Victoria Harding BMedSci, MBBS, MRCP, PGDip(Onc)
MRC Clinical Research Fellow

Department of Surgery and Cancer
ICTEM Building, Hammersmith Hospital Campus
Imperial College London
W12 0HS

Transcripts and Transcription at an Estrogen Regulated Enhancer of *CCND1*

A Thesis submitted in accordance with the requirements of Imperial College London for the degree of Doctor of Philosophy

June 2021

Supervisors:
Professor Justin Stebbing
Dr Leandro Castellano

Declaration of Originality

I declare that all of the experimentation presented and written in this thesis has been performed by me unless explicitly stated otherwise and acknowledged.

RNA-Sequencing was performed by the ICTEM Genomics Facility run by the Medical Research Council under the guidance of Dr Lawrence Game. Bioinformatic analysis of all RNA-sequencing data generated was undertaken by Dr Leandro Castellano.

Copyright Declaration

Abstract

Enhancer sequences have been well documented for over a decade, and whilst their function as gene expression regulators is widely appreciated, the mechanism by which they exert their control is not yet understood. Transcription of enhancer regions is linked to enhancer activity, but it is unclear if the enhancer RNA (eRNA) transcript is necessary for *cis* regulation or merely a by-product of transcription.

Through RNA-Sequencing of estradiol (E2) treated MCF7 cells and the use of publicly available sequencing data, we have identified a region neighbouring the *CCND1* gene locus which contains at least one enhancer whose bi-directional transcription is upregulated with E2 treatment in estrogen receptor (ER) positive breast cancer cell lines. This enhancer region is known to have *cis*-regulatory effects on the neighbouring *CCND1* gene, whose amplification and overexpression is linked to a poorer prognosis and treatment resistance in ER positive breast cancers.

To determine the role of the bi-directionally transcribed eRNAs arising from this enhancer region, we have identified their cellular location and used appropriate siRNA techniques to knockdown both transcripts. We show that siRNA knockdown of either eRNA does not affect regulation of the neighbouring *CCND1* gene but premature termination of transcription of the antisense enhancer not only knocks-down the eRNA but also down regulates *CCND1* and may have a more global effect on ER regulation. We discuss the challenges encountered in CRISPR/Cas9 mediated knock-in of a polyadenylation signal and compare the resultant effects of knockdown of these eRNA with the premature transcription termination of the enhancer from which they arise and discuss these findings in the context of alternative possible roles for eRNAs.

Acknowledgements

I offer my most sincere thanks to my wonderful supervisors Professor Justin Stebbing and Dr Leandro Castellano who through their kindness and encouragement have enabled me to explore this new world. I am forever grateful for the time and support they have given me over the many years and everything I have learnt from them both.

I would like to thank Dr Lori Pellegrino, Dr Laura Roca-Alonso, Dr Alex de Giorgio and Dr Adam Frampton for welcoming me into their lab family with such open arms and particular thanks to Dr Jon Krell for his friendship, music and support through all of those early mistakes.

And to the group I returned to after parental leave; Dr Paul Cathcart for always being kind and generous with his time and humour, Drs Silvia Ottaviani and Ola Dabrowska for keeping us all in order, to Dr Angela Yiu for her constant willingness to help and to Dr Sladjana Zagorac for her friendship and encouragement through the long days of CRISPR.

My sincere thanks to Action Against Charity and their generous donors for their financial support for the early months of my work, and their on going support of the group, and to the Medical Research Council for awarding me a Clinical Research Fellowship, without which this project would not have been possible.

I am indebted to Professor Buluwela for his CRISPR expertise, kind advice and for support through the many knock-in failures.

A special thank you to my wonderful husband Amit and our children Olivia, Georgia and Raife who have arrived at various stages of this PhD and brought me so much happiness. I would never have been able to juggle my many roles without your unwavering support, patience and love. And finally, to my mum, to whom I owe everything.

Full manuscripts arising

Below are complete papers that have been published during the time of this PhD and which relate to the non-coding genome.

“Gene editing in the context of an increasingly complex genome” Blighe K, DeDionisio L, Christie, KA, Chawes B, Shareef S, Kakouli-Duarte T, Chao-SHern C, **Harding V**, Kelly RS, Castellano L, Stebbing J, Lasky-Su JA, Nesbit MA, Moore CBT. *BMC Genomics* 2018 Aug8;19(1):595

“LMTK3 escapes tumour suppressor miRNAs via sequestration of DDX5.” Jacob J, Favicchio R, Karimian N, Mehrabi M, **Harding V**, Castellano L, Stebbing J, Giamas G. *Cancer Lett*. 2016 Mar 1;372(1):137-46.

“TP53 regulates miRNA association with AGO2 to remodel the miRNA-mRNA interaction network”. Krell J, Stebbing J, Carissimi C, Dabrowska AF, de Giorgio A, Frampton AE, **Harding V**, Fulci V, Macino G, Colombo T, Castellano L. *Genome Res*. 2016 Mar;26(3):331-41

“The role of TP53 in miRNA loading onto AGO2 and in remodelling the miRNA-mRNA interaction network” Krell J, Stebbing J, Frampton AE, Carissimi C, **Harding V**, De Giorgio A, Fulci V, Macino G, Colombo T, Castellano L.. *Lancet*. 2015 Feb 26;385 Suppl 1:S15.

“Non-coding RNAs and the control of hormonal signalling via nuclear receptor regulation.” Ottaviani S, de Giorgio A, **Harding V**, Stebbing J, Castellano L. *J Mol Endocrinol*. 2014 Oct;53(2):R61-70.

“Growth arrest-specific transcript 5 associated snoRNA levels are related to p53 expression and DNA damage in colorectal cancer” Krell J, Frampton AE, Mirnezami R, **Harding V**, De Giorgio A, Roca Alonso L, Cohen P, Ottaviani S, Colombo T, Jacob J, Pellegrino L, Buchanan G, Stebbing J, Castellano L. *PLoS One*. 2014 Jun 13;9(6)

"MicroRNAs cooperatively inhibit a network of tumor suppressor genes to promote pancreatic tumor growth and progression" Frampton AE, Castellano L, Colombo T, Giovannetti E, Krell J, Jacob J, Pellegrino L, Roca-Alonso L, Funel N, Gall TM, De Giorgio A, Pinho FG, Fulci V, Britton DJ, Ahmad R, Habib NA, Coombes RC, **Harding V**, Knösel T, Stebbing J, Jiao LR. *Gastroenterology*. 2014 Jan;146(1):268-77

"Emerging roles of competing endogenous RNAs in cancer: insights from the regulation of PTEN". de Giorgio A, Krell J, **Harding V**, Stebbing J, Castellano L. *Mol Cell Biol*. 2013 Oct;33(20):3976-82.

"The p53 miRNA interactome and its potential role in the cancer clinic" Krell J, Frampton AE, Colombo T, Gall TM, De Giorgio A, **Harding V**, Stebbing J, Castellano L. *Epigenomics* 2013, Aug 5(4):417-28

Table of Contents

VII. Table of Figures

VIII. Table of Tables

IX: Abbreviations

Abbreviation	Definition
3'RACE	3' Rapid Amplification of cDNA Ends
3C	Chromosome Conformation Capture
5C	Carbon Copy Chromosome Conformation Capture
ANOVA	Analysis of variance
ASOs	Anti Sense Oligonucleotide
BRD4	Bromodomain-containing protein 4
CAGE	Cap analysis gene expression
Cas9	CRISPR associated protein 9
Cdk4	Cyclin dependent kinase 4
Cdk6	Cyclin dependent kinase 6
Cdk9	Cyclin dependent kinase 9
ceRNA	Competing endogenous RNA
CHIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-seq	Chromatin Immunoprecipitation with massively parallel sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
dCas9	Dead Cas9
DMEM	Dulbecco modified eagle medium
DNA	Deoxyribonucleotide acid
DNase I	Deoxyribonuclease I
DNase-Seq	DNase I hypersensitive sites sequencing
DSB	Double strand break
dsDNA	Double Stranded DNA
DSS	Dextran serum supplement
E2	Estrogen
EDTA	Ethylenediaminetetraacetic acid
ERE	Estrogen response element
eRNA	Enhancer RNA
ERα	Estrogen receptor alpha
Erβ	Estrogen receptor beta
FACS	Flourescence Activating Cell Sorting
FAIRE-Seq	Formaldehyde-Assisted Isolation of Regulatory Elements – DNA sequencing
FCS	Fetal Calf Serum
FISH	Flourescence in-site hybridisation
FP	Flavopiridol
G1 phase	Growth phase 1
GFP	Green Flourescent Protein
GRO-Seq	Global Run On sequencing
GWAS	Genome Wide Association Studies
H3K27ac	Histone 3 Lysine 27 acetylation
H3K4me1	Histone 3 Lysine 4 monomethylation
HDR	Homology Directed Recombination
Hi-C	High throughput chromatin conformation capture
HOTAIR	HOX transcript antisense RNA
HR	Homologous recombination
KRAB	Kruppel- associated box
lincRNA	Long intergenic non-coding RNA

lncRNA	Long non-coding RNA
LSD1	Lysine Specific demethylase 1
MiR-7	micro-RNA 7
miRNA	Micro-RNA
mL	millilitre
mRNA	Messenger RNA
ncRNA	Non-coding RNA
NELF	Negative elongation factor
NHEJ	Non Homologous End Joining
nt	Nucleotide
ORF	Open reading frame
PAM	Protospacer adjacent motif
PAS	Polyadenylation signal
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
piRNA	Piwi RNA
Pol II	Polymerase II
pRb	Retinoblastoma protein
PRC2	Polycomb repressive complex 2
Rb	Retinoblastoma
RISC	RNA induced silencing
RNA	Ribonucleic acid
RNAi	RNA interference
RNase H	Ribonuclease H
RNP	Ribonucleoprotein
rpm	Revolutions per minute
rRNA	Ribosomal RNA
RuvC	Recombination UV C- resolvase
S phase	Synthesis phase
siRNA	Short interfering RNA
snoRNA	Small nucleolar RNA
SpCas9	Streptococcus pyogenes Cas9
ssDNA	Single stranded DNA
ssODN	Single stranded oligonucleotide
TALENs	Transcription activator-like effector nucleases
TFs	Transcription Factors
TSS	Transcriptional Start Site
ZFNs	Zinc finger nucleases

Chapter 1: Introduction

1.1 The Non-coding Genome

In recent years, high throughput sequencing has challenged the central dogma of biology by revealing that the vast proportion of the eukaryotic genome is actively transcribed, with only 2% being transcribed into protein coding genes. Much of the rest generates an array of non-protein coding RNA (ncRNA) classes, some of which have been shown to play a pivotal role in regulating cellular and organism complexity. The best example to date is that of microRNAs, a 22nucleotide(nt) single stranded RNA molecule that mediates post transcriptional regulation of gene expression through partial or full complementary binding to target mRNA, resulting in their degradation or inhibition of translation[1-3]. Other members of the ncRNA class include other small RNAs such as small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), piwi-interacting RNA (piRNA), ribosomal RNA (rRNA) and the longer transcripts such as long non-coding RNA (lncRNA) and enhancer RNA (eRNA). Genome-wide association studies (GWAS) have shown that most genetic variants that predispose a cell to cancer are outside of the protein coding genome and lie within transcribed non-coding regions that play a pivotal role in regulation of gene expression.

1.2 Long Non-coding RNA

1.2.1 Defining long non-coding RNA

Whilst much is still unknown about the pervasively transcribed group of non-coding RNA, it is generally agreed that they no longer account for transcriptional "background noise". Long non-coding RNAs (lncRNAs) are one described group within the class of non-coding RNA and have been defined by their length of greater than 200nt, their absence of a reasonable sized (greater than 100 codons) open reading frame (ORF) and a lack of homology to any protein. Despite a general lack of sequence conservation and low expression levels, many lncRNAs have reported functionality in a wide range of both biological processes and in disease, although the majority are likely non functional. Challenges in the annotation of lncRNAs have led to some difficulties in estimating their true number in the human genome with estimates ranging from as many as 58000[4] to the more conservative GENCODE7 estimate of 9640 lncRNA loci, representing 15,512 transcripts[5]. Although the majority of these lncRNA remain uncharacterized, gene expression profiling and in situ hybridization studies have shown that lncRNA expression is highly cell specific and responsive to external stimuli and some are undoubtedly key players in cellular control.

In contrast to mRNAs, lncRNAs are generally shorter in length and significantly less expressed[6-8]. Although they are generated through similar pathways to that of protein coding genes, with similar histone modifications and splice site sequences, they are predominantly located in the nucleus and associated with the chromatin and are biased toward two-exon transcripts compared to an average 10 of mRNA. Nonetheless, like other RNAs, most lncRNAs are transcribed by RNA polymerase II (Pol II), are often 5'capped and may be polyadenylated (16.8% of lncRNAs)[5]. In addition, they often exhibit similar histone modifications such as trimethylation of histone 3 lysine 36 (H3K36me3) along their length, and enrichment of histone 3 lysine 4 trimethylation (H3K4me^3) at their promoters which are enriched in transcription factor binding sites[9]. As our knowledge of lncRNAs increases it is likely that the nomenclature used will change and they will be further sub categorised according to their functions. At present the class is highly heterogenous and despite the many thousands of transcripts very few have clearly described roles and mechanisms of action.

1.2.2 Classification of lncRNA

Having been defined by their length of greater than 200nt, long non-coding RNAs are most commonly described according to the genomic location from which they are transcribed (Figure 1.1). Intronic lncRNAs are transcribed from within the intron of a coding gene but do not intersect an exon on the same strand whilst intergenic lncRNAs arise from between two coding genes (and may be called long intergenic RNA (lincRNA)). Sense lncRNAs are transcribed from the coding gene strand and might overlap one or several introns and exons whereas antisense lncRNA intersect a coding gene on the opposite strand. Bidirectional transcripts arise where transcription occurs from both strands in a divergent manner and are commonly associated with enhancer RNAs which are further described below.

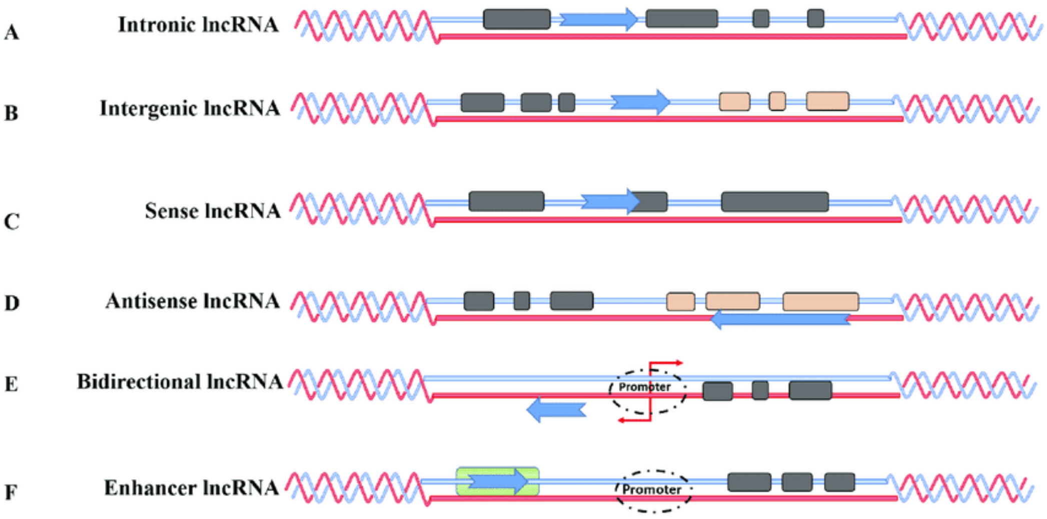


Figure 1.1 Genomic organization of lncRNAs.
Long non-coding RNA can be further classified based on their location with respect to nearby protein coding genes or the regulatory region from which they arise into (A) Intronic (B) Intergenic (C) Sense (D) Antisense (E) Bidirectional and (F) Transcripts arising from enhancer regions. Figure adapted from Richard and Eichorn[10].

1.2.3 Reported functions of lncRNA

Although most lncRNA remain uncharacterised, several mechanisms of action have been attributed to lncRNAs in both transcriptional and post transcriptional gene expression. Whilst most lncRNAs are enriched in the nucleus, with a considerable number of those associated with the chromatin, several roles are also described in the cytoplasm (Figure 1.2). Some nuclear lncRNA molecules have been shown to interact with the chromatin modifying complexes essential for mobilising and restructuring nucleosomes and thereby controlling transcriptional activity through access to condensed DNA [11] [12, 13]. Other nuclear lncRNA may act as a decoy through the sequestration of DNA helicases to prevent chromatin remodelling[14]; or as a guide to recruit and anchor TFs and effector proteins [15] or repressive complexes [16, 17] to the gene promoter and hence induce or repress gene expression; or through an interaction with the nuclear architecture to help co-localisation of distal chromosomal interacting loci[18].

In the cytoplasm, lncRNA are involved in modulating mRNA stability[19],[20] and regulating their translation[21]. They are also known to function as competing endogenous RNAs (ceRNA)[22] and to be precursors of microRNAS[23]. Competitive endogenous RNAs (ceRNAs) have been shown to compete with mRNAs for the binding of miRNAs, acting as molecular sponges and downregulating the effects of the miRNA. Recently, it has also been shown that some putative lncRNAs may have been misannotated as non-coding when in fact they contain short open reading frames encoding small proteins or micropeptides with biological importance[24].

One of the most well characterised lncRNAs to date is HOTAIR, a long intergenic antisense RNA arising from the antisense strand of the HOXC gene. HOTAIR is known to participate in several different processes of normal cell development and is a diagnostic, prognostic and predictive biomarker in many human cancers. HOTAIR acts as scaffold in the binding of two different chromatin modifiers; polycomb repressive complex 2 (PRC2) which leads to the trimethylation of the histone complex H3K27 resulting in transcription repression[15,25,26], and lysine specific histone demethylase 1A (LSD1)[27]. In bringing the two complexes together, it exerts significant repressive control over gene expression. In the cytoplasm HOTAIR functions as a competitive endogenous RNA where it can regulate gene expression through interactions with a range of microRNAs[28, 29]. HOTAIR overexpression is associated with high metastatic potential and poor survival in breast cancer, in part, through its interactions with miR-7 which inhibits cell migration and invasion[30]. The heterogenous functionality of HOTAIR and shared by other lncRNAs is afforded by their ability to fold and alter their 3-dimensional structure and the presence of multiple functional domains within that structure which allow interaction with RNA, DNA and proteins. Understanding the roles that lncRNAs play in epigenetic and transcriptional regulation, and the implications of gain and loss of function of individual lncRNAs may explain the complexity of the human genome despite a relatively small number of protein coding genes and may offer a novel approach to targeting disease.

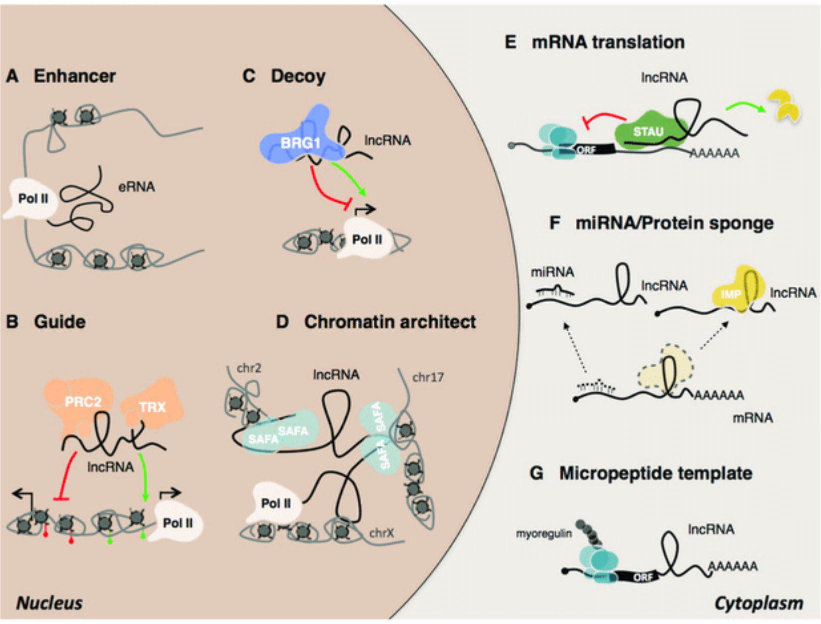


Figure 1.2. Mechanisms of action of lncRNAs in the nucleus and cytoplasm In the nucleus, lncRNAs can act as (A) enhancer RNAs; (B) protein guides and (C) decoys for transcription factors and other proteins and (D) to assist in chromatin architecture. In the cytoplasm they can (E) regulate mRNA translation and stability; (F) act as miRNA sponges and (G) micropeptide templates. Figure adapted from Cipriano& Ballarino [31].

1.3 Enhancers

1.3.1 Enhancer identification and activation

Enhancers are non-coding *cis*-acting DNA elements that play a critical role in transcriptional regulation of tissue and cell-type specific gene expression[32-34]. They are typically 200-2000bp in length[35] and contain numerous closely spaced recognition motifs for sequence specific transcription factors, the binding of which leads to nucleosome remodelling, recruitment of cofactors and initiation of transcription at a target promoter. Enhancers can be either up or down stream of a gene's promoter and can regulate gene expression independent of their distance and orientation to the target gene[36]. Whilst they are known to work in a *cis* configuration, they are also able to bypass neighbouring genes to regulate those located at a considerable distance and sometimes even on another chromosome[37, 38]. Although weakly conserved across species, enhancers are among the most highly constrained sequences across humans and are key elements in establishing spatiotemporal patterns of gene expression.

Cell development, lineage determination and cellular functions all rely on the precise step wise control of enhancer activation. To enable this tight regulation each enhancer contains sequence specific binding sites for transcription factors, chromatin remodelling complexes and coactivators. Extracellular stimuli initiate enhancer activation through signalling pathways and result in stimulus dependent and/or cell type specific transcription factors binding to the closed chromatin at closely spaced recognition motifs [39, 40]. In some cases this leads to the chromatin becoming more accessible for the nucleosome remodelling complexes[39], whilst in others the binding occurs at constitutively accessible DNA. Following their binding to the DNA, TFs recruit coactivators and other complexes (such as the MegaTrans complex in breast cancer cells) including the histone modifiers and DNA demethylating components of coactivator complexes and together they result in further sequential binding of TFs, cofactors and RNA Pol II with consequent cell and signal specific gene expression.

This recognised pathway of enhancer activation has enabled their identification despite enhancers lacking a well defined sequence code, as these epigenetic features have been used to indirectly detect their location[41,42]. Histones are primary protein components of eukaryotic chromatin and they play an important role in gene regulation; H3 and H4 histone tails protrude from the nucleosome and can hence be modified to alter the histone's interactions with DNA and nuclear proteins. Enhancers have been shown to be located in the open chromatin of DNase I hypersensitive regions and flanked by histone H3 covalently modified with monomethylation of lysine 4 (H3K4me^1)[41,43]. Other histone modifications may also be present and can indicate the activity state of the enhancer; poised enhancers often exhibit trimethylation of H3K27 (H3K27me^3) in addition to monomethylation of H3K4 (H3K4me^1) whereas active enhancers usually show acetylation of the lysine 27 (H3K27ac) with H3K4me^1 [44, 45] (Figure 1.3). These chromatin features can be assessed using ChIP-Seq analysis and together with massively parallel reporter assays and cap analysis of gene expression (CAGE) to detect nascent enhancer RNA transcription, can help to identify active enhancers whilst interference of these histone modifications has been shown to have consequences for enhancer activation and their function[46,47].

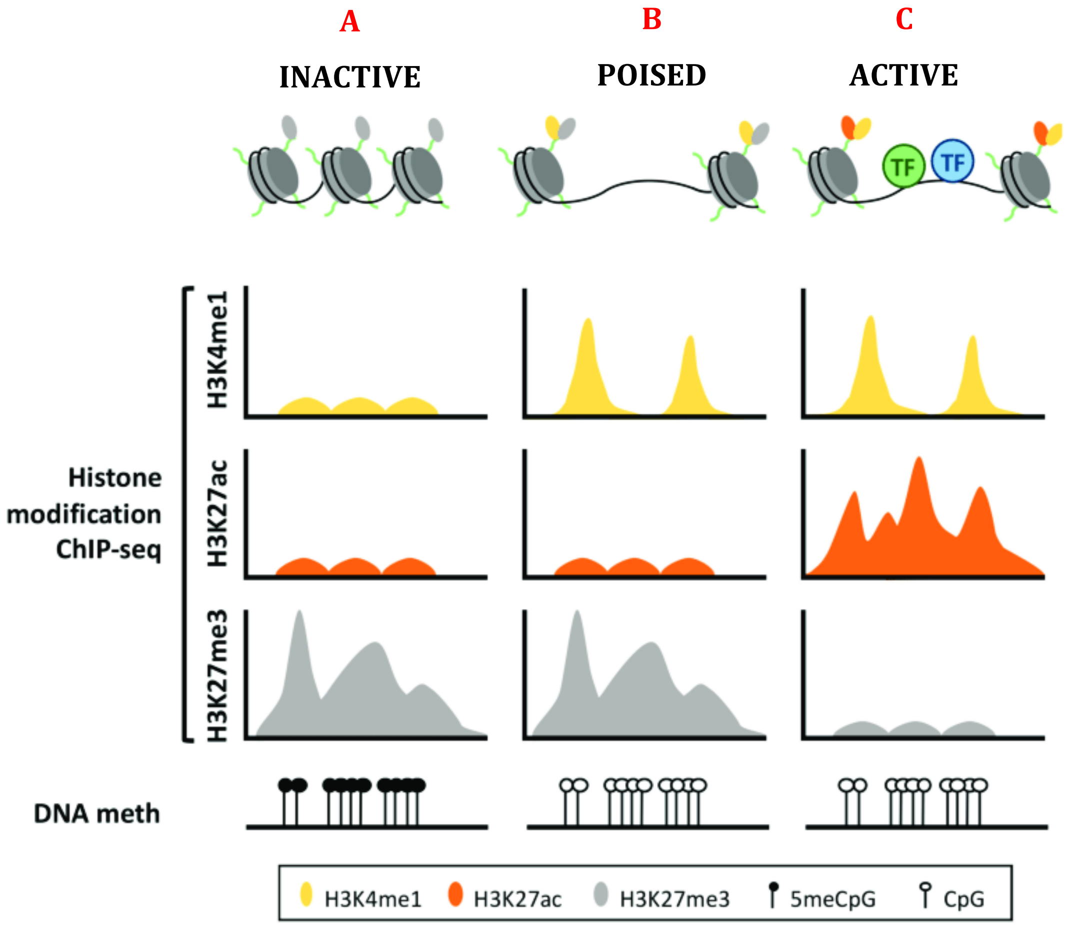


Figure 1.3 Chromatin modifications at inactive, poised and active enhancers.

(A) The inactive DNA is tightly packed around histone proteins marked with H3K27me3 modifications preventing interactions between transcription factors and the DNA.

(B) Monomethylation of H3K4 (H3K4me1) makes the nucleosomes more mobile, allowing their displacement to form highly accessible DNA regions, and the enhancer poised for activation.

(C) Upon activation of the enhancer region, nucleosomes flanking this region acquire H3K27ac, losing the repressing H3K27me3 mark, which subsequently recruits the corresponding transcription factors. Figure adapted from Ordonez et al, 2019[48].

1.3.2 Transcription at Enhancers

In 2010, genome-wide studies revealed that a number of enhancers were occupied by RNA Polymerase II (Pol II) and subsequently transcribed into a class of non-coding RNA called enhancer RNA (eRNA), of various length, polyadenylation status and strand specificity [49] [50]. It has since been shown that transcription at enhancers is pervasive[51] although eRNA abundance varies significantly across tissues, with immune cells, neural tissue and hepatocytes amongst those with the most [52]. Those exhibiting a lower enhancer:gene ratio (such as a smooth muscle and fibroblasts) are typically less responsive to environmental stimuli indicating that active transcription at enhancers is vital in their role in gene regulation.

When genome-wide transcription of enhancers was first reported, the transcripts arising were thought to be a mere by-product of the presence of Pol II at the open chromatin[50],[49]. However whilst their mechanisms of action are yet to be fully understood, and it remains unclear how generalizable their functions are, it is recognised that some eRNAs play a pivotal role in cellular control.

Enhancer transcription, initiation and elongation is similar to that seen at gene promoters with phosphorylation of the C-terminal domain of Pol II being critical and associated with Pol II stability[53]. Poly(A) signals (PAS) are also found immediately downstream of the transcriptional start site of actively transcribed enhancers[52, 54], which have been shown to promote exosome recruitment and instability of Pol II and bring about transcription termination[55]. Complexes involved in termination and cleavage of the eRNA transcripts have been shown to also facilitate in their activation[56] suggesting that eRNA transcription control is tightly regulated and very unlikely to be biological noise.

When initially reported, eRNAs were described as non-polyadenylated, bidirectionally transcribed RNA transcripts shorter than 2kb in length[50]. However, it has seen been shown that eRNAS can be polyadenylated or unilaterally transcribed and indeed, those transcribed from only one strand are usually polyadenylated, longer and arise from more active enhancers[57], although these remain the minority with most being short (with a median of 346nucleotides), bidirectionally transcribed, unspliced and non-polyadenylated[52]. Clearly any structural or functional attributes of eRNAs cannot be assigned to their class as a whole as nearly all combinations of directionality, length, splicing and polyadenylation have been reported. However, generally it is agreed that eRNAs are transcribed from enhancer regions harbouring high levels of H3K4me^{1} and H3K4me^{2} relative to H3K4me^{3} and their expression is positively correlated with an enrichment of histone marks typical of an active enhancer, notably H3K27ac, and a lack of (the repressive) H3K27me^{3}. However, eRNA transcription may be a better prediction of enhancer activity than the described epigenetic modifications[58] although how they play a role in enhancer function is still unclear.

1.3.3 Functional roles of enhancer RNA and their interaction with gene promoters

Despite a decade of active research, the function of eRNAs remain elusive. Whilst it is recognised that their transcription is indicative of an active enhancer [58], their presence does not necessarily indicate a function as they could merely be by-products of transcription events under high concentrations of Pol II. Alternatively their active transcription could enable the critical chromatin architecture alterations required to recruit the necessary proteins for gene transcription; or their transcription could interfere with the transcriptional elongation of mRNA from whose genetic location they arise [59]. Hence, although their presence is correlated with nearby gene transcription, the eRNA molecule itself is not necessarily functional. That said, in recent years there is growing evidence that the eRNA transcript does indeed play a mechanistic role in regulating gene expression.

Large scale analysis has shown a positive correlation between the expression level of eRNAs and their nearby or target genes [39, 40, 50, 52, 60], and knockdown of many (although not all) eRNAs results in downregulation of the target gene [60-65]. In addition, several recent reports have also shown that overexpression of an individual eRNA results in a corresponding increase in their target mRNA [47 62]. How these enhancer derived transcripts regulate this gene expression is still not clear, but the diverse mechanisms of action of individual eRNAs being reported is raising the likelihood that any functional characteristics are unlikely to be generalizable to their class.

Several studies have described interactions of eRNAs with the recruitment and control of RNA polymerase II. Schaukowitch et al described an eRNA interaction with the Negative Elongation Factor (NELF) complex in which it releases the paused Pol II and enables continuation of gene transcription. [61]. On knockdown of the eRNA, they report downregulation of the gene because NELF had not released Pol II. Others have demonstrated roles for eRNA in the recruitment [66] and binding [67] of Pol II to a gene enhancer as well as increasing its phosphorylation to facilitate transcription [68].

It is well reported that transcribed enhancers are more likely to be involved in chromatin looping between enhancers and gene transcriptional start sites (TSS) [69-71]and these events are seen using Chromatin Interaction Analysis by Paired-End Tag sequencing (ChiA-PET) following ligand treatment. Li et al [60] were the first to propose a role for eRNAs in this enhancer–promoter looping when they used chromosomal conformation capture (3C) technology and eRNA knockdown to investigate the effects on enhancer promoter interactions in response to estrogen stimuli. They found that knockdown of the eRNA decreased recruitment of cohesin to the enhancer region and since that early study many others have reported functional roles for enhancer RNAs in the stabilization[72], looping and recruitment of necessary complexes including cohesin, Mediator [64, 73]and Integrator[74] of enhancer-promoter looping.

It is generally thought that enhancer-promoter looping serves to deliver factors such as TFs, trans-activators and Pol II to the gene promoter in the right tissue, at the right time, and those bound factors could help to stabilise the chromatin loops. The mechanism by which the large flexible polymer of chromatin undergoes such looping conformation is unknown but the formation of such large loops by active bending would require considerable energy, and an explanation for this has yet to be found. However, chromatin is constantly moving by constrained diffusion and the radius of this constraint is sufficiently large that any two sequences within approximately 1Mb of each other could randomly encounter each other. If the bound complexes at these sequences had an affinity for one another, it is possible that the chromatin loop could be stabilised through this more passive mechanism and the proximity of the factors to each other could promote further recruitment and binding.

In recent years, the proximity of promoters and enhancers can be visualised using fluorescent in site hybridisation (FISH) in which fluorescent probes are bound to sequences of interest and observed under fluorescence microscopy. Alternatively, chromosome confirmation capture (3C) and its derivative technologies such as 5C, Hi-C and ChiA-PET, have been used to quantify the frequency with which DNA-DNA interactions occur. In some cases, the two methods support each other but in others cross-linked enhancer and promoters captured in 3C do not appear to have a significant spatial co-localization when observed through FISH [75]. One suggestion for such findings is the transient nature of the chromatin loops, which have "un-looped" too quickly to be detected by FISH. Another is that the ligation products of 3C techniques are the result of indirect cross-linking of large sub-structures rather than a true reflection of direct enhancer-promoter linkage.

Whilst knockdown of some eRNA transcripts has been shown to reduce enhancer-promoter interactions and to downregulate gene transcription[60], transcription elongation by Pol II has been chemically inhibited by the cyclin dependent kinase 9 (Cdk9) inhibitor flavopiridol (FP) and shown no consequent effect on enhancer-promoter looping or the assembly of the polymerase initiation complex, enhancer complexes or histone modifications [61 76]. Of course whilst such transcription inhibition experiments offer some insight, the impact of flavopiridol on transcription across the whole genome limits its ability to decipher how eRNA transcription affects target gene regulation.

In summary, whilst much has been discovered about the mechanics of individual enhancer transcripts in the past decade, our understanding of eRNAs role in regulating gene expression remains poor. My CRISPR editing of an enhancer region was thus generated from this ongoing question; is it the transcription of eRNA or the eRNA itself that is primarily responsible for its enhancing function?

1.4 Estrogen Receptor binding at Enhancers

The estrogen receptors are members of a ligand-regulated nuclear receptor superfamily that play a pivotal role in many aspects of human physiology including sexual development and reproduction, cardiovascular health and bone metabolism. Whilst the two ER isoforms, ERα and ERβ, share almost identical DNA binding domains, their tissue expression, tertiary structure and biological functions all differ markedly. Despite these differences both are primarily transcription factors which either hetero- or homo-dimerize following binding of their ligand 17β estradiol (E2). Once activated by E2, the ER regulates target gene expression by binding the DNA at specific sequences called estrogen response elements. However, dysregulation of these pathways is responsible for the progression of hormone sensitive cancers including breast, endometrial and ovarian cancer, as well as osteoporosis, neurodegenerative disease and insulin resistance[77, 78].

In cancer cells, the binding of E2 to ERα stimulates unregulated cellular proliferation and hence increases the potential for tumour formation. Whilst many studies have shown ERα to be responsible for the estrogen dependent changes in breast cancer, ERβ may also play a role [79]. In addition to direct binding to the genomic DNA, ERs are also found in smaller quantities associated with the cytoplasmic membrane. When activated by the lipophilic and freely diffusing E2, these membrane-associated ERs stimulate a kinase-mediated signalling pathway resulting in changes to the localization and activity of nuclear transcription factors and consequent binding to the DNA through other regulatory elements [80 81]. Both pathways ultimately result in the recruitment of co-regulators, histone modifications and other remodelling events which regulate gene transcription. However, despite extensive research into how ERα regulates this transcription, many questions have remained about the exact mechanisms of control. Nevertheless, in 2013 two groups[60, 76] reported on the importance of enhancers in the role of ER regulation and sought to determine the properties of ER bound enhancers and the effects of their active transcription and their work has been instrumental in the work I present in this thesis.

Using ERα chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-Seq) analysis, Li et al [60] identified over 31 000 activated ER binding sites in the E2 treated breast cancer cell line MCF7. Surprisingly, only 902 of these were within gene promoter regions whereas more than 7000 sites exhibited the modifications of an enhancer (HsK4me^1 and H3K27ac). Further still, almost all of the upregulated protein coding genes did not bind ER at their promoter, but instead were found to have ER bound enhancer regions within 200kb of their transcriptional start site. These findings suggested a key role for enhancers in E2 regulated gene expression, particularly those located close to a regulated gene. Li et al [60] also reported that these ER bound enhancers were rapidly transcribed into eRNAs (almost all bi-directionally) and their expression strongly correlated with nearby genes. Knockdown of those E2 regulated eRNAs resulted in reduced expression of the neighbouring gene but had no effect on ER binding at the enhancer.

At the same time, Hah et al [76] used Global Run On sequencing (GRO-Seq) to generate a global profile of active transcription at ER binding sites in MCF7 cells. GRO-Seq assays the location and orientation of all active RNA polymerases to generate a global profile of active transcription at a given time. Like many studies that have come since, they found ER binding sites with nascent eRNA production strongly correlated with chromosomal looping to target gene promoters but found no looping at ER binding sites where they were not. In addition, in using the CdK inhibitor flavopiridol to block enhancer transcription elongation they saw no effects on TF binding, histone modification or clustering of co-activators at the enhancer suggesting that neither the eRNAs nor their transcription is required for enhancer activation.

Hah et al [76] identified all E2 regulated eRNAs transcribed from ER bound enhancers within 40 minutes of E2 treatment. They observed that ER bound enhancers producing short bi-directional eRNAs positively correlated with levels of pioneering factors such as FOXA1, co-regulators such as P300 and active histone modifications, as well as the most accessible chromatin (defined by DNase-seq and FAIRE-seq) and were twice as likely to be involved in chromosomal looping when compared with ER bound enhancers producing no transcripts. Looping from ER bound enhancers correlates with estrogen dependent target gene activation [82] and thus the transcription of E2 bound enhancers is clearly an important feature in E2 regulated gene expression.

Although these findings indicate that eRNAs are indicators of active enhancers, the function of the eRNA itself remains unknown. Whilst Hah et al[76] found that the assembly of enhancer complexes can be dissociated from eRNA production, suggesting that eRNA production occurs after the assembly of the active enhancer complex, rather than being a prerequisite for its formation, Li et al [60] and others [83] have shown that knockdown of some eRNAs transcribed from proximal E2 bound enhancers downregulates target gene expression. Hence I sought to identify individual E2 regulated eRNAs and further investigate the mechanisms by which they may exert their control

1.5 CCND1

1.5.1 The Cell Cycle and Cyclin D1

Cyclins are the regulatory subunits of a protein kinase family and through their association with cyclin-dependent kinase (cdk) subunits they can regulate cell cycle progression and proliferation. Cyclin D1 (*CCND1*) is encoded by the *CCND1* gene located on 11q13 and has multiple roles during cell cycle, propagation and tumorigenesis [84]. It forms a complex with the cyclin dependent kinases (CDKs) cdk4 and cdk6, initiating phosphorylation of the retinoblastoma (Rb) family of transcriptional repressors and results in the synthesis of S-phase genes [85] (Figure 1.4A). This *CCND1*-Cdk4/6 activity peaks in the late G1 phase and is one of the main determinants in initiation of DNA replication and completion of the cell cycle.

Amplification of the genomic locus at 11q13 and/or overexpression of *CCND1* occurs in a substantial proportion of human cancers, including breast[86], colon cancer[87] and head and neck cancers[88]. *CCND1* is overexpressed in up to 50% of human breast cancers but deleted in only 5%[89] suggesting that transcriptional or post translational activation of the gene could be responsible for its oncogenic properties. E2 bound ERα is positively correlated with high *CCND1* expression levels in breast cancer cells and studies have shown that induction of *CCND1* mRNA expression in breast cancer cells can mimic the pro-proliferative effects of estrogen and can drive cell proliferation[90]. Furthermore, when either induced or overexpressed, *CCND1* can overcome the anti-proliferative effects of anti-estrogen treatment raising the possibility of its role in hormone therapy resistant breast cancer [91].

In addition to its oncogenic properties when dysregulated, it has also been shown that *CCND1* is important in the repair of DNA damage through a cdk-independent manner via its recruitment to DNA damage sites by *BRCA2*[92]. In combination with *BRCA2* and *RAD51,* important proteins in DNA repair, *CCND1* is required by cells to utilise the homologous recombination (HR) DNA repair pathway in which double strand breaks can undergo high fidelity repair (Figure

1.4B)[92]. This requirement for *CCND1* in repair of DNA damage is at odds to its tumorigenic properties when overexpressed but it is possible that high levels of *CCND1* in cancer cells are acting to support the survival of oncogene induced DNA damage. Whilst much research is still needed in this apparent juxtaposition, it is relevant to the CRISPR work conducted during this work because of the intentional double strand breaks caused by CRISPR in the enhancer of *CCND1*.

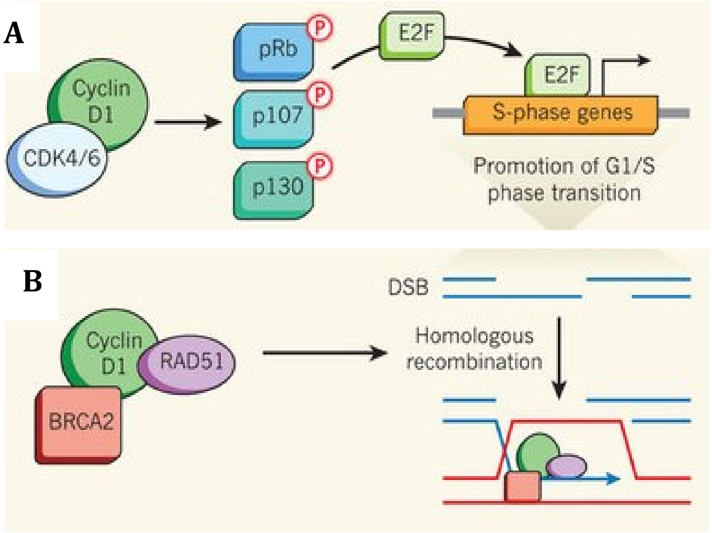


Figure 1.4 Roles of *CCND1*

CCND1 is important in controlling cell cycle through G1 and also has a role in homologous recombination repair of double strand breaks

A. *CCND1* interacts with the cdk enzymes cdk4 and cdk6, initiating phosphorylation of the retinoblastoma family (pRb, p107 and p130). *E2F* transcription factors are subsequently liberated enabling genes of the S phase of cell cycle to be transcribed. This activity peaks in late G1 phase.
B. By binding to the enzyme *RAD51*, *CCND1* co operates with *BRCA2* to stabilise an important complex for HR repair of double strand breaks.

Both mechanisms may contribute to the oncogenic properties of *CCND1* when expressed at high levels. Adapted from Bartek and Lukas, 2011 [92].

1.5.2 Targeting cell cycling in the systemic treatment of ER positive breast cancer

In recent years the cyclin-dependent kinases have become a successful target in the treatment of advanced ER positive breast cancer. The cdk4/6 inhibitors act at the G1/S cell cycle checkpoint to prevent cycle progression and ultimately lead to cell cycle arrest. Clinical studies have shown an impressive increase in overall survival in patients treated with cdk4/6 inhibitors with ER positive advanced breast cancer (in both first and second line settings) compared with endocrine therapy alone[93-95]. These clinical outcomes give further credence to the close interplay between ER regulation, *CCND1* and cdk4/6 and the possible benefits of *CCND1* as a therapeutic target.

1.6 Modulation of eRNA function

1.6.1 Knockdown of eRNA using RNA interference and Antisense Oligomirs

Loss of function models are commonly used to investigate the role of RNA transcripts and can be generated by transcript knockdown with RNA interference (RNAi). Short interfering RNAs (siRNAs) are double stranded RNA synthesized with a complementary sequence to the ncRNA of interest and, after transfection into the cell, associate with multiple protein factors to form the RNA-induced silencing complex (RISC). From the RISC, siRNAs base pair to their target RNA and cleave it, thus preventing translation or down-stream processing of the RNA transcript (Figure 1.5).

RNAi has several limitations in generating loss of function models with lncRNAs. Firstly, the RNAi machinery is mainly cytoplasmic and hence those lncRNAs found primarily in the nucleus or are chromatin bound are likely to escape the protein factors which mediate RNA degradation. Secondly, siRNAs are less efficient at targeting RNAs with a strong secondary structure, which is typical of lncRNAs; and thirdly, RNAi machinery acts on the post-transcription product rather than interrupting the process of transcription, which may be more relevant in the role of enhancer RNAs where it is postulated that the act of transcription is important.

