A cross-platform analysis of lncRNA expression associated with estrogen receptor status and methylation in breast cancer

Katrine Bull Evensen

Thesis for the Master of Science degree in Molecular Biosciences

Department of Molecular Biosciences
Faculty of Mathematics and Natural Biosciences

UNIVERSITY OF OSLO

August 2019

A cross-platform analysis of lncRNA expression associated with estrogen receptor status and methylation in breast cancer

Katrine Bull Evensen

Master thesis

60 study points

Department of Molecular Biosciences

Faculty of Mathematics and Natural Biosciences

University of Oslo

Department of Cancer Genetics

Institute for Cancer Research

The Norwegian Radium Hospital

August 2019

UiO : Universitetet i Oslo

Oslo universitetssykehus

Autumn 2019

A cross-platform analysis of lncRNA expression and associations to estrogen receptor status and methylation in breast cancer

Main supervisor: Sunniva Maria Stordal Bjørklund[1]
Co-supervisors: Miriam Ragle Aure[1], Vessela N. Kristensen[1,2], Fahri Saatcioglu[3]

1) Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway;

2) Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, Norway.

3) University of Oslo, Department of Molecular Biosciences

Print: Reprosentralen, Universitetet i Oslo

Abstract

The part of the genome not coding for proteins has moved from being viewed as junk DNA to being investigated as vast, complicated areas that produce diverse non-coding RNAs with different functions. In contrast to protein coding genes, these sequences in the genome lack known domains and motifs. While the majority of lncRNAs are uncategorized, previous efforts have shown that lncRNAs can function as transcriptional activators, -repressors, as scaffolds for chromatin modification complexes, RNA splicing- or degradation regulators or miRNA sequesters or -blockers.

The two most clinically relevant subgroups of breast cancer are estrogen receptor positive (ER+) and negative (ER-) tumors. DNA methylation display distinct patterns in breast cancer, and is central for development of ER+ breast cancer. And several lncRNAs have been reported to be involved in epigenetic mechanisms. In this study we wanted to identify lncRNAs associated with the ER subgroups and with DNA methylation of CpGs. We analyzed RNA-sequencing data from The Cancer Genome Atlas Breast Cancer Cohort (TCGA-BRCA) using the StringTie Ballgown tool suite, and made a catalog of lncRNAs. Differential expression analysis of 1907 lncRNAs in 1045 breast cancer patients from TCGA identified 1386 significant lncRNAs up- or down-regulated in ER+ versus ER- patients. Among the identified lncRNAs were DSCAM-AS1, NEAT1 and MALAT1, previously known to be associated with breast cancer. We also identified new candidates, such as NRAV and NCK1-AS1. Of the 1386 lncRNAs, 743 and 658 were identified and detected in two independent breast cancer cohorts, respectively, based on microarray data. Of these, 533 and 513 significant lncRNAs were found to be differentially expressed in the two clinical groups in the two cohorts, respectively; 354 lncRNAs were identified in all three cohorts. The results were further used in a genome wide correlation (emQTL) analysis, which identified 265 lncRNAs associated with methylation of CpGs, both in *cis* and in *trans*, in TCGA and OSL2. We intersected the lncRNAs most upregulated in ER+ tumors and the lncRNAs with highest number of associations to methylation. We further quantified levels of lncRNAs in the nuclear and cytosolic fractions of MCF7 and identified candidates with higher expression in the nucleus, which could suggest functions connected to chromatin structure. Furthermore, focusing on high confidence lncRNAs not previously associated to breast cancer in the literature, the above criteria identified three candidate lncRNAs; GATA3-AS1, FAM198B-AS1 and DRAIC. A knockdown of the candidates in the ER+ breast cancer cell line MCF7 resulted in a ~50% knockdown. Viability was measured after three days, and there was no significant effect of the knockdown of the three lncRNAs. However, relative expression of GATA3 in GATA3-AS1 knockdown showed a significant reduction.

The analysis of lncRNAs confirms previous observations that many lncRNAs are expressed dependently on ER status. We identified a subset of lncRNAs with strong associations to

methylation, and this group could be explanatory to the epigenetic events defining these clinical subgroups. Future experiments should include methods to assess binding sites of the lncRNA candidates. Experiments should also assess effects after knockdown and over-expression of the candidates on functions such as drug resistance, migration, or other effects important for cancer progression, including the effect on global methylation of CpGs.

Acknowledgements

This thesis was written as a part of a master's degree in molecular biosciences at the University of Oslo. The work presented here was conducted at the Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, from June 2018 to August 2019.

I would like to express my deepest gratitude to my main supervisor, Dr. Sunniva Maria Stordal Bjørklund, for accepting me as her student, for sharing her knowledge of bioinformatics and long non-coding RNA, for her extraordinary ability to never give up on solving complicated problems, and for all the interesting conversations.

I am also very grateful to my co-supervisor, Dr. Miriam Ragle Aure, for her excellent guidance, clear overview and kindness, for her major contribution towards this thesis, and for all the time she has invested in me.

I want to express my gratitude to professor Vessela N. Kristensen for including me in her Cancer Genome Variation Group, for creating such a positive and interesting work environment, and for being a truly inspirational scientist and person.

I also want to thank all the members of Vessela's group for making me feel welcome from the start, and for the including and friendly environment. A special thanks to Daniel Nebdal for his computer skills, to Grethe I. G. Alnæs for always being so helpful in the lab, and to Marie Fongaard for teaching me cell culture work and being so educational and patient.

I want to thank my fellow students, especially Siril, Amanda and Malene at NMBU, for all the fun during our bachelor's, and to Jørgen for his kind help during our master's projects.

I also want to thank my family, especially my mother, Mette, who was diagnosed with cancer and passed away during my first year at NMBU. I am grateful for all her encouragement and support when I decided to begin a new education.

Oslo, August 2019

Katrine Bull Evensen

List of abbreviations

ADH	Atypical ductal hyperplasia
BH	Benjamini Hochberg
bp	Base pair
cDNA	Complementary DNA
Ct	Cycle threshold
CTG	CellTiter-Glo
DCIS	Ductal carcinoma *in situ*
DNMT	DNA methyltransferase
E2	Estradiol
EGA	European Genome-phenome Archive
emQTL	Expression methylation quantitative trait loci
ENCODE	ENCyclopedia Of DNA Elements
eRNA	Enhancer RNA
ER	Estrogen receptor
fc	Fold change
FDR	False discovery rate
FPKM	Fragments Per Kilobase Million
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GEO	Gene Expression Omnibus
GR	Glucocorticoid receptor
GRC	Genome Reference Consortium
GRCh	Genome Reference Consortium human genome
GRE	Glucocorticoid response element
GSEA	Gene set enrichment analysis
HER2	Human epidermal growth factor receptor 2
HOXD	Homebox D cluster
IHC	Immuno-histochemical staining
IQR	Interquartile Range
lincRNA	Long intergenic non-coding RNA
lncRNA	Long non-coding RNA
lORF	long open reading frame
Metabric	Molecular Taxonomy of Breast Cancer International Consortium
NAT	Natural antisense transcripts
ncRNA	Non-coding RNA
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
ORF	Open reading frame
OSL2	Oslo2

PCR	Polymerase chain reaction
PR	Progesterone receptor
PRC2	Polycomb repressive complex 2
QmRLFS	Quantitative model of R-loop forming sequences
RNAi	RNA interference
RNase H	Ribonuclease hybrid
RNA-seq	RNA sequencing
RPKM	Reads per kilobase million
RPM	Reads per million
RT	Reverse transcription / reverse transcriptase
siRNA	Small interfering RNA
SF	Splicing factor
SNIP1	Smad nuclear-interacting protein 1
SNP	Single nucleotide polymorphisms
sORF	Small open reading frame
TCGA	The Cancer Genome Atlas
TCGA-BRCA	The Cancer Genome Atlas Breast Cancer Cohort
TF	Transcription factor
Th2	T-helper 2
UTR	Tntranslated region
UCSC	University of California Santa Cruz
+	Positive
-	Negative

Table of contents

Abstract ..... *I*

Acknowledgements ..... *III*

List of abbreviations ..... *IV*

Table of contents ..... *VI*

1 Introduction *1*

1.1 Cancer **1**
1.1.1 The hallmarks of cancer 1
1.1.2 Oncogenes and tumor suppressor genes. 3

1.2 Breast cancer **3**
1.2.1 Incidence and mortality 3
1.2.2 Female human breast anatomy 4
1.2.3 Breast cancer progression 5
1.2.4 Breast cancer classification 6
1.2.5 Molecular markers in breast cancer 6

1.3 lncRNA **10**
1.3.1 lncRNA classification based on location and transcription start site 11
1.3.2 The diverse functions of lncRNAs 13
1.3.3 Functions of lncRNAs in cancer 14

1.4 Epigenetics **17**
1.4.1 DNA methylation (and histone tail methylation) 18
1.4.2 Chromatin structure and gene expression 20
1.4.3 DNA methylation in cancer 20

2 Aims *22*

3 Materials *23*

3.1 Discovery cohort; TCGA BRCA **23**

3.2 Validation cohorts **23**
3.2.1 Metabric 23
3.2.2 Oslo2 24

3.3 Databases **24**
3.3.1 Reference genome 24
3.3.2 Reference transcriptome 24

4 Methods *26*

4.1. Computational language and software environment R **26**

4.2 RNA-sequencing technology **26**

4.3 RNA-seq bioinformatic pipeline **28**
4.3.1 RNA-seq data. 28
4.3.2 Quantification 28
4.3.3 Normalization. 28

4.4 Workflow for lncRNA transcript assembly and quantification (TCGA) **29**

4.5 Microarray **30**

4.6 DNA methylation by Illumina Infinium Methylation450K beadchip **30**

4.7 Nuclear/cytosolic localization data 31
4.8 Log2 transformation 32
4.9 Differential expression analysis 32
4.10 p-value 32
4.10.1 Multiple testing correction 33
4.11 Genome-wide correlation analysis of lncRNA expression and DNA methylation of CpGs 33
4.12 Hierarchical clustering and heatmaps 34
4.13 Box plots 34
4.14 Pathway enrichment analysis 34
4.15 Quantitative model of R-loop forming sequences 35
4.16 Cell line 35
4.17 RNase H mediated knockdown 36
4.17.1 LNA GapmeR 36
4.17.2. Transfection 37
4.18 Cell viability assay 38
4.19 RNA extraction 38
4.20 First strand cDNA synthesis 39
4.21 PCR 40
4.21.1 Realtime quantitative PCR for validation of knockdown 40
4.21.2 Analysis of RT qPCR results 42
5 Results 43
5.1 Identification of lncRNAs in the TCGA BRCA cohort 43
5.2 Differential expression of lncRNAs in ER+ versus ER- patients in the discovery 43
cohort TCGA BRCA 43
5.3 Validation of results in independent cohorts and matching of lncRNA IDs 44
5.4 Associations between lncRNA expression and methylation status 48
5.5 Functional characterization of candidate lncRNAs in the ER+ cell line MCF7 49
5.6 Knockdown of candidate lncRNAs in MCF7 55
5.6.1 The effect of GATA3-AS1 knockdown on GATA3 transcription 56
5.6.2 The effect of candidate knockdown on cell viability with CellTiter-Glo 57
5.7 Pathway enrichment analysis of highly correlated genes 57
5.8 R-loop formation sequence (RLFS) prediction for candidate lncRNAs 59
6 Discussion 60
6.1 Methodological considerations 60
6.1.1 Expanding annotation databases 60
6.1.2 The challenges of cross-platform analysis 61
6.1.3 Cell lines 62
6.1.4 Cell viability knockdown experiment in MCF7 62
6.1.5 Data driven research 63

6.2 Biological considerations 64
6.2.1 Clinical material 64
6.2.2 lncRNAs associated with ER status 65
6.2.3 Associations to methylation of CpGs 66
6.2.4 Possible functions of GATA3-AS1 67
6.2.5 FAM198B-AS1 is previously uncharacterized 68
6.2.6 Possible functions of DRAIC 68
7 Conclusion and future perspectives 70
References 71
Supplementary data 79

1 Introduction

1.1 Cancer

Cancer is a term used to describe many diseases characterized by uncontrolled cell growth. Robert A. Weinberg's two definitions of cancer are 1: A clinical condition that is manifested by the presence of one or another type of neoplastic growth; 2: A malignant tumor [1]. Cancer can arise in all organs, but the distribution of the different types is related to sex and age, and the risk of cancer increases with age. According to Anand et al. [2] 5-10% of all cancer cases are caused by genetic defects, and the remaining 90-95% can be attributed to environment and lifestyle. There are more than 200 different types of cancer. The networks of causation in cancers are very complex, and the diseases have possibly large individual differences, making it a challenging area of research.

1.1.1 The hallmarks of cancer

In order to systematize common characteristics of the many diseases that cancer represents, Hanahan and Weinberg defined in 2000 six hallmarks of cancer [3], and the list was further updated to ten hallmarks in 2011 [4]. The hallmarks describe capabilities that normal cells acquire during their way to become cancerous (Figure 1).

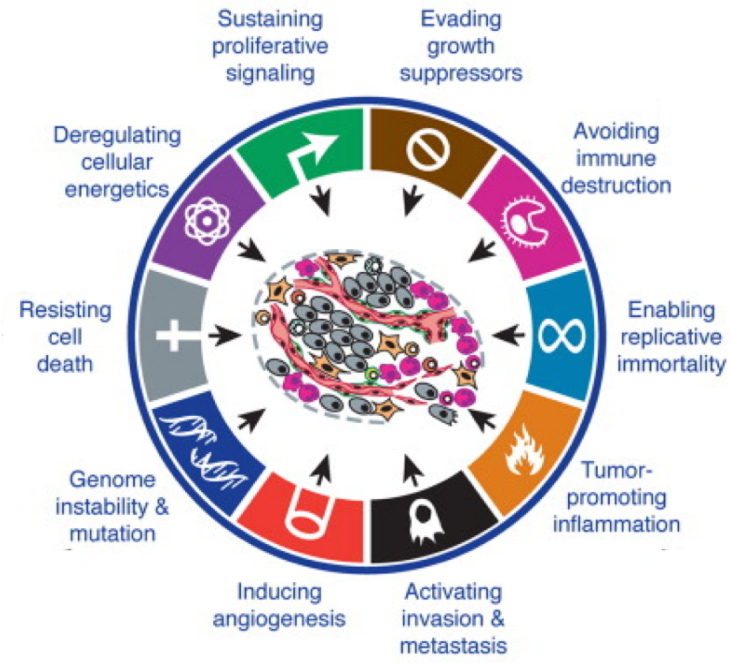


Figure 1. The ten hallmarks of cancer. The updated version of the hallmarks of cancer as described by Hanahan and Weinberg in 2011. Adapted from [4].

1) Evading growth suppressors
In normal cells, proliferation is negatively regulated by growth suppressors such as p53, RB1 and TGF-β. In cancerous cells some of these genes can be inactivated, and cancer cells escape the negative regulation of the cell cycle.

2) Enabling replicative immortality
The length of the telomeres (the ends of the chromosomes) normally inhibits the number of times a cell can divide. Tumor cells can exploit the enzyme telomerase, which extends the telomeres, and make them capable of dividing infinitely, without entering senescence, by e.g. overexpression.

3) Avoiding immune destruction
To avoid recognition by immune cells, cancer cells express or suppress certain molecules normally recognized by the immune system at their surface.

4) Tumor promoting inflammation
Chronic inflammation can cause pro-neoplastic mutations and resistance to apoptosis. In addition tumor cells can overexpress inflammation-promoting molecules. They can also attract immune cells such as macrophages which express cytokines and growth factors that nourish the tumor.

5) Activating invasion and metastasis
Metastasis is the most common cause of cancer mortality. For metastasis to happen, the cells must gain epithelial-mesenchymal transition (EMT) characteristics to detach from the primary tumor, get increased motility, and then develop mesenchymal-epithelial transition (MET) characteristics to attach and start growing in a new tissue.

6) Inducing angiogenesis
Formation of new blood vessels induced by molecules such as VEGF and bFGF is necessary to provide the elevated need for growth nutrients and flow of oxygen to the tumor cells.

7) Genome instability and mutations
Genome instability such as a high frequency of mutations and larger DNA alterations are caused by external DNA damage or errors in DNA repair genes. Accumulation of mutations in oncogenes or tumor suppressor genes is thought to often be the starting point for cancers. Small scale mutations can include insertions, deletions, or substitutions. Larger scale mutations include gene amplifications, deletions of larger chromosomal sequences, or chromosomal rearrangements. The latter include inversions or translocations, leading to fusion-proteins, translocations of promoters leading to altered protein concentration or the duplication or loss of a whole chromosome leading to aneuploidy. Epigenetic alterations such as promoter hypo- or hypermethylation can also be seen as part of this hallmark.

8) Resisting cell death
In normal cells, there is a fine-tuned balance between pro-apoptotic and anti-apoptotic proteins. In cancer cells there can for example be activation of pro-survival genes that inhibit activation of apoptotic proteins.

9) Deregulating cellular energetics
In tumors, both when oxygen is and is not present, the glycolysis converts most of the pyruvate into lactate instead of delivering it to the Krebs cycle as in healthy cells when oxygen is present. This process yields less ATP, which makes the cancer cell import more glucose. The cell gains more intermediate products from the glycolysis which it uses for building blocks in proliferation, instead of for energy.

10) Sustaining proliferative signaling
Cancer cells are able to produce their own growth signals and to overexpress receptors for these signals, which keep the cells in a continually prolific state.

1.1.2 Oncogenes and tumor suppressor genes

Two of the main types of genes involved in cancer development and progression are oncogenes and tumor suppressor genes. An oncogene can be defined as a gene that when it is altered, often by mutation or increased concentration of its protein product, has the potential to cause cancer [5]. These genes typically promote hallmark characteristics. An oncogene that is not activated is called a proto-oncogene. Weinberg's definition of a tumor suppressor gene is "a gene whose partial or complete inactivation, occurring in either the germ line or the genome of a somatic cell, leads to an increased likelihood of cancer development" [5]. The most commonly mutated oncogenes in breast cancer are *PIK3CA*, *GATA3* and *MAP3K1 [6]*. When it comes to concentration, the gene for estrogen receptor (ER) 1, *ESR1* is overexpressed in 65-70% of breast cancers, known as the ER+ subgroup, and *ERBB2* (coding for the HER2 protein) is amplified in 20% of breast cancers, the HER2+ subgroup. The most commonly mutated tumor suppressor genes in breast cancer are *TP53, CDH1, PTEN* and *BRCA1/2* [7].

1.2 Breast cancer

1.2.1 Incidence and mortality

In Norway, the four most common cancers across both sexes are cancers of the breast, prostate, lung and colon, accounting for 45 % of the total number of incidences. Of these, breast cancer has the third highest rate of incidence, but 1 out of 11 women will develop

breast cancer before the age of 75, making it the most common cancer for women between 25 and 69 years of age [8]. Men can also develop breast cancer, but it occurs in less than 1% of the incidences [9]. In 2017 there were 3589 incidences in Norway, and by the end of 2017 47568 persons were living with the disease. Between both sexes, breast cancer had the fourth highest mortality rate in 2016, accounting for 6% of the total cancer related deaths. Between 2013 and 2017 breast cancer had a 90,4% five-year relative survival rate. In 2016 623 women and 6 men died from breast cancer [8].

1.2.2 Female human breast anatomy

The female breast is made up of glandular tissue surrounded by adipose tissue, connective tissue, blood- and lymph vessels. The glandular tissue is composed of 12-20 lobes, and each lobe is made up of lobules, containing alveoli, the glands that produce milk (Figure 2). Each lobe drains into ducts that lead the milk to the nipple. The lobules have an outer layer of myoepithelial (smooth muscle) cells that will contract and assist in ejection of milk during lactation. The ducts and lobes are supported by connective tissue coming together in ligaments that anchor the breast to the chest wall. The surrounding adipose tissue is also supporting and creating a framework for the other structures [10].

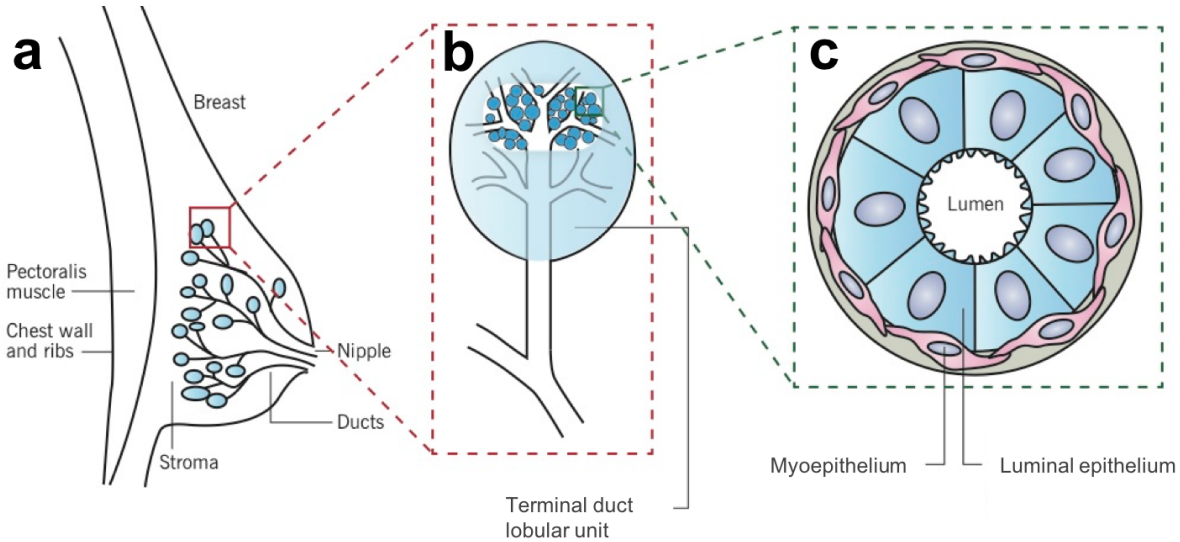


Figure 2. Overview of the glandular tissue in a healthy female breast. **a.** Lobes drain into ducts that lead milk to the nipple. **b.** Each lobe is composed of lobules containing milk-producing alveoli. **c.** The ducts and lobules have two layers of cells, an inner luminal epithelium and an outer myoepithelium that contract to eject the milk. Adapted from [11].

1.2.3 Breast cancer progression

A normal duct has two layers of cells, an inner luminal epithelium and an outer myoepithelium, where an abnormal cell growth can develop in stages going from a normal duct to invasive ductal carcinoma (Figure 3). When there are more than two layers of cells it is called ductal hyperplasia. Hyperplasia denotes an excessive number of otherwise normally appearing cells. This may progress to atypical ductal hyperplasia (ADH) characterized by a histomorphological abnormality in the arrangement of the cells. This condition is not yet defined as cancer, but is associated with an increased risk (4 to 5 times) [5]. If the cells progress to the next stage, ductal carcinoma *in situ* (DCIS), the characteristics are the same as in ADH, but with a larger amount of cells. *In situ* means "in place", and this is a precancerous state where the cancer cells still are non-invasive and confined to the duct [12]. A further progression to the condition defined as cancer is when the cancer cells invade and break through the basement membrane and into the surrounding tissue [13].

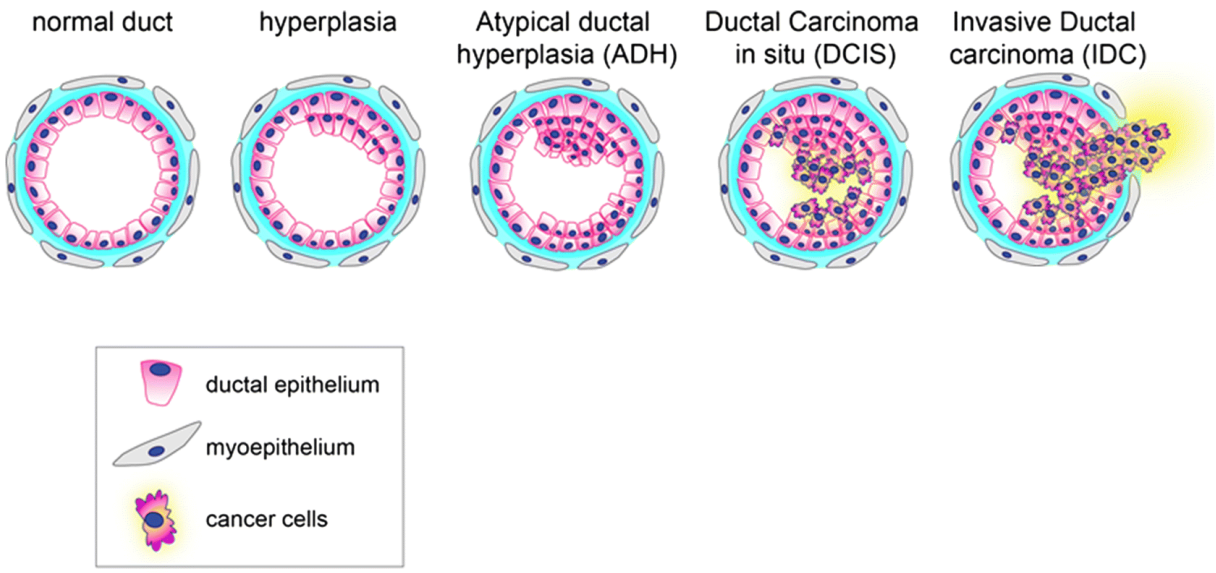


Figure 3. Breast cancer progression. The stages going from a normal duct to invasive ductal carcinoma. Adapted from [14].

1.2.4 Breast cancer classification

Histopathological classification

When using histopathological classification, two main types of breast cancer are visually observed; ductal and lobular carcinomas, that together make up >95% of all breast carcinomas. The term carcinoma is used for cancers originating in epithelial cells. In addition to the main types lobular and ductal carcinomas other less frequent types of invasive carcinoma include tubular, mucinous, adenoid cystic, micropapillary, secretory and apocrine breast cancer [15].

Grade

In addition to the histopathological classification, the tumors are evaluated for the histologic grading which reflects how closely the tumor cells resemble normal breast epithelial cells; the degree of differentiation. The cancer cells may be de-differentiated, disorganized, and do not line up in an orderly way. The cell nuclei change, and cell division is no longer controlled. The grades of theses changes are used to classify tumors: Differentiated (low-grade), moderately differentiated (intermediate-grade), and poorly differentiated (high-grade). More differentiated cells resemble more normal tissue, and less differentiated cells therefore gives a worse prognosis [16].

Stage

Further classification is done by staging the tumor where information on tumor (T) size, lymph node (N) involvement and presence of metastasis (M) is combined (TNM) [17]. Stage 0 is a pre-cancerous condition. Stages 1 to 3 are tumors within the breast or with cancer cells found in lymph nodes. Stage 4 is metastatic cancer, where the cancers have spread and tumors have grown in distant tissue beyond the axillary lymph nodes. There is no cure for metastatic breast cancer.

1.2.5 Molecular markers in breast cancer

Estrogen receptor

The two most clinically relevant subgroups of breast cancer are estrogen receptor positive (ER+) and estrogen receptor negative (ER-) tumors. The classification is done by immuno-histochemical staining (IHC). The cut-off value for ER+ is 1% staining in the nucleus, while cytoplasmic staining is regarded as unspecified [18]. There are two main forms of ER: ERα and ERβ, encoded by two separate genes, *ESR1* and *ESR2*, on different chromosomal locations [19]. In women, estrogen is part of the signaling pathways that makes cells in the reproductive tissue proliferate. Estrogen's role in breast cancer is complex, but it is evident

that estrogen periodically induces cell proliferation in a way that enables progression of mammary epithelial cells into cells characteristic for breast cancer [5]. The repeated fluctuations in estrogen level inducing proliferation and apoptosis of breast tissue during each menstrual cycle thereby increase the risk of breast cancer. A compelling illustration of this is the doubled risk of breast cancer of women who continue to menstruate to age 55 or beyond, compared to those who reach menopause before age 45 [5]. Estrogens, such as estradiol (E2) diffuse across the cell membrane and bind to both nuclear and extranuclear ERα in breast cancer cells (Figure 4). One of the most known ER functions is as a ligand-dependent transcription factor that activates genes involved in cell proliferation and invasion. The E2-ERα complex induces dimerization and phosphorylation of ERα and binds to estrogen receptor elements (ERE) of numerous target enhancers and promoters, such as *CCND1* (Cyclin D1) which is important for regulation of the cell cycle [20]. In addition growth factors can activate ERα, or induce signaling pathways such as MAPK or PI3K, involving ERα downstream [7] and lead to transcription of genes promoting cell proliferation, survival and metastasis [21]. For some transcription factors, ERα acts as co-regulator. ERα can interact with transmembrane receptors or cytoplasmic proteins complexes, as well as participate in kinase-dependent signaling for the activation of some transcription factors, such as FOS, SP1 and JUNB [22]. Some isoforms are also thought to associate with the membrane and trigger signaling cascades of several kinases [21].

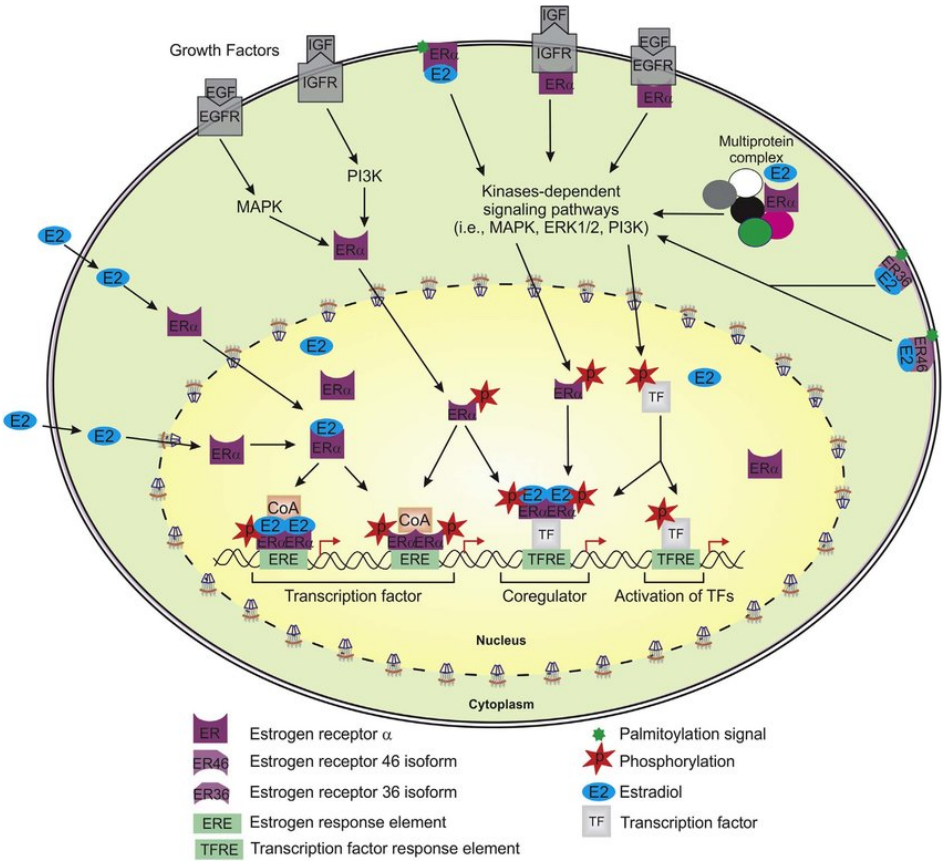


Figure 4. Signaling pathways of ERα in BC cells. E2 binds to ERα nuclear and extranuclear. E2 and ERα are involved in numerous pathways activating transcription of genes promoting proliferation and cell growth. Reprinted from [21].

Endocrine therapy is a form of therapy that aims to inhibit ER signaling, and Tamoxifen is a drug that binds to the ER [23]. ER+ patients experience a better prognosis and outcome, compared to ER-. Still, the mortality rate is higher among ER+ patients, partly due to the higher number of patients in this group, and partly because resistance to endocrine therapy can be inherent or acquired [24]. There is a pressing need to understand more of the mechanisms through which ER-driven breast cancers become resistant, and this is subject of intense scientific investigation. In post menopausal women with ER+ tumors, aromatase inhibitors are used [23]. The presence of the progesterone receptor (PR) is also used in some clinical classifications, but the prognostic value of PR is uncertain [25].

Human epidermal growth factor receptor 2 (HER2)

Amplification of the *ERBB2* gene encoding the HER2 protein occurs in ~20% of breast cancers. HER2-enriched breast cancers can have up to 25–50 copies of the *ERBB2* gene, giving up to 40–100-fold increase in HER2 protein resulting in 2 million receptors expressed at the tumor cell surface [26]. The protein signals to the cell to continue to grow and divide and protects it from apoptosis [5, 27]. This upregulation is correlated with poor prognosis [28].

Ki-67

The Ki-67 protein is a cell marker associated with cell proliferation. It can be used as a marker because it is present in all the active phases of the cell cycle, but not in resting cells. High levels of Ki-67 is associated with poor prognosis. Patients with high Ki-67 have a good chance of response to chemotherapy [29].

BRCA1* and *BRCA2

Mutations in the *BRCA1* and *BRCA2* genes are strongly correlated with breast cancer. The hereditary cases make up ~10% of breast cancers, and more than half of the hereditary cases carry germ line mutations in *BRCA1* or *BRCA2*. Furthermore, more than half of the sporadic breast carcinomas have inactivated *BRCA1* genes caused by promoter hypermethylation [5].

TP53

The top most mutated gene in breast cancer is *TP53*, being mutated in 35% of all breast cancers [7]. *TP53* is an important tumor suppressor gene, rapidly increasing in expression levels in response to varying cell-physiological stress, such as oncogene signaling, hypoxia or double stranded DNA breakage. The p53 protein induces cell cycle arrest, allowing time for DNA repair proteins to work, or as an alternative pathway, p53 induces apoptosis. This

makes it a crucial protein for suppressing progression of cancer cell characteristics [5]. Accordingly, *TP53* status has high prognostic value, where mutations in this gene give worse disease-free and overall survival [30].

Gene expression and the five subtypes of breast cancer

Breast cancer was separated into five molecular subtypes with different prognosis based on gene expression in 2000 [31]. In 2009 the PAM50 classification method was established to separate breast cancer into these five molecular subtypes based on the expression of 50 genes [32]. These subtypes correspond closely to other clinical markers such as hormone receptor status and amplification of the HER2 locus. The five subtypes are luminal A, luminal B, Her2, basal, and normal-like.

Luminal A

Luminal A breast cancer is the most common subtype (~50%). This cancer is HER2 negative, and hormone-receptor positive, meaning that it is both ER and/or PR positive, and it has low Ki-67 [33, 34]. Luminal A tumors have the best prognosis of all the subtypes, both in overall survival rate, and relapse free survival, although the risk of delayed metastasis persists [33, 35]. Because this cancer is hormone receptor positive, it can be treated with endocrine treatment alone, endocrine therapy plus other targeted therapies, or chemotherapy [36].

Luminal B

Luminal B breast cancer is hormone-receptor positive. This cancer has upregulation of Ki-67. The cancer can be both HER2 positive or HER2 negative. Because of faster tumor growth, luminal B patients have a worse prognosis than luminal A patients [35]. However, they have better prognosis than patients with triple negative and HER2-enriched breast cancer. Because of the high level of Ki-67 these patients receive chemotherapy [36].

Basal-like

Basal like breast cancer essentially consists of tumors that are triple negative based on immunohistochemistry. Triple-negative means that the tumor is both hormone-receptor negative (ER- and PR-) as well as HER2 negative. The basal-like class is associated with *BRCA1* gene mutations. This cancer is more often found in African-American women and younger women [37]. Basal-like cancer has the worst prognosis of the five subtypes [35]. A study published in 2018 found that the 5-year overall and disease-free survival were ~20% less for triple negative patients compared to non-triple-negative [38]. Treatments for this subclass can be both chemotherapy and potentially also immune therapy [36, 39].

HER2-enriched

HER2-enriched breast cancers are mainly positive for HER2 amplification. It is mostly hormone-receptor negative. Overexpression of HER2 is a well-known factor contributing to a poor survival rate. These tumors are faster growing than the luminal cancers. HER2-enriched was the subtype with the poorest prognosis next after triple-negative at the time of Sørlie and Perou's seminal articles in the beginning of the 2000's [31, 35]. However, because of new treatment, HER2-enriched breast cancer now has a better prognosis [40]. Patients can receive chemotherapy and be treated with HER2 protein-targeted therapies. Examples are Trastuzumab, Pertuzumab, Lapatinib, and T-DM1 or Ado-trastuzumab Emtansine [36].

Normal-like

Normal-like breast cancers show similarities to luminal A tumors. They are mainly hormone-receptor positive, HER2 negative, and have down-regulated Ki-67 levels, and otherwise expression profiles similar to normal breast tissue [41]. This subtype has intermediate prognosis; worse than luminal A, but better than luminal B [35].

1.3 lncRNA

Recent advances in sequencing technologies have enabled more in-depth analyses of the genome and transcriptome. The part of the genome not coding for proteins has moved from being viewed as junk DNA to being investigated as vast, complicated areas that produce diverse non-coding RNA (ncRNA) with different functions [42, 43]. Studies have revealed that ~80% of the human genome is transcribed. But only 1% of the human genome consists of protein-coding regions, meaning that most of the transcriptome is non-coding [44]. In contrast to protein coding genes, these sequences in the genome lack known domains and motifs. Among the ncRNAs shorter than 200 bp there are for example microRNA (miRNA), small interfering RNA (siRNA) and transfer RNA (tRNA). Long non-coding RNAs (lncRNA) are defined by a minimum base pair length of 200, and the lack of a long open reading frame (lORF), defined by more than 100 bp. However, lncRNAs have been found to contain short open reading frame (sORF), less than 100 codons [45, 46]. These have the possibility to produce small peptides, which have until recently been overlooked because of the cutoff of 100 codons as a minimum for coding capacity. This discovery of bifunctional genes containing both protein-coding and coding-independent transcripts with distinct functions has made the previous distinction even more unclear [47]. Even though next generation sequencing (NGS) has led to the discovery of novel lncRNAs, of the ~55 000 lncRNAs annotated to date (Lncipedia5 [43]), still only ~1500 have been functionally characterized. The need for more knowledge about the functions of lncRNAs has motivated this thesis, especially because of the mounting evidence of their roles in transcriptional regulation, epigenetic regulation and disease, including breast cancer [48, 49].

1.3.1 lncRNA classification based on location and transcription start site

There is large variation in how lncRNAs can be located in the genome and in relation to coding genes. These locations, different ways of transcription and the placement of the promoters have been used to classify lncRNA. Different examples of this is illustrated in Figure 5 [47, 50, 51]. lncRNAs can be both completely intronic, partly overlap with exons of coding genes, and they can be transcribed both from the sense and the antisense strands (Figure 5 a). lncRNAs can completely reside within a coding region of the gene (CDS), and they can be located in the untranslated regions (UTRs) in both the 3' and the 5' ends (Figure 5 b). lncRNAs can either overlap with coding genes, or when the whole lncRNA resides outside coding genes they are called long intergenic non coding RNA, or lincRNAs (Figure 5 c and d).

