Capture-seq and small RNA-seq to identify noncoding RNAs in the mouse ribosomal RNA gene repeat intergenic spacer

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

In

Genetics

At Massey University (Albany), New Zealand

Jessica Leigh Fitch

2018

Abstract

Cancer is a leading cause of mortality in developed countries. Due to the genetic and epigenetic heterogeneity of this disease, we still don't have effective long-term therapies for many cancers. A characteristic of many cancer cells is an alteration in the structure of the nucleolus - the primary location of the ribosomal DNA (rDNA). The rDNA encodes ribosomal RNA, which is the major structural and catalytic component of ribosomes – the cellular machinery responsible for protein biosynthesis. Accordingly, the rDNA and its transcription is a key regulator of cell proliferation. Despite this critical role, the highly repetitive nature of the rDNA has made it difficult to study, thus it remains an attractive target for anti-cancer therapies. Indeed, the promising anti-cancer drug, CX-5461, developed by our collaborators, targets the rDNA through the inhibition of the rDNA dedicated RNA polymerase I (currently in clinical trials).

In preliminary experimentation, there is a dramatic change in expression of non-coding RNAs (ncRNAs) from the rDNA during the transition to malignancy. Although the function of rDNA ncRNAs is almost entirely unknown, ncRNAs from other regions of the genome have a multitude of regulatory functions, including involvement in cancer. We hypothesise that these transcripts play a role in malignancy and CX-5461 sensitivity.

Utilising a mouse B-lymphoma model (Eμ-myc), we first applied a high throughput hybridisation-based RNA-sequencing approach (capture-seq), to enrich for rDNA intergenic spacer (IGS) ncRNA transcripts within 11 cDNA sequencing libraries. Regions of transcription throughout the IGS were identified using several bioinformatic tools, and qPCR was performed to validate transcription status as well as assess for CX-5461-dependent transcriptional changes. We also utilised other bioinformatics tools, to predict small RNAs arising from the IGS and other regions of the Eμ-myc genome, and briefly assessed their response to CX-5461 treatment. miRNAs of interest were assessed for potential pathway targets using several bioinformatic targets. Lastly, we aimed to further characterise the Eμ-myc model. With this, we assessed efficacy of methods that could be used for downstream knockdown/over expression analysis.

Overall, using the capture-seq method we identified 8 major clusters of exons (known as exon cluster groups), that were consistently predicted between RNA library preparations. These were confirmed to be transcriptionally active by qPCR, with one of these clusters. Additionally, we identified several sites in the mouse rDNA IGS that may express small RNAs, with small RNA reads aligning to these sites with some consistency between library preparations. Some of these, due to presence and absence patterns in either CX-5461 treated or control libraries, may show some signs of treatment-dependent differential expression. We also identified miRNAs from other regions of the genome which show similar patterns. We assessed potential small RNAs for gene target enrichment. No pathways/cellular components appeared to be biologically significant . We assessed a method of viral-mediated gene knockdown in a number of cell lines, which did not show efficacy in the mouse lines we had available.

In conclusion, if these exons produce ncRNAs that contribute to malignancy, the ncRNAs will form attractive new targets for therapy, independently or in combination with CX-5461, and could be used as diagnostic and prognostic markers of cancer. The future trajectories of this project include selecting promising IGS transcripts, particularly those differentially expressed, to confirm their size by northern blot. Then, to assess their role in malignant cells, to perform knockdown/overexpression assays and assess cellular response. Further, we would target the rDNA ncRNAs in several cancer and non-cancer cell lines, to broaden our understanding of anti-cancer application.

Acknowledgements

I'd like to extend my sincerest gratitude to my supervisors, Austen and Sebastian, for their on-going support and guidance throughout my masters. I have always been one to ask a lot of questions, some better than others, so I have appreciated your patience. My confidence in my own ability has greatly improved over the last few years, and they (as well as all my colleagues and friends) are the reason behind it. I will always be immensely grateful.

I'd also like to thank the entire Ganley Lab for their help, by means of suggestions, critique and some laughs along the way; it has been a pleasure being part of the team. Particularly, thank you Daria Chudakova and Diksha Sharma for all your help and advice in the lab and around computers. A lot of the work completed was completely foreign (and in some cases extremely intimidating). I wouldn't have been able to finish this without you.

Thank you to my parents, family and friends for all your love and encouragement over the last few years. Your support (in all the many shapes and forms) will always be cherished.

Lastly, thank you to all our new friends at The University of Auckland. Special acknowledgements to Liam Williams and Kristine Boxen, who day and night work tirelessly sequencing and helping with genomics equipment. Your tips and tricks have been fantastic, and it has been great working with you.

Contents

List of tables

List of Figures

List of abbreviations

Unit abbreviations

µl/ µlg : micro-litre/-gram

G: gram

Hrs: hours

L: litre

Mins : minutes

ml/mg: mili-litre/-gram

nl/ng: nano-litre/-gram

Frequently used abbreviations

cDNA: complementary DNA

IGS: intergenic spacer

miRNA: micro-RNA

mRNA: messenger RNA

qPCR: quantitative PCR

rDNA: ribosomal DNA

RIN: RNA integrity number

RNA pol I : RNA polymerase I

RNA-seq: RNA- sequencing

shRNA: short hairpin RNA

1.Introduction

1.1 A brief introduction to cancer and current cancer therapies

Cancer is a group of devastating diseases, killing millions worldwide independent of socio-economic status, with these numbers continuing to grow as global populations age ((Torre et al., 2015)). Cancer is characterised by cells that exhibit abnormal growth (which can result in the formation of tumours) and metastatic invasion ability into other tissues (Fidler, 2003; Stratton, Campbell, & Futreal, 2009). Cancers, specifically tumours, generally show high levels of genetic heterogeneity (Burrell, McGranahan, Bartek, & Swanton, 2013; Fidler, 1978). Not only can a tumour in one individual exhibit a distinct panel of oncogenic mutations different to those in another individual's tumour from the same tissue origin (inter-heterogeneity), but additionally a single tumour can also manifest genetically diverse cell subpopulations (intra-heterogeneity). Examples of common mutations in cancers include mutations in pathways associated with normal programmed cell death, like inactivation of the key pro-apoptotic transcription factor p53 (Ouyang et al., 2012), and mutations in pathways associated with promoting growth and survival, like the constitutive activation of RAS proteins (promoting survival and cell cycle progression (Downward, 2003)) or of an anti-apoptotic factor BCL2 (P. N. Kelly & Strasser, 2011).

Classical treatment methods of cancer include chemotherapy, radiation and surgery. Chemotherapy includes treatments using a variety of chemical agents which target rapidly proliferating cells. In many cases, chemotherapy agents induce cellular stress and instigate cell

death (Kaufmann & Earnshaw, 2000). Radiation, which harnesses radioactive beams or isotopes, can induce apoptosis in cancer cells through the DNA damage pathway (Eriksson & Stigbrand, 2010). Finally, surgery to physically remove tumour tissue is often combined with chemotherapy or radiation to increase the likelihood that all tumour tissues are removed. It has been shown that there is a 10% survival rate increase when surgery was coupled with chemotherapy in the treatment of gastric cancer (Sasako et al., 2011). These classical therapies have been linked to poor long-term prognosis because of their non-specific nature resulting in devastating side effects of treatment, and possibility of tumour relapse accompanied with acquired resistance to the therapy. For example, studies have shown that polynucleated cells (atypically large cells with aneuploidy greater than 2n) and cancer stem cells (cancer cells that exhibit stem-like characteristics) acquire resistance to some of the classical therapies and can lead to tumour relapse (Lu et al., 2015; Sharma et al., 2013).

Currently a new era of targeted cancer therapy using highly specific inhibitors designed to target critical oncogenic pathways has emerged (Haber, Gray, & Baselga, 2011). Such inhibitors can be selected specifically for a particular type of cancer in a mutation-dependent manner (reviewed in (Sawyers, 2004)). As an example, Cetuximab is a drug based of a monoclonal antibody targeting epidermal growth factor receptor (EGFR) which plays a role in the control of differentiation and proliferation in a number of tissues(Olayioye, Neve, Lane, & Hynes, 2000). Mutations in EGFR that result in the constitutive activation of the receptor, have been associated with aberrant cell growth and are linked to the formation and progression of several cancer types ((Hynes & Lane, 2005; Tebbutt, Pedersen, & Johns, 2013)), particularly colorectal cancers (Wong, 2005). The use of Cetuximab, either in combination with other treatments or alone, has been shown in several studies to significantly prolong *EGFR* mutant cancer patient survival ((Cunningham et al., 2004; Jonker et al., 2007)). A second example is the drug Trastuzumab, an antibody targeting human epidermal growth factor receptor 2 (*HER2* or *ERBB2,* a well characterised oncogene (Baselga & Swain, 2009)) which has been shown to be overexpressed in ~25% of breast cancers, has shown efficacy in the treatment of *HER2* mutant breast cancer resulting in prolonged survival times(Vogel et al., 2002).

Importantly, due to the rapid accumulation of a variety of mutations and chromosomal rearrangements (Duesberg & Li, 2003; Jackson & Loeb, 1998; Vogelstein & Kinzler, 2004), cancer cells recurrently acquire resistance to once effective drugs. For example in relation to the drugs mentioned earlier, Cetuximab resistance has been shown to occur through gene

amplification of *HER2* reducing the dependence of the cancer cell on the aberrant EGFR expression by inducing a bypass pathway (Yonesaka et al., 2011). Consequently, there remains a persistent requirement for the development of new targeted drugs to continue combating cancers, with many targets yet to be exploited. The next sections of my thesis will outline two novel targets in cancer therapy, the ribosomal DNA, and noncoding RNA (ncRNA).

1.2 Targeting the ribosomal DNA and noncoding RNA for the treatment of cancers

1.2.1 Ribosomal DNA (rDNA)

The eukaryotic ribosomal DNA (rDNA) encodes the ribosomal RNA (rRNA); the main structural and catalytic component of ribosomes (Long & Dawid, 1980). The rDNA is arranged in head-to-head tandemly repeated arrays, with all units being identical in a cell which is maintained by the phenomena of concerted evolution (Eickbush & Eickbush, 2007; Ganley & Kobayashi, 2007). In Eukaryotes , a single rDNA gene comprises of a 13.3 kb coding region, consisting of the 28S, 5.8S and 18S rRNA genes (together producing a 47S pre-RNA), and 30 kb noncoding spacer (known as the intergenic spacer, or IGS)(Zentner, Balow, & Scacheri, 2014). In mice, the rDNA arrays are found on the chromosomes 12, 15, 16, 18 and 19 (Ito, Tsuchiya, Osawa, Shibata, & Kanda, 2008) . Chromosome arms containing the rDNA arrays colocalise in the nucleolus, a membrane-less sub compartment of the nucleus, which is the major site of ribosome biogenesis in the cell (Schwarzacher & Wachtler, 1993). Recently, using a combination of digital droplet PCR (ddPCR) and bioinformatic methods, the mouse genome has been estimated to have around 14-200 copies, with copy number estimates varying depending on tissue type and breeding status of the strain (Xu et al., 2017).

Ribosome biosynthesis is responsible for roughly 80% of a cells metabolic expenditure; and consequently rRNA transcripts are the most abundant in the transcriptome (Kusnadi et al., 2015; Yan et al., 2017). In Eukaryotes, the 47S pre-rRNA is synthesised by RNA polymerase I (RNA pol I). RNA pol I is a 590 kDA enzyme, composed of 14 subunits (12 of which are homologous RNA polymerase II and III subunits), and has the dedicated role of rDNA

transcription(Grummt, 2003; C.-D. Kuhn et al., 2007; Russell & Zomerdijk, 2005) . Pol I transcription initiation from the rDNA promoter requires a number of co-factors, including but not limited to SL1 (RNA polymerase I selectivity factor), RRN3 (a RNA polymerase I transcription factor) and Upstream binding factor (UBF), which assist in Pol I recruitment at the rDNA promoter, transcription initiation and/or elongation at the of the 47S pre-rRNA (Jordan, Mannervik, Tora, & Carmo-Fonseca, 1996; Moorefield, Greene, & Reeder, 2000; Peyroche et al., 2000) (**Figure 1**). SL1, also known as TIF-IB, features several (TATA-binding proteins)-associated factors (or TAFs). SL1 is recruited to the rDNA promoter via several of the TAF's(Gorski et al., 2007). Acetylation of one particular TAF subunit (TAF_I68), via TFF1 (transcription termination factor 1) recruitment of an acetyltransferase to the promoter, further facilitates SL1 localisation (Muth, Nadaud, Grummt, & Voit, 2001) . SL1 then recruits RRN3 (or TIF-1A) via TAF_I63 and TAF_I110, a second transcription factor that directly interacts with RNA pol I(Miller et al., 2001). UBF, which can exist in 2 different isoforms (1 and 2), localises at rDNA and shows no specific sequence requirement for binding (Bell, Learned, Jantzen, & Tjian, 1988; Copenhaver, Putnam, Denton, & Pikaard, 1994; A. Kuhn et al., 1994). UBF has been shown to be heavily present across the mouse rDNA coding region (Herdman et al., 2017). UBF is essential for rDNA transcription, as when UBF is knocked down it can result in complete loss of localisation of RNA pol I to the rDNA promoter thus silencing of the rDNA (Herdman et al., 2017; Sanij et al., 2008). At the rDNA, UBF has additional roles in maintaining rDNA chromatin state, through the displacement of nucleosomes (see (Sanij & Hannan, 2009)).

After transcription by RNA Pol I, the 47S pre-RNA is processed down into the 18S, 28S and 5.8S units (Yan et al., 2017), that along with a number of ribosomal proteins(transcribed by RNA pol II) and the 5S rRNA (transcribed by RNA pol III(C Mayer & Grummt, 2006; Weinmann & Roeder, 1974)), fold to produce the small 40S subunit (from 18S rRNA) and the large 60S subunit (from the 28S, 5.8S and 5S), which complex with mRNA to form a mature (80S) ribosome (Pestova et al., 2001; Strezoska, Pestov, & Lau, 2000) . Alterations in essential genes or pathways required for ribosome biosynthesis has been linked to a number of diseases (reviewed in (Freed, Bleichert, Dutca, & Baserga, 2010)), emphasising the role of the ribosomal RNA and ribosomes in the maintenance of a typical cellular state.

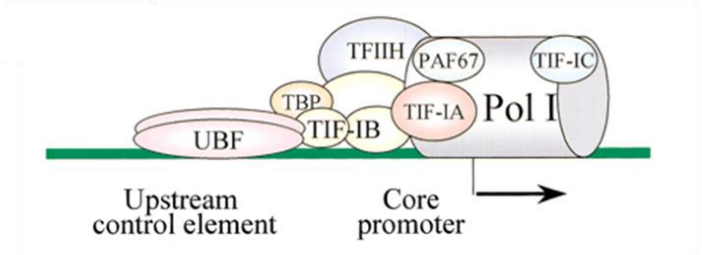


Figure 1 RNA polymerase I (RNA pol I) complex at the rDNA promoter
RNA pol I is recruited to the mouse rDNA promoter by several co-factors. Absence of some co-factors, for example UBF or TIF-1B (SL1) , inhibits/reduces pol I transcription (see text). Schematic adapted from Grummt (2003)

1.2.1.1 The rDNA intergenic spacer (IGS)

Currently, though the mammalian rDNA IGS has yet to be fully characterised, it has been shown to contain conserved elements that are important for rDNA maintenance and transcription. The IGS 5' end begins at the site of rRNA transcription termination, that in mice, is marked by the first of a series of conserved 18bp repetitive sequences that contain a Sal restriction site (hence referred to as Sal boxes) (Diermeier, Németh, Rehli, Grummt, & Längst, 2013; Grummt, Maier, Öhrlein, Hassouna, & Bachellerie, 1985). These small repeats provide a binding site for the transcription terminator factor TTF-1, which hinders RNA polymerase I transcription through preventing elongation (Grummt et al., 1985). Humans have been shown to share a Sal box-like element at a similar position (Bartsch, Schoneberg, & Grummt, 1987), which can bind both human and mouse orthologs of TTF-1, though this is not seen conversely in mice (Bartsch et al., 1987). Some Sal boxes of the IGS also function as sites for the production of a replication fork barrier (RFB). First characterised in yeast, the RFB functions as a polar barrier, preventing collisions between replication and transcription machineries by forcing unidirectional synthesis by the different polymerases (Rothstein, Michel, & Gangloff, 2000). In mice and humans, it has been shown that the DNA replication fork is hindered by Sal Box 2 with TTF-1 bound (as well as other factors) consequently acting as a RFB site (Gerber et al., 1997) (Lopez-Sánchez et al., 2014). Mutations around the Sal Box 2 sequence hinders RFB activity, supporting its role as a RFB(Gerber et al., 1997).

There has been reported to be hotspots of recombination in the Eukaryotic IGS. So far this observation has been comprehensively studied in yeast, where homologous recombination (HR) levels are increased at RFB by HOT1 (a stimulator of recombination (Lin & Keil, 1991)) in the IGS , and this is thought to be required for rDNA copy number maintenance (Kobayashi, 2011). The human IGS has been shown to also contain 9 hotspots of double strand breaks (DSB), which overlap with CTCF binding sites and active epigenetic marks (Tchurikov et al., 2014). These, more recently have been refined to 2 dominant peaks, where the highest observations of DSB are seen. It has been proposed that these DSBs may play a role in transcriptional control (Tchurikov et al., 2013) , as there have been links with coordinated gene expression of gene clusters defined both DSB sites with PARP1 binding sites (a regulator of gene expression, (Kraus, 2008)).

Within the mouse IGS, there is evidence of a spacer promoter around 2kB upstream of the transcription start site. This spacer promoter shows significant sequence similarity to the rRNA coding region promoter, and produces a noncoding transcript which is processed to around ~150-300 nt (the promoter RNA or pRNA, more information on noncoding RNAs in section *1.4.2*)(A. Kuhn & Grummt, 1987). In mammals, the epigenetic state of the rDNA promoter is established by the nucleolar remodelling complex (NoRC), which is associated with inactive rDNA repeat promoters (Santoro, Li, & Grummt, 2002). It has been found these pRNAs , recruits the NoRC complex and poly(ADP-ribose)-polymerase-1 (PARP1) to the promoter, along with an additional factors including TFF-1 (that binds to the T_0 site upstream of the promoter) , which in turn induces silencing of the rDNA unit(Guetg, Scheifele, Rosenthal, Hottiger, & Santoro, 2012; Christine Mayer, Schmitz, Li, Grummt, & Santoro, 2006; Zhou et al., 2009). This relationship is necessary for maintenance of inactive and active chromatin states with each round of cell division (Guetg et al., 2012). Therefore, this pRNA is essential for maintaining rDNA activity states within the rDNA.

Along with the spacer promoter RNA, the IGS of different organisms have been shown to produce other noncoding RNA transcripts (see *1.3.4* for information on noncoding RNA). Two IGS transcripts, IGS1-F and IGS1-R, have been well characterised in yeast. These are transcribed bidirectionally from the yeast rDNA promoter within IGS1 (Houseley, Kotovic, El Hage, & Tollervey, 2007). It has been suggested that these IGS transcripts (specifically IGS1-R) plays a role in recruiting Trf4 (part of the TRAMP complex which facilitates exosome degradation of

abnormal rRNA) to help stabilise rDNA copy number (Houseley et al., 2007). Currently the human rDNA IGS has three well characterised ncRNAs (IGS_{16}, IGS_{22} and IGS_{28}), that were identified to be differentially expressed upon sensing different environmental stresses (Audas, Jacob, & Lee, 2012). Upon stress, some proteins are immobilised to the nucleolus to prevent their functionality (Mekhail et al., 2005). Collectively, these IGS transcripts play a role in localisation of some proteins that contain the nucleolar detention sequence (or NoDs) to the nucleolus. IGS_{28} (encoded ~28kB in the rDNA IGS), shows temporary transcription upon acidosis, and this transcription resulted in VHL (a tumour suppressor that is activated upon sensing oxygen (Mekhail, Gunaratnam, Bonicalzi, & Lee, 2004)) binding at this site of the IGS. IGS_{16} and IGS_{22} (encoded at ~16kB and ~22kB of the IGS respectively) showed temporary transcription upon heat shock, and consequently induced the localisation of Hsp70 (a factor that prevents heat-induced apoptosis (Mosser, Caron, Bourget, Denis-Larose, & Massie, 1997)) at both their respective transcription origins of the IGS. Knockdown IGS_{28} by shRNA showed reduced or abolished localisation of VHL to the nucleolus in an acidic environment. A similar observation was seen in the knockdown of IGS_{22}, where again Hsp70 localisation to the nucleolus was hindered after heat shock. Together this highlights the ncRNA-based roles of protein localisation in the nucleolus. Currently, aside from the presence of the spacer promoter, there is little known about mouse IGS transcripts.

1.2.1.2 The rDNA and cancer

Abnormal nucleoli morphology is a common phenotype in many cancer cell types(Quin et al., 2014), and consequently there has been a link to changes in the rDNA in cancer cells. One valuable model to study changes in the rRNA in malignancy is the mouse B-lymphoma Eμ-myc model. This model is based off the observation that in Burkitt B cell lymphoma, the c-*Myc* oncogene (normally located on chromosome 8), often translocates to a immunoglobin μ loci (Dalla-Favera et al., 1982; Taub et al., 1982) . The transgenic mice of this model were designed to mimic this observation, and have the *c-Myc* oncogene overexpressed under the immunoglobulin promoter which consistently induces the development lymphoid tumours from birth (Harris et al., 1988). The *Myc* oncogene encodes c-Myc protein which is a central transcription regulator, and along with an essential co-factor Max binds to target E-boxes DNA sequences to influence transcription ((Conacci-Sorrell, McFerrin, & Eisenman, 2014)). Gene targets transcriptionally regulated by the Myc-Max complexes have been extensively studied((Dang, 2012) (Pelengaris, Khan, & Gerard, 2002)), and include factors involved in proliferation,

stem-cell state maintenance, and DNA replication. Interestingly, c-Myc is a key factor in the control of ribosome biogenesis. It is required for RNA polymerase I -dependent transcription, and regulates is via binding at active rDNA promoters with SL1 (Arabi et al., 2005b; Grandori et al., 2005). Due to the varied but largely growth inducing nature of c-Myc targets, c-Myc has been shown to be upregulated in 50% of cancers, and consequently plays a dominant role in cancer establishment and maintenance (Dang, 2012).

Using the Eμ-Myc model (as well as other models not described here), significant findings have been made establishing a dominant role of changes in rDNA in cancer. Changes in the nucleolus in cancer has been attributed to increased rates of rDNA transcription, accelerating ribosome biosynthesis, which is required of highly proliferating cells (Hein, Hannan, George, Sanij, & Hannan, 2013; Montanaro, Treré, & Derenzini, 2008). In normal cells, rRNA transcription rate is regulated by altering the availability of RNA pol I transcription factors, or gradually through epigenetic activation/inactivation of rDNA units (see (Grummt, 2003; McStay & Grummt, 2008). Increased rates of rRNA synthesis in cancer may result from a multitude of cellular changes, including overexpression of c-Myc allowing for rapid SL1 and UBF recruitment of RNA pol I (Arabi et al., 2005a; Dang, 2012; Grandori et al., 2005), or epigenetic alterations in rDNA units forcing more into a transcriptional active state (Ghoshal et al., 2004). Using the Eμ-myc model, it has been shown that deregulated rRNA synthesis is enough to maintain an oncogenic phenotype, as the loss of RNA pol I alone by either small molecular inhibitors or RNA interference can induce apoptosis in some cancer cells (Bywater et al., 2012). In preliminary data, our collaborators have also seen an increase of transcription from the Eμ-myc rDNA IGS as the cells transition from a pre-malignant to a malignant phenotype (Prof. Ross Hannan, unpublished results, personal communication).

In normal cells, it has been shown that generally half the rDNA repeats are transcriptionally silent or pseudo-silent. Silent repeats are silenced through epigenetic mechanisms, and pseudo-silent repeats have a transcriptionally active chromatin state but lack UBF binding, consequently hindering RNA pol I recruitment and transcription (Sanij & Hannan, 2009). As Eμ-myc cells transition into more malignant phenotype, more repeats enter a transcriptionally active state (Prof. Ross Hannan, unpublished results, personal communication), suggesting a requirement for either a more open rDNA chromatin state or increased numbers of active repeats in malignant cells.

With the correlation of changes in rDNA chromatin state and transcription levels and a cancer cells survival and proliferation, rDNA has become an attractive anticancer therapy target.

1.2.1.3 Targeting ribosomal RNA synthesis with small molecule inhibitor, CX-5461

Researchers in recent years have developed small molecule inhibitors specific to RNA pol I, for a targeted cancer therapy. One drug candidate is CX-5461, discovered using a cell-based screening assay, inhibits RNA pol 1 transcription by blocking the SL1 subunit binding and consequently preventing formation of an active pre-initiation complex (Drygin et al., 2011). Currently in clinical trials, CX-5461 treatment results in either cancer-cell specific senescence and autophagy, or apoptosis (depending on tissue type). CX-5461-driven senescence and autophagy, common in solid tumour types, has been shown to be independent of p53 status (Drygin et al., 2011). CX5461-induced apoptosis is generally p53-dependent, a result of either nucleolar stress signalling (typically through release RPs binding to MDM2 prevent p53 degradation) or G2 -cell cycle arrest (via ATM/ATR pathway) (Bywater et al., 2012; Negi & Brown, 2015; Yan et al., 2017). p53 independent apoptosis has also been shown in cell lines with mutant p53 status, but this mechanism has yet to be characterised (Negi & Brown, 2015) . CX-5461 has been comprehensively characterised in the Eμ-myc model. That is of particular interest because B-cell lymphomas and other Myc-driven haematological cancers generally exhibit resistance to many conventional chemotherapy/radiation treatments (Hein et al., 2013; Tallman, Gilliland, & Rowe, 2005).

CX-5461 efficacy in targeting cancer cells was originally hypothesised to be purely a result of cancer cells dependence on elevated levels of rRNA transcription for ribosome biogenesis and protein biosynthesis. However, Eμ-myc cells can be rescued from CX-5461 induced apoptosis by the overexpression of anti-apoptotic protein BCL2 and continue growing normally even in presence of CX-5461(Bywater, Pearson, McArthur, & Hannan, 2013). Thus, CX-5461 effect on cancer cells is not a solely result of ribosome shortage. Consequently, our group and our collaborators are testing a number of alternative hypothesis which may explain cancer cells sensitivity to CX-5461. One hypothesis is that CX-5461 may change transcription of ncRNAs located in the rDNA IGS, which may be drivers of malignancy.

1.2.2 Non-coding RNAs

Noncoding transcripts are not translated into a functional protein, and are generally transcribed at low levels (Derrien et al., 2012; Mercer, Dinger, & Mattick, 2009). These transcripts can be produced in a number of ways, including from alternatively spliced mRNAs, transcription of the non-template strand, and transcription from coding regulatory regions (i.e. promoters and enhancers) (Kapranov et al., 2007). Given that the large proportion of the mammalian transcriptome has been shown to be noncoding (Mattick, 2005), often very few are characterised. This is often due to their low expression, lack of consistent motifs for bioinformatic screening, and variability in size, stability and location within/around coding regions (Clark et al., 2012; Dinger, Pang, Mercer, & Mattick, 2008; Hon et al., 2017). Bioinformatic approaches can predict whether a transcript is non-coding using several criteria. For example, as mRNAs generally are encoded in a longer open reading frame (ORF), so transcripts mapping to short ORFs may be noncoding (Dinger et al., 2008). Similarly, as generally nucleotide patterns are not-random in coding RNAs, the absence of a variety of motifs may suggest the transcript is noncoding (Housman & Ulitsky, 2016).

NcRNAs can be categorised into two groups, small and long, with each group having distinct cellular roles. Small non-coding RNAs typically range between <20-200 bp, include microRNAs (miRNA), Piwi -interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and short interfering RNAs (siRNAs). These have been shown to often have significant regulatory roles in cellular gene expression, particularly siRNAs and miRNAs. miRNAs are produced from hairpin precursors, and siRNAs from long double stranded DNA; and both carry out RNA interference (RNAi) gene regulation as part of the RNA induced silencing complex (RISC) (Valencia-Sanchez, Liu, Hannon, & Parker, 2006). MicroRNAs and siRNAs regulate target mRNAs through perfect or partial hybridisation of a seed sequence site within the mRNA, typically in the 3'UTR (Djuranovic, Nahvi, & Green, 2011). When bound, they are shown to promote silencing via several mechanisms. Binding between the miRNA/siRNA and mRNA in some cases has been shown to allow recruitment of endonucleases to induce RNA cleavage and consequently post transcriptional silencing(Valencia-Sanchez et al., 2006). Evidence has shown that some miRNAs may impede mRNA cap recognition, for example by preventing binding of eIF4E translation

initiation factor, consequently silencing through translational repression (L. Wu & Belasco, 2008). Due to their well characterised mechanism of targeting, miRNAs and siRNAs have been exploited within a laboratory setting, where they are used to knockdown target transcripts of interest to assess cellular response.

Long non-coding RNA's (or lncRNAs) are noncoding transcripts greater than 200bp in length. In reference to the Wang & Chang 2011 review, lncRNA can be broken into 4 main types based on their functionality (though some could feature in more than one) : signals, decoys, guides and scaffolds (Wang & Chang, 2011). Signals encompass long noncoding RNAs which signal molecular events, often also exerting some functional role in the event. A well-established example is the lncRNA XIST, which plays a role in DNA imprinting (Costa, 2008). The lncRNA XIST is transcribed from only the second X chromosome (Xi) in females during X inactivation. After transcription, it has been shown to coat the Xi chromosome , consequently signalling the recruitment of the Polycomb protein complex for H3K27me3 mediated silencing (Plath et al., 2003). XIST, along with its antisense equivalent Tsix, work together to mediate the X inactivation process; where Tsix actively represses XIST on the active copy thus preventing the X inactivation signal (Navarro, Page, Avner, & Rougeulle, 2006; Ohhata et al., 2015). Lee (2000), found the deletion of *Tsix* disrupted maternal X imprinting by losing the XIST antagonist function, and further lead to the silencing of all X chromosomes in the early mice embryo(J. T. Lee, 2000). Consequently, this highlights the coordinated roles in the signalling and initiating X-inactivation in normal cells.

The second class, decoys, can be defined as lncRNAs which function in competitive binding/interference scenarios. An example is the control of the human dihydrofolate reductase (*dhfr*) gene, required for the formation of metabolites like amino acids (Chen et al., 1984), which is regulated by a negative feedback loop with a lncRNA. This noncoding transcript is transcribed 5' upstream from the *dhfr* start site from a minor promoter (Martianov, Ramadass, Barros, Chow, & Akoulitchev, 2007). Upon transcription, it acts as a decoy by competitively binding at the promoter and sequestering core transcription factors , consequently repressing *dhfr* transcription (Blume, Meng, Shrestha, Snyder, & Emanuel, 2003; Martianov et al., 2007).

The third class of lncRNAs are the guides lncRNA, which guide proteins or protein complexes to a site in either *cis* or *trans*. The lncRNA HOTAIR, which is a well-characterised lncRNA guides in *trans*. It has been shown to play an active role in many types of cancer, in that its increased

expression guides the polycomb repressor complex 2 (PRC2) localisation, changes chromatin states to resemble a fibroblast-like phenotype allowing for greater metastatic capabilities (Gupta et al., 2010; Kaneko et al., 2010)

The final, scaffold lncRNAs, bring multiple genomic regions together to exert their functionality. An example is the lncRNA *Kcnq1ot1*, which via 3C studies (an assay which captures chromosomal interactions via cross-linking and assessment of crosslinked chromatin (Simonis, Kooren, & De Laat, 2007)), was shown to act as a scaffold for the formation of an interchromosomal loop between KvDMR1 and the *Kcnq1* promoter (Zhang et al., 2014). It has been shown to be required for the maternal methylation and consequent imprinting by PRC2, where loss of imprinting prevents specific parent-associated gene expression (Thakur et al., 2004). Loss of this noncoding RNA was shown to inhibit imprinting, and is frequent in individuals with Beckwith-Wiedemann syndrome(M. P. Lee et al., 1999), suggesting its crucial role in this particular epigenetic interaction.

1.2.2.1 Noncoding RNAs role in cancer

Noncoding RNAs are emerging as key players in disease initiation and progression, with focus particularly in cancer(Iorio & Croce, 2012; Meltzer, 2005; Prensner & Chinnaiyan, 2011). In cancers, ncRNAs have been shown to play two general roles, in which they can either act as tumour suppressors or promote the oncogenic phenotype.

The miRNA miR-200 family have been shown to be upregulated in some mouse mammary tumours lines with characteristic highly invasive (metastatic) phenotypes (Dykxhoorn et al., 2009), and consequently is an example of a ncRNA that promotes cancer. miR-200 transcripts can repress the ZEB family protein production (a group of transcription repressors) which in turn upregulates ZEB family targets including E-cadherin, required for cell to cell adhesion (Brabletz & Brabletz, 2010). This upregulation facilitates metastasis by transitioning migrating cancer cells back into an epithelial-like state for secondary tumour formation(Park, Gaur, Lengyel, & Peter, 2008). Tumours with upregulation of the miR-200 family showed greater metastatic rate than those without (Dykxhoorn et al., 2009).

The lncRNA PCA3 is another example of a ncRNA that has been shown to promote cell proliferation and an oncogenic phenotype. PCA3 (sometimes referred to as DD3) is a long

noncoding RNA, encoded in an antisense orientation within an intron of the PRUNE2 gene on chromosome 9 in humans (Clarke et al., 2009). For some time, it has been used as an effective biomarker for prostate cancer (Bussemakers et al., 1999; Hessels et al., 2003). More recently, it has been established as a regulator of PRUNE2 expression, where decreased PRUNE2 expression is correlated increased cell proliferation capabilities(Salameh et al., 2015) . PCA3 regulates PRUNE2 through forming a dsRNA hybrid, which is then proposed to act as a guide for ADAR (adenosine deaminase acting on RNA) members for RNA editing of PRUNE2(Salameh et al., 2015) . PCA3 is upregulated in prostate cancers which are dependent on survival signalling through the androgen receptor (AR). PCA3 has been shown to be directly upregulated through androgen stimulation, suggesting that PCA3 may play a pivotal role in AR dependent prostate cancers (Ferreira et al., 2012).

An example of a characterised ncRNA with a tumour suppressor activity is GAS5. This lncRNA is shown to be frequently downregulated in several cancer tissues and is often linked to a poor prognosis (Cao, Liu, Li, Zhao, & Qin, 2014; Mourtada-Maarabouni, Pickard, Hedge, Farzaneh, & Williams, 2009; Sun et al., 2014). Overexpression of GAS5 leads to decreased cell invasion and increased apoptosis (Mourtada-Maarabouni et al., 2009; Yin et al., 2014). Specifically, it has been shown that GAS5 induces cycle arrest and induction, through binding to the DNA binding domain of the glucocorticoid receptor and inhibiting its function (Kino, Hurt, Ichijo, Nader, & Chrousos, 2010), where activation of the glucocorticoid receptors have been linked to promoting cell growth (W. Wu, Pew, Zou, Pang, & Conzen, 2005).

2. Project aims

Currently, aside from the promoter RNA, there is little known about the transcriptional status of the mouse rDNA IGS. We hypothesize that there are noncoding RNA transcripts located within the mouse rDNA IGS that have yet to be identified. Preliminary data showed that there is transcription from the Eμ-myc rDNA IGS which increases when these cells transition into a more malignant phenotype. Additionally, we hypothesize the IGS ncRNA transcripts may be differentially expressed upon treatment with CX-5461. Utilising the Eμ-myc model, this project aims to achieve the following:

1) **To identify mouse noncoding transcripts transcribed from the rDNA IGS, which are not currently described in the literature**
2) **To assess transcriptional response of these rDNA IGS ncRNAs to treatment with CX-5461**

To achieve these aims, we will perform both capture-seq and RNA-seq on RNA from both CX-5461 treated and untreated Eμ-myc cells. Utilising a bioinformatic approach, we will identify regions of transcription within the mouse rDNA IGS. Validation of their expression and assessment of differential expression will be performed using qPCR.

3. Materials and Methods

3.1 Culturing mammalian suspension cell lines (Eμ-myc) *in vitro*

Frozen suspension aliquots of Eμ- myc cell variants (**table 1** for more information) were defrosted at +37° C until liquid, then diluted 1:10 with suspension cell culture media (Gibco DMEM 1x media (Life Technologies, ref 1960-004), supplemented with 10% Fetal Bovine Serum (FBS, Mediray) , 1% 100mM sodium pyruvate (Sigma, CAS number 113-24-6), 1% 100x Pen Strep (Gibco, Ref 15140-122), 1% 100x Glutamax (Gibco, ref 35050-061), 0.1% 1000x β-Mercaptoethanol (Gibco, ref 21985-023), and 0.1% 100mM L-asparagine (Sigma, A4284) . Cells were grown at +37 °C (10% CO_2) in 75 cm^2 flasks. To maintain growth, suspension cell lines were passaged every 2 days (1:10 cell suspension to culture media), at which point they have become highly confluent (dense). Cell viability was measured using a hemacytometre, using a 1:1 ratio of 0.4% Trypan Blue solution (Gibco, cat 15250061) which permeates dead cells. Live cell count per ml was calculated by the standard protocol:

$$\frac{cells}{ml} = live\ cells\ counted \times \left(\frac{dilution\ factor}{squares\ counted}\right) \times 10^4$$

Suspension cells were prepared for harvesting, by first centrifuging culture at 1,500rpm for 10 mins to pellet. The supernatant was removed, and the pellet washed in 1 µl of 1x PBS and pelleted as before. Remaining supernatant was then removed from the pellet. Cells were either cold killed (-80 °C for 10 min) for immediate use or combined with a lysis buffer for extraction. Pellet/lysis buffer mixtures were stored at -80 °C until extractions could be carried out.

Table 1 Mammalian cell lines used in experiments

Cell line	Acronym in text	Specifics of line	Source
Eμ-Myc 4242 GFP LMP	4242	Mouse B-cell Lymphoma Line, with LMP vector carrying a green fluorescent protein (GFP). Suspension cells	Dr Ross Hannan, Australian National University
Eμ-Myc 4242 shUBF GFP LMP	shUBF	Mouse B-cell Lymphoma Line, with LMP vector carrying a green fluorescent protein (GFP) and a	Dr Ross Hannan, Australian National University

		short hair pin to UBF. Suspension cells	
MEF NIH 3T3	MEF	Mouse Embryonic Fibroblasts. Adherent cells	Dr Evelyn Sattleleger, Massey University Albany & Dr. Jo Perry, University of Auckland
HEK293T	HEK	Human Embryonic Kidney. Adherent cells	Dr Ross Hannan, Australian National University
MDA-MB-231	MDA	Human Breast cancer line. Adherent cells	Dr Evelyn Sattleleger, Massey University Albany & Dr. Jo Perry, University of Auckland

3.2 Culturing adherent cell lines

Adherent cells (**table 1**) were grown in adhesion culture media (Gibco 1x DMEM , 10% FBS, 1% Glutamax and 1% Sodium pyruvate) in a 75 cm^2 flasks at +37 °C (5% CO_2), and passaged every 2 days. To passage or harvest, cells were washed with 1x PBS, detached using 1x trypsin (Gibco, 15090-046)(trypsin exposure time differed per cell line), and trypsin was deactivated using adhesion culture media. HEK cells were passaged 1:10 every 2 days at which point they were densely packed on flask surface, while MEF and MDA cells were passage between 1:1 - 1:4 cells every 2 days to reach a similar confluency. Confluency were determined visually by microscope.