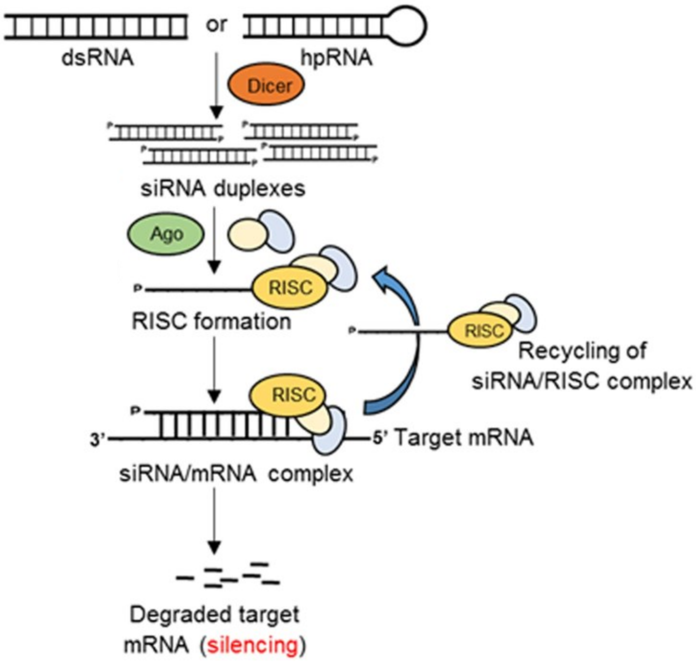


Figure 1.5 Schematic of RNAi-mediated gene silencing.
Double-stranded RNAs or hairpin RNAs (hpRNAs) generate small siRNA duplexes by the action of Dicer. The guide RNA strand binds with Argonaute (Ago) and other proteins to form an RNA-induced silencing complex (RISC). The siRNA/RISC complex then binds the complementary sequence of the target mRNA resulting in the degradation of the target transcript or inhibition of translation. The components of siRNA/mRNA complex can be recycled to the RISC complex or generate siRNA duplexes. Adapted from Majumdar et al[96].

Antisense oligomirs (ASOs) do not rely on the RNAi machinery to knockdown lncRNA expression. Instead, these synthetic oligonucleotides to bind the target RNA and trigger degradation by endogenous RNAse H, an enzyme that cleaves the RNA strand in a DNA/RNA heteroduplex. Chemical modifications have improved the functionality of ASOs and slowed their rapid degradation in the intracellular environment. GapmeRs are ASOs with melting temperature enhancing modifications at the ends to increase binding affinity to their target whilst retaining a central DNA core to form a substrate for RNAse H when it is duplexed with the RNA target. Such modifications can greatly improve the potency of the ASOs. In comparison to the siRNAs, ASOs can reportedly target nuclear RNAs as well as nascent transcripts but are generally short lived in their efficacy and tend to be more toxic than siRNA. This may be relevant as some lncRNA can be exclusively found in the nucleus where siRNAs will have little effect.

1.6.2 Genomic editing with CRISPR/Cas9

The CRISPR/Cas9 system is a programmable, sequence-specific genome editing system with enables endonucleases to precisely edit genomic loci[97-99]. Clustered regularly interspaced short palindromic repeats (CRISPRs) are repeating DNA sequences in the genome of prokaryotes, such as bacteria and archaea. These sequences are derived from the DNA of bacteriophages that have previously infected the prokaryote. When first discovered, they were thought to be a novel DNA repair mechanism but such CRISPR systems are actually part of the adaptive immune response systems seen in prokaryotes in which they are able to cleave the nucleic acids of invading viruses, thus protecting themselves from repeated viral infection. Cas9 (CRISPR-associated protein 9) is an enzyme that uses the CRISPR sequence as a guide to recognise and cleave invading complementary double stranded viral DNA. However, in 2012 George Church, Jennifer Doudnam, Emmanuelle Charpentier and Feng Zhang discovered that they could use the CRISPR -Cas9 system to target a specific region in the genome using a sequence specific guide RNA (gRNA) and the CRISPR-Cas9 system as the DNA editing tool.

The CRISPR/Cas9 genome editing system contains two essential components: a single guide RNA (sgRNA) and an endonuclease (Cas9) which contains two conserved nuclease domains, HNH and RuvC, which cleave the DNA strand complementary and non-complementary to the guide RNA respectively. The gRNA is a synthetic RNA molecule composed of a scaffold for Cas9 binding and a user designed specific sequence of 20nt (called a spacer) that defines the genomic target to be modified[100]. The endonuclease can be directed to any genomic locus by the gRNA sequence (which base pairs with the target DNA), and there scans the local DNA for a short sequence known as a protospacer adjacent motif (PAM) which is found at the 3' end of the target sequence[101]. Cas9 then generates a double strand break 3 base pairs upstream of the PAM (Figure 1.6) The most commonly used Cas9 originates from *Streptococcus pyogenes* (*Sp*Cas9) for which the PAM is 5'-NGG although other variants of Cas9 are available. Most Cas9 nucleases can tolerate up to five mismatches between the sgRNA and the target genomic sequence, but the 10-12 most proximal bases to the PAM are the main determinant of sgRNA specificity[102]. As a result of imperfect matching, sgRNA can bind to other areas of the genome and result in off target effects although imperfect matching of the sgRNA may not be sufficient for DNA cleavage and hence may not result in insertions and deletion (indels).

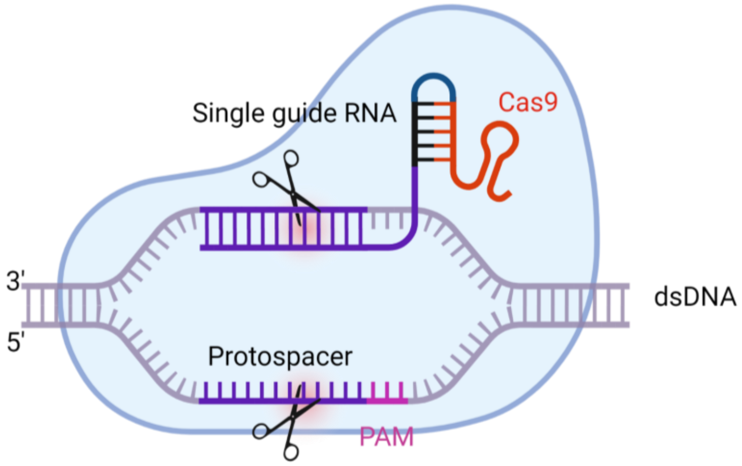


Figure 1.6 Cas9 is guided by sgRNA to target sequence
Cas9 endonuclease is guided by a single guide RNA (sgRNA) consisting of a 20nt spacer sequence homologous to the target sequence. On complementary binding to the DNA, Cas9 scans the local DNA for a protospacer adjacent motif (PAM) and cleaves a double strand break 3nt upstream. Created with BioRender.com

The CRISPR/Cas9 system exploits the double strand break (DSB) repair pathway to create mutations at specific genomic locations within the DNA. These breaks are repaired either through the error-prone non homologous end joining (NHEJ), which can result in random insertion and deletion mutations (usually smaller than 10bp), or by homology directed repair (HDR), which results in an exact repair but is much less efficient (Figure 1.7)[97, 99, 100].

NHEJ occurs more frequently within a cell and is effective at mediating gene knockouts because the indels can result in frameshifts and premature stop codons[103]. If the target sequence is still recognisable to the gRNA, Cas9 may continue to cut and re-ligate at the site (called re-targeting), leaving a scar until it is no longer identifiable. The NHEJ pathway can also be used to remove stretches of the genome by directing two sgRNAs to either end of the sequence and generating multiple double strand breaks[104]. However, removal of DNA sequences is an inefficient modification method, with a reported efficiency of targeted deletions by paired sgRNAs being approximately 25%, and decreasing with increasing deletion size plus the additional chance of off target effects with more sgRNAs[105].

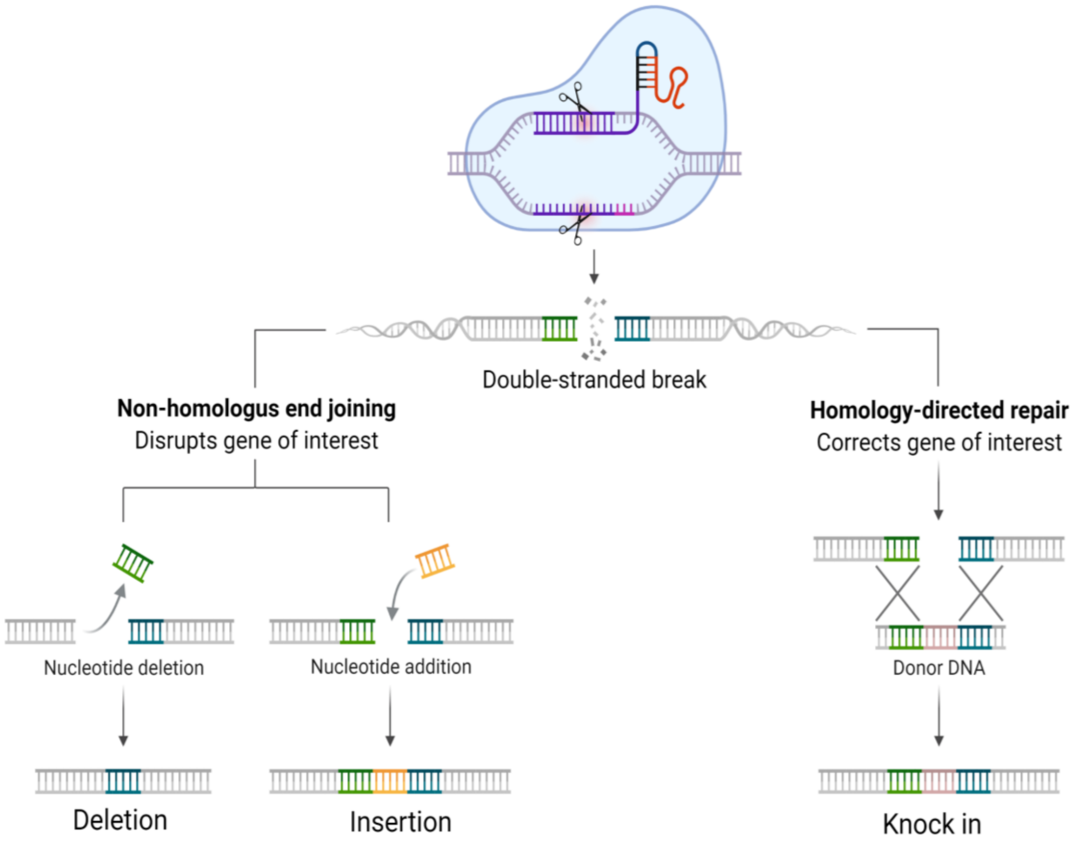


Figure 1.7 Cas9 mediated DSB can undergo repair via NHEJ or HDR

DNA repaired through the NHEJ pathway can result in the insertion of additional nucleotides, or in the deletion of some and longer insertions can result in frameshifts. These indels may result in nonsense mutations and consequent knockout of the gene/non-coding region. If a donor template is available, the DNA can be repaired via homology directed repair pathway which can be utilised by CRISPR to insert a sequence of interest into a specific region of the genome. Created with BioRender.com

Whilst NHEJ is the preferred method of repair by the cell, if provided with a donor template, the DNA may repair with homology-directed repair (HDR) and the recombination can result in a knock-in of a desired sequence. The donor template can be in the form of double stranded (ds)DNA plasmids, single stranded (ss)DNA oligonucleotides (ssODNs) and dsDNA linear fragments with each having their own advantages and disadvantages. Plasmids are more useful for insertion of longer sequences whilst risking plasmid integration and greater cell toxicity, whilst ssODNs are good for shorter inserts but are likely to be less efficient because of shorter homology arms[106].

Overall, it is recognised that HDR mediated knock-in using CRISPR has poor efficiency and is considerably less efficient than NHEJ mediated genome editing. Whilst many optimisation techniques have been developed since the widespread use of CRISPR, the reported efficiency of knock-in mediated HDR is less than 10% in most experimental conditions, and often much lower[107]. Even when knock-in has been successful, the low efficiency makes isolation of a positive clone challenging.

1.6.3 Modifications of the CRISPR/Cas9 system

Due to the variable efficacy and off target effects of the CRIPSR/Cas9 system, many groups have reported adaptations to all aspects of the process. For example, modifications to the Cas9 endonuclease domain have enabled fusion with activator and repressor proteins, enabling the delivery of such effectors direct to the target gene. Transcriptional activators (CRISPRa) can assist in the recruitment of cofactors or histone modifications with a resultant up regulation of the target, or alternatively, repressive elements fused to Cas9 such as KRAB can downregulate the target[108-110]. In fact, Cas9-KRAB complexes have been shown to repress the expression of protein coding and non-coding genes on a genome-wide scale, in part through reducing chromatin accessibility at both enhancers and promoters[111]. A point mutation of the Cas9 nuclease can render it incapable of cleaving the DNA and is thus termed dead Cas9 (dCas9). dCas9 is still able to bind to the gRNA and to the DNA and although it does not mediate a double strand break, its 3 dimensional structure can be enough to block TF binding and RNA Pol II initiation and elongation[112, 113]. This is thus called steric hindrance and is used in CRISPR interference (CRISPRi).

Although only in widespread use for less than a decade, there are many modifications to the CRISPR/Cas9 system that have enabled groups to harness the power of CRISPR with optimising HDR being a primary focus. Modifications such as asymmetric homology arms complementary to a non-target locus (long arm on the PAM side) have been reported to induce higher HDR efficiency[114] as well as the use of paired sgRNAs with a Cas9 nickase mutant in which simultaneous nicks mediated by offset gRNAs can result in double strand breaks with considerably less off target activity and much higher fidelity[98].

1.6.4 CRISPR-Cas9 mediated engineering of Enhancers

Due to the design of the CRISPR system, Cas9 can be directed to any part of the genome, including non-coding transcriptional regulatory elements such as enhancers. There have been numerous studies reporting CRISPR/Cas9 mediated editing of enhancer regions. Much of the work has centred around interrupting the binding of known TFs to the enhancer, although it has been shown that the repressing dCas9-KRAB directed to enhancer regions may have significant off target effects on histone modifications at their target promoters, thus actually silencing the promoter rather than the enhancer. Others have reported that targeting of an enhancer by a single sgRNA to produce random indels and mutations can result in comparable genetic effects to the deletion of an entire enhancer using two sgRNA targeting each end of the sequence [104 115].

Other dCas9 fusions developed to date and used in enhancer regions include a dCas9-LSD1 fusion[116] which catalyses the removal of H3K4 methylation and a dCas9-p300[117] fusion which enables the modulation of H3K27ac at either end of the enhancer region. Although such manipulation of gene expression via histone modification has been shown in vivo, the wider implications of off target effects has yet to be fully determined. CRISPR/Cas9 has also been used in functional screening to identify endogenous enhancers elements. Korkmaz et al[118] used CRISPR-Cas9 to identify functional elements by disrupting the transcription factor binding sites of well recognised TFs and screening for phenotypic changes. It was in this way that the enhancer element neighbouring *CCND1* that I identified through the active transcription of bidirectional eRNAs was confirmed as being a *CCND1* enhancer, and the CRISPR mediated genome editing by them, in part, led to the work carried out and described in this thesis.

1.7 Aims and Hypothesis

ERα regulation of the genome is extensive and not limited to protein coding genes. lncRNA and eRNA regulation by ERα may contribute to breast cancer progression and I hope to reveal their mechanism of action. This is important because most currently available therapies are aimed at preventing ligand-receptor binding or reducing endogenous production of estrogen, rather than targeting the genes and regulatory elements of the genome that are responsible for the estrogen signalling pathway. Understanding the role of these regulatory elements in the up regulation of cell proliferation and acquisition of metastatic potential may identify novel pathways for druggable targets in the treatment of ER driven cancers. I hope to add to the growing body of literature suggesting that transcripts arising from enhancer regions play a functional role in the cell and are not biological noise.

Hypothesis:

Estrogen responsive enhancer RNAs play an important role in gene regulation in ER positive breast cancer cells. The eRNA transcript arising from the enhancer of CCND1 is involved in CCND1 gene regulation.

Aims:

- Genome-wide identification of estrogen regulated lncRNAs and eRNA in the breast cancer cell line MCF7.

- Validate RNA-seq expression and identify the cellular location of the estrogen regulated eRNA transcript arising from the enhancer of CCND1.

- Generate loss of function models to better understand the cellular function and mechanism of action of the eRNA transcript.

- Use CRISPR to prematurely terminate transcription of the CCND1 enhancer and assess its impact on both CCND1 gene regulation and genome wide.

Chapter 2: Materials and Methods

2.1 Mammalian cell culture and growth media

CELL TYPE	TISSUE	MORPHOLOGY	TUMOURIGENICITY
MCF7	Breast	Epithelial	Human breast cancer cell line- ductal ER positive
MCF7 LUC	Breast	Epithelial	Human breast cancer cell line- ductal ER positive
T47D	Breast	Epithelial	Human breast cancer cell line- ductal ER positive
MDA-MB-231	Breast	Epithelial	Human breast cancer cell line-adenocarcinoma – Her2 negative, ER negative

Table 2.1: Mammalian cell lines used

CELL TYPE	MEDIA	ADDITIVES	STORAGE
MCF7 MCF7 -LUC MDA-MB- 231	DMEM medium (modified) (Gibco®)	2mM Glutamine 50 units/ml Penicillin 50µg/ml Streptomycin 10% fetal calf serum (FCS)	4°C, used within one month
T47D	RPMI medium (modified) (Gibco®)	2mM Glutamine 50 units/ml Penicillin 50µg/ml Streptomycin 10% fetal calf serum (FCS)	4°C, used within one month

Table 2.2: Normal growth media

2.2 Primers used for reverse transcription

Primers were designed using PerlPrimer design software and checked in silico using the UCSC genome browser (http://genome.ucsc.edu)[119].

GENE	**Primer sequence (5' to 3')**
GAPDH	forward TGAAGGTCGGAGTCAACGGATTT reverse GCCATGGAATTTGCCATGGGTGG
GREB1	forward CAAAGAATAACCTGTTGGCCCTGC reverse GACATGCCTGCGCTCTCATACTTA
PS2 (TFF-1)	forward CCAGACAGAGACGTGTACAG reverse GTGTCGTCGAAACAGCAG
PUM1	forward CAGATCATTCAGTTTCCCAG reverse GACAGTACAGAATTGACCTC
CCND1	forward CCTGTCCTACTACCGCCTCA reverse TCCTCCTCTTCCTCCTCCTC
MALAT1	forward TGGGAGTGGTAGGATGAAAC reverse CCTTCCCGTACTTCTGTCTT
ACTB	forward AGCACAGAGCCTCGCCTT reverse CATCATCCATGGTGAGCTGG
miR-17-92	forward AAAGGCAGGCTCGTCGTTG reverse CGGGATAAAGAGTTGTTTCTCCAA
FOXC1e (1)	forward CATGAAAGGTGAAGCGGAAATAC reverse TGAAGGAGCAGGTGAAACG
FOXC1e (2)	forward CTGAGGAACACAAGACTAGCC reverse ACTGGACTCATTTTGGGACATC
rs614367 (1)	forward CTTGGCTTCTCTGCAACTCC reverse CTGTCTTGTCGGAGGAAGTC

rs614367 (2)	forward GAGCCTCTTCCTTGGCTT reverse ACTGAGAGAGGTGGAGACAG
GAPDH for genomic DNA	forward TCAAAGGTGGAGGAGTGG reverse ACATCATCCCTGCCTCTAC
CCND1e(sense) 1	forward ATCTGCTTGGCTTCTGGT reverse GGTGACTGTCCTCAAGATAGTG
CCND1e(sense) 2	forward CGGTCATGTGTATTCAGCAG reverse CTCACCAGAAGCCAAGCA
CCND1e (sense) 3	forward GCATTGAGGCCATCTTTCTG reverse ACCCTTCTTTGACTCAGCAT
CCND1e(antisense)	forward ATGGGAGTGGAACTGAAGG reverse CTGCTGAATACACATGACCG
Agami rd 1	forward CTCTGGGATCCTGTTTACCT reverse TCAGGTATGCCTCTTGTTTCC

Table 2.3: Primer sequences used for reverse transcription

2.3 Mammalian cell culture

2.3.1 Cell culture

MCF7 and MCF7-Luc cells were kindly provided by Dr L Magnani and Professor L Buluwela respectively (Department of Surgery and Cancer, Imperial College London). ZR 75-1, T47D and MDA-MB 231 were purchased from the American Tissue Type Culture Collection (ATCC). Cells were regularly tested to ensure no mycoplasma infection (MycoAlert, Lonza, UK).

Cells were maintained at 37°C in a humidified 5% CO2 incubator and cultured in 150 cm^2, 75cm^2 or 25cm^2 flasks, 100-mm or 60mm dishes or 6, 12 and 24-well plates unless otherwise specified. Prior to reaching 80% confluence, cells were passaged; medium was aspirated, cells washed with warm PBS solution and then detached from the culture vessel with EDTA-trypsin (1X trypsin in 0.02% EDTA) at 37°C for 3-10 minutes depending on cell line. Media containing 10% FCS was added to inactivate the trypsin at a 1:1 ratio before transferring the cell suspension to a 15 mL sterile centrifuge tube. Cell clumps were disrupted with gentle pipetting. The cell suspension was then centrifuged for 3 minutes at 1300 rpm. Subsequently, the supernatant was removed and discarded and the cell pellet re-suspended in an appropriate volume of medium for seeding into new flasks with fresh media. Cells were counted with a haemocytometer with a 0.1mm sample depth and light microscope.

In order to maintain cell stocks, aliquots were cryo-frozen regularly. When at approximately 80% confluency, cells were trypsinised and quenched as above, but the pellet was resuspended in 4.5mL of freezing media. 1.5ml of this cell suspension was then aliquoted into three 2mL cryogenic vials and frozen by storing in -80C in a suitable container, such as a Mr. Frosty™. The samples were later moved to liquid nitrogen for long term storage. When required again, the cells were rapidly thawed in a water bath at 37°C and transferred to pre warmed media. The cells were then centrifuged at 300 x*g* for 3 minutes, supernatant discarded, cell pellet resuspended in 5mL of culture media and pipetted into a T25 tissue culture flask and maintained as usual.

2.3.2 Starving cells of estradiol

Following the same procedure as for passaging, cells were washed with PBS and trypsinised at 37°C for 3-10 minutes. Phenol red free DMEM media containing double charcoal stripped FCS was added to inactivate the trypsin at a 1:1 ratio. Cell suspensions were centrifuged as for normal passage and resuspended in the stripped media. The process of centrifuging and resuspending was repeated a total of 2 times, before cells were plated in the appropriate culture vessel and left to adhere under normal growth conditions. Cells were starved of estradiol stimulation in this manner for 72 hours prior to further treatments.

2.3.3 Treating cells with estradiol, tamoxifen and vehicle

Following 72 hours of estradiol deprivation as described above, media was removed from adherent cells and replaced with one of the following. Estradiol was added to phenol red free DMEM supplemented with PSG and 10% double stripped activated charcoal FBS at a concentration of 10nM and 100% ethanol at 10mM as vehicle treatment. Cells were treated for 0,1,3,6 or 24 hours, at which time they underwent extraction of DNA, RNA or protein as described below, and analysed accordingly.

2.4 Quantitative real-time Reverse Transcription-PCR

2.4.1 RNA preparation

Using appropriate personal precaution and working inside the fume hood, cells were lysed with Trizol reagent (Invitrogen) as per the manufacturer's instructions. Briefly, media was removed and cells washed with ice cold PBS and then lysed with 1mL of trizol per 10cm^2 for cells grown in monolayer. The lysate was pipetted several times prior to incubation for 5 minutes in eppendorf tubes.

Two methods of RNA extraction have been used in this work. The majority has been conducted using the Zymo Direct-zol RNA extraction kit as per manufacturer's instructions (Zymo Research, Irvine, CA, USA). Briefly, 1 volume of 100% ethanol was added to the sample lysed in trizol and mixed thoroughly. Zymo-Spin columns were used to elute the RNA in 36µL DNase and RNase free water. Prior to elution, columns were treated with on column- DNase I. RNA was always kept on ice and subsequently stored at -80°C. Further DNase treatment of the RNA is discussed below.

If larger culture vessels or tumour samplers were used, RNA extraction was performed using conventional phase separation with phenol chloroform. 200µl of chloroform was added to the eppendorf per 1ml of Trizol reagent. Samples were vortexed for 15 seconds, incubated at room temperature for 2 minutes, and then centrifuged at 12,000 x*g* for 15 minutes at 4°C. Following centrifugation, the sample separates into three; a lower pink phenol-chloroform phase, an interphase, and a colourless upper aqueous phase. RNA is contained exclusively within the upper aqueous phase. This was carefully removed to a fresh tube prior to mixing with 500µl of isopropyl alcohol and incubating for 10 minutes at room temperature. The samples were left overnight at -80°C to allow small RNA precipitation and subsequently centrifuged at 12,000 x*g* for 10 minutes at 5°C to form a pellet. Having removed the supernatant, RNA pellets were washed in 1ml of 75% ethanol. Samples were centrifuged at 7,500 x*g* for 5 minutes at 5°C, and again supernatant was removed. Pellets were air-dried for 5 minutes to remove residual ethanol. Finally, the RNA pellets were re-suspended in an appropriate volume of DNase and RNase-free water. RNA was always kept on ice and subsequently stored at -80°C. Throughout all RNA work, appropriate RNase precautions were used, including the use of filter pipette tips and RNaseZap (Applied Biosystems).

2.4.2 Spectrophotometry

RNA concentration was determined using the NanoDrop ND-100-Spectrophotometer, (Thermo Fisher Scientific UK Ltd, UK). An optical density (OD) 260nm of 1.0 corresponds to a concentration of 40µg/mL for RNA. An estimate of nucleic acid purity was obtained by measuring the absorption at

260nm and 280nm. The ration between these two readings for pure nucleic acid should be approximately 2.0 for RNA. Significantly different readings can indicate protein or phenol contamination.

2.4.3 Agarose gel electrophoresis of RNA

RNA quality was established using agarose gel electrophoresis. 1% agarose gels were made using agarose and 1XTAE and 10µL of nucleic acid gel stain (Sigma-Aldrich). Samples were loaded with 6X gel loading solution (Thermo Fisher Scientific). Intact total RNA will produce sharp 28S and 18S rRNA bands, with the former being twice as intense as the 18S band, visible on the UV transillumінator. Partially degraded RNA shows as a smear. These bands represent 28S and 18S rRNA and indicate successful RNA preparation.

2.4.4 DNase treatment of total RNA

Effective DNase treatment of RNA was imperative due to the nature of the RNA products which were under investigation. It became clear that in column DNase I (Zymo Research) treatment during RNA precipitation was not sufficient to remove all traces of genomic DNA. In order to optimize the DNase treatment the following methods were compared; In column DNse I (Zymo Research); RQ1 DNase; TurboDNase; Quantitect Reverse Transcription Kit Genomic DNA Wipeout (Qiagen) and phenol:chloroform:isoamyl alcohol extraction.

For the purposes of DNase treatment prior to RT-qPCR, I found TurboDNase (ThermoFischer) to be the most effective but rather than using the inactivation buffer as per manufacturer's instructions, I found phenol:chloroform:isoamyl alcohol extraction to be the most robust inactivator despite additional spinning after inactivation as recommended. After extracting and quantifying RNA, 0.1 vol 10 x Turbo DNase buffer and 1ul Turbo DNase was mixed gently and incubated at 37°C for 25 minutes. The RNA was then extracted with phenol:chloroform:isoamyl alcohol and washed in 70% ethanol and subsequently resuspended in RNase-free water and analysed for downstream application.

RNA quality was confirmed using a 1% non-denaturing agarose gel electrophoresis prepared in TAE buffer to assess for contamination or degradation following DNase treatment. Bands corresponding to 28S and 18S ribosomal RNA were seen on the transilluminator at 4.5kb and 1.9kb.

2.4.5 cDNA preparation by Reverse Transcription

cDNA was synthesized from 500-1000ng of purified DNase-treated RNA by the First Strand cDNA Synthesis Kit (Sigma Aldrich) with Oligo(dT)$_{12-18}$ primers incubated in a 7900Ht Thermal Cycler (Applied Biosystems). After RT cycles, the cDNAs samples were placed in ice and then prepared for quantitative real-time PCR.

For the purpose of the RT-qPCR, Quantitect Reverse Transcription Kit (Qiagen) was used as per manufacturer's instructions. Briefly, the reaction occurred at 42°C for 5 minutes and inactivated at 95°C for 3 minutes. An RT- negative control was performed concurrently without the use of the transcriptase. In contrast to other methods, additional steps for RNA denaturation, primer annealing and RNase H digestion were not necessary.

2.4.6 Quantitative real-time PCR (qPCR)

SYBR green real time quantitative PCR was carried out according to manufacturer's instructions on an Applied Biosystems StepOnePlus Real-Time PCR. Each PCR reaction was carried out in triplicate in a Microamp fast optical 96 well reaction plate sealed using the MicroAmp Optical adhesive film.

In each 20ul reaction, 10ng of cDNA and 10ul Fast SYBR green PCR master mix (2X) were mixed with 1ul each of 10µM forward and reverse specific primers. PCR conditions were as per Fast SYBR green protocol;

Stage 1 (1cycle)	95°C	20 seconds
Stage II (40cycles)	95°C	3 seconds
	60°C	30seconds

2.4.7 Analysis of Quantitative PCR

Background threshold levels were set at the number of cycles before any Fast SYBR-Green fluorescence was detected. This threshold was set at the point were the increase in fluorescence became exponential, assuming that the cycle number at which a sample's fluorescence intersected the detection threshold was directly proportional to the amount of DNA in the sample. This is expressed as a cycle-threshold, or C_T value6+. In order to determine relative abundance of a transcript, two ubiquitously expressed genes were used as housekeeping genes, GAPDH and PUM1. Absolute comparative expression levels were analysed using Microsoft Excel using the formula 2^-(C_T target gene-C_T housekeeping gene) and, if appropriate, relative to the vehicle control.

2.5 Library preparation

In total, two separate stranded RNA libraries were prepared during the course of this work as per the manufacturer's instructions using standard Illumina sequencing primers and 6bp single indices.

RNA-Seq: MCF7 cells treated for 3,6 and 24 hours with oestradiol or vehicle
RNA extraction as described in 2.4.1
Samples: 3hE2, 6hr E2, 24hr E2, 24hr Veh (in triplicate)
Illumina TruSeq Stranded **Total** RNA LS Prep Kit (Illumina, San Diego, USA)

CRISPR: MCF7 Luc clone harbouring Poly(A) knock-in and WT control
Cells treated for 6 hours with oestradiol or vehicle
CRISPR as described below
Samples: Clone and Wild type (in triplicate)
Illumina Truseq Stranded **mRNA** LS Prep kit (Illumina, San Diego, USA)

Prior to pooling of samples, they were quantified using the Qubit® Quant-iT™ dsDNA HS Assay Kit and the Qubit® 2.0 or 3.0 Flourometer, as per the manufacturer's instructions. The quality of the DNA libraries was validated using an Agilent Technologies 2100 Bioanalyzer and an Agilent HS DNA chip.

Paired-end sequences of 100nt in length were generated using a HiSeq 2000 machine (Illumina), operated by colleagues in the MRC (Medical Research Council) facility at Imperial College, London.

2.6 Cellular Fractionation

Cell fractionation into two compartments was achieved following a protocol kindly provided by Dr Mark Kalisz and Professor Jorge Ferrer (Beta Cell Genome Regulation lab, Imperial College, London). All steps were performed on ice and using pre-cooled buffers.

MCF7 cells were plated in starvation medium in 10cm dishes for 72 hours and then media replaced with either vehicle or 10nM estradiol and cells incubated for 0, 3, 6 and 24 hours. The cells were then trypsinised, washed once in starvation medium and washed once in 1mL ice-cold PBS. The cells were centrifuged, supernatant aspirated and packed cell volume (PCV) estimated (approximately 100ul). 1X lysis buffer (5M Sodium Chloride, 1M Tris-HCl, Nonidet P-40, Sodium deoxycholate (10%), SDS (10%), ddH20) was added at 5x volume of the PCV and the cell pellet vortexed for 10 seconds before being incubated on ice for 15 minutes, allowing the cells to swell. IGEPAL CA-630 solution was then added to a final concentration of 0.6% and vortexed for 10 seconds before being centrifuged for 5 minutes at 16 000 x*g* at 4°C. The supernatant was transferred to a new tube as the cytoplasmic fraction.

The remaining crude nuclei pellet was resuspended in Nuclear extraction buffer (plus additives) without using pipette tips, but by mounting the eppendorf on a vortex mixer and agitating at high speed for 30 minutes at 4°C. The eppendorf was then centrifuged at 16 000 x*g* for 10 minutes at 4°C and the supernatant transferred to a clean chilled eppendorf as the nuclear extract. Both fractions were stored at -80°C.

Cell fractionation into three compartments (cytoplasmic, nuclear and chromatin) was not successful despite following a well regarded protocol [120].

2.7 RNA Flourescence in situ hybridization (FISH)

Cellular localization of the eRNAs of interest was investigated with both cellular fractionation techniques (see above) and RNA visualization with Stellaris® RNA FISH (Biosearch Technologies, Inc., Petaluma, CA), which enables simultaneous detection, localization and quantification of individual mRNA molecules at the cellular level in fixed samples using fluorescence microscopy.

24-48 Custom Stellaris® RNA FISH probes were designed against the eRNA transcripts (nucleotide length at the 3'end) using the Stellaris® RNA FISH Probe Designer (Biosearch Technologies, Inc., Petaluma, CA) available online at www.biosearchtech.com/stellarisdesigner. The MCF7 cells were hybridized with the eRNA Stellaris RNA FISH Probe set labeled with Quasar® 670 dye and Human MALAT1 RNA FISH probe set labeled with Quasar® 570 Dye (Biosearch Technologies, Inc.), following the manufacturer's instructions available online at www.biosearchtech.com/stellarisprotocols.

Briefly, MCF7 cells were grown on 18mm round coverglass in a 12 well plate in starvation media at 70% confluence. After 72 hours of starvation, they were treated with 10nM estradiol or vehicle. After 6 and 24 hours, the protocol was followed as per manufacturer's instructions; incubation steps were performed for the maximum time suggested.

Samples were then imaged using a widefield fluorescence microscope (Zeiss AxioObserver), courtesy of the Facility for Imaging by Light Microscopy (FILM) at Imperial College London. Single and z-stack images were obtained, deconvoluted using Huygens software (SVI) and analysed using ImageJ Fiji [121].

2.8 Candidate Knockdown using siRNA and Antisense Oligonucleotides

Enhancer Transcript knockdown was attempted using both small interfering RNA (siRNA) and Antisense oligonucleotide (ASO) technology. siRNAs and ASO (LNA GapmeRs) were designed to target the bidirectionally transcribed eRNA and the neighbouring gene *CCND1*. Appropriate commercially available negative controls were used. Sequences of successful siRNAs and ASO GapmeRs can be found in tables 2.4 and 2.5 respectively.

Several transfection methods were employed in these experiments including Hiperfect, Optimem and Lipofectamine 2000, but the final method involved reverse transfection using Lipofectamine™ RNAiMax as per the manufacturer’s instructions.

siRNA ID	**Sequence (5’->3’)**
si*CCND1* (A) s229	Commercially available *Thermofisher*
si*CCND1* (B) s201129	Commercially available *Thermofisher*
siRNA A targeting antisense *CCND1*e (s501508) *Thermofisher*	sense ACUGUGCAGUGGACCCUUAtt antisense UAAGGGUCCACUGCACAGUta
siRNA B targeting antisense *CCND1*e (s501509) *Thermofisher*	sense: AAGCCUUGUCUAUAACAUATT antisense: UAUGUUAUAGACAAGGCUUCC
si501510 targeting antisense *CCND1*e *Thermofisher*	sense GAUGCAUGCUUGUUGGAAATT antisense: UUUCCAACAAGCAUGCAUCAC
siRNA A sense targeting sense *CCND1*e (s501511) *Thermofisher*	sense: UGAUUAAACAUGAUGCUGATT antisense: UCAGCAUCAUGUUUAAUCATG
siRNA B sense targeting sense *CCND1*e (s501512) *Thermofisher*	sense: CAGCUGAAGGUGAUAAAAATT antisense UUUUUAUCACCUUCAGCUGCT
All star negative control	Commercially available Qiagen
Negative Control 1	Commercially available *Thermofisher*

Table 2.4 Sequences of successful siRNAs

LNA GapmeR ID	**Sequence (5’->3’)**
ASO A targeting sense *CCND1*e (AC12-1)	CAACAAGCATGCATCA
ASO B targeting sense *CCND1*e (AC12-7)	GGCATGAATAGTCTAT
ASO A targeting antisense *CCND1*e (A4)	TCTTGCTTCCACTTTA
ASO B targeting antisense *CCND1*e (A6)	CAGGATCCCTCATCTA
Negative Control A (supplied Exiqon)	AACACGTCTATACGC
Negative Control B (supplied Exiqon)	GCTCCCTTCAATCCAA

Table 2.5 Sequences of successful ASO GapmeRs and negative controls

100ul OPTIMEM was added to each well of a 24 well plate. 10nM of the appropriate siRNA (stock concentration of 20uM) or 10nM or 25nM ASO (stock concentration 50uM) was added. The plates were rocked for 1 minute. 1ul of RNAiMax was added to each well and the plate rocked again. The plates were left at room temperature for 20 minutes.

MCF7 cells were then added to the 24 well plate at density of 5x10^4 cells per well seeded in Antibiotic free DMEM supplemented with 10% FCS. The plate was rocked and then incubated for 48 hours at 37°C in a humidified 5% CO2 incubator.

Cells were then either trypsinized for RNA extraction and RT-qPCR or used for further downstream experiments. DNase treatment was carried out using 2µl TurboDNase followed by inactivation buffer as previously described. Knockdown experiments were conducted in triplicate.

2.9 Cell Cycle Analysis

To quantify the percentage of the cell population in different phases of the cell cycle, the Millipore Muse® Cell Analyzer was used with Muse® Cell cycle kit. This system uses miniaturized fluorescence detection and microcapillary cytometry to enable rapid quantitative measurements of the percentage of cells in the G0/G1, S and G2/M phases of the cell cycle. The Muse® system uses a premixed reagent including the nuclear DNA intercalating stain propidium iodide (PI) which discriminates cells at distinct stages of the cell cycle based on the DNA content. Resting cells in G0/G1 contain 2 copies of each chromosome but as the cell moves into S phase, they synthesize chromosomal DNA and the fluorescence intensity from PI increases until all chromosomal DNA has doubled (G2/M phase), at which point fluorescence is double and then the cells divide into two and the fluoresce falls again.

Cell cycle analysis was performed on MCF7 cells successfully transfected with siRNA or GapmeR with resultant knockdown on either eRNA or *CCND1*. The same technique was also used to analyse CRISPR edited MCF7 cells with Poly A knock-into the antisense enhancer of *CCND1*.

The Cell analyzer was used as per manufacturer's instructions, but briefly; 200ul of cells were added to each tube and centrifuged at 300 x*g* for 5 minutes and then washed once with 1 X PBS. 200ul ice cold 70% ethanol added to cells and mixed slowly and then incubated overnight at -20C. Cells were then centrifuged again at 300 x*g* for 5 minutes and washed with 1 X PBS. 200ul Muse® Cell Cycle reagent was added and incubated in the dark at room temperature for 30 mins. Each sample was then analysed in the Muse Cell analyzer.

2.10 Analysis of Cellular Proliferation

To accurately measure cell proliferation and cell viability, the WST-1 Assay Reagent (Abcam ab65473) was employed. The protocol is based on the cleavage of the tetrazolium salt WS-1 to formazan by cellular mitochondrial dehydrogenases such that their higher activity results in greater amounts of formazan dye which can be analyzed by measurement at 440nm in a microplate reader. The manufacturer's instructions were followed to analyze cell proliferation of wild type MCF7, MCF7 following knockdown with siRNAs and GapmeRs targeting *CCND1* and *CCND1* enhancer transcripts and CRISPR engineered mutations within the *CCND1* enhancer region.

2.11 Extracting RNA from Tumour Derived Xenografts

RNA was extracted from snap frozen tumour derived xenografts (a kind gift from Champions Oncology https://championsoncology.com/xenograft-tumor-models/oncology-pdx-models/) as described below: A pestle and mortar was chilled in an ice box containing dry ice. The tumour was weighed and if necessary, cut using a sterile scalpel to remove and use approximately 25mg frozen tissue. The snap frozen tumour was added to the pestle and a small volume of liquid nitrogen added until the pestle was 7/8 full. The mortar was then used to grind the frozen tumour to a fine powder and a sterile spatula used to transfer the powder to a 50ml falcon tube containing Trizol (Invitrogen, Paisley, UK) reagent (1mL trizol per 100mg tissue) on ice. The sample was vortexed vigorously for 10 seconds.

RNA was then extracted by conventional phase separation with phenol chloroform as described in 2.4.1 and RNA concentrations were determined using the Nanodrop spectrophotometer. Genomic DNA was eradicated using TurboDNase as described in 2.4.4

2.12 CRISPR/Cas9

The CRISPR/Cas9 gene editing system was used to cut the genome and insert a template for homology directed repair (HDR) 3' to publicly available CHIP-Seq data of known transcription factor binding sites at the transcriptional start site (TSS) of the two enhancer RNA transcripts of interest. The donor template was designed such that it would deliver a polyadenylation (polyA) sequence and consequently, bring about termination of transcription[122]. The intention was to prevent elongation of transcription of the eRNAs whilst not interrupting their TF binding or initiation of the transcriptional machinery.

Following identification of the target sequence, the protocol consists of several steps:

Design and produce sgRNAs and clone them into an empty plasmid with subsequent bacteria transformation

Isolate significant amounts of sgRNA containing plasmid DNA using MaxiPrep technique

Transfect the cell line of interest with the plasmid containing the sgRNA, the Cas9 plasmid and the single stranded oligonucleotide template for HDR.

Select for transfected cells

Validate the presence of cells within the cell pool harbouring the desired "knocked-in" Poly(A) sequence

Isolate individual cells harbouring the "knock-in" and clone them

2.12.1 Selecting Target Sequence

ChIP-Seq data sets from MCF7 cells were obtained from the NCBI Gene Expression Omnibus (GSE) (http://www.ncbi.nlm.nij.gov/geo/) accession number GSE43836 [76] which was used to identify the genomic location of ER and other TF binding at the enhancer sequence. The site for genomic editing with polyadenylation signal insertion was within the first 200 nucleotides just 3' to these TF binding sites.

Following DNA extraction from (initially) MCF7 and later MCF7-Luc cells as described above, the predicted genomic sequence was validated using PCR amplification followed by sanger sequencing.

DNA was extracted from 5 x 10^6 using Zymo Quick gDNA Mini Prep as per manufacturer's protocol (Zymo Research). DNA amplification was then performed using Thermofisher 2x ReddyMix PCR Master Mix with forward and reverse primers as listed in Table 2.6 as per protocol. PCR amplification proceeded using the recommended thermal cycling conditions for 35 cycles. Thermo fisher CleanSweep PCR Purification reagent was used to dephosphorylate unincorporated nucleotides and digest unused primers prior to running a 1.5% agarose gel with 10µL SYBR stain to ensure a product of the expected length (approximately 400bp for sense strand and 272bp for antisense). The PCR product was then sent to GeneWhiz for Sanger sequencing along with a forward or reverse primer and the final sequence checked for deviation from the predicted sequence.

OLIGO NAME	**SEQUENCE (5' - 3')**
VH *CCND1*E 1	Forward AGCAGTTTCACATCAATATA
	Reverse TAACAGAATACCTGAAACTA
VH *CCND1*E 2	Forward GCTTGCTCTTCCTGGACACT
	Reverse CTGGCACAACTGCTGCAGTT
VH *CCND1*E 3	Forward TCAACGTAGCAGGATGGAGG
	Reverse TAGAGGAAGCTTCTGGCACC
VH SNP C12 1	Forward TTCCTTCGCTCCCTCTCATC
	Reverse GGTGGGACTTTGTGACACCA
VH SNP C12 2	Forward TTACATAGAAGGGGGTGAGC
	Reverse GAGAGTCACCCCTCCTTCTG
VH *CCND1*E NEG 1	Forward CTTGGTGCTGTCCTCAAGAT
	Reverse CCACCCCATCTGGAGATCTT
VH *CCND1*E NEG 2	Forward AATGGCTTGGTGCTGTCCTC
	Reverse AAAGCTCAGTGCTGGTGTCC
VH *CCND1*E NEG 3	Forward ACCAGAAGCCAAGCAGATGT
	Reverse ATGGGAAGCGAGGGAGATTT
C12 NEG ROUND 1	Forward ATGCTTCTTGCTTCACAGAGG
	Reverse GGGATTTCAGTTCAACAGGAGG

Table2.6 Primers to check genomic sequence at point of knock-in pre CRISPR

2.12.2 sgRNA design

Following validation of the genomic sequence, sgRNAs were designed using the appropriate sequence. The freely available webtool crispr.mit.edu[123] (https://*cis*r.mit.edu/) was used to scan the target DNA sequence for possible CRISPR guides (20nucleotides followed by the PAM sequence NGG), and also for possible off target matches throughout the genome.