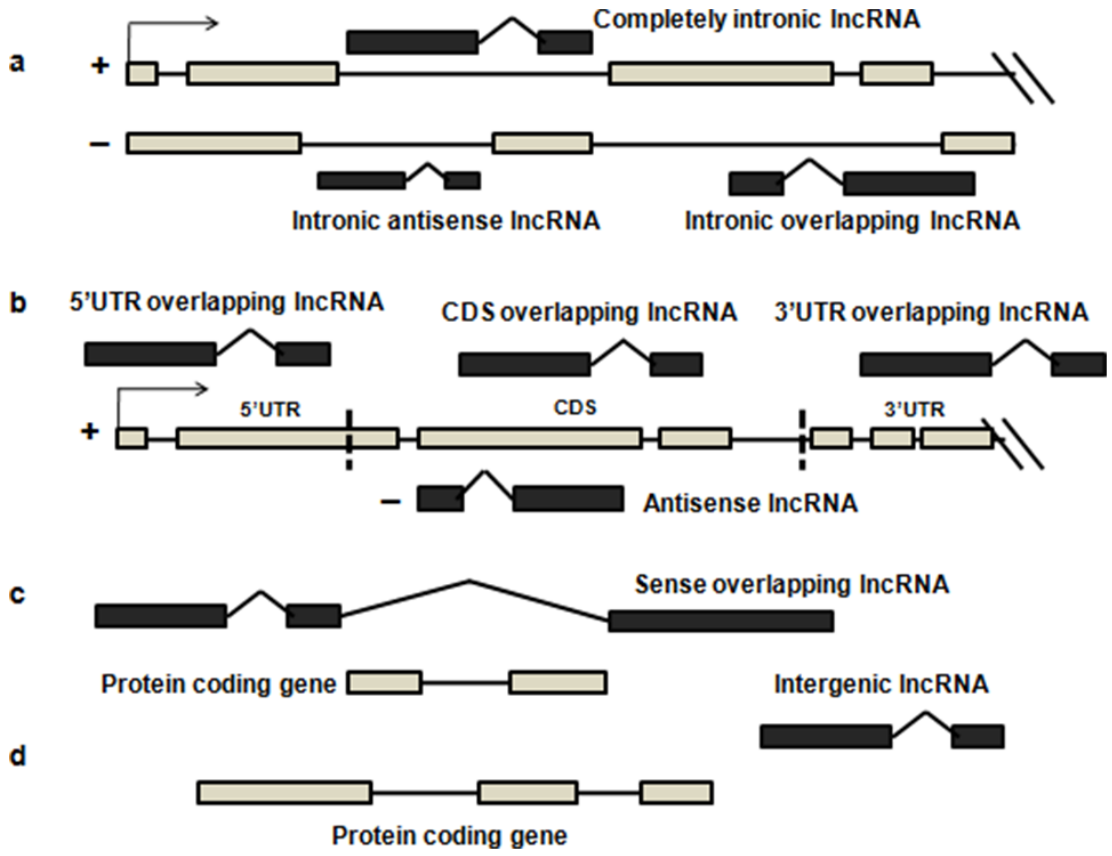


Figure 5. The large variety of how lncRNAs can be located in the genome. See text for details. Reprinted from [52].

1.3.1.1 Natural antisense transcripts

The first recognized group of ncRNA were natural antisense transcripts (NATs) [53]. These are lncRNAs transcribed from the antisense thread close to or overlapping with a coding gene. NATs can be transcribed in different ways relating to coding genes as illustrated in Figur 6. Divergent and convergent lncRNAs are two commonly used terms for bidirectionally transcribed sense/antisense pairs.

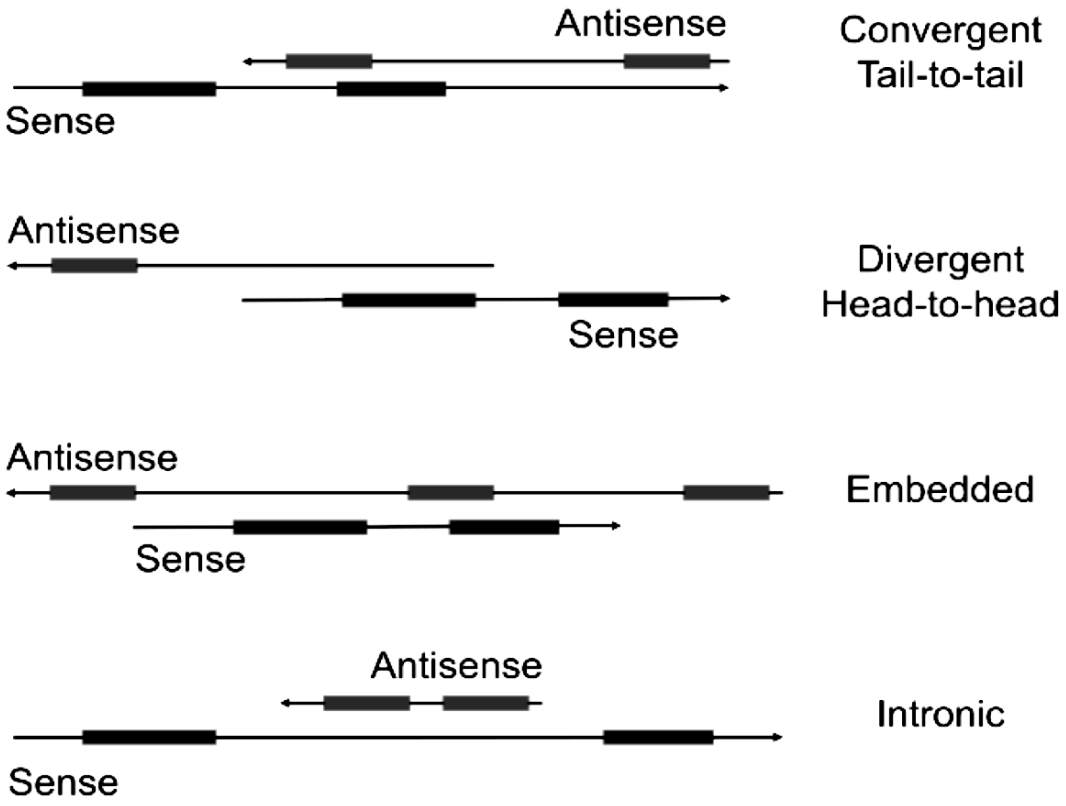


Figure 6. Commonly used terms for sense and antisense pairs. Convergent: The transcription start sites are on each side, and the sense/antisense pair overlap in between, also called "tail-to-tail". Divergent: The transcription start site is in the middle of the sense/antisense pair, also called "head-to-head". Other used terms are "embedded" and "intronic". Adapted from [53].

The exact different molecular mechanisms of NATs are poorly understood, but they have been reported to regulate their corresponding sense transcript either positively or negatively. When convergently expressed, a NAT can for example function as a gene repressor, causing RNA polymerase (Pol) collision and stop the gene transcription. Also these pairs are thought to be able to form sense–antisense hybrids that can trigger RNA interference (RNAi) [53]. In addition to these examples of sense/antisense pairs, the transcription of lncRNAs can be initiated at enhancers further away, and they can be both *cis*- and *trans*-acting.

1.3.2 The diverse functions of lncRNAs

Compared to the more well-studied functions of miRNAs, lncRNA are less understood, but a general view emerging is that they are fundamental regulators in transcription and post-transcriptional processes [50]. While the majority of lncRNAs are uncategorized, previous efforts have shown that lncRNAs can function as transcriptional activators, -repressors, as scaffolds for chromatin modification complexes, RNA splicing- or degradation regulators. They may also interact with miRNA, for example by acting as miRNA sequesters- or blockers [54] (Figure 7).

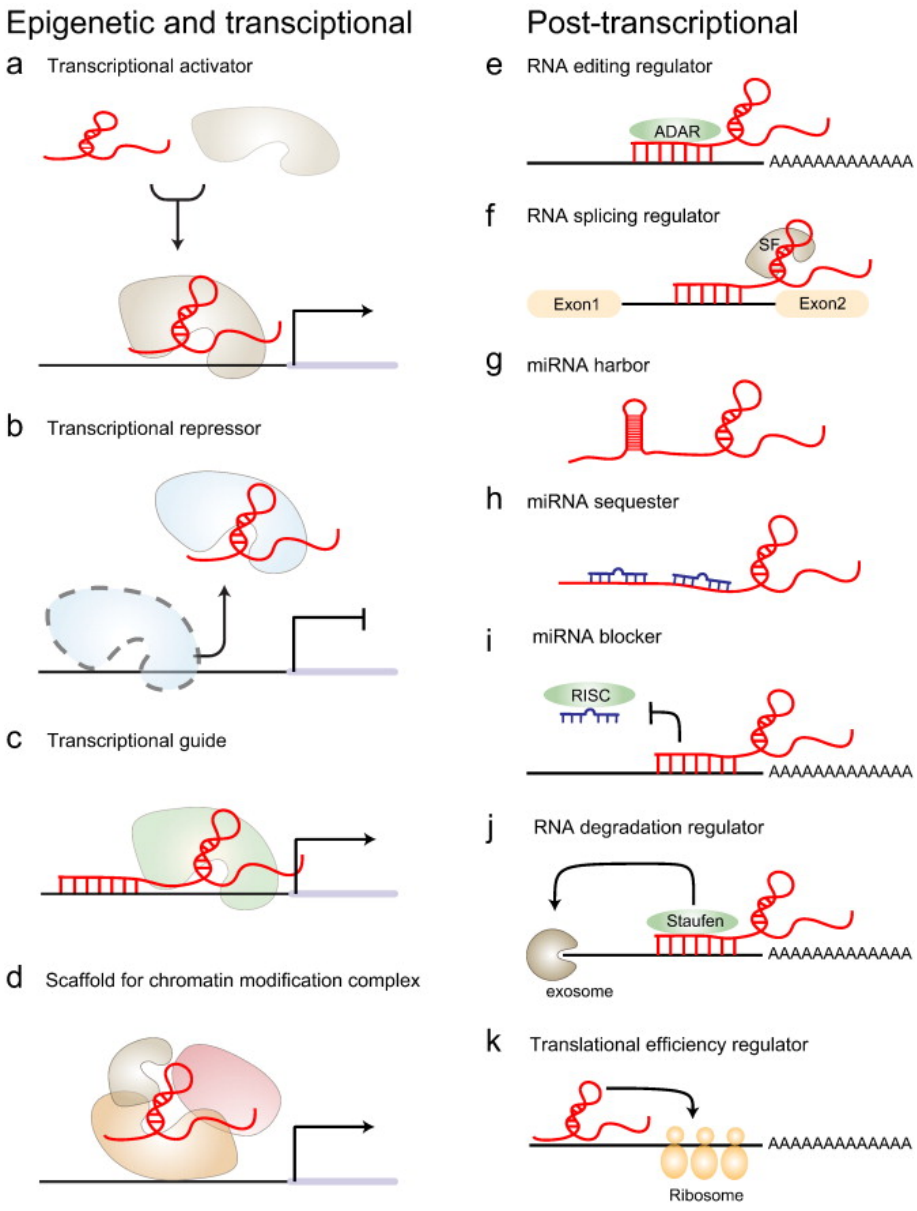


Figure 7. Epigenetic, transcriptional and post- transcriptional functions of lncRNA. **a.** Transcriptional activator. **b.** Transcriptional repressor. **c.** Transcriptional guide. **d.** Scaffold for chromatin modification complex. **e.** RNA editing regulator. **f.** RNA splicing regulator. **g.** mRNA harbor. **h.** miRNA sequester. **i.** miRNA blocker. **j.** RNA degradation regulator. **k.** Translational efficiency regulator.

Most lncRNAs (81%) have poorly conserved DNA sequences and are primate-specific. But 3% of lncRNAs are ultra-conserved, appear in a range of organisms, and may have originated more than 300 million years ago [54]. Many lncRNAs are associated with human diseases, and in particular, cancers. For example is XIST dysregulated in various cancers, HULC is upregulated in hepatocellular carcinoma, GAPLINC is associated with poor prognosis in gastric cancer, MALAT1 is associated with poor prognosis and metastasis in liver-, lung- and colorectal cancers, HOTAIR is associated with metastasis in breast-, colorectal-, liver-, pancreatic- and gastric cancers, and ANRIL is upregulated in prostate cancer [50].

1.3.3 Functions of lncRNAs in cancer

The antisense lncRNA HOX transcript HOTAIR is an example of a *trans*-acting lncRNA working both as a scaffold and a guide for proteins involved in epigenetic regulation as illustrated in (Figure 8 A). HOTAIR is thought to form multiple double stem-loop structures that bind two different histone-modification complexes. The 5′ end binds polycomb repressive complex 2 (PRC2), and HOTAIR guides it to homebox D cluster (HOXD) loci where PRC2 facilitates histone methylation that silences ~800 target genes (Figure 8 B) [55]. The 3′ end of HOTAIR binds lysine-specific demethylase 1 (LSD1)/CoREST/REST complex which executes histone demethylation activating gene expression [56]. HOTAIR is upregulated in certain subgroups of breast cancer and is thought to promote metastasis [55].

MALAT1 is reported to be involved in regulation of alternative splicing and transcriptional repression. MALAT1 can form a transcription repressive complex with HuR (Figure 8C) which binds to the promoter of CD133, a marker for cancer stem cells and inducer of EMT [57]. The lncRNA assembles serine/arginine splicing factors (SF) mainly in nuclear speckles as illustrated in Figure 8 D. When MALAT1 was depleted in a knockout mouse model, it resulted in increased SR protein levels, less phosphorylation, an abnormal distribution of SR proteins, and interference with normal mRNA processing [58]. MALAT1 is upregulated in several cancer types, including breast cancer, where there have been reported conflicting results regarding its precise role [59].

BCAR4 is another epigenetic regulator. One of its functions is reported to be binding of Smad nuclear-interacting protein 1 (SNIP1), an inhibitor for p300, a histone acetyltransferase (HAT) (Figure 8 E). SNIP1 binding to BCAR4 releases the inhibition of p300, leading to histone acetylation inducing gene transcription associated with migration [60]. In breast cancer BCAR4 has been reported to be related to the ERBB2/3 pathway and tamoxifen resistance [61].

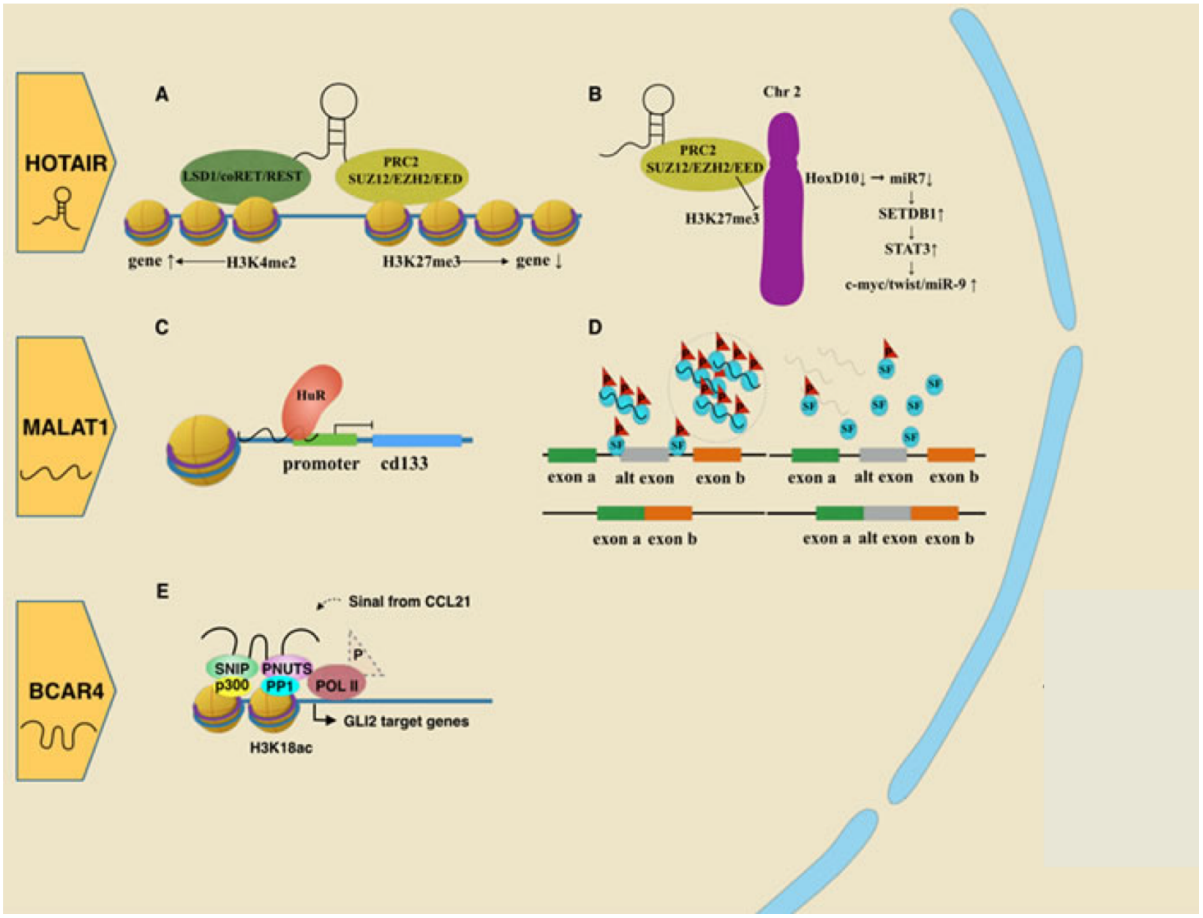


Figure 8. Examples of mechanisms of known oncogenic lncRNAs in breast cancer. **A.** HOTAIR binds PRC2 facilitating histone methylation, and LSD1/CoREST/REST complex which executes histone demethylation. **B.** HOTAIR guides PRC2 to the HOXD locus. **C.** MALAT1 forms a transcription repressive complex with HuR and binds to the promotor of CD133. **D.** MALAT1 is involved in regulation of alternative splicing and transcriptional repression, assembling SFs in nuclear speckles. **E.** SNIP1's binding to BCAR4 liberates the activity of p300, leading to histone acetylation inducing transcription of target genes. Adapted from [62].

GAS5 is an example of a decoy lncRNA, a miRNA sponge and degradation regulator. It is thought to be a tumor suppressor downregulated in multiple cancers, including breast cancer [63]. GAS5 has a structure resembling glucocorticoid response element (GRE), a DNA sequence that binds the glucocorticoid receptor (GR). As a result, GAS5 competes with GREs for binding to GR as illustrated in Figure 9 A. This leads to suppression of the glucocorticoid-induced genes, including one gene encoding an inhibitor of apoptosis [64]. GAS5 also works as a molecular sponge for miR21, inhibiting its mediation of mRNA degradation of several tumor suppressors (Figure 9 B) [65].

One of the most known lncRNAs is XIST which is essential for X chromosome inactivation (XCI), the mechanism for dosage compensation in females. In all differentiated somatic cells, XIST is transcribed from the X inactivation center (XIC) on one of two X chromosomes, from where it spreads and coats the chromosome (Figure 9 D) [66]. Studies have suggested that

XIST recruits chromatin remodeling complexes, including PRC2, responsible for trimethylation of lysine 27 on histone H3 (H3K27me3), a mark for heterochromatin, silencing its nearly 1000 genes [67]. The silent state is stably and irreversibly inherited to daughter cells. Loss of X inactivation and downregulation of XIST have been reported in breast cancer [68]. Together with co-repressors SHARP/SPEN and SMRT, XIST is thought to function as a decoy for histone deacetylase 3 (HDAC3), inhibiting deacetylation of the promoter of PH domain and leucine-rich repeat phosphatase 1 (PHLPP1), keeping transcription levels of PHLPP1 high enough for it to dephosphorylate pAKT into AKT, resulting in limited cell viability (Figure 9 C) [69].

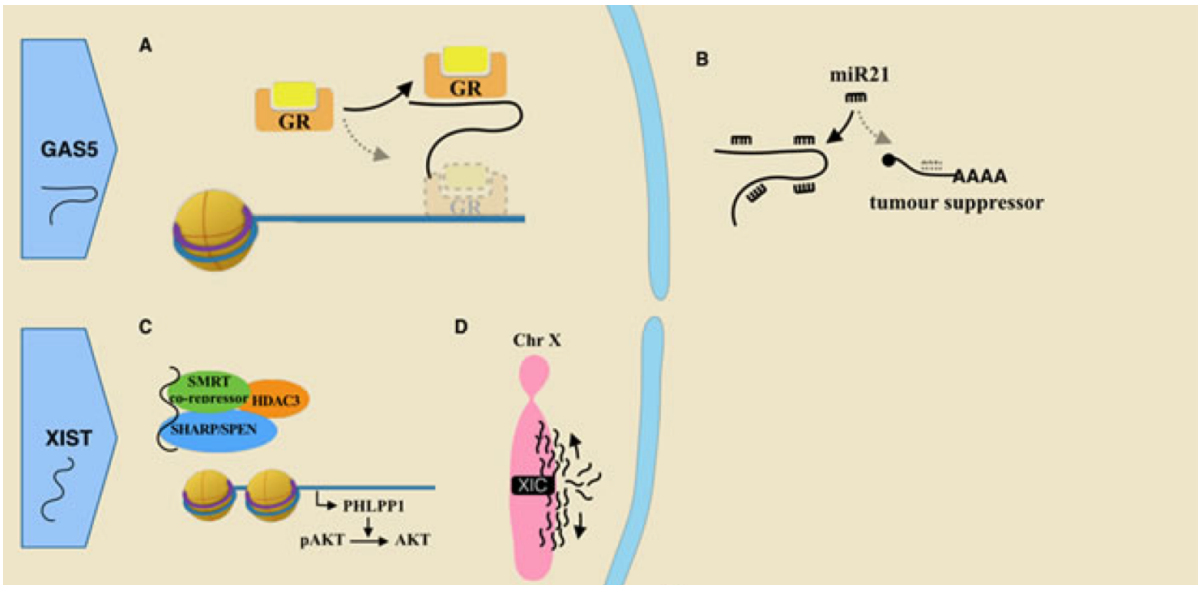


Figure 9. Examples of mechanisms of tumor suppressive lncRNAs GAS5 and XIST, known tumor suppressive lncRNAs in breast cancer. **A.** GAS5 resembles GRE, and competes with binding to GR. **B.** GAS5 works as a molecular sponge for miR21, inhibiting its mediation of mRNA degradation of tumor suppressors. **C.** Together with co-repressors SHARP/SPEN and SMRT, XIST functions as a decoy for histone deacetylase 3 (HDAC3), inhibiting deacetylation of the promoter of PHLPP1, ensuring enough transcription and turning active pAKT into AKT. **D.** XIST is transcribed from XIC, from where it spreads and coats the chromosome. Adapted from [62].

Another example of a lncRNA function is chromatin loop formation. HOTTIP is an enhancer RNA (eRNA), a transcription activating lncRNA transcribed at an enhancer region, enabling the distantly located enhancer to connect with its transcription factor through formation of a chromatin loop as illustrated in Figure 10 [47].

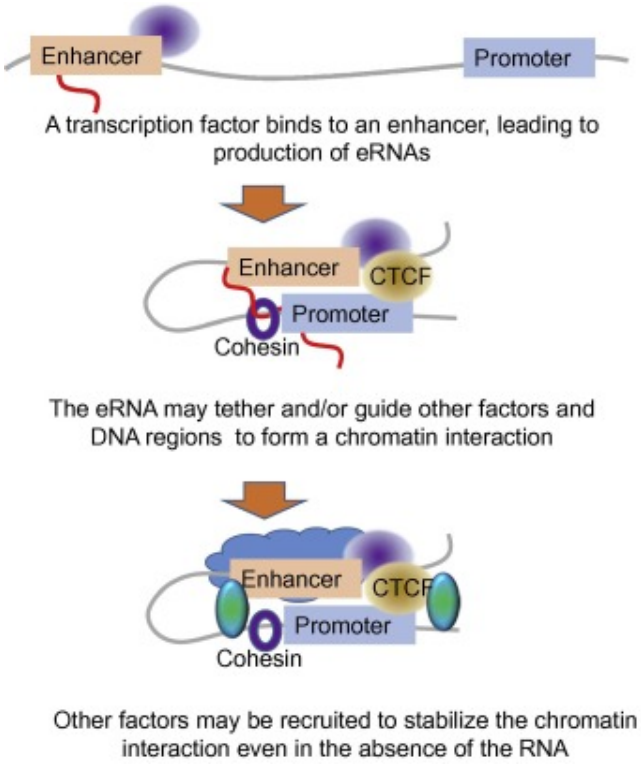


Figure 10. Illustration of an enhancer RNA (eRNA) involved in chromatin loop formation. Adapted from [47].

1.4 Epigenetics

Conrad Hal Waddington introduced the term epigenetics in 1942 in his publication "The epigenotype" [70]. Epi, from Greek means "over", and the term was meant to include the interactions between the environment and the genes leading to the development of phenotype. The concept has been developing, and todays definitions are more molecular [71]. Russo et al.'s definition from 1996 [72] "Mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" is close to a consensus [71]. Commonly the term is used in relation to nucleosome positioning and histone modifications that leads to chromatin changes, and DNA methylation, the two main components of epigenetic coding (Figure 11). Together they affect the accessibility of the DNA molecule for RNA polymerase, and thus the regulation of gene transcription and cell phenotype [73].

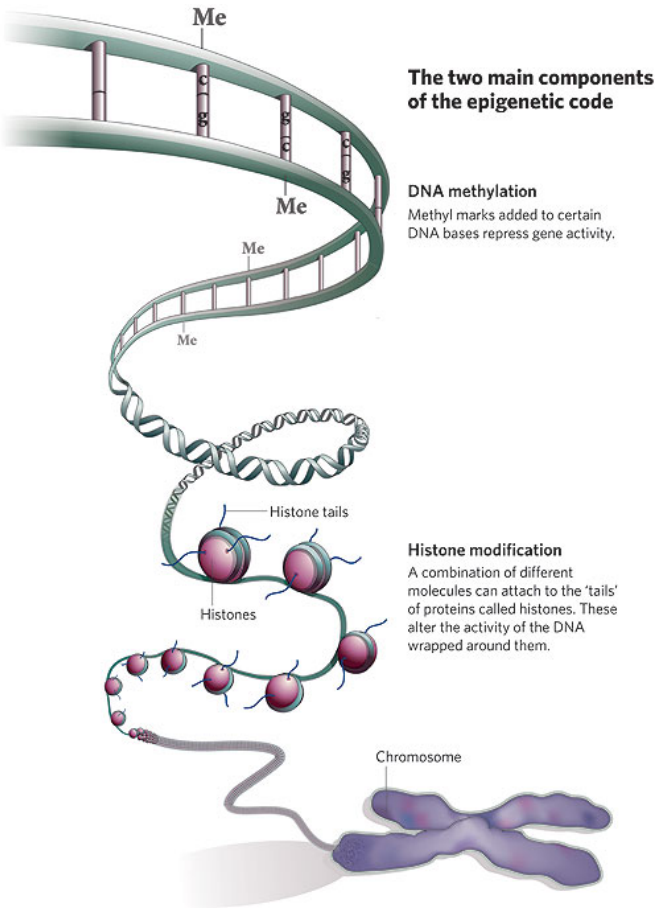


Figure 11. The main components of the epigenetic code: Methylation of certain DNA bases, which repress transcription, and histone modification, the addition of different molecular marks to histone tails, signals for changes in the density of chromatin structure, which leads to more or less access to the DNA for the RNA polymerase. Reprinted from [73].

1.4.1 DNA methylation (and histone tail methylation)

DNA methylation is the most studied form of epigenetics. In mammalian cells, methylation occurs mainly at a cytosine followed by a guanine, a so called CpG. CpG methylation is a pre-transcriptional modification where the family of DNA methyltransferases (DNMTs) adds a methyl group (CH3) to the fifth carbon in the pyrimidine ring of the base cytosine in DNA (Figure 12). DNMT1 is thought to catalyze the maintenance of the methylation following DNA replication, and is thereby responsible for the inheritance of methylation patterns. DNMT3 is thought to add CH3 *de novo* early in development [74]. Regions in the genome with many CpGs are called CpG islands, and these areas are therefore interesting for investigating methylation. In healthy, normal cells ~80% of CpGs are methylated [75]. Methylation of promotors inhibits transcription factor binding and thereby suppresses transcription as illustrated in Figure 12. Methyl-CpG–binding proteins (MBPs) interact with

methylated DNA, and these proteins may also facilitate certain histone modifications after DNA replication by forming complexes with histone methyltransferases (HMTs) causing methylation of histone tails [74]. The effect of histone methylation is dependent on which residues of the histone tails that is marked. H3K4me3 is normally associated with transcriptional activation and H3K27me3 is generally associated with repression [76]. Also, the effect of histone methylation can be context dependent [77]. Histone methylation may also promote DNA methylation, the two processes enforcing each other [78]. The terms hypermethylation, from hyper (more) and hypomethylation, from hypo (less) are used to describe more and less methylation of certain CpGs when compared to the normal methylation level.

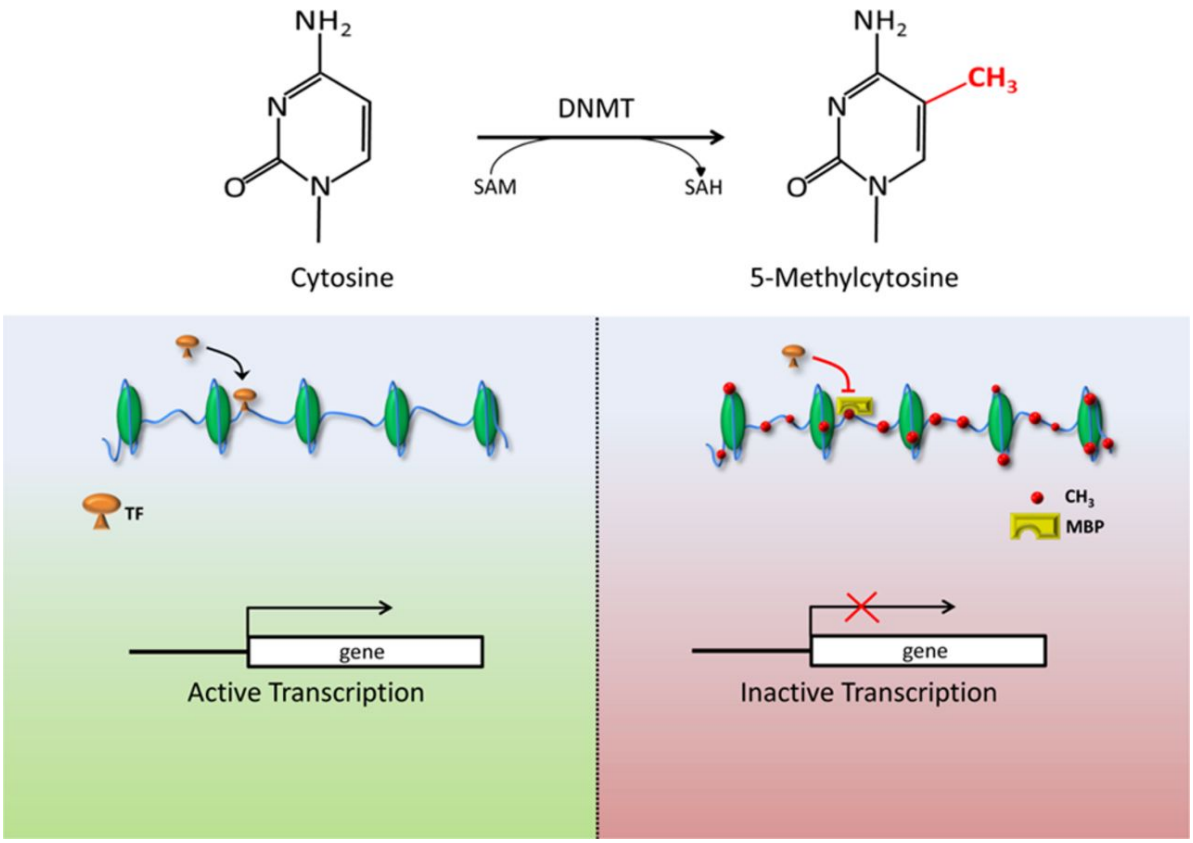


Figure 12. DNA methylation and its effect on transcription. DNA methyltransferase enzyme (DNMT) adds a methyl group (CH3) to the fifth carbon in the pyrimidine ring of the base cytosine in DNA. When methyl CpG-binding proteins (MBPs) bind to methylated DNA they prevent access to this sequence by transcription factors (TFs), thereby repressing transcription. Reprinted from [79].

1.4.2 Chromatin structure and gene expression

In eukaryotic cells the core histones pack and order the DNA molecule into nucleosomes, the basic element of chromatin. The histones have tails with amino acid residues that can be marked with a set of chemical modifications, including acetylation, methylation and phosphorylation [80]. "The histone code" is part of what regulates the density of the packaged DNA. De-condensed chromatin, euchromatin, is thought to be open to entry of the transcriptional machinery including RNA polymerase, and allows transcription of genes. Densely packed chromatin, heterochromatin, is associated with less accessibility for transcription, though studies indicate that the different chromatin states is part of a more complex picture [81].

1.4.3 DNA methylation in cancer

The patterns of methylation is essential for normal tissue differentiation and development of cell phenotypes. Aberrant methylation patterns are often found in cancer cells [82, 83], including breast cancer [84-87]. Both hypomethylation of promotors of oncogenes, which leads to a more transcribed gene, as well as silencing of tumor suppressor genes by promotor hypermethylation increases the risk of developing hallmark cancer cell characteristics. Some genes have been reported to be more hypermethylated in ER- than ER+ breast cancers, such as the tumor suppressor gene *CDKN2A* [88]. Another example is the tumor suppressor gene *BRCA1*, most known for being mutated in familiar breast cancer, but also frequently silenced by promoter hypermethylation in sporadic breast cancers [89]. Hypomethylation of repetitive genomic sequences can cause reactivation of transposable elements [90], and disruption of DNMTs may cause aneuploidy [91]. Global DNA hypomethylation in tumors can lead to chromosomal instability, further leading to deletions, amplifications and translocations [92], and several studies have investigated global DNA methylation differences associated to breast cancer clinical subgroups, including ER status [86, 93, 94]. DNA methylation patterns were for instance used by Fleischer et al. to further split tumors of the luminal A subtype into two subgroups with different prognosis [95]. DNA methylation research in cancer has until recently mostly focused on CpGs in *cis*. When introducing their genome-wide expression methylation quantitative trait loci (emQTL) analysis in 2017, Fleischer, Tekpli et al. included CpGs in the intergenic areas; in *trans* [96]. emQTL is a correlation analysis used to identify associations between levels of DNA methylation at CpGs and gene expression. Part of the importance of their discovery was that many of the CpGs in the ER-associated emQTLs were found in enhancers, not in genes, as earlier observed, indicating their role as part of gene regulatory networks and importance for phenotype. Two clusters were identified, one of them related to ER signaling. They demonstrated that DNA methylation at enhancers is linked to transcription factor activity and central for development of ER positive breast cancer. Furthermore, they found that

gene regulatory networks within the estrogen-related cluster were strongly dominated by ER, FOXA1 or GATA3. They conclude that hypomethylation of ER, FOXA1 and GATA3 binding sites is specific for ER+ tumors and may be central for the development of this disease.
One of the layers making this field even more complex is the recent advancement in lncRNA studies. Many lncRNAs have been shown to have direct interactions with chromatin-modifying proteins, influencing the epigenetic state [97, 98], several of them with Polycomb repressive complex 2 (PRC2). PRC2 catalyzes trimethylation of lysine 27 on histone H3 (H3K27me3), a mark that mediates the formation of heterochromatin, hence PRC2 can be recruited by lncRNAs to silence targeted genes [99, 100]. Aberrant expression, deficiency or mutation of both PRC2 and interacting lncRNAs have been associated with cancer [97]. Another example is the lncRNA H19, which has been reported to be overexpressed in breast cancer, and to prevent histone and DNA methylation, promoting cell growth, migration and invasion [101].

2 Aims

Long non-coding RNAs are a diverse class of RNA involved in crucial biological processes, both in healthy development and in disease progression. The development of next-generation sequencing has made discovery of novel forms of RNA more within reach, generating a large amount of newly annotated lncRNAs in recent years. While only a fraction of these have been functionally characterized, several have been linked to breast cancer and its known molecular subtypes, as well as to epigenetic mechanisms. The overall aim of this thesis is to identify and gain more knowledge about lncRNAs' role in breast cancer. The more specific goals of this project are:

- To identify lncRNAs differentially expressed between ER+ and ER- breast cancer
- To identify a robust set of lncRNAs across different cohorts and compare data between different platforms
- To identify lncRNAs with a strong association to DNA methylation in breast cancer
- To functionally characterize candidate lncRNAs both in *silico* and in breast cancer cell lines

3 Materials

3.1 Discovery cohort; TCGA BRCA

The Cancer Genome Atlas (TCGA) is a cooperation between the National Cancer Institute's Center for Cancer genomics and the National Human Genome Research Institute in the US, where multiple participating institutions have collected tissue and pre-processed data [6]. Molecular and clinical data for a range of cancer types have been made publicly available at the GDC data portal [102]. RNA-seq data and clinical data from 1095 samples from The Cancer Genome Atlas Breast Cancer Cohort (TCGA-BRCA, from here on only referred to as TCGA) were used for lncRNA discovery. Of these, 1045 patients had clinical ER annotation based on immunohistochemistry, of whom 237 were ER- and 808 were ER+, and these were used in the further analysis. TCGA methylation data were obtained by Illumina HumanMethylation450K beadchip arrays that measure DNA methylation levels of more than 450,000 CpG sites. Methylation data (level 3) were downloaded from the TCGA Data Portal [102].

3.2 Validation cohorts

Metabric and Oslo2 (OSL2), were used for validation of the results from TCGA. In validation the results obtained in the discovery cohort are compared with independent cohorts to identify the most robust biological findings.

3.2.1 Metabric

The Molecular Taxonomy of Breast Cancer International Consortium (Metabric) cohort is a Canadian and British project that aims to classify breast cancer tumors into further subcategories, based on molecular signatures [103]. Illumina HT-12 microarray expression data and clinical data from 1980 breast cancer samples, of which 1506 were ER+ and 474 were ER-, were used for validation of the differential expression analysis. Expression data were obtained from the European Genome-phenome Archive (EGA) with accession number EGAD00010000210.

3.2.2 Oslo2

The Oslo2 (OSL2) breast cancer cohort is an ongoing project started in 2006 at several south-eastern Norwegian hospitals. At the time of this project it contained all together 425 breast cancer samples with molecular data available, of which 349 had expression data with known ER status; 284 ER+ and 65 ER- [104]. Agilent SurePrint G3 Human GE 8x60K microarray expression data and clinical data were obtained from the Gene Expression Omnibus database (GEO) with accession number GSE58215 [104]. Illumina Infinium HumanMethylation450K methylation data from 330 patients were available from GEO with accession number GSE84207 [96]. 266 patients had both expression and methylation data.

3.3 Databases

3.3.1 Reference genome

A reference genome is a digital database of nucleic acid sequences as a representative example of a species' set of genes, a mosaic of DNA sequences from different donors. Reference genomes can be accessed online, using browsers such as Ensembl [105] or University of California Santa Cruz (UCSC) Genome Browser [106]. The Genome Reference Consortium (GRC) is an international collective of research institutes working on improving the representation of reference genomes. Several updates have been made of the human genome since the first release in 2003. The GRC human genome build 37 (GRCh37) is derived from 13 anonymous volunteers from Buffalo, New York. It corresponds to UCSC's hg19, and was released in 2009. The most updated version, GRCh38, corresponding to hg38, was released in 2016 [107]. The RNA-seq data from the TCGA used in this project were aligned to GRCh38.

3.3.2 Reference transcriptome

The Encyclopedia Of DNA Elements (ENCODE) is a research consortium launched by The National Human Genome Research Institute (NHGRI) in 2003 with the aim of identifying all functional elements in the human genome sequence [108]. Their annotated transcripts include both coding and non-coding RNA. Their selection is more conservatively curated than for example Ensembl, through a combination of computational analysis, manual annotation, and experimental validation [109]. The latest version, Gencode v27, is mapped to GRCh38 and contains annotation of 16 000 lncRNAs that can be downloaded as a FASTA file and used for quality confirmation of transcripts from RNA-seq data.