3.3 Cytotoxicity assay

Cytotoxicity assays assessing CX-5461 efficacy and to determine IC_{50} was performed using 4 different concentrations of CX-5461 (125 nM, 12.5 nM, 1.25 nM and 0.125 nM) from a 25 µM stock solution in DMSO. 4242 and/or shUBF cell suspension were seeded into 96 well plate, 100µl/well at a seeding density of $1x10^4$ cells/well. In duplicates, 100 µl of media containing CX-5461 or 100 µl of media with the corresponding volume of DMSO per treatment concentration was added to 4242 and shUBF cells.

200 µl Suspension culture media (blank)	100 µl Cell suspension + 100 µl of media with DMSO volume corresponding to 125 nM CX-treatment	100 µl Cell suspension + 100 µl of media with DMSO volume corresponding to 125 nM CX-treatment replicate	100 µl Cell suspension +100 µl DMSO with 125 nM CX-5461 in DMSO	100 µl Cell suspension + 100 µl 125 nM CX-5461 replicate in DMSO

This format was repeated for all CX-5461 concentrations and for each cell variant. All remaining wells were filled with 100µl of culture media only.

The plates are incubated for 72 hrs at +37 °C (10% CO_2). 20µl of Resazurin Blue, (prepared by dissolving 3mg of Resazurin salt (Sigma-Aldrich, K7017-1G) in 27.15ml 1x PBS) was added to all wells and left for 5 hr at +37 °C (10% CO_2)a, and the plate was read for cell viability (fluorescence) using an EnVision plate reader , with 540-570 nm excitation wavelength and 580-610 nm fluorescence emission settings (J. Li, Zhang, Ward, Prendergast, & Ayene, 2012).

The average fluorescence between replicates were taken for all concentrations of treatment and corresponding DMSO control, and background fluorescence (media blank) was subtracted using Excel. Cell viability for each CX-5461 concentration was then calculated as percentage of the corresponding DMSO control, and a trend line was plotted (CX-5461 concentrations (nM) vs Cell viability(%)). The IC_{50} was calculated from the trend line, as the value of X (CX-5461 treatment) when Y (cell viability) was 50%.

3.4 Treating cells with CX-5461

4242 and shUBF cell lines were passaged for several weeks before treatment. For CX-5461 treatment 4242 and shUBF cells were seeded into 6 flasks each at seeding density of 1-2 x 10^5 cell/ml. 50 µM of CX-5461 was added to 3 flasks with both cell lines (CX-5461 treatment), equal volumes of DMSO were added to the remaining 3 flasks (DMSO control). Cells were incubated for 3 hrs at growth conditions and harvested following the standard protocol (section *3.2.1*).

3.5 Preparing RNA for Capture-Seq

3.5.1 RNA extractions for total RNA

RNA was extracted from harvested cells using both the Machery-Nagel Nucleospin RNA (Ref 740955.50, lot 1604/003, referred in the text as Nucleospin), and Qiagen miRNeasy Mini Kit (Cat 217004, Lot 154010383, referred in the text as miRNeasy). Extractions were performed following the manufacturers protocols, and RNA was eluted in PCR grade water and stored at -80C until use.

RNA samples were concentrated using the method outlined in (Walker & Lorsch, 2013). Briefly, 0.1x volume of 3M sodium acetate (Sigma) was mixed with each RNA sample. 2.5x volume of absolute ethanol (~99.9%, EMSURE) was added, and left overnight at -20 °C. RNA is then pelleted via centrifugation at 12,000 xg for 15 mins. The supernatant was removed, leaving behind ~20 µl to ensure the pellet is not disturbed, and washed with 70% ethanol for 2 mins. The RNA was pelleted again via centrifugation for 2 mins at 12,000 x g, ethanol removed using a pipette and the pellets are left to further dry at room temperature for ~25 mins. Samples were then resuspended in 10 µl of RNase free water and stored at -80 °C until use.

3.5.2 Ribodepletion of total RNA

Ribodepletion on total RNA was carried out using the Invitrogen Ribominus Transcriptome Isolation Kit, human/mouse (Cat K155002), following the protocol outlined by the manufacturers, with input differing depending on sample and extraction attempt. Ribodepleted RNA was precipitated following the Ribominus kit protocol, without the addition of glycogen.

3.5.3 Measuring RNA quality

Quality of RNA pre and post depletion was measured using Agilent Bioanalyser and the RNA 6000 Nano kit. The chips were prepared using the RNA 6000 Nano Kit Quick Start Guide, using 1 µl of RNA per well. Quality was assessed by concentration, and the RIN (RNA integrity number) output, which measures the ratio of the 28S to the 18S rRNA peaks.

3.5.4 DNA extraction

DNA was extracted using GF1- Blood DNA extraction Kit (Vivants, GF-BD-100 Lot #12224c), following the online protocol with some minor modifications. Briefly, cells were harvested and mixed with the kit lysis buffer and proteinase K and left to incubate at +65 °C for 5 mins following instructions. The sample was combined with 20 µl of RNAse A (10mg/ml) and

treated at +37 °C for 30 mins. 200 µl of absolute ethanol was added, and the sample was halved into two different columns, and washing was carried out following the kit protocol. DNA was eluted from the columns using PCR grade water, and the final elution were combined and measured on a nanophotometer (IMPLEN, N60 model).

3.5.5 DNA sonication

Extracted DNA was sheared using the Covaris M220 Focused-Ultrasonicator and microTUBE AFA Fiber Snap-Cap (PN 520045). DNA was aliquoted into 8 equal volumes, and each aliquot was sheared using a different default setting, which when all aliquots were pooled together resulted in a range of DNA fragment lengths between 200-1500bp. The same sonication tube was used for all fragmentation protocols and remaining small volume of DNA left in the tube produced the range of smaller fragments. To concentrate, the pooled fragmented DNA was combined with 1x volume 100% isopropanol and 0.1x volume of 3M sodium acetate (Sigma, 126-96-5), and left to precipitate for 48 hrs at – 20 °C. The mix was centrifuged at 13,000x rpm for 45 mins at +4 C, the supernatant was removed then the pellet washed with 500 µl of fresh 70% ethanol. This was left for an additional 30 mins at -20 °C to re-precipitate. The washed pellet was centrifuged for 30 mins at 13,000x rpm, all supernatant removed, and pellet was dried in a shaking heat plate at +21 °C. The pellet was then dissolved in 30 µl of PCR grade water and stored at -20 °C until use.

Sonication efficiency of DNA was assessed by gel electrophoresis. 1 µl of pooled fragmented DNA was run on 1% agarose SB gel at 100V for 40 minutes with a 1kB ladder (Thermo scientific, SM0311). The gel was visualised on a Gel Doc EZ Imager (Bio-Rad) , and DNA fragments of a desired size range (200-1500 bp) were detected. DNA concentration post-precipitation was measured using a N60 nanophotomer (IMPLEN).

3.6 Western blot analysis for analysing UBF knockdown efficiency

4242 and shUBF cell lines were grown and harvested in replicates (following *3.2.1*). Cell pellets were lysed using SDS lysis buffer (see *3.1.2*) and RIPA buffer (Cell Signalling Technology, cat 9806), both with 1x Protease Inhibitor Cocktail (Sigma-Aldrich/Merck, 11697498001/11836145001), and left on ice for 40 mins. Cell debris was pelleted via centrifugation at 13000x RPM for 15 mins, and the supernatant (cell lysate) transferred to a clean tube.

Pierce BCA Protein Assay Kit (Thermofisher) was used to measure total protein concentration. BSA protein standards were prepared following the manufactures protocol. 25 µl of A-I protein standards were loaded into a 96 well plate. Dilutions of the 4242 and shUBF cell lysates in the two different buffers (SDS lysis buffer and RIPA) were used, with all volumes being made to a total of 25 µl with buffer. 200 µl of 50:1 reagent A: B from the Pierce BCA protein kit (Thermofisher, cat 23227) was added to all wells (protein standards and samples), and the plate was left to incubate at +37 °C for 30 mins. Protein concentrations from samples were measured via EnVision plate reader (absorbance at 562nm).

A protein standard curve, with a line of best fit, was produced using absorbance readings plotted against known concentrations of the protein standards. Protein concentrations (in µg/ml) in samples were estimated using their absorbance reading and the line of best fit. Volumes with equal protein concentrations for 4242 and shUBF samples were made to a final volume of 10 µl using SDS lysis buffer. Protein extracts were combined 1:1 with 2x Laemmli buffer and incubated at + 95 °C for 5 mins. The mixtures were transferred into wells of a precast gel (Bio Rad, cat 4561083), along with a 5 µl protein ladder lane (Precision Plus Protein dual colour standards, Bio-Rad cat#161-0374), and run in 1x SDS running buffer at 100 V for 1hr (or until ladder bands were well separated) in a Mini-PROTEAN Tetra Vertical Electrophoresis Cell (Bio-Rad, 1658005). Using an electrotransfer method, proteins were transferred from the gel to an Immunoblot PVDF membrane (Bio-Rad, 162-0177), which was activated with methanol and washed in 1x transfer buffer. Elecrotransfer was carried out in the same vertical gel electrophoresis cell, using the Mini Trans-Blot Module (Bio-Rad, 1703935), with 1x Transfer buffer (*3.1.2*) at 200Ma for 1 hr. Next, membrane was incubated in 1x TBS and 5% dry milk overnight at +4 °C then washed in TBS-T. Membrane probing was performed by diluting rabbit anti-UBF specific antibody (kind gift from Dr. Ross Hannan, Australian National University) diluted 1:200 in 1x TBS-T with 0.05% dry milk , and incubating overnight at +4 °C. The membrane was washed with 1x TBS-T for 15 mins, followed by incubation for 2 hrs with the secondary HRP-conjugated anti-Rabbit antibody (cat# 31458, Thermo-Fisher), diluted 1:10000 in 1xTBS-T + 5% dry milk . After incubation, the membrane was washed for an additional 30 minutes in 1x TBS-T and visualised using the ECL-Plus Western Blotting Substrate (ThermoFisher, 32132) and the Amersham Imager 600 machine (GE Healthcare Life Sciences, cat 29083461). For a loading control, additional probing with anti-

alpha Tubulin (Thermo-Fisher, cat# A11126) or anti-histone H2A (abcam, Cat# ab13923), was carried out on the same membrane, following the same protocol.

3.7 Quantitative polymerase chain reaction (qPCR) for shUBF knockdown efficiency

4242 or shUBF RNA concentrations were measured using the nanophotometer, and equal concentrations were achieved via dilution with PCR grade water. cDNA was synthesised in replicates using the SensiFast cDNA synthesis kit (Bioline, cat BIO-6506#), following the manufactures protocol.

Two UBF primer sets which could distinguish UBF1 and UBF2 isoforms were designed using Geneious software. Sequences are shown in **appendix 1**.

qPCR assays were performed using the Sensifast Sybr Hi-ROX (Bioline, Cat BIO-92005/92020) and PerfeCTa SYBR green Supermix (Quantabio, cat 95054-100) and the 7900HT Fast Real-Time PCR system (Thermofisher, cat 4329001) in a 384 well plate. The default settings were used (see manual for details: https://tools.thermofisher.com/content/sfs/manuals/cms_042252.pdf). cDNA concentrations were estimated using the nanophotometer, and 4242 and shUBF cDNA were diluted to equal concentrations. qPCR reactions were set up in duplicates following the manufacturers protocol for each qPCR kit, for all UBF primer sets (Mouse UBF1 isoform specific, Mouse UBF 1 and 2 isoforms (long product),Mouse UBF 1 and 2 isoforms (short product)) using both 4242 and shUBF cDNA. Additional qPCR reactions (for both 4242 and shUBF cDNA) were performed using GAK and GAPDH primers, used for normalisation as housekeeping genes (sequences are shown **appendix 1**).

Relative gene expression change was calculated as fold change of treatment (or UBF knockdown) relative to control normalised to housekeeping gene expression, following the standard 2^{-ddct} method (Livak & Schmittgen, 2001).

3. 8 cDNA synthesis, library preparation and hybridisation-based enrichment for lncRNAs and small RNAs.

Capture-seq library preparation and capture for sequencing was performed using the SeqCap RNA Developer Enrichment Kit (Roche, cat 07279213001), following the SeqCap RNA Enrichment System User Guide (Roche Nimblegen, http://netdocs.roche.com/DDM/Effective/07279337001_RNG_SeqCapRNA-UGuide_v1p0.pdf). The following protocol briefly outlines the steps of the procedure outlined in the guide, with the major steps being summarised in **Figure 2.**

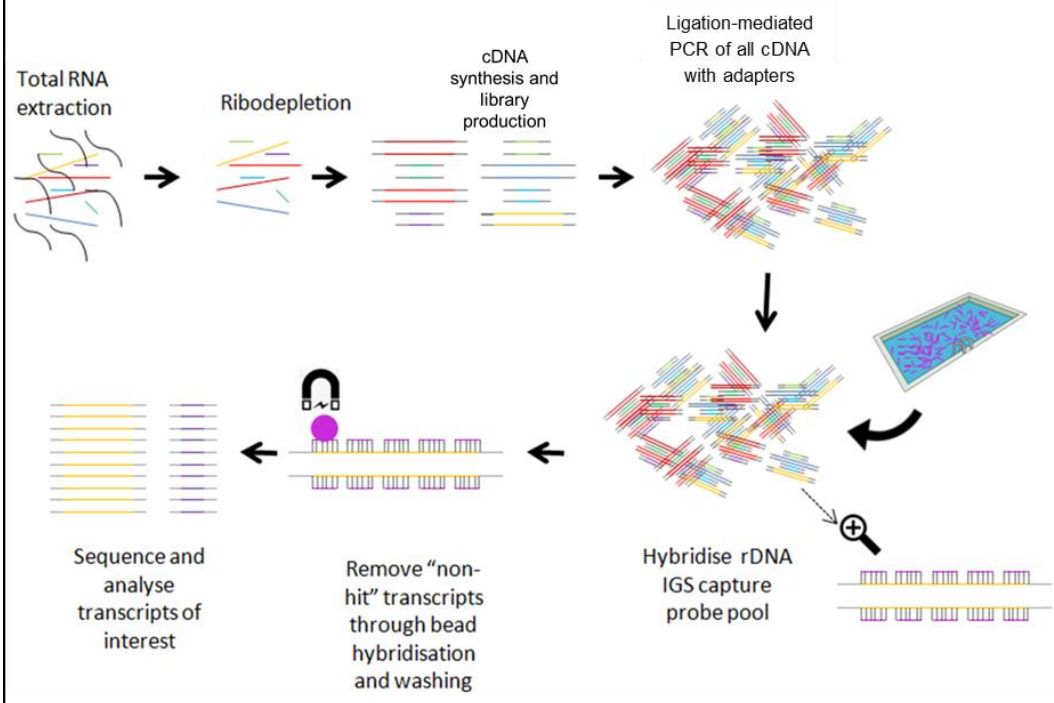


Figure 2 Schematic summarising RNA extraction, Ribodepletion and the SeqEZ RNA Enrichment System User Guide
Schematic summarising the steps before and during the capture protocol (SeqEZ RNA Enrichment System). Before the SeqEZ RNA enrichment system is used, total RNA is extracted and ribodepleted to reduce amount of rRNA transcripts. The SeqEZ RNA Enrichment system protocol starts with RNA being converted into cDNA and cDNA is prepared into libraries (via A-tailing and adapter/oligo ligation steps). Ligation mediated PCR (lm-PCR) is performed, priming from adapters that have been ligated to the end of the cDNA. cDNA transcripts are captured via hybridisation to probes targeting a region of interest. Captured cDNA transcripts are bound to beads, and non-targeted cDNA is washed away, leaving behind an enriched sample of cDNA from the target region. Enriched samples again undergo lm-PCR in preparation for sequencing.

100ng (or the maximum available amount) of CX-5461 treated or control RNA samples from 4242 or shUBF cells (see section *3.5.*2) were aliquoted into fresh microcentrifuge tubes. ERCC spike in mixes (Thermofisher, cat 4456740) were diluted and added in volumes corresponding to pre-ribodepletion concentrations (see section 4.1.2.1 and 4.1.2.2 for specifics), and ERCC spike in mix 1 was added to 4242 DMSO control RNA samples and shUBF CX-5461 treated RNA samples, and ERCC spike in mix 2 was added to 4242 CX-5461 treated RNA samples and shUBF control DMSO RNA samples.

Following the SeqCap RNA Enrichment System User Guide, and briefly described here, RNA samples were mixed with 10 µl of 2x Fragmented Primer and Elute Buffer and fragmented at + 94 °C for 8 mins. The RNA was synthesised into first and second strand cDNA using the KAPA Stranded RNA-Seq Library Preparation Kit (KAPPA Biosystem, Cat KK8400). A sample of 100ng of sonicated DNA (*3.5.*5) was included from this stage, and all underwent the following cleaning, A-tailing, adapter ligations steps, where a different adapter was used per sample replicate. 11 cycles of ligation-mediated PCR (LM-PCR) was performed using the 2X KAPA Hifi Hotstart Ready Mix and 10X KAPA library primer mix (found in the cDNA synthesis kit), and the resulting lm-PCR amplified libraries were cleaned using Agencourt beads included in the SeqCap Pure Capture Bead Kit. cDNA integrity was measured on the Agilent Bioanalyser, using the DNA high-Sensitivity Bioanalyzer Chip and kit (Agilent, 5067-4626). cDNA was diluted 1:10 in water, with 1 µl loaded into wells.

Samples were then concentrated using the Savant DNA 120 concentrator, set at auto and medium temperature (+65 °C) until completely dried, and resuspended in 4.2 µl of PCR grade water. These were pooled equally (83.3ng/sample) to make a final pooled cDNA library of 1 µg in a 50 µl total volume.

The pooled cDNA library was enriched for IGS noncoding RNA using the hybridisation approach as outlined in the SeqCap RNA Enrichment manual. A probe pool was designed specifically by Roche to enrich for transcripts from the mouse IGS, independent of synthesis strand origin (see **Figure 3**). The hybridisation step was carried out according to the manual and left for 20

hrs. Targeted cDNA was enriched using the provided Seqcap capture beads. The beads and the pooled hybridised cDNA library were incubated together for 45 mins at +45 °C/heated lid + 57 °C .Un-bound cDNA was washed away. The IGS enriched multiplex library was amplified using the LM-PCR method provided in the protocol, using the primers included within the SeqCap EZ Accessory Kit v2, for a total of 14 cycles. The final enriched pooled library was cleaned using AMP XP Beads included in the kit and eluted in 50 µl of PCR-grade water. Quality of the captured library was determined using an Agilent DNA High sensitivity chip (diluted 1:10) and the Bioanalyser. Sequencing was carried out on the MiSeq Illumina platform, with paired-end reads (250x2) by New Zealand Genomics ltd (NZGL).

3.9 cDNA synthesis and library preparation for <120 bp small RNAs.

RNA extracted using the miRNeasy kit in *3.5.1* from 4242 and shUBF CX-5461 treated and untreated cell was submitted to a NZGL for small library construction using the NEXTflex Small RNA-Seq V3 kit (Bioo Scientific, Cat 5132-05/6). The libraries were sequenced on the MiSeq platform with single end reads 50 bp x 1 by NZGL.

3.10 Bioinformatic analysis Capture long ncRNA RNA-Seq Multiplex libraries

3.10.1 Producing reference genome

All bioinformatic analysis was carried out on Biolinux (version x86_64). The latest mouse primary assembly reference genome (last modified 3/6/17) was downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-88/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz). A mouse rDNA unit sequence generated in our lab was used as the rDNA reference (unpublished). The ERCC sequences were downloaded from (https://tools.thermofisher.com/content/sfs/manuals/cms_095047.txt) and were converted into the FASTA format.

The rDNA sequence was converted from lower case to upper case using

```
less Mouse_rDNA_unit_corrected_Feb2013.fasta | tr 'a-z' 'A-Z' >
Mouse_rDNA_unit_corrected_Feb2013-uppercase.fasta
```

The ERCC sequence and rDNA sequence were added to the mouse reference genome using cat command:

```
Cat /path/to/ERCC_Controls_Annotation.fasta
/path/to/Mus_Musculus_GRCm38_Primary_Assembly.fasta
/path/to/mouse_rDNA_sequence.fasta > Mus_Musculus_GRCm38_rDNA+ERCC
```

3.10.2 Indexing reference genome

The newly produced reference genome was indexed using STAR (ver 2.5.3a, see (Dobin et al., 2013)), following the manual and with the command:

```
Nohup STAR
 - -runThreadN 1
 - -runMode genomeGenerate
 - -genomeDir
 - -GenomeFastaFiles
```

Where genomeDir and genomeFastaFiles specify the path to the output directory and path to genome FASTA files respectively. Read length was left as default (100), due to variation in average read lengths between libraries.

3.10.3 Aligning Capture-seq reads to reference genome, and sorting/cleaning alignment outputs using Samtools

Reads were trimmed by NZGL using the BBmaps tool. Trimmed reads were aligned to the indexed reference genome using STAR aligner as follows:

```
Nohup STAR
- - runThreadN 2
- - genomeDir /path/to/indexed/genome
 - - readFilesIn /path/to/R1/trimmed/reads/ /path/to/R2/trimmed/reads/
```

```
 - -outSAMstrandField intronMotif
- -readFilesCommand zcat
```

Where outSAMstrandField specifies unstranded data, and zcat avoids issues with zipping.

STAR aligner output SAM format files were converted into BAM and sorted using Samtools (ver 1.4.1, see (H. Li et al., 2009)) as follows:

```
Samtools view -Sb Aligned.out.sam > Aligned.out.bam
```

And

```
Samtools sort -O bam -o Aligned.out.sorted.bam Aligned.out.bam
```

Samtools was used to clean BAM files to remove low quality and unmapped reads , as follows

```
 samtools view -q 255 -f 2  -h -b Aligned.sorted.bam >
Aligned.cleaned.sorted.bam
```

3.10.4 Using Stringtie to assemble reads into Transcripts

Transcripts were assembled for each library independently with Stringtie (ver 1.3.3, see (Pertea et al., 2015)) , following the manual and incorporating the ERCC GTF file (from Thermofisher website https://www.thermofisher.com/order/catalog/product/4456739) using the BAM sorted aligned reads as follows

```
Stringtie
/path/to/library/Aligned.out.sorted.bam
-p 2
-o /path/to/output/directory/libraryname_stringtie.gtf
```

```
-G ERCC_Annotation. GTF

-l libraryname_transcripts

-A libraryname_abund.tab
```

Where -o specifies output name, -p specifies thread number, -m specifies minimum transcript length , -l specifies gives the assembled transcripts a reference name, and -A produces a transcripts abundance file. Transcript GTF files produced by Stringtie were visualised using Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir, Robinson, & Mesirov, 2013).

3.10.5 Determining per base coverage levels of the rDNA IGS to assess for areas of high transcription

Using the aligned bam output, read coverage per base of the entire rDNA unit was found using Bedtools (version 2.19.0 (Quinlan, 2014)) as follows

```
~/Bedtools genomecov

-ibam Alignedout.sorted.bam

-g GenomeFastaFile.fa

-d

> library#_bedtools_coverage.txt
```

The rDNA was extracted using:

```
grep "RDNA" library#_bedtoolsOutputFile.txt >
library#_rDNAcoverage.txt
```

Each library rDNA coverage file was plotted in Rstudio (ver 3.3.3 x86_64-w64-mingw32/x64 (64 -bit) using the ggplot2 package (ver 2.2.1), with script below (as an example)

```
a <- read.table("library_DNA_IGS.txt")
```

```
names(a) <- c("genomeID","basePosition","depth")

head(r)

ggplot(data=a, aes(x=basePosition, y=depth))+

  geom_line(colour="black")+

  xlab("rDNA Base Position (bp)")+

  ylab("Coverage Depth")+

  ggtitle("library  IGS coverage plot")+

  scale_x_continuous(breaks = seq( 0, 45400, by = 4000))+

  scale_y_continuous(breaks= seq ( 0, 10000, by = 800))
```

The start of the rDNA IGS was determined by using BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) to find the first Sal box sequence (Grummt et al., 1985). The rDNA IGS was then extracted from the full rDNA unit bedtools output and plots were created as previously with the full rDNA unit, aside the altering the X and Y axis limits to fit the new data.

To produces graphs comparing different libraries, a variation of the following script was used (here specifying 3 libraries).

```
d <-read.table("library1_rDNAcov_IGS.txt")

e <- read.table("library2_rDNAcov_IGS.txt")

f <- read.table("library3_rDNAcov_IGS.txt")

names(d) <- c("genomeID","basePosition","depth")

names(e) <- c("genomeID","basePosition","depth")

names(f) <- c("genomeID","basePosition","depth")

ggplot()+

  geom_line(data= d, aes(x = basePosition, y = depth), colour="blue")+

  geom_line(data= e, aes(x = basePosition, y = depth), colour="red")+
```

```
  geom_line(data= f, aes(x = basePosition, y = depth),
colour="green")+
  xlab("rDNA Base Position (bp)")+
  ylab("Coverage Depth")+
  ggtitle("replicates rDNA IGS coverage plot")+
  scale_x_continuous(breaks = seq ( 13400,45400, by = 5000))+
  scale_y_continuous(breaks= seq ( 0, 5000, by = 500))
```

Average coverage plots of triplicates/duplicates for each treatment condition were made by taking the average coverage value of triplicates/duplicates within a treatment scheme at each base of the rDNA IGS, and these values were plotted using the geom_line function in ggplot2 by altering the single line script above to plot 4 lines (script not shown).

Average coverage values were normalised to the DNA sample (library 12) by dividing the coverage value at each base position of the cDNA libraries rDNA IGS by the corresponding base position in the DNA library. These were again plotted as geom_line function in ggplot2 by a variation of the 4-line script above (script not shown), using an arbitrary log scale for the Y axis.

3.10.6 Comparing IGS aligned read numbers to reads aligning to the whole genome in the DNA-derived library to estimate theoretical enrichment potential

Each chromosome from the DNA library bedtools file was extracted individually using

```
Awk '{if ($1== "chromosome"){print$1,$2,$3}}' S12_cleancov.txt >
chromosome.txt
```

Where *chromosome* refers to an individual chromosome number. Large chromosomes were split into smaller files using the following split command

```
Split -l number of lines    chromosome.txt
```

As before, the rDNA IGS bedtools coverage values were extracted from the full rDNA unit using Excel. In R, the average coverage of each chromosome (or rDNA IGS and coding region) was taken using the following functions

```
Chromosome <- read.table("chromosome".txt)

Mean(chromosome$V3)
```

The average coverage of the rDNA IGS compared by the average coverage of the rDNA coding region (or the rest of the genome), to give the theoretical enrichment potential of the system for enriching for the rDNA IGS.

3.10.7 Normalising between libraries using ERCC spike ins

Using the ERCC Controls analysis document (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_095046.txt), the actual concentration of each ERCC transcript within each ERCC spike in mix (1 or 2) put into the library preparation (in Attomole/µl) could be calculated by the following equation

$$Actual\ ERCC\ concentration\ in\ sample\ \left(\frac{attomole}{\mu l}\right) = Known\ concentration\ of\ ERCC\ x \times \frac{Volume\ added\ (\mu l)}{Dilution\ factor}$$

Where the known concentration of each ERCC is specified in the ERCC analysis documents. ERCC transcripts and their corresponding FPKM values were extracted from the full Stringtie GTF file, and each ERCC transcript FPKM value was plotted against their attomole/µl concentration in R using ggplot2 geom_points and geom_smooth following the script

```
#plotting a library

library <- read.table("Library.txt")
```

```
names(library) <- c("ERCC_ID","FPKM","Conc")


#defining the regression line

fit_library <- lm(library$FPKM ~ library$Conc)

#finding the coefficients (y intercept, and slope)

slope <- fit_library$coefficients[[2]]

y_intercept<- fit_library$coefficients[[1]]


#r-squared

R2<- signif(summary(fit_library)$r.squared)


#call to give values

R2

Y_intercept

slope


#plotting dotplot and linear regression line and paste values on
graph

ggplot(library, aes(x=Conc, y= FPKM)) + geom_point()+

  xlab("ERCC concentration(atto/ul")+

  ylab("FPKM")+

  scale_x_log10()+

  scale_y_log10()+

  geom_smooth(method ="lm", col="colour")+

  ggtitle("library ERCC plot")+
```

```
  annotate("label", x=1000, y=1, label= "R^2: value, slope:value
, y_intercept: value")
```

The difference in ERCC (FPKM vs concentration) slopes between library preparations were used for normalisation of Stringtie Exon coverage.

3.10.8 Normalising exons to the captured-DNA to reduce effect of capture bias

Exons from RNA derived libraries were normalised to the DNA (cDNA exon coverage/DNA) derived library to account for capture and sequencing bias. Average DNA coverage from the DNA captured library spanning across each exonic location was calculated from the bedtools output using R and the script as follows

```
#set table

DNA <- read.table("captured DNA library")

#set names of columns

Names (DNA) <- c("genomeID","basePosition","depth")

P <- DNA [c(exon start position: exon end position)]

Mean (p$depth)
```

Exons were ranked from highest to lowest coverage.

3.10.9 Producing a Repeatmasker GTF file for the mouse rDNA IGS

The mouse rDNA FASTA file was loaded into the RepeatMasker Webserver (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) (Smit, 2013-2015). The output was converted into a GTF file in Excel.

3.11 Designing and optimising qPCR primers to validate expression of IGS exons

3.11.1 Assessing efficiency/specificity of IGS exon qPCR primers and the preliminary qPCR testing validating exon transcription

Using the Geneious software (ver 10.2.3, (Kearse et al., 2012)), primers were designed to amplify within IGS exon clusters. Amplicon sizes were designed to be preferentially between 100-200bp, and primers aimed to have similar melting temperate (T_m), GC content, and lack any significant end symmetry to avoid concatenation. Primers were synthesised by Integrated DNA technologies (IDT), and sequences can be found in **Appendix 1** (Capture seq-exons).

Primers were resuspended to produce a stock concentration of 100µM in water. 10 µM working stock concentrations of primer sets were validated on extracted Eµ-myc DNA. Primer efficiency of each primer set was assessed by qPCR using PerfeCTA Sybr green or Sensifast-SYBR Hirox qPCR mix on the 7900H qPCR machine with default settings (section 3.7). Eµ-myc genomic DNA that was diluted 10-fold four times, and each dilution was amplified using default thermocycler settings (section 3.7) with each primer set (in duplicates) in a 10µl reaction volume. Efficient primer sets were classified as having equal (preferably 3) cycle numbers between dilutions, with no/minimal water amplification . Primers were also assessed for off-target amplification using conventional PCR. A master mix of the Ex-Taq Hotstart kit (TaKaRa, cat RR001) or KAPPA 2G Rhobust Hotstart kit (sigma,KK5023) was prepared following the kit instructions. 1µl of Eµ-myc DNA was combined with the master mix, and the forward and reverse primers for each primer set and amplified following the kit protocol. The output was run on a 1% Agarose 1x SB gel for 30 mins (100v) and stained and destained in ethidium bromide (8µg/ml) for 5 mins each. The gel was visualised using a Gel Doc EZ Gel system (BIO RAD). Non-specific primer sets showed more than one band >100bp and were not used downstream.

For qPCR, cDNA synthesis was performed using the SensiFast cDNA synthesis kit, following the kit protocol, on RNA from CX-5461 treated and untreated 4242 cells. qPCR was performed on the 7900H qPCR machine with default thermocycler settings (section 3.7), using the LUNA Universal qPCR mix (New England Biolabs, cat #M3003) or SensiFast SYBR Hi-ROX kit (Bioline, cat BIO-92005) in replicates on cDNA from CX5641-treated and untreated cells using all IGS exon primer sets. GAPDH and GAK were used as a housekeeping genes for normalisation , where relative expression was calculated using the standard 2^{-ddct} method (described above). Differences in Ct values were used to assess for transcriptional differences in DMSO control and CX-5461 treated samples in IGS regions where ncRNA transcripts were predicted.

3.11.2 Testing for presence of gDNA contamination

Presence of gDNA contamination was tested by treating 1µg of RNA with 1µl RNAse A (stock concentration 10 mg/ml) for 30 mins at +37 C. 1µg of RNase treated RNA and 1µg of the corresponding RNAse-untreated RNA were used for cDNA synthesis (Sensifast cDNA synthesis kit) and qPCR (Sensifast SYBR Hi-ROX kit) on the 7900H qPCR following standard methods as mentioned previously, along with a water (negative control).

3.11.3 RNA extraction and final to validate transcription from predicted transcribed regions of the mouse rDNA IGS

4242 cells were seeded in four 75cm^2 flasks (9X10^5 cells per flask) and treated with CX-5461 or DMSO control following the standard treatment protocol described in *3.4*. RNA was extracted from cell pellets using the Nucleospin RNA kit (as outlined in *3.5.1*), and RNA concentration was quantified using a nanophotometer. cDNA was synthesised using the Sensifast cDNA synthesis kit, with 1µg from each extraction being used as input. 500ng of output cDNA (in replicates) from each extraction was used as template for qPCR, amplifying with capture exon primer sets 11-19 (**appendix 1**) using the SensiFast SYBR Hi-ROX qPCR kit on the 7900H qPCR machine with default settings (section 3.7. A water control for each primer set was incorporated as a negative control.

To estimate base level of gDNA contamination in RNA samples, 1 µg of RNA from each RNA extraction was diluted to 20 µl (the same volume as reaction volume for cDNA synthesis). This

was a No Reverse Transcription negative control sample (NRT). The equal volumes of cDNA sample and NRT sample were analysed by qPCR with IGS exon primer sets 11-19 using the SensiFast Hirox Sybr qPCR kit on the 7900H qPCR machine with default settings (section 3.7). Ct values were compared for both samples.

3.11.4 Assessing rRNA transcription changes with CX-5461 treatment via qPCR

cDNA and gDNA contamination control outputs from section *3.11.5* were amplified using Mouse ETS primer set (**appendix 1**) using 7900H qPCR and the SensiFast Hirox Sybr qPCR kit following previously described methods for qPCR set up.

3.12 Bioinformatic analysis of small (>120bp) RNA-seq data

3.12.1 Quality assessment and trimming raw reads

Raw sequencing data of libraries produced in *3.9* was first quality checked using FastQC (0.11.5) (Andrews, 2010). The Illumina small RNA 3' adapter sequence was downloaded from the Illumina support page (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-02.pdf) , and converted into FASTA format. 3' adapter sequences were trimmed using Trimmomatic (ver 0.36, (Bolger, Lohse, & Usadel, 2014)) from all libraries, following the online manual, and using the command as follows specifying for single end reads

```
Java -jar Trimmomatic-version.jar SE -phred33 data.fastq
/path/to/ouputDirectory/data_trimmed.fa ILLUMINACLIP:
small_RNA_3prime_adapter.fasta:2:30:10
```

The adapter trimming success was confirmed using FastQC. The rDNA unit was added to the mouse genome as an extra chromosome using a cat function as before.

3.12.2 Aligning and visualising small RNA reads to identify potential small RNAs

3.12.2.1 Aligning small RNA reads using STAR

The reference genome was indexed using STAR using the script outlined in 3.10.2. Trimmed reads were aligned to the indexed reference using STAR (2.5.3a) with the command, suggested by both the developer and in other published work (Khurana et al., 2017), as follows

```
STAR

--runThreadN 3

--genomeDir /path/to/STAR_INDEX/

--readFilesIn /path/to/trimmed_data.fa

--outFilterMismatchNoverLmax 0.05

--outFilterMismatchNmin  16

 --outFilterScoreMinOverLread  0

 --outFilterMatchNminOverLread 0

  --alignIntronMax 1
```

Samtools was used to convert aligned SAM to aligned sorted BAM following the protocol in *2.10.3*, and mapped reads were extracted using the command as follows

```
Samtools view -F 4 sorted.bam > cleaned_sorted.bam
```

Aligned sorted bam files produced in section 3.12.2 were visualised in IGV. Small RNA candidates from the rDNA IGS were selected due to having the following features

1. A read was present in more than one library that exhibit the same start site/length
2. Reads showed no variation/polymorphisms from the reference sequence
3. Reads were not multimappers (white bars in IGV)

3.12.2.2 Finding gene/pathway targets of potential rDNA IGS miRNAs

Seed sequences were predicted using nucleotides 2-8 for the small RNA candidates reads fitting the criteria from section *3.12.2*. Seed sequences were assessed for gene targets using Targetscan Custom(Friedman, Farh, Burge, & Bartel, 2009), with mouse as the selected species. Target scan gene outputs were adopted as inputs for gene enrichment analysis using the Gene Ontology Consortium online resource (Ashburner et al., 2000; Gene Ontology Consortium, 2017), and specifically executing the "Go-Slim Biological Processes" and "Go-Slim Cellular Component" functions. Those with a p value of ≥0.05 were considered significant.

3.12.3 Finding miRNA from the rDNA IGS using Bowtie and mirDeep2

3.12.3.1 Indexing the reference genome using Bowtie

The reference genome built in *3.12.1.1* was edited to not include spaces in the FASTA file title using the function

```
 Awk '/^/{print">1"++i; next } { print } ' original_reference.fa >
modified_reference.fa
```

Bowtie (1.0.0) , was used to index the modified reference file

```
Bowtie-build
modified_reference.fa referenceName
```

3.12.3.2 Using miRDeep2 to identify miRNA transcripts within the data

miRDeep2 ver 2.0.0.8-pl5.22.05((Friedländer et al., 2008)) mapper function was used to compress all trimmed small library reads (from *3.12.1.1*) with the mapper.pl executor using the command as follows

```
mappler.pl
sample_trimmed.fq
```

```
-e
-h
-j
-m
-s sample_collapsed.fa
```

miRDeep2 mapper function was then used to map the compressed reads against the Bowtie indexed genome using the command as follows

```
mapper.pl
sample_collapsed.fa
-c
-p path/to/indexedReference/directory/referenceName
-t sample_readsCollapsed_vs_genome.arf
```

mirDeep2 was used to find known and novel miRNA within the data using the following command

```
miRDeep2.pl
sample_collapsed.fa
modified_reference.fa
sample_readsCollapsed_vs_genome.arf
none
none
none
-t Mouse
2>sample_report.log
```

Where each “none” represents no GTF input of known miRNAs of the species , no GTF input of known miRNAs of a related species, and no GTF input of known hairpin precursors for the species, respectively.

3.12.3.3 GO enrichment of Targetscan target genes of miRNA identified by miRDeep2

Gene targets and GO enrichment analysis (of the gene targets) of miRNAs identified using miRDeep2 was performed following the description in *3.12.2.2*, using seed sequences (2-8) of the predicted in the mature miRNAs. GO slim pathways/components were considered significant using the same threshold.

3.13 Lentiviral transfection and infection optimisation

3.13.1 Lentiviral transfection of HEK and MEF cells

Lentiviral transfection and infection was carried out in three cell lines (see **table 1** for cell line details) using the Lenti-X shRNA Expression system and the pLVX-shRNA2 vector with an RFP kindly provided by Dr. John Taylor and Carol Wang (University of Auckland, School of Biological Sciences).

HEK packaging cells were grown to 80 % confluency overnight at +37 °C in 6 well plates (5% CO_2), and the growth media was removed carefully. Following the protocol for Lipofectamine 2000 transfection (https://tools.thermofisher.com/content/sfs/manuals/Lipofectamine_2000_Reag_protocol.pdf) , the vector and expression system packaging plasmid were mixed in 200µl of OptiMEM media (Gibco, cat 11058-021) in a 1.5ml centrifuge tube (total of 3.69 µg vector and expression system per well). In a separate tube, 7.38 µl of Lipofectamine 2000 (Thermofischer, 11668027)was diluted in 200 µl of OptiMEM media in a 2:1 ratio (to vector & plasmid DNA). Each mixture was left to incubate at room temperature separately for 5 minutes, and then the mixtures were combined (total of 400 µl) and left to incubate for an additional 15 minutes. The total 400 µl of Lipofectamine 2000 with vector/plasmid solution was diluted again in OptiMEM to make a total of 3ml and added to the 6 well plates for 4-hr incubation (+37 C). Following incubation, transfection media was removed, and replaced with equal volumes of growth media for overnight growth. Transfection efficiency was assessed the next day using an Inverted microscope (Nikon Ti-E, with Nikon Elements software and a Nikon DsRiE camera) and the TxRed turret.