The highest 8-10 ranking sgRNAs, based on faithfulness to on–target activity and predicted off-target scores were selected. The same DNA sequence was run through the DESKGEN™ CRISPR webtool (http://deskgen.com), and 4-8 of the top ranking gRNAs matching the guides designed by the crispr.mit tool were chosen for in vitro validation.

Using NCBI Blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [124], these sgRNA sequences (5'-NNNNN NNBBB BBBBB BBBBB NGG-3', where B represents the actual bases of the target genomic location) were checked to ensure that no other location existed in the human genome with the same sequence. If the sgRNA did not start with the base G then one was added as the U6 promoter prefers G.

The 20-21bp sgRNA sequence was incorporated into two 60mer oligonucleotides using the 40bp scaffold sequences shown in table 2.7 (Invitrogen at 100uM stock). The forward and reverse strands pair with each other and result in two overhangs that can be ligated as seen in Figure 2.1 which shows sgRNA 107. All sgRNA 60mer ssODN sequences are found in Table 2.8.

Scaffold	Sequence
SA3984CRISPRf	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC
SA3985CRISPRr	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC

Table 2.7 gRNA scaffold sequences (from Invitrogen)

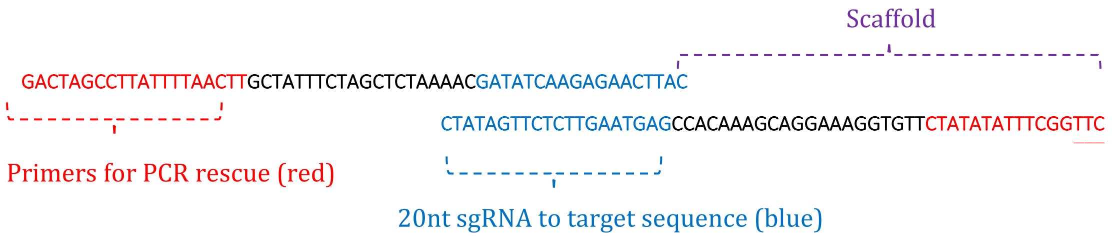

Scaffold
GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGATATCAAGAGAACTTAC
CTATAGTTCTCTTGAATGAGCCACAAAGCAGGAAAGGTGTTCTATATATTTCGGTTC
Primers for PCR rescue (red)
20nt sgRNA to target sequence (blue)

Figure 2.1 An example of the 60mer sgRNA oligonucleotide pair sg107

ID	**sgRNA 107**
Target sequence	GAGTAAGTTCTCTTGATATC
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGTAAGTTCTCTTGATATC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGATATCAAGAGAACTTACTC
ID	**sgRNA 123**
Target sequence	GCTGAGAAGTCCAGATCGAG
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTGAGAAGTCCAGATCGAG
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACCTCGATCTGGACTTCTCAGC
ID	**sgRNA 127**
Target sequence	GGCGAGGGCCTTCTTGCTGC
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGAGGGCCTTCTTGCTGC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGCAGCAAGAAGGCCCTCGCC

Table 2.8 CRISPR gRNAs targeting sense strand

ID	**sgRNA 11**
Target sequence	GTGAGACTCAGTGTCTAGTCC
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGACTCAGTGTCTAGTCC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGGACTAGACACTGAGTCTCAC
ID	**sgRNA 12**
Target sequence	GTCTCCATGTGGGGCCACGGC
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCTCCATGTGGGGCCACGGC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGCCGTGGCCCCACATGGAGAC
ID	**sgRNA 13**
Target sequence	GAGAGGTTGTGCTACTTGCCT
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCAGGCAAGTAGCACAACCTCTC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGAGAGGTTGTGCTACTTGCCT
ID	**sgRNA 14**
Target sequence	GCAATGCTAGAGCCATGCTGT
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAATGCTAGAGCCATGCTGT
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACACAGCATGGCTCTAGCATTGC

Table 2.9 CRISPR gRNAs targeting antisense strand

ID	**sgRNA ER:A**
Target sequence	GGCGGAGTCATGCCAGCTCA
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGACTCAGTGTCTAGTCC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGGACTAGACACTGAGTCTCAC
ID	**sgRNA ER:B**
Target sequence	GTCAGGATGACTGAGAGCTC
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCTCCATGTGGGGCCACGGC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGCCGTGGCCCCACATGGAGAC
ID	**sgRNA ER:C**
Target sequence	GCTCTCAGTCATCCTGACCT
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCAGGCAAGTAGCACAACCTCTC
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACGAGAGGTTGTGCTACTTGCCT
ID	**sgRNA ER:D**
Target sequence	GTGGAGACACCTGGAAGCTC
Oligo (forward)	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAATGCTAGAGCCATGCTGT
Oligo (reverse)	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACACAGCATGGCTCTAGCATTGC

Table 2.10 CRISPR gRNAs targeting ER binding site

2.12.3 sgRNA template production

The following protocol was kindly provided by Professor Buluwela (Department of Cancer and Surgery, Imperial College London) and is an optimized protocol of the recognised PCR approach for sgRNA template production and transformation into an empty plasmid vector. In brief, this involves annealing the two sgRNA oligonucleotide primers, extending them using NEB Phusion® DNA Polymerase, followed by PCR amplification of the 100bp products. The 100bp DNA fragments are then incorporated into a linearized gRNA cloning vector (Addgene p41824) using Gibson assembly.

2.12.4 Anneal two sgRNA primers

For each designed sgRNA, the following were added to an eppendorf and left to cool slowly, floating in a beaker of boiled water overnight:

2ul Forward primer

2ul Reverse primer

8ul 5x Phusion® HF Buffer (NEB)

28ul H20

2.12.5 Extension and end repair of DNA product

After brief centrifugation of the annealed product, 20ul was added to the following mix in a new eppendorf:

4ul Phusion® High Fidelity Buffer (deletes any incorrectly attached nucleic acids)

0.8ul 10mM dNTP mix

0.4ul Phusion® DNA Polymerase

14.8ul H20

The mix was gently pipetted and placed on a heat block at 72°C for 10 minutes. It was briefly centrifuged and kept on ice.

2.12.6 PCR rescue of double stranded DNA template

5μl of the two 100μM primers used for PCR amplification (SA3984CRISPRf and SA3985CRISPRr) were mixed with 40μL H_20. For each sgRNA, 5ul of this mix was added to the following in a new micro-eppendorf:

5ul of 5X Phusion® HF Buffer

1ul of extended and end repaired DNA product from previous step

1ul of 10mM dNTP mix

0.5ul Phusion® DNA Polymerase

37.5ul H_20

This was placed in the 7900Ht Thermal Cycler (Applied Biosystems) and the following program run:

Stage 1 (1cycle)	98°C	30 seconds
Stage II (25cycles)	98°C	15 seconds
	50°C	15seconds
	72°C	15seconds
Stage III (1cycle)	72°C	2 minutes
Hold	4°C	Hold

2.12.7 Validation of 100bp DNA product

5ul of the product from PCR repair was run on a 1.5% TBE agarose gel with the MassRuler™ DNA Ladder, low range, (Fermentas) to ensure the presence of 100bp products.

2.12.8 Purification of PCR Products

The 100bp DNA products of PCR repair underwent purification using the Qiagen PCR Purification kit, as per manufacturer's instructions, and subsequently kept on ice. DNA concentrations were determined using the Nanodrop spectrophotometer.

2.12.9 Cloning of sgRNA into a linearized empty vector

The empty gRNA expression vector plasmid 41824 (a gift from George Church[125] (Addgene)) is a kanamycin resistant vector which was linearized using the restriction enzyme AflII in the following reaction which was left overnight at 37°C and then 65°C for 10 minutes before being kept on ice.

10ul CutSmart® buffer (NEB)
1ul 100X BSA
1.9ul of Plasmid p41824
1ul AflII restriction enzyme (NEB)
87.1ul H20

2ul of the plasmid was run on 1.5% agarose gel with MassRuler™ DNA Ladder, low range, (Fermentas) to ensure the presence of the linearized product. 6.37ng of the 100bp DNA purified product was mixed with 50ng of the linearized vector DNA and made up to 10ul total volume. 10ul of Gibson Assembly® master mix was added and pipetted gently. The sample was incubated at 50°C for 60 minutes and subsequently stored at -20°C.

2.12.10 Transformation of bacterial competent cells with plasmid DNA

2ul of the Gibson assembled vector containing the sgRNA was transformed into NEB 5-alpha Competent E.Coli (NEB C2987) cells according to the manufacturer's instructions. Briefly, the chemically competent cells were thawed on ice in a chilled eppendorf. 2ul of Gibson assembled DNA was added to 50ul of cells, gently pipetted and placed on ice for 30 minutes. The sample was then heat shocked at 42C in a water bath for 30 seconds and transferred back to ice for 2 minutes prior to being incubated at 37°C for 60 minutes with vigorous shaking in 950ul of room temperature S.O.C. medium (Invitrogen, Paisley UK). 100ul of the bacterial culture was then spread on pre-warmed LB-agar kanamycin plates (and ampicillin for control) and incubated upside down at 37°C overnight.

Up to 15 individual colonies were picked from the LB-agar plates using a sterile 200ul pipette tip and the tip added to 1.5mL LB broth containing 50ug/ml kanamycin (LB-kanamycin) in a 7ml bijou tube (Sterilin, UK) and left in a shaking incubator overnight.

2.12.11 Isolation of Plasmid DNA

Small quantities of plasmid DNA (up to 20ug) were isolated using the Qiaprep Spin Miniprep Kit (Qiagen Ltd, Crawley, UK). Briefly, 1ml of the bacterial culture (the remaining 500ul of bacterial culture was stored at 4°C) from each bijou tube was moved to an eppendorf and pelleted at 5000rpm for 5 minutes prior to following the protocol according to manufacturer's instructions. The final product was eluted in 50ul of elution buffer provided with the kit and kept on ice. DNA concentrations were determined using the Nanodrop spectrophotometer. These samples were sent for Sanger sequencing as described below to validate their successful incorporation of the sgRNA.

Following validation of the presence of the sgRNA within the plasmid (as described below), at least two of the bacterial cultures containing each sgRNA underwent the maxi-prep procedure to isolate a large quantity of DNA for further work. First, 2ml of LB-kanamcyin broth was added to the remaining 500ul bacterial culture that had been stored at 4°C, and incubated at 37°C for 5 hours with vigorous shaking.

200ul of the bacterial culture was set aside in Copan Diagnostics CRYOBANK™ Bead system tubes (Fisher Scientific) at -80° for long term storage. The remaining bacterial culture was added to a 1litre conical flask containing 200mL of LB-kanamycin broth and left in a shaking incubator at 37°C overnight.

The PureLink™ Expi Endotoxin-Free Maxi Plasmid Purification (Thermo Fischer) was then used to isolate the plasmid DNA from the bacterial culture as per manufacturer's instructions. Bacterial pellets were centrifuged for 5 minutes at 15°C using the Sorval SLA-1500 rotor (DuPont, Herts, UK) and subsequent centrifugation took place in the Sorval SS34 rotor (DuPont, Herts, UK). The final DNA was precipitated in 70% ethanol and resuspended in 400ul of endotoxin-free TE buffer provided. DNA concentrations were determined using the Nanodrop spectrophotometer, and stored at 4°C.

These samples were again sent for Sanger sequencing as described below to ensure that the sgRNA sequence was still present in the plasmid.

2.13 Screening for successful transformation by Sanger sequencing

The above isolated DNA was sequenced using the Sanger DNA sequencing service provided by Genewiz, Takeley, UK. The facility uses ABI 3730xl DNA

Analyzers for capillary electrophoresis and fluorescent dye terminator detection. 5uM TOPO4 forward primer (5' – CTTTATGCTTCCGGCTCGTA – 3') was provided along with at least 40ng of the DNA product.

The sequence chromatograms returned by the service were analysed using SnapGene software (GSL Biotech LLC, Chicago USA) and the chromatograms checked to ensure incorporation of the sgRNA sequence into the empty vector.

2.14 Cas9 selection

As the gRNA design was successful in identifying appropriate guides with neighbouring PAM sequences NGG, S. Pyogenes Cas9 (SpCas9) was considered an appropriate nuclease for this CRISPR work. Initially, wild type spCas9 (Cas9 (Addgene PX 260)) was used (a kind gift from Professor Buluwela), but the very low "knock-in" rate required a selection method, and both Green Fluorescent Protein (GFP) (pSpCas9(BB)-2A-GFP (Addgene PX458)) and puromycin resistant (pSpCas9(BB)-2A-Puro (Addgene PX259)) Cas9 plasmids were subsequently used.

2.15 Poly(A) oligonucleotide template design

For each predicted genomic cut site based on the gRNAs designed above, a single stranded oligonucleotide (ssODN) template was designed to provide a template for HDR and subsequent insertion of the polyadenylation signal. Oligonucleotide templates were designed for both the sense and antisense strand for each cut site in order to reduce the chance of NHEJ on the opposite strand. The central sequence motif, AAUAAA, and the flanking auxiliary elements were previously described by Proudfoot[122]; shown as the central 49 nucleotides highlighted in red.

Poly(A) Oligonucleotide *sense strand* for use with sgRNA107

AAGGGGGAGGGCCACACACTTTTAAGCAACCAGATaataaaatatctttattttcattacatctgt gtgttggtttttgtgtgATCAAGAGAACTTACTCACTATCTTGAGGACAGCA

Poly(A) Oligonucleotide *antisense strand* for use with sgRNA 107

TGCTGTCCTCAAGATAGTGAGTAAGTTCTCTTGATcacacaaaaaccaacacacagatgta atgaaaataaagatattttattATCTGGTTGCTTAAAAGTGTGTGGCCCTCCCCCTT

Homologous arms of 35 nucleotides were used to flank the central sequence. These homologous arms were designed around the predicted cut site at 3nt 3' to the PAM site of the appropriate sgRNA. Figure 2.2 shows the ssODN designed as the donor template for sgRNA 107. The 119 nucleotide donor oligonucleotides were manufactured by Sigma.

GGAAGCAAGAGAGAG*AAGGGGGAGGGCCACACACTTTTAAGCAACCAgat*★*atcaagagaacttact*
*cACTATCTTGAGGACAGCA*CCAAGCCATTCATGAGGAGTCCAC

Figure 2.2 Sequence showing ssODN donor homologous arms designed around the intended cut site for co-transfection with sgRNA 107

Predicted cut site marked with ★

Target sequence to match sgRNA 107 in blue

ssODN homologous arms are underlined

PAM site in red

ID	Sequence (5’-3’)
107_AS	TGCTGTCCTCAAGATAGTGAGTAAGTTCTCTTGATcacacaaaaaaccaacacacagatgtaatgaaaataaagatattttattATCTGGTTGCTTAAAAGTGTGTGGCCCTCCCCCTT
107_S	AAGGGGGAGGGCCACACACTTTTAAGCAACCAGATaataaaatatctttattttcattacatctgtgtgttggtttttgtgtgATCAAGAGAACTTACTCACTATCTTGAGGACAGCA
123_AS	TATCTTACAGTTCTGTAGGCTGAGAAGTCCAGATCcacacaaaaaaccaacacacagatgtaatgaaaataaagatattttattGAGGGGTTGCATCTGGCGAGGGCCTTCTTGCTGCT
123_S	AGCAGCAAGAAGGCCCTCGCCAGATGCAACCCCTCaataaaatatctttattttcattacatctgtgtgttggtttttgtgtgGATCTGGACTTCTCAGCCTACAGAACTGTAAGATA
127_AS	GATCGAGGGGTTGCATCTGGCGAGGGCCTTCTTGCcacacaaaaaaccaacacacagatgtaatgaaaataaagatattttattTGCTGGGGACTCTCTGCCAAGTCTCAAGGCAGTGC
127_S	GCACTGCCTTGAGACTTGGCAGAGAGTCCCCAGCAaataaaatatctttattttcattacatctgtgtgttggtttttgtgtgGCAAGAAGGCCCTCGCCAGATGCAACCCCTCGATC

Table 2.11 Sense strand ssODN donor sequences

ID	Sequence (5’-3’)
12_AS	TGTGTGGACGTTTCCCTGTCTCCATGTGGGGCCACcacacaaaaaaccaacacacagatgtaatgaaaataaagatattttattGGCCGGCAGGTCCAGCTCTCTTGGGCACATTCAAT
12_S	TAGCATTGAATGTGCCCAAGAGAGCTGGACCTGCCaataaaatatctttattttcattacatctgtgtgttggtttttgtgtgGTGGCCCCACATGGAGACAGGGAAACGTCCACACAG
14_AS	GCTCTCTTGGGCACATTGCAATGCTAGAGCCATGCcacacaaaaaaccaacacacagatgtaatgaaaataaagatattttattTGTGGGGTTGTGAGACTCAGTGTCTAGTCCTGGCT
14_S	AGCCAGGACTAGACACTGAGTCTCACAACCCCACAaataaaatatctttattttcattacatctgtgtgttggtttttgtgtgGCATGGCTCTAGCATTGCAATGTGCCCAAGAGAGC

Table 2.12 Antisense strand ssODN donor sequences

2.16 Identifying the best transfection method

Given the suspected low “knock-in” efficiency of this CRISPR experiment, optimizing the transfection rate was imperative. Several transfection methods and reagents were compared. GFP pmax GFP-vector or pSpCas9(BB)-2A-GFP plasmid was used to visualize and estimate the transfection efficiency.

2.16.1 Nucleofection

Nucleofection using the 4D-Nucleofector™ System (Lonza,) and the Amaxa® Cell Line Nucleofector Kit V (Lonza) was performed. The protocol had previously been optimized for MCF7-Luc cells by Professor Buluwela’s (Department of Cancer and Surgery, Imperial College) group, and was followed for MCF7 cells using program EN-130 and 0.4ug of pmax GFP-vector (provided) in 20ul Nuclovette™ Strips.

2 x10^6 MCF7 cells were required per nucleofection. The adherent MCF7 cells in culture were trypsinised and centrifuged at 90 *xg* for 10 minutes at room temperature before being washed in 1ml PBS. The cells were counted and 2x10^6 cells were placed in an eppendorf. These were centrifuged at 90 *xg* for 10 minutes and the supernatant removed. The cells were resuspended in 100ul of room temperature 4D-Nucleofector™ Solution provided with the kit.

DNA in the form of Cas9 plasmid, sgRNA plasmid, ssOligo template and pmax GFP-vector was added to the cell suspension and the sample transferred into a Nucleocuvette™ Vessel. In order to optimize the nucleofection, different concentrations and ratios of Cas9:sgRNA:SSODN were trialed;

1:1:1   5:5:1   10:10:1   50:50:1   100:100:1   1:1:2

Three ssOligo template amounts were compared in an attempt to optimize the nucleofection but minimize toxicity: 10ng, 100ng and 1000ng. I also transfected both ssODNs for each double strand break so that both strands would have access to a donor template and hence minimize the chance of NHEJ on one strand. The maximum volume of total DNA to be added was less than 5ul.

The 4D-Nucleofector™ System was run as per instructions and immediately afterwards to cells gently resuspended in pre-warmed DMEM using the supplied pipettes. The cells suspension was then plated in 5ml of antibiotic free DMEM in one well of a 6 well plate and incubated under usual conditions.

2.16.2 Transfection Reagents

Lipofectamine® 2000. Lipofectamine® 3000, Lipofectamine® LTX with PLUS™ reagent (Invitrogen, Paisley, UK) and GeneJuice® (Merck) transfection reagents were all used according to manufacturer's instructions. In order to establish the best transfection reagent for MCF7-Luc cells, each was used to transfect 1000ng pSpCas9(BB)-2A-GFP plasmid DNA and 1000ng sgRNA plasmid DNA. Cells were subsequently visualized using an epifluorescence microscope at 24, 48 and 72 hours post transfection.

2.16.3 Final transfection method

Transient transfection with the plasmid DNA and ssOligo template was ultimately conducted using GeneJuice® transfection reagent. Briefly, 3.5×10^5 MCF7-Luc cells were plated in 3ml complete growth media in a 6 well plate and incubated overnight in normal growth conditions, such that they were 50% confluent at the time of transfection. For each transfection, 6ul of GeneJuice® was added drop-wise to 100ul serum free Opti-MEM reagent in a sterile tube and vortexed before incubating at room temperature for 5 minutes.

For each transfection, 1000ng puro-Cas9 (pSpCas9(BB)-2A-Puro (Addgene PX259) plasmid and 1000ng sgRNA plasmid plus/minus 100ng ssoligo template was added to each GeneJuice®/OptiMEM tube. The mixture was incubated at room temperature for 20 minutes. The entire volume was then added drop-wise to the cells in complete growth medium and the cells incubated in normal growth conditions for 24 hours.

2.17 Cell selection

After 24 hours, puromycin was added to the 3ml full growth media and transfection mixture to a final concentration of 1mg/ul puromycin. The puromycin was left for 48 hours, after which, all of the media was removed, the cells washed twice with warm PBS, and the 3ml full growth media replaced. The cells that survived puromycin exposure were assumed to have been effectively transfected and as such were incubated until the 6 well plates reached 60-80% confluency. Cells were harvested after 48 hours to limit the chance of stable integration of Cas9 plasmid.

2.18 Validating the genome edit

The puromycin resistant transfected cells were trypsinised when they reached 60-80% confluency and 50% of the cells were re-plated in a 24 well dish in full growth medium. The remaining 50% were taken for genomic DNA preparation.

2.18.1 Genomic DNA preparation

Cells were first pelleted in eppendorfs at 300 *xg* for 5 minutes. Genomic DNA was extracted from whole cell pellets using the column-based PureLink™ Genomic DNA Mini kit (Thermo Fisher). MCF7 and MCF7 Luc cells cultured in a 6 well plate and transfected and selected as described above (plasmids containing gRNA and Cas9 and poly(A) oligonucleotides) underwent the DNA extraction protocol as per the manufacturer's instructions. Briefly, cell pellets were resuspended in 200ul PBS and 20ul RNase R and Proteinase K added and the mix incubated at room temperature for 2 minutes. 200ul of PureLink™ genomic lysis/binding buffer was added and vortexed and then incubated at 55°C for 10 minutes to promote protein digestion. 200ul of ethanol was added and vortexed and the 640ul samples added to the PureLink™ spin columns. Binding of the DNA to the columns occurs during centrifuging. The columns were then washed in 500ul of two provided wash buffers and finally eluted using the kit provided elution buffer into a new eppendorf. The purified DNA was stored at -20C.

2.18.2 PCR amplification of edited genomic DNA

10ul of the prepared genomic DNA underwent "first round" PCR using primers flanking the expected knock-in region and 2X ReddyMix PCR Master Mix (Thermo Scientific), in the following reaction as per manufacturer's instructions.

12.5ul 2X ReddyMix PCR Master Mix
1.25ul 10uM Forward primer
1.25ul 10uM Reverse primer
10ul template DNA

This was placed in the 7900Ht Thermal Cycler (Applied Biosystems) and the following program run:

Stage 1 (1cycle)	95°C	2 minutes
Stage II (**40**cycles)	95°C	25 seconds
	55°C	35seconds
	72°C	65seconds
Stage III (1cycle)	72°C	5 minutes
Hold	4°C	Hold

<u>Sense strand knock-in</u>

Round 1 Forward Primer	GAGAAGTCCAGATCGAGGGG
Round 1 Reverse Primer	TTGGCTCACAGTTCTGCAG
Round 2 Forward Primer	CATCTGCTTGGCTTCTGGTG
Round 2 Reverse Primer	CCAACACACAGATGTAATGAA

<u>Antisense strand knock-in</u>

Round 1 Forward Primer	GAGAGGTTGTGCTACTTGCC
Round 1 Reverse Primer	ATGCTTCTTGCTTCACAGAGG
Round 2 Forward Primer	GAGAGGTTGTGCTACTTGCC
Round 2 Reverse Primer	CCAACACACAGATGTAATGAA

2.18.3 Validating sgRNA cutting

5ul of the PCR product was then purified using CleanSweep™ PCR purification reagent as per manufacturer's instructions and sent for Sanger sequencing as described using the same forward and/or reverse primers used for the PCR. The sequence chromatograms returned by the service were analysed using SnapGene software (GSL Biotech LLC, Chicago USA) and the chromatograms checked to identify the presence of genome editing at the expected cut site, 3bp from the PAM site.

Using the chromatograms, identifying the most efficient sgRNA was possible using the free webtool for easy quantitative assessment of genome editing provided by The University of Netherlands (https://tide.nki.nl) [126].

2.18.4 Validating presence of "knock-in"

2ul of the PCR product from above was used for further "second round" PCR amplification using a reverse primer specific to the ssOligo template sequence as shown below. Hence amplification of this genomic region would only occur in the presence of the knocked in sequence.

Forward strand 5'-3'

AGTGGGAGCAGGCATCACATGGTTAAAGTGGAAGCAAGAGAGAGAAGGGGGAGGG
CCACACACTTTTAAGCAACCAGATaataaaatatctttattttcattacatctgtgtgttggtttttgtgt
gATCAAGAGAACT

12.5ul 2X ReddyMix PCR Master Mix
1.25ul 10uM Forward primer
1.25ul 10uM Reverse primer
2ul template DNA from "first round PCR"
8ul H20

This was placed in the 7900Ht Thermal Cycler (Applied Biosystems) and the following program run:

Stage 1 (1cycle)	95°C	2 minutes	
Stage II (**35**cycles)	95°C	25 seconds	
	55°C	35seconds	(optimized)
	72°C	65seconds	
Stage III (1cycle)	72°C	5 minutes	
Hold	4°C	Hold	

2ul of the “second round” PCR product was added to 8ul H20 and run in a 1.5% TBE agarose gel as previously described, along with NEB Quick-Load® 100bp DNA Ladder (NEB).

2.19 Clonal expansion

Once presence of the knock-in within the heterogeneous cell pool had been proven, individual cells harbouring the knock-in were sought through single cell cloning.

The cells growing following puromycin selection were washed once in PBS and trypsinised and resuspended in full growth media. 5000 of these puromycin resistant MCF7-Luc cells were then plated in a 15cm dish in full growth media and incubated in normal conditions for 10 to 14 days, changing the media every 3-4 days.

At 10-14 days, a single cell colony was visible to the naked eye. The media was removed from the 15cm dish and the cells washed gently in PBS and the PBS removed. 4.8mm Sterile Cloning Discs (Sigma-Aldrich) were immersed in trypsin and carefully placed over visible colonies using tweezers. Between ten and twenty colonies were picked for expansion.

The plate with trypsinised cloning discs was placed in the incubator for 5 minutes to enable the colony to detach. After 5 minutes, the cloning disc was carefully removed with the use of tweezers and each disc placed in an individual well within a 24 well plate, containing 500ul of full growth medium. The plate was then incubated under normal growth conditions for 3 days, at which time the media was replaced. The cells were incubated until they reached 60-80% confluency within the well, with the median time being approximately 10 days.

2.20 Identifying clones harbouring the "knocked-in" sequence

Genomic DNA was extracted from approximately 30% of the cells growing in each well in the 24 well plates in the same way as described in 2.18.1 genomic DNA preparation. The remaining 70% of the clonal population was left in culture and expanded whilst further investigation of their genome editing was undertaken.

Following DNA preparation, "first round" PCR was performed in the same way as described in 2.18.2 PCR amplification of edited genomic DNA.
Clones were identified through three methods:

- "first round" and "second round" PCR as described above followed by agarose gel looking for DNA product harbouring knock-in (as described in 2.18.4 validating presence of knock-in);
- an agarose gel run after "first round" PCR looking for a product 49bp larger than the wild type;
- DNA used for RT-qPCR at concentration of 10ng/ul DNA and run as previously described

2.21 Elution of DNA band from agarose gel for Sanger Sequencing

In order to sequence the band visible on the 1.5% agarose gel indicating the presence of the polyadenylation knock-in, I used the QIAquick Gel Extraction kit (Qiagen). This enables removal of nucleotides, enzymes, salts, agarose, ethidium bromide and other impurities from samples, allowing up to 80% recovery of DNA. Using a silica membrane assembly to bind DNA in high salt buffer and then elute in low salt buffer, the kit enables recovery of pure DNA suitable for sequencing.

First, a 1.5% agarose gel (as described previously) in 1XTAE buffer. MassRuler™ DNA Ladder, low range, (Fermentas) was run along with the samples at 80V for 60 minutes. Using a lightbox in the dark room, clean scalpels were used to cut out the DNA band at approximately 49bp larger than the wild type. Each appropriate band was individually weighed in a colourless tube and 3X its volume of Buffer QG provided in the kit was added to the gel. Manufacturer's instructions were followed and the DNA eluted from the QIAquick columns in 30ul of the provided Buffer EB. The DNA was stored at -20C or sent for Sanger Sequencing.

2.22 Dual-Luciferase Reporter Assay

CRISPR knock-in clones and wild type cells were starved and plated in 24 well plates and treated with E2 or vehicle for 6 hours as previously described. Cells were washed three times with PBS. Luciferase activity was measured using Dual-Glo® luciferase assay system (Promega) as per manufacturer's instructions. 100μl of passive lysis buffer was added to each well and the plate shaken for 30 minutes at room temperature. 50μl of the lysate was transferred to each well of a 96 well OptiPlate (PerkinElmer) with 50 μl of luciferase. The plate was shaken for 10 minutes in darkness and analysed using a fluorescence spectrophotometer. Stimulation of the oestrogen responsive element by oestrogen receptor bound to E2 causes an increase in luciferase activity.

Chapter 3 Results: Identification of estrogen regulated non-coding RNA in MCF7 breast cancer cell line

3.1 Global transcriptome changes in response to estrogen treatment

In 2009, this group reported a subset of microRNAs (miRNAs) modulated by ERα and identified a role for these miRNAs as part of a negative autoregulatory feedback loop in the cellular response to estrogen in ERα breast cancer cells[3]. With the subsequent identification of a vast repertoire of other non-coding genes in the genome, whose functions were still to be discovered in 2013, this project first set to look for the subset of long non-coding RNAs regulated by estrogen in ERα expressing breast cancer cells.

It is known that ERα binds to more than 10 000 sites across the human genome, and in doing so exerts at least two functions. The first to promote recruitment of co-regulators involved in the post translational modification of histones or other transcription factors, and the second, to regulate the activity of RNA Polymerase II and its associated machinery[76]. [60] [127] .

To investigate the effects of estrogen on the global transcriptome of ERα human breast cancer cells, I treated estrogen deprived ERα positive MCF7 cells with 17β-estradiol (E2) for 0, 3, 6 and 24 hours. The expression of well documented estrogen-regulated transcripts were first quantified with RT-qPCR to ensure that estrogen treatment had been successful (Figure 3.1). GREB1 is a known early response gene regulated by ERα both in vitro and in vivo and is a key regulator of estrogen induced breast cancer growth. GREB1 is induced in a dose dependent manner in ERα positive breast cancer cell lines[128]. pS2 (or TFF1 (Trefoil factor 1)) is a stable secretory protein of unknown function usually expressed in the gastrointestinal mucosa but also found in human tumours. It is a direct estrogen inducible transcript in MCF7 cells and is rapidly up-regulated by estrogen exposure [129]. GAPDH is a recognised housekeeping gene in MCF7 cells and for many years has been used as a control in RT-qPCR analyses. Whilst studies[130] have recently suggested that GAPDH expression is associated with breast cancer cell proliferation and aggressiveness of tumours, given its ubiquitous use I chose to use it as a control, and used PUM1 as an internal control for the GAPDH. I found no statistically significant difference between analyses using PUM1 and GAPDH and I subsequently used both.

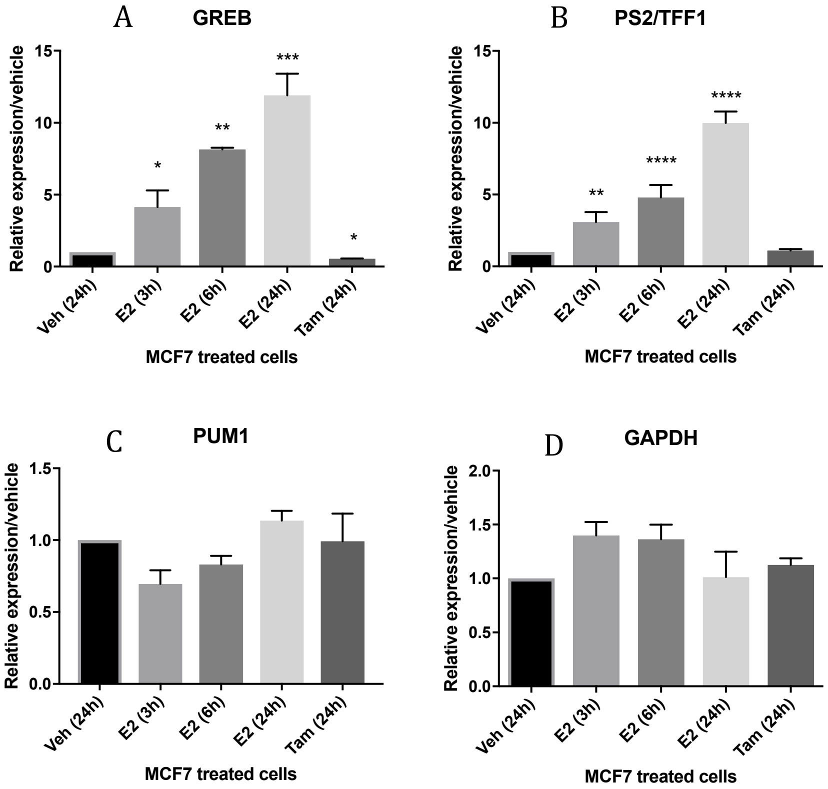


Figure 3.1 The estrogen response in MCF7 cells.

MCF7 cells were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours. Cells were then plated onto 6 well plates (day 0) with either vehicle (ethanol), 100μM tamoxifen or 10nM 17β estradiol for 24 hours. RT-qPCR was performed from RNA prepared from the estrogen treated MCF7 cells at 0, 3, 6 and 24 hours and the vehicle and tamoxifen treated at 24 hours. Data for GREB (A), PS2/TFF (B) and PUM1 (C) were normalised to GAPDH levels and the expression level of each gene is shown relative to expression of vehicle at 24 hours. GAPDH (D) was normalised to PUM1. The mean and standard error (s.e.m.) of three biological replicates are shown *p<0.1, **p<0.01, ***p<0.001.

3.2 Differential Gene Expression using RNA-Seq

In order to identify the entire transcriptome modulated by ERα, I used RNA-sequencing (RNA-seq), the gold standard method for evaluation of transcriptomic changes. RNA extracted from MCF7 cells exposed to E2 was subjected to ribosomal depletion and RNA sequencing using Illumina TruSeq Stranded Total RNA LS Prep kit as described in 2.5 and sequencing performed with the Hiseq 2000 instrument (Illumina). Dr Castellano performed the following RNA-Seq analysis on samples collected at 3, 6 and 24 hours and compared with the vehicle (flowchart of analysis shown in Figure 3.2).

3.2.1 RNA Sequencing data analysis

The raw sequencing reads produced by the Illumina sequencer first underwent a quality assurance check using the fastqc program (www.bioinformatics.babraham.ac.uk/projects/fastqc/). The sequencing reads from the FASTQ files were aligned to the human genome (hg38) using the program TopHat (http://tophat.cbcb.umd.edu/), such that the amount of a transcript could be determined via the number of reads produced (mapping summary is shown in Table 3.1). TopHat was used to align the triplicate FASTQ files from the four estrogen treated time points (0h, 3h, 6h and 24 hours), resulting in a BAM file. Next, using Cufflinks, reads were normalised according to their length and quantity identified in numerous runs, resulting in a fragment per kilobase of exon per million fragments mapped (FPKM) score which is directly proportional to their abundance. To identify only genes significantly modulated by E2 we used an FPKM of 0.5 and excluded all transcripts with an FPKM below this in order to exclude those with very low expression. The program CuffDiff then calculated the differential expression of transcripts over the four time points and tested their statistical significance with CuffMerge. In this way, the transcripts were grouped into biologically meaningful groups, such as those with the same transcriptional start site (TSS), thus enabling the identification of differentially regulated transcripts which included only those with a false discovery rate (FDR) or q value ≤ 0.05 (p value, adjusted for multiple testing with the Benjamini-Hochberg correction).

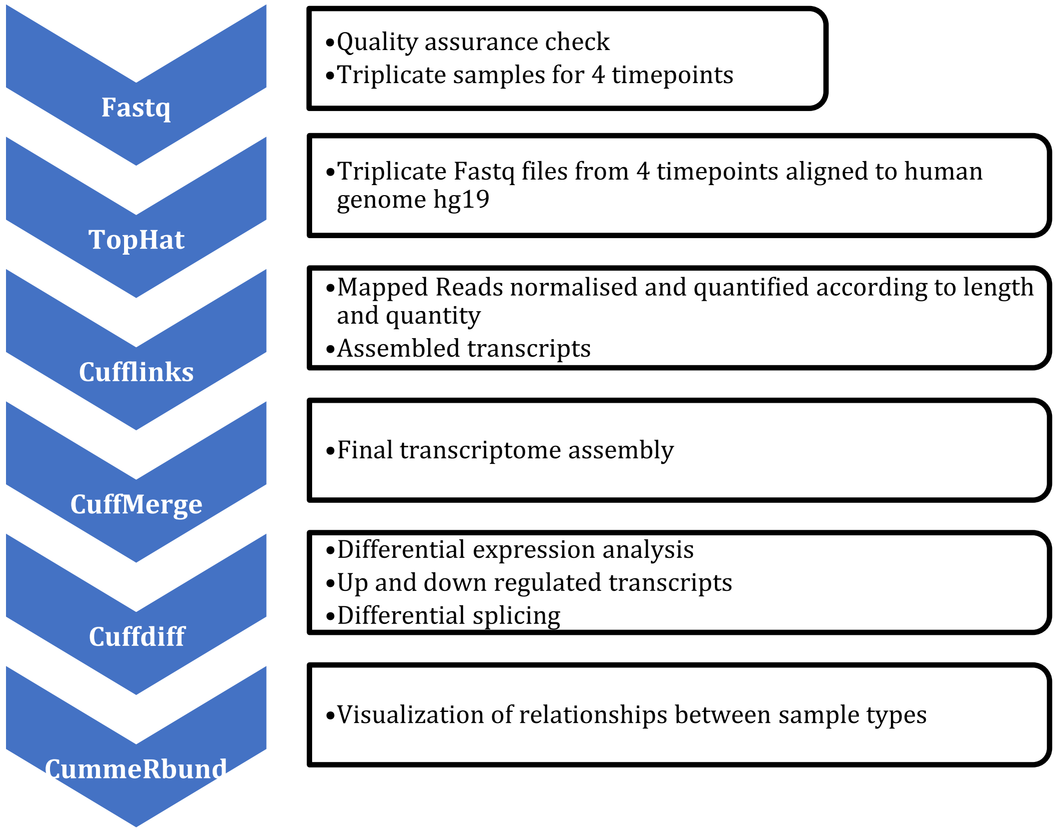

Fastq
•Quality assurance check
•Triplicate samples for 4 timepoints
TopHat
•Triplicate Fastq files from 4 timepoints aligned to human genome hg19
Cufflinks
•Mapped Reads normalised and quantified according to length and quantity
•Assembled transcripts
CuffMerge
•Final transcriptome assembly
Cuffdiff
•Differential expression analysis
•Up and down regulated transcripts
•Differential splicing
CummeRbund
•Visualization of relationships between sample types

Figure 3.2 Flowchart showing use of the Cufflinks suite to perform RNA-Seq analysis.

Time point	Number of Mapped reads	Total number of mapped reads per Time point
0h_1	45,936,498	
0h_2	33,502,304	
0h_3	35,720,908	115,159,710
3h_1	37,232,030	
3h_2	42,615,426	
3h_3	31,015,528	110,862,984
6h_1	33,368,540	
6h_2	37,674,250	
6h_3	34,707,762	105,750,552
24h_1	34,246,340	
24h_2	41,038,964	
24h_3	37,615,438	112,900,742

Table 3.1 Summary of the RNA-Seq data.
Number of reads mapped to human genome hg38 from MCF7 cells treated over 24hour time period with esotrogen and vehicle (3 biological replicates).

As it has been reported that some transcripts thought to be lncRNAs may in fact code for short peptides [131], non-annotated transcripts regulated by E2 were assessed for their coding potential using the Coding Potential Calculator (CPC) [132], which assesses coding potential based on six sequence features and PhyloCSF score[133], which analyses a multispecies nucleotide sequence alignment to determine its coding potential based on statistical comparison of phylogenetic codon models.

From a total of 78 051 estrogen regulated transcripts initially generated by CuffNorm, 60 552 had an FPKM greater than 0.5 and of these, 4889 an FDR less than 0.05. 3594 were known RefSeq coding transcripts and 1147 were transcripts greater than 200 base pairs in length and previously unannotated (Figure 3.3). 134 previously annotated long non-coding RNAs such as MALAT1 and NEAT1 were identified as being significantly regulated by estrogen.

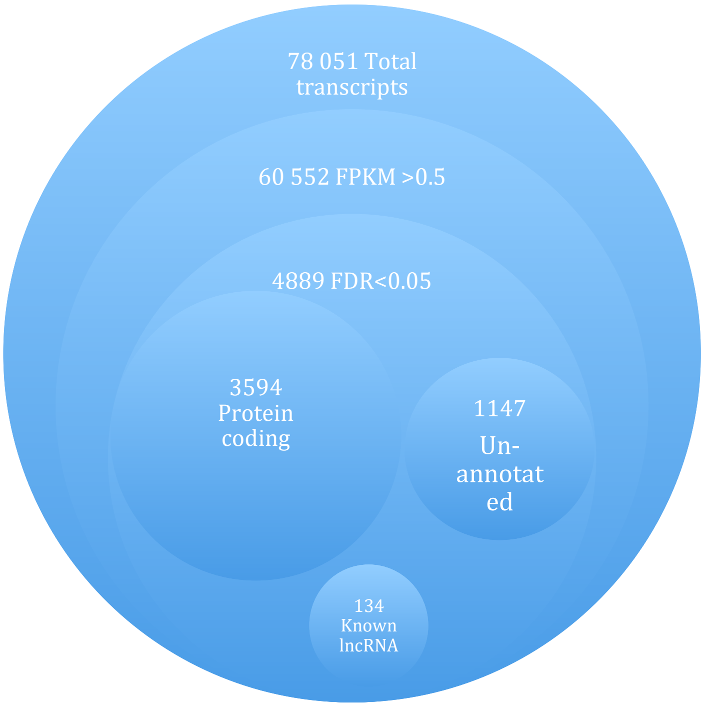


Figure 3.3 Stacked Venn Diagram of RNA Seq analysis.
Diagram represents the number of transcripts at each point of initial analysis following RNA-Seq of MCF7 cells treated with E2 over a 24 hour time course.

Initial analysis of the RNA-Seq data used the visualization package CummeRbund to show relationships between the sample types. Both a multidimensional scaling (MDS) plot and a cluster dendrogram (Figure 3.4 A and B respectively) show gene expression differences of the four conditions and are in keeping with the expected outcome that those samples exposed to estrogen (3h, 6h and 24h) are more similar that those treated with vehicle (0h), and that the gene expression within the technical replicates are generally similar. Gene expression at 24 hours varied the most between the replicate samples.

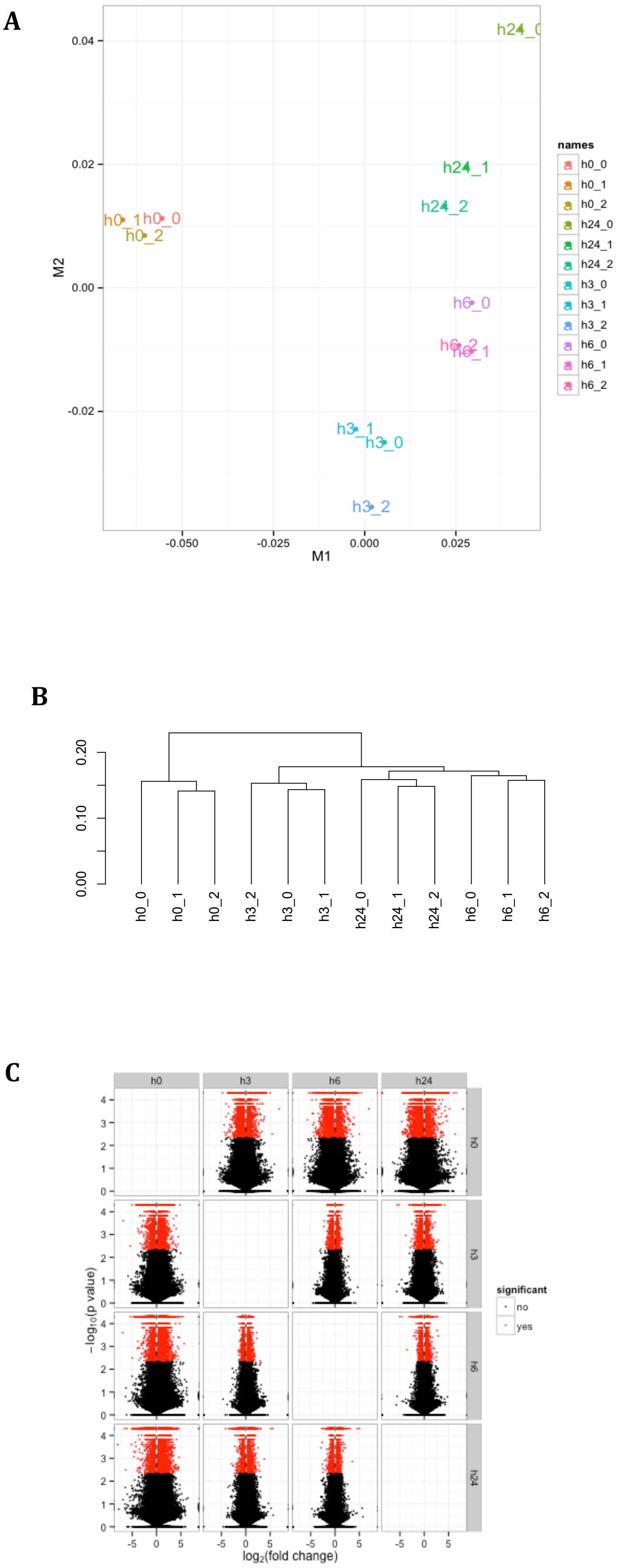


Figure 3.4 Global analysis of cufflinks data.

