DISCOVERY AND CHARACTERIZATION OF A CANCER STEM CELL-ENRICHED LONG NON-CODING RNA IN TRIPLE-NEGATIVE BREAST CANCER

by

Dejan Vidovic

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
July 2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Triple-negative breast cancer (TNBC) patients face poor prognoses, possibly due to the presence of highly tumorigenic cancer stem cells (CSCs) within their tumors. The search for possible CSC targets is limited, but may be aided by identification of novel unstudied factors that regulate stemness, such as long noncoding RNAs (lncRNAs), which are often highly overexpressed in cancer making them attractive therapeutic targets. Here, we identify oncogenic LINC00284 as highly enriched in TNBC and CSCs, and for the first time, target this CSC-enriched lncRNA *in vivo* to slow tumor growth and kill CSC populations. We also functionally interrogate LINC00284 to identify its yet unknown mechanism of action in cancer, revealing a complex mechanism involved in gene expression regulation through interactions with multifunctional nuclear protein SND1.

LIST OF ABBREVIATIONS USED

EMT	epithelial mesenchymal transition
IDC	invasive ductal carcinoma
ILC	invasive lobular carcinoma
DCIS	ductal carcinoma in situ
LCIS	lobular carcinoma in situ
NST	invasive carcinoma of no special type
ER	estrogen receptor
PR	progesterone receptor
HER2	human epidermal growth factor 2 receptor
TNBC	triple-negative breast cancer
CSCs	cancer stem cell(s)
TICs	tumor-initiating cell(s)
ALDH	aldehyde dehydrogenase
ALDH1A3	aldehyde dehydrogenase isoform 1A3
ALDE	Aldefluor activity
DEAB	diethylaminobenzaldehyde
ATRA	all-trans retinoic acid
RA	retinoic acid
ncRNA	non-coding RNA
lncRNA	long non-coding RNA
lincRNA	long intergenic non-coding RNA
miRNA	micro RNA

ceRNA	competitive endogenous RNA
PRC2	polycomb repressive complex 2
shRNA	short hairpin RNA
RNAi	RNA interference
siRNA	small interfering RNA
ASO	anti-sense oligonucleotide
PS	phosphorothioate backbone
LNA	locked nucleic acid
ATCC	American Type Culture Collection
PDX	patient-derived xenograft
FACS	fluorescence-activated cell sorting
RCF	relative centrifugal force
PBS	phosphate-buffered salid
BAAA	BODIPY-aminoacetaldehyde
APC	allophycocyanin
SSC	side scatter
FSC	forward scatter
CCAC	Canadian Council for Animal Care
RT-QPCR	real-time quantitative polymerase chain reaction
cDNA	complementary DNA
gDNA	genomic DNA
QPCR	quantitative polymerase chain reaction
ChIRP	chromatin isolation by RNA purification

RIP	RNA immunoprecipitation
PIC	protease inhibitor cocktail
CCLE	Cancer Cell Line Encyclopedia
GO	Gene Ontology term(s)
GSEA	Gene Set Enrichment Analysis
RNA-seq	RNA-sequencing
TCGA	The Cancer Genome Atlas
FAO	fatty acid oxidation
SND1	staphylococcal nuclease domain-containing 1
MTDH	metadherin
ChIP	chromatin immunoprecipitation

ACKNOWLEDGEMENTS

First, I would like to acknowledge my supervisor, Dr. Paola Marcato, for her endless support throughout the last four years that I have been working with her, in both my undergraduate degree and my Masters' degree. Paola has allowed me to delve into a project that has given me an enormous passion for science, which will likely follow me throughout the rest of my life. As my mentor, she has taught me countless valuable lessons both inside and outside of the lab.

I would also like to thank my graduate committee members, Dr. Shashi Gujar and Dr. Roy Duncan for their valuable input on this project. A special thanks goes to lab members Thomas Huynh, Cheryl Dean, Brianne Cruickshank, Mo Sultan, and collaborator Prathyusha Konda for all their insight and support. Lastly, I would like to acknowledge and extend an enormous thanks to my partner, Ana Jemcov, who has been one of the sole moral supports during this journey, and whose constant patience and support has made this all possible.

CHAPTER 1 INTRODUCTION

1.1 Cancer

Cancer is fundamentally a genetic disease: the result of dysregulations that arise in the genome and prevent the normal functioning of our genes, ultimately leading to uncontrolled cellular proliferation. While all cancers share this general definition, in reality cancer is a group of very diverse pathologies. Two patients of the same age with the same diagnosis of breast cancer could be treated with the same drug, and in an all too familiar scenario, one may survive and the other may not. Therefore, cancer is truly a complex disease: no single mutation alone can be attributed to be the cause of cancer. Rather, cancer arises from a collection of mutations that push a cell into a tumorigenic process. Often, these mutations arise in three kinds of genes: proto-oncogenes, tumor suppressors, and stability genes[1].

In normal cells, proto-oncogenes are genes which are tightly regulated and are therefore often lowly expressed[1]. In many cases, these proto-oncogenes normally promote cellular division, cell survival, and proliferative signaling[1,2]. Through chromosomal translocations, gene amplifications, or point mutations in their gene bodies, these oncogenes can become constitutively turned on – leading cells down the path to resist cell death, sustain proliferative signaling, and enable replicative immortality[2]. However, our genome has safeguards in place: excessive oncogene signaling may lead cells to die through apoptosis or oncogene-induced senescence[2,3]. For example, Ras oncogene overexpression in the presence of normal tumor suppressors p53 and p16 leads to permanent cell cycle arrest; when p53 and p16 are lost, however, the oncogenic effects of Ras prevail and cancer progresses[4]. Thus, mutations in oncogenes alone are not

enough to lead to tumorigenesis – they must be concurrent with the loss of tumor suppressors.

Tumor suppressors, as their name suggests, are the key drivers of anti-tumor cellular programs, and are often expressed at basal levels in most cells. However, tumor suppressor genes may acquire loss-of-function mutations that render them unable to perform their normal function[1]. For example, the multifunctional tumor suppressor p53, upon sensing DNA damage or oncogenic stress, induces cell cycle arrest and apoptosis [5,6]. Therefore, when a loss-of-function mutation in p53 occurs, cells are unable to recognize when to arrest, and fail to undergo apoptosis. This, coupled with oncogene activation, can lead not just to a multistep tumorigenic process, but also to widespread, irreversible chromosomal damage[7].

Stability genes include genes which normally uphold genomic stability and integrity. Normal cell division results in hundreds to thousands of new mutations within our genomes[8]. If unchecked, the presence of these mutations means cancer would likely arise at a higher rate. Stability genes encompass those involved in DNA repair processes, such as mismatch repair, base-excision repair, and nucleotide-excision repair[1]. A classic example of a stability gene mutation is that of BRCA1, in which germline mutations predispose women to breast and ovarian cancer. Mutations in BRCA1 prevent formation of a DNA damage-sensing complex, which prevents cells from recognizing double-strand chromosomal breaks[9,10]. Cells with BRCA1 mutations will ultimately fail to repair these breaks, leading to widespread DNA mutation; which unfortunately, may arise in the proto-oncogenes and tumor suppressors described above.

Mutations in proto-oncogenes, tumor suppressors, and stability genes are what allow cells to transition from a normal to a cancerous state. The first mutations in important proto-oncogenes and tumor suppressors promote clonal expansion of cells[1]. Then, in a Darwinian process, further mutations give cells selective advantages such as: continuous proliferative signaling, apoptosis resistance, replicative immortality, evading growth suppressors, and promoting invasiveness[2]. Mutations, however, are not solely what promote these selective advantages. A portion of these advantages come from aberrant pathways that cancer cells hijack to promote progression. A classic example is the epithelial-to-mesenchymal transition (EMT) pathway, which in the normal developmental context, is responsible for movement of cells into different embryonic regions to form differentiated highly organized tissues[11]. Cancer cells, however, hijack the EMT pathway to promote metastasis and spread to different organs[11]. Thus, cancer arises from an amalgamation of successive advantageous mutations in proto-oncogenes, tumor suppressors, and stability genes, complemented by hijacked gene expression programs. For these reasons, it is clear why cancer is not one single disease, but rather a collection of complex and diverse pathologies: depending on the genes mutated and pathways dysregulated, two types of cancer can be vastly different.

1.2 Breast cancer

Breast cancer is the most common cancer affecting Canadian women today. One in eight women are predicted to develop breast cancer during their lifetime, representing approximately 25% of all new diagnosed cancer cases in women[12]. Furthermore, it carries a high mortality rate, with approximately 13% of all cancer deaths in women being

attributed to breast cancer[12]. Fortunately, advances in screening, improved diagnoses, and research has significantly improved patient outcomes.

Breast cancer can generally be stratified into classes based on histological and morphological assessment: the most common are invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), ductal carcinoma *in situ* (DCIS), and lobular carcinoma *in situ* (LCIS). ILC arises from the milk-producing lobules, and is characterized by diffuse infiltrative growth[13]. Similarly, IDC, sometimes called invasive carcinoma of no special type (NST), is an invasive carcinoma arising from some part of the milk-carrying ducts of the breast[14]. Lastly, *in situ* carcinomas are, by definition, precursor lesions – the tumor has not yet invaded into other regions in the breast. With enough time, however, they may become invasive[14]. The histopathological typing of tumors in this way is crucial for understanding patient prognosis, clinical features, and predicting responses to treatment[14].

Besides histopathological typing based on morphology, patients' tumors are often subtyped based on expression of hormone receptors. Tumors are classified based on their histological expression of the estrogen receptor (ER), progesterone receptor (PR) and/or human epidermal growth factor 2 (HER2) receptor. Subtyping by hormone receptor status has clinical value, as different hormone subtypes dictate treatment and prognosis[15]. ER^+ tumors can be favorably treated with ER antagonists, such as tamoxifen. Similarly, $HER2^+$ tumors can be treated with trastuzumab, a monoclonal antibody against the HER2 receptor. Some patients have tumors that lack cell surface expression of all three receptors, termed triple-negative breast cancers (TNBC). These patients face a disproportionately poor prognosis, in part because these tumors are resistant to drugs like

trastuzumab or tamoxifen[16]. Thus, subtyping based on hormone receptor status is an attempt at targeted therapy; using a drug that is specifically potent in inhibiting a patient's unique tumor, as opposed to treating the patient with a non-specific cytotoxic chemotherapeutic. In the case of TNBC patients, hormone receptor subtyping fails to identify therapeutic targets.

Even within similar hormone receptor subtypes, there remains a fair amount of phenotypic diversity. Most of this diversity is due to differential gene expression patterns and pathways[17]. Subtyping tumors based on gene expression patterns reveals five distinct subtypes, each associated with a prognoses and gene expression profile; basal-like, luminal A, luminal B, Her2-enriched, and claudin-low[17,18]. Luminal A and luminal B tumors tend to be ER^+/PR^+ by histology[19]; therefore, targetable with ER antagonists. Similarly, tumors subtyped by gene expression as Her2-enriched tend to overexpress HER2 histologically[19] and are targetable with trastuzumab. Basal-like and claudin-low tumors are most often TNBC by histology[20], and, like TNBC, generally lack therapeutic targets. However, with the inception of gene expression profiling and subtyping, pathways that are enriched in these subtypes can be exploited to identify novel targets. Of note, dysregulated pathways enriched in basal-like breast cancers (which represents 70-80% of all TNBC[21,22]) can be studied, giving promise to identifying TNBC-specific treatments.

The successive subtyping of breast cancer from histopathological to histological hormone receptor status to gene expression pathways highlights the complexity behind this disease. Furthermore, it reveals the need for novel therapies for those subtypes for which there are currently none, such as basal-like and TNBC. These patients suffer most:

not only do their tumors progress to high grade rapidly and are more resilient than other breast tumors[23], but the non-specific cytotoxic drugs currently used may actually worsen prognosis[16]. With the exception of olaparib, a small-molecule targeted therapy approved for treatment of BRCA1-mutant TNBC[24] (ranging from 10-20% of TNBC[21]), currently no other targeted therapies exist for the treatment of basal-like or TNBC.

1.3 Barriers to effective treatment: cancer stem cells

A major barrier to successful treatment of cancer is the presence of specialized cells within tumors called cancer stem cells (CSCs). Many tumors do not generally exist as monoclonal populations of one "cancerous" cell, but rather, are a heterogeneous population of differentiated cells originating from a rare, highly tumorigenic, non-differentiated CSC[25]. Thus, CSCs within most tumors share the property of self-renewal and differentiation capacity. Additionally, CSCs may readily mediate tumor recurrence, as they tend to be resistant to radiotherapies, chemotherapies, and immunotherapies[26–29].

The true definition of a CSC has been contentious in recent years. Some early evidence showed that CSCs can be malignant forms of normal stem cells, where mutagenic events have occurred to transform them into their malignant form[30]. Others have suggested, especially within most solid tumors, that CSCs are not truly transformed normal stem cells but rather a more tumorigenic population of cells with properties of stemness (i.e. differentiation, self-renewal, and drug-resistance), often called tumor-initiating cells (TICs; the term is generally interchangeable with CSC).

However, the most current model proposed is that CSCs are indirectly derived from normal stem cells through successive processes of de-differentiation[31]. First, normal stem cells differentiate into a transit-amplifying pool (the first set of differentiated cells), which, as a large pool of cells, acquire tumorigenic mutations and survive in a Darwinian manner (i.e. the cell with the most advantageous traits will survive). These mutated cells then de-differentiate back into a more stem-like form, upon which the process is repeated. Each intermediate stem cell product of de-differentiation may give rise to its own downstream differentiated population of cells, thus accounting for the heterogeneity present in tumors. Ultimately, at the end of multiple rounds of de-differentiation, the stem cell with the most advantageous cancerous mutations may give rise to a highly aggressive tumor[31].