HEK or MEF cells were seeded for 50% confluency. 3ml of media from infected packaging HEK cells was filtered using a 0.22 µm syringe filter and transferred to HEK and MEF cells for

infection with 3 µl of 1000xHexadimethrine bromide (Sigma, cat 107689). This was incubated overnight, and infection efficiency was assessed the following day using the same microscope settings.

3.13.2 Lentiviral transduction of 4242 cells

Viral particles were packaged by HEK as previously described in 3.13.1. The plate was then incubated further overnight with 1% BSA at + 4 °C. Infection of the 4242 was performed following confidential protocol provided by our collaborators from Peter McCallum Cancer Centre. $1x10^6$ of 4242 cells and $5x10^5$ of 4242 cells were spun down and resuspended in media from packaging HEK cells in presence of 1xHexadimethrine bromide . The suspension was transferred into 6-well plates. Some of the wells were pre-coated with retronectin, to improve efficiency of infection. Infection was performed twice, and cells were visualised after 72hrs of incubation.

4. Results

Very little is known about the mouse rDNA IGS-derived noncoding transcriptome. The aim of this project was to characterise the mouse rDNA IGS noncoding profile, utilising the Eμ-myc mouse model of lymphoma. Additionally, we wanted to assess if any regions of the mouse rDNA IGS showed expression changes following CX-5461 treatment. To address these main aims, we performed a combination of capture RNA sequencing (capture-seq), small RNA-seq, bioinformatic analysis and qPCR.

Section 4.1 Identifying regions of lncRNA transcription in the mouse rDNA IGS

4.1.1 Capture-seq experimental design

The first goal of this project was to identify ncRNAs from the mouse rDNA IGS and compare expression level in the presence or absence of RNA polymerase I inhibition. An Eμ-myc variant line which had a short hairpin targeting UBF (shUBF) was also utilised along with the normal Eμ-myc cell (4242). This was included as a means to assess potential changes in the rDNA IGS transcription with a different form of RNA polymerase I inhibition, in either the presence or absence of CX-5461 treatment. Due to the estimated low abundance of these long noncoding transcripts within the transcriptome, we decided to take a capture-seq approach, where sequencing is performed on a 'captured' library of transcripts enriched from an area of interest. To achieve this enrichment, we opted for a probe-based method designed by Roche (the SeqCap EZ RNA enrichment system), in which probes are designed to a region of interest (here being the mouse rDNA IGS) and are pooled together (the probe pool). During library preparation, the probe pool is introduced, and mouse RNAs (converted to cDNAs) that are transcribed from the region of interest hybridise with the probes. Un-bound cDNA can be washed away using a bead-based wash protocol, leaving behind an enriched library derived from the region of interest. The regions of the mouse rDNA IGS where the probes were designed to target is presented in **Figure 3A**. The probes are designed to avoid targeting regions with high levels of similarity to other regions of the genome. In the case of our mouse rDNA unit, it was because these regions were often enriched with repeats as shown by our Repeatmasker IGV visualisation (**Figure 3B**). Our probe pool also consisted of probes designed to capture noncoding RNAs from the human rDNA IGS. Similar to the mouse rDNA IGS, the human IGS had some regions were probes were not designed to target (**Figure 3C**), likely due these regions overlapping with elements found elsewhere in the genome.

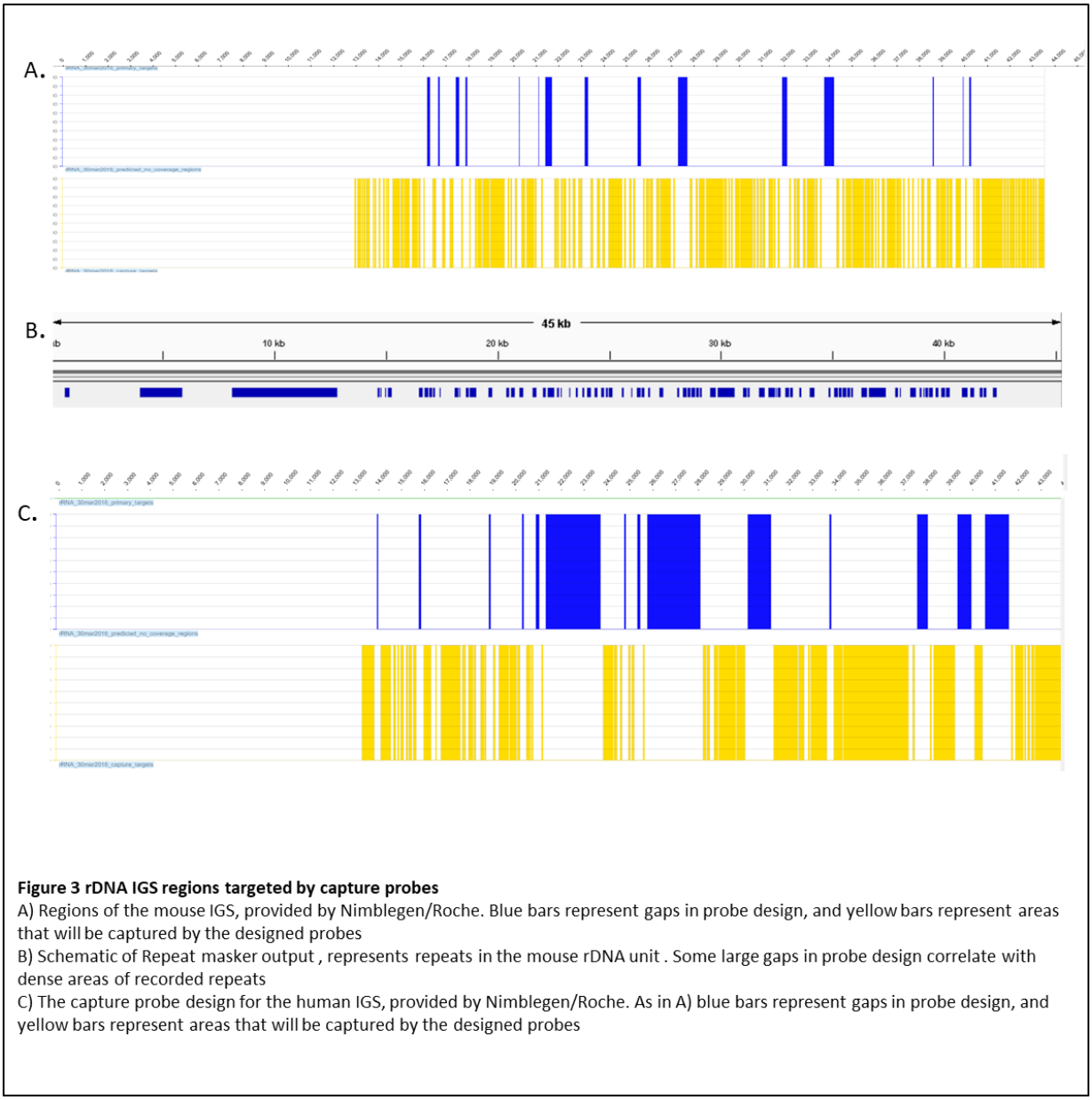


Figure 3 rDNA IGS regions targeted by capture probes
A) Regions of the mouse IGS, provided by Nimblegen/Roche. Blue bars represent gaps in probe design, and yellow bars represent areas that will be captured by the designed probes
B) Schematic of Repeat masker output , represents repeats in the mouse rDNA unit . Some large gaps in probe design correlate with dense areas of recorded repeats
C) The capture probe design for the human IGS, provided by Nimblegen/Roche. As in A) blue bars represent gaps in probe design, and yellow bars represent areas that will be captured by the designed probes

4.1.2 Producing the captured libraries enriched for mouse rDNA IGS transcripts

The capture-seq method can be broken into several steps: RNA extraction and preparation, library preparation for capture (cDNA synthesis and ligation-mediated PCR), probe hybridisation and "captured" RNA enrichment, and finally sequencing for bioinformatic analysis. Here I performed steps of optimisation which were required in order to produce our pooled IGS transcript enriched library, along with the output from the bioinformatic analysis . This revealed a number of potential noncoding-exons within the mouse rDNA IGS.

4.1.2.1 First attempt cDNA library synthesis from high quality ribodepleted RNA

To increase the success of the capture-seq method, high quality and highly concentrated RNA is required as the initial input for cDNA synthesis and the capture. To extract RNA used for cDNA library production, control (4242) and UBF knockdown (shUBF) Eμ-myc mouse lymphoma cells (**Table 1** for more information) were seeded at 8 million cells in 75cm^2 for CX-5461 or control treatment (in triplicates), resulting in 12 flasks. RNA was extracted using two different kits (4 million cells per extraction per kit), the Nucleospin kit and the miRNeasy kit, which were used to get a range of RNA sizes. A Bioanalyser run was performed to assess quality and concentration of RNA output, which showed extracted RNA generally had high ribosomal integrity numbers (or RIN, which accounts for 28S to 18S rRNA ratios) and varying concentrations depending on sample and kit used for extraction. miRNeasy values often did not give a RIN number, but peaks reflected tended to reflect those with a RIN of 10, so we understood RNA quality to be high in these samples also. Each RNA extraction concentration and RIN results can be found in **Table 2 (columns 2/3)**. **Figure 4A** is a representation of a Bioanalyser Nucleospin extraction, that has 28S and 18S peaks corresponding to a RIN score of 10. **Figure 4B** is a representation of a RIN score of 10 in a miRNeasy extraction. The miRNeasy extraction also had a peak early into the trace , which we understand is a reflection of small RNA species. In this figure, both results are from 4242 untreated cells extracted with the different kits.

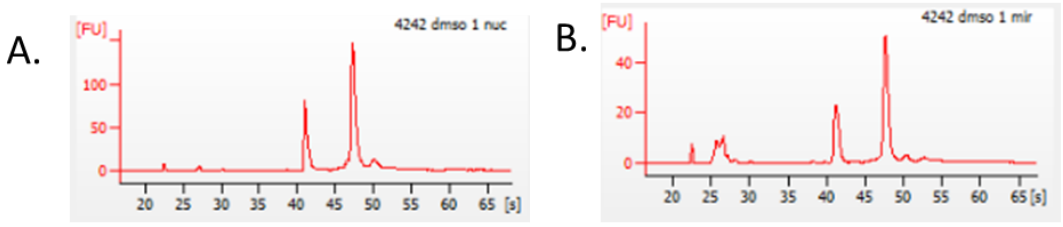


Figure 4 Examples of Bioanalyser outputs with high RNA quality
The Bioanalyser can estimate quality of RNA output by comparing the 28S and 18S peak ratios and determining the RNA integrity number (RIN). A) represents a Nucleospin extraction with a RIN score of 10, while B) represents a miRNeasy extraction with a RIN score of 10. Note miRNeasy extraction has small band at >25 secs, which we understand is a reflection of small RNAs successfully extracted.

To reduce the rRNA abundance in the RNA samples, we performed ribodepletion using the maximum input volume allowed for the ribodepletion, taking equal RNA amounts from each extraction kit of each treatment condition triplicate (**Table 2 column 4**) **.** Consequently, inputs into ribodepletion varied depending on pre-depletion concentration. As shown in **Figure 5,** ribodepletion efficiency was high, as all RNA samples show an absence of 28S and 18S bands seen in **Figure 4**. Ribodepleted RNA outputs (ng) were diverse, with some samples having less than the 100ng input required for capture (**see Table 2 column 5**). Nevertheless, it was decided to continue to cDNA library preparation with these samples, using either the recommended input of 100ng , or the RNA input concentration in the maximum volume allowed for cDNA synthesis.

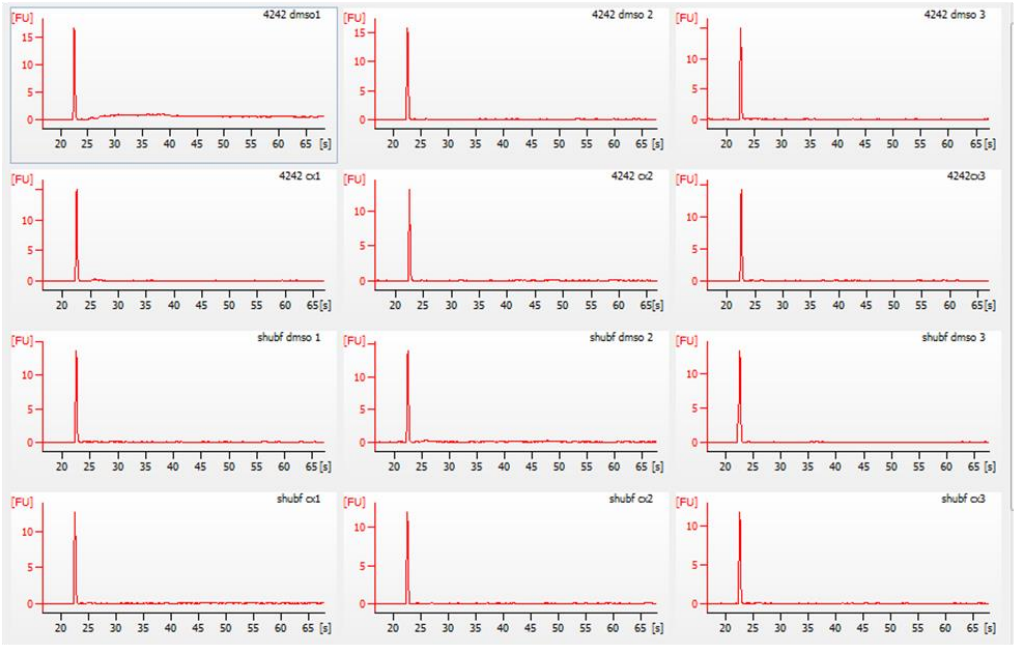
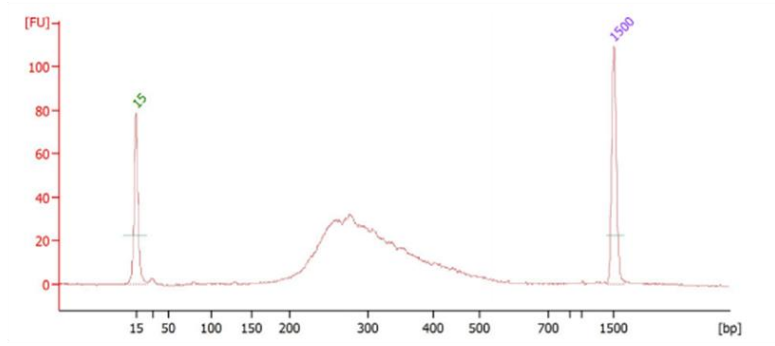


Figure 5 Assessing ribodepletion efficiency in first RNA extractions
Ribodepletion efficiency can be visually assessed using a Bioanalyser. Complete loss of 28S and 18S rRNA peaks suggests high ribodepletion efficiency. Figure shows ribodepletion efficiency, of all RNA samples from the first extraction attempt, was high due to the absence of rRNA bands.

To begin the preparation of cDNA libraries, ERCC spike in mixes (synthetic transcripts used for downstream library preparation quality and differential expression analysis) were added to the RNA samples with ERCC mix being dependent on both cell variant and presence/absence of CX-5461 treatment (section 3.8). The dilution of ERCC mix and volume added was appropriate for 5 µg RNA. cDNA library preparation was then carried out on ribodepleted RNA. Only the two more concentrated shUBF CX-5461 treated ribodepleted RNA sample triplicates were prepared into cDNA libraries, due to restrictions with the number of sequencing adapters available (where sequencing adapters are required both for the process of sequencing and distinguishing cDNA libraries within a pooled library). A DNA sample, which was fragmented via sonication (section 3.5.5), was prepared into a library as well (incorporated into the workflow after the cDNA synthesis step) as a control for assessing downstream capture-efficiency. After libraries were processed, they underwent several cycles of ligation mediated PCR (PCR of cDNA via priming to ligated adapters). According to the SeqCap manual , we expected to see a peak at around 300bp for successfully produced cDNA libraries (**Figure 6A**), measured using a Bioanalyser. Though some peaks were at the correct position (**Figure 6B**), some libraries lacked clear peaks, and others had concentrations lower than required 20.825 ng/µl for the downstream hybridisation step (**Table 2 column 6**). This suggests cDNA synthesis or lm-PCR efficiency differed between cDNA library preparations. These libraries were referred to downstream as day one prepared libraries.

A.



B.

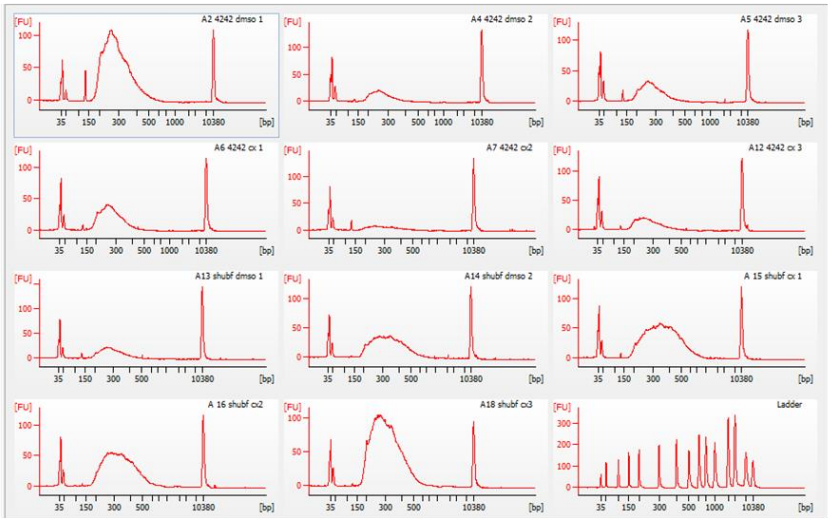


Figure 6 Comparing Bioanalyser results from day one cDNA library preparations after ligation-mediated PCR (lm-PCR) to ideal capture-seq cDNA libraries

A. Ideal cDNA libraries for this capture-seq system after lm-PCR (PCR, priming to adapters ligated to cDNA sequences) were stated by the manual to have a peak in the Bioanalyser trace between 150 and 500 bp. Image adapted from (http://netdocs.roche.com/DDM/Effective/07279337001_RNG_SeqCapRNA-UGuide_v1p0.pdf)

B. The Bioanalyser output from our library preparations post lm-PCR (in order top left to bottom right: 4242 cells control treatment triplicate 1 to triplicate 3, 4242 CX-5461 treated cells triplicate 1 to triplicate 3, shUBF cells control treatment triplicates 1-2. shUBF cells CX-5461 treatment triplicates 1-3). Peak intensity varied, with some lacking obvious peaks at 150-500 bp.

Table 2 First extraction RNA concentrations pre- ,post- ribodepletion and post lm-PCR

Treatment scheme	**Nucleospin RNA conc (ng/µl)**	**miRNeasy RNA conc (ng/µl)**	**Ng input into ribodepletion**	**Ng output from ribodepletion (in 14 µl)**	**Concentration after lm-PCR/DNA concentration (ng/µl)**
4242 DMSO control triplicate 1	304 RIN 10	146 RIN 9.9	3970	440	30.6
4242 DMSO control triplicate 2	172 RIN 10	91 RIN N/A	2401	60	10.2
4242 DMSO control triplicate 3	86 RIN 10	108 RIN N/A	1914	120	109
4242 CX-5461 treated triplicate 1	156 RIN 10	89 RIN N/A	2267	120	16.4
4242 CX-5461 treated triplicate 2	191 RIN 10	330 RIN N/A	4839	120	20.8
4242 CX-5461 treated triplicate 3	114 RIN 10	115 RIN N/A	2290	120	31.6
shUBF DMSO control triplicate 1	85 RIN 10	84 RIN N/A	1690	120	24.4
shUBF DMSO control triplicate 2	402 RIN 10	244 RIN N/A	6073	180	35.7
shUBF DMSO control triplicate 3	122 RIN 10	120 RIN 10	2420	60	N/A
shUBF CX-5461 treated triplicate 1	125 RIN 10	98 RIN 9.9	2196	180	63.7
shUBF CX-5461 treated triplicate 2	210 RIN 10	389 RIN N/A	5453	60	25.1
shUBF CX-5461 treated triplicate 3	189 RIN 10	112 RIN N/A	2813	60	51.0

4.1.2.2 Second attempt cDNA library synthesis from high quality ribodepleted RNA

To produce more concentrated libraries for capture, extractions were repeated with cells seeded and treated at 11 million cells per treatment flask (or 5.5 million cells for each extraction kit). RNA output yield from extractions from both Nucleospin and miRNeasy kits were higher than the previous extraction attempts while maintaining high RNA quality with peaks similar to those seen in **Figure 4A** (**table 3 column 2/3** for concentrations) . Ribodepletion was performed on 7 µg of pooled RNA (3.5 µg from each extraction kit) of each treatment scheme triplicate as before, to allow for a greater output for library preparation. The post-ribodepleted RNA output was more concentrated than the previous attempts output (**table 3 column 5**), though depletion was less efficient than the first attempt, as represented **Figure 7** by the presence of residual rRNA peaks in some RNA samples (compared to **figure 5**). cDNA library preparation was repeated using these ribodepleted RNA samples as completed previously, using either the recommended input of 100ng or the RNA input concentration in the maximum volume allowed for cDNA synthesis. Here, ERCC in spike transcripts were added in a dilution and volume appropriate for 7 µg input. Again, only the two shUBF-CX-5461 treated RNA samples with the highest concentration were used for library preparation. As before, the sonicated DNA was prepared into a library in parallel with the RNA-derived libraries after the cDNA synthesis steps were carried out on the. Post lm-PCR Bioanalyser electrographs of the cDNA libraries can be seen in **Figure 8**, showing cDNA libraries generally having peaks in the ideal area of between 150-500bp (as shown in **Figure 6A**). Again, we saw varying degrees of intensity in this region between cDNA libraries, with some libraries again lacking clear peaks, suggesting that cDNA synthesis or lm-PCR efficiency again differed between cDNA library preparations. These libraries were referred to as day two prepared libraries.

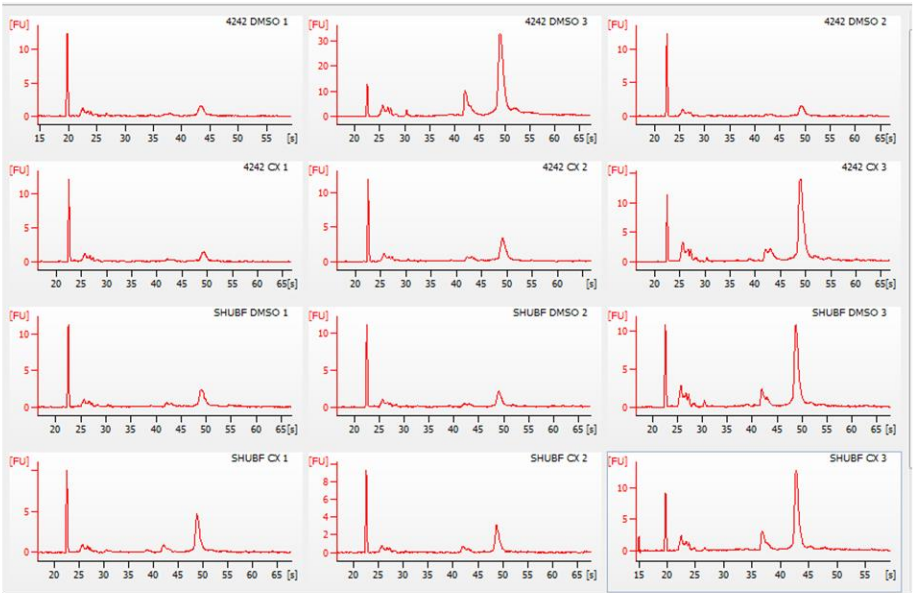


Figure 7 Assessing ribodepletion efficiency in second RNA extractions
Ribodepletion efficiency can be visually determined using a Bioanalyser by the presence or absence of rRNA 18S and 28S peaks. Residue of peaks suggests ribodepletion wasn't 100% efficient

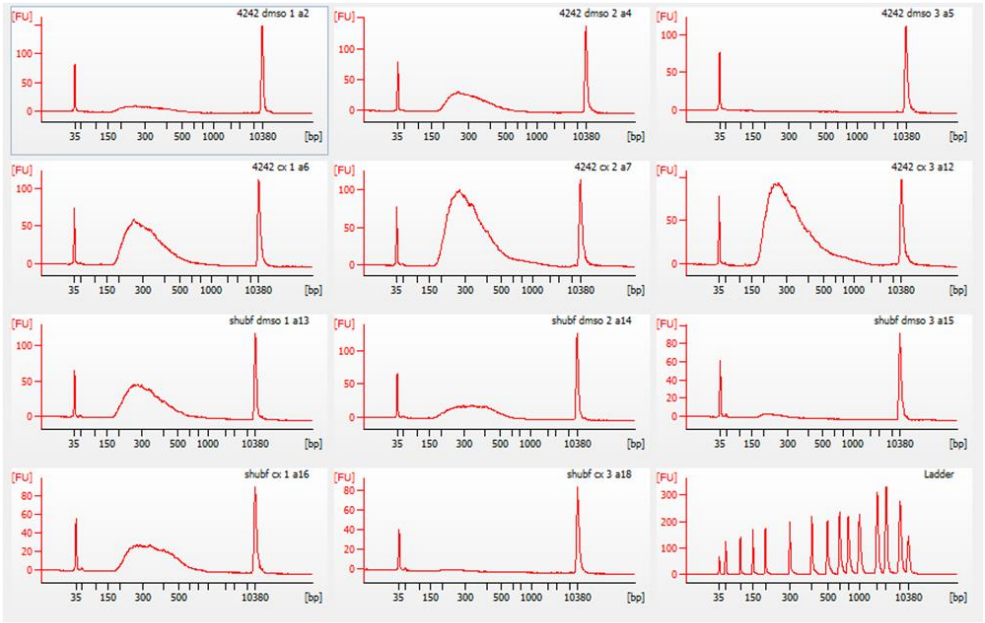


Figure 8 Bioanalyser results of day two cDNA library preparations after ligation-mediated PCR (lmPCR)
Bioanalyser results of the lmPCR of cDNA produced from the different treatment scheme RNA preparations. Peaks should ideally be between 150-500bp (see figure 6A). Some libraries showed low efficiency lmPCR resulting in low intensity peaks (adapters A2, A5, A14, A15 and A18). Adapter numbers and library treatment scheme (cell line, treatment, triplicate in top left of each graph.

Table 3 Second extraction RNA concentrations pre- ,post- ribodepletion and post lm-PCR

Treatment scheme	**Nucleospin RNA conc (ng/µl)**	**miRNeasy RNA conc (ng/µl)**	**Ng input into ribodepletion**	**Ng output from ribodepletion (in 14 µl)**	**Concentration after lm-PCR/DNA concentration (ng/µl)**
4242 DMSO control triplicate 1	438 RIN 10	RIN Not recorded	7000	196	22.8
4242 DMSO control triplicate 2	438 RIN 10	RIN	7000	112	55.8
4242 DMSO control triplicate 3	310 RIN 10	RIN	7000	1344	10
4242 CX-5461 treated triplicate 1	612 RIN 422	RIN	7000	126	160
4242 CX-5461 treated triplicate 2	422 RIN 10	RIN	7000	182	263
4242 CX-5461 treated triplicate 3	397 RIN 10	RIN	7000	546	318
shUBF DMSO control triplicate 1	465 RIN 10	RIN	7000	182	138
shUBF DMSO control triplicate 2	385 RIN 10	RIN	7000	112	46
shUBF DMSO control triplicate 3	445 RIN 10	RIN	7000	392	21.3
shUBF CX-5461 treated triplicate 1	652 RIN 10	RIN	7000	210	560
shUBF CX-5461 treated triplicate 2	457 RIN 10	RIN	7000	98	N/A
shUBF CX-5461 treated triplicate 3	511 RIN 10	RIN	7000	434	40

4.1.2.3 Measuring efficacy of CX-5461 treatment on 4242 and shUBF cells

In order to ensure that CX-5461 treatment was effective in causing cell death in both 4242 and shUBF cell variants, a cytotoxicity assay (section 3.3) to measure IC_{50} (treatment concentration resulting in 50% of cells being viable) was performed in parallel to treatment for RNA extractions, using the same cells and CX-5461 treatment density in a 96 well plate format. An example of the cytotoxicity assay and IC_{50} results, specifically from cells that were treated and used for day one library preparations, is shown in **Figure 9.** Increasing concentrations of CX-5461 treatment resulted in decreasing percentages of viable cells , suggesting that CX-5461 treatment resulted in decreased cell viability corresponding to increasing CX-5461 concentrations. IC_{50} concentrations were calculated as 3.5nM for 4242 cells and 3.1nM for shUBF.

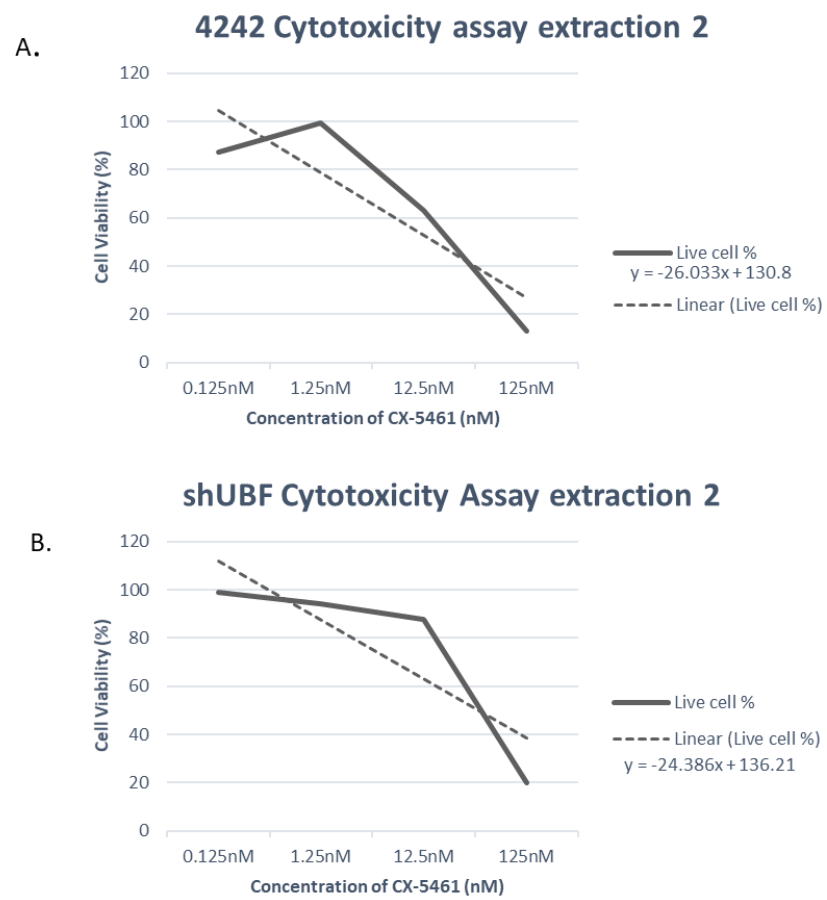


Figure 9 Example of Cytotoxicity assay results
Cytotoxicity assays measure cell viability after treatment with increasing concentrations of a drug. Cell viability (%) after treatment with CX-5461 decreases, with increasing concentrations of CX-5461 in both 4242 and shUBF cell variants.

4.1.2.4 Preparing and capturing pooled cDNA libraries

To prepare the libraries for capture, cDNA libraries were pooled. As a result of different cDNA or lm-PCR efficiencies producing variable concentrations in cDNA libraries prepared specifically in section 4.1.2.2, we selected two triplicates from day two library preparation with one triplicate from the day one library preparation from each treatment scheme (excluding shUBF CX-5461 cDNA libraries where one was selected from each day preparation). Library selection was dependant on output concentration (where the highest output concentrations were generally selected), as well as adapter number as it was critical not to have any adapter

replicates. **Table 4** outlines which library preparation day (one or two) cDNA libraries used for capture originated from. The DNA library produced on day two was used selected for capture.

Table 4 Outline of day one and day two libraries pooled for capture

cDNA library	Library preparation day	Adapter name	Library name
4242 DMSO Triplicate 1	day two	A2	S1
4242 DMSO Triplicate 2	day two	A4	S2
4242 DMSO Triplicate 3	day one	A5	S3
4242 CX-5461 Triplicate 1	day two	A6	S4
4242 CX-5461 Triplicate 2	day two	A7	S5
4242 CX-5461 Triplicate 3	day one	A12	S6
shUBF DMSO Triplicate 1	day two	A13	S7
shUBF DMSO Triplicate 2	day one	A14	S8
shUBF DMSO Triplicate 3	day two	A15	S9
shUBF CX-5461 Triplicate 1	day two	A16	S10
shUBF CX-5461 Triplicate 3	day one	A18	S11
DNA library	day two	A19	S12

Most cDNA samples required concentration to the meet the required concentrations for capture. Consequently, we compared two methods of concentration: evaporation by DNA vacuum and evaporation using a heat block. ~820ng replicates of Eμ-myc 4242 DNA (extracted previously) were diluted to 50µl in water. 1 replicate was left over the weekend on a shaking heat block at + 23C (at 400rpm), while the other completely dried in a Savant 120 DNA concentration. Passive drying showed no significant reduction in volume over a 48-hr period and is likely to result in a greater risk of contamination (not shown). In contrast, after complete drying with the DNA concentrator and resuspension in the original input volume of water, we

saw only a 12.5% loss of DNA was shown (820ng to 717ng). It was decided to then use the DNA evaporation concentration in the vacuum as the method of concentrating all pre-capture libraries. Consequently, we used the vacuum method to concentrate all pre-capture libraries.

All concentrated libraries were pooled at equal concentrations, producing a single multiplexed pooled library. The multiplex pooled library was used for the capture using the rDNA IGS targeting probe pool(described in section 4.1.2), following the protocol outlined in section 3.8, resulting captured multiplexed pooled library enriched for rDNA IGS transcripts. The captured multiplexed library pool was measured both by gel electrophoresis and Bioanalyser (with Bioanalyser results shown in **figure 10**). According to the capture protocol followed throughout this section, the majority of peaks in the captured multiplexed library pool should be between 150-500bp. Our results (**figure 10A**), showed the majority of products being of this size, with some larger. The negative control (**figure 10B**) showed no amplification, suggesting our results were not an artefact of random amplification or some other external factor. The captured multiplexed library pool was sent for sequencing by New Zealand Genomics Ltd (NZGL).

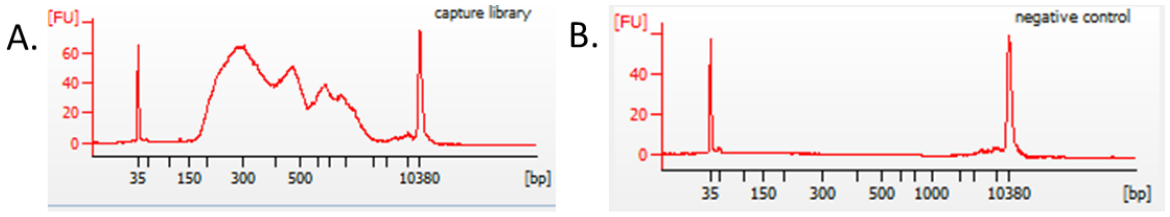


Figure 10 Final pooled captured library output
Captured multiplexed library pool are assessed by Bioanalyser. According to the SeqEZ RNA enrichment manual (section 3.8), the majority of fragments should be between 150-500bp. Our libraries show this (A.), along with some of larger size . Our negative control (B.)which was prepared alongside showed no peaks, suggesting that there was no contamination nor were our results a reflection of random amplification.

4.1.3 Bioinformatic analysis of capture-RNA-seq for IGS noncoding-transcripts

To identify potential mouse rDNA IGS noncoding RNAs and assess their transcriptional changes in control and CX-5461 treated RNA libraries, sequenced data produced in section 4.1.2 was bioinformatically analysed in a number of ways. The initial raw sequencing output (and sequencing statistics) can be found in **Table 5**. Of worthy note, it appeared that libraries produced on day one (S3, S6, S8, S11) had on average a lower read output. 4242 DMSO triplicate 3 had particularly low read output. This variability produced in sequencing had an effect downstream in terms of normalising between libraries.

Table 5 Raw read and STAR aligner outputs from rDNA IGS capture sequencing data

Sample	Library name	GC%	Average length	Total reads	STAR Aligner uniquely mapped reads
4242 DMSO triplicate 1	S1	58	143.67222	1817640	1202288
4242 DMSO triplicate 2	S2	57	151.39281	2172674	1510219
4242 DMSO triplicate 3	S3	46	141.12672	93835	68115
4242 CX-5461 triplicate	S4	56	155.08229	2156277	1482284
4242 CX-5461 triplicate 2	S5	56	156.78795	2628097	1833497
4242 CX-5461 triplicate 3	S6	48	132.82546	249822	183436
shUBF DMSO 1	S7	58	159.67718	2099849	1449374
shUBF DMSO triplicate 2	S8	51	161.26273	430707	316185
shUBF DMSO triplicate 23	S9	51	120.76444	958589	655017
shUBF CX-5461 triplicate 1	S10	56	175.03224	1407617	974549
shUBF CX-5461 triplicate 3	S11	52	150.92825	986154	723962
DNA	S12	43	219.27666	3158404	2062302

To begin the bioinformatic workflow, the rDNA unit and the ERCC files (where ERCCs are used to normalise for sequencing variability between libraries) were concatenated to the newest mouse genome assembly. This genome was indexed using STAR, in which the genome is compressed to allow for faster alignment. STAR was also used for read alignment to the genome, and the final numbers uniquely mapped reads can be found in **Table 5 column 5**. We

did see some multimapping reads and some unmapped reads, which were removed at a later stage. From here, we further assessed our sequencing data using a number of different tools, to address several different questions.

4.1.3.1 The theoretical efficiency of the Capture-seq method at enriching for noncoding RNA from the IGS

We wanted to assess the theoretical efficiency of capturing the rDNA IGS (and consequently IGS transcripts) using the capture method performed in section 4.1.2.4. To do this, we utilised the sequenced DNA-derived library (library S12) captured in parallel with cDNA libraries that was mapped to the reference genome. The theoretical capture efficiency is reflected by the difference in average read depth (coverage) of the IGS (target region) compared to the rest of the genome. Here, a high coverage of the IGS compared to the low coverage of rest of the genome would suggest that reads aligning to the IGS are comparatively more enriched in the sequencing data, indicating a high capture efficiency.

To calculate the average coverage of the genome, we used Bedtools on the S12 aligned read output. Bedtools is a bioinformatic tool which calculates average coverage for every base position across the reference after read alignment. From this output, each chromosome's coverage value (per bp) was extracted, and the average coverage of a chromosome was calculated. The average coverage depth of rDNA coding region and IGS was also calculated independently. Average coverage values per chromosome can be found in **Table 6**. The average coverage of the genome was calculated excluding the rRNA coding region, mitochondria and the sex chromosomes. Sex chromosomes had a coverage of half, reflective of existing as a single copy in this cell line (male specimen).

The average coverage depth of the IGS was ~91950 higher than the average coverage of the genome. Since coverage is proportional to the read abundance, we propose that reads aligning to the rDNA IGS are enriched in the S12 library, and consequently we estimate efficiency of capturing reads aligning to the IGS was high proportional to the rest of the genome.

