Open access • Dissertation • DOI:10.14264/UQL.2016.158

The long range regulation of breast cancer associated genes — Source link

Joshua A. Betts

Published on: 13 Mar 2016

Topics: Cancer, Breast cancer, Locus (genetics), Gene and Candidate gene

THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

The Long Range Regulation of Breast Cancer Associated Genes

Joshua Betts

MBBS, MMolBiol, FRACGP

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

School of Chemistry and Molecular Biology

Breast cancer is the most common non-cutaneous cancer of women and a major cause of mortality [1]. Genetic factors are the single biggest risk factor and 75% of the risk derives from common but low penetrance single nucleotide polymorphisms (SNPs) [2]. These are found using genome-wide association studies (GWAS) and the majority localize to non-coding regions of the genome [3]. A 2009 GWAS identified an estrogen receptor (ERα) positive breast cancer susceptibility locus in the 11q13 gene desert [4]. Previous fine mapping of this locus by our group revealed three independent risk signals containing five candidate causal single nucleotide polymorphisms (SNPs). Four of these clustered in an enhancer element (called PRE1) and one in a silencer element (PRE2). Interactions were demonstrated between these elements and the nearby cyclin D1 gene (*CCND1*), one of the most commonly amplified genes in breast tumours and a known oncogene [5]. Enhancers commonly regulate multiple genes however, suggesting that PRE1 or PRE2 may have other targets than *CCND1* [6]. Such targets may be nearby protein coding genes or previously unrecognised non-coding transcripts within the gene desert.

The identification of *CCND1* as an interacting partner of PRE1 and PRE2 was done using a candidate gene approach with 3C (chromosome conformation capture) targeted confirmation of interaction. To detect other potential targets of PRE1 and PRE2 at the 11q13 locus using an agnostic approach, the 4Cseq and 5C techniques were used. These revealed a number of local interaction regions covering six gene promoters. A novel strategy using knockdown of enhancer RNA transcribed from PRE1 then identified the *IGHMBP2* and *CPT1A* genes as likely additional targets of PRE1. Genome editing with transcription activator-like effector nucleases (TALENS) was also used, creating isogenic cell lines to clarify the effects of the SNPs in their native genomic context. Further functional work is required, however the range of techniques employed has substantially expanded our understanding of the 11q13 breast cancer risk locus and forms a template for future investigations of GWAS risk loci.

To identify noncoding RNAs expressed at the 11q13 locus that may be interacting with PRE1 or PRE2 required the use of RNA Capture-seq. RNA Capture-seq involves a targeted enrichment step that greatly increases the sequencing depth at regions of interest and can reveal lowly expressed transcripts that may have not been found by traditional RNA-seq [7]. This identified one novel long non-coding RNA expressed from the (+) strand that was named *CUPID1* and a second arising from the same locus on the (–) strand named *CUPID2*. They were located in the nucleus, oestrogen induced and both expressed relatively highly in ERα positive breast cancer cell lines but not in ERα

negative cell lines or normal breast tissue. 3C revealed that PRE1 exhibits allele specific interactions with the lncRNA promoter and functional assays demonstrated that PRE1 markedly increased promoter activity. The increased activity was partially abrogated following incorporation of the risk SNPs. SiRNA mediated knockdown of both lncRNAs impaired the normal DNA damage response of breast cancer cells and the ChIRPseq technique confirmed that *CUPID1* was preferentially bound to promoters of genes involved in DNA repair. Two oestrogen responsive lncRNAs that interact with the 11q13 breast cancer susceptibility locus have thus been identified that may mediate the associated risk of oestrogen responsive breast cancer.

The 11q13 locus is amplified in around 20% of breast cancers and is thought to contain a number of driver genes including *CCND1* [8]. Given that *CUPID1* and *CUPID2* are widely over-expressed in ERα positive breast cancer cell lines, it was hypothesized that they may also have a role in driving tumour growth. Publically available RNAseq data showed *CUPID2* to be highly expressed in breast and renal cancer but not in normal tissue. Stable cell lines over-expressing the lncRNAs were then generated and subsequent oncogenic assays revealed that *CUPID2* increased cellular proliferation and the efficiency of colony formation in breast cancer cells. A murine xenograft model confirmed these findings *in vivo* by demonstrating a marked increase in tumour size for the mice injected with cells over-expressing *CUPID2*. There is thus evidence that *CUPID2* behaves as an oncogene and may be an additional driver of the 11q13 amplicon in ERα positive breast cancer.

In summary, *CPT1A* and *IGHMBP2* were identified as potential mediators of risk at the 11q13 breast cancer susceptibility locus. Further investigations of the locus then identified two novel non-coding transcripts called *CUPID1* and *CUPID2* that may also mediate the effects of the causal SNPs through downregulation of cellular DNA damage repair pathways. Finally, *CUPID2* was shown to function as a putative oncogene in ERα positive breast cancer. These findings add to our knowledge of breast cancer risk and progression and may ultimately lead to more effective treatments and prevention programs to better manage this disease.

Declaration by Author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Peer-Reviewed Publications

1. **Betts JA**, French JD, Brown MA and Edwards SL (2013), 'Long-range transcriptional regulation of breast cancer', *Genes, Chromosomes and Cancer*, vol. 52(2), p113-125.

2. French JD, Goussaini M, Edwards SL, Meyer KM, Michailidou K, Ahmed S, Khan S, Maranian MJ, O'Reilly M, Hillman KM, **Betts JA**, (BCAC consortium 200 authors), Brown MA, Chenevix-Trench G. Easton DF, Dunning AM et al. (2013), 'Functional Variants at the 11q13 Breast Cancer Risk Loci Regulate *CCND1* Expression through Long-Range Enhancers', *American Journal of Human Genetics*, vol. 92, p489-503.

3. Glubb DM, Maranian MJ, Michailidou K, Pooley KA, Meyer KB, Nord S, Carlebur S, O'Reilly M, **Betts JA**, Hillman KM, Kaufmann S, Beesley J, BCAC consortium, Brown MA, Ponder BAJ, Chenevix-Trench G, Edwards SL, Easton DF, Dunning AM and French JD (2014), 'Fine scale mapping of the 5q11.2 breast cancer locus reveals three independent risk signals regulating *MAP3K1*', *American Journal of Human Genetics*, vol. 96, p5-20.

Submitted/Due for Submission

1. Bailey PJ, Milevskiy M, **Betts JA**, Dowhan D, Muscat G, Dray E, Shewan AM, French JD, Edwards SE, and Brown M (2014), 'Long-range regulation of *HOTAIR* reveals a network of transcription factors associated with reduced metastasis-free survival in breast cancer', Journal to be decided.

Conference Presentations

2012 Lorne Cancer
Mantra Lorne, Lorne VIC
Poster :*'An Exploration of the 11q13 Breast Cancer Susceptibility Locus Using 3C and 4C Chromatin Conformation Capture Technologies.'*

2013 Australian Society of Medical Researchers student conference
Translational Research Institute, Brisbane, QLD
Poster : *'An Exploration of the 11q13 Breast Cancer Susceptibility Locus Using 3C and 4C Chromatin Conformation Capture Technologies.' – Prize for best poster.*

2013 QIMR student conference
Cedar Creek Conference Facility, Mt Tambourine, QLD
Poster : *'An Exploration of the 11q13 Breast Cancer Susceptibility Locus Using 3C and 4C Chromatin Conformation Capture Technologies.'*

2014 Australian Society of Medical Researchers student conference
Translational Research Institute, Brisbane, QLD
Poster : *'RNA Capture-seq and Chromatin Conformation Capture (3C) technologies identify a novel functional long non-coding RNA at the 11q13 Breast Cancer Susceptibility Locus.'*

2014 Keystone Symposia on Cancer Epigenetics and Transcriptional Regulation
Santa Fe Community Convention Centre, Santa Fe, New Mexico, USA
Poster : '*RNA Capture-seq and Chromatin Conformation Capture (3C) technologies identify a novel functional long non-coding RNA at the 11q13 Breast Cancer Susceptibility Locus.'*

2014 Brisbane Breast Cancer Meeting
QIMR Berghofer, Brisbane, QLD
Talk : *'RNA Capture-seq and breast cancer.'*

2014 School of Chemistry and Molecular Biology Student Conference
University of Queensland, Brisbane, QLD
Talk : *'RNA Capture-seq and Chromatin Conformation Capture (3C) technologies reveal a novel lncRNA at the 11q13 Breast Cancer Susceptibility Locus.' - Prize for best talk.*

2015 Lorne Genome
Mantra Lorne, Lorne VIC
Poster : *'RNA Capture-seq and Chromatin Conformation Capture (3C) technologies identify a novel functional long non-coding RNA at the 11q13 Breast Cancer Susceptibility Locus.'*

No publications are included in this thesis.

Overall Thesis

The project was conceived and designed with the help of my supervisors Dr Stacey Edwards and Dr Juliet French who also provided feedback on the interpretation and presentation of the obtained data. Assistance with statistical analysis was provided by Dr Leesa Wockner (biostatistician at QIMR).

Chapter 1: Literature Review

No significant contribution from others.

Chapter 2: Methods

Routine reagents used to perform experiments were prepared by Edwards/French lab research assistants (Mrs Kristine Hillman, Ms Susanne Kaufmann and Ms Hayley Mills).

Chapter 3: Experiments

3.2.4 Assistance in setting up r3Cseq program for processing the 4Cseq data was provided by Dr Nicole Cloonan.

3.2.5 The 5C experiment and data analysis was performed by Dr Haran Sivakumaran. Initial 3C cell culture and library preparation was performed by myself.

3.2.8 Modified antisense oligonucleotides for silencing of PRE1 eRNA were designed and synthesized by Dr Steve Wilton.

3.2.9 DNA from the TALEN clones was screened using the Sequenom iPLEX Mass ARRAY by Dr Qing Chen. The FISH for PRE1 amplification in T47D cells was performed by Mrs Kristine Hillman. Initial BAC preparation and subsequent microscopy was performed by myself.

Chapter 4: Experiments

4.2.1 The capture and sequencing steps of the RNA Capture-seq were performed by Dr Michael Clark, Dr Tim Mercer and Assoc. Prof Marcel Dinger. Initial cell culture, RNA extraction and sample preparation were performed by myself as was the final Cufflinks assembly of the mapped reads.

4.2.14 The Ion Torrent sequencing of the ChIRP-seq libraries was performed by Dr Darren Korbie and the data processed by Dr Mahdi Moradi. The initial ChIRP process and Ion Torrent library preparation were performed by myself.

4.2.16 Cell culture and siRNA knockdowns for the BO2 Rad51 inhibition assay was performed by myself and the final MTS assay and data analysis performed by Dr Adrian Wiegmans.

Chapter 5: Experiments

5.2.6 Initial preparation of T47D xenograft cells performed by myself. Cells injected into mice and subsequent animal care by Dr Shu Shu Wen and Dr Christina Wong. Mammary glands dissected together with Dr Juliet French. Mounting, staining and imaging of glands performed by myself.

Chapter 6: Final Discussion

No significant contribution by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

This thesis is dedicated to my six year old daughter Imogen, a scientist and poet in the making:

"If everything is made of atoms, then what about dreams and shadows?"

Your efforts in trying to coax me away from the computer to play with you have finally come to fruition. To my wife Nga and two boys Jacob and Reuben, thanks for all your support and understanding over the past four years. Sorry for all those late nights and weekends in the lab when I was 'looking after my cells', as suspicious as that may have sounded!

I would like to extend a huge thank you to my supervisors Juliet French and Stacey Edwards for all their support and advice over the years. The many hours you have spent reading this thesis and my other milestones to give suggestions and feedback have been most appreciated. It must have been a little weird to have a geriatric student who was beyond you in years but of course not in wisdom! I am sure the lab will go from strength to strength in the years ahead.

To the rest of the Edwards/French group, thank you all for your professional advice and personal friendships through my time there and in particular, the research assistants Kris, Susanne and Hayley who helped my project in so many ways. All those times when you gave me technical tips or took a culture or PCR out of the machines for me when I had to leave early to pick up the kids made a huge difference to my ability to finish experiments and complete this PhD. Similarly, the post-docs Dylan, Haran and Mahdi have offered invaluable advice through my time there and I wish you all the best in your careers ahead. To my fellow former students Cherie and Hamish, great working with you and exchanging ideas and I hope your careers follow the desired path.

Thank you also to my associate supervisor Melissa Brown who has given me many sage words of wisdom through the years and also to the rest of the Brown lab: Michael Milevskiy (student extraordinaire), Gurveen Sandhu, Ania Wronski, Brooke Brewster, Annette Shewan, Peter Bailey, Eloise Dray and Darren Korbie. Thanks for never losing your cool Darren despite all those endless iterations of Ion Torrent library preps and sequencing – got there in the end!

I also acknowledge with gratitude the financial support provided by the Commonwealth government through the Australian Postgraduate Award for the duration of my candidature and the Top Up Scholarship provided by QIMR Berghofer for my second and third years.

11q13, breast cancer, gwas, lncrna, chromatin looping, *cupid1*, *cupid2*.

ANZSRC code: 111203, Cancer Genetics, 40%

ANZSRC code: 111201, Cancer Cell Biology, 40%

ANZSRC code: 060199, Biochemistry and Cell Biology not elsewhere classified, 20%

FoR code: 0601, Biochemistry and Cell Biology, 50%.

FoR code: 0604, Genetics, 30%

FoR code: 0699, Other Biological Sciences, 20%

Table of Contents

3'RACE …… 3' rapid amplification of cDNA ends
3'UTR ……3' untranslated region
3C……..chromosome conformation capture
4C…….….circular chromosome conformation capture
5C…….. carbon copy chromosome conformation capture
γH2AX…….…..H2A histone family, member X, phosphorylated on Serine 139
AGRF…… Australian genome research facility
ANKRD16……..…ankyrin Repeat Domain 16
ANOVA……. analysis of variance
ANRIL ……antisense non-coding RNA in the INK4 locus
AP-2……...adaptor protein 2
ATCC ……American type culture collection
ATM……ataxia telangiectasia mutated
ATP……adenosine triphosphate
BAC……bacterial artificial chromosome
BC200…….brain cytoplasmic RNA 1
BCAC ……breast cancer association consortium
BCAS1 …….……..breast cancer amplified sequence 1
BCL2 …….……B-Cell CLL/lymphoma 2
Bp…….……...base pairs
BRCA1…….……breast cancer 1
BRCA2…….…....breast cancer 2
BRD4 ……...……..bromodomain 4
BSA.…….…...…..bovine serum albumen
C5orf35…….……..chromosome 5 open reading frame 35
CAGE ……..…..cap analysis gene expression
CARLo ……..……..cancer-associated region long noncoding RNA
CASP8 ……..…..caspase 8, apoptosis-related cysteine peptidase
CBP ……....cyclic AMP-responsive element binding protein
CCAT1 ……..…..colon cancer associated transcript 1
CCND1……..…..cyclin D1
cDNA ……..complementary DNA
ChIP …….……...chromatin immunoprecipitation
CTCF ……..CCCTC-binding factor
CDKN2A……...cyclin-dependent kinase inhibitor 2A
CDKN2B…….…..…cyclin-dependent kinase inhibitor 2B
CFTR …….……..cystic fibrosis transmembrane conductance regulator
CHARTseq……...…capture hybridization of RNA targets
ChIA-PET……..……chromatin interaction analysis by paired-end tag sequencing
ChIRP ……...……chromatin isolation by RNA purification
COGS ……...…collaborative oncologic gene-environment study
CORE …….....cluster of open regulatory elements
CPT1A …….....carnitine palmitoyltransferase 1A
CRISPR……..……clustered regularly interspaced short palindromic repeats
CUPID ……..……*CCND1* upstream intergenic DNA damage
CYP24A1…….....1,25-dihydroxyvitamin D3 24-hydroxylase
DAPI …….……4',6-diamidino-2-phenylindole
DIRC3 ……...……disrupted in renal carcinoma 3
DMSO……dimethyl sulfoxide
DNA ……..……de-oxy ribose nucleic acid
DNase I HS……...……DNAse I hypersensitivity

DROSHA…….....……drosha, ribonuclease type III
DSB……...…double stranded break
EDTA ……...……Ethylenediaminetetraacetic acid
EGFP……....……enhanced green fluorescent protein
EMSA ……...……electrophoretic mobility shift assay
ENCODE …….....……encyclopedia of DNA elements
eQTL……....……expression quantitative trait loci
ERα…….....… (o)estrogen receptor alpha
ERBB……..……Erb-B2 Receptor Tyrosine Kinase 2
ERE ……..… (o)estrogen response element
eRNA ……enhancer RNA
ESC…….……...…embryonic stem cells
ESR1……...(o)estrogen receptor 1
EXO1……..……exonuclease 1
FACS …….....……fluorescence activated cell sorting
FAL1…….....…focally amplified lncRNA on chromosome 1
FANTOM ……..…...…functional annotation of the human genome
FCS....……foetal calf serum
FDR.....false discovery rate
FGF .....…..……fibroblast growth factor
FGFR2 ……..……..fibroblast growth factor receptor
FISH.....……fluorescence in situ hybridization
FITC....……fluorescein isothiocyanate
GAPDH……..…….……glyceraldehyde-3-phosphate dehydrogenase
GAS5…….....……growth arrest-specific 5
GATA3 ……..…….....…GATA Binding Protein 3
GENCODE…..……...…encyclopedia of DNA elements sub project
GRO-seq.…...……...…global run on and sequencing
GUS…….....……β-glucuronidase
GWAS……....…...…genome-wide association study
H3K4me1/2/3.....histone H3 mono/di/trimethylation at lysine 4
H3K27me3 …….....……histone H3 trimethylation at lysine 27
HEATR6 ……..……amplified in breast cancer protein 1
HEK293……...……human embryonic kidney
HEPES ……...……4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HER2….....…….……human epidermal growth factor receptor 2
HIF…….....……hypoxia induced factor
HOTAIR……...……hox antisense intergenic RNA
HMEC …….....…...…human mammary epithelial cells
iCHAV…….....……independent set of correlated, highly trait-associated variants
IGFBP5……....……insulin-like growth factor binding protein 5
IGHMBP2……....…....immunoglobulin mu binding protein 2
LacZ…….....…β-galactosidase
LCR …….....……locus control region
LD ……..... linkage disequilibrium
LncRNA …….....……long non-coding RNA
LSP1…….....……lymphocyte-specific protein 1
Luc …….....……luciferase
LUNAR ……....……...…leukemia-induced noncoding activator RNA
MALAT ……....……metastasis associated lung adenocarcinoma transcript 1

MAP3K1……..……mitogen-activated protein kinase kinase kinase 1
M-ASO……..……modified antisense oligonucleotides
meRNA……..……multi exonic RNA
miRNA……..……microRNA
MRPL21……..……mitochondrial ribosomal protein L21
MTL5……..……metallothionein-like 5, testis-specific
MTT.....…… (4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
MYC ….....……v-myc avian myelocytomatosis viral oncogene homolog
MYEOV ….....……myeloma overexpressed
ncRNA ……..……non coding RNA
NCOA3…….....…nuclear receptor coactivator 3
NEAT ….....……nuclear enriched abundant transcript
NEB ….....……New England biosystems
NOD ….....……non-obese diabetic
NSG ……..……NOD SCID gamma
NTP……..……nucleoside triphosphate
OCT4 ……..……octamer-binding transcription factor 4
Optim-MEM…..……optimized minimal essential media
ORAOV1 ……...……oral cancer over-expressed
P300 ……..……E1A binding protein
PAK1 ….....……P21 protein (Cdc42/Rac)-activated kinase 1
PALB2……..……partner and localizer Of BRCA2
PARP ……..……poly (ADP-Ribose) polymerase 1
PAUPER ……...……PAX6 upstream antisense RNA
PBS……..……phosphate buffered saline
PC3……..…..……prostate cancer 3
PCR ……..……polymerase chain reaction
PDZK1……..……PDZ domain containing 1
PI …….....…propidium iodide
PIWI……..……P-element induced wimpy testis in *Drosophila*
PPF1A ……..……polypeptide interacting protein alpha 1
PPP6R3 …....……...…protein phosphatase 6, regulatory subunit 3
PR …….....……progesterone receptor
PRE…….....……putative regulatory element
PRC2 ……..……polycomb 2 repressor complex
PRNCR1 ….....……prostate cancer associated non-coding RNA 1
PROMPTs …..……promoter upstream transcript
PSA…….....….....prostate specific antigen
PTEN……..……phosphatase and tensin homolog
PTF1A ……..……pancreas specific transcription factor, 1a
RAP……..…… RNA antisense purification
RFLP.....restriction fragment length polymorphism
RNA ……..……ribose nucleic acid
RNAPolII……..……RNA polymerase 2
RNF115……..……ring finger protein 115
RNF21 ……..……ring finger protein 21
RPKM ……..……reads per kilobase per million
rRNA ……..……ribosomal RNA
SCID……..……severe combined immunodeficiency
SD……..……standard deviation

SEM .....……..……standard error of the mean
SETBP1…….....……SET binding protein 1
shRNA …….....……short hairpin RNA
siRNA …….....……short interfering RNA
SNP…….....……single nucleotide polymorphism
TAD ….....……..……topologically associating domain
TALEN …....……transcription activator-like effector nuclease
TATA .....……5'-TATAAAA-3'
TBP …….....….....…TATA binding protein
TBX5…….....….....……T-box 5
TCGA….....….....……the cancer genome atlas
TERT…….....……telomerase reverse transcriptase
TF …….....……transcription factor
TFF1.....……tre-foil factor 1
TIC …….....……transcription initiation complex
TLS …….....…..……translocated in liposarcoma
TNF …….....…tumour necrosis factor
TNRC9 …….....……CAG trinucleotide repeat-containing gene F9 protein
TOX3 …….....…...……TOX high mobility group box family member 3
TP53…….....……tumor protein P53
TRITC…….....……tetramethylrhodamine
TSS…….....……transcription start site
UCSC …….....……University of California Santa Cruz
U-enhancer …….....……ubiquitous enhancer
UTR…….....……untranslated region
VEGF …….....…...……vascular endothelial growth factor
VEL.....……variant enhancer loci
VSV…….....……vesicular stomatitis virus
ZMIZ1 …….....……zinc finger, MIZ-type containing 1
ZNF365….....……...…zinc finger protein 365

CHAPTER 1

Literature Review

1.1. Breast Cancer

Breast cancer is the most common non-cutaneous cancer in women, with more than 450,000 women worldwide dying of the disease each year [1]. The predisposing factors can be divided into environmental and genetic influences. Environmental factors include obesity, alcohol intake, endogenous hormone exposure and physical activity [2]. The inherited genetic risk possessed by an individual is the strongest single risk factor and forms the basis of this thesis [2, 9]. Current treatment regimes for breast cancer include surgery (local lumpectomy or more radical mastectomy), radiotherapy and anti-oestrogen therapies [10, 11]. Many of the current therapies are inadequate however and there is a pressing need to improve the clinical management of the disease.

Initial classifications divided breast cancer into four intrinsic groups based on the immunohistochemistry and gene expression patterns derived from microarrays of human breast tumours. These are the luminal A (low grade and ERα positive), luminal B (higher grade and ERα positive), HER2 positive (high expression of the *ERBB2* gene) and triple negative cancers (ERα negative, PR negative, HER2 negative) [12]. The advent of high throughput next generation sequencing technologies has also allowed new classification systems to be developed with greater discrimination between the groups giving improved prognostic value [13]. Prominent amongst these is a study by Curtis *et al.* who profiled gene expression and copy number in ~2000 breast cancers and divided them into ten subgroups with distinct molecular drivers and clinical outcomes [14]. Identifying these underlying molecular drivers of cancer is a priority in current cancer research and has led to the development of targeted therapies that directly inhibit tumour growth [15]. Examples in common use include inhibitors of key drivers of breast cancer proliferation such as Herceptin to block the human epidermal growth factor receptor 2 (HER2) and Bevacizumab for the vascular endothelial growth factor (VEGF) [16-18]. It is hoped that further functional analysis of genetic risk loci may improve our understanding of breast cancer and lead to the development of new therapies.

1.1.1. Genetic Susceptibility to Breast Cancer

Genetic variants can be classified into three broad categories: highly penetrant mutations of genes critical for genome stability such as *BRCA1, BRCA2* and *TP53* which impart a lifetime risk of breast cancer of 50-70%; a middle tier of moderately penetrant mutations in genes such as *PALB2* and *ATM* that increase relative cancer risk by two to five fold; and a final group of common but low penetrance variants (*Figure 1.1*) [19]. All of the genes in the high and moderate penetrance group

have a role in DNA damage repair, highlighting the importance of genomic instability in the development of breast cancer [2]. Germline mutations of *BRCA1* and *BRCA2* are estimated to cause around 5% of all breast cancers annually in the United States, typically due to post translational inactivation or truncation of the protein [20]. Such mutations also increase the risk of ovarian cancer before the age of 70 in affected women up to 39% for *BRCA1* mutation carriers and 11% for those carrying a mutation of *BRCA2* [21]. Those women carrying pathogenic *BRCA1* or *BRCA2* mutations are frequently offered prophylactic mastectomy and oophorectomy to prevent the development of cancer or placed under a more intensive regime of surveillance [22].

The high and moderate penetrance mutations only account for around 25% of genetic risk. The remaining 75% is proposed to consist of the common, low penetrance single nucleotide polymorphisms (SNPs) which influence risk in a polygenic manner proportional to the number of low-risk alleles carried by a woman [2, 22-24]. This can lead to a lifetime breast cancer risk of up to 50% for those women carrying a high number of such variants [25]. A priority in cancer research is to identify the SNPs responsible for this risk, thus providing valuable information on the lifetime risk of breast cancer faced by an individual, and also more generally as a basis for uncovering new molecular mechanisms underlying disease [26]. Improved risk stratification allows more personalised breast cancer screening and may indicate the need for prophylactic therapy as is done for *BRCA1* and *BRCA2* mutation carriers, including surgery or oestrogen receptor antagonists for those women found to be at high risk for the development of breast cancer [2].

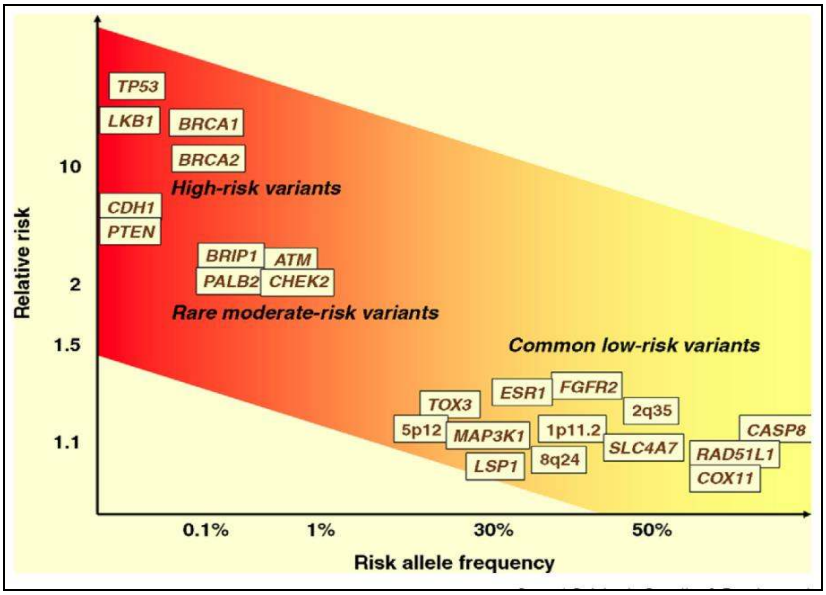


Figure 1.1 Genetic loci identified for breast cancer. The relative risk of breast cancer increases in a logarithmic manner up the y axis as the risk allele frequency increases along the x axis. Genes are clustered into high, moderate or low risk variants [19].

1.2. Genome Wide Association Studies (GWAS)

1.2.1. The Genetic Principles of GWAS

GWAS are a powerful method used to identify genetic variants within the human genome that may be associated with a particular disease. They are based on the principle that DNA is inherited in blocks called haplotypes, a phenomenon called linkage disequilibrium (LD). Marker SNPs are then used as surrogates for the whole haplotype and the association measured between the marker SNP and the phenotype of interest. Due to haplotype structure, there will be many SNPs in very high LD with the marker SNP and these can then be 'imputed' onto the original SNP array as though they had actually been genotyped [27]. If one or more marker SNPs are significantly over-represented in the case group (as compared to the control group), then it suggests that there are causative SNPs for the particular phenotype present within the haplotype containing the marker SNP [28]. It is important to note that the marker SNP used for the initial GWAS is merely a proxy for the entire haplotype block and is not usually itself responsible for the phenotype of interest.

A strategy combining functional, bioinformatic and statistical methodologies is used to identify the potential causative SNPs within a haplotype [29]. The bioinformatics and laboratory based functional components will be discussed in *Section 1.4*. The statistical genetics component is called 'fine mapping'. It requires larger patient numbers and more dense SNP arrays including further variants present in the haplotype marked by the original GWAS SNP [30]. The likely causal SNPs present in a haplotype will be in linkage disequilibrium with other nearby SNPs and those in close LD comprise an iCHAV (independent set of correlated, highly trait-associated variants) [26]. Sophisticated statistical methods such as conditional logistic regression may be employed to aid in the assignation of SNPs into iCHAVs (described in [31]). Typically, SNPs within each iCHAV are then removed from further consideration if they have likelihood ratios less than 100x that of the best candidate SNP, leaving a group of high probability SNPs that can be examined further to find those that are likely to be causal [32]. Finally, programs such as HaploView can be used to compare haplotype structure between populations at a particular locus [33]. This comparison of LD blocks across different ethnic populations can reduce the number of potential causative SNPs due to overlaps in haplotype structure when similar associations are found for a SNP in each group [34].

GWAS studies to date have revealed modest effects of risk SNPs on the phenotype of interest, with most odds-ratios being below 1.5, though there is an allele-dosage effect in which the risk for a

homozygote is double that for a heterozygote. Such low odds ratios are consistent with the theory that cancer risk is largely polygenic and normally involves significant numbers of these low penetrance SNPs rather than the better known high penetrance *BRCA1* or *BRCA2* mutations [19]. Given the importance of low penetrance SNPs to disease susceptibility, their identification and functional characterisation is a priority to better understand the mechanisms underlying breast cancer.

1.2.2. GWAS in Breast Cancer

GWAS have identified more than 90 genetic loci that harbour SNPs predisposing to breast cancer and it is estimated that there may be over 1000 more loci that influence risk but with very low odds ratios [35, 36]. Easton *et al.* performed the first large breast cancer GWAS and provided evidence for five independent loci conferring an increased risk of developing breast cancer [37]. The associated LD block for four of the loci contained several genes that were plausible candidates for disease (*FGFR2, TNRC9, MAP3K1* and *LSP1*), whilst the fifth locus lay in an intergenic region (8q24). A follow up GWAS by Turnbull *et al.* identified a further five risk loci and confirmed the associations with *FGFR2, MAP3K, LSP1* and *8q24* [4]. The loci overlapped genes on 9p21 (*CDKN2A, CDKN2B* and the ncRNA, *ANRIL*), 10q21 (*ZNF365*), 10p14 (*ANKRD16*) and 10p22 (*ZMIZ1*), whilst the marker SNP rs614367 at 11q13 fell in an intergenic LD block of around 166kb in size. Other susceptibility loci for breast cancer that have been identified using by GWAS are summarised in *Table 1.1*.

1.2.3. The Collaborative Oncologic Gene-environment Study (COGS)

Despite the increasing number of GWAS for breast cancer, up until 2013 only a few of the identified loci were fine mapped to locate the candidate causal SNPs [4, 37-44]. Thus, only 30% of the estimated inherited risk of breast cancer had been identified, most of which comprised moderate or high penetrance variants [38]. In an effort to improve this figure, an international collaboration termed COGS was formed. COGS aimed to find novel breast cancer susceptibility loci; confirm previously discovered GWAS associations or proposed moderate penetrance variants; and to fine map existing risk haplotypes [45]. Collectively, COGS identified 64 new breast loci and fine mapped 24 known risk loci using a custom designed iCOGS chip which contained 211,155 SNPs to genotype 114,255 breast cancer cases and controls sourced by BCAC (Breast Cancer Association Consortium) [35].

Locus	Marker SNP	OR	Special Groups	Candidate Gene	Ref.
10q26	rs2981582	1.26	Nil	*FGFR2*	[37]
16q12	rs3803662	1.2		*TOX3*	
5q11	rs889312	1.13		*MAP3K1*	
8q24	rs13281615	1.08		*FAM84B/cMYC*	
11p15	rs3817198	1.07		*LSP1*	
2q35	rs13387042	1.2	nil	Intergenic	[41]
16q12	rs3803662	1.27		*TNRC9*	
10p14	rs1045485	0.89	nil	*CASP8 (D302H)*	[46]
	rs1982073	1.07		*TGFB1 (L10P)*	
5p12	rs10941679	1.19	nil	*MRPS30*	[42]
	rs4415084	1.16		*FGFR10*	
3p24	rs4973768	1.11	nil	*NEK10, SLC4A7*	[39]
17q23.2	rs6504950	0.95		*COX11*	
1p11.2	rs11249433	1.16	nil	*NOTCH2, FCGR1B*	[43]
14q24.1	rs999737	0.94		*RAD51L1*	
6q25.1	rs2046210	1.29	nil	*ESR1*	[44]
9p21	rs1011970	1.09	nil	*CDKN2A, CDKN2B*	[4]
11q13.3	rs614367	1.15		*CCND1*	
10q21.2	rs10995190	0.86		*ZNF365(intron 4)*	
10q22.3	rs704010	1.07		*ZMIZ1*	
6q25.1	rs3757318	1.3		*CCDC170*	
8q24.2.1	rs1562430	1.17		*CASC21, CASC8*	
10p15.1	rs2380205	0.94		*ANKRD16, FBX018*	
11p15.5	rs909116	1.17		*TNNT3*	
19p13	rs8170	1.26	*BRCA1* mutation	*MERIT40, ANKLE1,*	[47]
	rs2363956	0.84		*ABHD8*	
5p15	rs10069690	1.18	TNBC, ERα neg	*TERT*	[48]
9q31.2	rs865686	0.89	nil	*KLF4*	[49]
6q25.1	rs9383938	1.18		*ESR1*	
10q21.2	rs10822013	1.12	nil	*ZNF365*	[50]
12p11	rs10771399	0.85	nil	*PTHLH*	[38]
12q24	rs1292011	0.92		intergenic	
21q21	rs2823093	0.94		*NRIP1*	
6q14	rs17530068	1.12	ERα neg	*intergenic*	[51]
20q11	rs2284378	1.08		*RALY*	
6q25.1	rs9485372	0.9	nil	*TAB2*	[52]
2q34	rs13393577	1.53	nil	*ERBB4*	[53]
1q32.1	rs2290854	ne	ERα neg	*MDM4*	[54]
6p24	rs9348512	0.85	*BRCA2* mutation	*TFAP2A*	[55]

Table continued on next page.

1q32.1	rs6678914	ne	ERα neg	*LGR6*	[56]
2p24.1	rs12710696	ne		*intergenic*	
16q12.2	rs11075995	ne		*FTO*	
Multiple	41 SNPs		iCOGS	multiple	[35]
1q32.1	rs4951011	1.09	nil	*ZC3H11A*	[57]
5q14.3	rs10474352	1.09		*intergenic*	
15q26.1	rs2290203	1.08		*PRC1*	
3p14.1	rs1053338	1.07	nil	*ATXN7*	[58]
7q21.2	rs6964587	1.05		*AKAP9*	
1q21.1	rs124505132	0.95	imputed iCOGS	*NBPF10, RNF115*	[36]
1q21.2	rs12048493	1.07		*intergenic*	
1q43	rs72755295	1.15		*EXO1*	
3p21.3	rs6796502	0.92		*intergenic*	
5p15.1	rs13162653	0.95		*intergenic*	
5p15.3	rs2012709	1.05		*intergenic*	
5q14	rs7707921	0.93		*ATG10, RPS23*	
6p22.1	rs9257408	1.05		*intergenic*	
7q32.3	rs4593472	0.95		*FLJ43663*	
8p11.23	rs13365225	0.95		*intergenic*	
8q23.3	rs13267382	1.05		*LINC00536*	
14q32.12	rs11627032	0.94		*RIN3*	
17q11.2	chr17:29230520:D	0.93		*ATAD5*	
17q25.3	rs745570	0.95		*intergenic*	
18q12.3	rs6507583	0.91		*SETBP1*	

Table 1.1 Breast cancer risk loci identified by GWAS

OR = odds ratio of minor allele relative to major allele. Ne = not estimated. The 41 SNPs from the initial iCOGS study are: rs616488, rs11552449,rs4849887, rs2016394, rs1550623, rs16857609, rs6762644, rs12493607, rs9790517, rs6828523, rs10472076, rs1353747, rs1432679, rs11242675, rs204247, rs720475, rs9693444, rs6472903, rs2943559, rs11780156, rs10759243, rs7072776, rs11814448, rs7904519, rs11199914, rs3903072, rs11820646, rs12422552, rs17356907, rs11571833, rs2236007, rs2588809, rs941764, rs17817449, rs13329835, rs527616, rs1436904, rs4808801, rs3760982, rs132390, and rs6001930 [35].

Analysis of the loci identified by previous GWAS confirmed the association with breast cancer in 23 out of 27 loci, with only weak association demonstrated for three loci and the remaining locus not included on the iCOGS chip. Overall, the SNPs present on iCOGS explained an estimated 28% of familial breast cancer risk with the individual SNPs possessing odds ratios ranging from 1.02 up to 1.26 [35]. A second study using extensive imputation to expand the pool of iCOGs genotypes added another 15 new loci, explaining a further 2% of familial breast cancer risk [36]. A number of iCOGS follow up studies have been performed involving fine mapping of the loci to identify the target genes mediating the SNP-associated risk (detailed in *Section 1.4.2*) [5, 36, 59-62]. One of the most interesting findings to come out of GWAS in general and further confirmed by analysis of the

COGs data, is that the majority of risk-associated SNPs fall in non-coding regions of the genome [34]. This suggests that effects on transcriptional regulation involving distal enhancers or non-coding RNAs may be contributing to inherited disease.

1.3. Long-range transcriptional regulation

1.3.1. Gene Transcription

Control of gene expression is primarily achieved through regulation of transcription initiation, a process dependent on interactions between *cis* regulatory elements and *trans* acting transcription factors (TFs) [63, 64]. Gene promoters are present at the transcription start site of all eukaryotic genes and serve as platforms for binding of the transcription initiation complex (TIC) and its associated RNA polymerase (*Figure 1.2*). Once the TIC is engaged it remains paused at the promoter proximal region until the appropriate regulatory input allows release and subsequent transcription [65]. It has become apparent that many promoters exhibit a complex pattern of transcription, with antisense transcription present at 77% of active genes, though most such transcripts are rapidly cleaved after production to maintain appropriate promoter output [66, 67]. Recent studies have found that this transcription actually derives from a separate core promoter oriented in the antisense direction and that individual promoters themselves have an inherent directionality [68]. Short, unstable, divergent bursts of transcription called PROMPTs (promoter upstream transcripts) are also present at active transcriptional start sites and may have a role in transcriptional regulation or maintaining open chromatin [69].

The chromatin state at eukaryotic promoters as defined by local histone modifications is predominantly uniform across various cell types with tissue specificity being provided by distal enhancer elements [70]. The promoters of housekeeping genes tend to be regulated by local TF binding whilst more dynamic genes require TF binding to distal enhancers which then participate in chromatin looping to activate the promoter [71]. Zhang *et al*. found that 74-87% of promoters regulating such dynamic genes have enhancers specific for that promoter, whilst the more ubiquitous housekeeping gene promoters can interact with a wide range of enhancer elements [72]. Transcription can be repressed by CpG methylation at the promoter, providing a mechanism of stable gene silencing [73]. Inappropriate methylation of tumour suppressor gene promoters however, is a common mechanism driving cancer initiation and subsequent tumour growth [74]. From the risk perspective, studies of genome-wide transcription across 975 human cell types by the

FANTOM group has recently revealed overlap of promoter elements with a number of GWAS SNPs that had not been previously noted with less comprehensive experiments [75]. It is likely that a significant proportion of the SNPs discovered by iCOGS will be fine mapped to gene promoters.

Promoters interact with distal *cis*-regulatory elements as part of the process of transcriptional control [63]. There are four main classes of distal *cis*-regulatory elements; enhancers, silencers, insulators and locus control regions (*Figure 1.2*). Enhancers and silencers consist of TF binding site clusters, with the specific combination of TF binding acting to either repress or activate gene transcription [76]. Their regulatory function is dependent on the availability of the appropriate TFs and some enhancers are even able to act as silencers depending on the local *trans* environment [77].

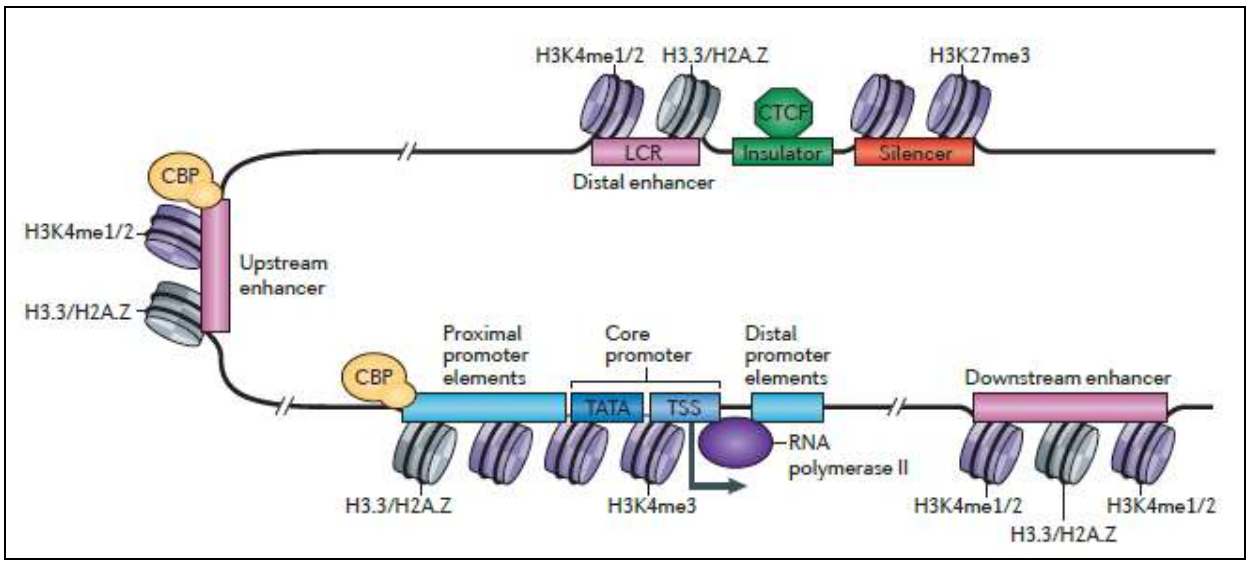


Figure 1.2 Cis-regulatory elements in the human genome. The black right angled arrow indicates active transcription which initiates from the core promoter (blue boxes). Transcription is induced by enhancers (pink boxes) and repressed by silencers (red boxes). Insulators (green blocks) restrict the extent of enhancer or silencer activity. Elements may be grouped into a LCR (locus control region). The grey disks represent nucleosomes. CTCF (CCCTC-binding factor) and CBP (cyclic AMP-responsive element binding protein) are regulatory proteins. The core promoter contains a TATA (5'-TATAAAA-3') sequence and TSS (transcription start site). H3.3/H2A.Z are histone variants. H3K4me1/2 (histone H3 mono/dimethylation at lysine 4); H3K4me3 (histone H3 trimethylation at lysine 4); and H3K27me3 (histone H3 trimethylation at lysine 27) are histone modifications associated with regulatory elements. Adapted from [78].

1.3.2. Cis Regulatory Elements

Enhancers

Enhancers contain a concentration of DNA motifs that are recognised by specific TFs and thus allow the assembly of a protein complex able to activate transcription of an interacting gene [79]. One of the principal differences between promoter and enhancer elements is that they can exert their transcriptional activating properties independent of their orientation or location in the genome relative to a target promoter [80]. They can be described as genomic hubs integrating the TF milieu of a particular cell type to produce the required temporal and tissue specific pattern of gene expression [81]. Enhancers are characterised by their histone marks and may be active (H3K4me1 and H3K27ac), poised (H3K4me1 and H3K27me3) or inactive (H3K27me3), with this state correlating with the activity of genes in their vicinity [82]. The acetyl-transferase p300 is present at both active and poised enhancers and was the first modification used to map enhancers genome-wide using ChIP-chip [83]. Enhancers may also be mapped using the presence of DNAse I hypersensitivity sites which are more general mark of active chromatin [84].

Studies have estimated that the human genome contains over 50,000 enhancers, more than double the number of protein coding genes [70]. Comprehensive approaches to classify enhancers according to their activity and strength usually incorporate data relating to the binding of proteins such as CTCF in addition to histone modifications including H3K4me3 (promoter associated), H3K4me2 (enhancer or promoter associated), H3K4me1 (enhancer associated), H3K9ac and H3K27ac (active regulatory regions), H3K36me3 and H4K20me1 (transcription associated); and H3K27me3 (repressed regions) [85]. The FANTOM group took a different approach however, and assessed levels of transcription from putative enhancers (eRNAs) using CAGE tags in 432 primary cells and 241 human cell lines. Their data indicated that the level of transcription from an enhancer was the best measure of its activity in a particular tissue, allowing improved differentiation between enhancers in an active or poised state [86]. A similar strategy involving global run-on and sequencing (GRO-seq) can also detect active enhancers and has been used in breast cancer cells to demonstrate the effect of oestrogen on enhancer activity [87]. Notably, regardless of the method used, most genome-wide enhancer mapping studies find that GWAS SNPs are enriched in the regions predicted to be active enhancer elements [86, 88-90].

Enhancer Dysregulation and Disease

Regulatory element function relies on the appropriate temporal and spatial binding of TFs. This may be disrupted by SNPs which alter recognition motifs and thus increase or decrease affinity for the TF, consequently altering the activity of the regulatory element [5, 60]. These effects can be screened for using functional assays on a candidate basis or by novel high throughput methods such as gene-centred yeast one-hybrid assays which are particularly applicable for screening the copious data emerging from GWAS [91]. Pathogenic mutations can be found at enhancer elements leading to diseases such as pancreatic agenesis caused by *PTF1A* enhancer mutations [92], or congenital heart disease from mutations in *TBX5* enhancers [93]. Complete deletions of regulatory elements have also been reported with the best known being at the β-globin locus where a large deletion results in β-thalassemia syndrome [94]. More broadly, mutations in any of the factors involved in enhancer function and looping may be associated with disease as they disrupt the normal *trans* and *cis* regulation of gene expression (reviewed in [95]).

Enhancer Synergies

Two enhancers may have differing effects on a gene with both being required to give the correct temporal-spatial expression of their target [96]. Additionally, some distal enhancers may not show activity when tested in isolation but can be shown to augment the effect of a more proximal enhancer when assayed together [97]. This clustering of enhancer elements is mediated by cohesin, a complex of proteins that form a ring shaped structure around DNA to link sister chromatids or interacting enhancer-promoter segments [98]. Knockout experiments reveal a decrease in interactions between enhancer interactions when cohesion is not present [99]. It is suggested that synergy between multiple enhancers provides robustness in gene expression throughout evolution and may compensate for any inherent failure rate of an individual enhancer [76]. Disease-associated SNPs often affect a number of clustered enhancers in a locus which then target the same gene, giving rise to the 'multiple enhancer variant' hypothesis [100]. This is in agreement with the work by Hnisz *et al*. which reveals enrichment of SNPs in highly active enhancers that are grouped together in super enhancer elements [101]. Investigating the action of such SNPs in producing a phenotype will add to our understanding of how enhancer networks function to regulate gene expression.

Super Enhancers

Enhancers conferring tissue specificity are often grouped into COREs (Clusters of Open Regulatory Elements) of around 20kb in size which are thought to coordinate together to regulate a nearby gene [89]. Recent studies have defined a new class of enhancer called super-enhancers which shows overlap with LCRs or COREs depending on the precise definition used [102]. These are usually found at key cell identity genes and are characterised by extremely high levels of master TF binding, DNAse I hypersensitivity, H3K27ac and Mediator [101, 103]. Studies in multiple myeloma and lymphoblastic leukaemia have shown that super-enhancers are located near key oncogenes and are likely to be important factors driving cancer progression [104, 105]. GWAS risk SNPs are enriched in super enhancers as is seen for the breast cancer risk locus at 6q25 (*ESR1*) [101]. Interestingly, the transcriptional co-activator BRD4 is also highly enriched at super enhancers and inhibition of BRD4 leads to a preferential reduction in activity of super enhancer controlled genes making such enhancers an attractive therapeutic target [104].

Silencers

Silencers are regulatory elements that bind repressor TFs to reduce the transcriptional output of their target genes and are present at a far lower frequency than enhancer elements. This may be because the default state for most genes is inactive and thus enhancer elements are usually required to activate transcription [63]. There appears to be a functional overlap with insulator elements as shown by Petrykowska *et al.* who used luciferase assays to characterise regulatory elements in the *CFTR* locus and found CTCF binding and insulator properties associated with 4/15 of the identified silencers [106]. A more rigorous test for silencer function is through transgenic reporter assays where the putative *cis*-regulatory element is introduced into an embryo along with a reporter gene such as *LacZ*, allowing precise spatial localisation of regulatory activity [107]. Such assays have revealed a network of enhancer and silencer elements controlling limb development in a mouse model and highlighted the delicate balancing act required for appropriate temporal and spatial gene expression [108].

Insulators (CTCF)

Mammalian insulators bind CTCF, the only known insulator protein in mammals [109]. Insulator elements act to limit promiscuity of enhancer elements, preventing unwanted gene activation and can also create heterochromatin domains that repress local gene transcription [110]. CTCF is often found at the boundaries of euchromatic and heterochromatic loci, preventing the spread of epigenetic silencing [111]. Loss of CTCF binding upstream of the $p16^{INK4a}$ gene has been observed in many cancers and is associated with methylation of the gene promoter leading to transcriptional

inhibition [112]. *p16*^{INK4a} down regulation is frequently observed early in breast cancer [113]. Aberrant binding of CTCF is also linked to *BRCA1* promoter methylation [114]. Interestingly, the *CTCF* gene at 16q is within a locus often lost in breast cancer, with low *CTCF* expression being a frequent histological finding in surgical specimens [115]. These phenomena may be explained by new research showing that intra-chromosomal interactions are limited by megabase sized Topologically Active Domains (TADS) that partition the genome into defined regions [116]. CTCF is found at the domain boundaries and deletion of the CTCF binding site leads to ectopic intra-chromosomal contacts outside the TAD and a decrease in pre-existing interactions within the TAD [117]. Disruption of wild type enhancer-promoter interactions in this way has been shown to produce a pathological phenotype in murine models of limb growth [118]. Although GWAS SNPs have been found in CTCF binding sites [119], they have not yet been demonstrated to function by destabilising TAD boundaries and this would represent a novel mechanism of inherited disease.

Locus Control Regions

Cis-elements may also be arranged into Locus Control Regions (LCR). These are regions containing multiple regulatory elements and are binding sites for chromatin re-modellers with nuclear matrix attachment points to induce alterations of the local chromatin architecture [120]. Although they may contain silencer elements, LCRs usually have a potent enhancing effect on their target genes, acting in a position and orientation independent manner (*Figure 1.3*) [121]. The most extensively characterised LCRs are the β-globin and T_H2 cytokine loci [122, 123]. As previously stated, new research suggests that many LCRs may be better described as super-enhancers, however there is not yet consensus as to the definition of such entities and the field awaits further clarification [102].

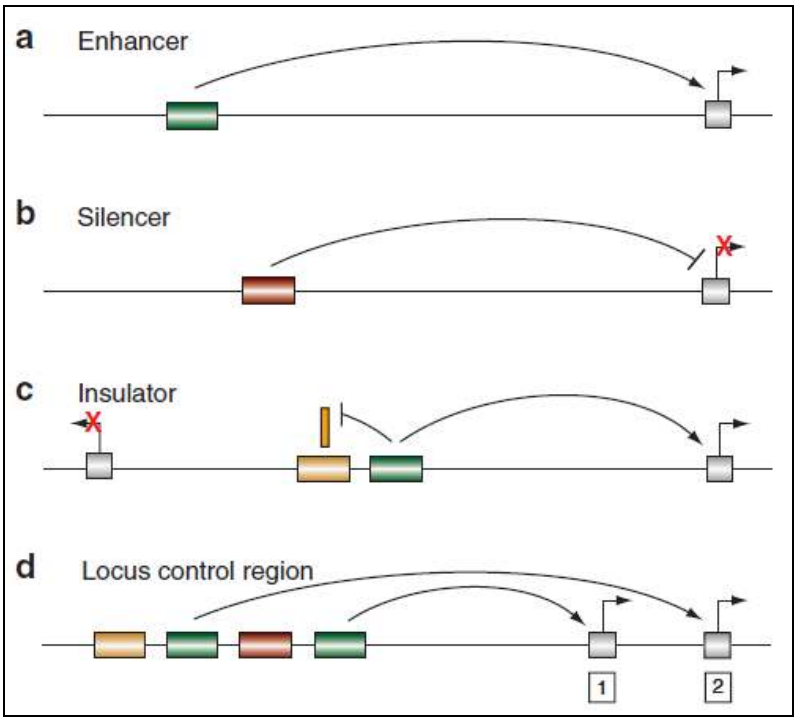


Figure 1.3 Cis-regulatory element function. (a) Enhancers (green box) increase the transcription (right angled arrow) of target genes (grey box). (b) Silencers (brown box) reduce the transcription of target genes. (c) Insulators (yellow box) block enhancer or silencer function. (d) Insulators, silencers and enhancers may be grouped into locus control regions that regulate nearby genes. Adapted from [63].

1.3.3. Trans factors in gene regulation

The binding of TFs to regulatory modules such as enhancers is required for transcriptional regulation. The composition of TFs available in a cell determines which regulatory elements are able to be activated and is a major influence on cell-type identity [124-126]. TF binding may be directed to a specific DNA motif or may involve indirect binding to other TFs already bound to the target element [127]. Collaborative binding of factors is very important, with one study in mouse macrophages finding that more than 70% of enhancers relied on such collaboration to function appropriately [128]. Disruption of a TF binding motif by a SNP may have a snowball effect with reduced binding by one key TF leading to the exclusion of many other co-binding TFs, thus greatly impacting enhancer activity.

Oestrogen Receptor Alpha (ERα)

ERα has been described as 'the master transcriptional regulator of breast cancer phenotype' and is the major target for clinical therapeutics [129]. ERα, which is encoded by *ESR1*, is activated by the steroid hormone oestrogen to mediate oestrogen regulated transcription in various tissues [130]. It interacts with chromatin via direct binding to oestrogen response element (ERE) DNA motifs and also via indirect binding to intermediate TFs [131]. The majority of ERα binding genome-wide is concentrated at distal enhancers providing temporal and tissue specific expression of oestrogen regulated genes [129]. Chromatin looping brings the distal enhancers into contact with their target promoter and many of these loops can be abrogated by ERα knockdown or removal of the enhancer ERE indicating a role for ERα in mediating loop formation [132]. Interestingly, SNPs at 6q25.1 (which contains *ESR1*) are also associated with breast cancer susceptibility and have been identified in several GWAS, however the exact causal variants in this region have not been clarified [44].

FoxA1

FoxA1 is regarded as a pioneer TF which is able to open up chromatin to allow subsequent binding of nuclear receptors such as ERα [133, 134]. It is particularly important in breast cancer biology with knockdown of FoxA1 causing repression of the majority of oestrogen responsive genes and prevention of oestrogen induced progression through the cell cycle [135, 136]. FoxA1 is able to de-methylate DNA at repressed enhancers and induce mono-methylation of adjacent histone tails to give the poised H3K4me1 epigenetic state indicative of enhancer activity [137]. It is also thought to be directly involved in a subset of breast cancers that either exhibit amplification of the FoxA1 genomic locus or possess somatic mutations in key FoxA1 binding locations [138].

Additional Factors

A number of other TFs are required for appropriate functioning of the ERα interactome. These include the AP-2γ cofactor which interacts with FoxA1 and has a role in chromatin looping and transcriptional regulation [139]. AP-2 family members have previously been linked to a variety of cancers [140]. Another co-factor is the steroid receptor co-activator SRC3 which directly binds ERα and mediates interactions with the epigenetic modifiers p300 and CBP [141]. Retinoic acid receptor α and GATA3 are required for full activity of the ERα transcriptional complex and transcription from oestrogen regulated genes is inhibited following their silencing [142-144]. Non-coding RNAs are also increasingly found to be important in *trans* interactions, mediating TF binding and chromatin looping [145]. GWAS risk-SNPs affecting such non coding RNAs could thus have widespread secondary effects on gene transcription across the genome.

1.3.4. Chromatin Looping

As *cis*-regulatory elements may potentially be located over a megabase from their target genes, regulation of transcription often involves long range chromatin interactions which occur through the formation of chromatin loops (*Figure 1.4*) [146]. The loops are mediated by a variety of TFs depending on the tissue, however the mediator and cohesion proteins have been found to play a central role [98, 147]. These two factors and additional tethering proteins allow an enhancer to contact its target promoter by 'looping out' the intervening DNA, resulting in transcription of the target gene (*Figure 1.4*) [148, 149]. Tissue specific loops often incorporate TFs that define that cell type, as is the case with oestrogen receptor mediated looping seen in ERα positive breast cancer cells [132, 150].

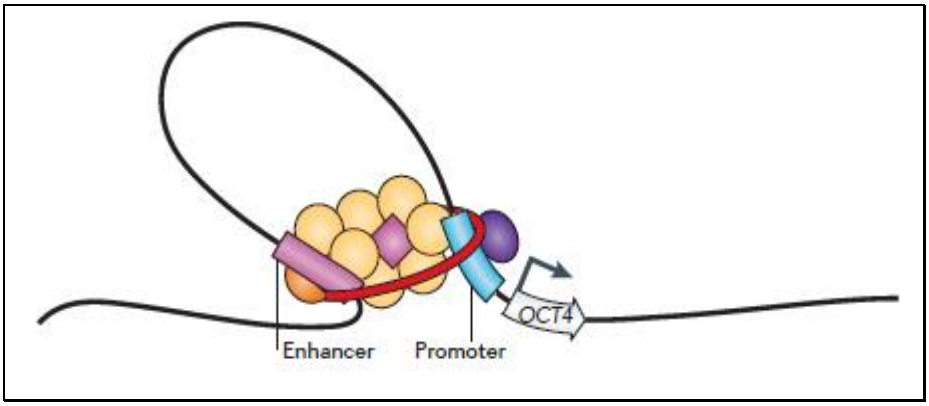


Figure 1.4 Chromatin looping between regulatory elements. The looped DNA is represented by a black line with the promoter of the *OCT4* gene in blue. The pink enhancer box loops around to contact the promoter, directed by a variety of protein co-factors (orange circles) including mediator (pink square) and the loop is stabilised by cohesion (red ring). This allows transcription to proceed (black right angled arrow). Adapted from [78] .

There has been debate over whether gene looping is a primary event or merely a consequence of the transcriptional process. An elegant study by Deng *et al.* resolved this issue by demonstrating that the forced juxtaposition of enhancer and promoter elements using zinc finger nucleases at the β-globin locus results in active transcription of the target gene [151]. The β-globin locus has been the most extensively studied in regards to looping between enhancer and promoter elements, initially using modified FISH methods [152], and then more precisely using chromosome conformation capture (3C) [122]. The tissue specific nature of interactions has also been explored at this locus,

with reconfiguration of looping contacts demonstrated as erythroblasts differentiate [153]. It is important to note however that looping alone will not induce transcription if the appropriate transcriptional co-factors are absent. This is exemplified by the large proportion of loops found to be pre-existing in fibroblasts by Jin *et al.* that required addition of the TNF ligand to allow transcriptional activation to take place [154].

The functional characterisation of breast cancer GWAS hits has confirmed a number of *in cis* enhancer-promoter interactions relevant to breast cancer. These include looping interactions between an enhancer containing risk SNPs and *MYC* at the 8q24 risk locus [155], and similar interactions between two regulatory elements containing risk-SNPs and the *CCND1* gene at the 11q13 risk locus [5]. Previous studies of genes implicated in breast cancer have also revealed functional looping between distal enhancers and the *CDKN1A* promoter mediated by the vitamin D receptor [156], and between the *BCL2* gene promoter and an enhancer element present in its own 3' UTR (untranslated region) [157].

Looping contacts have also long been reported between enhancers and promoters on different chromosomes [158, 159], with Zhang *et al.* finding that 40% of interactions involving enhancers were inter-chromosomal in an embryonic stem cell model [72]. There has however, only recently been evidence provided that such contacts *in trans* can be functional [160]. Multiple interactions *in trans* have been found in experiments on breast cancer cells, including looping involving GWAS loci, however these studies failed to provide experimental confirmation of any biological function mediated by the loops [161, 162]. The majority of post–GWAS studies have only considered interactions *in cis* leaving open the possibility that there may be interactions *in trans* awaiting discovery if the appropriate techniques are used to find them.

1.3.5. Identifying and Characterising Chromatin Interactions

Determining the target genes that interact with a newly identified regulatory element can be a major challenge [163]. The FANTOM consortium found that only 40% of enhancers targeted the nearest gene, with 64% linked to a TSS within 500kb [86]. Therefore, a combination of *in silico* and experimental techniques is required to identify potentially functional enhancer-promoter interactions [164]. The *in silico* approaches will be discussed further in *Section 1.4.1*.

Experimental Approaches

Enhancer-promoter interactions can be directly assessed via chromatin conformation capture (3C) techniques (*Figure 1.5*). 3C, pioneered by Dekker in 2002, is a method based on an initial formaldehyde step to crosslink the DNA and proteins, followed by restriction enzyme digestion and re-ligation to create a DNA fragment composed of two stretches of DNA that were in close proximity at the time of fixation [165]. Identification and quantitation of the novel junctions created is then assessed by quantitative PCR (qPCR), chip hybridisation or sequencing. The relative number of novel junctions present at a particular genomic location gives information as to the structure of local chromatin interactions (reviewed in [166]). Subsequent Sanger sequencing of interacting fragments can also be performed to determine whether the detected interaction preferentially involves a particular allele (allele specific 3C).

Interactions found by 3C may represent random collisions of nearby loci, functional interactions of regulatory elements, or a less specific interaction where two loci share a common sub-nuclear space within such structures as transcription factories or nuclear speckles [167]. On a broader scale, the 3C techniques have revealed a hierarchy of chromatin interactions which range up through topologically associated domains to the compartmentalisation of chromatin into territories within the nucleus [168]. 3C and its related techniques have revolutionised the study of enhancer-promoter interactions but need to be coupled with functional assays to determine whether the observed interactions are functionally relevant and actually alter transcriptional activity [169]. The most common method to demonstrate a potential *in vivo* function is through the *in vitro* use of luciferase plasmid constructs, where the candidate promoter is cloned upstream from a luciferase reporter gene and transfected into a target cell. An increase in luciferase transcription following inclusion of the enhancer in the construct indicates that the enhancer is able to drive promoter activity in that cell type [170].

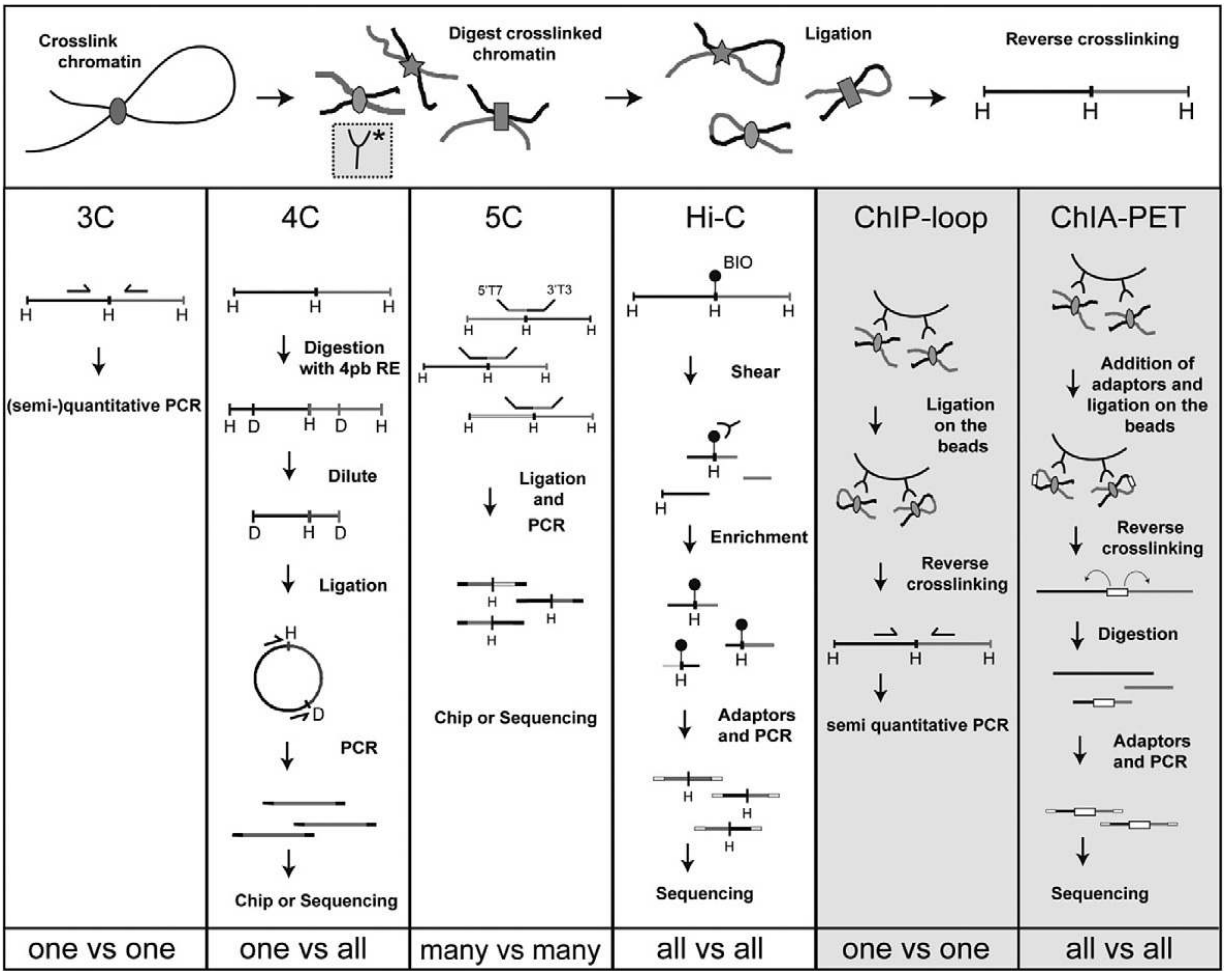


Figure 1.5 The family of Chromosome Conformation Capture (3C) techniques. The top panel shows the initial library preparation common to all the techniques. Tethering proteins are represented by the shaded shapes and the recognition sites for restriction enzymes (RE) used to digest the DNA are indicated by the letters H for the 6bp RE *HindIII* and D for the 4bp cutter *DpnII*. Adapted from Dekker *et al.* [166].

4C (Circular Chromosome Conformation Capture)

3C requires primers to be designed against all the restriction digest fragments that are to be tested for interactions and thus involves an initial prediction as to which interactions are likely to be present [166]. To locate interactions involving a nominated bait region (usually an enhancer or promoter) in an unbiased manner, requires the use of 4C. As described in *Figure 1.5*, it involves a step where the interacting DNA fragments are circularised. Primers facing away from each other are then designed at either end of the bait region, with a subsequent PCR step to amplify any interacting DNA forming part of the circle alongside the bait sequence [171]. The power of this technique has been greatly enhanced by next-generation sequencing which allows the creation of a genome wide map of loci that interact with the genomic element being studied [172]. It is of particular use to investigate those GWAS loci where the candidate causal SNPs can be mapped to one or two

regulatory elements allowing genome-wide interaction data to be generated from a single experiment [173]. It has not yet been employed to investigate breast cancer GWAS loci and should prove a powerful tool for post-GWAS analysis.

5C (Chromosome Conformation Capture Carbon Copy) and HiC

Another related technique called 5C (*Figure 1.5*) was used extensively by the ENCODE project and is able to analyse thousands of chromatin interactions simultaneously, allowing a three dimensional map of chromatin to be constructed [6, 174]. HiC extends the 5C technique genome wide by pulling down and sequencing all interacting fragments (*Figure 1.5*), and has revealed that chromatin is arranged in a 'fractal globule' conformation with active and inactive domains clustering separately in the nucleus [175]. HiC may be combined with a capture step to enrich the sequencing library for interactions containing specified regions of interest such as GWAS loci. This 'Capture-HiC' greatly increases the obtained resolution and allows multiple interacting regions to be investigated simultaneously [162]. Capture HiC shows great promise for post-GWAS work and has been used to follow up risk-SNPs at both breast and colorectal cancer loci [162, 173].

ChIA-PET (Chromosome Interaction Analysis with Paired End Tags)

3C techniques can also be combined with a ChIP step to pull down interacting fragments bound to a particular protein, a process called ChIP-loop [176]. Chip-loop can then be scaled up to analyse chromatin interactions genome-wide that are bound by that target protein, a process called ChIA-PET. This was first performed using ERα as the antibody target and uncovered a looping network of 689 interacting regions mediated through binding of the ERα receptor [161]. It confirmed the involvement of genes previously known to be important in breast cancer and revealed many more requiring further study, providing an invaluable resource for breast cancer research. It is of particular interest in post-GWAS work as it has been found that many GWAS SNPs overlap major regions of interaction [177]. ChIA-PET can thus be used to prioritise SNPs for further characterisation and simultaneously determine which genes are controlled by the regulatory elements harbouring the risk SNPs. Interestingly, further ChIA–PET studies using RNAPolII as the pull-down target in MCF7 cells have revealed that 90% of the PolII binding sites actually lie proximal to known promoters and almost half of these display RNA production indicating the presence of ncRNAs or transcribed enhancers [178]. This suggests that there are many uncharacterised non coding transcripts that may be interacting with regions containing risk-SNPs and hence such transcripts could be mediating some of the observed risk.

GWAS and Chromatin Conformation Capture

A weakness of previous post-GWAS studies is that they either rely solely on *in silico* evidence to determine interactions between enhancer containing GWAS SNPs and potential gene targets [36, 61], or they merely employ 3C to confirm a proposed interaction with a nearby gene [62, 155]. This may be misleading given the fact that 60% of enhancers do not interact with the closest gene and there are thus potential long range interaction partners that may be overlooked [86]. Techniques such as 4C-seq, 5C, ChIA-PET and Capture-HiC allow an agnostic approach to determining such interactions and should form integral parts of future post-GWAS functional studies.

1.4. Mechanisms Underlying SNP Associations and Breast Cancer

As mentioned previously, 88% of fine mapped GWAS SNPs fall in intronic or intergenic regions and 71% overlap a DNAse I hypersensitivity site [3]. This suggests that they lie in regulatory elements and can thus cause transcriptional dysregulation of target genes [163]. Several recent papers by my group and others have identified the target gene and explored potential underlying mechanisms at GWAS breast cancer risk loci. These include 10q26 (*FGFR2*) [62], 2q35 (*IGFBP5*) [60], 5q11 (*MAP3K1*) [59], 2q33 (*CASP8*) [61], and 11q13 (*CCND1*) [5]. The 11q13 study will shortly be described in more detail as it forms the basis for this thesis. Three other studies followed up a number of loci simultaneously; Dryden *et al.* (2q35, 8q24 and 9q31) [162]; Khan *et al.* (42 SNPs involved in miRNA regulation) [179]; and Michailadou *et al.* (15 loci including 18q12.3 and 1q21.1) [36]. A variety of technical approaches may be used to characterise risk loci identified through GWAS and a recommended workflow for such analysis and interpretation has recently been outlined by Edwards *et al.* (*Figure 1.6*) [26]. The genetic fine mapping component has been previously discussed in section 1.2.1.