Several lncRNA databases have been developed especially for annotated lncRNAs. FANTOM CAT is a group that in 2017 released a collection of lncRNAs based on cap analysis of gene expression (CAGE) data. This provided annotation with more accurate 5' end positions, and generated a comprehensive atlas of 27,919 human lncRNA genes with high-confidence [110]. For lncRNA curation Gencodev27 and the FANTOM CAT database were used in this project.

4 Methods

4.1. Computational language and software environment R

In this thesis the statistical analyses were conducted and graphics were made using R version 3.5.1 [111] and RStudio version 1.1.383 [112]. R is a free statistical computation and graphics software. R provides a large set of packages, allowing a wide range of specialized statistical and graphical techniques. R contains a core set of packages that is included in the installation, but more than 15000 user-created packages are available, including many customized for biological data.

4.2 RNA-sequencing technology

RNA-sequencing (RNA-seq) utilizes next generation sequencing (NGS) to study the expressed part of a genome. Different methods for RNA extraction can be used for isolation of RNA from cells or tissue of interest at specific time points, developmental stages or after specific treatments. Following extraction, cDNA synthesis is performed. Starting with total RNA, different selections of RNA species can be performed, for example depletion of ribosomal RNA, which makes up ~80% of the total RNA. The RNA-seq libraries can be made strand specific by incorporation of uracils (U's) in first strand synthesis. In lncRNA research this is an important step, as there can be produced transcripts from both the sense and the antisense strands of the same locus, for example a coding transcript on one strand and a non-coding transcript on the other.

Several methods for NGS can then be used for sequencing the transcripts. One such workflow is illustrated in Figure 13 a. In library construction RNA is fragmented into ~200-500 base pairs (bp) by the use of ultrasound waves. PCR is used for amplification prior to sequencing. Adapters are ligated to both ends of the fragments. A flow cell is a specialized glass slide with lanes covered with two types of oligos. One of the types of oligos is complementary to one side of the adapters on the fragments with which it hybridizes. Polymerase synthesizes a new strand complementary to the fragment, and the original template is washed away. The end of the synthesized strand hybridizes with the second type of oligos, and polymerase synthesizes a second strand, clonally amplified through bridge amplification. Clusters of identical molecules are generated, and the process can be repeated, creating millions of fragments. The sequencing by synthesis takes place by polymerase adding four kinds of fluorescently labeled nucleotides, only one per cycle (Figure 13 b). For each cycle an imaging step registers the newly added nucleotides at each cluster. In this way millions of fragments are sequenced in a massively parallel process. The sequenced fragments can further be mapped to a reference genome [113].

In this project RNA-seq data from TCGA was used for discovery of lncRNA and curation of the data set used as the starting point for the analyses [6, 114].

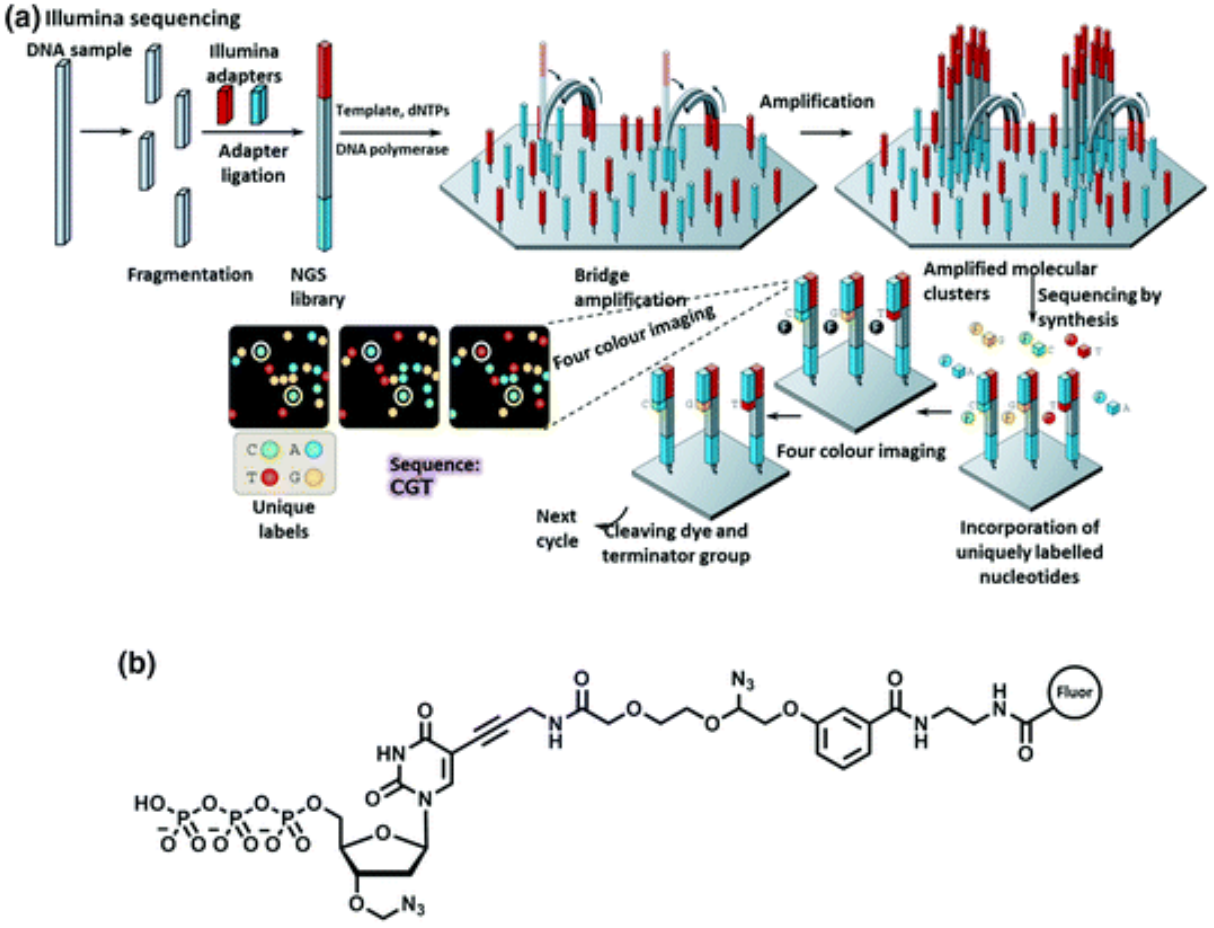


Figure 13. Overview of Illumina sequencing by synthesis technology. **a.** The process from cDNA through bridge amplification to sequencing by synthesis. **b.** Fluorescently labelled azidomethyl dNTP used in Illumina sequencing. Reprinted from [115].

4.3 RNA-seq bioinformatic pipeline

4.3.1 RNA-seq data

Publicly available RNA-seq data from breast cancer patients were obtained from TCGA [6]. Transcriptome sequence alignment files mapped with STAR to GRCh38 were used as a starting point for curating lncRNAs using the Tuxedo StringTie pipeline.

4.3.2 Quantification

The StringTie and Ballgown tool suite are free, open-source software tools for RNA-seq analysis, and parts of the Tuxedo StringTie pipeline [116]. StringTie is a tool to quantify genomic features. By providing StringTie with a reference transcriptome, the program will output quantifications for all features present in the file. Gencodev27 was used as a reference transcriptome. In addition StringTie will quantify features not present in the reference such as unknown genes and transcripts. The program builds isoform models to best fit the data. The isoform models that deviate from the reference transcriptome will be given a separate StringTie ID. In the downstream analysis isoforms for the same gene were pooled for total gene expression analysis.

4.3.3 Normalization

Fragments Per Kilobase Million (FPKM) was used to normalize for sequencing depth and gene length. The total RNA is fragmented before sequencing, resulting in more fragments from longer genes than shorter ones, which is why gene length is taken into account during quantification. The number of reads for a gene will also vary between samples due to batch effects and different factors such as DNA/RNA quantification and PCR efficiencies during library preparation, giving different sequencing depth. To attempt to normalize for sequencing depth, the total number of reads in a sample is divided by 1,000,000, giving the per million scaling factor. The read counts (the total number of reads mapping to a specific gene) is divided by the per million scaling factor, giving reads per million (RPM). To normalize for gene length, the RPM value for a specific gene is divided by the length of this gene, in kilobases, giving reads per kilobase million (RPKM). RPKM was made for single-end RNA-seq, where each read correspond to one fragment. FPKM is an adaption for paired-end RNA-seq, and takes into account that two reads correspond to one single fragment.

4.4 Workflow for lncRNA transcript assembly and quantification (TCGA)

StringTie was used to create a catalog of lncRNAs which was the starting point for the data used in this thesis. The workflow of lncRNA transcript assembly and quantification of RNA-seq data from TCGA with Gencode v27 as the reference transcriptome is illustrated Figure 14. First all transcripts shorter than 200 bp were removed. Transcripts that overlapped with annotated lncRNAs in Gencode, or the long non-coding RNA atlas FANTOM CAT, or both were kept for further analysis. The machine learning tool CPAT [117] was used to identify transcripts without coding potential. Finally the transcripts were filtered on expression level, keeping only transcripts above 1 FPKM in more than 5 % of the samples. lncRNA curation was performed by Sunniva Maria Stordal Bjørklund.

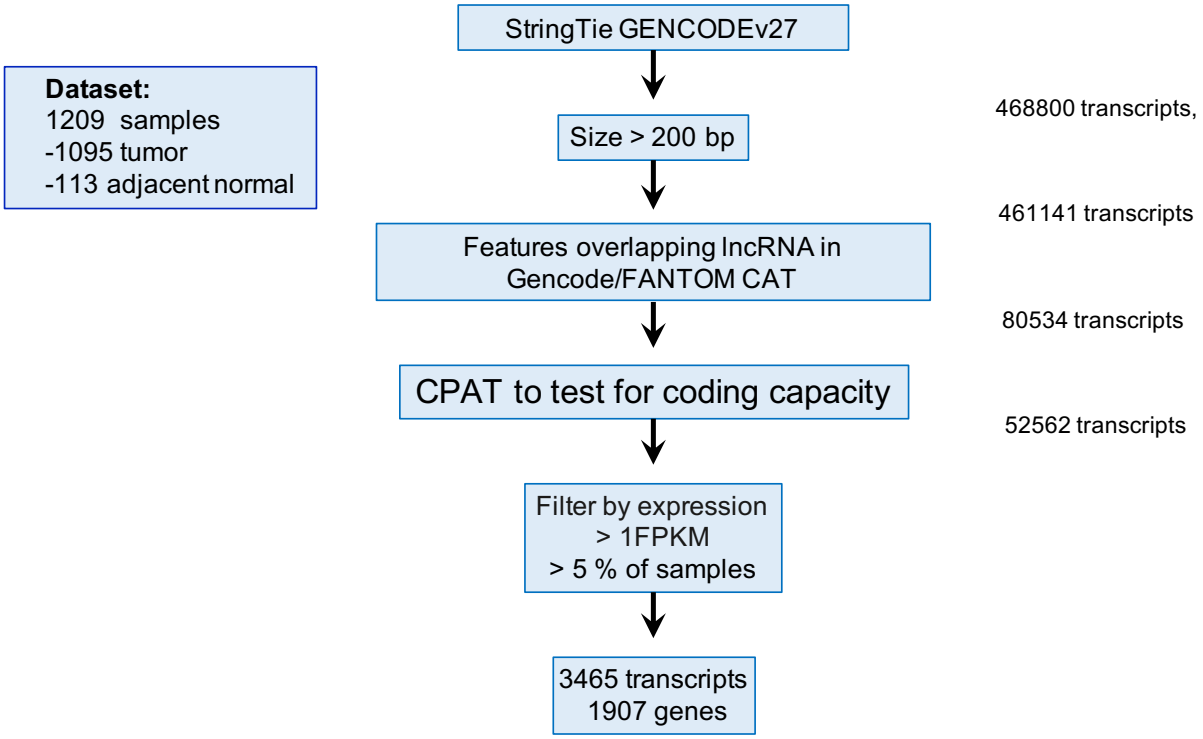


Figure 14. Workflow for lncRNA transcript assembly and quantification of RNA-seq data from TCGA to create a catalog of lncRNAs that was the starting point of this thesis.

4.5 Microarray

A DNA microarray is a chip that contains many spots of oligonucleotides attached to a solid surface. Fluorescently labeled cDNA in a sample hybridizes to the oligos on the chip, and the fluorescence signal from each spot on the chip is measured for quantification of each specific probe [118]. The DNA microarray technique was introduced in the mid 1990s and revolutionized the field of gene expression analysis, making it possible to compare the expression levels of a large number of genes in many samples. In this project, differential expression analysis is performed with data obtained from microarrays in OSL2 and Metabric. The method has some limitations. Because the technique relies on hybridization, one need to know the sequences of the targets before quantification. In contrast to NGS used in RNA-seq that relies on synthesis, where there is no need of known sequences in advance. Also the specificity of the probe and choice of location within the gene, affecting which transcripts will be targeted influence the quantification [119].

4.6 DNA methylation by Illumina Infinium Methylation450K beadchip

Illumina Infinium HumanMethylation450K beadchip assays [96] uses an adaption of a technique originally designed to assess single nucleotide polymorphisms (SNPs) to assess the methylation of single CpGs. The 450K version contains more than 450 000 CpG probes through out the human genome. The method uses bisulfite conversion where unmethylated cytosine (C) residues are deaminated to uracil (U) (Figure 15 a). Methylated cytosine residues remain unaffected, making the treated DNA contain only cytosines which are methylated. In whole-genome amplification the Us are converted to tyrosines (Ts), and then the DNA is fragmented [120]. The fragments are applied on arrays containing beads covered in probes for CpG sites. For most of the CpG sites there are two different types of beads, one unmethylated (U) and one methylated (M) (Figure 15 b). When an unmethylated CpG target site matches with a U probe, the converted base T matches with the probe's base A, enabling single-base extension with fluorescently red labeled nucleotides. When the same CpG target site hybridizes with a methylated probe, the single-base mismatch between the T and the G will inhibit extension, giving no signal. If the CpG locus of interest is methylated, the reverse occurs, giving extension of fluorescently green labeled nucleotides [121]. The amounts of red and green signals are processed to give a beta-value representing the degree of methylation for each CpG site. This is a continuous variable between one and zero, and can be correlated with other continuous variables such as expression levels.

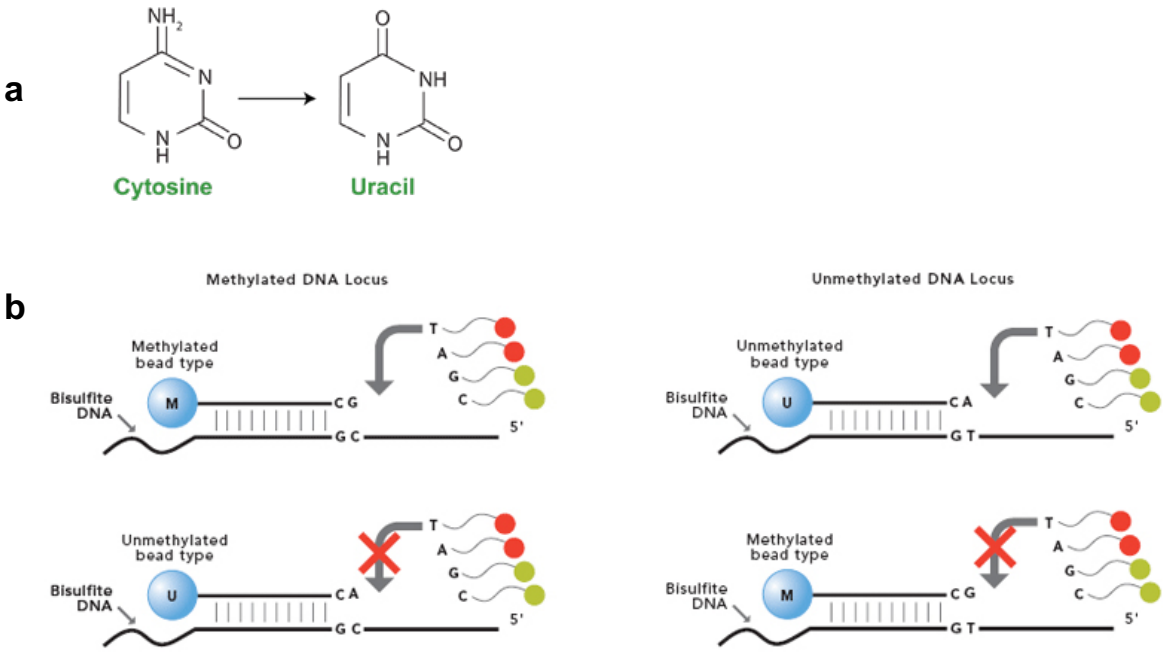


Figure 15. Beadchip array. **a.** In bisulfite conversion unmethylated cytosine residues are deaminated to uracil. **b.** The Infinium Methylation Assay uses two different bead types to detect CpG methylation. Only when unmethylated DNA containing C converted to T hybridize with the unmethylated bead type, the single-base binding is correct and enables extension with a red labeled nucleotide that can be imaged. For methylated DNA, correct single-base binding gives extension with a green labeled nucleotide. Single-base mismatch will not give extension of any color and therefor no signal. The illustration is adapted from [121, 122].

Methylation data sets generated from Illumina Infinium HumanMethylation450K beadchip arrays from TCGA [102] and OSL2 [96] were used for emQTL analysis to find significant associations between methylation of CpGs and lncRNA expression in the two cohorts.

4.7 Nuclear/cytosolic localization data

The raw RNA-sequencing reads from the nuclear and cytosolic fractions of the MCF7 BC cell line were obtained from the Short Read Archive (SRA) with the accession number GSE26284 [123].The ENCODE The data was generated as part of The ENCODE (ENCyclopedia Of DNA Elements) Project [124]. All protein coding genes and non coding genes in Ensembl v93 were quantified in house using Kallisto [125], an alignment free quantification tool for RNA-seq data.

4.8 Log2 transformation

Log2 transformation is performed to distribute the elements of data closer to a normal distribution, a prerequisite for many of the tests in the downstream analysis. The expression data from OSL2 and Metabric were already log2 transformed from the cohorts, and expression levels in TCGA were log2 transformed and centered when performing hierarchical clustering, generating heatmap and boxplots.

4.9 Differential expression analysis

In differential expression analysis statistical tests are used to discover quantitative changes in expression levels, identifying genes significantly up- or downregulated under varying conditions, such as different tissues, time points, cell stages, treated/untreated samples, or between patient groups. In this project the method was used to identify differentially expressed lncRNAs between ER+ and ER- tumors in three cohorts. Many different tools can perform the tests, according to different kinds of data distribution. Some tools can only perform pair-wise comparison, and others can perform multiple comparisons.

The Ballgown package from the Tuxedo StringTie pipeline is a package for downstream analysis of StringTie results. It is especially suitable for RNAseq data and was used to identify differentially expressed lncRNAs between ER+ and ER- samples in TCGA. The Stattest function in Ballgown uses a standard linear model-based comparison to identify transcripts that show statistically significant differences between groups [116].

A t-test is a statistical hypothesis test where the mean of two groups are compared. The null-hypothesis is that the means are equal to each other. The test is suited for normally distributed data. The t-test function in R was used to test differential expression of probes corresponding to lncRNAs in the OSL2 and Metabric cohorts. The t-test function was also used in Excel to calculate the significance of knockdown of candidate genes.

4.10 p-value

A cut off value for significance is decided, and if the resulting p-value is lower than the decided cut off, the null hypothesis is rejected. For differential expression analysis, a significance threshold of 0.05 was chosen.

4.10.1 Multiple testing correction

When multiple hypothesis are tested, as when comparing many genes in large data sets, the chance of false positives, Type 1 errors, increases with the number of tests. There are several ways of correcting for multiple testing. The Benjamini Hochberg (BH) method [126] is a procedure to control for the false discovery rate (FDR). The raw p values are ranked from low to high, and each each p value is multiplied by the number of variables, and divided by its rank order. If the corrected p value is less that the significance level, the variable is considered statistically significant. When using the t-test and stattest function in R, BH correction was used with p<0,05 considered statistically significant. Bonferroni correction is another method to control for the problem with FDR described above [126]. Here, the raw p values are multiplied by the number of tests performed.

4.11 Genome-wide correlation analysis of lncRNA expression and DNA methylation of CpGs

Correlation analyses are methods to calculate both the direction of relationships and the strength of them, because the scale is independent of the variables themselves. The correlation coefficient is the numerical measure of the strength and direction, a number between -1 and +1. Two widely used correlation statistics are Pearson and Spearman correlation. For linear relationships Pearson coefficient is used, and for non-linear, monotonic correlations Spearman can be used. Spearman can also be used for not normally distributed values and ordinal values. It is important to not confuse correlation with causation, as the correlating variables can have other causes that are not part of the analysis.

Genome-wide correlation analysis using Spearman correlation was performed with the R *cor* function to identify significant associations between the levels of DNA methylation at CpGs (Infinium Human Methylation 450K BeadChip) and expression levels of lncRNAs in the TCGA and OSL2 data set. Here, both associations between CpGs and lncRNAs on the same chromosome (*cis*), and CpGs and lncRNAs located on different chromosomes (*trans*) were included. Because variation between data points is necessary for identifying correlations, Interquartile Range (IQR) is used as a measure of variation, and data with less variation can be filtered out before the analysis. Here, CpGs with an IQR >0,1 were filtered out, and Bonferroni-corrected Spearman correlation p-values <0,05 were considered as statistically significant.

4.12 Hierarchical clustering and heatmaps

Clustering is a technique that groups similar data points into clusters, and this can be used to show the similarity between e.g. groups of genes and groups of patients. Hierarchical clustering can be divided into two types: Agglomerative and divisive. Agglomerative clustering is the most commonly used technique, a bottom up strategy. Each data point is considered as an individual cluster, and the closest clusters are merged. The merging of the new clusters is repeated as one moves up the hierarchy of similarity. Divisive clustering is the opposite strategy, a top down approach. All of the data points are considered as one individual cluster which is then split, and the splitting of the new clusters is repeated for each move down the hierarchy of similarity. A measure of similarity is required to decide the distance between observations. Different distance metrices and linkage rules can be applied. Distance metrices are functions that define the distance between elements in the form of a number. Different metrices that can be used are for example Manhattan distance, Euclidean distance, Binary, Maximum and Correlation distance. Finally different linkage criterions can be chosen to determine the distance between each cluster. The result is often presented as a dendrogram; a hierarchical tree, combined with a heatmap. A heatmap is a matrix were the values of each data points are visualized as colors. Hierarchical clustering and heatmap were used to show expression values of lncRNAs found differentially expressed between ER+ and ER- breast tumors in the TCGA data set. The clustering was performed using Correlation distance and Average linkage.

4.13 Box plots

Box plots are graphical representations of groups of numeric data by their quartiles. Each box displays the spread of data points within a group, reaching from the first quartile (Q1) to the third quartile (Q3). The horizontal middle line depicts the median. Vertical lines extend from the box to the smallest and largest non-outliers. Outliers are often displayed as individual points. Boxplots were used to visualize differentially expressed lncRNAs between ER+ and ER- patients.

4.14 Pathway enrichment analysis

Gene set enrichment analysis (GSEA) is a method to identify classes of genes that are overrepresented in a large set of genes. Statistical tests are used to identify significantly enriched or depleted groups of genes within the gene set that could be associated with specific biological phenotypes, for example subgroups within a cancer type [127]. ToppGene is an online GSEA tool for mammalian gene sets [128]. It contains libraries for transcription regulation, pathways and protein interactions, drug treated cell signatures and expression of

genes in different cells and tissues. The lists of genes are manually curated by ToppGene to be included in pathways based on different databases. As a part of functional characterization of candidate lncRNAs, ToppGene was used in this thesis. First, correlation analyses were performed between the expression of each lncRNA and the expression of all other genes, using the TCGA dataset. Then, highly positively and negatively correlated genes were separately retained (Spearman correlation>0,4 and < -0,4). These lists were then used as input to ToppGene to identify biological pathways associated with the candidate lncRNAs. ToppGene was run using default settings.

4.15 Quantitative model of R-loop forming sequences

An R-loop is a co-transcriptionally formed three-stranded hybrid nucleic acid structure, which consists of a newly formed RNA transcript still attached to its DNA template, and a fragment of a displaced non-template single stranded DNA. The RNA strand pairs with one of the two DNA strands in a region of homology. The R loop is then formed consisting of one RNA:DNA duplex and one single-stranded DNA [129]. Such formations are mostly determined by nucleic acid sequence and has their highest stability when guanine-rich RNA sequences hybridize with cytosine-rich ssDNA. [130], and these sequences are called R-loop forming sequences (RLFS). Kuznetsov et al. has proposed a quantitative model of RLFS (QmRLFS) that has predicted strand-specific chromosome coordinates of putative RNA:DNA hybrids and R-loops in the human genome, and the model has predicted RLFSs in 664 774 regions [131]. The QmRLFS can be used through an online tool [132], and this model was used for RLFS prediction for the candidate lncRNAs.

4.16 Cell line

The ER+ breast cancer cell line MCF7 was isolated from patient Frances Mallon by Herbert Soule and colleagues at The Michigan Cancer Foundation in 1970 [133]. It was the first long living, stable mammary cell line, and it is the source of much of the knowledge about breast cancer. MCF7 cells (catalog nr ECACC 86012803) were purchased from Sigma-Aldrich (Saint-Louis, MO, USA). The cells were grown without antibiotics in DMEM supplemented with 10% fetal bovine serum (FBS). The cells were incubated at 37°C in a humidified 5% CO2, 95% air incubator.

4.17 RNase H mediated knockdown

Gene knockdown is a term used for methods reducing the amount of a gene product; RNA or protein. Whereas knockout is a deletion of a gene resulting in no transcripts, knockdown only reduces the amount of transcripts. Knockdown experiments can be used for functional analysis of genes. In DNA replication Ribonuclease hybrid (RNase H) removes the RNA primers hybridized to the newly formed DNA strand. RNase H mediated knockdown utilizes this mechanism to hydrolyze the RNA strand in a DNA-RNA duplex. This technology is supposed to have more specific knockdown with fewer off-target effects than RISC mediated knockdown [134]. In order to functionally characterize the three candidate lncRNAs, a RNase H mediated knockdown was performed to investigate the effect on cell proliferation and viability.

4.17.1 LNA GapmeR

Antisense LNA™ GapmeR Standard (Qiagen, Hilden, Germany, catalog nr 339511) was used for knockdown of the candidate lncRNAs GATA3-AS1 (assay nr LG00222051-DDA), FAM198B-AS1 (assay nr LG00222009-DDA) and DRAIC (assay nr LG00221976-DDA). LNA GapmeRs are single stranded antisense oligonucleotides, 16 nucleotides long. They are introduced directly into the cytoplasm through transfection. LNA GapmerR is a special form of antisense oligonucleotides (ASOs) suitable for lncRNA. Probes hybridize only when the RNA is in the N configuration, not in the S configuration. LNA stands for “locked nucleic acids”, and they have a methylene bridge that keeps the RNA in a stable N configuration, making the probes hybridize more stably with the target. As Figure 16 shows LNA GapmeRs contain a short part of DNA flanked by LNA. The DNA “gap” in the center activates RNase H cleavage of the target RNA [135].

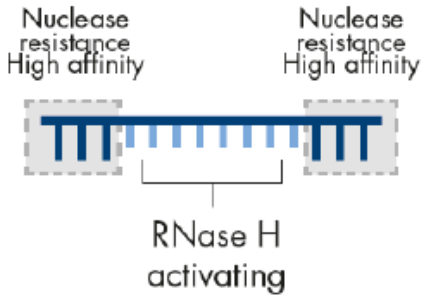


Figure 16. Antisense LNA GapmeRs are 16mer oligonucleotides. A DNA sequence is flanked by LNA, which increases the hybridization to the RNA target. The DNA “gap” in the center activates RNase H cleavage of the target RNA. Reprinted from [135].

4.17.2. Transfection

Transfection is to introduce nucleic acids into a cell, and it involves creating pores in the cell membrane to allow uptake of the genetic material. Different techniques can be used to disturb the cell surface, for example electroporation or transfection agents such as calcium phosphate, or as in this experiment, a cationic lipid that forms liposomes entrapping the DNA and fusing with the cell membrane. In revers transfection cells in suspension are added to pre-plated transfection complexes. In this experiment forward transfection was used, were the cells are seeded a day prior to transfection.

Lipofectamine® RNAiMAX Reagent (catalog nr 13778150) purchased from Thermo Fisher (Waltham, MA, USA) was used as the transfection agent. Allstars Hs Cell Death siRNA (catalog nr 1027298) purchased from Qiagen was used as a positive control to ensure that a transfection had taken place. Cell Death is a blend of siRNA targeting human genes essential for survival. As a negative control Antisense LNA™ GapmeR Control (catalog nr 339515) was used, which is a scrambled sequence of nucleotides that will not hybridize with the target genes. This was used to control for any effect on viability of the transfection itself. Untreated cells and cells treated with Lipofectamine without GapmeRs were also used as negative controls to ensure that other parts of the transfection process did not have an effect on the viability.

Three biological replicates were performed, though the last two were performed at the same day with two different cell concentrations. Each biological replicate contained three technical replicates of each of the three lncRNAs. The cells were seeded in 96 well plates for luminometry and 6 well plates for qPCR. Cells were plated at the following densities; 8000-10000 cells per well in 96 well plates, and 160000-200000 cells per well in 6 well plates.

After optimizing the protocol, forward transfection was chosen. The transfection complexes were mixed and pipetted onto the plated cells after 24 hours. For each well in the 96-well plate 0,2 μl Lipofectamine, 25 μl Opti-MEM™ I Reduced Serum Medium, no phenol red (Thermo Fisher, catalog nr 11058021) and 0,075 μl GapmeR (50μM) /negative-/positive control were mixed and added. For each well in the 6-well plate 4,0 μl Lipofectamine, 500 μl Opti-MEM and 1,5 μl GapmeR (50μM) /negative-/positive control were added. The plates were incubated at 37°C for 72 hours.

72 hours after transfection MCF7 cells were harvested in Phosphate-buffered saline (PBS), spun down, and pellets were kept at minus 80°C.

4.18 Cell viability assay

Luminometry is a type of spectrophotometry that uses bioluminescence, the form of luminescence produced by the interference of an enzyme. Luciferase can, in the presence of magnesium, oxygen and ATP, catalyze oxidation of luciferin, which results in light emission (Figure 17). For each ATP molecule reacting, one photon is released. The detection of the wavelengths of the released photons can therefor be used to measure the amount of ATP in cells, which is a measure of cell viability.

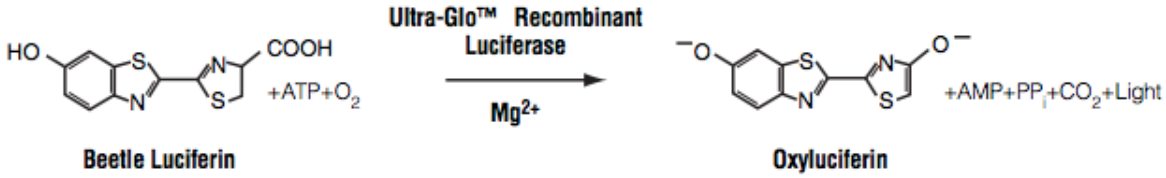


Figure 17. The luciferase reaction used in CellTiter-Glo. The assay contains a recombinant luciferase that catalyzes oxidation of luciferin in the presence of Mg2+, ATP and oxygen. Reprinted from [136].

CellTiter-Glo® Luminescent Cell Viability Assay (Promega, Madison, WI, USA) was used 72 hours after transfection with LNA GapmeRs. CellTiter-Glo® Reagent (CTG reagent) was diluted 1:1 with double distilled water (ddH2O). 75 μL of the medium in the 96 wells was pipetted out, and 50 μL with the diluted CTG reagent was added to make a 1:1 solution of medium and diluted CTG reagent. The luciferase reaction was started by placing the plate on a shaker without light. After 10 minutes luminometry was performed using VICTOR Multilabel Plate Reader X3 (PerkinElmer, Waltham, MA, USA).

4.19 RNA extraction

MCF7 pellets were thawed on ice, and total RNA was extracted using an automated version of column chromatography, QIAcube Connect (Qiagen). AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, catalog nr 80204) was used after the manufacturer's instructions. The principle used in this chromatography is forcing a liquid phase through a stationary phase in the form of a silica membrane by centrifugation. The lysis step (disruption of cell membranes) was done by vortexing and centrifugation in Buffer RLT Plus. Homogenization was done in an integrated shaker. The lysate is first passed through AllPrep DNA Mini spin columns. The nucleic acids adsorb to the silica, and the rest of the material passes as flow through. A denaturing and chaotropic guanidine-isothiocyanate-containing buffer creates a hydrophobic environment which makes adsorption of genomic DNA to the silica more

favorable than binding to water. The chaotropic salt also gives the silica positive charges that increase the affinity between the nucleic acids and the silica under hydrophobic conditions. A washing buffer is passed through the column to remove proteins or other impurities that have not bound to the membrane. An elution buffer without the chaotropic agent is then passed through, eluting the nucleic acids. DNA was here eluted in 30 μL buffer. The flowthrough was digested with Proteinase K together with ethanol to remove protein contaminants and passed through RNeasy Mini spin columns for binding of RNA. DNase 1 was used for DNA digestion. Finally RNA was eluted in 30 μL RNasefree water. NanoDrop™ One Microvolume UV-Vis Spectrophotometers (Thermo Fisher) was used for quantity and quality control of the eluted RNA samples.

4.20 First strand cDNA synthesis

Reverse transcription (RT) is the process of making complementary DNA (cDNA) from single stranded RNA. The reaction uses RNA as template, reverse transcriptase (RT), the heat stable enzyme taq DNA polymerase, primers and dNTPs. dNTPs are the four nucleotides dATP, dGTP, dCTP, dTTP, the building blocks that makes up the DNA thread. RNA is first incubated with primers at 70°C to denature the RNA secondary structure. Then a quick cooling makes the primers anneal to the RNA. One option is to use a primer called Oligo-dT which binds to the 3 prime (3') poly A tale of eukaryotic mRNA transcripts and adds dTTPs to the 3' dATPs. A limitation is that one can get a 3' bias due to where the primer binds. Another option is to use gene specific primers to yield cDNA only from specific targeted sequences. In this experiment random binding primers were used containing randomized hexamers which bind all along the RNA fragment. Then, at 37°C, RT synthesizes a complimentary strand by adding dNTPs. When reheated to 70°C, the enzyme is inactivated. Now the original single strand of RNA has a new built single stranded DNA thread bound to it, a double stranded molecule consisting of one RNA thread and one DNA thread. The single stranded cDNA can serve as a template in PCR reactions.

Total RNA from MCF7 cells was reversed transcribed using High-Capacity cDNA Transcription Kit (Thermo Fisher, catalog nr 4368814). Each PCR reaction contained 1,0 μg RNA, 2 μL RT Buffer, 0,8 μL dNTP Mix, 2,0 μL RT Random Primers, 1,0 μL Multiscribe™ Reverse Transcriptase, 0,5 μL Rnase Inhibitor, and 0,5 μL Nuclease-free H2O. 13,2 μL RNA soluted in water was added to 6,8 μL RT master mix. 20 μL was pipetted to each well for PCR in LifeECO™ (Hangzhou Bioer Technology, Binjiang, China). The thermal cycler was set to 25°C for 10 minutes, 37°C for 120 minutes, 85°C for 5 minutes, and then the samples was kept at 4°C over night.

4.21 PCR

Polymerase chain reaction (PCR) is a method for amplifying small amounts of DNA. The components needed for the process is DNA template, dNTPs, heat stable taq DNA polymerase and primers. A specific segment of a DNA template can be amplified by using primers designed to target a sequence of interest. In automated PCR, a program is set to change the temperature in a thermal cycler. The reaction is heated to denature the DNA template allowing DNA strands to separate. Then the temperature is lowered allowing the primers to anneal to their targets. The temperature is then raised for extension, allowing DNA polymerase to add nucleotides to the template strand between the primers, and synthesize a new double strand. The process is repeated, and for each cycle, ideally the number of DNA strands is doubled.

4.21.1 Realtime quantitative PCR for validation of knockdown

Real-time quantitative PCR (RT qPCR) is a method to quantify the PCR product during the PCR process. The curves in Figure 18 show the increase of PCR product in real time during the temperature changing cycles. It is a sensitive method, and therefor suitable for showing small changes in gene expression. Relative quantification shows the change in mRNA products of the PCR process relative to a control mRNA and relative between samples. In real-time PCR the threshold cycle (Ct) is the PCR cycle number where the amplification curve crosses a signal threshold. Ct is used as a relative measure of the concentration of the target/gene of interest. In a 100% efficient PCR the number of transcripts would double for each cycle in an exponential curve. That means that the number of cycles to reach the signal threshold is dependent on the number of target transcripts in the start of the reaction. When two samples, one treated and one untreated are run together, the two Ct values are compared, and a difference in Ct value on 1 means that one had twice as much transcripts from the start. The lower the Ct, the higher amount of transcripts, because it takes fewer cycles to reach the threshold.

For validation of knockdown, the levels of transcripts from both the treated and the untreated samples were measured in real-time and compared. If the Ct values are significantly lower for the treated samples, there has been a successful knockdown. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is called a housekeeping gene because it is often constitutively expressed at high levels in different tissues unaffected by varying conditions. It is much used as a reference gene in PCR, as a GAPDH background. To compare the two targets, the treated and the untreated samples, the background is subtracted as a form of normalization, to be able to compare the values without for example pipetting differences skewing the difference one wants to investigate.

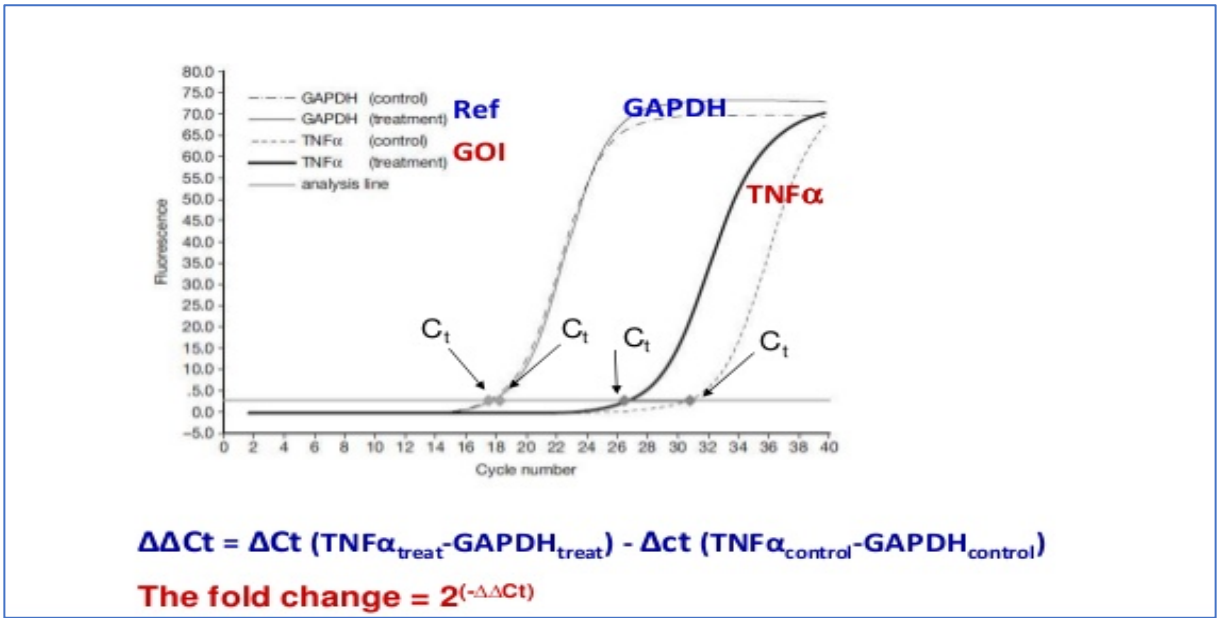


Figure 18. The curves show the increase of PCR product in real time during the temperature change cycles. The threshold cycle (Ct) is the PCR cycle number where the amplification curve rises above the threshold for the background signal. Reprinted from [137].