We incorporated only a single rDNA reference copy into the reference genome. Given that the rDNA exists in multiple copies, reads from all rDNA copies will map to the single reference copy, which will give higher coverage signals at the rDNA. Consequently, to better assess theoretical capture efficiency, we then compared average coverage of the IGS to the rRNA coding region. In this case, the rDNA IGS had a coverage 171.91 times higher than the rRNA coding region. This suggests that reads aligning to the IGS were more abundant in the data than reads aligning to the rRNA which results in a 171.91 times higher coverage of the IGS, consequently the capture was concluded to have high theoretical efficiency.

From this data, we also estimated the rDNA and the mtDNA copy number. Like the rDNA, the mtDNA exists in multiple copies in the genome, but align only to a single reference copy. Copy number could be calculated by dividing the coverage of the rDNA or mtDNA by coverage of the rest of the genome (having an average coverage of 0.2435). We estimated the copy number of rDNA was ~535 , and the mitochondria DNA copy number was ~34. This was a much higher estimation of rDNA copy number than previous estimates which have suggested that in mouse the rDNA can exist in up 410 copies((Gibbons, Branco, Godinho, Yu, & Lemos, 2015)), and much lower for the mitochondria DNA, which copy number varies dependent on tissue of origin but generally estimated to be in the hundreds (R. D. Kelly, Mahmud, McKenzie, Trounce, & St John, 2012). This may not be surprising, as a previous study in humans found a negative correlation between rDNA copy number and mtDNA copy number (Gibbons, Branco, Yu, & Lemos, 2014), which would explain the high rDNA copy number and low mtDNA copy number.

Table 6 Theoretical capture efficiency using DNA derived library comparing all mouse chromosomes (and rDNA coding region) to the IGS

Chromosome	Average Depth/coverage (X)	Chromosome size (bp, NCBI)
Chromosome 1	0.224139	195471971
Chromosome 2	0.2747	182113224
Chromosome 3	0.215091	160039680
Chromosome 4	0.37334	156508116
Chromosome 5	0.4205	151834684
Chromosome 6	0.181738	149736546
Chromosome 7	0.23855	145441459
Chromosome 8	0.214176	129401213

Chromosome 9	0.252	124595110
Chromosome 10	0.201439	130694993
Chromosome 11	0.271195	122082543
Chromosome 12	0.21181	120129022
Chromosome 13	0.242176	120421639
Chromosome 14	0.177596	124902244
Chromosome 15	0.255041	104043685
Chromosome 16	0.203962	98207768
Chromosome 17	0.224116	94987271
Chromosome 18	0.198078	90702639
Chromosome 19	0.246911	61431639
Chromosome X	0.105637	171031299
Chromosome Y	0.103462	91744698
Mitochondria	8.236149	20000
rDNA coding region	130.2447	13427
rDNA IGS	22389.83	31888

4.1.2.2 Using bedtools coverage to assess for areas of transcription within the mouse rDNA IGS

To begin to assess which regions of the mouse rDNA IGS show transcriptional activity based on the Bedtools coverage data, we further assessed sequencing coverage per base in all RNA derived libraries (S1-S11) across the mouse rDNA IGS. To do this, we extracted the coverage values of each library for every base position of the rDNA using Bedtools (section 3.10.5). Using this data and the ggplot2 package within R, we first plotted the coverage values corresponding to each base of the rDNA for each library (**Figure 11**). There remained significant read coverage associated with the rRNA coding region as depicted by peaks in the graphs. This indicated that some rRNA remained in samples even with ribodepletion and capture-based enrichment. Together, this suggested that though theoretical capture efficiency is relatively high (section 4.1.2.1), transcripts derived from the IGS are likely to be in low abundance comparative to any remaining rRNA. Nevertheless, independent of library preparation quality (i.e. day 1 or day 2), we saw peaks at similar positional coordinates throughout the IGS. This suggested consistency

in regions that have greater coverage outputs, which may represent regions of higher transcriptional rate.

To better visualise high coverage peaks in the IGS which may reflect regions of transcription, we plotted the IGS alone, where the start was determined from the first of the Sal boxes (section 1.3.2). The first Sal box (searching for the first 11 bases of the Sal box with BLAST) started at 13427 bp, so Bedtools outputs were edited to remove all values before 13427 bp. These were again graphed as coverage against base position (**Figure 12**).Peaks seen in the previous figure became clearer with the reduction in the y-axis maximum value. Though the depth values differed greatly dependent on library preparation day, we saw a general trend in coverage peaks. From 13400 bp to before 25400 bp, coverage values remained low (with a few smaller peaks scattered throughout), at which point there was a spike in coverage. From 25400 bp to around 41000 bp, the graphs plateau but at consistently higher coverage values than between 13400-25400 bp, with several peaks consistently arising roughly at 30000bp, 32000bp, 39000bp and 40000bp. Finally, there are 2 larger groups of peaks, with the largest corresponding to around 42000bp, and 44000bp towards the end of the IGS. From these graphs, we rationalise that each major peak is attributed to a region of higher transcription levels, with the peaks with the highest y-values having the highest transcriptional levels. Lower peaks or plateaus may also represent areas of transcription, but the transcripts have lower abundances.

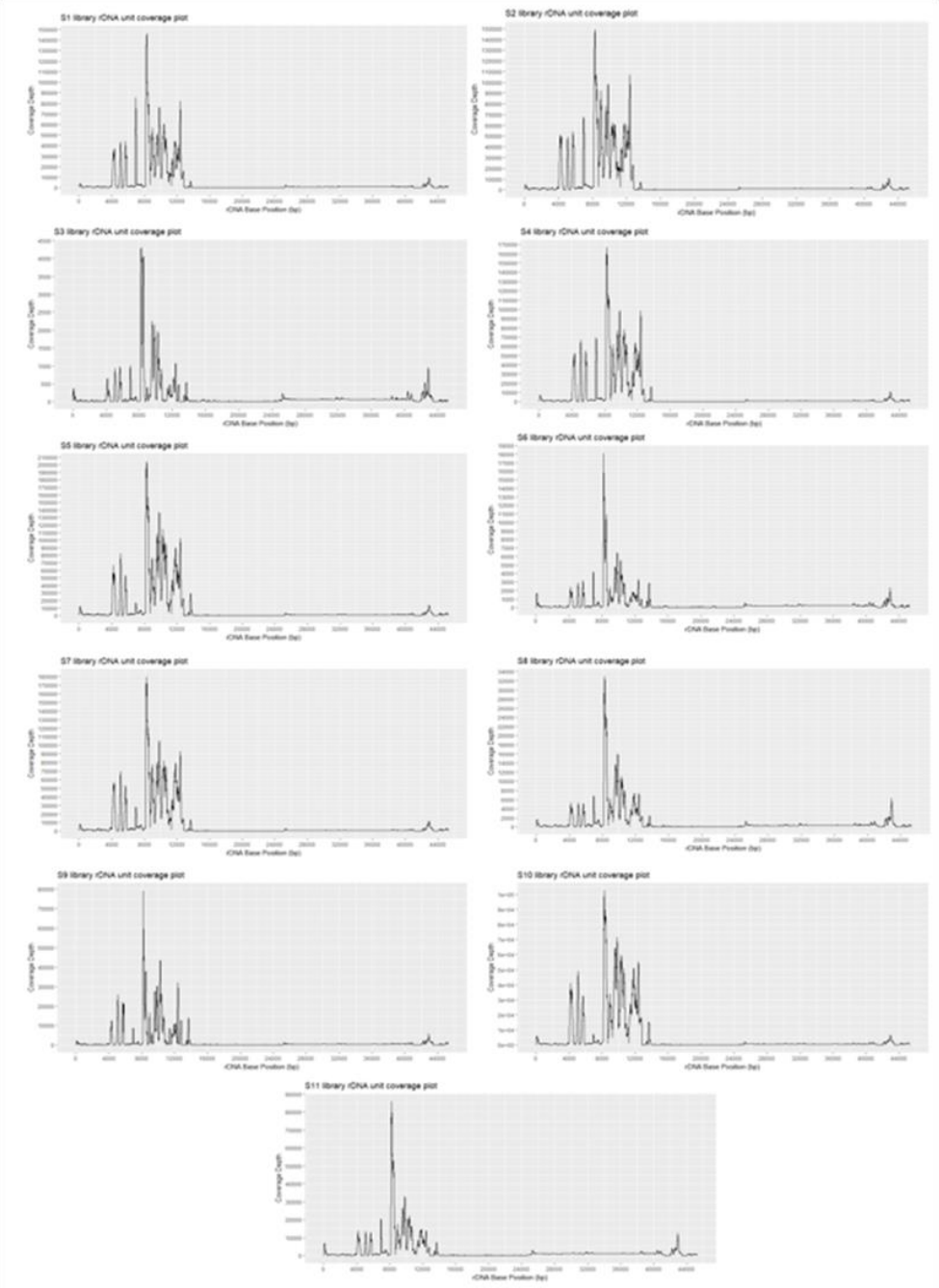


Figure 11 Coverage graphed against rDNA base position (coding and IGS)
Bedtools outputs graphically represented as coverage at each base of the full rDNA unit, for all cDNA libraries produced (named here S1 through S11, see table 4 for names). Peaks with higher coverage were seen in the coding region suggest that rRNA was still present in our samples post-depletion and capture. Peaks after coding region may represent areas of transcription.

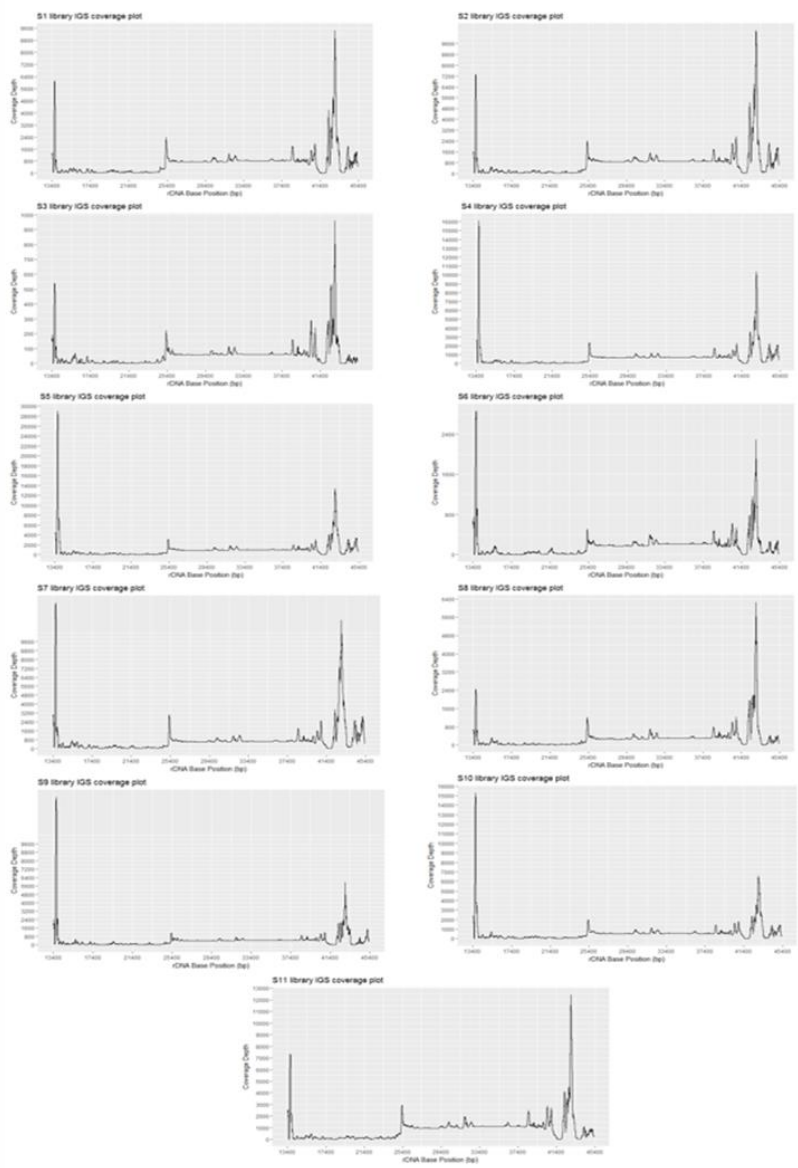


Figure 12 Coverage graphed against IGS base position
Base position of the rDNA specifically starting at the IGS (X axis) was plotted against their corresponding coverage values (Y axis, from Bedtools output) for all cDNA libraries. Peaks become more obvious when rRNA coding region was removed. Peaks in coverage correspond with increased read abundance and consequently correlated with increased transcription.

Some sequences in the genome show sequencing bias reflective of base composition (Ross et al., 2013). To assess if peaks were artefacts of sequencing bias within regions of the IGS, we normalised the coverage values in cDNA libraries to the S12 DNA library. This would account

for regions of the IGS which naturally sequence better, consequently giving a higher coverage and artificial peaks in the coverage plots. First, we took the average coverage of all RNA libraries within each treatment scheme (4242 DMSO control, 4242 CX-5461, shUBF DMSO control, shUBF CX-5461), to produce 4 (average) data sets. Then, the average coverage per base of the IGS (within each treatment scheme) was normalised to the coverage values at each corresponding base of the IGS in the DNA library. This is graphically represented in **Figure 13**. We found the majority of normalised peaks remained at the end of the rDNA IGS, suggesting this region is likely to produce transcripts. Regions that showed peaks before, but upon normalisation these peaks are lost or plateau (i.e. in the middle of the rDNA IGS, see **Figure 12**), may represent regions with either high capture or sequencing bias.

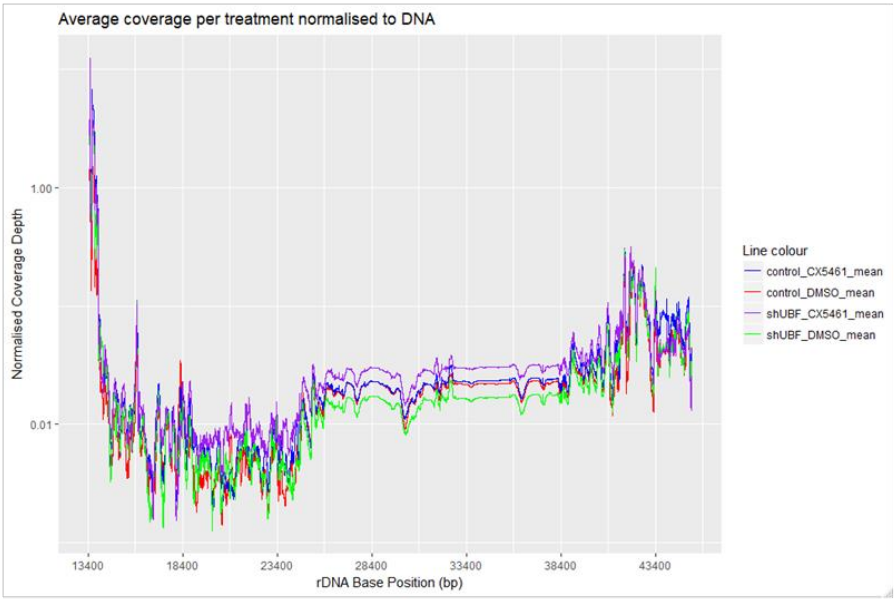


Figure 13 Coverage graphed against IGS base position after normalisation to DNA capture
Average coverage values at each base of the rDNA IGS for each treatment scheme (4242 DMSO control, 4242 CX-5461 treatment, shUBF DMSO control, shUBF CX-5461 treated) were normalised to the S12 captured DNA library. This was performed to normalise for any bias occurring during sequencing that might produce artificial peaks (or troughs) in data.

This data together revealed some clear peaks in coverage across the mouse rDNA IGS in captured cDNA libraries, where coverage can reflect transcription. These regions are consequently likely to produce ncRNAs, which was validated downstream.

4.1.2.3 Finding rDNA IGS exons using a bioinformatic approach

To identify transcripts and exons within our aligned sequencing data and assess changes in their coverage values (a reflection of abundance) in the different treatment schemes, we utilised the Stringtie software (see section 3.10.4). Stringtie has manipulatable parameters which allow users to better adjust stringency in what can be predicted as an exon/transcript. We used the default parameters, with an example of the output being shown in **Appendix 2** (the S1 library). IGS transcripts were often very long and showed signs of potential splicing events as many IGS transcripts were made up of the same predicted exons. Further, exons tended to be predicted in the same general regions across the rDNA IGS. Due to a combination of the length of the transcripts, and more consistency in exon prediction across libraries, we decided to focus on assessing and validating the potential ncRNA exons within the IGS. Stringtie predicted rDNA exons were extracted from the output, and can be seen in **Figure 14**, where predicted exons are represented by bars on tracks. Some larger bars on tracks are a representation of overlapping predicted exons**.** To insure exons predicted by Stringtie do not show major overlap with repeats, which could suggest predicted exons are potential reflection of sequencing biases , we used the RepeatMasker software (section 3.10.9) to identify mouse rDNA repeats. The Repeatmasker output was transferred to an excel file, where it was made into a GTF file. This was loaded into IGV, along with the Stringtie output (**Figure 14**, RepeatMasker output green track). The Repeatmasker track shows a scattering of repeats throughout the rDNA IGS. Exon predicted by Stringtie(blue tracks) are much longer, and though there is some overlap between repeats and exons, no full predicted exon can be explained by the presence of a repeat.

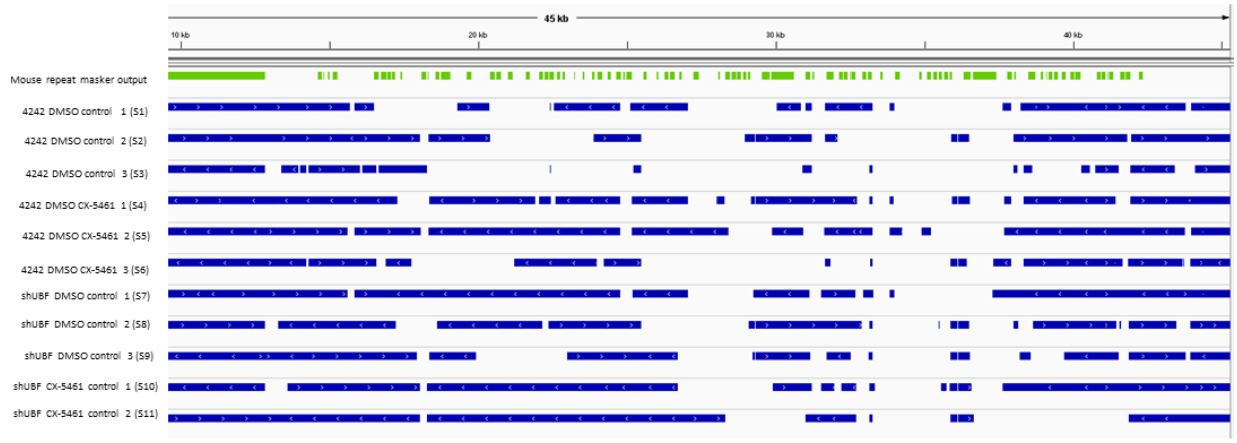


Figure 14 Stringtie-predicted exon output visualised in IGV
Stringtie predicted exons are in .gtf file format, which can be visualised in IGV. Each track represents a captured cDNA library preparation, and exons are represented as blue bars. The green track is repeat masker (converted into .gtf format). Exon are not called always called consistently between libraries, though there are clear areas that lack transcription shared between libraries.

4.1.2.4 ERCC analysis for normalising Stringtie FPKM outputs between cDNA library samples

cDNA libraries may show variation in exon coverage values reflecting differences produced by a number of external factors, for example differences in sequencing efficiencies. To normalise for these differences, we compared ERCC spike ins between the different libraries. ERCC spike in mixes contain artificial transcripts at different concentrations, which during sequencing will result in concentration-dependent FPKM outputs. Each mix (1 or 2) contains the same amount of transcripts overall, but the concentration ratios of all transcripts differ. As the ERCC concentration input is known, the relationship ERCC transcript input concentrations correlated with ERCC transcript FPKM output can be used an exon coverage normalisation factor for downstream differential expression analysis. Concentrations of all ERCCs were calculated specifically for day 1 or day 2 library preparations, and for input mix (Mix 1 or Mix 2) (**see appendix 7**, section 3.10.7 for method). ERCC concentration input/ERCC FPKM output graphs for all cDNA libraries are represented in **Figure 15A**, and slopes were calculated for individual cDNA libraries. To assess how slopes differed dependent on the two library preparation days (section 4.1.2.1 and 4.1.2.2 for day one and day two prepared libraries respectively), ERCC concentration input/ERCC FPKM output slopes were plotted separately for day one and day

two prepared libraries. Here, we found slopes from day two cDNA libraries clustered into two groups, hence forth known as the upper and lower slope clusters groups (**SCG**) (**Figure 15C)**. Day one prepared libraries had highly variable slopes (**Figure 15B**). To investigate what is responsible for the slope clustering in day two libraries, we assessed whether the two groups were associated with input volume into capture or read number output from sequencing. We found no clear correlation between either of these factors and slope clustering (**Figure 16 A and B respectively**).

We propose that the lower SCG groups had lower sequencing efficiency, resulting in lower FPKM outputs. It was decided to continue analysis with day 2 prepared libraries, as the libraries were generally of higher sequencing output and quality (see **Table 4** for day two library preparation libraries) .

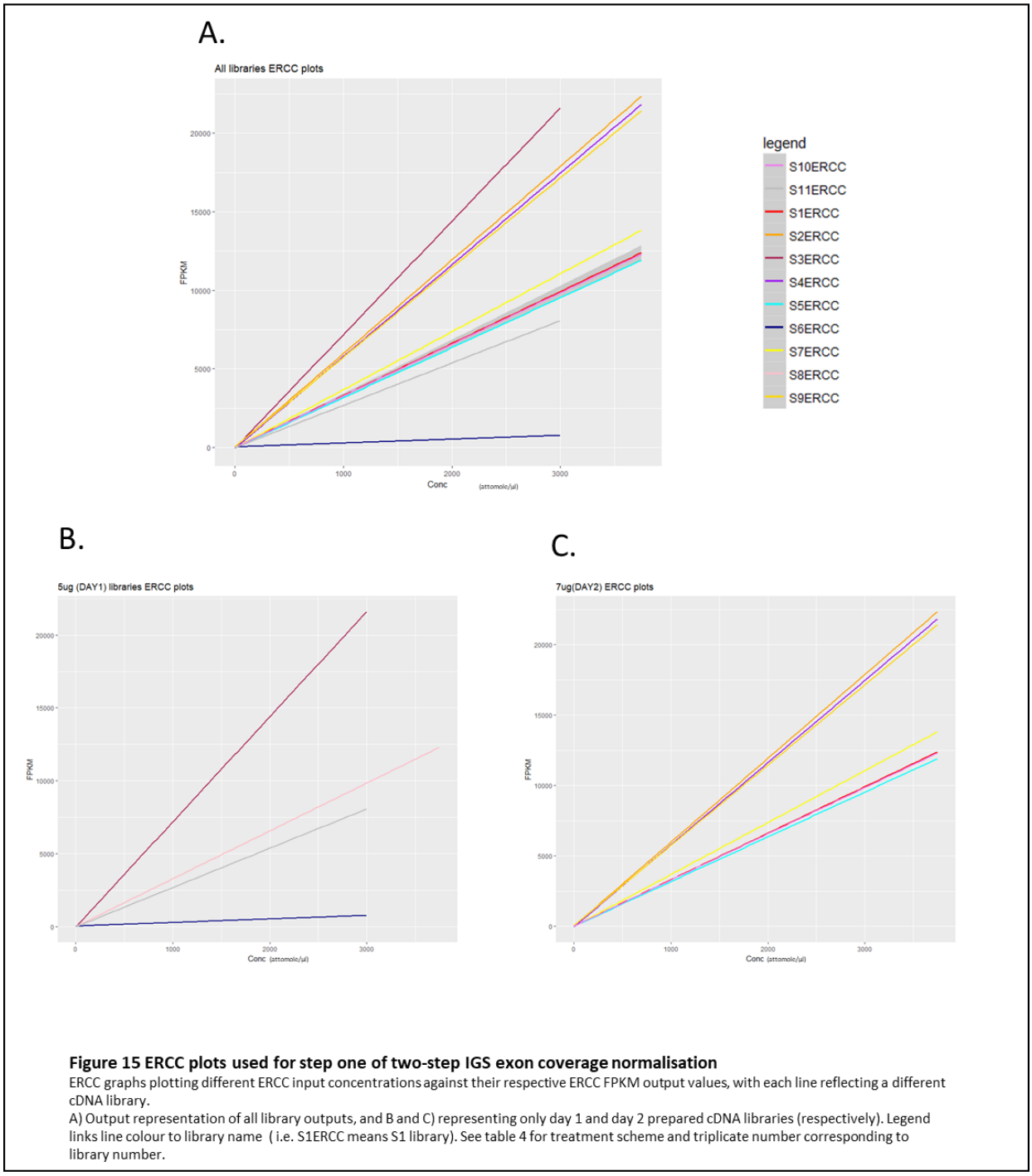


Figure 15 ERCC plots used for step one of two-step IGS exon coverage normalisation
ERCC graphs plotting different ERCC input concentrations against their respective ERCC FPKM output values, with each line reflecting a different cDNA library.
A) Output representation of all library outputs, and B and C) representing only day 1 and day 2 prepared cDNA libraries (respectively). Legend links line colour to library name (i.e. S1ERCC means S1 library). See table 4 for treatment scheme and triplicate number corresponding to library number.

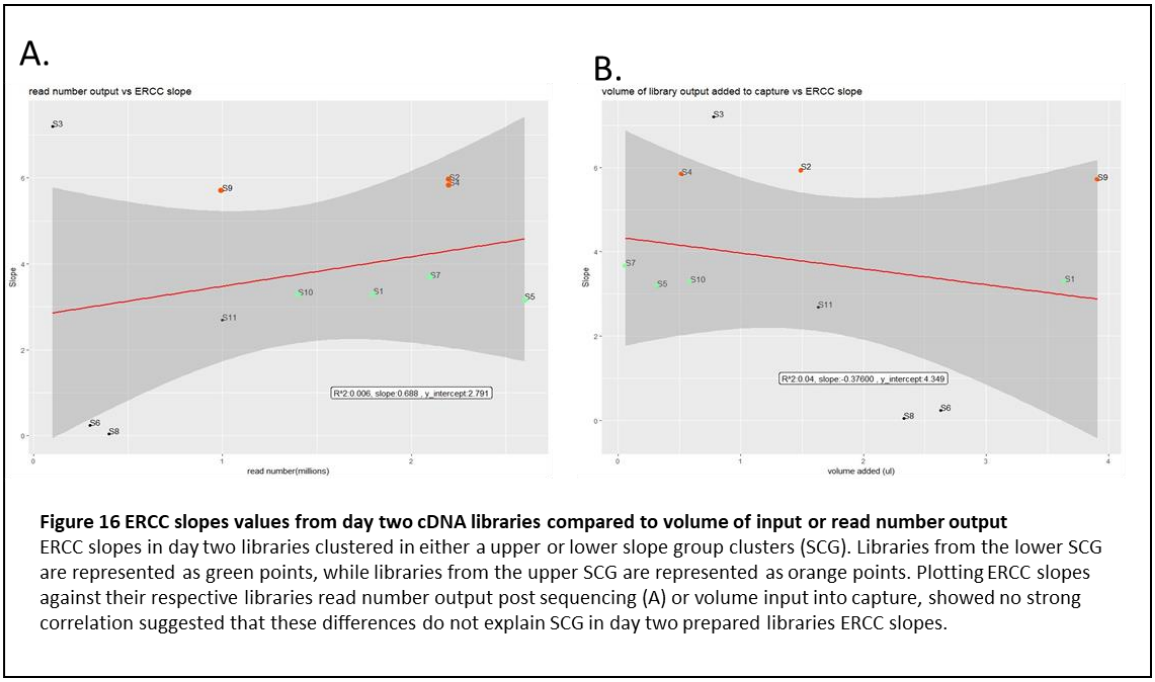


Figure 16 ERCC slopes values from day two cDNA libraries compared to volume of input or read number output
ERCC slopes in day two libraries clustered in either a upper or lower slope group clusters (SCG). Libraries from the lower SCG are represented as green points, while libraries from the upper SCG are represented as orange points. Plotting ERCC slopes against their respective libraries read number output post sequencing (A) or volume input into capture, showed no strong correlation suggested that these differences do not explain SCG in day two prepared libraries ERCC slopes.

To normalise exon coverage between libraries and reduce the effect of sequencing differences for later identification of highest coverage exons (which are more likely to be exons of interest), we performed a two-step normalisation. Firstly, to normalise within the ERCC SCGs, all slopes within each SCG were average. This resulted in an average slope for the upper and lower SCG. Then, the normalisation factor was calculated as the average slope of the upper SCG divided by the average slope of the lower SCG. All Stringtie exon coverage values in libraries in the lower SCG were multiplied by the normalisation factor, to account for any differences in exon coverage that may be explained by lower sequencing output (reflected by lower FPKM). This completed the first step of the two-step normalisation, with this differences with sequencing efficiency were accounted.

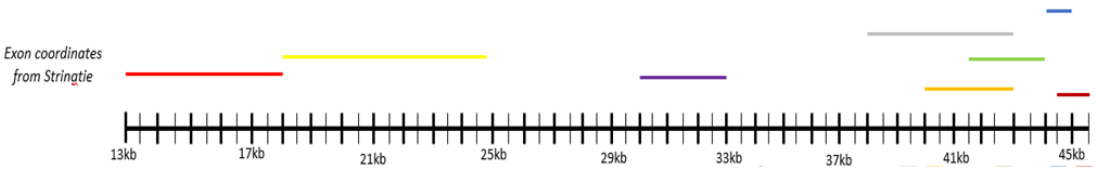
To normalise for sequencing bias across the IGS, we performed the second step of the two-step normalisation, which required us to normalise for regions of the IGS that capture or sequence more or less efficiently depending on the sequence. In reference to **Figure 13**, we saw that some regions showed peaks in coverage, that were reduced by normalising to the DNA, which may suggest these regions either capture or sequence better. To account for this, exons were normalised to average coverage of the DNA sample spanning the region the exon

is predicted. The final coverage values for all exons in each of the treatment libraries were ranked from highest to lowest coverage, where the highest ranked exons may be of particular interest due to reflecting regions of higher transcriptional activity. The full list of two-steps normalised Stringtie predicted exons from all day two cDNA libraries (specifically including only exons which overlap with the IGS), ranked from highest to lowest coverage, can be found in **Appendix 3**.

From the normalised Stringtie data, we found many higher coverage exons were predicted consistently between libraries, meaning that they had either a common start or stop sites, or both. Further, even when exons did not share the start/stop site within the rDNA IGS, predicted exons generally tended to cluster/overlap in several main regions in most libraries, suggesting these areas of the IGS have higher rates of transcription. We decided to focus on 8 potentially transcription rich regions, which consistently showed high numbers of exons predicted within their boundaries. These regions were called exon cluster 1 through to exon cluster 8 for clarity , and a rough location of these can be seen in **Figure 17**. These 8 particular cluster regions were selected as either all exons fitting the clusters had the same stop or start site predicted between libraries, or there was a particular enrichment in exons predicted in the boundaries of the clusters which generally showed higher predicted normalised coverage values.

We propose that exon clusters might show differential expression dependent on treatment scheme. We consequently assessed normalised exon coverage data, to see if there were any patterns in exons with high or low coverage in specific treatment conditions. We found where exons ranked in regard to their coverage, differed dependant on library preparation. There were no apparent trends in where exons ranked (from highest or lowest coverage) correlated with specific treatment schemes. For example, in libraries S1 and S5 (4242DMSO and CX-5451 cDNA respectively), the top-ranking exon fit into exon cluster group 5. In S2 and S4 libraries (4242DMSO and CX-5461 cDNA libraries respectively), the top-ranking exon fit into exon cluster group 1. This trend was similar across all library preparations, where there was no clear pattern in exon rank (reflecting one of the 8 exon clusters) and treatment scheme.

Collectively, using this approach, we found many potential exons within the mouse rDNA IGS. Exons tended to cluster/overlap into specific regions between libraries, which we refer to as exon clusters. At this point, we found no signs of differential expression from exon clusters with CX-5461 treatment.



Legend
Group 1, Group 2 , Group 3 , Group 4, Group 5, Group 6, Group 7, Group 8, Gap
Note: represents the earliest start point and latest stop point of exon groups, as called by Stringtie.

Exon cluster	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Location in IGS kB	13000-18000	44972-45315	18000-mid 25000	40030 - 41800	41800-43800	38000-41000	31000-32000	>43000-44708

Figure 17 Location of exon clusters within the mouse rDNA IGS
In each captured library, Stringtie predicted many exons that consistently cluster in several general regions of the mouse rDNA IGS. These general regions were called exon clusters for clarity. Due to exons showing consistent start and/or stop sites predicted towards the end of the rDNA IGS, these exons were classed as separate exon clusters despite showing some overlap.

4.1.3 Initial qPCR assessing IGS exon transcription from exon clusters

We performed qPCR to validate the normalised mouse IGS exons clusters found during bioinformatic analysis (section 4.1.2.4) and to further assess for signs of expression changes upon CX-5461 treatment. Primer sets were designed using the Geneious tool to amplify regions within the exon cluster groups, aiming to avoid any overlap with other clusters.

To validate our IGS exon cluster primers for specificity and efficiency, we performed qPCR and conventional PCR using the primers on Eμ-myc DNA. Primers were first tested for qPCR amplification efficiency using untreated Eμ-myc DNA which was diluted 10-fold four times. Primer pairs were determined as being efficient if they showed a 3-4 cycle difference between each dilution. Capture-seq primer sets 2, 4, 5, 6 , 7, 9, 12, 13, 14, 15, 16 and 17 were

determined as being efficient, while others showed issues with large cycles between the dilutions of DNA (see **Figure 18B** and **18C** for examples.). IGS exon cluster primers were further tested for efficiency, by assessing off-target amplification using conventional PCR and gel electrophoresis. The presence of more than one band after amplification of genomic DNA suggests that the primers can amplify elsewhere in the genome, and consequently they are not specific at amplifying noncoding RNA transcription from the IGS. Capture-seq exon primer sets 2, 3, 4, 5, 6, 8, 10, 12, 13, 14, 15, 17, 18 and 19 showed a single product (see **figure 18A** for an example), while the others produced more than one band and were excluded from use downstream. To note, primer set 16 didn't show any product, but due to amplifying in the qPCR, we suggest this be a result of an error during conventional PCR.

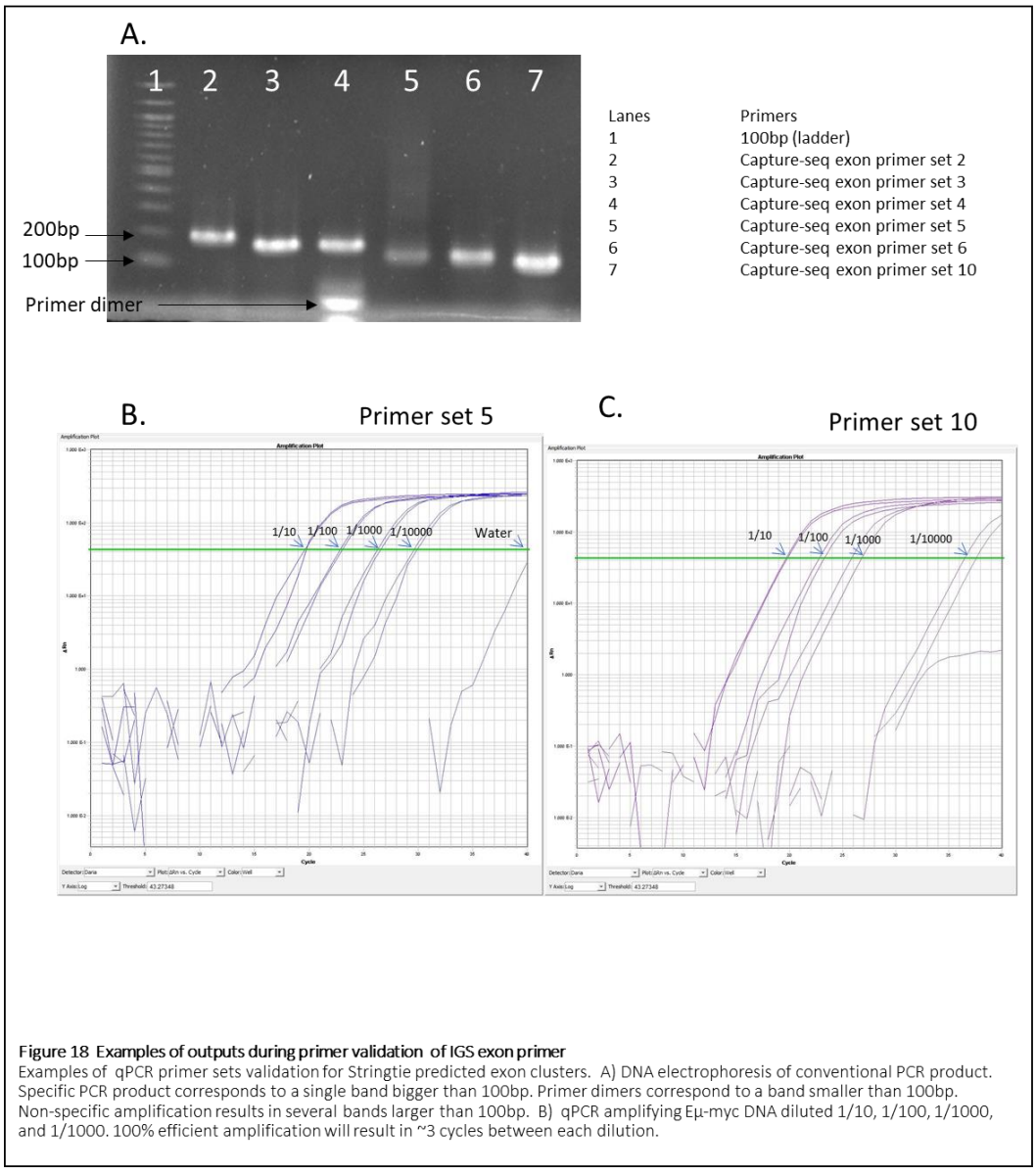


Figure 18 Examples of outputs during primer validation of IGS exon primer
Examples of qPCR primer sets validation for Stringtie predicted exon clusters. A) DNA electrophoresis of conventional PCR product. Specific PCR product corresponds to a single band bigger than 100bp. Primer dimers correspond to a band smaller than 100bp. Non-specific amplification results in several bands larger than 100bp. B) qPCR amplifying Eμ-myc DNA diluted 1/10, 1/100, 1/1000, and 1/1000. 100% efficient amplification will result in ~3 cycles between each dilution.