MDS and clustering plots showing relationship between RNA-Seq samples

A) MDS plot showing similarity in gene expressions in each of the three technical replicates in the four groups

B) Cluster Dendrogram showing that vehicle treated samples share fewer similarities to those treated with E2, and technical replicates share most similarities.

C) Volcano plot matrix generated from the gene expression data showed that a number of genes were differentially expressed between samples and both up and down regulated over the 24 hours of E2 exposure. Points in red indicate a statistical significance q value of < 0.05 (y axis, -log10 of p value). 0h=vehicle, 3h=3 hours, 6h=6 hours, 24h=24 hours E2.

3.2.2 Identification of differentially expressed coding genes between 3hours estrogen exposure and vehicle

Using the criteria q<0.05 and log2 fold change ≤-0.5 or ≥ 0.5, 2495 coding genes were identified as significantly differentially regulated. 1317 were up regulated and 1178 down regulated in MCF7 cells treated with estrogen over 3 hours compared to vehicle. The log2 fold changes ranged from 1.4 to 32.5. Gene Ontology (GO) term enrichment and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway analysis were applied for the identification of key genes and pathways regulated by oestrogen at 3 hours. To determine the functions of differentially expressed genes, all were mapped to terms in the GO database and compared to the whole transcriptome background. GO terms with corrected P value less than 0.05 were considered significantly enriched. As expected, these results showed that synthesis of proteins involved in cellular proliferation processes such as cell cycling, DNA replication and mitosis were all significantly upregulated following exposure to estrogen when compared with vehicle. GO enrichment analysis of the most regulated biological processes are shown (Figure 3.5 A and B respectively).

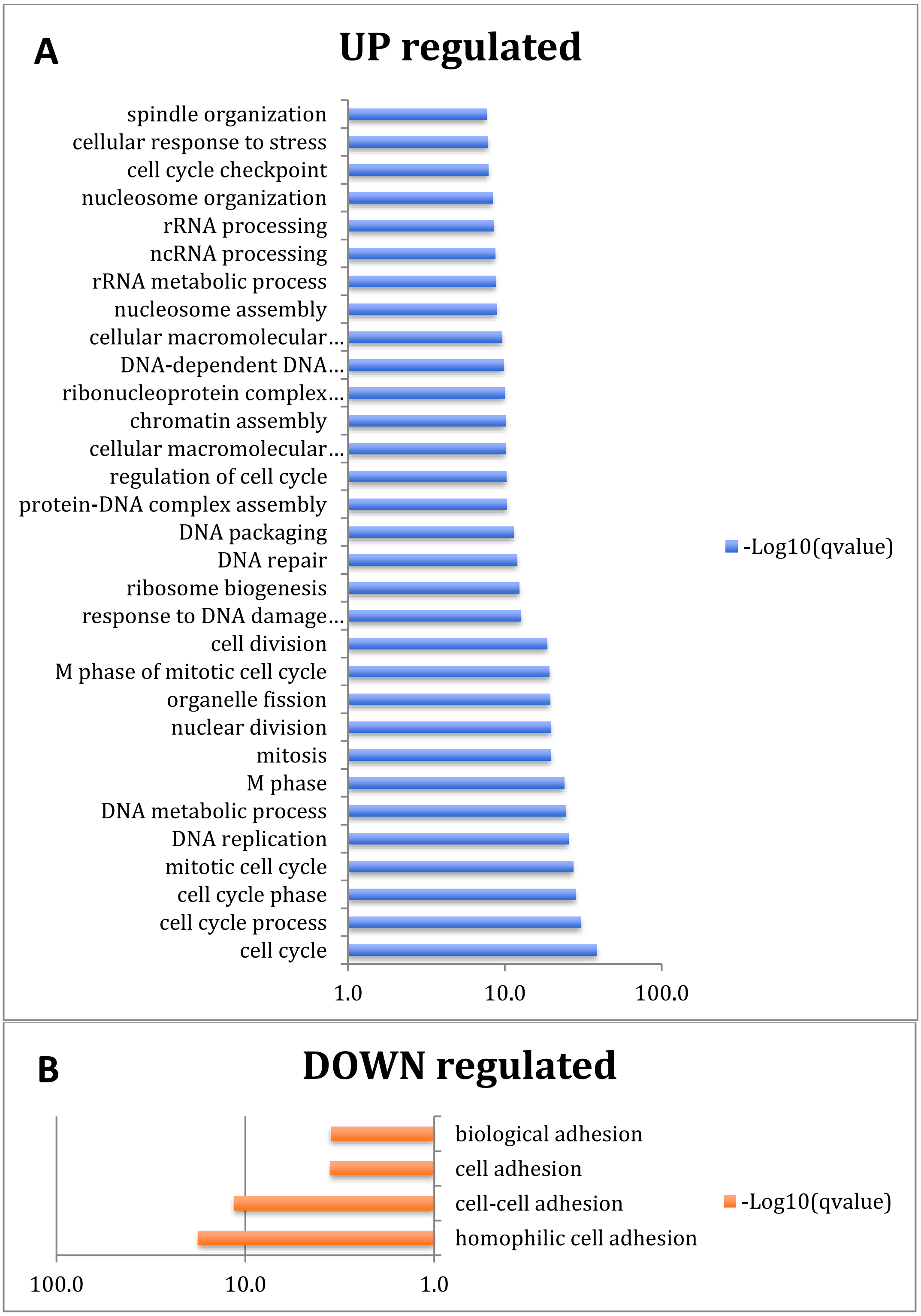


Figure 3.5 Gene Ontology enrichment analysis of differentially expressed biological processes.

Bar plot showing enrichment Q values of 20 biological processes upregulated (A) and 4 down regulated (B) in MCF7 cells at 3 hours estrogen exposure compared to vehicle. Up and down regulated genes were defined based on CuffDiff FDR cut off of 0.05.

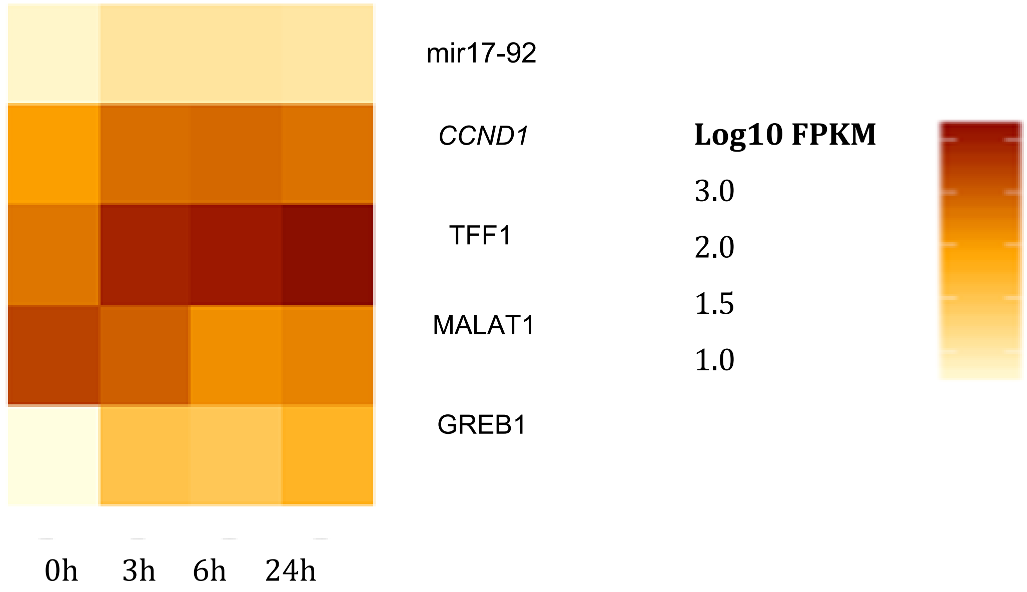


Figure 3.6 Heatmap generated from RNA-Seq analysis.
Known estrogen regulated genes in estrogen treated MCF7 cells over 24 hours, where h0 is vehicle treatment.

3.3 Validation of differentially expressed genes by quantitative PCR

In vitro validation of the RNA-seq data was performed using Reverse Transcription-quantitative PCR (RT-qPCR) of at least three biological replicates. Figure 3.6 shows the RNA-Seq generated heat map of five known estrogen regulated genes and lncRNAs over the 24 hour time course, and Figure 3.7 the validation by RT-qPCR results of those same genes. The RT-qPCR data shows a gradual increment in GREB and TFF1 expression over 6 hours, mirroring that seen in the heatmap, as well as the peak of *CCND1* expression at 6 hours. Rapid induction of mir-17-92, an estrogen regulated pre miR described by this group, and down regulation of MALAT1, a well described lncRNA, are both clearly seen in the RNA-Seq generated heat map and in vitro validation.

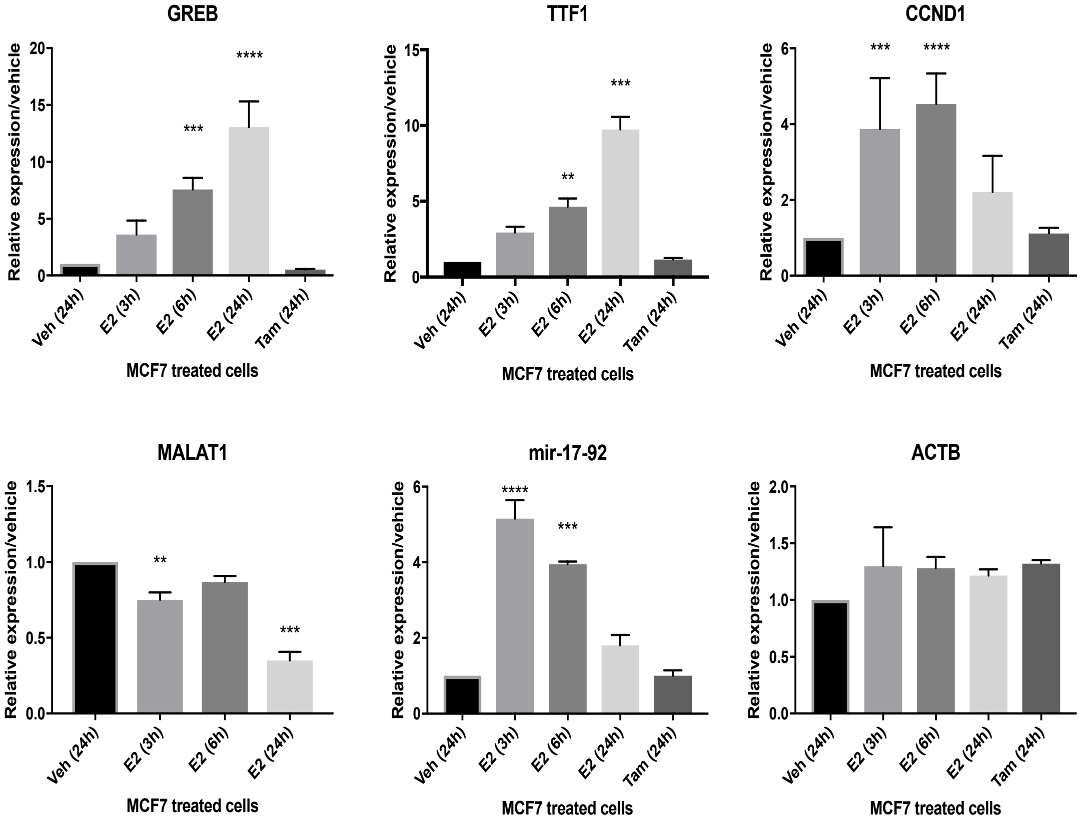


Figure 3.7 RT-qPCR of E2 regulated genes and non-coding transcripts. MCF7 cells were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours. Cells were then plated onto 6 well plates (day 0) with either vehicle (ethanol), tamoxifen or 17β estradiol for 24 hours. RT-qPCR was performed from RNA prepared from the treated MCF7 cells at 0, 3, 6 and 24 hours and the vehicle treated at 24 hours. Data were normalised to GAPDH and expressed relative to vehicle. The mean and standard error (s.e.m.) of three biological replicates are shown *p<0.1, **p<0.01, ***p<0.001.

3.4 Non-coding RNA

3.4.1 Identifying non-coding transcripts of interest

Having used RNA-Seq to identify the transcriptome regulated by estrogen, and having concluded that the RNA-Seq data was reproducible with well-recognised estrogen regulated genes and lncRNAs, we compiled a list of previously un-annotated non-coding RNA transcripts longer than 200 nucleotides with significant up or down regulation in the presence of estrogen for further analysis. These lncRNA all had a log2 fold change of greater than 1 or less than -1 and an FPKM of greater than 1, indicating a significant change in expression in the presence of estrogen. We then looked further at the 12 candidates for evolutionary sequence conservation (although as expected for lncRNA the candidates did not show high interspecies conservation, with conservation only to chick or fish) and performed further in -silico analysis to see if cellular location, splicing and stability data was available through publicly available data sets.

Two of the 12 potential candidates were discarded as they were felt to have protein coding potential based on their PhyloCSF Score[133] and one of the transcripts was so short that I was unable to design a primer that could adequately and reliably identify it. Attempts to validate the remaining nine candidates with RT-qPCR proved problematic, despite the design of many primers. It is possible that the primers were suboptimal, or that the actual number of transcripts within the cell were very low, despite the log2 fold change, hence making identification difficult. Optimising the DNase treatment to ensure complete eradication of DNase during the RNA precipitation process did little to help with the reproducibility of the RT-qPCR results for many of the transcripts.

However through the course of the validation attempts I identified several transcripts of interest with reproducible expression over the 24 hour estrogen timecourse, two of which were the subject of the rest of this work. It was at this time in 2013 that the newly identified ncRNA arising from enhancers (eRNAs) were becoming of interest, and as shown below, it appeared that these two bi-directionally transcribed transcripts bore some of the hallmarks of eRNAs.

3.4.2 Enhancer RNA identification

Enhancers are genomic elements that contain specific recognition sequences for TF binding which regulate transcription of distantly located genes and are involved with enhancer-promoter looping. Genome-wide studies have shown that non-coding RNA are transcribed from these regions in a signal dependent manner and enhancers enriched with certain histone marks and involved in looping are more likely to be actively transcribed[129].

I used publicly available data sets of ChIP-Seq, GRO-Seq[60] and histone modifications from estrogen treated MCF7 cells to look for actively transcribed enhancers in my RNA-Seq data. Using my RNA-Seq data as well as ChIP-seq data obtained from public datasets for the proteins Mediator, p300 and histone modifications H3K4me1 and H3K27ac (all markers of active enhancers) in MCF7 I was able to map 103 ncRNA arising from enhancer regions in response to estrogen treatment. Figure 3.8 shows the bi-directionally transcribed eRNAs that became the focus of the rest of this work arising from 11q13 in the MCF7 genome.

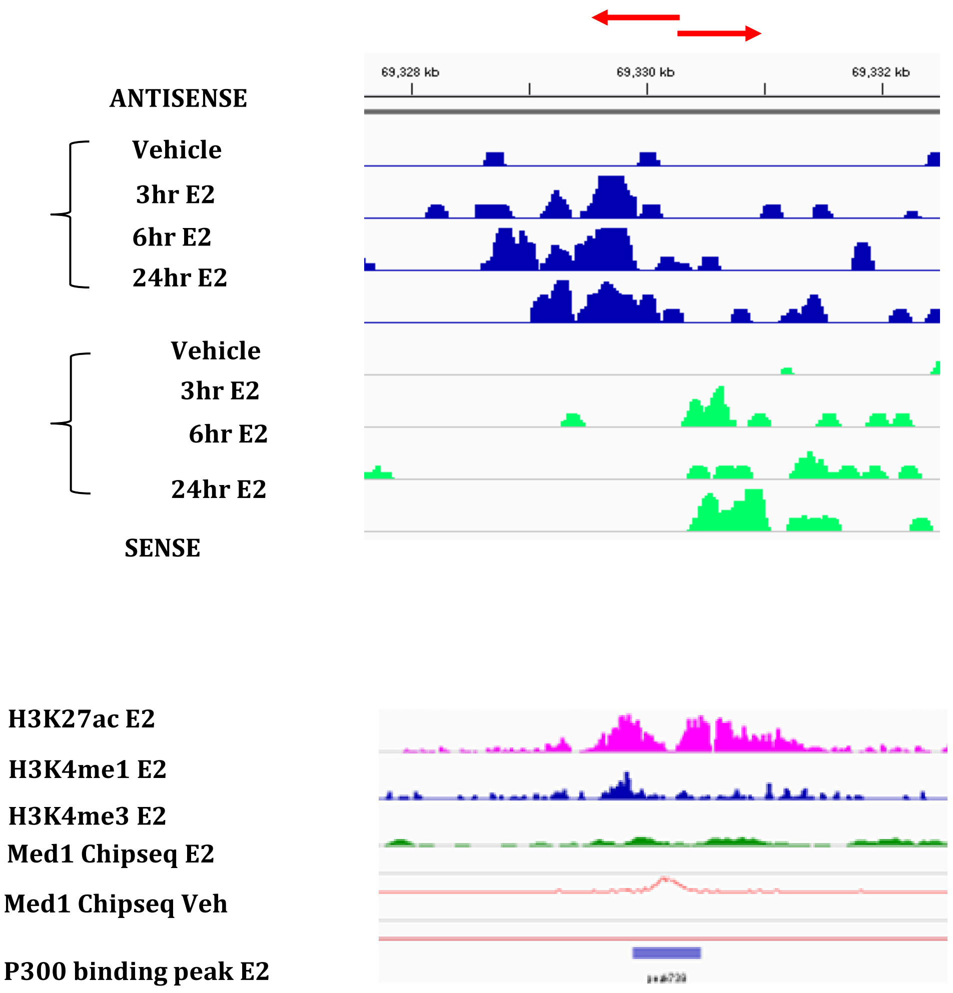


Figure 3.8 Transcription at enhancer region over 24hours following E2.
IGV 2.3 [134] view of 5kb around the TSS of bidirectional transcription of the eRNAs of interest at 11q13 (antisense in blue and sense in green) and shows increased transcription over the 24 hours, with negligible transcription occurring with vehicle treatment.
Tracks show the flanking peaks of H3K27ac at the start of bidirectional transcription (pink), as well as a peak of H3K4me1 (blue) and lack of H3K4me3 (dark green) with E2. Red lines show Med 1 binding at the TSS in presence of E2 (and not occurring in the presence of vehicle) along with P300 (blue box indicating binding peak). These features are in keeping with the recognised features of active enhancers.

3.5 Splicing of Enhancer RNAs

Long non-coding RNAs in general are known to have far fewer exons than coding mRNAs and the majority of enhancer RNAs (eRNAs) are thought to be unspliced and mono-exonic. My MCF7 cell RNA-Seq data suggests that indeed, whilst the majority of estrogen regulated eRNAs are mono-exonic (Figure 3.9 A), there are numerous transcripts of multiple exons and existing in multiple isoforms, with some evidence of alternative splicing (Figure 3.9 B).

The heatmap generated from the RNA-Seq data of spliced and unspliced eRNA arising from enhancer regions (Figure 3.10) shows that there is both up and down regulation of these transcripts over the 24 hours of estrogen exposure. Splicing indicates evolution to a more stable transcript, suggesting that the transcripts arising from enhancer regions play a physiological role and are not just “noise” resulting from the transcription of the enhancer region.

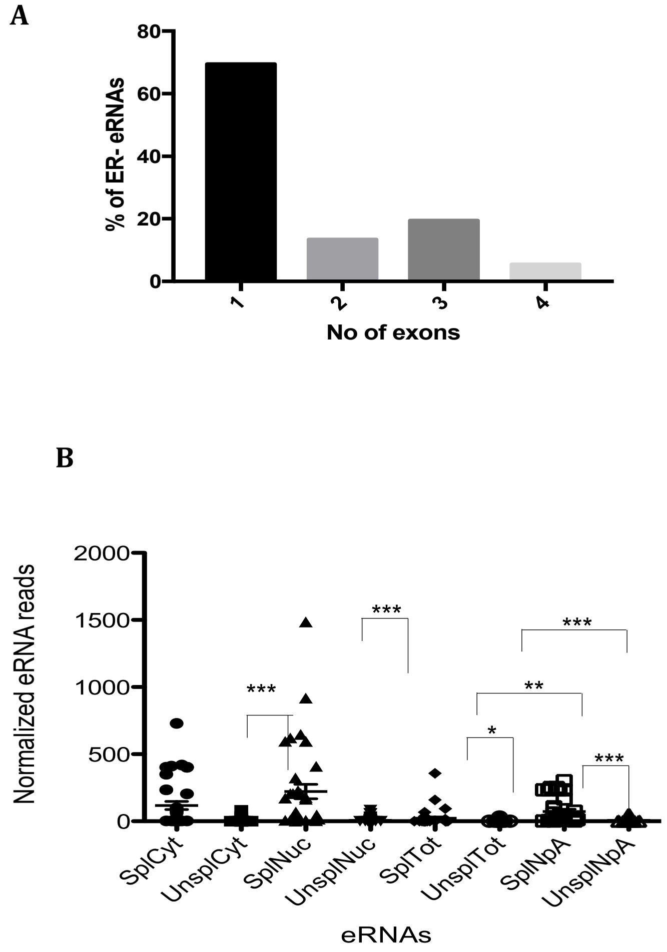


Figure 3.9 RNA-Seq analysis of transcript splicing.

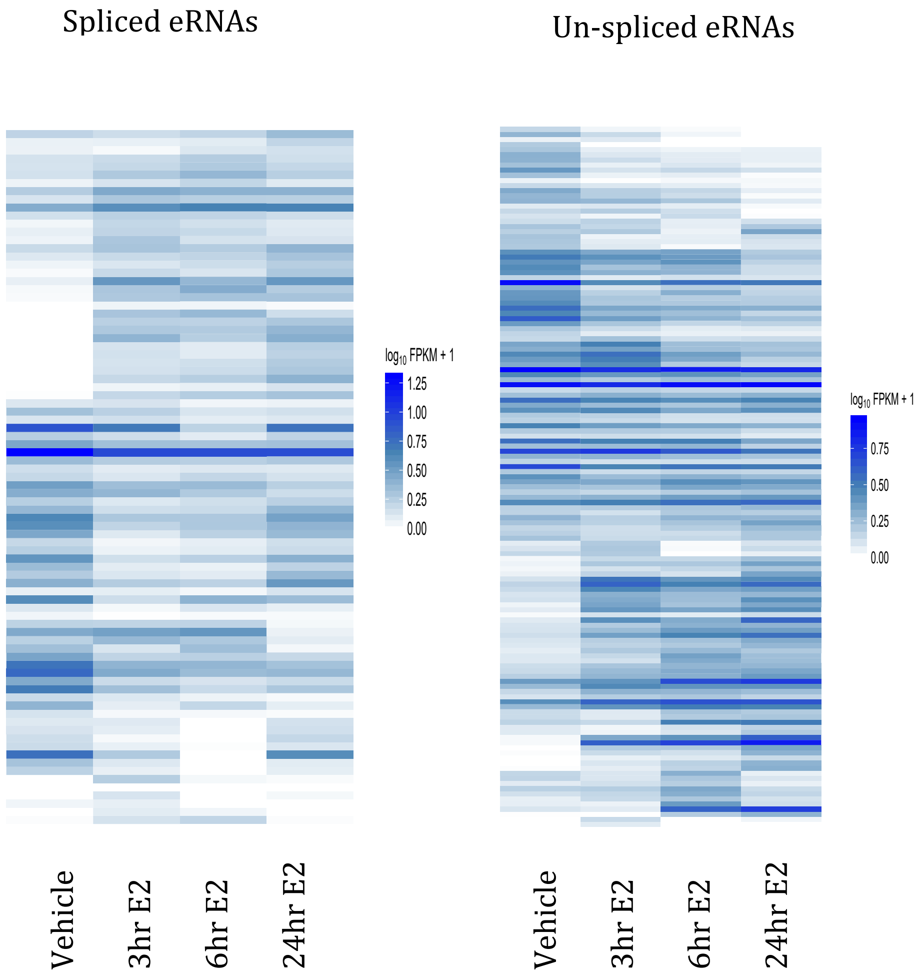
A) Global analysis of RNA-Seq data showed that 65% of eRNAs identified at 6 hours of E2 treatment are mono-exonic.

B) Bioinformatical analysis has identified spliced eRNA transcripts differentially expressed in the nucleus and the cytoplasm at 6 hours E2 treatment. * $p<0.05$, ** $p<0.01$, ***$p<0.0001$

splcyt=spliced in cytoplasm, unsplcyt = unspliced in cytoplasm, splnuc= spliced in nucleus, unspnuc= unspliced in nucleus, spltot = spliced total, unspltot=unspliced total, splnpa = spliced in the nucleus and polyadenylated, unsplnpa = unspliced in nucleus and polyadenylated

Figure 3.10 Hierarchical clustering showing normalised expression of spliced and unspliced eRNAs in MCF7 with estrogen treatment.

This shows the presence of both spliced and unspliced enhancer RNAs regulated by estrogen over a 24hour estrogen treatment. Both types of transcripts can be up or down regulated in the presence of estrogen, with many showing differential expression within 3 hours.



Chapter 4 Results: Identification of estrogen regulated RNA transcribed from the enhancer of CCND1

4.1 Bi-directionally transcribed eRNAs from 11q13 in MCF7

The bi-directionally transcribed transcripts arising from the enhancer region 150kb upstream (5’) of *CCND1* became the focus of my research. The start of transcription lies approximately 150kb from the TSS of the gene *CCND1*. *CCND1* is known to be E2 regulated and is overexpressed in more than 50% of breast cancers, amplified in 15-20% and deleted in 5%. It is known to play a pivotal role in the regulation of the cell cycle and more recently it has also been shown to play a role in the DNA damage repair pathway, possibly by localising to damaged chromosomes and recruiting the protein RAD51, a key enzyme in homologous recombination. *CCND1*’s function within the cell is discussed further in chapter 1.

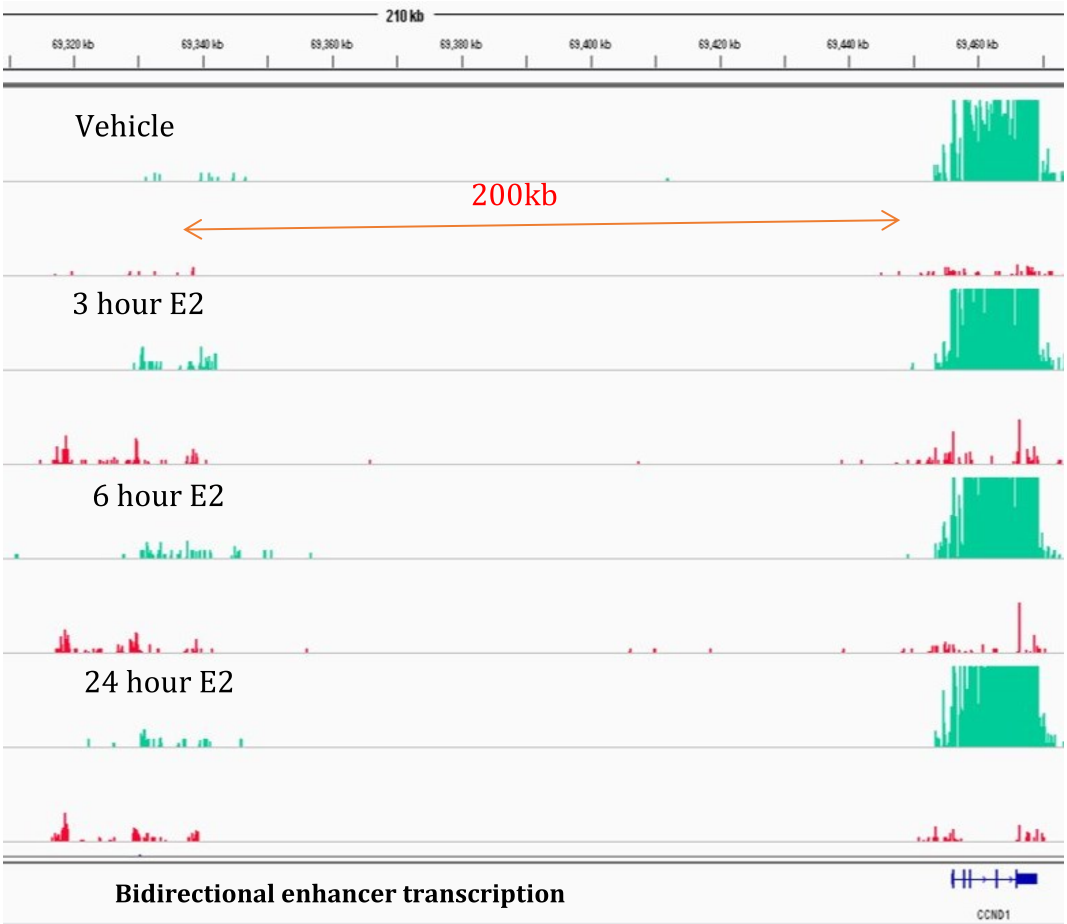


Figure 4.1 Proximity of eRNA bidirectional transcription to neighbouring gene *CCND1* around Ch11:696330 (GRCh38/hg38)
IGV 2.3[134] View of RNA-Seq data; Sense (green) and antisense (red) strands from the RNA-Seq data generated in MCF7 cells treated with E2 over a 24 hour time course are overlaid. *CCND1* is transcribed from the sense strand less than 200kb upstream from the site of bidirectional transcription of *CCND1*e(sense) and *CCND1*e(antisense).

This site of bidirectional transcription at 11q13 is also recognised as a susceptibility locus for breast cancer in human mammary epithelial cells and the single nucleotide polymorphism (SNP) rs614367 is found at location Chromosome 11:69328764 (GRCh38) in MCF7 cells. This particular SNP, with a variant allele of C to T, is reported to have an increased association with breast cancer and its location is found within the first exon of the sense eRNA (*CCND1*e (sense)) [135].

Analysis of publicly available GRO-Seq, CHIP-Seq and CHIA-PET data supported the hypothesis that the bi-directionally transcribed ncRNAs arose from an enhancer region, and indeed in 2016 (after I had commenced this work) Korkmaz [118] identified the same genomic location in MCF7 as an enhancer of *CCND1*. I named the two transcripts *CCND1*e(sense) and *CCND1*e(antisense), after the accepted nomenclature for enhancer RNA and to describe their direction of transcription.

4.1.1 Coding potential of the bi-directionally transcribed eRNA

Determination of the coding potential of both these transcripts was conducted by my colleague Dr Paul Cathcart using both PhyloCSF [136] and CPAT, with *CCND1* as a positive control. PhyloCSF showed no peak greater than 0 for either sense or antisense transcript, and the calculated CPAT score for both was below the optimum human coding probability of 0.364 (sense coding probability of 0.030895 and antisense 0.010763589). His assessments deemed that the coding potential of these eRNA transcripts was low with no evolutionary evidence that they are protein coding and he performed further in vitro ribosome profiling experiments which excluded their ability to code for peptides.

4.1.2 RT-qPCR analysis of the enhancer RNA transcripts arising from 11q3 in ER positive and negative breast cancer cell lines

Having identified transcripts of interest from the RNA-Seq analysis of the estrogen treated time course of MCF7 cells, I sought to validate their estrogen regulation with RT-qPCR. Figure 4.2 shows estrogen regulation of both the sense and antisense transcripts arising from the enhancer region, as well as regulation of both *GREB1* and *CCND1* in ER positive breast cancer cell lines MCF7 (A) and T47D (B). Induction of ER responsive *GREB1* was not seen in the ER negative cell line MDA-231 (C) and as I predicted, the enhancer RNAs *CCND1*e(sense) and *CCND1*e(antisense) were similarly not induced by estrogen treatment. Interestingly, the RT-qPCR data suggests that the estrogen regulated coding gene *GREB1* is downregulated by estrogen in the ERα negative cell line MDA-231.

Although MCF7 and T47D cell lines both express the estrogen receptor and are considered to be estrogen sensitive, relative expression of the ER, as well as the progesterone receptor (PR) and other nuclear receptors, do differ. Relative expression of *CCND1* and *CCND1e*(antisense) is not statistically significantly in the T47D cell line (Figure 4.2 (B)) which expresses more androgen receptor (AR) and PR than MCF7, with a lower ratio of ER:AR in the former. It is likely that these other steroid nuclear receptors, as well as other genetic factors such as HER2 positivity and p53 mutations will influence the estrogen response on cellular epigenetics, accounting for the differences seen in expression response of the enhancer RNA.

A

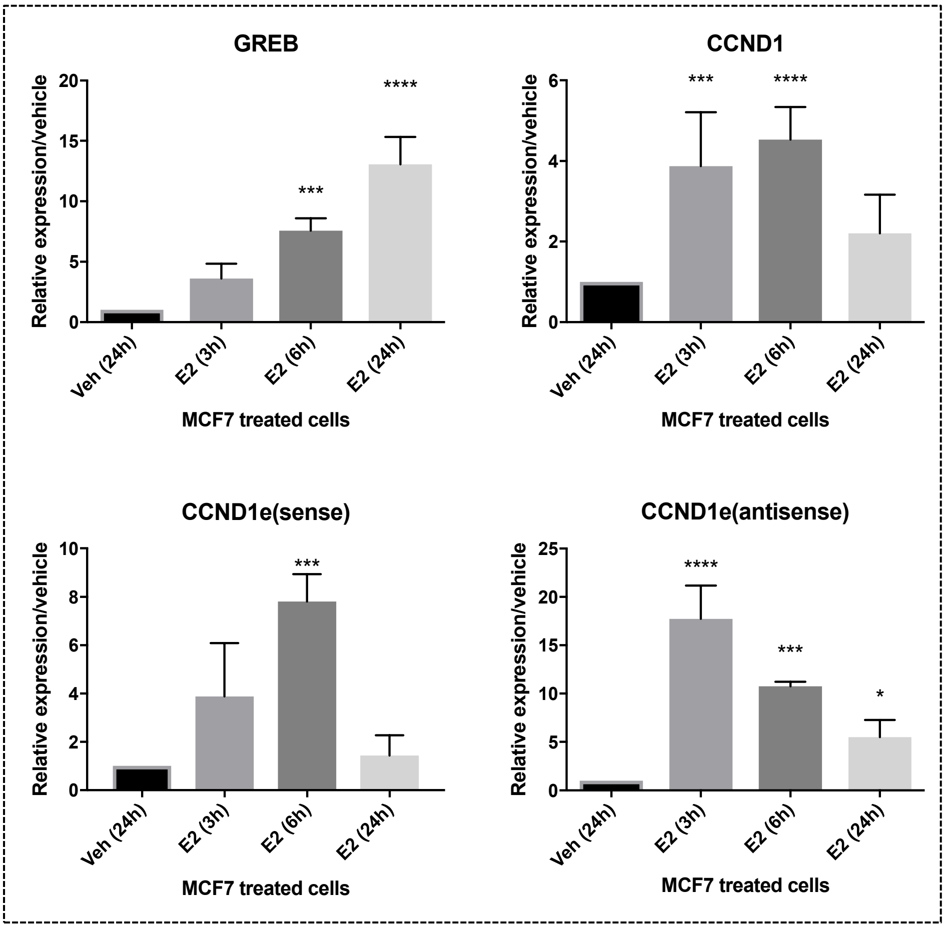
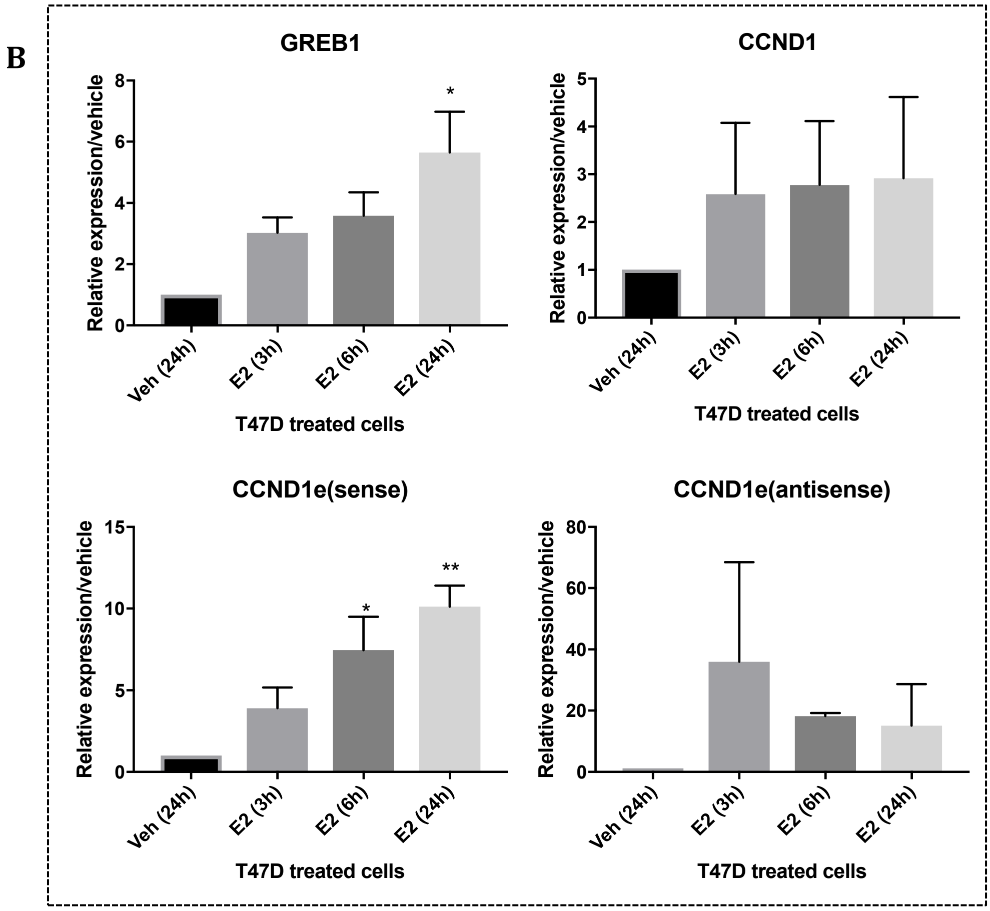
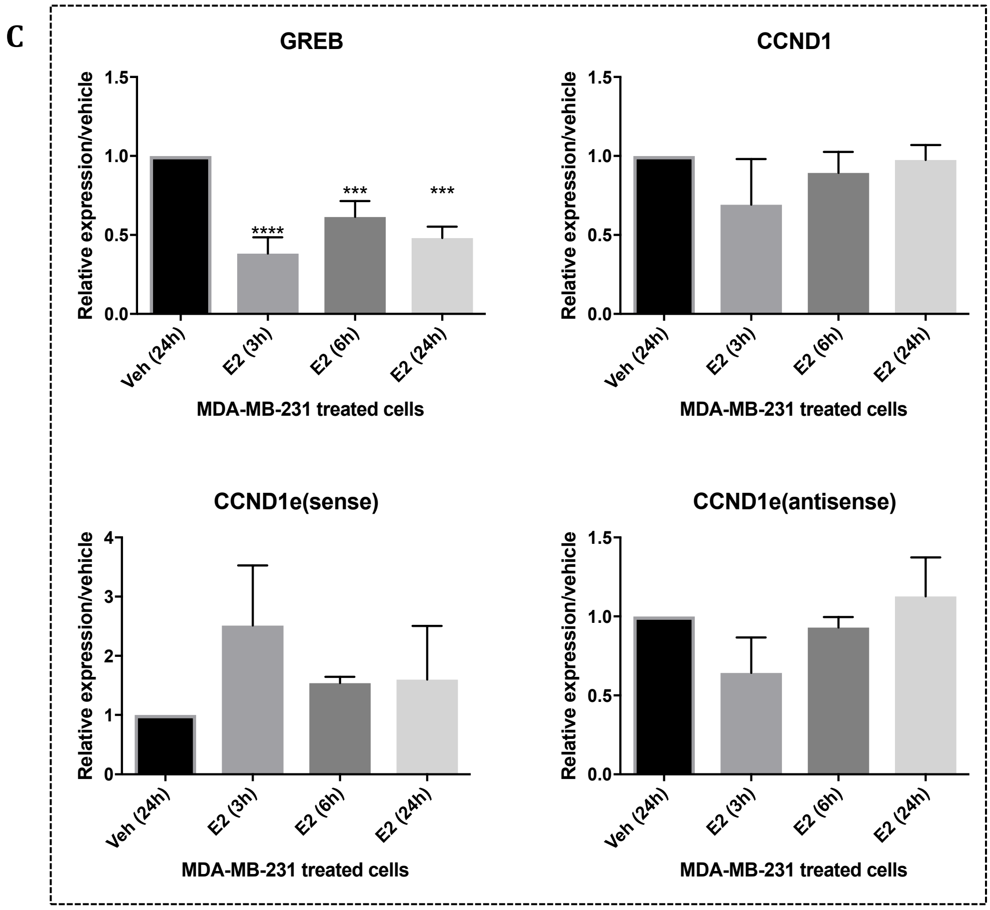


Figure 4.2 RT-qPCR analysis of the eRNA transcription in ER positive breast cancer cell lines MCF7 (A) and T47D (B) and ER negative breast cancer cell line MDA-231 (C).

Breast cancer cells were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours. Cells were then plated onto 6 well plates (day 0) with either vehicle (ethanol) or 10nM 17β estradiol for 24 hours. RT-qPCR was performed from RNA prepared from the estrogen treated MCF7 cells at 0, 3, 6 and 24 hours and the vehicle at 24 hours. Data were normalised to GAPDH levels and the expression level of each gene is shown relative to expression of vehicle at 24 hours. The mean and standard error (s.e.m.) of three biological replicates are shown. Ordinary One way ANOVA *p<0.1, **p<0.01, ***p<0.001.


B
GREB1
Relative expression/vehicle
0
2
4
6
8
*
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
T47D treated cells
CCND1
Relative expression/vehicle
0
1
2
3
4
5
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
T47D treated cells
CCND1e(sense)
Relative expression/vehicle
0
5
10
15
*
**
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
T47D treated cells
CCND1e(antisense)
Relative expression/vehicle
0
20
40
60
80
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
T47D treated cells

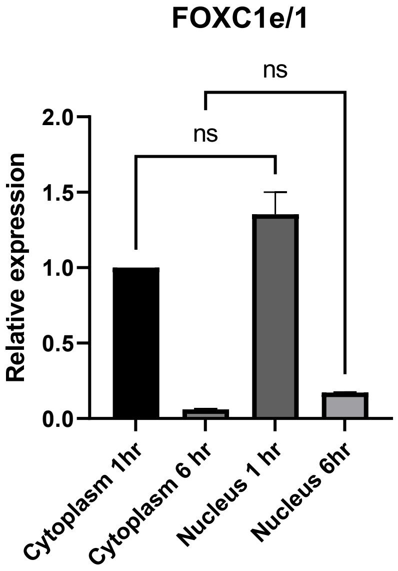

C
GREB
Relative expression/vehicle
0.0
0.5
1.0
1.5
****
***
***
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
MDA-MB-231 treated cells
CCND1
Relative expression/vehicle
0.0
0.5
1.0
1.5
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
MDA-MB-231 treated cells
CCND1e(sense)
Relative expression/vehicle
0
1
2
3
4
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
MDA-MB-231 treated cells
CCND1e(antisense)
Relative expression/vehicle
0.0
0.5
1.0
1.5
Veh (24h)
E2 (3h)
E2 (6h)
E2 (24h)
MDA-MB-231 treated cells

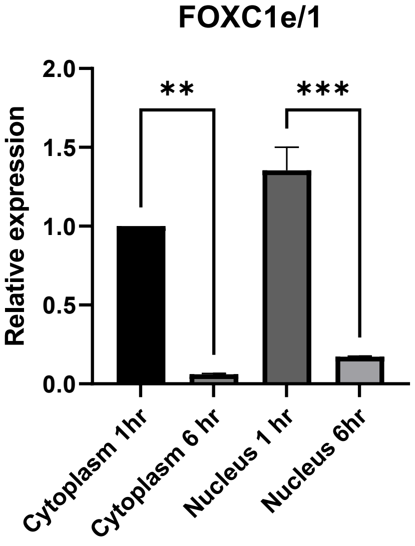
4.2 Cellular location of eRNAs *CCND1*e(sense) and *CCND1*e(antisense)

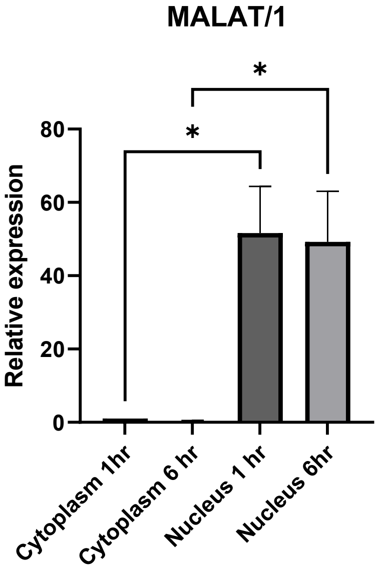
If enhancer RNAs are purely a by-product of the transcription of enhancers and have no function of their own, then transportation of the transcript out of the nucleus and into the cytoplasm would be a waste of cellular energy. Indeed, a substantial proportion of the non-coding RNA transcribed is rapidly degraded in the nucleus, suggesting either no function, or only a fleeting one. Analysis of publicly available RNAseq data of fractionated MCF7 cells suggests that the eRNA *CCND1*e(antisense) is found primarily in the cytoplasm, whilst *CCND1*e(sense) is almost entirely found in the nucleus (NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nij.gov/geo/) accession number GSE63189) [137].

