The Role of Long noncoding RNAs in Tamoxifen Resistance in Oestrogen Receptor Positive Breast Cancer

By:

Fatma A. Bucklain

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Medicine, Dentistry and Health
Department of Oncology and Metabolism

July 2022

Acknowledgments

I would like to first and foremost thank my supervisor, Dr Helen Bryant, for continued help and guidance throughout my PhD, most particularly during the past two years, when a life-changing event along with a global pandemic was very challenging to finish my journey. I like to describe Helen as a super-supervisor, she always showed understanding, pushed me to do what I have to do, and provided me with support. Helen, this thesis would not exist without you, I will be eternally grateful.

I would like to thank my second supervisor Dr Dennis Wang and Dr Mark Dunning for their help and expertise in the bioinformatics analysis in this thesis. I would also like to profoundly thank Dan Harrison for his valuable help and time.

I would not have been able to complete the work in this thesis without the friendship and support of my fellow Bryant and Collis lab members. Special mentions go to Emma, Leona, Tim, Kathryn, Polly, Katie, Callum, Ola, Chris, Tom, and Saskia, would not forget my friend Marwa as well, my deepest thanks to you all.

My sincerest gratitude to the many unnamed patients and research groups who gave the gift of data.

I cannot thank my one and only, Hamza, enough. The past 2 years have been really hard. You were my solid rock, you pulled me when I was down, praised me when I was up, and took care of everything when I was not there. Thank you from the bottom of my heart, I love you. To my Elias, you are my lighting joy in the dark. To my friends Fatma and Nadeen, your friendship kept me going through the years. This work is dedicated to you all.

Abstract

Endocrine resistance in ER-positive breast cancers is a common occurrence with hormonal therapies leading to accelerated tumour progression and metastatic complications. The exact molecular mechanisms driving resistance are still unclear and need further elucidation. In this study, long noncoding RNAs were highlighted as key players in orchestrating the complicity of different molecular pathways of resistance. LncRNAs are pervasively transcribed across the genome, measuring > 200 nucleotides in length, and found to be dysregulated in several tumorigenic pathways.

We hypothesised that tamoxifen resistance is influenced by dysregulated lncRNAs, and they constitute novel targets for nominating new biomarkers and therapeutics. To first investigate this, we generated a list of lncRNAs differentially expressed between tamoxifen resistant and sensitive MCF-7 cells using RNA sequencing. Of which, LUCAT1, SOX21-AS1, NR2F1-AS1, and HOTAIRM1 were selected to undergo further molecular validation studies, mainly siRNA mediated depletion of expression in tamoxifen resistant cells, then assessing tamoxifen sensitivity in vitro. While the 4 nominated lncRNAs' expression was significantly high in tamoxifen resistant MCF-7 compared to tamoxifen sensitive MCF-7, no observable change in the tamoxifen sensitivity was observed when silencing any of the lncRNAs. To aid further candidate lncRNAs selection for invitro studies, GEO and TCGA databases were searched for data related to tamoxifen resistant phenotype, analyzing the data allowed for expanding our understanding of the genotypic changes related to tamoxifen resistance, as many genes known in the pathway of resistance were confirmed. Also, this offers grounds for nominating promising candidate lncRNAs in

the pathway of tamoxifen resistance. In conclusion, RNA-seq analysis identified many dysregulated lncRNAs and protein coding genes, of which LUCAT1, NR2F1-AS1, SOX21-AS1 and HOTAIRM1 were prioritized as lncRNAs driving tamoxifen resistance in breast cancer. LUCAT1 was promising after bioinformatics investigation and remains a significant candidate for invitro validation. It was showed that comparisons, relating and overlapping between differentially expressed genes from RNA-seq and publicly available datasets resulted in nominating lncRNA ZNRD1ASP as possible candidate, additional in-depth analysis is needed.

Table of Contents

Table of Figure

Table of Table

Abbreviations

4-OHT	4-hydroxy tamoxifen
ADGRV1	Protein coding genes adhesion G protein-coupled receptor V1
AI	Aromatase inhibitors
AP-1	Activator protein 1
APS	Ammonium persulphate
ARRDC3	Arrestin domain- containing 3
ATRA	All-trans retinoic acid
BD	Becton Dickinson
BRCA	Breast cancer project
BSA	Bovine serum albumin
CCLE	The Cancer Cell Line Encyclopaedia
CPAT	Coding Potential Assessment Tool
DAPI	Diamidino-2-phenylindole
dATP	Adenosine
DBD	DNA-binding domain
ddH2O	Double-distilled purified water
DDR	DNA damage response
DEA	Differential expression analysis
DGE	Differential gene expression
DMSO	Dimethyl sulphoxide
DNA	Deoxyribose Nucleic Acid
dTTP	Hymidine
E1	Estron
E2	Estradiol
E3	Estriol
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
EMT	Epithelial-mesenchymal transition
ENCODE	The Encyclopaedia of DNA Elements
ER	Oestrogen hormone receptors
ERAR	Oestrogen receptor agitation related
ERE	Oestrogen responsive elements
ERK	Extracellular signal-regulated kinase
ERα	Oestrogen receptor-alpha
ERβ	Oestrogen receptor-beta
ES	Enrichment score
FACS	Fluorescence-activated cell sorting
FSH	Follicular-releasing hormone
GEO	The Gene Expression Omnibus
GFRs	Growth factor receptors

GnRH	Gonadotropin-releasing hormone
GPCRs	G-protein coupled receptors
GSEA	Gene set enrichment analysis
HGP	The Human Genome Project's
hnRNA	Homogeneous nuclear RNA
HOXA	Homeobox cluster A
HRT	Hormone replacement therapy
LH	Luteinizing hormone
lncRNAs	Long non-coding RNA
LUCAT1	The lung cancer-related transcript 1
MAKP	Mitogen-activated protein kinase
mRNA	Protein coding messenger
MSigDB	Molecular Signature Database
MTT	Thiazolyl blue tetrazolium bromide
NB	The negative binomial model
NEAA	Non-essential amino acids
NES	Normalised enrichment score
NF-ĸB	Nuclear factor kappa B
NR2F1-AS1	Nuclear receptor subfamily 2 group F member 1 antisense RNA 1
OD	The optical density
PAR1	Protease-activated receptor 1
PBS	Phosphate buffered saline
PCA	Principal component analysis
PFS	Progress free survival
PI	Propidium iodide
piRNA	Piwi-interacting RNA
PMEPA1	Prostate transmembrane protein androgen induced 1
PMSF	Phenylmethanesulphonyl fluoride
RA	Retinoic acid
RFS	Relapse free survival time
RIPA	Radio-immunoprecipitation assay
RNA	Ribonucleic acid
RPKM	Reads per kilobase of transcript per million mapped reads
RPMI	Roswell Park Memorial Institute
rRNA	Non-coding - ribosomal RNA
SC	Scrambled siRNA
SDS	Sodium dodecyl sulphate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SDS-PAGE	Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis
SERMs	Selective oestrogen receptor modulators
siRNA	Small interfering RNA
snoRNAs	Small nucleolar RNA
SNPs	Single nucleotide polymorphisms

SOX21-AS1	SRY-box transcription factor 21 antisense RNA 1
SP1	Specificity protein 1
TAMR	Tamoxifen resistant MCF-7
TBS	Tris-buffered saline
TCGA	The Cancer Genome Atlas
TF	Transcription factors
TNM	Staging of breast cancer follows tumour-node-metastasis
Tris-HCL	1 M Tris hydroxymethyl aminomethane hydrochloride
VST	Variance stabilizing transformation.
YAP	Yes-associated protein.
β-ME	β-Mercaptoethanol

Chapter 1 Introduction

1.1 Breast Cancer Biology

1.1.1 incidence and development

About 1 in 8 women will be affected by breast cancer in their lifetime, this is a cumulative risk of 12.5%. According to the latest statistics, about 2.3 million women were diagnosed with breast cancer in 2020, and it accounted for more than 600,000 cancer related deaths (Sung *et al.*, 2021). In the UK, breast cancer is considered a major public health issue, with about 55,000 new cases annually and more than 11,000 deaths associated with the disease (Cancer Research UK, 2017). While familial breast cancer cases constitute only 5% of breast cancers, most of breast cancer cases are sporadic. Unlike mono-genetic diseases, where an easily identified mutated single gene is responsible for pathology, breast cancer is a multi-genetic disorder. Many somatic gene mutations and gene expression aberrations are found to be directly and indirectly associated with the development and progression of breast cancer (Nandy, Gangopadhyay, and Mukhopadhyay, 2014). The complexity of breast cancer is further enhanced by the strong association with environmental mutagens and social factors that alter many molecular pathways such as cell signalling and DNA repair (Ferrucci *et al.*, 2009).

Breast cancer originates in either the ductal or lobular compartments of the mammary glands; with most breast cancers classified as invasive ductal carcinomas (Goh *et al.*, 2019). Carcinogenesis in breast tissue starts usually with abnormal proliferation (hyperplasia) that progresses to carcinoma in-situ then becomes invasive to adjacent breast tissue (local disease), breast lymph nodes

(regional disease), to invade distant organs (metastatic or secondary disease) (L. Chen *et al.*, 2013). Further pathological examination of the breast cancer histological biopsies labels it with a grade according to cancer cellular degree of differentiation compared to normal cells. Staging of breast cancer follows tumour-node-metastasis (TNM) system, where T expound the size of primary breast tumour, N is for the number of regional lymph nodes involved, and M for presence or absence of distant metastases, based on the TNM score breast cancer get a stage from 0 – 4 (Li *et al.*, 2018).

Breast cancer receptor status together with pathological properties and staging impact the treatment plan dramatically. Treatment is usually multi-strategic combining surgery, chemotherapy, radiation, and targeted therapy (Schmitz *et al.*, 2012). Owning to advancement in treatment options, targeted therapies, and early detection, the mortality rate has improved considerably over the past decades. With larger numbers of survivors, long-term complications like drug resistance and progression can be observed more in the population (Zhang *et al.*, 2013).

Better understanding of the genomic signatures of breast cancer will lead to comprehension of the cellular and molecular mechanisms that give rise to the complexity of disease. In addition, this understanding will further advance clinical practice towards more personalised medicine.

1.1.2 Molecular subtypes of breast cancer

The difference in genetic profiles of breast cancers creates an extensive heterogeneity in-terms of clinical response and progression in phenotypically very similar cases. This demanded the application of molecular classification of breast

tumours to allow for better allocation of treatment plans and targeted therapies (Cianfrocca and Gradishar, 2009).

There are three main categories of breast cancer based on pathological markers: the most prevalent type, hormone-receptor positive, exhibit either or both of oestrogen hormone receptors (ER) and progesterone hormone receptors (PgR). ER positive tumours comprise more than 70% of cases, such tumour cells depend on oestrogen hormone for growth and expansion (ZHANG *et al.*, 2014). Based on this principle, over the past five decades endocrine therapies have been used successfully to prevent cell proliferation in this type of cancer and to improve survival rates (Vasconcelos *et al.*, 2016). Human epidermal growth factor receptor 2 (HER2) also called HER2/neu or Erb-B2 receptor tyrosine kinase 2 (ERBB2) is over expressed in about 20% of breast cancers. The prognosis for this type of breast cancer was considered poor until the introduction of targeted immunotherapies that dramatically improved survival (Incorvati *et al.*, 2013). HER2 expression can occur with or without ER/PR expression. A third type of breast cancer is triple negative breast cancer, these tumours lack hormone and HER2 receptors. Most patients with genetic predispositions like BRCA1 and BRCA2 gene mutations get diagnosed with this type of breast cancer, it is the least common type with least favourable prognosis due to the lack of targeted therapies (Mayer *et al.*, 2014).

Creating an effective categorisation model for breast cancer has been complex and many classification techniques have been developed over the past two decades. Using gene expression profiles, based on the degree of expression of ER, PgR, HER2 and Ki-67, five main intrinsic molecular subtypes were identified: ER-positive

(luminal A and luminal B +/-HER2) and ER-negative (HER2-positive and basal-like) (Table 1.1 (Vasconcelos et al., 2016; Charles et al., 2000).

Table 1.1. Molecular subtypes of breast cancer

Molecular subtypes	ER	PgR	HER2	Ki-67
Luminal A	+	+	-	low
Luminal B-HER2	+	-*	-	High*
Luminal B+HER2	+	+ or -	+	High or low
HER2 +ve	-	-	+	n/a
Basal like (triple negative)	-	-	-	n/a

(*) either PR is absent, or law or Ki-67 is high. (n/a) not associated. (+) positive, (-) negative.

1.1.3 Oestrogen receptor signalling

Oestrogen hormone is the main steroid hormone initiating and maintaining sexual growth in females. It regulates development of secondary sexual characteristics (e.g., growth and proliferation enhancing effects on mammary cells), menstrual cycles, and gestation (Khan, 2019). It's vital systemic action supports homeostasis, healthy bones, heart, and brain, both in males and females (Bartos, 2009) (Figure 1.1). Throughout reproductive years, ovaries serve as an endocrine gland producing oestrogen under the influence of gonadotropin-releasing hormone (GnRH) from the hypothalamus, and luteinizing hormone (LH) and follicular-releasing hormone (FSH) from the anterior pituitary. Gonadotropins stimulate theca

and granulosa cells in ovary to metabolite androstenedione to oestrogen (Cui *et al.*, 2018) (Figure 1.1). Oestrogen gets secreted in a cyclical manner through the menstrual cycle, the lowest level of oestrogen is reached during menstruation and highest is reached peri-ovulation (Gaskins *et al.*, 2012).

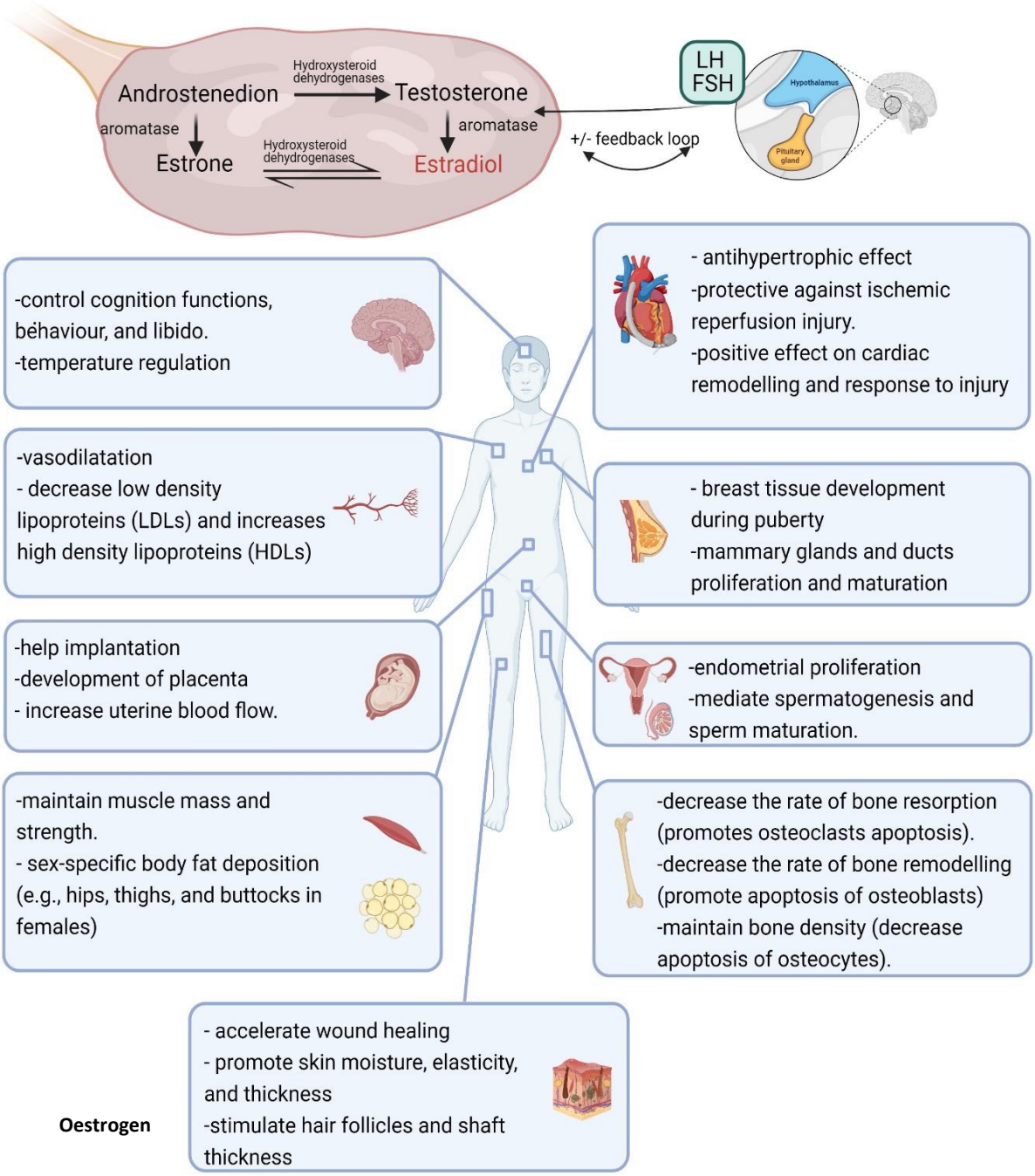


Figure 1.1. production and physiological actions. TOP: Scheme of estrogen synthesis and control by ovary and brain, where LH stands for luteinising hormone, FSH for follicle stimulating hormone. **BOTTOM:** organ and systems regulated by estrogen, affecting almost all tissues in the body with gender specific and nonspecific manner.

There are three different forms of oestrogen, Estradiol (E2) is the most potent and the predominant form circulating during female reproductive age, Estron (E1) is the main oestrogen post-menopause, and Estriol (E3) that is primarily produced during pregnancy (Bartos, 2009). After menopause, ovarian oestrogen production stops and much lower quantities of oestrogen get produced from nongonadal sites, mainly adipose tissue but also from skin, liver and adrenal glands (Rachoń and Teede, 2010).

Oestrogenic effects are classically exerted through E2 interactions with binding sites on ERs in the target tissue. These are ligand-activated nuclear receptors that function as transcription factors (TF) regulating gene expression (Fuentes and Silveyra, 2019). Oestrogen receptors, like other members of the nuclear receptor superfamily, hold a well-conserved and central structural component containing a DNA-binding domain (DBD) that interacts with the DNA in a sequence-specific manner at oestrogen response elements (ERE). There are two subtypes of ERs, both transcribe from different genomic locations, oestrogen receptor-alpha (ERα) is transcribed from ESR1 gene on 6q24-q27-chromosome six and oestrogen receptor-beta (ERβ) which is transcribed from ESR2 on q22-24-chromosome 14. Both ESR1 and ESR2 encode eight exons and generate three and five isoforms respectively by alternative splicing (Greene et al., 1986; Enmark et al., 1996). The structural organisation of both consists of five domains: A/B which form the ligand-independent domain, C which forms the DBD, D which forms the hinge domain and E/F domain that forms the ligand-dependant activation area as shown in (Figure 1.2) (Nilsson *et al.*, 2001). ERα and ERβ share a lot of resemblance in structure,

but have differences in localisation, activity, and interactions, details of which are given in (Table 1.2).

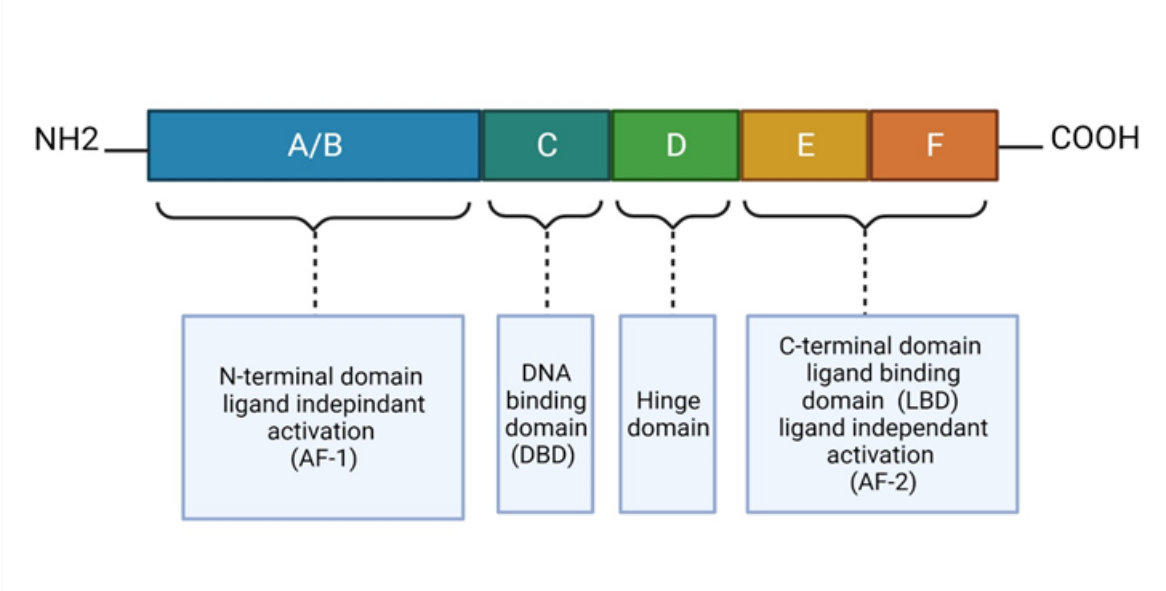


Figure 1.2. Structure of wild type ER. There are five structural domains A/B, C, D, E and F, three important functional areas in addition to 1 structural or D hinge domain (1) A/B covers ligand-independent activation function (AF1) and mediates many protein-protein interactions, (2) C is the DNA binding domain which mediates ER-ERE interaction, (3) E/F which forms the ligand binding domain, binds coactivators and is involved in signal transformation.

Table 1.2. Comparison of general properties of ERα and ERβ

Difference	ERα	ERβ
Gene and location	ESR1 / 6q25.1	ESR2 /14q22–24
Tissue distribution	Breast (Epithelial cells nuclei) Bone, Adipose tissue, liver, kidney.	Breast (Epithelial and myoepithelial cells nuclei) Colon, lungs, male reproductive organs, neurological tissue
Domains	Different N-terminal domain but similar DBD and LBD	
Response to tamoxifen	Partial agonist	Pure antagonist
Expression in normal mammary cells	~5-15 %	Widespread
Expression in breast cancerous cells	~70%	Far smaller than Erα
Association with breast cancer	Well demonstrated	Unclear and needs further investigation

The physical association between E2 and ER initiates series of molecular events that induce oestrogen receptor conformational change and receptor dimerization, where ERα and ERβ homodimers (ERαα or ERββ) or heterodimers (ERβα) are formed according to the specific tissue distribution of ERα and ERβ (Cowley *et al.*, 1997) (Table 1.2).

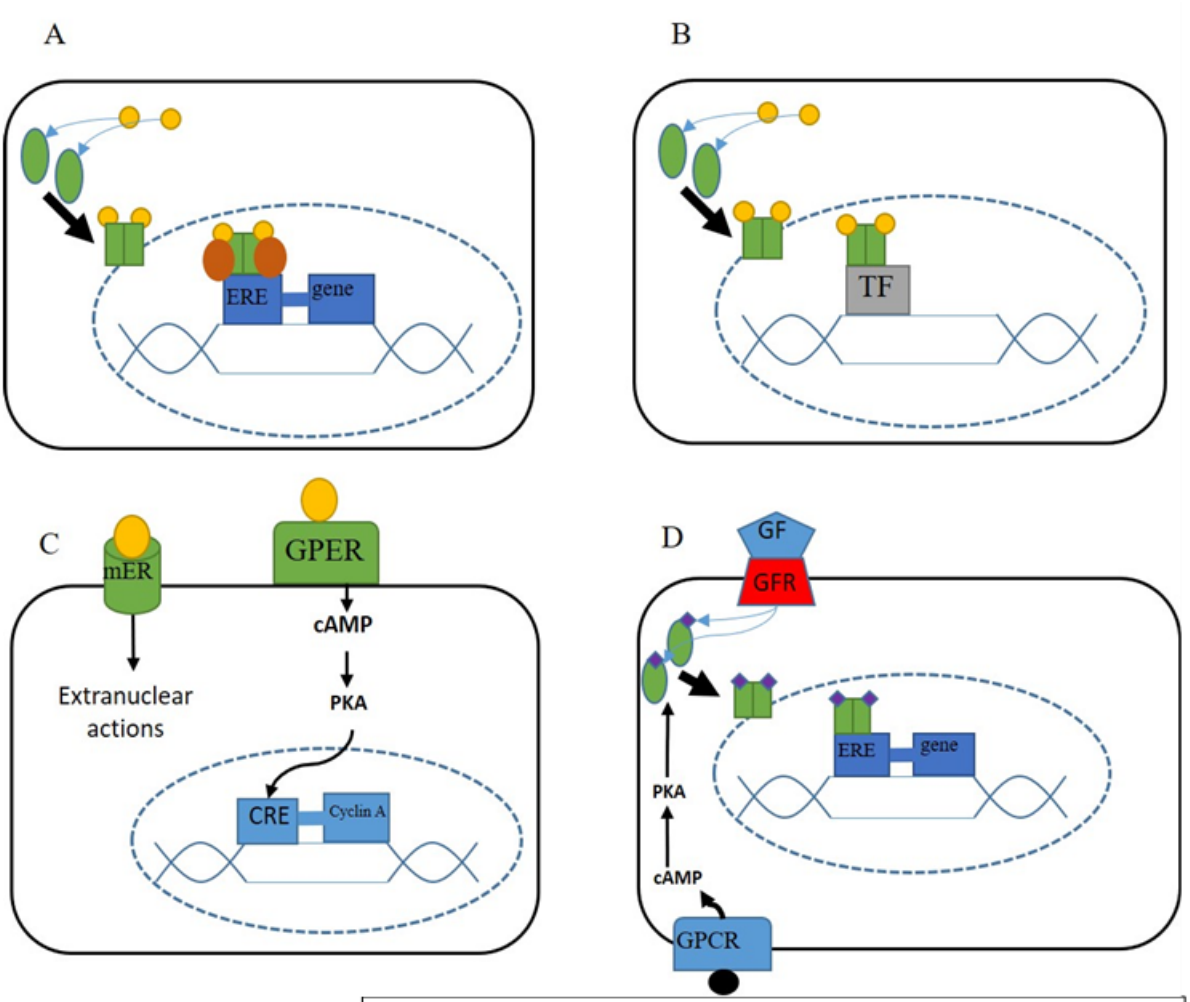
Once activated the ER-E2 complex undergoes nuclear trans-localization where it binds to a specific region on cis-acting transcripts proximal to oestrogen targeted genes on the DNA called oestrogen responsive elements (ERE), this binding is

followed by recruitment of co-activators and co-suppressors to directly stimulate or supress the desired genes transcription (Nilsson *et al.*, 2001) (Figure 1.3.A). Moreover, ER-E2 binding is not exclusive to ERE, under certain conditions, it can interact with alternative response elements (figure 1.3.B), such as nuclear factor kappa B (NF-ĸB), activator protein 1 (AP-1) and specificity protein 1 (SP1) and mediate the transcription of their targeted genes (deGraffenried, 2004).

In addition to the above-mentioned pathways that mediate oestrogen genomic signalling through nuclear ERs, oestrogen is known to have a nongenomic signalling pathways through membranous ERs. Membrane ERs are the same protein product as nuclear ERs but post-translational chemical modification (i.e., palmitoylation of cysteine on the C-terminal LBD (E/F domains)) is responsible for the ability of ER to translocate to the cell membrane and attach to signalling molecules such as G-proteins and c-Src (Razandi *et al.*, 2003) (Levin, 2008). G-protein coupled ERs (GPER) mediate rapid oestrogen signalling by activating several intracellular cascades through second messenger signalling to many growth-related kinases such as extracellular signal-regulated kinase (ERK) and phosphatidylinositol 3-kinase (PI3K) in a similar manner to classic G-protein coupled receptors (GPCR) (Levin, 2008). E2 is also a known ligand for orphan G-protein coupled receptor 30 mediates rapid oestrogen signalling through protein kinase A/cyclic adenosine monophosphate (Camp/PKA) signaling pathway (Mo et

al., 2013; Filardo et al., 2002) (Figure 1.3.C). Functional studies using truncated forms of ER protein showed that, in the absence of E2, ER-activation can be achieved in a ligand-independent manner through growth factor receptors (GFRs) (e.g. EGFR, HER2 and IGF-R1) and G-protein coupled receptors (GPCRs) (e.g. protease-activated receptor 1 (PAR1) and GPR116) signalling cascades. GFRs activate downstream protein kinases (e.g., mitogen-activated protein kinase (MAPK), phosphatidylinositol-3 kinase (PI3K)) that in-turn mediate the phosphorylation of AF-1 (A/B) domain of ER and subsequent ER-related transcriptional activation (El-Tanani and Green, 1997) (Maggi, 2011) (Figure 1.3.D) . GPCRs ER-activation by phosphorylation is achieved mainly through cAMP/PKA pathway but it can also crosstalk with GFR pathway components at the level of mTOR and MAPK (Thomas *et al.*, 2006) (Singh, Nunes and Ateeq, 2015). Thus, while E2-nuclear ER interaction is the dominant driver of growth signalling in mammary tissue, non-classical ER signalling contributes to a complex network of molecules that have the capacity to override E2-ER pathway and facilitate cellular growth and proliferation.

Well-controlled and balanced oestrogen signalling is essential for functional breast physiology, while dysregulated signalling and prolonged E2 exposure are known to gives rise to uncontrolled growth, proliferation, and progression of breast tumours (Colditz, 1998).



	E2
	Inactive ER
	active dimerized ER
	Phosphate
	Mediator for GPCR
	Co-activators /co-suppressors
ERE	Estrogen reactive element
TF	Transcription factor
mER	Membranous ER
GPER	G-protein coupled ER
CRE	Cyclin reactive element
cAMP	Cyclic adenosine monophosphate
PKA	Protein kinase A
GPCR	G-protein coupled receptor
GF	Growth factor
GFR	Growth factor receptor

Figure 1.3 Oestrogen signalling pathways in breast cancer. A) classical ER signalling pathway: in response to E2 binding, ER undergoes dimerization, activation, and nuclear translocation, to bind DNA directly on ERE and recruit coactivators/cosupressors upstream of estrogen-targeted genes and mediate their transcription. Alternative growth pathways are B) ERE-independent pathway: where activated ER binds TFs such as SP1 or NF-κB, which go on to bind to promoters and thus activate downstream associated genes. C) nuclear estrogen receptor independent pathway: other types of estrogen receptors (cell membrane ER or GPER) mediate extra nuclear or nuclear (cAMP-PKA pathway) actions. D) ligand-independent pathway: growth factor receptors (EGFR, IGF-R1 and HER2) and GPCR cause phosphate-dependant activation of ER, which then goes on to bind ERE and alter transcription

1.2 Endocrine therapies and resistance in Breast Cancer

1.2.1 Hormones and breast cancer

In the past it was the standard of care to prescribe hormone replacement therapy (HRT) for post-menopausal women. These were mainly combined oestrogen and progesterone hormones and aimed to reduce menopause symptoms, such as: hot flushes, mood swings and vaginal dryness. However, a correlation between combined HRT and increased risk of breast cancer has been established necessitating reconsiderations such as shortening the exposure duration, using oestrogen-only HRT or alternative therapies (Vinogradova, Coupland and Hippisley-Cox, 2020; Stuenkel et al., 2015; Bassuk and Manson, 2014; Crosignani, 2003).

The direct relationship between oestrogen hormone and breast cancer was firmly evidenced many decades ago (Dimitrios, Brian and Philip, 1972). Oestrogen's major role in breast tissue is to enhance mammary cell proliferation and differentiation (Yaghjyan and Colditz, 2011). It was proved that the longer the exposure to endogenous oestrogen, the higher the risk of breast cancer – e.g., early menarche and late menopause (Hamajima *et al.*, 2012). After menopause, while oestrogen production cesses from the ovaries, oestrogen gets produced from peripheral body parts, mainly adipose tissue, by converting circulating androstenedione to oestrogen by enzyme aromatase (Hetemäki *et al.*, 2017). Hence obesity is more significant as a risk factor for breast cancer in post- than in pre-menopause (Gravena *et al.*, 2018). It was clearly concluded that the cumulative, excessive, and prolonged exposure to oestrogen is one of the most important predisposing factors to developing oestrogen receptor positive breast cancer.

1.2.2 Targeted Endocrine therapies for ER-positive breast cancer

The treatment strategies for breast cancer have evolved dramatically over the past decades. Early on, breast cancer would be treated with surgical (radical mastectomy), irradiation and chemical castration, hypophysectomy (removal of pituitary glands) and bilateral adrenalectomy (removal of adrenal glands) in a sequential manner (Kennedy, 1965). Use of combined chemotherapeutic regimens was then in fashion during the seventies (Cooper, 1969) (Hortobagyi, 2000). Such old, nonspecific treatment strategies of breast cancer were harsh, invasive, delayed and based on trial and error, with modest survival improvement.

Given that oestrogen is the main influencer of ER-positive breast cancer cells and that it functions by reacting with oestrogen receptors forming an active oestrogen-ER complex, hindering this interaction is a logical target in treating ER-positive breast cancer. Three hormonal drug classes are available, each with a distinct mechanism of action: (1) selective oestrogen receptor modulators (SERMs): e.g., Tamoxifen and Raloxifene, these competitively bind the oestrogen receptor blocking it's growth and proliferation enhancing actions. (2) selective oestrogen receptor down-regulators (SERD): e.g. Fulvastrant, which mediates oestrogen receptor destruction and degradation, counteracting ER upregulation, and (3) aromatase inhibitors (AI): e.g. Letrozole, Anastrozole and Exemestane, these block the action of aromatase enzyme and thus reduce oestrogen synthesis (Burstein *et al.*, 2014). The focus of this thesis will be on tamoxifen resistance, other types of endocrine resistance are covered in many reviews (e.g., (Macedo, Sabnis and Brodie, 2008) (Ma *et al.*, 2015) (Huang *et al.*, 2017) (de Marchi *et al.*, 2016).

1.2.3 Tamoxifen

Previously known as compound I.C.I. 46,474, Tamoxifen is a trans-isomer of a substituted triphenylethylene, nonsteroidal antioestrogen (Walpole and Harper, 1966). It was first synthesized and tested as postcoital contraception (HARPER and WALPOLE, 1967). However, on the contrary, it was shown to stimulate ovulation function (Williamson and Ellis, 1973). It's potential as an anticancer was first tested in advanced breast cancer patients, but results were modest (Cole, Jones and Todd, 1971). Despite the unencouraging results, researching tamoxifen potential as an antioestrogen therapy for breast cancer continued through the 1970s (Jordan, 2008). It gained approval in the 1977, initially as a palliative therapy for metastatic breast cancer (Kiang, 1977). Subsequent evidence showed unprecedented qualities of tamoxifen compared to standard chemotherapies, such as: selectivity - tissue specific oestrogen regulatory function - (Radin and Patel, 2016) and chemoprotection, reducing the incidence of breast cancer for high-risk patients (Fisher *et al.*, 1998). The eventual introduction of tamoxifen as standard therapy dramatically improved the clinical outcomes of breast cancer patients, most importantly in terms of decreasing mortality rate by 31% and improving relapse-free survival (Abe *et al.*, 2005). Currently, a highly recommended 5-year adjuvant course, extendable to 10 years, is now the standard of care for oestrogen receptor positive breast cancers (Burstein *et al.*, 2014). A timeline of development of tamoxifen is shown in Figure 1.4.

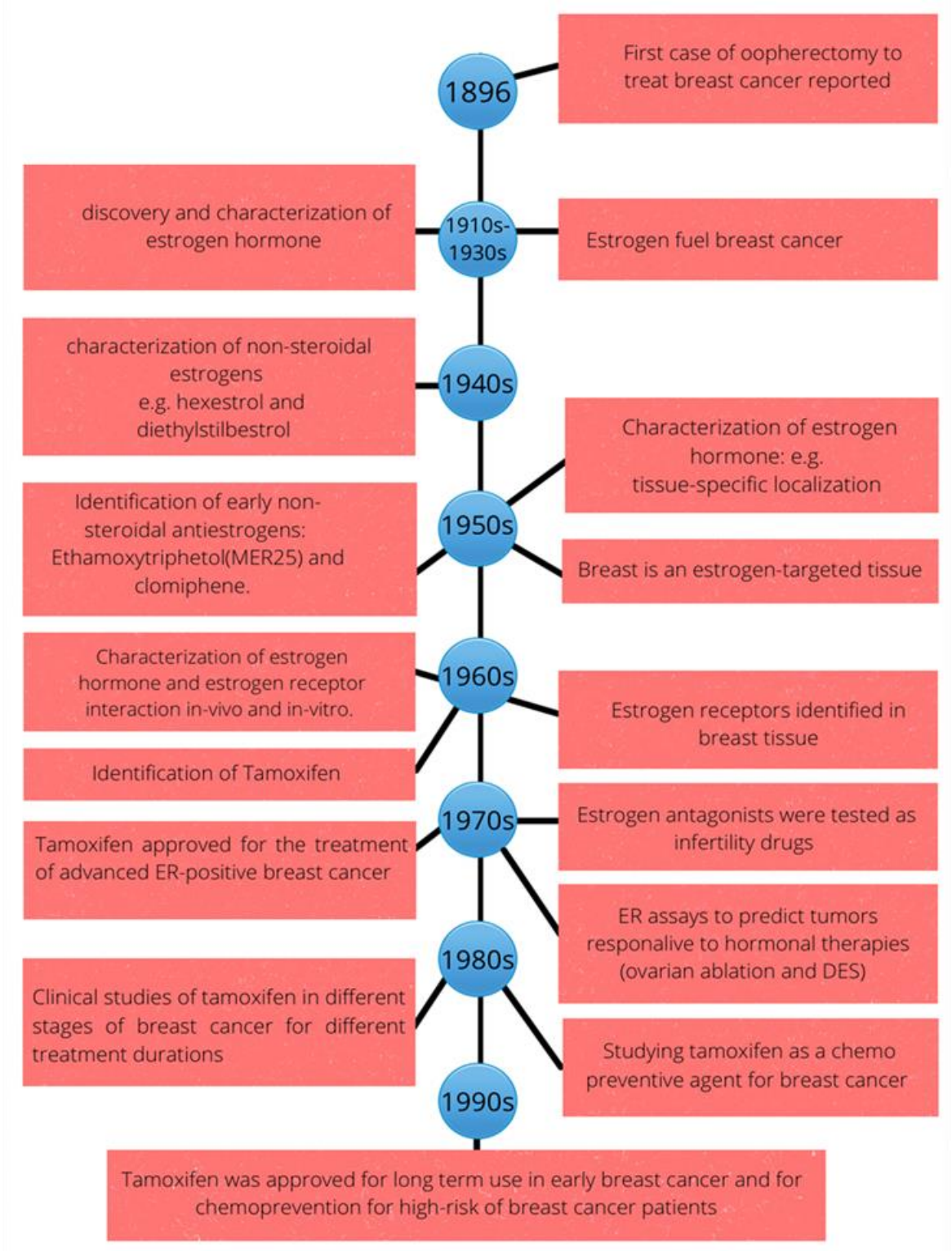


Figure 1.4. Historical sequence of events leading to the approval of Tamoxifen for breast cancer.

Tamoxifen is known to be a pro-drug so therapeutic effects are exhibited through hepatic metabolic activation by Cytochrome P450 enzymes, to produce active metabolites, mostly 4-hydroxy tamoxifen (4-OHT) and endoxifen (Lim *et al.*, 2006). It's unique chemical structure closely resembles that of oestrogen with some chemical and structural differences making it of higher affinity for ER than oestrogen hormone (Nilsson *et al.*, 2001). Tamoxifen mainly acts as a cytostatic agent, terminating pre-cancerous cell cycle progression at G0, rather than being cytotoxic to the already matured cancerous cells (Dalvai and Bystricky, 2010).

1.3 Endocrine resistance in breast cancer

While endocrine therapies are proven effective in ER-positive breast cancer treatment and prevention, endocrine resistance, *de-novo* and acquired is reported to be as high as 40 % (Lei *et al.*, 2019). In theory, the overexpression of ERα in breast cancer should be the definite indication of positive response to endocrine therapy. However, this is not always the case, approximately half of ERα positive advanced breast cancers is not responsive to tamoxifen and more than one third of initially responsive patients end up with endocrine resistance, interrupting their course of treatment. These patients often feature local or distant metastasis, and/or primary tumour progression and recurrence (Normanno *et al.*, 2005).

1.3.1 Mechanisms of tamoxifen/endocrine resistance

Tamoxifen competitively binds and effectively blocks ERα, However, with continuous prolonged exposure, cancer cells start to desensitize and become resistant regardless of their oestrogen receptor positivity (de Marchi *et al.*, 2016).

Well-defined pathways of resistance include defective ER expression and aberrant molecular interactions with growth-related transcription factors resulting in uncontrolled ER-signaling. In addition, the presence of crosstalk between ERs and growth factor receptors pathways (e.g., EGFR and HER2) that act through many intracellular signaling cascades (PI3K/AKT/mTOR and MAPK/ERK) to control differentiation and cell cycle progression can contribute to resistance (details follow in the following sections). It should be noted that these pathways are tightly regulated under normal physiological conditions, but they get out of balance usually due to carcinogenic transformations in the breast cancer genome or in response to external stimulus such as endocrine therapies. Overview of mechanism of resistance is shown in Figure 1.5.

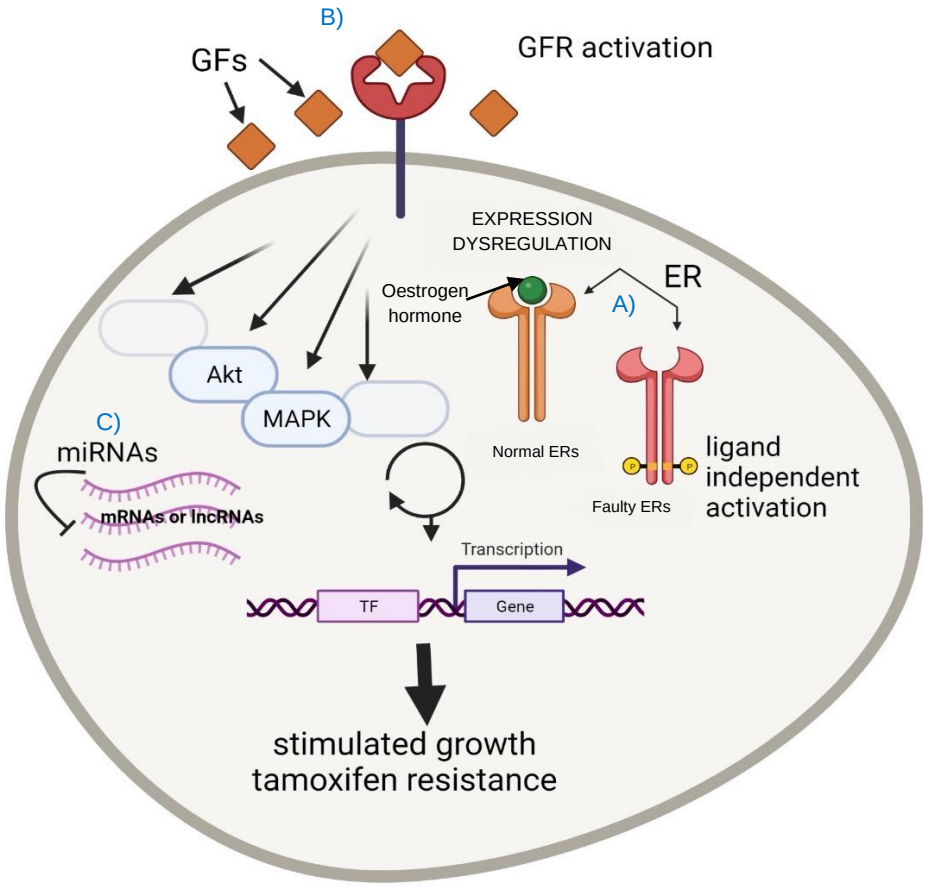


Figure 1.5. Mechanisms of tamoxifen resistance in Oestrogen receptor positive breast cancers. This includes a) loss or hyperactivation of ESR1 expression or producing faulty ER prone to ligand independent activation by phosphorylation. B) GFRs are central contributors in driving tamoxifen resistance, where growth factors such as IGF, EGF and TGF bind GFR initiating cellular growth signalling cascades (AkT and MAPK and others). C) miRNAs have a widespread regulatory function and can participate in resistance by freeing RNAs from their control. Tamoxifen resistance is a complex interconnected pathways that can involve one or a combination of mechanisms that leads to cancer related TF and genes stimulation. GFR denotes growth factor receptor, GF is growth factor, ER is estrogen receptor, p is phosphate, AkT is activated tyrosine kinase, MAPK is mitogen-activated protein kinase, empty oblongs denotes other protein kinases and TF is transcription factor.

1.3.2 Mutations/ alterations in ER molecules

Many genomic aberrations in the gene encoding ERα itself, ESR1, cause structural and functional alterations in the receptor and have been linked to endocrine resistance pathways. Loss of ER expression as a genotype is the main cause of de novo endocrine resistance phenotype in triple negative breast cancers. In acquired endocrine resistance, this relation is much more complicated as ER downregulation has been observed in only ~15% of resistant patients (Encarnacion *et al.*, 1993; Yao *et al.*, 2013). While it is not completely understood if ER loss is the cause or a consequence in the resistance pathway, it has been correlated with growth factor receptor (GFR) (e.g., HER2 and EGFR) overexpression that seem to supress ER expression to take over the control on growth and proliferation (Gutierrez *et al.*, 2005; Osborne and Schiff, 2011).

Somatic point mutations, mostly affecting the part encoding LBD on the ESR1 gene, give rise to a truncated form of ERα, capable of ligand-independent autoactivation and causing conformational change that makes ligand binding sites inaccessible to tamoxifen. Most frequent ESR1 mutations are Y537S and D538G (in 20 – 30 % of cases) (Merenbakh-Lamin *et al.*, 2013; Fanning *et al.*, 2016), less frequent mutations include Y537N, Y537C and L536R, (in 3 - 13 % of cases). Interestingly mutations were mostly observed in advanced breast cancers pre-treated with endocrine therapies rather than endocrine therapy-unexposed ER-positive or ER-negative subtypes (Jeselsohn *et al.*, 2014; Fanning *et al.*, 2016). This observation suggests a probable ESR1 genetic scarring caused by endocrine therapies making tumours prone to endocrine resistance.

Other possible genomic alterations related to resistance include alternative splicing and exon skipping that produce functionally distinctive isoforms such as ERα-36, ERα-46 and types1, 2 and 3 exon skipping splice variants that have reluctant binding abilities to tamoxifen (Klinge *et al.*, 2010; Li *et al.*, 2013; Lee *et al.*, 2016). In addition, genomic rearrangements and translocations that give raise to ESR1–CCDC170 gene fusions were observed in 4% of breast cancer cases examined (8 out of 200 ER-positive tumours) and were especially reported in the aggressive resistant form of breast cancer, this genetic defect produced ΔCCDC170 protein that is believed to participate in malignant progression and resistant transformation (Veeraraghavan *et al.*, 2014).

1.3.3 Dysregulation of alternative reactive elements signalling pathway

The transcription of oestrogen responsive genes can be activated through other mechanisms and can compensate for the blockage of ER signalling by antioestrogens. Upregulation of TFs known to act on genes associated with proliferation and tumorigenesis, has been linked to endocrine resistance (Dixon, 2014). For example, NF-κB. NF-κB is a TF normally found in an inactive form in the cytoplasm under the inhibitory effect of IκB. Upstream stimuli such as stress and growth factor receptor activation (e.g., EGFR, FGFR and IGF-1R) induce the proteolysis of IκB. Dissociation of IKK from NF-κB, activates NF-κB, that in turn translocates to the nucleus and acts as an alternative TF for ER signalling (deGraffenried, 2004). ER-negative and tamoxifen resistant breast cancer cell lines show much greater DNA-binding affinities to NF-κB than ER-positive and tamoxifen sensitive cell lines (Yamamoto and Gaynor, 2001). Furthermore, combining IKK

inhibitor PA with tamoxifen re-sensitized resistant cell lines (HER2-positive and tamoxifen resistant ER-positive) (Zhou *et al.*, 2005).

1.3.4 Upregulation of ligand-independent signalling pathway

One of the most prominent mediators of tamoxifen resistance is the overexpression of transmembrane tyrosine kinase growth factor receptors, that cause growth and proliferation in a ligand-independent manner (Figure 1.3.D) (Paplomata and O'Regan, 2014). GFRs such as EGFR, FGFR, HER2 and IGF-1R bind their ligands and cause subsequent crosstalk between ERα and many downstream kinases: mitogen-activated protein kinase (MAKP), phosphoinositide 3-kinase (PI3K) and protein kinase B (AKT) that lead to ER activation by phosphorylation. All these molecules have been reported to be upregulated in breast cancer and linked to the development of endocrine resistance (Davis *et al.*, 2014). Another subset of receptors that crosstalk with the ER signaling pathway and are found to be upregulated in endocrine resistant breast cancer cell lines are GPCRs, that act through second messenger signalling (cAMP, Ca2+, IP3), to activate PKA and induce ERα phosphorylation and activation of cyclic A (Kulkoyluoglu and Madak-Erdogan, 2016) (Figure 1.3.C).

1.3.5 Overactivation of membrane ER signaling pathway.

Studying ERα knockout mice models proved that malignant breast tumour growth is not solely dependent on the expression of ERα (Bocchinfuso *et al.*, 1999). Another proposed alternative is E2 binding to cell surface G-protein coupled oestrogen receptors (GPER) (Figure 1.3.C). Many studies have argued the identity of GPER as an oestrogen adapter and E2 as a potential ligand (Otto *et al.*, 2008,

2009; Kang *et al.*, 2010). Others have suggested GPER may have a tumour suppressor role in ER-positive breast (Ariazi *et al.*, 2010). On the other hand, some studies have linked its overexpression to triple negative resistant phenotype and to HER-2/EGFR overexpression (Ignatov *et al.*, 2011). This controversy is strongly inclined toward expression being an endocrine resistance initiator and promoter of invasion and migration (Otto *et al.*, 2008), and might be explained by the presence of other factors controlling this receptor activation, suppression, interactions with other molecules.

1.3.6 Altered microRNAs (miRNA)

miRNAs are short single stranded non-coding RNA molecules of ~ 22 nucleotides, they predominantly function through post-transcriptional silencing of targeted RNA transcripts. Regulation of gene expression occurs through two main mechanisms, either by cleaving RNA sequences to induce their degradation or by altering the sequence, blocking the process of translation (Shyu, Wilkinson and van Hoof, 2008). Genomic analysis found more than half of miRNAs originate from cancer-associated genomic regions (Calin *et al.*, 2004) and many miRNAs have been linked to cancer related cellular processes. For example, miR-126 has been shown to have a role in proliferation (Guo *et al.*, 2008), miR-24 in apoptosis (Qin *et al.*, 2010) and miR-181d in DNA repair (Zhang *et al.*, 2012). Classified as oncogenes or tumour suppressors, miRNAs can directly or indirectly release inhibited pro- or anti- carcinogenesis pathways (Zhang *et al.*, 2007). There are several reports of miRNAs alterations being significantly altered in tamoxifen resistant phenotypes. For example, manipulation of miR-342-5p expression level impacted tamoxifen sensitivity and was linked to changes in several cancer pathways (Cittelly *et al.*,

2010). Another tamoxifen response related miRNA where reduced expression linked to poor prognosis is miR-26a (Jansen *et al.*, 2012). On the other hand, miR-210, miR-126 and mir-10a were all more highly expressed in ER-positive breast cancers with worse prognosis and tamoxifen response (Rothé *et al.*, 2011; Hoppe *et al.*, 2013).

1.3.7 DNA damage Response

An essential biological strategy to contain acquired mutational events that could lead to resistance, and which might be caused spontaneously, or by radiotherapy or cytotoxic drugs is the DNA damage response (DDR) (Jeggo, Pearl and Carr, 2016). DDR is a complex pathway responsible for detecting and repairing DNA damage, essential for genomic stability and ultimately restraining the propagation of damaged cancerous cells(reference). Following DNA damage detection, series of signaling cascades are initiated, aiming to repair the DNA damage, regulation of cell proliferation and cell cycle, or opting for programmed cell death (apoptosis). In the context of tamoxifen resistance, defective DDR can lead to genomic damage not effectively repaired and consequently the accumulation of pathological mutations related to resistance pathway(reference). Defective DDR especially homologous recombination (HR) has been observed in most resistant breast cancer subtypes (Triple negative and hereditary breast cancer mutations: BRCA1 and BRCA2) (Turner and Reis-Filho, 2006). HR-associated protein p53 accumulation and TP53 mutations are frequent in breast cancers (~30%) and were linked to increased susceptibility to developing endocrine resistance (Yamashita *et al.*, 2006).

1.4 Genomic information

1.4.1 Central dogma of molecular biology

While DNA may contain the permanent copy of the genome itself, RNA has the final say on protein production. RNA is a polymeric molecule of nucleotides of single stranded sequence, made from DNA by RNA polymerase and transcribed transcripts possess a diverse set of functions. The structural properties of RNA allow for freedom of movement and adoption of different functional structures; however, this negatively affects the stability and degradability. Many types of RNAs have been identified, either protein coding (messenger RNA) or non-coding - ribosomal RNA (rRNA) the predominant form in cells, transfer RNA (tRNA), long noncoding RNAs, microRNA (miRNA), small interfering RNA (siRNA), small nucleolar RNA (snoRNAs), and Piwi-interacting RNA (piRNA). RNA has a diverse range of functions either as a distinct entity (rRNA and tRNA) or as an intermediary molecule serving as a catalyst of cellular reactions (siRNA and microRNA) or having a regulatory role (lncRNA, snoRNAs and piRNA) (Table 1.3).

RNA Type	Size	Function	Example	Reference
mRNA	>200 nucleotides	Curry genetic sequence for protein synthesis.	BRCA, ESR1 and PD-1.	(Mei et al., 2020)
rRNA	Variable sizes	Form ribosome organelles responsible for protein synthesis	5S rRNA, 16S rRNA and 23S rRNA.	(Harold et al., 2021)
tRNA	75 to 90 nucleotides	Intermediate molecule between mRNA and protein, mainly transfer amino acids to the ribosome during protein synthesis.	$tRNA^{Arg(CCU)}$, $tRNA^{Tyr(GUA)}$, $tRNA^{Ser(GCU)}$	(Pavon-Eternod et al., 2009)
miRNA	20-25 nucleotides	Gene expression regulation by mRNA degradation.	miR-92a-3p, miR-23b-3p and miR-191-5p	(Sharifi et al., 2022)
siRNA	20-26 nucleotides	RNA interference-repressing translation process.	Related to the targeted gene.	(Tyagi et al., 2017)
lncRNA	>200 nucleotides	Molecular regulation of cellular processes.	MALAT1, HOTAIR and GATA3	(Xu et al., 2017)
piRNAs	24-31 nucleotides	regulation of transposable elements, epigenetics, and gene expression.	piR-36712, piR-62011, piR-49145.	(Tan et al., 2019)
snoRNA	60-300 nucleotides	biogenesis and maturation of other RNA molecules.	SNORA38	(Song et al., 2022)

Table 1.3. RNA species and their known functions.

The concept of central dogma describes the basic logic of the genetic code (Burian and Barbieri, 2015). However, the continuous information flow (DNA > RNA > Protein) has been revisited many times due to a lot of findings that extended or contradicted the way the central dogma was conceived originally. In alternative pathways genetic information flow is reversed. For example, reverse transcription in retroviruses, where DNA is synthesised from RNA, not the reverse, it gets integrated in the host original genome, pathologically affecting cellular integrity (Weiss, 1998) Additionally, the transcription step can be skipped such as in the case of RNA viruses where upon infection, viral RNA is translated into a protein directly (Wu and White, 2007). Most interestingly, the discovery of non-coding RNAs showed that some transcripts skip the process of translation and act directly on targeted molecules driving their regulation (Ponting, Oliver and Reik, 2009). In this thesis, we studied the role of one class of non-coding RNAs, long non-coding RNAs (lncRNAs), in endocrine resistance in breast cancer using *in-silico* and *in-vitro* approaches.

1.4.2 lncRNAs

lncRNAs are classified according to their relationship with the adjacent protein coding genes: linc RNA is intergenic lncRNA and is transcribed from the region between two genes; antisense lncRNA is transcribed from the antisense strand; sense lncRNA is transcribed from the sense strand and intronic lncRNA is transcribed from an intronic region (Ponting, Oliver and Reik, 2009). Influencing protein-coding genes' expression by lncRNAs may occur in cis (proximal) or in trans (distal) relative to their site of transcription(Wang and Chang, 2011). This fact guided the search for functional lncRNAs, and many were discovered through

searching the active transcription sites for lncRNA signatures. Correlating the expression of lncRNAs to protein coding genes known to be related to endocrine resistance is a useful technique to predict distinct functions and behaviours of lncRNAs.

1.4.2.1 Functions of lncRNAs

LncRNAs have a specific pattern of expression and unique distribution in a wide spectrum of cellular events and compartments, where the activation and suppression of their transcription is tightly regulated, and varies with cellular conditions such as proliferation, development, aging and disease (Ponting, Oliver and Reik, 2009). The expression of lncRNAs can be controlled by classical transcription factors (e.g., NF-KB regulates AK019103) (Wan, Mathur, *et al.*, 2013), coding genes (e.g., c-Myc regulates HOTAIR) (Li et al., 2016); miRNA (e.g. miR-141 regulates HOTAIR) (Chiyomaru *et al.*, 2014); and epigenetic processes (e.g. hypermethylation supresses MEG3) (Wang and Chang, 2011). In turn, lncRNAs can control the processes of transcription, translation, and posttranslational modification, either directly or by interacting with other regulatory molecules (transcription factors, proteins and micro RNAs). This bidirectional regulatory mechanism forms a complex network around many recognized cancer pathways. This is evident with the network of lncRNAs around p53 which has a significant role as a tumour suppressor by directing the transcription of many genes involved in apoptosis, DNA repair and cell cycle arrest in response to DNA damage, and many cancers associated lncRNAs found upstream and downstream of p53 as shown in (figure 1.7).

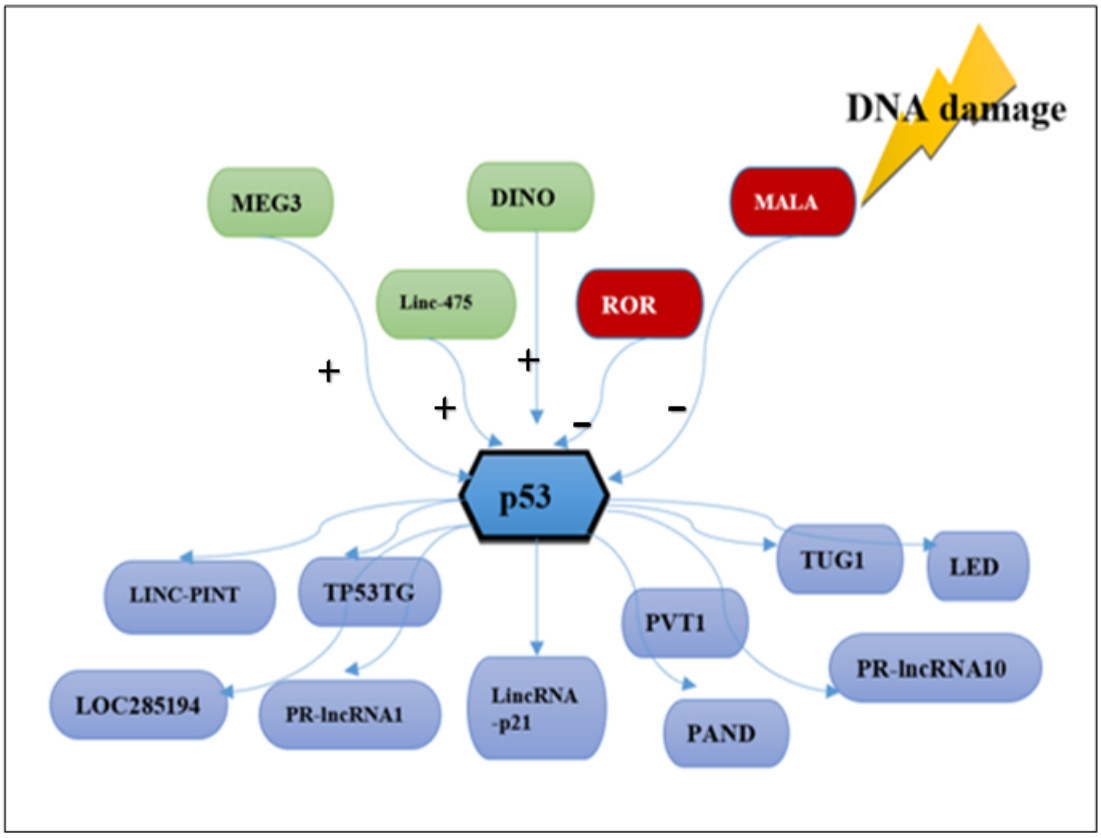


Figure 1.6 p53 and lncRNAs network. DNA damage induces the activity of many p53 associated lncRNAs that either activate or supress its activity. When activated, p53 in turn affect the expression of other lncRNAs that regulate the transcription of p53-related genes involved in DNA repair response.

LncRNAs act mainly by orchestrating chromatin modifications to enhance or silence target gene transcription (Khalil *et al.*, 2009), for example lncRNA XIST directs X chromosome inactivation in females during embryogenesis via chromosome structural changes (Sun, Deaton and Lee, 2006). Also, lncRNAs can serve as scaffolds that bind and guide the interaction of several molecules at once such as the HOTAIR (HBXIP/LSD1) complex during regulation of c-Myc (Li et al., 2016). In

addition, many lncRNAs can influence alternative mRNA splicing, protein localization, act as precursors to small RNAs, act as decoys directing the destruction of many RNA species and many other functions, reviewed in:(Wang and Chang, 2011). With their diverse functions, lncRNAs have become increasingly recognized to be a significant contributor to the development and progression of pathological conditions. Pertaining to treatment resistance, lncRNAs was found to alter drug response through several mechanisms. Drug efflux and drug metabolism was affected by many lncRNAs such as MALAT1, that negatively affects the expression of the efflux transporters MRP1 and MDR1, contributing to chemoresistance (Fang et al., 2018). Drug targets expression such as apoptosis-related proteins was confirmed to be a target of lncRNAs HOTAIR (Liu et al., 2013) and XIST (Xu et al., 2020) inhibiting apoptosis and disrupting cell cycle prpgression. The regulatory role of lncRNAs extend to function as an Immunomodulators affecting various immune processes and anti-inflammatory factors, such as lncRNAs NEAT1, LINC00473 and NKX2-1-AS1 that create an immune escape resisting immunotherapy (Zhou et al., 2019).

1.4.2.2 lncRNAs role in endocrine resistance

Given what is already known about the function of lncRNAs, they form suitable candidates as regulators of endocrine resistance in breast cancer (Terai *et al.*, 2016). Indeed, emerging data from deep sequencing of multiple breast cancer genomes, suggests the presence of a considerable number of lncRNAs differentially expressed between sensitive and resistant tissues. The predicted functions of these lncRNAs span the range of oncogenic events linked to endocrine resistance including ER signalling, enhanced growth factor receptor signalling and

altered DNA damage response (Table 1.3). In addition to those listed, NF-κB interacting LncRNA (NKILA) was found to supress NF-κB signaling by interfering with IKB phosphorylation (Liu *et al.*, 2015). However, this mechanism of action was questioned by (Dijkstra and Alexander, 2015), who suggested that NKILA is transcribed from and regulates the prostate transmembrane protein androgen induced 1 (PMEPA1) gene that in turn directly affects the NF-κB pathway. Regardless of this dispute, decreased expression of NKILA, increased expression of MALAT1, HOTAIR and lincRNA-p21 and others all enhance the expression of NF-κB in many types of cancers (Mao, Su and Mookhtiar, 2017). Further support for lncRNAs being associated with resistance comes from a study that used a different approach to identify lncRNAs potentially associated with endocrine resistance. They identified more than 30 oestrogen receptor agitation related (ERAR) lncRNAs, many could be used to classify ER+ tumours as high or low risk of endocrine resistance (Wu *et al.*, 2016). A table of current literature is included Table 1.3.

Considering the issue of endocrine resistance from a wider perspective, defective DNA damage response (DDR) is a logical causative factor for many genomic defects. Many lncRNAs have been linked to DDR pathways, for example the complex network of lncRNAs around p53 (Figure 1.7). Another effector molecule in DDR is Ataxia-telangiectasia mutated (ATM) kinase, a key respondent to DNA double-strand breaks. Among the lncRNAs associated with this molecule are LncRNA-JADE and ANRIL (antisense non-coding RNA in the INK4 locus), both are induced by DNA damage in an ATM dependent manner and mediate subsequent

regulation of downstream proteins such as p53 and BRCA1 (Wan, Hu, *et al.*, 2013; Wan, Mathur, *et al.*, 2013). However, their role in endocrine resistance in still to be investigated.

Endocrine resistance is a molecular defence mechanism of cancer cells against endocrine therapies, compensating for the growth restrictions imposed. Studying the molecular mechanism of endocrine resistance and different variables contributing to tamoxifen response is an active and promising field of research. the expression status of hundreds of lncRNAs was found to be distinctively up or downregulated in comparative analysis of drug resistance conditions, implying that lncRNAs can provide insights into the underlying mechanisms of gene regulation and provide reliable biomarkers and effective therapeutic targets. However, the number of studies investigating the mechanistic role of lncRNAs in endocrine resistance in breast cancer is relatively minor compared to the massive number of differentially expressed lncRNAs in different forms of breast cancer either sensitive or resistant to endocrine therapies.

lncRNA	Type	Role in endocrine resistance	Reference
BCAR4	Antisense Linc RNA	-overexpression of BCAR4 promotes the phosphorylation and activation of growth signalling pathways independent of ER this includes (ERBB2, ERBB3, MAPK and AKT)	(Godinho *et al.*, 2010)
UCA1	Linc RNA	-tamoxifen treatment increases the expression of HIF1α that upregulates UCA1 that inhibits miR-18a that is in turn an important inhibitor of HIF1α (miR-18a HIF1α feedback loop) and DNA repair.	(X. Li *et al.*, 2016)
ROR	Linc RNA	-upregulation of ROR leads to decreased expression of miRNA (miR-205-5p) and increased expression of the protein coding genes (ZEB1 and ZEB2). -act as repressor to p53	(Zhang *et al.*, 2013, 2017)
HOTAIR	Antisense linc RNA	Upregulated HOTAIR increases ESR1 transcription and leads to overexpression of ER.	(Xue *et al.*, 2016)
MIR2052 HG	-	Overexpression of MIR2052HG increases expression of ER by enhancing the AKT/FOXO3 pathway and supresses ER proteolysis mediated inactivation.	(Ingle *et al.*, 2016)
CCAT2	Sense Linc RNA	-upregulation of CCAT2 promotes endocrine resistance through the ERK/MAPK signalling pathway	(Caia, He and Zhang, 2016)
H19	Maternally imprinted linc RNA	-upregulated H19 is related to the NOTCH4 receptor (NR4) signaling pathway	(Basak *et al.*, 2017)
DSCAM-AS1	Antisense	-by interacting with hnRNPL and PCBP2 that are protein known to be associated with RNA stability and processing.	(Niknafs *et al.*, 2016)
LINC00978	Linc RNA	-upregulation is associated with decreased expression of ER.	(Deng *et al.*, 2016)

Table 1.4 lnc RNAs linked directly to endocrine resistance.

ROR (regulator of reprogramming), EMT (epithelial mesenchymal transition), UCA1 (Urothelial carcinoma associated 1), BCAR4 (Breast cancer anti-estrogen resistance 4), HIF1α (Hypoxia-inducible factor 1-alpha), HOTAIR (HOX transcript antisense RNA), ZEB1,2 (Zinc finger E-box-binding homeobox 1,2), CCAT2 (Colon Cancer Associated Transcript 2), DSCAM-AS1 (DSCAM antisense RNA 1)

1.5 Studying the genome(techniques)

.1.5.1 The human genome project

The Human Genome Project's (HGP) main aim was to study the DNA molecules in human cells physically and functionally. Genes were sequenced, identified, mapped, and assembled. HGP revolutionised scientific research by revealing the genetic root of cellular processes and dysfunctions, hence laying the ground for biomarkers and drug discoveries. The greatest accomplishment for HGP was the creation of the human reference genome; allowing for comprehensive and comparative genetic research and considered the long-waited solid foundation of molecular biology (Gates *et al.*, 2021). The project was initiated in 1990 and released the final version 13 years later in 2003, costing the US government about 3 billion dollars (National Human Genome Research Institute (NHGRI), 2020). Further development of the project produced improved versions with less error rates and less gaps. The latest complete version was released publicly in early 2022 and consisted of 3,117,275,501 sequenced base pairs representing the final sequence of the 23 chromosomes (T2T Consortium, 2022). Research in the aftermath of the HGP has a focus bias towards the protein coding landscape of the genome with most studies revolving around widely known genes (Gates *et al.*, 2021). After characterising the permanent hub of genomic information by HGP, where only a tiny part of the DNA encodes protein coding genes (Birney *et al.*, 2007), the worth of the profound amount of 'junk' DNA sequence now needs to be studied. In addition, study of what was determined later as 'the effector molecule' (the RNA or transcriptome) is needed to understand the convoluted physiology of biological processes such as cellular proliferation and differentiation, and the pathophysiology

of complex diseases such as cancer and autoimmune disorders. Ever-growing research technologies in many areas including sequencing platforms and bioinformatics permitted The Encyclopaedia of DNA Elements (ENCODE) project to launch as follow up to HGP. ENDCODE Transcriptome sequencing aims to inspect functional genomics, gene discovery and novel protein characterisation (de Souza, 2012). Compared to the DNA molecule, RNA and other genomic elements show a profound depth and diverse range of activity. Type of cell, pathological condition, and even external factors like environment, affect the transcriptome activity in a sequential manner (Kotsantis *et al.*, 2016). Gene expression quantification is a fundamental way to study how genotypic alterations are expressed as a specific phenotype (Deonarine *et al.*, 2007). Gene expression profiling or sequencing is the process that allows detection of simultaneous changes in a large set of genes in controlled condition at a certain moment (Kotsantis *et al.*, 2016).

1.5.2 Next generation sequencing (NGS)

The most recent development in sequencing technologies is next-generation sequencing, that enable high throughput quantitative genomic and transcriptomic study. Here DNA/RNA is fragmented to multiple bits (called libraries), adapters are added, the whole lot is sequenced and then a genomic sequence is computationally reassembled. In the field of next generation sequencing, Illumina is the lead manufacturer and developer globally. The platform's series includes Genome Analyzer, HiSeq, MiSeq and the HiScanSQ (Bentley et al., 2008). More recently, the company introduced NovaSeq sequencing system, with a wider range of

applications and noticeably improved performance(Liu et al., 2022). As per illumine data, NovaSeq6000 machine can produce up to 6 terabases in a single run, it offers more simplified and scalable workflow with shorter running time at lower cost (Illumina, 2016).

1.5.3 Determining the transcriptome

RNA molecules are dynamic, that means they undergo multiple changes at different times such as alternative splicing (Gallego-Paez *et al.*, 2017), alternative polyadenylation (APA) (Xia *et al.*, 2014) and post-transcriptional modification (Seelam, Sharma and Mitra, 2017). Most importantly, the functional spectrum of RNA extends to controlling the activity of other genomic elements including DNA methylation and histone modifications and post translation modification. Modern technologies enable gene expression quantification at a global level (Takahashi *et al.*, 2015). An unprecedented flow of genomic information provides a full-scale catalogue of sequencing data from different organisms and disease-phenotypic samples. Many bioinformatic technologies and datasets are available for studying many genomic processes.

