10.1007/s00403-019-01909-w.
- Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." *Nucleic Acids Research* 37 (1): 1–13. doi:10.1093/nar/gkn923.
- Huang, Di, Hanna M. Petrykowska, Brendan F. Miller, Laura Elnitski, and Ivan Ovcharenko. 2019. "Identification of Human Silencers by Correlating Cross-Tissue Epigenetic Profiles and Gene Expression." *Genome Research* 29 (4): 657–67. doi:10.1101/gr.247007.118.
- Huang, Yao-ming, Weiwei Hu, Eddie Rustandi, Kevin Chang, Helena Yusuf-makagiansar, and Thomas Ryll. 2010. "Maximizing Productivity of CHO Cell-Based Fed-Batch Culture Using Chemically Defined Media Conditions and Typical Manufacturing Equipment." doi:10.1002/btpr.436.
- Hultqvist, Greta, Daniel Ocampo Daza, Dan Larhammar, and Manfred W. Kilimann. 2012. "Evolution of the Vertebrate Paralemming Gene Family: Ancient Origin of Gene Duplicates Suggests Distinct Functions." *PLoS ONE* 7 (7). doi:10.1371/journal.pone.0041850.
- Hung, Tiffany, Yulei Wang, Michael F Lin, Ashley K Koegel, Yojiro Kotake, Gavin D Grant, Hugo M Horlings, et al. 2011. "Extensive and Coordinated Transcription of Noncoding RNAs within Cell-Cycle Promoters." *Nature Genetics* 43 (7). Nature Publishing Group: 621–29. doi:10.1038/ng.848.
- Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang. 2018. "Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines." *Experimental and Molecular Medicine* 50 (8). Springer US. doi:10.1038/s12276-018-0071-8.
- Hwang, Sun Ok, and Gyun Min Lee. 2009. "Effect of Akt Overexpression on Programmed Cell Death in Antibody-Producing Chinese Hamster Ovary Cells" 139: 89–94. doi:10.1016/j.jbiotec.2008.09.008.
- Ilott, Nicholas E., James A. Heward, Benoit Roux, Eleni Tsitsiou, Peter S. Fenwick, Luca Lenzi, Ian Goodhead, et al. 2014. "Long Non-Coding RNAs and Enhancer RNAs Regulate the Lipopolysaccharide-Induced Inflammatory Response in Human Monocytes." *Nature Communications* 5 (June): 3979. doi:10.1038/ncomms4979.
- Ilott, Nicholas E, and Chris P Ponting. 2013. "Predicting Long Non-Coding RNAs Using RNA Sequencing." *Methods* 63 (1). Elsevier Inc.: 50–59. doi:10.1016/j.ymeth.2013.03.019.
- Iyer, Matthew K, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, et al. 2015. "The Landscape of Long Noncoding RNAs in the Human Transcriptome" 47 (3). doi:10.1038/ng.3192.
- Jadhav, Vaibhav, Matthias Hackl, Aliaksandr Druz, Smriti Shridhar, Cheng-yu Chung, Kelley M Heffner, David P Kreil, et al. 2013. "CHO microRNA Engineering Is Growing up : Recent Successes and Future Challenges." *Biotechnology Advances* 31 (8). The Authors: 1501–

13. doi:10.1016/j.biotechadv.2013.07.007.

- Jadhav, Vaibhav, Matthias Hackl, Gerald Klanert, Juan A. Hernandez Bort, Renate Kunert, Johannes Grillari, and Nicole Borth. 2014. "Stable Overexpression of miR-17 Enhances Recombinant Protein Production of CHO Cells." *Journal of Biotechnology* 175 (1). Elsevier B.V.: 38–44. doi:10.1016/j.jbiotec.2014.01.032.
- Jain, Shilpa, Xiaoling Zhang, Preeti J. Khandelwal, Aleister J. Saunders, Brian S. Cummings, and Peter Oelkers. 2009. "Characterization of Human Lysophospholipid Acyltransferase 3." *Journal of Lipid Research* 50 (8): 1563–70. doi:10.1194/jlr.M800398-JLR200.
- Jayavelu, Naresh Doni, Ajay Jajodia, Arpit Mishra, and R. David Hawkins. 2018. "An Atlas of Silencer Elements for the Human and Mouse Genomes." *bioRxiv*, 252304. doi:10.1101/252304.
- Jeon, Min Kyoung, Da Young Yu, and Gyun Min Lee. 2011. "Combinatorial Engineering of Ldh-a and Bcl-2 for Reducing Lactate Production and Improving Cell Growth in Dihydrofolate Reductase-Deficient Chinese Hamster Ovary Cells." *Applied Microbiology and Biotechnology* 92 (4): 779–90. doi:10.1007/s00253-011-3475-0.
- Jia, Guangshuai, Jens Preussner, Xi Chen, Stefan Guenther, Xuejun Yuan, Michail Yekelchik, Carsten Kuenne, et al. 2018. "Single Cell RNA-Seq and ATAC-Seq Analysis of Cardiac Progenitor Cell Transition States and Lineage Settlement." *Nature Communications* 9 (1). Springer US. doi:10.1038/s41467-018-07307-6.
- Jiang, Wenjun, Dandan Zhang, Baoning Xu, Zhuo Wu, Siyang Liu, Lei Zhang, Ye Tian, Xu Han, and Dali Tian. 2015. "Long Non-Coding RNA BANC1 Promotes Proliferation and Migration of Lung Carcinoma via MAPK Pathways." *Biomedicine & Pharmacotherapy* 69 (February). Elsevier Masson SAS: 90–95. doi:10.1016/j.biopha.2014.11.027.
- Johnson, L Steven, Sean R Eddy, and Elon Portugaly. 2010. "Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure." *BMC Bioinformatics* 11 (1): 431. doi:10.1186/1471-2105-11-431.
- Johnsson, Per, Amanda Ackley, Linda Vidarsdottir, Weng-onn Lui, Martin Corcoran, Dan Grandér, and Kevin V Morris. 2013. "A Pseudogene Long-Noncoding-RNA Network Regulates PTEN Transcription and Translation in Human Cells" 20 (4). doi:10.1038/nsmb.2516.
- Johnsson, Per, Leonard Lipovich, Dan Grandér, and Kevin V Morris. 2014. "Evolutionary Conservation of Long Non-Coding RNAs; Sequence, Structure, Function." *Biochimica et Biophysica Acta* 1840 (3): 1063–71. doi:10.1016/j.bbagen.2013.10.035.
- Jung, Chang Hwa, Seung-Hyun Ro, Jing Cao, Neil Michael Otto, and Do-Hyung Kim. 2010. "mTOR Regulation of Autophagy." *FEBS Letters* 584 (7): 1287–95. doi:10.1016/j.febslet.2010.01.017.
- Kalvari, Ioanna, Joanna Argasinska, Natalia Quinones-Olvera, Eric P. Nawrocki, Elena Rivas, Sean R. Eddy, Alex Bateman, Robert D. Finn, and Anton I. Petrov. 2017. "Rfam 13.0: Shifting to a Genome-Centric Resource for Non-Coding RNA Families." *Nucleic Acids Research*, no. November. Oxford University Press: 1–8. doi:10.1093/nar/gkx1038.
- Kandoi, Gaurav, and Julie A. Dickerson. 2019. "Tissue-Specific Mouse mRNA Isoform Networks." *Scientific Reports* 9 (1). Springer US: 1–24. doi:10.1038/s41598-019-50119-x.
- Kang, Yu-Jian, De-Chang Yang, Lei Kong, Mei Hou, Yu-Qi Meng, Liping Wei, and Ge Gao.