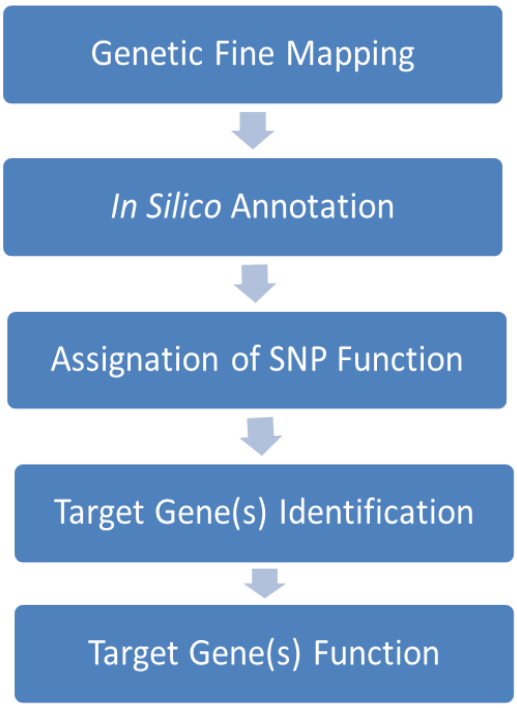


Figure 1.6 A workflow for characterising GWAS risk loci. Adapted from [26].

1.4.1. In Silico Annotation

The initial phase of functional characterisation typically includes the use of publically available ChIP-seq or DNAse I hypersensitivity data sets to predict the location of regulatory elements and allow prioritisation of candidate SNPs (reviewed in [26]). The ENCODE and ROADMAP projects are major resources for such work as they cover chromatin structure, histone modifications, transcription, chromatin interactions and TF binding giving functional predictions for over 80% of the genome [3, 180, 181]. This information can be accessed directly or via the many web-based tools integrating such datasets with SNP annotations, allowing the data to be used in predicting the effect of SNPs at defined disease loci. These tools include HaploReg [182], RegulomeDB [183], GWAS 3D [184], and FunciSNP [185].

The features most commonly used to identify regulatory elements are DNAse I hypersensitivity and the binding of TFs such as p300 and CBP. These are general markers of an enhancer or promoter, with the more specific histone modifications described in *Section 1.3.2* (also *Figure 1.7*). The simultaneous appearance of such marks at a distal enhancer element and nearby promoter can be used as evidence that they are likely to be interacting [85, 186, 187]. The PreSTIGE (Predicting Specific Tissue Interactions of Genes and Enhancers) method developed by Corradin *et al.* is a variation on this approach and combines H3K4me1 ChIPseq with RNAseq expression data across a

wide variety of cell lines. A consistent association observed across cell lines between the gain of a H3K4me1 mark at an enhancer with active transcription from its potential target gene provides evidence for their interaction [100]. More recently however, the demonstration of correlated transcription between an enhancer and nearby genes has been shown to provide even more reliable evidence that the two are interacting and that the enhancer is active [86]. Datasets of large scale chromosome conformation capture type experiments are also available to search for chromatin interactions, including multiple ChIA-PET datasets in MCF7 ERα positive breast cancer cells and a hi-resolution HiC datasets in the HMEC non-cancer breast cell line [161, 177, 188].

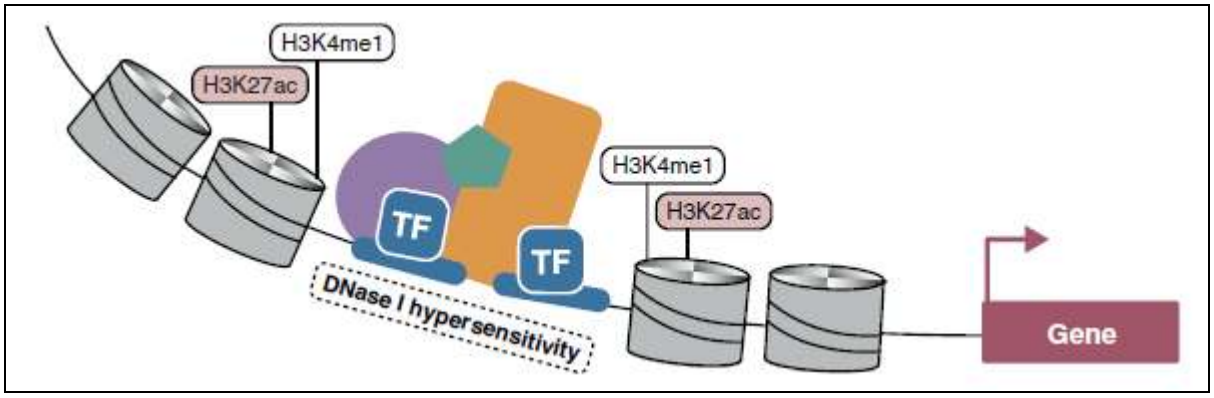


Figure 1.7 Characteristic marks of active enhancers. The identification of enhancer elements can be performed by mapping histone marks (H3K4me1 or H3K27ac) on nucleosomes (gray cylinders), open chromatin (DNase I hypersensitivity) or TF binding (p300, CBP) to DNA motifs (blue oblongs). The coloured purple, green and yellow shapes represent proteins of the pre-initiation complex. Adapted from [189].

The rapid rise in the number of available genome-wide datasets mean that increasingly accurate functional predictions can be made for candidate causal SNPs before any confirmatory lab work has been performed. Michailadou *et al.* used an imputation approach to expand the SNP genotypes used in the iCOGS chip from 211,155 up to over 11,000,000 and found 15 new breast cancer susceptibility loci [36]. They then combined breast cancer cell data from Corradin *et al.* (H3K4me1) [100], and Hnisz *et al.* (H3K27ac) [101], with ChIA-PET ERα or RNAPolII interactome datasets to prioritise SNPs which fell in an enhancer element and consistently interacted with a gene promoter [161, 177]. This process predicted three SNPs at the 1q21.1 locus to lie in enhancers interacting with either the *RNF115* or *PDZK1* genes, and another set of SNPs present in an enhancer at 18q12.3 that interacted with the *SETBP1* gene.

The Hnisz *et al.* (H3K27ac) dataset was also used by Lin *et al.* to study SNPs at the 2q33 locus and provided evidence that several SNPs fell in enhancers of the *CASP8* gene [61, 101]. An analysis of eQTL (expression quantitative trait locus) data then supported an effect of the SNPs on *CASP8* expression. eQTLs are SNPs that are associated with a change in expression of a nearby gene (usually within a megabase) and often represent disruption of an enhancer element *in cis*, though they also may act *in trans* [26]. An interesting study by Li et al. used eQTL datasets from breast tumours to confirm the hypothesis that many breast cancer risk SNPs are also eQTLs, either acting *in cis* on nearby genes in their locus or *in trans* by affecting the expression of TFs such as *MYC* or *ESR1* (oestrogen receptor α-1) [190]. They were unable to confirm effects *in trans* but found significant associations *in cis* for the 2q35 (*IGFBP5*), 5q11 (*C5orf35*) and 16q12 (*TOX3*) loci. Although these *in silico* analyses can be of great assistance in prioritising SNPs for further study, more definitive assignment of a causal SNP requires a direct functional approach to confirm that the properties predicted *in silico* are reflected first *in vitro* and ultimately *in vivo*.

1.4.2. Assignation of SNP Function

The information gained from an *in silico* interrogation of the locus is combined with direct experimental approaches to determine the mechanism by which the SNPs may act to alter breast cancer risk. The most frequently observed effect is a change in TF binding, usually due to the SNP affecting a specific TF recognition motif [191]. Alterations in TF binding can be experimentally confirmed using ChIP-seq to compare differential enrichment between the alleles or with EMSAs (electrophoretic mobility shift assays) which assess the ability of a DNA sequence to be bound by a target protein, thus changing its speed of migration through a gel [192]. Alterations in TF binding may lead to the disruption of tissue specific chromatin looping or changes in activity of an affected regulatory element [26].

Allele specific looping

The most robust breast cancer association to date is with *FGFR2* at 10q26 (p value $2x10^{-76}$) [37]. Fine mapping of this locus identified three iCHAVs associated with the risk of developing ERα positive breast cancer. As part of the functional follow up of risk SNPs, preferential binding of the FoxA1 TF to the one risk allele was observed and E2F1 binding to a second risk allele [62]. Stronger binding of FoxA1 led to an increased recruitment of ERα to the locus. Notably, altered FoxA1 binding appears to be a common mechanism by which SNPs can influence ERα positive breast cancer. Cowper-Sallari *et al.* found an enrichment of breast cancer associated SNPs altering the binding affinity of FoxA1 at distal regulatory sites of the ERα cistrome, leading to allele specific

changes in gene expression [193]. One of the mechanisms may involve allele specific chromatin looping with changes in enhancer-promoter interactions underlying the altered gene expression. This was seen by Ghoussaini *et al.* at the 2q35 locus where a regulatory element containing the cancer protective allele preferentially bound FoxA1 and was involved in increased looping interactions with the *IGFBP5* gene promoter compared to that seen for enhancers containing the risk allele [60]. Given that they had previously demonstrated that the regulatory element acted to enhance *IGFBP5* promoter activity, such an increase in interaction frequency would be expected to drive *IGFBP5* expression [148].

Dysregulation of regulatory elements

Changes in TF binding may also affect the ability of distal regulatory elements to enhance or repress activity of their target promoter [127]. This was seen at the 5q11.2 locus where the risk allele preferentially bound GATA3, which was then associated with an increase in activity of the surrounding enhancer. Another risk allele at the locus increased the activity of a different enhancer, with both enhancers acting on the *MAP3K1* gene to increase promoter activity [59]. A similar alteration in enhancer function was found for the risk SNPs at the 11q13 locus leading to reduced activity of the *CCND1* gene [5].

Direct Disruption of Coding Sequences

A SNP may directly affect gene function by altering the sequence of transcribed RNA, thus creating a non-synonymous change in amino acid codons, disrupted miRNA recognition sites or a structural change in the RNA itself. One of the few variants that directly affect gene function by disrupting an amino acid codon was found at 1q43. This mapped to an exon of the DNA damage repair gene *EXO1*, causing a substitution of an amino acid that was predicted to be deleterious to normal protein functioning [36]. SNPs may also have effects on transcribed RNA that is not mediated by changes in protein composition. Wan *et al.* defined riboSNitches as SNPs that can alter RNA structure and thus influence the ability of the RNA to bind micro RNAs, undergo correct splicing or properly function as a long non-coding RNA [194]. GWAS risk SNPs were enriched for riboSNitches and it seems likely that this will prove another important mechanism by which these SNPs alter normal cellular functioning and promote a disease phenotype. A retrospective analysis of iCOGS data was performed by Khan *et al.* to search for SNPs affecting miRNA binding sites and found five SNPs significantly associated with the risk of breast cancer [179]. These affected genes including the previously discussed *CASP8* at 2q33 and also *DROSHA* which is involved in miRNA processing [195].

1.4.3. Genome Editing Techniques

It is of critical importance to keep a clear chain of causation from the initial fine mapping of a risk SNP through to a demonstration of its functional effect and not to fall into the trap of prematurely declaring SNP functionality and causality [34]. Genome editing techniques are likely to prove valuable in this regard and should form a part of any post GWAS functional follow up study [26]. These allow the creation of parallel cell lines or model animals that only differ in the presence of the SNPs of interest and allow a direct assessment of the SNP effects in their native genomic context [196]. Due to their low cost and ease of use, CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) are likely to be the principle technique used in the future, taking over from the previously used Zinc finger nucleases and TALENS (Transcription Activator-Like Effector Nucleases) [197]. These tools have been previously utilised in post-GWAS work to investigate the effects of a SNP linked to foetal haemoglobin levels and two SNPs linked to Parkinson's disease [198, 199].

1.4.4. The 11q13 Breast Cancer Susceptibility Locus

The first fine mapping and functional characterisation study of a breast cancer susceptibility locus following the iCOGS GWAS release was performed by our group in 2013 [5]. This involved the 11q13 region originally identified in the 2010 Turnbull *et al*. GWAS, with the risk SNP (rs614367) tagging a susceptibility locus for ERα positive breast cancer within a large gene desert [4]. The genetic fine mapping of this locus was performed as part of the iCOGs chip, genotyping 731 SNP variants within the tagged haplotype block for 89,050 European subjects in 41 studies and 12,893 Asian subjects from 9 case-control studies. This revealed three independent signals (iCHAVs) associated with ERα positive breast cancer, though the effect size was far smaller for iCHAV3 and it was not followed up further. iCHAV1 had four candidate causative, highly correlated SNPs (rs661204 = SNP1, rs78540526 = SNP2, rs554219 = SNP3 and rs657686 = SNP4) and iCHAV2 had only one SNP (rs75915166 = SNP5) remaining after the selection process. The SNPs were located in two putative regulatory elements (PRE1 and PRE2) possessing the characteristic epigenetic and TF binding profiles of enhancers or silencers as shown in *Figure 1.8* [5].

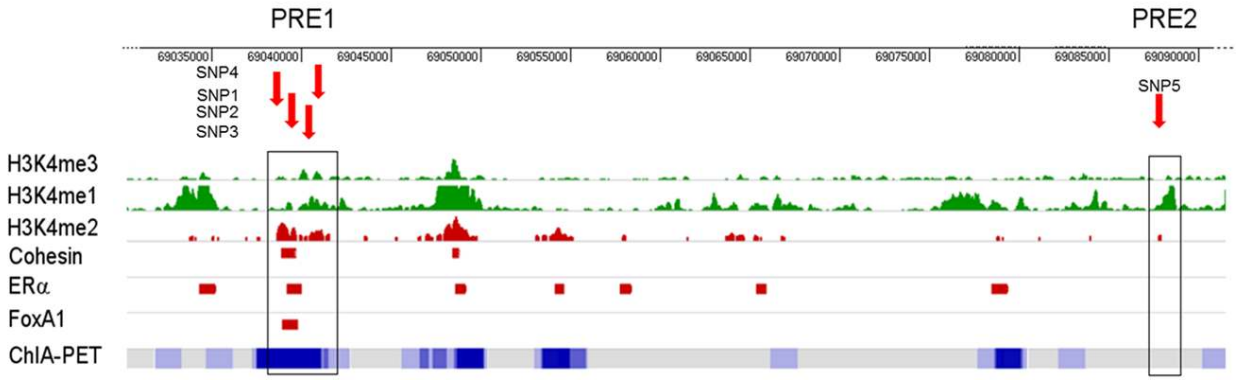


Figure 1.8 The 11q13 breast cancer susceptibility locus. PRE1 contains four candidate causal SNPs and PRE2 contains one candidate causal SNP (red arrows). Promoter activity is indicated by HeK4me3 enrichment (1st green track). Enhancer activity is indicated by H3K4me1 (2nd green track) and H3K4me2 (1st red track) marks (GEO #GSM594606). TFs: ERα (GEO #GSM365926), cohesion (ArrayExpress; #ETABM-828) and FoxA1 [133] are marked as red rectangles. ChIA-PET data derived from [161], is represented as a heat map, with the darker blue representing increased numbers of chromatin interactions. Adapted from [5].

One element (PRE1) was shown to act as an enhancer of transcription and the other (PRE2) as a silencer. Chromosome conformation capture experiments revealed interactions between these regulatory elements and the nearby cell cycle gene cyclin D1 (*CCND1*), one of the most commonly amplified genes in breast tumours and a known oncogene [200]. ERα mediated ChIA-PET data also confirmed the interactions which were consistent with previous studies showing that the action of oestrogen on *CCND1* is a primary driver for ERα positive breast cancer proliferation [126, 161]. Luciferase assays demonstrated that two of the SNPs led to reduced *CCND1* promoter activity *in vitro*. This finding that the SNPs reduced *CCND1* promoter activity was unexpected given that *CCND1* is usually over-expressed in ERα positive breast cancer, however two subsequent studies demonstrated that cyclin D1 associates with the Rad51 protein and has a critical role in oestrogen mediated DNA damage repair [201, 202]. This provides a putative mechanism of disease as interference with these pathways would be consistent with the associated increase in breast cancer risk.

CCND1 was chosen as the likely target of PRE1 and PRE2 due to evidence of interactions provided by the ChIA-PET data and by the fact that it was such a well characterised oncogene. It may not be the only gene mediating risk however, as the ENCODE study found that distal regulatory elements such as PRE1 and PRE2 may interact with greater than ten transcription start sites [6]. This seems particularly likely for PRE1 given that it is such a hot spot for ChIA-PET chromatin interactions

(*Figure 1.8*) [161]. Investigations at 8q24 which is a multiple cancer risk locus much like 11q13, have found other targets for the causal risk SNPs apart from the primary candidate *MYC* oncogene. These include several lncRNAs in the 8q24 gene desert which appear to be involved in both cancer risk and subsequent tumour proliferation [203, 204]. Further characterisation is similarly required for the 11q13 susceptibility locus to uncover additional protein coding genes or non-coding RNAs that may be mediating the breast cancer risk conferred by the five risk SNPs.

1.5. Long Non-Coding RNA

It is now generally accepted that the human genome is pervasively transcribed, with the pilot ENCODE study demonstrating primary transcription involving 74-93% of all bases [205]. As well as representing splice variants of known genes, such transcription encompasses a variety of non-coding RNAs (ncRNAs) which may be functional in themselves or be the precursors of smaller, active RNA molecules [206-208]. NcRNAs can be divided into two broad groups; short, regulatory RNAs and long (>200bp), non-coding RNAs (lncRNAs). Short RNAs include short interfering RNAs (siRNA), PIWI-acting RNAs (piRNA) and microRNAs (miRNA), with many new subgroups added every year [209]. A number of miRNAs have been linked to breast cancer development and metastasis, with tumour miRNA profiling becoming increasingly used as a prognostic indicator and also to identify potential therapeutic targets (reviewed in [210]). LncRNA biology however is still in its infancy and functions have been difficult to assign given their vast numbers and frequent redundancy [211] .

1.5.1. The Function of LncRNAs

Although lncRNAs are estimated to comprise around 80% of genomic transcription, there is debate as to the biological relevance and function of such transcripts with some arguing that much is merely transcriptional noise [212-214]. Guttman *et al*. have provided evidence against this hypothesis, firstly by identifying over 1600 large ncRNAs in mice with strong evidence for conservation and functionality, and secondly by silencing 147 lncRNAs in embryonic stem cells which led to significant effects on gene expression networks in 93% of cases [215, 216]. Similarly, Ponjavic *et al*. analysed 3122 ncRNAs in mice and found evidence of purifying selection consistent with functional transcription [217]. Blurring the issue of sequence conservation being a proxy measure of transcript function however, is the fact that many transcripts overlap functional enhancers which are themselves conserved, making it difficult to determine whether either or both elements are under active selection [214, 218].

Comprehensive studies by Cabili *et al.* and Derrien *et al.* have curated reference catalogues comprising 8195 and 9277 lncRNAs respectively [219, 220]. These demonstrate that lncRNAs are far more tissue specific than protein coding genes and tend to be found in the nuclear compartment, frequently bound to chromatin [220]. *Figure 1.9* describes an array of intra-nuclear functions for lncRNAs. Their transcription alone has also been shown to directly affect local promoter and enhancer function and such transcriptional interference may be a common mode of action (reviewed in [221]). Although most lncRNAs act in the nucleus, some have cytoplasmic function as has been described for *HOTAIR* which appears to have multiple modes of action [222].

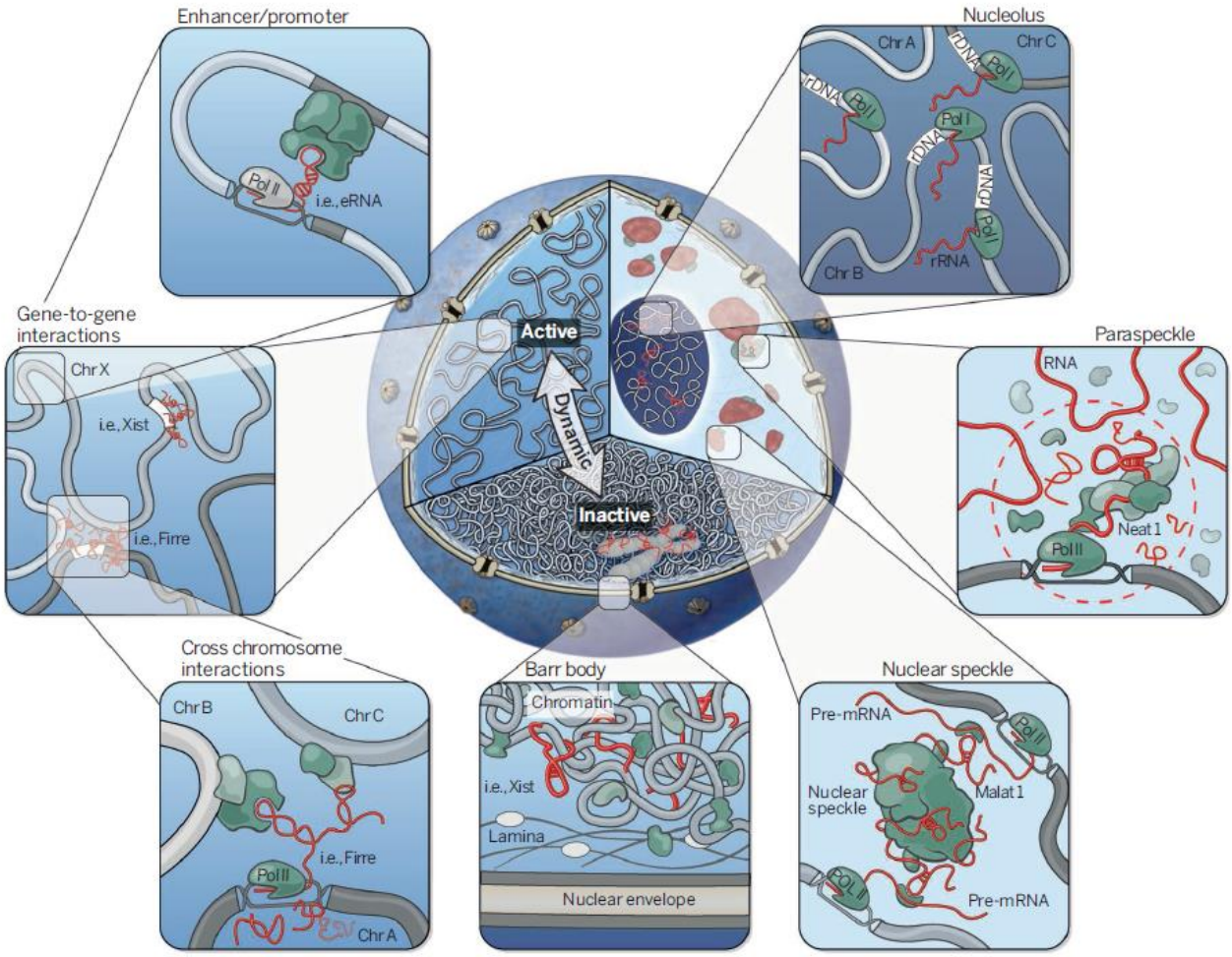


Figure 1.9 Nuclear functions of lncRNAs. The central sphere represents the nucleus containing active (left) and repressed (lower) chromatin with nuclear bodies depicted on the right. Clockwise from upper right: ribosomal RNA (rRNA) is found within the nucleolus; *NEAT1* arranges paraspeckles and *MALAT* is found in speckles which associate with transcribed genes for splicing; *XIST* represses the inactive X chromosome; chromatin looping is mediated by lncRNAs between chromosomes, genes and finally enhancers and promoters. Adapted from [223].

The GENCODE v7 catalog of human lncRNA found a correlation between the expression of lncRNAs and nearby genes, suggesting that many may act in a similar manner to enhancers and regulate local gene expression *in cis* [220]. This supports earlier work by Orom *et al.* who described a class of ncRNAs called enhancer-like RNAs [224]. Proposed mechanisms include guiding the assembly of chromatin modifying complexes to regulatory elements or as an integral component of such complexes (reviewed in [145]). Recent research has shown that a large percentage of lncRNAs arise from enhancers and are better defined as enhancer RNAs (eRNAs) [219].

1.5.2. Enhancer RNA (eRNA)

Transcription from enhancer regions has been documented in the β-globin locus since the early 1990s but was not confirmed on a genome-wide scale until 2010 [225]. De Santa *et al.* demonstrated by RNAPolII ChIP and RNA-seq that 70% of extragenic transcription in macrophages was from enhancers, and this transcription preceded mRNA production from nearby genes [225-227]. Globally, 6.8% of the transcripts in oestrogen (Ez) stimulated MCF7 cells derive from ERα binding enhancers, with maximal production at 40 minutes post Ez exposure [87]. Enhancer RNAs can be divided into two groups; short, non-polyadenylated, bidirectional, 'eRNA' transcripts and longer, polyadenylated, multiexonic transcripts called meRNAs although there is not yet clear consensus on this classification system [228, 229]. Interestingly, intronic enhancers can act as promoters and produce sense meRNAs that represent abridged transcripts of the original mRNA, often contributing over 50% of the total mRNA for that particular gene [229]. The division of regulatory elements into proximal promoters and distal enhancers may therefore be quite artificial as the histone modifications and function of putative enhancers and promoters exist on a spectrum, with strong enhancers having a weak promoter ability and weak enhancers possessing stronger promoter ability when tested using luciferase assays [178]. Antisense meRNAs from intronic enhancers may directly interfere with transcription and favour the production of particular gene isoforms [230].

Enhancer RNAs were initially suggested to be incidental transcription either arising from RNAPolII coming in contact with the open chromatin found at active enhancers, or as the product of recruited RNAPolII as it opens up the chromatin to allow TF binding [218]. Against the former explanation is the observation that RNAPolII binding to enhancers still occurs in the absence of eRNA production [228]. Natoli *et al*. have suggested that the purpose of enhancer transcription by RNAPolII may bring it into the transcription factory where its target gene is being transcribed (albeit at low levels),

thus increasing the chance of the two productively interacting [218]. Regardless of any underlying role, the induction of eRNAs at an enhancer has been shown to be 'the most precise mark of final, functional looping between an activated enhancer and its regulated gene promoter' [231]. Confirming the function of eRNAs is complicated by the lack of consensus as to what actually constitutes an eRNA as opposed to a lncRNA. Given the overlapping spectrum of transcriptional and functional activity exhibited by enhancers and promoters there may however be no clear cut division between the two and new classifications are awaited to clarify our understanding of gene regulation [86, 229].

In support of a functional role for eRNAs, a number of studies have demonstrated reductions in target gene expression following eRNA silencing [232, 233]. Li *et al.* combined eRNA silencing with 5C in oestrogen induced MCF7 cells to show that reduced eRNA expression was associated with a dramatic loss of chromatin interactions, supporting the hypothesis that eRNAs are involved in chromatin looping between their enhancer of origin and its target promoter [234]. Arner *et al*. proposed that looping and transcriptional initiation are also likely to be regulated by eRNAs with their demonstration of a wave of eRNA production preceding subsequent gene expression by around 15mins rather than it being concurrent as previously thought [235]. Additionally, an intriguing new study by Pefanis *et al*. has discovered a role for the RNA exosome in titrating levels of enhancer RNA and antisense RNAs from interacting genes or secondary enhancers to fine tune chromatin interactions and subsequent gene expression [236]. This further supports the involvement of eRNAs in regulating genomic organisation however it is likely that many other functions await discovery. At the 11q13 locus of interest, probable eRNAs arise from regulatory elements upstream of the *CCND1* gene in response to DNA damage, allosterically modifying the histone acetylase inhibitor TLS to reduce *CCND1* expression [237].

From the post-GWAS perspective it follows that any function ascribed to an eRNA may be disrupted by causal risk-SNPs, making the identification and consideration of eRNAs an important part of future analyses. This is emphasized by the fact that the majority of functionally characterised breast cancer GWAS SNPs affect enhancer elements, with a large percentage of these likely to produce eRNA [238].

1.5.3. LncRNAs and GWAS

Despite being more numerous than coding genes, lncRNAs in general have been historically under-represented in post-GWAS studies [239]. They may be affected by GWAS risk SNPs either directly by incorporation of the SNP as part of their RNA sequence, or indirectly where the SNP affects an interacting transcriptional regulatory element. Studies have shown lncRNAs to be strongly enriched with disease associated SNPs, with up to 7% of lncRNAs containing risk-SNPs identified by GWAS [240, 241]. Most GWAS–lncRNA studies in cancer have focussed on the 8q24 multiple cancer risk locus, with the first described transcript being *PRNCR1,* an intronless lncRNA that contains SNPs linked to prostate cancer risk [203]. Knockdown of *PRNCR1* reduced prostate cell viability and androgen receptor transactivation but it is unclear whether the transcript may instead represent an eRNA and thus knockdown is actually acting to disrupt enhancer activity. Similarly, the *CCAT1-L* lncRNA in the same locus is associated with colorectal cancer and appears to arise from a super-enhancer [204]. Another transcript at a locus found in multiple GWAS studies for diabetes, intracranial aneurysm, coronary artery disease and breast cancer is the lncRNA *ANRIL* which may be involved in cell proliferation pathways [4, 242, 243]. A link between a breast cancer associated risk SNP at 2q35 and the expression of the *DIRC3* lncRNA was found by Dryden *et al.* in their capture-HiC paper [162]. GWAS identified SNPs within non-coding transcripts are called snpRNAs and studies by Glinskii *et al.* have shown many to have allele-specific biological effects including the alteration of miRNA levels [244, 245]. Most studies on GWAS risk-SNPs affecting lncRNAs have however provided no direct functional evidence that the SNPs produce a biologically significant effect on the lncRNA and this remains a major deficit in the post-GWAS literature.

1.5.4. LncRNAs and Breast Cancer

An increasing number of lncRNAs are implicated in breast cancer (reviewed in [246]), illustrating the relevance of non-coding transcripts in breast cancer research, though many other less abundant transcripts may await discovery. The first lncRNA to be associated with breast cancer is also the best characterised, the X-inactive-specific-transcript (*XIST*) that inactivates one of the female X chromosomes *in cis* [247]. It was discovered over 20 years ago but has only recently been classified as a lncRNA [248]. The absence of appropriate silencing of an X chromosome is frequently seen in breast tumours and appears to be related to the duplication of the existing active X with loss of its inactivated partner [247]. More recent research in MCF7 breast cancer cells suggests that inappropriate and unstable expression of *XIST* by the active X chromosome may then spread into

nearby chromosome territories, potentially causing dysregulation of genes involved in tumorigenesis [249].

Other lnRNAs of relevance to breast cancer include *H19, MALAT, BC200, GAS5* and *HOTAIR*, as reviewed in Gibb *et al.* [246]. *H19* was the first lncRNA to be shown to be involved in regulation of the cell cycle and is over-expressed in many breast cancers to promote cell proliferation [250]. *MALAT* expression is increased in primary breast cancers [251]. *BC200* is a lncRNA normally only expressed in neurones however it exhibits high expression in invasive breast carcinomas and shows potential as a bio-marker to assist in the diagnosis and staging of tumours [252]. There is also a correlation with high-grade ductal carcinoma *in situ*, raising the possibility that it could be used as a early screening tool, much in the same way as the PSA (Prostate Specific Antigen) is used for prostate cancer [252]. *GAS*5 over-expression in breast cancer cells can trigger apoptosis, raising the possibility that it may function as a tumour suppressor [253].

Perhaps the most topical of the lncRNA family is *HOTAIR* (Hox Antisense Intergenic RNA), which exhibits greatly increased expression (over 125x normal) in around one third of primary breast tumours. This expression can climb more than ten-fold again in metastasis and represents an independent risk factor for breast cancer mortality [254]. *HOTAIR* was discovered by Rinn *et al*. who demonstrated its role in silencing the various *HOX* genes via recruitment of the polycomb 2 repressor complex (PRC2) [255]. From a treatment perspective, the inhibition of *HOTAIR* has been demonstrated to reduce the invasive potential of tumours [254]. Expression profiling of lncRNAs such as *HOTAIR* may also be used to classify tumours and provide prognostic information for patients [256]. Sorensen *et al*. were the first group to do the latter in breast cancer and achieved more than 90% sensitivity and 64% specificity in their classification of patients into metastatic and non-metastatic groups [257]. Though the lncRNA field is very new, the explosion in the number of annotated transcripts makes it highly likely that many more will be implicated in breast cancer biology [220, 241].

HOTAIR stands out amongst those breast cancer associated lncRNAs in that a distinct pathogenic mechanism has been described. The majority of lncRNAs have only been identified due to their over-expression in breast cancer and mechanistic detail is lacking. The precise functional role of these transcripts requires further research, with the ultimate goal of developing novel biomarkers or anti-lncRNA chemotherapeutic agents to improve our management of breast cancer.

1.6. Outstanding Issues and Thesis Aims

This thesis concerned the functional follow up of GWAS risk-SNPs determined to be likely causal variants in the initiation of breast cancer. The identification of genes that may mediate this risk has been historically been undertaken using a candidate gene approach [5, 60, 62]. This may however overlook low probability target genes that had not been previously been implicated in breast cancer. Traditional approaches are also biased towards protein coding genes and have thus neglected the larger non-coding component of the transcriptome. Given that the majority of GWAS risk-SNPs fall in non-coding regions of the genome it is highly likely that a substantial proportion of these SNPs will be affecting non-coding RNA [239]. This project therefore aimed to explore strategies for the characterisation of GWAS loci that maximises the ability to detect functional transcripts of relevance to breast cancer, whether these be coding or non-coding.

The breast cancer associated susceptibility locus at 11q13 was the specific focus, to build on the previous study by our group which identified five candidate causal SNPs in two regulatory elements (PRE1 and PRE2) that interacted with the *CCND1* gene. Chromosome conformation capture techniques were first used to find additional protein coding genes that may also be regulated by PRE1 or PRE2 and prioritise them for further investigation. This identified *CPT1A* and *IGHMBP2* as possible novel gene candidates mediating breast cancer risk in the region. RNA Capture-seq was then used to reveal the full complexity of transcription in the locus and further functional techniques employed to demonstrate that two of the transcripts (*CUPID1* and *CUPID2*) were lncRNAs affected by the risk SNPs contained within PRE1. A follow up chapter provided evidence that *CUPID2* also had many properties consistent with an oncogene and may have a role in driving breast cancer proliferation. This finding addressed another unresolved issue regarding the 11q13 regions which is amplified in around 20% of breast cancer and thought to contain multiple oncogenic drivers which have not yet been definitively determined [8]. *CUPID2* is thus proposed to be a novel driver of the 11q13 amplicon along with the well characterised oncogene *CCND1*.

The overall hypothesis of this thesis is that common genetic variants act by modulating long-range regulatory elements that are found at breast cancer susceptibility loci and that the target genes (coding or non-coding) have a role in the pathogenesis of breast cancer. The evidence presented here supports that hypothesis and the findings presented from the 11q13 locus suggest that many more risk loci will harbor non-coding RNAs that mediate the increased risk of breast cancer imparted by identified causal SNPs.

CHAPTER 2

Materials and Methods

2.1. Cell culture

2.1.1. General conditions

HS578T (ATCC#HTB-126), MDA MB231 (ATCC#HTB-26), MDA MB415 (ATCC#HTB-128), BT474 (ATCC#HTB-20), T47D (ATCC#HTB-133) and MCF7 (ATCC#HTB-22) cells were cultured at 37°C with 5%CO_2, using RPMI 1640 media + L-glutamine (Gibco Invitrogen) with 10% FBS (Gibco Invitrogen) and 10ug/ml antibiotic/antimycotic cocktail (Gibco Invitrogen). 10ug/ml bovine pancreatic insulin (Sigma Aldrich), 1mM Na pyruvate (TRACE Scientific) and 20mM HEPES (TRACE Scientific) was also added when culturing BT474, T47D and MCF7 cells. HMECs (Human Mammary Epithelial Cells) were obtained from Life Technologies and grown in the supplied media.

2.1.2. Fulvestrant/Oestrogen Induction

MCF7 cells were induced with oestrogen (Sigma Aldrich) by seeding them at a confluence of 15% then replacing the media with fresh media containing 10nM Fulvestrant (ICI 182780, Sigma Aldrich) and culturing for 48 hours. The media was then replaced by media containing either 10nM 17B-Estradiol (Sigma) dissolved in DMSO (Ajax Finechem) or an equivalent volume of DMSO vehicle. The cells were grown for a further 12 or 24 hours prior to RNA harvesting or lysate preparation. QPCR was then performed on the synthesized cDNA using intron spanning primers to determine gene expression.

2.2. Luciferase Assays

Promoter (*ORAOV1* or *CUPID*) and enhancer DNA were amplified from MCF7 gDNA using the Kappa HiFi system (Kapa Biosystems) and sub-cloned into the pBLUNT (Thermo Fisher) plasmid before a final transfer into the pGL3-Basic (Promega) plasmid. The promoter fragments were amplified with primers that added *KpnI* and *HindIII* restriction enzyme sites to assist cloning into pGL3 Basic. Constructs containing the SNP variants had been already generated by other lab members for a previous study [5]. Sequencing of constructs by AGRF (Queensland) confirmed them to be correct. Cells were seeded to 80% confluence in 24 well plates together with equimolar quantities of luciferase reporter plasmids, 50ng of pRLTK (renilla) and 1µL of Lipofectamine 2000 (Invitrogen) in a final volume of 600µL per well and incubated at 37°C. The empty vector pUC19 was used to make the final DNA amount (650ng) constant between wells. After 24hrs the media was removed and luciferase/renilla activity measured using the Dual-Glo Luciferase system on a

Beckman Coulter DTX880 multimode detector. The final values were obtained by dividing the luciferase result by the corresponding renilla result to control for variation in cell lysis or transfection efficiency. A one-way ANOVA test with Dunnett's correction for multiple comparisons with reference to a defined construct was used to analyse the data.

2.3. Chromosome Conformation Capture

2.3.1. 3C Library Preparation

Lysates were prepared from cells grown to 70% confluence in 10cm plates, washed with PBS, and fixed for 10mins using 10ml of 1% formaldehyde. The formaldehyde was inactivated with a 10ml 0.125M PBS-Glycine wash then the cells scraped off in 10ml of fresh 0.125M PBS-Glycine. After a 6min centrifuge at 600g at 4°C, the cell pellets were washed with 25ml of cold PBS before another 400g centrifuge for 6 mins at 4°C. The pellet was then resuspended in 10ml of ice-cold lysis buffer (see Appendix) and incubated on ice for 30mins with intermittent mixing. Cell lysis was completed using 10strokes of a Dounce homogenizer and the nuclei spun down with a 6 min centrifuge at 800g at 4°C. The pellets were then resuspended in 120µL of NEB restriction enzyme buffer, 575µL of dH_20 and 24µL of 10% SDS (sodium dodecyl sulphate - Sigma)(0.3% final), mixed and incubated for 30mins at 37° C with shaking. 200µL of 10% Triton X-100 (Sigma)(2% final) was added, the solution mixed and then incubated for another 30mins at 37°C with shaking to inactivate the SDS. Three aliquots of the desired restriction enzyme (purchased from NEB) were then added over a 6 hour period (1500U total for *HindIII*, 1000U for *NcoI*) and the samples incubated overnight with shaking at 37°C. Digestion efficiencies were assessed using both qPCR on a 20µL aliquot of purified digested DNA and agarose gel electrophoresis. 160µL of 10% SDS was then added to each tube and the samples incubated for 30mins at 65°C with intermittent shaking to inactivate the enzymes. After transferring to a 50ml Falcon tube, 5.4ml of dH_20, 750µL 10% Triton, 920µL ligase buffer (see appendix), 80µL 10mg/ml BSA (NEB) and 80µL of freshly prepared 100mM ATP was added and the tubes incubated in a waterbath at 37°C for 45mins followed by 15mins on the bench at room temperature. Ligation of the DNA fragments was performed by adding 2µL of 2000U/µL T4 DNA ligase (NEB) and incubating for 4 hours in a 16°C waterbath, then 30mins on the bench at room temperature. De-crosslinking was performed by incubating overnight at 65°C after the addition of 30µL 10mg/ml proteinase K (Astral Scientific) then 30µL 10mg/ml RNAseA (Sigma Aldrich) was added followed by a 45 min incubation at 37°C. The samples were cleaned up using 8ml of phenol chloroform (Invitrogen) then 8ml of chloroform and the DNA precipitated with 5µL

glycoblue (Ambion), 1ml 3M Na acetate, 5ml water and 24ml of 100% ethanol in a centrifuge tube. After 45min incubation at -80^{o}C, the tubes were centrifuged at 20,000g for 45mins followed by 4-6 washes with 70% ethanol to remove precipitated salts. The pellet was then dissolved in 150µL of Tris-HCL (pH 7.5) and left on a roller at 4^{o}C overnight. Further DNA purification was achieved by the use of Amicon-Ultra spin columns (EMD-Millipore). After library quantitation using qPCR with comparison to diluted gDNA standards, the libraries were aliquoted and stored at -20^{o}C, ready to be used as a PCR template for 3C.

2.3.2. 3C QPCR

300ng of DNA template was used in a qPCR reaction together with 10x MyTaq buffer, 0.05µM bait primer, 0.05µM variable primer, 0.1µL of MyTaq (Bioline), 1.25mM SYTO9 (Life Technologies) made up to 20µL with dH_2O. The reactions were run at 95^{o}C for 2mins; 50 cycles of [95^{o}C for 20sec, 66^{o}C for 30sec and 72^{o}C for 30sec]; 72^{o}C for 4mins then held at 16^{o}C. The PCR products were run out on a 2% agarose and bands corresponding to low Ct results were excised, extracted using the Qiagen gel extraction kit and sent for Sanger sequencing (AGRF QLD) to confirm their identity. Interaction frequencies were determined by adjusting the qPCR converted Ct values according to the relative primer efficiency as measured on serial dilutions of an artificial BAC ligation product library.

2.3.3. Generation of BAC (Bacterial Artificial Chromosomes) 3C Controls

DNA from BAC clones RP11-614E9, RP11-378E8, RP11-156B3 and RP11-825J6 (Invitrogen) was combined to give a total of 10ug, digested by the appropriate 3C enzyme and ligated with T4 DNA ligase (NEB) to create an artificial ligation product library.

2.3.4. Allele Specific 3C

3C library preparation was performed as described previously in MB415 cells using the *HindIII* restriction enzyme, however the final PCR was performed using primers designed to span a ligation junction and a risk SNP of interest, such that the subsequent amplicon contained the SNP as well as DNA from the bait and interacting restriction fragments. The products were then run on a 2% agarose gel, extracted using the Qiagen gel extraction kit and sent for Sanger sequencing (AGRF QLD). The chromatograms were compared to those from PCR products containing the SNP in

genomic DNA to determine whether a particular allele was preferentially participating in the chromatin interaction.

2.3.5. Circular Chromosome Conformation Capture

For the 4C-seq, 2x20ug aliquots of DNA from prepared 3C libraries were taken and digested either with 10U/µL *DpnII* (NEB) or *NlaI* (NEB) in 2ml snap lock tubes. Each tube contained 20ug DNA, 50µL of the appropriate NEB buffer, 5µL of enzyme, 5µL of 100xBSA (NEB) for the *NlaI* samples and dH_20 up to 500µL. The tubes were incubated overnight at 37°C with shaking and then the enzymes heat inactivated at 65°C for 30mins. DNA was extracted with 600µL phenol-chloroform (Invitrogen) followed by a 10min centrifuge at 130000rpm, then a further cycle using 500µL chloroform. The supernatants were transferred to fresh tubes along with 2µL glycoblue (Ambion), 50µL 3M Na acetate and 950µL of 100% ethanol before a 30min incubation at -80°C. After a 20min centrifuge at 13000rpm at 4°C, the pellets were washed using 200µL 70% ethanol then air dried and resuspended in 200µL dH_20. After 80mins on ice, 3.5µL of each sample was run on a 2% gel to confirm adequate digestion and the libraries stored at -20°C pending ligation. Once thawed, 2µL of Quick T4 DNA Ligase was added along with 240µL of 1:2 Quick Ligation buffer (132mM TrisHCl, 20mM $MgCl_2$, 2mM DTT, 2mM ATP, 15% PEG, pH 7.60). Samples were ligated overnight in a waterbath at 16°C and then on the bench for 30mins. The DNA was precipitated using 5µL glycoblue, 50µL 3M Na acetate and 1250µL of 100% ethanol and incubated for 45mins at -80°C followed by a 30min centrifuge at 13000rpm at 4°C. The pellets were washed using 500µL of 70% ethanol, centrifuged for 15mins at 13000rpm at 4°C and air dried for 5mins before resuspending in 100µL of 10mMTris (pH 7.5) and allowing it to dissolve for 30mins at 37°C. An aliquot of each sample was run on a 2% gel to confirm ligation and the remaining sample combined with 800µL of binding buffer and processed using a High Pure PCR Product Purification Kit (Roche).

The purified libraries were quantified using qPCR to compare them to diluted DNA standards of a known concentration and 12.5/25/50/100/200ng/µL dilutions made for use in a PCR reaction to determine the optimal library concentration for the final PCR. Each 25µL reaction contained 16.75µL of dH_20, 2.5µL buffer1, 0.5µL 10mM dNTPs (Bioline), 0.5µL of the 35µM forward and reverse primers, 0.25µL of Expand Long Template Taq (Roche) and 4µL of the diluted library. Cycling conditions were 2mins at 94°C, (30s at 94°C, 30s at 60°C, 3mins at 68°C) x 30 cycles then 5mins at 68°C. The products were run on a 2% gel and the DNA template concentration with the

best spread of PCR products used for the final reaction. This comprised 40µL buffer1, 8µL 10mM dNTPs, 8µL of the 35µM forward and reverse primers, 5µL of Expand Long Template Taq (Roche), 16x the optimal library quantity used in the trial reaction and dH_20 to 400µL. The sample was split into 8 PCR tubes and cycled as follows: 2mins at 94°C, (30s at 94°C, 30s at 70°C, 3mins at 68°C) x10 cycles with the annealing temperature reducing by 1°C per cycle, then (30s at 94°C, 30s at 60°C, 3mins at 68°C) x30cycles and finally 5mins at 68°C. The primers were then digested off by incubating for 90mins at 37°C after adding 40µL of the recommended NEB buffer, 3µL of either *NlaI* or *DpnII* depending on the sample origin and 3.8µL of BSA for the *NlaI* derived libraries. Finally, the samples were split into 200µL aliquots and processed using the High Pure PCR Product Purification Kit (Roche) before being processed by the Ion Torrent fragment kit and sent for Ion Torrent sequencing using a 318v2 chip producing 5-6 million usable reads per library.

2.3.6. 4C Data Processing

The *.bam* files generated from sequencing the 4C-seq libraries were used as input for the published r3Cseq pipeline to normalize the libraries and determine interactions [258]. The output *.bedgraph* files were uploaded to the Galaxy platform (Garvan), filtered using RPKM>10 and clustered using a 20kb cutoff. The clustered files were then intersected to give the final regions of interaction in common between the 4C-seq libraries.

2.4. RNA Experiments

2.4.1. RNA Extraction

Cells were harvested at 50-60% confluency using Trizol (Ambion) and processed according to the manufacturer's instructions. The RNA was then Turbo DNAse treated (Ambion) and further purified using the RNeasy MinElute Cleanup Kit (Qiagen) ready for Capture-seq or RNA-seq library preparation. Quality control involved running the RNA on a QSep RNA column to assess purity and PCR to ensure there was no significant residual DNA contamination. For the latter, a 200ng aliquot of RNA from each sample was used as a template along with the 3C-*GAPDH* primers and MyTaq polymerase (Bioline). The products were run on a 1% agarose gel with a positive and negative control, with no bands expected from the RNA samples.

2.4.2. ***RNA Capture-seq***

The capture step and subsequent sequencing were performed by our collaborators Tim Mercer *et al*., according to their published protocol using capture arrays previously designed by the Dinger lab [259]. Approximately 395kb of the intergenic region between *MYEOV* and *CCND1* was captured on the chip (69,061,622-69,455,873). The *.bam* files were then assembled by myself on the Galaxy (Garvan) platform using cufflinks and viewed using the UCSC and IGV browsers [260].

2.4.3. ***siRNA Knockdowns***

Dharmacon On-TARGET siRNAs were used to knockdown *CCND1, ORAOV1, CUPID1* and *CUPID2* with a validated non-targeting siRNA (see *Appendix Table 8.2* for sequences) as a negative control. The siRNAs were designed against *CUPID1* and *CUPID2* using the Dharmacon siDesign tool (http://dharmacon.gelifesciences.com/design-center/) then transfected into the cells using Lipofectamine 3000 (Invitrogen) at a final concentration of 50nmol. After 48hours the RNA was extracted using Trizol (Ambion) or the cells were processed for luciferase expression depending on the experiment. Efficacy of knockdown was assessed by comparing the gene expression in the knockdown sample against that obtained with the negative control. Seven different siRNAs were trialled against *CUPID1* and *CUPID2*, with the most efficient chosen for subsequent experiments.

2.4.4. ***QPCR for Gene Expression***

RNA was harvested from cells with Trizol (Ambion) following culture +/- siRNA knockdown and cDNA synthesized with random hexamers and the Superscript III kit (Life Technologies). Matching negative control samples were also prepared by omitting reverse transcriptase from the reaction. Gene expression was measured on a RotorGene6000 (Corbett Research) either using Taqman Gene Expression assays (list in *Appendix Table 8.1*) normalising against *B-glucuronidase* (MIM611499; Cat# 4326320E), or qPCR with MyTaq (Bioline) and SYTO9 normalising against *TBP*.

2.4.5. ***ChIRPseq (chromatin isolation by RNA purification)*** (Adapted from [261]).

16 anti-sense DNA probes with 3'biotin TEG were designed for *CUPID1* with an average GC% of 45 and length 20bp, tested for specificity using NCBI Blast and ordered from Integrated DNA Technologies (Singapore). 16 negative control probes targeting *LacZ* were also ordered as per Chu *et al.* The probes were diluted to 100uM concentration in water, divided into even and odd pools

and stored at -20°C. BT474 cells were grown as described previously, trypsinised, washed with PBS x2 then resuspended in fresh 1% glutaraldehyde (Sigma Aldrich), with 10ml per 10 million cells in 50ml Falcon tubes. These were incubated for 10mins in a rotator at room temperature then quenched by the addition of glycine to a final concentration of 0.125M for 5mins at room temperature (RT). The tubes were then centrifuged for 5mins at 400g and the pellets washed with 20ml chilled PBS before a repeat centrifuge. The pellets were finally resuspended in 1ml of PBS, centrifuged and all the PBS removed before flash freezing in liquid N_2 and storage at -80°C until use. The lysates were thawed, resuspended in 10x volume of Lysis Buffer *(Appendix 8.1)* and sonicated in a Covaris S220 (175W, 200cycles, DF10, 15mins) in 130µL batches which were then combined and centrifuged at 16100RCF for 10mins at 4°C to pellet cell debris. 3x10µL aliquots were taken to for input RNA and input DNA, with the other 10µL undergoing phenol-chloroform clean up and the DNA extracted and run on a gel to confirm adequate sonication to 100-500bp.

Hybridisation was performed at 37°C with 2ml Hybridisation buffer (*Appendix 8.1*) and 100pmol of probes added per 1ml of sonicate then incubated overnight in a rotating oven. 100µL of streptavidin C-1 magnetic beads were then washed, resuspended in Lysis Buffer and added per ml of chromatin. The tubes were incubated with shaking for a further 30mins before being placed on a magnet and the supernatant discarded. The pellet was washed 5 times in 1ml of Wash Buffer (*Appendix 8.1*) with 5mins of shaking at 37°C between washes. 100µL was removed for RNA after the final wash, proteinase K (Ambion) treated and the RNA extracted using Trizol (Ambion) to check for enrichment compared with the input RNA that was processed similarly. The remaining pellet was resuspended in 150µL DNA Elution Buffer and incubated for 30mins before removing the supernatant and repeating the process. The 10µL DNA input aliquot was also processed in DNA Elution Buffer along with the post-ChIRP samples. 15µL of proteinase K (Ambion) was added and the samples incubated for 45mins at 55°C with shaking. DNA extraction was achieved using phenol chloroform then a sodium acetate precipitation. The samples were further sonicated in the Covaris S220 for 100sec (DF10, 200cycles, 175W) and prepared for Ion Torrent Proton Sequencing using the Ion Plus Fragment Library Kit (Life Technologies). To compensate for the low starting DNA concentration, ligation of adaptors was performed overnight at 16°C and the libraries amplified for 12 cycles before gel size selection.

The even and odd ChiRP-Seq datasets were analysed for peak detection using the MACS (Model-based Analysis of ChIP-Seq) version 1.4.2 peak-calling algorithm with a p-value cut-off of 0.00001

[262]. The *LacZ* dataset was used as a control to assign each detected peak an FDR (false discovery rate) score. The peaks with minimum overlap fractions of 70% between the even and odd datasets which also overlapped with the surrounding regions of the transcription start sites (TSS ± 500bp and TSS ± 1kb) were identified using BEDTools [263]. The corresponding genes were evaluated for pathway and network enrichment using QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

2.4.6. Nuclear/Cytoplasmic Fractionation

MCF7 cells were grown to 65% confluence in a 10cm plate, washed with PBS, scraped off in 10ml of PBS then centrifuged for 5mins at 1000g. The pellet was then re-suspended in 1ml of cold cell disruption buffer (10mM KCl, 1.5mM $MgCl_2$, 20mM TrisHCl pH7.5, 1mM DTT (Sigma)) plus 150µL dH_20 to enhance lysis and incubated on ice for 10mins. Once the cells had swollen to >3x normal size they were subject to 30 strokes of a Dounce homogenizer to give >90% cell lysis. Triton-X-100 was added to give a 0.1% solution, the tube inverted 5 times and centrifuged for 5mins at 1000RCF. The supernatant was then separated from the pellet and 500µL Trizol (Invitrogen) added to the nuclear sample for RNA extraction. 1/9 volume of 10x SDS solubilisation buffer was added to the cytoplasmic supernatant with 100ug Proteinase K per ml of solution then a 20min incubation was performed at 42°C. 1/10 volume 3M Na acetate plus 1 volume of phenol chloroform was then added, the tube shaken, and centrifuged for 5mins at 12000g before a chloroform cleanup and isopropranol precipitation. Both RNA samples were Turbo DNAse (Life Technologies) treated and cDNA made using random hexamers and SuperscriptIII (Life Technologies). Equal proportions of RNA were used from each cellular compartment to make the cDNA. QPCR was then performed in parallel using the cDNA template from both samples and the nuclear/cytoplasmic RNA ratio calculated by dividing the result for the nuclear sample by that for the cytoplasmic sample.

2.4.7. Sub-nuclear Fractionation

MCF7 cells were trypsinised, pelleted, PBS washed and counted then $2.5x10^6$ cells were resuspended in Lysis Buffer (*Appendix 8.1)* before 20mins incubation on ice. The solution was then centrifuged for 10mins at 2000g with the supernatant reserved as the cytoplasmic compartment. The pellet was resuspended in 50µL of the Nuclei Lysis Buffer (*Appendix 8.1*) and incubated for 5mins on ice. After a 5min centrifuge at 17000RCF at 4°C, the supernatant was reserved as the nucleoplasm compartment. The pellet was resuspended in 50µL Salt Extraction Buffer *(Appendix*

8.1) and incubated for 30mins with rotation at 4°C before a final centrifuge at 17,000g for 20mins at 4°C. The supernatant was reserved as the salt extracted fraction and the pellet resuspended in 50µL of Salt Extraction Buffer. Trizol (Ambion) was used to extract the RNA from all the fragments and cDNA made using random hexamers and SuperscriptIII (Life Technologies). Adapted from [264].

2.5. CUPID1 and CUPID2 over-expression

2.5.1. Determining the sequence of CUPID1 and CUPID2

The 5' end of the lncRNAs was determined using publically available RIKEN CAGE data from MCF7 cells. 3'RACE was used to find the 3' end. A series of sequentially nested forward primers were designed to be used with a reverse primer complementary to a sequence incorporated in a polyT bait primer used to synthesize cDNA from MCF7 RNA. After 2-3 rounds of nested PCR, the product was run on a 1% agarose gel, stained with Midori Green (Nippon Genetics) and the bands excised, purified using a Qiagen gel extraction kit, cloned into pBLUNT and multiple clones sequenced via Sanger sequencing. Exons were delineated by sequencing a PCR product from primers designed against regions just within the determined 5' and 3' ends. Kappa HiFi was then used to amplify the two main isoforms of *CUPID2* using primers starting at each end and containing *EcoRI* and *XhoI* recognition sequences. The products were run on a 1% agarose gel, extracted using the Qiagen kit and subcloned into pBLUNT. The two main isoforms of *CUPID1* were ordered as pre-synthesized sequences within a pUC57 vector obtained from Genscript (USA).

2.5.2. Stable over-expression of CUPID1 and CUPID2

A dual UbC-driven promoter vector pCDH (SystemBio) containing an EGFP (enhanced green fluorescent protein) and puromycin selection cassette was used for lncRNA over-expression. *HindIII* and *XhoI* enzymes were used to digest the lncRNA amplicons out of pBLUNT or pUC57 and ligate them into the pCDH vector. For experiments involving transient transfection, the plasmids were transfected into MCF7 cells using Lipofectamine 2000 (Life Technologies) as previously described, RNA extracted after 6 hours and the expression measured by qPCR. Stable cell lines over-expressing the lncRNAs were created using the ERα positive cell line T47D. Initially, a second generation lentiviral system utilising the pVSV envelope plasmid and pdeltaR8 packaging plasmid was used to generate infective viral particles in 90% confluent HEK293 cells (ATCC CRL-1573) by transfecting 3µg pVSV, 6µg pdeltaR8 and 3µg of the pCDH plasmids with 36µL of Lipofectamine 2000 (Life Technologies) in 6500ml of OptiMEM per T75 flask. The media

was removed after an overnight incubation at 37°C with 5%CO_2 and replaced with 8ml DMEM + 5% FCS and antibiotic/antimycotic (Gibco). The supernatant was removed after 48hours growth, centrifuged to pellet any cell debris and snap frozen at 80°C in 1ml aliquots. For viral transduction, T47D cells were grown till 50% confluent in a 6 well plate then the media was removed. An aliquot of viral supernatant was thawed, mixed with 1ml of fresh 5%FCS DMEM media and 7µL of a 2µg/µL Polybrene (Sigma-Aldrich) stock before filtering through a 0.45µM syringe filter onto a well of the T47D cells. The cells were incubated for 24 hours at 32°C to improve transduction efficiency before a media change and ongoing culture at 37°C. After 3 days 2µg of puromycin was added per well and the media changed daily until all the cells of a non-transduced T47D well had been killed. 400ng of puromycin per 2.5ml media was then used as maintenance. After 5 passages the cells were trypsinised and sorted by a FACS-ARIA (BD Biosciences) machine, discarding the lower 50% of GFP expressing cells. The remainder were cultured further then frozen down at -80°C after confirming successful lncRNA overexpression via RNA extraction and qPCR.

2.5.3. Xenograft mouse model

30 NOD/SCID mice were obtained from the Animal Resources Centre. One day prior to xenograft injection, a17ß-Estradiol 60 Day Release 0.72mg/pellet (Innovative Research) was inserted subcutaneously into the inter-scapular region of each mouse to support ERα positive tumour growth. Stable lncRNA expressing cells were trypsinised, counted and washed with PBS. Just prior to injection, $10x10^6$ cells were resuspended in 1ml of PBS then 100µL injected into the left lower mammary fat pad of the SCID mice (5 mice per group). Mammary injection and mouse care were performed by Doctor Shu Wen Wen and Doctor Christina Wong. Water and chow was provided *ad libitum* and 5 mice of the same xenograft group were housed per cage. Tumours were palpated twice weekly and the mice monitored for signs of metastasis.

Once the tumours reached a size of 1cm, the mice were sacrificed using an overdose of isoflourane. The mammary glands were then dissected off the abdominal wall and spread over a glass slide. They were fixed overnight in 40ml of Canoy's fixative (6 parts ethanol/3 parts chloroform/1 part acetic acid) then moved from 70% to 50% to 20% solutions of ethanol for 15mins per wash, finishing in plain water for 5mins. 40ml of Carmine stain (1g Carmine red (Sigma) + 2.5g aluminium potassium sulphate (Sigma) boiled in 500ml dH_2O) was used to stain the tissue overnight and the slides finally transferred to 70% ethanol for storage and imaging using a Leica MZ6 stereo microscope.

2.5.4. MTT Assay

5000 cells per well were seeded into a 96 well plate and grown for 72 hours in a 37°C incubator with 5%CO_2. The cells were then washed twice with PBS, the media replaced with 100µL phenol red free media (Gibco) and 10µL of MTT solution (5mg MTT powder (Life Technologies M6494) in 1ml of dH_2O) added per well. After 4 hours further incubation the media was replaced with 100µL of DMSO (Ajax Finechem) and mixed well to dislodge the purple granules. The plates were incubated for another 30mins at 37°C until all the granules had dissolved, mixed again and the absorbance at 540nm read using a Biorad Benchmark Plus microplate spectrophotometer. Results from a set of wells containing no cells were subtracted from the obtained values to control for variance in plate and MTT absorbance.

2.5.5. Colony formation assay

Cells were initially transfected with pCDH plasmids that overexpressed *CUPID1, CUPID2* or the vector alone. 48 hours after transfection the cells were trypsinised, counted using a Tali Image cytometer then seeded in equal numbers into a 6 well plate with 3 replicates per condition. After 24 hours 2ug of puromycin was added to each well, including a control well containing non-transfected cells. The media and puromycin were changed daily until all the cells in the non-transfected well were dead and then the puromycin concentration was reduced to a maintenance dose of 400ng per well. The cells were cultured for 2-3 weeks until large, distinct colonies had formed and the plates were then washed with cold PBSx2 and placed on ice. They were then fixed for 10mins with ice-cold 100% methanol and removed off the ice. 0.5% crystal violet in 25% methanol was added and the cells incubated for a further 10mins before aspiration of the solution. The cells were washed repeatedly with tap water until no further purple dye came off. Once dry, the plates were scanned with an E BoxVX2 (Vilber Lourmat) and the images processed using the OpenCFU program [265] to count the average number of colonies per plate.

2.6. Functional lncRNA Assays

2.6.1. Cell cycle assay

MC7 cells were transfected for 48hours with Lipofectamine2000 (Life Technologies) in a 6 well plate as previously described. They were then trypsinised, washed with PBS and resuspended in 1

ml of PBS. Cell clumps were removed by passing the solution through a 40uM filter and the filtrate added to 9ml of 70% ethanol for storage at -20°C. For processing, the cells were centrifuged for 5mins at 1200RPM and the pellet resuspended in 5ml of PBS. After 60seconds the tube was centrifuged again for 5mins at 1200RPM and the pellet resuspended in PI buffer (1:100 Triton-X-100, 1:5000 propidium iodide and 1:50 RNAseA (10mg/ml) all in PBS). The cells were processed using a FACSCantoII (BD Biosciences) machine within 30mins. Data analysis was via the Modfit LT 3.2 (Verity Software House, Topsham, ME) suite using default settings to determine G_0/G_1 and G_2/M phases.

2.6.2. Immunofluorescence assay for DNA damage

$6.5x10^4$ MCF7 cells were reverse transfected with Lipofectamine3000 (Invitrogen) and 50nM siRNA against *CUPID1, CUPID2* and *CCND1* on coverslips placed in 24well plates. After 48hours, half the plates were irradiated at 6Gy with the remaining plates grown without irradiation as negative controls. After 6 hours of further culture, the cells were fixed by firstly washing with cold PBS then incubating on ice for 5mins in Cytoskeleton Buffer (*Appendix 8.1*). This was then replaced with Cytoskeleton Stripping Buffer (*Appendix 8.1*) for a 5min incubation on ice and the cells washed with cold PBSx2. They were then fixed in 4% PFA at RT for 15mins, washed again with cold PBSx2 and permeabilised with 0.5% Triton-X-100 at RT for 15mins. After a PBS wash x2, the cells were blocked using FBT buffer (*Appendix 8.1*) for 30mins at RT. Primary antibody staining was performed for 1 hour at RT in 1:500 Genetex Rad51 (GTX70230) mouse 1^o antibody in FBT buffer, the cells washed with cold PBSx2, then a further 1 hour incubation with 1:1000 Abcam γH2AX (ab11174) rabbit 1^o antibody in FBT buffer. After cold PBS washing x2, the cells were stained for 30mins in 1:500 Alexa Flour 488 (A21202) donkey anti-mouse 2^o antibody in FBT buffer, the cells washed with cold PBSx2, then a further 30min incubation with 1:500 Alexa Flour 546 (A11010) goat anti-rabbit 2^o antibody in FBT buffer. After a final cold PBS wash x2, the coverslips were mounted using Vectashield mounting medium with DAPI (Vectorlabs). Once dry, the slips were sealed using nail polish and were viewed with a GE DeltaVision Deconvolution microscope. Slides were randomised before viewing and the intranuclear foci for Rad51 and γH2AX counted in 5 sets of 3 nuclei per field with 2 coverslips per condition. DAPI was used to stain the nuclei and specific antibodies used to mark either γH2AX or Rad51 foci before imaging using TRITC or FITC excitation respectively. The number of foci in the non-irradiated group was subtracted from the irradiated group to adjust for variation between the groups. The final adjusted

figure for the number of Rad51 foci was divided by the total number of adjusted γH2AX foci. Final image processing was done using the SoftWorx program.

2.6.3. Rad51 inhibition assay

$5x10^4$ MCF7 cells were seeded per well of a 24 well plate and reverse transfected with siRNA against *CUPID1, CUPID2* and *CCND1*. After 24 hours the Rad51 Inhibitor B02 (SML0364-Sigma) was added to the media in increasing concentrations (0.0, 0.1, 0.3, 1.0, 3.0, 10.0uM all in duplicate) and the cells grown for another 48 hours. The cells were then washed twice with PBS and the media replaced with 100µl of MTS Buffer (10% FBS, 10% cell titre aqueous one solution cell proliferation assay buffer (Promega) in PBS). 5µL of 20% SDS was used to stop the reaction after 15-30min. The absorbance was read at 500nm using a Biorad Benchmark Plus microplate spectrophotometer and used as a proxy for cell number. Results from a set of wells containing no cells were subtracted from the values to control for variance in plate and MTS absorbance. The final result was expressed as a percentage of the value obtained from the B02 free well.

2.7. TALEN genome editing (Transcription activator like effector nucleases)

2.7.1. TALEN process

The initial transfection of T47D cells was performed in a 6 well plate using 15µL of Lipofectamine 2000 (Life Technologies) per well together with 800ng of each TALEN plasmid, 2400ng of donor plasmid and 168ng of EGFP plasmid made up to 500µL with OptiMEM (Life Technologies). 2500µL of antibiotic free media containing 15x10^5 T47D cells was then added and the wells incubated for 48 hours. The media was then changed and the cells allowed to recover for 72 hours before being FACS sorted into single cell cultures in 96 well plates, using high GFP expression as a positive selection marker. The clones were cultured until they reached 90% confluence then split into 3 wells with 1 well used for DNA extraction and the remainder frozen at -80^{o}C until the identity of the clones containing the donor sequence had been confirmed. DNA extraction was performed using a Blood Genomic DNA Extraction kit (Sigma) and the DNA then tested by Dr Qing Chen using a Sequenom Mass-Array IPlex system for the presence of the minor alleles of SNP1, SNP2 and SNP3. Clones found to be positive were recovered from -80^{o}C and grown up in 6 well plates before a final DNA extraction using the DNeasy Blood and Tissue Kit (Qiagen) and RNA extraction using Trizol (Life Technologies). The DNA was used as a template in a TaqMan SNP assay with probes specific for the minor allele or major allele of SNP3. 50ng of DNA was also

used as a template to amplify a 1615bp product over SNP1. 500ng was then digested with *DpnII* and run on a 2% agarose gel with the digest pattern compared to digested PCR products amplified off DNA from a known SNP1 heterozygous cell line (Cal51) and a cell line homozygous for the major allele (T47D). The RNA was Turbo DNAse (Ambion) treated and cDNA made using random primers before measuring gene expression using specific TaqMan gene expression probes.

2.7.2. T7 Endonuclease-1 Assay

To confirm TALEN activity, DNA was extracted from cells following FACS enrichment using the DNeasy Blood and Tissue Kit (Qiagen) and 100ng used as a template for a PCR reaction using KappaHF (Kapa Biosystems) producing a 740bp amplicon across the predicted TALEN cut site. The products were run on an agarose gel and the 740bp band excised and purified using a Qiagen gel extraction kit. 200ng was then prepared in 1x NEB2 buffer to a total volume of 19µL and annealed in a thermocycler (95^{o}C for 10 mins; 95 ^{o}C to 85 ^{o}C dropping 2 ^{o}C/sec; 85 ^{o}C to 25 ^{o}C dropping 0.3/sec; 25 ^{o}C hold for 10sec). 1µL of T7 Endonuclease-1 (New England Biolabs) was then added and the reaction incubated for 15mins at 37^{o}C before terminating the reaction with 2µL of 0.25M EDTA. The products were then run on a 1% agarose gel and compared with a control sample untreated with T7 Endonuclease-1.

2.8. Computational analysis

2.8.1. Statistical analysis

Data analysis was performed using the GraphPad Prism 6 (GraphPad Software, La Jolla, CA, USA) program with the test specified in the appropriate figure legends. Results were considered statistically significant when the p value was <0.05. Significance in figures is denoted as follows: * p<0.05, **p<0.001, ***p<0.0001, ****p<0.00001, ns = not significant p>0.05.

2.8.2. Images and Figures

All graphs were produced using GraphPad Prism 6 (GraphPad Software, La Jolla, CA, USA). Genome track images were generated from the UCSC browser [266], and annotated in Microsoft Powerpoint (Microsoft, Redmond, WA, U.S.A.). Gel images were processed using the EBox program then further annotated in Microsoft Powerpoint.

CHAPTER 3

Functional Characterisation of the 11q13 Breast Cancer Susceptibility Locus

The 11q13 breast cancer associated locus was fine mapped and partially characterised in a previous study by our group [5]. Four candidate causal SNPs mapped to an enhancer element (called PRE1) and one independent SNP mapped to a silencer element (called PRE2) Both PREs were shown to physically interact with and regulate *CCND1,* their likely target gene (*Figure 1.8*). However, there are a number of other genes in the locus which also may be interacting with PRE1 or PRE2 and thus contribute to breast cancer risk. These include *ORAOV1*, a poorly characterised gene that is postulated to act via regulating *CCND1* levels [267]; *MYEOV*, a putative transforming gene that is over expressed in myeloma [268]; and three fibroblast growth factor (*FGF)* genes *FGF3, FGF4* and *FGF19*. Members of the *FGF* gene family are involved in cell proliferation and have been implicated in a number of cancers [269]. The most promising gene in this list is *ORAOV1,* the only other gene in the locus that has open chromatin at its promoter and is significantly transcribed in MCF7 breast cancer cells (*Figure 3.1*). It is also a potential oncogene, being over-expressed in oral squamous carcinoma and cervical carcinoma [270].

The 11q13 interval also contains risk-associated SNPs for other types of cancers including prostate and renal carcinoma which are located centromeric to the breast cancer risk SNPs (*Figure 3.1*) [271, 272]. This raises the question of whether PRE1 and PRE2 are tissue specific and whether prostate or renal risk at the locus may similarly be mediated through *CCND1*[5]. In addition to coding genes there may also be non-coding RNAs at the locus mediating risk. A subset of ncRNAs are expressed from enhancer elements and are called enhancer RNAs (eRNAs). Further examination of the 11q13 locus revealed prominent transcription consistent with eRNAs, particularly from PRE1 (*Figure 3.1*) [87]. The knockdown of such eRNAs has been shown to affect chromatin looping and cause a reduction in the expression of nearby genes regulated by the enhancer producing the eRNA [232, 233].

Identifying regulatory elements and their target genes is a complex task. Many studies have used a candidate approach, relying on the hypothesis that the nearest gene is the most likely candidate. For these studies, standard chromosome conformation capture (3C) assays are used to confirm physical interactions between the element and the known gene [5, 60]. A broader approach is to use additional 3C-based techniques such as 4Cseq (Circular Chromosome Conformation Capture) and 5C (Chromosome Conformation Capture Carbon Copy) to find novel interacting targets (*Figure 1.5*). 4Cseq relies on using one bait locus to interrogate the entire genome for interactions present at the time of fixation [171]. Notably, this approach allows detection of inter-chromosomal contacts

which would be otherwise difficult to predict [273]. 4Cseq has already been successfully used in a post-GWAS study to identify genes interacting with a cardiac enhancer that contains a risk SNP linked to Brugada syndrome [274]. 5C examines interactions within a defined genomic region and can identify enhancer-promoter looping or give a broader picture of local chromatin structure [174]. 5C has not yet been used to characterise risk loci although a related technique (capture Hi-C) has recently been employed to find targets of putative enhancers at the 2q35, 8q24 and 9q31 breast cancer risk loci and at 14 colorectal carcinoma loci [162, 173].