To assess non-coding expression from the exon clusters described, we performed qPCR using capture-seq exon primer sets 1-10 on both 4242 DMSO triplicate 1 and 4242 CX-5461 treated triplicate 1 cDNA samples (used for day two prepared cDNA libraries). qPCR analysis could allow us to both validate general transcription from these areas, as well as assess for expression differences in the presence or absence of CX-5461 treatment. To normalise for cDNA loading differences when comparing treatment and nontreatment, we measured the relative expression of transcription from the IGS normalised to expression of both GAK and GAPDH housekeeping in genes, which in theory should have the same expression levels independent of drug treatment. The relative expression results can be found in **figure 19A**. Primer sets 1, 7 and 9 are not included in the graph due to being characterised as either non-specific or inefficient. Firstly, all primer pairs showed cDNA amplification, which would suggest that all exon clusters are transcriptionally active. Further, we saw a reduction in relative expression in CX-5461 treatment cDNA relative to untreated (**figure 19A**) , suggesting that all exon clusters reduce transcription with CX-5461 treatment. We believed that this may have been a reflection of an indirect change in housekeeping gene transcription upon CX-5461 treatment stress. This was shown in the raw data, where we saw a consistent decrease in Ct for GAK and GAPDH housekeeping genes in CX-5461 treatment compared to control (data not shown) . To account for this, we plotted the qPCR Ct values for treatment compared to nontreated. As seen in **Figure 19(B)**, we found minor variation between the treatment compared to control, suggesting no change in transcription in these regions upon CX-5451 treatment. We also saw amplification using the primers designed to the gap in our Stringtie output. Thus, we rationalised there may be gDNA contamination in our cDNA samples.

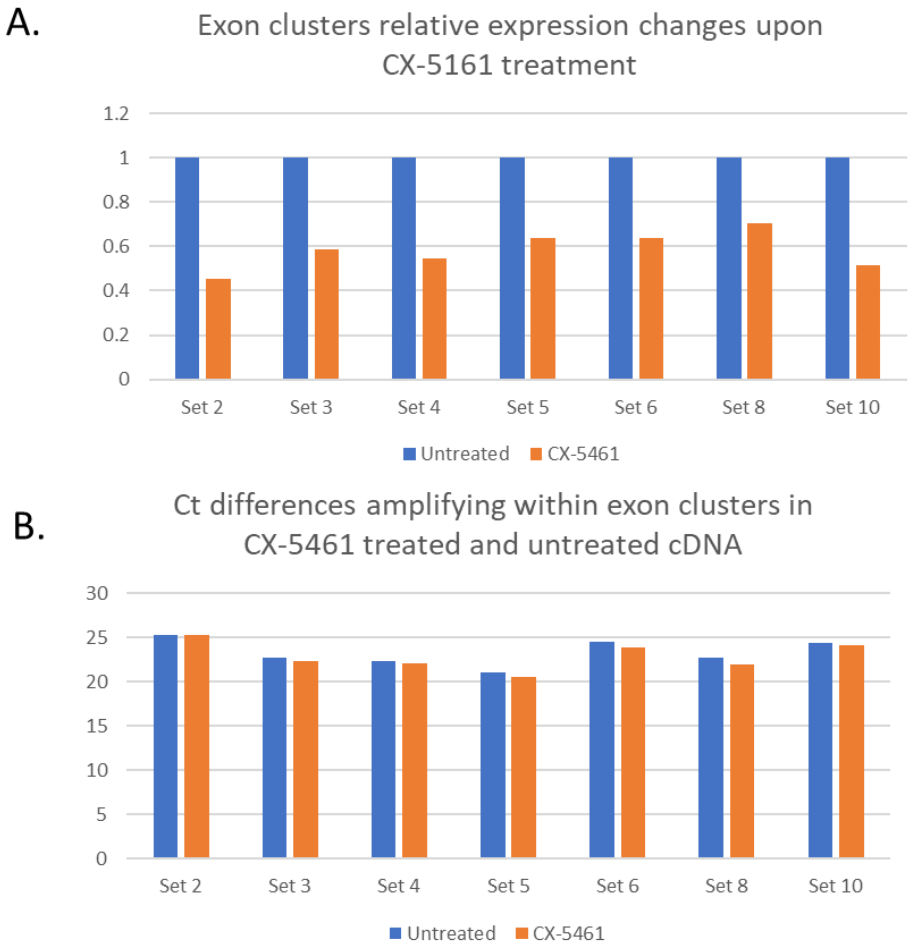


Figure 19. qPCR validating transcription from predicted IGS noncoding exons clusters
qPCR using primers designed to amplify predicted transcriptionally-active regions of the mouse IGS (capture-seq exon primers 1-10), with each primer set designed to amplify a specific exon cluster (see supplementary table 1 and figure 18) A) Relative expression normalising to house keeping genes B) No normalisation to housekeeping genes.

4.1.3.1 Assessing gDNA contamination in qPCR RNA samples

To assess for the presence of gDNA in RNA samples, we RNAse treated RNA from 4242 DMSO triplicate 1 and 2 and 4242 CX-5461 triplicates 1 and 2 RNA extraction and tested this alongside cDNA produced from the same samples by qPCR. Assuming no genomic DNA was present, RNAse treated RNA should show amplification similar to the water control. In all cases, we saw amplification in the RNAse treated wells , suggesting some gDNA contamination. Contamination varied between libraries, as for example the difference in Ct between cDNA input and RNase treated RNA input varied between libraries with primer set two was 6.269595 for 4242 DMSO triplicate 1, 13.5629455 for 4242 DMSO triplicate 2 , 5.2487185 for 4242 CX-5461 treatment triplicate 1, and 6.6597655 4242 CX-5461 treatment triplicats 2, suggesting 4242 DMSO triplicate 2 is the least gDNA contaminated assuming there was no loading error. The remaining differences in cycles between cDNA and RNAse treated RNA inputs can be seen in **Table 7**. Importantly, given that there was a Ct difference between RNAse treated RNA and RNA (both used for cDNA synthesis), we can suggest that underlying gDNA contamination cannot entirely explain our results, supporting these regions are likely to be producing transcripts. Though of note, we cannot exclude that some remaining RNA was present in the samples after RNAse treatment that were used for cDNA synthesis.

To remove gDNA, we attempted several DNAse treatment methods on our RNA and assessed for remaining contamination. It was important that all DNAse was removed after treatment, to prevent subsequent cDNA synthesis steps from being compromised. Both DNase I (origin unknown) which was removed for the RNA either by heat inactivation or RNA precipitation methods, and the use of TUBRO DNAse (Thermofisher, AM1907) following standard protocols (not included in text) were tested. In both cases, we failed to remove all gDNA (measured by qPCR with housekeeping primers comparing no reverse transcription RNA input to water control) without compromising integrity or concentration of the RNA output (measured by nanophotometer or Qubit), so the data was not included.

Table 7 Cycle differences between RNA to cDNA input and RNase treated RNA input into qPCR with capture-seq noncoding exon primers 1-10

Primers	4242DMSO1	4242DMSO2	4242 CX1	4242 CX2	H20
1	3.998392	12.92892	3.335491	4.872675	na
2	6.269595	13.56295	5.2487185	6.659766	na
3	7.0713225	15.57941	6.531418	7.596578	na
4	4.990355	12.09823	4.260963	5.783729	na
5	7.5296105	14.7417	7.3225105	8.027001	33.60457
6	5.7332285	10.30548	5.579123	6.526817	32.31664
7	6.047087	13.31495	5.276643	7.791553	34.42954
8	4.1881435	9.710773	3.834789	5.237281	34.58596
9	4.3941905	12.58312	4.192106	5.588607	32.95678
10	2.440312	8.173154	2.4226655	3.556055	30.68901

4.1.4 Final qPCR validation of mouse rDNA IGS transcription within exon clusters

As the final method used to validate our predicted IGS exon clusters and assess for differential expression between CX-5461 treated and untreated RNA, we performed a variation of the previous qPCR attempt, but this time accounting for gDNA contamination. To reduce gDNA contamination, we extracted RNA from a smaller input of cells and using the Nucleospin kit. Using less cells is thought to reduce clogging of the column, which reduces the specificity of RNA alone being retained in the column (Liam Williams, University of Auckland, personal communication). The RNA outputs were assessed for gDNA contamination via qPCR with no cDNA synthesis step, using GAPDH primers. Amplification ahead of the water control was observed, suggesting that gDNA was still present in the sample (data not included).

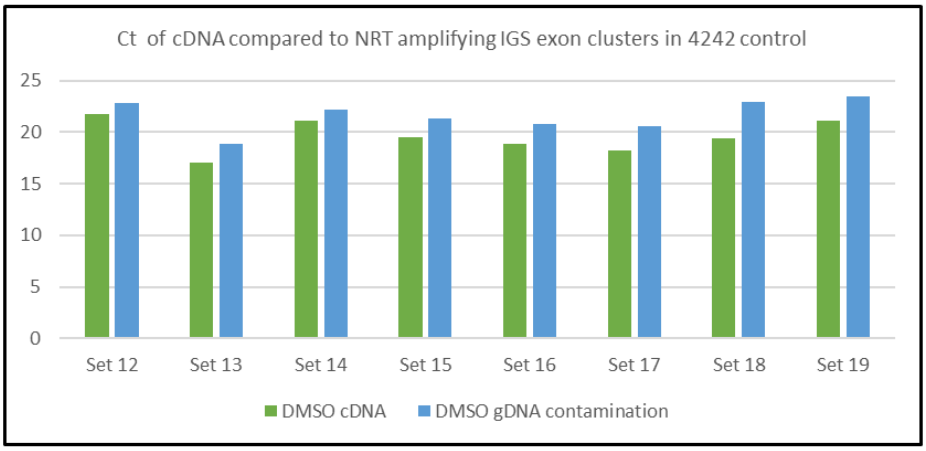
To perform qPCR and assess for IGS exon transcription whilst accounting for gDNA amplification, we performed qPCR on 500ng cDNA input from CX-5461 treated and untreated samples (in duplicates, cDNA was synthesised from 1 µg of RNA). In parallel to this, we performed qPCR on RNA that wasn't reverse transcribed (NRT control). The NRT control was made by diluting 1µg of RNA to the cDNA synthesis reaction volume, then using the volume of the diluted RNA proportional to volume of 500ng of cDNA used for qPCR. This is proportional to the amount of residual gDNA in the cDNA samples put into qPCR, and consequently accounts for background gDNA amplification. The primer sets used were Capture-seq exon

primer sets 12-19, which showed generally better amplified efficiency and produced a single product (by conventional PCR). The target regions of these primer sets can be seen in **Figure 20C**, as well as in **Appendix 1**. From the qPCR results and in reference to **Figure 20 A and B**, we saw amplification using the majority of primer sets earlier in cDNA wells than in NRT control (or gDNA background contamination) wells in both CX-5461 treated and DMSO control RNA extractions (figure depicts average of replicates), supporting that there is amplification of cDNA that originates from these predicted exon clusters in the IGS (above background gDNA signal). When assessing all the Ct values (**Figure 20D**) , for some replicates there are inconsistencies in Ct values that may reflect pipetting error (I.e. Set 13 4242 CX-5461 duplicate 2 and Set 18 4242 CX-5461 duplicate 2), that are effecting final normalised results. To address this, samples should be assessed several times with more technical and biological replicates of CX-5461 treated and untreated. Otherwise, most primer sets show consistent patterns of Ct differences between NRT and cDNA samples. Importantly, primer sets with Ct difference consistently higher between cDNA and NRT (primer sets 17, 18, 19 amplifying exon clusters 2, 4, 8) amplifying within predicted exon clusters towards the end of the IGS. We have yet to confirm transcription from exon cluster 5 due to lacking a primer set. In reference to **Figure 13**, exon clusters 2, 4 and 8 (as well as 5) are in regions with show higher normalised coverage values. We propose that these regions are the most transcriptionally active exon clusters in the IGS. Because different primer sets will have different Ct values, we cannot use this data to confirm which exon clusters are more highly transcribed.

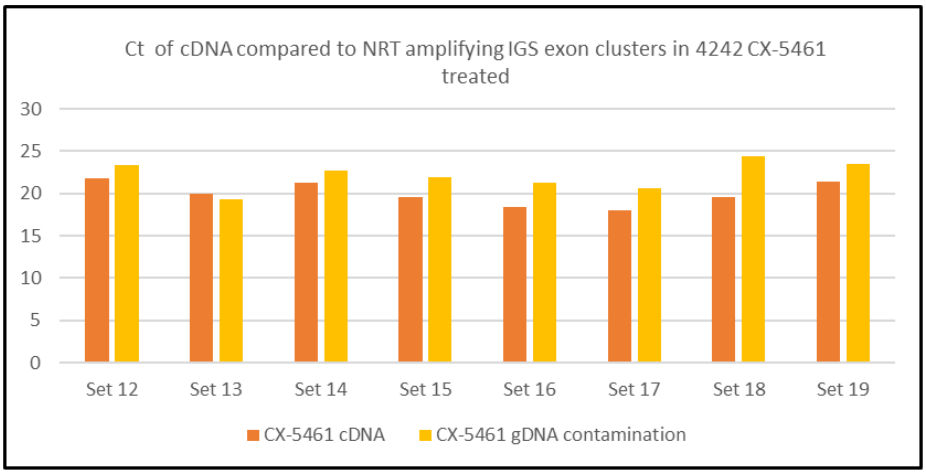
To assess for any potential differential expression of exon clusters between CX-5461 and control DMSO RNA extractions, we briefly compared differences in the average Ct values of the cDNA and gDNA wells between CX-5461 treated and untreated RNA extractions. We class an exon cluster that shows differential expression in the presence or absence of treatment, as a qPCR result that shows limited/no difference between cDNA and gDNA Ct values in one treatment scheme (Ct differences of~0), and an obvious difference in the other (Ct difference of >0). The results can be seen in **Figure 20 C**. This data does not support that any of the exons show clear differential expression upon CX-5461 treatment. The dramatic difference in set 13 is likely to be a result of pipetting error.

To briefly assess the extent of the CX-5461 driven inhibition of rRNA synthesis via qPCR, we performed qPCR with primers designed to the 47S pre-rRNA ETS (primer set ETS in **Appendix 1**). Similar to before, we compared qPCR Ct values of cDNA and underlying gDNA contamination (by direct RNA input) for 4242 DMSO control and CX-5461 treatment extractions in duplicates after amplification with the primers, as well as with a water control. The results can be seen in **Figure 21** . In terms of the average Ct values, we see cDNA samples amplifying ahead of NRT controls, again suggesting amplification is not solely a reflection of gDNA contamination. Upon normalising to NRT and in reference to **Figure 21 B** , we see very little difference in Ct values between CX-5461 and control libraries. Consequently, these results suggest that there were limited changes in levels 47S pre-rRNA levels, suggesting inhibition of RNA pol I by CX-5461 was not effective at this time. The small difference of ~0.3 of a cycle is likely to be explained by one of the NRT controls failing for the CX-5461 cycle.

A.



B.



C.

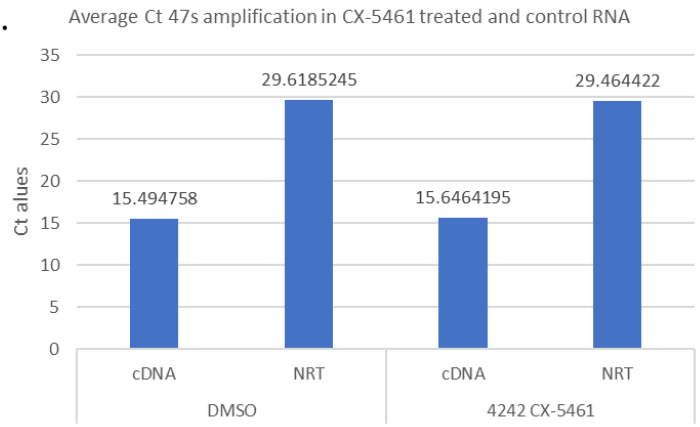
	Target region in IGS	Δ in Ct in untreated	Δ in Ct in CX-5461
Set 12	16234-16353 bp	1.16	1.59
Set 13	22074-22419 bp	1.78	-0.67
Set 14	25288-25444 bp	1.1	1.44
Set 15	31444- 31591 bp	1.83	2.29
Set 16	39300 - 39444 bp	1.92	2.89
Set 17	40061 - 40209 bp	2.39	2.58
Set 18	44526 - 44678 bp	3.53	4.79
Set 19	45075 - 45254 bp	2.34	2.06

D.

	4242 DMSO1 cDNA		NRT	4242 DMSO2 cDNA		NRT
Set 12	21.82677	22.25977	22.44409	21.11706	21.64277	23.29061
Set 13	17.63054	17.62666	18.60504	16.48836	16.61911	19.14198
Set 14	20.95828	21.56979	21.69037	20.84982	20.84644	22.60931
Set 15	19.56673	19.81871	20.79215	19.21372	19.24756	21.78372
Set 16	19.24852	19.53251	20.31512	18.36786	18.29731	21.238
Set 17	18.68225	18.30993	21.18885	17.89072	17.8195	19.93288
Set 18	19.81055	19.70281	22.67022	18.9878	19.21999	23.25408
Set 19	21.49831	21.39077	23.27993	20.49133	20.89545	23.54551
	4242 CX5461-1 cDNA		NRT	4242 CX5461-2 cDNA		NRT
Set 12	21.5019	21.54049	22.45631	21.95572	22.16446	23.47161
Set 13	16.32685	16.42461	17.98679	16.97198	30.10067	19.42109
Set 14	20.9036	20.99208	21.6655	21.59252	21.68329	22.86379
Set 15	19.40632	19.44137	21.28449	19.60113	19.8552	21.95714
Set 16	17.94966	18.38689	20.51838	18.79022	18.63457	21.41332
Set 17	17.66481	17.93275	20.77537	18.31312	18.26824	21.31092
Set 18	19.45939	19.26004	22.63668	19.73155	19.87248	25.49505
Set 19	21.23389	21.3123	23.54905	21.6228	21.57568	23.4437

Figure 20 qPCR validation of bioinformatically identified mouse rDNA IGS ncRNA exon clusters
A) and B) Graphical representations of average Ct values of cDNA and NRT control for control (A) and CX-5461 (B) treated cells amplifying cDNA from exon clusters. C) Table depicting differences in Ct values between DMSO and CX-5461 treated cDNA samples amplifying within exon clusters normalised to NRT control background amplification. D) All Ct values outputs from qPCR. Blue and red) represent untreated cDNA and NRT inputs (biological replicates) , green and navy) represent CX-5461 treated cDNA and NRT (biological replicates).

A.



B.

	ΔCt (NRT-cDNA)
DMSO	14.1237665
CX-5461	13.8180025

Figure 21 Measuring 47S pre-rRNA expression difference in CX-5461 treated and untreated samples
A) Graphical representation of CX-5461 treated and untreated (DMSO) Ct values in cDNA and NRT samples. B) Ct change differences (average NRT- average cDNA Ct values) between CX-5461 and untreated (DMSO samples)

Section 4.2 Identifying small RNAs derived from the mouse rDNA IGS and assessing small RNA response to CX-5461

In this section, we aimed to identify mouse rDNA IGS derived small RNAs in both control and CX-5461 treatment schemes and assess for differential expression. We used two different methods to achieve these aims, the first being a standard alignment approach and second by using a miRNA prediction software, miRDeep2. To assess the targets of potential miRNA species discovered, we performed GO enrichment analysis from predicted seed sequence targets.

4.2.1 Preparing small RNA for library preparation, and early sequencing output cleaning and manipulation

Small RNA sequencing was performed, as per the method, on CX-5461 treated and untreated 4242 and shUBF Eμ-myc RNA samples (in triplicates) extracted using the miRNeasy kit only (final extraction as mentioned in section 4.1.2.2). The miRNeasy kit extracted smaller RNAs, as shown by the presence of a peak to the far left of all miRNeasy extractions electrograms, which are absent in the Nucleospin extraction (see **Figure 4B**). As mentioned in the method, library preparation and sequencing were carried out by a third party, where we specified selection for sequencing of small RNAs less than 120bp in size. This, similar to that of our capture, was to avoid the sequencing the highly abundant 5S rRNA (Hori & Osawa, 1987)which is 120 nt long, and enrich for reads that may correspond to less abundant small RNA transcripts.

To quality check the small RNA raw read data, we used the FastQC software (see section 3.12.1), and as expected due to no previous cleaning carried out third parties, there was an abundance of 3' adapters from sequencing (see **Figure 22A** shows represents a single library). The read outputs for each library can be found in **Table 8**. To remove the 3' adapter, we first searched for its sequence within the raw read files using a `grep` command (confirming its presence), and then using Trimmomatic we trimmed the adapters sequence from the reads. The success of the trimming was confirmed by FastQC, which showed a reduction in 3' adapter signal (**Figure 22B**). The resulting trimmed data was used for downstream analysis.

Table 8 Small RNA library raw read output numbers

Sample name	Read count
4242DMSO1	503706
4242DMSO2	310673
4242DMSO3	380257
4242CX1	2373877
4242CX2	2486716
4242CX3	3621053
shUBFDMSO1	992046
shUBFDMSO2	1212094
shUBFDMSO3	1284558
ShUBFCX1	1107296
shUBFCX2	348743
shuBFCX3	746087

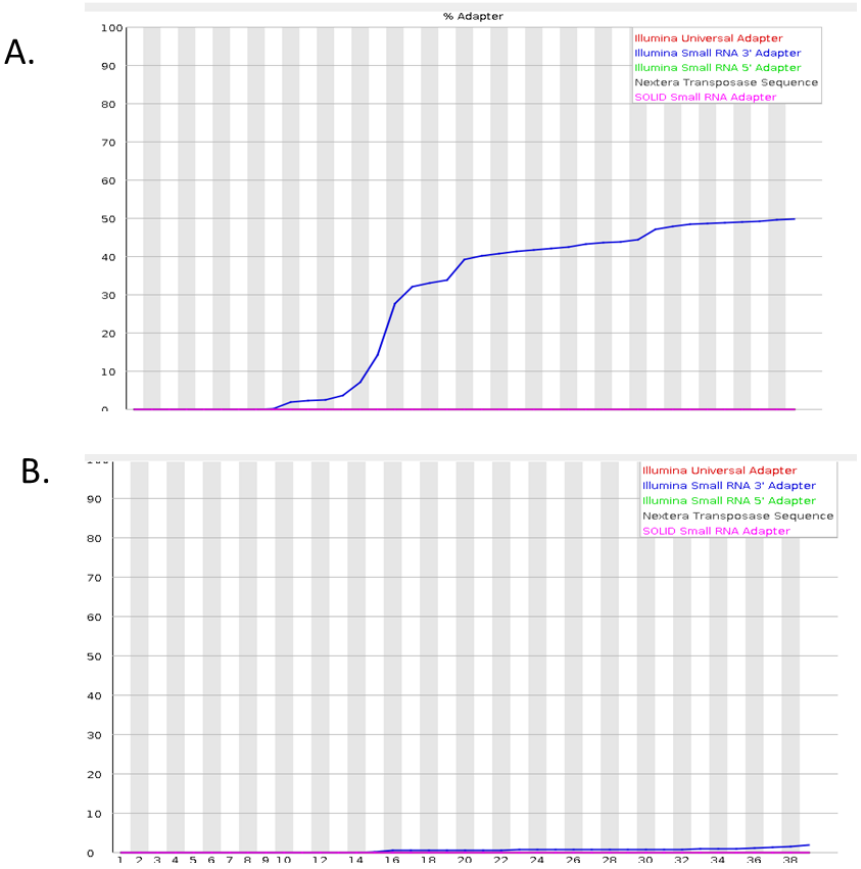


Figure 22 Pre- and post- trimming FastQC graphs of a raw small RNA sequencing output
A) Example of a raw small RNA sequencing output shows an abundance of small RNA 3' adapter . This is anticipated as the treads were not cleaned by the sequencing provider . B) Successful 3' adapter trimming is depicted by drop in blue line, after using Trimmomatic

4.2.2.1 Small RNA analysis using standard alignment/visualisation, and downstream target and GO enrichment analysis

Potential small RNAs were identified within the mouse rDNA IGS, using an alignment (STAR) and visualisation (IGV) approach. As small RNAs are classed as being smaller than 200bp, a single read (or a small collection of overlapping reads) may be a sign of a small RNA, and consequently the previous approach used during long noncoding assessment (in the previous section) would not be suitable for identifying these. To align trimmed raw read output to the mouse genome built, we utilised the STAR aligner. This aligner was selected due to having a set of small RNA parameter which have been published in the literature (see section 3.12.2.1). The

alignment output of both uniquely mapped and unmapped reads can be found in **Table 9**, compared against input alignment reads. 4242 CX-5461 treated libraries had a higher average number of uniquely mapped reads (2827215 reads on average), compared to 4242 DMSO control libraries (398212 reads on average). This observation was reversed in the shUBF line, with DMSO control libraries having on average higher read output then CX-5461 treated libraries (1162899 reads and 734042 respectively). Importantly, we saw no correlation between number the of total reads and the corresponding proportion of uniquely mapped reads (**Figure 23A**) or unmapped reads (**Figure 23B**).

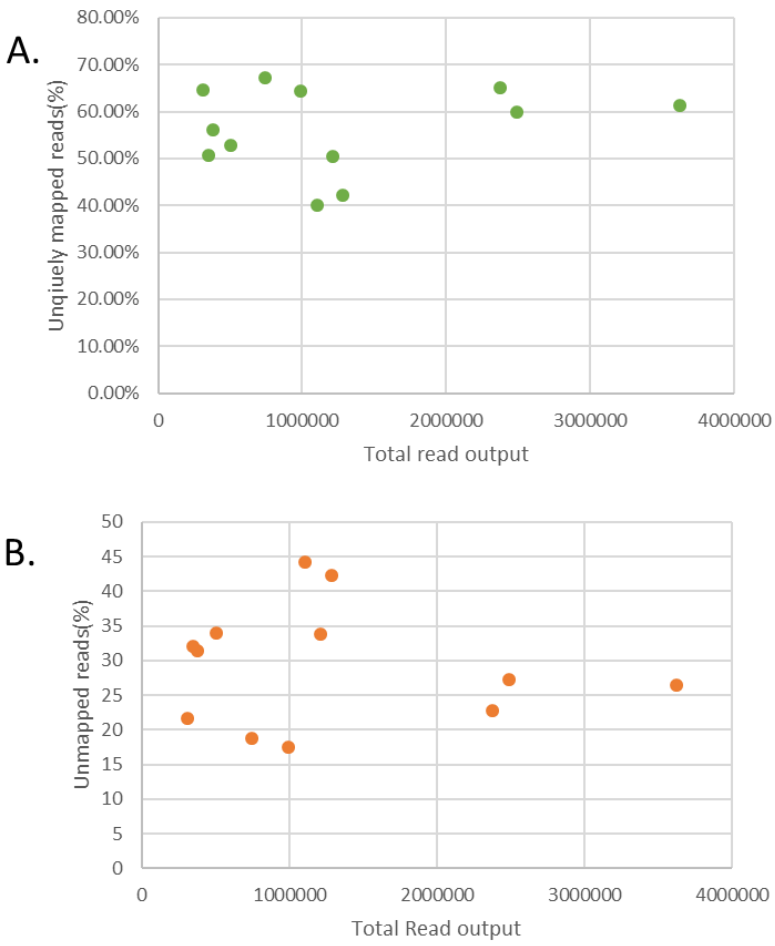


Figure 23 Small library read outputs compared to numbers of uniquely mapped and unmapped
Graphical comparison of small RNA read output to A) uniquely mapped reads or B) unmapped reads.

To convert the STAR aligner output to a readable downstream format for other tools, SAMtools (section 3.10.3 for details) was again used to produce a sorted BAM-file. Multimapping reads were not removed at this stage, as with reads as short as 50bp, the stringency of cleaning criteria was more difficult to set without a sizeable proportion of the data potentially being lost. Samtools output for the rDNA alone was visualised in IGV. As expected, there were prominent levels of alignment to the coding region (1-13400bp), that was likely to reflect partial or degraded coding rRNA products. Importantly, there was scattered alignment throughout the IGS, indicating potential small RNAs (see **Figure 24**).

Table 9 STAR aligner read outputs from small library sequencing for 4242 and shUBF lines either treated or untreated (DMSO) with CX-5461

Sample name	Input reads	STAR - uniquely mapped	Percentage of uniquely mapped	STAR- Unmapped (too short, other)
4242DMSO1	503706	263805	52.7%	7.03%, 26.91%
4242DMSO2	310673	200725	64.6%	5.21%, 16.39%
4242DMSO3	380257	213162	56.1%	6.47%, 24.98%
4242CX1	2373877	1546119	65.1%	5.70%, 17.05%
4242CX2	2486716	1492373	60.0%	5.8%, 21.39%
4242CX3	3621053	2223161	61.3%	6.33%, 20.14%
shUBFDMSO1	992046	639232	64.4%	6.71%, 10.76%
shUBFDMSO2	1212094	613233	50.5%	7.48%, 26.35%
shUBFDMSO3	1284558	540540	42.1%	8.55%, 33.78%%
ShUBFCX1	1107296	442973	40.0%	8.87%, 35.39%
shUBFCX2	348743	177246	50.8%	7.41%, 24.71%
shuBFCX3	746087	501865	67.2%	4.11%, 14.64%

In order to select reads that are more likely to belong to potential small RNAs, we used a set of criteria to a sub-select reads for further analysis. This included a requirement for the presence of a read, more specifically the start site of the read, to be the same (plus or minus 1) in more than one library, and the reads should show no polymorphism from the reference sequence. This resulted in 12 potential small RNA exons reflected by reads aligning to 12 specific regions being identified scattered across with rDNA IGS (**see Table 10**), with particular density at

~43kb. The read length associated with these locations were variable (16bp->48bp), which may reflect sequencing of a mixture of both mature small miRNA, precursor miRNAs and other species of small RNA. There was a particular clustering of potential small RNAs, noted in Table as exon 9-12, which cluster in a relatively small area around between 43320-43370 bp. Some potential small RNAs may show differential expression upon CX-5461 treatment. For example, reads determining potential small RNA exons 2, 3, 4 ,6 are only found in CX-5461 treated libraries.

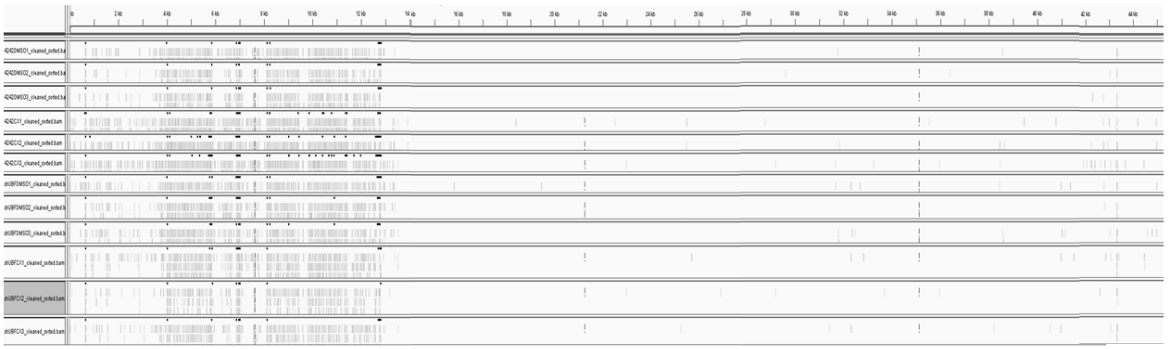


Figure 24 IGV visualisation of all small RNA reads aligning across the mouse rDNA unit
IGV alignment from the STAR aligner/SAMTOOLs output. Reads are represented by grey lines, and tracks represent library

As stated in the introduction, there resides a spacer promoter 2 Kb upstream of a rRNA promoter in the IGS of the previous rDNA repeat which encodes a promoter RNA. To assess if any of the potential small RNA exons (or combination of exons) may be in the spacer promoter region and consequently may be spacer promoter transcript, we first used BLAST (section 3.10.5) the reference sequence to clarify the region of the spacer promoter sequence. The spacer promoter sequence used for BLAST was extracted from (A. Kuhn & Grummt, 1987), and in our reference was found to span between 43296 to 43318 bp. Reads determine small RNA exons 9-12 all fit into this range, as well as a number of other reads which were not further assessed due to not fitting the criteria. We believe that reads that determined small RNA exons 9-12 are likely to be partial reads of the spacer promoter transcript. To see reads aligning to this region showed differential expression after CX-5461 treatment, we determined the

number of reads that correlate with the spacer promoter region, which can be found in **Table 11.** Dividing the average number of reads fitting this promoter region by the average of reads output for each cell variant and treatment condition, we found that on average there was little difference in the proportion of reads corresponding with the pRNA location between CX-5461 treatment and control in 4242 libraries (~$1.25x10^{-5}$ and ~$1.26x10^{-5}$ respectively). This suggests there is no differential expression of the pRNA after CX-5461 treatment. Interestingly, we saw a reduction in the proportion of reads corresponding with the pRNA in the shUBF extracts (DMSO control ~$8x10^{-6}$ and CX-5461 treatment $9x10^{-6}$). Consequently, we removed small RNA exons 9-12 from our assessment.

Table 10 Small RNA reads fitting our set criteria that may reflect small RNA exons, their location in the IGS and seed sequence

Small RNA exons	**Start site bp (read length)**	**Direction**	**Seed sequence name**
1	13523 (18bp)	Forward	A
2	22998 (16bp)	Reverse	B
3	25497 (23/46bp)	Forward	C
4	29219 (16bp)	Reverse	B
5	32352 (35 or ~42bp)	Reverse	D
6	35991 (16bp)	Reverse	B
7	41016 (35 or ~42bp)	Reverse	D
8	43068 (28bp)	Reverse	E
9	43326 (46bp)	Forward	F
10	43347 (33bp)	Forward	G
11	43351 (28bp)	Forward	H
12	43361 (28bp)	Forward	I

Table 11 Comparing read numbers aligning to spacer promoter region between different libraries

Library	Read number	Read boundaries
4242 DMSO triplicate 1	5	43321-43383
4242 DMSO triplicate 2	3	43326-43378
4242 DMSO triplicate 3	7	43326-43400
4242 CX-5462 triplicate 1	30	43306-44408
4242 CX-5462 triplicate 2	33	43323-43378
4242 CX-5462 triplicate 3	~44	43321-43404
shUBF DMSO triplicate 1	14	43322-43382
shUBF DMSO triplicate 2	6	43326-43379
shUBF DMSO triplicate 3	8	43323-43380
shUBF CX-5461 triplicate 1	9	43325-43378
shUBF CX-5461 triplicate 2	2	43323-43375
shUBF CX-5461 triplicate 3	11	43323-43382

The seed sequences of siRNA and miRNAs span from nucleotides 2-8 from the start (Grimson et al., 2007). Consequently, these coordinates were used to predict the seed sequence for the potential small RNAs. At this point, we classed all our potential small RNA reads as miRNAs for the purpose of downstream analysis. Interestingly, we found that some of our potential small RNAs share the same seed sequence. Due to the high stringency of alignment criteria, the possibility of the same seed sequence being a feature more than once is not likely a consequence of a read aligning to multiple regions. This suggested that some miRNAs within the rDNA IGS may share targets as determined by their seed sequences, and potentially IGS small RNAs may work together to regulate specific pathways. As the mouse rDNA IGS has been shown to have a series of larger and smaller duplications(Grozdanov, Georgiev, & Karagyozov, 2003) and as shown in **Figure 25** by lines representing self-symmetry , we rationalised that the maybe our small RNAs regions may reside in duplicated regions giving rise to common seed sequences. We saw that the duplicated regions and potential small RNA exons sharing a common seed sequence do not overlap, suggesting that the common seed sequence shared between the small RNAs are likely not a reflection of these major regions of sequence duplications.

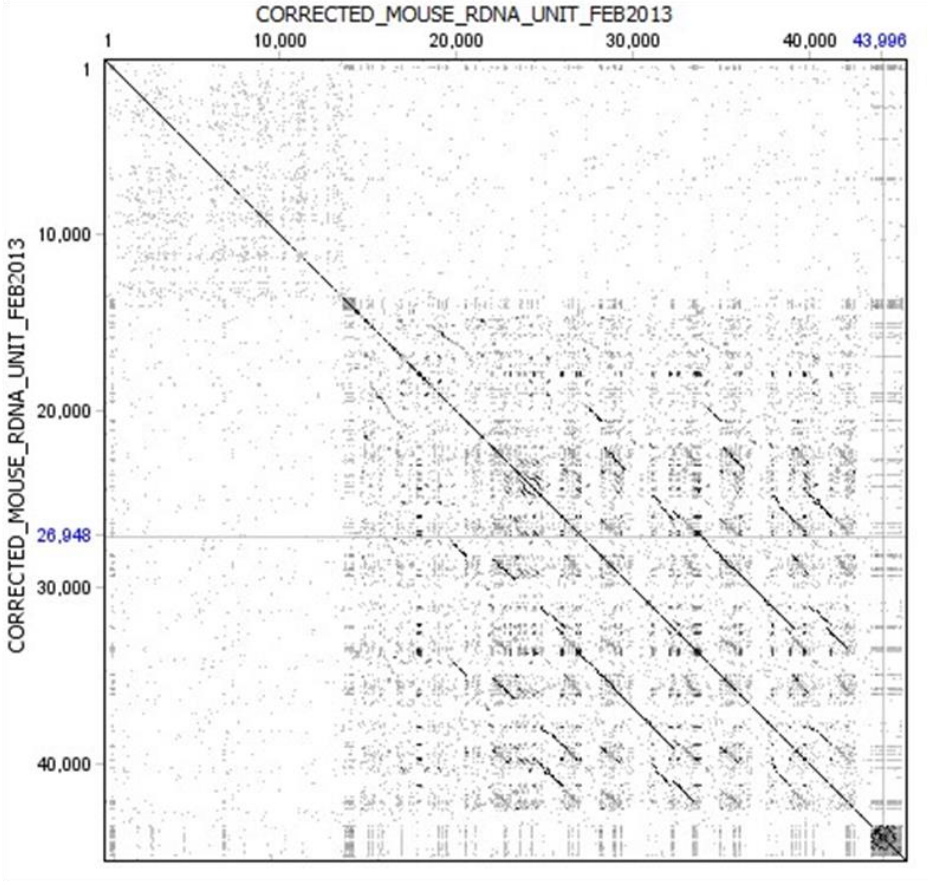


Figure 25 Self-dotplot of mouse rDNA unit produced by Geneious software
Self dotplot of the mouse rDNA unit as a visual representation of large and smaller sequence duplicates within the mouse rDNA IGS.

With the common seed sequence shared between some reads, we had a final total of 5 seed sequences of interest for further analysis (**Table 12**). Interestingly, based on the common seed sequences, some of the predicted small RNAs may be differentially expressed. For example, seed sequences B and C was only present in 4242 CX-5461 treated samples, though this may be influenced by total read output as these also had the highest read counts. This may suggest that these small RNAs are produced in response to CX-5461 treatment, which would need to be confirmed by a technical repeat of the experiment and downstream analyses, which are covered in the discussion.

Table 12 Seed sequences of our predicted small RNAs, their sequence and treatment scheme found in

Seed name	Sequence (nucleotides 2-8)	Treatment schemes present in
A	CCCCCCG	4242CX-5461 Trip 1 4242CX-5461 Trip 3 shUBFCX-5461 Trip 1 shUBFCX-5461 Trip 3 shUBFDMSO Trip 1
B	UGAUGGU	4242CX-5461 Trip 2 4242CX-5461 Trip 3
C	CAACCAG	4242CX-5461 Trip 1 4242CX-5461 Trip 2
D	CCUGCUU	shUBFDMSO Trip 1 shUBFCX-5461 Trip 3
E	UCUGUCU	4242DMSO Trip 2 shUBF DMSO Trip 3

4.2.2.2 GO enrichment analysis of potential seed sequences as determined from STAR aligner

To assess potential target pathways of our predicted seed sequences (and consequently miRNAs), we utilised Targetscan to predict gene targets, and gene ontology GO SLIM enrichment analysis to assess the gene targets for functional pathways or locations in the cell. Target scan gave highly varied outputs of matched targets dependent on each seed sequence. In the following section, we have outlined the GO SLIM analysis outputs for each seed sequence. We classed a pathway/cellular component that had an enrichment p value of ≤ 0.05 as significant which are outlined in this section, but for the purpose of discussion, we also included the lowest p value provided if no “significant” pathway or component was identified.