- 2017a. "CPC2: A Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features." *Nucleic Acids Research* 45 (W1): W12–16. doi:10.1093/nar/gkx428.
- Kapustin, Yuri, Alexander Souvorov, Tatiana Tatusova, and David Lipman. 2008. "Splign: Algorithms for Computing Spliced Alignments with Identification of Paralogs." *Biology Direct* 3 (1): 20. doi:10.1186/1745-6150-3-20.
- Karnuta, Jaret M, and Peter C Scacheri. 2018. "Enhancers : Bridging the Gap between Gene Control and Human Disease" 27 (May): 219–27. doi:10.1093/hmg/ddy167.
- Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, et al. 2003. "The UCSC Genome Browser Database." *Nucleic Acids Research* 31 (1): 51–54. doi:10.1093/nar/gkg129.
- Kern, Colin, Ying Wang, James Chitwood, Ian Korf, Mary Delany, Hans Cheng, Juan F Medrano, et al. 2018. "Genome-Wide Identification of Tissue- Specific Long Non-Coding RNA in Three Farm Animal Species." *BMC Genomics*, 1–14.
- Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, et al. 2018. "JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework." *Nucleic Acids Research* 46 (D1): D260–66. doi:10.1093/nar/gkx1126.
- Khan, Dilshad H., and James R. Davie. 2013. "HDAC Inhibitors Prevent the Induction of the Immediate-Early Gene FOSL1, but Do Not Alter the Nucleosome Response." *FEBS Letters* 587 (10). Federation of European Biochemical Societies: 1510–17. doi:10.1016/j.febslet.2013.03.029.
- Kim, Jee Yon, Yeon Gu Kim, and Gyun Min Lee. 2012. "CHO Cells in Biotechnology for Production of Recombinant Proteins: Current State and Further Potential." *Applied Microbiology and Biotechnology* 93 (3): 917–30. doi:10.1007/s00253-011-3758-5.
- Klemm, Sandy L, Zohar Shipony, and William J Greenleaf. 2019. "Chromatin Accessibility and the Regulatory Epigenome." *Nature Reviews. Genetics* 20 (4). Springer US: 207–20. doi:10.1038/s41576-018-0089-8.
- Kondili, Maria, Annika Fust, Jens Preussner, Carsten Kuenne, and Thomas Braun. 2017. "OPEN UROPA : A Tool for Universal RObust Peak Annotation." Springer US, 1–12. doi:10.1038/s41598-017-02464-y.
- Kong, Lei, Yong Zhang, Zhi Qiang Ye, Xiao Qiao Liu, Shu Qi Zhao, Liping Wei, and Ge Gao. 2007. "CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine." *Nucleic Acids Research* 35 (SUPPL.2): 345–49. doi:10.1093/nar/gkm391.
- Kornfeld, Jan-Wilhelm, and Jens C Brüning. 2014. "Regulation of Metabolism by Long, Non-Coding RNAs." *Frontiers in Genetics* 5: 57. doi:10.3389/fgene.2014.00057.
- Kozomara, Ana, and Sam Griffiths-Jones. 2014. "MiRBase: Annotating High Confidence microRNAs Using Deep Sequencing Data." *Nucleic Acids Research* 42 (D1): 68–73. doi:10.1093/nar/gkt1181.
- Krchňáková, Zuzana, Prasoon Kumar Thakur, Michaela Krausová, Nicole Bieberstein, Nejc Haberman, Michaela Müller-McNicoll, and David Staněk. 2019. "Splicing of Long Non-Coding RNAs Primarily Depends on Polypyrimidine Tract and 5' Splice-Site Sequences due to Weak Interactions with SR Proteins." *Nucleic Acids Research* 47 (2): 911–28.

doi:10.1093/nar/gky1147.

- Kumar, Niraj, Patrick Gammell, Paula Meleady, Michael Henry, and Martin Clynes. 2008. "Differential Protein Expression Following Low Temperature Culture of Suspension CHO-K1 Cells." *BMC Biotechnology* 8 (1): 42. doi:10.1186/1472-6750-8-42.
- Kung, Johnny T Y, David Colognori, and Jeannie T Lee. 2013. "Long Noncoding RNAs: Past, Present, and Future." *Genetics* 193 (3): 651–69. doi:10.1534/genetics.112.146704.
- Lamble, Sarah, Elizabeth Batty, Moustafa Attar, David Buck, Rory Bowden, Gerton Lunter, Derrick Crook, Bassam El-Fahmawi, and Paolo Piazza. 2013. "Improved Workflows for High Throughput Library Preparation Using the Transposome-Based Nextera System." *BMC Biotechnology* 13. doi:10.1186/1472-6750-13-104.
- Landt, Stephen G., Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, et al. 2012. "ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia." *Genome Research* 22 (9): 1813–31. doi:10.1101/gr.136184.111.
- Lawrence, Moyra, Sylvain Daujat, and Robert Schneider. 2016a. "Lateral Thinking: How Histone Modifications Regulate Gene Expression." *Trends in Genetics* 32 (1). Elsevier Ltd: 42–56. doi:10.1016/j.tig.2015.10.007.
- Le, Huong, Santosh Kabbur, Luciano Pollastrini, Ziran Sun, Keri Mills, Kevin Johnson, George Karypis, and Wei-Shou Hu. 2012. "Multivariate Analysis of Cell Culture Bioprocess Data--Lactate Consumption as Process Indicator." *Journal of Biotechnology* 162 (2–3). Federation of European Biochemical Societies: 210–23. doi:10.1016/j.jbiotec.2012.08.021.
- Lee, Bum-Kyu, Nadima Uprety, Yu Jin Jang, Scott K Tucker, Catherine Rhee, Lucy LeBlanc, Samuel Beck, and Jonghwan Kim. 2018. "Fosl1 Overexpression Directly Activates Trophoblast-Specific Gene Expression Programs in Embryonic Stem Cells." *Stem Cell Research* 26: 95–102. doi:10.1016/j.scr.2017.12.004.
- Lee, Catherine, and Nobuaki Kikyo. 2012. "Strategies to Identify Long Noncoding RNAs Involved in Gene Regulation," 1–6.
- Lee, Jae Seong, Tae Kwang Ha, Jin Hyoung Park, and Gyun Min Lee. 2013. "Anti-Cell Death Engineering of CHO Cells : Co-Overexpression of Bcl-2 for Apoptosis Inhibition , Beclin-1 for Autophagy Induction" 110 (8): 2195–2207. doi:10.1002/bit.24879.
- Lee, J T, L S Davidow, and D Warshawsky. 1999. "Tsix, a Gene Antisense to Xist at the X-Inactivation Centre." *Nature Genetics* 21 (april): 400–404. doi:10.1038/7734.
- Legnini, Ivano, Gaia Di Timoteo, Francesca Rossi, Mariangela Morlando, Francesca Briganti, Olga Sthandier, Alessandro Fatica, et al. 2017. "Circ-ZNF609 Is a Circular RNA That Can Be Translated and Functions in Myogenesis." *Molecular Cell* 66 (1): 22–37.e9. doi:10.1016/j.molcel.2017.02.017.
- Leung, Karen N, and Barbara Panning. 2014. "X-Inactivation: Xist RNA Uses Chromosome Contacts to Coat the X." *Current Biology* 24 (2). Elsevier: R80–82. doi:10.1016/j.cub.2013.11.052.
- Lewis, Amanda M, Nicholas R Abu-Absi, Michael C Borys, and Zheng Jian Li. 2016. "The Use of 'Omics Technology to Rationally Improve Industrial Mammalian Cell Line Performance." *Biotechnology and Bioengineering* 113 (1): 26–38. doi:10.1002/bit.25673.