To date, highly tumorigenic CSCs have been identified in breast[32,33], brain[34], ovarian[35], liver[36], lung[37], colorectal[38], pancreatic[39], thyroid[40], and other solid tumors[41]. These CSCs are identified by using flow cytometry to isolate tumorigenic cell populations using cell surface markers. Isolated putative CSCs are then implanted into mice at low concentrations, to prove that they can form a heterogeneous tumor – a hallmark of stemness. Tumors are then harvested and re-passaged to screen for secondary tumor formation, thus proving self-renewal and differentiation capacity. Recent evidence shows that CSCs may retain some normal embryonic stem cell gene signatures throughout the de-differentiation process[41]. Furthermore, pluripotent stem cells markers Sox2, Nanog, and Oct4 are often overexpressed in many different CSCs[42]. Importantly, however, in most solid tumors there are CSC markers that are specific to only the CSC population (i.e. not expressed on non-CSC cancer cells or the normal stem cell

counterparts). Some of these CSC-specific markers only become expressed once CSCs have become markedly tumorigenic[43].

1.4 Breast cancer stem cells

Within heterogeneous breast tumors, CSCs are enriched within populations of cells defined by expression of cell surface markers ($CD24^{-}CD44^{+}$)[32], or by a more functional readout, increased aldehyde dehydrogenase (ALDH) enzyme activity ($Aldefluor^{high}$)[33]. In breast cancer, high levels of these CSC markers significantly predicts clinical response and metastases[33,44–46]. Of note, the $Aldefluor^{high}$ phenotype is a significant indicator of breast cancer recurrence[47–49], potentially due to the role of breast CSCs in chemoresistance[50–53], radioresistance[51], and immunotherapy resistance[28]. Thus, patients treated with therapies that are not CSC-targeting will see their tumors shrink, but still recur due to the persistence of the CSC pool.

In general, CSCs exist within all subtypes of breast cancer[54]. However, evidence suggests that basal-like/TNBC in particular is enriched for CSCs; higher numbers of CSCs are often present within these tumors[54–62]. With the importance of tumor heterogeneity in basal-like/TNBC progression and metastasis[23], it may be that the poor survival and prognoses of these patients can be attributed, at least in some part, to the enrichment of CSCs within these tumors. Therefore, successful treatment of basal-like/TNBC tumors and prevention of recurrence may be mediated by using therapies that specifically target breast CSCs.

1.5 Novel therapies for breast cancer: targeting CSCs

Numerous attempts have been made in recent years to develop CSC-specific therapies for breast cancer. Many of these potential therapies have attempted to inhibit key stemness pathways in CSCs, such as the Wnt signaling pathway, which normally promotes self-renewal and cell fate determination[63]. For example, a recent approach aimed to use salinomycin, a Wnt inhibitor, along with paclitaxel, a non-specific chemotherapeutic, to inhibit CSC and non-CSC pools, respectively[64]. The drugs were loaded onto nanoparticles, to promote maximum half-life and tumor targeting effects, thus maximizing Wnt inhibition and DNA damage to both CSCs and non-CSCs alike.

Unfortunately, this approach still suffers from some major drawbacks. While salinomycin alone was enough to prevent metastases, mice still died from their primary tumors[64]. Paclitaxel was required to de-bulk the non-CSCs in the primary tumor, and was necessary for therapeutic success. The heavy reliance on a non-specific chemotherapeutic like paclitaxel does therefore not necessarily reflect a truly targeted therapy. Furthermore, salinomycin has moderate toxic effects in normal fibroblasts[65], and some reports have shown devastating neuronal[66,67] and systemic toxicities in horses[66], turkeys[68], and pigs[69], raising the question of drug efficacy and safety. Considering the Wnt pathway is vital for survival of normal embryonic stem cells[70], a Wnt inhibitor coupled to paclitaxel may not be the most targeted approach, and will certainly need much further investigation before being used clinically.

Clearly, there is a need for robust CSC biomarkers which both: identify CSCs over non-CSCs and normal tissue; and are “druggable” or have “druggable” effectors. Approaches that target stemness-enriched pathways, such as the Wnt pathway mentioned

above, therefore may not be ideal. As mentioned previously, breast CSCs are identified using CD44 positivity or high ALDH activity. CD44 is a cell surface marker involved in cell adhesion, and is thus less of a druggable CSC marker than the enzyme ALDH1[33]. Regardless, some recent work has tried to develop CD44-based inhibitors to abrogate $CD24^{-}CD44^{+}$ breast CSCs with some, but limited, pre-clinical success[71–73].

In general, the ALDH superfamily of 19 ALDH enzymes function in removal of toxic aldehydes from cells. Of the 19 enzymes, each individual isoform has its own substrate specificity and specific function[74]. In breast cancer, the high ALDH activity that identifies CSCs is primarily through isoform 1A3 (ALDH1A3)[46]. Importantly, ALDH1A3 has a functional role in terms of breast cancer progression and chemoresistance[75,76]. Recent attempts have been made to inhibit the ALDH1A3 enzyme to varied success. Non-specific ALDH inhibitor diethylaminobenzaldehyde (DEAB) was able to decrease breast tumor formation *in vivo* in mice[77]; however, DEAB inhibition is preferential for other ALDH members ALDH1A1, ALDH1A2, and ALDH2[78], raising concerns for specificity. Another group achieved ALDH1A3 inhibition using disulfiram, an anti-alcoholism drug, which abrogated CSC numbers and reversed chemoresistance in one cell line model of TNBC[79]. However, in other models of TNBC, disulfiram was not successful in ALDH1A3 inhibition and did not decrease Aldefluor activity[75]. While the drug citral was successfully able to specifically abolish ALDH1A3-mediated $Aldefluor^{high}$ TNBC CSCs, the use of citral *in vivo* was limited, and nanoparticle encapsulation was required to see a moderate effect on CSC inhibition[75]. Therefore, novel therapeutic approaches for ALDH1A3 inhibition are required to prevent CSC survival.

To identify novel ALDH1A3 inhibitors, the exact mechanism behind ALDH1A3 function in cancer progression, chemoresistance, and conferred benefit upon CSCs must be elucidated, as it is still currently unclear. However, it is likely due to ALDH1A3-mediated gene expression changes[76,80,81]. As an enzyme, ALDH1A3 metabolizes the production of all-trans retinoic acid (ATRA or RA) from retinal. ATRA, a signaling molecule, binds to nuclear hormone receptors bound to ATRA-response elements in gene promoters to regulate expression of hundreds of genes, to ultimately alter cellular fate[74,82,83]. Recent evidence has shown that ALDH1A3-regulated gene expression changes involves additional, yet unknown effectors[80]. To therefore successfully target ALDH1A3 for basal-like/TNBC treatment[84], identification and inhibition of these unknown effectors must first be performed. With the rise of whole transcriptome analyses, the importance of non-coding RNAs (ncRNAs) is being realized. Therefore, it is possible some of these unknown effectors are ncRNAs.

1.6 Long noncoding RNAs & functions

The classic central dogma of molecular biology defines the final fate of all genetic information: DNA is transcribed to RNA, and then translated into functional proteins. Outside of proteins, the role of some abundant functional ncRNAs, such as ribosomal RNA, transfer RNA, and nuclear/nucleolar RNAs had been known for some time, due to their vital role in key cellular processes such as transcription or translation. It was assumed that if the rest of the genome did not code for protein, then it was simply non-coding "junk" DNA[85]. Only in the last two decades, since the inception of whole-genome technologies and tiling arrays, has it become clear that many of these genomic "junk"

regions actually possess genes which code for RNAs that are not translated but may still exhibit important, more nuanced cellular functions[85]. These ncRNA genes can generally be categorized into long non-coding RNAs (lncRNAs) and microRNA (miRNAs); we will focus on the former.

Long non-coding RNAs (lncRNAs) are generally defined as non-protein coding transcripts that are greater than 200 nucleotides in length. Genes that transcribe for lncRNAs are present throughout the genome, sometimes even within protein-coding genes, as sense or antisense transcripts that overlap introns or exons. They are more commonly distinct transcriptional units with their own promoter and/or enhancer regions, as intergenic lncRNAs (often called large intergenic noncoding RNA, or lincRNA)[85,86]. It is estimated that within our genome, there are just under 28,000 lncRNA genes, a number that surpasses the estimated 20,000 protein-coding genes[85,87,88]. However, the functions of only a few hundred are known[89], reflecting the need for increased understanding of lncRNA biology.

The function of a lncRNA is often dependent on its sub-cellular location (i.e. nucleus or cytoplasm)[90]. So far, many nuclear lncRNAs have been shown to function in modulating gene expression either in *cis* or in *trans*, by acting as transcriptional guides, transcription factor decoys, or as scaffolds for molecular interactions[91]. Cytoplasmic lncRNA often play roles in modulating: mRNA stability; altering translation control; and modulating miRNA levels, by either acting as competitive endogenous RNAs (ceRNA) to inhibit miRNA function, or as miRNA precursors[92].

Many of the first lncRNAs discovered were identified to play roles in vital cellular developmental processes. For example, the lncRNA XIST is responsible for X

chromosome inactivation in humans, a process which is required for proper embryogenesis of females to occur[93,94]. XIST acts as a large molecular scaffold, to coat the inactive X and bring together chromatin-silencing proteins to maintain the inactive state[95]. Another important developmentally involved lncRNA is HOTAIR, which regulates the expression of the HOX family of genes during development[96]. Tightly regulated HOX gene expression is essential for dictating the position of cells during embryogenesis; HOTAIR acts as a guide to the polycomb repressive complex 2 (PRC2) to epigenetically silence certain HOX loci as development progresses[96].

An overwhelming number of lncRNA have been identified as being dysregulated in cancer[97,98]. Akin to how cancer cells hijack the EMT pathway from the developmental context to promote metastasis, they may also hijack lncRNAs from lncRNA-mediated developmental pathways to use them to their advantage to aid in tumorigenesis[99], metastasis[100], and drug resistance[101]. For example, breast cancer cells overexpress HOTAIR to cause recruitment of PRC2 to hundreds of other genes, causing epigenetic silencing of PRC2-targeted genes. Many of those targets are tumor suppressors whose silencing ultimately leads to metastasis[100]. With the recent increased understanding of the role of lncRNAs in cancer and tumorigenesis came the discovery of two intrinsic properties of most lncRNAs: they tend to be highly expressed in certain cancers, and they exhibit polarized tissue- or cell-specific gene expression[102].

1.7 Long noncoding RNAs in CSCs

The search for basal-like/TNBC CSC targets has focused mainly on proteins (i.e. Wnt inhibitors, CD44 inhibitors, ALDH1A3 inhibitors). However, protein coding genes

only represent 2% of the total genome, with up to 70% of the genome ultimately transcribed into RNA[86,103]. It is therefore possible that some of the effectors responsible for basal-like/TNBC progression (mediated by ALDH1A3$^+$ CSCs) are lncRNAs. Coupled with their property of being overexpressed in cancer and specific cell types, targeting these CSC-enriched oncogenic lncRNAs could potentially be used as a novel, highly specific therapeutic approach. However, very few breast CSC-enriched lncRNAs have so far been discovered.

Studies exploring the landscape of CSC-enriched lncRNAs in breast cancer have generally not employed *de novo* lncRNA discovery approaches. Rather, they rely on exploring whether already functionally annotated lncRNAs (i.e. the several hundreds of lncRNAs whose function has been identified) play a role in stemness pathways in breast CSCs (i.e. a top-down approach). For example, HOTAIR was recently found to be upregulated in breast cancer mammospheres (an *in vitro* readout of stemness), promoting proliferation and self-renewal of the mammosphere-forming cells[104]. The lncRNA LINC-ROR, originally studied in induced pluripotent stem cells[105], was able to induce EMT when overexpressed in normal breast cells, and lead to higher levels of CD44 positivity[106]. Similarly, the lncRNA TUNAR, first discovered as regulating pluripotency and neural lineage commitment in mice[107], additionally upregulates pluripotency factor SOX2 in breast cancer cells to promote EMT and metastasis[108]. While the importance of these lncRNAs in stemness-related pathways is clear, their enrichment has not been directly measured in breast CSC populations.

The use of bottom-up approaches when identifying CSC-enriched lncRNAs is much less common, with very few examples in the literature. The lncRNA GAS1RR was

discovered *de novo* by microarray as a highly upregulated transcript in mammospheres positive for Twist (an EMT marker) and ALDH1[109]. Tumor cells deficient in GAS1RR grew much slower *in vivo* than those with endogenous levels of GAS1RR, and expressed lower levels of stemness markers. A similar microarray-based approach was used to identify the lncRNA H19, which was the first lncRNA to ever be discovered[110], as the most upregulated lncRNA in ALDH1$^+$ breast CSC populations[111,112]. Importantly, H19 regulates breast CSC stemness through acting as a ceRNA to sponge the miRNA let-7, causing increases in expression of pluripotency factor LIN28[112]. H19 expression was vital for maintenance of breast CSCs *in vivo*, suggesting that targeting this lncRNA (or its ceRNA network) could prove to be an effective therapeutic strategy[112].

Despite the recent work identifying these lncRNAs as being CSC-enriched and potential targets, no one has yet administered a drug to inhibit a CSC-enriched lncRNA *in vivo* in a breast tumor model. This is likely attributed to both the novelty of the field and the difficulty of lncRNA inhibition *in vivo*.

1.8 Noncoding RNA therapies

The prospect of lncRNA inhibition *in vivo* to abrogate tumor formation or growth has been met with significant difficulties. While *in vitro* inhibition of a lncRNA may be achieved using short hairpin RNAs (shRNAs) to induce RNA interference pathways (RNAi), this approach cannot work *in vivo* for a number of reasons. First, shRNAs function primarily through a cytoplasmic pathway, meaning nuclear lncRNAs will be relatively unaffected[113]. Furthermore, shRNA-based modulation of expression is often achieved using viral vectors to induce a stable (i.e. permanent) transfection, causing

concerns for immune-related toxicities and off-target effects[114]. Others have tried using short interfering RNAs (siRNAs), another form of RNAi, which has been met with similar difficulties – off-target effects and poor stability *in vivo*[115]. Additionally, siRNAs are not likely to be efficacious in knockdown of nuclear lncRNAs, since they also function primarily through the cytoplasmic RNAi pathway[115].