1.5.4 RNA-seq

RNA-seq has emerged as a strong NGS technique to study the gene expression and the molecular regulatory relationships within the genome. Following genetic material isolation from target cells and genomic DNA removal, total RNA is further purified and assessed for integrity, quality, and amount. The predominant form of RNA is ribosomal RNA (rRNA) (80-90%), a non-protein coding RNA that functions in the process of protein translation (Henras *et al.*, 2015). Since sequencing rRNA

is not needed for gene profiling, sequencing raw total RNA without targeted population selection is a waste of time and finance. So, to ensure a strong signal is attained from sequencing (Flannigan *et al.*, 2017), the first step in the sequencing workflow is depletion of unwanted RNA subtypes and enrichment for the wanted RNA-population. Several library prep kits and techniques are available for different purposes and platforms (Aigrain, Gu and Quail, 2016). Magnetic bead-based methods are the most popular, of the large pool of total RNA, poly A tail on target RNA transcripts gets captured onto complementing poly T oligos on the beads, loose RNA then gets washed away (Z. Li et al., 2021). Alternatively, probes can be hybridized on rRNA sequences, the reaction is then mixed with a specific type of bead that pulls down unwanted rRNAs to be discarded (Hinahon *et al.*, 2013). To be compatible with the sequencing platform, the now enriched RNA is sheared, into fragments appropriately sized for the sequencer (e.g., by nebulization, sonication, hydrodynamic shear, or transposase) (Sambrook and Russell, 2006) and converted to double stranded cDNAs, if required. Adapter ligation is the next step; specially designed oligos (barcoded adapters) are attached to both ends of the fragment. This introduces sequencing primer hybridization sites for the next step and allows for indexing of the fragments so that many samples can be run in the same sequencing reaction (Ring *et al.*, 2017). A more progressive option is Tagmentation, that is transposomes technology, which combines both approaches (fragmentation and tagging), allowing reduced library preparation time and reducing the amount of genomic material required and resulting in more uniform and consistent library yield (Adey and Shendure, 2020). Finally, satisfactory library quantification and quality control results in approval of samples for sequencing. Sequencing is initiated by a

cluster generation step that occurs within the specially designed glass slide called a ‘flowcell’. It starts with cDNA fragments ‘flowing’ into the flowcell where the ligated adapters get captured onto complementary surface-bound oligos. There are then several rounds of isothermal bridging amplification which prepares the cDNA fragments for sequencing (Holt and Jones, 2008). Illumine platforms predominantly use sequencing-by-synthesis (SBS) technology, a reversible terminator–based method. Here, only forward, or reverse strand clusters are massively parallel sequenced simultaneously in one read. Florescent tagged nucleotides are added to the sequencing primer, after each sequencing cycle, tagged nucleotides get excited by a light source emitting a specific signal, the image is then captured, and the corresponding nucleobases are called. This is a base-by-base sequencing, so the number of cycles reflects the length of the sequenced fragment (Turcatti *et al.*, 2008). Other next-generation sequencing manufacturers and systems commercially available include Oxford Nanopore Technologies (Wood *et al.*, 2019), Ion Torrent (Vanni *et al.*, 2015), Roche (Nielsen *et al.*, 2014), Complete Genomics (Weißbach *et al.*, 2021), Helicos BioSciences (Milos, 2008).

In general, simultaneous measurement of gene expression provides insight into the continuously changing cellular transcriptome. Carefully constructed experimental designs make use of comparative analyses to relate genotype to biological diversity and dysregulated cellular functions. Utilizing RNA-seq data is one of the most powerful methods in the path of dissecting mechanisms of biological processes.

1.5.3 Microarray technology

Another method to measure gene expression levels and determine DNA/RNA sequence is microarrays, a hybridisation-based method (Wang, Gerstein and

Snyder, 2010). Microarray is a chip base genome array that contains a certain number of known gene sequence probe sets, this allows for selecting the genetic traits under query. By simultaneously detecting thousands of genes parallelly, the relative abundance of genes in many samples can be determined precisely in a single experiment (Alonso-Betanzos, 2019). Microarray chip or slide's unique design comprise spots featuring oligonucleotide probes to which complementary target sequence binds (Tsoi and Zheng, 2007). Extracted sample RNA/DNA is fluorescently tagged and sequence hybridisation generates a signal from fluorescence emission during light excitation when slides are scanned (Alonso-Betanzos, 2019). The sample's gene expression profile is identified from intensities produced relative to the degree of hybridization detected on the chip (Alonso-Betanzos, 2019). The output is an image incorporating signal intensities that can be further analysed and gene expression profiles can then be compared under different conditions (Levant, 2005). While both microarray and RNA-seq are used to measure gene expression levels, they are totally different techniques. Depending on the research question, experimental design, and the budget. In general, RNA-seq is considered superior in term of speed and versatility. Still microarray is cheaper per sample and have well-established/characterised workflow, and has a wide use in thee field of research.(Weißbach *et al.*, 2021),(Levant, 2005).

1.6 Hypotheses

- LncRNAs play a significant role in the development of Tamoxifen resistance in breast cancer.
- Altered expression of candidate lncRNAs can change breast cancer cell response to tamoxifen.

- LncRNAs are potential biomarkers, prognostic factors, and therapeutic targets in endocrine resistance breast cancer.

1.7 Project Aims

1. To use the previously generated RNA-seq data from tamoxifen resistant and non-resistant breast cancer cell lines to identify candidate lncRNAs for subsequent investigation.
2. To use different bioinformatics tools and publicly available breast cancer microarray and RNA-seq data sets to identify lncRNAs involved in tamoxifen resistance.
3. To alter these candidate lncRNA expression levels in breast cancer cell lines and investigate response to tamoxifen.
4. To identify any other functional consequences of alteration of candidate lncRNAs in the context of breast cancer.

2. Materials and methods

2.1 Materials

2.1.1 Water

Double-distilled purified water (ddH_2O) was used. Laboratory grade ultra-pure ddH_2O was produced by triple tree water technology that uses chemical-free electro-deionisation system.

2.1.2 Sterilisation

Glass vessels and solutions were sterilised by autoclaving at 120 °C under pressure of 15 p.s.i for 15 minutes. Solutions that could not be autoclaved were filter sterilised using a sterile syringe filter to remove fine particles and microorganisms.

2.1.3 Drugs

4-Hydroxy Tamoxifen

4-OH tamoxifen supplied by Santa Cruz Biotechnologies was diluted in ethanol or DMSO to reach a stock concentration of 10 mM, aliquoted and stored at -20°C.

All-trans retinoic acid (ATRA)

ATRA stock was purchased from ACROS organics and was diluted in DMSO (sigma) to reach a stock concentration of 10mM and stored at -20° C. all preparations and handlings of ATRA were done under subdued lighting.

2.1.4 PCR probes:

For each gene, when there are more than one transcript, RefSeq database were used to select the one with the most reliable sequence information.

Gene Target	Assay ID	Technology	Supplier
HORAIRM1	LPH10483A	SYBR® Green	Qiagen
HOXA1	PPH01464B	SYBR® Green	Qiagen
HOXA10	LPH41929A	SYBR® Green	Qiagen
HOXA5	LPH31164A	SYBR® Green	Qiagen
HOXA9	LPH29723A	SYBR® Green	Qiagen
LUCAT1	LPH16113A	SYBR® Green	Qiagen
NR2F1-AS1	LPH12924A	SYBR® Green	Qiagen
SOX21-AS1	LPH06753A	SYBR® Green	Qiagen
β-actin	PPH00073G	SYBR® Green	Qiagen

Table 2.1 PCR probes

2.1.5 Antibodies

2.1.5.1 Primary antibodies:

Antibody Name	Host Animal	Manufacturer (Catalogue Number)	Application (Dilution)
Phosphorylated H2A histone family member X (γH2AX) S139	Mouse	Millipore (JBW301)	IF (1:500)
γH2AX S139	Rabbit	Cell Signalling Technology (2577)	IF (1:500)
Epithelial cadherin (E cadherin)	Rabbit	Cell Signalling (24E10)	IF (1:200)
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)	Mouse	Proteintech (60004-1-Ig)	WB (1:20,000)
Vinculin	Mouse	R & D Systems (MAB68961)	IF (10 µg/mL)
Yes kinase-associated protein (YAP)	Rabbit	Santa Cruz (sc-154-07)	IF (1:100)
β-tubulin	Mouse	Sigma-Aldrich (T8328)	WB (1:10000)
β-catenin	Mouse	Santa Cruz Biotechnology (sc-7963)	IF (1:200)

Table 2.2 primary antibodies.

IF: Immunofluorescence, WB: Western blot

2.1.5.2 Secondary antibodies

Antibody Name	Host Animal	Manufacturer (Catalogue Number)	Application (Dilution)
Anti-mouse Alexa 488	Goat	Thermo Fischer Scientific (A11017)	IF (1:1000)
Anti-mouse Alexa 594	Goat	Thermo Fischer Scientific (A11005)	IF (1:500)
Anti-rabbit Alexa 488	Donkey	Life Technologies (A21206)	IF (1:500)
Anti-rabbit Alexa 594	Goat	Thermo Fischer Scientific (A11037)	IF (1:1000)
Anti-mouse IgG horse radish peroxidase (HRP)	Horse	Cell Signalling Technology (7076)	WB (1:2000)
Anti-rabbit IgG HRP	Goat	Cell Signalling Technology (7074)	WB (1:2000)

Table 2.3 Secondary antibodies.

IF: Immunofluorescence, WB: Western blot

2.1.6 Cell lines:

Name	Supplier	Genotype	Receptor Expression
CAL-51	DSMZ	Human mammary gland adenocarcinoma, derived from metastatic site: pleural effusion	ER - PR - ERBB2-
MCF-7	ATCC	Human mammary gland adenocarcinoma, derived from metastatic site: pleural effusion	ER+ PR+ ERBB2-
TAMR	Cardiff University	Human mammary gland adenocarcinoma, developed by exposing MCF-7 cells to tamoxifen continuously for an extended period of time.	ER+ PR+ ERBB2-
MDA-MB-231	ATCC	Human mammary gland adenocarcinoma, derived from metastatic site: pleural effusion	ER - PR - ERBB2-
MDA-MB-468	ATCC	Human mammary gland adenocarcinoma, derived from metastatic site: pleural effusion	ER - PR - ERBB2-
ZR-75-1	ATCC	Human mammary gland ductal carcinoma, derived from metastatic site: ascites	ER+ PR - ERBB2-
T47D	ATCC	Human mammary gland ductal carcinoma, derived from metastatic site: pleural effusion	ER + PR+ ERBB2-

Table 2.4 Cell lines. ATCC: American Type Culture Collection, DSMZ: Deutsche Sammlung von Mikroorganismen und Zellkulturen

Tamoxifen resistant MCF-7 cell line.

TAMR (tamoxifen resistant MCF-7) cell lines were kindly provided by Dr Julia Gee, Cardiff University. Briefly, TAMR cell line was created by culturing MCF-7 cell line in RPMI 1640 phenol-red free media supplemented with 4-OH Tamoxifen at a concentration of 1uM for several months (> 6 months) (Knowlden *et al.*, 2003).

2.1.7 Cell culture medium

MCF-7 cells: Roswell Park Memorial Institute (RPMI)-1640 media (Lenzo) supplied with 5% FCS (foetal calf serum) (Gibco), 1% PenStrep (Gibco) and 1% amphotericin B (Gibco) or RPMI phenol red free media supplied with 5% charcoal stripped FCS, 2% L-glutamax, 1% PenStrep and 1% amphotericin B, when challenged with tamoxifen.

TAMR cells: RPMI-1640 phenol red free media (Gibco) supplied with 5% charcoal stripped FCS (sigma), 2% L-glutamax (Gibco), 1% PenStrep (Gibco) and 1% amphotericin B (Gibco), 4 OH-tamoxifen was added fresh to cell culture flasks at a concentration of 1 µmol.

CAL-51, MDA-MB-231 and ZR-75-1 cell lines: High glucose Dulbecco's Modified Eagle Medium (DMEM) containing L-glutamine and 1% of 1x non-essential amino acids (NEAA) (Sigma Aldrich).

MDA-MB-468 and T-47D cell lines: RPMI-1640 medium containing L-glutamine and 1% of 1x non-essential amino acids (NEAA) (Sigma Aldrich).

All prepared media were stored at 5 ºC in tissue culture lab and used under sterile conditions.

2.1.8. Trypsin and versene/EDTA

To release the adherent cells from culture vessel surfaces, Trypsin EDTA with 0.5 g/L Trypsin and 0.2 g/L versene (EDTA) was supplied by Lonza. Stock was stored at -20°C and thawed when needed.

2.1.9 Short interfering RNAs (siRNAs)

siRNA name	ID	siRNA sequence	Company
Control si-RNA	SiGENOME	Non-targeting siRNA pool	Dharmacon
si-HOTAIRM1	HOTAIRM1 siRNA#1	Sense: ACUGGUAGCUUAUUAAAGAtt Antisense: UCUUUAAUAAGCUACCAGUct	Ambion®
si-HOTAIRM1	HOTAIRM1 siRNA#2	Sense: GAAAUGUGGGUGUUUGAAAtt Antisense: UUUCAAACACCCACAUUUCaa	Ambion®
si-HOTAIRM1	HOTAIRM1 siRNA#3	Sense: ACUUAGUUAUUGACCUCCatt Antisense: UCCAGGUCAAUAACUAAGUta	Ambion®
Si-LUCAT1	LUCAT1 siRNA#	Sense CCCAUCAGAAGAUGUCAGAAGAUAA Antisense UUAUCUUCUGACAUCUUCUGAUGGG	Eurofin
Si-LUCAT1	LUCAT1 siRNA#2	Sense CAAGCUCUUGCAGUCAACAAGAACU Antisense AGUUCUUGUUGACUGCAAGAGCUUG	Eurofin
Si-SOX21-AS1	SOX21-AS1 siRNA#1	Sense AACAGAAACAGAGGCUUCUCGCAUU Antisense AAUGCGAGAAGCCUCUGUUUCUGUU	Eurofin
Si-SOX21-AS1	SOX21-AS1 siRNA#2	Sense CAGUUAACUUACAGUGUCUCACUUA Antisense UAAGUGAGACACUGUAAGUUAACUG	Eurofin
Si-NR2F1-AS1	NR2F1-AS1 siRNA#1	Sense ACCACAAUAUUAACCAGGAtt Antisense UCCUGGUUAAUAUUGUGGUca	Ambion®
Si-NR2F1-AS1	NR2F1-AS1 siRNA#2	Sense GAAUUGGCUAGAUCAGGAAtt Antisense UUCCUGAUCUAGCCAAUUCta	Ambion®

Table 2.5 Short interfering RNAs

2.1.10. Matrigel matrix

ECM basement membrane matrix gel from Engelbreth-Holm-Swarm Murine Sarcoma was supplied by Corning.

2.1.11. Buffers

Phosphate buffered saline (PBS)

PBS was produced by dissolving 1 Oxoid PBS tablet per every 100 ml of ddH_2O to prepare balanced salt solution without calcium and magnesium. PBS was sterilised by autoclave and stored at room temperature.

Tris-buffered saline (TBS)

10x TBS solution was made by dissolving 24.2 g Tris Base (200 mM) with 80 g NaCl (1.4 M) in 800 ml ddH_2O. 5 M HCl was then used to acidify solution pH to 7.6 before increasing the volume to 1 L by adding ddH_2O. 10x solution was kept at room temperature. 1x solution was achieved by diluting a part of 10x solution with 9 parts of ddH2O.

1 M Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCL) pH 6.8 and 8.0

To prepare 1M Tris-HCL solution, 121.1 g of Tris base (C4H11NO3, CAS Number: 77-86-1) was dissolved in 800 ml ddH_2O using magnetic stirrers. Then, 5 M HCl was added to adjust solution pH to 6.8 or 8.0. finally, volume was adjusted to 1 L with ddH_2O. 1 M Tris pH 6.8 and pH 8.0 were stored at room temperature for several weeks.

1.5 M Tris-HCL pH 8.8

1.5 M Tris solution was made by mixing 181.7 g of Tris base in 800 ml ddH2O. Then, 5 M HCl was used to adjust solution pH to 8.8, then, final volume was topped

up to 1 L with ddH_2O. this Tris solution was stored at room temperature for several weeks.

10 mM Tris-HCL 1 mM EDTA solution pH 9.0

To produce 10 mM Tris-HCL 1 mM EDTA solution, 1.21 g of Tris base and 0.37 g EDTA were dissolved in 800 ml ddH2O. 5 M HCl was then used to modulate solution pH to 9.0. 0.5 ml Tween 20 (0.05%) was then added to the solution before bringing up volume to 1 L with ddH_2O. 10 mM Tris-HCL 1 mM EDTA solution pH 9.0 was stored at 4°C.

10% SDS

10% SDS solution was prepared by diluting 500 mL 20% SDS solution with 500 mL ddH2O. The solution was then stored at room temperature. In case of the presence of precipitation, 10%SDS was heated to 60 °C until the precipitated SDS become homogenised again.

5 x Radio-immunoprecipitation assay (RIPA) lysis buffer

100 ml 5 x RIPA lysis buffer was produced by adding 25 ml 1 M Tris-HCL pH 8.0 (250mM), 15 ml 5 M NaCl (750mM), 5 ml 10% SDS (0.5%), 5 ml NP-40 (5%), 2.5 g Sodium deoxycholate (2.5%) and ddH_2O to total volume to 100 ml. 5 x RIPA lysis buffer was autoclaved and stored at room temperature for several weeks.

SDS-polyacrylamide gel electrophoresis (SDS-PAGE) running buffer

10 x SDS-PAGE running buffer was produced by mixing 30 g Tris-HCL base (250 mM) with 144 g Glycine (1.9 M) in 900 mL ddH_2O. Then 100 ml of 10% SDS (1%) was added to bring volume up to 1 L. SDS-PAGE running buffer was then stored at room temperature for several weeks.

5 x SDS sample buffer

5 X SDS sample buffer was produced by mixing 25 ml 1 M Tris pH 6.8 (250mM), 10 g SDS (10%), 50 ml glycerol (50%), 5 ml β-mercaptoethanol (5%), 20 mg bromophenol blue (0.02%) and ddH_2O to bring volume to 100 ml. 5 x SDS sample buffer was stored at room temperature for several weeks.

1 x SDS-PAGE running buffer

1 x SDS-PAGE running buffer was produced by diluting 100 ml 10 x SDS-PAGE running buffer with 900 ml ddH_2O.

10 x Towbin transfer buffer

10 x Towbin transfer buffer was made by dissolving 30.3 g Tris base (250 mM) and 144 g Glycine (1.9 M) in 800 ddH_2O, then, solution was brought to a final volume of 1 L with ddH_2O.

1 x Towbin transfer buffer

1 x Towbin transfer buffer was made by mixing by order 100 ml of 10 x Towbin buffer, 700 ml of ddH_2O and 200 ml methanol. The buffer was chilled to 4◦C before using.

Crystal violet stain

To stain cells in static adhesion assay, 0.1% crystal violet with 20% methanol was prepared. First, 20% methanol was prepared by mixing 10 mL 100% methanol with 40 mL ddH_2O. Then, 0.05 g of crystal violet powder was added and mixed well by magnetic stirrer. Solution then was passed through a sterile syringe filter to remove fine particles and was stored at room temperature.

2.2 Methods

2.2.1 Cell passaging

All cell lines were kept at 37 °C, 5% CO2 growth environment in cell culture incubator from Sanyo Electric Co. Growth was evaluated using light microscope, once cells reached 80-90% confluency they were ready for passaging. Media were removed, and cells were washed with 10 ml PBS. Cells were then removed from the flasks by trypsinisation, this step included the addition of 1ml of trypsin/EDTA solution followed by incubation for 5 minutes. When the cells started to detach, 9mL of the appropriate maintenance media was added and the solution was pipetted up and down several times to form a single cell suspension. According to the desired subculture ratio, usually ranging from 1:3 to 1:6, cell suspension was aliquoted into fresh T75 cell culture flasks and resuspended in media to reach a total volume of 10mL. Passage number was considered in all experiments, biological repeats of the same experiment were done within 10 passages to avoid biological variation.

2.2.2 Cells long term storage

Performed on cells of as low passage number as possible. Once the cells in the culture flask reach 80-90% confluence, standard trypsinization protocol was followed, the cells were then re-suspended in 10 ml of media and transferred to sterile 15 ml tube to be centrifuged in 2000 rmp for 3 minutes after that the supernatants were removed. 1 ml of medial and 100 µl of DMSO were added slowly to the cells and the solution was transferred into 1.8 ml CryoPure tube and kept at -80 °C to be stored afterward in liquid nitrogen tank for long term reservation.

2.2.3 Transfection of cells with siRNA

SiRNA was transfected into cells using cationic lipid-based transfection reagent. DharmFECT 1 was purchased from Dharmacon or Lipofectamine™ 2000 transfection Reagent was purchased from Invitrogen™ as transfection reagents. For better siRNA stability during storage, 100 µM stock for all siRNAs were prepared by briefly centrifuging the provided 5 nmole powder to ensure all the dried RNA oligonucleotides are collected in the bottom of the tube. Then, it was mixed with 50µL of RNase-free 1x siRNA buffer or provided RNase free water. The solution then aliquoted and kept at -20◦C to reduce the freezing-thawing cycles when preparing the working solution. To prepare working solution, 25µL of stock solution was diluted in into RNase-free 1x siRNA buffer to reach 5 µM and aliquoted into Eppendorf tubes to reduce the number freezing-thawing cycles and stored at -20°C.

2.2.4 Cell pellet preparation

When the cells reached 60-70 % confluence, they were washed with 10 ml PBS, trypsinised and 10 ml of media was added and transferred to 15 ml Falcon tubes and centrifuged for 3 minutes at 1000 rpm, supernatants were then discarded. Cells then were resuspended in 10 ml of PBS and centrifuged again, supernatants removed, and cells resuspended in 1 ml of PBS and transferred to 1.5ml Eppendorf tube, centrifuged in cooled microcentrifuge at 3000 rpm for 5 minutes, supernatants were then removed, and cell pellets kept at -70◦C.

2.2.5 RNA extraction

Total RNA was extracted from cell pellets using RNeasy Mini Kit (Qiagen Cat No.: 74104) according to manufacturer instructions. For each cell pellet, the first step was cell lysis and homogenization, this involved the following: RLT buffer was prepared by mixing 10µl of β-ME with every 1ml of RLT and 600 µl of the mixture were added to every sample, then cell lysate was disrupted and homogenized by passing through 20-gauge needle attached to a sterile syringe 10 times followed by the addition of 600µl of 70% Ethanol. Next step was RNA purification through centrifugation to remove contaminants such as genomic DNA and small RNAs leaving only RNA species >200 bases attached to the spin column membrane. Here, 700 µl of the sample were transferred to RNA spin column and centrifuged for 15 seconds at ≥10,000 rpm then flow-through were discarded, this step was repeated for the remainder of the sample. This is followed by three washes by adding 700 µl of buffer RW1 to RNase spin column for the first wash then 500 µl of buffer RPE for each of the second and third washes with centrifuging the spin column for 15 s at ≥10,000 rpm and discarding flow-through from the collection tube after each wash. The last step is purified RNA elution by adding 50 µl of RNase free water to the spin column placed in 1.5 ml collection tube and centrifuging the sample for 1 minute at ≥10,000 rpm. To achieve a higher RNA concentration, the eluate produced were re-loaded into the spin column and centrifuged at ≥10,000 rpm for 1 minute. RNA concentration for each sample was measured by NanoDrop spectrometer ND-1000 by loading 1 µl of sample after water blank then samples were stored at -70°C.

2.2.6 Quantification of RNA concentration

The concentration and quality of the extracted RNA were assessed using NanoDrop spectrometer ND-1000, After thoroughly cleaning the NanoDrop pedestals, 1 µL of RNase free water was used to blank the instrument prior to initiation. Then the proper setting for measuring RNA concentration on the available software were selected. 1 µL from each sample was loaded and RNA concentration for each in ng/µL were recorded. Also, the ratio of absorbance at 260/280 nm and 260/230 nm were noted, and ~2.0 were indicative of accepted purity of the sample. In-between each measurement, the lower and upper pedestals were wiped with soft laboratory wipes and re-blanked with RNase free water.

2.2.7 cDNA synthesis

For cDNA synthesis RT2 first strand kit (Qiagen Cat No./ID: 330404) was used. Reagents and RNA samples were thawed on ice then briefly centrifuged to get the contents at the bottom of the tubes. First, for each RNA sample, genomic DNA elimination mix was prepared by mixing 2 µg of RNA and 2 µl of Buffer GE and enough amount of Nuclease-free water to reach a final volume of 10 µl in 0.5 ml Eppendorf tubes, samples then were centrifuged briefly and incubated in heating block at 42℃ for 5 minutes then on ice while preparing the Reverse transcription (RT) mix. For RT mix enough for 12 samples, in one 1.5 ml tube, 48 µl 5x buffer BC3, 12 µl Control P2, 24 µl RE3 Reverse Transcriptase mix and 36 µl Nuclease-free water were mixed by pipetting and 10 µl of RT mix were mixed in with genomic DNA elimination mix. Samples then were incubated for 60 min at 37℃ directly followed by 5 minutes at 95℃ to stop the reaction. Finally, 91 µl of Nuclease-free water were added and the samples finally were stored at -20°C.

2.2.8 Real-time PCR

RT² SYBR Green ROX qPCR Mastermix (Qiagen, Cat No./ID: 330523) were used according to manufacturer instruction to detect HOTAIRM1 expression. ACTB (B-actin) was used as an internal control, as ACTB expression does not change under experimental conditions. 1:10 dilution of HOTAIRM1, ACTB in RNase free water were prepared. For example, volume for 14 reactions (12 reactions (3 technical replicate) + 10% pipetting errors) of PCR components were mixed in RNase free 0.5 ml tube for each sample, this consists of: 8.4 µl Nuclease free water, 70 µl SYBR Green Mastermix and 5.6 µl sample cDNA. 6 µl of sample's PCR mix and 4 µl of diluted primers were loaded in each well of 384 well-. Plate then was centrifuged for 2 minutes at 1000 rpm in PCR plates centrifuge to remove bubbles and was sealed with optical film to avoid evaporation of samples in the PCR machine.

2.2.9 Drug treatment of cells

ATRA treatment

It has been shown that HOTAIRM1 expression can be induced by exposing NB4 cells to ATRA (X Zhang et al., 2009). Following the same experimental conditions, three biological replicates of MCF-7 cells were cultured in 10mL of experimental media with 1µL of 10mM ATRA to reach a final concentration of 1µM in 75cm2 cell culture flask. Cells were divided into 4 groups according to treatment period :48 hours, 24 hours, and 2 hours and untreated (control) group. Cell pellets were harvested and kept at -70℃ for RNA extraction, cDNA synthesis and HOTAIRM1 expression analysis using RT-qPCR.

.

Tamoxifen treatment

Cells were passaged once in experimental media prior to addition of increasing doses of 4-hydroxy tamoxifen or ethanol control. Cells were divided into 4 groups according to treatment period :48 hours, 24 hours, and 2 hours and untreated (control) group. Cell pellets were harvested and kept at -70℃ for RNA extraction, cDNA synthesis and HOTAIRM1 expression analysis using RT-qPCR.

2.2.10 Candidate lncRNAs depletion from cells

To optimise protocol for transfection of siRNA targeting lncRNAs we started with three different siRNAs for HOTAIRM1 (section 2.1.13) and the Bryant lab optimized protein coding gene siRNA transfection protocol for MCF-7 cell line was followed. TAMR cells were seeded in 6-well plate at cellular density of 5×10^5 cell/well in 2 ml of maintenance media. Once reached 60-70% confluency, transfection complexes were prepared by diluting 10 µl of each siRNA (for HOTAIRM1 and scrambled control) in 190 µl of SAFM (serum-free and antibiotic-free media) to reach a total volume of 200 µl for each well. In another tube, 4 µl of DharmaFECT 1 were diluted in 196 µl of SAFM to reach a total volume of 200 µl for each well, this mixture was then incubated for 5 minutes at room temperature then 200 µl of diluted DharmaFECT were mixed gently with each diluted siRNA tube and this were incubated for 20 minutes at room temperature. Cells were washed twice with 2 ml of PBS and filled with 1600 µl of antibiotic-free media and 400 µl of transfection complex were added drop by drop to each labelled well and incubated for 48 hours. Cellular toxicity was evaluated visually under light microscope, by observing changes in cell morphology, cell density, and cell confluency.

This transfection protocol was further optimized to reach the optimum cell density with least toxicity. Using 6-well plates, three different cell densities of TAMR cells (300,000/ 500,000/ 700,000 cells per well) and three different DharmaFECT concentrations (4/6/8 µl per well) were used following the same protocol mentioned above. For lncRNAs knockdown in 96-well plates, the number of cells were scaled down and the procedure were repeated according to the manufacturer's instructions.

2.2.11 MTT cell viability assay

The main purpose of MTT assay is to monitor the effect of tamoxifen on growth and proliferation of different cell lines under different experimental conditions. Cells were allowed to reach 80-90% confluence, then they were detached by the addition of 1 ml of trypsin, re-suspended in steroid-free media and counted by haemocytometer. Cells were plated as 6000 in 96 well-plates (200µl/well) by an 8-chanel pipet. Four hours post seeding the cultivated cells were treated with different concentrations of 4-hydroxy tamoxifen (i.e., 10µM, 1µM, 0.1µM, 0.01µM), and 10 wells were treated with 1µL DMSO as vehicle control. MTT assay was performed four days later, when the control group reached 70-80 % confluence. MTT solution was prepared at 3 mg MTT reagent per 1 ml of PBS, 50µl of the MTT solution were added to each well, plates were placed in the incubator for four hours. Media were removed and replaced by 100µl of DMSO the plates were placed in plate reader to calculate the optical density at a wavelength of 560nm and subtract background at 670nm.

2.2.12 Immunofluorescent staining.

Slide preparation and fixation

22 × 22mm Coverslips were dipped in ethanol, allowed to dry, and lay flat onto 6 well plate, 400,000 cells were added directly on the coverslips. Plates were kept at 37°C and 5%CO_2 overnight to adhere to coverslips before knockdown. Optimised HOTAIRM1 depletion protocol was followed. Now cells were ready for fixation.

Fixation step was performing by media removal from the treated cells then washed with 2 ml PBS for 5 minutes. Cells then were fixed with 4% paraformaldehyde for 10 minutes; this step was followed by 5 minutes TBS wash twice. Cells were permeabilised with 1 mL of 0.2% v/v Triton-X-100 in PBS for 10 minutes then washed with TBS 3 times each for 5 minutes. Next, cells were blocked for 1 hour with 200 µL 3% w/v BSA in PBS as blocking buffer. Finally, cells were washed twice with 2 mL TBS each for 10 minutes. If the slides were not to be stained immediately, 1 mL of TBS was added to each well to prevent dryness, plates covered and wrapped in plastic wrap and kept at 4 °C up to 1 month ready for staining and imaging. All the washings were done gently by adding buffers to the walls of wells and all done at room temperature on the orbital shaker.

Staining method

TBS was removed from fixed slides and coverslips were inverted onto 100 µl 2% w/v BSA (E-cadherin and β-catenin) or 1% w/v BSA (γH2AX) in TBS containing the appropriate concentration of primary antibodies for 1 hr at room temperature (E-cadherin and β-catenin) or overnight at 4^{o} C (H2AX) in a humidified chamber. Coverslips then were inverted back into 6 well plates (the side with cells attached facing up) and washed 3 times in 2 mL TBS for 10 minutes. Coverslides were then

inverted (the side with cells attached facing down) on 100 µL of Alexa Fluor® 594–conjugated goat anti-mouse IgG secondary antibody and Alexa Fluor® 488–conjugated donkey anti rabbit IgG secondary antibody diluted in 1% w/v BSA in TBS with 1 µg/mL DAPI and kept for 1 hour at room temperature in a humidified chamber protected from light. Coverslides were inverted onto 6 well dish then washed 2 times with 2 mL TBS. Finally, each Coverside were inverted on a drop of Shandon™ Immu-Mount on labelled glass microscope slide. Mounted coverslips were allowed at least 2 hours to dry in the dark at room temperature before being stored at 4°C protected from light prior to imaging.

Analysis.

Images of 100 cells per condition were acquired on a Nikon TE200 inverted fluorescent microscope using a 60x/1.4 oil immersion objective lens with Volocity software. Fluorescent channels of corresponding images were merged using FIJI and the signal intensity of E-cadherin and β-catenin assessed for each condition. Alternatively, the nuclear γH2AX foci intensity per nuclei determined using FIJI software for each condition.

2.2.13 Fluorescence-activated cell sorting (FACS)

Cell harvesting

Optimised HOTAIRM1 depletion protocol was followed. Cells were harvested for FACS analysis 48 hours post transfection. Cells were gently washed with 2 mL PBS and dislodged with 0.5 mL trypsin EDTA. Cells were collected in 5 mL of the appropriate media in a 15 mL falcon tube. Cells were pelleted at 1200 RPM for three minutes then media was poured off, tubes were kept on ice all the time. Next, cells were resuspended in 5 mL PBS, pelleted at 1200 RPM for 3 minutes, PBS

was poured off. Cells were resuspended in 1 ml of PBS and centrifuged at 2000 RPM for 5 minutes at 4°C, PBS was discarded completely. Next, the cell pellet was resuspended in 1 ml ice cold 100% methanol gradually with frequent vortexing. Cells were stored in 100% methanol for overnight minimum up to a month at -20°C.

Propidium iodide (PI) and S10 p-Histone 3 co-staining

Cells were retrieved from -20°C and pelleted at 2000 RPM for 5 minutes at 4°C before methanol was gently poured off and cells were washed twice in PBS on ice. Then, were resuspended in 1 mL PBS. Cells were incubated with anti-pH3 antibody for 1 h at RT then after washing 3 x in PBS PI solution was added and cells ran though the FACS machine.

Analysis

First step was doublets exclusion, by forward scatter height vs area and gating FL3-Width low cells. 10,000 events were collected in this gating region per sample and a FL3-height distribution histogram was produced. The G1 proportion of cells was defined by gating across the base of the first peak in FL3-H plot (~200 FL3-H), the G2 proportion of cells was defined by gating across the base of the second peak in FL3-H plot (~400 FL3-H). S phase population was defined as all signal between first and second peaks.

2.2.14 Cell migration

Optimised HOTAIRM1 depletion protocol was followed. Cells were harvested for cell migration assessment 48 hours post-transfection. Cells were gently washed with 2 mL PBS and dislodged with 0.5 mL trypsin EDTA, resuspended in media, and counted using a hemocytometer. In-order to achieve the appropriate cell density, different seeding densities of control siRNA cells were used (5×10^4, $4 \times$

10^4 and 3×10^4) cells per well. The appropriate cell density was selected when the cells reached a confluent monolayer after overnight incubation. For gap creation, cells were seeded at an optimised density of 3×10^4 cells per well in culture-Insert 2-Well. After cells were allowed to attach overnight, culture-insert was removed, and cells were gently washed with 2 mL of warm (37◦C) PBS twice to remove any debris or non-adherent cells. Then, 2 mL 0.5% FCS supplied culture media was provided; to exclude the effect of proliferation and ensure only migration is measured. Baseline image (0 H) and subsequent gap closure images were captured at 24-, 48- and 72 hours using Nikon TE200 inverted fluorescent microscope at 10× magnification.

Analysis

The gap closure was measured over time using ImageJ software. To calculate the remaining clear area after each time point compared to the control time of zero hour. The gap areas were measured, and the percentage of the closed area was calculated. These steps were repeated for each condition. The percentage of gap closure was calculated using the following equation:

$$\text{Gap remaining (\%)} = \frac{(\text{area of the initial gap} - \text{area of the remaining gap at } 0,24,48 \text{ or } 72\text{h}) \times 100}{\text{width of initial gap}}$$

The experiments were repeated three times for each experimental condition.

2.2.15 Cell adhesion

To prepare culture plates, EMC Corning Matrigel mix was thawed overnight at 4◦C, 6 well-plates and pipets were precooled at 4◦C before use. EMC gel was diluted as 1:2 in cold media cell culture medium, then 35 µL of EMC mix was dispensed in each well as 5 technical repeats per condition, 1 plate was made per time point and

left in incubator for 30 mins to set, then plates were kept at 4°C for up to two weeks. The product will gel within 5 minutes at 20°C. For prolonged manipulations, work should be conducted below 10°C. Cells can be plated on top of a thin gel layer of 0.5 mm, 96 well-plates were coated with stored at 4° C for up to two weeks. Optimised HOTAIRM1 depletion protocol was followed. Transfected cells were harvested for cell adhesion assay 48 hours post transfection. Cells were gently washed with 2 mL PBS and dislodged with 0.5 mL trypsin EDTA, resuspended in media and counted using hemocytometer. 200 µL of optimised number of transfected cells' solution (25,000 cells per well) was seeded in Matrigel coated wells. Cells were incubated at 37°c and 5% CO_2 for 1 hour. Then, gently the media was discarded, and plates washed three times gently with 200 µL PBS without touching gel layer, to remove non-adherent cells. Attached cells then were fixed with 100 µL 4%paraformaldehyde for 15 minutes, after the time passed paraformaldehyde was removed gently and cells were washed twice with 200 µL PBS. To stain cells, 100 µL of 0.1% crystal violet with 20% methanol were added for 15 minutes. Subsequently, stain was removed, and cells were washed twice with 200 µL PBS to remove excess crystal violet. Gels/with cells attached were solubilised with 100 µL 1% SDS added carefully, to avoid any bubbles, for 30 minutes on the belly dancer. Absorbance was recorded at 590 nm on plate reader.

2.2.16 Western blotting.

Lysate production.

Optimised HOTAIRM1 depletion protocol was followed. About 1 million cells were harvested for protein extraction 48 hours post transfection (3 culture wells of 6 well-plates per condition). Transfected cells were gently washed with 2 mL PBS and

dislodged with 0.5 mL trypsin EDTA. Cells were collected in 5 mL of appropriate media in 15 mL falcon tube. Cells were pelleted at 1200 RPM for three minutes then media was poured off, tubes were kept on ice all the time. Next, cells were resuspended in 5 mL PBS, pelleted at 1200 RPM for 3 minutes, PBS was poured off. Cells were resuspended in 1 ml of PBS and centrifuged at 2000 RPM for 5 minutes at 4°C, PBS was discarded completely. 1 X lysis buffer was prepared by mixing 700 µL ddH2O, 200 µL 5x RIPA lysis buffer, 10 µl 100 mM PMSF, 10 µl 100x protease inhibitor, 10 µL 100x phosphatase inhibitor and 2 µl Benzonase per 1 mL. Then, the pellet was resuspended in one pellet volume of 1x lysis buffer. The resuspended pellet was incubated on ice and periodically vortexed every 10 minutes for 30 minutes, then homogenisation step followed by passing cell lysate through a 25G needle 10-15 times. This solution was then centrifuged for 10 minutes at 13,400 RPM at 4 °C to separate protein lysates. The supernatant was transferred into a new 1.5 mL Eppendorf tube before being stored at -20 °C until required.

Protein quantification.

To accurately determine protein concentration from the lysate, a standard curve was produced using commercially available protein standards: bovine serum albumin (BSA) concentrations (Bradford, 1976). One microlitre of each lysate was diluted in 799 µl of ddH_2O followed by addition of 200 µl of Bio-Rad protein assay dye reagent concentrate in 1.5 mL Eppendorf tubes (see Table). 200 µL of each solution moved to a well in 96 well plate. The optical density (OD) was measured using a Multiskan™ FC microplate photometer, absorbance was measured at 595 nm. ODs of BSA standards were plotted against their known concentrations to

produce a standard curve. The standard curve line from this allowed calculation of total protein in lysates from their OD values. The amount of protein was calculated to be loaded per lane for western blot.

Total Protein (µg)	0.1mg/ml BSA (µl)	ddH_2O (µl)	Biorad Protein Assay Dye Reagent Concentrate (µl)
0	0	800	200
1	10	790	200
5	50	750	200
10	100	700	200
15	150	650	200
20	200	600	200

Table 2.6 Volumes needed for producing BSA standard curve.

SDS-PAGE.

Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) is the method used to separate proteins based on their mass (molecular weights). The highest of interest protein molecular weight was E-cadherin (135 kDa), while the lowest is GAPDH (37 kDa). Referring to Table 2.2 it was decided to make 8% percent polyacrylamide resolving gel as detailed in Table x. Volumes were scaled up and down as required to produce multiple gels. When making the gels, TEMED and APS were added immediately before pouring the gels in moulds. The resolving gel was allowed to set completely. Then 5 mL of stacking gel was poured on top and a comb of desired well number and size placed into it. The wells were washed with water following setting to ensure that they were clear of any debris. Lysates

achieved by RIPA lysis were loaded in equal amounts of 30-60 µg per lane with 1x SDS sample buffer. In addition, 5 µL of Precision Plus protein standard was loaded to one lane to run parallel to the samples so it can be used as a marker of molecular weights and samples' proteins can be identified relatively. More 1x SDS-PAGE running buffer were added carefully and power were run at 180 V for 45 minutes to 1 hr 30 minutes until the proteins of interested were migrated and sufficiently separated.

Protein molecular weight (kDa)	**Resolving gel percentage (%)**
4-40	20
12-45	15
10-70	12.5
15-100	10
25-200	8

Table 2.7. The appropriate resolving gel percentages for certain ranges of protein sizes.

Solution	8% Resolving gel (10 mL)	Solution	5% Stacking g mL)
ddH_2O (ml)	4.6	ddH_2O (ml)	3.4
30% Acrylamide (ml)	2.7	30% Acrylamide (ml)	0.83
1.5 M Tris pH 8.8 (ml)	2.5	1 M Tris pH 6.8 (ml)	0.63
10% SDS	0.1	10% SDS	0.05
10% APS (ml)	0.1	10% APS (ml)	0.05
TEMED (ml)	0.006	TEMED (ml)	0.005

Table 2.8. Volumes of solutions for preparing resolving and stacking gels.

Protein transfer.

Proteins fractionated by SDS-PAGE were then electrophoretically transferred (Figure 2.1) from the gels onto 0.45 µm Protran nitrocellulose transfer membrane (GE Healthcare) using a Criterion blotter and 1x pre-chilled Towbin transfer buffer on ice. A current of 85 V was used for 2 hours.

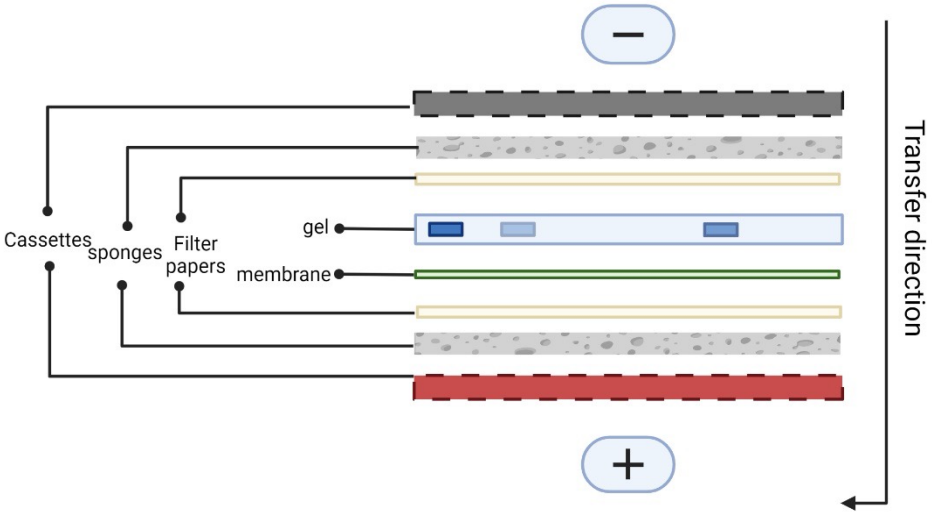


Figure 2.1. assembly Illustration of transfer sandwich setup. The sandwich cassette layers as observed from the top for the western blotting experiment. It is advisable to build on the black side of the sandwich holder, which will go towards the black side of the electrode box.

Antibody detection

To reduced background noise and increase specificity of primary antibodies, following protein transfer, membranes were blocked with a neutral protein (5% w/v skimmed milk powder in TBS for 1 hr at room temperature). To identify target proteins, membranes were then probed with primary antibody diluted in 5% w/v skimmed milk powder in TBS at 4 °C for 16-24 hrs, then washed three times with 0.05% v/v Tween-20 in TBS (TBS-T) every 10 minutes. Membranes were then incubated with the appropriate HRP-labelled secondary antibody that act as a reporter for targeted proteins, secondary antibodies were diluted in 5% w/v skimmed milk powder in TBS for 1 hr at room temperature. Following secondary antibody incubation membranes were washed three times with TBS-T every 10 minutes prior to chemiluminescent detection.

Enhanced chemiluminescence (ECL).

Equal volumes of reagent 1 and reagent 2 of the Amersham ECL Western blotting detection reagent kit were mixed in a falcon tube and then incubated with the membrane for 1 minute at room temperature with gentle agitation to ensure full coverage of the membrane. The membrane was then moved to a developing cassette and exposed to X-ray film in a dark room for varying amounts of time to acquire different exposures. The signal was developed and fixed in the film processor using RG universal X-ray developer and using RG universal X-ray fixer respectively.

Western blot quantification.

Developed films were scanned in on an EPSON EXPRESSION 1680 pro scanner in JPEG format at 1600 dpi. Relative protein expression levels in each sample were determined by first quantifying the band densitometry of the target protein in FIJI software. This value was then normalised to the band densitometry of an internal control for the same sample. A medium exposure was chosen for each protein to ensure it was within the detection range and that signals were not saturated.

2.2.17 CAL51 cell lines for RNA-seq

Optimised HOTAIRM1 knockdown protocol was followed (section) aiming to get the recommended number of cells for sequencing (10^6 cells per sample). 48 hours post transfection, cells were gently washed with PBS, trypsinised, pelleted and stored at -70°C. frozen cell pellets were sent on dry ice to Genewiz (sequencing service) to be sequenced.

2.3 Computational methods

All codes used in this project can be accessed through the following this URL: https://drive.google.com/drive/folders/1JuHXVMtmcnoRc5UdozUVZS4_-gGiqVIc?usp=sharing

2.3.1 RNA sequencing

2.3.1.1 data processing

RNA-seq data was obtained as BAM files of aligned reads that was annotated using supplied Gene Transfer Format (GTF) file, that is an essential annotation file that contains information about genes and their corresponding features (chromosome name, starting point, ending point, strand…, etc). Sequenced reads were then counted by applying FeatureCounts command with the help of Dr Mark Dunning (personal communication). Sample information matrix was created defining each sample and what phenotype it represents. In R software, matrixes were submitted, and data was re-arranged following the pipeline requirements. Next stage was data processing, it included filtering unexpressed or lowly expressed genes followed by data quality control assessment and finally normalisation. The process of identifying genes with differences in expression levels between different experimental conditions was done through differential expression analysis (DEA), that was performed using Deseq2 package following package manual. DEA results file was annotated using BioMart package to get gene biotypes and gene names.

2.3.1.2 Prioritisation of lncRNAs for molecular studies

Differentially expressed lncRNAs were filtred on three levels. The first filter applied was the consideration of the combined effect of fold change (FC)= 1.5, as a

measure of change in gene expression level, and p-value of < 0.005, as a measure of statistical significance of FC value, to increase the validity of lncRNAs presumed to be differentially expressed between MCF-7 and TAMR (Conesa *et al.*, 2016). The second filter was the protein coding probability score generated using Coding Potential Assessment Tool (CPAT), a score of < 0.3 was determined the optimal cut-off to select the true noncoding transcripts according to (Wang *et al.,* 2013). For the third filter, annotation, /and characterisation of lncRNAs were evaluated. Next, the shortened list of prioritised lncRNAs generated from the above filtering steps were assessed in term of the published literature, annotation, and availability of reagents upon choosing the first lncRNA for molecular evaluation.

Four candidate lncRNAs (LUCAT1, SOX21-AS1, NR2F1-AS1 and HOTAIRM1) were selected and further reviewed in the published literature. Inspection of the genomic location was done using ensemble genomic browser (Genome assembly: GRCh38.p10) to survey neighbouring genes and transcription factors binding sites (Zerbino *et al.*, 2018).

2.3.1.3 GSEA

Gene set enrichment analysis (GSEA) was performed with software ((Debrabant, 2017). Raw counts were normalised using Deseq2 software, input files were prepared according to GSEA protocol, the first file is the expression dataset file that contain Contains features of consistent gene identifiers (ensemble IDs) in rows, samples one in each column, and a normalised number of counts for each gene in each sample. The second file was a phenotype labels file contains GSEA format of categorical labels that define a discrete phenotype associating each sample with a specific phenotype. Final file needed was a gene sets file selected from file servers

hosted on GSEA associated Molecular Signature Database (MSigDB), list of selected gene sets selected are listed in Table 2.7. These gene sets represent published lists of genes known to be associated with tamoxifen resistance pathways. In this type of files, each gene set has a given name and list of genes related evidentially in this set.

	genes	description
BECKER_TAMOXIFEN_RESISTANCE_DN	54	Genes downregulated in a breast cancer cell line resistant to tamoxifen [PubChem=5376] compared to the parental line sensitive to the drug.
BECKER_TAMOXIFEN_RESISTANCE_UP	54	Genes up regulated in a breast cancer cell line resistant to tamoxifen [PubChem=5376] compared to the parental line sensitive to the drug.
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_DN	88	Genes bound by ESR1 [GeneID=2099] and down-regulated by Estradiol [PubChem=5757] in MCF-7 cells (breast cancer).
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	211	Genes bound by ESR1 [GeneID=2099] and up-regulated by Estradiol [PubChem=5757] in MCF-7 cells (breast cancer).
BHAT_ESR1_TARGETS_VIA_AKT1_DN	82	Genes bound by ESR1 [GeneID=2099] and down-regulated by Estradiol [PubChem=5757] in MCF-7 cells (breast cancer) expressing constitutively active form of AKT1 [GeneID=207].
BHAT_ESR1_TARGETS_VIA_AKT1_UP	281	Genes bound by ESR1 [GeneID=2099] and up-regulated by estradiol [PubChem=5757] in MCF-7 cells (breast cancer) expressing constitutively active form of AKT1 [GeneID=207].
BOWIE_RESPONSE_TO_EXTRACELLULAR_MATRIX	17	Genes up-regulated by growing HMEC-E6 cells (mammary epithelial cells damaged by expression of HPV-16 E6 [GeneID=1489078]) in extracellular matrix (ECM).
BOWIE_RESPONSE_TO_TAMOXIFEN	18	Genes up-regulated by tamoxifen [PubChem=5376] in HMEC-E6 cells (mammary epithelial cells damaged by expression of HPV-16 E6 [GeneID=1489078]).
CLIMENT_BREAST_CANCER_COPY_NUMBER_DN	8	Genes from the most frequent genomic losses and homozygous deletions in a panel of patients with lymph node negative breast cancer (NNBC).
CLIMENT_BREAST_CANCER_COPY_NUMBER_UP	23	Genes from the most frequent genomic gains and amplifications in a panel of patients with lymph node negative breast cancer (NNBC).
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_1	533	The 'group 1 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 and ERBB2 [GeneID=2099;2064].
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_2	470	The 'group 2 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 and ERBB2 [GeneID=2099;2064].
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_3	726	The 'group 3 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 and ERBB2 [GeneID=2099;2064].
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_4	309	The 'group 4 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 but not ERBB2 [GeneID=2099;2064].
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_5	493	The 'group 5 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 but not ERBB2 [GeneID=2099;2064].
FRASOR_RESPONSE_TO_ESTRADIOL_DN	77	Genes down-regulated in MCF-7 cells (breast cancer) by estradiol (E2) [PubChem=5757].
FRASOR_RESPONSE_TO_ESTRADIOL_UP	35	Genes up-regulated in MCF-7 cells (breast cancer) by estradiol (E2) [PubChem=5757].
FRASOR_RESPONSE_TO_SERM_OR_FULVESTRANT_D	49	Genes down-regulated in MCF-7 cells (breast cancer) by selective estrogen receptor modulators (SERM) 4-hydroxytamoxifen, raloxifene, or ICI 182780 but not by estradiol [PubChem=44959;5035;3478439;5757].
FRASOR_RESPONSE_TO_SERM_OR_FULVESTRANT_U	23	Genes up-regulated in MCF-7 cells (breast cancer) by selective estrogen receptor modulators (SERM) 4-hydroxytamoxifen, raloxifene, or ICI 182780 but not by estradiol [PubChem=44959;5035;3478439;5757].
FRASOR_TAMOXIFEN_RESPONSE_DN	12	Genes preferentially down-regulated in MCF-7 cells (breast cancer) by tamoxifen [PubChem=5376] but not by estradiol or fulvestrant (ICI 182780) [PubChem=5757;3478439].
FRASOR_TAMOXIFEN_RESPONSE_UP	51	Genes preferentially up-regulated in MCF-7 cells (breast cancer) by tamoxifen [PubChem=5376] but not by estradiol or fulvestrant (ICI 182780) [PubChem=5757;3478439].
MASRI_RESISTANCE_TO_TAMOXIFEN_AND_AROMAT	20	Genes down-regulated in derivatives of MCF-7aro cells (breast cancer) that developed resistance to tamoxifen [PubChem=5376] or inhibitors of aromatase (CYP19A1) [GeneID=1588].
MASRI_RESISTANCE_TO_TAMOXIFEN_AND_AROMAT	20	Genes up-regulated in derivatives of MCF-7aro cells (breast cancer) that developed resistance to tamoxifen [PubChem=5376] or inhibitors of aromatase (CYP19A1) [GeneID=1588].
MASSARWEH_RESPONSE_TO_ESTRADIOL	62	Genes rapidly up regulated in breast cancer cell cultures by estradiol [PubChem=5757].
MASSARWEH_TAMOXIFEN_RESISTANCE_DN	257	Genes down-regulated in breast cancer tumors (formed by MCF-7 xenografts) resistant to tamoxifen [PubChem=5376].
MASSARWEH_TAMOXIFEN_RESISTANCE_UP	581	Genes up-regulated in breast cancer tumors (formed by MCF-7 xenografts) resistant to tamoxifen [PubChem=5376].
RIGGINS_TAMOXIFEN_RESISTANCE_DN	221	Genes down-regulated SUM44/LCCTam cells (breast cancer) resistant to 4-hydroxytamoxifen [PubChem=63062] relative to the parental SUM44 cells sensitive to the drug.
RIGGINS_TAMOXIFEN_RESISTANCE_UP	67	Genes up-regulated SUM44/LCCTam cells (breast cancer) resistant to 4-hydroxytamoxifen [PubChem=63062] relative to the parental SUM44 cells sensitive to the drug.
WP_TAMOXIFEN_METABOLISM	21	Tamoxifen metabolism

Table 2.8. List of selected MSigDB gene sets.

After loading data into GSEA, analysis was performed using the default settings except for “perpetuation type” which was set to “gene_set”. Gene sets were identified as significant if false discovery rate (FDR) q-value < 0.05 to control the expected number of false positive genes.

2.3.2 Analysis of publicly available sequencing data

2.3.2.1 The Gene Expression Omnibus (GEO) Project

GEO is a public database that archives high throughput genomic data (microarray and RNA-seq) from different independent studies supplied by research (Barrett *et al.*, 2013). Steps of bioinformatic analyse of genes expression in breast cancer datasets are

Dataset Selection

Selecting the right datasets is of great importance. Datasets built upon studies that has the appropriate experimental conditions suitable for our project. Consequently, the resulted lists of differentially expressed genes would be considered highly related to the phenotype under investigation.

Planning and creating a systematic search consist of the following steps:

1. Formulating a focused search question.
2. Select an appropriate database to search in.
3. formulate the key concepts that explain different elements of the question.
4. Document the search process.
5. Identify synonyms for search key words.
6. Use variations with search words such as truncation, spelling differences and abbreviations.

7. Use database-appropriate syntax, parentheses, and Boolean operators.
8. Collect the articles that appeared in the search.

Data analysis

Systematic search for microarray studies resulted in 120 GEO series, for a study to be included it has to have the appropriate number of samples which is three in each experimental condition, microarray platform should be either illumina or Affymetrix and raw data files (.CEL) should be available in the database. 5 datasets were included in the analysis (3 clinical biopsies and 2 in-vitro cell culture models). GEOquery package manual were followed to perform the DEA.

2.3.2.2 The Cancer Genome Atlas (TCGA)

This project collects high throughput sequencing analysis of tumour samples and the clinical information of participants diagnosed with more than 20 types of cancers (Li *et al.*, 2017). Gene expression profiles from the TCGA study breast cancer project (BRCA) were accessed using the TCGAbiolinks package in R software. Using samples barcodes (Figure 2.2) cases were identified to have matched normal solid and solid tumour tissues samples were selected to be compared. Then PAM50 classified barcodes of BRCA data were compared as follows, (luminal A tumours vs basal tumours), (luminal A normal vs luminal A tumours), (Basal normal vs Basal tumours). HER2 enriched and luminal B subtypes were excluded. Considering the aim of this analysis, to select lncRNAs and protein-coding genes of interest from the DEA results table: lncRNAs were divided into up- or down- regulated genes based on fold change direction. However, due to the high number of differentially expressed protein-coding genes, the cut-off of a significant p-value and FC is 0.005 and 1.5 respectively.

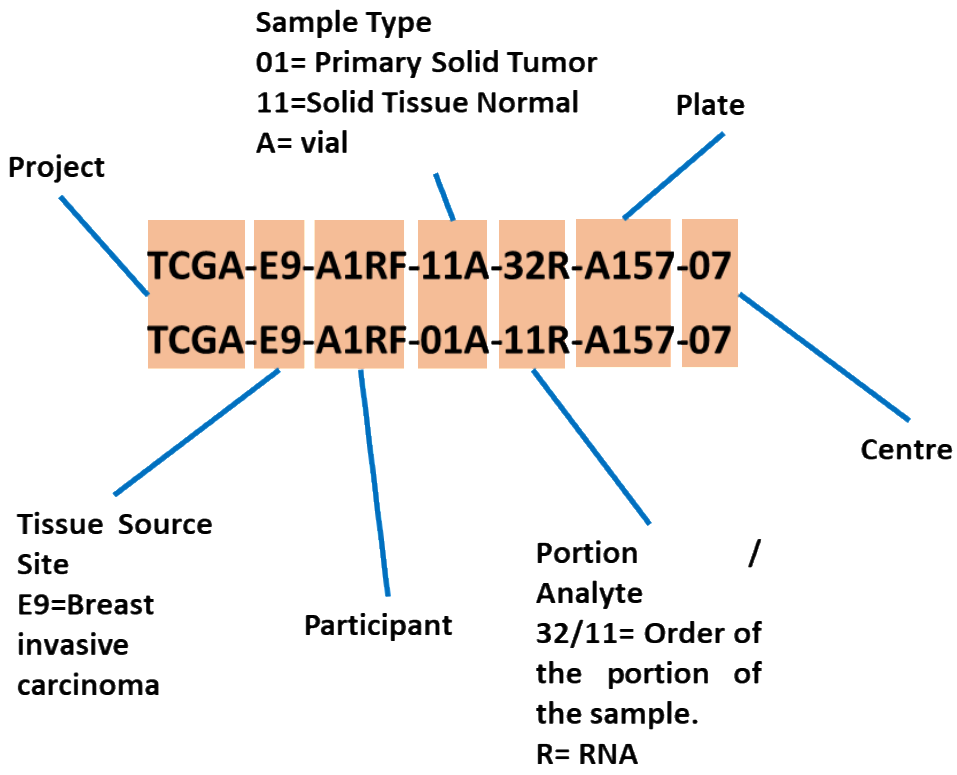


Figure 2.2. illustration of how a sample is assigned a TCGA barcode at each processing step. This example is for matched tumour and normal samples from the same case.

2.3.2.3 The cancer cell line encyclopaedia (CCLE) data

RNA-seq Gene expression profiles of breast cancer cell lines were downloaded from the CCLE database (Barretina *et al.*, 2012) Data included gene expression for 1019 cell lines, of which 51 breast cell lines were selected that were marked ER-positive or ER-negative based on the provided metadata. Gene expressions were provided in Reads Per Kilobase Million mapped reads (RPKM) that defined as the output value after normalizing the read counts in term of sequence depth and gene length. Log2 RPKM was calculated and candidate lncRNAs expression and correlation to ESR1 gene were evaluated.

Chapter 3. Bioinformatic analysis of tamoxifen resistant and sensitive breast cancer cell lines.

3.1 Introduction

Transcription is the flow of genomic information from DNA to RNA, it is a crucial biological step that controls cellular homeostasis and the manifestation of different normal or pathological traits (Mattick, 2003). The transcriptional landscape encompasses the whole population of RNA molecules consisting of protein coding and a variety of non-protein coding transcripts; it forms an intermediate passageway connecting genomic DNA to the protein end-product (Djebali *et al.*, 2012). The classic assumption of protein coding transcripts being the functional unit in the cell has changed dramatically. Noncoding regions in the genome were found to be pervasively transcribed, producing mainly lncRNAs among others (Szymański *et al.*, 2003). Furthermore, across-organisms genome analyses showed quite similar numbers of protein coding genes in multicellular organisms, but the increasing ratio of non-protein coding genes relative to the total genome size in higher organisms was correlated to the evolutional complexity (Taft, Pheasant and Mattick, 2007; Liu, Mattick and Taft, 2013). In addition, the ENDCODE project proposed that about 90% of pathological single nucleotide polymorphisms (SNPs) lay in noncoding areas (Myers *et al.*, 2011). Based on these arguments, non-coding RNAs are believed to be key functional molecules orchestrating arrays of biological pathways. Functional genomic analysis of many candidates long noncoding RNAs revealed their broad functionality, acting as mediators of biological developmental and differentiation (Dinger *et al.*, 2009). Recent data suggested their driving role in

complex diseases especially in cancer and lncRNAs have emerged as potential biomarkers and drug targets (Silva, Bullock and Calin, 2015). For example, lncRNAs UCA1 (urothelial carcinoma associated 1) and telomerase RNA component (TERC) were found be bladder carcinoma non-invasive biomarkers with a strong sensitivity and specificity (70-80%) (Srivastava et al., 2014; Chen et al., 2022). One clinically implemented lncRNA biomarker is prostate cancer antigen 3 (PCA3), a test is available commercially in the form of Progensa® PCA3 test kit and is used for diagnosing and follow up in prostate cancer (Haese *et al.*, 2008; Hologic, 2012). Identifying more cancer relevant lncRNAs for practical use in the field of cancer diagnostics and treatment is facilitated through understanding the functional role of lncRNAs is tumorigenesis.

RNA-seq is a fundamental next generation sequencing technology and one of the most reliable methods used for studying lncRNA. Combined with other experimental methods it deciphers the interconnected biochemical, physiological, and biological systems controlling different cellular organisation. This process provides a wide range of molecular information on the input samples. Through measuring the global abundancy of RNA transcripts, gene expression analysis is carried out. It offers quantitative measurement, hence perspectives on the transcriptional patterns. When detecting the similarity and diversity in transcriptional profiles between two distinctive experimental conditions, dysregulated RNA signaling is detected and a cause effect relationship can be established and further investigated.

The hypothesis of this chapter is:

Gene expression analysis of tamoxifen resistant (TAMR) and parent tamoxifen responsive breast cancer cell lines (MCF-7), will produce a list of candidate lncRNAs that may have biological function relevant to tamoxifen acquired resistance after long term tamoxifen treatment.

The aims of this chapter are:

1- To process RNA-seq data of TAMR and MCF-7 cell lines using quality control and differential expression bioinformatics pipelines.

2- To construct a list of statistically significant differentially expressed genes.

The objectives of this chapter include:

1- Validating input raw sequencing data by:

   A) Assessing samples library sizes and adjusting for any detected systematic biases.

   B) Checking sample to sample relationship to each other and global distribution of count data.

2- Performing differential expression analysis (DEA).

3- Ranking lncRNAs and protein coding genes based on DEA results.

4- Choosing a few lncRNAs for further investigation.

3.2. Results

3.2.1. Sequencing data generation

Transcriptome quantification is one of the core RNA-seq activities, enabling other downstream analyses. In this project, RNA--sequencing of breast cancer cells aimed to establish a regulatory role of lncRNAs in tamoxifen resistance. Cell lines were kindly provided by Dr Julia Gee, Cardiff University (Knowlden *et al.*, 2003). Before sequencing the tamoxifen sensitivity of each cell line was confirmed. Both cell lines were exposed to increasing concentrations of tamoxifen and sensitivity determined by MTT assay (Figure 3.1). MCF-7 cell lines were significantly more sensitive to tamoxifen at all tested concentrations. Mean lethal concentration (50% of cell mortality observed (LC50) of tamoxifen in MCF-7 was 1 µM, while for TAMR cells, even with the highest concentration of tamoxifen (10 µM), only 30.6% of cells were affected, compared to 87.7% of MCF-7 cells.

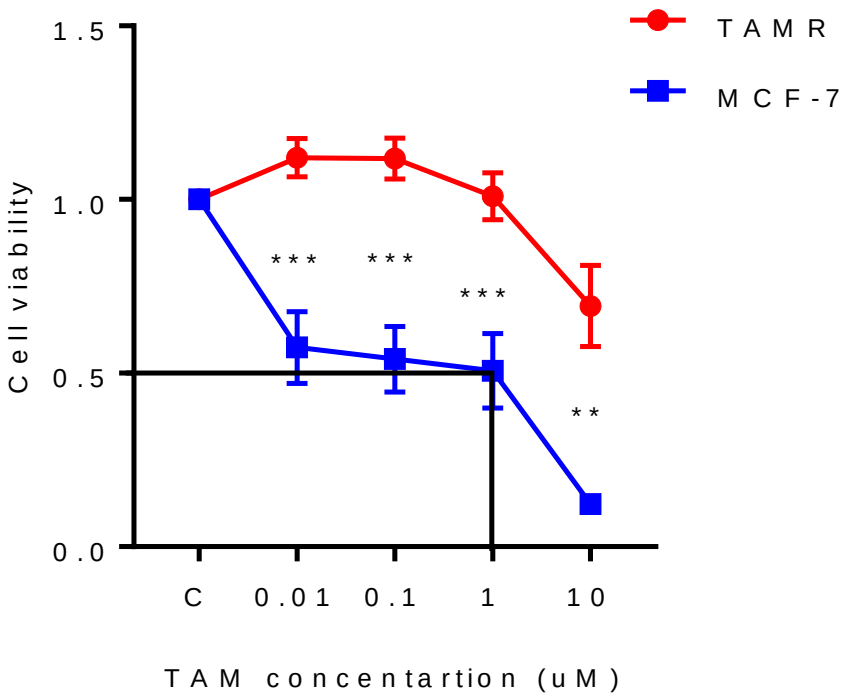


Figure 3.1 TAMR and MCF-7 cells sensitivity to tamoxifen. Cells were cultured in 96-well plates, treated with increasing concentrations of tamoxifen (0.01 µm, 0.1 µm, 1 µm and 10 µm) and vehicle control (C). MTT was performed 4 days post treatment, by reading optical densities in the plate reader. Cell viability was calculated by dividing tamoxifen treated well reads by vehicle control read. Data points represent mean cell viability of each treatment group. Error bars depict standard deviation of the mean (N=3). Statistical significance was determined using unpaired one-tailed Student's t-test at each concentration * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p=≤0.001

Following conformation of tamoxifen resistance in TAMR vs parent MCF-7 cells. Total RNA was extracted, then in-house quality and concentration assessments were performed. Further in-depth quality control assessments were carried out in the sequencing facility. As a result, biological samples were confirmed of high-quality RNA with no gDNA contamination or RNA degradation. Samples were prepped and sequenced following Illumina-based RNA-seq protocol. FASTQ-format files were generated, sequenced reads were aligned to the human reference

genome and stored in the form of BAM files that then were annotated, and sequenced reads were counted. Next, computational analysis of RNA-seq data in R software started by reading and formatting the data to give 2 basic files: 1) Count matrix of measured gene expression data that included 58,381 genes (also called feature counts) identified by the official Entrez gene ID and their number of reads in each of the 6 samples, 3 repeats of each cell line. 2) a comma separated file containing meta data or sample information to identify each sample and provide information that help determine the type of statistical comparison desired. Both files, used together, form the foundation of this bioinformatics analysis (Figure 3.2).

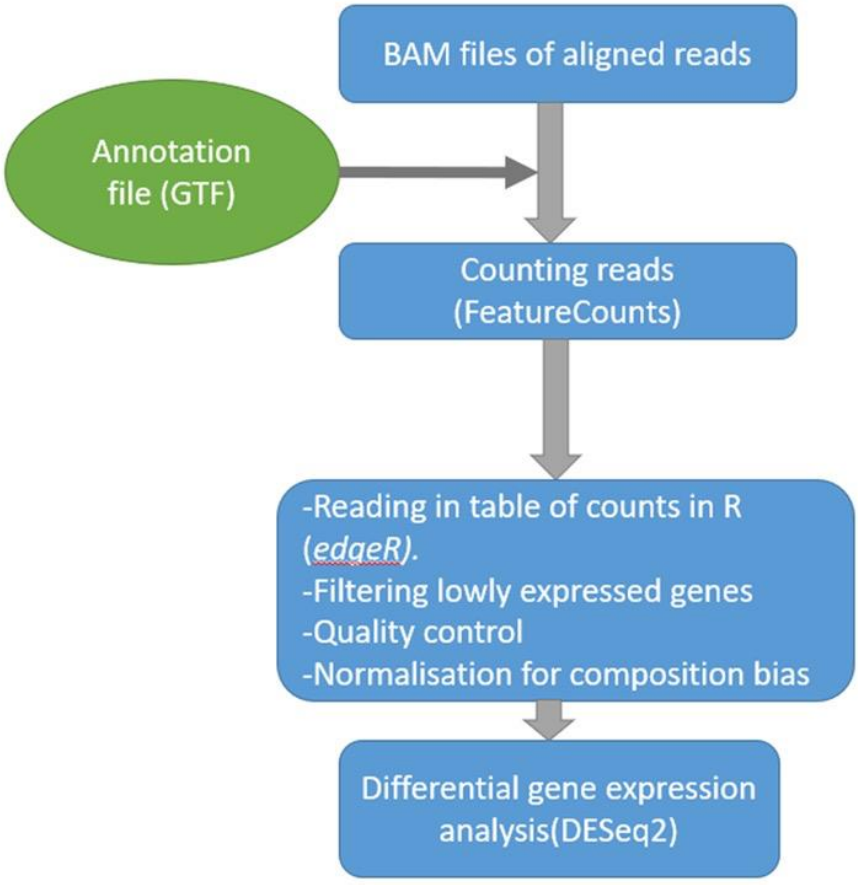


Figure 3.2 Flow diagram of pipeline used to analyze RNA-seq data.

3.2.2. Data filtering

Although RNA-seq is very reliable deep-sequencing technology, the amount of data produced is huge and presents many statistical challenges. It is very important to address such issues early on, carefully considering the upcoming biological and computational limitations. The extensive statistical analyses the data undergo, demanded setting an expression standard value to help us demarcate gene expression standards. Filtering is a common practice that allows for excluding lowly expressed genes, alleviating the statistical burden they procure. As a result, more sensitive detection of deferentially expressed genes was obtained, increasing both confidence and precision of downstream analyses and increasing the number of detected DEGs.

For this thresholding method was used filtering low-expression genes using the edgeR package. Initially, untranscribed genes were defined to have 0 count reads in the input sequenced genes matrix were removed. Count-per million (CPM) values were calculated from raw counts. CPM were used as it corrects for sample library size as a method of read count normalisation. Filtering threshold was sited at 0.5 CPM, as it fulfils the recommended limit for the number of counted reads in each gene to be considered expressed (5-10 counts in each sample library). In addition, a gene should be expressed in at least 3 samples to avoid being filtered out. As a result, from a total of 58,381 genes, 14,780 met expression inclusion conditions while 43,601 genes were excluded being below expression threshold and/or expressed in less than 3 samples at the same time. Figure 3.3. depicts how the expression distribution changed dramatically after filtering of lowly expressed

genes. Before filtering, a large proportion of sequenced genes clustered around areas defining them as not or lowly expressed. After applying filtering standards (0.5 CPM expression threshold in at least 3 of the 6 samples), the plotted curve of filtered data has changed focus onto highly expressed genes, excluding all uninformative genes and clearly showing overall data uniformity across all samples.

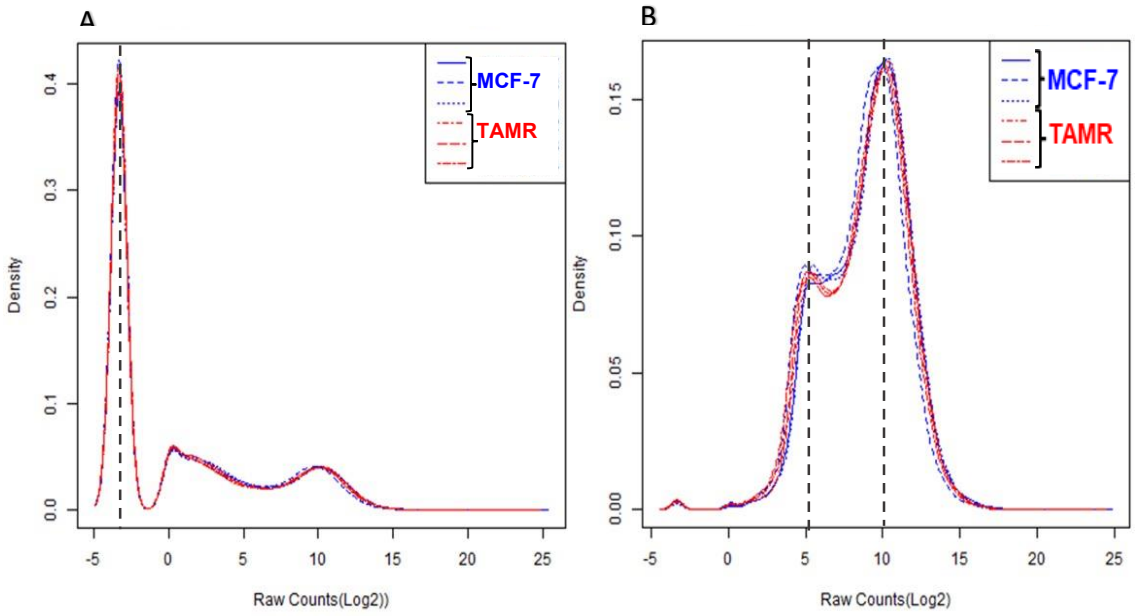


Figure 3.3. Raw counts distribution before and after filtering of lowly expressed genes . Density plot illustrating the effect of Filtering lowly expressed genes on the number of counts in each sample. A) majority of genes have zero reads, manifested in the peak at -3.322 corresponding to log2 (0+0.1) (dotted line). B) After excluding genes with a number of counts less than 0.5 CPM in at least 3 of the 6 samples, the area under the curve corresponding to (35-1000 reads) expands, (dotted lines), this reflects the shifted focus onto genes with a satisfactory number of counts. Blue lines are MCF-7 samples, and red lines are TAMR samples. X-axis is log transformed raw counts, y-axis represent the amount of genes bearing certain number of counts.

