Integrated multiomic characterisation of the CHO cell transcriptome and translatome using next generation sequencing

A thesis submitted for the degree of Ph.D

Dublin City University

By

Craig Monger, M.Sc., B.Sc. (Hons)

The research work described in this thesis was performed under the supervision of
Dr. Paula Meleady and Dr. Colin Clarke (NIBRT, Dublin, Ireland)

National Institute for Cellular Biotechnology

School of Biotechnology

Dublin City University

February 2019

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, and 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.:** 14213049

Date: ________________________

This thesis is dedicated to my loving wife,

Zoe Thorp

Acknowledgements

First of all a huge thank you to my supervisor Dr. Colin Clarke for all the work you have put into moulding me into a PhD candidate over the past 4 years! I've enjoyed the time spent carrying out research in your lab and cannot thank you enough for giving me the opportunity to work on this project. It has been an honour and a privilege to be the first of hopefully a long line of students to carry out their research under your wing. Your guidance throughout all aspects of this work has been invaluable. Most of all I have to thank you for the time spent reading this thesis and providing feedback even when away on leave!

I would like to thank Professor Martin Clynes, Professor Niall Barron and Dr Paula Meleady for the supervision you have provided me over the past 4 years. Thank you for welcoming me into the CHO lab group in NICB, assisting in designing this project and giving me feedback and advice throughout my PhD.

Thank you to Dr Paul Kelly, Justine Meiller, Alan Costello and Clair Gallagher for carrying out CHO cell cultures and library preparations to generate the sequencing data analysed throughout this thesis. Thank you to Orla Coleman and Michael Henry for the LC-MS/MS proteomics data generated which was utilised in this thesis and for the multiple reanalyses of the data required for all the databases I sent you! Thank you to "The best post –doc wet-lab scientist in Colins group" Dr Ioanna Tzani for carrying out additional CHO cultures, metabolic profiling and qPCR validation of splicing results. The work carried out in this thesis literally would not have been possible without your squishy wet lab work and I am forever indebted to you all.

Krishna Motheramgari, thank you for putting up with me sitting next to you over the past 3 years with headphones blaring! You have become a close friend throughout our studies and I cannot thank you enough for being there whenever I've needed to bounce ideas and results of someone! Also thank you to the newer members of our lab, Andrew Watson and Marina Castro for the encouragement and support you've given me in times of need!

To all my colleagues who have made NIBRT a great environment to work, there are too many of you to name individually but a particular thanks to the bakers who have provided delicious treats, the facilities team for keeping the place open and running, and the coffee break gang including Adam, Flo, Craig, Geo, Sinead, Josh, Gordon, Roisin and Sarah for keeping me entertained and caffeinated over the past 4 years.

Thank you to my Godfather and science teacher Alan Dival, who sadly passed during this research, and my inspiring teacher Angela Curtis who instilled an interest in science throughout my life and led me towards a career in biology.

To my mother Tracy, my grandparents June and George and my brother and sisters. Thank you for all the encouragement you have given me throughout my life and this research. You have been extremely supportive with my move to Cardiff and then to Dublin to carry out my studies and I couldn’t have done so without your help.

Finally, a huge thank you to my recently wed wife Zoe Thorp. Thank you from the bottom of my heart for all you have done for me and being so understanding while I have been unable to spend the time with you that you deserve during this research. You have gone above and beyond supporting me and doing more than your fair share of the chores and I look forward to spending the rest of my life making it up to you.

Table of contents

Abbreviations

3’UTR – 3 prime untranslated region

A site – Aminoacyle site

BH – Benjamini Hochberg

BHK – Baby Hamster Kidney

cDNA – Complementary DNA

CDS – Coding sequence

CHO – Chinese Hamster Ovary

CLIP-Seq - Cross-linking and immunoprecipitation sequencing

CPM – Counts per million mapped reads

CT – Cycle threshold

ddNTP – dideoxynucleotides

DE – Differentially expressed

dsRNA – double stranded RNA

E site – Exit site

EST – Expressed Sequence Tag

FBS – Foetal Bovine Serum

FC – Fold change

FPKM - Fragments per kilobase of transcript per million mapped reads

HEK – Human Embryonic Kidney

IgG – Immunoglobulin G

LC-MS/MS - liquid chromatography-mass spectrometry

m6A - N6-methyladenosine

mAbs – Monoclonal Antibodies

miRNAs – micro RNAs

mTOR – Mammalian target of rapamycin complex

NMD – Non-sense mediated decay

NGS – Next Generation Sequencing

Neu5Gc - N-glycolylneuraminic acid

NTS – Non-temperature shifted

OXPHOS – Oxidative Phosphorylation

P site – Peptidyl site

PCA – Principal component analysis

piRNA – Piwi-interacting RNA

PSI – Percent spliced in

PTM – Post translational modifications

RISC - RNA-induced silencing complex

RNA-Seq - High-throughput sequencing of transcripts

RPF - Ribosome protected footprint

rRNA - Ribosomal RNA

siRNA – Small interfering RNA

snRNA – Small nuclear RNA

snoRNA – Small nucleolar RNA

TE – Translation efficiency

TS – Temperature shifted

tsRNA – tRNA-derived small RNA

Tables

Figures

Abstract

Craig Monger MSc. BSc. : Integrated multiomic characterisation of the CHO cell transcriptome and translatome using next generation sequencing

Chinese hamster ovary (CHO) cells are the predominant mammalian cell line for the production of recombinant therapeutic proteins. Following the publication of the CHO cell genome and with the availability of next generation sequencing (NGS) technologies there is a significant opportunity for the field to increase the performance of cellular factories for biopharmaceutical production. Despite the tremendous promise of rational cell line development, genetic engineering approaches to improve manufacturing performance have yet to be routinely adopted across the biopharmaceutical industry. It is imperative that the field continues to improve the annotation of the CHO cell genome by characterising transcript populations and deepening our understanding of the influence of post-transcriptional gene expression control.

In this thesis, a unique high resolution multiomics dataset comprised of RNA expression data as well as ribosome footprint profiling (RiboSeq) and proteomics was used to study global transcriptional and translational processes in CHO cells. The first stage in this analysis was to characterise the expressed splice variants including the assembly of novel transcripts and the identification of differential transcript expression. The influence of post-transcriptional control of gene expression was determined utilising RiboSeq, revealing a significant number of proteins undergoing differential translation. Finally, non-miRNA-mediated translational repression was elucidated by integrating results of small RNA sequencing and differential expression analysis. In summary, the work undertaken in this PhD has significantly improved completeness of genome annotation, characterisation of alternative splicing and improved our understanding of post-transcriptional gene expression control in CHO cells.

Publications

“Towards next generation CHO cell biology: Bioinformatics methods for RNA-Seq based expression profiling” (2015) Monger, C., Kelly, P., Gallagher, C., Clynes, M., Barron, N., Clarke, C. *Biotechnol. J.* **7:** pp.950-66 doi 10.1002/biot.201500107

“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

“Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in Chinese hamster ovary cells” (2017) Kelly, P., Clarke, C., Costello, A., Monger, C., Meiller, J., Dhiman, H., Borth, N., Betenbaugh, MJ., Clynes, M., Barron, N. *Metab Eng.* **41:** pp.11-22 doi: 10.1016/j.ymben.2017.02.001

“Understanding biopharmaceutical production at single nucleotide resolution using ribosome footprint profiling.” (2018) Tzani, I., Monger, C., Kelly, P., Barron, N., Kelly, RM., Clarke, C. *Curr Opin Biotechnol.* **53:** pp. 182-190 doi: 10.1016/j.copbio.2018.01.030.

“High-throughput RNA-Sequencing reveals extensive alternative splicing induced by temperature shift in Chinese hamster ovary cells.” - Manuscript in preparation

“Parallel application of multi-omics techniques reveals extensive post-transcriptional gene regulation regulating the CHO cell response to mild hypothermia” - Manuscript in preparation

Talks and Posters

“High resolution RNA-Seq reveals widespread alternative splicing induced by temperature shift in CHO cells.” 8th Irish Next Generation Sequencing Meeting- Awarded the runners up prize from Roche for best talk

“Next generation RNA Sequencing reveals extensive alternative splicing of the CHO cell transcriptome following temperature shift” - ESACT 2017

“Towards next generation CHO cell biology: Bioinformatics methods for RNA-Seq-based expression profiling” Opportunities in Biopharma Research 2015- Awarded SFI prize for best poster

Chapter 1

Literature Review

1.1 - The biopharmaceutical industry

Since the approval of the first biopharmaceutical product in 1982, human insulin for the treatment of diabetes, the industry has rapidly expanded and as of 2014 there are 166 products on the market (Walsh, 2014). The increase in the number of different products available has been paired with rapid growth in the profits from the sale of these products. Annual sales have increased over 20 fold in a 20 year period from 1994 to 2014, with revenues from biopharmaceutical products totalling $7.7 billion and $162 billion respectively (Figure 1.1) (Lawrence, 2005; Walsh, 2014). Growth of this industry is projected to increase at a compounding rate of 10% each year - with 2020 projections estimated at $278 billion (Persistence Market Research, 2015).

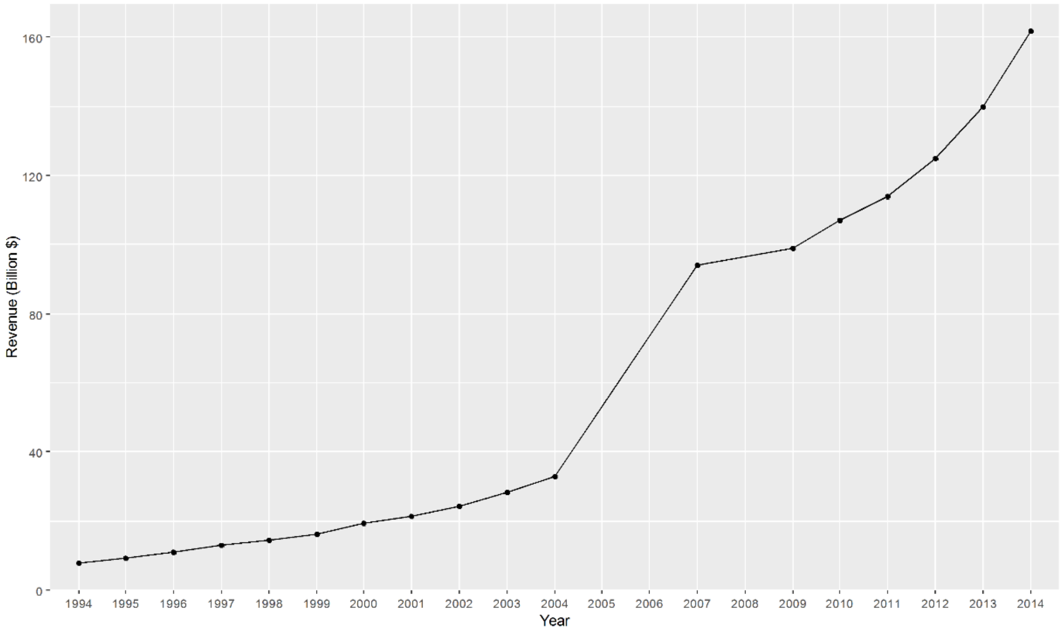

Figure 1.1– 20 years of biopharmaceutical sales growth. The sales revenue of biopharmaceutical products has increased dramatically over the past 20 years from $7.7 billion in 1994 to over $162 billion per year in 2014. Data from (Lawrence, 2005; Walsh, 2006, 2010a, 2014). Data for years 2005 and 2006 unavailable.

Much of the growth in the industry can be attributed to the production and sales of monoclonal antibodies (mAbs) and fusion proteins which account for a large proportion of the sales (54% of the 2013 sales). mAbs dominate the list of blockbuster biopharmaceutical drugs with 6 out of 10 of the highest grossing products a mAb (Table 1.1) (Walsh, 2014).

Table 1.1 – The top 10 selling biopharmaceutical products in 2013. Of the top 10 best-selling biopharmaceutical products on the market in 2013, 7 were produced in a mammalian cell line (6 of which were produced in CHO cells) (Walsh, 2014).

Product Name	**Host**	**Sales (Billion$)**	**Class**	**Indication**	**Company**
Humira	CHO cells	11	Monoclonal antibody	Autoimmune diseases including Crohn's disease, ulcerative colitis, psoriasis, ankylosing spondylitis, and arthritis	AbbVie & Eisai
Enbrel	CHO cells	8.76	Fusion Protein	Autoimmune diseases including ankylosing spondylitis, psoriasis and arthritis	Amgen, Pfizer, Takeda Pharmaceuticals
Remicade	Sp2/0 cells	8.37	Monoclonal antibody	Autoimmune diseases including Crohn's disease, ulcerative colitis, psoriasis, and arthritis	J&J, Merck & Mitsubishi Tanabe Pharma
Lantus	E. coli	7.95	Insulin analogue	Diabetes	Sanofi
Rituxan	CHO cells	7.91	Monoclonal antibody	Lymphomas, leukaemia, transplant rejection, autoimmune diseases	Biogen-IDEC, Roche
Avastin	CHO cells	6.97	Monoclonal antibody	Colon, lung, renal, ovarian and brain cancer	Roche/Genentech
Herceptin	CHO cells	6.91	Monoclonal antibody	Breast Cancer	Roche/Genentech
Neulasta	E. coli	4.39	PEGylated recombinant protein	Reduces risk during chemotherapy	Amgen
Lucentis	E. coli	4.27	Monoclonal antibody	Age related vision loss	Roche/Genentech
Epogen	CHO cells	3.35	Erythropoietin recombinant protein	Anaemia	Amgen

The most abundant class of biopharmaceutical products approved for market is mAbs, accounting for 22% of the total number of approvals (Figure 1.2). Market trends show approvals of mAbs to be increasing, with mAbs accounting for 30% of the drugs approved for market between 2010-2014 (Walsh, 2014).

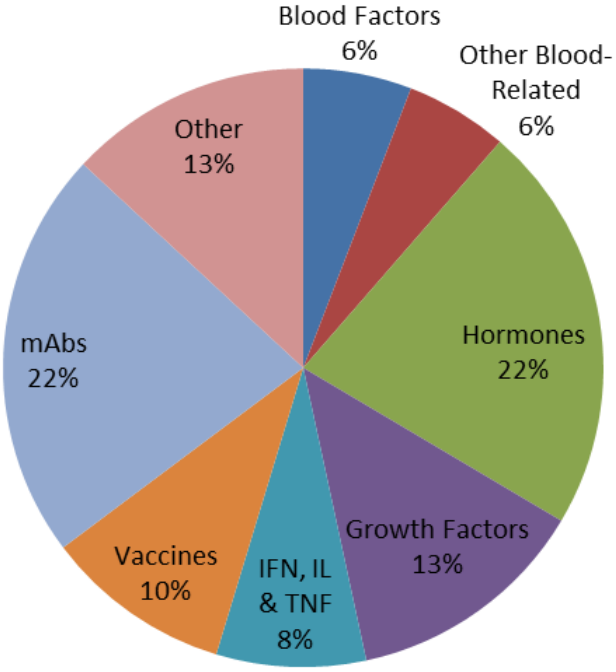

Figure 1.2 – The classes of biopharmaceuticals approved since 1982. Illustrated in this figure are the proportions of biopharmaceutical product classes approved for market since 1982 as of 2013. mAbs and hormones are the most abundant type of biopharmaceutical approved. Data from (Walsh, 2014).

Although mAbs are crucial for the treatment of many life threatening or debilitating diseases, they are among the most expensive drugs in the world. The most expensive mAb, eculizumab (sold with the product name Soliris), costs over $400,000 per patient per year and the average cost of the top 10 selling biologics is approximately $200,000 (Shaughnessy, 2012). High drug prices are largely contributed to by costly research and development processes and a high failure rate of products in clinical trials as approximately 86% of drugs which reach phase 1 of trials fail to reach the market (Wong et al., 2018). Further contributing to the price of biopharmaceuticals is the cost of developing production pipelines for new products – the cost of this stage of manufacturing could be decreased if this process is optimised by using more efficient expression hosts with rapid growth and productivity.

1.2 – Host cell lines for biopharmaceutical production

Several host organisms are used for the production of recombinant therapeutic proteins and vary in their biological complexity from simple prokaryotic organisms to mammalian cell lines. Different classes of biopharmaceuticals require different degrees of post-translational modification (PTM) to produce a safe and active product. For example, disulphide bonds may be required for protein stability or glycosylation to aid protein folding and immunoglobulin G (IgG) antibody binding affinity (Daëron and Nimmerjahn, 2014; Walsh, 2010b). The selection of an appropriate host organism is therefore important to ensure the safety and functionality of the product (Balbás and Lorence, 2004). The best host is typically the host cell line of the lowest evolutionary complexity capable of producing a product, as less complex organisms such as bacteria and single celled eukaryotes have rapid growth rates and produce large amounts of recombinant protein product (Balbás and Lorence, 2004). Figure 1.3 contains a breakdown of the organisms used to produce approved biotherapeutics.

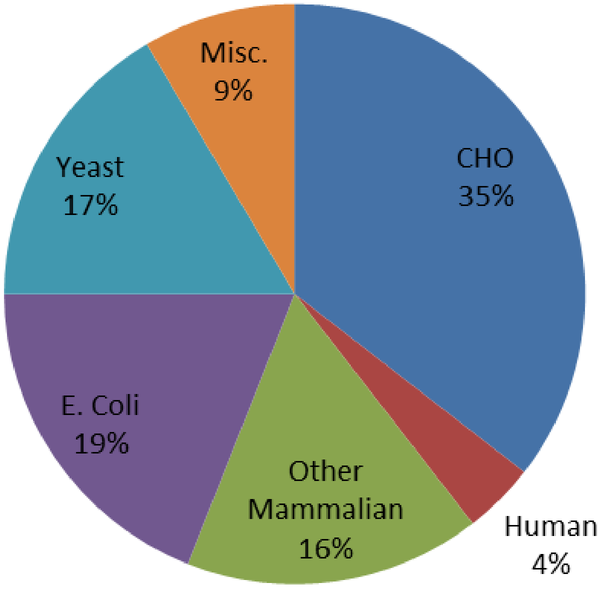

Figure 1.3 – The prevalence of expression systems for production of recombinant therapeutic proteins. Illustrated in this figure is the percentage of approved biopharmaceutical products for host cell type. The most popular expression systems are mammalian cells, making up 55% of the total – of which 35% comes from CHO cells. Non-CHO cell mammalian cell lines include hybridomas, baby hamster kidney (BHK21), mouse Sp2/0, mouse C127 cells, and mouse NSO cells. Human cell lines include human embryonic kidney (HEK), HT-108-0 and PER.C6 cells (Dumont et al., 2016).

1.2.1 - Production of recombinant therapeutics in bacterial cells

In the past bacterial cells have been the preferred choice of expression host for the production of biopharmaceuticals due to their rapid growth rate in large scale cultures and well developed genome editing techniques which allow the efficient expression of recombinant proteins (Balbás and Lorence, 2004). *Escherichia coli* is a popular choice of host expression system for recombinant therapeutics due to doubling times of 20 minutes and the high cell densities that can be achieved (Rosano and Ceccarelli, 2014). *E. coli* cell lines are used for the production of 19% of all recombinant therapeutic proteins (Walsh, 2014) including insulin, glucagon and parathyroid hormones for the treatment of diabetes, hypoglycaemia and osteoporosis, as well as interferons for treatment of cancer, hepatitis, genital warts, leukemia and multiple sclerosis (Baeshen et al., 2015). Two other bacterial organisms are used for the production of recombinant proteins: *Vibrio cholera* for the production of the cholera vaccine (Dukoral) and *Bordetella pertussis* for the diphtheria, tetanus and pertussis triple vaccine (Triacelluvax) (Walsh, 2014).

1.2.2 - Production of recombinant therapeutics in yeast cells

The use of yeast cells for the production of recombinant proteins has a number of advantages over bacterial cells. Yeast cells are eukaryotic and therefore have organelles which provide an environment for recombinant proteins to correctly fold and can produce proteins with post-translational modifications such as removal of signal peptides, formation of disulphide bonds, acylation and simple glycosylation (Nielsen, 2013). Although single celled eukaryotic yeast cells are incapable of producing more complex mammalian glycosylation patterns, the use of yeast cells can be advantageous for producing simple biotherapeutic products due to rapid proliferation and high product yields (Lalonde and Durocher, 2017). The budding yeast *Saccharomyces cerevisiae* is the most widely used yeast cell line for the production of biopharmaceuticals (Kim et al., 2015b). Biological therapeutics produced in yeast include the hepatitis B surface antigen, human papillomavirus vaccine, as well as human serum albumin for the treatment of liver disease (Nielsen, 2013). Although bacterial and yeast cells are capable of rapid growth and high production levels of recombinant protein, they cannot produce the complex post-translational modifications such as mammalian glycosylation and therefore a cell line derived from a higher eukaryote is often selected for manufacturing (Walsh, 2010b). The correct glycosylation of peptides in mAbs is required for the product to be safe, stable and efficacious (Zheng et al., 2011b). Without the correct glycosylation pattern stability can be impacted with the mAb being more susceptible to unfolding, proteolytic cleavage and aggregation (Zheng et al., 2011b).

1.2.3 - Production of recombinant therapeutics in plant cells

Plant cell lines have been used as an expression host for over 30 viral antigens for use in vaccines against viruses such as hepatitis B, rotavirus, smallpox, HIV, avian flu, papillomavirus and swine fever which have been shown to confer protection to the target virus if administered as an oral vaccine. However none of these vaccines have received market approval for humans (Laere et al., 2016). The majority of these vaccines are produced in the tobacco plant *Nicotiana tobacum*. The use of plant expression systems enables lower production costs as plants can be grown in grow rooms at the site of disease outbreak wherever required instead of in bioreactors which need large production facilities, growth media and expensive equipment (Waheed et al., 2016). In addition, less downstream processing is required as products can be administered orally in the form of an edible plant instead of purified and injected intravenously (Daniell et al., 2009). Plants also cannot be affected by many diseases which affect mammals and can process proteins requiring post-translational modifications including disulphide bonds and glycosylation if targeted to the endoplasmic reticulum with a signal peptide (Daniell et al., 2009).

Although plant-based vaccines have yet to be approved for human use, plants have been used to produce recombinant therapeutic proteins. The first recombinant therapeutic protein produced in a plant was approved in 2016 – the recombinant glucocerebrosidase enzyme (market name Eleyso) produced in a carrot root cell line. Eleyso is used for the treatment of a lysosomal storage disorder called Gaucher disease (Lee and Ko, 2017) and is efficiently targeted to macrophages (the cell type affected in Gaucher disease) due to an immune response to the plant-like terminal mannose glycosylation (Walsh, 2014). Perhaps the most notable utilisation of plants as biopharmaceutical factories in recent years is for the production of Zmapp, a product containing 3 mAbs targeting the Ebola surface protein to treat haemorrhagic fever which is produced in the tobacco species *Nicotiana benthamiana* (Budzianowski, 2015).

1.2.4 - Production of recombinant therapeutics in insect cell lines

Three recombinant proteins are produced in insect cell lines: a vaccine for human papilloma virus (a cause of cervical cancer) called Cervarix, which is produced in high-five cells, as well as the influenza vaccine Flublok and the fusion protein Provenge for treatment of prostate cancer which are produced in Sf9 cells (Walsh, 2014). The use of insect cells provides several production advantages such as growth in serum and protein free media containing high concentrations of amino acids and glucose. Insect cells can utilise amino acids as an energy source and produce less metabolic by-products such as lactate, which they are also highly tolerant to compared to mammalian cells (Ikonomou et al., 2003). Finally, insect cells can be used for the production of gene therapies in baculovirus expression vectors (Airenne et al., 2013). Baculoviruses can be engineered to harbour large gene inserts and enter mammalian cells *in vivo* to deliver gene therapies while being unable to replicate in mammalian cells and without triggering an immune response (Airenne et al., 2013). Baculovirus based gene therapies have not been used in clinical trials but pre-clinical trials have shown they are promising candidates for gene therapies to deliver vaccinations, for regenerative medicine and cancer treatments (Kwang et al., 2016).

1.2.5 - Production of recombinant therapeutics in mammalian cell lines

There are a variety of mammalian cell lines used for the production of complex recombinant proteins including baby hamster kidney (BHK) cells, the mouse derived NS0 and SP2/0 cell lines, human embryonic kidney (HEK) cells and Chinese hamster ovary (CHO) cells (Walsh, 2014).

BHK cells are capable of growth in long-term fermentations and can be maintained for up to 6 months in perfusion cultures (i.e. antihemophilic factor VIII manufacture (Wurm, 2004)). NS0 cell lines are derived from murine myeloma cells – a cancer of white blood cells which naturally secretes antibodies and are not part of a tissue and therefore can be cultivated in suspension without adaption (Wurm, 2004). NS0 cells are used for the production of Arzerra and Soliris, mAbs administered for the treatment of chronic lymphocytic leukemia and atypical haemolytic uremic syndrome (Geigert, 2014) respectively. SP2/0 cells are mouse hybridoma cells produced from the fusion of a spleen and myeloma cell which are used for the production of mAbs including Erbitux for the treatment of colorectal cancer and Stelara for the treatment of the auto immune disease plaque psoriasis (Geigert, 2014).

Human cell lines such as HEK cells are used for the manufacture of 4% of biopharmaceuticals (Walsh, 2014). Human cell lines are typically used to produce complex glycoproteins which require extensive post-translational modification such as rFVIIIFc (trade name Alprolix) used to treat haemophilia A (Lalonde and Durocher, 2017). rFVIIIFc is an FC fusion protein which combines the domains of the recombinant factor VIII clotting protein to the FC domain of IgG to reduce immunogenicity of the rFVIII treatment - increasing the half-life of the protein and preventing tolerance to treatment (Krishnamoorthy et al., 2016). rFVIIIFc requires sulfation at 6 tyrosine residues in order to be functional – complete sulfation of these residues cannot be efficiently achieved when rFVIIIFc is produced in non-human cell lines (Kannicht et al., 2013).

1.3 - Chinese hamster ovary cells

1.3.1 – The Chinese hamster ovary cell lineage

The Chinese hamster ovary (CHO) cell line was derived from the ovary tissue of an outbred Chinese hamster in 1957 (Wurm, 2013). The cells from the tissue sample were treated with trypsin and cultured for 10 months during which the cells spontaneously become immortalised as a fibroblast-like culture. The parental CHO cell line grew quickly and healthily as an adherent culture in foetal bovine serum (FBS) containing media and was distributed to many labs. The original cell line was proline synthesis deficient and referred to as CHO/Pro(-) (Kao and Puck, 1968). Three major lineages of CHO cell lines were derived from CHO/Pro(-) cells, CHO-K1, CHO DG44 and CHO-S (Lewis et al., 2013). All CHO cell lines used for the production of recombinant proteins are derived from one of these 3 lineages (Figure 1.4).

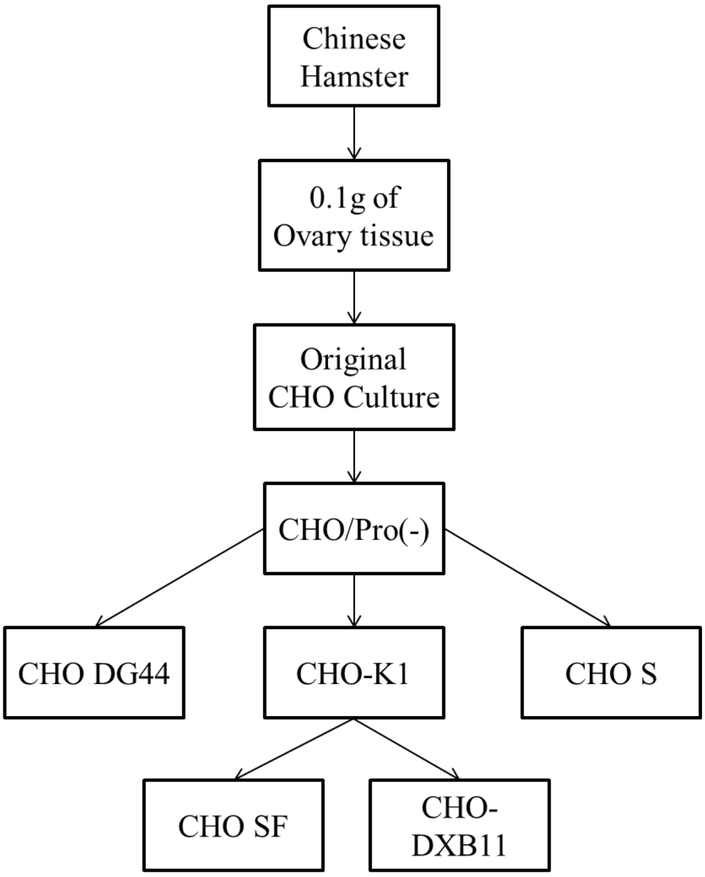

Figure 1.4 – Origin of major CHO cell lineages. The major CHO cell lineages used for the production of recombinant therapeutic proteins are all derived from the same parental clonal colony. The CHO colony became proline dependant and was sub-cloned in many laboratories around the world to create the 3 major lineages of CHO cells used in the biopharmaceutical industry, CHO DG44, CHO-K1 and CHO S.

CHO-K1 cells have been adapted to grow in serum free media (CHO SF) and the first recombinant product synthesised in CHO cells used a host derived from CHO-K1 cells, the DXB11 cell line (Wurm, 2013). DXB11 cell lines were produced by irradiating CHO-K1 cells with gamma radiation to induce mutations, causing the dual knockout of the dihydrofolate reductase (*Dhfr*) gene. The *Dhfr* gene is completely deleted in one genomic loci of DXB11 cells and contains a missense mutation in the other which enables the use of the DHFR expression system in CHO cells. Cells deficient in *Dhfr* expression cannot reduce dihydrofolic acid to tetrahydrofolic acid, a requirement for the synthesis of purine nucleic acid precursors, glycine and thymidine and therefore require their addition as media supplements to grow (Urlaub and Chasin, 1980). *Dhfr* deficiency can be exploited by transfecting *Dhfr* deficient CHO cells with an expression vector containing the *Dhfr* gene and the gene of interest (the recombinant protein) and selecting for successful transfections through cells capable of growth in nucleoside-free media (and therefore expressing the *Dhfr* gene) (Kingston et al., 2001). An inhibitor of *Dhfr* activity (methotrexate) is then added to the culture in increasing concentrations over a period which may last months to obtain cell lines which have amplified the *Dhfr* gene and the gene of interest to levels of up to 1000 copies per cell (Kingston et al., 2001). It is possible for DXB11 cells to regain *Dhfr* functionality by recovering from the missense mutation and therefore the DG44 cell line was produced by completely deleting both *Dhfr* alleles from CHO/Pro(-) cells (Urlaub et al., 1983).

CHO-S cells were the first cell line derived from CHO-Pro(-) capable of growth in suspension culture which enables scale up of cultures to high cell densities as cells are not limited to an adherent monolayer where density is limited by surface area (Kim et al., 2012a; Wurm, 2013). Collectively the lineages of CHO cells are used for the production of 35.5% of the total approved biopharmaceutical products including 6 of the top 10 best-selling products (Walsh, 2014).

1.3.2 – The advantages of the CHO cell line

Adaptability to culture: A variety of CHO cell lines are available which are adapted to various culture conditions such as suspension cultures - a key adaptation which enabled the industry to move from limited adherent culture scale to stirred tank reactors of up to 25,000 L (Kelley, 2009). CHO cell lines have also been adapted to growth in serum free media which does not contain FBS – a nutrient source mammalian cell lines have been historically dependant on (Wurm, 2013). Serum free CHO cell lines have been reported as early as 1977 and media has been developed which is completely chemically defined in terms of its amino acid, vitamin, salts, lipids and growth factor contents (Li et al., 2010). The presence of FBS in media is a risk factor for contamination by viruses, mycoplasma or prions such as the cause of bovine spongiform encephalopathy (commonly known as "mad cow" disease - a risk factor for contracting the human prion disease Creutzfeldt-Jakob disease) (Schröder et al., 2004; Wurm, 2013). CHO cells have been adapted to protein free media formulations by omitting the protein components of serum-free media solutions and weaning cells off proteins such as insulin (Schröder et al., 2004).

Rapid proliferation to a high cell density: The past 3 decades of cell line development strategies have selected for CHO cell lines with industrially beneficial phenotypes such as enhanced growth rate. Rapid proliferation allows for production cultures to quickly reach the maximum viable cell density once a culture has been seeded. A typical proliferating mammalian cell such as HEK cells double approximately every 24 hours (Cervera et al., 2011), however, CHO cells in exponential growth phase can divide as quickly as every 14 hours (Sitton and Srienc, 2008). Maximum viable cell densities have improved over the last 30 years, increasing the number of CHO cells in culture actively producing product. Cell densities of 2×10^6 viable cells/ml were typical in the 1980's (Wurm, 2004) and these have improved tenfold to over 2×10^7 viable cells/ml (Huang et al., 2010). Perfusion cultures can be utilised to increase density further and densities of over 1×10^8 have been reported with comparable levels of productivity to fed-batch cultures (Clincke et al., 2013).

High productivity: Productivity is the amount of recombinant protein produced by a culture. It is measured on the level of the individual cell (pg per cell per day or pg/cell/day) and at the level of final product titre (grams per litre of culture, g/L) (Kumar et al., 2007). Advances to gene amplification, expression and secretion systems in CHO cells has resulted in over a 10 fold increase in cell specific productivity from 10pg/cell/day in the 1980's (Wurm, 2004) to over 100pg/cell/day in todays optimised cell lines (Tabuchi, 2013). Final product titres have also increased from 50mg/L in 1986 (Wurm, 2004) to over 10g/L today which can be attributed to a combination of improved cell specific productivity, increased maximum viable cell densities, increased culture durations, optimised feeding strategies and downstream product purification (Kim et al., 2012b).

Manipulation of culture phase with temperature shift: CHO cells in industrial batch/fed-batch cultures have 3 distinct phases – the exponential growth phase, the stationary phase and the death phase (Pan et al., 2017). Exponential growth occurs from the onset of culture seeding when nutrients are abundant and cells rapidly divide to a high cell density while primarily consuming glucose as an energy resource using glycolysis and producing lactate as a by-product (Pan et al., 2017). During the stationary phase cells stop dividing, are larger and have a higher productivity and switch their metabolism to consume lactate and utilise the TCA cycle and oxidative phosphorylation (OXPHOS) to produce energy (Pan et al., 2017). When nutrients are depleted or metabolic by-products such as ammonia reach intolerable levels cells enter the death phase and begin apoptosis. As cells have a higher productivity in the stationary phase it is advantageous to maximise the amount of time a culture spends in this phase. CHO cells can be inherently switched from exponential growth phase to stationary phase when cell density has reached the maximum viable level by altering culture conditions (Wurm, 2004). A temperature shift is often applied to industrial cell cultures from 37°C to 30-33°C which induces a shift in metabolism and growth rates. Temperature shift also enhances longevity of batch culture (Vergara et al., 2014) and can result in increasing the viability of cells from 1 week to over 3 weeks (Wurm, 2004).

Post-translational modification: Not only are CHO cells capable of producing protein in large quantities, CHO cells are also capable of producing more complex proteins than bacteria or yeast cells such as those requiring mammalian post-translational modifications. Examples of these proteins include mAbs, bispecific antibodies and fusion proteins which require extensive post-translational modifications such as glycosylation to function correctly (Jefferis, 2009). The correct glycosylation of recombinant proteins is important to aid in the correct folding of proteins, to target cell surface receptors, prevent aggregation, maintain biological activity, stabilize proteins, maximise their *in-vivo* half-life and prevent degradation in the body (Walsh, 2010b). Mammalian cell lines such as CHO cells produce glycosylation patterns compatible with and active in humans with the exception of N-glycolylneuraminic acid (Neu5Gc), a terminal residue on mammalian N-linked glycans which is not produced by humans due to a deletion in the *CMAH* gene (Irie et al., 1998). Recombinant glycoprotein therapeutics containing Neu5Gc can illicit an immune response in humans and it has been suggested that this modification should be avoided in biotherapeutics (Ghaderi et al., 2010, 2012). CHO cells express less Neu5Gc modifications than other non-human mammalian cell lines and can be eliminated from recombinant proteins by feeding CHO cell cultures sialic acid, the human form of the terminal glycan which is incorporated into recombinant proteins by metabolic competition (Ghaderi et al., 2012).

Viral resistance: Viral contamination of cell culture can have negative impact on the production of recombinant therapeutic proteins, with viruses affecting potential product quality and yields as well as the growth and viability of cell lines (Xu et al., 2011). Viral contamination is a concern to patient health as viruses could be present in product produced by infected cell lines and be transferred to the patient. CHO cells are immune to infection from many viruses as they do not express the cell surface receptors necessary for viral entry such as the Nectin-1/HveC and herpes virus entry mediator receptors responsible for herpes simplex virus entry which makes them an attractive host for producing recombinant proteins (Xu et al., 2011). 162 of 388 known human genes related to viral infection are deleted or not expressed in the CHO genome (Xu et al., 2011). The viruses CHO cells are resistant to include the herpes simplex virus, pseudorabies virus, HIV, hepatitis B and Vaccina viruses (the cause of cow and small pox) (Xu et al., 2011).

1.4 – Trends in the use of CHO cells for the industrial production of biopharmaceutical products

1.4.1 – The drive to produce molecules in a predictably cheaper way

Despite the large growth of biopharmaceutical sales over the past 20 years, a number of challenges face the biopharmaceutical industry. The largest challenge is mounting pressure from government organisations, regulatory agencies, insurance companies and the public to increase access to biopharmaceutical therapeutics. The average cost of treating a patient with a chemically synthesized drug is $2 per day in the United States, but over $45 per day for a recombinant therapeutic protein (Walsh, 2014). Recently high cost medicines such as Orkambi for the treatment of cystic fibrosis have been rejected by the UKs National Health Service and Irelands National Centre for Pharmacoeconomics due to its price of €160,000 per patient per year.

Prices of therapies currently on the market are already being challenged due to the prices necessary to cover the costs of developing and producing these complex medicines however the complexity of recombinant proteins is ever increasing. Examples of increased complexity include antibody-drug conjugates and bispecific antibodies. Designer antibodies are able to deliver drugs to specific cells or tissues or able to target multiple cell types respectively (Beck et al., 2010). With the ever-increasing complexity of recombinant therapeutic proteins, efforts must be made to decrease the costs associated with developing and producing biopharmaceuticals. Gene therapy treatments developed in recent years have demonstrated the importance of this issue as the first approved gene therapy Alipogene tiparvovec (trade name Glybera) was priced over $1 million per treatment (Scott, 2015). Although Glybera treatment was effective and approved by the European medicines agency it has since been withdrawn from market due to a lack of demand due to the high price.

One method which is being increasingly used to reduce the costs of producing mAbs is to reduce infrastructure costs by making bioprocessing facilities more flexible and to use single use bioreactors (Hernandez, 2015). In the past the biopharmaceutical industry has focused on the production of blockbuster drugs which are used by a large number of people affected by a disease and therefore require continuous production in large quantities. Dedicated production facilities are therefore typically built for a single product to meet the demands of these molecules required. The industry is however beginning to shift production to flexible modular manufacturing facilities which are smaller in size, can rapidly switch production to different molecules and do not require large fixed infrastructure such as 25,000L stainless steel bioreactors (Jacquemart et al., 2016). A major advantage of using a flexible facility is the ability to manufacture recombinant proteins to the current demand rather than a traditional facility producing in excess if demand drops. The flexibility of these modular facilities is largely enabled by the utilisation of single-use disposable plastic bag bioreactors which can reach 2000L in volume and facilities using single-use reactors can meet the yields of traditional stainless steel bioreactor facilities if used in parallel (Jacquemart et al., 2016). Single use bioreactors also offer the advantage of requiring no cleaning therefore meaning there is less production downtime between batches, less piping required in the facility as water, steam and cleaning fluids do not need to be delivered to the bioreactor, and less money required to be spent on utilities. Flexible bioprocessing facilities will enable quick turnaround between the production of recombinant therapeutics which are required in smaller quantities such as for personalised medicine applications and treatment of orphan diseases (Sharma et al., 2010). Optimisation of the cell line vectors themselves is still however of vital importance to increase the amount of product which can be produced using single use technologies as well as to reduce the amount of time required for cell line development phases of production.

1.4.2 – Biosimilars

Many of the first recombinant therapeutic proteins produced have now lost the protection for exclusive ownership and production granted by their patents due to their expiration - opening a market for companies to produce these products and sell them at a lower cost, potentially decreasing the costs of some of the most expensive medicines. The first biosimilar mAbs were approved for market in Europe in 2013 (trade names remsima and inflectra, biosimilars for the mAb infliximab) (Walsh, 2014) and launched in 2015 when infliximab's patent protection expired. Inflectra was also recently approved as the first biosimilar mAb for the US market by the FDA (FDA, 2016).

Strict regulations are in place for the approval of biosimilar products which ensure they match the quality, safety and efficacy of the original product - meaning supplying sufficient evidence to show the products identity, structure, bioactivity and impurity levels are comparable to regulators is required. The most successful biosimilar products have also conferred an additional advantage such as ease of drug administration or patient compliance (reduced dose frequency) (Walsh, 2010a). The biosimilar market is growing steadily and $33 billion of biologics have lost patent protection (Walsh, 2014). As more mAbs lose patent protection in the coming years, there will be a market focus on producing biosimilars for these products and it is important that the characteristics of the original mAbs including amino acid sequence, protein folding, glycosylation, and absence of degradation products, aggregation and impurities are maintained to ensure immunogenicity is not a concern (Brinks, 2013).

1.4.3 – Better understanding of CHO cell biology

Over the last decade there has been an increasing interest in understanding CHO cell biology in order to enable rapid and predicable process development. At present, the cell line development to acquire optimal production phenotypes is laborious and time consuming - taking as long as 6-18 months to develop a high production stable cell line (Kuystermans and Al-Rubeai, 2015). Cell line development processes can screen thousands of sub-cloned cell lines and current strategies often incorporate trial and error approaches such as random mutagenesis followed by selection of beneficial traits to develop candidate sub clones optimised to grow quickly to a high cell density and produce high quality recombinant protein at a large quantity (Kuystermans and Al-Rubeai, 2015).

By increasing our understanding of the molecular pathways which contribute to these beneficial phenotypes it may be possible to predict growth rate and cell specific productivity of cell lines using next-generation sequencing (Clarke et al., 2011) to select faster and if appropriate modify the cell line for high performance.

Genetic modifications to cell lines to improve performance have begun to be utilised such as to prevent product aggregation (Le Fourn et al., 2014), engineer glycosylation (Chenu et al., 2003; Walsh, 2010b) and enhance CHO cell viral resistance (Haines et al., 2012). When expressed at high levels, some difficult to express recombinant proteins such as IgG's may form polypeptide aggregates inside cells in many host systems rather than being secreted due to the overburdening of the protein processing and secretion pathways (Le Fourn et al., 2014). CHO cells have been engineered to improve the secretion efficiency of therapeutic proteins and prevent this aggregation from occurring by overexpressing components of the secretion pathway such as signalling receptor proteins, allowing high cell specific productivity of mAbs in CHO cells without protein aggregation (Le Fourn et al., 2014). The *Cmah* gene encoding the enzyme which catalyses non-human Neu5Gc glycosylation has been knocked down using RNA interference to cause the degradation of *Cmah* mRNA and reduce the immunogenicity of recombinant products (Chenu et al., 2003). Other glycoengineering strategies have been used in CHO cells to produce glycosylation patterns which affect the activity of mAbs, for example the knockout of the *Fut8* gene prevents fucosylation of amino acids and increases the activity of the antibody 100 fold (Walsh, 2010b). CHO cells have also been engineered to gain resistance to the ebolavirus, removing potential viral contamination of products produced in CHO cells (Haines et al., 2012). Engineering of cell lines resistant to ebolavirus were generated by mutational screening to knock out the *Npc1* gene, which encodes a membrane protein required for ebolavirus entry (Haines et al., 2012). The examples of genetic modifications presented illustrate the potential of engineering to enhance production cell lines for enhanced productivity and product quality however in order to further rationally engineer cell lines an increased understanding of the CHO cell molecular biology is required.

Over the last decade progress has been made towards increasing our understanding of CHO cell molecular biology to inform rational cell line engineering strategies. Early investigations of the transcriptome relied on the homology between the Chinese hamster and mouse as no CHO specific tools or transcript sequences were available (Ernst et al., 2006; Yee et al., 2008). Using mouse microarrays these studies were however able to investigate the CHO cell response to heat shock (Ernst et al., 2006) and sodium butyrate treatment (Yee et al., 2008). Studies utilising mouse microarrays were limited in their ability to profile the CHO transcriptome and have been found to inaccurately report the expression with a false negative rate of 57% and a false positive rate of 12% (Melville et al., 2011). Expressed sequence tag (EST) sequencing experiments, which were used to generate cDNA libraries (collections of the sequences of reverse transcribed mRNAs) using Sanger sequencing, enabled the creation of CHO cell specific microarrays (Melville et al., 2011). The WyeHamster2a microarray was generated through combination of CHO cell sequences in public databases with proprietary sequences from Wyeth (acquired by Pfizer) to create a microarray consisting of 3,714 sequences. Although limited to approximately 20% of the protein coding genes in the CHO cell genome this array was capable of identifying widespread changes in gene expression between high producing cell lines and their parental cell line, demonstrating the application of transcriptomics techniques to understand and potentially improve cell line development strategies at a molecular level (Doolan et al., 2008). The power of transcriptomic analysis of CHO cells during cell line development were further demonstrated utilising the WyeHamster2a microarray to predict cell-specific productivity using transcriptomics measurements of 287 genes (Clarke et al., 2011). Biopharmaceutical industry interest in the application of ‘omics techniques has continued, with further sequencing projects resulting in the generation of the WyeHamster3a microarray containing probes to profile 19,809 sequences (Clarke et al., 2012). Progress towards understanding the production of biotherapeutics in CHO cells at a molecular level has accelerated since the publication of the CHO cell and Chinese hamster genome sequences (Brinkrolf et al., 2013; Lewis et al., 2013; Xu et al., 2011).

1.5 - The CHO genome

1.5.1 – Chinese hamster genome sequencing projects

Through a combination of the efforts from an intercontinental collaboration of scientists the CHO cell line genome was sequenced and published in 2011 (Xu et al., 2011). Sequencing of the CHO cell genome had been made possible due to advances to next generation sequencing (NGS) technology since the publication of the first mammalian genomes which significantly reduced the time and cost of carrying out genome sequencing experiments. The availability of the CHO cell genome sequence offers enormous potential for the understanding of CHO cell molecular biology as studies are no longer limited to reliance on homology to other model species.

The cell lines used in this experiment were clones derived from the original CHO-K1 cell line isolated in culture. Sequencing was carried out using an Illumina HiSeq 2000 instrument with genome sequencing libraries containing insert sizes varying from 200bp up to 20kb. A total of 343Gb of DNA was sequenced, representing an estimated 132X coverage of the CHO-K1 cell line genome assuming a genome size of 2.6Gb. Using the *SOAPdenovo* assembler, the reads with a short insert size (<500bp) were assembled using this *De Brujin* graph based approach to generate long contigs (Li et al., 2010). Read pairs with the largest insert sizes were then aligned to these contigs and read pair concordance data used to bridge contigs into scaffolds. The final assembly is made up of 2.45Gb of sequence in 14,122 scaffolds ranging from 200 nucleotides to 8,779,783 nucleotides in size.

In addition to the genome of CHO-K1 cells in 2011, the genome sequence of the CHO-SEAP lineage (a cell line expressing secreted alkaline phosphatase generated from CHO-DUK cells) was also published in 2011. However this genome was sequenced to a depth of just 1 fold and the lack of coverage meant assembly resulted in 3.57million contigs (Hammond et al., 2011). The lack of sufficient coverage of this genome means sequencing errors are likely to have been incorporated into this sequence and the genome is in an extremely fragmented stage which makes the use of this genome inappropriate as a reference sequence.

Although the publication of the CHO-K1 cell line genome was a significant step in generating the genome sequence of the Chinese hamster, CHO cell lines have been shown to have unstable genomes which are susceptible to chromosomal translocations and loss (Wurm, 2013). The potential for lost or rearranged genetic information therefore means a genome of a given cell line may not be optimal for use as a reference when studying other CHO cell lines (Borth, 2014; Lewis et al., 2013). In response to this the genome of the parental Chinese hamster strain from which the CHO lineage was derived as well as a further 6 CHO cell lines have been sequenced which were simultaneously published in 2013 by 2 groups (Brinkrolf et al., 2013; Lewis et al., 2013). While both research groups published the sequence of the parental Chinese hamster they used distinct methodologies to sequence the genome.

Lewis et al. followed the approach used by Xu et al. to sequence the CHO-K1 cell line genome, generating a 2.4Gb genome by assembling 347.5Gb of data sequenced on an Illumina HiSeq 2000 using *SOAPdenovo*. The Lewis et al. Chinese hamster genome consists of 286,619 scaffolds ranging in size from 201-8,324,132 nucleotides in size, of which 6,356 are over 2kb in length and 1,091 scaffolds contain 90% of the genome (Lewis et al., 2013). The latter group utilised methodology in which the genomic sequence was first sorted into chromosomes by size separation with flow cytometry. Library preparations were carried out for each individual chromosome separately prior to sequencing on an Illumina Genome Analyzer IIx. The genome assembly software *ALLPATHS-LG* was used to assemble the ~200Gb of raw sequence data into a 2.33Gb genome in 28,764 scaffolds ranging in size from 830-14,658,418 nucleotides (Brinkrolf et al., 2013; Gnerre et al., 2011).

In addition to the Chinese hamster genome, the Lewis group also sequenced the genomes of six CHO cell lines – C0101 (a recombinant protein producing suspension cell line), a CHO protein free suspension cell line, adherent CHO-K1 cells, CHO-S cells, CHO-K1/SF serum free adherent cells, and DG44 *Dhfr*- mutant suspension cells. More recently, the CHO DXB11 (DUKX) cell line was sequenced using an Illumina HiSeq 2000 in order to study copy number variations and SNPs between CHO cell lines and the parental Chinese hamster (Kaas et al., 2015). None of these CHO cell line genomes were however made publicly available in an assembled and annotated form.

Despite 10 genome sequences of the Chinese hamster and the derived CHO cell lines being published in the literature, the genomes contain less than 2.4Gb of sequence data assembled into thousands of contigs. The Chinese hamster genome is predicted to have a genome size of 2.8Gb and therefore ~400Mb of sequence data is unsequenced. In order to close the gaps between contigs and sequence these hard to sequence regions of the genome, community funded third generation sequencing from Pacific Biosciences is underway to improve the genome (Borth, 2014). In 2017 the genome sequence of the Horizon Discovery CHO-K1GS cell line was released by Ensembl (Zerbino et al., 2018). The Ensembl assembly of the Horizon cell line is assembled to a greater standard in terms of the number of chromosome scaffolds of which there are 8,265 however this genome sequence contains 50mb less genomic data than the Xu et al. CHO-K1 genome. The genome sequences discussed in this section are summarised in Table 1.2.

Table 1.2 – The genomic sequences of the Chinese hamster. To date, 11 Chinese hamster or CHO cell line genome sequences have been released into the public domain. The best assembled in terms of scaffolds is the CHO-K1GS genome while the most complete in terms of genome length is the original CHO-K1 reference sequence. The genome assemblies of some of these cell lines were not released into the public domain and therefore unavailable data is denoted with an 'NA'.

Cell Line	**Scaffolds**	**Genome length (Mb)**	**Reference**
CHO-K1	109,152	2,399.79	(Xu et al., 2011)
C. griseus 17A/GY	28,749	2,332.77	(Brinkrolf et al., 2013)
C. griseus	52,710	2,360.13	(Lewis et al., 2013)
C0101 (antibody producing CHO-S derived)	N/A	N/A	(Lewis et al., 2013)
CHO Protein free (ECACC 85051005)	N/A	N/A	(Lewis et al., 2013)
CHO-s cGMP banked (A1136401)	N/A	N/A	(Lewis et al., 2013)
CHO-K1/SF (EXACC 93061607)	N/A	N/A	(Lewis et al., 2013)
DG44 cGMP banked (A1097101)	N/A	N/A	(Lewis et al., 2013)
CHO DXB11	N/A	N/A	(Kaas et al., 2015)
CHO SEAP	3,570,000	N/A	(Hammond et al., 2011)
CHO-K1GS	8,295	2,358.17	(Zerbino et al., 2018)

1.5.2 - Annotation of CHO genomes

Alongside each publication of the genome, gene predictions have been calculated using homology to other species, algorithms for the detection of conserved domains such as *HMMER* (Eddy, 2011) and by aligning sequenced mRNA transcripts to the genome. The initial publication of the CHO cell line genome in the 2011 contained 24,383 gene predictions of which 19,711 had homologs in human, 20,612 in mouse and 21,229 in rat. 83% of predicted CHO-K1 cell protein coding genes were assigned an annotated function based on their homology to human, mouse and rat transcripts. Despite these published findings, the annotation for the CHO-K1 cell line genome was not made publicly available to the community in any format other than raw sequence reads (Xu et al., 2011). Raw sequence reads were later submitted to the NCBI assembly and annotation pipeline RefSeq where 28,978 genes and 35,628 transcripts were predicted (RefSeq 2018).

The release of the Chinese hamster genome in 2013 was annotated with 24,044 genes using a variety of methods for *de novo* gene annotation – *AUGUSTUS*, *genscan* & *glimmerHMM* along with an accompanying Trinity RNA-Seq transcriptome assembly. 82% of genes were functionally annotated using homology based approaches to proteins in databases such as Swiss-Prot (Lewis et al., 2013). Submission of this sequence to NCBI RefSeq (RefSeq assembly accession GCA_000419365.1, 2018) for annotation resulted in 28,446 gene predictions encoding 33,465 transcripts by the NCBI RefSeq annotation pipeline.

The publication of the Chinese hamster genome using a chromosome sorting approach was not accompanied by any gene or transcriptome level annotation, however, the research group responsible for the publication released a public CHO cell line transcript database containing 65,561 transcripts from 17,598 genes (Rupp et al., 2014). Annotation for this genome has since been deposited in NCBI RefSeq, containing 21,779 genes (of which 14,576 are unique) and 29,144 proteins (RefSeq2018).

The Horizon CHO-K1GS genome assembly was released by Ensembl (Zerbino et al., 2018) as a full Ensembl genome build with genes annotated with standardised gene accession identifiers which can be readily converted to orthologous species in the Ensembl database. 25,072 genes and 32,575 transcripts are present in the Ensembl Horizon CHO-K1GS genome annotation and are available in the GTF annotation format.

Ensembl also reannotated the original CHO-K1 genome assembly (Xu et al., 2011) using publicly available sequence data in the European Nucleotide Archive (Zerbino et al., 2018). The Ensembl reannotation of the CHO-K1 assembly contains 26,668 genes and 34,472 transcripts – more than the Horizon CHO-K1GS annotation. While the Ensembl release has less annotations than the NCBI annotation of the CHO-K1 genome the Ensembl version has more uniquely annotated genes and the advantage of use in the Ensembl Biomart tool to find orthologous genes. The genome annotations presented in this section are summarised in Table 1.3.

Table 1.3 - The annotations of the Chinese hamster and CHO genomes. Presented in this table are the most recent GTF/GFF annotations for CHO and Chinese hamster genomes available. The CriGri1 (NCBI) genome contains the most genes and transcripts. The CHO-K1GS genome however has the most uniquely annotated genes. The Cgr1.0 and PICR genomes do not have gene biotype or gene name/identifier information and therefore incomplete information is denoted N/A.

Genome	Genes	Transcripts	Exons	ncRNAs	Gene Names	Unique Genes
CriGri1(NCBI)	28,978	35,628	425,408	5,738	15,695	14,149
C_griseus_v1.0 (NCBI)	28,446	33,465	390,848	5,725	15,979	14,116
Cgr1.0 (Sorted chromosomes, NCBI)	21,779	29,144	327,934	N/A	N/A	N/A
CHO-K1GS (Ensembl)	25,072	32,575	301,692	4,248	20,020	16,722
CriGri1 (Ensembl)	26,668	34,472	291,060	7,487	18,340	15,067
PICR	24,686	24,948	185,943	N/A	N/A	N/A

1.5.3 - Comparison of CHO cell genome assembly and annotation to model organisms

The Chinese hamster genome releases are made up of tens/hundreds of thousands of scaffolds however the Chinese hamster has 12 pairs of chromosomes, indicating the assembly is fragmented with a large number of gaps. The gaps in the assembly sequence are a result of limitations of second generation sequencing technologies which are unable to efficiently sequence heterochromatin, repeating microsatellites, transposable elements and duplicated regions of a genome. The large number of scaffolds highlight how the Chinese hamster genomes are currently in a relatively early draft stage and work must be done to achieve a reference genome standard. Third generation sequencing methods achieving long reads with single molecule real time sequencing (SMRT, or PacBio) can be utilised in combination with the high throughput, high quality sequence data produced by Illumina sequencing instruments to achieve this (Pendleton et al., 2015).

Although there are 5 CHO cell line/Chinese hamster genomes with their own annotation available, none of these annotations contain as many genes as are annotated for model organisms such as the mouse and rat, or that of the human genome. For example, the human genome annotation is currently annotated with 59,644 gene features and the mouse 46,062, approximately 2 times the amount annotated in the CHO cell line genomes discussed in the previous section (NCBI genome ID 51 & 52). A large discrepancy is also seen at the transcript level, with 160,474 annotated transcripts in the human genome (of which 113,620 are mRNA transcripts) and 108,126 transcripts (76,203 mRNA) in the mouse, compared to 41,318 transcripts (34,912 mRNA) in the Chinese hamster (NCBI Genome database 2018). Assuming an approximately equal number of transcripts encoded in the Chinese hamster and the mouse genomes (which are both rodents), these figures suggest the annotation for the Chinese hamster is incomplete. A large percentage of the unannotated features in CHO cells are likely to be lncRNAs, an area of research which has not been extensively studied in CHO cells to date, as there are 27,119 of this RNA species confirmed in the human genome and 21,616 in the mouse while just 4,974 are annotated in CHO cell lines. The genome annotation statistics presented in this section are summarised in Table 1.4.

Table 1.4 – Comparison of NCBI reference genome annotations of Human, Mouse and Chinese hamster. The genome sequences of the most well studied mammalian organisms (human and mouse) have considerably more transcripts annotated than the Chinese hamster. Annotation information from NCBI Genome database (2018).

Genome	Proteins/mRNAs	ncRNAs	Genes	Pseudogenes	lncRNAs
Human (Release 109)	113,620	46,854	59,644	16,152	27,119
Mouse (Release 106)	76,203	31,923	46,062	9,575	21,616
Chinese hamster (Release 102)	34,912	6,406	28,438	4,071	4,974

1.5.4 - How publication of the genome has improved our understanding of the molecular biology of CHO cells

Glycosylation: Ensuring recombinant therapeutic proteins have the correct glycosylation patterns is vital to ensure the safety, stability and activity of biological drugs. The availability of the CHO cell line genome enabled an assessment of the genes for glycosylation pathways which are found within the CHO cell genome. Findings from this study showed that CHO-K1 cells have 297 of the 300 known glycosylation genes in its genome, suggesting that glycoproteins produced in CHO cells will be of high quality (Xu et al., 2011). However, the authors also showed that only half of these genes are expressed in CHO-K1 cells during exponential growth. The unexpressed genes included sulfotransferases, fucosyltransferases and N-acetylgalactosamine transferases. A particularly noteworthy finding of this study is that although present in the genome, the *Mgat3* gene is not expressed in the CHO-K1 cell line. Mgat3 dependant glycosylation modification is found on 10% of human IgG antibodies and therefore engineering a gain-of-function mutation directed to this gene would improve the cell line to produce humanised monoclonal antibodies and glycoproteins.

Viral susceptibility: An advantageous trait of using the CHO-K1 cell line is that they are not susceptible to many viral infections which could contaminate a culture, posing a risk to the health of the cell line or humans taking the therapeutics produced by it. By utilising homology-based approaches comparing the sequence of human viral susceptibility genes to the CHO cell line genome, 4 genes necessary for viral entry had no homologue in the CHO cell genome. A further 158 genes of the 384 identified viral susceptibility genes were not expressed in the CHO-K1 cell line. Many of these genes prevent infection by stopping viral entry such as the Herpes Virus entry Mediator (HVeM) or the CD4 protein (used by the HIV virus for entry) (Xu et al., 2011). The sequence of these genes revealed to be expressed in this study could be targeted by genetic knockdown approaches to further the immunity of CHO cell lines to viral susceptibility.

Apoptosis: Apoptosis is a factor in determining the maximum duration of culturing CHO cells as preventing apoptosis results in cells which live longer and can therefore produce recombinant proteins for longer. Genomics studies identified 57 of the 63 known apoptosis genes to be present in the CHO cell line genome. By upregulating anti-apoptosis genes and knocking down/out pro-apoptosis genes, it should be possible to control cell death to increase the maximum duration of cell culture (Lewis et al., 2013).

Identification of miRNAs: The availability of the CHO cell genome allowed for the identification of the genomic loci and sequence of 319 miRNAs by utilising homology between known human and mouse miRNAs with the CHO cell genome (Hackl et al., 2012). miRNAs have been shown to be a powerful tool for engineering CHO cell phenotypes such as rapid proliferation (Barron et al., 2011).

Prediction of genome editing targets: The genome sequence of CHO cells has enabled site-specific genome editing using CRISPR-CAS9 technology, allowing the identification of target sites for guide RNAs to mutate a target locus. A database containing approximately 2 million CRISPR targets within 27,533 CHO cell genes has been published (Ronda et al., 2014).

1.6 - RNA-Seq

Section 1.6 presents a review of the available methodologies for analysis RNA-Seq data. Sections of this literature review are in part adapted from the work published in the following review paper and protocol book chapter:

1. "Towards next generation CHO cell biology: Bioinformatics methods for RNA-Seq based expression profiling" (2015) Monger, C., Kelly, P., Gallagher, C., Clynes, M., Barron, N., Clarke, C. *Biotechnol. J.* **7:** pp.950-66 doi 10.1002/biot.201500107
2. "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

1.6.1 – Introduction to RNA-Seq

The high-throughput sequencing of transcripts, or RNA-Seq, is an analytical technique while applies next generation sequencing (Section 1.6.5.2) to enable the simultaneous discovery of novel mRNA and ncRNA transcripts and measurement of gene expression levels (Mortazavi et al., 2008). Expression levels can be compared between sample conditions in order to identify differences in gene expression in response to a stimulus or change in environment. While microarrays can also be utilised to investigate changes in gene expression, the application of RNA-Seq offers additional advantages which arise because RNA-Seq is not reliant on the binding of transcripts top pre-defined probe-sets (discussed in 1.6.5).

In order to explain the methodology presented in subsequent sections some RNA-Seq specific terminology need to be defined. The sequence of an individual transcript is obtained as a string of nucleotides referred to as a read. The total number of reads which can be sequenced by an instrument is referred to as the throughput. The process of mapping reads to a location on the genome from which the transcript they are derived from is expressed based on homologous sequence matching is known as alignment. The number of reads aligned to a genomic locus is known as the depth while the percentage of nucleotides with read alignments crossing them over a genomic range is referred to as coverage. Abundance refers to the estimated expression level of a gene based on coverage at a gene locus. Sequencing-by-synthesis describes the process of exploiting DNA replication during sequencing methodology.

The RNA-Seq protocol begins by first enriching RNA libraries for biologically useful information, which is necessary as over 90% of the RNA in a sample is ribosomal RNA (rRNA), and reverse transcribing the remaining RNA to cDNA (Section 1.6.2.1) (Zhao et al., 2014b). Adapter sequences such as sample barcodes and sequencing primer binding sites are ligated to the cDNAs and libraries are amplified by PCR (Pease and Sooknanan, 2012). Next generation sequencing is then used to sequence RNA libraries (Section 1.6.2.2) resulting in as many as 100 million sequence 'reads' per sample (Conesa et al., 2016). A crucial step in RNA-Seq data analysis is determining the expression levels of each expressed transcript in a sample. Abundance measurements are achieved by aligning reads to a reference genome or transcriptome sequence to determine the genomic loci from which the transcript was derived (Section 1.6.3.2) (Engström et al., 2013). The depth of reads aligned to a gene or transcript is then used to estimate the expression abundance (Section 1.6.3.3) and can be compared to the expression levels of transcripts in other samples to identify differential expression (Section 1.6.3.4) (Love et al., 2014). RNA-Seq data can also be assembled into full length transcript sequences to detect novel transcript isoforms and the presence of alternatively spliced transcripts between samples (Grabherr et al., 2011; Trapnell et al., 2013). If a ribosomal depletion strategy was used to enrich RNA samples rather than a polyA selection during sequence library preparation it is also possible to identify the expression of ncRNA species (Friedländer et al., 2012). RNA-Seq methodology including the preparation of sequencing libraries, sequencing instruments and software for analysis of RNA-Seq data is discussed in detail in the following sections.

1.6.2 – RNA-Seq Methodology

1.6.2.1 Sample preparation

RNA is typically extracted from cells with phenol-chloroform reagents such as TRIzol (Life Technologies) or a silica gel column (Qiagen). DNA contamination can be removed by treating samples with DNAase and then ribosomal RNA must be removed from samples as rRNA accounts for up to 90% of the RNA in eukaryotic cells (Costa et al., 2010) which would take up too much of the available throughout of the sequencing instrument. rRNA can be removed by either directly depleting rRNA with a kit such as Illuminas Ribo-Zero rRNA removal kit or by selecting mRNAs by their polyA tail using oligo-dT beads such as the TruSeq RNA Library Prep Kit V2 (Illumina). rRNA removal has the benefit of retaining ncRNA species such as miRNAs and lncRNAs, however if only protein coding mRNAs are being analysed then a polyA selection kit is more appropriate (Sultan et al., 2014). Extracted RNA sample quality and purity can be analysed using spectrophotometry based instruments such as a NanoDrop (Thermo Scientific) and RNA integrity and size distribution can be examined using electrophoresis methods such as on a Bioanalyzer 2100 instrument (Agilent).

Purified RNA is prepared for sequencing by fragmenting RNAs to an appropriate size using either nebulization, sonication or enzymatic digestion (Linnarsson, 2010). Fragment length distribution can be reassessed using electrophoresis methodology after which an optional size selection stage can take place to enrich libraries for RNA fragments of a given size. A library preparation kit can then be used to add sequencing primers and indices to RNA fragments, such as a TruSeq Stranded Total RNA kit (Illumina) for sequencing on Illumina instruments. RNA-Seq strategies can utilise paired-end sequencing protocols which provides a sequencing read from both the 5’ and 3’ end of each RNA fragment sequenced. When using paired-end sequencing a stranded library preparation kit such as the TruSeq Stranded Total RNA kit retains information on whether a read in a read pair comes from the 5’ and 3’ direction by incorporating dUTPs in the second sequenced strand (Levin et al., 2010). Stranded sequencing provides higher accuracy in downstream analyses stages when processing reads, particularly when carrying out assembly or studying lncRNA.

1.6.5.2 Next Generation Sequencing Technologies

Sequencing was first developed in 1975 by Frederick Sanger (Sanger and Coulson, 1975) using a process which has been named Sanger sequencing. Sanger sequencing utilises PCR to replicate a fragment of DNA in the presence of dideoxynucleotides (ddNTPs) which block the further incorporation of nucleotides in subsequent rounds of PCR if incorporated into an elongating chain of nucleotides (Heather and Chain, 2016). As DNA synthesis is utilised in this reaction the process can be referred to sequencing-by-synthesis. After multiple rounds of PCR the result is strings of nucleotides of different lengths based on when a ddNTP was incorporated instead of a standard deoxynucleotide. By using this process in 4 separate reactions, each using a different ddNTP, it is possible to obtain sequences of all lengths of a fragment. When separated based on size using electrophoresis (with a separate well for each of the 4 ddNTPs), the nucleotide sequence can be 'read' by the order of bands of DNA visible (Heather and Chain, 2016). The process of Sanger sequencing has been refined, automated and scaled up by the use of sequencing machines capable of simultaneously providing the nucleotide sequence of millions of input DNA or cDNA fragments – known as massively parallel or next generation sequencing. It is this increase in sequencing throughput which enables the use of RNA-Seq – where one sequence read is representative of one RNA molecule within the sample which can be quantified and interpreted as relative levels of gene expression. A number of sequencing instruments are available which each use different technologies and have varied throughputs and properties of output reads.

454 Sequencing

The basis of 454 sequencing is the detection of pyrophosphates released from replicating DNA when dNTP is incorporated into the replicating strand which leads to the commonly used name to describe this sequencing technique – pyrosequencing (Ronaghi et al., 1998). 454 Pyrosequencing was commercialized in 2005 and used to sequence the second personal genome and first genome costing less than $1 million (the genome of James Watson - the co-discoverer of the structure of DNA) (Wheeler et al., 2008). The pyrosequencing process begins with ligation of cDNA to a micron-sized streptavidin bead followed by utilisation of emulsion PCR to amplify each cDNA into millions of clonal copies on each bead (with one unique cDNA per bead) (Rothberg and Leamon, 2008). Once amplification is complete beads which do not contain any cDNA are filtered out and emulsion PCR oil is cleaned from beads. The remaining beads are loaded onto a PicoTitre plate containing 1.6 million wells of sufficient diameter to allow only a single DNA loaded capture bead in each well. Beads containing the enzymes required for the sequencing reaction are also loaded onto the PicoTitre plate packing each well. The sequencing reaction then begins by sequentially washing the PicoTitre plate with nucleotide triphosphates in the order TACG. As the nucleotides enter each well, if it is complementary to the template strand of DNA being replicated it is incorporated into the elongating chain. Pyrophosphate released is used as a substrate by the enzyme ATP sulfurylase to convert adenosine 5' phosphosulfate into ATP. Resulting ATP is then used as a substrate by the enzyme luciferase which converts luciferin to oxyluciferin, releasing visible light. The light signal from each of the 1.6 million wells is recorded by a CCD camera and apyrase degrades any unincorporated nucleotides and ATP so that the reaction can be repeated for the next nucleotide. The process cycle of sequentially adding each of the nucleotides is repeated hundreds of times. When there are consecutive occurrences of the same nucleotide in tandem on the DNA molecule being sequenced, multiple nucleotides are incorporated during the same sequencing cycle and a proportionate level of ATP and light is generated. For example, if TTT is sequenced, when tyrosine is washed across a plate, 3 nucleotides are incorporated into the replicating strand in a single sequencing cycle and 3 times the level of light is recorded. 454 sequencing with the GS FLX Titanium XL+, can deliver read lengths of up to 1000bp with a run time of 23 hours and at an accuracy of 99.997%. However 454 sequencing throughput is limited by the relatively low read sequence capacity per run of 1 million due to the limitation of wells on a PicoTitre plate. 454 sequencing also suffers from being unable to accurately determine the amount of nucleotides in long homopolymer regions. In 2015, Roche announced the phasing-out of this technology due to it no longer being competitive with other sequencing instruments.

Ion Torrent Sequencing

Released in 2010, the Ion Torrent sequencing instruments also utilise sequencing by synthesis and a process similar to that of 454 sequencing. Rather than using optics to detect visible light emitted from phosphate reactions, the Ion Torrent system utilises semiconductors to detect hydrogen ions released during replication of DNA (Rothberg et al., 2011). Library preparation is also similar to 454 and uses emulsion PCR to amplify each cDNA fragment on beads which are loaded onto an Ion Sequencing Chip – a chip containing millions of sensor wells each fitting a single bead with amplified DNA. Sequencing begins by sequentially adding the 4 nucleotide triphosphates in turn. If a nucleotide is incorporated into the extending primer a proton is released shifting the pH in the well which is detected by a sensor and converted to a voltage recorded by sequencing software. Excess nucleotides are washed away and the cycle is repeated for the next nucleotide. For homopolymer regions where 2 or more bases are identical the voltage shift is proportional to the number of inserted nucleotides which is interpreted by the signal-processing software (Rothberg et al., 2011). Advantages of this technology include a fast runtime of less than 4 hours and a decreased cost of sequencing as expensive enzymes, fluorescent tagged nucleotides and optics are not required. Ion torrent technology suffers from a higher error rate than other second generation sequencing platforms (approximately 1.8%) (Quail et al., 2012). The latest high throughput Ion Torrent instrument (Ion S5 XL) outputs up to 80 million 200bp or 20 million 400bp reads and their benchtop sequencer (Ion PGM) outputs 5.4 million 400bp reads.

SOLiD Sequencing

As with 454 and Ion Torrent sequencing SOLiD also utilises an emulsion PCR bead-based library amplification step, however, the SOLiD system allows for the sequencing of mate-paired libraries. Mate-pair libraries are circularized by joining of the 5’ and 3’ ends of a cDNA by the ligation of an insert of predetermined length to each end. A selection process enriches for beads containing DNA fragments and they are covalently attached to a glass slide. Rather than sequencing by synthesis, SOLiD sequencing uses the ligation and measurement of fluorescently labelled probes to determine sequence (Shendure et al., 2005). The set of fluorescent probes are 16 distinct probes of degenerate 8-mer oligos which can bind a specific 2-mer of DNA. The 16 probes are labelled with 4 fluorophores which allow for a di-based encoding of sequence.

A separate sequencing reaction occurs for each bead on the glass slide initiated by a universal sequencing primer annealing to the 5’ adapter. The first round of sequencing begins by the ligation of a single fluorescent probe binding the template sequence. The fluorescent label is excited and its emission spectra recorded before being cleaved off. Cleavage of the fluorescent label allows for the incorporation of the next probe, offset by 5 nucleotides, and this process is repeated until the desired read length is met. Once enough cycles have finished to read the full length, the extended sequence is denatured to reset the template and a new universal sequence primer is added which binds to n-1 on the adapter sequence. The entire sequencing reaction is repeated until 5 rounds of sequencing by ligation have occurred, each with a further n-1 offset, meaning each base is read twice. The di-based encoding system is then used to determine the full length sequence (Breu, 2010). The latest SOLiD system, the 5500xl, is able to sequence up to

100 million reads per run with 99.99% accuracy. Reads produced with this system are relatively short, with reads reaching a maximum of just 60bp in mate-pair libraries.

Illumina Sequencing

The Solexa sequencing technology, later commercialized by and henceforth referred to as Illumina sequencing, is a sequencing technology released in 2008 (Bentley et al., 2008). Illumina sequencing incorporates principles of both sequencing-by-synthesis and SOLiD sequencing, utilising fluorescently labelled dideoxynucleotide terminators to sequence strands of DNA during their synthesis. To prepare a genomic library for Illumina sequencing, DNA is first fragmented and sequencing adapters, indices and primers are ligated to both ends of the fragments. The library is then washed across a flow-cell - a glass array with clusters of oligos complimentary to the sequencing adapters. The DNA fragments are denatured into single stranded molecules and a single fragment is incorporated into each cluster of oligos on the flow-cell. Through a process known as bridge-amplification, the single fragment of DNA is replicated to generate clonal clusters of identical fragments.

Dideoxynucleotides with a distinct fluorescent tag are washed across the flow-cell and a single nucleotide is incorporated into each fragment of DNA in every cluster. Excess unincorporated fluorescent nucleotides are washed away and a laser excites the fluorescent tags of the nucleotides incorporated into each cluster. The emission spectra from each cluster is captured by a CCD camera and tris(2-carboxyethyl)phosphine (TCEP) is used to cleave the fluorescent tag and generate a free 3’ hydroxyl group primed for subsequent rounds of sequencing. The process is repeated until the desired read length has been achieved, adding and sequencing one nucleotide per cycle. For paired-end sequencing, the template strands are replicated and denatured. The forward strand is washed away and the sequencing process is repeated for the reverse strand.

The current Illumina sequencing instruments available offer varied sequencing throughputs and read-lengths and are each tailored towards different applications. The Illumina MiSeq outputs up to 25 million paired-end reads at a read-length of 300bp, the NextSeq 500 outputs up to 400million paired-end reads at 150bp and the Illumina HiSeq 4000 can sequence up to 5 billion paired-end 250bp reads in a single run. Multiple Illumina HiSeq instruments can be paired into the HiSeq X5 or X10 instrument, giving the power to sequence genomes at a population scale at a cost of less than $1000 per human genome (Illumina Inc. 2017).

Table 1.5– Advantages and disadvantages of second generation sequencing platforms. Presented in this table is an overview of the 4 second generation sequencing platforms and their advantages and disadvantages.

	Advantages	**Disadvantages**
454	Long read length, Low cost per Experiment, Low error rate	High cost per Mb, Low reads/sequencing run, Discontinued platform
SOLiD	Low error rate, Adaptable flow-chip, Paired-end/Mate-paired reads	Software often not compatible with colourspace reads, Short reads
Illumina	Variety of options for throughput and read length, Paired-end reads, High reads/sequencing run	Shorter reads than 454, Expensive for small scale sequencing experiments
Ion Torrent	Fast runtime, Low cost small scale experiments, Low cost instrument	High error rate, Shorter reads than 454, Lower reads/sequencing run than SOLiD & Illumina

1.6.3 - Software for the processing of RNA-Seq data

After sequencing has been completed, instruments output strings of nucleotides known as reads, where a given read is representative of a single cDNA which was in the original sample. Before this data can be interpreted in a meaningful way a number of analysis steps must first take place. Stages of a typical RNA-Seq analysis are summarized in Figure 1.5. Software for each stage of analysis is summarised in this section.

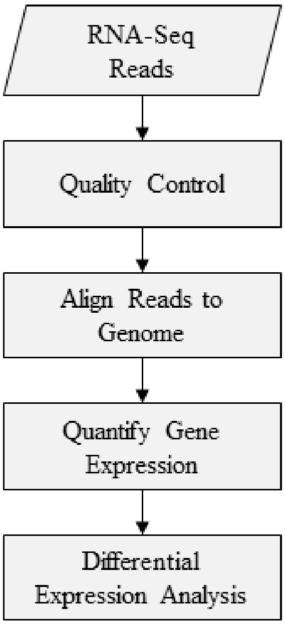

Figure 1.5 – Stages of an RNA-Seq differential expression analysis pipeline. RNA-Seq reads must first be quality controlled to remove any issues which can arise during sequencing. Reads are then aligned to a reference genome and gene expression is quantified. Differential expression analysis algorithms are then used to identify significant fold changes of genes between conditions.

1.6.3.1 Quality Control

Prior to the computational analysis of a RNA-Seq experiment it is vital to ensure raw data is of high quality. There are several sources of error which can occur in NGS reads and arise from library preparation or sequencing itself. Phasing is a phenomenon which results in base quality decreasing in the 3’ of sequence reads due to imperfect sequencing chemistry causing some of the nucleotides in a sequencing cluster failing to incorporate into the elongating nucleotide chain or incorporating too many nucleotides in a single sequencing cycle (Fuller et al., 2009). Phasing causes reads in a cluster to become out of sync with each other and causes an accumulative effect on the confidence of a base call – therefore resulting in lower confidence and quality in later cycles. Ambiguous base calls can also affect sequence quality due to the emission spectra of the fluorescently labelled nucleotides used in Illumina sequencing overlapping and therefore the base-calling software cannot confidently distinguish between two nucleotides (Ledergerber and Dessimoz, 2011). Base calling software assigns a quality score to each nucleotide based on the confidence of the base call which can be analysed before processing RNA-Seq data to ensure high quality.

FastQC is a user friendly piece of software which summarises raw sequence data quality metrics into easily interpretable plots (Andrews, 2010). If any quality issues are detected in *FastQC*, it is best practice to pre-process RNA-Seq data to trim or remove reads with poor sequence quality or adapter content. A number of tools are available for the pre-processing of RNA-Seq data including the *fastx-toolkit*, *Trimmomatic* and *cutadapt* (Bolger et al., 2014; Hannon Lab; Martin, 2011). *FastX* trims all reads in a sample to a user defined threshold (either a specified number of nucleotides from the 5' and 3' end or all nucleotides below a quality threshold from both ends). The approach used by *FastX* is naïve and results in the unnecessary loss of information as all reads are trimmed by the same amount. *Trimmomatic* applies a sliding window approach in which each read is processed individually by taking the average quality score across a specified number of nucleotides and trimming when the quality falls below the specified threshold. Both *cutadapt* and *Trimmomatic* are capable of removing adapter sequences, quality trimming and length filtering to remove short reads in a single command.