TaqMan® Gene Expression Assay (Thermo Fisher, catalog nr 4351372) was used to assess the relative expression of the three candidate lncRNAs GATA3-AS1 (assay nr Hs04401807_g1), FAM198B (assay nr, Hs04975817_m1), DRAIC (assay nr, Hs05011039_m1), as well as GATA3 (catalog nr 4331182, assay nr Hs00231122_m1) in MCF7 cells. The TaqMan Assays are based on 5' nuclease chemistry, where a fluorogenic probe enables the detection of the specific PCR products as they accumulate. The probe has a FAM™ dye label on the 5' end, a minor groove binder (MGB) and on the 3' end a nonfluorescent quencher (NFQ), a compound that can absorb the energy from the fluorophore. When the polymerase reaches a TaqMan probe, its 5' nuclease activity cleaves the probe, separating the dye from the quencher, and the fluorescent energy can be released and detected. GAPDH was used as the endogenous control gene. The primers were designed by the manufacturer, and chosen to detect the transcripts targeted by the GapmeRs.

Each PCR reaction contained 2 µL cDNA/GAPDH, 5,0 µL TaqMan® Gene Expression Master Mix (catalog nr 4369016), 0,5 µL TaqMan Assay and 2,5 µL nuclease-free water was mixed. For standard curve a sample from untreated cells was diluted in nuclease-free water as follows: 1:1, 1:10, 1:100, 1:1000.

10 μL was pipetted into wells in 384 well plates with three technical replicates for each cDNA sample. Cells from three biological replicate knockdown experiments for each lncRNA were used. H2O and a negative control from the reverse transcription were used to ensure that the effects measured were not due to other circumstances of the experimental setup. The PCR was performed in Applied Biosystems 7900HT Version 2.3 Sequence Detection Systems (Thermo Fisher). PCR conditions included a first step of 50°C for 2 minutes, then the initial activation step at 95°C for 10 minutes, followed by 40 cycles of 95°C denaturation for 15 seconds, 60°C annealing and extension for 1 minute.

4.21.2 Analysis of RT qPCR results

The ΔΔCt method was used to analyze the Ct values from RT qPCR [138]. The following formula was used to calculate relative expression:

$$\text{ratio} = \frac{(E_{\text{target}})^{\Delta \text{CP}_{\text{target}}(control - sample)}}{(E_{\text{ref}})^{\Delta \text{CP}_{\text{ref}}(control - sample)}}$$

CP is defined as the point at which the fluorescence rises above the background fluorescence. Efficiency (E) is each primer set's capacity for amplification, where perfect efficiency is a doubling of the number of transcripts for each cycle. The standard curve is used for defining the efficiency of the primer set. E_{target} is the RT PCR efficiency of target gene transcript, and E_{ref} is the RT PCR efficiency of the reference gene transcript. ΔCP_{target} is the CP deviation of control minus sample of the target gene transcript. And ΔCP_{ref} is CP deviation of control minus sample of reference gene transcript, in this case GAPDH. Control is here the negative control of the knockdown.

5 Results

5.1 Identification of lncRNAs in the TCGA BRCA cohort

lncRNAs are emerging as an important class of RNA transcripts, several with known gene regulatory functions. lncRNAs are expressed with great specificity in regards to cell type and tissue [139]. There are several databases with annotations of lncRNAs, and with each update, that catalog of lncRNAs grows. In order to analyze a large collection of lncRNAs we used RNA sequencing from the TCGA BRCA, and the newest versions of the Gencode and the FANTOM CAT annotation databases. All together these databases had annotation of 23810 unique lncRNA genes. The workflow for lncRNA transcript assembly and quantification of RNA-seq data from TCGA (n=1209) as previously shown in Figure 14. Among the samples analyzed were 113 adjacent normal samples that were part of lncRNA-identification, but these samples were not a part of the further analysis. The StringTie Ballgown tool suite [116] was used to create a catalog of lncRNAs that was the starting point for the data used in this thesis. Raw sequencing files were aligned to the human genome (GRCh38), and transcripts were assembled and quantified using StringTie provided with the Gencode v27 transcriptome annotation. This generated 468800 transcripts. Curation of lncRNAs by removing transcripts shorter than 200 bp resulted in 461141 transcripts. From these, transcripts that overlapped with annotated lncRNAs in Gencode or the long non-coding RNA atlas FANTOM CAT, or both, were kept, leaving 80534 transcripts. The machine learning tool CPAT was trained to select genes with coding capacity, leaving 52562 transcripts that were predicted to be non-coding. Filtering on expression level, keeping transcripts above 1 FPKM in more than 5% of the samples left 3465 transcripts corresponding to 1907 genes.

5.2 Differential expression of lncRNAs in ER+ versus ER- patients in the discovery cohort TCGA BRCA

The two most clinically relevant subgroups of BC is ER+ ER-. Because of the pressing need to understand more of the mechanisms through which ER-driven breast cancers become aggressive, as well as the emerging view of lncRNAs as essential for gene regulation, we wanted to identify ER-associated lncRNAs. Based on immunohistochemical staining (IHC) of ERα, as annotatd in the clinical file, TCGA BRCA cohort was divided into ER+ (808) and ER- (237) patients. Differential expression analysis using Ballgown stattest to identify significant associations between expression levels of 1907 lncRNAs and ER status resulted in 1386 significant lncRNAs of which 857 were up-regulated in ER+ and 529 up-regulated in ER-. Genes with FDR-corrected p<0,05 were considered significant. Hierarchical clustering of the expression values of the resulting lncRNAs was performed to create the heatmap shown in Figure 19, which illustrates, as expected, a quite clear separation of tumors based on ER status. When it comes to the PAM50 classification, the luminal groups (luminal A and luminal

B) are mixed, but separated from the Basal like. The patients in the Her2 subtype show no clear pattern. The heatmap shows a clear 3-part separation of expression levels of lncRNAs, the one group highly expressed in almost all patients, while the other groups has low and moderately expression levels in almost all patients.

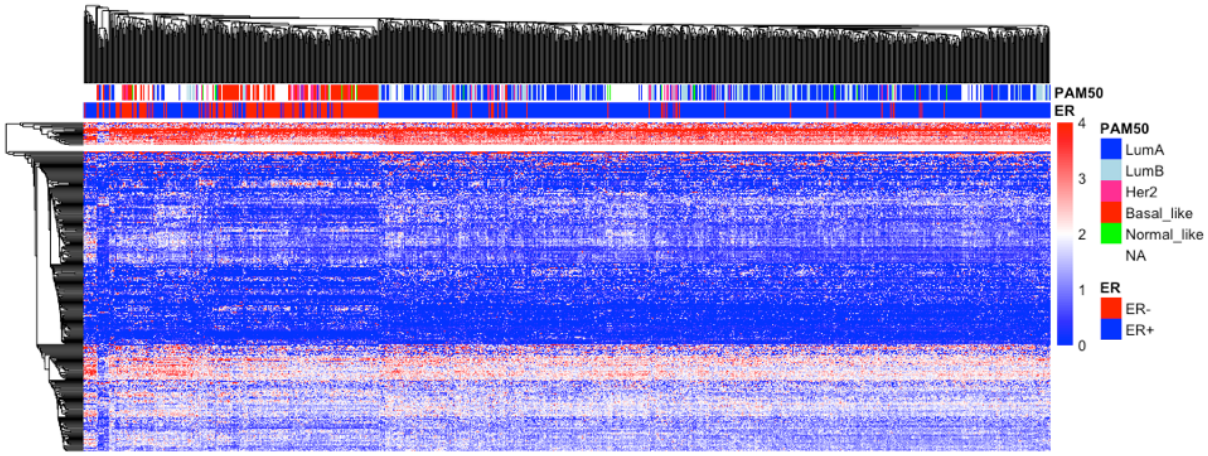


Figure 19. Hierarchical clustering of expression values of the 1386 lncRNAs found differentially expressed between ER+ and ER- breast tumors in the TCGA data set. The clustering was performed using correlation distance and average linkage, and expression levels were log2 (FPKM+1) transformed and centered.

5.3 Validation of results in independent cohorts and matching of lncRNA IDs

Previous efforts by Niknafs et al. has identified a long list of ER-associated lncRNAs in TCGA [24], but these results have not been validated in independent cohorts. Comparing results between cohorts is important all the time the presence and pattern of certain lncRNAs can be specific for TCGA. In order to identify robust lncRNAs we wanted to validate the significant results using microarray expression data from two independent cohorts, OSL2 (n=349) and Metabric (n=1980) (Figure 20). To be able to compare lncRNAs, gene IDs had to be matched between the cohorts. The matching of StringTie IDs to microarray probes was performed based on genomic location; leading in many cases to several microarray probe IDs corresponding to the same StringTie-ID (Figure 20). Of the 1386 StringTie IDs that were significantly differentially expressed between the two patient groups in TCGA, 743 genes corresponding to 2198 Agilent probes were present in the OSL2 data. In Metabric 658 StringTie IDs were found, corresponding to 1341 Illumina probes (Figure 20). To identify lncRNAs with differential expression in ER+ and ER- patients t-tests were performed on the 2198 and the 1341 lncRNA probes in OSL2 and Metabric, respectively (Figure 20). 1150 and 894 lncRNA probes were significantly differentially expressed in these two patient groups in the two cohorts (BH corrected t-test p-values < 0.05). The significant probes mapped to 533

StringTie IDs in OSL2 and 513 StringTie IDs in Metabric, meaning 72 % (533/743) was validated in OSL2 and 78 % (513/658) was validated in Metabric.

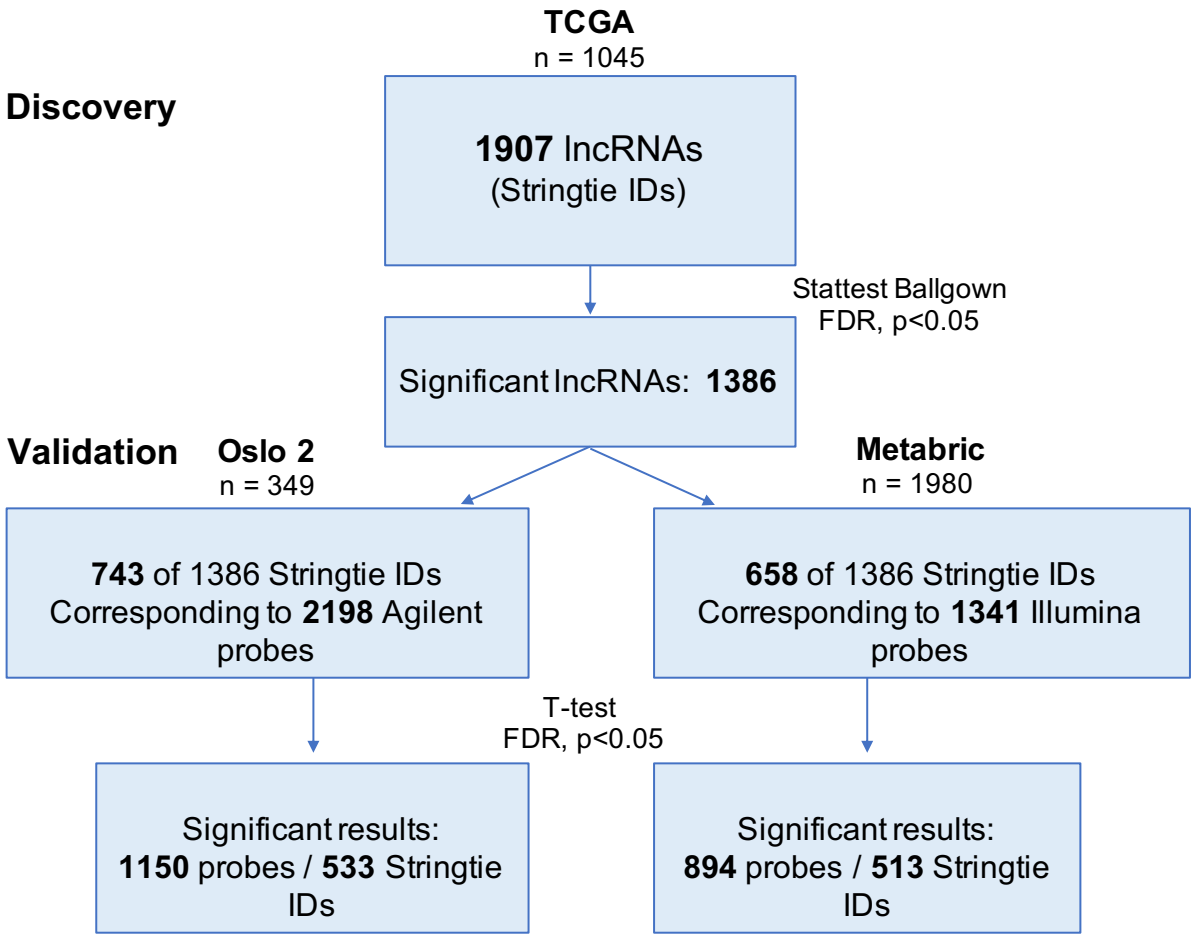


Figure 20. Workflow of discovery and validation, and matching of IDs between cohorts. 1386 lncRNAs was found differentially expressed between ER+ and ER- breast tumors in the TCGA data set. 72 % (533/743) was validated in Oslo 2 and 78 % (513/658) was validated in Metabric.

When comparing the results from the t-tests in OSL2 and Metabric and from the stattest in TCGA, all together 354 lncRNAs were associated with ER status in all the three cohorts, of which 227 were up-regulated in ER+ compared to ER- patients, and 127 were down-regulated (Figure 21 and Supplementary table 1). These results represent a robust identification of ER-associated lncRNAs across the three cohorts.

- 533 validated lncRNAs in OSL2
- 513 validated lncRNAs in Metabric

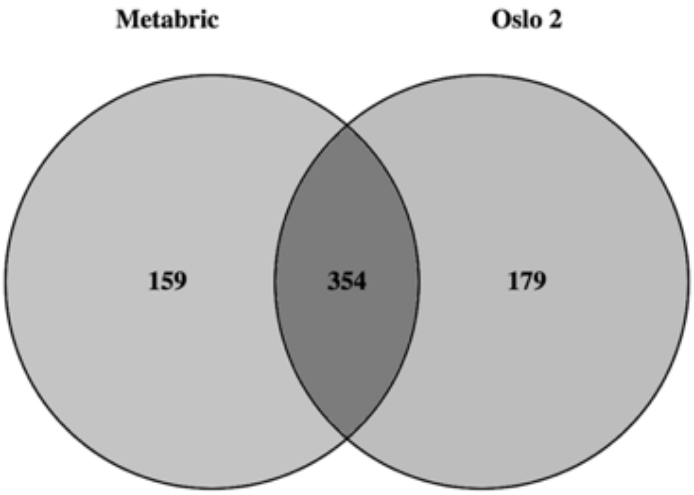


Figure 21. Overlap of validated results. The overlap between StringTie-IDs of the validated results from OSL2 and Metabric resulted in 354 lncRNAs significant in all 3 cohorts.

The analysis of lncRNAs confirms previous observations which shows that many lncRNAs are expressed dependently on ER status. Among the results were lncRNAs previously known to be associated with breast cancer in addition to novel candidates. Examples of lncRNAs showing consistent expression patterns in the two main clinical groups across several cohorts are NRAV (AC012531), previously known to be associated with immune infiltration and MIR193BHG, associated with breast cancer. Examples of novel results are AL590133.2, LINC01116 and AC244205.1 (Figure 22).

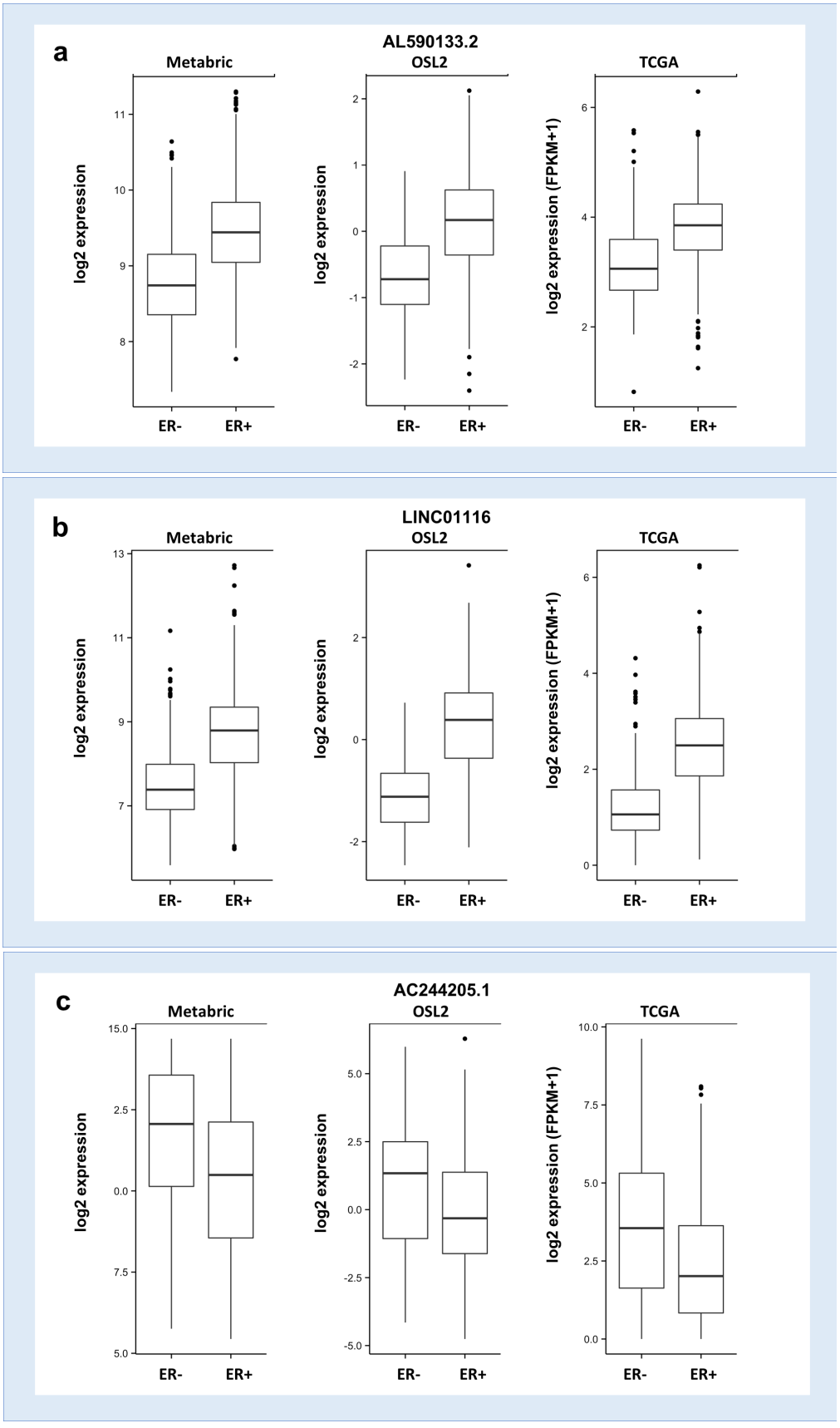


Figure 22. Examples of significant lncRNAs showing the same trend of different expression in ER- versus ER+ samples across three cohorts (Metabric, OSL2, TCGA). **a.** AL590133.2 **b.** LINC0116 **c.** AC244205.1.

5.4 Associations between lncRNA expression and methylation status

Many associations found between a lncRNA's expression and methylation at CpGs substantiates its involvement in epigenetic regulation. Therefore we wanted to add an additional layer of analysis to the previous efforts by Niknafs et al. [24], who focused only on ER-associated lncRNAs. Using the emQTL method, where both *cis*- and *trans* associations are identified, the associations between DNA methylation at CpGs and expression of the 354 identified ER-associated lncRNAs were quantified in the TCGA and OSL2 cohorts. As the Metabric cohort did not have methylation data available, this cohort was not considered in this analysis. The CpGs included in the analysis was initially filtered independently in each cohort to only retain CpGs with an IQR > 0,1. Considering only lncRNAs with at least one significant association to a CpG, 265 significant lncRNAs were identified in OSL2 and 298 in TCGA, resulting in 265 significant lncRNAs altogether when the results were intersected (Figure 23 and Supplementary Table 1).

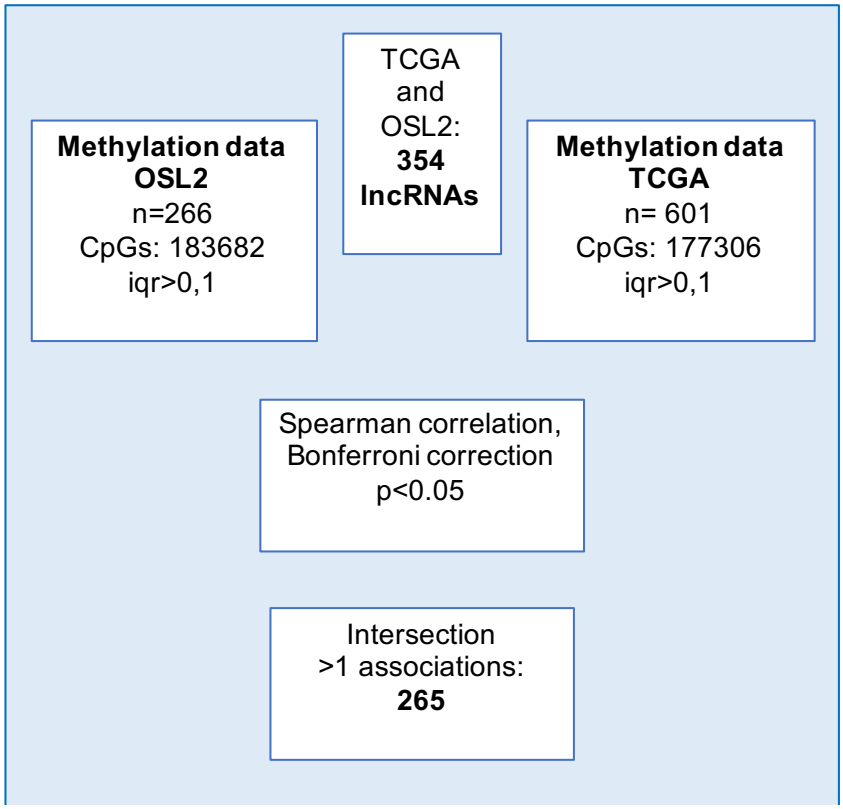


Figure 23. Workflow of the emQTL analysis with ER-associated lncRNAs. Genome-wide correlation analysis between levels of CpG methylation and expression level of lncRNAs resulted in 265 significant lncRNAS with >1 association OSL2 and 298 in TCGA, and 265 when intersected.

The distribution of significant CpG-lncRNA associations shows that a few lncRNAs account for most of the associations (Figure 24).

The lncRNA with the highest number of associations in both TCGA and OSL2 was PRKCQ-AS1 with 54391 associations in TCGA and 25691 in OSL2. This lncRNA was downregulated in ER+ in TCGA and OSL2. The second and third highest ranked in TCGA were the lincRNA AL928654.5 with 47331 associations, downregulated in ER+, and LINC02095/SOX9-AS1 with 43783 associations, also downreg<ulated in ER+. In OSL2 the second and third highest ranked were MIR3142HG, a lincRNA with 24692 associations, that was downregulated in ER+, and AC098679,1 a lncRNA with 15399 associations, found downregulated in ER+.

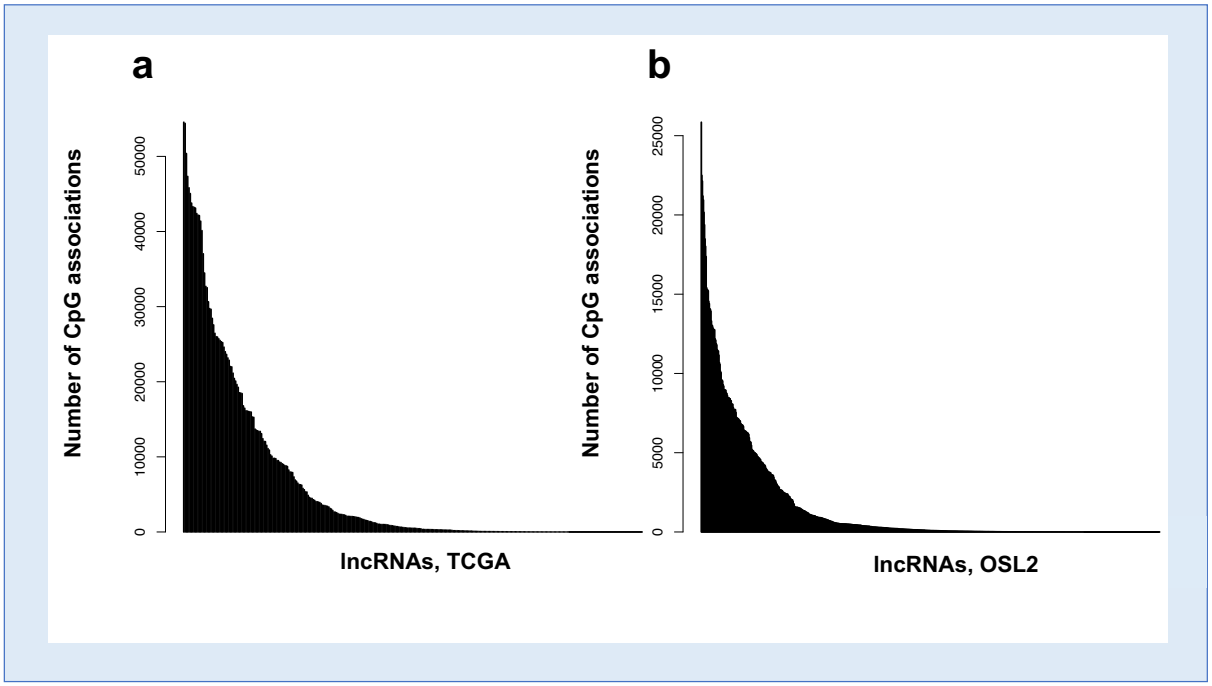


Figure 24. The distribution of significant correlations between lncRNAs and CpGs. **a.** TCGA. **b.** OSL2.

5.5 Functional characterization of candidate lncRNAs in the ER+ cell line MCF7

In order to identify lncRNAs with possible driver functions in ER+ breast cancer we wanted to perform a knockdown viability assay. To select candidates for the assay, lncRNAs were first prioritized among the results from testing on ER status, and then chosen based on high expression in ER+ tumors. To further narrow it down, the lncRNAs with potential functions in epigenetic regulation were prioritized by choosing lncRNAs with many associations to CpG methylation levels (Figure 25 and Table 1).

As an additional step to ensure that the candidates were truly non-coding, the resulting probes from the emQTL analysis were filtered by blasting the microarray probe sequences to annotated lncRNAs in the Ensembl93 database, keeping only transcripts with high confidence annotation as lncRNAs. This left 84 high confidence lncRNAs for candidate selection.

To be able to have a possible function related to chromatin structure, a lncRNA has to be located to the nucleus where the DNA is located. To identify lncRNAs with higher expression in the nucleus, we quantified levels of specific lncRNAs in the nuclear and cytosolic fractions of MCF7 [123] and further intersected these values with the other results (Figure 25 and Table 1). Documented expression levels in MCF7 were also taken into account, considering the functional study planned in this cell line. Finally, search in existing literature was done to identify interesting candidates still uncharacterized in breast cancer. When all the criteria were taken into consideration, it led to the selection of three candidates: GATA3-AS1, FAM198B-AS1 and DRAIC.

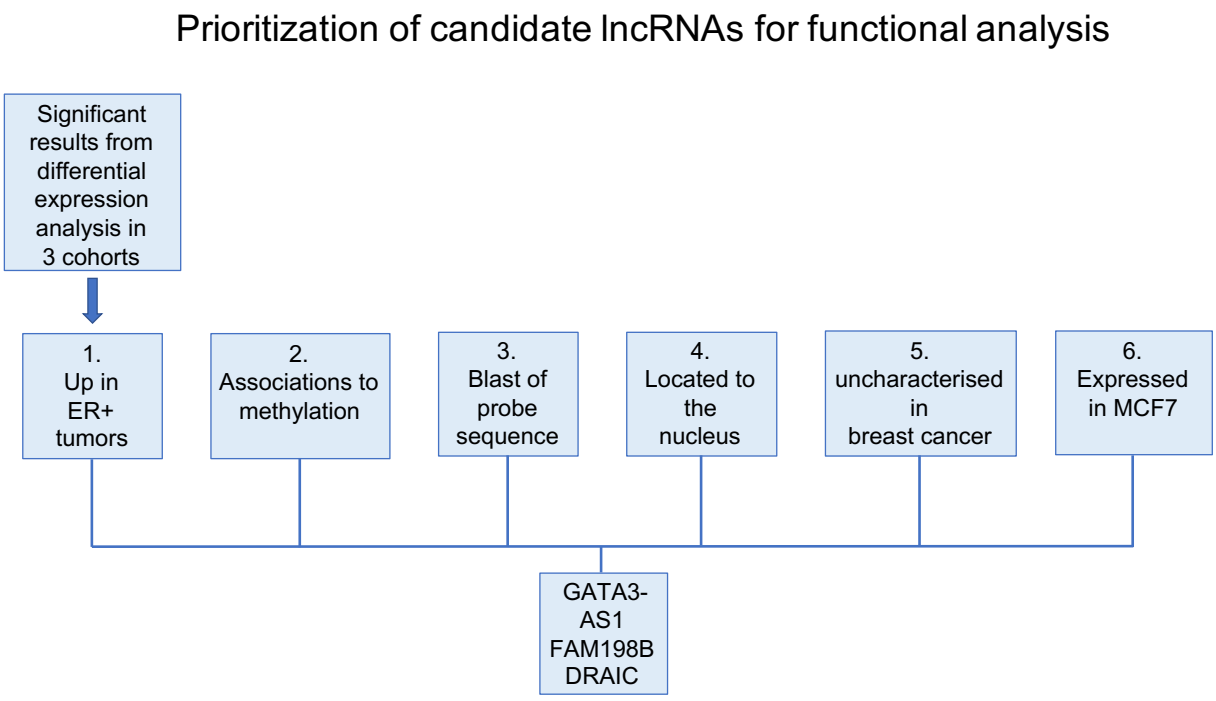


Figure 25. The selection process of the candidates for functional study. The selection of the candidates GATA3-AS1, FAM198B and DRAIC were chosen among the significant in all 3 cohorts after differential expression analysis and further based on criteria in the following order: **1.** Upregulation in ER+ tumors in TCGA. **2.** Associations to more than 15 000 methylated CpGs in TCGA and more than 8 000 in OSL2. **3.** Blast of microarray probes to ensure high confidence lncRNAs. **4.** Localization of transcripts with a ratio nucleus over cytosol >1. **5.** Manual search in existing literature to find candidates not yet characterized in breast cancer. **6.** Expression in MCF7 cells.

Table 1. Sorting values for prioritization of lncRNAs. The table is here sorted after number of associations to methylated CpGs in OSL2, and excludes results with less than 1000 associations in both OSL2 and TCGA. Selected candidates are highlighted in bold. Prioritization is shown by column numbers: **1.** Fold change in TCGA. **2.** Associations to methylated CpGs in TCGA. **3.** Associations to methylated CpGs in OSL2. Here the frequencies are shown for the array probe with the highest number of significant associations. **4.** Ratio of transcripts; nucleus over cytosol in MCF7. Only candidates with probe blast match to Ensemb.93 ncRNAs are presented here.

lncRNA name	Fc TCGA	CpGs TCGA	CpGs OSL2	Nucleus/cytosol
PRKCQ-AS1	0,81	54391	25691	0,00
FAM30A	0,56	47331	18313	27,41
ZNF213-AS1	1,12	2066	15189	4,67
GATA3-AS1	**1,29**	**23166**	**14163**	**2,09**
FAM198B-AS1	**1,25**	**15957**	**13789**	**3,40**
RMDN2-AS1	0,81	1215	13245	9,01
DRAIC	**1,90**	**42119**	**8523**	**3,67**
LINC01235	0,82	7326	8216	2,15
EIF3J-AS1	1,28	10255	7772	1,23
AC015922.4	1,46	22022	6150	0,41
AC093525.6	1,21	6750	5131	12,65
LINC01184	1,35	19630	5006	0,83
CRNDE	1,28	25995	3136	3,87
AL031429.2	1,45	13695	2681	7,89
AC008124.1	1,36	25319	2622	1,00
LINC01116	2,34	25710	2479	1,15
PCAT18	1,42	3127	2140	0,93
AC004847.1	0,91	43307	2112	17,35
AL589935.1	1,25	3411	1780	21,81
AL096870.2	1,21	2056	1547	0,63
LINC02095	0,51	43783	1528	6,58
MIR17HG	0,85	6915	1501	22,99
AL590617.2	0,83	9231	1259	9,51
LINC01503	0,77	8795	1109	7,17

The three selected candidates had the following fold change (fc) between ER+ and ER- tumors in TCGA: GATA3-AS1: 1,29, FAM198B-AS1: 1,25 and DRAIC: 1,90, giving them the 2., 3. and 4. highest fc in TCGA of all the tested lncRNAs that were upregulated in ER+ tumors. Differential expression of the candidates is consistent in all the three cohorts compared (Figure 26). Furthermore GATA3-AS1 had a nuclear to cytosol ratio of 2,09, FAM198B-AS1 had 3,40, and DRAIC had 3,67. The candidates were not ranked as the highest in the localization data, but all had a ratio >1.

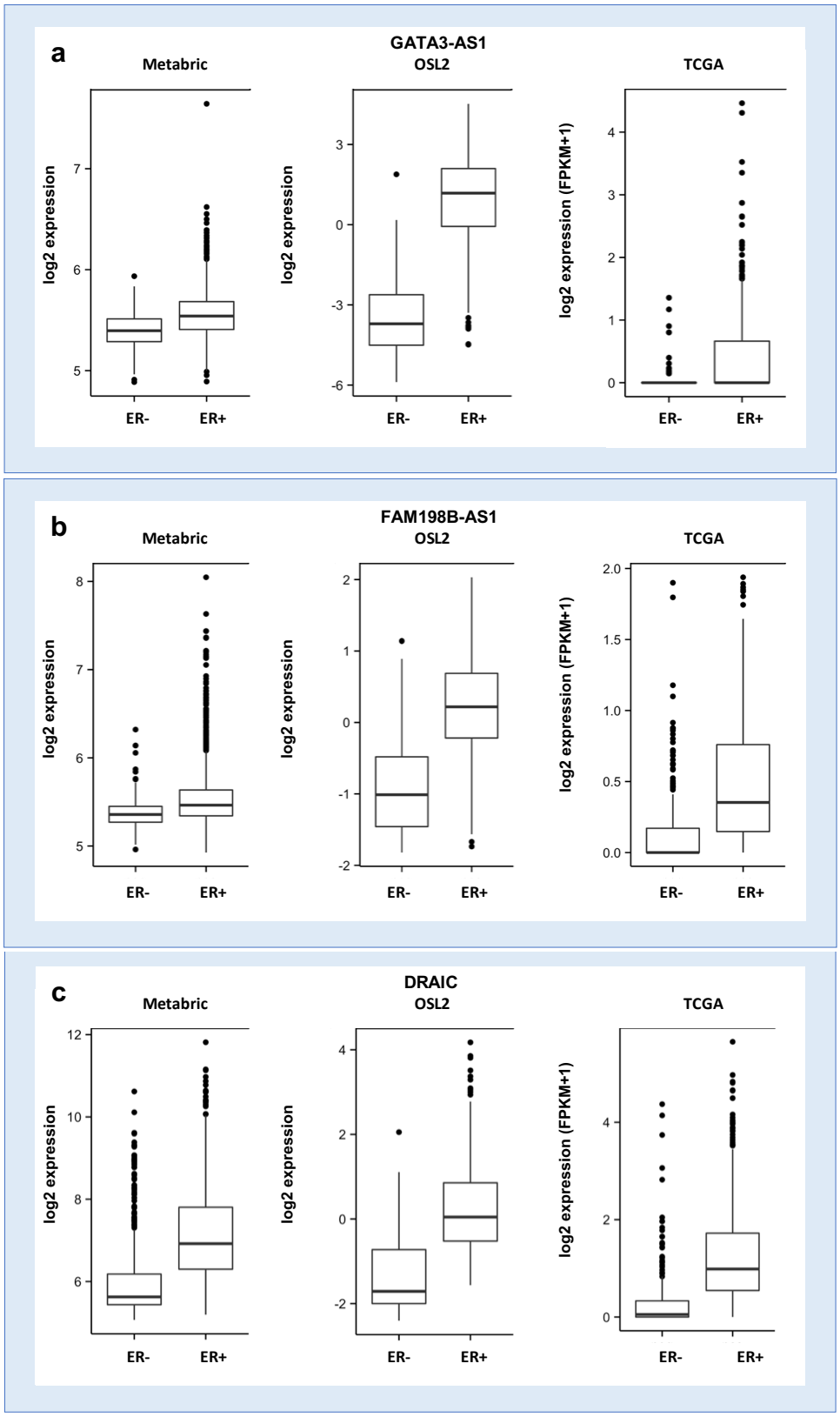
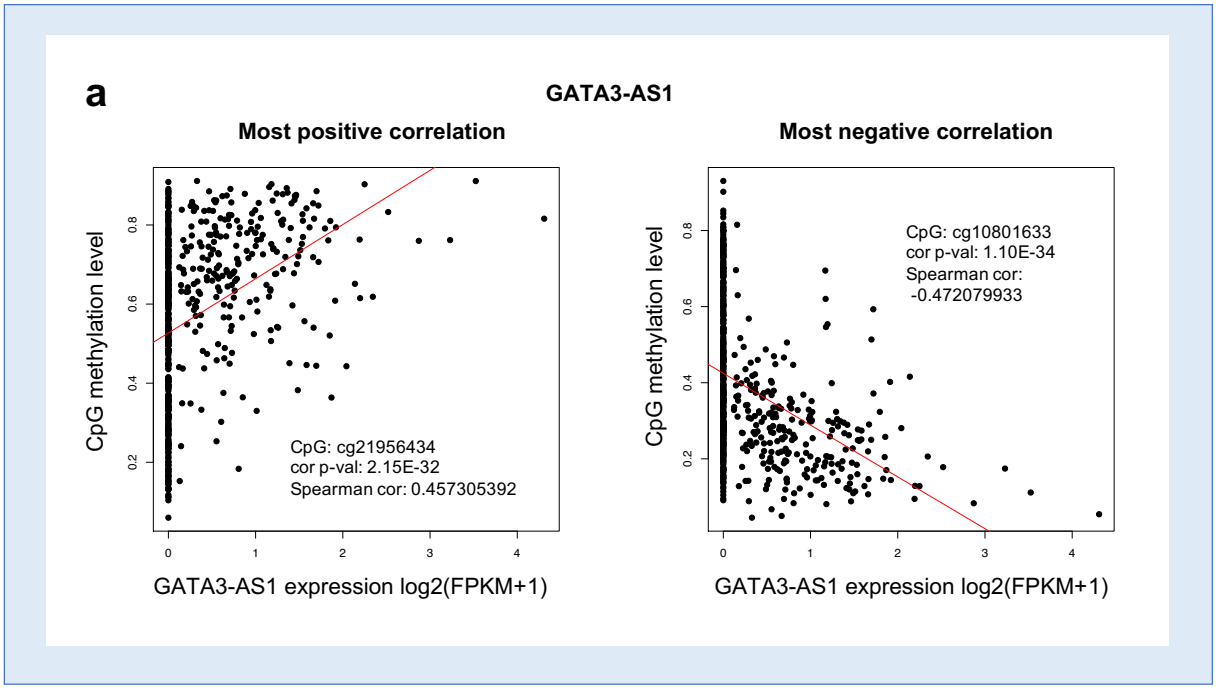
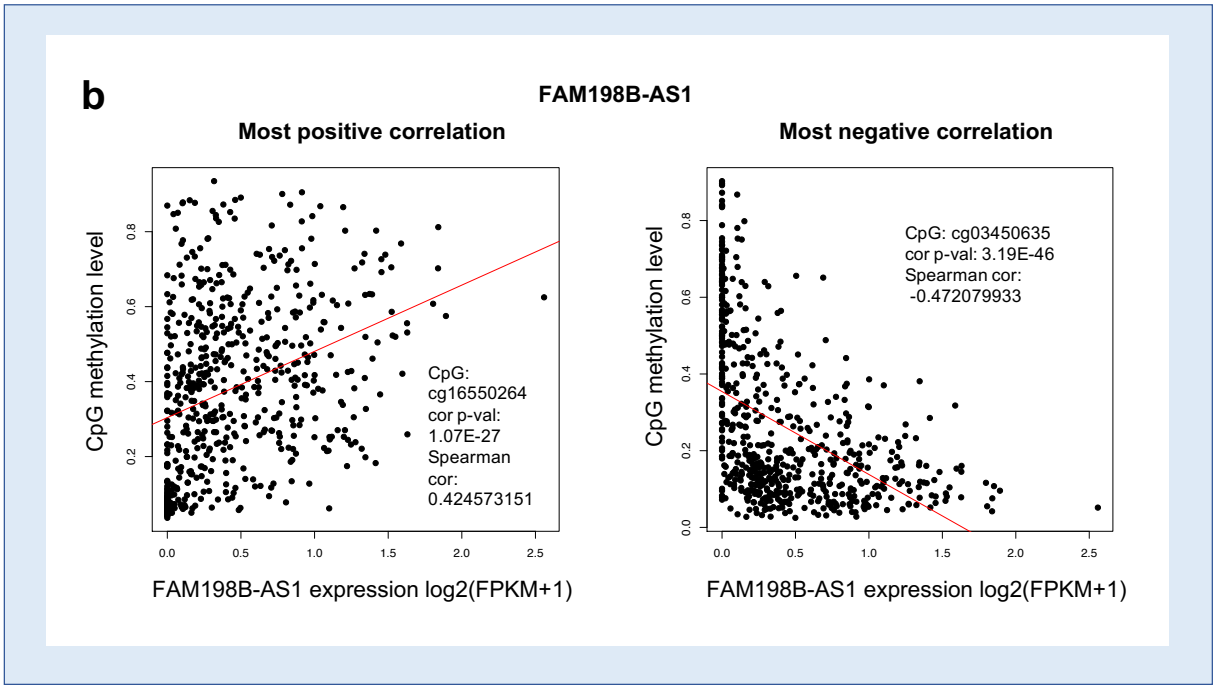


Figure 26. Differential expression of DRAIC, FAM198B-AS1 and GATA3-AS1 in ER- versus ER+ samples in three cohorts compared (Metabric, OSL2, TCGA); **a.** DRAIC. **b.** FAM198B-AS1. **c.** GATA3-AS1.