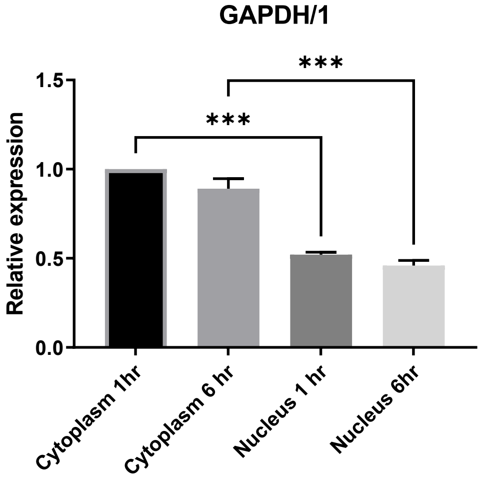
Initial attempts to validate this RNA-seq data and identify ncRNAs a) associated with the chromatin, b) localised to the nucleus and c) in the cytoplasm through fractionation of the cell into three compartments, were unsuccessful. This is possibly because the true number of transcripts in some of the three compartments is very low and difficult to reproduce with the technique used. Hence, after numerous attempts to isolate the chromatin bound transcripts, the two compartment model technique was used and the RNA from these compartments analysed with RT-qPCR.

It is well recognised that the lncRNA MALAT1 is found only in the nucleus whilst the enhancer of FOXC1 (FOXC1e) is found in both the cytoplasm and the nucleus shortly after its transcription but is quickly degraded in both[137]. These two lncRNAs were used to validate the RT-qPCR results of cellular fractionation after 1 and 6 hours of estrogen exposure in MCF7 cells and can be seen in Figure 4.3.


FOXC1e/1
ns
ns
Relative expression
2.0
1.5
1.0
0.5
0.0
Cytoplasm 1hr
Cytoplasm 6 hr
Nucleus 1 hr
Nucleus 6hr


FOXC1e/1
**
***
Relative expression
2.0
1.5
1.0
0.5
0.0
Cytoplasm 1hr
Cytoplasm 6 hr
Nucleus 1 hr
Nucleus 6hr


MALAT/1
*
*
Relative expression
80
60
40
20
0
Cytoplasm 1hr
Cytoplasm 6 hr
Nucleus 1 hr
Nucleus 6hr


GAPDH/1
***
***
Relative expression
1.5
1.0
0.5
0.0
Cytoplasm 1hr
Cytoplasm 6 hr
Nucleus 1 hr
Nucleus 6hr

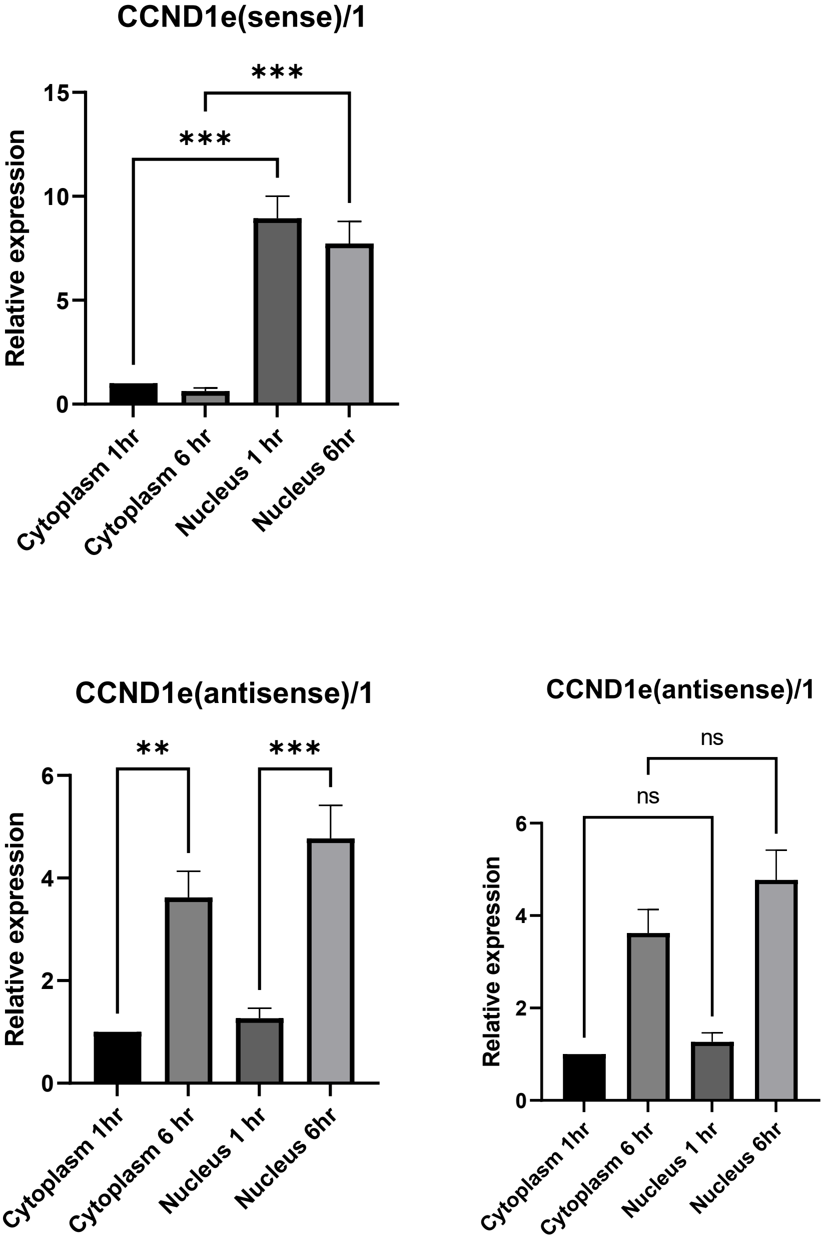


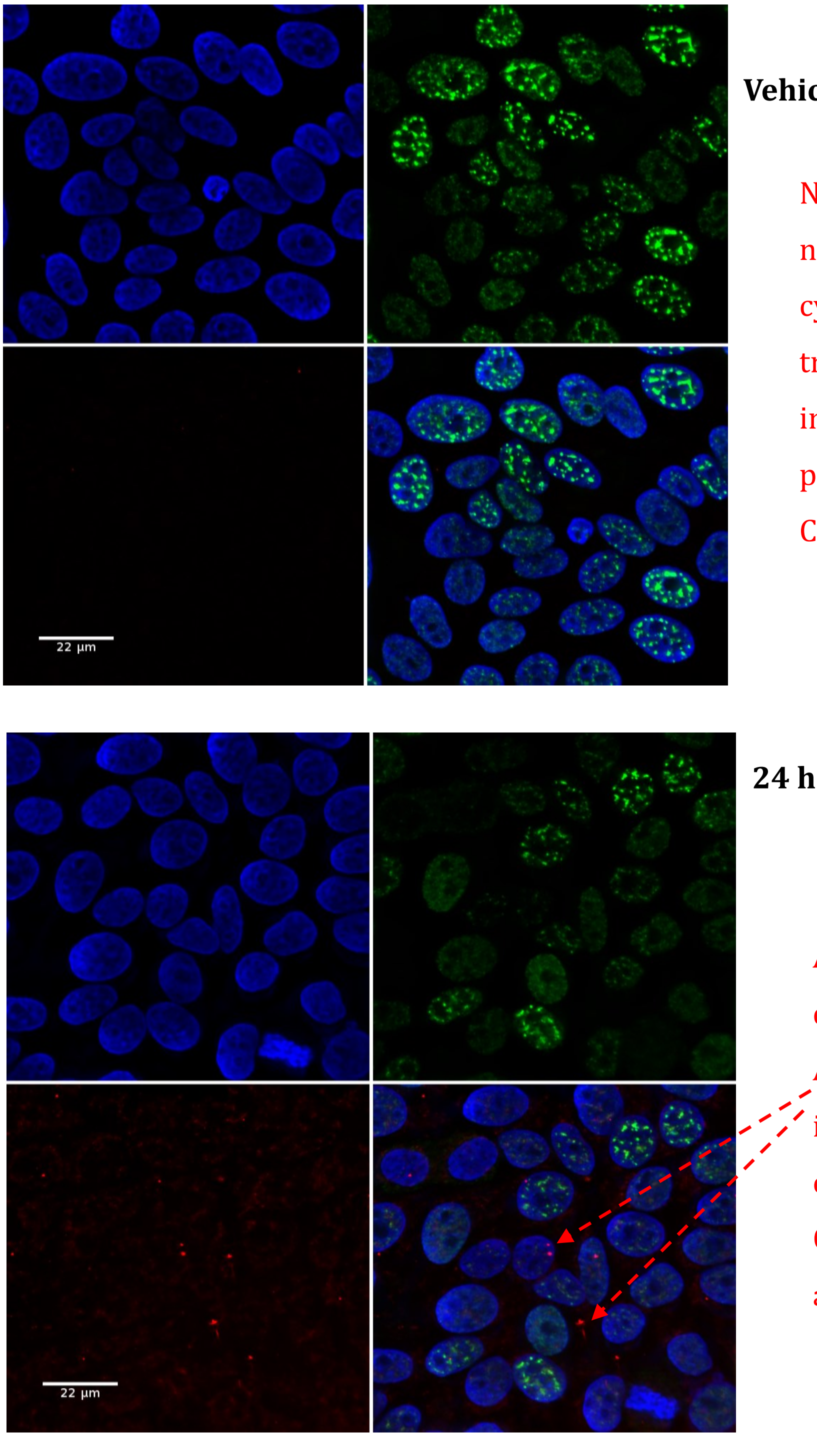
Figure 4.3 RT-qPCR analysis of cellular location of RNA transcripts.

Following cellular fractionation into nucleus and cytoplasmic fractions as described in chapter 2, RT-qPCR analysis was performed from RNA prepared from the oestrogen treated MCF7 cells at 0 and 6 hours.

The mean and standard error (s.e.m) of 2 or 3 biological replicates are shown relative to expression in the cytoplasm at 1 hour. One way ordinary anova comparing means in each compartment at the two timepoints (and between compartments at 1 timepoint for CCND1e(antisense) and FOXC1e) *p<0.1, **p<0.01, ***p<0.001.

The presence of the antisense eRNA transcript in the cytoplasm both early (1 hour) and much later (6 hours) raises the possibility that it plays a physiological role in the cytoplasm. If it were a by-product of transcription, one would expect rapid degradation which would usually occur within the nucleus.

The cellular location of the sense and antisense eRNA transcripts were further investigated using confocal staining with Stellaris RNA FISH, in which the MCF7cells were hybridized with *CCND1*e(antisense) FISH probe set labelled with Quasar 670 dye (red), and a MALAT1 probe set labelled with Quasar 570 dye (green). The protocol had been optimised by other members of the host laboratory with special thanks to Dr Angela Yu for her assistance in the use of her optimised protocol. The MALAT1 probe identifies the exclusively nuclear RNA MALAT1. The results at 6 and 24 hours of estrogen exposure, and 24 hours vehicle are shown in Figure 4.4. The RNA FISH images suggest that the antisense RNA transcript is found in the cytoplasm, as well as the nucleus, following exposure to estrogen. It is not possible from these images to quantify its expression across the time course, but they suggest a presence at both 6 and 12 hours. The host laboratory group subsequently performed RNA immunoprecipitation of the antisense transcript in the cytoplasm followed by mass spectrometry to identify any protein bound to the transcript. This work was performed by my colleague Dr Paul Cathcart and the full extent of this work can be found in his thesis [138].



Vehicle

No red dots in nucleus or cytoplasm in vehicle treated cells indicating minimal presence of CCND1e(antisense)

24 hour E2

Arrow points to red dots in nucleus AND cytoplasm indicating presence of CCND1e(antisense) at 24 hours

Figure 4.4 In hybridisation using Stellaris Fluorescent RNA probes reveals ncRNA localisation.

Red dots denote the Quasar 670 dye hybridized to the *CCND1*e(antisense) transcript and green dots are Quasar 570 dye hybridized to MALAT1, known to be found only in the nucleus of the MCF7 cell.

4.3 Exploration of the eRNAs in breast cancer using The Atlas of Non-coding RNAs in Cancer (TANRIC)

I used the open-access web resource TANRIC (http://ibl.mdanderson.org/tanric/_design/basic/main.html) to characterise the expression profiles of the two unannotated eRNAs *CCND1*e(sense) and *CCND1*e(antisense) in breast cancer cohorts. The breast cancer data is generated by the The Cancer Genome Atlas (TCGA) Research Network (http://www.cancer.gov/tcga) in which ncRNA expression is quantified using paired end sequencing results from 105 normal samples and 837 tumour samples. Figure 4.5 shows the expression data of the *CCND1*e(antisense) eRNA transcript in breast tumours known to be both positive and negative for the estrogen receptor (A) and across the staging of all breast cancers (not ER related)(B). The data shows that both the sense and antisense eRNA expression is much higher in ER positive tumours than negative (y axis, lncRNA expression log2), although no statistically significant difference is seen between the early stages of disease (Stage 1) and the advanced metastatic stage (stage IV) in both ER positive and negative breast cancer.

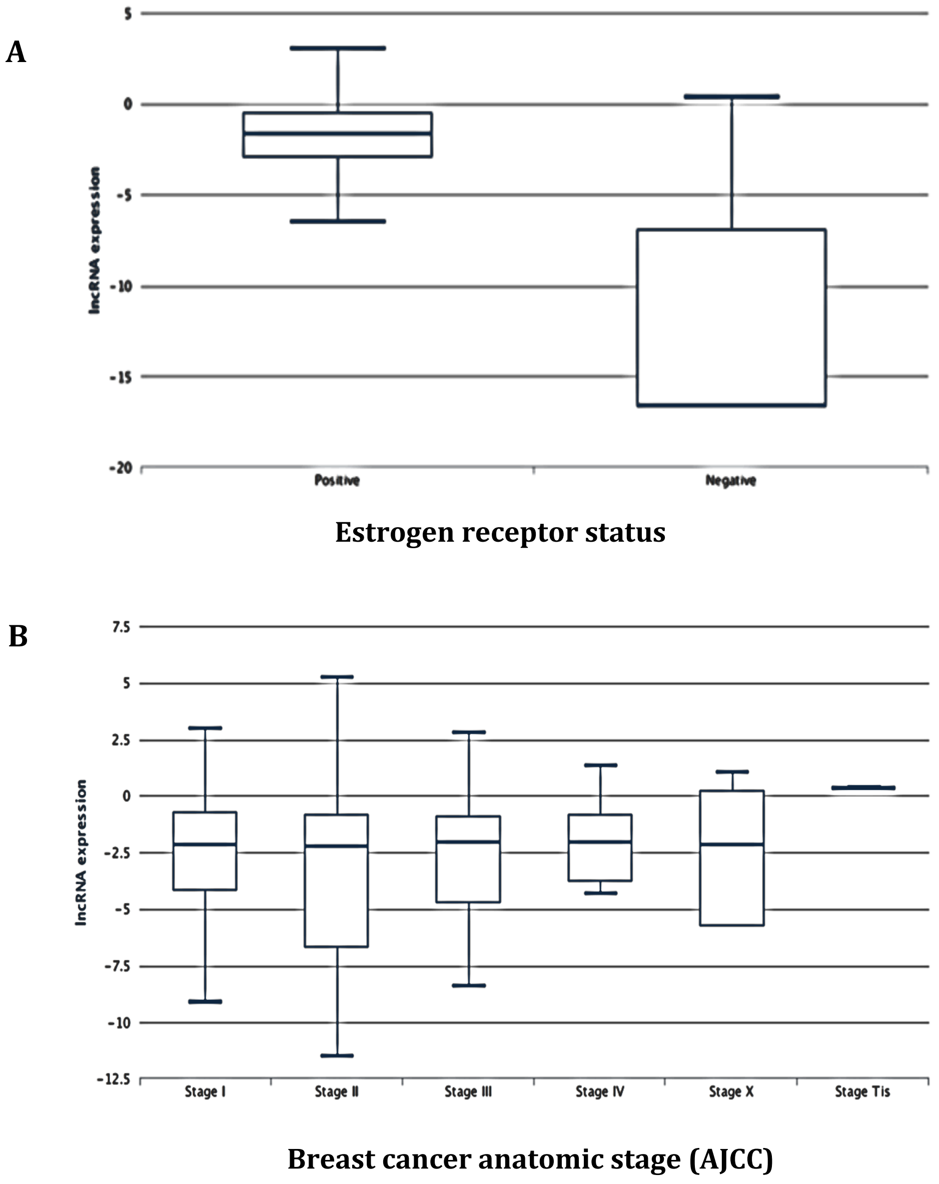


Figure 4.5 TCGA analysis of eRNA *CCND1*e(antisense) across ER receptor status and Breast Cancer anatomic staging.

eRNA relative expression (log2 FPKM) shown on y axis.

(A) The expression of *CCND1*e(antisense) in ER positive and ER negative breast cancers

(B) The expression of *CCND1*e(antisense) across the recognised stages of all breast cancer, regardless of hormone receptor status

Breast cancer staging according to AJCC Anatomic Staging groups as described in AJCC Cancer Staging Manuel, 8th ed. [139]

4.4 Knockdown of the sense and anti-sense eRNA transcripts with siRNA and antisense oligomer (ASO) technology does not affect *CCND1* mRNA levels

To investigate the role of the enhancer RNA transcripts on *CCND1* mRNA, I first sought to generate loss of function models using knockdown of the transcript with RNA interference (RNAi) using short interfering RNAs (siRNAs) and antisense oligomirs (ASOs). In keeping with the known limitations of siRNAs (as discussed in Chapter 1), I found that I was able to successfully knockdown the more cytoplasmic antisense *CCND1*e using two siRNAs at 10nM (siRNA A (si501508) and siRNA B (s501509) both from Life Technologies) (Figure 4.6 A). However, only one of many siRNAs designed to target the more nuclear sense *CCND1*e was successful (siRNA A sense (s501511) (Figure 4.6 B). In order to overcome the difficulty of knockdown of the nuclear transcript, I tried ASOs (GapmeRs) targeting the sense eRNA, at both 10nM, 25nM, 50nM and 80nM. As shown in Figure 4.6 D-F I was unable to identify any ASOs capable of statistically significant knock down of the sense transcript despite the higher and potentially more toxic concentration.

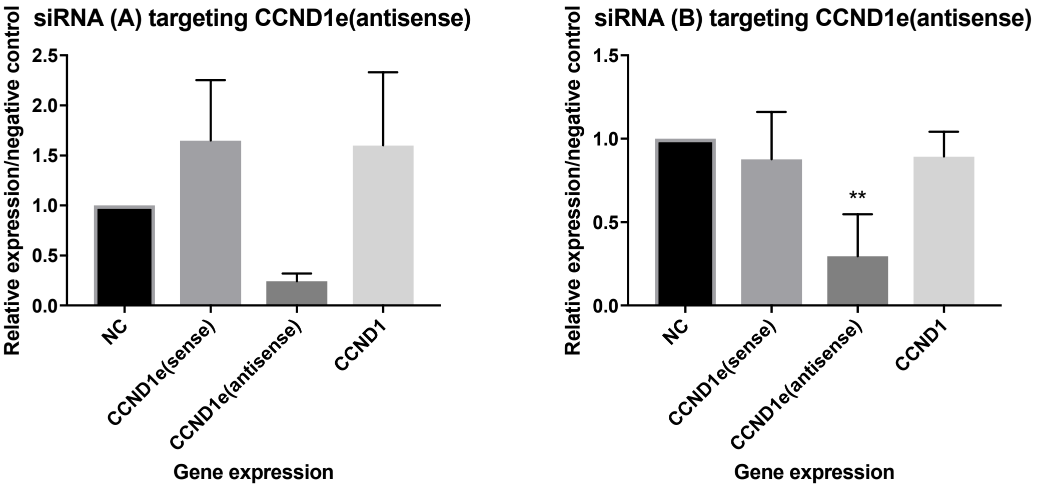
I used a number of transfection reagents including Hiperfect, Lipofectamine and RNAiMAX as well as varying cell seeding concentration and cell confluence at transfection, transfection timing, transfection reagents, concentration of siRNA/ASO and both forward and reverse transfection methods.

Whilst I was somewhat successful in knocking down the eRNAs with siRNA, I found that it had no effect on the mRNA levels of *CCND1*. The literature reports many examples of enhancer derived transcript knockdown downregulating the neighbouring gene [60],[63-65] and hence I expected the same and so spent a long time repeating these experiments and trying to optimise their conditions and using different siRNAs and ASOs. However, I showed clearly that knocking down any single eRNA had no effect on the neighbouring gene *CCND1* or on levels of the other divergently transcribed eRNA. Instead I actually found that with knockdown of *CCND1* mRNA using a commercially available siRNA against *CCND1* (Life Technologies), I saw a significant upregulation of both of the eRNAs

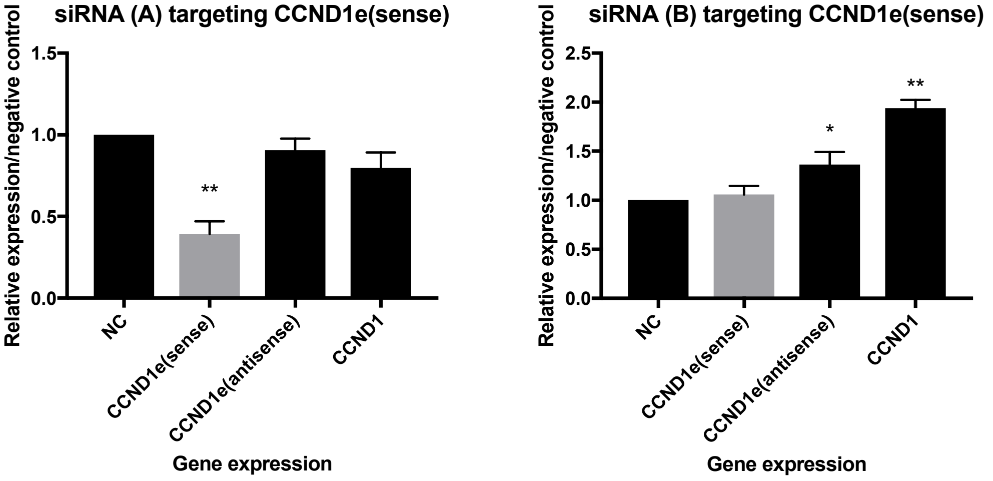
(Figure 4.6 C). These findings would suggest that the enhancer transcripts do not themselves play a role in regulation of *CCND1* expression, but it is possible that through a feedback loop the *CCND1* mRNA may be involved in the transcriptional regulation of the eRNAs. Further analysis of siRNA and ASO knockdown of both the eRNAs and *CCND1* is discussed in Chapter 6.

Figure 4.6 RT-qPCR of eRNA knockdown using siRNAs and ASO.
Reverse transfection of siRNA/ASO into each transfection was performed in triplicate. RT-qPCR was performed from RNA prepared from MCF7 cells after reverse transfection with the siRNA/GapmeR. Cells initially underwent 72 hours of estrogen starvation prior to transfection followed by 24 hour treatment with either vehicle or 10nM 17β estradiol afterwards. Expression was normalised to expression levels in the negative control. The mean and standard error (s.e.m.) of three biological replicates are shown *p<0.1, **p<0.01, ***p<0.001.

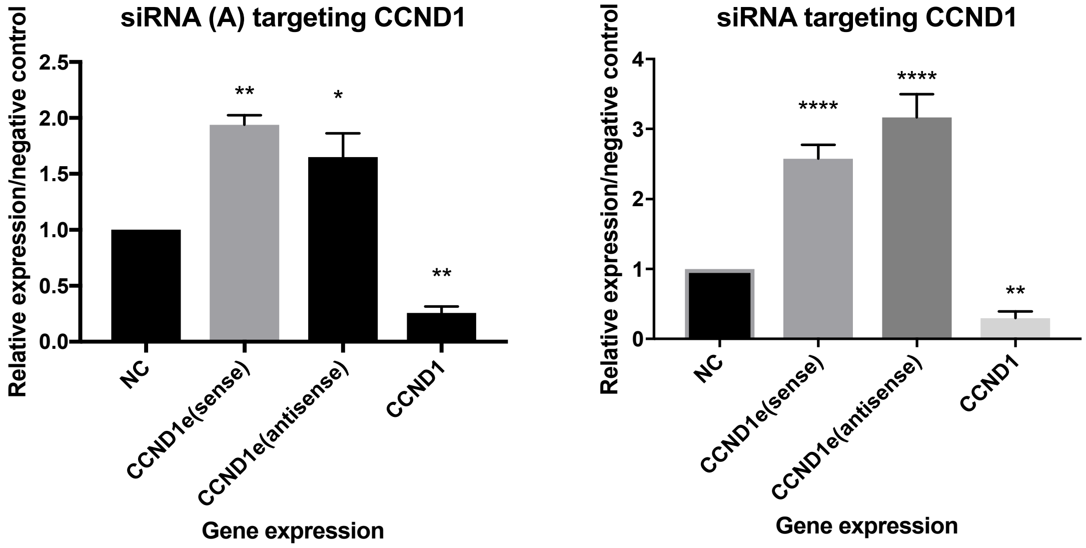
A) Knockdown of *CCND1*e(antisense) using siRNAs s501508 (A) and s501509 (B) at 10nM



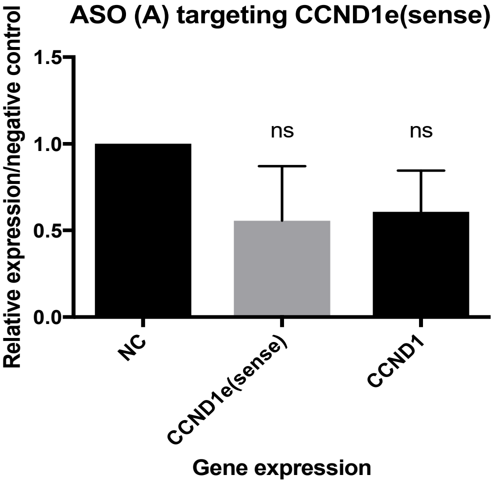
B) Knockdown of *CCND1*e(sense) using siRNAs s501511(A) and s501512(B) at 10nM

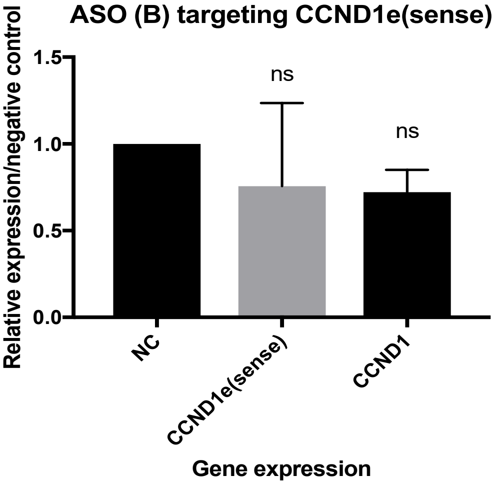


C) Knockdown of *CCND1* using siRNAs (s229(A) and S201129(B)) at 10nM

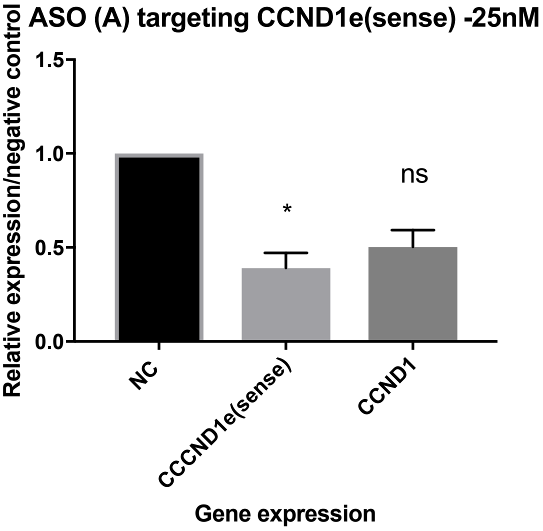


D) Knockdown of *CCND1*e(sense) using ASOs (C12-1 (A) and C12-7(B)) at 10nM

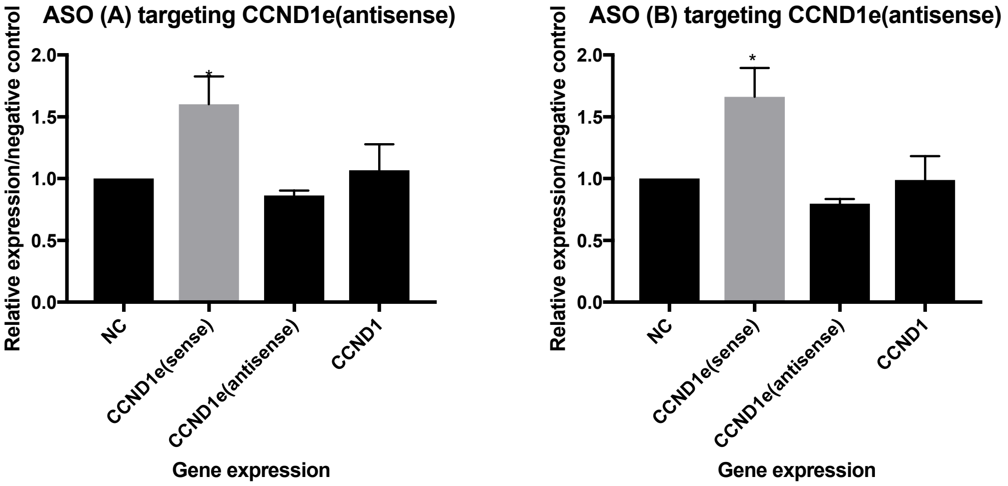




E) Knockdown of *CCND1*e(sense) using ASO (C12-1) at 25nM



F) Knockdown of *CCND1*e(antisense) using ASOs (A(A4) and B(A6)) at 25nM



4.5 siRNA knockdown of *CCND1*e RNA does not affect the cell cycle

CCND1 plays an important role in regulation of the cell cycle and it is known that knockdown of *CCND1* results in the cell being held in G0/G1 phase [140, 141]. I confirmed this finding using the Millipore Muse® Cell Cycle kit which showed that with transfection of cells with a negative control 38.15% of all live cells were held in G0/G1 phase, but with transfection of an siRNA targeted against *CCND1*, 52.45% of all live cells were held in this phase (Table 4.1).

In keeping with the findings that knockdown of the eRNA transcripts *CCND1*e(sense) and *CCND1*e(antisense) had minimal effect on *CCND1* mRNA levels, I found that knockdown of the enhancer RNA did not affect the percentage of cells held at G0/G1 with knockdown of both transcripts resulting in approximately 37% of live cells in the G0/G1 phase, similar to that with negative control.

	% G0/G1	*% S*	*% G2*
si Negative Control	38.15 ± 2.61	24.05 ± 4.1	22.65 ± 3.09
siCCND1	52.45 ± 3.13	22.5 ± 3.54	15.85 ± 2.77
siCCND1e(sense)	37.25 ± 3.67	25.2 ± 2.76	27.15 ± 3.45
siCCND1e(antisense	37.45 ± 2.98	26.15 ± 2.81	25.95 ± 1.99

Table 4.1 Cell cycle analysis following eRNA knockdown.
Percentage of all live cells held at each phase of the cell cycle following MCF7 transfection with the siRNA noted. Cells initially underwent 72 hours of estrogen starvation followed by 24 hour treatment with either vehicle or 10nM 17β estradiol. Results calculated by Milliport Muse™ Cell Analyzer. Mean ± s.e.m. of three replicates.

4.6 siRNA knockdown of *CCND1*e RNA does not affect cellular proliferation

I used colimetric assay of WST-1 in a 96 well plate using a multi-well ELISA reader to assess effect of eRNA knockdown on cellular proliferation. *CCND1* is a regulatory subunit of the cyclin-dependent kinases whose activity controls cell proliferation and development and its knockdown by siRNA has previously been shown to cause significant reduction in cell proliferation [140, 141]. I report the same finding in MCF7 cells transfected with si*CCND1* (Figure 4.7). However, i did not find a similar effect on cellular proliferation when silencing either the sense or antisense transcripts of eRNA, as shown in Figure 4.7.

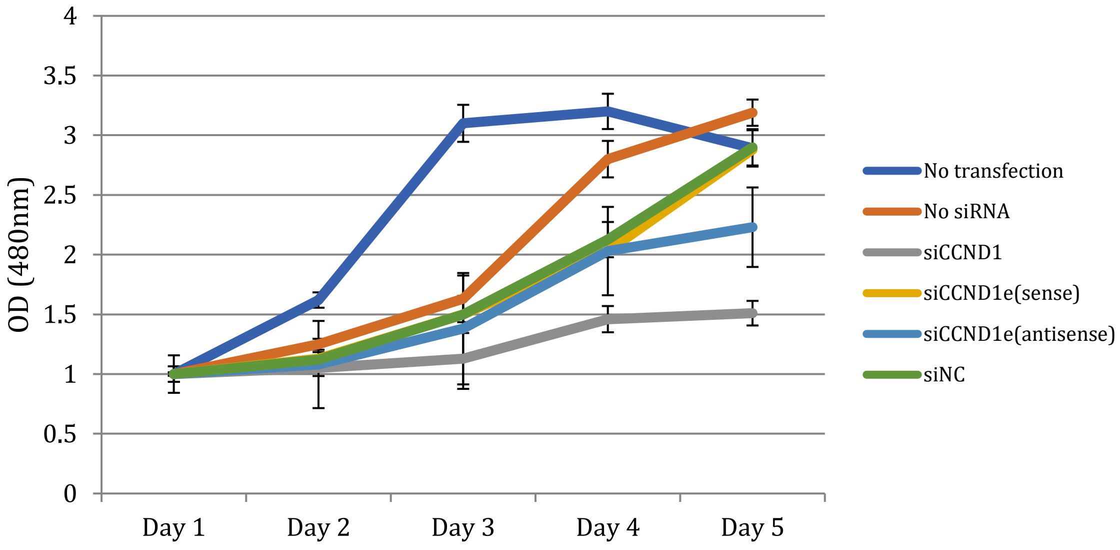


Figure 4.7 MCF7 Cellular proliferation measured by WST-1 at 480nm. Controls include no transfection reagents (No transfection), transfection reagent but no siRNA (no siRNA) and siNegative Control (siNC). siRNAs at 10nM targeted at *CCND1*e(sense) and *CCND1*e(antisense) and *CCND1*.

4.7 Overexpression *CCND1*e RNA does not affect expression of *CCND1*

Capped Analysis of Gene Expression (CAGE) is a technique used to identify the capped 5'end of RNA transcripts to a single base resolution to identify the transcriptional start site of transcripts, TFs, promoters and enhancers [142]. The sequencing analysis data is available as a result of the FANTOM (Functional ANnoTation Of the Mammalian genome) research consortium led by RIKEN and is publically available through ENCODE. The FANTOM5[51] project used CAGE technology in more than 1000 human and mouse primary cells, cell lines and tissues to accurately map the 5' ends sets in a large variety of primary cell lines and tumours and I thus used the call set from the ENCODE portal[143] (http://www.encodeproject.org/) with the Accession code ENCFF000USH to identify the exact start site of both eRNA transcripts for generation of gain of function models and for the work in chapter 5 using CRISPR/Cas9 genome editing. This analysis was performed by Dr L Castellano and gain of function models and 3'RACE were carried out by my colleague Dr P Cathcart and I am grateful for their sharing of the data [138].

Gain of function, or overexpression models, can be used to investigate the effect of increasing the expression of a particular RNA on the cell and can be used to assess the effect on neighbouring genes. Using CAGE to determine the transcriptional start site for both transcripts and either 3'RACE (for *CCND1*e(antisense)) or predicted sequence (for *CCND1*e(sense)) for the 3' end of each eRNA, the full sequence of each enhancer RNA was used to generate cDNA and cloned into an expression vector for transient overexpression of 48 hours. RT-qPCR analysis of the transfected MCF7 cells showed significant overexpression of the eRNA expressed in the plasmid but no effect on either the neighbouring gene *CCND1* or the divergently transcribed eRNA (data not shown) [138].

4.8 Summary

Having determined that the transcripts arising from the enhancer region upstream of *CCND1* bore the hallmarks of enhancer RNAs, but knockdown (which may not have been sufficient) did not result in the consequent knock down of the *CCND1* mRNA, or on phenotypic changes seen with *CCND1* knockdown, I considered alternative loss of function models. Of particular interest was the notion that the transcript themselves may not be important in regulating *CCND1*, but instead the act of transcription of the enhancer might be the regulating factor and this is explored further in chapter 5.

Chapter 5 Results: CRISPR/Cas9 mediated knock-in of a transcription termination sequence into the CCND1 enhancer

5.1 Using the CRISPR/Cas9 gene editing system to engineer premature transcription termination of the *CCND1* enhancer

Having shown through knockdown experiments that the eRNA transcripts themselves are not necessary for regulation of *CCND1* expression, I thus hypothesised that the act of transcription of the *CCND1* enhancer region may be important in its regulatory function. Hence, with the assistance of Professor Buluwela (Department of Cancer and Surgery, Imperial College, London), I sought to engineer premature termination of the transcription of the enhancer region and used CRISPR/Cas9 technology to "knock-in" a polyadenylation (polyA) transcription termination signal as described by Proudfoot. In 2011 Proudfoot published a 49 base polyadenylation sequence capable of terminating transcription by RNA polymerase II[122] and I used this sequence as a donor template for homologous directed repair at the 3' end of the target enhancer TSS. I used publicly available ChIP-Seq data of known transcription factor binding sites at the TSS and designed sgRNAs to target the genome approximately 200 bases downstream of that region with the intention to terminate transcription of the enhancer region shortly after RNA polymerase II activity but not to interfere with known transcription factor binding to the genome. Illustration of the intended cut site is shown in Figure 5.1 and Figure 5.2.

At the time of attempting this technique in 2016, CRISPR/Cas9 knock-in of a polyadenylation signal into an enhancer had only been published once[144], and the process of "knocking-in" a specific sequence was considered one of the more challenging techniques of CRISPR/Cas9 genome editing. CRISPR/Cas9 has recently emerged as a powerful tool for genome editing; on cleavage of a DNA double stranded break (DSB) with the Cas9 nuclease the cell will attempt to fix the break with endogenous machinery. Non Homologous End Joining (NHEJ) is the preferred repair pathway by the cell but often results in the introduction of insertions or deletions which can lead to missense mutations. However, if a donor sequence is provided to the site of DSB, this can induce the cell to use homology-directed repair (HDR) for more precise genome editing. It is this process that can be utilised to knock-in a specific sequence.

During the course of this CRISPR work, Korkmaz et al[118] published the results of a functional genetic screen for enhancer elements using ERα and p53 as TFs of interest. They used the top 2000 Chip-Seq ERα binding sites reported by Li et al [60] and intersected 406 of them (those with ERα consensus motif and able to be targeted by CRISPR) with those transcribing bidirectional nascent eRNA according to GRO-Seq data[60]. In doing so they identified 73 ER bound bi-directionally transcribed enhancers, one of which was the same previously uncharacterised genomic region as that I had been investigating. CRISPR/Cas9 mediated mutation of the ER binding site at this E2 regulated enhancer resulted in reduced cellular proliferation, down regulation of eRNA expression and down regulation of the *CCND1* mRNA and protein (which was seen on ChIA-PET to interact directly with the enhancer). They report that activation of *CCND1* expression by estrogen in MCF7 cells requires the fully active and non- mutated ER binding site at this enhancer[118]. Their identification of this enhancer region validated my bioinformatical finding that this region is indeed an active enhancer of *CCND1* and thus I sought to further identify the mechanism by which the transcription and resultant eRNA play a role in its function.

Whilst the functional screening validated the enhancer region, the ER binding site mutation prevented ER binding and hence activation of the enhancer and transcription of the eRNA. With this in mind, the CRISPR work was designed such that ER and all other TFs known to bind to the TSS of the enhancer should still be allowed to bind and initiate the activation cascade including initiation of transcription as I wanted to determine if the act of transcription of the enhancer plays a role in its regulatory function. Thus I planned to create a double strand break using *Streptococcus Pyogenes* Cas9 (SpCas9, which recognises the PAM site NGG) as close to the transcription start site as feasibly possible whilst avoiding the regions of TF binding visualised in UCSC genome browser (http://genome.ucsc.edu)[119] from publicly available CHIP-Seq data performed by the ENCODE project [145-147] (Figure 5.1). I used UCSC to visualise the expected binding sites of TFs including ER (*ESR1)*, *E2F*1, *ZNF217, GATA3, MYC, FOXA1, SPI1, STAT3, FOS* and *TCF7L2*.

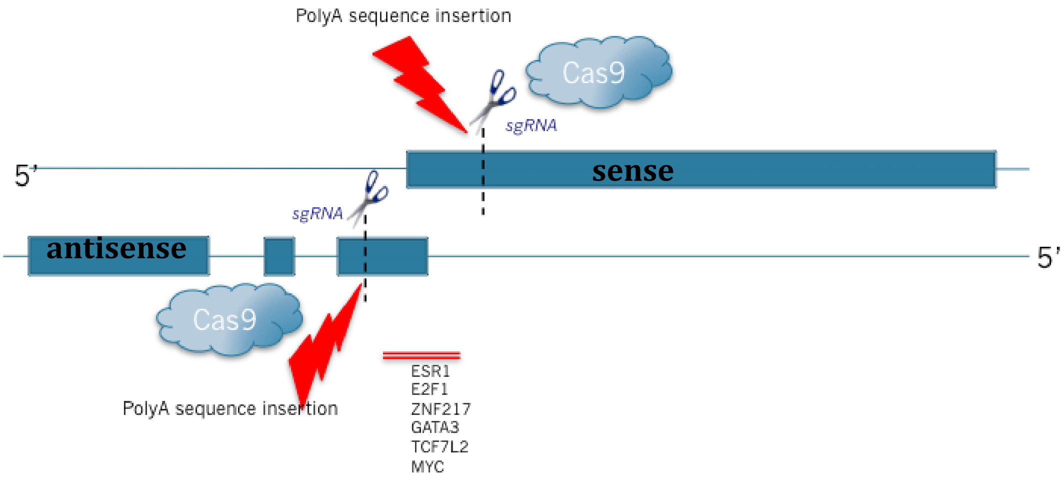


Figure 5.1 Diagrammatic representation of intended site for insertion of polyadenylation signal.

sgRNAs were designed to target both the sense and antisense strand of the enhancer and lead Cas9 to generate a DSB just downstream of the recognised TF binding sites.

5.1.1. Designing single guide RNAs

Single guide RNAs (sgRNAs) are an artificial fusion of CRISPR RNA (crRNA) and transactivating RNA (tracrRNA) which have the necessary secondary structures for loading onto Cas9 and for hybridization to the genomic DNA. In CRISPR/Cas9, the sgRNA substitutes for the crRNA-tracrRNA complex that occurs in natural CRISPR systems. Because my system utilised the *Streptococcus pyogenes* Cas9 (*sp*Cas9), the sgRNAs contained a 20 base pair (bp) guide sequence homologous to the target DNA and adjacent to the 3 base pair protospacer adjacent motif 'NGG' (other Cas9 systems require alternative sgRNA design). Each of the sgRNAs required a G base at the 5'end for integration into the CRISPR tracrRNA/crRNA scaffold (because the human U6 promoter prefers a G at the TSS to have higher expression[99]) making them 21bp in total. Although able to tolerate up to 5 mismatches between the sgRNA and the target genomic sequence, the 10-12 most proximal bases of the guide are the main determinants of specificity and as such steps are required to ensure that the predicted genomic sequence is in fact representative of the cell line being used.