1.6.3.2 Alignment

Following quality pre-processing, analysis can begin by aligning reads to a reference genome. RNA-Seq reads derived from mature mRNAs contain exon-exon junctions due to the removal of introns from pre-mRNA transcripts and therefore alignment algorithms need to consider splicing when aligning RNA-Seq reads to a genome sequence (Kim et al., 2013). Gapped alignment algorithms have been designed specifically for the purpose of aligning RNA-Seq reads including *TopHat2*, *STAR*, *CRAC* & *HISAT2* (Dobin et al., 2013; Kim et al., 2013, 2015a; Philippe et al., 2013).

The *Tophat2* alignment process begins by first aligning reads to the reference genome in end-end mode (meaning large gaps due to introns are not considered) (Kim et al., 2013) using *Bowtie2* as the alignment engine (Langmead and Salzberg, 2012). During this first stage of alignment, only reads which map in their entirety to a single exon are aligned. Reads which did not align (and therefore likely to be derived from a section of mRNA containing an exon-exon junction) are then split into shorter segments and the alignment process is repeated. If two segments align to regions of a gene flanking an intron then the whole unsegmented read is used in a local alignment to that region of the genome, allowing the discovery of novel exons and splice sites.

STAR is considerably faster at aligning reads to a genome than *Tophat2,* however, the genome index required to use *STAR* on a mammalian genome is >30Gb and therefore a computer with a large amount of RAM is required (Dobin et al., 2013). The large index size enables the usage of "Maximal Mappable Prefix" (MMP) methodology to rapidly identify the location in the genome matching the longest prefix (5' end of a read onwards) perfectly matching a read. Genome locations at which mapped prefixes aligned are considered seed regions from which alignment is extended. The bases in a read following the mapped prefix are used in a second MMP alignment in which the next exon is typically identified. MMP alignments are iteratively repeated until the full length of a read has been aligned, at which point mapped seeds are joined to identify the exons making up the full gene.

CRAC is an RNA-Seq alignment algorithm capable of predicting sequencing errors, mutations and chimeric splice junctions (Philippe et al., 2013). *CRAC* applies a kmer based alignment strategy in which reads are broken up into all possible substrings of nucleotides of length k – where k should be set as the length at which it is statistically likely that the kmer will align to just 1 location in the reference genome (22 for a typical mammalian genome). A unique property of the *CRAC* aligner is the usage of support profiles. Both the genome locations at which kmers of a given read match and their support profiles (the number of reads which share a given kmer, allowing the approximation of coverage) are used to simultaneously score alignments. Kmers with low support are likely sequencing errors and are not aligned. Imperfectly aligned kmers which have high support (of equal coverage to overlapping kmers) are true mutations such as insertions, deletions or SNVs. Kmers of a read which overlap by k-1 nucleotides in their alignment position are used to determine the genome alignment location and identify mutations and splices. *CRAC* is not as fast as *STAR* and uses large genome indices requiring up to 64Gb of RAM but has been shown to have higher sensitivity and precision (Philippe et al., 2013).

HISAT2 utilises a dual genome indexing strategy known as Hierarchical Graph FM indexing (HGFM) in order to rapidly align reads (Kim et al., 2015a). HGFM indexing uses a whole genome index to seed alignments to a region of the genome and then uses a second smaller 64,000bp index to extend alignments. Small 64,000bp indices are generated to cover the whole reference genome and overlap by 1,024bp to allow ease of alignment of reads aligning over index boundaries. *HISAT2* is faster than the STAR algorithm, as sensitive and precise as other competing algorithms and can be ran on a standard computer due to the indexing strategy requiring just 4Gb of RAM.

1.6.3.3 Feature quantification

Once reads have been aligned to a reference genome, they must then be assigned to annotated gene features in order to calculate expression abundance. Expression quantification algorithms can be categorised into in two main categories – those which use gene level counting and those which use transcript isoform deconvolution. Gene level counting is the simplest approach which is applied by algorithms such as *HTSeq-counts* and *featureCounts* (Anders et al., 2014; Liao et al., 2014). Gene counting algorithms simply assign aligned reads to genes based on features present in an annotation file which overlap their aligned genome location and output counts of reads per gene feature. Gene counts are typically used by gene level differential expression algorithms and are subject to further normalisations before analysis occurs. Isoform deconvolution approaches used by programs such as *Cufflinks* and *Stringtie* (Pertea et al., 2015; Trapnell et al., 2010) attempt to estimate the expression of each individual transcript isoform expressed from a gene locus. Isoform deconvolution uses coverage of exon-exon junctions unique to a given transcript isoform to distribute a proportionate number of reads aligning to shared regions of a transcript to give consistent across of a transcript. Expression estimates output from isoform deconvolution algorithms are not raw read counts and are typically output as normalised FPKM or TPM units in order to account for different transcript lengths and library size.

1.6.3.4 Differential expression

In order to identify statistically significant changes in expression among conditions using count data, differential gene expression (DGE) software must be used. Raw read counts are often used as input to DGE software and normalised before a statistical test is applied. Examples of the most widely utilised DGE software packages include *DESeq2* and *edgeR* (Love et al., 2014; Robinson et al., 2010). Both *DESeq2* and *edgeR* employ similar processes to identify differentially expressed genes and typically output concordant results but differ in their methods of normalisation used.

DESeq2 applies a normalisation to gene counts by calculating the geometric mean for each gene across the samples in all conditions and dividing the counts of each sample for a gene with the mean counts to generate a scaling factor. The median scaling factor of all genes in each sample is then used to normalise the read counts of each gene, outputting counts which have been normalised for differences in sequencing depth among samples. Usage of median size factors in this way ensures genes with extremely high expression abundance do not strongly influence library size normalisation. Fitting of a negative binomical generalized linear model to each gene is then applied before a Wald test is used to identify significant differential gene expression between experimental conditions (Anders and Huber, 2010). *edgeR* utilises a trimmed means of M values (TMM) normalisation in which library size scaling factors are calculated using the means of all but the most highly and lowly expressed genes, such as removal of the 0-5% and 95-100% of the data. TMM normalisation achieves the same aim as *DESeq2* normalisation and prevents outlying genes with extremely high expression abundances over influencing the normalisation. A Poisson exact test is then applied to identify differential expressed genes among conditions (Robinson et al., 2010).

1.6.4 - The advantages of RNA-Seq over microarray

With over 2 decades of published research, microarray technology is the most widely utilised method of measuring global gene expression changes (Schena et al., 1995). The use of microarrays has the advantages of robust established statistical analysis methods and relatively low cost, however the use of RNA-Seq as a method of measuring the transcriptome brings many advantages:

No prior sequence data required to measure gene expression: RNA-Seq reads can be assembled into full length transcripts allowing the creation of *de novo* transcriptomes which can then be used as a reference for aligning reads and determining their abundance (Robertson et al., 2010). While this process is enhanced by the availability of a reference genome on which to initially align reads, this is by no means required (Trapnell et al., 2010). Microarrays however rely on homology based binding of transcripts to probe-sets which must be designed prior to the onset of the experiment and therefore only transcripts with a known nucleotide sequence can be studied. In addition this advantage enables RNA-Seq datasets to be reanalysed when updated annotation or genome assemblies are available while microarrays are limited to the probes physically present on the microarray chip.

Increased dynamic detection range: Not only can RNA-Seq identify novel transcripts, the sensitivity of the measurement of low abundance transcripts has also been demonstrated to be superior when compared to microarray technology with the number of reads sequenced the only limiting factor (Zhao et al., 2014a). The detection of fluorescent intensity of microarray probes limits the dynamic range at which fold-changes can be determined, as background levels of signal prevent the accurate measurement of low abundance transcripts and signal saturation affects the most highly abundant of transcripts. In RNA-Seq a separate nucleotide sequence is generated for each transcript in a sequence library which are utilised as the exact counts of transcripts to a gene rather than an estimation of expression based on fluorescent intensity and therefore has no such limitation.

Differential expression at any level: The ability to assemble full length transcripts from sequence data enables differential expression analyses at the level of the whole gene, individual transcript isoform or even individual exon to identify alternative splicing or differential exon usage without prior knowledge of transcript isoforms expressed (Anders et al., 2012b; Trapnell et al., 2013). The detection of alternative splicing and differential exon usage is possible using microarrays such as GeneChip ST arrays (ThermoFisher Scientific) however detection of novel splice junctions is not possible using this technology, Experiments have shown that differential gene expression fold changes from RNA-Seq experiments are more consistent with qPCR results than microarrays (93% concordance for RNA-Seq, 75% for microarray) (Wang et al., 2014).

1.6.5 - Challenges in applying RNA-Seq to study CHO cell biology

While RNA-Seq is a powerful technique which has many advantages over techniques such as microarrays there are a number of challenges which must be overcome during data analysis in order to obtain biologically informative results. First, the selection of an appropriate reference genome and annotation is unclear. With 5 draft genomes each with their own annotation files publicly available, there needs to be a consensus on which genome and annotation is most suitable. One study has evaluated the genome sequences available, coming to the conclusion that the best approach is likely to sequentially align to the Chinese hamster genome followed by the human and mouse genomes to maximise read alignment (Le et al., 2015). The study was however not comprehensive as the authors failed to include all available genomes and transcriptomes available at the time of publication. Progress has been made towards improving the annotation of the CHO cell genome with an Ensembl genome annotation now available however this annotation has yet to be evaluated or utilised in any published research (Zerbino et al., 2018).

Second, there is an overwhelming number of tools for the analysis of RNA-Seq data, each of which come with the option to modify parameters and data assumptions which must be configured and optimised to correctly analyse a dataset. Chen et al. published a pipeline for analysing CHO cell RNA-Seq data, however this study includes software specifically configured for studies with low replicates and does not include any software for pre-processing, aligning or quantifying RNA-Seq data (Chen et al., 2015). A pipeline developed to perform RNA-Seq analyses in CHO cells which incorporates all stages of analysis from pre-processing to differential expression during the work carried out in this thesis has been published (Monger et al., 2017).

Finally, RNA-Seq experiments produce a large amount of data which requires high-powered computers to store and process it as well as expertise in the field of bioinformatics to utilise the tools which are invoked using the UNIX command line and R programming language. Access to a graphical interface from which analysis pipelines are selected may make the analysis of RNA-Seq data more user-friendly for biologists who are not skilled in informatics. Distribution of computing power using cloud-based technologies could also help overcome the challenge of storing and analysing data.

1.6.6 – The progress of RNA-Seq in CHO cells

The publication of the Chinese hamster genome (Brinkrolf et al., 2013; Lewis et al., 2013; Xu et al., 2011) has enabled the use of RNA-Seq to study CHO cells using genome alignment RNA-Seq strategies and this technology has begun to be used to study the CHO cell transcriptome and identify expression patterns associated with industrially beneficial phenotypes (Fomina-Yadlin et al., 2015; Könitzer et al., 2015). In this section the challenges and progress of applying RNA-Seq to CHO cells are discussed.

Despite the challenges associated with performing an RNA-Seq analysis in CHO cell lines, a handful of research groups have published studies utilising the technology. The first experiment to apply RNA-Seq analysis to CHO cells was published in 2010, a year before the publication of a reference genome, and therefore used *de novo* assembly to overlap RNA-Seq reads into full length transcripts (Birzele et al., 2010). In this study an Illumina Genome Analyzer II was used to sequence 36bp single-end reads which were assembled using the *de novo* short read assembler *Velvet* to assemble RNA-Seq reads into contigs without prior sequence knowledge (Zerbino and Birney, 2008). The authors also used a guided approach in which alignment to annotated mouse transcripts was used prior to assembly. By using this knowledge-based approach as well as homology searching, 13,000 CHO cell line genes were identified of which 5,000 were novel. Although the sequence and annotation of transcripts was not released to the public, this study served as a proof of concept that short-read sequencing technology RNA-Seq approaches can be used to study the CHO transcriptome.

By cultivating CHO cells in concentrations of 0.5mM, 1.0mM and no butyrate and collecting samples at culture day 0, 4, 6 and 8, this 12 sample experiment enabled the simultaneous study of the changing transcriptome patterns over time and on how the presence of butyrate affects the transcriptome dynamics of CHO cells. Reads were aligned to the generated CHO cell reference transcriptome as well as mouse transcripts using *Bowtie* and differential expression calculated using *SAGEBetaBin* (Langmead et al., 2009; Vêncio et al., 2004). The effect of butyrate treatment resulted in differential expression of 1,785 genes of which 1,037 were upregulated and 748 downregulated. While successful in assembling RNA transcripts and performing a differential gene expression analysis (the results of which were shown to be in accordance with those found using microarrays), this study was limited by a lack of publicly available genome and transcriptome sequences and annotations and therefore relied on homology to mouse transcripts to infer biological information.

Along with the publication of the CHO cell line genome in 2011, Xu et al. published the first RNA-Seq study utilising a genome-guided RNA-Seq approach (Xu et al., 2011). RNA-Seq reads were sequenced using an Illumina Genome Analyzer II in paired-end mode, aligned to the reference genome with the splice aware alignment algorithm *Tophat* and assembled into transcripts with *Cufflinks* (Trapnell et al., 2009, 2010). 24,282 genes and 29,291 transcripts were found to be expressed in CHO-K1 cells. The authors published results concentrating on the expression of viral susceptibility and glycosylation gene expression. A similar approach was used by Lewis et al. to annotate the genome sequence of the Chinese hamster genome, in which an Illumina HiSeq 2000 was used to sequence paired-end reads which were assembled into 98,116 transcript contigs with *Trinity* and aligned to the reference genome (Grabherr et al., 2011; Lewis et al., 2013).

RNA-Seq has also been used in 2 further attempts to characterise the transcriptome of CHO cells by the same research group, first in 2011 and again in 2014 (Becker et al., 2011; Rupp et al., 2014). The first attempt used the raw RNA-Seq data published by Birzele et al. as well as data acquired from their own cell lines following heat and cold shock on a 454 GS FLX sequencing instrument. RNA-Seq data was assembled into 28,995 transcript contigs which were deposited in NCBI Genbank with accession numbers JP037157-JP066151. Resulting transcript contigs were however largely unannotated, not easily searchable and not available in a format for use as genomic annotation in an RNA-Seq experiment. A second transcriptome assembly and genome annotation was published by this research group utilising a variety of assembly strategies, including *Cufflinks*, *Trinity* and *Velvet* (*Oases*) and reference genome guided assembly with the CHO-K1 cell line genome to assemble 65,561 transcripts from 17,598 genes in the ‘GFF format’ which can be used by RNA-Seq splice aware alignment algorithms as splice junction annotation (https://gendbe.cebitec.uni-bielefeld.de/cho.html). The availability of the reference genomes and transcriptome annotation by these studies have enabled RNA-Seq studies investigating a variety of transcript patterns including codon optimisation (Chung et al., 2013), effects of copper on lactate metabolism (Yuk et al., 2014), single nucleotide variants (Vishwanathan et al., 2014), glycoprofiles (Könitzer et al., 2015) and temperature shift (Bedoya-López et al., 2016).

1.7 Motivation, Design and Aims

1.7.1 – The utilisation of temperature shift in recombinant protein producing CHO cell cultures

As outlined in section 1.4.1, a challenge facing the biopharmaceutical industry is the drive to produce recombinant therapeutics in a predictably cheaper way. One process utilised to achieve higher yields and therefore lower production costs in the biopharmaceutical industry is the exposure of mammalian cell cultures to mild hypothermia conditions after a critical mass of cells have accumulated towards the end of the exponential growth phase termed "temperature shift" (Kumar et al., 2007). Reports of the effects of temperature shift on recombinant antibody production show improved titres of 2-5 fold (Oguchi et al., 2006; Tan et al., 2008) due to this process however the molecular biology underpinning the improvements to productivity have yet to be fully elucidated (Oguchi et al., 2006). The phenotypic effects of temperature shift include delayed apoptosis (and therefore a longer culture time during which cells are producing recombinant proteins) (Bedoya-López et al., 2016), enhanced cell specific productivity (Vergara et al., 2014), a halting of cell cycle in the G0/G1 phase (Moore et al., 1997), an increase to recombinant mRNA levels (Tan et al., 2008), and a change in metabolism from lactate production to consumption (Martínez et al., 2015). While a decrease in proliferation rates seems counterintuitive towards increasing productivity, the G1 phase is the time of the cell cycle at which cells have the highest cell volume (Martínez et al., 2015). Additionally the drawback of reducing growth rates is overcome by using the "biphasic culture" technique (Tan et al., 2008) where cultures are initially inoculated at 37°C (a condition at which cells can rapidly proliferate) before reduction of temperature to between 28-34°C (Farrell et al., 2014; Martínez et al., 2015).

A number of studies to have begun to enhance our understanding of the biological processes behind the temperature shift phenotype by utilising omics techniques (discussed in section 2.1). It is currently known that the cold inducible binding protein (*Cirbp*) gene plays a role in the temperature shift response and is expressed in CHO cells following a reduction in culture temperature (Tan et al., 2008). Overexpression of *Cirbp* has been demonstrated to improve recombinant protein titres in cultures inoculated at

37°C (Tan et al., 2008) however overexpression of this gene alone is not sufficient to fully mimic the effects of temperature shift and results in only a 40% increase in titre compared to the 2-5 fold increased reported for temperature shift and therefore other genes and processes likely contribute to the increase in productivity (Tan et al., 2008). Studies have shown rates of global translation are impacted by temperature shift due to a decrease in translation elongation factor expression (Baik et al., 2006) and translation initiation factor activity at temperatures below 32°C (Masterton et al., 2010). Additionally, temperature reduction has shown to increase mRNA levels due to enhanced mRNA stability (Masterton et al., 2010) which may be due to *Cirbp* activity (Xia et al., 2012). miRNA expression also plays a role in the temperature shift response with miRNAs including miR-21 and miR-24 which inhibit growth shown to be upregulated (Gammell et al., 2007) which may play a role in regulating translation of proliferation inducing genes during mild hypothermia.

CHO cells growing in the exponential growth phase also exhibit a Warburg type metabolism (Martínez et al., 2015) where glucose is inefficiently metabolised into lactate rather than pyruvate (the fuel for the TCA cycle and oxidative phosphorylation) despite the presence of oxygen (aerobic glycolysis) (Locasale and Cantley, 2011). A Warburg type metabolism is advantageous during the exponential growth phase as increased glucose uptake and glycolysis rates provides a carbon source for lipid and amino acid biosynthesis required to increase the biomass of growing cells (Locasale and Cantley, 2011; Pereira et al., 2018). Continued usage of a Warburg type metabolism is however undesirable during the stationary growth phase for two reasons: first, the build-up of acidic lactate and ammonium by-products of glycolysis are toxic to CHO cells and can inhibit growth and induce cell death (Pereira et al., 2018) and second, the energy requirements of unproliferating cells to synthesise proteins required for cellular maintenance as well as recombinant protein production are higher than that of proliferating cells (Locasale and Cantley, 2011). The inefficient utilisation of glucose primarily in glycolysis rather than as fuel for oxidative phosphorylation is unwarranted as it produces approximately 20 fold less ATP per glucose (Locasale and Cantley, 2011). Switching the metabolism of CHO cells from a Warburg type metabolism to lactate consumption at the exponential growth to stationary phase transition is therefore advantageous and this can be induced by temperature shift (Martínez et al., 2015).

By discovering more about the temperature shift response at the levels of transcription and translation on a genome wide scale it may be possible to identify targets for cellular engineering to enhance or mimic the temperature shift response in CHO cells with the aim of producing cell lines with improved productivity. The potential benefits of engineering a temperature shift-like phenotype with increased productivity has been demonstrated previously such as reports of upregulating *Cirbp* expression (Tan et al., 2008), the use of miR-7 to regulate proliferation (Barron et al., 2011) and recently a report of reducing the Warburg phenotype by inhibiting inhibitors of the pyruvate dehydrogenase complex which converts pyruvate into acetyl-CoA fuel for the TCA cycle (Buchsteiner et al., 2018). As previous studies have demonstrated a role of post-transcriptional mechanisms contributing to the temperature shift response, a multi-omics approach needs to be applied to profile regulation of gene expression at the levels of transcription and translation.

1.7.2 – Experimental Design

In order to assess post-translational control occurring in temperature shifted CHO cells the experiment detailed throughout this thesis applies multi-omic techniques to capture gene expression at levels ranging from the transcription of mRNA from a gene to the translated protein. 8 biological replicates of antibody producing CHO cells are cultured at 37°C for 48hrs at which point 4 samples are temperature shifted to 31°C to induce a mild hypothermia phenotype (detailed in Section 2.2.1). 24hrs later cells are collected from each culture and utilised in parallel multi-omics experiments detailed in the following chapters. First, the levels of mRNA and protein expression are assessed using RNA-Seq and proteomics to identify differential gene and protein expression as well to identify the number of genes where expression at the mRNA and protein level do not correlate (Chapter 2). Previous studies have shown a lack of correlation between mRNA and protein levels in CHO cells (Clarke et al., 2012; Courtes et al., 2013) and it is therefore expected this will be the case in this study. The following chapters then aim to elucidate the levels at which post-transcriptional control occurs in temperature shifted CHO cells, beginning with the role of alternative splicing utilising RNA-Seq data (Chapter 3), the role of miRNAs using small RNA-Seq (Chapter 4) and the impact of differential translation efficiency using RiboSeq (Chapter 5). The experimental design of this study is illustrated in Figure 1.6.

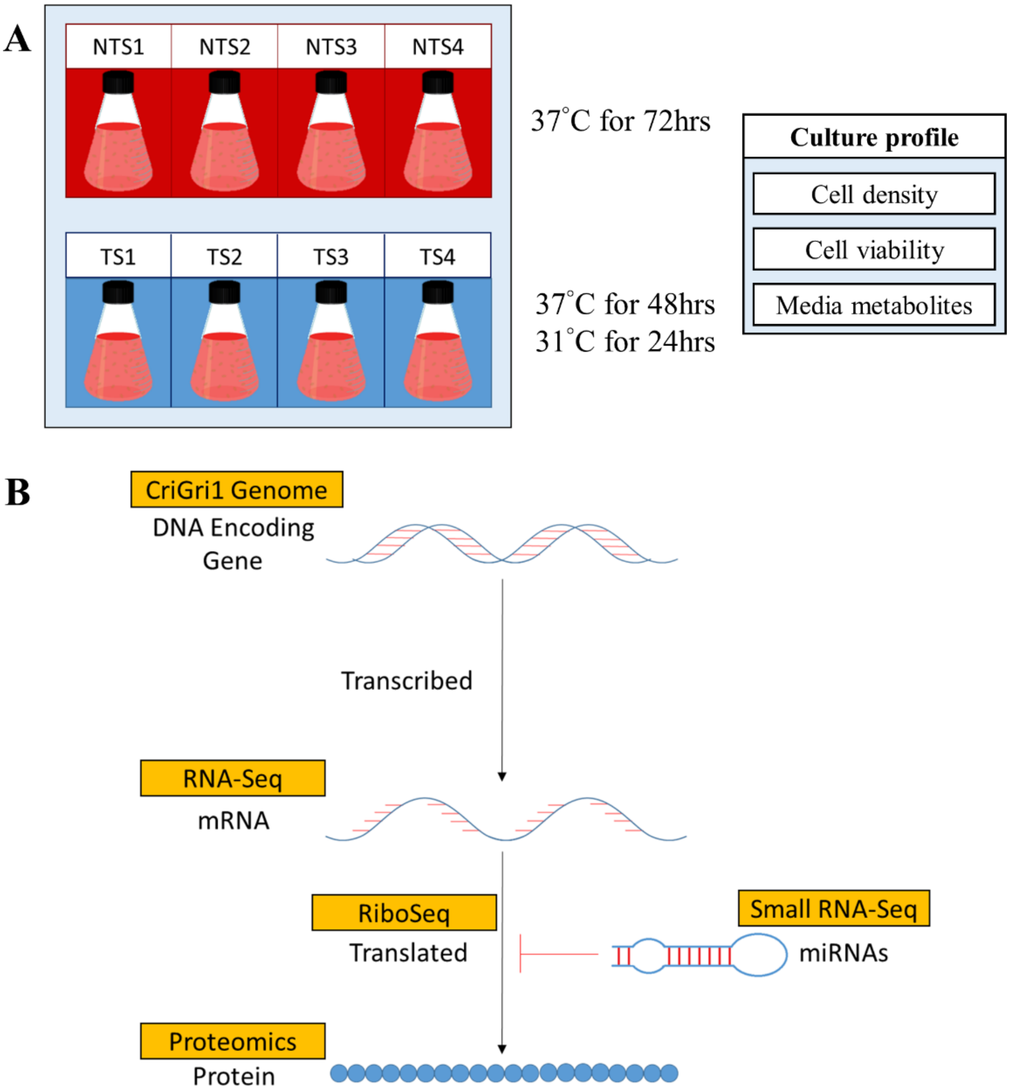

Figure 1.6 - Experimental design for parallel multi-omics study of temperature shift in CHO cells (A) Biological replicate recombinant antibody producing CHO cells are cultured at 37°C for 48hrs in 8 shaker flasks. The temperature of 4 samples is reduced to 31°C and after a further 24hrs cells are harvested for multi-omics profiling. (B) Different omics techniques are applied (yellow boxes) to profile factors contributing to, and measure, gene expression at the mRNA and protein levels.

By integrating datasets which represent different levels of gene expression, complex interactions of gene expression mechanisms can be elucidated. For example in Chapter 3 the role of alternative splicing is assessed which may result in changes to gene expression without affecting the overall levels of mRNA transcribed from a given gene. Alternative splicing resulting in altered 3'UTRs may also affect miRNA targeting of transcripts which is assessed in Chapter 4. Production of spliced isoforms with codons more efficiently translated may result in altered translation efficiency of transcripts which can be assessed in Chapter 5. Differentially expressed miRNAs identified in Chapter 4 can be investigated using target prediction databases and potential activity in temperature shifted CHO cells can be identified if target genes are differentially expressed at the mRNA or protein level in Chapter 2. miRNA mediated translation repression may also be identified through identification of miRNA targets deemed differentially translated in Chapter 5 or genes without correlated expression levels at the mRNA and protein level identified in Chapter 2. Finally, differentially translated genes identified in Chapter 5 can be compared to genes without correlated gene and protein expression measurements identified in Chapter 2 and the potential mechanism resulting in differential translation can be identified if a gene is identified as alternatively splice or as a target of a differentially expressed miRNA in Chapters 3 and 4 respectively. The interactions of the multi-omics datasets assessed throughout this thesis are summarised in Figure 1.7.

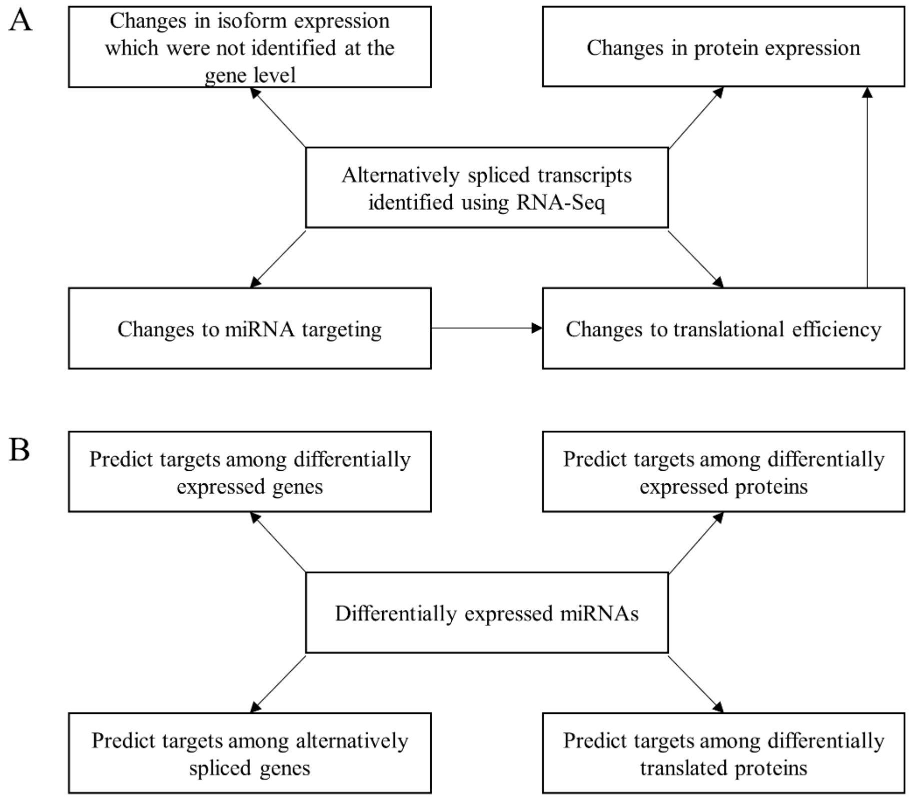
A
Changes in isoform expression which were not identified at the gene level
Changes in protein expression
Alternatively spliced transcripts identified using RNA-Seq
Changes to miRNA targeting
Changes to translational efficiency
B
Predict targets among differentially expressed genes
Predict targets among differentially expressed proteins
Differentially expressed miRNAs
Predict targets among alternatively spliced genes
Predict targets among differentially translated proteins

Figure 1.7 Multi-omics dataset integration Shown in this figure are the comparisons which will be made between datasets in this experiment. (A) Alternatively spliced transcripts can be investigated in high throughput RNA-Seq data and isoform level differential expression compared to gene and protein level differential expression as well as to find altered miRNA targets and genes differentially translated. (B) Differentially expressed miRNAs can be used in target predictions to assess their impact on differential gene and protein expression as well as on alternatively spliced transcripts and differentially translated genes.

Generation and analysis of these datasets will present a number of key challenges which must be overcome in order to gain biologically meaningful information. First, generation of multi-omics datasets can be hindered by batch effects which occur due to the timing of sampling for each dataset. Sampling biases will be overcome by harvesting cells for each ‘omics experiment simultaneously and storing samples by freezing until used in library preparation and sequencing. Second, batch effects due to sample storage, shipping and sequencing platform among the NGS based studies are limited by shipping samples together to the same sequencing core and sequencing samples on the same Illumina sequencing platform. Third, using different types of data (based on RNA-Seq and Proteomics methodologies) means it may be difficult to integrate results between the methods. Utilisation of a protein database generated using the genome sequence means this issue can be overcome by using the same gene identifiers across experiments such that results can be integrated. Finally, the different ‘omics techniques each have inherent biases and provide differing coverages and therefore integrations will be limited to the number of gene measurements observed by the technique with the lowest coverage.

1.7.3 - Project aims

- To review the software available for the analysis of RNA-Seq data and identify software best suited for CHO cell RNA-Seq analysis.
- To assess the available Chinese hamster genomic data and annotation to determine the optimal reference sequence for use in RNA-Seq analysis.
- To evaluate the optimal parameters for software to process RNA-Seq data.
- To apply NGS to CHO cells to improve our biological understanding of the transcriptome patterns associated with an industrially relevant reduction in cell culture temperature.
- To identify differential expression of mRNA and ncRNA transcripts associated with reduced temperature.
- To perform global identification of alternative splicing of transcripts in CHO cells for the first time.
- To integrate RNA-Seq data with proteomics data to identify correlation of expression levels between mRNAs, ncRNAs and proteins.
- To apply a ribosomal sequencing technique (RiboSeq) to study transcripts being actively translated and combine this data with mRNA, ncRNA and protein expression data to study CHO cells undergoing a temperature shift at a systems level.
- To characterise novel mRNA transcripts and alternatively spliced isoforms as well as novel ncRNAs and use these data to improve the annotation of the Chinese hamster genome.

Chapter 2

Differential expression analysis of temperature shifted CHO cells

2.1 Introduction

A reduction in cell culture temperature is widely used throughout the biopharmaceutical industry to increase culture longevity and production yields in CHO cells (Trummer et al., 2006). In response to temperature shift, CHO cell growth is arrested while maintaining or increasing recombinant protein productivity (Kumar et al., 2007). Although the biological mechanisms driving improved product titres achieved in response to temperature shift have not yet been fully characterised, a number of mechanisms are known to contribute to enhanced bioprocess performance. Reduced glucose and oxygen consumption, reduced lactate and ammonia production, delayed apoptosis and increased levels of recombinant mRNA via increased transcription or mRNA stability have all been observed during mild hypothermia (Kumar et al., 2007). By further characterising the CHO cell response to reduced culture temperature at a molecular level, it may be possible to identify targets for engineering cell lines with enhanced productivity (Borth, 2014; Laux, 2014; Wright and Estes, 2014). For example, identifying the pathways through which a cell increases RNA stability, longevity, or productivity could enable the enhancement of these traits to produce designer cell lines with improved productivity through manipulating the expression of key genes.

To date, a number of studies have been carried out to understand the effect of mild hypothermic conditions on CHO cell behaviour (Kumar et al., 2007). The first 'omics study to profile CHO cells cultured at a reduced temperature utilised mass spectrometry to identify proteins differentially expressed after temperature shift from 37°C to 33°C (Baik et al., 2006). In this study high abundance proteins were selected from 2D-PAGE gels and measured using mass spectrometry resulting in identification of 26 proteins of which 7 were upregulated and 2 downregulated. Temperature shift from 37°C to 31°C has also been profiled using mass spectrometry with 53 proteins identified of which 18 were upregulated and 5 downregulated (Kumar et al., 2008). Cold induced expression of proteins with roles in growth, glycoprotein quality, metabolic shift (Baik et al., 2006) and translation (Kumar et al., 2008) was identified in these studies. The methodology used in these studies was however limited in terms of the number of proteins which could be identified and quantified. The UniProt-KB database for the Chinese hamster holds 34,958 proteins (2018), indicating only a fraction of the proteins expressed in CHO cells were surveyed in these studies. Transcriptomics experiments performed previously to investigate the effects of reduced culture temperature were limited due to reliance on microarrays with pre-defined probe-sets to profile transcript expression. For example, probe-sets to profile expression of 1,655 and 4,643 genes were present on the rat and mouse microarrays utilised in pioneering CHO temperature shift microarray experiments (Baik et al., 2006) and 6,822 genes on a CHO cell specific cDNA library microarray (Yee et al., 2009). Despite limitations to the number of genes which were able to be surveyed, microarray experiments successfully identified differentially expressed cold shock genes as well as genes with roles in energy metabolism, cell cycle regulation and apoptosis which may contribute to the CHO cell response to mild hypothermia (Baik et al., 2006; Yee et al., 2009).

While previous studies have contributed to our understanding of the molecular biology coordinating the CHO cell response to decreased culture temperature, the mechanisms underpinning a mild hypothermia phenotype are not fully understood (Masterton and Smales, 2014). By applying techniques which provide a global view of transcription we may be able to further elucidate alterations to gene expression which contribute to the increased product yields achieved by culturing CHO cells at a reduced temperature. In this chapter we apply RNA-Seq and LC MS/MS proteomics to CHO cells which have been temperature shifted from 37°C to 31°C during the exponential phase of culture and remained at 31°C for 24 hours. A key aspect of the experimental design is the use of deep RNA-Seq to profile gene expression at high resolution on a global level which can serve as a baseline to compare measurements from other ‘omics studies including proteomics. RNA-Seq and proteomics have not been utilised in parallel previously to study mild hypothermia in CHO cells and application of this technique may shed light on the contribution of post-transcriptional control of gene expression regulating the temperature shift response.

2.2 Materials and Methods

2.2.1 Cell Culture & experimental design

A recombinant antibody producing CHO-K1 cell line was grown in 20ml of Serum-Free Media (ThermoFisher Scientific) in 250ml shaker flasks at 37°C. Cultures were seeded at a density of 2×10^5 cells/ml in 8 biological replicates. Sample identifiers were assigned to each culture, summarised in Table 2.1. The temperature of 4 of the 8 cultures was reduced from 37°C to 31°C after 48hrs and all samples were grown for a further 24hrs. Cell density and viability was measured in 24hr intervals throughout the duration of the culture using a Guava easyCyte 5HT instrument (Merck). After 72 hours from seeding, 1ml aliquots of cells were collected for RNA-Seq and stored for use in further experiments. Cell culture was carried out by Dr. Paul Kelly in NICB, Dublin City University.

Table 2.1 – Experimental design and cell culture. Detailed in this table are the conditions used to generate each of the CHO cell cultures from which RNA was harvested in this experiment. The sample identifiers given to each sample are used throughout this chapter.

Sample Identifier	Condition	Culture configuration
NTS_1	Non-temperature shifted	37°C for 72hrs
NTS_2	Non-temperature shifted	37°C for 72hrs
NTS_3	Non-temperature shifted	37°C for 72hrs
NTS_4	Non-temperature shifted	37°C for 72hrs
TS_1	Temperature shifted	37°C 48hrs, 31°C 24hrs
TS_2	Temperature shifted	37°C 48hrs, 31°C 24hrs
TS_3	Temperature shifted	37°C 48hrs, 31°C 24hrs
TS_4	Temperature shifted	37°C 48hrs, 31°C 24hrs

2.2.2 Library preparation and next generation sequencing

Samples of cells were harvested 72hrs post-seeding from each culture and TRIzol (ThermoFisher Scientific) used to extract RNA. Ribo Zero Gold (Illumina Inc.) was used to deplete rRNA and a TruSeq Stranded Total RNA Library Prep Kit (Illumina Inc.) used to generate sequencing libraries. RNA library quality and fragment size were assessed using a high sensitivity assay on a 2100 Bioanalyzer instrument (Agilent Technologies) prior to sequencing. Libraries were sequenced on a HiSeq 2500 instrument (Illumina Inc.) configured to generate a minimum of 50 million 150bp paired-end reads per sample.

RNA-Seq libraries were generated by Dr. Clair Gallagher and Alan Costello in NICB, Dublin City University.

2.2.3 Quality assessment and pre-processing of raw sequence data

FastQC was used to assess quality of raw sequence data (Andrews, 2010). The number of reads, per nucleotide sequence quality, GC content, sequence length distribution, presence of overrepresented sequences and adapter content was examined for each sample. The Illumina universal sequence prefix 'AGATCGGAAGAGC' was used to remove adapters using *cutadapt* (Martin, 2011). Nucleotides with a sequence quality below Q10 were trimmed from each end of the reads. Reads below 20 nucleotides in length post trimming were filtered. *FastQC* analysis was repeated to analyse read quality post trimming.

2.2.4 Alignment of pre-processed reads to the CHO-K1 genome

HISAT2 (Kim et al., 2015a) was used to align reads to the 'CriGri_1.0' CHO-K1 genome (ENSEMBL genome build accession GCA_000223135.1) (Xu et al., 2011). Alignment parameters were set to align in the 'reverse-forward' stranded mode to ensure the Illumina second strand dUTP marking library preparation method was correctly accounted for. To allow for compatibility with *Stringtie* (Pertea et al., 2015), the '--dta' (downstream transcriptome assembly) parameter was used which rejects read alignments with short-anchors on de novo splice junctions, thus reducing the false positive rate of assembled transcripts in downstream analyses. Aligned reads were converted from the SAM to BAM format, sorted by chromosome position and indexed using *Samtools* (Li et al., 2009).

2.2.5 Feature quantification

Aligned reads were assigned to existing genome annotation for the CHO cell genome available on the ENSEMBL FTP site (ftp://ftp.ensembl.org/pub/release-93/gtf/cricetulus_griseus_crigri/Cricetulus_griseus_crigri.CriGri_1.0.93.gtf.gz). *featureCounts* from the *SubRead* package (Liao et al., 2014) was used to assign the reads to genes based on annotated exon features with the parameters '–s 2' and '-p' for stranded and paired counting.

2.2.6 Gene level DE analysis

Output from *featureCounts* was parsed to create a gene count table for import into R. Prior to differential expression analysis, counts were filtered to remove genes with an average CPM (counts per million mapped reads) less than 1. Principal component analysis (PCA) was carried out using the *FactoMineR* R package on count data after the removal of low abundance genes. The *DESeq2* Bioconductor package (Love et al., 2014) was utilised to analyse differential gene expression. Those genes with a BH adjusted p value of less than ≤0.05 and fold change ≥±1.5 were designated as differentially expressed. Genes identified as differentially expressed were used as input for gene ontology analysis using *DAVID* to identify enriched GO terms with a Benjamini Hochberg (BH) adjusted P value of ≤ 0.05.

2.2.7 Metabolic profiling

Growth media taken from aliquots of cells grown as described in section 2.2.1 for each of the 8 cultures was utilised for metabolic profiling. Expended media from each sample was centrifuged to pellet cells and debris and 1ml of supernatant was harvested for analysis on a Bioprofile FLEX instrument (Nova Biomedical). The automated analysers “Chemistry and Gas Module” was utilised to measure levels of metabolites including glucose, lactate, glutamine, glutamate, ammonium, sodium, potassium and calcium as well as pH levels. A Students T-test was applied to identify temperature induced differences in metabolite consumption with a P value ≤ 0.05. Metabolic profiling was carried out by Dr. Ioanna Tzani in NIBRT.

2.2.8 Proteomics analysis

A lysis buffer containing 7M urea, 2M Thiourea, 4% CHAPS and 30mM Tris at pH 8.5 was applied to aliquots of cell cultures generated as described in 2.2.1. Samples were centrifuged to remove cell debris and protein containing supernatant was collected. Protein concentration was measured prior to label-free liquid chromatography-mass spectrometry (LC-MS/MS) using a Quick Start Bradford protein assay kit (Bio-rad). 10 µg from each sample was prepared for trypsin digest by reduction using 50mM ammonium bicarbonate (Sigma-Aldrich). Samples were alkylated using 5mM Dithiothreitol for 20 minutes at 56C° and 15mM lodoacetamide for a further 15 minutes in the dark. ProteaseMAX (Promega) was added to each fraction at a volume of 0.1% (v/v) for efficient digestion and further enzymatic digestion of proteins was carried out by adding Trypsin Gold (Promega) at a ratio of 1:20 to each sample and incubating overnight at 37°C. Trifuoroacetic acid (Sigma-Aldrich) was added to stop the enzymatic digestion reaction.

LC-MS/MS analysis was carried out on an Ultimate 300nanoLC instrument coupled with a LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific). 5μL of each digested sample was loaded onto C18 PepMap trap columns (300 μm id × 5 mm, 5 μm particle size, 100 Å pore size; Thermo Fisher Scientific) which were desalted for 5 minutes at a flow rate of 25 μL/min in 0.1% TFA and 2% Acetonitrile (Sigma-Aldrich). The trap column was switched on-line with an analytical C18 PepMap column (75 μm id × 500 mm, 3 μm particle, and 100 Å pore size; Thermo Fisher Scientific). A binary gradient of 2 solvents was used to elute peptides - solvent A (2% ACN and 0.1% formic acid (Sigma-Aldrich) in LC-MS grade water (Sigma-Aldrich) and 0%–25% solvent B (80% ACN and 0.08% formic acid in LC-MS grade water) for 160 min and 25%–50% solvent B for a further 20 min. The column flow rate was set to 350 nL/min and the Orbitrap configured to survey MS scans in the 400–1800 m/z range with a resolution of 30,000 at m/z 400 and lock mass set to 445.120025 u. Collision-induced dissociation fragmentation was carried out in the linear ion trap on the three most intense ions per scan. A dynamic exclusion window of 40s was applied. A normalised collision energy of 32%, an isolation window of 3 m/z, and one microscan were used to collect suitable tandem mass spectra.

Mass spectra were processed using *Progenesis QI* software (Non-linear dynamics) with the run with the most peptide ions selected as the reference run. Any drift in retention time between replicate samples was normalised for by aligning peak intensities to the reference run. A MASCOT file was exported from Progenesis and imported to *Proteome Discoverer* (Thermo Fischer Scientific) and searched against a database containing 26,985 proteins generated by retrieving the predicted amino acid sequences of transcripts annotated in the CHO cell Ensembl genome reference 93 using Biomart (Durinck et al., 2009). Proteins were identified using the MASCOT and SEQUEST search algorithms with parameters set to allow for a MS/MS mass tolerance of 0.6Da, 2 missed cleavages, a MASCOT ion score of ≥40, SEQUEST XCorrs of 1.9, 2.2 and 3 for +1-3 ions respectively, cysteine carbamidomethylation set as a static modification and methionine oxidation set as a dynamic modification. An ANOVA test was applied to identify differential protein expression between temperature shifted and non-temperature shifted samples. Thresholds of a ≤ 0.05 P value, a ≥ 1.2 fold change and a minimum of 2 identified peptides were applied to identify proteins considered differentially expressed. Proteomics analysis including sample preparation and LC-MS/MS were carried out by Orla Coleman and Michael Henry in NICB, Dublin City University.

2.3 Results

2.3.1 Cell culture

A 24.12% decrease in cell density was reported in temperature-shifted cells compared to non-temperature-shifted cells at 72hrs which had an average of 1.5 × 10^6 cells/ml and 2.06 × 10^6 cells/ml respectively (Table 2.2, Figure 2.1). No change in viability was observed in temperature shifted samples, with an average of 97% viable cells in both conditions at each time point. Metabolic shift was observed after temperature shift, with lower levels of lactate in the media of temperature shifted cells and higher glucose – indicating cells had transitioned from aerobic glycolysis to OXPHOS (Figure 2.2, Appendix 2A).

Table 2.2 – Cell density and viability. Cell density and viability were measured in 24hr intervals throughout the duration of cell culture. Reducing the culture temperature from 37°C to 31°C decreased cell density after 24hrs with no significant change in viability.

Culture ID	Day 1 Density (cells/ml)	Day 2 Density (cells/ml)	Day 3 Density (cells/ml)	Day 1 Viability (%)	Day 2 Viability (%)	Day 3 Viability (%)
NTS_1	5.29 × 105	1.06 × 106	2.13 × 106	96.4	97	97.6
NTS_2	5.53 × 105	1.01 × 106	2.10 × 106	96.2	96	97.3
NTS_3	5.55 × 105	1.06 × 106	2.11 × 106	97.6	97.2	97.6
NTS_4	5.18 × 105	1.02 × 106	1.91 × 106	97.5	97	96.6
TS_1	5.44 × 105	1.01 × 106	1.61 × 106	96.3	97.5	96.9
TS_2	5.37 × 105	9.80 × 105	1.66 × 106	96.6	96.1	96.9
TS_3	5.39 × 105	9.50 × 105	1.54 × 106	97.4	96.7	96.5
TS_4	5.34 × 105	9.57 × 105	1.45 × 106	97.8	96	95.5

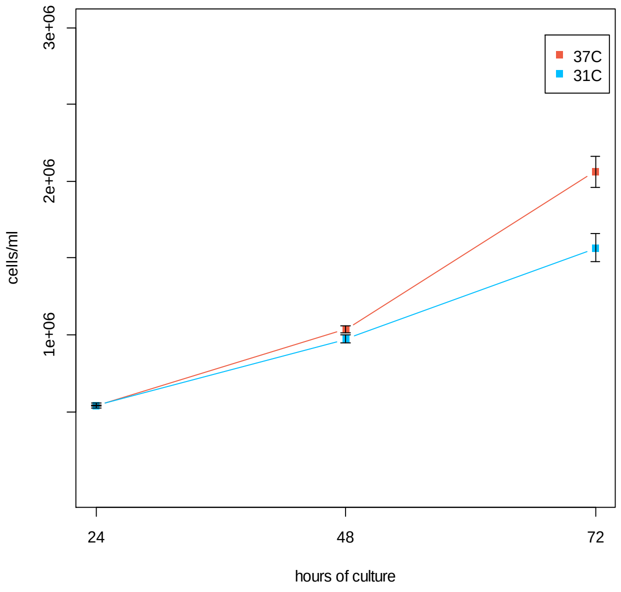

Figure 2.1 – Cell density of temperature shifted and non-temperature shifted CHO cells. Shown in this figure is the average cell densities of samples measured in 24hr intervals for temperature shifted (31°C, blue) and non-temperature shifted (37°C, red) samples. A 24% decrease in cell density was observed after 24hrs of temperature shift which took place during the 48-72hr culture period.

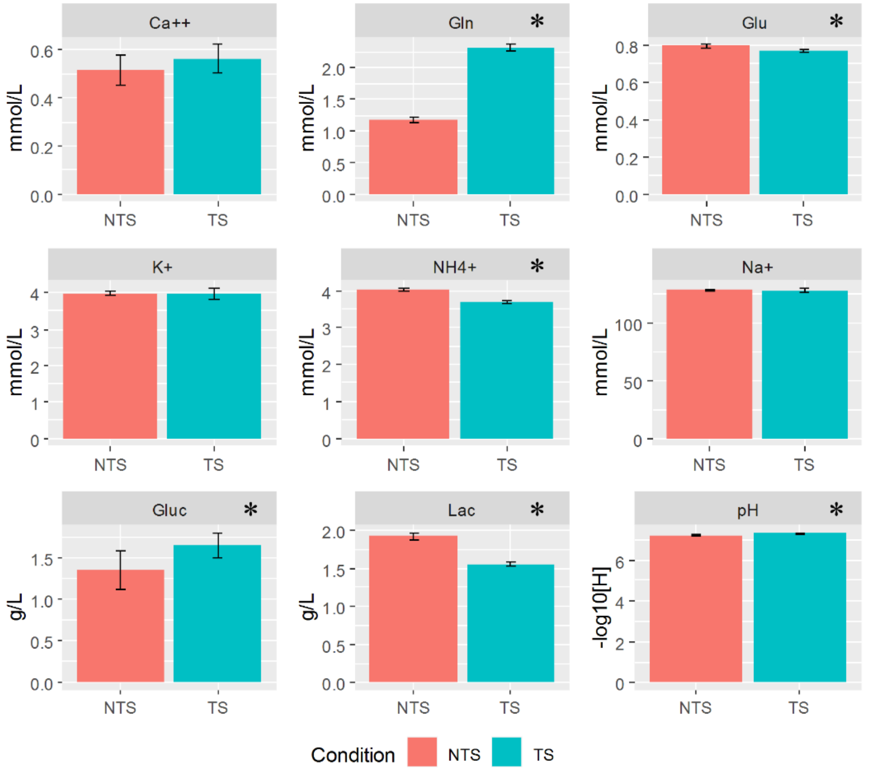

Figure 2.2 – Metabolite concentrations in temperature shifted and non-temperature shifted cell culture media. Reducing cell culture temperature resulted in the decrease in lactate and ammonium in media while glutamine, glucose and pH levels increased. Metabolites with a statistically significant difference in P value calculated by T test are denoted with an asterisk.

Following confirmation of the typical behaviour of CHO cells induced by a reduction bioreactor temperature (i.e. a decrease in growth rate and a switch from lactate production to consumption) RNA from cells at 72hrs was subjected to next generation sequencing.

2.3.2 Quality Control and Pre-processing

All 8 samples were successfully sequenced to a minimum sequencing depth of 50 million read-pairs per sample with an average of 69.7 million read pairs (50.9-80 million) acquired for each sample (Table 2.3). Reads were of high quality with an average Phred score of Q33 in all samples (Figure 2.3). 99.9% of reads in all samples survived trimming which successfully removed adapters and low quality nucleotides from the ends of reads, resulting in an increase in overall quality to Q34 (Figure 2.4). The quality of reads diminished at the 3' ends of reads, as is typically observed in reads generated by Illumina sequencing technologies, and the average nucleotide quality in the 3' ends of reads improved by trimming (Figure 2.4). An average GC of 48% was observed in reads. Prior to trimming all reads were of length 150bp and ranged from 20-150bp post trimming with the majority of reads above 145bp.

Table 2.3 – Sequence read quality summary statistics. The quality of sequence reads was inspected using *FastQC*. Reads were of overall high quality and samples were sequenced to a high depth (> 50 million reads). The presence of adapter content identified the need for quality pre-processing to remove adapters from the ends of reads.

Sample	Reads	Reads surviving	GC (%)	Adapter content (%)
NTS_1	70,447,918	70,441,755	48	21
NTS_2	76,600,082	76,592,061	48	18
NTS_3	69,654,137	69,645,676	47	22
NTS_4	70,644,823	70,637,585	48	18
TS_1	67,864,940	67,857,487	48	18
TS_2	50,979,659	50,975,133	47	17
TS_3	73,203,827	73,197,167	48	18
TS_4	77,979,243	77,971,134	49	17

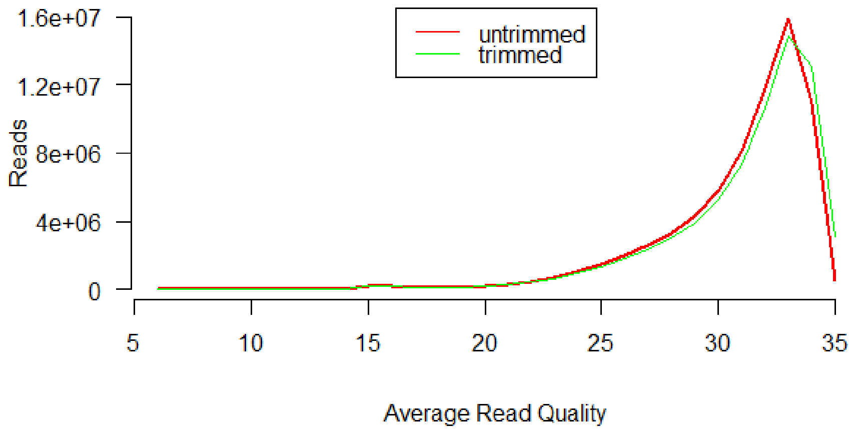

Figure 2.3 – Average read quality distribution for untrimmed and trimmed reads. FastQC was used to summarise the per-read average quality for each sample. The average of all samples is plotted here for untrimmed (red) and trimmed (green reads). An increase in the number of reads >Q33 in trimmed reads is visible.

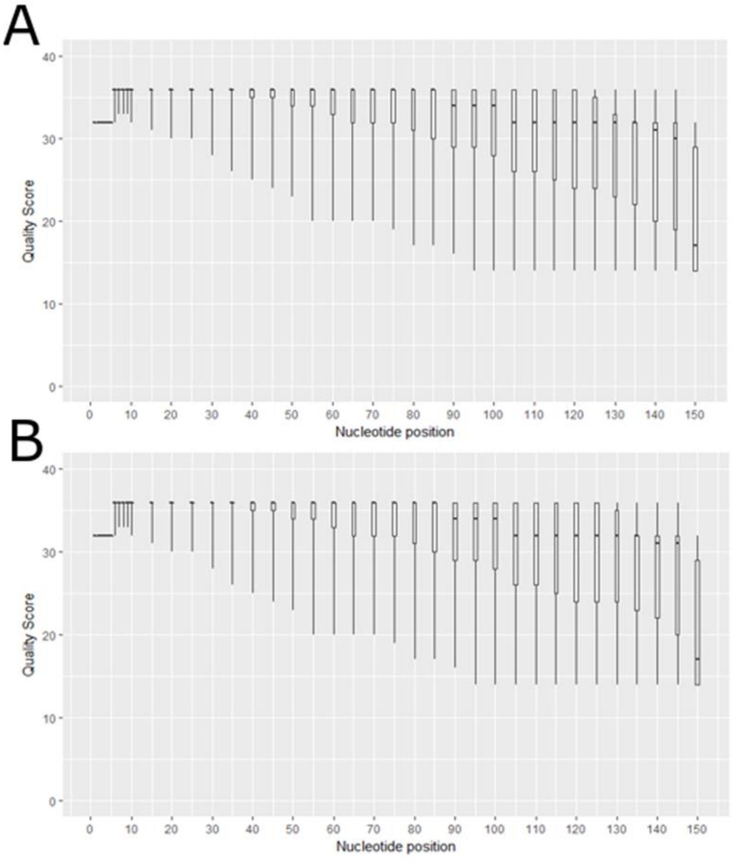

Figure 2.4 – Average per nucleotide sequence quality across the length of reads. The presented boxplots show the sequence quality of all samples (A) before and (B) after trimming. The 5' ends of reads were of particularly high sequence quality with all reads >Q30. Sequence quality decreases along the length of the reads but is improved slightly after trimming. Boxes represent the 1st-3rd quartiles of read quality and whiskers show the 10% and 90% intervals.

2.3.3 Alignment of RNA-Seq reads to the CHO cell genome

The optimal alignment algorithm and genome assembly to use for analysis throughout this thesis were assessed using publically available CHO cell RNA-Seq datasets and computationally simulated reads (Appendix 1). The findings of these experiments revealed *HISAT2* provided the best balance of runtime, low memory requirements and high sensitivity. The CriGri_1.0 genome was the most complete out of the available Chinese hamster and CHO cell genome assemblies and the ENSEMBL version of the genome annotation provides the most biological information on the genome.

An average of 84.46% (82.93-86.17%) of reads were aligned to the CHO-K1 cell CriGri_1.0 genome using *HISAT2* (Figure 2.5, Appendix 2B). Alignments to CHO cell rRNA and mitochondrial sequences were used to evaluate contamination levels and the efficiency of the rRNA depletion stage of library preparation. We observed low levels of both rRNA and mitochondrial derived reads with average alignment rates of 0.19% (0.12-0.33%) and 1% (0.87-1.34%) respectively (Appendix 2B). The reads mapping to known genome features annotated by Ensembl were determined using *featureCounts* (Liao et al., 2014). An average of 54.1% of reads were assigned to annotated gene features (51.57-56%, Figure 2.5, Appendix 2B). Read counts for each genome feature with gene abundance over 1CPM were used for principal component analysis. The resulting PCA plot is displayed in Figure 2.6.

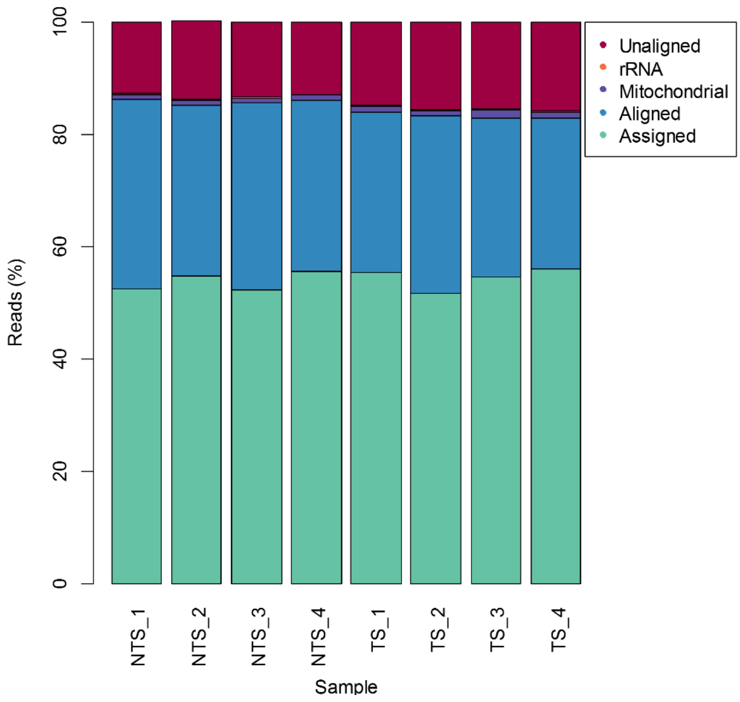

Figure 2.5 – Alignment statistics. An average of 84.46% of reads were aligned to the genome and 54.1% assigned to annotated gene features. Unaligned reads are represented in red, reads aligned to CHO cell rRNA sequences are in orange (not visible), reads aligned to the Chinese hamster mitochondria in purple, reads aligned to the CHO-K1 CriGri_1.0 genome in blue and reads assigned to annotated gene features in the Ensembl genome annotation in green.

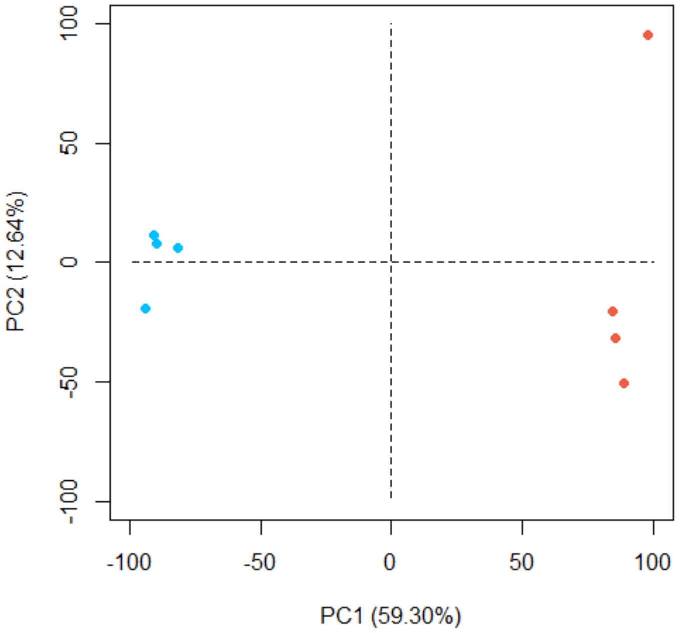

Figure 2.6 Principal component analysis plot of read counts assigned to genes. PCA was carried out on count data of genes expressed at a minimum of 1CPM. The first 2 principle components are plotted in this figure. Temperature shifted samples are plotted in blue and non-temperature shifted in red.

2.3.4 Differential gene expression with DESeq2

DESeq2 was used to identify differentially expressed genes using the counts generated in Section 2.5.3.4 as input. At a ≥ ±1.5 fold change and a BH p value of ≤ 0.05, 1,953 genes were identified as differentially expressed. Of these differentially expressed genes 937 were upregulated and 1,016 were downregulated at 31˚C (Appendix 2C, Table 2.4). Alignments of reads to highly upregulated and downregulated genes are illustrated in Figures 2.7 and 2.8. 53 GO categories were identified as enriched with a BH adjusted p value ≤ 0.05 (Appendix 2E, Table 2.5).

Table 2.4 – Top 10 significantly up and downregulated genes. Contained in this table are the most significantly differentially expressed genes which had an annotated gene symbol. baseMean refers to the DESeq2 normalised expression.

Gene ID	**Gene Symbol**	**Gene Name**	**Base Mean**	**FC**	**BH P Value**
ENSCGRG00000005663	*Nphp1*	Nephrocystin 1	3,760	2.84	7.05×10^{-302}
ENSCGRG00000008084	*Lgals3bp*	Galectin 3 Binding Protein	4,148	3.23	2.55×10^{-273}
ENSCGRG00000014728	*Tmem123*	Transmembrane Protein 123	3,287	3.52	8.13×10^{-270}
ENSCGRG00000012454	*Scn2a*	Sodium Voltage-Gated Channel Alpha Subunit 2	2,374	3.37	8.66×10^{-238}
ENSCGRG00000013654	*Tmem63b*	Transmembrane Protein 63B	3,521	4.20	1.10×10^{-224}
ENSCGRG00000010485	*Epha2*	EPH Receptor A2	4,241	2.66	1.35×10^{-221}
ENSCGRG00000011491	*Zmat3*	Zinc Finger Matrin-Type 3	2,773	3.07	3.02×10^{-211}
ENSCGRG00000008014	*Dzip1*	DAZ Interacting Zinc Finger Protein 1	2,093	3.52	4.57×10^{-205}
ENSCGRG00000007833	*Fam198b*	Family With Sequence Similarity 198 Member B	2,668	2.57	1.90×10^{-204}
ENSCGRG00000018207	*Tax1bp3*	Tax1 Binding Protein 3	1,777	2.90	9.42×10^{-200}
ENSCGRG00000012297	*Clspn*	Claspin	2,772	-3.21	1.56×10^{-267}
ENSCGRG00000019090	*Hist2h2bb*	Histone Cluster 2 H2B Family Member B	14,730	-4.99	6.45×10^{-267}
ENSCGRG00000015009	*Hist1h3f*	Histone Cluster 1 H3 Family Member F	3,037	-5.62	2.49×10^{-264}
ENSCGRG00000015066	*Kntc1*	Kinetochore Associated 1	4,932	-3.25	1.44×10^{-255}
ENSCGRG00000010342	*Cdca7*	Cell Division Cycle Associated 7	3,150	-2.51	2.09×10^{-255}
ENSCGRG00000012764	*Fanci*	FA Complementation Group I	2,803	-2.72	9.09×10^{-231}
ENSCGRG00000017959	*Kif14*	Kinesin Family Member 14	4,973	-2.25	1.69×10^{-222}
ENSCGRG00000003462	*Cdc45*	Cell Division Cycle 45	1,602	-2.95	2.28×10^{-219}
ENSCGRG00000003449	*Mcm7*	Minichromosome Maintenance Complex Component 7	5,854	-2.67	1.58×10^{-215}
ENSCGRG00000014843	*Pold1*	DNA Polymerase Delta 1, Catalytic Subunit	2,921	-3.03	2.87×10^{-209}

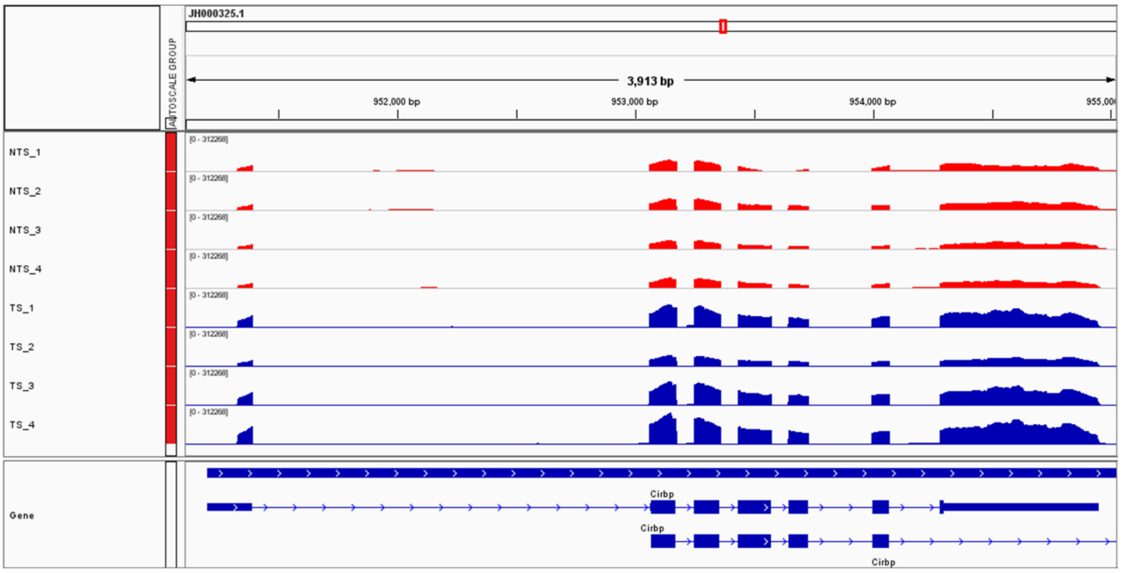

Figure 2.7 – Genome alignment of reads to the *Cirbp* gene. The cold inducible RNA binding protein was upregulated in the temperature shift condition (blue tracks) with a 2.3 fold change in expression.

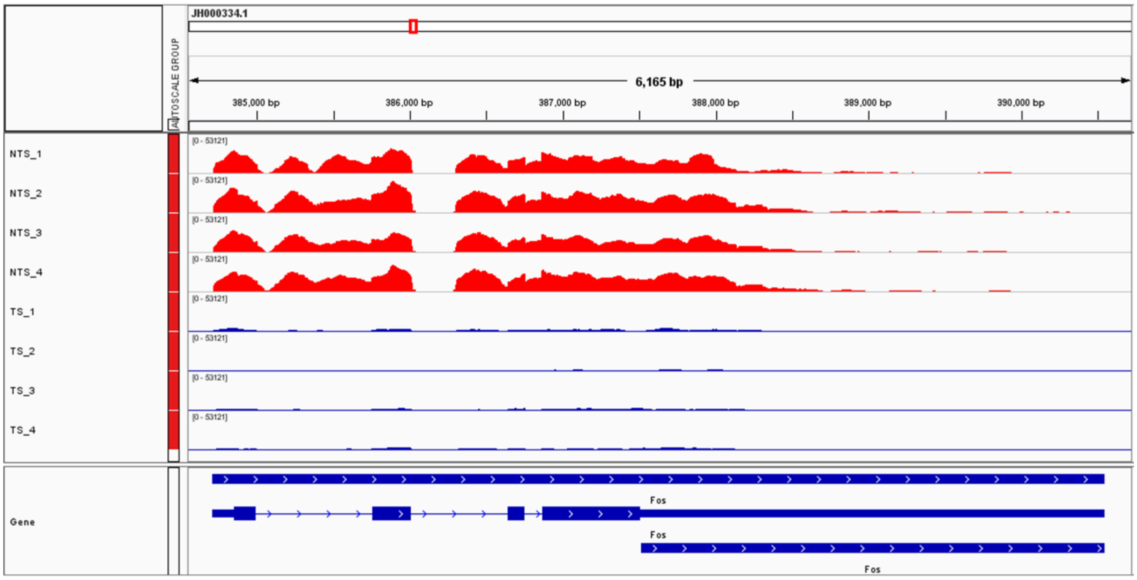

Figure 2.8 – Genome alignment of reads to the *Fos* gene. The *Fos* gene was highly expressed in non-temperature shifted samples (red tracks) however temperature shift induced a -10 fold downregulation of expression.

Table 2.5 – Enriched GO terms for genes with differential expression after temperature shift. The top 10 enriched GO terms of differentially expressed genes are presented in this table. GO analysis was performed using DAVID. Full table in Appendix 2E.

Term ID	Term	Count	%	Fold Enrichment	BH P value
GO:0006260	DNA replication	67	4	5.13	3.75×10^{-27}
GO:0006281	DNA repair	68	4.1	3.43	2.06×10^{-16}
GO:0051301	cell division	85	5.1	2.88	$4.97 \times 10\text{-}^{16}$
GO:0000722	telomere maintenance via recombination	23	1.4	8.53	5.87×10^{-14}
GO:0006270	DNA replication initiation	22	1.3	8.16	1.08×10^{-12}
GO:0007067	mitotic nuclear division	63	3.8	3.01	2.12×10^{-12}
GO:0000731	DNA synthesis involved in DNA repair	22	1.3	7.46	1.09×10^{-11}
GO:0007062	sister chromatid cohesion	37	2.2	4.26	2.05×10^{-11}
GO:0000082	G1/S transition of mitotic cell cycle	35	2.1	4.07	4.48×10^{-10}
GO:0006974	cellular response to DNA damage stimulus	52	3.1	2.97	8.19×10^{-10}

2.3.5 Differential gene and protein expression

348 proteins were detected with at least 2 peptides mapped to their amino acid sequence using mass spectrometry. 108 of these were identified as significantly upregulated after temperature shift and 63 downregulated (Appendix 2F). Gene ontology enrichment analysis revealed enrichment for cell cycle process, G1/S transition, mRNA stability, gene expression and protein transport (Table 2.6, Appendix 2G).

Table 2.6 – Enriched GO terms for proteins with differential expression after temperature shift. The top 10 enriched GO terms of differentially expressed proteins are presented in this table. GO analysis was performed using DAVID. Full table in Appendix 2G.

Term ID	Term	Count	%	Fold Enrichment	BH P value
GO:0010608	posttranscriptional regulation of gene expression	22	14.97	5.79	5.06×10^{-7}
GO:0046907	intracellular transport	37	25.17	2.88	9.01×10^{-6}
GO:1901566	organonitrogen compound biosynthetic process	34	23.13	3.07	7.06×10^{-6}
GO:0019752	carboxylic acid metabolic process	24	16.33	3.61	1.17×10^{-4}
GO:0043436	oxoacid metabolic process	24	16.33	3.59	1.04×10^{-4}
GO:0006082	organic acid metabolic process	25	17.01	3.42	1.03×10^{-4}
GO:0043933	macromolecular complex subunit organization	44	29.93	2.22	1.12×10^{-4}
GO:0043488	regulation of mRNA stability	11	7.48	9.43	9.82×10^{-5}
GO:0043487	regulation of RNA stability	11	7.48	8.99	1.35×10^{-4}
GO:0010033	response to organic substance	47	31.97	2.07	1.91×10^{-4}

Just 27 of the genes identified as differentially expressed at the protein level were identified with differential expression in the same direction in RNA-Seq data. 93 of the upregulated proteins and 51 of the downregulated proteins were not detected as differentially expressed at the gene level using *DESeq2*. 5 upregulated genes and 8 downregulated genes did not have corresponding differential expression at the protein level. 4 genes were identified with anti-correlated expression at the protein level (gene expression up but protein expression down or vice versa) (Table 2.7). The full results of genes and proteins without corresponding expression levels (153) are presented in Appendix 2H).

Table 2.7- Differentially expressed genes with anti-correlated protein expression. 4 genes were identified with anti-correlated expression between the gene and protein level with at least a ≥±1.5 fold change at the gene level and ≥±1.2 fold change at the protein level.

Gene	**Gene name**	**Gene FC**	**Gene P**	**Protein FC**	**Protein P**
Lmnb1	Lamin B1	-1.88	4.41×10^{-24}	1.71	2.9×10^{-2}
Cdk1	Cyclin-dependant kinase 1	-1.9	2.91×10^{-146}	1.21	1.4×10^{-2}
Hnrnpc	Heterogeneous Nuclear Ribonucleoprotein C	-1.66	1.29×10^{-120}	1.26	3.5×10^{-3}
H2afj	H2A Histone Family Member J	1.9	8.3×10^{-48}	-1.35	5.5×10^{-3}

2.4 Discussion

The experimental design of this study utilised CHO K1 cells cultured in biological quadruplicates for two conditions – those grown at 37°C and cells temperature shifted to 31°C. The effects of decreased culture temperature for 24hrs is sufficient to have a phenotypic effect on CHO cell growth rate without being lethal. We observed a 24% reduction in cell density 72hrs post-seeding in temperature shifted cells compared to those without temperature shift with no significant change in viability – indicating a decreased rate of growth and replication rather than apoptosis is responsible for the decreased cell density. In addition to a decreased growth rate, measurements of metabolites in the media indicated an increase in glucose concentration and decrease in lactate and ammonia levels in temperature shifted cells. The observed change in metabolites in media suggested a change from a Warburg type metabolism to an oxidative phosphorylation (OXPHOS) metabolic phenotype as is typical in temperature shifted CHO cells which have halted exponential growth and entered the production phase of culture and is associated with enhanced productivity (Buchsteiner et al., 2018). RNA-Seq data obtained through this experiment is of high sequence quality – with low levels of mitochondrial and ribosomal RNA contamination and adapter sequences successfully removed during pre-processing. Principal component analysis on gene count data revealed separation of the samples by their experimental condition in the first principal component, evidence that temperature shift has induced global changes to the transcriptome which have been captured in sequencing and that replicate samples are representative of each other.

Temperature shift resulted in widespread gene expression changes and 1,953 genes were identified as differentially expressed. Widespread expression changes in response to temperature shift have been previously reported and therefore this is an expected observation (Bedoya-López et al., 2016). Consistent with Bedoya-López et al. we identified differentially expressed genes associated with cold shock, apoptosis, cell cycle and metabolism as differentially expressed 24 hours after temperature shift. Differentially expressed genes were enriched for gene ontology terms including DNA replication, cell division, G1/S transition of mitotic cell cycle and the cellular response to DNA damage. Enrichment of differentially expressed genes regulating the cell cycle, particularly at the G1/S phase transition, are in agreement with existing knowledge of the temperature shift response in CHO cells which have been observed to pause at the G1 stage of the cell cycle

(Underhill and Smales, 2007). Increased viability and productivity of CHO cells has been linked to spending an increased amount of time in the G1 phase - at which cells are the largest and therefore have a higher capacity for producing proteins (Moore et al., 1997). Although the proportions of cells in each phase of the cell cycle and productivity were not measured, it is likely decreased growth rate induced by reduced culture temperature is due to an elongated G1 phase and temperature shifted cells would likely produce increased yields if batch cultures extended until the maximum duration.

Previous studies of mild hypothermia in CHO cells have revealed upregulation of cold shock genes including cold inducible RNA binding protein (*Cirbp*) and RNA binding motif protein 3 (*Rbm3*) (Bedoya-López et al., 2016). Cirbp is a well-studied cold shock protein which is known to play a role in the CHO cell response to reduced cold temperature and has been previously utilised as a genetic engineering target to enhance productivity in CHO cells (Tan et al., 2008). When *Cirbp* is overexpressed 90 fold compared to *Cirbp* expression in CHO cells cultured at 37°C, recombinant protein yields were found to increase by 40% relative to cell culture strategies incorporating a temperature shift to 32°C (Tan et al., 2008). The mechanism through which Cirbp impacts productivity is not yet fully understood however there is evidence to suggest Cirbp acts as an RNA chaperone during cold shock which binds mRNAs to regulate their expression post-transcriptionally (Zhu et al., 2016). Cirbp has been shown to associate with ribosomes as well as the untranslated regions of its target mRNAs and result in increased translation (Zhu et al., 2016). Cirbp has also been shown to have a role in regulating splicing and alternative polyadenylation of target transcripts which may enhance the stability of mRNAs (Morf et al., 2012; Zhu et al., 2016). In this study *Cirbp* was identified among the most significantly differentially expressed genes – upregulated with a 2.3 fold change and a P value of 5.76×10^{-188}. Although *Rbm3* is not as well studied as *Cirbp* in CHO cells the two genes are highly similar in terms of their protein structure, molecular function and cold specific upregulation (Zhu et al., 2016). Rbm3 also regulates splicing and translation of its targets through association with the spliceosome and ribosome (Zhu et al., 2016), however Rbm3 also regulates the expression of temperature specific miRNAs (Wong et al., 2016) – potentially through a role in miRNA biogenesis (Pilotte et al., 2011). *Rbm3* was also among the most significantly upregulated genes with a 2.99 fold change in expression and 1.56×10^{-112} P value. In addition Rbm3 was the most upregulated protein detected using LC-MS/MS with a 1.99 fold change and 5.7 × 10^{-5} P value. Upregulation of *Cirbp* and *Rbm3* is further evidence of a role of both these genes in the CHO cell response to mild hypothermia and is consistent with previous RNA-Seq experiments on temperature shifted CHO cells (Bedoya-López et al., 2016) - instilling confidence in this dataset.

In this experiment a 24% in cell density was observed in temperature shifted cells which we can infer is because of a decrease in growth rate due to sustained levels of viability. A halting in the G1 phase of growth has been reported previously in cold shocked CHO cells (Kumar et al., 2007), however the molecular mechanisms behind this change to growth rates remains to be elucidated. Gene ontology analysis revealed enrichment for "G1/S transition of mitotic cell cycle" (GO:0000082) among the most significantly enriched GO terms (1.5 × 10^{-10} P value). The 35 differentially express genes annotated with this process include cell division cycle 6 (*Cdc6*), cell division cycle 7 (*Cdc7*), cell division cycle 45 (*Cdc45*), cyclin D1 *(Ccnd1)*, cyclin E1 (*Ccne1*), cyclin E2 (*Ccne2*) and cyclin dependant kinase inhibitor 1a (*Cdkn1a*). Each of these cell cycle regulating genes were downregulated in response to temperature shift with the exception of *Cdkn1a* which is among the most highly upregulated genes (4.47 fold change and 7.64 × 10^{-122} P value) and *Ccnd1* (1.52 fold change, 5.65 × 10^{-82} P value). Progression from the G1 phase of the cell cycle in which cells grow to the S phase where DNA is replicated requires expression and activity of E cyclins which bind to and activate cyclin dependant kinase 2 (CDK2) (Teixeira et al., 2016). The mammalian E-type cyclins *Ccne1* and *Ccne2* were both downregulated in temperature shifted CHO cells with a -1.71 and -2.51 fold change respectively (2.21 × 10^{-22} and 4.59 × 10^{-47} P values). Ccne1/Cdk2 inhibition, either by knockdown or expressing inhibitors, has been shown to result in lengthened G1 phase in mouse embryonic stem cells (Li et al., 2012). The downregulation of *Ccne1* and *Ccne2* in CHO cells has been previously studied and found to be regulated by expression of the Nfat1 transcription factor to control the G1/S phase transition (Teixeira et al., 2016). The *Nfatc1* gene was not identified as differentially expressed in this study and therefore Cyclin E expression is likely regulated through other means in the response to decreased culture temperature. A potential mechanism is due to downregulation of the E2F transcription factors which are known to regulate the expression of *Ccne1* and *Ccne2* (Caldon and Musgrove, 2010) the *E2f1* and *E2f2* genes were downregulated in response to mild hypothermia with a -2.05 and -2.18 fold change respectively. *E2f1* and *E2f2* expression is supressed by *Ccnd1* which is identified as upregulated with a 1.5 fold change in temperature shifted cells. In addition to regulation at the level of gene expression, Cyclin E activity is regulated by p21 – a protein expressed from the *Cdkn1a* gene (Bertoli et al., 2013). p21 inhibits Cyclin E from forming a complex with Cdk2 to enter the S phase of the cell cycle and expression of *Cdkn1a* has been shown have a larger impact on the cell cycle than increased Cyclin E expression (Caldon et al., 2009). *Cdkn1a* is among the most significantly overexpressed genes in temperature shifted cells with a 4.43 fold change. Most interestingly p21 expression may be directly regulated during temperature shift by Cirbp activity (Morf et al., 2012). CLIP-Seq (Cross-linking and immunoprecipitation sequencing) experiments have shown Cirbp binds *Cdkn1a* mRNA in the 3' UTR (Morf et al., 2012). The reduced growth rate we observe in temperature shifted CHO cells may therefore be a direct result of *Cirbp* expression regulating cyclin genes to control the cell cycle.