Among the most studied RNA therapy is antisense oligonucleotides (ASOs), DNA molecules that inhibit RNAs by binding to them antisense to form a DNA-RNA duplex, which is recognized by RNase H and degraded[113]. Forty years ago, the first ASOs used were successful in preventing RNA maturation, but were unmodified[116] and thus had very poor stability and efficacy in serum, and were often degraded by nucleases[117]. Over the years, ASOs have undergone many modifications to improve their stability, antisense activity, and efficacy. First, a phosphorothioate (PS) backbone was added to confer nuclease resistance to the ASO; in doing so, the melting temperature of the ASO was decreased, lowering efficacy[113]. Further ASO evolutions use a central DNA "gap" region that is flanked by the 2' regions[115], to significantly improve melting temperature stability. The next generation of ASOs utilized synthetic nucleotides containing a methylene bridge between the 2'-O of the ribose ring and 4'-C of the nucleic acid, causing a "locked" formation[118]. The combination of a central DNA gap, PS backbone, and locked nucleic acids (LNA) creates an ASO with high efficacy, stability, and nuclease resistance[115] – these synthetic, highly potent ASOs are often called LNA GapmeRs.

Because the development of effective *in vivo* LNA GapmeRs has only occurred in the last few years, there are only a few examples of their successful use in inhibiting

lncRNAs *in vivo*. An LNA GapmeR targeting the highly oncogenic lncRNA MALAT1 was used to dramatically decrease lung tumor metastases *in vivo*[119]. Similarly, LNA GapmeRs targeting BCAR4, another oncogenic lncRNA that promotes Hedgehog-mediated cell migration, were successful in significantly suppressing breast cancer metastases *in vivo*[120]. Another study identified a novel lncRNA, SAMMSON as being critical for melanoma tumorigenesis, and administered LNA GapmeRs *in vivo* to mice bearing patient-derived melanoma tumors. Tumors treated with GapmeRs against SAMMSON were significantly smaller and contained higher proportions of dead cells compared to control tumors[101]. Together, the evidence shows that successful anti-cancer targeting of lncRNAs pre-clinically is possible, and if the correct lncRNA target is identified, it may be a viable therapeutic approach.

1.9 Rationale & hypothesis

Ultimately, we aim to discover novel oncogenic lncRNAs that mediate basal-like/TNBC progression and CSC maintenance, which may be used potential therapeutic targets for basal-like/TNBC. We hypothesize that our screen will reveal at least one basal-like/TNBC-enriched oncogenic lncRNA that may potentiate the role of ALDH1A3 in CSC function. Furthermore, any hypothetical lncRNA target identified will likely be enriched in ALDH1A3$^+$ CSCs and will be associated with worse patient outcomes. Therefore, to generate a shortlist of lncRNAs for functional study, we developed the following prioritization strategy: first, highly expressed in basal-like/TNBC patients' tumors; second, highly enriched in ALDH1A3$^+$ breast CSCs; and third, demonstrably associated with worse patient outcomes. We hope to functionally characterize the

mechanism behind our hypothetical lncRNA target, and assess its targetability pre-clinically.

CHAPTER 2 MATERIALS & METHODS

2.10 Cell lines, cell culture, patient-derived xenografts

BT20, BT549, HCC70, MDA-MB-453, HCC1599, HCC1143, MCF7, SKBR3, HCC1937, HCC38, HS578T, MDA-MB-231, MDA-MB-436, HCC1806, HCC1187, HCC1599, BT474, T47D, HS578Bst, and MCF10A cell lines were obtained from the American Type Culture Collection (ATCC) and cultured as per their recommendations. SUM149, SUM159, and SUM1315MO cells were obtained from BioIVT (previously Asterand) and cultured as per their recommendations.

The patient-derived xenograft (PDX), PDX 7482 (a TNBC PDX) was obtained as a frozen sample from the Dr. Michael Lewis laboratory (Baylor College of Medicine). PDX 7482, upon receipt, was surgically engrafted into the 2nd thoracic mammary fat pad of 8-week old NOD/SCID female mice, and subsequently expanded for experimental use. Additional samples of PDX 7482 were preserved in liquid nitrogen.

For retinoic acid treatment, cell lines were cultured to approximately 40% confluency, treated with 100 nM all-trans retinoic acid (ATRA; Sigma-Aldrich) for 24 hours, and then collected in TRIzol reagent (Invitrogen Thermo Fisher Scientific) for RNA purification using a RNA purification kit (Invitrogen Thermo Fisher Scientific) as per manufacturer's instructions.

2.11 Retrovirus stable transfection

All stably transfected cell lines with plasmids were puromycin-resistant and were cultured in the presence of 0.25 ug/mL puromycin (Sigma-Aldrich). Stable SUM149 and MDA-MB-468 ALDH1A3 knockdown clones were generated as previously described [76]

[46], using two pSMP-shRNA clones (shRNA 1, shRNA 3). Knockdown of ALDH1A3 was confirmed using western blot and RT-QPCR.

In order to generate stable ALDH1A3-overexpressing MCF7 cells, pMSCVpuro-ALDH1A3 plasmids previously generated[76] were packaged into retroviral particles as previously described[76] and used to transduce MCF7 cells. Overexpression of ALDH1A3 was confirmed using western blot and RT-QPCR.

2.12 GapmeR treatments, cell proliferation and apoptosis assays

For all *in vitro* GapmeR treatments, GapmeRs were mixed with OptiMEM reduced serum media (Invitrogen Thermo Fisher Scientific) and TransIT-BrCa transfection reagent at a ratio of 10:1:2 parts OptiMEM:GapmeR:TransIT, to a final treatment concentration of approximately 15 nM GapmeR.

To quantify the GapmeR-mediated decrease in lncRNA expression, SUM149, MDA-MB-468, and MCF7 cells were treated with control GapmeR, GapmeR#3, or GapmeR#4, and then collected in TRIzol reagent 48 hours after treatment for RNA purification and expression quantification via RT-QPCR.

For cell proliferation experiments, SUM149, MDA-MB-468, and MCF7 cells were seeded at optimized cell numbers (SUM149, 1 x 10^6; MDA-MB-468, 5 x 10^5; MCF7, 1 x 10^6) and treated with GapmeRs then counted 48 hours after the endpoint (SUM149 1 treatment; MDA-MB-468, 3 treatments; MCF7, 3 treatments).

For apoptosis assays, the same treatment protocol was used for each cell line as for the cell proliferation experiments, but cells were instead collected, washed in PBS, and then resuspended in annexin V binding buffer (components in Table 2). The cells

were then incubated with annexin V conjugated to Alexa-Fluor 488 (Invitrogen Thermo Fisher Scientific) at room temperature for 30 minutes. Following the incubation, we added 7-AAD viability marker at a final dilution of 1/25. Flow cytometry was then performed and data collected using a FACSCalibur (BD BioSciences), and processed with FCSExpress 4 RE software (De Novo Software).

2.13 Western blots

Following trypsinization, cells were washed with ice cold PBS and resuspended in an appropriate volume of ice-cold RIPA buffer supplemented with PIC and phosphatase inhibitors (Sigma). For ALDH1A3 blots, protein lysates were centrifuged to fractionate nuclei from cytoplasm, and cytoplasmic fraction was used. For SND1 blots, the nuclear fraction was used, and nuclei were sonicated for 5 mins under the following conditions; temperature = 4˚C, amplitude = 70%, ON time = 7 min 30 seconds; cycles = 15S ON, 45S OFF. The protein lysates (cytoplasmic or nuclear) were then quantified using a BCA Protein assay kit (Thermo Scientific). Lysates were boiled in Laemmli buffer at 95˚C for 5 min, and then 20 ug loaded into a 4-15% Mini-PROTEAN TGX Stain-Free gel (BioRad) and run at 100V for 1 hour. The gels were then transferred to a PVDF membrane using a TurboBlot system (BioRad). The membranes were probed using primary antibodies: rabbit anti-human SND1 antibody (Abcam, ab65078, 1/1000; 10 min transfer time), mouse anti-human ALDH1A3 (Origene, clone 4E8, 1/1000), or IgG antibody (Diagenode, C15410206, 1/1000). Secondary species-specific horseradish peroxidase (HRP)-conjugated goat anti-mouse IgG (1/1000) were used to detect protein levels by chemiluminescence following staining with ECL substrate (BioRad) on a

ChemiDoc imager (BioRad). For a loading control, total protein load was assessed using the stain-free luminescence ability of the PROTEAN TGX gels.

2.14 Aldefluor sorting

Cell sorting of distinct cell populations based on Aldefluor activity was performed for tumorigenicity assays and RNA collection as per the manufacturer's instructions (Aldefluor assay kit, StemCell Technologies) and as previously described[28]. Briefly, prior to cell sorting with a FACSAria cell sorter (BD BioSciences), PDX tumors passaged in mice were minced mechanically and then digested with 15 mL 225U/mL collagenase III (BioShop) for one hour at 37°C with rotation. The mix was then passed through a 70 uM strainer, pelleted at 500 relative centrifugal force (RCF) for 5 min, and resuspended in 5 mL of red blood cell lysis buffer (components in Table 2). Cells were incubated for 2 min, then pelleted at 500RCF and with phosphate buffered saline (PBS) before resuspension in 10 mL Aldefluor buffer and staining with BODIPY® - aminoacetaldehyde (BAAA) substrate (as per manufacturer's instructions) and allophycocyanin (APC)-conjugated mouse lineage-specific antibodies (anti-H2Kd). Cells were then stained with 7-AAD (BioLegend) to a final dilution of 1/10, a cell viability dye used to discard dead cells during sorting. To eliminate debris, side scatter (SSC) and forward scatter (FSC) were used. A control sample treated with Aldefluor activity-inhibitor diethylaminobenzaldehyde (DEAB; as per manufacturer's instructions) was included to ensure that an Aldefluor+ population of cells had been correctly gated. The same experimental workflow was followed for single cell suspensions of SUM149 cell line; however, mouse-specific lineage antibodies were not used, and 7AAD concentration

was adjusted to 1/25. Aldefluor+ and Aldefluor- cells destined for RNA purification were centrifuged at 600RCF and resuspended in TRIzol reagent.

2.15 Animal GapmeR studies

All animal studies detailed in this manuscript have been conducted in accordance with the ethical standards and according to the Declaration of Helsinki and according to national and international guidelines. All experiments were conducted in accordance with the Canadian Council on Animal Care (CCAC) standards using a protocol approved by Dalhousie University Committee on Laboratory Animals (#17-011).

To assess the efficacy of LINC00284-targeting GapmeRs on tumor growth, we obtained *in vivo*-ready GapmeRs with the same sequence as our *in vitro*-validated GapmeRs from Exiqon (now under Qiagen). To establish tumors, 8-week old NOD/SCID female mice were injected with 1 x 10^6 MDA-MB-468 or SUM149 cells admixed 1:1 with Matrigel-HC (BD BioScience) into the right 4^{th} inguinal mammary fat pad. Once palpable tumors formed (21 days for MDA-MB-468, 7 days for SUM149), the mice were treated subcutaneously (as per Exiqon *in vivo* guidelines, to maximize distribution in mouse tissues) with 15 mg kg^{-1} control GapmeR or GapmeR#4 twice per week. Tumor size was measured weekly (length*width*height)/2, and mice were sacrificed as soon as one mouse reached tumor endpoints.

For PDX 7482 tumors, palpable tumor formation was achieved by surgically engrafting equal 2 mm^3 chunks into the 2^{nd} thoracic mammary fat pad of 8-week old NOD/SCID female mice, and allowing tumors to form for 14 days. At 14 days post-implantation, GapmeR treatments for PDX 7482 tumors were performed as stated above,

and mice sacrificed when at least one mouse reached tumor and humane endpoints. For *ex vivo* mammosphere assays, tumors were excised from the mice, briefly minced mechanically, and digested in 15 mL 225U/mL collagenase III for 45 min at 37°C. The PDX 7482 cells were then filtered through a 70 uM filter, centrifuged at 500 RCF, and washed with red blood cell lysis buffer. The cells were pelleted again and washed with PBS then filtered through a 70 uM filter a second time. Cells were resuspended in a small volume of complete MammoCult media, counted, and then seeded in triplicate in 24-well ultralow adherence plates (Corning) containing complete MammoCult media.

2.16 Aldefluor-sorted tumorigenicity studies

FACS-isolated Aldefluor+ and Aldefluor- cells were quantified and either 5,000 or 50,000 cells were admixed 1:1 with Matrigel-HC and injected into the left and right 4th inguinal mammary fat pads (left: Aldefluor+, right: Aldefluor-) of 8-week old NOD/SCID female mice. Tumor volume was measured weekly, mice were sacrificed when at least one mouse reached tumor and/or humane endpoints, and tumors were then processed as describe above for downstream analyses.

2.17 Real-time quantitative PCR

For all real-time quantitative polymerase chain reactions (RT-QPCR), cells were collected in TRIzol reagent and total RNA purified using a PureLink RNA kit (Invitrogen Thermo Fisher Scientific) following the manufacturer's instructions, with the addition of an on-column DNase I (Invitrogen Thermo Fisher Scientific) treatment step. RNA was reverse transcribed with the iScript cDNA Synthesis Kit (Bio-Rad) as per manufacturer's

instructions. RT-QPCR was performed using SsoAdvanced Universal SYBR Supermix (Bio-Rad) and gene-specific primers (primer sequences are listed in Table 1) as per manufacturer's recommended protocol using a CFX96 Touch RealTime PCR Detection System (Bio-Rad). Standard curves were generated for each primer pair, and primer efficiencies were incorporated into the CFX Manager software (Bio-Rad). Melt curves were analyzed for quality control to ensure single amplicon generation. Gene expression of all samples was calculated relative to references genes (GAPDH, B2M, PUM1, ARF1, or RPL13A) and, in some cases, normalized to respective controls.