3.2.3 Quality control

After filtering below-standard genes, the following analysis focused on assessing the quality of satisfactorily expressed genes. The presence of non-biological factors can have a major effect on sequencing quality, resulting in noticeable variations in

the data produced. Batch effects may result from subsets of samples exposed to technical variabilities due to practical restrictions. This might occur due to laboratory conditions, such as different technicians handling the samples or even changes in lab atmosphere. Also, the extensive list of reagents, kits used in RNA-seq preparation varying in lots or providers or different sequencing platforms used can contribute to creating this batch effect. This is a major issue that needs to be addressed early, to avoid letting easily recognisable sources of variation result in incorrect conclusions. In our sequencing study, the number of paired-end reads were plotted for each sample. Bearing in mind the principle understanding that sequencing data is not normally distributed, and a bell-shaped continuous bar plot is not expected. So, while it is normal for sequencing data of independent repeats to have different library sizes no obvious or extreme differences should be recognised. Using raw count measures to visualise library size laying (Figure 3.4), we assumed that a good degree of read distribution with no major discrepancies or imbalanced coverage had been achieved and that further analysis downstream could occur. Furthermore, variance stabilizing transformation (VST) of the data was carried out to investigate more exploratory plots for the quality control step. The *vst* function from DESeq2 package was used to calculate log2 cpm and other calculations in order to eliminate internal factors that could create bias such as outliers. Other, benefits of this transformation are that it adjusts for data skewness making it conform to a normal distribution model and reduces variability. The output of this step is a matrix of normalised counts. This method accounts for extreme reads resulting from variations in sequencing depths and library composition, where many genes' counts fluctuate between samples. Added to that, the process is

blinded for study design, to further eliminate any hint from the provided sample information allowing for an unbiased comparison. This is especially important to observe in our experimental design, that compares between tamoxifen sensitive and resistant cell lines. Tamoxifen, a selective oestrogen receptor modulator exerts widespread effects as an oestrogen agonist/antagonist through up- and down regulating hundreds of genes (Frasor *et al.*, 2003) (Frasor *et al.*, 2006). This might result in wide disruptions in genetic activity, requiring robust methods of normalisation and inquiring data quality from different angles. Looking into the scaled data (figure 3.5), the distribution of scaled reads counts is uniform, with very close clustering around the median (blue line). Both analyses above conclude the input data is of good sequencing quality.

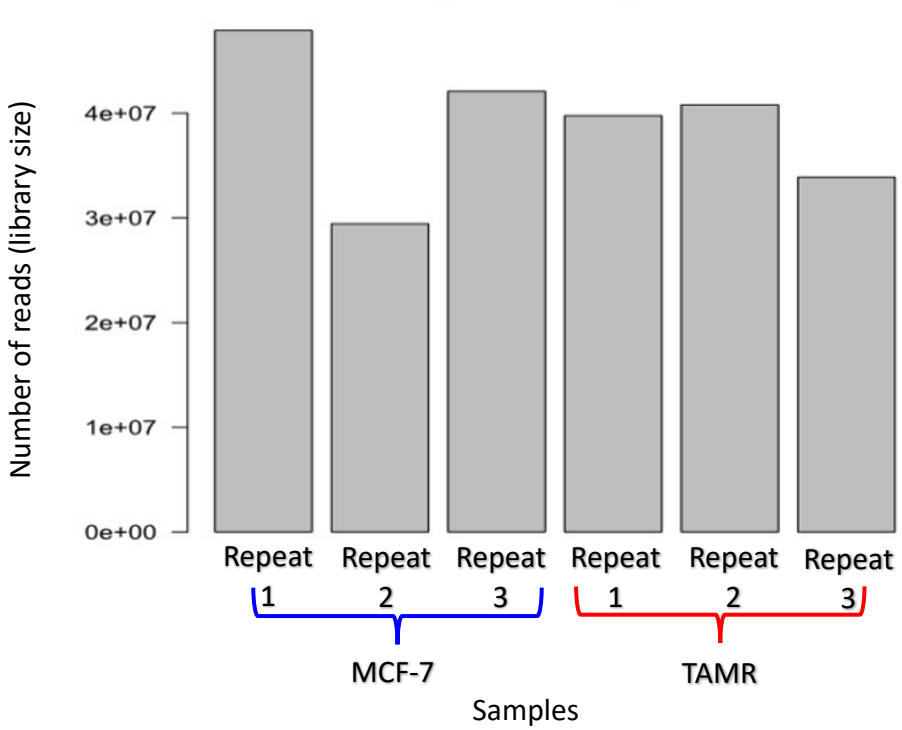


Figure 3.4 Library sizes of RNA-seq samples. Samples were sequenced, 3 samples in each cells group and the total number of raw counts in each sample adapted a bar indicating the library size.

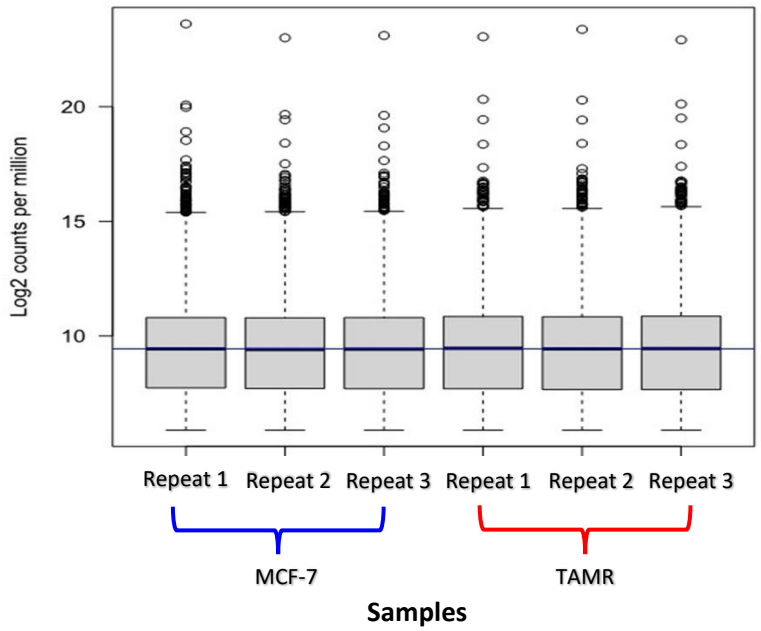


Figure 3.5. Similar distribution of expression profiles were confirmed by boxplot of scaled RNA-seq data. samples were sequenced and variance-stabilising transformation was applied on CPM transformed expression values to check all genes distributions on the log2 scale across all samples. blue horizontal line corresponds to the median log2 CPM.

The next part of quality assessment focused on evaluating the data at the sample level. The aim was to build a hierarchy of samples by deciding which samples to combine in a cluster and which to split apart, this is done based on a measure of distances between the observations in samples, following their overall degree of similarity or dissimilarity is determined. For this purpose, *dist* function was used on the filtered normalised data to create a matrix of Euclidean metrics that help performing hierarchical agglomerative clustering in the form of a heatmap of sample-to-sample distances between rows and columns. This showed independent repeats of the same cell line belonging to the same cluster, indicating a considerable correlation within the same biological group of samples. On the other hand, a weak correlation between MCF-7 and TAMR can be detected (figure 3.6). Also, unsupervised principal component analysis was performed; this is a data transformation technique. After variable value standardisation, the data is compressed and the dimensional distances between samples is calculated, this reflects the degree of variance each group of samples processes. This method is another way to separate samples based on how different they are in term of intrinsic gene expression. Consistent with previous findings, our data appear to cluster well and spread across the two main principal components (PC1 and PC2), that account for 97% and 2% of the total variation respectively (figure 3.7). The TAMR samples cluster closely together, reflecting how similar their transcription profiles are. At the other end of the PC1 axis, the MCF-7 samples look less clustered. This finding does not negatively affect MCF-7 clustering, because the samples diverge on the PC2 axis that is inferior to PC1 in determining cluster distinction (only 2% of variation). Taken all together, quality control output support data validity. The data suggests

that biological properties are the main source of variation, which is consistent with the experimental design, hence we can conclude that the data is fit for subsequent analyses.

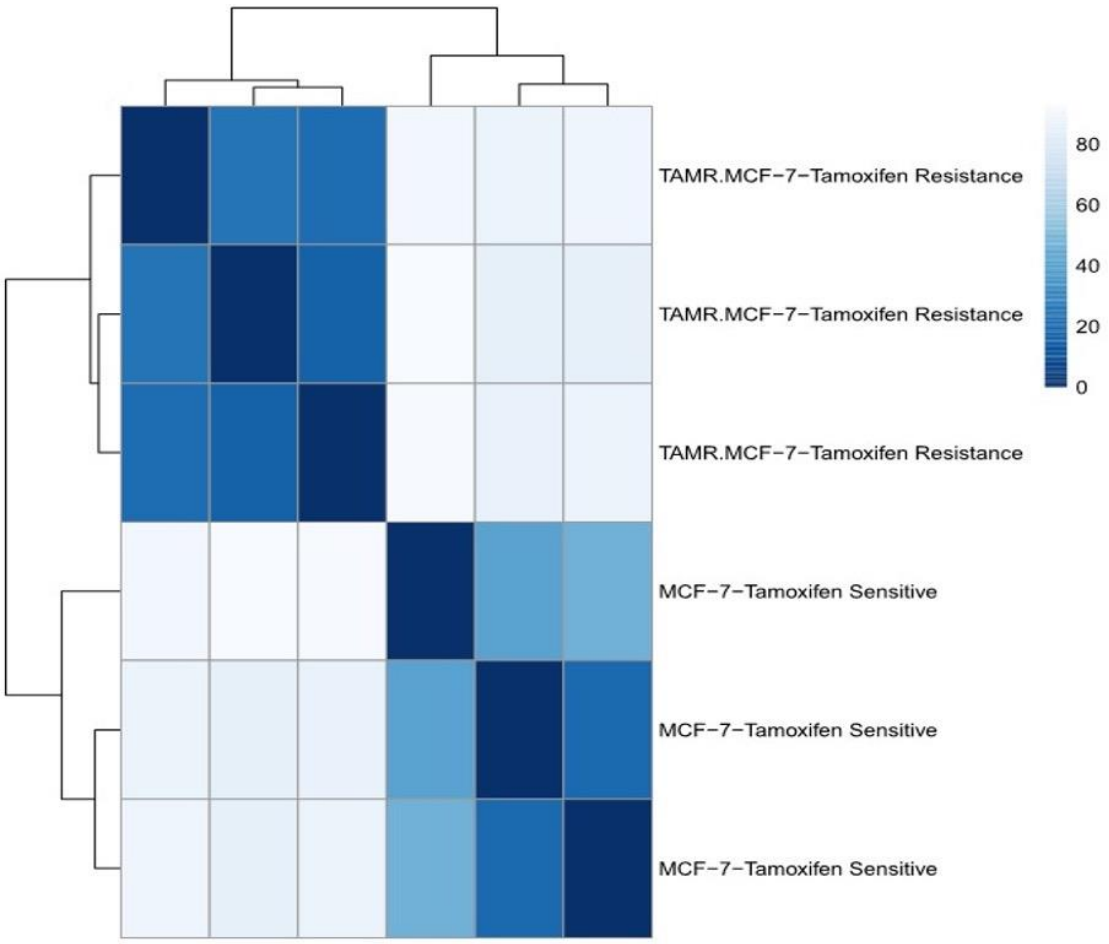


Figure 3.6. Heatmap of sample-sample distances. Variance-stabilising transformation was applied to raw count data to calculate distances between samples. RNA-seq samples are plotted against each other in a two-dimensional grid, Each cell in the matrix contains a calculated sample-sample distance value that corresponds to a specific colour gradient the shorter the distance between samples shows higher intensity of the colours in cells, ranging from dark blue to white (refer to the color scale on the right upper side of the figure). The dendrogram on the sides of the heat map shows clustering of similar samples.

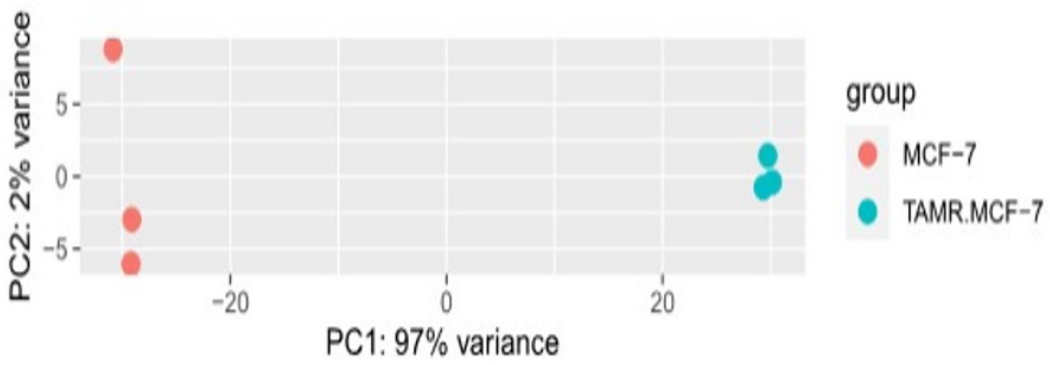


Figure 3.7. Principle component analysis plot showing variability in global gene expression between samples. Variance-stabilising transformed gene expressions were determined, then, the pattern of unsupervised variability between samples was plotted on a 2D plan as percentage, where, PC1 on x-axis were plotted against a PC2 on y-axis. Sample groups are indicated in coloured dots (red: MCF-7, blue: TAMR). The relationships between samples are determined by judging the distance between samples on the horizontal PC1 axis, the further the distance between samples, the highest degrees of variability.

3.2.4 Differential expression analysis (DEA)

The main utilization of transcriptomic data produced by RNA-seq is for DEA. Sequencing data provides quantitative information about the genes in the samples, when performing DEA, the observable trend can be explained.

The uploaded data has two sample levels, tamoxifen resistant (TAMR samples) and tamoxifen sensitive (MCF-7 samples). The MCF-7 group of samples were assigned as the reference level, the goal of this analysis is to find out which genes changed when cells adopted the tamoxifen resistant phenotype.

After filtering off lowly expressed genes and testing the quality of samples at different levels, now we have the final count matrix as an input for DEA. The basic principle of statistical detection of differentially expressed genes is to measure the dispersion of gene counts. Dispersion parameter estimates the metric location of each repeat relative to the mean. By correctly estimating the dispersion, more reliable DEA results are obtained (Landau and Liu, 2013).

3.2.4.1 RNA-seq data properties and distribution

Initially, the statistical properties of the RNA-seq data were assessed. To carry out a DEA, the data distribution needs to be determined and fitted to an appropriate model. As shown in figure 3.8.A, count data clearly did not fit within a normal distribution curve. Also, additional properties about the underlying data were observed. Firstly, the values were normal numbers, never including minus or decimals, as a fraction or subtract of a read cannot be mapped to a gene. Secondly, the data was continuous and its range dynamic, this means it does not fall within a fixed range of values (the long right tail). Taken all together, the RNA-seq data suggested it could be a Poisson distribution, that mainly evaluates the relationship between mean and variance in the data. A vector of mean and variance was created by calculating each for each variable, then both were plotted against each other. However, when the data were fitted into a Poisson model, the observed trend did not fit as the model requires the mean counts to be equal to the variance counts (Witten, 2011). As shown in figure 3.8.B, across samples, genes in our data have high mean and variance expressions, meaning the Poisson model is inappropriate. Considering the data properties and the observed degree of variability in data, the next candidate model was negative binomial distribution figure 3.8.C. Plotting of per-gene dispersion estimates in R allowed us to see that dispersion decreased as the mean increases, which is the expected trend of RNA-seq data of this nature (Ren & Kuan, 2020).

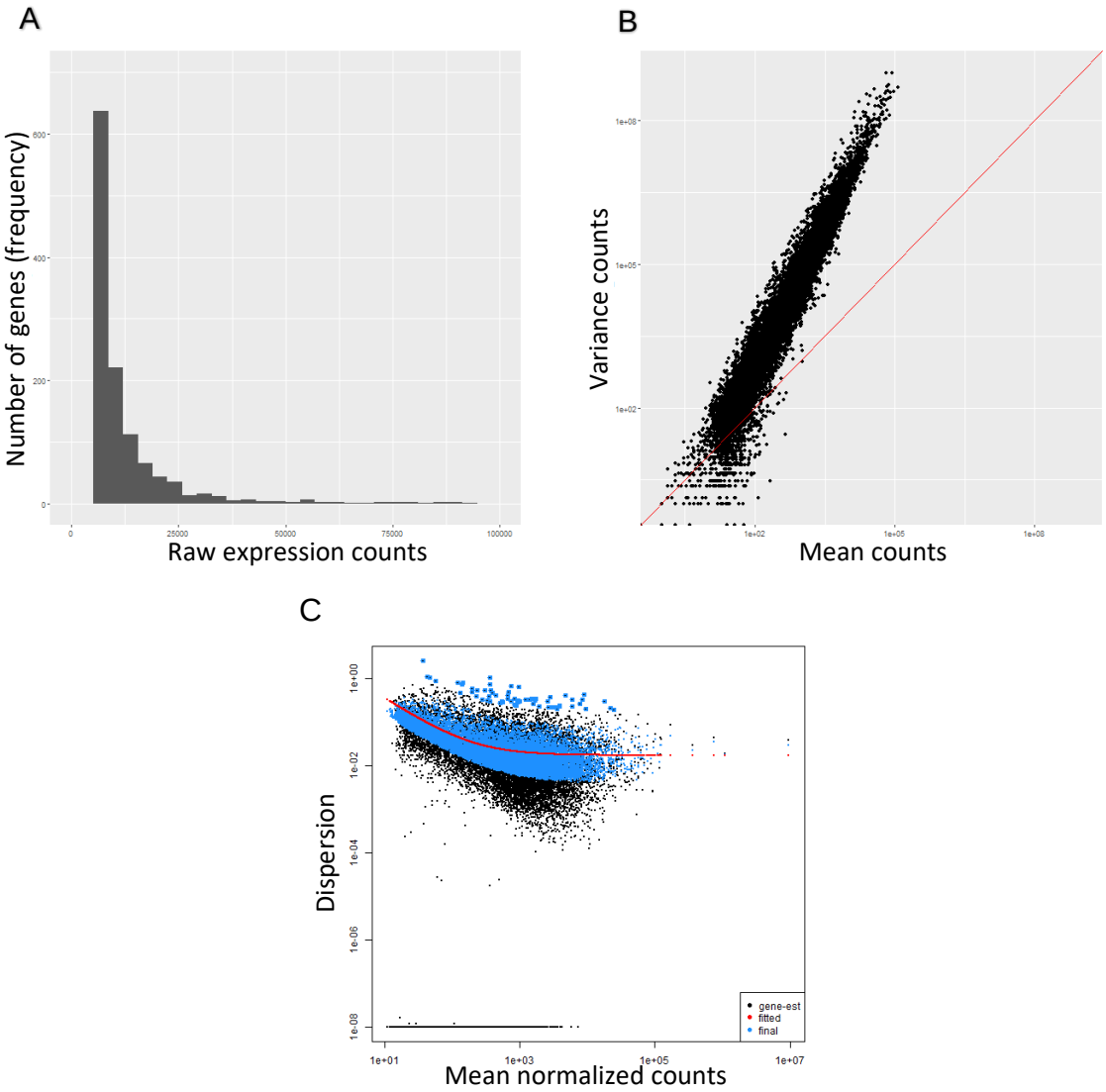


Figure 3.8 Distribution of gene expression levels in RNA-seq samples. (A) Histogram of count data illustrating basic properties of the expression set. X-axis represents the number of reads mapped to a particular gene, y-axis indicates the frequency by which a particular number of counts recuring in the data set. (B) Poussin distribution (PD) plot, counts were fitted within PD parameters, mean counts on x-axis and variance counts on y-axis. Red diagonal line is where mean equals variance for a particular gene. Black dots represent individual genes. Genes falling to fall along the red line indicates unfit to PD model. (C) Deseq2 plot of dispersion estimates, the higher the mean value on x-axis , the more probability the gene is found in tested samples, y-axis is the dispersion parameter that illustrate how spread samples are. Solid blue dotes are the genes with the final adjusted dispersion values, black dotes surrounded by blue hue are genes that have a very high dispersion compared to other genes, the red curve is the genes expected value of dispersion.

3.2.4.2 DEseq2 analysis

DEseq2 is a popular bioconductor package that adopts the negative binomial distribution (NBD) to measure differential gene expression across genes count matrix (Love, Huber and Anders, 2014). NBD uses a linear model to account for the extra variability in the variance produced by the experimental design. By accounting for library size and composition, DEseq2 estimates a size factor for each sample by applying the median of ratio method. Subsequently, gene-wise dispersion or spread relative to mean counts is calculated and log2 fold change is calculated based on the linear model equations programmed with in the package.

The principal concept is shown in the first equation below where (Kij) is the number of mapped sequenced reads (K) for gene (i) in sample (j). Kij is to be represented by the negative binomial model (NB) which has two main parameters, sample and gene specific fitted mean (μij) and extra variability resulting from experimental design per-gene (i) is accounted for by the dispersion parameter (αi). The fitted mean (equation 2 – below) in turn consists of two parameters, firstly sample specific normalisation, or size factor (Sj) - a constant to be used for all genes in one sample, and secondly the per-sample (j) concentration of gene (i) fragments (qij). Finally, log2 fold change between conditions is calculated based on a generalized linear model in the form of equation 3 (below). Where xj is the design factor that depends on the initial experimental design, in our case a one factor design comparing between samples (conditions), with two values tamoxifen sensitive and tamoxifen resistant conditions, and βi is the gene expression coefficient whereas explained in equation 4 (Table 3.1), where β_0 refers to the expression in the reference samples (tamoxifen resistant) or β_1 for the opposite condition.

1	$\mathrm{Kij} \sim \mathrm{NB}(\mu\mathrm{ij},\alpha\mathrm{i})$
2	$\mu ij = sjqij$
3	$log2(qij) = \sum \beta i.xj$
4	$log2(qij) = \sum (\beta_0 + \beta_1 \times xj)$

Table 3.1 Equations used for estimating data distribution.

3.2.4.3 DEseq2 results

After confirming our data fit using the DEseq2 basic statistical handling method, the raw sequencing data matrix was loaded into R software together with the sample metadata files. Subsequently, a differential expression results table was generated where rows are genes and columns consist of 6 calculated DEA parameters including "baseMean" (average normalised counts of all samples)," log2FoldChange"(change in gene expression in tamoxifen resistant samples group compared to tamoxifen sensitive samples group), "lfcSE" (Standard error estimates for log2fold changes), "stat"(Wald test values), "pvalue" (p-value or statistical significance), "padj" (statistical significance adjusted for multiple testing). (Table 3.2)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ENSG00000196208	12711.20049	9.2538810	0.24465640	37.823990	0.000000e+00	0.000000e+00
ENSG00000144476	8836.81731	4.9377768	0.13086139	37.732878	0.000000e+00	0.000000e+00
ENSG00000146072	5005.19198	-4.0108301	0.10282149	-39.007702	0.000000e+00	0.000000e+00
ENSG00000152402	5794.01529	-6.4286241	0.12461303	-51.588701	0.000000e+00	0.000000e+00
ENSG00000164932	1820.32602	-5.3454821	0.14891294	-35.896694	3.440031e-282	1.016873e-278
ENSG00000198467	3336.63613	-6.5182849	0.19598972	-33.258300	1.548568e-242	3.814639e-239
ENSG00000101144	92785.40099	-3.1716646	0.09704211	-32.683384	2.689997e-234	5.679737e-231
ENSG00000116016	5827.33507	-4.4070514	0.13614955	-32.369194	7.450336e-230	1.376450e-226
ENSG00000197959	1550.42319	-6.8507710	0.21324550	-32.126215	1.898274e-226	3.117389e-223
ENSG00000183421	2103.40423	-3.3836234	0.10819742	-31.272681	1.097850e-214	1.622622e-211

Table 3.2. Example of differential gene expression analysis results as an output of Deseq2 package.

If a p-value <0.5 is used, this means 5% of statistically significant genes are false positives, relating to our data, 7786 genes had p-value <0.05, implying 398 genes are a result of natural random variation, tamoxifen resistance had no effect on them. Here the importance of adjusting p-values can be seen, due to the massive number of low p-values and multiple statistical testing data undergo. For this, DESeq2 package uses Benjamini-Hochberg method to determine the false discovery rate adjusted p-value (padj), aiming control the probability of making at least one false positive, reducing the chance of making type one error.

Log2foldchange is the main element in the results table, it allows for evaluating gene expression between samples. log2foldchange values were used for hypothesis testing, the null hypothesis (H_0) states no change in gene expression between TAMR and MCF-7 samples, this was true for 9641 genes with log2foldchange of 0. On the other hand, 2506 genes had log2foldchange values more than 0 while 2633 genes had values less than 0 (Figure 3.9).

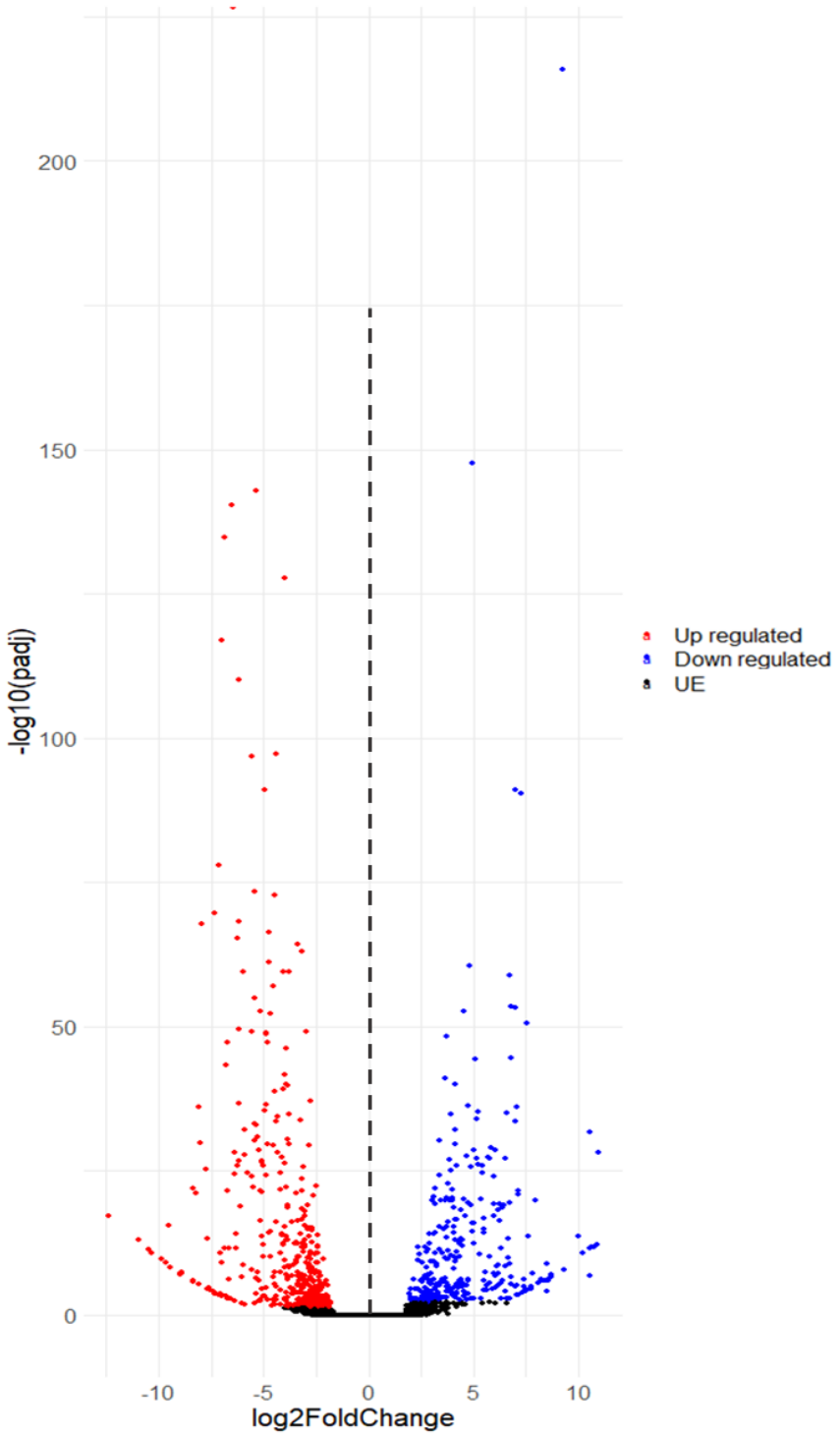


Figure 3.9. Volcano plot for differentially expressed genes in RNA-seq. The scatter plot was created by reading DEA results matrix in ggplot2 package. Data points represent genes coloured based on log2 fold change direction (red is up-regulated in TAMR and blue is down-regulated in TAMR and black dotes are filtered out genes due to below cut-off adjusted p-value or log2 fold change. X-axis is log2 fold change, y-axis is statistical significance (-log10 (adjusted p-value).

3.2.4.3 Annotation of DEseq2 results

The next step was to annotate the DEA results table. For this the BioMart data base was used, a query was built with ensemble ids, calling for the corresponding gene names and biotypes. The DEA annotated table included data for 12439 protein coding genes, 1511 lncRNAs, 7 Mitochondrial DNA (Mt RNA), 21 miscellaneous RNA (miscRNA), 10 micro RNAs(miRNA), 102 To be Experimentally Confirmed (TEC), 542 pseudogenes (199 processed pseudogenes, 195 transcribed unprocessed pseudogenes, 61 transcribed processed pseudogenes, 48 unprocessed pseudogenes, 36 transcribed unitary pseudogenes, 1 unitary pseudogene, 1 rRNA pseudogene, 1 polymorphic pseudogene), 1 Constant chain T cell receptor gene (TR_C) gene, 5 Small Cajal body-specific RNAs (scaRNA), 1 small conditional RNA (scRNA), 47 Small nucleolar RNAs (snoRNA), 36 Small nuclear RNAs (snRNA) and 57 un annotated IDs (NA values). Taken all together, from the 14780 genes that underwent DEA, 9544 genes were excluded for having an adjusted p-value >0.005, of the remaining 5235 genes, 770 were upregulated in TAMR, 1911 genes were upregulated but by less than 1.5 log2 fold change, 804 were downregulated in TAMR and 1750 genes were downregulated in TAMR but by less than 1.5 log2 fold change, visual expression of these results can be seen in (Figure 3.9). By observing the gene clustering pattern, we can get an indication of data integrity and of the expected load of statistically significant results, this is because genes were stratified visually based on holding an adjusted p-value of less than 0.005 and 1.5 log2 fold change.

When looking at the relationship between significant change in gene expression and expression strength of the genes, the DEA data reflect the observed raw number of genes to

be included and excluded in downstream analyses.

3.2.4.4 DEseq2 result filtering.

In this project our interest was focused primarily on lncRNAs and protein coding genes due to the nature of lncRNA regulatory function (G. Chen *et al.*, 2013). To increase the probability of detecting truly differentially expressed genes, we considered our experimental design that includes three samples per condition with a sequencing depth of about 15-20 million reads, consequently, we chose to change DEseq2 default settings that automatically apply log2 fold change of greater than or less than 0, to a log2 fold change of greater than 1.5 or less than -1.5, and from using default p-value of < 0.1 to using an adjusted p-value of < 0.005, that indicated 0.5% of final list of genes are likely false positives not tamoxifen resistance related genes. After filtering out unwanted genes, a total of 88% of lncRNAs and 90% of protein coding genes were excluded (Table 3.3).

	Before filtering		After filtering	
Gene biotype	Upregulated in TAMR	Downregulated in TAMR	Upregulated in TAMR	Downregulated in TAMR
lncRNAs	620	891	79	104
Protein coding genes	6316 Table 2	6123	594	627

Table 3.3 number of genes in results before and after filtering.

Finally, lncRNAs and protein coding genes were counted and divided into groups based on range of log fold changes (Table 3.4). LncRNAs upregulated in TAMR have log2 fold changes ranging from 8.8 to 0.42, and protein coding genes upregulated in TAMR have log2 fold changes ranging from 11.8 to 0.25. LncRNAs downregulated in TAMR have log2 fold changes ranging between 11.7 and 0.39 and protein coding genes of the same category range between 12.2 and 0.25.

	lncRNAs		**Protein coding genes**	
log2 fold change	**Upregulated in TAMR**	**Downregulated in TAMR**	**Upregulated in TAMR**	**Downregulated in TAMR**
≥5	10	16	85	84
5≥4	2	4	53	60
4≥3	21	16	97	81
3≥2	28	34	224	154
2≥1.5	17	35	167	215
<1.5	79	145	1841	1636

Table 3.4 Numbers of lncRNAs and protein coding genes according to log2 fold change values.

After filtering genes below the determined threshold, ensuring that genes with significant fold changes between tamoxifen sensitive and resistant experimental conditions have biological relevance. the next step was prioritisation of genes identified. This followed the pipeline in Figure 3.10.

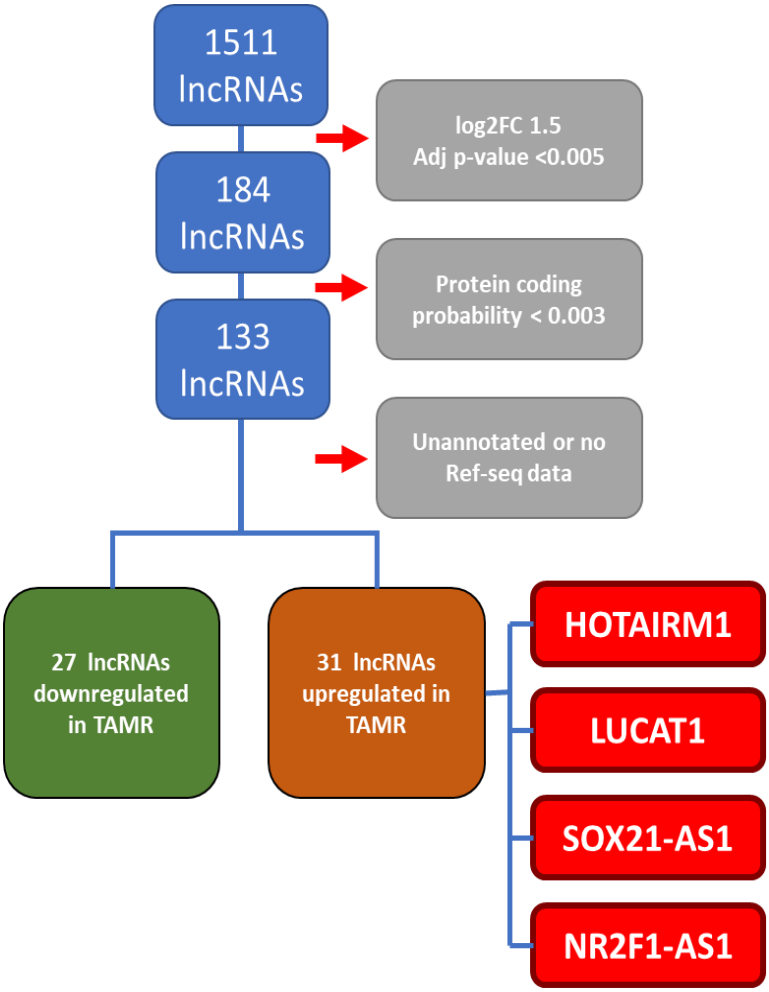


Figure 3.10 Flow chart of lncRNAs prioritisation. Diagrammatic representation of lncRNAs filtering, from DEA result table, we started this workflow with 1511 lncRNAs, after passing this number of lncRNAs through 3 filters 1.5 log2 fold change and FDR adjusted pvalue (adj p-value) of <0.005, excluded 1328 gens, protein coding probability <0.003 excluded 50 genes then filtering out genes missing annotations and ref.seq indexing excluded 45 genes. From the 43 lncRNAs upregulated in TAMR samples group, HOTAIRM1, LUCAT1, SOX21-AS1, NR2F1-AS1 lncRNAs were selected for further investigation.

3.2.4.5 Most dysregulated genes following all analysis – top 10s.

In addition, protein coding probability were calculated using CPAT, genes with a score less than 0.003 were selected to insure them being noncoding transcripts (Wang *et al.*, 2013). The top 10 lncRNAs and protein coding genes up- and downregulated in TAMR group of samples were exported (Tables 3.5, 3.6, 3.7 and 3.8)

Tables 3.5 and 3.6 Up and down regulated protein coding genes.

Ranking	ensembl_gene_id	log2FoldChange	padj	gene_name	gene_biotype
1	ENSG00000185008	-12.2383	5.28E-19	ROBO2	protein_coding
2	ENSG00000174343	-12.2057	8.20E-24	CHRNA9	protein_coding
3	ENSG00000041982	-11.4903	2.05E-16	TNC	protein_coding
4	ENSG00000175329	-10.9376	4.95E-15	ISX	protein_coding
5	ENSG00000155966	-10.6765	2.19E-14	AFF2	protein_coding
6	ENSG00000138435	-10.4532	3.17E-13	CHRNA1	protein_coding
7	ENSG00000187527	-10.2186	3.90E-17	ATP13A5	protein_coding
8	ENSG00000042980	-9.91368	9.34E-12	ADAM28	protein_coding
9	ENSG00000175556	-9.80161	4.55E-12	LONRF3	protein_coding
10	ENSG00000185008	-12.2383	5.28E-19	ROBO2	protein_coding

Ranking	ensembl_gene_id	log2FoldChange	padj	gene_name	gene_biotype
1	ENSG00000126947	11.8554	2.07E-17	ARMCX1	protein_coding
2	ENSG00000139352	11.62548	2.00E-17	ASCL1	protein_coding
3	ENSG00000213988	11.23954	2.40E-16	ZNF90	protein_coding
4	ENSG00000171587	10.81729	2.14E-39	DSCAM	protein_coding
5	ENSG00000095627	10.43166	6.81E-45	TDRD1	protein_coding
6	ENSG00000152977	10.33537	1.19E-17	ZIC1	protein_coding
7	ENSG00000180155	10.18176	7.41E-13	LYNX1	protein_coding
8	ENSG00000075213	9.785136	2.12E-20	SEMA3A	protein_coding
9	ENSG00000183036	9.326026	1.26E-10	PCP4	protein_coding
10	ENSG00000196208	9.240923	0	GREB1	protein_coding

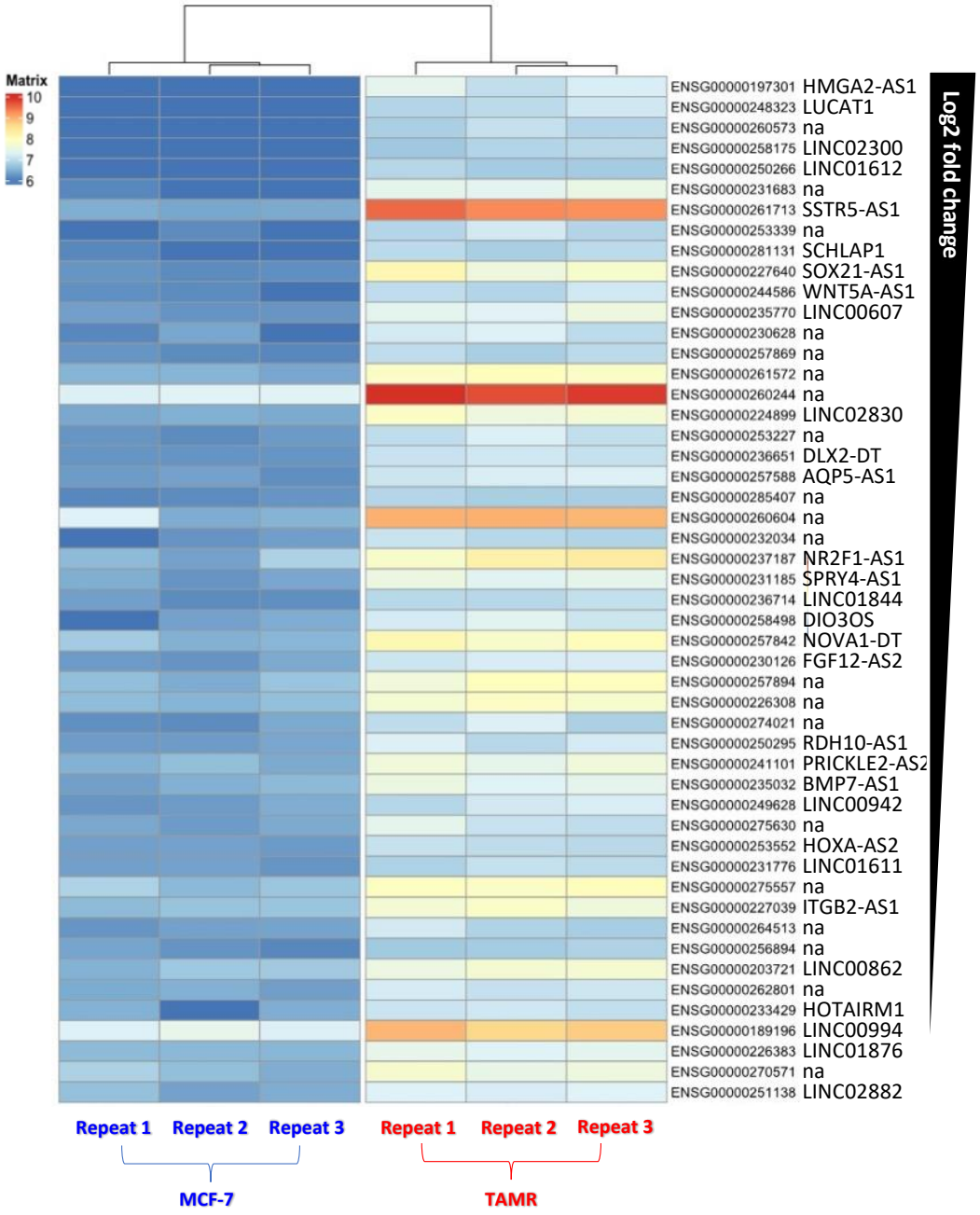
Tables 3.7 and 3.8 Up and down regulated lncRNA genes.

Ranking	ensembl_gene_id	log2FoldChange	padj	gene_name	gene_biotype
1	ENSG00000197301	-8.85136	1.02E-09	HMGA2-AS1	lncRNA
2	ENSG00000248323	-8.17352	2.09E-08	LUCAT1	lncRNA
3	ENSG00000260573	-7.88631	7.47E-08	Na	lncRNA
4	ENSG00000258175	-7.64873	2.25E-07	LINC02300	lncRNA
5	ENSG00000250266	-7.49781	3.99E-07	LINC01612	lncRNA
6	ENSG00000231683	-7.21264	3.78E-09	Na	lncRNA
7	ENSG00000261713	-6.15492	7.14E-66	SSTR5-AS1	lncRNA
8	ENSG00000253339	-6.0558	1.38E-06	Na	lncRNA
9	ENSG00000281131	-5.83702	3.20E-06	SCHLAP1	lncRNA
10	ENSG00000227640	-5.81193	3.18E-17	SOX21-AS1	lncRNA

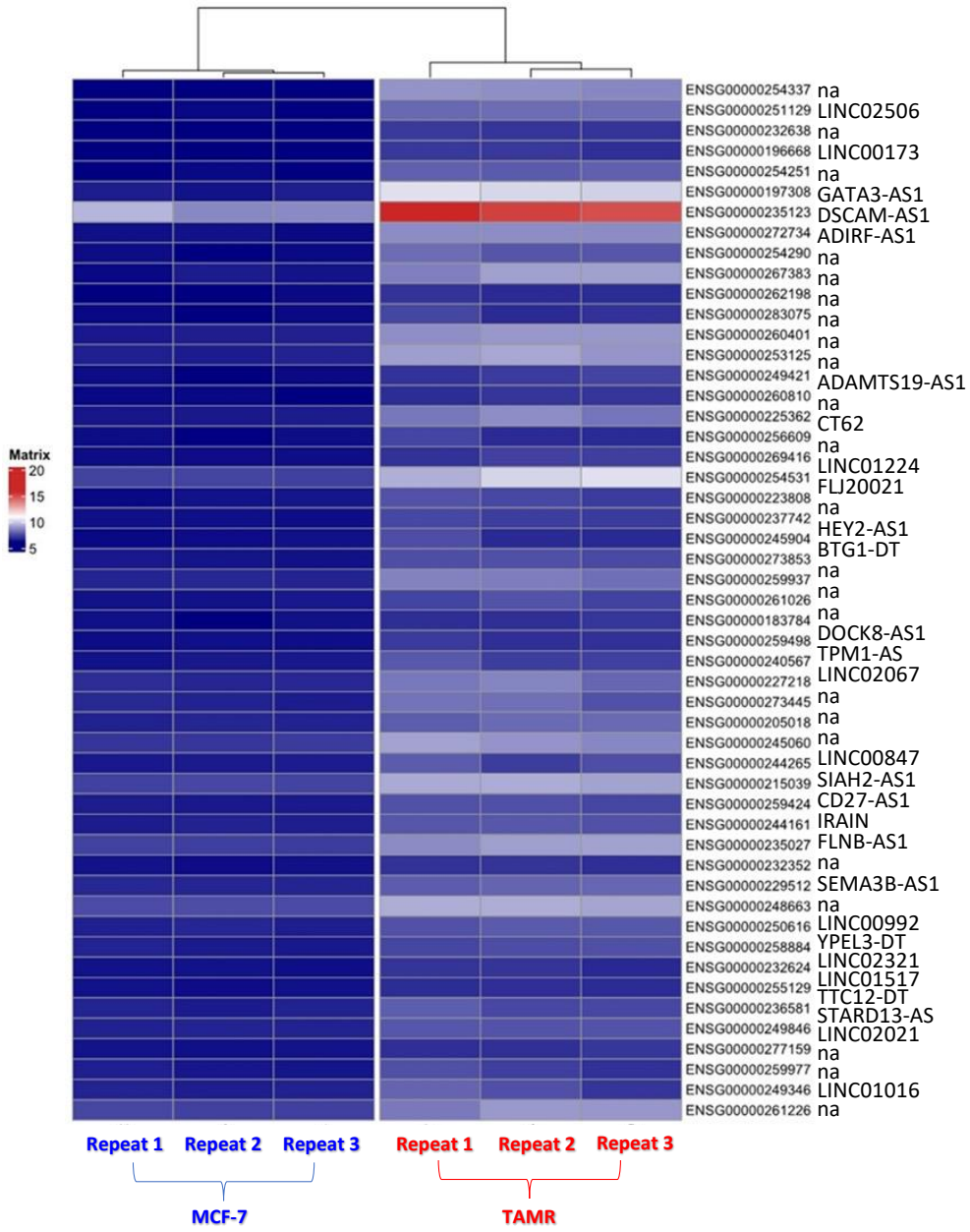
Ranking	ensembl_gene_id	log2FoldChange	padj	gene_name	gene_biotype
1	ENSG00000254337	11.73365	1.11E-17	Na	lncRNA
2	ENSG00000251129	8.335559	5.92E-12	LINC02506	lncRNA
3	ENSG00000232638	8.07619	1.86E-08	Na	lncRNA
4	ENSG00000196668	8.008662	3.17E-08	LINC00173	lncRNA
5	ENSG00000254251	7.897944	7.69E-11	Na	lncRNA
6	ENSG00000197308	6.932705	1.60E-88	GATA3-AS1	lncRNA
7	ENSG00000235123	6.735506	3.08E-90	DSCAM-AS1	lncRNA
8	ENSG00000272734	6.657618	4.10E-34	ADIRF-AS1	lncRNA
9	ENSG00000254290	6.525568	1.02E-12	Na	lncRNA
10	ENSG00000267383	6.110617	1.23E-33	Na	lncRNA

3.2.4.6 Most dysregulated genes following all analysis – top 50s.

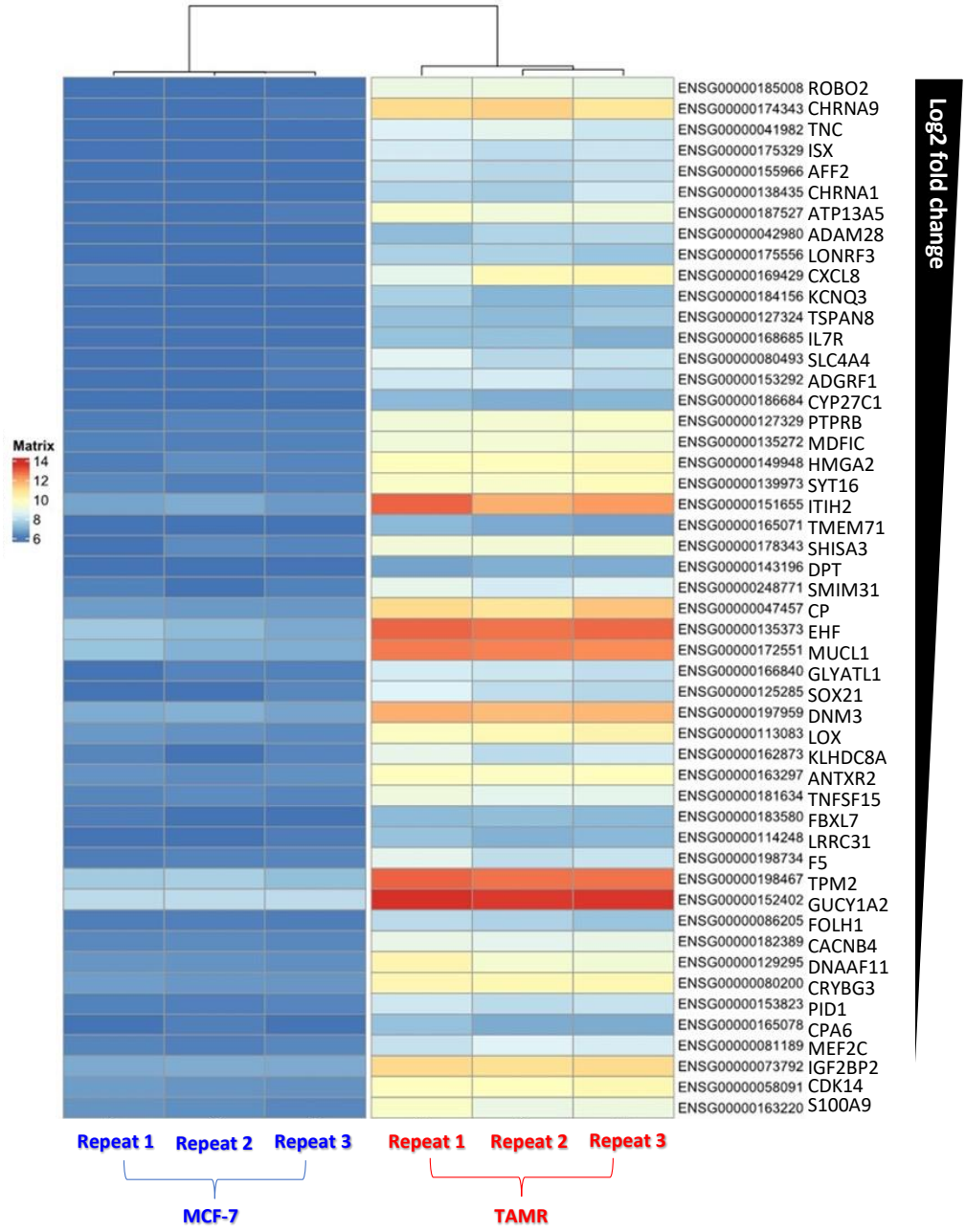
Expression discrepancies in lncRNAs and protein coding genes were screened visually using heatmap representation; to gather insights into RNA-seq large dimensionality, addressing gene expression in each sample individually, where expression matrix were displayed in a range of colours scaled to expression intensity. We opted to turn off default DEseq2 automatic clustering; to present the descending ranking of top 50 genes based on the value of fold change. However, clustered heatmaps are available in (Figures 1, 2 ,3 and in appendix), showing genes re-ordered based on degree of similarity in expression profile, creating groups of genes that might have biological associations. Four heat maps were produced, top 50 upregulated lncRNA genes, top 50 upregulated protein coding genes, top 50 downregulated lncRNA genes and top 50 downregulated protein coding genes in TAMR samples (heat maps 3.1, 3.2, 3.3, 3.4). The process of scanning through colour-hue coded gene expressions of samples separately and relative to each other indicated the degree of expression variability and hints at gene-cluster wise functionality.



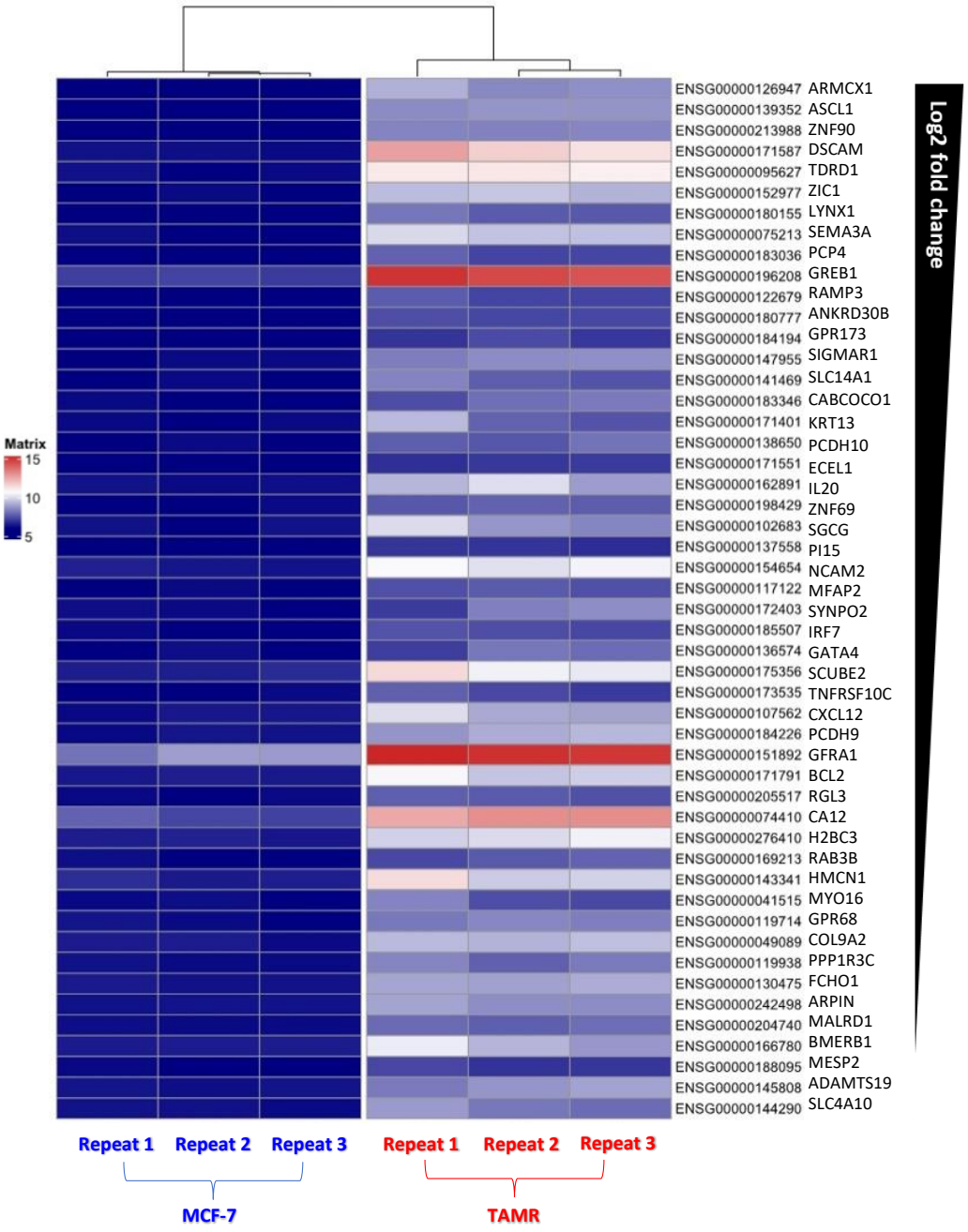
Heat map 3.1 of top 50 lncRNAs up-regulated in TAMR samples. From the DEA result table, we chose 50 differentially expressed lncRNAs with the highest expression in TAMR compared to MCF-7 samples. lncRNAs were ranked in fold change decreasing order. In the grid heat map, coloured tiles are scaled to the strength of gene expression, as seen in the colour key (Matrix), ranging from dark red (higher expression) to dark blue (lower expression) . Rows correspond to genes annotated by ensemble IDs and gene names (na indicated missing annotation), and columns correspond to samples (Three MCF-7 biological repeats and three TAMR biological repeats).



Heat map 3.2 of top 50 lncRNAs downregulated in TAMR samples. From DEA result table, we chose 50 differentially expressed lncRNAs with the highest expression in TAMR compared to MCF-7 samples. lncRNAs were ranked in fold change decreasing order. In the grid heat map, coloured tiles are scaled to the strength of gene expression, as seen in the colour key (Matrix), ranging from dark red (higher expression) to dark blue (lower expression) . Rows correspond to genes annotated by ensemble IDs and gene names (na indicated missing annotation), and columns correspond to samples (Three MCF-7 biological repeats and three TAMR biological repeats).



Heat map 3.3 of top 50 protein coding genes up-regulated in TAMR samples. From DEA result table, we chose 50 differentially expressed protein coding genes with the highest expression in TAMR compared to MCF-7 samples. protein coding genes were ranked in fold change decreasing order. In the grid heat map, coloured tiles are scaled to the strength of gene expression, as seen in the colour key (Matrix), ranging from dark red (higher expression) to dark blue (lower expression) . Rows correspond to genes annotated by ensemble IDs and gene names (na indicated missing annotation), and columns correspond to samples (Three MCF-7 biological repeats and three TAMR biological repeats).



Heat map 3.4 of top 50 protein coding genes up-regulated in TAMR samples. From DEA result table, we chose 50 differentially expressed protein coding genes with the highest expression in TAMR compared to MCF-7 samples. protein coding genes were ranked in fold change decreasing order. In the grid heat map, coloured tiles are scaled to the strength of gene expression, as seen in the colour key (Matrix), ranging from dark red (higher expression) to dark blue (lower expression) . Rows correspond to genes annotated by ensemble IDs and gene names (na indicated missing annotation), and columns correspond to samples (Three MCF-7 biological repeats and three TAMR biological repeats).

3.2.5 Mapping Interactions across gene populations

lncRNA-miRNA-mRNA correlation is an important mechanism implicating the function of these lncRNAs. To identify the relationship between nominated lncRNAs and miRNAs, miRcode database was used to find the target relationship between a candidate lncRNA and different miRNA families, so basically the output is a list of microRNAs that are conserved in the genomic location of these lncRNAs of interest. Next, TargetScan database was applied (using miRcode results), to find all the protein coding genes targets of these miRNAs, for this, the Aggregated PCT score of 0.5 was considered, defined as the probability of conserved targeting, which is the estimation score of the probability that a site is conserved due to the miRNA targeting rather than by chance. So, for example a score of 0.99 means the site has a 1 % chance of being conserved by chance. Shown in figures 3.11, 3.12. 3.13 and 3.14. The interaction networks were built around significantly differentially expressed lncRNAs as central nodes, connected through microRNAs to protein coding genes enriched in carcinogenesis and oestrogen signaling-related gene ontology gene families (Table 3.10)

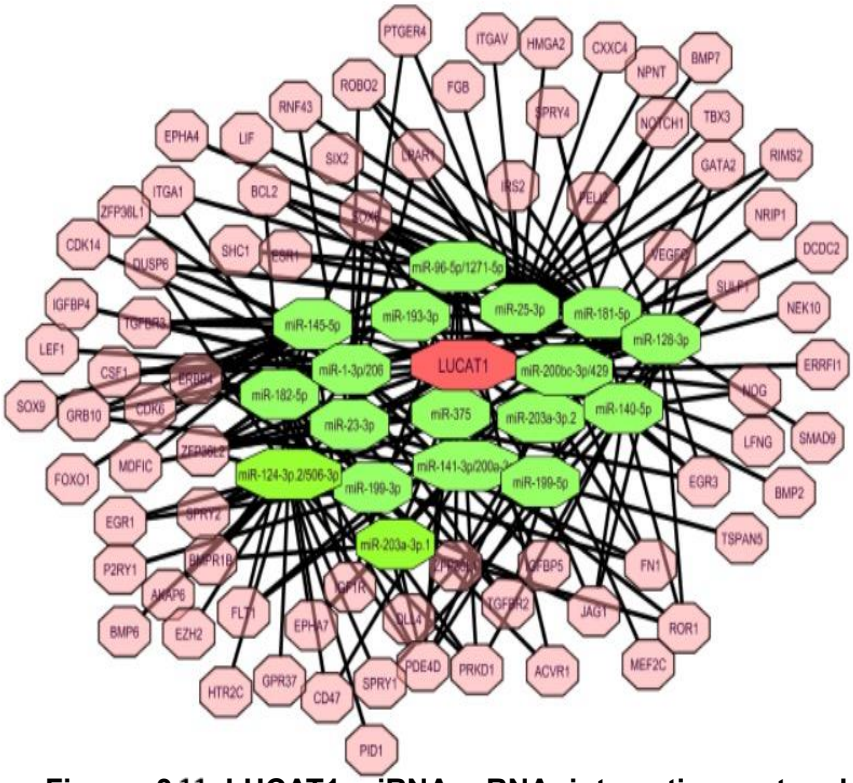


Figure 3.11 LUCAT1-miRNA-mRNA interaction network. The figure shows relationships between coexpression modules. Candidate lncRNA (red), protein coding genes (pink) and miRNAs (green).

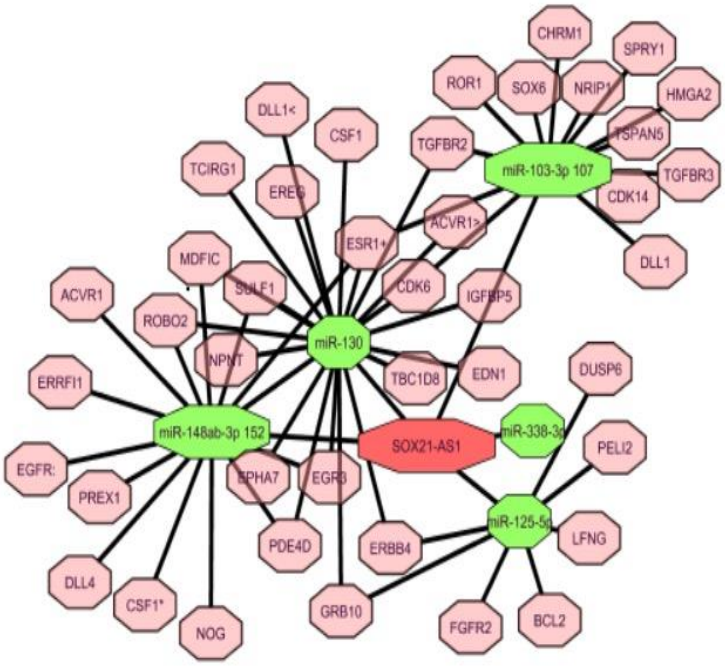


Figure 3.12. SOX21-AS1-miRNA-mRNA interaction network. The figure shows relationships between coexpression modules. Candidate lncRNA (red), protein coding genes (pink) and miRNAs (green).

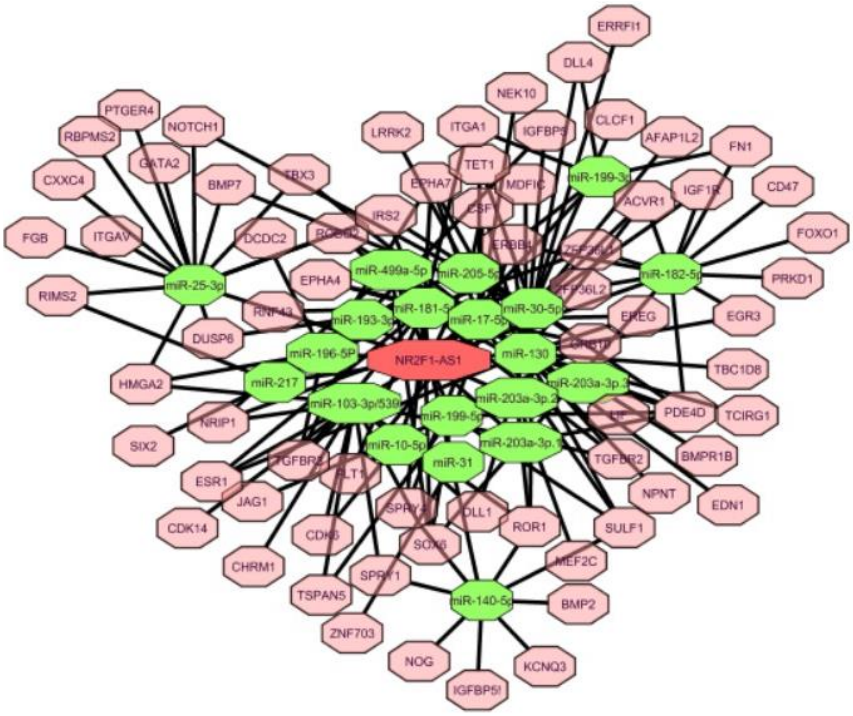


Figure 3.13 NR2F1-AS1-miRNA-mRNA interaction network. The figure shows relationships between coexpression modules. Candidate lncRNA (red), protein coding genes (pink) and miRNAs (green).

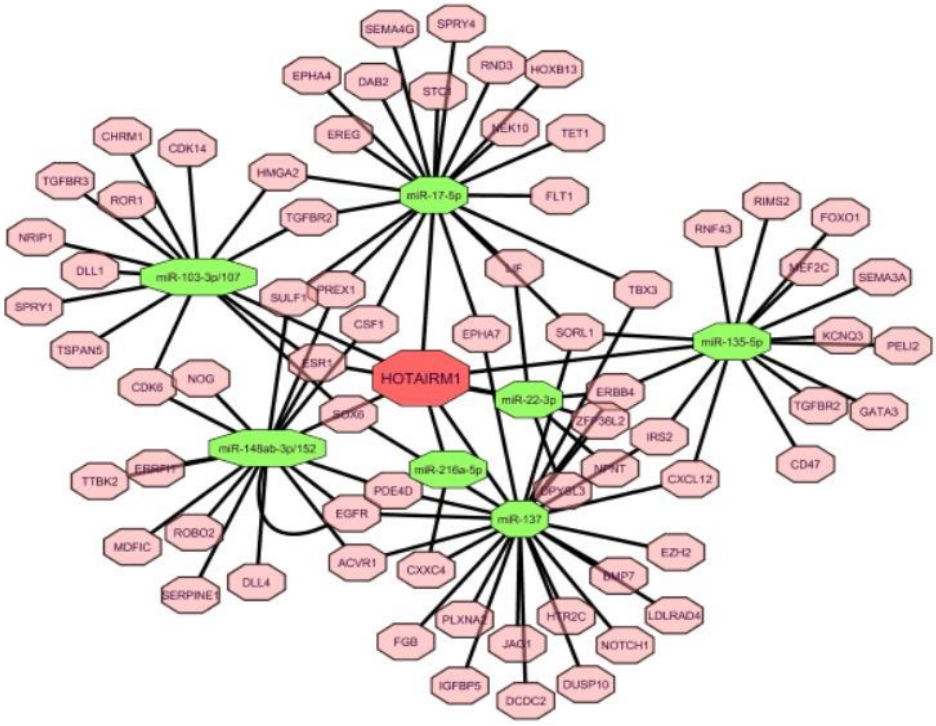


Figure 3.14 HOTAIRM1-miRNA-mRNA interaction network. The figure shows relationships between coexpression modules. Candidate lncRNA (red), protein coding genes (pink) and miRNAs (green).

3.2.6 Gene set enrichment analysis

Gene set enrichment analysis was performed to identify genes involved in overrepresented functional groups. Protein coding genes were used for this; as functional and gene ontology analysis of lncRNAs is still not well developed. So, it was opted here to predict lncRNAs function by looking to protein coding genes, presuming they were affected by dysregulated lncRNAs. DEseq2 normalised matrix of counts were loaded into GSEA software along with tamoxifen related gene set files downloaded from MSigDB molecular signatures database (Debrabant, 2017). 10 out of 29 gene sets were enriched with genes raked to be highly expressed in TAMR samples and 10 were enriched with genes raked to be lower in TAMR samples, (Tables 3.9 and 3.10).

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_5	474	-0.401	-1.810	0	0.001	tags=32%, list=16%, signal=37%
CLIMENT_BREAST_CANCER_COPY_NUMBER_UP	18	-0.655	-1.770	0.005	0.004	tags=50%, list=16%, signal=59%
BHAT_ESR1_TARGETS_VIA_AKT1_DN	80	-0.362	-1.330	0.058	0.201	tags=21%, list=8%, signal=23%
FRASOR_RESPONSE_TO_SERM_OR_FULVESTRANT_UP	20	-0.475	-1.294	0.169	0.191	tags=35%, list=10%, signal=39%
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_DN	84	-0.343	-1.272	0.077	0.177	tags=23%, list=9%, signal=25%
RIGGINS_TAMOXIFEN_RESISTANCE_DN	202	-0.290	-1.209	0.089	0.225	tags=26%, list=16%, signal=30%
RIGGINS_TAMOXIFEN_RESISTANCE_UP	57	-0.340	-1.170	0.213	0.245	tags=23%, list=7%, signal=25%
FRASOR_RESPONSE_TO_ESTRADIOL_DN	72	-0.306	-1.108	0.265	0.317	tags=28%, list=11%, signal=31%
BECKER_TAMOXIFEN_RESISTANCE_UP	45	-0.322	-1.055	0.382	0.380	tags=20%, list=10%, signal=22%
BECKER_TAMOXIFEN_RESISTANCE_DN	46	-0.274	-0.895	0.660	0.695	tags=30%, list=17%, signal=36%

Table 3.9. Gene sets enriched with genes upregulated in TAMR samples

SIZE: number of genes, ES: enrichment score, NES: normalized enrichment score, NOM p-val: nominal p-value, FDR q-val: false discovery rate.

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
MASSARWEH_TAMOXIFEN_RESISTANCE_DN	232	-0.547	-2.404	0	0
MASSARWEH_RESPONSE_TO_ESTRADIOL	59	-0.675	-2.389	0	0
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_4	282	-0.510	-2.316	0	0
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	195	-0.526	-2.296	0	0
FRASOR_RESPONSE_TO_ESTRADIOL_UP	33	-0.672	-2.130	0	0
BHAT_ESR1_TARGETS_VIA_AKT1_UP	261	-0.439	-1.962	0	0.0001
MASRI_RESISTANCE_TO_TAMOXIFEN_AND_AROMATASE_INHIBITORS_UP	16	-0.670	-1.833	0.004	0.0002
FRASOR_RESPONSE_TO_SERM_OR_FULVESTRANT_DN	49	-0.479	-1.673	0.002	0.004
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_2	357	-0.264	-1.222	0.026	0.140
FRASOR_TAMOXIFEN_RESPONSE_UP	51	-0.309	-1.089	0.302	0.279

Table 3.10 Gene sets enriched with genes downregulated in TAMR samples.
SIZE: number of genes, ES: enrichment score, NES: normalized enrichment score, NOM p-val: nominal p-value, FDR q-val: false discovery rate.