Reducing the temperature of CHO cell cultures also impacted metabolic activity with an increase in glucose concentrations in temperature shifted cultures and a decrease in lactate levels. A hallmark of CHO cell energy metabolism in the exponential growth phase is a "Warburg" type phenotype in which cells inefficiently utilise glucose in glycolysis as the primary source of ATP and produce lactate as a toxic metabolic by-product (Buchsteiner et al., 2018). Temperature shift can reverse the Warburg phenotype and switches cells to consuming lactate and utilising OXPHOS as the primary source of energy metabolism - increased glucose and decreased lactate levels in the media suggests the cells in this experiment have undergone this "metabolic switch" (Buchsteiner et al., 2018). Differential expression was observed in genes which may contribute to this change in metabolism including lactate dehydrogenase genes (Li et al., 2016) and drivers of glycolysis (Salazar-Roa and Malumbres, 2017). The lactate dehydrogenase C (*Ldhc*) gene was among the 10 genes with the greatest fold change in gene expression with a 32.6 fold increase in expression induced by temperature shift. In addition the lactate dehydrogenase genes *Ldha*, *Ldhal6b* were upregulated (2.13 & 2.52 fold) and *Ldhd* was downregulated (-1.82 fold). Lactate dehydrogenases catalyse the reversible conversion between lactate and pyruvate (used to fuel the TCA cycle and OXPHOS). Alterations to lactate dehydrogenase gene expression results in reduced levels of anaerobic glycolysis (Li et al., 2016) and therefore differential expression of these genes is likely the major mechanism through which increased lactate consumption occurs. The phosphofructokinase gene (*Pfkm*) is a rate limiting enzyme in the glycolysis pathway which commits fructose to use in glycolysis by catalysing conversion to fructose 1,6-biphosphate (Tang et al., 2012). Upregulation of *Pfkm* has shown to induce anaerobic respiration and the accumulation of lactate while miRNA knockdown of *Pfkm* results in decreased lactate production and therefore the *Pfkm* gene is likely a regulator of metabolic switching (Tang et al., 2012). *Pfkm* was downregulated -1.5 fold in response to temperature shift which may contribute to the metabolic shift observed. Metabolic switch from glycolysis to OXPHOS is also regulated by adenosine monophosphate-activated protein kinase (Ampk) which phosphorylates and activates 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3 (Pfkfb3) during mitotic arrest (Doménech et al., 2015). The Ampk subunit genes *Prkab2* and *Prkag1* were identified as differentially expressed with a 1.59 and -1.78 fold change respectively. Prkag1 is the γ subunit of Ampk which senses energy levels through binding of AMP which changes the structure of Ampk and activates the catalytic domain to phosphorylate Pfkfb3 (Adams et al., 2004). Reduced expression of *Prkag1* may therefore reduce Ampk activity and prevent Pfkfb3 activity from inducing anaerobic glycolysis (Salazar-Roa and Malumbres, 2017).

Gene expression profiling using high resolution RNA-Seq to study the response to reduced temperature during CHO cell cultures has improved our understanding of alteration energy metabolism induced by temperature shift. Metabolic switch regulating genes including *Pfkm* and *Prkag1* have not been previously linked to the CHO cell response to reduced culture temperature. However, through proteomics analysis of matched samples carried out to identify differentially expressed proteins it is clear that gene level expression changes are not always conferred to the protein level. While only 2.75% of the genes profiled using RNA-Seq were confidently observed as expressed in proteomics (340 of the 12,291 genes identified in RNA-Seq > 1CPM) a lack of overlap between differentially expressed genes and proteins was apparent. Fold changes of 45% (153) of proteins detected by LC-MS/MS did not correlate with values from differential gene expression experiments indicating a high degree of post-transcriptional regulation of gene expression contributes to the response to mild hypothermia.

The cell surface glycoprotein Cd44 (*Cd44*) is a transmembrane receptor which is activated by multiple different ligands to induce cell signalling for pathways with roles in cell proliferation, adhesion and apoptosis protection (Chen et al., 2018). Cd44 can also activate the hypoxia inducible factor 1α (Hif1α) transcription factor (Chen et al., 2018) and increase glucose uptake (Pastò et al., 2014) – inducing a Warburg metabolic phenotype. *Cd44* was identified as upregulated with a 2.07 fold change in cells cultured at a reduced temperature. As growth rate and glucose uptake are down in temperature shifted cells this was an unexpected observation. Cd44 was however not identified as differentially expressed at the protein level using LC-MS/MS (1.18 fold change) indicating post-transcriptional regulation controls *Cd44* expression.

In this Chapter a number of potential targets for engineering enhanced CHO cells have been identified. The *Rbm3* gene was identified as one of the most highly upregulated genes and was the most upregulated protein in temperature shifted CHO cells. *Rbm3* is known to play a role in cold shock and acts in tandem to *Cirbp*, which shares similar protein structure (Zhu et al., 2016). *Cirbp* has been utilised previously as a genetic engineering target in CHO cells for enhancing productivity (Tan et al., 2008), however *Rbm3* has yet to be utilised as such. Overexpressing *Rbm3* may also result in enhanced productivity through regulating the splicing and translation of cold specific mRNAs (Zhu et al., 2016) which results in elongated stationary phase of culture and therefore yield. The *Cdkn1a* gene which regulates the cell cycle through cyclin inhibition (Caldon et al., 2009) was identified as one of the most upregulated genes in temperature shift. As a target of Cirbp (Morf et al., 2012) and a master regulator of cell cycle progression at the G1/S phase (Caldon et al., 2009), the *Ckdn1a* gene may be one of the main regulators of the temperature shift phenotype which results in a halting of the cell cycle. Inducing overexpression of *Cdkn1a* once cells have reached optimal cell density to transition cells into the stationary phase may enhance the beneficial effects of temperature shift and result in elongated cultures. The decrease in lactate accumulation presented in temperature shifted cells is beneficial as lactate has a negative impact on culture viability (Bort et al., 2010). One of the most highly downregulated genes in temperature shifted cells was the lactate dehydrogenase *Ldhd*. Knockout or knockdown of *Ldhd* may further reduce lactate production and further increase cellular viability due to decreased media acidification. In addition, the *Pfkm* and *Prkag1* genes which induce a Warburg metabolism (Salazar-Roa and Malumbres, 2017; Tang et al., 2012) were identified as downregulated in temperature shifted cells. Decreased expression of these genes at the start of the stationary phase could be utilised to revert cells from a Warburg metabolism to the primary use of oxidative phosphorylation and increased energy availability for recombinant protein production. Each of these gene targets have not been previously attributed to temperature shift in CHO cells and therefore this experiment has improved our understanding and enabled identified of novel genetic engineering targets.

Comparisons of gene and proteomics expression data has revealed a high degree of post-transcriptional regulation of gene expression in addition to regulation at the level of transcription. While this experiment was informative in discovering the degree of post-translational regulation which occurs, this methodology is limited when it comes to discovering the mechanisms through which post-transcriptional regulation is occurring. A reliance on mass spectrometry proteomics also means this experiment suffers from low resolution of the protein landscape as only 3% of the genes observed to be expressed at the gene level were confidently identified with a minimum of 2 peptides at the protein level. Altered expression of splice factors indicate temperature shift induced post-transcriptional regulation of gene expression may be contributed to by alternative splicing (Wang et al., 2015). In addition the lack of correlation between mRNA and protein levels indicates regulation of translation efficiency contributes to the temperature shift response through mechanisms such as miRNA activity (Catalanotto et al., 2016). Profiling gene expression through microarray or RNA-Seq and measurements of the proteome through mass spectrometry alone are insufficient to profile the mechanism of post-transcriptional regulation of gene expression. Alternative methods must therefore be applied such as the profiling of alternative splicing, the activity of miRNAs and to measure translational efficiency in order to obtain a complete overview of the regulation of gene expression in the temperature shift response which we will explore in the following chapters throughout this thesis.

2.5 Conclusion

In conclusion, presented in this Chapter are the results of differential gene and protein expression on CHO cells exposed to mild hypothermia. Throughout this experiment an effective pipeline was developed for the analysis of RNA-Seq data and utilised to identify significant differential expression in 1,953 genes. The findings of this experiment demonstrate the model used is representative of the phenotype achieved by temperature shift in industrial cell culture – with reduced growth rates and a metabolic shift from Warburg to OXPHOS metabolism. Differential gene expression enhanced our understanding of the control of energy metabolism genes in temperature shifted CHO cells - with *Pfkm*, *Prkab2*, and *Prkag1* regulation linked to mild hypothermia conditions for the first time. Integration of mass spectrometry proteomics results revealed a lack of correlation between transcript and protein levels for 45% of detected genes. The lack of correlation between gene and protein expression levels indicates a high degree of post-transcriptional control of gene expression is responsible for regulating the temperature shift response in CHO cells. Further in-depth analysis of this dataset using other algorithms may have the potential to undercover more transcription patterns associated with the CHO differential expression response such as alternative splicing and the role of ncRNAs. A hypothesised role of Cirbp and Rbm3 activity in the temperature shift response is the regulation of splicing. Alternative splicing is therefore a particularly interesting subject and as such is the focus of Chapter 3.

Chapter 3

Analysis of temperature induced alternative splicing in CHO cells

3.1 Summary

In Chapter 2 widespread differential gene expression induced by temperature shift in CHO cells in growth, apoptosis and energy metabolism genes was identified. Mass spectrometry based proteomics revealed that gene expression changes were not always correlated to protein abundance. In this chapter the roles of alternative splicing and differential transcript usage are explored using the RNA-Seq described in Chapter 2. To maximise the identification of these events novel transcript isoforms and gene loci were identified using genome guided transcriptome assembly.

3.2 Introduction

Alternative splicing is a post-transcriptional process which results in the generation of multiple different transcripts from a single gene by the skipping of exons, retention of introns, or use of alternative splice or polyadenylation sites during pre-mRNA processing (Wang et al., 2015). Splicing was first observed by purifying mRNA fractions encoding adenovirus hexon polypeptide using electrophoresis and observing differing lengths of mRNA hybridising to its corresponding DNA fragments using electron microscopy (due to the absence of introns) (Berget et al., 1977). Technology to profile splicing has significantly progressed since this time to improve throughput and information gained. EST sequencing was used to generate large libraries of cDNAs which were used in early bioinformatics analyses to cluster mRNAs according to their genomic location (Modrek et al., 2001). Meta-analysis of EST libraries enabled the identification of over 6,200 alternative splicing events in a single experiment (Modrek et al., 2001) as well as 667 tissue-specific isoforms (Xu et al., 2002). The scale at which splicing occurs in mammalian genomes was however not realised until the application of deep RNA-Seq to globally profile the transcriptome - with over 95% of multi-exonic genes expressing at least 2 transcript isoforms (Pan et al., 2008). Alternative splicing increases the complexity of the mammalian transcriptome enabling, for example, the expression of over 90,000 diverse proteins from ~20,000 protein coding genes in Humans (Nilsen and Graveley, 2010). The most extreme reported case of alternative splicing is in the *Drosophila* DSCAM gene, where alternative splicing of 4 variable mutually exclusive exon clusters gives rise to over 38,000 potential different isoforms to generate diverse neuron receptors (Schmucker et al., 2000). Alternative splicing is also known to occur in CHO cells, with as many as 10 transcript isoforms documented for a single gene in the Ensembl genome annotation (*Gstm1*, Ensembl ID: ENSCGRG00000000476).

Splicing of pre-mRNAs expressed from multi-exonic genes is catalysed by the spliceosome, one of the largest macromolecules in eukaryotic cells, which consists of 5 non-protein-coding RNAs (snRNPs, small nuclear ribonucleo proteins) and >150 proteins (Cieply and Carstens, 2015). The spliceosome assembles on pre-mRNA by the binding of snRNPs to the 5’ splice site and 3’ splice site (donor and acceptor site, at the end of and beginning of neighbouring exons respectively) which bring the exons in close proximity for removal of the intron). The consensus sequences of the donor and acceptor splice sites can vary in their strength (the homology of the sequence to the consensus) and therefore weak sites require the binding of splicing factors to intronic or exonic splicing enhancer sequence elements to regulate the efficiency of spliceosome assembly (or conversely the efficiency of splicing can be reduced by the binding of splice factors to intronic or exonic splicing silencer elements) (Cieply and Carstens, 2015). Expression of splicing factors in a particular tissue, cell type or stage of development therefore allows for the control of splicing to regulate the dominant transcript isoform in a cell by regulating the rates of exon skipping, intron inclusion, or use of alternative splice sites (Grosso et al., 2008). Alternative splicing can also be induced in response to stimuli including temperature changes - such as the skipping of exons 6 and 7 reported in the U2AF26 splice factor in response to 32°C cold shock in mouse N2A cells mediated by the SRSF2 and SRSF7 splice factors (Preußner et al., 2017).

Both EST sequencing and microarray profiling technologies have been utilised previously for the study of alternative splicing events (Bumgarner, 2013) however each of these techniques have disadvantages compared to RNA-Seq. EST sequencing provides no quantification data on the mRNA which means rare isoforms with non-biologically relevant expression levels and incompletely processed mRNAs cannot be distinguished from highly expressed functional isoforms (Lee and Roy, 2004). Conversely, microarrays can be utilised to obtain transcript expression data but cannot be used to retrieve the nucleotide sequence of novel variants and cannot distinguish similar splice variants from each other (Bumgarner, 2013). To distinguish different isoforms from each other using microarrays it is possible to design probes targeting every exon-exon junction which can possibly splice together in a given transcript, but for complex genes with many exons this would be an exhaustive process (Bumgarner, 2013). Microarrays targeting every possible exon-exon junction combination would also be unable to capture subtle changes to the sequence of a spliced transcript such as inclusion of a small number of nucleotides due to the usage of an alternative splice site. Microarrays have been designed and utilised for studying alternative splicing which target exons and exon junctions in panels of genes (Su et al., 2008) and commercial microarrays are available such as the GeneChip 2.1 which can capture exon level measurements on >418,000 human exons (Affymetrix).

RNA-Seq overcomes the drawbacks of both EST sequencing and microarrays enabling simultaneous quantification and identification of every transcript expressed in a sample when sufficient sequencing depth is utilised (Bryant et al., 2012). RNA-Seq reads aligned to a genome can be quantified on a transcript isoform level using an approach known as “isoform deconvolution” in which the abundance of reads crossing exon:exon junctions unique to a given transcript are used to proportionally distribute reads aligned to regions shared by multiple transcripts (Trapnell et al., 2010). Novel exon:exon junctions are utilised to simultaneously discover unannotated transcript isoforms. One challenge of using RNA-Seq reads for the assembly of isoforms is the rates of false positives generated by many of the algorithms which may arise due to incomplete splicing, partially degraded transcripts or alignment errors and incorporated into novel transcripts erroneously (Soneson et al., 2015). New algorithms such as *Stringtie* are however improving the specificity and sensitivity of transcript assembly (Pertea et al., 2015). Expression abundances estimated by isoform deconvolution can be used to calculate differential expression at the whole gene, individual transcript or exon level. Comparing the results of genes with differentially expressed transcripts with gene level differential expression results enables the identification of “isoform switching” – a change in the proportions of the expressed isoforms of a gene without necessarily affecting the overall gene expression (Vitting-Seerup and Sandelin, 2017). Changes in the proportions of expressed transcript isoforms of a gene, resulting from differential alternative splicing, can result in dramatic and even antagonistic effects on the functionality of the translated protein. For example isoform switching of the *Bcl-x* gene results in a switch from pro-apoptosis to anti-apoptosis and is often a hallmark of cancers (Merdzhanova et al., 2008). Although anti-apoptotic functionality and rapid cellular replication are cancer causing traits in human cells, these are advantageous for CHO cells in the context of biopharmaceutical production in order to increase batch culture duration and to reach optimal cell density quickly.

Alternative splicing has not yet been extensively studied in CHO cells, however previous transcriptome sequencing experiments have resulted in a collection of transcript variants which have been deposited into genome annotation databases (Birzele et al., 2010; Lewis et al., 2013; Rupp et al., 2014). No study to date has yet made use of these existing transcript variants to study the contribution of alternative splicing to a phenotype in CHO cells. One previous RNA-Seq study on temperature shifted CHO cells used the transcripts annotated in the CHO-K1 genome for quantifying aligned read counts prior to differential expression analysis, however their results were only analysed in the context of the individual gene and no insight into observed alternative splicing or differential transcript expression was reported (Bedoya-López et al., 2016). Other RNA-Seq experiments have also been performed on CHO cells (Chen et al., 2017; Chung et al., 2013; Hsu et al., 2017; Könitzer et al., 2015; Vishwanathan et al., 2014; Yuk et al., 2014) however each of these experiments analysed their data at the level of gene rather than the transcript and therefore potentially missed the presence of differential transcript usage and isoform switching in their results (Vitting-Seerup and Sandelin, 2017).

In this chapter the RNA-Seq data from Chapter 2 was re-analysed in order to identify changes in the isoform population induced in response to the culture temperature. In addition splicing events were classified on at the level of exons to infer potential biological effects of observed alternative splicing.

3.3 Methods

3.3.1 Genome-guided assembly

The ENSEMBL reference genome annotation release 93 for CriGri_1.0 was used as an annotation for genome guided assembly with *Stringtie* (Pertea et al., 2015). Each of the 8 samples aligned reads in the sorted BAM format (generated in Chapter 2) were individually assembled into transcripts, requiring a minimum junction coverage of 5 aligned reads to accept a novel splice junction. The resulting transcripts from each sample were merged with the reference GTF annotation file using the '*Stringtie --merge'* function to generate a comprehensive GTF file containing both the existing annotation as well as novel transcripts assembled from our RNA-Seq data. Only novel transcripts with a minimum isoform fraction of 0.1 (i.e. those transcripts detected ≤10% of the expression of the most abundant transcript of a particular gene) and an average coverage of 10 reads were retained.

3.3.2 Differential transcript expression analysis

A second pass of *Stringtie* using the comprehensive merged annotation GTF file was used to quantify the expression of each transcript (Pertea et al., 2015). The '-e' parameter was used to disable the assembly of new transcripts, only estimating the abundance of transcripts contained within the annotation file. The '-B' parameter was specified to generate tables containing expression values for each transcript. FPKM (fragments per kilobase of transcript per million mapped reads) normalised expression values for each transcript were used in *ballgown* to calculate differential transcript expression between the temperature-shifted and non-temperature-shifted conditions (Frazee et al., 2015). Transcripts with a fold change exceeding ≥±1.5 and a Benjamini-Hochberg (BH) adjusted P value ≤ 0.05 were identified. Gene ontology enrichment analysis was also carried out on genes with differentially expressed transcripts using *DAVID* (Ashburner et al., 2000).

3.3.3 Characterisation of alternative splicing events

rMATS (Shen et al., 2014) was used to identify and classify alternative splicing events in genes contained in the Ensembl defined annotation as well as novel isoforms identified using the *Stringtie* assembly. *rMATS* parameters were configured for input of paired reads of length 150bp with a fr-secondstrand library type. Following completion of rMATs only those splicing events identified by assessment of reads spanning the annotated splice junctions were considered (events identified from both junctions reads and reads on the target exon were not included). Skipped exons, alternative 5' and 3' splice sites, retained introns and mutually exclusive exon splicing events were considered to be significantly different if a ≥ 10% difference in percent spliced in (PSI) inclusion rate and < 0.05% P value was observed. Sashimi plots were generated for differential exon usage in genes of interest using *rmats2sashimi* (Shen et al., 2014). Protein domains overlapping skipped exons were identified using the *biomaRt* Bioconductor R package (Durinck et al., 2009) to retrieve Intropro protein domain information and overlapping with the co-ordinates of skipped exons using the *GenomicRanges* R package (Lawrence et al., 2013). An overview of the bioinformatics pipeline used in this experiment is depicted in Figure 3.1.

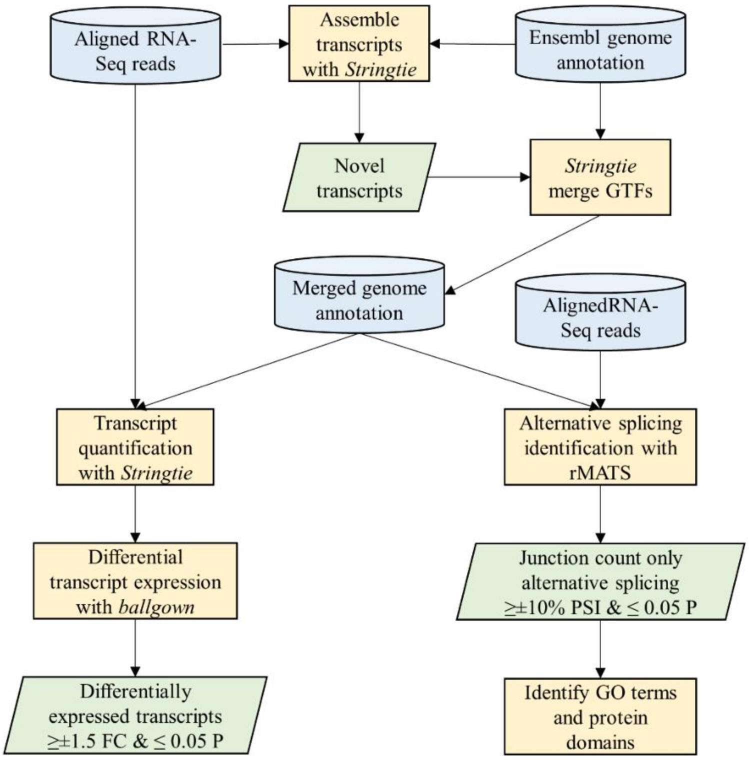
Aligned RNA-Seq reads
Assemble transcripts with Stringtie
Ensembl genome annotation
Novel transcripts
Stringtie merge GTFs
Merged genome annotation
AlignedRNA-Seq reads
Transcript quantification with Stringtie
Alternative splicing identification with rMATS
Differential transcript expression with ballgown
Junction count only alternative splicing ≥±10% PSI & ≤ 0.05 P
Differentially expressed transcripts ≥±1.5 FC & ≤ 0.05 P
Identify GO terms and protein domains

Figure 3.1 Pipeline for the identification of differentially expressed transcripts and alternative splicing. Aligned RNA-Seq data from Chapter 2 was used as input to *Stringtie* to assemble transcripts which were merged with the reference genome annotation. Transcript expression was quantified with a 2nd pass of *Stringtie* and *ballgown* was used to identify transcript level differential expression with ≥±1.5 fold change and ≤0.05 Benjamini-Hochberg adjusted P value. The merged genome annotation was used in *rMATs* to identify alternative splicing with a 10% difference in percent spliced in between conditions and ≤0.05 adjusted P value. GO terms of interest and protein domains were identified in genes with alternative splicing using *biomaRt* and the Interpro database.

3.3.4 qPCR validation

CHO 16F cells, CHO-K1 cells and CHO-DP12 cells were cultured in 6 replicate samples in 20ml of CHO-S-SFM II media (Thermo Fisher Scientific) in 250ml shaker flasks at 37°C. After 48hrs the temperature of 3 samples was reduced to 31°C. 24hrs later cells were harvested from each culture. PrimerQuest (IDT) was used to design primers using default parameters on regions of transcripts unique to a given transcript isoform. Where possible, primers were designed to cross splice junctions unique to a given transcript. Primer-Blast (Ye et al., 2012) was used to assess primer designs for self-complementarity and specificity. Further specificity checks were carried out using BLAST searches against the assembled transcriptome. The primers used for this analysis are listed in Appendix 3O. Primers were validated with melting curve analysis, electrophoresis and Sanger sequencing of qPCR products. Primer sequences used for each transcript are shown in Appendix 3. The *Actb* and *Gusb* genes were included in qPCR for normalization and control purposes.

Total RNA was extracted from cell pellets using Trizol (Invitrogen) and contaminating DNA degraded using DNase I (Sigma). A Nanodrop 1000 (Thermo Scientific) instrument was used to assess the purity and quality of RNA, before cDNA generation using a High-capacity cDNA Reverse Transcription Kit (ThermoFisher Scientific). Real Time PCR was carried out on samples in technical triplicate for each biological replicate using an AB7500 instrument and Fast SYBR® Green Master Mix (Applied Biosystems). Negative controls of mRNA from each gene with no reverse transcriptase and water controls were included to monitor contamination by genomic DNA or introduced during sample preparation. The *Cirbp* gene was included as an internal control to ensure cultures had responded to temperature shift. Ct values for each transcript isoform were quantified using the $2^{-\Delta\Delta}$Ct method (Livak and Schmittgen, 2001) and normalised using *Actb* and *Gusb.* A two-tailed t-test was used to identify differences in dCT values between conditions with a statistical significance threshold of ≤0.05. Cell culture, primer validation and qPCR was carried out by Dr Ioanna Tzani in Dublin City University.

3.4 Results

3.4.1 Genome guided transcriptome assembly

Stringtie was used to assemble novel transcripts following alignment of RNA-Seq reads to the CHO-K1 genome. Assembly resulted in the identification of 27,687 novel transcripts and 10,019 novel gene loci (locations on the genome where a transcript was previously unannotated) including both protein coding and non-protein coding sequences. Figure 3.2 illustrates the number of genes in both the original Ensembl reference annotation and the assembled genome annotation bringing the total annotated sequences to 40,692 gene loci and 62,159 transcripts. The increase to the number of genes in the assembled annotation with 2+ transcript isoforms per gene locus. An example of a novel annotated transcript region is shown in Figure 3.3.

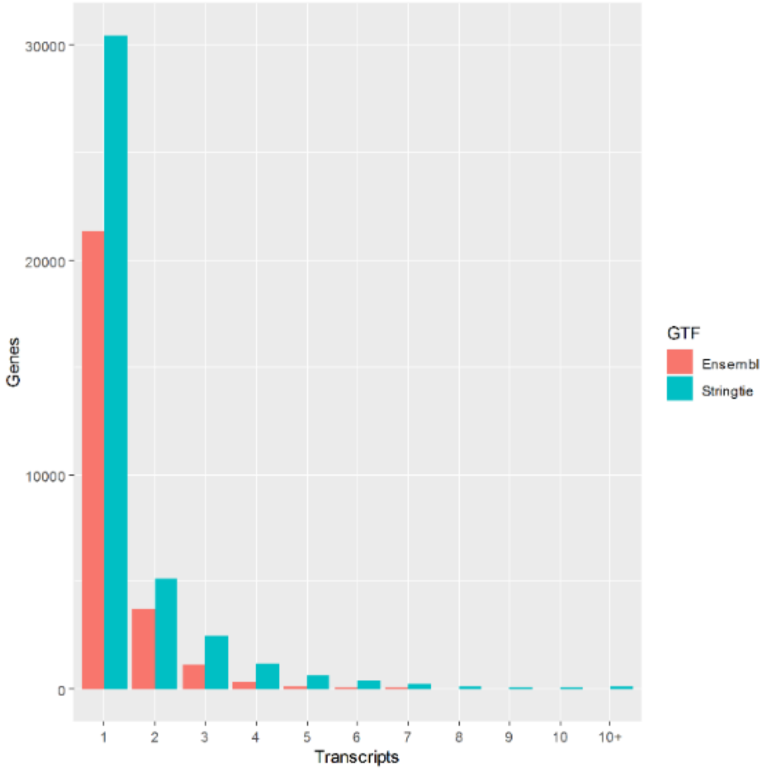

Figure 3.2 –Assembled and reference genome annotation statistics. The number of genes in the Ensembl reference genome annotation (red) and the assembled annotation generated using *Stringtie* (turquoise) are plotted in this figure. Genes are categorised by the number of transcripts per gene locus. 10,019 novel gene loci and 27,687 novel transcript isoforms were assembled.

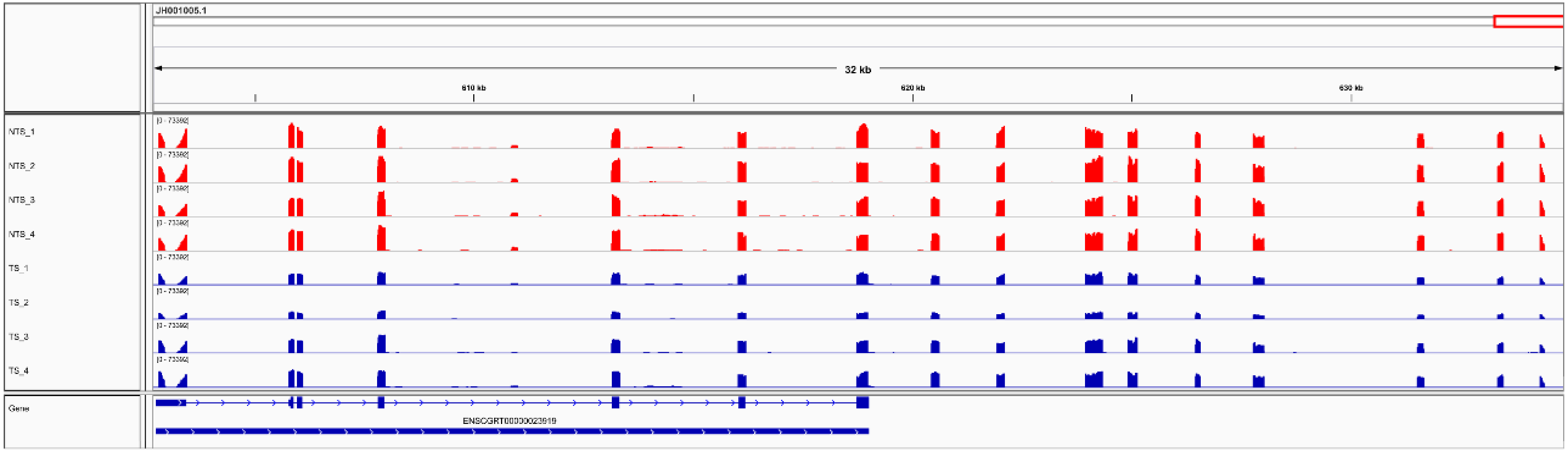

Figure 3.3 –Alignment to the *Atm* gene. The Ensembl annotation for the *Atm* gene does not represent the full transcript isoform. The annotated gene region can be seen in the bottom track of this plot with a further nine exons identified using *Stringtie*.

3.4.2 Differentially expressed transcripts

1,166 transcripts were identified as upregulated and 1,277 downregulated at a BH adjusted p value ≤0.05 and ≥ ±1.5 fold change using *ballgown*. 1,105 gene loci had an upregulated transcript and 1,216 downregulated (Appendix 3A, Table 3.1). A total of 2,288 gene loci were identified as differentially expressed, a value less than the combined up and downregulated gene loci due to the presence of some genes in both the up and down lists. 1,381 genes with differential transcript expression were annotated with an Ensembl gene ID and used in gene ontology analysis to identify enriched ontologies and pathways with transcripts differentially expressed (Table 3.2, Appendix 3B).

Table 3.1 – Most significant differentially expressed transcripts. The top 10 most up and downregulated transcripts identified by *ballgown* are presented in this table. Previously unannotated genes are among this list (those with an MSTRG ID) and therefore incomplete data is denoted with an N/A.

Gene ID	**Gene Symbol**	**Gene Name**	**BH Adj. P**	**FC**
ENSCGRG00000009838	*Coro1c*	Coronin 1C	4.26×10^{-03}	14.26
ENSCGRG00000012558	*Hap1*	Huntingtin Associated Protein 1	1.46×10^{-03}	8.90
MSTRG.18664.1	N/A	N/A	3.37×10^{-03}	8.68
ENSCGRG00000008472	*Vps35*	Vacuolar Protein Sorting 35	2.61×10^{-02}	8.22
ENSCGRG00000013098	*Eda2r*	Ectodysplasin A2 Receptor	3.95×10^{-03}	8.09
ENSCGRG00000010025	*Fads6*	Fatty Acid Desaturase 6	1.58×10^{-02}	8.07
ENSCGRG00000015407	*Rgs12*	Regulator Of G Protein Signaling 12	1.32×10^{-02}	7.64
ENSCGRG00000011992	*Pdia6*	Protein Disulfide Isomerase Family A Member 6	3.72×10^{-03}	7.57
ENSCGRG00000005111	*Smg5*	SMG5, Nonsense Mediated MRNA Decay Factor	4.60×10^{-03}	7.34
MSTRG.14873.1	N/A	N/A	4.20×10^{-03}	7.27
ENSCGRG00000001881	*Dglucy*	D-Glutamate Cyclase	3.59×10^{-02}	7.08
ENSCGRG00000017132	*Efemp1*	EGF Containing Fibulin Extracellular Matrix Protein 1	4.58×10^{-02}	-4.09
ENSCGRG00000001676	*Pter*	Phosphotriesterase Related	4.40×10^{-02}	-4.21
MSTRG.26120.2	N/A	N/A	2.48×10^{-02}	-4.49
ENSCGRG00000013200	*Ywhah*	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Eta	3.59×10^{-02}	-4.99
ENSCGRG00000007933	*Naxd*	NAD(P)HX Dehydratase	3.14×10^{-02}	-5.25
ENSCGRG00000009282	*Kcnd1*	Potassium Voltage-Gated Channel Subfamily D Member 1	2.80×10^{-02}	-5.50
ENSCGRG00000016118	*Gga2*	Golgi Associated, Gamma Adaptin Ear Containing, ARF Binding Protein 2	1.66×10^{-03}	-6.20
ENSCGRG00000023559	*RF00100*	N/A	4.56×10^{-02}	-7.06
ENSCGRG00000016041	*Gphn*	Gephyrin	1.43×10^{-03}	-7.15
ENSCGRG00000004829	*Smox*	Spermine Oxidase	8.19×10^{-03}	-7.21
ENSCGRG00000015009	*Hist1h3f*	Histone Cluster 1 H3 Family Member F	6.94×10^{-03}	-9.11

Table 3.2 – Enriched GO biological processes among genes with differential transcript expression. DAVID was used to identify enriched GO terms among those transcripts that were altered upon reduction of cell culture temperature. The top 20 most significant GO terms with a BH adjusted P value ≤0.05 were included in this table.

GO ID	Term	Count	%	BH Adj. P
GO:0000278	mitotic cell cycle	171	13.55	2.14×10^{-31}
GO:0022402	cell cycle process	210	16.64	2.29×10^{-31}
GO:1903047	mitotic cell cycle process	161	12.76	$5.42 \times 10^{-3}1$
GO:0007049	cell cycle	234	18.54	5.45×10^{-29}
GO:0006259	DNA metabolic process	150	11.89	2.97×10^{-20}
GO:0006260	DNA replication	70	5.55	5.43×10^{-19}
GO:0000280	nuclear division	104	8.24	1.31×10^{-18}
GO:0044770	cell cycle phase transition	100	7.92	1.21×10^{-18}
GO:0051276	chromosome organization	163	12.92	2.2×10^{-18}
GO:0048285	organelle fission	107	8.48	3.84×10^{-18}
GO:0006261	DNA-dependent DNA replication	47	3.72	1.06×10^{-17}
GO:0006974	cellular response to DNA damage stimulus	125	9.90	1.03×10^{-17}
GO:0044772	mitotic cell cycle phase transition	94	7.45	1.36×10^{-17}
GO:0006281	DNA repair	93	7.37	3.52×10^{-16}
GO:0007067	mitotic nuclear division	82	6.50	3.43×10^{-16}
GO:0007059	chromosome segregation	68	5.39	1.16×10^{-14}
GO:0071103	DNA conformation change	61	4.83	4.61×10^{-14}
GO:1902589	single-organism organelle organization	187	14.82	1.74×10^{-13}
GO:0033554	cellular response to stress	200	15.85	2.47×10^{-13}
GO:0051301	cell division	92	7.29	2.35×10^{-13}

3.4.3 Identification of Isoform Switching

Genes with differentially expressed transcript isoforms were compared with lists of genes which were identified as differentially expressed at the gene level using *DESeq2* (Chapter 2). From the analysis conducted in this chapter 597 differentially expressed transcripts transcribed from previously annotated genes (present in the Ensembl genome annotation) were not identified as differentially expressed at the gene level (Appendix 3D). 351 of these transcripts had upregulated and 246 downregulated expression. 37 transcripts identified as differentially expressed with no gene level change were annotated with GO terms of interest (Appendix 3E, Table 3.3). Figures 3.4 illustrates transcripts of the *Ybx3* gene which was identified as differentially expressed at the transcript level only. A cold specific transcript of *Ybx3* was expressed which does not have an intron retention event.

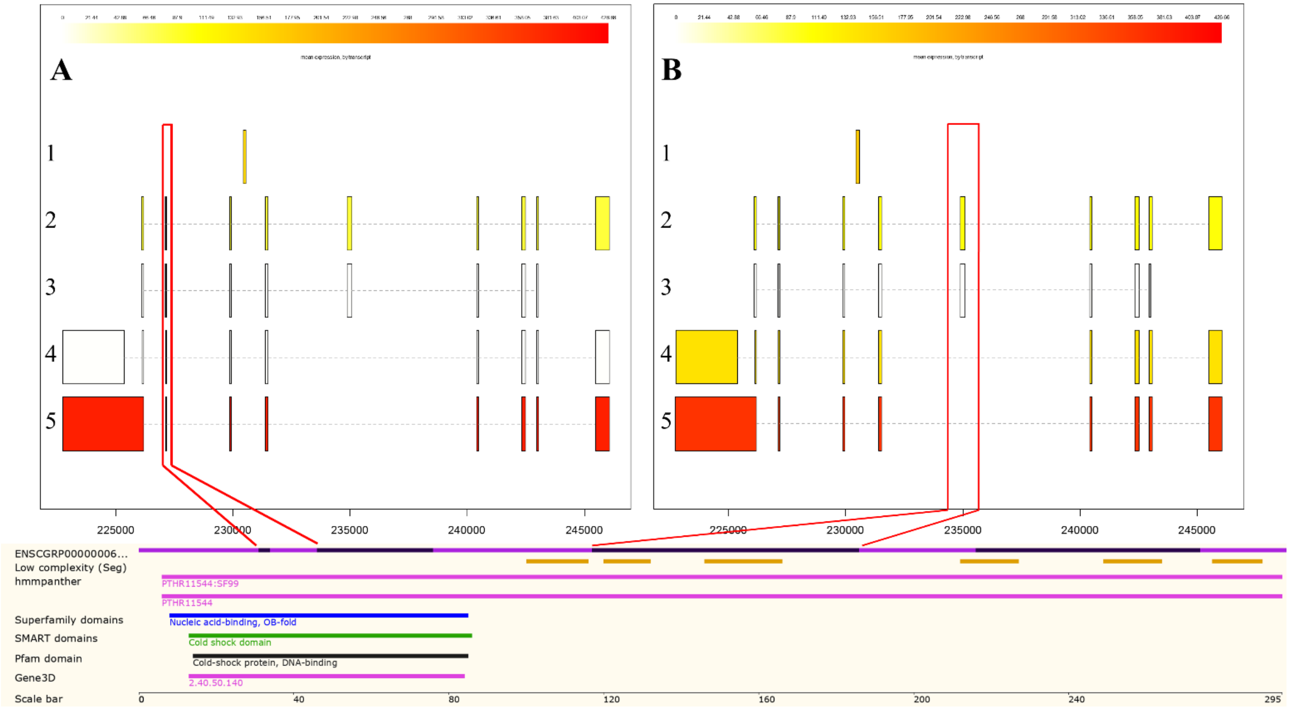

Figure 3.4 Differential transcript usage in the *Ybx3* gene. The *Ybx3* gene was identified as differentially expressed at the transcript level but not the gene level. (A) The mean coverage of non-temperature shifted transcript isoforms. (B) The mean coverage of temperature shifted samples. Temperature shift induced expression of a cold specific transcript isoform (4) which does not have an intron retention event between the first and second exons and skips exons 3 and 7 (red boxes in A and B respectively). The amino acid sequence encoded by ex on 3 of the *Ybx3* transcript is part of a cold shock DNA-binding domain. *Ybx3* is transcribed from the sense strand and therefore exon 1 is on the left of this plot (ENSCGRG00001016148).

Table 3.3 – Genes with significant differential transcript expression which were not identified as differentially expressed at the gene level. RNA-Seq data analysis using *Stringtie* and *Ballgown* resulted in the identification of 597 differentially expressed transcripts which were not identified as differentially expressed at the gene level using *DESeq2* in Chapter 2. Presented in this table are a selection of 10 differentially expressed transcripts with roles in apoptosis, regulation of oxidative phosphorylation, response to cold and glucose metabolism GO annotations.

GeneID	Gene Name	Transcript ID	Bh Adj. P	Fold Change	GO Terms
ENSCGRG00000016072	*Dnm1l*	MSTRG.22559.5	0.014	1.63	GO:0001836 RELEASE OF CYTOCHROME C FROM MITOCHONDRIA, GO:0043065 POSITIVE REGULATION OF APOPTOTIC PROCESS
ENSCGRG00000001056	*Slc25a33*	MSTRG.20029.2	0.01	3.6	GO:0002082 REGULATION OF OXIDATIVE PHOSPHORYLATION
ENSCGRG00000007974	*Nipsnap2*	ENSCGRT00000011122	0.01	-1.56	GO:0006119 OXIDATIVE PHOSPHORYLATION
ENSCGRG00000004563	*Ybx3*	MSTRG.19515.2	0.001	Infinite	GO:0009409 RESPONSE TO COLD, GO:0043066 NEGATIVE REGULATION OF APOPTOTIC PROCESS
ENSCGRG00000003763	*Pdk3*	MSTRG.13664.2	0.011	-1.56	GO:0010906 REGULATION OF GLUCOSE METABOLIC PROCESS
ENSCGRG00000010104	*Rnps1*	ENSCGRT00000019526	0.039	1.61	GO:0043065 POSITIVE REGULATION OF APOPTOTIC PROCESS
ENSCGRG00000002300	*Ddx20*	MSTRG.21486.1	0.0073	-1.55	GO:0043065 POSITIVE REGULATION OF APOPTOTIC PROCESS
ENSCGRG00000017082	*Atm*	MSTRG.17450.1	0.047	-1.54	GO:0043065 POSITIVE REGULATION OF APOPTOTIC PROCESS
ENSCGRG00000010377	*Arnt2*	MSTRG.23256.2	0.026	-1.54	GO:0043066 NEGATIVE REGULATION OF APOPTOTIC PROCESS
ENSCGRG00000017539	*Prkca*	MSTRG.8880.1	0.016	-1.83	GO:0097190 APOPTOTIC SIGNALING PATHWAY

In total, 33 cases were identified where both a transcript was observed as upregulated and another transcript identified as downregulated which were transcribed from the same gene. 24 of these genes were annotated with an Ensembl GeneID and were not identified as differentially expressed at the gene level, indicating the presence of isoform switching (Appendix 3F, Table 3.4). An example of isoform switching in the *Acp1* gene is depicted in Figures 3.5 and 3.6.

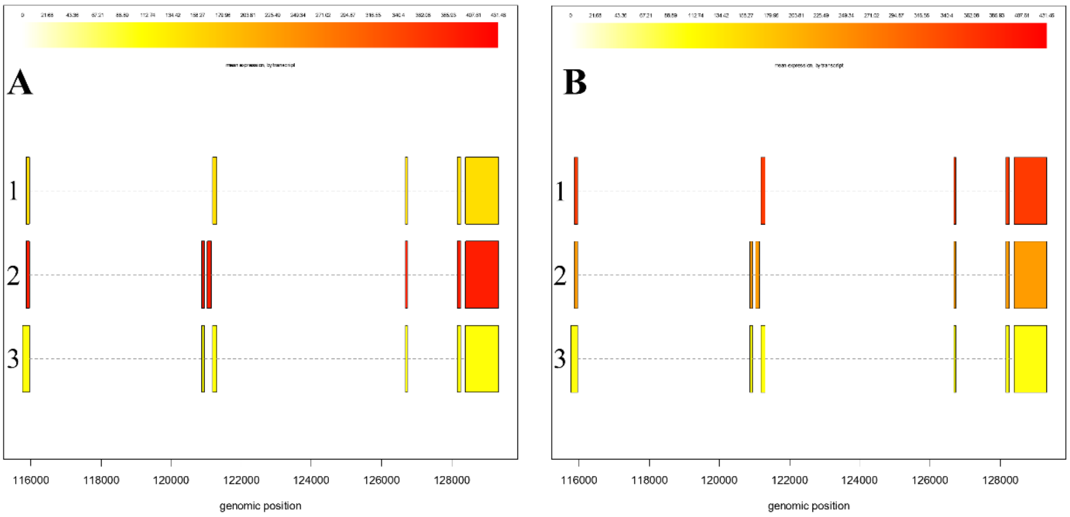

Figure 3.5 – Isoform Switching in the *Acp1* gene. The *Acp1* gene had both an upregulated and downregulated isoform found differentially expressed. (A) Non-temperature shifted cells expressed an isoform containing exon 2 and 3 at a higher level. (B) Temperature shift induced expression of an isoform which skips exon 2 and 3. *Acp1* is transcribed from the sense strand of the genome and therefore exons are numbered from left to right.

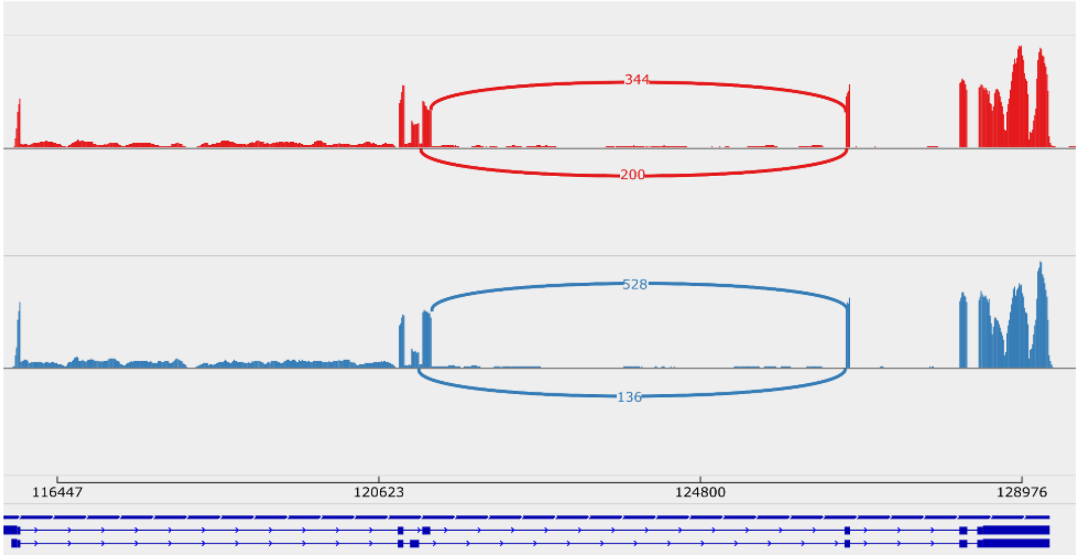

Figure 3.6 – Isoform switching in transcripts with mutually exclusive exons in the *Acp1* gene. Temperature shift resulted in isoform switching in the *Acp1* gene with upregulation of a transcript isoform containing exon 4 and downregulation of an isoform containing exon 3. Counts for exon junctions other than 3:5 and 4:5 are omitted for clarity.

Table 3.4 - Isoform switching. 34 genes were identified with a gene level fold change below 1.5 but at least 1 transcript isoform significantly differentially expressed in both the up and downregulated fold change directions. A selection of genes undergoing this isoform switching are presented in this table.

GeneID	Gene Symbol	Gene Name	Base Mean	DESeq2 FC	DESeq2 BH P	Transcript ID	Ballgown FC	Ballgown BH P
ENSCGRG00000000521	*Eps8*	Epidermal Growth Factor Receptor Pathway Substrate 8	16,996	1.34	4.51×10^{-32}	MSTRG.7794.2	4.78	4.15×10^{-2}
						MSTRG.7794.1	2.11	3.40×10^{-2}
						MSTRG.7794.5	-2.11	3.70×10^{-2}
ENSCGRG00000000960	*Siah1*	Siah E3 Ubiquitin Protein Ligase 1	1,391	1.43	1.16×10^{-16}	ENSCGRT00000001283	26.89	3.35×10^{-2}
						MSTRG.16790.1	-3.17	6.00×10^{-3}
ENSCGRG00000003711	*Gatd3a*	Glutamine Amidotransferase Like Class 1 Domain Containing 3A	1,634	-1.19	7.58×10^{-5}	MSTRG.7372.2	2.27	2.18×10^{-2}
						MSTRG.7372.6	2.16	1.72×10^{-2}
						MSTRG.7372.1	-1.68	9.35×10^{-3}
ENSCGRG00000005111	*Smg5*	SMG5, Nonsense Mediated MRNA Decay Factor	3,415	-1.37	1.94×10^{-24}	MSTRG.10246.1	161.58	4.60×10^{-3}
						ENSCGRT00000007145	-2.21	5.78×10^{-3}
ENSCGRG00000006394	*Gga1*	Golgi Associated, Gamma Adaptin Ear Containing, ARF Binding Protein 1	963	-1.26	2.47×10^{-8}	ENSCGRT00000008893	10.27	2.80×10^{-2}
						MSTRG.19185.1	-1.67	4.25×10^{-2}
ENSCGRG00000006500	*Cers6*	Ceramide Synthase 6	1,049	1.23	6.69×10^{-7}	MSTRG.1243.1	7.87	3.29×10^{-3}
						ENSCGRT00000009035	-2.37	3.41×10^{-2}
ENSCGRG00000009339	*Ubqln1*	Ubiquilin 1	7,013	-1.14	4.29×10^{-6}	MSTRG.21216.1	1.56	1.89×10^{-3}
						ENSCGRT00000013032	-1.78	2.38×10^{-2}
ENSCGRG00000012304	*Acp1*	Acid Phosphatase 1	3,571	1.14	1.87×10^{-5}	MSTRG.8383.2	2.53	8.15×10^{-3}
						ENSCGRT00000017163	-1.61	4.81×10^{-2}
ENSCGRG00000013039	*Bud23*		2,438	-1.33		MSTRG.7804.1	6.9	8.22×10^{-3}

		BUD23, RRNA Methyltransferase And Ribosome Maturation Factor			1.45×10^{-17}	ENSCGRT00000018195	-1.57	1.36×10^{-2}
ENSCGRG00000013201	*Haus7*	HAUS Augmin Like Complex Subunit 7	2,027	-1	8.56×10^{-1}	MSTRG.14637.1	3.35	3.41×10^{-2}
						ENSCGRT00000018414	-1.78	8.27×10^{-3}
ENSCGRG00000016118	*Gga2*	Golgi Associated, Gamma Adaptin Ear Containing, ARF Binding Protein 2	1,301	-1.44	7.72×10^{-19}	MSTRG.23841.2	48.51	5.85×10^{-4}
						MSTRG.23841.1	-73.34	1.66×10^{-3}
ENSCGRG00000016703	*Aldh6a1*	Aldehyde Dehydrogenase 6 Family Member A1	1,815	-1.2	1.07×10^{-7}	MSTRG.13065.1	2.14	4.26×10^{-2}
						ENSCGRT00000023382	-3.93	3.29×10^{-3}
ENSCGRG00000017169	*Trip12*	Thyroid Hormone Receptor Interactor 12	24,950	1.12	6.04×10^{-9}	ENSCGRT00000024026	1.56	1.49×10^{-2}
						MSTRG.20560.7	-1.68	1.32×10^{-2}
						MSTRG.20560.4	-4.59	4.73×10^{-3}
ENSCGRG00000017660	*Cib2*	Calcium And Integrin Binding Family Member 2	451	-1.17	8.05×10^{-3}	MSTRG.3798.2	1.86	4.91×10^{-2}
						ENSCGRT00000024694	-1.93	2.85×10^{-2}

3.4.4 Identification and classification of splicing events

3,158 differential exon usage events were identified with an adjusted p value < 0.05 and a minimum of 10% change in inclusion level (the difference in prevalence of splicing events between conditions) (Figure 3.7, Appendix 3G & 3H). Skipped exon events were the most prevalent type of alternative splicing observed and account for 61.72% of splicing events.

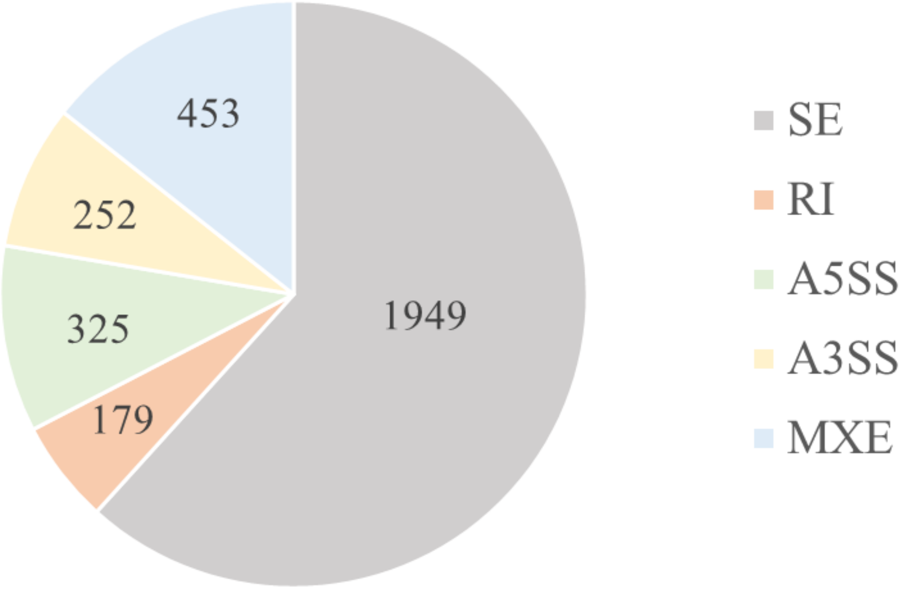

Figure 3.7 – Differential exon usage events identified by rMATs. 3,158 splicing events were identified with a ≥±10% difference in inclusion level between conditions and a ≤0.05 adjusted P value. A3SS= Alternative 3’ Splice Site, A5SS = Alternative 5’ Splice Site, MXE = Mutually Exclusive Exons, RI = Retained Introns, SE = Skipped Exons.

2,467 of the differential splicing events identified occurred in genes which were not identified as differentially expressed at the gene level (Appendix 3I & 3J). An exon skipping event was identified in the *Slirp* gene which was upregulated in response to temperature shift. A transcript isoform with this exon skipping event was not annotated prior to this experiment in the CHO cell genome. Transcripts of the *Slirp* gene are depicted in Figure 3.8.

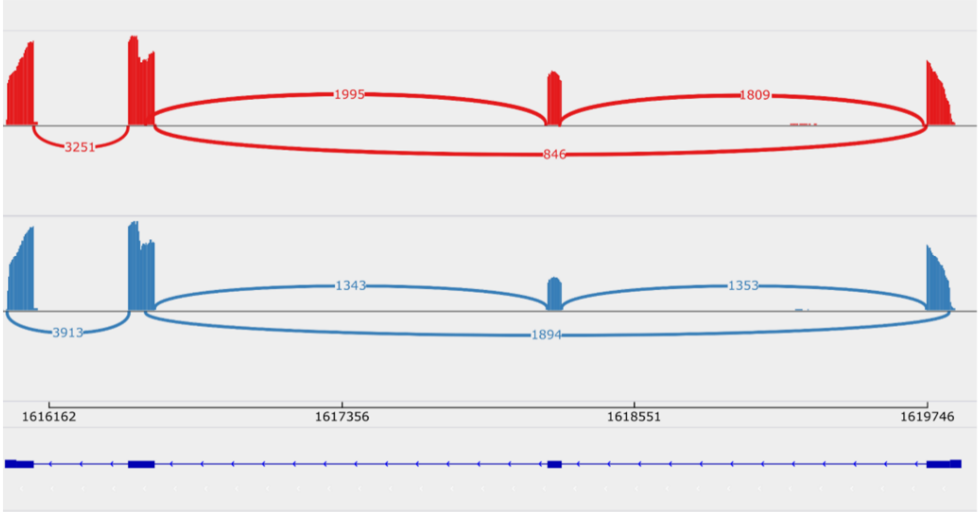

Figure 3.8 Temperature shift induced skipping of exon 2 in the *Slirp* gene. Temperature shift induced exon skipping in the *Slirp* gene. Sashimi plots display the number of reads which align across exon:exon junctions. In the non-temperature shift condition (red) the exon 1 and 2 junctions as well as the exon 2 and 3 junctions had over 1,800 reads aligning, more than twice the 864 which align to the 1:3 junction. In temperature shift (blue) the exon 1:3 junctions have higher support than the 1:2 and 2:3 junctions.

The gene with the most differentially expressed splicing events was *Cd44*. 8 transcript isoforms for *Cd44* were present in the Ensembl genome annotation and no new isoforms were assembled however exons 6-15 were extensively spliced with 9 skipped exon and 11 mutually exclusive exon events identified (Figure 3.9).

862 of the genes with skipped exons which were not identified as differentially expressed at the gene level had an annotated protein domain in InterPro (Appendix 3L). A skipped exon event in *Dnm1l* was identified as overlapping a GTPase domain in the transcripts encoded amino acid sequence (Figure 3.10).

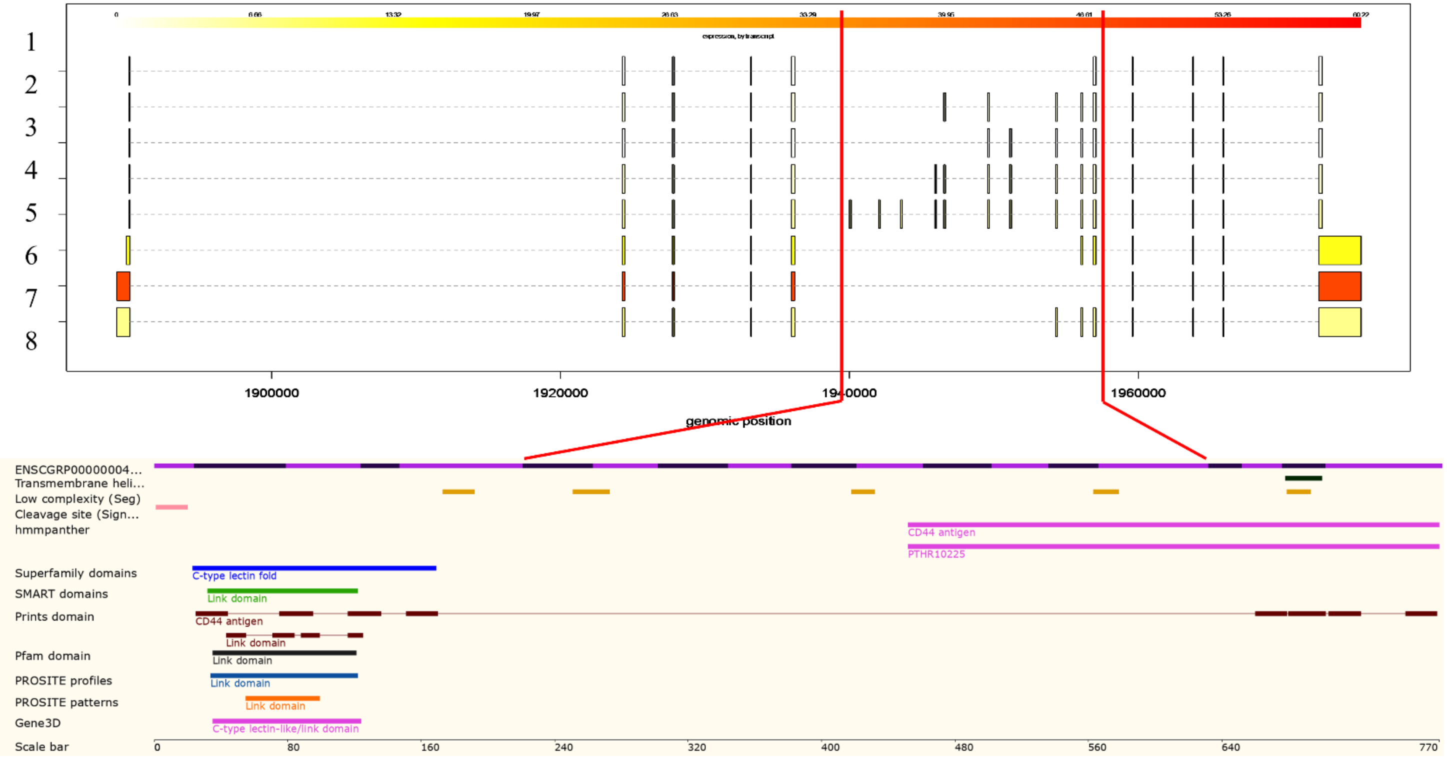

Figure 3.9 - The complex splicing of the *Cd44* gene. 8 transcript isoforms were identified for the *Cd44* gene. Transcript 5, representing full length *Cd44*, was identified as upregulated in reduced culture temperature. Alternative splice isoforms of *Cd44* vary due to differential incorporation of exons 6-15 (region highlighted by red bars). The spliced regions translated amino acid sequence does not contain any domains annotated in the Pfam or PROSITE databases which would be affected by differential transcript usage (Ensembl transcript ENSCGRT00000004981.1).

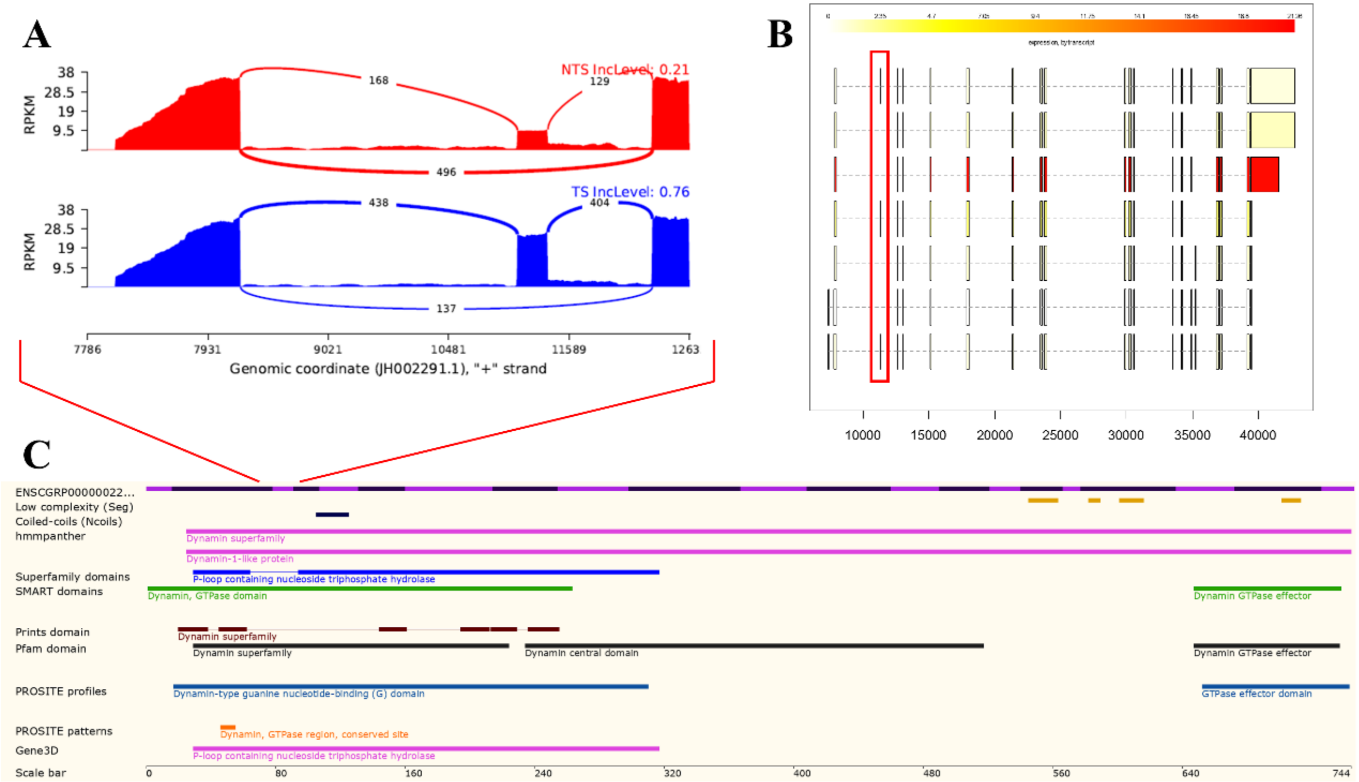

Figure 3.10 – Skipped exon in the GTPase domain encoding region of *Dnm1l.* One of the largest changes in exon inclusion levels between temperature shifted and non-temperature shifted cells occurred in the *Dnm1l* gene. (A) Inclusion of exon 3 in *Dnm1l* is significantly higher in the 31°C condition with an inclusion rate of 0.76 while an inclusion rate of 0.21 was observed in non-temperature shifted cells. (B) The location of this skipped exon event in the assembled transcripts of *Dnm1l*. (C)The translated protein sequence of *Dnm1l* with the region of the sequence affected by this alternative splicing event shown by the red bars (Ensembl transcript ENSCGRT00000022493.1).

3.4.5 Alternative splicing in genes with uncorrelated gene and protein expression

The 153 genes identified in Chapter 2 which had uncorrelated differential gene and protein expression were overlapped with genes identified with differential transcript expression or alternative splicing using *ballgown* or *Rmats*. 25 genes were identified as differentially regulated at the transcript isoform level (Appendix 3M). 2 of these genes (*Eps8* and *Ogdh*) had both up and downregulated transcripts. The *Cd44* gene is among these genes – which has the most splicing events identified as described in Section 3.4.4. A selection of these genes with roles in the cell cycle, apoptosis, splicing and translation are presented in Table 3.5.

Table 3.5 – Differential transcript usage and alternative splicing in genes which did not have differential gene and protein expression correlated. In Chapter 2 153 genes were identified which were considered differentially expressed at the gene level with a ≥1.5 fold change but did not have a ≥1.2 fold change in protein expression or vice versa. Differential expression of transcript isoforms of these genes is presented in the DE transcript column with the fold change in parentheses. Individual differential splicing events identified using *rMATs* are in the "DE Splicing Events" column where numbers in parentheses indicate the amount of times a splice event occurred. Full table in Appendix 3M.

Gene ID	**Gene Symbol**	**Gene Name**	**Gene FC**	**Protei n FC**	**DE Transcripts**	**DE Splicing Events**
ENSCGRG00000000521	*Eps8*	Epidermal Growth Factor Receptor Pathway Substrate 8	1.34	1.32	MSTRG.7794.1 (2.1) MSTRG.7794.2 (4.78) MSTRG.7794.5 (-2.11)	SE, MXE
ENSCGRG00000001864	*Eno1*	Enolase 1	1.14	1.2		SE

ENSCGRG00000003592	*Cd44*	Cell Surface Glycoprotein CD44	2.07	1.19	ENSCGRT00000004981 (4.64)	SE(8)
						MXE(11)

ENSCGRG00000009833	*Srsf5*	Serine And Arginine Rich Splicing Factor 5	-1.09	1.34	MSTRG.4290.3 (1.59)	RI
ENSCGRG00000014628	*Eef1g*	Eukaryotic Translation Elongation Factor 1 Gamma	1.11	-1.47		SE
ENSCGRG00000016658	*Eif5*	Eukaryotic Translation Initiation Factor 5	1.48	1.33	ENSCGRT00000023320 (1.52)	
ENSCGRG00000016747	*Cdk1*	Cyclin Dependent Kinase 1	-1.9	1.21	ENSCGRT00000023447 (-1.93)	
ENSCGRG00000017777	*Ogdh*	Oxoglutarate Dehydrogenase	-1.66	-1.19	ENSCGRT00000024856 (61.52) ENSCGRT00000024854 (14.58) MSTRG.12544.2 (-3.7)	SE(2) MXE

3.4.5 qPCR Validation of temperature shift induced alternative splicing

The *Acp1*, *Sc5d*, *Slirp* and *Srfbp1* genes were selected for validation using qPCR. Alternative splicing detected in the RNA-Seq data was validated in at least one transcript isoform from each of these genes. The *Acp1* transcript isoform downregulated in temperature shift was not found to be differentially expressed when cycle threshold (CT) values were normalised using either *Actb* or *Gusb*. The *Sc5d*, *Slirp* and *Srfbp1* transcripts downregulated in response to reduced culture temperatures were only identified as significantly differentially expressed when either *Actb* or *Gusb* was used to normalise CT values. All transcript isoforms of these genes predicted as downregulated in mild hypothermia were validated using qPCR with statistical significance and a ≥1.5 fold change when normalised with either *Actb* or *Gusb* (Figure 3.11, ΔCT value measurements for each biological replicate in Appendix 4O). In order to identify if these splicing events occur in CHO cell lines besides CHO-16F, qPCR was also carried out on CHO-K1 and CHO DP12 cell lines confirming alternative splicing in these genes is conserved across different CHO cell lineages (Figures 3.12 & 3.13).

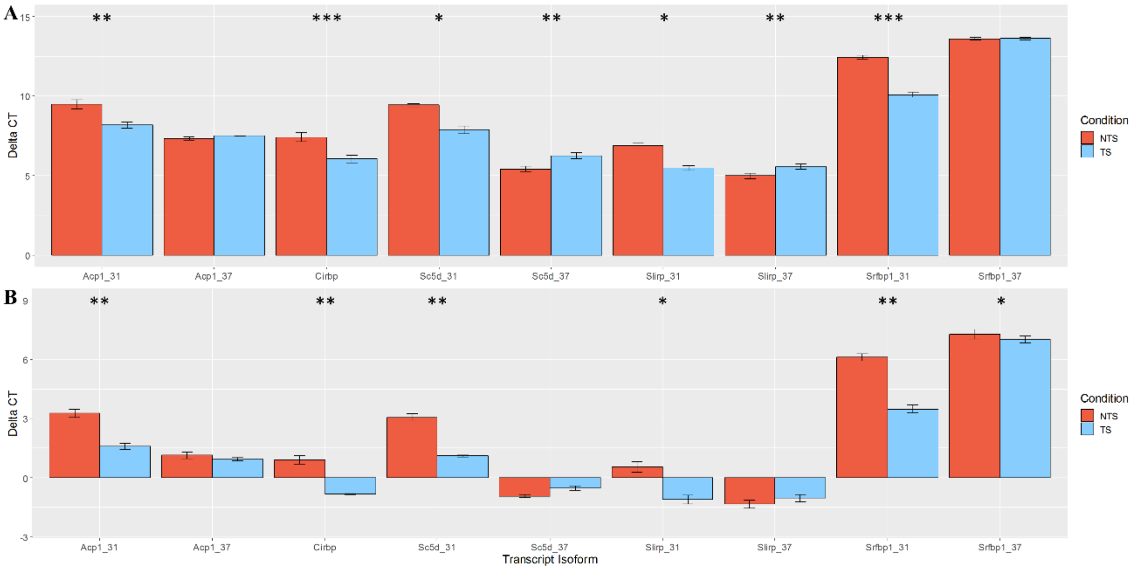

Figure 3.11 – qPCR validation of temperature shift induced alternative splicing. qPCR was carried out on genes identified as alternatively spliced with primers designed to cross exon junctions. *ActB* (A) and *GusB* (B) were used as housekeeping genes to normalise CT values. P values calculated using a two tailed T-test are illustrated with asterisks (*=≤0.05, **=≤0.01, ***≤0.001). All transcript isoforms predicted as upregulated in temperature shift (isoforms given _31 identifiers) were validated as such with statistical significance when either normalisation method was used.

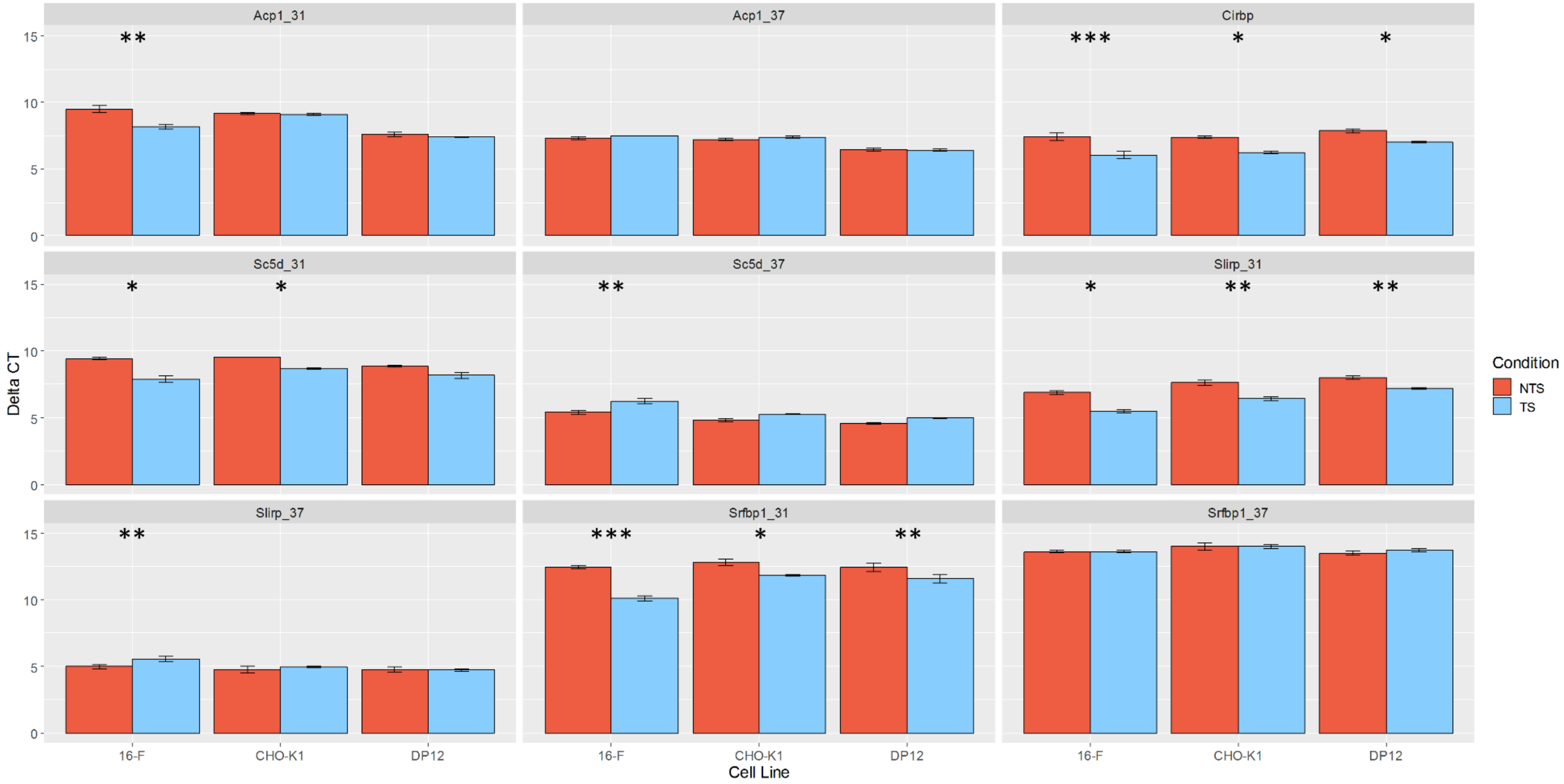

Figure 3.12 – qPCR normalised using *Actb* shows temperature shift induced alternative splicing is conserved across cell line lineages. qPCR was performed on the 16-F, K1 and DP12 CHO cell lines. The Srbp1_31 and Slirp_31 isoforms were validated to be upregulated across the 3 cell lines with statistical significance.

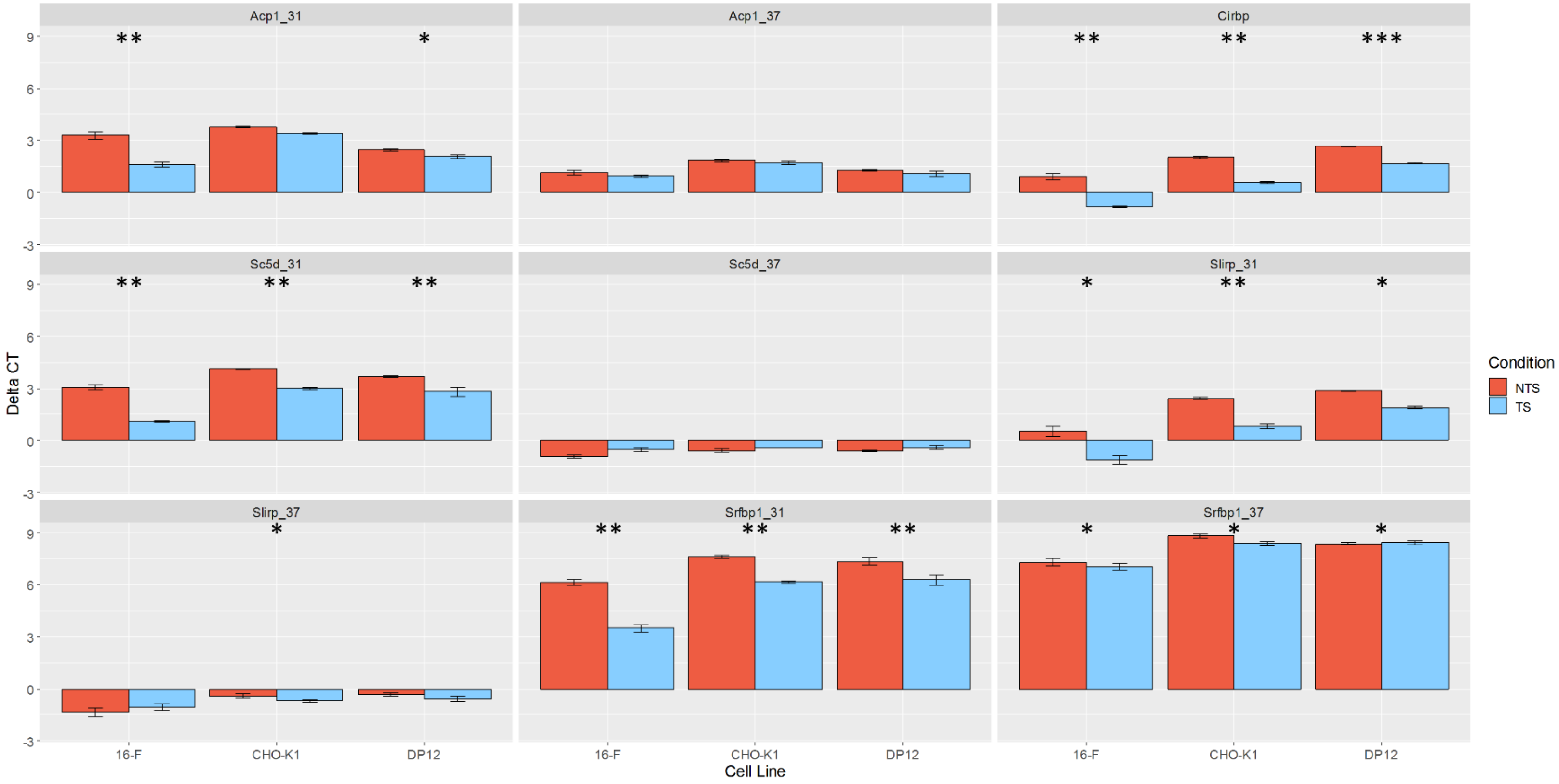

Figure 3.13 - qPCR normalised using *Gusb* shows temperature shift induced alternative splicing is conserved across cell line lineages. qPCR was performed on the 16-F, K1 and DP12 CHO cell lines. The Sc5d_31, Srbp1_31 and Slirp_31 isoforms were validated to be upregulated across the 3 cell lines with statistical significance.