The aforementioned techniques were all employed as part of a comprehensive functional characterisation of the 11q13 breast cancer locus to identify further genes in the region affected by the SNPs that may be relevant to breast cancer biology. Investigation of *ORAOV1* as a potential gene mediating the effect of the SNPs was inconclusive, however the use of additional 3C based techniques revealed another four candidate genes including *MTL5*, *CPT1A*, *IGHMBP2* and *MRPL21*. An eRNA knockdown approach to confirm interactions between PRE1 and the identified genes then prioritised *CPT1A* and *IGHMBP2* as likely additional targets of PRE1. Finally, genome editing with transcription activator-like effector nucleases (TALENS) was used in an attempt to confirm the effect of the SNPs. Further functional work is required, however the range of techniques employed has substantially expanded our understanding of the 11q13 breast cancer risk locus and forms a template for future investigations of GWAS risk loci.

Figure 3.1 ***The 11q13 multiple cancer susceptibility locus*** - PRE1 and PRE2 are part of the breast cancer risk locus and interact (red lines) with the *CCND1* promoter. A renal cancer risk locus lies just centromeric with a prostate risk locus on the other side of the *MYEOV* (myeloma overexpressed) gene. The other genes in the locus include *ORAOV1* (oral cancer overexpressed 1 gene), and three fibroblast growth factor genes *FGF19, FGF4* and *FGF3*. Oestrogen receptor α (ERα) is shown by the blue rectangles whilst the orange track depicts areas of localised DNAse I hypersensitivity [275], usually seen at enhancers or promoters. The blue dashes of the lowermost track indicate GROseq reads from nascent transcription [87]. Figure generated using the IGV browser [276].

3.2.1. PRE2 interacts with CCND1 in renal and prostate cell lines.

To determine whether the previously demonstrated interactions between PRE1 and PRE2 and the *CCND1* promoter at the 11q13 locus were specific to breast tissue [5], 3C was performed in the 786-0 renal and PC3 prostate carcinoma cell lines. The results showed that PRE2 interacted strongly with the *CCND1* promoter in both cell types (*Figure 3.1B,D*). In contrast, PRE1 interacted with the *CCND1* terminator region but not with the promoter (*Figure 3.2A,C*). Luciferase reporter assays using pGL3 constructs containing the *CCND1* promoter were then conducted to assess the regulatory activity of PRE1 and PRE2 in a prostate cell line (*Figure 3.2E*). Similar assays were also attempted in renal 786-0 cells but these were unsuccessful as this cell line proved difficult to transfect. Luciferase assays showed that PRE1 acted as a silencer of the *CCND1* promoter in prostate cancer cells and that inclusion of the minor alleles of SNPs 1-4 reduced the magnitude of silencing. A similar transfection performed in MCF7 breast cancer cells by French *et al.* showed that PRE1 acted as a strong enhancer on *CCND1* promoter activity, emphasizing a marked difference in response between breast and prostate tissue [5]. PRE2 also acted as a silencer, however inclusion of the minor allele of SNP5 did not alter its activity (*Figure 3.2E*). Whilst PRE2 acted as a silencer in MCF7 cells, the absolute values of these results fall at the lowest limit of what can be reliably discriminated by the assay. Further 3C assays were also performed in prostate cells using a bait fragment containing the top prostate risk SNP to confirm the validity of the PC3 3C library and to determine whether the identified prostate cancer risk at the locus is mediated through *CCND1* in a similar manner to breast cancer risk (*Figure 3.2F*) [272]. The 3C in prostate cancer cells confirmed the predicted interaction between SNP containing enhancers in the prostate risk locus and the *CCND1* promoter, suggesting that prostate cancer SNPs may mediate risk through dysregulation of *CCND1* in a similar manner to the breast cancer risk SNPs (*Figure 3.2F*).

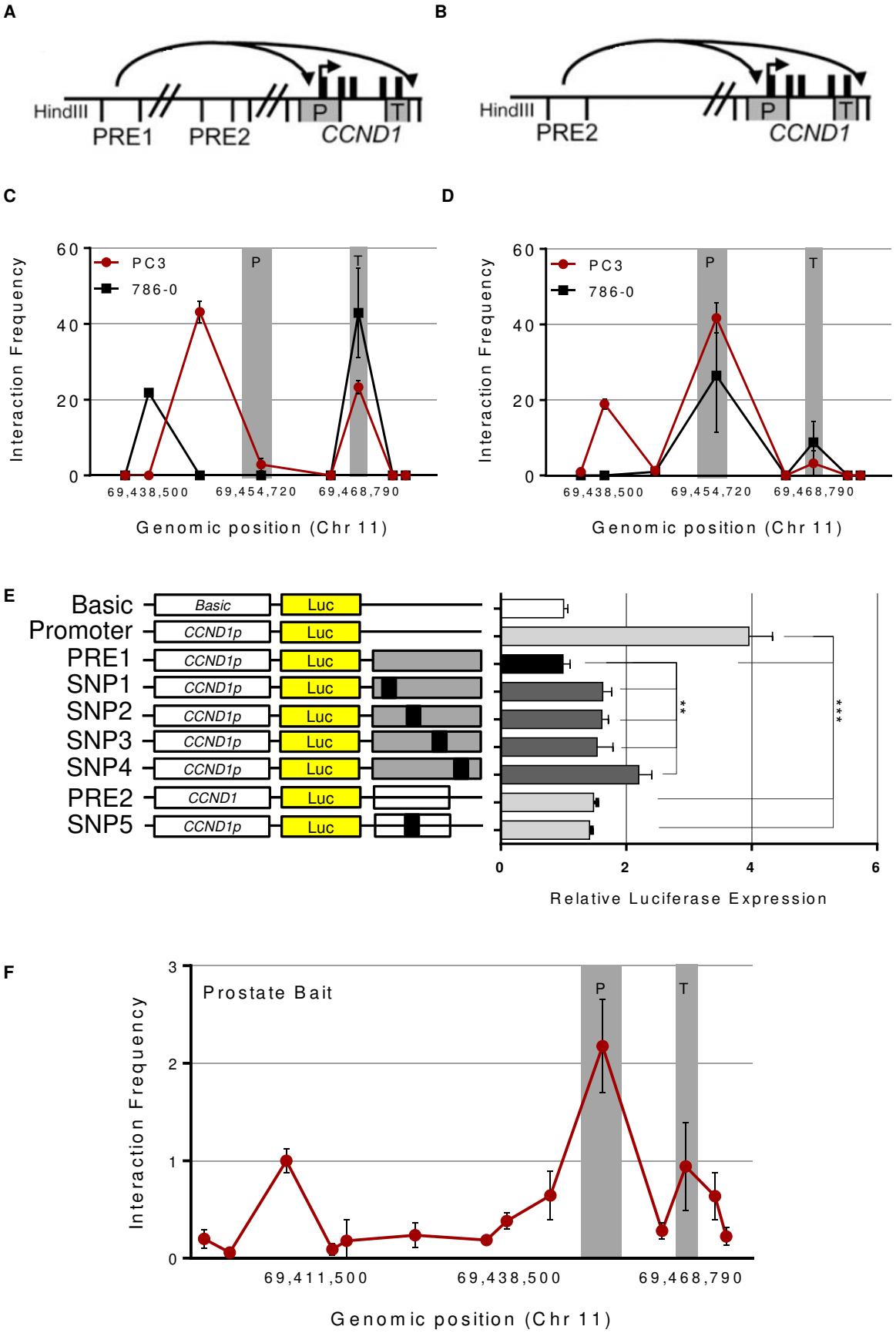

A
HindIII
PRE1
PRE2
P
T
CCND1
B
HindIII
PRE2
P
T
CCND1
C
Interaction Frequency
60
40
20
0
PC3
786-0
P
T
69,438,500
69,454,720
69,468,790
Genomic position (Chr 11)
D
Interaction Frequency
60
40
20
0
PC3
786-0
P
T
69,438,500
69,454,720
69,468,790
Genomic position (Chr 11)
E
Basic
Promoter
PRE1
SNP1
SNP2
SNP3
SNP4
PRE2
SNP5
Basic
CCND1p
CCND1
Luc
**
***
0
2
4
6
Relative Luciferase Expression
F
Prostate Bait
Interaction Frequency
3
2
1
0
P
T
69,411,500
69,438,500
69,468,790
Genomic position (Chr 11)

Figure 3.2 3C and luciferase assays in renal and prostate cell lines. (A, B) Schematics of the 11q13 locus interrogated by 3C. (C, D) 3C analysis of interactions between *HindIII* fragments containing PRE1 (C) or PRE2 (D) and the *CCND1* gene in renal (786-0) and prostate (PC3) cell lines. Grey bars depict the position of the *CCND1* promoter (P) and terminator (T), matching them with the schematics above each panel. (E) PRE1 and PRE2 were cloned downstream of a *CCND1* promoter-driven luciferase reporter +/- SNPs1-5. Luciferase activity is shown normalised to *Renilla* and expressed relative to the Promoter construct. (F) 3C interactions between the main prostate risk locus and *CCND1* in PC3 cells. Data shown are the mean ± SEM from three biological replicates. For the luciferase assays, significance was determined using a one way ANOVA test with Dunnett's correction for multiple comparisons. ** $p<0.01$,*** $p<0.001$.

3.2.2. ORAOV1 is induced by oestrogen and is widely expressed in breast cancer cell lines.

To prioritise additional 11q13 genes for further study, the expression of five local genes was measured in a panel of breast cancer cell lines by TaqMan assays (*Figure 3.3*). *MYEOV* and *ORAOV1* were expressed at similar levels to the house keeping gene *GUS* in the majority of ERα negative breast cell lines examined, however *MYEOV* expression was extremely low or undetectable in the ERα positive breast cell lines (*Figure 3.3A*). This is significant as the breast cancer risk-SNPs at 11q13 are associated with ERα positive breast cancer [5]. *FGF3* and *FGF4* were not detected in any cell line examined, whilst *FGF19* was expressed at low levels in only one of the four ERα positive breast cell lines (*Figure 3.3B*). *CCND1* was expressed at very high levels in all the cells examined, with most lines expressing over 100x more *CCND1* than the housekeeping gene *GUS* (*Figure 3.3C*).

The expression of *ORAOV1* was further examined following oestrogen stimulation as 11q13 is a risk locus for ERα positive breast cancer, and PRE1 displays extensive ERα binding (*Figure 3.1*). A robust increase in levels of the highly oestrogen responsive gene *TFF1* was seen following oestrogen exposure confirming a successful induction (*Figure 3.4A*). *CCND1* expression increased after 6-12 hours post oestrogen (*Figure 3.4B*) and *ORAOV1* was induced significantly at 12 hours (*Figure 3.4C*). The breast cancer cell expression data and oestrogen expression data support the further investigation of *ORAOV1* as a possible additional target gene mediating the effects of the breast cancer risk SNPs at 11q13. In addition, ChIP-seq and DNase I hypersensitivity data showed that the *ORAOV1* promoter was the only gene (apart from *CCND1*) to display open chromatin in MCF7 breast cancer cells (*Figure 3.1*).

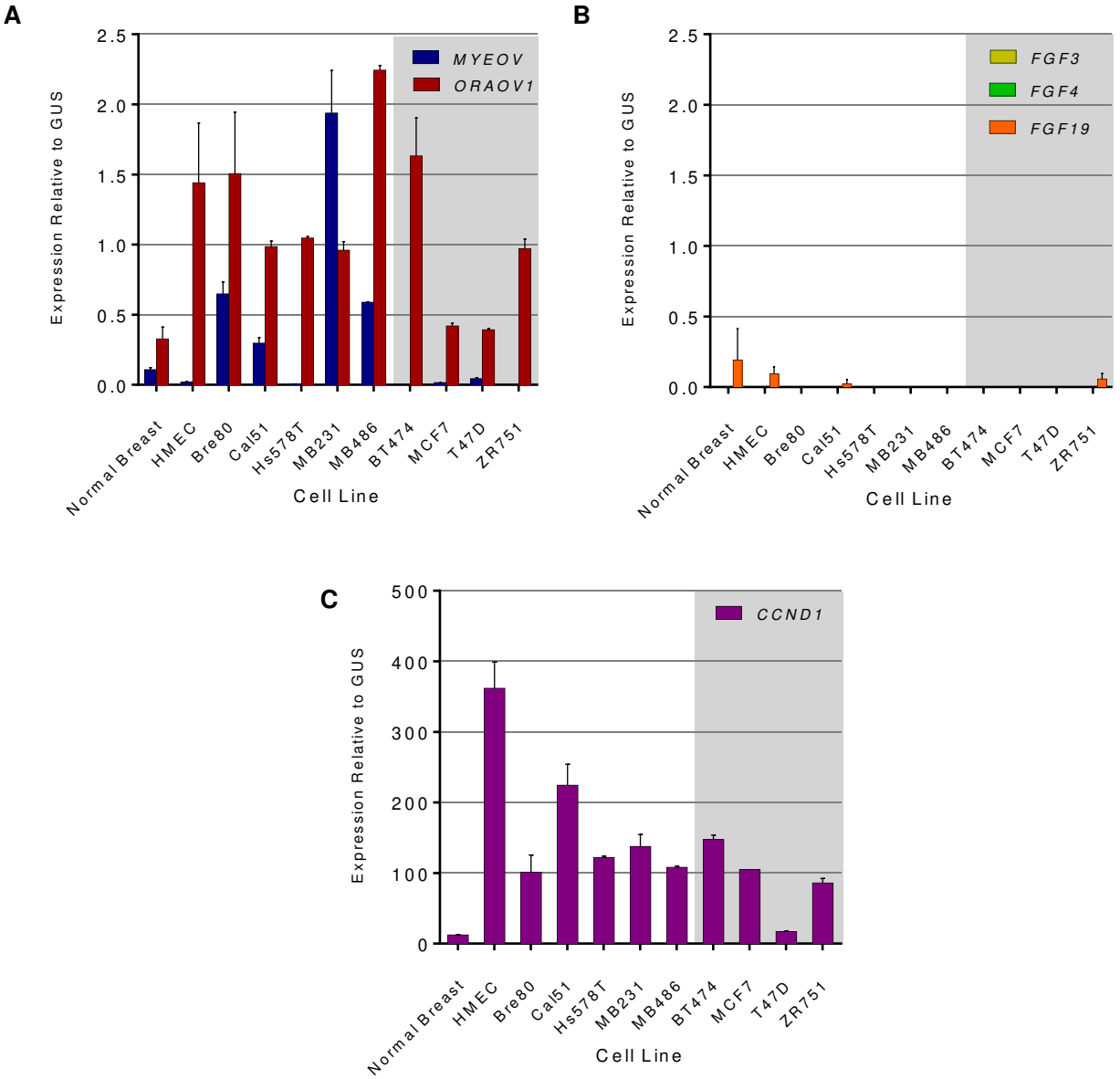


Figure 3.3 Local gene expression at the 11q13 locus in a panel of breast cancer cell lines. Expression is presented normalised to the house keeping gene *β–glucuronidase* (*GUS)*. The grey shading highlights ERα positive cell lines. (A) *MYEOV* and *ORAOV1* expression, (B) *FGF3*, *FGF4* and *FGF*19 expression and (C) *CCND1* expression, which is presented with a scale 200x that of the other two graphs reflecting its far higher expression levels. Data shown is the mean +/- SD of a single biological replicate.

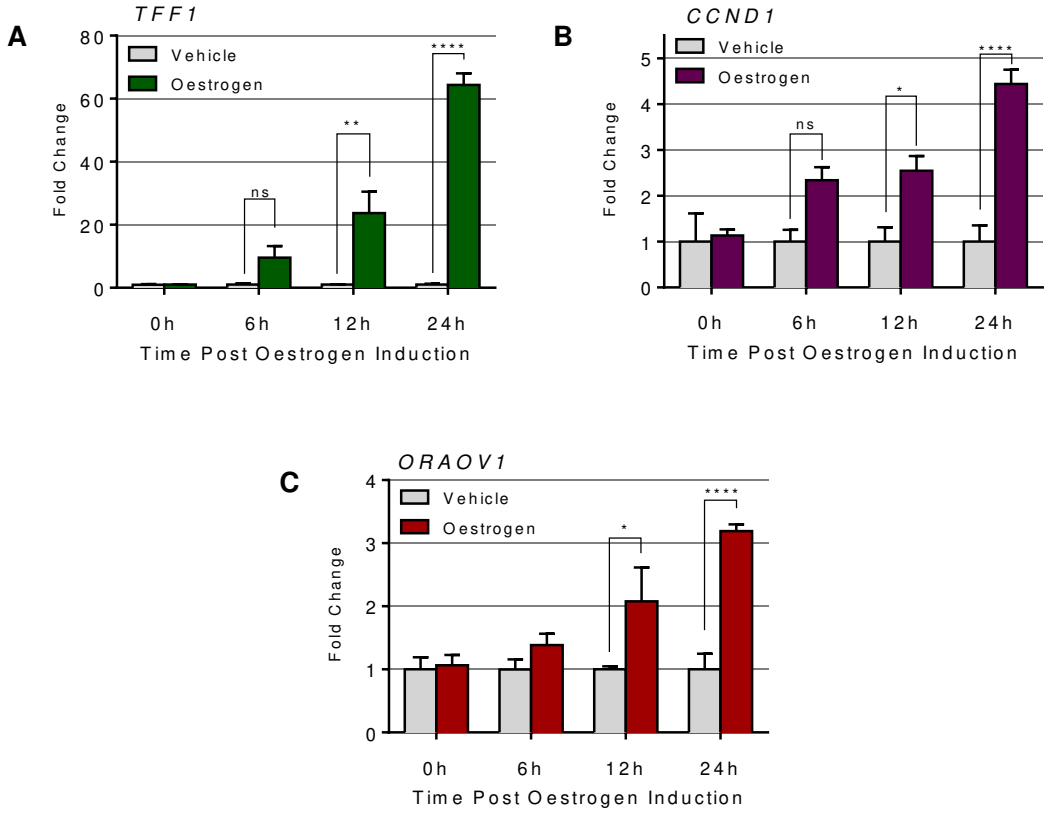


Figure 3.4 TFF1, CCND1 and ORAOV1 are induced by oestrogen. Gene expression at each time point is expressed as a fold change over vehicle, with vehicle arbitrarily set to one. (A) The highly oestrogen responsive gene *TFF1* acts as a positive control, (B) *CCND1* and (C) ORAOV1 expression over a 24 hour oestrogen induction. All mRNA levels were normalised to *β-glucuronidase* (*GUS*). Data shown is the mean ± SEM from three biological replicates. Significance was determined using a 2 way ANOVA test with Dunnett's correction for multiple comparisons. * $p<0.05$, ** $p<0.001$, ***$p<0.0001$, ****$p<0.00001$.

3.2.3. PRE1 interacts with the ORAOV1 gene in a non-oestrogen dependent manner.

3C was performed in the ERα positive cell line BT474, to examine interactions between PRE1 or PRE2 and the *ORAOV1* promoter. Given that *ORAOV1* is Ez responsive (*Figure 3.4C*), the 3C was repeated in Ez induced MCF7 cells to investigate whether the chromatin looping interactions were regulated by ERα. The results showed that PRE1 interacts strongly with the *ORAOV1* promoter region in MCF7 and BT474 cells, indicating that it is highly likely to be involved in the regulation of *ORAOV1* expression (*Figure 3.5*). In contrast, the absence of an interaction peak over PRE2 indicates that it does not interact in the cell lines tested and is thus unlikely to be involved in regulating *ORAOV1*. The interactions were not significantly different between libraries induced

with oestrogen or vehicle alone, demonstrating that the chromatin looping is pre-existing and not oestrogen dependent.

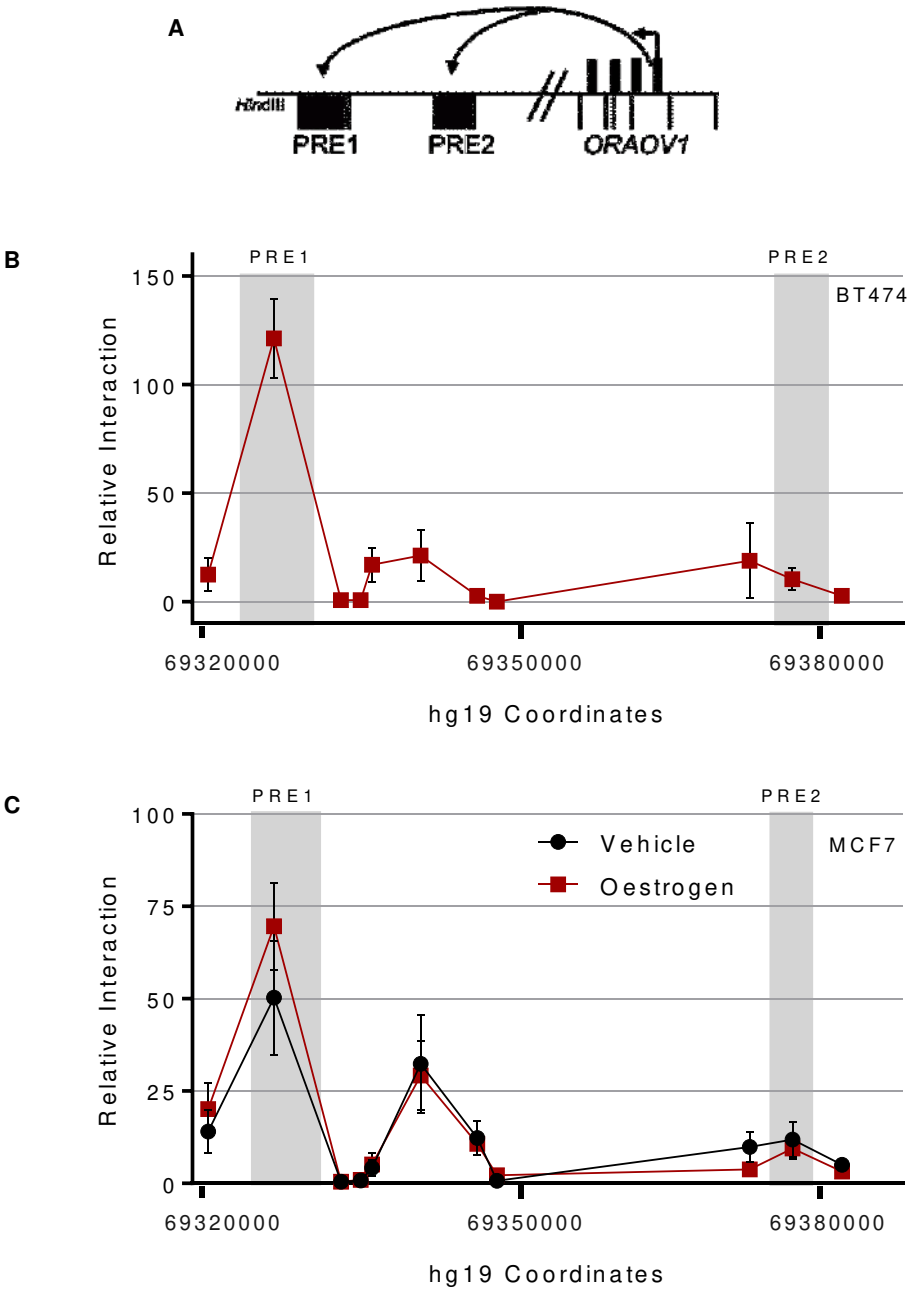


Figure 3.5 Chromatin interactions are present between the ORAOV1 gene and PRE1. (A) A schematic diagram of the 11q13 genomic region interrogated by 3C. Black vertical lines at the bottom indicate *HindIII* sites, black arrows indicate direction of transcription and black blocks indicate exons. (B) 3C interaction profiles between the *ORAOV1* promoter and the locus containing PRE1 and PRE2. 3C libraries were generated with *HindIII* in MCF7 cells +/- oestrogen (B) or BT474 cells (C). Data shown is the mean ± SEM derived from three biological replicates.

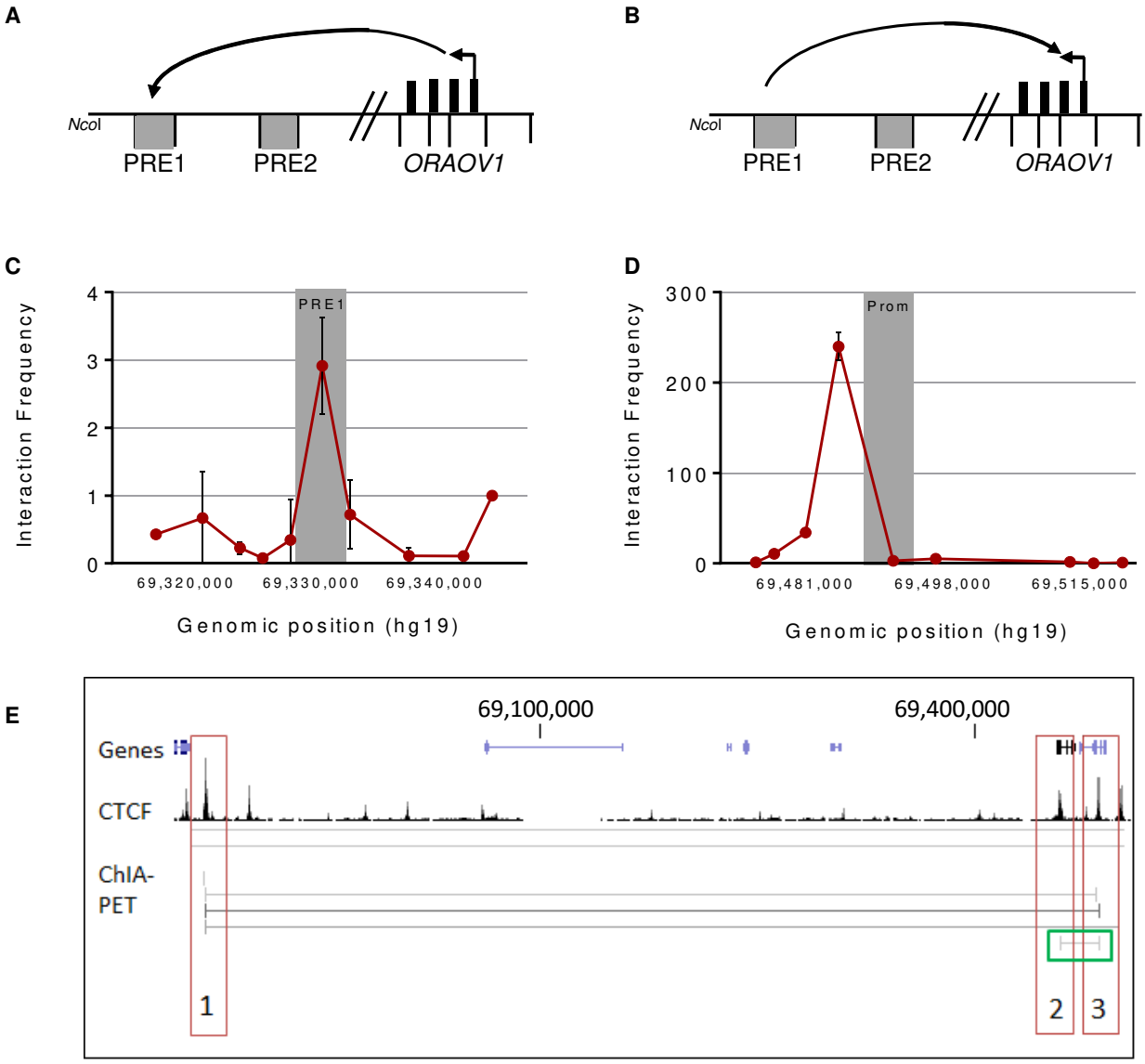


Figure 3.6 The interaction between PRE1 and ORAOV1 is maximal within the gene body. (A)(B) Schematic diagrams of the 11q13 genomic region and *Nco*I fragments used for 3C analysis. Black vertical lines at the bottom indicate *Nco*I sites, the right angled arrow indicates the promoter and direction of transcription, black blocks indicate exons. (A+C) 3C interaction profiles between the *ORAOV1* promoter and *Nco*I fragments around PRE1 in MCF7 cells. (B+D) 3C interaction profiles between PRE1 and *Nco*I fragments around the *ORAOV1* gene in MCF7 cells. The grey shaded box indicates the PRE1-*ORAOV1* promoter interaction which has the equivalent value in each graph for ease of comparison. Data shown are the mean ± SEM derived from three biological replicates. (E) A screenshot from the UCSC browser of RNAPolII mediated ChIA-PET data in MCF7 cells [277]. The horizontal lines represent interactions between genomic loci and the black peaks indicate the frequency of interaction. The 11q13 TAD extends from box1 across to box 3 which corresponds to the 3C interaction peak in (D). Interactions are also present between box2 (*CCND1* promoter) and box3 (*ORAOV1* intron).

Further 3C was performed using libraries prepared with the *Nco1* enzyme in MCF7 cells, using either the *ORAOV1* promoter or PRE1 as bait to allow improved localisation of the interaction (*Figure 3.6*). This demonstrated that the maximal chromatin interaction is actually present between PRE1 and a region within the 2nd intron of *ORAOV1*, rather than the promoter region itself (*Figure 3.6D*). ChIA-PET data of CTCF mediated interactions in MCF7 cells was used to examine this region more thoroughly and revealed that there is a CTCF binding site within the *ORAOV1* gene that forms the boundary of a topologically associated domain (TAD) encompassing the 11q13 gene desert (*Figure 3.6E*) [277]. The boundary localises to the same restriction enzyme fragment found to contain the maximal interaction with PRE1. Given the absence of a direct promoter-enhancer interaction, *ORAOV1* was not studied further and other potentials targets in the region were sought.

3.2.4. Genome-wide 4C-seq identifies additional target genes of PRE1.

3C is a very useful technique for identifying *cis*-interactions, however it does have one significant limitation. That is, 3C can only detect interactions between pre-specified regions as it relies on user defined PCR primers. To overcome this issue, a more high-throughput 3C-based method called circular chromosome conformation capture (4C-seq) was performed to detect interactions genome wide in an unbiased manner (*Figure 1.5*) [278]. 4C-seq uses 3C libraries created using a 6bp restriction digest enzyme (*EcoRI*) and digests them further using a more frequent cutter (*CviQI).* The products can then be circularised using DNA ligase and interrogated using sets of nested primers (*Figure 3.7A*). Before sequencing, the PCR amplicons were run on an agarose gel (*Figure 3.7B*). The E/EN and C/CN primer pair produced a range of products including the expected bands at 3800bp (representing self-ligation of the *EcoRI* fragment during library production) and 1280bp (representing incomplete digestion of the *EcoRI* fragment during library production). The smears indicate a range of products, each corresponding to a restriction digest fragment that was interacting with the bait fragment at the time of fixation. A contaminating band at 300bp was noted in the control library however it was not seen in the subsequent 4C library and was of a size which would be removed at the size selection step prior to sequencing.

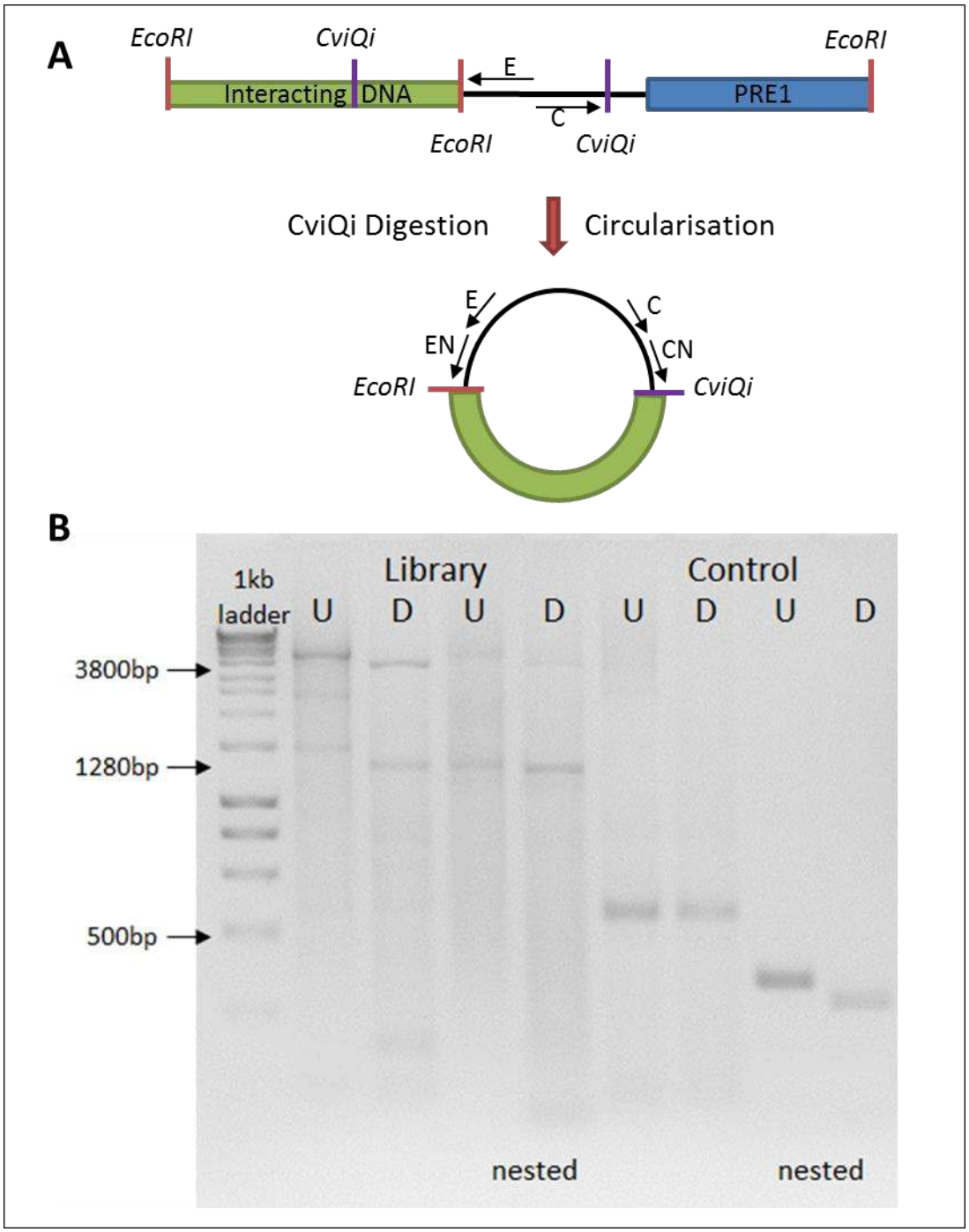


Figure 3.7 An extended smear of 4C products is produced by nested PCR. (A) A schematic showing the position of the *EcoRI* and *CviQi* sites used to circularise the products of the initial 3C (top). Following *CviQi* digestion and ligation the interacting DNA may be captured by the E and C primers which face their respective restriction enzyme sites. Each primer has a nested primer (EN, CN) to increase specificity of the PCR. (B) Primers E and C were used to amplify a range of products from the 4C library which was run on an agarose gel. PCR products from a 2nd nested PCR using the EC and NC primers were run on the gel in parallel. The primers were digested (D) off using the appropriate restriction enzyme to remove concatamers, or left undigested (U). A control library lacking digestion and ligation steps was included to identify non-specific amplification.

Two MCF7 4C libraries and one BT474 4C library were interrogated using the E and C PRE1 primers and the subsequent PCR products sequenced using an Ion Torrent 318v2 chip (Life Technologies) producing 5-6 million useable reads per library. The sequencing data was then processed and normalised using the published R-3Cseq program [258]. The Galaxy platform was used to find consistent regions of overlap between the three libraries and rank the most significant interaction regions [279]. A clustering analysis identified 27 regions common to the three libraries (*Table 3.1*). Four interaction regions were identified in the vicinity of the 11q13 locus overlapping a number of genes including *CCND1* and *ORAOV1* (*Figure 3.8A*). The top ranked hits *in trans*, contained *HEATR6* (amplified in breast cancer protein 1) and a number of pseudogenes including *TBC1D3P1-DHX40P1*at chromosome 17q23 (*Figure 3.8B*). A cluster of hits was also seen around the *NCOA3* gene (nuclear receptor coactivator amplified in breast cancer). At chromosome 20q13 were the *CYP24A1* (1,25-dihydroxyvitamin D3 24-hydroxylase) and *BCAS1* (breast cancer amplified sequence 1) genes (*Figure 3.8C*). All of the identified genes are associated with breast cancer, however the 17q23 and 20q13 genomic loci are heavily amplified in many breast cancer cell lines, including the MCF7 and BT474 cells that were used for this experiment [280, 281], making it unclear whether the detected interactions were merely a consequence of local amplifications. 4Cseq was also performed in Bre80 cells which are a non-cancer derived breast cell line and thus less prone to the massive amplifications seen in cancer cells. This revealed no significant, reproducible interactions beyond the local 11q13 region.

The two MCF7 4C libraries and one BT474 4C library were also interrogated using inverse primers adjacent to PRE2 and the PCR products sequenced and analysed as for PRE1. This only revealed 17 interaction regions, 12 of which overlapped centromeres (*Table 3.2*). Of the remaining regions, none localised to any characterised genes at 11q13 apart from regions 3 and 4 which contained *CCND1* and a number of uncharacterised transcripts (*Figure 3.7A*). Further investigation of the locus therefore focused on PRE1 which also possessed a number of features not seen for PRE2 making it a better candidate for functional interactions relevant to ERα positive breast cancer. In particular, it is a hub for multiple ERα mediated ChIA-PET interactions (*Figure 4.9*) and exhibits extensive binding of the breast associated transcription factors FoxA1 and ERα (*Figure 3.1*).

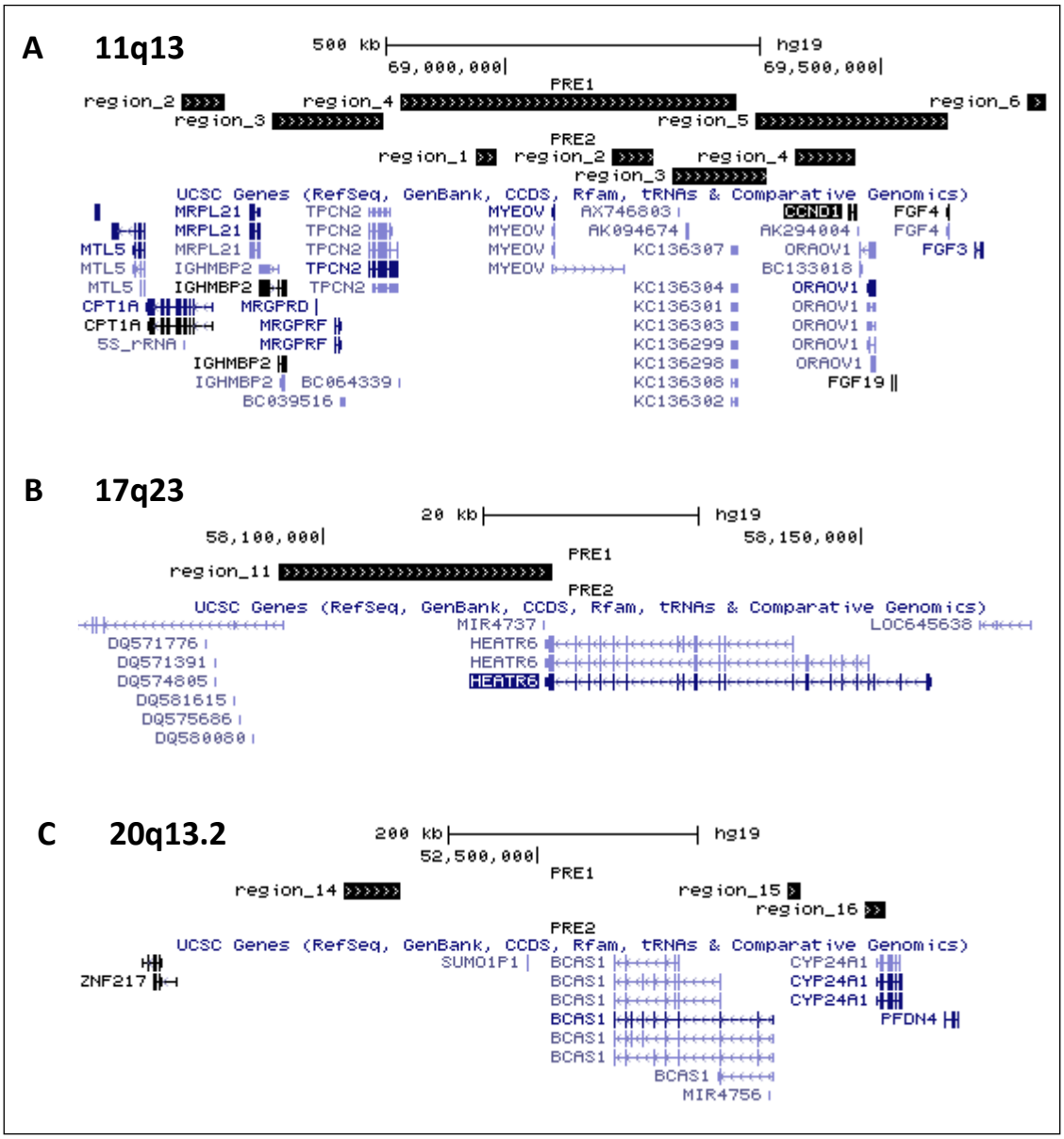


Figure 3.8 Regions of interaction for PRE1 and PRE2 in the 11q13, 17q23 and 20q13 loci. Each locus (A,B,C) contains clustered interaction regions for PRE1 and PRE2 above the isoforms of local genes. (A) The 11q13 locus with regions 2-6 for PRE1 and regions 1-4 for PRE2. (B) The 17q23 locus with region 11 for PRE1. (C) The 20q13.2 locus with regions 14-16 for PRE1. Data visualised using the UCSC browser.

Interaction Region	Chromosome	Genes within 50kb
region_0	chrY	centromere
region_1	chrX	centromere
region_2	chr11	*CPT1A*
region_3	chr11	*MRPL21, IGHMBP2, MRGPRF, TPCN2*
region_4	chr11	*MYEOV*
region_5	chr11	*CCND1, ORAOV1, FGF19, FGF4*
region_6	chr11	*FGF3*
region_7	chr10	centromere
region_8	chr10	centromere
region_9	chr10	centromere
region_10	chr17	centromere
region_11	chr17	*HEATR6*
region_12	chr19	centromere
region_13	chr20	*ZMYND8, NCOA3*
region_14	chr20	*ZNF217, BCAS1*
region_15	chr20	no genes
region_16	chr20	*CYP24A1*
region_17	chr20	*DOK5*
region_18	chr7	centromere
region_19	chr7	centromere
region_20	chr6	centromere
region_21	chr4	centromere
region_22	chr4	*CNEPC1, STAP1*
region_23	chr2	centromere
region_24	chr1	centromere
region_25	chr1	*PP1A4F, LOC645166*
region_26	chr8	*REXO1L2P*

Table 3.1 Genome-wide interaction regions detected by 4C-seq for PRE1. Genes with promoters within 50kb of the interacting region are listed in the right hand column. Regions containing genes potentially associated with breast cancer are highlighted grey.

Interaction Region	Chromosome	Genes within 50kb
region_0	chrY	centromere
region_1	chr11	no genes
region_2	chr11	*MYEOV*
region_3	chr11	no genes
region_4	chr11	*CCND1*
region_5	chr10	centromere
region_6	chr10	centromere
region_7	chr10	centromere
region_8	chr17	centromere
region_9	chr16	no genes
region_10	chr19	centromere
region_11	chr7	centromere
region_12	chr6	centromere
region_13	chr4	centromere
region_14	chr4	no genes
region_15	chr2	centromere
region_16	chr 1	centromere

Table 3.2 Genome-wide interaction regions detected by 4C-seq for PRE2. Genes with promoters within 50kb of the interacting region are listed in the right hand column. Regions containing genes potentially associated with breast cancer are highlighted grey.

3.2.5. 5C reveals five main loci interacting with PRE1

The 4C technique provided information genome-wide for interactions with the chosen bait. *Figure 3.8* however illustrates the wide regions of interactions identified, preventing identification of individual promoters. 5C (*Figure 1.5*) was therefore employed to further interrogate the proximal 11q13 region and identify genes that may be regulated by PRE1. I acknowledge that Dr Haran Sivakumaran prepared, sequenced and analysed the 5C experiments shown in *Figure 3.9*. The most consistent interaction is shown in box (*B*), with four black bars lined up over a bidirectional promoter of *MRP21* and *IGHMBP2*. This encompassed region 3 from the 4C-seq (*Figure 3.8A*). Box (*A*) is another interaction hotspot containing the *CPT1A* gene and overlapped region 2 from the 4C-seq. *MTL5* is a potential interacting partner of PRE1 on the edge of region 2 in the 4Cseq but has only two red bars over the promoter in the 5C, indicating low-moderate interactions. The final box (*C*) shows that *CCND1* is marked by four red bars indicating a moderate interaction, whilst *ORAOV1* only has three yellow and one red indicating low-moderate interaction frequency. Both of these genes are covered by the 4C-seq region 5 (*Figure 3.8A*).

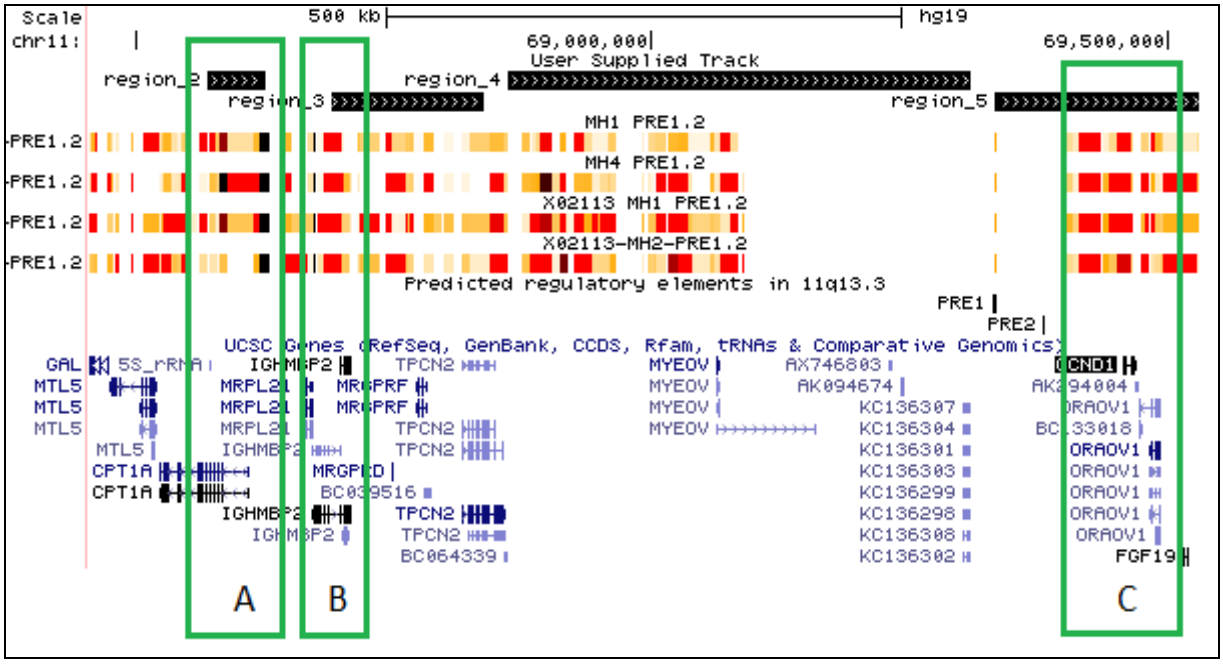


Figure 3.9 5C identifies interactions between PRE1 and multiple genes at 11q13. A snapshot from the UCSC browser showing four biological replicates of a 5C library with PRE1 as the bait. Interaction data is presented as a heatmap with yellow = few interactions; red = moderate interactions and black = high interactions. The top track contains the interaction regions detected by 4C-seq. The location of PRE1 and PRE2 are provided for reference in the middle track. USCS genes are displayed in the lower track. The green boxes highlight genes of interest: **(A)** *CPT1A*, **(B)** *MRP21* and *IGHMBP2*, and **(C)** *CCND1* and *ORAOV1*.

3.2.6. An oestrogen responsive eRNA is transcribed from PRE1.

To prioritise the genes which lay in the interaction regions identified by 4Cseq and 5C (*Figures 3.8, 3.9*), an enhancer RNA (eRNA) knockdown strategy was employed. Knockdown of the RNA transcribed from an enhancer has been shown to reduce the transcription of genes regulated by that enhancer [232]. The expression of the interacting genes was predicted to decrease following knockdown of eRNAs associated with PRE1 if the genes were indeed regulated by PRE1. First, it was necessary to determine whether an eRNA was transcribed from PRE1. RNA Capture-seq data (discussed in chapter 4) was examined for evidence of enhancer transcription arising from the

11q13 locus in a panel of 20 human tissues [282]. A cluster of reads was seen at PRE1 in the breast cancer cell datasets but not in other tissues. This suggested that PRE1 expresses a breast tissue specific eRNA. Examination of the assembled reads and CAGE tag data revealed the PRE1 eRNA to be bidirectional, with the sense transcript containing SNP1 (*Figure 3.10*). No transcription was found at PRE2 to support production of an eRNA at that locus.

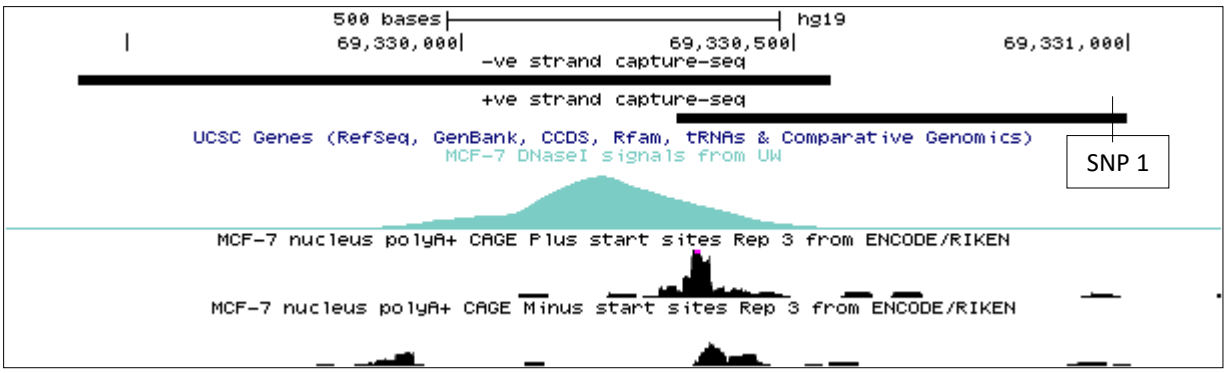


Figure 3.10 A bidirectional transcript is produced from PRE1 consistent with an eRNA. A screenshot from the UCSC browser with the Capture-seq reads for the (-) and (+) strands assembled into bidirectional single exon transcripts. The blue track shows a DNase I hypersensitivity peak overlaying PRE1 and the CAGE tags below this show the predicted transcription start sites for each of the eRNA transcripts. The position of SNP1 within the (+) transcript is indicated by a vertical line.

Further experiments focussed on the sense eRNA transcript, given that it was more highly expressed as indicated by the CAGE data and also contained one of the risk SNPs (*Figure 3.10*). The expression of an eRNA at PRE1 was supported by GROseq data which also revealed extensive nascent transcription at PRE1 in response to oestrogen stimulation (*Figure 3.1*) [87]. To confirm the presence of an oestrogen regulated eRNA at PRE1, qPCR was performed on RNA extracted from oestrogen induced MCF7 cells (*Figure 3.11*). This demonstrated significant eRNA transcription peaking at 6 hours and then falling back to baseline by 24 hours post oestrogen stimulation, confirming the eRNA to be an oestrogen regulated transcript.

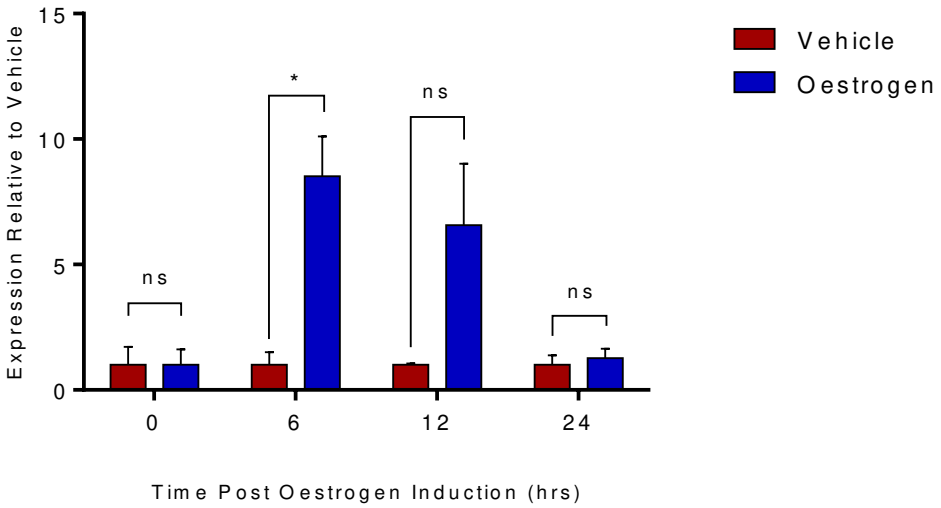


Figure 3.11 The PRE1 derived eRNA is oestrogen responsive. eRNA expression was measured using qPCR in oestrogen stimulated or vehicle control groups. Expression relative to the vehicle is shown for each time point. Data shown is the mean ± SEM from three biological replicates. Significance was determined with a two-way ANOVA test with Dunnett's test for multiple comparisons. * $p<0.05$, ns = not significant.

3.2.7. The PRE1 enhancer has promoter activity which is altered by the risk SNPs.

To investigate whether the SNPs may affect any intrinsic promoter activity of PRE1 that is required to drive transcription of the enhancer RNAs, pGL3 luciferase constructs containing PRE1 as a promoter were transfected into the ERα positive cell line MCF7 and the ERα negative cell line Bre80 (*Figure 3.12*). PRE1 exhibited marked promoter activity over the empty pGL3 Basic plasmid in MCF7 cells to a level one tenth that of the highly active pGL3 Control plasmid. The incorporation of SNP1 reduced this effect whilst SNP2 increased promoter activity. In the Bre80 cells the promoter activity was minimal (one thousandth that of the control plasmid) except in the presence of SNP3 however such low absolute luciferase levels are at the boundary of reliable discrimination for the assay. This suggests that PRE1 activity is oestrogen dependent, as supported by the eRNA oestrogen induction (*Figure 3.11*).

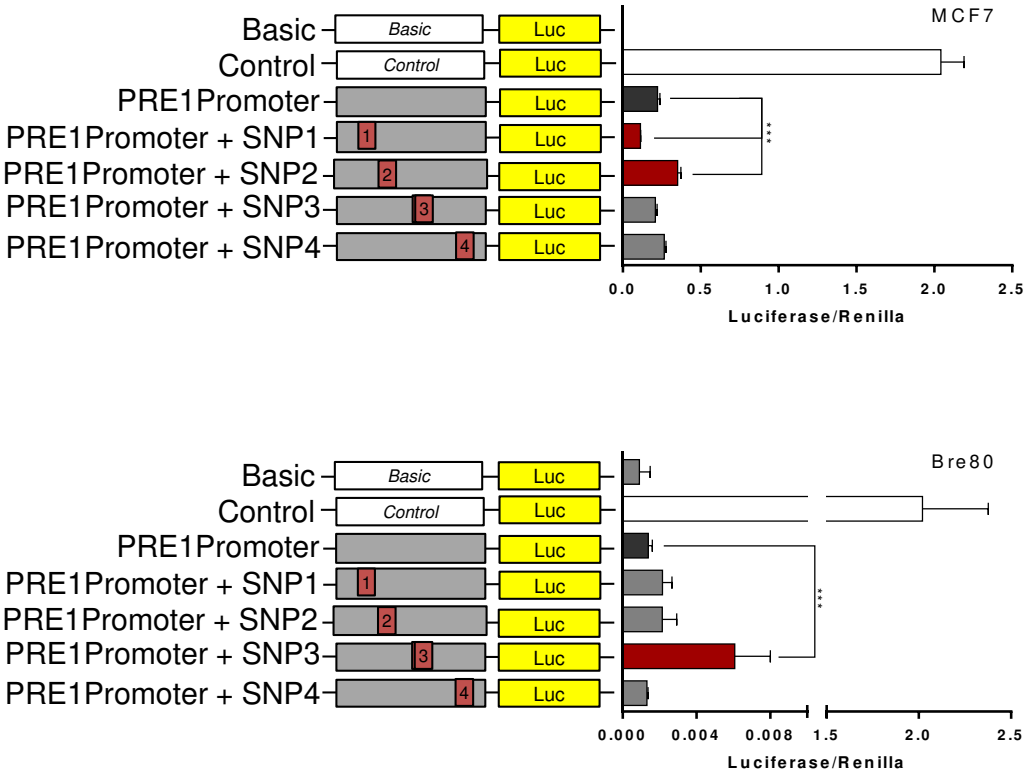


Figure 3.12 PRE1 exhibits promoter activity which is altered by the risk SNPs. The schematic to the left depicts components of the pGL3Basic plasmid with the first box containing PRE1 positioned as a promoter for the luciferase gene (yellow) and each construct containing one of the four SNPs (numbered red boxes). The basic construct lacks a promoter and enhancer, whilst the control plasmid has a strong CMV promoter to drive luciferase expression in the absence of an enhancer. The luciferase activity of the control construct is set as 2 for all experiments and the other results normalised to this to allow comparison between the MCF7 (top) and Bre80 (bottom) cell lines. Data shown is the mean ± SEM from 3 biological replicates. Significance was determined using a one way ANOVA incorporating Dunnett's test for multiple comparisons. * $p<0.05$, ** $p<0.001$, ***$p<0.0001$.

3.2.8. Silencing of the PRE1 eRNA reduces expression of CPT1A and IGHBPM2.

To confirm the role of PRE1 in regulating the 11q13 target genes identified by the 4C-seq and 5C experiments, the PRE1 eRNA was silenced using siRNA and gene expression measured using a TaqMan assay (*Figure 3.13*). The concentrations of two siRNA were first optimised and compared with the silencing obtained using two modified antisense oligonucleotides (M-ASO) which have a phospho-thiorate backbone for extra stability [283]. Following optimisation, the relative silencing efficacy was assessed using eRNA specific primers in a qPCR assay. The siRNA-2 consistently produced a 20-25% reduction in eRNA expression however siRNA-1 did not produce an effect when compared to the scrambled control siRNA (*Figure 3.13A*). The M-ASO mediated silencing

achieved a 15-20% reduction in eRNA levels. RNA obtained from the best three eRNA silencing experiments using siRNA-2 was used for subsequent TaqMan gene expression analysis of *CCND1, CPT1A, IGHBPM2, MRP12, MTL5* and *ORAOV1* (*Figure 3.13B*). Notably, *CPT1A* and *IGHBPM2* expression were significantly reduced following silencing of the PRE1 eRNA using siRNA-2. *CCND1* expression was also reduced as predicted by previous luciferase experiments [5], however this did not reach statistical significance. No significant effects were seen in the other genes assessed.

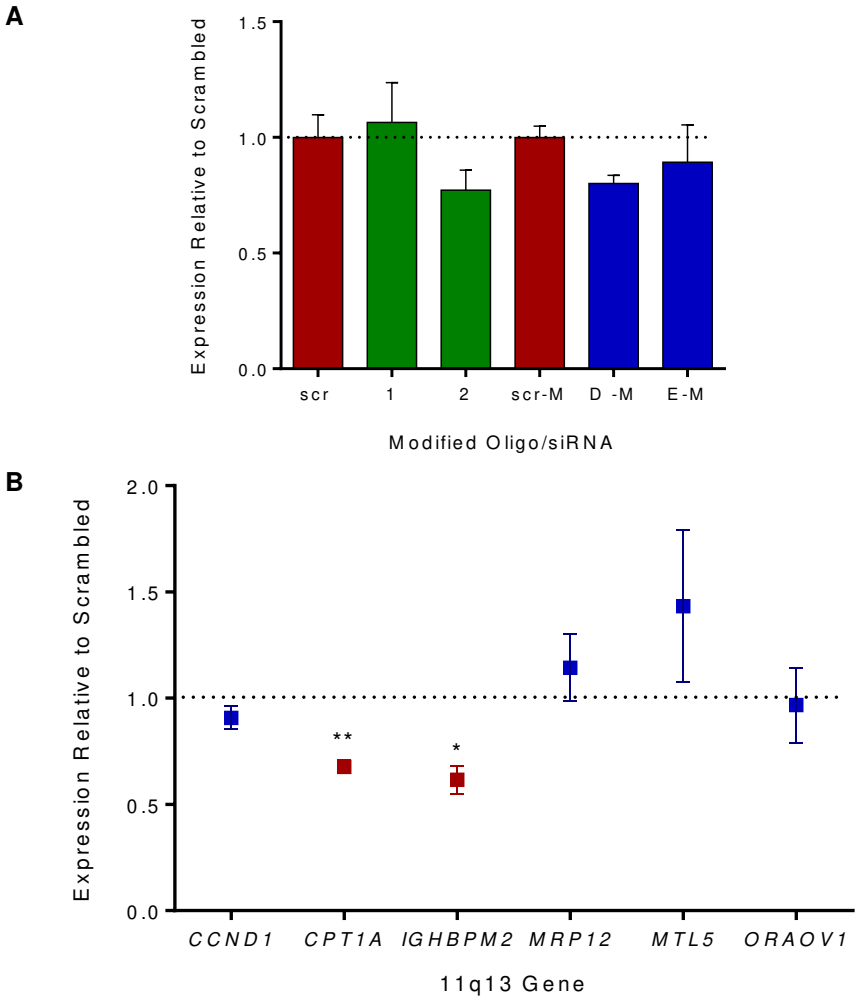


Figure 3.13 Silencing of the PRE1 eRNA reduces CPT1A and IGHBPM2 expression. (A) eRNA expression following knockdown with the modified oligonucleotides D-M and E-M relative to the scrambled control scr-M; and siRNA-1 and siRNA-2 in comparison with a scrambled siRNA (scr). Results are mean +/- SD expressed relative to the respective scrambled controls. (B) Expression of the target genes following eRNA knockdown with the dotted line indicating the level of a scrambled control. Results are mean +/- SEM of three biological replicates expressed relative to the scrambled control. Significance calculated using a one sample t test, comparing values to a hypothetical mean of 1. *p<0.05, **p<0.001.

3.2.9. Paired heterozygous cell lines were created using TALENs (Transcription activator like effector nucleases).

To assess the effect of the risk alleles on PRE1 function in their native genomic context, the SNP1, SNP2 and SNP3 risk alleles were genome edited into the T47D breast cancer cell line. This cell line was chosen as it is an ERα positive cell line without the high frequency of amplification found at the PRE1 locus in other ERα positive cell lines such as MCF7 and BT474 [284]. A pair of TALENs was designed to create a double stranded break (DSB) between SNP1 and SNP2 which could then be repaired using homologous recombination (HR). The HR process used a donor plasmid containing a PRE1 sequence with the risk alleles of SNPs 1-3 as a template to repair the DSB, in the process replacing the common alleles with the minor (risk) alleles (*Figure 3.14A*). SNP4 and SNP5 were too far from the DSB to be altered and were not explored further using this approach. To optimise the assay, the TALENs were transfected along with a donor plasmid and EGFP into T47D cells and then sorted by FACS to produce an enriched transfected population that could then be assessed using the T7 endonuclease assay (*Figure 3.14B*). This assay involved cutting the DNA at the site of mutagenesis to determine the frequency of successful TALEN activity in the cell population (*Figure 3.14C*). The presence of bands at 480bp and 260bp in the TALEN column was consistent with active cutting of the PCR template at the predicted location and validated the TALEN pair used for genome editing.

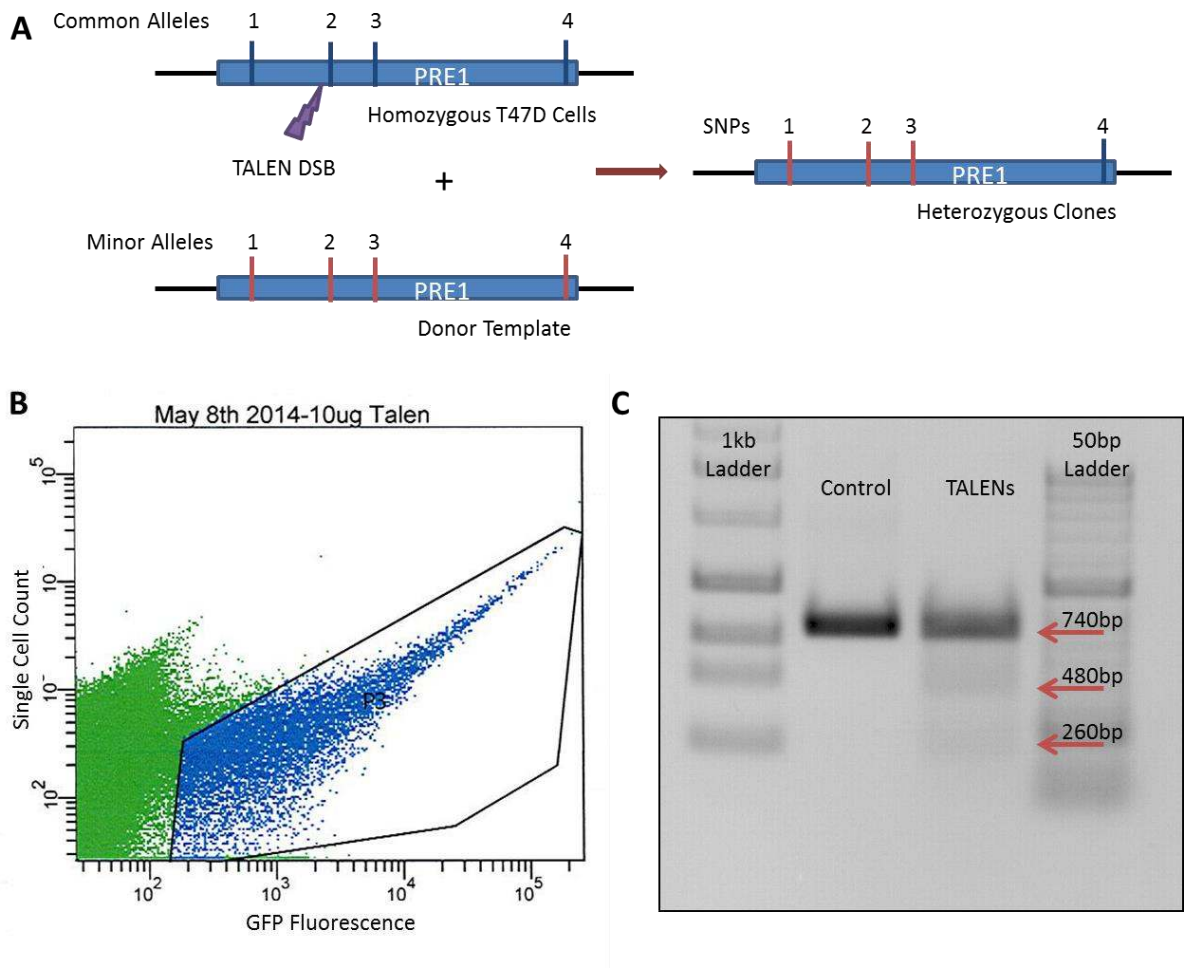


Figure 3.14 The TALENs cut PRE1 between SNP1 and SNP2. (A) A schematic of the TALEN process with a DSB (blue lightning bolt) induced at PRE1 in the T47D cells repaired using the donor template containing the minor (risk) alleles. (B) A scatter plot of individual cells during FACS sorting with the gate used to select GFP expressing cells in blue. (C) T7 endonuclease assay demonstrating the generated PRE1 amplicon at 745bp and the predicted products following digestion of the TALEN edited cells compared to a non-edited control.

A combination of approaches was required to identify individual clones which had successfully incorporated the risk alleles as all methods were found to yield false positive results when checked using Sanger sequencing. The Sequenom iPLEX Mass ARRAY was used initially to screen a large number of transfected and FACS sorted clones simultaneously. This assay uses a mass spectrometry approach to determine the identity of a selected base [285]. After DNA extraction and preparation, the samples were given to Dr Qing Chen to perform the assay. Heterozygosity was identified in 12 clones for SNP1, 7 clones for SNP2 and 8 clones for SNP3 (*Figure 3.15*). Four homozygous clones were also identified but were later found to be invalid results caused by a low DNA input. Clones that were high confidence for the heterozygous genotype were screened further using a TaqMan SNP assay to confirm the presence of the SNP3 risk allele (*Figure 3.16A*). Finally, a restriction fragment length polymorphism (RFLP) method was employed, as the common allele for SNP1

formed part of a *DpnII* recognition site which was lost with conversion to the risk allele. Digested amplicons spanning SNP1 were run on an agarose gel and compared with known heterozygotes and homozygotes. Clones 11 and D11 thus appeared to have incorporated the minor allele of SNP1 and had a digest pattern similar to the heterozygous cell line Cal51 (*Figure 3.16B*).

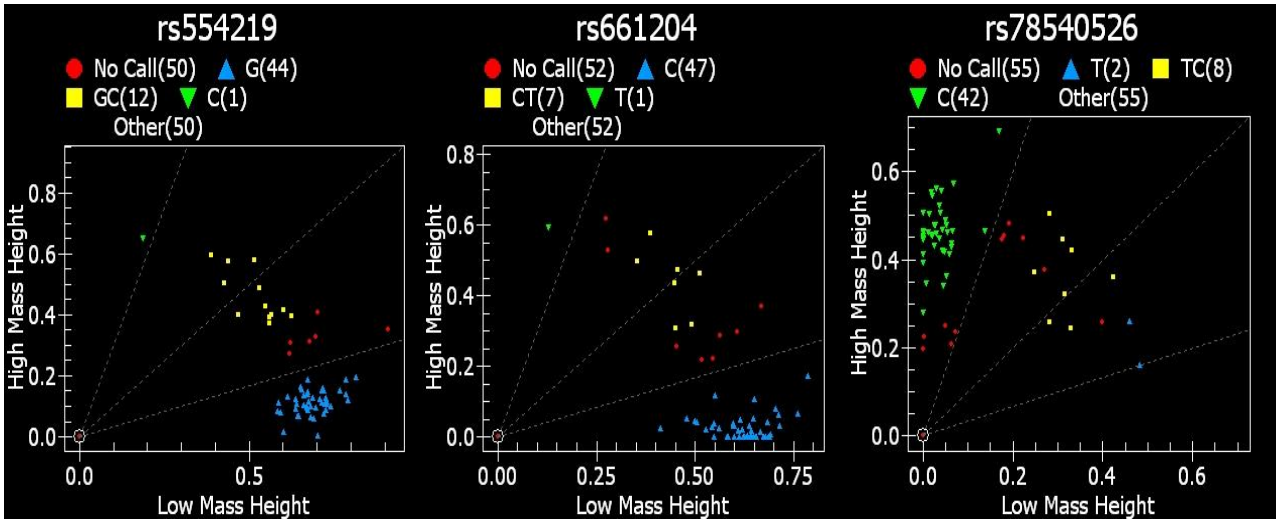


Figure 3.15 Screening of the TALEN clones by Sequenom iPlex Mass Array. The Sequenom assay separated the clones by mass of the SNP allele into homozygotes (blue or green), heterozygotes (yellow) or uncertain (red). Panels depict SNP1 (rs554219), SNP2 (rs661204) and SNP3 (rs78540526).

The final selection of clones for Sanger sequencing included only those clones with concordance between their Sequenom, RLFP and SNP TaqMan results. Clones D11 and 11 were confirmed to be heterozygous for SNPs 1,2 and 3 whilst G3 and G9 were heterozygous for SNPs 2 and 3 but retained the common allele of SNP1 (*Figure 3.17A*). The frozen culture of clone D11 failed to grow after thawing and was unable to be used. The ratio between the peaks seen for the common allele vs the introduced risk allele on Sanger sequencing was markedly biased towards the existing common allele in all clones. This suggested that the T47D cells may be polyploid for the PRE1 locus and that only a small proportion of the SNP1 common alleles had been converted to the risk allele. DNA-FISH was performed to assess the number of copies of the 11q13 locus per cell (*Figure 3.17B*). The FISH hybridisation process was performed by Mrs Kristine Hillman. The results revealed that at least 3 copies of 11q13 are present in the T47D cells that can be distinguished using the resolution of the FISH probes.