When considering possible epigenetic functions, by inspecting the number of CpG methylation associations, GATA3-AS1 had 23166 associations in TCGA and 14163 in OSL2, FAM198B-AS1 had 15957 in TCGA and 13789 in OSL2, and DRAIC had 42119 in TCGA and 8523 in OSL2, ranking them as number 4, 5 and 7 in OSL2. The number of associations was much higher in TCGA in general. For illustration the most significant positive and negative correlations between the expression of the candidate genes and methylation level at the CpGs in TCGA are plotted in Figure 27.





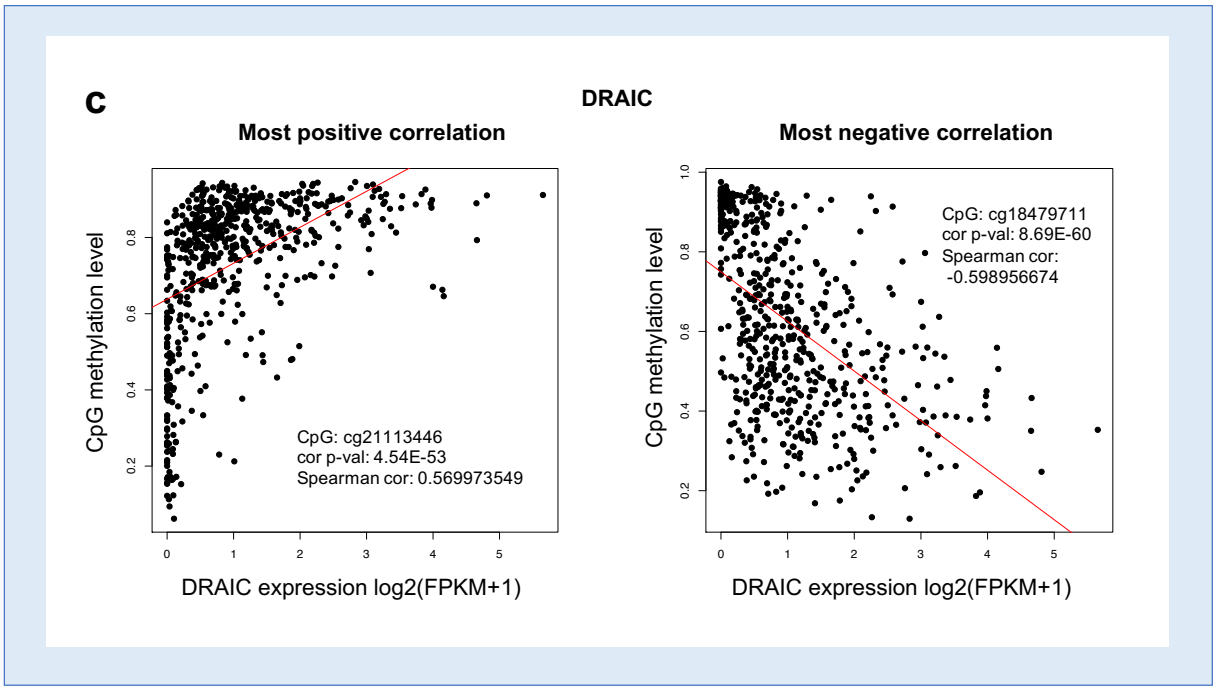


Figure 27. The most significant positive and negative correlations between the expression of the candidate genes and methylation level at the CpGs in TCGA. **a.** GATA3-AS1**. b.** FAM198B-AS1. **c.** DRAIC.

An overlap of the CpGs associated to the three candidates are shown in Figure 28. Almost 3000 CpGs were overlapping between all the three lncRNAs in both cohorts.

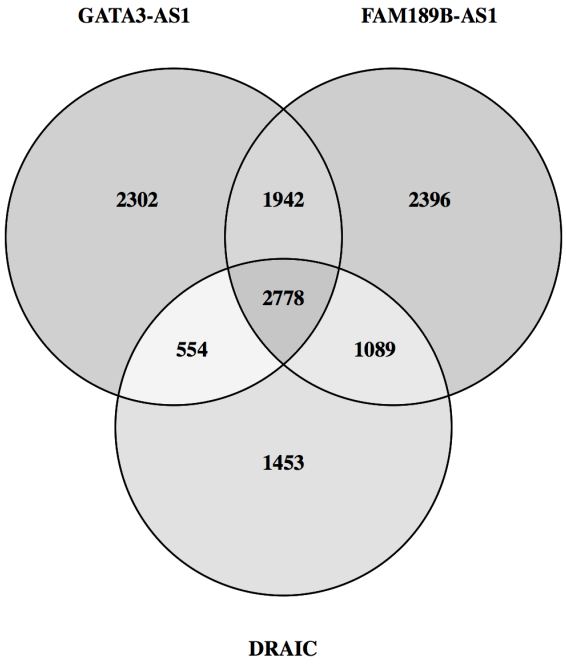


Figure 28. An overlap of CpGs associated to the three lncRNAs shows there are 2778 CpGs that overlap in all three genes.

5.6 Knockdown of candidate lncRNAs in MCF7

In order to functionally characterize GATA3-AS1, DRAIC and FAM198B-AS1, we wished to assess possible driver functions of the three lncRNAs in ER+ breast cancer through a viability assay. To investigate the candidates' effect on cell viability, a knockdown of the three lncRNAs was performed in the ER+ cell line MCF7. Cells were transfected with LNA GapmeRs for 72 hours (see Methods for details) before RNA was extracted in order to quantify the three lncRNAs by RT q-PCR. Nanodrop values are listed in Supplementary Table 2. Inspection of the treated cells compared to cells transfected with a positive control that induces cell death three days after transfection indicates a successful transfection. Relative expression in MCF7 72 hours after transfection shows a significant reduction of all 3 targeted transcripts in treated cells relative to negative control GapmeR (Figure 29). Analysis of RT q-PCR values using the $\Delta\Delta$Ct method shows a ~50% reduction of GATA3-AS1 transcripts, a ~50% reduction of FAM198B-AS1 transcripts, and a ~50% reduction of DRAIC transcripts. Ct values from q-PCR are listed in Supplementary Table 3.

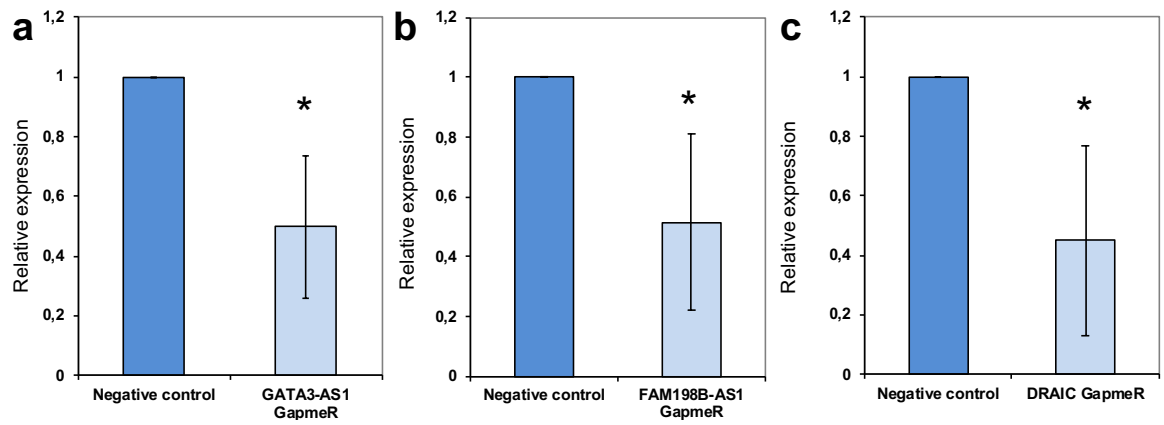


Figure 29. Validation of knockdown of candidate lncRNAs by qPCR. Relative expression in MCF7 72 hours after transfection shows a significant reduction, ~50%, of all 3 targeted transcripts in treated cells relative to negative control GapmeR. **a.** GATA3-AS1. **b.** FAM198B-AS1. **c.** DRAIC. Asterisk (*) indicates t-test p-value <0,05.

5.6.1 The effect of GATA3-AS1 knockdown on GATA3 transcription

A recent study [140] has showed that GATA3-AS1 regulates transcription of divergently expressed GATA3 in T-helper 2 cells (Th2). GATA3 is also an important transcription factor in ER+ breast cancer contributing to the regulation of genes associated with estrogen dependent tumor growth [141]. Therefor we wanted to investigate whether GATA3-AS1 had the same function of regulating GATA3 in BC as in Th2. We used RT qPCR to assess the effect of the knockdown of GATA3-AS1 on GATA3 expression. Relative expression 72 hours after transfection showed a ~30% reduction of GATA3 in GATA3-AS1 knockdown compared to the negative control (Figure 30).

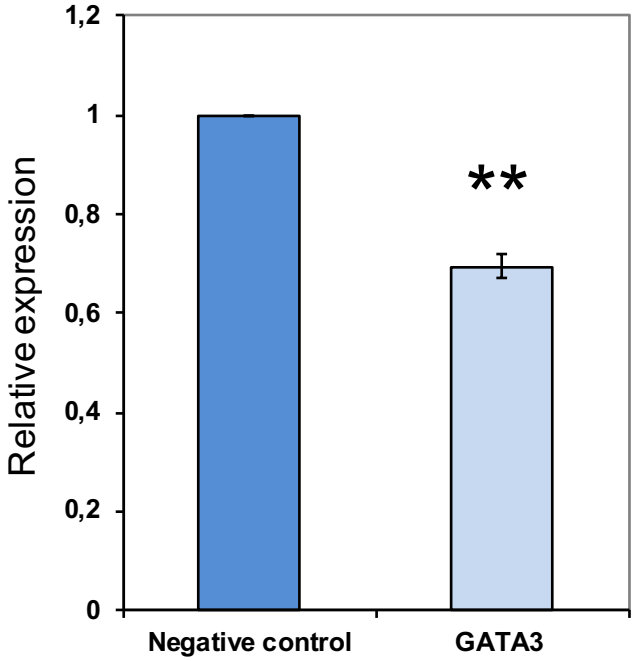


Figure 30. Knockdown of GATA3-AS1 effect on GATA3-expression. Relative expression 72 hours after transfection showed a ~30% reduction of GATA3 in GATA3-AS1 knockdown compared to the negative control. This is the mean relative fold change from two independent experiments. Two asterisks (**) indicates t-test p-value <0,005.

5.6.2 The effect of candidate knockdown on cell viability with CellTiter-Glo

To assess the effect of the knockdowns on cell viability, CTG was used as described in the Methods section, and the values are listed in Supplementary Table 4. Transfection with positive control resulted in ~30% reduction in cell viability, which shows that the transfections have taken place. However, knockdown of the three lncRNAs shows no reduction in cell viability after 72 hours, meaning that we did not observe any significant reduction in cell viability in any of the candidates at the time of measurement (Figure 31).

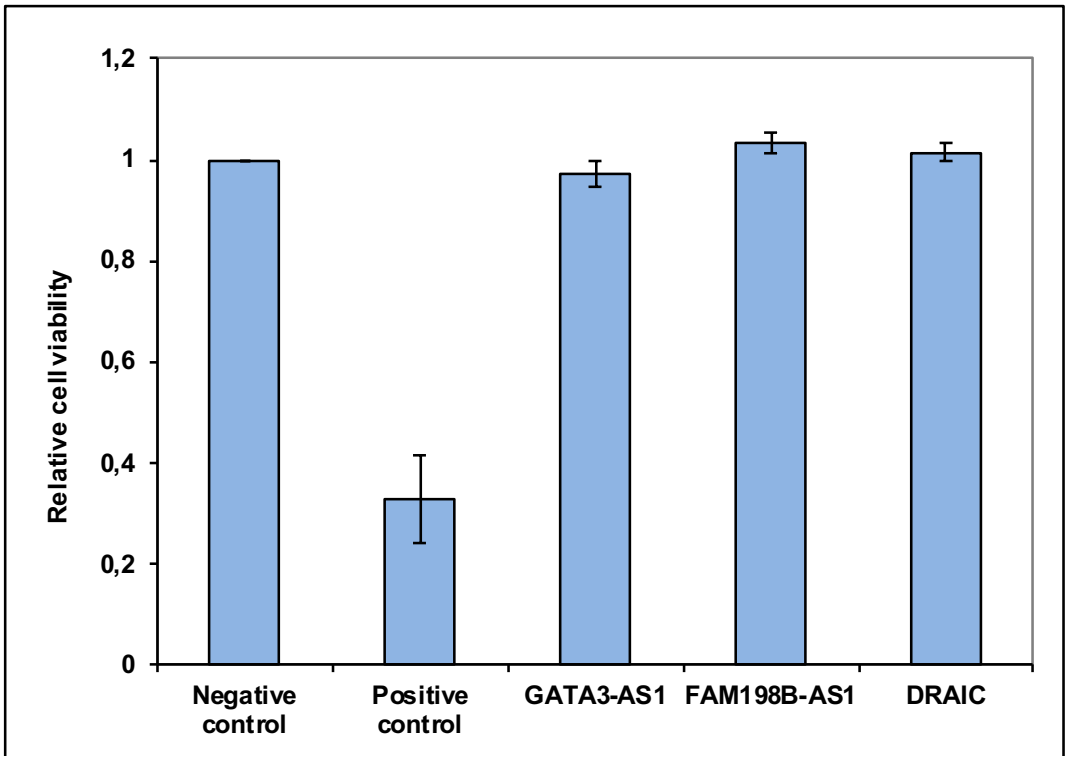


Figure 31. The effect of knockdown of candidate lncRNAs on cell viability. The diagram shows mean values from 3 experiments performed in triplicates. There were no significant difference in cell viability.

5.7 Pathway enrichment analysis of highly correlated genes

To further explore possible functions of the three candidate lncRNAs we performed correlation analysis between expression of all human genes and expression of the candidates. For each lncRNA the highly positively and negatively correlated genes were separately retained. Here, high correlation was defined as Spearman correlation > 0,4 and < -0,4. All p-values were highly significant. The gene lists (6 in total) were used to perform GSEA using ToppGene using default settings [128]. The top ten identified pathways for genes positively and negatively correlated to each candidate lncRNA are shown in table Tables 2 A

and 2 B. Both GATA3-AS1 and FAM198B-AS1 positively correlated genes showed enrichment in for pathways related to cilium assembly, which plays a role in cell cycle [142].

2 A. Top ten enriched pathways for genes positively correlated to each of the lncRNAs.

GATA3-AS1			FAM198B-AS1			DRAIC		
Pathway	**Name**	**q-value FDR B&H**	**Pathway**	**Name**	**q-value FDR B&H**	**Pathway**	**Name**	**q-value FDR B&H**
REACTOME 1268846	Cilium Assembly	1.3E-03	REACTOME 1269650	Generic Transcription Pathway	1.7E-22	SMPDB SMP00032	Valine, Leucine and Isoleucine Degradation	2.6E-02
Pathway Ontology PW:0000674	insulin secretion pathway	1.3E-03	REACTOME 1269877	Membrane Trafficking	8.0E-15	KEGG 524496	Ceramide biosynthesis	3.8E-02
REACTOME 1268838	Organelle biogenesis and maintenance	3.4E-02	REACTOME 1269876	Vesicle-mediated transport	9.6E-12	GenMAPP MAP00280 Valine_leucine_and_isoleucine_degradation	MAP00280 Valine leucine and isoleucine degradation	4.4E-02
BIOCYC 142359	leucine degradation	3.4E-02	REACTOME 1268846	Cilium Assembly	1.1E-05	KEGG 524497	Sphingosine biosynthesis	4.4E-02
REACTOME 1268850	BBSome-mediated cargo-targeting to cilium	3.4E-02	KEGG 102279	Endocytosis	3.1E-05	KEGG 82952	Valine, leucine and isoleucine degradation	4.4E-02
KEGG 413354	Leucine degradation, leucine => acetoacetate + acetyl-CoA	3.4E-02	REACTOME 1268848	Cargo trafficking to the periciliary membrane	5.7E-05	Pathway Ontology PW:0000674	insulin secretion pathway	4.4E-02
REACTOME 1268848	Cargo trafficking to the periciliary membrane	3.4E-02	REACTOME 1383038	Intra-Golgi and retrograde Golgi-to-ER traffic	1.6E-04	Pathway Ontology PW:0000677	altered insulin secretion pathway	4.4E-02
REACTOME 1268853	Intraflagellar transport	5.0E-02	REACTOME 1268725	Transport to the Golgi and subsequent modification	2.7E-04	BIOCYC 142334	beta-alanine degradation	4.4E-02
			REACTOME 1269649	Gene Expression	2.7E-04	REACTOME 1339134	Defective ABCC8 can cause hypoglycemias and hyperglycemias	4.4E-02
			REACTOME 1268726	ER to Golgi Anterograde Transport	7.8E-04	BIOCYC 142420	anandamide degradation	4.4E-02

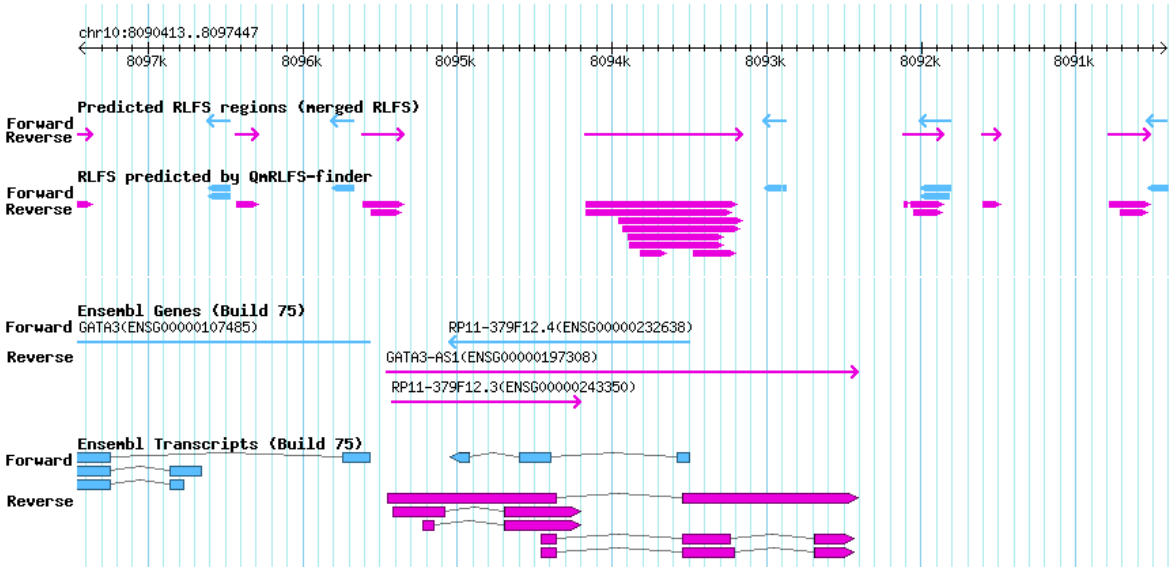
2 B. Top ten enriched pathways for genes negatively correlated to each of the lncRNAs.

GATA3-AS1			FAM198B-AS1			DRAIC negative		
Pathway	**Name**	**q-value FDR B&H**	**Pathway**	**Name**	**q-value FDR B&H**	**Pathway**	**Name**	**q-value FDR B&H**
REACTOME 1269530	Signaling by NOTCH	3.0E-02	REACTOME 1269741	Cell Cycle	3.7E-03	Pathway Interaction Database 169351	Validated targets of C-MYC transcriptional activation	5.0E-02
REACTOME 1269535	Signaling by NOTCH1	3.5E-02	REACTOME 1269763	Cell Cycle, Mitotic	3.7E-03			
KEGG 82995	Glycosphingolipid biosynthesis - lacto and neolacto series	5.0E-02	REACTOME 1269851	APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	3.7E-03			
			REACTOME 1269837	Regulation of mitotic cell cycle	4.4E-03			
			REACTOME 1269838	APC/C-mediated degradation of cell cycle proteins	4.4E-03			
			KEGG 83054	Cell cycle	1.3E-02			
			REACTOME 1269691	mRNA Splicing - Minor Pathway	1.5E-02			
			KEGG 373901	HTLV-I infection	1.6E-02			
			REACTOME 1269849	APC/C:Cdc20 mediated degradation of Securin	2.1E-02			

5.8 R-loop formation sequence (RLFS) prediction for candidate lncRNAs

In order to identify possible RLFS within our candidates and strand specific chromosome coordinates of putative lncRNA:DNA hybrids, we used R-loop database (R-lopp DB), with an integrated quantitative model of R-loop forming sequences (QmRLFS) [143]. R-loop DB predicted RLFSs and possible chromosome locations for GATA3-AS1 and for DRAIC (Figure 32 A and B). GATA3-AS1 have both predicted sequences within its own gene and in the neighboring GATA3 gene, while DRAIC shows predicted binding within its own sequence. No RLFSs was predicted for FAM198B-AS1.

A.



B.

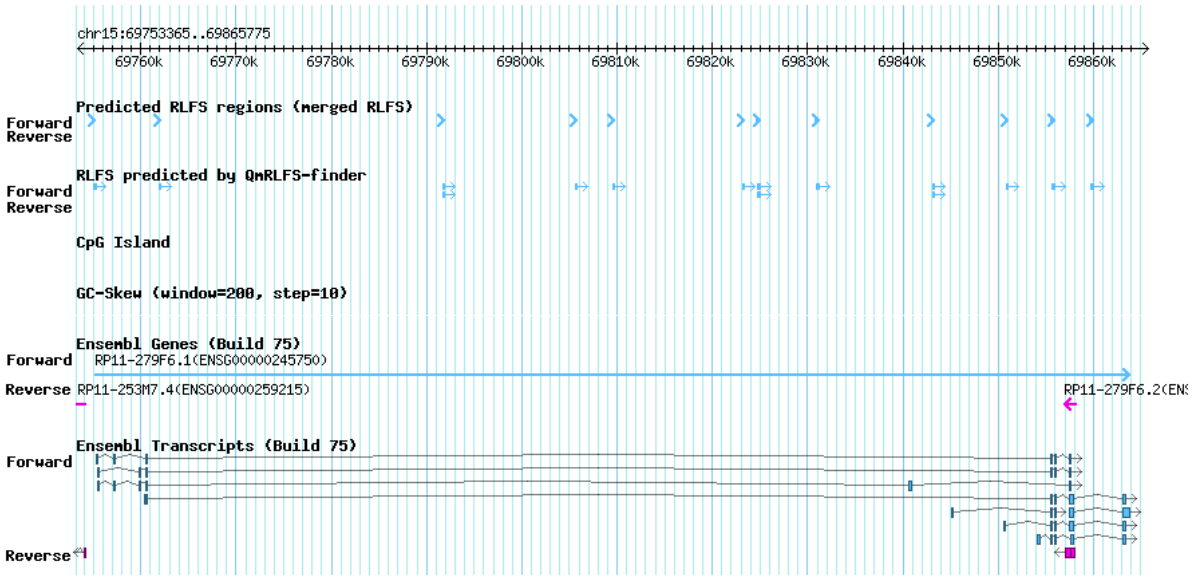


Figure 32. R-loop formation sequence (RLFS) prediction for candidate lncRNAs. **A.** GATA3-AS1. **B.** DRAIC. Blue arrows indicate binding on the forward strand, and pink arrows indicate binding on the reverse strand.

6 Discussion

6.1 Methodological considerations

6.1.1 Expanding annotation databases

While the advent of RNA sequencing technologies drastically has increased the number of identified lncRNAs, precise annotation is a big challenge. Development of different assemblies especially for lncRNA has been started, and Ensembl [105] and Refseq [144] are highly curated and updated databanks for lncRNA together with coding genes. Despite advances in lncRNA annotation, current resources are still incomplete and contain many predicted, but not validated, transcript and gene models, with important consequences for the lncRNA research field. Another challenge is lack of official gene names, making it hard to keep updated on what is published about specific genes. In addition, the coding potential of numerous genes is still unclear, clouding the fundamental differentiation between coding and non-coding RNA. Efforts of manual curation of genes from literature has resulted in datasets such as Lnc2Cancer [145], LncRNADisease [146], the pan-cancer lncRNA co-expression atlas LncMAP [147] and the Mammal ncRNA Disease Repository (MNDR) [148]. In 2012, LNCipedia [43], a database with the aim to collect human lncRNA sequences and annotation was released. The fifth updated version, LNCipedia 5, was released in 2019, and the database now contains 56946 lncRNA genes curated from Ensembl, Refseq and FANTOM CAT [110]. A new filtering pipeline for removing transcripts with coding potential was added, and resulted in a high confidence set, containing 49372 of the 56946 lncRNA genes. Of these, 1555 lncRNA genes are annotated with functional information. LNCipedia's authors claim that while previous releases included large increases in the number of lncRNA gene loci, it appears that the number is stabilizing around 55 000 [43]. This is a resource that could be of great use in the future. We used the annotated lncRNAs from Gencode v27 (15777 lncRNAs), in addition to the lncRNAs in the FANTOM CAT database (8985), a total of 23810 non-overlapping lncRNA loci. This is twice as many lncRNAs as those analyzed by Niknafs et al. [24], where Ensembl v67 (2012) was used (11790 annotated lncRNAs). Illustrating the complex and fast developing nature of the field, Ensembl released its latest update, Ensembl97, on July 3 2019. This version, equivalent to Gencodev31, contains in their own words "major changes" to lncRNA annotations. Several thousand new transcripts have been added as a result of a new pipeline created by a team working for GENCODE called TAGENE [149]. In this version Ensembl announced that they retire all of the previous used terms "non-coding, lincRNA, macro lncRNA, antisense, sense intronic, sense overlapping, 3' overlapping ncRNA, bidirectional promoter lncRNA and retained intron". They will from now on simply classify all these transcripts as "lncRNAs".

The challenge with separating coding and non-coding genes that overlap on the same or opposite strands makes the discovery and annotation of novel lncRNA very complex and

convoluted. To add to the complexity, the definition of lncRNA genes is also under development. The lack of a long open reading frame (lORF), >100 codons, is a part of the definition of lncRNA, but recent studies have reported that some lncRNAs may contain short ORFs (sORF), less than 100 codons, thus producing small peptides [45, 46]. These peptides used to be neglected because of the use of the classical definition of ORFs. Another aspect of this separation is that many protein coding genes can contain transcripts with non-coding capacity [150]. These transcripts could for example be generated by alternative splicing, or alternative transcription start sites. For instance a non-coding mRNA isoform of the coding gene *ASCC3* has been shown to have the opposite effect of the protein coding gene after UV damage [151]. Mounting examples of these phenomena is making the lncRNA field even more complex. We did not include search for sORF in our study, but this could be interesting to do in the future. In our approach, a filtration step was added in the analysis after size selection, keeping only transcripts >200 base pairs. Here, transcripts were filtered keeping only genes overlapping lncRNAs annotated in Gencode v27 in addition to the lncRNAs in the FANTOM CAT database. Then, the CPAT tool was further used to filter out coding transcripts that may have been included due to partial overlap. LncRNAs have generally low expression levels, and they display large differences in expression between cell types and tissue [152]. This could also be seen in the TCGA cohort when ~50000 transcripts were reduced to ~3500 transcripts after we further filtered the data, retaining only lncRNAs expressed in more than 5% of the samples (and FPKM>1). This shows that the vast majority of transcripts were expressed in only a few samples.

6.1.2 The challenges of cross-platform analysis

RNA-sequencing has advantages for the study of non-coding genes, by the possibility for large scale and fast quantification of transcripts. One of the aims of this project was to validate the results from TCGA in two independent cohorts, OSL2 and Metabric, and to do this with the most updated lncRNA annotation at the time of analysis. RNA-seq is a relatively new method, and not all cohorts have this available, making comparison with data based on microarray probes not exact. To be able to compare expression levels, lncRNA IDs had to be matched between the different cohorts. The StringTie-IDs from RNA-seq (TCGA) were generated based on location of transcript sequences in the reference genome. The microarray probes from the Agilent 60k expression array (OSL2) and Illumina HT-12 (Metabric) were matched to the lncRNA transcripts based on genomic location. As the microarray probes in OSL2 and Metabric did not include strand information in the annotation, we observed some incidences where the microarray probes actually were located in the coding gene on the opposite strand of the lncRNA in the initial analysis. Therefore the resulting probes from the differential expression analysis and the emQTL analysis were filtered by blasting the sequences to annotated lncRNAs in the Ensembl93 database, and this time only transcripts with high confidence annotation as non-coding

transcripts were kept. In this filtration we observed that this affected many of our resulting lncRNAs, leaving only 84 lncRNAs in OSL2 among the emQTL results. It is possible that by including only high confidence lncRNAs in the differential expression tests, we could have avoided to correct for coding genes in the results, with the possibility of identifying more differentially expressed lncRNAs. Considering the many possible important roles of lncRNAs in cancer development, these challenges highlight the need for more precise and coordinated lncRNA annotation.

6.1.3 Cell lines

Immortalized cell lines are important model systems for functional studies in cancer. Cell lines can be treated without ethical considerations compared to treatment of patients. Other factors making cell lines valuable tools are little need for space, low costs and the acquired immortality and continuous growth. It must be taken into consideration, however, that genetic drift and mutations can be acquired over time and through many passages. One does not know which passage the cells were at when obtained from manufacturers, and how altered it is from the primary tumor originally isolated from the patient. It is important to keep in mind that a cell culture is a model system, and especially the homogeneity of the cells is different from the natural conditions where tumors are heterogeneous, complex systems where the different cell types influence each other. The gene expression and cellular pathways can change in response to culture conditions, affecting the results. When preparing the knockdown study in this project in the MCF7 cell line, two vials were discharged after observing cell growth uncharacteristic for MCF7, and a third culture had to be established before the experiment could be executed. Finally the cells used in the experiment were proliferating more than normally expected, even when starved, which is not expected for MCF7. This can have affected our results in the cell viability assay.

6.1.4 Cell viability knockdown experiment in MCF7

In order to functionally characterize GATA3-AS1, DRAIC and FAM198B-AS1, we wanted to assess possible driver functions of the three lncRNAs in ER+ breast cancer through a cell viability assay. LNA GapmeRs were used to knock down the genes in the MCF7 cell line. The GapmeRs were chosen to ensure best possible knockdown of the genes, as one does not know which transcripts accounts for the significant difference. Ideally the GapmeR should target the region of the probes from the Agilent array, to ensure that the exact same transcript were knocked down, but this was not possible for technical reasons concerning GapmeR design. Importantly, both GapmeRs and TaqMan probes used for RNA quantification covered the same transcripts as the ones measured by the array probes.

We observed a ~50% knockdown of the three lncRNAs in the MCF7 cell line. It is not expected to achieve a 100% knockout with this form of knockdown experiment, but a significantly lower amount of RNA transcripts. In GapmeR knockdown the gene is not removed. Many transcripts are produced and the probes might not bind to all of them. In contrast, CRISPR-based procedures, where the gene is removed, meaning no transcripts are produced, give a full knockout of the target. When using siRNA normally three versions are tested to choose the one that causes the best knockdown. This was outside the time frame of this project, but optimally we could have tested several GapmeRs per lncRNA. It is possible that the knockdown efficiency could have been further optimised, and that a higher level of knockdown could have influenced the results of the cell viability assay. The reduced cell viability of the cells transfected with the Cell Death positive control showed that the cells had been transfected successfully. There was on the other hand no effect on cell viability when the three lncRNAs were knocked down. There can be several explanations for this. For instance the cells were growing at a faster rate than expected, which resulted in the cells reaching confluence in the first out of three experiments. The confluence could have masked a reduction in proliferation caused by the knockdown, since the proliferation plateaus when cells become confluent. Also, viability was measured after three days, and there could have been an effect in viability detectable after shorter time, but undetectable after three days. The experiment should therefore be repeated with fewer cells and include tracking over time.

Considering these factors together, we cannot definitely conclude that the three candidate lncRNAs do not influence cell viability, even if though we did not detect it in this experiment. This is further strengthened by the documented effect that knockdown of GATA3-AS1 had on cell viability in a study by Liu et al. [153]. Additionally, even though we did not see any effect on cell viability, the knockdown could have had other effects such as drug resistance, migration, or other effects important for cancer progression, that were not included in the frame of this thesis.

6.1.5 Data driven research

The advantages of RNA sequencing technologies have increased the number of identified novel lncRNAs, and is generating an enormous amount of data. At the same time, functional characterization of most lncRNAs remains to be clarified. Bioinformatics has also changed the field of biology dramatically [154]. The pioneers of DNA microarray stated that "Exploration means looking around, observing, describing and mapping undiscovered territory, not testing theories or models. The goal is to discover things we neither knew or expected, and to see relationships and connections among the elements, whether previously suspected or not. It follows that this process is not driven by hypothesis and should be as model-independent as possible» [155]. To find patterns in large amounts of data, many

methods are used, and correlation analysis is a broadly used method in genetics. The method has some limitations, for example are the associations to emQTLs hard to validate, but by observing similar trends across different cohorts one gets an indication that there are important biological associations that can be further investigated. When doing correlation analysis, the chance of finding passenger associations is very high, a correlation in itself does not necessarily tell us about the causality and the inwards relationship of the events in a biological pathway and interconnected pathways. Therefore it is important to follow up with further analysis downstream to try to find the underlying causality, and to formulate a clear and testable hypothesis, and try to construct the experiments in a way that actually tests the hypothesis, excludes possible unconscious or hidden helper hypothesis to better have the chance to find the causality. In this project the starting point was quite explorative, with large data sets and search for correlations between expression, ER status and methylation. Then the selection of candidates and the experiments performed in this thesis was the first steps in finding hypotheses for functions of the candidate lncRNAs that can be further tested experimentally and explored in future studies.

6.2 Biological considerations

6.2.1 Clinical material

In this thesis data from several breast cancer cohorts were analyzed. Different aspects have to be taken into consideration when these data are being used to look for patterns. Different cohorts have collected tumor samples from patients over time and in TCGA, OSL2, and Metabric - also at different hospitals. Protocols for tissue handling, RNA extraction and quantification can differ between institutions and also different time periods in the same lab, and the methods used are not always documented in detail. In addition, criteria for which patients are included in the studies can vary over time, between hospitals and cohorts. Population differences between different countries is another factor that can influence the data. For example TCGA is an American cohort, and contains a larger fraction of African-American patients than the Norwegian cohort OSL2, who are known to have a higher percentage of triple negative breast cancer. The different sizes of cohorts affect the results as well. A larger sample size will result in more statistical power as we see in the higher number of significant associations of lncRNA expression to methylated CpGs in TCGA compared to OSL2. When using tumor tissue it is important to keep in mind that the tissue consists of different cell types, and the data values will represent a mean of the signals from different cells, including for example fibroblasts, immune cells and normal breast epithelial cells together with the cancer cells. lncRNAs are specifically expressed in different cell types and tissue. Also, what part of the resected tumor that is used for different analyses can vary and be difficult to reproduce between patients. In the data sets from the three cohorts we used, the expression data was not scaled in the same way between the three cohorts,

making the values in themselves not directly comparable, although we could observe the similar trends in expression. All of these factors are important to keep in mind, and a full overlap when validating results in an independent cohort should not be expected. However, a strength of this project was the contribution to previous knowledge by including validation across cohorts.

6.2.2 lncRNAs associated with ER status

Research on ER-driven breast cancer has until recently mainly focused on protein-coding genes, such as abnormalities of *ESR1* and related signaling pathways [156, 157]. In recent years studies of breast cancer-associated lncRNAs have been published, for example HOTTIP has been suggested to be involved in cell growth, migration and apoptosis in breast cancer. And several studies have reported different oncogenic functions of the lncRNA *H19* in breast cancer [158], one study found it to be ten fold higher expressed in ER+ than ER- breast cancer [159]. The emerging view of lncRNAs as essential for gene regulation makes it even more interesting to investigate their role in relation to ER status. Niknafs et al. [24] demonstrated that lncRNA expression can be categorized by the known molecular subtypes of breast cancer in the TCGA cohort, and several estrogen regulated lncRNAs have been identified. They report a long list of ER-associated lncRNAs [24] including DSCAM-AS1, an estrogen-driven lncRNA deregulated and mediating tumor progression and tamoxifen resistance in ER+ breast cancer. In our study we saw many of the same trends reported in Niknafs et al.'s paper, but we included approximately the double amount of lncRNAs and approximately 300 more patient samples in our analysis. This in addition to the different lncRNA annotation made it difficult to compare the results directly.