- Lewis, Nathan E, Xin Liu, Yuxiang Li, Harish Nagarajan, George Yerganian, Edward O'Brien, Aarash Bordbar, et al. 2013. "Genomic Landscapes of Chinese Hamster Ovary Cell Lines as Revealed by the Cricetulus Griseus Draft Genome." *Nature Biotechnology* 31 (8). Nature Publishing Group: 759–67. doi:10.1038/nbt.2624.
- Li, Feng, Natarajan Vijayasankaran, Amy Shen, Robert Kiss, and Ashraf Amanullah. 2010. "Cell Culture Processes for Monoclonal Antibody Production." *mAbs* 2 (5): 466–79. doi:10.4161/mabs.2.5.12720.
- Li, Rui, Hongliang Zhu, and Yunbo Luo. 2016. "Understanding the Functions of Long Non-Coding RNAs through Their Higher-Order Structures." *International Journal of Molecular Sciences* 17 (5). doi:10.3390/ijms17050702.
- Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30. doi:10.1093/bioinformatics/btt656.
- Lin, Michael F., Irwin Jungreis, and Manolis Kellis. 2011. "PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-Coding Regions." *Bioinformatics* 27 (13): 275–82. doi:10.1093/bioinformatics/btr209.
- Liu, Liang, Guangxu Jin, and Xiaobo Zhou. 2015. "Modeling the Relationship of Epigenetic Modifications to Transcription Factor Binding" 43 (8): 3873–85. doi:10.1093/nar/gkv255.
- Liu, Xiangfan, Huapeng Li, Mihir Rajurkar, Qi Li, Jennifer L Cotton, Jianhong Ou, Lihua J Zhu, et al. 2016. "Tead and AP1 Coordinate Transcription and Motility." *Cell Reports* 14 (5): 1169–80. doi:10.1016/j.celrep.2015.12.104.
- Lorenz, Ronny, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. 2011. "ViennaRNA Package 2.0." *Algorithms for Molecular Biology* 6 (1): 26. doi:10.1186/1748-7188-6-26.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of fold-change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. doi:10.1186/s13059-014-0550-8.
- Love, Michael I, Charlotte Soneson, Rob Patro, Kristoffer Vitting-seerup, and Alicia Oshlack. 2019. "Swimming Downstream : Statistical Analysis of Differential Transcript Usage Following Salmon Quantification [Version 1 ; Peer Review : 3 Approved with Reservations] Referee Status : This Article Is Included in the RPackage Gateway .," no. May: 1–50.
- Luo, Jun, Natarajan Vijayasankaran, Jennifer Autsen, Rodell Santuray, and Terry Hudson. 2012. "Comparative Metabolite Analysis to Understand Lactate Metabolism Shift in Chinese Hamster Ovary Cell Culture Process" 109 (1): 146–56. doi:10.1002/bit.23291.
- Madrigal, Pedro, and Kaur Alasoo. 2018. "AP-1 Takes Centre Stage in Enhancer Chromatin Dynamics." *Trends in Cell Biology* 28 (7). Elsevier Ltd: 509–11. doi:10.1016/j.tcb.2018.04.009.
- Malphettes, Laetitia, Yevgeniy Freyvert, Jennifer Chang, Pei-qi Liu, Edmond Chan, Jeffrey C Miller, Zhe Zhou, et al. 2010. "Highly Efficient Deletion of FUT8 in CHO Cell Lines Using Zinc-Finger Nucleases Yields Cells That Produce Completely Nonfucosylated Antibodies" 106 (5): 774–83. doi:10.1002/bit.22751.
- Mathiassen, Søs Grønbæk, Daniela De Zio, and Francesco Cecconi. 2017. "Autophagy and the Cell Cycle: A Complex Landscape." *Frontiers in Oncology* 7 (MAR): 1–16.

doi:10.3389/fonc.2017.00051.