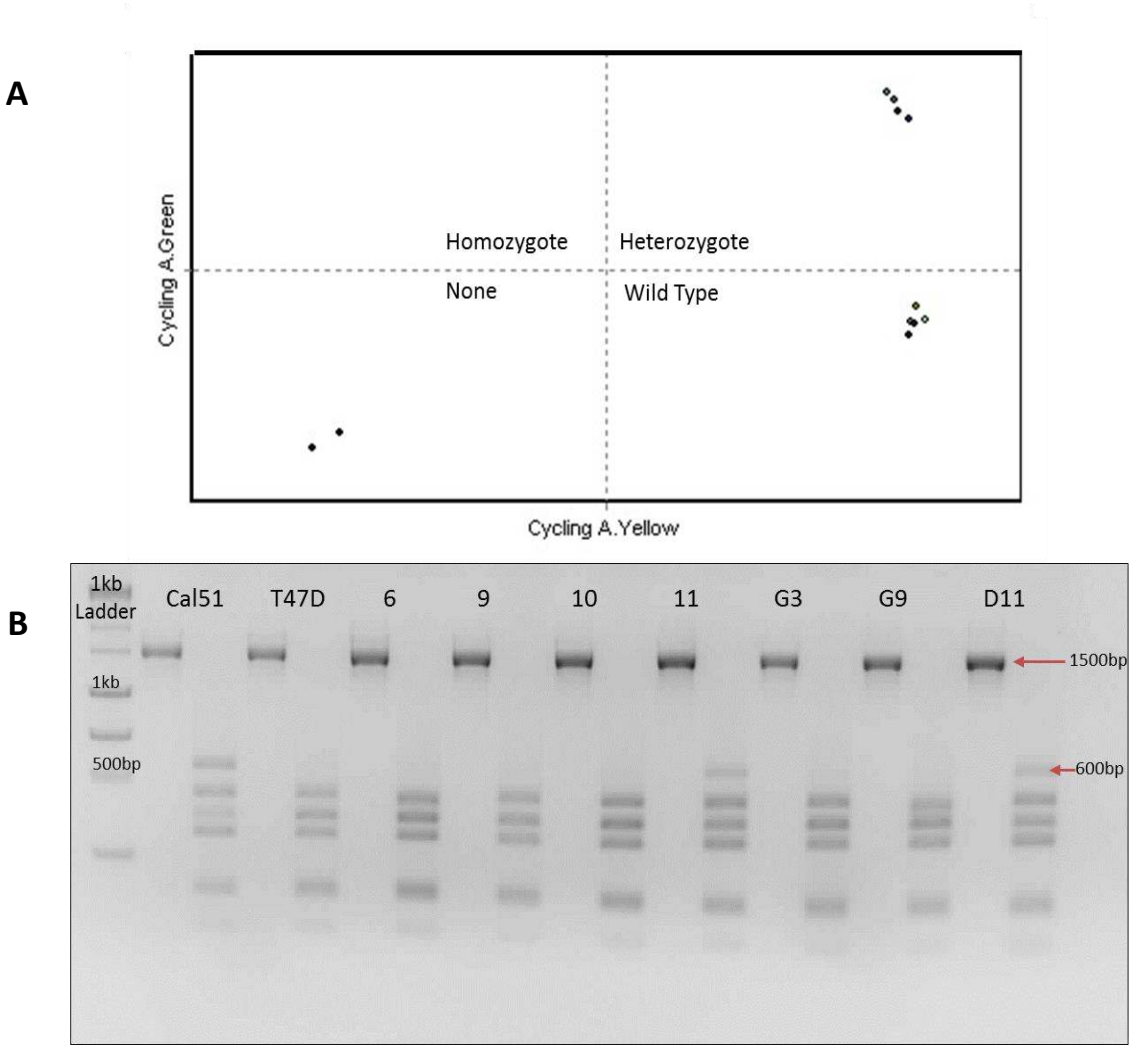
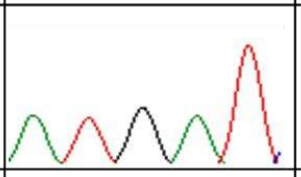
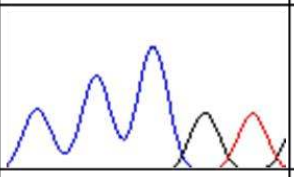
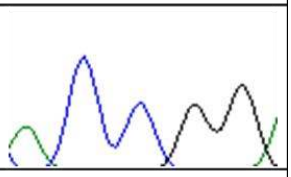
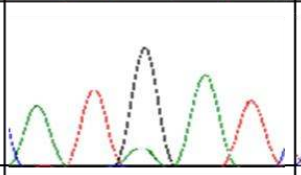
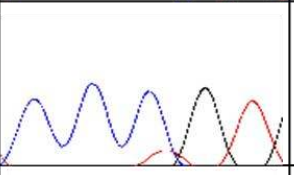
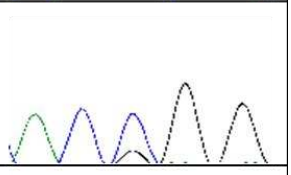
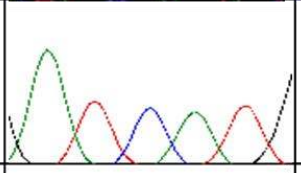
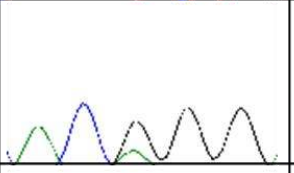
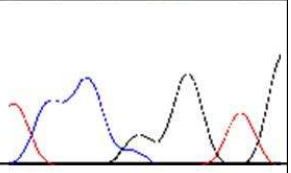
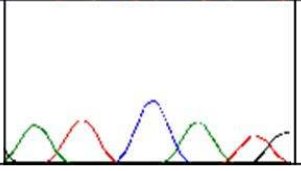
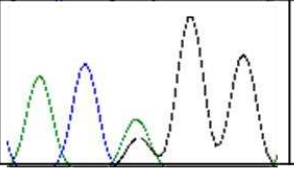
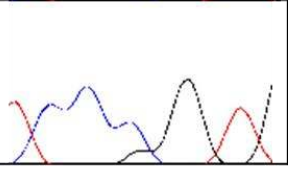


Figure 3.16 Screening of the TALEN clones detects heterozygotes for the risk SNPs. (A) The TaqMan SNP assay separated the clones according to the allelic status of SNP3. The ratio of signal from probes designed against the common allele (yellow) or risk allele (green) clustered according into three groups. Heterozygotes are found in the top right and include the heterozygous Cal51 control. There are no homozygotes (top left) for the risk allele amongst the clones. Wild type (homozygous for common allele) cluster in the bottom right and include the homozygous T47D control. The two negative (no DNA) controls are in the bottom left quadrant. (C) An agarose gel image of the products from each clone following digestion of a 1500bp amplicon with *DpnII*. The red arrow indicates the location of a 600bp band which remains undigested in the presence of the risk allele of SNP1. DNA from homozygous T47D cells and heterozygous Cal51 cells is included for comparison.

A

Clone	SNP1 (G/A)	SNP2 (G/A)	SNP3 (G/C)
T47D Wild Type			
11			
G3			
G9			

B

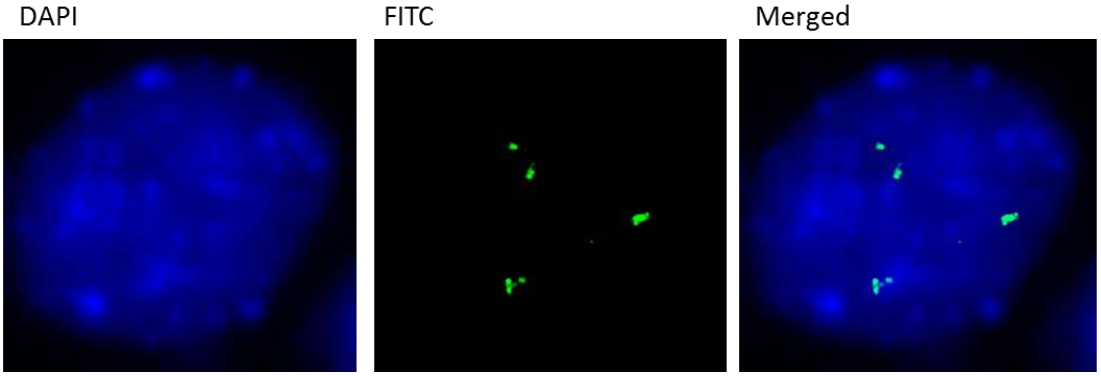


Figure 3.17 Sanger sequencing and DNA FISH confirms multiple copies of the PRE1 locus. (A) Representative chromatograms are shown for the unedited T47D cell (1st row) and the three positive clones. The SNP alleles are positioned at the central peak, with two adjacent peaks on either side. The colours of the peaks vary depending on the identity of the base and the orientation of the primer used for sequencing. The presence of two overlapping coloured peaks is indicative of heterozygosity for the allele at that location. SNP1 = rs661204, SNP2 = rs78540526, SNP3 = rs554219. (B) Images from the GE Deltavision deconvolution microscope following DNA-FISH of T47D cells. The frames are split into DAPI (left) to demarcate the nucleus; FITC (middle) to recognise the labelled BAC probe; and a merged view on the right panel.

3.2.10. No significant change in gene expression was found in PRE1 heterozygous clones.

To determine the effect of the risk SNPs in their native genomic context, expression levels of the target genes identified by 4Cseq and 5C were measured in three heterozygous clones using a TaqMan assay. Gene expression was compared to the average obtained for the three control cell cultures (clones 6, 9 and 10) that had gone through the TALEN process but remained homozygous for the major allele (*Figure 3.18*). None of the changes in gene expression reached statistical significance although there was a trend for reduced *CCND1* levels in clones G3 and G9 in the presence of SNP2 and SNP3 as predicted by previous luciferase assays [5].

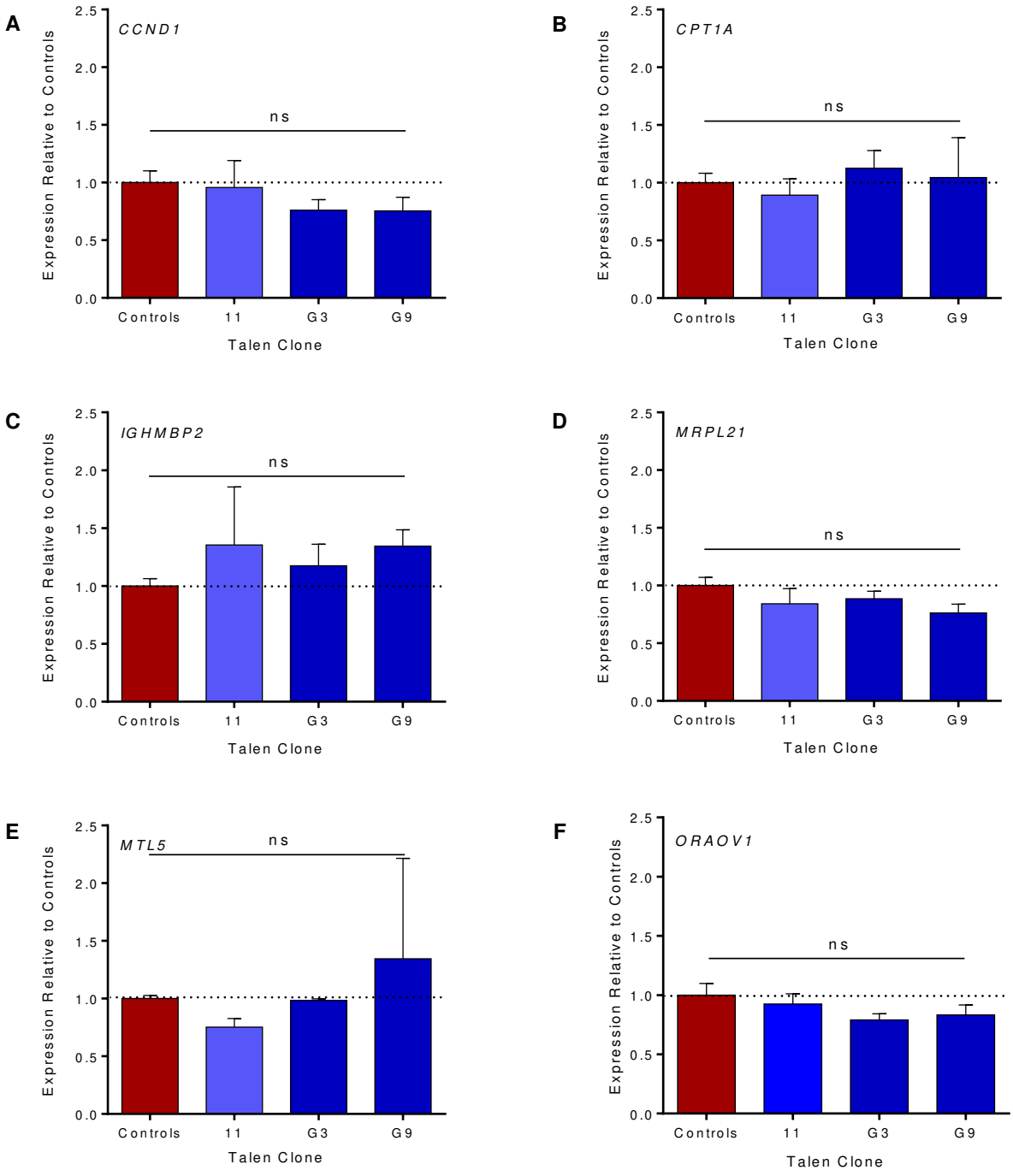


Figure 3.18 Relative gene expression in heterozygous clones compared to the homozygous wild type. The average expression of three combined control clones is shown by the red bar and normalised to a value of 1. Target gene expression is shown in blue for each clone relative to the control value. Note that clone11 is a heterozygote for all three SNPs whilst clones G3 and G9 are heterozygous for SNP2 and SNP3. Data obtained is mean +/- SEM of three independent RNA extractions per clone. Significance was assessed using a one way ANOVA with Dunnett's correction for multiple comparisons. ns = not significant.

The 11q13 breast cancer susceptibility locus was fine mapped and characterised by our group in 2013. That study identified five candidate causal SNPs driving the association and their likely target gene, *CCND1* [5]. There remain however, additional genes in the region with known roles in cancer that needed to be included or excluded as contributors to breast cancer risk using a more agnostic approach. In this chapter, the PRE1 and PRE2 risk-associated regulatory elements were further characterised using a combination of chromosome conformation capture variants with other molecular techniques. This demonstrated that PRE1 was breast specific and produced a novel eRNA containing one of the risk SNPs. It also revealed interactions between PRE1 and the *CPT1A* and *IGHMBP2* genes. These findings represent novel mechanisms by which the breast cancer associated SNPs may be mediating disease at this locus.

The 11q13 interval contains risk-associated SNPs for multiple cancer types including breast, renal and prostate cancer [4, 271, 272]. The breast SNPs were previously shown to lie in two regulatory elements called PRE1 and PRE2 and regulate *CCND1* [5]. The 3C results indicated that the PRE1/*CCND1* interaction is breast specific, consistent with the extensive binding of the breast specific TFs ERα, SRC3 and FoxA1 at PRE1 [135, 286]. The tissue specificity was also reflected in the luciferase results with PRE1 inducing a minimal change in *CCND1* promoter activity when assayed in prostate cells (*Figure 3.2E*), whereas it exhibited high activity in breast cancer cell lines [5]. In contrast, PRE2 can interact with *CCND1* in several different tissue types consistent with the lack of breast specific TF binding (*Figure 3.2B,D, Figure 3.1*). It also retained its ability to act as a silencer in prostate cancer cells (*Figure 3.2E*). This may indicate that a conserved chromatin loop between PRE2 and *CCND1* is important to regulate cell proliferation by the repression of *CCND1* expression. The repression would then be overcome by the activation of tissue specific enhancers such as PRE1 in breast tissue following oestrogen stimulation. Similarly, androgen stimulation in prostate tissue may activate prostate specific enhancers in the locus to allow activation of *CCND1*. A genome editing approach in breast and prostate cells where *CCND1* expression is assessed following the inactivation of PRE2 would assist in confirming this hypothesis.

Interestingly, the 11q13 renal SNPs are also located in enhancers, affect the binding of HIF (hypoxia induced factor) and are involved in looping interactions to *CCND1*. These loops were renal specific and not present in MCF7 breast cancer cells [271]. The prostate cancer locus has not yet been fully characterised, however 3C using a bait containing enhancers affected by prostate risk associated SNPs demonstrated interactions with the *CCND1* promoter in prostate cancer cells (*Figure 3.2F*). This suggests that the risk of different cancers at the locus is likely mediated though

a common target gene (*CCND1*). A similar situation is present at the 8q24 locus which contains risk SNPs for multiple epithelial cancers in the same manner as the 11q13 locus [287]. Studies at 8q24 have shown that colon, breast and prostate cancer risk at that locus are likely to be due to independent SNPs in distal enhancer elements that regulate the *MYC* oncogene [155, 271, 288, 289]. The enhancers in the region also largely tissue specific when assessed using *in vivo* assays and the risk loci do not overlap apart from a single haplotype block associated with both prostate and colorectal cancer [287, 289]. The finding of risk SNPs for multiple cancers in the same locus is likely to be a recurrent theme in cancer GWAS as more studies become available given the common genetic mechanisms underlying the development of cancer in different cell types [290].

We have previously demonstrated that *CCND1* is a target gene at the 11q13 locus, however recent analyses by the ENCODE and FANTOM consortia indicate that regulatory elements can engage in multiple long-range interactions [3, 86]. A candidate approach was first used to identify additional target genes at the 11q13 locus. *In silico* and expression analysis of five local genes indicated that *ORAOV1* was the next most active gene in the region after *CCND1* (*Figure 3.3*) [87, 275]. *ORAOV1* also plays an important role in cell proliferation and angiogenesis in other tumours, in particular oral squamous cell carcinoma [270]. The initial 3C using the *ORAOV1* promoter as bait revealed a strong interaction peak over PRE1 that was not altered by the presence of oestrogen despite *ORAOV1* expression being oestrogen responsive (*Figure 3.5C*). This is frequently seen with chromatin looping, where the loops are often pre-existing but then require recruitment of regulatory TFs for transcription to occur [154, 161]. Further 3C experiments using PRE1 as the bait however, demonstrated that the interaction between PRE1 and *ORAOV1* actually localised to the second intron rather than the promoter (*Figure 3.6C*). This suggested that *ORAOV1* was not a direct target of PRE1 but it did not exclude the possibility that PRE1 may still have some regulatory effect on *ORAOV1* by their proximity in nuclear space, such as within a transcription factory [291]. ChIA-PET data indicated that *ORAOV1* is on the boundary of the TAD containing the 11q13 gene desert and further looping interactions connected this boundary with *CCND1* (*Figure 3.6E*). Given that PRE1 interacts strongly with *CCND1*, this provides a mechanism by which PRE1 and *ORAOV1* may be indirectly brought into contact [5, 277].

PRE1 was considered particularly likely to have additional gene targets. In contrast to PRE2, it is a chromatin interaction hub as revealed by the ERα mediated CHIA-PET data (*Figure 4.9*), and has very high levels of DNase I hypersensitivity indicating open, active chromatin (*Figure 3.1*)[161, 275]. This is consistent with a subgroup of enhancers identified by the FANTOM CAGE study that

bound high levels of p300 and cohesion, had higher H3K4me3 levels than expected for an enhancer and produced longer eRNAs (530bp median vs 346bp) [86]. These were associated with a high frequency of ChIA-PET interactions with other enhancers and gene promoters. PRE1 shares all these features, making it highly likely that other interactions will be present. These may be local interactions *in cis* or distal interactions *in trans*, potentially on other chromosomes. Local interactions at 11q13 may be explored using 5C, whilst interaction *in trans* can be best explored using an agnostic 4C-seq approach [174, 292].

The 4Cseq was initially performed in the ERα positive breast cancer cell lines MCF7 and BT474 to detect interactions that may be relevant to ERα positive breast cancer. The *trans* interactions all involved genomic loci that were amplified in breast cancer and contained genes previously known to be important in breast cancer biology (*Figure 3.8*). These interactions may be a true representation of the chromatin interactome or may potentially be an artefact produced by the fixation of random interactions between the highly abundant DNA of the amplified regions [284]. Interestingly, the interactions on chromosomes 17q23 and 20q13 were not found for the PRE2 bait, suggesting that they are a valid representation of the PRE1 interactome. This is supported by Hsu *et al*. who describe the occurrence of amplified clusters of oestrogen receptor elements (EREs) in MCF7 cells at 17q23 and 20q13 which were hotspots for inter-chromosomal interactions [280]. They speculate that these elements may associate in transcription factories with other similarly amplified ERα binding regions which would explain the observed result for PRE1 as it lies in a heavily amplified region and is highly bound by ERα [8, 129]. Such associations may be functional, leading to the deregulation of tumour suppressor genes interacting with the amplified EREs and hence drive tumour growth [280]. Hsu *et al.* note that MCF7 and BT474 cells in particular have high levels of ERE copy number gain at 20q13 and that this amplification is driven by oestrogen exposure which is an unavoidable part of culturing such cells in the laboratory [280].

The use of a control library in future experiments may provide some clarification on whether the results were due to captured proximal interactions or an artefact due to high levels of amplified DNA being available to interact in the ligation assay. 4C-seq protocols typically do not include the use of control libraries that have not been formaldehyde fixed to enrich for interacting regions [172, 293], however this may be required when dealing with cancer cell lines containing high levels of copy number variation. Interactions present in the control libraries could thus be subtracted from any analysis in the actual fixed libraries. Instructively, the vast majority of published studies using 4C-seq do not use cancer cell lines [274, 294-296]. One exception is a study by Zeitz *et al*. that also

used MCF7 cells to find genome-wide interactions for an enhancer of *IGFBP3* on chromosome 7 and detected a high number of interactions with the regions adjacent to those detected by the PRE1 4C-seq (17q22 and 20q12) [297]. No control library was used, however the interactions were also present in the non-cancer breast cell line HMEC and confirmed using DNA-FISH.

The interactions may also be a result of translocations as all the detected regions are frequently translocated in breast cancer cells [298]. The 11q13 region in particular is often a translocation partner with the 17q23 region where *HEATR6* is located [299], and both 17q23 and 20q13 are among the four most translocated regions in MCF7 cells [298]. Indeed, 23.6% of all fusion events mapped throughout the MCF7 genome involve 17q23 and 20q13, with amplified EREs enriched at the breakpoints [280]. The translocations between 11q13, 17q23 and 20q13 may be a direct result of their *in trans* interactions as co-localisation of active genomic loci is a key driver in the formation of translocations [300, 301]. Importantly however, the interactions seen *in trans* for PRE1 and PRE2 were not found when a non-cancer breast cell line (Bre80) was used for the 4C-seq. This indicates that although the PRE1 interactome described for MCF7 and BT474 cells may be functional and important in cancer progression, it is unlikely to be relevant to the SNP-associated risk of developing breast cancer which is the focus of this chapter.

The obtained interactions patterns *in cis* and *in trans* did not enable accurate identification of any potentially interacting promoters but instead demonstrated larger scale interactions between genomic regions as is obtained using the HiC technique [302]. The lack of resolution may be related to inadequate digestion or ligation steps in the preparation of the 4C libraries, although the standard quality control checks of such measures appeared adequate when assessed [303]. Similar results have also been obtained by other groups, for example Apostolou *et al.* used 4C-seq to identify interactions involving the *NANOG* promoter in embryonic stem cells and found 'broad domains' of contact rather than detailed, discrete interactions [296]. More recent 4C-seq protocols advise that resolution can be improved with the use of two 4bp cutter restriction enzyme digests rather than the initially recommended 6bp cutter restriction enzyme digest followed by a 4bp cutter restriction enzyme final digest [304]. Adoption of this revised method may have improved discrimination for the assay, allowing better identification of the interacting targets of PRE1 and PRE2.

One prominent feature of the 4C interaction data was the high frequency of interactions near centromeres seen for both PRE1 and PRE2 which were interpreted as artefactual (*Tables 3.1, 3.2*). This has not been commented on in previous studies using the 4C technique [294-296], however the

majority of such studies were performed in embryonic or pluripotent stem cells which have fewer centromeres positioned peripherally compared to cells which have undergone differentiation [305, 306]. This may affect the frequency of pericentromeric interactions involving PRE1 and PRE2 as these elements are located on chromosome 11 which in turn localises to the nuclear periphery [307]. Centromeres also tend to cluster and thus any interactions may be expected to involve multiple centromeres as was indeed observed [307]. Alternatively, the interactions may represent functional contacts with significance in cellular metabolism or cancer. Despite pericentromeric regions being enriched for inactive heterochromatin, an increasing number of genomic regions have been found to reside there as condition for their normal functioning and it is possible that *CCND1* and the 11q13 gene desert fall into that category [308]. A third possibility is informed by the fact that 4C data represents an average of contacts within a cell population and there thus may be a proportion of cells in which the 11q13 region is moved into a pericentromeric region for silencing when the cells are not dividing and another population of cells in which 11q13 has relocated to a more active region of the nucleus to drive the high levels of *CCND1* expression required for progression through the cell cycle [306, 309]. Experiments involving single cell transcriptomics and chromosome conformation capture may help to clarify this further [310, 311].

PRE1 was found to express relatively high levels of eRNAs which are indicative of significant enhancer activity [87]. This eRNA production at PRE1 was confirmed using Capture-seq data (*Figure 3.10*) and also direct qPCR, which demonstrated that the eRNA was oestrogen responsive (*Figure 3.11*). Furthermore, publically available GROseq data showed nascent transcription at PRE1 increasing rapidly over the first hour post oestrogen stimulation and then falling by the 3 hour mark [87]. These early changes were not explored in the oestrogen induction described in *Figure 3.11* as the data time points were assigned to six hour intervals, but may have been observed using more frequent measurements taking into account the fact that mature expression of a transcript will obviously lag behind nascent transcription.

Transcriptional enhancers such as PRE1 can also display promoter activity which drives their eRNA production [229]. The breast cancer risk SNPs within PRE1 have previously been shown to markedly alter the ability of PRE1 to enhance transcription, raising the possibility that they may act by reducing production of a functional eRNA [5]. PRE1 was shown to drive luciferase production when cloned into a pGL3 construct as a promoter, confirming that it had promoter activity. This activity was altered by inclusion of the SNPs, with SNP1 causing a 50% reduction in promoter activity. Lam *et al.* demonstrated that a 50% reduction in eRNA expression reduces the expression of target genes by a similar magnitude so the disruption of eRNA production may represent a novel

mechanism by which risk SNPs act to alter local gene expression [232]. It is also interesting to note that SNP1 is contained within the sequence of the (+) strand eRNA raising the question whether it may represent a riboSNitch inducing a conformational change in RNA structure and hence alter the function of the eRNA [312].

Given that eRNA transcription from an enhancer precedes transcription from its target gene it has been proposed that there may be a causal relationship and knocking down the eRNA will therefore abrogate expression of the target gene [232, 235]. This has been confirmed by several groups and appeared to be a specific effect as eRNA silencing reduced expression of only those genes regulated by the enhancer producing the eRNA [232, 234]. The effect of eRNA silencing was assessed on the genes identified through 4Cseq and 5C to confirm that they are regulated by PRE1 and prioritise them for further functional assessment. Optimisation of the eRNA knockdown included a comparison between standard siRNA and modified antisense oligonucleotides (M-ASO) possessing a phosphoro-thioate backbone, which makes them far more resistant to nucleases than standard RNA molecules [283]. They are able to activate the RNaseH pathway in the nucleus which efficiently hydrolyzes the bound target RNA [313]. Both the siRNA and M-ASO were designed against the (+) strand eRNA as this contained the risk SNP and was more highly expressed according to CAGE data (*Figure 3.10*). The (+) and (–) eRNA transcripts may have different activities however and the experiments are currently being repeated by another lab member including (–) strand eRNA knockdown to ensure no strand specific effects are missed [232]. A comparison between the different M-ASOs and siRNAs found that one of the siRNAs (2) produced the most reliable knockdown and this was then used to assess gene expression changes (*Figure 3.13*). Significant reductions of *CPT1A* and *IGHMBP2* were observed along with the predicted reduction in *CCND1* levels, though this did not reach significance. *CPT1A* encodes a protein (carnitine palmitoyltransferase 1A) required for transporting fatty acids into mitochondria and has recently been identified as being a driver of proliferation in luminal breast cancer [314, 315]. *IGHMBP2* is an immunoglobulin mu binding protein that is predicted to have a role in DNA damage repair [316-318]. Both genes are promising candidates as breast cancer associated genes, however further functional characterisation is still required before they can be confirmed as mediating risk at the 11q13 locus.

Functional characterisation of GWAS identified SNPs has been historically limited by the need to work with tumour derived cell lines and artificial plasmid based assays such as luciferase expression constructs. These approaches are unable to fully address the actual effect of a SNP in its proper genomic context and this needs to be kept in mind when interpreting such data [26]. We now

however have a suite of tools, including TALENS, Zinc Finger Nucleases and CRISPRs, to alter SNPs *in situ* and create isogenic cell lines which only differ in the particular SNP(s) of interest [196]. TALENs were used to create heterozygous cell T47D cells containing SNPs1, 2 and 3, with SNP4 being too far from the TALEN recognition site to be changed without employing multiple TALEN pairs. The efficiency of homologous recombination mediated editing rapidly falls off with increasing distance from an induced double stranded break (DSB) with 3% of clones converting at 400bp from the DSB, compared to 40% converting at 46bp [319]. Sanger sequencing of the clones suggested that there were likely to be multiple copies of PRE1 and this was confirmed using DNA-FISH (*Figure 3.17*). Subsequent assessment of target gene expression in the heterozygous clones did not demonstrate a significant effect, however any changes would have been greatly diluted by the presence of multiple copies of the common allele, making the results inconclusive (*Figure 3.18*).

It proved difficult to find an ERα positive breast cancer cell line that was not amplified at 11q13. This may be related to the fact that oestrogen induced proliferation of such cells is mediated by *CCND1* which is present at 11q13 and drives a positive clonal selection of cells amplified in this region [200]. T47D cells were selected as they were the least amplified of the available cell lines [284], but the efficiency of genome editing was not sufficient to produce homozygotes for the minor allele or even heterozygotes containing an even proportion of minor and major alleles. CRISPRs may provide improved efficiency of genome editing for future work as studies report conversion rates approaching 50% in optimized conditions [320]. Ideally genome editing should be performed in a non-cancer cell line to avoid 11q13 amplification, however it needs to be a cell line that expresses ERα to be an appropriate model of ERα disease and such cell lines are not currently available.

This chapter aimed to find additional genes interacting with PRE1 or PRE2 that may mediate the effect of the risk SNPs at the 11q13 breast cancer risk locus. *ORAOV1* was initially examined however its role was not validated by the 5C or eRNA knockdown experiments and there was insufficient evidence to pursue it further as a candidate. The 4C-seq and 5C in combination with the eRNA knockdown did suggest a role for the *CPT1A* and *IGHMBP2* genes. The latter was of particular interest given its possible involvement in DNA damage repair which is a common feature of genes conferring an increase in breast cancer susceptibility [2]. The genome editing strategy was not successful in demonstrating an effect of the SNPs in their native genomic context, however the remainder of the functional approaches used to characterize the region provide an experimental framework for further investigations of breast cancer associated risk loci.

CHAPTER 4

Novel non-coding transcripts contribute to breast cancer susceptibility at the 11q13 locus

.

The majority of breast cancer GWAS SNPs fall in regions of the genome that do not contain known protein-coding genes and are predicted to affect regulatory elements such as transcriptional enhancers [163]. Recent evidence has shown that these gene poor areas are also extensively transcribed and contain multiple non-coding transcripts such as long non-coding RNAs (lncRNAs) [7, 212, 220]. These lncRNAs may be directly affected by the incorporation of risk-SNPs into the transcript itself or indirectly if regulated by an enhancer that contains functional risk-SNPs [162, 321]. The 11q13 breast cancer risk locus contains four candidate causal SNPs located in an enhancer element (called PRE1) and a fifth SNP falling in a silencer (called PRE2), with both elements regulating the nearby *CCND1* gene [5]. Interactions between PRE1 or PRE2 and additional protein coding genes have already been explored (Chapter 3), however non-coding transcripts may also be contributing to risk at this locus. Such transcripts can be missed by a traditional RNA-seq process as they are frequently tissue specific and expressed at low levels [322].

To address the limitations of RNA-seq, Mercer *et al*. developed a technique called RNA Capture-seq which incorporates an enrichment step to amplify transcriptional products from a defined portion of the genome, followed by deep-sequencing and mapping to identify the transcripts (*Figure 4.1*) [7]. Enrichment is achieved by the use of hybridisation tiling arrays with the bound transcripts then eluted for sequencing [323]. As an example of the improved detection power, 204 novel isoforms were found for proteins in the regions tiled by Mercer *et al*. and only 31.3% of the intergenic transcripts found were also seen in the corresponding RNA-seq dataset. Notably, they achieved an estimated 380-fold enrichment over the tiled region which would require around 10 billion standard RNA-seq reads to give comparable coverage [7].

Several noncoding RNAs have already been identified at cancer GWAS loci, particularly from the 8q24 region which like 11q13 is a susceptibility locus for multiple cancers [324]. These include the lncRNA *CARLo-5* which shares an enhancer containing a colon risk SNP with the *MYC* oncogene and exhibits increased expression in the colon tissue of people possessing the risk allele [325]. *CCAT1-L* is another lncRNA implicated in colorectal cancer that is found at the locus, with overexpression increasing tumour size in xenograft experiments [204].

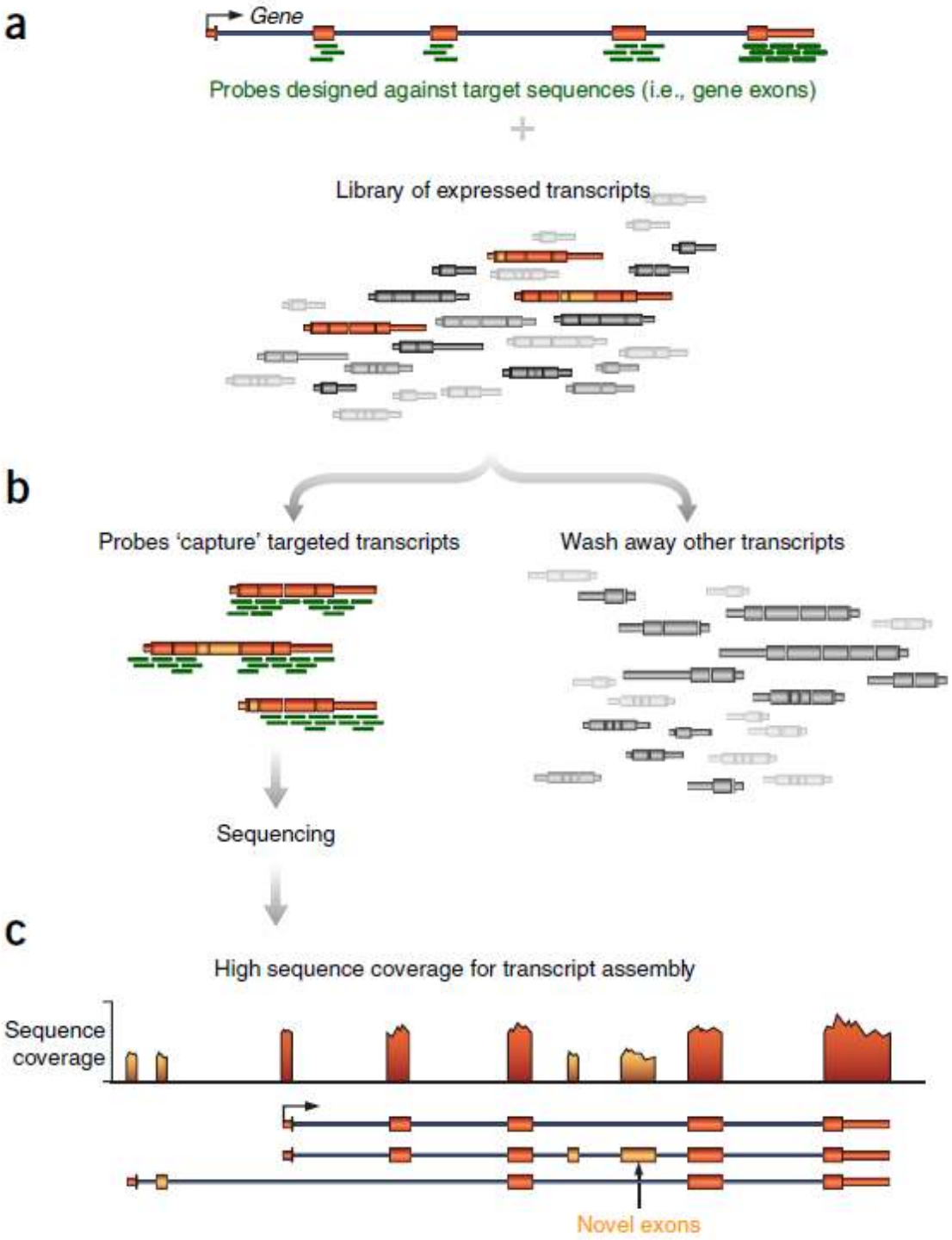


Figure 4.1 RNA Capture-seq (a) Probes are designed against exons in the regions of interest (or intergenic regions if novel transcript discovery is intended); (b) The transcripts are captured through hybridisation and pulled down; (c) The retained reads are assembled and analysed. Captured exons are shown in orange and unwanted sequences in dark gray. Adapted from [259].

This chapter investigated the hypothesis that novel lncRNAs are expressed at the 11q13 breast cancer risk locus and can contribute to breast cancer biology. The locus was interrogated by RNA Capture-seq which identified two oestrogen induced lncRNAs that were regulated by PRE1. The risk SNPs reduced chromatin looping between PRE1 and the lncRNA promoter which is predicted to reduce their transcription. The lncRNAs were named *CUPID1* and *CUPID2* and functional analysis revealed that they had a probable role in DNA damage repair. A reduction in their expression as mediated by the risk SNPs is predicted to increase breast cancer susceptibility due to the vital role of DNA damage repair genes in the maintenance of genome stability [2].

4.2.1. RNA Capture-seq reveals novel transcripts CUPID1 and CUPID2 are transcribed from the 11q13 gene desert.

RNA Capture-seq was performed on six breast cancer cell lines and one normal human breast tissue sample to identify lowly expressed transcripts that may have been missed by previous RNA-seq studies. Assembly of the reads revealed a number of single exon transcripts and two complex divergent transcripts arising from the 11q13 intergenic region (*Figure 4.2*). The H3K4me3 histone modification and DNase I hypersensitivity was present at the origin of the divergent transcripts suggesting promoter activity (*Figure 4.2*). A corresponding large DNase I hypersensitivity peak was present over PRE1 indicating highly active chromatin (blue box). The two transcripts were prioritised for closer examination and their raw mapped read densities are displayed in *Figure 4.3*. Transcription was largely confined to ERα positive cell lines for the transcript on the (+) strand whilst the (-) transcript was expressed in all the cell lines examined. A marked increase in read depth was seen for the MCF7 cells following their induction by oestrogen (*Figure 4.3*).

These two transcripts were prioritised for further study based on their complexity, oestrogen responsiveness and evidence of enrichment in breast tissue as shown by the CAGE (Cap Analysis of Gene Expression) and ChIPseq data (*Figure 4.2*) [3, 326, 327]. Their coding potential was then assessed using an online tool, the Coding Potential Calculator (Peking University - http://cpc.cbi.pku.edu.cn/). The (+) transcript returned a score of -1.03427 indicating very low coding potential. The (-) transcript returned a score of -0.975903 which also suggested a low coding potential, however there was a 373bp long fragment which corresponded to an unreliable ORF (open reading frame) with the potential to produce a peptide. As putative lncRNAs, the transcripts were named according to lncRNA nomenclature as *CUPID1* and *CUPID2* (*CCND1* Upstream Intergenic DNA Damage associated RNA 1 and 2) [328]. The evidence for their association with DNA damage repair will be presented in *Sections 4.2.14-16*.

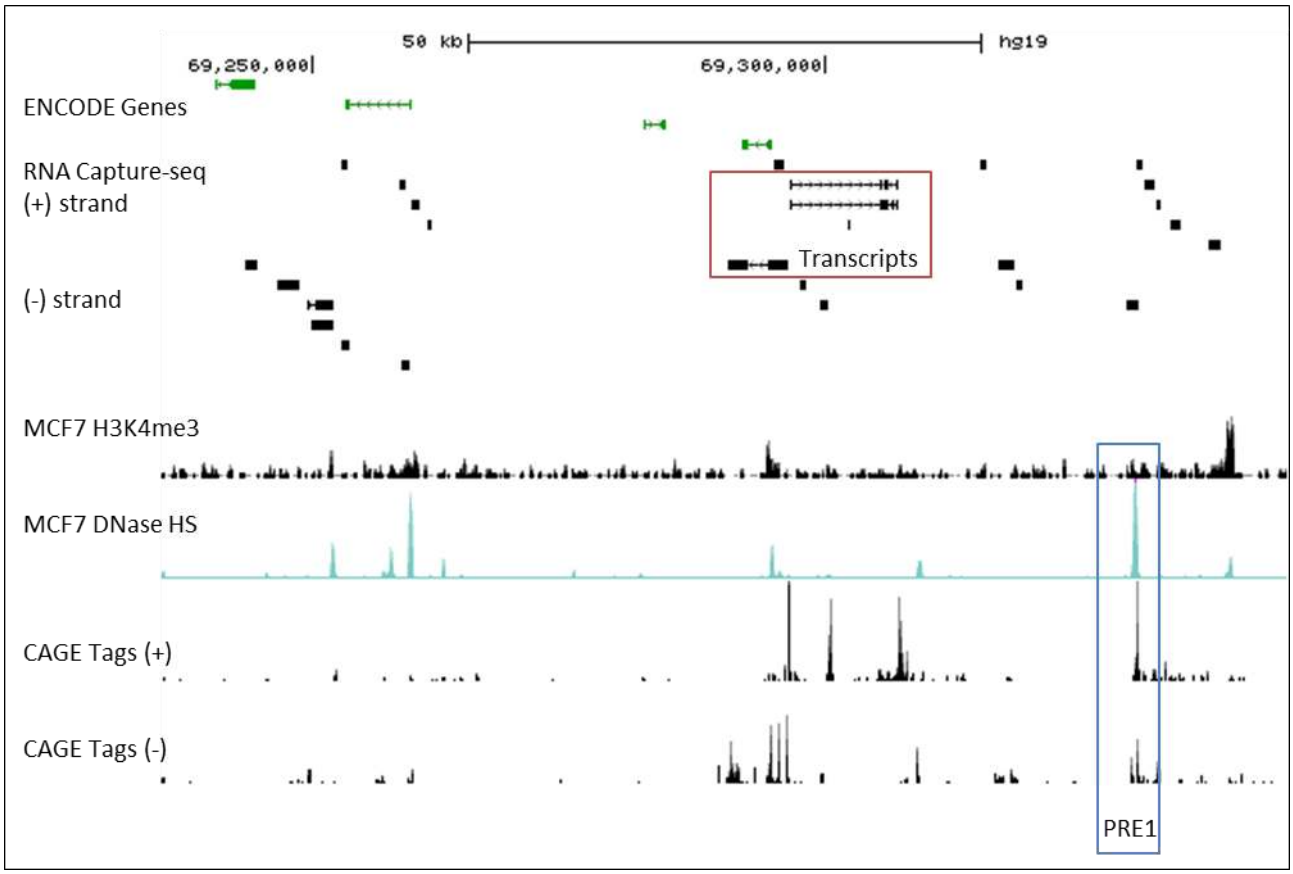


Figure 4.2 Transcripts detected by RNA-Capture-seq at the 11q13 intergenic region. The top track displays partially annotated transcripts from ENCODE (green ideograms) [277]. A *C*ufflinks assembly of the RNA Capture-seq data separated into (+) and (–) strands reveals multiple transcripts across the locus. The prioritised transcripts are marked by a red box with two isoforms shown for the (+) strand transcript. H3K4me3 enrichment peaks indicate promoter activity in MCF7 breast cancer cells [327]. DNase I hypersensitivity peaks indicate open chromatin in MCF7 cells [3]. CAGE tags in MCF7 cells indicate transcription start sites (TSS) [326]. The blue box encloses PRE1. Data visualised using the UCSC browser.

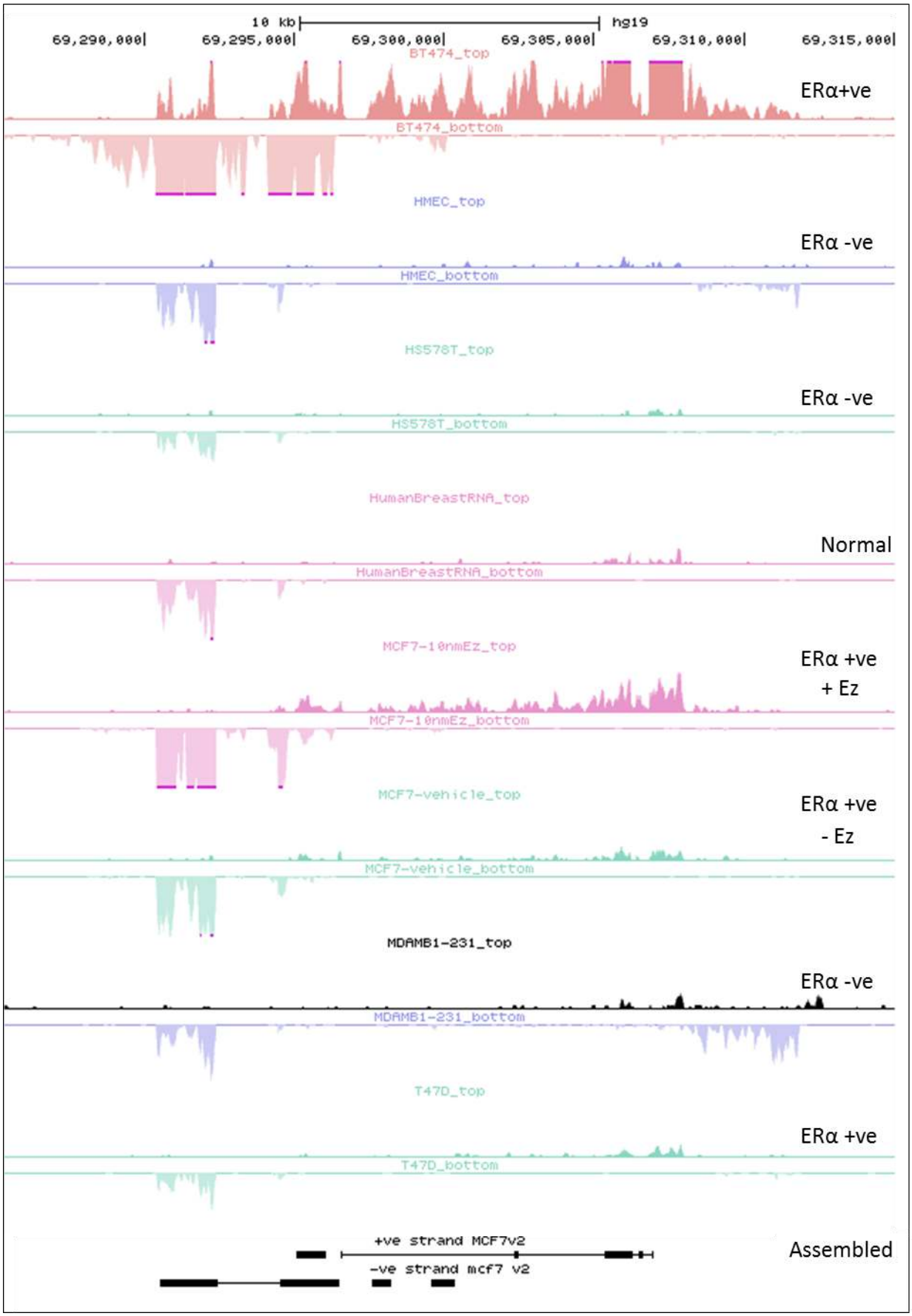

10 kb
hg19
69,290,000
69,295,000
69,300,000
69,305,000
69,310,000
69,315,000
BT474_top
ERα+ve
BT474_bottom
HMEC_top
ERα -ve
HMEC_bottom
HS578T_top
ERα -ve
HS578T_bottom
HumanBreastRNA_top
Normal
HumanBreastRNA_bottom
MCF7-10nmEz_top
ERα +ve
+ Ez
MCF7-10nmEz_bottom
MCF7-vehicle_top
ERα +ve
- Ez
MCF7-vehicle_bottom
MDAMB1-231_top
ERα -ve
MDAMB1-231_bottom
T47D_top
ERα +ve
T47D_bottom
+ve strand MCF7v2
-ve strand mcf7 v2
Assembled

Figure 4.3 RNA Capture-seq reads over the divergent transcripts. Mapped reads for the breast cancer cell lines and normal breast sample examined by RNA Capture-seq. Each cell line has a top track for transcription of the (+) strand and bottom track for (-) transcription. The oestrogen receptor status of the cell lines is indicated on the right; oestrogen receptor positive (ERα+ve) or oestrogen receptor negative (ERα-ve). The MCF7 cells data is divided into oestrogen induced (+Ez) or non-induced (-Ez). The assembled reads are shown on the lowermost track. Data visualised using the UCSC browser.

Interrogation of publically available ChIP-seq datasets suggested that the putative lncRNAs arose from a bidirectional promoter that bound a number of transcription factors (TFs) relevant to breast cancer biology including the oestrogen receptor (ERα)[129], the pioneer factor FoxA1 [133], and the steroid receptor co-activator SRC3 [286] (*Figure 4.4*). H3K4me3 enrichment indicative of an active promoter is seen in MCF7 cells (*Figure 4.4*). *CUPID1* corresponded to a recently discovered isoform in hepatocellular carcinoma that had not been described prior to the commencement of the RNA Capture-seq experiment [329]. *CUPID2* overlapped a recently annotated, but uncharacterised, predicted lncRNA (Gene ID ENSG00000255774) [330].

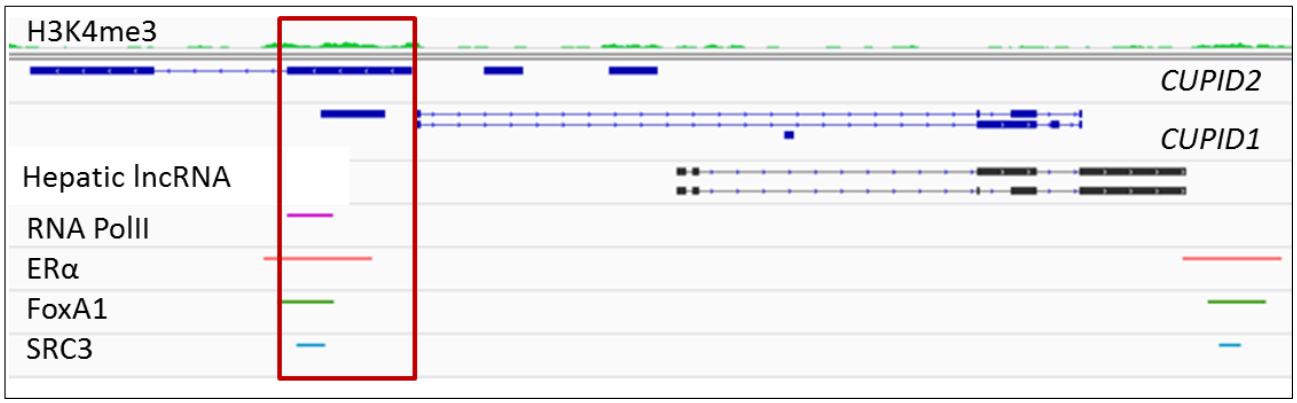


Figure 4.4 Breast associated TFs bind at the putative lncRNA promoter. The assembled RNA Capture-seq reads for MCF7 cells are shown in blue. The putative promoter region is indicated by the red box, showing enrichment for the binding of RNAPolII, ERα [129], FoxA1 [133], and SRC3 [286]. The uppermost green track is H3K4me3 ChIP data from MCF7 cells, with a small peak of enrichment seen overlapping the TF binding sites [331]. The black track depicts isoforms of the hepatocellular lncRNA. Data visualised using the IGV browser.

4.2.2. Determining the transcripts for CUPID1 and CUPID2

Confirmation of the transcript sequences as indicated by the Cufflinks assembly required determination of their 5' origin, 3' terminus and mapping of alternative splicing. Publically available CAGE data was used to obtain a consensus 5' TSS for each transcript (*Figure 4.5A2*) [332]. The CAGE tags for *CUPID2* indicated two similarly utilised TSSs, with the more proximal used to define the 5' end for subsequent cloning (*Figure 4.5A1*). The CAGE tags for *CUPID1* also showed two main TSSs, with the distal, more prevalent site taken as the consensus 5' start (*Figure 4.5A2*). 3'RACE (Rapid Amplification of cDNA Ends) was used to determine the 3' terminus of each transcript with the PCR products run on an agarose gel prior to excision and Sanger sequencing (*Figure 4.5B,C*). The position of the primers used for the 3'RACE PCR are indicated by blue arrows on the isoform schematic (*Figure 4.5D,E*). Two bands were seen in each of the *CUPID1* and *CUPID2* 3'RACE PCRs indicating two main isoforms for each lncRNA. Further PCR was performed to delineate the structure of the isoforms using one primer located adjacent to the predicted 5' end of each transcript and a second primer located adjacent to the 3' end as indicated by the red arrows (*Figure 4.5D,E*) . A number of minor variations were present within each major isoform band and the most abundant chosen as the consensus sequence. A typical poly-A recognition sequence (AAUAAA) was found for each transcript.

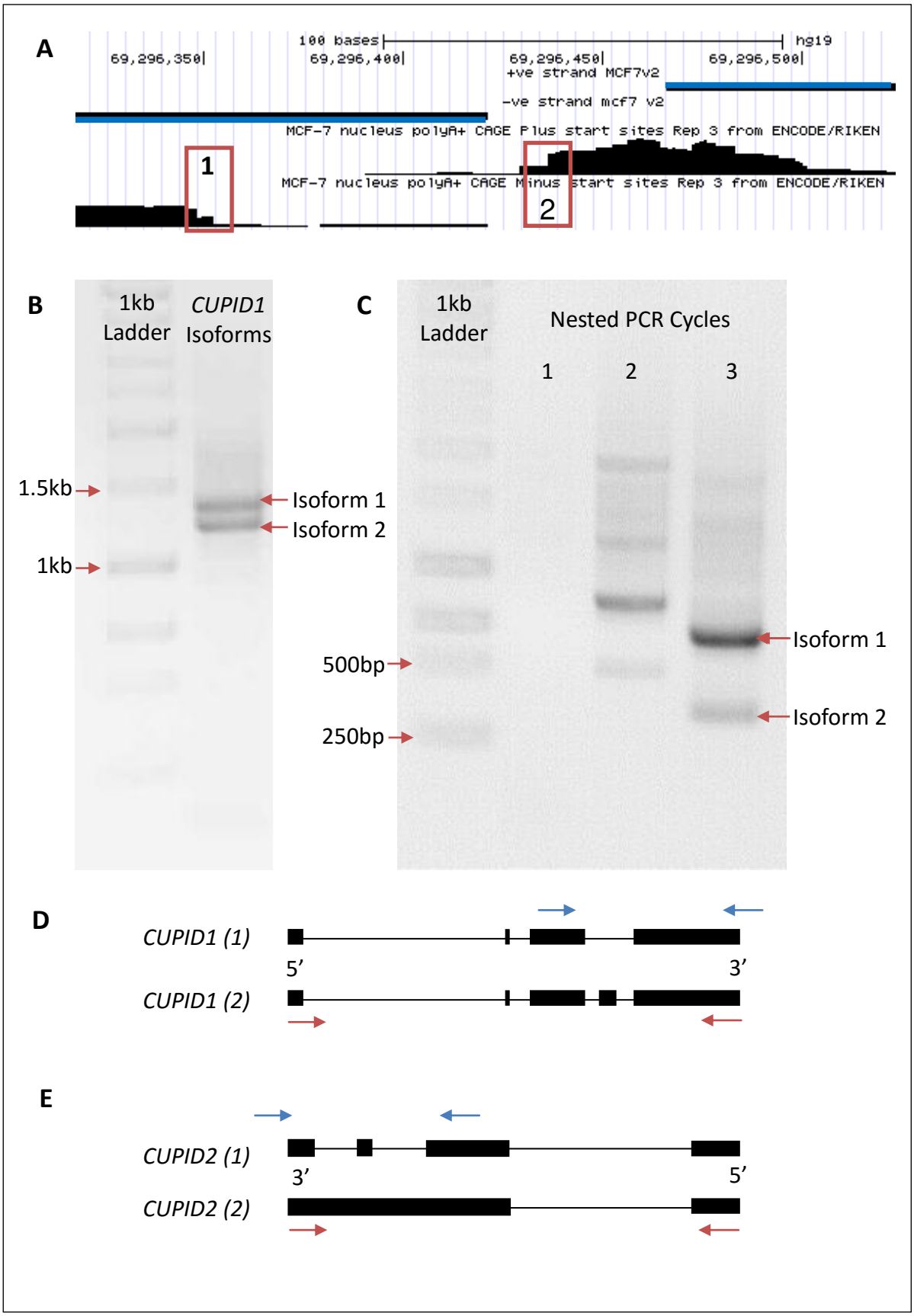

A
100 bases
hg19
69,296,350
69,296,400
69,296,450
69,296,500
+ve strand MCF7v2
-ve strand mcf7 v2
MCF-7 nucleus polyA+ CAGE Plus start sites Rep 3 from ENCODE/RIKEN
MCF-7 nucleus polyA+ CAGE Minus start sites Rep 3 from ENCODE/RIKEN
1
2
B
1kb Ladder
CUPID1 Isoforms
1.5kb
1kb
Isoform 1
Isoform 2
C
1kb Ladder
Nested PCR Cycles
1
2
3
500bp
250bp
Isoform 1
Isoform 2
D
CUPID1 (1)
5'
3'
CUPID1 (2)
E
CUPID2 (1)
3'
5'
CUPID2 (2)

Figure 4.5 Determining the full sequence of the CUPID1 and CUPID2 Isoforms. (A) The 5' consensus TSS was determined using CAGE tags. The cufflinks assembly for *CUPID1* and *CUPID2* are shown by the upper and lower blue/black lines respectively. (1) CAGE tags for *CUPID2.* (2) CAGE tags for *CUPID1*. Data from 2012 RIKEN CAGE release visualized using the UCSC browser [332]. (B) *CUPID1* 3'RACE. Two bands are present on the agarose gel between 1kb and 1.5kb indicating the major isoforms. (C) *CUPID2* 3' RACE. Two cycles of nested PCR were required to isolate the two major isoforms of *CUPID2,* each cycle using progressively more proximal primers against the fixed bait primer. (D) Schematic of the two main isoforms of *CUPID1* (black ideograms) with the lower isoform having an additional exon. (E) *CUPID2* has two main isoforms (black ideograms) with the lower isoform having lost the two terminal introns present in the upper isoform. The blue arrows indicate primer positioning for the 3'RACE. The red arrows indicate primer positioning for PCR across the whole transcript.

4.2.3. CUPID 1 and CUPID 2 are oestrogen regulated.

To determine the distribution of the lncRNAs, a panel of breast cell lines was screened for *CUPID1* and *CUPID2* expression using qPCR (*Figure 4.6*). The transcripts were predominantly expressed in ERα positive cell lines, particularly in those cell lines containing high level amplifications of the 11q13 region such as BT474 and MCF7 cells [284]. *CUPID2* expression in BT474 cells approaches that of the common housekeeping gene *TBP* (TATA-box binding protein). The expression data was supported by the RNA Capture-seq results, with high raw read counts seen in BT474 cells and the oestrogen stimulated MCF7 cells (*Figure 4.3*). The other ERα positive cell line T47D had low levels of *CUPID1* and *CUPID2* transcription, possibly due to its low level of expressed oestrogen receptor (unpublished data from Michael Milevskiy). The ERα positive cell lines and normal human RNA all had minimal evidence of transcription detected by qPCR which is consistent with the RNA Capture-seq data.

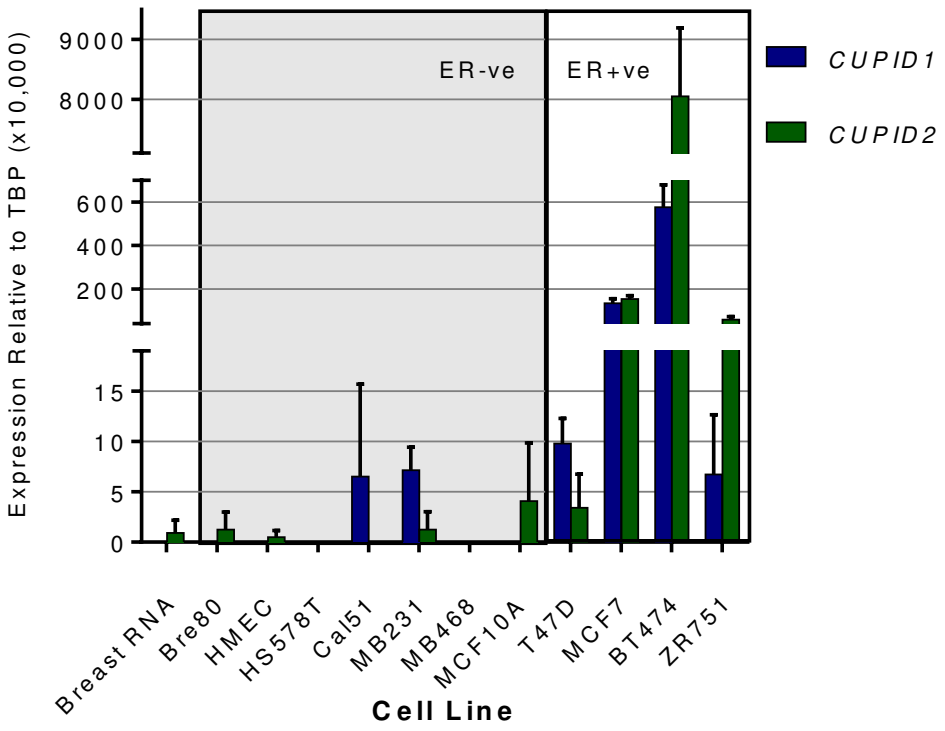


Figure 4.6 CUPID1 and CUPID2 are preferentially expressed in ERα positive breast cell lines. RNA expression levels were determined using qPCR with the results normalised to *TATA box binding protein* (*TBP*). Data shown is mean +/- SD from a single biological replicate. The grey shading indicates ERα negative cell lines. Note the split scale on the y axis.

Additional RNA Capture-seq data was then analysed to determine whether the two novel transcripts were breast specific. Data from other cell lines was provided by our collaborator A/Prof Marcel Dinger [282]. *CUPID1* reads overlapped those from hepatocellular (liver) cell lines though with a breast specific first exon (*Figure 4.7*). There was no expression noted for the colon and prostate cell lines (*Figure 4.7*). Analysis of all 24 cell lines and human tissues included in the capture panel revealed that *CUPID1* was expressed to a significant degree in 2/24 samples and *CUPID2* in 8/24 following RNA Capture-seq. This suggests that *CUPID1* is more tissue specific than *CUPID2*.

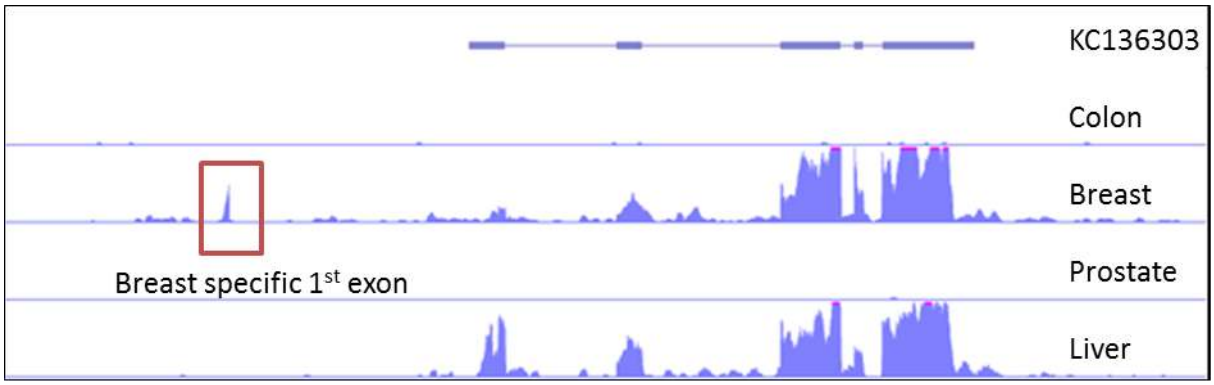


Figure 4.7 – CUPID1 is expressed in breast and liver cells. Raw reads mapped over the *CUPID1* locus following RNA Capture-seq of colon, breast, prostate and hepatocellular carcinoma cells are shown in purple. The top track is an ideogram of the hepatocellular transcript KC136303 from the UCSC browser.

To determine whether *CUPID1* and *CUPID2* were oestrogen regulated as predicted by their preferential expression in ERα positive breast cancer cell lines, qPCR was performed on MCF7 breast cell lines induced with oestrogen. This revealed a robust increase in expression following 6-12 hours of oestrogen exposure for each transcript (*Figure 4.8*). The finding was supported by the RNA Capture-seq data which included RNA from paired MCF7 cell lines grown for 24 hours +/- oestrogen and demonstrated a marked increase in raw read count following oestrogen induction (*Figure 4.3*).

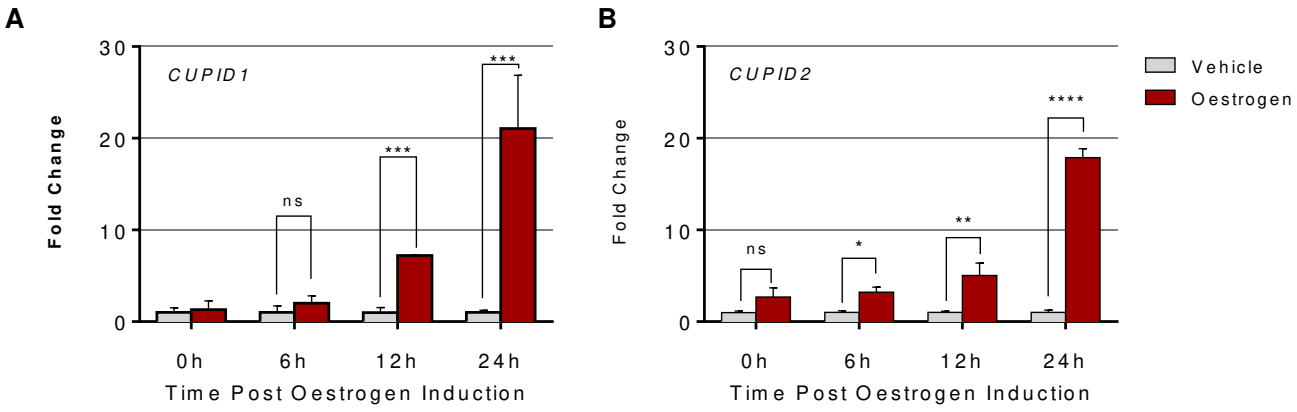


Figure 4.8 CUPID1 and CUPID2 are induced by oestrogen in MCF7 breast cancer cells. Gene expression at each time point is expressed as a fold change over vehicle, with vehicle arbitrarily set at one. *CUPID1* (A) and *CUPID2* (B) expression over a 24 hour oestrogen induction. RNA levels are normalised to *TATA box binding protein* (*TBP*). Data shown is the mean ± SEM from three biological replicates. Significance determined with a 2 way ANOVA test with Dunnett's correction for multiple comparisons. * p<0.05, ** p<0.001, ***p<0.0001, ****p<0.00001.

4.2.4. PRE1 interacts with the putative promoter of CUPID1 and CUPID2.

To determine whether the *CUPID* bidirectional promoter was regulated by PRE1 or PRE2, publically available ChIA-PET data was accessed that defined chromatin interactions mediated by the oestrogen receptor in MCF7 cells [161]. A clear interaction was detected between the *CUPID* promoter and PRE1 but not with PRE2 (*Figure 4.9*). 3C was used to confirm the ChIA-PET findings in the ERα positive breast cell lines BT474 and T47D. This revealed an increased frequency of chromatin contacts between a bait fragment containing PRE1 and distal fragments overlapping the putative lncRNA bidirectional promoter (*Figure 4.10A,B*).

The additional two peaks identified by 3C downstream of the *CUPID* promoter were most likely due to interactions between PRE1 and other distal regulatory elements (*Figure 4.10A,B*). To investigate whether PRE2 participates in chromatin looping in a similar manner to PRE1, 3C was performed in BT474 cells using PRE2 as the bait (*Figure 4.10C*). This showed a far more irregular interaction pattern and peak proximal to the location of the *CUPID* promoter. A significant interaction is not excluded by the 3C data but the absence of a clear interaction peak and the lack of ChIA-PET interactions meant further analysis was confined to PRE1.

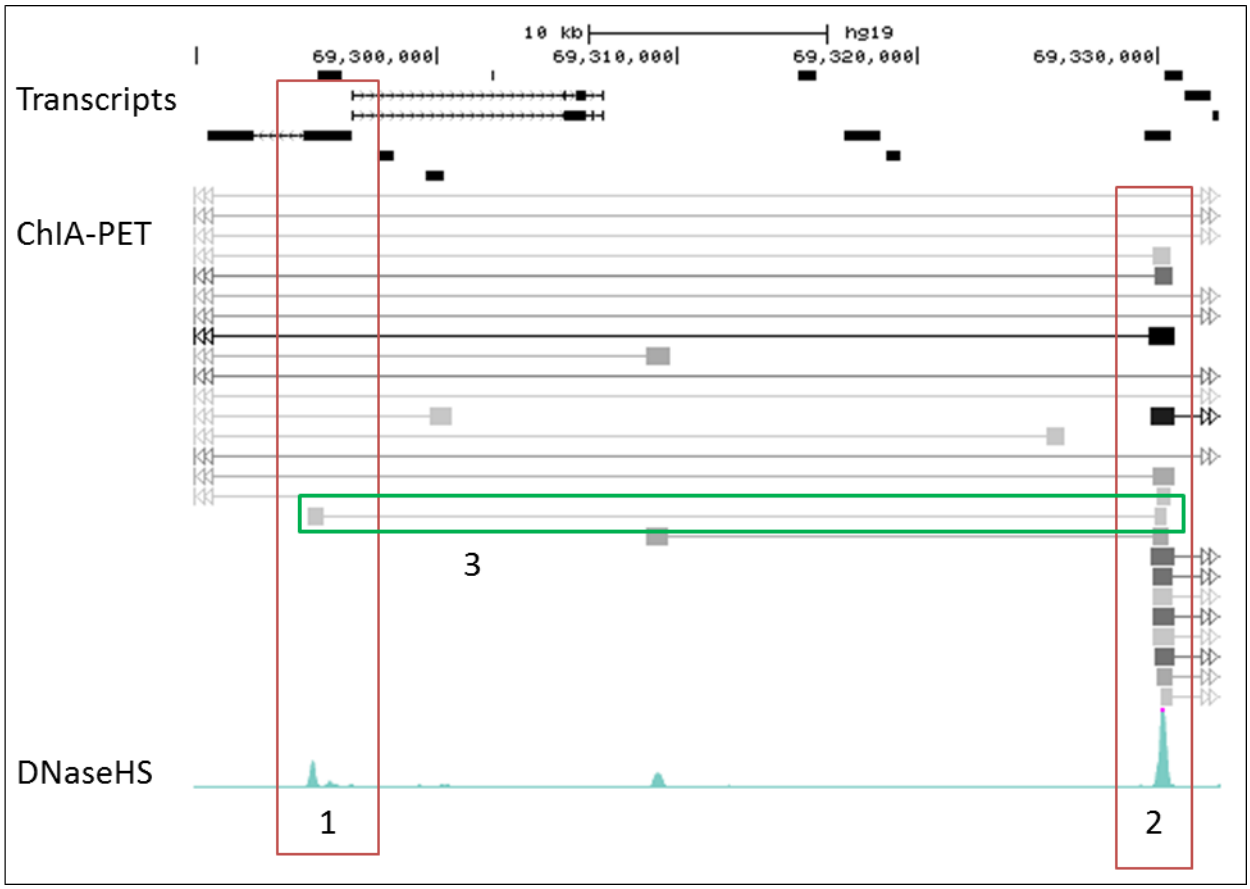


Figure 4.9 Interactions between PRE1 and the CUPID promoter in the ERα mediated ChIA-PET dataset. Box (1) marks the position of the putative bidirectional promoter of *CUPID1* and *CUPID2*. Box (2) denotes PRE1. The horizontal lines indicate chromatin interactions between loci on chromosome 11 [161]. The increasing darkness of the line or box is proportional to the number of interactions. Box (3) encloses a line indicating interactions between PRE1 and the *CUPID* promoter. The RNA Capture-seq transcript assembly is shown in the top tracks (black ideograms). The lower track shows DNase I hypersensitivity peaks indicative of active, open chromatin [275]. Data visualised using the UCSC browser.