3.5 Discussion

One of the most powerful applications of RNA-Seq data is the ability to discover novel transcripts and genes with transcriptome assembly algorithms. The utilisation of *Stringtie* in this chapter yielded an expanded transcriptome assembly with 32.66% more annotated gene loci and an 80.32% increase in the number of transcripts identified in comparison to the Ensembl CHO K1 cell genome annotation. The minimum isoform fraction of 0.1 and coverage thresholds used during assembly reduce the likelihood of false positives which are often present in assembled RNA-Seq transcripts - which may be the result of partially processed pre-mRNAs, partially degraded mRNA or transcripts with too low expression level to be biologically relevant (Soneson et al., 2016). In the Ensembl reference annotation there are 5,330 genes with more than one transcript isoform reported while 10,238 multi-transcript encoding genes were identified post-assembly indicating and increase in the annotation of alternative splice variants in CHO cells. Using this improved transcriptome annotation alterations in transcript expression and alternative splicing were analysed. Enrichment analysis of differentially expressed transcripts revealed results broadly in agreement with differential gene expression results from Chapter 2 with regulation of cell cycle, cell division and the G1/S transition of the mitotic cell cycle enriched.

The *Cd44* gene was unexpectedly identified in Chapter 2 as differentially expressed at the gene level with a 2.07 fold change. Upregulation of *Cd44* was unexpected as expression of *Cd44* drives phenotypic responses opposite to those observed such as cellular proliferation (Chen et al., 2018) and increased glucose uptake (Pastò et al., 2014). Cd44 was however not identified as significantly differentially expressed at the protein level – suggesting the presence of post transcriptional control of *Cd44*. Alternative splicing was found to be highly prevalent in *Cd44* and was identified as the gene with the most differential splicing events (9 skipped exon and 11 mutually exclusive exon events identified). 8 transcript isoforms of *Cd44* are present in the Ensembl genome annotation for CHO cells with 19 annotated exons. The first 5 and last 4 exons of *Cd44* are present in all transcripts however exons 6-15 are differentially incorporated into *Cd44* splice variants. Alternative splicing in *Cd44* has been reported previously and exons 6-15 also represent a varied splice region in the human, mouse and rat versions of this gene (Chen et al., 2018). The "standard" *Cd44* isoform (*Cd44s)* does not incorporate any of the variant exons (Chen et al., 2018) and is the highest expressed isoform but its expression does not significantly change in response to reduced culture temperatures (-1.005 fold change). Full length *Cd44* (containing all variant exons) is however upregulated 4.64 fold in response to temperature shift. The splicing of *Cd44* may be too complex to capture on the level of the whole transcript using isoform deconvolution approaches with a 0.1 minimum isoform fraction and reads from low expressed variants may have been incorrectly assigned to this longer transcript variant. Utilisation of *Rmats* to identify specific differential splicing events in *Cd44* reveals 6 exon skipping events were upregulated in temperature shift with isoforms containing splices between exon junctions 5:9 or 6:9 particularly prevalent. Exon skipping events downregulated in response to temperature shift were also identified - with higher incorporation of splices between exon:exon junctions 10:12 and 11:14. Proteomics results may also have been impacted by the complex splicing of *Cd44* as only 3 peptides were identified which covered 8% of the Cd44 amino acid sequence. If these peptides aligned to the first 5 or last 4 exons of *Cd44* rather than the variable region then differential expression may have been missed. Regulation of *Cd44* variant expression has been shown to be regulated by the epithelial splicing regulatory proteins Esrp1 and Esrp2 (Warzecha et al., 2009). *Esrp2* was identified in Chapter 2 as upregulated in mild hypothermia with a 2.25 fold change at the gene level. The cold shock protein *Rbm3* (identified as highly expressed in response to reduced culture temperature in Chapter 2) has also been shown to regulate *Cd44* splicing and inhibit growth by repressing *Cd44* variants containing exons 13:15 and upregulation of standard *Cd44* (Zeng et al., 2013) but these changes to *Cd44* expression were not observed in this data. The complex splicing of *Cd44* and the potential regulation of *Cd44* splicing by *Rbm3* in the response to mild hypothermia therefore requires further investigation. Widespread alterations to *Cd44* splicing may however contribute to the temperature shift response by impacting growth or energy metabolism (Chen et al., 2018; Zeng et al., 2013).

Significant differential transcript usage was observed in 597 genes which were not identified as differentially expressed at the gene level including genes with roles in cold shock, mitochondrial function and energy metabolism. The Y box protein 3 gene (*Ybx3*) is a cold shock domain containing transcription factor which has been shown to both positively (Zhang et al., 2017a) and negatively (Lima et al., 2010; Lindquist et al., 2014) regulate transcription of its targets. Ybx3 has been found to both positively regulate the cell cycle through targeting of genes such as *Cdk1*, *Ccnd1* and *Ccne2* (Zhang et al., 2017a), and negatively by binding *Cdkn1a* mRNA and enhancing its mRNA stability and translation (Nie et al., 2012). Alternative splice variants of *Ybx3* have been reported previously in humans which skip exon 6 (ENST00000279550.11), a 207bp exon downstream of the cold shock domain, which results in deletion of 69 amino acids from the protein sequence (Lindquist and Mertens, 2018). A transcript containing the same exon skipping event was assembled and only expressed (on/off) in CHO cells cultured at a reduced temperature (annotated as exon 7 in Figure 3.4 due to usage of an alternative transcription start site adding a previously unannotated exon in the 5' end). The upregulated transcript isoform also contained an exon skipping event in exon 3 – a region of the transcript which partially encodes the cold-shock nucleic-acid binding site. Upregulation of this shorter transcript isoform may therefore have decreased efficiency for binding the promoters of its target genes and impact the expression of cell cycle regulators. It is also possible that one of these transcript isoforms encodes a protein with DNA binding transcription factor functionality while the other encodes a protein which carries out Ybx's mRNA binding (translation regulating) functionality.

Transcript isoform specific differential expression was also identified in genes with roles in mitochondrial biogenesis including dynamin-1 like (*Dnm1l*) (Vanstone et al., 2016), mitochondrial fission factor (*Mff*) (Otera et al., 2010), inner mitochondrial membrane peptidase subunit 1 (*Immpl1*) (Burri et al., 2005). *Dnm1l* is among the genes with the most significant differential splicing events – with a skipped exon event occurring in exon 3 of the longest transcript isoform. In non-temperature shifted cells exon 3 is included with an inclusion rate of 0.21 however rates of exon 3 inclusion significantly increase to 0.75 in temperature shifted cells. *Dnm1l* regulates mitochondrial fission by recruiting to the mitochondrial membrane and oligomerising to form a ring around the mitochondria (Kraus and Ryan, 2017). Upon activation by hydrolysis of GTP Dnm1l changes shape, shrinking the ring two-fold and constricting the mitochondria to result in fission (Kraus and Ryan, 2017). Alternative splicing of *Dnm1l* has been previously studied in mouse embryonic fibroblasts where expression of an isoform including exon 3 (which we identified as upregulated) was found to result in decreased fission activity (Strack et al., 2013). When the nucleotide sequence of *Dnm1l* is translated, exon 3 encodes a section of the Dnm1l GTPase binding domain. The assembly of Dnm1l around the mitochondrial membrane and GTPase activity is regulated by the fission factor Mff (Kraus and Ryan, 2017). *Mff* was also observed to undergo extensive splicing in response to temperature shift, with 5 differentially expressed skipped exons and 2 mutually exclusive exon events. Splice variants of *Mff* have been reported previously (Ducommun et al., 2015) but none containing the most significant alternative splicing event identified in *Mff* in this study. Incorporation of the assembled exon 9 of *Mff* was identified as upregulated in mild hypothermia with an increase in inclusion level of 47%. The amino acid sequence of no domains annotated in the Mff protein are interrupted by inclusion of this exon and therefore the effect of upregulating an isoform of *Mff* containing exon 9 on mitochondrial fission activity is unclear. The differential splicing of *Dnm1l* and *Mff*, which were both unidentified as differentially expressed at the gene level, may alter the efficiency of mitochondrial fission in temperature shift which may have implications in apoptosis (Kraus and Ryan, 2017) and cell cycle progression (Salazar-Roa and Malumbres, 2017; Strack et al., 2013). Mitochondrial fusion and fission is coupled with the cell cycle and during the G1 phase Dnm1l (Drpl) is degraded to prevent fission and induce fusion until the G1/S phase transition of the cell cycle where Dnm1l is upregulated and activated to induce fission (Salazar-Roa and Malumbres, 2017).

In addition to regulation of mitochondrial fission the activity of mitochondria may also be impacted by differential splicing occurring in genes which were not identified as differentially expressed at the gene level. Genes with a role in mitochondrial activity or energy metabolism include SRA stem-loop interacting RNA binding protein (*Slirp*) (Lagouge et al., 2015), Pyruvate dehydrogenase lipoamide kinase isozyme 3 (*Pdk3*) and peroxisome proliferator activated receptor gamma (*Pparg*) (Wenz et al., 2011). Slirp impacts mitochondrial protein synthesis by binding mitochondrial RNAs and chaperoning them to the mitochondrial ribosome (Baughman et al., 2009; Lagouge et al., 2015). Slirp is an important regulator of OXPHOS as it maintains mitochondrial localisation of transcripts encoding OXPHOS subunits (Baughman et al., 2009). Prevalence of an exon skipping event in exon 2 was upregulated in response to temperature shift, resulting in a higher proportion of smaller transcripts which have an alternate open reading frame. An RNA recognition motif classified by Interpro is present in the 5' region of the Slirp transcript and the sequence of this motif is altered due to the isoform switch. The truncated protein may harbour altered binding affinity to its target RNAs or may target different RNAs which could affect the rates of translation of OXPHOS components. Confirmation of the existence and upregulation of transcripts skipping exon 3 was achieved using qPCR. *Pdk3* is a kinase which regulates the activity of the pyruvate dehydrogenase complex which carries out the conversion of pyruvate to acetyl-Coa – linking glycolysis to the TCA cycle and OXPHOS (Lu et al., 2008). Overexpression of *Pdk3* in cancer cells inhibits mitochondrial respiration and lactate production while suppression of *Pdk3* results in a switch from a Warburg type metabolism to primary use of OXPHOS (Kluza et al., 2012; Lu et al., 2008). A transcript isoform of *Pdk3* was identified as downregulated with a -1.55 fold change which skips exons 8 and 9 – sections of the transcript annotated with the histidine kinase-like ATPase domain using Pfam (Ensembl gene ID ENSCGRG00000003763). Transcripts variants which skip exons 9 and 10 or 10 and 11 have been previously identified but the effects of splicing in this region of *Pdk3* has not yet been studied. The *Pparg* gene was also identified as undergoing splicing but was not identified as differentially expressed at the gene level. Pparg is a regulator of pyruvate kinases and decreased Pparg is associated with increased pyruvate kinase activity and therefore decreased pyruvate dehydrogenase activity and mitochondrial respiration (Lecarpentier et al., 2017). Previously unannotated transcript isoforms were assembled for the *Pparg* gene in which an exon skipping event and 2 mutually exclusive exon events were identified - all upregulated in temperature shifted samples. Altered activity of Pparg in temperature shifted cells resulting as an occurrence of alternative splicing may contribute to metabolic switching if Pparg activity is negatively affected by expression of this alternative isoform. Analysis of *ballgown* differential transcript expression results suggest this is likely the case as transcript isoforms with the full intact protein sequence are downregulated and upregulated transcripts encode a truncated protein missing the peroxisome ligand binding domain. Decreased Pparg activity in temperature shifted cells may therefore contribute to the temperature shift response by increasing Pdk3 activity to reverse the Warburg metabolism – converting more pyruvate into acetyl-CoA for the citric acid cycle instead of converting it to lactate (Lecarpentier et al., 2017).

The results presented in this chapter are the first reported observation of alternative splicing on a global scale in CHO cells. In order to validate the occurrence of temperature shift induced splicing qPCR was applied to transcript isoforms of genes which had splicing events predicted as both up and downregulated. The *Slirp*, *Acp1*, *Sc5d* and *Srfbp1* genes were selected for validation as it was possible to design primers uniquely targeting specific transcript isoforms. At least one of the transcript isoforms for each of these genes were validated to be differentially expressed by qPCR. qPCR on the CHO-K1 and CHO-DP12 cell lines was also applied and found temperature shift induced splicing to be consistent across CHO lineages – with upregulated transcript isoforms of the *Slirp* and *Srfbp1* genes validated across all three lineages.

Utilisation of high depth RNA-Seq enabled the first global analysis of splicing in CHO cells with alternative splicing occurring at an astounding degree in response to reduced culture temperature. Complex alternative splice variants were identified in genes such as *Cd44* which have not previously been characterised in CHO cells but are conserved among other mammalian species (Chen et al., 2018). Splicing events were observed in genes regulating mitochondrial biogenesis and a switch from a Warburg phenotype of energy metabolism to use of OXPHOS including *Dnm1l*, *Mff*, *Pdk3* and *Prkag1*. Changes to the expression of these genes would not have been identified if only a differential gene expression approach was utilised as was performed in Chapter 2 (which is the most widely utilised method of analysing RNA-Seq data). The presence of post-transcriptional regulation of gene expression was suspected due to utilisation of parallel RNA-Seq and proteomics profiling in Chapter 2 which revealed genes with differential expression at either the gene or protein level which was not identified in the other experiment.

Analysis of differential transcript expression and alternative splicing has revealed targets for engineering improved CHO cells for the production of biopharmaceuticals. The *Cd44* gene was identified as undergoing complex splicing and regulates proliferation and glucose uptake (Chen et al., 2018; Pastò et al., 2014). *Cd44* could be engineered to only express the full length isoform and supress expression of alternative splice variants (Zeng et al., 2013), potentially through knocking out the endogenous gene and inserting the cDNA of spliced, full length *Cd44* into the genome. Alternatively splicing could be regulated through further increasing expression of the splice regulators *Esrp2* or *Rbm3* which were observed as upregulated in Chapter 2 and have been shown to regulate *Cd44* splicing (Warzecha et al., 2009; Zeng et al., 2013) which may impact growth and energy metabolism. The *Ybx3* gene was identified as alternatively spliced but not as differentially expressed at the whole gene level, illustrating the importance of utilising a transcript isoform approach when analysing the transcriptome. The findings of this chapter suggest that genetic engineering approaches should also take a transcript level approach and aim to engineer expression changes to individual isoforms which contain domains required for protein activity. Through upregulating the shorter isoform of *Ybx3* characterised in the temperature shift condition, the Ybx3 proteins transcription factor activity may be comprised – resulting in decreased expression of cell cycle promoting targets (Zhang et al., 2017a) and maintaining Cdkn1a binding activity (Nie et al., 2012) to halt the cell cycle. The *Dnm1l* and *Mff* genes were identified as spliced and are regulators of mitochondrial fission which impacts the cell cycle, apoptosis and energy metabolism (Kraus and Ryan, 2017; Salazar-Roa and Malumbres, 2017). Utilisation of *Dnm1l* and *Mff* as engineering targets may yield CHO cells with enhanced productivity due to altered metabolism and longevity. Finally, splicing was identified in the *Pparg* gene which likely reduces the activity of Pparg and therefore increases Pdk2 activity to result in more pyruvate being converted to fuel for the TCA cycle rather than accumulating lactate (Lecarpentier et al., 2017). Diminishing *Pparg* activity may be a useful tool for regulating the Warburg metabolism in CHO cells to increase energy availability for producing recombinant proteins and increase culture durations.

While the findings discussed in this chapter have focused on potential alterations to protein function due to alternative splicing, differential transcript isoform expression can also act as a mechanism to regulate gene expression post-transcriptionally. Modifications to the 5’ or 3’ ends of a transcript as a result of differential exon usage may alter miRNA binding sites present in transcripts (Han et al., 2018b), produce transcripts isoforms with premature stop codons which are degraded by non-sense mediated decay (NMD) before being translated (Nickless et al., 2017) or result in the inclusion of internal ribosomal entry sites (IRES) containing sequences with enhanced translation (Zhou et al., 2014a). Cold specific induction of Cirbp transcript isoforms containing IRES sites have been hypothesised as a potential mechanism through which mammalian cell lines respond to temperature (Al-Fageeh and Smales, 2009). Of the 153 genes identified in the previous chapter with differential expression identified as uncorrelated at the gene and protein levels, 25 were found to be alternatively spliced. The alternative splicing observed in these genes may contribute to the post-transcriptional regulation through altering miRNA targeting, NMD or enhanced translation. In the following chapter the role of miRNAs in post-transcriptional control of gene expression in the temperature shift response is investigated.

3.6 Conclusion

We have demonstrated the benefits of using Next Generation Sequencing to produce RNA-Seq libraries of high depth in CHO cells as well as the benefits of using a transcript or exon level analysis over the gene level approach utilised in the majority of RNA-Seq experiments. Through the assembly of RNA-Seq data the genome annotation of CHO cells was significantly improved with 27,687 transcripts and 10,019 gene loci identified which were previously unannotated. Improved annotation of the genome will serve as a valuable resource for the wider community of CHO cell researchers by enabling the study of genes at a transcript isoform level. Reduced culture temperature resulted in widespread differential transcript usage in 2,288 genes including 543 genes which were not identified as differentially expressed at the gene level. Characterisation of individual splicing events revealed alternative splicing in regulators of the cell cycle and energy metabolism including *Ybx3, Dnm1l, Pparg* and *Pdk3*. Splicing was also identified in genes which did not have correlated differential expression between the gene and protein level including *Cd44*.

Chapter 4

Elucidating the role of miRNA mediated post transcriptional regulation in the CHO cells following temperature shift

4.1 Summary

Comparison of RNA-Seq data and mass spectrometry based proteomics conducted in Chapter 2 indicated the presence of post-transcriptional control of translation plays a role in the CHO cell response to temperature reduction. In Chapter 3 post-transcriptional gene regulation was identified in the form of alternative splicing in 16.3% of genes which did not have correlated differential gene and protein expression - suggesting splicing is not the predominant mechanism responsible for post-transcriptional regulation of gene expression in temperature shifted CHO cells. In this chapter the expression of miRNAs in temperature shifted CHO cells are profiled using small RNA sequencing. From these data we identify targets of miRNAs which contribute to explanation of the lack of correlation between transcriptomic and proteomic measurements of gene expression.

4.2 Introduction

4.2.1 Small RNAs

Small RNAs are a class of non-protein coding RNAs between 18 and 200 nucleotides in length which typically play a role in regulating gene expression through RNA:RNA interactions (Kaikkonen et al., 2011; Nguyen et al., 2016). The major classes of small RNAs include microRNAs (miRNA), Piwi-interacting RNA (piRNA), small interfering RNA (siRNA) (Axtell, 2013a), small nucleolar RNAs (snoRNA) (Matera et al., 2007), tRNA-derived small RNAs (tsRNA) (Lee et al., 2009), and small nuclear RNAs (snRNA) (Valadkhan and Gunawardane, 2013). miRNAs, piRNAs, tsRNAs and siRNAs are capable of guiding nucleases to mRNAs to control gene expression post-transcriptionally (Felekkis et al., 2010; Martinez et al., 2017; Shen et al., 2018). snRNAs drive mRNA splicing during transcription by associating with donor and acceptor splice sites and forming the spliceosome complex to bring splice sites into close proximity for intron removal to occur (Shi, 2017). snoRNAs also act as guide RNAs by recruiting proteins to form snoRNP complexes to direct methylation and pseudouridine modifications to RNAs such as ribosomal RNA (Bachellerie et al., 2002).

miRNAs and siRNAs are of particular interest for use as tools for bioengineering as they can be used to manipulate gene expression (Felekkis et al., 2010; Valdés-Bango Curell and Barron, 2018). Both miRNAs and siRNAs utilise the RNAi pathway to carry out their silencing activity and are of similar length however these molecules differ in their function and biogenesis (Lam et al., 2015). siRNAs act as a defence mechanism to protect cells from infection of viruses with double stranded RNA (dsRNA) genomes. In the antiviral siRNA pathway viral dsRNA is recognised and processed by the Dicer protein into a 21-24nt RNA duplex which is loaded into the RNA-induced silencing complex (RISC). One of the strands is degraded leaving a single strand of siRNA which is used to target the RISC to RNA with perfect sequence complementarity to the siRNA for degradation (Felekkis et al., 2010). miRNAs are conversely generated from endogenously expressed RNAs which form a hairpin structure containing a double stranded RNA region due to the self-complimentary of the RNA sequence in the 5’ and 3’ ends of the pri-miRNA (Ding and Voinnet, 2007). Pri-miRNAs are expressed from the genome and processed in the nucleus by the Drosha protein to form a 70nt hairpin before being exported into the cytoplasm by Exportin5 where they are then bound by Dicer. Dicer cleaves the hairpin loop region leaving 18-25 nucleotides of dsRNA, a single strand of which (the mature miRNA) is loaded into the RISC complex and directs suppression of gene expression (Felekkis et al., 2010). miRNA mediated regulation of gene expression can be a result of either degradation of target mRNA transcripts through endonucleolytic cleavage or translation suppression (Lam et al., 2015). miRNA mediated degradation however rarely occurs – only when near perfect Watson-Crick base pairing occurs between a miRNA and a target mRNA. miRNAs typically bind the 3' untranslated region (3'UTR) of a target transcript with only partial complementarity required – a property which enables a single miRNA to regulate the expression of hundreds of transcripts (Lam et al., 2015). The region of a mature miRNA transcript which regulates target recognition is known as the seed region and ranges from 2-7nt of the mature miRNA transcript. Base pairing between the seed region and a transcript alone is however not sufficient to determine miRNA activity as other characteristics contribute including thermodynamics, the secondary structure of the miRNA:mRNA duplex and the location in the mRNA at which the target site occurs (Martinez-Sanchez and Murphy, 2013). Given the complexity of miRNA:mRNA interactions, as well at the laborious nature of experientially validating direct targets, a range of bioinformatics approaches have been developed to identify and predict the roles of miRNAs in silico.

4.2.2 Bioinformatics software for predicting small RNAs and targets using NGS data

4.2.2.1 Identifying miRNAs from next generation sequencing

miRDeep was the first algorithm released for the prediction of miRNAs from NGS data sets (Friedländer et al., 2008). *miRDeep* utilises aligned reads from small RNA-Seq experiments to identify potential miRNA precursor locations in the genome. The first stage of this analysis results in thousands of potential precursors by identifying regions of the genome with aligned read coverage from small RNA-Seq. If the read coverage region is longer than 30 nucleotides the region is extended by 22 nucleotides both upstream and downstream to generate a potential precursor miRNA. If the coverage is shorter than 30bp the precursor is generated by extending the region 22bp in one direction and then extension is continued in the opposite direction until a total length of 110bp is achieved. Potential precursors are then filtered on their likelihood to be false positives based on their predicted secondary RNA structure. The properties of miRNA biogenesis are then exploited to find likely miRNAs utilising the alignment positions of small RNA-Seq reads to genomic miRNA precursor loci. When pre-miRs are processed by dicer the nucleotides are cleaved into the mature miRNA, the star sequence and the loop region. The mature miRNA is predominantly incorporated into RISC and carries out biological activity as a guide miRNA, the star miRNA is typically degraded but can also be a functional guide in some miRNAs and the loop region is degraded (Guo and Lu, 2010). Therefore it is expected for the majority of miRNA reads in a small RNA-Seq data set to map wholly within one of these 3 regions while reads derived from non-miRNAs would map across the boundaries. In addition, the mature miRNA sequence is typically more abundant than the star sequence. Potential miRNA precursors are scored based on how well read alignments to the precursor fit this model of miRNA biogenesis to identify both known and novel miRNAs (Friedländer et al., 2008). The *miRDeep* algorithm was improved and rereleased as *miRDeep2* which has updated methods for identifying and scoring potential precursors (Friedländer et al., 2012). Precursors are only selected if they are the highest alignment stack (the location with the highest coverage) within 70 nucleotides upstream and downstream. Precursors are identified for each strand of the genome and therefore each potential pre-miR is extracted from the genome twice. The excised pre-miR is extracted with 70 nucleotides upstream and downstream of the alignment stack. If too many candidates are predicted in this stage (over 50,000) the stack height stringency is increased by 1 and subsequent rounds of repetition of this process are applied until less than 50,000 candidates are predicted. *Bowtie* is then used to align reads to the predicted precursor, only accepting perfect alignments, and *RNAFold* is used to predict the secondary structure of the candidate pre-miRs. Candidate miRs are scored using the same method as *miRDeep* with the exception of only using sense alignments and precursors with >60% of the mature miRNAs nucleotides base-paired in the predicted pre-miR secondary structure (Friedländer et al., 2012). The mapper and quantifier modules of *miRDeep* can be used for aligning reads to a genome and quantifying the expression of detected miRNAs respectively. The *miRDeep* tools have also been adapted into the *miRDeep** package which has a graphical user interface and alters the *miRDeep* algorithm during the stages for precursor generation by extending predicted mature miRNAs by 15nt upstream, 15nt downstream for the loop region, an equivalent number of nucleotides to the predicted mature miRNA length downstream for the star miRNA and a further 22nt for the miRNA offset (An et al., 2013). *miRDeep*2 is available on the *mirTools2.0* web server (Wu et al., 2013).

miRanalyzer (Hackenberg et al., 2009, 2011) is a pipeline for the detection and analysis of miRNAs from sequencing data available both as an online web-server (now integrated into the sRNA workbench (Mohorianu et al., 2017)) and a downloadable standalone package which can be configured to be applied to any organism. The *miRanalyzer* pipeline first aligns sequence data to known miRNAs in *miRBase*, and then to Refseq and RFam to remove known miRNAs, mRNA transcripts and non-coding RNAs. Unaligned reads are mapped to the genome for novel miRNA detection. First, overlapping aligned reads are clustered and the positions of the most upstream and downstream nucleotide with coverage in the cluster are noted alongside the position of the most expressed read (which likely represents the mature miRNA). Candidate precursors are generated for each cluster with a range of lengths between 65 and 135 nucleotides for mammalian miRNAs. Candidate precursors are filtered based on criteria similar to miRdeep, with potential precursors removed if they do not form a hairpin, have 19 base pair bindings in the stem of the pre-miRNA (of which at least 11 must be in the mature miRNA) and to ensure read clusters do not overlap the boundaries between the 5' arm and the loop. Likely pre-miRNA candidates are selected based on the number of self-complementary binding sites, the free energy of the predicted secondary structure, the number of reads in alignment clusters, and the asymmetry of bulbs in the secondary structure (Hackenberg et al., 2011). Other algorithms are available for the prediction of miRNAs from RNA-Seq datasets however their use has become obsolete due to becoming unsupported or their accuracy superseded by more recent software including *MIReNA* (Mathelier and Carbone, 2010). In addition some software pipelines only exist in web server form and are configured to be used with a small number of model organisms such as *DARIO* (Fasold et al., 2011).

As well as these software packages which have been developed for the comprehensive study of miRNAs, algorithms have also been developed which can simultaneously predict small RNAs of multiple classes. *Shortstack* is one such algorithm which has been shown to accurately predict miRNAs, piRNAs and siRNAs (Axtell, 2013b). *Shortstack* applies a similar approach to the miRNA prediction algorithms described above, but does so in a modified way to account for non-miRNA small RNAs. To identify potential small RNAs, reads are aligned to a genome sequence and loci with coverage above a user specified depth are identified. The identified loci are extended upstream and downstream to account for positional heterogeneities in mature small RNAs which occur during precursor loading to the Piwi or Argonaute proteins during processing. Regions which overlap after this extension are clustered together and are separated into dicer-derived and non-dicer-derived sequences based on their length. Secondary structural analysis is performed on unique dicer-derived sequences and folded short RNA predictions are compared to those in miRBase by structural criteria including the number and length of loops and hairpins, as well as base-pair bindings between potential mature and star miRNAs. Precursors matching the criteria for miRNAs are annotated and quantified (Axtell, 2013b). Small RNAs with no hairpin are considered candidate siRNAs.

4.2.2.2 miRNA target prediction

Several algorithms have been developed to predict targets of miRNAs which rely on different methods and criteria to identify and score miRNA-mRNA target pairs (Riffo-Campos et al., 2016). The first algorithm created for identifying miRNA targets was *TargetScan* (Lewis et al., 2003). *TargetScan* relies both on base pairing of the miRNA to UTR regions of mRNAs in an organism as well as conservation of the target region between species. First the UTR regions of a transcript are searched for perfect base pairing to the seed region (bases 2-8) of a given mature miRNA and base pairing is extended in each direction until a mismatch occurs. The base-pairing of a section of a miRNA 3’ to any mismatches is then optimised by predicting the secondary structure of the mRNA and assessing the binding of the miRNA to the target site and regions downstream when folding is accounted for. Potential binding sites are assessed based on their free energy, number of seed matches and the likelihood of interaction of a given miRNA:mRNA pair are ranked. The process is repeated for other organisms and ranks for orthologous miRNA:mRNA pairs are compared to find conserved miRNA targets. The current version of *TargetScan* uses 14 features of miRNA:mRNA interactions to predict sites which likely have biological function (Agarwal et al., 2015). Features of mRNA targets include the proximity of the target site to the path of the ribosome (how close it is to the translation stop site), the GC content of the UTR (lower GC increases the accessibility of the UTR to the silencing complex), UTR length and stability of predicted secondary structure. Features of the miRNA are also considered including the number of predicted sites throughout the transcriptome and the strength of the seed-pairing between the miRNA and its predicted target. The *TargetScan* website also hosts a database of predicted miRNA interactions.

Alternatives to *TargetScan* include the *DIANA-microT* webserver (Paraskevopoulou et al., 2013) and *MiRanda* (Enright et al., 2004; John et al., 2004). The *MiRanda* algorithm utilises a similar method to *TargetScan* – applying alignment of miRNAs to mRNA sequences to identify targets followed by calculation of free energy and identification of evolutionary conservation. During the alignment stage *MiRanda* applies an algorithm similar to Smith-Waterman adapted to score based on complementarity rather than homology. Further rules are applied to account for what was known about the properties of miRNA targeting at the time of development - prohibiting mismatches at positions 2-4, a maximum of 5 mismatches between positions 3-12, at least 1 mismatch between position 9 and the end of the alignment and less than 2 mismatches in the last 5 nucleotides of the match. The secondary structure of the mRNA:miRNA interaction while hybridized is predicted using the *Vienna* package and the minimum free energy calculated. Evolutionary conservation is assessed by repeating the target identification and free energy calculation stages in closely related organisms and identifying targets in the same position in orthologous mRNA UTRs. Complementarity scores are multiplied in the first 11 nucleotides of the target site to account for seed region binding (John et al., 2004). *DIANA-microT* (Paraskevopoulou et al., 2013) identifies miRNA recognition sites with a match of at least 7 consecutive base pairs between a 3'UTR of an mRNA and the first 9nt of a miRNA (Maragkakis et al., 2009). If the 3' (non-seed region) of a miRNA sequence also matches a target site then a mismatch is allowed in the seed region. Evolutionary conservation of potential targets sites is assessed using the sequences of 27 species and scored by comparing against mock target sites generated using mock miRNA sequences. Mock miRNA sequences are generated for each real input miRNA and generated to have the same number of predicted target sites. Scores of mock sequences are used as a cut-off for identifying false positive interactions predicted for input miRNAs. The *DIANA-microT* algorithm is only available as a web server configured to identify miRNA targets in human, mouse, *C. elegans* and *D. melanogaster*. An updated release of *DIANA-microT* enables identification of miRNA target sites in CDS regions (Paraskevopoulou et al., 2013).

4.2.3 Small RNAs in CHO cells

4.2.3.1 Identification of CHO cell miRNAs

miRNAs were first identified in CHO cells by the cloning of miR-21 - discovered using array based methodology to profile miRNAs conserved between humans, mice and rats (Gammell et al., 2007). Since the publication of this study the potential of miRNAs for engineering desirable phenotypes in production CHO cells has begun to be explored (Valdés-Bango Curell and Barron, 2018). Several miRNAs have been investigated including mmu-miR-446h-5p which has been shown to improve apoptosis resistance and productivity when knocked down (Druz et al., 2013), miR-557, miR-1287 and miR-17 which improve productivity when upregulated (Jadhav et al., 2014; Strotbek et al., 2013), improving expression of difficult to express proteins through miR-557 expression (Fischer et al., 2017) and the role of miR-23 in metabolic switching to oxidative phosphorylation in CHO cells (Kelly et al., 2015a).

Several studies have begun to apply techniques to survey expression of miRNAs in CHO cells. Sanger based EST sequencing experiments were applied to the CHO transcriptome and used to discover the sequence of 28,000 transcripts of which 260 were unique miRNAs (Kantardjieff et al., 2009). Next generation sequencing has also been utilised to identify miRNAs in CHO cells with 350 miRNAs discovered using small RNA-Seq in one study (Johnson et al., 2011) and 387 in another (Hackl et al., 2011). A further 71 novel mature miRNAs were identified in a follow up study by the same authors (Diendorfer et al., 2015) using conservation of miRNAs in miRBase to identify candidate miRs in the CHO genomes and then providing proof of expression by their presence in NGS or array data. In the current miRBase release (version 22) there are 351 unique mature CHO cell miRNAs with sufficient evidence of expression to be included in the database (Table 4.1) (Griffiths-Jones et al., 2006).

Table 4.1 – Annotated miRNAs in miRBase for human, mouse rat and Chinese hamster. The Chinese hamster has considerably less miRNAs annotated than human, mouse and rat. Data from miRBase release 22 (March 2018)

Species	Precursor miRNAs	Mature miRNAs
Human	1,917	2,654
Mouse	1,234	1,978
Rat	496	764
Chinese hamster	245	353

4.2.3.2 Differential expression of miRNAs in CHO cells

Microarray profiling methods were originally reliant on utilising microarrays designed for profiling human miRNAs but due to the conservation of mature miRNA sequences in mammalian species these were also applicable for use on CHO cells (Barron et al., 2011). Using cross-species arrays as many as 350 miRNAs could be detected and quantified in a single experiment which enabled detection of differentially expressed genes such as miR-7 and 9 other miRNAs in response to temperature shift (Barron et al., 2011). Arrays have also been applied to identify differentially expressed genes in CHO cells with varied productivity levels (Harreither et al., 2015), culture stage (Gammell et al., 2007) and growth rates (Klanert et al., 2016) revealing 127, 26 and 12 miRNAs associated with each phenotype respectively. Next generation sequencing has also been utilised to study miRNA expression patterns - Illumina sequencing of small RNAs was used to identify the expression of 387 miRNAs in 6 biotechnologically relevant CHO cell lines (Hackl et al., 2011). In this study 18 miRNAs were found to be differentially expressed between serum-dependant and serum-free adapted cell lines.

4.2.4 Potential for miRNAs as tools for cellular engineering

A single miRNA has the potential to regulate the expression of hundreds of mRNA targets and have the capacity to regulate biological processes by targeting many genes in a given pathway simultaneously (Pritchard et al., 2012). For example, an miRNA can be capable of controlling the process of cellular differentiation during embryonic development (Alvarez-Garcia and Miska, 2005) or control cellular growth and apoptosis (Zhao et al., 2015). Gene silencing can be achieved by upregulating endogenous miRNAs which target a given gene or pathway or by constructing artificial miRNAs which can be transfected into cells (Macovei et al., 2012). Expression of miRNAs at a particular time can be achieved by utilising inducible promoters in which transcription is only initiated in response to a stimuli such as a drug treatment or heat-shock (Weber and Fussenegger, 2007; Yang and Paschen, 2008).

miRNA expression can also be downregulated by gene knockout or knockdown approaches to increase the translation of their target mRNAs. Depletion of miRNA expression can be achieved by the expression of miRNA sponges or decoys (Banks et al., 2012; Gentner and Naldini, 2012). miRNA sponges are engineered RNA sequences which contain multiple copies of a given miRNA binding site which results in its target miRNA becoming sequestered and unable to target its intended mRNA (Gentner and Naldini, 2012). A recently discovered role of lncRNAs is that they may act as endogenous miRNA decoys to sequester miRNAs and fine tune gene expression (Banks et al., 2012). For example the lncRNA linc-MD1 sequesters miR-133 and miR-135 to regulate expression of the Myocyte-specific enhancer factor *Mef2c* during muscle differentiation (Cesana et al., 2011). Endogenous miRNAs could also serve as engineering targets by using gene editing techniques such as CRISPR-Cas9 to edit their genetic sequence (Aquino-Jarquin, 2017). Such gene editing techniques have been used to disrupt the functionality of specific miRNAs and promote apoptosis in ovarian cancer cells (Li et al., 2017) and have been applied to delete miRNAs from the genome to drive tumorigenesis (Yoshino et al., 2017).

4.2.4.1 Engineering of CHO cells utilising miRNAs

miR-7 (Barron et al., 2011) has been utilised as a tool for engineering enhanced productivity in CHO cells. When transfected with exogenous copies of the pre-miR-7 precursor which are processed and result in upregulated mature miR-7, cellular growth rate was reduced and individual cell productivity increased (Barron et al., 2011). miR-7 contributes to the cell cycle arrest in the G1 phase in temperature shifted CHO cells which results in enhanced yields through increasing the length of the stationary phase of culture where cells are more productive (Sanchez et al., 2013). Further examples of CHO miRNA engineering include upregulation of miR-483 and miR-30 to enhance recombinant protein production (Emmerling et al., 2016; Fischer et al., 2015), upregulation of miR-17 to increase proliferation (Jadhav et al., 2012) and dual expression of miR-557 and miR-1287 to enhance product yield via improved cell density and specific productivity respectively (Strotbek et al., 2013). The mouse miR-466h-5p induces apoptosis by inhibiting the expression of anti-apoptotic genes and therefore miR-466h was used as a candidate for knockdown by anti-miR expression (Druz et al., 2013). Inhibition of miR-446h-5p resulted in a 53.8% increase in viable cells compared to untransfected cells in apoptotic conditions through the upregulation of anti-apoptotic genes such as *Bcl2l2* and delay of Caspase activation (Druz et al., 2013).

4.2.4.2 Improving our understanding of the role of miRNAs in CHO cells

Identification of miRNA expression patterns associated with reduced temperature has been studied previously using human, mouse and rat array probes (Gammell et al., 2007). However this study, and all previously mentioned experiments which utilised microarrays, were limited to previously characterised miRNAs which are conserved between species. The current version of miRbase (Version 22, March 2018) contains just 353 mature miR entries for the Chinese hamster but 2,654 for human and 1,978 for mouse - suggesting incomplete annotation of the miRNA landscape of CHO cells (Griffiths-Jones et al., 2006). An opportunity therefore remains to apply deep next generation sequencing methods to discover novel miRNAs which yet to be discovered in CHO cells. A recent miRNA sequencing experiment however demonstrated the capacity of next generation sequencing to discover miRNAs in CHO cells in which 71 novel miRNA sequences were identified (Diendorfer et al., 2015). No study to date has yet applied deep Next Generation Sequencing to small RNAs expressed in temperature shifted CHO cells and therefore there may be miRNAs involved in the temperature shift response which have yet to be catalogued. Performing next generation sequencing on temperature shifted CHO cells could result in the identification of miRNA targets for engineering enhanced phenotypes such as apoptosis resistance, enhanced growth rates or increased productivity (Müller et al., 2008). The experimental design used to generate sequencing libraries throughout this thesis is also beneficial for the purpose of discovering more about the function of miRNAs as total RNA-Seq, proteomics and RiboSeq experiments can be integrated to identify potential mRNA targets of miRNA regulation.

In this chapter small RNA-Seq is applied to survey the miRNA landscape of CHO cells cultured at a reduced temperature. The utilisation of deep sequencing enables the potential to identify novel miRNA transcripts and accurately quantify differential expression of miRNAs induced by mild hypothermia. By predicting miRNA:mRNA target interactions it may be possible to elucidate post-transcriptional control of gene expression contributing to the inconsistency between mRNA and protein measurements of gene expression identified in Chapter 2.

4.3 Methods

4.3.1 Cell culture and sequencing

Total RNA for small RNASeq was harvested in parallel from the CHO cell cultures descried in Section 2.2.1. Libraries were generated using the TruSeq Small RNA Library Preparation Kit (Illumina Inc.). RNA was sequenced on an Illumina NextSeq 500 instrument, multiplexed across the 4 lanes on the flow cell and configured to produce a minimum of 10 million reads per sample at a minimum of 50bp. Sequence libraries were constructed by Dr Clair Gallagher and Alan Costello in NICB, Dublin City University.

4.3.2 Pre-processing of small RNA data

Adapter sequences were trimmed using *cutadapt* (Martin, 2011) using the adapter sequence ‘TGGAATTCTCGG’ a second pass of *cutadapt* was used to trim 4 nucleotides from the starts and ends of reads to remove barcode sequences. Reads below 17 nucleotides in length were discarded at this stage to reduce contamination as reads below this length would be too small to be derived from miRNAs. *FastQC* (Andrews, 2010) was used to assess the quality of samples – ensuring a peak was present between 18-23 nucleotides in length in the read length distribution chart and to determine the overrepresentation of sequences such as ribosomal RNA or remaining adapters.

4.3.3 Identification of novel miRNAs

miRDeep2 (Friedländer et al., 2012) was utilised to identify novel miRNAs from the pre-processed data. RNA-Seq reads were first aligned to the CHO cell genome and alignments output in the ARF format for use in the main *miRDeep2* algorithm. Previously annotated Chinese hamster miRNAs in miRBase were provided as input to *miRDeep2* as well as the mature miRNA sequences for mammalian organisms with over 500 annotated miRNAs. miRNAs with a *miRDeep2* score > 10 were accepted as candidate novel miRNAs.

4.3.4 Differential expression of miRNAs

Bowtie (Langmead et al., 2009) was used to align reads to the Ensembl CHO-K1 genome (Xu et al., 2011; Zerbino et al., 2018) with parameters set to disallow mismatches and output alignments in the SAM format, which was converted to BAM and sorted using *Samtools* (Li et al., 2009). The BED file containing predicted miRNA output from *miRDeep2* was used as input to count the number of aligned reads to genomic loci with

Bedtools multicov (Quinlan and Hall, 2010). *DESeq2* was used to identify differential expression in miRNAs with a minimum of 10 aligned read counts and considered differentially expressed if the miRNA had a minimum fold change >±1.2 and Benjamini Hochberg adjusted p value of < 0.05 between conditions.

4.3.5 Prediction of miRNA targets

Ensembl GeneIDs for differentially expressed genes identified in Chapter 2 and the miRBase identifiers for human orthologues of differentially expressed miRNAs from Section 4.3.4 were imported to R. The *biomaRt* Bioconductor R package was used to convert CHO cell Ensembl GeneIDs to orthologous human Ensembl GeneIDs (Durinck et al., 2009). The *multiMiR* Bioconductor R package (Ru et al., 2014) was used to retrieve miRNA:mRNA interactions between differentially expressed genes and miRNAs. All interactions in databases with experimentally validated targets (miRecords (Xiao et al., 2009), miRTarBase (Chou et al., 2018) and TarBase (Vergoulis et al., 2012)) and evolutionarily conserved targets (miRanda (Betel et al., 2008), PITA (Kertesz et al., 2007), TargetScan (Agarwal et al., 2015)) were retrieved. In addition the top 20% scoring *in silico* predicted miRNA:mRNA interactions in databases without evolutionary conservation support were included (miRanda, PITA, TargetScan, DIANA_microT (Paraskevopoulou et al., 2013), ElMMo (Gaidatzis et al., 2007), MicroCosm (Griffiths-Jones et al., 2006), miRDB (Wang, 2008) and PicTar (Anders et al., 2012a)). Predictions were retained if an interaction was present in a database with biological validation as well as predictions identified in TargetScan and at least 1 other database. The described process was repeated on alternatively spliced genes identified in Chapter 3 as well as differentially expressed proteins identified in Chapter 2.

4.4 Results

4.4.1 Quality and pre-processing

Libraries were sequenced to an average of 55.4 million reads per sample (range 44.7-62.9 million). Adapter and quality trimming resulted in the removal of an average of 4.4 million reads per sample (3.67-5.3 million, 7.9%) (Figure 4.1). An average of 28.7 million reads (22.3-35.4 million, 53.7%) per sample were discarded during length filtering due to falling outside the 17-25 nucleotide range (Figure 4.2). After this stringent pre-processing an average of 21.3 million reads remained per sample (16.05-24.29 million, 38.4%) to be used in assembly of miRNAs and differential expression analyses (Tables of full pre-processing results in Appendix 4A).

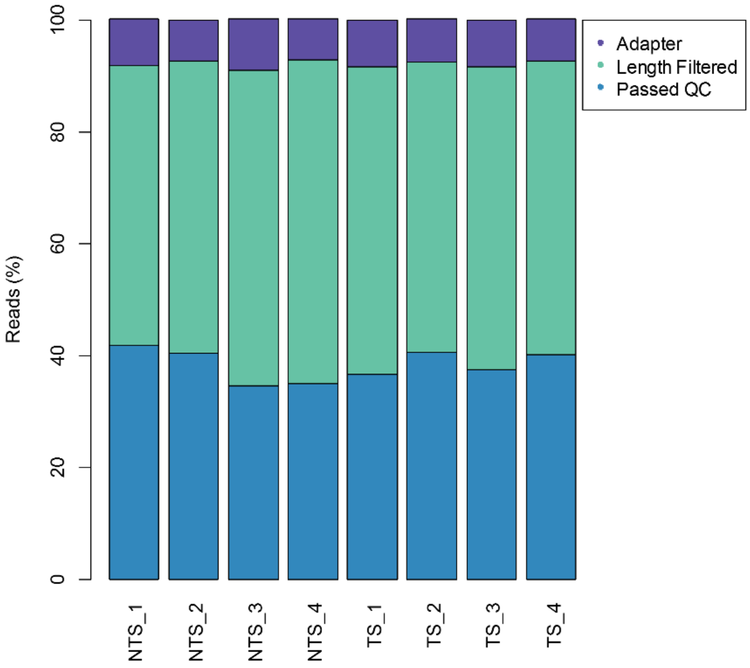

Figure 4.1 – Small RNA-Seq Read Pre-processing. Reads were trimmed to remove barcodes and adapters resulting in the removal of 7.9% of reads (purple). A further 53.7% of reads were discarded due to being outside the 17-25 nucleotide range (green). An average of 38.4% of reads passed pre-processing (blue).

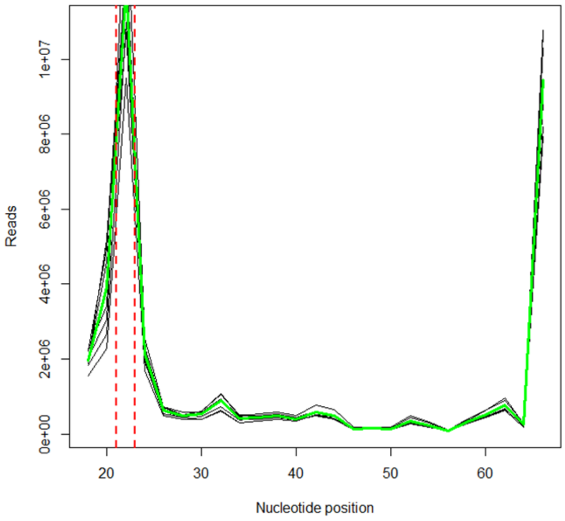

Figure 4.2 – Sequence length distribution. A large peak is present in the 21-23 nucleotide range of trimmed reads - the size of the majority of miRNAs annotated in CHO cells annotated in miRBase.

4.4.2 Detection and assembly of miRNAs with *miRDeep2*

70% of the preprocessed reads were aligned to the CHO-K1 genome and utilised for the *miRDeep2* discovery process. Known Chinese hamster miRNAs and those from other mammalian species (e.g. human, mouse, rat) in miRBase were provided as input and resulted in the detection of 1,507 candidate miRNAs of which 1,280 were novel to the Chinese hamster (Appendix 4B). 251 of these novel candidate miRNAs had a miRDeep2 score ≥10 which has a true positive rate of 85 ± 2% associated with it which is calculated by randomly permuting read alignments to potential precursors as aligned to other potential precursors in the same position to check the hypothesis that the miRNA is derived from dicer processing (Figure 4.3, Appendix 4C). 169 of the 251 candidates have a significant *randfold* p value and no hits to the rfam database, further evidence they are likely miRNAs as they form a secondary structure consistent with known miRNAs and do not contain non-coding RNA sequence elements such as tRNAs or rRNAs.

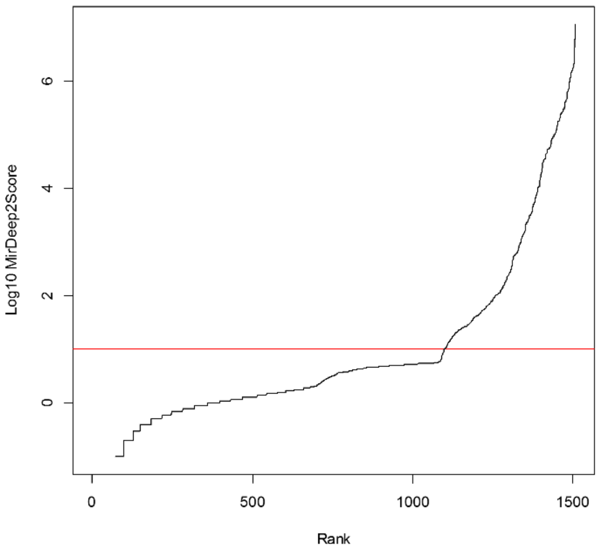

Figure 4.3 – Detection of miRNAs with *miRDeep2*. 1,507 miRNAs were assembled of which 1,280 were novel. 251 of these have a miRDeep2 score ≥10 and are likely true positive miRNA candidates (Score 10 indicated by red line).

For 119 of the miRNAs with a score ≥ 10 there are no reported miRNAs with orthologous seed sequences in miRBase for the mammalian organisms used as input to *miRDeep2*. The highest scoring of these novel miRNAs which has a significant Randfold P value, has a miRDeep2 score of 24,514 – higher than well studied miRNAs such as miR34a (21,547), miR30a-5p (10,228) and miR-9 (2,316) and therefore this may present a true positive miRNA which is uniquely expressed in CHO cells. The coverage and predicted hairpin structure of novel_miR_24,227 is depicted in Figure 4.4.

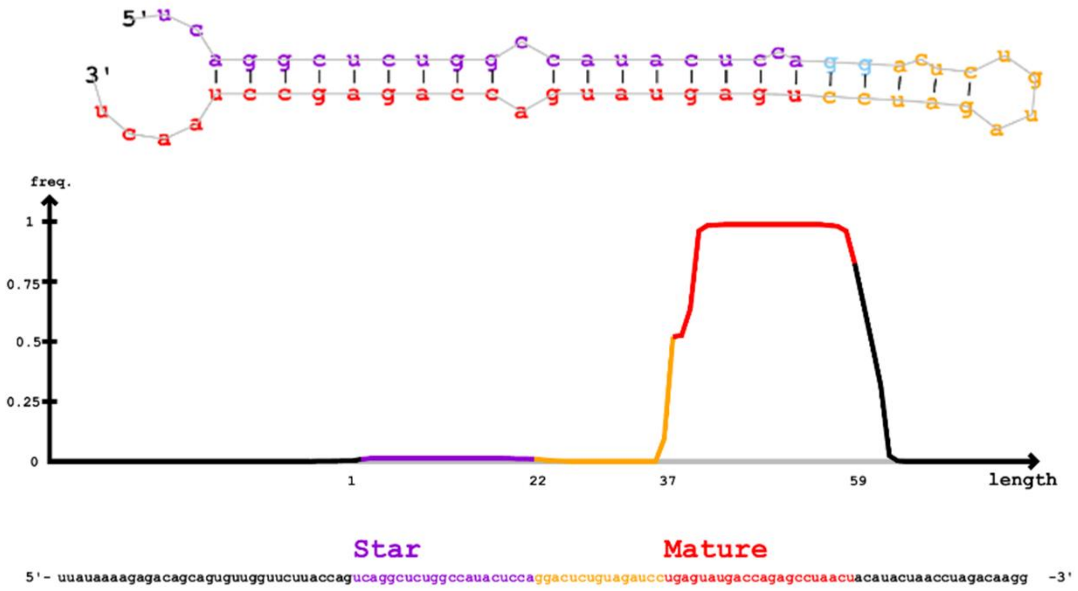

Figure 4.4 – Predicted hairpin structure and coverage of novel_miR_24227 Presented in this figure is a novel miRNA with the highest *miRDeep2* score which has no annotated orthologue in any species in miRBase. The secondary structure (top) is typical of a miRNA, with consistent base pairing in the seed region (5' of the red sequence) and forming a hairpin. The majority of reads aligning to this hairpin cover the mature sequence region (bottom, red). Figure generated using *miRDeep2*.

132 of the novel miRNAs identified have orthologues reported in other mammalian organisms. miRNAs were identified from all species used as input to *miRDeep2* with the exception of armadillo, opossum and chimpanzee. The highest scoring novel miRNA with an orthologue is novel_mir_482, orthologous to mdo-miR-7388c (Figure 4.5). Among the orthologous miRNAs assembled which have previously not been reported in CHO cells are 11 miRNAs which are highly conserved among species and prevalent in the literature (Table 4.2). The most cited of these miRNAs is miR-335, presented in Figure 4.6.

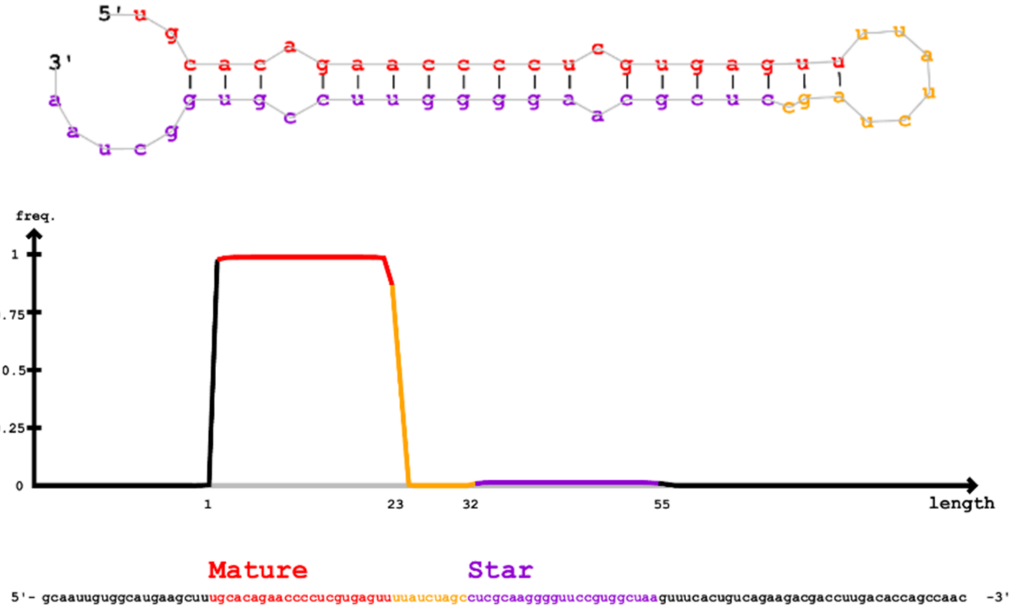

Figure 4.5- Predicted hairpin structure and coverage of novel_mir_482. Presented in this figure is the novel miRNA novel_mir_482, orthologous to mdo-miR-7388c. The mature sequence is highly expressed relative to the Star sequence as can be seen by the peak in the coverage coloured red.

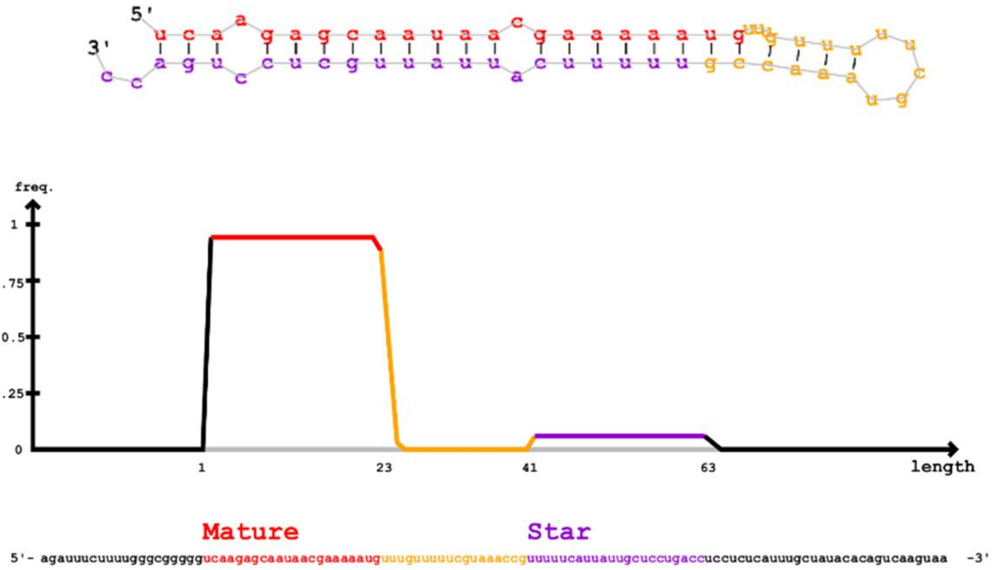

Figure 4.6 – Predicted structure of novel_miR_ 5944. Novel_miR_5944 is orthologous to miR_335, a miRNA highly cited in the literature and previously reported in 19 species.

Table 4.2 – Assembled orthologous miRNAs which have not been previously annotated in CHO cells. The 11 miRNAs presented in this table are orthologous to known miRNAs in other mammalian species and their functionality has been studied and reported in the literature.

miRNA	ID	Genomic Coordinates	Score	miRNA present in species	Number of publications in Pubmed
miR-207	JH001133.1_29824	JH001133.1:479663..479725:-	980.7	Mouse Rat Dog	7
miR-290a-5p	JH058547.1_43020	JH058547.1:108..165:+	88.8	Mouse Rat Guinea Pig	64
miR-295-5p	JH000876.1_26769	JH000876.1:427933..427996:-	26	Mouse Rat	6
miR-335	JH000073.1_5944	JH000073.1:2901468..2901530:+	21.6	Human Mouse Rat + 16	255
miR-383-3p	JH001533.1_33225	JH001533.1:148968..149031:+	20.5	Human Mouse Rat + 24	72
miR-448-3p	JH000200.1_11719	JH000620.1:1030049..1030126:+	135	Human Mouse Rat + 15	45
miR-873-3p	JH000060.1_4751	JH000060.1:723280..723352:+	180.1	Human Mouse Rat + 14	38
miR-939-5p	JH040064.1_42766	JH040064.1:201..262:+	47.7	Human Macaque Chimp Orangutan	32
miR-1231	JH000489.1_20324	JH000489.1:1218229..1218294:-	312.5	Human Mouse	8
miR-1294	JH000196.1_11622	JH000196.1:1407600..1407660:-	1578.9	Human Chimp Marmoset	6
miR-2909	JH000836.1_26319	JH000836.1:312746..312805:+	126.4	Human	16

When the novel miRNAs identified in this experiment are combined with those already present in miRBase, the number of miRNAs reported in the Chinese hamster increases to 583, closer to parity with other closely related mammalian species such as the rabbit and guinea pig (Figure 4.7).

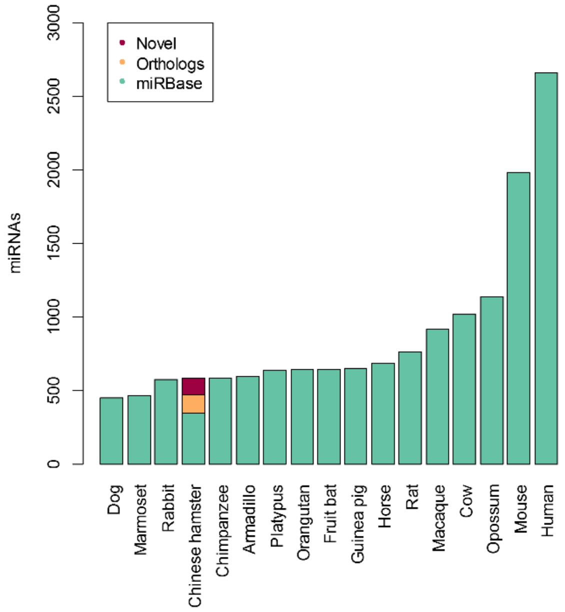

Figure 4.7 – miRNAs reported in mammalian species used as input to *miRDeep2*. miRNAs reported in miRBase are coloured in green. Chinese hamster miRNAs assembled in this chapter are coloured orange (those with mammalian orthologues) and red (novel miRNAs).

4.4.3 Differential expression of miRNAs

Reads were aligned to the CriGri1 CHO-K1 genome (Xu et al., 2011; Zerbino et al., 2018) using *Bowtie* resulting in an average of 13.9 million aligned reads per sample (65%, range 12-17.5 million). Of these aligned reads, 6.9 million reads per sample (32.7%, range 6.6-8.9 million) were assigned to annotated miRNAs generated using miRDeep2 in the previous section (Figure 4.8, full alignment statistics in Appendix 4A). An average of 12% fewer reads aligned to the genome in temperature shifted samples. Investigation of unaligned reads revealed 7.32% (7.06%-7.56%) of the reads in non-temperature shifted and 22.38% (21.63%-23.45%) of temperature shifted reads aligned to the 18S, 5.8S, 28S ribosomal subunits or the internal transcribed spacer sequences between these rRNA genes (GenBank: DQ235090.1). After removal of reads aligned to the transcribed spacer region an average of 13.83 million reads (11.45-15.34 million) aligned to the CriGri1 CHO-K1 genome with consistent alignment rates between 75.61% and 77.41%.

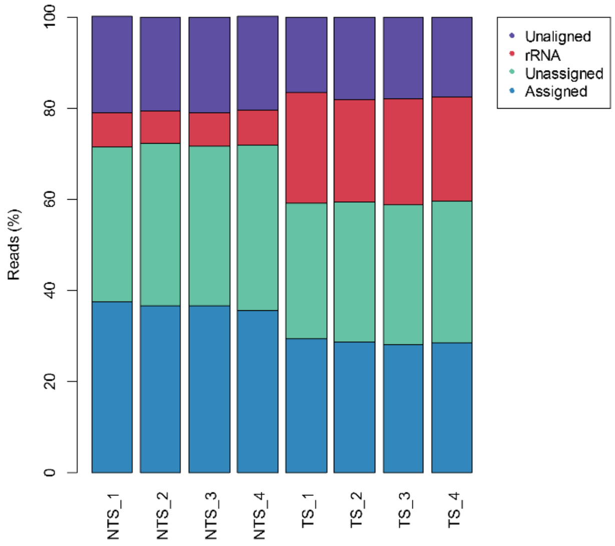

Figure 4.8 – Read alignment statistics. Preprocessed reads were aligned to the CriGri1 genome with no mismatches. 65% of reads were successfully aligned (green). 32.7% of the preprocessed reads were assigned to miRNAs detected using *miRDeep2*.

30 miRNAs were detected as differentially expressed (DE) at a ±≥ 1.2FC and a Benjamini Hochberg adjusted P value ≤ 0.05 - 10 upregulated and 2 downregulated. 7 of the upregulated and 17 of the downregulated were previously reported in CHO cells. 2 upregulated and 2 of the downregulated miRNAs were novel with orthologous mRNAs in mammals. 1 up and 1 down miRNA were novel and not previously reported in any species. miRNAs with a ≥±1.5 fold change are presented in Table 4.3. Full differential expression results are in Appendix 4D.

Table 4.3 Differentially expressed miRNAs with a ≥±1.5 fold change. 12 miRNAs were identified as differentially expressed with a ≥1.5 fold change and BH adjusted P value of ≤ 0.05.

miRDeep2 ID	miRDeep2 Score	miRBase miRNA	Orthologous miRBase Seed	Base Mean	Fold Change	BH Adj. P
JH001693.1_34177	5,046	-	miR-6895-3p	600	3.72	1.24×10^{-53}
JH007589.1_41503	10,990	cgr-miR-34a	miR-34a-5p	2,399	3.33	2.74×10^{-38}
JH000282.1_14656	1,35	cgr-miR-7b	miR-7-5p	251	3.01	2.49×10^{-23}
JH000003.1_357	198	cgr-miR-628	miR-628-5p	43	2.97	1.04×10^{-02}
JH000315.1_16059	1,325	-	bta-miR-2292	293	2.13	4.76×10^{-05}
JH000002.1_210	86,845	cgr-miR-215-5p	miR-192-5p	20,140	1.91	2.32×10^{-25}
JH000275.1_14328	2,808	-	-	6056	1.74	3.58×10^{-09}
JH000136.1_9442	269,390	cgr-miR-103-5p	miR-103a-3p	326	1.55	4.47×10^{-03}
JH000656.1_23790	44,731	cgr-miR-374-5p	miR-374a-5p	9,793	-1.52	4.35×10^{-04}
JH000344.1_17048	2,316	cgr-miR-9	mmu-miR-9-5p	240	-1.56	1.90×10^{-02}
JH000263.1_13955	3,112	cgr-miR-671-5p	miR-671-3p	445	-1.68	6.20×10^{-10}
JH000054.1_4298	6,638	cgr-miR-298-5p	mmu-miR-298-5p	1,262	-1.86	5.89×10^{-15}

4.4.4 Target prediction on differentially expressed genes

Differentially expressed miRNAs with a ≥±1.2 fold change which could be assigned a human miRBase identifier were used as input to *multiMiR* to identify miRNA targets in differentially expressed genes identified in Chapter 2 (Section 2.3.4). Criteria to accept miRNA:mRNA interactions were 1) present in a database validated with experimental evidence or 2) present in the TargetScan database plus at least one other database. Target predictions for upregulated miRNAs were found in 213 upregulated and 383 downregulated genes. Downregulated miRNAs were predicted to target 265 upregulated and 347 downregulated genes (Figure 4.9, Appendix 4E).

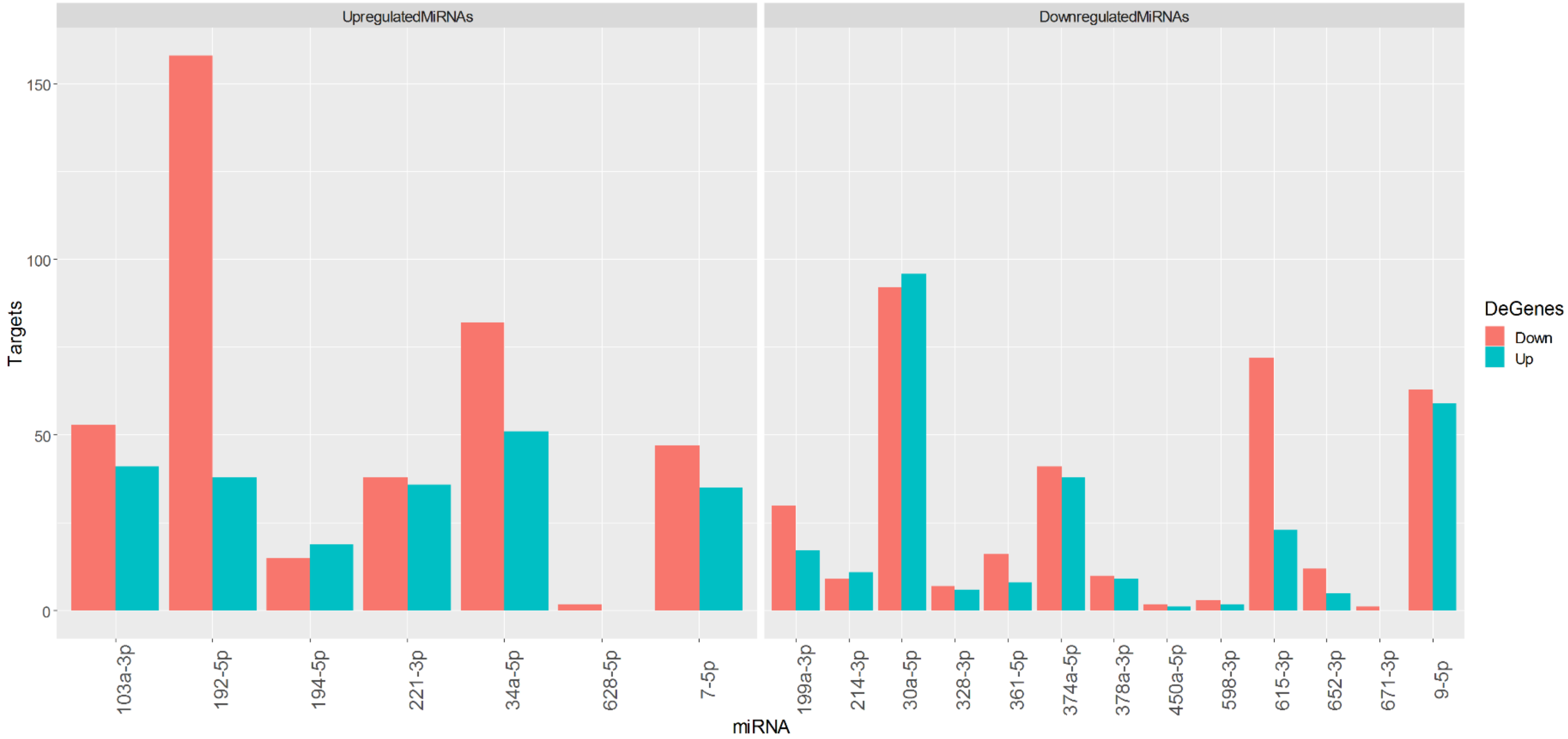

Figure 4.9 –Predicted targets for differentially expressed miRNAs. Target prediction using multiMiR identified 1,248 miRNA:mRNA interactions between differentially expressed miRNAs and genes. 'mir-' is omitted from the start of miRNA names on the x axis for clarity. Differentially expressed genes downregulated in response to mild hypothermia are coloured red and upregulated genes in turquoise.

A total of 726 differentially expressed genes were identified as targets of DE miRNAs – 441 by upregulated miRNAs and 456 by downregulated miRNAs. 195 genes were predicted as targets of both up and downregulated miRNAs. Gene ontology enrichment analysis identified 501 enriched GO terms with a BH adjusted P value ≤ 0.05 (Appendix 4F) with cell cycle, DNA replication and response to stress among the top enriched GO terms (Table 4.4).

Table 4.4 – Top 10 enriched GO terms for differentially expressed genes predicted as targets of miRNAs. *DAVID* was used to identify enriched GO terms among the 702 differentially expressed genes predicted as targets of miRNAs. 501 GO terms were identified as enriched with cell cycle related processes dominating the list.

Term	**Count**	**Fold Enrichment**	**BH Adj. P**
GO:0007049~cell_cycle	192	3.08	4.13×10^{-45}
GO:0022402~cell_cycle_process	170	3.36	2.75×10^{-44}
GO:0000278~mitotic_cell_cycle	133	3.69	1.40×10^{-37}
GO:1903047~mitotic_cell_cycle_process	123	3.71	1.08×10^{-34}
GO:0006259~DNA_metabolic_process	122	3.29	2.67×10^{-29}
GO:0006260~DNA_replication	63	5.83	5.86×10^{-27}
GO:0006974~cellular_response_to_DNA_damage_stimulus	101	3.36	2.37×10^{-24}
GO:0051276~chromosome_organization	125	2.84	4.07×10^{-24}
GO:0033554~cellular_response_to_stress	152	2.30	9.69×10^{-21}
GO:0000280~nuclear_division	79	3.64	1.24×10^{-20}

miRNA interactions between miRNAs and genes with anti-correlated expression (gene expression upregulated ≥ 1.5 fold and miRNA down ≥ 1.2 fold or vice versa) were identified in 491 genes. 30 of these genes were identified as targets of 3 or more anti-correlated miRNAs (Appendix 4G). Differentially expressed genes targeted by miRNAs DE in the opposite direction with roles in the cell cycle, energy metabolism or miRNA processing are presented in Table 4.5.

Table 4.5 – Differentially expressed genes targeted by miRNAs with differential expression in the opposite direction. Differentially expressed genes with roles in the cell cycle, energy metabolism or miRNA activity targeted by miRNAs are presented in this table. Full table in Appendix 4G.

CHO GeneID	Gene Symbol	Gene Name	Gene FC	miRBase ID	miRNA	miRNA FC
ENSCGRG00000006182	*Ccne1*	Cyclin E1	-1.72	MIMAT0000101	miR-103a-3p	1.55
				MIMAT0000222	miR-192-5p	1.91
				MIMAT0000252	miR-7-5p	3.01
ENSCGRG00000013335	*Ago4*	Argonaute 4, RISC Catalytic Component	-1.92	MIMAT0000101	miR-103a-3p	1.55
				MIMAT0000255	miR-34a-5p	3.33
				MIMAT0000278	miR-221-3p	1.40
ENSCGRG00000006644	*Pkia*	CAMP-Dependent Protein Kinase Inhibitor Alpha	-1.53	MIMAT0000101	miR-103a-3p	1.55
				MIMAT0000252	miR-7-5p	3.01
				MIMAT0000255	miR-34a-5p	3.33
				MIMAT0000278	miR-221-3p	1.40
ENSCGRG00000003592	*Cd44*	Cell Surface Glycoprotein CD44	2.07	MIMAT0000087	miR-30a-5p	-1.33
				MIMAT0000232	miR-199a-3p	-1.32
				MIMAT0000752	miR-328-3p	-1.41
ENSCGRG00000014167	*Cpeb4*	Cytoplasmic Polyadenylation Element Binding Protein 4	1.70	MIMAT0000087	miR-30a-5p	-1.33
				MIMAT0000232	miR-199a-3p	-1.32
				MIMAT0000271	miR-214-3p	-1.29
				MIMAT0000441	miR-9-5p	-1.56
				MIMAT0000727	miR-374a-5p	-1.52
ENSCGRG00000014709	*Ddit4*	DNA Damage Inducible Transcript 4	2.66	MIMAT0000087	miR-30a-5p	-1.33
				MIMAT0000232	miR-199a-3p	-1.32
				MIMAT0003283	miR-615-3p	-1.39

4.4.5 Target prediction on alternatively spliced genes

Genes identified as alternatively spliced in Chapter 3 were used for target prediction using *multiMiR* identifying 2,337 interactions on 1,274 genes. 445 GO terms were identified as enriched in these genes (Appendix I) with the most significantly enriched GO terms related to the cell cycle (Table 4.6).

Table 4.6 – Enriched GO terms in alternative spliced genes targeted by differentially expressed miRNAs. 445 GO terms were identified among the 1,274 genes targeted by differentially expressed miRNAs. The top 10 enriched GO terms are presented in this table (Full table in Appendix 4I).

GO Term	Count	Fold Enrichment	BH adj. FC
GO:0051276~chromosome_organization	197	2.43	2.87 × 10-^{29}
GO:0022402~cell_cycle_process	208	2.24	2.05 × 10^{-26}
GO:0000278~mitotic_cell_cycle	167	2.52	2.28 × 10^{-26}
GO:0006259~DNA_metabolic_process	169	2.48	4.83 × 10^{-26}
GO:0007049~cell_cycle	237	2.07	4.89 × 10^{-26}
GO:1903047~mitotic_cell_cycle_process	156	2.56	1.98 × 10^{-25}
GO:0033554~cellular_response_to_stress	239	1.97	3.23 × 10^{-23}
GO:0006974~cellular_response_to_DNA_damage_stimulus	141	2.55	1.25 × 10^{-22}
GO:0006260~DNA_replication	71	3.57	2.79 × 10^{-18}
GO:0006281~DNA_repair	99	2.73	3.05 × 10^{-17}

Of the 1,274 alternatively spliced genes targeted by differentially expressed miRNAs 838 were not identified as differentially expressed at the gene level. 43 of these genes were found to have a splicing event in the 3'UTR (Appendix 4J). Genes with a role in the cell cycle, energy metabolism and miRNA processing with predicted miRNA targets and splicing in the 3'UTR are presented in Table 4.7.

Table 4.7 – Alternative splicing events in 3’UTRs of genes not differentially expressed predicted as targets of differentially expressed miRNAs. 43 genes were identified which were alternatively spliced in the 3’UTR region, were not differentially expressed at the gene level and were targets of differentially expressed miRNAs. Presented in this table are 6 of these genes which have roles in the cell cycle, energy metabolism or miRNA processing (Full table in Appendix 4J).

GeneID	Gene Symbol	Gene Name	3'UTR	Splice Type	Alt Spliced Exon	Inc level Change	miRNA	miRNA FC
ENSCGRG00000010825	*Atp11a*	ATPase Phospholipid Transporting 11A	JH000452.1 613497-620166	SE	JH000452.1 618286-618390	0.15	miR-9-5p	-1.56
ENSCGRG00000004154	*Atp5e*	ATP Synthase F1 Subunit Epsilon	JH000054.1 2720910-2724092	RI	JH000054.1 2719746-2720912	-0.22	miR-615-3p	-1.39
ENSCGRG00000009745	*Acsl6*	Acyl-CoA Synthetase Long Chain Family Member 6	JH000154.1 333464-374921	SE	JH000154.1 344303-344381	-0.31	miR-34a-5p	3.33
							miR-374a-5p	-1.52
ENSCGRG00000001663	*CcnI*	Cyclin I	JH001903.1 117041-122170	SE	JH001903.1 121996-122227	-0.15	miR-9-5p	-1.56
ENSCGRG00000013702	*Grb2*	Growth Factor Receptor Bound Protein 2	JH000019.1 251982-299535	RI	JH000019.1 296047-297830	0.19	miR-7-5p	3.01
							miR-34a-5p	3.33
							miR-378a-3p	-1.28
ENSCGRG00000015964	*Hnrnpa2b1*	Heterogeneous Nuclear Ribonucleoprotein A2/B1	JH000116.1 674150-675059	SE	JH000116.1 674149-675059	-0.19	miR-30a-5p	-1.33
							miR-103a-3p	1.55
				A3SS	JH000116.1 674149-675476	-0.19	miR-361-5p	-1.3
							miR-615-3p	-1.39

Alternative splicing in the 3’UTR of transcripts has the potential to alter miRNA activity if splicing results in an interrupted miRNA binding site. An example of this occurring was identified in the *Hnrnpa2b1* gene where a skipped exon event results in decreased incorporation of a segment of the 3’UTR containing a binding site for miR-103-3p (Figure 4.10).