- Matlin, Arianne J., Francis Clark, and Christopher W. J. Smith. 2005. "Understanding Alternative Splicing: Towards a Cellular Code." *Nature Reviews Molecular Cell Biology* 6 (5): 386–98. doi:10.1038/nrm1645.
- Melé, Marta, Kaia Mattioli, William Mallard, David M. Shechner, Chiara Gerhardinger, and John L. Rinn. 2017. "Chromatin Environment, Transcriptional Regulation, and Splicing Distinguish lincRNAs and mRNAs." *Genome Research* 27 (1): 27–37. doi:10.1101/gr.214205.116.
- Mercer, Tim R, and John S Mattick. 2013. "Structure and Function of Long Noncoding RNAs in Epigenetic Regulation." *Nature Publishing Group* 20 (3). Nature Publishing Group: 300–307. doi:10.1038/nsmb.2480.
- Meyer, Laurence R., Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Robert M. Kuhn, Matthew Wong, Cricket A. Sloan, et al. 2013. "The UCSC Genome Browser Database: Extensions and Updates 2013." *Nucleic Acids Research* 41 (D1): 918–23. doi:10.1093/nar/gks1048.
- Misaghi, Shahram, Andrew Snowden, Jennifer Chang, and Brad Snedecor. 2013. "Resilient Immortals , Characterizing and Utilizing Bax / Bak Deficient Chinese Hamster Ovary (CHO) Cells for High Titer Antibody Production," 727–37. doi:10.1002/btpr.1722.
- Montefiori, Lindsey, Liana Hernandez, Zijie Zhang, Yoav Gilad, Carole Ober, Gregory Crawford, Marcelo Nobrega, and Noboru Jo Sakabe. 2017. "Reducing Mitochondrial Reads in ATAC-Seq Using CRISPR/Cas9." *Scientific Reports* 7 (1). Springer US: 1–9. doi:10.1038/s41598-017-02547-w.
- Moore, A, J Mercer, G Dutina, C J Donahue, K D Bauer, J P Mather, T Etcheverry, and T Ryll. 1997. "Effects of Temperature Shift on Cell Cycle, Apoptosis and Nucleotide Pools in CHO Cell Batch Cultures." *Cytotechnology* 23 (1–3): 47–54. doi:10.1023/A:1007919921991.
- Mudge, Jonathan M., and Jennifer Harrow. 2015. "Creating Reference Gene Annotation for the Mouse C57BL6/J Genome Assembly." *Mammalian Genome* 26 (9–10). Springer US: 366–78. doi:10.1007/s00335-015-9583-x.
- Müller, Dethardt, Hermann Katinger, and Johannes Grillari. 2008. "MicroRNAs as Targets for Engineering of CHO Cell Factories." *Trends in Biotechnology* 26 (7): 359–65. doi:10.1016/j.tibtech.2008.03.010.
- Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C Baker, Frank Grützner, and Henrik Kaessmann. 2014. "The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods." *Nature* 505 (7485): 635–40. doi:10.1038/nature12943.
- Neil, Dominic O, Heike Glowatz, and Martin Schlumpberger. 2013. "Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity," no. July: 1–8. doi:10.1002/0471142727.mb0419s103.
- Nie, Lei, Hsing-Ju Wu, Jung-Mao Hsu, Shih-Shin Chang, Adam M LaBaff, Chia-Wei Li, Yan Wang, Jennifer L Hsu, and Mien-Chie Hung. 2012. "Long Non-Coding RNAs: Versatile Master Regulators of Gene Expression and Crucial Players in Cancer." *American Journal of Translational Research* 4 (2): 127–50. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3353529/>.

- Niemczyk, Malwina, Yoko Ito, Joanna Huddleston, Anna Git, Sayeda Abu-Amero, Carlos Caldas, Gudrun E. Moore, Lovorka Stojic, and Adele Murrell. 2013a. "Imprinted Chromatin around DIRAS3 Regulates Alternative Splicing of GNG12-AS1, a Long Noncoding RNA." *American Journal of Human Genetics* 93 (2). The American Society of Human Genetics: 224–35. doi:10.1016/j.ajhg.2013.06.010.
- Nikolayeva, Olga, and Mark D. Robinson. 2014. "edgeR for Differential RNA-Seq and ChIP-Seq Analysis: An Application to Stem Cell Biology." In , 45–79. doi:10.1007/978-1-4939-0512-6_3.
- Noh, Soo Min, Seunghyeon Shin, and Gyun Min Lee. 2018. "Comprehensive Characterization of Glutamine Synthetase-Mediated Selection for the Establishment of Recombinant CHO Cells Producing Monoclonal Antibodies." *Scientific Reports* 8 (1). Springer US: 1–11. doi:10.1038/s41598-018-23720-9.
- O'Brien, Sofie A., Kyoungcho Lee, Hsu Yuan Fu, Zion Lee, Tung S. Le, Christopher S. Stach, Meghan G. McCann, et al. 2018. "Single Copy Transgene Integration in a Transcriptionally Active Site for Recombinant Protein Synthesis." *Biotechnology Journal* 13 (10): 1–10. doi:10.1002/biot.201800226.
- Osterlehner, Andrea, Silke Simmeth, Ulrich Go, Early Development, and Roche Diagnostics Gmbh. 2011. "Promoter Methylation and Transgene Copy Numbers Predict Unstable Protein Production in Recombinant Chinese Hamster Ovary Cell Lines" 108 (11): 2670–81. doi:10.1002/bit.23216.
- Pan, Xiao, Mathieu Streefland, and Ciska Dalm. 2017. "Selection of Chemically Defined Media for CHO Cell Fed-Batch Culture Processes," 39–56. doi:10.1007/s10616-016-0036-5.
- Park, Eddie, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. 2018. "The Expanding Landscape of Alternative Splicing Variation in Human Populations." *American Journal of Human Genetics* 102 (1). The Authors: 11–26. doi:10.1016/j.ajhg.2017.11.002.
- Patrucco, Laura, Andrea Chiesa, Maria Felicia Soluri, Francesca Fasolo, Hazuki Takahashi, Piero Carninci, Silvia Zucchelli, et al. 2015. "Engineering Mammalian Cell Factories with SINEUP Noncoding RNAs to Improve Translation of Secreted Proteins." *Gene* 569 (2). Elsevier B.V.: 287–93. doi:10.1016/j.gene.2015.05.070.
- Pelechano, Vincent, and Steinmetz. Lars M. 2013. "Gene Regulation by Antisense Transcription." *Nature Publishing Group* 14 (12). Nature Publishing Group: 880–93. <http://dx.doi.org/10.1038/nrg3594>.
- Penny, G D, G F Kay, S a Sheardown, S Rastan, and N Brockdorff. 1996. "Requirement for Xist in X Chromosome Inactivation." *Nature*. doi:10.1038/379131a0.
- Pertea, Mihaela, Geo M Pertea, Corina M Antonescu, Tsung-cheng Chang, Joshua T Mendell, and Steven L Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads" 33 (3). doi:10.1038/nbt.3122.
- Pilotte, Julie, Esther E. Dupont-Versteegden, and Peter W. Vanderklis. 2011. "Widespread Regulation of miRNA Biogenesis at the Dicer Step by the Cold-Inducible RNA-Binding Protein, RBM3." *PLoS ONE* 6 (12). doi:10.1371/journal.pone.0028446.
- Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2007. "NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins." *Nucleic Acids Research* 35 (SUPPL. 1): 61–65. doi:10.1093/nar/gkl842.