This was done to reduce noise of large target gene groups which may play roles in numerous cellular pathways by chance or may be present homogenously throughout the cell. A summary of the GO SLIM output for each sequence is described in **Table 13** (longer output in **Appendix 4**).

Table 13 GO slim biological processes and cellular compartment outputs from Targetscan results of seed sequences (table 11)

Predicted seed sequence	**GO slim Biological process (with p <0.05, or lowest)**	**GO slim Cellular compartment (with p <0.05, or lowest)**
Seed sequence A	embryo development (5.35E-03)	All with P of 1
Seed sequence B	Sensory perception of sound (2.93E-01)	All with P of 1
Seed sequence C	All with P of 1	All with P of 1
Seed sequence D	developmental processes (0.104), perception of sound (0.28), and systems development (0.694)	intracellular space (0.135) , cellular components (0.649)
Seed sequence E	cell-cell signalling (0.0797), embryo development (0.0891)	postsynaptic membrane (0.0311)

In order to have a significant p value according to our set criteria, many enriched GO biological pathways or cellular components became too broad to have any clear significance. At this point, we saw no significant link between the identified mouse rDNA IGS potential miRNAs and specific pathway targets.

4.2.3 Small RNA analysis using miRDeep2 software

The miRDeep2 tool is a bioinformatics tool with the capacity to identify both known and novel miRNAs from a variety of animal clades and species (section 3.12.3.2 for references). miRDeep2 was used as an additional method of predicting any novel miRNAs in our sequencing data that might show variations in abundance as a result of CX-5461 treatment. Functions within the miRDeep2 software were used for both small RNA read alignment and miRNA prediction. The output predicted several miRNAs (**Table 14**). Using this software, no predicted miRNAs mapped back to the rDNA IGS. Nevertheless, some patterns became obvious within this miRDeep2 output data. Firstly, as expected, predicted miRNAs found in more than one library were more often found within libraries that had the highest read count. For example, as 4242CX-5461 and shUBF DMSO libraries tended to have higher read numbers, and most miRNAs further investigated were found in these libraries. Predicted miRNAs 1 and 6 were found in either all, or almost all libraries, suggesting they may have relatively higher expression compared to the other miRNAs predicted. Some predicted miRNAs were only found in specific treatment schemes: miRNA 7 was predicted only in CX-5461 treated libraries, miRNA 8 was predicted only DMSO treated libraries, and miRNA 9 was predicted only in shUBF libraries. Together, this may suggest that some miRNAs identified are differentially expressed. Additionally, some predicted miRNAs showed alignment to more than one chromosome. For example, predicted miRNA 2 showed alignment to both chromosome 9 and 5 in 4242 CX-5461 triplicates 2 and 3 libraries. In all other libraries it was predicted in, it showed alignment only to chromosome 9. Together, from the miRDeep2 output, we have identified potential miRNAs from the mouse genome which may show some differential expression driven by presence or absence CX-5461 treatment.

As some of these predicted miRNAs showed potential for differential expression upon CX-5461 treatment, we performed some additional steps of investigation, regardless that none originated from the rDNA IGS. Firstly, mature miRNAs sequences were compared to previously found miRNAs in the literature, using the miRBase tool ((Griffiths-Jones, 2006; Griffiths-Jones, Grocock, Van Dongen, Bateman, & Enright, 2006; Griffiths-Jones, Saini, van Dongen, & Enright, 2007; Griffiths-Jones, 2004; Kozomara & Griffiths-Jones, 2013)). Strangely, searching the sequences of our predicted miRNAs showed no perfect 100% matches, suggesting these predicted miRNAs might be novel. Then, to assess if these miRNAs may target pathways that

may show some potential relevance to the rDNA or RNA pol I inhibition by CX-5461, we again predicted targets genes using the seed sequences (as determined by the miRDeep2 output) and used the output for GO enrichment analysis. The results of this GO slim analysis can be found in **Table 15**. We chose to include biological processes or cellular compartment enrichment outputs with p values of less than 0.05. The full list of GO enrichment summaries (where the top three with the lowest p values were included) for each miRDeep2 predicted miRNA can be found in the **Appendix 5**. The significance of these conclusions will be discussed.

Table 14 miRNAs predicted by miRDeep2 found in more than one library

Number	Name	Sequence	Libraries found in	Align to Chr	miRDeep2 score	Matches to miRBase
1	mir1	CCACCACUGCCACCAGGCC	4242DMSO1	10		no match
			4242DMSO2	10		
			4242DMSO3	10		
			4242CX1	10		
			4242CX2	10		
			4242CX3	10		
			shUBFDMSO1	10		
			shuBFDMSO2	10		
			shUBFDMSO3	10		
			shUBFCX1	10		
			shUBFCX2	10		
			shUBFCX3	10		
2	mir2	CAGGAGGUCCCUGGUAAGC	4242CX1	9		no matches
			4242CX2	9		
				5		
			4242CX3	9		
				5		
			shUBFDMSO1	9		
			shUBFDMSO3	9		
3	mir3	AGGGAGGUCCCUGGUGGUU	4242CX1	7		no matches
			4242CX2	7		
			4242CX3	7		
			shUBFDMSO3	7		
			shUBFCX1	7		
			shUBFCX3	7		
4	mir4	AGGCAGGUCCCUGGUCCUC	4242CX1	4		no matches
			4242CX3	4		
			shUBFDMSO1	4		
			shUBFDMSO3	4		
5	mir5	CCGCCGCUGCCACCAGCCC	shUBFDMSO1	17		no matches
			shUBFDMSO3	17		
			shUBFCX1	17		

6	mir6	CCACCACUGCCACCACAGU	4242DMSO1	16		no 100% matches, similar to chromosome 2 miRNA (3 mismatches)
			4242CX1	16		
			4242CX2	16		
			4242CX3	16		
			shUBFDMSO1	16		
			shUBFDMSO2	16		
			shUBFDMSO3	16		
			shUBFCX1	16		
			shUBFCX2	16		
			shUBFCX3	16		
7	mir7	CCACAGCUGCCACCAGGGC	4242CX2	14		no matches
			4242CX3	14		
8	mir8	CCGGACGAGCCCCCAAAUG	4242DMSO1	14		no matches
			4242DMSO3	14		
9	mir9	CCACCGCUGCCACUAACAC	shUBFDMSO3	13		no matches
			shUBFCX3	13		
10	mir10	CCACCACUGCCACCAGGUU	4242CX2	12		no matches
			shUBFDMSO1	12		
11	mir11	CCACAGCUGCCACCACAAC	4242DSMO3	12		no matches
			shUBFCX1	12		
12	mir12	CCGGACGAGCCCCCAAUGU	4242DMSO2	17		no matches
			shUBFCX3	17		

Table 15 miRDeep2 predicted miRNAs GO slim pathway enrichment output from Targetscan-predicted targets

Predicted miRNA	**GO slim Biological process (with p value <0.05)**	**GO slim Cellular compartment (with p value<0.05)**
1	-	-
2	Developmental process (p value 0.042)	-
3	Muscle Organ development (p value 1.18×10^{-4}) Mesoderm development (p value 3.41×10^{-3}) Cellular component morphogenesis (p value 1.93×10^{-2})	Postsynaptic membrane (p value 2.78×10^{-2})
4	Ectoderm development (p value 3.99×10^{-7})	-

	MAPK cascade (p value 2.16 x 10^{-2})	
5	-	-
6	-	-
7	Nucleobase containing compound metabolic process (p value 2.32x10^{-3})	Cell junction (p value 1.84x10^{-2})
8	-	-
9	-	-
10	-	-
11	Nucleobase-containing compound metabolic process (p value 2.96x10^{-2})	Nucleus (p value 4.37x10^{-2})
12	-	-

We also attempted miRDeep2 mapping and prediction using the rDNA alone for indexing, to see if this produced some novel miRNA prediction from the rDNA. Mapping was successful for all libraries, but during miRDeep2 prediction it was found ¾ libraries gave an error and prediction wasn't successfully completed. As all libraries were processed in the same way, the reason why some executed without error is unclear. Only libraries 4242DMSO triplicate 1, 4242CX-5461 triplicate 3, shUBF DMSO triplicate 3 and sHUBFCX-5461 triplicate 3 continued without error. Unfortunately, the output of these libraries using miRDeep2 predicted no novel (or known) miRNAs aligning to the rDNA.

4.3 shRNA characterisation in the Eμ-myc model

To further characterise our mouse lymphoma model used in the previous experiments, we performed several experiments. Firstly, to check if UBF levels were reduced in our Eμ-myc line carrying the LMP vector with a short hairpin to UBF, we assessed UBF levels both in terms of mRNA (via qRT-PCR) and protein (via western blot). For future experiments which may require us to knockdown or overexpress noncoding transcripts of interest, we tested a Lentiviral transduction system in several cell lines (including mouse cell lines), where infection could be assessed by a fluorescent reporter. The results of these experiments will be discussed in the following paragraphs.

4.3.1 shUBF knockdown confirmation analysis

To confirm the UBF knockdown in 4242 shUBF GFP LMP line which we received from another laboratory , we performed qPCR and Western blot assays to measure the abundance of the UBF mRNA or protein (respectively).

Using qPCR, we measured relative UBF gene expression in shUBF compared to Eμ-Myc 4242 control. We amplified both UBF1 and UBF2 isoforms, that differ via alternative splicing generating an additional exon in UBF1 isoform, and consequently using primers designed to this unique region allowed UBF1 specific amplification. We performed normalisation to several housekeeping genes, including β-actin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and GAK. All primer set sequences can be found in **Appendix 1**. qPCR was performed several times with 2 different qPCR kits using RNA isolated from 4242 or UBF shUBF cells. The results are shown in **Figure 26 (B)**. When looking at the raw data, there seems to be a pipetting error, resulting in reading inconsistencies between qPCR experiments (data not shown). Overall, from the qPCR data, we believe that the shUBF knockdown line shows relative UBF gene expression similar or marginally less than the 4242 control line, suggesting that the knockdown

has been lost. The level of UBF1 relative expression shows an increase of mRNA abundance in UBF lines, but this is likely to be a result of the large variability in qPCR readings.

We also performed western blots using protein extracts from 4242 control and shUBF cell lines. These were performed on multiple occasions by myself, as well as several colleagues in the lab. We used both α-UBF and α-H2A antibodies, with the latter being used as a loading control. In **Figure 26A**, we show an example of the western blot UBF protein is detected as a doublet ~ 75kDa -100 kDA (UBF 1 and UBF 2 isoforms). By eye, we could not detect any difference in brightness of the protein bands between the control and shUBF knockdown samples. This is not due to the different protein loading, as shown by the lack of difference in the H2A band acting as a loading control.

Together, both the qPCR and Western blotting results, show that UBF knockdown efficiency in shUBF cells was low (if any).

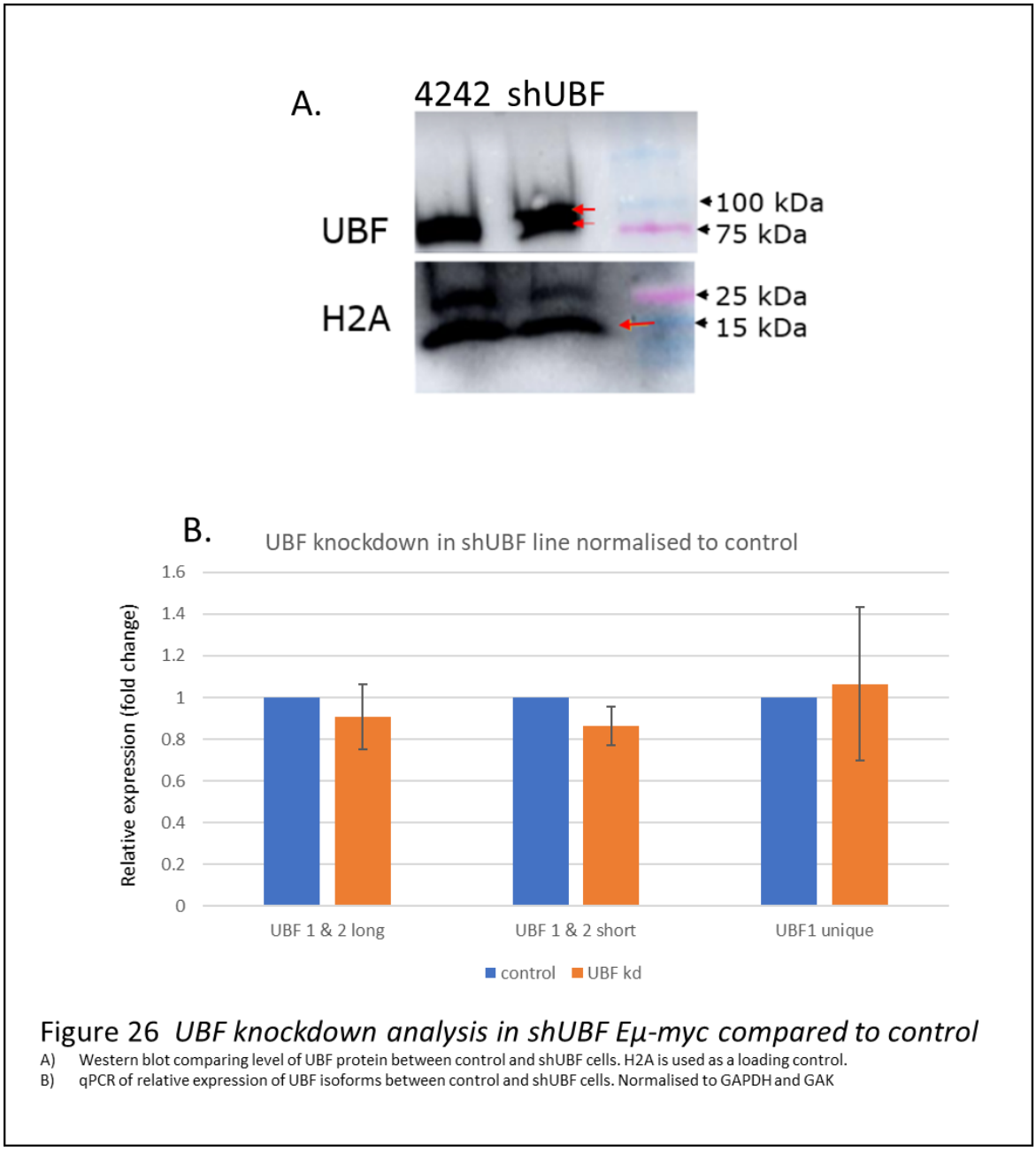


Figure 26 *UBF knockdown analysis in shUBF Eμ-myc compared to control*

A) Western blot comparing level of UBF protein between control and shUBF cells. H2A is used as a loading control.
B) qPCR of relative expression of UBF isoforms between control and shUBF cells. Normalised to GAPDH and GAK

44.3.2 Lentiviral transduction

The role of noncoding RNAs of interest, found in sections 4.1 and 4.2, could be further characterised by assessing cellular response to their knockdown/overexpression. In preparation of this, we performed lentiviral transduction assays, which could be applied to knockdown or overexpress the ncRNAs, in a number of cell lines.

To test the efficiency of the lentiviral transduction in a panel of cell lines, we first attempted virus packaging and subsequent infection using reportedly highly transfection/infection efficient HEK cells (Swift, Lorens, Achacoso, & Nolan, 2001). 24 hours after transfection, we saw a large number of HEK cells fluorescing with the lentiviral RFP (**Figure 27 A**). Importantly, efficiency of infection of low confluency HEK target cells was relatively high (**Figure 27 B**).

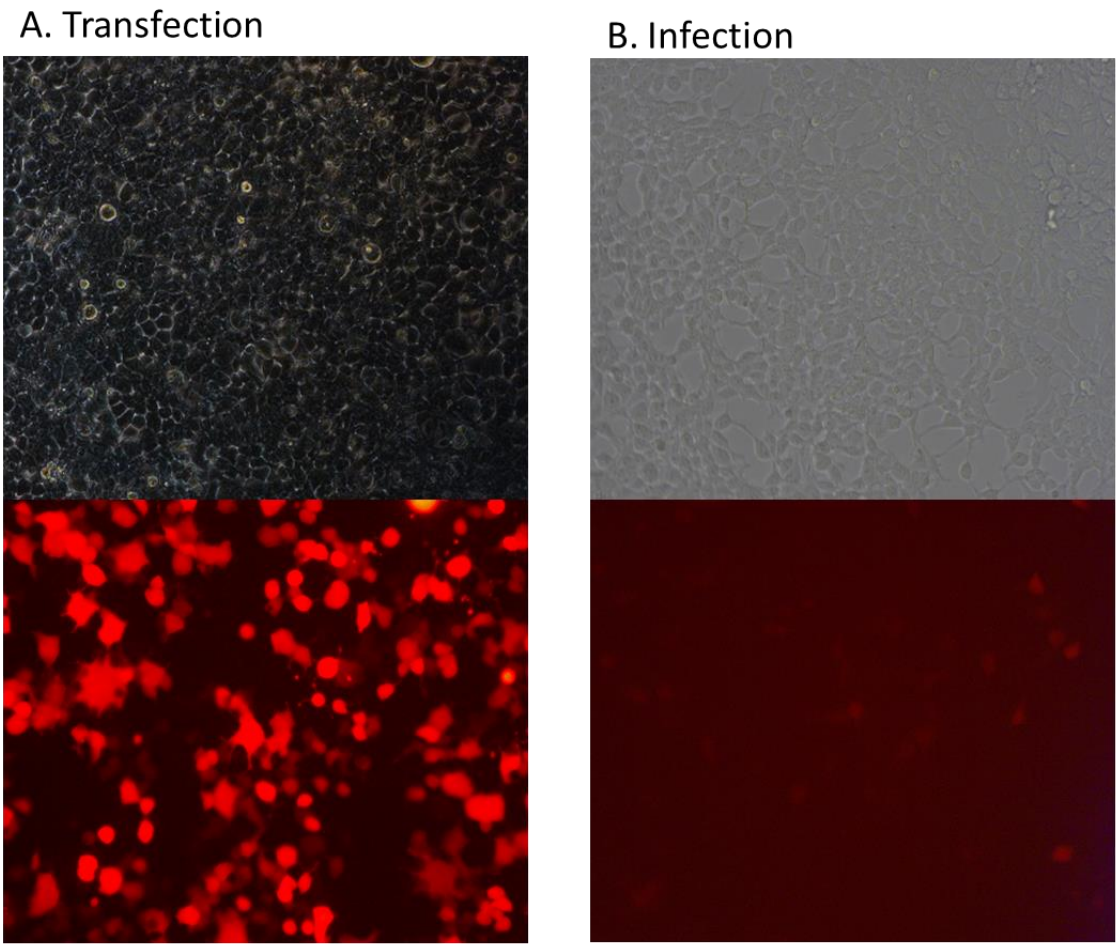


Figure 27 Transfection and infection into HEK using lentiviral system

A)HEK packaging cells 24h after transfection with pIVX shRNA RFP lentiviral system (top image: brightfield view, bottom image: txRED filter). B) HEK packaging cells 18h after infection with lentiviral pIVX shRNA RFP system (top image: brightfield view, bottom image: txRED filter)

Due to the efficacy of infection in HEK cells, and the downstream aim of assessing the effect of changing the expression of our ncRNAs of interest, we attempted to repeat the lentiviral transduction procedure but using the 4242 lymphoma line as the target cells. Again, packaging appeared successful in the HEK, as assessed by red fluorescence (**Figure 28 A**). Due to the low infection efficiency previously reported in suspension lines, we added an extra step of pre-coating 6 well plate to be used for incubation of 4242 target cells with retronectin. As the standard protocol of lentiviral transduction given to us by our collaborators had a large range suggested for target cells confluency, we used both the minimum and maximum suggested confluency's for the cells ($5x10^5$ cells per well, and $1x10^6$ cells per well). As mentioned in the methods, wells containing 4242 cells that were not infected with viral media, and 4242 in a well with no rectronectin, were included as a controls for background fluorescence level and the effect of retronectin respectively. After one day of infection, the 4242 target cells had to be transferred to a new 6-well plate, due to technical issues with contamination leaving us with only one of the seeding densities. After several days, we detected no difference in fluorescence between control (uninfected, **Figure 28 C**), infected Eμ-myc (**Figure 28 B**), and no rectronectin (**Figure 28 D**) with all cells appearing red. Consequently, we deduced that the infection of the Eμ-myc line was not successful, and rather the cells appearing red was a result of auto fluorescence.

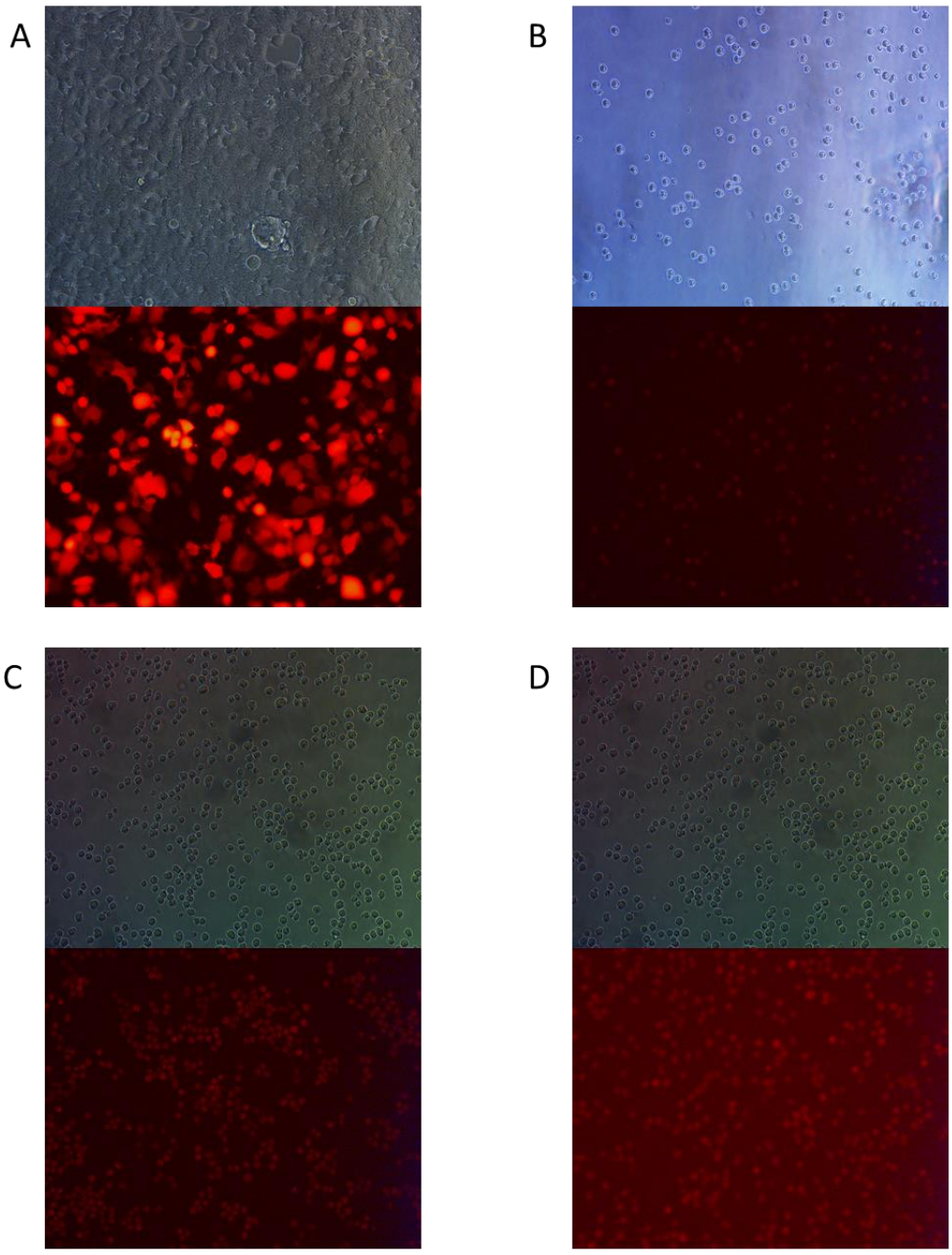


Figure 28 Lentiviral infection into Eμ-myc from HEK packaging cells

A) HEK packaging cells 24h after transfection with plVX shRNA RFP system (top image: brightfield view, bottom image: txRED filter). B) HEK packaging cells 18h after infection with lenti-viral plVX shRNA RFP system (top image: brightfield view, bottom image: txRED filter) B) Eμ-myc cells after infection with viral media with all lentiviral packaging plasmids but no plVX ShRNA RFP in rectronectin coated wells (top image: brightfield view, bottom image: txRED filter). C) Eμ-myc cells after infection with lentiviral plVX shRNA RFP system in rectronectin coated wells (top image: brightfield view, bottom image: txRED filter D) Eμ-myc cells after infection with lentiviral plVX shRNA RFP system without retronectin (top image: brightfield view, bottom image: txRED filter)

Due to the lack of detectable infection of the Eμ-myc cell line, we decided to attempt infection of a second mouse cell line we had available, MEFs, which could also be used in downstream validation of the role of IGS ncRNA. HEK cells were used for packaging, as well as a positive control for infection alongside the MEFs. Transfection of HEK was highly effective, as in previous experiments . Similar to the Eμ-myc cell line, MEFs infected with virus-containing media from HEK packaging cells showed the same level of red fluorescence (**figure 29 B**) as MEF controls not infected with virus-containing media (**Figure 29 D**), suggesting infection was unsuccessful.

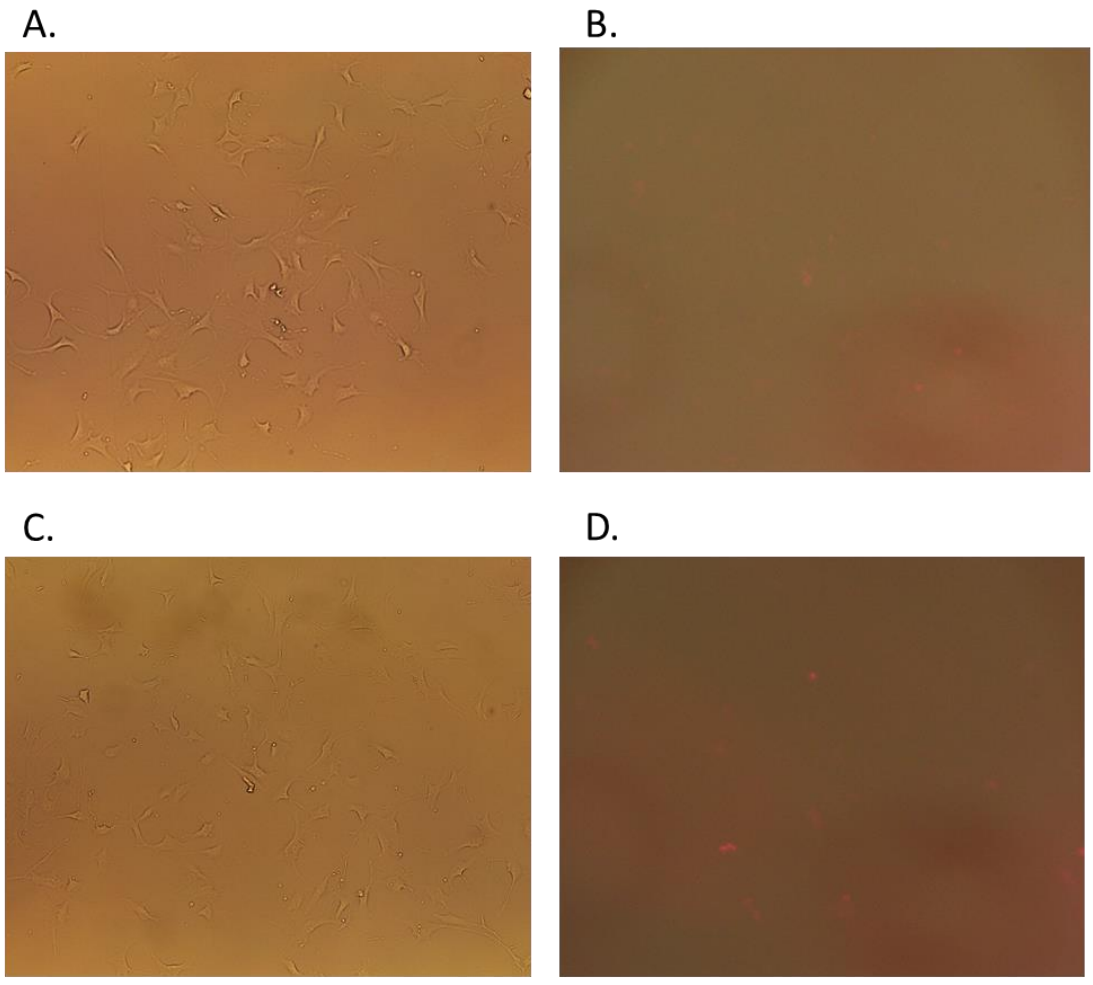


Figure 29 Lentiviral infection into MEF from HEK packaging cells
A) and B) MEF cells after infection with plVX ShRNA RFP lentiviral system (A: brightfield view, B: txRED filter). C) and C) MEF (no infection) (A: brightfield, B: txRED filter)

Together, this suggested that though our lentiviral system was highly effective in the human kidney cell line, the infection of mouse lines tested was unsuccessful. Further, this proposes that using our Lentiviral system is not the feasible approach for knockdown or overexpression analysis of discovered IGS ncRNAs of interest in 4242 and MEF cell lines.

5. Discussion

5.1.1 Identification of noncoding exons within the Eμ-myc rDNA IGS

Currently, very few mammalian rDNA IGS ncRNAs have been characterised in literature, and to the best of our knowledge, only one, the pRNA (A. Kuhn & Grummt, 1987; Santoro et al., 2002), has been characterised in the mouse rDNA. Interestingly, all characterised IGS ncRNAs have been proposed to serve roles in rDNA transcription and copy number regulation, and nucleolar protein sequestration, suggesting that the IGS has some functionality through ncRNA expression (Jacob, Audas, Mullineux, & Lee, 2012). Here, we aimed to broaden our understanding of the transcriptional status of the mouse rDNA IGS using a capture-seq approach that enriched for rDNA IGS lncRNA transcripts. Utilising capture-seq RNA data from 11 mouse cDNA libraries, we have predicted many noncoding exons within the mouse rDNA IGS. Exons were not always consistently predicted between libraries, but instead clustered in several different areas. For simplicity, we characterised these regions as exon clusters. Exon clusters at the end of the rDNA IGS (specifically exon clusters 2, 4, 5 and 8) are predicted with more consistency (i.e. are more often predicted with the same start and end sites) and are present in regions that overlap with peaks in normalised coverage plots (section 4.1.2.2) where higher normalised coverage supports higher rates of transcription. Therefore, they are more likely to reflect potential transcriptionally active regions of the rDNA IGS. We have yet to confirm these observations but could do so with a variety of downstream analyses to be discussed later. We also have yet to establish how exon clusters are transcribed: whether each predicted exon cluster acts as a single transcriptional unit, linked to the production of one/a few transcripts produced from exons within an exon cluster, or whether each predicted exon is an independent transcriptional unit.

Other predicted exon clusters (exon clusters 1, 3 and 7) were present inow regions that show either plateau or low peaks in normalised coverage values (suggesting these regions have low RNA-seq read alignment reflecting low transcriptional activity), as well as exons in these

clusters were predicted with less consistency and tend to overlap between libraries. We believe that some of the exons predicted in these clusters may be a reflection of gDNA contamination of the RNA extractions.

The exon clusters that we identified at the 3' end of the IGS, which we believe are transcriptionally active regions, co-locate with signals of H3K36me3, H3K4me1 and H3K4me2 active methylation marks that were previously observed in the mouse rDNA (Zentner et al., 2014). Of note, some active chromatin mark signals at the 3' end of the IGS are not as high as other regions of the rDNA IGS, specifically between 13kB to ~25kB. We had many exons predicted between 13 kB and ~25 kB, but as this region is influenced by higher mapability according to Zentner et al (2014) (where high mapability reflects better sequencing ability of a region), along with us observing low normalised coverage in this area (reflecting low level of transcription), we propose this region of the IGS may have less transcriptional activity than what can be predicted by the presence of active histone marks. Some exon clusters we identified overlap with regions identified by Zentner et al to be enriched for chromatin marks that represent transcriptional silencing. However, the rDNA is a mosaic of active and silent repeats, therefore these silent marks may derive from the silent rDNA fraction and do not automatically preclude transcriptional activity producing noncoding transcripts from the active copies in this region.

To validate transcription from the predicted exon clusters, we performed qPCR on cDNA. Our qPCR results supported some transcription from all predicted exon clusters, with exon clusters towards the 3' end of the IGS showing the most convincing signs of transcription (in support of the higher normalised coverage values discussed earlier). Transcription from the exon clusters was measured by comparing the Ct value of the cDNA to the Ct value of a NRT control. This method was used due to the presence of residual gDNA found in all RNA samples, which amplified in our cDNA samples. gDNA contamination was not apparently present in the Bioanalyser data assessing RNA used for library preparation. gDNA contamination predicted by a Bioanalyser analysis have been reported to appear as a large peak between the two ribosomal RNA peaks, or a large peak following the 28S peak, but these representations are likely to reflect high levels of contamination with gDNA (see (Caruana & McInnes, 2004). But, by using qPCR we could detect the presence of gDNA in RNA samples used for library

preparation. Residual DNA was also detected in independent RNA extractions from Eμ-myc cells which were treated following the same protocol for qPCR. Despite attempts to remove this gDNA contamination by several methods (described in 4.1.3), we were not able to 100% eliminate it. Comparing Ct values of cDNA and NRT control accounted for amplification attributed to gDNA contamination, and with this we found exon cluster Ct values were between 1 to ~4.8 cycles lower than the corresponding NRT controls. Exon clusters 2, 4 and 8 that reside at the 3' end of the rDNA IGS had larger differences in Ct when comparing cDNA to NRT control (amplified by primers 17-19), which we propose further supports that these regions are transcriptionally active. We have yet to test exon cluster 5 via qPCR, but as it is also in the high coverage region of the IGS, we might expect similar results, but this remains to be tested. Though Ct values when comparing cDNA to NRT controls may not be dramatically different, we have to consider the repetitive nature of the rDNA. When amplifying transcripts from the rDNA, gDNA contamination will have a stronger signal than when amplifying transcripts from a single copy gene, thus can skew the significance of the Ct difference between background and sample. Further, the smaller Ct differences between cDNA and NRT amplifying exon clusters earlier in the IGS is additional support that these exons may be a reflection of gDNA contamination and sequencing bias. It is difficult to compare qPCR results between primer sets to assess relative transcriptional levels between exon clusters, as different primer pairs can amplify with different efficiencies, thus giving different Ct values for the same amount of template.

The small molecule inhibitor CX-5461 targets RNA pol I transcription to inhibit rRNA synthesis and consequently induce cell death in cancer cells whilst having negligible effects on normal cells. Preliminary results by collaborators (Prof. Ross Hannan, personal communication, unpublished) observed both an increase of transcription from the rDNA IGS and an increased ratio of active (to inactive) rDNA repeats in malignant Eμ-myc cells compared to pre-malignant. It has been suggested that small molecule inhibitors could be utilised to interfere with the function of ncRNAs that play leading roles in the development and progression of some cancers (Tsai, Spitale, & Chang, 2011). Our group hypothesised that if IGS ncRNA transcripts have a critical role in Eμ-myc malignancy (leading to their increased expression in malignant cells), their expression may be affected by CX-5461 driven pol I inhibition which would provide sensitivity to CX-5461. We assessed for CX-5461-treatment induced IGS ncRNA expression changes by performing Capture-seq on RNA from both from CX-5461 treated and non-treated

control Eμ-myc cells. We ranked all exons in the captured cDNA libraries according to the normalised coverage. We found no apparent changes in exon coverage patterns following treatment with CX-5461, which was supported by qPCR. These results may not be surprising, as although we calculated an IC_{50} similar to the reported concentration in cytotoxicity assays, and we used a drug concentration double that of the IC_{50} required for RNA pol I inhibition after 1 hr (Bywater et al., 2012), we saw no change in 47S pre-rRNA in our CX-5461 treated samples compared to control via qPCR. This would suggest that CX-5461 was not effectively reducing RNA pol I transcription in the treatment timeframe we used. Consequently, our results do not distinguish whether the IGS exons show differential expression upon CX-5461 treatment.

A reduction of UBF protein has been shown to reduce RNA pol I localisation to the rRNA and consequently hinder rRNA transcription. However, UBF binds across the entire mouse rDNA unit (Herdman et al., 2017), and consequently may play other regulatory roles in the rDNA. We wondered whether this may include chromatin structure modification that affects IGS ncRNA transcription. Thus, we performed capture-seq on shUBF Eμ-myc variant, to assess differences in rDNA IGS ncRNA expression upon RNA pol I inhibition by UBF depletion, CX-5461 treatment or both. We saw no apparent changes in normalised exon coverage patterns in shUBF libraries compared to control libraries. However, when we later checked the extent of UBF knockdown in shUBF and 4242 cells, both by Western blot and qPCR assays, we found UBF mRNA and protein levels were similar in 4242 and shUBF cells. The shUBF line was shown to still express GFP that is present on the shUBF construct (data not shown), suggesting that cells which retained the construct but silenced the shRNA may have been selected for in the shUBF line. A similar observation has been reported in the literature (Fish & Kruithof, 2004). Consequently, we have yet to determine if our exon clusters show expression changes upon UBF knockdown-dependent RNA pol I inhibition, but this could be assessed by repeating capture-seq or qPCR using a Eμ-myc cell line with a definite UBF knockdown potentially through another form of RNAi.).

We have yet to test if exon clusters predicted at the 3' end of the IGS correspond solely to the pRNA. The pRNA is transcribed from the spacer promoter to ~200bp upstream of the pre rRNA transcription start site (Christine Mayer et al., 2006). This transcript is processed to 150-300bp (Santoro et al., 2002). Consequently, some of the 3' exon clusters predicted downstream from

the start of the spacer promoter (specifically exon clusters 2, 5 and 8) may represent partial pRNA transcripts. The pRNA is of a size detectable by our system, and consequently its detection serves as a positive control of the capture. Consequently, we still need to discern if all 3' exon cluster groups represent the pRNA, or if other transcripts are also transcribed from this region.

5.2 Small RNA-seq data reveals small RNAs with potential differential expression upon CX-5461 treatment

We performed size selected small RNA-seq to identify mouse rDNA IGS small RNA exons (corresponding to RNAs less than 120bp long). The small RNA-seq data was analysed in two ways. The first method applied a basic alignment strategy using parameters designed for small RNA reads. Due to both the concept that many small RNAs (including miRNA and siRNAs) are smaller than the shortest current Illumina read length (50bp), along with the assumption that mouse rDNA IGS small RNAs will be of low abundance, we proposed that any read aligning to the IGS in the small RNA-seq data set may reflect a potential small RNA exon. This method revealed several regions with read alignment in the IGS that are present in more than one library preparation. These reads were of a variety of lengths (**Table 10**). There were a large number of reads that align around spacer promoter region, which may reflect partial spacer promoter transcripts. Importantly, there was a high level of read alignment across the rRNA coding region. These may reflect small RNAs coming from the coding region, or (due to size selection) may represent degraded rRNA transcripts. Therefore, at this point we cannot rule out that the reads aligning to this region do not derive from degraded longer ncRNAs.