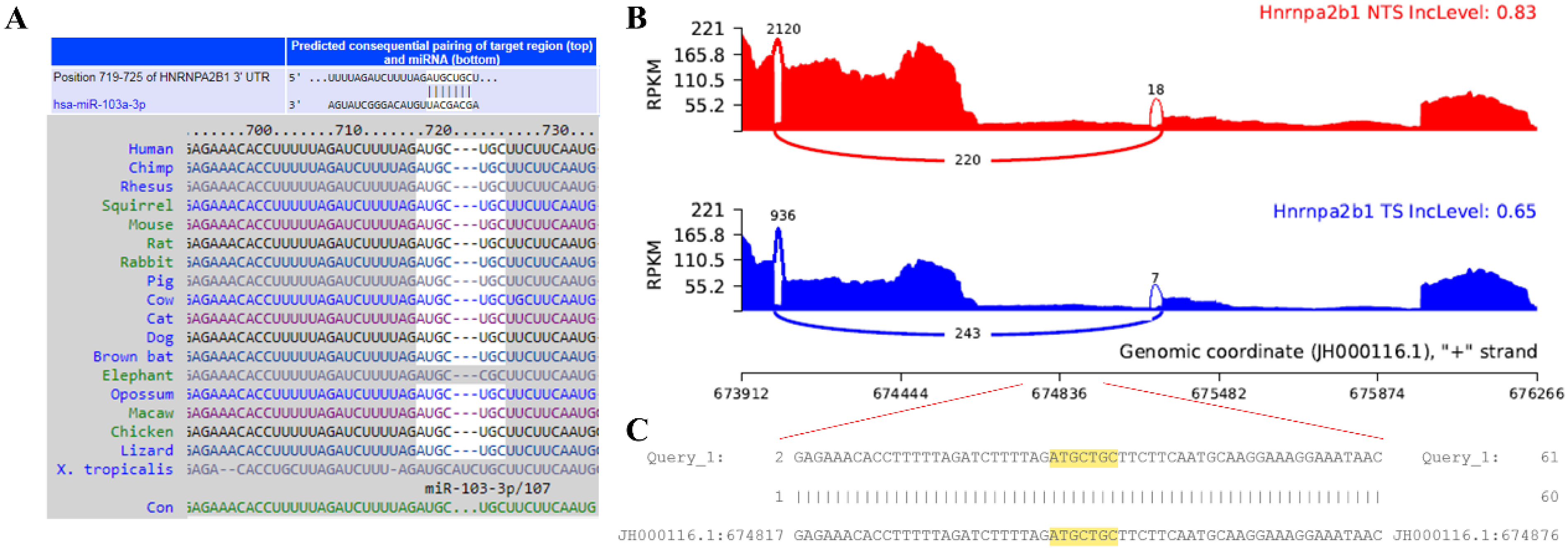

Figure 4.10 – Alternative splicing alters miRNA targeting of miR-103-3p on *Hnrnpa2b1*. An interaction between miR-103-3p and *Hnrnpa2b1* was identified using the approach outlined in Section 4.3.5. *Hnrnpa2b1* was not identified as differentially expressed in Chapter 2 but was identified as undergoing alternative splicing in Chapter 3. **(A)** The predicted target site of miR-103-3p on *Hnrnpa2b1* is conserved among mammalian species (Agarwal et al., 2015). (**B**) A skipped exon event is identified in the 3'UTR of the transcript which results in an exon being skipped at a higher rate in cells cultured at a reduced temperature (0.65 percent spliced in) than non-temperature shifted cells (0.83 PSI). (**C**) The target site for miR-103-3p is conserved in CHO cells and present in the skipped exon.

4.4.6 Target Prediction on differentially expressed proteins

Target predictions were carried out on the 171 genes encoding differentially expressed proteins identified in Chapter 2 (Section 2.3.5), revealing 139 miRNA interactions on 86 proteins (Appendix K). Gene ontology analysis was applied to differentially expressed proteins with predicted miRNA targets identifying 49 enriched GO terms (Appendix 4L). The most significantly enriched GO terms include those related to protein localisation (Table 4.8).

Table 4.8 – Top 10 enriched GO terms for differentially expressed proteins with miRNA targets. *DAVID* was used to identify enriched GO terms among the 86 differentially expressed proteins with miRNA targets predicted. The full table of 49 enriched GO terms is presented in Appendix 4L.

Term	**Count**	**Fold Enrichment**	**BH Adj. P**
GO:0046907~intracellular_transport	25	3.18	8.54×10^{-04}
GO:1902589~single-organism_organelle_organization	25	3.15	5.25×10^{-04}
GO:0008104~protein_localization	31	2.57	5.18×10^{-04}
GO:0043933~macromolecular_complex_subunit_organization	31	2.56	4.36×10^{-04}
GO:0034613~cellular_protein_localization	24	3.05	7.33×10^{-04}
GO:0070727~cellular_macromolecule_localization	24	3.02	7.03×10^{-04}
GO:0045184~establishment_of_protein_localization	27	2.71	7.07×10^{-04}
GO:0044085~cellular_component_biogenesis	33	2.30	8.07×10^{-04}
GO:0072594~establishment_of_protein_localization_to_organelle	15	4.57	1.08×10^{-03}
GO:0033036~macromolecule_localization	32	2.31	9.75×10^{-04}

In Chapter 2 (Section 2.3.5), 153 genes were identified with uncorrelated differential expression at the gene and protein levels. Of the differentially expressed proteins identified with miRNA targets, 68 did not have corresponding differential expression at the gene level (Appendix 4M). In addition, 9 genes identified as differentially expressed at the gene level but not the protein level were predicted as targets of differentially expressed miRNAs (Appendix 4M). Potential miRNA activity was identified by overlapping fold changes of genes, proteins and predicted miRNA targets to find cases where 1) an upregulated miRNA targets a post-transcriptionally downregulated gene or 2) a downregulated miRNA targets a post-transcriptionally upregulated gene. By applying these criteria 64 cases of miRNA regulation were identified in 45 genes which were not correlated at the gene and protein level (Appendix 4M). Genes post-transcriptionally regulated by differentially expressed miRNAs with roles in RNA-processing, energy metabolism and translation are presented in Table 4.9.

Table 4.9 – miRNA mediated control of gene expression contributes to the lack of correlation between levels of gene and protein expression. Presented in this table are post-transcriptionally upregulated genes targeted by downregulated miRNAs which have roles in energy metabolism, RNA processing and translation.

Gene ID	Gene Symbol	Gene Name	Gene FC	Protein FC	miRBaseID	miRNA	miRNA FC
ENSCGRG00000000751	*Acly*	ATP Citrate Lyase	1.03	1.28	MIMAT0000703	miR-361-5p	-1.30
ENSCGRG00000014672	*Aldh2*	Aldehyde Dehydrogenase 2 Family Member	1.08	1.22	MIMAT0003283	miR-615-3p	-1.39
ENSCGRG00000010880	*Dsx3x*	DEAD-Box Helicase 3 X-Linked	1.07	1.37	MIMAT0000703	miR-361-5p	-1.30
					MIMAT0000441	miR-9-5p	-1.56
ENSCGRG00000016658	*Eif5*	Eukaryotic Translation Initiation Factor 5	1.48	1.33	MIMAT0000441	miR-9-5p	-1.56
ENSCGRG00000001864	*Eno1*	Enolase1	1.14	1.20	MIMAT0000732	miR-378a-3p	-1.28
					MIMAT0003283	miR-615-3p	-1.39
ENSCGRG00000015236	*Fsan*	Fatty Acid Synthase	1.05	1.20	MIMAT0003283	miR-615-3p	-1.39
ENSCGRG00000013953	*Got1*	Glutamic-Oxaloacetic Transaminase 1	-1.06	1.37	MIMAT0000441	miR-9-5p	-1.56
ENSCGRG00000014824	*Hk1*	Hexokinase 1	1.15	1.38	MIMAT0000087	miR-30a-5p	-1.33
					MIMAT0003283	miR-615-3p	-1.39
ENSCGRG00000017777	*Ogdh*	Oxoglutarate Dehydrogenase	-1.66	-1.19	MIMAT0000087	miR-30a-5p	-1.33
					MIMAT0000703	miR-361-5p	-1.30
ENSCGRG00000005301	*Sf3b1*	Splicing Factor 3b Subunit 1	1.19	1.34	MIMAT0003283	miR-615-3p	-1.39
ENSCGRG00000010128	*Sfpq*	Splicing Factor Proline And Glutamine Rich	1.09	1.37	MIMAT0003283	miR-615-3p	-1.39

4.5 Discussion

Multi-omics has proven previously to be valuable in identifying miRNA activity in the study of CHO cell growth rate (Clarke et al., 2012) however this technique has yet to be utilised using next generation sequencing to study the role of miRNAs in the CHO cell response to reduced culture temperature. Presented in this Chapter is the usage of the highest sequencing depth to date for profiling small RNAs in CHO cells - over 4 times as deeply sequenced as other miRNA profiling experiments carried out in CHO cells (Hackl et al., 2011). Sequencing samples to a depth of 55 million reads and the usage of 4 biological replicates ensures confidence in the assembly of novel miRNAs as well as identification of miRNAs which are expressed at a low enough expression level such they would have been missed in previous experiments. In addition to a high sequencing depth the sequencing libraries generated in this study are of extremely high quality, with an average sequencing quality of Q34 (a probability of 1 in 4,000 bases sequenced called incorrectly). After trimming adapters and barcode sequences, libraries were heavily enriched for RNAs 21-23 nucleotides in size (the same range as the majority of CHO cell miRNAs in miRBase) suggesting library preparation successfully enriched for miRNAs. Read s were aligned to the genome for the identification and quantification of miRNAs revealing 10% higher alignment rates to the genome in non-temperature shifted samples. Alignment of these unaligned reads to the rRNA sequences revealed the presence of reads aligned to the external transcribed spacer sequence which is not present in the CriGri1 genome assembly. After removal of these reads consistent alignment rates to the genome were observed across samples with an average of 13.83 million reads aligned per sample – indicating the 10% difference in alignment rate was due to increased rRNA derived reads in temperature shifted samples which were removed during this pre-processing.

In total 251 miRNAs not currently annotated in miRBase were identified with scores exceeding the *miRDeep2* thresholds for likely miRNAs with a confidence of 85±2%. At this confidence level a maximum of 36±6 of these miRNAs are estimated to be false positives however *miRDeep2* applies a conservative approach which overestimates false positives and therefore the true number of false positives is likely lower than this estimate (Friedländer et al., 2008). Given the worst case-scenario of false positives, an estimated 209 true positive miRNAs were identified in this study - increasing the number of mature miRNA identified in CHO cells by at least 60% and the number of precursors by 85%

(Kozomara and Griffiths-Jones, 2014). Among these candidate miRNAs are 132 which have seed regions orthologous to miRNAs reported in other mammalian species used as input to *miRDeep2*. Orthologous miRNAs include miR-290a-5p, miR-335, miR-383-3p, miR-448-3p, miR-873-3p and miR-939-5p. Each of these miRNAs play a role in inhibiting cellular proliferation and apoptosis (Han et al., 2018a; Shu et al., 2011; Su et al., 2018; Zhao et al., 2017) and therefore illustrate the importance of identifying previously uncharacterised miRNAs in understanding the CHO cell response to temperature shift.

miR-290a-5p is a rodent-specific miRNA which has only been reported in mouse, guinea pig and rat previously in miRBase (Paikari et al., 2017). miR-290 is expressed in a cluster of the miRs-290:295 in rodent germ cells – we also identified miR-295 for the first time in CHO cells through this experiment suggesting miR-290 and miR-295 are co-expressed during temperature shift (Wu et al., 2014). miR-290-295 cluster miRNAs are highly expressed in mouse embryonic stem cells and plays a role in early cellular differentiation (Wu et al., 2014). miRNAs from this cluster are known to regulate the G1/S phase transition of the cell cycle and to promote rapid proliferation through post-transcriptionally regulating the expression of *Cdkn1a* (Wang et al., 2008). miR-290-295 cluster miRNAs also play a role in suppression of apoptosis through targeting of caspase mRNAs (Zheng et al., 2011a). miR-939 has previously only been reported in Simiiformes of the Primate order in miRBase (humans, chimpanzees, orangutans and macaques). Expression of miR-939 supresses the cell cycle through the inhibition of the Raf/MEK/ERK pathway and with under expression of miRNA associated with cancers in humans as well as chemo-resistance (Zhang et al., 2017b). The expression of a novel miR with the same mature seed as miR-939 presents the first evidence of miR-939 expression in mammalian species outside the simian order.

Differential expression analysis revealed 30 miRNAs differentially expressed in response to mild hypothermia - 10 upregulated and 20 downregulated. Upregulated miRNAs include miR-7b, miR-34a, miR-103, miR-215 and miR-628-5p which have roles in the regulation of proliferation (Li et al., 2018; Vychytilova-Faltejskova et al., 2017). miR-7 expression has previously been shown to result in decreased growth in CHO cells (Sanchez et al., 2013) as well as increased recombinant protein productivity (Barron et al., 2011). miR-7 has previously been identified as downregulated in temperature shifted CHO cells but when overexpressed transiently results in a halting of the cell cycle (Barron et al., 2011). Conflicting reports show that both overexpression (Barron et al., 2011; Meleady et al., 2012) and depletion (Sanchez et al., 2014) of miR-7 in CHO cells can result in increased productivity. The mechanism of miR-7s effect on cell growth is through targeting genes such as *Skp2* and *Psme3* which regulate p27^{KIP} expression to control of the G1/S phase transition of the cell cycle (Sanchez et al., 2013). Studies have shown miR-7 to regulate the expression of the *Fos* gene (Lee et al., 2006), a component of the transcription factor AP-1 which regulates cell proliferation and apoptosis. *Fos* was identified as the most downregulated gene in Chapter 2 (Section 2.3.4) with a -10.17 FC which may be a result of miR-7 activity. As conflicting reports have shown both upregulation and depletion of miR-7 to positively impact productivity in CHO cells, miR-7 expression alone may not be sufficient to mimic the effects of temperature shift on productivity. miR-34a is a master regulator of tumour suppression whose expression is induced by p53 and regulates proliferation and cell survival genes including *Ccnd1* and *Cdk6* (Kelly et al., 2015b; Saito et al., 2015; Slabáková et al., 2017). miR-34a has previously been utilised as a target for cellular engineering in CHO cells, with ectopic overexpression resulting in a halting of the cell cycle and depletion of miR-34a increasing product yields (Kelly et al., 2015b). mir-34a was not observed as overexpressed in previous CHO cell temperature shift experiments (Barron et al., 2011; Gammell et al., 2007) however our findings suggest differential expression of this highly expressed miRNA contributes to the CHO cell response to mild hypothermia in conjunction with miR-7. Upregulation of miR-34a and miR-7 may contribute to the decreased growth rate we observed in temperature shifted cells through regulation of genes which regulate the G1/S phase transition including *Skp2, Psme3* and *Ccnd1* (Sanchez et al., 2013; Sun et al., 2008).

The observed reduction in growth rate induced by temperature shift may also be contributed to by downregulation of proliferation inducing miRNAs. miR-9 is a positive regulator of proliferation which has been shown to target *Sox2* (Zeng et al., 2018) and *Foxo1* (Liu et al., 2017). *Sox2* is a transcription factor which regulates differentiation and growth genes including the epidermal growth factor receptor *Egfr* (Chou et al., 2013) and *Ccnd1* (Luo et al., 2018) while the *Foxo1* transcription factor regulates the cell cycle through supressing expression of *Cdkn1a* (Xie et al., 2012) and *Cdkn1b* (Coomans de Brachène and Demoulin, 2016). Inhibition of miR-9 in MCF7 cells has previously been shown to supress proliferation (Liu et al., 2017) and therefore temperature shift induced downregulation of miR-9 at a -1.56 FC may contribute to reduced growth rate we observe in cells cultured at a reduced temperature.

Downregulation of miRNAs may also contribute to the temperature shift induced alterations to metabolism observed. miR-199a is a regulator of metabolic switching between primary use of a Warburg-type metabolism to OXPHOS (el Azzouzi et al., 2013; Zhang et al., 2015). miR-199a has been shown to target hexokinase-2 (*Hk2*), pyruvate kinase-M2 (*Pkm2*) (Zhang et al., 2015) and peroxisome proliferator-activated receptor delta (*Pparδ*) (el Azzouzi et al., 2013) which regulate Warburg metabolism. Knockdown of miR-199a using miR antagonists results in decreased glucose uptake (Zhang et al., 2015) and a shift from glycolysis to increased utilisation of fatty acids and OXPHOS (el Azzouzi et al., 2013). miR-199a was identified as downregulated with a -1.32 FC in temperature shifted CHO cells. Downregulation of miR-199a has been previously reported in temperature shifted CHO cells (Barron et al., 2011) however the phenotypic effects of altering miR-199a expression in CHO cells has not yet been studied.

While differential expression results were used to identify both up and downregulated miRNAs which may contribute to the CHO cell response to reduced culture temperatures the gene targets through which these differentially expressed miRNAs have not been studied in CHO cells. Target prediction algorithms are known to suffer from a high false positive rate when targets are experimentally tested for verification (Pinzón et al., 2017). Additionally databases such as TargetScan which contain target predictions do not have Chinese hamster specific interactions predicted and therefore to use these predictions we must rely on homology between mammalian species. Mammalian miRNAs are typically highly conserved, particularly in the seed region and their target sites on mRNAs, and therefore identified miRNA:mRNA interactions should be transferable between mammalian species (Friedman et al., 2009). To reduce false positives, target predictions were restricted to differentially expressed genes, miRNAs and proteins as anti-correlated expression can be used as evidence that a predicted interaction is biologically occurring. In addition only target predictions from databases containing experimental evidence of an interaction such as luciferase reporter assays or interactions present in *TargetScan* and minimum of one other database were utilised. *TargetScan* has been shown to outperform other algorithms for the predication of miRNA targets and therefore predicted targets were limited to only those in the *TargetScan* database (Agarwal et al., 2015).

Target predictions were identified in 726 differentially expressed genes (286 upregulated genes and 441 downregulated) which were enriched for cell cycle related and stress response GO terms. Target predictions between miRNAs and genes with opposite fold change directions were identified in 491 genes of which 30 had at least 3 anti-correlated miRNA target predictions. Differentially expressed genes targeted by at least 3 anti-correlated miRNAs include those with roles in the cell cycle and energy metabolism. The -1.72 fold downregulated gene Cyclin E1 (*Ccne*1) was highlighted in Chapter 2 as a crucial regulator of the cell cycle in the G1/S phase (Krivega et al., 2015). The upregulated miRNAs miR-7, miR-103a and miR-192 (1.55, 1.91 & 3.01 FC respectively) were identified as targets of *Ccne1*. Collectively these miRNAs may contribute to the downregulation of *Ccne1* by miRNA mediated transcript decay (Iwakawa and Tomari, 2015). miR-7 and miR-103 have been experimentally validated as directly targeting the 3'UTR of *Ccne1* using luciferase reporter assays (Liao and Lönnerdal, 2010; Zhang et al., 2014). The cAMP-dependent protein kinase inhibitor alpha (*Pkia*) gene is an inhibitor of cAMP-dependent protein kinase A (Pka) which regulates OXPHOS by phosphorylating mitochondrial proteins (Acin-Perez et al., 2009). Downregulation of *Pkia* has been shown to result in metabolic switching from a Warburg type metabolism to OXPHOS in cancer cells (Xing et al., 2017). The *Pkia* gene was identified as downregulated -1.53 fold in temperature shifted cells (Section 2.3.4) and predicted as a target of 4 upregulated miRNAs – miR-103a, miR-7, miR-34a and miR-221 (1.55, 3.01, 3.33, 1.4 FC respectively). Upregulation of these miRNAs in response to decreased culture temperature may contribute to decreased *Pkia* expression resulting in increased Pka activity and higher rates of OXPHOS.

The potential impact of alternative splicing identified in Chapter 3 on miRNA targeting was assessed by performing target predictions on differentially spliced genes. 1,274 genes were predicted as targets of differentially expressed genes – of which 838 were not differentially expressed at the gene level. Alternative splicing has the potential to alter miRNA binding sites if splicing affects the 3'UTR region of a transcript (Han et al., 2018b). Of the genes undergoing alternative splicing and predicted as targets of differentially expressed miRNAs 43 were spliced in the 3'UTR which may alter their binding efficiency. Genes with roles in the cell cycle and energy metabolism were among these targets including ATP synthase subunit epsilon (*Atp5e*), acyl-CoA synthetase long chain family member 6 (*Acsl6*), cyclin I (*Ccni*) and growth factor receptor-bound protein 2 (*Grb2*). Atp5e is a subunit of the ATP producing complex V of the mitochondrial electron transport chain which carries out OXPHOS (Hurtado-López et al., 2015). Downregulation of *Atp5e* is associated with reduced rates of OXPHOS and a Warburg type metabolism in cancers (Hurtado-López et al., 2015). An intron retention event was observed as downregulated with a -0.22 inclusion rate of the alternative splicing event in temperature shifted cells. *Atp5e* was predicted as a target of the downregulated miR-615 which was differentially expressed at a -1.39 FC. The cyclin gene *Ccni* has a role in promoting progression of the cell cycle through the G2/M phase (Nagano et al., 2013) and was identified as alternatively spliced with an exon skipping event in the 3'UTR upregulated in temperature shift with a 0.15 increased inclusion rate. *Ccni* was predicted as a target of the downregulated miR-9 (-1.56 FC) – the targeting efficiency of miR-9 on *Ccni* may be altered by this alternative splicing event. Alternative splicing may therefore regulate the binding of miRNAs to their mRNA targets and contribute to the CHO cell response to reduced temperature through altering post-transcriptional regulation of cell cycle and energy metabolism.

Analysis of the location at which an miRNA is predicted to target human miRNAs in the TargetScan database which are conserved across species revealed a binding site for miR-103a in the 3'UTR of the *Hnrnpa2b1* gene. The miRNA binding site was found to be conserved in CHO cells at the chromosomal location JH000116.1:674,817-674,876 with 100% sequence identity conserved between mammals, birds and reptiles. A differential splicing event was observed to overlap this region resulting in the exon being skipped at a higher rate in CHO cells cultured at a reduced temperature (-0.2 PSI). Upregulation of this exon skipping event may result in *Hnrnpa2b1* evading post-transcriptional silencing by the upregulated miR-103a (1.55 FC). Hnrnpa2b1 is an RNA processing protein which binds RNA containing N6-methyladenosine (m^6A) modifications to regulate alternative splicing (Alarcón et al., 2015) or m^6A modified miRNAs to promote their processing by recruitment of Dgcr8 and Drosha (Wu et al., 2018). m^{6}A modifications are particularly prevalent in pri-miRNAs and the efficient processing of the majority of miRNAs is dependent on m^{6}A marking by the methyltransferase like 3 (Mettl3) enzyme (Alarcón et al., 2015). If *Hnrnpa2b1* transcripts evade post-transcriptional suppression due to alternative splicing resulting in skipping of the miR-103a binding site then increased *Hnrnpa2b1* activity may contribute to the differential expression of miRNAs and alternative splicing events observed in CHO cells cultured at reduced temperature. Examples of alternative splicing eliminating a miRNA binding site have been reported previously in plant species (where miRNAs binding sites are typically in the 5'UTR or coding sequence) (Park and Grabau, 2017; Yang et al., 2012) however this is the first reported case of an exon skipping in the 3'UTR affecting mammalian miRNA targeting.

Overlapping differential gene and protein expression results in Section 2.3.5 revealed a set of 153 genes identified as differentially expressed at only either the protein or gene level in the same direction suggesting the influence of post-transcriptional regulation. miRNA regulation likely contributes to the post-transcriptional regulation of these genes as miRNA targeting was able to explain 29.4% of these cases due to up or downregulated miRNAs predicted to target genes post-transcriptionally regulated in the opposite direction. Likely cases of miRNA targeting were identified in genes with roles in energy metabolism, regulation of splicing and translation. Increased translational efficiency was identified in regulators of mitochondrial energy metabolism including aldehyde dehydrogenase 2 (*Aldh2)* (Kang et al., 2016), aspartate aminotransferase (*Got1)* (Abrego et al., 2017) and oxoglutarate decarboxylase (*Ogdh*) (Bunik et al., 2016). Each of these enzymes are involved in mitochondrial metabolism by generating fuel for OXPHOS and have been implicated in the Warburg metabolic phenotype in cancer cells (Abrego et al., 2017; Bunik et al., 2016; Kang et al., 2016). *Aldh2* was unchanged at the gene (1.08 FC) but differentially expressed at the protein level (1.22 FC) and is targeted by miR-615 (-1.39 FC). *Got1* was also unchanged at the gene level (-1.06 FC) but upregulated at the protein level (1.37 FC) and predicted as a target of the downregulated miR-9 (-1.56 FC). *OGDH* was downregulated at the gene level (-1.66 FC) but did not pass the threshold set for differential protein expression (-1.19 FC) and identified as a target of the miRNAs mir-30 and miR-361 (-1.33 and -1.3 FC respectively). Decreased post-transcriptional suppression of these genes due to downregulation of miRNAs targeting them may increase CHO cells capacity to carry out OXPHOS during mild hypothermia by producing acetyl-CoA and NADH at a faster rate.

Assembly of novel miRNAs, identification of differential miRNA expression and target prediction on differentially expressed and spliced mRNAs has resulted in the identification of miRNAs expressed in the temperature shift response which may serve as valuable targets for genetic engineering. miR-290a-5p and miR-295-5p were both identified for the first time in CHO cells and their expression in other species has been demonstrated to regulate the cell cycle (Wang et al., 2008) and supress apoptosis (Zheng et al., 2011a). Increasing the expression of miR-290a-5p and miR-295-5p at the onset of the stationary phase may enhance the effects of temperature shift in CHO cells and further increase viability and productivity. Similarly, inhibition of miR-9 expression which was identified as downregulated in temperature shifted CHO cells may suppress proliferation (Liu et al., 2017) and enhance the temperature shift phenotype to result in more efficient producer cell lines. The Warburg type metabolism may also be able to be reverted to primary utilisation of OXPHOS through manipulating the expression of miRNAs such as miR-199a (el Azzouzi et al., 2013; Zhang et al., 2015). Downregulation of miR-199a in temperature shifted CHO cells was first reported in 2011 (Barron et al., 2011) and confirmed in this experiment. Finally, expression of miR-103a-5p may enhance the temperature shift phenotype and increase productivity through inhibiting the cell cycle (Liao and Lönnerdal, 2010), and driving a switch from a Warburg to OXPHOS metabolism by regulating *Pkia* (Xing et al., 2017).

While miRNA target predictions were able to explain the lack of correlation between gene and protein levels of expression in enzymes involved in mitochondrial energy metabolism (which may contribute to the decrease in glucose consumption and lactate production observed in temperature shifted CHO cells), 71.6% of the genes with uncorrelated expression were not predicted as targets of differentially expressed miRNAs. In addition, proteomics experiments were able to detect only 2.75% of the genes identified as expressed in transcriptomics. In order to obtain a more complete view of post-transcriptional gene expression alternative techniques must be utilised which are capable of identifying forms of post-transcriptional regulation of gene expression besides miRNA targeting such as differential translation efficiency.

4.6 Conclusion

Presented in this Chapter is an analysis of small RNA sequencing data generated from CHO cell cultures shifted from 37°C to 31°C. RNA-Seq data generated during this experiment represents the highest sequencing depth utilised to study CHO cell miRNAs to date at over 4 times the sequencing depth of previous experiments - enabling characterisation of 251 previously unannotated miRNAs. miRNA target prediction on differentially expressed genes revealed miRNA targets on genes involved in regulation of the cell cycle and energy metabolism which had inversely correlated miRNA target expression. miRNA target prediction on alternatively spliced genes revealed altered miR-103a targeting in the *Hnrnpa2b1* due to a skipped exon event in temperature shifted CHO cells resulting in the lack of a miRNA binding site – the first event of its kind to be reported outside of plant species. Finally, miRNA targeting was able to explain 29.4% of the genes undergoing post-transcriptional regulation of gene expression identified in Chapter 2. Genes with functions involved in mitochondrial energy metabolism were identified which had no gene level expression change induced by temperature shift but were upregulated at the protein level – potentially due to downregulation of miRNAs predicted to be their targets. Together these findings enhance our understanding of post-transcriptional control of gene expression in CHO cells cultured at a reduced temperature however further investigation is required to elucidate the mechanisms contributing to the lack of correlation between measurements of the transcriptome and proteome in genes which were not identified as targets of differentially expressed miRNAs.

Chapter 5

Translational response associated with reduced CHO cell culture temperature

5.1 Summary

The response of CHO cells to a reduction in cell culture temperature is mediated at both the transcriptional and post-transcription levels as evidenced by mRNA profiling and protein expression analysis. It is likely that this phenomenon is due to a combination of both biological and technical factors – the role of post-transcriptional gene expression, as well as an incomplete profile of the proteome. In Chapter 4 we examined the potential role of miRNAs in regulating gene expression and a number of miRNAs were identified as differentially expressed and found to target anti-correlated mRNAs which were not differentially expressed at the protein level. 58.8% of the genes undergoing post-transcriptional control of gene expression were however not explained by miRNA activity or alternative splicing. In this Chapter, a cutting edge next generation sequencing method called ribosome footprint profiling (RiboSeq) is applied.

5.2 Introduction

5.2.1 Translation

In order for a transcribed mRNA to be processed into a protein its nucleotide sequence must be translated by a ribosome (Watson et al., 2014). The eukaryotic ribosome is a large molecular machine consisting of 6 ribosomal RNAs which contribute approximately half of the mass of a ribosome as well as 80 proteins including structural proteins and those with roles in unwinding mRNA and signalling (Wilson and Doudna Cate, 2012). Translation is a carefully orchestrated sequence of events modulated by the sequential binding of RNA-binding proteins to mRNAs and recruitment of ribosomal RNAs which can be split into 3 major phases – initiation, elongation and termination. Initiation is the process of assembling initiation factors within the 5'UTR of the mRNA to recruit the ribosome to the start codon to begin translation. The majority of translation events are initiated by binding of translation initiation factors to the 7-methylguanylate cap and poly(A)-tail which are added to the 5' and 3' ends of a transcript respectively during the processing of a pre-mRNA. (Chu et al., 2016). mRNAs are prepared for ribosome binding by the eIF4F complex which includes eIF4E, eIF4G and eIF4A. The 5' cap is first bound by the initiation factor eIF4E while poly(A) binding protein (PABP) binds the poly(A)-tail (Cheng and Gallie, 2013). The 5' and 3' ends of the transcript are bridged by the binding of eIF4G, a large scaffolding protein, to eIF4E and PABP which results in the circularisation of the transcript (Chu et al., 2016). eIF4G also contains binding sites for the helicase eIF4A which is recruited and simulated by eIF4B to unwind any secondary RNA structures in the 5’UTR into an open-state to allow access ribosomal access (Andreou et al., 2016; Sokabe and Fraser, 2017).

Two ribosomal subunits known as the large (60S) and small (40S) subunits assemble into the functional ribosome complex (80S) capable of translating protein. The major sites of translation activity occur in 3 clefts within the 80S ribosome which contain binding sites for tRNAs - known as the aminoacyl (A), peptidyl (P) and exit (E) sites (Watson et al., 2014). During translation amino acids are transported to the ribosome by binding of aminoacyl-tRNAs to the A site by complementarity between codons in mRNA and the anticodons on the tRNA. While in the cytoplasm, the small ribosomal subunit is assembled into the 43S pre-initiation complex through binding of the ternary complex, eIF1, eIF3 and eIF5 (Chu et al., 2016). The ternary complex consists of a methionyl-tRNA, GTP and the initiation factor eIF2 which positions the methionyl-tRNA in the P site of the small subunit (Kolitz and Lorsch, 2010). The initiation factors eIF1 and eIF1a assist in the assembly of the pre-initiation complex by binding the E Site and A site of the 40S subunit and interacting with eIF2 to recruit methionyl-tRNA to the P site (Aitken and Lorsch, 2012). eIF1 however blocks the tRNA from fully entering the P site, keeping the 40S subunit in an open conformation which mRNA can move through (Obayashi et al., 2017). eIF3 binds the 40S subunit and eIF5 binds eIF3 to complete the assembly of the pre-initiation complex (Aitken and Lorsch, 2012). The 43S pre-initiation complex can then bind mRNA in the 5’UTR through interaction of eIF3 and eIF5 with eIF4G to form the 48S initiation complex (Chu et al., 2016; Pisareva and Pisarev, 2014).

Once bound to the 5’UTR of an mRNA the 48S initiation complex scans the mRNA in an open conformation form the 5’ to the 3’ direction until an AUG start codon is present in the P site which binds to the anticodon of the methionyl-tRNA through complementary base pairing (Lomakin and Steitz, 2013). The base pairing between the start codon and anticodon results in the GTP bound to the tertiary complex to be hydrolysed into GDP, releasing components of the pre-initiation complex including eIF1, eIF1A, eIF2, eIF3 and eIF5, freeing the E and A sites and enabling the tRNA to fully enter the P site. Binding of the initiator methionyl-tRNA results in a conformational change to a closed form of the 40S subunit which the 60S subunit can bind to form the 80S ribosome (Obayashi et al., 2017). The process of initiation is illustrated in Figure 5.1.

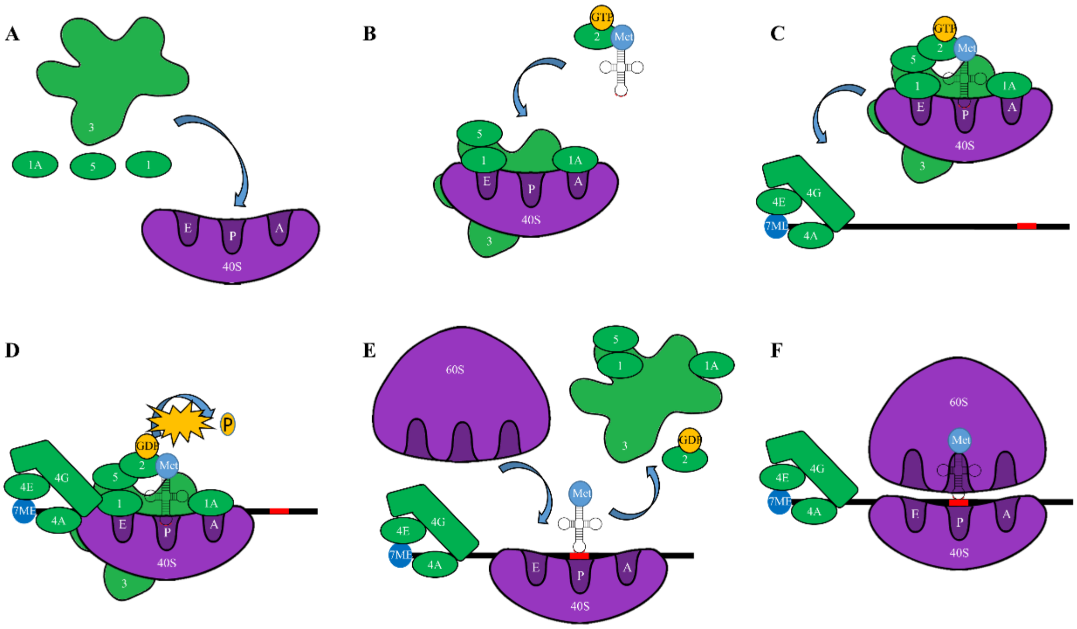

Figure 5.1 – Translation Initiation. (A) The pre-initiation complex is assembled on small ribosomal subunits in the cytoplasm by binding of the initiation factors eIF3, eIF1a, eIF1 and eIF5. (B) Methylacyl-tRNA-GTP-eIF2 is incorporated into the P site, assisted by the interaction of eIF2 and eIF5. (C) The pre-initiation complex is then recruited to the 5' of an mRNA with eIF4F assembled. (D) Upon binding of the complex to mRNA, GTP is hydrolysed to GDP. (E) GTP hydrolysis signals for the release of initiation factors eIF3, eIF1, eIF1a, eIF5 and eIF2 which releases the 40S subunit from eIF4F to scan the mRNA until the Methylacyl-tRNA base pairs to a start codon in the P site. (F) The 60S ribosomal subunit can then form a complex with the 40S subunit ready for elongation to begin.'eIF' omitted from initiation factors (green) for clarity. Adapted from (Watson et al., 2014).

Translation elongation follows initiation and is the stage of translation during which amino acids are incorporated into a growing peptide chain (Watson et al., 2014). Translation elongation factors (eEF's) assist the ribosome in driving the process of elongation by recruiting aminoacyl-tRNAs to be incorporated (Li et al., 2013). Aminoacyl-tRNAs in the cytoplasm are bound by eEF1a-GTP which upon binding of the tRNA anticodon to the codon in the A site of the ribosome hydrolyse to GDP. The aminoacyl-tRNA is then released from eEF1a-GDP and a peptide bond can be formed between the amino acid in the A and P site, releasing it from the tRNA (Li et al., 2013). The tRNA in the P and A site are then shifted to the E and P site respectively while the mRNA within the ribosome is also shifted by 3 nucleotides to present the next codon at the A site – a process called translocation which is driven by the elongation factor eEF2 (Zhou et al., 2014b). The formation of peptide bonds between amino acids in the P site and the A sites contribute to the movement of an mRNA through a ribosome as when the peptide bond is formed the tRNA in the P site becomes deacetylated. The deacetylated tRNA in the P site then has a higher affinity for the E site whilst the tRNA in the A site (still attached to the peptide chain) has a higher affinity for the P site (Watson et al., 2014). eEF2 undergoes a conformational change when GTP is hydrolysed resulting in affinity for the A site, rotating the small ribosomal subunit as it moves towards the A site and moving the mRNA with it through the ribosome channel (Sengupta et al., 2008). The A site is then free to be bound by another aminoacyl-tRNA-GTP-eEF1 for the next round of elongation.

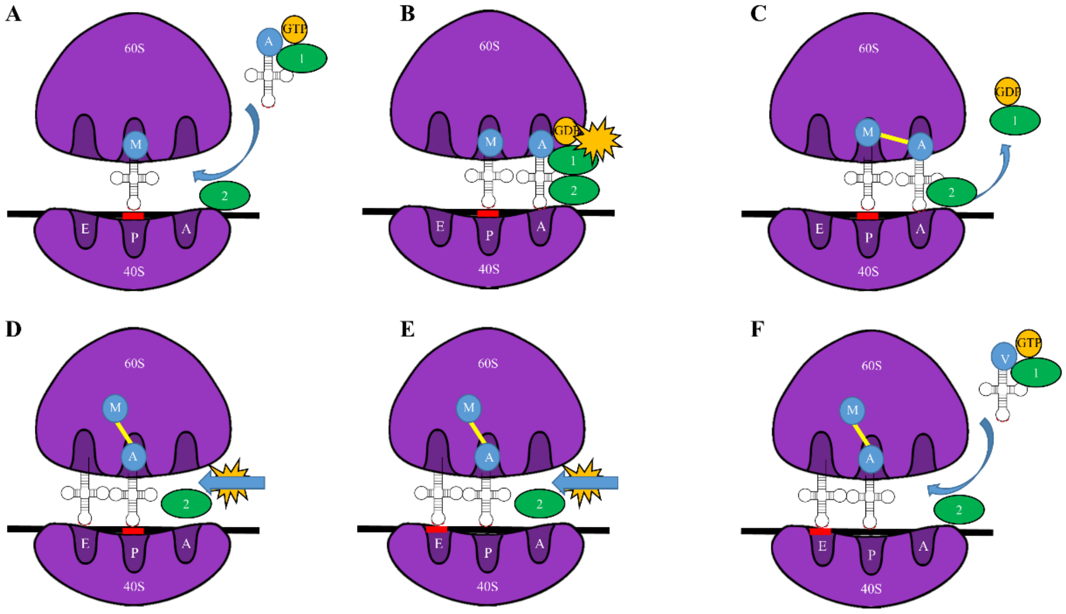

Figure 5.2 Translation Elongation. (A) Once the ribosome is assembled and a methylacyl-trna is present in the P site, aminoacyl-tRNA-GTP-eEF1 is recruited to the A site. (B) GTP bound to eEF1 is hydrolysed. (C) eEF1-GDP leaves the aminoacyl-tRNA enabling a peptide bond to form between the amino acid in the P and A sites. (D) Now deacetylated, the tRNA in the P site shifts to the E site and eEF2 shifts to the A site. (E) The shift results in translocating the tRNA from the A site into the P site and shifting the mRNA through the ribosome by 1 codon. (F) The A site is then free for this process to be repeated until a stop codon is present in the P site. 'eEF' omitted from elongation factors (green) for clarity. Adapted from (Watson et al., 2014).

Subsequent rounds of elongation occur until the stop codon (UAA, UAG or UGA) is reached at the A site. Stop codons are recognised and bound by the tRNA shaped release factor protein eukaryotic translational termination factor 1 (ETF1) (Dever and Green, 2012). ETF1 forms a complex with GTP-bound eukaryotic translational termination factor 3 (ETF3) which results in release of the active domain of ETF1 into the P site, catalysed by GTP hydrolysis. The active domain of ETF1 results in hydrolysis of the ester bond linking the tRNA in the P site to the elongating peptide chain (Dever and Green, 2012). The release factor bound stop codon blocks further tRNAs entering the A site and therefore no aminoacyl-tRNA can accept a peptide bond and the peptide chain is released (Watson et al., 2014). Ribosomal subunits and tRNAs then dissociate from the mRNA and are recycled for use in further rounds of translation, a process coordinated by recycling factors (Dever and Green, 2012).

5.2.2 Translational Efficiency

A single mRNA can be translated multiple times through subsequent rounds of translation or even translated by multiple ribosomes at the same time before it is degraded, amplifying the number of proteins in a cell from the number of mRNA transcripts expressed (Rodnina, 2016). The rate at which an mRNA is translated into protein (referred to as translational efficiency) can be selectively regulated to control gene expression post-transcriptionally to adapt protein expression in response to environmental stimuli (Gobet and Naef, 2017). Regulation of translational rates contributes to a widespread lack of correlation between mRNA and protein levels when measured by omics techniques (Rodnina, 2016). Correlational coefficients of between 0.36 and 0.76 have been reported for mRNA:protein in bacteria, yeast and mouse samples (Maier et al., 2009). Lack of correlation between mRNA and protein levels has also been observed in CHO cells, such as a study which identified 158 differentially expressed proteins with no change in the mRNA level between fast and slow growing cells (Clarke et al., 2012). Another study which compared rates of translation efficiency (TE) to the levels of mRNA in CHO cells have shown a lack of correlation of mRNA and mRNA undergoing translation in 95% of genes (Courtes et al., 2013).

A number of mechanisms contribute to rates of translational efficiency at both the stages of translational initiation and elongation (Gobet and Naef, 2017). miRNA mediated translational suppression can block translation initiation through the trinucleotide repeat containing 6A (TNRC6A) protein which binds the Argonaut component of the RISC complex and is recruited to target mRNAs (Bridge et al., 2017). TNRC6A recruits the initiation inhibitor eIF4E2 to the eIF4E binding site and prevents eIF4E binding and therefore the pre-initiation complex cannot be assembled (Chen and Gao, 2017). Translation efficiency can also be influenced during the stage of elongation due to differences in decoding rates between codons and the incorporation rates of particular amino acids (Joazeiro, 2017). Different codons which encode the same amino acid are utilised at different frequencies in different organisms and codons which have a higher bias can often be translated at a faster rate (Chung et al., 2013). Codon specific elongation rates can be influenced by the availability of tRNAs in a cell and the upregulation of particular tRNAs can positively regulate the translation rates of mRNAs containing a corresponding codon (Gobet and Naef, 2017).

Control of translation efficiency can regulate protein expression in response to environmental stimuli and contribute to phenotypic changes such as during stress (Gobet and Naef, 2017). The mammalian target of rapamycin (mTOR) kinase is a regulator of translation which phosphorylates eIF4E binding proteins to allow eIF4E to bind initiation factors required to assemble eIF4F and the pre-initiation complex (Morita et al., 2015). mTOR activity can be induced in response to insulin or growth factors and regulates mitochondrial activity during the cell cycle by increasing the translational efficiency of components of the electron transport chain (Morita et al., 2015). The initiation factors *eIF2α* and *eIF3* are also involved in regulating the translation response (Chu et al., 2016; Lee et al., 2015). Phosphorylation of eIF2α is a mechanism which can be utilised by cells to decrease global translation in times of stress or amino acid deficiency (Chu et al., 2016). The pre-initiation complex component eIF3 can also directly bind specific mRNAs involved in processes such as the cell cycle and apoptosis which selectively increases or decreases translation of mRNAs based on stem-loop structures of its targets (Lee et al., 2015).

The impact of post-transcriptional control on gene expression means techniques for the study of active translation are required to bridge the gap in our knowledge between transcription and protein synthesis as profiling either mRNA or protein levels may not be sufficient to understand protein synthesis in CHO cell cultures on a genome-wide scale.

5.2.3 Polysome Profiling

Polysome profiling is a technique which is commonly used for the study of mRNA translation (Chassé et al., 2017). Highly translated mRNAs are bound by more than one ribosome at a given time in a complex known as a polysome while mRNAs with a low rate of translation are bound by a single ribosome (a monosome). Polysomes have a greater molecular mass than monosomes- a property which can be exploited to separate mRNAs with high and low translation rates. When centrifuged through a sucrose gradient, polysomes are pulled towards the bottom of the gradient with higher sucrose densities while monosomes remain near the top in low density sucrose (Arava et al., 2003). The gradient can then be separated into sucrose fractions which contain monosomes and polysomes respectively and represent mRNAs with low and high translation rates. mRNA can be released from the sucrose fractions and techniques such as qPCR and microarrays applied to quantify mRNAs in each fraction to identify transcripts with high or low levels of translation (Chassé et al., 2017; Vyas et al., 2009). Polysome profiling can be applied to identify increases or decreases in translation efficiency for thousands of genes (Sampath et al., 2008).

5.2.4 RiboSeq - Sequencing of ribosome-protected mRNA fragments

While polysome profiling can provide an overview of transcripts which are undergoing high and low levels of translation and enable the detection of differential translation efficiency, the technique can be enhanced via the application of next generation sequencing. Sequencing of ribosome-protected mRNA fragments (referred to as RiboSeq) is a technique which can also be applied to study rates of translation (Ingolia et al., 2009). The RiboSeq protocol aims to capture fragments of mRNA which are protected by a ribosome and therefore undergoing active translation (a ribosome protected footprint (RPF)) at either the stages of ribosomal initiation, elongation or both (MGlincy and Ingolia, 2017).

5.2.4.1 Translation inhibitors

To arrest translation and capture ribosomal footprints, inhibitors must first be used. Several inhibitors have been used to arrest translation which enrich for different types of ribosomal footprints depending on the stage of translation which they inhibit (Figure 5.3) (Tzani et al., 2018). Treatment with antibiotics including lactimidomycin, harringtonine and puromycin results in the stalling of ribosomes during initiation. Lactimidomycin is large in size relative to other chemicals used for inhibiting translation and can only enter the ribosome during initiation when the E site (its binding site) is free from tRNAs (Lee et al., 2012). Presence of lactimidomycin in the E site therefore blocks translational elongation from occurring as tRNAs are blocked from shifting from the P site to the E site. Harringtonine prevents translation proceeding from the initiation phase to elongation by binding free 60S ribosomal subunits prior to assembly of the 80S ribosome. The exact mechanism of action for harringtoinines role in inhibiting translation is unknown however harringtonine is known to bind the A site of ribosomes and hypothesized to block formation of peptide bonds between the aminoacyl-tRNAs in the A site and P site (Gürel et al., 2009). Puromycin carries out its translational inhibitory effect by acetylating the 3' hydroxyl group on the end of a growing peptidyl-tRNA chain in the P site, resulting in the release of the peptide chain during the early stages of elongation and remaining bound to the ribosome (Clamer et al., 2017).

RPFs from ribosomes in the elongation phase of translation be captured using chemical inhibitors such as cycloheximide, emetine, chloramphenicol, anisomycin or by flash freezing samples to stall translation (Andreev et al., 2017; Tzani et al., 2018). Cycloheximide works in a similar fashion to lactimidomycin (binding to the E site of the 80S ribosomal subunit), however, cycloheximide is smaller in size and can enter the E site in the presence of deacylated-tRNA (which would be present during the elongation phases of translation) (Lee et al., 2012). Emetine also inhibits protein synthesis through preventing the translocation of the mRNA through the ribosome (Ozaki, 1998). Emetine irreversibly carries out this function through an incompletely understood mechanism which does not rely on binding to the E site (Dang et al., 2011; Graber et al., 2013). Anisomycin blocks ribosomes in an alternative rotated conformation which results in a small footprint of 20-22 nucleotides trapped in the ribosome during elongation (Wei and Qian, 2015) while chloramphenicol inhibits peptidyl transfer similar to harringtonine (Bougas et al., 2017; Michel and Baranov, 2013).

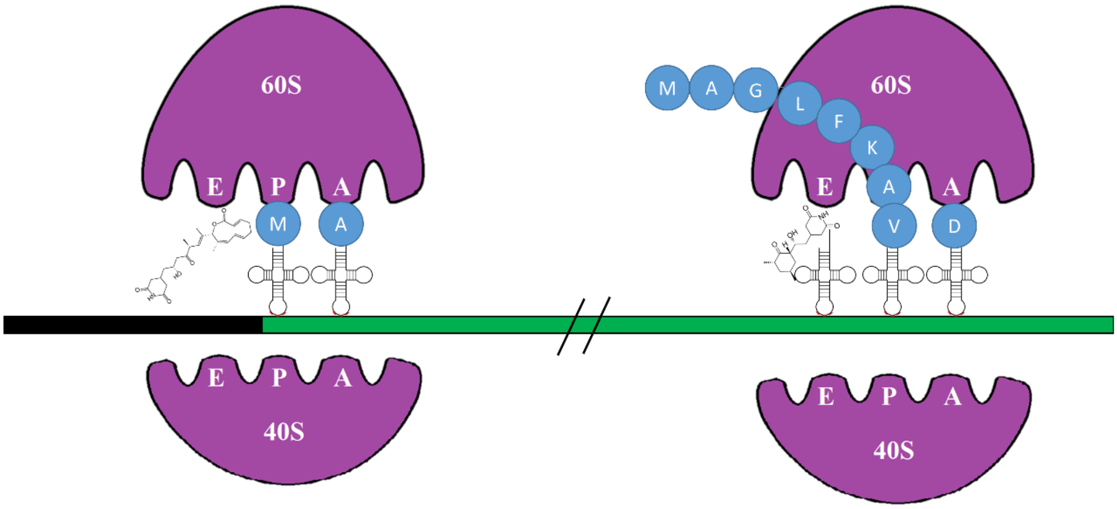

Figure 5.3 – Blocking of initiating and elongating ribosomes. Different inhibitors can be used to block translation in RiboSeq which enables libraries to be enriched for either initiating or elongating ribosomes. Both lactimidomycin and cycloheximide block translation by binding and blocking the ribosome E site. Lactimidomycin (left) is however a larger antibiotic than cycloheximide (right) which means lactimidomycin can only enter the ribosomal E site in the absence of a tRNA during initiation. Cycloheximide is smaller and can enter the E site during elongation with a tRNA present in the E site. Coding region of mRNA transcript depicted in green.

5.2.4.2 – Purification of RPFs

With translation stalled and mRNA fixed inside ribosomes, cells are lysed and subsequently treated with RNase I (MGlincy and Ingolia, 2017). Ribosomal protected fragments (footprints) typically between 28-31 nucleotides in length are protected from RNase degradation while unbound RNA is digested. Next an RNase inhibitor is added and then ribosomes are added to a sucrose cushion or size exclusion chromatography column to separate ribosomes from the lysate. Ribosomes are pelleted using ultracentrifugation and resuspended in TRIzol. During this step the footprints are released from the ribosome and a standard RNA-Seq protocol can be followed from this point to prepare sequencing libraries. Size selection of fragments of the appropriate size can be achieved by utilising electrophoresis.

5.2.5 The applications of RiboSeq data

The major advantage of RiboSeq data over other methods which analyse translation such as polysome profiling is the application of next generation sequencing technology provides the nucleotide sequence of each RPF captured inside a ribosome (Michel and Baranov, 2013). RPFs can be mapped to the genome or transcriptome of an organism to determine the exact location of a transcript which was inside a ribosome at the time of profiling. Polysome profiling does not provide any information about the locations within a transcript occupied by a ribosome and it is not possible to differentiate between active translation of open reading frame coding regions and, for example, ribosomes stalled in 5'UTR regions or non-protein coding transcripts. Failure to distinguish between mRNAs undergoing active translation and those bound to stalled ribosomes means transcripts bound but not being translated by ribosomes can be incorrectly interpreted as high translation efficiency in polysome profiling experiments. Examples of bound RNA which could be captured as a false positive in polysome fractions include transcripts with stalled translation such as ribosomes present in stress granules (Kimball et al., 2003) or ribosomes queued for translation (Yordanova et al., 2018).

A property of RiboSeq reads is that the majority of reads in high quality data should map to protein coding regions of transcripts (Calviello and Ohler, 2017). Libraries enriched for translational elongation will align to the entire coding region while those treated with an initiation blocker have alignments enriched to transcription initiation sites. Many algorithms have been developed to utilise this property to accurately identify the coding regions and UTRs of an mRNA transcript. Algorithms which utilise RiboSeq reads to identify transcription start sites include *SPECtre* (Chun et al., 2016), *ORF-Rater* (Fields et al., 2015) and *RiboHMM* (Raj et al., 2016). *ORF-Rater* requires input from both initiation and elongation RiboSeq libraries as well as untreated libraries in order to identify open reading frames. In addition to aligning to coding regions, RiboSeq reads also have a unique property in that they align with triplet periodicity to coding regions – the majority of reads align in the first open reading frame of the coding sequence when a transcript is undergoing active translation. Utilisation of these properties of RiboSeq have revealed unexpected findings of translation which could not have been observed using polysome profiling (Jackson and Standart, 2015). First – in addition to the canonical coding region of a transcript, the presence of upstream open reading frames (uORFs) which likely play a role in regulating the translation of the main open reading frame by queuing poised ribosomes for translation upstream of the transcription start site have been identified (Jackson and Standart, 2015). Second, alternative translation initiation sites have been identified on transcripts which can increase or decrease the length of a transcript (Calviello and Ohler, 2017). Finally, utilising the triplet periodicity property of aligned reads has enabled the identification of out of frame overlapping CDS regions in a transcript which encode dual coding transcripts (Malone et al., 2017). Other translational ambiguities can be identified from the analysis of RiboSeq data using software such as *RiboTools* (Legendre et al., 2015). *RiboTools* can be used to identify unexpected levels of ribosomal occupancy such as altered overall levels of a given amino acid or codon enriched in RPFs between sample conditions which may indicate ribosomal stalling. *Ribotools* can also identify stop codon read-through if alignment of RPFs to a gene extends past the stop codon which result in transcripts encoding proteins with C-terminal extensions (Loughran et al., 2014).

Application of sequencing technology as opposed to microarray or qPCR for footprint quantification also provides the same benefits RNA-Seq has over these techniques for analysis of the transcriptome to the study of the translatome. The throughput of sequencing technology enables quantification of ribosomal occupancy in all transcripts in a sample as opposed to the lower throughput of qPCR based polysome profiling which is used for targeted investigation of a given gene (King and Gerber, 2016). In addition next generation sequencing is not dependent on pre-defined probes or primers and therefore it is possible to study translation of novel transcripts. Finally, a more accurate and informative calculation of translational efficiency of transcripts is possible by pairing RiboSeq data with matched control RNA-Seq to identify differences in expected levels of translation given observed expression of mRNAs for a given gene (Brar and Weissman, 2015; Xiao et al., 2016). A number of algorithms have been developed for identifying differential translation including *Anota* (Larsson et al., 2010), *Babel* (Olshen et al., 2013) and *Xtail* (Xiao et al., 2016).

5.2.6 Previous utilisation of polysome profiling and RiboSeq to study CHO cell biology

5.2.6.1 Application of Polysome profiling to CHO cells

Polysome profiling has been previously applied to the study of the CHO cell translatome to study growth (Courtes et al., 2013), the contribution of the mTOR pathway to translation (Courtes et al., 2014), codon bias (Ang et al., 2016) and recombinant mRNA loading (Godfrey et al., 2017). Courtes et al. 2013 applied polysome profiling followed by microarray analysis to study the translatome for the first time in CHO cells during different days of culture. 1,079 genes were identified with high translational efficiency of which 43 were associated with growth and likely contribute to the characteristics of CHO cell growth in the exponential phase. The authors of this study then went on to investigate the cause of changes to the translational efficiency of transcripts by studying the role of the mTOR pathway (Courtes et al., 2014). The mTOR pathway was induced in fed-batch cultures by increasing nutrient availability and inhibited using rapamycin treatment, revealing enhanced translational efficiency of recombinant monoclonal antibodies light and heavy chain in fed-batch conditions. The findings of Courtes et al. were also incorporated into the first CHO cell multi-omics study to utilise translatomics data to calculate codon biases in high and low translated transcripts with the aim of optimising the codon usage in interferon gamma to enhance recombinant protein production (Ang et al., 2016). Godfrey et al. used polysome profiling to investigate translation of IgG, finding increased overall levels of translational efficiency correlated with cell lines displaying the highest specific growth rate and demonstrated a link between high levels of heavy chain translation with high product titre (Godfrey et al., 2017).

5.2.6.2 Application of RiboSeq to CHO cells

Despite the promise of the technique to discover more about the production of recombinant proteins using RiboSeq, few studies have been carried out which have applied this technique to CHO cells. To date a single example has been reported in the literature highlighting the potential for the application of RiboSeq to CHO cells to inform rational engineering of cell lines to increase production of recombinant proteins (Kallehauge et al., 2017). In that study, samples of CHO cells were taken at two different culture time points representing exponential (day 3) and stationary growth phase (day 6) to observe differences in translation between cells growing rapidly and metabolising glucose with those that have undergone metabolic shift to consume lactate.

The use of RiboSeq in this study enabled the analysis of the efficiency of recombinant protein translation, finding it is translated as efficiently as host cell proteins. Although recombinant protein dominated the occupancy of ribosomes in both time points, the Neomycin resistance (*NeoR*) gene sequestered over 5% of the ribosomal capacity. NeoR was introduced as a resistance marker for selection of host cells which had been successfully transfected with the recombinant protein and is unnecessary for both cell survival/growth and recombinant protein production. After identifying unnecessary translation of this protein the authors used siRNA knock-down of NeoR resulting in increased cellular growth and product titres attributed to increased ribosomal capacity to translate mRNA for growth and protein production transcripts (Kallehauge et al., 2017). Successful application of RiboSeq in this study demonstrates how RiboSeq can uncover more information about the CHO cell translatome in varied culture conditions to inform rational cellular engineering.

In this chapter we apply RiboSeq to CHO cells undergoing temperature shift from 37°C to 31°C with samples generated in parallel to the RNA-Seq data utilised in Chapters 2, 3 and 4. The experimental design and aim of this experiment differs from that of Kallehauge et al. in that their study focused on differences in ribosomal occupancy between culture phases. The experiment presented in this chapter aims to identify differential translational efficiency of genes in temperature shifted CHO cells by utilising matched RNA-Seq controls. An additional aim is to identify post-transcription gene regulation such as alternative splicing or miRNA control which contributes to the CHO cell response to mild hypothermia utilising the sequencing data from the previous chapters which was generated in parallel.

5.3 Methods

5.3.1 Library Preparation & Sequencing

RNA for the preparation of RiboSeq and matched control RNA-Seq libraries were harvested from temperature shifted CHO cells cultured using the strategy described in Section 2.5. RNA for this RiboSeq experiment was taken in parallel to the RNA used for RNA-Seq experiments described in Section 2.5 .Harvested RNA samples taken from biological quadruplicates of temperature shifted and non-temperature shifted CHO cells were aliquoted into 2 separate fractions for the generation of RiboSeq and matched control RNA-Seq libraries. RiboSeq libraries were generated using the TruSeq Ribo Profile kit (Illumina Inc.) and RNA-Seq libraries generated using the TruSeq RNA Library Prep Kit v2 (Illumina Inc.). Ribosomal RNA was depleted prior to sequencing using the Ribo-Zero Gold rRNA Removal Kit (Illumina Inc.). Illumina HiSeq2500 instruments were configured to sequence libraries to a minimum depth of 30 million single end reads for 75 cycles. RiboSeq sample preparation was carried out by Dr Paul Kelly in Dublin City University.

5.3.2 Quality assessment & Preprocessing

FastQC (Andrews, 2010) was utilised for the analysis of sequence quality of both the RiboSeq and matched control RNA-Seq reads. Reads < 28 and >34 nucleotides in length were removed from the RiboSeq data (Diament and Tuller, 2016). A *Bowtie* (Langmead et al., 2009) index was created containing the CHO rRNA sequences in RefSeq. RiboSeq reads were aligned to this index with parameters configured for the alignment of short reads –seed length reduced to 23 (from the default of 28) and up to 2 mismatches permitted in the seed alignment (Ingolia et al., 2009; Lintner et al., 2017). Reads which could not be aligned (and were therefore non-ribosomal RNA) were written to a FASTA file by specifying the "--un" parameter. Matched RNA-Seq data for each RiboSeq sample was preprocessed in the same way as RiboSeq data. The only exception was the removal of the upper limit on the length filter processing step as mRNA reads should be above 34 nucleotides in length. Pre-processing of matched RNA-Seq data filtered out reads too short for accurate alignment and ribosomal contamination prior to further analysis.

5.3.3 Alignment of reads

Preprocessed reads from Section 5.3.2 were aligned to the CHO-K1 genome (Xu et al., 2011). The *STAR* '--genomeGenerate' mode was used to create a genome index utilizing the splice junctions generated in Chapter 3 from the RNA-Seq data using *Stringtie.* Splice junction overhangs were set to 33 nucleotides – one less than the max read length of the RiboSeq data to allow for optimal mapping. Reads were aligned to this index using *STAR* with parameters configured for the optimal mapping of small ~30bp reads. The anchor multimap reads parameter was increased to 200, the seed search length reduced to 15 and primary alignments for multimapped reads randomly assigned among multimappers with the highest alignment score. A second genome index was generated using a splice junction overhang parameter of 66 (one less than the length of the matched RNA-Seq data) and matched control RNA-Seq data was aligned without the alterations to the multi-mapping and seed search length parameters.

5.3.4 Qualitative analysis of RiboSeq read alignments

Qualitative analysis of RiboSeq read alignments can be used to further examine the quality of the dataset as correctly produced RiboSeq libraries should display characteristics such as triplet periodicity within open reading frame regions (Michel et al., 2016). The software *RiboTaper* (Calviello et al., 2016) was used to create plots for each read length summarizing the coverage of the first aligned nucleotide around start codons and stop codons for each gene. The presence of triplet periodicity was analysed in these plots to determine the lengths of alignments which display the characteristics of ribosome protected fragments based on the proportion of reads with the first nucleotide aligned in the first open reading frame.

5.3.5 Differential translation efficiency analysis

Once RPFs and RNA-Seq reads were mapped to the genome, the number of alignments for each protein coding gene was determined using *featureCounts* (Liao et al., 2014). Only primary read alignments were counted for genes based on the Ensembl gene identifier. Read counts for genes with >1 CPM generated in the previous section were used as input to *DESeq2* using the experimental design '~ protocol + condition + protocol:condition' where protocol represented the sequencing type (RiboSeq or control RNA-Seq) and condition the experimental condition (37°C or 31°C). The *DESeq2* algorithm (Love et al., 2014) was utilised to identify genes with a ≥±1.5 fold change in translational efficiency and ≤ 0.05 BH adjusted P value. The Bioconductor R package *xtail* (Xiao et al., 2016) was also applied to identify genes with a ≥±1.5 fold change in translational efficiency and ≤ 0.05 BH adjusted P value to confirm *DESeq2* results.

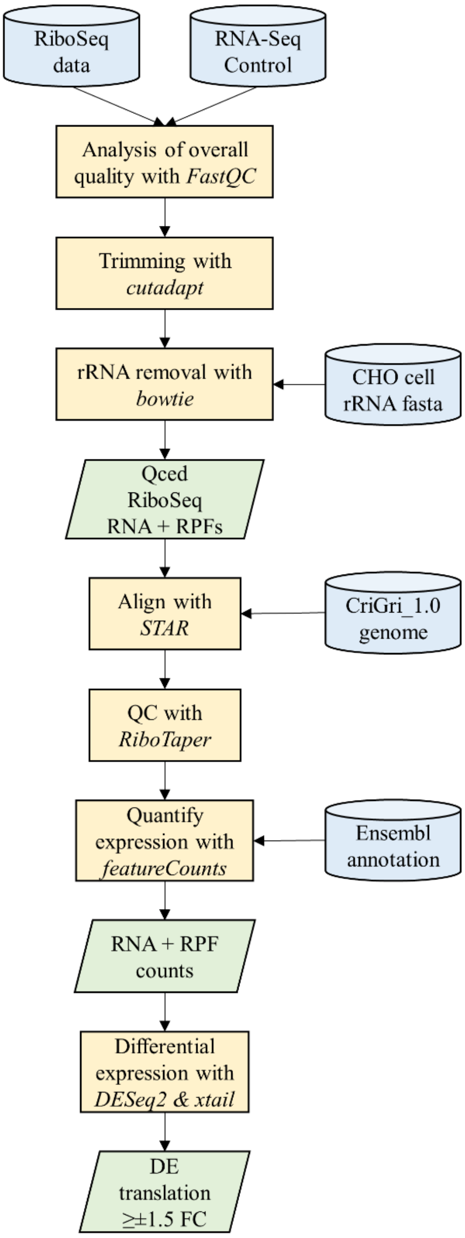

Figure 5.4 – Overview of RiboSeq analysis pipeline. RiboSeq and matched RNA-Seq control reads were preprocessed by removing sequence adapters and applying a read length and rRNA filters. Reads were aligned using *STAR* to the CHO-K1 genome and quantified against annotated gene features. Triplet periodicity of aligned reads was examined using *RiboTaper* and *DESeq2* and *xtail* were used for identifying differential translational efficiency.

5.4 Results

5.4.1 Quality Assessment and Alignment of RiboSeq and Control RNA-Seq data

RiboSeq and matched control RNA-Seq data reads were pre-processed to remove any potential contamination of non-ribosomal protected reads and select only those within the expected size range for RPFs. Application of a length filter and removal of reads mapping to rRNA resulted in the elimination of an average of 57.5% (48.33-71.1%) (Figure 5.6) and 18.61% (7.25-28.51%) of the RiboSeq (Figure 5.7) and matched control RNA-Seq data respectively. Following pre-processing an average of 29.3 million (17.8-35.8 x 10^6) RiboSeq and 46.3 million (33.1-55.8 x 10^6) RNA-Seq reads remained- 42.45% (28.9-51.67%) and 81.39% (71.5-92.8%) of the total sequenced reads respectively (See Appendix 5A for complete QC analysis).

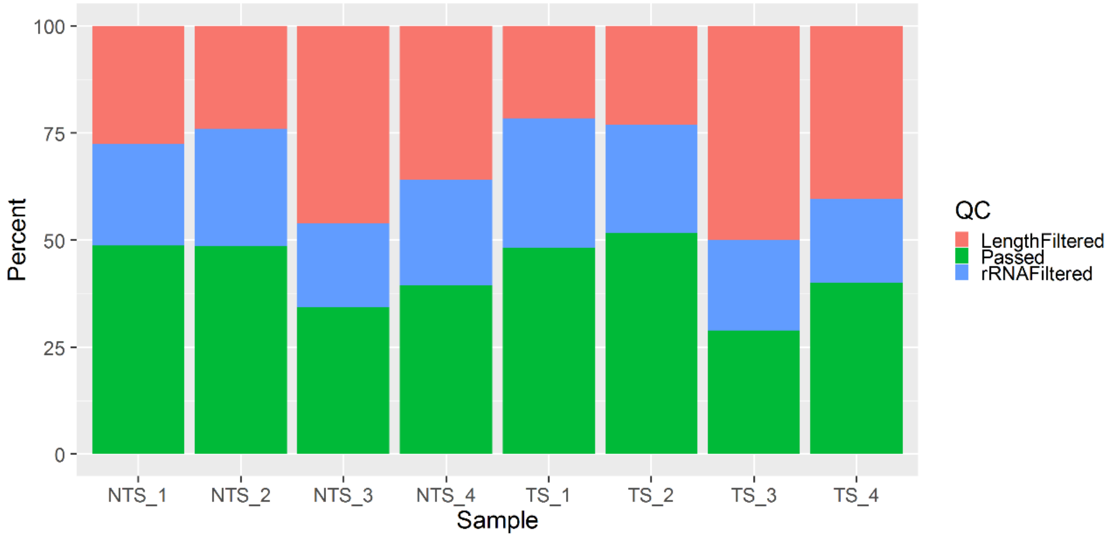

Figure 5.5 – Pre-processing of RiboSeq data. A length filter and alignment to rRNA to remove non-ribosomal protected fragments was applied to all samples raw sequence reads. An average of 42% (28.9-51.67%) of the input reads were retained after QC.

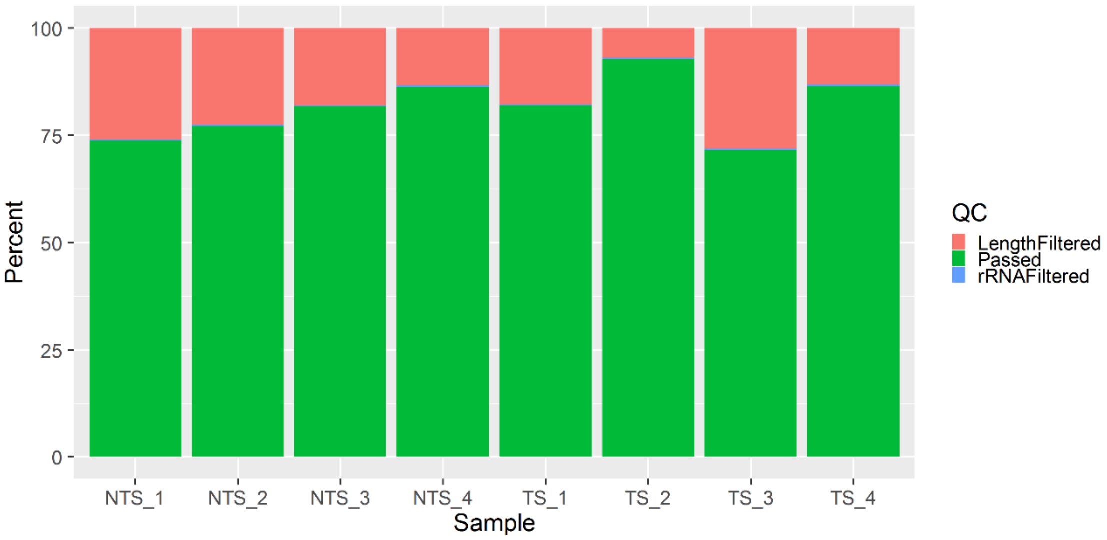

Figure 5.6 – Pre-processing of RNA-Seq matched control data. Raw reads were pre-processed by applying a length filter and aligning to rRNA to remove contamination and ensure consistency of pre-processing between the RiboSeq and matched RNA-Seq control data. An average of 81.4% (71.5-92.8%) of input reads survived these stages of pre-processing.

After initial pre-processing, both RiboSeq and matched RNA-Seq reads were aligned to the CHO cell genome using annotation generated in Chapter 3. Secondary read alignments of multi-mapping alignments were discarded resulting in an average of 22.56 million (13.2-28.8 million) RiboSeq (76.86% alignment rate) and 40.7 million (29.18-50.14 million) RNA-Seq reads (87.85% alignment rate) aligned. The average alignment length of RiboSeq reads was in the range of the expected ribosomal protected fragment size – 29.3nt. *featureCounts* was used to assign aligned reads to annotated protein coding genes resulting in an average of 14.49 million (7.5-20.3 million) RiboSeq reads assigned to genes (48.83% of the total, 63.46% of the aligned reads – Figure 5.7). 22.43 million RNA-Seq reads were assigned to protein coding genes (48.49% of the total, 55.2% of aligned – Figure 5.8). Full read alignment statistics in Appendix 5A.

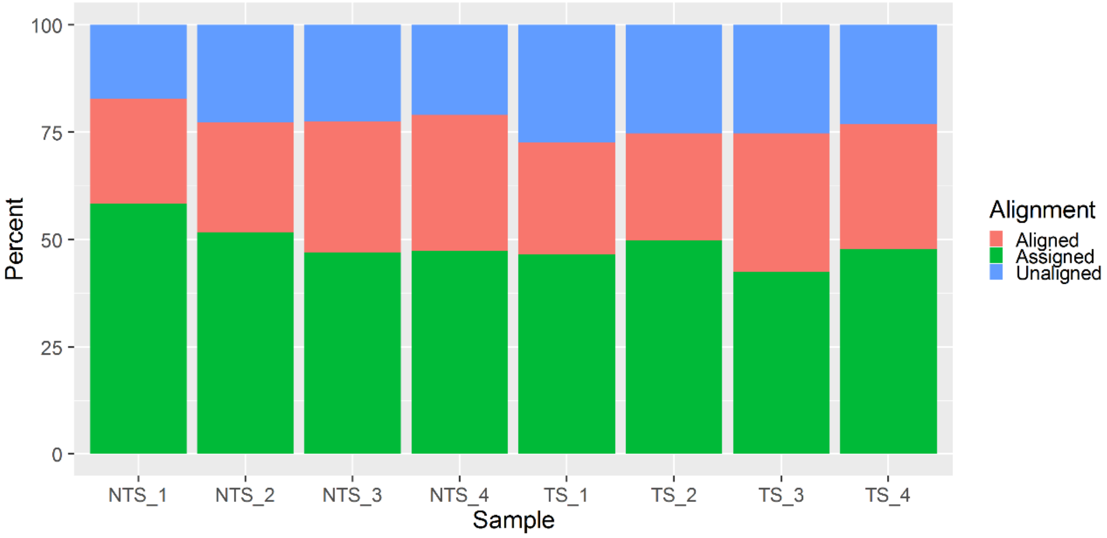

Figure 5.7 – Alignment of RiboSeq data. RiboSeq reads were aligned to the CHO genome using *STAR* with parameters configured to allow up to 200 seed search anchors and to reduce the seed search size to 15 (to allow for short read alignment). 75.57% (73.7-81.9%) of reads were successfully aligned and 43.16% (37.68-51.54%) assigned to protein coding transcripts.

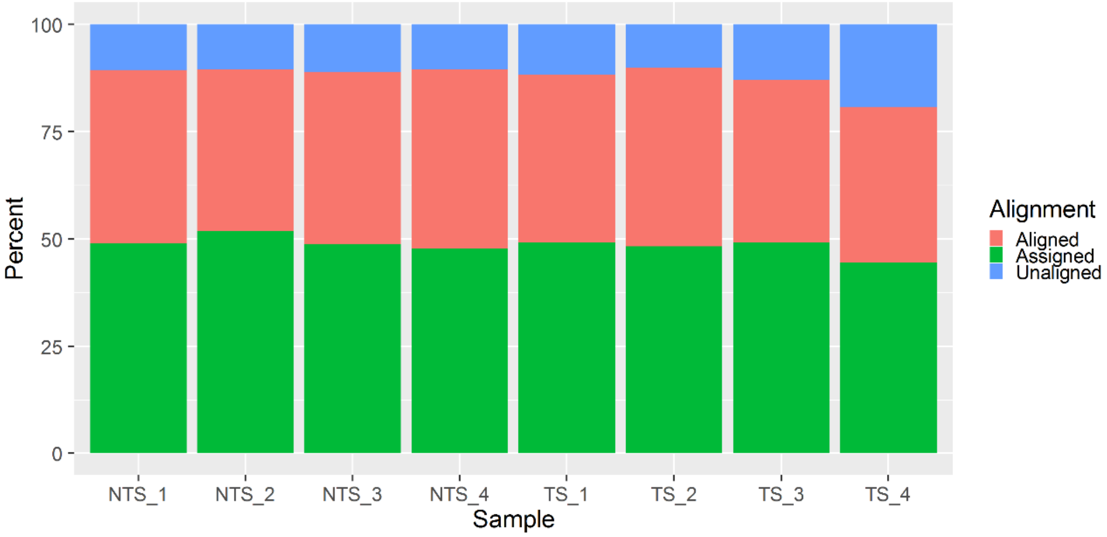

Figure 5.8 – Alignment of control RNA-Seq data. RNA-Seq reads were aligned to the CHO genome using *STAR*. 85% (78.83-88%) of reads were successfully aligned and 45.94% (42.4-48.88%) assigned to protein coding transcripts.

The quality of RiboSeq libraries was analysed by examining the triplet periodicity of read alignments in the CDS regions of annotated transcripts. Triplet periodicity was visible in kmers of length 28-31 (Figure 5.9). P value offsets were also calculated using the distance of read alignments upstream from start codons and identified as the expected 12 nucleotides in kmers of length 28-31 and 13nt for 32-34. The drop-off in coverage upstream of -12 and downstream of the stop codon at -18 indicates the majority of RiboSeq reads align to CDS regions.

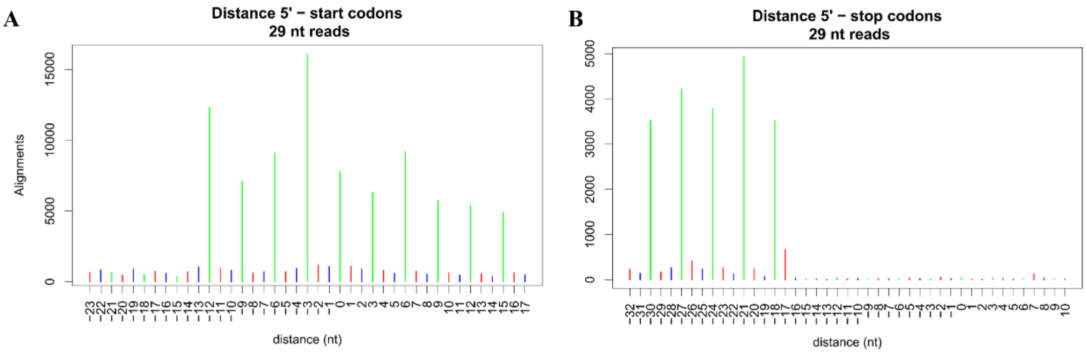

Figure 5.9 Triplet periodicity analysis in reads of length 29nt. RiboSeq libraries were of high quality with triplet periodicity visible in kmers of lengths 28-31. (A) Shown in green are reads down-sampled to 10% with the first aligned nucleotide in the first open reading frame indicating the majority of reads of length 29 are in the first frame. (B) The drop off in coverage before the stop codon indicates RiboSeq read alignments were enriched for coding sequences.

5.4.2 Differential translation efficiency with *DESeq2*

DESeq2 was applied to identify genes with a BH adjusted P value ≤0.05 and ≥ ±1.5 fold change in translational efficiency - revealing differential translation efficiency in 385 genes (Appendix 5B). The translation efficiency of 142 genes increased following a reduction in cell culture temperature while 243 decreased (Figure 5.10). The top 10 genes with increased and decreased translational efficiency induced by temperature shift are presented in Table 5.1.

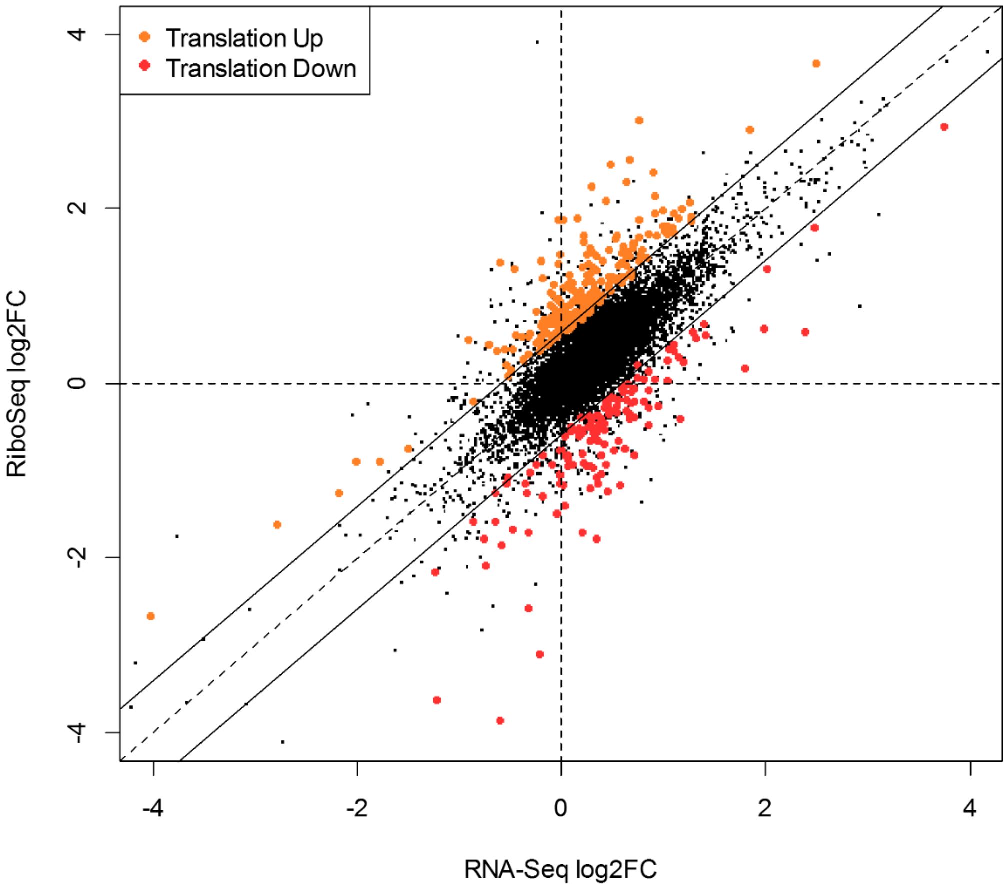

Figure 5.10 Transcriptome-wide translational efficiency. In this plot are the fold changes between levels of mRNA in RNA-Seq and ribosomal footprints using RiboSeq for the 11,434 genes which were detected at a CPM >1. Differential translational efficiency was identified in 385 genes with statistical significance (<0.05 BH P value, coloured red and orange). Solid lines represent a ±1.5 fold change in translational efficiency.