GSEA results show that, out of the 10 gene sets upregulated in TAMR, two sets had a significant FDR q-value<0.05. “CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_5” was the top gene set enritched. It includes 474 genes involved with acquired endocrine therapy resistance in ESR1 expressing breast tumours (Creighton *et al.*, 2008). Enrichment score (ES) is the the primary result of GSEA mediated ove-representation analysis, it shows if an assigned gene list positively correlates to TAMR phenotype. A visual representation of ES is shown through an enrichment plot of the Creighton gene set (figure 3.15), that demonstartes the enrichment profile or curve peaking at an ES of 0.401 as it runs through gene list, starting with the ones ranked to be most upregulated in TAMR samples and fading as this correlation decreases. The second gene set enriched was “CLIMENT_BREAST_CANCER_COPY_NUMBER_UP “. This includes 23 genes found to gain DNA copy number abberations in breast cancers (Climent *et al.*, 2007). The ES of the Climent gene set was 0.655. While reflecting a higher

association, it is still ranked inferior to the Creighton gene set. This can be explained when looking into the normalised enrichment score (NES) results. NES is used to compare between gene sets as it accounts for the size of gene set and corrlates it to the expression set, showing the true rank of a functional gene set (Subramanian *et al.*, 2005). The results above show statistically significant enrichnet of Creighton and Climent pathways in TAMR cells, validating that a large number of the expression changes observed are related to the tamoxifen resistance phenotype.

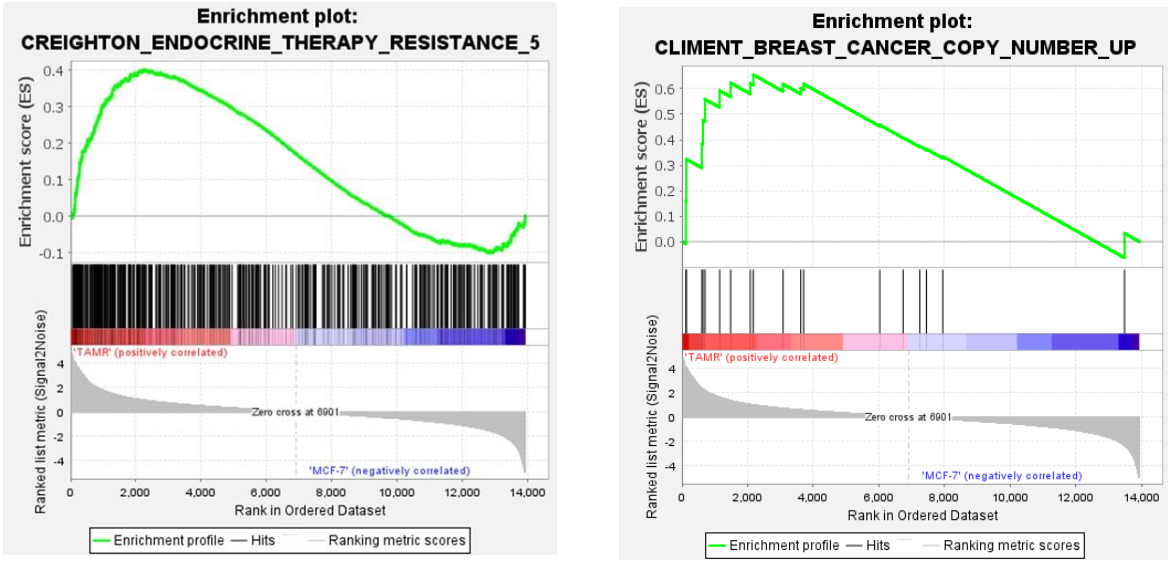


Figure 3.15. GSEA analysis results of top gene sets enriched for TAMR phenotype. RNA-seq expression set file was uploaded in GSEA software and tested against tamoxifen related gene sets. Only gene sets with FDR q-value < 0.05 were considered. (A) Enrichment score plots for Creighton gene signature. Positive enrichment is observed for TAMR group (FDR q-val = 0.001), negative enrichment is observed for MCF-7 group. (B)Enrichment score plots for Climent gene signature. Positive enrichment is observed for TAMR group (FDR q-val = 0.004), negative small enrichment is observed for MCF-7 group. Along the x-axis is RNA-seq genes ranked by their expression from the left (high in TAMR) to right (low in TAMR compared to MCF-7 .y-axis is the enrichmet metrics related to ranked genes per particular pathway or gene set.

3.2.7 Final filtering step to identify 4 genes for further investigation.

Further filtering of lncRNAs was done. The protein coding probability of transcripts was examined and by applying a limit of < 0.3 score a further 50 lncRNAs were excluded. Finally, we decided to exclude any unannotated genes or those lacking Ref.seq data, this was 45 lncRNAs. Thus, at the end of the filtering lncRNAs process, 31 upregulated lncRNAs and 26 downregulated lncRNAs in TAMR samples were identified (Table 1 and 2 appendix).

From this list it was opted to nominate 4 lncRNAs (LUCAT1, SOX21-AS1, HOTAIRM1 and NR2F1-AS1) from TAMR upregulated lncRNA genes group to further investigate them *in-vitro* experimentally; because each of these lncRNAs showed statistically significant high fold change score, low coding probability and acceptable annotation profiles.

3.3 Discussion

In this chapter, the aim was to provide an insight into the genomic variation between tamoxifen sensitive and resistant breast cancers by investigating differential gene expression between matched tamoxifen sensitive and resistant cell lines - the tamoxifen sensitive MCF-7 and an offspring resistant TAMR cell line. Resistance was created by long-term exposure of MCF-7 cells to increasing amounts of tamoxifen until cleared resistant (Knowlden *et al.*, 2003). Tamoxifen sensitivity patterns seen were consistent with the literature, validating the model for subsequent testing (Gu et al., 2016; Fagan et al., 2017; Men et al., 2021). Our cell line model has many advantages, being reproducible, reliable, and feasible, but has many limitations, such as the inadequate or even non-existent interactions between

the cellular and extracellular environments, and the fact that cells have been grown in culture for a long time; and are therefore likely adapted and thus changed compared to patient tissues.

Studies reporting tamoxifen resistance genotype-phenotype correlations have showed countless gene-level dysregulations in key regulators of intracellular signaling pathways that control several cellular processes, including proliferation, apoptosis, and survival (Rondón-Lagos *et al.*, 2016). RNA-seq is a very reliable method for accurately measuring gene expression, making it possible to determine the absolute quantification of RNAs and directly compare results between experiments. However, RNA-seq is a lengthy multistep process, exhibiting various types and degrees of bias. Experimentally, wet-lab processes such as samples (e.g., siRNA-transfection, RNA extraction) and library preparation (library development, fragmentation, and size selection) steps are often error-prone, involving many procedures, each with technical bias, batch effect, and sample loss (Ross *et al.*, 2013). We put this into consideration from the initial steps of bioinformatics analysis by performing quality control data assessment on our data at different levels, on the gene level, variance stabilizing transformation was applied on raw counts to get normalized and variance-stabilized log CPM values of gene expressions, while VST method we used is one of the highly recommended normalization methods, there is no settlement in the literature on the best normalization method to use (Zwiener, Frisch and Binder, 2014; Noel-MacDonnell et al., 2018). On the samples level, sample-sample distance-dependent clustering and PCA were applied. Taken all together, the quality control assessment used

concluded that RNA-seq data are suitable, and declared accepted by published statistical standards, for subsequent analyses.

For differential gene expression analysis, while there are several statistical, algorithmic models available to fit RNA-seq data, it was opted to use Deseq2, which is one of the most potent and popular models, mainly when applied to the matrix of raw counts, designed to account for variations in library size internally and know to yield a higher number of differentially expressed genes (Pan et al., 2021; Tong, 2021; Y. Li et al., 2021). From there, different sequential filtering steps were systematically applied on DEA results to rank genes. Referring to the highly cited article (Conesa *et al.*, 2016), it was taken into consideration the sequenced number of replicates, library sizes, and quality control assessments based on this, adjusted p-value score limit was set to <0.005 (Noble, 2009) ;to construct a proper statistical power analysis. From TAMR upregulated lncRNAs prioritized list, already published breast cancer drug resistance-related genes emerged, for example, LINC00942 (Sun *et al.*, 2020), HOXA-AS2 (Cui et al., 2020), and NKILA (Wu *et al.*, 2018), this is supportive of our approach and of the model. In addition, looking at the overall DEA results, differentially expressed protein coding genes were utilized for GSEA analysis as an indirect was to predict lncRNAs functions as functional analysis is not yet developed and reliable solely for lncRNAs, this revealed significant enrichment for tamoxifen resistance-related genes. Results from the GSEA analysis therefore suggested a strong potential of the input data for creating tamoxifen resistance functional signature. However, unexpectedly some well-characterized lncRNAs such as DSCAM-AS1 (Ma *et al.*, 2019), H19 (Wang et al., 2019), and GATA3-AS1 (Zhang et al., 2020), featured in the downregulated

lncRNAs, these are reported previously as oncogenes and drivers of tamoxifen resistance, implying an opposite functional role in the constructed model. This contradiction highlights a significant potential limitation of RNA-seq being vulnerable to multiple sources of bias both experimentally and statistically (Ross et al., 2013; Costa-Silva, Domingues and Lopes, 2017) or could reflect differences in the models used.

It's been shown that lncRNAs function mainly as regulatory molecules of protein coding genes by acting as a sponge for miRNAs, interrupting their guided negative control of oncogenes (X. Zhang *et al.*, 2019), or their role as inhibitory decoys for tumour suppressors (Singh *et al.*, 2022). Such comprehensive assembly of our data compared to single genes prioritization practice is believed to be more representative of biological function, stable, and accurate (Barter et al., 2014). This is especially true with lncRNAs, as shown in many studies related to tamoxifen resistance (Y. Liu et al., 2019; Feng et al., 2020; Fang et al., 2022).

Summary

In summary, we have conducted a bioinformatics analysis of sequenced tamoxifen resistant and sensitive cell lines. Following the analysis pipeline, our RNA-seq data passed the quality checkpoints at every stage. Indicating the integrity of our DEA results. In addition to lncRNAs being the main focus of our project, differentially expressed protein coding genes were also considered for gene set enrichment analysis and were used to build co-expression networks. The resulted gene lists will be used for subsequent analysis.

Chapter 4. Role of LUCAT1, SOX21-AS1, NR2F1-AS1 and HOTAIRM1 in TAMR cells

4.1 Introduction

In the previous chapter RNA sequencing (RNA seq) was used as a high-throughput in-depth sequencing method to reveal the transcriptomic changes in breast cancer cells under the experimental conditions of interest. Quality assured total RNA was extracted from TAMR and parent MCF-7 cells, RNA sequencing was performed, and in-silico bioinformatics analysis performed, differentially expressed lncRNAs were only included if they had a log2FC of 1.5 and an adjusted p-value >0.005. From the results of this RNA-seq analysis, four genes were selected from the list of lncRNA genes differentially expressed between TAMR and MCF-7 cell lines - LUCAT1, SOX21-AS1, NR2F1-AS1 and HOTAIRM1. Each was highly expressed in tamoxifen resistant TAMR compared to tamoxifen sensitive MCF-7 cells (log2FC of 7.5, 5.9, 3.5 and 2.7 consecutively). Statistically significant differential expression in addition to high fold change in expression suggests a role in tamoxifen resistance.

An introduction to each gene follows.

4.1.1 Candidate lncRNA LUCAT1

The lung cancer-related transcript 1 (LUCAT1) is a lincRNA that has 65 transcripts, the longest is 3072 bp in length (Cunningham *et al.*, 2022). It transcribes from the antisense strand in the genomic location, chromosome 5: 91,054,834-91,314,547, opposite the area in between protein coding genes adhesion G protein-coupled

receptor V1 (ADGRV1) and arrestin domain- containing 3 (ARRDC3). LUCAT1 was first identified as upregulated in bronchial epithelium of cigarette smokers and panel of lung cancer cell lines (Hsien and Journal, 2013). Higher expression of LUCAT1 was found in non-small cell lung cancer tissues (NSCLC), enhancing cisplatin resistance, and was connected to worsened patient survival (Renhua et al., 2016; Shen, Xu and Xu, 2020). Many studies have linked dysregulated LUCAT1 to the initiation and progression of multiple cancers. In colorectal cancer, LUCAT1 was seen to increase cellular proliferation by regulating myc transcription via interaction with RNA-binding protein Nucleolin (NCL) (Wu *et al.*, 2020). In addition, hypoxia has been shown to induce LUCAT1 transcription that in turn stimulates expression of DNA damage-related genes and as a result combats DNA damage inducing chemotherapeutic agents such as Adriamycin and Oxaliplatin (Huan *et al.*, 2020). LUCAT1 was also shown to promote metastasis and poorer survival in ovarian cancer (Yu *et al.*, 2018). Consistent with these findings, it was reported that LUCAT1 depletion in papillary thyroid cancer compromises cell proliferation, cell cycle progression and invasion while inducing apoptosis (Luzón-Toro *et al.*, 2019). Similar effects on carcinogenesis have been observed in pancreatic ductal adenocarcinoma (Nai *et al.*, 2020), clear cell renal cell carcinoma (Wang et al., 2018), cervical cancer (L. Zhang et al., 2019) and prostate cancer (C. Liu *et al.*, 2019).

In breast cancer, high expression of LUCAT1 has been related to advanced clinical stage especially in the triple negative subtype, it has been shown to promote proliferation, invasion, migration and EMT. One of the molecular mechanisms by which LUCAT1 can promote carcinogenicity was via a negative correlation with

micro-RNA miR-7-5p, that acts as a sponge for the oncogene SOX2. Finally, LUCAT1 has been shown to act as a competing endogenous RNA (ceRNA) impairing the regulatory function of miR-5702 (Mou and Wang, 2019; Li et al., 2020).

4.1.2 Candidate lncRNA SOX21-AS1

SRY-box transcription factor 21 antisense RNA 1 (SOX21-AS1) is an intergenic lncRNA that has 7 transcripts, the longest is 3287 bp in length. It transcribes from the forward strand in the genomic location, Chromosome 13: 94,703,454-94,803,430, from within protein coding gene SOX21 (Cunningham *et al.*, 2022). The first reported SOX21-AS1 role in cancer implied it as a tumour repressor in Oral squamous cell carcinoma (Yang *et al.*, 2016). Following studies contradicted this, showing that SOX21-AS1 has tumour promoting roles in a variety of cancers. For example, its high expression in colorectal cancer was correlated to worsen prognosis, and directly related to cancer cellular proliferation, invasion, and migration. In this context, Wei, et al. found that SOX21-AS1 antagonizes miR-145, detaching colorectal cancer related oncogene MYO6 from microRNA mediated regulation. Similarly, in lung adenocarcinoma, SOX21-AS1 was found to have a similar effect on survival and tumorigenesis. Here, depletion of SOX21-AS1 caused a cell cycle arrest in S phase by hindering the expression of cell cycle-dependent kinase inhibitor p57 (Lu *et al.*, 2017). Another demonstrated mechanism of action involves the sponging of miR-24-3p, which was seen to dysregulate the expression of its downstream target oncogene PIM2 (Wang et al., 2021).

4.1.3 Candidate lncRNA NR2F1-AS1

Nuclear receptor subfamily 2 group F member 1 antisense RNA 1 (NR2F1-AS1) is an intergenic lncRNA that has 75 transcripts, the longest is 8117 bp in length. It transcribes from the antisense strand in the genomic location, Chromosome 5: 93,360,779-93,585,649, opposite NR2F1 gene (Cunningham *et al.*, 2022). The first study to report a role for NR2F1-AS1 in cancer corelated high NR2F1-AS1 expression to increased tumorigenicity in hepatocellular carcinoma (Huang *et al.*, 2018). In addition, NR2F1 has been shown to predispose to oxaliplatin resistance via interfering with miR-363's inhibitory function on the ABCC1 oncogene. This function is supported by (Ji *et al.*, 2021; Management, 2021; Xu *et al.*, 2021), who linked high levels of NR2F1-AS1 in hepatocellular carcinoma to advanced stage disease and poor prognosis. In contrast, low NR2F1-AS1 expression has been associated to worse prognosis both in colorectal and cervical squamous cell cancers (Peng et al., 2020; Wang et al., 2020).

4.1.4 Candidate lnc RNA HOTAIRM1

HOX antisense intergenic RNA myeloid 1 (HOTAIRM1) is a linc RNA, with 6 transcripts the longest one is 889 bp in length, it transcribes antisense to the HOXA gene cluster from in between HOXA1 and HOXA2, from the genomic location: chromosome 7: 27,095,647-27,100,265 (Cunningham *et al.*, 2022). Unlike most lncRNAs HOTAIRM1 is highly conserved between species (Gardner *et al.*, 2015). The HOXA gene cluster is an active transcription site for many protein coding and noncoding RNAs associated with cell proliferation, cell cycle progression and apoptosis, and a considerable number of these genes are altered in several malignancies (Wei et al., 2016).

HOTAIRM1 was first described as being involved in granulocytic differentiation under the influence of retinoid acid (Zhang et al., 2009). Further studies have indicated the significance of this lncRNA as a prognostic indicator of many types of leukaemia, colorectal cancer and glioma (Díaz-Beyá et al., 2014, Chen et al., 2015, Wan et al., 2016). It has also been found to be highly upregulated in pancreatic adenocarcinoma and ovarian cancer cells (Zhou et al., 2016, Yang et al., 2017).

While the exact molecular functions of HOTAIRM1 have not been fully addressed, it has been seen to interact with several molecules such as microRNAs (e.g., miR-3960 and miR-196b) where it acts as a sponge, restraining microRNAs from mRNA silencing or as a precursor for small RNAs (Diaz-Beya et al., 2015), and proteins (e.g., PML-RARA oncoprotein and PRC2) to regulate the expression of many genes (Chen et al., 2017). In addition, HOTAIRM1 has been demonstrated to regulate the transcription of neighbouring HOXA genes in a temporal co-linear manner through chromatin reorganization (Wang and Dostie, 2017). Hence more comprehensive functional characterisation of HOTAIRM1 can be achieved by focusing on a) the genes closely related to HOTAIRM1 in term of physical location, b) microRNAs and proteins known to play an intermediate role between HOTAIRM1 and genes liked to resistance pathways.

4.1.5 Hypothesis and Aims of this chapter.

The Hypothesis of this chapter is that depleting each candidate lncRNA expression in tamoxifen resistant cell lines will re-sensitize them to tamoxifen.

The aims of this chapter are:

1. To validate candidate lncRNA expression in tamoxifen resistant cell line models.

2. To determine the response of breast cancer cells to tamoxifen after depletion of each lincRNA.

The objectives of this chapter are:

1. To perform siRNA mediated depletion of each of the four candidate lncRNA.

2. To assess the sensitivity of a range of tamoxifen resistant cell lines to tamoxifen treatment after relevant lncRNA expression depletion and investigate if resistance or proliferation can be reversed.

4.2 Results

4.2.1 In-vitro validation of candidate lncRNAs differential expression in MCF7 compared to TAMR cell lines.

Next generation sequencing generates a high output of data that goes through a pipeline of bioinformatics and statistical analyses, increasing the chances of genes false discovery rate. Before proceeding with any functional studies confirmation of differential expression was therefore an essential step. Quantitative PCR (qPCR) is the gold standard method to validate sequencing results. qPCR is superior in-terms of being specific for the gene of interest, it is cheaper and with much less data generated there is less statistical burden so less chances of false discovers errors. Validation is important in two main aspects: most importantly, measuring gene expression in a different cohort of samples establishes biological reproducibility. Also, qPCR validation allows for easy assessment of same-sample replicates that establishes technical reproducibility of the findings.

To validate the RNA-seq data in chapter 3, total RNA was extracted from cells, cDNA was synthesised, and qPCR was performed on a set of three biological

samples each in three technical repeats, using (Figure 4.1 A). The results of qPCR agreed with the RNA-seq results, all lncRNAs were significantly highly expressed in TAMR cells compared to MCF-7 cells across all the biological repeats. LUCAT1 had the highest relative expression in TAMR cells at 27.7 log2 fold increase compared to MCF-7 cells (Figure 4.1 B), followed by SOX2-AS1 - 26.6 log2 fold increase (Figure 4.1 C), NR2F1-AS1 - 14 log2 fold increase (Figure 4.1 D), and HOTAIRM1 by 7.2 log2 fold increase (Figure 4.1 E) . These data therefore confirm RNA-seq findings, not only in-terms of increased expression in TAMR vs MCF-7 but also in the order by which candidate lnRNAs fold change in gene expression rank in relation to each other.

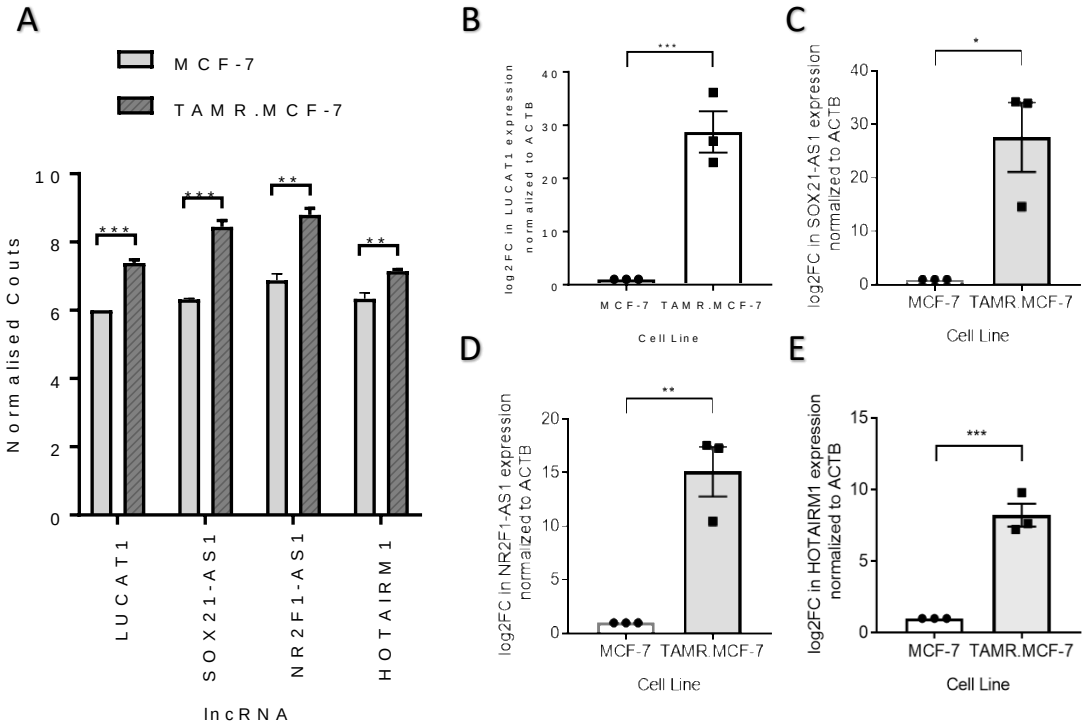


Figure 4.1. High expression of candidate lncRNAs in TAMR compared to parent MCF-7. A. Expression of LUCAT1, SOX21-AS1, NR2F1-AS1 and HOTAIRM1 lncRNAs in RNA-seq analysis. Data represent mean normalised raw counts +/- SD . qPCR confirmation of pattern of expression in TAMR.MCF-7 and MCF-7 cells of B. LUCAT1, C. SOX21-AS1, D. NR2F1-AS1, E. HOTAIRM1 lncRNAs. Each lncRNA expression was normalised to ACTB housekeeping gene expression within the same repeat, each data point represent log2FC in lncRNA expression relative to same gene expression in MCF-7 as control condition. Data points represent mean relative lncRNA expression +/- SD. * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p= ≤0.00, (Student's independent samples unpaired two-tailed t-test).

4.2.2 Expression of candidate lncRNAs in other breast cancer cell lines

Given our hypothesis that high expression of lncRNAs is associated with tamoxifen resistance in breast cancer. Publicly available next generation sequencing data for a large panel of breast cancer cell lines was used to evaluate the expression of each candidate lncRNA and the degree of co-expression of the gene encoding oestrogen receptor alpha (ERα) ESR1. RNA-seq data for breast cancer cell lines was obtained from the Cancer Cell Line Encyclopaedia (CCLE) project.

A GCT file of gene expression data, in this case RNA-seq data of 51 sequenced breast cancer cell lines was downloaded. Gene expression was displayed as reads per kilobase of transcript per million mapped reads (RPKM) values. The matrix file included gene expression values of 56,205 annotated genes, both protein-coding and non-protein-coding. The HMEL cell line was excluded from analysis due to being identified as a non-cancerous. The final list included 32 oestrogen-receptor negative and 18 oestrogen-receptor positive breast cancer cell lines. Values then were log2 transformed in preparation for downstream analyses.

LUCAT1 had a positive RPKMs (log2) value in six cell lines MDA-MB-436, HMC18, HCC1395, HCC1143, MDA-MB-231 and BT20 (figure 4.2. A). Interestingly each of these cell lines was also ER- negative and LUCAT1 was not detected in any of the oestrogen receptor positive cell lines. Correlation analysis between LUCAT1 and ESR1 expression resulted in Pearson's correlation coefficient (r = - 0.394, p-value of 0.004) which implies a significant negative correlation (figure 4.2. B). LUCAT1 mean expression was overall higher in oestrogen receptor negative compared to oestrogen receptor positive cell lines (-1.472 and -2.951 RPKMs (log2) respectively). To confirm the in-silico observed trend of LUCAT1 expression across different breast cancer cell line, RT-qPCR was performed on seven breast cancer cell lines available within the lab sorted into 3 groups: ER-negative (CAL51, MDA-MBA-231 and MDA-MB-468), ER-positive (ZR751 and T47D) and our tamoxifen resistance model (parent MCF-7 and TAMR), that was included in the experiment to see how it relates to other examined cell lines in terms of LUCAT1 expression level. Averaged delta CT values of LUCAT1 normalised to ACTB were plotted (Figure 4.2.C). In contrast to the data from the CCLE, LUCAT1 could be detected

in all cell lines, however except for the MDA-MB-231, MDA-MB-468 and TAMR cell lines, CT values indicated very low expression. MDA-MB-231 and MDA-MB-468 did not show expression in the CCLE data set but this difference could be due to the sensitivity of RNA seq compared to qPCR. MDA-MB-231 and MDA-MB-468 are oestrogen receptor negative. These data are therefore also generally supportive of an association between oestrogen receptor expression and LUCAT1, as both bioinformatics data (CCLE and RNA-seq) and invitro validation shows higher LUCAT1 expression in estrogen receptor negative and tamoxifen resistant phenotypes.

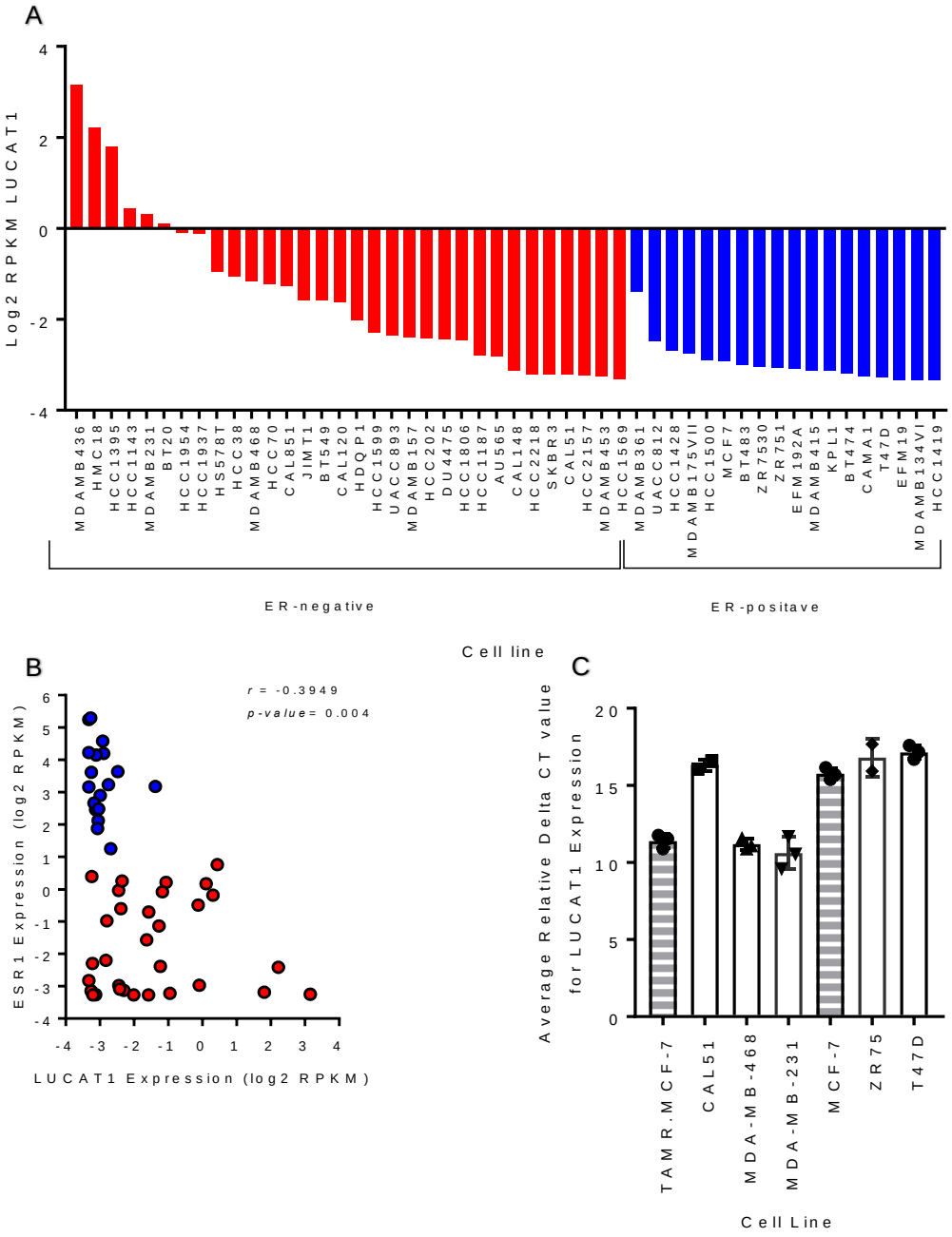


Figure 4.2. LUCAT1 expression in breast cancer cell line CCLE panel
(A) LUCAT1 expression presented in log2 RPKM values in 32 estrogen receptor negative (ER-negative) and 18 estrogen receptor positive (ER-positive) cell lines, RNA-seq data were obtained from CCLE database, cell lines in each group are presented in descending order of LUCAT1 expression. (B) Pearson correlation investigating the relationship between estrogen receptor gene (ESR1) and LUCAT1 expression levels across breast cancer cell lines. Each data point represent a cell line, coloured red if classified estrogen receptor negative and blue if estrogen receptor positive. r is Pearson correlation coefficient and p-value is statistical significance. (C) LUCAT1 expression in the breast cancer cell lines determined by RT-qPCR. Data points represent average relative delta CT values relative to β-actin. Plotted in black is the mean and standard deviation of 3 independent experiments (N=3) each tested in three technical repeats.

The second lncRNA investigated was SOX21-AS1. The same analysis as above was performed. In the CCLE data set, all cell lines had log2 RPKM values lower than 0 regardless of oestrogen receptor status(Figure 4.3 A). Oestrogen receptor negative cells had slightly higher mean expression level of -3.003 vs -3.227 in oestrogen receptor positive by there was no significant correlation between ESR1 and SOX21-AS1. Low SOX21-AS1 expression was confirmed when measured by qPCR in our cell lines (Figure 4.3 C), with high or undetermined delta CT values in all but TAMR cell line. Undetermined delta CT value means the gene is below detection level either due to low amount of cDNA due to not enough genetic material or low expression level, failed reaction due to disrupted thermal cycles, bad quality of SOX21-AS1 primers or sample RNA. The same RNA was used for analysis of all lncRNAs we therefore discounted bad sample RNA, the primers amplified SOX21-AS1 in TAMR we therefore discounted primer quality or design. Three biological replicates each in triplicate was performed and undetermined was seen consistently suggesting the result was not from thermal cycling failure. In addition, we tried adding more cDNA and using a second machine. We therefore conclude that very low levels of SOX21-AS1 are present in our cells in line with previously reported RNA-seq data. This limits our ability to draw conclusions about expression and oestrogen receptor status. However, our findings are contradictory to findings of Liu et al. (2020) who found SOX21-AS1 to be overexpressed in triple negative breast cancer especially in MDA-MB-231 and MDA-MB-468 cell lines. In that study depletion of SOX21-AS1 expression in both cell lines reduced carcinogenic characteristics and promoted apoptotic fate of cells. Correlation analysis between

SOX21-AS1 and ESR1 expression (Figure 4.3 B) was a negative correlation with Pearson's correlation coefficient ($r = -0.169$), though statistically insignificant (p-value>0.05). It is not clear why these differences were seen.

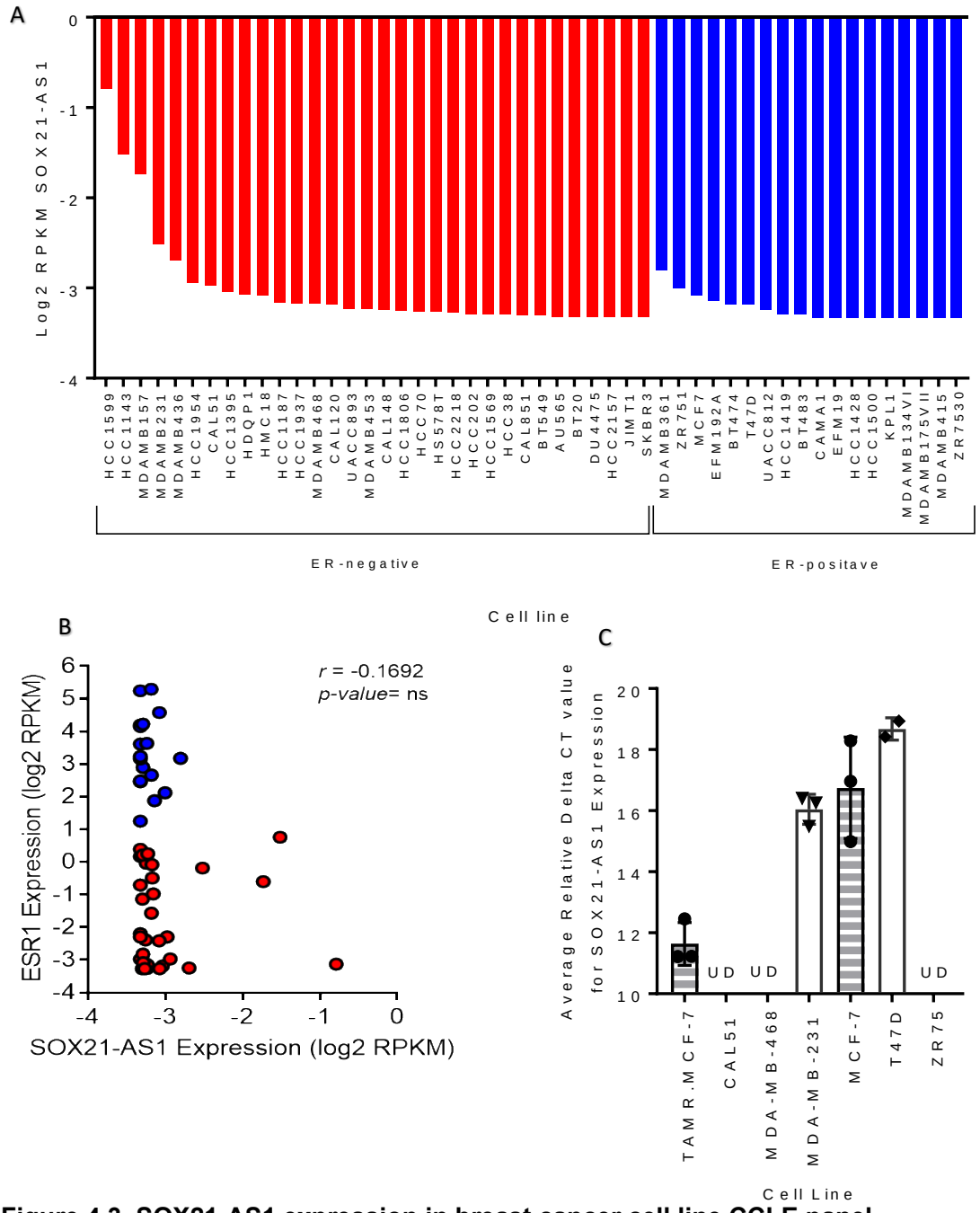


Figure 4.3. SOX21-AS1 expression in breast cancer cell line CCLE panel
(A) SOX21-AS1 expression presented in log2 RPKM values in 32 estrogen receptor negative (ER-negative) and 18 estrogen receptor positive (ER-positive) cell lines, RNA-seq data were obtained from CCLE database, cell lines in each group are presented in descending order of SOX21-AS1 expression. (B) Pearson correlation investigating the relationship between estrogen receptor gene (ESR1) and SOX21-AS1 expression levels across breast cancer cell lines. Each data point represent a cell line, coloured red if classified estrogen receptor negative and blue if estrogen receptor positive. r is Pearson correlation coefficient and p-value is statistical significance. (C) SOX21-AS1 expression in the breast cancer cell lines determined by RT-qPCR. Data points represent average relative delta CT values relative to β-actin. Plotted in black is the mean and standard deviation of 3 independent experiments (N=3) each tested in three technical repeats.

Thirdly NR2F1-AS1 expression was examined. In the CCLE data (Figure 4.4 A), NR2F1-AS1 positive RPKMs (log2) values were reported in eight oestrogen negative cell lines (MDA-MB-453, HCC1937, HS578T, BT20, JIMT1, CAL51, MDA-MB-468 and MDA-MB-231) and one oestrogen receptor positive cell line (HCC1500). Overall mean expression was -1.278 in oestrogen receptor negative vs -2.537 in oestrogen receptor positive cell lines with a trend but no significant correlation (Figure 4.4 B) (Pearson's correlation coefficient r = -0.246, *p-value* >0.05). NR2F1-AS1 expression trend was validated using qPCR in our cell line panel, lncRNA expression was detectable at different levels in all cell lines except in oestrogen receptor positive ZR75 cell line. Highest expression of NR2F1-AS1 was detected in oestrogen receptor negative CAL51 and MDA-MB-231 and our TAMR cell line, with lower levels in the oestrogen receptor positive cell lines, suggesting a correlation with oestrogen receptor status (Figure 4.4 C). The exception to this was MDA-MB-468 which had lower levels in our hands but had a high expression of NR2F1-AS1 in the RNA-seq data. These data do suggest an association between oestrogen receptor status and NR2F1-AS1 although it appears weak.

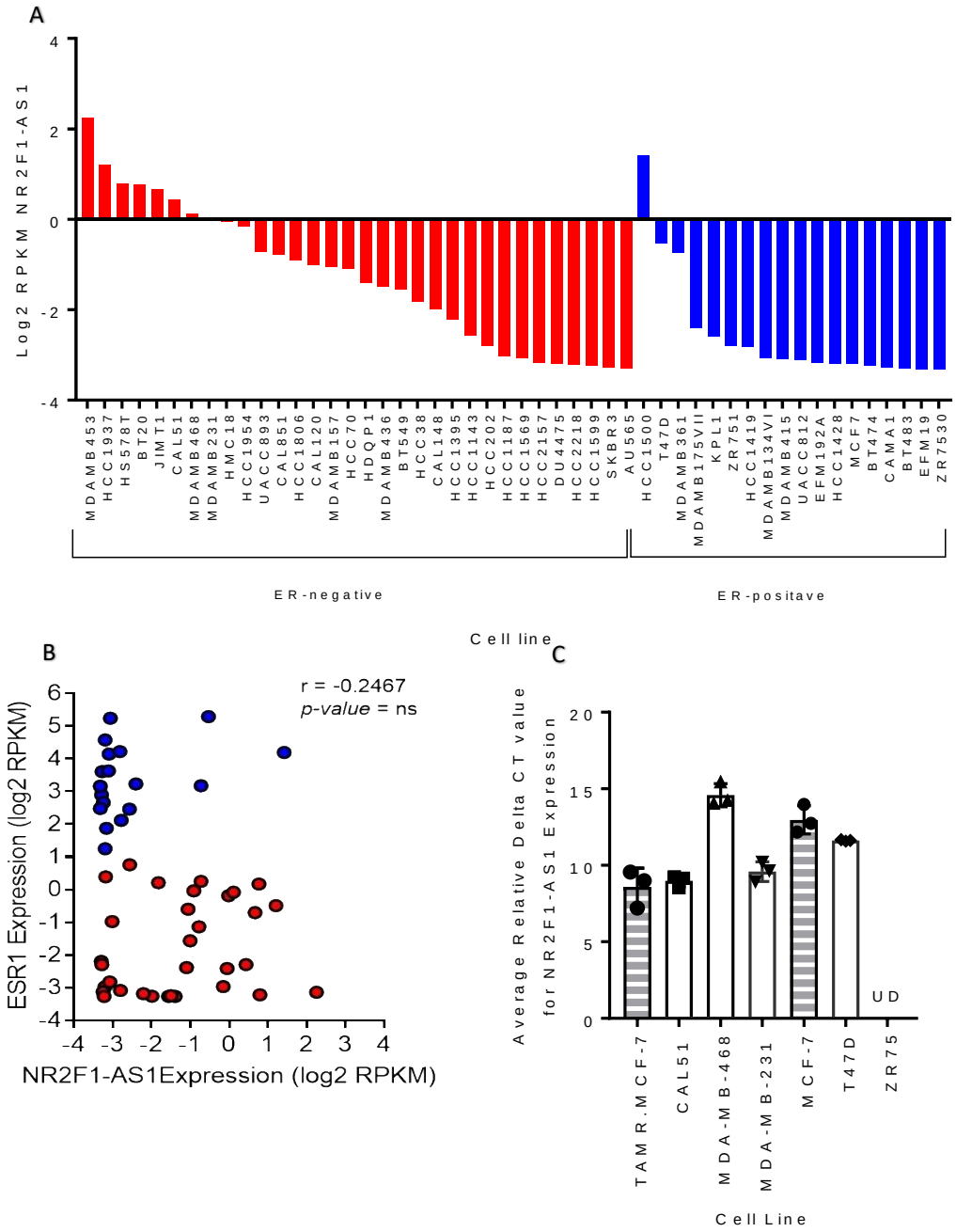


Figure 4.4. NR2F1-AS1 expression in breast cancer cell line CCLE panel
(A) NR2F1-AS1 expression presented in log2 RPKM values in 32 estrogen receptor negative (ER-negative) and 18 estrogen receptor positive (ER-positive) cell lines, RNA-seq data were obtained from CCLE database, cell lines in each group are presented in descending order of NR2F1-AS1 expression. (B) Pearson correlation investigating the relationship between estrogen receptor gene (ESR1) and NR2F1-AS1 expression levels across breast cancer cell lines. Each data point represent a cell line, coloured red if classified estrogen receptor negative and blue if estrogen receptor positive. r is Pearson correlation coefficient and p-value is statistical significance. (C) NR2F1-AS1 expression in the breast cancer cell lines determined by RT-qPCR. Data points represent average relative delta CT values relative to β-actin. Plotted in black is the mean and standard deviation of 3 independent experiments (N=3) each tested in three technical repeats.

Finally, HOTAIRM1 expression was examined. Mean expression in the CCLE was lower in oestrogen receptor positive compared to oestrogen receptor negative cell lines, 0.094 and -2.100 respectively. 16 oestrogen receptor negative cell lines had high expression of HOTAIRM1, and all oestrogen receptor positive cell lines had expression values of less than zero except for MDA-MB-361, MCF7 and KPL1 cell lines (figure 4.5 A). Correlation analysis between HOTAIRM1 and ESR1 expression resulted in Pearson's correlation coefficient (r = - 0.2058) which implies a negative correlation. However, with p-value>0.05 this relationship is not statistically significant (p>0.05) (figure 4.5 B). To confirm the in-silico observed trend of HOTAIRM1 expression across different breast cancer cell line, RT-qPCR was performed *in-vitro* on our panel of 7 cell lines, two oestrogen receptor negative cell lines had the highest HOTAIRM1 expression (CAL51 and MDA-MB-468), even higher than TAMR but the MDA-MB-231 had non-detectable expression In the oestrogen receptor positive cells HOTAIRM1 expression was non-detectable in ZR751 and T47D cells and although HOTAIRM1 expression was detectable in MCF-7 cells, it was significantly low compared to TAMR, CAL51 and MDA-MB-468 cell lines (figure 4.5 C). These data therefore also suggest an association between oestrogen receptor status and HOTAIRM1 expression.

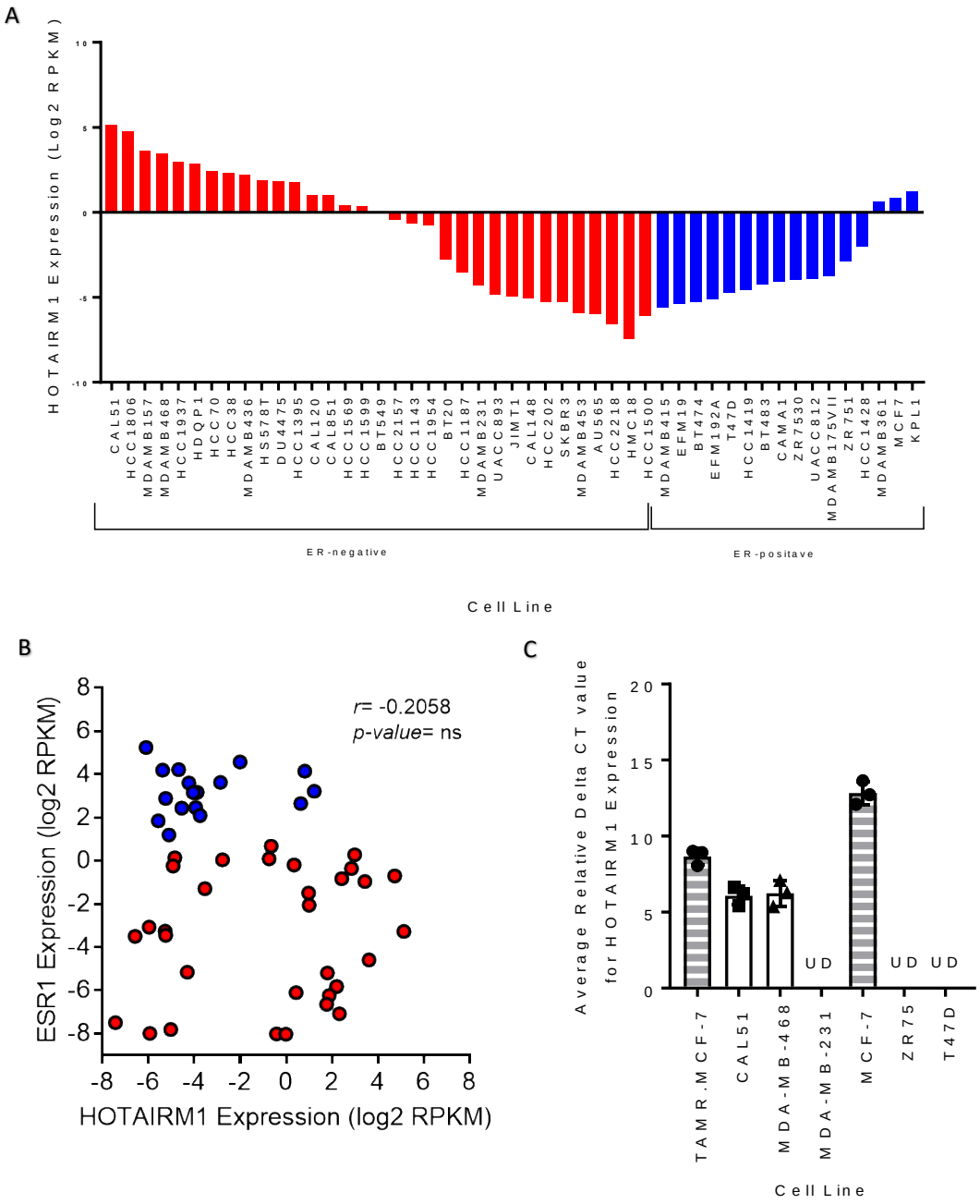


Figure 4.5. HOTAIRM1 expression in breast cancer cell line CCLE panel
(A) HOTAIRM1 expression presented in log2 RPKM values in 32 estrogen receptor negative (ER-negative) and 18 estrogen receptor positive (ER-positive) cell lines, RNA-seq data were obtained from CCLE database, cell lines in each group are presented in descending order of HOTAIRM1 expression. (B) Pearson correlation investigating the relationship between estrogen receptor gene (ESR1) and HOTAIRM1 expression levels across breast cancer cell lines. Each data point represent a cell line, coloured red if classified estrogen receptor negative and blue if estrogen receptor positive. r is Pearson correlation coefficient and p-value is statistical significance. (C) HOTAIRM1 expression in the breast cancer cell lines determined by RT-qPCR. Data points represent average relative delta CT values relative to β-actin. Plotted in black is the mean and standard deviation of 3 independent experiments (N=3) each tested in three technical repeats.

ESR1 expression is a strong predictor of breast cancer response to tamoxifen (Kim *et al.*, 2011). Based on this, ESR1 expression was correlated with each candidate lncRNA expression in sequencing data available in CCLE for 50 breast cancer cell lines. While correlation coefficient always had negative value, indicating lncRNAs have an inverse relationship with ESR1, i.e. the higher lncRNA gene expression, the lower ESR1 expression, this was not always statistically significant and thus association may be weak. Negative trend in association supports the project's main hypothesis that lncRNAs are contributors to tamoxifen resistance. However, *p-values* were nonsignificant for correlation analyses except for LUCAT1. This finding may be explained by the variations in lncRNA expression across many cell lines in RNA-seq data that has other major molecular characteristics such as HER2 status.

4.2.3 Depletion of candidate lncRNAs in tamoxifen resistant MCF-7 (TAMR) cells

One of the most important tools to investigate lncRNAs of interest is using gene-specific siRNAs to interfere with lncRNA expression. siRNA mediated depletion of lncRNAs was chosen due to our special interest in studying lncRNAs upregulated in TAMR cell lines. The double-stranded sequence of a transfection siRNA binds to RNA molecule encoding targeted lncRNA gene to direct it to physical degradation resulting in gene silencing. Based on the results obtained from in-house RNA-seq experiment and cell line investigation (CCLE and qPCR data), our interest in lncRNAs was ordered as follows: LUCAT1, HOTAIRM1, SOX21-AS then NR2F1-AS1.

4.2.3.1 Effect of silencing LUCAT1 expression on tamoxifen sensitivity and proliferation

Two different siRNA sequences targeting different regions of LUCAT1 sequence in the optimised siRNA transfection protocol for MCF-7 cells were used. LUCAT1 siRNAs#1 and #2 and scrambled control siRNA were delivered into TAMR cells using DharmaFECT1 transfection reagent. 48 hours post-transfection, cells were evaluated for cellular toxicity then harvested to confirm siRNA mediated depletion of LUCAT1. Surprisingly, initial experiments showed that both siRNAs increased rather than decreased LUCAT1 expression compared to SC control (Figure 4.6 A). We changed transfection reagent, plating density and incubation period, however, the same upregulated expression pattern was observed (Figure 4.6 B).

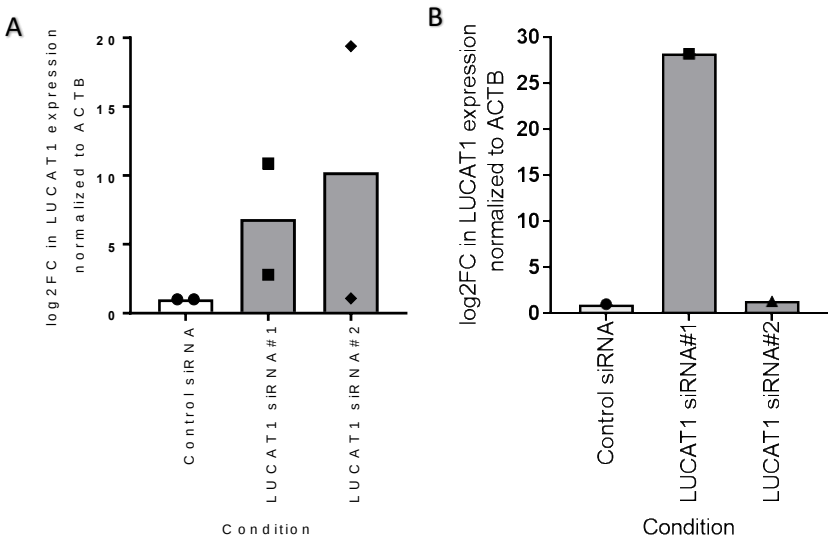


Figure 4.6 . LUCAT1 expression does not change after treatment with siRNA. (A) TAMR cells were transfected with scrambled siRNA Control siRNA), LUCAT1 siRNA#1 AND siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and LUCAT expression was measured by RT-qPCR. LUCAT1 CT values were averaged and normalised to β-actin. For each condition, 2 independents pellets were collected (N=2) and 3 technical replicates were processed. (B) TAMR cells were transfected with scrambled siRNA Control siRNA), LUCAT1 siRNA#1 AND siRNA#2 using transfection reagent Lipofectamine 2000. For each condition, one pellet was collected (N=1) and 3 technical replicates were processed. Data points are log2 fold change (log2FC) in LUCAT1 expression relative to control siRNA treated samples.

4.2.3.2 of silencing SOX21-AS1 expression on tamoxifen sensitivity and proliferation

TAMR cells were transfected with two different siRNA transcripts targeting SOX21-AS1 sequence. Scrambled control siRNA, SOX21-AS1 siRNA #1 and siRNA#2 were transfected into seeded TAMR cells using DharmaFECT1, 48 hours post treatment, cells were collected into a pellet, RNA isolated and cDNA made. Substantial level of expression demotion was achieved with both siRNAs in all repeats (89%-98%) (figure 4.7.A). MTT assay was then used to examine the effect

of SOX21-AS1 depletion on response to tamoxifen, transfected TAMR cells (control siRNA, sRNA#1 and siRNA#2) were treated with increasing doses of tamoxifen for 3-4 days and cell viability was determined relative to vehicle control. Cell viability through all treatment conditions was almost identical, implying SOX21-AS1 depletion, has no effect on TAMR cells sensitivity to tamoxifen (figure 4.7.B). Furthermore, TAMR cell proliferation under the three transfection conditions was examined. We observed that cell viability decreased in cells treated with SOX21-AS1 siRNA#1 by about 17%. However, no statistically significant change in cell viability was observed with SOX21-AS1 siRNA#2 (figure 4.7.C). Considering the comparability in individual cell viability values between both siRNAs, the reduction in proliferation in siRNA#1 rather than siRNA#2 can be explain by nonspecific siRNA targeting or technical variability.

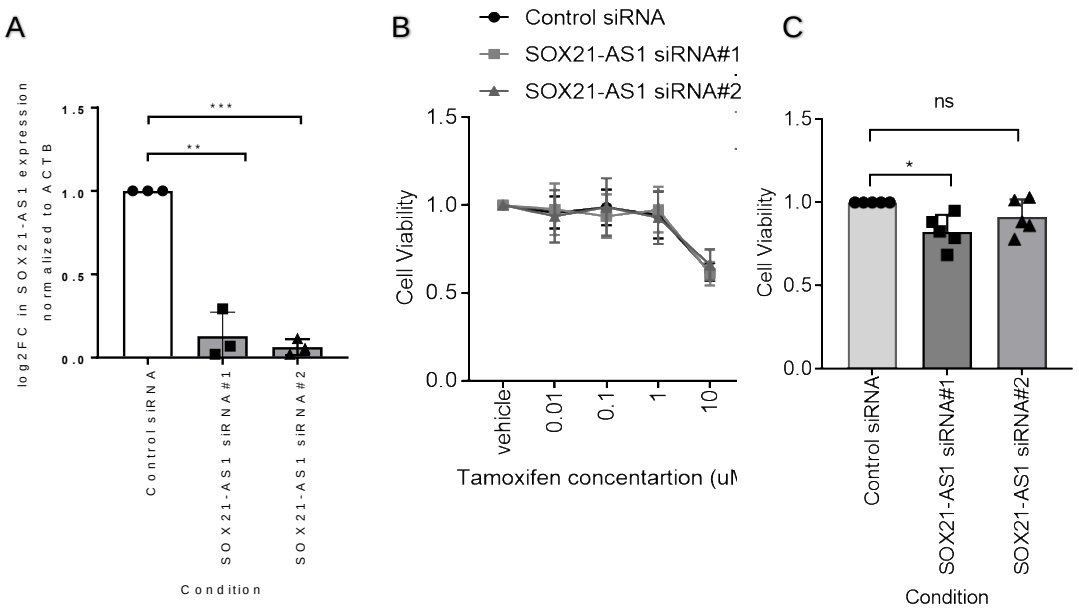


Figure 4.7. SOX21-AS1 depletion does not affect proliferation and Tamoxifen sensitivity in TAMR cells. (A) siRNA attenuated SOX21-AS1 Expression: TAMR cells were transfected with scrambled siRNA (Control siRNA), SOX21-AS1 siRNA#1 AND siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and SOX21-AS1 expression was measured by RT-qPCR, SOX21-AS1 CT values were averaged and normalised to β-actin. For each condition, data points represent log2 fold change of relative SOX21-AS1 expression +/- SD, 3 independents experiemnts were collected (N=3) and 3 technical replicates were processed. (B) TAMR cells transfected with scrambled siRNA (Control siRNA), SOX21-AS1 siRNA#1 and siRNA#2 were tested for sensitivity with different tamoxifen concentrations and vehicle control. Data points represent mean relative cell viability +/- SD, for each condition 3 experiments were performed (N=3) and 5-10 technical replicates were processed. (C) TAMR cells transfected with scrambled siRNA (Control siRNA), SOX21-AS1 siRNA#1 and siRNA#2 were tested for SOX21-AS1 depletion effect on proliferation. Data points represent mean cell viability +/- SD, for each condition, 5 experiments were performed (N=5) and 5-10 technical replicates were processed * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

4.2.3.3 Effect of silencing NR2F1-AS1 expression on tamoxifen sensitivity and proliferation

Two different siRNA transcripts targeting NR2F1-AS1 sequence were used to transfect TAMR cells. Scrambled control siRNA, NR2F1-AS1 siRNA #1 and siRNA#2 were transfected into seeded TAMR cells using DharmaFECT1, 48 hours post treatment, cells were collected into a pellet, RNA isolated and cDNA made.

NR2F1-AS1 siRNA#1 showed superior depletion efficiency reducing NR2F1-AS1 level by 90.6%, compared to NR2F1-AS1 siRNA#2 which depleted NR2F1-AS1 expression by 67% (figure 4.8.A). MTT assay was used to examine the effect of NR2F1-AS1 depletion on response to tamoxifen, transfected TAMR cells (control siRNA, sRNA#1 and siRNA#2) were treated with increasing doses of tamoxifen for 3-4 days, cell viability was determined relative to vehicle control. Cell viability through all treatment conditions was almost identical implying NR2F1-AS1 expression depletion has no effect on TAMR cell sensitivity to tamoxifen (figure 4.7.B). Furthermore, TAMR cell proliferation under the three transfection conditions was examined. We observed that cell viability decreased with cells treated with NR2F1-AS1 siRNA#1 by 20.1% and NR2F1-AS1 siRNA#2 by 22.1% (figure 4.8.C). This supports a pro-proliferative role for NRSF1-AS1 in breast cancer.

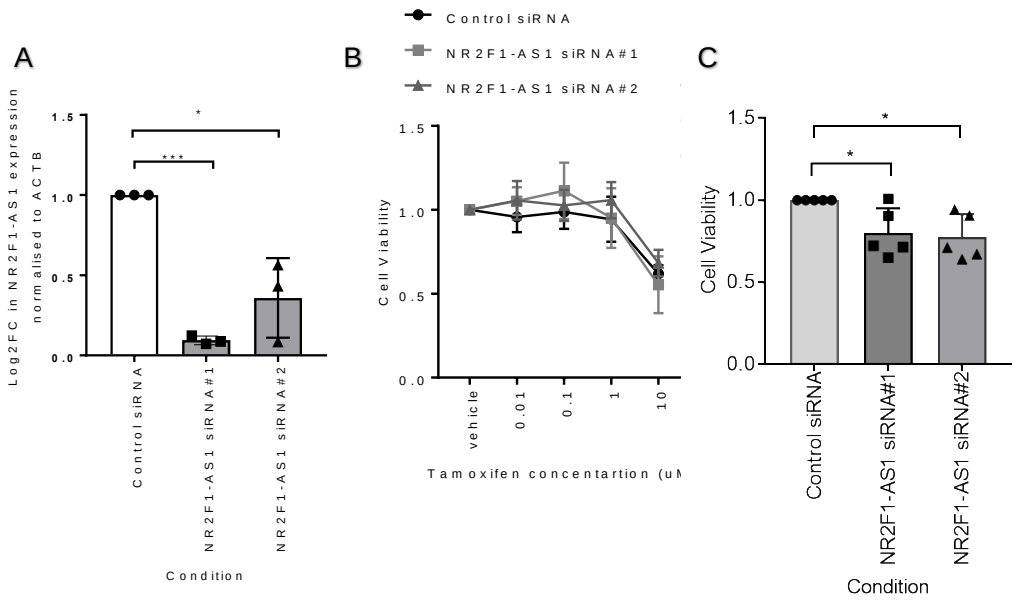


Figure 4.8. NR2F1-AS1 depletion does not affect proliferation and Tamoxifen sensitivity in TAMR cells. (A) siRNA attenuated NR2F1-AS1 Expression: TAMR cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 AND siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and NR2F1-AS1 expression was measured by RT-qPCR, NR2F1-AS1 CT values were averaged and normalised to β-actin. for each condition, data points represent log2 fold change of relative NR2F1-AS1 expression +/- SD, 3 independent experiments were performed (N=3) and 3 technical replicates were processed, . (B) TAMR cells transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 were tested for sensitivity with different tamoxifen concentrations and vehicle control. data points represent mean relative cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed. (C) TAMR cells transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 were tested for NR2F1-AS1 depletion effect on proliferation. data points represent mean cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

4.2.3.4 Effect of silencing HOTAIRM1 expression on tamoxifen sensitivity and proliferation

To prepare the cells for transfection, TAMR cells were seeded and allowed to attach overnight to 6-well plates, scrambled control siRNA, HOTAIRM1 siRNA#1 and siRNA#2 were transfected into TAMR cells using DharmaFECT1, 48 hours post

treatment, cells were collected into a pellet, RNA isolated and cDNA synthesised. Both siRNAs tested showed substantial degree of HOTAIRM1 depletion diminishing expression by 96.4% and 96.2% respectively compared to scrambles control siRNA (figure 4.9.A)

To examine HOTAIRM1 depletion on TAMR cells response to tamoxifen, in 96-well plates, TAMR cells were transfected with two HOTAIRM1 siRNAs and scrambles siRNA as a control. 48 hours following transfection, 4-OH tamoxifen was added at increasing concentrations and plates were incubated for three to four days. An MTT assay was then performed to assess whether HOTAIRM1 depletion can have a direct impact on TAMR response to tamoxifen. Cell survival and transfection efficiency was assessed in three independent experiments. There was not a statistically significant increase in TAMR cell sensitivity to tamoxifen following targeted depletion of HOTAIRM1. Indeed, there was even a small increase in resistance with siRNA#1 through this was non-significant (figure 4.9.B).

One of most distinctive traits of cancer cells, is their ability to sustain uncontrolled proliferation, when cells get stuck in intemperate cell growth and division cycle. To assess the role of HOTAIRM1 in promoting cellular proliferation in tamoxifen resistant cells. HOTAIRM1 siRNA#1 and siRNA#2 were exploited to supress HOTAIRM1 expression in TAMR cells. Using cell viability after 3-4 days as a marker of proliferation, compared to scrambled control siRNA, downregulating HOTAIRM1 expression had no effect on proliferation (figure 4.9.C).

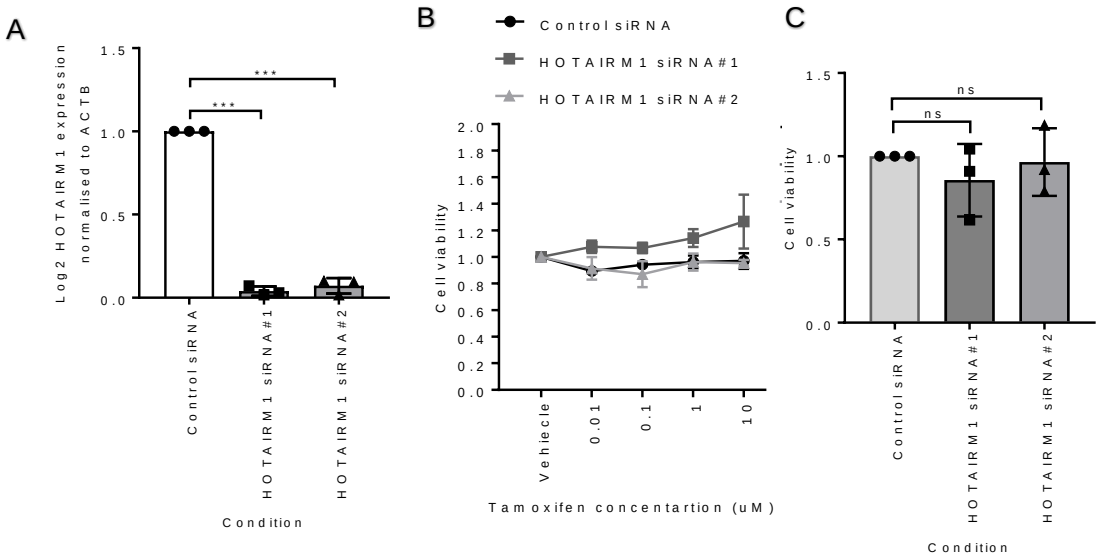


Figure 4.9. HOTAIRM1 depletion does not affect proliferation and Tamoxifen sensitivity in TAMR cells. (A) siRNA attenuated HOTAIRM1 Expression: TAMR cells were transfected with scrambled siRNA (Control siRNA), HOTAIRM1 siRNA#1 AND siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and HOTAIRM1 expression was measured by RT-qPCR, HOTAIRM1 CT values were averaged and normalised to β-actin. for each condition, data points represent log2 fold change of relative HOTAIRM1 expression +/- SD, 3 independent experiments were performed (N=3) and 3 technical replicates were processed. (B) TAMR cells transfected with scrambled siRNA (Control siRNA), HOTAIRM1 siRNA#1 and siRNA#2 were tested for sensitivity with different tamoxifen concentrations and vehicle control. data points represent mean relative cell viability +/- SD, for each condition for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed. (C) TAMR cells transfected with scrambled siRNA (Control siRNA), HOTAIRM1 siRNA#1 and siRNA#2 were tested for HOTAIRM1 depletion effect on proliferation. data points represent mean cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

4.2.4 Effect of silencing NR2F1-AS1 expression on tamoxifen sensitivity and proliferation in triple negative breast cancer cell lines

Triple negative cell lines are also resistant to tamoxifen. According to CCLE data (section 4.2.2), while SOX21-AS1 expression was below zero overall in different breast cancer cell lines, LUCAT1, NR2F1-AS1 and HOTAIRM1 expressions were

the highest in different cellular models of triple negative breast cancer (Figure 4.5). LUCAT1 siRNA mediating expression depletion was ineffective, so we tested the effect of NR2F1-AS1 and HOTAIRM1 silencing on tamoxifen sensitivity in thress triple negative cell lines. HOTAIRM1 silencing results will be presented in the chapter 5.

Tamoxifen resistance in each cell line was first confirmed (Figure 4.10).

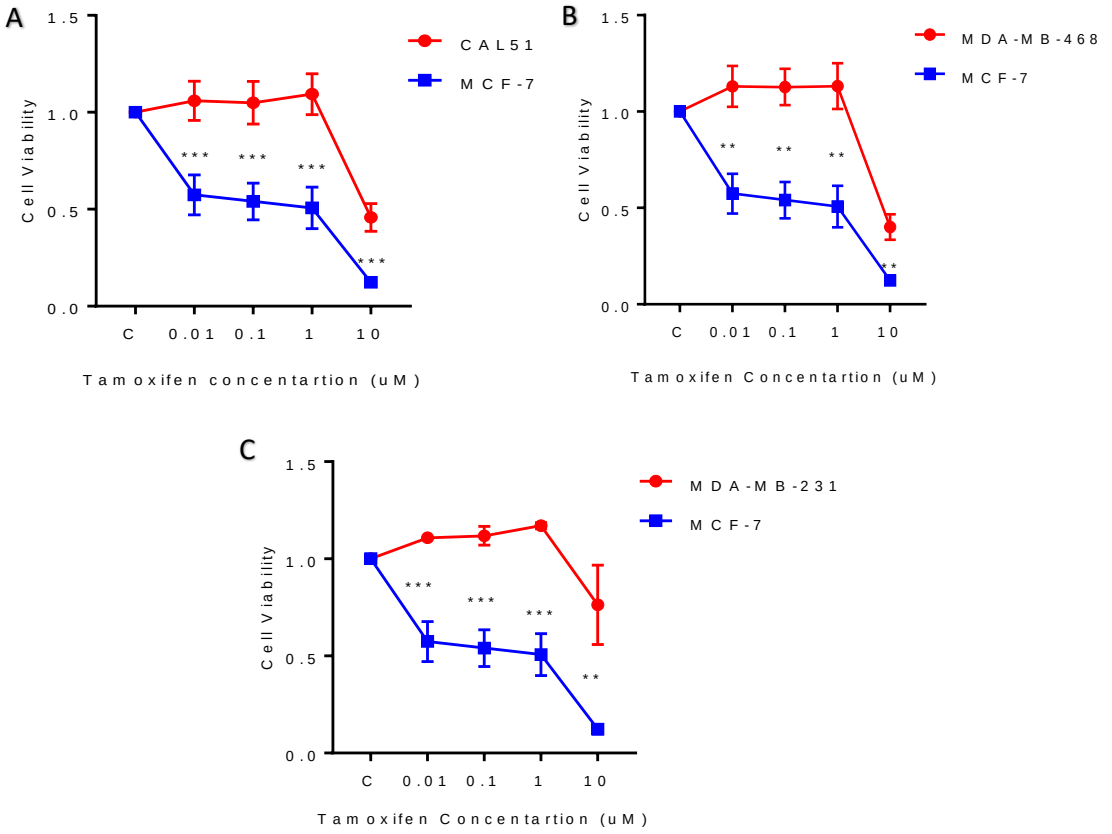


Figure 4.10. Triple negative breast cancer cells sensitivity to tamoxifen. (A) CAL51 cell line, (B)MDA-MB-468 cell line, and (C) MDA-MB-231 cell line. Cells were cultured in 96-well plates, treated with increasing concentrations of tamoxifen (0.01 µm, 0.1 µm, 1 µm and 10 µm) and vehicle control (C). MTT was performed 4 days post treatment, by reading optical densities in the plate reader. Cell viability was calculated by dividing tamoxifen treated well reads by vehicle control read. Data points represent mean cell viability of each treatment group. Error bars depict standard deviation of the mean (N=3). Statistical significance was determined using unpaired one-tailed Student's t-test at each concentration * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p=≤0.001.

Among triple negative breast cancer cell line available in-house, CAL51 cell line had the highest NR2F1-AS1 expression, so it was the first cell line examined. Optimised density of cells was seeded and transfected for 48 hours with three siRNAs, scrambled siRNA as a control, NR2F1-AS1 siRNA#1 and siRNA#2. As shown in (figure 4.11.A), both siRNAs effectively reduced NR2F1-AS1 with siRNA#2 more effective reducing lncRNA expression by 97% compared to 92.7% with NR2F1-AS1 siRNA#1. Progressing to test tamoxifen sensitivity, cells were seeded into a 96-well plate, transfect with siRNA and 48 hours post transfection, tamoxifen was added to the wells in increasing concentrations along with vehicle control. No change in response to tamoxifen was observed (figure 4.11.B). CAL51 cell proliferation was also assessed post transfection Proliferation was slightly reduced in cells treated with siRNA#1 compared to control cells, though this not statistically significant (figure 4.11.C). Thus, depletion of NR2F1-AS1 does not appear to alter tamoxifen sensitivity or baseline proliferation in CAL51 cells.