3C was then performed in the ERα negative cell line Bre80 and the non-breast ovarian cell line A2780 using libraries made with the *NcoI* enzyme to better localise the interactions and determine whether they were specific to ERα positive cell lines (*Figure 4.11*). This confirmed the interaction between PRE1 and the *CUPID* promoter region. Similar interaction profiles were seen in all cell lines tested, indicating that the interactions were not breast tissue specific and not dependent on the oestrogen receptor.

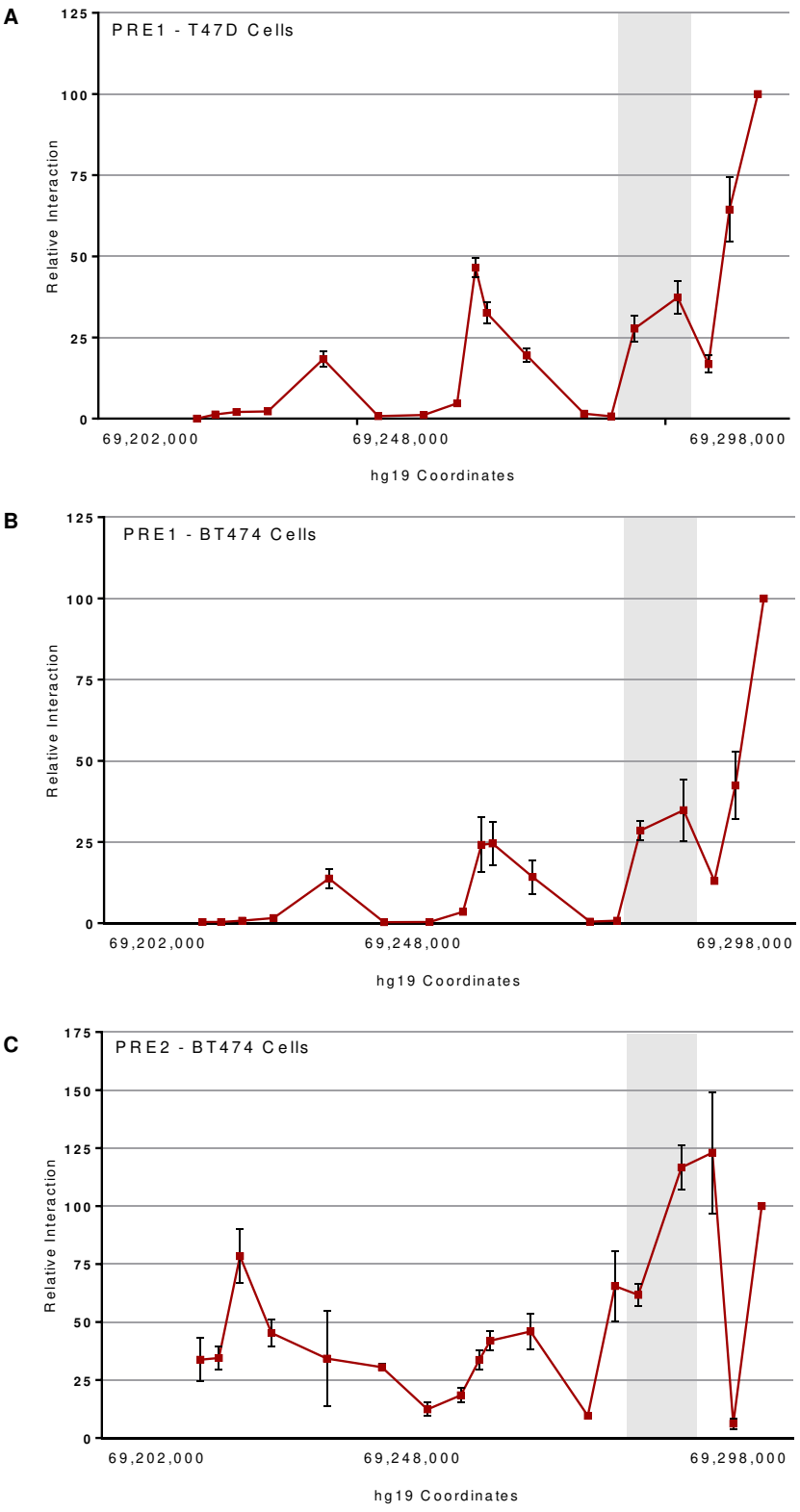


Figure 4.10 3C interactions between PRE1 or PRE2 and the putative lncRNA promoter. 3C interaction profiles between PRE1 and the 11q13 locus in BT474 cells (A) and T47D cells (B); and PRE2 and the 11q13 locus in BT474 cells (C) using libraries made with the restriction enzyme *EcoRI*. The position of the *CUPID* promoter is indicated by the grey box. Data shown is mean +/- SEM for three biological replicates with the results normalised to the most proximal fragment.

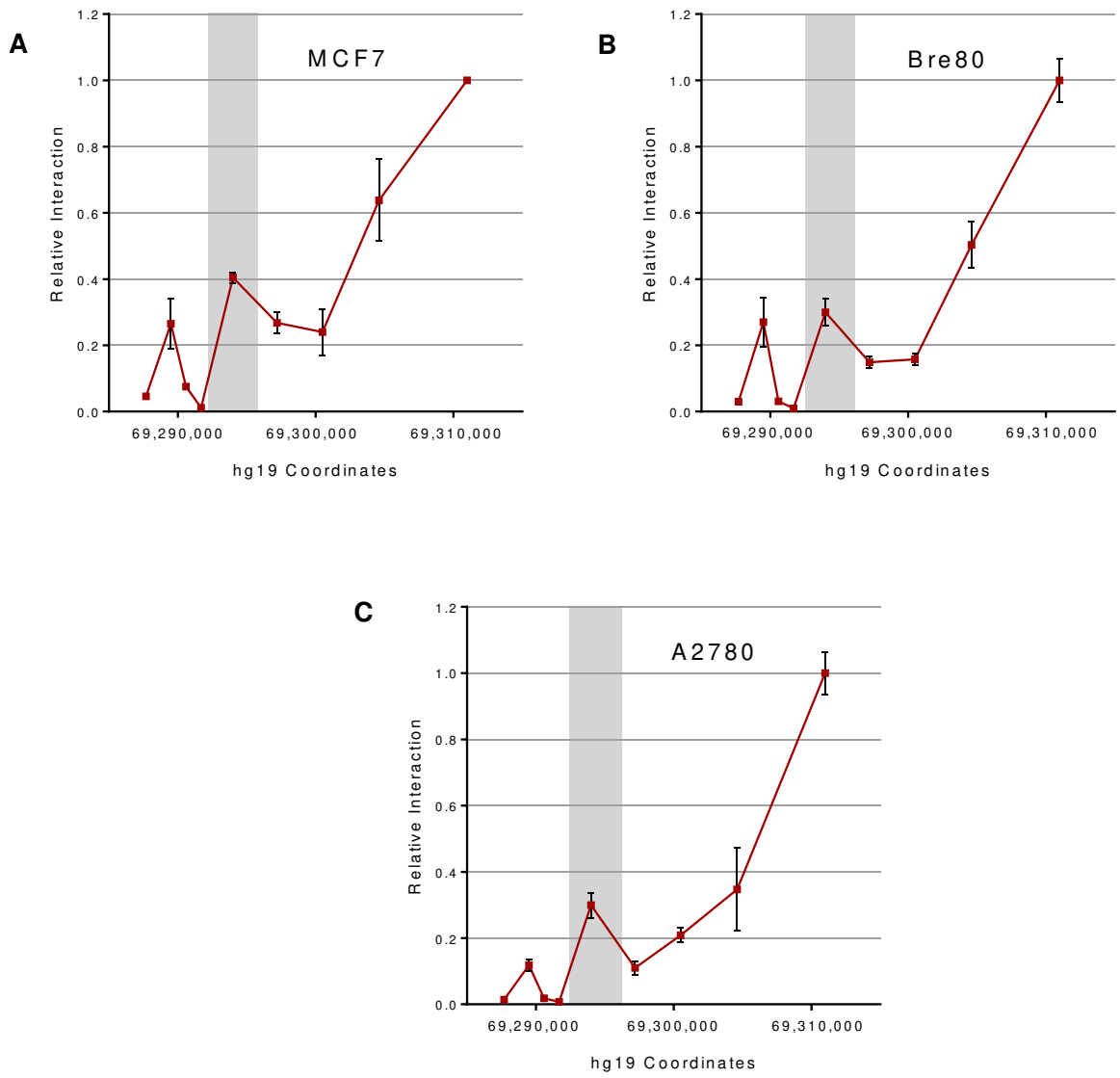


Figure 4.11 3C interactions between PRE1 and the CUPID promoter are present in other cell lines. 3C interaction profiles between PRE1 and the *CUPID1* and *CUPID2* locus in MCF7 (A), Bre80 (B), and A2780 cell lines (C), using libraries made with the restriction enzyme *NcoI*. The position of the *CUPID* promoter is indicated by the grey box. Data shown is mean +/- SEM for three biological replicates with the results normalised to the most proximal fragment.

4.2.5. Chromatin looping between PRE1 and the CUPID promoter is allele specific.

To determine whether the risk SNPs affected looping between PRE1 and the *CUPID* promoter it was first necessary to determine which part of PRE1 participated in the chromatin interactions. This was done by performing 3C using the putative lncRNA promoter region as a bait and analysing interactions across the PRE1 region (*Figure 4.12*). The chromatin interaction was shown to be

maximal over a restriction fragment containing the SNP1 risk allele. This is consistent with ChIA-PET data that also exhibited maximal interaction tags over the same segment of PRE1 (*Figure 4.12*) [161].

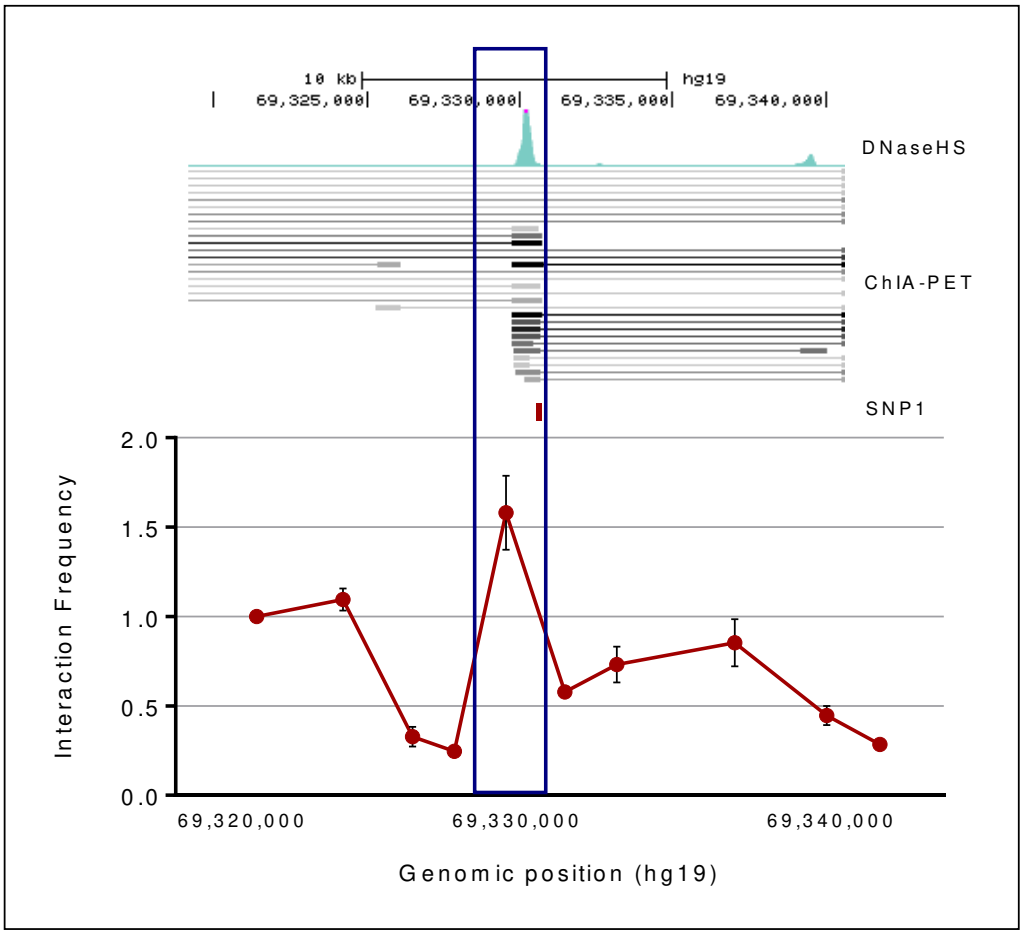


Figure 4.12 The maximal interaction between the CUPID promoter and PRE1 contain SNP1. 3C interaction profiles between the *CUPID* promoter and PRE1 in MCF7 cells using libraries made with the restriction enzyme *NcoI*. The peak interaction is indicated by the blue box and corresponds to a DNase I hypersensitivity peak (aqua) [275], a ChIA-PET interaction hub (grey and black lines) [161], and SNP1 (red line). Data shown is mean +/- SEM for three biological replicates with the results normalised to the most proximal fragment.

To evaluate allele specific looping between PRE1 and the *CUPID* promoter, 3C was performed in the ERα positive MDA-MB-415 cell line which is heterozygous for SNP1. The resulting PCR product was Sanger sequenced and the chromatograms compared with the genomic DNA input (*Figure 4.13*). The relative chromatogram peak heights over SNP1 were biased towards the common allele in the 3C product when compared to the input DNA. This indicated preferential chromatin looping between PRE1 and the putative lncRNA promoter in the presence of the major

(protective) allele and suggested that the minor (risk) allele may reduce chromatin looping and the subsequent transcription of *CUPID1* and *CUPID2*.

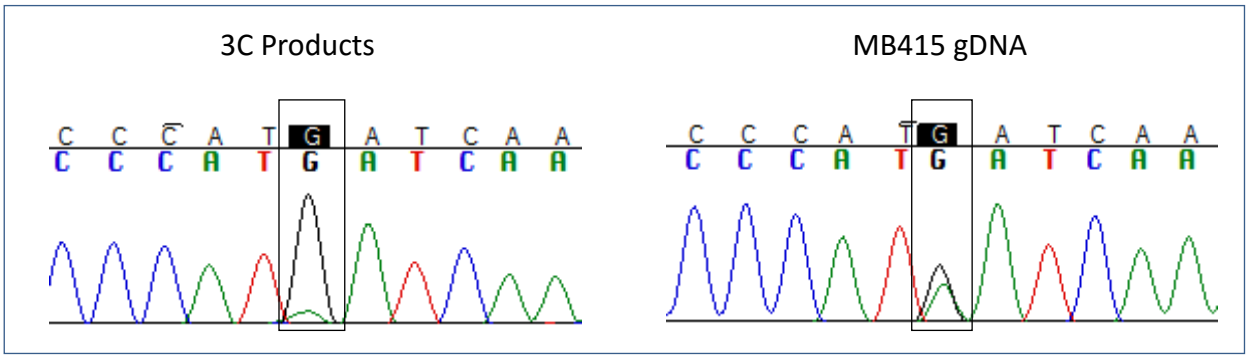


Figure 4.13 The protective allele preferentially participates in chromatin looping. The left chromatogram is derived from Sanger sequencing of 3C products generated from interactions between a PRE1 bait fragment and the lncRNA promoter. The right chromatogram is derived from a corresponding region of input genomic (gDNA). The black boxes highlight the position of SNP1. The common allele is G (black) and the minor (risk) allele is green. Representative chromatograms from three independent biological replicates are presented.

4.2.6. The risk-SNPs reduce activation of the CUPID1 promoter by PRE1.

To determine whether the risk SNPs affected PRE1 activity in addition to chromatin looping, luciferase assays were performed in ERα positive MCF7 and T47D cell. Constructs containing the *CUPID* promoter region (Promoter) in the sense orientation showed it to have a weak activity in both cell lines *(Figure 4.14*). This activity was increased 200-fold in the presence of PRE1 in MCF7 cells and 70-fold in T47D cells. A 25% reduction in enhancer activity was observed when PRE1 contained the minor (risk) alleles of SNP1 and SNP2 in both MCF7 and T47D cells (*Figure 4.14*). Constructs containing the *CUPID* promoter region (Promoter) in the antisense orientation also showed weak activity in both cell lines (*Figure 4.15*). This activity increased 25-100 fold in the presence of PRE1, however the activity of PRE1 was not altered by the presence of minor (risk) alleles of the SNPs.

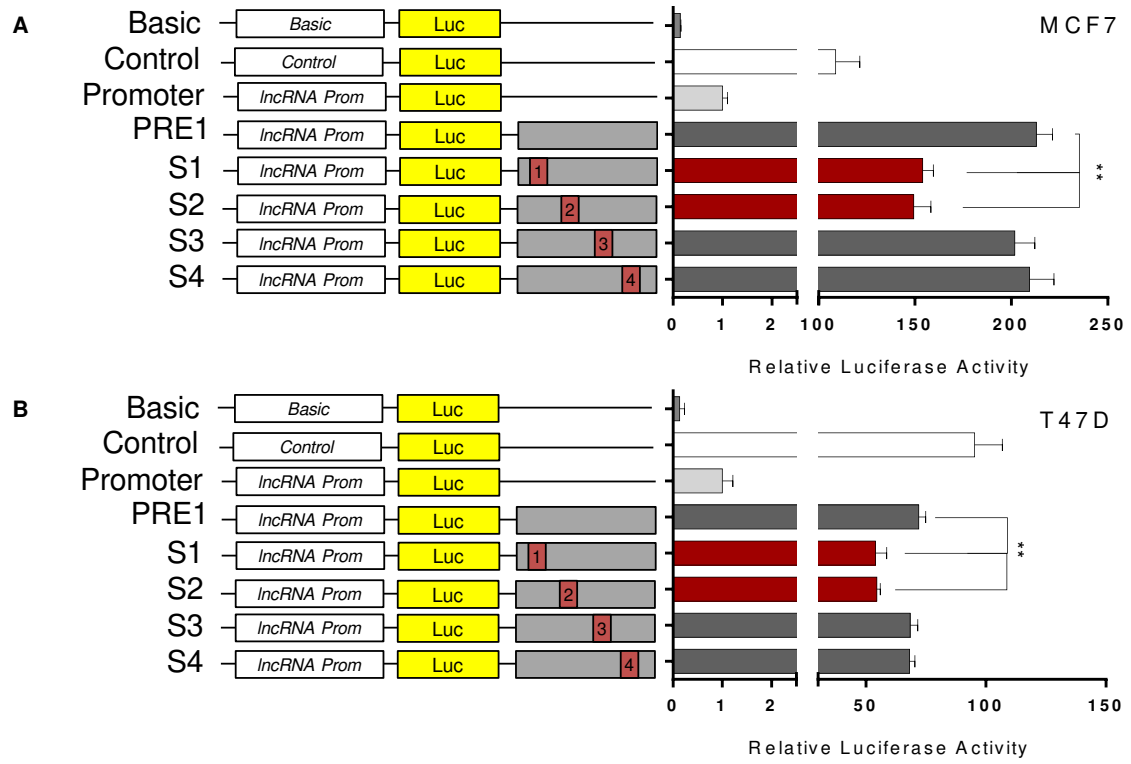


Figure 4.14 SNP1 and SNP2 reduce the ability of PRE1 to activate the CUPID1 promoter. PRE1 (grey boxes) was cloned downstream of a *CUPID1* promoter driven luciferase reporter with and without SNPs1-4. Luciferase activity following a 24 hour transfection in (A) MCF7 cells (ERα positive breast cancer cell line) and (B) T47D cells (ERα positive breast cancer cell line). The luciferase activity of the promoter construct was arbitrarily set as one for each experiment. Data shown is the mean ± SEM from 3 biological replicates. Significance determined using a one way ANOVA incorporating Dunnett's test for multiple comparisons ** $p<0.001$.

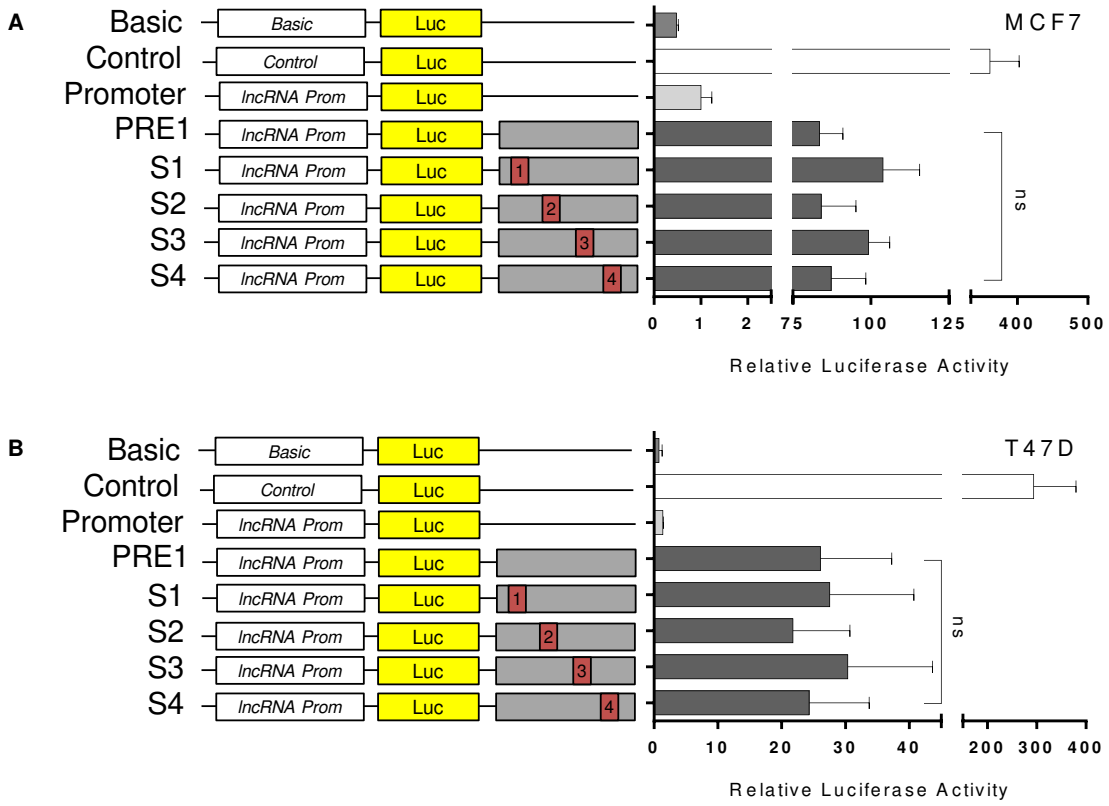


Figure 4.15 The risk SNPs do not affect the ability of PRE1 to activate the CUPID2 promoter. PRE1 (grey boxes) was cloned downstream of a *CUPID2* promoter driven luciferase reporter with and without SNPs1-4. Luciferase activity following a 24 hour transfection in (A) MCF7 cells (ERα positive breast cancer cell line) and (B) T47D cells (ERα positive breast cancer cell line). Luciferase activity of the promoter construct was arbitrarily set as one for each experiment. Data shown is the mean ± SEM from 3 biological replicates. Significance determined using a one way ANOVA incorporating Dunnett's test for multiple comparisons. ns = not significant.

4.2.7. The oestrogen induced activation of CUPID1 and CUPID2 is dependent on PRE1.

To assess whether the SNPs could modulate the response of PRE1 to oestrogen stimulation the luciferase experiments were repeated following a 24 hour oestrogen induction in MCF7 cells. A comparison of relative luciferase activity between the oestrogen induced and non-induced cells showed that neither the *CUPID1* nor the *CUPID2* promoter is directly regulated by oestrogen (*Figure 4.16*). In the presence of PRE1 however, luciferase activity increased 7-9 fold for *CUPID1* and 4-5 fold for *CUPID2* following the oestrogen induction, indicating marked oestrogen responsiveness (*Figure 4.16*). There was no significant change in this fold enhancement in the

presence of the risk SNPs for either promoter indicating that the SNPs do not affect the ability of PRE1 to increase *CUPID1* or *CUPID2* promoter activity following oestrogen stimulation.

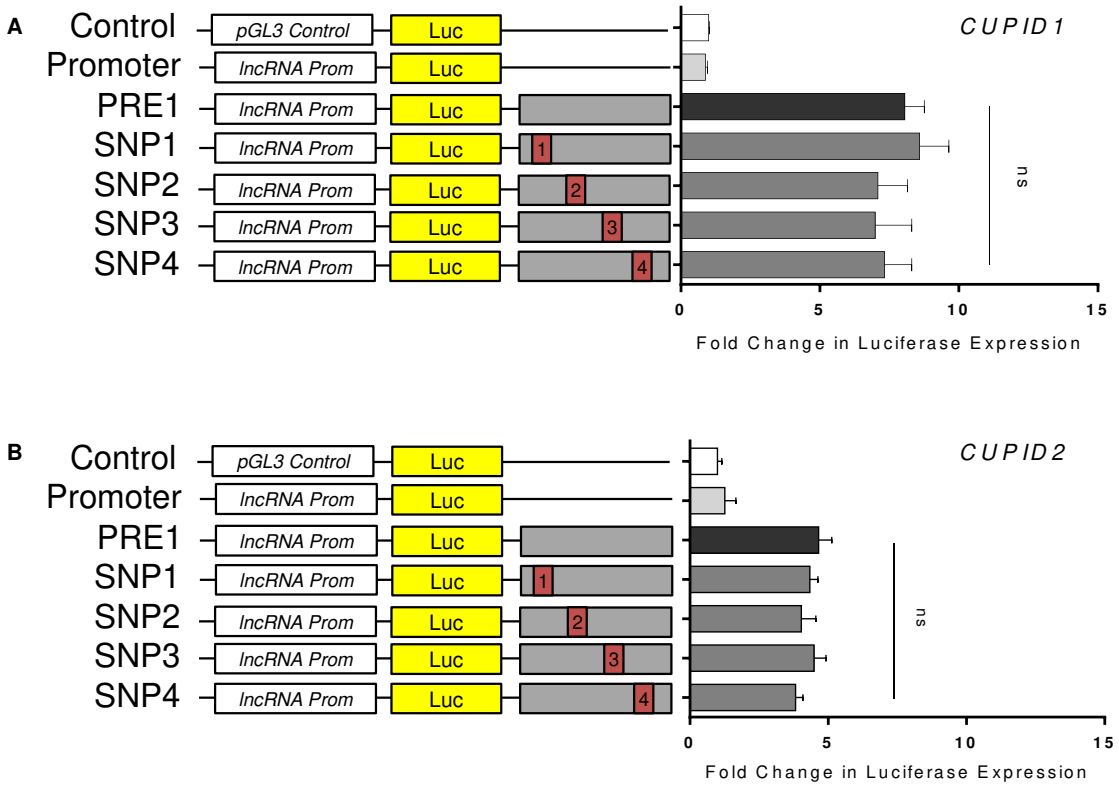


Figure 4.16 The risk SNPs do not alter the response of PRE1 to oestrogen stimulation. PRE1 (grey boxes) was cloned downstream of the *CUPID1* or *CUPID2* promoter driven luciferase reporter with and without SNPs1-4 and transfected into MCF7 cells following an oestrogen induction. For each reporter construct the luciferase value in the oestrogen induced cells was divided by the luciferase activity in the non-induced cells. The results were then normalised to the promoter only construct which was arbitrarily given a value of one. Data shown is the mean ± SEM from 3 biological replicates. Significance determined using a one way ANOVA incorporating Dunnett's test for multiple comparisons. ns = not significant.

4.2.8. CUPID1 associates with chromatin whilst CUPID2 is found in both nuclear and cytoplasmic compartments

To investigate the potential functions of *CUPID1* and *CUPID2*, it was first necessary to determine their subcellular location. MCF7 cells were fractionated into nuclear and cytoplasmic components then the expression of *CUPID1, CUPID2* and four control RNAs was compared between the compartments (*Figure 4.17A*). *CUPID1*was distributed principally in the nucleus whilst the distribution of *CUPID2* was biased towards the cytoplasm. By way of comparison, *B-ACTIN* and *GAPDH* RNA were predominantly found in the cytoplasm as expected for typical mRNAs (*Figure 4.17A*). Further sub-nuclear fractionation showed that *CUPID1* RNA is enriched in the chromatin fractions, either loosely associated (salt fraction), or more tightly bound (isolated chromatin fraction) (*Figure 4.17B*). The chromatin enrichment was greater than that observed for the nuclear lncRNA *MALAT* which is known to associate with chromatin in its role as a splicing cofactor (*Figure 4.17D*). In contrast, *CUPID2* was distributed between the nucleoplasm and chromatin, with a substantial fraction also found in the cytoplasm (*Figure 4.17C*). The mRNA *GAPDH* was distributed across all the compartments examined (*Figure 4.17E).* These results suggest that *CUPID1* interacts directly with chromatin whilst *CUPID2* is likely to have broader functions in both the cytoplasm and nucleus.

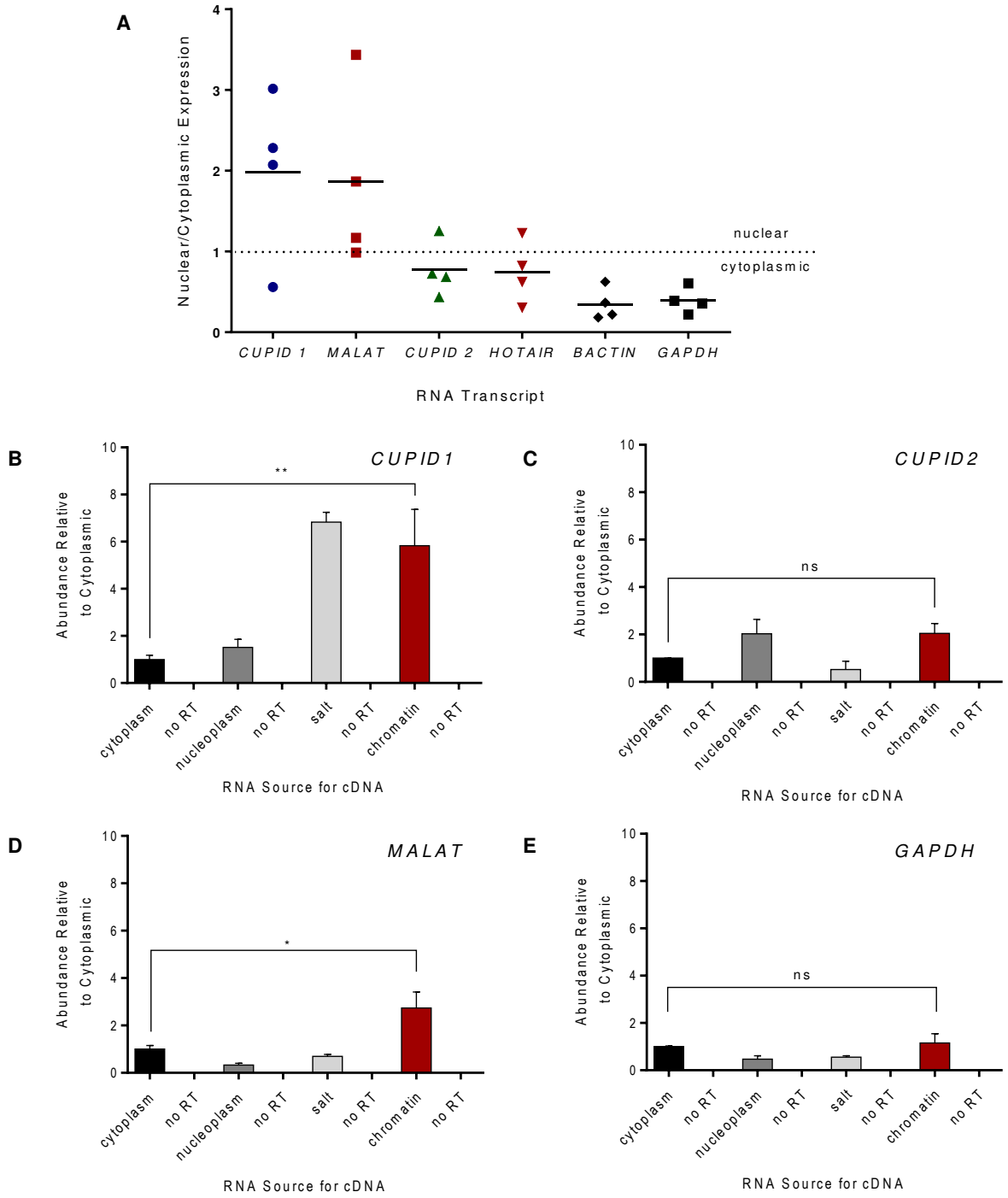


Figure 4.17 CUPID1 binds to chromatin whilst CUPID2 is distributed throughout the cell. (A) The relative abundance of RNA in the nucleus relative to the cytoplasm is shown for *CUPID1, CUPID2*, *MALAT, HOTAIR, GAPDH* and *B-ACTIN*. Expression is shown as a ratio of nuclear RNA abundance/cytoplasmic RNA abundance for each replicate with a value of 1 indicating equal distribution across the compartments. The black line indicates the mean of the 4 biological replicates. (B-E) A biochemical chromatin extraction of MCF7 cells compares the sub-nuclear localisation of *CUPID1, CUPID2, MALAT* and *GAPDH*. Values are expressed relative to the cytoplasmic abundance. RT = reverse transcriptase. Data shown is the mean ± SEM from 3 biological replicates. Significance was determined using a two way ANOVA test. * $p<0.05$, ** $p<0.001$, ns = not significant.

4.2.9. Silencing of CUPID1 and CUPID2 reduces the expression of CCND1.

To determine whether *CUPID1* and *CUPID2* affect the expression of local genes *in cis,* they were silenced using siRNA and gene expression measured using TaqMan assays. The efficiencies of four siRNAs targeting *CUPID1* and three siRNA targeting *CUPID2* were assessed and the best two siRNAs for each used in subsequent experiments (*Figure 4.18A*). Gene expression was assessed using siRNA C1-2 and siRNA C1-3 to silence *CUPID1,* and siRNA C2-3 and siRNA C2-4 to silence *CUPID2*. In addition to the nearby genes *CCND1, MYEOV* and *ORAOV1*, more distal genes (*PPF1A, PPP6R3, CPT1A* and *RNF21*) located >1Mb away were also examined to see whether lncRNA silencing also induced more global changes. *CUPID1* silencing using siRNA C1-2 significantly reduced the expression of *CCND1* and the nearby *ORAOV1* gene suggesting that it may act *in cis* to regulate local gene expression. The *CCND1* silencing achieved using siRNA C1-3 did not reach statistical significance (*Figure 4.18B*). There was a trend, although this was not statistically significant, for reduced *CCND1, CPT1A, MYEOV* and *ORAOV1* expression following silencing of *CUPID2* with both siRNAs (*Figure 4.18C*). Only the silencing of *MYEOV* using siRNA C2-4 reached statistical significance however. Interestingly, all these genes are oestrogen regulated, suggesting that *CUPID2* may play a role in the oestrogen response.

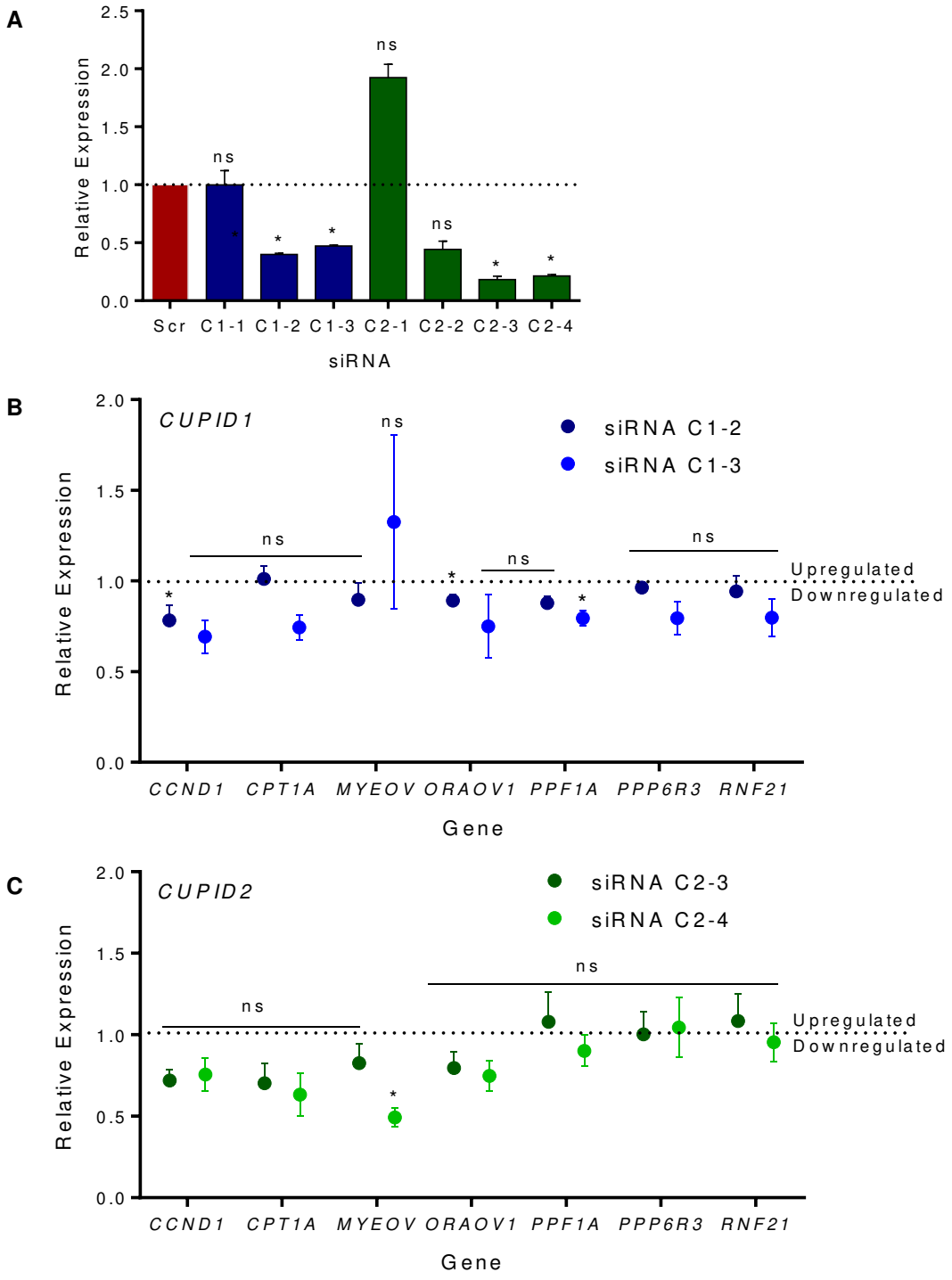


Figure 4.18 Gene expression following siRNA-mediated knockdown of CUPID1 and CUPID2. (A) Silencing efficiencies of the siRNAs targeting *CUPID1* (C1-1, C1-2 and C1-3) and *CUPID2* (C2-1, C2-2 and C2-3). Data shown is mean +/- SD from a single biological replicate. (B) 11q13 gene expression following siRNA C1-2 and C1-3 knockdown of *CUPID1*. (C) 11q13 gene expression following siRNA C2-3 and C2-4 knockdown of *CUPID2*. Data shown is mean +/- SEM relative to a non-targeting control siRNA for three biological replicates. Significance was assessed using a one sample t-test, comparing values to a hypothetical mean of 1.0. * $p<0.05$, ns = not significant.

4.2.10. CUPID1 or *CUPID2 silencing does not significantly alter the magnitude of the oestrogen response.*

To investigate whether *CUPID1* or *CUPID2* affect the oestrogen response as suggested by the changes in gene expression, an oestrogen induction was performed in the presence of *CUPID1* or *CUPID2* siRNA silencing. The degree of oestrogen induction was assessed using expression of the known oestrogen responsive gene *TFF1* [333] (*Figure 4.19*). Target RNA knockdown of >50% per siRNA was confirmed for each replicate. No significant changes in *TFF1* induction was observed in the presence of *CUPID1* or *CUPID2* silencing when compared to a non-targeting control, though both siRNAs targeting *CUPID2* showed a consistent trend for suppressed *TFF1* induction.

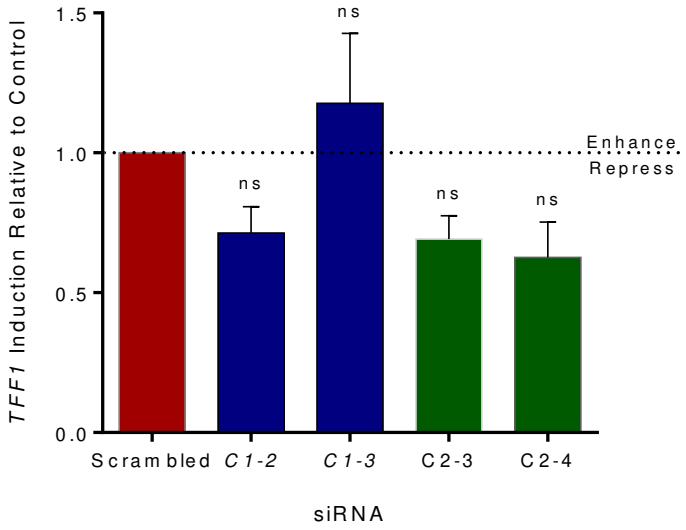


Figure 4.19 CUPID1 or CUPID2 silencing does not alter the oestrogen response. TaqMan assays were used to assess *TFF1* expression before and after a 24 hour oestrogen induction in MCF7 cells. The change in *TFF1* expression is expressed relative to the change observed in the non-targeting control. Blue bars indicate the *CUPID1* knockdowns (siRNA C1-2 and C1-3). Green bars indicate *CUPID2* silencing (siRNA C2-3 and C2-4). Data shown is mean +/- SEM for three biological replicates. Significance calculated using a one sample t test, comparing values to a hypothetical mean of 1.0. ns = not significant.

4.2.11. CUPID1 or CUPID2 silencing does not alter transactivation of the CCND1 promoter.

To investigate whether *CUPID1* or *CUPID2* may be involved in mediating TF interactions at regulatory elements such as the *CCND1* promoter or PRE1, a transactivation assay was performed. Constructs containing multiple oestrogen response elements (EREs) were also included to further

assess the effects of the lncRNAs on the oestrogen response (*Figure 4.20*). Target RNA knockdown of >50% per siRNA was confirmed for each replicate. No significant difference was seen for transactivation following silencing of either *CUPID1* or *CUPID2* when combined with transfection of the *CCND1* promoter +/- PRE1 constructs. This makes it unlikely for *CUPID1* or *CUPID2* to be involved in the assembly of TF complexes at either *CCND1* or PRE1. The ERE transactivation was also not significant due to the marked variation between replicates and thus failed to provide further evidence that *CUPID2* acts on the oestrogen response pathway.

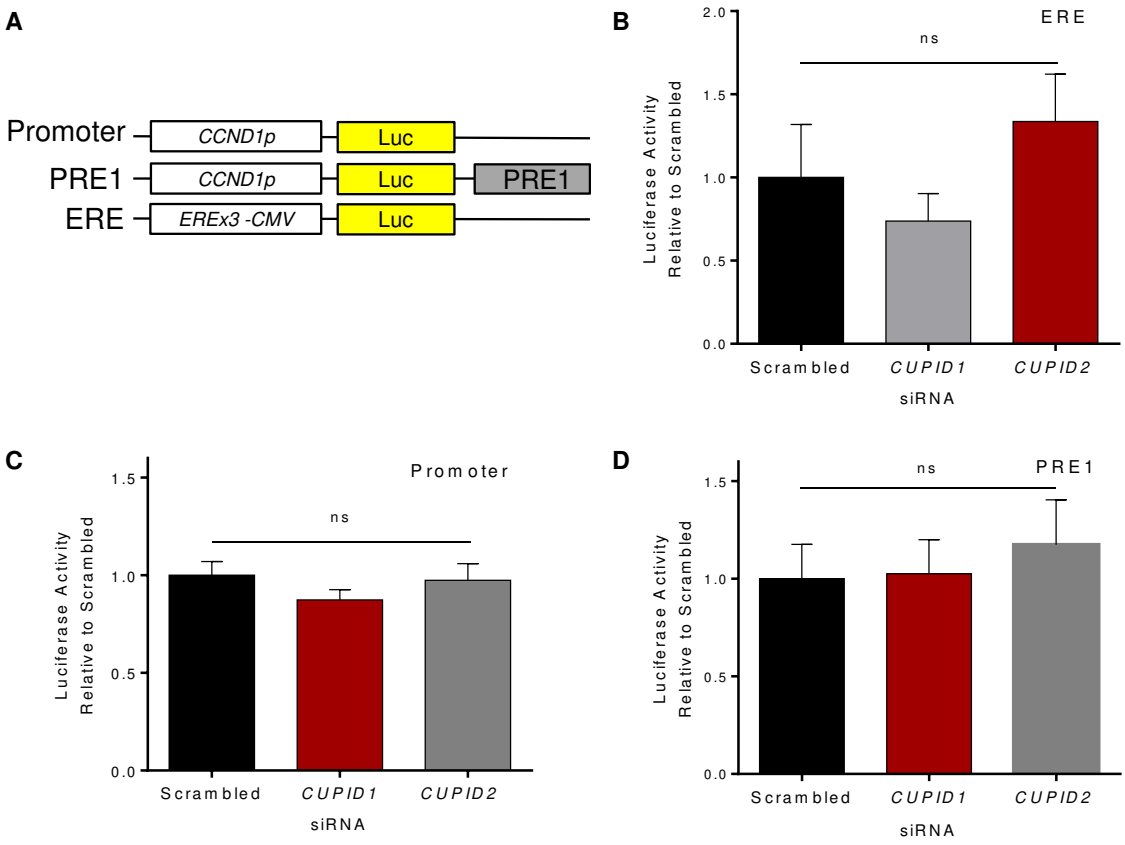


Figure 4.20 Silencing of CUPID1 and CUPID2 does not alter transactivation of the CCND1 promoter. Luciferase activity relative to the non-targeting control is shown for each group following transfection into MCF7 cells. (A) Schematic showing the different pGL3 plasmids used in the transfection. (B) *CCND1* promoter alone. (C) *CCND1* promoter + PRE1. (D) ERE construct with three oestrogen response elements linked to a CMV promoter. Data shown is the mean +/- SEM for three biological replicates. Significance assessed using a one way ANOVA with Dunnett's correction for multiple comparisons. ns = not significant.

4.2.12. CUPID1 or CUPID2 silencing does not affect chromatin looping between PRE1 and the CCND1 promoter.

To determine whether the lncRNAs may have a role in mediating looping between PRE1 and *CCND1*, 3C assays were performed on cells that had been treated with siRNAs targeting *CUPID1* or *CUPID2* (*Figure 4.21*). Target RNA knockdown of >50% per siRNA was confirmed for each replicate. A strong interaction between PRE1 and the *CCND1* promoter was seen for all groups as demonstrated previously [5]. No significant difference was found between the libraries made from the *CUPID1* or *CUPID2* silenced cells compared to a non-targeting control. These results suggest the lncRNAs are not involved in chromatin looping between PRE1 and *CCND1,* however variation between replicates reduced the ability to discriminate subtle differences between the three libraries.

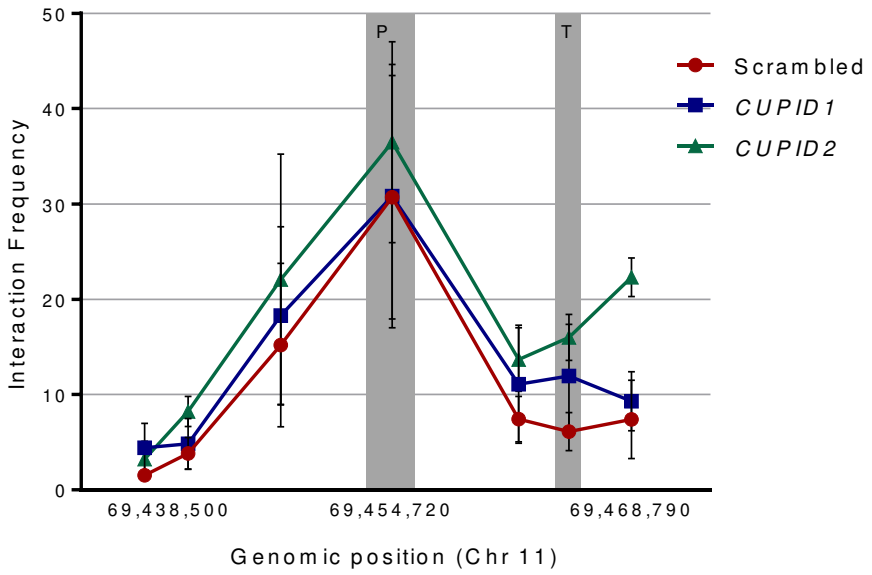


Figure 4.21 CUPID1 or CUPID2 silencing does not alter local chromatin interactions. 3C interaction profiles between PRE1 and the *CCND1* promoter (P) and terminator (T) in MCF7cells using libraries generated with the restriction enzyme *HindIII*. Relative interactions are shown for the non-targeting control group (red) and in libraries with siRNA targeting *CUPID1* (blue) or *CUPID2* (green). Data shown is mean +/- SEM for three biological replicates.

4.2.13. CUPID1 or CUPID2 silencing does not significantly alter progression through the cell cycle.

To determine whether *CUPID1* or *CUPID2* may regulate the cell cycle, they both were silenced using siRNA C1-2 or C2-3 and the proportion of cells in each stage of the cell cycle measured relative to a non-targeting siRNA control. A FACS CANTOII was used to assess the cells with subsequent analysis by the Modfit suite (*Figure 4.22A,B*). Adequate silencing of >50% per siRNA was confirmed for each replicate. *CCND1* levels were measured for all samples using a TaqMan assay and confirmation of cyclinD1 protein knockdown was checked by Western blotting (*Figure 23*). The Western blot was performed by Dr Haran Sivakumaran. A reduction in the proportion of cells found in S and G2 phase was observed for the *CCND1* silenced group with a corresponding increase in the proportion remaining in G1 phase (*Figure 4.22A,B*). This was expected given that cyclinD1 is crucial for cell cycle progression from G1 to S phase [334]. In contrast, no difference was observed in cell cycle following *CUPID1* or *CUPID2* silencing (*Figure 4.22A,B*). Despite differences in the cell cycle, the *CCND1* mRNA levels following siRNA silencing of *CUPID1, CUPID2* or *CCND1* were comparable between the groups (*Figure 4.22C*). However, Western blot analysis showed there was no significant change in the cyclinD1 protein levels within the cell (*Figure 4.23*). This contrasts with the marked reduction in band intensity seen for the silencing of *CCND1*.

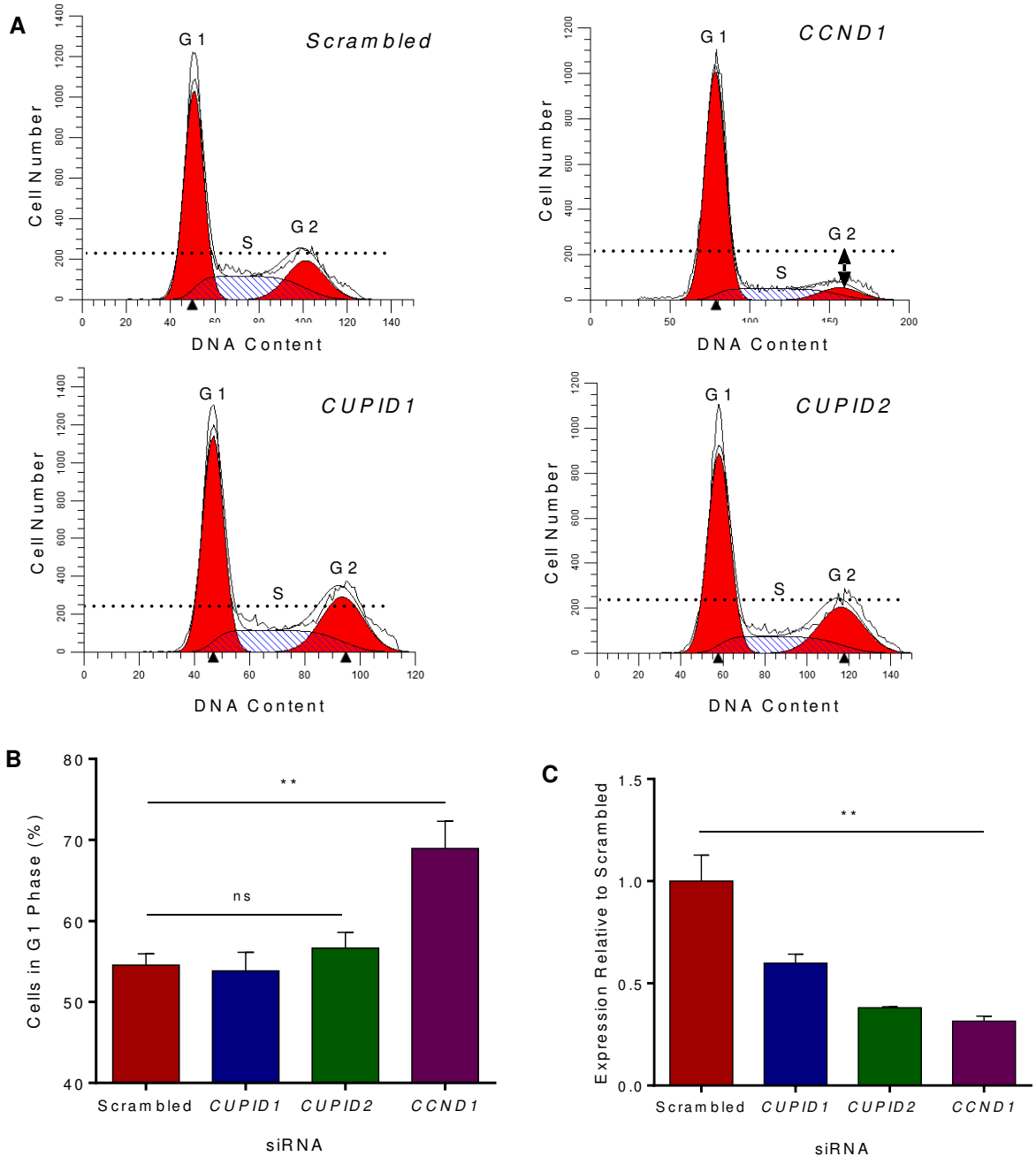


Figure 4.22 CUPID1 and CUPID2 silencing does not alter cell cycle progression. (A) Cell cycle distribution following siRNA silencing of *CCND1, CUPID1, CUPID2* or a non-targeting control. (B) The percentage of cells remaining in G1 phase was assessed for the four groups. (C) The degree of *CCND1* silencing following siRNA knockdown is expressed relative to the non-targeting control. Experiments were performed in MCF7 cells using siRNA-C1-2 for *CUPID1* and siRNA-C2-3 for *CUPID2*. Data shown is the mean ± SEM for three biological replicates. Significance assessed using a one way ANOVA incorporating Dunnett's test for multiple comparisons. ** $p<0.01$.

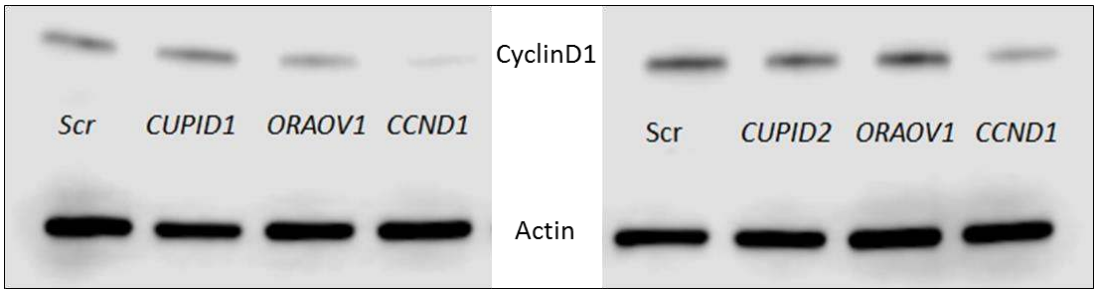


Figure 4.23 CyclinD1 levels are not altered by CUPID1 or CUPID2 silencing. A Western blot of cyclinD1 protein abundance performed in parallel with the siRNA silencing. The cyclinD1 band intensity is compared to the non-targeting (Scr) for each set of four samples. *CUPID1* silenced samples are shown on the left and *CUPID2* samples on the right. Experiments were performed in MCF7 cells using siRNA C1-2 for *CUPID1* and siRNA C2-3 for *CUPID2*.

4.2.14. ChIRP-seq (Chromatin Isolation by RNA Purification) for CUPID1.

CUPID1 was predicted to bind chromatin as shown by the nuclear fractionation assay (*Figure 4.17*). To determine the chromatin binding sites for *CUPID1* genome-wide, the ChIRP-seq technique was performed. ChIRP-seq involves the use of biotin labelled oligonucleotides to pull down a target RNA from a sample of sonicated cross-linked chromatin. Interacting DNA can then be isolated and sequenced to determine where the RNA was located at the time of fixation (*Figure 4.24A*) [261]. A primer pool targeting *CUPID1* RNA was designed and split into even and odd probe sets, with detected interactions only accepted if found in both sets. A pool of *LacZ* primers was used as a negative control. RNA and DNA enrichment compared to an input control that did not undergo the ChIRP process was first confirmed prior to sequencing. DNA enrichment of the *CUPID1* locus was achieved following ChIRP compared with the input sample and negative *LacZ* control (*Figure 4.24B*). The overall RNA yield of *CUPID1* was low (*Figure 4.24C*), reflecting extensive loss of sample through the ChIRP process which involves a number of rigorous wash steps to remove non-hybridised RNA. The Even probe set appears more efficient at pulling down RNA and any associated DNA. Although the total RNA yield is low, the overall DNA and RNA enrichment is acceptable, indicating adequate specificity of the probes and validation of the ChIRP process. The enrichment of *CUPID1* RNA was compared to control mRNAs *TBP* and *GAPDH* and the lncRNA *HOTAIR* via PCR using cDNA made from RNA extracted from the pre and post ChIRP samples (*Figure 4.24D*). Relatively high levels of *CUPID1* RNA were found in both the even and odd probe sets whilst the control RNAs were only found in the pre-ChIRP input sample, indicating a successful pulldown and a specific enrichment for *CUPID1*. A faint band was seen for *CUPID1* in the *LacZ* negative control, however all other reactions were negative as expected.

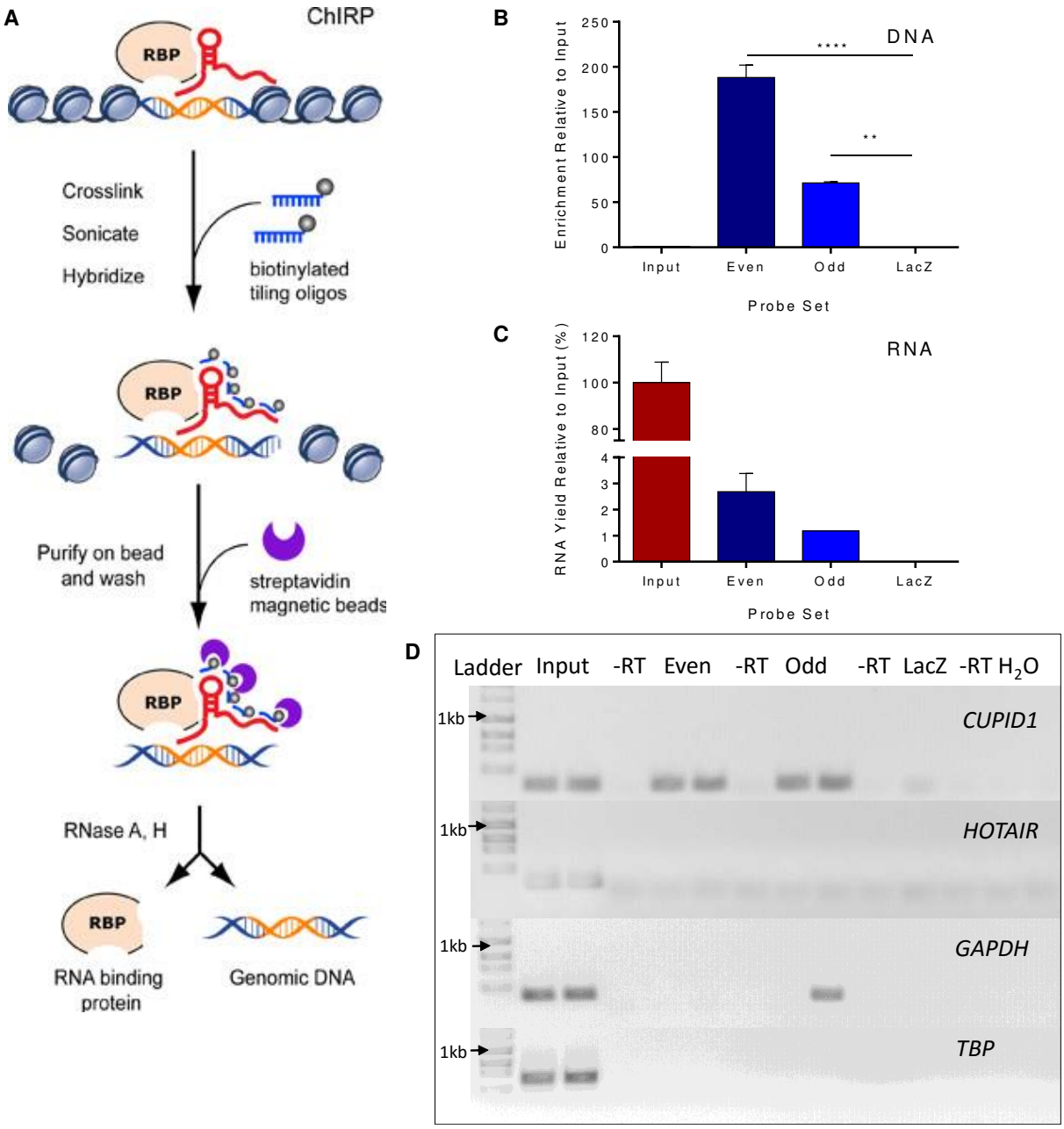


Figure 4.24 DNA and RNA enrichment is confirmed in the ChIRP libraries. (A) A description of the ChIRP-seq process demonstrating a lncRNA (red) binding to DNA with its associated histones (grey spheres) and an RNA binding protein (RBP). The blue oligos bind the lncRNA and are pulled down using streptavidin beads. Following clean up and RNase treatment the purified protein or DNA may then be analysed. Figure taken from [261]. (B) DNA enrichment following ChIRP in the different probe sets. (C) Final RNA yield following the ChIRP process expressed relative to the input sample. (D) RT-PCR products showing specific enrichment of *CUPID1* RNA from two independent probe libraries (Odd and Even). Specific primers were used for *CUPID1*, *HOTAIR*, *GAPDH* and *TBP*. Results are mean +/- SD from the biological replicates that were sent for sequencing. Significance assessed using a one way ANOVA with Dunnett's correction for multiple comparisons. **p<0.01, ****p<0.0001.

The purified post-ChIRP DNA was processed using the Ion Torrent Fragment Library kit. Sequencing was performed using the Ion Torrent Proton platform by Dr Darren Korbie and 11 million reads were generated for the even sample, 19 million for the odd, and 27 million for the *LacZ* –ve control. The ChIRP-seq data was then processed by Dr Mahdi Moradi and the highest confidence genes with a false discovery rate (FDR) > 0.5% are shown below (*Table 4.1*). A larger pool of 60 genes was selected as the input for an Ingenuity Pathway Analysis, using a p value < 0.00001 as a cut off for inclusion. The network analysis revealed a highly significant enrichment for genes involved in DNA replication and repair, cellular assembly, tissue development, cancer and endocrine functioning (*Table 4.2*).

Chromosome	Fold Enrichment	FDR(%)	Gene
chr2	129.77	0.07	*CALM2*
chr20	88.45	0.07	*NRSN2*
chr19	20.18	0.29	*SNAR A3-A14*
chr1	47.28	0.35	*TMEM69*
chr2	64.88	0.35	*PGAP1*
chr2	27.14	0.36	*DUSP28*
chr8	43.59	0.37	*EXOSC4*
chr17	8.67	0.39	*STARD3*
chr1	44.24	0.43	*EPS15*
chr1	26.54	0.45	*LOC100129534*
chr7	47.19	0.45	*CLDN15*
chr7	47.19	0.45	*CLDN15*
chr8	24.39	0.45	*ADRB3*
chr17	35.39	0.45	*CHRNE*
chr17	14.1	0.45	*MED24*
chr7	41.28	0.47	*UPK3B*
chr1	23.59	0.48	*LOC100129924*
chr17	35.39	0.48	*MAP2K3*
chr11	31.97	0.49	*FTH1*
chr6	41.29	0.5	*GNL1*
chr17	28.36	0.5	*LINC00482*

Table 4.1 Gene promoters binding CUPID1. Fold enrichment is a ratio of the read count to a calculated local background. FDR (false discovery rate) indicates the likelihood of the peaks representing a false positive. A FDR of 0.5% has been used for the cut off to remove low probability peaks.

Score	Associated Network Functions
48	DNA Replication, Recombination, and Repair, Energy Production, Nucleic Acid Metabolism
34	Cellular Assembly and Organization, Tissue Development, Cell Morphology
25	Cancer, Developmental Disorder, Endocrine System Disorders
22	Endocrine System Development and Function, Small Molecule Biochemistry, Cell-To-Cell Signaling and Interaction
15	Cellular Development, Metabolic Disease Skeletal and Muscular System Development and Function,

Table 4.2 Gene networks enriched in CUPID1 binding. The score is the $-\log_{10}$(p value) with the initial p value calculated using Fisher's exact test to determine whether the identified ChIRP-seq peaks fall in the associated network by chance.

4.2.15. CUPID1 and CUPID2 silencing impairs Rad51 foci formation in MCF7 cells following irradiation.

Dysregulation of DNA repair is a common feature of genes mediating breast cancer risk [2]. To determine whether *CUPID1* is also involved in the DNA damage repair process as predicted by the ChIRP-seq analysis (*Table 4.2*), immunofluorescence assays for γH2AX and Rad51 foci were performed following silencing of *CUPID1, CUPID2* and *CCND1* in irradiated MCF7 cells (*Figure 4.25*). γH2AX localises to, and is a marker of, double stranded breaks (DSBs) in DNA, whilst Rad51 is a DNA damage repair protein recruited to help repair the DSB [335]. Target RNA knockdown of >50% per siRNA was confirmed for each replicate. The non-targeting siRNA control group displayed a relatively even distribution of γH2AX and Rad51 foci as expected with numerous overlapping foci on the merged window indicating co-localisation of the signal (*Figure 4.25*). Silencing of *CUPID1* (using siRNA C1-2 and C1-3) resulted in widespread γH2AX foci however a number of nuclei lacked associated Rad51 foci, indicating an impairment of the Rad51 recruitment to the DSB. A similar result was also observed following silencing of *CUPID2* (siRNA C2-3 and C2-4). The data from all replicates was collated and the Rad51: γH2AX ratio expressed for each experimental group as shown in *Figure 4.26*. A significant reduction in the ratio of foci was seen following silencing of *CCND1, CUPID1*or *CUPID2* compared to the non-targeting siRNA control.

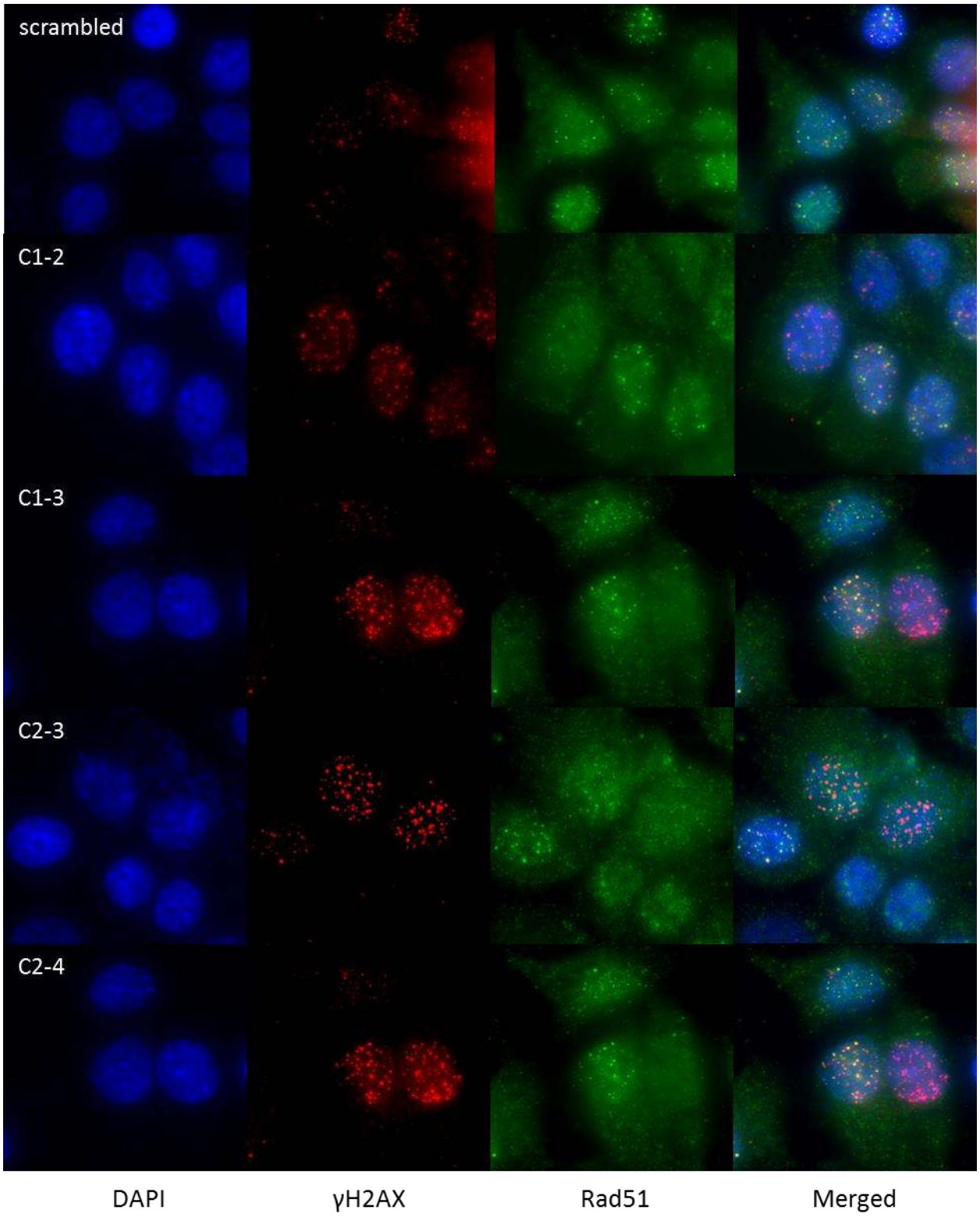


Figure 4.25 Rad51 foci are reduced following siRNA silencing of CUPID1 or CUPID2. Representative images are shown post irradiation and silencing of *CUPID1* (*C1-2 or C1-3), CUPID2 (C2-3 or C2-4*), or a non-targeting control in MCF7 cells. The nuclei are stained with DAPI (blue) in the first column; anti-γH2AX (red) in the second column; anti-Rad51 antibody in the third column; and the images are merged in the fourth column (overlapping foci yellow).