Table 5.1 – Differential translation efficiency identified with DESeq2. Presented in this table are the 10 genes with the most up and downregulated translational efficiency induced by temperature shift identified using *DESeq2*. Full *DESeq2* differential translation results in Appendix 5B.

Increased translation after TS				
Gene ID	**Gene Symbol**	**Gene Name**	**FC**	**BH Adj. P**
ENSCGRG00000005370	*Dchs1*	Dachsous Cadherin-Related 1	8.81	2.02×10^{-2}
ENSCGRG00000014053	*Nnat*	Neuronatin	6.99	3.77×10^{-3}
ENSCGRG00000008782	*Il25*	Interleukin 25	5.44	1.11×10^{-2}
ENSCGRG00000010865	*C130050O18Rik*	Riken cDNA	5.06	2.83×10^{-2}
ENSCGRG00000014447	*Apln*	Apelin	4.68	4.98×10^{-2}
ENSCGRG00000002033	*Hist1h2bl*	Histone Cluster 1 H2B Family Member L	3.58	1.35×10^{-22}
ENSCGRG00000019757	*Naa11*	N(Alpha)-Acetyltransferase 11, NatA Catalytic Subunit	3.34	3.34×10^{-3}
ENSCGRG00000014337	*Arid3b*	AT-Rich Interaction Domain 3B	3.29	2.91×10^{-2}
ENSCGRG00000001044	*Il31ra*	Interleukin 31 Receptor A	3.13	5.01×10^{-3}
ENSCGRG00000019323	*Hist1h1e*	Histone Cluster 1 H1 Family Member E	2.98	4.72×10^{-38}
Decreased translation after TS				
Gene ID	**Gene Symbol**	**Gene Name**	**FC**	**BH Adj. P**
ENSCGRG00000014262	*Ebf4*	Early B Cell Factor 4	-4.77	2.73×10^{-02}
ENSCGRG00000011397	*Wnt4*	Wnt Family Member 4	-4.28	2.52×10^{-02}
ENSCGRG00000017996			-4.25	1.86×10^{-02}
ENSCGRG00000002677	*Dhfr*	Dihydrofolate Reductase	-4.08	4.96×10^{-67}
ENSCGRG00000016690	*Ccdc125*	Coiled-Coil Domain Containing 125	-4.00	4.55×10^{-02}
ENSCGRG00000006451	*Vegfd*	Vascular Endothelial Growth Factor D	-3.85	1.47×10^{-02}
ENSCGRG00000005261			-3.62	6.35×10^{-04}
ENSCGRG00000016594	*Nr4a1*	Nuclear Receptor Subfamily 4 Group A Member 1	-3.47	1.40×10^{-03}
ENSCGRG00000019716	*Rplp1*	Ribosomal Protein Lateral Stalk Subunit P1	-3.36	1.03 × 10-^{32}
ENSCGRG00000001695	*Naa16*	N(Alpha)-Acetyltransferase 16, NatA Auxiliary Subunit	-3.20	4.95×10^{-08}

5.4.3 Differential translation efficiency identified using *xtail*

In addition to *DESeq2*, the *xtail* algorithm was utilised to identify differential translation efficiency induced by temperature shift. Differential translation efficiency was identified in 430 genes -160 upregulated and 270 downregulated. The top ten 10 genes with up and downregulated translation efficiency are presented in Table 5.2.

Table 5.2 -– Differential translation efficiency identified with *xtail*. Presented in this table are the 10 genes with the most up and downregulated translational efficiency induced by temperature shift identified by *xtail*. Full *xtail* differential translation efficiency results in Appendix 5C.

Increased translation after TS				
Gene ID	**Gene Symbol**	**Gene Name**	**FC**	**BH Adj. P**
ENSCGRG00000005370	*Dchs1*	Dachsous cadherin-related 1	7.59	3.55×10^{-2}
ENSCGRG00000014447	*Apln*	apelin	4.60	2.42×10^{-2}
ENSCGRG00000002033		Histone H2B type 1	3.57	4.87×10^{-36}
ENSCGRG00000019757	*Naa11*	N(alpha)-acetyltransferase 11	3.35	8.22×10^{-4}
ENSCGRG00000014337	*Arid3b*	AT rich interactive domain 3B	3.30	2.88×10^{-2}
ENSCGRG00000001438	*Tex33*	Testis expressed 33	3.22	$4.81 \times 10^{-}2$
ENSCGRG00000005554			3.10	1.05×10^{02}
ENSCGRG00000019323	*Hist1h1e*	Histone cluster 1, H1e	2.98	4.68×10^{-52}
ENSCGRG00000018540		Nucleoporin SEH1	2.95	9.04×10^{-3}
ENSCGRG00000012285	*Zscan22*	Zinc finger and SCAN domain containing 22	2.93	5.45×10^{-3}
Decreased translation after TS				
Gene ID	**Gene Symbol**	**Gene Name**	**FC**	**BH Adj. P**
ENSCGRG00000009639	*Cspg5*	Chondroitin sulfate proteoglycan 5	-4.19	3.55×10^{-2}
ENSCGRG00000002677	*Dhfr*	Dihydrofolate reductase	-4.08	1.53×10^{-97}
ENSCGRG00000006451	*Vegfd*	Vascular endothelial growth factor D	-3.91	1.86×10^{-3}
ENSCGRG00000005261			-3.65	9.47×10^{-4}
ENSCGRG00000016594	*Nr4a1*	Nuclear receptor subfamily 4, group A, member 1	-3.47	4.55×10^{-3}
ENSCGRG00000019716		60S acidic ribosomal protein P1	-3.37	2.10×10^{-30}
ENSCGRG00000009590	*Bmp6*	Bone morphogenetic protein 6	-3.25	5.71×10^{-3}
ENSCGRG00000001695	*Naa16*	N(alpha)-acetyltransferase 16, NatA auxiliary subunit	-3.20	2.00×10^{-4}
ENSCGRG00000004712		Sodium channel and clathrin linker 1	-3.10	7.85×10^{-3}
ENSCGRG00000007809	*Nlrp6*	NLR family, pyrin domain containing 6	-3.03	4.45×10^{-2}

5.4.4 Overlap of *DESeq2* and *xtail* differential translation efficiency results

Two approaches were utilised for identifying differential translational efficiency – *xtail* and *DESeq2*. Translational efficiency fold changes calculated by these two algorithms were highly correlated with an R^2 value of 0.998 and P value of 2.2 × 10-16 (Figure 5.11).

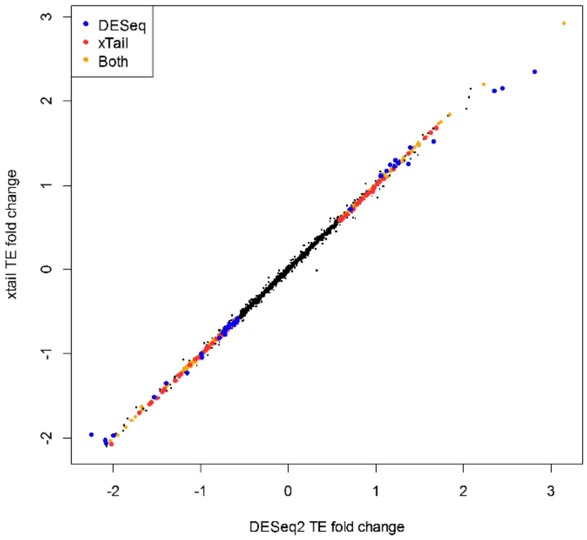

Figure 5.11 – Correlation of translational efficiency fold changes calculated using *DESeq2* and *xtail*. A Pearson's correlation test was carried out on the fold changes of *DESeq2* and *xtail*. Fold changes were consistent between the two algorithms with an (R^2 of 0.998 and a P value of 2.2 × 10^{-16}.

Despite high correlation of fold changes, there were differences in the number of genes identified with statistically significant differential translation efficiency. Differential translation efficiency identification with *DESeq2* revealed 385 genes while 430 were identified with *xtail*. A high degree of overlap was found in the genes identified as differentially translated using these two algorithms – with 351 genes identified using both algorithms (Figure 5.12).

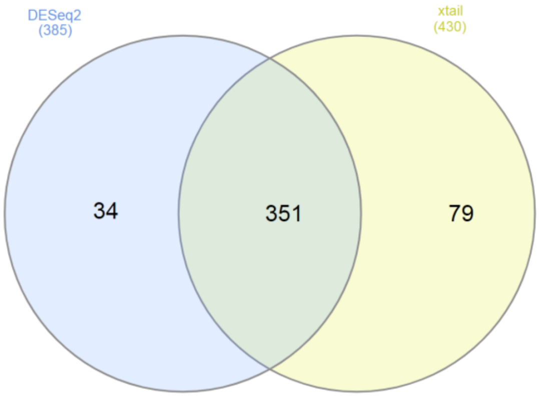

Figure 5.12 – Overlap of the genes identified with differential translation efficiency using *DESeq2* and *xtail*. 91.17% of the genes identified with differentially translated by *DESeq2* were also identified using *xtail*.

While the results of *DESeq2* and *xtail* were largely in accordance with each other – 34 genes were identified by *DESeq2* and 79 by *xtail* which were not identified with differential translation efficiency by the other algorithm. No discernible pattern could be obtained to explain why this is the case from fold changes alone (Figure 5.12) but for each of the cases where a gene was identified as differentially expressed in *xtail* but not *DESeq2* they were borderline cases and close to the thresholds for significance (Figure 5.13).

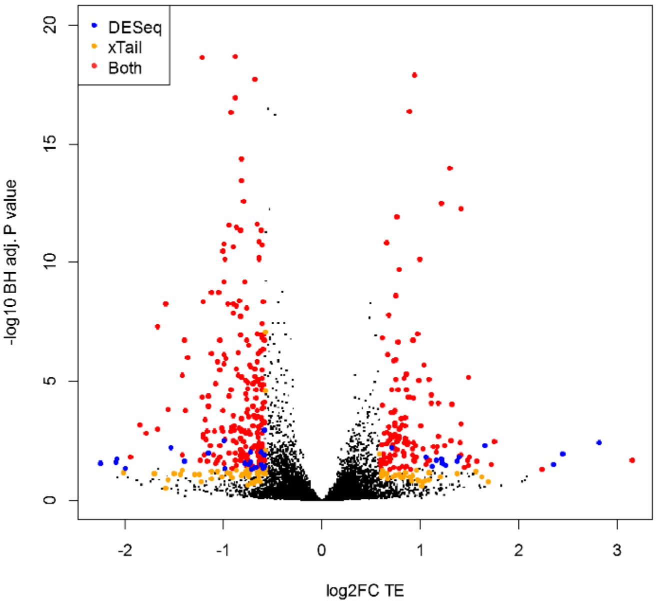

Figure 5.13 Volcano Plot of differentially translated genes. In this plot is the log2 fold changes of translational efficiency induced by temperature shift against their adjusted p values. In red are genes identified as differentially expressed by both *xtail* and *DESeq2* while blue and orange points are uniquely identified as significantly differentially expressed by *DESeq2* and *xtail* respectively.

In order to reduce false positives further analyses were restricted to the 351 genes identified with differential translation efficiency by both *DESeq2* and *xtail* (Appendix 5D). The top 10 genes identified by both algorithms are presented in Table 5.3.

Table 5.3 - The top 10 genes with up and downregulated translation efficiency identified by *DESeq2* and *xtail*. Differential translation efficiency was identified in 351 genes using both *DESeq2* and *xtail*. Full differential translation results in Appendix 5D.

GeneID	Gene Symbol	Gene Name	Fold Change	DESeq2 BH Adj. P	Fold Change	Xtail BH Adj. P
ENSCGRG00000005370	*Dchs1*	Dachsous Cadherin-Related 1	8.81	2.02×10^{-2}	7.59	3.55×10^{-2}
ENSCGRG00000014447	*Apln*	Apelin	4.68	4.98×10^{-2}	4.60	2.42×10^{-2}
ENSCGRG00000002033	*HIST1H2BL*	Histone H2B type 1	3.58	1.35×10^{-22}	3.57	4.87×10^{-36}
ENSCGRG00000019757	*Naa11*	N(Alpha)-Acetyltransferase 11, NatA Catalytic Subunit	3.34	3.34×10^{-3}	3.35	8.22×10^{-4}
ENSCGRG00000014337	*Arid3b*	AT-rich interaction domain 3b	3.29	2.91×10^{-2}	3.30	2.88×10^{-2}
ENSCGRG00000019323	*Hist1h1e*	Histone Cluster 1 H1 Family Member E	2.98	4.72×10^{-38}	2.98	4.68×10^{-52}
ENSCGRG00000012285	*Zscan22*	Zinc Finger And SCAN Domain Containing 22	2.95	2.15×10^{-2}	2.93	5.45×10^{-3}
ENSCGRG00000002164	*Tdg*	Thymine-DNA glycosylase	2.81	1.45×10^{-2}	2.81	1.33×10^{-3}
ENSCGRG00000008928	*Smpd2*	Sphingomyelin phosphodiesterase 2	2.80	3.25×10^{-2}	2.82	1.28×10^{-2}
ENSCGRG00000002900	*Art3*	ADP-Ribosyltransferase 3	2.77	1.89×10^{-2}	2.81	2.18×10^{-2}
ENSCGRG00000002677	*Dhfr*	Dihydrofolate reductase	-4.08	4.96×10^{-67}	-4.08	1.53×10^{-97}
ENSCGRG00000006451	*Vegfd*	Vascular Endothelial Growth Factor D	-3.85	1.47×10^{-2}	-3.91	1.86×10^{-3}
ENSCGRG00000005261			-3.62	6.35×10^{-4}	-3.65	9.47×10^{-4}
ENSCGRG00000016594	*Nr4a1*	Nuclear Receptor Subfamily 4 Group A	-3.47	1.40×10^{-3}	-3.47	4.55×10^{-3}
ENSCGRG00000019716	*Rplp1*	60S acidic ribosomal protein P1	-3.36	1.03×10^{-32}	-3.37	2.10×10^{-30}
ENSCGRG00000001695	*Naa16*	N(Alpha)-Acetyltransferase 16, NatA Auxiliary Subunit	-3.20	4.95×10^{-8}	-3.20	2.00×10^{-4}
ENSCGRG00000004712	*Sclt1*	Sodium Channel And Clathrin Linker 1	-3.18	9.34×10^{-4}	-3.10	7.85×10^{-3}
ENSCGRG00000009505			-3.02	5.10×10^{-9}	-3.02	6.40×10^{-5}
ENSCGRG00000000054			-2.96	1.40×10^{-4}	-2.96	7.94×10^{-8}
ENSCGRG00000016892	*Moxd1*	Monooxygenase DBH like 1	-2.69	1.22×10^{-2}	-2.70	9.97×10^{-4}

DAVID was used for gene ontology analysis on the 351 genes identified as differentially translated by both *DESeq2* and *xtail*. 19 GO categories were identified as enriched with a BH adjusted P value of < 0.05 (Table 5.4). Full *DAVID* results are presented in Appendix 5E.

Table 5.4 – Enriched gene ontologies in differentially translated genes. Presented in this table are the enriched GO categories in differentially translated genes sorted by their adjusted P value.

ID	Term	Count	%	BH adj. P
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	6	2	2.81×10^{-4}
GO:0034645	cellular macromolecule biosynthetic process	107	35	1.11×10^{-2}
GO:0071357	cellular response to type I interferon	9	3	1.48×10^{-2}
GO:0060337	type I interferon signaling pathway	9	3	1.48×10^{-2}
GO:0034340	response to type I interferon	9	3	1.64×10^{-2}
GO:0072331	signal transduction by p53 class mediator	12	4	2.44×10^{-2}
GO:0006355	regulation of transcription, DNA-templated	79	26	6.14×10^{-2}
GO:0051171	regulation of nitrogen compound metabolic process	92	31	5.96×10^{-2}
GO:1903506	regulation of nucleic acid-templated transcription	79	26	5.68×10^{-2}
GO:0006974	cellular response to DNA damage stimulus	27	9	5.54×10^{-2}
GO:0051252	regulation of RNA metabolic process	81	27	5.34×10^{-2}
GO:2001141	regulation of RNA biosynthetic process	79	26	4.95×10^{-2}
GO:0097659	nucleic acid-templated transcription	81	27	4.84×10^{-2}
GO:0033554	cellular response to stress	46	15	4.55×10^{-2}
GO:0007049	cell cycle	44	15	4.54×10^{-2}
GO:0006357	regulation of transcription from RNA polymerase II promoter	47	16	4.39×10^{-2}
GO:0006351	transcription, DNA-templated	78	26	4.14×10^{-2}
GO:0010468	regulation of gene expression	90	30	4.29×10^{-2}
GO:0019219	regulation of nucleobase-containing compound metabolic process	86	29	4.21×10^{-2}

The fold changes of genes with differential translation efficiency with roles in growth or mitochondrial activity are presented in Figure 5.14. Read alignments to the *Dhfr* and *Hist1h1e* genes which are among the most significantly differentially translated are presented in Figure 5.15 and 5.16.

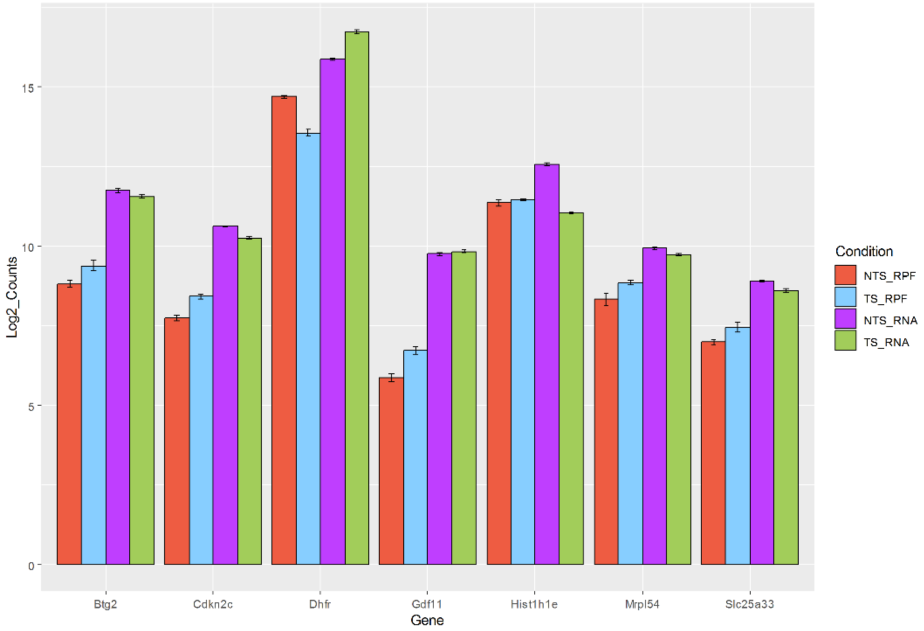

Figure 5.14 – Read counts of genes with differential translation efficiency. Presented in this plot are library size normalised log2 read counts for RiboSeq data and matched RNA-Seq controls. For each of these genes the RPF counts are significantly different in each condition while the RNA-Seq counts either do not change or are changed in the opposite direction.

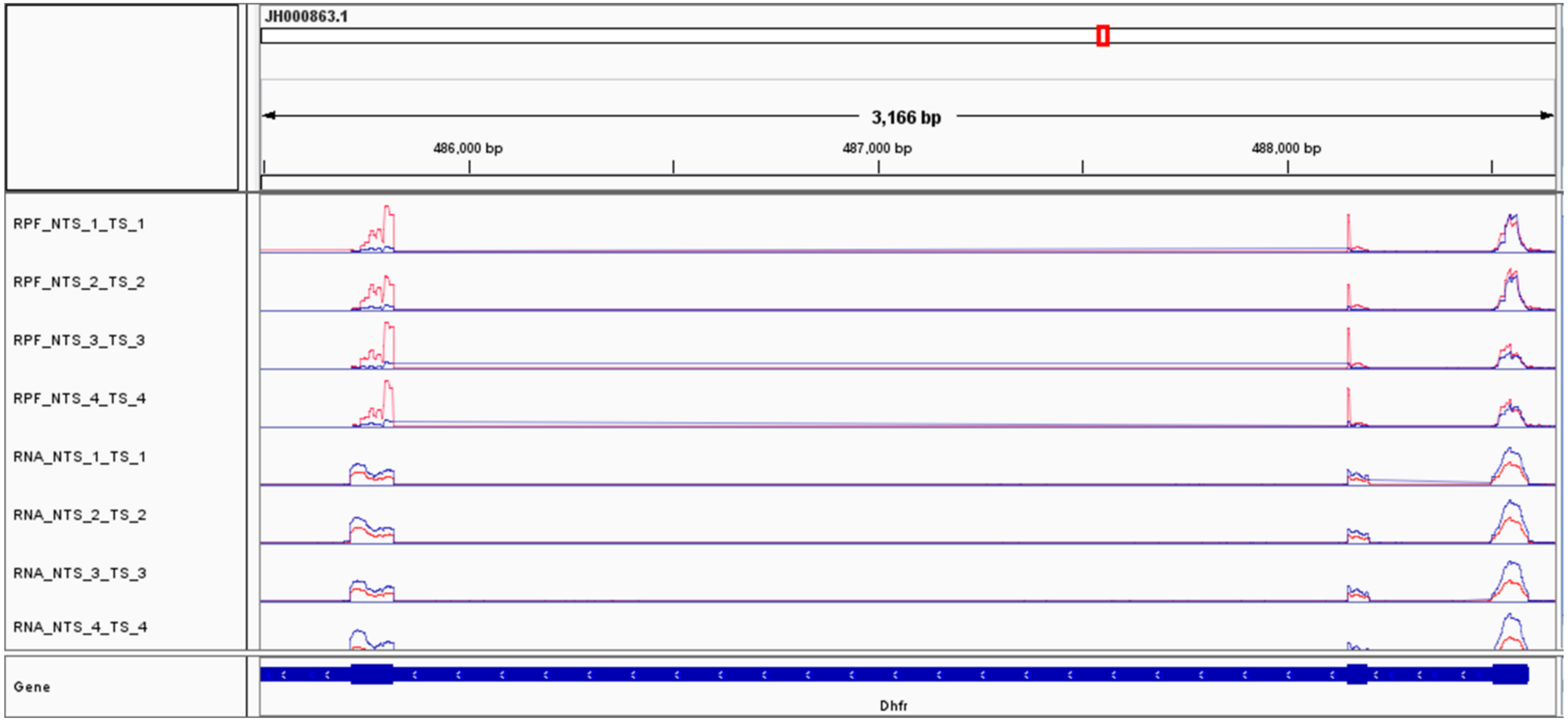

Figure 5.15 – Alignment of RiboSeq & matched control RNA-Seq to *Dhfr*. The *Dhfr* gene is the gene with the most significant downregulation of translation efficiency in temperature shift. In this plot tracks are overlaid to show the differences in read alignment between the NTS and TS conditions. Translation is decreased in temperature shifted cells (RPF tracks) however mRNA levels are increased (RNA tracks). TS samples shown in blue, NTS in red. Only exons 1-3 are depicted for clarity.

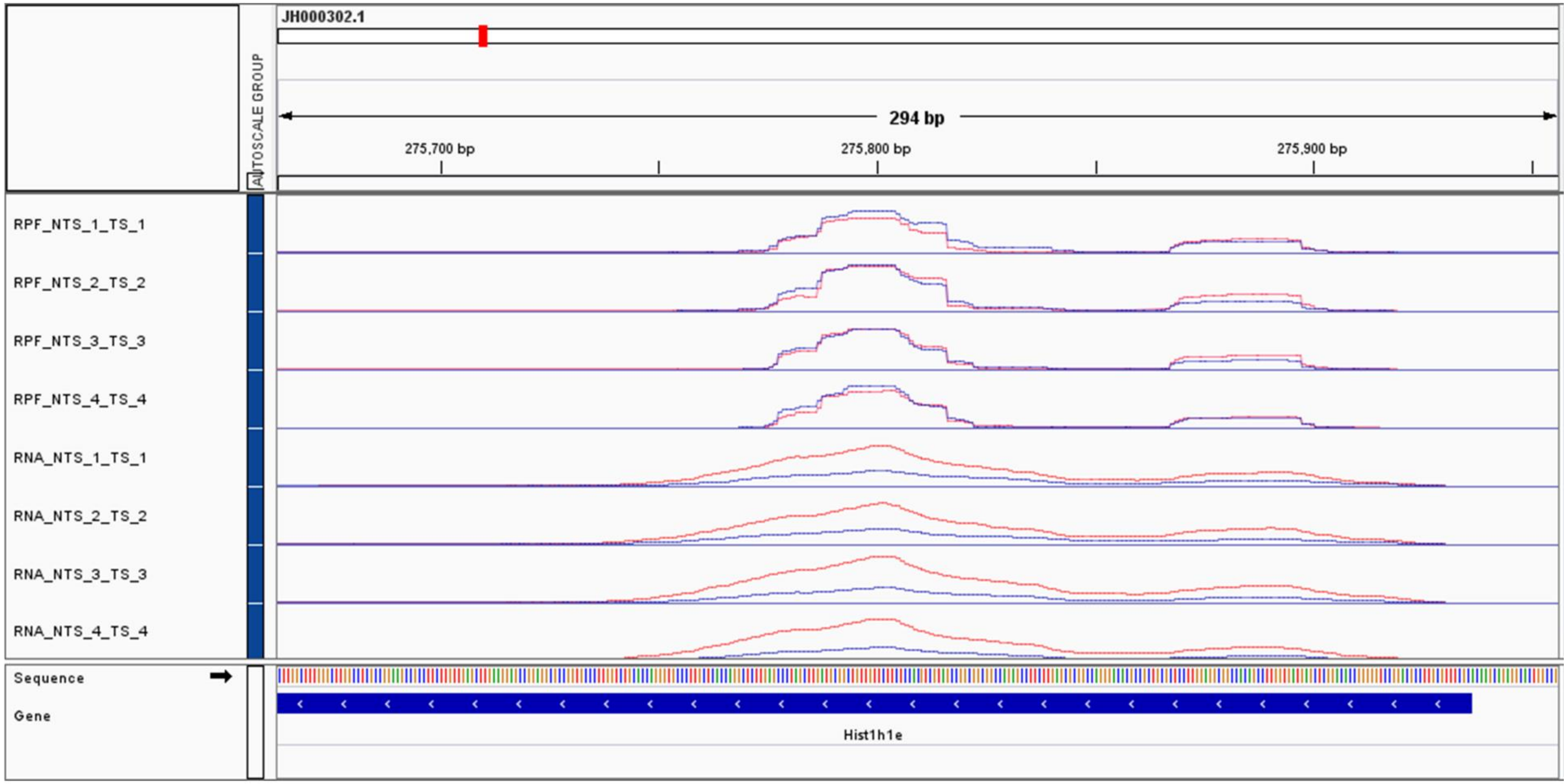

Figure 5.16 – Alignment of RiboSeq & matched control RNA-Seq to *Histh1e.* *Histh1e* is among the genes with the most significantly upregulated translation efficiency. In this single exon gene the mRNA levels are significantly increased in non-temperature shifted cells (RNA tracks) while translation is higher in temperature shifted cells (RPF tracks). TS samples shown in blue, NTS in red.

5.4.5 Comparison of differentially translated genes to those previously reported as differentially translated in rapidly proliferating cells

Genes identified as differentially translated using both DESeq2 and xtail were compared to a previous report which utilised polysome profiling to identify genes associated with rapid proliferation in CHO cells (Courtes et al., 2013). Of the differentially translated genes in Courtes et al. 8 were identified as differentially translated in this experiment (Table 5.5).

Table 5.5 - Differentially translated genes which were differentially translated in rapidly proliferating genes. The 657 differentially translated genes in (Courtes et al., 2013) were compared those differentially translated in temperature shifted CHO cells revealing the 8 genes in this table which were common between the datasets. Intensity and TE refer to the values observed in Courtes et al. while DESeq2 FC and DESeq2 P value are those observed in this experiment.

Gene ID	Gene Name	Intensity	TE	Courtes et al. result	DESeq2 FC	DESeq2 P Value	Temperature shift result
Dph3	Diphthamide Biosynthesis 3	2254	0.38	Translation down	2.46	0.0028	Translation up
E2f1	E2F Transcription Factor 1	1795	0.25	Translation down	1.8	2.15×10^{-5}	Translation up
Smarce1	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily E, Member 1	4793	1.57	Translation up	1.9	0.0065	Translation up
Spsb2	SplA/Ryanodine Receptor Domain And SOCS Box Containing 2	3735	1.67	Translation up	1.78	7.12×10^{-6}	Translation up
Ctdnep1	CTD Nuclear Envelope Phosphatase 1	2984	0.58	Translation down	-1.55	0.026	Translation down
Hivep1	Human Immunodeficiency Virus Type I Enhancer Binding Protein 1	217	0.34	Translation down	-1.7	0.012	Translation down
Rnf123	Ring Finger Protein 123	3017	1.57	Translation up	-1.52	0.018	Translation down
Tspan17	Tetraspanin 17	3288	0.48	Translation down	-1.87	1.36×10^{-8}	Translation down

5.4.6 Highly translated genes with stable translation unaffected by temperature shift

Genes with a fold change in translation efficiency between conditions less than ±0.1 were isolated revealing 3,653 genes with translation unaffected by temperature shift. Of these, 634 were highly translated with a translation efficiency log2 fold change within a given condition greater than 1 and 479 were poorly translated with a >-1log2FC (Appendix 5F). Annotated genes with a translation efficiency over 10 fold in a condition with a less than ±0.1 fold change between conditions are presented in Table 5.6.

Table 5.6 – Genes with over 10 fold translation efficiency unaffected by temperature shift. Genes with a less than 0.1 fold change in translation efficiency and a greater than 10 fold translation efficiency within a condition are presented in this table.

Gene ID	DESeq2 BaseMean	Gene Symbol	Gene Name	NTS TE FC	TS TE FC	FC	P adj
ENSCGRG00000018026	375	Myadm	Myeloid associated differentiation marker	46.37	43.36	-1.07	0.84
ENSCGRG00000016583	1550	Hmgn5	High mobility group nucleosome binding domain 5	38.08	35.57	-1.07	0.87
ENSCGRG00000012106	2879	Fus	FUS RNA binding protein	37.43	36.11	-1.04	0.88
ENSCGRG00000019525	55	Dda1	DET1 and DDB1 associated 1	23.17	22.1	-1.05	0.89
ENSCGRG00000020028	2912	Hist1h2bg	Histone H2B type 1	21.77	22.35	1.03	0.89
ENSCGRG00000015025	156	Tfap2a	Transcription factor AP-2, alpha	20.9	20.73	-1.01	0.98
ENSCGRG00000011089	944	Sf3a2	Splicing factor 3a subunit 2	16.29	15.26	-1.07	0.75
ENSCGRG00000012583	220	Ndrg1	N-myc downstream regulated 1	14.38	14.33	-1	0.99
ENSCGRG00000002209	650	Acsf2	Acyl-CoA synthetase family member 2	13.34	13.28	-1	0.99
ENSCGRG00000012849	79	Creb5	CAMP responsive element binding protein 5	13	13.9	1.07	0.93
ENSCGRG00000010208	283	Prrx1	Paired related homeobox 1	11.37	10.81	-1.05	0.84
ENSCGRG00000015567	457	Crtc1	CREB regulated transcription coactivator 1	10.84	11.18	1.03	0.9
ENSCGRG00000008147	3971	Lpl	Lipoprotein lipase	10.5	10.48	-1	0.99
ENSCGRG00000002549	434	Terf2	Telomeric repeat binding factor 2	10.43	9.75	-1.07	0.77

5.4.7 Differential translation efficiency in alternatively spliced genes

Of the 351 genes with differential translation efficiency, 99 were identified as alternatively spliced in Chapter 3 (Appendix G). Gene ontology analysis revealed 50 enriched GO terms among these genes including the G1/S phase transition checkpoint, regulation of cell cycle arrest and cellular response to stress (Table 5.7, Appendix H).

Table 5.7 – GO terms enriched in alternatively spliced and differentially translated genes In Chapter 3 temperature shift induced alternative splicing was identified in 1,274 genes. Differential translation efficiency was observed in 99 of these genes. *DAVID* was used for gene ontology analysis on these 99 genes revealing the GO terms presented in this table as the top 10 enriched biological processes.

Term	**Count**	**Fold Enrichment**	**BH. Adj. P**
GO:0000077~DNA_damage_checkpoint	9	12.98	8.24×10^{-4}
GO:0031570~DNA_integrity_checkpoint	9	12.16	6.78×10^{-4}
GO:0006974~cellular_response_to_DNA_damage_stimulus	17	4.50	4.63×10^{-4}
GO:0044773~mitotic_DNA_damage_checkpoint	7	14.94	3.45×10^{-3}
GO:0044774~mitotic_DNA_integrity_checkpoint	7	13.96	4.08×10^{-3}
GO:0000075~cell_cycle_checkpoint	9	8.39	3.58×10^{-3}
GO:0071156~regulation_of_cell_cycle_arrest	7	13.84	3.07×10^{-3}
GO:0072413~signal_transduction_involved_in_mitotic_cell_cycle_checkpoint	6	20.33	2.69×10^{-3}
GO:1902402~signal_transduction_involved_in_mitotic_DNA_damage_checkpoint	6	20.33	2.69×10^{-3}
GO:1902400~intracellular_signal_transduction_involved_in_G1_DNA_damage_checkpoint	6	20.33	2.69×10^{-3}

Alternatively spliced genes with differential translation efficiency include the BTG Anti-proliferation Factor 2 gene *Btg2* which has role in regulation of the cell cycle (GO :0000075). The translation efficiency of the *Btg2* gene was upregulated 1.66 fold with a BH adjusted P value of 1.59×10^{-3} (Figure 5.17).

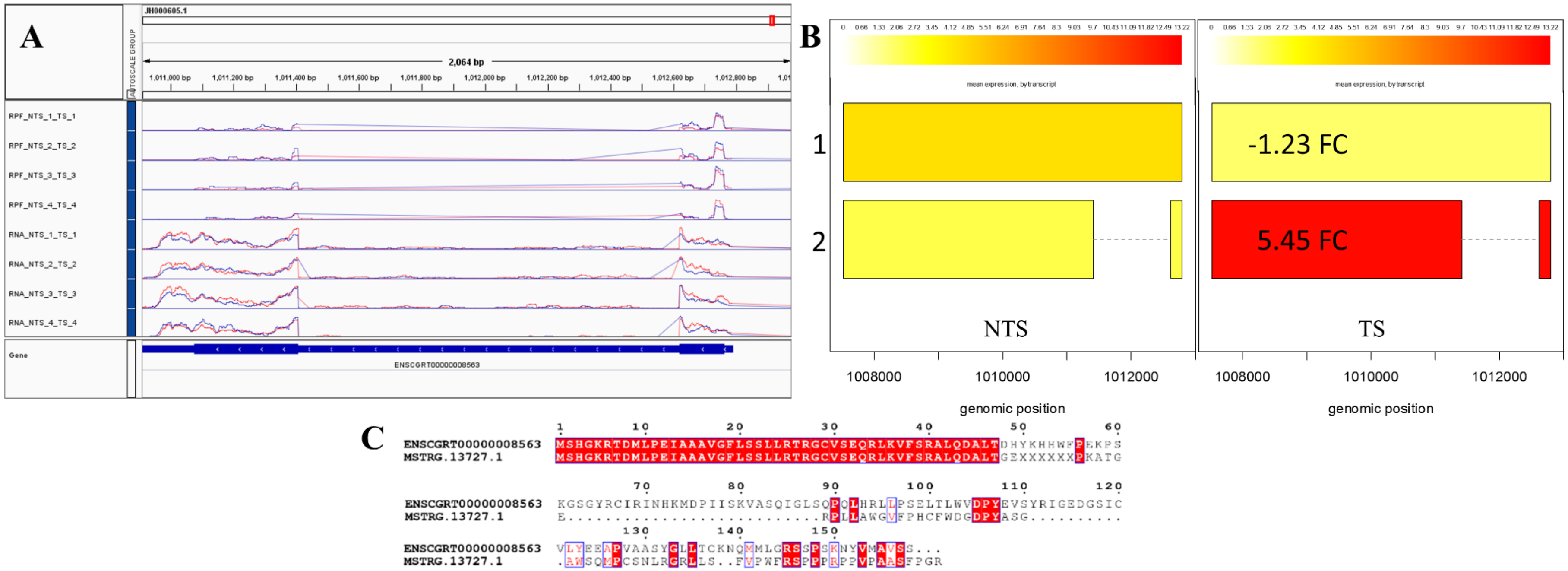

Figure 5.17 Alternative splicing induced differential translation efficiency of *Btg2*. (A) The translation efficiency of the *Btg2* gene was upregulated 1.66 fold in temperature shifted cells (higher coverage in RPF tracks). NTS samples represented in red, TS in blue. (B) *Btg2* was identified as alternatively spliced, with decreased intron retention in TS samples. (C) The intron retention event alters the open reading frame of the transcript, resulting in a truncated and altered amino acid sequence (The canonical sequence is labelled with the transcript ID ENSCGR0000008563).

5.4.8 miRNA targeting of genes with differential translation efficiency

The process described in Section 4.3.5 to identify miRNA targets was repeated on genes with differential translation efficiency revealing 254 miRNA target predictions of differentially expressed miRNAs on 136 genes with differential translation efficiency (Appendix 5I). Cases of miRNAs differentially expressed in the opposite direction to the translation efficiency of their predicted target gene were found in 88 genes (Figure 5.18). 197 genes were identified with differential translation efficiency which were not a target of miRNAs or identified as alternatively spliced in Chapter 3 (Appendix J).

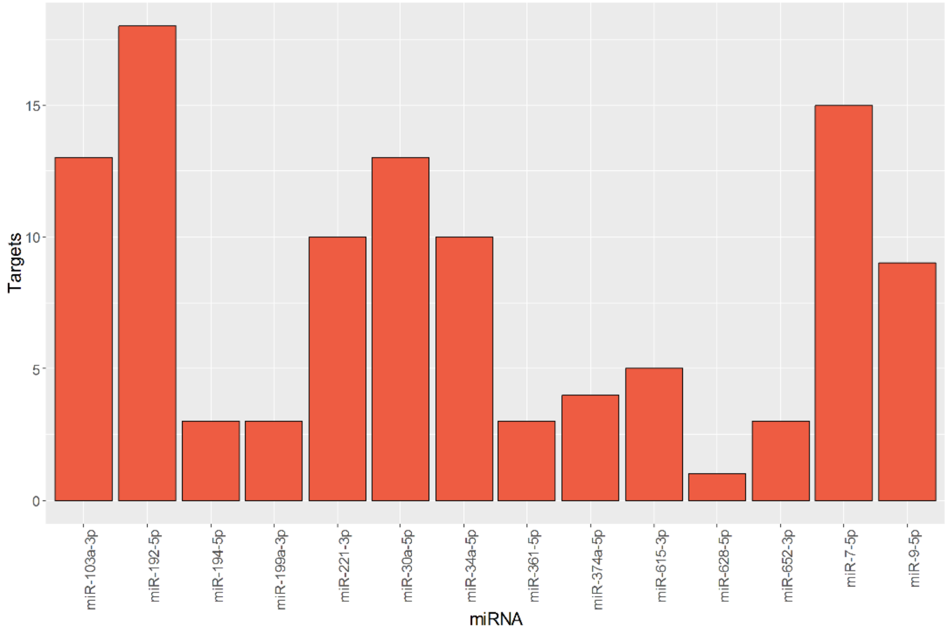

Figure 5.18 – Differential miRNA expression may be responsible for the differential translation efficiency of 25% of genes. miRNA targets were predicted on 88 genes with differential translation efficiency in the opposite direction to the differential expression of the miRNA. miR-192-5p was found to target the most genes, with 18 predicted targets downregulated.

miR-192-5p (1.91 FC) was found to target the most genes (18) with differential expression in the opposite direction – including dihydropholate reductase (*Dhfr)* which was the gene with the largest decrease in translational efficiency (-4.08 FC). Differentially translated genes with roles in translation, the cell cycle and apoptosis targeted by miRNAs differentially expressed in the opposite direction are presented in Table 5.8.

Table 5.8 – Differentially expressed miRNAs predicted to target genes with anti-correlated differential translation efficiency. Anti-correlated miRNA expression was able to explain 88 of the 351 cases of differential translation efficiency identified. Presented in this table are genes identified with a role in translation, apoptosis or regulation of the cell cycle which had differential translation efficiency and a predicted anti-correlated miRNA target (Full table in Appendix 5I).

Gene ID	Gene Symbol	Gene Name	TE FC	BH adj. P	miRBase ID	miRNA	miRNA FC
ENSCGRG00000015542	*Gdf11*	Growth Differentiation Factor 11	1.69	3.22×10^{-3}	MIMAT0003283	miR-615-3p	-1.39
					MIMAT0000441	miR-9-5p	-1.56
ENSCGRG00000014709	*Ddit4*	DNA Damage Inducible Transcript 4	1.60	1.65×10^{-3}	MIMAT0000232	miR-199a-3p	-1.32
					MIMAT0000087	miR-30a-5p	-1.33
					MIMAT0003283	miR-615-3p	-1.39
ENSCGRG00000000784	*Eif2s2*	Eukaryotic Translation Initiation Factor 2 Subunit Beta	1.77	8.41×10^{-3}	MIMAT0000727	miR-374a-5p	-1.52
ENSCGRG00000012487	*Eif4g3*	Eukaryotic Translation Initiation Factor 4 Gamma 3	-1.77	6.10×10^{-6}	MIMAT0000278	miR-221-3p	1.40
ENSCGRG00000005818	*Apopt1*	Apoptogenic 1, Mitochondrial	-2.00	7.30×10^{-7}	MIMAT0000255	miR-34a-5p	3.33
ENSCGRG00000016368	*Sh3glb1*	SH3 Domain Containing GRB2 Like, Endophilin B1	-1.63	1.03×10^{-4}	MIMAT0000252	miR-7-5p	3.01
ENSCGRG00000013456	*Src*	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase	-1.58	1.66×10^{-4}	MIMAT0000255	miR-34a-5p	3.33

5.4.9 Differential translation efficiency in genes uncorrelated at the gene and protein levels

Of the 153 genes identified in Chapter 2 as uncorrelated at the gene and protein level, 3 were identified as differentially translated (Table 5.9). Each of these genes have protein fold changes in the opposite direction to what is expected given the gene fold changes even when translational efficiency is accounted for. In addition to these genes, the Ribosomal protein S12 (*Rps12)* gene was detected in proteomics (-1.13 FC, 0.18 P value), identified in gene expression data at a 1.09 FC (0.01 P value) and had downregulated translational efficiency (-1.8 FC, 4.08×10^{-22}).

Table 5.9 – Gene, translation efficiency and protein fold changes of genes identified as differentially translated and uncorrelated at the gene and protein levels. Presented in this table are gene level fold changes identified using *DESeq2* to analyse RNA-Seq data, translation efficiency calculated using *DESeq2* to analyse RiboSeq data and protein level fold changes calculated using an ANOVA test on LC-MS/MS data.

Gene ID	**Gene Symbol**	**Gene Name**	**Gene FC**	**BH Adj. P**	**TE FC**	**TE BH Adj. P**	**Protein FC**	**P Value**
ENSCGRG00000002677	*Dhfr*	Dihydrofolate reductase	-1.11	1.43×10^{-2}	-4.08	4.96×10^{-67}	1.36	3.09×10^{-5}
ENSCGRG00000013906	*Rplp0*	Ribosomal protein lateral stalk subunit P0	1.22	$1.09 \times 10^{-}8$	-1.84	2.17×10^{-19}	1.40	2.28×10^{-3}
ENSCGRG00000016980	*Eif3j*	Eukaryotic Translation Initiation Factor 3 Subunit J	1.15	5.49×10^{-5}	1.66	1.61×10^{-4}	-1.57	8.54×10^{-3}

5.5 Discussion

In this Chapter an exciting, revolutionary technique to monitor how cells control the production of proteins through regulation of translation was utilized to understand the discrepancies between gene and protein expression. While differences in translation were identified through this experiment, there were data analysis challenges which had to first be overcome. A major quality issue occurring during RiboSeq library generation is the contamination of samples by incomplete degradation of ribosomal RNA (Chung et al., 2015). Contamination levels were reduced before library preparation by performing size exclusion - excising bands on a gel corresponding to the 28-31nt size ranges on a ladder and by treating RNA libraries with Ribo-Zero Gold (Bazzini et al., 2014).

To remove contamination during bioinformatics analysis stringent filters were applied to remove potential non-RPF derived reads and reduce false positives (Diament and Tuller, 2016). Reads above 34nt and below 28nt in length were removed to preserve only sequences falling into the expected range for RPFs resulting in the loss of 38.9% of the data. In addition, a further 18.6% of RPF reads were removed by alignment to rRNA sequences. rRNA accounts of >90% of the RNA in a cell and RiboZero kits are not 100% effective and therefore it is expected for rRNA derived reads to be present at some level even after measures are taken to reduce contamination. A 20% level of rRNA is a typical level of contamination in RiboSeq data sets which have used RiboZero approaches to deplete rRNA (Chung et al., 2015). While a loss of 57.5% of sequencing reads due to removal of rRNA and length filtering represents a large portion of RiboSeq data, samples were sequenced to a sufficient depth that an average of 29.3 million high quality reads remained after pre-processing. RiboSeq experiments have previously been performed with considerably lower read depth, for example, 4.67 million non-quality controlled reads per sample were used in the first RiboSeq experiment (SRA accession SRP000637) (Ingolia et al., 2009). Consistent with the data obtained in this study, more recent experiments which have applied RiboSeq to mammalian cell lines had 30 million RPFs remaining after quality pre-processing (Calviello et al., 2016).

When aligned to the CHO-K1 genome, 56% of reads uniquely mapped which is considerably higher than the low levels (10-20%) of uniquely mapped reads often observed in RiboSeq experiments (Gonzalez et al., 2014; Wang et al., 2016). A further

19.3% of reads mapped to multiple locations in the genome which arise due to ambiguous alignment of short reads. A single primary alignment was randomly selected to distribute these reads evenly among potential genome loci. Reads were assigned to genes annotated in the Ensembl annotation resulting in 11,433 genes with sufficient read depth (over 1CPM) for differential translation analysis. The R Bioconductor packages *DESeq2* and *xtail* were selected to identify differential translation efficiency using RiboSeq and matched RNA-Seq control data as these algorithms have been shown to outperform other software in sensitivity and specificity when using simulated datasets (Oertlin et al., 2018). A Wald test was utilised using *DESeq2* (Love et al., 2014) while a joint probability matrix was used in *xtail* (Xiao et al., 2016) to identify differential translation efficiency. Despite differences in the statistical methodology used in these software packages differential translation efficiency results were highly correlated (99.98 R^2). To increase the stringency of differential translation efficiency results, only those genes called as such with statistical significance in both these algorithms were carried forward for further investigation.

Differential translation analysis revealed 351 genes with altered translation efficiency induced by temperature shift – 129 upregulated and 222 downregulated. Gene ontology enrichment analysis of differentially translated genes reveals that the most significantly enriched GO categories include “cell cycle” and “cellular response to stress”. Previous studies have investigated the temperature shift response at the level of the transcriptome and reported decreased global transcription (Du et al., 2015), and differential expression of energy metabolism, apoptosis and cell cycle genes (Bedoya-López et al., 2016). Differential translation efficiency results therefore show not only are cell cycle genes regulated at the level of transcription but also undergo differential translation induced by temperature shift.

Differentially translated genes which may contribute to the decreased growth rate observed in CHO cells cultured at a reduced temperature include *Cdk11b*, *Cdkn2c, Fbxw7*, *Gdf11*, and *Btg2*. Cyclin dependant kinase 11b (*Cdk11b,* -1.72 FC in TE*)* is expressed in the G2 phase of growth and regulates duplication of the centrosome during mitosis through recruitment of the proteins Centrosomal protein 192 (Cep192) and Serine/threonine-protein kinase (Plk4) to the centrosome (Franck et al., 2011). Decreased *Cdk11b* translation suggests less cells are undergoing mitosis and may contribute to the reduced growth rates observed in temperature shifted cells. For the cell cycle to progress past the G1 phase Ccnd1 must form a complex with the kinases Cdk4 or Cdk6 (Mende et al., 2015). Translation efficiency of the Cyclin dependent kinase inhibitor 2C (*Cdkn2c*), an inhibitor of Cdk4 and Cdk6 (Jalili et al., 2012), increased 1.98 fold in response to mild hypothermia. In addition to alterations to Cyclin D activity, differential translation may also impact the activity of Cyclin E – also required for the G1/S phase transition of the cell cycle (Hwang and Clurman, 2005). Expression of the F-box and WD repeat domain containing 7 (*Fbxw7*) gene results in ubiquitination and degradation of Cyclin E to prevent cell cycle progression (Ibusuki et al., 2011). *Fbxw7* translation efficiency increased 2.66 fold in temperature shifted cells which would result in increased Fbxw7 protein expression and degradation of Cyclin E, contributing to the halting of the cell cycle.

Suppression of miRNA mediated translation repression due to downregulation of miRNAs may also impact the growth rate of CHO cells cultured at a decreased temperature. The downregulated miRNAs miR-615-3p (-1.39 FC) and miR-9-5p (-1.56 FC) were predicted to target the Growth differentiation factor 11 (*Gdf11*) gene which had an increased translation efficiency of 1.69 in temperature shifted cells. Gdf11 is a negative regulator of the cell cycle that induces expression of the Cyclin D and E inhibitor *Cdkn1b* (Ray et al., 2009; Shi and Liu, 2011; Xu et al., 1999). The DNA damage inducible transcript 4 (*Ddit4*) gene had a 1.6 fold increase in translation efficiency and was predicted as a target of 3 downregulated miRNAs - miR-199a-3p (-1.32 FC), miR-30a-5p (-1.33) and miR-615-3p (-1.39 FC). Recent emerging evidence has identified the mechanism of p53 action to halt the cell cycle (Janic et al., 2018), revealing p53 driven cell arrest is induced through the *Ddit4* gene. Decreased expression of the miRNAs miR-199a-3p, miR-30a-5p, miR-615-3p and miR-9-5p may therefore contribute to the decreased growth rate induced by mild hypothermia through increased translation of their target genes with roles in the cell cycle such as *Gdf11* and *Ddit4*.

The BTG anti-proliferation factor 2 (*Btg2)* gene has a 1.66 fold change of translation efficiency and was identified with differential transcript usage in Chapter 3. A transcript that does not contain an intron retention event in this single intronic gene was found to be upregulated 5.46 fold in temperature shifted cells. Retention of this intron results in an insertion to the coding sequence of the transcript which causes a truncated protein with a premature stop codon. As the transcript containing the interrupted open reading frame is downregulated -1.2 fold it is expected that translation efficiency of *Btg2* would increase. *Btg2* inhibits proliferation in the G1 phase through downregulation of Cyclin E genes *(Ccne1* & *Ccne2*) - which were observed in Chapter 2 with -1.71 and -2.51 fold changes respectively (Boiko et al., 2006). *Btg2* was also identified as a target of miR-9 (-1.56 FC) and the increased translation efficiency of this gene may be a result of both differential transcript usage and decreased miRNA mediated repression.

Differential translation efficiency was also observed in genes which may impact the metabolic efficiency of CHO cells. Mitochondrial Ribosomal Protein L54 (*Mrpl54*) encodes the large 39S subunit of the mitochondrial ribosome required for translation of mitochondrial proteins and is required for efficient oxidative metabolism and general mitochondrial functionality (Andreux et al., 2012). While *Mrpl54* was not identified as differentially expressed at the gene level (-1.17 FC) a 1.57 fold increase in translational efficiency was observed. The Translocase of inner mitochondrial membrane 8A (*Timm8a*) gene encodes a chaperone responsible for transporting proteins through the inner mitochondrial membrane (Kulawiak et al., 2013). Timm overexpression results in increased mitochondrial function and energy metabolism due to an increase in solute carrier trafficking across the inner mitochondrial membrane (Lin et al., 2017). The Solute Carrier Family 25 Members 12 and 13 (Slc25a12 and Slc25a13) are transported by Timm8a (Roesch, 2004) and have roles in mitochondrial DNA replication (Favre et al., 2010), to transport substrates of the TCA cycle (Wongkittichote et al., 2013), and glucose metabolism (Contreras et al., 2016). Slc25a12 overexpression has been previously reported to be involved in metabolic switching to lactate consumption in CHO cells (Zagari et al., 2013). The *Timm8a* gene had a 1.69 fold increase of translation efficiency induced by temperature shift but gene level expression was unchanged (-1.07 FC). *Timm8a* was predicted as a target of miR-30a-5p (-1.33 FC) and Mrpl54 as a target of miR-615-3p (-1.39 FC) and therefore the increased translation of *Timm8a* and *Mrpl54* may be due to decreased miRNA mediated repression. Increased translation of *Timm8a* and *Mrpl54* may contribute to the switch in metabolism observed in temperature shifted CHO cells by enhancing translation of mitochondria proteins and their transport thereby increasing the cells capacity to carry out OXPHOS.

miRNA target prediction was carried out on differentially translated genes revealing 110 target predictions on 88 genes between differentially expressed miRNAs and genes with differential translation efficiency in the opposite direction. The differences in translation efficiency of these 88 genes is likely due to these miRNA interactions through miRNA mediated translational repression by inhibiting assembly of translation initiation factors (Gebert and MacRae, 2018). Translation initiation factors were among the genes predicted as targets of miRNAs – with upregulated translation efficiency identified in the Eukaryotic translation initiation factor 2 subunit 2 (*Eif2s2*) (1.77 FC) and downregulated translation in the Eukaryotic translation initiation factor 4 gamma 3 (*Eif4g3*) gene (-1.77 FC). *Eif2s2* is the subunit of Eif2 which binds GTP required for Eif2 to initiate translation (Liao et al., 2014) while Eif4g3 is a component of the Eif4f complex required for cap-dependant translation initiation (Watson et al., 2014). *Eif2s2* was predicted as a target of miR-374-5p (downregulated with a -1.52 FC) and *Eif4g3* was identified as a target of miR-221-3p (upregulated with a 1.4 FC). Alterations to the translation efficiency of these genes may impact global translation which has been shown to be decreased in CHO cells cultured at a reduced temperature (Al-Fageeh and Smales, 2009). In addition to translation initiation factors, miRNA mediated regulation of translation efficiency was identified in regulators of apoptosis. The Apoptogenic 1, Mitochondrial (*Apopt1*) gene was identified with decreased translation efficiency (-2 FC) and predicted as a target of the upregulated miR-34a-5p (3.33 FC). Apopt1 is able to release cytochrome C from mitochondria into the cytosol and induce apoptosis (Yasuda et al., 2006). The Endophilin-B1 (*Sh3glb1*) (-1.63 FC TE) was identified as a target of miR-7-5p (3.01 FC). *Sh3glb1* is also an inducer of apoptosis which causes release of cytochrome C through activation of the BCL2 associated x apoptosis regulator (Bax) protein (Takahashi et al., 2005). Decreased translation efficiency of *Apopt1* and *Sh3glb1* mediated by miRNA upregulation may result in delayed apoptosis. Temperature shift in CHO cells results in increased viability and elongated culture duration due to delayed apoptosis (Kumar et al., 2008) which may be a result of differential translation of these genes.

Also identified as a target of miRNA mediated translation repression was the Dihydrofolate reductase gene (*Dhfr).* The translation efficiency of the *Dhfr* gene was the most significantly decreased with a -4.08 fold change and was identified as a target of miR-192-5p (1.91 FC). *Dhfr* is used as a selection marker in CHO cell engineering and is present in cells at a high copy number due to amplification of the *Dhfr* gene alongside the transfected product sequence (Voronina et al., 2016). Expression of Dhfr at a high level is no longer required in cells in production culture as methotrexate treatment to inhibit Dhfr activity is only used during the cell line development phases of CHO cell culture. A previous experiment in CHO cells applied RiboSeq and identified unnecessary translation in remnants of selection markers sequestered ribosomes and reduced productivity (Kallehauge et al., 2017). Reduced translation efficiency of *Dhfr* in temperature shifted CHO cells may therefore increase productivity by freeing ribosomal capacity for translation of recombinant proteins. Further reduction of *Dhfr* translation using siRNA knockdown or increasing miR-192-5p expression may result in increases to productivity.

Surprisingly, the *Dhfr* gene was downregulated at the gene level (-1.11 FC) and had decreased translation efficiency, but was identified as upregulated at the protein level with a 1.36 FC. Unexpected fold changes were also identified in the Eukaryotic translation initiation factor 3 subunit J (*Eif3J*) which had a 1.15 FC at the gene level, a 1.66 FC increase in translation efficiency but a -1.57 decrease in protein levels. Temperature shift may have induced a halting to the production of these proteins however the proteins have not yet degraded to match relative levels of transcription and translation (Liu et al., 2016). Differences in gene and protein levels may then be due to differences in the half-life of mRNAs and proteins (Liu et al., 2016). In addition, 197 genes were identified with differential translation efficiency which were not a target of miRNAs or identified as alternatively spliced in Chapter 3 including the *Cdk11b* and *Fbxw7* genes discussed for their roles in regulation of the cell cycle. Other mechanisms of post-transcriptional control likely contribute to the regulation of gene expression in addition to alternative splicing and miRNA activity in temperature shifted CHO cells. Potential mechanisms through which this may occur include selective regulation of the translation of individual genes by lncRNAs (Dimartino et al., 2018; Zucchelli et al., 2015), translation activators (Roilo et al., 2018) or translation inhibitors (Lyabin and Ovchinnikov, 2016).

The cold shock response in mammalian cells has previously been reported to be regulated by *Cirbp* and *Rbm3* (Masterton and Smales, 2014), which have upregulated expression in response to temperature shift as shown in Chapter 2. Hypothesised roles of Cirbp and Rbm3 include maintaining the stability of their mRNA targets, inhibiting miRNA activity, inhibiting apoptosis pathways and acting as a regulator of translation (Masterton and Smales, 2014). The role of Cirbp in translation is thought to be mediated by binding the 3'UTR of its mRNA targets and directly interacting with Eif4g (Yang et al., 2010), therefore resulting in increased transcription of its protein targets via assembly of the pre-initiation complex. Recently emerging evidence has shown Cirbp binding to the p27 5'UTR enhances its translation either through releasing p27 mRNA from being sequestered in stress granules or by directly promoting its translation, resulting in increased p27 protein which inhibits the cell cycle (Roilo et al., 2018).

There has been one previous report of the analysis of a CHO cell growth related phenotype using methodology to profile the translatome published in the literature (Courtes et al., 2013). Courtes et al. identified 657 genes with uncoupled mRNA levels and translation rates throughout the exponential growth phase of culture including 5 key growth genes which were highlighted as highly translated – *Utp6*, *Pcna*, *Hnrnpc*, *Vcp* and *Mcm5*. The 5 highlighted growth genes were not identified as differentially translated in this experiment which suggests alternative mechanisms are responsible for regulating the cell cycle during exponential growth phase and in response to mild hypothermia at the level of the translatome. While none of these 5 genes were identified as differentially translated between temperature shifted and non-temperature shifted cells the *Mcm5* gene was found to be stably translated regardless of the condition in this experiment.

Of the 657 genes identified with uncoupled mRNA and translation rates in Courtes et al. 4 genes were found to have increased translation in temperature shifted cells (*Dph3*, *E2f1*, *Smarce1* and *Spsb2*) and 4 with decreased translational efficiency (*Ctdnep1*, *Hivep1*, *Rnf123* and *Tspan17*). The Diphthamide biosynthesis 3 (*Dph3*) gene is required for diphthamide biosynthesis – a posttranslational modification required for Eef2 activity in translation elongation and impacts the efficiency of protein synthesis in CHO cells (Liu et al., 2012). *Dph3* was identified as poorly translated in CHO cells in the exponential growth phase by Courtes et al. but was differentially translated in this experiment with upregulated translation efficiency after temperature shift. Increased translation of Dph3 may further increase a cells capacity to synthesise proteins and affect the rates of recombinant protein production in temperature shifted cells. In addition the E2F transcription factor 1 (*E2f1*) gene (a master regulator of the cell cycle in the G1/S phase transition (Caldon and Musgrove, 2010) which was identified as downregulated at the gene level in Chapter 2) also had increased translation efficiency after temperature shift but decreased translation efficiency in Courtes et al.. Increased translation of E2f1 in temperature shifted cells which have a halted cell cycle and decreased translation efficiency of E2f1 in CHO cells in the exponential growth phase is unexpected and suggests E2f1 alone is not sufficient to regulate the cell cycle in CHO cells. E2f activity is dependent on forming a heterodimer with transcription factor Dp-1 (Tfdp1) (Zaragoza et al., 2010) which was observed as undergoing temperature shift induced alternative splicing in Chapter 3 (ENSCGRG00000002395) with 2 retained intron events. *Tfdp1* was however not identified as differentially expressed or translated in this experiment and the intron retention events do not interrupt the encoded amino acid sequence and therefore the role of *E2f1* in the CHO cell cycle requires further investigation. The SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily E, member 1 (*Smarce1)* gene was identified with high translation efficiency in exponentially growing CHO cells (Courtes et al., 2013) and had upregulated translation efficiency in temperature shifted CHO cells. *Smarce1* encodes a chromatin remodelling complex which is often overexpressed in cancer cell lines (Yamaguchi et al., 2015) and regulates the G1 phase of growth. As with *E2f1*, enhanced translation of *Smarce1* is unexpected in CHO cells with a halted cell cycle and inhibiting the translation of *Smarce1* mRNA may have the potential to enhance the temperature shift phenotype.

The CTD nuclear envelope phosphatase 1 *(Ctdnep1*) gene was poorly translated in Courtes et al. and differentially translated in temperature shifted CHO cells with downregulated expression. The Lipin 1 and Lipin 2 phosphatases which convert phosphatidic acid to diacylglycerol are regulated by Ctdnep1 (Han et al., 2012) and have roles in the cell cycle (Jiang et al., 2015), fatty acid and energy metabolism (Zhang and Reue, 2017) and mitochondrial fission (Frohman, 2015). Downregulation of *Ctdnep1* may reduce a cells capacity to both proliferate in temperature shifted cells and also reduce rates of mitochondrial respiration in cells in the exponential growth phase. Engineering CHO cells to upregulate or downregulate *Ctdnep1* transcription and translation may result in enhanced cell lines with altered energy metabolism and growth rates. Finally, the Ring finger protein 123 (*Rnf123*) gene was identified as efficiently translated by Courtes et al. but decreased in translational efficiency after temperature shift in this experiment. Rnf123 is a ubiquitin ligase which regulates the degradation of p27^{Kip1} (encoded by *Cdkn1b*) in the G1 phase (Kamura et al., 2004) to prevent inhibition of cyclins and cyclin dependant kinases which progress the cell cycle. As *Rnf123* has been previously shown to be efficiently translated in rapidly proliferating CHO cells and *Rnf123* is differentially translated in temperature shifted CHO cells it likely plays a role in regulating the cell cycle and may be a viable target for cellular engineering enhanced CHO cells.

Genes identified with high translational efficiency in Courtes et al. were utilised in a subsequent multi-omics investigation to identify codon bias in efficiently translated genes (Ang et al., 2016). Codon bias was identified in CHO cells between genes with high and low transcript or protein expression or translation efficiency utilising codon context to establish a reference codon bias and used in optimisation of the human IFNγ sequence. It may be possible to utilise the findings of this experiment to enhance translation of recombinant protein products by optimising codons to reflect those in genes identified as most efficiently translated in both temperature shifted and non-temperature shifted samples. Doing such may ensure recombinant product is stably translated at a high rate throughout the duration of culture both before and after temperature shift. Translation efficiency fold changes within a given condition as high as 46.37 fold were observed, with the Myeloid associated differentiation marker (*Myadm*) gene the most efficiently translated. Further investigation into the properties of these highly translated genes may reveal findings useful for engineering recombinant products in the biopharmaceutical industry which are highly translated throughout the duration of a culture.

Recent evidence has shown the ribosome itself can be heterogeneous as opposed to an unchangeable inflexible structure (Genuth and Barna, 2018) and not all ribosomal proteins are present in every ribosome within a cell (Shi et al., 2017). Ribosomal heterogeneity can be tissue specific (Kearse et al., 2011) and ribosomes with particular ribosomal proteins incorporated into their structure may have higher affinity for the translation of specific mRNAs such as those containing IRES elements (Shi et al., 2017). RNA-Seq has been utilised to profile the heterogeneity of ribosome component expression in different tissues (Guimaraes and Zavolan, 2016; Gupta and Warner, 2014) revealing widespread differential expression and heterogeneity, however, this has not been studied using proteomics or at the level of individual ribosome complexes (Genuth and Barna, 2018). Differential translation was identified in the Ribosomal protein lateral stalk subunit P0 (Rplp0) which was downregulated with a -1.84 FC in TE, not differentially expressed at the gene level (1.22 FC) but upregulated at the protein level (1.4 FC). Rplp proteins have been shown to exhibit tissue specific expression and knockout leads to cell cycle arrest through dysregulation of cyclins, p21, p27 and p53 (Perucho et al., 2014). *Rplp0* has been shown to be overexpressed in replicating cells (Artero-Castro et al., 2011) and therefore decreased translational efficiency of *Rplp0* in mild hypothermia conditions may result in ribosomal heterogeneity - with ribosomes less efficient at translating mRNAs encoding cell cycle progression proteins and therefore contribute to cell cycle arrest.

Through the translatomics experiment detailed throughout this Chapter, gene targets for engineering enhanced producer CHO cell lines were identified. The *Cdkn2c*, *Fbxw7*, *Gdf11* and *Btg2* genes which negatively regulate the cell cycle (Boiko et al., 2006; Ibusuki et al., 2011; Jalili et al., 2012; Shi and Liu, 2011) were all identified with enhanced translation efficiency in temperature shifted CHO cells. Further enhancing the expression and translation of these genes may result in increased productivity through halting the cell cycle and resulting in an elongated stationary phase of culture in which cells are most productive (Moore et al., 1997). In addition, inhibiting the expression and translation of the *Apopt1* and *Sh3glb1* genes may result in a suppression of apoptosis (Takahashi et al., 2005; Yasuda et al., 2006) to increase culture duration and therefore increase recombinant protein yields. Finally, capacity for OXPHOS may be enhanced through increasing expression and translation of *Timm8a* and *Mrpl54* (Andreux et al., 2012; Lin et al., 2017; Zagari et al., 2013) thereby increasing the resources available for recombinant protein production and decreasing lactate production to increase yields.

5.6 Conclusion

Integration of multi-omics datasets in this Chapter has revealed measurements of gene expression using RNA-Seq or proteomics in isolation are not sufficient for understanding the complex regulation of gene expression in CHO cells. Differential translation efficiency was identified in key apoptosis, cell cycle and energy metabolism genes which may contribute to the observed reduction in growth rate and metabolic switching in cells cultured at a decreased temperature. The genes identified in this chapter with differential translational efficiency may serve as potential targets for genetic engineering enhanced producer cell lines. Further diminishing the expression of apoptosis activators such as *Apot1* or enhancing the expression of cell cycle progression inhibitors such as *Btg2* may be able to further extend the viability of CHO cells for increased batch culture yields. While our knowledge of the CHO cell response to temperature shift at the level of translation was enhanced through this analysis, further studies are required to understand the translational regulation of the 198 differentially translated genes which were not predicted as targets of differentially expressed miRNAs or undergoing alternative splicing.

Chapter 6

Summary & Future Work

6.1 Summary

6.1.1 Significance of this dataset

The dataset generated and analysed throughout this thesis is of high significance for enhancing our understanding of CHO cells for the purposes of biopharmaceutical manufacturing. As discussed in Section 1.7.1, temperature shift is widely applied throughout the biopharmaceutical industry as a means of increasing culture duration and productivity to produce more recombinant proteins however the biological mechanisms behind this phenotype are not fully understood. The dataset in this experiment represents the largest multi-omics study carried out to date in CHO cells with gene expression, translation, protein expression and post-transcriptional regulation captured in parallel in addition to metabolic profiling. Application of these techniques in parallel has for the first time enabled a global study from alterations to gene expression from the level of the mRNA up to the protein and enhanced our understanding of the levels of alternative splicing, miRNA activity and differential translation efficiency. In addition to being the most comprehensive study in terms of the number of ‘omics techniques applied in parallel, each of the individual datasets produced are of high quality and sequencing depth. The RNA-Seq and small RNA-Seq datasets produced are of the highest sequencing depth carried out to date in published literature on CHO cells which allowed for the characterisation of previously unannotated genes and therefore have applications beyond improving our understanding of temperature shift. One major finding of this experiment is the level of post-transcriptional activity which occurs in CHO cells resulting in a lack of correlation between mRNA and protein levels. In particular, alternative splice isoforms in genes such as *Slirp* and *Dnm1l* illustrate how although the overall expression of a gene may not change, the expressed transcript isoforms may be altered in a way which affects activity of their protein product by interrupting protein domains. The widespread alternative splicing identified in temperature shift also suggests there may be an additional layer of heterogeneity in production CHO cell cultures which should be monitored. The findings of this thesis using this dataset suggests a multi-omics approach is utilised in CHO cell ‘omics experiments going forward as a single level of omics is not enough to provide an overview of gene expression alterations.

6.1.2 Novel contributions

Through the work described in this thesis a number of novel contributions have been made to enhancing our understanding of CHO cell biology. Contributions to the field include improving our characterisation of the transcriptomic landscape of CHO cells and improving the annotation of the CHO cell genome, generating a pipeline to analyse multi-omics datasets, carrying out the first multi-omics analysis to reveal the scale of alternative splicing, miRNA activity and impact on differential translation in parallel. Utilising extensive literature searching, expression patterns in genes which have not been previously attributed to temperature shift induced altered energy metabolism and regulation of the cell cycle were identified, as were cases of alternative splicing altering protein domains which likely impact activity and miRNA targeting in CHO cells. Finally, a panel of high priority targets for genome engineering were selected which have the potential to enhance productivity in CHO cells by reinforcing the beneficial effects of the temperature shift phenotype. Each of these novel contributions are summarised further throughout this section.

6.1.2.1 Progress towards improving the assembly and annotation of the Chinese hamster genome

At the onset of the work carried out in this thesis in 2014 there were 3 Chinese hamster genome sequences which were each at an early stage of their assembly and annotation (Brinkrolf et al., 2013; Lewis et al., 2013; Xu et al., 2011). While attempts were made to identify transcripts present in these genomes (Becker et al., 2011; Rupp et al., 2014) the annotation for these genomes remained at an early stage of completeness throughout this research. August 2017 saw the release of Ensembl version 90 which included annotation for the Chinese hamster (Zerbino et al., 2018). Release of the Ensembl genome annotation of the CHO cell genome (Xu et al., 2011) served as a major landmark in the progress of CHO cell bioinformatics resources bringing curated genome annotation, stable gene identifiers, and enabling usage of Biomart and conversion between identifiers of orthologous species genes and CHO cells (Durinck et al., 2009). The ability to query Biomart to retrieve information on genes throughout this thesis was an invaluable resource for the integration of multi-omics databases. While this genome annotation bought useful functionality, annotation progress in terms of annotated genes is still at an early stage. 30,673 gene loci and 34,472 transcripts are annotated in the latest version of the CHO cell genome in Ensembl (release 93). The mouse however has 51,086 gene loci and 136,630 transcripts annotated in Ensembl. The discrepancy between the number of gene and transcript annotations of these two closely related rodent species suggests a large amount of transcripts are unannotated in the CHO cell genome, particular alternative transcript isoforms. In Chapter 3 RNA-Seq reads from non-temperature shifted and temperature shifted CHO cell cultures were assembled into novel transcripts using the Ensembl genome annotation as a guide. The dataset utilised represents the most deeply sequenced RNA-Seq libraries produced from CHO cells to date – enabling the annotation of 10,019 novel gene loci,27,687 novel transcripts and 251 previously uncharacterised miRNAs. The improved genome annotation produced in this experiment could serve as a valuable resource for future ‘omics research in CHO cells.

6.1.2.2 A pipeline for multi-omics analysis of CHO cells

In order to carry out experiments detailed in this thesis the vast library of bioinformatics software needed to be tested, optimised and configured for use with CHO cell data. Software for the quality control, assembly of transcripts, identification of differential gene expression, differential isoform usage, differential exon usage, identification of novel miRNAs and miRNA:mRNA target interactions, and analysis of RiboSeq data was tested throughout this research project. Chinese hamster genome sequences were also benchmarked to find the genome which provides the most biologically informative results. The code used to replicate this analysis can be found in Appendix 6. The pipeline can be readily adapted for use in other experiments or by other lab groups. The pipeline used in this experiment is summarised in Figure 6.1.

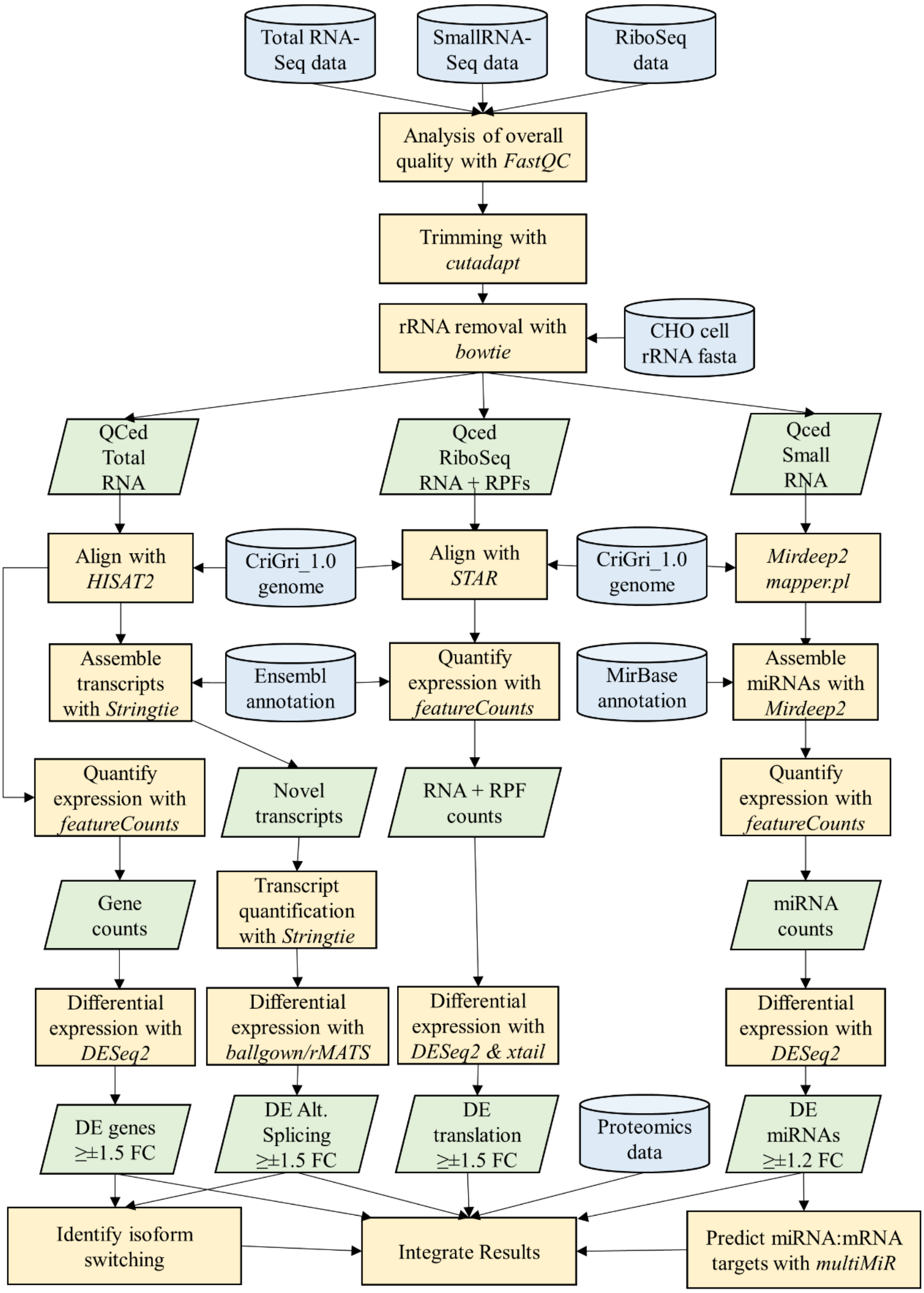
Total RNA-Seq data
SmallRNA-Seq data
RiboSeq data
Analysis of overall quality with *FastQC*
Trimming with *cutadapt*
rRNA removal with *bowtie*
CHO cell rRNA fasta
QCed Total RNA
Qced RiboSeq RNA + RPFs
Qced Small RNA
Align with *HISAT2*
CriGri_1.0 genome
Align with *STAR*
CriGri_1.0 genome
Mirdeep2 mapper.pl
Assemble transcripts with *Stringtie*
Ensembl annotation
Quantify expression with *featureCounts*
MirBase annotation
Assemble miRNAs with *Mirdeep2*
Quantify expression with *featureCounts*
Novel transcripts
RNA + RPF counts
Quantify expression with *featureCounts*
Gene counts
Transcript quantification with *Stringtie*
miRNA counts
Differential expression with *DESeq2*
Differential expression with *ballgown/rMATS*
Differential expression with *DESeq2 & xtail*
Differential expression with *DESeq2*
DE genes ≥±1.5 FC
DE Alt. Splicing ≥±1.5 FC
DE translation ≥±1.5 FC
Proteomics data
DE miRNAs ≥±1.2 FC
Identify isoform switching
Integrate Results
Predict miRNA:mRNA targets with *multiMiR*

Figure 6.1 – Multi-omics pipeline used to analyse CHO cells. In this thesis 3 NGS datasets were analysed to find differential gene, transcript and miRNA expression as well as isoform switching, differential translation and predicted miRNA targets.

6.1.2.3 Identification of correlations among omics levels

Little correlation was observed between fold changes at the gene, protein and translational efficiency levels in temperature shifted and non-temperature shifted cells (R^2 values of 0.0038, 3.66×10^{-5} and 0.028 for gene vs protein, gene vs TE and protein vs TE respectively). A lack of correlation between these fold changes is expected due to the influence of post-transcriptional regulation as is captured in the RiboSeq and small RNA-Seq experiments. Despite the lack of correlation, trends in the multi-omics data were observed which have contributed to furthering our understanding of the temperature shift phenotype in CHO cells.

Cyclin dependant kinase inhibitors were observed to be differentially regulated at the gene, transcript isoform, miRNA and translation level throughout this experiment. The *Cdkn1a* gene, an inhibitor of the cell cycle which regulates Cyclin E activity to prevent activation of Cdk2 and progression from the G1 to the S phase of the cell cycle (Bertoli et al., 2013; Caldon et al., 2009), was identified as one of the most significantly upregulated genes in temperature shift. Alternative splicing may contribute to the regulation of *Cdkn1a* through the *Ybx3* gene which enhances the mRNA stability of *Cdkn1a* (Nie et al., 2012) which was not identified as differentially expressed, but was found to be alternatively spliced in Chapter 3. In addition the miRNAs miR-290-5p and miR-295 were observed as expressed for the first time in CHO cells which are known to regulate *Cdkn1a* expression (Wang et al., 2008). The downregulated miR-9 targets the transcription factor *Foxo1* which also regulates expression of *Cdkn1a* and *Cdkn1b* (Coomans de Brachène and Demoulin, 2016; Xie et al., 2012). *Cdkn1b* expression is also regulated by *Gdf11* which was observed to be differentially translated with increased efficiency, as were the *Cdkn2c* and *Fbxw7* genes – all of which inhibit cyclin activity (Ibusuki et al., 2011; Jalili et al., 2012; Ray et al., 2009; Shi and Liu, 2011). Inhibition of cyclin activity is therefore likely the major mechanism through which temperature shift induces a halting of the cell cycle in the G1 phase where cells are the most productive (Moore et al., 1997) and is orchestrated through alterations to gene expression, mRNA processing and post-transcriptional regulation. In addition the *Cdkn1a* gene has been predicted to be bound by Cirbp in previous experiments (Morf et al., 2012) and therefore the CHO cell temperature shift induced halting of the cell cycle may be in part due to *Cirbp* expression however overexpression of *Cirbp* alone has been shown to be insufficient to result in a halting of the cell cycle (Tan et al., 2008).