- Puvvula, Pavan Kumar, Rohini Devi Desetty, Pascal Pineau, Agn s Marchio, Anne Moon, Anne Dejean, and Oliver Bischof. 2014. "Long Noncoding RNA PANDA and Scaffold-Attachment-Factor SAFA Control Senescence Entry and Exit." *Nature Communications* 5: 5323. doi:10.1038/ncomms6323.
- Pybus, Leon P, Greg Dean, Nathan R West, Andrew Smith, Olalekan Daramola, Ray Field, Stephen J Wilkinson, and David C James. 2014. "Model-Directed Engineering of "Difficult-to-Express" Monoclonal Antibody Production by Chinese Hamster Ovary Cells" 111 (2): 372–85. doi:10.1002/bit.25116.
- Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools : A Flexible Suite of Utilities for Comparing Genomic Features" 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.
- Quinn, Jeffrey J., and Howard Y. Chang. 2015. "Unique Features of Long Non-Coding RNA Biogenesis and Function." *Nature Reviews. Genetics* 17 (1). Nature Publishing Group: 47–62. doi:10.1038/nrg.2015.10.
- Quinn, Jeffrey J., Qiangfeng C. Zhang, Plamen Georgiev, Ibrahim A. Ilik, Asifa Akhtar, and Howard Y. Chang. 2016. "Rapid Evolutionary Turnover Underlies Conserved lncRNA-Genome Interactions." *Genes and Development* 30 (2): 191–207. doi:10.1101/gad.272187.115.
- Ram rez, Fidel, Devon P. Ryan, Bj rn Gr ning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike D ndar, and Thomas Manke. 2016. "deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis." *Nucleic Acids Research* 44 (W1): W160–65. doi:10.1093/nar/gkw257.
- Rashid, Farooq, Abdullah Shah, and Ge Shan. 2016. "Long Non-Coding RNAs in the Cytoplasm." *Genomics, Proteomics & Bioinformatics* 14 (2). Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China: 73–80. doi:10.1016/j.gpb.2016.03.005.
- Ratter, Jacqueline M, Hanne M M Rooijackers, and Guido J Hooiveld. 2018. "In Vitro and in Vivo Effects of Lactate on Metabolism and Cytokine Production of Human Primary PBMCs and Monocytes" 9 (November): 1–13. doi:10.3389/fimmu.2018.02564.
- Reinhart, David, Lukas Damjanovic, Christian Kaisermayer, and Renate Kunert. 2015. "Benchmarking of Commercially Available CHO Cell Culture Media for Antibody Production," 4645–57. doi:10.1007/s00253-015-6514-4.
- Renaud, Stephen J, Kaiyu Kubota, M A Karim Rumi, and Michael J Soares. 2014. "The FOS Transcription Factor Family Differentially Controls Trophoblast Migration and Invasion." *The Journal of Biological Chemistry* 289 (8): 5025–39. doi:10.1074/jbc.M113.523746.
- Reyes, Alejandro, and Wolfgang Huber. 2018. "Alternative Start and Termination Sites of Transcription Drive Most Transcript Isoform Differences across Human Tissues." *Nucleic Acids Research* 46 (2): 582–92. doi:10.1093/nar/gkx1165.
- Ritter, Deborah I, Zhiqiang Dong, Su Guo, and Jeffrey H Chuang. 2012. "Transcriptional Enhancers in Protein-Coding Exons of Vertebrate Developmental Genes" 7 (5). doi:10.1371/journal.pone.0035202.
- Ross-Innes, Caryn S, Rory Stark, Andrew E Teschendorff, Kelly A Holmes, H Raza Ali, Mark J Dunning, Gordon D Brown, et al. 2012. "Differential Oestrogen Receptor Binding Is Associated with Clinical Outcome in Breast Cancer." *Nature* 481 (7381): 389–93.

doi:10.1038/nature10730.

Rupp, Oliver, Jennifer Becker, Karina Brinkrolf, Christina Timmermann, Nicole Borth, Thomas Noll, and Alexander Goesmann. 2014. "Construction of a Public CHO Cell Line Transcript Database Using Versatile Bioinformatics Analysis Pipelines" 9 (1). doi:10.1371/journal.pone.0085568.

Rupp, Oliver, Sven Griep, Kelley Heffner, Karina Brinkrolf, Nathan E Lewis, Brewster Kingham, Alexander Goesmann, Michael J Betenbaugh, Nicole Borth, and Kelvin H Lee. 2018. "A Reference Genome of the Chinese Hamster Based on a Hybrid Assembly Strategy," no. March: 2087–2100. doi:10.1002/bit.26722.

Sanchez-Garcia, Laura, Lucas Martín, Ramon Mangués, Neus Ferrer-Mirallès, Esther Vázquez, and Antonio Villaverde. 2016. "Recombinant Pharmaceuticals from Microbial Cells: A 2015 Update." *Microbial Cell Factories* 15 (1). BioMed Central: 33. doi:10.1186/s12934-016-0437-3.

Sanchez, Noelia, Paul Kelly, Clair Gallagher, Nga T Lao, Colin Clarke, Martin Clynes, and Niall Barron. 2014. "CHO Cell Culture Longevity and Recombinant Protein Yield Are Enhanced by Depletion of miR-7 Activity via Sponge Decoy Vectors," 396–404. doi:10.1002/biot.201300325.

Schaub, Jochen, Christoph Clemens, Peter Schorn, Tobias Hildebrandt, Werner Rust, Detlev Mennerich, Hitto Kaufmann, and Torsten W Schulz. 2010. "E DITOR CHO Gene Expression Profiling in Biopharmaceutical Process Analysis and Design" 105 (2): 431–38. doi:10.1002/bit.22549.

Schep, Alicia N., Beijing Wu, Jason D. Buenrostro, and William J. Greenleaf. 2017. "ChromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data." *Nature Methods* 14 (10): 975–78. doi:10.1038/nmeth.4401.

Sitton, Greg, and Friedrich Srienc. 2008. "Growth Dynamics of Mammalian Cells Monitored with Automated Cell Cycle Staining and Flow Cytometry." *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* 73 (6): 538–45. doi:10.1002/cyto.a.20567.

Soneson, Charlotte, Michael I Love, and Mark D Robinson. 2018. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences [Version 1 ; Referees : 2 Approved] Referee Status :," no. 2: 1–18. doi:10.12688/f1000research.7563.1.

Song, Lingyun, Zhancheng Zhang, Linda L Grasfeder, Alan P Boyle, Paul G Giresi, Bum-Kyu Lee, Nathan C Sheffield, et al. 2011a. "Open Chromatin Defined by DNaseI and FAIRE Identifies Regulatory Elements That Shape Cell-Type Identity." *Genome Research* 21 (10): 1757–67. doi:10.1101/gr.121541.111.

Souvorov, a, Y Kapustin, B Kiryutin, V Chetvernin, T Tatusova, and D Lipman. 2010. "Gnomon—NCBI Eukaryotic Gene Prediction Tool." *National Center for Biotechnology Information*, 1–24.

Stewart, Alexander J., and Joshua B. Plotkin. 2012. "Why Transcription Factor Binding Sites Are Ten Nucleotides Long." *Genetics* 192 (3): 973–85. doi:10.1534/genetics.112.143370.