The expression of the significant lncRNAs separated the patients into two clinical groups. The generated heatmap showed a clear 3-part separation of expression levels of lncRNAs, one group highly expressed in almost all patients, while the other groups had low and moderate expression levels in almost all patients. The fold changes of many of the differentially expressed genes were low, which can also be observed by low variation in the heatmap. They were however statistically significant possibly partly explained by the large cohort size of TCGA.

While Niknafs et al's publication is an important contribution, a disadvantage is the lack of validation in independent cohorts - the data is only derived from one discovery cohort, TCGA. Comparing results between cohorts is important all the time the presence and expression pattern of certain lncRNAs can be specific for TCGA. In this study we used two validation cohorts with the aim of identifying the most robust lncRNAs differentially expressed between ER positive and negative tumors. When comparing the results from the t-tests in OSL2 and Metabric and from the stattest in TCGA, all together 354 lncRNAs were

associated with ER status in all the three cohorts, of which 227 were up-regulated in ER+ compared to ER- patients, and 127 were down-regulated. These results represent a robust identification of lncRNAs in the three cohorts. Our results included some new candidates together with lncRNAs previously known to be associated with breast cancer. Examples of lncRNAs showing consistent expression patterns in the two main clinical groups across several cohorts were LINC01116, previously described to regulate ESR1 expression [160], and NRAV previously known to be associated with immune infiltration [161]. Considering the well-established PAM50 classification, numerous gene expression studies focusing on ER status, emerging epigenetic characterization and studies identifying lncRNAs related to ER status, quite much is known about the characteristics of ER+ and ER- tumors. For example in a recent study, Rueda et al. identifies new subgroups of ER+ patients, giving prognostic value in relation to the risk of relapse [162]. But the internal causational relationship of all the different factors in the complex gene regulatory networks and pathways of the ER-driven tumors remains to be elucidated. Our analysis confirms previous observations which shows that many lncRNAs are expressed dependently on ER status, but considering that most of these lncRNAs are functionally uncharacterized, only ~1500 of ~55000 in lncipedia5, this is still a vast area of possible important discoveries. The possibility that the identified lncRNAs have roles that could help elucidate gene regulatory networks, and the possibility to discover targets for therapy among them are important reasons to continue to study the identified lncRNAs bioinformatically and characterize more of them functionally.

6.2.3 Associations to methylation of CpGs

Several lncRNAs have been show to have functions related to DNA methylation and chromatin modification. A known example is *H19*, that is both reported to be regulated by methylation and to regulate many other genes by methylation, as well as being overexpressed in ER+ breast cancer, and being part of an estrogen signaling pathway [158, 159]. Fleischer, Tekpli et al. showed that associations between expression and methylation is important for breast cancer [96], and we have here adapted the same method to quantify the number of associations between DNA methylation levels of CpGs and expression of lncRNAs. The hypothesis here was that if a lncRNA has many associations, it substantiates its involvement in epigenetic regulation. Of our three cohorts only TCGA and OSL2 had available methylation data, and thus Metabric was not part of this analysis. The number of associations varied greatly from 0 to 54550 in TCGA, and 0 to 23059 in OSL2. In TCGA the median was 1517 associations, and in OSL2 the median 450. The lncRNA with most positive correlations in both OSL2 and TCGA was PRKCQ-AS1, previously reported to be an oncogene in a subset of triple negative breast cancer [163]. However, this lncRNA was not considered for functional studies in this thesis because of downregulation in ER+ patients, but could be interesting for future studies. Another lncRNA with many associations and downregulation in ER+ was SOX9-AS1, which has been connected to cancer stem cells and suggested as a

possible therapeutic target [164]. Fleischer, Tekpli et al.'s study showed a large difference in methylation levels between ER+ and ER- tumors. This could have a potential relevance for development of epigenetic drugs. They demonstrated that DNA methylation at enhancers is linked to transcription factor activity and is central for development of ER+ breast cancer. Some of the CpGs with many associations were within transcription factor binding sites of ER, and CpGs in enhancers and binding sites of ER had lost methylation in ER+ breast cancer. Fleischer, Tekpli et al.'s downstream analysis further explained this by demonstrating that some of the associations were overlapping with chromatin loops connecting the enhancers and transcription factors. They conclude that hypomethylation of ER, FOXA1 and GATA3 binding sites is specific for ER+ tumors and may be central for the development of this disease. In the same way the subset of lncRNAs with strong associations to methylation in our results may be associated with the epigenetic events defining clinical subgroups. As the previous study produced a list of gene emQTLs, we now have a list of lncRNA emQTLs. When we overlapped the CpGs associated to GATA3-AS1, FAM198B-AS1 and DRAIC with CpGs in Fleischer, Tekpli et al.'s ER-cluster, we observed 2791, 2499 and 2408 overlapping CpGs respectively. One interesting aspect of this is that we found that the lncRNAs have a much larger number of associations compared to the protein coding genes identified by Fleischer, Tekpli et al. They found that the median number of CpG associations for protein coding genes were ~10, while the median for lincRNAs were <1500 in TCGA. This argues for a special role of lncRNAs in relation to methylation, and emphasize the importance of decoding their functions.

6.2.4 Possible functions of GATA3-AS1

Although the catalog of annotated lncRNAs has grown in recent years, most lncRNAs still have an unknown function. The function of the three candidate lncRNAs in breast cancer is, to our knowledge, not known. However, a previous study by Gibbons et al [140] indicates that GATA3-AS1 controls transcription of GATA3 by recruitment of a methyl-transferase in human PBMC cells. They show that GATA3-AS1 is located primarily in the nucleus, and that the expression level correlates with IL-4, IL-5 and IL-13 in Th2 under T helper cell differentiation. In the study, knockdown of the antisense transcript inhibited expression of GATA3 on the sense strand. Overexpression of the antisense transcript enhanced sense transcription. The intron between exon 2 and 3 of GATA3-AS1 is predicted to have high probability of forming an R loop because of high GC content. An R-loop is a co-transcriptionally formed hybrid, an RNA:DNA duplex and one single-stranded DNA. It is made of one newly formed RNA transcript still attached to its DNA template, and a fragment of a displaced ssDNA [129]. The authors hypothesized that GATA3-AS1 lncRNA may tether to this intronic region on the DNA as an R loop. By immunoprecipitation they found that the lncRNA binds to the histone methylating MLL H3K4-methyltransferase complex and tethers it to the locus [140]. GATA3 plays an important role as a transcription factor in ER+ breast cancer,

contributing to the regulation of genes associated with estrogen dependent tumor growth [141]. Considering this we wanted to investigate whether GATA3-AS1 had the same function of regulating GATA3 in breast cancer cells as it has in Th2 cells. We did observe a significant reduction in GATA3 expression upon GATA3-AS1 knockdown in MCF7 cells, indicating that GATA3-AS1 regulates GATA3 transcription also in this cell line. In our knockdown experiment with GATA3-AS1 we did not observe an effect on cell viability. In a recent study, a CRISPR interference (CRISPRi) platform was developed to target a library of lncRNAs in 7 different cell lines [153]. Interestingly, in contrast to our results, GATA3-AS1 was one of the lncRNAs with the largest negative effect on cell proliferation in MCF7. Due to factors in the cell line experiment discussed above it may be worth repeating this experiment.

Gibbons et al.'s hypothesis of GATA3-AS1's regulation of GATA3 by tethering a methyltransferase in an R-loop formation in Th2 cells [140]. In silico RLFS prediction of the GATA3-AS1 sequence showed binding both within its own gene and in the neighboring GATA3 gene. These predictions and our results showing that GATA3-AS1 regulates GATA3 in MCF7, makes it interesting to conduct future experiments to investigate the effect on methylation of GATA3 when GATA3-AS1 is knocked down.

6.2.5 FAM198B-AS1 is previously uncharacterized

One of our three candidates FAM198B-AS1, is annotated as an antisense lncRNA in Ensembl, but is to our knowledge previously uncharacterized. FAM198B-AS1 was one of the top results after differential expression analysis on ER status, found upregulated in ER+ tumors. The lncRNA has many associations to methylation of CpGs, 15957 associations in TCGA and 13789 in OSL2. It also has a 15 to 4 ratio of presence in the nucleus, making it possible for it to have a role related to chromatin modifications. Knockdown of FAM198B-AS1 had no effect on cell viability, as discussed above, and it would be interesting to use other strategies such as migration-, or apoptotic assays in further functional studies to try to decipher its biological role. The genes most correlated and anticorrelated to FAM198B-AS1 showed the most significant enrichment of pathways of the three candidates. Enriched pathways for genes positively correlated were related to membrane trafficking and cilium assembly pathways. Deregulation of cilium assembly which is important for cell cycle regulation has previously been suggested to be related to cancer characteristics such as migration [142].

6.2.6 Possible functions of DRAIC

DRAIC is reported to be downregulated in many cancers [165], and a previous study by Sakurai et al. identified DRAIC as a tumor suppressor and good prognostic marker in prostate cancer [166]. They report that DRAIC is downregulated in prostate cancer cells when they

lose androgen dependency and that higher levels of DRAIC were associated with longer disease-free survival. By functional experiments they discovered that when androgen receptor (AR) binds to the DRAIC locus it represses DRAIC expression. They also report that FOXA1 and NKX3-1 are recruited to the DRAIC locus and induces DRAIC, and that decrease of FOXA1 and NKX3-1 leads to the decrease of DRAIC during prostate cancer progression, and that DRAIC prevents cellular migration and invasion [166].

In our results DRAIC had many associations to methylation of CpGs; 42119 in TCGA and 8225 in OSL2. Considering its important tumor suppressor function and prognostic value in prostate cancer it would be interesting to elucidate its role in breast cancer further. In silico RLFS prediction of the DRAIC sequence showed binding within its own gene. This could be functionally tested in future experiments. Enriched pathways for genes positively correlated were related to metabolism, specifically that of branched chain amino acids in addition to pathways related to altered insulin secretion. Tumors have been shown to take up branched chain amino acids which can be utilized for protein synthesis or for energy production [167]. It could be interesting to investigate whether DRAIC plays a role in cell metabolism.

7 Conclusion and future perspectives

Today many lncRNAs are reported to be involved in breast cancer as well as epigenetic regulation. In this study we identified differentially expressed lncRNAs between ER+ and ER- breast cancer tumors. We expanded the pool of lncRNAs as well as patient samples included in the analysis compared to previous studies. RNA-seq data from TCGA was prosessed to create a catalog of lncRNAs. We used annotated lncRNAs from Gencode v27, in addition to the lncRNAs in the FANTOM CAT database, twice as many lncRNAs as those analysed by Niknafs et al. The analysis of lncRNAs confirms previous observations that many lncRNAs are expressed dependently on ER status. Previous studies on ER-association also lack validation across cohorts. We confirmed our results in two independent cohorts, and identified lncRNAs with robust differential expression across three cohorts. Among the results were lncRNAs previously known to be associated with breast cancer in addition to novel candidates.

Several studies have reported associations between DNA methylation and breast cancer carcinogenesis. We adapted the emQTL method developed by Fleischer, Tekpli et al. and identified a subset of lncRNAs with strong associations to methylation of CpGs within the group of ER-related lncRNAs. This subset of lncRNAs could be explanatory for the epigenetic events defining the ER-related clinical subgroups. The distribution of significant CpG-lncRNA associations shows that a few lncRNAs account for most of the associations. Also, we observed that lncRNAs have more associations to CpGs compared to protein coding genes.

We selected three candidate lncRNAs; GATA3-AS1, FAM198B-AS1 and DRAIC for functional studies. Our criteria were high expression in ER+ tumors and many associations to methylation of CpGs. In addition we quantified levels of specific lncRNAs in the nuclear and cytosolic fractions of MCF7 to identify lncRNAs with possible functions connected to chromatin structure.

When we performed a knockdown of the candidates, we observed no significant reduction in live cells after 72 hours. However, we show that knockdown of the antisense GATA3-AS1 reduces transcription of the sense GATA3 in MCF7, contributing to similar results reported in Th2 cells. Further experiments should be performed to study the character of the mechanism behind this effect. Future experiments should also include RIP-seq, to assess binding sites of the lncRNA candidates. Also other end points after knockdown and over-expression should be investigated, such as the effect on drug resistance, migration, or other effects important for cancer progression, including the effect on global methylation of CpGs.

References

1. Weinberg, R.A., *Weinberg, Robert A. The Biology of Cancer. 2nd ed. New York: Garland Science, 2014. Print.*
2. Anand, P., et al., *Cancer is a preventable disease that requires major lifestyle changes.* Pharm Res, 2008. **25**(9): p. 2097-116.
3. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
4. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
5. Weinberg, R.A., *The biology of Cancer*. Second edition. ed. 2014, New York: Garland Science, Taylor & Francis Group. xx, 876, A 6, G 30, I 28 pages.
6. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
7. Kornelia Polyak, O.M.F., *SnapShot: Breast Cancer.* Cancer Cell, 2012. **22Issue 4p417-562**.
8. Cancer Registry of Norway. IK Larsen, B.M., TB Johannesen, TE Robsahm, TK Grimsrud, S Larønningen, E Jakobsen, G Ursin, *Cancer in Norway 2017 - Cancer incidence, mortality, survival and prevalence in Norway.* 2018.
9. *American Cancer Society. Key Statistics for Breast Cancer in Men*. 2019; Available from: https://www.cancer.org/cancer/breast-cancer-in-men/about/key-statistics.html.
10. Russo, J. and I.H. Russo, *Development of the human breast.* Maturitas, 2004. **49**(1): p. 2-15.
11. Eric Wong, M.O.B.C. *Breast Cancer*. Available from: http://www.pathophys.org/breast-cancer/.
12. Allred, D.C., *Ductal carcinoma in situ: terminology, classification, and natural history.* J Natl Cancer Inst Monogr, 2010. **2010**(41): p. 134-8.
13. Cowell, C.F., et al., *Progression from ductal carcinoma in situ to invasive breast cancer: revisited.* Mol Oncol, 2013. **7**(5): p. 859-69.
14. Villanueva, H., et al., *The Emerging Roles of Steroid Hormone Receptors in Ductal Carcinoma in Situ (DCIS) of the Breast.* J Mammary Gland Biol Neoplasia, 2018. **23**(4): p. 237-248.
15. Dieci, M.V., et al., *Rare breast cancer subtypes: histological, molecular, and clinical peculiarities.* Oncologist, 2014. **19**(8): p. 805-13.
16. Jogi, A., et al., *Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors.* Ups J Med Sci, 2012. **117**(2): p. 217-24.
17. Koh, J. and M.J. Kim, *Introduction of a New Staging System of Breast Cancer for Radiologists: An Emphasis on the Prognostic Stage.* Korean J Radiol, 2019. **20**(1): p. 69-82.
18. *Diagnostisering og utredning. Utredning og diagnostikk ved påvist invasiv brystkreft*. 2019; Available from: https://www.helsebiblioteket.no/retningslinjer/brystkreft/diagnostisering-og-utredning/utredning-og-diagnostikk/patologidiagnostikk.
19. Deroo, B.J. and K.S. Korach, *Estrogen receptors and human disease.* J Clin Invest, 2006. **116**(3): p. 561-70.
20. Altucci, L., et al., *17beta-Estradiol induces cyclin D1 gene transcription, p36D1-p34cdk4 complex activation and p105Rb phosphorylation during mitogenic stimulation of G(1)-arrested human breast cancer cells.* Oncogene, 1996. **12**(11): p. 2315-24.

21. Tecalco-Cruz, A.C., et al., *Nucleo-cytoplasmic transport of estrogen receptor alpha in breast cancer cells.* Cell Signal, 2017. **34**: p. 121-132.
22. Wang, S., et al., *Genome-Wide Investigation of Genes Regulated by ERalpha in Breast Cancer Cells.* Molecules, 2018. **23**(10).
23. Lumachi, F., D.A. Santeufemia, and S.M. Basso, *Current medical treatment of estrogen receptor-positive breast cancer.* World J Biol Chem, 2015. **6**(3): p. 231-9.
24. Niknafs, Y.S., et al., *The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression.* Nat Commun, 2016. **7**: p. 12791.
25. Van Asten, K., et al., *Prognostic Value of the Progesterone Receptor by Subtype in Patients with Estrogen Receptor-Positive, HER-2 Negative Breast Cancer.* Oncologist, 2019. **24**(2): p. 165-171.
26. Iqbal, N. and N. Iqbal, *Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications.* Mol Biol Int, 2014. **2014**: p. 852748.
27. Slamon, D.J., et al., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.* Science, 1987. **235**(4785): p. 177-82.
28. Wolff, A.C., et al., *American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer.* Arch Pathol Lab Med, 2007. **131**(1): p. 18-43.
29. Urruticoechea, A., I.E. Smith, and M. Dowsett, *Proliferation marker Ki-67 in early breast cancer.* J Clin Oncol, 2005. **23**(28): p. 7212-20.
30. Langerod, A., et al., *TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer.* Breast Cancer Res, 2007. **9**(3): p. R30.
31. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
32. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
33. Yersal, O. and S. Barutca, *Biological subtypes of breast cancer: Prognostic and therapeutic implications.* World J Clin Oncol, 2014. **5**(3): p. 412-24.
34. Kennecke, H., et al., *Metastatic behavior of breast cancer subtypes.* J Clin Oncol, 2010. **28**(20): p. 3271-7.
35. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
36. Norwegian Breast Cancer Group. *Norwegian Breast Cancer Group. Behandlingsskjemaer. Medikamentell behandling*. 2018; Available from: https://nbcg.no/behandlingsskjemaer/medikamentell-behandling/.
37. Amend, K., D. Hicks, and C.B. Ambrosone, *Breast cancer in African-American women: differences in tumor biology from European-American women.* Cancer Res, 2006. **66**(17): p. 8327-30.
38. Gonçalves H Jr, G.M., Duarte Cintra JR, Fayer VA, Brum IV, Bustamante Teixeira MT. , *Survival Study of Triple-Negative and Non–Triple-Negative Breast Cancer in a Brazilian Cohort.* Clin Med Insights Oncol. , 2018.
39. Vikas, P., N. Borcherding, and W. Zhang, *The clinical promise of immunotherapy in triple-negative breast cancer.* Cancer Manag Res, 2018. **10**: p. 6823-6833.
40. Arteaga, C.L., et al., *Treatment of HER2-positive breast cancer: current status and future perspectives.* Nat Rev Clin Oncol, 2011. **9**(1): p. 16-32.
41. Dai, X., et al., *Breast cancer intrinsic subtype classification, clinical use and future trends.* Am J Cancer Res, 2015. **5**(10): p. 2929-43.

42. Fernandes, J.C.R., et al., *Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease.* Noncoding RNA, 2019. **5**(1).
43. Volders, P.J., et al., *LNCipedia 5: towards a reference set of human long non-coding RNAs.* Nucleic Acids Res, 2019. **47**(D1): p. D135-D139.
44. Zhao, R.F. *ENCODE: Deciphering Function in the Human Genome*. 2012.
45. Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**(3): p. 193-204.
46. Landry, C.R., et al., *Found in translation: functions and evolution of a recently discovered alternative proteome.* Curr Opin Struct Biol, 2015. **32**: p. 74-80.
47. Fang, Y. and M.J. Fullwood, *Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer.* Genomics Proteomics Bioinformatics, 2016. **14**(1): p. 42-54.
48. Huarte, M., *The emerging role of lncRNAs in cancer.* Nat Med, 2015. **21**(11): p. 1253-61.
49. Ponting, C.P., P.L. Oliver, and W. Reik, *Evolution and functions of long noncoding RNAs.* Cell, 2009. **136**(4): p. 629-41.
50. Gutschner, T. and S. Diederichs, *The hallmarks of cancer: a long non-coding RNA point of view.* RNA Biol, 2012. **9**(6): p. 703-19.
51. Salviano-Silva, A., et al., *Besides Pathology: Long Non-Coding RNA in Cell and Tissue Homeostasis.* Noncoding RNA, 2018. **4**(1).
52. Chakraborty, S., et al., *LncRBase: an enriched resource for lncRNA information.* PLoS One, 2014. **9**(9): p. e108010.
53. Zinad, H.S., I. Natasya, and A. Werner, *Natural Antisense Transcripts at the Interface between Host Genome and Mobile Genetic Elements.* Front Microbiol, 2017. **8**: p. 2292.
54. Yang, G., X. Lu, and L. Yuan, *LncRNA: a link between RNA and cancer.* Biochim Biophys Acta, 2014. **1839**(11): p. 1097-109.
55. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.
56. Tsai, M.C., et al., *Long noncoding RNA as modular scaffold of histone modification complexes.* Science, 2010. **329**(5992): p. 689-93.
57. Latorre, E., et al., *The Ribonucleic Complex HuR-MALAT1 Represses CD133 Expression and Suppresses Epithelial-Mesenchymal Transition in Breast Cancer.* Cancer Res, 2016. **76**(9): p. 2626-36.
58. Zhang, B., et al., *The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult.* Cell Rep, 2012. **2**(1): p. 111-23.
59. Arun, G. and D.L. Spector, *MALAT1 long non-coding RNA and breast cancer.* RNA Biol, 2019. **16**(6): p. 860-863.
60. Xing, Z., et al., *lncRNA directs cooperative epigenetic regulation downstream of chemokine signals.* Cell, 2014. **159**(5): p. 1110-1125.
61. Godinho, M.F., et al., *Relevance of BCAR4 in tamoxifen resistance and tumour aggressiveness of human breast cancer.* Br J Cancer, 2010. **103**(8): p. 1284-91.
62. Wang, J., et al., *Dysregulation of long non-coding RNA in breast cancer: an overview of mechanism and clinical implication.* Oncotarget, 2017. **8**(3): p. 5508-5522.
63. Ma, C., et al., *The growth arrest-specific transcript 5 (GAS5): a pivotal tumor suppressor long noncoding RNA in human cancers.* Tumour Biol, 2016. **37**(2): p. 1437-44.
64. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Sci Signal, 2010. **3**(107): p. ra8.
65. Zhang, Z., et al., *Negative regulation of lncRNA GAS5 by miR-21.* Cell Death Differ, 2013. **20**(11): p. 1558-68.

66. Plath, K., et al., *Xist RNA and the mechanism of X chromosome inactivation.* Annu Rev Genet, 2002. **36**: p. 233-78.
67. Brockdorff, N., *Polycomb complexes in X chromosome inactivation.* Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1733).
68. Pageau, G.J., et al., *The disappearing Barr body in breast and ovarian cancers.* Nat Rev Cancer, 2007. **7**(8): p. 628-33.
69. McHugh, C.A., et al., *The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3.* Nature, 2015. **521**(7551): p. 232-6.
70. Waddington, C.H., *The epigenotype. 1942.* Int J Epidemiol, 2012. **41**(1): p. 10-3.
71. Tronick, E. and R.G. Hunter, *Waddington, Dynamic Systems, and Epigenetics.* Front Behav Neurosci, 2016. **10**: p. 107.
72. V.E.A. Russo, R.A.M., A.D. Riggs, *Epigenetic Mechanisms of Gene Regulation*. 1996, Plainview, New York: Cold Spring Harbor Laboratory Press.
73. Qiu, J., *Epigenetics: unfinished symphony.* Nature, 2006. **441**(7090): p. 143-5.
74. Jin, B., Y. Li, and K.D. Robertson, *DNA methylation: superior or subordinate in the epigenetic hierarchy?* Genes Cancer, 2011. **2**(6): p. 607-17.
75. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development.* Nat Rev Genet, 2013. **14**(3): p. 204-20.
76. Greer, E.L. and Y. Shi, *Histone methylation: a dynamic mark in health, disease and inheritance.* Nat Rev Genet, 2012. **13**(5): p. 343-57.
77. Shi, X., et al., *ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression.* Nature, 2006. **442**(7098): p. 96-9.
78. Fuks, F., *DNA methylation and histone modifications: teaming up to silence genes.* Curr Opin Genet Dev, 2005. **15**(5): p. 490-5.
79. Zhong, J., G. Agha, and A.A. Baccarelli, *The Role of DNA Methylation in Cardiovascular Risk and Disease: Methodological Aspects, Study Design, and Data Analysis for Epidemiological Studies.* Circ Res, 2016. **118**(1): p. 119-131.
80. Jenuwein, T. and C.D. Allis, *Translating the histone code.* Science, 2001. **293**(5532): p. 1074-80.
81. Luger, K., M.L. Dechassa, and D.J. Tremethick, *New insights into nucleosome and chromatin structure: an ordered state or a disordered affair?* Nat Rev Mol Cell Biol, 2012. **13**(7): p. 436-47.
82. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer, 2004. **4**(2): p. 143-53.
83. van Hoesel, A.Q., et al., *Assessment of DNA methylation status in early stages of breast cancer development.* Br J Cancer, 2013. **108**(10): p. 2033-8.
84. Brooks, J., P. Cairns, and A. Zeleniuch-Jacquotte, *Promoter methylation and the detection of breast cancer.* Cancer Causes Control, 2009. **20**(9): p. 1539-50.
85. Sharma, G., et al., *CpG hypomethylation of MDR1 gene in tumor and serum of invasive ductal breast carcinoma patients.* Clin Biochem, 2010. **43**(4-5): p. 373-9.
86. Ronneberg, J.A., et al., *Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer.* Mol Oncol, 2011. **5**(1): p. 61-76.
87. Jeronimo, C., et al., *Quantitative hypermethylation of a small panel of genes augments the diagnostic accuracy in fine-needle aspirate washings of breast lesions.* Breast Cancer Res Treat, 2008. **109**(1): p. 27-34.
88. Tao, M.H., et al., *DNA hypermethylation and clinicopathological features in breast cancer: the Western New York Exposures and Breast Cancer (WEB) Study.* Breast Cancer Res Treat, 2009. **114**(3): p. 559-68.

89. Azzollini J, P.C., Pizzamiglio S, et al. , *Constitutive BRCA1 Promoter Hypermethylation Can Be a Predisposing Event in Isolated Early-Onset Breast Cancer.* Cancers (Basel), 2019. **2019 Jan; 11(1): 58.**
90. Esteller, M., *Epigenetics in cancer.* N Engl J Med, 2008. **358**(11): p. 1148-59.
91. Karpf, A.R. and S. Matsui, *Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells.* Cancer Res, 2005. **65**(19): p. 8635-9.
92. Eden, A., et al., *Chromosomal instability and tumors promoted by DNA hypomethylation.* Science, 2003. **300**(5618): p. 455.
93. Holm, K., et al., *Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns.* Breast Cancer Res, 2010. **12**(3): p. R36.
94. Bediaga, N.G., et al., *DNA methylation epigenotypes in breast cancer molecular subtypes.* Breast Cancer Res, 2010. **12**(5): p. R77.
95. Fleischer, T., et al., *DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival.* Oncotarget, 2017. **8**(1): p. 1074-1082.
96. Fleischer, T., et al., *DNA methylation at enhancers identifies distinct breast cancer lineages.* Nat Commun, 2017. **8**(1): p. 1379.
97. Achour, C. and F. Aguilo, *Long non-coding RNA and Polycomb: an intricate partnership in cancer biology.* Front Biosci (Landmark Ed), 2018. **23**: p. 2106-2132.
98. Morlando, M. and A. Fatica, *Alteration of Epigenetic Regulation by Long Noncoding RNAs in Cancer.* Int J Mol Sci, 2018. **19**(2).
99. Laugesen, A., J.W. Hojfeldt, and K. Helin, *Role of the Polycomb Repressive Complex 2 (PRC2) in Transcriptional Regulation and Cancer.* Cold Spring Harb Perspect Med, 2016. **6**(9).
100. Tsai MC, M.O., Wan Y, et al. , *Long noncoding RNA as modular scaffold of histone modification complexes.* Science, 2010. **2010;329(5992):689–693. doi:10.1126/science.1192002**.
101. Vennin, C., et al., *The long non-coding RNA 91H increases aggressive phenotype of breast cancer cells and up-regulates H19/IGF2 expression through epigenetic modifications.* Cancer Lett, 2017. **385**: p. 198-206.
102. Yamanaka, Y., et al., *Aberrant overexpression of microRNAs activate AKT signaling via down-regulation of tumor suppressors in natural killer-cell lymphoma/leukemia.* Blood, 2009. **114**(15): p. 3265-3275.
103. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.
104. Aure, M.R., et al., *Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome.* Breast Cancer Res, 2017. **19**(1): p. 44.
105. Zerbino, D.R., et al., *Ensembl 2018.* Nucleic Acids Res, 2018. **46**(D1): p. D754-D761.
106. *UCSC genome browser*. Available from: http://genome.ucsc.edu/.
107. Roush, S. and F.J. Slack, *The let-7 family of microRNAs.* Trends in Cell Biology, 2008. **18**(10): p. 505-516.
108. Qian, P., et al., *Pivotal Role of Reduced let-7g Expression in Breast Cancer Invasion and Metastasis.* Cancer Research, 2011. **71**(20): p. 6463-6474.
109. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
110. Hon, C.C., et al., *An atlas of human long non-coding RNAs with accurate 5′ ends.* Nature, 2017. **543**(7644): p. 199-204.
111. R Core Team. *R: A language and environment for statistical computing.*; Available from: https://www.R-project.org/.

112. RStudio Team. *RStudio: Integrated Development for R.* . Available from: http://www.rstudio.com/.
113. Mardis, E.R., *Next-generation sequencing platforms.* Annu Rev Anal Chem (Palo Alto Calif), 2013. **6**: p. 287-303.
114. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. **513**(7517): p. 202-9.
115. A., C., *Overview of Next-Generation Sequencing Technologies and Its Application in Chemical Biology. In: Advancing Development of Synthetic Gene Regulators.* 2018, Springer, Singapore: Springer Theses.
116. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nat Protoc, 2016. **11**(9): p. 1650-67.
117. Wang, L., et al., *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.* Nucleic Acids Res, 2013. **41**(6): p. e74.
118. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
119. Hurd, P.J. and C.J. Nelson, *Advantages of next-generation sequencing versus the microarray in epigenetic research.* Brief Funct Genomic Proteomic, 2009. **8**(3): p. 174-83.
120. Li, Y. and T.O. Tollefsbol, *DNA methylation detection: bisulfite genomic sequencing analysis.* Methods Mol Biol, 2011. **791**: p. 11-21.
121. Illumina. *Infinium Methylation Assay Overview*. Available from: https://emea.illumina.com/science/technology/beadarray-technology/infinium-methylation-assay.html?langsel=/no/.
122. *Epigentek. Bisulfite conversion* Available from: https://www.epigentek.com/catalog/dna-bisulfite-conversion-c-75_21_47.html.
123. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
124. Parkhomchuk, D., et al., *Transcriptome analysis by strand-specific sequencing of complementary DNA.* Nucleic Acids Res, 2009. **37**(18): p. e123.
125. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.
126. McDonald, J.H., *Handbook of Biological Statistics*. Vol. 3. 2014, Baltimore, Maryland: Sparky House Publishing.
127. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
128. Chen, J., et al., *ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W305-11.
129. White, R.L. and D.S. Hogness, *R loop mapping of the 18S and 28S sequences in the long and short repeating units of Drosophila melanogaster rDNA.* Cell, 1977. **10**(2): p. 177-92.
130. Ratmeyer, L., et al., *Sequence specific thermodynamic and structural properties for DNA.RNA duplexes.* Biochemistry, 1994. **33**(17): p. 5298-304.
131. Kuznetsov, V.A., et al., *Toward predictive R-loop computational biology: genome-scale prediction of R-loops reveals their association with complex promoter structures, G-quadruplexes and transcriptionally active enhancers.* Nucleic Acids Res, 2018. **46**(15): p. 8023.
132. Enerly, E., et al., *miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors.* PLoS ONE, 2011. **6**(2): p. e16915.

133. Soule, H.D., et al., *A human cell line from a pleural effusion derived from a breast carcinoma.* J Natl Cancer Inst, 1973. **51**(5): p. 1409-16.
134. Qiagen, *PROM-12604-001_RNA_Functional_Analysis_Brochure*
. 2018.
135. *Antisense LNA® GapmeRs Handbook. LNA-optimized oligonucleotides for strandspecific*
knockdown of mRNA and lncRNA. 2017: Qiagen.
136. *CellTiter-Glo® Luminescent Cell Viability Assay handbook*. 2015, Madison, WI: Promega.
137. *Qiagen*. Available from: https://www.qiagen.com/no/.
138. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR.* Nucleic Acids Res, 2001. **29**(9): p. e45.
139. Jiang, C., et al., *Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs.* Oncotarget, 2016. **7**(6): p. 7120-33.
140. Gibbons, H.R., et al., *Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells.* Front Immunol, 2018. **9**: p. 2512.
141. Theodorou, V., et al., *GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility.* Genome Res, 2013. **23**(1): p. 12-22.
142. Fabbri, L., F. Bost, and N.M. Mazure, *Primary Cilium in Cancer Hallmarks.* Int J Mol Sci, 2019. **20**(6).
143. Kuznetsov, V.A., et al., *Toward predictive R-loop computational biology: genome-scale prediction of R-loops reveals their association with complex promoter structures, G-quadruplexes and transcriptionally active enhancers.* Nucleic Acids Res, 2018. **46**(15): p. 7566-7585.
144. *RefSeq*. Available from: https://www.ncbi.nlm.nih.gov/refseq/.
145. Ning, S., et al., *Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers.* Nucleic Acids Res, 2016. **44**(D1): p. D980-5.
146. Chen, G., et al., *LncRNADisease: a database for long-non-coding RNA-associated diseases.* Nucleic Acids Res, 2013. **41**(Database issue): p. D983-6.
147. Li, Y., et al., *LncMAP: Pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations.* Nucleic Acids Res, 2018. **46**(3): p. 1113-1123.
148. Wang, Y., et al., *Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network.* Cell Death Dis, 2013. **4**: p. e765.
149. *Ensembl Blog. What's coming in Ensembl 97 & Ensembl Genomes 44*. 2019; Available from: http://www.ensembl.info/2019/05/20/whats-coming-in-ensembl-97-ensembl-genomes-44/.
150. Dhamija, S. and M.B. Menon, *Non-coding transcript variants of protein-coding genes - what are they good for?* RNA Biol, 2018. **15**(8): p. 1025-1031.
151. Williamson, L., et al., *UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene.* Cell, 2017. **168**(5): p. 843-855 e13.
152. Chen, L., et al., *Tissue Expression Difference between mRNAs and lncRNAs.* Int J Mol Sci, 2018. **19**(11).
153. Liu, S.J., et al., *CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells.* Science, 2017. **355**(6320).
154. Mazzocchi, F., *Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science.* EMBO Rep, 2015. **16**(10): p. 1250-5.
155. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 33-7.

156. Toy, W., et al., *ESR1 ligand-binding domain mutations in hormone-resistant breast cancer.* Nat Genet, 2013. **45**(12): p. 1439-45.
157. Robinson, D.R., et al., *Activating ESR1 mutations in hormone-resistant metastatic breast cancer.* Nat Genet, 2013. **45**(12): p. 1446-51.
158. Berteaux, N., et al., *H19 mRNA-like noncoding RNA promotes breast cancer cell proliferation through positive control by E2F1.* J Biol Chem, 2005. **280**(33): p. 29625-36.
159. Sun, H., et al., *H19 lncRNA mediates 17beta-estradiol-induced cell proliferation in MCF-7 breast cancer cells.* Oncol Rep, 2015. **33**(6): p. 3045-52.
160. Hu, H.B., Q. Chen, and S.Q. Ding, *LncRNA LINC01116 competes with miR-145 for the regulation of ESR1 expression in breast cancer.* Eur Rev Med Pharmacol Sci, 2018. **22**(7): p. 1987-1993.
161. Ouyang, J., et al., *NRAV, a long noncoding RNA, modulates antiviral responses through suppression of interferon-stimulated gene transcription.* Cell Host Microbe, 2014. **16**(5): p. 616-26.
162. Rueda, O.M., et al., *Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups.* Nature, 2019. **567**(7748): p. 399-404.
163. Byerly, J., et al., *PRKCQ promotes oncogenic growth and anoikis resistance of a subset of triple-negative breast cancer cells.* Breast Cancer Res, 2016. **18**(1): p. 95.
164. Domenici, G., et al., *A Sox2-Sox9 signalling axis maintains human breast luminal progenitor and breast cancer stem cells.* Oncogene, 2019. **38**(17): p. 3151-3169.
165. Zhao, D. and J.T. Dong, *Upregulation of Long Non-Coding RNA DRAIC Correlates with Adverse Features of Breast Cancer.* Noncoding RNA, 2018. **4**(4).
166. Sakurai, K., et al., *The lncRNA DRAIC/PCAT29 Locus Constitutes a Tumor-Suppressive Nexus.* Mol Cancer Res, 2015. **13**(5): p. 828-38.
167. Ananieva, E., *Targeting amino acid metabolism in cancer growth and anti-tumor immune response.* World J Biol Chem, 2015. **6**(4): p. 281-9.