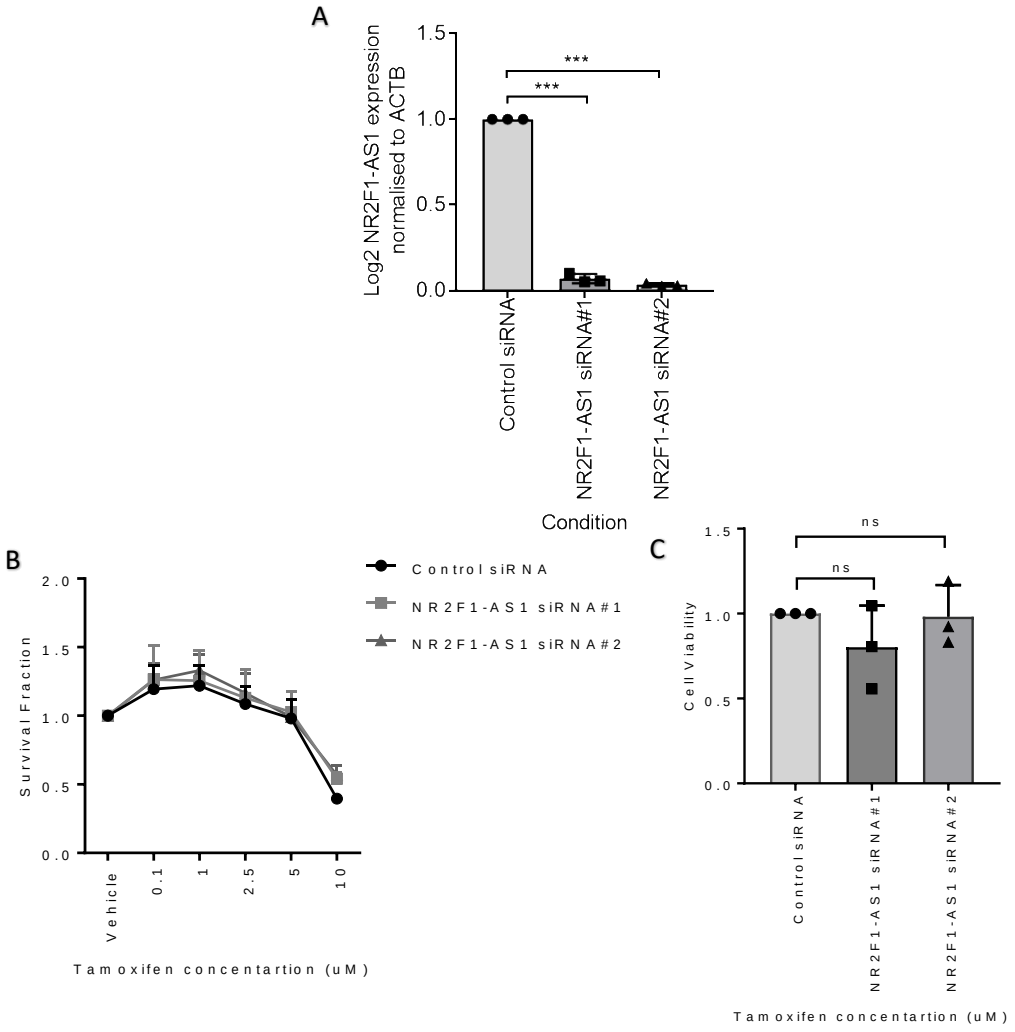


Figure 4.11. NR2F1-AS1 depletion does not affect proliferation and Tamoxifen sensitivity in CAL51 cells. (A) CAL51 cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and NR2F1-AS1 expression was measured by RT-qPCR, NR2F1-AS1 CT values were averaged and normalised to β-actin. for each condition, data points represent log2 fold change of relative NR2F1-AS1 expression +/- SD, 3 independent experiments were performed (N=3) and 3 technical replicates were processed. (B) CAL51 cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2, then were tested for sensitivity with different tamoxifen concentrations against vehicle DMSO control. data points represent mean relative cell viability +/- SD, for each condition for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed. (C) CAL51 cells was transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 were tested for NR2F1-AS1 depletion effect on proliferation. data points represent mean cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed ** denotes p= ≤0.01, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

MDA-MB-468 and MDA-MB-231 cell lines were examined next. Cells were treated as above. RT-qPCR results showed effective NR2F1-AS1 depletion (figure 4.12 and 4.12). However once again no change in tamoxifen sensitivity and very little change in baseline proliferation was observed.

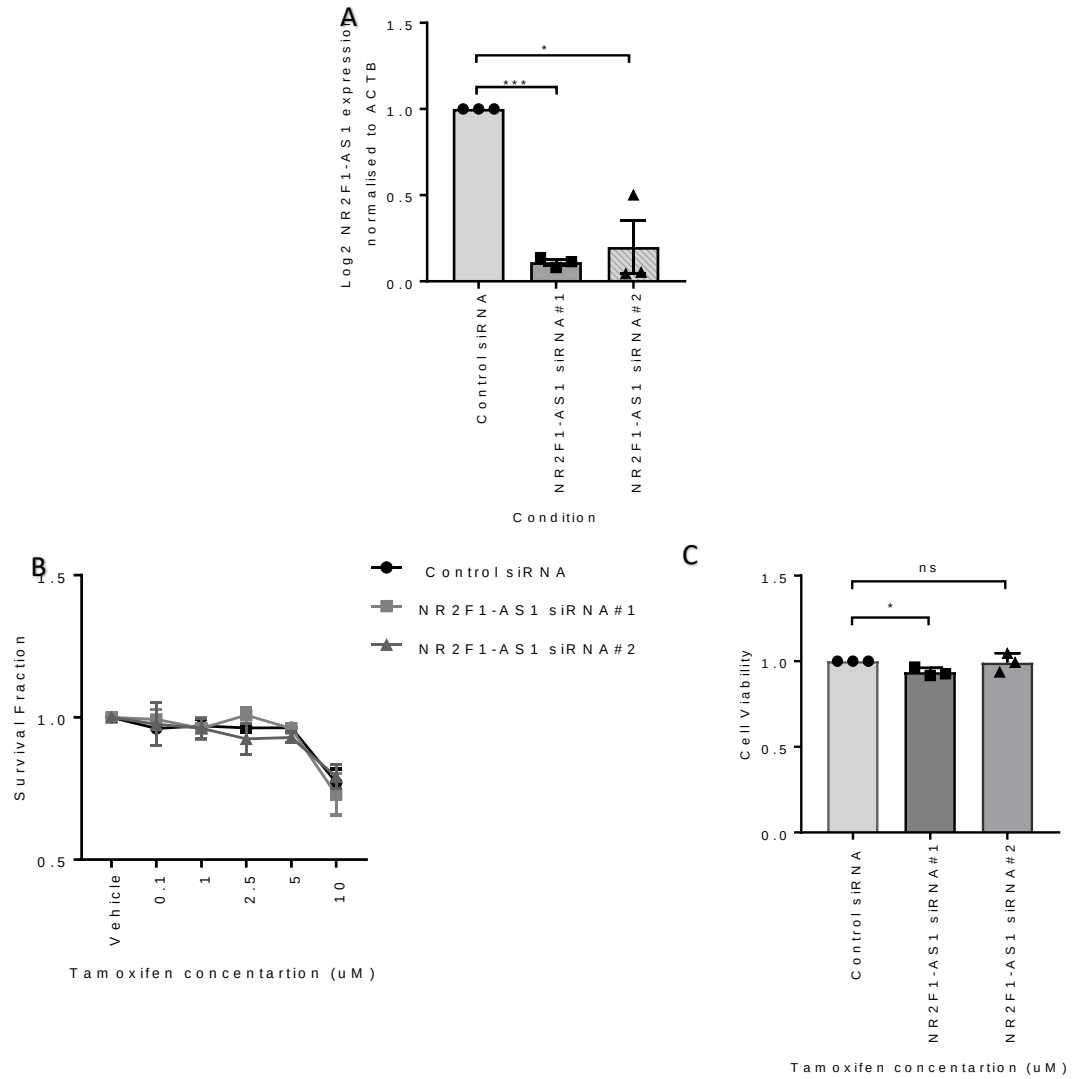


Figure 4.12. NR2F1-AS1 depletion does not affect proliferation and Tamoxifen sensitivity in MDA-MB-468 cells. (A) MDA-MB-468 cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and NR2F1-AS1 expression was measured by RT-qPCR, NR2F1-AS1 CT values were averaged and normalised to β-actin. for each condition, data points represent log2 fold change of relative NR2F1-AS1 expression +/- SD, 3 independent experiments were performed (N=3) and 3 technical replicates were processed. (B) MDA-MB-468 cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2, then were tested for sensitivity with different tamoxifen concentrations against vehicle DMSO control. data points represent mean relative cell viability +/- SD, for each condition for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed. (C) MDA-MB-468 cells transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 were tested for NR2F1-AS1 depletion effect on proliferation. data points represent mean cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed * denotes p= ≤0.05, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

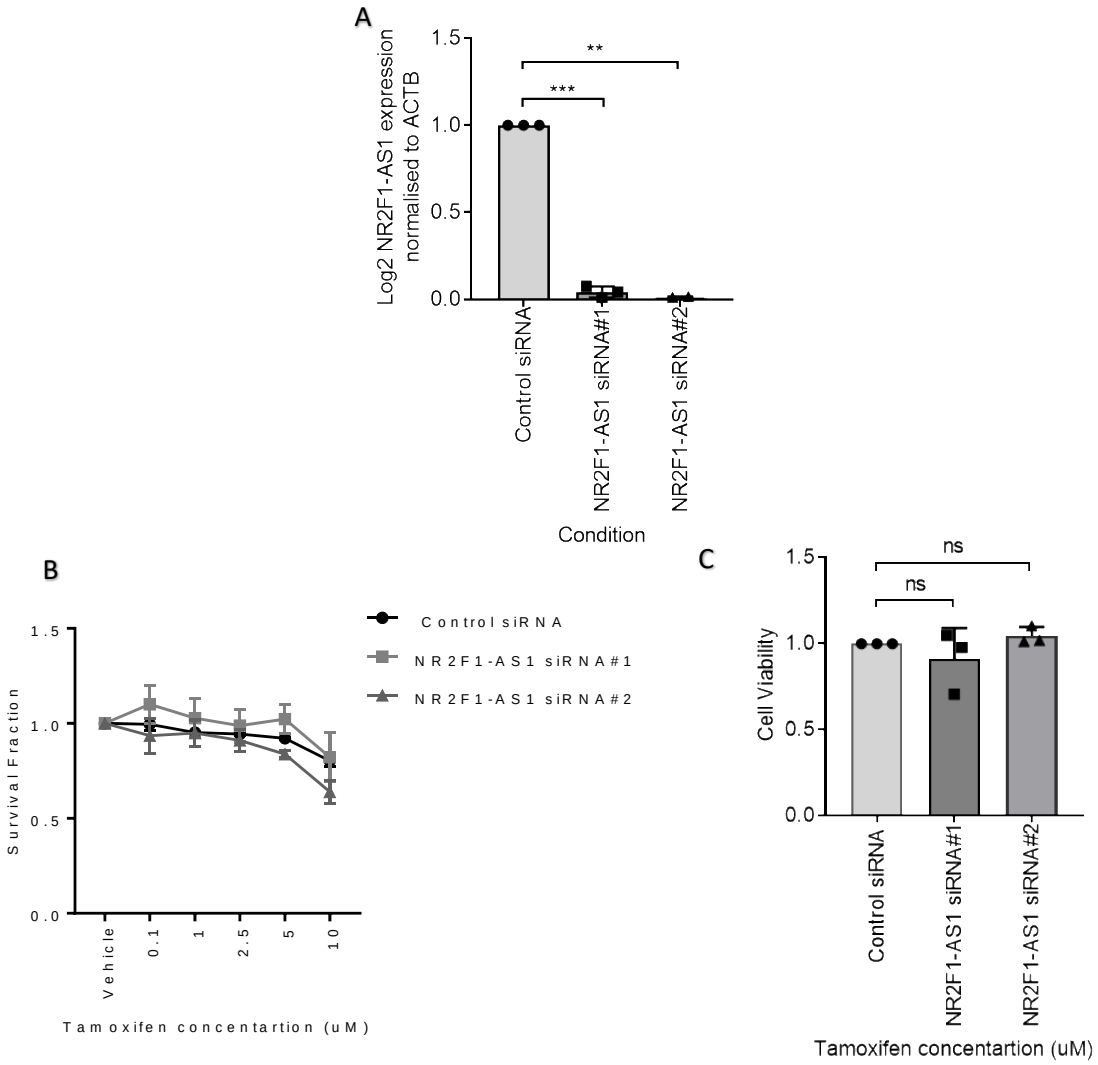


Figure 4.13. NR2F1-AS1 depletion does not affect proliferation and Tamoxifen sensitivity in MDA-MB-231 cells. (A) MDA-MB-231 cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post transfection, cell pellets were collected, RNA isolated, cDNA synthesised and NR2F1-AS1 expression was measured by RT-qPCR, NR2F1-AS1 CT values were averaged and normalised to β-actin. for each condition, data points represent log2 fold change of relative NR2F1-AS1 expression +/- SD, 3 independent experiments were performed (N=3) and 3 technical replicates were processed. (B) MDA-MB-231 cells were transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2, then were tested for sensitivity with different tamoxifen concentrations against vehicle DMSO control. data points represent mean relative cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed. (C) MDA-MB-231 cells transfected with scrambled siRNA (Control siRNA), NR2F1-AS1 siRNA#1 and siRNA#2 were tested for NR2F1-AS1 depletion effect on proliferation. data points represent mean cell viability +/- SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed. ** denotes p= ≤0.01, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

This data together shows no direct cause-effect relationship between NR2F1-AS1 and tamoxifen resistance in acquired or de-novo tamoxifen resistance cell models tested and does not support a major role for NR2F1-AS1 in breast cancer proliferation.

In summary none of the genes identified a high priority for testing as tamoxifen resistance genes in chapter 3 could be validated in chapter 4.

4.3 Discussion

in chapter 3, we nominated four lncRNAs (LUCAT1, SOX21-AS1, NR2F1-AS1, and HOTAIRM1) to further investigate their role in tamoxifen resistance. In this chapter, RT-qPCR results found that all four lncRNAs' expressions were significantly overexpressed in TAMR compared to the parent MCF-7 cell line. In this chapter we validated expression and tested functional significance by depleting each lncRNA. Candidate lncRNAs expression across CCLE breast cancer cell lines RNA-seq data was investigated. All candidate lncRNAs were expressed the highest in ER-negative cell lines, these results were further validated using RT q-PCR in three ER-negative cell lines (CAL51, MDA-MB-468 and CAL51, MDA-MB-468) and ER-positive cell lines (T47D and ZR-75-1). The same trend of expression as seen in RNA-seq was also seen in these results, adding confidence to the conclusions of Chapter 3. Correlation of these lncRNAs CCLE expression to ESR1 expression were not statistically significant (except for LUCAT1), but were always negative, positively adding to the correlation of these lncRNAs upregulation to tamoxifen-resistant phenotype.

LUCAT1 was the first lncRNA investigated, beginning with depleting of expression in TAMR cell line. Surprisingly, LUCAT1 expression rather went up rather than down even when changing the transfection reagent/conditions. This contrasts with the findings of Mou & Wang (2019), where LUCAT1 depletion was successful resulting in suppression of many carcinogenic characteristics (e.g., proliferation, migration, invasion, EMT and dysregulated apoptosis). In contrast to the method we used, (Mou and Wang, 2019) used a DNA plasmid to deliver short hairpin RNA (shRNA) and achieved a good level of LUCAT1 depletion in triple-negative breast cancer cell lines. The increased expression seen in our approach could be explained by potential off-target effects of the siRNAs used. Perhaps rather than targeting the LUCAT1 transcript for degradation, a feedback loop of related coding or noncoding genes was activated (Scacheri *et al.*, 2004). As we did not achieve depletion, we did not continue with analysis of tamoxifen sensitivity.

SOX21-AS1 expression was depleted successfully in the TAMR cell line but no effect was observed on cell proliferation and viability in response to tamoxifen. In breast cancer, SOX21-AS1 depletion were found reduce the carcinogenic properties of breast cancer stem cells (Li, Meng and Wang, 2021) and triple negative breast cancer (Liu et al., 2020b). Both of these studies were carried out primarily on triple negative breast cancer cells, highlighting SOX21-AS1 role as an oncogene in triple negative subtype were supportive of our hypothesis. Findings of Sheng et al. (2020), regarding SOX21-AS1 depletion inhibiting proliferation, while done on MCF-7 that has the same genetic background as our TAMR model, it does

not necessarily mean they respond the same way, this is evidenced by the large number of RNA-seq differentially expressed genes from chapter 3. Moreover, they used a different test to measure cell viability (CCK8) rather than what we used (MTT), which have been shown to produce unconcordant results in-terms of assessing cell viability (Jiao *et al.*, 2015).

NR2F1-AS1 siRNA-mediated depletion was very successful in all cell lines especially MDA-MB-231. Still, no change was observed in sensitivity to tamoxifen in any of the cell lines. A difference was found in TAMR proliferation in response to NR2f1-AS1 depletion similar to results found by Zhong and Zeng (2022). Liu et al. (2021) revealed a cis regulatory link between NR2F1-AS1 and NR2F1 oncogene that interestingly is significantly upregulated by more than two folds in TAMR cell line in our RNA-seq dataset. It would be interesting to further explore the effect of NR2F1-AS1/NR2F1 dysregulation on tamoxifen resistance as the relationship might be reversed (NR2F1 → NR2F1-AS1 rather than what we assumed (NR2F1-F1-AS1 → NR2F1); another possibility is that as shown in chapter 3, a miRNA is an intermediate functional molecule (Huang *et al.*, 2018; Zhong and Zeng, 2022). Whilst we did not observe a significant difference in tamoxifen sensitivity in any tamoxifen resistant cell lines tested after NR2F1-AS1 depletion, we did observe a difference in TAMR cells proliferation but with some variation in cell viability results between repeats. So, further repeats of cell viability post NR2F1-AS1 depletion in TAMR are required to confirm this.

Cell viability in response to tamoxifen treatment, and TAMR cells proliferation did not show any change after HOTAIRM1 depletion. Our findings are completely

opposite to those previously reported (Kim et al., 2020), as they found a direct relationship between HOTAIRM1 expression and tamoxifen resistance. In the previous study, HOTAIRM1 KD was performed using a pool of four different siRNAs rather than using multiple different single siRNAs targeting alternative areas of HOTAIRM1, using separate siRNA reduces the risk of off-target effects (Brown *et al.*, 2022). There is a risk that previously observed effects are through off target effects. So, it would be useful to further investigate HOTAIRM1 depletion in different cell lines.

Summary

The role of the four upregulated candidate lncRNAs in tamoxifen resistance in breast cancer has not been investigated previously and our data highlights this as an area which warrants further investigation. Following successful siRNA mediated depletion of each candidate lncRNA (except for LUCAT1). However, this had no effect on breast cancer cells proliferation or response to tamoxifen. Recently published studies linked genomic dysregulations in these lncRNAs to multiple pathways of tamoxifen resistance. This warrant revisiting of our functional investigation pipeline to explore their role in tamoxifen resistance from different angles.

Chapter 5. HOTAIRM1 Molecular Studies

5.1. Introduction

HOTAIRM1 was one of the four lncRNAs upregulated in TAMR cells, that were selected for the *in-vitro* validation of their role in tamoxifen resistance in breast cancer. Following effective depletion in TAMR cells, resistance to tamoxifen did not change as shown in chapter 4. However, HOTAIRM1 was the most consistently upregulated gene and was also upregulated in more in ER negative compared to ER positive tumour cell lines. It has been reported that HOTAIRM1 acts as an oncogene in a wide range of cancer processes (Zhao *et al.*, 2020). However, not many studies have considered the role of HOTAIRM1 in breast cancer so we decided to see if HOTAIRM1 depletion in cell lines with high levels of HOTAIRM1 would have any cancer related phenotypic effects separate to tamoxifen resistance. Further, the homeobox cluster A (HOXA) is a group of highly conserved 11 protein-coding genes located in chromosome 7, many reported to regulate gene expression and cellular differentiation, morphogenesis, and proliferation (Bhatlekar, Fields and Boman, 2014). HOTAIRM1 is known to transcribe from within the HOXA gene cluster (Wei *et al.*, 2016), but as yet, there is limited evidence linking HOTAIRM1 to HOXA genes or elucidating their mutual regulatory relationship, so we decided to investigate this as well.

The hypothesis of this chapter is:

Manipulation of HOTAIRM1 expression in breast cancer cell lines (with high levels of HOTAIRM1) will alter carcinogenic features and growth characteristics.

The aim of this chapter is

To evaluate different carcinogenic properties of HOTAIRM1 in TAMR and CAL51 cell lines

The objectives of this chapter are:

1. Further validation of HOTAIRM1 depletion

2. Determine the effect of HOTAIRM1 depletion on proliferation in CAL51 and MDA-MB-468 cell lines

3. Determine the effect of increased HOTAIRM1 expression on tamoxifen sensitivity in MCF-7 cell line.

4. Determine effect of HOTAIRM1 depletion on cell cycle progression in TAMR and CAL51 cell lines.

5. Determine effect of HOTAIRM1 depletion on EMT, endogenous DNA damage, and HOXA genes expression in TAMR and CAL51 cell lines

6. Determine the effect of HOTAIRM1 depletion on global differential gene expression in CAL51 cell line.

5.2 Results

5.2.1. Confirmation of HOTAIRM1 depletion

Depletion efficiency differs according to the cell line, transfection reagent used, and the sequence of the siRNA used. Three commercially available HOTAIRM1

targeting siRNAs were compared. The standard Bryant lab protocol for siRNA transfection of the parental MCF-7 cell line was used (methods). Cells were left untreated, treated with the transfection reagent alone, transfection reagent plus a control siRNA (which is reported not to target any known sequence in the genome), or transfection reagent with one of the three HOTAIRM1 targeting siRNAs. In each case, HOTAIRM1 expression was determined by RT-PCR.

Comparing HOTAIRM1 expression in non-treated or transfection reagent alone treated cells, Untreated TAMR cells showed a small but non-significant increase in gene expression. Transfection with each of the HOTAIRM1 targeting siRNAs reduced expression to 80%, 84%, and 75% of the scrambled siRNA control treated sample, for HOTAIRM1 siRNA#1, HOTAIRM1 siRNA#2, and HOTAIRM1 siRNA#3 respectively (Figure 5.1). Toxicity was assessed by eye and appeared to be very low under all conditions (approximately 15-20% cell death). Subsequent experiments used 500,000 cells as transfection efficacy was not altered but a greater number of cells could be obtained (data not shown).

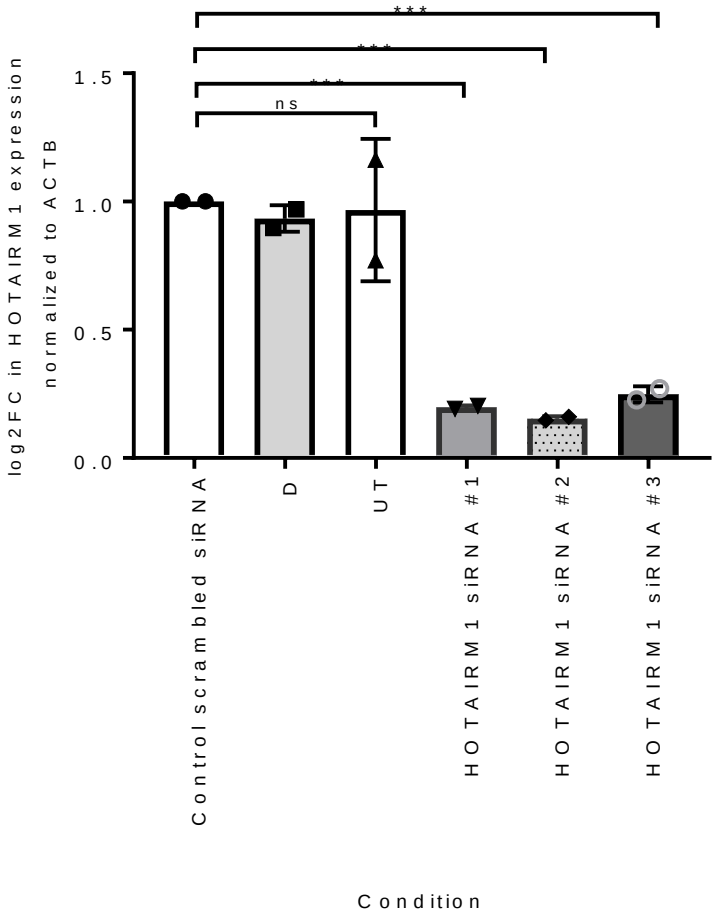


Figure 5.1. HOTAIRM1 depletion in TAMR cells following transfection with siRNA. HOTAIRM1 expression as log2 fold change of scrambled siRNA control was determined in cells untreated (UT), DharmaFECT 1 only treated (D), control non-targeting scrambled siRNA treated (Control siRNA), or treated with HOTAIRM1 targeting siRNAs: HOTAIRM1 siRNA#1, HOTAIRM1 siRNA#2, HOTAIRM1 siRNA#3. HOTAIRM1 expression was determined by qRT-PCR relative to ACTB reference gene 48 h post-transfection. The mean ± SD of two independent repeats is shown. Knockdown data are expressed relative to data from cells transfected with scrambled siRNA control. denotes p= >0.05, * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p=≤0.001.

5.2.2 Duration of HOAIRM1 silencing post-transfection

Given the short life cycle of siRNAs, they usually lose their effect on gene expression in a few days following transfection (Grimm, 2009). The persistence of siRNA efficiency was investigated, to ensure the knockdown stayed effective during subsequent experiments that might take place over up to five days.

The two most effective HOTAIRM1 siRNAs (siRNA#1, and siRNA#2) and scrambled control siRNA were transfected in to TAMR cells under optimised conditions (5 nM siRNA, 4 µL DharmaFECT1, and 500,000 cells per well). Then harvested at 24 h, 48 h, 72 h, 96 h, and 120 h post-transfection (Figure 5.2). Maximal knockdown was reached by 48h, when HOTAIRM1 expression was reduced by 90% and 95% in siRNA#1 and siRNA#2 transfected cells respectively and remained low for up to 120 hours.

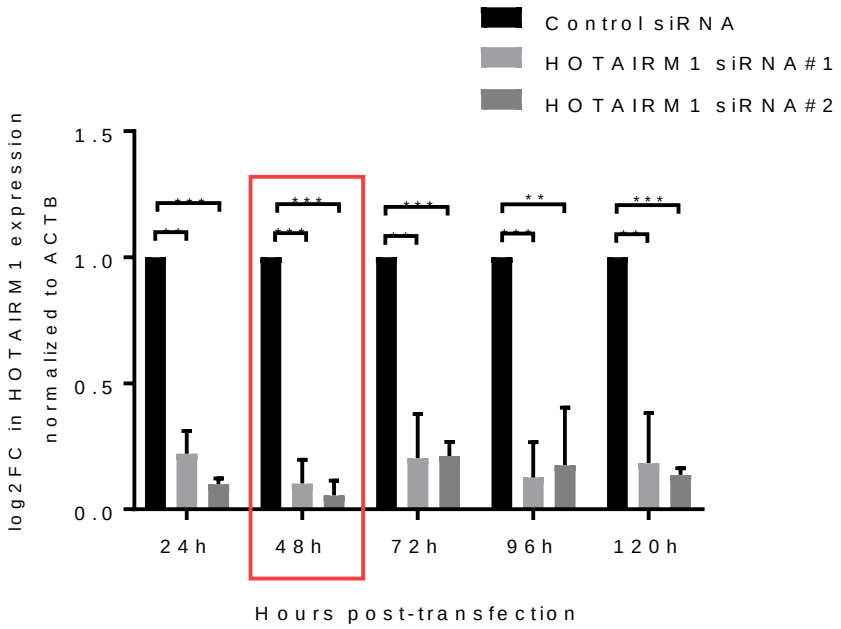


Figure 5.2. Duration of HOTAIRM1 silencing in TAMR cells after a single siRNA transfection. Cells were transfected with two different HOTAIRM1 siRNAs at standard concentrations. Cells were lysed 24 to 120 hours post-transfection. RT-qPCR was performed to detect the level of HOTAIRM1 expression at 24h, 48h, 72h, 96h, and 120h. The mean ± SD of three independent repeats is shown. Knockdown data are expressed relative to data from cells transfected with scrambled siRNA control. denotes p= >0.05, * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p=≤0.001

5.2.3 Effect of silencing HOTAIRM1 expression on tamoxifen sensitivity and proliferation in triple negative breast cancer cell lines

A distinctive characteristic of cancer cells is an increased rate of proliferation. Depletion of HOTAIRM1 has already been shown to have no effect on proliferation of TAMR cells (Figure 4.9). Here the effect of depletion of HOTAIRM1 was examined in 2 other cell lines that were seen to have high levels of HOTAIRM1 i.e. CAL51 and MDA-MB-468 cells. In addition, we decided to check the sensitivity to tamoxifen as it could be done in parallel. Cells were transfected with HOTAIRM1 siRNAs (siRNA#1 and siRNA#2) and scrambled siRNA as a control in 96-well plates. 48 hours following transfection, where appropriate 4-OH tamoxifen was added at increasing concentrations, and plates were incubated for three to four days. An MTT assay was then performed to assess cell viability. There was not a statistically significant change in proliferation or tamoxifen sensitivity in either cell line (Figure 5.3).

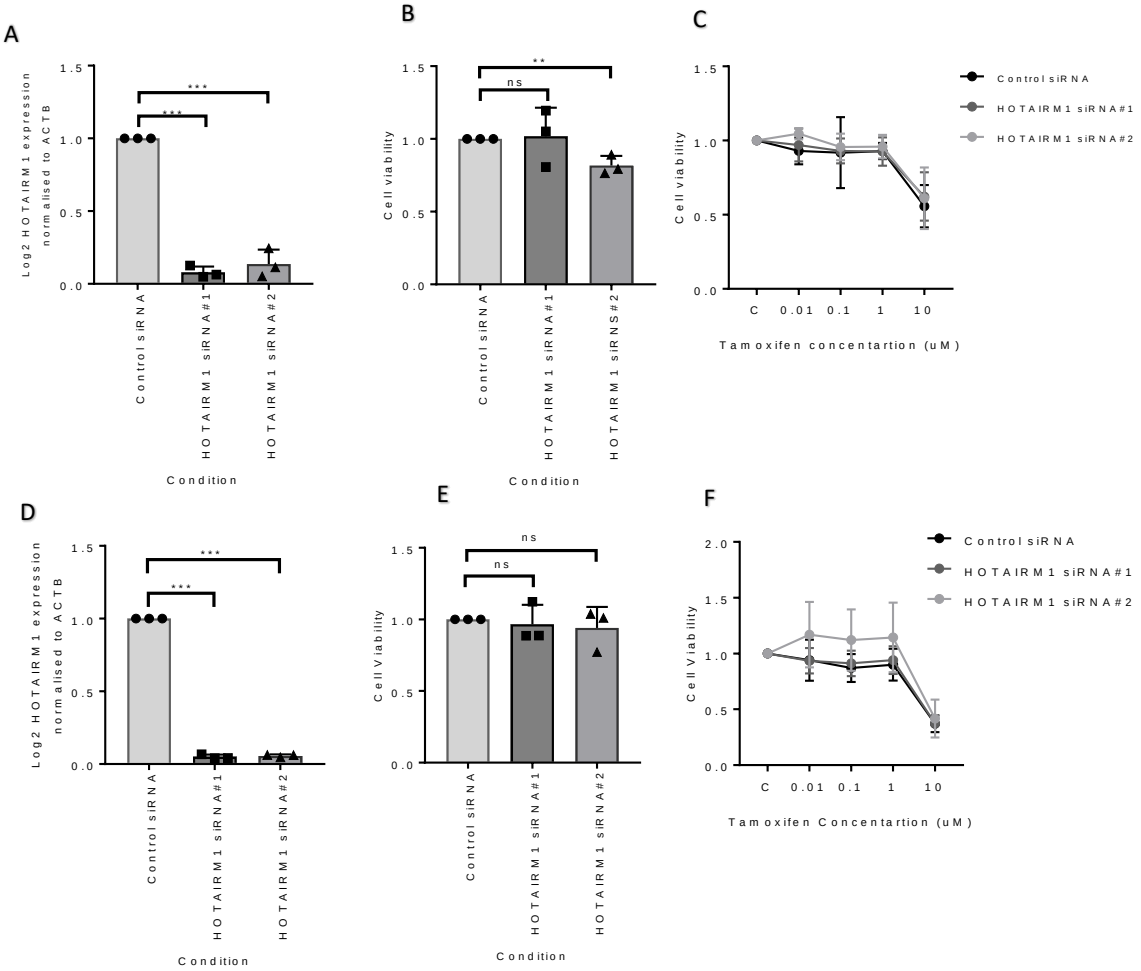


Figure 5.3. HOTAIRM1 depletion does not affect proliferation and Tamoxifen sensitivity in CAL51 or MDA-MB-468 cells**.** (A/D) siRNA attenuated HOTAIRM1 Expression: CAL51 cells (A-C) and MDA-MB-468 cells (D-F) were transfected with scrambled siRNA (Control siRNA), HOTAIRM1 siRNA#1 AND siRNA#2 using transfection reagent DhrmaFECT1. 48 hours post-transfection, cell pellets were collected, RNA isolated, cDNA synthesised, and HOTAIRM1 expression was measured by RT-qPCR, HOTAIRM1 CT values were averaged and normalised to β-actin. for each condition, data points represent log2 fold change of relative HOTAIRM1 expression +/- SD, 3 independent experiments were performed (N=3) and 3 technical replicates were processed. (B/E) cells transfected with scrambled siRNA (Control siRNA), HOTAIRM1 siRNA#1 and siRNA#2 were tested for HOTAIRM1 depletion effect on proliferation. (C/F) cells transfected with scrambled siRNA (Control siRNA), HOTAIRM1 siRNA#1, and siRNA#2 were tested for sensitivity with different tamoxifen concentrations and vehicle control. data points represent mean cell viability ± SD, for each condition 3 independent experiments were performed (N=3) and 5-10 technical replicates were processed * denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

5.2.4 Effect of HOTAIRM1 depletion on cell cycle in TAMR and CAL51 cell lines

Linked to proliferation is cell cycle progression. Cell cycle analysis was carried out to investigate the influence of HOTAIRM1 on breast cancer cells cell cycle progression. TAMR and CAL51 cells were transfected with control siRNA, HOTAIRM1 siRNA#1 and #2. Forty-eight hours post-transfection, cells were collected and stained with DNA-binding PI dye. Samples then run through the flow cytometer and data were then analysed in FlowJo software. As shown in (Figures 5.4), no changes were observed in cell cycle distribution in either cell line when depleting HOTAIRM1.

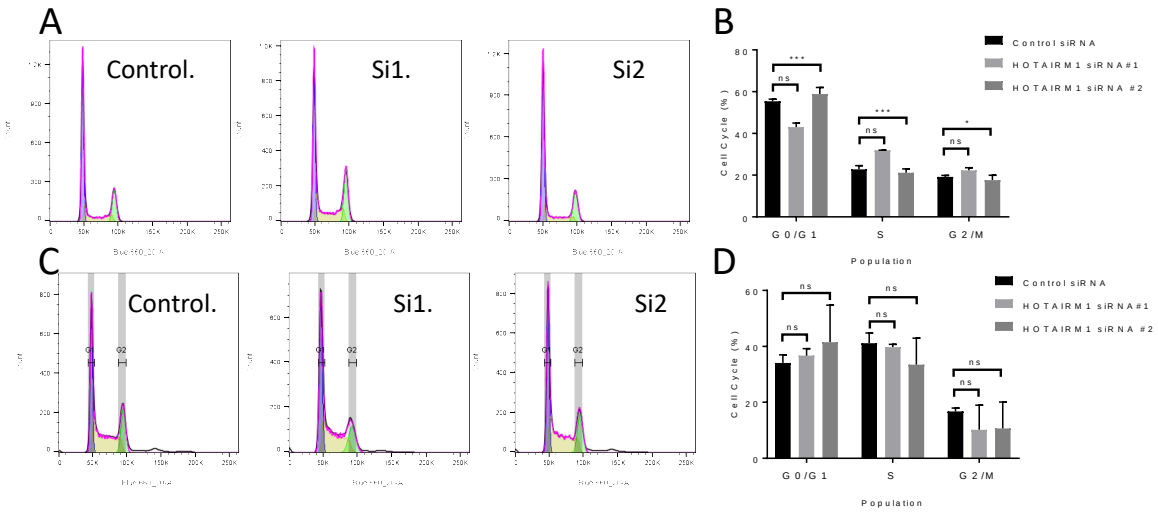


Figure 5.4. HOTAIRM1 depletion does not alter the cell cycle in TAMR and CAL51 cell lines. (A) Representative images of HOTAIRM1 depleted TAMR cell cycle analysis using flow cytometry following propidium iodide staining (B) quantification of the cell cycle populations. (C) Representative images of HOTAIRM1 depleted CAL51 cell cycle analysis using flow cytometry following PI staining. (D) quantification of the cell cycle populations. Data are expressed as the mean ± SD of three independent repeats. ns denotes non-significant, *** denotes p= ≤0.001, (Student's independent samples unpaired two-tailed t-test).

5.2.5 Effect of HOTAIRM1 depletion on TAMR and CAL51 cell adhesion

Metastasis occurs in breast cancer and is associated with worse prognosis. Endocrine resistant breast cancer is known to have more metastatic potential. One of the processes in metastasis is a change in adhesion of cells to sold matrixes. Analysis of cell adhesion in TAMR and CAL51 cells depleted of HOTAIRM1 was performed on matrigel-coated 96-well plates. After depletion cells were reseeded in the prepared 96 well plates and were allowed time to bind to the extracellular matrix prior to washing and staining with crystal violet. Following HOTAIRM1 depletion in CAL51 and TAMR cells showed no change in cell adhesion (Figure 5.5 A/B).

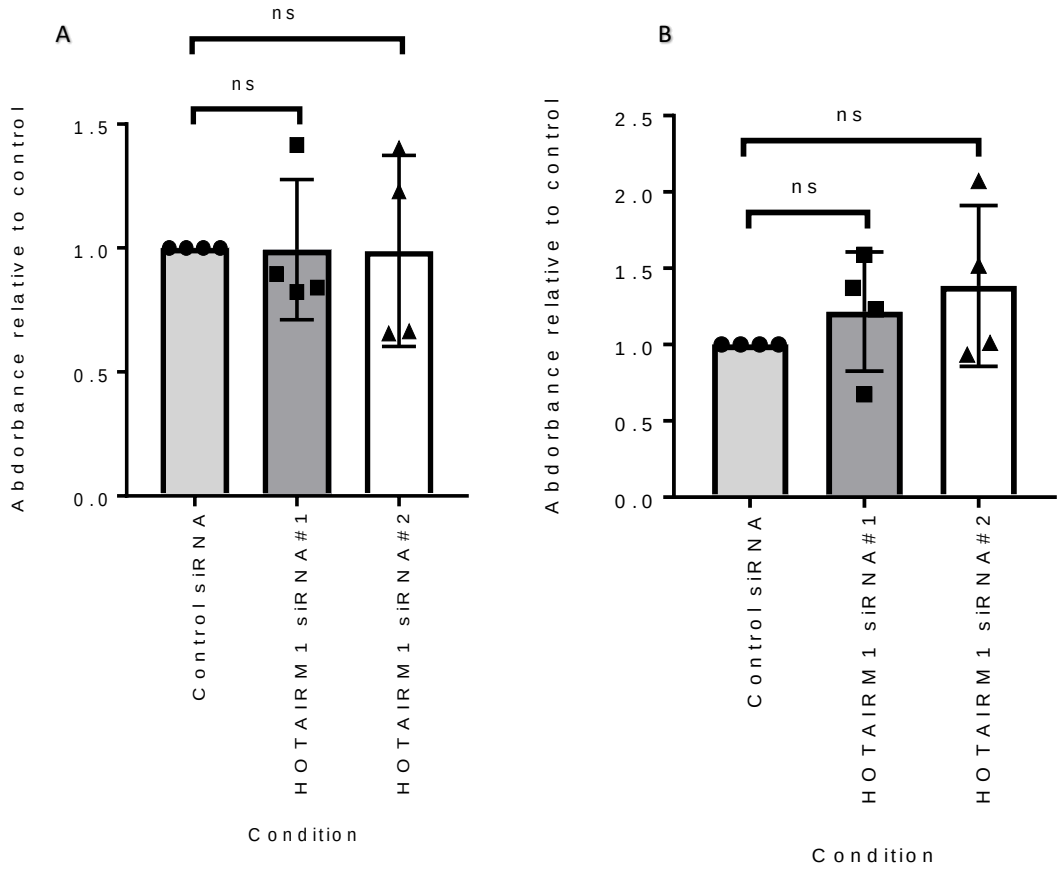


Figure 5.5. Depletion of HOTAIRM1 does not alter the adhesion of cells to Matrigel. (A) TAMR or (B) CAL51 cells. Cells were transfected with two independent HOTAIRM1 siRNAs and scrambled control siRNAs. Cells were then seeded on matrigel-coated wells in 96-well dish in 5 technical repeats per condition and incubated for 1 hour at 37℃. After washing adherent cells were fixed and stained with crystal violet. Light absorbance of stained cells is proportional to the number of cells. Absorbance relative to the control siRNA was calculated. The mean ± SD of four independent repeats is shown.

5.3.6 Effect of HOTAIRM1 depletion on CAL51 cell migration

Cell migration is also linked to metastasis. Therefore, cell migration rates were assessed in tamoxifen resistant breast cancer cells to investigate the influence of HOTAIRM1 on breast cancer cell movement. Cal51 cells were transfected with control siRNA, HOTAIRM1 siRNA#1 and #2. Forty-eight hours post-transfection, cells were seeded at 3×10^4 cells per well in cell inserts 2-wells. After achieving a confluent monolayer, a gap was created by removing the culture inserts and cell

left to move into the gap in low serum media. Images were captured at 0 h, 24 h, 48 h, 72 h, and 96 h post-gap creation (Figure 5.6). The wound area in percentage represents the rate of migration of TAMR across the created gap. Depleting HOTAIRM1 with HOTAIRM1 targeting siRNA did not change the pattern of migration compared to control.

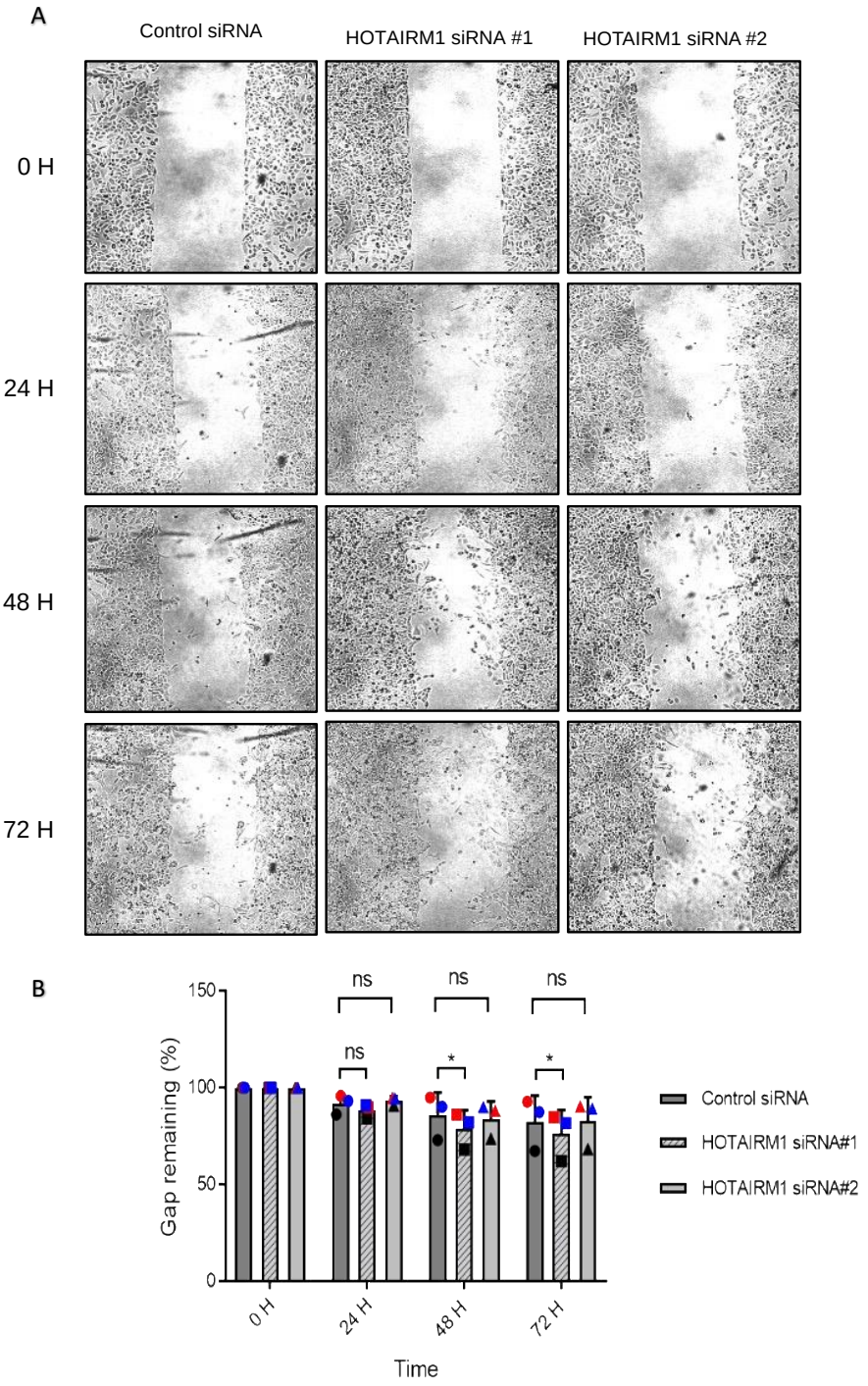


Figure 5.6. Depletion of HOTAIRM1 does not alter cell migration. (A) representative images of CAL51 cells following treatment with control siRNA, HOTAIRM1 siRNA#1, and HOTAIRM1 siRNA#2. Cells were transfected with siRNA and left 48 h prior to replating in culture cell dishes. The culture well was then removed, and images were taken. (B) Quantification of cell migration as indicated by 100% gap remaining relative to time 0. The mean ± SD of 3 independent experiments are shown. ns denotes non-significant Student's T.test comparing control siRNA to HOTAIRM1 siRNAs at each time point.

5.3.7 Effect of HOTAIRM1 depletion on EMT

TNBC are well established to be more aggressive and invasive. Epithelial-mesenchymal transition (EMT) enables some cancer cells to suppress epithelial characteristics and adopt mesenchymal like features. This facilitates cell movement and thus metastasis. Several proteins are associated with EMT. First, we chose to investigate E-Cadherin, a protein known to be vital to maintaining tissue integrity. E-Cadherin is a transmembrane protein that interacts extracellularly with adjacent cells to maintain cell to cell adhesion. The second protein investigated was β-catenin which links the intracellular domain of E-cadherin to the cellular cytoskeleton. Also, together with TGF-1 and Wnt signalling pathway, B-Catenin acts as a transcription factor participating in tissue morphogenesis and carcinogenesis. Yes-associated protein (YAP) was the third protein investigated, that functions in Hippo pathway, primarily acting as a transcriptional co-activator for many genes involved in proliferation and apoptosis.

Two siRNAs targeting HOTAIRM1 transcript, and a scrambled control siRNA were transfected into Cal51 cells. 48 hours post-transfection cells were stained for E-Cadherin, and β-Catenin. Images were taken and analysed for signal intensity as a quantitative measure of corresponding marker abundance in cells. As shown in (Figure 5.7). No significant change was observed in any condition.

In addition, 48 h post transfection cells were lysed, and western plot was performed using antibodies specific to β-catenin, E-cadherin, and YAP proteins, no change was detected (Figure 5.8). This data suggests no apparent effect of HOTAIRM1 depletion on the assessed proteins.

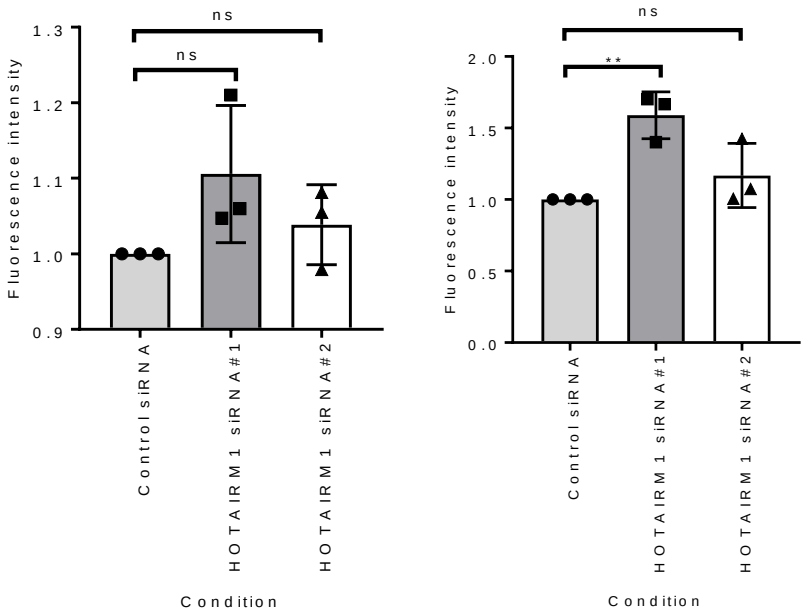


Figure 5.7. HOTAIRM1 depletion does not affect β-catenin or E-cadherin signals in CAL51 cells. (A) Representative images of CAL51 stained for DAPI (cyan), E-cadherin (red), and β-catenin (green) (B) quantification of E-cadherin signal intensity. (C) quantification of β-catenin signal intensity 100 cells were counted per condition per repeat, data points represent the mean ± SD of three independent repeats. Imaged were taken with a 40x microscope objective, scale bars represent 20 µm

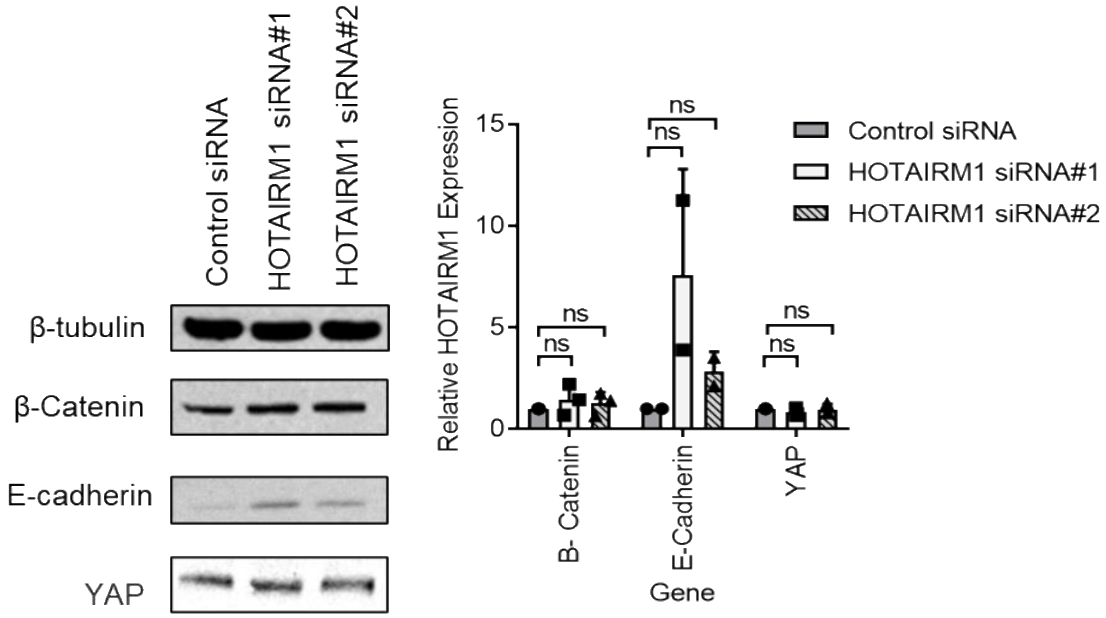
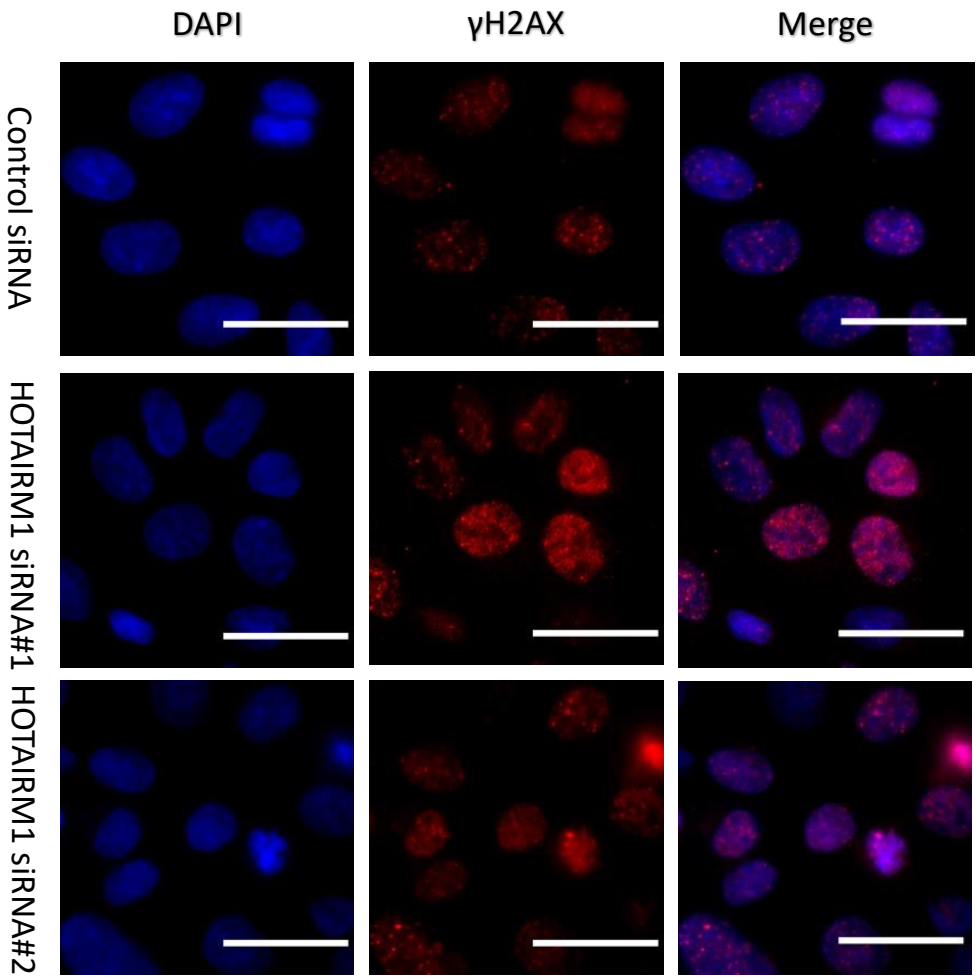


Figure 5.8. Depletion of HOTAIRM1 does not alter the expression of EMT-related protein-coding genes. CAL51 cells were seeded on coverslides placed in 6-well dish in 3 technical repeats per condition and incubated overnight at 37°C and transfected the next day. cells were fixed and exposed to the appropriate primary and secondary antibodies. Band intensity is proportional to the expression of the gene of interest. expression relative to the control ACTB band was calculated. Mean ± SD of three independent repeats is shown, except for E-cadherin (n=2). Ns denotes nonsignificant.

5.3.8 Effect of HOTAIRM1 depletion on DNA damage

A common phenomenon of cancer is increased genomic instability this can often be seen as an increase in DNA damage in cancer cells. We looked at H2AX foci as a marker of damage again staining 48h post transfection of siRNA. No differences were seen (figure 5.9)



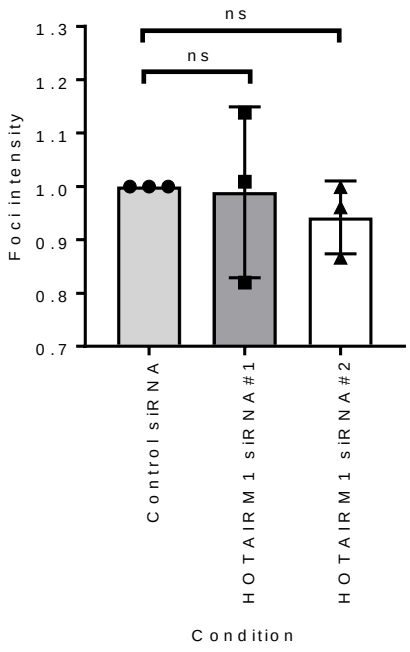


Figure 5.9. HOTAIRM1 depletion does not affect γH2AX foci signal in CAL51 cells. (A) Representative images of CAL51 stained for DAPI (cyan) and γH2AX (red) (B) quantification of γH2AX foci intensity. 100 cells were counted per condition per repeat., data points represent the mean ± SD of three independent repeats. Imaged were taken with a 40x microscope objective, scale bars represent 20 µm

5.3.9 Effect of HOTAIRM1 depletion on HOXA genes cluster

Considering the regulatory role lncRNA usually play on neighbouring genes and that HOTAIRM1 is transcribed from chromosome 7 within the HOXA genes cluster and that HOXA is an established role in cancer and drug resistance (Bhatlekar, Fields and Boman, 2014). TAMR and CAL51 cells were transfected with control siRNA, HOTAIRM1 siRNA#1 and #2. Forty-eight hours post-transfection, cells were collected, RNA was extracted, and cDNA was synthesised. Samples were then run through RT-qPCR machine to test for changes in HOXA1, HOXA5, HOXA9, and HOXA10. Fold changes in expression of each gene compared to scrambled control sample were calculated. As shown in Figure 5.10, regardless of successful HOTAIRM1 expression depletion, no significant change was observed in gene expression when depleting HOTAIRM1 with HOTAIRM1 targeting siRNA compared to control in both cell lines (TAMR and CAL51).

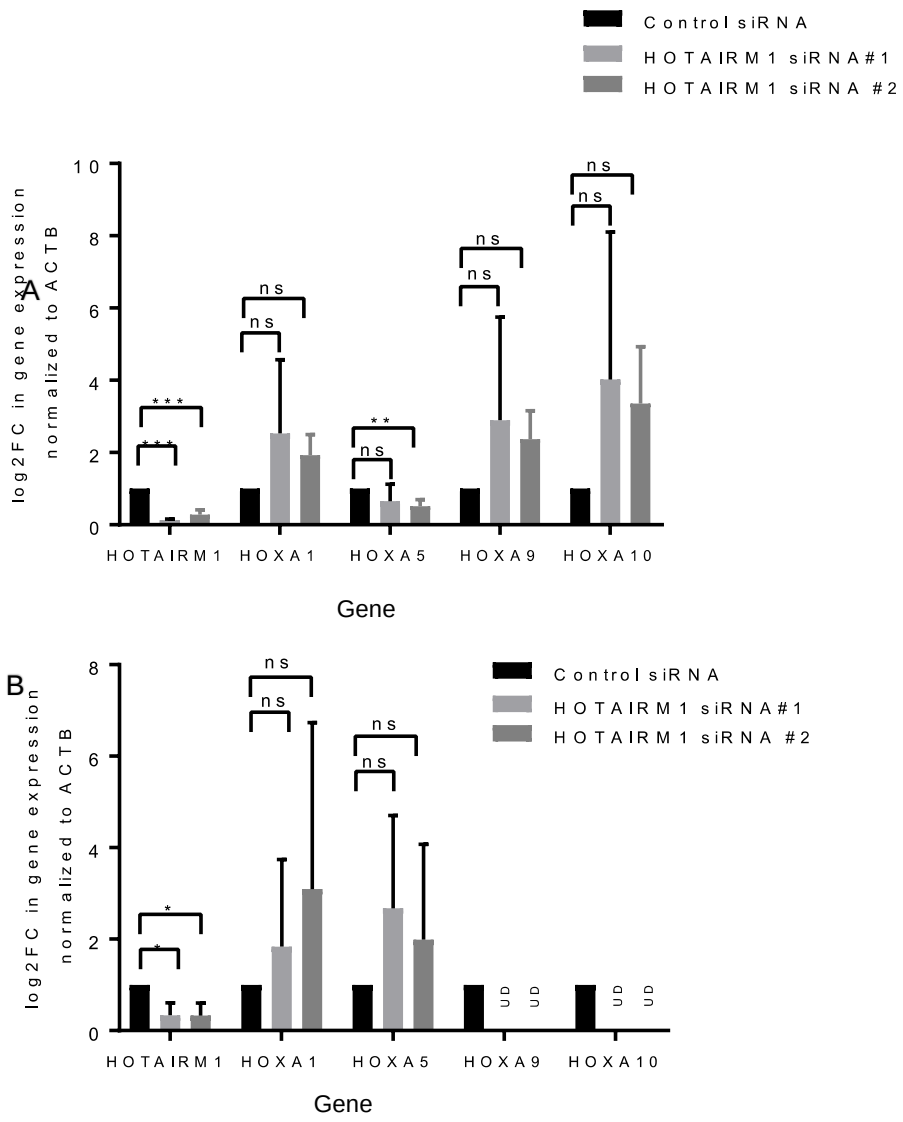


Figure 5.10. HOTAIRM1 depletion effect on HOXA genes. Following HOTAIRM1 siRNA mediated depletion, HOXA1, HOXA5, HOXA9, and HOXA10 expression was measured in TAMR (A) and CAL51 (B) cell lines. Lines Indicate mean ± SD of three independent repeats. *denotes p= ≤0.05, ** denotes p= ≤0.01, *** denotes p= ≤0.001, ns denotes nonsignificant, and UD denotes undetermined expression value.

5.3.10 Effect of HOTAIRM1 upregulation using ATRA treatment on MCF-7 response to tamoxifen.

Given that HOTAIRM1 is upregulated in tamoxifen resistant MCF-7 cells we sought to induce resistance by inducing expression of HOTAIRM1 in normal tamoxifen sensitive MCF-7 cells. Considering the location of HOTAIRM1 transcript is in HOXA gene cluster, transcription of this region is well established to be enhanced by retinoic acid (RA), an interaction known to play important role in myeloid cells differentiation (Wei *et al.*, 2016). For the purpose of HOTAIRM1 expression manipulation studies, it was decided to initially try ATRA treatment (section 2.1.7), to examine how it affects HOTAIRM1 expression in MCF-7 cells, where HOTAIRM1 expression is lower than that in the resistant cell line. MCF-7 cells were incubated with 1 µM ATRA for either two hours, 24 hours, or 48 hours. As shown in (Figure 5.11 A) 2 hours, 48 hours, and 24 hours of exposure gave a pronounced increase in the log 2-fold change of HOTAIRM1 expression (5, 40, and 69) respectively. Therefore, considering the time factor, 2 hours and 24 hours exposure were chosen as the optimal ATRA exposure period for MCF-7 cells to express sufficiently upregulated levels of HOTAIRM1. These results were encouraging to take a further step into testing the viability of ATRA-treated cells in tamoxifen, and whether the change in HOTAIRM1 level might enhance their survival.

The impact of HOTAIRM1 upregulation in MCF-7 cells on their response to tamoxifen treatment was determined using the MTT assay. MCF-7 cells were cultured in the absence and presence of 1uM ATRA for two hours and 24 hours, where-after increasing concentrations of tamoxifen were added and cells incubated for a further four days (Figure 5.11). Upregulation of HOTAIRM1 did not alter sensitivity to tamoxifen.

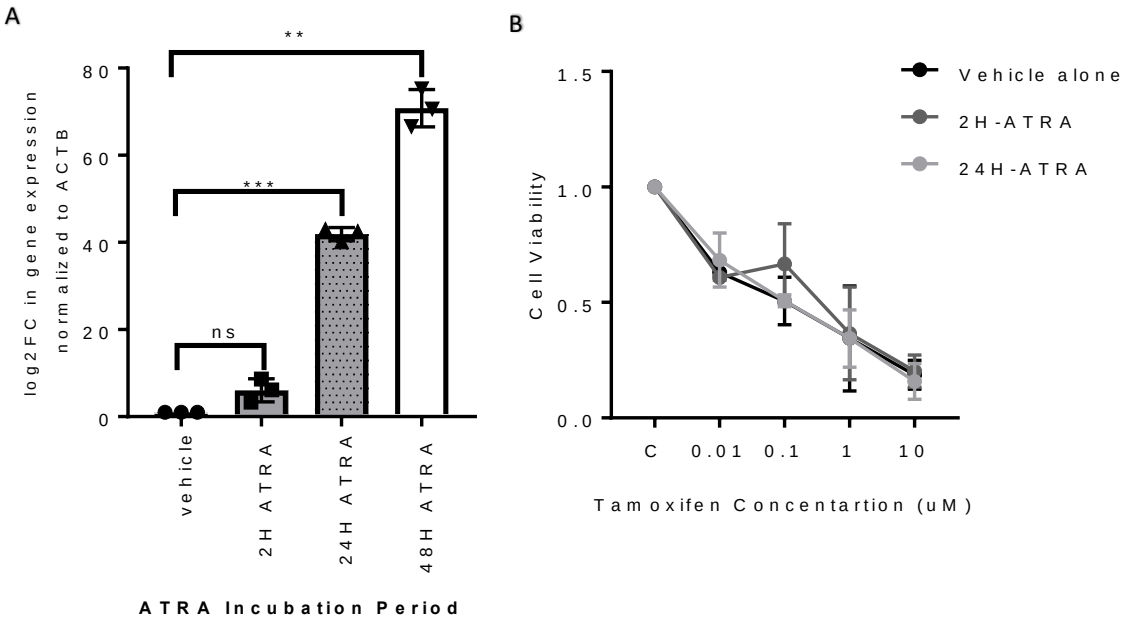


Figure 5.11. Upregulation of HOTAIRM1 expression using ATRA does not alter MCF-7 cells response to tamoxifen. (A) The log2 fold change in HOTAIRM1 expression shows an increase with longer incubation time. 2 hours (2H), 24 hours (24H) and 48 hours (48H). HOTAIRM1 expression was determined by qRT-PCR relative to ACTB reference gene. (B) ATRA-treated MCF-7 cells viability with tamoxifen treatment. MCF-7 cells were treated with ATRA for two hours (2H-ATRA) or 24 hours (24H-ATRA) prior to the addition of tamoxifen. Cell viability was then determined by MTT after a further four days. lines indicate mean ± SD of technical repeats on one occasion (n=3).

5.3.11 Effect of HOTAIRM1 depletion on global gene expression in CAL51 cell line

Given that depletion of HOTAIRM1 failed to produce any significant changes in cells with high levels of expression but that HOTAIRM1 did appear to be aberrantly

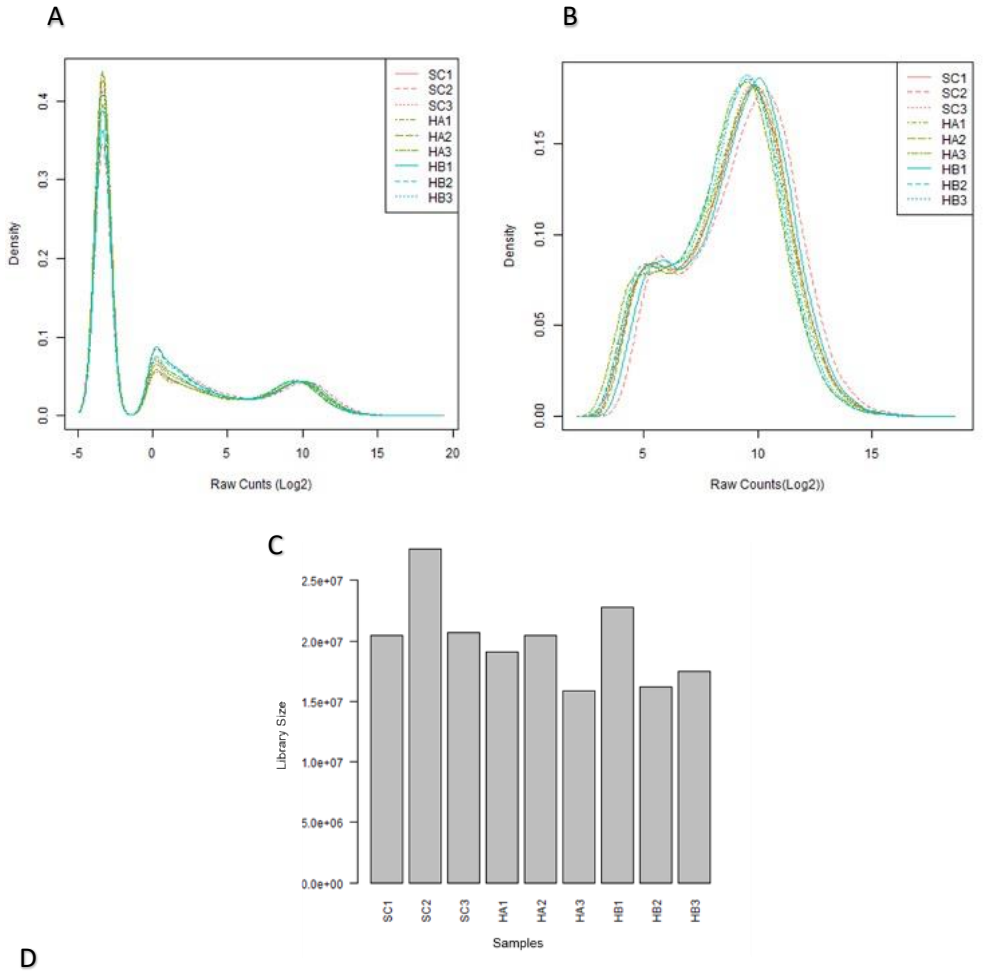
expressed in TNBC cell lines and patient samples. We decided to take an unbiased approach to look at the effect of depletion on global transcriptomics.

The expression profiles of 3 groups of samples were compared: CAL51 transfected with control scrambled siRNA (SC), CAL51 transfected with HOTAIRM1 siRNA#1 (HA) and CAL51 transfected with HOTAIRM1 siRNA#2 (HB). Each group contained three biological replicates. CAL51 cells were treated, and cell pellets sent for external RNA-seq. Sequencing read primary data were trimmed to remove adapters and poor-quality nucleotides then aligned to Homo sapiens GRCh38 reference genome. Gene counts were then calculated by using featureCounts command.

For analysis raw reads counts were inputted with a reference file that contains, basic biological information essential to identify samples and their criteria comparisons (below).

Sample ID	Cell Type	Condition
SC1	CAL51	Control siRNA
SC2	CAL51	Control siRNA
SC2	CAL51	Control siRNA
HA1	CAL51	HOTAIRM1 siRNA#1
HA2	CAL51	HOTAIRM1 siRNA#1
HA3	CAL51	HOTAIRM1 siRNA#1
HB1	CAL51	HOTAIRM1 siRNA#2
HB2	CAL51	HOTAIRM1 siRNA#2
HB3	CAL51	HOTAIRM1 siRNA#2

One of the most important steps in data prepossessing is filtering low-expression genes. There are many reasons for doing this step: physically, genes with law counts are not long enough to produce any considerable biological effect. (From reads to genes to pathways). In addition, from a statistical point of view, lowly expressed sequenced reads have a higher rate of measurement errors and noise. Hence, excluding this type of genes at the very start, is vital to reduce the burden of the multiple testing. This simultaneously improves sensitivity of downstream differential gene expression (DGE) analysis and increase the number and quality of detected genes. To filter data, count-per-million (CPM) values were obtained using *cpm* function in edgeR package. Looking at the trend of library sizes in our samples (Figure 5.12 A). With library sizes between 15-27 million reads, it was required that for a gene to be included it have at least 0.5 CPM in all 9 samples. The instant effect of removing unwanted genes was observed in density plots, generated using *plotDensity* function from affy package (Figure 5.12 B). This function calculates the density for each observation in the input log base 2 scaled raw count matrix and estimates the probability density function of a random variable creating a smooth, continuous distribution curve. The peak of the curve represents the maximum concentration of raw counts. Unfiltered data has 3 peaks the far biggest is at -3.321 which equals log base 2 of 0, the value that makes the most of data (figure 5.12 A). After filtering, samples' curves are smoother and show better overlap and similar peaks with no strange deviations (Figure 5.12 B). Library size distribution was also evaluated for any aberrating data (Figures 5.12 C and D).



D

Sample ID	SC1	SC2	SC3	HA1	HA2	HA3	HB1	HB2	HB3
Library size (bp)	20483155	27630701	20680817	19084970	15857378	22772696	16161720	20423209	17479654

Figure 5.12. Quality control assessment of HOTAIRM1 depleted CAL51 RNA-seq data. (A) Raw counts distribution before filtering of lowly expressed genes. (B) Raw counts distribution after filtering of lowly expressed genes. (C)/(D) Library sizes of RNA-seq samples. Samples were sequenced, 3 samples in each cells group and the total number of raw counts in each sample was adapted into a bar indicating the library size.

Raw data also underwent a quality control step post-filtering; to assess the quality of each samples' biological annotation and how they relate to each other statistically. To check for batch effect in the experiment, principal component analysis (PCA) was conducted using R package DESeq2 function *plotPCA.* PCA is a statistical procedure used to observe the direction and distance of which samples diverge based on their statistical variation. In a two-dimensional co-ordinate system, samples were plotted relative to the two features with maximum variance called principal component 1 and 2 (PC1 and PC2). The PCA plot produced for sequenced HOTAIRM1 depleted CAL51 cells raised concern with this data set as a whole. Samples were showed to vary based on their biological repeat number rather than depletion status (Figure 5.13).