Stiefel, Fabian, Simon Fischer, Alexander Sczyrba, Kerstin Otte, and Friedemann Hesse. 2016. "miRNA Profiling of High , Low and Non-Producing CHO Cells during Biphasic Fed-Batch Cultivation Reveals Process Relevant Targets for Host Cell Engineering." *Journal of*

- Biotechnology* 225. Elsevier B.V.: 31–43. doi:10.1016/j.jbiotec.2016.03.028.
- Struhl, Kevin. 2007. “Transcriptional Noise and the Fidelity of Initiation by RNA Polymerase II.” *Nature Structural & Molecular Biology* 14 (2): 103–5. doi:10.1038/nsmb0207-103.
- Su, Ye, Haijiang Wu, Alexander Pavlosky, Ling-Lin Zou, Xinna Deng, Zhu-Xu Zhang, and Anthony M Jevnikar. 2016. “Regulatory Non-Coding RNA: New Instruments in the Orchestration of Cell Death.” *Cell Death & Disease* 7 (8): e2333. doi:10.1038/cddis.2016.210.
- Sultan, Marc, Vyacheslav Amstislavskiy, Thomas Risch, Moritz Schuette, Simon Dökel, Meryem Ralser, Daniela Balzareit, Hans Lehrach, and Marie-laure Yaspo. 2014. “Influence of RNA Extraction Methods and Library Selection Schemes on RNA-Seq Data.”
- Sumit, Madhuresh, Sepideh Dolatshahi, An Hsiang Adam Chu, Kaffa Cote, John J. Scarcelli, Jeffrey K. Marshall, Richard J. Cornell, et al. 2019. “Dissecting N-Glycosylation Dynamics in Chinese Hamster Ovary Cells Fed-Batch Cultures Using Time Course Omics Analyses.” *iScience* 12. Elsevier Inc.: 102–20. doi:10.1016/j.isci.2019.01.006.
- Sun, Xinghui, and Danny Wong. 2016. “Long Non-Coding RNA-Mediated Regulation of Glucose Homeostasis and Diabetes” 6 (2): 17–25.
- Tan, Hong Kiat, May May Lee, Miranda G. S. Yap, and Daniel I. C. Wang. 2008. “Overexpression of Cold-Inducible RNA-Binding Protein Increases Interferon- γ Production in Chinese-Hamster Ovary Cells.” *Biotechnology and Applied Biochemistry* 49 (4): 247. doi:10.1042/BA20070032.
- Tarbell, Evan D, and Tao Liu. 2019. “HMMRATAC: A Hidden Markov Modeler for ATAC-Seq.” *Nucleic Acids Research* 47 (16). Oxford University Press: 1–14. doi:10.1093/nar/gkz533.
- Thomas, Reuben, Sean Thomas, Alisha K Holloway, and Katherine S Pollard. 2017. “Features That Define the Best ChIP-Seq Peak Calling Algorithms.” *Briefings in Bioinformatics* 18 (3): 441–50. doi:10.1093/bib/bbw035.
- Thomson, Daniel W, and Marcel E Dinger. 2016. “Endogenous microRNA Sponges : Evidence and Controversy.” *Nature Publishing Group* 17 (5). Nature Publishing Group: 272–83. doi:10.1038/nrg.2016.20.
- Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5). Nature Publishing Group: 511–15. doi:10.1038/nbt.1621.
- Tripathi, Vidisha, Jonathan D Ellis, Zhen Shen, David Y Song, Qun Pan, Andrew T Watt, Susan M Freier, et al. 2010. “Article The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation.” *Molecular Cell* 39 (6). Elsevier Ltd: 925–38. doi:10.1016/j.molcel.2010.08.011.
- Tsao, Y, A G Cardoso, R G G Condon, M Voloch, P Lio, J C Lagos, B G Kearns, and Z Liu. 2005. “Monitoring Chinese Hamster Ovary Cell Culture by the Analysis of Glucose and Lactate Metabolism” 118: 316–27. doi:10.1016/j.jbiotec.2005.05.016.
- Tsompana, Maria, and Michael J. Buck. 2014. “Chromatin Accessibility: A Window into the Genome.” *Epigenetics and Chromatin* 7 (1): 1–16. doi:10.1186/1756-8935-7-33.

- Turk, Casey M., Katerina D. Fagan-Solis, Kristin E. Williams, Joseph M. Gozgit, Sallie Smith-Schneider, Sharon A. Marconi, Christopher N. Otis, et al. 2012. "Paralemm-1 Is over-Expressed in Estrogen-Receptor Positive Breast Cancers." *Cancer Cell International* 12. doi:10.1186/1475-2867-12-17.
- Turnover, Stabilizing Schlafen- Protein, Dianrong Li, Jie Chen, Youwei Ai, Dianrong Li, Jie Chen, Youwei Ai, et al. 2019. "Estrogen-Related Hormones Induce Apoptosis by Article Estrogen-Related Hormones Induce Apoptosis by Stabilizing Schlafen-12 Protein Turnover," 1103–16. doi:10.1016/j.molcel.2019.06.040.
- Tzani, Ioanna, Craig Monger, Krishna Motheramgari, Clair Gallagher, Ryan Hagan, Alan Costello, Justine Meiller, et al. 2019. "Article Sub Physiological Temperature Induces Pervasive Alternative Splicing in Chinese Hamster Ovary Cells," 1–29.
- Urlaub, Gail, K Emmanuel, Adelaide M Carothers, and Lawrence A Chasin. 1983. "D + -- D + dR D + dR" 33 (June): 405–12.
- Vacik, Tomas. 2017a. "Alternative Intronic Promoters in Development and Disease." *Protoplasma*. Protoplasma. doi:10.1007/s00709-016-1071-y.
- Vance, Keith W., and Chris P. Ponting. 2014. "Transcriptional Regulatory Functions of Nuclear Long Noncoding RNAs." *Trends in Genetics* 30 (8). Elsevier Ltd: 348–55. doi:10.1016/j.tig.2014.06.001.
- Veith, Nathalie, Holger Ziehr, Roderick A F Macleod, and Stella Marie Reamon-buettner. 2016. "Mechanisms Underlying Epigenetic and Transcriptional Heterogeneity in Chinese Hamster Ovary (CHO) Cell Lines." *BMC Biotechnology*. BMC Biotechnology, 1–16. doi:10.1186/s12896-016-0238-0.
- Vergara, Mauricio, Mauro Torres, Andrea Mu, Cristian Acevedo, Julio Berrios, Juan G Reyes, Norma A Valdez-cruz, and Claudia Altamirano. 2018. "High Glucose and Low Specific Cell Growth but Not Mild Hypothermia Improve Specific R-Protein Productivity in Chemostat Culture of CHO Cells," 1–22.
- Villegas, Victoria E, and Peter G Zaphiropoulos. 2015. "Neighboring Gene Regulation by Antisense Long Non-Coding RNAs," 3251–66. doi:10.3390/ijms16023251.
- Vito, Davide, and Christopher Mark Smales. 2018. "The Long Non-Coding RNA Transcriptome Landscape in CHO Cells Under Batch and Fed-Batch Conditions." *Biotechnology Journal* 13 (10): 1–12. doi:10.1002/biot.201800122.
- Vitting-seerup, Kristoffer, and Albin Sandelin. 2019. "IsoformSwitchAnalyzeR : Analysis of Changes in Genome-Wide Patterns of Alternative Splicing and Its Functional Consequences," 2–3.
- Vitting-Seerup, Kristoffer, and Albin Sandelin. 2017. "The Landscape of Isoform Switches in Human Cancers." *Molecular Cancer Research* 15 (9): 1206–20. doi:10.1158/1541-7786.MCR-16-0459.
- Wahrheit, Judith, Averina Nicolae, and Elmar Heinzle. 2014. "Dynamics of Growth and Metabolism Controlled by Glutamine Availability in Chinese Hamster Ovary Cells." *Applied Microbiology and Biotechnology* 98 (4): 1771–83. doi:10.1007/s00253-013-5452-2.
- Walsh, Gary. 2014. "Biopharmaceutical Benchmarks 2014." *Nature Biotechnology* 32 (7): 992–1000. doi:10.1038/nbt0910-917.