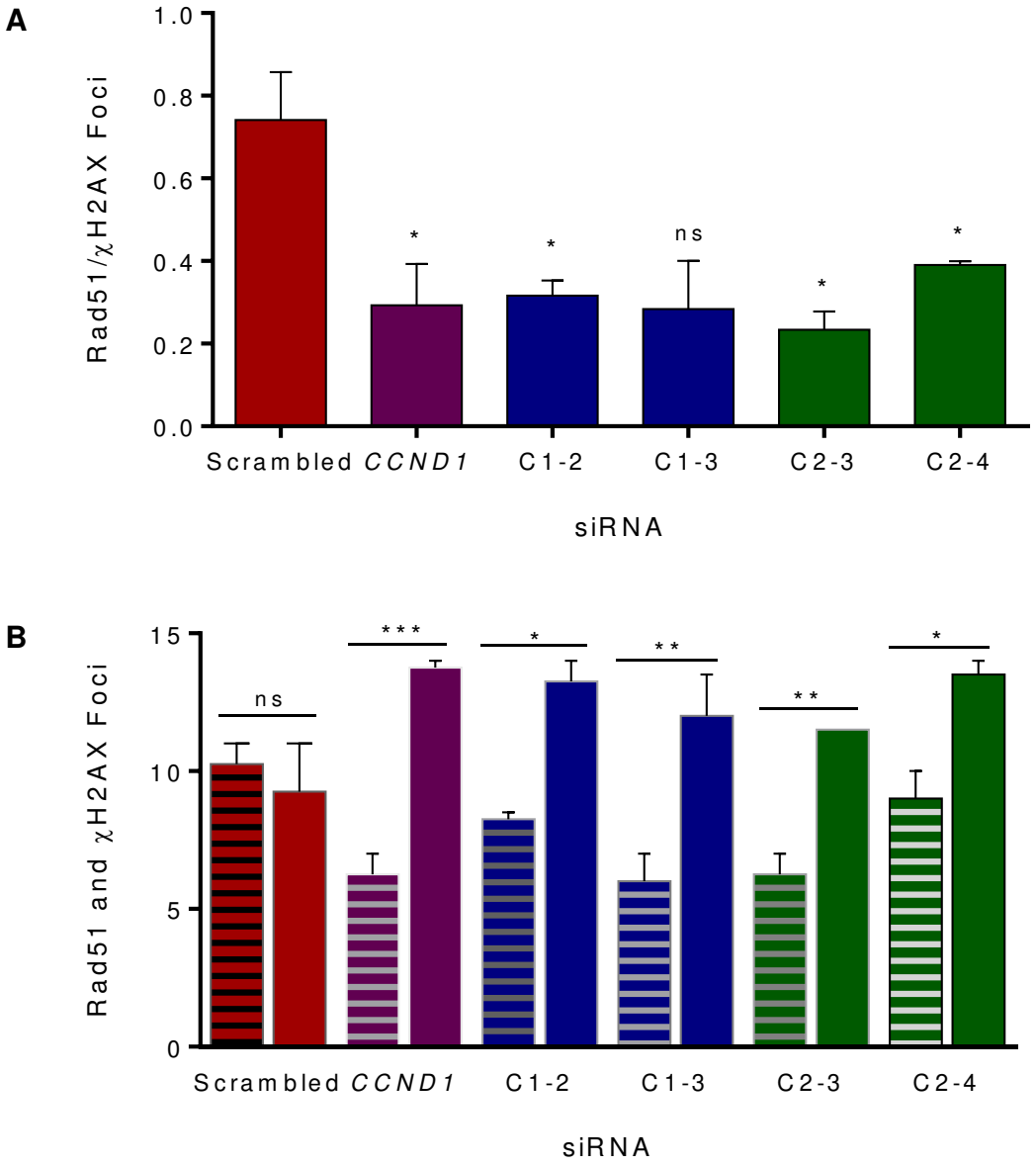


Figure 4.26 The Rad51/γH2AX ratio is reduced following silencing of CUPID1 and CUPID2.

(A) The ratio of Rad51: γH2AX foci number post irradiation in MCF7 cells transfected with siRNA against *CUPID1 (C1-2 and C1-3), CUPID2 (C2-3 and C2-4), CCND1* or a scrambled control. (B) The raw, uncorrected data indicating total Rad51 foci (horizontal hatching) and γH2AX foci (solid colour) counted post irradiation for each siRNA. Data shown is the mean +/- SEM of two biological replicate experiments. Each experiment was performed in duplicate with data taken from 5 randomly chosen groups of 3 nuclei per group. Significance assessed for (A) using a one way ANOVA incorporating Dunnett's test for multiple comparisons and using a two way ANOVA test incorporating Sidak's test for multiple comparisons for (B). * $p<0.05$, ** $p<0.001$, ***$p<0.0001$, ns = not significant.

4.2.16. CUPID1 and CUPID2 silencing augments the effect of Rad51 inhibition to reduce cell survival

To determine whether the effect of *CUPID1* and *CUPID2* in reducing Rad51 foci formation was due to a direct interaction with Rad51 or an effect upstream in the DNA damage pathway, they were silenced in the presence of increasing concentrations of a Rad51 inhibitor. The previously published Rad51 inhibitor BO2 was used and cell proliferation measured using an MTS assay [336]. Cell culture and siRNA knockdowns were performed as previously and the final MTS assay component completed by Dr Adrian Wiegmans. Target RNA knockdown of >50% per siRNA was confirmed for each replicate. All samples exhibited a gradual reduction in cell viability as the Rad51 inhibitor concentration increased (*Figure 4.27A*). Silencing of *CCND1* reduced cell viability slightly compared to the non-targeting control however this did not reach statistical significance. Silencing of *CUPID1* and *CUPID2* knockdowns induced a rapid reduction in cell viability, which was highly significant for siRNA C1-2 and C2-3 (*Figure 4.27B*). This suggests that *CUPID1* and *CUPID2* affect components of the DNA damage pathway upstream from Rad51 and provides further evidence that they do not act through an effect on *CCND1*.

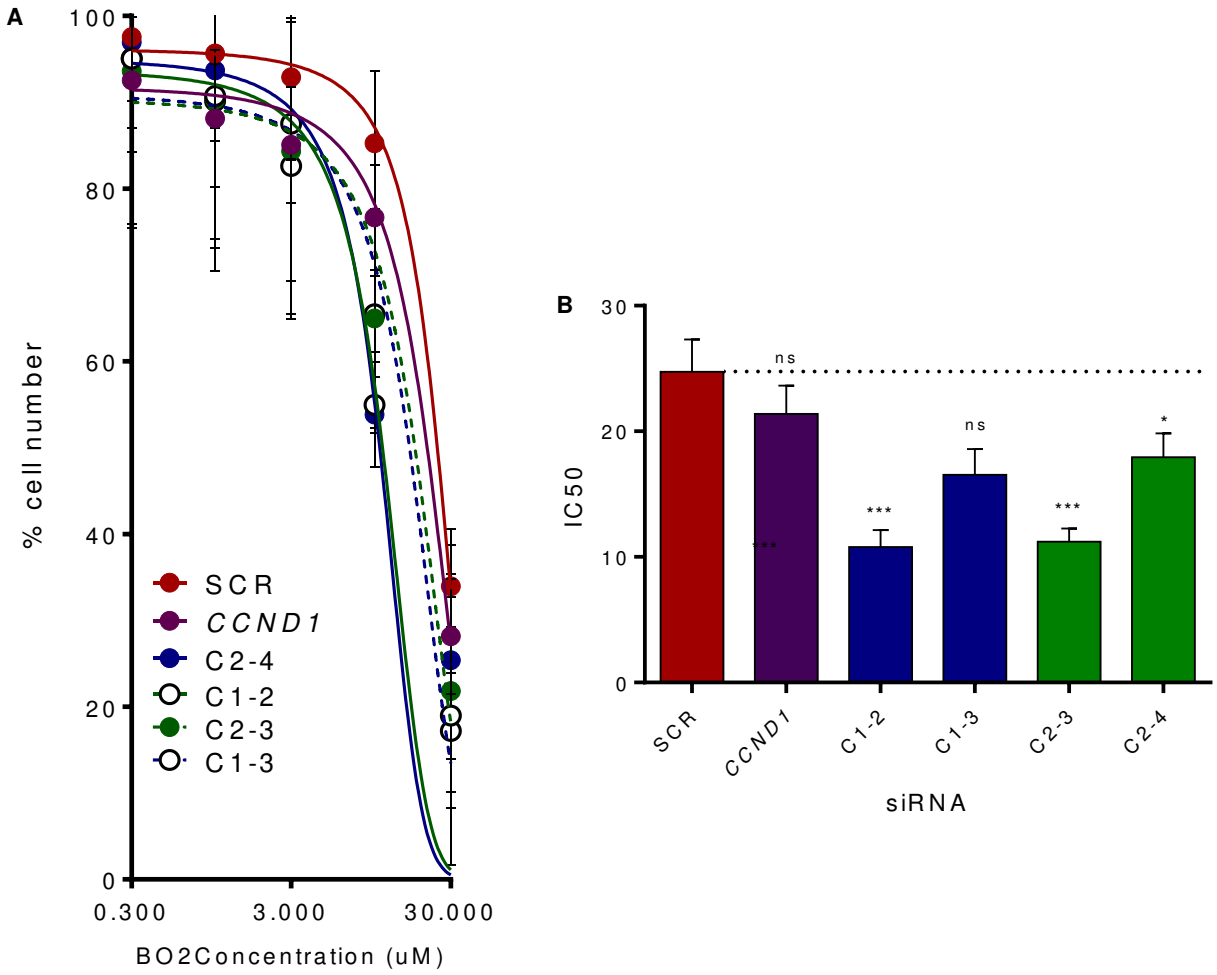


Figure 4.27 Knockdown of CUPID1 and CUPID2 enhances the effect of Rad51 inhibition to reduce cell viability. Survival of MCF7 cells following silencing of *CCND1, CUPID1* or *CUPID2* with increasing concentrations of the Rad51 inhibitor (BO2). (A) The x axis indicates the log of BO2 concentration and the y axis indicates the percentage of cells surviving compared to a BO2 free control group. The samples are compared to a non-targeting siRNA control (red circles and lines). A line of best fit has been drawn for each group. (B) The concentration of BO2 which reduces the cell number by 50% (IC50) is shown for the each siRNA group. The dotted line allows comparison with the scrambled control. Data shown is mean +/- SEM from three biological replicates. Significance determined using a one way ANOVA incorporating Dunnett's test for multiple comparisons. * $p<0.05$, ***$p<0.0001$, ns = not significant.

Despite the pervasive transcription of lncRNAs throughout the genome, they have been poorly represented in post-GWAS studies. This relates to their low abundance and tissue specificity which has historically made them more challenging to identify [219, 220]. In this chapter, RNA Capture-seq was used at the 11q13 breast cancer risk locus to identify novel non coding transcripts for further characterisation. Two lncRNAs were prioritised for functional studies based on their complexity, tissue specificity and oestrogen responsiveness and named *CUPID1* and *CUPID2* due to their demonstrated effect on DNA damage repair. They represent the first lncRNAs to be characterised as potentially mediating risk at a fine-mapped GWAS-identified breast cancer susceptibility locus.

RNA Capture-seq was first proposed as a new approach to post-GWAS follow up studies in 2012 however its use has not yet been described in a published study [7]. In this chapter, RNA Capture-seq was used to identify *CUPID1* and *CUPID2* from the 11q13 gene desert. Prior to the commencement of this work, an uncharacterized lncRNA corresponding to *CUPID2* was present in the GENCODE database (ENST00000542064.1), however there is no published literature regarding its function. *CUPID1* was initially thought to be a novel transcript, however identification of a corresponding lncRNA associated with hepatocellular carcinoma was published by another group in 2014 [329]. The overlap of exons seen between *CUPID1* and the hepatic transcript suggests that they are isoforms of same lncRNA (*KC136297*). The hepatocellular transcript has not been characterized beyond the identification of a putative promoter which is not shared by the breast transcript (*Figure 4.7*). Notably, the *CUPID* promoter is enriched for TFs typical of breast tissue (SRC3, ERα and FoxA1) which is a feature characteristic of cell specific promoters [75, 129, 135, 141]. This tissue specificity was supported by expression studies in various cell lines (*Figure 4.6*) and examination of the additional RNA Capture-seq data [282].

Further evidence for the role of *CUPID1* in breast cancer has come from a GWAS study by Ahsan *et al.* which found a SNP (rs537626) in linkage ($r^2 = 0.41$) with the original 11q13 marking SNP rs614367 defined by Turnbull *et al.* that increased early onset breast cancer risk (odds ratio of 1.29) [4, 337]. It was independently associated with early onset breast cancer as opposed to the late onset breast cancer cohort used for the Turnbull *et al.* GWAS [4]. Rs537626 falls in the terminal exon of *CUPID1* suggesting it may affect post-transcriptional regulation, though no follow up experimental work has been published to date [312]. Such a mechanism of action has precedence in the breast cancer literature with Li *et al*. describing a SNP associated with breast cancer risk in Chinese

women that disrupted the binding site of miRNA-370. This altered the activity of a lncRNA (ENST00000515084) and led to significant changes in tumour growth [321].

CUPID1 and *CUPID2* were initially identified following an assembly of the RNA Capture-seq reads using the Cufflinks program, however such assemblies may not accurately reflect the underlying transcript structure and should ideally be confirmed using alternate methods [260]. Cap analysis of gene expression (CAGE) allows quantification of the most commonly used 5' transcription start sites of a transcript and an extensive database of such information has been compiled by the RIKEN consortium [75]. This resource uses the HeliScopeCAGE sequencing platform which reduces systematic bias in determining the TSS at single nucleotide resolution when compared to other CAGE approaches [338]. RNAseq and RNA Capture-seq are similarly poor at delineating the 3' ends of transcripts, due to the transcript end depletion bias induced by RNA fragmentation [339]. 3'RACE was therefore used on the assumption that the lncRNAs were polyadenylated, which is a common finding for lncRNAs [224, 241]. A typical polyA consensus sequence was found for each transcript, with previous studies finding such sequences present in about 41% of lncRNAs studied [224]. PCR was then performed to delineate the structure of internal exons and introns and found two main isoforms per lncRNA (*Figure 4.5*). More than 25% of lncRNAs have multiple isoforms with an average of 2.3 per locus. The average number of exons is 2.9, with 42% of lncRNAs possessing only two exons [219, 220]. *CUPID1* and *CUPID2* thus have slightly more exons than the average (4-5 for *CUPID1* and 2-4 for *CUPID2*).

The expression data and promoter TF binding suggest that *CUPID1* and *CUPID2* are oestrogen regulated, consistent with their expression at a risk locus specific for ERα positive breast cancer [4]. Oestrogen has long been known to drive the proliferation of ERα positive breast cancer and oestrogen antagonists such as tamoxifen are the main medications used to manage such cancers [340]. This was confirmed following an oestrogen induction in MCF7 cells which revealed a 20-fold increase in expression of both lncRNAs following 24hour of oestrogen exposure (*Figure 4.8*). The induction appears to require the presence of PRE1 however, as oestrogen stimulation of a luciferase construct containing the promoter alone does not increase its activity (*Figure 4.16*). This differs from the lncRNA *HOTAIR* which is also induced by oestrogen in breast tissue, however contains multiple oestrogen response elements (EREs) in its promoter which allow a direct response to oestrogen stimulation [341].

For PRE1 or PRE2 to be activating the *CUPID* promoter *in cis* requires chromatin interactions to be present between them and the target promoter [147]. PRE1 was found to interact with the *CUPID*

promoter in the ERα mediated ChIA-PET data, however the spatial resolution of such methods is not completely reliable and it is important to confirm significant findings using traditional chromosome conformation capture (3C) [161, 342]. 3C assays in the ERα positive cell lines BT474, T47D and MCF7 using PRE1 as the bait showed a consistent, prominent peak at the promoter of *CUPID1* and *CUPID2* (*Figures 4.10, 4.11*). PRE2 appeared to interact with a region just proximal to the *CUPID* promoter with distal interactions decreasing across the promoter region (*Figure 4.10 C*). Previous 3C has shown that PRE1 and PRE2 themselves interact so it is possible that they are all part of an interaction hub together with any genes that they may regulate [86, 161].

Current models of chromatin looping propose that looping contacts are mediated via TFs which recognize specific DNA binding motifs [343]. These motifs may be disrupted by SNPs leading to an reduction in the frequency of enhancer-promoter interactions [344]. Allele specific looping was demonstrated between PRE1 and the *CUPID* promoter with a clear bias for preferential looping involving the common allele (*Figure 4.13*). A reduction of looping contacts in cells containing the risk allele would thus reduce promoter activity and the subsequent transcription of *CUPID1* and *CUPID2* [151]. Such an affect has previously been described at the 2q35 breast cancer locus where a risk allele was associated with decreased looping interactions between an enhancer and the *IGFBP5* gene [60].

SNPs in *cis*-regulatory elements such as PRE1 and PRE2 that disrupt TF binding motifs can also alter the binding of factors required for transcriptional activation of a target promoter [91, 345]. This was assessed using luciferase assays which examined the effect of the SNPs on the ability of PRE1 to activate the bidirectional lncRNA promoter and revealed a 25% reduction in PRE1 activation in the presence of SNP1 and SNP2. The magnitude of this effect is consistent with studies showing that 22% of enhancer SNPs cause more than a 1.2x alteration in enhancer activity, with only 3% altering it by more than 2 fold [346]. The *CUPID1* and *CUPID2* promoters were minimally active by themselves, which is consistent with the low observed expression of the transcripts in ERα negative and non-breast cell lines (*Figure 4.6*). Robust expression is likely to require oestrogen and the presence of the PRE1 enhancer. This is supported by the marked difference in luciferase activity seen following the addition of PRE1 to the *CUPID1* and *CUPID2* promoter constructs in the ERα positive cell lines MCF7 and T47D (*Figure 4.14*). PRE1 mediated oestrogen regulation of promoter activity represents another mechanism by which the SNPs could potentially affect the expression of *CUPID1* and *CUPID2*. Breast cancer associated risk SNPs in an enhancer of *MAP3K1* at 5q11.2 have been previously shown to affect the oestrogen response in luciferase assays [59]. Similar assays incorporating the *CUPID* promoters in the presence of an

oestrogen induction did not demonstrate any effect of the SNPs in this instance, however it did confirm that PRE1 itself was strongly oestrogen responsive (*Figure 4.16*). This is consistent with the high enrichment for ERα seen with ChIP-seq data and suggests that the genes it controls will also be regulated by oestrogen as was demonstrated for *CUPID1* and *CUPID2* [129].

The data thus far supported the hypothesis that *CUPID1* and *CUPID2* have a role in mediating the increase in breast cancer risk however the mechanism behind such an increase still required further exploration. The subcellular location of *CUPID1* and *CUPID2* was first determined by a biochemical fractionation of the cellular and nuclear compartments (*Figure 4.17*). Enrichment of a transcript in a particular compartment provides information as to the function of that transcript and whether it may be classified as an mRNA or lncRNA, with the latter being largely localised to the nucleus [347]. *CUPID1* was bound to chromatin in the nucleus along with the nuclear paraspeckle associated lncRNA *MALAT* [348], whilst *CUPID2* was distributed similarly to *HOTAIR,* a lncRNA known to have both cytoplasmic and nuclear functions [222]. It is not unusual for lncRNAs originating from a divergent promoter to be found in different locations or even to be expressed at different levels, with Cabili *et al.* finding that 7/8 lncRNAs studied using RNA-FISH that arose from such divergent promoters were expressed and localised independently to their associated transcript [347]. The tissue specific expression, nuclear localisation and bioinformatic predictions all support the identification of *CUPID1* and *CUPID2* as non-coding transcripts [220, 347].

It is important to note that determining the subcellular location of a transcript by chromatin fractionation or RNA-FISH does not provide definitive evidence for the function of a particular transcript, however it does give an indication as to what experiments should be prioritised for its further characterisation [347]. These may include ChIRP to define lncRNA-chromatin interactions if the transcript is enriched in the chromatin fraction [261], or ribosome profiling to confirm productive translation if the transcript is enriched in the cytoplasm [349]. Another technique with great promise in exploring lncRNA function is ChIRP-MS, combining ChIRP and mass spectrometry to determine the protein binding partners of an RNA in an agnostic manner [350]. LncRNAs can of course have multiple functions which may or may not include the production of active peptides and activities both inside and outside the nucleus [351, 352]. Transcript copy number can provide additional information in this regard, with a very low transcript abundance consistent with an activity *in cis*, either at the site of transcription or at a target location brought into proximity by chromatin looping [145]. The latter phenomenon is seen for the lncRNA *HOTTIP*, which is estimated to be functional when present at an average of 0.3 copies per cell [353]. RNA function may also vary between cellular compartments and in response to changing conditions

within the cell such as nutrient deprivation or DNA damage [351]. A full characterisation of *CUPID1* and *CUPID2* is beyond the scope of this thesis and will need to be addressed in subsequent studies as has been done for the lncRNA *HOTAIR* [254, 341, 354].

The luciferase and allele specific 3C assays indicated that the risk-associated SNPs may reduce the expression of *CUPID1* and *CUPID2 in vivo.* To recapitulate this effect *in vitro*, siRNAs were designed against the two lncRNAs using the Dharmacon siDesign tool (http://dharmacon.gelifesciences.com/design-center/). SiRNAs have previously been used to reduce the expression of nuclear lncRNAs and it is now known that the factors required for RNAi are indeed present in the nuclear compartment [254, 355-357]. **S**iRNA C1-2 and siRNA C1-3 were chosen for further experiments to silence *CUPID1*, with siRNA C1-3 being less efficient but still producing a consistent 50-70% reduction in *CUPID1* levels (*Figure 4.18A*). siRNA C2-3 and siRNA C2-4 were chosen for *CUPID2* silencing. Interestingly, silencing efficiencies were consistently better for siRNA targeting *CUPID2* compared to those that targeted *CUPID1*, presumably due to the more cytoplasmic distribution of *CUPID2* allowing better access to the silencing machinery (*Figure 4.18A*) [358]. Note that target RNA knockdown of >50% was confirmed for all subsequent experiments involving siRNA.

The effect of siRNA knockdown on local gene expression was first assessed as lncRNAs frequently act *in cis* to modify the expression of genes in the vicinity of their own transcriptional origin [359, 360]. Acting in proximity of their transcriptional locus requires far lower levels of the transcript per cell which is consistent with the extremely low levels of *CUPID1* and *CUPID2* seen in normal breast tissue (*Figure 4.3*) [223]. A reduction in *CCND1* and *ORAOV1* expression was seen following *CUPID1* silencing with siRNA C1-2 but no change was seen in the more distal genes (*Figure 4.18B*). This is consistent with *CUPID1* acting locally on chromatin *in cis.* The failure of siRNA C1-3 silencing to reach statistical significance may relate to the consistently poorer silencing achieved using siRNA C1-3 compared to siRNA C1-2 (*Figure 4.18A*). A broader pattern of reduced gene expression was seen following *CUPID2* silencing with reductions in *CCND1, ORAOV1, MYEOV* and *CPT1A* though none reached statistical significance (*Figure 4.18C*). Interestingly, these four genes are oestrogen responsive, suggesting that *CUPID2* may have a wider effect on ERα regulatory pathways, in keeping with its more even distribution through the cytoplasm and nucleus (*Figure 4.17*). In a similar fashion, *HOTAIR* affects gene transcription genome-wide by recruiting the polycomb repressive complex to gene regulatory elements and also possesses extra-nuclear activities [222, 261]. Further clarification of the effects of *CUPID1* and *CUPID2* silencing is required, and RNA-seq comparing gene expression in MCF7 cells that have been exposed to siRNA

silencing of *CUPID1, CUPID2* or a negative control is underway. This aims to confirm the observed local trends in gene expression and also determine broader effects on biological pathways. The data will not be available before the submission date of this thesis.

LncRNAs have previously been shown to globally regulate androgen receptor target genes by the recruitment of TFs to androgen-response element containing enhancers [361]. The effects of *CUPID1* and *CUPID2* knockdown on the oestrogen response were assessed to see whether they may play a similar role (*Figure 4.19*). The results were inconclusive, however a trend was seen for reduced oestrogen induction with the *CUPID2* siRNA silencing, consistent with the reduced expression of oestrogen regulated genes seen in the TaqMan expression assays (*Figure 4.18*). Many lncRNAs function as molecular scaffolds, enabling protein-protein or protein-DNA interactions and may mediate TF binding to regulatory elements [362]. Disruption of this process may therefore affect enhancer activity as seen with the *E2f2* enhancer following silencing of the lncRNA *PAUPER* [264]. Further experiments directly assessing the effect of lncRNA knockdown on transactivation of the oestrogen responsive enhancer PRE1 or a construct containing multiple oestrogen response elements did not reveal a significant effect (*Figure 4.20*). There was also no effect on the activation of the *CCND1* promoter suggesting that *CUPID1* and *CUPID2* are unlikely to directly affect assembly of the pre-initiation complex.

Another mechanism by which the lncRNAs could affect *CCND1* expression is by altering chromatin looping between PRE1 and the *CCND1* promoter, as certain lncRNAs bind the mediator complex and direct it via chromatin looping to the promoter of their target genes [363]. PRE1 has been shown to significantly increase the activity of the *CCND1* promoter so disruption of looping interactions that facilitate this activation would be expected to have a marked effect on *CCND1* expression [5]. 3C confirmed the interaction between PRE1 and the *CCND1* promoter but did not show a significant effect on this looping following *CUPID1* or *CUPID2* silencing (*Figure 4.21*). Cell cycle assays were then performed to determine whether knockdown of the lncRNAs would affect the cell cycle, either directly or via a reduction in *CCND1* levels. *CCND1* itself is an important regulator of the cell cycle, governing progression from G1 to S phase to drive the proliferation of ERα positive breast cancer [200]. It has been previously shown to be regulated by an RNA binding protein (TLS) which is directed to the promoter region by low copy number non coding RNAs then released in response to DNA damage [237]. *CUPID1* and *CUPID2* knockdown did not however affect progression of the MCF7 cells from G1 through to G2 phase in contrast to the predicted effect seen following *CCND1* silencing (*Figure 4.22B*).

Despite reducing *CCND1* mRNA levels to a similar level to that achieved with direct *CCND1* knockdown, a corresponding reduction in cyclin D1 protein was not seen (*Figures 4.22C and 4.23*). This is not likely to be related to excessive stability of the cyclin D1 protein which has a very short half-life (<1 hour) so a 48 hour knockdown should be more than enough time to see a response [364]. The relationship between *CCND1* mRNA and cyclin D1 protein has been shown to be complex, with experiments involving over-expression of *CCND1* mRNA not giving rise to a corresponding protein increase and post-transcriptional processes appear critical [365]. Off target effects of the *CUPID1* and *CUPID2* siRNAs are possible as such effects may reduce mRNA levels detectable by PCR in the absence of protein changes, however it would be unlikely for both siRNA to be causing the same off target effects [366]. Regardless of the underlying cause, the effects of *CUPID1* and *CUPID2* do not appear to be mediated through an effect on cyclinD1.

CUPID1 was found to primarily interact with chromatin, suggesting that it may influence gene expression by binding regulatory elements or interacting with TFs and chromatin modifying complexes [145]. The targets of a particular transcript may be mapped genome-wide with a variety of novel techniques, which include ChIRPseq (chromatin isolation by RNA purification) [261], CHARTseq (capture hybridization of RNA targets) [367], and RAP (RNA antisense purification) [368]. All three techniques rely on labelled antisense oligonucleotides to pull down a lncRNA of interest which has been chemically linked to adjacent DNA. The DNA can then be purified and next generation sequenced or interrogated using qPCR to find the binding sites of the original lncRNA. These methods have been used to map interactions of the lncRNAs *XIST* [368], *HOTAIR* [261], and *MALAT* [369]. ChIRPseq was used to determine interactions with *CUPID1* in BT474 cells which have the highest expression levels of both lncRNAs. More lowly expressed lncRNAs may require induced overexpression prior to ChIRP to ensure enough RNA is present however care needs to be taken with ectopic over-expression as it may alter the native binding patterns of the transcript [204, 261]. A key feature of ChIRP involves splitting the tiling oligos into two pools, with the final sequencing peaks only considered valid if they are replicated in both pools [261]. Good enrichment was achieved in both even and odd pools compared to the *LacZ* negative control and sequencing identified a number of genes possessing *CUPID1* binding within 1kb of the promoter (*Table 4.1*).

The top ranked genes were *CALM2* and *NRSN2*, both with a FDR of 0.07% and fold enrichments of 129.77 and 88.45 respectively. *CALM2* is a calcium binding protein involved in cell cycle progression, cell motility and proliferation [370]. It is highly expressed in breast cancer cells though its exact role in tumorigenesis is uncertain [371]. *NRSN2* is a neuronal protein of uncertain function that may have roles in hepatocellular and non-small cell lung cancer [372, 373]. The top 60 genes

were analyzed using the Ingenuity program, revealing enrichment of genes involved in the DNA damage response, cellular organization, cancer, endocrine functioning and cellular development (*Table 4.2*). To confirm that the observed binding peaks were functional, *CUPID1* was silenced using siRNA in MCF7 breast cancer cells and RNA extracted for RNA-seq analysis as stated previously. A comparison of the differential expression between control cells and those exposed to *CUPID1* silencing will be made with particular reference to the genes identified by ChIRP-seq. The Ingenuity program will then again be used to analyse broader changes in biological pathways, looking for overlap with the networks enriched following ChIRP-seq. The results will not however be available prior to the submission date of this thesis.

The most significant gene network identified by the ChIRPseq analysis involved DNA damage repair (*Table 4.2*). Interestingly, *CCND1* has recently been shown to be involved in this pathway and genes sharing a regulatory mechanism (PRE1) tend have similar functions [201, 202, 374]. A possible role for *CUPID1* and *CUPID2* in DNA damage repair was investigated firstly using an immunofluorescence assay to assess appropriate Rad51 accumulation at the site of DSBs and then with a survival assay in association with Rad51 inhibition. The immunofluorescence assay revealed a reduction in Rad51 accumulation at the radiation induced DSBs following silencing of *CUPID1, CUPID2* and *CCND1* compared to a non-targeting control (*Figure 4.25*). Previous studies have shown that cyclin D1 is also recruited to DSBs and induces Rad51 expression [202]. As expected, *CCND1* knockdown reduced Rad51 recruitment (*Figure 4.26*). The effects seen with *CUPID1* and *CUPID2* knockdown are likely to be due to different pathways however, given that no effect on cyclin D1 protein levels was demonstrated by the previous Western blot (*Figure 4.24)*. Ingenuity Pathway Analysis (Ingenuity) of the RNAseq data obtained from *CUPID1* silenced breast cancer cells may provide further clarification as to the underlying mechanism behind the observed effect by identifying genes or pathways indirectly regulated by *CUPID1*.

A survival assay demonstrated a significant reduction in breast cancer cell survival when *CUPID1* and *CUPID2* silencing was combined with a potent Rad51 inhibitor, suggesting that both lncRNAs influence additional upstream components of the DNA repair pathway (*Figure 4.27*). Only a minimal effect of *CCND1* silencing was seen, despite the previously demonstrated reduction in Rad51 recruitment, suggesting that there may be compensatory mechanisms involved to buffer the observed phenotype (*Figure 4.26, Figure 4.27*). The role of *CUPID1* and *CUPID2* in DNA damage repair requires further clarification and relevant assays are currently being conducted by our collaborators. These include a GFP reporter assay to assess the efficiency of homologous recombination following *CUPID1* and *CUPID2* silencing [375], and combination knockdowns with

PARP inhibition which would be expected to lead to a markedly increased cytotoxicity were the lncRNAs to indeed have a role in the HR pathway of DNA damage repair [376].

This chapter used RNA Capture-seq and a variety of functional assays to uncover two novel transcripts at the 11q13 breast cancer locus that are affected by the risk SNPs. The transcripts were found to be oestrogen regulated lncRNAs that were preferentially expressed in ERα positive breast cancer cells and named *CUPID1* and *CUPID2* due to their effects on DNA damage repair. SNPs increasing the risk of breast cancer frequently affect genes involved in DNA damage repair pathways [2]. The findings presented here would fit a mechanism in which the risk SNPs reduce the expression levels of *CUPID1* and *CUPID2* leading to an impairment of the DNA damage response and hence an increased risk of breast cancer. Given the pervasive nature of transcription and the proportion of GWAS hits in gene deserts, *CUPID1* and *CUPID2* are likely to represent just the tip of the iceberg for lncRNAs at GWAS-identified loci [219, 243]. The broader application of techniques such as RNA Capture-seq will undoubtedly discover many more such lncRNAs that may then be added to the growing list of genes underlying the genetic causes of breast cancer.

CHAPTER 5

CUPID2 is a putative oncogene that may drive ERα positive breast cancer

In addition to the SNPs at 11q13 conferring an increased risk of developing breast cancer, the locus in which the SNPs lie is heavily amplified in approximately 20% of breast cancers (*Figure 5.1*) [8]. This amplicon contains the cell cycle gene *CCND1* which is one of the best characterised oncogenes in breast cancer and thought to be the major driver promoting focal 11q13 amplification [8]. Tumour amplicons may have more than one driver oncogene however, that function independently or cooperatively within a locus to favour clonal selection [377, 378]. The high number of individual genes that are amplified within the 11q13 amplicon suggests that there may be multiple driver genes present in addition to *CCND1* [14]. These additional drivers may be protein coding genes or non-coding RNA transcripts. The identification of the lncRNAs *CUPID1* and *CUPID2* as novel genes at the 11q13 locus raises the question whether they may also function as oncogenes driving focal amplification and subsequent tumour growth.

LncRNAs have only recently been characterised as a class and are under-represented in studies on cancer despite being present in numbers more than twice that of protein coding genes [241]. With the more comprehensive data now available from recent cancer sequencing projects it seems inevitable that a substantial proportion of the vast numbers of lncRNAs awaiting characterisation will be found to have central roles in cancer biology [379, 380]. Thus far, studies examining the role of non-coding transcripts in human cancer have found an enrichment of lncRNAs at sites exhibiting copy number variation and identified more than 80 lncRNA genes as potential drivers of tumour progression [355, 380]. Functional work on lncRNAs has confirmed their role in key oncogenic processes such as cell migration [254, 381], differentiation [215, 216], cell proliferation [348, 382], and apoptosis [383].

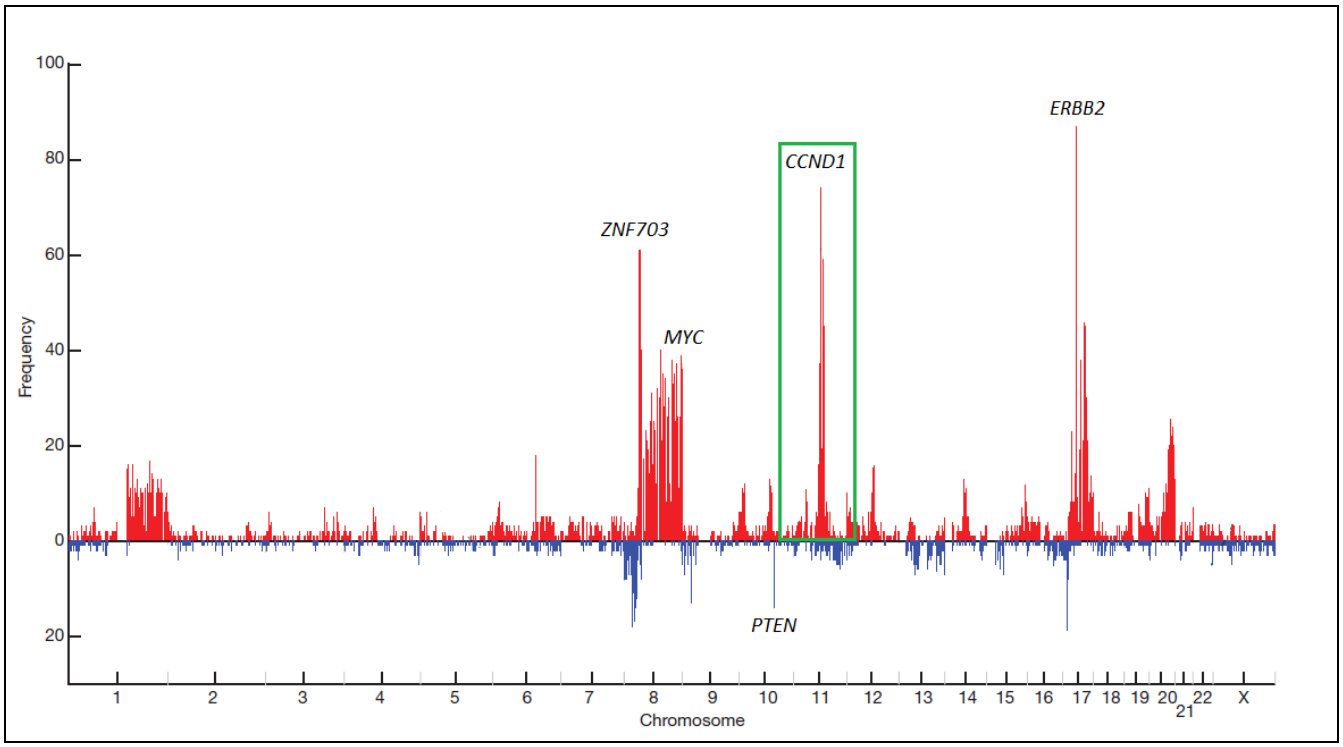


Figure 5.1 Genome wide copy number variations and associated gene expression changes in 2000 human breast cancers. The chromosomal locations are on the x axis with the observed frequency of gene expression at those locations on the y axis. Red indicates high level amplification and blue indicates homozygous deletion. The green box highlights the 11q13 locus where *CUPID1* and *2* are located, along with *CCND1*. The location of significant oncogenes (*ZNF703, MYC, ERBB2*) and the tumour suppressor *PTEN* are indicated. Figure adapted from [14].

The hypothesis of this chapter is that *CUPID1* or *CUPID2* have a role in driving the proliferation of cancer cells and represent novel oncogenes. They may thus be potential therapeutic targets or have a role as biomarkers [241]. Over-expression of *CUPID2* was shown to increase cell proliferation *in vitro* and accelerated tumour growth in a murine xenograft model of human breast cancer. *CUPID2* therefore represents a potential novel oncogene and driver of the 11q13 amplicon.

5.2.1. CUPID2 is overexpressed in human breast cancer.

To determine whether *CUPID1* and *CUPID2* were highly expressed in human breast cancers as expected by their presence in the highly amplified 11q13 region, the mitranscriptome cancer database was searched for equivalent transcripts [241]. This comprehensive database features 7256 curated RNA-seq data sets of human tumours, cell lines and normal tissues principally collated from the TCGA (The Cancer Genome Atlas) [241, 379]. *CUPID2* was shown to be highly expressed in breast cancer (pink) and renal cancer (yellow), with minimal expression in normal tissues (*Figure 5.2*). No transcripts corresponding to *CUPID1* were present in the database.

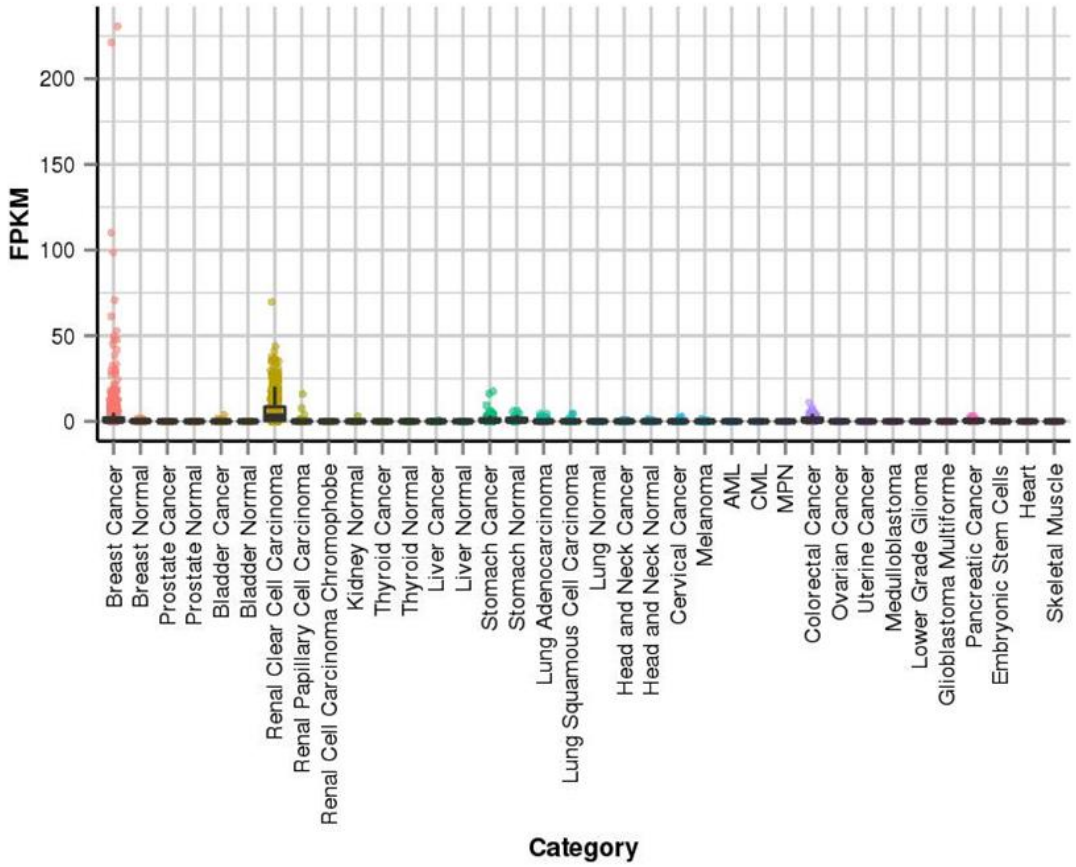


Figure 5.2 CUPID2 is highly expressed in breast and renal cancers. The expression of a transcript overlapping *CUPID2* is shown for multiple tissue types. Breast cancer samples (pink) are found in the first column, adjacent to normal breast tissue for comparison. Each dot represents an individual tumour sample. Expression is measured using fragments per kilobase of exon per million reads (FPKM). Figure generated from the mitranscriptome database [241].

5.2.2. Ectopic expression of CUPID1 and CUPID2.

Given that *CUPID1* and *CUPID2* are over-expressed in breast cancer cell lines (*Figure 4.6*) and *CUPID2* is highly expressed in breast tumours (*Figure 5.2*), an equivalent cell culture model was required to investigate these effects *in vitro.* The dual promoter pCDH plasmid (System Bio) was used to generate over-expression constructs of the identified *CUPID1* and *CUPID2* isoforms (*Figure 5.3A,B*) and these were assessed by a trial transient transfection in MCF7 cells (*Figure 5.3C,D*). *CUPID1* RNA levels were elevated 200x in the construct expressing the first isoform *(CUPID1 (1)*), however there was minimal change after transfection of the second construct *(CUPID1 (2)*). Both *CUPID2* constructs led to increased expression of the lncRNA to 200x vector alone for either isoform. Further cloning and plasmid preparation did not alter the activity of the second *CUPID1* construct and thus the first construct was used for over-expression experiments. The second *CUPID2* construct was chosen for further use as it encoded the more prevalent isoform as indicated by examination of the RNA Capture-seq data.

To determine the optimal time for measuring the effects of *CUPID1* and *CUPID2* over-expression on the 11q13 genes it was necessary to perform a time course of RNA expression following the initial transfection. Expression of *CUPID1* rises rapidly to 100,000X that of the vector after 24 hours whilst the increase for *CUPID2* is more gradual up to a level 3000X that of the vector (*Figure 5.3E,F*). A 6 hour transfection was chosen to measure subsequent gene expression changes with 1000X and 100X expression over baseline for *CUPID1* and *CUPID2* respectively**.** The selected time point produces a similar magnitude of expression change to that achieved in equivalent experiments on the *HOTAIR* lncRNA [254].

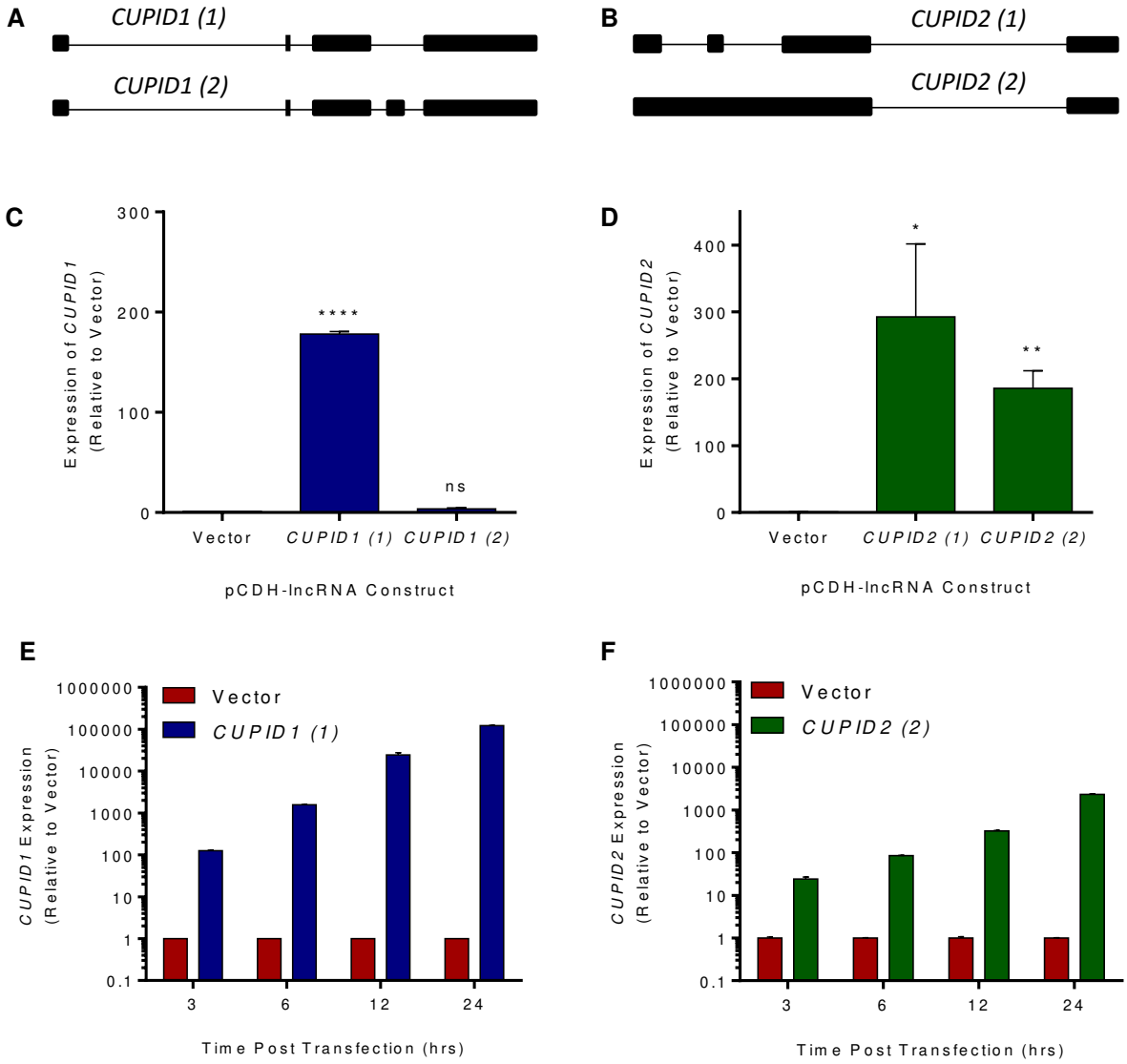


Figure 5.3 CUPID1 and CUPID2 are highly expressed following transient transfections in MCF7 breast cancer cells. (A) A schematic of the *CUPID1* isoforms depicted 5' to 3'. (B) A schematic of the *CUPID2* isoforms depicted 3' to 5'. (C) *CUPID1* isoforms are shown in blue for the trial transfection. (E) *CUPID1 (1) i*soform expression over 24 hours. (D) *CUPID2* isoforms are shown in green for the trial transfection. (F) *CUPID2 (2) i*soform expression over 24 hours. RNA levels were normalized to *TBP* and then expressed relative to the value obtained from transfection of the vector alone. Results are mean +/- SD from a single experiment.

5.2.3. Stable over-expression of CUPID1 and CUPID2 is maintained in a breast cancer cell line.

To assess the oncogenic potential of the lncRNAs in later murine xenograft studies, a more stable over-expression of *CUPID1* and *CUPID2* was required. T47D cells were used as they are an ERα positive cell line with relatively low expression of the two lncRNAs (*Figure 4.6*), thus providing a greater contrast between the wild type cells and those cells engineered to over-express *CUPID1* and *CUPID2*. They have also been previously utilised to create xenograft models of human breast cancer [384-386]. A second generation lentiviral system was employed to transduce T47D cells with infective viral particles and the cells selected through growth in puromycin containing media. The cells were then FACS sorted to select the top 50% of GFP expressing cells (*Figure 5.4C*).

Introduced transgenes are often silenced by epigenetic mechanisms following incorporation into the genome and thus it is important to confirm that transgene expression is sustained before they are used *in vivo* [387]. The transduced cells were therefore cultured under puromycin selection for 3 weeks, FACS sorted and then cultured for a further 3 weeks before assessment of gene expression (*Figure 5.4*). LncRNA expression is depicted relative to *TBP* to illustrate that they are now being expressed at a level similar to that for a common housekeeping gene. *CUPID1* and *CUPID2* RNA levels were elevated to a level 30% and 25% higher respectively than that for *TBP* after 6 weeks, compared to a previously demonstrated baseline expression 1/1000th that of *TBP* (*Figure 4.6*). This confirmed that transgene expression was elevated to high levels for a sufficient length of time to perform xenograft experiments.

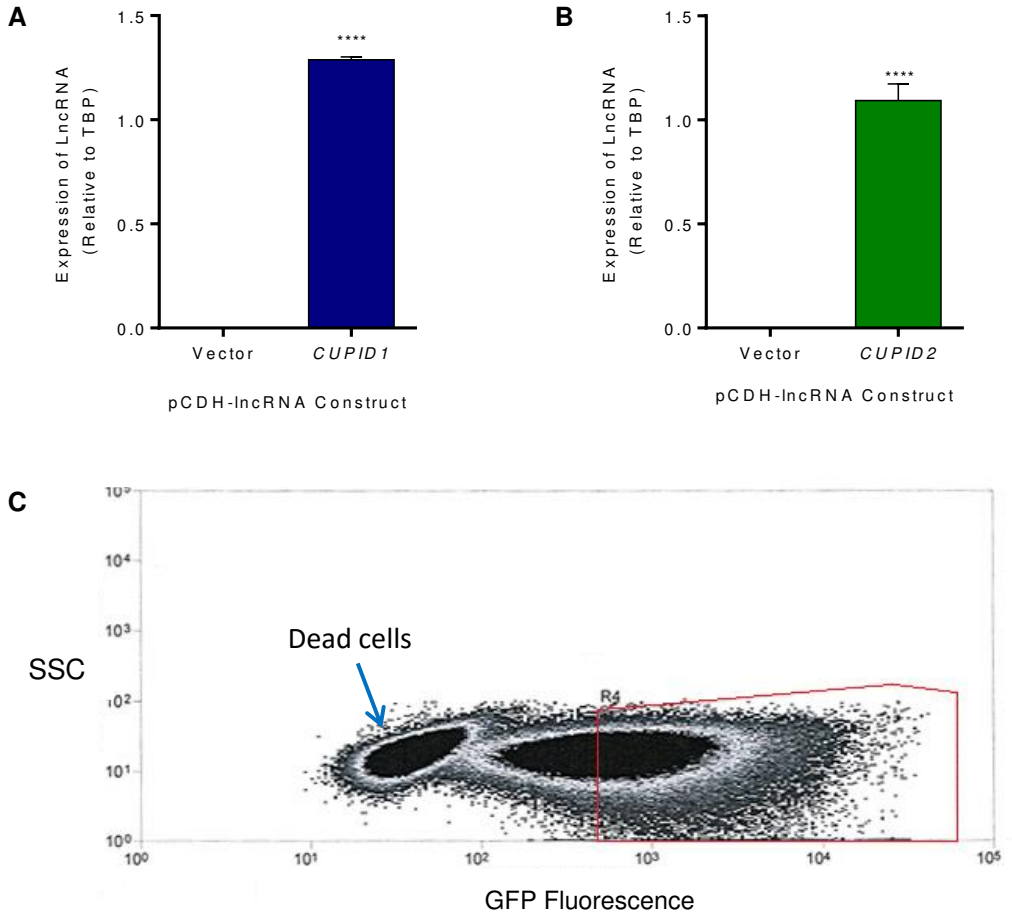


Figure 5.4 CUPID1 and CUPID2 expression is sustained in transduced T47D cells The lncRNA expression is shown relative to *TBP.* (A) *CUPID1* RNA expression compared to the empty vector control. (B) *CUPID2* RNA expression compared to the empty vector control. (C) FACS plot of the cell sorting procedure. GFP (Green Fluorescent Protein) activity along the x axis was used as a proxy for the lncRNA expression and the cells gated such that only the top 50% of GFP expressing cells were retained as indicated by the red box. SSC (side scatter) is on the y axis.

5.2.4. CUPID1 and CUPID2 over-expression does not alter the expression of local genes

LncRNAs frequently act *in cis* to regulate the expression of genes in the vicinity of their own transcriptional loci [360, 388]. To determine whether *CUPID1* and *CUPID2* may act by increasing the expression of nearby genes, TaqMan assays were used to measure 11q13 gene expression following the over-expression of *CUPID1* and *CUPID2*. This was done both transiently and stably using a lentiviral expression system in the ERα positive cell line T47D. The expression of six genes was assessed, with both proximal genes (*CCND1* and *ORAOV1*), and more distal 11q13 genes (*CPT1A, PPF1A, PPP6R3* and *RNF21),* included to determine whether the lncRNAs may be affecting gene expression locally or outside the immediate *cis*

environment. Cells with differing ERα status were used for the transient transfection (ERα positive MCF7 cells or ERα negative MDA MB231 cells) to see whether the hormonal context may influence the effect of over-expression. No significant changes in gene expression were seen for either transient or stable over-expression of *CUPID1* and *CUPID2* in any of the cell lines examined (*Figure 5.5*).

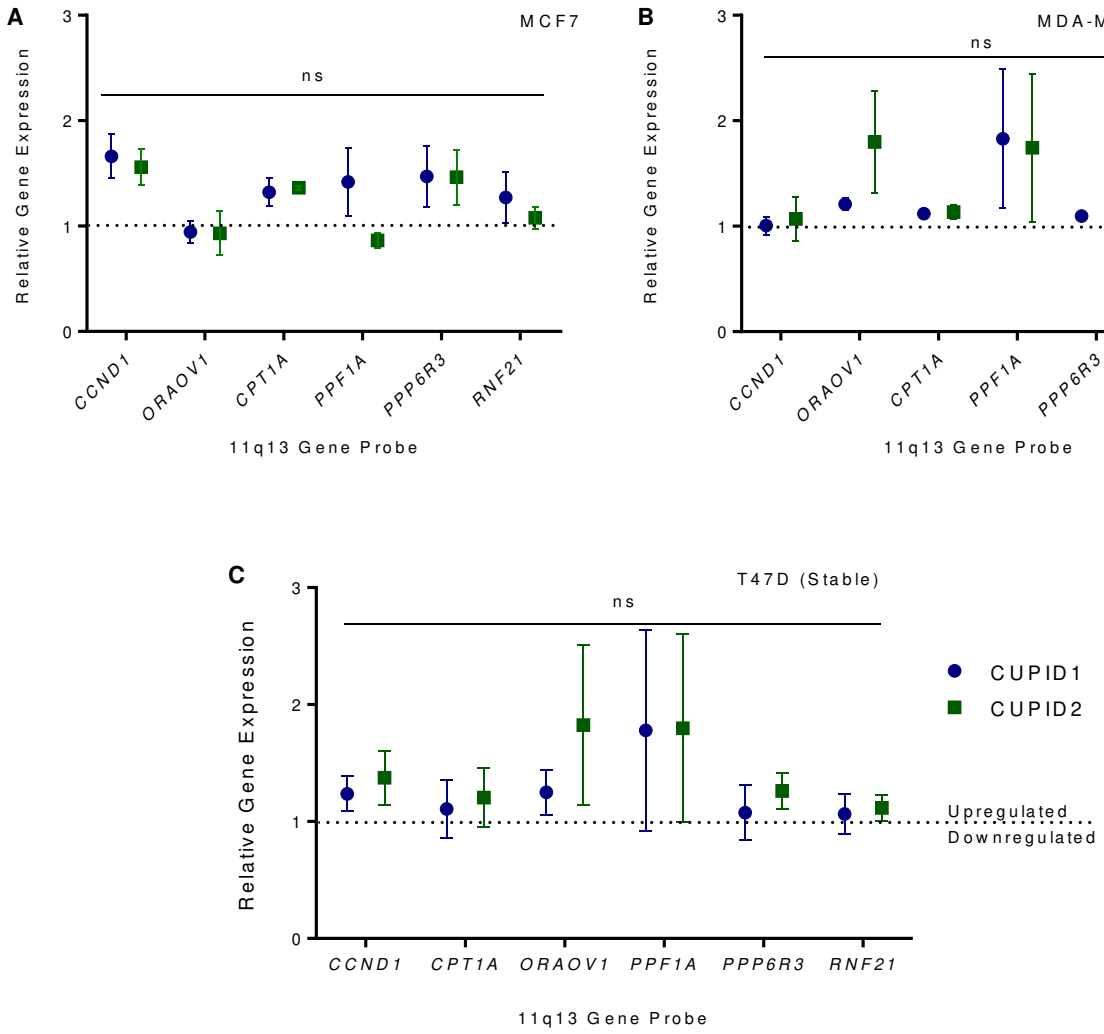


Figure 5.5 11q13 gene expression is not significantly altered following CUPID1 and CUPID2 over-expression. Gene expression was measured using Taqman assays. (A) Expression of 11q13 genes in MCF7 cells following *CUPID1* (blue) and *CUPID2* (green) transient overexpression. (B) Expression of 11q13 genes in MDAMB-231 cells following *CUPID1* and *CUPID2* transient overexpression. (C) Expression of 11q13 gene in T47D cells that stably over-express *CUPID1* and *CUPID2.* Expression data for each gene is the average of 3 biological replicates. Results are the mean expression +/- SEM presented relative to the empty pCDH vector. Significance calculated using a one sample t test, comparing values to a hypothetical mean of 1.0 as represented by the dotted black line. ns = p>0.05.

5.2.5. CUPID2 over-expression increases breast cancer cell proliferation and survival in vitro.

To determine whether *CUPID1* and *CUPID2* possessed the characteristics of oncogenes they were assessed *in vitro* using a colony assay and MTT assay. The colony assay provides a measure of cell survival and the ability to sustain clonal replication [389]. For the colony assay, MCF7 cells were transiently transfected with *CUPID1* and *CUPID2* expression constructs and seeded into 6 well plates as single colonies. The plates were cultured for 3 weeks, stained with crystal violet then imaged (*Figure 5.6A,B*). Significantly more colonies were found following *CUPID2* but not *CUPID1* over-expression compared to the vector only control suggesting that *CUPID2* supports breast cancer cell survival and clonal replication.

The effect of *CUPID1* and *CUPID2* over-expression on the proliferation of breast cancer cells was then assessed using the MTT assay which provides a proxy measure for the number of living cells in a culture [390]. T47D cells either stably expressing the lncRNAs or a selection vector only control were cultured for 4 days then processed according to the MTT protocol. Over-expression of *CUPID2* produced a significant increase in cell number whilst *CUPID1* overexpression did not alter cellular proliferation (*Figure 5.6C*).

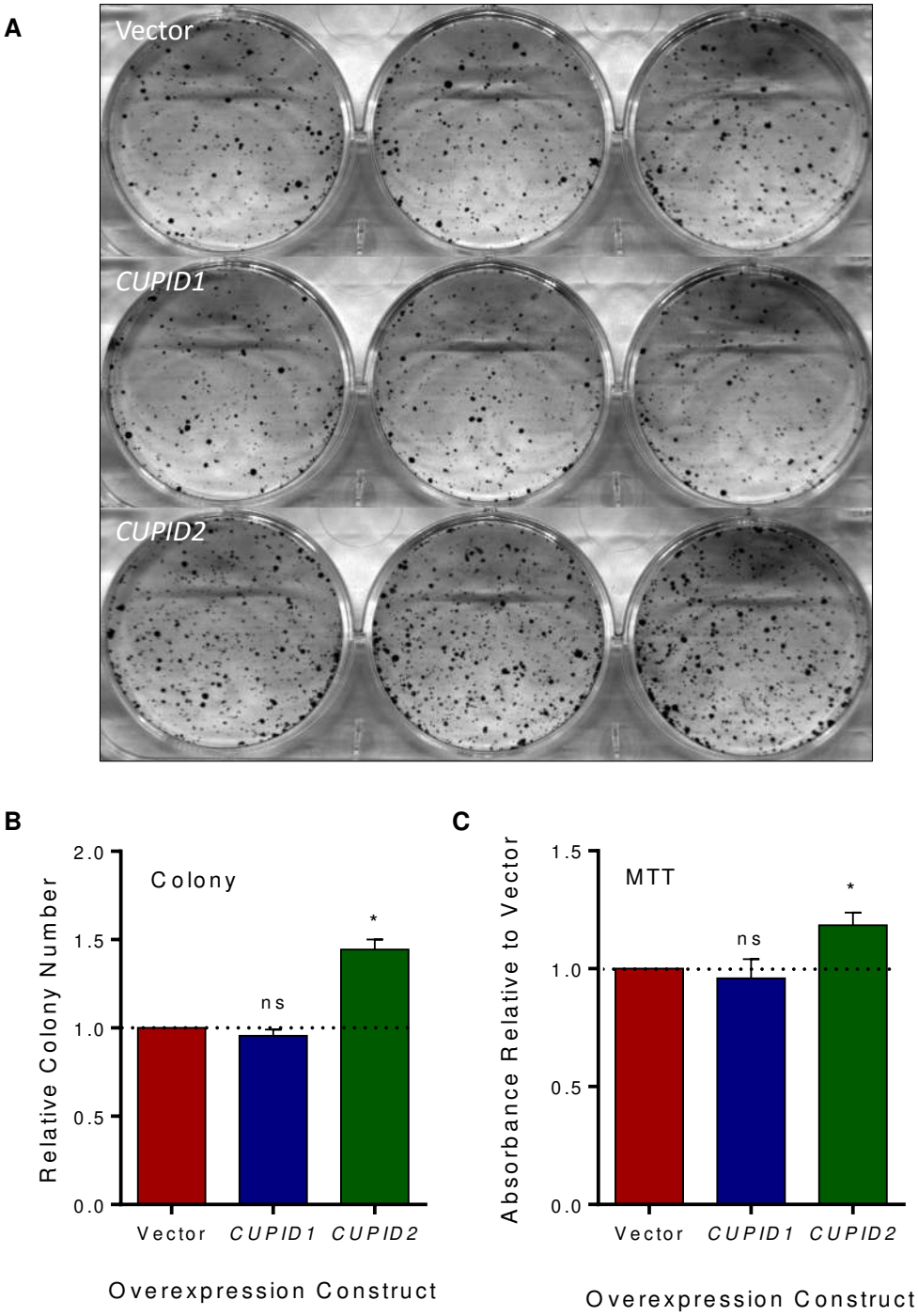


Figure 5.6 CUPID2 enhances breast cancer cell growth in vitro. (A) Scans of a representative biological replicate of the colony formation assay in MCF7 cells with the vector only transfection group (top), *CUPID1* over-expression (middle), and *CUPID2* over-expression (lower). (B) Mean colony numbers expressed relative to that for the vector +/- SEM. Three biological replicates of the experiment were performed with 3 technical replicates per condition. (C) Relative absorbance for *CUPID1* and *CUPID2* over-expressing cells in the MTT assay. Data shown is mean +/- SEM relative to the pCDH vector only control group for three biological replicates. Significance for both assays calculated using a one sample t test, comparing values to a hypothetical mean of 1.0. * $p<0.05$, ns = not significant.

5.2.6. CUPID 2 promotes tumour growth in a mouse xenograft model.

To determine whether *CUPID1* and *CUPID2* over-expression affected tumour formation *in vivo*, a murine xenograft tumour was used [391]. One million T47D cells stably over-expressing either the lncRNAs or vector alone were injected into the left lower mammary fat pad of SCID (severe combined immune-deficient) mice, with five mice per group. An estradiol pellet was also inserted to support tumour growth as T47D cells are oestrogen dependent [392]. After 7 weeks there were no palpable tumours in the vector only and *CUPID1* over-expressing mice so the mammary glands were removed for a better assessment of tumour growth (*Figure 5.7*). Mammary fat pad injection and animal care was performed by Dr Shu Shu Wen and Dr Christina Wong and animals then returned after culling for processing of the tumours.

Small tumours were visible in specimen 1 and 3 of the vector only expressing group, with a larger tumour in specimen 5. Only specimen 2 exhibited tumour formation in the *CUPID1* over-expression group and two mice in this cohort died of unknown causes prior to the final cull. Large tumours are visible in all specimens of the *CUPID2* over-expression group, with the tumours from specimens 2-4 being 2-3 times larger than the largest tumour found in the vector group. The lungs and liver were removed from the mice during the dissection process and examined macroscopically, however no signs of hepatic or pulmonary metastasis were found. There were also no neurological symptoms indicative of cerebral metastases seen in any of the mice during the experiment. Statistical analysis revealed a significant difference in tumour size seen between the control group and *CUPID2* over-expressing group when all data was collated (*Figure 5.8*).

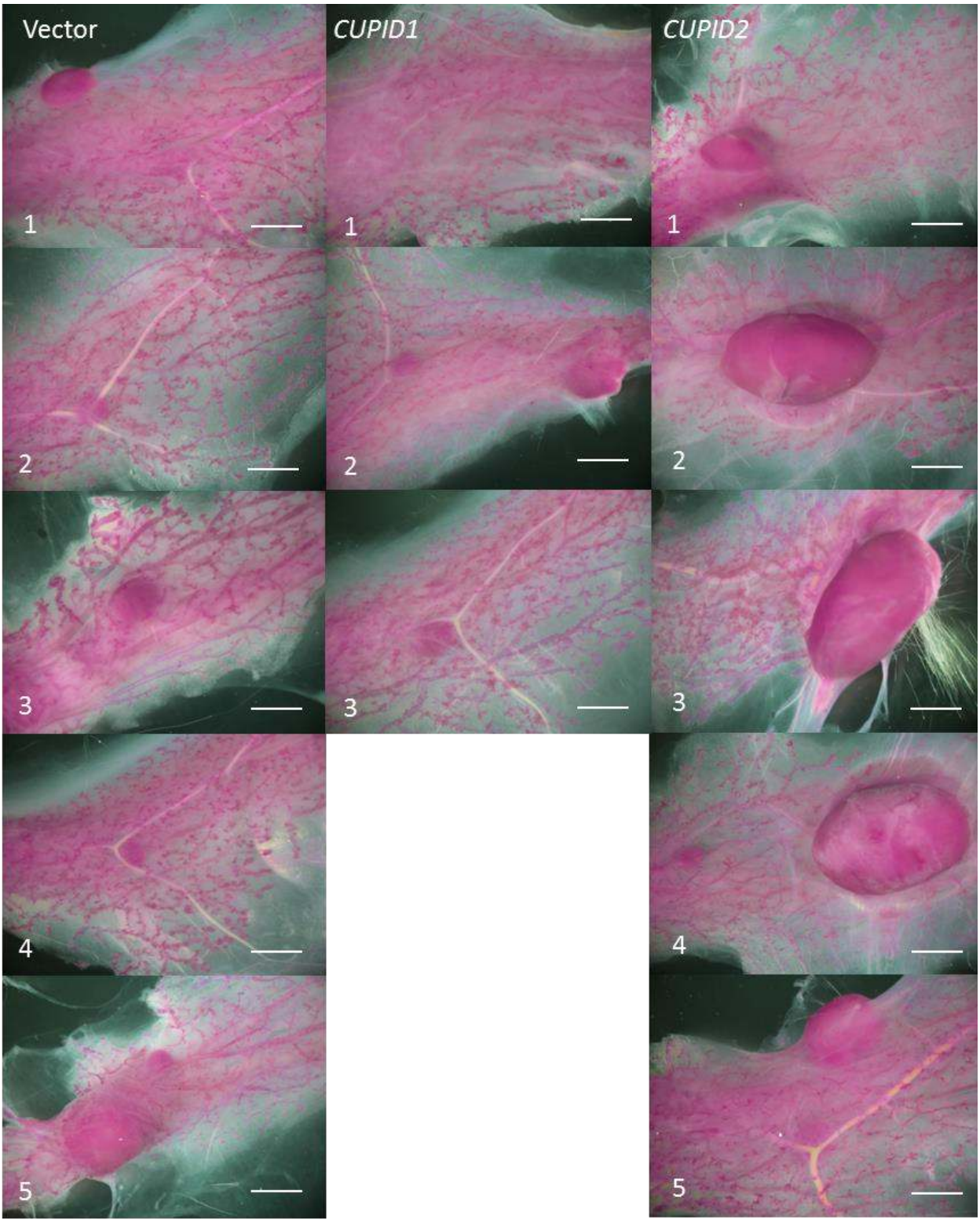


Figure 5.7 Large mammary tumours are produced by xenografts over-expressing CUPID2. A Leica MZ6 stereo microscope was used to capture images of the tumours. Specimens from each group are labelled 1-5, with vector only controls in the left column, *CUPID1* over-expressing cells in the middle column; and *CUPID2* over-expressing cells in the right column. The mammary glands were whole mounted onto coverslips and stained with carmine red. Specimens are displayed for all mice except the two from the *CUPID1* group that died prematurely. Scale bar = 2mm.

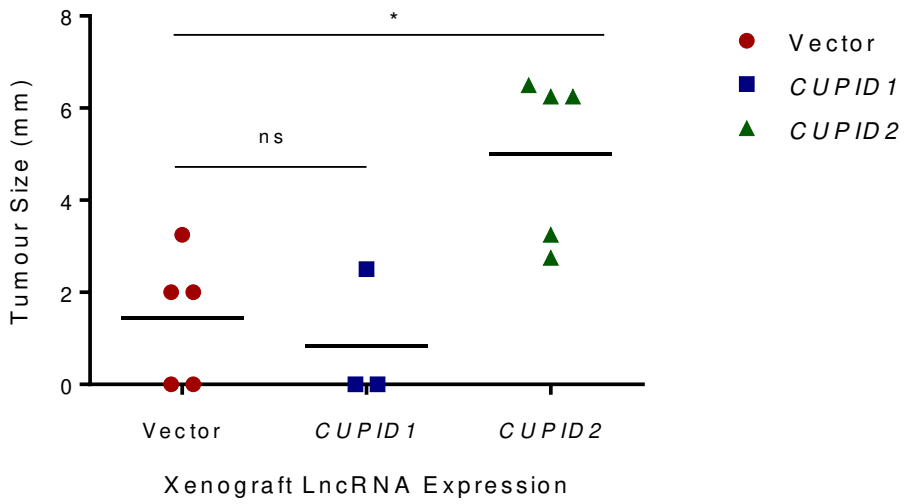


Figure 5.8 An interleaved scatter plot of relative tumour size from the xenograft mouse model. Each data point represents a tumour from an individual mouse with the solid black line indicating the mean size for each group. Significance was calculated using a one way ANOVA with Dunnett's correction for multiple comparisons. *p<0.05.

This chapter explored whether *CUPID1* and *CUPID2* amplification may act as a driver of breast tumorigenesis. Expression data in breast cancer cell lines (*Figure 4.6*) and tumour specimens (*Figure 5.2*) supported this hypothesis and their effect on measures of oncogenic potential was assessed both *in vitro* and *in vivo*. The MTT and colony formation assays demonstrated an increase in cell proliferation and survival associated with *CUPID2* (*Figure 5.6*). A murine xenograft model of breast tumour growth then showed a significant *in vivo* effect with an increase in tumour size seen for *CUPID2* over-expression (*Figure 5.7*). These findings support a role for *CUPID2* as a putative oncogene and highlight the potential for lncRNAs such as *CUPID2* to be used as disease biomarkers or future therapeutic targets [380].

Genomic regions subject to copy number changes such as amplification frequently contain key oncogenes [393]. Evidence that a particular gene may drive expansion of an amplicon requires that the expression of the gene is linked to its amplification and secondly that over-expression of the gene enhances a malignant phenotype, thus identifying it as a potential oncogene [8]. Oncogenic drivers tend to be found at the middle of amplicon cores as is the case with *CCND1*, *CUPID1* and *CUPID2* [14, 394]. Studies indicate that 11q13 is amplified in 9-24% of tumours and that *CCND1* expression is strongly linked to amplification [8]. This is also the case with other putative oncogenes at the 11q13 locus including *PAK1* and *EMSY* [8, 14, 395]. The determination of a relationship between amplification and expression similarly needs to be demonstrated for *CUPID1* and *CUPID2* and it is also of interest whether their expression may be correlated with that of *CCND1*. A bioinformatics analysis on TCGA breast cancer expression and copy number data is currently being conducted by another member of our group to resolve these issues. Analysis of the cell line expression data however (*Figure 3.3, Figure 4.6*) showed that high levels of *CCND1* expression are often seen in the absence of *CUPID1* and *CUPID2* expression, presumably due to the many mechanisms by which *CCND1* can be over-expressed in breast cancer cells [396].