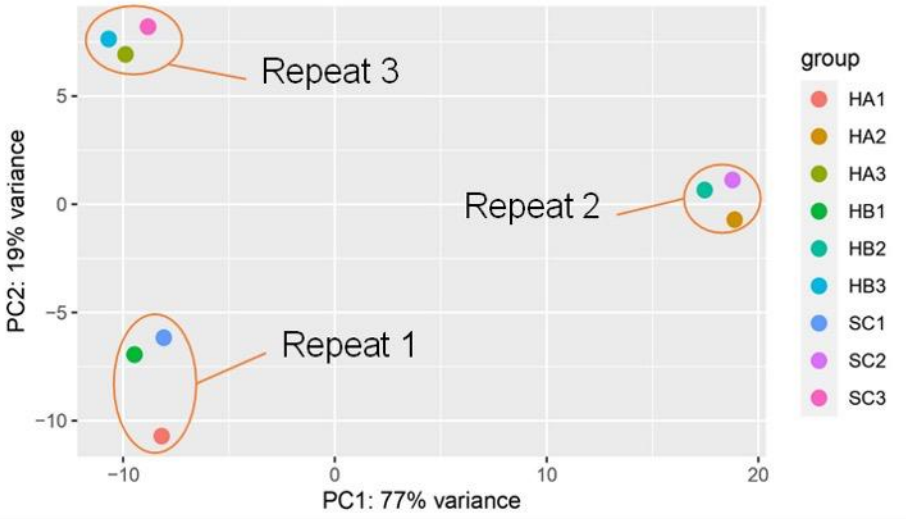


Figure 5.13. Principle component analysis plot showing variability in global gene expression between samples. variability between samples was plotted on a 2D plan as a percentage, where, PC1 on the x-axis was plotted against a PC2 on the y-axis. Samples are indicated in coloured dots. The relationships between samples are determined by judging the distance between samples on the horizontal PC1 axis, the further the distance between samples, the highest degrees of variability.

5.3 Discussion

The aim of this chapter was to provide insight into the role of HOTAIRM1 in carcinogenesis using TAMR and CAL51 cell lines as a representation of HOTAIRM1 highly expressing cells. Of the four candidate lncRNAs initially tested in chapter 3, we opted to further explore HOTAIRM1 in detail for many reasons, fundamentally, the consistent level of expression among different cell lines and the considerable depletion level we were able to achieve. The planned subsequent experiments required more persistent silencing of HOTAIRM1 that might extend to 5 days. So, our siRNA-mediated silencing protocol was further optimised, and the duration of silencing was assessed; to cover the duration of our experiments without affecting cellular wellbeing. Proliferation, adhesion to solid matrix, migration, EMT marker expression and DNA damage were all assessed. These were chosen due to previous literature linking HOTAIRM1 to each of the processes in other cancers. In none of the HOTAIRM1 highly expressing cell lines did depletion of HOTAIRM1 lead to changes in proliferation. This is in contrast to findings in thyroid cancer (Li et al 2021) glioblastoma (Lin et al 2020) and endometrial cancer (Li et al 2019).

Aggressive cancer cells are also known to possess less ability to adhere to surrounding tissue and increased motility, making them more prone to metastasise (Osuchowska *et al.*, 2021). In this regard lncRNA has suggested as a prognostic marker for metastasis in oestrogen receptor positive breast cancer (Sørensen et al 2013). We did not see any change in EMT type gene expression, adhesion, or mobility upon HOTAIRM1 depletion in either CAL51 or TAMR cells. Again, this is not consistent with what is seen in other cancers where lncRNA has been shown to in other cancers e.g. thyroid cancer (Li et al 2021). Our observations on EMT-

related phenotypes could be expanded to more repeats and different assays such as the trans-well test.

(Xueqing Zhang et al., 2009) and (Zhang, Sherman M. Weissman and Newburger, 2014) proposed that the knockdown of HOTAIRM1 maintains cell cycle activation from G1 to S phases. Though, our cell cycle analysis showed no difference in cell cycle population quantification between HOTAIRM1 depleted and unmodified cells in either TAMR or CAL51 cell lines. Both studies were in leukaemia cells, suggesting a different cell cycle role in breast cancer cells.

The HOXA gene cluster enfolds HOTAIRM1 in addition to many protein-coding genes. Following HOTAIRM1 depletion in TAMR and CAL51 cell lines, the expression of each of the four HOXA genes (HOXA1, HOXA5, HOXA9, HOXA10) were compared between HOTAIRM1 depleted and highly expressing cells. HOXA5 has been previously shown to participate in tamoxifen resistance in breast cancer (Kim *et al.*, 2021), in our data, the expression of HOXA5 was decreased in response to HOTAIRM1 silencing, although this was only statistically significant only with one of the siRNAs, it is still encouraging as a potential axis for driving resistance, and further experimenting on more repeats would clarify this more. Recently, Kim et al., (2020), published the direct relationship between HOTAIRM1 being a driving cause of tamoxifen resistance in breast cancer, acting through HOXA1 transcriptional regulation HOXA1 down-regulation might be an off-target effect of the used HOTAIRM1 siRNA silencing technique, as using a pool of multiple siRNAs rather than individual siRNAs increase this risk, especially with genomic location; both genes being in close proximity to each other (Brown *et al.*, 2022)

We used ATRA treatment to induce the expression of HOTAIRM1 in a cell line where HOTAIRM1 expression was low (MCF-7) following the same protocol mentioned in the literature for acute promyelocytic leukemia cell line (Wei *et al.*, 2016). This was done mainly to have an idea about the degree of inducibility of HOTAIRM1 in this cell line. HOTAIRM1 was shown to be highly induced, especially after overnight and two-day incubation. ATRA is known to induce the proliferation of leukemia cells and induction of HOTAIRM1 is frequently seen with ATRA treatment (Zhang, Sherman M Weissman and Newburger, 2014; Wei *et al.*, 2016), due to its global effect on many HOX genes (Chen *et al.*, 2012). Here ATRA did regulate HOTAIRM1 expression but no change in proliferation was observed. However, to really answer whether HOTAIRM1 expression in a low background can drive resistance this experiment would need to be refined by pure HOTAIRM1 targeted amplification for example by HOTAIRM1 over-expression using an expression vector.

By sequencing the total RNA extracted from HOTAIRM1 depleted CAL51 cells we sought to achieve a comprehensive view of the effect HOTAIRM1 might have on global gene expression and molecular pathways. So, more targeted experimentation can be planned. Though the knockdown was successful. We faced a major obstacle in the quality control step while analysing RNA-seq data computationally. Samples are clustered by biological repeat rather than by experimental condition. This inconvenience is common as a result of batch-effect or sample cross-contamination while handling the physical samples. Unfortunately, it was not possible to make any conclusions from these data as the aberrated

clustering attenuated proceeding with the downstream analysis of this data. This could be repeated.

Summary

In summary, we have screened for several aspects of cancer in HOTAIRM1 depleted TAMR and CAL51 cell lines. Although we saw no differences in the processes examined, the HOTAIRM1/HOXA5 axis is a tempting pathway for resistance, needing further in-depth investigation.

Chapter 6. Analysis of Publicly Available Datasets

6.1 Introduction

In chapter 3 a list of differentially expressed genes was produced by comparing the transcriptomic profiles of tamoxifen-resistant and sensitive breast cancer cell lines. Functional biological assessment of these genes - LUCAT1, SOX21-AS1, NR2F1-AS1, and HOTAIRM1 – was carried out in chapters 4 and 5. Unfortunately, none of these candidates could be validated as having a role in tamoxifen resistance. Our next step was to analyse publicly available datasets that include different endocrine resistance phenotypes in order to try again to identify common genes involved in the complex pathophysiology of tamoxifen resistance in breast cancer.

Two databases were selected

1. The Gene Expression Omnibus (GEO) database

GEO (Edgar, Domrachev and Lash, 2002) is a public repository that stores next generation sequencing data submitted by different scientific groups. Archived datasets are available for researchers and can be accessed through different direct and indirect modalities. Large data input is preferable as it retains statistical power, however, it is not always accessible. Instead, what's readily available is a huge number of small sample size datasets hosted in a database such as GEO. When metadata available in the GenomeBrowser database (www.ncbi.nlm.nih.gov/sites/GDSbrowser) (about 4348 GEO datasets) were analysed, most had a sample size of 12. (Figure 6.1)

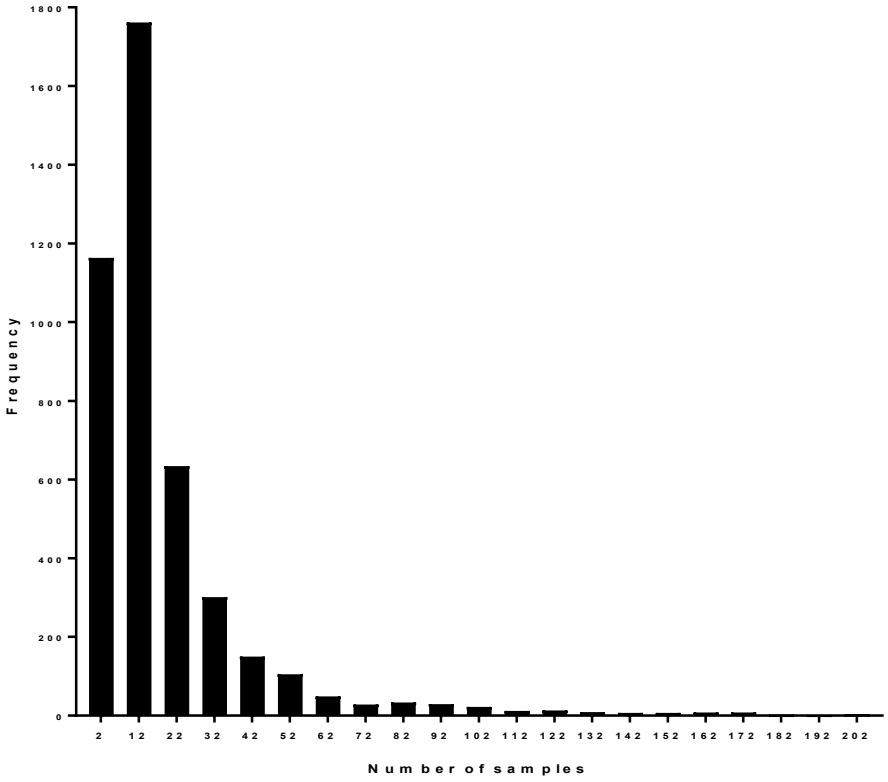


Figure 6.1. Sample size frequency for datasets in the GEO database. The histogram illustrates the frequency of datasets containing the number of samples.

2. The Cancer Genome Atlas (TCGA) data base

TCGA (Tomczak, Czerwińska and Wiznerowicz, 2015)is one of the biggest next generation sequencing datasets. TCGA collects different types of data, clinical information (e.g., patient data, treatment, and survival data), molecular analyte metadata (e.g., samples collection and processing) and molecular experimental data (e.g., gene expression and copy number microarray).

The hypothesis of this chapter is:

Gene expression analysis of publicly available datasets related to tamoxifen response, will produce a list of lncRNAs and protein coding genes that may overlap with RNA-seq in chapter 3.

The aims of this chapter are:

1- To systematically search GEO dataset for tamoxifen resistance related phenotypes

2- To process GEO microarray expression data using quality control and differential expression bioinformatics pipelines, construct a list of statistically significant differentially expressed genes for each study.

3- To process TCGA RNA-seq expression data using quality control and differential expression bioinformatics pipelines and construct a list of statistically significant differentially expressed genes.

The objectives of this chapter include:

1- analysing GEO raw sequencing data by:

A) Assessing sample library sizes and adjusting for any detected systematic biases.

B) Performing DEA and ranking lncRNAs and protein-coding genes based on DEA results.

2- analysing TCGA sequencing data based on PAM50 classification of samples

3- comparing GEO and TCGA data internally and to RNA-seq results.

6.2 GEO data set analysis

6.2.1 GEO Data set selection

Selecting the right datasets to be analysed was critical, as studying a disease with complex genotype requires finding studies that have asked the right research questions. A search strategy was created, and appropriate datasets were selected from a systematic search of the GEO while including several inclusion and exclusion criteria.

Our search question was formulated using P.I.C.O.T format (***P*** symbolizes the target population, ***I*** is the Intervention of interest, ***C*** is the control group, ***O*** is the key outcomes, and ***T*** is the Time frame over which the outcomes took place) (Riva *et al.*, 2012) as follows:

- **P**opulation: endocrine resistant breast cancer tissue or cells
- **I**ntervention/exposure: dysregulated of gene expression
- **C**omparison/control: endocrine sensitive breast cancer
- **O**utcome: relapse/recurrence/endocrine resistance

The time factor was excluded from the search strategy as it is not applicable in the case of endocrine resistance.

Based on that, our search question was “What are the genes that are differentially expressed between endocrine resistant and sensitive breast cancers leading to relapse of patients”.

Inclusion and exclusion criteria were chosen as follows: For a data set to be included, firstly, samples can be either clinical from human patients, or they can be from human cancer cell lines, as they are all considered biological models

representing the phenotype under investigation. Samples from other species were excluded, being biologically disparate and too divergent to be included. Secondly, the minimum number of samples were assigned as three controls and three comparative phenotype samples, this allows for establishing a solid argument of statistical significance. In addition, this cut-off fulfils the requirements of many software packages, that reject number of samples lower than three per condition. Thirdly, microarray data sets were only included when raw .cel files were submitted and were generated using platforms from Affymetrix and Illumina only, due to their compatibility with the used software pipelines in this chapter.

Target key words were produced for terms “breast cancer” and “endocrine therapy” or terms known to be associated with these phenotypes such as “triple negative breast cancer”, “metastatic breast cancer”, “Tamoxifen”, “Antineoplastic Hormonal therapy” and others. Our search query was built using appropriate variations in spelling and abbreviations following GEO database guidelines (GEO, 2021).

Studies were checked to fit the established inclusion criteria by referring to the metadata and experimental design information provided by the depositor. Five microarray datasets were selected to undergo further analysis, details of the selected data sets can be found in Table 6.1.

GEO ID	Number of samples	Type	Samples background	Platform	Publication
GSE67916	18	Cell lines	MCF-7 and TAMR were used in microarray experiment	GPL570	(Elias et al., 2015)
GSE27473	6	Cell lines	Wiled type MCF7 cell compared to estrogen receptor silenced MCF-7	GPL570	(al Saleh et al., 2011)
GSE124647	140	Clinical	biopsied metastatic hormone receptor positive breast cancer underwent expression profiling by array	GPL96	(Sinn et al., 2019)
GSE58644	321	Clinical	Biopsies of different types of breast cancer	GPL6244	(Tofigh et al., 2014)
GSE9195	77	Clinical	Biopsies from estrogen receptor-positive breast cancer patients treated with tamoxifen	GPL570	(Loi et al., 2010)

Table 6.1. Datasets selected from GEO data repository. GEO ID is the GSE reference that is one of GEO entities of the corresponding study. Number of samples indicates the size of the dataset. Type is the samples origin. Samples background provides a brief description about samples. Platform indicates the technical array record used for microarray data processing. Publication contains the reference supplied with the data on GEO.

Three datasets were formed of breast cancer tissue samples from patients with associated clinical data and two datasets of breast cancer cell lines. These microarray data were programmatically accessed and analysed from GEO using GEOquery package.

6.2.2 The GEO data sets

6.2.2.1 GSE67916.

In this dataset, there are 18 samples, 10 are tamoxifen resistance breast cancer cell lines and 8 are parent MCF-7 cell lines. Tamoxifen resistant cells were established by gradually exposing MCF-7 cells to increasing doses of tamoxifen until they adapt to growing normally in 1 μM of tamoxifen. The 10-tamoxifen resistant breast cancer cell lines were divided in 4 groups (3 TamR1 samples, 3 TamR4 samples, 2 TamR7 samples and 2 TamR8 samples), after quality assessment of the data and normalisation, MCF-7 samples were compared to TamR1 and TamR4 separately, since they satisfy the inclusion criterias. Raw microarray data format or CEL files and corresponding data frame of samples metadata were loaded, processed, and arranged in a way fitting the used pipeline. Gene expression measurements were calculated, then normalised, using the robust multichip average (RMA) normalization method. RMA includes the following steps, background correction, normalization to account for technical variations between individual samples (Figure 6.2 A, B, C and D), and finally calculation of gene expressions. To perform DEA, the first step is to describe the experiment in the analysis pipeline. So, what's called a sample level were assigned to fit samples based on their point of disparity; to design the study layout, making tamoxifen resistant phenotype the reference in the analysis. The multiple independent

statistical tests data undergo demand that it is filtered, going through the probe-level intensities, the MAS 5.0 absolute detection method was used to calculate the probability of observing probe intensity and assign a call for a probe being ‘present’ or expressed. We decided to include a probe if it was expressed in at least 9 samples at the same time, this excluded almost half of the probe population. In addition, probes that cross hybridized with other genes and Affymetrix control probes were also removed. Before proceeding to DEA, quality control analysis was also performed to compare replicates by observing their clustering behaviour relative to each other. As shown in (Figure 6.2 E) tamoxifen resistant samples clustered opposite to MCF-7 samples along the principal component 1 axis (representing the most distinguishable variation between samples). There was some concern around one MCF-7 sample (MCF-7.8) and one tamoxifen resistant sample (TR1.1), however we decided to sill include them for the following reasons. First though they are relatively close to each other, they are still on the same side of their group of samples, second there are 7 other MCF-7 samples and finally, we will analyse each tamoxifen resistant group of samples independently then compare and find the intersection with other comparisons. After calculating differential expression statistics, annotations were added. A summary of number of genes in each category comparing MCF-7 versus TR1 and TR4 is produced in (Table 6.2).

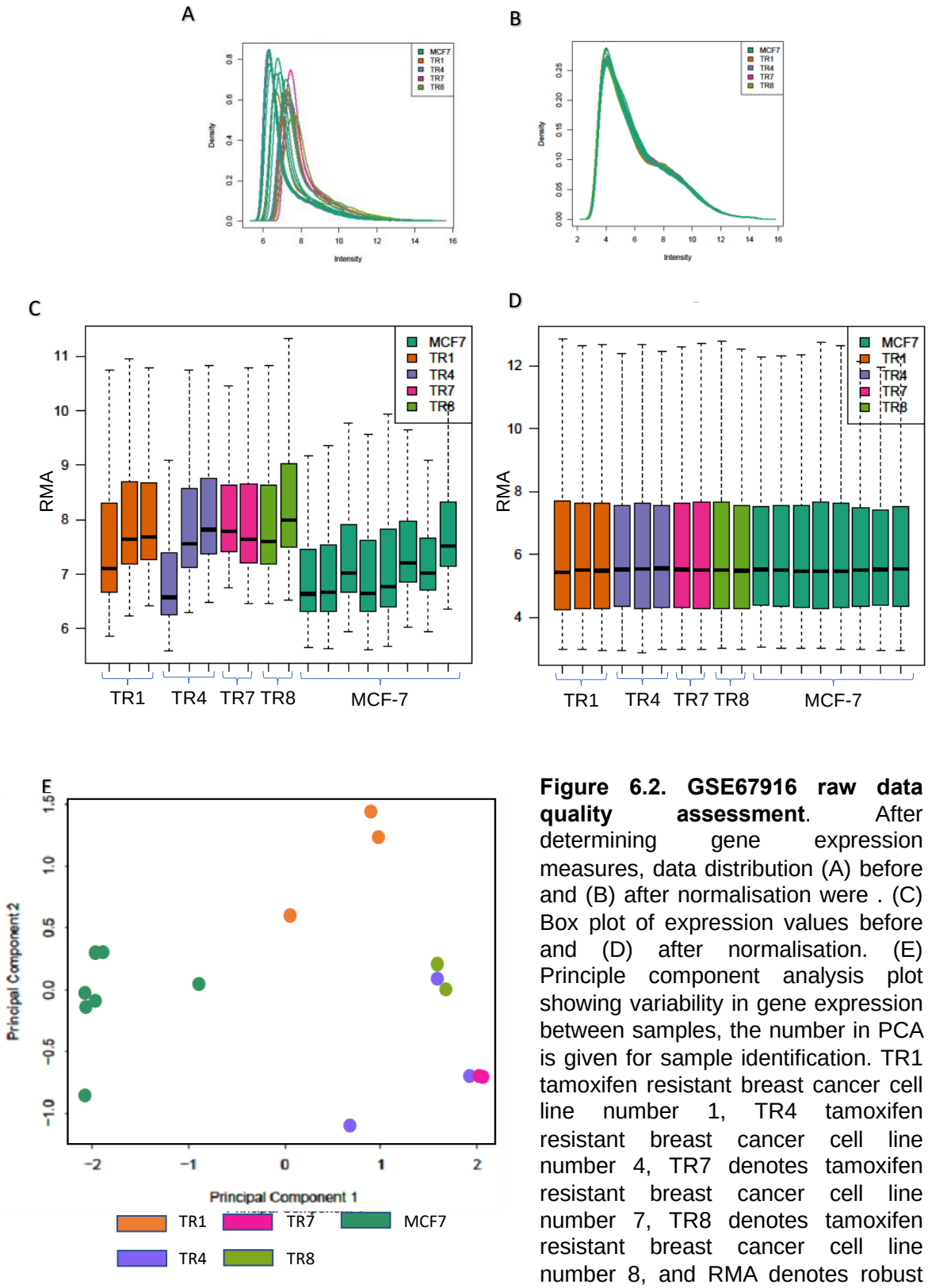
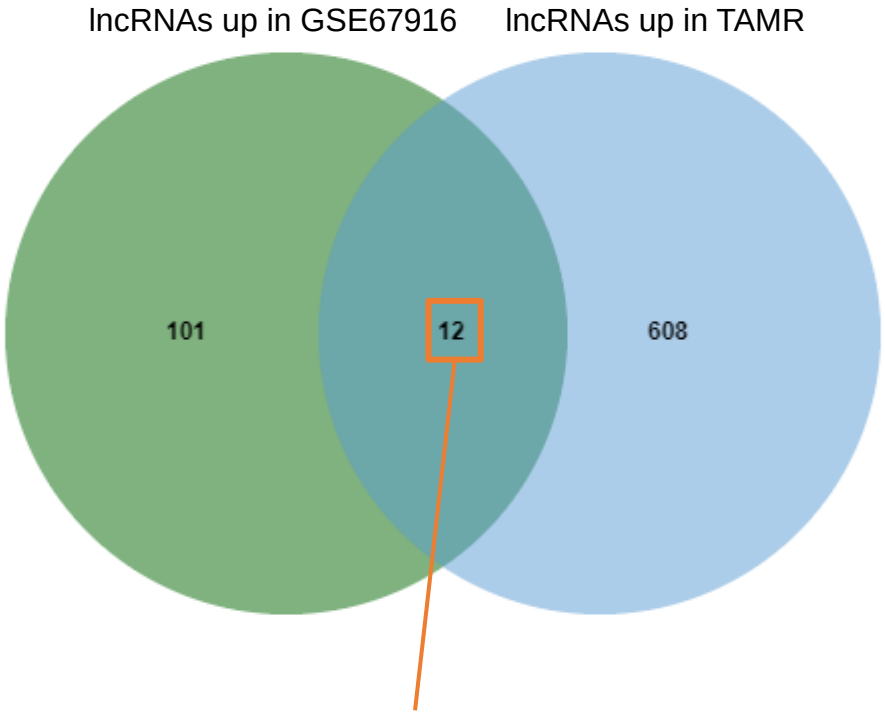


Figure 6.2. GSE67916 raw data quality assessment. After determining gene expression measures, data distribution (A) before and (B) after normalisation were . (C) Box plot of expression values before and (D) after normalisation. (E) Principle component analysis plot showing variability in gene expression between samples, the number in PCA is given for sample identification. TR1 tamoxifen resistant breast cancer cell line number 1, TR4 tamoxifen resistant breast cancer cell line number 4, TR7 denotes tamoxifen resistant breast cancer cell line number 7, TR8 denotes tamoxifen resistant breast cancer cell line number 8, and RMA denotes robust multichip average

Selection criteria	lncRNAs				protein coding genes			
	TR1		TR4		TR1		TR4	
	Up regulated	Down regulated	Up regulated	Down regulated	Up regulated	Down regulated	Up regulated	Down regulated
log fold change 1.5 and adjusted p-value <0.005	4	1	3	1	86	36	116	59
difference in expression level	154	87	141	99	6620	4603	5906	5317

Table 6.2. Number of differentially expressed lncRNAs and protein coding genes in GSE67916 dataset. TR1 tamoxifen resistant breast cancer cell line number 1, TR4 denotes tamoxifen resistant breast cancer cell line number 4.

To compare DEA results of this dataset to other GEO data sets, we decided to find the common differentially expressed genes between TR1 and TR4. The number of differentially expressed lncRNAs after applying 1.5-fold change and 0.005 p-value cut off was very low, so the common genes were found by intersecting lncRNAs that had a fold change in expression above or below zero. Conversely, protein coding genes were high in number, so we used the 1.5-fold change and <0.005 p-value cut off to find the commonly differentially expressed genes. The shared genes are shown in (Figures 6.3 and 6.4).



ENSG00000203497	PDCD4-AS1
ENSG00000218510	LINC00339
ENSG00000230551	CTB-89H12.4
ENSG00000233937	CTC-338M12.4
ENSG00000240032	RP11-274H2.3
ENSG00000240498	CDKN2B-AS1
ENSG00000245532	NEAT1 ONC
ENSG00000246731	CTD-2514K5.2
ENSG00000260669	AL136419.6
ENSG00000265282	RP11-269G24.4
ENSG00000272711	RP11-259N19.1
ENSG00000282977	PCBP2-OT1

Figure 6.3. Common lncRNAs up-regulated both in GSE67916 tamoxifen resistant cells and our chapter 3 TAMR cells.

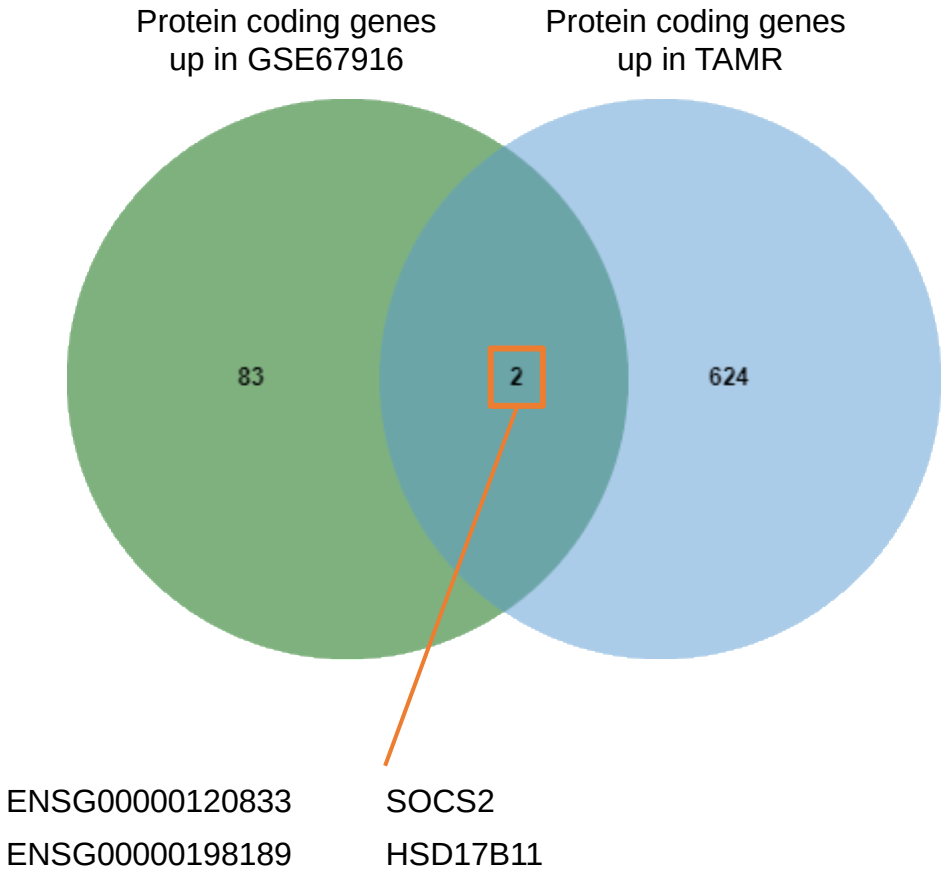


Figure 6.4 Common protein coding genes up-regulated both in GSE67916 tamoxifen resistant cells and our chapter 3 TAMR cells.

6.2.2.2 GSE27473

The second study aimed to test the global transcriptomic effect of oestrogen receptor silencing on oestrogen receptor positive breast cancer cell line. Three samples of wild type MCF-7 cells and 3 samples of MCF-7 cells depleted of ERS1 (oestrogen receptor alpha) gene by siRNA. Raw data were quality assessed and normalised as previously described (Figures 6.5 A, B, C and D). As part of quality control step, samples were clustered relative to each other in respect to the most accounted variations (principal components 1 and 2). It is clear the two sample

groups cluster far from each other (Figure 6.5 E), confirming their biological diversity. The initial DEA results are shown in (Table 6.3).

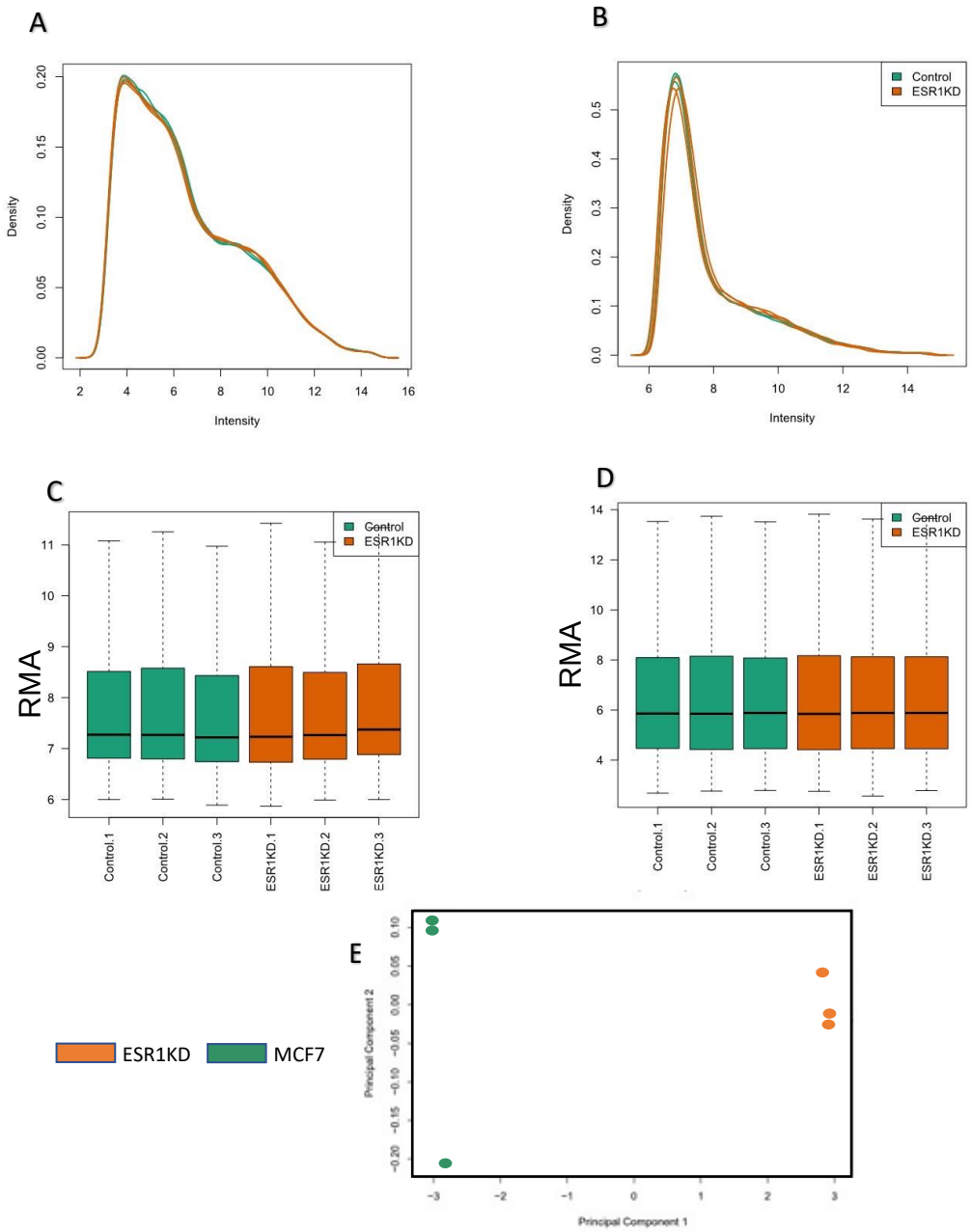
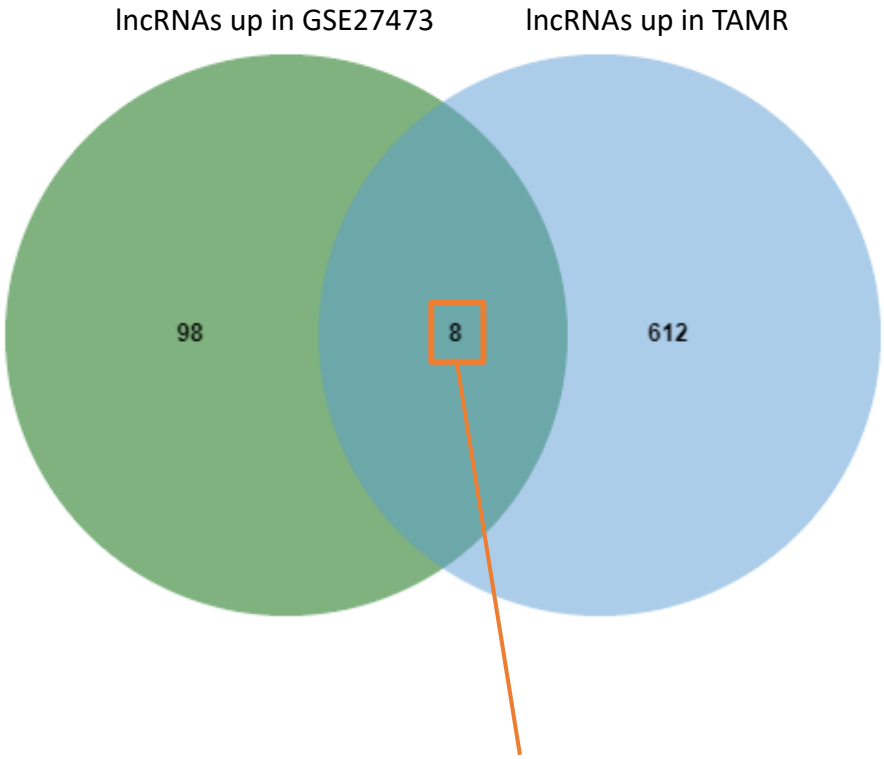


Figure 6.5. GSE27473 raw data quality assessment. After determining gene expression measures, data distribution (A) before and (B) after normalisation were . (C) Box plot of expression values before and (D) after normalisation. (E) Principle component analysis plot showing variability in gene expression between samples, the number in PCA is given for sample identification. Control denotes wild type MCF-7 cells, ESR1KD denotes MCF-7 modified with ESR1 depletion, and RMA denotes robust multichip average.

Selection Criteria	lncRNAs		protein coding genes	
	Up regulated in ESR1KD	Down regulated in ESR1KD	Up regulated in ESR1KD	Down regulated in ESR1KD
log fold change 1.5 and adjusted p-value <0.005	18	14	1181	41
difference in expression level	109	106	6013	4854

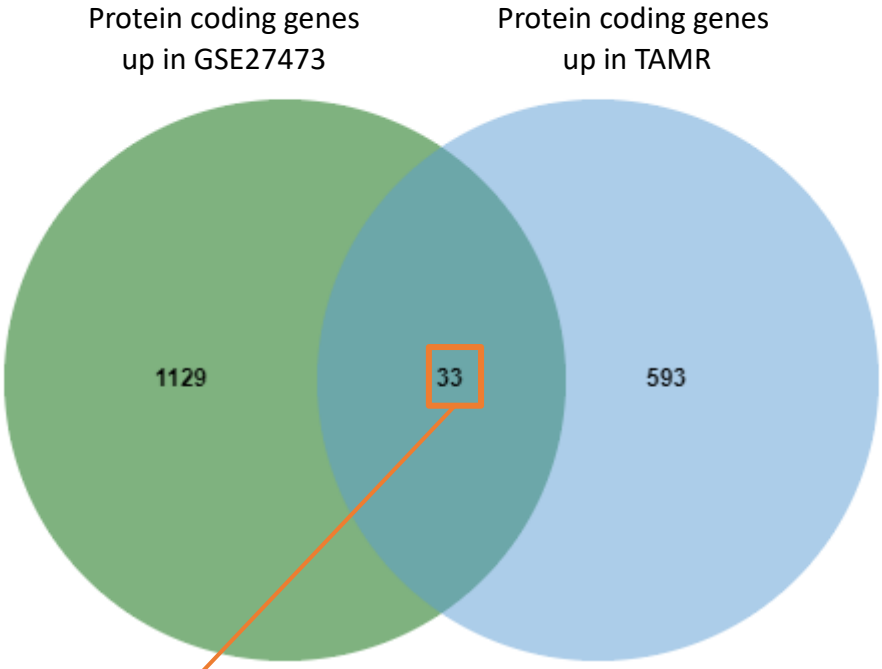
Table 6.3. Number of differentially expressed lncRNAs and protein coding genes in GSE27473 dataset. ESR1KD denotes MCF-7 modified with ESR1 depletion

To formulate the final list of differentially expressed genes from the initial DEA results, we decided to include differentially expressed lncRNAs regardless of the fold change and p-value cut off, accounting only for the fact that a gene was up or down regulated in ESR1 silenced group. However, with protein coding genes we used the 1.5-fold change and 0.005 p-value cut off to form the list of genes for downstream analysis. These genes were compared to genes upregulated in TAMR samples, the genes laying in the intersection as common genes between both ate shown and listed in (Figures 6.6 and 6.7).



ENSG00000157306	RP11-66N24.4
ENSG00000182165	TP53TG1
ENSG00000203497	PDCD4-AS1
ENSG00000218510	LINC00339
ENSG00000234912	LINC00338
ENSG00000240498	CDKN2B-AS1
ENSG00000261087	KB-1460A1.5
ENSG00000267296	CEBPA-AS1

Figure 6.6. Common lncRNAs up-regulated both in GSE27473 ESR1 depleted MCF-7 samples and our chapter 3 TAMR cells.



ENSG00000041353	RAB27B
ENSG00000060656	PTPRU
ENSG00000073792	IGF2BP2
ENSG00000088538	DOCK3
ENSG00000100504	PYGL
ENSG00000105388	CEACAM5
ENSG00000110660	SLC35F2
ENSG00000114251	WNT5A
ENSG00000128512	DOCK4
ENSG00000131389	SLC6A6
ENSG00000131711	MAP1B
ENSG00000132334	PTPRE
ENSG00000133083	DCLK1
ENSG00000138640	FAM13A
ENSG00000138795	LEF1
ENSG00000141404	GNAL
ENSG00000148700	ADD3
ENSG00000151690	MFSD6
ENSG00000153250	RBMS1
ENSG00000154822	PLCL2
ENSG00000162409	PRKAA2
ENSG00000162496	DHRS3
ENSG00000168675	LDLRAD4
ENSG00000175745	NR2F1
ENSG00000175832	ETV4
ENSG00000178726	THBD
ENSG00000183421	RIPK4
ENSG00000185650	ZFP36L1
ENSG00000196139	AKR1C3
ENSG00000196954	CASP4
ENSG00000213654	GPSM3
ENSG00000221988	PPT2
ENSG00000244274	DBNDD2

Figure 6.7. Common protein coding genes up-regulated both in GSE27473 ESR1 depleted MCF-7 samples and our chapter 3 TAMR cells.

6.2.2.3 GSE124647

This data set included microarray data of 140 biopsied metastatic tissues taken from stage IV breast cancer patients. Referring to the clinical data provided by the submitters, samples were split into two groups depending on progress free survival (PFS), a mean of 5.53 months were chosen as a reference of comparison. 70 samples had a PFS more than 5.53 months, they were assigned to the high PFS (Hpfs) group, and 70 samples had a PFS less than 5.53 months, they were assigned to the low PFS (Lpfs) group. CEL files were loaded, and raw data were extracted, quality assessed and normalised as previously described (Figures 6.8 A, B, C and D). As part of quality control step, samples were clustered relative to each other in respect to the most accounted variations (principal components 1 and 2) (Figure 6.8.E). However, it appears to be hard to separate samples according to their assigned group. No distinguishable clustering pattern could be recognised as distances between samples were closer than expected. Since this dataset were formed of sequenced tissue samples, this observation can be explained by genotypic diversity between patients.

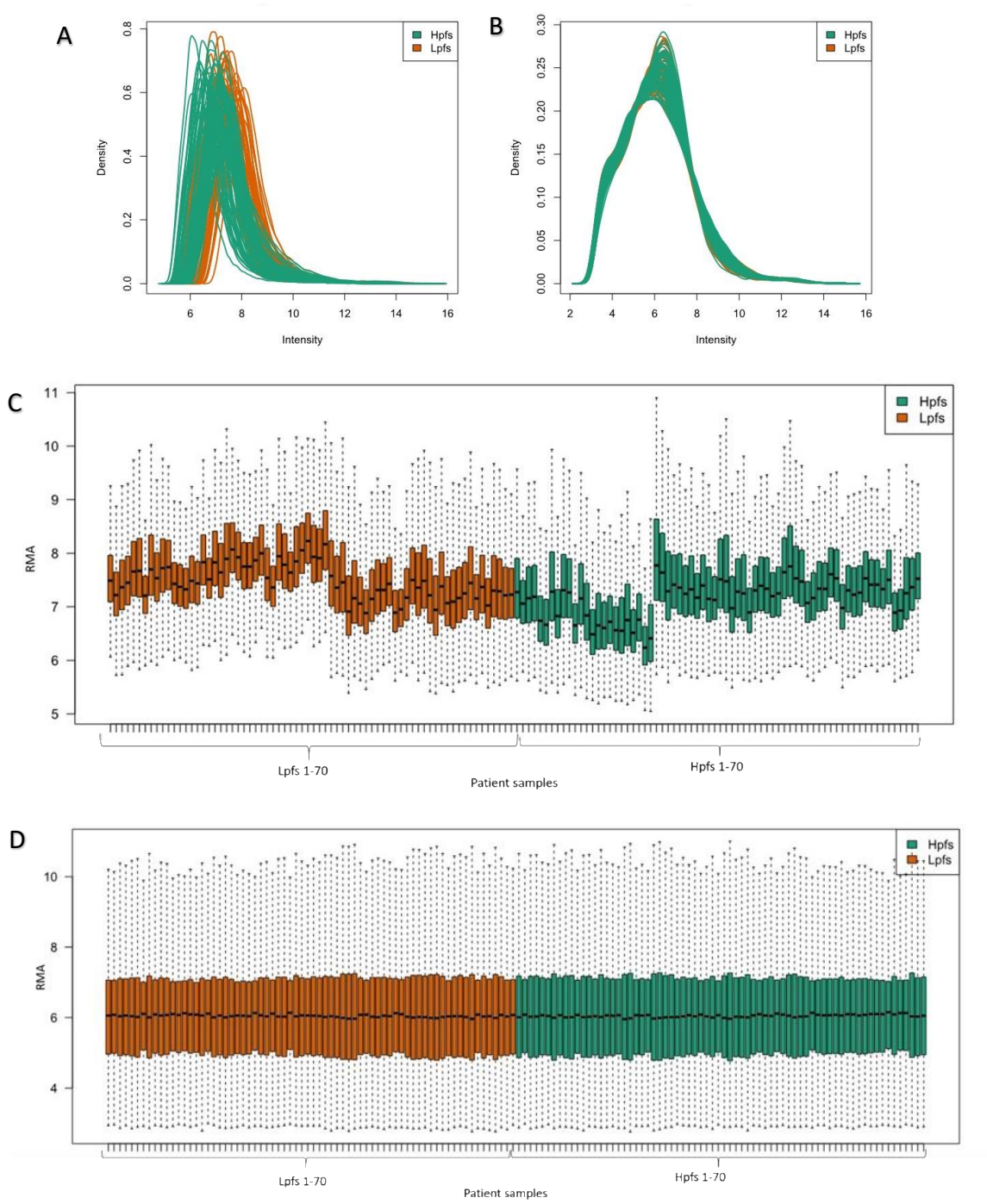


Figure 6.8. GSE124647 raw data quality assessment. After determining gene expression measures, data distribution (A) before and (B) after normalisation were . (C) Box plot of expression values before and (D) after normalisation. Continued on next page

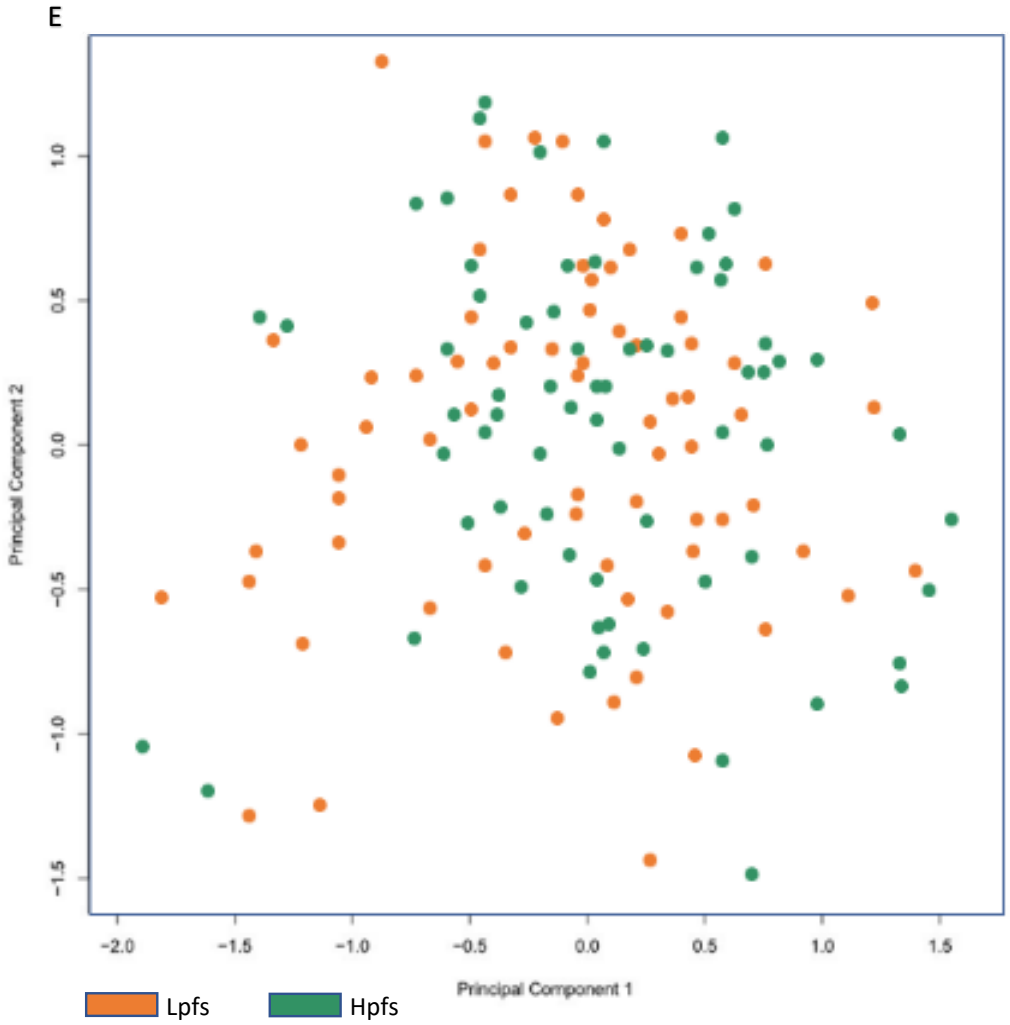


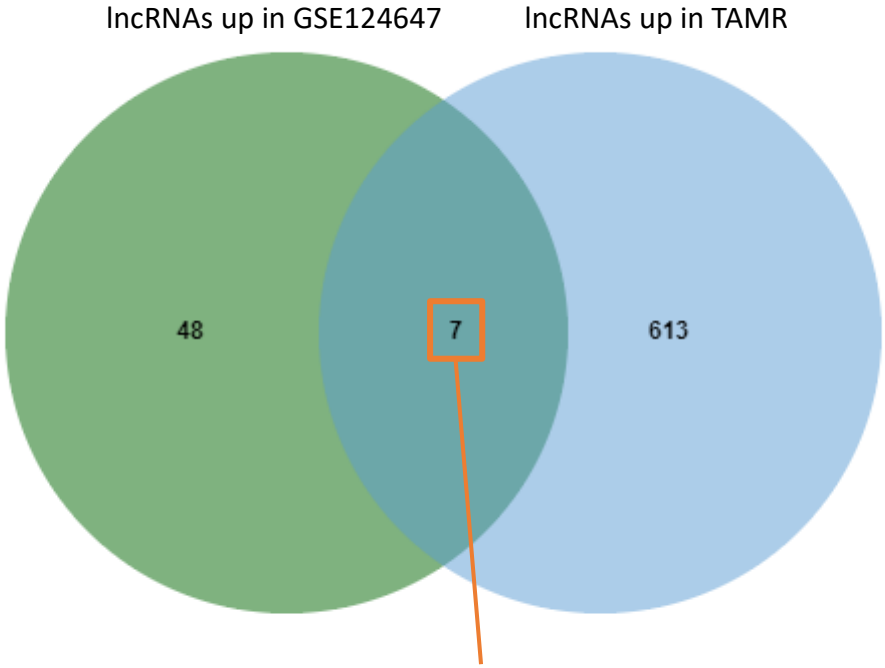
Figure 6.8. GSE124647 raw data quality assessment. Continued (E) Principle component analysis plot showing variability in gene expression between samples, the number in PCA is given for sample identification. Lpfs denotes low progression free survival, and Hpfs denotes high progression free survival, RMA denotes robust multichip average.

To perform DEA, Hpfs group were set as a reference and the results are summarised in Table 6.4.

Selection Criteria	lncRNAs		protein coding genes	
	Up regulated in Lpfs	Down regulated Lpfs	Up regulated Lpfs	Down regulated Lpfs
adjusted p-value <0.005	1	6	81	319
difference in expression level	57	83	2464	4493

Table 6.4. Number of differentially expressed lncRNAs and protein coding genes in GSE67916 dataset. TR1 tamoxifen resistant breast cancer cell line number 1, TR4 denotes tamoxifen resistant breast cancer cell line number 4.

To produce the list of differentially expressed lncRNAs and protein coding genes from the initial DEA results, we decided to include differentially expressed lncRNAs regardless of the fold change and p-value cut off, considering only whether a gene was up or down regulated in the Lpfs group. Also, for protein coding genes we used 0.005 adjusted p-value cut off to form the list of genes for downstream analysis. These genes were compared to genes upregulated in TAMR samples, the genes laying in the intersection as common genes between both ate shown and listed in (Figures 6.9 and 6.10).



ENSG00000182165	TP53TG1
ENSG00000224078	SNHG14
ENSG00000230551	CTB-89H12.4
ENSG00000233937	CTC-338M12.4
ENSG00000262879	RP11-156P1.3
ENSG00000272821	CTA-384D8.36
ENSG00000282977	PCBP2-OT1

Figure 6.9. Common lncRNAs up-regulated both in GSE124647 low progress free survival samples and our chapter 3 TAMR cells.

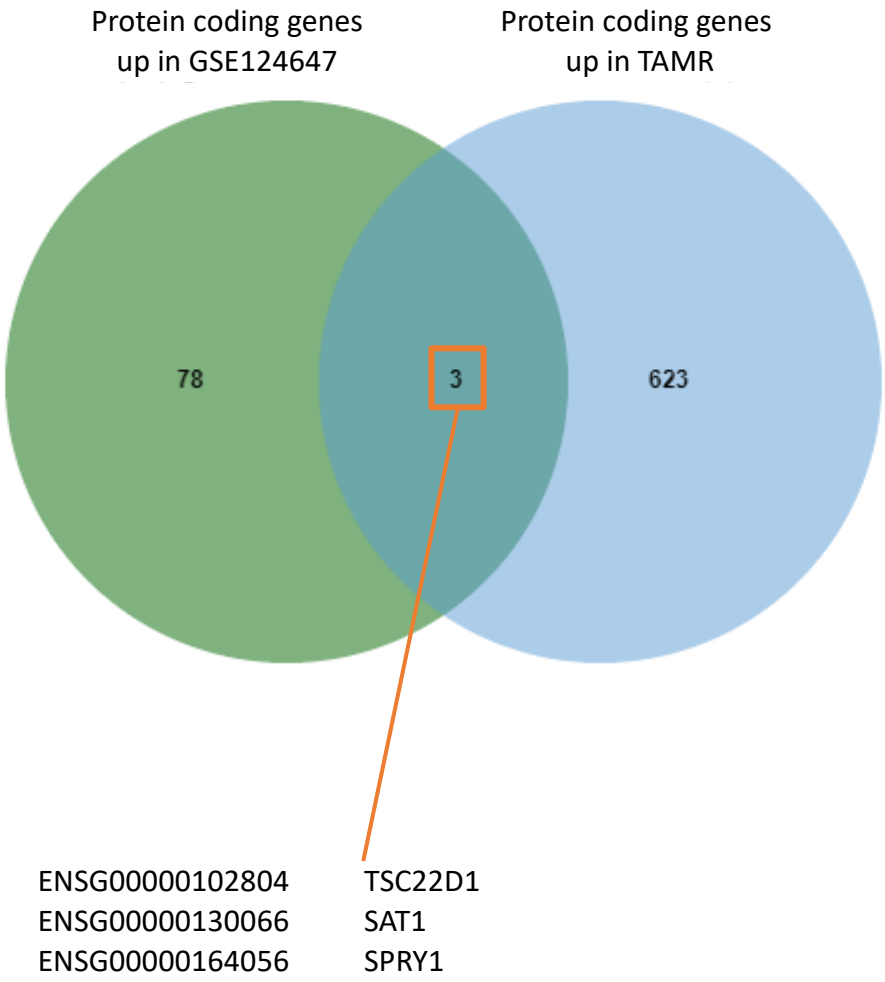


Figure 6.10. Common protein coding genes up-regulated both in GSE124647 low progress free survival samples and our chapter 3 TAMR cells.

6.2.2.4 GSE58644

This data set included microarray data of 321 breast cancer tissues taken from patients of different subtypes and stages of breast cancer. The clinical data file provided by the submitters, included a lot of details about patients, sample collection and study protocol. The samples were split into two groups depending on their oestrogen receptor status. 251 oestrogen receptor positive samples were

compared to 70 oestrogen receptor negative samples. CEL files were loaded, and raw data were extracted, quality assessed and normalised as previously described (Figures 6.11 A, B, C and D). As part of quality control step, samples were clustered relative to each other in respect to the most accounted variations (principal components 1 and 2) (Figure 6.11. E). No distinguishable clustering pattern could be recognised as distances between samples were closer than expected. Since this dataset were formed of sequenced tissue samples, this observation can be explained by genotypic diversity between patients.

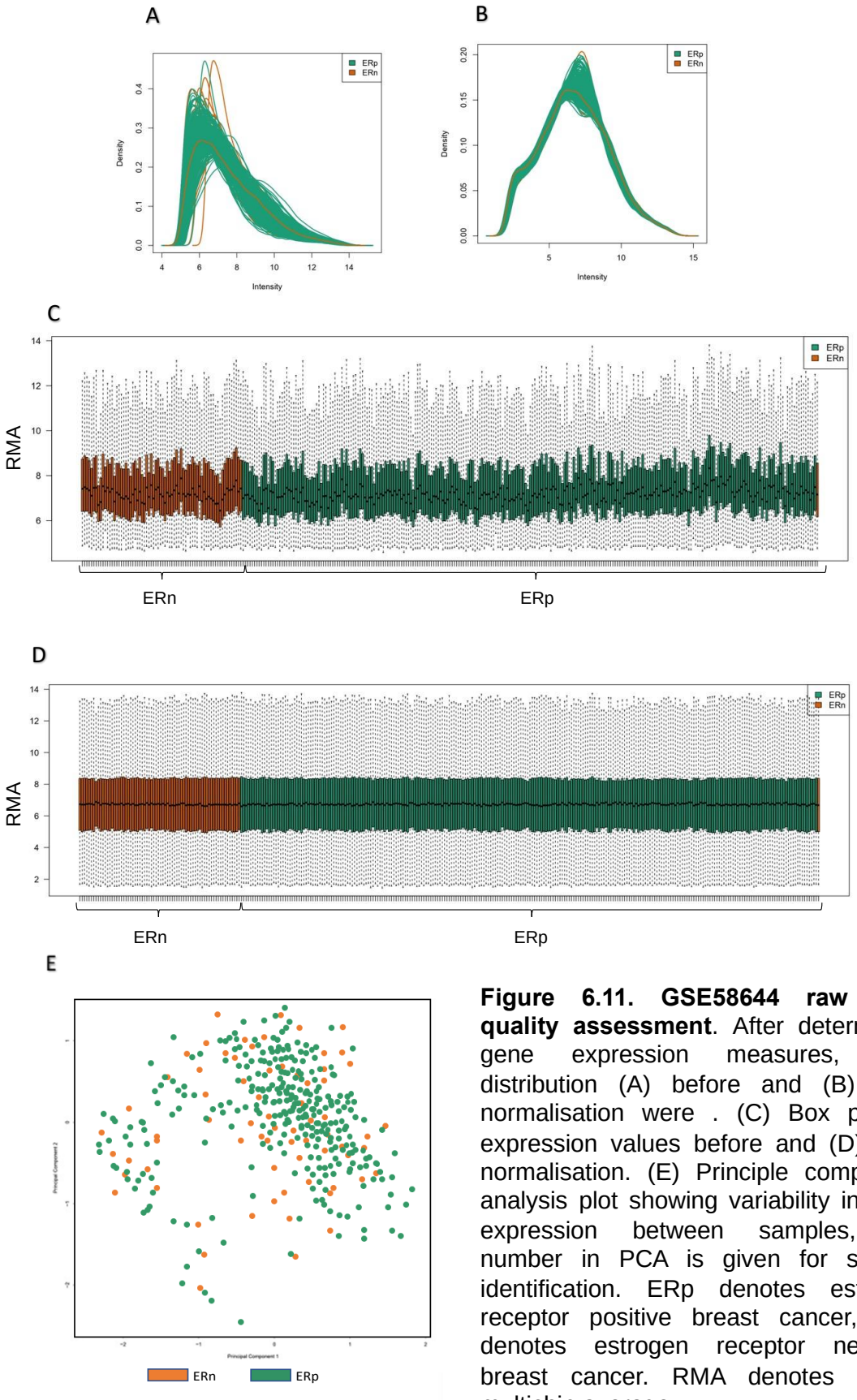


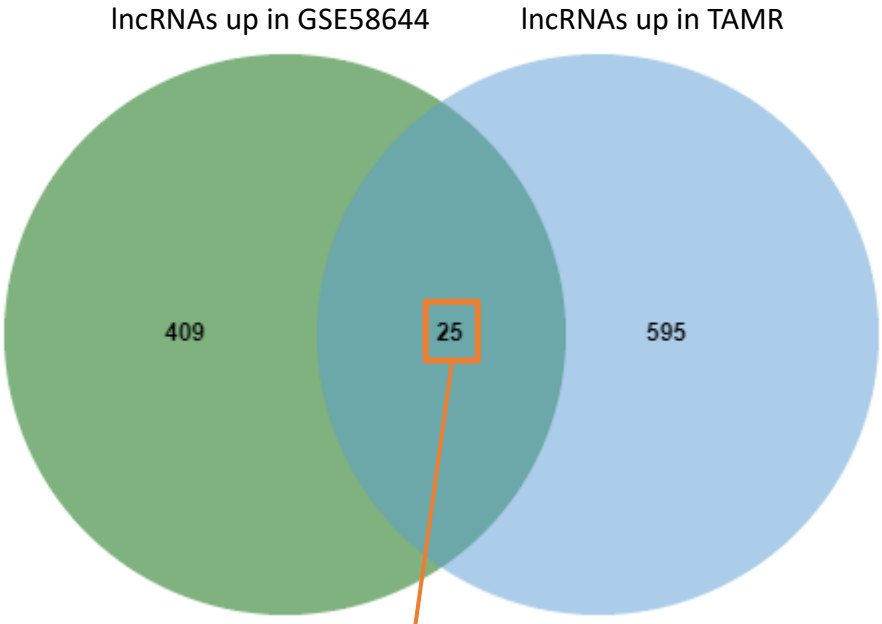
Figure 6.11. GSE58644 raw data quality assessment. After determining gene expression measures, data distribution (A) before and (B) after normalisation were . (C) Box plot of expression values before and (D) after normalisation. (E) Principle component analysis plot showing variability in gene expression between samples, the number in PCA is given for sample identification. ERp denotes estrogen receptor positive breast cancer, ERn denotes estrogen receptor negative breast cancer. RMA denotes robust multichip average

To perform DEA, oestrogen receptor positive group were set as a reference. A summary of number of lncRNAs and protein coding genes differentially expressed between oestrogen receptor positive and negative samples is shown in (Table 6.5).

Selection criteria	lncRNAs		protein coding genes	
	Up regulated in ERn	Down regulated in ERn	Up regulated in ERn	Down regulated in ERn
adjusted p-value <0.005	53	59	869	1000
difference in expression level	434	340	6171	4622

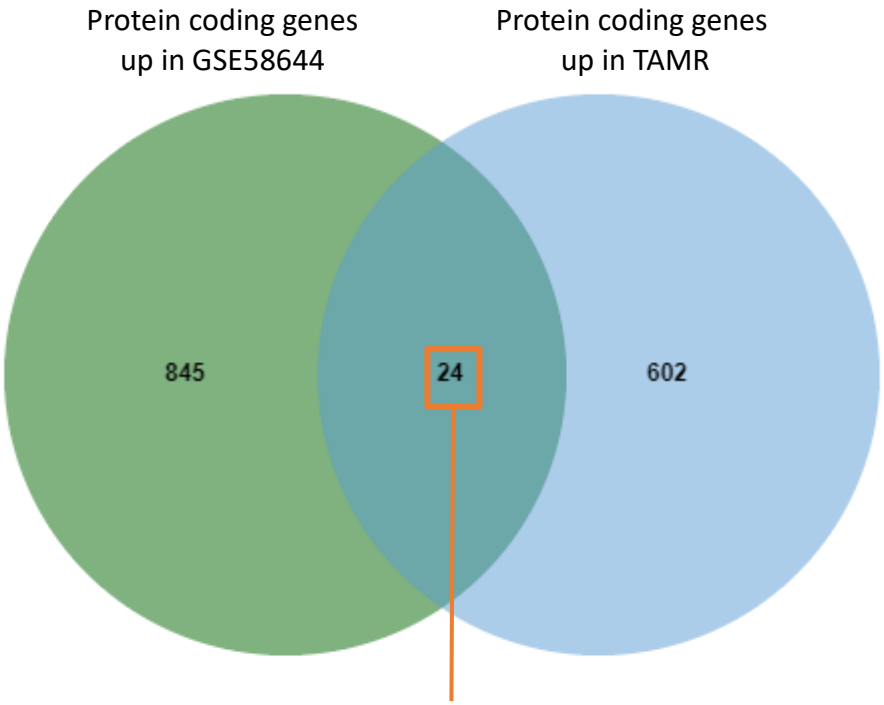
Table 6.5. Number of differentially expressed lncRNAs and protein coding genes in GSE58644 dataset. ERn denotes estrogen receptor negative breast cancer

A list of differentially expressed lncRNAs and protein coding genes was then formulated from the DEA results. We Included differentially expressed lncRNAs regardless of the fold change and p-value cut off while for protein coding genes we only used 0.005 adjusted p-value cut off to form the list of genes for downstream analysis. These genes were compared to genes upregulated in TAMR samples, the genes laying in the intersection as common genes between both ate shown and listed in (Figures 6.12 and 6.13)



ENSG00000082929	C4orf6
ENSG00000170161	RP11-262H14.4
ENSG00000174365	SNHG11
ENSG00000177410	ZFAS1
ENSG00000185904	LINC00839
ENSG00000197670	RP4-724E16.2
ENSG00000206195	AP000525.9
ENSG00000224078	SNHG14
ENSG00000228794	LINC01128
ENSG00000231312	AC007246.3
ENSG00000232973	CYP1B1-AS1
ENSG00000235865	GSN-AS1
ENSG00000237491	RP11-206L10.9
ENSG00000244041	LINC01011
ENSG00000249087	C1orf213
ENSG00000253552	HOXA-AS2
ENSG00000255198	SNHG9
ENSG00000257621	RP11-349A22.5
ENSG00000258655	ARHGAP5-AS1
ENSG00000260669	AL136419.6
ENSG00000261713	SSTR5-AS1
ENSG00000264247	LINC00909
ENSG00000267249	RP11-973H7.3
ENSG00000272142	RP11-428J1.5
ENSG00000278175	GLIDR

Figure 6.12. Common lncRNAs up-regulated both in GSE58644 estrogen receptor negative samples and our chapter 3 TAMR cells.

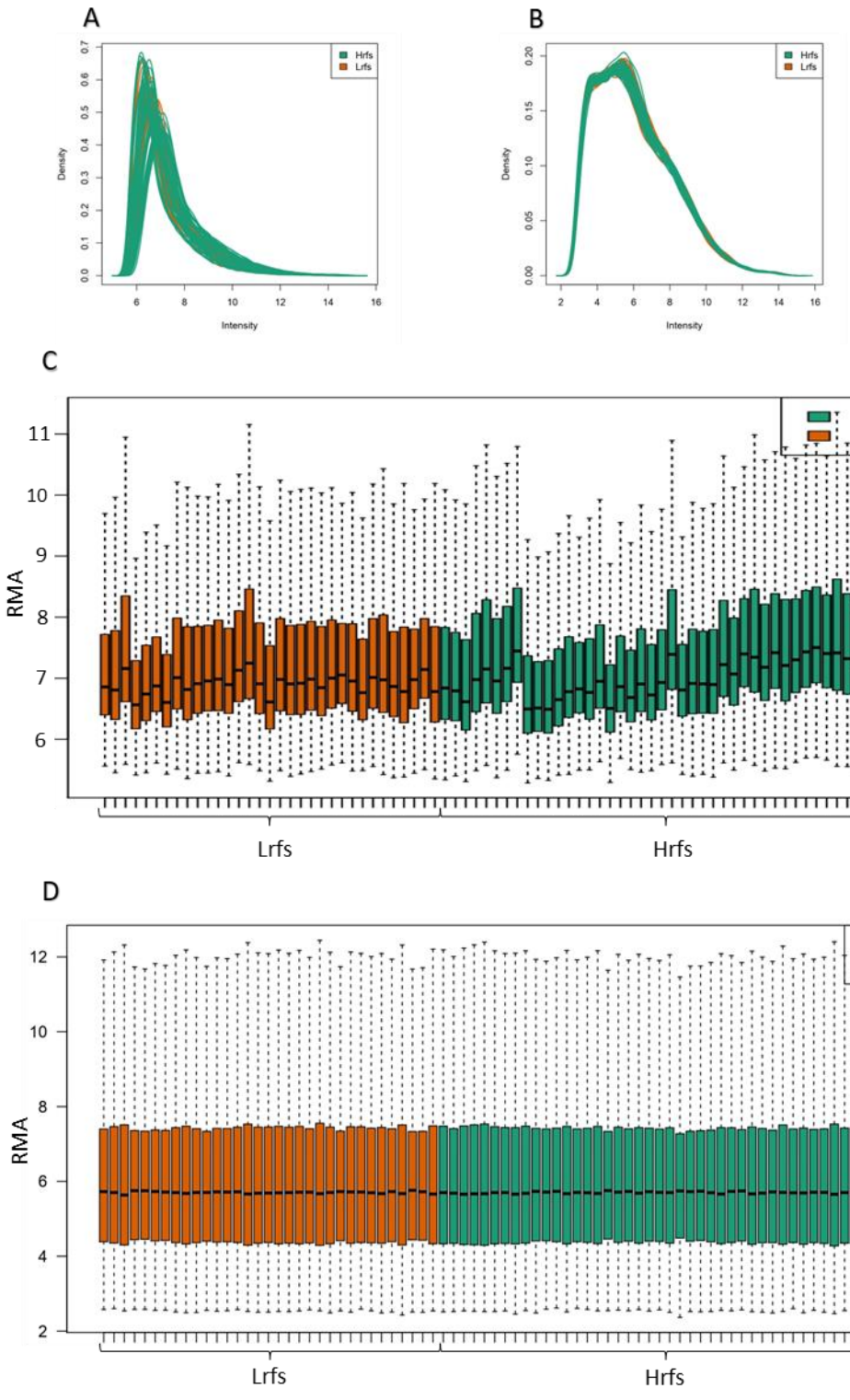


ENSG00000067167	TRAM1
ENSG00000088881	EBF4
ENSG00000101230	ISM1
ENSG00000102265	TIMP1
ENSG00000107159	CA9
ENSG00000110693	SOX6
ENSG00000111341	MGP
ENSG00000121039	RDH10
ENSG00000135363	LMO2
ENSG00000138759	FRAS1
ENSG00000147251	DOCK11
ENSG00000148468	FAM171A1
ENSG00000154065	ANKRD29
ENSG00000162576	MXRA8
ENSG00000167772	ANGPTL4
ENSG00000168032	ENTPD3
ENSG00000168487	BMP1
ENSG00000168685	IL7R
ENSG00000175329	ISX
ENSG00000175899	A2M
ENSG00000175928	LRRN1
ENSG00000181104	F2R
ENSG00000182568	SATB1
ENSG00000196139	AKR1C3

Figure 6.13. Common protein-coding genes up-regulated both in GSE58644 estrogen receptor negative samples and our chapter 3 TAMR cells.

6.2.2.4 GSE9195

This data set included microarray data of 77 oestrogen receptor positive breast cancer tissues taken from patients of different diagnostic stages of breast cancer. The clinical data file provided by the submitters, included a lot of details about patients, sample collection and study protocol. samples were split into two groups depending on the value of their relapse free survival time (RFS). Mean RFS was determined as 2835 days, and samples were classified as having RFS higher than this number (high RFS group) or lower than the mean (low RFS group). As a result, 44 high RFS samples were compared to 33 low RFS samples. Initially, CEL files were loaded, and raw data were extracted, quality assessed and normalised as previously described (Figures 6.14 A, B, C and D). As part of quality control step, samples were clustered relative to each other in respect to the most accounted variations (principal components 1 and 2). As shown in Figure 6.14.E, no distinguishable clustering pattern can be recognised as distances between samples were closer than expected.


A
Hrfs
Lrfs
Density
Intensity
B
Hrfs
Lrfs
Density
Intensity
C
Hrfs
Lrfs
RMA
Lrfs
Hrfs
D
Hrfs
Lrfs
RMA
Lrfs
Hrfs

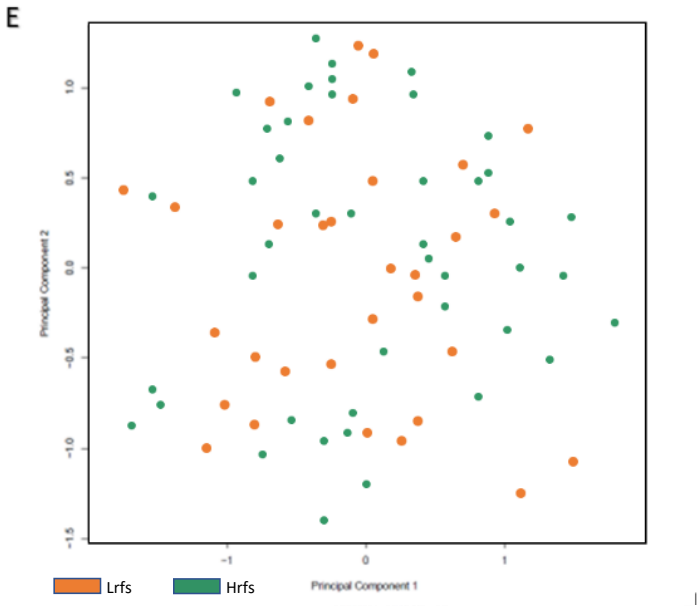


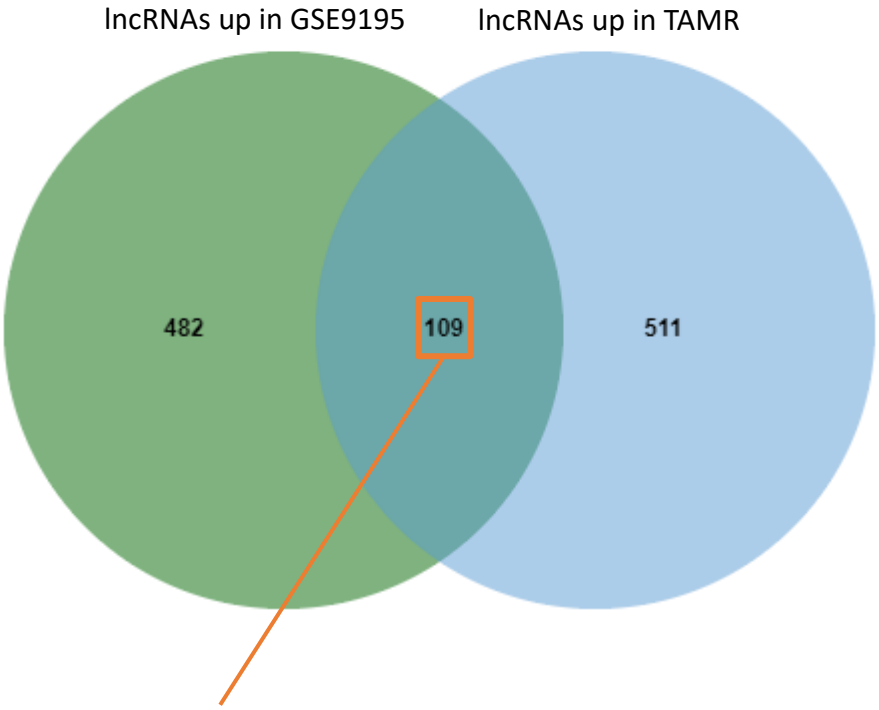
Figure 6.14. GSE9195 raw data quality assessment. After determining gene expression measures, data distribution (A) before and (B) after normalisation were . (C) Box plot of expression values before and (D) after normalisation. (E) Principle component analysis plot showing variability in gene expression between samples, the number in PCA is given for sample identification. Lrfs denotes low relapse free survival, and Hrfs denotes high relapse free survival,. RMA denotes robust multichip average.