2.18 Chromatin isolation by RNA purification

ChIRP-seq (chromatin isolation by RNA purification) experiments were performed as previously described by Chu et al[121]. Briefly, tiling antisense oligo probes spanning the LINC00284 sequence were generated using the Stellaris FISH Probe Designer (https://www.biosearchtech.com/support/tools/design-software/stellaris-probe-designer) using the following parameters: 20-mer probe every 100 bp of RNA, spacing between probes of 80 nt, target GC% = 45. The probes were purchased from BioSynthesis with HPLC purification. MDA-MB-468 cells were collected and 40 million cells split into two tubes (20M cells/tube), for an even probeset and an odd probeset. Cell pellets were cross-linked with 1% glutaraldehyde (Sigma) for 10min and then quenched with 1.25M glycine (Sigma) for 5 mins. The pellets were resuspended in complete lysis buffer (containing protease inhibitor cocktail (PIC) and RNase inhibitor) and then weighed and prepared for sonication (QSonica Q800R2) in polystyrene tubes using the following conditions; temperature = 4°C, amplitude = 70%, ON time = 1 hr 30 mins;

cycles = 15S ON, 45S OFF (to ensure constant 4°C temperature is maintained). Chromatin size (between 100-400 bp) was confirmed using a BioAnalyzer (QIAxcel). RNA and DNA input samples were taken from the sonicated chromatin, and the experimental "odd" or "even" tubes of chromatin in hybridization buffer were incubated with odd or even probesets. Lysis buffer-washed Streptavidin C1 magnetic beads (Invitrogen Thermo Fisher Scientific) were incubated with the biotinylated probes and chromatin overnight at 37°C. Following 5 washes with wash buffer containing fresh PIC and RNase inhibitors, a small sample of RNA-bound chromatin was aliquoted for RNA purification. This RNA sample, along with the RNA input, were de-crosslinked in proteinase K buffer containing proteinase K (ThermoFisher) at 50°C for 45 mins. The samples were then heated at 95°C for 10 min, chilled on ice for 5 min, and resuspended in TRIzol reagent. RNA purification was performed as previously described, and LINC00284 retrieval confirmed via RT-QPCR using LINC00284-specific primers and GAPDH primers as a negative control.

The rest of the RNA-bound chromatin on beads, along with the DNA input, was incubated with 100 mg RNase A (Invitrogen Thermo Fisher Scientific), 100U RNase H (New England Bioscience), and DNA elution buffer for 30 min at 37°C. Supernatants were removed from beads using a magnetic stand. The elution was performed twice and eluates from each sample were pooled. Proteinase K was added to each DNA sample, and left to incubate at 50°C for 45 min. The DNA was purified using a DNA purification kit (Invitrogen Thermo Fisher Scientific). Purified ChIRP'd DNA was then used for high throughput next generation sequencing (Illumina HiSeq 2000) to identify LINC00284-bound chromatin.

2.19 Cloning and generation of pcDNA3-LINC00284

Standard techniques were used to construct the pcDNA3-LINC00284 plasmid. All restriction enzymes were obtained from New England Biolabs. The size and integrity of restriction enzyme-cut products was confirmed using a 1% agarose gel, and where applicable, was purified from the gel using a gel extraction kit (Qiagen). All plasmids were purified from bacteria using a plasmid purification kit, as per manufacturer's instructions (Qiagen). All constructs were characterized by polymerase chain reaction (PCR) and sequences were confirmed using Sanger sequencing (GeneWiz).

First-strand cDNA was synthesized from MDA-MB-468 cell RNA, using SuperScript III, and oligo$DT_{12\text{-}18}$ (Invitrogen Thermo Fisher Scientific), as per manufacturer's instructions. The LINC00284 sequence was retrieved and amplified using LINC00284-specific primers (Table 1) with flanking 5'-*Hind*III (forward) and 3'-*Bam*HI (reverse) sites, in the presence of Platinum SuperFi enzyme (Invitrogen Thermo Fisher Scientific). The amplified LINC00284 cDNA with the primer-introduced restriction sites were digested with *Bam*HI and *Hind*III and ligated into double-digested (*Bam*HI and *Hind*III) pcDNA3 plasmid using T4 ligase (New England Bioscience). Competent TOP10 *E. coli* cells were transformed and incubated overnight at 37°C. The plasmids were purified with a plasmid purification kit, and the resulting plasmid was designated pcDNA3-LINC00284. The clone was verified to have the correct 3208-bp insert by sequencing (GeneWiz).

2.20 RNA-Protein pulldown assays

RNA-protein pulldown assays were performed as previously described[122], with some modifications. Briefly, 1.2 x 10^8 MDA-MB-468 cells were isolated, washed with PBS, and then resuspended in a 1:1:3 ratio of PBS:nuclear isolation buffer:ddH_2O (components in Table 2), to a total of 20 mL. Cells were lysed on ice for 20 min, and nuclei were pelleted at 600RCF. The nuclear pellet was then resuspended in buffer A (components in Table 2) and sonicated to lyse nuclei under the following conditions; temperature = 4°C, amplitude = 70%, ON time = 7 min 30 seconds; cycles = 15S ON, 45S OFF (to ensure constant 4°C temperature was maintained). The sonicated nuclei were collected, centrifuged, and supernatant saved as purified nuclear lysate. Next, pcDNA3-LINC00284 (generation of vector described in section 2.20) was *in vitro* transcribed in the presence of biotin (Biotin Labelling Mix, Roche) using either T7 polymerase (generating a sense LINC00284 transcript) or Sp6 polymerase (generating an antisense LINC00284 transcript). Biotinylated RNA was incubated with DNase I and purified using G50 sephadex columns (Roche). The biotinylated RNA was denatured at 65°C, then slowly cooled to 4°C in a thermocycler, to ensure proper folding. The nuclear lysate (approximately 3 mg) was pre-cleared with buffer A-washed magnetic C1 streptavidin beads and following removal of beads, 25 ug of biotinylated sense or antisense LINC00284 was added and left to incubate rotating overnight at 4°C. Biotinylated RNA bound to proteins were then isolated by incubating with 30 uL buffer A-washed magnetic beads. Beads were washed five times with buffer A supplemented with fresh PIC and RNase inhibitors, and RNA-protein complexes eluted by incubating with 3 mM biotin (Sigma) in PBS at room temperature for 30 min. Eluted proteins were

then TCA-precipitated and subjected to mass spectrometry to identify LINC00284-bound proteins.

2.21 RNA Immunoprecipitation

RNA immunoprecipitation (RIP) was performed as previously described[96], with some adaptations. Briefly, 1 x 10^7 MDA-MB-468 cells were grown to confluency, trypsinized, washed in PBS, resuspended in a 1:1:3 ratio of ice cold PBS, fresh nuclear isolation buffer, and ice cold ddH_2O, and then incubated on ice for 20 min. The nuclei were then pelleted at 2500RCF for 15 min, and resuspended in RIP buffer (RNase inhibitors and PIC added fresh). Nuclei were then sonicated under the following conditions; temperature = 4°C, amplitude = 70%, ON time = 7 min 30 seconds; cycles = 15S ON, 45S OFF. Nuclear membranes and debris were pelleted at 13,000 rpm for 10 min, and supernatant was collected as nuclear lysates. A 5% input sample was collected. The nuclear lysates were precleared by incubating with RIP buffer-washed (containing fresh PIC and RNase inhibitors) magnetic protein A dynabeads (Invitrogen Thermo Fisher Scientific) for 30 min at 4°C. Rabbit polyclonal anti-human SND1 antibody (6 ug, Abcam, cat#)) or rabbit IgG (negative control; 6 ug, company, cat#) antibodies were added to the precleared lysates and incubated overnight, rotating at 4°C. The next day, 30 uL protein A beads (per RIP) were added to the lysates and incubated for 1 hour at 4°C, gently rotating. Beads were then precipitated, and resuspended in RIP buffer (with PIC and RNase inhibitors) five times followed by one wash in PBS – at the last RIP wash, 5% of the bead volume was saved for SDS-PAGE analysis to confirm protein retrieval (anti-SND1 vs anti-IgG). The remaining beads were resuspended in TRIzol reagent and

subjected to RNA extraction, cDNA preparation, and RT-QPCR. SND1-bound LINC00284 was confirmed using LINC00284-specific primers (see Table 1), and compared to IgG-bound LINC00284. The SND1 pulldown was confirmed using western blot; antibody details in Table 1.

2.22 Subcellular fractionation

In order to fractionate cells into nuclear and cytoplasmic fractions, 2 x 10^6 MDA-MB-468 cells were grown to confluency, trypsinized, and resuspended in ice-cold fractionation lysis buffer (100 uL per 10^6 cells; components in Table 2). The tube was inverted gently to mix, then placed on ice to incubate for 5 min. The sample was then centrifuged at 5000 RCF for 5 min at 4°C to pellet nuclei. The supernatant (cytoplasmic fraction) was transferred to a new RNase-free tube, and EDTA was added to a final concentration of 10 mM. An appropriate amount of TRIzol reagent was added to each fraction (1 mL for nuclei, 0.8 mL for cytoplasmic fraction). RNA purification was then performed as previously discussed, and lncRNA expression was assessed in the fractions by RT-QPCR, using gene specific primers for LINC00284, NEAT1 (positive nuclear control), and DANCR (positive cytoplasmic control). See Table 1 for primer sequences.

2.23 Mammosphere assay

Mammosphere assays were performed by seeding optimized numbers of SUM149 (4 x 10^3) or PDX 7482 (5 x 10^3) cells. SUM149 cells were trypsinized, counted, and then seeded in triplicate in 24-well ultralow adherence plates in complete MammoCult media (StemCell Technologies). PDX 7482 cells were isolated 2 weeks' post-implantation from

8-week old NOD/SCID mice, and incubated in 225 U/mL collagenase for 45 min at 37°C. The PDX 7482 cells were then filtered through a 70 uM filter, centrifuged at 500 RCF, and washed with red blood cell lysis buffer. The cells were pelleted again and washed with PBS then filtered through a 70 uM filter a second time. Cells were resuspended in a small volume of complete MammoCult media, counted, and then seeded in triplicate in 24-well ultralow adherence plates. Cells were then treated 4 hours later with control GapmeR, GapmeR#3, or GapmeR#4 to final concentration of 15 nM. Spheres were counted following either 5 days (SUM149) or 5-11 days (PDX 7482) of incubation.

2.24 Bioinformatics analyses

Samples were sequenced on the Illumina platform with read lengths of approximately 130 bp. After quality control, raw reads were uniquely mapped to the reference genome (hg38 assembly) using Bowtie software. Peaks were called using MACS 2.0 and shifted bedgraphs were generated. Reads were normalized by finding concordance between the even and odd lane sequences. To do this, a consensus track was generated by taking the lower value of the two at each coordinate. Per-base coverage was normalized to a total of 150M mappable reads. The consensus track was assumed as the true coverage for every coordinate (i.e. true coverage = min(even coverage, odd coverage)). A SAM file was generated based on the combined lane, and peaks were called using MACS against the corresponding total input with a p-value cutoff of 1e-5. Peaks were then filtered based on: window size of the peaks (around the peak summit),

peak length, fold enrichment vs input of >2.0, average coverage >1.5, and Pearson correlation of >0.30.

All lncRNA survival data was extracted from the KMPlotter breast cancer database[123]. LncRNA expression (RNA-seq) data from the TCGA Cell 2015 database was retrieved using TANRIC online software[124]. RNA-seq expression data of SND1 and ALDH1A3 in the TCGA Cell 2015 database was retrieved using cBioportal online software[125,126]. RNA-seq expression of LINC00284 in the CCLE database was retrieved using the CCLE Broad Institute portal (portals.broadinstitute.org/ccle). GO term analysis was performed using Gene Set Enrichment Analysis software (GSEA)[127,128].

2.25 Microarray analyses

For microarray analyses, MDA-MB-468 cells were treated with either control GapmeR or GapmeR#4 for 48 hours and then collected in TRIzol reagent, and RNA purification was performed. Samples (in triplicate; n=3) were reverse transcribed into cDNA and hybridized to an Affymetrix Human Gene 2.0 ST microarray platform, and gene expression differences quantified. Data was collected in raw CEL format and processed through Transcriptome Analysis Console (Affymetrix) to reveal differential gene expression. In the case of control shRNA vs shALDH1A3 cells, cells were simply collected in TRIzol once confluent and processed as described above, using the same microarray platform.

2.26 Statistical analyses

All statistical analyses were performed in GraphPad Prism 7. Statistical tests varied based on the experimental setup. In all cases where two samples are compared, a student's t test was performed. When three or more samples were compared, a one-way ANOVA with Dunnett's post-test was performed. An alpha of 0.05 was used for all statistical tests and correlations. Significance is listed as follows: * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$, **** = $p<0.0001$).

Table 1. Primer sequences used and antibody details.

Primer sequences	Forward	Reverse
PUM1	GGCGTTAGCATGGTGGAGTA	CATCCCTTGGGCCAAATCCT
ARF1	GTGTTCGCCAACAAGCAGG	CAGTTCCTGTGGCGTAGTGA
GAPDH	GGAGTCAACGGATTTGGTCGTA	TTCTCCATGGTGGTGAAGAC
B2M	AGGCTATCCAGCGTACTCCA	CGGATGGATGAAACCCAGACA
LINC00284	CAGATAAGCCTCTCGGCACC	TTTGCCCTTCCCTGTTACCC
ALDH1A1	TGTTAGCTGATGCCGACTTG	TTCTTAGCCCGCTCAACACT
ALDH1A3	TCTCGACAAAGCCCTGAAGT	TATTCGGCCAAAGCGTATTC
ZFAS1	CAACTACTAGAGCGCCTCGG	CCAAAGATGGCTTTCGCACC
SNHG6	CTGTCTTCCGATGTCGCTCT	CGGCATGACTAACGGCTCTT
LRCC75A-AS1	TTCCCGTTGTTATGGAGGGC	TCCAGTTCTCTCGGGTTTGC
AC004542.2	CAGAAAGGCCAGCCATACCA	ACCCCTGAGTACCCAGAGTAA
AC0093001	TCGCCAAGCAATTACCTACCA	CGGCTGCGGGTATATTCCAA
UCA1	CCGAGAGCCGATCAGACAAA	GGGATGGCCATTTGGAAGGA
TUG1	AGCGTGGGTGTACGTAAAGG	CCAAGGATTGGGGAACTGCT
NORAD	CTTAAGGGGCCTGGAAGGTG	AGAATGAAGACCAACCGCCC
MALAT1	GCAAAACGTGTGGCTGTCTT	GTGGCAAAATGGCGGACTTT
LINC-ROR	GAATCAGAGTGCTGGGCAGT	TCAGCAGCTCATGCCCTAAC
HULC	ACTCTGAAGTAAAGGCCGGAA	TGCCAGGAAACTTCTTGCTTG
HOTAIR	GGGACAGAAGGAAAGCCCTC	GAGTCAGAGTTCCCCACTGC
H19	GAGAGCTTGTGGGAGCCAAG	CCTACTCCACACTCCTCACTG
CYTOR	TTCCAACCTCCGTCTGCATC	GGGGGCTGAGTCGTGATTTT
TUNAR	ACCGGCAGCGTTATTGTTTC	CTAATCCCGAGCTTTCCCCC
GAS1RR	AAGGGGGCAACTGTATACGC	GGAGGTGCTTGATCACTGGT
NEAT1	CCTCCCTTTAACTTATCCATTCAC	TCTCTTCCTCCACCATTACCA
DANCR	AGGAGTTCGTCTCTTACGTCT	TGAAATACCAGCAACAGGACA
PART1	CAGGGTACGCCAACTATAGGAC	TTCAGCTTTCCAGAGCCAGT
LINC00511	TTCCCACAGGAAACCCACAC	CATCACCTGTCTCCTTGCCA
LINC01198	TGCCGAATAGCTCTGACCTG	GTTGAAAAAGGTGCAGTGGCA
LINC01354	ATGCACACATTCGAGGGGAA	CTGTGGAGGACGCTTGAGAG
LINC02159	TCCCATCGGCTTTTGGCTTT	GACACCTGTCCTGCCTCTTAG
AC022509.2	AGTTGGAACTTGTGACCAGCA	GCTTTCCAGCCTCACTTTGG
LINC00518	ACCTAACCTGCGAATGCTGT	GCCTAAACATTTGCTGCCCC
LINC00162	ATTGCTTAGGTGGGGAGCAC	TCATGACTGAGCTTTCCGTCG