Over-expression of genes known to be amplified in breast cancer such as *CCND1* or *HER2,* can be used to create an *in vitro* model of disease to investigate mechanisms by which potential oncogenes may act *in vivo* [397]. Due to experimental limitations, only the most prevalent isoforms of *CUPID1* and *CUPID2* were chosen for further study however it is important to note that gene isoforms can have different functions as has been shown for *CCND1* [398]. It is therefore possible that other isoforms of the lncRNAs may have performed differently in subsequent assays. High levels of *CUPID1* and *CUPID2* expression were seen following both transient transfection and stable transduction of the target cell lines however this did not result in a significant change in

expression of the assayed genes. This may be an accurate reflection of the biology involved or it may be cell type dependent as secondary phenotypes produced by forced gene over-expression can alter depending on the cell line utilized [397]. A non-significant trend was observed for an increase in *CCND1* expression levels following *CUPID1* and *CUPID2* over-expression which was in keeping with previous data showing that knockdown of the lncRNAs decreased the levels of *CCND1* (*Figure 4.19, Figure 5.5*). Confirmation of a local regulatory effect on genes *in cis* raises the possibility that *CUPID1* and *CUPID2* may belong to the enhancer-like lncRNA subset described by Orom *et al*. [224].

Another consideration is that ectopic expression of chromatin acting lncRNAs may not recapitulate normal function, which requires a local action *in cis* for appropriate biological effect [399, 400]. This was seen by Xiang *et al.* at the 8q24 multiple cancer risk locus where ectopic over-expression *in trans* of the *CCAT1-L* lncRNA had no effect on the expression of the nearby *MYC* oncogene whilst expression *in cis* using genome editing techniques increased *MYC* levels 1.3-1.8 fold [204]. Despite not showing an effect on oncogenic assays *CUPID1* may therefore still be found to have a role as an oncogene if its overexpression is re-examined *in cis*. This may be less of an issue for *CUPID2,* which unlike the heavily chromatin bound *CUPID1*, is spread across nuclear and cytoplasmic compartments and thus more likely to act distal to its transcriptional origin (*Figure 4.17*). Effects from over-expression of an oncogene may also be complex and mechanisms of action better seen by examining perturbations in oncogenic pathways rather than individual genes [401]. RNAseq has therefore been employed to explore such global changes following stable over-expression of *CUPID2* in the T47D cell model. Extracted RNA from three biological replicates of T47D cell populations independently transduced to either over-express *CUPID2* or the pCDH vector alone control has been sent for sequencing and will be analysed by the Ingenuity program (Qiagen) to assess changes in cellular gene networks.

The effect of lncRNA over-expression was assessed using two slightly different *in vitro* measures of cell proliferation. The colony formation or clonogenic cell survival assay measures the ability of an individual cell to form colonies and demonstrates that an individual cell retains the ability to replicate over many generations [389]. The MTT assay however is a more broad assessment of proliferation in a mixed group of cells [390]. Increased cell survival and proliferation are both hallmarks of cancer and characteristics of oncogenes [402]. *CUPID2* over-expression caused a significant increase in both measures indicating enhancement of the oncogenic potential of the cells *(Figure 5.6). In vitro* assays can be misleading however, with Aapro *et al*. demonstrating that cells taken from breast cancer specimens showed no relationship between the ability to form colonies and

subsequent survival [403]. It was therefore important to investigate the lncRNAs *in vivo* using a xenograft model to mimic the over-expression of *CUPID2* observed in breast cancer.

Xenograft model have been previously used to assess putative drivers of the 11q13 amplicon such as *EMSY* and allow a more biologically relevant assessment of breast cancer cell growth in three dimensions [404, 405]. An orthotopic injection of breast cancer cells was utilized as this provides a better mimic of the native mammary stromal environment [404]. A significant increase in tumour growth was observed for those xenografts containing *CUPID2* over-expressing cells. This was consistent with the increases in cell survival and proliferation obtained from the *in vitro* experiments, providing further evidence that *CUPID2* may be a novel oncogene (*Figure 5.7*). Unfortunately, many of the mice in all groups exhibited diarrhoea, abdominal distension and blood in the perineal region 5-6 weeks following injection, with the death of two mice in the *CUPID1* over-expression group. This may have been due to bladder calculus formation which is a known side effect of estradiol supplementation and a common cause of mortality in xenograft experiments using ERα positive tumours [406].

To confirm these findings, the xenograft experiments will be repeated using a higher initial number of cells in the orthotopic injection and with the addition of matrigel to enhance tumour generation, growth and metastatic potential [407]. The aim would be to accelerate the development of tumours and enable the mice to be culled prior to the development of symptomatic bladder calculi. Another option would be to use NSG (NOD SCID gamma) mice which have impaired cytokine production in addition to a lack of competent T, B and NK cell activity [408]. NSG mice are claimed to be more resistant to the effects of estradiol supplementation according to their supplier (The Jackson Laboratory http://jaxmice.jax.org/nod-scid-gamma/) and may thus be less likely to succumb to bladder calculi, however direct evidence for this was not found in the literature. They are also more susceptible to metastasis than SCID mice, presumably due to the absence of natural killer cells [409].

This chapter has provided evidence that *CUPID2* may represent a novel oncogene in breast cancer and a potential co-driver of the 11q13 amplicon. Further confirmation of the oncogenic stratus of *CUPID2* may be obtained by demonstrating that mechanisms other than amplification are involved in its over-expression or increased activity. Given that *CUPID2* is oestrogen regulated (*Figure 4.8*), an obvious candidate mechanism would be over-expression of ERα. This is a common finding in ERα positive breast cancer and often seen in relation to *CCND1* expression in the absence of amplification [8, 410]. Xenograft experiments combining over-expression of *CUPID2* with

selective knockdown would also support the hypothesis that *CUPID2* has a direct role in tumour growth [411]. Should that be the case, *CUPID2* may have utility as a biomarker for ERα positive breast cancer and represent a promising target for therapeutics against this challenging disease.

CHAPTER 6

Final Discussion

My group has previously fine-mapped the 11q13 breast cancer risk locus and identified *CCND1* as one of the target genes mediating risk [5]. In an effort to investigate the 11q13 locus in a more comprehensive and agnostic manner, several different 3C-based techniques were used in this thesis to identify chromatin interactions between PRE1 or PRE2 and candidate target genes. 4C-seq was used to detect distal interactions with PRE1 *in trans* and 5C to detect proximal interactions *in cis*. 4C-seq was a reasonable approach for the 11q13 risk locus as the five candidate causal SNPs could be genetically fine mapped to two discrete regulatory elements that were then used as bait for the 4C-seq experiment. Whilst the 4C-seq did not provide sufficient resolution to map proximal interactions, it did identify clusters of interactions between PRE1 and other regions commonly amplified in breast cancer that may be of relevance to disease progression (*Figure 3.8*). These regions included 17q23 and 20q13 which have been previously shown to form extensive inter-chromosomal interactions that dysregulate genes critical to breast cancer proliferation [280]. The interactions were only seen in the highly amplified breast cancer cell lines MCF7 and BT474, whilst no distal interactions *in trans* were seen in the karyotypically normal Bre80 breast cell line. This is consistent with the capture HiC study performed on the 2q35, 8q24 and 9q31 breast cancer risk loci which found no interactions *in trans* for the normal GM06990 cells but 11 and 20 interactions respectively for the amplified BT483 and SUM44 cancer cell lines [162]. Such findings emphasize the need for caution in interpreting genome-wide 4C data obtained from cancer cell lines. It is less of an issue for the 3C and 5C techniques which usually characterise interactions over shorter genomic distances that may be entirely contained within a region of focal amplification.

5C was then used to determine proximal interacting partners of PRE1 and demonstrated interactions with several target genes including *CPT1A*, *MRP21, IGHMBP2*, *CCND1* and *ORAOV1*(*Figure 3.9*). Of these, the two genes prioritised for follow up are plausible candidates for mediating an increase in breast cancer risk. *CPT1A* (carnitine palmitoyltransferase 1A) encodes a mitochondrial protein involved in fatty acid metabolism and has been shown to regulate cell motility and tumour growth in alveolar rhabodomyosarcoma and a transgenic mouse model of Burkitt's lymphoma [412, 413]. Upregulation of fatty acid synthesis and oxidation is crucial to provide the energy and molecular substrates required to sustain increased cancer cell proliferation and *CPT1* inhibitors are promising adjuvant treatments for breast cancer [413-416]. Related genes involved in fatty acid metabolism have previously been implicated in mediating the risk of breast cancer at the 6q22.33 locus and colorectal cancer at the 11q12.2 locus [417, 418]. *IGHMBP2* (immunoglobulin μ-binding protein 2) encodes a helicase associated with ribosomes and is suggested to have additional roles in DNA damage repair [317, 318, 419, 420]. Such a role would be consistent with the functions of high and

moderate penetrance genes previously associated with breast cancer risk, the majority of which are components of the DNA damage repair pathways [2]. Further functional work is still required to confirm that *CPT1A* and *IGHMBP2* are involved in breast cancer risk at 11q13, including 3C to confirm the interaction between PRE1 and the gene promoters and luciferase expression assays to assess the effects of the SNPs on promoter transactivation.

An alternative approach to validating the identified interactions was also trialled, involving silencing of the eRNA transcribed from PRE1 (*Figure 3.10*). Several studies have shown that silencing of the eRNA produced by an enhancer can reduce the expression of target genes and dramatically reconfigure chromatin interactions connecting enhancers and promoters in a locus [234]. Silencing of an eRNA can be achieved using siRNA, modified antisense oligonucleotides (M-ASO), or by direct demethylation of the enhancer to inactivate it [232, 234, 421]. A comparison between siRNAs and M-ASOs determined that the most consistent PRE1 eRNA reduction was obtained using siRNA and subsequent gene expression analysis showed a reduction in expression of *CPT1A* and *IGHMBP2* (*Figure 3.13B*). Targeting eRNAs as a therapeutic strategy has already been demonstrated at the *MYC* locus where silencing of the *CCAT1* and *CCAT2* eRNAs reduced colon cancer cell proliferation, presumably through down-regulation of the *MYC* oncogene [204, 381]. The silencing of eRNA or direct enhancer decommissioning as therapeutic options have the advantage of being very tissue specific and would avoid the problems of inhibiting important genes such as *MYC* or *CCND1* in normal tissues [233, 422]. A suite of eRNAs regulating key oncogenes could theoretically be silenced to down-regulate drivers of tumour growth. The experiments in chapter three demonstrated that it can however be challenging to obtain sufficient reduction in eRNA levels at amplified regions, many of which would contain the relevant target oncogenes (*Figure 3.13A*) [14].

Our initial fine mapping study of the 11q13 risk locus focussed on *CCND1,* a well characterised oncogene in breast cancer [5, 200]. However, the emerging role of non-coding transcripts in breast cancer prompted a search for nearby lncRNAs that may be influenced by the risk SNPs and hence potentially involved in breast cancer biology [254, 423]. RNA Capture-seq was used to comprehensively probe the LD block containing the original marker SNP at 11q13, and uncovered two multi-exonic transcripts called *CUPID1* and *CUPID2* (*Figure 4.2*), that were oestrogen regulated (*Figure 4.8*), and enriched in ERα breast cancer cells (*Figure 4.7*). Chromatin interactions were demonstrated between PRE1 and the *CUPID* promoter (*Figure 4.10*) and the PRE1 enhancer was shown to cause a robust increase in promoter activity (*Figures 4.14, 4.15*).

The findings at the 11q13 breast cancer risk locus have many parallels with the 8q24 breast cancer risk locus. The risk haplotypes both fall in gene desert regions with nearby risk loci for multiple cancer subtypes and the candidate causal SNPs were both proposed to affect enhancers regulating nearby oncogenes; *MYC* at 8q24 or *CCND1* at 11q13 [5, 155, 271, 272, 287]. Multiple lncRNAs have also been identified in the 8q24 region which are affected by SNPs associated with prostate cancer risk (*PRNCR1*), colorectal cancer risk (*CCAT1* and *CCAT2*), or both (*CARLo-5*) [203, 204, 357, 381]. Interestingly, the *CCAT1* and *CCAT2* lncRNAs both arise from enhancers of *MYC* and are thus better classified as eRNAs [204, 381]. Only *CCAT2* contains a disease SNP variant, with the risk allele increasing expression of the *CCAT2* eRNA [381]. The expression of the 11q13 PRE1 eRNA has not been confirmed to vary with the risk SNPs in this manner, however it seems likely given the changes in PRE1 promotor activity observed with SNP1 and SNP2 (*Figure 3.12*). Further characterisation of the PRE1 eRNA is still required as described for *CCAT1* and *CCAT2*, including 3C following eRNA knockdown to assess the role of the eRNA in mediating chromatin looping between PRE1 and *CCND1* or the *CUPID* promoters [204]. Given that the eRNA contains SNP1, it would also be interesting to explore the effects of structural differences between eRNAs containing the major (protective) allele of SNP1 or the minor (risk) allele as such 'riboSNitches' have been shown to have an important role in lncRNA function [194, 312]. This would require *in silico* and *in vitro* approaches to predict the effect of SNP1 on the eRNA structure [424, 425], its ability to bind miRNA [179, 424], and its potential effect as an eQTL [239, 425]. LncRNAs in general have a lower density of SNPs than comparative protein coding genes, possibly due to extra constraints imposed by the tight relationship between their structure and function [426, 427].

The minor (risk) alleles of SNP1 and SNP2 were also shown to reduce the ability of PRE1 to enhance *CUPID1* promoter activity when incorporated into a luciferase construct (*Figure 4.14*). Such transactivation effects of risk-SNPs on regulatory elements are frequently observed in post GWAS studies of breast cancer including at the 2q35 and 5q11.2 loci, and appear to be a common mechanism by which SNPs may promote a disease phenotype [5, 59, 60]. The 25% reduction of enhancer activation seen is consistent with previous high throughput studies of SNPs that dysregulate transcription, with a 1.3-2 fold change in activity being typical [189, 346]. As part of the risk haplotype, the minor allele of SNP1 was also shown to reduce chromatin looping between PRE1 and the *CUPID* promoter compared to the major (protective) allele (*Figure 4.13*). PRE1 increases *CUPID* promoter activity 25-200 fold over baseline so even a small reduction in interaction frequencies should cause a marked reduction in both *CUPID1* and *CUPID2* expression

(*Figure 4.14*). Allele specific looping is another mechanism previously observed both at the 2q35 breast cancer locus and also in a GWAS locus associated with skin pigmentation where multiple allele specific loops were found to connect two enhancers and the *IRF4* gene to alter transcription [60, 428]. The changes in transactivation and chromatin looping associated with the risk-SNPs are predicted to reduce the expression of *CUPID1* and *CUPID2*. A subsequent siRNA mediated reduction in *CUPID1* and *CUPID2* levels was in turn shown to impair the DNA damage response (*Figures 4.26 and 4.27*). Interestingly, genetic variants reducing the activity of other components of the DNA damage repair pathway such as *BRCA1, BRCA2, CHEK* and *ATM* are well characterised causes of breast cancer [2, 429, 430]. Unfortunately it proved technically challenging to create isogenic cell lines for the 11q13 risk-SNPs which would have allowed confirmation of a reduction in *CUPID1* and *CUPID2* expression as predicted by the *in vitro* assays (*Figure 3.17*). Another means of validation would be to demonstrate that the risk-SNPs represented eQTLs that reduced *CUPID1* and *CUPID2* expression however available datasets are currently inadequate for this purpose as they neither contain sufficient samples from breast tissue nor coverage of lowly expressed lncRNAs [431, 432].

CUPID1 was shown to associate with chromatin on a nuclear fractionation assay however it was unclear whether it may act *in cis* or *in trans* (*Figure 4.17*). An action *in cis* may include recruiting regulatory proteins to the *CUPID* locus or acting as a scaffold to mediate looping between the *CUPID* locus and interacting DNA [223, 433]. These functions have been demonstrated at the 8q24 risk locus for *CCAT1* [204], the lncRNA *LUNAR* at the *IGF1R* locus [434], and more globally for eRNAs regulating the oestrogen response [234]. An action *in trans* is also possible, as seen for the lncRNA *HOTAIR* which recruits the polycomb complex genome-wide to silence target genes and is an important oncogene in breast cancer [254, 433]. ChIRP-seq was performed to identify the binding sites of *CUPID1* (*Figure 4.18*), and identified 21 gene promoters with high confidence *CUPID1* binding peaks (FDR<0.5%) within 1kb (*Table 4.1*) [261]. A functional network analysis of the binding peaks revealed enrichment for genes associated with DNA replication and repair as the most significant result, with additional changes in networks associated with cellular organisation, cancer and endocrine functioning (*Table 4.2*). RNA-seq data from *CUPID1* silenced breast cancer cells is awaited to confirm that *CUPID1* binding to the identified gene promoters is indeed associated with a change in gene expression.

All the experiments incorporating silencing to assess the function of *CUPID1* and *CUPID2* used siRNA and obtained a knockdown efficiency of 50-60% for *CUPID1* and 70-80% for *CUPID2*

(*Figure 4.19A*). The efficiency for *CUPID2* is similar to that demonstrated in other studies involving siRNA silencing of lncRNAs implicated in cancer such as *CARLo5* and *MALAT* [357, 435]. The *CUPID1* silencing efficiency was lower however, which may be a function of its marked localisation to chromatin, making it less accessible to the siRNA [197, 435]. Though a trend was seen for reduced expression of *CCND1, CPT1A, MYEOV* and *ORAOV1* with *CUPID2* silencing, the levels did not reach statistical significance except *MYEOV* with siRNAC2-4 (*Figures 4.19B and C*). There were also no consistent changes in gene expression shared between the different siRNAs used to silence *CUPID1*. This may be due to an absence of a functional effect or it may reflect difficulty in achieving sufficient silencing of the lncRNAs which are already highly over-expressed in MCF7 cells compared to ERα negative breast cell lines [197, 435]. Studies on the highly expressed lncRNA *MALAT* have returned differing phenotypes depending on the silencing method employed [435-438], and one recent guideline recommends at least two methods of silencing be employed to achieve confidence in any obtained results [358]. The options for lncRNA silencing include deletion of the promoter or entire locus; insertion of a premature transcriptional terminator; knockdown using siRNA, shRNA or M-ASO; and epigenetic modification at the promoter to repress transcription [358]. Many of these are however impractical to use for *CUPID1* and *CUPID2* silencing in available ERα positive cell lines due to local amplification. The inconsistent results seen for gene expression (*Figure 4.19*) and negative results for chromatin looping (*Figure 4.22*) and involvement of the lncRNAs in the oestrogen response (*Figure 4.20*) will need to be revisited in later studies once an improved system of *CUPID1* and *CUPID2* silencing has been developed.

A robust phenotype was seen for the DNA damage assays in support of the ChIRP-seq results, with reduced Rad51 foci formation at induced DSBs (*Figure 4.26*) and a highly significant reduction in MCF7 cell proliferation (*Figure 4.27*) following silencing of *CUPID1* and *CUPID2*. Further assays are awaited to confirm an involvement of *CUPID1* and *CUPID2* in DNA damage repair, however the magnitude of the observed response so far suggests that MCF7 cells may be 'addicted' to high levels of *CUPID1* and *CUPID2* for their ongoing proliferation. Breast cancer cells often have high rates of replication in combination with an impairment of various components of the DNA damage pathways such as the recombination co-mediators *BRCA1* and *BRCA2* [439-441]. To prevent excessive genome instability and subsequent apoptosis they often compensate by having increased levels of other critical factors of the DNA repair pathways such as the homologous recombination (HR) mediator Rad51 [440, 442]. Inhibition of Rad51 thus holds great promise as a breast cancer therapeutic (reviewed in [443]). In the same way, therapies targeting *CUPID1* and *CUPID2* may be a complementary strategy for breast cancer treatment alongside the current efforts to inhibit critical

DNA repair proteins such as Rad51 or PARP [443, 444]. It would therefore be interesting to combine *CUPID1* and *CUPID2* silencing with PARP inhibition to assess whether this has an additive effect on toxicity as is seen with inhibition of PARP in cells deficient in other components of HR repair such as *BRCA1* [376, 444].

Proposing *CUPID1* and *CUPID2* inhibition as a potential therapy for breast cancer seems counter-intuitive when a reduction in their expression caused by the SNPs is predicted to increase breast cancer risk (*Figures 4.13 and 4.14*). The same dichotomy occurs with *CCND1* which is a highly over-expressed driver of ERα positive breast cancer progression, however the GWAS risk-SNPs at 11q13 cause a reduction in its expression [5, 200]. The mechanism described above for Rad51may also be applicable to cyclinD1as it is known to be a crucial component of oestrogen mediated DNA repair and could theoretically be over-expressed to compensate for other defects in the DNA damage pathway [201, 202]. Genes identified through breast cancer GWAS are often found to have a role in breast cancer progression, although the SNPs usually alter the expression of the target gene in the same direction as is seen in breast tumours. This is the case with the *FGFR2* locus where the minor (risk) allele increased FoxA1 binding to a nearby enhancer and was thus predicted to increase *FGFR2* expression [62]. Over-expression of *FGFR2* or the presence of activating mutations in FGFR2 kinase are able to drive breast cancer progression and clonal amplification [445]. Conversely, *MAP3K1* expression is increased by the germline risk alleles at 5q11.1, whilst somatic mutations of *MAP3K1* in breast cancer are predicted to be inactivating [59].

An interesting contrast to the breast cancer associated risk locus at 11q13 is the nearby renal cancer risk locus where the risk allele has been shown to increase the activity of a distal enhancer element regulating *CCND1* instead of reducing it as seen for the breast associated risk-SNPs [5, 271]. A feature shared between renal and breast cancer however is that they both exhibit over-expression of *CUPID2* (*Figure 5.2*) [241]. Further characterisation of the renal locus and adjacent prostate cancer locus is required to determine whether they may also mediate risk through an effect on *CUPID1* and *CUPID2* and whether the risk alleles alter lncRNA expression in the same direction as is seen in tumours. Although the underlying mechanisms may be complex and even seem contradictory at times, these examples all highlight the potential for breast cancer GWAS to find important and novel breast cancer genes, with the ultimate aim of developing new therapeutics or biomarkers of disease [26]. From a clinical perspective however, even without extensive fine mapping and any mechanistic understanding of the causal SNP effects, GWAS data can be used to guide the repositioning of existing drugs which target genes within the identified risk locus [446, 447].

An intriguing finding from 11q13 copy number variation studies is that *CCND1* amplification is strongly associated with an adverse effect of tamoxifen in premenopausal breast cancer, however this effect is independent of cyclinD1 protein levels [448]. This suggests that other genes co-amplified with *CCND1* (such as *CUPID1* and *CUPID2*) may be affecting the tamoxifen response [394]. Tamoxifen is an ERα antagonist and the most commonly used drug in breast cancer treatment, whilst the *CCND1* locus is one of the most heavily amplified in ERα positive breast cancer so an explanation for this effect is of major importance to a large percentage of breast cancer patients [8, 394, 449]. In chapter four, a trend for a reduced response to oestrogen was observed following *CUPID2* knockdown, though this did not reach statistical significance (*Figure 4.20*). It would be interesting to investigate this further using a more comprehensive silencing approach and also to combine over-expression of *CUPID1* and *CUPID2* in breast cancer cells with tamoxifen treatment to see whether tamoxifen resistance could be induced through over-expression of the lncRNAs in a previously sensitive cell line. Consistent with this, the ERα positive cell lines with the highest levels of *CUPID1* and *CUPID2* expression (MCF7 and BT474) are both resistant to tamoxifen and are amplified at the 11q13 locus (*Figure 4.6*) [450].

CUPID2 was shown to increase cell proliferation in the MTT and colony formation *in vitro* assays (*Figure 5.6*). It also produced a marked increase in tumour size when over-expressed in a xenograft murine model of breast cancer (*Figure 5.7*). Although *CUPID1* was not shown to affect cell proliferation *in vitro* or tumour growth *in vivo* (*Figures 5.6, 5.7, 5.8*), it may require over-expression *in cis* for an effect to be seen and this should be attempted before discounting a role as a potential driver of ERα positive breast cancer [204]. An action *in cis* would also explain why it is not seen in the TCGA lncRNA data set as it would only require low level overexpression to have a biological effect and this may not have been detected using standard RNA-seq testing [241]. The presence of *CUPID2* within an amplified region and its demonstrated biological effects are consistent with a role as a putative oncogene [355]. Santarius *et al*. provide a set of criteria for defining oncogenes which synthesizes clinical information, molecular changes such as amplification, biological evidence from over-expression and silencing experiments and finally data from animal models [411]. Taking these guidelines into account, more information is required before *CUPID2* can be considered an oncogene, including expression data from further cancer cell lines and mouse models using xenografts with *CUPID2 k*nockdown to complement the overexpression model. The alternative possibility is that *CUPID2* is merely one of the many passenger genes over-expressed as part of the 11q13 amplicon due to another driver gene such as *CCND1* [8].

Cancer cells often become dependent on high levels of particular oncogenes for maintenance of a tumour phenotype as is the case with *CCND1* in oesophageal cancer [388]. This 'oncogene addiction' provides the opportunity to design therapies which specifically target a tumour but have minimal effect on normal cells which do not depend on the overexpressed oncogene [451, 452]. Interestingly, in a mouse model using sarcomas overexpressing *MYC*, even temporary inhibition of *MYC* was enough to cause the tumours to regress and subsequently undergo apoptosis when the inhibition was lifted [453]. A feature of such 'addictive' oncogenes is that their overexpression consistently follows their amplification due to a survival advantage conferred on the clones where it occurs [454-456]. Further work is required to determine the relationship between expression and amplification with *CUPID2* and TCGA data is currently being analyzed by another group member in this regard. This information could be combined with xenograft models incorporating a brief or sustained knockdown of *CUPID2* with an inducible shRNA system to confirm whether *CUPID2* is required for the maintenance of tumour growth [457]. Such experiments have been performed *in vitro* for a number of other oncogenic lncRNAs including *HOTAIR* [254, 341], *HNF1A-AS1* [458], *CARLo-5* [357], and *FAL1* [355] with a resultant reduction in the malignant phenotype. These examples are just the beginning however, as there are over 60,000 lncRNA genes identified to date compared to less than 30,000 protein coding genes making it is highly likely that many of these non-coding transcripts will have a role in cancer biology [241].

The data obtained through this study can be used to propose a model explaining how the risk-SNPs lead to an increased risk of breast cancer at the 11q13 locus (*Figure 6.1*). Under the influence of oestrogen, the regulatory element PRE1 makes chromatin interactions with the *CUPID, CCND1, CPT1A* and *IGHMBP2* promoters. Subsequent activation of the *CUPID* promoter by PRE1 induces the expression of *CUPID1* and *CUPID2* which are required for the DNA damage response and additional functions which are yet to be confirmed. The presence of the risk-SNPs reduces both the activity of PRE1 and chromatin looping between PRE1 and the *CUPID* promoter. This is predicted to reduce the expression of *CUPID1* and *CUPID2* and impair the DNA damage response, leading to genomic instability and an increased risk of breast cancer.

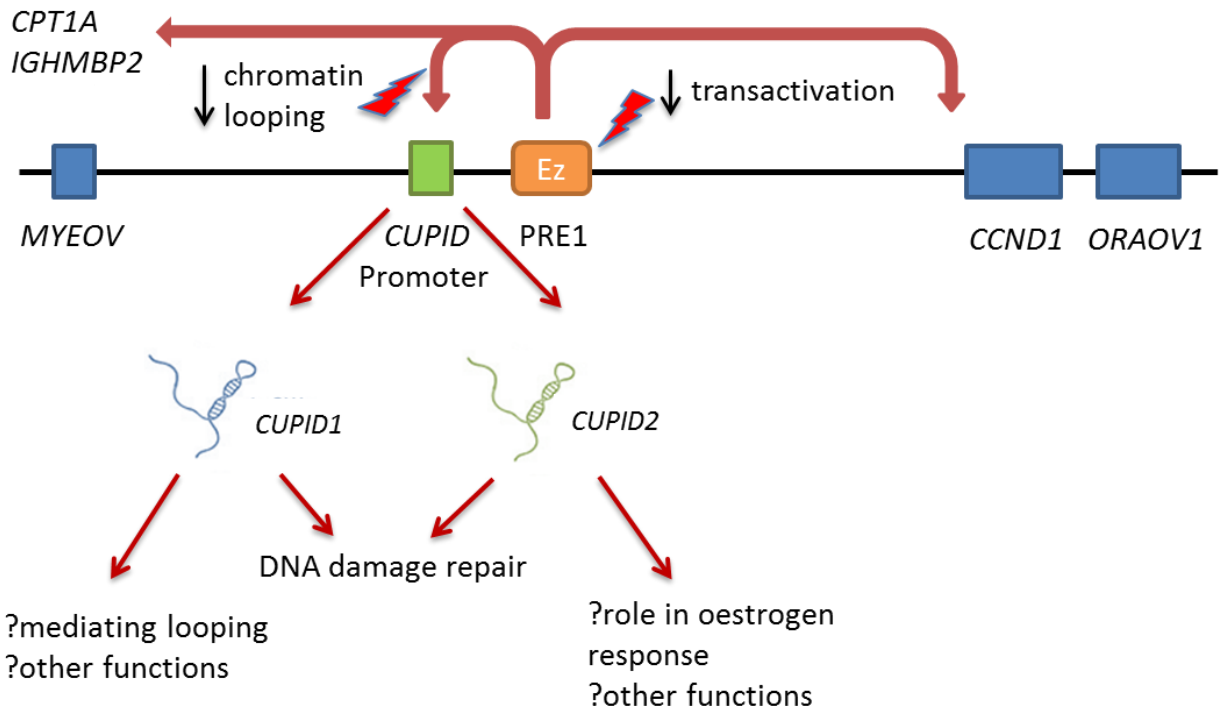


Figure 6.1 A model of risk at the 11q13 breast cancer susceptibility locus. The blue boxes denote genes. PRE1 (orange box) binds oestrogen (Ez) and participates in chromatin interactions (thick red arrows) with target gene promoters. Activation of the *CUPID* promoter (green box) by PRE1 induces expression of *CUPID1* and *CUPID2*. The risk SNP effects are depicted by red lightning bolts and indicate reduced transactivation of PRE1 and reduced chromatin looping between PRE1 and the *CUPID* promoter.

The ultimate goal of post-GWAS analysis is to fully understand how identified risk-SNPs increase disease risk and to apply this knowledge through prevention or treatment strategies in a clinical practice environment [34]. In regards to breast cancer, the prevention component aims to use the number of causal SNPs present in an individual to estimate their breast cancer risk and tailor screening recommendations accordingly [459]. Early attempts at using this data with the limited number of GWAS SNPs available at the time did not provide enough information to be of clinical use individually or at a population level [460, 461]. Post iCOGS however, a follow up study using a risk score derived from the status of 72 risk SNPs including fine mapped SNPs from the 11q13 locus, found that the breast cancer risk for women in the bottom quintile was 5.2% compared to 16.6% for those in the top quintile [462]. Combining this data with family history and lifestyle factors would thus provide sufficient power to guide screening at a population level and also allow better management of individual prevention strategies such as prophylactic medication or an increased frequency of mammography [459, 462]. An increasing number of private providers now offer gene panels containing a variety of different high, moderate and low penetrance variants, however many have not been appropriately validated and more work is needed to determine how best to proceed with preventative genetic counselling (reviewed in [463]). The second component of post-GWAS studies is to uncover biological mechanisms and novel pathways underlying disease which may be potentially targeted for therapies [34]. This requires distinguishing the functional SNPs from other tightly correlated SNPs that may be inherited together as a risk haplotype. That has been the focus of this thesis with the prioritization of the *CPT1A* and *IGHMBP2* genes for further investigation in Chapter 3 and the identification of *CUPID1* and *CUPID2* in Chapter 4.

When this project was being planned, 4C-seq and 5C were considered novel and powerful techniques with great potential for application in areas such as GWAS loci [172, 464]. A more recent trend however has been the use of a hybridisation capture step to enrich for interactions in the loci of interest which may have provided higher quality data if used at the 11q13 locus. This involves DNA capture following a modified 3C protocol including a sequencing step (Capture-C) [465], or capture combined with the standard HiC process (Capture HiC) [162]. Either approach then allows interactions to be mapped genome-wide for multiple loci simultaneously. Capture HiC has been used to characterise both breast cancer and colon cancer risk loci, with the breast cancer study confirming previously demonstrated interactions with *MYC* and *IGFBP5* at the 8q24 and 2q35 loci respectively [60, 155, 162, 173]. Capture-C has not yet been applied to GWAS loci in a published study, however a modification of the technique which incorporates a biotin pull down

step to improve the signal to noise ratio has been used at the 8q24 prostate cancer associated locus [466]. Further optimisation of such techniques is likely to see them widely adopted in future post-GWAS functional characterisation studies. Chapter Four demonstrates that these methods should ideally be performed in parallel with RNA Capture-seq to ensure all potentially relevant genes in the loci are identified otherwise potentially significant interactions may be ignored when they do not map to a previously known gene promoter [7].

All the approaches discussed so far, and indeed all the post-GWAS studies of breast cancer risk loci that have been performed to date rely on indirect evidence to demonstrate the functional effect of a risk SNP [5, 59, 60, 62]. The gold standard for such confirmation requires the creation of isogenic normal cell lines that differ only in the SNPs of interest [26]. This may be done using such tools as Zinc Finger Nucleases, TALENs and CRISPRs and should be part of any post-GWAS functional characterisation now that the cost and technical proficiency required is within the reach of most labs [196, 197]. These techniques enable a more biologically relevant assessment of SNP function in their native genomic context and can be extended further to mouse models, though inter-species differences limit how much information can be obtained from such comparisons [467]. Such techniques were technically challenging in this study due to the high levels of 11q13 amplification found in ERα positive breast cancer cell lines and the lack of availability of ERα positive normal breast cell lines. The experiments thus failed to yield useful information as to the *in vivo* effects of the SNPs in PRE1 (*Figure 3.17*) [8, 26]. Multiple cycles of editing may be required to create a sufficient number of risk alleles for a significant effect on gene expression to be seen and though possible, it is extremely labour and time intensive with the current available methods [197].

The 11q13 project described here provides proof of principle that low abundant transcripts expressed from noncoding regions should be sought using RNA Capture-seq at other GWAS loci as initially proposed by Mercer *et al*. [7]. Such an approach is validated by a recent RNA-Capture-seq study which discovered over 1000 novel transcripts within LD blocks containing GWAS risk SNPs, with many predicted to represent lncRNAs [468]. This demonstrates the way forward in uncovering lncRNAs associated with disease risk as previous studies have merely used existing databases of known transcripts and intersected these with GWAS SNPs or risk loci of interest [469, 470]. Many lncRNA transcripts that are tissue specific and lowly expressed are likely to be missed by this approach, even by those studies accessing extremely comprehensive RNA-seq datasets [240, 241]. The next step for the post-GWAS breast cancer community is to extend the pilot RNA Capture-seq

experiment described in Chapter 4 to cover all known breast cancer loci, providing a valuable resource for future follow up studies.

Chapter 5 examined the role of the lncRNAs as potential oncogenes influencing breast cancer progression. If *CUPID2* is confirmed as an independent driver of breast cancer then it may have a role as a biomarker for ERα positive disease. The use of lncRNAs as tumour biomarkers in bodily fluids or tissue is an area of active research and shows great promise for future management of cancer [241, 471, 472]. As a therapeutic target, the tissue specificity of lncRNAs such as *CUPID2* would allow drugs to have a more selective effect and hence reduce the effects on non-breast cancer cells which has proved an issue with previous attempts to target genes such as *CCND1* and *MYC* [233, 396, 473]. A case in point is the lncRNA *PVT1* which is required for the functional upregulation of *MYC* and is highly specific to *MYC* amplified cancer cells. Drugs specifically targeting *PVT1* would therefore bypass the toxicity seen with direct *MYC* inhibition [422]. It would be interesting to determine whether *CUPID2* overexpression is similarly required to co-exist with *CCND1* overexpression or whether the two are merely co-expressed due their co-regulation by PRE1.

Targeting interactions between lncRNAs such as *CUPID2* and their protein binding partners is of particular interest as this represents a promising therapeutic approach which would theoretically reduce the off target effects seen with targeting the protein or lncRNA in isolation [473, 474]. The field is in its infancy however and a substantial amount of work is required to provide the necessary structural information on the lncRNA-protein interaction and also the initial identification of the proteins involved for a specific lncRNA of interest (reviewed in [473]). Another approach under investigation is the injection of antisense oligonucleotides (ASO) as has been explored in a mouse model of Angelman syndrome where silencing of the lncRNA *UBE3A-ATS* led to an improvement in cognitive functioning [475]. Overexpression of *MALAT* in lung cancer has also been reduced by subcutaneous ASO injection leading to reduced proliferation of tumour cells [436]. One of the few methods to target a lncRNA trialed in the clinical setting thus far, used the injection of a plasmid expressing diphtheria toxin under control of the lncRNA *H19* promoter directly into ovarian, bladder and pancreatic tumours [476]. This caused localized production of the toxin only in the cells which pathologically overexpressed *H19*. SiRNA may also be used to directly target an lncRNA as was demonstrated in chapter 4 for *CUPID1* and *CUPID2* and work is ongoing to optimize delivery systems for this method *in vivo* [423, 473, 477].

Another treatment possibility is based on the observation that PRE1 has many properties consistent with a super enhancer including high levels of p300 and master TFs such as ERα and FoxA1 [135]; DNase I hypersensitivity levels a magnitude higher than the average enhancer [275]; and it is part of a DNA region containing multiple enhancer elements with epigenetic marks consistent with activity in breast tissue [103, 277]. Super enhancers are predicted to emerge near genes that contribute to the Hanahan and Weinberg 'hallmarks of cancer' which would include the cell cycle regulator *CCND1* [101, 290]. This is the case in colorectal cancer where such an enhancer is created at the 11q13 locus to drive the over expression of *CCND1* and subsequent cell proliferation [101]. This is of more than just academic interest as super enhancers are also characterized by their exquisite sensitivity to inhibition of the transcriptional co-activator BRD4 which induces a dramatic reduction in expression levels of key super-enhancer driven oncogenes such as *MYC* [104]. Super enhancers have also been targeted by genome editing strategies which caused collapse of the super enhancer cluster and up to 85% reduction in expression of the target oncogene [105, 478]. Further studies are required to demonstrate whether this will translate to a significant effect on tumour growth in the clinical setting, however initial work on BRD4 inhibitors has shown promise in reducing cancer cell growth *in vitro* [479, 480]. Given that the presence of PRE1 is required for significant expression of *CUPID1* and *CUPID2* (*Figures 4.14 and 4.15*), the effect of such BRD4 inhibitors on PRE1 activity and subsequent *CUPID1, CUPID2* and *CCND1* expression would be interesting to explore.

This thesis proposed that common genetic variants at 11q13 increase breast cancer risk by disrupting long-range regulatory elements and that the target genes (coding or non-coding) have a role in the pathogenesis of breast cancer. With the release of COGS data and the many functional post-GWAS studies which have been carried out on breast cancer it has become clear that the majority of risk at GWAS loci is mediated through the effects of SNPs on regulatory elements (reviewed in [45, 481]). Chapter Three explored different approaches to finding genes that may interact with the risk-SNP containing regulatory elements PRE1 and PRE2. As a result, two genes (*CPT1A* and *IGHMBP2*) potentially mediating risk at the locus were prioritized for further follow up. Chapter Four used RNA Capture-seq to identify non-coding transcripts in the 11q13 region on the basis that such transcripts may have important contributions to human disease but are often lowly expressed and missed by standard RNA-seq [7]. Two novel lncRNAs, *CUPID1* and *CUPID2* were characterized and evidence provided that their expression could be reduced by the risk-SNPs. They were also shown to have a role in DNA damage repair which is a common feature of genes previously shown to mediate the genetic risk of breast cancer [2]. Chapter Five explored the biological significance of *CUPID1* and *CUPID2* in relation to breast cancer progression as the importance of lncRNAs is being increasingly recognized in this regard [241, 355, 423]. *CUPID2* was found to drive cell proliferation *in vitro* and promote tumour growth in a murine xenograft model raising the possibility that it represents a novel oncogene.

The results obtained thus support the original hypothesis by demonstrating that the SNPs at PRE1 alter the regulation of target protein coding (*CPT1A* and *IGHMBP2)* and non-coding (*CUPID1* and *CUPID2)* genes. Evidence is also provided that the novel non-coding genes (*CUPID1* and *CUPID2)* have a role in breast cancer, though further work is required in this regard. Overall, this thesis highlights the need for breast cancer researchers to look more broadly at the potential genes mediating risk in GWAS loci and not be tempted to focus on the nearest or most obvious gene target. The approaches described here can be expanded upon for future post-GWAS studies, with a broader application of techniques such as RNA Capture-seq and variants of the chromosome conformation capture method. The goal of such work is to further our understanding of the underlying mechanisms behind breast cancer initiation and progression [26, 34]. This will lead to superior targeted therapies to reduce mortality and an improved ability to personalise screening and prevention programs based on individual genetic risk. Such a dual pronged approach to reduce the development of breast cancer and optimize the treatment of existing tumours will ultimately lead to better outcomes for breast cancer patients.

1. Ferlay J., S.H.R., Bray F., Forman D., Mathers C. and Parkin D.M. *GLOBOCAN 2008, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10* 2008 15/1/13].

2. Mavaddat, N., et al., *Genetic susceptibility to breast cancer.* Mol Oncol, 2010. **4**(3): p. 174-91.

3. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

4. Turnbull, C., et al., *Genome-wide association study identifies five new breast cancer susceptibility loci.* Nat Genet, 2010. **42**(6): p. 504-7.

5. French, J.D., et al., *Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers.* American Journal of Human Genetics, 2013. **92**(4): p. 489-503.

6. Sanyal, A., et al., *The long-range interaction landscape of gene promoters.* Nature, 2012. **489**(7414): p. 109-13.

7. Mercer, T.R., et al., *Targeted RNA sequencing reveals the deep complexity of the human transcriptome.* Nat Biotechnol, 2012. **30**(1): p. 99-104.

8. Wilkerson, P.M. and J.S. Reis-Filho, *the 11q13-q14 amplicon: Clinicopathological correlations and potential drivers.* Genes Chromosomes Cancer, 2012. **52**(4): p. 333-355.

9. Lichtenstein, P., et al., *Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland.* New England Journal of Medicine, 2000. **343**(2): p. 78-85.

10. Fisher, B., et al., *Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer.* New England Journal of Medicine, 2002. **347**(16): p. 1233-1241.

11. Clarke, M., et al., *Tamoxifen for early breast cancer: An overview of the randomised trials.* Lancet, 1998. **351**(9114): p. 1451-1467.

12. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-752.

13. Dawson, S.J., et al., *A new genome-driven integrated classification of breast cancer and its implications.* Embo Journal, 2013. **32**(5): p. 617-628.

14. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.

15. Bange, J., E. Zwick, and A. Ullrich, *Molecular targets for breast cancer therapy and prevention.* Nature Medicine, 2001. **7**(5): p. 548-552.

16. Molina, M.A., et al., *Trastuzumab (Herceptin), a humanized anti-HER2 receptor monoclonal antibody, inhibits basal and activated HER2 ectodomain cleavage in breast cancer cells.* Cancer Research, 2001. **61**(12): p. 4744-4749.

17. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* New England Journal of Medicine, 2001. **344**(11): p. 783-792.

18. Miller, K., et al., *Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer.* New England Journal of Medicine, 2007. **357**(26): p. 2666-2676.

19. Varghese, J.S. and D.F. Easton, *Genome-wide association studies in common cancers--what have we learnt?* Curr Opin Genet Dev, 2010. **20**(3): p. 201-9.

20. O'Donovan, P.J. and D.M. Livingston, *BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair.* Carcinogenesis, 2010. **31**(6): p. 961-967.

21. Antoniou, A., et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies.* American Journal of Human Genetics, 2003. **72**(5): p. 1117-1130.

22. Turnbull, C. and N. Rahman, *Genetic predisposition to breast cancer: past, present, and future.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 321-45.

23. Antoniou, A., et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies.* Am J Hum Genet, 2003. **72**(5): p. 1117-30.

24. Pharoah, P.D., et al., *Polygenic susceptibility to breast cancer and implications for prevention.* Nat Genet, 2002. **31**(1): p. 33-6.

25. Houlston, R.S. and J. Peto, *The search for low-penetrance cancer susceptibility alleles.* Oncogene, 2004. **23**(38): p. 6471-6476.

26. Edwards, S.L., et al., *Beyond GWASs: Illuminating the Dark Road from Association to Function.* American Journal of Human Genetics, 2013. **93**(5): p. 779-797.

27. Servin, B. and M. Stephens, *Imputation-based analysis of association studies: Candidate regions and quantitative traits.* Plos Genetics, 2007. **3**(7): p. 1296-1308.

28. Manolio, T.A., *Genomewide Association Studies and Assessment of the Risk of Disease.* New England Journal of Medicine, 2010. **363**(2): p. 166-176.

29. Hardy, J. and A. Singleton, *CURRENT CONCEPTS Genomewide Association Studies and Human Disease.* New England Journal of Medicine, 2009. **360**(17): p. 1759-1768.

30. Easton, D.F. and R.A. Eeles, *Genome-wide association studies in cancer.* Human Molecular Genetics, 2008. **17**: p. R109-R115.

31. Chatterjee, N., et al., *Analysis of Case-Control Association Studies: SNPs, Imputation and Haplotypes.* Statistical Science, 2009. **24**(4): p. 489-502.

32. Udler, M.S., J. Tyrer, and D.F. Easton, *Evaluating the Power to Discriminate Between Highly Correlated SNPs in Genetic Association Studies.* Genetic Epidemiology, 2010. **34**(5): p. 463-468.

33. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-265.

34. Freedman, M.L., et al., *Principles for the post-GWAS functional characterization of cancer risk loci.* Nat Genet, 2011. **43**(6): p. 513-8.

35. Michailidou, K., et al., *Large-scale genotyping identifies 41 new loci associated with breast cancer risk.* Nature Genetics, 2013. **45**(4): p. 353-361.

36. Michailidou, K., et al., *Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer.* Nature Genetics, 2015. **47**(4): p. 373-U127.

37. Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(7148): p. 1087-93.

38. Ghoussaini, M., et al., *Genome-wide association analysis identifies three new breast cancer susceptibility loci.* Nat Genet, 2012. **44**(3): p. 312-8.

39. Ahmed, S., et al., *Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2.* Nat Genet, 2009. **41**(5): p. 585-90.

40. Haiman, C.A., et al., *A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer.* Nat Genet, 2011. **43**(12): p. 1210-4.

41. Stacey, S.N., et al., *Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.* Nat Genet, 2007. **39**(7): p. 865-9.

42. Stacey, S.N., et al., *Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer.* Nat Genet, 2008. **40**(6): p. 703-6.

43. Thomas, G., et al., *A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1).* Nat Genet, 2009. **41**(5): p. 579-84.

44. Zheng, W., et al., *Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1.* Nat Genet, 2009. **41**(3): p. 324-8.

45. Maxwell, K.N. and K.L. Nathanson, *Common breast cancer risk variants in the post-COGS era: a comprehensive review.* Breast Cancer Research, 2013. **15**(6): p. 212-229.

46. Cox, A., et al., *A common coding variant in CASP8 is associated with breast cancer risk.* Nat Genet, 2007. **39**(3): p. 352-8.

47. Antoniou, A.C., et al., *A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population.* Nat Genet, 2010. **42**(10): p. 885-92.

48. Haiman, C.A., et al., *A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer.* Nature Genetics, 2011. **43**(12): p. 1210-U61.

49. Fletcher, O., et al., *Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study.* J Natl Cancer Inst, 2011. **103**(5): p. 425-35.

50. Cai, Q.Y., et al., *Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium.* Human Molecular Genetics, 2011. **20**(24): p. 4991-4999.

51. Siddiq, A., et al., *A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11.* Human Molecular Genetics, 2012. **21**(24): p. 5373-5384.

52. Long, J., et al., *Genome-Wide Association Study in East Asians Identifies Novel Susceptibility Loci for Breast Cancer.* Plos Genetics, 2012. **8**(2): p. e1002532.

53. Kim, H.-c., et al., *A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study.* Breast Cancer Research, 2012. **14**(2): p. R56.

54. Couch, F.J., et al., *Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk.* Plos Genetics, 2013. **9**(3): p. e1003212.

55. Gaudet, M.M., et al., *Identification of a BRCA2-Specific Modifier Locus at 6p24 Related to Breast Cancer Risk.* Plos Genetics, 2013. **9**(3): p. e1003173.

56. Garcia-Closas, M., et al., *Genome-wide association studies identify four ER negative-specific breast cancer risk loci.* Nature Genetics, 2013. **45**(4): p. 392-398.

57. Cai, Q.Y., et al., *Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1.* Nature Genetics, 2014. **46**(8): p. 886-890.

58. Milne, R.L., et al., *Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042.* J Natl Cancer Inst, 2009. **101**(14): p. 1012-8.

59. Glubb, D.M., et al., *Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1.* American Journal of Human Genetics, 2015. **96**(1): p. 5-20.

60. Ghoussaini, M., et al., *Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation.* Nature Communications, 2014. **5**: p. 4999.

61. Lin, W.Y., et al., *Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk.* Human Molecular Genetics, 2015. **24**(1): p. 285-298.

62. Meyer, K.B., et al., *Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1.* American Journal of Human Genetics, 2013. **93**(6): p. 1046-1060.

63. Maston, G.A., S.K. Evans, and M.R. Green, *Transcriptional regulatory elements in the human genome.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 29-59.

64. Lenhard, B., A. Sandelin, and P. Carninci, *REGULATORY ELEMENTS Metazoan promoters: emerging characteristics and insights into transcriptional regulation.* Nature Reviews Genetics, 2012. **13**(4): p. 233-245.

65. Henriques, T., et al., *Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals.* Molecular Cell, 2013. **52**(4): p. 517-528.

66. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.* Science, 2008. **322**(5909): p. 1845-1848.

67. Almada, A.E., et al., *Promoter directionality is controlled by U1 snRNP and polyadenylation signals.* Nature, 2013. **499**(7458): p. 360-U141.

68. Duttke, S.H.C., et al., *Human Promoters Are Intrinsically Directional.* Molecular Cell, 2015. **57**(4): p. 674-684.

69. Preker, P., et al., *RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters.* Science, 2008. **322**(5909): p. 1851-1854.

70. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-12.

71. Heidari, N., et al., *Genome-wide map of regulatory interactions in the human genome.* Genome Research, 2014. **24**(12): p. 1905-1917.

72. Zhang, Y.B., et al., *Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations.* Nature, 2013. **504**(7479): p. 306-310.

73. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes & Development, 2002. **16**(1): p. 6-21.

74. Herman, J.G. and S.B. Baylin, *Mechanisms of disease: Gene silencing in cancer in association with promoter hypermethylation.* New England Journal of Medicine, 2003. **349**(21): p. 2042-2054.

75. Forrest, A.R.R., et al., *A promoter-level mammalian expression atlas.* Nature, 2014. **507**(7493): p. 462-470.

76. Barolo, S., *Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy.* Bioessays, 2012. **34**(2): p. 135-41.

77. Perissi, V., et al., *A corepressor/coactivator exchange complex required for transcriptional activation by nuclear receptors and other regulated transcription factors.* Cell, 2004. **116**(4): p. 511-526.

78. 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.

79. Shlyueva, D., G. Stampfel, and A. Stark, *Transcriptional enhancers: from properties to genome-wide predictions.* Nature Reviews Genetics, 2014. **15**(4): p. 272-286.

80. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote sv40 DNA-sequences.* Cell, 1981. **27**(2): p. 299-308.

81. Buecker, C. and J. Wysocka, *Enhancers as information integration hubs in development: lessons from genomics.* Trends Genet, 2012. **28**(6): p. 276-84.

82. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-283.

83. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.

84. Crawford, G.E., et al., *Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).* Genome Res, 2006. **16**(1): p. 123-31.

85. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.

86. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.

87. Hah, N., et al., *A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells.* Cell, 2011. **145**(4): p. 622-634.

88. Akhtar-Zaidi, B., et al., *Epigenomic enhancer profiling defines a signature of colon cancer.* Science, 2012. **336**(6082): p. 736-9.

89. Gaulton, K.J., et al., *A map of open chromatin in human pancreatic islets.* Nat Genet, 2010. **42**(3): p. 255-9.

90. Nord, A.S., et al., *Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development.* Cell, 2013. **155**(7): p. 1521-1531.

91. Fuxman Bass, J.I., et al., *Human gene-centered transcription factor networks for enhancers and disease variants.* Cell, 2015. **161**(3): p. 661-73.

92. Weedon, M.N., et al., *Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis.* Nature Genetics, 2014. **46**(1): p. 61-64.

93. Smemo, S., et al., *Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease.* Human Molecular Genetics, 2012. **21**(14): p. 3255-3263.

94. Vanderploeg, L.H.T., et al., *Gamma-beta-thalassaemia studies showing that deletion of the gamma-genes and delta-genes influences beta-globin gene-expression in man.* Nature, 1980. **283**(5748): p. 637-642.

95. Smith, E. and A. Shilatifard, *Enhancer biology and enhanceropathies.* Nature Structural & Molecular Biology, 2014. **21**(3): p. 210-219.

96. Perry, M.W., A.N. Boettiger, and M. Levine, *Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(33): p. 13570-13575.

97. Guerrero, L., et al., *Secondary enhancers synergise with primary enhancers to guarantee fine-tuned muscle gene expression.* Developmental Biology, 2010. **337**(1): p. 16-28.

98. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**(7314): p. 430-5.

99. Ing-Simmons, E., et al., *Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin.* Genome Research, 2015. **25**(4): p. 504-513.

100. Corradin, O., et al., *Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits.* Genome Research, 2014. **24**(1): p. 1-13.

101. Hnisz, D., et al., *Super-Enhancers in the Control of Cell Identity and Disease.* Cell, 2013. **155**(4): p. 934-947.

102. Pott, S. and J.D. Lieb, *What are super-enhancers?* Nature Genetics, 2015. **47**(1): p. 8-12.

103. Whyte, W.A., et al., *Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.* Cell, 2013. **153**(2): p. 307-319.

104. Loven, J., et al., *Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers.* Cell, 2013. **153**(2): p. 320-334.

105. Mansour, M.R., et al., *An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element.* Science, 2014. **346**(6215): p. 1373-1377.

106. Petrykowska, H.M., C.M. Vockley, and L. Elnitski, *Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus.* Genome Research, 2008. **18**(8): p. 1238-1246.

107. Hardison, R.C. and J. Taylor, *Genomic approaches towards finding cis-regulatory modules in animals.* Nat Rev Genet, 2012. **13**(7): p. 469-83.

108. Vokes, S.A., et al., *A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb.* Genes & Development, 2008. **22**(19): p. 2651-2663.

109. Phillips, J.E. and V.G. Corces, *CTCF: Master Weaver of the Genome.* Cell, 2009. **137**(7): p. 1194-1211.

110. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-1245.

111. Handoko, L., et al., *CTCF-mediated functional chromatin interactome in pluripotent cells.* Nat Genet, 2011. **43**(7): p. 630-8.

112. Witcher, M. and B.M. Emerson, *Epigenetic silencing of the p16(INK4a) tumor suppressor is associated with loss of CTCF binding and a chromatin boundary.* Mol Cell, 2009. **34**(3): p. 271-84.

113. Foster, S.A., et al., *Inactivation of p16 in human mammary epithelial cells by CpG island methylation.* Molecular and Cellular Biology, 1998. **18**(4): p. 1793-1801.

114. Butcher, D.T. and D.I. Rodenhiser, *Epigenetic inactivation of BRCA1 is associated with aberrant expression of CTCF and DNA methyltransferase (DNMT3B) in some sporadic breast tumours.* Eur J Cancer, 2007. **43**(1): p. 210-9.

115. Green, A.R., et al., *Loss of expression of chromosome 16q genes DPEP1 and CTCF in lobular carcinoma in situ of the breast.* Breast Cancer Res Treat, 2009. **113**(1): p. 59-66.

116. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre.* Nature, 2012. **485**(7398): p. 381-385.

117. Giorgetti, L., et al., *Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription.* Cell, 2014. **157**(4): p. 950-963.

118. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-25.

119. Dai, J.C., et al., *Systematical analyses of variants in CTCF-binding sites identified a novel lung cancer susceptibility locus among Chinese population.* Scientific Reports, 2015. **5**(7833).

120. Noordermeer, D. and W. de Laat, *Joining the Loops: beta-Globin Gene Regulation.* Iubmb Life, 2008. **60**(12): p. 824-833.

121. Li, Q.L., et al., *Locus control regions.* Blood, 2002. **100**(9): p. 3077-3086.

122. Tolhuis, B., et al., *Looping and interaction between hypersensitive sites in the active beta-globin locus.* Molecular Cell, 2002. **10**(6): p. 1453-1465.

123. Spilianakis, C.G., et al., *Interchromosomal associations between alternatively expressed loci.* Nature, 2005. **435**(7042): p. 637-645.

124. Bergsland, M., et al., *Sequentially acting Sox transcription factors in neural lineage development.* Genes & Development, 2011. **25**(23): p. 2453-2464.

125. Lin, Y.C., et al., *A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate.* Nature Immunology, 2010. **11**(7): p. 635-U109.

126. Eeckhoute, J., et al., *A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer.* Genes Dev, 2006. **20**(18): p. 2513-26.

127. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-26.

128. Heinz, S., et al., *Effect of natural genetic variation on enhancer selection and function.* Nature, 2013. **503**(7477): p. 487-492.

129. Carroll, J.S., et al., *Genome-wide analysis of estrogen receptor binding sites.* Nat Genet, 2006. **38**(11): p. 1289-97.

130. McDonnell, D.P. and J.D. Norris, *Connections and regulation of the human estrogen receptor.* Science, 2002. **296**(5573): p. 1642-1644.

131. Joseph, R., et al., *Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha.* Mol Syst Biol, 2010. **6**: p. 456-469.

132. Pan, Y.F., et al., *Regulation of Estrogen Receptor-mediated Long Range Transcription via Evolutionarily Conserved Distal Response Elements.* Journal of Biological Chemistry, 2008. **283**(47): p. 32977-32988.

133. Lupien, M., et al., *FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription.* Cell, 2008. **132**(6): p. 958-70.

134. Hurtado, A., et al., *FOXA1 is a key determinant of estrogen receptor function and endocrine response.* Nat Genet, 2011. **43**(1): p. 27-33.

135. Carroll, J.S., et al., *Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1.* Cell, 2005. **122**(1): p. 33-43.

136. Laganiere, J., et al., *From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response.* Proc Natl Acad Sci U S A, 2005. **102**(33): p. 11651-6.

137. Serandour, A.A., et al., *Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers.* Genome Res, 2011. **21**(4): p. 555-65.

138. Koboldt, D.C., et al., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

139. Tan, S.K., et al., *AP-2gamma regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription.* EMBO J, 2011. **30**(13): p. 2569-81.

140. Eckert, D., et al., *The AP-2 family of transcription factors.* Genome Biol, 2005. **6**(13): p. 246-256.

141. Xu, J., R.C. Wu, and B.W. O'Malley, *Normal and cancer-related functions of the p160 steroid receptor co-activator (SRC) family.* Nat Rev Cancer, 2009. **9**(9): p. 615-30.

142. Eeckhoute, J., et al., *Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer.* Cancer Res, 2007. **67**(13): p. 6477-83.

143. Ross-Innes, C.S., et al., *Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer.* Genes Dev, 2010. **24**(2): p. 171-82.

144. Theodorou, V., et al., *GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility.* Genome Res, 2013. **23**(1): p. 12-22.

145. Vance, K.W. and C.P. Ponting, *Transcriptional regulatory functions of nuclear long noncoding RNAs.* Trends in Genetics, 2014. **30**(8): p. 348-355.

146. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Hum Mol Genet, 2003. **12**(14): p. 1725-35.

147. Kadauke, S. and G.A. Blobel, *Chromatin loops in gene regulation.* Biochim Biophys Acta, 2009. **1789**(1): p. 17-25.

148. Deng, W., et al., *Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor.* Cell, 2012. **149**(6): p. 1233-44.

149. Sexton, T., F. Bantignies, and G. Cavalli, *Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation.* Semin Cell Dev Biol, 2009. **20**(7): p. 849-55.

150. Bretschneider, N., et al., *E2-mediated cathepsin D (CTSD) activation involves looping of distal enhancer elements.* Mol Oncol, 2008. **2**(2): p. 182-90.

151. Deng, W.L., et al., *Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor.* Cell, 2012. **149**(6): p. 1233-1244.

152. Carter, D., et al., *Long-range chromatin regulatory interactions in vivo.* Nature Genetics, 2002. **32**(4): p. 623-626.

153. Palstra, R.J., et al., *The beta-globin nuclear compartment in development and erythroid differentiation.* Nature Genetics, 2003. **35**(2): p. 190-194.

154. Jin, F.L., et al., *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, 2013. **503**(7475): p. 290-294.

155. Sotelo, J., et al., *Long-range enhancers on 8q24 regulate c-Myc.* Proc Natl Acad Sci U S A, 2010. **107**(7): p. 3001-5.

156. Saramaki, A., et al., *Cyclical chromatin looping and transcription factor association on the regulatory regions of the p21 (CDKN1A) gene in response to 1alpha,25-dihydroxyvitamin D3.* J Biol Chem, 2009. **284**(12): p. 8073-82.

157. Gong, F.R., et al., *The BCL2 gene is regulated by a special AT-rich sequence binding protein 1-mediated long range chromosomal interaction between the promoter and the distal element located within the 3'-UTR.* Nucleic Acids Research, 2011. **39**(11): p. 4640-4652.

158. Lomvardas, S., et al., *Interchromosornal interactions and olfactory receptor choice.* Cell, 2006. **126**(2): p. 403-413.

159. Bateman, J.R., J.E. Johnson, and M.N. Locke, *Comparing enhancer action in cis and in trans.* Genetics, 2012. **191**(4): p. 1143-55.

160. Patel, B., et al., *Aberrant TAL1 activation is mediated by an interchromosomal interaction in human T-cell acute lymphoblastic leukemia.* Leukemia, 2014. **28**(2): p. 349-361.

161. Fullwood, M.J., et al., *An oestrogen-receptor-alpha-bound human chromatin interactome.* Nature, 2009. **462**(7269): p. 58-64.

162. Dryden, N.H., et al., *Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C.* Genome Research, 2014. **24**(11): p. 1854-1868.

163. Visel, A., E.M. Rubin, and L.A. Pennacchio, *Genomic views of distant-acting enhancers.* Nature, 2009. **461**(7261): p. 199-205.

164. Daniel, B., G. Nagy, and L. Nagy, *The intriguing complexities of mammalian gene regulation: How to link enhancers to regulated genes. Are we there yet?* Febs Letters, 2014. **588**(15): p. 2379-2391.

165. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.

166. de Laat, W. and J. Dekker, *3C-based technologies to study the shape of the genome.* Methods, 2012. **58**(3): p. 189-91.

167. Dekker, J., *The three 'C' s of chromosome conformation capture: controls, controls, controls.* Nat Methods, 2006. **3**(1): p. 17-21.

168. Gibcus, J.H. and J. Dekker, *The Hierarchy of the 3D Genome.* Molecular Cell, 2013. **49**(5): p. 773-782.

169. Dean, A., *In the loop: long range chromatin interactions and gene regulation.* Briefings in Functional Genomics, 2011. **10**(1): p. 3-10.

170. Shifera, A.S. and J.A. Hardin, *Factors modulating expression of Renilla luciferase from control plasmids used in luciferase reporter gene assays.* Anal Biochem, 2010. **396**(2): p. 167-72.

171. Zhao, Z., et al., *Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.* Nat Genet, 2006. **38**(11): p. 1341-7.

172. van de Werken, H.J., et al., *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nat Methods, 2012. **9**(10): p. 969-72.

173. Jager, R., et al., *Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci.* Nat Commun, 2015. **6**: p. 6178-86.

174. Dostie, J., et al., *Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.* Genome Res, 2006. **16**(10): p. 1299-309.

175. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

176. Horike, S., et al., *Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome.* Nature Genetics, 2005. **37**(1): p. 31-40.

177. Li, G., et al., *Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation.* Cell, 2012. **148**(1-2): p. 84-98.

178. Li, G.L., et al., *Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation.* Cell, 2012. **148**(1-2): p. 84-98.

179. Khan, S., et al., *MicroRNA Related Polymorphisms and Breast Cancer Risk.* Plos One, 2014. **9**(11): p. e109973.

180. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nature Biotechnology, 2010. **28**(10): p. 1045-1048.

181. Zhou, X., et al., *Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser.* Nature Biotechnology, 2015. **33**(4): p. 345-346.

182. Ward, L.D. and M. Kellis, *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.* Nucleic Acids Research, 2012. **40**(D1): p. D930-D934.

183. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Research, 2012. **22**(9): p. 1790-1797.

184. Li, M.J., et al., *GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications.* Nucleic Acids Research, 2013. **41**(W1): p. W150-W158.

185. Coetzee, S.G., et al., *FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs.* Nucleic Acids Research, 2012. **40**(18).

186. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.

187. Natarajan, A., et al., *Predicting cell-type-specific gene expression from regions of open chromatin.* Genome Research, 2012. **22**(9): p. 1711-1722.

188. Rao, S.S.P., et al., *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.* Cell, 2014. **159**(7): p. 1665-1680.

189. Corradin, O. and P.C. Scacheri, *Enhancer variants: evaluating functions in common disease.* Genome Medicine, 2014. **6**(10:85).

190. Li, Q., et al., *Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci.* Cell, 2013. **152**(3): p. 633-641.

191. Huang, D. and I. Ovcharenko, *Identifying causal regulatory SNPs in ChIP-seq enhancers.* Nucleic Acids Research, 2015. **43**(1): p. 225-236.

192. Smith, A.J.P. and S.E. Humphries, *Characterization of DNA-Binding Proteins Using Multiplexed Competitor EMSA.* Journal of Molecular Biology, 2009. **385**(3): p. 714-717.

193. Cowper-Sal lari, R., et al., *Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression.* Nat Genet, 2012. **44**(11): p. 1191-8.

194. Wan, Y., et al., *Landscape and variation of RNA secondary structure across the human transcriptome.* Nature, 2014. **505**(7485): p. 706-709.

195. Bernardo, B.C., et al., *A MicroRNA Guide for Clinicians and Basic Scientists: Background and Experimental Techniques.* Heart Lung and Circulation, 2012. **21**(3): p. 131-142.

196. Gaj, T., C.A. Gersbach, and C.F. Barbas, *ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering.* Trends in Biotechnology, 2013. **31**(7): p. 397-405.

197. Boettcher, M. and M.T. McManus, *Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR.* Molecular Cell, 2015. **58**(4): p. 575-585.

198. Bauer, D.E., et al., *An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level.* Science, 2013. **342**(6155): p. 253-257.

199. Soldner, F., et al., *Generation of Isogenic Pluripotent Stem Cells Differing Exclusively at Two Early Onset Parkinson Point Mutations.* Cell, 2011. **146**(2): p. 318-331.

200. Roy, P.G. and A.M. Thompson, *Cyclin D1 and breast cancer.* Breast, 2006. **15**(6): p. 718-27.

201. Jirawatnotai, S., et al., *A function for cyclin D1 in DNA repair uncovered by protein interactome analyses in human cancers.* Nature, 2011. **474**(7350): p. 230-4.

202. Li, Z.P., et al., *Cyclin D1 Integrates Estrogen-Mediated DNA Damage Repair Signaling.* Cancer Research, 2014. **74**(14): p. 3959-3970.

203. Chung, S.Y., et al., *Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility.* Cancer Science, 2011. **102**(1): p. 245-252.

204. Xiang, J.-F., et al., *Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus.* Cell Research, 2014. **24**(5): p. 513-531.

205. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

206. Halvardson, J., A. Zaghlool, and L. Feuk, *Exome RNA sequencing reveals rare and novel alternative transcripts.* Nucleic Acids Res, 2013. **41**(1): p. e6.

207. Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.

208. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

209. Taft, R.J., et al., *Non-coding RNAs: regulators of disease.* Journal of Pathology, 2010. **220**(2): p. 126-139.

210. Tang, J., A. Ahmad, and F.H. Sarkar, *The Role of MicroRNAs in Breast Cancer Migration, Invasion and Metastasis.* International Journal of Molecular Sciences, 2012. **13**(10): p. 13414-13437.

211. Li, L. and H.Y. Chang, *Physiological roles of long noncoding RNAs: insight from knockout mice.* Trends in Cell Biology, 2014. **24**(10): p. 594-602.

212. Clark, M.B., et al., *The reality of pervasive transcription.* PLoS Biol, 2011. **9**(7): p. e1000625; discussion e1001102.

213. van Bakel, H., et al., *Most "dark matter" transcripts are associated with known genes.* PLoS Biol, 2010. **8**(5): p. e1000371.

214. Mattick, J.S., *The Genetic Signatures of Noncoding RNAs.* Plos Genetics, 2009. **5**(4): p. e1000459.

215. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.

216. Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation.* Nature, 2011. **477**(7364): p. 295-U60.

217. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Research, 2007. **17**(5): p. 556-565.

218. Natoli, G. and J.C. Andrau, *Noncoding transcription at enhancers: general principles and functional models.* Annu Rev Genet, 2012. **46**: p. 1-19.

219. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & Development, 2011. **25**(18): p. 1915-1927.

220. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression.* Genome Research, 2012. **22**(9): p. 1775-1789.

221. Kornienko, A.E., et al., *Gene regulation by the act of long non-coding RNA transcription.* Bmc Biology, 2013. **11**(1): p. 59.

222. Khalil, A.M., et al., *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(28): p. 11667-11672.

223. Rinn, J. and M. Guttman, *RNA and dynamic nuclear organization.* Science, 2014. **345**(6202): p. 1240-1241.

224. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.

225. Tuan, D., S.M. Kong, and K. Hu, *Transcription of the hypersensitive site hs2 enhancer in erythroid-cells.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(23): p. 11219-11223.

226. De Santa, F., et al., *A large fraction of extragenic RNA pol II transcription sites overlap enhancers.* PLoS Biol, 2010. **8**(5): p. e1000384.

227. Collis, P., M. Antoniou, and F. Grosveld, *Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high-level expression.* Embo Journal, 1990. **9**(1): p. 233-240.

228. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-7.

229. Kowalczyk, M.S., et al., *Intragenic enhancers act as alternative promoters.* Mol Cell, 2012. **45**(4): p. 447-58.

230. Onodera, C.S., et al., *Gene isoform specificity through enhancer-associated antisense transcription.* PLoS One, 2012. **7**(8): p. e43511.

231. Wang, D., et al., *Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA.* Nature, 2011. **474**(7351): p. 390-4.

232. Lam, M.T.Y., et al., *Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription.* Nature, 2013. **498**(7455): p. 511-515.

233. Melo, C.A., et al., *eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription.* Molecular Cell, 2013. **49**(3): p. 524-535.

234. Li, W.B., et al., *Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation.* Nature, 2013. **498**(7455): p. 516-+.

235. Arner, E., et al., *Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells.* Science, 2015. **347**(6225): p. 1010-1014.

236. Pefanis, E., et al., *RNA Exosome-Regulated Long Non-Coding RNA Transcription Controls Super-Enhancer Activity.* Cell, 2015. **161**(4): p. 774-789.

237. Wang, X., et al., *Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription.* Nature, 2008. **454**(7200): p. 126-30.

238. Leveille, N., C.A. Melo, and R. Agami, *Enhancer-associated RNAs as therapeutic targets.* Expert Opinion on Biological Therapy, 2015. **15**(5): p. 723-734.

239. Kumar, V., et al., *Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression.* Plos Genetics, 2013. **9**(1).

240. Hangauer, M.J., I.W. Vaughn, and M.T. McManus, *Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs.* Plos Genetics, 2013. **9**(6).

241. Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nature Genetics, 2015. **47**(3): p. 199-208.

242. Helgadottir, A., et al., *The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm.* Nature Genetics, 2008. **40**(2): p. 217-224.

243. Pasmant, E., et al., *ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS.* Faseb Journal, 2011. **25**(2): p. 444-448.