- Wan, Ledong, Jianlu Kong, Jinlong Tang, Yihua Wu, Enping Xu, and Maode Lai. 2016. "HOTAIRM1 as a Potential Biomarker for Diagnosis of Colorectal Cancer Functions the Role in the Tumour Suppressor" 20 (11): 2036–44. doi:10.1111/jcmm.12892.
- Wang, Kevin C., and Howard Y. Chang. 2012. "Molecular Mechanisms of Long Noncoding RNAs Kevin." *Mol Cell* 43 (6): 904–14. doi:10.1016/j.molcel.2011.08.018.Molecular.
- Wang, Liguang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean Pierre Kocher, and Wei Li. 2013. "CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model." *Nucleic Acids Research* 41 (6): 1–7. doi:10.1093/nar/gkt006.
- Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1): 57–63. doi:10.1038/nrg2484.
- Wellmann, Sven, Matthias Truss, Elisabeth Bruder, Luigi Tornillo, Andrea Zelmer, Karl Seeger, and Christoph Bührer. 2010. "The RNA-Binding Protein RBM3 Is Required for Cell Proliferation and Protects Against Serum Deprivation-Induced Cell Death." *Pediatric Research* 67 (1): 35–41. doi:10.1203/PDR.0b013e3181c13326.
- West, Jason A., Christopher P. Davis, Hongjae Sunwoo, Matthew D. Simon, Ruslan I. Sadreyev, Peggy I. Wang, Michael Y. Tolstorukov, and Robert E. Kingston. 2014. "The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites." *Molecular Cell* 55 (5). Elsevier Inc.: 791–802. doi:10.1016/j.molcel.2014.07.012.
- Wight, Megan, and Andreas Werner. 2013. "The Functions of Natural Antisense Transcripts." *Essays In Biochemistry* 54: 91–101. doi:10.1042/bse0540091.
- Wittkopp, Patricia J., and Gizem Kalay. 2012. "Cis-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence." *Nature Reviews Genetics* 13 (1). Nature Publishing Group: 59–69. doi:10.1038/nrg3095.
- Wu, Shyh-Jong, Mathilde Soulez, Ya-Hui Yang, Chih-Sheng Chu, Shih-Chuan Shih, Marie-Josée Hébert, Mei-Chuan Kuo, and Ya-Ju Hsieh. 2015. "Local Augmented Angiotensinogen Secreted from Apoptotic Vascular Endothelial Cells Is a Vital Mediator of Vascular Remodelling." Edited by Gianfranco Pintus. *PLOS ONE* 10 (7): e0132583. doi:10.1371/journal.pone.0132583.
- Wucher, Valentin, Fabrice Legeai, Benoit Hedan, Guillaume Rizk, Laetitia Lagoutte, Tosso Leeb, Vidhya Jagannathan, et al. 2016. "FEELnc: A Tool for Long Non-Coding RNAs Annotation and Its Application to the Dog Transcriptome." *bioRxiv*. doi:10.1101/064436.
- Wulhfard, Sarah, Sophie Bouchet, Jean Cevey, Maria De Jesus, David L Hacker, and Florian M Wurm. 2008. "Mild Hypothermia Improves Transient Gene Expression Yields Several Fold in Chinese Hamster Ovary Cells." *Biotechnology Progress* 24: 458–65.
- Wurm, Florian M. 2004. "Production of Recombinant Protein Therapeutics in Cultivated Mammalian Cells" 22 (11): 1393–98. doi:10.1038/nbt1026.
- Xia, Tian, Qi Liao, Xiaoming Jiang, Yongfu Shao, Bingxiu Xiao, Yang Xi, and Junming Guo. 2014. "Long Noncoding RNA Associated-Competing Endogenous RNAs in Gastric Cancer." *Scientific Reports* 4 (August): 6088. doi:10.1038/srep06088.
- Xu, Jianlin, Peifeng Tang, Andrew Yongky, Barry Drew, Michael C Borys, Shijie Liu, and Zheng Jian Li. 2019. "Systematic Development of Temperature Shift Strategies for Chinese Hamster Ovary Cells Based on Short Duration Cultures and Kinetic Modeling." *mAbs* 11 (1): 191–204. doi:10.1080/19420862.2018.1525262.