Supplementary data

1. Supplementary table 1
2. Supplementary table 2
3. Supplementary table 3
4. Supplementary table 4

Supplementary table 1

id	Genename	gencode_gene_id	fantom_gene_id	fc TCGA	qval TCGA	probe.id METABRIC	qval.ER METABRIC	probe.id OSL2	qval.ER OSL2	frquency CpG associations TCGA	frquency CpG assotiations OSL2 (probe with most ass)	blast of probe 100% match to ncRNA
MSTRG.8487	PRKCQ-AS1	ENSG00000237943.6	ENSG00000237943.2	0.81	0.0E+00	ILMN_1893633	9.3E-48	A_32_P232559	1.2E-07	54391	25849	yes
MSTRG.66629	MIR3142HG	ENSG00000253522.5	NA	0.81	0.0E+00	ILMN_1882512	2.1E-17	A_33_P3718269	1.5E-07	1	22493	yes
MSTRG.27835	FAM30A	ENSG00000279495.1,ENSG00000244620.1,I	CATG00000020298.1,ENSG00000244620.1	0.56	0.0E+00	ILMN_1656920	1.3E-86	A_33_P3331193	1.6E-06	47331	20942	yes
MSTRG.62195	FAM198B-AS1	ENSG00000248429.5	ENSG00000248429.1	1.25	0.0E+00	ILMN_1708110	3.5E-113	A_33_P3304372	2.5E-17	15957	19376	yes
MSTRG.29187	ZNF213-AS1	ENSG00000263072.6	ENSG00000263072.1	1.12	1.7E-11	ILMN_1676577	1.2E-03	A_23_P432448	1.7E-17	2066	12863	yes
MSTRG.63950	AC093297.1	ENSG00000248779.1,ENSG00000272335.1	NA	2.02	0.0E+00	ILMN_1726743	2.5E-151	A_19_P00803393	5.5E-26	853	12747	yes
MSTRG.8508	GATA3-AS1	ENSG00000197308.9	ENSG00000243350.1,ENSG00000197308.4	1.29	0.0E+00	ILMN_1679767	8.3E-55	A_33_P3227269	1.8E-31	23166	12093	yes
MSTRG.42704	RMDN2-AS1	ENSG00000235848.4,ENSG00000231367.5,I	ENSG00000231367.1,CATG00000048152.1	0.81	0.0E+00	ILMN_1716384	3.5E-67	A_33_P3394040	1.3E-12	1215	11411	yes
MSTRG.28009	DRAIC	ENSG00000245750.7,ENSG00000259641.5	ENSG00000245750.3	1.90	0.0E+00	ILMN_1688228	7.1E-79	A_19_P00322780	1.4E-22	42119	7538	yes
MSTRG.65778	AC104109.4	ENSG00000273345.5	ENSG00000273345.1	1.07	6.0E-02	ILMN_2196097	1.7E-75	A_23_P110643	7.4E-10	279	7065	yes
MSTRG.80598	LINC01235	ENSG00000270547.5	ENSG00000270547.1	0.82	1.4E-13	ILMN_1694458	4.0E-15	A_33_P3266739	3.1E-06	7326	6797	yes
MSTRG.27196	EIF3J-AS1	ENSG00000179523.4	ENSG00000179523.4	1.28	0.0E+00	ILMN_1689749	2.7E-56	A_32_P22501	3.6E-13	10255	6257	yes
MSTRG.32915	AC015922.4	ENSG00000276855.1	NA	1.46	0.0E+00	ILMN_2205896	3.7E-134	A_32_P45168	2.4E-24	22022	4997	yes
MSTRG.29424	AC093525.6	ENSG00000261613.2,ENSG00000260176.1,I	ENSG00000260565.2	1.21	0.0E+00	ILMN_1653793	2.2E-09	A_33_P3354796	1.3E-05	6750	3902	yes
MSTRG.70920	AL590617.2	ENSG00000225177.5	ENSG00000225177.1	0.83	0.0E+00	ILMN_1773553	1.6E-56	A_33_P3340490	1.4E-12	9231	3889	yes
MSTRG.65600	LINC01184	ENSG00000245937.7	ENSG00000245937.3	1.35	0.0E+00	ILMN_1665571	1.5E-47	A_33_P3313830	2.4E-11	19630	3777	yes
MSTRG.31093	CRNDE	ENSG00000245694.9,ENSG00000259725.1	ENSG00000261171.1,ENSG00000245694.4	1.28	0.0E+00	ILMN_1671800	1.3E-27	A_19_P00322533	1.2E-07	25995	3746	yes
MSTRG.23708	DHRS4-AS1	ENSG00000215256.3	ENSG00000215256.3	1.15	9.2E-04	ILMN_1800465	5.8E-21	A_32_P118250	3.8E-10	154	3745	yes
MSTRG.58658	AC093323.1	ENSG00000170846.16	ENSG00000170846.11	1.12	3.7E-04	ILMN_1797181	2.1E-73	A_19_P00317789	1.4E-17	8715	3600	yes
MSTRG.18039	AC008124.1	ENSG00000273015.2	NA	1.36	0.0E+00	ILMN_1780773	5.2E-66	A_19_P00316861	9.6E-12	25319	2664	yes
MSTRG.35724	LINC02095	ENSG00000228639.6,ENSG00000234899.9	ENSG00000234899.5	0.51	0.0E+00	ILMN_1876620	5.3E-28	A_24_P497464	3.3E-07	43783	2450	yes
MSTRG.69453	AL589935.1	ENSG00000232295.7,ENSG00000269966.1,I	ENSG00000232295.3,ENSG00000233237.2	1.25	0.0E+00	ILMN_1666617	1.2E-45	A_19_P00319404	3.7E-08	3411	2393	yes
MSTRG.46256	LINC01116	ENSG00000163364.9	ENSG00000163364.5	2.34	0.0E+00	ILMN_1674782	5.8E-108	A_23_P302787	7.2E-25	25710	2390	yes
MSTRG.83232	LINC01503	ENSG00000233901.5,ENSG00000179082.3	ENSG00000233901.1	0.77	0.0E+00	ILMN_1777477	7.6E-33	A_19_P00322442	2.3E-08	8795	2383	yes
MSTRG.27981	AC027088.1	ENSG00000137808.12,ENSG00000259222.2	ENSG00000259222.1,ENSG00000212766.5	1.36	0.0E+00	ILMN_1665876	2.9E-139	A_19_P00322662	5.0E-27	15257	2306	yes
MSTRG.23729	AL096870.2	ENSG00000260669.2,ENSG00000278784.1	ENSG00000260669.2	1.21	0.0E+00	ILMN_1662263	1.1E-98	A_23_P99811	2.0E-13	2056	2212	yes
MSTRG.25825	MEG3	ENSG00000214548.14,ENSG00000258399.7	ENSG00000214548.10,ENSG00000225746.	1.05	2.3E-02	ILMN_1731673	1.6E-02	A_19_P00805614	3.3E-02	669	2108	yes
MSTRG.2734	AL031429.2	ENSG00000269933.1	NA	1.45	0.0E+00	ILMN_1677384	5.0E-117	A_33_P3265744	7.5E-23	13695	2018	yes
MSTRG.37031	PCAT18	ENSG00000265369.3,ENSG00000277534.1,I	ENSG00000265369.2	1.42	0.0E+00	ILMN_1759984	5.3E-03	A_33_P3219095	2.2E-18	3127	1603	yes
MSTRG.7024	LINC00467	ENSG00000153363.12	ENSG00000153363.8	1.10	7.8E-06	ILMN_1740493	6.3E-74	A_23_P201731	3.7E-10	9799	1554	yes
MSTRG.73116	AC004847.1	ENSG00000260997.1	ENSG00000232956.4	0.91	5.1E-08	ILMN_1692295	1.9E-42	A_33_P3734384	7.9E-03	43307	1452	yes
MSTRG.21727	AL138963.3	ENSG00000273149.1,ENSG00000228886.1,I	ENSG00000273149.1,ENSG00000170919.1	0.93	1.1E-05	ILMN_1800276	4.8E-26	A_24_P110914	7.0E-04	3	1374	yes
MSTRG.3902	ATP1A1-AS1	ENSG00000203865.9	ENSG00000203865.5	0.89	1.3E-04	ILMN_2415189	2.6E-21	A_33_P3229215	1.5E-11	27	1333	yes
MSTRG.51174	LINC00160	ENSG00000230978.1,ENSG00000230794.1	NA	1.16	4.4E-09	ILMN_1801504	2.1E-26	A_24_P96403	7.3E-09	9758	1228	yes
MSTRG.22552	MIR17HG	ENSG00000215417.12	NA	0.85	0.0E+00	ILMN_1660754	1.9E-15	A_33_P3313622	9.2E-07	6915	956	yes
MSTRG.13027	LINC00294	ENSG00000280798.1	NA	0.84	1.2E-12	ILMN_1713174	5.9E-25	A_33_P3209869	5.3E-06	480	901	yes
MSTRG.65321	EPB41L4A-AS1	ENSG00000224032.6,ENSG00000251187.5	ENSG00000224032.2	1.35	0.0E+00	ILMN_1654609	1.3E-45	A_23_P58538	7.4E-07	451	865	yes
MSTRG.18518	AC012531.1	ENSG00000260597.1,ENSG00000249388.1,I	ENSG00000260597.1,ENSG00000248265.1	1.75	0.0E+00	ILMN_1831106	3.4E-57	A_32_P169406	3.1E-11	4487	698	yes
MSTRG.74150	TP53TG1	ENSG00000182165.17	ENSG00000182165.13	1.69	0.0E+00	ILMN_1658469	1.2E-31	A_33_P3250963	1.2E-10	23991	615	yes
MSTRG.86626	XIST	ENSG00000229807.10	ENSG00000229807.5	1.48	7.2E-13	ILMN_1764573	1.7E-29	A_19_P00331623	2.9E-06	4200	515	yes
MSTRG.87943	LINC00893	ENSG00000241769.7,ENSG00000238039.1	NA	1.10	2.9E-07	ILMN_1804775	6.1E-26	A_19_P00320989	4.3E-09	336	505	yes
MSTRG.36184	AC124319.4	ENSG00000280248.1,ENSG00000262979.1	NA	1.09	2.1E-05	ILMN_2289093	2.5E-09	A_19_P00315499	1.1E-07	638	488	yes
MSTRG.49344	SNHG11	ENSG00000174365.19	ENSG00000174365.15	1.16	5.5E-08	ILMN_1791884	1.5E-48	A_23_P325119	1.2E-07	16	475	yes
MSTRG.11515	CASC2	ENSG00000177640.15	NA	1.16	0.0E+00	ILMN_1808546	3.7E-05	A_33_P3213362	3.5E-10	30639	452	yes
MSTRG.75646	AC004918.3	ENSG00000270157.1,ENSG00000244701.1,I	ENSG00000228775.3	0.90	1.0E-04	ILMN_1772645	2.5E-21	A_23_P309224	8.3E-05	8	448	yes
MSTRG.9367	LINC00839	ENSG00000185904.11	ENSG00000185904.7	0.80	1.7E-15	ILMN_1759186	2.5E-10	A_23_P63736	9.1E-05	45	440	yes
MSTRG.16629	AP000844.2	ENSG00000238117.1	ENSG00000238117.1	1.14	1.2E-10	ILMN_1882519	6.7E-29	A_33_P3347477	4.2E-09	2356	370	yes
MSTRG.79818	PVT1	ENSG00000249859.9	ENSG00000249859.3	1.07	2.7E-03	ILMN_1732926	3.8E-04	A_19_P00317900	3.6E-07	292	337	yes
MSTRG.133	TNFRSF14-AS1	ENSG00000238164.6	ENSG00000238164.2	1.28	0.0E+00	ILMN_1832920	4.7E-06	A_33_P3415247	5.7E-13	3561	325	yes
MSTRG.26439	AC090983.1	ENSG00000280118.1,ENSG00000271347.1,I	ENSG00000224078.8,ENSG00000272846.1	1.21	1.0E-13	ILMN_2372082	1.3E-20	A_19_P00321504	1.0E-06	992	313	yes
MSTRG.15746	AP001830.1	ENSG00000255337.1	ENSG00000255337.1	0.87	1.4E-13	ILMN_1827283	7.2E-12	A_32_P110016	2.7E-06	1649	280	yes
MSTRG.11257	AL133355.1	ENSG00000260461.1	NA	1.37	0.0E+00	ILMN_1906158	3.0E-20	A_24_P759674	3.2E-07	18527	270	yes
MSTRG.29854	MIR193BHG	ENSG00000262454.3	NA	1.19	7.9E-15	ILMN_1859127	1.1E-08	A_32_P226646	1.8E-03	2286	209	yes
MSTRG.20495	LINC01089	ENSG00000212694.8	ENSG00000212694.4	1.24	3.7E-06	ILMN_2098947	5.2E-37	A_23_P116743	7.4E-09	18	204	yes
MSTRG.46073	AC007405.3	ENSG00000239467.5	ENSG00000239467.1	1.12	3.7E-09	ILMN_1775584	3.1E-33	A_32_P197340	7.5E-07	1381	182	yes
MSTRG.7782	LINC01132	ENSG00000227630.3	NA	0.92	2.9E-05	ILMN_1902123	9.8E-09	A_19_P00316324	1.0E-02	626	135	yes
MSTRG.24528	KTN1-AS1	ENSG00000186615.10	ENSG00000186615.6	1.20	0.0E+00	ILMN_1779832	3.9E-36	A_19_P00808704	1.4E-04	16034	119	yes
MSTRG.48092	LINC01237	ENSG00000233806.7,ENSG00000237940.3,I	ENSG00000233806.3	1.22	0.0E+00	ILMN_1656266	5.3E-17	A_19_P00322212	2.0E-08	4042	96	yes
MSTRG.44682	AC017002.1	ENSG00000224959.1,ENSG00000172965.14	ENSG00000172965.10	0.92	4.0E-05	ILMN_2143795	4.2E-28	A_19_P00319050	1.9E-03	777	91	yes
MSTRG.58816	AC005674.2	ENSG00000261490.1	ENSG00000261490.1	0.96	7.3E-03	ILMN_1675844	3.1E-03	A_23_P213000	2.9E-03	13474	85	yes
MSTRG.51709	AP000525.1	ENSG00000272872.1,ENSG00000206195.10	NA	0.79	4.4E-12	ILMN_1729197	2.4E-33	A_33_P3349455	1.7E-10	1679	82	yes
MSTRG.39279	BISPR	ENSG00000282851.1	ENSG00000269640.1	1.38	6.0E-16	ILMN_2121207	5.2E-19	A_23_P393758	3.5E-03	2080	60	yes
MSTRG.12163	H19	ENSG00000130600.17	ENSG00000130600.11	1.23	1.1E-02	ILMN_2148527	1.8E-02	A_19_P00323082	2.7E-02	1118	52	yes
MSTRG.42791	AC007388.1	ENSG00000231312.6	NA	1.06	1.9E-02	ILMN_1864059	4.5E-03	A_33_P3405285	5.8E-04	4	52	yes
MSTRG.66332	AC021078.1	ENSG00000230551.4	NA	1.57	0.0E+00	ILMN_1785988	1.2E-44	A_24_P183292	2.6E-05	41374	51	yes
MSTRG.55783	DUBR	ENSG00000243701.5	ENSG00000243701.1	1.42	0.0E+00	ILMN_1696701	2.4E-87	A_19_P00315753	3.1E-11	4460	49	yes
MSTRG.86647	JPX	ENSG00000225470.7,ENSG00000228906.1	ENSG00000225470.2	1.10	5.9E-05	ILMN_1704431	2.6E-27	A_19_P00324814	9.2E-08	233	49	yes
MSTRG.19663	SOCS2-AS1	ENSG00000246985.7	ENSG00000246985.3	1.14	2.4E-08	ILMN_1699188	8.5E-11	A_33_P3375002	2.8E-04	244	33	yes
MSTRG.5274	RUSC1-AS1	ENSG00000225855.6	ENSG00000225855.2	0.88	3.6E-08	ILMN_1725906	4.2E-61	A_24_P114183	5.3E-08	56	28	yes
MSTRG.21178	NA	NA	ENSG00000205861.7	1.19	4.3E-14	ILMN_1713038	2.3E-18	A_33_P3531206	1.7E-03	3529	26	yes
MSTRG.10203	ZNF503-AS2	ENSG00000237149.5	ENSG00000272692.1	0.90	4.3E-07	ILMN_1660288	3.4E-06	A_33_P3331641	1.5E-04	7990	19	yes
MSTRG.29434	SRRM2-AS1	ENSG00000205913.6	ENSG00000205913.2	1.22	0.0E+00	ILMN_1894057	2.6E-108	A_33_P3324755	1.0E-05	8832	7	yes
MSTRG.14165	SNHG1	ENSG00000255717.6	ENSG00000255717.2	0.71	0.0E+00	ILMN_1784328	1.8E-04	A_24_P785293	3.4E-02	1743	7	yes
MSTRG.43631	PCBP1-AS1	ENSG00000179818.13,ENSG00000231327.1	ENSG00000231327.1,ENSG00000179818.9	1.10	7.9E-08	ILMN_1800197	1.2E-13	A_19_P00317631	2.3E-03	150	7	yes
MSTRG.53513	THUMPD3-AS1	ENSG00000206573.8,ENSG00000254485.5	ENSG00000206573.4,ENSG00000254485.1	0.84	0.0E+00	ILMN_2400644	1.3E-26	A_23_P358410	2.3E-02	57	7	yes
MSTRG.43748	AC073263.1	ENSG00000217702.2	NA	0.77	0.0E+00	ILMN_1869390	1.7E-33	A_33_P3341524	2.0E-05	9795	5	yes
MSTRG.56198	LINC02035	ENSG00000273033.2	NA	1.15	3.1E-11	ILMN_1793743	9.8E-07	A_33_P3413098	1.3E-04	167	4	yes
MSTRG.25669	DICER1-AS1	ENSG00000235706.7	ENSG00000235706.3	1.06	1.5E-02	ILMN_1764619	7.3E-61	A_24_P678418	2.9E-05	1	3	yes
MSTRG.73157	SNHG15	ENSG00000232956.8	ENSG00000232956.4	0.90	4.4E-05	ILMN_1898124	2.1E-17	A_19_P00331316	4.9E-05	300	2	yes
MSTRG.80109	AC105202.1	ENSG00000247317.3	ENSG00000247317.3	0.89	1.2E-04	ILMN_1695404	4.7E-21	A_33_P3251492	6.1E-04	127	2	yes
MSTRG.4833	ADAMTSL4-AS1	ENSG00000203804.4	NA	0.89	4.6E-06	ILMN_1803988	2.6E-09	A_23_P115011	2.2E-03	92	2	yes
MSTRG.8195	LINC01341	ENSG00000227953.6,ENSG00000260855.1	NA	1.19	9.3E-11	ILMN_1741506	4.3E-32	A_19_P00320979	6.3E-04	47	2	yes
MSTRG.76138	LINC01006	ENSG00000182648.11	ENSG00000182648.7	1.05	4.1E-02	ILMN_1737663	2.3E-29	A_23_P31218	6.1E-05	23	1	yes
MSTRG.47206	LINC01963	ENSG00000260804.3,ENSG00000279348.1	NA	1.15	1.8E-05	ILMN_2105983	5.3E-06	A_33_P3628409	2.0E-02	17	1	yes
MSTRG.30469	AC009133.1	ENSG00000238045.9	ENSG00000261690.1,ENSG00000261962.1	1.29	0.0E+00	ILMN_2194828	2.5E-46	A_33_P3357620	1.5E-06	NA	NA	no
MSTRG.64052	AC008914.1	ENSG00000262211.1,ENSG00000227908.3	ENSG00000227908.3	0.82	4.3E-02	ILMN_2331890	1.8E-07	A_23_P128067	1.7E-08	NA	NA	no
MSTRG.8533	AC044784.1	ENSG00000223808.1,ENSG00000234752.1	ENSG00000223808.1	1.82	0.0E+00	ILMN_1912417	4.5E-105	A_33_P3244998	1.3E-02	NA	NA	no
MSTRG.87448	AC005052.1	ENSG00000279601.1	CATG00000112301.1	0.87	1.2E-06	ILMN_1737015	4.2E-03	A_23_P34018	6.1E-03	NA	NA	no
MSTRG.24092	AL121790.1	ENSG00000258414.1	NA	1.11	5.9E-05	ILMN_1715505	4.2E-152	A_33_P3226575	3.1E-17	5791	20160	no
MSTRG.16880	AC006064.4	ENSG00000269968.1,ENSG00000215039.6,I	ENSG00000269968.1,ENSG00000215039.2	0.82	2.6E-03	ILMN_1802252	1.2E-41	A_33_P3256660	6.3E-05	35	18020	no
MSTRG.41599	NA	ENSG00000269867.1,ENSG00000283103.1	CATG00000041516.1,ENSG00000268230.1	1.18	1.9E-09	ILMN_1700563	2.9E-207	A_23_P38830	6.8E-29	7	15308	no
MSTRG.68252	HCP5B	ENSG00000281831.1,ENSG00000231074.8,I	ENSG00000204623.4,ENSG00000214922.5	0.79	5.7E-15	ILMN_1762861	4.5E-42	A_19_P00318213	2.9E-08	32687	15262	no
MSTRG.28433	AC021483.1	ENSG00000259548.1,ENSG00000279394.1,I	CATG00000025682.1	1.07	1.9E-03	ILMN_1772302	2.2E-34	A_23_P152002	1.8E-07	19	15210	no
MSTRG.4861	AL590133.2	ENSG00000259357.2	NA	1.60	0.0E+00	ILMN_1726108	1.0E-75	A_23_P63010	5.3E-12	50380	14546	no
MSTRG.38645	AC008763.1	ENSG00000268204.1	ENSG00000268204.1	1.23	0.0E+00	ILMN_1720093	1.7E-167	A_24_P399500	6.1E-22	20083	14405	no
MSTRG.11497	AC023283.1	ENSG00000225302.2	ENSG00000225302.1	1.33	0.0E+00	ILMN_1782861	2.4E-60	A_33_P3374289	1.2E-24	12059	13927	no
MSTRG.26916	AC037198.1	ENSG00000259279.1,ENSG00000259345.5,I	ENSG00000259450.1,ENSG00000259345.1	0.85	2.1E-06	ILMN_1849033	1.0E-78	A_23_P353125	1.5E-35	NA	13041	no
MSTRG.29332	AL133297.1	ENSG00000260954.1,ENSG00000280231.1	NA	0.94	1.6E-04	ILMN_1687037	1.3E-140	A_23_P140725	3.1E-22	704	12999	no
MSTRG.68296	AL662797.3	ENSG00000272540.1,ENSG00000214894.6	ENSG00000272540.1,ENSG00000214894.2	0.70	0.0E+00	ILMN_1732967	8.4E-63	A_23_P331479	3.4E-09	2852	12178	no
MSTRG.17365	RERG-IT1	ENSG00000256650.1	NA	1.10	9.8E-04	ILMN_1746359	4.8E-184	A_24_P321525	9.0E-24	NA	10572	no
MSTRG.33995	AC243829.1	ENSG00000274767.1,ENSG00000277089.4	NA	0.52	0.0E+00	ILMN_1716276	3.6E-27	A_23_P55270	7.1E-11	42388	10136	no
MSTRG.69115	C2-AS1	ENSG00000281756.1,ENSG00000233627.2,I	NA	1.94	0.0E+00	ILMN_1719250	2.9E-101	A_24_P684183	3.1E-14	11533	10047	no
MSTRG.69167	PSMB8-AS1	ENSG00000204261.8	ENSG00000204261.4,CATG00000083589.1	0.82	1.2E-08	ILMN_1751079	8.9E-52	A_19_P00321143	5.8E-06	43209	9591	no
MSTRG.30461	AC009133.5	ENSG00000280607.1	ENSG00000260719.1	1.20	2.3E-07	ILMN_1724424	2.0E-111	A_23_P66017	2.3E-16	268	9570	no
MSTRG.5554	AL391001.1	ENSG00000269887.1,ENSG00000271917.1	NA	1.63	0.0E+00	ILMN_1784678	6.9E-14	A_24_P173823	3.6E-13	26446	9522	no

MSTRG.52501	AC244250.1	ENSG00000223999.1,ENSG00000274422.1	CATG00000058281.1	0.56	0.0E+00	ILMN_1678859	1.2E-12	A_33_P3372647	1.2E-04	45818	9284	no
MSTRG.61156	AC023886.1	ENSG00000249509.1	ENSG00000249509.1	1.23	2.3E-12	ILMN_1700257	1.1E-112	A_23_P58390	8.1E-14	906	9195	no
MSTRG.34393	AC019349.1	ENSG00000229732.1	CATG00000031386.1	1.50	2.9E-07	ILMN_1736760	2.8E-55	A_23_P96158	2.0E-04	11	8975	no
MSTRG.12134	AC068580.3	ENSG00000235027.1,ENSG00000227306.1,I	ENSG00000235027.1,ENSG00000233930.3	1.14	1.7E-02	ILMN_2169261	3.1E-24	A_23_P24784	4.4E-06	NA	8932	no
MSTRG.31125	AC026461.1	ENSG00000259827.1,ENSG00000259923.1	ENSG00000259827.1	0.51	1.1E-10	ILMN_1775170	5.1E-47	A_33_P3236868	2.3E-06	11097	8772	no
MSTRG.26113	AL583722.1	ENSG00000256050.2,ENSG00000258430.1	ENSG00000256050.2,ENSG00000258430.1	1.07	4.1E-02	ILMN_1727248	4.4E-05	A_33_P3289128	6.1E-11	9	8722	no
MSTRG.29342	AL031710.2	ENSG00000275092.1,ENSG00000260107.1,I	ENSG00000255198.3	1.09	1.8E-04	ILMN_1669456	1.8E-128	A_23_P152115	2.0E-14	75	8460	no
MSTRG.41645	AC012313.3	ENSG00000268230.5	ENSG00000268230.1	1.23	2.2E-03	ILMN_1736007	7.3E-03	A_33_P3323323	2.4E-10	NA	8404	no
MSTRG.40151	AC016590.2	ENSG00000267682.1	ENSG00000226686.3	0.80	2.8E-10	ILMN_1675150	5.6E-21	A_23_P429020	2.8E-07	3	8404	no
MSTRG.35832	NA	NA	CATG00000034547.1	1.59	0.0E+00	ILMN_1680925	1.6E-52	A_33_P3378835	1.8E-09	24606	8286	no
MSTRG.8723	NA	ENSG00000229124.6,ENSG00000234961.1	ENSG00000229124.2,ENSG00000234961.1	0.81	0.0E+00	ILMN_2404112	3.1E-26	A_23_P115636	3.6E-08	34471	8083	no
MSTRG.38805	AC011511.5	ENSG00000267607.1,ENSG00000266978.1	ENSG00000267607.1	0.79	0.0E+00	ILMN_1681296	1.2E-09	A_33_P3363355	2.9E-04	40116	7724	no
MSTRG.6823	BLACAT1	ENSG00000281406.1	NA	0.67	0.0E+00	ILMN_1826921	2.7E-42	A_19_P00319808	4.2E-10	54550	7720	no
MSTRG.69278	AL021368.4	ENSG00000283352.1,ENSG00000272316.1	ENSG00000215190.4	0.88	4.9E-07	ILMN_1705991	8.7E-42	A_24_P332739	6.0E-08	NA	7715	no
MSTRG.8338	AL451164.3	ENSG00000278419.1,ENSG00000237399.7	CATG00000102496.1	0.70	0.0E+00	ILMN_1805737	7.8E-116	A_23_P46928	2.4E-13	20433	7154	no
MSTRG.36556	AP005060.1	ENSG00000279885.1	NA	0.87	2.2E-03	ILMN_1712155	4.5E-02	A_24_P391918	3.1E-07	9072	7152	no
MSTRG.40425	AC008537.3	ENSG00000279108.1,ENSG00000269353.1	NA	1.15	1.2E-15	ILMN_2191192	1.5E-209	A_24_P339514	9.7E-40	1966	7117	no
MSTRG.53337	U62317.1	ENSG00000226738.1,ENSG00000273272.1	ENSG00000273272.1,ENSG00000226738.1	0.56	0.0E+00	ILMN_1661314	1.8E-13	A_24_P117410	4.9E-07	22852	7016	no
MSTRG.5375	NA	NA	CATG00000020712.1,CATG00000020742.1	0.92	4.1E-02	ILMN_1706413	1.1E-87	A_23_P115064	2.6E-08	NA	6999	no
MSTRG.44281	AC244205.1	ENSG00000240040.6	NA	0.47	1.4E-13	ILMN_1680274	2.9E-29	A_24_P357847	1.3E-03	43120	6840	no
MSTRG.65814	AC006077.2	ENSG00000279799.1,ENSG00000248753.1,I	CATG00000077624.1	1.19	7.2E-13	ILMN_1651364	4.7E-138	A_33_P3661631	1.4E-04	2005	6760	no
MSTRG.29845	AC130650.2	ENSG00000276564.1	NA	1.15	0.0E+00	ILMN_1690807	9.5E-95	A_23_P54556	7.8E-21	8974	6749	no
MSTRG.55947	AC093010.3	ENSG00000259976.3	ENSG00000259976.1	1.18	2.0E-06	ILMN_1878029	3.5E-15	A_19_P00805237	1.2E-03	978	6376	no
MSTRG.18493	AC023509.1	ENSG00000257379.1,ENSG00000282977.1	ENSG00000257379.1	1.12	7.5E-03	ILMN_2344204	6.8E-50	A_23_P139575	1.5E-10	NA	6370	no
MSTRG.69155	NA	ENSG00000232080.5,ENSG00000223534.1	CATG00000083570.1,CATG00000083577.1	0.84	4.4E-03	ILMN_1808405	7.6E-15	A_24_P852756	1.3E-03	28434	6321	no
MSTRG.6208	AL078645.1	ENSG00000230470.1	NA	1.17	4.3E-06	ILMN_1797735	6.5E-103	A_23_P113161	2.0E-12	1491	6294	no
MSTRG.46023	CERS6-AS1	ENSG00000227617.8	NA	1.32	0.0E+00	ILMN_1767662	2.3E-93	A_32_P5480	2.0E-14	42182	5708	no
MSTRG.58409	AC107464.3	ENSG00000272927.1,ENSG00000233799.1,I	ENSG00000272927.1,ENSG00000249592.1	1.12	2.5E-11	ILMN_1801703	6.1E-116	A_24_P51909	1.5E-16	2190	5461	no
MSTRG.37762	AC022726.1	ENSG00000267766.1	NA	1.60	1.2E-15	ILMN_1801119	3.4E-151	A_23_P352266	2.7E-18	264	5203	no
MSTRG.69054	AL671883.2	ENSG00000256166.1	ENSG00000272501.1	0.76	7.1E-04	ILMN_1778401	1.5E-23	A_33_P3330109	1.8E-03	317	4984	no
MSTRG.29042	AC118658.2	ENSG00000278022.1,ENSG00000278090.1,I	CATG00000024257.1	0.85	6.0E-16	ILMN_1744023	1.8E-187	A_23_P417282	7.0E-16	21976	4936	no
MSTRG.34673	AC138150.1	ENSG00000224505.2,ENSG00000267288.2	ENSG00000267288.1,ENSG00000224505.2	1.21	1.7E-13	ILMN_1701006	1.1E-106	A_23_P377214	2.7E-11	239	4777	no
MSTRG.69051	HCG27	ENSG00000206344.7,ENSG00000206337.1C	ENSG00000206337.6	0.87	3.2E-06	ILMN_1883997	8.0E-23	A_32_P220770	4.0E-04	4018	4564	no
MSTRG.7056	AL360091.1	ENSG00000229832.1,ENSG00000230063.1	NA	1.20	2.0E-02	ILMN_1738784	1.5E-42	A_24_P251221	1.4E-07	NA	4446	no
MSTRG.32666	PRAL	ENSG00000279296.1	NA	0.89	7.6E-03	ILMN_1742618	1.1E-08	A_24_P557479	2.1E-03	3320	4259	no
MSTRG.10356	AL450306.1	ENSG00000228748.2	NA	1.22	2.5E-07	ILMN_1744709	1.4E-23	A_24_P129632	5.8E-11	NA	4215	no
MSTRG.83128	AL157935.1	ENSG00000227218.7	NA	1.07	4.6E-04	ILMN_2401779	6.6E-58	A_23_P112531	1.3E-12	49	3801	no
MSTRG.49532	AL118522.1	ENSG00000276223.1	NA	1.19	0.0E+00	ILMN_1746031	2.2E-104	A_32_P38645	9.0E-44	12027	3775	no
MSTRG.10852	AL157387.1	ENSG00000234918.1	NA	1.50	0.0E+00	ILMN_1651329	6.8E-127	A_23_P12533	5.5E-22	25490	3638	no
MSTRG.58475	AC016773.1	ENSG00000270195.1	ENSG00000270195.1	0.94	4.9E-03	ILMN_1724407	7.1E-36	A_23_P212844	5.9E-09	31	3578	no
MSTRG.54655	ZMYND10-AS1	ENSG00000235058.1,ENSG00000271858.5	ENSG00000235058.1,ENSG00000271858.1	1.30	0.0E+00	ILMN_1738718	3.2E-64	A_23_P121326	1.6E-09	13078	3145	no
MSTRG.37466	AC100778.2	ENSG00000263916.1	CATG00000037523.1	1.12	2.3E-02	ILMN_2196588	1.2E-20	A_24_P271049	8.5E-05	5	2926	no
MSTRG.37573	AC007673.1	ENSG00000267311.1	NA	1.15	4.8E-04	ILMN_1837428	2.7E-80	A_23_P107612	2.2E-08	NA	2846	no
MSTRG.7451	NA	NA	CATG00000037366.1,CATG00000037271.1	1.18	1.9E-04	ILMN_2099594	7.0E-24	A_24_P66528	2.5E-05	432	2598	no
MSTRG.28898	AC106028.3	ENSG00000260361.1,ENSG00000272298.1,I	ENSG00000258922.1	1.37	5.8E-05	ILMN_1652797	7.9E-62	A_23_P100001	1.7E-09	NA	2517	no
MSTRG.30022	AC130456.5	ENSG00000260934.1,ENSG00000279747.1	NA	1.05	2.3E-02	ILMN_1730117	2.2E-27	A_23_P15101	3.1E-08	NA	2507	no
MSTRG.27843	AL928654.3	ENSG00000257270.1	CATG00000022219.1,CATG00000022222.1	1.22	3.7E-07	ILMN_1694432	5.7E-23	A_33_P3251703	1.8E-12	6502	2421	no
MSTRG.40779	AC073548.1	ENSG00000279861.1	NA	1.20	0.0E+00	ILMN_1713285	2.5E-26	A_23_P55990	9.8E-07	1909	2403	no
MSTRG.35246	AC099850.3	ENSG00000265415.1	NA	0.79	2.9E-12	ILMN_1782403	2.3E-62	A_33_P3403075	8.6E-09	334	2217	no
MSTRG.69209	HCG25	ENSG00000232940.5	ENSG00000232940.1	1.07	2.0E-07	ILMN_1666632	3.1E-25	A_23_P70583	2.2E-05	5616	2074	no
MSTRG.70511	Z99129.3	ENSG00000279114.1,ENSG00000272472.1	NA	1.57	6.0E-16	ILMN_2337263	2.0E-111	A_23_P145529	6.3E-15	7908	2027	no
MSTRG.14412	NA	NA	ENSG00000256514.1	1.71	0.0E+00	ILMN_1803772	4.5E-72	A_23_P127367	5.1E-16	16469	1556	no
MSTRG.54293	AC099329.1	ENSG00000235288.3	NA	1.29	1.8E-14	ILMN_1674522	1.1E-26	A_24_P548866	2.3E-08	6230	1555	no
MSTRG.3618	AL390038.1	ENSG00000224698.1	NA	1.69	0.0E+00	ILMN_1704065	5.0E-180	A_33_P3435493	6.0E-30	18422	1542	no
MSTRG.52230	AC004832.5	ENSG00000273350.1	NA	1.31	0.0E+00	ILMN_1684211	4.7E-118	A_23_P17811	2.5E-26	16005	1527	no
MSTRG.64814	AC018754.1	ENSG00000272367.1	NA	1.12	3.7E-07	ILMN_1742250	1.4E-53	A_23_P30338	8.6E-10	NA	1507	no
MSTRG.51028	GRIK1-AS1	ENSG00000174680.9	NA	1.14	7.2E-10	ILMN_1658375	7.9E-67	A_33_P3353471	4.7E-12	2333	1417	no
MSTRG.36378	AC145207.3	ENSG00000262831.1,ENSG00000263859.1,I	ENSG00000262831.1,ENSG00000262413.1	0.86	8.6E-05	ILMN_1719303	4.4E-19	A_33_P3351664	3.3E-04	NA	1354	no
MSTRG.46572	NA	NA	CATG00000050961.1,CATG00000050959.1	1.25	1.4E-03	ILMN_1773079	5.8E-03	A_24_P935491	1.2E-04	1918	1337	no
MSTRG.35020	AC007485.2	ENSG00000279059.1	NA	1.11	5.0E-05	ILMN_1666280	5.9E-09	A_23_P118493	3.5E-10	821	1307	no
MSTRG.10584	AL136982.6	ENSG00000272508.1,ENSG00000271880.1,I	ENSG00000271880.1,CATG00000115873.1	1.27	3.0E-06	ILMN_1680110	6.4E-94	A_23_P161439	7.0E-11	2331	1297	no
MSTRG.29422	AC141586.5	ENSG00000279162.1,ENSG00000261288.1,I	NA	1.08	6.0E-05	ILMN_2387799	2.7E-46	A_24_P222599	3.0E-04	310	1250	no
MSTRG.66271	AC011352.1	ENSG00000248362.1	NA	0.79	0.0E+00	ILMN_1718866	2.5E-41	A_33_P3392052	3.0E-08	3036	1212	no
MSTRG.74361	NA	NA	CATG00000096485.1,CATG00000096487.1	1.41	1.3E-05	ILMN_2104356	1.7E-04	A_24_P277934	1.2E-03	3446	1173	no
MSTRG.31931	AC092718.7	ENSG00000278985.1,ENSG00000260213.5	ENSG00000260213.1	0.71	0.0E+00	ILMN_1783333	4.7E-101	A_23_P106544	6.0E-09	1332	1118	no
MSTRG.36072	TNRC6C-AS1	ENSG00000204282.4	NA	0.82	2.2E-09	ILMN_1794677	5.5E-26	A_33_P3302423	1.3E-05	13385	1099	no
MSTRG.29141	AC023024.2	ENSG00000279970.1	NA	1.08	1.8E-03	ILMN_1676384	2.9E-43	A_23_P151907	2.6E-07	NA	1048	no
MSTRG.45257	AC013269.2	ENSG00000225341.1,ENSG00000282944.1	NA	0.77	1.2E-15	ILMN_1659631	1.8E-35	A_24_P237804	1.6E-15	1188	1034	no
MSTRG.48990	ABALON	ENSG00000281376.1,ENSG00000236559.1	CATG00000052997.1	1.26	5.0E-12	ILMN_1796949	2.7E-55	A_23_P68610	4.5E-08	711	1025	no
MSTRG.30594	NA	ENSG00000278133.1,ENSG00000261385.1,I	ENSG00000262766.1	1.24	3.5E-06	ILMN_1804679	6.7E-37	A_33_P3287745	7.3E-07	845	1024	no
MSTRG.41470	AC024580.1	ENSG00000279794.1	CATG00000039871.1	1.20	0.0E+00	ILMN_1727300	1.4E-31	A_33_P3360555	3.9E-05	4866	1019	no
MSTRG.3693	AC000032.1	ENSG00000260246.1	NA	1.27	1.2E-04	ILMN_1762255	1.2E-06	A_23_P115407	1.3E-09	10	1013	no
MSTRG.36394	AC145207.5	ENSG00000263731.1,ENSG00000265692.1,I	ENSG00000263731.1,ENSG00000265692.1	0.87	1.5E-07	ILMN_1750518	1.9E-32	A_23_P152984	7.0E-08	148	943	no
MSTRG.66440	AC011374.1	ENSG00000260581.1,ENSG00000249035.6,I	ENSG00000254226.1	1.14	6.2E-03	ILMN_1796734	1.9E-12	A_33_P3382924	1.0E-05	493	937	no
MSTRG.74768	LNCPRESS1	ENSG00000232301.1,ENSG00000232445.1	CATG00000093299.1,ENSG00000232445.1	0.88	6.1E-07	ILMN_1787951	1.8E-31	A_24_P286527	1.1E-11	317	914	no
MSTRG.12066	AC131934.1	ENSG00000255158.1,ENSG00000255142.1,I	CATG00000001307.1	1.03	6.7E-02	ILMN_1689868	2.8E-48	A_23_P332960	3.1E-07	25	909	no
MSTRG.56963	AC092979.1	ENSG00000240521.1	NA	1.40	7.3E-10	ILMN_1766551	5.7E-36	A_23_P6615	1.3E-20	15300	907	no
MSTRG.50099	LINC01711	ENSG00000268941.2	NA	1.05	3.0E-02	ILMN_1749414	5.6E-11	A_23_P102681	3.0E-05	NA	902	no
MSTRG.19085	AC025423.4	ENSG00000257181.1	NA	1.05	6.2E-02	ILMN_2299795	6.3E-24	A_24_P925664	3.1E-07	NA	900	no
MSTRG.34345	AC018629.1	ENSG00000279806.1	CATG00000033638.1,CATG00000033637.1	1.92	3.3E-12	ILMN_1665865	4.2E-79	A_24_P382187	2.4E-13	13379	888	no
MSTRG.7182	NA	NA	CATG00000094282.1	1.31	0.0E+00	ILMN_1831834	3.0E-15	A_33_P3355743	1.2E-04	1562	842	no
MSTRG.35109	AC027801.5	ENSG00000262837.1,ENSG00000236472.1,I	CATG00000031710.1,CATG00000031711.1	1.20	3.7E-02	ILMN_1705397	1.6E-47	A_23_P207517	1.4E-05	516	831	no
MSTRG.50617	MIR99AHG	ENSG00000215386.11,ENSG00000280432.1	NA	0.95	6.5E-02	ILMN_1690703	1.5E-05	A_33_P3406886	6.8E-04	2078	806	no
MSTRG.19258	AC121761.2	ENSG00000273987.1	NA	1.06	3.0E-04	ILMN_1747105	2.5E-03	A_32_P209208	1.4E-15	2	760	no
MSTRG.52819	AL022326.2	ENSG00000280216.1	NA	1.12	1.3E-04	ILMN_1727805	1.0E-75	A_23_P348063	4.0E-08	1	684	no
MSTRG.50543	AL050303.4	ENSG00000279773.1	NA	1.46	0.0E+00	ILMN_1719826	5.7E-94	A_33_P3391316	7.9E-27	18494	675	no
MSTRG.12080	AP006621.1	ENSG00000255108.1,ENSG00000250397.2	ENSG00000255108.1	1.13	1.5E-02	ILMN_1670037	4.5E-70	A_24_P295806	5.6E-05	NA	568	no
MSTRG.34186	AC006441.3	ENSG00000265784.1	NA	1.09	2.1E-09	ILMN_1665909	3.8E-09	A_23_P89187	6.8E-03	3775	550	no
MSTRG.26938	NA	ENSG00000259409.1	ENSG00000248508.2,CATG00000022687.1	1.11	6.3E-02	ILMN_1809347	2.5E-36	A_23_P14876	1.9E-04	166	536	no
MSTRG.58959	TAPT1-AS1	ENSG00000263327.6	ENSG00000263327.2	1.13	1.2E-11	ILMN_1903049	1.6E-63	A_33_P3335984	1.7E-08	39	524	no
MSTRG.196	NA	NA	ENSG00000227372.6	1.22	6.4E-10	ILMN_2090040	3.7E-03	A_33_P3317948	3.0E-03	41	523	no
MSTRG.85294	AC078993.1	ENSG00000238178.6	ENSG00000238178.2	1.74	0.0E+00	ILMN_2119123	1.4E-214	A_24_P357266	1.0E-51	19252	510	no
MSTRG.29924	AF001548.1	ENSG00000263065.1,ENSG00000280206.1,I	ENSG00000263335.1	0.90	7.2E-04	ILMN_1739805	1.6E-25	A_23_P206901	8.1E-08	NA	505	no
MSTRG.42904	NA	NA	CATG00000048295.1,CATG00000048300.1	0.78	2.6E-04	ILMN_1662417	3.7E-07	A_24_P674924	3.3E-04	2	503	no
MSTRG.5023	NA	ENSG00000272030.1,ENSG00000271853.5,I	CATG00000076852.1,ENSG00000272030.1	1.43	1.3E-06	ILMN_1765459	1.9E-38	A_23_P372874	1.3E-05	244	498	no
MSTRG.27098	AC091117.1	ENSG00000259352.1,ENSG00000259418.1	NA	1.29	6.0E-16	ILMN_1692517	3.4E-40	A_23_P77103	3.1E-03	25206	486	no
MSTRG.48388	AL035461.2	ENSG00000275632.1	NA	0.92	3.2E-03	ILMN_2407124	3.5E-23	A_23_P68547	2.0E-09	31	482	no
MSTRG.30477	AC120114.2	ENSG00000260114.2,ENSG00000261416.1,I	ENSG00000261416.1	1.10	1.5E-07	ILMN_1796244	1.6E-71	A_23_P61551	2.6E-05	303	476	no
MSTRG.4995	AL589765.1	ENSG00000227045.1,ENSG00000232536.1,I	ENSG00000232536.1,ENSG00000234614.1	1.09	4.5E-07	ILMN_1781374	4.5E-26	A_23_P371824	7.5E-07	113	467	no
MSTRG.36613	AP001033.2	ENSG00000273284.1,ENSG00000265257.5	ENSG00000265257.1	1.05	8.1E-04	ILMN_1661833	9.1E-19	A_23_P130418	3.0E-04	NA	449	no
MSTRG.10600	AL136982.7	ENSG00000273413.1,ENSG00000229969.1,I	ENSG00000273413.1	1.10	1.1E-06	ILMN_1680110	6.4E-94	A_23_P161439	7.0E-11	3	446	no
MSTRG.7207	TGFB2-OT1	ENSG00000281453.1	NA	0.69	1.8E-05	ILMN_1812526	1.2E-09	A_24_P402438	3.1E-02	1517	439	no
MSTRG.58657	LINC02482	ENSG00000251580.1	ENSG00000251580.1	1.23	2.4E-07	ILMN_2055165	1.2E-70	A_23_P144453	2.7E-09	529	435	no
MSTRG.53136	LINC01644	ENSG00000218357.3	NA	1.17	7.5E-12	ILMN_1900573	1.2E-55	A_33_P3371509	1.1E-25	9291	426	no
MSTRG.5186	AL513302.2	ENSG00000260766.1	NA	1.24	2.7E-06	ILMN_1799860	8.5E-12	A_33_P3386970	2.5E-03	22	420	no
MSTRG.16039	AP002840.2	ENSG00000270179.1	NA	1.19	2.0E-06	ILMN_2183784	2.6E-48	A_33_P3336053	4.9E-09	89	414	no
MSTRG.18411	NA	ENSG00000257671.1	ENSG00000257829.1,ENSG00000257137.1	0.80	7.1E-10	ILMN_2163723	1.8E-48	A_33_P3350748	2.4E-08	9491	403	no
MSTRG.14917	AP001972.5	ENSG00000279117.1	CATG00000006169.1	1.54	2.9E-14	ILMN_2325168	6.6E-43	A_24_P386622	6.2E-06	500	394	no
MSTRG.64062	LINC01948	ENSG00000248727.5	NA	1.10	3.6E-05	ILMN_1822495	3.2E-92	A_19_P00320193	1.1E-11	NA	376	no
MSTRG.25479	AL137230.2	ENSG00000259053.1,ENSG00000258792.4	ENSG00000259053.1	1.12	2.6E-03	ILMN_2106902	3.7E-41	A_23_P129014	2.7E-06	NA	362	no
MSTRG.66011	AC011379.1	ENSG00000250069.1,ENSG00000254363.6	NA	1.23	1.1E-14	ILMN_2127328	1.3E-59	A_33_P3411296	6.8E-06	47	359	no
MSTRG.27161	AC011330.2	ENSG00000275601.1	NA	1.12	3.4E-06	ILMN_2333766	6.4E-52	A_33_P3420020	1.2E-07	NA	358	no