244. Glinskii, A.B., et al., *Identification of intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell cycle progression/differentiation pathways in multiple common human disorders.* Cell Cycle, 2009. **8**(23): p. 3925-3942.

245. Glinskii, A.B., et al., *Networks of intergenic long-range enhancers and snpRNAs drive castration-resistant phenotype of prostate cancer and contribute to pathogenesis of multiple common human disorders.* Cell Cycle, 2011. **10**(20): p. 3571-97.

246. Gibb, E.A., C.J. Brown, and W.L. Lam, *The functional role of long non-coding RNA in human carcinomas.* Mol Cancer, 2011. **10**(1): p. 38.

247. Sirchia, S.M., et al., *Loss of the inactive X chromosome and replication of the active X in BRCA1-defective and wild-type breast cancer cells.* Cancer Res, 2005. **65**(6): p. 2139-46.

248. Brown, C.J., et al., *A gene from the region of the human x-inactivation center is expressed exclusively from the inactive x-chromosome.* Nature, 1991. **349**(6304): p. 38-44.

249. Sirchia, S.M., et al., *Misbehaviour of XIST RNA in breast cancer cells.* PLoS One, 2009. **4**(5): p. e5559.

250. Berteaux, N., et al., *H19 mRNA-like noncoding RNA promotes breast cancer cell proliferation through positive control by E2F1.* J Biol Chem, 2005. **280**(33): p. 29625-36.

251. Guffanti, A., et al., *A transcriptional sketch of a primary human breast cancer by 454 deep sequencing.* BMC Genomics, 2009. **10**: p. 163.

252. Iacoangeli, A., et al., *BC200 RNA in invasive and preinvasive breast cancer.* Carcinogenesis, 2004. **25**(11): p. 2125-33.

253. Mourtada-Maarabouni, M., et al., *GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer.* Oncogene, 2009. **28**(2): p. 195-208.

254. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.

255. 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.

256. Su, X., et al., *Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes.* Oncotarget, 2014. **5**(20): p. 9864-9876.

257. Sorensen, K.P., et al., *Long non-coding RNA expression profiles predict metastasis in lymph node-negative breast cancer independently of traditional prognostic markers.* Breast Cancer Research, 2015. **17**: p. 55.

258. Thongjuea, S., et al., *r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data.* Nucleic Acids Research, 2013. **41**(13).

259. Mercer, T.R., et al., *Targeted sequencing for gene discovery and quantification using RNA CaptureSeq.* Nature Protocols, 2014. **9**(5): p. 989-1009.

260. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (vol 7, pg 562, 2012).* Nature Protocols, 2014. **9**(10): p. 2513-2513.

261. Chu, C., et al., *Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions.* Molecular Cell, 2011. **44**(4): p. 667-678.

262. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

263. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.

264. Vance, K.W., et al., *The long non-coding RNA Paupar regulates the expression of both local and distal genes.* Embo Journal, 2014. **33**(4): p. 296-311.

265. Geissmann, Q., *OpenCFU, a New Free and Open-Source Software to Count Cell Colonies and Other Circular Objects.* Plos One, 2013. **8**(2): p. e54072.

266. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

267. Jiang, L., et al., *Oral cancer overexpressed 1 (ORAOV1) regulates cell cycle and apoptosis in cervical cancer HeLa cells.* Mol Cancer, 2010. **9**: p. 20.

268. Janssen, J.W.G., et al., *Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t(11;14)(q13;q32).* Blood, 2000. **95**(8): p. 2691-2698.

269. Turner, N. and R. Grose, *Fibroblast growth factor signalling: from development to cancer.* Nature Reviews Cancer, 2010. **10**(2): p. 116-129.

270. Jiang, L., et al., *Oral cancer overexpressed 1 (ORAOV1): a regulator for the cell growth and tumor angiogenesis in oral squamous cell carcinoma.* Int J Cancer, 2008. **123**(8): p. 1779-86.

271. Schodel, J., et al., *Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression.* Nat Genet, 2012. **44**(4): p. 420-5, S1-2.

272. Chung, C.C., et al., *Fine mapping of a region of chromosome 11q13 reveals multiple independent loci associated with risk of prostate cancer.* Hum Mol Genet, 2011. **20**(14): p. 2869-78.

273. Simonis, M., et al., *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).* Nat Genet, 2006. **38**(11): p. 1348-54.

274. van den Boogaard, M., et al., *A common genetic variant within SCN10A modulates cardiac SCN5A expression.* Journal of Clinical Investigation, 2014. **124**(4): p. 1844-1852.

275. He, H.H., et al., *Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics.* Genome Res, 2012. **22**(6): p. 1015-25.

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

277. Myers, R.M., et al., *A User's Guide to the Encyclopedia of DNA Elements (ENCODE).* Plos Biology, 2011. **9**(4): p. e1001046.

278. van de Werken, H.J.G., et al., *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nature Methods, 2012. **9**(10): p. 969-972.

279. Blankenberg, D. and J. Hillman-Jackson, *Analysis of Next-Generation Sequencing Data Using Galaxy*, in *Stem Cell Transcriptional Networks: Methods and Protocols*, B.L. Kidder, Editor. 2014. p. 21-43.

280. Hsu, P.Y., et al., *Amplification of Distant Estrogen Response Elements Deregulates Target Genes Associated with Tamoxifen Resistance in Breast Cancer.* Cancer Cell, 2013. **24**(2): p. 197-212.

281. Wu, G.J., et al., *Structural analysis of the 17q22-23 amplicon identifies several independent targets of amplification in breast cancer cell lines and tumors.* Cancer Research, 2001. **61**(13): p. 4951-4955.

282. Clark, M.B., et al., *Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing.* Nature Methods, 2015. **12**(4): p. 339-U92.

283. Dean, N.M. and R. McKay, *Inhibition of protein-kinase c-alpha expression in mice after systemic administration of phosphorothioate antisense oligodeoxynucleotides.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(24): p. 11762-11766.

284. Shadeo, A. and W.L. Lam, *Comprehensive copy number profiles of breast cancer cell model genomes.* Breast Cancer Research, 2006. **8**(1): p. R9.

285. Bradic, M., J. Costa, and I.M. Chelo, *Genotyping with Sequenom.* Molecular Methods for Evolutionary Genetics, 2011. **772**: p. 193-210.

286. Lanz, R.B., et al., *Global Characterization of Transcriptional Impact of the SRC-3 Coregulator.* Molecular Endocrinology, 2010. **24**(4): p. 859-872.

287. Ghoussaini, M., et al., *Multiple loci with different cancer specificities within the 8q24 gene desert.* J Natl Cancer Inst, 2008. **100**(13): p. 962-6.

288. Wright, J.B., S.J. Brown, and M.D. Cole, *Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells.* Mol Cell Biol, 2010. **30**(6): p. 1411-20.

289. Wasserman, N.F., I. Aneas, and M.A. Nobrega, *An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer.* Genome Res, 2010. **20**(9): p. 1191-7.

290. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

291. Sutherland, H. and W.A. Bickmore, *Transcription factories: gene expression in unions?* Nat Rev Genet, 2009. **10**(7): p. 457-66.

292. Gheldof, N., et al., *Detecting long-range chromatin interactions using the chromosome conformation capture sequencing (4C-seq) method.* Methods Mol Biol, 2012. **786**: p. 211-25.

293. Splinter, E., et al., *Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation.* Methods, 2012. **58**(3): p. 221-30.

294. Gao, F., et al., *The interactomes of POU5F1 and SOX2 enhancers in human embryonic stem cells.* Scientific Reports, 2013. **3**: p. R1588.

295. Wei, Z., et al., *Klf4 Organizes Long-Range Chromosomal Interactions with the Oct4 Locus inReprogramming andPluripotency.* Cell Stem Cell, 2013. **13**(1): p. 36-47.

296. Apostolou, E., et al., *Genome-wide Chromatin Interactions of the Nanog Locus in Pluripotency, Differentiation, and Reprogramming.* Cell Stem Cell, 2013. **12**(6): p. 699-712.

297. Zeitz, M.J., et al., *Genomic Interaction Profiles in Breast Cancer Reveal Altered Chromatin Architecture.* Plos One, 2013. **8**(9): p. e73974.

298. Hampton, O.A., et al., *A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome.* Genome Research, 2009. **19**(2): p. 167-177.

299. Rondon-Lagos, M., et al., *Unraveling the chromosome 17 patterns of FISH in interphase nuclei: an in-depth analysis of the HER2 amplicon and chromosome 17 centromere by karyotyping, FISH and M-FISH in breast cancer cells.* Bmc Cancer, 2014. **14**: p. e922.

300. Roix, J.J., et al., *Spatial proximity of translocation-prone gene loci in human lymphomas.* Nature Genetics, 2003. **34**(3): p. 287-291.

301. Lin, C., et al., *Nuclear Receptor-Induced Chromosomal Proximity and DNA Breaks Underlie Specific Translocations in Cancer.* Cell, 2009. **139**(6): p. 1069-1083.

302. Belton, J.M., et al., *Hi-C: A comprehensive technique to capture the conformation of genomes.* Methods, 2012. **58**(3): p. 268-76.

303. van de Werken, H.J.G., et al., *4C Technology: Protocols and Data Analysis*, in *Nucleosomes, Histones & Chromatin, Pt B*, C. Wu and C.D. Allis, Editors. 2012. p. 89-112.

304. van de Werken, H.J.G., et al., *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nature Methods, 2012. **9**(10): p. 969-+.

305. Wiblin, A.E., et al., *Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells.* Journal of Cell Science, 2005. **118**(17): p. 3861-3868.

306. Wijchers, P.J., et al., *Characterization and dynamics of pericentromere-associated domains in mice.* Genome Research, 2015. **25**(7): p. 958-969.

307. Weierich, C., et al., *Three-dimensional arrangements of centromeres and telomeres in nuclei of human and murine lymphocytes.* Chromosome Research, 2003. **11**(5): p. 485-502.

308. Yasuhara, J.C. and B.T. Wakimoto, *Oxymoron no more: the expanding world of heterochromatic genes.* Trends in Genetics, 2006. **22**(6): p. 330-338.

309. Bickmore, W.A., *The Spatial Organization of the Human Genome*, in *Annual Review of Genomics and Human Genetics, Vol 14*, A. Chakravarti and E. Green, Editors. 2013. p. 67-84.

310. Nagano, T., et al., *Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.* Nature, 2013. **502**(7469): p. 59-64.

311. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-240.

312. Halvorsen, M., et al., *Disease-associated mutations that alter the RNA structural ensemble.* PLoS Genet, 2010. **6**(8): p. e1001074.

313. Dias, N. and C.A. Stein, *Antisense oligonucleotides: Basic concepts and mechanisms.* Molecular Cancer Therapeutics, 2002. **1**(5): p. 347-355.

314. Bonnefont, J.-P., et al., *Carnitine palmitoyltransferases 1 and 2: biochemical, molecular and medical aspects.* Molecular aspects of medicine, 2004. **25**(5-6): p. 495-520.

315. Gatza, M.L., et al., *An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer.* Nature Genetics, 2014. **46**(10): p. 1051-1059.

316. Fukita, Y., et al., *The human s-mu-bp-2, a DNA-binding protein-specific to the single-stranded guanine-rich sequence related to the immunoglobulin-mu chain switch region.* Journal of Biological Chemistry, 1993. **268**(23): p. 17463-17470.

317. Mizuta, T.R., et al., *Isolation of cdna-encoding a binding-protein specific to 5'-phosphorylated single-stranded-DNA with g-rich sequences.* Nucleic Acids Research, 1993. **21**(8): p. 1761-1766.

318. Savas, S., et al., *Identifying functional genetic variants in DNA repair pathway using protein conservation analysis.* Cancer Epidemiology Biomarkers & Prevention, 2004. **13**(5): p. 801-807.

319. Elliott, B., et al., *Gene conversion tracts from double-strand break repair in mammalian cells.* Molecular and Cellular Biology, 1998. **18**(1): p. 93-101.

320. Hsu, P.D., et al., *DNA targeting specificity of RNA-guided Cas9 nucleases.* Nature Biotechnology, 2013. **31**(9): p. 827-832.

321. Li, N., et al., *A Polymorphism rs12325489C > T in the LincRNA-ENST00000515084 Exon Was Found to Modulate Breast Cancer Risk via GWAS-Based Association Analyses.* Plos One, 2014. **9**(5): p. e98251.

322. Dinger, M.E., *lncRNAs: finding the forest among the trees?* Mol Ther, 2011. **19**(12): p. 2109-11.

323. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.

324. Ghoussaini, M., et al., *Multiple loci with different cancer specificities within the 8q24 gene desert.* Journal of the National Cancer Institute, 2008. **100**(13): p. 962-966.

325. Kim, T., et al., *Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(11): p. 4173-4178.

326. Kawaji, H., et al., *CAGE basic/analysis databases: the CAGE resource for comprehensive promoter analysis.* Nucleic Acids Research, 2006. **34**: p. D632-D636.

327. Sabo, P.J., et al., *Discovery of functional noncoding elements by digital analysis of chromatin structure.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(48): p. 16837-16842.

328. Wright, M.W., *A short guide to long non-coding RNA gene nomenclature.* Human Genomics, 2014. **8**: p. e7.

329. Tang, K., et al., *Cloning and Functional Characterization of a Novel Long Non-coding RNA Gene Associated With Hepatocellular Carcinoma.* Progress in Biochemistry and Biophysics, 2014. **41**(2): p. 153-162.

330. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Research, 2012. **40**(D1): p. D84-D90.

331. Joseph, R., et al., *Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha.* Molecular Systems Biology, 2010. **6**: p. e456.

332. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

333. Carroll, J.S. and M. Brown, *Estrogen receptor target gene: an evolving concept.* Mol Endocrinol, 2006. **20**(8): p. 1707-14.

334. Bertoli, C., J.M. Skotheim, and R.A.M. de Bruin, *Control of cell cycle transcription during G1 and S phases.* Nature Reviews Molecular Cell Biology, 2013. **14**(8): p. 518-528.

335. Polo, S.E. and S.P. Jackson, *Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications.* Genes & Development, 2011. **25**(5): p. 409-433.

336. Huang, F., et al., *Identification of Specific Inhibitors of Human RAD51 Recombinase Using High-Throughput Screening.* Acs Chemical Biology, 2011. **6**(6): p. 628-635.

337. Ahsan, H., et al., *A Genome-wide Association Study of Early-Onset Breast Cancer Identifies PFKM as a Novel Breast Cancer Gene and Supports a Common Genetic Spectrum for Breast Cancer at Any Age.* Cancer Epidemiology Biomarkers & Prevention, 2014. **23**(4): p. 658-669.

338. Kawaji, H., et al., *Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing.* Genome Research, 2014. **24**(4): p. 708-717.

339. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.

340. Osborne, C.K., et al., *Effects of tamoxifen on human-breast cancer cell-cycle kinetics - accumulation of cells in early g1-phase.* Cancer Research, 1983. **43**(8): p. 3583-3585.

341. Bhan, A., et al., *Antisense Transcript Long Noncoding RNA (lncRNA) HOTAIR is Transcriptionally Induced by Estradiol.* Journal of Molecular Biology, 2013. **425**(19): p. 3707-3722.

342. Reeder, C., et al., *High Resolution Mapping of Enhancer-Promoter Interactions.* Plos One, 2015. **10**(5): p. e0122420.

343. Nolis, I.K., et al., *Transcription factors mediate long-range enhancer-promoter interactions.* Proc Natl Acad Sci U S A, 2009. **106**(48): p. 20222-7.

344. Qiu, X.W., et al., *A complex deoxyribonucleic acid looping configuration associated with the silencing of the maternal Igf2 allele.* Molecular Endocrinology, 2008. **22**(6): p. 1476-1488.

345. Butter, F., et al., *Proteome-Wide Analysis of Disease-Associated SNPs That Show Allele-Specific Transcription Factor Binding.* Plos Genetics, 2012. **8**(9): p. e1002982.

346. Patwardhan, R.P., et al., *Massively parallel functional dissection of mammalian enhancers in vivo.* Nature Biotechnology, 2012. **30**(3): p. 265-270.

347. Cabili, M.N., et al., *Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution.* Genome Biology, 2015. **16**: p. e20.

348. Tripathi, V., et al., *The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation.* Molecular Cell, 2010. **39**(6): p. 925-938.

349. Guttman, M., et al., *Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins.* Cell, 2013. **154**(1): p. 240-251.

350. Chu, C., et al., *Systematic Discovery of Xist RNA Binding Proteins.* Cell, 2015. **161**(2): p. 404-416.

351. Ulveling, D., C. Francastel, and F. Hube, *When one is better than two: RNA with dual functions.* Biochimie, 2011. **93**(4): p. 633-644.

352. Geisler, S. and J. Coller, *RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts.* Nature Reviews Molecular Cell Biology, 2013. **14**(11): p. 699-712.

353. Wang, K.C., et al., *A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.* Nature, 2011. **472**(7341): p. 120-4.

354. Yoon, J.H., et al., *Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination.* Nature Communications, 2013. **4**: p. e2939.

355. Hu, X., et al., *A Functional Genomic Approach Identifies FAL1 as an Oncogenic Long Noncoding RNA that Associates with BMI1 and Represses p21 Expression in Cancer.* Cancer Cell, 2014. **26**(3): p. 344-357.

356. Gagnon, K.T., et al., *Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading.* Nature Protocols, 2014. **9**(9): p. 2045-2060.

357. Kim, T., et al., *Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(11): p. 4173-4178.

358. Bassett, A.R., et al., *Considerations when investigating lncRNA function in vivo.* Elife, 2014. **3**.

359. Ma, L., V.B. Bajic, and Z. Zhang, *On the classification of long non-coding RNAs.* Rna Biology, 2013. **10**(6): p. 925-934.

360. Kutter, C., et al., *Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression.* Plos Genetics, 2012. **8**(7): p. e1002841.

361. Yang, L., et al., *lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs.* Nature, 2013. **500**(7464): p. 598-602.

362. Guttman, M. and J.L. Rinn, *Modular regulatory principles of large non-coding RNAs.* Nature, 2012. **482**(7385): p. 339-346.

363. Lai, F., et al., *Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.* Nature, 2013. **494**(7438): p. 497-501.

364. Alao, J.P., *The regulation of cyclin D1 degradation: roles in cancer development and the potential for therapeutic invention.* Molecular Cancer, 2007. **6**: p. e24.

365. Rosenwald, I.B., et al., *Elevated levels of cyclin d1 protein in response to increased expression of eukaryotic initiation-factor 4e.* Molecular and Cellular Biology, 1993. **13**(12): p. 7358-7363.

366. Aleman, L.M., J. Doench, and P.A. Sharp, *Comparison of siRNA-induced off-target RNA and protein effects.* Rna-a Publication of the Rna Society, 2007. **13**(3): p. 385-395.

367. Simon, M.D., et al., *The genomic binding sites of a noncoding RNA.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(51): p. 20497-20502.

368. Engreitz, J.M., et al., *The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome.* Science, 2013. **341**: p. e123973.

369. West, J.A., et al., *The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites.* Molecular Cell, 2014. **55**(5): p. 791-802.

370. Toutenhoofd, S.L., et al., *Characterization of the human CALM2 calmodulin gene and comparison of the transcriptional activity of CALM1, CALM2 and CALM3.* Cell Calcium, 1998. **23**(5): p. 323-338.

371. Calaluce, R., et al., *The RNA binding protein HuR differentially regulates unique subsets of mRNAs in estrogen receptor negative and estrogen receptor positive breast cancer.* Bmc Cancer, 2010. **10**: p. e126.

372. Zhang, X.-Y., et al., *NRSN2 promotes non-small cell lung cancer cell growth through PI3K/Akt/mTOR pathway.* International Journal of Clinical and Experimental Pathology, 2015. **8**(3): p. 2574-2581.

373. Ma, H.-Q., et al., *Decreased expression of Neurensin-2 correlates with poor prognosis in hepatocellular carcinoma.* World Journal of Gastroenterology, 2009. **15**(38): p. 4844-4848.

374. Yu, H.Y., et al., *Genomic analysis of gene expression relationships in transcriptional regulatory networks.* Trends in Genetics, 2003. **19**(8): p. 422-427.

375. Pierce, A.J., et al., *XRCC3 promotes homology-directed repair of DNA damage in mammalian cells.* Genes & Development, 1999. **13**(20): p. 2633-2638.

376. Javle, M. and N.J. Curtin, *The role of PARP in DNA repair and its therapeutic exploitation.* British Journal of Cancer, 2011. **105**(8): p. 1114-1122.

377. Zender, L., et al., *Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach.* Cell, 2006. **125**(7): p. 1253-1267.

378. Kao, J. and J.R. Pollack, *RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes.* Genes Chromosomes & Cancer, 2006. **45**(8): p. 761-769.

379. Weinstein, J.N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nature Genetics, 2013. **45**(10): p. 1113-1120.

380. Du, Z., et al., *Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer.* Nature Structural & Molecular Biology, 2013. **20**(7): p. 908-913.

381. Ling, H., et al., *CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer.* Genome Research, 2013. **23**(9): p. 1446-1461.

382. Hung, T., et al., *Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters.* Nat Genet, 2011. **43**(7): p. 621-9.

383. Huarte, M., et al., *A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response.* Cell, 2010. **142**(3): p. 409-419.

384. Schafer, J.M., et al., *Analysis of cross-resistance of the selective estrogen receptor modulators arzoxifene (LY353381) and LY117018 in tamoxifen-stimulated breast cancer xenografts.* Clinical Cancer Research, 2001. **7**(8): p. 2505-2512.

385. Hoffmann, J., et al., *Characterization of new estrogen receptor destabilizing compounds: Effects on estrogen-sensitive and tamoxifen-resistant breast cancer.* Journal of the National Cancer Institute, 2004. **96**(3): p. 210-218.

386. Hartman, J., et al., *Estrogen receptor beta inhibits angiogenesis and growth of T47D breast cancer xenografts.* Cancer Research, 2006. **66**(23): p. 11207-11213.

387. Herbst, F., et al., *Extensive Methylation of Promoter Sequences Silences Lentiviral Transgene Expression During Stem Cell Differentiation In Vivo.* Molecular Therapy, 2012. **20**(5): p. 1014-1021.

388. Weinstein, I.B., *Disorders in cell circuitry during multistage carcinogenesis: the role of homeostasis.* Carcinogenesis, 2000. **21**(5): p. 857-864.

389. Rafehi, H., et al., *Clonogenic Assay: Adherent Cells.* Jove-Journal of Visualized Experiments, 2011(49).

390. van Meerloo, J., G.J.L. Kaspers, and J. Cloos, *Cell Sensitivity Assays: The MTT Assay*, in *Cancer Cell Culture: Methods and Protocols, Second Edition*, I.A. Cree, Editor. 2011. p. 237-245.

391. Kim, J.B., M.J. O'Hare, and R. Stein, *Models of breast cancer: is merging human and animal models the future?* Breast Cancer Research, 2004. **6**(1): p. 22-30.

392. Holliday, D.L. and V. Speirs, *Choosing the right cell line for breast cancer research.* Breast Cancer Research, 2011. **13**(4): p. e215.

393. Beroukhim, R., et al., *The landscape of somatic copy-number alteration across human cancers.* Nature, 2010. **463**(7283): p. 899-905.

394. Holm, K., et al., *Characterisation of amplification patterns and target genes at chromosome 11q13 in CCND1-amplified sporadic and familial breast tumours.* Breast Cancer Research and Treatment, 2012. **133**(2): p. 583-594.

395. Reis, J.S., et al., *Cyclin D1 protein overexpression and CCND1 amplification in breast carcinomas: an immunohistochemical and chromogenic in situ hybridisation analysis.* Modern Pathology, 2006. **19**(7): p. 999-1009.

396. Casimiro, M.C., et al., *Overview of cyclins D1 function in cancer and the CDK inhibitor landscape: past and present.* Expert Opinion on Investigational Drugs, 2014. **23**(3): p. 295-304.

397. Korkaya, H., et al., *HER2 regulates the mammary stem/progenitor cell population driving tumorigenesis and invasion.* Oncogene, 2008. **27**(47): p. 6120-6130.

398. Holley, S.L., J. Heighway, and P.R. Hoban, *Induced expression of human CCND1 alternative transcripts in mouse Cyl-1 knockout fibroblasts highlights functional differences.* International Journal of Cancer, 2005. **114**(3): p. 364-370.

399. Zhang, Y., et al., *Circular Intronic Long Noncoding RNAs.* Molecular Cell, 2013. **51**(6): p. 792-806.

400. Yin, Q.F., et al., *Long noncoding RNAs with snoRNA ends.* Mol Cell, 2012. **48**(2): p. 219-30.

401. Desai, K.V., et al., *Initiating oncogenic event determines gene-expression patterns of human breast cancer models.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(10): p. 6967-6972.

402. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.

403. Aapro, M.S., et al., *Colony formation invitro as a prognostic indicator for primary breast-cancer.* Journal of Clinical Oncology, 1987. **5**(6): p. 890-896.

404. Vargo-Gogola, T. and J.M. Rosen, *Modelling breast cancer: one size does not fit all.* Nature Reviews Cancer, 2007. **7**(9): p. 659-672.

405. Vire, E., et al., *The Breast Cancer Oncogene EMSY Represses Transcription of Antimetastatic microRNA miR-31.* Molecular Cell, 2014. **53**(5): p. 806-818.

406. Kang, J.S., et al., *Low Dose Estrogen Supplementation Reduces Mortality of Mice in Estrogen-Dependent Human Tumor Xenograft Model.* Biological & Pharmaceutical Bulletin, 2009. **32**(1): p. 150-152.

407. Fridman, R., et al., *Increased initiation and growth of tumor cell lines, cancer stem cells and biopsy material in mice using basement membrane matrix protein (Cultrex or Matrigel) co-injection.* Nature Protocols, 2012. **7**(6): p. 1138-1144.

408. Ito, M., et al., *NOD/SCID/gamma(null)(c) mouse: an excellent recipient mouse model for engraftment of human cells.* Blood, 2002. **100**(9): p. 3175-3182.

409. Iorns, E., et al., *A New Mouse Model for the Study of Human Breast Cancer Metastasis.* Plos One, 2012. **7**(10): p. e47995.

410. Albertson, D.G., *Gene amplification in cancer.* Trends in Genetics, 2006. **22**(8): p. 447-455.

411. Santarius, T., et al., *Epigenetics and genetics: a census of amplified and overexpressed human cancer genes.* Nature Reviews Cancer, 2010. **10**(1): p. 59-64.

412. Liu, L.L., et al., *Carnitine palmitoyltransferase 1A (CPT1A): a transcriptional target of PAX3-FKHR and mediates PAX3-FKHR-dependent motility in alveolar rhabdomyosarcoma cells.* Bmc Cancer, 2012. **12**: p. e154.

413. Pacilli, A., et al., *Carnitine-Acyltransferase System Inhibition, Cancer Cell Death, and Prevention of Myc-Induced Lymphomagenesis.* Jnci-Journal of the National Cancer Institute, 2013. **105**(7): p. 489-498.

414. Carracedo, A., L.C. Cantley, and P.P. Pandolfi, *Cancer metabolism: fatty acid oxidation in the limelight.* Nature Reviews Cancer, 2013. **13**(4): p. 227-232.

415. Currie, E., et al., *Cellular Fatty Acid Metabolism and Cancer.* Cell Metabolism, 2013. **18**(2): p. 153-161.

416. Ren, X.R., et al., *Perhexiline promotes HER3 ablation through receptor internalization and inhibits tumor growth.* Breast Cancer Research, 2015. **17**.

417. Zhang, B., et al., *Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk.* Nature Genetics, 2014. **46**(6): p. 533-542.

418. Gold, B., et al., *Genome-wide association study provides evidence for a breast cancer risk locus at 6q22-33.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(11): p. 4340-4345.

419. Guenther, U.P., et al., *IGHMBP2 is a ribosome-associated helicase inactive in the neuromuscular disorder distal SMA type 1 (DSMA1).* Human Molecular Genetics, 2009. **18**(7): p. 1288-1300.

420. Shen, J., et al., *IGHMBP2 Thr671Ala polymorphism might be a modifier for the effects of cigarette smoking and PAH-DNA adducts to breast cancer risk.* Breast Cancer Research and Treatment, 2006. **99**(1): p. 1-7.

421. Mendenhall, E.M., et al., *Locus-specific editing of histone modifications at endogenous enhancers.* Nature Biotechnology, 2013. **31**(12): p. 1133-+.

422. Tseng, Y.-Y., et al., *PVT1 dependence in cancer with MYC copy-number increase.* Nature, 2014. **512**(7512): p. 82-+.

423. Cheetham, S.W., et al., *Long noncoding RNAs and the genetics of cancer.* British Journal of Cancer, 2013. **108**(12): p. 2419-2425.

424. Gong, J., et al., *lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse.* Nucleic Acids Research, 2015. **43**(D1): p. D181-D186.

425. Mirza, A.H., et al., *Effects of GWAS-Associated Genetic Variants on lncRNAs within IBD and T1D Candidate Loci.* Plos One, 2014. **9**(8): p. e105723.

426. Chen, G., et al., *Genome-Wide Analysis of Human SNPs at Long Intergenic Noncoding RNAs.* Human Mutation, 2013. **34**(2): p. 338-344.

427. Ning, S., et al., *A global map for dissecting phenotypic variants in human lincRNAs.* European Journal of Human Genetics, 2013. **21**(10): p. 1128-1133.

428. Visser, M., R.-J. Palstra, and M. Kayser, *Allele-specific transcriptional regulation of IRF4 in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the IRF4 promoter.* Human Molecular Genetics, 2015. **24**(9): p. 2649-2661.

429. Meijers-Heijboer, H., et al., *Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations.* Nature Genetics, 2002. **31**(1): p. 55-59.

430. Renwick, A., et al., *ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles.* Nature Genetics, 2006. **38**(8): p. 873-875.

431. Popadin, K., et al., *Genetic and Epigenetic Regulation of Human lincRNA Gene Expression.* American Journal of Human Genetics, 2013. **93**(6): p. 1015-1026.

432. Yang, T.-P., et al., *Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies.* Bioinformatics, 2010. **26**(19): p. 2474-2476.

433. Rinn, J.L., *lncRNAs: Linking RNA to Chromatin.* Cold Spring Harbor Perspectives in Biology, 2014. **6**(8): p. a018614.

434. Trimarchi, T., et al., *Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia.* Cell, 2014. **158**(3): p. 593-606.

435. Gutschner, T., M. Baas, and S. Diederichs, *Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases.* Genome Res, 2011. **21**(11): p. 1944-54.

436. Gutschner, T., et al., *The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells.* Cancer Research, 2013. **73**(3): p. 1180-1189.

437. Schmidt, L.H., et al., *The Long Noncoding MALAT-1 RNA Indicates a Poor Prognosis in Non-small Cell Lung Cancer and Induces Migration and Tumor Growth.* Journal of Thoracic Oncology, 2011. **6**(12): p. 1984-1992.

438. Tano, K., et al., *MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes.* Febs Letters, 2010. **584**(22): p. 4575-4580.

439. Dobrovic, A. and D. Simpfendorfer, *Methylation of the BRCA1 gene in sporadic breast cancer.* Cancer Research, 1997. **57**(16): p. 3347-3350.

440. Schild, D. and C. Wiese, *Overexpression of RAD51 suppresses recombination defects: a possible mechanism to reverse genomic instability.* Nucleic Acids Research, 2010. **38**(4): p. 1061-1070.

441. Wilson, C.A., et al., *Localization of human BRCA1 and its loss in high-grade, non-inherited breast carcinomas.* Nature Genetics, 1999. **21**(2): p. 236-240.

442. Maacke, H., et al., *Over-expression of wild-type Rad51 correlates with histological grading of invasive ductal breast cancer.* International Journal of Cancer, 2000. **88**(6): p. 907-913.

443. Ward, A., K.K. Khanna, and A.P. Wiegmans, *Targeting homologous recombination, new pre-clinical and clinical therapeutic combinations inhibiting RAD51.* Cancer Treatment Reviews, 2015. **41**(1): p. 35-45.

444. Farmer, H., et al., *Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy.* Nature, 2005. **434**(7035): p. 917-921.

445. Reintjes, N., et al., *Activating Somatic FGFR2 Mutations in Breast Cancer.* Plos One, 2013. **8**(3): p. e60264.

446. Zhang, J.H., et al., *Use of Genome-Wide Association Studies for Cancer Research and Drug Repositioning.* Plos One, 2015. **10**(3).

447. Sanseau, P., et al., *Use of genome-wide association studies for drug repositioning.* Nature Biotechnology, 2012. **30**(4): p. 317-320.

448. Jirstrom, K., et al., *Adverse effect of adjuvant tamoxifen in premenopausal breast cancer with cyclin D1 gene amplification.* Cancer Research, 2005. **65**(17): p. 8009-8016.

449. Palmieri, C., et al., *Breast cancer: Current and future endocrine therapies.* Molecular and Cellular Endocrinology, 2014. **382**(1): p. 695-723.

450. Shou, J., et al., *Mechanisms of tamoxifen resistance: Increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer.* Journal of the National Cancer Institute, 2004. **96**(12): p. 926-935.

451. Luo, J., N.L. Solimini, and S.J. Elledge, *Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction.* Cell, 2009. **136**(5): p. 823-837.

452. Weinstein, I.B., *Cancer: Addiction to oncogenes - The Achilles heal of cancer.* Science, 2002. **297**(5578): p. 63-64.

453. Jain, M., et al., *Sustained loss of a neoplastic phenotype by brief inactivation of MYC.* Science, 2002. **297**(5578): p. 102-104.

454. Bernard-Pierrot, I., et al., *Characterization of the recurrent 8p11-12 amplicon identifies PPAPDC1B, a phosphatase protein, as a new therapeutic target in breast cancer.* Cancer Research, 2008. **68**(17): p. 7165-7175.

455. Reis, J.S., et al., *FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas.* Clinical Cancer Research, 2006. **12**(22): p. 6652-6662.

456. Turner, N., et al., *Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets.* Oncogene, 2010. **29**(14): p. 2013-2023.

457. Seibler, J., et al., *Reversible gene knockdown in mice using a tight, inducible shRNA expression system.* Nucleic Acids Research, 2007. **35**(7): p. e54.

458. Yang, X., et al., *Long non-coding RNA HNF1A-AS1 regulates proliferation and migration in oesophageal adenocarcinoma cells.* Gut, 2014. **63**(6): p. 881-890.

459. Veron, A., S. Blein, and D.G. Cox, *Genome-wide association studies and the clinic: a focus on breast cancer.* Biomarkers in Medicine, 2014. **8**(2): p. 287-296.

460. Huesing, A., et al., *Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status.* Journal of Medical Genetics, 2012. **49**(9): p. 601-608.

461. Wacholder, S., et al., *Performance of Common Genetic Variants in Breast-Cancer Risk Models.* New England Journal of Medicine, 2010. **362**(11): p. 986-993.

462. Mavaddat, N., et al., *Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants.* Jnci-Journal of the National Cancer Institute, 2015. **107**(5): p. djc036.

463. Easton, D.F., et al., *Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk.* New England Journal of Medicine, 2015. **372**(23): p. 2243-2257.

464. Ferraiuolo, M.A., et al., *From cells to chromatin: Capturing snapshots of genome organization with 5C technology.* Methods, 2012. **58**(3): p. 255-67.

465. Hughes, J.R., et al., *Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment.* Nature Genetics, 2014. **46**(2): p. 205-212.

466. Du, M.J., et al., *Prostate cancer risk locus at 8q24 as a regulatory hub by physical interactions with multiple genomic loci across the genome.* Human Molecular Genetics, 2015. **24**(1): p. 154-166.

467. Ermann, J. and L.H. Glimcher, *After GWAS: mice to the rescue?* Current Opinion in Immunology, 2012. **24**(5): p. 564-570.

468. Bartonicek, N.C., MB; Maag, JL; Gloss, BS; Mattick, JS; Dinger, ME., *Uncovering hidden genes in intergenic regions.* Poster Presentation, Lorne Genome, 2015.

469. Jin, G., et al., *Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk.* Carcinogenesis, 2011. **32**(11): p. 1655-1659.

470. Hrdlickova, B., et al., *Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity.* Genome Medicine, 2014. **6**: p. e88.

471. Cui, W., et al., *Discovery and characterization of long intergenic non-coding RNAs (lincRNA) module biomarkers in prostate cancer: an integrative analysis of RNA-Seq data.* BMC genomics, 2015. **16 Suppl 7**: p. S3-S3.

472. Qi, P. and X. Du, *The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine.* Modern Pathology, 2013. **26**(2): p. 155-165.

473. Fatemi, R.P., D. Velmeshev, and M.A. Faghihi, *De-repressing LncRNA-Targeted Genes to Upregulate Gene Expression: Focus on Small Molecule Therapeutics.* Molecular Therapy-Nucleic Acids, 2014. **3**: p. e196.

474. Mohamadkhani, A., *Long Noncoding RNAs in Interaction With RNA Binding Proteins in Hepatocellular Carcinoma.* Hepatitis Monthly, 2014. **14**(5): p. e18794.

475. Meng, L., et al., *Towards a therapy for Angelman syndrome by targeting a long non-coding RNA.* Nature, 2015. **518**(7539): p. 409-12.

476. Smaldone, M.C. and B.J. Davies, *BC-819, a plasmid comprising the H19 gene regulatory sequences and diphtheria toxin A, for the potential targeted therapy of cancers.* Current Opinion in Molecular Therapeutics, 2010. **12**(5): p. 607-616.

477. Oh, Y.-K. and T.G. Park, *siRNA delivery systems for cancer treatment.* Advanced Drug Delivery Reviews, 2009. **61**(10): p. 850-862.

478. Hnisz, D., et al., *Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers.* Molecular Cell, 2015. **58**(2): p. 362-370.

479. Garcia, P.L., et al., *JQ1 suppresses tumor growth in tumorgraft models of pancreatic ductal adenocarcinoma.* Cancer Research, 2014. **74**(19): p. S4612.

480. Lee, D.H., et al., *Synergistic effect of JQ1 and rapamycin for treatment of human osteosarcoma.* International Journal of Cancer, 2015. **136**(9): p. 2055-2064.

481. Fachal, L. and A.M. Dunning, *From candidate gene studies to GWAS and post-GWAS analyses in breast cancer.* Current opinion in genetics & development, 2015. **30**: p. 32-41.

8.1 Buffers

3C lysis buffer
10mM Tris pH 7.5
10mM NaCl
48.9ml of water
IGEPAL (Sigma Aldrich) to final 0.0004% concentration
1 tablet of 'Complete' protease inhibitor (Roche)

3C ligation buffer
660mM Tris pH 7.5
50mM DTT (dithiothreitol)
50mM $MgCl_2$
10mM ATP (freshly made from 100mM stock)
made up to 10ml total volume with dH_2O

ChIRP-seq lysis buffer
50mM Tris-Cl pH 7.0
10mM EDTA
1% SDS
additives only added fresh, just before use

ChIRP-seq proteinase K buffer (for DNA)
100mM NaCl
1mM EDTA
10mM TrisCl pH 8.0 (Use pH 7.0 for RNA)
add 5% by volume of Proteinase K (Ambion) 20mg/ml fresh before use

ChIRP-seq hybridization buffer
750mM NacCl
1% SDS
50mM TrisCl pH 7.0
1mM EDTA
15% formamide
additives fresh just before use

ChIRP-seq wash buffer
2x NaCl and Sodium citrate (SSC) diluted from 20x Invitrogen stock
0.5% SDS
fresh PMSF just before use

ChIRP-seq DNA elution buffer
50mM $NaHCO_3$
1% SDS

1:100 RNAseA 10mg/ml (Sigma)
1:50 RNAseH 5mg/ml (NEB)

ChIRP-seq additives for hybridisation and lysis buffers
1 pellet of Complete Protease Inhibitor (Roche) dissolved in 1ml water (50x stock)
100mM PMSF (Sigma) in isopropranol (100x stock)
Superase-in (Ambion) as 200x stock

Cell Fractionation Lysis Buffer
15mM HEPES pH7.5
10mM KCl
5mM $MgCl_2$
0.1mM EDTA
0.5mM EGTA
250mM Sucrose
0.4% Igepal
1mM DTT
40U/ml RNaseOUT (Invitrogen)
Protease inhibitor cocktail (Roche)

Cell Fractionation Nuclei Lysis Buffer
10mM HEPES pH7.5
0.1mM EDTA
0.1mM EGTA
1mM DTT
40U/ml RNaseOUT (Invitrogen)
Protease inhibitor cocktail (Roche)

Cell fractionation salt extraction buffer
25mM HEPES pH7.5
10% glycerol
420 mM NaCL
5mM $MgCl_2$
0.1mM EDTA
1mM DTT
40U/ml RNaseOUT (Invitrogen)
Protease inhibitor cocktail (Roche)

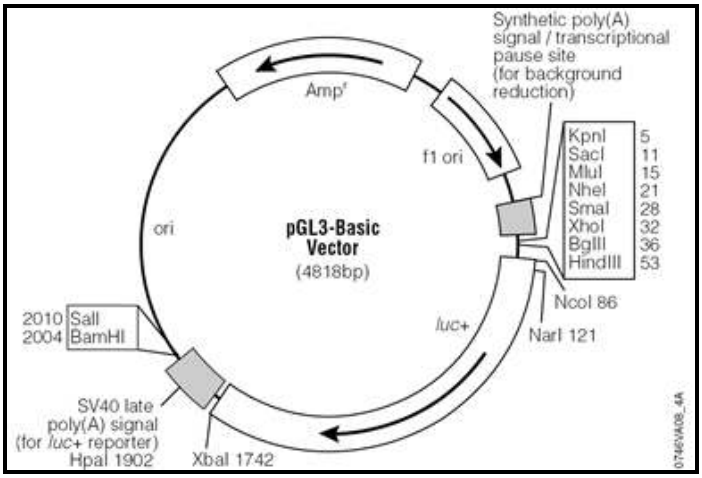
Immunofluorescence cytoskeleton buffer
10mM PIPES pH6.8
100mM NaCl
30mM Sucrose
3mM $MgCl_2$
1mM EGTA

Immunofluorescence cytoskeleton stripping buffer
10mM Tris HCl pH7.4
1mM NaCl
3mM $MgCl_2$
1% Tween-20
0.5% Sodium Deoxycholate

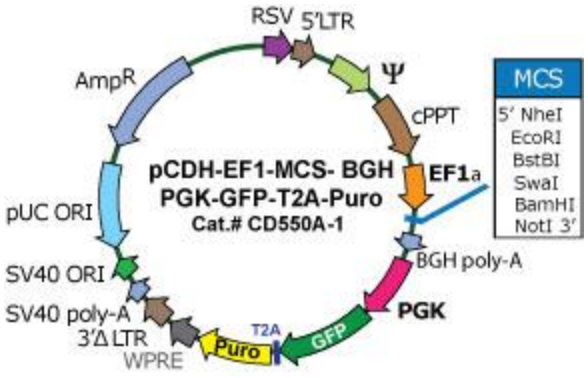
Immunofluorescence FBT buffer
10mM Tris HCl pH7.4
5% foetal bovine serum
1% bovine serum albumin
0.05% Tween-20
100mM $MgCl_2$
All resuspended in PBS and prepared fresh.

8.2 Plasmids

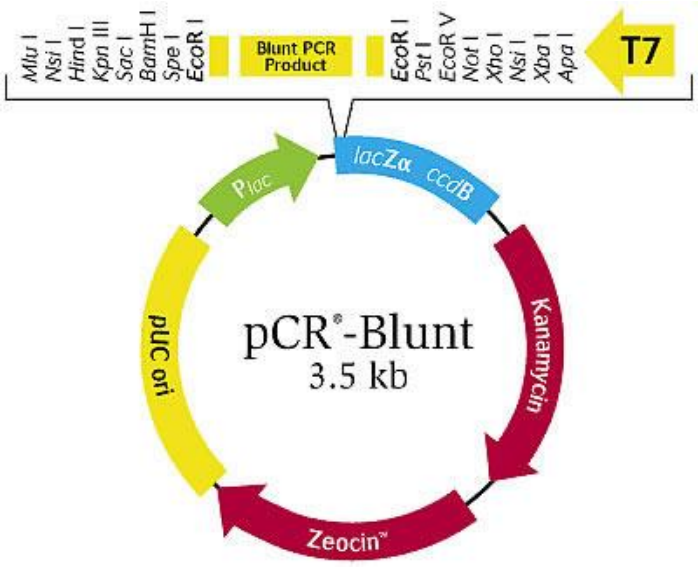
1. pGL3-Basic (Promega) for luciferase assays



2. pCDH (SystemBio) for lncRNA overexpression



3. pCR-Blunt (Life Technologies) subcloning vector



8.3 Taqman Gene Expression Assays

TaqMan Probes	
CCND1	Hs00765553_m1
ORAOV1	Hs00411598_m1
MYEOV	Hs00993153_g1
TFF1	Hs00907239_m1
GUS	Hs00939627_m1
FGF3	Hs00173742_m1
FGF4	Hs00999691_m1
FGF19	Hs00192780_m1
PPP6R3	Hs00217759_m1
PPFIA1	Hs01548999_m1
MTL5	Hs01127481_m1
MRPL21	Hs00698959_m1
CPT1A	Hs00912671_m1
IGHMBP2	Hs00158054_m1

Table 8.1 TaqMan probes. *Used with TaqMan gene expression master mix.*

8.4 siRNA Sequences

siRNA	
C1-1	GGAGAUCCCUACAGAGGAAUU
C1-2	GGAUGUGAAAUGAGGGAAAUU
C1-3	GCUUAUAAGUGGAGGAGGAUU
C2-1	UGUAAAGUGUGGUGAGCAUUU
C2-2	GGAUGUGAAAUGAGGGAAAUU
C2-3	GAGCAAUGUUACACGAUAAUU
C2-4	GCAGUGAGUCCCACGGUGAUU
C2-5	GGUCCUGAGGGAAGCGGAAUU
CCND1	GUUCGUGGCCUCUAAGAUG
eRNA-1	GCGAAAUGUUUAAAGGAAAUU
eRNA-2	CAUUCUUAAGGCUGCGAAAUU
Scrambled	UGGUUUACAUGUCGACUAA
Modified Oligos	
M-ASO-D	GUCUCCACCCCAGGAGUGAUCCUCU
M-ASO-E	CCUCCGAAACUCGUGUUGAAAUGUA
M-ASO scrambled	CCUCUUACCUCAGUUACAAUUUAUA

Table 8.2 siRNA and M-ASO sequences

8.5 ChIRP-seq probes

Chirp-seq Probes	
LACZ_1	CCAGTGAATCCGTAATCATG
LACZ_2	TCACGACGTTGTAAAACGAC
LACZ_3	ATTAAGTTGGGTAACGCCAG
LACZ_4	AGGTTACGTTGGTGTAGATG
LACZ_5	AATGTGAGCGAGTAACAACC
LACZ_6	GTAGCCAGCTTTCATCAACA
LACZ_7	AATAATTCGCGTCTGGCCTT
LACZ_8	AGATGAAACGCCGAGTTAAC
LACZ_9	AATTCAGACGGCAAACGACT
LACZ_10	TTTCTCCGGCGCGTAAAAAT
LACZ_11	ATCTTCCAGATAACTGCCGT
LACZ_12	AACGAGACGTCACGGAAAAT
LACZ_13	GCTGATTTGTGTAGTCGGTT
LACZ_14	TTAAAGCGAGTGGCAACATG
LACZ_15	AACTGTTACCCGTAGGTAGT
LACZ_16	ATAATTTCACCGCCGAAAGG

LNCRNA_1	TTGTATTTGCTGGATGGAGC
LNCRNA_2	CTTCGATGTCTTCATGTCTC
LNCRNA_3	ATCACAGGGTGTGTTTTGCA
LNCRNA_4	ATCTACTCCTTCATGCTGTG
LNCRNA_5	TCCAGAAGGTCCAAGTTCAT
LNCRNA_6	CTAAGCAAATGTTGACCAGG
LNCRNA_7	TCTCTCCAGTCTCTTCTCAT
LNCRNA_8	CACATCCACTCAGTTCAGAA
LNCRNA_9	AACCTTCATGCCTGCAAAGT
LNCRNA_10	TTCTATCCTCACATCTGCTC
LNCRNA_11	GCTTAAAACAGCACACGTGT
LNCRNA_12	ATGCGTCTTTCATCTCCATG
LNCRNA_13	CCCAGCCACTTTCTCTTTTT
LNCRNA_14	TTGGGTGCCATTTCTTTGAC
LNCRNA_15	ACCCAAACCAAGCTGACAAA
LNCRNA_16	GTGCTTTGGAAATAGTGCAG

Table 8.3 ChIRP-seq probes. *The lncRNA probes were split into even and odd pools for the experiment.*

8.6 Primer sequences (all primers presented 5' to 3')

Primer Name	Primer Sequence
***CUPID* promoter cloning primers**	
CUPID1cloningKpnF1	GGTACCGAGAGAAAAGAAAAAACAAGTGGGTGG
CUPID1cloningXhoR1	CTCGAGGCTTACGATCTCAAAGCTGTTAGAGACC
CUPID2cloningXhoF1	CTCGAGGAGAGAAAAGAAAAAACAAGTGGGTGG
CUPID2cloningKpnR1	GGTACCGCTTACGATCTCAAAGCTGTTAGAGACC
***CUPID* expression primers**	
CUPID2expressionF1	AGGCTAGGAAGATGTGCACCTGC
CUPID2expressionR1	CCTCTGCCTCTTCTCATCTGACG
CUPID1expressionF1	AGACCTGGTCAACATTTGCTTAGAACC
CUPID1expressionR1	TCACATCCACTCAGTTCAGAACCCTGC
***CUPID2* Cloning Primers**	
CUPID2NheIF1	GCTAGCAGACTTTGCCCTCACAGGCAAG
CUPID2NotIR1	GCGGCCGCGCGGTTTAAATCGCAAACATTTATTTC
Gene expression primers	
HOTAIR expressionF1	GCAGTGGGGAACTCTGACTCG
HOTAIR expressionR1	TCAGTGCCTGGTGCTCTCTTACC

MALAT expressionF1	GAATTGCGTCATTTAAAGCCTAGTTAAC
MALAT expressionR1	GTTTCATCCTACCACTCCCAATTAATC
GAPDH expressionF1	GAAGGTGAAGGTCGGAGTCAACG
GAPDH expressionR1	GCCATGGGTGGAATCATATTGG
TBP expressionF1	CATGGATCAGAACAACAGCCTGC
TBP expressionR1	TTGTGAGAGTCTGTGAGTGGAAGAGC
BACTIN expressionF1	GGAAATCGTGCGTGACATTAAGG
BACTIN expressionR1	AGTACTTGCGCTCAGGAGGAGC

3C NCO PRIMERS FOR *CUPID* PROMOTER

11q13Ncofrag9up4F1	ATTTGTGCAGAGGAGACCCAGAAAGCAGG
11q13Ncofrag9up3R1	GAAGTGCCTCCTTGTCTGCCTTCCTGG
11q13Ncofrag9up2R1	GAGACAGCATCTGATTTGGGAAGCGACC
11q13Ncofrag9up1R1	GTCCTCGGTGCACAGGCTGTAATCAGC
11q13Ncofrag9int1R1	CTCCTGGAATAGAGATGACACCGTGCTGC
11q13Ncofrag9int1F1	CTGATGCAGCTCACTGTGGCAAAAGACC
11q13Ncofrag9dn1F1	AGTCCCAGCAATGCTACCCAGAGCAGG
11q13Ncofrag9dn2R1	CAAAAGGGACAATTGCAGACCATCACGC
11q13Ncofrag9dn3F1	TCCATTGAAGCCTCTGCCAGCCTCC
11q13Ncofrag9dn3R1	CTTGCTGTGTCTCCCTCTTACCCTGGTCC
11q13Ncofrag9dn4R1	TAATCCCTGTCCACTGTGCTCCAAATGTCC
NcoDpnDigEffREV	GTGCTGTTCTGGGACAATCTGG
NcoDpnDigEffFOR	CAGCATGCCCTCAGTTCTCG
3CGAPDHConFor	TGATGACATCAAGAAGGTGGTGAAGCAGG
3CGAPDHConRev	GAAATGAGCTTGACAAAGTGGTCGTTGAGG

TALEN

TalenbreakF1	TCTTGCACTGCCTTGAGACTTGG
TalenbreakR1	GGTGACTCACACAAGTGGTCTTCC

Sequenom probes

rs78540526_F	ACGTTGGATGCCTCTAAACTGCAGCAGTTG
rs554219_F	ACGTTGGATGAGGTGCTGGGTTGACTGTG
rs661204_F	ACGTTGGATGTGTCCTCTCCAAATCTCAC
rs78540526_R	ACGTTGGATGATGTTCCCATGGACCTGAGC
rs554219_R	ACGTTGGATGTGTTGTGTGATTCCACTCCC
rs661204_R	ACGTTGGATGAGCCATTCATGAGGAGTCCA
rs78540526_E	AGCTCTTCCCAGCAC
rs554219_E	GCAGGGAAATCCTCAC
rs661204_E	TGGTGGAAAGAGTTTTGAT

***CUPID EcoRI* 3C primers**

11q13EcoRIPRE1intF1	TTAGCCCAAGCATCTCTTCCCTGGATGG
11q13EcoRIup4R1	AGGAGCTCTCTATTCAGCAGGTCACCTCAGC
11q13EcoRIup5R1	CCAACCACTGTGTGGGTTACAGCAAAGAGG
11q13EcoRIup6R1	CCATCCCACCACAGAGCCTGATATCTGG
11q13EcoRIup7R1	TATGACTCCAACCCTGTGACATACGGAAAAGG

11q13EcoRIup8R1	CCTGGGGTCTTCGATGTCTTCATGTCTCC
11q13EcoRIup9R1	CTTGCTAGGGGTCAGCACTGTTCCAATATCC
11q13EcoRIup10R1	GGTGACGTGTGCATGCGATGTATCTGG
11q13EcoRIup11R1	CTCCTGAGTCCCAAAACCTGTCCTCACC
11q13EcoRIup12R1	ACTCCGTGACCACCCTGTCACCTTGTACC
11q13EcoRIup13R1	CTCCCACCACTCTGACTACCTGCTTCTGTCC
11q13EcoRIup14R1	GAATGGGACCCAGGATACCTGCTTCCTATACC
11q13EcoRIup15R1	TCAAGTGGGTGGCTGCACTAGAGGTCC
11q13EcoRIup16R1	CTCGAGGAAGTGAAAGCCAGATGGTGC
11q13EcoRIup17R1	GATGAAGTGATTGTGTGGCTGACAGAATGG
11q13EcoRIup18F1	TGTGGGACCCTTTGCAGATTACTCTGAACC
11q13EcoRIup19R1	TAGGGAGTGCTGTGAAGGGTGCATTTGG
11q13EcoRIup21R1	AGGTGTTAGGACCTTGAAGTCAGGACAGTTGG
11q13EcoRIup22R1	AAGGAGGAAAACGTAGCTCTCAGACCAAGTCC
11q13EcoRIup23R1	GCATTGTGCACCTACTGCACACCAGG
11q13EcoRIup24R1	GGGACATTTTAGACGTGGGAAGGAGAGTGG
11q13EcoRIup25F1	CAAAGACCCCTCCTCACATAACCCTGAGC
11q13EcoRIup26F1	CGTGAGAGTAACGCTGACCTCATGGAACG
11q13EcoRIup28R1	TTCTCCTACTTCCCCAAGTCAGTCTCCCTGC
11q13EcoRIup29F1	GGTGCGATGACTGGAAAGGATGCTGG
11q13EcoRIup31F1	AATGCCTGCTATTTAGGCACTCGTGTGTGG
11q13EcoRIup34F1	AAGGCACTTCCACTGTGGGAGAGAGAGG
11q13EcoRIup35F1	CTTTAAGTCTTGGGTGGGAACTCAAACTGTGG
11q13EcoRIup36F1	GGAAATCCCTCCTGCATCCCTCTTCC
11q13EcoRIup38F1	GTCCACAGCAACTGACGCACAGAGAAACC
11q13EcoRIup41F1	GGATGGTGTTCACCCACCCTGAGAAGG
11q13EcoRIup42R1	CACACCACACCACGGATGAATTTCAAAGC
11q13EcoRIup43R1	AGCAGTTTTATCAACCTGTGAGCAGCCTCTCC
11q13EcoRIup44F1	CTGCTGGTATGCTCTGTGCCTTCCTTGG
11q13EcoRIup45F1	ACAGATGGTGCCAAAGAGAAGCCCAGG
11q13EcoRIPRE2F1	GGAAATGGCTTTGAACACCATCCACAGG
11q13EcoRIPRE2R1	AATGATGGCAGAGTTCTTATGGGCACTTTGG

PRE1 eRNA

PRE1eRNAR1	AGAGTTTTGATCATGGGGTGGACTCC
PRE1eRNAF1	TTGCTCTTCCTGGACACTTGGTGC

***ORAOV1* cloning primers**

ORAOV1PromoterkpnF1	GGTACCGAACCAAGCTTTTCCCAAACAACC
ORAOV1PromoterxhoIR1	CTCGAGCAGCTTCAGGCACAAATGCTCC

***ORAOV1/CCND1 HindIII* 3C primers**

11q13ORAOV1HindInt1F1	TCACATTCACAGATCGGGTGCTACCAAGG
11q13HindFrag6R	CCTGCACCCTGGAGTGTCTGACAATTCTT
11q13HindFrag9R	CGTGGTCTGTGCTTAGGAGAAGACGTGG
11q13HindFrag10R2 (PRE1)	CCAGAGAGTCCCGTTATGGGATTCCCC
11q13HindFrag12R	CCAGAGATCCAGAGACCTACAGTACAGCATGAG

11q13HindFrag13R	GAGAGTGGGAGTTGCCACATACTTTGAAATGAC
11q13HindFrag14R	CCACCTGAGACTTTGGCAGCCTAGACTTCA
11q13HindFrag15F2	GGACGACCTGGGATTCGTAGATGACTCC
11q13HindFrag16F	GGAGACCAAAGCAGCGAGAGCTGG
11q13HindFrag17R	GCTCCGTTTCCTCTGGTGCGAGAATAAGAT
11q13HindFrag18F	CCCTCCCTTCACTTGTTTAGAAATGCACTGG
11q13HindFrag19F1	TTGGGTGGCCGCCTAAGTCAAGAGG
11q13HindFrag20F	CCATCCTCTGGAAGACAGGCACAGTGC
11q13HindFrag21F	AGACTGTGCAACTGAGAGAACGGTGCTGC
11q13HindFrag22R	GGAGCAACGAAAGTAGATGCCACTTCAAACC
11q13HindFrag23F	GTATGAAAAGCAGCTACAGCAGGGCTTAGAGG
11q13HindFrag25F	CCACTTCTTGGACCCTAGAGCCTCGAAATCC
11q13HindFrag27R	CCAAGAGGGAGAGCGAAGTCACAACAATAC
11q13HindFrag28R	CCAGGCACGCTTGGATGCGTTAGTCTGAGA
11q13HindFrag29R2	GCGGAGGTTGTACATGAGGTTAAGGAAAGAAGC
11q13HindFrag30F2	GGTCATCGGGATCCACAGGTTCATGG
11q13HindFrag31R2	GCAGAAGCCTTTGGCCCAACTTTGC
11q13HindFrag32R1	CTTTACTCTTTCCAAGCGCGGCAGAGC
11q13HindFrag33R1	ATTGCCCAGGCTTGCCGAATACACC
11q13HindFrag34F	GCTGTGCTGGTTGACTCCTGGTATTCAAAGG
11q13HindFrag35R	CAACAGGAACCTTGGTCTTGGACGTCTGG
11q13HindFrag37R2	CCTAGAAGTCCGAGGAGCCTGGCTGTACATAT
CCND1hindfragUp3F2	TTAAAACCAAAAGCCACCAGGCACAGTGG
CCND1HindFragUp2F1	TGACTTGGGTGAAGAGGGTCTTGGAACG
CCND1HindFragUpF1	CCCGTGACTCCCCTACATTGCATCTCC
CCND1HindFragIntF1	CAGCAGATGCTATCTAGGGTCCACCTGCC
CCND1HindFragDnR1	GCTCTCTCAGGAAAATGGCTCAGAAACACC
CCND1HindFragDn2F1	CAGGGATTGGCAGGTTTTCTGGAAAGG
CCND1hindfragDn4F2	GGCTTGGACAACTCACTGAACCCTTGAGG

***ORAOV1 NcoI* 3C primers**

ORAOV1Nco1up1F1	ACAGGCTCCCAAACCCAAGCACACC
ORAOV1Nco1up4F1	CCAAGTCTTGTCAATAACTTGCTGTGT
ORAOV1Nco1up5F1	CTGTCATGGATGCTGTCCTGAGAATAA
ORAOV1Nco1up6F1	GGACAGTCTGTAGCACAAACGTCCAT
ORAOV1Nco1up7F1	GGTTGGGTCACCTGAAATCCCTCTGTG
ORAOV1Nco1up7R1	CGTGGTAAATGCGTGATAAATGCTCTT
ORAOV1Nco1up8F1	AGACGCAGCACAGCCTGGATTTACGA
ORAOV1Nco1up11F1	GTGACTCTGGGCAAGGAGCAGTTTCT
ORAOV1Nco1up12R1	GTAGTTCTTGCAGCAGCGGCTCTGATG
ORAOV1Nco1dn2F1	GCTTGCCAAGTGGAAAGTGCGTGG
ORAOV1Nco1dn3F1	GTTGCATGAACGATTCGCTTCCTGAGG
ORAOV1Nco1dn4F1	TGACCGCAAGGCTCACTCTCCAAGAG
ORAOV1Nco1dn5R1	GAATTTCAACATGCCTGTCTTGCTTAG

Table 8.4 Primers used for thesis. *The sequences are given 5' to 3'.*

8.7 CUPID1 and CUPID2 Sequences

*CUPID1*Isoform(1)

tgtggggcagagggcagcctgagcatctgcaaacagagtggctcacagagaggacacctgccctgaggacacctgcaggagatccctacag
aggaagggctagaccaggctccatccagcaaatacaaactcatctggaagcattaaaacaatggaagcagggcctgctgtgcgccagctgct
ggagacatgaagacatcgaagaccccaggccttgacctccaagagcccccagtttggtggaagaaacaaacaaggagactgcaaaacacac
cctgtgatggggagggttgaggtggccagtgggagaggcgagctggcagggctcagtggtgaggtgaaggtgtgggagcagggacacag
catgaaggagtagatggccaggttgagctttgcaggctccagggggaggataaagagctgacacctggaacgtgcttggaacagcaggtgct
cggtcagtgcagccacgaccatcccgtcctgcatccccaccatgaacttggaccttctggaaggaccttgcttatcccagcatccacacatccttg
accagacctggtcaacatttgcttagaaccgcctggctgctgggatgagaagagactggagagagaccacgatggaagccaggaggccaac
gaggagttggggtagtcctccagggagacatgatgtggcttcagccccagtggtgatggcggggtagagccagcagggttctgaactgagtg
gatgtgaaatgagggaaagagggatgctgcggtcggcagaatcgtgcctcccgcaaagatgccccatcctaacccgtcacccttgacagtga
gacctccaccacaaaggggactttgcaggcatgaaggttaaggaccctgagaggaggagatggccctggactgtctggcgggcccagcatc
atcacaagggtgcttataagtggaggaggaggcaggggggagccgagcagatgtgaggatagaagcagaggtcggagggacgtgctgttgg
cttgagatggaggaggacctatgagccaaggagcatgggcggcctctggaagccagaaagggcaaggcctcttcccggagcctccagcag
aacccagccccgctgccacctcgactgcagaacggtgacaccgtgtccgattcccgacctccagaaccgtgagataacacgtgtgctgttttaa
gccactgcaattgtggtatttgtcacagcagccataggaaactgtacgggggttaaagaggacctgctgacctctgagctcacactgaagcctcc
tggaatgctgtgggttgtgagctgggatgagaacgcagcatggagatgaaagacgcatccacccaggaaggcccactggcatcgcaatctca
gggaagattcccaggctgaaaaaagagaaagtggctggggtggcactgaatgccccacgtcaaagaaatggcacccaagccgtccgtcagt
ccctcacctgccacacagacagcaccttggacccagcaagatgtcactcaccacattgatgctgtccacaggcatcgccttgcttttgcagctctg
tgtgagatttgtcagcttggtttgggtctggggctgtattagagctgtcagcctaggagcctaccctagttttgtgtctgcactatttccaaagcactg
gaataaaaggcattttaaccaac

CUPID1 Isoform (2)

tgtggggcagagggcagcctgagcatctgcaaacagagtggctcacagagaggacacctgccctgaggacacctgcaggagatccctacag
aggaagggctagaccaggctccatccagcaaatacaaactcatctggaagcattaaaacaatggaagcagggcctgctgtgcgccagctgct
ggagacatgaagacatcgaagaccccaggccttgacctccaagagcccccagtttggtggaagaaacaaacaaggagactgcaaaacacac
cctgtgatggggagggttgaggtggccagtgggagaggcgagctggcagggctcagtggtgaggtgaaggtgtgggagcagggacacag
catgaaggagtagatggccaggttgagctttgcaggctccagggggaggataaagagctgacacctggaacgtgcttggaacagcaggtgct
cggtcagtgcagccacgaccatcccgtcctgcatccccaccatgaacttggaccttctggaaggaccttgcttatcccagcatccacacatccttg
accagacctggtcaacatttgcttagcctctgccagcctcctctcacccattcctgcatgtgccgcaccaatctgtgtggagcgcccactccagcc
aggccctgctctatcagcaaactcagcaggcagaaccccctgccccatggcgctgcttcagcgtggtgagacaggaaccgcctggctgctggg
atgagaagagactggagagagaccacgatggaagccaggaggccaacgaggagttggggtagtcctccagggagacatgatgtggcttca
gccccagtggtgatggcggggtagagccagcagggttctgaactgagtggatgtgaaatgagggaaagagggatgctgcggtcggcagaat
cgtgcctcccgcaaagatgccccatcctaacccgtcacccttgacagtgagacctccaccacaaaggggactttgcaggcatgaaggttaagg
accctgagaggaggagatggccctggactgtctggcgggcccagcatcatcacaagggtgcttataagtggaggaggaggcaggggagc
cgagcagatgtgaggatagaagcagaggtcggagggacgtgctgttggcttgagatggaggaggacctatgagccaaggagcatgggcgg
cctctggaagccagaaagggcaaggcctcttcccggagcctccagcagaacccagccccgctgccacctcgactgcagaacggtgacaccg
tgtccgattcccgacctccagaaccgtgagataacacgtgtgctgttttaagccactgcaattgtggtatttgtcacagcagccataggaaactgta
cgggggttaaagaggacctgctgacctctgagctcacactgaagcctcctggaatgctgtgggttgtgagctgggatgagaacgcagcatgga
gatgaaagacgcatccacccaggaaggcccactggcatcgcaatctcagggaagattcccaggctgaaaaaagagaaagtggctggggtgg
cactgaatgccccacgtcaaagaaatggcacccaagccgtccgtcagtccctcacctgccacacagacagcaccttggacccagcaagatgtc
actcaccacattgatgctgtccacaggcatcgccttgcttttgcagctctgtgtgagatttgtcagcttggtttgggtctggggctgtattagagctgt
cagcctaggagcctaccctagttttgtgtctgcactatttccaaagcactggaataaaaggcattttaaccaac

CUPID2 Isoform (1)

Gcaagacagccacaccttgaactctgactcaaggaagagagtgggcagctcaccagagaggccagaggaggcggcaggaaggtggggtc
tctccagccaggggcttcctgaaagagctgccttctgagcaggcccccccaggctaggaagatgtgcacctgccgagctgggcaggagagggt
ggcccagtggaggacccagcatgggcagaggtttggaggtggcaggtcctgagggaagcggaaggcagccaggttgtgacagcctgtgag
gacctcttggctcccggagccttctctctcctggaatagagatgacaccgtgctgccattgaactcaacccgaaaaaatgatctgttgcaaagctc
acaggaaataatggagcaatgttacacgataaactgtaaagtgtggtgagcatgtgaacaagcgtccaggacccagcgaagccccaggctact
gggctaggcgtcagatgagaagaggcagagggcaacctcctgtaggatctgaggccttcatgtgctcagcccatggctgtcgtcacggtgacg
tgtgcatgcgatgtatctgggccatccttgttcttgaacgcgggcagtgagtcccacggtgatgtggcatgtcacgagctgaaggaggggcacc
tgcattcaggcttgcgaaccagcgagacagcctttggacctttcggagcctcactttcccacactaagaattggagtggaattcccaaggtgctg
gaggttgaaaggcattgagaatctaaaacgctttgcagacagcaagaccctctgtgtaagttaagaatgctcactcagtcccagcaggtactcag
agggatccttttatttctccagaacattctgcaggtgagctggggaagcccactggctccagggactgccaggagggccccagggaagggga
gccgggaagcgttttggctgcctccggcgtgaagtcgtgcaaacagccattgaatctgccagtgacttggtcgtggagttccagcctccagaact
gtgagaaataaatgtttgcgatttaaaccgca

CUPID2 Isoform (2)

cagactttgccctcacaggcaagacagccacaccttgaactctgactcaaggaagagagtgggcagctcaccagagaggccagaggaggcg
gcaggaaggtggggtctctccagccaggggcttcctgaaagagctgccttctgagcaggcccccccaggctaggaagatgtgcacctgccgag
ctgggcaggagagggtggcccagtggaggacccagcatgggcagaggtttggaggtggcaggtcctgagggaagcggaaggcagccag
gttgtgacagcctgtgaggacctcttggctcccggagccttctctctcctggaatagagatgacaccgtgctgccattgaactcaacccgaaaaa
atgatctgttgcaaagctcacaggaaataatggagcaatgttacacgataaactgtaaagtgtggtgagcatgtgaacaagcgtccaggaccca
gcgaagccccaggctactgggctaggcgtcagatgagaagaggcagagggcaacctcctgtaggatctgaggccttcatgtgctcagcccat
ggctgtcgtcacggtgacgtgtgcatgcgatgtatctgggccatccttgttcttgaacgcgggcagtgagtcccacggtgatgtggcatgtcacg
agctgaaggaggggcacctgcattcaggcttgcgaaccagcgagacagcctttggacctttcggagcctcactttcccacactaagaattggag
tggaattcccaaggtgctggaggttgaaaggcattgagaatctaaaacgctttgcagacagcaagaccctctgtgtaagttaagaatgctcactca
gtcccagcaggtactcagagggatccttttatttctccagaacattctgcaggtgagctggggaagcccactggccctggcctcacccccgagc
ccgagccttcagcctgggaatggtctggcctcttccttgctggctctccagcctggcctcgcaggaccctgggcgggaccccaggaaccctca
cccgctatgtgcttccccaagaagcgctccctgtggcctaatttgagaaaacaatgggcctcaatccatattaatgaccctagagggacccttttgt
cctcggtgcacaggctgtaatcagcggggctccgggggctctggcctaatttggagggacaggttttatcatcaccttgattcgggtgacccaa
tctgacaggcccacgacccccctgtatgcggggtccacgtcagagatgggcttcctcccagcggcccaccccagcgggctggggagagggaa
gggggaggtggtggccatgggggaagttcggggtgaagaaggggtcacagccagacccccacttggatgggcctgtgatcgggttctcggg
aggagcaggatattgattagatcagtgaatggtgtggaggcagctctccccaggcacgtgctcccgaccacccaccagcaagcgtctgttgcct
gccggtgccagggctgggtggggactctgggacaggccggctgcctaggaggggccaggcaggagccaaggggctggctccagggact
gccaggagggccccagggaaggggagccgggaagcgttttggctgcctccggcgtgaagtcgtggaaacgtgtttgcaggttagggccgtg
gcatcctcgttgcaccatttgagcgcctatggggtgccagatacatgacaaagggccacctcatttaatgcttccctcaaccttcccaggccccga
tccccatgttctgggtcaggaggcccattcgaggtcaaggggtgggcgaaggctgacgagacaaacccagggactagctcctgagtcccaaa
acctgtcctcacccacccaggggtcccctgcacgctctccagggtgtgcggtgcgtggctgctgggtttgggtgcattagagcctgaggcaag
ggcttgagtctgaattctcgctacggaccaaatgtgtgcgtcctccctaaatacatatgttgaagccttaactcccaatgtgatggcattaggagag
gggacctgtggaatgtagttgagtcctgagggtgcggccccgtgatgaaattagtggcctgataagaagaggcccagagaaccggctctgtct
ctgctctctgtcgcacgaggaccccacaagaagacggccacctgcaaacagccattgaatctgccagtgacttggtcgtggagttccagcctcc
agaactgtgagaaataaatgtttgcgatttaaaccgca