RP11-120J1.1	TTCCCAAACCCCATGACTCTG	AGCCCATCACAGTGTTCCTT
LINC01315	CCAATTCCCCAGCGTTTTCC	GGCATCCACTTCATCGCTCA
AP000851.1	TGCTTGGCACATAGCATCCA	GGGTCCATGTTTTCCAAGACG
LNC-DPMI-1:1-2	CCCTGCATGTGCTAAGTGCT	GTGGAAAAATCCTGACATGGTGA
LINC01833	CTTTGTTGGGTTTGGGAGGC	AGGTGAGCTGGCGAATACTG
LINC00839	GGCCCAGATTGTTCCAGGAT	TGGTGAAAAGGCAGATCCCA
LINC00880	CGGGAAAGGTGTACCTCGTG	CCAGGGGCTTTGATCAACCT
LINP1	GACCAGGGCACTCTGTAAGG	GCAGTGGAGTCTGAAGTCCC
AP001626.1	CCCCAGTGGAGGAACCTTCT	GTATGGCAGCCAGGCGATT
AC016995.3	ACTCTGTCCTCTGTCACTGC	CTACTCACGTCCCCTTGGTC
LINC01956	AGTGTTGACTTGGGGACTGC	CAAAAGTCGCGCACTACCTG
LINC00707	CCCAGACATGACCCGATGAC	ATTTTGGTTTGCTGGCCCTG
LNC-NKX-1-2-1:1	GACCCTTCTGGTTTCCACAGA	GGATTTGGTTGGAAAGCCACT
LINC02188	TTGAAGGACCCCTGAATGGC	TCCACGGCTTTGTCTGTCTG
AC091053.1	GAAGCCAGCAAACATCTGAGAC	GAGAGGCCTTCCGCAAATCT
FOXP4-AS1	CTCTGTTTCCGTGGCAACCT	GACCTGGAGCTGTCATCGAG
DGCR5	CCATGGTTCGACGCCATTC	CCAGGGGGCCTTCTTTTTCT
SOX9-AS1	CAACATCTGCATTGGCGGAG	TGCTGAGGCTTCACTCATGG
LINC01819	CTAGGCACACTTGCCCTACC	GTGGGTCTACCCTGTATGCC
LINC00092	CCCATTCTCTTAGGCCCGTC	AGAAACATGCTTTGCGCTGG
LNC-DSC2-1:1	AGCTGCTCCTTATTTCTCCTGTAG	TTGCATGGAGAATGCGATGC
AC027031.2	GTGCAGGATCCGAAACAGGA	AACATGGCAAGCTGGATGGA
VIM-AS1	GCCCAGGCATTGAGTACCAT	CGACGTGTTGTCCTGATGGA
LINC02487	GTCTGTGTCCCTCAGAAGGC	AGAACCAGCGTTTCGGATGT
AC025154.2	AACCACAGGTCGCCACATAG	CAGCTGCCCTGTGTGAGTAA
AC009041.2	GTCCAACCGCGGGTCC	GGTTGAAAGGATCCGAGCCA
AC015712.4	AAGAACCAAGGTGCAACAGGA	GCACTCTTGAGGTTCCCCTG
CYP4Z1	GAATCCTGGGTTGGTCGAGG	AGGTTTCACAATCTGGCGGT
CTNNAL1	CCATGATGGCTCTCTTAGTCCA	ACCCATCCGTTATTTTCCATCTGA
EIF5A2	AGAACGGCTTCGTGGTACTG	CGTGCTTTCCCGTCTTGGA
FOSL1	CTGGTGCCAAGCATCAACAC	ACTGAGGGTAGGTCAGAGGC
FSCN1	GCAAGAATGCCAGCTGCTAC	ACAAACTTGCCATTGGACGC
IL7R	TTCTCTGTCGCTCTGTTGGTC	ACTGGGCCATACGATAGGCT
SERPINE2	ATTGAACTGCCCTACCACGG	GTGTGGGATGATGGCAGACA

SLITRK6	TGCTGCAGGGATAGTGGTTC	TGCACAGGACTGTTGTCTCTC
Cloning primers	**Forward primer**	**Reverse primer**
LINC00284-pcDNA3	ACGTAAGCTTTACAAAGCAAACCAGGAGAA	ACGTGGATCCATTTCTGAAATTTTTATGTTTCATTT
Western blot antibodies	**Details, catalogue number**	**Concentrations**
ALDH1A3	Mouse monoclonal antibody, clone OTI4E8. Origene catalogue number: TA502841	WB: 1/1000
SND1	Rabbit polyclonal antibody, Abcam catalogue number ab65078	WB: 1/1000. RIP: 6 ug
RIP IgG	Rabbit polyclonal antibody, Diagenode catalogue number C15410206	WB: 1/1000. RIP: 6 ug

Table 2. Composition of buffers for various protocols.

Protocol	Buffer	Ingredients
Aldefluor assay/PDX 7482 cell collection	Red blood cell lysis buffer	157 mM NH_4Cl, 2.5 mM $KHCO_3$, ddH_2O; pH adjusted to 7.30.
ChIRP	Lysis buffer	50 mM Tris-HCl pH 7.0, 10 mM EDTA, 1% SDS, ddH_2O. **Add fresh:** 1 mM AEBSF, 100x PIC, and 100x RNase inhibitor fresh each use
	Hybridization buffer	750 mM NaCl, 50 mM Tris-HCl pH 7.0, 1 mM EDTA, 1% SDS, 15% formamide, ddH_2O. **Add fresh:** 1 mM AEBSF, 100x PIC, and 100x RNase inhibitor fresh each use.
	Wash buffer	2X SSC, 0.5% SDS, ddH_2O. **Add fresh:** 1 mM AEBSF, 100x PIC, and 100x RNase inhibitor fresh each use.
	Proteinase K buffer	100 mM NaCl, 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 0.5% SDS, ddH_2O. **Add fresh:** 1 mM AEBSF, 100x PIC, and 100x RNase inhibitor fresh each use.
	DNA elution buffer	50 mM $NaHCO_3$, 1% SDS, ddH_2O.
RNA pulldown	Nuclear isolation buffer	1.28 M sucrose, 40 mM Tris-HCl pH 7.5, 20 mM $MgCl_2$, 4% Triton X-100, ddH_2O.
	Buffer A	150 mM KCl, 25 mM Tris-HCl pH 7.4, 5 mM EDTA, 0.5% IGEPAL, ddH_2O. **Add fresh:** 0.5 mM DTT, 1 mM AEBSF, 100x PIC, 100x RNase inhibitor fresh each use.
RIP	Nuclear isolation buffer	1.28 M sucrose, 40 mM Tris-HCl pH 7.5, 20 mM $MgCl_2$, 4% Triton X-100.
	RIP buffer	150 mM KCl, 25 mM Tris-HCl pH 7.4, 5 mM EDTA, 0.5% IGEPAL. **Add fresh:** 0.5 mM DTT, 1 mM AEBSF, 100x PIC, 100x RNase inhibitor fresh each use.
Subcellular fractionation	Fractionation lysis buffer	140 mM NaCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl pH 8.5, 0.5% IGEPAL, ddH_2O.

CHAPTER 3 RESULTS

3.1 A clinically-guided RNA-sequencing screen identifies lncRNAs enriched in TNBC CSCs

Our aim was to generate a list of potential lncRNAs for further study that could be important in TNBC growth and CSC maintenance. Thus, we employed the following prioritization strategy: first, enriched in TNBC/basal-like patient tumors; second, enriched in TNBC CSC populations; and third, demonstrably predicting patient survival, and thus, associated with worse outcomes. Utilizing The Cancer Genome Atlas (TCGA) RNA-seq data (Breast Cancer Cell 2015), Zhang et al.,[103] recently published lists of the most highly enriched lncRNAs in TNBC and basal-like breast cancer. With the knowledge that, in general, TNBC/basal-like tumors harbor higher numbers of CSCs[54,56–62,129], we wondered if at least some of the top 50 TNBC/basal-like enriched lncRNAs identified by Zhang et al. would also be enriched in the CSCs of TNBC/basal-like tumors. Thus, we employed the Aldefluor assay to isolate populations of CSCs in two TNBC models: SUM149 cells and PDX 7482 cells. Aldefluor^{high} (ALDE+) and Aldefluor^{low} (ALDE-) in the two TNBC models were identified (Fig. 1A,B top), and confirmed through inclusion of a sample treated with the ALDH inhibitor, diethylaminobenzaldehyde (DEAB) (Fig. 1A,B bottom). Subsequent implantation of equal numbers of sorted cells in the mammary fat pads of NOD/SCID female mice confirmed that the isolated Aldefluor^{high} cells were significantly more tumorigenic than the Aldefluor^{low} cells (Fig. 1C). RNA was isolated from the sorted cell populations and using QPCR, we assessed expression of ALDH1A1, ALDH1A3, and the 50 most enriched lncRNAs in TNBC/basal-like patient tumors. Consistent with previous reports[46], the Aldefluor^{high} cells had increased ALDH1A3 levels (and low/undetectable ALDH1A1

levels) compared to Aldefluor^{low} cells (Fig. 2A, 2B, first bar). Of the 50 lncRNA, we were unable to detect expression of 13 lncRNAs in either TNBC model by QPCR. Of the remaining 37 QPCR-detectable lncRNAs, 26 and 12 lncRNAs were enriched >2-fold (i.e. $\log_2$>1) in the Aldefluor^{high} populations of SUM149 (Fig. 2A) and PDX 7482 cells (Fig. 2B), respectively. Across both TNBC CSC models, 10 lncRNAs were commonly enriched >2-fold (PART1, LINC00511, LINC001198, LINC01354, LINC00284, LINC02159, AC022509.2, LINC00518, LINC00162, and RP11-120J1.1; Fig. 2A, 2B, green bars) and were thus prioritized for further analyses.

We also assessed whether previously studied CSC-enriched lncRNAs were highly expressed in our ALDE+ populations. Importantly, breast CSC-associated lncRNAs (LINC-ROR, HOTAIR, H19, TUNAR, and GAS1RR), but not the lncRNAs enriched in CSCs from other cancers, were enriched in at least one of our CSC models (Fig 3A). This is in contrast to the five most expressed lncRNAs in breast cancer in general (identified using online database TANRIC), which were not enriched in the Aldefluor^{high} populations of the two models (Fig. 3B). Thus, our model and approach seems to be reliable in identifying novel lncRNAs enriched in CSCs.

Consistent with the analysis of Zhang et al[103]., the 10 Aldefuor^{high}-enriched lncRNAs are significantly enriched in both basal-like (Fig. 4A) and TNBC (Fig. 4B) patient tumors (RNA-seq data on lncRNA expression retrieved from online database TANRIC). Next, we assessed if any of the 10 lncRNAs are associated with worse outcomes among basal-like patients, since this may suggest possible oncogenic function. Using both KMPlotter[130] and analysis of TCGA breast cancer Cell 2015 dataset (data extracted from cBioportal), high expression of LINC00284 was most associated with

decreased survival compared to the other lncRNAs (Fig. 4C,D). Therefore, LINC00284 best fulfills the three aforementioned prioritization criteria; highly expressed in TNBC/basal-like tumors, enriched in TNBC/basal-like CSC populations, and demonstrably associated with worse patient outcomes. For these reasons, it was on the top of our list for functional studies of its role in TNBC and CSCs.