- Xu, Xun, Harish Nagarajan, Nathan E Lewis, Shengkai Pan, Zhiming Cai, Xin Liu, Wenbin Chen, et al. 2011. "The Genomic Sequence of the Chinese Hamster Ovary." *Nature Biotechnology* 29 (8). Nature Publishing Group: 1–8. doi:10.1038/nbt.1932.
- Yan, Kun, Yasir Arfat, Dijie Li, Fan Zhao, Zhihao Chen, Chong Yin, Yulong Sun, Lifang Hu, Tuanmin Yang, and Airong Qian. 2016. "Structure Prediction: New Insights into Decrypting Long Noncoding RNAs." *International Journal of Molecular Sciences* 17 (1). doi:10.3390/ijms17010132.
- Yang, Yuansheng, Mariati, Janet Chusainow, and Miranda G S Yap. 2010. "DNA Methylation Contributes to Loss in Productivity of Monoclonal Antibody-Producing CHO Cell Lines." *Journal of Biotechnology* 147 (3–4). Elsevier B.V.: 180–85. doi:10.1016/j.jbiotec.2010.04.004.
- Yoon, Je-Hyun, Kotb Abdelmohsen, and Myriam Gorospe. 2014. "Functional Interactions among microRNAs and Long Noncoding RNAs." *Seminars in Cell & Developmental Biology* 34. Elsevier Ltd: 9–14. doi:10.1016/j.semcdb.2014.05.015.
- Yoon, Je-hyun, Kotb Abdelmohsen, Subramanya Srikantan, Xiaoling Yang, Jennifer L Martindale, Supriyo De, Maite Huarte, Ming Zhan, Kevin G Becker, and Myriam Gorospe. 2012. "Short Article LincRNA-p21 Suppresses Target mRNA Translation." *MOLCEL* 47 (4). Elsevier: 648–55. doi:10.1016/j.molcel.2012.06.027.
- Yu, Guangchuang, Li Gen Wang, and Qing Yu He. 2015. "ChIP Seeker: An R/Bioconductor Package for ChIP Peak Annotation, Comparison and Visualization." *Bioinformatics* 31 (14): 2382–83. doi:10.1093/bioinformatics/btv145.
- Yu, Yang, Mingjiong Zhang, Jie Liu, Boming Xu, Jian Yang, Ni Wang, Shuai Yan, et al. 2018. "Long Non-Coding RNA PVT1 Promotes Cell Proliferation and Migration by Silencing ANGPTL4 Expression in Cholangiocarcinoma." *Molecular Therapy - Nucleic Acids* 13 (December). Elsevier Ltd.: 503–13. doi:10.1016/j.omtn.2018.10.001.
- Yuan, Hua, Hongliang Liu, Zhensheng Liu, Kouros Owzar, Younghun Han, Li Su, Yongyue Wei, et al. 2016. "A Novel Genetic Variant in Long Non-Coding RNA Gene NEXN-AS1 Is Associated with Risk of Lung Cancer." *Scientific Reports* 6 (1): 34234. doi:10.1038/srep34234.
- Yuan, Xuan, Zhe Li, Andrea C. Baines, Eleni Gavrilaki, Zhaohui Ye, Zhexing Wen, Evan M. Braunstein, et al. 2017a. "A Hypomorphic PIGA Gene Mutation Causes Severe Defects in Neuron Development and Susceptibility to Complement-Mediated Toxicity in a Human iPSC Model." Edited by Majlinda Lako. *PLOS ONE* 12 (4): e0174074. doi:10.1371/journal.pone.0174074.
- Zabidi, Muhammad A., and Alexander Stark. 2016. "Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors." *Trends in Genetics* 32 (12). Elsevier Ltd: 801–14. doi:10.1016/j.tig.2016.10.003.
- Zalai, Dénes, Krisztina Koczka, László Párta, Patrick Wechselberger, Tobias Klein, and Christoph Herwig. 2015. "Combining Mechanistic and Data-Driven Approaches to Gain Process Knowledge on the Control of the Metabolic Shift to Lactate Uptake in a Fed-Batch CHO Process." *Biotechnology Progress* 31 (6): 1657–68. doi:10.1002/btpr.2179.
- Zampetaki, Anna, Andreas Albrecht, and Kathleen Steinhofel. 2018. "Long Non-Coding RNA Structure and Function: Is There a Link?" *Frontiers in Physiology* 9: 1201. doi:10.3389/fphys.2018.01201.

- Zhang, Fang, Xiangming Sun, Xiaoping Yi, and Yuanxing Zhang. 2006. "Metabolic Characteristics of Recombinant Chinese Hamster Ovary Cells Expressing Glutamine Synthetase in Presence and Absence of Glutamine." *Cytotechnology* 51 (1): 21–28. doi:10.1007/s10616-006-9010-y.
- Zhang, Hu, Weixiang Wang, Chunshan Quan, and Shengdi Fan. 2010. "Engineering Considerations for Process Development in Mammalian Cell Cultivation," 103–12.
- Zhang, Jian, Shiyao Sui, Hao Wu, Jinfeng Zhang, Xingda Zhang, Shouping Xu, and Da Pang. 2019a. "The Transcriptional Landscape of lncRNAs Reveals the Oncogenic Function of LINC00511 in ER-Negative Breast Cancer." *Cell Death and Disease* 10 (8). Springer US. doi:10.1038/s41419-019-1835-3.
- Zhang, Jie, Pin Liu, Junyan Tao, Pan Wang, Yi Zhang, Xinhua Song, Li Che, et al. 2019. "TEA Domain Transcription Factor 4 Is the Major Mediator of Yes-Associated Protein Oncogenic Activity in Mouse and Human Hepatoblastoma." *American Journal of Pathology* 189 (5). American Society for Investigative Pathology: 1077–90. doi:10.1016/j.ajpath.2019.01.016.
- Zhang, Kailiang, Xiaotian Sun, Xuan Zhou, Lei Han, Luyue Chen, Zhendong Shi, Anling Zhang, et al. 2015. "Long Non-Coding RNA HOTAIR Promotes Glioblastoma Cell Cycle Progression in an EZH2 Dependent Manner." *Oncotarget* 6 (1): 537–46. doi:10.18632/oncotarget.2681.
- Zhang, Song, and Caifeng Xie. 2017. "The Role of OXCT1 in the Pathogenesis of Cancer as a Rate-Limiting Enzyme of Ketone Body Metabolism." *Life Sciences* 183 (2016). Elsevier Inc: 110–15. doi:10.1016/j.lfs.2017.07.003.
- Zhao, Shanrong, Ying Zhang, William Gordon, Jie Quan, Hualin Xi, Sarah Du, David von Schack, and Baohong Zhang. 2015. "Comparison of Stranded and Non-Stranded RNA-Seq Transcriptome Profiling and Investigation of Gene Overlap." *BMC Genomics* 16 (1). BMC Genomics: 675. doi:10.1186/s12864-015-1876-7.
- Zhou, Keda, Guillaume Gaullier, and Karolin Luger. 2019a. "Nucleosome Structure and Dynamics Are Coming of Age." *Nature Structural and Molecular Biology* 26 (1). Springer US: 3–13. doi:10.1038/s41594-018-0166-x.
- Zhu, Xinshou, Christoph Buhner, and Sven Wellmann. 2016. "Cold-Inducible Proteins CIRP and RBM3 , a Unique Couple with Activities Far beyond the Cold." *Sciences, Molecular Life*, 3839–59. doi:10.1007/s00018-016-2253-7.
- Zhu, Zhonglin, Zhilong Yu, Zeyin Rong, Zai Luo, Jing Zhang, Zhengjun Qiu, and Chen Huang. 2019. "Theranostics The Novel GINS4 Axis Promotes Gastric Cancer Growth and Progression by Activating Rac1 and CDC42" 9 (26). doi:10.7150/thno.36256.
- Ziegler, Christian, and Markus Kretz. 2017. "The More the Merrier-Complexity in Long Non-Coding RNA Loci." *Frontiers in Endocrinology* 8 (APR): 1–6. doi:10.3389/fendo.2017.00090.
- Zucchelli, Silvia, Francesca Fasolo, Roberta Russo, Laura Cimatti, Laura Patrucco, Hazuki Takahashi, Michael H Jones, et al. 2015. "SINEUPs Are Modular Antisense Long Non-Coding RNAs That Increase Synthesis of Target Proteins in Cells." *Frontiers in Cellular Neuroscience* 9 (May): 174. doi:10.3389/fncel.2015.00174.