Since this dataset were formed of sequenced tissue samples, this observation can be explained by genotypic diversity between patients. To perform DEA, higher RFS group of samples were set as a reference. A summary of number of lncRNAs and protein coding genes differentially expressed between oestrogen receptor positive and negative samples is shown in (Table 6.6).

Selection Criteria	lncRNAs		protein coding genes	
	Up regulated in Lrfs	Down regulated Lrfs	Up regulated Lrfs	Down regulated Lrfs
adjusted p-value <0.005	0	0	383	202
difference in expression level	657	648	9204	9420

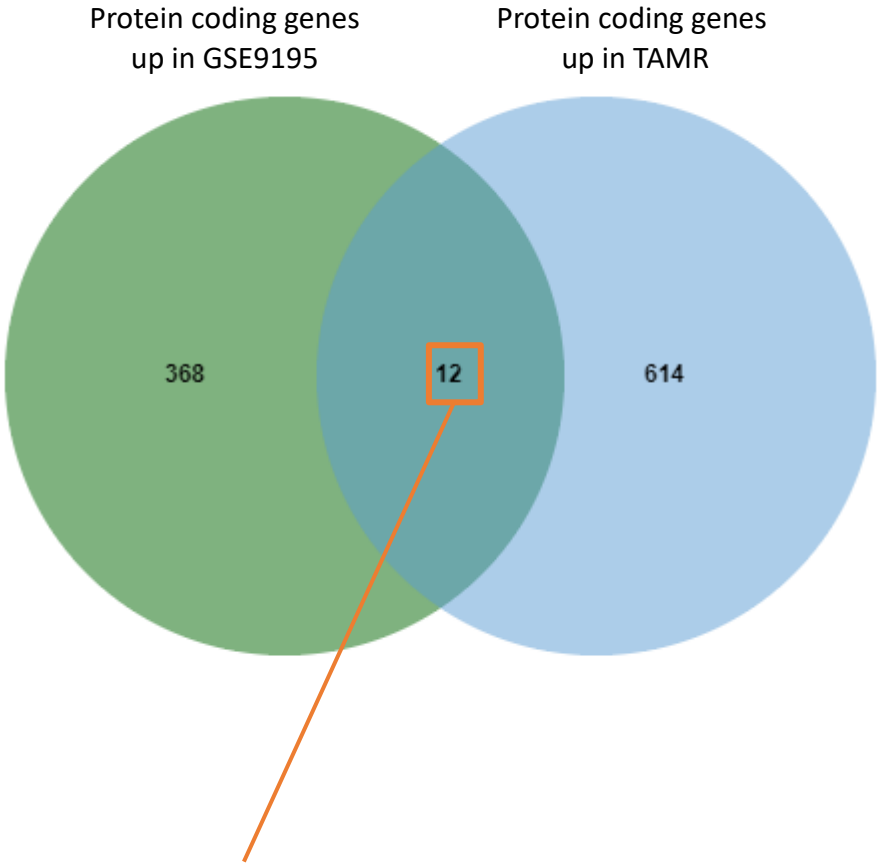
Table 6.6. Number of differentially expressed lncRNAs and protein coding genes in GSE9195 dataset. Lrfs denotes low release free survival breast cancer

A list of differentially expressed lncRNAs and protein coding genes was formulated from the DEA results. Differentially expressed lncRNAs were included regardless of the fold change and p-value cut off, while for protein coding genes used 0.005 p-value was used as a cut off to form the list of genes for downstream analysis. These genes were compared to genes upregulated in TAMR samples, the genes laying in the intersection as common genes between both ate shown and listed in (Figures 6.15 and 6.16)



ENSG00000153363	LINC00467	ENSG00000233184	RP11-421L21.3	ENSG00000257894	RP1-78O14.1	ENSG00000267280	RP11-332H18.4
ENSG00000157306	RP11-66N24.4	ENSG00000233429	HOTAIRM1	ENSG00000258634	RP4-773N10.4	ENSG00000267296	CEBPA-AS1
ENSG00000163364	LINC01116	ENSG00000233461	RP11-295G20.2	ENSG00000258701	LINC00638	ENSG00000267322	SCARNA17
ENSG00000167920	TMEM99	ENSG00000233621	LINC01137	ENSG00000258708	SLC25A21-AS1	ENSG00000267575	CTC-459F4.3
ENSG00000170161	RP11-262H14.4	ENSG00000233937	CTC-338M12.4	ENSG00000259248	USP3-AS1	ENSG00000268205	CTC-444N24.11
ENSG00000170919	TPT1-AS1	ENSG00000234684	SDCBP2-AS1	ENSG00000259642	C15orf37	ENSG00000268713	CTC-444N24.8
ENSG00000180769	WDFY3-AS2	ENSG00000237310	GS1-124K5.4	ENSG00000260231	JHDM1D-AS1	ENSG00000270673	YTHDF3-AS1
ENSG00000182165	TP53TG1	ENSG00000237491	RP11-206L10.9	ENSG00000260317	RP11-48B3.4	ENSG00000270820	RP11-355B11.2
ENSG00000189223	PAX8-AS1	ENSG00000237753	AC079922.3	ENSG00000260329	RP11-412D9.4	ENSG00000272086	CTD-2186M15.3
ENSG00000196756	SNHG17	ENSG00000237807	RP11-400K9.4	ENSG00000260669	AL136419.6	ENSG00000272288	RP11-140K17.3
ENSG00000206195	AP000525.9	ENSG00000237976	RP11-126K1.6	ENSG00000260793	RP5-882C2.2	ENSG00000272335	RP11-53O19.3
ENSG00000214293	RSBN1L-AS1	ENSG00000239911	PRKAG2-AS1	ENSG00000260912	RP11-363E7.4	ENSG00000272341	RP1-151F17.2
ENSG00000224281	SLC25A5-AS1	ENSG00000242086	LINC00969	ENSG00000260966	RP11-690D19.3	ENSG00000272841	RP3-428L16.2
ENSG00000224870	RP4-758J18.2	ENSG00000244879	GABPB1-AS1	ENSG00000261512	RP11-46D6.1	ENSG00000273314	RP5-1136G13.2
ENSG00000226686	AC012309.5	ENSG00000245213	RP11-10K16.1	ENSG00000261654	RP11-96K19.4	ENSG00000273344	PAXIP1-AS1
ENSG00000228506	RP11-98I9.4	ENSG00000245532	NEAT1	ENSG00000261754	CTC-523E23.1	ENSG00000273456	RP11-686O6.2
ENSG00000228630	HOTAIR	ENSG00000246731	CTD-2514K5.2	ENSG00000261824	LINC00662	ENSG00000273680	NA
ENSG00000229152	ANKRD10-IT1	ENSG00000247271	ZBED5-AS1	ENSG00000262580	RP11-334C17.5	ENSG00000273888	FRMD6-AS1
ENSG00000229953	RP11-284F21.7	ENSG00000247950	SEC24B-AS1	ENSG00000262879	RP11-156P1.3	ENSG00000275202	NA
ENSG00000229970	AC007128.1	ENSG00000250988	RP11-752G15.6	ENSG00000264112	RP11-159D12.2	ENSG00000275234	NA
ENSG00000230551	CTB-89H12.4	ENSG00000251562	MALAT1	ENSG00000265282	RP11-269G24.4	ENSG00000275557	NA
ENSG00000231074	HCG18	ENSG00000252690	SCARNA15	ENSG00000266916	CTD-3064H18.1	ENSG00000276116	FUT8-AS1
ENSG00000231312	AC007246.3	ENSG00000253200	RP11-582J16.5	ENSG00000266962	RP11-400F19.6	ENSG00000277534	NA
ENSG00000231721	LINC-PINT	ENSG00000253716	RP13-582O9.5	ENSG00000267106	C19orf82	ENSG00000278156	TSC22D1-AS1
ENSG00000232533	AC093673.5	ENSG00000253982	CTD-2336O2.1	ENSG00000267152	CTD-2528L19.6	ENSG00000278175	GLIDR
ENSG00000232677	LINC00665	ENSG00000257621	RP11-349A22.5	ENSG00000267216	AC010642.1	ENSG00000278291	NA
ENSG00000232956	SNHG15	ENSG00000257702	LBX2-AS1	ENSG00000267254	CTD-2162K18.5	ENSG00000280206	NA
						ENSG00000283103	NA

Figure 6.15. Common lncRNAs up-regulated both in GSE9195 low relapse free survival samples and our chapter 3 TAMR cells.



ENSG00000110660	SLC35F2
ENSG00000113448	PDE4D
ENSG00000118263	KLF7
ENSG00000133687	TMTC1
ENSG00000135363	LMO2
ENSG00000156113	KCNMA1
ENSG00000164116	GUCY1A3
ENSG00000166002	SMCO4
ENSG00000166173	LARP6
ENSG00000169851	PCDH7
ENSG00000171004	HS6ST2
ENSG00000198121	LPAR1

Figure 6.16 Common protein coding genes up-regulated both in GSE9195 low relapse free survival samples and our chapter 3 TAMR cells.

Finally, tamoxifen response related up-regulated lncRNAs and protein coding genes from all GEO studies were all related to each other in a Venn diagram (Figures 6.17 and 6.18); to find what genes are common between them all. Only one lncRNA (ENSG00000196299 ZNRD1ASP) was upregulated in all the data sets. Searching literature about this gene was trick as its annotation and

classification keep on changing, there were about 10 synonymous of its ensemble ID (C6orf12, Em:AB023056.3, HCG8, HCGVIII, HCGVIII-1, HTEX4, NCRNA00171, ZNRD1-AS, ZNRD1-AS1, ZNRD1AS), its biotype also went prom a processed transcript to a lncRNA to a pseudogene (Ensemble, 2022), few studies were found concerning this gene role in carcinogenesis (Wang et al., 2017; Ba et al., 2021). Since none were related to endocrine positive breast cancer(H. W. Kim *et al.*, 2020), this data while very interesting, it inspires further *in-silico* and *in-vitro* sequence and functional analysis of this gene.

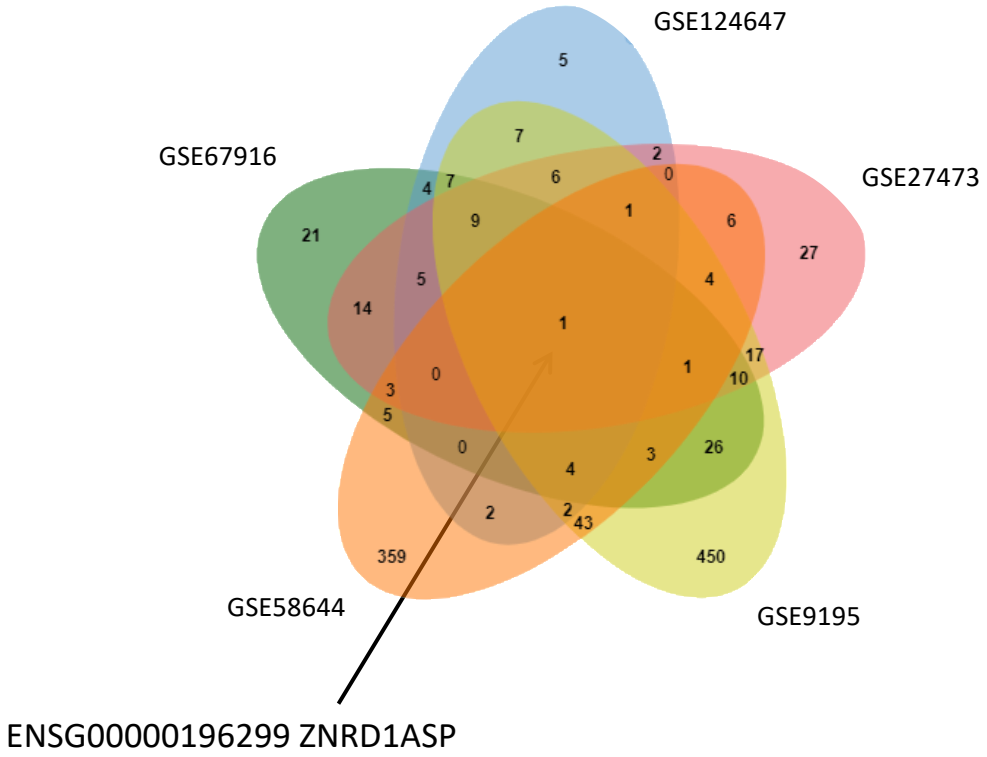


Figure 6.17 Common lncRNAs up-regulated in all GEO analysed datasets

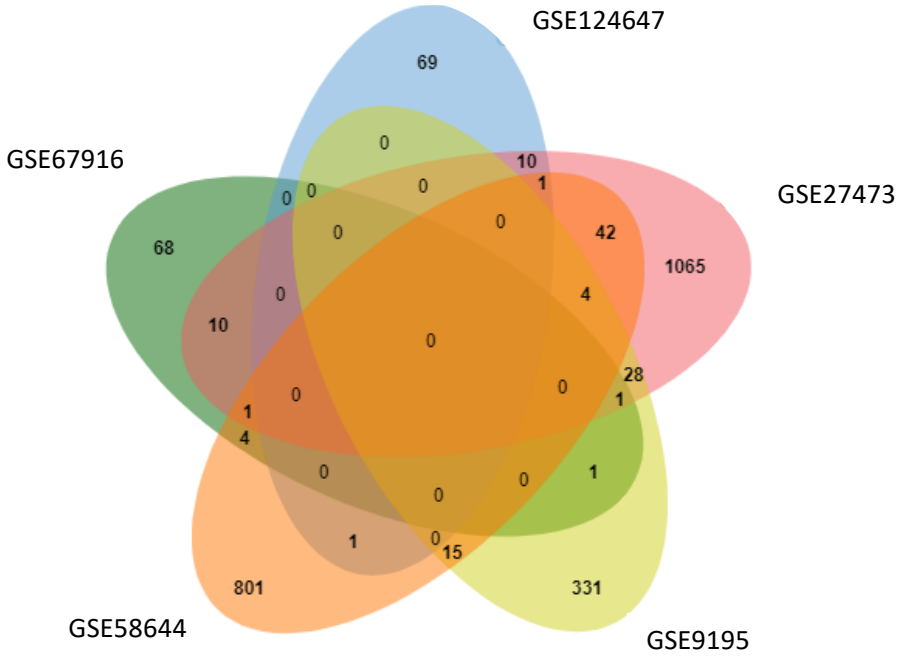


Figure 6.18 Common protein coding genes up-regulated in all GEO analysed datasets

6.3 TCGA data set analysis

The TCGA is another publicly available data set that we wanted to take advantage of. TCGA-BRCA project were accessed through GDC open-access data consortium using TCGAbiolinks package. TCGA-BRCA samples were then classified into different groups based on their barcode identification and clinical metadata. After dividing samples to normal and tumour, tumour was further divided based on PAM50 categorisation system into luminal A, luminal B, HER2-enriched and basal-like breast cancer subtypes. In order to look at genes related to a tamoxifen resistance phenotype, a series of comparisons was then carried out, using the assumption that Basal-like tumours can be used as a proxy for tamoxifen resistant tumours and Luminal A breast cancer samples represent tamoxifen sensitive samples.

Firstly 58 luminal A normal tissues were compared to 412 luminal A tumour tissue samples, the resulting DEA table (Table 6.7) included 1495 protein coding genes and 64 lncRNAs up regulated in luminal A tumours. Secondly, 13 Basal-like normal tissue was compared to 131 Basal-like tumour tissue samples, the resulting DEA table (Table 6.8) included 1221 protein coding genes and 68 lncRNAs up regulated in Basal-like tumours.

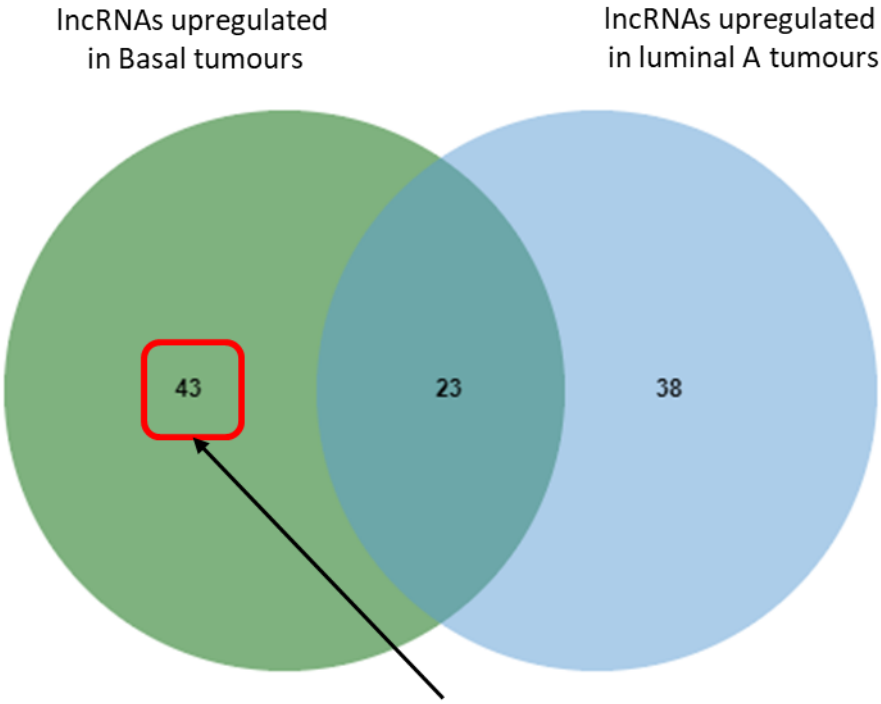
	lncRNAs		protein coding genes	
	Up regulated in basal tumours	Down regulated in basal tumours	Up regulated in basal tumours	Downregulated in basal tumours
log fold change 1.5 and adjusted p-value <0.005	7	21	1189	5793
difference in expression level	65	121	7430	1190

Table 6.7. Number of differentially expressed lncRNAs and protein coding genes between Luminal A normal and tumour samples .

	lncRNAs		protein coding genes	
	Up regulated in basal tumours	Down regulated basal tumours	Up regulated basal tumours	Down regulated basal tumours
log fold change 1.5 and adjusted p-value <0.005	12	37	1189	5793
difference in expression level	68	118	8015	10492

Table 6.8. Number of differentially expressed lncRNAs and protein coding genes between Basal normal and tumour samples .

lncRNAs upregulated in tumour samples in Lum A and Basal-like subtypes were then cross compared in Venn diagram to find common and differentially up and down regulated genes (Figure 6.19 and 6.20). Genes common between both data sets were considered general cancer related genes.



ENSG00000171987 C11orf40
ENSG00000172460 PRSS30P
ENSG00000174365 SNHG11
ENSG00000176728 TTTY14
ENSG00000177640 CASC2
ENSG00000182165 TP53TG1
ENSG00000186493 C5orf38
ENSG00000196166 C8orf86
ENSG00000196366 C9orf163
ENSG00000198496 NBR2
ENSG00000203463 SLC44A4
ENSG00000204091 TDRG1
ENSG00000204661 C5orf60

ENSG00000205861 PCOTH
ENSG00000213057 C1orf220
ENSG00000224975 INE1
ENSG00000225362 CT62
ENSG00000225391 NHEG1
ENSG00000225725 FAM66E
ENSG00000225937 PCA3
ENSG00000225978 HAR1A
ENSG00000226746 SMCR5
ENSG00000228147 HCG22
ENSG00000228630 HOTAIR
ENSG00000229236 TTTY10
ENSG00000230223 ATXN8OS
ENSG00000231133 HAR1B

ENSG00000235947 EGOT
ENSG00000245532 NEAT1
ENSG00000248265 FLJ12825
ENSG00000249054 FAM138D
ENSG00000251562 MALAT1
ENSG00000253521 HPYR1

ENSG00000255794 RMST
ENSG00000260551 PWRN2
ENSG00000262117 BCAR4
ENSG00000271816 BMS1P4
ENSG00000271858 CYB561D2
ENSG00000273018 FAM106A
ENSG00000274454 SLC22A18AS
ENSG00000281670 CABIN1
ENSG00000287151 C2orf27A
ENSG00000288547 C3orf36

Figure 6.19 Common lncRNAs up-regulated in TCGA Basal tumours and Luminal A tumours, surrounded by the red square is the genes up regulated only in Basal tumours (pure Basal).

Figure 6.20. Common protein coding genes up-regulated in TCGA Basal tumours and Luminal A tumours, surrounded by the red square is the genes up regulated only in Basal tumours (pure Basal).

Examples of protein coding genes in this set include: CMA1 which is associated with poor prognosis and immunological infiltrations in gastric cancer (Shi et al., 2020), ANGPT4 involved with promotion of a pro-cancer microenvironment in ovarian cancer (Brunckhorst *et al.*, 2014), and PCK1 and IGFBP1 both of which have roles in colorectal cancer metastasis (Kim et al., 2016; Yamaguchi et al., 2019). Examples of lncRNAs include: H19 and XIST which are both thought to have pan-cancer oncogenic roles (Eldesouki et al., 2022; Ma et al., 2014; Wei et al., 2017; Yu et al., 2020; Q. Zhang et al., 2018), and SNHG8 which has a reported role in ovarian and prostate cancers (Shi et al., 2021; Xuan et al., 2021).

To focus solely on basal-like breast cancer, we opted to include only the set of genes differentially expressed in Basal-like but not in luminal A tumour samples. 43 and 1111 upregulated lncRNAs and protein coding genes respectively, were in this

group. When these were compared to the genes identified in chapter 3, 7 lncRNAs (Figures 6.21) and 64 protein coding genes (Figure 6.22) were in common.

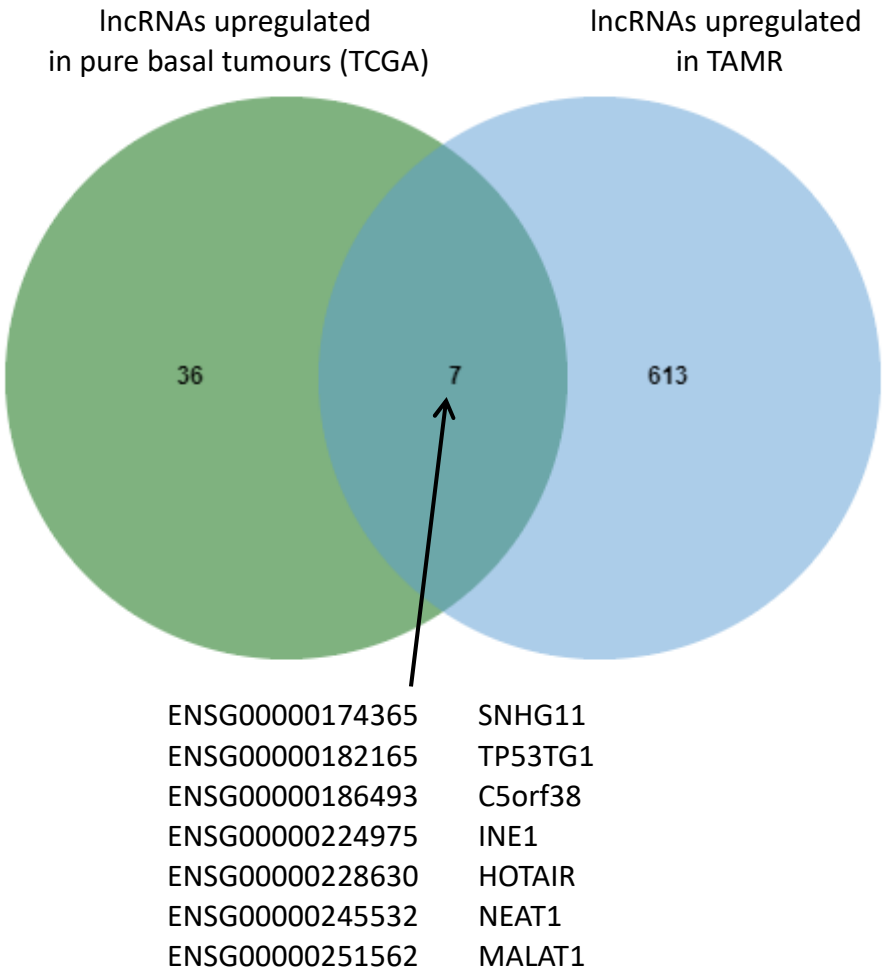
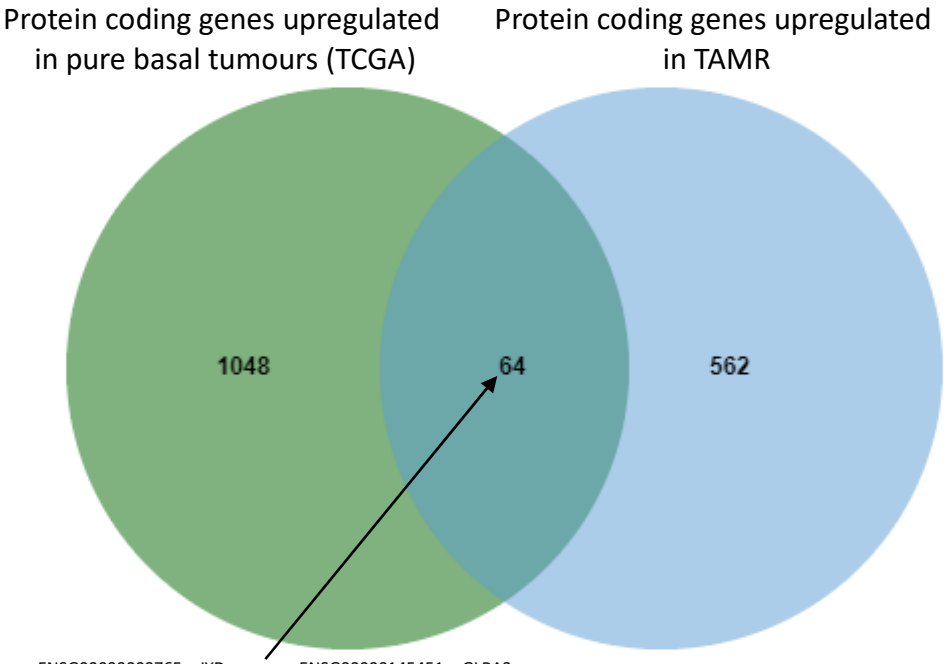


Figure 6.21. Common lncRNAs up-regulated both in TCGA basal tumours and our chapter 3 TAMR cells.



ENSG00000009765	IYD	ENSG00000145451	GLRA3
ENSG00000010310	GIPR	ENSG00000145703	IQGAP2
ENSG00000016402	IL20RA	ENSG00000150527	CTAGE5
ENSG00000041353	RAB27B	ENSG00000151715	TMEM45B
ENSG00000073910	FRY	ENSG00000156113	KCNMA1
ENSG00000079337	RAPGEF3	ENSG00000162009	SSTR5
ENSG00000080493	SLC4A4	ENSG00000165078	CPA6
ENSG00000086548	CEACAM6	ENSG00000165548	TMEM63C
ENSG00000088881	EBF4	ENSG00000166106	ADAMTS15
ENSG00000100003	SEC14L2	ENSG00000166840	GLYATL1
ENSG00000103460	TOX3	ENSG00000168594	ADAM29
ENSG00000105388	CEACAM5	ENSG00000168743	NPNT
ENSG00000106003	LFNG	ENSG00000170786	SDR16C5
ENSG00000106018	VIPR2	ENSG00000171385	KCND3
ENSG00000109472	CPE	ENSG00000171564	FGB
ENSG00000109738	GLRB	ENSG00000172602	RND1
ENSG00000110169	HPX	ENSG00000174343	CHRNA9
ENSG00000113924	HGD	ENSG00000175745	NR2F1
ENSG00000114248	LRRC31	ENSG00000178401	DNAJC22
ENSG00000115461	IGFBP5	ENSG00000179023	KLHDC7A
ENSG00000115844	DLX2	ENSG00000180616	SSTR2
ENSG00000122584	NXPH1	ENSG00000183580	FBXL7
ENSG00000130643	CALY	ENSG00000185008	ROBO2
ENSG00000130997	POLN	ENSG00000186197	EDARADD
ENSG00000133083	DCLK1	ENSG00000187550	SBK2
ENSG00000135111	TBX3	ENSG00000188817	SNTN
ENSG00000137968	SLC44A5	ENSG00000197208	SLC22A4
ENSG00000137976	DNASE2B	ENSG00000197249	SERPINA1
ENSG00000139910	NOVA1	ENSG00000203697	CAPN8
ENSG00000139970	RTN1	ENSG00000203952	CCDC160
ENSG00000140284	SLC27A2	ENSG00000249853	HS3ST5
ENSG00000144355	DLX1	ENSG00000253485	PCDHGA5

Figure 6.22. Common protein coding genes up-regulated both in TCGA basal tumours and our chapter 3 TAMR cells.

Examples include lncRNAs such as ATXN8OS which has been reported previously to have a role in promoting tamoxifen resistance in breast cancer (Zhang et al.,

2021), and TP53TG1 which has been shown independently to be more abundant in tamoxifen resistant cell lines (Muluhngwi and Klinge, 2021).

In addition, 412 Luminal A tumour samples were directly compared to 131 Basal tumour samples (Table 6.9)

	lncRNAs		protein coding genes	
	Up regulated in basal tumours	Down regulated in basal tumours	Up regulated in basal tumours	Down regulated in basal tumours
log fold change 1.5 and adjusted p-value <0.005	24	32	1189	5793
difference in expression level	68	118	8014	10491

Table 6.9. Number of differentially expressed lncRNAs and protein coding genes between Luminal A tumour and Basal tumour samples .

Global breast cancer genotypic profile changes have also been considered by performing DEA comparing gene expression profiles of all TCGA normal breast samples to all breast cancer tumour samples. Finally, only matching solid normal and primary solid tumour samples corresponding to 220 patient samples were compared results for DEA are in (Table 6.10)

	lncRNAs		protein coding genes	
	Up regulated in tumours	Down regulated in tumours	Up regulated in tumours	Down regulated in tumours
log fold change 1.5 and adjusted p-value <0.005	8	25	1141	6195
difference in expression level	67	118	6621	11750

Table 6.10. Number of differentially expressed lncRNAs and protein coding genes between matched normal and tumour samples .

6.4 Discussion

This chapter aimed to utilize publicly available next generation sequencing databases to help nominate better candidate lncRNAs for *in-vitro* study of their role in tamoxifen resistance. A systematic search for relevant datasets was conducted in the GEO database to analyse conditions representative of tamoxifen response in breast cancer. Following the P.I.C.O.T design to formulate the search question, 120 datasets were found, after applying inclusion criteria the total number was down to five studies. At this early stage, a challenge was highlighted as the number of included studies was smaller than expected. Including a larger number of appropriate datasets would be possible by expanding the inclusion criteria and adding more time and resources. On the other hand, selected studies had the element of diversity covering cell lines and patient biopsies, adding more power to the analysis.

Gene lists were generated by running each GEO dataset through the decided DEA pipeline, results included different populations of genes of which we chose lncRNAs and protein-coding genes to undergo more detailed analysis. Conventionally, after differential expression analysis, a threshold is set to produce a statistically significant ranked list of differentially expressed genes. For protein-coding genes, applying the same threshold we used in RNA-seq data in chapter 3 (FDR<0.005 and fold change =1.5), produced a satisfactory number of differentially expressed genes. When applying the same threshold to lncRNA, DEA results were usless as the number of lncRNAs was very low or even nill. So, it was clear that applying the conventional threshold will not work in favour of producing enough lncRNAs from

all datasets. Consequently, we decided to treat this class of genes differently. After differential expression analysis of each of our GEO datasets, fold change and direction was identified for each lncRNA. Genes were assigned to up-regulated or down-regulated groups depending on their fold change sign being positive or negative according to each study design. While this method has many downsides, from a statistical perspective, making the results more vulnerable to false positives and false negative conclusions. Yet, we took into consideration two facts, firstly, lncRNAs are known to have a wider dynamic range of activity and function at low expression, where small changes in expression level produce a pronounced pathological effect (Ahadi, 2021). Secondly, this chapter studies differential gene expression profile similarities between GEO datasets and our tamoxifen resistance model from chapter 3, and lncRNAs selected by the fold change direction method will not be extensively studied directly unless they were consistently dysregulated across different studies and validated independently. Here, we assessed the performance of five microarray datasets, consistent results were obtained from all datasets from bioinformatics quality control processing (e.g., normalization and batch-effect assessment). This was regardless of being sourced by five different centres. Much of this good quality data observed may stem from adhering to the set inclusion criteria especially starting with raw(.CEL) files and restricting platforms used to Illumina and Affymetrix suppliers that algorithms fit perfectly with most DEA pipelines. We observed randomness in the distribution only with samples from opposite groups in clinical datasets (GSE124647, GSE58644 and GSE9195) across PCA axes that represent the degree of variance in the dataset. Higher dimensionality is expected in clinical samples like biopsies sourced from different

individuals. This is due to them featuring gene expression values that are the result of a long-term accumulation of physiological and pathological stimuli. In addition, as RNA is highly unstable with a high rate of decay, biopsy tissue handling, and storage is an extensive source of technical bias and batch effect compared to cell line highly controlled experiments.

Upregulated lncRNAs and protein-coding genes in each GEO study were compared to genes upregulated in TAMR group from chapter 3. The literature showed conflicting statements about combining microarray and RNA-seq data due to fundamental differences in the technical ground between them (Machlus *et al.*, 2010). For this reason and a lot of others, the results of comparing datasets should be interpreted carefully. Many relevant genes have been identified to be common between TAMR and multiple GEO datasets, for example, multi cancer-associated oncogene identified as Erα target lncRNA NEAT1 (Chakravarty *et al.*, 2014; Pang *et al.*, 2019), RP11-156p1.3 (Ali *et al.*, 2020) and CDKN2B-AS1 (Zhuang *et al.*, 2019). TP53TG1 was found commonly upregulated in three GEO datasets in addition to our TAMR data set, it had an oncogenic role in hepatocellular Carcinoma(Lu *et al.*, 2021), pancreatic cancer (Y. Zhang *et al.*, 2019) and glioma (Gao, Qiao and Luo, 2021) but tumour suppressor in lung (Xiao *et al.*, 2018) and breast cancer (Diaz-Lagares *et al.*, 2016; Shao *et al.*, 2020). The same trend of contradiction was also observed in LINC00339 literature (Wu *et al.*, 2022). A number of breast cancer protein-coding oncogenes overlapped between chapter 3 TAMR dataset and multiple GEO datasets such as AKR1C3 (Penning, 2019), and LMO2 (Liu *et al.*, 2017) and SLC35F2 (Winter *et al.*, 2014). Next, we compared

lncRNAs marked upregulated in tamoxifen resistance-related phenotypes in each GEO dataset to other GEO datasets. Interestingly, one lncRNA emerged as mutually upregulated in all geo data sets (ensemble ID: ENSG00000196299), However, it was not differentially expressed in our TAMR dataset by any degree, taking us back to the questionable degree of concordance between microarray and RNA-seq output (Perkins *et al.*, 2014). Moreover, when searching the literature for verified information referencing this gene, we faced the classic lncRNA annotation problem,10 gene name synonyms were found for this ensemble ID and even the biotype classification was diverse from a lncRNA to a pseudogene to a processed transcript. In general, when interpreting the results, we noted that the published literature about lncRNA is lacking and holds contradictions to *in-silico* results. In contrast, search results describing protein-coding genes were much clearer and more decisive, which is very expected, a possible explanation for this might be that the latter being far well-studied and characterised by all means compared to lncRNAs.

The other publicly accessible dataset searched in detail was TCGA-BRCA dataset. Our interest centred around the Basal PAM50 subtype; as it represents triple negative tamoxifen resistance breast cancer phenotype. DEA was applied to normal versus tumour transcriptomic profiles of samples tagged Luminal A and basal breast cancer cases. The output of this comparison (Figures 6.19 and 6.20) was a list of genes purely upregulated in Basal subtype; when compared to TAMR list of genes, many lncRNAs (e.g., HOTAIR (Xue *et al.*, 2016) and MALAT1 (Huang *et al.*, 2016)) and protein-coding genes (e.g., CEACAM6 (Cummings *et al.*, 2007)

and TOX3 (Seksenyan *et al.*, 2015)) emerged in the literature as being well characterised with relation to tamoxifen resistance and carcinogenesis pathways. Additional comparisons were also informative, most prominently tamoxifen resistance associated BCAR4 was found to be down-regulated in luminal A tumours while up-regulated in basal tumours, these results are similar to those reported by (Godinho et al., 2010, 2011).

Summary

In summary, we have identified multiple lists of differentially expressed lncRNAs and protein-coding genes believed to be related directly or indirectly to tamoxifen resistance and aggressiveness of breast cancer. Results from comparing relevant GEO and TCGA DEA results to TAMR upregulated genes generated common genes that can be utilised to construct novel interaction networks. To make up for the contradiction observed it is important when a gene is prioritised, its expression trend should not be taken at face value. Planning the next steps should be supported by other sources of information.

7. Discussion

Breast cancer is the most common cancer in the female population and is considered to be a major health issue with high morbidity and mortality rates worldwide (Cancer Research UK, 2017). The clear majority of breast tumours express ERα, indicating a positive response to endocrine therapies such as tamoxifen and aromatase inhibitors. However, endocrine resistance in ER-positive breast cancers is one of the major obstacles needing to be resolved. While many mechanisms of resistance have been proposed, many of the key triggers and regulators of these pathways are yet to be identified in order to establish highly sensitive and specific prognostic biomarkers and new therapeutic targets and hence improve survival rates. Most of the studies elaborate on the role of alterations in protein-coding genes on the molecular mechanism of endocrine resistance, such as ESR1, ErbB-family (EGFR, HER2, and HER3), and IGFR (Osipo *et al.*, 2007; Jeselsohn *et al.*, 2014; Murphy and Dickler, 2016).

Many lncRNAs having been found to be altered in endocrine-resistant breast cancers such as HOTAIR and BCAR4 (Godinho *et al.*, 2010; Bhan *et al.*, 2013) However, the role of lncRNAs in the pathways of resistance is still obscure, and many aspects of their integrated functional association with other molecules are missing. Generally, the main function of lncRNAs is directing genomic transcriptional, post-transcriptional, and translational processes (Wang and Chang, 2011). This is achieved by regulatory chromatin looping and histone modification to control targeted protein-coding gene expression, many lncRNAs also act as precursors of small non-coding RNAs (Cai and Cullen, 2007) and microRNA sponges to suppress or enhance their repressive activities (Huang *et al.*, 2017).

Taken together, this gives the impression that lncRNAs work on multiple levels and locations on the genome. The complexity of lncRNAs' mechanism of action and the dynamic nature of their expression, where active transcription and degradation occur according to continually changing cellular conditions, form a major challenge in the functional dissection of lnc RNA transcripts.

The overall aim of this project is to assess the role of lncRNAs in driving tamoxifen resistance in breast cancer. Our objective was to nominate a group of lncRNAs to further investigate *in-vitro*. Our entire project consisted of four main stages. Stage one was the foundation that was established by identifying a list of lncRNAs using RNA-seq technology. Stage 2 was to study a few of the nominated lncRNAs *in-vitro*, by manipulating their expression and assessing the effect on tamoxifen sensitivity in TAMR cells. Unfortunately, none of the investigated lncRNAs could be confirmed as regulating tamoxifen resistance. Therefore, for the next stage, we investigated the function one of the lncRNAs identified HOTAIRM1 in detail in pathways other than resistance. The final stage of our project was to utilize publicly available next generation sequencing datasets to help strengthen and refine our lncRNA selection criteria. This general discussion section summarizes the main findings of each results section, outlines the limitations of these studies, and discusses ideas for future research development.

The results from RNA-seq bioinformatics analysis identified differentially expressed lncRNAs and protein-coding genes between tamoxifen-sensitive and tamoxifen-resistant MCF-7 cell lines. From this we proposed that dysregulated lncRNAs act as oncogenes if their expression is increased or tumor suppressors if their expression is decreased. lncRNAs upregulated in TAMR cells were prioritised

mainly based on the degree of difference in expression and the statistical significance of this change. Of this list, LUCAT1, SOX21-AS1, NR2F1-AS1, and HOTAIRM1 lncRNAs were selected to undergo further *in-silico* analysis. GSEA functional characterisation of the list of differentially expressed lncRNAs and protein-coding genes were also performed. Initially general carcinogenesis pathways produced a list of significantly enriched pathways larger than expected. To gain more functional context for our list of differentially expressed protein-coding coding genes of which found to enrich carcinogenesis/endocrine resistance-associated gene sets were used to seed a lncRNA-miRNA-mRNA interaction network for each of the four lncRNAs.

Analysis of the 4 prioritised genes indicated a number of promising potentials for in-vitro validation. Silencing of candidate lncRNAs was successful except for LUCAT1. However, no increase in tamoxifen sensitivity was observed after depleting the expression of SOX21-AS1, NR2F1-AS1, or HOTAIRM1 which is contrary to what was expected. For the next stage of our project, HOTAIRM1 was selected to undergo further molecular investigations in an effort to explain our previous observations in the light of enhanced carcinogenic properties is that drive the aggressiveness of breast cancer cells with dysregulated HOTAIRM1 expression. The results of the effect of HOTAIRM1 depletion on EMT-related molecular expressions and HOXA5 expression affection in triple-negative breast cancer (CAL51) cells. We considered these areas to be needing further investigation as some changes were observed, such as increased migration and higher relative expression of β-catenin. Though it happens only with one of the siRNAs and might

be an off-target, non-specific effect, however, we can consider this area inconclusive and can be a candidate for further exploration.

Finally, datasets pertaining to tamoxifen responsiveness were sourced from the GEO and TCGA databases. We aimed to develop a methodological pipeline capable of producing multiple lists of differentially expressed genes and comparing different phenotypes and then highlighting the genes shared across comparable populations. Matchings between differentially expressed lists of genes between our list of RNA-seq data, GEO, and TCGA data, will further validate the prioritised list of candidate genes and give flexibility when evaluating any candidate gene stand relative to tamoxifen response in breast cancer.

Strengths and Limitations

The present study was designed to prioritise and then study the function of one of the identified lncRNAs generated using RNA-seq technology. While RNA-seq is a very reliable technology, bias can be introduced at any level along the pathway of the experiment. From technical bias during wet lab handling to the sequencing machine technical bias and to the bioinformatics analysis itself. Here, we noticed that RNA-seq produced a huge number of significantly expressed lncRNAs and protein-coding genes. When prioritising the genes, setting the fold change and p-value cut-off was a risky step of introducing false positives or excluding false negatives. Also, there was the well-established issue of lncRNA annotation and non-coding genes biotype determination, which has potential for losing true candidates, in this field we noticed the annotation resources to be very limiting, so, validating the annotation (automatically) was not possible.

There was also the challenge of choosing a tamoxifen resistant model for studying tamoxifen resistance in breast cancer. Due to the time limit, only a 2D monolayer culture model was used. While it is easier, and more readily used, it does not account for the tumour's microenvironment nor the patient physiological variables. It is undeniable that the affected supply chains due to the recent global pandemic affect some of our decisions regarding directing the project, while HOTAIRM1 was a considerable candidate for driving tamoxifen resistance, other candidates were also serious contenders for molecular analyses.

Transcriptomic analysis of GEO and TCGA datasets enabled the exploration of a broad diversity of genes, together with our RNA-seq data. It created a genotypic foundation that has the flexibility to be used and re-analysed for following omics studies.

Future work

For the currently nominated lncRNAs, further exploration and refinement of the constructed lncRNA-miRNA-mRNA interaction networks of each in areas such as prioritising an entire axis to be tested *in-vitro*. LUCAT1 expression negatively correlated with ESR1 expression in the analysed CCLE data and was significantly upregulated in TAMR cells. It would be therefore interesting to investigate the response of TAMR cells to LUCAT1 down-regulation. HOTAIRM1 characterization by expression manipulation was done using ATRA treatment for gene amplification. ATRA -mediated amplification is not specific for HOTAIRM1 only and carries an unignorable risk for off-target modifications. Therefore, HOTAIRM1 targeted amplification is to be achieved using vector mediated HOTAIRM1 overexpression.

There are a number of very promising candidates lncRNAs, and the process of nominating the true proper candidate is a vast area for improvement. For this matter, further exploration of more (raw datasets) from the publicly available databases; will give much uniform and reliable results, especially when paired with clinical data from the same samples that can give us the opportunity to stratify patients based on hormone receptor status, type of treatment given, the outcome of treatment (survival analysis), and associated genomic alterations. Furthermore, correlation analysis between HOTAIRM1 and protein-coding genes and miRNAs, known to be involved in the resistance pathways would be used to identify upstream regulators and downstream targets of any nominated lncRNA.

Conclusion

Our study aimed to examine the contribution of lncRNAs in the development of tamoxifen resistance in breast cancer. The in-silico study model used, was a comprehensive bioinformatics approach using R programming, utilizing RNA-seq technology; to pinpoint candidate lncRNAs with potential implications in driving tamoxifen resistance. we identified many dysregulated lncRNAs, of which we prioritized LUCAT1, SOX21-AS1, NR2F1-AS1, and HOTAIRM1 for further investigation. These four genes were significantly upregulated in tamoxifen resistant breast cancer cell line. Therefore, we proceeded with in-vitro analyses, that showed no association between downregulating any of the four lncRNAs and the overall response to tamoxifen. Consequently, we sought to refine our selection criteria by exploring publicly available GEO and TCGA datasets relative to our RNA-seq results. Despite research limitations due to the time factor and the effects of global pandemic, an intriguing list of lncRNAs has been identified, that might prove

useful for the development of novel therapeutic targets and prognostic biomarkers. By improving the experimental models and expanding the bioinformatics analyses we can enhance our understanding of the role of lncRNAs in tamoxifen resistance and ultimately improve the survival rates and quality of life of breast cancer patients.

References

Abe, O. *et al.* (2005) “Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials,” *The Lancet*, 365(9472), pp. 1687–1717. Available at: https://doi.org/10.1016/S0140-6736(05)66544-0.

Adey, A. and Shendure, J. (2020) “Bisulfite Sequencing,” *Definitions*, pp. 1139–1143. Available at: https://doi.org/10.32388/vje46h.

Ahadi, A. (2021) “Functional roles of lncRNAs in the pathogenesis and progression of cancer,” *Genes & Diseases*, 8(4), pp. 424–437. Available at: https://doi.org/https://doi.org/10.1016/j.gendis.2020.04.009.

Aigrain, L., Gu, Y. and Quail, M.A. (2016) “Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing,” *BMC Genomics*, 17(1), pp. 1–11. Available at: https://doi.org/10.1186/s12864-016-2757-4.

Ali, H.S. *et al.* (2020) “lncRNA- RP11-156p1.3, novel diagnostic and therapeutic targeting via CRISPR/Cas9 editing in hepatocellular carcinoma.,” *Genomics*, 112(5), pp. 3306–3314. Available at: https://doi.org/10.1016/j.ygeno.2020.06.020.

Alonso-Betanzos, V.B.-C. and A. (2019) *Microarray Bio-informatics analysis*, *Molecular Biology*.

Ariazi, E.A. *et al.* (2010) “The G protein-coupled receptor GPR30 inhibits proliferation of oestrogen receptor-positive breast cancer cells,” *Cancer Res*, 70(3), pp. 1184–1194. Available at: https://doi.org/10.1158/0008-5472.CAN-09-3068.

Ba, M.-C. *et al.* (2021) “Knockdown of lncRNA ZNRD1-AS1 suppresses gastric cancer cell proliferation and metastasis by targeting the miR-9-5p/HSP90AA1 axis.,” *Aging*, 13(13), pp. 17285–17301. Available at: https://doi.org/10.18632/aging.203209.

Barretina, J. *et al.* (2012) “The Cancer Cell Line Encyclopaedia enables predictive modelling of anticancer drug sensitivity,” *Nature*, 483(7391), pp. 603–607. Available at: https://doi.org/10.1038/nature11003.

Barrett, T. *et al.* (2013) “NCBI GEO: archive for functional genomics data sets—update,” *Nucleic Acids Research*, 41(D1), pp. D991–D995. Available at: https://doi.org/10.1093/nar/gks1193.

Barter, R.L. *et al.* (2014) “Network-based biomarkers enhance classical approaches to prognostic gene expression signatures.,” *BMC systems biology*, 8 Suppl 4(Suppl 4), p. S5. Available at: https://doi.org/10.1186/1752-0509-8-S4-S5.

Bartos, J.R. (2009) *Oestrogens: Production, Functions and Applications*. New York, UNITED STATES: Nova Science Publishers, Incorporated. Available at: http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=3018573.

Bassuk, S.S. and Manson, J.A.E. (2014) “Women’s Health Initiative Hormone Therapy Trials,” *Methods and Applications of Statistics in Clinical Trials*, 1(13), pp. 918–930. Available at: https://doi.org/10.1002/9781118596005.ch77.

Bentley, D.R. *et al.* (2008) “Accurate whole human genome sequencing using reversible terminator chemistry,” *Nature*, 456(7218), pp. 53–59. Available at: https://doi.org/10.1038/nature07517.

Bhan, A. *et al.* (2013) “Antisense transcript long noncoding RNA (lncRNA) HOTAIR is transcriptionally induced by estradiol,” *J Mol Biol*, 425. Available at: https://doi.org/10.1016/j.jmb.2013.01.022.

Bhatlekar, S., Fields, J.Z. and Boman, B.M. (2014) “HOX genes and their role in the development of human cancers.,” *Journal of molecular medicine (Berlin, Germany)*, 92(8), pp. 811–823. Available at: https://doi.org/10.1007/s00109-014-1181-y.

Birney, E. *et al.* (2007) “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project,” *Nature*, 447(7146), pp. 799–816. Available at: https://doi.org/10.1038/nature05874.

Bocchinfuso, W.P. *et al.* (1999) “A mouse mammary tumor virus-Wnt-1 transgene induces mammary gland hyperplasia and tumorigenesis in mice lacking oestrogen receptor-alpha,” *Cancer Res*. 1999/04/23, 59(8), pp. 1869–1876.

Brown, D. *et al.* (2022) *Are siRNA Pools Smart?* Available at: Screening with small interfering RNAs (siRNAs) is an important method for identifying genes involved in biological pathways (Accessed: July 1, 2022).

Brunckhorst, M.K. *et al.* (2014) “Angiopoietins promote ovarian cancer progression by establishing a procancer microenvironment,” *American Journal of Pathology*, 184(8), pp. 2285–2296. Available at: https://doi.org/10.1016/j.ajpath.2014.05.006.

Burian, R. and Barbieri, M. (2015) “Crick, F . H. C. (1958) On protein synthesis Related papers,” (1958).

Burstein, H.J. *et al.* (2014) “Adjuvant endocrine therapy for women with hormone receptor-positive breast cancer: American society of clinical oncology clinical practice guideline focused update,” *J Clin Oncol*, 32(21), pp. 2255–2269. Available at: https://doi.org/10.1200/JCO.2013.54.2258.

Cai, X. and Cullen, B.R. (2007) “The imprinted H19 noncoding RNA is a primary microRNA precursor,” *RNA*, 13(3), pp. 313–316. Available at: https://doi.org/10.1261/rna.351707.

Calin, G.A. *et al.* (2004) “Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers,” *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), pp. 2999–3004. Available at: https://doi.org/10.1073/pnas.0307323101.

Cancer Research UK (2017) *Breast cancer statistics*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Eight.

Chakravarty, D. *et al.* (2014) “The ooestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer,” *Nature Communications*, 5. Available at: https://doi.org/10.1038/ncomms6383.

Chen, C. *et al.* (2022) “Urinary Exosomal Long Noncoding RNA TERC as a Noninvasive Diagnostic and Prognostic Biomarker for Bladder Urothelial Carcinoma,” *Journal of Immunology Research*, 2022. Available at: https://doi.org/10.1155/2022/9038808.

Chen, G. *et al.* (2013) “LncRNADisease: a database for long-non-coding RNA-associated diseases,” *Nucleic Acids Res*, 41. Available at: https://doi.org/10.1093/nar/gks1099.

Chen, L. *et al.* (2013) “Role of Deregulated microRNAs in Breast Cancer Progression Using FFPE Tissue,” *PLoS ONE*, 8(1). Available at: https://doi.org/10.1371/journal.pone.0054213.

Chen, S. *et al.* (2012) “Realgar-induced apoptosis and differentiation in all-trans retinoic acid (ATRA)-sensitive NB4 and ATRA-resistant MR2 cells.,” *International journal of oncology*, 40(4), pp. 1089–1096. Available at: https://doi.org/10.3892/ijo.2011.1276.

Chiyomaru, T. *et al.* (2014) “Long non-coding RNA HOTAIR is targeted and regulated by miR-141 in human cancer cells,” *The Journal of biological chemistry*, 289(18), p. 12550. Available at: https://doi.org/10.1074/jbc.M113.488593.

Cianfrocca, M. and Gradishar, W. (2009) “New Molecular Classifications of Breast Cancer,” *CA: A Cancer Journal for Clinicians*, 59(5), pp. 303–313. Available at: https://doi.org/10.3322/caac.20029.

Cittelly, D.M. *et al.* (2010) “Downregulation of miR-342 is associated with tamoxifen resistant breast tumors.,” *Molecular cancer*, 9, p. 317. Available at: https://doi.org/10.1186/1476-4598-9-317.

Climent, J. *et al.* (2007) “Deletion of Chromosome 11q Predicts Response to Anthracycline-Based Chemotherapy in Early Breast Cancer,” *Cancer Research*, 67(2), pp. 818–826. Available at: https://doi.org/10.1158/0008-5472.CAN-06-3307.

Colditz, G.A. (1998) “Relationship Between Oestrogen Levels, Use of Hormone Replacement Therapy, and Breast Cancer,” *Journal of the National Cancer Institute*, 90(11), pp. 814–823. Available at: https://doi.org/10.1093/jnci/90.11.814.

Cole, M.P., Jones, C.T.A. and Todd, I.D.H. (1971) “A new anti-ooestrogenic agent in late breast cancer an early clinical appraisal of ICI46474,” *British Journal of Cancer*, 25(2), pp. 270–275. Available at: https://doi.org/10.1038/bjc.1971.33.

Conesa, A. *et al.* (2016) “A survey of best practices for RNA-seq data analysis,” *Genome Biol*, 17, p. 13. Available at: https://doi.org/10.1186/s13059-016-0881-8.

Cooper, R.G. (1969) “Combination chemotherapy in hormone resistant breast cancer,” in *Proceedings of the American Association for Cancer Research*. AMER ASSOC CANCER RESEARCH PO BOX 11806, BIRMINGHAM, AL 35202, p. 15.

Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017) “RNA-Seq differential expression analysis: An extended review and a software tool.,” *PloS one*, 12(12), p. e0190152. Available at: https://doi.org/10.1371/journal.pone.0190152.

Cowley, S.M. *et al.* (1997) “Oestrogen Receptors α and β Form Heterodimers on DNA,” *Journal of Biological Chemistry*, 272(32), pp. 19858–19862. Available at: https://doi.org/10.1074/jbc.272.32.19858.

Creighton, C.J. *et al.* (2008) “Development of Resistance to Targeted Therapies Transforms the Clinically Associated Molecular Profile Subtype of Breast Tumor Xenografts,” *Cancer Research*, 68(18), pp. 7493–7501. Available at: https://doi.org/10.1158/0008-5472.CAN-08-1404.

Crosignani, P.G. (2003) “Breast cancer and hormone-replacement therapy in the Million Women Study,” *Maturitas*, 46(2), pp. 91–92. Available at: https://doi.org/10.1016/j.maturitas.2003.09.002.

Cui, H. *et al.* (2018) “Follicle-stimulating hormone promotes the transformation of cholesterol to oestrogen in mouse adipose tissue,” *Biochemical and Biophysical Research Communications*, 495(3), pp. 2331–2337. Available at: https://doi.org/10.1016/j.bbrc.2017.12.120.

Cui, H. *et al.* (2020) “Inferences of Individual Drug Response-Related Long Non-coding RNAs Based on Integrating Multi-omics Data in Breast Cancer,” *Molecular Therapy - Nucleic Acids*, 20, pp. 128–139. Available at: https://doi.org/https://doi.org/10.1016/j.omtn.2020.01.038.

Cummings, M. *et al.* (2007) “O-5 Carcinoembryonic antigen cell adhesion molecule (CEACAM6) predicts breast cancer recurrence following adjuvant tamoxifen,”

European journal of cancer supplements, 5(3), p. 2. Available at: https://doi.org/10.1016/S1359-6349(07)71695-6.

Cunningham, F. *et al.* (2022) “Ensembl 2022,” *Nucleic Acids Research*, 50(D1), pp. D988–D995. Available at: https://doi.org/10.1093/nar/gkab1049.

Dalvai, M. and Bystricky, K. (2010) “Cell cycle and anti-oestrogen effects synergize to regulate cell proliferation and ER target gene expression,” *PLoS One*, 5(6), p. e11011. Available at: https://doi.org/10.1371/journal.pone.0011011.

Davis, N.M. *et al.* (2014) “Deregulation of the EGFR/PI3K/PTEN/Akt/mTORC1 pathway in breast cancer: possibilities for therapeutic intervention,” *Oncotarget*, 5(13), pp. 4603–4650. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4148087/.

Debrabant, B. (2017) “The null hypothesis of GSEA, and a novel statistical model for competitive gene set analysis,” *Bioinformatics*, 33(9), p. 1271.

deGraffenried, L.A. (2004) “NF- B inhibition markedly enhances sensitivity of resistant breast cancer tumor cells to tamoxifen,” *Annals of Oncology*, 15(6), pp. 885–890. Available at: https://doi.org/10.1093/annonc/mdh232.

Deonarine, K. *et al.* (2007) “Gene expression profiling of cutaneous wound healing,” *Journal of Translational Medicine*, 5, pp. 1–11. Available at: https://doi.org/10.1186/1479-5876-5-11.

Diaz-Lagares, A. *et al.* (2016) “Epigenetic inactivation of the p53-induced long noncoding RNA TP53 target 1 in human cancer.,” *Proceedings of the National Academy of Sciences of the United States of America*, 113(47), pp. E7535–E7544. Available at: https://doi.org/10.1073/pnas.1608585113.

Dijkstra, J.M. and Alexander, D.B. (2015) “The ‘NF-kB interacting long noncoding RNA’ (NKILA) transcript is antisense to cancer-associated gene PMEPA1,” *F1000Res*, 4, p. 96. Available at: https://doi.org/10.12688/f1000research.6400.1.

Dimitrios, T., Brian, M. and Philip, C. (1972) “Menopause and breast cancer risk,” *Journal of the National Cancer Institute*, 48(3), pp. 605–613. Available at: https://doi.org/10.1093/jnci/48.3.605.

Dinger, M.E. *et al.* (2009) “Pervasive transcription of the eukaryotic genome: Functional indices and conceptual implications,” *Briefings in Functional Genomics and Proteomics*, 8(6), pp. 407–423. Available at: https://doi.org/10.1093/bfgp/elp038.

Dixon, J.M. (2014) “Endocrine Resistance in Breast Cancer,” *New Journal of Science*, 2014, pp. 1–27. Available at: https://doi.org/10.1155/2014/390618.

Djebali, S. *et al.* (2012) “Landscape of transcription in human cells.,” *Nature*, 489(7414), pp. 101–108. Available at: https://doi.org/10.1038/nature11233.

Edgar, R., Domrachev, M. and Lash, A.E. (2002) “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Res*, 30(1), pp. 207–210. Available at: https://doi.org/10.1093/nar/30.1.207.

Eldesouki, S. *et al.* (2022) “XIST in Brain Cancer,” *Clinica Chimica Acta*, 531(April), pp. 283–290. Available at: https://doi.org/10.1016/j.cca.2022.04.993.

El-Tanani, M.K.K. and Green, C.D. (1997) “Two Separate Mechanisms for Ligand-Independent Activation of the Oestrogen Receptor,” *Molecular Endocrinology*, 11(7), pp. 928–937. Available at: https://doi.org/10.1210/mend.11.7.9939.

Encarnacion, C.A. *et al.* (1993) “Measurement of steroid hormone receptors in breast cancer patients on tamoxifen,” *Breast Cancer Res Treat*. 1993/01/01, 26(3), pp. 237–246.

Enmark, E. *et al.* (1996) “Cloning of a novel oestrogen receptor expressed in rat prostate and ovary,” *Proceedings of the National Academy of Sciences of the United States of America*, 93(12), p. 5925. Available at: https://doi.org/10.1073/pnas.93.12.5925.

Ensemble (2022) *Gene: ZNRD1ASP ENSG00000196299*. Available at: https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000196299;r=CHR_HSCHR6_MHC_SSTO_CTG1:29990282-30052027.

Fagan, D.H. *et al.* (2017) “Acquired Tamoxifen Resistance in MCF-7 Breast Cancer Cells Requires Hyperactivation of eIF4F-Mediated Translation.,” *Hormones & cancer*, 8(4), pp. 219–229. Available at: https://doi.org/10.1007/s12672-017-0296-3.

Fang, J. *et al.* (2022) “LncRNA TTN-AS1 confers tamoxifen resistance in breast cancer via sponging miR-107 to modulate PI3K/AKT signaling pathway,” *Am J Transl Res*, 14(4), pp. 2267–2279.

Fanning, S.W. *et al.* (2016) “Oestrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation,” *eLife*, 5(FEBRUARY2016), pp. 1–25. Available at: https://doi.org/10.7554/eLife.12792.

Feng, J. *et al.* (2020) “Cross-talk between the ER pathway and the lncRNA MAFG-AS1/miR-339-5p/ CDK2 axis promotes progression of ER+ breast cancer and confers tamoxifen resistance,” *Aging (Albany NY)*, 12(20), pp. 20658–20683. Available at: https://doi.org/10.18632/aging.103966.

Ferrucci, L.M. *et al.* (2009) “Intake of meat, meat mutagens, and iron and the risk of breast cancer in the prostate, lung, colorectal, and ovarian cancer screening trial,” *British Journal of Cancer*, 101(1), pp. 178–184. Available at: https://doi.org/10.1038/sj.bjc.6605118.

Filardo, E.J. *et al.* (2002) “Oestrogen action via the G protein-coupled receptor, GPR30: stimulation of adenylyl cyclase and cAMP-mediated attenuation of the epidermal growth factor receptor-to-MAPK signaling axis,” *Molecular endocrinology (Baltimore, Md.)*, 16(1), p. 70.

Fisher, B. *et al.* (1998) “Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study,” *JNCI: Journal of the National Cancer Institute*, 90(18), pp. 1371–1388. Available at: https://doi.org/10.1093/jnci/90.18.1371.

Flannigan, R. *et al.* (2017) “Mp07-02 Polya Tag Library Preparation for New Generation Sequencing (Ngs) in Human Testis Fails to Detect Non-Coding and Translated RNAs Important in Testicular Function As Compared To Ribosomal RNA Depletion Method.,” *Journal of Urology*, 197(4S), p. e82. Available at: https://doi.org/10.1016/j.juro.2017.02.268.

Frasor, J. *et al.* (2003) “Profiling of oestrogen up- and down-regulated gene expression in human breast cancer cells: Insights into gene networks and pathways underlying oestrogenic control of proliferation and cell phenotype,” *Endocrinology*, 144(10), pp. 4562–4574. Available at: https://doi.org/10.1210/en.2003-0567.

Frasor, J. *et al.* (2006) “Gene expression preferentially regulated by tamoxifen in breast cancer cells and correlations with clinical outcome,” *Cancer Research*, 66(14), pp. 7334–7340. Available at: https://doi.org/10.1158/0008-5472.CAN-05-4269.

Fuentes, N. and Silveyra, P. (2019) “Oestrogen receptor signaling mechanisms.,” *Advances in protein chemistry and structural biology*, 116, pp. 135–170. Available at: https://doi.org/10.1016/bs.apcsb.2019.01.001.

Gallego-Paez, L.M. *et al.* (2017) “Alternative splicing: the pledge, the turn, and the prestige: The key role of alternative splicing in human biological systems,” *Human Genetics*, 136(9), pp. 1015–1042. Available at: https://doi.org/10.1007/s00439-017-1790-y.

Gao, W., Qiao, M. and Luo, K. (2021) “Long Noncoding RNA TP53TG1 Contributes to Radioresistance of Glioma Cells Via miR-524-5p/RAB5A Axis.,” *Cancer biotherapy & radiopharmaceuticals*, 36(7), pp. 600–612. Available at: https://doi.org/10.1089/cbr.2020.3567.

Gardner, P.P. *et al.* (2015) “Conservation and Losses of Non-Coding RNAs in Avian Genomes,” *PLoS One*, 10(3), p. e0121797. Available at: https://doi.org/10.1371/journal.pone.0121797.

Gaskins, A.J. *et al.* (2012) “Endogenous reproductive hormones and C-reactive protein across the menstrual cycle: The BioCycle Study,” *American Journal of Epidemiology*, 175(5), pp. 423–431. Available at: https://doi.org/10.1093/aje/kwr343.

Gates, A.J. *et al.* (2021) “A wealth of discovery built on the Human Genome Project — by the numbers,” *Nature*, 590(7845), pp. 212–215. Available at: https://doi.org/10.1038/d41586-021-00314-6.

GEO (2021) *Querying GEO DataSets and GEO Profiles*. Available at: https://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html.

Godinho, M. *et al.* (2011) “Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells,” *J Cell Physiol*, 226. Available at: https://doi.org/10.1002/jcp.22503.

Godinho, M.F.E. *et al.* (2010) “Relevance of BCAR4 in tamoxifen resistance and tumour aggressiveness of human breast cancer,” *British Journal of Cancer*, 103(8), p. 1284. Available at: https://doi.org/10.1038/sj.bjc.6605884.

Goh, C.W. *et al.* (2019) “Invasive ductal carcinoma with coexisting ductal carcinoma in situ (IDC/DCIS) versus pure invasive ductal carcinoma (IDC): a comparison of clinicopathological characteristics, molecular subtypes, and clinical outcomes,” *Journal of Cancer Research and Clinical Oncology*, 145(7), pp. 1877–1886. Available at: https://doi.org/10.1007/s00432-019-02930-2.

Gomes, A. and Korf, B. (2018) “Chapter 5 - Genetic Testing Techniques,” in N.H. Robin and M.B.B.T.-P.C.G. Farmer (eds). Elsevier, pp. 47–64. Available at: https://doi.org/https://doi.org/10.1016/B978-0-323-48555-5.00005-3.

Gravena, A.A.F. *et al.* (2018) “The obesity and the risk of breast cancer among pre and postmenopausal women,” *Asian Pacific Journal of Cancer Prevention*, 19(9), pp. 2429–2436. Available at: https://doi.org/10.22034/APJCP.2018.19.9.2429.

Greene, G.L. *et al.* (1986) “Sequence and expression of human oestrogen receptor complementary DNA,” *Science*. 1986/03/07, 231(4742), pp. 1150–1154.

Grimm, D. (2009) “Small silencing RNAs: State-of-the-art,” *Advanced Drug Delivery Reviews*, 61(9), pp. 672–703. Available at: https://doi.org/10.1016/j.addr.2009.05.002.

Gu, Y. *et al.* (2016) “Lower beclin 1 downregulates HER2 expression to enhance tamoxifen sensitivity and predicts a favorable outcome for ER positive breast cancer,” *Oncotarget*, 8(32), pp. 52156–52177. Available at: https://doi.org/10.18632/oncotarget.11044.

Guo, C. *et al.* (2008) “The noncoding RNA, miR-126, suppresses the growth of neoplastic cells by targeting phosphatidylinositol 3-kinase signaling and is frequently lost in colon cancers,” *Genes, Chromosomes and Cancer*, 47(11), pp. 939–946. Available at: https://doi.org/https://doi.org/10.1002/gcc.20596.

Gutierrez, M.C. *et al.* (2005) “Molecular Changes in Tamoxifen-Resistant Breast Cancer: Relationship Between Oestrogen Receptor, HER-2, and p38 Mitogen-Activated Protein Kinase,” *Journal of Clinical Oncology*, 23(11), pp. 2469–2476. Available at: https://doi.org/10.1200/JCO.2005.01.172.

Haese, A. *et al.* (2008) “Clinical Utility of the PCA3 Urine Assay in European Men Scheduled for Repeat Biopsy,” *European Urology*, 54(5), pp. 1081–1088. Available at: https://doi.org/10.1016/J.EURURO.2008.06.071.

Hamajima, N. *et al.* (2012) “Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies,” *The Lancet Oncology*, 13(11), pp. 1141–1151. Available at: https://doi.org/10.1016/S1470-2045(12)70425-4.

HARPER, M.J.K. and WALPOLE, A.L. (1967) “MODE OF ACTION OF I.C.I. 46,474 IN PREVENTING IMPLANTATION IN RATS,” *Journal of Endocrinology*, 37(1), pp. 83–92. Available at: https://doi.org/10.1677/joe.0.0370083.

Henras, A.K. *et al.* (2015) “An overview of pre-ribosomal RNA processing in eukaryotes,” *Wiley Interdisciplinary Reviews: RNA*, 6(2), pp. 225–242. Available at: https://doi.org/10.1002/wrna.1269.

Hetemäki, N. *et al.* (2017) “Oestrogen metabolism in abdominal subcutaneous and visceral adipose tissue in postmenopausal women,” *Journal of Clinical Endocrinology and*

Metabolism, 102(12), pp. 4588–4595. Available at: https://doi.org/10.1210/jc.2017-01474.

Hinahon, C. *et al.* (2013) "Improvements to RiboMinus™ Eukaryote rRNA Depletion Probe Design and Functionality to Enable a Faster and More Complete Workflow," *Journal of Biomolecular Techniques - JBT*, 24(May), p. 2013.

Hologic, Inc. (2012) "Progensa® PCA3 Assay," p. December. Available at: http://www.hologic.com/sites/default/files/package inserts/502083-IFU-PI_001.pdf.

Holt, R.A. and Jones, S.J.M. (2008) "The new paradigm of flow cell sequencing," *Genome Research*, 18(6), pp. 839–846. Available at: https://doi.org/10.1101/gr.073262.107.

Hoppe, R. *et al.* (2013) "Increased expression of miR-126 and miR-10a predict prolonged relapse-free time of primary ooestrogen receptor-positive breast cancer following tamoxifen treatment," *Eur J Cancer*, 49. Available at: https://doi.org/10.1016/j.ejca.2013.07.145.

Hortobagyi, G.N. (2000) "Developments in chemotherapy of breast cancer," *Cancer*, 88(12 SUPPL.), pp. 3073–3079. Available at: https://doi.org/10.1002/1097-0142(20000615)88:12+<3073::aid-cncr26>3.0.co;2-r.

Hsien, C. and Journal, R.A. (2013) "Characterization of a Novel Long Noncoding RNA, SCAL1 , Induced by Cigarette Smoke and Elevated in Lung Cancer Cell Lines," 49(Aug), pp. 1–11.

Huan, L. *et al.* (2020) "Hypoxia induced LUCAT1/PTBP1 axis modulates cancer cell viability and chemotherapy response," *Molecular Cancer*, 19(1), pp. 1–17. Available at: https://doi.org/10.1186/s12943-019-1122-z.

Huang, D. *et al.* (2017) "Mechanisms of resistance to selective oestrogen receptor down-regulator in metastatic breast cancer," *Biochimica et Biophysica Acta - Reviews on*

Cancer, 1868(1), pp. 148–156. Available at: https://doi.org/10.1016/j.bbcan.2017.03.008.