Differential expression of regulators of a switch from a Warburg metabolism to OXPHOS were also prevalent in the output of all the ‘omics experiments in this thesis. Literature review of genes identified as differentially expressed or undergoing post-transcriptional regulation identified genes which may impact metabolism and contribute to the switch from Warburg metabolism to use of OXPHOS in temperature shifted CHO cells. The lactate dehydrogenase genes *Ldha*, *Ldhc* and *Ldha16b* were identified as upregulated at the gene level in Chapter 2 while *Ldhd* was identified as downregulated. In addition *Ldha* and *Ldha16b* were identified as alternatively spliced in Chapter 3 and *Ldha* was found to be upregulated at the protein level. Alterations to the expression patterns of lactate dehydrogenases has been shown to result in reduced levels of anaerobic glycolysis previously (Li et al., 2016) and may therefore contribute to the change in metabolism observed in temperature shifted CHO cells. *Pkfm*, a rate limiting enzyme involved in the regulation of glycolysis (Tang et al., 2012) was observed as downregulated at the gene level and alternatively spliced. The *Pdk3* and *Pparg* genes also regulate glycolysis rates by controlling the conversion of pyruvate to acetyl-Coa to fuel the TCA cycle (Lecarpentier et al., 2017; Lu et al., 2008) and were both identified as alternatively spliced in Chapter 3. *Pdk3* and *Pkfm* have been previously shown to regulate the switch from a Warburg metabolism to OXPHOS in other species (Kluza et al., 2012; Tang et al., 2012) – as have the *Prkab2* and *Prkag1* subunits of Ampk (Adams et al., 2004; Salazar-Roa and Malumbres, 2017) which were differentially expressed at the gene level. In addition the *Cd44* gene, which has also been shown to regulate the Warburg metabolism (Chen et al., 2018) was found to be upregulated at the gene but not the protein level and to undergo complex alternative splicing. Metabolic switch inducing miRNAs were also found to be differentially expressed including miR-199a (el Azzouzi et al., 2013; Zhang et al., 2015) and the *Pkia* gene was found to be downregulated and a target of the upregulated miR-103a, miR-7, miR-34a and miR-221 genes which has also been shown to result in switching to primary use of OXPHOS (Xing et al., 2017). In order to integrate the findings of each experiment in this thesis and illustrate how differential gene, protein, splicing, translation and miRNA targeting impacts energy metabolism in temperature shift, results were projected onto the RECON1 Glycolysis and TCA metabolic pathway (King et al., 2015) using gene identifiers assigned to each reaction in the iCHOV1 model (Hefzi et al., 2016) (Figure 6.2).

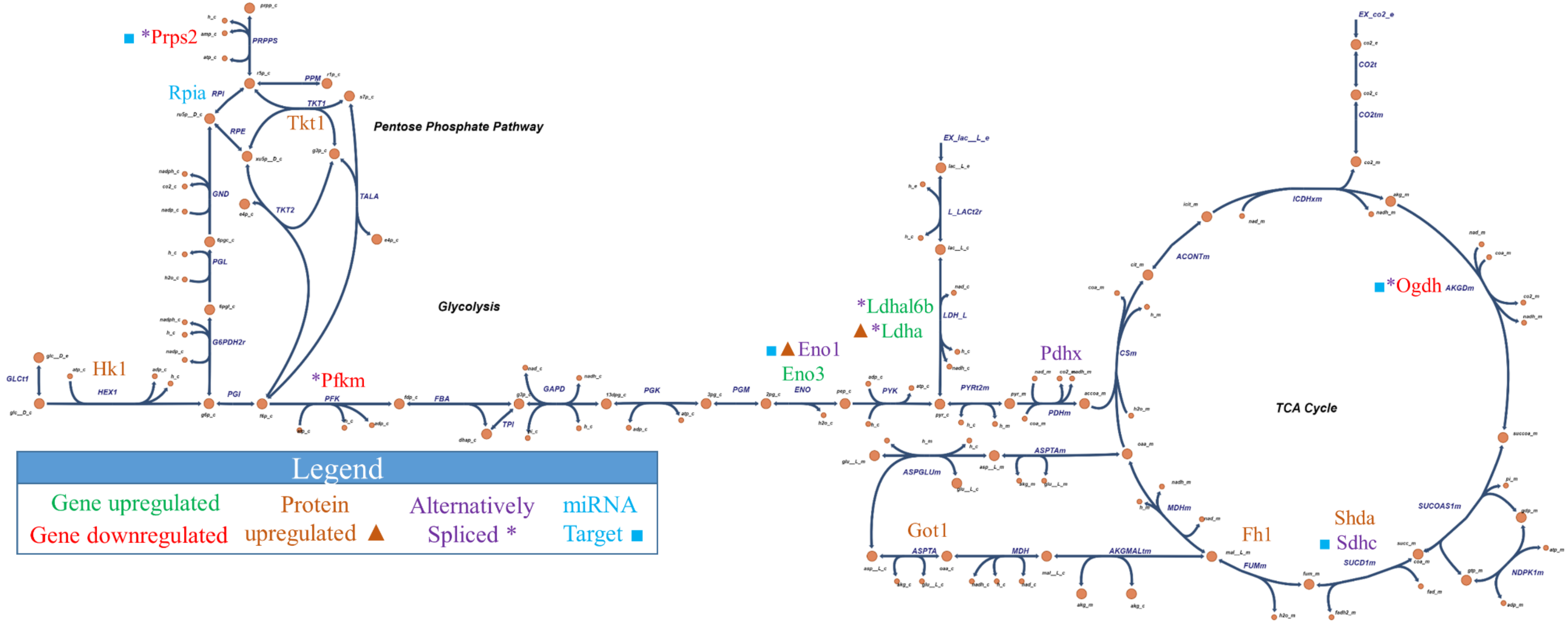

Figure 6.2 – Regulation of the central glycolysis and TCA pathways in CHO cells during temperature shift. Presented in this figure is the glycolysis and TCA cycle pathway with genes differentially expressed or post-transcriptionally regulated illustrated. Genes are coloured as depicted in the legend for the type of regulation they are affected by, and denoted with symbols if affected by multiple types of regulation. Pathway image generated using RECON1 Glycolysis and TCA metabolic model on Escher (King et al., 2015).

In summary these findings illustrate for the first time in CHO cells not only does differential expression play a role in regulating metabolic switching and the cell cycle during temperature shift, but also post-transcriptional regulation. The findings of this thesis also suggests Cyclin dependant kinase inhibitors play a central role in regulating the cell cycle during temperature shift in CHO cells and have identified many potential regulators of metabolic switching from a Warburg metabolism to primary use of OXPHOS.

6.1.2.4 Multiomics integration enhances our understanding of temperature shift in CHO cells

Temperature shift results in widespread differential splicing in CHO cells

In Chapter 3 the transcriptome was surveyed for the presence of alternative splicing and differential transcript usage between temperature shifted and non-temperature shifted CHO cells. Widespread alternative splicing was identified in 1,789 genes including those with roles in the cell cycle, apoptosis and energy metabolism. Overlap of differential gene expression results from Chapter 2 with differential transcript expression results in Chapter 3 revealed a striking amount of alternative splicing in genes which were not identified as differentially expressed at the gene level. Identification of alternative splicing in these 1,492 genes illustrates the potential for missed information in RNA-Seq analyses performed at the gene level and the usage of an isoform level approach for RNA-Seq experiments performed in the future is recommended. Application of qPCR was utilised to confirm the presence of alternative splicing in assembled transcripts, validating temperature shifted induced alternative splicing for the first time in CHO cells and revealing alternative splicing in the *Slirp* and *Srfbp3* genes is conserved across CHO cell lineages. In addition, cases of splicing altering protein domains were identified such as in the *Slirp, Dnm1l* and *Ybx3* genes as well as the first case observed in an animal species of splicing in the 3'UTR altering miRNA binding sites.

Multi-omics integration of CHO cell data reveals extensive post-transcriptional regulation of genes controlling the G1/S phase transition of the cell cycle

Temperature shift has previously been shown to result in stalling of the cell cycle in the G1 phase of the cell cycle in CHO cells (Underhill and Smales, 2007). A 24% reduction in growth rate was observed between cells shifted from 37°C to 31°C with no change in cellular viability which suggests a decrease in proliferation rate. Increased viability and productivity in temperature shifted CHO cells has been previously attributed to an elongated G1 phase (Moore et al., 1997) however the molecular mechanisms underpinning this phenotype are not fully understood. Multi-omics integration of RNA-Seq, RiboSeq, small RNA-Seq and LC-MS/MS data revealed temperature shift induced widespread changes to the transcription and translation of genes controlling the G1/S phase. The interactions of differentially expressed genes regulating the cell cycle identified throughout this thesis are summarised in Figure 6.3.

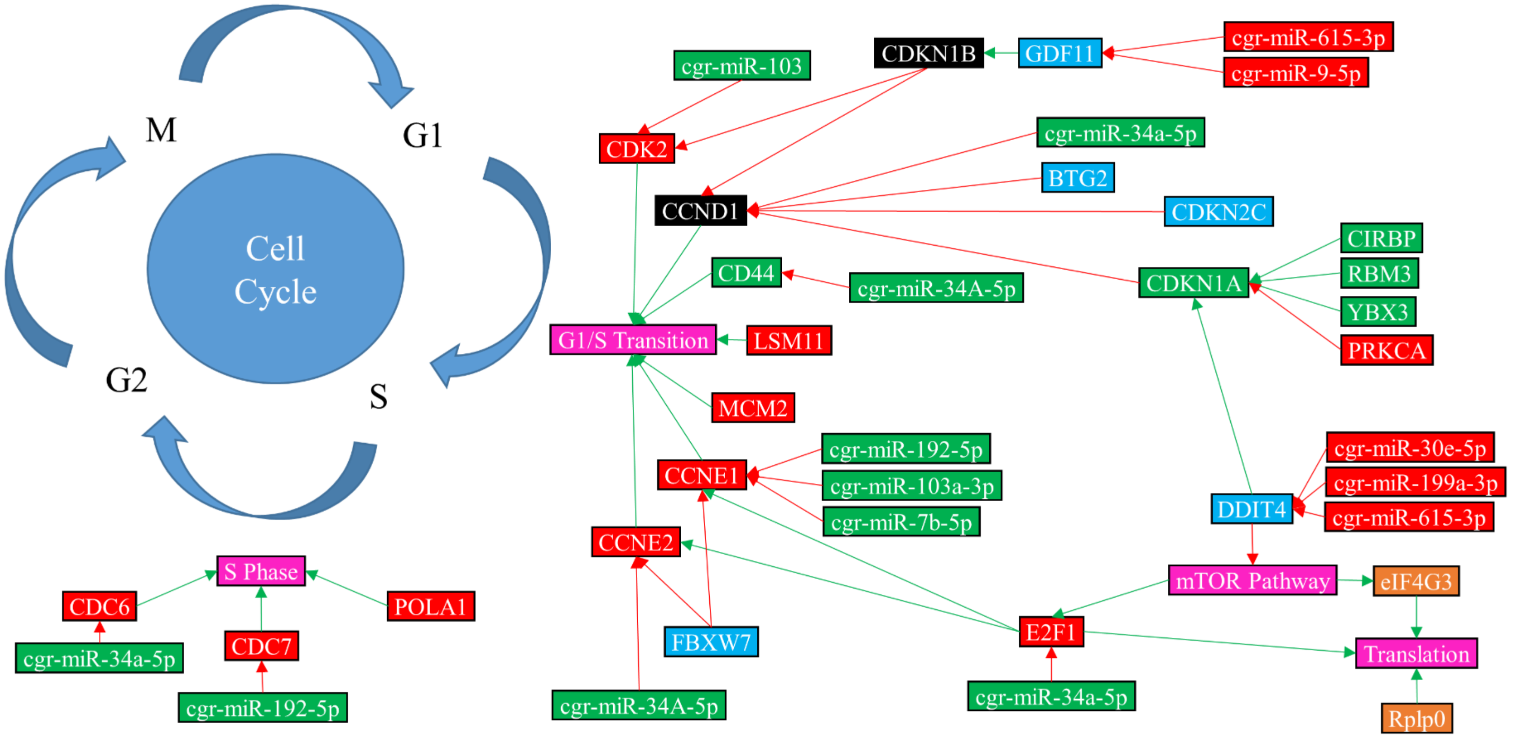

Figure 6.3 – Multi-omics analysis of proliferation genes in temperature shifted CHO cells. Presented in this figure are the interactions of differentially expressed genes regulating the cell cycle throughout this thesis. Positive interactions are denoted by green arrows while negative interactions are depicted by red arrows. Upregulated genes or transcripts are coloured green and downregulated red. Genes with upregulated translation are coloured blue and downregulated translation orange. Non-differentially expressed genes are coloured black and pathways/biological processes pink.

Multi-omics integration of CHO cell data reveals regulation of glycolysis and oxidative phosphorylation genes in the temperature shift response

CHO cells in the exponential growth phase exhibit a Warburg type metabolism in which glycolysis is the primary source of ATP generation despite being less efficient than oxidative phosphorylation in terms of ATP produced per glucose molecule and generating lactate as a by-product which accumulates in culture media and is toxic to cells (Buchsteiner et al., 2018). Temperature shift has been shown to reverse Warburg metabolism and CHO cells undergoing cold shock stop producing lactate and using glycolysis as the primary fuel source (Fogolín et al., 2004; Young, 2013). A reduction in lactate and ammonia levels in temperature shifted samples as well as increased levels of glucose remaining in media suggest this is the case in the temperature shift model used for multi-omics analysis throughout this thesis. Integration of multi-omics datasets was utilised to investigate the cause of this metabolic shift in CHO cells revealing regulation of gene expression at the transcription and translation level which may impact rates of glycolysis, lactate consumption and mitochondrial activity. Interactions of differentially expressed and translationally regulated genes are presented in Figure 6.4.

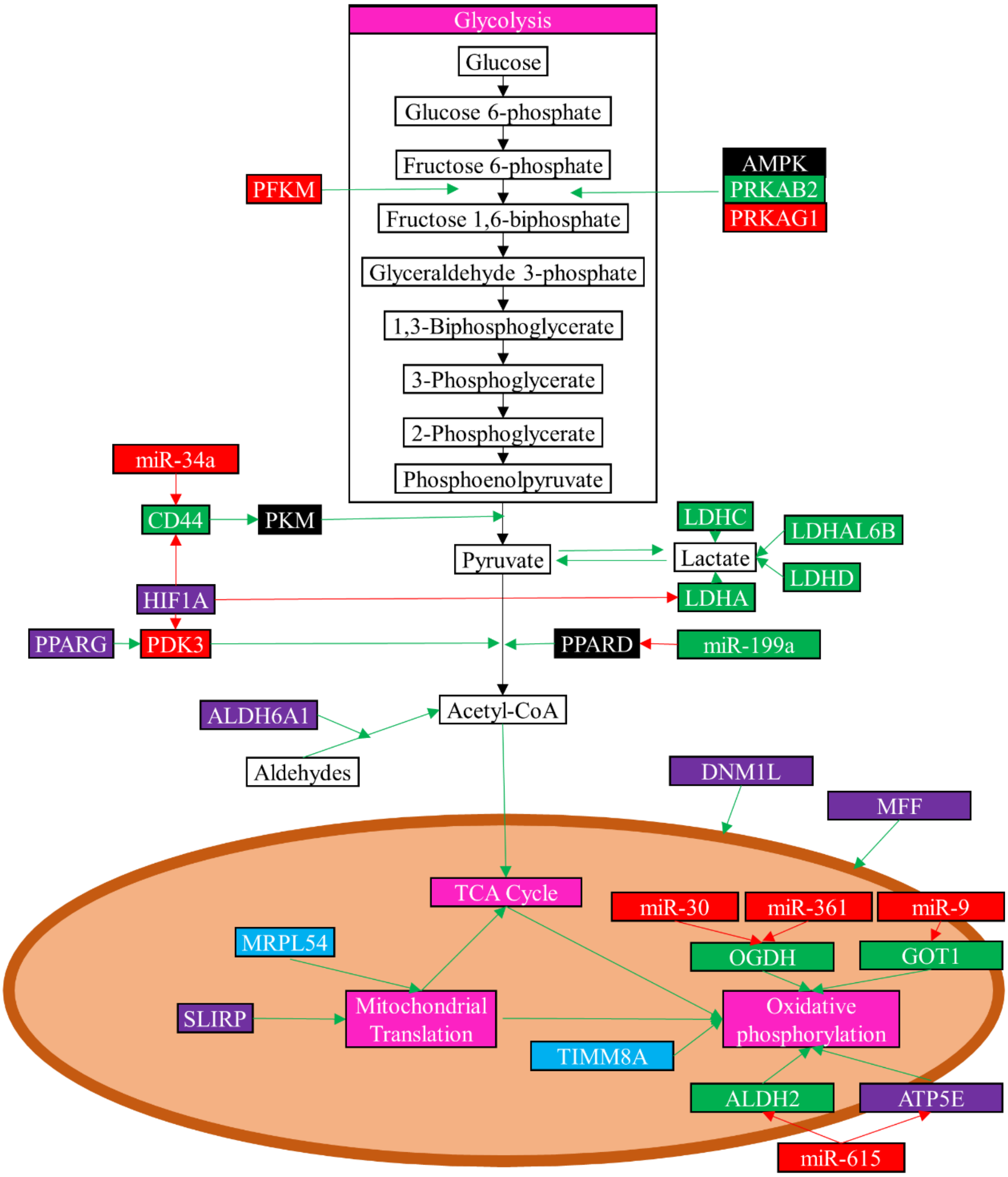

Figure 6.4 - Multi-omics analysis of energy metabolism genes in temperature shifted CHO cells. Integration of RNA-Seq, small RNA-Seq, proteomics, metabolomics, and RiboSeq data revealed regulation of energy metabolism was altered in temperature shifted CHO cells. Positive interactions are denoted by green arrows while negative interactions are depicted by red arrows. Upregulated genes or transcripts are coloured green and downregulated red. Isoform switching is coloured purple. Genes with upregulated translation are coloured blue. Non-differentially expressed genes are coloured black and pathways/biological processes pink.

6.1.2.5 - Selection of a panel of key engineering targets

Through the multi-omic investigation carried out as described in this thesis, a panel of high priority gene targets has been selected which may result in enhanced producer cell lines with enhanced productivity and culture longevity. Table 6.1 presents this panel of gene targets, why they were selected and suggests methodology which could be utilised to test if engineering these genes results in enhanced producer cell lines.

Table 6.1 – Engineering targets identified through multi-omics characterisation of the temperature shift response. Presented in this table are engineering targets which have not previously been utilised for engineering enhanced CHO cells in the literature and are selected as high priority targets for engineering due to their identification as differentially regulated in this experiment and their known functions in other species.

Gene	**Reason**	**Method**
Rbm3	Highest upregulated protein, known to be involved in temperature shift response and regulates splicing and translation but hasn't been utilised as an engineering target.	Overexpression of *Rbm3* may enhance productivity. Can be achieved by transfecting cells with an expression vector such as pcDNA3.1(+).
Cdkn1a	Highly upregulated cell cycle inhibitor. May be regulated by *Cirbp* and therefore may regulate cold induced halting of cell cycle.	Induce overexpression of *Cdkn1a* in a temperature sensitive expression vector to induce further overexpression during temperature shift (Such as Thaisuchat 2011).
Ldhc	Highly upregulated in temperature shift. Regulates lactate production.	Induce overexpression of Ldhc in temperature sensitive expression vector.
Pfkm	Downregulated in temperature shift. Induces Warburg metabolism.	Design siRNA in temperature sensitive expression vector to knockdown *Pfkm* during temperature shift.
Prkag1	Downregulated in temperature shift. Induces Warburg metabolism.	Design siRNA in temperature sensitive expression vector to knockdown *Prkag1* during temperature shift.
Cd44	Most spliced gene, differentially expressed at gene but not protein level. Regulates proliferation and glucose uptake.	Utilise CRISPR-Cas9 to knockout endogenous *Cd44* gene and introduce cDNA of full length spliced *Cd44* transcript.
Ybx3	Cold shock protein which regulates transcription and translation of cell cycle genes. Was identified as alternatively spliced but not DE at gene level.	Overexpress cDNA of *Ybx3* isoforms both with and without exon 7.
Dnm1l	One of the most significantly spliced genes. Regulator of mitochondrial fission which may impact cells capacity to carry out OXPHOS. Not identified as DE at gene level.	Overexpress cDNA of full length Dnm1l in temperature sensitive promoter containing expression vector.

Mff	Dnm1l partner which was also not identified as DE at gene level. Splicing may regulate rates of mitochondrial fission and therefore contribute to switch from Warburg to OXPHOS.	Overexpress cDNA of isoform containing exon 9 in temperature specific expression vector.
Pparg1	Alternatively spliced to increase proportion of transcripts with interupted amino acid sequence. May result in increased Pdk2 activity to result in increased fuel for OXPHOS and less lactate production.	Design siRNA to knock down *Pparg1* and express in temperature sensitive expression vector.
miR-290a-5p	Identified for the first time in CHO cells. Regulates the cell cycle and supresses apoptosis.	Overexpress miR-290a-5p in temperature sensitive expression vector
miR-295-5p	Identified for the first time in CHO cells. Regulates the cell cycle and supresses apoptosis.	Overexpress miR-295-5p in temperature sensitive expression vector
miR-9	Downregulated in temperature shift. Induces proliferation.	Overexpress miR-9 sponge/decoy RNA in temperature specific expression vector to downregulate miR-9 activity.
miR-199a	Downregulated in temperature shift. Inducer of Warburg metabolism.	Express miR-199a sponge/decoy RNA in temperature shift specific expression vector to downregulate miR-199a activity.
miR-103a-5p	Upregulated in temperature shift. Regulates switch from Warburg metabolism to OXPHOS and inhibits the cell cycle.	Overexpress miR-103a-5p in temperature shift specific expression vector.
Cdkn2c	Increased translation efficiency in temperature shift. Halts the cell cycle.	Overexpress *Cdkn2c* in temperature shift specific expression vector.
Fbxw7	Increased translation efficiency in temperature shift. Halts the cell cycle.	Overexpress Fbxw7 in temperature shift specific expression vector.
Gdf11	Increased translation efficiency in temperature shift. Halts the cell cycle.	Overexpress *Gdf11* or knockdown of miR-615-3p/miR-9 using miR sponge in temperature shift specific expression vector.
Btg2	Increased translation efficiency in temperature shift. Halts the cell cycle.	Overexpress *Btg2* in temperature shift specific expression vector, utilise isoform specific siRNA to knockdown transcript isoform with intron retention, or knockdown miR-9 in temperature shift specific expression vector.
Apopt1	Increased translation efficiency in temperature shift. Supressor of apoptosis.	Knockout *Apopt1* with CRISPR-Cas9 or upregulate miR-34a-5p.
Sh3glb1	Increased translation efficiency in temperature shift. Supressor of apoptosis.	Knockout *Sh3glb1* with CRISPR-Cas9 or upregulate miR-7-5p.
Timm8a	Increased translation efficiency in temperature shift. May increase capacity for OXPHOS.	Overexpress *Timm8a* or knockdown miR-30a-5p.
Mrpl54	Increased translation efficiency in temperature shift. May increase capacity for OXPHOS.	Overexpress *Mrpl54* or knockdown miR-615-3p.

6.1.3 Challenges of multi-omics integration

6.1.3.1 Challenges and potential improvements to this experiment

One of the biggest challenges when carrying out bioinformatics analyses is the lack of established analysis pipelines and the overwhelming amount of software, thresholds and parameters which can be selected from to carry out analyses. The problem of having undefined pipelines as well as the large amounts of data which must be stored, processed and analysed are compounded when a multi-omics experiment is carried out due to the need to handle each data type in a different way. Multi-omics software capable of handling multiple datatypes have been developed which utilise machine learning approaches to predict phenotypic effects based on multiple layers of genomics and transcriptomics measurements (Huang et al., 2017) however these algorithms are focused on Human cancer genetics and do not have support for Chinese hamster ovary cells. Our current understanding of CHO cell molecular biology was also a challenge which had to be overcome as we lack a complete understanding of ncRNAs, a comprehensively annotated genome reference and the phenotypic effects of alternative splice variants. As we understand these levels of CHO cell gene expression in the future in greater detail, it may be possible to move towards a complete systems understanding of CHO cells and predict phenotypic outcome based on ‘omics experiments to optimise production cultures.

Integration of multiple-omics techniques has begun to be applied to CHO cells and have discovered a lack of correlation between mRNA expression and proteomics (Clarke et al., 2012; Orellana et al., 2015) and to utilise transcriptomics, proteomics and metabolomics to build a genome scale metabolic model (Hefzi et al., 2016). Existing multi-omics integration tools are however not currently capable of integrating the different measurements taken in this experiment and estimating phenotypic outcome or enabling a statistical analysis of these data types – in particular, the effects of alternative splicing, miRNA targeting and differential translation efficiency. Due to this limitation, pipelines had to be developed to analyse each dataset in isolation and the results of each experiment integrated post-processing. Such an approach has also been utilised in the only other CHO cell multi-omics experiment in which translatomics measurements were incorporated (Ang et al., 2016).

While this was the most comprehensive experiment carried out to date in terms of the number of ‘omics techniques applied in parallel to CHO cells, there are gaps in this experiment which could be filled if repeated again to gain a more complete overview of CHO cells undergoing temperature shift. For example, ChiP-Seq or Methyl-Seq (Feichtinger et al., 2016) could be applied to further our understanding of chromatin state changes and transcription factor activity, phosphoproteomics could enhance our understanding of how the transcriptomics and translatomics alterations observed are carried forward in terms of protein activity (Henry et al., 2017), glycomics analysis could reveal any alterations to product quality (Zhang et al., 2016) and metabolomics approaches would enable observation of alterations to energy production and cellular metabolism (Hefzi et al., 2016). Expanding this experiment to multiple time points after temperature shift would also provide a greater understanding of how the alterations to miRNA expression and alternative splicing affect translation and protein levels over time (Hsu et al., 2017). Finally, the metabolic profile of the culture media is not comprehensive and expanding this culture profile to contain additional measurements such as amino acid concentrations could provide greater understanding of how amino acid availability has impacted differential translation.

The techniques used in this experiment to generate multi-omics datasets each have different levels of coverage and a dynamic range at which it is possible to detect changes in expression. For example, just 350 proteins were detected at a level of confidence with 2 peptides while 12,291 genes were detected as expressed in RNA-Seq. Integrations with the proteomics dataset were therefore limited to the common genes identified between the techniques. If a proteomics instrument providing a higher coverage was utilised such as a modern Orbitrap Fusion Lumo Tribrid Mass Spectrometer (ThermoFisher Scientific), it may have been possible to identify more cases of gene and protein expression which did not overlap and therefore indicating the presence of post-transcriptional regulation. Different thresholds for differential expression also had to be set due to the differences in dynamic range and accuracy of the techniques, such as the 1.5 fold change in differential gene expression and translation, and the 1.2 fold change used for differential protein expression.

6.1.3.2 miRNA targeting, alternative splicing and differential translation efficiency do not fully explain the lack of correlation between gene and protein levels of expression

Of the 153 genes identified in Chapter 2 as undergoing post-transcriptional regulation of gene expression, 66 were identified as alternative spliced in Chapter 3, with anti-correlated miRNA expression in Chapter 4 or differentially translated in the previous section. The 87 genes with non-correlated gene and protein level measurements which could not be explained using the techniques applied throughout this thesis are presented in Appendix 5J. Among these genes are those with roles in splicing, transcription, translation, energy metabolism, the cell cycle and proteasomal degradation (Table 6.2).

Table 6.2 – Uncorrelated differential gene and protein expression in genes with no alternative splicing, targets of differentially expressed miRNAs or differential translation efficiency. Of the 153 genes with uncorrelated gene and protein expression identified in Chapter 2, 87 were not identified as alternatively spliced, as targets of DE miRNAs. Genes with a role in splicing, transcription, translation, energy metabolism, the cell cycle and proteasomal degradation are presented in this table.

Gene ID	**Gene name**	**Gene name**	**Gene FC**	**Gene BH Adj. P**	**Protein FC**	**Protein P value**
ENSCGRG00000003468	*Sdha*	Succinate dehydrogenase complex, subunit A	1.09	7.96×10^{-04}	1.22	1.88×10^{-03}
ENSCGRG00000003825	*Eftud2*	Elongation Factor Tu GTP Binding Domain Containing 2	-1.12	4.33×10^{-04}	1.28	7.22×10^{-03}
ENSCGRG00000005590	*Srsf1*	Serine and Arginine Rich Splicing Factor 1	-1.35	1.04×10^{-13}	1.20	1.05×10^{-02}
ENSCGRG00000007210	*Psmd4*	Proteasome 26S Subunit, Non-ATPase 4	1.23	4.01×10^{-12}	1.46	1.27×10^{-03}
ENSCGRG00000006662	*Eif5a*	Eukaryotic Translation Initiation Factor 5A	1.28	5.29×10^{-11}	1.33	5.21×10^{-03}
ENSCGRG00000008227	*Eef1a1*	Eukaryotic Translation Elongation Factor 1 Alpha 1	1.15	5.50×10^{-05}	1.35	1.55×10^{-02}
ENSCGRG00000009521	*Atp5c1*	ATP Synthase F1 Subunit Gamma	1.02	3.82×10^{-01}	1.33	2.32×10^{-02}
ENSCGRG00000012678	*Psmb1*	Proteasome Subunit Beta 1	1.14	4.86×10^{-04}	1.28	3.84×10^{-03}
ENSCGRG00000014336	*Pa2g4*	Proliferation-Associated 2G4	1.16	4.95×10^{-09}	1.28	6.72×10^{-03}

The number of genes uncorrelated at the protein and gene levels which were identified as differentially translated was considerably lower than expected – with just 3 genes identified. If the 87 unexplained genes were undergoing post-transcriptional control of gene regulation by mRNA degradation or alteration of translation efficiency due to translation activators and suppressors then these genes should have been identified as differentially translated using RiboSeq. Technical reasons aside, the only mechanisms through which this could occur is due to differences in the turnover of mRNAs and proteins. The average half-life of an mRNA in a mammalian cell is approximately 7 hours (Sharova et al., 2009) while the average half-life of protein is 43 hours (Toyama and Hetzer, 2013). The disconnect between gene and protein measurements may therefore be in part due to the time it takes for levels of protein to mirror changes to gene expression (Haider and Pal, 2013). In addition, proteins are also differentially degraded post-translation by the ubiquitin-proteasome pathway (Zheng and Shabek, 2017) and autophagy (Dikic, 2017). Components of the proteasome were among the genes which were not correlated at the gene and protein level which could not be explained by alternative splicing, miRNA targeting or differential translation including Proteasome subunit beta 1 (*Psmb1*) and Proteasome 26S Subunit, Non-ATPase 4 (*Psmd4*). Both of these genes were not differentially expressed but were upregulated at the protein level with a 1.28 fold and 1.46 fold change respectively.

6.2 Future Work

6.2.1 Validation of differential gene expression and translation

Differential gene expression was identified in regulators of the cell cycle, apoptosis and metabolic switching. Validation was carried out for a selection of genes undergoing alternative splicing in Chapter 3 however no validation has been performed on results of Chapter 2, 4 and 5. It would be interesting to measure the change in gene expression induced by temperature shift in regulators of the cell cycle including *Ccne1*, *Ccne2*, *Mcm2*, *Lsm11*, *E2f1* and *Cdkn1a* using an alternative technique such as qPCR.

In Chapter 4 251 previously unannotated miRNA loci were identified in the CHO cell genome. 119 of these have not been previously reported in any other species and therefore require validation. As this was the first analysis of differential translation in CHO cells using RiboSeq and the first to study the temperature shift response on a global scale it would be interesting to confirm results using an alternative technique such as polysome profiling.

6.2.2 Validation of miRNA targets

In Chapter 4 miRNA target prediction was used to identified anti-correlated expression of genes and miRNAs for 491 gene:miRNA pairs and in Chapter 5 88 genes with differential translation were identified with anti-correlated differential miRNA expression. It would be interesting to validate these targets by knocking down or upregulating miRNAs predicted to be differentially expressed and measuring the expression of its predicted targets using qPCR. Knockdown of miRNAs could be achieved using antimiRs, miRNA sponges or genetic knockout while upregulation could be achieved by transiently transfecting cells with a synthetic miRNA mimic (Stenvang et al., 2012).

6.2.3 Engineering of CHO cells with enhanced phenotypes using panel of gene targets

Before proceeding with attempting to genetically engineer enhanced phenotypes using the panel of genes presented in Table 6.1, the temperature shift induced expression of these targets should first be validated utilising alternative methodologies. qPCR could be utilised to confirm the overexpression of *Rbm3, Cdkn1a, Ldhc,* and miR-103a-5p, downregulation of *Pfkm, Prkag1,* miR-9 and miR-199a as well as the expression of miR-290a-5p and miR-295-5p (Lin et al., 2015). qPCR could also be utilised to confirm the expression of specific transcript isoforms of *Cd44*, *Ybx3*, *Dnm1l, Mff* and *Pparg1* by designing primers crossing exon:exon junctions specific to a given transcript isoform as was carried out in Chapter 3 to validate alternative splicing in CHO cells (Section 3.3.4). To confirm protein level quantifications of Rbm3 and Cd44 achieved using LC MS/MS, Western blotting (Meleady et al., 2008) or Multiple Reaction Monitoring (Albrecht et al., 2018) could be applied as alternative methods of assessing differential expression of these proteins in temperature shifted cells. Finally, differential translation of *Cdkn2c*, *Fbxw7*, *Gdf11*, *Btg2*, *Apopt1*, *Sh3glb1*, *Timm8a* and *Mrpl54* in temperature shifted CHO cells could be confirmed by utilising polysome profiling (Courtes et al., 2013).

In order to engineer cell lines which may produce improved recombinant protein yields utilising the engineering targets identified throughout this thesis, varied techniques will need to be implemented depending on the expression pattern required. Genes such as *Rbm3, Timm8a,* and *Mrpl54* could be overexpressed constitutively by transfecting cells with the gene of interest in a cytomegalovirus promoter containing expression vector such as pcDNA3.1(+) (Invitrogen) as was utilised for overexpressing *Cirbp* (Tan et al., 2008). In addition, supressing the translation of *Apopt1* and *Sh3glb1* could be achieved by overexpressing miR-34a-5p and miR-7-5p in a constitutive expression vector to reduce apoptosis. Alternatively these apoptosis inducing genes could be knocked out utilising CRISPR/Cas9 with guide RNAs designed to disrupt the open reading frame of *Apot1* or *Sh3glb1* (Ronda et al., 2014). The other engineering targets however will require more complex engineering strategies to drive their expression during only the stationary phase of biphasic culture as their overexpression would diminish a cells ability to proliferate quickly during the exponential growth phase. In order to achieve this the genes will need to be expressed in an inducible expression vector such as utilising a tetracycline (Zhao et al., 2012), or cumate (Poulain et al., 2017) inducible promoter and treating cells with tetracycline or cumate once optimal cell density has been reached. Alternatively an expression vector could be constructed containing the temperature sensitive S100a6 promoter to induce expression only under mild hypothermia conditions (Thaisuchat et al., 2011). Utilising one of these expression systems, *Cdkn1a, Ldhc,* miR-290a-5p, miR-295-

5p, miR-103a-5p, *Cdkn2c, Fbxw7, Gdf11* and *Btg2* gene expression could be induced at the onset of the stationary phase to have their anti-proliferation, anti-apoptosis and metabolic switching effects without impacting proliferation and metabolism in the exponential phase.

Inducible expression vectors can also be utilised to selectively knock down gene expression during the stationary phase by designing inducible expression vectors containing siRNAs to target the gene of interest (Valdés-Bango Curell and Barron, 2018). The *Pparg1*, *Pfkm* and *Prkag1* genes could be knocked down using this approach to induce a switch in metabolism at the stationary phase and reduce lactate production further. A similar approach can be utilised to knock down the activity of miRNAs by expressing miRNA sponge decoy vectors containing target sites for a given miRNA to sequester endogenous miRNAs and prevent them from acting on their targets (Sanchez et al., 2014). Induced expression of miRNA sponges at the stationary phase could be utilised to sequester miR-9 and miR-199a to prevent proliferation and induce a switch from a Warburg metabolism to OXPHOS respectively. As targets of miR-9, this approach could also result in increased translation of *Btg2* and *Gdf11* and may also be utilised to sequester miR-615-3p to enhance *Gdf11* translation further.

To utilise the findings of Chapter 3 and engineer CHO cells with isoform specific expression in the *Cd44*, *Ybx3*, *Dnm1l* and *Mff* genes a combination of these approaches will need to be utilised. First, the endogenous gene would need to be knocked out by utilising CRISPR-Cas9 (Ronda et al., 2014) and then the isoform of interest reverse transcribed to cDNA which is transfected in an inducible expression vector. Utilising this approach it would be possible to engineer full length *Cd44* and *Dnm1l* to prevent expression of alternatively spliced isoforms which may impact the cell cycle and energy metabolism. Alternatively it may be possible to engineer artificial splicing factors to regulate the splicing of transcript isoforms as has been previously utilised to regulate the splicing of *Bcl-x* between the pro and anti-apoptosis isoforms (Wei et al., 2017). Artificial splice factors have the potential to control the inclusion or exclusion of particular exons and could be utilised to assess the phenotypic impact of upregulating isoforms of *Ybx3* containing exon 7 or *Mff* containing exon 9.

6.2.4 Identification of lncRNAs in total RNA

The RNA-Seq data generated and used in Chapters 2 and 3 was produced using rRNA-depletion rather than polyA-selection and therefore contains total RNA rather than just mRNA transcripts. In addition to the novel transcript isoforms assembled in Chapter 3, it would be possible to identify ncRNAs such as lncRNAs. lncRNAs are non-protein coding RNAs over 200nt in length which can silence gene expression by inhibiting transcription or regulate splicing by binding pre-mRNAs and blocking the splicing machinery (Yoon et al., 2013). Software packages are available for the identification of lncRNAs in RNA-Seq data such as *UClncR* (Sun et al., 2017).

A class of lncRNAs which has been discovered to be biologically functional in mammals relatively recently but have yet to be characterised in CHO cells is circRNAs (Barrett and Salzman, 2016). circRNAs are generated during the removal of introns from an mRNA transcript in back-splicing events to generate a circularised transcript with a splice donor and acceptor site bound. circRNAs may function as miRNA sponges and sequester miRNAs to regulate miRNA activity of their mRNA targets (Toit, 2013). Algorithms such as *CIRI* can be utilised to detect circRNA expression in RNA-Seq data (Gao et al., 2015). Identification of this class of lncRNAs would be particularly interesting as the circRNA bioprocessing gene ILF3 was identified as undergoing alternative splicing in Chapter 3 and therefore circRNA expression may be involved in the temperature shift response (Lin and Chen, 2018).

6.2.5 Identification of other classes of ncRNAs in small RNA-Seq data

Chapter 4 focused on the identification, differential expression and target prediction of miRNAs however other forms of small RNA can be detected in RNA-Seq data. piwiRNAs are small RNAs which repress the transcription of transposable elements in the genome to prevent replication and insertion of transposable elements which can result in mutations (Ishizu et al., 2012). piwiRNAs have also been shown to regulate gene expression in mRNAs post-transcriptionally using similar mechanisms to miRNAs (Watanabe and Lin, 2014) and may therefore contribute to the lack of correlation between observed mRNA and protein levels in temperature shifted CHO cells. Software such as *ShortStack* can be utilised to identify piwiRNAs and other small RNAs such as endogenous siRNAs in RNA-Seq datasets (Axtell, 2013b).

6.2.6 Identification of differential translation control in RiboSeq data

In Chapter 5 we identified differential translation efficiency of mRNAs. miRNA interactions were identified which explains the differential translation efficiency of 88 genes. Alternative splicing may contribute to the post-transcriptional regulation of a further 66 of the differentially translated genes however splicing or miRNA targeting was not found in 197 of the identified genes undergoing temperature shift induced differential translation. Other mechanisms must therefore be impacting the translation of these genes. Translational ambiguities can be identified in RiboSeq data such as ribosomal stall sites and stop codon read through using software such as *RiboTools* (Legendre et al., 2015) and *RiboProfiling* (Popa et al., 2016). Small open reading frames (smORFs) can also be identified in RiboSeq data which are small peptides which have been observed to have roles in apoptosis and metabolism (Saghatelian and Couso, 2015). smORFs can be present upstream of translation start sites of protein coding ORFs in mRNAs and can regulate translation of the mRNA by preventing ribosome assembly or inducing stalling (Saghatelian and Couso, 2015). smORFs can be identified by applying algorithms such as *uORF-seqr* (Spealman et al., 2018).

6.2.7 Investigation of the impact of alternative splicing in the 3' UTR on miRNA targeting

In Chapter 4 alternative splicing was identified in the 3'UTR of 43 genes with anti-correlated miRNA targets. In the *Hnrnpa2b1* gene the miRNA binding site of miR-103a-3p was predicted to be within a skipped exon which was spliced out at a higher rate in temperature shifted cells and therefore a higher percentage of the expressed transcripts would not be translationally repressed. The impact of the alternative splicing in the other 42 genes may also affect miRNA targeting as one of the properties effecting the efficiency of miRNA regulation is the proximity of a binding site to the stop codon (Agarwal et al., 2015). Alternative splicing in the 3'UTR may alter the distance between the target site and the stop codon and therefore impact miRNA activity. Investigation of the impact of 3'UTR splicing on miRNA targeting efficiency could be investigated by expressing individual alternatively spliced isoforms in cells alongside a predicted miRNA target and measuring the resulting translation efficiency.

Bibliography

Abrego, J., Gunda, V., Vernucci, E., Shukla, S.K., King, R.J., Dasgupta, A., Goode, G., Murthy, D., Yu, F., and Singh, P.K. (2017). GOT1-Mediated Anaplerotic Glutamine Metabolism Regulates Chronic Acidosis Stress In Pancreatic Cancer Cells. Cancer Lett *400*, 37–46.

Acin-Perez, R., Salazar, E., Kamenetsky, M., Buck, J., Levin, L.R., and Manfredi, G. (2009). Cyclic AMP produced inside mitochondria regulates oxidative phosphorylation. Cell Metab *9*, 265–276.

Adams, J., Chen, Z.-P., Van Denderen, B.J.W., Morton, C.J., Parker, M.W., Witters, L.A., Stapleton, D., and Kemp, B.E. (2004). Intrasteric control of AMPK via the gamma1 subunit AMP allosteric regulatory site. Protein Sci. *13*, 155–165.

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. ELife Sciences *4*, e05005.

Airenne, K.J., Hu, Y.-C., Kost, T.A., Smith, R.H., Kotin, R.M., Ono, C., Matsuura, Y., Wang, S., and Ylä-Herttuala, S. (2013). Baculovirus: an insect-derived vector for diverse gene transfer applications. Mol. Ther. *21*, 739–749.

Aitken, C.E., and Lorsch, J.R. (2012). A mechanistic overview of translation initiation in eukaryotes. Nature Structural & Molecular Biology *19*, 568–576.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene annotation system. Database (Oxford) *2016*.

Alarcón, C.R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S., and Tavazoie, S.F. (2015). HNRNPA2B1 is a mediator of m6A-dependent nuclear RNA processing events. Cell *162*, 1299–1308.

Albrecht, S., Kaisermayer, C., Reinhart, D., Ambrose, M., Kunert, R., Lindeberg, A., and Bones, J. (2018). Multiple reaction monitoring targeted LC-MS analysis of potential cell death marker proteins for increased bioprocess control. Anal Bioanal Chem *410*, 3197–3207.

Al-Fageeh, M.B., and Smales, C.M. (2009). Cold-inducible RNA binding protein (CIRP) expression is modulated by alternative mRNAs. RNA *15*, 1164–1176.

Alvarez-Garcia, I., and Miska, E.A. (2005). MicroRNA functions in animal development and human disease. Development *132*, 4653–4662.

An, J., Lai, J., Lehman, M.L., and Nelson, C.C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res *41*, 727–737.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*, R106.

Anders, G., Mackowiak, S.D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012a). doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res. *40*, D180-186.

Anders, S., Reyes, A., and Huber, W. (2012b). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics btu638.

Andreev, D.E., O’Connor, P.B.F., Loughran, G., Dmitriev, S.E., Baranov, P.V., and Shatsky, I.N. (2017). Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. Nucleic Acids Res *45*, 513–526.

Andreou, A.Z., Harms, U., and Klostermeier, D. (2016). eIF4B stimulates eIF4A ATPase and unwinding activities by direct interaction through its 7-repeats region. RNA Biol *14*, 113–123.

Andreux, P.A., Williams, E.G., Koutnikova, H., Houtkooper, R.H., Champy, M.-F., Henry, H., Schoonjans, K., Williams, R.W., and Auwerx, J. (2012). Systems Genetics of Metabolism: The Use of the BXD Murine Reference Panel for Multiscalar Integration of Traits. Cell *150*, 1287–1299.

Andrews, S. (2010). FastQC: A Quality Control tool for High Throughput Sequence Data.

Ang, K.S., Kyriakopoulos, S., Li, W., and Lee, D.-Y. (2016). Multi-omics data driven analysis establishes reference codon biases for synthetic gene design in microbial and mammalian cells. Methods *102*, 26–35.

Aquino-Jarquin, G. (2017). Emerging Role of CRISPR/Cas9 Technology for MicroRNAs Editing in Cancer Research. Cancer Research *77*, 6812–6817.

Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. PNAS *100*, 3889–3894.

Artero-Castro, A., Castellvi, J., García, A., Hernández, J., Cajal, S.R. y, and LLeonart, M.E. (2011). Expression of the ribosomal proteins Rplp0, Rplp1, and Rplp2 in gynecologic tumors. Human Pathology *42*, 194–203.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat Genet *25*, 25–29.

Axtell, M.J. (2013a). Classification and Comparison of Small RNAs from Plants. Annu. Rev. Plant Biol. *64*, 137–159.

Axtell, M.J. (2013b). ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA *19*, 740–751.

Bachellerie, J.-P., Cavaillé, J., and Hüttenhofer, A. (2002). The expanding snoRNA world. Biochimie *84*, 775–790.

Baeshen, M.N., Al-Hejin, A.M., Bora, R.S., Ahmed, M.M.M., Ramadan, H.A.I., Saini, K.S., Baeshen, N.A., and Redwan, E.M. (2015). Production of Biopharmaceuticals in E. coli: Current Scenario and Future Perspectives. J. Microbiol. Biotechnol. *25*, 953–962.

Baik, J.Y., Lee, M.S., An, S.R., Yoon, S.K., Joo, E.J., Kim, Y.H., Park, H.W., and Lee, G.M. (2006). Initial transcriptome and proteome analyses of low culture temperature-induced expression in CHO cells producing erythropoietin. Biotechnol. Bioeng. *93*, 361–371.

Balbás, P., and Lorence, A. (2004). Recombinant Gene Expression: Reviews and Protocols (Springer Science & Business Media).

Banks, I.R., Zhang, Y., Wiggins, B.E., Heck, G.R., and Ivashuta, S. (2012). RNA decoys. Plant Signal Behav *7*, 1188–1193.

Barrett, S.P., and Salzman, J. (2016). Circular RNAs: analysis, expression and potential functions. Development *143*, 1838–1847.

Barron, N., Kumar, N., Sanchez, N., Doolan, P., Clarke, C., Meleady, P., O'Sullivan, F., and Clynes, M. (2011). Engineering CHO cell growth and recombinant protein productivity by overexpression of miR-7. J. Biotechnol. *151*, 204–211.

Baughman, J.M., Nilsson, R., Gohil, V.M., Arlow, D.H., Gauhar, Z., and Mootha, V.K. (2009). A Computational Screen for Regulators of Oxidative Phosphorylation Implicates SLIRP in Mitochondrial RNA Homeostasis. PLOS Genetics *5*, e1000590.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., et al. (2014).

Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. The EMBO Journal *33*, 981–993.

Beck, A., Wurch, T., Bailly, C., and Corvaia, N. (2010). Strategies and challenges for the next generation of therapeutic antibodies. Nat Rev Immunol *10*, 345–352.

Becker, J., Hackl, M., Rupp, O., Jakobi, T., Schneider, J., Szczepanowski, R., Bekel, T., Borth, N., Goesmann, A., Grillari, J., et al. (2011). Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. Journal of Biotechnology *156*, 227–235.

Bedoya-López, A., Estrada, K., Sanchez-Flores, A., Ramírez, O.T., Altamirano, C., Segovia, L., Miranda-Ríos, J., Trujillo-Roldán, M.A., and Valdez-Cruz, N.A. (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*, e0151529.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A *74*, 3171–3175.

Bertoli, C., Skotheim, J.M., and de Bruin, R.A.M. (2013). Control of cell cycle transcription during G1 and S phases. Nat Rev Mol Cell Biol *14*, 518–528.

Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. Nucleic Acids Res *36*, D149–D153.

Birzele, F., Schaub, J., Rust, W., Clemens, C., Baum, P., Kaufmann, H., Weith, A., Schulz, T.W., and Hildebrandt, T. (2010). Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucl. Acids Res. *38*, 3999–4010.

Boiko, A.D., Porteous, S., Razorenova, O.V., Krivokrysenko, V.I., Williams, B.R., and Gudkov, A.V. (2006). A systematic search for downstream mediators of tumor suppressor function of p53 reveals a major role of BTG2 in suppression of Ras-induced transformation. Genes Dev. *20*, 236–252.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bort, J.A.H., Stern, B., and Borth, N. (2010). CHO-K1 host cells adapted to growth in glutamine-free medium by FACS-assisted evolution. Biotechnology Journal *5*, 1090–1097.

Borth, N. (2014). Opening the black box: Chinese hamster ovary research goes genome scale. Pharmaceutical Bioprocessing *2*, 367–369.

Bougas, A., Vlachogiannis, I.A., Gatos, D., Arenz, S., and Dinos, G.P. (2017). Dual effect of chloramphenicol peptides on ribosome inhibition. Amino Acids *49*, 995–1004.

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where, and how of protein synthesis. Nat Rev Mol Cell Biol *16*, 651–664.

Bridge, K.S., Shah, K.M., Li, Y., Foxler, D.E., Wong, S.C.K., Miller, D.C., Davidson, K.M., Foster, J.G., Rose, R., Hodgkinson, M.R., et al. (2017). Argonaute Utilization for miRNA Silencing Is Determined by Phosphorylation-Dependent Recruitment of LIM-Domain-Containing Proteins. Cell Reports *20*, 173–187.

Brinkrolf, K., Rupp, O., Laux, H., Kollin, F., Ernst, W., Linke, B., Kofler, R., Romand, S., Hesse, F., Budach, W.E., et al. (2013). Chinese hamster genome sequenced from sorted chromosomes. Nat Biotech *31*, 694–695.

Brinks, V. (2013). Immunogenicity of biosimilar monoclonal antibodies. Generics and Biosimilars Initiative Journal *2*, 188–193.

Bryant, D.W., Priest, H.D., and Mockler, T.C. (2012). Detection and quantification of alternative splicing variants using RNA-seq. Methods Mol. Biol. *883*, 97–110.

Buchsteiner, M., Quek, L.-E., Gray, P., and Nielsen, L.K. (2018). Improving culture performance and antibody production in CHO cell culture processes by reducing the Warburg effect. Biotechnology and Bioengineering *0*.

Budzianowski, J. (2015). Tobacco against Ebola virus disease. Prz. Lek. *72*, 567–571.

Bumgarner, R. (2013). DNA microarrays: Types, Applications and their future. Curr Protoc Mol Biol *0 22*, Unit-22.1.

Bunik, V.I., Mkrtchyan, G., Grabarska, A., Oppermann, H., Daloso, D., Araujo, W.L., Juszczak, M., Rzeski, W., Bettendorff, L., Fernie, A.R., et al. (2016). Inhibition of mitochondrial 2-oxoglutarate dehydrogenase impairs viability of cancer cells in a cell-specific metabolism-dependent manner. Oncotarget *7*, 26400–26421.

Burri, L., Strahm, Y., Hawkins, C.J., Gentle, I.E., Puryer, M.A., Verhagen, A., Callus, B., Vaux, D., and Lithgow, T. (2005). Mature DIABLO/Smac is produced by the IMP protease complex on the mitochondrial inner membrane. Mol. Biol. Cell *16*, 2926–2933.

Caldon, C.E., and Musgrove, E.A. (2010). Distinct and redundant functions of cyclin E1 and cyclin E2 in development and cancer. Cell Div *5*, 2.

Caldon, C.E., Sergio, C.M., Schütte, J., Boersma, M.N., Sutherland, R.L., Carroll, J.S., and Musgrove, E.A. (2009). Estrogen Regulation of Cyclin E2 Requires Cyclin D1 but Not c-Myc. Mol Cell Biol *29*, 4623–4639.

Calviello, L., and Ohler, U. (2017). Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. Trends Genet. *33*, 728–744.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 165–170.

Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics *48*, 4.11.1-4.11.39.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res *18*, 188–196.

Catalanotto, C., Cogoni, C., and Zardo, G. (2016). MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. Int J Mol Sci *17*.

Cervera, L., Gutiérrez, S., Gòdia, F., and Segura, M.M. (2011). Optimization of HEK 293 cell growth by addition of non-animal derived components using design of experiments. BMC Proc *5*, P126.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. Cell *147*, 358–369.

Chassé, H., Boulben, S., Costache, V., Cormier, P., and Morales, J. (2017). Analysis of translation using polysome profiling. Nucleic Acids Res *45*, e15.

Chen, S., and Gao, G. (2017). MicroRNAs recruit eIF4E2 to repress translation of target mRNAs. Protein Cell *8*, 750–761.

Chen, C., Le, H., and Goudar, C.T. (2015). An automated RNA-Seq analysis pipeline to identify and visualize differentially expressed genes and pathways in CHO cells. Biotechnol. Prog. *31*, 1150–1162.

Chen, C., Le, H., and Goudar, C.T. (2017). Evaluation of two public genome references for chinese hamster ovary cells in the context of rna-seq based gene expression analysis. Biotechnology and Bioengineering *114*, 1603–1613.

Chen, C., Zhao, S., Karnad, A., and Freeman, J.W. (2018). The biology and role of CD44 in cancer progression: therapeutic implications. Journal of Hematology & Oncology *11*, 64.

Cheng, S., and Gallie, D.R. (2013). Eukaryotic initiation factor 4B and the poly(A)-binding protein bind eIF4G competitively. Translation (Austin) *1*.

Chenu, S., Grégoire, A., Malykh, Y., Visvikis, A., Monaco, L., Shaw, L., Schauer, R., Marc, A., and Goergen, J.-L. (2003). Reduction of CMP-N-acetylneuraminic acid hydroxylase activity in engineered Chinese hamster ovary cells using an antisense-RNA strategy. Biochim. Biophys. Acta *1622*, 133–144.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. *46*, D296–D302.

Chou, Y.-T., Lee, C.-C., Hsiao, S.-H., Lin, S.-E., Lin, S.-C., Chung, C.-H., Chung, C.-H., Kao, Y.-R., Wang, Y.-H., Chen, C.-T., et al. (2013). The emerging role of SOX2 in cell proliferation and survival and its crosstalk with oncogenic signaling in lung cancer. Stem Cells *31*, 2607–2619.

Chu, J., Cargnello, M., Topisirovic, I., and Pelletier, J. (2016). Translation Initiation Factors: Reprogramming Protein Synthesis in Cancer. Trends in Cell Biology *26*, 918–933.

Chun, S.Y., Rodriguez, C.M., Todd, P.K., and Mills, R.E. (2016). SPECtre: a spectral coherence--based classifier of actively translated transcripts from ribosome profiling sequence data. BMC Bioinformatics *17*, 482.

Chung, B.K.-S., Yusufi, F.N.K., Mariati, Yang, Y., and Lee, D.-Y. (2013). Enhanced expression of codon optimized interferon gamma in CHO cells. Journal of Biotechnology *167,* 326–333.

Chung, B.Y., Hardcastle, T.J., Jones, J.D., Irigoyen, N., Firth, A.E., Baulcombe, D.C., and Brierley, I. (2015). The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. RNA *21*, 1731–1745.

Cieply, B., and Carstens, R.P. (2015). Functional roles of alternative splicing factors in human disease. Wiley Interdiscip Rev RNA *6*, 311–326.

Clamer, M., Tebaldi, T., Lauria, F., Bernabo, P., Gomez-Biagi, R.F., Perenthaler, E., Gubert, D., Pasquardini, L., Guella, G., Groen, E.J.N., et al. (2017). Active ribosome profiling with RiboLace. BioRxiv 179671.

Clarke, C., Doolan, P., Barron, N., Meleady, P., O'Sullivan, F., Gammell, P., Melville, M., Leonard, M., and Clynes, M. (2011). Predicting cell-specific productivity from CHO gene expression. J. Biotechnol. *151,* 159–165.

Clarke, C., Henry, M., Doolan, P., Kelly, S., Aherne, S., Sanchez, N., Kelly, P., Kinsella, P., Breen, L., Madden, S.F., et al. (2012). Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. BMC Genomics *13*, 656.

Clincke, M.-F., Mölleryd, C., Samani, P.K., Lindskog, E., Fäldt, E., Walsh, K., and Chotteau, V. (2013). Very High Density of Chinese Hamster Ovary Cells in Perfusion by Alternating Tangential Flow or Tangential Flow Filtration in WAVE Bioreactor^{TM}—Part II: Applications for Antibody Production and Cryopreservation. Biotechnol Prog *29*, 768–777.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology *17*, 13.

Contreras, L., Ramirez, L., Du, J., Hurley, J.B., Satrústegui, J., and de la Villa, P. (2016). Deficient glucose and glutamine metabolism in Aralar/AGC1/Slc25a12 knockout mice contributes to altered visual function. Mol. Vis. *22*, 1198–1212.

Coomans de Brachène, A., and Demoulin, J.-B. (2016). FOXO transcription factors in cancer development and therapy. Cell. Mol. Life Sci. *73*, 1159–1172.

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the Complexity of Transcriptomes with RNA-Seq.

Courtes, F.C., Lin, J., Lim, H.L., Ng, S.W., Wong, N.S.C., Koh, G., Vardy, L., Yap, M.G.S., Loo, B., and Lee, D.-Y. (2013). Translatome analysis of CHO cells to identify key growth genes. Journal of Biotechnology *167*, 215–224.

Courtes, F.C., Vardy, L., Wong, N.S.C., Bardor, M., Yap, M.G.S., and Lee, D.-Y. (2014). Understanding translational control mechanisms of the mTOR pathway in CHO cells by polysome profiling. N Biotechnol *31*, 514–523.

Daëron, M., and Nimmerjahn, F. (2014). Fc Receptors (Springer).

Dang, Y., Schneider-Poetsch, T., Eyler, D.E., Jewett, J.C., Bhat, S., Rawal, V.H., Green, R., and Liu, J.O. (2011). Inhibition of eukaryotic translation elongation by the antitumor natural product Mycalamide B. RNA *17*, 1578–1588.

Daniell, H., Singh, N.D., Mason, H., and Streatfield, S.J. (2009). Plant-made vaccine antigens and biopharmaceuticals. Trends Plant Sci *14*, 669–679.

Dever, T.E., and Green, R. (2012). The Elongation, Termination, and Recycling Phases of Translation in Eukaryotes. Cold Spring Harb Perspect Biol *4*.

Diament, A., and Tuller, T. (2016). Estimation of ribosome profiling performance and reproducibility at various levels of resolution. Biology Direct *11*, 24.

Diendorfer, A.B., Hackl, M., Klanert, G., Jadhav, V., Reithofer, M., Stiefel, F., Hesse, F., Grillari, J., and Borth, N. (2015). Annotation of additional evolutionary conserved microRNAs in CHO cells from updated genomic data. Biotechnol Bioeng *112*, 1488–1493.

Dikic, I. (2017). Proteasomal and Autophagic Degradation Systems. Annu. Rev. Biochem. *86*, 193–224.

Dimartino, D., Colantoni, A., Ballarino, M., Martone, J., Mariani, D., Danner, J., Bruckmann, A., Meister, G., Morlando, M., and Bozzoni, I. (2018). The Long Non-coding RNA lnc-31 Interacts with Rock1 mRNA and Mediates Its YB-1-Dependent Translation. Cell Rep *23*, 733–740.

Ding, S.-W., and Voinnet, O. (2007). Antiviral Immunity Directed by Small RNAs. Cell *130*, 413–426.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Doménech, E., Maestre, C., Esteban-Martínez, L., Partida, D., Pascual, R., Fernández-Miranda, G., Seco, E., Campos-Olivas, R., Pérez, M., Megias, D., et al. (2015). AMPK and PFKFB3 mediate glycolysis and survival in response to mitophagy during mitotic arrest. Nat Cell Biol *17*, 1304–1316.

Doolan, P., Melville, M., Gammell, P., Sinacore, M., Meleady, P., McCarthy, K., Francullo, L., Leonard, M., Charlebois, T., and Clynes, M. (2008). Transcriptional profiling of gene expression changes in a PACE-transfected CHO DUKX cell line secreting high levels of rhBMP-2. Mol. Biotechnol. *39*, 187–199.

Druz, A., Son, Y., Betenbaugh, M., and Shiloach, J. (2013). Stable inhibition of mmu-miR-466h-5p improves apoptosis resistance and protein production in CHO cells. Metabolic Engineering *16*, 87–94.

Du, Z., Treiber, D., McCarter, J.D., Fomina-Yadlin, D., Saleem, R.A., McCoy, R.E., Zhang, Y., Tharmalingam, T., Leith, M., Follstad, B.D., et al. (2015). Use of a small molecule cell cycle inhibitor to control cell growth and improve specific productivity and product quality of recombinant proteins in CHO cell cultures. Biotechnology and Bioengineering *112*, 141–155.

Ducommun, S., Deak, M., Sumpton, D., Ford, R.J., Núñez Galindo, A., Kussmann, M., Viollet, B., Steinberg, G.R., Foretz, M., Dayon, L., et al. (2015). Motif affinity and mass spectrometry proteomic approach for the discovery of cellular AMPK targets:

Identification of mitochondrial fission factor as a new AMPK substrate. Cellular Signalling *27*, 978–988.

Dumont, J., Euwart, D., Mei, B., Estes, S., and Kshirsagar, R. (2016). Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. Crit Rev Biotechnol *36*, 1110–1122.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184–1191.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLOS Computational Biology *7*, e1002195.

el Azzouzi, H., Leptidis, S., Dirkx, E., Hoeks, J., van Bree, B., Brand, K., McClellan, E.A., Poels, E., Sluimer, J.C., van den Hoogenhof, M.M.G., et al. (2013). The Hypoxia-Inducible MicroRNA Cluster miR-199a∼214 Targets Myocardial PPARδ and Impairs Mitochondrial Fatty Acid Oxidation. Cell Metabolism *18*, 341–354.

Emmerling, V.V., Fischer, S., Stiefel, F., Holzmann, K., Handrick, R., Hesse, F., Hörer, M., Kochanek, S., and Otte, K. (2016). Temperature-sensitive miR-483 is a conserved regulator of recombinant protein and viral vector production in mammalian cells. Biotechnol. Bioeng. *113*, 830–841.

Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., The RGASP Consortium, Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Meth *10*, 1185–1191.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2004). MicroRNA targets in Drosophila. Genome Biol *5*, R1.

Ernst, W., Trummer, E., Mead, J., Bessant, C., Strelec, H., Katinger, H., and Hesse, F. (2006). Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. Biotechnol J *1*, 639–650.

Farrell, A., McLoughlin, N., Milne, J.J., Marison, I.W., and Bones, J. (2014). Application of Multi-Omics Techniques for Bioprocess Design and Optimization in Chinese Hamster Ovary Cells. J. Proteome Res. *13*, 3144–3159.

Fasold, M., Langenberger, D., Binder, H., Stadler, P.F., and Hoffmann, S. (2011). DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res *39*, W112–W117.

Favre, C., Zhdanov, A., Leahy, M., Papkovsky, D., and O'Connor, R. (2010). Mitochondrial pyrimidine nucleotide carrier (PNC1) regulates mitochondrial biogenesis and the invasive phenotype of cancer cells. Oncogene *29*, 3964–3976.

Feichtinger, J., Hernández, I., Fischer, C., Hanscho, M., Auer, N., Hackl, M., Jadhav, V., Baumann, M., Krempl, P.M., Schmidl, C., et al. (2016). Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. Biotechnol Bioeng *113*, 2241–2253.

Felekkis, K., Touvana, E., Stefanou, C., and Deltas, C. (2010). microRNAs: a newly described class of encoded molecules that play a role in health and disease. Hippokratia *14*, 236–240.

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., et al. (2015). A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. Mol Cell *60*, 816–827.

Fischer, S., Mathias, S., Schaz, S., Emmerling, V.V., Buck, T., Kleemann, M., Hackl, M., Grillari, J., Aschrafi, A., Handrick, R., et al. (2015). Enhanced protein production by microRNA-30 family in CHO cells is mediated by the modulation of the ubiquitin pathway. Journal of Biotechnology *212*, 32–43.

Fischer, S., Marquart, K.F., Pieper, L.A., Fieder, J., Gamer, M., Gorr, I., Schulz, P., and Bradl, H. (2017). miRNA engineering of CHO cells facilitates production of difficult-to-express proteins and increases success in cell line development. Biotechnol. Bioeng. *114*, 1495–1510.

Fogolín, M.B., Wagner, R., Etcheverrigaray, M., and Kratje, R. (2004). Impact of temperature reduction and expression of yeast pyruvate carboxylase on hGM-CSF-producing CHO cells. Journal of Biotechnology *109*, 179–191.

Fomina-Yadlin, D., Mujacic, M., Maggiora, K., Quesnell, G., Saleem, R., and McGrew, J.T. (2015). Transcriptome analysis of a CHO cell line expressing a recombinant therapeutic protein treated with inducers of protein expression. J. Biotechnol. *212*, 106–115.

Franck, N., Montembault, E., Romé, P., Pascal, A., Cremet, J.-Y., and Giet, R. (2011). CDK11p58 Is Required for Centriole Duplication and Plk4 Recruitment to Mitotic Centrosomes. PLoS One *6*.

Frazee, A.C., Pertea, G., Jaffe, A.E., Langmead, B., Salzberg, S.L., and Leek, J.T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotechnol *33*, 243–246.

Friedländer, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology *26*, 407–415.

Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. *40*, 37–52.

Friedman, R.C., Farh, K.K.-H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92–105.

Frohman, M.A. (2015). Role of mitochondrial lipids in guiding fission and fusion. J Mol Med (Berl) *93*, 263–269.

Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C., et al. (2009). The challenges of sequencing by synthesis. Nat. Biotechnol. *27*, 1013–1023.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics *8*, 69.

Gammell, P., Barron, N., Kumar, N., and Clynes, M. (2007). Initial identification of low temperature and culture stage induction of miRNA expression in suspension CHO-K1 cells. Journal of Biotechnology *130*, 213–218.

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biology *16*, 4.

Gebert, L.F.R., and MacRae, I.J. (2018). Regulation of microRNA function in animals. Nature Reviews Molecular Cell Biology 1.

Geigert, J. (2014). The Challenge of CMC Regulatory Compliance for Biopharmaceuticals (Springer Science & Business Media).

Gentner, B., and Naldini, L. (2012). Exploiting microRNA regulation for genetic engineering. Tissue Antigens *80*, 393–403.

Genuth, N.R., and Barna, M. (2018). The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. Molecular Cell *71*, 364–374.

Ghaderi, D., Taylor, R.E., Padler-Karavani, V., Diaz, S., and Varki, A. (2010). Implications of the presence of N-glycolylneuraminic acid in recombinant therapeutic glycoproteins. Nat Biotech *28*, 863–867.

Ghaderi, D., Zhang, M., Hurtado-Ziola, N., and Varki, A. (2012). Production platforms for biotherapeutic glycoproteins. Occurrence, impact, and challenges of non-human sialylation. Biotechnol. Genet. Eng. Rev. *28*, 147–175.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A. *108*, 1513–1518.

Gobet, C., and Naef, F. (2017). Ribosome profiling and dynamic regulation of translation in mammals. Current Opinion in Genetics & Development *43*, 120–127.

Godfrey, C.L., Mead, E.J., Daramola, O., Dunn, S., Hatton, D., Field, R., Pettman, G., and Smales, C.M. (2017). Polysome profiling of mAb producing CHO cell lines links translational control of cell proliferation and recombinant mRNA loading onto ribosomes with global and recombinant protein synthesis. Biotechnology Journal *12*, 1700177.

Gonzalez, C., Sims, J.S., Hornstein, N., Mela, A., Garcia, F., Lei, L., Gass, D.A., Amendolara, B., Bruce, J.N., Canoll, P., et al. (2014). Ribosome Profiling Reveals a Cell-Type-Specific Translational Landscape in Brain Tumors. J Neurosci *34*, 10924–10936.

Graber, T.E., Hébert-Seropian, S., Khoutorsky, A., David, A., Yewdell, J.W., Lacaille, J.-C., and Sossin, W.S. (2013). Reactivation of stalled polyribosomes in synaptic plasticity. Proc Natl Acad Sci U S A *110*, 16205–16210.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech *29*, 644–652.

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. *40*, 10073–10083.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res *34*, D140–D144.

Grosso, A.R., Gomes, A.Q., Barbosa-Morais, N.L., Caldeira, S., Thorne, N.P., Grech, G., von Lindern, M., and Carmo-Fonseca, M. (2008). Tissue-specific splicing factor gene expression signatures. Nucleic Acids Res *36*, 4823–4832.

Guimaraes, J.C., and Zavolan, M. (2016). Patterns of ribosomal protein expression specify normal and malignant human cells. Genome Biology *17*, 236.

Guo, L., and Lu, Z. (2010). The Fate of miRNA* Strand through Evolutionary Analysis: Implication for Degradation As Merely Carrier Strand or Potential Regulatory Molecule? PLOS ONE *5*, e11387.

Gupta, V., and Warner, J.R. (2014). Ribosome-omics of the human ribosome. RNA.

Gürel, G., Blaha, G., Moore, P.B., and Steitz, T.A. (2009). U2504 determines the species specificity of the A-site cleft antibiotics: the structures of tiamulin, homoharringtonine, and bruceantin bound to the ribosome. J. Mol. Biol. *389*, 146–156.

Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J.M., and Aransay, A.M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res *37*, W68–W76.

Hackenberg, M., Rodríguez-Ezpeleta, N., and Aransay, A.M. (2011). miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res *39*, W132–W138.

Hackl, M., Jakobi, T., Blom, J., Doppmeier, D., Brinkrolf, K., Szczepanowski, R., Bernhart, S.H., Siederdissen, C.H. zu, Bort, J.A.H., Wieser, M., 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*, 62–75.

Hackl, M., Jadhav, V., Jakobi, T., Rupp, O., Brinkrolf, K., Goesmann, A., Pühler, A., Noll, T., Borth, N., and Grillari, J. (2012). Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines. Journal of Biotechnology *158*, 151–155.

Haider, S., and Pal, R. (2013). Integrated Analysis of Transcriptomic and Proteomic Data. Curr Genomics *14*, 91–110.

Haines, K.M., Vande Burgt, N.H., Francica, J.R., Kaletsky, R.L., and Bates, P. (2012). Chinese hamster ovary cell lines selected for resistance to ebolavirus glycoprotein mediated infection are defective for NPC1 expression. Virology *432*, 20–28.

Hammond, S., Swanberg, J.C., Kaplarevic, M., and Lee, K.H. (2011). Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. BMC Genomics *12*, 67.

Han, G., Zhang, L., Ni, X., Chen, Z., Pan, X., Zhu, Q., Li, S., Wu, J., Huang, X., and Wang, X. (2018a). MicroRNA-873 Promotes Cell Proliferation, Migration, and Invasion by Directly Targeting TSLC1 in Hepatocellular Carcinoma. CPB *46*, 2261–2270.

Han, S., Bahmanyar, S., Zhang, P., Grishin, N., Oegema, K., Crooke, R., Graham, M., Reue, K., Dixon, J.E., and Goodman, J.M. (2012). Nuclear Envelope Phosphatase 1-Regulatory Subunit 1 (Formerly TMEM188) Is the Metazoan Spo7p Ortholog and Functions in the Lipin Activation Pathway. J. Biol. Chem. *287,* 3123–3137.

Han, S., Kim, D., Shivakumar, M., Lee, Y.-J., Garg, T., Miller, J.E., Kim, J.H., Kim, D., and Lee, Y. (2018b). The effects of alternative splicing on miRNA binding sites in bladder cancer. PLOS ONE *13,* e0190708.

Harreither, E., Hackl, M., Pichler, J., Shridhar, S., Auer, N., Łabaj, P.P., Scheideler, M., Karbiener, M., Grillari, J., Kreil, D.P., et al. (2015). Microarray profiling of preselected CHO host cell subclones identifies gene expression patterns associated with increased production capacity. Biotechnol J *10*, 1625–1638.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1–8.

Hefzi, H., Ang, K.S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C.A., Baycin-Hizal, D., Huang, Y., Ley, D., et al. (2016). A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary (CHO) Cell Metabolism. Cell Syst *3*, 434-443.e8.

Henry, M., Power, M., Kaushik, P., Coleman, O., Clynes, M., and Meleady, P. (2017). Differential Phosphoproteomic Analysis of Recombinant Chinese Hamster Ovary Cells Following Temperature Shift. J. Proteome Res. *16*, 2339–2358.

Hernandez, R. (2015). Modular Manufacturing Platforms for Biologics. BioPharm International *5*, 18–25.

Hsu, H.-H., Araki, M., Mochizuki, M., Hori, Y., Murata, M., Kahar, P., Yoshida, T., Hasunuma, T., and Kondo, A. (2017). A Systematic Approach to Time-series Metabolite Profiling and RNA-seq Analysis of Chinese Hamster Ovary Cell Culture. Scientific Reports *7*, 43518.

Huang, S., Chaudhary, K., and Garmire, L.X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet *8*.

Huang, Y.-M., Hu, W., Rustandi, E., Chang, K., Yusuf-Makagiansar, H., and Ryll, T. (2010). Maximizing productivity of CHO cell-based fed-batch culture using chemically defined media conditions and typical manufacturing equipment. Biotechnol Progress *26*, 1400–1410.

Hurtado-López, L.M., Fernández-Ramírez, F., Martínez-Peñafiel, E., Ruiz, J.D.C., and González, N.E.H. (2015). Molecular Analysis by Gene Expression of Mitochondrial

ATPase Subunits in Papillary Thyroid Cancer: Is ATP5E Transcript a Possible Early Tumor Marker? Med Sci Monit *21*, 1745–1751.

Hwang, H.C., and Clurman, B.E. (2005). Cyclin E in normal and neoplastic cell cycles. Oncogene *24*, 2776–2786.

Ibusuki, M., Yamamoto, Y., Shinriki, S., Ando, Y., and Iwase, H. (2011). Reduced expression of ubiquitin ligase FBXW7 mRNA is associated with poor prognosis in breast cancer patients. Cancer Sci. *102*, 439–445.

Ikonomou, L., Schneider, Y.-J., and Agathos, S.N. (2003). Insect cell culture for industrial production of recombinant proteins. Appl Microbiol Biotechnol *62*, 1–20.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science *324*, 218–223.

Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T., and Suzuki, A. (1998). The Molecular Basis for the Absence ofN-Glycolylneuraminic Acid in Humans. J. Biol. Chem. *273*, 15866–15871.

Ishizu, H., Siomi, H., and Siomi, M.C. (2012). Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes Dev. *26*, 2361–2373.

Iwakawa, H., and Tomari, Y. (2015). The Functions of MicroRNAs: mRNA Decay and Translational Repression. Trends in Cell Biology *25*, 651–665.

Jackson, R., and Standart, N. (2015). The awesome power of ribosome profiling. RNA *21*, 652–654.

Jacquemart, R., Vandersluis, M., Zhao, M., Sukhija, K., Sidhu, N., and Stout, J. (2016). A Single-use Strategy to Enable Manufacturing of Affordable Biologics. Comput Struct Biotechnol J *14*, 309–318.

Jadhav, V., Hackl, M., Bort, J.A.H., Wieser, M., Harreither, E., Kunert, R., Borth, N., and Grillari, J. (2012). A screening method to assess biological effects of microRNA overexpression in Chinese hamster ovary cells. Biotechnol. Bioeng. *109*, 1376–1385.

Jadhav, V., Hackl, M., Klanert, G., Hernandez Bort, J.A., Kunert, R., Grillari, J., and Borth, N. (2014). Stable overexpression of miR-17 enhances recombinant protein production of CHO cells. Journal of Biotechnology *175*, 38–44.

Jalili, A., Wagner, C., Pashenkov, M., Pathria, G., Mertz, K.D., Widlund, H.R., Lupien, M., Brunet, J.-P., Golub, T.R., Stingl, G., et al. (2012). Dual Suppression of the Cyclin-

Dependent Kinase Inhibitors CDKN2C and CDKN1A in Human Melanoma. J Natl Cancer Inst *104*, 1673–1679.

Janic, A., Valente, L.J., Wakefield, M.J., Stefano, L., Milla, L., Wilcox, S., Yang, H., Tai, L., Vandenberg, C.J., Kueh, A.J., et al. (2018). DNA repair processes are critical mediators of p53-dependent tumor suppression. Nature Medicine 1.

Jefferis, R. (2009). Glycosylation as a strategy to improve antibody-based therapeutics. Nat Rev Drug Discov *8*, 226–234.

Jiang, W., Zhu, J., Zhuang, X., Zhang, X., Luo, T., Esser, K.A., and Ren, H. (2015). Lipin1 Regulates Skeletal Muscle Differentiation through Extracellular Signal-regulated Kinase (ERK) Activation and Cyclin D Complex-regulated Cell Cycle Withdrawal. J. Biol. Chem. *290*, 23646–23655.

Joazeiro, C.A.P. (2017). Ribosomal Stalling During Translation: Providing Substrates for Ribosome-Associated Protein Quality Control. Annual Review of Cell and Developmental Biology *33*, 343–368.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). Human MicroRNA Targets. PLoS Biol *2*.

Johnson, K.C., Jacob, N.M., Nissom, P.M., Hackl, M., Lee, L.H., Yap, M., and Hu, W.-S. (2011). Conserved microRNAs in Chinese hamster ovary cell lines. Biotechnol. Bioeng. *108*, 475–480.

Kaas, C.S., Kristensen, C., Betenbaugh, M.J., and Andersen, M.R. (2015). Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy. BMC Genomics *16*, 160.

Kaikkonen, M.U., Lam, M.T.Y., and Glass, C.K. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res *90*, 430–440.

Kallehauge, T.B., Li, S., Pedersen, L.E., Ha, T.K., Ley, D., Andersen, M.R., Kildegaard, H.F., Lee, G.M., and Lewis, N.E. (2017). Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion. Sci Rep *7*.

Kamura, T., Hara, T., Matsumoto, M., Ishida, N., Okumura, F., Hatakeyama, S., Yoshida, M., Nakayama, K., and Nakayama, K.I. (2004). Cytoplasmic ubiquitin ligase KPC regulates proteolysis of p27(Kip1) at G1 phase. Nat. Cell Biol. *6*, 1229–1235.

Kang, J.H., Lee, S.-H., Hong, D., Lee, J.-S., Ahn, H.-S., Ahn, J.-H., Seong, T.W., Lee, C.-H., Jang, H., Hong, K.M., et al. (2016). Aldehyde dehydrogenase is used by cancer cells for energy metabolism. Exp Mol Med *48*, e272.