5.1.2 Validating predicted target genomic sequence

sgRNA design is dependent on knowing the precise genomic sequence at the first 5-12 bases of the intended cut site because a single genomic variation (including SNPs) could render the sgRNA ineffective. Hence, validating the expected genome sequence around the target region is the first step in the process of the CRISPR/Cas9 protocol as although the MCF7 cell line has been sequenced and published by the Encyclopaedia of DNA Elements (ENCODE) project (http://www.encodeproject.org)[143], it is possible that my MCF7 cell line had developed single point mutations. To verify the sequence of the MCF7 cell line being used I first performed PCR amplification and Sanger sequencing of the intended CRISPR site. This was done for each new batch of the MCF7 cell line and for every intended genomic region. The Sanger sequencing confirmed no deviations from the ENCODE published sequence for each region[143].

5.1.3 Targeting CCND1e(sense) with CRISPR/Cas9

For the intended knock-in site within *CCND1*e(sense), I first generated a predicted sequence for genomic editing from UCSC genome browser (http://genome.ucsc.edu)[119] (Figure 5.2) and used this to design primers which would amplify this region with PCR.

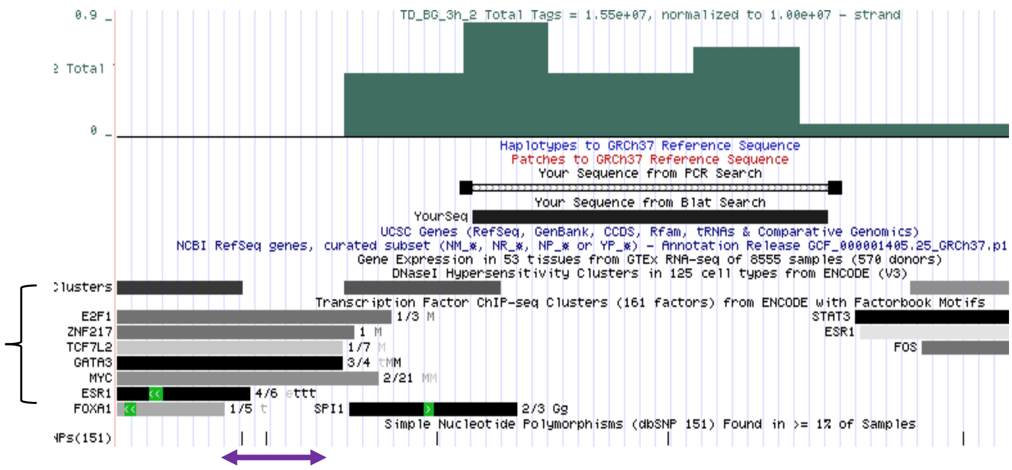

0.9
2 Total
0
TD_BG_3h_2 Total Tags = 1.55e+07, normalized to 1.00e+07 - strand
Haplotypes to GRCh37 Reference Sequence
Patches to GRCh37 Reference Sequence
Your Sequence from PCR Search
Your Sequence from Blat Search
YourSeq
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001405.25_GRCh37.p1
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Clusters
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
E2F1
1/3 M
ZNF217
1 M
TCF7L2
1/7 M
GATA3
3/4 tMM
MYC
2/21 MM
ESR1
4/6 ettt
FOXA1
1/5 t
SPI1
2/3 Gg
STAT3
ESR1
FOS
Simple Nucleotide Polymorphisms (dbSNP 151) Found in >= 1% of Samples
NPs(151)

Figure 5.2 UCSC genome browser view showing intended site for CRISPR knock-in.
The green track indicates transcription on the sense strand at *CCND1*e(sense) at 3 hours following E2. The red arrow shows the intended region for insertion of the polyadenylation sequence, avoiding the known TF binding sites. The purple arrow indicates the enhancer ER binding site targeted by Korkmaz et al[118].

The target genomic sequence of *CCND1*e(sense) for insertion of the polyA transcription termination signal (PAS) is shown in Figure 5.3(A). This sequence relates to the red arrow shown in Figure 5.2. Figure 5.3(B) shows the sequence generated following PCR amplification and Sanger sequencing by one set of primers (in red). The sequence aligns to the positive strand at chr11:69,330,695-69,331,218.

Figure 5.3 Predicted (A) and Sanger generated (B) genomic sequence for target site for CRISPR knock-in at *CCND1* enhancer.

5’ATTGGCTCACAGTTCTGCAGGCTGTGCAGGAAGCATGGTGCCCACATCTGCTTGGCTTCTGGTGAGGCCTCAG
GGAGCTTTTACTCATGATGAAAGGCCAAGTGGGAGCAGGCATCACATGGTTAAAGTGGAAGCAAGAGAGAGAAG
GGGGAGGGCCACACACTTTTAAGCAACCAGATATCAAGAGAACTTACTCACTATCTTGAGGACAGCACCAAGCCA
TTCATGAGGAGTCCACCCCATGATCAAAACTCTTTCCACCAGGCCCCACCTCCAACATTGGGGATTATGTTTCAA
TGTGAGATTTGGAGAGGACAAACATCCAAACCATATCACCATGTGATTTCTTGCACTGCCTTGAGACTTGGCAGA
GAGTCCCCAGCAGCAAGAAGGCCCTCGCCAGATGCAACCCCTCGATCTGGACTTCTCAGCCTACAGAACTGTA3’

(A) Predicted genomic sequence of knock-in target site (sense strand)

5’TTACGGAGATTGGGTCGTGAACCTCTGCCTCATGAATGGATTAAGACATTCATAGATGAGTGAATTAATAGT
TTAATGGATTGATGAGGTTGTCATGGGAGTGGAACTGAAGGCTTTATAAAAAGAGGAAGAGAGATCTGAGCCAG
CACACCCAGCCCCGGTCATGTGTATTCAGCAGTTTCACATCAATATAAAGACATTCTTAAGGCTGCGAAATGTTT
AAAGGAAAGAGGTTTAATTGGCTCACAGTTCTGCAGGCTGTGCAGGAAGCATGGTGCCCACATCTGCTTGGCTTC
TGGTGAGGCCTCAGGGAGCTTTTACTCATGATGAAAGGCCAAGTGGGAGCAGGCATCACATGGTTAAAGTGGAA
GCAAGAGAGAGAAGGGGGAGGGCCACACACTTTTAAGCAACCAGATATCAAGAGAACTTACTCACTATCTTGAG
GACAGCACCAAGCCATTCATGAGGAGTCCACCCCATGATCAAAACTCTTTCCACCAGGCCCCACCTCCAACATTGG
GGATTATGTTTCAATGTGAGATTTGGAGAGGACAAACATCCAAACCATATCACCATGTGATTTCTTGCACTGCCT
TGAGACTTGGCAGAGAGTCCCCAGCAGCAAGAAGGCCCTCGCCAGATGCAACCCCTCGATCTGGACTTCTCAGCC
TACAGAACTGTAAGATACAAATTTCTTTTCTTTATAGATGATTACCTAGTTTCAGGTATTCTGTTATAAGCAAT
AGAAAATGGGCAGAGACATCTAGCTCCTCTTGGTTTAGTCAAAGCCGACAGCAGAGAAATAAGATGAGGGATCC
TGGCTTCTAACCAAGTTGAAAGAAGAAAGTCGAGAGGCCCTCTCTCATCTCACTCTCATAAGAGGAAAAAAGCC
3’

(B)Sanger generated sequence of knock-in target site (sense strand). Forward and reverse primers for sense strand *CCND1*e in red. The sequence aligns to the positive (sense) strand at chr11:69,330,695-69,331,218.

The above sequence between the primers (shown in red) from Figure 5.3(B) was then used to generate sgRNAs using the freely available webtool crispr.mit.edu (https://*cis*r.mit.edu/) an online CRISPR/Cas9 sgRNA design webtool hosted by the Zhang lab and The Broad Institute [123]. The webtool generated a list of sgRNAs ranked according to likely on and off target effects. The same genomic sequence (Figure 5.3B) was used in a second webtool hosted by Desktop Genetics DESKGEN™ CRISPR webtool (http://www.deskgen.com/guide-picker) which utilises a machine learning algorithm to generate sgRNA. DESKGEN™ scores the sgRNA based on likely activity and minimal off target effects by looking at the relative proportion of each nucleotide, the position of the nucleotides and their neighbours and using this information to determine how likely a sgRNA is to result in a successful double strand break. The highest scoring three sgRNAs for each target with the least predicted off target effects were chosen for in vitro validation. Figure 5.4 shows the three chosen sgRNAs highlighted within the target sequence, with the neighbouring PAM in red text. Their location relative to the TF binding sites can be seen in Figure 5.5.

5’ATTGGCTCACAGTTCTGCAGGCTGTGCAGGAAGCATGGTGCCCACATCTGCTTGGCTTCTGGTGA
GGCCTCAGGGAGCTTTTACTCATGATGAAAGGCCAAGTGGGAGCAGGCATCACATGGTTAAAGTGGA
AGCAAGAGAGAGAAGGGGGAGGGCCACACACTTTTAAGCAACCAGATATCAAGAGAACTTACTCACT
ATCTTGAGGACAGCACCAAGCCATTCATGAGGAGTCCACCCCATGATCAAAACTCTTTCCACCAGGCC
CCACCTCCAACATTGGGGATTATGTTTCAATGTGAGATTTGGAGAGGACAAACATCCAAACCATATC
ACCATGTGATTTCTTGCACTGCCTTGAGACTTGGCAGAGAGTCCCCAGCAGCAAGAAGGCCCTCGCCA
GATGCAACCCCTCGATCTGGACTTCTCAGCCTACAGAACTGTA3’

Figure 5.4 Three sgRNAs targeting the genome within the target sequence of *CCND1*e(sense).
The sgRNAs are highlighted in yellow (sgRNA 107), blue (sgRNA 127) and pink (sgRNA 123). PAM sites are highlighted in red text.

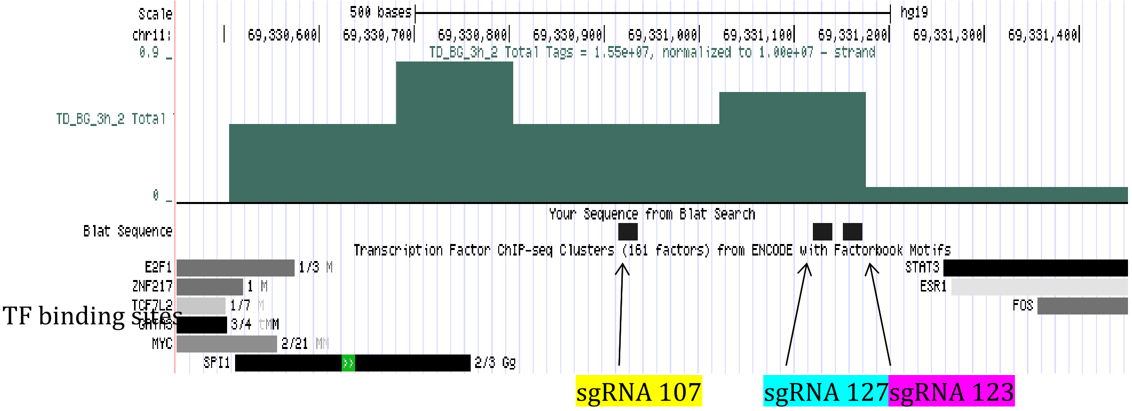


Figure 5.5 UCSC genome browser view showing location of sgRNAs relative to TF binding sites
Green track shows sense transcription at 3 hour E2 and lower tracks show known TF binding (ChIP-Seq) relative to sgRNAs targeting the sense enhancer.

5.1.4 Design of single stranded oligonucleotides HDR donors

Although the 49 base pair polyadenylation sequence remained consistent in each single stranded oligonucleotide (ssODN), in order for the homologous arms to have homology with the genomic sequence, a different ssODN was required for each sgRNA. Given that I was knocking in a 49 base sequence, I chose the ssODN design over plasmid donor as the literature reports improved efficiency with this technique with smaller donors [99]. I opted for symmetrical 35bp homologous arms to minimise the length of the ssODN and because at the time of design, 35bp was considered an optimal length. It was unknown if orientation of the donor was important in its HDR efficacy and so I designed both sense and antisense donors for each sgRNA predicted cut site. For the sense donor, the 5' homologous arm was generated from the 32 bases upstream of the sgRNA sequence and the first 3 bases of the sgRNA (because the predicted cut site is between the 3rd and 4th bases of the gRNA). The 3' homologous arm was generated from the remaining 17 bases of the sgRNA and the next 18 bases in the genomic sequence. I did not attempt knock-in with asymmetrical or longer homologous arms or the use of a donor plasmid. An example of the sense and antisense direction ssODN donors containing the 49bp polyadenylation sequence is shown in Figure 5.6. ssODN sgRNA 12 targets the first exon of the antisense enhancer.

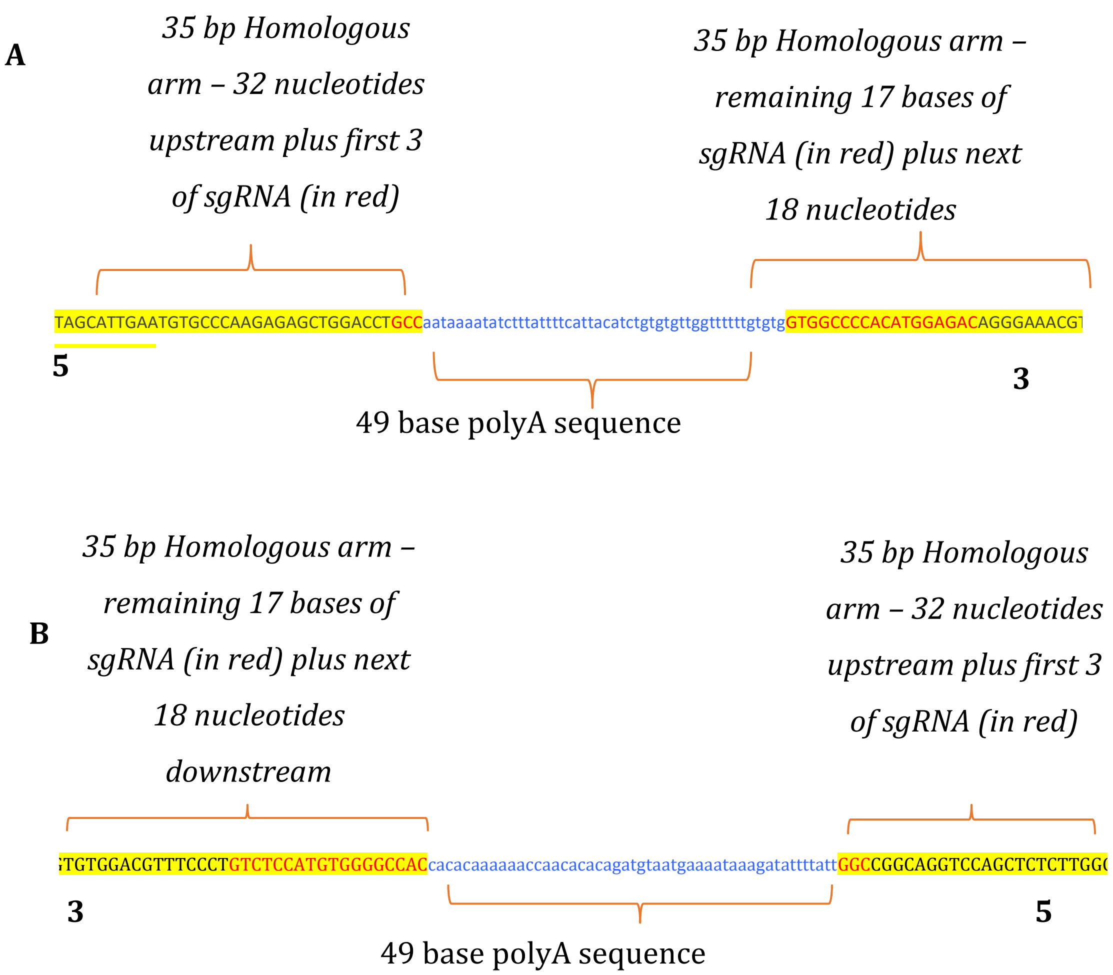

A
35 bp Homologous arm – 32 nucleotides upstream plus first 3 of sgRNA (in red)
35 bp Homologous arm – remaining 17 bases of sgRNA (in red) plus next 18 nucleotides
TAGCATTGAATGTGCCCAAGAGAGCTGGACCTGCCaataaaatatctttattttcattacatctgtgtgttggtttttgtgtgGTGGCCCCACATGGAGACAGGGAAACGT
5
3
49 base polyA sequence
B
35 bp Homologous arm – remaining 17 bases of sgRNA (in red) plus next 18 nucleotides downstream
35 bp Homologous arm – 32 nucleotides upstream plus first 3 of sgRNA (in red)
GTGTGGACGTTTCCCTGTCTCCATGTGGGGCCACcacacaaaaaaccaacacacagatgtaatgaaaataaagatattttattGGCCGGCAGGTCCAGCTCTCTTGGC
3
5
49 base polyA sequence

Figure 5.6 Single stranded Oligo donors designed for DSB associated with sgRNA 12 targeting the antisense enhancer.
A) shows the sense ssODN designed to cut the antisense strand
B) shows the antisense ssODN designed to cut the sense strand.
Highlighted in yellow are the 35bp homologous arms. In red text is the 21 bp sequence of sgRNA 12 and in blue text is the 49bp polyadenylation sequence knock-in donor.

5.1.5 Optimising the delivery method

Identification of a single clone with successful knock-in of a polyadenylation termination sequence using CRISPR/Cas9 took approximately 11 months from the start of experimental design. However, the process became much quicker following optimisation of the protocol, such that the host lab was able to generate similar models within a few weeks. With early attempts resulting in no evidence of sgRNA cutting or successful knock-in I found it better to optimise the transfection process and identify the most efficient sgRNA before attempting the knock-in. As MCF7 cells are notoriously difficult to transfect, I found that testing the sgRNAs in HEK297 cells first was a much quicker method to identify efficient sgRNAs (HEK297 data not shown).

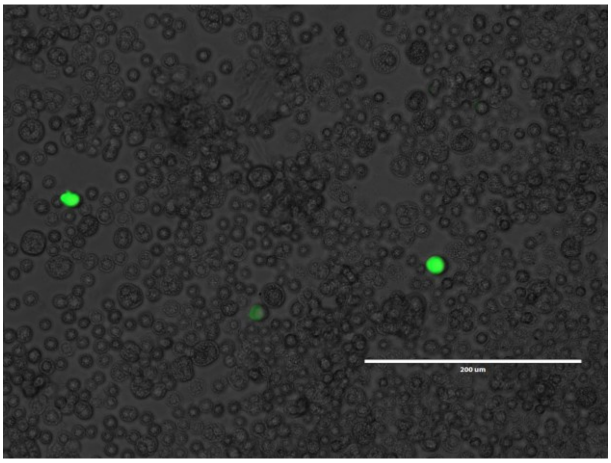
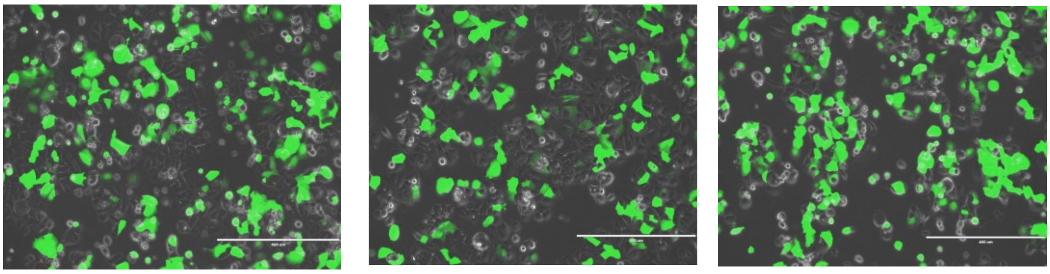


Figure 5.7 Assessment of electroporation efficiency using GFP visualisation.
MCF7 cells underwent electroporation with a GFP-expressing plasmid, sgRNA and two single stranded oligonucleotide templates for HDR. Cells were visualised 24 hours post nucleofection under fluorescence microscopy (GFP filter). The green cells indicate presence of GFP within the cell. The image shows poor efficiency and significant cell toxicity following this method.

It has been reported that electroporation is a successful method by which to transfect the CRISPR/Cas9 system into MCF7 cells, but I found that the significant amount of cell death prevented me from successfully identifying a positive clone (Figure 5.7). Time in the cuvette, the use of antibiotic free media after nucleofection and changing the voltage settings did little to improve the toxicity of the nucleofection process. Thus I compared the estimated cell death with other DNA transfection reagents such as Lipofectamine™ 2000, Lipofectamine™ 3000, Lipofectamine™ LTX (all ThermoFischer) and GeneJuice® (Merck) using a GFP plasmid as a proxy for efficient sgRNA transfection. GeneJuice® at 6µl for each transfection of 1000ng spCas9 and 1000ng sgRNA was found to be the most efficient and least toxic as shown in Figure 5.8.



Nucleofection 24 hours **Lipofectamine™ LTX 24 hours** **GeneJuice® 24 hours**

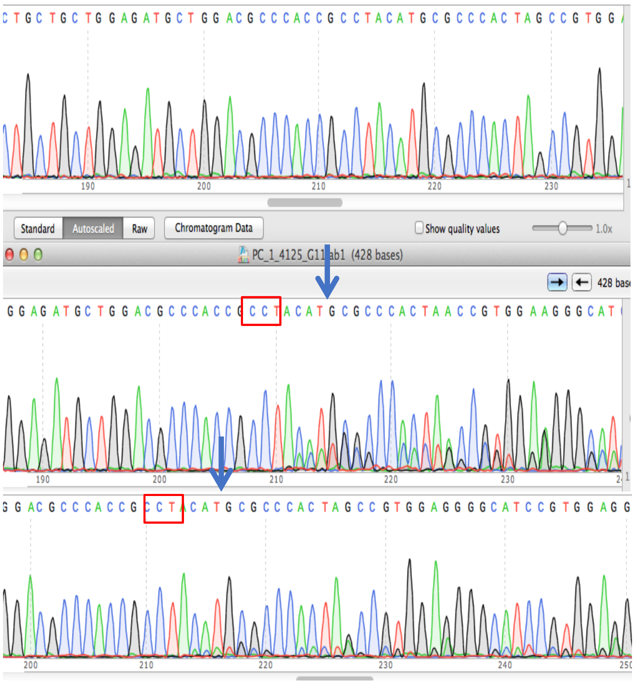
Figure 5.8 Assessment of transfection efficiency and estimated cell toxicity using GFP visualisation.
MCF7 cells were transfected with a GFP-expressing plasmid using nucleofection or the transfection reagents Lipofectamine™ LTX and GeneJuice®. Cells were visualised 24 hours post transfection under fluorescence microscopy (GFP filter). Green cells indicate the presence of GFP within the cell.

5.1.6 Identifying the most efficient sgRNA

Having designed numerous sgRNAs and single stranded oligonucleotide donors for each planned cut site and having failed initially to prove that any of the combinations were successful at knocking in, I opted to assess the sgRNAs alone. Transfecting the spCas9 plasmid and individual sgRNAs using GeneJuice® as the transfection reagent, the cutting efficiency of the sgRNAs was assessed using Sanger sequencing of the PCR product. Visualisation of the chromatogram data was possible using SnapGene Viewer software (from Insightful Science; available at snapgene.com) and I used this to align the CRISPR DNA to the wild type reference DNA to identify mismatches and signs of sgRNA cutting. The wild type cells had also undergone transfection with spCas9 but without an sgRNA.

The webtool TIDE by Desktop genetics (tide.deskgen.com) was helpful in estimating the cutting efficiency of the sgRNA. Using the Sanger sequencing data from the sample and the wild type, TIDE is able to provide an assessment of genome editing of a target locus by quantifying the editing efficacy between them [126]. It thus provides a profile of all insertions and deletions in the edited sample and estimates the efficiency of the sgRNA to guide Cas9 to cleave the DNA at the designated target. Using this tool I was able to focus on the most efficient sgRNAs for further experiments. I was also able to use the webtool TIDE to optimise sgRNA:Cas9 ratio used in transfection. I found that increasing the ratio from 1:1 to 2:1 did not improve efficiency and actually reduced it, as seen in Figure 5.9C.

The seven sgRNAs used to target the sense and antisense strands of the *CCND1* enhancer all proved to be relatively inefficient, ranging from 1% to 12% with a sgRNA:Cas9 1:1 ratio. My positive control sgRNA, kindly provided by Professor Buluwela (Imperial College, London) had a TIDE estimated efficiency of 31% in the same conditions (Figure 5.9). The estimated efficiency is a reflection of transfection efficiency and indel formation within the pooled cells. With this method I was able to select the most efficient sgRNAs and discard those I thought unlikely to produce successful knock-in clones, such as sgRNA 127 with an estimated efficiency of less than 1%.



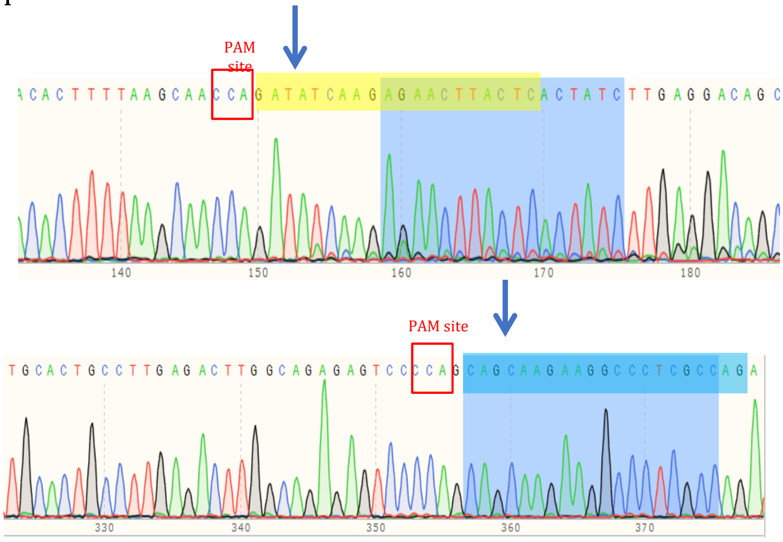
A Wild type control – MCF7 cells transfected with spCas9 but no sgRNA. This is the reference sequence.

B MCF7 cells transfected with spCas9 and 1000ng **positive control** sgRNA– TIDE estimated efficiency **31%**

C MCF7 cells transfected with spCas9 and 2000ng **positive control** sgRNA– TIDE estimated efficiency

Figure 5.9 CRISPR DNA chromatogram aligned to the A) wild type reference DNA using SnapGene Viewer software.

Increasing the concentration of sgRNA from B) 1000ng (sgRNA:Cas9 1:1) to C) 2000ng (2:1) did not increase the estimated transfection and indel efficiency as predicted by the deskgen TIDE webtool. Red box shows PAM site. Blue arrow predicted cut site.



A SnapGene view of sgRNA 107 in MCF7 pool showing overlapping peaks starting at the predicted cut site. sgRNA 107 sequence GATATCAAGAGAACTTACTC

B SnapGene view of sgRNA 127 showing no overlapping peaks starting at the predicted cut site. sgRNA 127 sequence CAGCAAGAAGGCCCTCGCC

Figure 5.10 DNA chromatograms of A) sgRNA 107 and B) sgRNA 127 using SnapGene Viewer software.

A) shows evidence of genomic sequence disturbance indicating successful cut in at least a proportion of the cells, whereas B) shows no sign of sgRNA cutting. Arrow indicates predicted cut site.

5.1.7 Improving transfection efficiency with the cell line MCF7/Luc

Due to the issues I was having with low transfection and indel formation I tried a variant MCF7 cell line called MCF7/Luc (kindly provided by Professor Buluwela) which in the hands of the providing lab group had proven easier to transfect. MCF7/Luc is derived from the same cell line MCF7 but stably expresses firefly luciferase gene and Neomycin resistant gene. It is estrogen responsive in the same way as MCF7 and testing previously performed by the providing lab had shown no difference in its genome to MCF7 apart from the known firefly luciferase gene and Neomycin resistant gene. Nonetheless I ensured E2 induction of known ER regulated genes GREB and CCND1 with RT-qPCR of the cell line prior to use and found their induction to be in keeping with my previous MCF7 work as seen in Figure 5.11 (comparative expression of MCF7-Luc and MCF7 not shown).

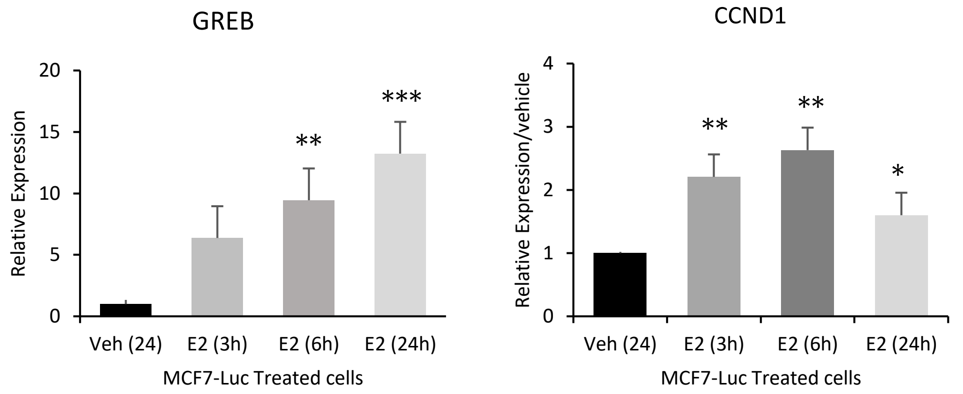


Figure 5.11 RT-qPCR of E2 regulated genes in MCF-Luc cell line.
MCF7-Luc cells were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours. Cells were then plated onto 6 well plates (day 0) with either vehicle (ethanol), or 17β estradiol for 24 hours. RT-qPCR was performed from RNA prepared from the treated MCF7-Luc cells at 0, 3, 6 and 24 hours and the vehicle treated at 24 hours. Data were normalised to GAPDH and expressed relative to vehicle. The mean and standard error (s.e.m.) of three biological replicates are shown *p<0.1, **p<0.01, ***p<0.001.

5.1.8 Identifying successful knock-in

Having identified the most efficient sgRNAs in MCF7/Luc cells and the best transfection method (GeneJuice®), I proceeded to attempt to knock-in my termination sequence using a single stranded oligonucleotide containing the polyadenylation sequence template. The final concentration found to be successful was 1000ng ssODN, using only one ssODN for each transfection, resulting in a 1:1:1 spCas9:sgRNA:ssODN.

I used a two round PCR method followed by gel electrophoresis to identify the presence of the knocked in sequence. Using a nested primer in the "second round PCR", the reverse primer being homologous to 21 bases within the polyadenylation sequence, (Figure 5.12A) I was able to identify the presence of the transcription termination sequence within the pool (Figure 5.12B).

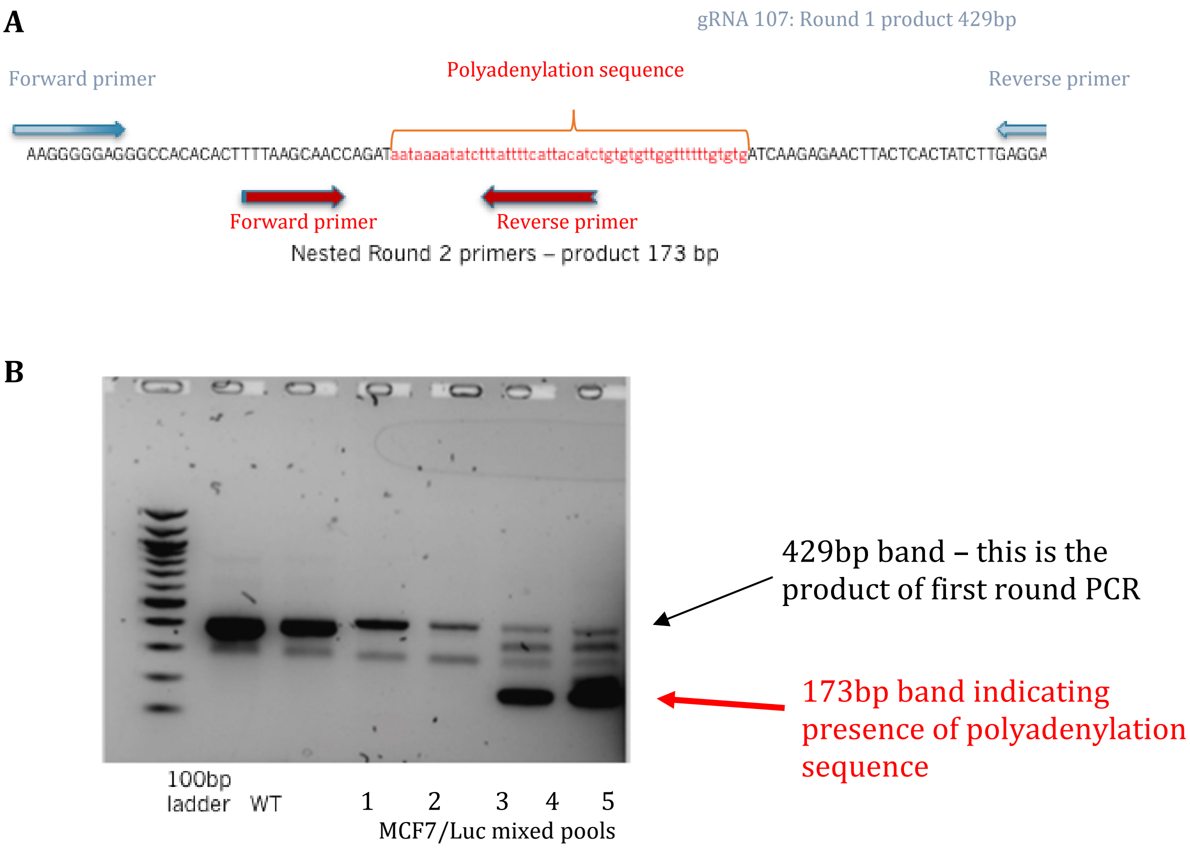

A
gRNA 107: Round 1 product 429bp
Forward primer
Polyadenylation sequence
Reverse primer
AAGGGGGAGGGCCACACACTTTTAAGCAACCAGATaataaaatatctttattttcattacatctgtgtgttggtttttttgtgtgATCAAGAGAACTTACTCACTATCTTGAGGA
Forward primer
Reverse primer
Nested Round 2 primers – product 173 bp
B
429bp band – this is the product of first round PCR
173bp band indicating presence of polyadenylation sequence
100bp ladder
WT
1
2
3
4
5
MCF7/Luc mixed pools

Figure 5.12 Identifying the presence of the knocked in polyadenylation sequence using nested primers.

A. Schematic representation of the primers used for first round PCR (blue arrows) and the "nested" primers (red arrows) used to identify the presence of the polyadenylation sequence (in red).

B. 1.5% agarose gel showing the presence of a 173bp band in MCF7/Luc mixed cell pools 4 and 5, indicating the presence of the polyadenylation sequence in some of the cells in the pool. The bands are indicative of the expected 429bp amplicon generated by the first round primers; and in the case of pool 4 and 5, the 173bp amplicon generated by the nested primer pair. A band indicating the presence of an amplicon approximately 385bp is likely to be the round 2 forward primer and the round 1 reverse primer.

5.1.9 Clonal selection of successful CRISPR mediated knock-in

It became apparent that if the transfection and knock-in efficiency was very low, and if the cells were passaged more than 2 or 3 times, the knock-in was actively selected out of the pool and hence the nested primer no longer detected evidence of the polyadenylation sequence.

Because of the recognised very low frequency of successful HDR, the identification of a single clone from a pool of cells without selection was unsuccessful. Plasmids encoding for Cas9 and selection markers, such as GFP and puromycin resistance have been used to enrich the population of successfully transfected cells and hence ease the identification of a successfully modified clone. Hence I used a spCas9-GFP plasmid (pSpCas9(BB)-2A-GFP (Addgene - PX458)) that allowed me to use fluorescence-activated cell sorting (FACS) to isolate only those cells successfully transfected with the SpCas9-GFP plasmid (although this did not indicate successful knock-in). Unfortunately, FACS proved unsuccessful due to significant amount of cell death prior to clonal expansion, presumably related to the toxicity of transfection and FACS when combined.

I subsequently tried the plasmid pSpCas9(BB)-2A-Puro (Addgene - PX459) which encodes spCas9 and a puromycin resistance cassette (puromycin N-acetyl-transferase which inactivates puromycin). I transfected the sgRNAs separately although it is possible for the plasmid to co-express the guide RNA as well. The use of this plasmid enabled me to select for cells successfully transfected with the plasmid using puromycin selection. I optimised the puromycin dosage using a puromycin kill curve and found that 1ug/ul was an adequate dosage as long as the cells were only treated for 48 hours before removal of puromycin (not shown). The switch to the pSpCas9(BB)-2A-Puro (PX459) plasmid resulted in a much more timely and efficient clonal selection method.

At the end of the optimisation steps above I was successful in identifying MCF7/Luc pools that exhibited evidence of the polyadenylation knock-in. The gel electrophoresis showed evidence of the knocked in polyadenylation sequence when using sgRNA 107 (which targeted the *CCND1*e(sense) enhancer) and using sgRNA 12 (which targeted the *CCND1*e(antisense) enhancer) and for each cut site I found that both the sense and antisense ssODN donors had been successful. TIDE analysis of the puromycin selected transfected cells showed an increased transfection and indel efficiency to 42% and 22% (from 12% and 8% respectively) such that I felt that clonal expansion of these mixed cell pools would be more likely to identify a successfully edited clone. The heterogeneous cell pools then underwent clonal expansion using cloning discs. From the CRISPR pools I plated over 2000 single cell colonies and subsequently checked each colony for presence of the amplicon indicating successful knock-in.

5.1.10 Identification of single clones harbouring the polyadenylation knock-in

From thousands of single cell clones arising from heterogenous pools of transfected MCF7/Luc cells, I identified two individual clones exhibiting evidence of successful incorporation of the polyadenylation sequence into at least one strand of DNA. Those clones, named B1 and A4, both came from the pool of MCF7/Luc cells transfected with pSpCas9(BB)-2A-Puro (PX459) conferring puromycin resistance, sgRNA 12 targeted at the *CCND1*e(antisense) enhancer and with insertion of the sense ssODN donor.

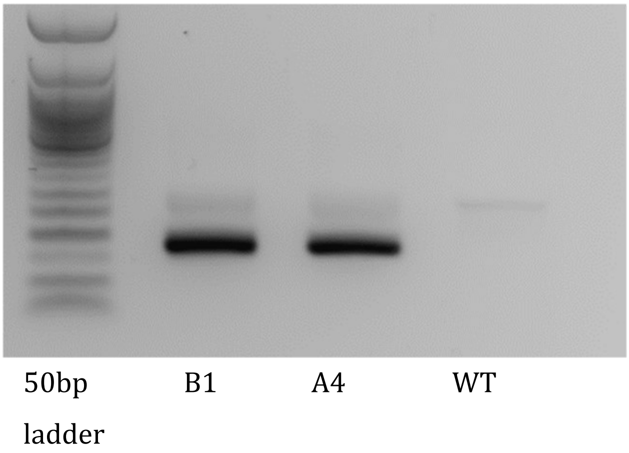


Figure 5.13 Gel electrophoresis showing amplicons of CRISPR mediated knock-in.
Clones B1 and A4: antisense enhancer targeted by sgRNA12. With successful knock-in, expected amplicon of 164bp is shown and not visible in wild type (WT).

Gel electrophoresis identified a 164bp amplicon in both clones B1 and A4, in keeping with the presence of the knocked in sequence (Figure 5.13). However, the DNA chromatogram from Sanger sequencing of the first round PCR product from both clones did not demonstrate evidence of the knock-in (Figure 5.14). The chromatogram for Clone A4 was identical to wild type throughout the length of the sequence with no suggestion of any indels. The Sanger sequencing results of Clone B1 suggested a more heterogenous population rather than a single cell clone, either suggesting clonal contamination and the presence of more than one clone, or the presence of indels in one or more strands of the DNA. This Sanger sequencing has been done on a single cell clonal expansion and hence all of the cells should have the same DNA sequence unless they have undergone mutation. However, it is known that HDR often only occurs on one strand of the DNA and the other may undergo NHEJ resulting in a heterogenous sequence on the chromatogram and this may be further complicated by the fact that the MCF7 cell line has been observed to have a highly rearranged karyotype containing 74 to 89 chromosomes per metaphase spread and more than 2 alleles may exist of the enhancer region at chromosome 11q13 [148].

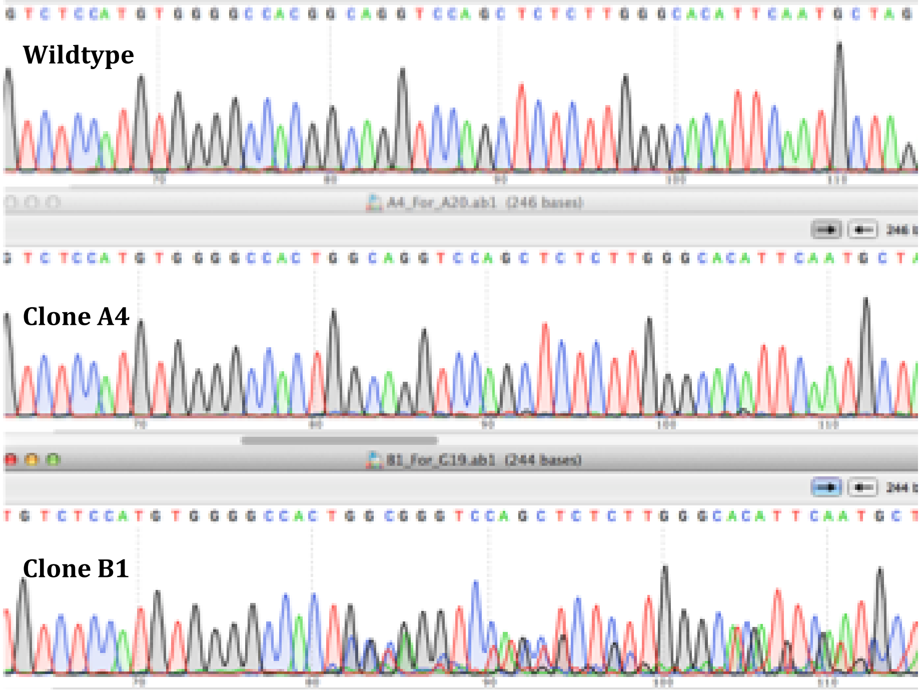


Figure 5.14 DNA Chromatograms visualised with SnapGene Viewer.
First round PCR product from wild type (WT), and clones A4 and B1.
Clone A4 shows the same sequence as wild type with no evidence of indels
Clone B1 shows atleast one allele has undergone a knock-in or indel at the expected site of DSB.

5.1.11 Proving the presence of the polyadenylation knock-in sequence

As the chromatogram was unable to prove presence of the knock-in, I sought alternative ways to prove the presence of the polyadenylation sequence at the target site.

5.1.11a Amplification of genomic DNA by PCR shows presence of knock-in

The reverse primer used for the second round nested PCR reaction was complementary to a sequence within the HDR donor template and thus should only be identified on genomic DNA PCR if the sequence has been knocked into the DNA. DNA from each clone was extracted and underwent amplification using PCR. I compared the expression of (part of) the polyadenylation sequence (the first half of which is found between the nested primers) in the sense enhancer region with its presence in the antisense region. As the successful sgRNA (sgRNA12) had targeted the antisense enhancer region I only expected to find the knock-in there.

Figure 5.15 shows the expression of at least part of the polyadenylation signal (that within the nested primers) in clones A4 and B1 at 80 times that found in wild type. The expression of the polyadenylation signal within the sense enhancer did not vary significantly across any of the clones or wild type.

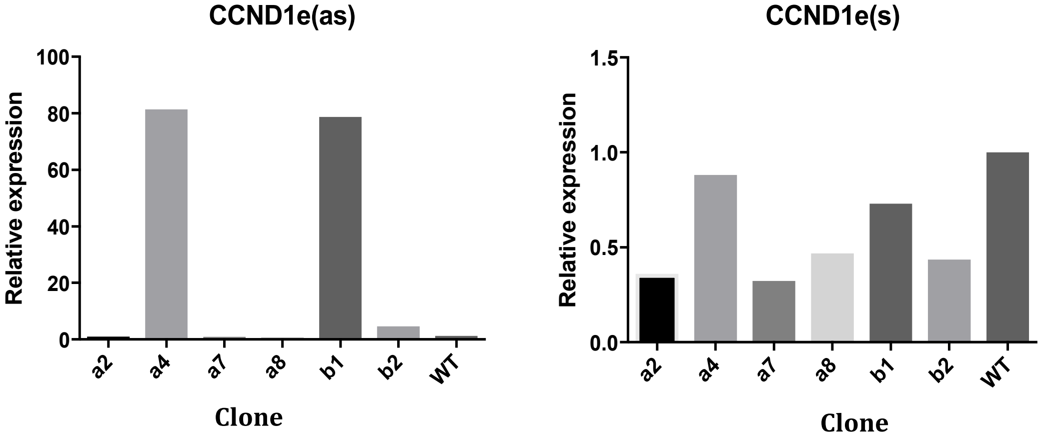


Figure 5.15 PCR of genomic DNA from CRISPR/Cas9 generated clones and wild type.