MSTRG.21904	INTS6-AS1	ENSG00000236778.7	ENSG00000236778.3	1.07	5.6E-04	ILMN_1905075	3.5E-07	A_32_P142652	1.9E-06	137	357 no
MSTRG.18237	AC011603.2	ENSG00000258017.1	ENSG00000258017.1	0.94	2.9E-02	ILMN_1742167	8.5E-23	A_24_P188071	8.5E-05	NA	336 no
MSTRG.74445	AC093799.2	ENSG00000272950.1	NA	0.93	4.8E-02	ILMN_1781580	1.4E-40	A_23_P122915	9.6E-07	NA	336 no
MSTRG.62888	NA	ENSG00000261434.1	ENSG00000249430.1	0.82	1.2E-08	ILMN_1702020	1.1E-15	A_33_P3746549	3.4E-04	NA	333 no
MSTRG.51778	AC006946.2	ENSG00000273203.1,ENSG00000273442.1	NA	0.81	0.0E+00	ILMN_1728724	1.4E-10	A_33_P3224809	1.5E-09	23636	309 no
MSTRG.30489	NA	NA	ENSG00000250616.2	1.20	1.4E-11	ILMN_1791147	7.9E-57	A_23_P15108	4.3E-07	114	306 no
MSTRG.13392	AC090589.3	ENSG00000270060.1	NA	1.12	0.0E+00	ILMN_1769250	3.2E-07	A_33_P3292840	1.7E-09	1244	304 no
MSTRG.49497	AL021578.1	ENSG00000275894.1	NA	1.07	9.5E-06	ILMN_1663042	1.2E-39	A_23_P109034	3.1E-05	31	291 no
MSTRG.28561	AC024270.3	ENSG00000260579.1	NA	1.08	1.5E-03	ILMN_1784256	4.6E-41	A_23_P344451	2.0E-08	25	277 no
MSTRG.27791	AC013553.4	ENSG00000278737.1,ENSG00000274383.1	NA	1.27	9.6E-09	ILMN_1723847	3.8E-16	A_23_P151895	9.8E-06	55	264 no
MSTRG.23719	NA	NA	ENSG00000259321.1	1.65	8.0E-09	ILMN_1806017	1.3E-28	A_23_P151614	2.2E-05	121	249 no
MSTRG.234	AL031848.1	ENSG00000231868.1,ENSG00000226944.1	NA	1.12	2.2E-03	ILMN_2079386	9.1E-27	A_24_P849801	5.3E-06	4	243 no
MSTRG.14075	AP001453.3	ENSG00000257086.1	ENSG00000257086.1	0.54	0.0E+00	ILMN_1659888	1.1E-85	A_23_P365817	8.5E-12	10088	231 no
MSTRG.1219	AL360012.1	ENSG00000270103.3,ENSG00000242125.3,I	ENSG00000229388.1,ENSG00000242125.2	0.78	0.0E+00	ILMN_1707493	7.1E-32	A_23_P46309	1.9E-06	2253	228 no
MSTRG.7319	AC093152.1	ENSG00000278610.1,ENSG00000228106.5	ENSG00000228106.1	1.06	2.3E-02	ILMN_1695745	5.6E-34	A_23_P201319	2.1E-04	319	216 no
MSTRG.28391	AC027228.2	ENSG00000261762.1,ENSG00000279421.1	NA	0.74	0.0E+00	ILMN_1770044	1.4E-50	A_23_P88691	7.6E-11	7947	213 no
MSTRG.19641	AC138123.2	ENSG00000257322.5,ENSG00000257252.5	ENSG00000257322.1,ENSG00000257252.1	1.30	2.9E-14	ILMN_2331890	1.8E-07	A_23_P128067	1.7E-08	990	212 no
MSTRG.64137	PART1	ENSG00000152931.7	ENSG00000152931.7	0.74	0.0E+00	ILMN_1724265	4.2E-03	A_33_P3215744	7.2E-05	12415	207 no
MSTRG.43740	AC073046.1	ENSG00000235499.1	ENSG00000235499.1,ENSG00000255989.1	0.89	1.4E-07	ILMN_1759352	3.9E-02	A_33_P3345001	8.5E-07	NA	194 no
MSTRG.21942	AL158066.1	ENSG00000217576.7	NA	1.08	1.3E-04	ILMN_2184525	4.7E-120	A_32_P48134	5.4E-16	NA	186 no
MSTRG.15425	AP001528.3	ENSG00000280339.1,ENSG00000255471.1	CATG00000006516.1,ENSG00000255471.1	1.14	1.0E-09	ILMN_1893728	3.2E-03	A_19_P00803777	3.1E-02	575	180 no
MSTRG.8154	SMYD3-IT1	ENSG00000230184.1	ENSG00000230184.1,CATG00000101171.1	1.16	2.6E-05	ILMN_1741954	1.2E-12	A_32_P103291	2.1E-05	1	177 no
MSTRG.35864	AC022211.1	ENSG00000263786.1	CATG00000034574.1	0.56	0.0E+00	ILMN_2384591	1.2E-36	A_23_P100632	4.8E-05	2048	172 no
MSTRG.28850	AC068831.2	ENSG00000224441.1	ENSG00000224441.1	1.32	5.6E-10	ILMN_1781638	4.3E-27	A_23_P65830	2.7E-03	3507	167 no
MSTRG.3614	AL359258.1	ENSG00000238122.1	NA	1.59	0.0E+00	ILMN_1704065	5.0E-180	A_33_P3435493	6.0E-30	16098	164 no
MSTRG.327	NA	ENSG00000230679.1	CATG00000008775.1,ENSG00000230679.1	0.60	0.0E+00	ILMN_1710756	4.3E-93	A_23_P137391	5.5E-05	16786	154 no
MSTRG.67267	AC008443.1	ENSG00000233937.6,ENSG00000248275.1,I	ENSG00000248275.1,ENSG00000248103.1	1.05	6.5E-03	ILMN_1754130	1.4E-21	A_23_P58579	7.9E-03	3	152 no
MSTRG.34428	AC125257.1	ENSG00000259623.1	NA	1.33	0.0E+00	ILMN_2371379	9.9E-07	A_23_P55388	1.0E-03	29735	149 no
MSTRG.40443	AC011462.4	ENSG00000277744.1,ENSG00000268475.1,I	ENSG00000268987.1	1.05	8.6E-03	ILMN_1735979	2.5E-24	A_33_P3363620	5.3E-07	NA	138 no
MSTRG.37492	AC090246.1	ENSG00000277310.1	NA	1.21	0.0E+00	ILMN_1746949	3.7E-41	A_23_P384532	7.0E-09	27564	129 no
MSTRG.17295	AC008115.3	ENSG00000275560.1	NA	1.33	0.0E+00	ILMN_1722811	6.4E-29	A_24_P81841	6.2E-11	516	99 no
MSTRG.25783	AL157871.1	ENSG00000258504.2	NA	1.06	9.7E-03	ILMN_1697544	2.2E-21	A_23_P77048	2.0E-03	49	94 no
MSTRG.29653	AC087190.4	ENSG00000263244.2	ENSG00000262944.1	1.46	0.0E+00	ILMN_1773407	1.2E-38	A_32_P4403	9.3E-06	8245	85 no
MSTRG.77117	AC102945.2	ENSG00000279041.1	NA	1.23	1.2E-15	ILMN_1728218	4.2E-23	A_23_P71316	1.5E-05	2661	84 no
MSTRG.48852	AL031673.1	ENSG00000278383.1,ENSG00000213742.6	ENSG00000213742.2	1.07	3.5E-03	ILMN_1663160	7.0E-28	A_33_P3333805	3.3E-04	NA	82 no
MSTRG.18416	AC021066.1	ENSG00000257830.1,ENSG00000257829.1	ENSG00000257829.1	0.76	0.0E+00	ILMN_1801442	6.2E-55	A_33_P3273534	5.9E-08	29639	71 no
MSTRG.3712	AL355816.2	ENSG00000273221.1	NA	1.18	1.2E-03	ILMN_1707088	2.5E-21	A_23_P85952	3.7E-04	41	71 no
MSTRG.27021	OIP5-AS1	ENSG00000247556.6	ENSG00000247556.2	1.11	2.6E-02	ILMN_2196984	2.3E-34	A_23_P379614	4.2E-04	NA	64 no
MSTRG.17092	LINC00987	ENSG00000237248.4	ENSG00000237248.3	0.76	3.4E-08	ILMN_1677906	2.0E-13	A_24_P264207	6.4E-08	1024	64 no
MSTRG.31788	AC009054.2	ENSG00000261783.1,ENSG00000259999.1,I	ENSG00000261476.1	0.93	7.1E-03	ILMN_2212590	6.2E-48	A_23_P100517	7.6E-09	1	62 no
MSTRG.7839	AL359921.1	ENSG00000230325.1,ENSG00000273058.2	ENSG00000273058.1	0.89	6.8E-04	ILMN_1787762	1.7E-47	A_23_P103628	5.8E-04	20	53 no
MSTRG.39748	AC092279.2	ENSG00000279425.1,ENSG00000268362.5	ENSG00000268362.1	1.08	6.1E-03	ILMN_1657229	1.8E-09	A_33_P3221443	1.8E-04	NA	52 no
MSTRG.23710	DHRS4L1	ENSG00000225766.9	CATG00000018317.1	1.09	7.4E-07	ILMN_1676014	1.1E-22	A_33_P3383422	2.4E-07	NA	46 no
MSTRG.41251	AC022150.2	ENSG00000268970.1	NA	1.07	3.0E-05	ILMN_1723542	1.1E-10	A_33_P3222753	3.4E-03	34	41 no
MSTRG.29293	AC120498.2	ENSG00000260182.1,ENSG00000277010.1	NA	1.20	1.1E-12	ILMN_1676256	5.1E-42	A_23_P37702	4.8E-11	10893	40 no
MSTRG.3041	AL603750.1	ENSG00000280099.1	NA	1.56	0.0E+00	ILMN_1840316	3.5E-50	A_33_P3412767	9.4E-03	25949	38 no
MSTRG.72752	AC007036.3	ENSG00000251660.1,ENSG00000281404.1	CATG00000091799.1	1.10	3.8E-02	ILMN_1854383	1.5E-34	A_33_P3312842	3.8E-04	NA	37 no
MSTRG.14788	AP002761.4	ENSG00000260401.1	NA	1.17	4.1E-03	ILMN_2372915	3.9E-07	A_23_P24903	3.1E-04	9	37 no
MSTRG.36768	AP005131.7	ENSG00000272746.1,ENSG00000267702.1	NA	1.15	3.2E-10	ILMN_2351795	1.6E-44	A_24_P130363	1.1E-10	1035	36 no
MSTRG.82262	AL161729.3	ENSG00000271384.1	ENSG00000237857.2	0.89	1.7E-12	ILMN_1683112	1.5E-06	A_33_P3319765	1.2E-09	20	35 no
MSTRG.20555	AC073857.1	ENSG00000280120.1,ENSG00000235423.8	ENSG00000235423.4	0.84	5.2E-14	ILMN_1654421	2.3E-04	A_23_P139486	1.9E-08	25	34 no
MSTRG.65823	C5orf66-AS1	ENSG00000249082.2	ENSG00000249082.1	0.92	6.4E-03	ILMN_1674386	8.8E-03	A_33_P3406796	9.3E-04	169	32 no
MSTRG.34906	AC091180.3	ENSG00000250186.3	NA	1.09	8.0E-06	ILMN_1692651	1.5E-04	A_23_P130040	2.7E-02	988	28 no
MSTRG.80323	AC105219.2	ENSG00000254548.1,ENSG00000254973.1,I	ENSG00000254973.1	1.06	7.0E-04	ILMN_1779404	3.3E-12	A_24_P149645	3.2E-03	2563	26 no
MSTRG.37145	AC011825.3	ENSG00000263917.1,ENSG00000274275.1	ENSG00000263917.1	1.10	5.9E-03	ILMN_1655165	9.6E-08	A_23_P89755	4.1E-05	64	24 no
MSTRG.14121	AP003306.1	ENSG00000254404.1,ENSG00000255118.1	NA	1.41	2.1E-08	ILMN_1723333	5.3E-42	A_23_P150555	4.7E-09	978	23 no
MSTRG.4893	AL450992.1	ENSG00000229021.2	ENSG00000229021.2	0.70	6.9E-11	ILMN_1750101	1.1E-46	A_23_P126593	3.1E-05	250	21 no
MSTRG.52263	TUG1	ENSG00000253352.8	ENSG00000253352.4	1.23	2.3E-02	ILMN_1682783	5.2E-07	A_19_P00320120	1.7E-05	NA	20 no
MSTRG.58452	CTBP1-AS2	ENSG00000196810.4	ENSG00000196810.4	1.12	2.4E-07	ILMN_1706893	6.1E-14	A_23_P7083	1.2E-05	NA	20 no
MSTRG.38987	AC092069.1	ENSG00000267458.1	NA	0.70	0.0E+00	ILMN_1736256	2.9E-52	A_33_P3292854	3.0E-06	1401	19 no
MSTRG.31411	NA	NA	CATG00000027630.1	0.91	3.0E-03	ILMN_1766408	2.7E-62	A_23_P113634	6.6E-05	2	19 no
MSTRG.73700	AC006001.2	ENSG00000226824.6	ENSG00000226824.2	0.82	0.0E+00	ILMN_2230577	3.2E-05	A_24_P329229	2.8E-02	3914	16 no
MSTRG.15027	AP002812.3	ENSG00000254691.1	NA	1.11	5.0E-02	ILMN_1779163	1.9E-23	A_24_P49383	3.9E-02	NA	15 no
MSTRG.41643	AC020915.3	ENSG00000268201.1	ENSG00000142396.6	1.21	3.4E-05	ILMN_1667330	2.8E-13	A_19_P00319584	1.6E-03	948	15 no
MSTRG.79509	AP003476.1	ENSG00000279535.1	NA	1.09	2.9E-05	ILMN_1824136	9.6E-04	A_33_P3326225	4.3E-05	55	14 no
MSTRG.27929	AC116913.1	ENSG00000261351.2	NA	1.04	2.6E-04	ILMN_1752285	1.6E-02	A_33_P3335746	1.7E-03	NA	13 no
MSTRG.71226	AL078581.1	ENSG00000233330.1,ENSG00000231760.4	NA	0.84	6.7E-05	ILMN_1661432	2.7E-02	A_33_P3420078	8.9E-03	596	13 no
MSTRG.44335	AC092636.1	ENSG00000273265.1	CATG00000049366.1	0.83	1.4E-04	ILMN_1662328	5.0E-38	A_33_P3255274	1.1E-04	NA	12 no
MSTRG.49104	NA	NA	CATG00000053050.1,CATG00000053051.1	0.57	0.0E+00	ILMN_2403994	8.9E-12	A_23_P109235	9.3E-04	6284	12 no
MSTRG.50250	AL121832.3	ENSG00000275437.1,ENSG00000273619.1	NA	0.89	4.1E-06	ILMN_1800573	3.6E-11	A_23_P307761	1.9E-02	2	12 no
MSTRG.68998	AL049555.1	ENSG00000261116.1	NA	0.59	0.0E+00	ILMN_1782863	2.6E-33	A_24_P66780	1.1E-05	2451	10 no
MSTRG.25246	NA	NA	CATG00000021599.1	1.54	1.8E-07	ILMN_1669523	5.1E-12	A_23_P106194	8.0E-03	761	9 no
MSTRG.68308	AL662797.2	ENSG00000272273.1	ENSG00000272273.1	1.21	7.5E-13	ILMN_1682717	7.2E-38	A_33_P3423551	8.2E-05	5315	7 no
MSTRG.14058	AP006333.1	ENSG00000256341.1	NA	1.11	1.9E-04	ILMN_1740960	3.6E-13	A_23_P86855	3.3E-02	74	7 no
MSTRG.36365	NA	ENSG00000229848.1	CATG00000032642.1,CATG00000032645.1	0.87	4.6E-02	ILMN_1713247	3.9E-56	A_19_P00811581	1.2E-02	NA	6 no
MSTRG.48188	AL136531.1	ENSG00000229728.1,ENSG00000234684.6	ENSG00000234684.2	1.06	7.5E-03	ILMN_1683658	2.3E-52	A_32_P50522	3.4E-06	NA	6 no
MSTRG.87041	AC234775.3	ENSG00000261101.2	NA	1.07	3.9E-05	ILMN_2334765	3.6E-27	A_24_P23411	4.0E-03	3	6 no
MSTRG.17842	AC026356.2	ENSG00000276115.1,ENSG00000277342.1,I	NA	0.88	6.4E-11	ILMN_2362034	6.9E-11	A_33_P3256425	4.5E-02	6318	5 no
MSTRG.33599	AC002094.1	ENSG00000258924.4,ENSG00000277450.1	NA	1.17	1.2E-07	ILMN_1746265	4.9E-10	A_33_P3370763	1.9E-02	4286	5 no
MSTRG.34796	AC018521.5	ENSG00000264920.1,ENSG00000234494.7,I	ENSG00000264920.1,ENSG00000234494.3	1.07	9.8E-05	ILMN_1769721	4.0E-19	A_33_P3282075	2.8E-02	116	5 no
MSTRG.35670	AC002546.1	ENSG00000267194.1,ENSG00000278972.1	NA	0.94	2.6E-02	ILMN_1660964	1.5E-02	A_23_P207445	1.4E-03	NA	4 no
MSTRG.44713	AC092645.1	ENSG00000278962.1	NA	1.13	2.1E-06	ILMN_1735743	9.2E-12	A_23_P102462	8.5E-04	NA	4 no
MSTRG.83124	AL157935.2	ENSG00000257524.5	NA	1.13	3.1E-05	ILMN_2291455	1.7E-18	A_23_P315892	3.3E-03	49	4 no
MSTRG.49166	AL121753.2	ENSG00000279253.1	ENSG00000261582.1	1.05	1.9E-02	ILMN_1787410	2.0E-12	A_24_P627306	3.2E-03	NA	3 no
MSTRG.81259	EBLN3P	ENSG00000281649.1	ENSG00000233137.2	1.34	1.6E-06	ILMN_1744980	2.4E-03	A_33_P3574055	2.3E-04	NA	3 no
MSTRG.57361	AC112491.1	ENSG00000269888.1,ENSG00000240567.1	NA	1.13	4.7E-04	ILMN_1683597	1.3E-05	A_23_P29747	2.9E-02	10	3 no
MSTRG.51557	AL844908.1	ENSG00000272825.1,ENSG00000276529.1	ENSG00000227039.2	0.66	0.0E+00	ILMN_2104924	2.1E-35	A_33_P3705907	1.0E-05	NA	2 no
MSTRG.63999	AC008873.1	ENSG00000279638.1	CATG00000079768.1	1.08	2.9E-02	ILMN_1774161	5.7E-06	A_24_P303199	3.4E-02	NA	2 no
MSTRG.38322	NA	NA	ENSG00000223573.2	0.69	1.7E-12	ILMN_1725831	1.5E-17	A_33_P3259153	1.8E-03	130	2 no
MSTRG.80759	AL158206.1	ENSG00000260912.1	ENSG00000260912.1	1.29	3.0E-06	ILMN_1885985	3.1E-46	A_33_P3413759	6.6E-03	1	2 no
MSTRG.30664	AC074050.4	ENSG00000276867.1,ENSG00000259810.2	NA	1.04	5.6E-02	ILMN_1656420	1.3E-12	A_23_P77661	7.2E-06	NA	1 no
MSTRG.44259	LINC00342	ENSG00000232931.5	NA	0.89	2.0E-02	ILMN_1811117	2.1E-02	A_33_P3244998	1.3E-02	NA	1 no
MSTRG.53522	AC018809.1	ENSG00000269894.1,ENSG00000230082.1	ENSG00000230082.1	0.66	0.0E+00	ILMN_1739558	9.3E-28	A_23_P6762	1.4E-03	13548	1 no
MSTRG.9114	SVIL-AS1	ENSG00000224597.9	ENSG00000224597.5	0.87	2.6E-11	ILMN_1679150	6.2E-10	A_33_P3257503	1.5E-05	3968	1 no
MSTRG.41652	AC012313.5	ENSG00000268912.1	ENSG00000232098.2	1.06	5.0E-03	ILMN_1812898	3.0E-03	A_33_P3411427	2.1E-02	2627	1 no
MSTRG.59196	AC104662.1	ENSG00000271172.1,ENSG00000281501.1	NA	1.11	8.7E-04	ILMN_1815134	4.8E-19	A_23_P18598	9.1E-04	2022	1 no
MSTRG.15601	AP000786.1	ENSG00000255893.1	NA	0.90	2.0E-11	ILMN_1709483	3.0E-25	A_23_P150189	1.5E-04	491	1 no
MSTRG.16803	AC125807.2	ENSG00000250899.3	NA	0.88	9.9E-07	ILMN_1705301	4.1E-40	A_23_P94795	1.1E-06	282	1 no
MSTRG.28793	ZNF710-AS1	ENSG00000259291.2	NA	0.91	2.1E-03	ILMN_1751753	8.5E-51	A_23_P129209	2.0E-02	23	1 no
MSTRG.47266	AC073321.1	ENSG00000225794.1,ENSG00000236886.2,I	ENSG00000236886.2	1.53	7.9E-12	ILMN_1725193	4.9E-58	A_23_P119943	4.0E-09	6	1 no
MSTRG.3730	AL390195.1	ENSG00000243960.1,ENSG00000260948.1	NA	0.78	7.6E-05	ILMN_1742889	1.9E-26	A_23_P103756	5.8E-03	1	1 no
MSTRG.15753	AP000851.1	ENSG00000256916.1,ENSG00000281655.1	ENSG00000256916.1	0.75	0.0E+00	ILMN_1881206	5.7E-24	A_33_P3398316	1.1E-02	NA	0 no
MSTRG.1795	AL139289.2	ENSG00000234694.1	ENSG00000234694.1	0.85	4.4E-08	ILMN_1776640	8.8E-03	A_23_P23380	4.7E-03	NA	0 no
MSTRG.18299	AC074032.1	ENSG00000272368.2	ENSG00000272368.1	1.08	7.9E-06	ILMN_1804656	5.4E-08	A_23_P151280	3.4E-02	NA	0 no
MSTRG.19835	AC126474.2	ENSG00000280088.1	NA	1.28	2.0E-08	ILMN_1737705	1.8E-20	A_23_P319808	9.6E-03	NA	0 no
MSTRG.24424	AL139317.3	ENSG00000259049.1	NA	1.07	2.9E-05	ILMN_1732475	1.8E-03	A_23_P429184	2.8E-02	NA	0 no
MSTRG.24818	NA	NA	CATG00000021387.1	1.08	1.5E-04	ILMN_1704236	2.2E-09	A_23_P205549	2.3E-04	NA	0 no
MSTRG.25022	NA	NA	ENSG00000258644.1	1.47	0.0E+00	ILMN_1706305	6.8E-13	A_23_P53957	1.9E-02	NA	0 no
MSTRG.28682	LINC00052	ENSG00000259527.2	ENSG00000259527.1	1.37	3.9E-07	ILMN_1770978	4.1E-14	A_23_P382094	1.5E-08	NA	0 no
MSTRG.3098	AC099062.1	ENSG00000284240.1,ENSG00000267272.5	CATG00000006749.1,ENSG00000267272.1	1.15	6.9E-05	ILMN_1699320	6.6E-04	A_23_P383590	1.5E-02	NA	0 no
MSTRG.31884	AC009145.2	ENSG00000260511.1,ENSG00000280274.1,I	ENSG00000260479.1,ENSG00000261537.1	1.06	2.1E-02	ILMN_1804841	2.3E-17	A_23_P71972	9.4E-03	NA	0 no
MSTRG.32623	AC012146.1	ENSG00000234327.7	ENSG00000234327.3	0.80	1.8E-10	ILMN_1782079	3.2E-15	A_33_P3258712	1.4E-03	NA	0 no

MSTRG.34765	AC068152.1	ENSG00000262879.5	ENSG00000262879.1	1.11	5.3E-11	ILMN_2393693	1.7E-41	A_19_P00803086	2.1E-02	NA	0	no
MSTRG.36541	LINC00667	ENSG00000263753.6	ENSG00000263753.2	1.05	3.3E-02	ILMN_1772522	9.7E-09	A_19_P00326065	2.4E-02	NA	0	no
MSTRG.37227	AC091060.1	ENSG00000278986.1	NA	0.91	1.7E-02	ILMN_1805998	1.6E-09	A_23_P101237	1.4E-03	NA	0	no
MSTRG.37945	LINC00909	ENSG00000264247.1	ENSG00000264247.1	1.10	6.4E-06	ILMN_1725528	1.4E-12	A_19_P00323008	3.4E-02	NA	0	no
MSTRG.38400	AC005786.3	ENSG00000267436.1,ENSG00000226800.5,I	ENSG00000226800.5	0.89	1.5E-03	ILMN_1813374	8.0E-23	A_33_P3352562	1.0E-03	NA	0	no
MSTRG.39205	AC008764.10	ENSG00000279977.1,ENSG00000267904.1,I	ENSG00000267275.1	1.07	2.8E-03	ILMN_1718171	4.4E-11	A_33_P3250438	9.8E-03	NA	0	no
MSTRG.39657	ERVK-28	ENSG00000267696.6,ENSG00000261770.1,I	ENSG00000261824.2	0.89	8.0E-11	ILMN_1902018	7.9E-07	A_19_P00320440	1.2E-02	NA	0	no
MSTRG.51211	CBR3-AS1	ENSG00000236830.6,ENSG00000230212.6	ENSG00000236830.2	1.26	2.5E-13	ILMN_1790819	1.0E-75	A_32_P143880	1.4E-03	NA	0	no
MSTRG.51743	AP001469.3	ENSG00000239415.1	NA	0.90	4.3E-08	ILMN_1784766	8.6E-08	A_23_P120744	1.4E-02	NA	0	no
MSTRG.51871	FP671120.1	ENSG00000278996.1,ENSG00000280441.2	NA	0.86	2.5E-02	ILMN_1733559	1.9E-07	A_33_P3336632	2.1E-02	NA	0	no
MSTRG.52035	Z99916.1	ENSG00000236641.1,ENSG00000272798.1,I	NA	0.92	3.6E-03	ILMN_1760708	1.6E-03	A_23_P425066	9.1E-03	NA	0	no
MSTRG.52736	AL031587.5	ENSG00000278948.1	NA	0.64	0.0E+00	ILMN_1721338	1.8E-04	A_24_P336137	6.1E-03	NA	0	no
MSTRG.53419	NA	NA	CATG00000059027.1	0.85	3.3E-02	ILMN_1752837	1.8E-09	A_23_P385217	2.4E-05	NA	0	no
MSTRG.61260	AC093752.2	ENSG00000250950.1	NA	1.19	2.0E-07	ILMN_1810836	8.5E-16	A_19_P00316362	3.0E-04	NA	0	no
MSTRG.65987	SNHG4	ENSG00000281398.2	NA	0.82	1.3E-09	ILMN_1661673	4.7E-17	A_33_P3280916	5.3E-04	NA	0	no
MSTRG.69662	AL359715.3	ENSG00000260645.1,ENSG00000272129.1,I	NA	1.07	1.9E-05	ILMN_1891324	5.7E-10	A_19_P00805367	1.7E-02	NA	0	no
MSTRG.71013	AL031320.2	ENSG00000278206.1	NA	0.92	4.2E-03	ILMN_2057389	2.5E-02	A_23_P259328	8.5E-03	NA	0	no
MSTRG.72039	RBAKDN	ENSG00000273313.1	ENSG00000273313.1	0.83	1.2E-08	ILMN_2054938	1.9E-02	A_24_P469641	4.1E-02	NA	0	no
MSTRG.732	AC004824.1	ENSG00000227751.1	NA	0.84	0.0E+00	ILMN_1720124	1.5E-42	A_23_P74950	2.3E-02	NA	0	no
MSTRG.74819	AC105052.3	ENSG00000279168.2,ENSG00000272949.1	NA	0.91	1.6E-04	ILMN_1726678	2.6E-06	A_33_P3363316	2.7E-02	NA	0	no
MSTRG.75492	NA	ENSG00000237243.1	CATG00000094161.1,CATG00000094163.1	1.30	1.9E-12	ILMN_2331890	1.8E-07	A_23_P128067	1.7E-08	NA	0	no
MSTRG.774	UBXN10-AS1	ENSG00000225986.1	ENSG00000225986.1	1.48	0.0E+00	ILMN_1656867	4.6E-07	A_33_P3344156	4.1E-02	NA	0	no
MSTRG.85871	ZNF674-AS1	ENSG00000230844.2	ENSG00000230844.2	0.91	5.2E-05	ILMN_1732831	1.4E-04	A_24_P256654	9.1E-04	NA	0	no
MSTRG.87434	AC004000.1	ENSG00000237903.1	NA	0.89	6.6E-04	ILMN_2142284	1.1E-24	A_23_P125668	5.3E-03	NA	0	no
MSTRG.9145	ZEB1-AS1	ENSG00000237036.4	ENSG00000237036.4	0.90	1.1E-06	ILMN_1829989	4.1E-05	A_33_P3365856	7.9E-03	NA	0	no

Supplementary table 2. Nanodrop quantity results from RNA extraction.

Well number, Gapmer identity and experiment number	Total RNA, ng/µL, in sample	Total RNA, ng, in sample (if eluted in 20 µL)
GATA3-AS1, well 1, experiment 3	433,778	8675,56
GATA3-AS1, well 2, experiment 3	447,951	8959,02
FAM198B-AS1, well 1, experiment 3	424,62	8492,4
FAM198B-AS1, well 2, experiment 3	286,7	5734
DRAIC, well 1, experiment 3	536,229	10724,58
DRAIC, well 2, experiment 3	249,84	4996,8
Negative control, well 1, experiment 3	315,043	6300,86
Negative control, well 2, experiment 3	225,695	4513,9
Only cells, experiment 3	276,802	276,802
GATA3-AS1, well 1, experiment 2	84,642	84,642
GATA3-AS1, well 2, experiment 2	119,665	119,665
FAM198B-AS1, well 1, experiment 2	86,165	86,165
FAM198B-AS1, well 2, experiment 2	80,699	80,699
DRAIC, well 1, experiment 2	97,678	97,678
DRAIC, well 2, experiment 2	70,86	70,86
Negative control, well 1, experiment 2	116,707	116,707
Negative control, well 2, experiment 2	69,386	69,386
GATA3-AS1, well 1, experiment 1	692,074	692,074
GATA3-AS1, well 2, experiment 1	742,079	742,079
FAM198B-AS1, well 1, experiment 1	490,865	490,865
FAM198B-AS1, well 2, experiment 1	456,728	456,728
DRAIC, well 1, experiment 1	358,495	358,495
DRAIC, well 2, experiment 1	546,133	546,133
Negative control ,well 1, experiment 1	307,907	307,907

Supplementary table 3. Ct values from RT qPCR.

A. Ct values from PCR of GATA3-AS1 and control GAPDH. F1, F2 and F3 are referring to experiment 1, 2 and 3. GAP scrambled is the negative control.

	Target gene				Reference gene			
PCR efficiency	1,9942			mean	1,9364			mean
Ct values	Ct Gata3 F1	Ct Gata3 F2	Ct Gata3 F3	Ct target	Ct GAPDH F1	Ct GAPDH. F2	Ct GAPDH F3	Ct reference
GAP Scrambled	25,48622933	26,210193	25,48622933	25,72755056	16,00019033	15,1126163	15,353128	14,05187067
GAP Gata3	27,49122033	26,677194	25,944201	26,70420511	16,021264	15,1252727	14,81403633	14,36979933

B. Ct values from PCR of FAM198B and control GAPDH. F1, F2 and F3 are referring to experiment 1, 2 and 3. GAP scrambled is the negative control.

	Target gene				Reference gene			
PCR efficiency	1,9942			mean	1,9364			mean
Ct values	Ct FAM198B F1	Ct FAM198B F2	Ct FAM198B F3	Ct target	Ct GAPDH F1	Ct GAPDH. F2	Ct GAPDH F3	Ct reference
GAP scramble	27,95097133	27,91130567	28,20771933	28,02333211	15,2889165	15,5128593	15,353128	15,38496794
GAP FAM198B	27,91137967	29,32647867	29,29519067	28,84434967	15,01043033	15,4485773	14,81403633	15,09101467

C. Ct values from PCR of DRAIC and control GAPDH. F1, F2 and F3 are referring to experiment 1, 2 and 3. GAP scrambled is the negative control.

	Target gene				Reference gene			
PCR efficiency	1,9942			mean	1,9364			mean
Ct values	Ct Draic F1	Ct Draic F2	Ct Draic F3	Ct target	Ct GAPDH F1	Ct GAPDH F2	Ct GAPDH F3	Ct reference
GAP Scrambled	31,77785067	30,07997467	28,20771933	30,02184822	16,00019033	15,1126163	15,353128	14,05187067
GAP Draic	32,493982	29,85889233	31,26959033	31,20748822	16,19032067	14,0301168	14,81403633	13,67002517

Supplementary table 4. Results from CellTiter-Glo cell viability assay.

A. Experiment 1
100 000 cells/mL

Cell Death Positive control	Scrambled GapmeR Negative control	Gata3-AS1	FAM198B-AS1	DRAIC	Lipofectamine negative control	Cells/ medium negative control	Medium
290846	426001	456618	482864	503293	488146	480031	1426
368417	469322	468890	365089	494941	472835	475471	53594
338714	469443	432039	472676	382249	456015	438762	40642

B. Experiment 2
100 000 cells/mL

Cell Death Positive control	Scrambled GapmeR Negative control	Gata3-AS1	FAM198B-AS1	DRAIC	Lipofectamine negative control	Cells/ medium negative control	Medium
105347	586527	540301	606256	590749	580284	543283	908108
150639	579725	577526	578450	568583	585311	536780	855199
193811	564575	552554	630805	570030	590206	561192	847770

C. Experiment 3
80 000 cells/mL

Cell Death Positive control	Scrambled GapmeR Negative control	Gata3-AS1	FAM198B-AS1	DRAIC	Lipofectamine negative control	Cells/ medium negative control	Medium
207809	643485	657965	685986	674809	676343	547078	642022
138744	659223	666326	691484	687477	707046	605069	578920
254329	688614	671807	699896	699888	702112	565681	620553