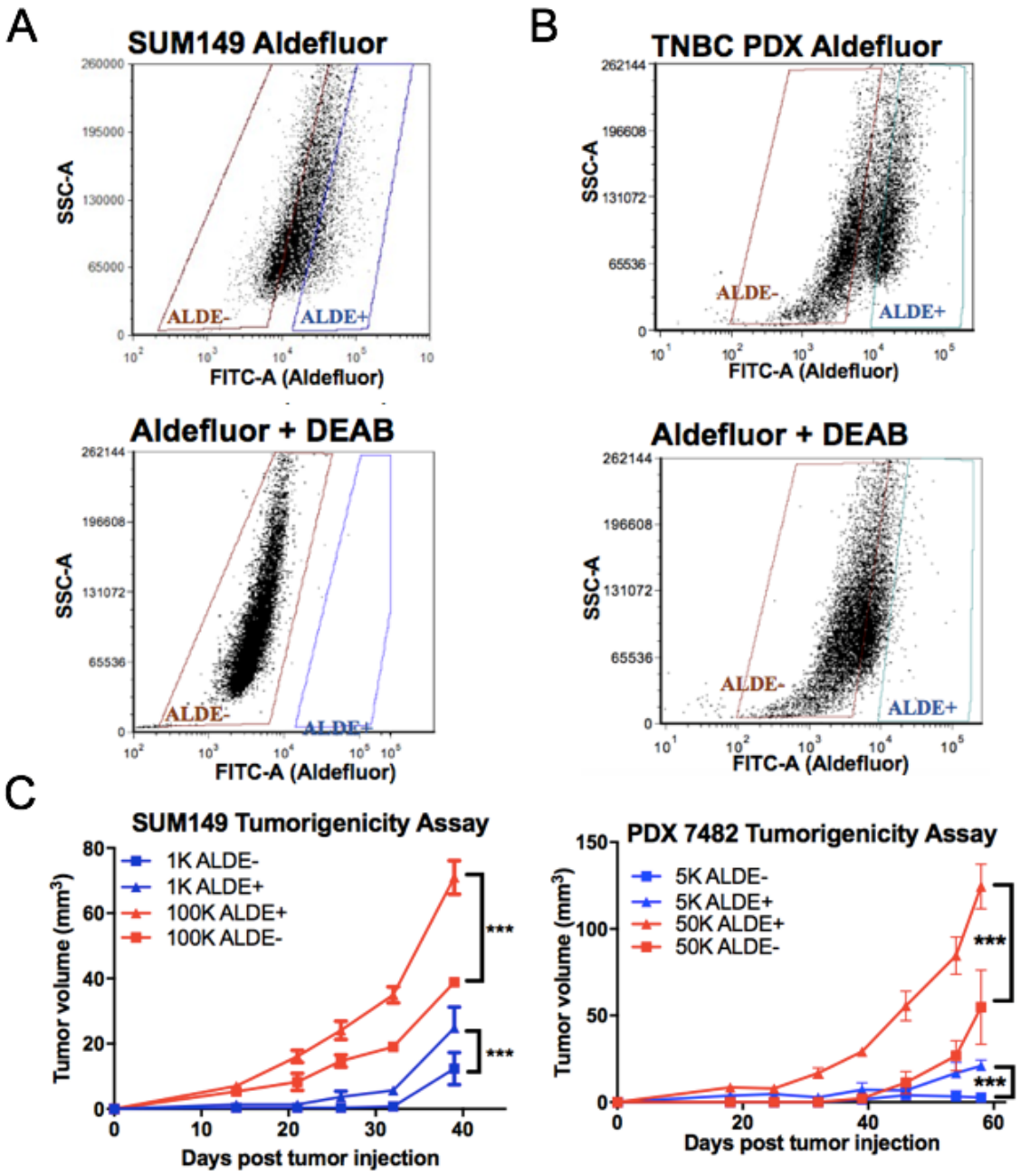

Figure 1. Isolation of Aldefluor^{high} CSCs that are highly tumorigenic. (A,B) Representative FACS plots of Aldefluor^{high} (ALDE+) cell populations that are isolated from SUM149 **(A)** and PDX 7482 **(B)** cells. The inclusion of a DEAB-treated control allows for proper gate setting during FACS. **(C)** Aldefluor^{high} or Aldefluor^{low} sorted populations of TNBC SUM149 (left) or PDX7482 (right) are assessed for tumorigenicity in mice. Significance was modelled using exponential regression (n=6 per group for SUM149, n=3 per group for PDX 7482). Error bars represent standard error.

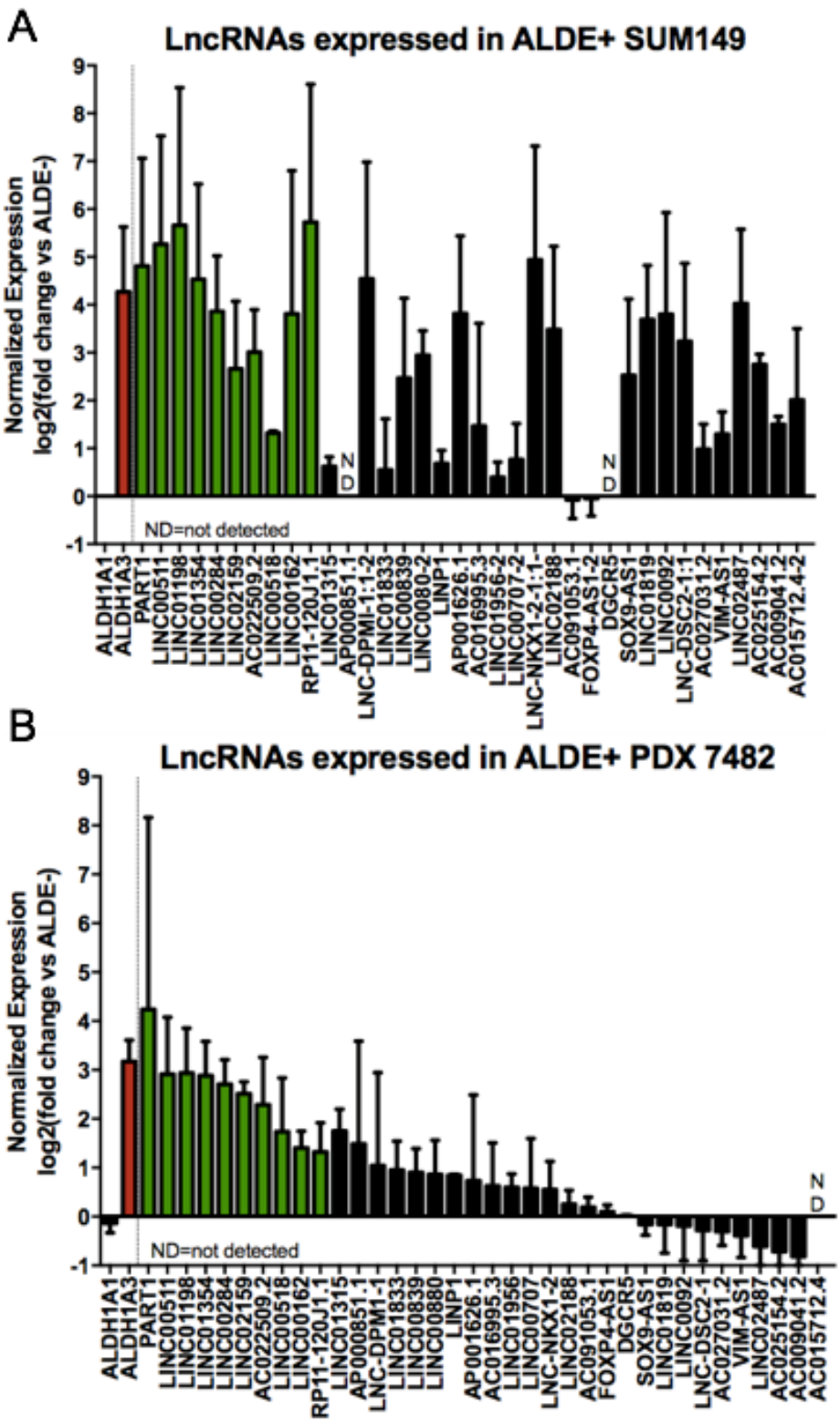

Figure 2. LncRNAs enriched in basal-like/TNBC patient tumors are enriched in TNBC CSCs.

Figure 2. LncRNAs enriched in basal-like/TNBC patient tumors are enriched in TNBC CSCs. (A,B) QPCR analysis of the Aldefluor^{high} cells vs the Aldefluor^{low} cells reveals that 10 lncRNA (green) are enriched in the Aldefluor^{high} cells of both SUM149 (n=3) **(A)** and PDX 7482 (n=2) **(B)**. Expression in both experiments is shown as normalized expression to reference genes PUM1 and ARF1, and then fold-change over the lncRNA levels in the Aldefluor^{low} cells. Error bars represent SD.

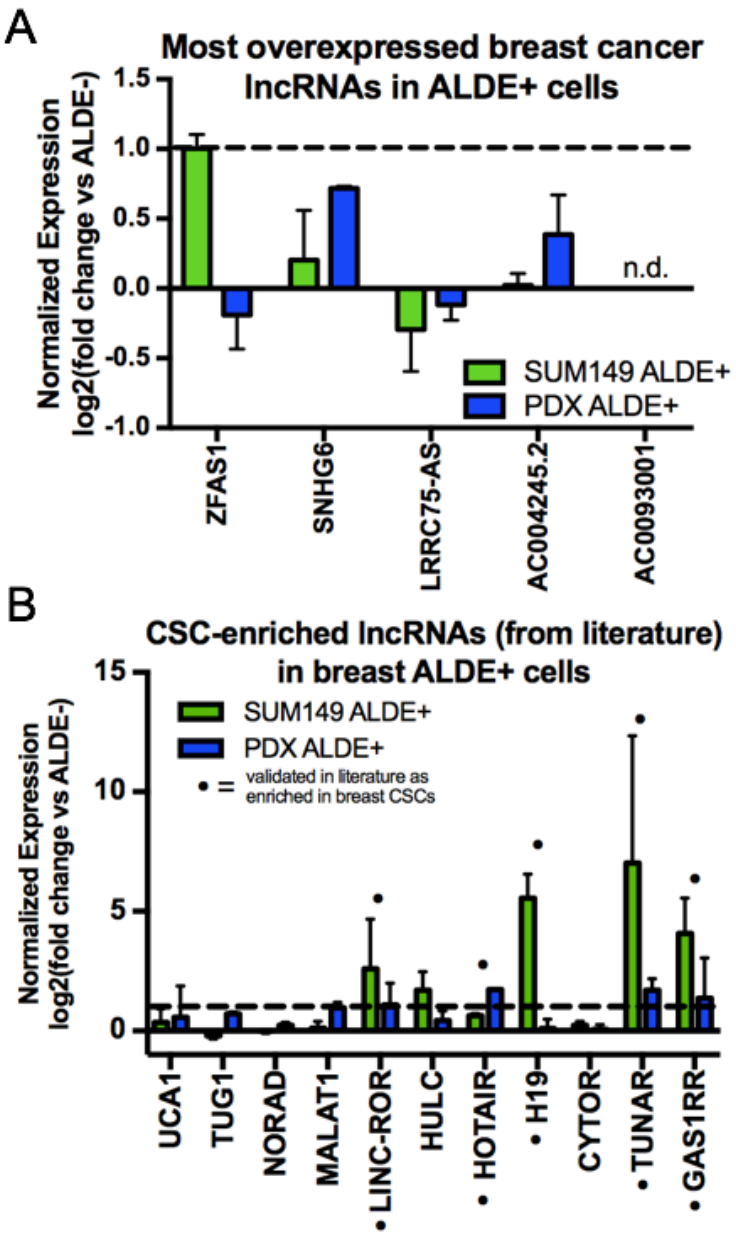

Figure 3. Highly expressed lncRNA in breast cancer are not enriched in TNBC models of Aldefluor^{high} CSCs, unlike CSC-enriched lncRNAs. (A) RNA-seq data retrieved from TCGA Cell 2015 dataset identified the most expressed lncRNA in breast cancer samples. QPCR analysis was performed to assess expression of the lncRNA in SUM149 (green) and PDX 7482 (blue) CSCs **(B)** Literature searches revealed numerous previously reported CSC-enriched lncRNAs. QPCR analysis was performed to assess expression of the lncRNAs in SUM149 and PDX 7482 CSCs. Expression in both experiments is shown as normalized expression to reference genes PUM1 and ARF1, and then fold-change over the lncRNA levels in the Aldefluor^{low} cells. Error bars represent SD (n=3 for SUM149, n=2 for PDX 7482).

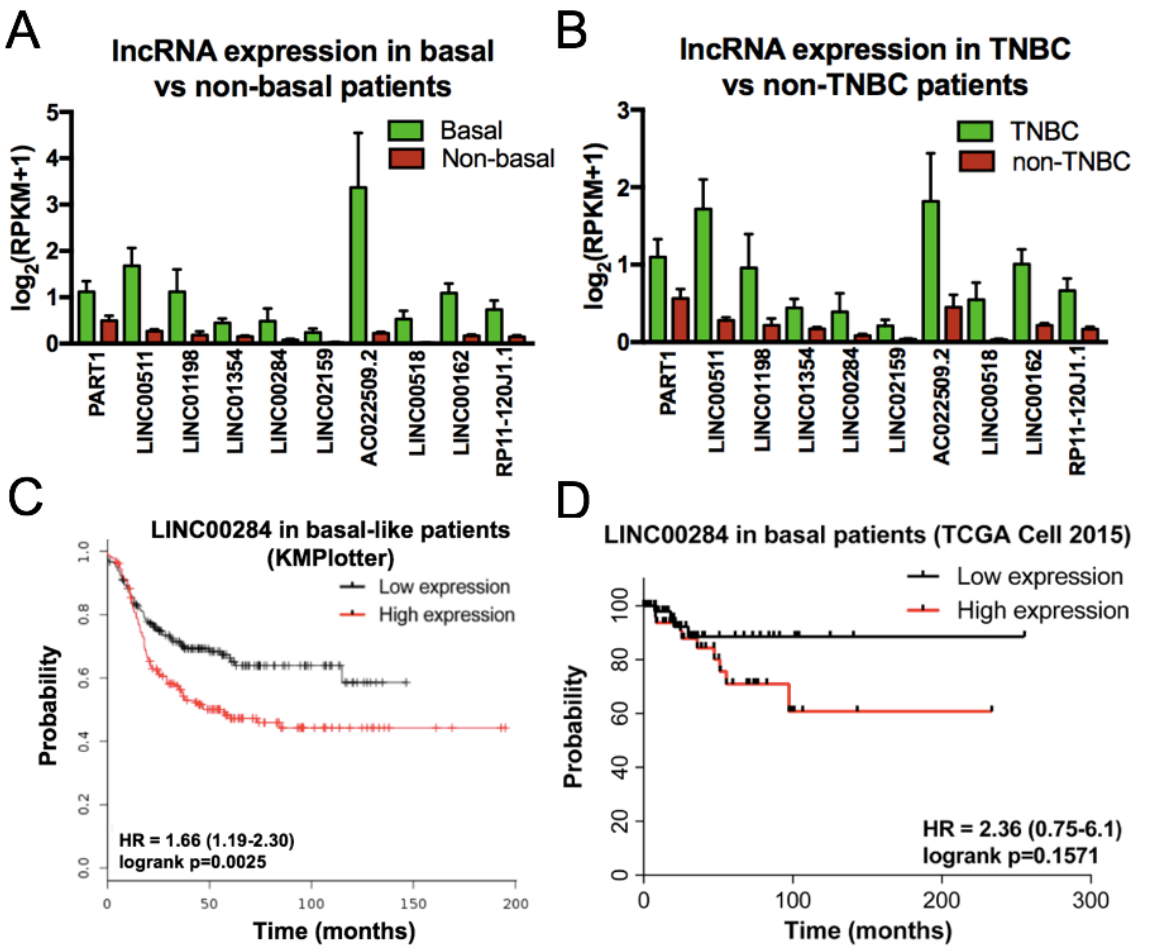

Figure 4. Amongst the 10 basal-like/TNBC-enriched screen-identified lncRNA expressed highly in CSCs, LINC00284 predicts patient outcomes. RNA-seq expression of the 10 lncRNAs enriched >2-fold in SUM149 and PDX 7482 CSCs in basal-like vs non-basal breast cancer **(A)** and TNBC vs non-TNBC **(B)** in the TCGA Cell 2015 dataset. **(C)** Regression-free survival in 360 basal-like breast cancer patients based on median expression of LINC00284 was analyzed using KMPlotter. **(D)** Regression-free survival in 106 basal-like breast cancer patients based on median expression of LINC00284 was analyzed using by extracting survival data from TCGA Cell 2015 dataset. HR = hazard ratio. Error bars in (A) and (B) represent 95% CI.