Kannicht, C., Ramström, M., Kohla, G., Tiemeyer, M., Casademunt, E., Walter, O., and Sandberg, H. (2013). Characterisation of the post-translational modifications of a novel, human cell line-derived recombinant human factor VIII. Thrombosis Research *131*, 78–88.

Kantardjieff, A., Nissom, P.M., Chuah, S.H., Yusufi, F., Jacob, N.M., Mulukutla, B.C., Yap, M., and Hu, W.-S. (2009). Developing genomic platforms for Chinese hamster ovary cells. Biotechnol. Adv. *27*, 1028–1035.

Kao, F.T., and Puck, T.T. (1968). Genetics of somatic mammalian cells, VII. Induction and isolation of nutritional mutants in Chinese hamster cells. Proc Natl Acad Sci U S A *60*, 1275–1281.

Kapustin, Y., Souvorov, A., Tatusova, T., and Lipman, D. (2008). Splign: algorithms for computing spliced alignments with identification of paralogs. Biology Direct *3*, 20.

Kearse, M.G., Chen, A.S., and Ware, V.C. (2011). Expression of ribosomal protein L22e family members in Drosophila melanogaster: rpL22-like is differentially expressed and alternatively spliced. Nucleic Acids Res *39*, 2701–2716.

Kelley, B. (2009). Industrialization of mAb production technology: the bioprocessing industry at a crossroads. MAbs *1*, 443–452.

Kelly, P.S., Breen, L., Gallagher, C., Kelly, S., Henry, M., Lao, N.T., Meleady, P., O’Gorman, D., Clynes, M., and Barron, N. (2015a). Re-programming CHO cell metabolism using miR-23 tips the balance towards a highly productive phenotype. Biotechnol J *10*, 1029–1040.

Kelly, P.S., Gallagher, C., Clynes, M., and Barron, N. (2015b). Conserved microRNA function as a basis for Chinese hamster ovary cell engineering. Biotechnol. Lett. *37*, 787–798.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nature Genetics *39*, 1278–1284.

Kim, B.J., Zhao, T., Young, L., Zhou, P., and Shuler, M.L. (2012a). Batch, fed-batch, and microcarrier cultures with CHO cell lines in a pressure-cycle driven miniaturized bioreactor. Biotechnology and Bioengineering *109*, 137–145.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology *14*, R36.

Kim, D., Langmead, B., and Salzberg, S.L. (2015a). HISAT: a fast spliced aligner with low memory requirements. Nat Meth *12*, 357–360.

Kim, H., Yoo, S.J., and Kang, H.A. (2015b). Yeast synthetic biology for the production of recombinant therapeutic proteins. FEMS Yeast Research *15*, 1–16.

Kim, J.Y., Kim, Y.-G., and Lee, G.M. (2012b). CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl. Microbiol. Biotechnol. *93*, 917–930.

Kimball, S.R., Horetsky, R.L., Ron, D., Jefferson, L.S., and Harding, H.P. (2003). Mammalian stress granules represent sites of accumulation of stalled translation initiation complexes. Am. J. Physiol., Cell Physiol. *284*, C273-284.

King, H.A., and Gerber, A.P. (2016). Translatome profiling: methods for genome-scale analysis of mRNA translation. Brief Funct Genomics *15*, 22–31.

King, Z.A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N.E., and Palsson, B.O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. PLOS Computational Biology *11*, e1004321.

Kingston, R.E., Kaufman, R.J., Bebbington, C. r., and Rolfe, M. r. (2001). Amplification Using CHO Cell Expression Vectors. In Current Protocols in Molecular Biology, (John Wiley & Sons, Inc.), p.

Klanert, G., Jadhav, V., Shanmukam, V., Diendorfer, A., Karbiener, M., Scheideler, M., Bort, J.H., Grillari, J., Hackl, M., and Borth, N. (2016). A signature of 12 microRNAs is robustly associated with growth rate in a variety of CHO cell lines. Journal of Biotechnology *235*, 150–161.

Kluza, J., Corazao-Rozas, P., Touil, Y., Jendoubi, M., Maire, C., Guerreschi, P., Jonneaux, A., Ballot, C., Balayssac, S., Valable, S., et al. (2012). Inactivation of the HIF-1α/PDK3 Signaling Axis Drives Melanoma toward Mitochondrial Oxidative Metabolism and Potentiates the Therapeutic Activity of Pro-Oxidants. Cancer Res *72*, 5035–5047.

Kolitz, S.E., and Lorsch, J.R. (2010). Eukaryotic Initiator tRNA: Finely Tuned and Ready for Action. FEBS Lett *584*, 396–404.

Könitzer, J.D., Müller, M.M., Leparc, G., Pauers, M., Bechmann, J., Schulz, P., Schaub, J., Enenkel, B., Hildebrandt, T., Hampel, M., et al. (2015). A global RNA-seq-driven analysis of CHO host and production cell lines reveals distinct differential expression patterns of genes contributing to recombinant antibody glycosylation. Biotechnology Journal *10*, 1412–1423.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucl. Acids Res. *42*, D68–D73.

Kraus, F., and Ryan, M.T. (2017). The constriction and scission machineries involved in mitochondrial fission. J. Cell. Sci. *130*, 2953–2960.

Krishnamoorthy, S., Liu, T., Drager, D., Patarroyo-White, S., Chhabra, E.S., Peters, R., Josephson, N., Lillicrap, D., Blumberg, R.S., Pierce, G.F., et al. (2016). Recombinant factor VIII Fc (rFVIIIFc) fusion protein reduces immunogenicity and induces tolerance in hemophilia A mice. Cell. Immunol. *301*, 30–39.

Krivega, M.V., Geens, M., Heindryckx, B., Santos-Ribeiro, S., Tournaye, H., and Van de Velde, H. (2015). Cyclin E1 plays a key role in balancing between totipotency and differentiation in human embryonic cells. Molecular Human Reproduction *21*, 942–956.

Kulawiak, B., Höpker, J., Gebert, M., Guiard, B., Wiedemann, N., and Gebert, N. (2013). The mitochondrial protein import machinery has multiple connections to the respiratory chain. Biochimica et Biophysica Acta (BBA) - Bioenergetics *1827*, 612–626.

Kumar, N., Gammell, P., and Clynes, M. (2007). Proliferation control strategies to improve productivity and survival during CHO based production culture. Cytotechnology *53*, 33–46.

Kumar, N., Gammell, P., Meleady, P., Henry, M., and Clynes, M. (2008). Differential protein expression following low temperature culture of suspension CHO-K1 cells. BMC Biotechnology *8*, 42.

Kuystermans, D., and Al-Rubeai, M. (2015). Mammalian Cell Line Selection Strategies for High-Producers. In Animal Cell Culture, M. Al-Rubeai, ed. (Springer International Publishing), pp. 327–372.

Kwang, T.W., Zeng, X., and Wang, S. (2016). Manufacturing of AcMNPV baculovirus vectors to enable gene therapy trials. Mol Ther Methods Clin Dev *3*, 15050.

Laere, E., Ling, A.P.K., Wong, Y.P., Koh, R.Y., Lila, M., Azmi, M., and Hussein, S. (2016). Plant-Based Vaccines: Production and Challenges.

Lagouge, M., Mourier, A., Lee, H.J., Spåhr, H., Wai, T., Kukat, C., Ramos, E.S., Motori, E., Busch, J.D., Siira, S., et al. (2015). SLIRP Regulates the Rate of Mitochondrial Protein Synthesis and Protects LRPPRC from Degradation. PLOS Genetics *11*, e1005423.

Lalonde, M.-E., and Durocher, Y. (2017). Therapeutic glycoprotein production in mammalian cells. Journal of Biotechnology *251*, 128–140.

Lam, J.K.W., Chow, M.Y.T., Zhang, Y., and Leung, S.W.S. (2015). siRNA Versus miRNA as Therapeutics for Gene Silencing. Molecular Therapy - Nucleic Acids *4*, e252.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.

Larsson, O., Sonenberg, N., and Nadon, R. (2010). Identification of differential translation in genome wide studies. PNAS *107*, 21487–21492.

Laux, H. (2014). Industry perspective on Chinese hamster ovary cell 'omics.' Pharmaceutical Bioprocessing *2*, 371-375377–375381.

Lawrence, S. (2005). Biotech drug market steadily expands. Nat Biotech *23*, 1466–1466.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.

Le, H., Chen, C., and Goudar, C.T. (2015). An evaluation of public genomic references for mapping RNA-Seq data from Chinese hamster ovary cells. Biotechnol. Bioeng. *112*, 2412–2416.

Le Fourn, V., Girod, P.-A., Buceta, M., Regamey, A., and Mermod, N. (2014). CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. Metabolic Engineering *21*, 91–102.

Lecarpentier, Y., Claes, V., Vallée, A., and Hébert, J.-L. (2017). Interactions between PPAR Gamma and the Canonical Wnt/Beta-Catenin Pathway in Type 2 Diabetes and Colon Cancer.

Ledergerber, C., and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. Brief Bioinform *12*, 489–497.

Lee, C., and Roy, M. (2004). Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol *5*, 231.

Lee, J.H., and Ko, K. (2017). Production of Recombinant Anti-Cancer Vaccines in Plants. Biomol Ther (Seoul) *25*, 345–353.

Lee, A.S.Y., Kranzusch, P.J., and Cate, J.H.D. (2015). eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. Nature *522*, 111–114.

Lee, H.-J., Palkovits, M., and Young, W.S. (2006). miR-7b, a microRNA up-regulated in the hypothalamus after chronic hyperosmolar stimulation, inhibits Fos translation. PNAS *103*, 15669–15674.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. PNAS *109*, 14728–14729.

Lee, Y.S., Shibata, Y., Malhotra, A., and Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. *23*, 2639–2649.

Legendre, R., Baudin-Baillieu, A., Hatin, I., and Namy, O. (2015). RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis. Bioinformatics *31*, 2586–2588.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods *7*, 709–715.

Lewis, B.P., Shih, I.-H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of Mammalian MicroRNA Targets. Cell *115*, 787–798.

Lewis, N.E., Liu, X., Li, Y., Nagarajan, H., Yerganian, G., O'Brien, E., Bordbar, A., Roth, A.M., Rosenbloom, J., Bian, C., et al. (2013). Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. Nat Biotech *31*, 759–765.

Li, D., Wei, T., Abbott, C.M., and Harrich, D. (2013). The Unexpected Roles of Eukaryotic Translation Elongation Factors in RNA Virus Replication and Pathogenesis. Microbiol. Mol. Biol. Rev. *77*, 253–266.

Li, F., Vijayasankaran, N., Shen, A. (Yijuan), Kiss, R., and Amanullah, A. (2010). Cell culture processes for monoclonal antibody production. MAbs *2*, 466–477.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J., Zhu, S., Tong, J., Hao, H., Yang, J., Liu, Z., and Wang, Y. (2016). Suppression of lactate dehydrogenase A compromises tumor progression by downregulation of the Warburg effect in glioblastoma. Neuroreport *27*, 110–115.

Li, M., Qian, Z., Ma, X., Lin, X., You, Y., Li, Y., Chen, T., and Jiang, H. (2018). MiR-628-5p decreases the tumorigenicity of epithelial ovarian cancer cells by targeting at FGFR2. Biochemical and Biophysical Research Communications *495*, 2085–2091.

Li, V.C., Ballabeni, A., and Kirschner, M.W. (2012). Gap 1 phase length and mouse embryonic stem cell self-renewal. Proc Natl Acad Sci U S A *109*, 12550–12555.

Li, X., Chen, W., Zeng, W., Wan, C., Duan, S., and Jiang, S. (2017). microRNA-137 promotes apoptosis in ovarian cancer cells via the regulation of XIAP. Br. J. Cancer *116*, 66–76.

Liao, Y., and Lönnerdal, B. (2010). Global MicroRNA Characterization Reveals That miR-103 Is Involved in IGF-1 Stimulated Mouse Intestinal Cell Proliferation. PLOS ONE *5*, e12976.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lima, W.R., Parreira, K.S., Devuyst, O., Caplanusi, A., N'Kuli, F., Marien, B., Van Der Smissen, P., Alves, P.M.S., Verroust, P., Christensen, E.I., et al. (2010). ZONAB Promotes Proliferation and Represses Differentiation of Proximal Tubule Epithelial Cells. J Am Soc Nephrol *21*, 478–488.

Lin, X., and Chen, Y. (2018). Identification of Potentially Functional CircRNA-miRNA-mRNA Regulatory Network in Hepatocellular Carcinoma by Integrated Microarray Analysis. Med Sci Monit Basic Res *24*, 70–78.

Lin, C.-C., Fang, C.-L., Sun, D.-P., Hseu, Y.-C., Uen, Y.-H., Lin, K.-Y., and Lin, Y.-C. (2017). High expression of mitochondrial intermembrane chaperone TIMM9 represents a negative prognostic marker in gastric cancer. Journal of the Formosan Medical Association *116*, 476–483.

Lin, N., Mascarenhas, J., Sealover, N.R., George, H.J., Brooks, J., Kayser, K.J., Gau, B., Yasa, I., Azadi, P., and Archer-Hartmann, S. (2015). Chinese Hamster Ovary (CHO) Host Cell Engineering to Increase Sialylation of Recombinant Therapeutic Proteins by Modulating Sialyltransferase Expression. Biotechnol Prog *31*, 334–346.

Lindquist, J.A., and Mertens, P.R. (2018). Cold shock proteins: from cellular mechanisms to pathophysiology and disease. Cell Commun Signal *16*.

Lindquist, J.A., Brandt, S., Bernhardt, A., Zhu, C., and Mertens, P.R. (2014). The role of cold shock domain proteins in inflammatory diseases. J Mol Med *92*, 207–216.

Linnarsson, S. (2010). Recent advances in DNA sequencing methods - general principles of sample preparation. Exp. Cell Res. *316*, 1339–1343.

Lintner, N.G., McClure, K.F., Petersen, D., Londregan, A.T., Piotrowski, D.W., Wei, L., Xiao, J., Bolt, M., Loria, P.M., Maguire, B., et al. (2017). Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. PLOS Biology *15*, e2001882.

Liu, D.-Z., Chang, B., Li, X.-D., Zhang, Q.-H., and Zou, Y.-H. (2017). MicroRNA-9 promotes the proliferation, migration, and invasion of breast cancer cells via down-regulating FOXO1. Clin Transl Oncol *19*, 1133–1140.

Liu, S., Bachran, C., Gupta, P., Miller-Randolph, S., Wang, H., Crown, D., Zhang, Y., Wein, A.N., Singh, R., Fattah, R., et al. (2012). Diphthamide modification on eukaryotic elongation factor 2 is needed to assure fidelity of mRNA translation and mouse development. PNAS *109*, 13817–13822.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell *165*, 535–550.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402–408.

Locasale, J.W., and Cantley, L.C. (2011). Metabolic Flux and the Regulation of Mammalian Cell Growth. Cell Metabolism *14*, 443–451.

Lomakin, I.B., and Steitz, T.A. (2013). The initiation of mammalian protein synthesis and the mechanism of scanning. Nature *500*, 307–311.

Loughran, G., Chou, M.-Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., Baranov, P.V., and Atkins, J.F. (2014). Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res *42*, 8928–8938.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Lu, C.-W., Lin, S.-C., Chen, K.-F., Lai, Y.-Y., and Tsai, S.-J. (2008). Induction of Pyruvate Dehydrogenase Kinase-3 by Hypoxia-inducible Factor-1 Promotes Metabolic Switch and Drug Resistance. J Biol Chem *283*, 28106–28114.

Luo, J., Yan, R., He, X., and He, J. (2018). SOX2 inhibits cell proliferation and metastasis, promotes apoptotic by downregulating CCND1 and PARP in gastric cancer. Am J Transl Res *10*, 639–647.

Lyabin, D.N., and Ovchinnikov, L.P. (2016). Selective regulation of YB-1 mRNA translation by the mTOR signaling pathway is not mediated by 4E-binding protein. Scientific Reports *6*, 22502.

Macovei, A., Gill, S.S., and Tuteja, N. (2012). microRNAs as promising tools for improving stress tolerance in rice. Plant Signaling & Behavior *7*, 1296–1301.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Letters *583*, 3966–3973.

Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H., and Dieterich, C. (2017). Bayesian prediction of RNA translation from ribosome profiling. Nucleic Acids Res *45*, 2960–2972.

Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., et al. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res *37*, W273–W276.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Martinez, G., Choudury, S.G., and Slotkin, R.K. (2017). tRNA-derived small RNAs target transposable element transcripts. Nucleic Acids Res. *45*, 5142–5152.

Martínez, V.S., Buchsteiner, M., Gray, P., Nielsen, L.K., and Quek, L.-E. (2015). Dynamic metabolic flux analysis using B-splines to study the effects of temperature shift on CHO cell metabolism. Metabolic Engineering Communications *2*, 46–57.

Martinez-Sanchez, A., and Murphy, C.L. (2013). MicroRNA Target Identification—Experimental Approaches. Biology (Basel) *2*, 189–205.

Masterton, R.J., and Smales, C.M. (2014). The impact of process temperature on mammalian cell lines and the implications for the production of recombinant proteins in CHO cells. Pharmaceutical Bioprocessing *2*, 49–61.

Masterton, R.J., Roobol, A., Al-Fageeh, M.B., Carden, M.J., and Smales, C.M. (2010). Post-translational events of a model reporter protein proceed with higher fidelity and accuracy upon mild hypothermic culturing of Chinese hamster ovary cells. Biotechnology and Bioengineering *105*, 215–220.

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat. Rev. Mol. Cell Biol. *8*, 209–220.

Mathelier, A., and Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics *26*, 2226–2234.

Meleady, P., Henry, M., Gammell, P., Doolan, P., Sinacore, M., Melville, M., Francullo, L., Leonard, M., Charlebois, T., and Clynes, M. (2008). Proteomic profiling of CHO cells with enhanced rhBMP-2 productivity following co-expression of PACEsol. Proteomics *8*, 2611–2624.

Meleady, P., Gallagher, M., Clarke, C., Henry, M., Sanchez, N., Barron, N., and Clynes, M. (2012). Impact of miR-7 over-expression on the proteome of Chinese hamster ovary cells. J. Biotechnol. *160*, 251–262.

Melville, M., Doolan, P., Mounts, W., Barron, N., Hann, L., Leonard, M., Clynes, M., and Charlebois, T. (2011). Development and characterization of a Chinese hamster ovary cell-specific oligonucleotide microarray. Biotechnol Lett *33*, 1773–1779.

Mende, N., Kuchen, E.E., Lesche, M., Grinenko, T., Kokkaliaris, K.D., Hanenberg, H., Lindemann, D., Dahl, A., Platz, A., Höfer, T., et al. (2015). CCND1-CDK4-mediated cell cycle progression provides a competitive advantage for human hematopoietic stem cells in vivo. J. Exp. Med. *212*, 1171–1183.

Merdzhanova, G., Edmond, V., Seranno, S.D., Broeck, A.V. den, Corcos, L., Brambilla, C., Brambilla, E., Gazzeri, S., and Eymin, B. (2008). E2F1 controls alternative splicing pattern of genes involved in apoptosis through upregulation of the splicing factor SC35. Cell Death and Differentiation *15*, 1815.

MGlincy, N.J., and Ingolia, N.T. (2017). Transcriptome-wide measurement of translation by ribosome profiling. Methods *126*, 112–129.

Michel, A.M., and Baranov, P.V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley Interdiscip Rev RNA *4*, 473–490.

Michel, A.M., Mullan, J.P.A., Velayudhan, V., O’Connor, P.B.F., Donohue, C.A., and Baranov, P.V. (2016). RiboGalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data. RNA Biology *13*, 316–319.

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res *29*, 2850–2859.

Mohorianu, I., Stocks, M.B., Applegate, C.S., Folkes, L., and Moulton, V. (2017). The UEA Small RNA Workbench: A Suite of Computational Tools for Small RNA Analysis. In MicroRNA Detection and Target Identification, (Humana Press, New York, NY), pp. 193–224.

Monger, C., Motheramgari, K., McSharry, J., Barron, N., and Clarke, C. (2017). A Bioinformatics Pipeline for the Identification of CHO Cell Differential Gene Expression from RNA-Seq Data. In Heterologous Protein Production in CHO Cells: Methods and Protocols, P. Meleady, ed. (New York, NY: Springer New York), pp. 169–186.

Moore, A., Mercer, J., Dutina, G., Donahue, C.J., Bauer, K.D., Mather, J.P., Etcheverry, T., and Ryll, T. (1997). Effects of temperature shift on cell cycle, apoptosis and nucleotide pools in CHO cell batch cultues. Cytotechnology *23*, 47–54.

Morf, J., Rey, G., Schneider, K., Stratmann, M., Fujita, J., Naef, F., and Schibler, U. (2012). Cold-Inducible RNA-Binding Protein Modulates Circadian Gene Expression Posttranscriptionally. Science *338*, 379–383.

Morita, M., Gravel, S.-P., Hulea, L., Larsson, O., Pollak, M., St-Pierre, J., and Topisirovic, I. (2015). mTOR coordinates protein synthesis, mitochondrial activity and proliferation. Cell Cycle *14*, 473–480.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth *5*, 621–628.

Müller, D., Katinger, H., and Grillari, J. (2008). MicroRNAs as targets for engineering of CHO cell factories. Trends in Biotechnology *26*, 359–365.

Nagano, T., Hashimoto, T., Nakashima, A., Hisanaga, S., Kikkawa, U., and Kamada, S. (2013). Cyclin I is involved in the regulation of cell cycle progression. Cell Cycle *12*, 2617–2624.

Nguyen, T.C., Cao, X., Yu, P., Xiao, S., Lu, J., Biase, F.H., Sridhar, B., Huang, N., Zhang, K., and Zhong, S. (2016). Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. Nature Communications *7*, 12023.

Nickless, A., Bailis, J.M., and You, Z. (2017). Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci *7*.

Nie, M., Balda, M.S., and Matter, K. (2012). Stress- and Rho-activated ZO-1-associated nucleic acid binding protein binding to p21 mRNA mediates stabilization, translation, and cell survival. Proc. Natl. Acad. Sci. U.S.A. *109*, 10897–10902.

Nielsen, J. (2013). Production of biopharmaceutical proteins by yeast. Bioengineered *4*, 207–211.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Obayashi, E., Luna, R.E., Nagata, T., Martin-Marcos, P., Hiraishi, H., Singh, C.R., Erzberger, J.P., Zhang, F., Arthanari, H., Morris, J., et al. (2017). Molecular Landscape of the Ribosome Pre-initiation Complex during mRNA Scanning: Structural Role for eIF3c and Its Control by eIF5. Cell Rep *18*, 2651–2663.

Oertlin, C., Lorent, J., Gandin, V., Murie, C., Masvidal, L., Cargnello, M., Furic, L., Topisirovic, I., and Larsson, O. (2018). Generally applicable transcriptome-wide analysis of translational efficiency using anota2seq. BioRxiv 106922.

Oguchi, S., Saito, H., Tsukahara, M., and Tsumura, H. (2006). pH Condition in temperature shift cultivation enhances cell longevity and specific hMab productivity in CHO culture. Cytotechnology *52*, 199–207.

Olshen, A.B., Hsieh, A.C., Stumpf, C.R., Olshen, R.A., Ruggero, D., and Taylor, B.S. (2013). Assessing gene-level translational control from ribosome profiling. Bioinformatics *29*, 2995–3002.

Orellana, C.A., Marcellin, E., Munro, T., Gary, P.P., and Nielsen, L.K. (2015). Multi-omics approach for comparative studies of monoclonal antibody producing CHO cells. BMC Proc *9*, O8.

Otera, H., Wang, C., Cleland, M.M., Setoguchi, K., Yokota, S., Youle, R.J., and Mihara, K. (2010). Mff is an essential factor for mitochondrial recruitment of Drp1 during mitochondrial fission in mammalian cells. The Journal of Cell Biology *191*, 1141–1158.

Ozaki, S. (1998). Hybridomas, T Cell. In Encyclopedia of Immunology (Second Edition), P.J. Delves, ed. (Oxford: Elsevier), pp. 1152–1154.

Paikari, A., D Belair, C., Saw, D., and Blelloch, R. (2017). The eutheria-specific miR-290 cluster modulates placental growth and maternal-fetal transport. Development *144*, 3731–3743.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Pan, X., Dalm, C., Wijffels, R.H., and Martens, D.E. (2017). Metabolic characterization of a CHO cell size increase phase in fed-batch cultures. Appl Microbiol Biotechnol *101*, 8101–8113.

Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., and Hatzigeorgiou, A.G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res *41*, W169–W173.

Park, S.-Y., and Grabau, E. (2017). Bypassing miRNA-mediated gene regulation under drought stress: alternative splicing affects CSD1 gene expression. Plant Mol Biol *95*, 243–252.

Pastò, A., Bellio, C., Pilotto, G., Ciminale, V., Silic-Benussi, M., Guzzo, G., Rasola, A., Frasson, C., Nardo, G., Zulato, E., et al. (2014). Cancer stem cells from epithelial ovarian cancer patients privilege oxidative phosphorylation, and resist glucose deprivation. Oncotarget *5*, 4305–4319.

Pease, J., and Sooknanan, R. (2012). A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing. Nat Meth *9*.

Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods *12*, 780–786.

Pereira, S., Kildegaard, H.F., and Andersen, M.R. (2018). Impact of CHO Metabolism on Cell Growth and Protein Production: An Overview of Toxic and Inhibiting Metabolites and Nutrients. Biotechnology Journal *13*, 1700499.

Persistence Market Research (2015). Global Market Study on Biopharmaceuticals: Asia to Witness Highest Growth by 2020.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotech *33*, 290–295.

Perucho, L., Artero-Castro, A., Guerrero, S., Cajal, S.R. y, LLeonart, M.E., and Wang, Z.-Q. (2014). RPLP1, a Crucial Ribosomal Protein for Embryonic Development of the Nervous System. PLOS ONE *9*, e99956.

Philippe, N., Salson, M., Commes, T., and Rivals, E. (2013). CRAC: an integrated approach to the analysis of RNA-seq reads. Genome Biol. *14*, R30.

Pilotte, J., Dupont-Versteegden, E.E., and Vanderklish, P.W. (2011). Widespread Regulation of miRNA Biogenesis at the Dicer Step by the Cold-Inducible RNA-Binding Protein, RBM3. PLOS ONE *6*, e28446.

Pinzón, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F., and Seitz, H. (2017). microRNA target prediction programs predict many false positives. Genome Res *27*, 234–245.

Pisareva, V.P., and Pisarev, A.V. (2014). eIF5 and eIF5B together stimulate 48S initiation complex formation during ribosomal scanning. Nucleic Acids Res. *42*, 12052–12069.

Popa, A., Lebrigand, K., Paquet, A., Nottet, N., Robbe-Sermesant, K., Waldmann, R., and Barbry, P. (2016). RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. F1000Res *5*, 1309.

Poulain, A., Perret, S., Malenfant, F., Mullick, A., Massie, B., and Durocher, Y. (2017). Rapid protein production from stable CHO cell pools using plasmid vector and the cumate gene-switch. J. Biotechnol. *255*, 16–27.

Preußner, M., Goldammer, G., Neumann, A., Haltenhof, T., Rautenstrauch, P., Müller-McNicoll, M., and Heyd, F. (2017). Body Temperature Cycles Control Rhythmic Alternative Splicing in Mammals. Molecular Cell *67*, 433-446.e4.

Pritchard, C.C., Cheng, H.H., and Tewari, M. (2012). MicroRNA profiling: approaches and considerations. Nature Reviews Genetics *13*, 358–369.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raj, A., Wang, S.H., Shim, H., Harpak, A., Li, Y.I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. ELife Sciences *5*, e13328.

Ray, A., James, M.K., Larochelle, S., Fisher, R.P., and Blain, S.W. (2009). p27Kip1 Inhibits Cyclin D-Cyclin-Dependent Kinase 4 by Two Independent Modes. Mol Cell Biol *29*, 986–999.

Riffo-Campos, Á.L., Riquelme, I., and Brebi-Mieville, P. (2016). Tools for Sequence-Based miRNA Target Prediction: What to Choose? Int J Mol Sci *17*.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., et al. (2010). De novo assembly and analysis of RNA-seq data. Nat Meth *7*, 909–912.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rodnina, M.V. (2016). The ribosome in action: Tuning of translational efficiency and protein folding. Protein Sci *25*, 1390–1406.

Roesch, K. (2004). The calcium-binding aspartate/glutamate carriers, citrin and aralar1, are new substrates for the DDP1/TIMM8a-TIMM13 complex. Human Molecular Genetics *13*, 2101–2111.

Roilo, M., Kullmann, M.K., and Hengst, L. (2018). Cold-inducible RNA-binding protein (CIRP) induces translation of the cell-cycle inhibitor p27Kip1. Nucleic Acids Res *46*, 3198–3210.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. Science *281*, 363, 365.

Ronda, C., Pedersen, L.E., Hansen, H.G., Kallehauge, T.B., Betenbaugh, M.J., Nielsen, A.T., and Kildegaard, H.F. (2014). Accelerating genome editing in CHO cells using

CRISPR Cas9 and CRISPy, a web-based target finding tool. Biotechnol. Bioeng. *111*, 1604–1616.

Rosano, G.L., and Ceccarelli, E.A. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol *5*.

Rothberg, J.M., and Leamon, J.H. (2008). The development and impact of 454 sequencing. Nat. Biotechnol. *26*, 1117–1124.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Ru, Y., Kechris, K.J., Tabakoff, B., Hoffman, P., Radcliffe, R.A., Bowler, R., Mahaffey, S., Rossi, S., Calin, G.A., Bemis, L., et al. (2014). The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. Nucleic Acids Res *42*, e133–e133.

Rupp, O., Becker, J., Brinkrolf, K., Timmermann, C., Borth, N., Pühler, A., Noll, T., and Goesmann, A. (2014). Construction of a Public CHO Cell Line Transcript Database Using Versatile Bioinformatics Analysis Pipelines. PLoS ONE *9*, e85568.

Saghatelian, A., and Couso, J.P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. Nat. Chem. Biol. *11*, 909–916.

Saito, Y., Nakaoka, T., and Saito, H. (2015). microRNA-34a as a Therapeutic Agent against Human Cancer. J Clin Med *4*, 1951–1959.

Salazar-Roa, M., and Malumbres, M. (2017). Fueling the Cell Division Cycle. Trends in Cell Biology *27*, 69–81.

Sampath, P., Pritchard, D.K., Pabon, L., Reinecke, H., Schwartz, S.M., Morris, D.R., and Murry, C.E. (2008). A Hierarchical Network Controls Protein Translation during Murine Embryonic Stem Cell Self-Renewal and Differentiation. Cell Stem Cell *2*, 448–460.

Sanchez, N., Gallagher, M., Lao, N., Gallagher, C., Clarke, C., Doolan, P., Aherne, S., Blanco, A., Meleady, P., Clynes, M., et al. (2013). MiR-7 Triggers Cell Cycle Arrest at the G1/S Transition by Targeting Multiple Genes Including Skp2 and Psme3. PLoS One *8*.

Sanchez, N., Kelly, P., Gallagher, C., Lao, N.T., Clarke, C., Clynes, M., and Barron, N. (2014). CHO cell culture longevity and recombinant protein yield are enhanced by depletion of miR-7 activity via sponge decoy vectors. Biotechnol J *9*, 396–404.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. *94*, 441–448.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. Cell *101*, 671–684.

Schröder, M., Matischak, K., and Friedl, P. (2004). Serum- and protein-free media formulations for the Chinese hamster ovary cell line DUKXB11. J. Biotechnol. *108*, 279–292.

Scott, L.J. (2015). Alipogene Tiparvovec: A Review of Its Use in Adults with Familial Lipoprotein Lipase Deficiency. Drugs *75*, 175–182.

Sengupta, J., Nilsson, J., Gursky, R., Kjeldgaard, M., Nissen, P., and Frank, J. (2008). Visualization of the eEF2-80S Ribosome Transition State Complex by Cryo-electron Microscopy. J Mol Biol *382*, 179–187.

Sharma, A., Jacob, A., Tandon, M., and Kumar, D. (2010). Orphan drug: Development trends and strategies. J Pharm Bioallied Sci *2*, 290–299.

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M.S.H. (2009). Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. DNA Res *16*, 45–58.

Shaughnessy, A.F. (2012). Monoclonal antibodies: magic bullets with a hefty price tag. BMJ *345*, e8346.

Shen, E.-Z., Chen, H., Ozturk, A.R., Tu, S., Shirayama, M., Tang, W., Ding, Y.-H., Dai, S.-Y., Weng, Z., and Mello, C.C. (2018). Identification of piRNA Binding Sites Reveals the Argonaute Regulatory Landscape of the C. elegans Germline. Cell *172*, 937-951.e18.

Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U.S.A. *111*, E5593-5601.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728–1732.

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nature Reviews Molecular Cell Biology *18*, 655–670.

Shi, Y., and Liu, J.-P. (2011). Gdf11 Facilitates Temporal Progression of Neurogenesis in the Developing Spinal Cord. J Neurosci *31*, 883–893.

Shi, Z., Fujii, K., Kovary, K.M., Genuth, N.R., Röst, H.L., Teruel, M.N., and Barna, M. (2017). Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. Molecular Cell *67*, 71-83.e7.

Shu, M., Zheng, X., Wu, S., Lu, H., Leng, T., Zhu, W., Zhou, Y., Ou, Y., Lin, X., Lin, Y., et al. (2011). Targeting oncogenic miR-335 inhibits growth and invasion of malignant astrocytoma cells. Mol. Cancer *10*, 59.

Sitton, G., and Srienc, F. (2008). Growth dynamics of mammalian cells monitored with automated cell cycle staining and flow cytometry. Cytometry A *73*, 538–545.

Slabáková, E., Culig, Z., Remšík, J., and Souček, K. (2017). Alternative mechanisms of miR-34a regulation in cancer. Cell Death Dis *8*, e3100.

Sokabe, M., and Fraser, C.S. (2017). A helicase-independent activity of eIF4A in promoting mRNA recruitment to the human ribosome. Proc Natl Acad Sci U S A *114*, 6304–6309.

Soneson, C., Matthes, K.L., Nowicka, M., Law, C.W., and Robinson, M.D. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biology *17*, 12.

Spealman, P., Naik, A.W., May, G.E., Kuersten, S., Freeberg, L., Murphy, R.F., and McManus, J. (2018). Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. Genome Res *28*, 214–222.

Stenvang, J., Petri, A., Lindow, M., Obad, S., and Kauppinen, S. (2012). Inhibition of microRNA function by antimiR oligonucleotides. Silence *3*, 1.

Strack, S., Wilson, T.J., and Cribbs, J.T. (2013). Cyclin-dependent kinases regulate splice-specific targeting of dynamin-related protein 1 to microtubules. J Cell Biol *201*, 1037–1051.

Strotbek, M., Florin, L., Koenitzer, J., Tolstrup, A., Kaufmann, H., Hausser, A., and Olayioye, M.A. (2013). Stable microRNA expression enhances therapeutic antibody productivity of Chinese hamster ovary cells. Metabolic Engineering *20*, 157–166.

Su, H.-Y., Lin, Z.-Y., Peng, W.-C., Guan, F., Zhu, G.-T., Mao, B.-B., Dai, B., Huang, H., and Hu, Z.-Q. (2018). MiR-448 downregulates CTTN to inhibit cell proliferation and promote apoptosis in glioma. Eur Rev Med Pharmacol Sci *22*, 3847–3854.

Su, W.-L., Modrek, B., GuhaThakurta, D., Edwards, S., Shah, J.K., Kulkarni, A.V., Russell, A., Schadt, E.E., Johnson, J.M., and Castle, J.C. (2008). Exon and junction microarrays detect widespread mouse strain- and sex-bias expression differences. BMC Genomics *9*, 273.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., and Yaspo, M.-L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics *15*.

Sun, F., Fu, H., Liu, Q., Tie, Y., Zhu, J., Xing, R., Sun, Z., and Zheng, X. (2008). Downregulation of CCND1 and CDK6 by miR-34a induces cell cycle arrest. FEBS Lett. *582*, 1564–1568.

Sun, Z., Nair, A., Chen, X., Prodduturi, N., Wang, J., and Kocher, J.-P. (2017). UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. Scientific Reports *7*, 14196.

Tabuchi, H. (2013). Novel strategy for a high-yielding mAb-producing CHO strain (overexpression of non-coding RNA enhanced proliferation and improved mAb yield). BMC Proc *7*, O3.

Takahashi, Y., Karbowski, M., Yamaguchi, H., Kazi, A., Wu, J., Sebti, S.M., Youle, R.J., and Wang, H.-G. (2005). Loss of Bif-1 suppresses Bax/Bak conformational change and mitochondrial apoptosis. Mol. Cell. Biol. *25*, 9369–9382.

Tan, H.K., Lee, M.M., Yap, M.G.S., and Wang, D.I.C. (2008). Overexpression of cold-inducible RNA-binding protein increases interferon-γ production in Chinese-hamster ovary cells. Biotechnology and Applied Biochemistry *49*, 247–257.

Tang, H., Lee, M., Sharpe, O., Salamone, L., Noonan, E.J., Hoang, C.D., Levine, S., Robinson, W.H., and Shrager, J.B. (2012). Oxidative stress-responsive microRNA-320 regulates glycolysis in diverse biological systems. FASEB J *26*, 4710–4721.

Teixeira, L.K., Carrossini, N., Sécca, C., Kroll, J.E., DaCunha, D.C., Faget, D.V., Carvalho, L.D.S., de Souza, S.J., and Viola, J.P.B. (2016). NFAT1 transcription factor regulates cell cycle progression and cyclin E expression in B lymphocytes. Cell Cycle *15*, 2346–2359.

Thaisuchat, H., Baumann, M., Pontiller, J., Hesse, F., and Ernst, W. (2011). Identification of a novel temperature sensitive promoter in cho cells. BMC Biotechnol *11*, 51.

Toit, A.D. (2013). RNA: Circular RNAs as miRNA sponges.

Toyama, B.H., and Hetzer, M.W. (2013). Protein homeostasis: live long, won't prosper. Nat Rev Mol Cell Biol *14*, 55–61.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech *28*, 511–515.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotech *31*, 46–53.

Trummer, E., Fauland, K., Seidinger, S., Schriebl, K., Lattenmayer, C., Kunert, R., Vorauer-Uhl, K., Weik, R., Borth, N., Katinger, H., et al. (2006). Process parameter shifting: Part I. Effect of DOT, pH, and temperature on the performance of Epo-Fc expressing CHO cells cultivated in controlled batch bioreactors. Biotechnol. Bioeng. *94*, 1033–1044.

Tzani, I., Monger, C., Kelly, P., Barron, N., Kelly, R.M., and Clarke, C. (2018). Understanding biopharmaceutical production at single nucleotide resolution using ribosome footprint profiling. Current Opinion in Biotechnology *53*, 182–190.

Underhill, M.F., and Smales, C.M. (2007). The cold-shock response in mammalian cells: investigating the HeLa cell cold-shock proteome. Cytotechnology *53*, 47–53.

Urlaub, G., and Chasin, L.A. (1980). Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. Proc Natl Acad Sci U S A *77*, 4216–4220.

Urlaub, G., Käs, E., Carothers, A.M., and Chasin, L.A. (1983). Deletion of the diploid dihydrofolate reductase locus from cultured mammalian cells. Cell *33*, 405–412.

Valadkhan, S., and Gunawardane, L.S. (2013). Role of small nuclear RNAs in eukaryotic gene expression. Essays Biochem. *54*, 79–90.

Valdés-Bango Curell, R., and Barron, N. (2018). Exploring the Potential Application of Short Non-Coding RNA-Based Genetic Circuits in Chinese Hamster Ovary Cells. Biotechnol J.

Vanstone, J.R., Smith, A.M., McBride, S., Naas, T., Holcik, M., Antoun, G., Harper, M.-E., Michaud, J., Sell, E., Chakraborty, P., et al. (2016). DNM1L-related mitochondrial fission defect presenting as refractory epilepsy. European Journal of Human Genetics *24*, 1084–1088.

Vêncio, R.Z., Brentani, H., Patrão, D.F., and Pereira, C.A. (2004). Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). BMC Bioinformatics *5*, 119.

Vergara, M., Becerra, S., Berrios, J., Osses, N., Reyes, J., Rodríguez-Moyá, M., Gonzalez, R., and Altamirano, C. (2014). Differential Effect of Culture Temperature and Specific Growth Rate on CHO Cell Behavior in Chemostat Culture. PLoS One *9*.

Vergoulis, T., Vlachos, I.S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A.G. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res *40*, D222–D229.

Vishwanathan, N., Yongky, A., Johnson, K.C., Fu, H.-Y., Jacob, N.M., Le, H., Yusufi, F.N.K., Lee, D.Y., and Hu, W.-S. (2014). Global insights into the Chinese hamster and CHO cell transcriptomes. Biotechnol. Bioeng. n/a-n/a.

Vitting-Seerup, K., and Sandelin, A. (2017). The Landscape of Isoform Switches in Human Cancers. Mol. Cancer Res. *15*, 1206–1220.

Voronina, E.V., Seregin, Y.A., Litvinova, N.A., Shvets, V.I., and Shukurov, R.R. (2016). Design of a stable cell line producing a recombinant monoclonal anti-TNFα antibody based on a CHO cell line. Springerplus *5*.

Vyas, K., Chaudhuri, S., Leaman, D.W., Komar, A.A., Musiyenko, A., Barik, S., and Mazumder, B. (2009). Genome-Wide Polysome Profiling Reveals an Inflammation-Responsive Posttranscriptional Operon in Gamma Interferon-Activated Monocytes. Mol. Cell. Biol. *29*, 458–470.

Vychytilova-Faltejskova, P., Merhautova, J., Machackova, T., Gutierrez-Garcia, I., Garcia-Solano, J., Radova, L., Brchnelova, D., Slaba, K., Svoboda, M., Halamkova, J., et al. (2017). MiR-215-5p is a tumor suppressor in colorectal cancer targeting EGFR ligand epiregulin and its transcriptional inducer HOXB9. Oncogenesis *6*, 399.

Waheed, M.T., Sameeullah, M., Khan, F.A., Syed, T., Ilahi, M., Gottschamel, J., and Lössl, A.G. (2016). Need of cost-effective vaccines in developing countries: What plant biotechnology can offer? Springerplus *5*.

Walsh, G. (2006). Biopharmaceutical benchmarks 2006. Nat Biotech *24*, 769–776.

Walsh, G. (2010a). Biopharmaceutical benchmarks 2010. Nat Biotech *28*, 917–924.

Walsh, G. (2010b). Post-translational modifications of protein biopharmaceuticals. Drug Discovery Today *15*, 773–780.

Walsh, G. (2014). Biopharmaceutical benchmarks 2014. Nat Biotech *32*, 992–1000.

Wang, X. (2008). miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA *14*, 1012–1017.

Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat. Biotechnol. *32*, 926–932.

Wang, H., McManus, J., and Kingsford, C. (2016). Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. Bioinformatics *32*, 1880–1882.

Wang, Y., Baskerville, S., Shenoy, A., Babiarz, J.E., Baehner, L., and Blelloch, R. (2008). Embryonic stem cell–specific microRNAs regulate the G1-S transition and promote rapid proliferation. Nature Genetics *40*, 1478–1483.

WANG, Y., LIU, J., HUANG, B., XU, Y.-M., LI, J., HUANG, L.-F., LIN, J., ZHANG, J., MIN, Q.-H., YANG, W.-M., et al. (2015). Mechanism of alternative splicing and its regulation. Biomed Rep *3*, 152–158.

Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B., and Carstens, R.P. (2009). ESRP1 and ESRP2 are epithelial cell type-specific regulators of FGFR2 splicing. Mol Cell *33*, 591–601.

Watanabe, T., and Lin, H. (2014). Post-transcriptional regulation of gene expression by Piwi proteins and piRNAs. Mol Cell *56*, 18–27.

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M., and Losick, R. (2014). Molecular Biology of the Gene, Chapter 15. In Molecular Biology of the Gene, (Pearson), pp. 509–571.

Weber, W., and Fussenegger, M. (2007). Inducible product gene expression technology tailored to bioprocess engineering. Current Opinion in Biotechnology *18*, 399–410.

Wei, S., and Qian, S.-B. (2015). Ribosome Profiling: Principles and Variations. In ELS, (American Cancer Society), pp. 1–7.

Wei, H.-H., Liu, Y., Wang, Y., Lu, Q., Yang, X., Li, J., and Wang, Z. (2017). Engineering Artificial Factors to Specifically Manipulate Alternative Splicing in Human Cells. J Vis Exp.

Wenz, T., Wang, X., Marini, M., and Moraes, C.T. (2011). A metabolic shift induced by a PPAR panagonist markedly reduces the effects of pathogenic mitochondrial tRNA mutations. J. Cell. Mol. Med. *15*, 2317–2325.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Wilson, D.N., and Doudna Cate, J.H. (2012). The Structure and Function of the Eukaryotic Ribosome. Cold Spring Harb Perspect Biol *4*.

Wong, C.H., Siah, K.W., and Lo, A.W. (2018). Estimation of clinical trial success rates and related parameters. Biostatistics.

Wong, J.J.-L., Au, A.Y.M., Gao, D., Pinello, N., Kwok, C.-T., Thoeng, A., Lau, K.A., Gordon, J.E.A., Schmitz, U., Feng, Y., et al. (2016). RBM3 regulates temperature sensitive miR-142–5p and miR-143 (thermomiRs), which target immune genes and control fever. Nucleic Acids Res *44*, 2888–2897.

Wongkittichote, P., Tungpradabkul, S., Wattanasirichaigoon, D., and Jensen, L.T. (2013). Prediction of the functional effect of novel SLC25A13 variants using a S. cerevisiae model of AGC2 deficiency. J. Inherit. Metab. Dis. *36*, 821–830.

Wright, C., and Estes, S. (2014). Next-generation bioprocess: an industry perspective of how the ‘omics era will affect future biotherapeutic development. Pharmaceutical Bioprocessing *2*, 371–375.

Wu, B., Su, S., Patil, D.P., Liu, H., Gan, J., Jaffrey, S.R., and Ma, J. (2018). Molecular basis for the specific and multivariant recognitions of RNA substrates by human hnRNP A2/B1. Nature Communications *9*, 420.

Wu, J., Liu, Q., Wang, X., Zheng, J., Wang, T., You, M., Sheng Sun, Z., and Shi, Q. (2013). mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. RNA Biol *10*, 1087–1092.

Wu, S., Aksoy, M., Shi, J., and Houbaviy, H.B. (2014). Evolution of the miR-290–295/miR-371–373 Cluster Family Seed Repertoire. PLOS ONE *9*, e108519.

Wurm, F.M. (2004). Production of recombinant protein therapeutics in cultivated mammalian cells. Nat Biotech *22*, 1393–1398.

Wurm, F.M. (2013). CHO Quasispecies—Implications for Manufacturing Processes. Processes *1*, 296–311.

Xia, Z., Zheng, X., Zheng, H., Liu, X., Yang, Z., and Wang, X. (2012). Cold-inducible RNA-binding protein (CIRP) regulates target mRNA stabilization in the mouse testis. FEBS Letters *586*, 3299–3308.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res *37*, D105–D110.

Xiao, Z., Zou, Q., Liu, Y., and Yang, X. (2016). Genome-wide assessment of differential translations with ribosome profiling data. Nat Commun *7*, 11194.

Xie, L., Ushmorov, A., Leithäuser, F., Guan, H., Steidl, C., Färbinger, J., Pelzer, C., Vogel, M.J., Maier, H.J., Gascoyne, R.D., et al. (2012). FOXO1 is a tumor suppressor in classical Hodgkin lymphoma. Blood *119*, 3503–3511.

Xing, F., Luan, Y., Cai, J., Wu, S., Mai, J., Gu, J., Zhang, H., Li, K., Lin, Y., Xiao, X., et al. (2017). The Anti-Warburg Effect Elicited by the cAMP-PGC1α Pathway Drives Differentiation of Glioblastoma Cells into Astrocytes. Cell Reports *18*, 468–481.

Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res *30*, 3754–3766.

Xu, X., Nakano, T., Wick, S., Dubay, M., and Brizuela, L. (1999). Mechanism of Cdk2/Cyclin E inhibition by p27 and p27 phosphorylation. Biochemistry *38*, 8713–8722.

Xu, X., Nagarajan, H., Lewis, N.E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., et al. (2011). The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotech *29*, 735–741.

Yamaguchi, T., Kurita, T., Nishio, K., Tsukada, J., Hachisuga, T., Morimoto, Y., Iwai, Y., and Izumi, H. (2015). Expression of BAF57 in ovarian cancer cells and drug sensitivity. Cancer Sci. *106*, 359–366.

Yang, W., and Paschen, W. (2008). Conditional gene silencing in mammalian cells mediated by a stress-inducible promoter. Biochemical and Biophysical Research Communications *365*, 521–527.

Yang, R., Zhan, M., Nalabothula, N.R., Yang, Q., Indig, F.E., and Carrier, F. (2010). Functional Significance for a Heterogenous Ribonucleoprotein A18 Signature RNA Motif in the 3′-Untranslated Region of Ataxia Telangiectasia Mutated and Rad3-related (ATR) Transcript. J Biol Chem *285*, 8887–8893.

Yang, X., Zhang, H., and Li, L. (2012). Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. Plant J. *70*, 421–431.

Yasuda, O., Fukuo, K., Sun, X., Nishitani, M., Yotsui, T., Higuchi, M., Suzuki, T., Rakugi, H., Smithies, O., Maeda, N., et al. (2006). Apop-1, a novel protein inducing cyclophilin D-dependent but Bax/Bak-related channel-independent apoptosis. J. Biol. Chem. *281*, 23899–23907.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134.

Yee, J.C., Wlaschin, K.F., Chuah, S.H., Nissom, P.M., and Hu, W.-S. (2008). Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. Biotechnol. Bioeng. *101*, 1359–1365.

Yee, J.C., Gerdtzen, Z.P., and Hu, W.-S. (2009). Comparative transcriptome analysis to unveil genes affecting recombinant protein productivity in mammalian cells. Biotechnol. Bioeng. *102*, 246–263.

Yoon, J.-H., Abdelmohsen, K., and Gorospe, M. (2013). Posttranscriptional gene regulation by long noncoding RNA. J. Mol. Biol. *425*, 3723–3730.

Yordanova, M.M., Loughran, G., Zhdanov, A.V., Mariotti, M., Kiniry, S.J., O'Connor, P.B.F., Andreev, D.E., Tzani, I., Saffert, P., Michel, A.M., et al. (2018). AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation. Nature *553*, 356–360.

Yoshino, H., Yonemori, M., Miyamoto, K., Tatarano, S., Kofuji, S., Nohata, N., Nakagawa, M., and Enokida, H. (2017). microRNA-210-3p depletion by CRISPR/Cas9 promoted tumorigenesis through revival of TWIST1 in renal cell carcinoma. Oncotarget *8*, 20881–20894.

Young, J.D. (2013). Metabolic flux rewiring in mammalian cell cultures. Curr Opin Biotechnol *24*.

Yuk, I.H., Zhang, J.D., Ebeling, M., Berrera, M., Gomez, N., Werz, S., Meiringer, C., Shao, Z., Swanberg, J.C., Lee, K.H., et al. (2014). Effects of copper on CHO cells: Insights from gene expression analyses. Biotechnol Progress *30*, 429–442.

Zagari, F., Jordan, M., Stettler, M., Broly, H., and Wurm, F.M. (2013). Lactate metabolism shift in CHO cell culture: the role of mitochondrial oxidative activity. New Biotechnology *30*, 238–245.

Zaragoza, K., Bégay, V., Schuetz, A., Heinemann, U., and Leutz, A. (2010). Repression of Transcriptional Activity of C/EBPα by E2F-Dimerization Partner Complexes. Mol Cell Biol *30*, 2293–2304.

Zeng, F., Xiong, J., Ke, W., Pu, J., and Zhan, Y. (2018). Expression of MiR-9 promotes proliferation, migration and differentiation of human neural stem cells. Tropical Journal of Pharmaceutical Research *17*, 213-217–217.

Zeng, Y., Wodzenski, D., Gao, D., Shiraishi, T., Terada, N., Li, Y., Griend, D.J.V., Luo, J., Kong, C., Getzenberg, R.H., et al. (2013). Stress-Response Protein RBM3 Attenuates the Stem-like Properties of Prostate Cancer Cells by Interfering with CD44 Variant Splicing. Cancer Res *73*, 4123–4133.

Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res *46*, D754–D761.

Zhang, P., and Reue, K. (2017). Lipin proteins and glycerolipid metabolism: Roles at the ER membrane and beyond. Biochimica et Biophysica Acta (BBA) - Biomembranes *1859*, 1583–1595.

Zhang, C., Mao, H.-L., and Cao, Y. (2017a). Nuclear accumulation of symplekin promotes cellular proliferation and dedifferentiation in an ERK1/2-dependent manner. Scientific Reports *7*, 3769.

Zhang, J.-X., Xu, Y., Gao, Y., Chen, C., Zheng, Z.-S., Yun, M., Weng, H.-W., Xie, D., and Ye, S. (2017b). Decreased expression of miR-939 contributes to chemoresistance and metastasis of gastric cancer via dysregulation of SLC34A2 and Raf/MEK/ERK pathway. Molecular Cancer *16*, 18.

Zhang, L.-F., Lou, J.-T., Lu, M.-H., Gao, C., Zhao, S., Li, B., Liang, S., Li, Y., Li, D., and Liu, M.-F. (2015). Suppression of miR-199a maturation by HuR is crucial for hypoxia-induced glycolytic switch in hepatocellular carcinoma. EMBO J. *34*, 2671–2685.

Zhang, P., Woen, S., Wang, T., Liau, B., Zhao, S., Chen, C., Yang, Y., Song, Z., Wormald, M.R., Yu, C., et al. (2016). Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic drugs. Drug Discovery Today *21*, 740–765.

Zhang, X., Hu, S., Zhang, X., Wang, L., Zhang, X., Yan, B., Zhao, J., Yang, A., and Zhang, R. (2014). MicroRNA-7 arrests cell cycle in G1 phase by directly targeting CCNE1 in human hepatocellular carcinoma cells. Biochemical and Biophysical Research Communications *443*, 1078–1084.

Zhao, L.-N., Wang, P., Liu, Y.-H., Cai, H., Ma, J., Liu, L.-B., Xi, Z., Li, Z.-Q., Liu, X.-B., and Xue, Y.-X. (2017). MiR-383 inhibits proliferation, migration and angiogenesis of glioma-exposed endothelial cells in vitro via VEGF-mediated FAK and Src signaling pathways. Cellular Signalling *30*, 142–153.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014a). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE *9*, e78644.

Zhao, S., Han, J., Zheng, L., Yang, Z., Zhao, L., and Lv, Y. (2015). MicroRNA-203 Regulates Growth and Metastasis of Breast Cancer. Cell. Physiol. Biochem. *37*, 35–42.

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N., and Perou, C.M. (2014b). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics *15*.

Zhao, X., Yu, Y., Zhao, Z., Guo, J., Fu, L., Yu, T., Hou, L., Yi, S., and Chen, W. (2012). Establishment of tetracycline-inducible, survivin-expressing CHO cell lines by an optimized screening method. Biosci. Biotechnol. Biochem. *76*, 1909–1912.

Zheng, N., and Shabek, N. (2017). Ubiquitin Ligases: Structure, Function, and Regulation. Annual Review of Biochemistry *86*, 129–157.

Zheng, G.X.Y., Ravi, A., Calabrese, J.M., Medeiros, L.A., Kirak, O., Dennis, L.M., Jaenisch, R., Burge, C.B., and Sharp, P.A. (2011a). A Latent Pro-Survival Function for the Mir-290-295 Cluster in Mouse Embryonic Stem Cells. PLOS Genetics *7*, e1002054.

Zheng, K., Bantog, C., and Bayer, R. (2011b). The impact of glycosylation on monoclonal antibody conformation and stability. MAbs *3*, 568–576.

Zhou, H.-L., Luo, G., Wise, J.A., and Lou, H. (2014a). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. Nucleic Acids Res *42*, 701–713.

Zhou, J., Lancaster, L., Donohue, J.P., and Noller, H.F. (2014b). How the Ribosome Hands the A-site tRNA to the P Site During EF-G-catalyzed Translocation. Science *345*, 1188–1191.

Zhu, X., Bührer, C., and Wellmann, S. (2016). Cold-inducible proteins CIRP and RBM3, a unique couple with activities far beyond the cold. Cell. Mol. Life Sci. *73*, 3839–3859.

Zucchelli, S., Fasolo, F., Russo, R., Cimatti, L., Patrucco, L., Takahashi, H., Jones, M.H., Santoro, C., Sblattero, D., Cotella, D., et al. (2015). SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. Front Cell Neurosci *9*.

Appendix

This thesis is accompanied by 6 appendices: 1 text appendix following this section, 4 multi-sheet excel workbooks and 4 text files containing bash computer code.

List of Appendices

1 - Benchmarking of quality control and alignment algorithms and selection of optimal reference genome (text following this section).

2 – Chapter 2 Supplementary results (Appendix2.xlsx)

A) Cell culture density, viability and metabolic measurements

B) Read quality & alignment statistics

C) Differentially expressed genes

D) CHO Pathways of interest

E) GO analysis of differentially expressed genes

F) Differentially expressed proteins

G) GO analysis of differentially expressed proteins

3 – Chapter 3 Supplementary results (Appendix3.xlsx)

A) Differential isoform expression with *stringtie*

B) Gene ontology analysis of genes with differential isoform usage

C) Differential isoform usage in genes in biologically relevant pathways

D) Differential isoforms not DE at gene level

E) Differential isoform usage in genes in biologically relevant pathways which were not identified as DE at the gene level

F) Isoform switching

G) Differential exon usage with *RMATS* – Skipped exons, alternative splice sites and retained introns

H) Differential exon usage with *RMATS* – Mutually exclusive exons

I) Differential exon usage in genes not identified as DE at gene level - Skipped exons, alternative splice sites and retained introns

J) Differential exon usage in genes not identified as DE at gene level - Mutually exclusive exons

K) Differential exon usage in genes in biologically relevant pathways not ID at gene level

L) Interpro protein domains in skipped exon genes

M) Differential splicing of genes which were not correlated at the gene and protein level

N) qPCR primers

O) qPCR results

4 – Chapter 4 Supplementary results (Appendix4.xlsx)

A) Read pre-processing and alignment statistics

B) *mirDeep2* Results

C) Significant novel *miRDeep2* Results

D) Differential expression analysis on *miRDeep2* Results

E) miRNA Targets on differentially expressed genes

F) Gene ontology analysis of miRNA interactions between differentially expressed genes and miRNAs

G) Differentially expressed genes targeted by 3 of more miRNAs with anti-correlated expression

H) miRNA Targets on differentially spliced genes

I) Gene ontology analysis of miRNA interactions on spliced genes

J) Genes differentially spliced in the 3'UTR with miRNA targets

K) Target prediction on differentially expressed proteins

L) Gene ontology analysis of miRNA interactions on differentially expressed proteins

M) miRNA target predictions on genes with uncorrelated expression between the gene and protein levels

5 – Chapter 5 & 6 Supplementary results (Appendix5.xlsx)

A) Quality control, preprocessing and alignment statistics

B) *DESeq2* differential translation

C) *xtail* differential translation

D) Differentially translated genes identified in both *DESeq2* and *xtail*

E) Gene ontology analysis on differentially translated genes identified in both *xtail* and *DESeq2*

F) Stable translated genes with high and low translation efficiency

G) Differentially spliced and translated genes

H) Gene ontology analysis on spliced and differentially translated genes

I) Differentially translated genes identified as targets of miRNAs

J) TE genes which cannot be explained by miRNA regulation or alternative splicing

K) Genes uncorrelated at the gene and protein level which are not spliced, targeted by miRNAs or differentially translated

6) Code to reproduce analysis

A) Chapter 2 Code (Appendix6A-Chapter2Code.txt)

B) Chapter 3 Code (Appendix6B-Chapter3Code.txt)

C) Chapter 4 Code (Appendix6C-Chapter4Code.txt)

D) Chapter 5 Code (Appendix6D-Chapter5Code.txt)

A1 – Benchmarking of quality control and alignment algorithms and selection of optimal reference genome

A1.1 Selection of optimal alignment algorithm

To identify the optimal alignment algorithm for aligning CHO cell RNA-Seq data, the datasets published in (Lewis et al., 2013; Rupp et al., 2014) and reads simulated with Flux simulator (Griebel et al., 2012) were aligned to the Chinese hamster genome (Lewis et al., 2013) using 4 alignment algorithms. Default alignment parameters were used across these experiments, with the exception of allowing for paired-end read input and multithreaded execution on 32 CPUs to minimise runtime. The alignment algorithm with the highest overall alignment percentage was CRAC, with 89.15%. This was followed by HISAT2 with 88.81% of reads aligned. The fastest alignment algorithm was STAR, completing alignments with an average speed of 3.1 minutes. This was followed by HISAT2 with 4.8 minutes. The Tophat2 and CRAC algorithms took considerably longer to complete alignments with runtimes of 102 minutes and 141 minutes respectively. Data presented in Table A1.

Table A1 – Benchmark analysis of alignment algorithms. 4 RNA-Seq alignment algorithms were selected for benchmarking analysis. *STAR* was the fastest algorithm in this test and *CRAC* aligned the most reads. *HISAT2* however had the best balance between speed and sensitivity.

	Tophat2		**HISAT2**		**STAR**		**CRAC**	
	Aligned	**Time**	**Aligned**	**Time**	**Aligned**	**Time**	**Aligned**	**Time**
Lewis et al.	61.9%	69m	65.37%	1m	64.97%	2m	72.45%	51m
Rupp et al.	77.4%	134m	80.42%	8m	78.53%	3m	85.36%	122m
Simulation 1	95.8%	97m	97.04%	5m	96.53%	6m	95.08%	219m
Simulation 2	96%	103m	97.23%	5m	96.71%	3m	95.34%	86m
Simulation 3	96.2%	110m	95.68%	5m	95.09%	2m	93.74%	184m
Simulation 4	95.9%	102m	97.13%	5m	94.08%	3m	92.92%	186m
Average	87.20%	102m	88.81%	4.8m	87.65%	3.16m	89.15%	141m

To determine the optimal alignment algorithm for aligning pre-processed reads to the Chinese hamster genome, a benchmarking study was used to identify the percentages of read alignment and runtime of alignment algorithms. Publically available Chinese hamster and CHO RNA-Seq data (Lewis et al., 2013; Rupp et al., 2014) was used alongside simulated reads generated using *Flux simulator* (Griebel et al., 2012) from transcript coding regions of the Chinese hamster genome using annotation available from UCSC. As these reads are derived from the genome, they should in theory align with or close to an alignment rate of 100%, however reads have been simulated to contain an error profile typical of RNA-Seq reads sequenced using Illumina instruments such as decreasing read quality in the 3' end and PCR and reverse transcription biases (Conesa et al., 2016). All aligners performed similarly in terms of the amounts of reads aligned with a range of 87.2%-89.15% read alignment. The *CRAC* alignment algorithm was the most successful in aligning the maximal amount of reads, with 89.15%, followed by *HISAT2* with 88.81%. The fastest alignment algorithm was *STAR*, with a runtime of just 3.1 minutes, followed by *HISAT2* in 4.8 minutes. The *Tophat2* and *CRAC* aligners were considerably slower, with runtimes of 102 minutes and 141 minutes respectively. The most memory efficient algorithm is *HISAT2*, requiring 4GB of RAM to hold a mammalian genome index (Kim et al., 2015a). *Tophat2* has similar memory requirements of 5.4Gb (Kim et al., 2013). *STAR* and *CRAC* have considerably higher memory requirements however of 28GB for *STAR* (Dobin et al., 2013) and over 40Gb for *CRAC*, increasing with the number of reads to be aligned (Philippe et al., 2013). With the best balance between read alignment rates, runtime and the lowest RAM requirement, *HISAT2* will be used as the alignment algorithm for future experiments and incorporated into analyses pipelines.

A1.2 Benchmarking of the available Chinese hamster genomes

Since the release of the first CHO genome in 2011, sequenced from the ancestral CHO-K1 cell line (Xu et al. 2011), additional genome sequences have since been assembled from various cell lines and the Chinese hamster organism itself. No comprehensive assessment of these genomes has yet taken place and it is unclear which of these genome sequences is most appropriate for use as a reference sequence for RNA-Seq and RIBO-Seq experiments. Furthermore, each of these genome sequences have existing reference annotation which have not been assessed and it is unclear which would best serve as a reference to guide and annotate genome-guided transcriptome assembly and differential gene expression experiments. In order to answer these questions, in this section we perform a comparative analysis of these genomes and identify which is most appropriate for use for differential gene and transcript expression experiments.

A1.2.1 Chinese hamster and CHO cell line genome sequences

For this analysis we selected available Chinese hamster and CHO cell line genome sequences which have accompanying reference. The 5 selected genome sequences were considered as they are each at a draft reference stage and sequenced to a high enough coverage for accurate assembly. The selected genomes include CriGri1 (Xu et al. 2011) – the first assembled CHO genome which was annotated by NCBI RefSeq in 2014, C_griseus_v1.0 (Lewis et al. 2013) – the parental Chinese hamster genome also annotated by NCBI RefSeq in 2014, CHOK1GS – the Horizon CHOK1GS cell line assembled and annotated by Ensembl in 2017, CriGri1 – the Ensembl 2017 reannotation of CriGri1, and PICR – the PacBio Chinese hamster genome corrected with Illumina reads annotated using the Maker pipeline (Cantarel et al., 2008).

A1.2.2 Assembly quality of the CHO and Chinese hamster genomes

An initial assessment of the genome assembly quality was carried out using the *BBtools* stats function (Joint Genome Institute). The results of this analysis are summarised in Table A2.

Table A2– Genome assembly statistics. 5 reference genomes of the Chinese hamster or CHO cell lines were assessed using *BBtools* to analyse the quality of their assembly and completeness

Genome	Scaffolds	Contigs	Bases (MB)	N50	L50	N90	N%	Gap%
CriGri1 (NCBI)	109,151	265,786	2,399	547	1,147,233	3061	3.4	3.403
C_griseus v1.0 (NCBI)	52,710	218,862	2,360	450	1,558,295	1558	2.49	2.492
CHOK1GS (Ensembl)	8,265	71,599	2,358	12	62,039,716	42	1.45	1.452
CriGri1 (Ensembl)	109,152	265,787	2,399	547	1,147,233	3061	3.4	3.403
PICR	1,830	4,878	2,368	33	19,581,774	122	0.12	0.121

These statistics show the largest genome is CriGri1. The 2 versions of CriGri1 differ by a single scaffold – the inclusion of the mitochondrial genome in the Ensembl version which is not present in the NCBI version. Although this genome is the largest in terms of bases, it is also the most fragmented genome and is comprised of considerably more scaffolds and contigs than the other assemblies and therefore the worst performing in terms of N50 and L50. It is possible that the increase in genome size comes from duplicated regions of the assembly or the higher percentage of N bases in this assembly and therefore the size of the genome in terms of bases is not the most important consideration when selecting a sequence to use as reference. The best assembled genome in terms of scaffolds and contigs is PICR which is made up of 1,830 scaffolds as a result of the advantage of using long read sequencing technology. Although CHOK1GS contains more scaffolds than the PICR assembly, CHOK1GS is also relatively well assembled and made up of 8,265 scaffolds. CHOK1GS contains the longest scaffolds with an L50 of 62,039,716 bases, resulting in 50% of the entire genome sequence being contained in just 12 scaffolds (N50) and 90% in 42 scaffolds (N90). CHOK1GS outperforms PICR in this regard which has N50 and N90 of 33 and 122 respectively. The C_griseus_v1.0 genome is neither the best or worst genome in any of the categories tested but has the advantage of being a genome sequence of the Chinese hamster itself rather than of a given CHO cell line. C_griseus_v1.0 may therefore be the most appropriate universal CHO reference due to the nature of CHO cell lines diverging from each other due to mutations in culture conditions and genetic modifications.

A1.2.3 Annotation of the CHO and Chinese hamster genomes

The annotations of the Chinese hamster genomes are each at different levels of completeness. The methodology used to annotate these genomes also vary. NCBI genomes use the NCBI Eukaryotic Genomic Annotation Pipeline which uses the *Splign* alignment algorithm to align transcripts and proteins in the RefSeq and Genbank databases in addition to RNA-Seq reads from the Sequence Read Archive (Kapustin et al., 2008). Genes are also predicted using *Gnomon* and other ncRNAs using *tRNAscan-SE* and *cmsearch*. The Ensembl gene annotation system uses a similar process with additional, more stringent assessment and filtering of target transcripts (Aken et al., 2016). The PICR genome was annotated using the MAKER pipeline (Campbell et al., 2014) which uses a variety of BLAST searches and Gene finders to annotate predicted genes. Table A3 contains the annotation statistics of each reference sequence.

Table A3 – Annotation available for Chinese hamster and CHO genomes. Transcriptome annotation is available for each of the 5 presented genomes. The annotations are at varying stages of completeness. CriGri1(NCBI) has the most genes and transcripts annotated.

Genome	GTF/ GFF3	Genes	Transcripts	ncRNAs	Gene Names	Non-redundant gene names
CriGri1(NCBI)	GFF3	28,978	35,628	5,738	15,695	14,149
C_griseus_v1.0 (NCBI)	GFF3	28,446	33,465	5,725	15,979	14,116
CHOK1GS (Ensembl)	GTF	25,072	32,575	4,248	20,020	16,722
CriGri1 (Ensembl)	GTF	26,668	34,472	7,487	18,340	15,067
PICR	GFF3	24,686	24,948	N/A	N/A	N/A

The CriGri1 (NCBI) genome has the most gene loci, transcripts and exons annotated and the CriGri1 (Ensembl) genome has the most ncRNAs annotated. The CHOK1GS (Ensembl) annotation has the highest rate of alternative splice variants annotated, with an average of 1.299 transcripts per gene locus. The PICR assembly has the fewest with 1.01 transcripts per gene locus. The CHOK1GS (Ensembl) assembly also has the most genes assigned a gene name and the most non-redundant gene names (excluding NCBI 'LOC' genes and Ensembl 'Rik' IDs as these are unknown transcripts). The PICR genome is not yet at a stage in its annotation at which it contains ncRNA/protein coding assignment or gene name annotation. Although the Ensembl genome annotations have less genes annotated than the NCBI genomes, they are of higher quality in that they have higher amounts of genes assigned annotation, more alternatively spliced transcripts and being in the GTF format means they have CDS and UTR regions annotated in the transcripts. An additional benefit of the Ensembl annotations is the use of stable Ensembl gene identifiers which can be used in other software packages such as *Biomart*.

A1.2.4 Use of genome annotation for differential gene expression experiments

The genomes tested in this section will be used in differential gene expression experiments and it is therefore important to test their performance using differential expression software to see which reference provides the most informative results. To test the different genome annotations, RNA-Seq reads were aligned to each genome using *HISAT2* and *featureCounts* was used to assign aligned reads to annotated gene features. *DESeq2* was then used to identify genes differentially expressed at a threshold of ≥±1.5 fold change and ≤0.05 adjusted P-value. Both the number of genes up and downregulated and the number of genes assigned gene symbol annotation were recorded for each genome.

Table A4 – Differentially expressed genes in reference annotation of 5 available genomes. Each of the 5 genomes in this table were used as reference in a differential gene expression analysis using temperature shifted CHO data. Genes were considered significantly differentially expressed if they had ≥±1.5 fold change and ≤0.05 P value. The number of total differentially expressed genes and only those with gene annotation are reported.

Genome	Upregulated	Downregulated	Upregulated Annotated	Downregulated Annotated
CriGri1(NCBI)	1,177	1,209	747	834
C_griseus_v1.0 (NCBI)	1,186	1,219	745	837
CHOK1GS (Ensembl)	850	955	726	781
CriGri1 (Ensembl)	938	1,017	629	691
PICR	877	941	0	0

A1.2.5 Use of genome sequences for genome guided transcript assembly

In Chapter 3 we detail the analysis of alternative splicing in CHO cells undergoing temperature shift. In this experiment we first assemble novel transcripts using genome guided assembly and therefore it is important to consider which genome has the greatest potential for use as a reference for assembling novel transcripts. Assembling transcripts using no reference annotation to guide assembly will provide data on which genome is the most complete as prior annotated genes will not confound analysis.

Table A5 – De novo genome-guided assembly results. Reads aligned to the 5 available genome sequences were assembled using *Stringtie* without the use of existing genome annotation as a guide. Transcripts were clustered using CD-Hit-EST to identify unique transcripts and remove the potential of paralogues and duplicated genome regions confounding analysis.

Genome	**Gene Loci**	**Transcripts**	**Exons**	**Clustered Transcripts**
CriGri1(NCBI)	26,261	33,202	209,449	27,759
C_griseus_v1.0 (NCBI)	23,936	31,446	212,608	26,154
CHOK1GS (Ensembl)	20,744	27,284	206,757	22,548
CriGri1 (Ensembl)	25,924	32,761	207,872	27,428
PICR	19,518	26,212	209,960	21,640

This experiment show the most complete genome is the CriGri1 (NCBI) reference sequence as the most gene loci and transcripts are assembled when it is used. The Chinese hamster C_griseus_v1.0 genome however has the most exons assembled. The results from this experiment do not account for duplicated regions of the genome (both paralogous genes and those resulting due to assembly errors) which cause reads to map in multiple gene loci resulting in duplicated transcripts. To overcome this limitation, the assembled transcripts from these experiments were converted into the FASTA format using gtf_to_fasta and clustered using *CD-Hit-EST*. Parameters were selected to cluster transcripts of over 0.90 sequence and length identity. Clustering of transcripts revealed the CriGri1 genome enables the assembly of the most unique transcripts, followed by the Chinese hamster C_griseus_v1.0. As the CHOK1GS and PICR genomes have considerably less transcripts assembled, these are not appropriate for use in our alternative splicing experiment which will rely on the assembly of novel transcripts.

A1.2.6 Differential expression using the de novo annotation generated using Stringtie

In Chapter 2 and Chapter 3 we will be performing differential gene expression and differential transcript expression analyses and therefore the identification of the optimal genome for use in these experiments is required. *Stringtie* will be used to assemble novel transcripts using the existing reference as a guide. However, in this section we perform differential expression analysis using *DESeq2* and *ballgown* on de-novo annotation generated in the previous section using *Stringtie* to assess the potential of each genome prior to using the existing annotation as a guide. The results of this experiment are presented in Table A6. The CriGri1 genome had both the most genes and transcripts identified as differentially expressed. The genes assembled in this experiment do not however have any gene symbol annotation or functionality assigned to them and therefore little biological inference could be attained from these results.

Table A6 – The use of the de novo transcriptome assemblies for differential gene and transcript expression. *Stringtie* was used to assemble transcripts on each genome which were used as reference in *DESeq2* and *ballgown* differential expression experiments. Differentially expressed genes with ≥±1.5 fold change and ≤0.05 P value are reported.

Genome	**DESeq2 Upregulated**	**DESeq2 Downregulated**	**Ballgown Upregulated**	**Ballgown Downregulated**
CriGri1(NCBI)	1,874	3,192	858	1,038
C_griseus_v1.0 (NCBI)	1,802	3,038	707	830
CHOK1GS (Ensembl)	1,475	2,699	512	707
CriGri1 (Ensembl)	1,830	3,137	783	972
PICR	1,417	2,599	602	808

A1.2.7 Which genome annotation is most improved by transcriptome assembly?

The RNA-Seq data generated in our experiment is from temperature shifted CHO cells lines which have not previously been studied using a high sequencing depth and therefore present an opportunity to identify novel transcripts which were previously uncharacterised in previous releases of the genome annotations. In Table 2.6 we presented an overview of the current stage of genome annotation however the genome sequences are also not at equivalent stages of completeness so this experiment will reveal which of the genome sequences have the potential to be the best annotated using our data. We investigate this by performing transcriptome assembly on each of the genomes using their existing reference as a guide then using *gffcompare* to identify the number of novel transcripts and gene loci (Table A7 & A8). Results of this experiment revealed the CriGri1 Ensembl genome had the most potential to be improved by assembly, however, the CriGri1 NCBI genome had the most transcripts and genes annotated both before and after assembly and is therefore a good candidate reference genome for use in further experiments.

Table A7 - Transcripts present in gene annotation before and after assembly with *Stringtie.* All of the genomes were annotated with over 20,000 novel transcripts. The Ensembl CriGri1 was most improved by assembly, however the NCBI CriGri1 contained the most annotated transcript both before and after assembly.

Genome	Annotated transcripts	Novel transcripts	Total	Average transcript length
CriGri1 (NCBI)	45,821	24,404	70,225	2,444
C_griseus v1.0 (NCBI)	43,576	23,357	66,933	2,417
CHOK1GS (Ensembl)	32,555	22,472	55,027	2,314
CriGri1 (Ensembl)	34,439	27,702	62,141	2,270
PICR	24,948	24,753	49,701	2,314

Table A8 - Gene loci annotated before and after assembly with *Stringtie.* Each of the genomes annotations were considerably improved by assembly with over 9000 novel genes annotated per sample. The CriGri1 ensembl genome was most improved by the assembly process but the CriGri1 NCBI genome contained the most gene loci annotated before and after assembly.

Genome	Annotated genes	Novel genes	Total
CriGri1 (NCBI)	28,835	13,627	42,462
C_griseus_v1.0 (NCBI)	28,270	11,459	39,729
CHOK1GS (Ensembl)	24,922	10,076	34,998
CriGri1 (Ensembl)	26,266	14,436	40,702
PICR	24,792	9,032	33,824

A1.2.8 Assessment of genomes best for the use in differential transcript expression experiments post transcriptome assembly?

To ensure the genome with the most assembled gene also provides the most informative view of differential transcript expression, *ballgown* was used to identify differentially expressed transcripts using each of the *Stringtie* assemblies using both existing reference annotation and novel transcripts. The number of differentially expressed transcript isoforms and the number of genes with differentially expressed transcripts were recorded (Table A9). The number of annotated genes with significant differential transcript expression which were not identified as differentially expressed at the gene level using *DESeq2* were also recorded, as were the genes with both an up and down regulated transcript isoform but no overall gene expression change. Use of the CriGri1 (NCBI) genome resulted in the identification of the most differentially expressed transcript isoforms.

Table A9 – The use of *ballgown* to identify differential transcript expression in assembled gene annotations. Differentially expressed transcripts with ≥±1.5 fold change and ≤0.05 P value are reported in this table for each of the 5 tested genomes. The number of genes which contained a differentially expressed transcript are also reported. The 'isoform only' columns contain the number of genes identified with differential transcript expression with *ballgown* but no differential gene expression with *DESeq2*. Some of the genes with isoform only differential expression had both an up and down regulated transcript isoform – reported as the number of genes with isoform switching. The CriGri1 (NCBI) genome had the most differentially expressed transcripts in all categories.

Genome	Isoforms Up	Isoforms Down	Genes Up	Genes Down	Isoform Only Up	Isoform Only Down	Isoform Switching
CriGri1 (NCBI)	1,611	1,734	1,095	1,003	553	388	77
C_griseus v1.0 (NCBI)	1,348	1,389	926	884	442	341	46
CHOK1 GS (Ensembl)	669	809	377	536	160	166	10
CriGri1 (Ensembl)	1,191	1,309	589	646	301	240	27
PICR	633	754	407	585	144	142	9

A1.2.9 Genome selection

The aim of this section was to identify the Chinese hamster/CHO genome out of those with available reference annotation which would be the most informative to use throughout experiments in this thesis. In order to benchmark the genomes, basic annotation and assembly statistics were assessed and each of the genomes were used in differential gene and transcript expression experiments, with and without the use of the available reference genome. The CriGri1 genome was identified as the most complete and has the highest numbers of existing gene and transcript annotations. Use of the CriGri1 genome enabled the identification of the most genes and transcripts in assembly experiments, with and without use of reference annotation, and the identification of the most significantly differentially expressed genes and transcripts. The CriGri1 genome is therefore the optimal choice of reference for differential gene and transcript experiments in later sections of this thesis. The ENSEMBL version of the genome annotation will however be used rather than the NCBI annotation due the more biologically informative view of the transcriptome it provides. ENSEMBL annotations contain annotations of 5' and 3’ UTRs of transcripts, stable gene identifiers and enables identification of orthologous genes and protein domains through the usage of Biomart.