The second method we used to identify miRNAs in the mouse rDNA was to employ miRDeep2, as it has been reported to successfully identify mouse miRNAs in the literature (Dhahbi et al., 2013; He et al., 2012). Our sequencing efforts resulted in several predicted miRNAs, none of which mapped to the mouse rDNA IGS. Strangely, none showed any 100% similarities to miRNAs within the miRbase database which contains all published miRNA's, and with only one showing some similarity to miRNAs in the database. 12 miRNAs were found in more than one library and were assessed further. Most (10 of 12) were found in a mixture of both CX-5461 treated and untreated libraries, while the others were predominantly found in either CX-5461

or untreated control libraries. Although CX-5461 treatment did not result in decrease pre-rRNA 47S levels (suggesting it had not yet inhibited RNA pol I transcription), it is possible that CX-5461 treatment may have induced some changes in small RNA expression that produced this observation, but it is also possible that this apparent difference is spurious. Using the number of libraries, they were found in as a measure of abundance, miR1, miR6, miR 3 and miR 2 are the most abundant predicted miRNAs and therefore may be of interest to assess whether the play any functional role. The miRDeep2 software did not predict any miRNAs derived from the mouse rDNA IGS. However, this does not mean there are no miRNAs encoded within this region, as the software requires a minimum of 5 reads mapping to a region to be able to predict a miRNA (Friedländer, Mackowiak, Li, Chen, & Rajewsky, 2011) , and rDNA IGS miRNAs may be of low abundance that they are unable to meet that threshold. Repeating the small RNA-seq with deeper coverage might identify miRNAs encoded in the IGS, as well as looking into publicly available data sets for IGS miRNAs which may have not been identified due to the common practice of excluding the rDNA from sequencing data processing because of its repetitive nature .

Using the seed sequences (as described in section 4.2.2.1) we identified targets of our miRNA/small RNAs predicted by the STAR aligner and miRDeep2 methods. Next, we performed GO enrichment analysis (independently for all miRNA/small RNAs predicted by either of the methods) to determine pathways which are targeted by the potential miRNAs. In regard to the small RNAs identified by STAR aligner, several shared the same predicted seed sequence. Consequently, if these small RNAs are miRNAs, it might suggest that the IGS produces a number of miRNAs that collectively target one or several pathways. The GO enrichment analysis revealed significant enrichment of certain pathways/cellular components, but these were typically too broad to give any clear insight into the potential roles of these miRNAs/small RNAs in the cell (i.e. sensory perception of sound).

5.3 Transduction into mouse cell lines

Using a viral system as means of introducing a shRNA or an expression plasmid for knockdown or overexpression of a transcript of interest, is an effective and well documented method for

studying the effect of the transcript *in vitro*. These methods, classically applied to transcripts with coding potentially, have more recently been used to study noncoding RNAs (Gupta et al., 2010; Jiao et al., 2014; Kogo et al., 2011). In this research, we attempted to use a Lentiviral system to assess efficiency of infection in mouse cell lines including the Eμ-myc line used in my work to identify IGS ncRNAs. This assessment is a precursor for future knockdown and over-expression approaches to probe the roles of the mouse rDNA IGS transcripts we have identified in this study. Although we achieved successful packaging and infection in the highly transfection/infection efficient HEK cells, we saw no obvious signs of infection into both mouse cell lines (Eμ-myc and MEF). This means that the Lentiviral system which we tested cannot be used for knockdown or overexpression in MEF and Eμ-myc cell lines. With regards to the Eμ-myc infection results, the absence of infection may not be surprising as Eμ-myc infection efficiency using a viral system has been reported to be as low as 1% (personal communication, unpublished). Measuring fluorescence via flow cytometry may be helpful in better quantifying if any cells have been infected. Altogether, as infection efficiency was also extremely low in MEFs, which have previously been reported to have generally a high level of infection using a different Lentiviral system (Carlotti et al., 2004), we showed that the Lentiviral system used in this research is not likely to be the best for altering expression of IGS ncRNA transcripts in Eμ-myc and MEF cell lines. However, it can be tested in other mouse cell lines that may have similar ncRNA profiles from the rDNA IGS. Also, even if infection efficiency is very low, this Lentiviral system still could be used when coupled with fluorescence flow cytometry sorting to enrich for a small sub-population of infected cells.

5.4 Future trajectories

In order to validate our results, determine size of transcripts arising from the identified IGS exons, and assess roles of any confirmed IGS transcripts, we could perform a number of downstream experiments.

As a biological replicate of the capture-seq performed in this project, we could repeat the capture-seq, to confirm the locations and boundaries of the mouse IGS exons found within this data. Several important optimisation steps would be included. Firstly, any gDNA presence would be identified by conventional PCR and only samples with no gDNA contamination would

be used for capture. Importantly, combining results from multiple capture-seq experiments may allow for more accurate predictions of exons/transcripts between library replicates, as well as potentially identify additional novel transcripts. ncRNAs of interest (identified in this study or future experiments) could also be validated for size using a Northern blot. Northern blots have been used to quantify size of a number of noncoding RNAs, like MALAT, which by northern blot supported the expected size of >8000nt (Ji et al., 2003). They can be used to not only identify size, but also to assess differential expression compared to a housekeeping gene transcript for example. Northern blot could also be applied to validating our small RNAs found in section 4.2 and this has been shown in the literature to be an effective means of doing so. For example, microRNA-21 which has been shown to be an overexpressed anti-apoptotic factor in cancer, was positively identified in several cell lines using Northern blot analysis (Chan, Krichevsky, & Kosik, 2005).

It would be of interest to know if IGS transcripts are differentially expressed. We have performed the first screen. First, we could assess differential expression after efficient inhibition of RNA pol I (either by UBF knockdown or by CX-5461 treatment). If transcripts show differential expression, it would suggest that IGS transcripts are RNA pol I dependent. Further, it would be of interest to compare IGS exon transcription within a primary mouse B -cell line, to Eμ-myc. If an IGS exon was differentially expressed, this may suggest the resulting transcript has an oncogenic or tumour suppressor role.

As mentioned earlier, knockdown and overexpression of ncRNAs of interest could also be utilised to understand their functional role in a cell. Due to the lack of efficacy in the viral-based systems we tested, transfection of siRNAs or full-length oligonucleotides of the transcripts themselves may be a more feasible approach for knockdown/overexpression. Upon knockdown/overexpression, several functional assays such as cell viability assays or cell cycle assays as examples. Additionally, the use of co-precipitation assays can unravel interactions between noncoding RNAs and other molecules (such as RNA or protein), which has been successfully used to determine many interactions for already classified ncRNA (i.e. HOTAIR in (Liu et al., 2014)). Together, this could allow us to determine the functional role of identified ncRNAs. Additionally, quantifying expression of these ncRNAs in different tissues will allow us

to determine if these ncRNAs show any tissue specific expression profiles, which has been shown to be a common phenomenon for lncRNAs (Cabili et al., 2011).

The rDNA unit across mammals, though sharing a similar general structure, have significant variation in sequence of the rDNA IGS as a result of less functional constraint forcing maintenance of sequence (see (Gonzalez & Sylvester, 1995; Grozdanov et al., 2003) for further IGS sequence information). Consequently, the exons identified in this study may encode transcripts that are not present in the human transcriptome. Therefore, another potential future direction to use the same capture-seq approach is to determine a full profile of ncRNAs from the human rDNA IGS. As shown in **Figure 3**, we already have available a range of probes designed to capture the human rDNA IGS sequence. This will allow us to apply the same capture-seq approach using human cells.

To further validate the identified small RNAs there are a number of potential approaches that could applied in the future. Small RNAs, particularly miRNAs and siRNAs, are often no longer than the size of a typical PCR primer, and consequently qPCR validation by standard protocols is not feasible. It has been reported that the addition of a poly(A)- or (U)-tailing step (where a string of adenosine/uracils is added to the 3' end of the transcript) can allow both the cDNA synthesis step and the qPCR step (Benes & Castoldi, 2010; Mei et al., 2012). In qPCR, A- or U-tailing provides a longer template, and degenerate primers can be used to amplify the 3' end while primers complementing the miRNA can be used for the 5' end, allowing specific amplification of miRNA. This would allow to amplify our predicted small RNAs, and if coupled with the Northern blot analysis would allow us to assess whether these predicted small RNAs are present in cells, and to determine their size. Finally, knockdown and overexpression of these small ncRNAs of interest may provide insight into their functional role in a cell.

5.5 Final summary

In this thesis, I set out to identify ncRNA transcripts from the Eμ-myc rDNA IGS using both capture-seq and small RNA-seq approaches. Several potential lncRNA exons were identified in the mouse rDNA IGS using Capture-seq, with the majority exons with the strongest evidence for their existence being located towards the 3' end of the rDNA IGS. These regions may be related to transcription of the pRNA, and together with the observation of small RNAs from this region, suggests that the major role of ncRNA transcription in the mouse rDNA IGS is to control rRNA synthesis through production of the pRNA. Experimental validation techniques such as Northern blotting, knockdown/overexpression assays, and qPCR approaches, could be used to clarify the nature and function of these rDNA IGS transcripts.

Small RNA reads aligning to the mouse rDNA IGS were also identified. We performed GO SLIM enrichment analysis on the predicted targets of these potential small RNAs, based on seed sequences predicted for each small ncRNA if they are acting as miRNAs, but the only enriched pathways/cellular components did not appear to be biologically significant. Validation of these potential small RNAs using similar methods to the lncRNAs could help to clarify whether they have any functional significance.

We did not find any evidence for differential expression of mouse rDNA IGS exons upon CX-5461 treatment. If true, this suggests that transcription from these exons are not dependent on RNA pol I. However, unexpectedly we also found no changes in the levels of 47S rRNA following CX-5461 treatment, therefore calling into question whether this treatment was effective in preventing RNA pol I activity in the timeframes used in this study. Therefore, we cannot make any conclusions regarding the effect of CX-5461 on mouse rDNA IGS transcription, and follow up studies are required to answer this question, as well as to determine the effects of UBF knockdown.

6. Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology, 215*(3), 403-410.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arabi, A., Wu, S., Ridderstråle, K., Bierhoff, H., Shiue, C., Fatyol, K., . . . Grummt, I. (2005a). c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription. *Nature cell biology, 7*(3), 303-310.

Arabi, A., Wu, S., Ridderstråle, K., Bierhoff, H., Shiue, C., Fatyol, K., . . . Grummt, I. (2005b). c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription. *Nature cell biology, 7*(3), 303.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Audas, T. E., Jacob, M. D., & Lee, S. (2012). Immobilization of proteins in the nucleolus by ribosomal intergenic spacer noncoding RNA. *Molecular cell, 45*(2), 147-157.

Banfi, B., Malgrange, B., Knisz, J., Steger, K., Dubois-Dauphin, M., & Krause, K.-H. (2004). NOX3, a superoxide-generating NADPH oxidase of the inner ear. *Journal of Biological Chemistry, 279*(44), 46065-46072.

Bartsch, I., Schoneberg, C., & Grummt, I. (1987). Evolutionary changes of sequences and factors that direct transcription termination of human and mouse ribsomal genes. *Molecular and cellular biology, 7*(7), 2521-2529.

Baselga, J., & Swain, S. M. (2009). Novel anticancer targets: revisiting ERBB2 and discovering ERBB3. *Nature Reviews Cancer, 9*(7), 463.

Bell, S. P., Learned, R. M., Jantzen, H.-M., & Tjian, R. (1988). Functional cooperativity between transcription factors UBF1 and SL1 mediates human ribosomal RNA synthesis. *Science, 241*(4870), 1192-1197.

Benes, V., & Castoldi, M. (2010). Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available. *Methods, 50*(4), 244-249.

Blume, S. W., Meng, Z., Shrestha, K., Snyder, R. C., & Emanuel, P. D. (2003). The 5'-untranslated RNA of the human dhfr minor transcript alters transcription pre-initiation complex assembly at the major (core) promoter. *Journal of cellular biochemistry, 88*(1), 165-180.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Brabletz, S., & Brabletz, T. (2010). The ZEB/miR-200 feedback loop—a motor of cellular plasticity in development and cancer? *EMBO reports, 11*(9), 670-677.

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature, 501*(7467), 338.

Bussemakers, M. J., van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F., Schalken, J. A., . . . Isaacs, W. B. (1999). DD3:: A New Prostate-specific Gene, Highly Overexpressed in Prostate Cancer. *Cancer research, 59*(23), 5975-5979.

Bywater, M. J., Pearson, R. B., McArthur, G. A., & Hannan, R. D. (2013). Dysregulation of the basal RNA polymerase transcription apparatus in cancer. *Nature Reviews Cancer, 13*(5), 299-314.

Bywater, M. J., Poortinga, G., Sanij, E., Hein, N., Peck, A., Cullinane, C., . . . Anderes, K. (2012). Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53. *Cancer cell, 22*(1), 51-65.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development, 25*(18), 1915-1927.

Cao, S., Liu, W., Li, F., Zhao, W., & Qin, C. (2014). Decreased expression of lncRNA GAS5 predicts a poor prognosis in cervical cancer. *International journal of clinical and experimental pathology, 7*(10), 6776.

Carlotti, F., Bazuine, M., Kekarainen, T., Seppen, J., Pognonec, P., Maassen, J. A., & Hoeben, R. C. (2004). Lentiviral vectors efficiently transduce quiescent mature 3T3-L1 adipocytes. *Molecular Therapy, 9*(2), 209-217.

Caruana, G., & McInnes, R. L. (2004). *Assessing genomic DNA contamination of total RNA isolated from kidney cells obtained by Laser Capture Microdissection using the Agilent RNA 6000 Pico assay Application*.

Chan, J. A., Krichevsky, A. M., & Kosik, K. S. (2005). MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer research, 65*(14), 6029-6033.

Chen, M.-J., Shimada, T., Moulton, A. D., Cline, A., Humphries, R. K., Maizel, J., & Nienhuis, A. (1984). The functional human dihydrofolate reductase gene. *Journal of Biological Chemistry, 259*(6), 3933-3943.

Clark, M. B., Johnston, R. L., Inostroza-Ponta, M., Fox, A. H., Fortini, E., Moscato, P., . . . Mattick, J. S. (2012). Genome-wide analysis of long noncoding RNA stability. *Genome research, 22*(5), 885-898.

Clarke, R. A., Zhao, Z., Guo, A.-Y., Roper, K., Teng, L., Fang, Z.-M., . . . Gardiner, R. A. (2009). New genomic structure for prostate cancer specific gene PCA3 within BMCC1: implications for prostate cancer detection and progression. *PloS one, 4*(3), e4995.

Conacci-Sorrell, M., McFerrin, L., & Eisenman, R. N. (2014). An overview of MYC and its interactome. *Cold Spring Harbor perspectives in medicine, 4*(1), a014357.

Copenhaver, G. P., Putnam, C. D., Denton, M. L., & Pikaard, C. S. (1994). The RNA polymerase I transcription factor UBF is a sequence-tolerant HMG-box protein that can recognize structured nucleic acids. *Nucleic Acids Research, 22*(13), 2651-2657.

Costa, F. F. (2008). Non-coding RNAs, epigenetics and complexity. *Gene, 410*(1), 9-17.

Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., . . . Verslype, C. (2004). Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *New England Journal of Medicine, 351*(4), 337-345.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C., & Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proceedings of the National Academy of Sciences, 79*(24), 7824-7827.

Dang, C. V. (2012). MYC on the path to cancer. *Cell, 149*(1), 22-35.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., . . . Knowles, D. G. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research, 22*(9), 1775-1789.

Dhahbi, J. M., Spindler, S. R., Atamna, H., Yamakawa, A., Guerrero, N., Boffelli, D., . . . Martin, D. I. (2013). Deep sequencing identifies circulating mouse miRNAs that are functionally implicated in manifestations of aging and responsive to calorie restriction. *Aging (Albany NY), 5*(2), 130.

Diermeier, S. D., Németh, A., Rehli, M., Grummt, I., & Längst, G. (2013). Chromatin-specific regulation of mammalian rDNA transcription by clustered TTF-I binding sites. *PLoS Genetics, 9*(9), e1003786.

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol, 4*(11), e1000176.

Djuranovic, S., Nahvi, A., & Green, R. (2011). A parsimonious model for gene regulation by miRNAs. *Science, 331*(6017), 550-553.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21.

Downward, J. (2003). Targeting RAS signalling pathways in cancer therapy. *Nature Reviews Cancer, 3*(1), 11.

Drygin, D., Lin, A., Bliesath, J., Ho, C. B., O'Brien, S. E., Proffitt, C., . . . Siddiqui-Jain, A. (2011). Targeting RNA polymerase I with an oral small molecule CX-5461 inhibits ribosomal RNA synthesis and solid tumor growth. *Cancer research, 71*(4), 1418-1430.

Duesberg, P., & Li, R. (2003). Multistep carcinogenesis: a chain reaction of aneuploidizations. *Cell Cycle, 2*(3), 201-209.

Dykxhoorn, D. M., Wu, Y., Xie, H., Yu, F., Lal, A., Petrocca, F., . . . Lieberman, J. (2009). miR-200 enhances mouse breast cancer cell colonization to form distant metastases. *PloS one, 4*(9), e7181.

Eickbush, T. H., & Eickbush, D. G. (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics, 175*(2), 477-485.

Eriksson, D., & Stigbrand, T. (2010). Radiation-induced cell death mechanisms. *Tumor Biology, 31*(4), 363-372.

Ferreira, L. B., Palumbo, A., de Mello, K. D., Sternberg, C., Caetano, M. S., de Oliveira, F. L., . . . Gimba, E. R. P. (2012). PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling. *BMC cancer, 12*(1), 507.

Fidler, I. J. (1978). Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer research, 38*(9), 2651-2660.

Fidler, I. J. (2003). The pathogenesis of cancer metastasis: the'seed and soil'hypothesis revisited. *Nature Reviews Cancer, 3*(6), 453-458.

Fish, R. J., & Kruithof, E. K. (2004). Short-term cytotoxic effects and long-term instability of RNAi delivered using lentiviral vectors. *BMC molecular biology, 5*(1), 9.

Freed, E. F., Bleichert, F., Dutca, L. M., & Baserga, S. J. (2010). When ribosomes go bad: diseases of ribosome biogenesis. *Molecular BioSystems, 6*(3), 481-493.

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., & Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology, 26*(4), 407-415.

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2011). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research, 40*(1), 37-52.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome research, 19*(1), 92-105.

Ganley, A. R., & Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome research, 17*(2), 184-191.

Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research, 45*(D1), D331-D338.

Gerber, J.-K., Gögel, E., Berger, C., Wallisch, M., Müller, F., Grummt, I., & Grummt, F. (1997). Termination of mammalian rDNA replication: polar arrest of replication fork movement by transcription termination factor TTF-I. *Cell, 90*(3), 559-567.

Ghoshal, K., Majumder, S., Datta, J., Motiwala, T., Bai, S., Sharma, S. M., . . . Jacob, S. T. (2004). Role of human ribosomal RNA (rRNA) promoter methylation and of methyl-CpG-binding protein MBD2 in the suppression of rRNA gene expression. *Journal of Biological Chemistry, 279*(8), 6783-6793.

Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S., & Lemos, B. (2015). Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proceedings of the National Academy of Sciences, 112*(8), 2485-2490.

Gibbons, J. G., Branco, A. T., Yu, S., & Lemos, B. (2014). Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nature communications, 5*, 4850.

Gonzalez, I. L., & Sylvester, J. E. (1995). Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics, 27*(2), 320-328.

Gorski, J. J., Pathak, S., Panov, K., Kasciukovic, T., Panova, T., Russell, J., & Zomerdijk, J. C. (2007). A novel TBP-associated factor of SL1 functions in RNA polymerase I transcription. *The EMBO journal, 26*(6), 1560-1568.

Grandori, C., Gomez-Roman, N., Felton-Edkins, Z. A., Ngouenet, C., Galloway, D. A., Eisenman, R. N., & White, R. J. (2005). c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature cell biology, 7*(3), 311-318.

Griffiths-Jones, S. (2006). miRBase: the microRNA sequence database. *MicroRNA Protocols*, 129-138.

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research, 34*(suppl_1), D140-D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2007). miRBase: tools for microRNA genomics. *Nucleic Acids Research, 36*(suppl_1), D154-D158.

Griffiths-Jones, S. (2004). The microRNA registry. *Nucleic Acids Research, 32*(suppl_1), D109-D111.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell, 27*(1), 91-105.

Grozdanov, P., Georgiev, O., & Karagyozov, L. (2003). Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer☆. *Genomics, 82*(6), 637-643.

Grummt, I. (2003). Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes & development, 17*(14), 1691-1702.

Grummt, I., Maier, U., Öhrlein, A., Hassouna, N., & Bachellerie, J.-P. (1985). Transcription of mouse rDNA terminates downstream of the 3' end of 28S RNA and involves interaction of factors with repeated sequences in the 3' spacer. *Cell, 43*(3), 801-810.

Guetg, C., Scheifele, F., Rosenthal, F., Hottiger, M. O., & Santoro, R. (2012). Inheritance of silent rDNA chromatin is mediated by PARP1 via noncoding RNA. *Molecular cell, 45*(6), 790-800.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., . . . Rinn, J. L. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature, 464*(7291), 1071-1076.

Haber, D. A., Gray, N. S., & Baselga, J. (2011). The evolving war on cancer. *Cell, 145*(1), 19-24.

Harris, A. W., Pinkert, C. A., Crawford, M., Langdon, W., Brinster, R., & Adams, J. (1988). The E mu-myc transgenic mouse. A model for high-incidence spontaneous lymphoma and leukemia of early B cells. *Journal of Experimental Medicine, 167*(2), 353-371.

He, M., Liu, Y., Wang, X., Zhang, M. Q., Hannon, G. J., & Huang, Z. J. (2012). Cell-type-based analysis of microRNA profiles in the mouse brain. *Neuron, 73*(1), 35-48.

Hein, N., Hannan, K. M., George, A. J., Sanij, E., & Hannan, R. D. (2013). The nucleolus: an emerging target for cancer therapy. *Trends in molecular medicine, 19*(11), 643-654.

Herdman, C., Mars, J.-C., Stefanovsky, V. Y., Tremblay, M. G., Sabourin-Felix, M., Lindsay, H., . . . Moss, T. (2017). A unique enhancer boundary complex on the mouse ribosomal RNA genes persists after loss of Rrn3 or UBF and the inactivation of RNA polymerase I transcription. *PLoS Genetics, 13*(7), e1006899.

Hessels, D., Gunnewiek, J. M. K., van Oort, I., Karthaus, H. F., van Leenders, G. J., van Balken, B., . . . Schalken, J. A. (2003). DD3 PCA3-based molecular urine analysis for the diagnosis of prostate cancer. *European urology, 44*(1), 8-16.

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., . . . Severin, J. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature, 543*(7644), 199-204.

Hori, H., & Osawa, S. (1987). Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. *Molecular biology and evolution, 4*(5), 445-472.

Houseley, J., Kotovic, K., El Hage, A., & Tollervey, D. (2007). Trf4 targets ncRNAs from telomeric and rDNA spacer regions and functions in rDNA copy number control. *The EMBO journal, 26*(24), 4996-5006.

Housman, G., & Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1859*(1), 31-40.

Hruz, T., Wyss, M., Docquier, M., Pfaffl, M. W., Masanetz, S., Borghi, L., . . . Laule, O. (2011). RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC genomics, 12*(1), 1.

Hynes, N. E., & Lane, H. A. (2005). ERBB receptors and cancer: the complexity of targeted inhibitors. *Nature Reviews Cancer, 5*(5), 341-354.

Iorio, M. V., & Croce, C. M. (2012). MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO molecular medicine, 4*(3), 143-159.

Ito, T., Tsuchiya, K., Osawa, S., Shibata, H., & Kanda, N. (2008). Mapping of rRNA gene loci in the mice, Mus musculus molossinus (Japan) and Mus musculus musculus (Russia) by double color FISH. *Journal of Veterinary Medical Science, 70*(9), 997-1000.

Jackson, A. L., & Loeb, L. A. (1998). The mutation rate and cancer. *Genetics, 148*(4), 1483-1490.

Jacob, M. D., Audas, T. E., Mullineux, S.-T., & Lee, S. (2012). Where no RNA polymerase has gone before: novel functional transcripts derived from the ribosomal intergenic spacer. *Nucleus, 3*(4), 315-319.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P. M., . . . Bulk, E. (2003). MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene, 22*(39), 8031.

Jiao, F., Hu, H., Yuan, C., Wang, L., Jiang, W., Jin, Z., . . . Wang, L. (2014). Elevated expression level of long noncoding RNA MALAT-1 facilitates cell growth, migration and invasion in pancreatic cancer. *Oncology reports, 32*(6), 2485-2492.

Jonker, D. J., O'callaghan, C. J., Karapetis, C. S., Zalcberg, J. R., Tu, D., Au, H.-J., . . . Simes, R. J. (2007). Cetuximab for the treatment of colorectal cancer. *New England Journal of Medicine, 357*(20), 2040-2048.

Jordan, P., Mannervik, M., Tora, L., & Carmo-Fonseca, M. (1996). In vivo evidence that TATA-binding protein/SL1 colocalizes with UBF and RNA polymerase I when rRNA synthesis is either active or inactive. *Journal of Cell Biology, 133*(2), 225-234.

Kaneko, S., Li, G., Son, J., Xu, C.-F., Margueron, R., Neubert, T. A., & Reinberg, D. (2010). Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes & development, 24*(23), 2615-2620.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., . . . Hofacker, I. L. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science, 316*(5830), 1484-1488.

Kaufmann, S. H., & Earnshaw, W. C. (2000). Induction of apoptosis by cancer chemotherapy. *Experimental cell research, 256*(1), 42-49.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649.

Kelly, P. N., & Strasser, A. (2011). The role of Bcl-2 and its pro-survival relatives in tumourigenesis and cancer therapy. *Cell death and differentiation, 18*(9), 1414.

Kelly, R. D., Mahmud, A., McKenzie, M., Trounce, I. A., & St John, J. C. (2012). Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A. *Nucleic Acids Research, 40*(20), 10124-10138.

Khurana, R., Ranches, G., Schafferer, S., Lukasser, M., Rudnicki, M., Mayer, G., & Hüttenhofer, A. (2017). Identification of urinary exosomal noncoding RNAs as novel biomarkers in chronic kidney disease. *RNA, 23*(2), 142-152.

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., & Chrousos, G. P. (2010). Noncoding RNA gas5 is a growth arrest–and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal., 3*(107), ra8-ra8.

Kobayashi, T. (2011). Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cellular and molecular life sciences, 68*(8), 1395-1403.

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., . . . Komune, S. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer research, 71*(20), 6320-6326.

Kozomara, A., & Griffiths-Jones, S. (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research, 42*(D1), D68-D73.

Kraus, W. L. (2008). Transcriptional control by PARP-1: chromatin modulation, enhancer-binding, coregulation, and insulation. *Current opinion in cell biology, 20*(3), 294-302.

Kuhn, A., & Grummt, I. (1987). A novel promoter in the mouse rDNA spacer is active in vivo and in vitro. *The EMBO journal, 6*(11), 3487.

Kuhn, A., Voit, R., Stefanovsky, V., Evers, R., Bianchi, M., & Grummt, I. (1994). Functional differences between the two splice variants of the nucleolar transcription factor UBF: the second HMG box determines specificity of DNA binding and transcriptional activity. *The EMBO journal, 13*(2), 416.

Kuhn, C.-D., Geiger, S. R., Baumli, S., Gartmann, M., Gerber, J., Jennebach, S., . . . Cramer, P. (2007). Functional architecture of RNA polymerase I. *Cell, 131*(7), 1260-1272.

Kusnadi, E. P., Hannan, K. M., Hicks, R. J., Hannan, R. D., Pearson, R. B., & Kang, J. (2015). Regulation of rDNA transcription in response to growth factors, nutrients and energy. *Gene, 556*(1), 27-34.

Lee, J. T. (2000). Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. *Cell, 103*(1), 17-27.

Lee, M. P., DeBaun, M. R., Mitsuya, K., Galonek, H. L., Brandenburg, S., Oshimura, M., & Feinberg, A. P. (1999). Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith–Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proceedings of the National Academy of Sciences, 96*(9), 5203-5208.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Li, J., Zhang, D., Ward, K. M., Prendergast, G. C., & Ayene, I. S. (2012). Hydroxyethyl disulfide as an efficient metabolic assay for cell viability in vitro. *Toxicology in Vitro, 26*(4), 603-612.

Lin, Y.-H., & Keil, R. L. (1991). Mutations affecting RNA polymerase I-stimulated exchange and rDNA recombination in yeast. *Genetics, 127*(1), 31-38.
Liu, X.-h., Sun, M., Nie, F.-q., Ge, Y.-b., Zhang, E.-b., Yin, D.-d., . . . Li, J.-h. (2014). Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Molecular cancer, 13*(1), 92.
Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *Methods, 25*(4), 402-408.
Long, E. O., & Dawid, I. B. (1980). Repeated genes in eukaryotes. *Annual review of biochemistry, 49*(1), 727-764.
Lopez-Sánchez, L. M., Jimenez, C., Valverde, A., Hernandez, V., Peñarando, J., Martinez, A., . . . Aranda, E. (2014). CoCl 2, a mimic of hypoxia, induces formation of polyploid giant cells with stem characteristics in colon cancer. *PloS one, 9*(6), e99143.
Lu, H., Samanta, D., Xiang, L., Zhang, H., Hu, H., Chen, I., . . . Semenza, G. L. (2015). Chemotherapy triggers HIF-1–dependent glutathione synthesis and copper chelation that induces the breast cancer stem cell phenotype. *Proceedings of the National Academy of Sciences, 112*(33), E4600-E4609.
Martianov, I., Ramadass, A., Barros, A. S., Chow, N., & Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature, 445*(7128), 666-670.
Mattick, J. S. (2005). The functional genomics of noncoding RNA. *Science, 309*(5740), 1527-1528.
Mayer, C., & Grummt, I. (2006). Ribosome biogenesis and cell growth: mTOR coordinates transcription by all three classes of nuclear RNA polymerases. *Oncogene, 25*(48), 6384-6391.
Mayer, C., Schmitz, K.-M., Li, J., Grummt, I., & Santoro, R. (2006). Intergenic transcripts regulate the epigenetic state of rRNA genes. *Molecular cell, 22*(3), 351-361.
McStay, B., & Grummt, I. (2008). The epigenetics of rRNA genes: from molecular to chromosome biology. *Annual review of cell and developmental biology, 24*, 131-157.
Mei, Q., Li, X., Meng, Y., Wu, Z., Guo, M., Zhao, Y., . . . Han, W. (2012). A facile and specific assay for quantifying microRNA by an optimized RT-qPCR approach. *PloS one, 7*(10), e46890.
Mekhail, K., Gunaratnam, L., Bonicalzi, M.-E., & Lee, S. (2004). HIF activation by pH-dependent nucleolar sequestration of VHL. *Nature cell biology, 6*(7), 642.
Mekhail, K., Khacho, M., Carrigan, A., Hache, R. R., Gunaratnam, L., & Lee, S. (2005). Regulation of ubiquitin ligase dynamics by the nucleolus. *J Cell Biol, 170*(5), 733-744.
Meltzer, P. S. (2005). Cancer genomics: small RNAs with big impacts. *Nature, 435*(7043), 745.
Mercer, T. R., Dinger, M. E., & Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews Genetics, 10*(3), 155-159.
Miller, G., Panov, K. I., Friedrich, J. K., Trinkle-Mulcahy, L., Lamond, A. I., & Zomerdijk, J. C. (2001). hRRN3 is essential in the SL1-mediated recruitment of RNA polymerase I to rRNA gene promoters. *The EMBO journal, 20*(6), 1373-1382.
Montanaro, L., Treré, D., & Derenzini, M. (2008). Nucleolus, ribosomes, and cancer. *The American journal of pathology, 173*(2), 301-310.
Moorefield, B., Greene, E. A., & Reeder, R. H. (2000). RNA polymerase I transcription factor Rrn3 is functionally conserved between yeast and human. *Proceedings of the National Academy of Sciences, 97*(9), 4724-4729.
Mosser, D. D., Caron, A. W., Bourget, L., Denis-Larose, C., & Massie, B. (1997). Role of the human heat shock protein hsp70 in protection against stress-induced apoptosis. *Molecular and cellular biology, 17*(9), 5317-5327.

Mourtada-Maarabouni, M., Pickard, M., Hedge, V., Farzaneh, F., & Williams, G. (2009). GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene, 28*(2), 195.

Muth, V., Nadaud, S., Grummt, I., & Voit, R. (2001). Acetylation of TAF I 68, a subunit of TIF-IB/SL1, activates RNA polymerase I transcription. *The EMBO journal, 20*(6), 1353-1362.

Navarro, P., Page, D. R., Avner, P., & Rougeulle, C. (2006). Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program. *Genes & development, 20*(20), 2787-2792.

Negi, S. S., & Brown, P. (2015). rRNA synthesis inhibitor, CX-5461, activates ATM/ATR pathway in acute lymphoblastic leukemia, arrests cells in G2 phase and induces apoptosis. *Oncotarget, 6*(20), 18094.

Ohhata, T., Matsumoto, M., Leeb, M., Shibata, S., Sakai, S., Kitagawa, K., . . . Wutz, A. (2015). Histone H3 Lysine 36 Trimethylation Is Established over the Xist Promoter by Antisense Tsix Transcription and Contributes to Repressing Xist Expression. *Molecular and cellular biology, 35*(22), 3909-3920.

Olayioye, M. A., Neve, R. M., Lane, H. A., & Hynes, N. E. (2000). The ErbB signaling network: receptor heterodimerization in development and cancer. *The EMBO journal, 19*(13), 3159-3167.

Ouyang, L., Shi, Z., Zhao, S., Wang, F. T., Zhou, T. T., Liu, B., & Bao, J. K. (2012). Programmed cell death pathways in cancer: a review of apoptosis, autophagy and programmed necrosis. *Cell proliferation, 45*(6), 487-498.

Park, S.-M., Gaur, A. B., Lengyel, E., & Peter, M. E. (2008). The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes & development, 22*(7), 894-907.

Pelengaris, S., Khan, M., & Gerard, E. (2002). c-MYC: more than just a matter of life and death. *Nature reviews. Cancer, 2*(10), 764.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology, 33*(3), 290-295.

Pestova, T. V., Kolupaeva, V. G., Lomakin, I. B., Pilipenko, E. V., Shatsky, I. N., Agol, V. I., & Hellen, C. U. (2001). Molecular mechanisms of translation initiation in eukaryotes. *Proceedings of the National Academy of Sciences, 98*(13), 7029-7036.

Peyroche, G., Milkereit, P., Bischler, N., Tschochner, H., Schultz, P., Sentenac, A., . . . Riva, M. (2000). The recruitment of RNA polymerase I on rDNA is mediated by the interaction of the A43 subunit with Rrn3. *The EMBO journal, 19*(20), 5473-5482.

Plath, K., Fang, J., Mlynarczyk-Evans, S. K., Cao, R., Worringer, K. A., Wang, H., . . . Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. *Science, 300*(5616), 131-135.

Prensner, J. R., & Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer discovery, 1*(5), 391-407.

Quin, J. E., Devlin, J. R., Cameron, D., Hannan, K. M., Pearson, R. B., & Hannan, R. D. (2014). Targeting the nucleolus for cancer intervention. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842*(6), 802-816.

Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, 11.12. 11-11.12. 34.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology, 29*(1), 24.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., . . . Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome biology, 14*(5), R51.

Rothstein, R., Michel, B., & Gangloff, S. (2000). Replication fork pausing and recombination or "gimme a break". *Genes & development, 14*(1), 1-10.
Russell, J., & Zomerdijk, J. C. (2005). RNA-polymerase-I-directed rDNA transcription, life and works. *Trends in biochemical sciences, 30*(2), 87-96.
Salameh, A., Lee, A. K., Cardó-Vila, M., Nunes, D. N., Efstathiou, E., Staquicini, F. I., . . . Hosoya, H. (2015). PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proceedings of the National Academy of Sciences, 112*(27), 8403-8408.
Sanij, E., & Hannan, R. D. (2009). The role of UBF in regulating the structure and dynamics of transcriptionally active rDNA chromatin. *Epigenetics, 4*(6), 374-382.
Sanij, E., Poortinga, G., Sharkey, K., Hung, S., Holloway, T. P., Quin, J., . . . Stefanovsky, V. (2008). UBF levels determine the number of active ribosomal RNA genes in mammals. *J Cell Biol, 183*(7), 1259-1274.
Santoro, R., Li, J., & Grummt, I. (2002). The nucleolar remodeling complex NoRC mediates heterochromatin formation and silencing of ribosomal gene transcription. *Nature genetics, 32*(3), 393.
Sasako, M., Sakuramoto, S., Katai, H., Kinoshita, T., Furukawa, H., Yamaguchi, T., . . . Ohashi, Y. (2011). Five-year outcomes of a randomized phase III trial comparing adjuvant chemotherapy with S-1 versus surgery alone in stage II or III gastric cancer. *Journal of Clinical Oncology, 29*(33), 4387-4393.
Sawyers, C. (2004). Targeted cancer therapy. *Nature, 432*(7015), 294.
Schwarzacher, H. G., & Wachtler, F. (1993). The nucleolus. *Anatomy and embryology, 188*(6), 515-536.
Sharma, S., Zeng, J.-Y., Zhuang, C.-M., Zhou, Y.-Q., Yao, H.-P., Hu, X., . . . Wang, M.-H. (2013). Small-molecule inhibitor BMS-777607 induces breast cancer cell polyploidy with increased resistance to cytotoxic chemotherapy agents. *Molecular cancer therapeutics, 12*(5), 725-736.
Simonis, M., Kooren, J., & De Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. *Nature methods, 4*(11), 895.
Smit, A. H. R. G., P. . (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.
Spandidos, A., Wang, X., Wang, H., & Seed, B. (2010). PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. *Nucleic Acids Research, 38*(suppl 1), D792-D799.
Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature, 458*(7239), 719.
Strezoska, Ž., Pestov, D. G., & Lau, L. F. (2000). Bop1 is a mouse WD40 repeat nucleolar protein involved in 28S and 5.8 S rRNA processing and 60S ribosome biogenesis. *Molecular and cellular biology, 20*(15), 5516-5528.
Sun, M., Jin, F.-y., Xia, R., Kong, R., Li, J.-h., Xu, T.-p., . . . De, W. (2014). Decreased expression of long noncoding RNA GAS5 indicates a poor prognosis and promotes cell proliferation in gastric cancer. *BMC cancer, 14*(1), 319.
Swift, S., Lorens, J., Achacoso, P., & Nolan, G. P. (2001). Rapid production of retroviruses for efficient gene delivery to mammalian cells using 293T cell–based systems. *Current protocols in immunology*, 10.17. 14-10.17. 29.
Tallman, M. S., Gilliland, D. G., & Rowe, J. M. (2005). Drug therapy for acute myeloid leukemia. *Blood, 106*(4), 1154-1163.
Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., . . . Leder, P. (1982). Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. *Proceedings of the National Academy of Sciences, 79*(24), 7837-7841.

Tchurikov, N. A., Fedoseeva, D. M., Sosin, D. V., Snezhkina, A. V., Melnikova, N. V., Kudryavtseva, A. V., . . . Kretova, O. V. (2014). Hot spots of DNA double-strand breaks and genomic contacts of human rDNA units are involved in epigenetic regulation. *Journal of molecular cell biology, 7*(4), 366-382.

Tchurikov, N. A., Kretova, O. V., Fedoseeva, D. M., Sosin, D. V., Grachev, S. A., Serebraykova, M. V., . . . Kravatsky, Y. V. (2013). DNA double-strand breaks coupled with PARP1 and HNRNPA2B1 binding sites flank coordinately expressed domains in human chromosomes. *PLoS Genetics, 9*(4), e1003429.

Tebbutt, N., Pedersen, M. W., & Johns, T. G. (2013). Targeting the ERBB family in cancer: couples therapy. *Nature Reviews Cancer, 13*(9), 663-673.