3.2 LINC00284 is a novel target of CSC marker ALDH1A3 and ATRA

Given the co-expression of CSC marker ALDH1A3 with the 10 lncRNAs in the Aldefluor^{high} cells (Fig. 2A,2B) and its role in gene expression regulation and tumor progression, we wondered if any of the 10 of Aldefuor^{high}-enriched lncRNA were regulated by ALDH1A3. We assessed expression of the lncRNAs in SUM149 cells with ALDH1A3 knocked down by short hairpin RNA (shRNA; Fig. 5A, knockdown confirmed via western blot). QPCR revealed that ALDH1A3 knockdown reduced LINC00284 levels, and LINC00162 to a lesser degree, in SUM149 cells. This suggests that high levels of LINC00284 (and possibly LINC00162) in Aldefluor^{high} cells may be ALDH1A3-dependent. To further investigate the link between LINC00284 and ALDH1A3, we examined the consequence of changing ALDH1A3 levels in two different breast cancer models. Knockdown of ALDH1A3 in basal-like TNBC MDA-MB-468 cells, which have relatively high levels of ALDH1A3[46], also reduced levels of LINC00284 (Fig. 5B, left). Inversely, overexpression of ALDH1A3 in MCF7 ER+ breast cancer cells, which harbor low basal levels of ALDH1A3[46], resulted in a corresponding increase in LINC00284 levels (Fig 5B, right). A correlation between LINC00284 and ALDH1A3 mRNA levels was also observed in breast cancer patient tumors (Fig. 5C, TCGA Cell 2015 dataset) and in cell line RNA-seq data obtained from the Cancer Cell Line Encyclopedia (CCLE, Fig. 5D). Together, these data suggest that LINC00284 levels are dependent upon ALDH1A3.

As an enzyme, ALDH1A3 has a role in gene expression regulation primarily through its production of all-trans retinoic acid (ATRA) from retinal, which binds nuclear hormone receptors leading to widespread gene expression changes[76,80]. Thus, we

wondered if ATRA treatment alone would be sufficient to induce LINC00284 expression. Treatment of a panel of breast cancer cell lines with 100nM ATRA consistently increased LINC00284 levels (Fig. 5E). Together, these results strongly suggest that ALDH1A3/ATRA regulate LINC00284 and that the lncRNA is novel downstream target of ALDH1A3, solidifying the importance of determining the function of LINC00284 in the context of breast cancer progression.

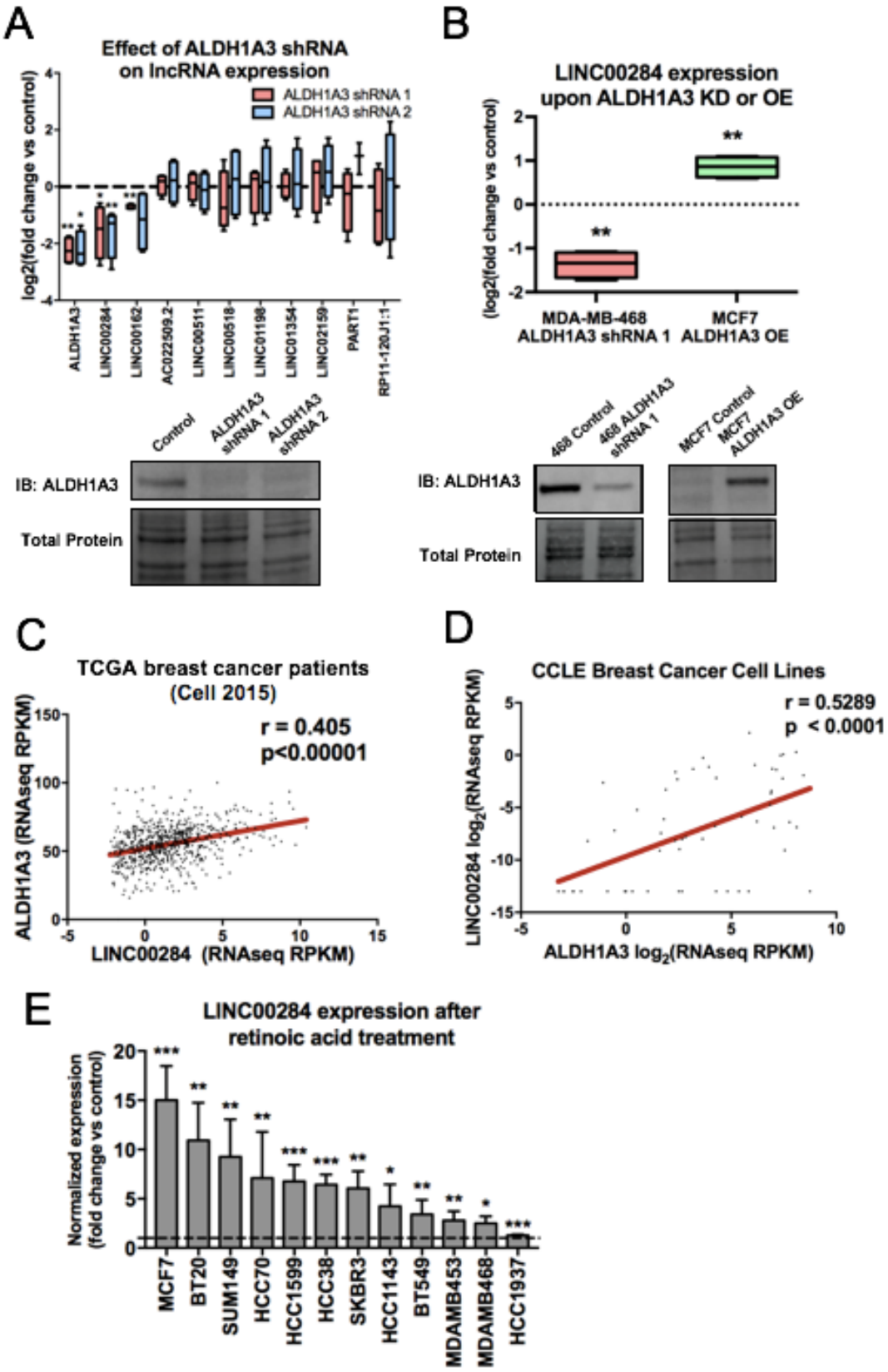

Figure 5. LINC00284 is regulated by CSC marker ALDH1A3.

Figure 5. LINC00284 is regulated by CSC marker ALDH1A3. (A) QPCR analysis of expression of the 10 CSC-enriched lncRNA from Fig. 2 in SUM149 cells following knockdown of ALDH1A3 with two separate shRNAs. Expression is shown normalized to reference genes PUM1 and ARF1 and then fold-change over the control cells (n=4). Western blot confirming a decrease in ALDH1A3 levels is below, using total protein as a loading control. **(B)** QPCR analysis of LINC00284 expression following knockdown of ALDH1A3 in MDA-MB-468 cells or overexpression (OE) of ALDH1A3 in MCF7 cells. Expression is shown normalized to reference genes B2M and GAPDH and then fold-change over the control cells (n=4). Western blots confirming changes in ALDH1A3 levels are below, using total protein as a loading control. **(C)** RNA-seq co-expression of LINC00284 and ALDH1A3 in the TCGA Cell 2015 dataset (all breast cancer patients) was retrieved using TANRIC. **(D)** RNA-seq co-expression of LINC00284 and ALDH1A3 in the Cancer Cell Line Encyclopedia (only breast cancer cell lines) was retrieved using the CCLE portal. **(E)** QPCR analysis of LINC00284 expression in a panel of cell lines treated with 100 nM retinoic acid for 24 hours versus no treatment (n=4). For (A), significance was determined using one-way ANOVA with Dunnett's post-test, and for (B) and (E), significance was determined using a student's t-test. For (C) and (D), Spearman correlation is shown, and significance was determined using linear regression. All error bars represent SD.

3.3 LINC00284 confers a survival advantage to breast cancer cells

Since our previous experiments suggested that LINC00284 played a functional role within cells, we wondered if inhibition of LINC00284 would cause a disadvantage to breast cancer cells. First, we assessed the levels of LINC00284 across different breast cancer cell lines. Consistent with LINC00284 enrichment in the basal-like subtype (Fig 4A), LINC00284 was enriched in basal-like cell lines in comparison to the non-basal like cell lines (Fig. 6). Using the same cell line models as in our previous experiments (MDA-MB-468, SUM149, and MCF7), we knocked down expression of LINC00284 using locked nucleic acid (LNA) GapmeRs, antisense oligonucleotides that achieve high knockdown efficiency of transcripts in both the nucleus and cytoplasm. We ensured that at least two GapmeRs were used to control for off-target effects. Compared to the scramble control GapmeR, LINC00284-targeting GapmeR#3 and GapmeR#4 both caused a significant decrease in LINC00284 expression in all three cell lines (Fig. 7A). This decrease in LINC00284 expression caused a corresponding significant decrease in cell proliferation in all three cell lines (Fig. 7B). We then used flow cytometry, staining cells with 7AAD (a cell death marker) and annexin V (an apoptotic marker) and found that there were significant increases in cellular apoptosis following LINC00284 knockdown (Fig 7C,D). We also wondered if LINC00284 was expressed outside of cancer cells, so we assessed LINC00284 levels using Genotype-Tissue Expression (GTEx) online software[131] and found that LINC00284 is poorly expressed in most normal body tissues (Fig 8).

Important for the proposed role of LINC00284 in CSCs, we assessed whether this decrease in cell survival following LINC00284 knockdown had any effect on tumor-

sphere forming cells (i.e. an *in vitro* readout of stemness). We seeded SUM149 or PDX 7482 cells in mammosphere assays and treated them with a single GapmeR treatment. We noted significant decreases in the tumor sphere forming capabilities of both SUM149 (Fig. 9A) and PDX 7482 (Fig. 9B) cells treated with GapmeR#3 and GapmeR#4. Together with our previous results, this suggests that LINC00284 inhibition is detrimental not just to non-stem-like cells, but also to the rarer stem-like cells which are capable of forming tumor spheres.

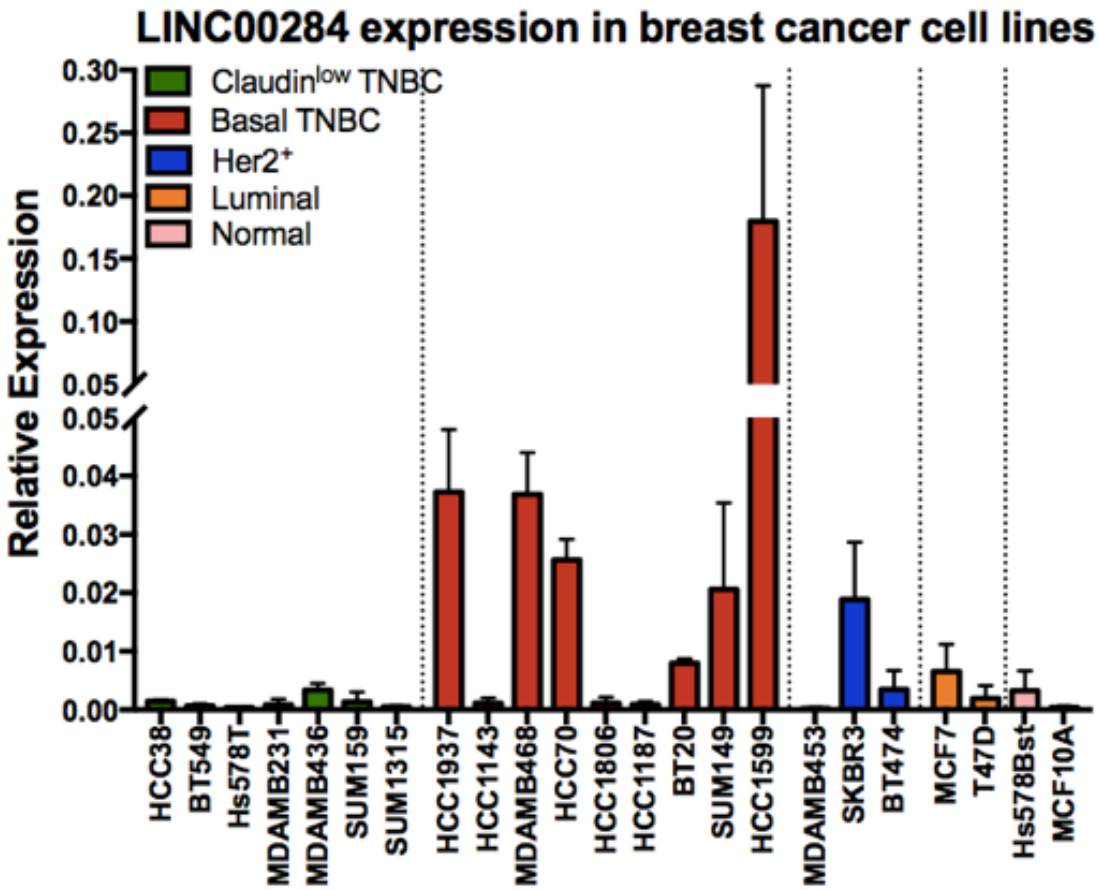

Figure 6. LINC00284 is predominantly overexpressed in basal-like breast cancer cell lines. QPCR was used to determine expression of LINC00284 in 21 different breast cancer cell lines and two normal immortalized breast cell lines. Expression relative to PUM1 and ARF1, which were used as reference genes in the panel due to high target stability values across all cell lines tested, is shown (n=4). Error bars represent SD.