Huang, H. *et al.* (2018) “LncRNA NR2F1-AS1 regulates hepatocellular carcinoma oxaliplatin resistance by targeting ABCC1 via miR-363,” *Journal of Cellular and Molecular Medicine*, 22(6), pp. 3238–3245. Available at: https://doi.org/10.1111/jcmm.13605.

Huang, N.-S. *et al.* (2016) “Long non-coding RNA metastasis associated in lung adenocarcinoma transcript 1 (MALAT1) interacts with oestrogen receptor and predicted poor survival in breast cancer,” *Oncotarget*, 7(25), pp. 37957–37965. Available at: https://doi.org/10.18632/oncotarget.9364.

Ignatov, A. *et al.* (2011) “G-protein-coupled oestrogen receptor GPR30 and tamoxifen resistance in breast cancer,” *Breast Cancer Research and Treatment*, 128(2), pp. 457–466. Available at: https://doi.org/10.1007/s10549-011-1584-1.

Illumina (2016) “NovaSeq 6000 Sequencing System,” *770-2016-025-H*, 4(February), pp. 1–4. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/novaseq-6000-system-specification-sheet-770-2016-025.pdf.

Incorvati, J.A. *et al.* (2013) “Targeted therapy for HER2 positive breast cancer,” *Journal of Hematology and Oncology*, 6(1), pp. 1–9. Available at: https://doi.org/10.1186/1756-8722-6-38.

Jansen, M.P. *et al.* (2012) “High miR-26a and low CDC2 levels associate with decreased EZH2 expression and with favorable outcome on tamoxifen in metastatic breast cancer,” *Breast Cancer Res Treat*, 133. Available at: https://doi.org/10.1007/s10549-011-1877-4.

Jeggo, P.A., Pearl, L.H. and Carr, A.M. (2016) “DNA repair, genome stability and cancer: a historical perspective,” *Nat Rev Cancer*, 16(1), pp. 35–42. Available at: https://doi.org/10.1038/nrc.2015.4.

Jeselsohn, R. *et al.* (2014) “Emergence of constitutively active oestrogen receptor-alpha mutations in pretreated advanced oestrogen receptor-positive breast cancer,” *Clin Cancer Res*, 20(7), pp. 1757–1767. Available at: https://doi.org/10.1158/1078-0432.CCR-13-2332.

Ji, W.C. *et al.* (2021) “Role of LncRNA NR2F1-AS1 and LncRNA H19 genes in hepatocellular carcinoma and their effects on biological function of Huh-7,” *Cancer Management and Research*, 13, pp. 941–951. Available at: https://doi.org/10.2147/CMAR.S284650.

Jiao, G. *et al.* (2015) “Limitations of MTT and CCK-8 assay for evaluation of graphene cytotoxicity,” *RSC Advances*, 5(66), pp. 53240–53244. Available at: https://doi.org/10.1039/c5ra08958a.

Jordan, V.C. (2008) “Tamoxifen: Catalyst for the change to targeted therapy,” *European Journal of Cancer*, 44(1), pp. 30–38. Available at: https://doi.org/10.1016/j.ejca.2007.11.002.

Kang, L. *et al.* (2010) “Involvement of oestrogen receptor variant ER-alpha36, not GPR30, in nongenomic oestrogen signaling,” *Mol Endocrinol*, 24(4), pp. 709–721. Available at: https://doi.org/10.1210/me.2009-0317.

Kennedy, B.J. (1965) “Hormone therapy for advanced breast cancer,” *Cancer*, 18(12), pp. 1551–1557. Available at: https://doi.org/10.1002/1097-0142(196512)18:12<1551::AID-CNCR2820181206>3.0.CO;2-1.

Khalil, A.M. *et al.* (2009) “Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression,” *Proc Natl Acad Sci U S A*, 106(28), pp. 11667–11672. Available at: https://doi.org/10.1073/pnas.0904715106.

Khan, W. (2019) “Oestrogen.” Rijeka: IntechOpen. Available at: https://doi.org/10.5772/intechopen.73419.

Kiang, D.T. (1977) “Tamoxifen (Antioestrogen) Therapy in Advanced Breast Cancer,” pp. 687–690.

Kim, C. *et al.* (2011) “Oestrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of oestrogen receptor-positive breast cancer.,” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29(31), pp. 4160–4167. Available at: https://doi.org/10.1200/JCO.2010.32.9615.

Kim, C.Y. *et al.* (2020) “The LncRNA HOTAIRM1 Promotes Tamoxifen Resistance by Mediating HOXA1 Expression in ER+ Breast Cancer Cells,” *J Cancer*, 11(12), pp. 3416–3423. Available at: https://doi.org/10.7150/jca.38728.

Kim, C.Y. *et al.* (2021) “HOXA5 confers tamoxifen resistance via the PI3K/AKT signaling pathway in ER-positive breast cancer.,” *Journal of Cancer*, 12(15), pp. 4626–4637. Available at: https://doi.org/10.7150/jca.59740.

Kim, H.W. *et al.* (2020) “ZNRD1 and Its Antisense Long Noncoding RNA ZNRD1-AS1 Are Oppositely Regulated by Cold Atmospheric Plasma in Breast Cancer Cells.,” *Oxidative medicine and cellular longevity*, 2020, p. 9490567. Available at: https://doi.org/10.1155/2020/9490567.

Kim, J.C. *et al.* (2016) “Complex behavior of ALDH1A1 and IGFBP1 in liver metastasis from a colorectal cancer,” *PLoS ONE*, 11(5), pp. 1–15. Available at: https://doi.org/10.1371/journal.pone.0155160.

Klinge, C.M. *et al.* (2010) “Oestrogen receptor alpha 46 is reduced in tamoxifen resistant breast cancer cells and re-expression inhibits cell proliferation and oestrogen receptor alpha 66-regulated target gene transcription,” *Molecular and Cellular Endocrinology*, 323(2), pp. 268–276. Available at: https://doi.org/10.1016/j.mce.2010.03.013.

Knowlden, J.M. *et al.* (2003) “Elevated levels of epidermal growth factor receptor/c-erbB2 heterodimers mediate an autocrine growth regulatory pathway in tamoxifen-resistant MCF-7 cells,” *Endocrinology*, 144(3), p. 1032.

Kotsantis, P. *et al.* (2016) “Increased global transcription activity as a mechanism of replication stress in cancer,” *Nature Communications*, 7, pp. 1–13. Available at: https://doi.org/10.1038/ncomms13087.

Kulkoyluoglu, E. and Madak-Erdogan, Z. (2016) “Nuclear and extranuclear-initiated oestrogen receptor signaling crosstalk and endocrine resistance in breast cancer,” *Steroids*, 114, pp. 41–47. Available at: https://doi.org/10.1016/j.steroids.2016.06.007.

Landau, W.M. and Liu, P. (2013) “Dispersion estimation and its effect on test performance in RNA-seq data analysis: A simulation-based comparison of methods,” *PLoS ONE*, 8(12). Available at: https://doi.org/10.1371/journal.pone.0081415.

Lee, H.B. *et al.* (2016) “Abstract P6-04-02: Identification of ESR1 splice variants associated with prognosis in oestrogen receptor positive breast cancer,” *Cancer*

Res, 76(4 Supplement), pp. P6-04–02. Available at: http://cancerres.aacrjournals.org/content/76/4_Supplement/P6-04-02.abstract.

Lei, J.T. *et al.* (2019) “Endocrine therapy resistance: new insights,” *Breast*, 48 Suppl 1(Suppl 1), pp. S26–S30. Available at: https://doi.org/10.1016/S0960-9776(19)31118-X.

Levant, B. (2005) “Applying Genomic and Proteomic Microarray Technology in Drug Discovery,” *Shock*, p. 194. Available at: https://doi.org/10.1097/00024382-200508000-00016.

Levin, E. (2008) *Rapid signaling by steroid receptors*. Available at: https://doi.org/10.1152/ajpregu.90605.2008.

Li, G. *et al.* (2013) “Oestrogen receptor-alpha36 is involved in development of acquired tamoxifen resistance via regulating the growth status switch in breast cancer cells,” *Mol Oncol*, 7(3), pp. 611–624. Available at: https://doi.org/10.1016/j.molonc.2013.02.001.

Li, L., Meng, D. and Wang, R. (2021) “Long non-coding RNA SOX21-AS1 enhances the stemness of breast cancer cells via the Hippo pathway.,” *FEBS open bio*, 11(1), pp. 251–264. Available at: https://doi.org/10.1002/2211-5463.13015.

Li, X. *et al.* (2018) “Validation of the newly proposed American Joint Committee on Cancer (AJCC) breast cancer prognostic staging group and proposing a new staging system using the National Cancer Database,” *Breast Cancer Research and Treatment*, 171(2), pp. 303–313. Available at: https://doi.org/10.1007/s10549-018-4832-9.

Li, Y. *et al.* (2016) “HBXIP and LSD1 Scaffolded by lncRNA Hotair Mediate Transcriptional Activation by c-Myc,” *Cancer Res*, 76(2), p. 293. Available at: https://doi.org/10.1158/0008-5472.CAN-14-3607.

Li, Y. *et al.* (2017) “A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data,” *BMC Genomics*, 18(1), p. 508. Available at: https://doi.org/10.1186/s12864-017-3906-0.

Li, Y. *et al.* (2021) “Identification and analysis of lncRNA, microRNA and mRNA expression profiles and construction of ceRNA network in Talaromyces marneffei -infected THP-1 macrophage,” *PeerJ*, 9, pp. e10529–e10529. Available at: https://doi.org/10.7717/peerj.10529.

Li, Y.L. *et al.* (2020) “Up-regulated lnc-lung cancer associated transcript 1 enhances cell migration and invasion in breast cancer progression,” *Biochemical and Biophysical Research Communications*, 521(2), pp. 271–278. Available at: https://doi.org/10.1016/j.bbrc.2019.08.040.

Li, Z. *et al.* (2021) “Research on magnetic bead motion characteristics based on magnetic beads preset technology,” *Scientific Reports*, 11(1), pp. 1–13. Available at: https://doi.org/10.1038/s41598-021-99331-8.

Lim, Y.C. *et al.* (2006) “Endoxifen, a secondary metabolite of tamoxifen, and 4-OH-tamoxifen induce similar changes in global gene expression patterns in MCF-7 breast cancer cells,” *J Pharmacol Exp Ther*, 318(2), pp. 503–512. Available at: https://doi.org/10.1124/jpet.105.100511.

Liu, B. *et al.* (2015) “A cytoplasmic NF-kappaB interacting long noncoding RNA blocks IkappaB phosphorylation and suppresses breast cancer metastasis,” *Cancer Cell*, 27(3), pp. 370–381. Available at: https://doi.org/10.1016/j.ccell.2015.02.004.

Liu, C. *et al.* (2019) “Long noncoding RNA LUCAT1 promotes migration and invasion of prostate cancer cells by inhibiting KISS1 expression,” *European Review for Medical and Pharmacological Sciences*, 23(8), pp. 3277–3283. Available at: https://doi.org/10.26355/eurrev_201904_17689.

Liu, G., Mattick, J.S. and Taft, R.J. (2013) “A meta-analysis of the genomic and transcriptomic composition of complex life,” *Cell Cycle*, 12(13), pp. 2061–2072. Available at: https://doi.org/10.4161/cc.25134.

Liu, X. *et al.* (2020a) “Long noncoding RNA SOX21-AS1 regulates the progression of triple-negative breast cancer through regulation of miR-520a-5p/ORMDL3 axis,” *J Cell Biochem*, 121(11), pp. 4601–4611. Available at: https://doi.org/10.1002/jcb.29674.

Liu, X. *et al.* (2020b) “Long noncoding RNA SOX21вЂђAS1 regulates the progression of triplevвЂђnegative breast cancer through regulation of miRввЂђ520avвЂђ5p/ORMDL3 axis,” *J Cell Biochem*, 121(11), pp. 4601–4611. Available at: https://doi.org/10.1002/jcb.29674.

Liu, Y. *et al.* (2017) “LMO2 promotes tumor cell invasion and metastasis in basal-type breast cancer by altering actin cytoskeleton remodeling.,” *Oncotarget*, 8(6), pp. 9513–9524. Available at: https://doi.org/10.18632/oncotarget.13434.

Liu, Y. *et al.* (2019) “lncRNA CYTOR promotes tamoxifen resistance in breast cancer cells via sponging miRвЂ‘125avЂ‘5p,” *International journal of molecular medicine*, 45(2), pp. 497–509. Available at: https://doi.org/10.3892/ijmm.2019.4428.

Liu, Y. *et al.* (2021) “Long non-coding RNA NR2F1-AS1 induces breast cancer lung metastatic dormancy by regulating NR2F1 and O”Np63,” *Nature communications*, 12(1), p. 5232. Available at: https://doi.org/10.1038/s41467-021-25552-0.

Liu, Y. *et al.* (2022) “Correction to: Comparative performance of the GenoLab M and NovaSeq 6000 sequencing platforms for transcriptome and LncRNA analysis (BMC Genomics, (2021), 22, 1, (829), 10.1186/s12864-021-08150-8),” *BMC Genomics*, 23(1), pp. 1–12. Available at: https://doi.org/10.1186/s12864-022-08307-z.

Love, M.I., Huber, W. and Anders, S. (2014) “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, 15(12), pp. 1–21. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Lu, Q. *et al.* (2021) “LncRNA TP53TG1 Promotes the Growth and Migration of Hepatocellular Carcinoma Cells via Activation of ERK Signaling,” *Non-Coding RNA* . Available at: https://doi.org/10.3390/ncrna7030052.

Lu, X. *et al.* (2017) “A Novel Long Non-Coding RNA, SOX21-AS1, Indicates a Poor Prognosis and Promotes Lung Adenocarcinoma Proliferation,” *Cellular Physiology and Biochemistry*, 42(5), pp. 1857–1869. Available at: https://doi.org/10.1159/000479543.

Luzón-Toro, B. *et al.* (2019) “LncRNA LUCAT1 as a novel prognostic biomarker for patients with papillary thyroid cancer,” *Scientific Reports*, 9(1), pp. 1–12. Available at: https://doi.org/10.1038/s41598-019-50913-7.

Ma, C. *et al.* (2014) “H19 promotes pancreatic cancer metastasis by derepressing let-7’s suppression on its target HMGA2-mediated EMT,” *Tumor Biology*, 35(9), pp. 9163–9169. Available at: https://doi.org/10.1007/s13277-014-2185-5.

Ma, C.X. *et al.* (2015) “Mechanisms of aromatase inhibitor resistance,” *Nature Reviews Cancer*, 15(5), pp. 261–275. Available at: https://doi.org/10.1038/nrc3920.

Ma, Y. *et al.* (2019) “LncRNA DSCAMвЂђAS1 acts as a sponge of miRвЂђ137 to enhance Tamoxifen resistance in breast cancer,” *J Cell Physiol*, 234(3), pp. 2880–2894. Available at: https://doi.org/10.1002/jcp.27105.

Macedo, L.F., Sabnis, G. and Brodie, A. (2008) “Preclinical modeling of endocrine response and resistance: focus on aromatase inhibitors,” *Cancer*, 112(3 Suppl), p. 679.

Machlus, K.R. *et al.* (2010) “A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data,” *Nat Biotechno*, 102(5), pp. 936–944. Available at: https://doi.org/10.1038/nbt.3001.A.

Maggi, A. (2011) “Liganded and unliganded activation of oestrogen receptor and hormone replacement therapies,” *Biochimica et biophysica acta*, 1812(8), pp. 1054–1060. Available at: https://doi.org/10.1016/j.bbadis.2011.05.001.

Management, C. (2021) “NR2F1-AS1 / miR-140 / HK2 Axis Regulates Hypoxia-Induced Glycolysis and Migration in Hepatocellular Carcinoma,” pp. 427–437.

Mao, X., Su, Z. and Mookhtiar, A.K. (2017) “Long non-coding RNA: a versatile regulator of the nuclear factor-κB signalling circuit,” *Immunology*, 150(4), pp. 379–388. Available at: https://doi.org/10.1111/imm.12698.

de Marchi, T. *et al.* (2016) “Endocrine therapy resistance in oestrogen receptor (ER)-positive breast cancer,” *Drug Discov Today*, 21(7), pp. 1181–1188. Available at: https://doi.org/10.1016/j.drudis.2016.05.012.

Mattick, J.S. (2003) “Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms,” *BioEssays*, 25(10), pp. 930–939. Available at: https://doi.org/https://doi.org/10.1002/bies.10332.

Mayer, I.A. *et al.* (2014) “New strategies for triple-negative breast cancer-deciphering the heterogeneity,” *Clinical Cancer Research*, 20(4), pp. 782–790. Available at: https://doi.org/10.1158/1078-0432.CCR-13-0583.

Men, X. *et al.* (2021) “Overexpression of TMEM47 Induces Tamoxifen Resistance in Human Breast Cancer Cells,” *Technology in Cancer Research and Treatment*, 20, pp. 1–10. Available at: https://doi.org/10.1177/15330338211004916.

Merenbakh-Lamin, K. *et al.* (2013) “D538G mutation in oestrogen receptor-alpha: A novel mechanism for acquired endocrine resistance in breast cancer,” *Cancer Res*, 73(23), pp. 6856–6864. Available at: https://doi.org/10.1158/0008-5472.CAN-13-1197.

Milos, P. (2008) “Helicos BioSciences,” *Pharmacogenomics*, 9(4), pp. 477–480. Available at: https://doi.org/10.2217/14622416.9.4.477.

Mo, Z. *et al.* (2013) “GPR30 as an initiator of tamoxifen resistance in hormone-dependent breast cancer,” *Breast Cancer Research*, 15(6). Available at: https://doi.org/10.1186/bcr3581.

Mou, E. and Wang, H. (2019) “LncRNA LUCAT1 facilitates tumorigenesis and metastasis of triple-negative breast cancer through modulating MIR-5702,” *Bioscience Reports*, 39(9), pp. 1–12. Available at: https://doi.org/10.1042/BSR20190489.

Muluhngwi, P. and Klinge, C.M. (2021) “Identification and roles of mir-29b-1-3p and mir29a-3p-regulated and non-regulated lncrnas in endocrine-sensitive and resistant breast cancer cells,” *Cancers*, 13(14), p. 3530. Available at: https://doi.org/10.3390/cancers13143530.

Murphy, C.G. and Dickler, M. (2016) “Endocrine resistance in hormone-responsive breast cancer: mechanisms and therapeutic strategies,” *Endocr.-Relat. Cancer*, pp. R337–R352. Available at: https://doi.org/10.1530/ERC-16-0121.

Myers, R.M. *et al.* (2011) “A user’s guide to the Encyclopedia of DNA elements (ENCODE),” *PLoS Biology*, 9(4). Available at: https://doi.org/10.1371/journal.pbio.1001046.

Nai, Y. *et al.* (2020) “LncRNA LUCAT1 contributes to cell proliferation and migration in human pancreatic ductal adenocarcinoma via sponging miR-539,” *Cancer Medicine*, 9(2), pp. 757–767. Available at: https://doi.org/10.1002/cam4.2724.

Nandy, A., Gangopadhyay, S. and Mukhopadhyay, A. (2014) “Individualizing breast cancer treatment-The dawn of personalized medicine,” *Experimental Cell Research*, 320(1), pp. 1–11. Available at: https://doi.org/10.1016/j.yexcr.2013.09.002.

National Human Genome Research Institute (NHGRI) (2020) *Human Genome Project FAQ*, *genome.gov*. Available at: https://www.genome.gov/human-genome-project/Completion-FAQ.

Nielsen, S.C.A. *et al.* (2014) “Near-complete genome sequencing of swine vesicular disease virus using the Roche GS FLX sequencing platform,” *PLoS ONE*, 9(5), pp. 1–7. Available at: https://doi.org/10.1371/journal.pone.0097180.

Nilsson, S. *et al.* (2001) “Mechanisms of oestrogen action,” *Physiol Rev*, 81.

Noble, W.S. (2009) “How does multiple testing correction work?,” *Nat Biotechnol*, 27(12), pp. 1135–1137. Available at: https://doi.org/10.1038/nbt1209-1135.

Noel-MacDonnell, J.R. *et al.* (2018) “Assessment of data transformations for model-based clustering of RNA-Seq data,” *PLoS ONE*, 13(2), pp. 1–12. Available at: https://doi.org/10.1371/journal.pone.0191758.

Normanno, N. *et al.* (2005) “Mechanisms of endocrine resistance and novel therapeutic strategies in breast cancer,” *Endocr Relat Cancer*, 12(4), pp. 721–747. Available at: https://doi.org/10.1677/erc.1.00857.

Osborne, C.K. and Schiff, R. (2011) “MECHANISMS OF ENDOCRINE RESISTANCE IN BREAST CANCER,” *Annual review of medicine*, 62, pp. 233–247. Available at: https://doi.org/10.1146/annurev-med-070909-182917.

Osipo, C. *et al.* (2007) “Role for HER2/neu and HER3 in fulvestrant-resistant breast cancer,” *International journal of oncology*, 30(2), p. 509.

Osuchowska, P.N. *et al.* (2021) “Adhesion of Triple-Negative Breast Cancer Cells under Fluorescent and Soft X-ray Contact Microscopy.,” *International journal of molecular sciences*, 22(14). Available at: https://doi.org/10.3390/ijms22147279.

Otto, C. *et al.* (2008) “G protein-coupled receptor 30 localizes to the endoplasmic reticulum and is not activated by estradiol,” *Endocrinology*, 149(10), pp. 4846–4856. Available at: https://doi.org/10.1210/en.2008-0269.

Otto, C. *et al.* (2009) “GPR30 does not mediate oestrogenic responses in reproductive organs in mice,” *Biol Reprod*, 80(1), pp. 34–41. Available at: https://doi.org/10.1095/biolreprod.108.071175.

Pan, J. *et al.* (2021) “Construction on of a Ferroptosis-Related lncRNA-Based Model to Improve the Prognostic Evaluation of Gastric Cancer Patients Based on Bioinformatics,” *Front Genet*, 12, p. 739470. Available at: https://doi.org/10.3389/fgene.2021.739470.

Pang, Y. *et al.* (2019) “NEAT1/miR-124/STAT3 feedback loop promotes breast cancer progression,” *Int J Oncol*, 55(3), pp. 745–754. Available at: https://doi.org/10.3892/ijo.2019.4841.

Paplomata, E. and O’Regan, R. (2014) “The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers,” *Therapeutic Advances in Medical Oncology*, 6(4), pp. 154–166. Available at: https://doi.org/10.1177/1758834014530023.

Patel, N. *et al.* (2016) “Cost analysis of standard Sanger sequencing versus next generation sequencing in the ICONIC study,” *The Lancet*, 388, p. S86. Available at: https://doi.org/10.1016/s0140-6736(16)32322-4.

Peng, J. *et al.* (2020) “lncRNA NR2F1-AS1 Regulates miR-17/SIK1 Axis to Suppress the Invasion and Migration of Cervical Squamous Cell Carcinoma Cells,” *Reproductive Sciences*, 27(7), pp. 1534–1539. Available at: https://doi.org/10.1007/s43032-020-00149-y.

Penning, T.M. (2019) “AKR1C3 (type 5 17β-hydroxysteroid dehydrogenase/prostaglandin F synthase): Roles in malignancy and endocrine disorders.,” *Molecular and cellular endocrinology*, 489, pp. 82–91. Available at: https://doi.org/10.1016/j.mce.2018.07.002.

Perkins, J.R. *et al.* (2014) “A comparison of RNA-seq and exon arrays for whole genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat.,” *Molecular pain*, 10, p. 7. Available at: https://doi.org/10.1186/1744-8069-10-7.

Ponting, C.P., Oliver, P.L. and Reik, W. (2009) “Evolution and functions of long noncoding RNAs,” *Cell*, 136(4), pp. 629–641. Available at: https://doi.org/10.1016/j.cell.2009.02.006.

Qin, W. *et al.* (2010) “miR-24 Regulates Apoptosis by Targeting the Open Reading Frame (ORF) Region of FAF1 in Cancer Cells,” *PLOS ONE*, 5(2), p. e9429. Available at: https://doi.org/10.1371/journal.pone.0009429.

Rachoń, D. and Teede, H. (2010) “Ovarian function and obesity-Interrelationship, impact on women’s reproductive lifespan and treatment options,” *Molecular and Cellular Endocrinology*, 316(2), pp. 172–179. Available at: https://doi.org/10.1016/j.mce.2009.09.026.

Radin, D.P. and Patel, P. (2016) “Delineating the molecular mechanisms of tamoxifen’s oncolytic actions in oestrogen receptor-negative cancers,” *European Journal of Pharmacology*, 781, pp. 173–180. Available at: https://doi.org/10.1016/j.ejphar.2016.04.017.

Razandi, M. *et al.* (2003) “Identification of a Structural Determinant Necessary for the Localization and Function of Oestrogen Receptor at the Plasma Membrane,” *Molecular and Cellular Biology*, 23(5), pp. 1633–1646. Available at: https://doi.org/10.1128/mcb.23.5.1633-1646.2003.

Ren, X. and Kuan, P.-F. (2020) “Negative binomial additive model for RNA-Seq data analysis,” *BMC Bioinformatics*, 21(1), p. 171. Available at: https://doi.org/10.1186/s12859-020-3506-x.

Renhua, G. *et al.* (2016) “165P: Long noncoding RNA LUCAT1 is associated with poor prognosis in human non-small cell lung cancer and affects cell proliferation via regulating p21 and p57 expression,” *Journal of Thoracic Oncology*, 11(4), p. S129. Available at: https://doi.org/10.1016/S1556-0864(16)30275-1.

Ring, J.D. *et al.* (2017) “A performance evaluation of Nextera XT and KAPA HyperPlus for rapid Illumina library preparation of long-range mitogenome amplicons,” *Forensic*

Science International: Genetics, 29, pp. 174–180. Available at: https://doi.org/10.1016/j.fsigen.2017.04.003.

Riva, J.J. *et al.* (2012) “What is your research question? An introduction to the PICOT format for clinicians.,” *The Journal of the Canadian Chiropractic Association*, 56(3), pp. 167–171.

Rondón-Lagos, M. *et al.* (2016) “Tamoxifen resistance: Emerging molecular targets,” *International Journal of Molecular Sciences*, 17(8), pp. 1–31. Available at: https://doi.org/10.3390/ijms17081357.

Ross, M.G. *et al.* (2013) “Characterizing and measuring bias in sequence data,” *Genome Biology*, 14(5), p. R51. Available at: https://doi.org/10.1186/gb-2013-14-5-r51.

Rothé, F. *et al.* (2011) “Global microRNA expression profiling identifies MiR-210 associated with tumor proliferation, invasion and poor clinical outcome in breast cancer,” *PLoS ONE*, 6(6). Available at: https://doi.org/10.1371/journal.pone.0020980.

Sambrook, J. and Russell, D.W. (2006) “Fragmentation of DNA by nebulization.,” *CSH protocols*, 2006(4). Available at: https://doi.org/10.1101/pdb.prot4539.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–5467. Available at: https://doi.org/10.1073/pnas.74.12.5463.

Scacheri, P.C. *et al.* (2004) “Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells,” *Proceedings of the National Academy of Sciences of the United States of America*, 101(7), pp. 1892–1897. Available at: https://doi.org/10.1073/pnas.0308698100.

Schmitz, K.H. *et al.* (2012) “Prevalence of breast cancer treatment sequelae over 6 years of follow-up: The pulling through study,” *Cancer*, 118(SUPPL.8), pp. 2217–2225. Available at: https://doi.org/10.1002/cncr.27474.

Seelam, P.P., Sharma, P. and Mitra, A. (2017) “Structural landscape of base pairs containing post-transcriptional modifications in RNA,” *Rna*, 23(6), pp. 847–859. Available at: https://doi.org/10.1261/rna.060749.117.

Seksenyan, A. *et al.* (2015) “TOX3 is expressed in mammary ER(+) epithelial cells and regulates ER target genes in luminal breast cancer,” *BMC Cancer*, 15(1), p. 22. Available at: https://doi.org/10.1186/s12885-015-1018-2.

Shao, M. *et al.* (2020) “Survival analysis for long noncoding RNAs identifies TP53TG1 as an antioncogenic target for the breast cancer.,” *Journal of cellular physiology*, 235(10), pp. 6574–6581. Available at: https://doi.org/10.1002/jcp.29517.

Shen, Q., Xu, Z. and Xu, S. (2020) “Long non-coding RNA LUCAT1 contributes to cisplatin resistance by regulating the miR-514a-3p/ULK1 axis in human non-small cell lung cancer,” *International Journal of Oncology*, 57(4), pp. 967–979. Available at: https://doi.org/10.3892/ijo.2020.5106.

Sheng, X.-Y. *et al.* (2020) “Long-Chain Non-Coding SOX21-AS1 Promotes Proliferation and Migration of Breast Cancer Cells Through the PI3K/AKT Signaling Pathway.,” *Cancer management and research*, 12, pp. 11005–11014. Available at: https://doi.org/10.2147/CMAR.S270464.

Shi, S. *et al.* (2020) “CMA1 is potent prognostic marker and associates with immune infiltration in gastric cancer,” *Autoimmunity*, 53(4), pp. 210–217. Available at: https://doi.org/10.1080/08916934.2020.1735371.

Shi, Z. *et al.* (2021) “Long non-coding RNA SNHG8 promotes prostate cancer progression through repressing miR-384 and up-regulating HOXB7,” *Journal of Gene Medicine*. Available at: https://doi.org/10.1002/jgm.3309.

Shyu, A. bin, Wilkinson, M.F. and van Hoof, A. (2008) “Messenger RNA regulation: To translate or to degrade,” *EMBO Journal*, 27(3), pp. 471–481. Available at: https://doi.org/10.1038/sj.emboj.7601977.

Silva, A., Bullock, M. and Calin, G. (2015) “The clinical relevance of long non-coding RNAs in cancer,” *Cancers*, 7(4), pp. 2169–2182. Available at: https://doi.org/10.3390/cancers7040884.

Singh, A., Nunes, J.J. and Ateeq, B. (2015) “Role and therapeutic potential of G-protein coupled receptors in breast cancer progression and metastases,” *European Journal of Pharmacology*, 763, pp. 178–183. Available at: https://doi.org/https://doi.org/10.1016/j.ejphar.2015.05.011.

Singh, A.P. *et al.* (2022) “A coordinated function of lncRNA HOTTIP and miRNA-196b underpinning leukemogenesis by targeting FAS signaling,” *Oncogene*, 41(5), pp. 718–731. Available at: https://doi.org/10.1038/s41388-021-02127-3.

de Souza, N. (2012) “Genomics: The ENCODE project,” *Nature Methods*, 9(11), p. 1046. Available at: https://doi.org/10.1038/nmeth.2238.

Srivastava, A.K. *et al.* (2014) “Appraisal of diagnostic ability of UCA1 as a biomarker of carcinoma of the urinary bladder,” *Tumor Biology*, 35(11), pp. 11435–11442. Available at: https://doi.org/10.1007/s13277-014-2474-z.

Stuenkel, C.A. *et al.* (2015) “Treatment of Symptoms of the Menopause: An Endocrine Society Clinical Practice Guideline,” *J Clin Endocrinol Metab*. 2015/10/08, 100(11), pp. 3975–4011. Available at: https://doi.org/10.1210/jc.2015-2236.

Subramanian, A. *et al.* (2005) “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545–15550. Available at: https://doi.org/10.1073/pnas.0506580102.

Sun, B.K., Deaton, A.M. and Lee, J.T. (2006) “A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization,” *Mol Cell*, 21(5), pp. 617–628. Available at: https://doi.org/10.1016/j.molcel.2006.01.028.

Sun, T. *et al.* (2020) “LNC942 promoting METTL14-mediated m(6)A methylation in breast cancer cell proliferation and progression.,” *Oncogene*, 39(31), pp. 5358–5372. Available at: https://doi.org/10.1038/s41388-020-1338-9.

Sung, H. *et al.* (2021) “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, 0(0), pp. 1–41. Available at: https://doi.org/10.3322/caac.21660.

Szymański, M. *et al.* (2003) “Noncoding RNA transcripts.,” *Journal of applied genetics*, 44(1), pp. 1–19.

T2T Consortium (2022) *T2T-CHM13v2.0 complete genome assembly*, *NCBI*. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4#/st.

Taft, R.J., Pheasant, M. and Mattick, J.S. (2007) “The relationship between non-protein-coding DNA and eukaryotic complexity,” *Bioessays*, 29(3), pp. 288–299. Available at: https://doi.org/10.1002/bies.20544.

Takahashi, M.K. *et al.* (2015) “Rapidly Characterizing the Fast Dynamics of RNA Genetic Circuitry with Cell-Free Transcription-Translation (TX-TL) Systems,” *ACS Synthetic Biology*, 4(5), pp. 503–515. Available at: https://doi.org/10.1021/sb400206c.

Terai, G. *et al.* (2016) “Comprehensive prediction of lncRNA–RNA interactions in human transcriptome,” *BMC Genomics*, 17(1), p. 12. Available at: https://doi.org/10.1186/s12864-015-2307-5.

Thomas, S.M. *et al.* (2006) “Cross-talk between G Protein–Coupled Receptor and Epidermal Growth Factor Receptor Signaling Pathways Contributes to Growth and Invasion of Head and Neck Squamous Cell Carcinoma,” *Cancer Res*, 66(24), p. 11831. Available at: http://cancerres.aacrjournals.org/content/66/24/11831.abstract.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.,” *Contemporary oncology (Poznan, Poland)*, 19(1A), pp. A68-77. Available at: https://doi.org/10.5114/wo.2014.47136.

Tong, Y. (2021) “The comparison of limma and DESeq2 in gene analysis,” *E3S Web of Conferences*, 271, p. 3058. Available at: https://doi.org/10.1051/e3sconf/202127103058.

Totomoch-Serra, A., Marquez, M.F. and Cervantes-Barragán, D.E. (2017) “Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome,” *F1000Research*, 6(0), pp. 1–7. Available at: https://doi.org/10.12688/f1000research.11610.1.

Tsoi, L.C. and Zheng, W.J. (2007) “A method of microarray data storage using array data type,” *Computational Biology and Chemistry*, 31(2), pp. 143–147. Available at: https://doi.org/10.1016/j.compbiolchem.2007.01.004.

Turcatti, G. *et al.* (2008) “A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis,” *Nucleic Acids Research*, 36(4). Available at: https://doi.org/10.1093/nar/gkn021.

Turner, N.C. and Reis-Filho, J.S. (2006) “Basal-like breast cancer and the BRCA1 phenotype,” *Oncogene*, 25(43), pp. 5846–5853. Available at: https://doi.org/10.1038/sj.onc.1209876.

Vanni, I. *et al.* (2015) “Next-generation sequencing workflow for NSCLC critical samples using a targeted sequencing approach by ion torrent PGM™ platform,” *International Journal of Molecular Sciences*, 16(12), pp. 28765–28782. Available at: https://doi.org/10.3390/ijms161226129.

Vasconcelos, I. *et al.* (2016) “The St. Gallen surrogate classification for breast cancer subtypes successfully predicts tumor presenting features, nodal involvement, recurrence patterns and disease free survival,” *Breast*, 29, pp. 181–185. Available at: https://doi.org/10.1016/j.breast.2016.07.016.

Veeraraghavan, J. *et al.* (2014) “Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of ooestrogen receptor-positive breast cancers,” *Nat Commun*, 5, p. 4577. Available at: https://doi.org/10.1038/ncomms5577.

Vinogradova, Y., Coupland, C. and Hippisley-Cox, J. (2020) “Use of hormone replacement therapy and risk of breast cancer: Nested case-control studies using the QResearch and CPRD databases,” *The BMJ*, 371. Available at: https://doi.org/10.1136/bmj.m3873.

Walpole, A.L. and Harper, M.J.K. (1966) “Contrasting Endocrine activities of cis and trans isomers in a series of substituted triphenylethylenes,” *Nature*, 212(5057), p. 87.

Wan, G., Hu, X., *et al.* (2013) “A novel non-coding RNA lncRNA-JADE connects DNA damage signalling to histone H4 acetylation,” *EMBO J*, 32(21), pp. 2833–2847. Available at: https://doi.org/10.1038/emboj.2013.221.

Wan, G., Mathur, R., *et al.* (2013) “Long non-coding RNA ANRIL (CDKN2B-AS) is induced by the ATM-E2F1 signaling pathway,” *Cell Signal*, 25(5), pp. 1086–1095. Available at: https://doi.org/10.1016/j.cellsig.2013.02.006.

Wang, F. *et al.* (2021) “Long non-coding RNA SOX21-AS1 modulates lung cancer progress upon microRNA miR-24-3p/PIM2 axis,” *Bioengineered*, 12(1), pp. 6724–6737. Available at: https://doi.org/10.1080/21655979.2021.1955578.

Wang, J. *et al.* (2019) “The long noncoding RNA H19 promotes tamoxifen resistance in breast cancer via autophagy,” *J Hematol Oncol*, 12(1), p. 81. Available at: https://doi.org/10.1186/s13045-019-0747-0.

Wang, J. *et al.* (2020) “LncRNA NR2F1-AS1 Regulates miR-371a-3p/TOB1 Axis to Suppress Proliferation of Colorectal Cancer Cells,” *Cancer Biotherapy and Radiopharmaceuticals*, 35(10), pp. 760–764. Available at: https://doi.org/10.1089/cbr.2019.3237.

Wang, K.C. and Chang, H.Y. (2011) “Molecular mechanisms of long noncoding RNAs,” *Mol Cell*, 43(6), pp. 904–914. Available at: https://doi.org/10.1016/j.molcel.2011.08.018.

Wang, L. *et al.* (2013) “CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model,” *Nucleic Acids Research*, 41(6), p. e74. Available at: https://doi.org/10.1093/nar/gkt006.

Wang, L.N. *et al.* (2018) “Long noncoding RNA lung cancer associated transcript 1 promotes proliferation and invasion of clear cell renal cell carcinoma cells by

negatively regulating miR-495-3p," *Journal of Cellular Biochemistry*, 119(9), pp. 7599–7609. Available at: https://doi.org/10.1002/jcb.27099.

Wang, P.-Y. *et al.* (2017) "Single nucleotide polymorphisms in ZNRD1-AS1 increase cancer risk in an Asian population.," *Oncotarget*, 8(6), pp. 10064–10070. Available at: https://doi.org/10.18632/oncotarget.14334.

Wang, Z., Gerstein, M. and Snyder, M. (2010) "RNA-Seq: a revolutionary tool for transcriptomics," 10(1), pp. 57–63. Available at: https://doi.org/10.1038/nrg2484.RNA-Seq.

Wei, A. (2017) "Biomedicine & Pharmacotherapy Long non-coding RNA SOX21-AS1 sponges miR-145 to promote the tumorigenesis of colorectal cancer by targeting MYO6," *Biomedicine & Pharmacotherapy*, 96(8), pp. 953–959. Available at: https://doi.org/10.1016/j.biopha.2017.11.145.

Wei, S. *et al.* (2016) "PU.1 controls the expression of long noncoding RNA HOTAIRM1 during granulocytic differentiation," *Journal of Hematology & Oncology*, 9(1). Available at: https://doi.org/10.1186/s13045-016-0274-1.

Wei, W. *et al.* (2017) "LncRNA XIST Promotes Pancreatic Cancer Proliferation Through miR-133a/EGFR," *Journal of Cellular Biochemistry*, pp. 3349–3358. Available at: https://doi.org/10.1002/jcb.25988.

Weiss, R.A. (1998) "Viral RNA-dependent DNA polymerase RNA-dependent DNA polymerase in virions of Rous sarcoma virus," *Reviews in Medical Virology*, 8(1), pp. 3–11. Available at: https://doi.org/10.1002/(sici)1099-1654(199801/03)8:1<3::aid-rmv218>3.0.co;2-%23.

Weißbach, S. *et al.* (2021) “Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines,” *BMC Genomics*, 22(1), pp. 1–15. Available at: https://doi.org/10.1186/s12864-020-07362-8.

Williamson, J.G. and Ellis, J.D. (1973) “the Induction of Ovulation By Tamoxifen,” *BJOG: An International Journal of Obstetrics & Gynaecology*, 80(9), pp. 844–847. Available at: https://doi.org/10.1111/j.1471-0528.1973.tb11230.x.

Winter, G.E. *et al.* (2014) “The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity.,” *Nature chemical biology*, 10(9), pp. 768–773. Available at: https://doi.org/10.1038/nchembio.1590.

Witten, D.M. (2011) “Classification and clustering of sequencing data using a poisson model,” *Annals of Applied Statistics*, 5(4), pp. 2493–2518. Available at: https://doi.org/10.1214/11-AOAS493.

Wood, E. *et al.* (2019) “Clinical long-read sequencing of the human mitochondrial genome for mitochondrial disease diagnostics,” *bioRxiv*, p. 597187. Available at: https://doi.org/10.1101/597187.

Wu, B. and White, K.A. (2007) “Uncoupling RNA virus replication from transcription via the polymerase: Functional and evolutionary insights,” *EMBO Journal*, 26(24), pp. 5120–5130. Available at: https://doi.org/10.1038/sj.emboj.7601931.

Wu, L. *et al.* (2016) “A new avenue for obtaining insight into the functional characteristics of long noncoding RNAs associated with oestrogen receptor signaling,” *Sci Rep*, 6, p. 31716. Available at: https://doi.org/10.1038/srep31716.

Wu, R. *et al.* (2020) “The long noncoding RNA LUCAT1 promotes colorectal cancer cell proliferation by antagonizing Nucleolin to regulate MYC expression,” *Cell Death and Disease*, 11(10). Available at: https://doi.org/10.1038/s41419-020-03095-4.

Wu, W. *et al.* (2018) “LncRNA NKILA suppresses TGFвЂђОІвЂђinduced epithelialвЂ“mesenchymal transition by blocking NFвЂђОєB signaling in breast cancer,” *Int J Cancer*, 143(9), pp. 2213–2224. Available at: https://doi.org/10.1002/ijc.31605.

Wu, Z. *et al.* (2022) “LINC00339: An emerging major player in cancer and metabolic diseases,” *Biomedicine & Pharmacotherapy*, 149, p. 112788. Available at: https://doi.org/https://doi.org/10.1016/j.biopha.2022.112788.

Xia, Z. *et al.* (2014) “Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3’2-UTR landscape across seven tumour types,” *Nature Communications*, 5. Available at: https://doi.org/10.1038/ncomms6274.

Xiao, H. *et al.* (2018) “TP53TG1 enhances cisplatin sensitivity of non-small cell lung cancer cells through regulating miR-18a/PTEN axis,” *Cell & Bioscience*, 8(1), p. 23. Available at: https://doi.org/10.1186/s13578-018-0221-7.

Xu, Y. *et al.* (2021) “Long Noncoding RNA NR2F1-AS1 Enhances the Migration and Invasion of Hepatocellular Carcinoma via Modulating miR-642a/DEK Pathway,” *Journal of Oncology*, 2021. Available at: https://doi.org/10.1155/2021/6868514.

Xuan, L. *et al.* (2021) “lncRNA SNHG8 promotes ovarian cancer progression through serving as sponge for miR-1270 to regulate S100A11 expression,” *The Journal of Gene Medicine*, pp. 0–2. Available at: https://doi.org/10.1002/jgm.3315.

Xue, X. *et al.* (2016) “LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer,” *Oncogene*, 35(21), pp. 2746–2755. Available at: https://doi.org/10.1038/onc.2015.340.

Yaghjyan, L. and Colditz, G.A. (2011) “Oestrogens in the breast tissue: A systematic review,” *Cancer Causes and Control*, 22(4), pp. 529–540. Available at: https://doi.org/10.1007/s10552-011-9729-4.

Yamaguchi, N. *et al.* (2019) “PCK1 and DHODH drive colorectal cancer liver metastatic colonization and hypoxic growth by promoting nucleotide synthesis,” *eLife*, 8, pp. 1–26. Available at: https://doi.org/10.7554/eLife.52135.

Yamamoto, Y. and Gaynor, R.B. (2001) “Therapeutic potential of inhibition of the NF-κB pathway in the treatment of inflammation and cancer,” *Journal of Clinical Investigation*, 107(2), pp. 135–142.

Yamashita, H. *et al.* (2006) “p53 protein accumulation predicts resistance to endocrine therapy and decreased post-relapse survival in metastatic breast cancer,” *Breast Cancer Research*, 8(4), pp. R48–R48. Available at: https://doi.org/10.1186/bcr1536.

Yang, C.M. *et al.* (2016) “Aberrant DNA hypermethylation-silenced SOX21-AS1 gene expression and its clinical importance in oral cancer,” *Clinical epigenetics*, 8, p. 129. Available at: https://doi.org/10.1186/s13148-016-0291-5.

Yao, Z. *et al.* (2013) “Discordance and clinical significance of ER, PR, and HER2 status between primary breast cancer and synchronous axillary lymph node metastasis,” *Medical Oncology*, 31(1), p. 798. Available at: https://doi.org/10.1007/s12032-013-0798-y.

Yu, H. *et al.* (2018) “Long noncoding RNA LUCAT1 promotes malignancy of ovarian cancer through regulation of miR-612/HOXA13 pathway,” *Biochemical and Biophysical Research Communications*, 503(3), pp. 2095–2100. Available at: https://doi.org/10.1016/j.bbrc.2018.07.165.

Yu, H. *et al.* (2020) “The prognostic value of long non-coding RNA H19 in various cancers: A meta-analysis based on 15 studies with 1584 patients and the Cancer Genome Atlas data,” *Medicine (United States)*, 99(2). Available at: https://doi.org/10.1097/MD.0000000000018533.

Zerbino, D.R. *et al.* (2018) “Ensembl 2018,” *Nucleic Acids Research*, 46(D1), pp. D754–D761. Available at: https://doi.org/10.1093/nar/gkx1098.

Zhang, B. *et al.* (2007) “microRNAs as oncogenes and tumor suppressors,” *Developmental Biology*, 302(1), pp. 1–12. Available at: https://doi.org/10.1016/j.ydbio.2006.08.028.

Zhang, H. *et al.* (2021) “LncRNA ATXN8OS enhances tamoxifen resistance in breast cancer,” *Open Medicine (Poland)*, 16(1), pp. 68–80. Available at: https://doi.org/10.1515/med-2021-0012.

Zhang, L. *et al.* (2019) “SP1-induced up-regulation of lncRNA LUCAT1 promotes proliferation, migration and invasion of cervical cancer by sponging miR-181a,” *Artificial Cells, Nanomedicine and Biotechnology*, 47(1), pp. 556–564. Available at: https://doi.org/10.1080/21691401.2019.1575840.

Zhang, M. *et al.* (2020) “LncRNA GATA3вЂђAS1 facilitates tumour progression and immune escape in tripleвЂђnegative breast cancer through destabilization of GATA3 but stabilization of PDвЂђL1,” *Cell Prolif*, 53(9), pp. e12855-n/a. Available at: https://doi.org/10.1111/cpr.12855.

ZHANG, M.H. *et al.* (2014) “Oestrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (Review),” *Biomedical Reports*, 2(1), pp. 41–52. Available at: https://doi.org/10.3892/br.2013.187.

Zhang, Q. *et al.* (2018) “XIST promotes gastric cancer (GC) progression through TGF-β1 via targeting miR-185,” *Journal of Cellular Biochemistry*, pp. 2787–2796. Available at: https://doi.org/10.1002/jcb.26447.

Zhang, S.J. *et al.* (2013) “Expression and significance of ER, PR, VEGF, CA15-3, CA125 and CEA in judging the prognosis of breast cancer,” *Asian Pacific Journal of Cancer Prevention*, 14(6), pp. 3937–3940. Available at: https://doi.org/10.7314/APJCP.2013.14.6.3937.

Zhang, W. *et al.* (2012) “miR-181d: a predictive glioblastoma biomarker that downregulates MGMT expression.,” *Neuro-oncology*, 14(6), pp. 712–719. Available at: https://doi.org/10.1093/neuonc/nos089.

Zhang, Xueqing *et al.* (2009) “A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster.,” *Blood*, 113(11), pp. 2526–2534. Available at: https://doi.org/10.1182/blood-2008-06-162164.

Zhang, X *et al.* (2009) “A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster,” *Blood*, 113(11), pp. 2526–2534. Available at: https://doi.org/10.1182/blood-2008-06-162164.

Zhang, X. *et al.* (2019) “LncRNA MACC1-AS1 sponges multiple miRNAs and RNA-binding protein PTBP1,” *Oncogenesis*, 8(12). Available at: https://doi.org/10.1038/s41389-019-0182-7.

Zhang, X., Weissman, Sherman M and Newburger, P.E. (2014) “Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells.,” *RNA biology*, 11(6), pp. 777–787. Available at: https://doi.org/10.4161/rna.28828.

Zhang, X., Weissman, Sherman M. and Newburger, P.E. (2014) “Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells,” *RNA Biology*, 11(6). Available at: https://doi.org/10.4161/rna.28828.

Zhang, Y. *et al.* (2019) “Long noncoding RNA TP53TG1 promotes pancreatic ductal adenocarcinoma development by acting as a molecular sponge of microRNA-96.,” *Cancer science*, 110(9), pp. 2760–2772. Available at: https://doi.org/10.1111/cas.14136.

Zhao, Y. *et al.* (2020) “Long noncoding RNA HOTAIRM1 in human cancers.,” *Clinica chimica acta; international journal of clinical chemistry*, 511, pp. 255–259. Available at: https://doi.org/10.1016/j.cca.2020.10.011.

Zhong, Y. and Zeng, W. (2022) “NR2F1-AS1 Acts as an Oncogene in Breast Cancer by Competitively Binding with miR-641,” *J Healthc Eng*, 2022, p. 6778199. Available at: https://doi.org/10.1155/2022/6778199.

Zhou, Y. *et al.* (2005) “The NFkappaB pathway and endocrine-resistant breast cancer,” *Endocr Relat Cancer*, 12 Suppl 1, pp. S37-46. Available at: https://doi.org/10.1677/erc.1.00977.

Zhuang, H. *et al.* (2019) “Overexpressed lncRNA CDKN2B-AS1 is an independent prognostic factor for liver cancer and promotes its proliferation.,” *Journal of B.U.ON. : official journal of the Balkan Union of Oncology*, 24(4), pp. 1441–1448.

Zwiener, I., Frisch, B. and Binder, H. (2014) “Transforming RNA-Seq data to improve the performance of prognostic gene signatures,” *PLoS ONE*, 9(1), pp. 1–13. Available at: https://doi.org/10.1371/journal.pone.0085150.

Appendix

Table 1 Prioritised lncRNAs upregulated in TAMR cells

	ensembl_gene_id	external_gene_name	baseMean	log2FoldChange	lfcSE	pvalue	padj
1	ENSG00000197301	HMGA2-AS1	25.05572602	-8.851355246	2.820294481	1.16E-10	1.02389E-09
2	ENSG00000248323	LUCAT1	17.0438206	-8.17352108	2.749795999	2.81E-09	2.08615E-08
3	ENSG00000250266	LINC01612	11.74716322	-7.497809528	2.687526209	6.49E-08	3.99482E-07
4	ENSG00000231683	na	34.79485654	-7.212641057	1.298035761	4.61E-10	3.77571E-09
5	ENSG00000227640	SOX21-AS1	61.81278581	-5.811933421	0.667590309	1.84E-18	3.17557E-17
6	ENSG00000235770	LINC00607	35.88862503	-4.256486108	0.56012439	6.56E-15	8.62992E-14
7	ENSG00000236651	DLX2-DT	21.10114095	-3.731187172	0.627504169	5.1E-10	4.13564E-09
8	ENSG00000237187	NR2F1-AS1	91.57338412	-3.476229928	0.453015277	1.73E-15	2.38034E-14
9	ENSG00000231185	SPRY4-AS1	37.51520264	-3.468792847	0.467410452	1.9E-14	2.41775E-13
10	ENSG00000236714	LINC01844	17.16702041	-3.382548377	0.671066262	7E-08	4.28236E-07
11	ENSG00000257842	NOVA1-DT	74.90697817	-3.298047728	0.359820312	6.2E-21	1.25509E-19
12	ENSG00000230126	FGF12-AS2	25.42660312	-3.207412817	0.531730025	2.37E-10	2.01336E-09
13	ENSG00000250295	RDH10-AS1	23.75927853	-3.062927828	0.565945714	7.43E-09	5.23465E-08
14	ENSG00000249628	LINC00942	23.11475092	-2.897886944	0.578723452	6.16E-08	3.80383E-07
15	ENSG00000231776	LINC01611	17.74706973	-2.7768074	0.597356625	3.77E-07	2.08372E-06
16	ENSG00000203721	LINC00862	55.88612405	-2.551037385	0.361738362	1.8E-13	2.08609E-12
17	ENSG00000233429	HOTAIRM1	22.56620131	-2.500997587	0.601596247	3.72E-06	1.78725E-05
18	ENSG00000189196	LINC00994	188.849314	-2.42264431	0.219269944	2.03E-29	6.41732E-28
19	ENSG00000226383	LINC01876	37.3554781	-2.418727981	0.381400628	2.84E-11	2.65278E-10
20	ENSG00000251138	LINC02882	30.92034239	-2.36288442	0.448552038	1.77E-08	1.17953E-07
21	ENSG00000241359	SYNPR-AS1	14.82027978	-2.333579518	0.638752144	3.03E-05	0.000124293
22	ENSG00000241111	PRICKLE2-AS1	342.2366951	-2.243213416	0.158677926	2.78E-46	1.6594E-44
23	ENSG00000227036	LINC00511	31.00389315	-2.098839981	0.425159653	9.86E-08	5.91245E-07
24	ENSG00000250548	LINC01303	22.138081	-1.895056661	0.46560589	6.61E-06	3.03661E-05
25	ENSG00000236432	MFF-DT	22.06083272	-1.765508291	0.462635198	2.09E-05	8.79911E-05
26	ENSG00000235529	AGAP1-IT1	38.9152265	-1.76442891	0.406227015	2.16E-06	1.07798E-05
27	ENSG00000225953	SATB2-AS1	47.06643079	-1.708380044	0.360652871	3.51E-07	1.9502E-06
28	ENSG00000251432	LINC02615	51.19226167	-1.669286165	0.336163491	1.19E-07	7.09357E-07
29	ENSG00000226067	LINC00623	180.5334226	-1.600593124	0.214993796	1.79E-14	2.28058E-13
30	ENSG00000250786	SNHG18	74.63359773	-1.578218131	0.311554919	7.28E-08	4.43883E-07
31	ENSG00000226476	LINC01748	28.51313219	-1.501403002	0.404168154	4.04E-05	0.00016167

Table 2 Prioritised lncRNAs doregulated in TAMR cells

	ensembl_gene_id	external_gene_name	baseMean	log2FoldChange	lfcSE	pvalue	padj
1	ENSG00000251129	LINC02506	73.17295375	8.335559324	1.29353	5.3196E-13	5.92492E-12
2	ENSG00000196668	LINC00173	16.2768872	8.008662123	2.669429	4.39555E-09	3.17218E-08
3	ENSG00000235123	DSCAM-AS1	20002.90821	6.735506235	0.332166	1.35483E-92	3.08066E-90
4	ENSG00000249421	ADAMTS19-AS1	20.97909443	5.040796291	0.912942	1.5268E-08	1.02248E-07
5	ENSG00000225362	CT62	113.3227372	4.897281009	0.413158	3.25219E-33	1.16668E-31
6	ENSG00000245904	BTG1-DT	19.72078462	3.855285689	0.891783	1.43301E-06	7.32108E-06
7	ENSG00000245060	LINC00847	183.8176812	3.189294663	0.267364	7.43818E-34	2.75529E-32
8	ENSG00000244161	FLNB-AS1	48.02732743	2.967889135	0.371596	1.73149E-16	2.55658E-15
9	ENSG00000232352	SEMA3B-AS1	15.54650222	2.816615484	0.63761	1.25457E-06	6.47436E-06
10	ENSG00000232624	LINC01517	14.75900804	2.713757205	0.659308	4.61249E-06	2.17752E-05
11	ENSG00000236581	STARD13-AS	44.18117173	2.643104395	0.443072	2.55502E-10	2.15175E-09
12	ENSG00000249846	LINC02021	48.68117328	2.623974679	0.340617	1.71872E-15	2.37408E-14
13	ENSG00000249346	LINC01016	44.32662342	2.527297159	0.555693	4.94355E-07	2.6902E-06
14	ENSG00000246695	RASSF8-AS1	34.32273203	2.233286257	0.393816	1.90526E-09	1.43892E-08
15	ENSG00000247809	NR2F2-AS1	19.75900329	2.110613376	0.535522	1.0475E-05	4.66953E-05
16	ENSG00000223749	MIR503HG	42.15565348	2.004706279	0.550357	3.09818E-05	0.00012667
17	ENSG00000267131	na	92.82991853	1.991241517	0.472151	2.95021E-06	1.4386E-05
18	ENSG00000224424	PRKAR2A-AS1	27.76847063	1.901905201	0.487985	1.31475E-05	5.74401E-05
19	ENSG00000243701	DUBR	72.64471061	1.901163348	0.338972	2.9592E-09	2.19233E-08
20	ENSG00000231107	LINC01508	47.16095784	1.88933689	0.451283	3.84281E-06	1.84106E-05
21	ENSG00000273018	FAM106A	15.12403021	1.852616361	0.583775	0.000207319	0.000728699
22	ENSG00000225077	ICMT-DT	25.83824463	1.844572261	0.542297	8.83504E-05	0.000333202
23	ENSG00000262155	LINC02175	26.49943504	1.787980611	0.47712	2.5429E-05	0.000105573
24	ENSG00000255100	TSKU-AS1	21.07523663	1.745081826	0.516882	0.000109731	0.000405761
25	ENSG00000245888	NSMCE1-DT	21.70646887	1.613683721	0.485786	0.000151045	0.000543834
26	ENSG00000229267	SNHG31	32.74171109	1.600995598	0.377066	4.06377E-06	1.9375E-05

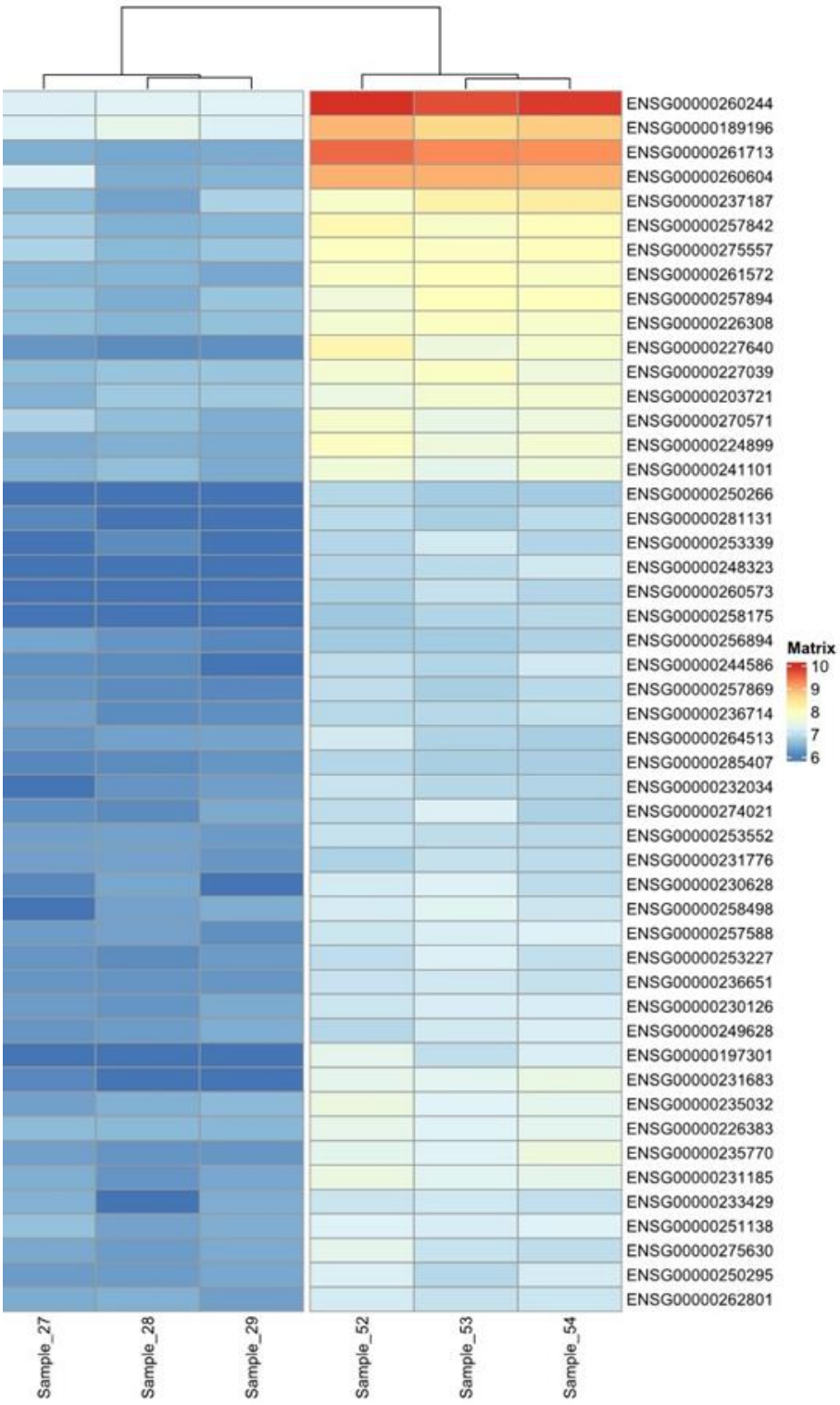


Figure 1 Top 50 lncRNAs up regulated in TAMR

Deseq2 Clustering on

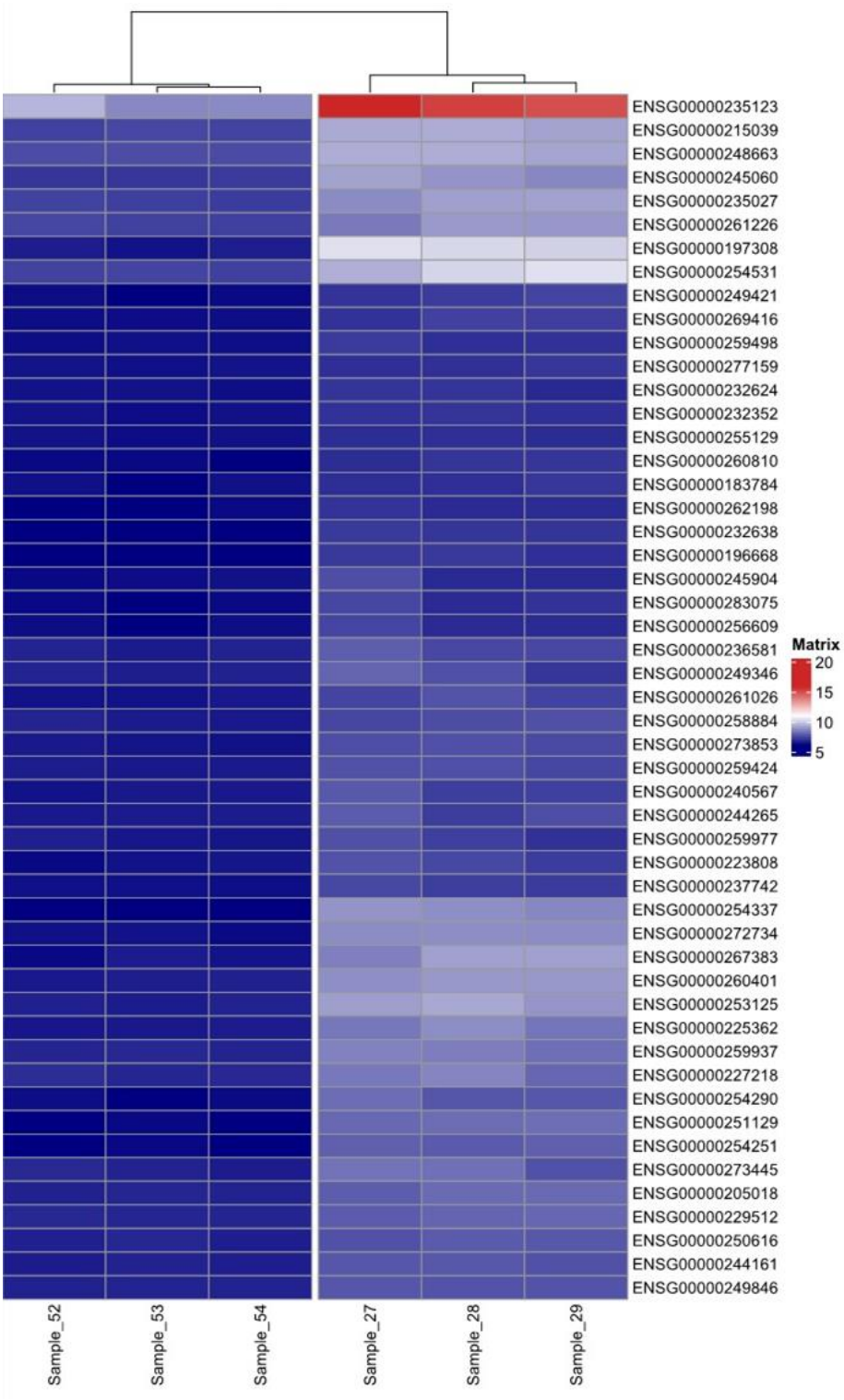


Figure 2 Top 50 lncRNAs down regulated in

TAMR Deseq2 Clustering on

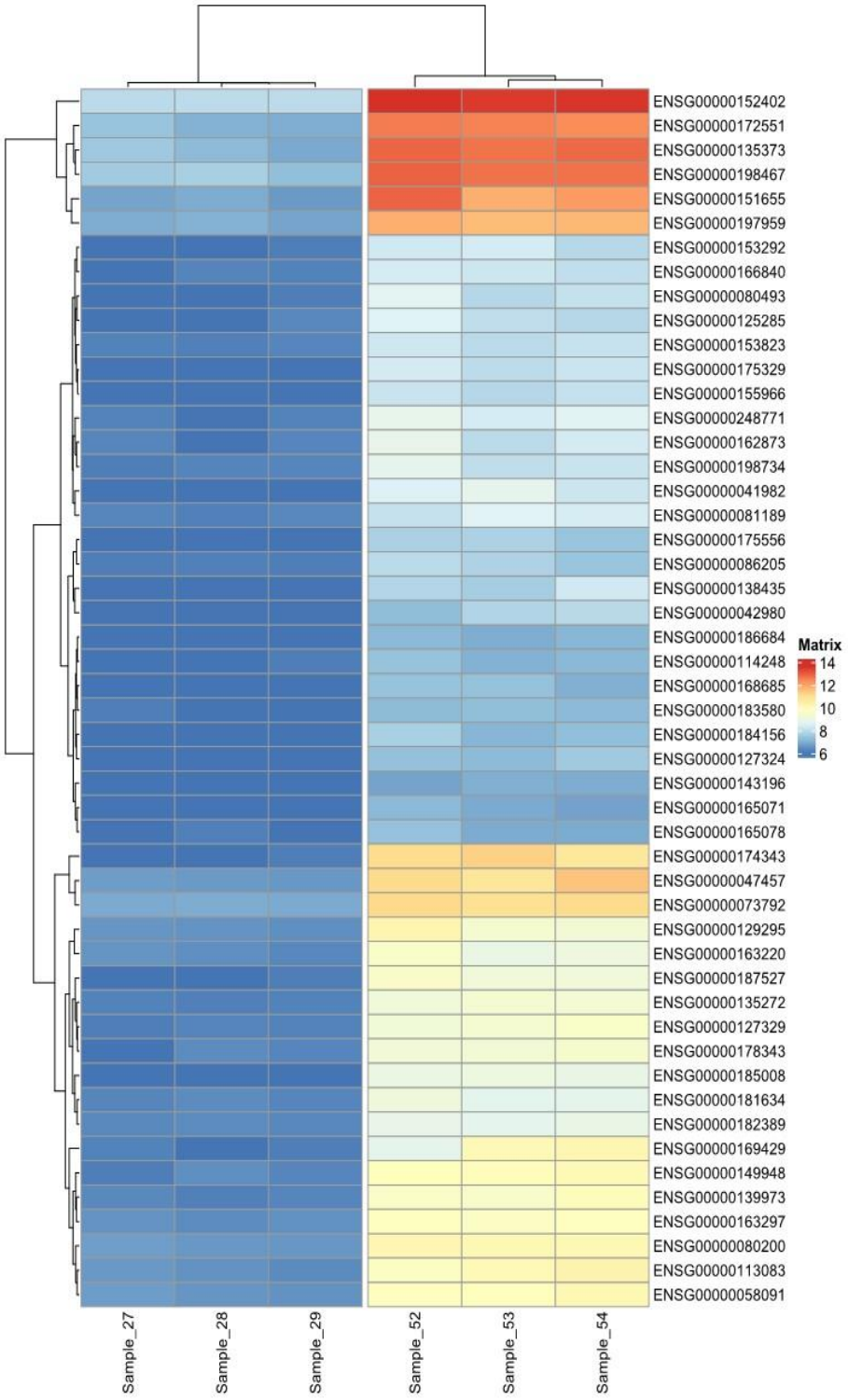


Figure 3 Top 50 protein coding genes up regulated in TAMR

Deseq2 Clustering on

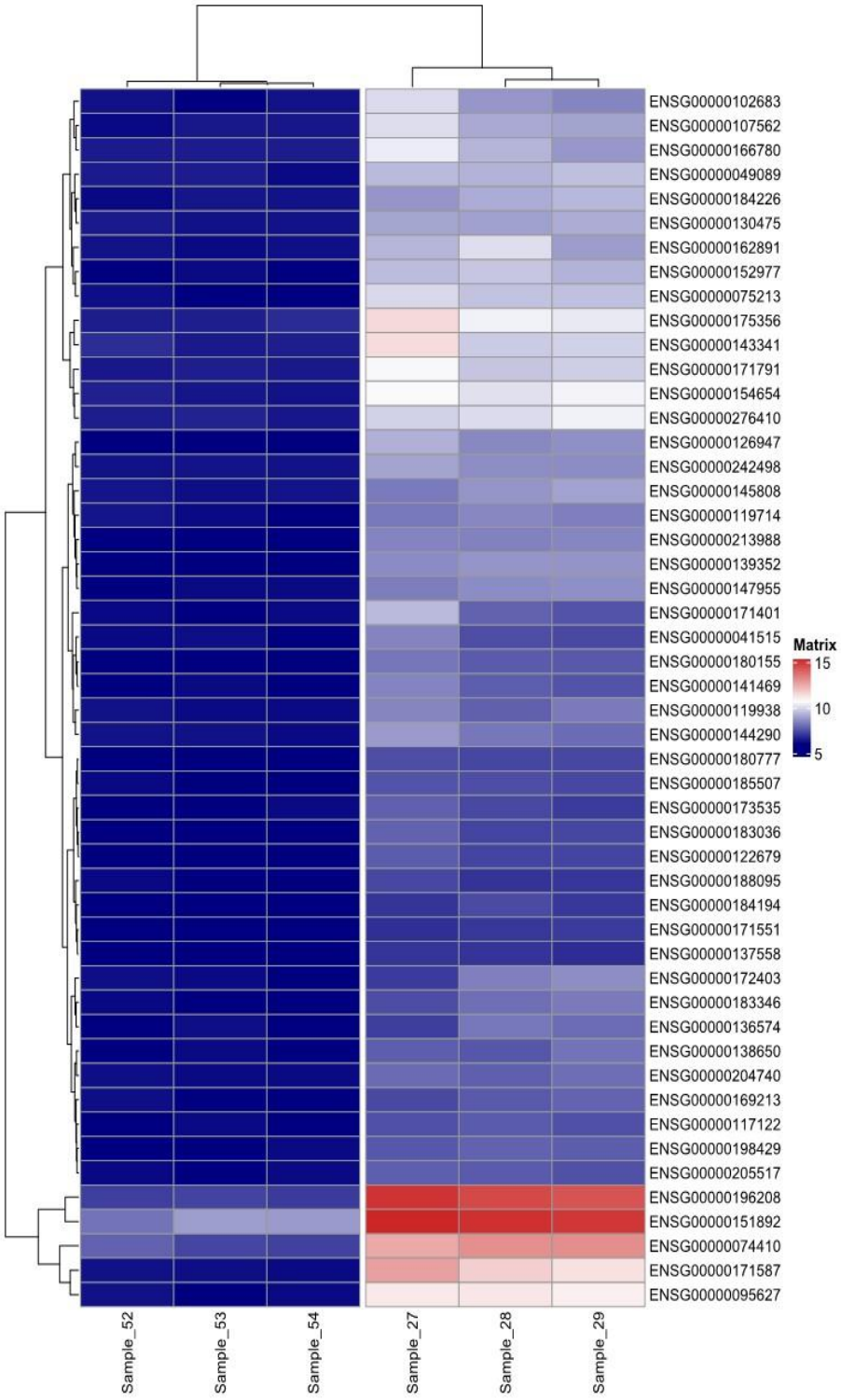


Figure 4 Top 50 protein coding genes down regulated in TAMR

Deseq2 Clustering on