PCR amplification of the knocked in polyadenylation signal is seen in DNA from antisense enhancer clones B1 and A4 in keeping with gel electrophoresis (A). Its presence is not significant in any other CRISPR clone (a2, a7, a8, b2) or wild type (WT) and is not seen in genomic DNA from the sense enhancer (B). Data was normalised to genomic GAPDH and expression is relative to wild type.

5.1.11b Sanger sequencing of DNA amplicon from gel electrophoresis

To further verify the presence of the poly(A) knock-in into the antisense enhancer region of clones B1 and A4 I performed gel electrophoresis with elution of the DNA band of interest and subsequent sequencing of that DNA. Figure 5.16 shows the DNA chromatogram visualised with SnapGene Viewer. The first 36 bases of the poly(A) sequence can be seen at the expected site of DSB and HDR in both clones. The sequence is only cut short by the length of the amplicon.

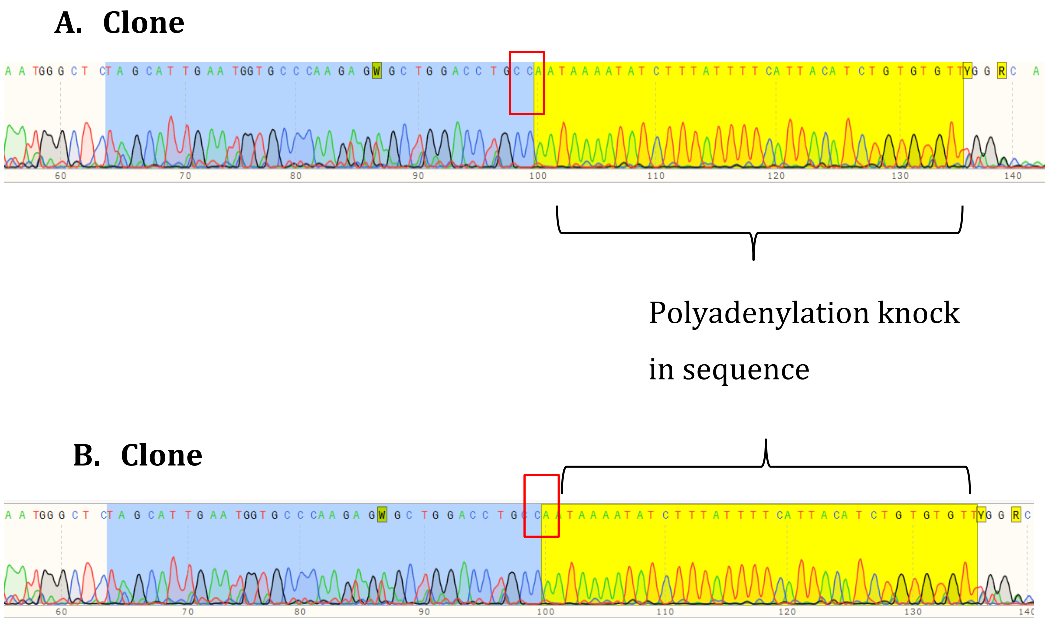


Figure 5.16 DNA Chromatogram visualised with SnapGene viewer of DNA sequence from the eluted DNA band from single cell clones following successful knock-in.

Clone B1 (A) and clone A4 (B). Highlighted in blue are the 32 bases of the upstream homologous recombination donor arm. Highlighted in yellow is the first 36 bases of the 49 base polyadenylation sequence (aataaaatatctttattttcattacatctgtgtgttggtttttgtgtg).

The GCC PAM of the gRNA are seen in the upstream donor arm inside of the red box.

5.2 Effect of knock-in of the polyadenylation sequence into *CCND1*e(antisense) on bidirectionally transcribed eRNAs and *CCND1* mRNA

Assessing the effect on transcript abundance of the eRNA *CCND1*e (antisense) and its bi-directionally transcribed partner *CCND1*e(sense), as well as the neighbouring gene *CCND1*, was an important first step in the assessment of the clones now proven to have a knock-in of the polyadenylation sequence.

I used RT-qPCR to compare differential expression in wildtype and clone B1 of CCND1, GREB and the enhancer RNA transcripts in the presence of estrogen at 6 hours. There was no statistically significant difference in expression in GREB OR the sense eRNA transcript.

To detect the antisense transcript, I designed primers downstream of the intended PAS and RT-qPCR proved a statistically significant reduction in expression of the antisense transcript in the B1 CRISPR clone when compared to wildtype. The same clone also showed a significant reduction in CCND1 expression in the presence of estrogen compared to wildtype, suggesting that the PAS knock in had resulted in less antisense eRNA expression and consequently less CCND1 expression (Figure 5.17).

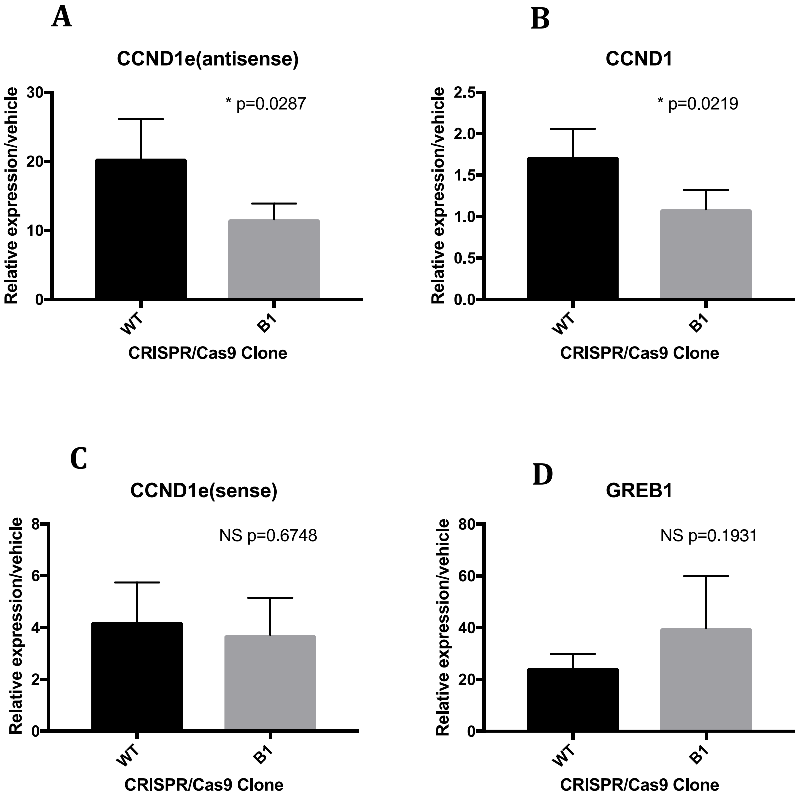


Figure 5.17 RT-qPCR of Wild type cells and Clone B1 following CRISPR mediated knock-in of PAS into antisense enhancer.

RT-qPCR was performed from RNA prepared from single cell clones WT and B1 which were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours prior to plating on 6 well plates with either vehicle or 10nM 17β estradiol. Cells were treated for 6 hours. Data was normalized to GAPDH levels and the expression of each gene is shown relative to expression of vehicle at 6 hours. The mean and s.e.m. of three biological replicates are shown; Unpaired student's t test: *p<0.1.

5.2.1 Cell cycle analysis of knock-in clone B1

Given the pivotal role of *CCND1* in the regulation of the cell cycle [84 89], I sought to assess the effect on the cell cycle of inserting the premature transcription termination sequence into the enhancer. Using the Millipore Muse™ Cell Cycle assay with the Muse™ Cell Analyzer I was able to quantitate the cell cycle phases of a cell suspension. The assay utilises the nuclear DNA intercalating stain propidium iodine (PI) which discriminates cells at distinct stages of the cell cycle and calculates the percentage of cells in each of the cycle phases. In this way I was able to show that following treatment with E2 for 24 hours, 52.61% of the knock-in clone B1 cells were still in phase G0/G1 compared to 45.35% of the wild type cells, and a lower percentage of B1 were in S phase (Table 5.1).

	% G0/G1	**% S**	**% G2**
Wild type Vehicle	58.8 ± 5.4	17.0 ± 1.3	24.2 ± 4.5
B1 clone Vehicle	59.6 ± 3.1	16.035 ± 0.7	24.35 ± 2.1
Wild type 24hr E2	45.35 ± 5.3	25.28 ± 0.7	29.36 ± 6.5
B1 clone 24hr E2	52.61 ± 4.9	20.42 ± 1.1	26.96 ± 3.6

Table 5.1 Percentage of all live cells in wild type and knock-in clone B1 held at each phase of the cell cycle following E2 treatment.
Cells initially underwent 72 hours of estrogen starvation followed by 24 hour treatment with either vehicle or 10nM 17β estradiol. Results calculated by Milliport Muse™ Cell Analyzer. Mean ± s.e.m. of three replicates.

5.2.2 RNA-Sequencing of polyadenylation knock-in clone B1 shows down-regulation of many estrogen regulated genes up-regulated in wild type clone.

To further characterise the effects of the knock-in of the premature transcription termination sequence, I performed stranded messenger RNA RNA-Sequencing of clone B1 and wild type cells starved of estrogen for 72 hours and treated with E2 for 6 hours using the Illumina TruSeq Stranded mRNA Library Prep kit.

The differential expression analysis of this RNA-Seq was undertaken by Dr Castellano and compared with the previous paired end RNA-sequencing of E2 treated MCF7 cells, as described in chapter 3. Differential expression analysis of E2 regulated messenger RNA (mRNA) in MCF7/Luc and the knock-in clone B1 was performed using the Bioconductor 3.12 software package. Correcting for multiple testing using the Benjamini-Hochberg adjustment, the p value (Fischer's exact test) for comparison was $<2.2\text{x}10^{-16}$. From a total of 17 650 estrogen regulated transcripts, excluding those with an FPKM less than 0.5, we found an overlap of genes that were significantly *up-regulated* by E2 at 6 hours in wild type MCF7/Luc cells but significantly *down-regulated* by E2 at 6hours in the polyadenylation insertion B1 clone (Figure 5.18).

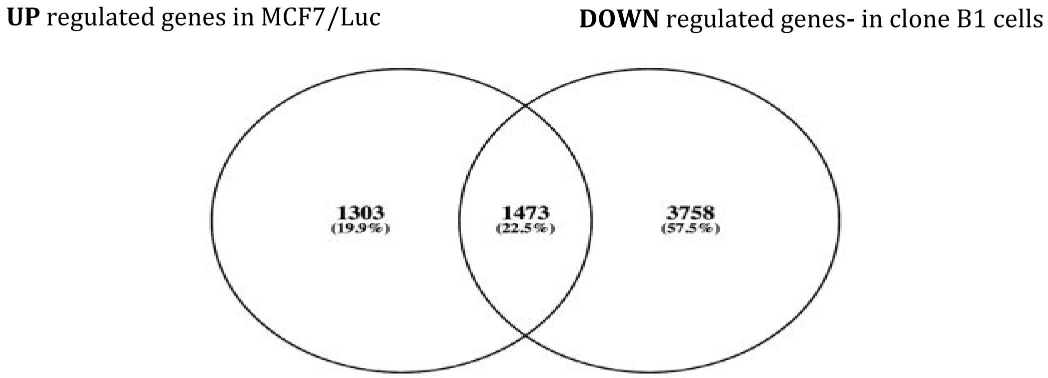

UP regulated genes in MCF7/Luc
DOWN regulated genes- in clone B1 cells
1303
(19.9%)
1473
(22.5%)
3758
(57.5%)

Figure 5.18 Venn diagram of differentially expressed genes B1 and MCF7/Luc.
Venn diagram representing the overlap of differentially expressed genes between E2 and treated MCF7/Luc cells at 6 hours and E2 treated B1 clone (poly(A) signal knock-in) at 6 hours from Total RNA RNA-Seq (left) and mRNA RNA Seq (right).

5.2.3 RT-qPCR validation of the RNA-Seq differential expression analysis

I conducted RT-qPCR for initial validation of the RNA-Seq discovery that, at 6 hours of E2 treatment, some ER up-regulated genes are down regulated in the

PAS knock-in B1 clone. Using the known ER regulated genes GREB1, TFF1, ERA and the miR-17-92 cluster, as shown in Figure 5.19, I compared the expression in CRISPR clones to expression in the wild-type line following 6 hours of estrogen treatment, as was the case for the RNA-Seq. I found that in most cases the differential expression was not statistically significant, with the exception of mir-17-92 which was expressed significantly less in both clones A4 and B1 compared to wildtype. Expression of TFF1 and ERα in clone A4 was also significantly less than in wild-type (Figure 5.19). Relative expression of estrogen regulated genes in these wild-type cells was validated with RT-qPCR of cells used in the initial MCF7-Luc validation time course (data not shown). This validation work needs further investigation in replicate and across the 24 hour timepoints.

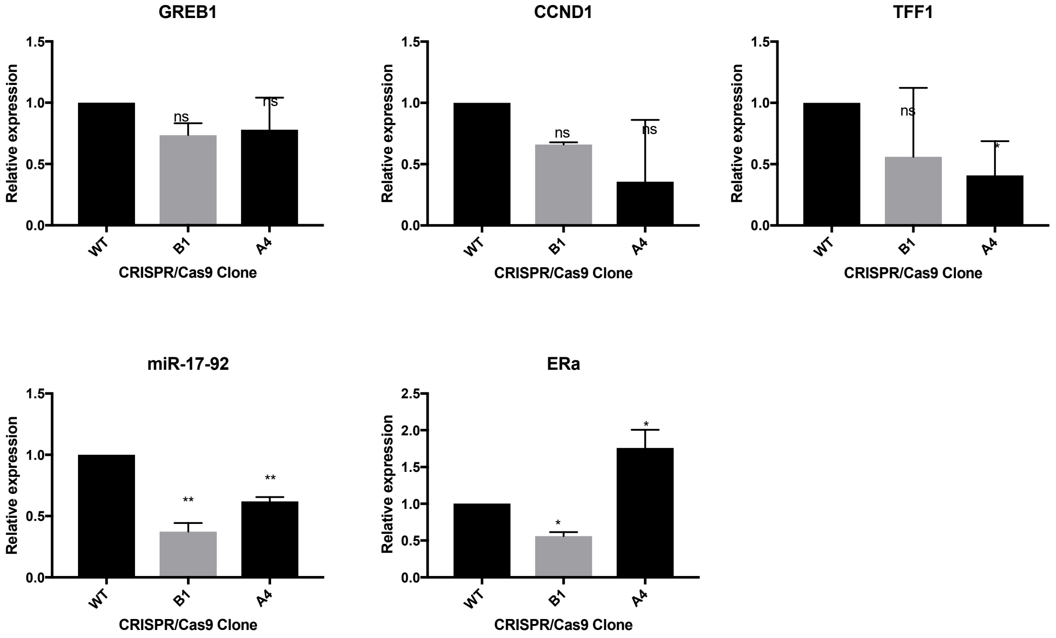


Figure 5.19 RT-qPCR analysis of known estrogen responsive genes in the knock-in clones B1 and A4 and the wild type cells.

Cells were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours. Cells were then plated onto 6 well plates (day 0) with either vehicle (ethanol) or 10nM 17β estradiol for 6 hours. RT-qPCR was performed from RNA prepared from the estrogen/vehicle treated cells. Data were normalised to GAPDH levels and the expression level of each gene is shown relative to expression of vehicle treated cells at 6 hours and expression in the wild type following E2 treatment. The mean and standard error (s.e.m.) of three biological replicates are shown *p<0.1, **p<0.01, ***p<0.001.

5.2.4 A dual-luciferase® reporter assay shows less ER activity in the knock-in clone B1 than MCF7/Luc

The RNA-Seq findings were surprising in the apparent global effect of premature transcription of the antisense eRNA arising from the *CCND1* enhancer. To better validate this I used a Cignal ERE Reporter Assay kit (Qiagen) with the dual-luciferase® reporter assay system (Promega) for a rapid quantitative assessment of the signal transduction pathway regulation in the B1 and wild type clones. The Cignal Reporter Assay consists of multiple repeats of the ER binding site and basic promoter elements to drive the expression of the firefly luciferase reporter gene. When the pathway is activated, the luciferase reporter activity is modulated and the change in activity is determined by comparing the normalized luciferase activity of the reporter between the clones. A negative control is a mixture of non-inducible firefly luciferase and constitutively expressing Renilla construct which serves as a specificity control. The positive control is a mixture of a constitutively expressing firefly luciferase construct and constitutively expressing Renilla luciferase construct and serves as a control for transfection efficiency. Figure 5.20 shows the relative luciferase activity in MCF7/Luc cells and in the B1 knock-in clone. There is significantly less relative luciferase activity in B1 reflecting a significant reduction in ER activity which can be used as readout for the activation status of the ER pathway. Such findings of down-regulation of the ER activation pathway in the B1 clone is in keeping with the results seen in the RNA-Seq analysis.

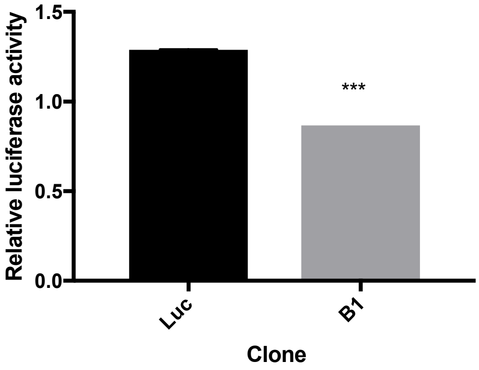


Figure 5.20 Assessment of activation of ER pathways in clone B1 and MCF7/Luc.
The Cignal ERE Reporter Assay kit (Qiagen) and the dual-luciferase® reporter assay system (Promega) were used for a rapid quantitative assessment of the signal transduction pathway regulation. The knock-in clone B1 shows significantly less relative luciferase activity than the MCF7/Luc cells indicating less ER activity in the knock-in clone. Mean and s.e.m. of duplicate experiments. Two-tailed students t test *p<0.5, **p<0.05, ***p<0.0005.

In light of the many questions that the RNA-Seq analysis has raised, there are many more experiments required to further investigate the finding that insertion of a premature transcription termination sequence into an enhancer of *CCND1* results in down regulation of multiple ER regulated genes, and these experiments are discussed in the next chapter.

5.3 CRISPR/Cas9 targeting ER binding site of enhancer

As previously discussed the aim of these CRISPR/Cas9 experiments was to terminate transcription of the eRNA early but not to interfere with the transcription initiation complex or with transcription factor binding at the enhancer. The latter had been achieved by Korkmaz et al [118] when they identified my region of interest as an enhancer of *CCND1*. They used CRISPR/Cas9 technology to target the ERE binding site at this enhancer and showed downregulation of *CCND1* mRNA as a consequence. From the outset it was my intention to use the sgRNAs reported by them [118] to compare the effects of ER binding interference (targeting ERE) to terminating transcription of the enhancer (early insertion of polyadenylation signal). However, despite multiple attempts to use their sgRNA sequences in my CRISPR/Cas9 system, I was unable to generate any cell models which showed successful interference of the ERE binding site.

I am thus grateful to the Agami group (Division of Oncogenomics at the Netherlands Cancer Institute) who very kindly provided two pool of cells from their CRISPR/Cas9 engineered clone in which they had mutated the ER binding site of the *CCND1* enhancer (Ag588) and two pools of their wild type clones (Ag WT). Unfortunately, during the time of this thesis preparation I was unable to reproduce the transcriptional activation of ER responsive genes upon estradiol treatment in their wild type cells for comparison (Figure 5.21) and as such this remains ongoing work.

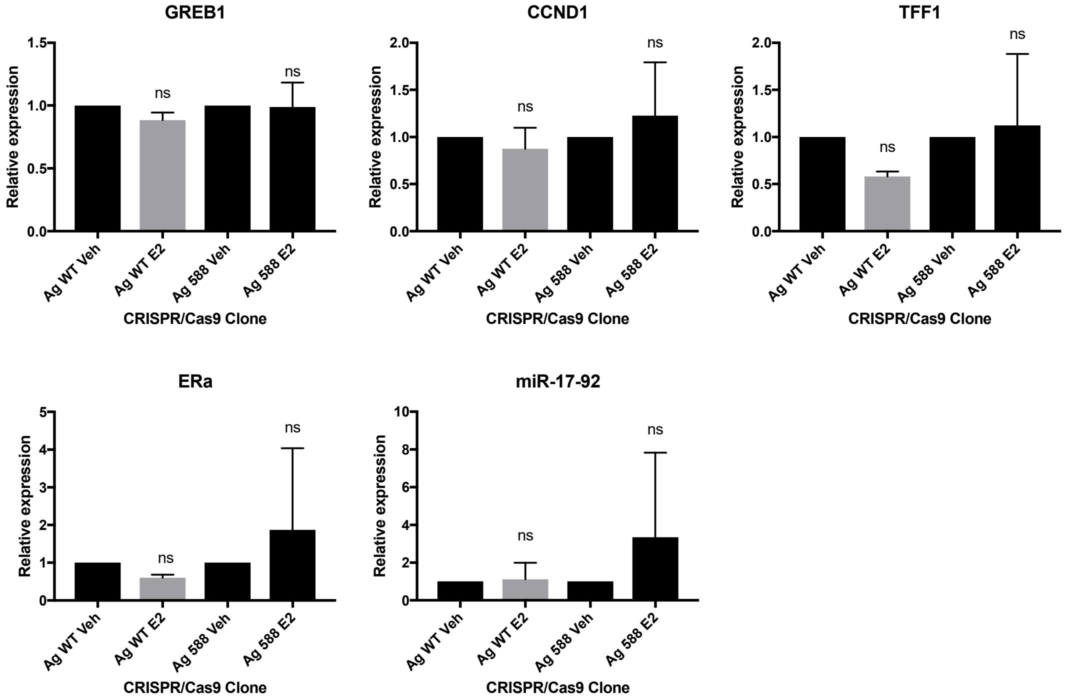


Figure 5.21 RT-qPCR analyses of estrogen response following CRISPR/Cas9 mutation of ER binding site at enhancer of interest.

Two cell lines kindly provided by the Agami lab: Ag588 cell line had undergone successful CRISP/Cas9 genomic editing of the TF binding site of the enhancer of *CCND1*. Ag WT is the wild type cell line provided.

Cells were cultured in phenol red free DMEM medium supplemented with 10% DSS for 72 hours. Cells were then plated onto 6 well plates (day 0) with either vehicle (ethanol) or 10nM 17b estradiol for 6 hours. RT-qPCR was performed from RNA prepared from the estrogen/vehicle treated cells. Data were normalised to GAPDH levels and the expression level of each gene is shown relative to expression of vehicle at 6 hours. The mean and standard error (s.e.m.) of three biological replicates are shown *p<0.1, **p<0.01, ***p<0.001.

5.4 Summary

Over the course of the work conducted and reported in these results chapters I have identified an enhancer region neighbouring the *CCND1* gene, in a region of high susceptibility to breast cancer. I have shown that this region is transcribed in a bidirectional manner and regulated by the ER.

To further characterise the nature of these transcripts I have identified their cellular location and performed loss and gain of function experiments. Knockdown, using siRNA and ASO technology, of the transcripts arising from the CCND1 enhancer region does not affect *CCND1* expression. To better determine if either the act of transcription along the enhancer or the transcript itself was required for CCND1 expression, I harnessed the genome editing technique of CRISPR/Cas9 and optimised a protocol in which I was able to insert a 21nt polyadenylation signal at a precise location within a short distance of the antisense transcriptional start site, whilst avoiding the known transcription factor binding sites at that region. There were many steps of optimisation in both identifying the site for insertion, transfection methodology and clonal screening and expansion. Ultimately, I have shown that introducing a polyadenylation sequence early into the transcribed enhancer region, which likely brings about premature transcription termination, knocks down *CCND1* expression with consequent reduced cellular proliferation, similar to that seen in CCND1 knockdown with siRNA. In addition to reduced CCND1 expression in the PAS knock-in clones, RNA-Seq analysis suggests that premature termination of transcription at this antisense enhancer of CCND1 may also have broader effects on other ER regulated genes. RNA-Seq showed a widespread knock down of ER regulated genes within the clones when compared to wild-type cells, and this work continues to be validated. This is discussed in the next chapter along with planned future work.

Chapter 6: Discussion

6.1 The estrogen regulated long non-coding RNA transcriptome

The genome regulated by ER is extensive and undoubtedly is not limited to protein coding genes and microRNAs. Long non-coding RNAs (lncRNAs), defined as non-coding transcripts greater than 200 nucleotides in length, encompass a diverse group of non-coding transcripts, and whilst many thousands have been identified, only a handful have been afforded functional characteristics. The importance of estrogen (E2) and the estrogen receptor (ER) in breast cancer in both its prognosis and its treatment is well recognised and many studies have been reported describing the effect of the ER on the coding genome, and more recently on the non-coding transcriptome[3, 76, 149]. However, at the outset of this work little was known about the regulation of lncRNA transcription by estrogen or the estrogen receptor.

Whilst RNA-Seq is the gold standard method in transcriptomics, the first generation of RNA-seq was unable to identify strand specificity of transcripts and thus it was difficult to accurately quantify gene expression for genes with overlapping genomic loci transcribed from opposite strands. However, with the use of the stranded RNA-seq library preparation kit (Illumina), I was able to maintain the strand orientation of the RNA transcripts and identify both sense and antisense transcription, much of which was previously thought to be biological or technical “noise”. This stranded RNA library incorporates adapter ligation in a predetermined orientation to the ends of the first strand of cDNA molecules, and in doing so provides the necessary information to orientate the RNA for sequencing.

I used this stranded RNA-sequencing technology to identify lncRNA transcripts regulated by E2 over 24 hours in the ER positive breast cancer cell MCF7 with the hope to better understand their relevance in the regulatory network. The MCF7 breast cancer derived cell line is a model system for studying estrogen signalling because of its substantial levels of ER and the considerable body of publicly available data describing its transcription factors, cofactors and histone marks both with and without estrogen treatment. The stranded RNA-Seq technique was important in our identification of enhancer RNAs arising from the enhancer region of *CCND1* as their bidirectional and divergent transcription was only visible due to the strand-specific technique.

6.1.1 Identification of E2 regulated unannotated lncRNAs in MCF7

I first sought to explore long non-coding RNA expression in MCF7 cells over a 24 hour time course after stimulation with E2. Using the accepted cut off of 200nt as the discriminator of long non-coding transcripts, I identified 1147 unannotated lncRNA transcripts that were significantly modulated by exposure to E2 over the time course. The RNA–seq data was also analysed to look for coding RNA and these differentially expressed genes were mapped to both the GO database and KEGG pathway analysis and, as expected, the most highly up regulated biological processes were those associated with cellular proliferation and passage through the cell cycle, whilst those most down regulated were associated with cell-cell adhesion. These findings correlate with what is already known about E2 stimulation in ER positive breast cancer cell lines and goes some way to validate my RNA-Seq analysis of ncRNA.

I further validated the RNA-Seq data using RT-qPCR to show a similar differential expression in well-recognised ER regulated genes and these mirrored my RNA-Seq analysis. Of the 1147 unannotated lncRNA that I identified as being regulated by E2, I looked for those with an expression level that would make further analysis feasible. Long non-coding RNAs, although pervasively transcribed, often exist at very low copy number per cell and as such, identification with in vitro techniques such as RT-qPCR can be problematic. Indeed, of the nine lncRNAs initially identified by the RNA-Seq for further investigation as being significantly differentially expressed, validation of their differential expression proved impossible in seven, which was likely related to their low abundance. I used numerous reverse transcriptase enzymes to improve the sensitivity of RT-qPCR and spent a long time optimising DNase treatment of RNA to ensure its complete eradication, but ultimately were able to validate only a few transcripts with reproducible expression. This could be a combination of inadequate RT, incomplete DNase treatment and poor primer design but it is likely to be exacerbated by the low copy number of the transcripts themselves.

Nevertheless, my RNA-Seq analysis enabled the identification of a number of E2 regulated ncRNA transcripts arising from regions bearing the hallmarks of enhancer regions. When it was first discovered in 2010 that enhancer regions are actively transcribed[49,50] it was assumed that the pervasive transcripts generated were simply the by-products of enhancer transcription, and it was the gathering of transcription factors and co-activators and physical proximity through looping to promoters that resulted in the enhanced regulation of neighbouring genes. However, although there are many thousands of eRNAs which have yet to be ascribed functional characteristics, several important mechanisms have since been reported, including transcription factor "trapping" [150], recruitment of proteins required for chromosomal looping[60, 71] and involvement in histone modifications[66, 151].

Whilst others have recently described the global increase in eRNA transcription at enhancers adjacent to E2 regulated coding genes[60], my unsupervised hierarchical clustering analysis of E2 treated MCF7 cells was surprising in that it showed evidence of both spliced and unspliced eRNAs with examples of alternative splicing and multiple eRNA isoforms. In very recent years, others have shown that splicing of multi-exonic RNAs arising from enhancers are associated with high enhancer activity[152] and the related chromatin modifications, DNase I accessibility and enrichment of p300 binding. It is possible that enhancer activity can be augmented through the process of splicing as it does in the coding genome[153] where elements of the spliceosome interact with and enhance initiation and transcription by RNA pol II. Or splicing could help with the chromatin modifications required for the subsequent recruitment of TFs and cofactors and activation of a previously poised enhancer. The bioinformatical identification of so many spliced E2 regulated eRNAs in MCF7 cells will add to this small body of literature and our ongoing investigation of one alternatively spliced lncRNA existing in two isoforms will hopefully contribute further. The reason for eRNA splicing remains unclear although interestingly recent evidence showing conservation of splicing related motifs and a lack of exonic constraint[154, 155] would suggest that it is a regulated and conserved process rather than a mere by-product of inconsequential transcription and our findings of alternative splicing and presence in the cytoplasm gives further credence to the argument that eRNAs are functional and not merely transcriptional noise.

6.1.2 Identification of a transcribed enhancer of *CCND1* which is upregulated by E2 in MCF7

One of the bi-directionally transcribed up regulated non-coding lncRNA transcripts identified by RNA-Seq analysis was of particular interest as it was transcribed from a region recognised as a susceptibility locus for hormone receptor breast cancer[156] and in close proximity to the *CCND1* gene on chromosome 11q13 which is overexpressed in up to 50% of breast cancers [89]. Notably the transcribed antisense strand is the site of a single nucleotide polymorphism (SNP rs614367 11q13.3 *intergenic*) associated with ER positive disease[135].

On analysis of publicly available data sets I was able to determine that the region from which the divergent transcripts arose bore many of the hallmarks of an ER bound enhancer [41, 44, 60] (Gene Expression Omnibus GSE 45822). I found p300 co-activator protein and Mediator binding at the bi-directional transcriptional start site of the enhancer as well as flanking peaks of H3K27ac and a high ratio of mono-methylation of histone H3 Lysine4 to tri-methylation (H3K4me1:H3K4me3), all in keeping with the recognised features of an active enhancer [41 43]. Analysis of our RNA-Seq data showed that the enhancer region of interest approximately 150kb 5' to *CCND1* was being transcribed, likely many other enhancers, in a bidirectional manner, generating two individual divergent transcripts and thus they were named *CCND1*e(sense) and *CCND1*e(antisense), relating to the strand from which they were transcribed. Whilst several other ncRNA transcripts were also initially investigated, it was this enhancer region and these divergent eRNA transcripts that generated much of the work described in this thesis.

Although originally described by Kim et al [50] as non-polyadenylated bidirectionally transcribed RNA transcripts of less than 2kb that originate from active enhancers, it has since been shown that some enhancer RNAs are polyadenylated and can be unilaterally transcribed. Generally those that are polyadenylated tend to be longer, transcribed in only one direction and from more active enhancers than those that are non polyadenylated [57]. More recently, reports show that eRNAs are in fact highly heterogenous with any number of combinations of direction, length, splicing and polyadenylation status [64] [157] [47]. Indeed, in 2019 Kouno et al[158] observed through single cell CAGE sequencing that whilst on a bulk level enhancers can be bi-directionally transcribed, on a single cell level enhancers are almost exclusively unidirectionally transcribed from either strand, further complicating our understanding of these enigmatic transcripts.

Our RNA-Seq data shows that both the mono-exonic sense transcript and the multi-exonic antisense transcript are upregulated within 3 hours of estrogen treatment, with expression peaking around 6 hours. This is confirmed in GRO-Seq data (NCBI's Gene Expression Omnibus (Accession no. GSE27463)) [159] in which the bidirectionally transcribed nascent eRNAs can be seen at 40 minutes following ligand treatment. The expression of eRNAs from this enhancer closely correlate with expression of its neighbouring gene *CCND1*, as shown in vivo in our sequencing data as well as analysis with The Atlas of Noncoding RNAs in Cancer (TANRIC; http://bioinformatics.mdanderson.org/main/TANRIC) and in vitro with RT-qPCR. This is in keeping with findings on a genomic scale, where eRNA transcription and the mRNA of neighbouring coding genes are correlated[39] [50] [60] [40, 52].

6.1.3 Cellular location of eRNAs arising from the enhancer of *CCND1*

The location of RNA in a cell can determine the outcome for that transcript; whether it be translated, preserved, modified or degraded. To better understand the potential function of the E2 regulated eRNAs arising from the enhancer of *CCND1* I sought to confirm their cellular location following induction. Published

RNA-Seq data of fractionated MCF7 cells already suggested a presence within the cytoplasm [137] (available from NCBI's GEO accession GSE63189) and I hoped to validate this. In light of the finding that thousands of non-coding RNAs had been found tethered to the chromatin adjacent to active genes [160], I had hoped to split the cell into three compartments (nuclear, cytoplasm and chromatin tethered) for further analysis of cellular location. Attempts to separate the chromatin from the nucleus proved challenging, in part due to the very sticky nature of chromatin, but also likely because the low copy number of eRNAs make their identification difficult. Hence, I settled for a two compartment method, and used 3 RNA controls whose cellular distribution have been well validated; MALAT1, an E2 regulated lncRNA found almost exclusively in the nucleus[161], GAPDH mRNA found equally distributed across both compartments[162] and FOXC1e eRNA which has been shown to be present in both the cytoplasm and the nucleus within 1 hour of E2 stimulation, but significantly diminished by 6 hours[60] .

RT-qPCR analysis of the fractionated MCF7 cells showed that the sense transcript *CCND1*e(sense) was found almost exclusively in the nucleus at both 1 and 6 hours following E2 treatment. However, the divergent transcript *CCND1*e(antisense) was found to be equally distributed between the nucleus and the cytoplasm at both 1 and 6 hours. RNA fluorescence in situ hybridization (RNA-FISH) is a method used to track and visualize a specific RNA by hybridizing labelled probes to the target RNA. RNA-FISH localisation using confocal staining validated my RT-qPCR findings (and that of published data [137]) of *CCND1*e(antisense) presence in the cytoplasm following E2 induction and a difference in the cellular distribution of the divergently transcribed eRNAs *CCND1*e(sense) and *CCND1*e(antisense).

Most eRNAs reported to date have been located in the nucleus where their function and mechanism of action remains an area of active research. Cytoplasmic long non-coding RNAs are usually involved in post-transcriptional control through the regulation of stability[20, 163] , decay[19] or translation[164] [165] of mRNAs. However, little is currently understood about the roles of cytoplasmic eRNAs or their propensity for forming protein RNA complexes but their presence in the cytoplasm suggests a role outside of the reported *cis* functions of chromosomal looping and transcription factor trapping. As I have shown, the antisense eRNA arising from the *CCND1* enhancer region is detected in the cytoplasm up to 6 hours following E2 induction indicating an ability to evade degradation and an active shuttling of the transcript out of the nucleus and thus I propose that this eRNA transcript is likely to perform a cellular function within the cytoplasm. Mass-spectrometry of the eRNA *CCND1*e(antisense) in the cytoplasm has been carried out by the host lab with the hope of uncovering a cytoplasmic function through the identification of its RNA bound protein partners. To date, such work has yet to definitively conclude any such function but is further discussed in the thesis of my lab colleague Dr Cathcart [138].

Mass spectrometry of other nuclear eRNAs has shown the binding of tens of proteins, many of which are known to function in chromatin remodelling and gene regulation [151, 157] but little has been reported on the protein interactions of enhancer RNAs in the cytoplasm. It is possible that eRNAs play different roles according to cellular location, like the long non-coding RNA HOTAIR which is known to participate in several different processes of normal cell development. The HOTAIR transcript has been shown at the chromatin level to act as a scaffold in the binding of the chromatin modifiers polycomb repressive complex 2 (PRC2)[15, 25, 26] and LSD1[27] thereby exerting significant repressive control over gene expression. However, in the cytoplasm, HOTAIR has been shown to act as a competitive endogenous RNA and regulate gene expression through interactions with a range of microRNAs[28, 29] including miR-7, which inhibits cell migration and invasion[30]. Its overexpression is associated with increased metastatic potential and poor survival in breast cancer, in part, through its interactions with miR-7 and thus its role in the cytoplasm is crucial to normal cell control. Whilst eRNAs have yet to have such location dependent functions identified, the presence of *CCND1*e(antisense), a spliced antisense multi-exonic enhancer RNA in the cytoplasm 6 hours following its induction raises the possibility that it plays a functioning role in the cytoplasm and is not solely involved in *cis*-mediated actions.

6.2 siRNA knockdown of either enhancer RNA did not affect transcription of *CCND1*

Whilst it has been shown by many that eRNA levels demonstrate a positive correlation with nearby genes[39, 40, 50, 52, 60], the functional role of the eRNA in gene regulation is still unclear. There are numerous examples in which knockdown of an individual eRNA results in reduced expression of the target gene[60-65] suggesting that their presence is important, but such findings cannot be attributed to the whole class of enhancer RNAs. Several studies have shown that knockdown of the eRNA negatively affects the enhancer:promoter looping [166-169], and others have shown that looping and gene transcription can be downregulated by preventing the release of nascent eRNA transcripts from RNA Pol II by reducing the Integrator protein nuclease activity. On the other hand, it has also been reported that enhancer:promoter looping can be maintained even when eRNAs are knocked-down or when eRNA transcription in inhibited by the chemical flavopiridol [61, 76] and so what remains unclear is how essential eRNA are and through what mechanism they might have any role in the stabilisation or recruitment of complexes to their gene targets and to enhancer functionality as a whole.

The cellular location of an individual transcript is known to affect the efficiency with which it can be knocked down; nuclear lncRNAs are more effectively suppressed using antisense oligonucleotides (ASOs including GapmeRs) whereas cytoplasmic transcripts are better silenced by siRNAs which utilise the endogenous cytoplasmic RNA interference machinery. Hence in my attempts to define a relationship between the eRNA transcripts and that of *CCND1* expression, I used both techniques to knock down the eRNA. I was successful in achieving statistically significant knock down of the antisense transcript using siRNA technology at relatively low concentrations (thereby limiting the off-target effects and activation of the innate immune system which can result in immunostimulation often misinterpreted as interference), presumably because of its cytoplasmic distribution. However, I found that statistically significant knockdown of the antisense eRNA transcript did not result in a reduction in expression of *CCND1*.

Knockdown of the exclusively nuclear eRNA *CCND1*e(sense) was more challenging despite using both siRNAs and ASOs at a range of concentrations and with a variety of cationic lipid transfection reagents, exposure time and cell culture conditions. Forward and reverse transfection methods were also used. GapmeRs (ASOs with melting temperature enhancing modifications which increase target binding affinity) were designed to target the eRNAs but ultimately the only successful knockdown of the nuclear eRNA *CCND1*e(sense) was achieved using a single siRNA. As with *CCND1*(antisense), although I experienced more difficulty in achieving significant knockdown of the sense transcript, I did not see an effect on *CCND1* mRNA expression. In keeping with this finding, I saw no phenotypic changes in cell cycling or proliferation following *CCND1* eRNA knockdown. Instead I found that siRNA knockdown of *CCND1* mRNA resulted in increased expression of the enhancer RNAs, suggesting a feedback loop between the gene and transcription of its enhancer. Such a mechanism could exist to keep expression of a gene within a limited range but the mechanism of such a loop remains uncertain. The transcriptional machinery would have to detect the cellular levels of mRNA and not just its level of transcription as has been proposed elsewhere [170] because siRNA acts post transcriptionally.

Interestingly, our overexpression experiments of the eRNAs *CCND1*e(sense) and (antisense) have also not shown an increase in *CCND1* expression suggesting that the transcript arising from this enhancer region may not in itself be important for regulation of the *CCND1* gene. I note others have reported that exogenous overexpression of eRNAs has been shown to increase target mRNAs [47] [62] [157] and thus I questioned if the active region identified was actually an enhancer of a more distant coding gene.

I was however encouraged to investigate further because whilst this body of work was being undertaken, the enhancer region identified by us was reported by others as being an enhancer region for *CCND1*. They [118] reported that mutating the ER binding site at the transcription start site of this enhancer resulted in significant downregulation of *CCND1*. I hypothesised that this downregulation could be caused by the consequent blockade of other TFs and coactivators (in addition to ER) binding at the enhancer region; prevention of the initiation of transcription by RNA Pol II and possible impact on enhancer: promoter looping. Hence I wondered if I could determine if the act of transcription by RNA Pol II was pivotal in the enhancer's role and sought to stop transcription of the enhancer with the insertion of a transcription termination sequence.

6.3 Using CRISPR/Cas9 to investigate the role of enhancer transcription

CRISPR/Cas9 has emerged in recent years as a powerful technique for genome editing. Cas9 nuclease cuts the DNA at a specific target sequence resulting in a double stranded break (DSB) which the cell then tries to repair using either non homologous end joining (NHEJ) or homology directed repair (HDR) if a donor template is available. NHEJ occurs more frequently within the cell but often introduces insertions and deletions which can result in missense mutations and consequent gene knockouts. On the other hand, HDR with a template donor involves the recombination of sequences with homologous ends and is thus capable of knocking in a specific pre-designed sequence. Having shown that knockdown of my eRNA transcripts did not affect nearby *CCND1* gene regulation, I chose to use CRISPR/Cas9 technology to try to stop the active transcription of the enhancer with the hypothesis that the act of transcription at the enhancer plays a role in its enhancer function.

Clearly the control of gene regulation is highly complex and multi-factorial and enhancers play only one role in the machinations. Enhancer RNA transcription occurs early in the gene transcription process and prior to mRNA [61 71 171 172] firstly through the recruitment of activated transcription factors which bind to the enhancer and promote nucleosome remodelling and further TF and cofactor complex binding. Briefly, the binding of these complexes such as P300 and Mediator enable histone modifications such as H3K27ac which result in a more open chromatin and additional protein binding including the recruitment of RNA Pol II. RNA Pol II is subsequently phosphorylated and through binding of additional proteins, such as BRD4, initiates transcription. Following its elongation, Pol II encounters polyadenylation signals shortly downstream of the TSS resulting in the recruitment of the polyadenylation machinery and Integrator and subsequent termination of transcription. RNA Pol II transcription is controlled in part by regulated pause mechanisms which are triggered initially by sequence-specific interactions between the DNA, RNA and Pol II and this transcriptional control at enhancers is thought to be less stable and more prone to early termination than when transcribing genes [55, 173].

I sought to bring about very early transcription termination in the *CCND1* enhancer through CRISPR mediated knock-in of a 49bp polyadenylation sequence (PAS)[122] to explore the possibility that RNA Pol II elongation at the enhancer played a role in *CCND1* gene regulation. I hoped to differentiate from other work[118] in which ER binding at this enhancer was prevented by CRISPR/Cas9 mediated disruption of the binding site by targeting the knock-in to occur outside of the recognised binding sites of known TFs in this region.

At the time of conducting this work, reports of successful CRISPR/Cas9 mediated knock-in were relatively scarce[174, 175]. The knock-in technique itself is problematic due to the low frequency of HDR repair, with efficiency reportedly ranging from 1-10%[176-178]. Because of the HDR requirement for a donor DNA template at the cleavage site, the cell will usually and preferentially use NHEJ to repair double strand breaks. Even when HDR editing has successfully occurred, others have reported the corruption of the knocked in sequence by unwanted insertions and deletions on the same allele, probably due to concomitant NHEJ repair. In some reports 90% of HDR edited alleles also had insertions or deletions, making identification of a successful uncorrupted knock-in almost impossible. To make matters even more challenging, CRISPR/Cas9 is sometimes bi-allelic, meaning that even if one allele is accurately edited, the other allele may contain unwanted indels, with the chances of knock-in into both (or more) alleles very low unless the targeting efficiency and HDR efficiency are both very high. Given the genetic instability associated with cancers, this becomes even more complicated when considering cancer cells exhibiting more than two alleles. The MCF7 cell line is reported to range from hypertriploidy to hypotetraploidy [148], and hence the enhancer region I hoped to knock-in to may exist in multiple alleles, adding further difficulty to identifying a successful knock-in.

6.3.1 Optimising CRISPR/Cas9 for knock-in of a polyadenylation signal

However, since the more widespread use of CRISPR/Cas9 in recent years, numerous approaches have been reported to improve HDR efficiency and increase the chances of successful knock-in. The strategies are primarily aimed at optimising sgRNA design and judicious choice of a DNA donor template format as well as techniques to enhance donor template delivery to the target site and shifting the balance within the cell from NHEJ to HDR.