Thakur, N., Tiwari, V. K., Thomassin, H., Pandey, R. R., Kanduri, M., Göndör, A., . . . Kanduri, C. (2004). An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region. *Molecular and cellular biology, 24*(18), 7855-7862.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics, 14*(2), 178-192.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians, 65*(2), 87-108.

Tsai, M.-C., Spitale, R. C., & Chang, H. Y. (2011). Long intergenic noncoding RNAs: new links in cancer progression. *Cancer research, 71*(1), 3-7.

Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., & Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development, 20*(5), 515-524.

Vogel, C. L., Cobleigh, M. A., Tripathy, D., Gutheil, J. C., Harris, L. N., Fehrenbacher, L., . . . Burchmore, M. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *Journal of Clinical Oncology, 20*(3), 719-726.

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine, 10*(8), 789.

Walker, S. E., & Lorsch, J. (2013). RNA purification--precipitation methods. *Methods in enzymology, 530*, 337.

Wang, K. C., & Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular cell, 43*(6), 904-914.

Weinmann, R., & Roeder, R. G. (1974). Role of DNA-dependent RNA polymerase III in the transcription of the tRNA and 5S RNA genes. *Proceedings of the National Academy of Sciences, 71*(5), 1790-1794.

Wong, S.-F. (2005). Cetuximab: an epidermal growth factor receptor monoclonal antibody for the treatment of colorectal cancer. *Clinical therapeutics, 27*(6), 684-694.

Wu, L., & Belasco, J. G. (2008). Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Molecular cell, 29*(1), 1-7.

Wu, W., Pew, T., Zou, M., Pang, D., & Conzen, S. D. (2005). Glucocorticoid receptor-induced MAPK phosphatase-1 (MPK-1) expression inhibits paclitaxel-associated MAPK activation and contributes to breast cancer cell survival. *Journal of Biological Chemistry, 280*(6), 4117-4124.

Xu, B., Li, H., Perry, J. M., Singh, V. P., Unruh, J., Yu, Z., . . . Gerton, J. L. (2017). Ribosomal DNA copy number loss and sequence variation in cancer. *PLoS Genetics, 13*(6), e1006771.

Yan, S., Frank, D., Son, J., Hannan, K. M., Hannan, R. D., Chan, K. T., . . . Sanij, E. (2017). The Potential of Targeting Ribosome Biogenesis in High-Grade Serous Ovarian Cancer. *International journal of molecular sciences, 18*(1), 210.

Yin, D., He, X., Zhang, E., Kong, R., De, W., & Zhang, Z. (2014). Long noncoding RNA GAS5 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer. *Medical oncology, 31*(11), 253.

Yonesaka, K., Zejnullahu, K., Okamoto, I., Satoh, T., Cappuzzo, F., Souglakos, J., . . . Takeda, M. (2011). Activation of ERBB2 signaling causes resistance to the EGFR-directed therapeutic antibody cetuximab. *Science translational medicine, 3*(99), 99ra86-99ra86.

Yu, F., Shen, X., Fan, L., & Yu, Z. (2015). Analysis of histone modifications at human ribosomal DNA in liver cancer cell. *Scientific reports, 5*.

Zentner, G. E., Balow, S. A., & Scacheri, P. C. (2014). Genomic characterization of the mouse ribosomal DNA locus. *G3: Genes, Genomes, Genetics, 4*(2), 243-254.

Zhang, H., Zeitz, M. J., Wang, H., Niu, B., Ge, S., Li, W., . . . Higgins, M. J. (2014). Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the Kcnq1 locus. *J Cell Biol, 204*(1), 61-75.

Zhou, Y., Schmitz, K.-M., Mayer, C., Yuan, X., Akhtar, A., & Grummt, I. (2009). Reversible acetylation of the chromatin remodelling complex NoRC is required for non-coding RNA-dependent silencing. *Nature cell biology, 11*(8), 1010.

7. Appendices

Appendix 1 :Primer sequences for qPCR and conventional PCR

Primer set name	Forward Sequence	Reverse Sequence	Approximate amplicon length (BP)	Target (if applicable)	Source
Mouse UBF1 isoform specific	ATGAGCCAACCTGGACCTGAG	GGGAGTCCTTCACCTCCTTC	92	-	Designed by our group
Mouse UBF 1 and 2 isoforms (long)	CAAAACTCCCCAGCAACTGT	ATGTTCAGCTCAGGGTGCTT	UBF1: 524, UBF2: 415	Long product	Designed by our group
Mouse UBF 1 and 2 isoforms	UBF1 isoform specific forward primer	UBF 1 and 2 isoform Reverse Primer	UBF1: 413, UBF2: 304	Short product	Designed by our group
Mouse GAPDH	GGCTGTGAACCACGAGAAAT	GTCATGAGCCTTCCACAAT	-	-	(Banfi et al., 2004)
Mouse B-ACTIN	GGCTGTATTCCCCTCCATCG	CCAGTTGGTAACAATGCCATG	-	-	(Spandidos, Wang, Wang, & Seed, 2010)
Mouse GAK	CTGCCCACCAGGCATTTG	CCATGTCACATACATATTCAATGT ACC T	-	-	(Hruz et al., 2011)
Capture-Seq Exon set 1	AAATCGCAGAGGTCGACCAG	ACCCAACACAACCCAACACA-	194	Mouse IGS 14072-14265bp, exon cluster 1	Designed by our group
Capture-Seq Exon set 2	CCCTCCCTCCTTCCACCATA	CCCTACACAAGGAGAT	189	Mouse IGS 16008-16196bp, exon cluster 1	Designed by our group

Capture-Seq Exon set 3	GGGCGCGGTTTTCTTTCATT	GGAAAGTGACAGGCCACAGA	160	Mouse IGS 45094-45254, exon cluster 2	Designed by our group
Capture-Seq Exon set 4	CTGTCTGTGGATGGTCGAGG	CTGCTAACTGAACTCCCGCA	165	Mouse IGS 25280-25444bp, exon cluster 3	Designed by our group
Capture-Seq Exon set 5	ATTACGGGTGGTTGTGAGCC	AAAAAGAAGGGGGAGGGCTG	117	Mouse IGS 40168-40284bp, exon cluster 4	Designed by our group
Capture-Seq Exon set 6	CCCCAAGCGGTAGAGTGTTT	ACAGTTAGGGGAAGGGAGCA	122	Mouse IGS 42687-42809bp, exon cluster 5	Designed by our group
Capture-Seq Exon set 7	GAAATCCTCCCATCCCTGCC	ATTAGGGTTCAAGGCCAGCC	193	Mouse IGS 39252-39444bp, exon cluster 6	Designed by our group
Capture-Seq Exon set 8	CCTCGTCACCATTCTGCACT	AACAAGATGGAGGTGGCTGG	110	Mouse IGS 31481-31591bp, exon cluster 7	Designed by our group
Capture-Seq exon set 9	GCTGTACTTCTGAGGCCGAG	GTGACAGCGACAGACAAGGT	129	Mouse IGS 44395-44523, exon cluster 8	Designed by our group
Capture-Seq exon set 10	GCTGACTGGCTAGTTTTCTGC	TTCAGTCAGTTGCCAGAGCC	113	Mouse IGS, 34512-34624, Amplifies gap in exons	Designed by our group
Capture-Seq	TGTGTTGGGTTGTGTTGGGT	GTTCGACCCGCAAAACACAA	154	Mouse IGS 14246-14399bp,	Designed by our group

exon set 11				exon cluster 1	
Capture-Seq exon set 12	TTTCTGGTCCACATCGCTCC	TGAGTGAGTGGAGCCTTCCT	111	Mouse IGS 16234-16353bp, exon cluster 1	Designed by our group
Capture-Seq exon set 13	TGGCTGTCCTGGAACTCAG	ACTAGGGAGGCAGAGATGGG	76	Mouse IGS 22074-22149bp, exon cluster 3	Designed by our group
Capture-Seq exon set 14	GGATGGTCGAGGCTGCTTTA	CTGCTAACTGAACTCCCGCA	157	Mouse IGS 25288-25444bp, exon cluster 3	Designed by our group
Capture-Seq exon set 15	GAGGCTGTCTGTGGATGGTC	AACAAGATGGAGGTGGCTGG	148	Mouse IGS 31444-31591bp, exon cluster 7	Designed by our group
Capture-Seq exon set 16	CCACCACTCCCCGGTATTTT	ATTAGGGTTCAAGGCCAGCC	145	Mouse IGS 39300-39444bp, exon cluster 6	Designed by our group
Capture-Seq exon set 17	TCTTTTCTCCCCTCCCCCTT	CAAATCCCAGCAACCACACG	149	Mouse IGS 40061-40209bp, exon clusters 4	Designed by our group
Capture-Seq exon set 18	GGGGCGCTTGTACTTCTGAT	CCTCAGATGTAAGGTGCCCC	153	Mouse IGS 44526-44678bp, exon cluster 8	Designed by our group
Capture-Seq exon set 19	GAGGCCGAGGGAAAGCTATG	GGAAAGTGACAGGCCACAGA	180	Mouse IGS 45075-45254, exon cluster 2	Designed by our group
Human UBF primer	GTCGGCCATGTTCATCTTCT	CTCAGACAGGTCGTTCCACA	-	Human UBF	(Yu, Shen, Fan, & Yu, 2015)

Mouse ETS primer	GGTTCGCGTGGTCCTTGT	CGACTCTGGGAACATGGTCAA		Primers to mouse 47S ETS	Designed by our group

Appendix 2 : Table of S1 library full Stringtie output example showing transcript coverage, exons contributing to the transcript and coverage of exons

Transcript start-end	Transcript coverage	Exons start-end	Exon coverage
1-1757	431.2	1 exon	
4020-20385		4020 7029 10210 15694 19310 20385	298.4 3479.8 31.9
6889-20385	3001.3	6889 15694 19310 20385	3362.47 45.065
20871-22175	47.53	1 exon	
25095-45315	237.8	25095 25487 40753 43491 44409 44708 44972 45315	225.92 284.33 57.2 49.9
28870-45315	36.03	28870 29310 29349 32521 41198 44708 44972 45315	2.61 12.97 55.03 95.6
22603-45315	38.5	22603 22718 28970 32521 41198 43522 44972 45315	0.9 5.7 94.9 12.5
22603-45315	38.5	28870 29310 29349 31253 40351 43491 44409 45315	4.7 16.7 58.1 23.4
28870-45315	39.14	28870 32485 41158 43773	8.9 96.8

		44409 45315	22.4
25095-45315	35.02	25095 25487 38444 43491 44409 45315	16.2 39.1 16.5
22603-36842	(4.6	22603 22718 35742 36110 36149 36842	3.04 47.7 29.1
25095-45315	87.7	25095 25379 31548 33270 41937 43592 43985 45315	100.2 209.02 19.97 12.04
8132-12836	(24488.9)	1 exon	

Appendix 3: Table of IGS exons predicted by Stringtie in day two library data, normalised to DNA and ranked from highest to lowest in regards to coverage. Library name in bold.

S1				S2				S4				S5				S7				S9				S10			
start	end	length	norm DNA	start	end	length	norm DNA	start	end	length	Norm DNA	start	end	length	norm DNA	start	end	length	norm DNA	start	end	length	norm DNA	start	end	length	Norm to DNA
41886	43522	1636	0.043465	12173	18057	5884	0.382877	13593	15922	2329	0.030238	41851	43491	1640	0.058065	43954	44708	754	0.056565	41979	43792	1813	0.047102	13593	18051	4458	0.036068
44409	44708	299	0.039766	12309	15603	3294	0.059057	13856	17276	3420	0.026401	41851	43773	1922	0.030094	43954	45315	1361	0.045479	13305	17945	4640	0.016057	41874	44708	2834	0.027195
44972	45315	343	0.02675	40753	41821	1068	0.02824	43954	44708	754	0.021533	44409	44708	299	0.008588	41839	43491	1652	0.044977	43955	45315	1360	0.014279	41874	43491	1617	0.020901
44972	45315	343	0.02675	41972	43897	1925	0.019217	41987	43491	1504	0.017446	41827	43592	1765	0.006139	40753	41458	705	0.042901	43955	44708	753	0.010653	44972	45315	343	0.017268
44972	45315	343	0.018001	41972	43522	1550	0.016101	40753	41442	689	0.015886	41937	43592	1655	0.004721	44972	45315	343	0.027955	40753	41542	789	0.008673	23884	24783	899	0.015494
44972	45315	343	0.018001	43829	44708	879	0.010769	43954	45315	1361	0.01569	38336	43592	5256	0.004704	41839	43631	1792	0.025003	44972	45315	343	0.006137	41874	43522	1648	0.013402
41886	43491	1605	0.015247	43954	44708	754	0.009023	44972	45315	343	0.009671	40645	43592	2947	0.003772	41839	43553	1714	0.022572	22991	26730	3739	0.003768	43954	45315	1361	0.011886
44409	45315	906	0.01214	43954	45315	1361	0.006428	40645	41442	797	0.008359	44972	45315	343	0.003513	25198	25487	289	0.014767	41937	43465	1528	0.003628	38653	43631	4978	0.011193
25126	25487	361	0.011226	41972	43840	1868	0.005698	41987	43773	1786	0.007077	25180	27058	1878	0.003206	40645	41520	875	0.013732	41194	41542	348	0.003385	40645	41525	880	0.010894
38789	43791	5002	0.009514	23886	25487	1601	0.005407	41198	41442	244	0.006883	25180	25379	199	0.003206	41839	43553	1714	0.012153	22991	25487	2496	0.003253	38336	41525	3189	0.006209
40030	43697	3667	0.008827	44972	45315	343	0.004799	25172	25487	315	0.006796	40030	43592	3562	0.002774	41198	41458	260	0.012038	43985	45315	1330	0.003051	44972	45315	343	0.00544
41937	43592	1655	0.008102	44972	45315	343	0.003922	41987	43553	1566	0.005069	25180	28139	2959	0.002432	41839	44708	2869	0.009475	38218	38585	367	0.002749	32213	32716	503	0.005351
40645	43697	3052	0.007336	31658	32084	426	0.003126	41987	43522	1535	0.004939	31648	33270	1622	0.001941	44409	44708	299	0.008921	18370	19946	1576	0.002089	31631	31864	233	0.004524
41886	43631	1745	0.007172	41972	43553	1581	0.002116	41194	41442	248	0.004784	18342	24783	6441	0.001756	15845	18065	2220	0.006391	14083	17945	3862	0.001794	36149	36595	446	0.004497
38336	43592	5256	0.005862	31658	31864	206	0.002107	41987	43553	1566	0.003463	18342	23883	5541	0.001313	38336	41520	3184	0.005959	31709	32525	816	0.001468	18297	25487	7190	0.00438
41937	43592	1655	0.005804	38877	41821	2944	0.001738	38336	41442	3106	0.00345	43985	45315	1330	0.001061	25198	26730	1532	0.005678	31709	32289	580	0.001313	44409	44708	299	0.004278
40030	43734	3704	0.005537	41972	43694	1722	0.001714	44409	44708	299	0.003215	39700	43592	3892	0.001014	31653	32521	868	0.005442	41198	41542	344	0.001276	38017	43631	5614	0.003828
40753	43631	2878	0.005037	41972	43553	1581	0.001483	40030	41442	1412	0.001694	18342	22449	4107	0.000992	41839	43522	1683	0.0052	36149	36540	391	0.00079	29902	31231	1329	0.003515
41158	43791	2633	0.004864	36149	36501	352	0.00132	25172	25379	207	0.001655	29900	30926	1026	0.000976	41960	43592	1632	0.005086	40030	41327	1297	0.000713	31548	31988	440	0.00337
40645	43592	2947	0.004743	29349	31231	1882	0.001244	25172	25379	207	0.00162	44506	45315	809	0.000914	40030	43592	3562	0.004952	29349	31104	1755	0.000688	40753	43631	2878	0.003158
40030	43697	3667	0.004566	35919	36110	191	0.000961	18355	20379	2024	0.001528	43985	45315	1330	0.000892	40030	41520	1490	0.004536	22991	23883	892	0.000597	40030	41525	1495	0.002706
43985	45315	1330	0.004168	38653	41821	3168	0.000906	29349	32521	3172	0.001491	18342	20379	2037	0.000799	44972	45315	343	0.004499	22991	24783	1792	0.000578	33871	36110	239	0.002408
43985	45315	1330	0.003758	44972	45315	343	0.000879	35949	36110	161	0.001465	33838	34262	424	0.000731	44972	45315	343	0.003944	31709	32521	812	0.000539	39700	41525	1825	0.002282
44409	45315	906	0.003743	18345	20409	2064	0.000472	36149	36537	388	0.001461	32213	32737	524	0.000566	40030	43592	3562	0.003548	39700	41327	1627	0.000431	18297	23883	5586	0.002072
41198	43791	2593	0.00343	28972	29310	338	0.000382	39700	41442	1742	0.001422	44638	45315	677	0.000463	41937	43592	1655	0.002977	35891	36110	219	0.000351	40030	41525	1495	0.001756
25126	27058	1932	0.002959	38017	41821	3804	0.000244	18355	21773	3418	0.001411	34916	35218	302	0.000458	31548	32685	1137	0.002962	29245	29310	65	0.000289	44972	45315	343	0.001717
31002	31211	209	0.002904	44151	45315	1164	0.000223	40030	41442	1412	0.001404	37703	43592	5889	0.000388	25198	25379	181	0.002733	41879	43592	1713	0.000107	18297	22449	4152	0.001267
40645	43709	3064	0.002736	44088	44708	620	0.000155	44972	45315	343	0.001401	43985	45315	1330	0.000382	40030	43592	3562	0.00258	31002	31181	179	6.17E-05	18297	23565	5268	0.0012
44506	45315	809	0.002723	15845	18057	2212	8.15E-05	22599	23883	1284	0.001349	43985	45315	1330	0.000277	16900	24783	7883	0.002025	33134	33274	140	4.45E-05	37653	41525	3872	0.001149
41937	43697	1760	0.002581	44972	45315	343	6.64E-05	41937	45315	3378	0.001336	44409	44708	299	0.000245	16900	22449	5549	0.001974	40955	41327	372	4.2E-05	18297	24783	6486	0.001056
13890	15694	1804	0.002556					25172	26730	1558	0.001295	43985	45315	1330	0.000168	43985	45315	1330	0.001762					41874	43553	1679	0.000898
44506	45315	809	0.00254					31600	32525	925	0.001125	28032	28427	395	0.000151	43985	45315	1330	0.00127					18297	26730	8433	0.000749
19310	20385	1075	0.002356					41937	43592	1655	0.001087	33152	33267	115	0.000143	37301	43592	6291	0.001024					38789	41525	2736	0.000715
43985	45315	1330	0.002234					23884	24783	899	0.001075	33152	33275	123	0.000133	40030	43592	3562	0.000999					29902	30860	958	0.000552
30049	30860	811	0.002104					41937	43592	1655	0.001041	33152	33268	116	0.000133	25198	27058	1860	0.000583					18297	25379	7082	0.000531
39700	43734	4034	0.002101					37703	37932	229	0.001004	33152	33269	117	0.000133	44409	44576	167	0.000523					24455	25611	1156	0.000413
25126	25379	253	0.002031					19488	21935	2447	0.000893	25180	25611	431	0.000126	25198	25379	181	0.000478					18297	25379	7082	0.000371
44638	45315	677	0.001897					32213	32752	539	0.000665	44972	45315	343	0.0001	37301	37932	631	0.000427					31002	31195	193	0.000193
19310	20385	1075	0.00176					44506	45315	809	0.000641	15845	18086	2241	9.62E-05	44506	45315	809	0.000382					23884	23919	35	0.000192
22566	23883	1317	0.001733					44972	45315	343	0.000619	43985	45315	1330	7.98E-05	33838	33994	156	0.000354					24455	25379	924	0.000168
44903	45315	412	0.001637					22599	24783	2184	0.000551	43985	45315	1330	4.56E-05	44972	45315	343	0.000349					18297	23565	5268	0.000125
44903	45315	412	0.001637					29183	29310	127	0.000545					44903	45315	412	0.00033					44409	45315	906	0.000105
37653	37932	279	0.001523					25172	27058	1886	0.000487					43985	45315	1330	0.000328					33152	33341	189	9.73E-05
38233	43705	5472	0.001324					43985	44369	384	0.00041					31002	31144	142	0.000328					23884	23919	35	6.8E-05
44506	45315	809	0.001291					43985	44369	384	0.000346					37653	37932	279	0.000318					18297	23565	5268	4.31E-05
32168	32444	276	0.001259					22599	23565	966	0.000306					44972	45315	343	0.000299					44409	45315	906	4.29E-05
40030	43734	3704	0.001189					13854	17276	3422	0.000302					44638	45315	677	0.000286					35583	35749	166	1.81E-05
25126	27058	1932	0.001183					33850	33978	128	0.00026					29267	31144	1877	0.000233					44409	44708	299	1.33E-05
44348	44369	21	0.001159					40030	41442	1412	0.000245					29267	30860	1593	0.000196					44972	45315	343	5.35E-06
44347	44369	22	0.001094					28032	28291	259	0.000242					37301	37932	631	0.00017								
41937	43701	1764	0.001077					13854	17276	3422	0.000217					45039	45315	276	0.000162								
33850	33995	145	0.001075					44409	45315	906	0.000163					33152	33295	143	0.000144								
19310	20385	1075	0.001041					30845	30926	81	0.000116					32942	33271	329	3.97E-05								
44903	45315	412	0.000937					25172	25611	439	0.000113																
44903	45315	412	0.000937					31002	31230	228	9.83E-05																
22566	24783	2217	0.000868					22045	22449	404	4.59E-05																
25126	25379	253	0.000855					44239	44665	426	3.58E-05																
44506	45315	809	0.000844					33152	33272	120	3.35E-05																
44506	45315	809	0.000799					33152	33268	116	3.22E-05																
25126	26730	1604	0.000722					33152	33270	118	3E-05																
44903	45315	412	0.000644					33152	33269	117	1.56E-05																
44903	45315	412	0.000644					44794	45315	521	1.39E-05																
44506	45315	809	0.000637																								
33838	33995	157	0.000598																								
25126	25379	253	0.00056																								
44349	44369	20	0.00055																								
22566	24783	2217	0.000541																								
31675	33270	1595	0.000514																								
22566	24783	2217	0.000448																								
44506	45315	809	0.000418																								
44360	44369	9	0.00039																								
44346	44369	23	0.000369																								
31717	31734	17	0.000361																								
22424	22449	25	0.000325																								
25126	25379	253	0.000318																								
44344	44369	25	0.00029																								
15845	16497	652	0.000285																								
31675	33270	1595	0.000238																								
32168	32485	317	0.000232																								
44344	44369	25	0.000191																								
32168	32521	353	0.000172																								
33152	33271	119	8.46E-05																								
31717	31734	17	5.78E-05																								
31717	31734	17	3.83E-05																								
33152	33271	119	9.26E-06																								

Appendix 4: Full GO SLIM enrichment outputs (biological processes and cellular components) from STAR alignment predicted seed sequences

Rules of selection: must be present in more than one library, start sites may deviate +/- 1									
	SeedA					**Seed B**			
Sequence	CCCCCCG	size: 18bp			**Sequence**	UGAUGGUU	size: 16bp		
Direction	_>				**Direction**	<-			
						22998			
						35991			
Start site/s	13523				**Start site/s**	29219			
	4242CX1								
	4242CX3								
	shUBFCX1								
	shUBFCX3					4242CX2			
Treatments present in	shUBF DMOS3				**Treatments present in**	4242CX3			
Goslim analysis (only overrepresented)					**Goslim analysis (only overrepresented)**				
Enriched bio processes (best p<0.05)		**Pvalue**	**#**	**FoldEnrich**	**Enriched bio processes (best p<0.05)**		**Pvalue**	**#**	**FoldEnrich**
	Embryo development	5.35E-03	5	14.94		sensory perception of sound	2.93E-01	3	14.5
Enriched cellular Components top 5 or p<0.05					**Enriched cellular Components top 5 or p<0.05**				
	All P of 1					All P of 1			
Sequence	**Seed C**					**Seed D**			
Direction	CAACCAGAGU	size: 23/46			**Sequence**	CCUGCUUGCC	size : 42/35		
Start site/s	_>				**Direction**	<-			
						32352			
Treatments present in	25497				**Start site/s**	41016			
						shUBF DMSO1			
	4242CX14242CX2					shUBF CX3			
Goslim analysis (only overrepresented)					**Treatments present in**				
Enriched bio processes (best p<0.05)									
		Pvalue	**#**	**FoldEnrich**	**Goslim analysis (only overrepresented)**				
	All P of 1				**Enriched bio processes (best p<0.05)**		**Pvalue**	**#**	**FoldEnrich**
						developmental process	0.104	27	1.99
						perception of sound	0.28	4	8.87
Enriched cellular Components top 5 or p<0.05						systems development	0.694	16	2.19
	All P of 1				**Enriched cellular Components top 5 or p<0.05**				
						intracellular	0.135	45	1.5
						cell part	0.649	55	1.33
	Seed E								
Sequence	UCUGUCU								
Direction	<-								
Start site/s	43068		size: 26						
	4242 DMSO 2								
Treatments present in	shUBF DMSO3								
Goslim analysis (only overrepresented)									
Enriched bio processes (best p<0.05)		**Pvalue**	**#**	**FoldEnrich**					
	Cell-cell signalling	7.97E-02	10	3.82					
	embryo development	8.91E-02	6	6.47					
	synaptic transmission	4.11E-01	7	4.12					
	cellular process	4.56E-01	71	1.33					
Enriched cellular Components top 5 or p<0.05									
	postsynaptic membrane	3.11E-02	2	61.73					
	protein complex	1.82E-01	18	2.05					
	macromolecular complex	1.94E-01	21	1.91					
	nucleus	2.60E-01	18	1.99					
	cellpart	3.05E-01	52	1.39					
	intracellular	3.06E-01	41	1.48					
	organelle	5.65E-01	29	1.57					

Appendix 5: Full GO SLIM enrichment biological processes (bio pro) and cellular components(cell comp) outputs from miRDeep2 predicted seed sequences

miRNA 1			
Seed	CACCACU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
Localisation	23	2.03	1.94E-01
Protein localisation	7	4.26	3.37E-01
Transport	21	2.03	3.46E-01
GO slim cell comp			
nuclear envelope	4	9.38	5.75E-02
nucleus	17	2.2	1.15E-01
intracellular	35	1.47	5.52E-01

miRNA 2			
Seed	AGGAGGU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
Development process	18	2.55	4.20E-02
induction of apoptosis	3	14.87	2.47E-01
systems development	11	2.88	3.56E-01
GO slim cell comp			
nucleus	11	2.11	9.11E-01

miRNA 3			
Seed	GGGAGGU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
Muscle organ development	11	7.19	1.18E-04
mesoderm development	12	4.59	3.41E-03
cellular component morphogenesis	11	4.15	1.93E-02
GO slim cell comp			
postsynaptic membrane	2	65.36	2.78E-02

miRNA 4			
Seed	GGCAGGU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
ectoderm development	20	5.38	3.99E-07
MAPK cascade	11	4.14	2.16E-02
GO slim cell comp			
All P of 1			

miRNA 5			
Seed	CGCCGCU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
Segment specification	2	77.77	5.91E-02
GO slim cell comp			
All P value of 1			

miRNA 6			
Seed	CACCACU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
localisation	23	20.5	1.73E-01
tansport	21	2.04	3.11E-01
protein localisation	7	4.29	3.22E-01
GO slim cell comp			
nuclear envelope	4	9.46	5.82E-02
nucleus	17	2.21	1.05E-01
intracellular	35	4.49	4.83E-01

miRNA 7			
Seed	CACAGCU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
nucleobase-containing compo	77	1.63	2.32E-03
GO slim cell comp			
cell junction	7	5.63	1.84E-02

miRNA 8			
Seed	CGGACGA		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
NA			
GO slim cell comp			
NA			

miRNA 9			
Seed	CACCGCU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
NA			
GO slim cell comp			
NA			

miRNA 10			
Seed	CACAGCU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
localisation	23	2.03	1.94E-01
protein localisation	7	4.26	3.37E-01
transport	21	2.03	3.46E-01
GO slim Cell comp			
nuclear envelope	4	9.38	5.75E-02
nucelus	17	2.2	1.15E-01

miRNA 11			
Seed	CACAGCU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
nucelobase-containing comp	32	1.96	2.96E-02
GO slim Cell comp			
nucleus	18	2.32	4.37E-02

miRNA 12			
Seed	CACAGCU		
Go enrichment			
GO slim Bio pro	**#**	**Fold enrichment**	**P**
NA			
GO slim cell comp			
NA			

Appendix 6: Buffer table

Buffer name	Recipe
1x PBS	8g NaCl, 0.2g KCl, 1.44g Na2HPO4, 0.2g KH2PO4/1L ddH2O
1x SDS lysis Buffer	50mM Tris (pH 8.1), 10mM EDTA, 1% SDS, with fresh proteinase inhibitors
10x TBS	37.3g KCl, 24.23g Tris Base, 87.66g NaCl to 800mls ddH2O. Adjust pH to 7.5-7.6 using HCl. Adjust volume to 1L ddH2O. Dilute to 1x for experimental use.
10x TBS-T	TBS buffer, with 10ml Tween20 added after pH adjustment. Dilute to 1x for experimental use.
1x Transfer buffer	4.54g Tris base and 21.7g Glycine to 500ml ddH2O and mix. Add 300mls methanol and 200ml ddH2O and mix. Adjust total volume to 1.5L with ddH2O
10x SDS running buffer	10g SDS, 30.2g Tris Base, 144g Glycine/1L of ddH2O
2x Laemmli Buffer	4% SDS, 10% β-mercaptoethanol, 20% glycerol, 0.004% bromphenol blue, 0.125M Tris-HCL to a pH of 6.8
Loading dye (10x)	3.9ml Glycerol, 500µl 10% SDS, 200µl 0.5 M EDTA, 0.025g bromophenol blue, 0.025g xylene cyanol, total volume to 10mls with water
DNAse I buffer	10mM Tris-HCL, 2.5 mM $MgCl_2$, 0.5 $CaCl_2$. PH 7.6
20x SB buffer	8g NaOH, 47g Boric Acid in 1L water

Appendix 7: Table of ERCC input concentrations calculated for day one or day two libraries (in ERCC mix 1 or mix 2)

ERCC ID	concentration in Mix 1 (attomoles/ul)	Starting concentration in Mix 2 (attomoles/ul)	Concentration in Mix 1 (attomoles/ul) in day 2 libraries	Concentration in Mix 2 (attomoles/ul) in day 2 libraries	Concentration in Mix 1 (attomoles/ul) in day 1 libraries	Concentration in Mix 2 (attomoles/ul) in day 1 libraries
ERCC-00002	15000	30000	1875	3750	1500	3000
ERCC-00003	937.5	1875	117.1875	234.375	93.75	187.5
ERCC-00004	7500	1875	937.5	234.375	750	187.5
ERCC-00009	937.5	937.5	117.1875	117.1875	93.75	93.75
ERCC-00012	0.11444092	0.17166138	0.014305115	0.021457673	0.011444092	0.017166138
ERCC-00013	0.91552734	1.83105469	0.114440918	0.228881836	0.091552734	0.183105469
ERCC-00014	3.66210938	7.32421875	0.457763673	0.915527344	0.366210938	0.732421875
ERCC-00016	0.22888184	0.34332275	0.02861023	0.042915344	0.022888184	0.034332275
ERCC-00017	0.11444092	0.02861023	0.014305115	0.003576279	0.011444092	0.002861023
ERCC-00019	29.296875	7.32421875	3.662109375	0.915527344	2.9296875	0.732421875
ERCC-00022	234.375	468.75	29.296875	58.59375	23.4375	46.875
ERCC-00024	0.22888184	0.34332275	0.02861023	0.042915344	0.022888184	0.034332275
ERCC-00025	58.59375	58.59375	7.32421875	7.32421875	5.859375	5.859375
ERCC-00028	3.66210938	0.91552734	0.457763673	0.114440918	0.366210938	0.091552734
ERCC-00031	1.83105469	1.83105469	0.228881836	0.228881836	0.183105469	0.183105469
ERCC-00033	1.83105469	0.45776367	0.228881836	0.057220459	0.183105469	0.045776367
ERCC-00034	7.32421875	7.32421875	0.915527344	0.915527344	0.732421875	0.732421875
ERCC-00035	117.1875	117.1875	14.6484375	14.6484375	11.71875	11.71875
ERCC-00039	3.66210938	5.49316406	0.457763673	0.686645508	0.366210938	0.549316406
ERCC-00040	0.91552734	1.37329102	0.114440918	0.171661378	0.091552734	0.137329102
ERCC-00041	0.22888184	0.45776367	0.02861023	0.057220459	0.022888184	0.045776367
ERCC-00042	468.75	468.75	58.59375	58.59375	46.875	46.875
ERCC-00043	468.75	937.5	58.59375	117.1875	46.875	93.75
ERCC-00044	117.1875	175.78125	14.6484375	21.97265625	11.71875	17.578125
ERCC-00046	3750	7500	468.75	937.5	375	750
ERCC-00048	0.01430512	0.02861023	0.00178814	0.003576279	0.001430512	0.002861023
ERCC-00051	58.59375	58.59375	7.32421875	7.32421875	5.859375	5.859375
ERCC-00053	29.296875	29.296875	3.662109375	3.662109375	2.9296875	2.9296875
ERCC-00054	14.6484375	21.9726563	1.831054688	2.746582038	1.46484375	2.19726563
ERCC-00057	0.01430512	0.02145767	0.00178814	0.002682209	0.001430512	0.002145767
ERCC-00058	1.83105469	2.74658203	0.228881836	0.343322754	0.183105469	0.274658203
ERCC-00059	14.6484375	29.296875	1.831054688	3.662109375	1.46484375	2.9296875
ERCC-00060	234.375	234.375	29.296875	29.296875	23.4375	23.4375
ERCC-00061	0.05722046	0.11444092	0.007152558	0.014305115	0.005722046	0.011444092
ERCC-00062	58.59375	14.6484375	7.32421875	1.831054688	5.859375	1.46484375
ERCC-00067	3.66210938	3.66210938	0.457763673	0.457763673	0.366210938	0.366210938
ERCC-00069	1.83105469	3.66210938	0.228881836	0.457763673	0.183105469	0.366210938
ERCC-00071	58.59375	87.890625	7.32421875	10.98632813	5.859375	8.7890625
ERCC-00073	0.91552734	0.91552734	0.114440918	0.114440918	0.091552734	0.091552734
ERCC-00074	15000	22500	1875	2812.5	1500	2250
ERCC-00075	0.01430512	0.01430512	0.00178814	0.00178814	0.001430512	0.001430512
ERCC-00076	234.375	351.5625	29.296875	43.9453125	23.4375	35.15625
ERCC-00077	3.66210938	7.32421875	0.457763673	0.915527344	0.366210938	0.732421875
ERCC-00078	29.296875	58.59375	3.662109375	7.32421875	2.9296875	5.859375
ERCC-00079	58.59375	117.1875	7.32421875	14.6484375	5.859375	11.71875

Appendix 7 continue: ERCC input concentrations calculated for day one or day two libraries (in ERCC mix 1 or mix 2)

ERCC-00081	0.22888184	0.45776367	0.02861023	0.057220459	0.022888184	0.045776367
ERCC-00083	0.02861023	0.00715256	0.003576279	0.00089407	0.002861023	0.000715256
ERCC-00084	29.296875	43.9453125	3.662109375	5.493164063	2.9296875	4.39453125
ERCC-00085	7.32421875	1.83105469	0.915527344	0.228881836	0.732421875	0.183105469
ERCC-00086	0.11444092	0.22888184	0.014305115	0.02861023	0.011444092	0.022888184
ERCC-00092	234.375	58.59375	29.296875	7.32421875	23.4375	5.859375
ERCC-00095	117.1875	29.296875	14.6484375	3.662109375	11.71875	2.9296875
ERCC-00096	15000	15000	1875	1875	1500	1500
ERCC-00097	0.45776367	0.11444092	0.057220459	0.014305115	0.045776367	0.011444092
ERCC-00098	0.05722046	0.08583069	0.007152558	0.010728836	0.005722046	0.008583069
ERCC-00099	14.6484375	21.9726563	1.831054688	2.746582038	1.46484375	2.19726563
ERCC-00104	0.22888184	0.22888184	0.02861023	0.02861023	0.022888184	0.022888184
ERCC-00108	937.5	234.375	117.1875	29.296875	93.75	23.4375
ERCC-00109	0.91552734	0.91552734	0.114440918	0.114440918	0.091552734	0.091552734
ERCC-00111	468.75	703.125	58.59375	87.890625	46.875	70.3125
ERCC-00112	117.1875	234.375	14.6484375	29.296875	11.71875	23.4375
ERCC-00113	3750	5625	468.75	703.125	375	562.5
ERCC-00116	468.75	117.1875	58.59375	14.6484375	46.875	11.71875
ERCC-00117	0.05722046	0.05722046	0.007152558	0.007152558	0.005722046	0.005722046
ERCC-00120	0.91552734	1.37329102	0.114440918	0.171661378	0.091552734	0.137329102
ERCC-00123	0.22888184	0.05722046	0.02861023	0.007152558	0.022888184	0.005722046
ERCC-00126	14.6484375	14.6484375	1.831054688	1.831054688	1.46484375	1.46484375
ERCC-00130	30000	7500	3750	937.5	3000	750
ERCC-00131	117.1875	29.296875	14.6484375	3.662109375	11.71875	2.9296875
ERCC-00134	1.83105469	0.45776367	0.228881836	0.057220459	0.183105469	0.045776367
ERCC-00136	1875	468.75	234.375	58.59375	187.5	46.875
ERCC-00137	0.91552734	1.83105469	0.114440918	0.228881836	0.091552734	0.183105469
ERCC-00138	0.11444092	0.11444092	0.014305115	0.014305115	0.011444092	0.011444092
ERCC-00142	0.22888184	0.22888184	0.02861023	0.02861023	0.022888184	0.022888184
ERCC-00143	3.66210938	5.49316406	0.457763673	0.686645508	0.366210938	0.549316406
ERCC-00144	29.296875	7.32421875	3.662109375	0.915527344	2.9296875	0.732421875
ERCC-00145	937.5	1406.25	117.1875	175.78125	93.75	140.625
ERCC-00147	0.91552734	0.22888184	0.114440918	0.02861023	0.091552734	0.022888184
ERCC-00148	14.6484375	14.6484375	1.831054688	1.831054688	1.46484375	1.46484375
ERCC-00150	3.66210938	3.66210938	0.457763673	0.457763673	0.366210938	0.366210938
ERCC-00154	7.32421875	1.83105469	0.915527344	0.228881836	0.732421875	0.183105469
ERCC-00156	0.45776367	0.11444092	0.057220459	0.014305115	0.045776367	0.011444092
ERCC-00157	7.32421875	10.9863281	0.915527344	1.373291013	0.732421875	1.09863281
ERCC-00158	0.45776367	0.45776367	0.057220459	0.057220459	0.045776367	0.045776367
ERCC-00160	7.32421875	14.6484375	0.915527344	1.831054688	0.732421875	1.46484375
ERCC-00162	58.59375	87.890625	7.32421875	10.98632813	5.859375	8.7890625
ERCC-00163	14.6484375	29.296875	1.831054688	3.662109375	1.46484375	2.9296875
ERCC-00164	0.45776367	0.68664551	0.057220459	0.085830689	0.045776367	0.068664551
ERCC-00165	58.59375	117.1875	7.32421875	14.6484375	5.859375	11.71875
ERCC-00168	0.45776367	0.91552734	0.057220459	0.114440918	0.045776367	0.091552734
ERCC-00170	14.6484375	3.66210938	1.831054688	0.457763673	1.46484375	0.366210938
ERCC-00171	3750	3750	468.75	468.75	375	375