6.3.1a Design of guide RNAs

Although in theory CRISPR/Cas9 is highly specific, off target effects are likely to be common and the design of sgRNAs is an important step. Computer algorithms are able to assist in creating a scoring algorithm based on predicted off target binding and on target Cas9 cleavage efficiency and I used two independent programs to design the sgRNAs. I examined the potential off target sequences to determine if they were within annotated genes and subsequently tested several of the highest scoring gRNAs appearing in both programs. The number of guides generated by the programs was relatively small because I wanted to limit the site of nuclease activity to a specific, limited region of the enhancer; the aim being to cut the genome and insert the transcription termination sequence as early into the transcribed enhancer as possible but not to interfere with TF binding or Pol II initiation. Hence I used two 400bp sequences of the enhancer genome (one from each strand) for guide RNA design. I subsequently sought to evaluate the efficiency of each of the sgRNAs in MCF7 using the most effective and least toxic transfection reagent (GeneJuice®) and used sanger sequencing and the webtool DESKGEN® to estimate the efficiency of the guides. I was disappointed to find that in an unselected MCF7 pool the efficiency of the sgRNA and Cas9 system used was 12 % or less. A positive control guide RNA used in the same system was estimated at 31%. In recent times, the general recommendation has been to use guide RNAs with an efficiency of at least 25% as the more DSBs generated by Cas9, the more cut sites available for repair with HDR. It is possible that widening the genomic region for knock-in could have generated more efficient guides in my model or alternatively I could have used multiple overlapping sgRNAs sharing at least 5 base pairs which have been shown to enhance HDR efficiency [179]. Interestingly, Graf et al[180] recently reported that the sequence motif in the four PAM proximal bases of the targeting sequence can be critical in the efficiency of the sgRNA and claim that two short motifs (TT- or GCC-) can result in a 10 fold reduction in gene knock out efficiency. sgRNA 13 targeting the antisense strand gas a GCC- motif and had an in vitro efficiency of 0% and was discarded. In time, such findings will be incorporated into the computer algorithms which assist in the design and ranking of sgRNAs and thus knock-in efficiency will continue to improve.

6.3.1b Consideration of the Cas9-sgRNA delivery method

The Cas9 nuclease and the gRNA components required for DNA cleavage can be delivered to the cell in a number of ways. For transient transfections, the guide and the nuclease can both be delivered as plasmid DNA, RNA or as a pre-complexed ribonucleoprotein (RNP) composed of the sgRNA and Cas9 nuclease. Like many at the time, I opted for a CRISPR/Cas9 protocol using a plasmid based technique in which the sgRNA and Cas9 were integrated into and delivered in separate plasmids.

Plasmids may not however be the best delivery method due to their ability for plasmid transfection, when all or part of the plasmid DNA is integrated into the genome (possibly outside of the target region so goes undetected). In addition, the plasmids are not rapidly degraded and as such have been reported to result in more off target effects than other delivery methods, presumably because the DNA is exposed to the cutting machinery for longer. Furthermore, plasmid dosage is inversely related to cell viability so the CRISPR may be limited to lower concentrations of Cas9 and gRNAs. At the outset, I found significant cell death when attempting transfection of the gRNA and Cas9 plasmids with electroporation despite altering many of the variables including voltage, time in cuvette and total plasmid concentrations. I subsequently switched to transfection reagents and using a GFP expressing plasmid as a surrogate marker of effective Cas9 and sgRNA transfection, I found GeneJuice® transfection reagent (Merck Millipore) to be the least toxic and most effective.

Having initially failed to identify any evidence of knock-in on nested PCR of pooled cells, I looked to optimise other variables such as increasing the ratio of sgRNA:Cas9:ssODN to 2:1:1 which had no demonstrable positive effect (although I note others have reported successful CRISPR at ratios of 10:1) and indeed, appeared to reduce the cutting efficiency of the positive control. Of note, when trying to transfect using electroporation I also tried a variety of different ratios from 1:1:1 up to 100:100:1 and even 1:1:2 but I believe the transfection method was too toxic for any of them to be successful. In order to streamline the transfection with so many variables, when using GeneJuice® transfection reagent I limited the ratio variable to 1:1:1 and 2:1:1 but I did try to use ssODNs designed for both strands for each transfection. This is because I hoped that the delivery of a donor template for each strand involved in the DSB would limit the opportunity for NHEJ on one strand and thus improve the efficacy of the polyA knock-in.

An alternative delivery method that I could have considered would be the use of ribonucleoprotein (RNP) complexes for delivery of the Cas9 and sgRNA into MCF7. RNP complexes have been associated with fewer off target effects, less cytotoxicity and improved delivery when compared with plasmids. However, they are unable to introduce a selectable marker and ultimately due to the low efficiency of the gRNAs and predicted low HDR mediated knock-in, I needed to enrich for the transfected cells.

In order to overcome the difficulties encountered in identifying positive knock-in clones, I used a plasmid co-expressing Cas9 and the fluorescent protein GFP (pSpCas9(BB)-2A-GFP (Addgene-PX458)) and attempted to isolate the successfully transfected cells using fluorescence-activated cell sorting (FACS). Unfortunately, this technique proved to be too toxic for the cells which had already undergone significant shock during electroporation or transfection and the vast majority of the single cells did not survive. Instead I found success using the plasmid co-expressing Cas9 and a puromycin resistance cassette (pSpCas9(BB)-2A-Puro (Addgene - PX459)) in which the cell is afforded resistance to the antibiotic puromycin (through puromycin N-acetyl-transferase). In selecting for those cells which had been successfully transfected with Cas9, I was able to significantly increase the proportion of positive knock-in clones within the selected pool and thus reduce the lengthy process of identifying a single cell knock-in clone.

6.3.1c Choice of donor template

In CRISPR, homology directed repair relies on the presence of a donor template with sufficient homology to the genome either side of the double strand break created by Cas9. In the case of CRISPR/Cas9 mediated knock-in, the homology arms are designed to flank the intended knocked in sequence and their length can influence the efficiency of the knock-in. There are several ways in which HDR templates can be delivered to the double strand break site, each having their own advantages and disadvantages, and many having been used in the pre-CRISPR era of TALENS and zinc finger nucleases (ZFNs). Single stranded oligo DNA nucleotides (ssODNs) are effective in delivering short donor templates[181, 182], however the ssODN themselves are limited to only a few hundred bases in length and thus the insert and the homology arms are both limited in length (although efficiency probably plateaus around 80-90nt[183]). However, longer homology arms increase the molecular weight of the donor and reduce the copy number of the templates introduced to the cell for the same total molecular weight. Nevertheless, for larger insertions, a donor plasmid is considered more efficient than ssODNs because of the possibility of longer homology arms but plasmids are associated with greater cytotoxicity and more off-target effects. The polyadenylation cassette designed by Proudfoot[122] that I hoped to knock-in to the enhancer was 49bp long, and hence I chose to use a single stranded ODN with short 35nt symmetrical arms. The ssODN was designed such that the donor template would be within 3 bases of the predicted cleavage site because HDR is most efficient when the insertion site and the DSB are within 10 nucleotides of each other. I designed two ssODNs for each DSB with each being complementary to a single strand as I did not know which would be more efficiently incorporated. As the donor template was 49bp long I did not introduce a blocking mutation as I felt that following successful knock-in, the sequence would no longer be recognised by the gRNAs for re-cutting, however, I did design a reverse RT-qPCR primer complementary to a sequence within the polyadenylation sequence so that I would be able to identify its presence using a nested PCR technique.

In recent years many new techniques have been developed to improve the efficiency of CRISPR mediated genome editing, and particularly the efficiency of HDR knock-in. Some have sought to improve donor template delivery to the target site by covalently binding it to Cas9[184] whilst others have utilised an alternative Cas9 requiring a different PAM recognition site or a deactivated Cas9 incapable of causing double strand breaks. Arresting the cell in G2/M phase with microtubule polymerization inhibitors[185] or inhibition of the repair pathway enzymes and hence inhibition of NHEJ[186, 187] have also been reported and many more novel techniques besides and if I was to repeat this knock-in attempt again there are many alternatives that I would explore further.

6.3.2 Identifying the CRIPSR/Cas9 mediated knock-in

Despite all of the optimisation steps and repeated CRISPR work I undertook, identification of a single cell harbouring the polyadenylation sequence knock-in proved challenging. Using a two round "nested" PCR technique to prove the presence of the polyadenylation sequence within the selected target genome, I was confident that my optimised technique was capable of knocking in the sequence but the frequency with which it was happening made single positive cell isolation very difficult. By switching to a plasmid co-expressing Cas9 and a puromycin resistance cassette I was able to select for only those cells that had been transfected with Cas9 with the hope of increasing the proportion of those with a successful knock-in. However, even after puromycin selection and evidence of the knock-in within the selected pool, from a total of over 2000 individual CRISPR MCF7 cells plated for clonal expansion in 96 well plates, I identified only 2 clones harbouring the knock-in. Despite DNA electrophoresis identifying the presence of the knocked in sequence, I was unable to identify the polyadenylation sequence on Sanger sequencing. As discussed above, HDR often only occurs within one allele and the second may undergo repair with NHEJ. Indeed, the Sanger sequencing of these clones suggest that at least 3 alleles exists, or, alternatively, that clonal contamination may have occurred and more than 1 clone was being sequenced. MCF7 is known to be at least tetraploidy and I think that only one allele underwent successful HDR with the sequencing being corrupted by the presence of two or more other alleles, either unaffected or with NHEJ mediated indels.

In order to gain further evidence of the presence of the knock-in sequence within the enhancer, I instead eluted and sequenced the DNA band produced during gel electrophoresis following nested PCR. Sanger sequencing thus confirmed the presence of the polyadenylation sequence embedded within the enhancer genome at the expected location, although the sequencing identified only the first 36 of the 49 knocked in nucleotides because the second round PCR primer is nested within the donor sequence.

6.4 Knock-in of the polyadenylation sequence results in knockdown of the enhancer RNA but not of the divergently transcribed enhancer RNA

Having shown definitively that the CRISPR/Cas9 mediated knock-in of the polyadenylation sequence had been successful, I sought to investigate the local effects of the knock-in. The intention of such a knock-in was to bring about early transcription termination of the enhancer. In both isolated clones the polyadenylation sequence was knocked in to the first exon of *CCND1*e(antisense) using an antisense donor template. Both of the clones identified as harbouring the knock-in had undergone successful editing using an antisense ssODN targeting the antisense enhancer, although I have no reason to suspect that the direction of transcription was important in its success as I had multiple examples of successful knock-into the sense enhancer within a mixed cell pool.

RT-qPCR of the two successfully knocked in clones identified a knockdown of the antisense eRNA but no knockdown of the divergently transcribed sense transcript. As the knock-in had delivered a transcription termination sequence to the antisense enhancer, and not to the sense enhancer, the RT-qPCR results were as I might have expected. As the divergently transcribed eRNAs arising from this region share a TSS and binding sites for the recognised TFs, I propose that I have not interfered with TF binding in the CRISPR design as the sense eRNA transcript appears unaffected by the knock-in.

Although the RT-qPCR findings suggest that the CRISPR mediated knock-in has prevented full nascent *CCND1*e(antisense) transcription it would be useful to visualise the initiation of its transcription using global run on sequencing (GRO-Seq) and to show the shortened *CCND1*e(antisense) transcript using 3' RACE and sequencing. Unfortunately attempts to show a much shorter antisense eRNA transcript using 3'RACE during the course of this work were unsuccessful because the 5' primer was distal to the insertion site. However, it may be difficult to identify such a short unstable RNA with 3' RACE because of rapid degradation by the RNA exosome.

6.5 knock-in of the polyadenylation sequence into the *CCND1*e antisense enhancer results in knockdown of *CCND1* mRNA and global reduction in expression of many ER regulated genes

In the clones shown to have successfully undergone PAS knock-in I found a significant reduction in the relative expression of *CCND1* mRNA following induction with E2. This is in contrast to my earlier siRNA experiments in which the antisense *CCND1*e(antisense) eRNA transcript was knocked down with siRNA and I saw no effect on *CCND1* gene regulation.

Further still, when one of these PAS knock-in clones underwent RNA-sequencing following 6hours E2 induction, I discovered that insertion of the polyadenylation cassette and subsequent early termination of transcription of the antisense *CCND1*e eRNA resulted in global reduction in expression of *many* ER regulated genes. Indeed, many genes usually up regulated by ER were found to be down regulated. However, RT-qPCR has not validated these findings as I cannot confirm statistically significant knockdown of recognised ER driven genes.

Nevertheless, a dual luciferase reporter assay confirms that there is considerably less luciferase activity in the knock-in clone compared to wild type, suggesting less ER activity in the clone harbouring the polyadenylation sequence in the antisense enhancer. The possibility that early termination of transcription of an enhancer of *CCND1* could have such a global effect is obviously interesting and the host lab will continue to investigate these findings and aim to validate, or not, the RNA-Seq findings. If such global knock down is not found on further investigation, it is possible that the RNA-seq findings are a consequence of clonal selection in which this expanded clone has become less responsive to E2 stimulation when compared to wild type cells. This could be as a consequence of CRISPR genome editing, either through off target effects or secondary to repeated passage and encountered toxicities. Of course, within an individual tumour there exists a heterogenous population of cells, some of which will exhibit different phenotypes and it is possible that the clone sequenced following CRISPR already exhibited an insensitivity to E2 induction. Single cell sequencing is likely to reveal more about such heterogeneity in the future. Indeed, single cell CAGE sequencing has recently identified that while *en masse* enhancers are bi-directionally transcribed, on a single cell level enhancers are almost exclusively unidirectionally transcribed from either strand [158], explaining the earlier findings that one of the divergent enhancer transcripts provides the majority of biological function [64].

In my MCF7 model I have identified a region upstream of *CCND1* which bears the hallmarks of an enhancer (which has since been confirmed in the literature [118]). The enhancer region is actively transcribed in a divergent bi-directional manner following induction by E2 producing two separate transcripts. These transcripts are present in the cell up to 24 hours following E2 induction and the multi-exonic antisense transcript is found in the cytoplasm at 24 hours too. siRNA knockdown of either transcript does not affect relative expression of its neighbouring gene *CCND1* but insertion of a transcription termination signal into the antisense enhancer not only knocks down the antisense eRNA but also significantly reduces expression of *CCND1* and *may* result in a global reduction in expression of many ER regulated genes.

The mechanism by which the *CCND1*e(antisense) eRNA may exert this influence over the cell has yet to be determined but my findings would suggest that in my model the eRNA transcript itself may not be required for enhancing *CCND1* transcription. Indeed the presence and abundance of this spliced eRNA transcript in the cytoplasm so many hours after its induction would suggest a functional role outside of the nucleus which as yet remains unknown. I suggest that a possible mechanism by which this *CCND1* enhancer exerts some control over *CCND1* gene expression is through the act of transcription of the enhancer. Having sought to disrupt enhancer transcription only after binding of all predicted TFs I hope to have differentiated my findings from others [118] who showed that CRISPR/Cas9 mediated editing of the ER binding site of this same enhancer resulted in downregulation of the *CCND1* gene.

Whilst I used spCas9 with an active endonuclease with the intention of causing DSB at the target site, it is possible that the large Cas9 protein has in fact resulted in steric hindrance similar to that used by dead dCas9 which has lost its endonuclease activity. If the sgRNA-Cas9 complex is held at the target site for long, it may be capable of blocking ER binding and further activation of the enhancer and thus the knockdown of *CCND1* may be as a result of an inactive enhancer rather than a prematurely terminated transcription. Chip-Seq of ER and other important proteins such as P300 and Mediator would clarify if this is the case, and Global RunOn sequencing would show that the nascent transcript is being initiated. I would like to see evidence not only of initiation of transcription but perhaps also of the truncated transcript through 3' RACE, although its short length would likely make it a target for rapid degradation.

However, one aspect that I would be keen to investigate is the impact of the CRISPR editing and eRNA knockdown on chromosomal looping between the enhancer and *CCND1* gene promoter. eRNAs are prominent at looped enhancers [69, 70] which occur prior to gene transcription[71], and whilst they have been shown to interact with and recruit many proteins involved in the formation and stabilization of such loops [72] [151] [168] their role in DNA looping remains unclear. Whilst some studies have shown decreases in chromosomal looping with knockdown of eRNA [60, 166, 168, 169] others have found no change despite eRNA knockdown or prevention of RNA Pol II transcription with flavopiridol and thus I would be keen to compare the impact of terminating transcription elongation of the enhancer (PAS knock-in) with prevention of TF binding (CRISPR mediated ER binding disruption clone) and knockdown of the transcript itself (siRNA knockdown). As others have previously stated, a potential explanation for the conflicting studies of eRNA involvement in enhancer:promoter looping is the basal level of eRNA transcription that continues prior to their disruption or knockdown which enables the establishment of looping [188].

At the outset of the CRISPR/cas9 work I had also intended to compare the outcome of PAS knock-in with CRISPR/Cas9 mediated disruption of ER binding at the same enhancer, in the same way as previously published[118].

Unfortunately, during the course of this work, I was unable to achieve disruption of the ER binding site at the *CCND1* enhancer through my own CRISPR but are grateful to the Agami group (Division of Oncogenomics at the Netherlands Cancer Institute) for their clones and their collaboration.

The technique and optimised protocol used in this work was subsequently used by others in the host laboratory with success and the many stages of optimisation resulted in a relatively speedy output in other work. Although there were other optimisation steps that I could have explored, and have since been reported in the literature, this work is one of very few reports of successful polyadenylation knock-into an enhancer region and I hope that my CRISPR work and findings will add to the body of literature investigating the potential mechanism of action of the ubiquitous enhancer RNAs.

6.6 Conclusion

From an original intention of identifying the lncRNA transcriptome regulated by ERα in hormone responsive breast cancer cell lines, I have identified a large number of previously unannotated lncRNAs some of which arise from *cis* regulatory regions bearing the hallmarks of active enhancers. I have shown that these transcripts can be spliced or unspliced and can be both up and down regulated in response to estrogen treatment. Interestingly, I find many transcripts are spliced and located in the cytoplasm, suggesting a functional relevance outside of *cis* regulation.

I have further identified the bidirectionally transcribed enhancer RNAs arising from an enhancer of *CCND1*, an important gene in cell cycle regulation and the DNA repair pathway that is overexpressed in over 50% of breast cancers. I have found that the divergently transcribed eRNAs are estrogen responsive, not rapidly degraded and that the antisense eRNA transcript is spliced and located in the cytoplasm at 6 hours after E2 induction. Although I have been unable to identify any protein binding partners or mechanism of action, I suggest that its location in the cytoplasm is likely to be functional.

I have successfully knocked down each individual eRNA transcript with siRNA technology and found no effect on expression of the neighbouring gene *CCND1*. However, using CRISPR/Cas9 I have successfully knocked in a polyadenylation signal to terminate transcription of the antisense enhancer and found that this not only downregulates the eRNA but also *CCND1* mRNA and may also have a more global effect of downregulating ER responsiveness. I also discuss the difficulties encountered in harnessing homology directed repair in the CRISPR system and ways in which I have optimised the technique.

Although there remains much more to investigate, my work adds to the growing body of literature around enhancer regulation and the functional relevance of the transcripts arising from them. My work suggests that, in the case of this enhancer, the eRNA transcripts themselves are not required for neighbouring gene regulation but at least one of them may have a functional role elsewhere in the cell. Instead, I propose that the act of transcription by RNA Pol II is responsible for the regulatory effects of this enhancer and I plan to explore this further in future work.

References

1. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
2. Blagden, S.P. and A.E. Willis, *The biological and therapeutic relevance of mRNA translation in cancer.* Nat Rev Clin Oncol, 2011. **8**(5): p. 280-91.
3. Castellano, L., et al., *The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15732-7.
4. Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nat Genet, 2015. **47**(3): p. 199-208.
5. Consortium, T.E.P., *6 Non coding RNA characterization.* Nature, 2019.
6. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
7. Guttman, M., et al., *Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.* Cell, 2013. **154**(1): p. 240-51.
8. Pauli, A., et al., *Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis.* Genome Res, 2012. **22**(3): p. 577-91.
9. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
10. Richard, J.L.C. and P.J.A. Eichhorn, *Deciphering the roles of lncRNAs in breast development and disease.* Oncotarget, 2018. **9**(28): p. 20179-20212.
11. Zhao, Z., et al., *lncRNA-Induced Nucleosome Repositioning Reinforces Transcriptional Repression of rRNA Genes upon Hypotonic Stress.* Cell Rep, 2016. **14**(8): p. 1876-82.
12. Wang, X.Q. and J. Dostie, *Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation.* Nucleic Acids Res, 2017. **45**(3): p. 1091-1104.
13. Bierhoff, H., et al., *Quiescence-induced LncRNAs trigger H4K20 trimethylation and transcriptional silencing.* Mol Cell, 2014. **54**(4): p. 675-82.
14. Han, P., et al., *A long noncoding RNA protects the heart from pathological hypertrophy.* Nature, 2014. **514**(7520): p. 102-106.
15. Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.
16. Wang, J., et al., *Imprinted X inactivation maintained by a mouse Polycomb group gene.* Nat Genet, 2001. **28**(4): p. 371-5.
17. Plath, K., et al., *Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome.* J Cell Biol, 2004. **167**(6): p. 1025-35.
18. Hacisuleyman, E., et al., *Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.* Nat Struct Mol Biol, 2014. **21**(2): p. 198-206.
19. Gong, C. and L.E. Maquat, *lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.* Nature, 2011. **470**(7333): p. 284-8.

20. Faghihi, M.A., et al., *Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase.* Nat Med, 2008. **14**(7): p. 723-30.
21. Yoon, J.H., et al., *LincRNA-p21 suppresses target mRNA translation.* Mol Cell, 2012. **47**(4): p. 648-55.
22. Wang, J., et al., *CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer.* Nucleic Acids Res, 2010. **38**(16): p. 5366-83.
23. Dey, B.K., K. Pfeifer, and A. Dutta, *The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration.* Genes Dev, 2014. **28**(5): p. 491-501.
24. Yin, X., Y. Jing, and H. Xu, *Mining for missed sORF-encoded peptides.* Expert Rev Proteomics, 2019. **16**(3): p. 257-266.
25. Bhan, A. and S.S. Mandal, *LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer.* Biochim Biophys Acta, 2015. **1856**(1): p. 151-64.
26. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.
27. Majello, B., et al., *Expanding the Role of the Histone Lysine-Specific Demethylase LSD1 in Cancer.* Cancers (Basel), 2019. **11**(3).
28. Dai, W., et al., *Long Noncoding RNA HOTAIR Functions as a Competitive Endogenous RNA to Regulate Connexin43 Remodeling in Atrial Fibrillation by Sponging MicroRNA-613.* Cardiovasc Ther, 2020. **2020**: p. 5925342.
29. Liu, X.H., et al., *Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer.* Mol Cancer, 2014. **13**: p. 92.
30. Zhang, H., et al., *MiR-7, inhibited indirectly by lincRNA HOTAIR, directly inhibits SETDB1 and reverses the EMT of breast cancer stem cells by downregulating the STAT3 pathway.* Stem Cells, 2014. **32**(11): p. 2858-68.
31. Cipriano, A. and M. Ballarino, *The Ever-Evolving Concept of the Gene: The Use of RNA/Protein Experimental Techniques to Understand Genome Functions.* Front Mol Biosci, 2018. **5**: p. 20.
32. Ghisletti, S., et al., *Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages.* Immunity, 2010. **32**(3): p. 317-28.
33. Ong, C.T. and V.G. Corces, *Enhancer function: new insights into the regulation of tissue-specific gene expression.* Nat Rev Genet, 2011. **12**(4): p. 283-93.
34. Woolfe, A., et al., *Highly conserved non-coding sequences are associated with vertebrate development.* PLoS Biol, 2005. **3**(1): p. e7.
35. Kvon, E.Z., et al., *Enhancer redundancy in development and disease.* Nat Rev Genet, 2021. **22**(5): p. 324-336.
36. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2 Pt 1): p. 299-308.
37. Geyer, P.K., M.M. Green, and V.G. Corces, *Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous*

chromosome: the molecular basis of transvection in Drosophila. EMBO J, 1990. **9**(7): p. 2247-56.

38. Lomvardas, S., et al., *Interchromosomal interactions and olfactory receptor choice.* Cell, 2006. **126**(2): p. 403-13.
39. Kaikkonen, M.U., et al., *Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription.* Mol Cell, 2013. **51**(3): p. 310-25.
40. Arner, E., et al., *Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells.* Science, 2015. **347**(6225): p. 1010-4.
41. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-12.
42. Hoffman, M.M., et al., *Integrative annotation of chromatin elements from ENCODE data.* Nucleic Acids Res, 2013. **41**(2): p. 827-41.
43. Zentner, G.E., P.J. Tesar, and P.C. Scacheri, *Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions.* Genome Res, 2011. **21**(8): p. 1273-83.
44. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.
45. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
46. Kanno, T., et al., *BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones.* Nat Struct Mol Biol, 2014. **21**(12): p. 1047-57.
47. Jiao, W., et al., *HPSE enhancer RNA promotes cancer progression through driving chromatin looping and regulating hnRNPU/p300/EGR1/HPSE axis.* Oncogene, 2018. **37**(20): p. 2728-2745.
48. Ordonez, R., et al., *DNA Methylation of Enhancer Elements in Myeloid Neoplasms: Think Outside the Promoters?* Cancers (Basel), 2019. **11**(10).
49. De Santa, F., et al., *A large fraction of extragenic RNA pol II transcription sites overlap enhancers.* PLoS Biol, 2010. **8**(5): p. e1000384.
50. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-7.
51. Noguchi, S., et al., *FANTOM5 CAGE profiles of human and mouse samples.* Sci Data, 2017. **4**: p. 170112.
52. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.
53. Descostes, N., et al., *Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells.* Elife, 2014. **3**: p. e02105.
54. Andersson, R., et al., *Nuclear stability and transcriptional directionality separate functionally distinct RNA species.* Nat Commun, 2014. **5**: p. 5336.
55. Henriques, T., et al., *Widespread transcriptional pausing and elongation control at enhancers.* Genes Dev, 2018. **32**(1): p. 26-41.
56. Dorighi, K.M., et al., *Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation.* Mol Cell, 2017. **66**(4): p. 568-576 e4.

57. Koch, F., et al., *Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters.* Nat Struct Mol Biol, 2011. **18**(8): p. 956-63.
58. Zhu, Y., et al., *Predicting enhancer transcription and activity from chromatin modifications.* Nucleic Acids Res, 2013. **41**(22): p. 10032-43.
59. Cinghu, S., et al., *Intragenic Enhancers Attenuate Host Gene Expression.* Mol Cell, 2017. **68**(1): p. 104-117 e6.
60. Li, W., et al., *Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation.* Nature, 2013. **498**(7455): p. 516-20.
61. Schaukowitch, K., et al., *Enhancer RNA facilitates NELF release from immediate early genes.* Mol Cell, 2014. **56**(1): p. 29-42.
62. Shii, L., et al., *SERPINB2 is regulated by dynamic interactions with pause-release proteins and enhancer RNAs.* Mol Immunol, 2017. **88**: p. 20-31.
63. Lam, M.T., et al., *Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription.* Nature, 2013. **498**(7455): p. 511-5.
64. Hsieh, C.L., et al., *Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation.* Proc Natl Acad Sci U S A, 2014. **111**(20): p. 7319-24.
65. Melo, C.A., et al., *eRNAs are required for p53-dependent enhancer activity and gene transcription.* Mol Cell, 2013. **49**(3): p. 524-35.
66. Mousavi, K., et al., *eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci.* Mol Cell, 2013. **51**(5): p. 606-17.
67. Maruyama, A., J. Mimura, and K. Itoh, *Non-coding RNA derived from the region adjacent to the human HO-1 E2 enhancer selectively regulates HO-1 gene induction by modulating Pol II binding.* Nucleic Acids Res, 2014. **42**(22): p. 13599-614.
68. Zhao, Y., et al., *Activation of P-TEFb by Androgen Receptor-Regulated Enhancer RNAs in Castration-Resistant Prostate Cancer.* Cell Rep, 2016. **15**(3): p. 599-610.
69. Sanyal, A., et al., *The long-range interaction landscape of gene promoters.* Nature, 2012. **489**(7414): p. 109-13.
70. Wang, D., et al., *Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA.* Nature, 2011. **474**(7351): p. 390-4.
71. Kim, Y.W., et al., *Chromatin looping and eRNA transcription precede the transcriptional activation of gene in the beta-globin locus.* Biosci Rep, 2015. **35**(2).
72. Pezone, A., et al., *RNA Stabilizes Transcription-Dependent Chromatin Loops Induced By Nuclear Hormones.* Sci Rep, 2019. **9**(1): p. 3925.
73. Lai, F., et al., *Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.* Nature, 2013. **494**(7438): p. 497-501.
74. Lai, F., et al., *Integrator mediates the biogenesis of enhancer RNAs.* Nature, 2015. **525**(7569): p. 399-403.
75. Williamson, I., et al., *Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization.* Genes Dev, 2014. **28**(24): p. 2778-91.

76. Hah, N., et al., *Enhancer transcripts mark active estrogen receptor binding sites.* Genome Res, 2013. **23**(8): p. 1210-23.

77. Deroo, B.J. and K.S. Korach, *Estrogen receptors and human disease.* J Clin Invest, 2006. **116**(3): p. 561-70.

78. Cadenas, C. and H.M. Bolt, *Estrogen receptors in human disease.* Arch Toxicol, 2012. **86**(10): p. 1489-90.

79. Hou, Y.F., et al., *ERbeta exerts multiple stimulative effects on human breast carcinoma cells.* Oncogene, 2004. **23**(34): p. 5799-806.

80. Shang, Y., et al., *Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription.* Cell, 2000. **103**(6): p. 843-52.

81. Metivier, R., et al., *Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter.* Cell, 2003. **115**(6): p. 751-63.

82. Holmes, K.A., et al., *Transducin-like enhancer protein 1 mediates estrogen receptor binding and transcriptional activity in breast cancer cells.* Proc Natl Acad Sci U S A, 2012. **109**(8): p. 2748-53.

83. Tan, Y., et al., *Dismissal of RNA Polymerase II Underlies a Large Ligand-Induced Enhancer Decommissioning Program.* Mol Cell, 2018. **71**(4): p. 526-539 e8.

84. Fu, M., et al., *Minireview: Cyclin D1: normal and abnormal functions.* Endocrinology, 2004. **145**(12): p. 5439-47.

85. Sherr, C.J. and J.M. Roberts, *CDK inhibitors: positive and negative regulators of G1-phase progression.* Genes Dev, 1999. **13**(12): p. 1501-12.

86. Barnes, D.M. and C.E. Gillett, *Cyclin D1 in breast cancer.* Breast Cancer Res Treat, 1998. **52**(1-3): p. 1-15.

87. Balcerczak, E., et al., *Cyclin D1 protein and CCND1 gene expression in colorectal cancer.* Eur J Surg Oncol, 2005. **31**(7): p. 721-6.

88. Ikeguchi, M., et al., *Cyclin D1 expression and retinoblastoma gene protein (pRB) expression in esophageal squamous cell carcinoma.* J Cancer Res Clin Oncol, 2001. **127**(9): p. 531-6.

89. Arnold, A. and A. Papanikolaou, *Cyclin D1 in breast cancer pathogenesis.* J Clin Oncol, 2005. **23**(18): p. 4215-24.

90. Ortiz, A.B., et al., *Prognostic significance of cyclin D1 protein expression and gene amplification in invasive breast carcinoma.* PLoS One, 2017. **12**(11): p. e0188068.

91. Hodges, L.C., et al., *Tamoxifen functions as a molecular agonist inducing cell cycle-associated genes in breast cancer cells.* Mol Cancer Res, 2003. **1**(4): p. 300-11.

92. Bartek, J. and J. Lukas, *DNA repair: Cyclin D1 multitasks.* Nature, 2011. **474**(7350): p. 171-2.

93. Hortobagyi, G.N., *Ribociclib for the first-line treatment of advanced hormone receptor-positive breast cancer: a review of subgroup analyses from the MONALEESA-2 trial.* Breast Cancer Res, 2018. **20**(1): p. 123.

94. Turner, N.C., et al., *Overall Survival with Palbociclib and Fulvestrant in Advanced Breast Cancer.* N Engl J Med, 2018. **379**(20): p. 1926-1936.

95. Sledge, G.W., Jr., et al., *MONARCH 2: Abemaciclib in Combination With Fulvestrant in Women With HR+/HER2- Advanced Breast Cancer Who Had Progressed While Receiving Endocrine Therapy.* J Clin Oncol, 2017. **35**(25): p. 2875-2884.

96. Majumdar, R., K. Rajasekaran, and J.W. Cary, *RNA Interference (RNAi) as a Potential Tool for Control of Mycotoxin Contamination in Crop Plants: Concepts and Considerations.* Front Plant Sci, 2017. **8**: p. 200.

97. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-2308.

98. Ran, F.A., et al., *Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity.* Cell, 2013. **154**(6): p. 1380-9.

99. Cong, L., et al., *Multiplex genome engineering using CRISPR/Cas systems.* Science, 2013. **339**(6121): p. 819-23.

100. Hsu, P.D., et al., *DNA targeting specificity of RNA-guided Cas9 nucleases.* Nat Biotechnol, 2013. **31**(9): p. 827-32.

101. Mojica, F.J.M., et al., *Short motif sequences determine the targets of the prokaryotic CRISPR defence system.* Microbiology (Reading), 2009. **155**(Pt 3): p. 733-740.

102. Liu, X., et al., *Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system.* Sci Rep, 2016. **6**: p. 19675.

103. Maquat, L.E., *When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells.* RNA, 1995. **1**(5): p. 453-65.

104. Groschel, S., et al., *A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia.* Cell, 2014. **157**(2): p. 369-381.

105. Aparicio-Prat, E., et al., *DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs.* BMC Genomics, 2015. **16**: p. 846.

106. Chu, V.T., et al., *Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells.* Nat Biotechnol, 2015. **33**(5): p. 543-8.

107. Salsman, J. and G. Dellaire, *Precision genome editing in the CRISPR era.* Biochem Cell Biol, 2017. **95**(2): p. 187-201.

108. Addgene. *CRISPR/Cas9 Guide*. 2022.

109. Jensen, T.I., et al., *Targeted regulation of transcription in primary cells using CRISPRa and CRISPRi.* Genome Res, 2021. **31**(11): p. 2120-2130.

110. Perez-Pinera, P., et al., *RNA-guided gene activation by CRISPR-Cas9-based transcription factors.* Nat Methods, 2013. **10**(10): p. 973-6.

111. Gilbert, L.A., et al., *CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes.* Cell, 2013. **154**(2): p. 442-51.

112. Dominguez, A.A., W.A. Lim, and L.S. Qi, *Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation.* Nat Rev Mol Cell Biol, 2016. **17**(1): p. 5-15.

113. Vojta, A., et al., *Repurposing the CRISPR-Cas9 system for targeted DNA methylation.* Nucleic Acids Res, 2016. **44**(12): p. 5615-28.

114. Richardson, C.D., et al., *Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA.* Nat Biotechnol, 2016. **34**(3): p. 339-44.

115. Canver, M.C., et al., *BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis.* Nature, 2015. **527**(7577): p. 192-7.

116. Kearns, N.A., et al., *Functional annotation of native enhancers with a Cas9-histone demethylase fusion.* Nat Methods, 2015. **12**(5): p. 401-403.

117. Hilton, I.B., et al., *Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers.* Nat Biotechnol, 2015. **33**(5): p. 510-7.
118. Korkmaz, G., et al., *Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9.* Nat Biotechnol, 2016. **34**(2): p. 192-8.
119. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
120. Mendez, J. and B. Stillman, *Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis.* Mol Cell Biol, 2000. **20**(22): p. 8602-12.
121. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.* Nat Methods, 2012. **9**(7): p. 676-82.
122. Proudfoot, N.J., *Ending the message: poly(A) signals then and now.* Genes Dev, 2011. **25**(17): p. 1770-82.
123. Zhang, F. *CISR.mit.edu* 2015 [cited 2016 26 August]; Available from: <https://cisr.mit.edu/>.
124. NCBI. [cited 2017 May]; Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.
125. Mali, P., et al., *RNA-guided human genome engineering via Cas9.* Science, 2013. **339**(6121): p. 823-6.
126. Brinkman, E.K., et al., *Easy quantitative assessment of genome editing by sequence trace decomposition.* Nucleic Acids Res, 2014. **42**(22): p. e168.
127. Siersbaek, R., S. Kumar, and J.S. Carroll, *Signaling pathways and steroid receptors modulating estrogen receptor alpha function in breast cancer.* Genes Dev, 2018. **32**(17-18): p. 1141-1154.
128. Rae, J.M., et al., *GREB 1 is a critical regulator of hormone dependent breast cancer growth.* Breast Cancer Res Treat, 2005. **92**(2): p. 141-9.
129. Li, Y., et al., *The histone modifications governing TFF1 transcription mediated by estrogen receptor.* J Biol Chem, 2011. **286**(16): p. 13925-36.
130. Kilic, Y., A.C. Celebiler, and M. Sakizli, *Selecting housekeeping genes as references for the normalization of quantitative PCR data in breast cancer.* Clin Transl Oncol, 2014. **16**(2): p. 184-90.
131. Ruiz-Orera, J., et al., *Long non-coding RNAs as a source of new peptides.* Elife, 2014. **3**: p. e03523.
132. Wang, L., et al., *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.* Nucleic Acids Res, 2013. **41**(6): p. e74.
133. Lin, M.F., I. Jungreis, and M. Kellis, *PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.* Bioinformatics, 2011. **27**(13): p. i275-82.
134. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
135. Turnbull, C., et al., *Genome-wide association study identifies five new breast cancer susceptibility loci.* Nat Genet, 2010. **42**(6): p. 504-7.
136. Mudge, J.M., et al., *Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci.* Genome Res, 2019. **29**(12): p. 2073-2087.

137. Sun, M., et al., *Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells.* Mol Cell, 2015. **59**(4): p. 698-711.
138. Cathcart, P., *Oestrogenic Regulation of lncRNA at Enhancer Regions*. 2019, Imperial College, London.
139. Amin, M.B.e.a., *AJCC Cancer Staging Manuel*. 8th ed. 2017, New York: Springer. 1032.
140. Arber, N., et al., *Antisense to cyclin D1 inhibits the growth and tumorigenicity of human colon cancer cells.* Cancer Res, 1997. **57**(8): p. 1569-74.
141. Wang, J., et al., *Knockdown of cyclin D1 inhibits proliferation, induces apoptosis, and attenuates the invasive capacity of human glioblastoma cells.* J Neurooncol, 2012. **106**(3): p. 473-84.
142. Lizio, M., et al., *Gateways to the FANTOM5 promoter level mammalian expression atlas.* Genome Biol, 2015. **16**: p. 22.
143. Sloan, C.A., et al., *ENCODE data at the ENCODE portal.* Nucleic Acids Res, 2016. **44**(D1): p. D726-32.
144. Paralkar, V.R., et al., *Unlinking an lncRNA from Its Associated cis Element.* Mol Cell, 2016. **62**(1): p. 104-10.
145. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
146. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Res, 2012. **22**(9): p. 1798-812.
147. Wang, J., et al., *Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium.* Nucleic Acids Res, 2013. **41**(Database issue): p. D171-6.
148. Davidson, J.M., et al., *Molecular cytogenetic analysis of breast cancer cell lines.* Br J Cancer, 2000. **83**(10): p. 1309-17.
149. Vasquez, Y.M., et al., *Genome-wide analysis and functional prediction of the estrogen-regulated transcriptional response in the mouse uterusdagger.* Biol Reprod, 2020. **102**(2): p. 327-338.
150. Sigova, A.A., et al., *Transcription factor trapping by RNA in gene regulatory elements.* Science, 2015. **350**(6263): p. 978-81.
151. Tsai, P.F., et al., *A Muscle-Specific Enhancer RNA Mediates Cohesin Recruitment and Regulates Transcription In trans.* Mol Cell, 2018. **71**(1): p. 129-141 e8.
152. Tan, J.Y., et al., *Splicing of enhancer-associated lincRNAs contributes to enhancer activity.* Life Sci Alliance, 2020. **3**(4).
153. Meininger, I., et al., *Alternative splicing of MALT1 controls signalling and activation of CD4(+) T cells.* Nat Commun, 2016. **7**: p. 11292.
154. Falanga, A., et al., *Exonic splicing signals impose constraints upon the evolution of enzymatic activity.* Nucleic Acids Res, 2014. **42**(9): p. 5790-8.
155. Plass, M. and E. Eyras, *Differentiated evolutionary rates in alternative exons and the implications for splicing regulation.* BMC Evol Biol, 2006. **6**: p. 50.
156. Lambrechts, D., et al., *11q13 is a susceptibility locus for hormone receptor positive breast cancer.* Hum Mutat, 2012. **33**(7): p. 1123-32.

157. Alvarez-Dominguez, J.R., et al., *The Super-Enhancer-Derived alncRNA-EC7/Bloodlinc Potentiates Red Blood Cell Development in trans.* Cell Rep, 2017. **19**(12): p. 2503-2514.
158. Kouno, T., et al., *C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution.* Nat Commun, 2019. **10**(1): p. 360.
159. Hah, N., et al., *A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells.* Cell, 2011. **145**(4): p. 622-34.
160. Werner, M.S. and A.J. Ruthenburg, *Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes.* Cell Rep, 2015. **12**(7): p. 1089-98.
161. Wilusz, J.E., S.M. Freier, and D.L. Spector, *3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA.* Cell, 2008. **135**(5): p. 919-32.
162. Tristan, C., et al., *The diverse functions of GAPDH: views from different subcellular compartments.* Cell Signal, 2011. **23**(2): p. 317-23.
163. Kretz, M., *TINCR, staufen1, and cellular differentiation.* RNA Biol, 2013. **10**(10): p. 1597-601.
164. Cesana, M., et al., *A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA.* Cell, 2011. **147**(2): p. 358-69.
165. Carrieri, C., et al., *Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.* Nature, 2012. **491**(7424): p. 454-7.
166. Yang, Y., et al., *Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster.* Sci Rep, 2016. **6**: p. 20961.
167. Tan, S.H., et al., *The enhancer RNA ARIEL activates the oncogenic transcriptional program in T-cell acute lymphoblastic leukemia.* Blood, 2019. **134**(3): p. 239-251.
168. Pnueli, L., et al., *RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin alpha-subunit gene.* Proc Natl Acad Sci U S A, 2015. **112**(14): p. 4369-74.
169. Liang, J., et al., *Epstein-Barr virus super-enhancer eRNAs are essential for MYC oncogene expression and lymphoblast proliferation.* Proc Natl Acad Sci U S A, 2016. **113**(49): p. 14121-14126.
170. Brown, J.C., *Involvement of promoter/enhancers in a feedback loop to regulate human gene expression.* Heliyon, 2020. **6**(9): p. e04934.
171. Tyssowski, K.M., et al., *Different Neuronal Activity Patterns Induce Different Gene Expression Programs.* Neuron, 2018. **98**(3): p. 530-546 e11.
172. Baillie, J.K., et al., *Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease.* PLoS Genet, 2017. **13**(3): p. e1006641.
173. Gressel, S., B. Schwalb, and P. Cramer, *The pause-initiation limit restricts transcription activation in human cells.* Nat Commun, 2019. **10**(1): p. 3603.
174. Lin, S., et al., *Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery.* Elife, 2014. **3**: p. e04766.

175. Jinek, M., et al., *RNA-programmed genome editing in human cells.* Elife, 2013. **2**: p. e00471.

176. Hisano, Y., et al., *Precise in-frame integration of exogenous DNA mediated by CRISPR/Cas9 system in zebrafish.* Sci Rep, 2015. **5**: p. 8841.

177. Nakade, S., et al., *Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9.* Nat Commun, 2014. **5**: p. 5560.

178. Sakuma, T., et al., *MMEJ-assisted gene knock-in using TALENs and CRISPR-Cas9 with the PITCh systems.* Nat Protoc, 2016. **11**(1): p. 118-33.

179. Jang, D.E., et al., *Multiple sgRNAs with overlapping sequences enhance CRISPR/Cas9-mediated knock-in efficiency.* Exp Mol Med, 2018. **50**(4): p. 1-9.

180. Graf, R., et al., *sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing.* Cell Rep, 2019. **26**(5): p. 1098-1103 e3.

181. Elliott, B., et al., *Gene conversion tracts from double-strand break repair in mammalian cells.* Mol Cell Biol, 1998. **18**(1): p. 93-101.

182. Yang, L., et al., *Optimization of scarless human stem cell genome editing.* Nucleic Acids Res, 2013. **41**(19): p. 9049-61.

183. Okamoto, S., et al., *Highly efficient genome editing for single-base substitutions using optimized ssODNs with Cas9-RNPs.* Sci Rep, 2019. **9**(1): p. 4811.

184. Aird, E.J., et al., *Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template.* Commun Biol, 2018. **1**: p. 54.

185. Yang, D., et al., *Enrichment of G2/M cell cycle phase in human pluripotent stem cells enhances HDR-mediated gene repair with customizable endonucleases.* Sci Rep, 2016. **6**: p. 21264.

186. Maruyama, T., et al., *Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining.* Nat Biotechnol, 2015. **33**(5): p. 538-42.

187. Wienert, B., et al., *Timed inhibition of CDC7 increases CRISPR-Cas9 mediated templated repair.* Nat Commun, 2020. **11**(1): p. 2109.

188. Arnold, P.R., A.D. Wells, and X.C. Li, *Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate.* Front Cell Dev Biol, 2019. **7**: p. 377.