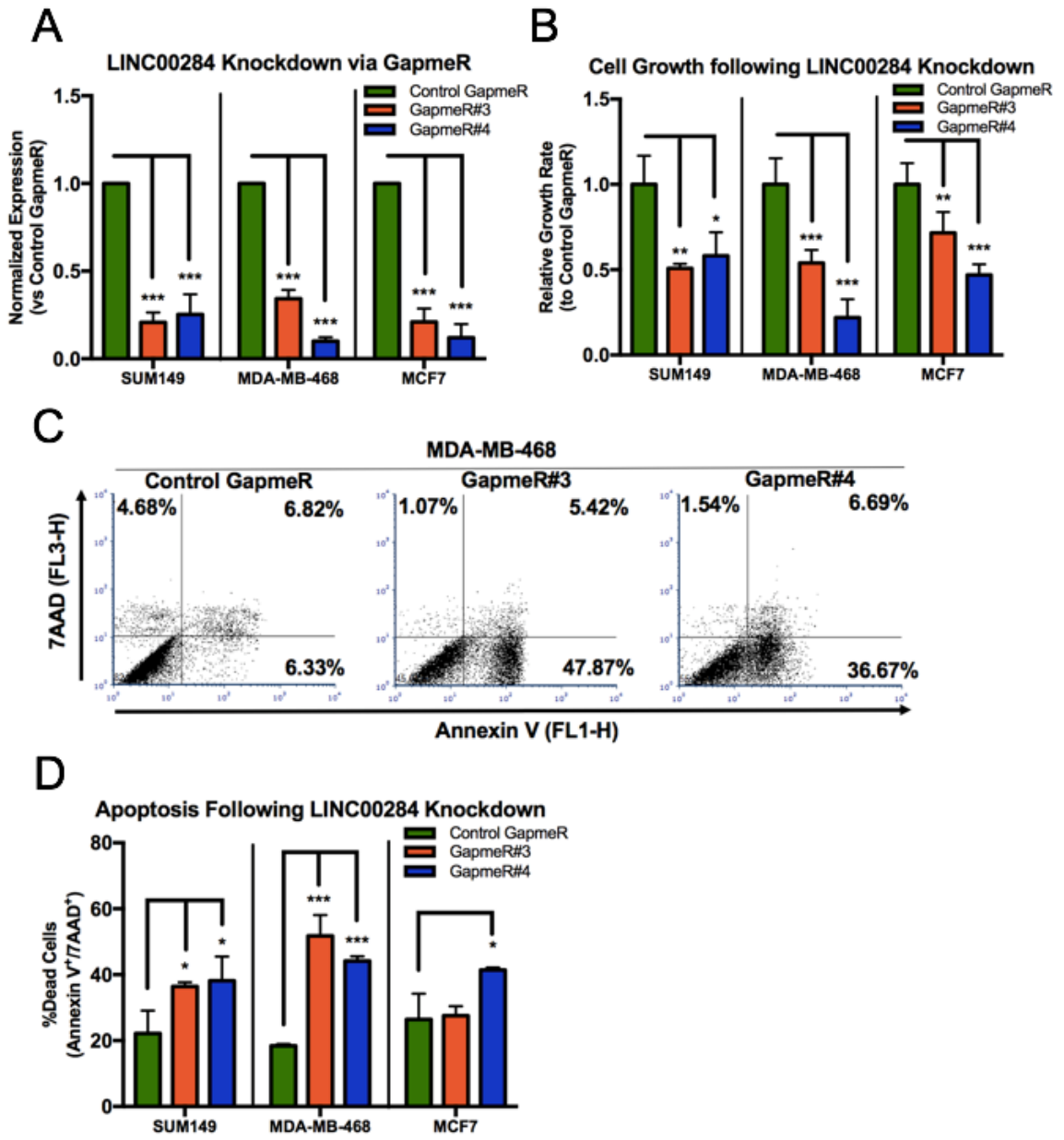

Figure 7. LINC00284 contributes to breast cancer cell viability. (A) QPCR analysis of LINC00284 expression following knockdown in SUM149, MDA-MB-468, and MCF7 cells treated with LINC00284-specific GapmeRs or control GapmeR (n=4). Expression is shown normalized to reference genes B2M and GAPDH. **(B)** The effect of LINC00284 inhibition via two GapmeRs versus the control GapmeR was quantified by counting the relative number of viable cells 3 days after treatment, using a trypan blue exclusion assay (n=4). **(C)** Representative flow cytometry of MDA-MB-468 cells treated with either control GapmeR or LINC00284-specific GapmeRs, quantified by staining with cell death marker 7AAD and apoptosis marker annexin V-488 (10,000 events shown). **(D)** Flow cytometry was used to quantify levels of cell death and apoptosis across SUM149, MDA-MB-468, and MCF7 cells (n=4). Significance was determined using one-way ANOVA with Dunnett's post-test. Error bars represent SD.

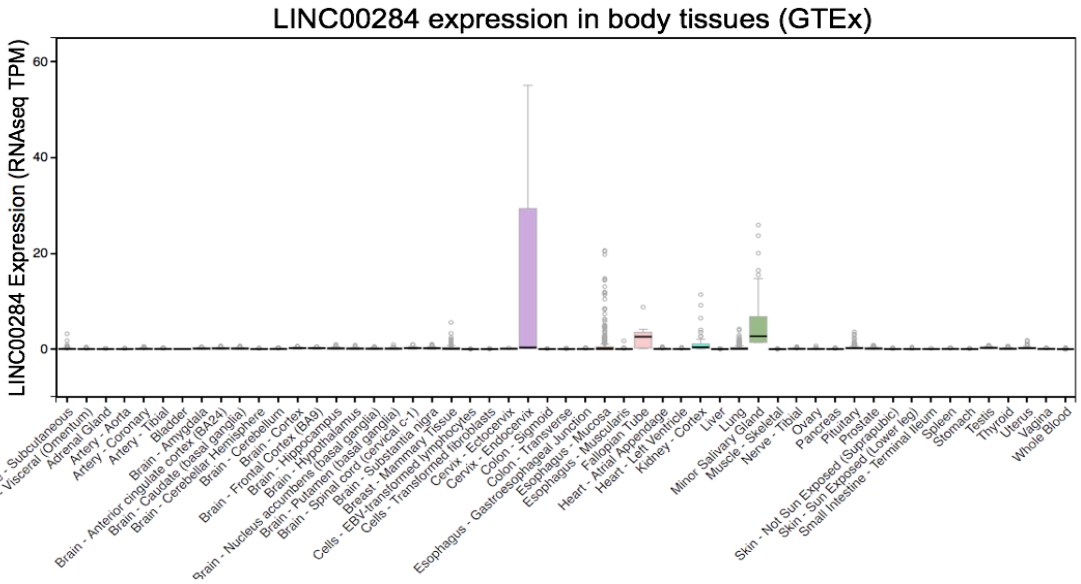

Figure 8. LINC00284 is poorly expressed across most human bodily tissues. RNA-seq data (transcripts per million, TPM) for LINC00284 was extracted from GTEx to assess LINC00284 levels across normal tissues.

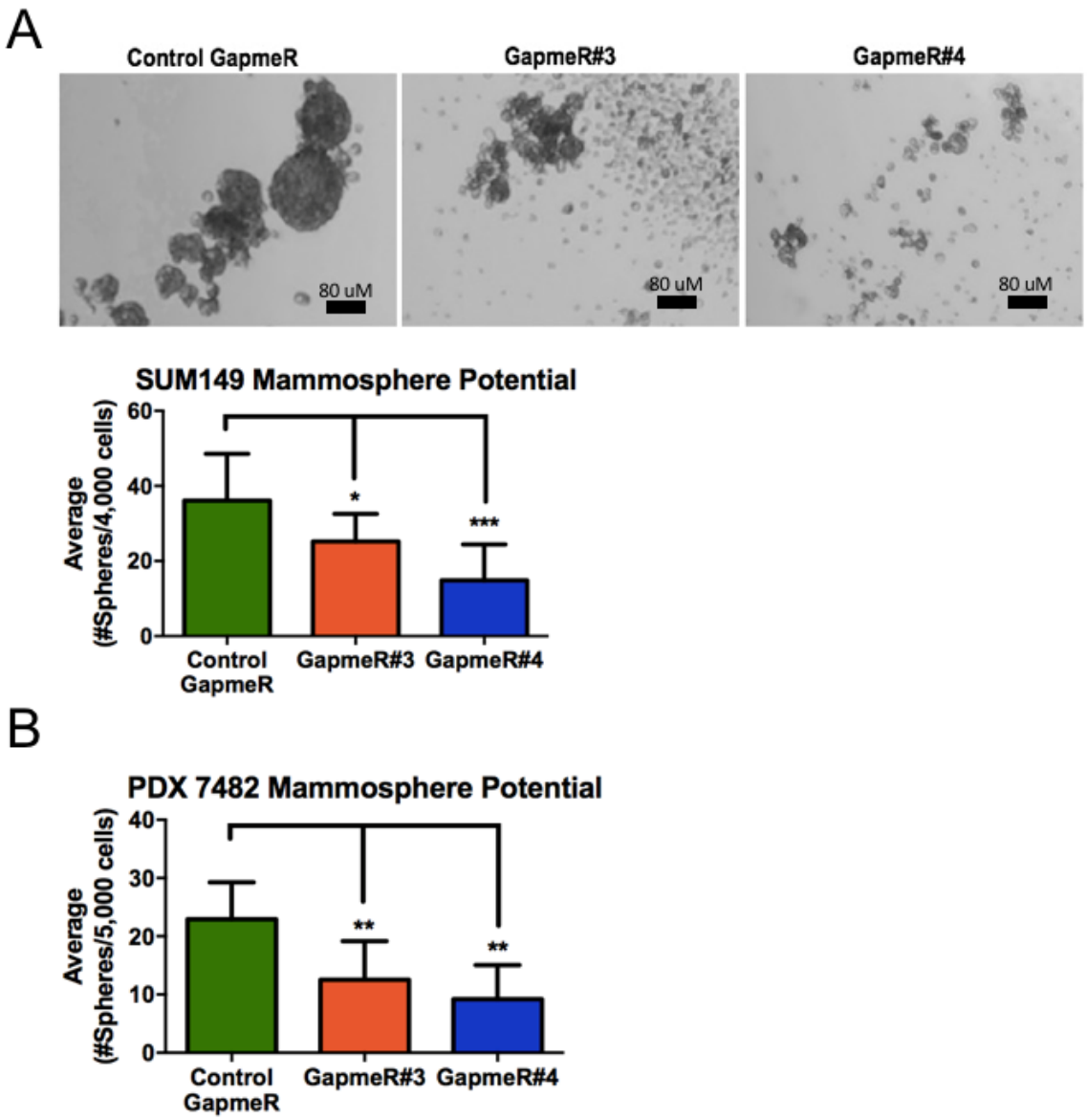

Figure 9. LINC00284 affects mammosphere-forming potential of TNBC cells. (A) SUM149 cells and PDX 7482 cells (**B)** were treated with 15 nM GapmeR (control or two GapmeRs specific to LINC00284) and seeded at 4,000 cells/well (SUM149, n=6) or 5,000 cells/well (PDX 7482, n=5) in MammoCult media in an ultra-low adherence plate. Spheroids greater than 50 uM in size were counted. Significance was determined by one-way ANOVA with Dunnett's post-test. Error bars represent SD.

3.4 LINC00284 and ALDH1A3 regulate common genes within their respective pathways

Since sub-cellular localization of a lncRNA in part defines its function[90], we first performed subcellular fractionation experiments followed by QPCR to identify where LINC00284 is localized. Like nuclear-localized lncRNA NEAT1[113], and in contrast to cytoplasmic-localized lncRNA DANCR[113], LINC00284 is predominantly localized to the nucleus (Fig. 10A). Nuclear lncRNA often function in gene expression changes[91]. Therefore, to identify genes regulated by LINC00284, we knocked down its expression using LNA GapmeRs in MDA-MB-468 cells and performed microarray gene expression analyses, using a fold change of 1.70 as a cutoff. A large number of genes were either upregulated (383) or downregulated (232) by LINC00284 (Fig. 10B), supporting the notion that it is important in gene expression regulation. Gene ontology (GO) term analysis showed that genes downregulated by LINC00284 tended to be involved in certain developmental and differentiation processes (Fig. 10C). Interestingly, genes that positively regulate muscle and cell differentiation and cellular development are suppressed by LINC00284. Contrarily, LINC00284 upregulates genes involved in alpha amino acid metabolism and lipid metabolism.

To further tease apart the ability of LINC00284 to control genes involved in ALDH1A3-mediated programs, we knocked down ALDH1A3 in the same cell line, MDA-MB-468, and performed microarray gene expression analysis, again with a fold change of 1.70 as a cutoff. ALDH1A3 knockdown caused a much smaller change in gene expression; with only 54 and 75 genes downregulated and upregulated, respectively, by ALDH1A3 (Fig. 10D). When overlapping the gene lists that are upregulated, however, 7 of the 75 ALDH1A3-upregulated genes are also upregulated by LINC00284 (Fig. 10E),

indicating that approximately 10% of positively-regulated ALDH1A3-mediated gene expression is also controlled by LINC00284. However, when looking at genes downregulated by ALDH1A3, we found that expression of approximately 25% - 13 of the 54 genes – was also controlled by LINC00284 (Fig. 10F). To show the robustness of our microarray analyses, we confirmed that these overlapping hits were, in fact, up or downregulated (Fig. 11A,B) via QPCR. Together, these data strongly suggest that a large portion of ALDH1A3-mediated gene expression changes are also controlled by LINC00284.

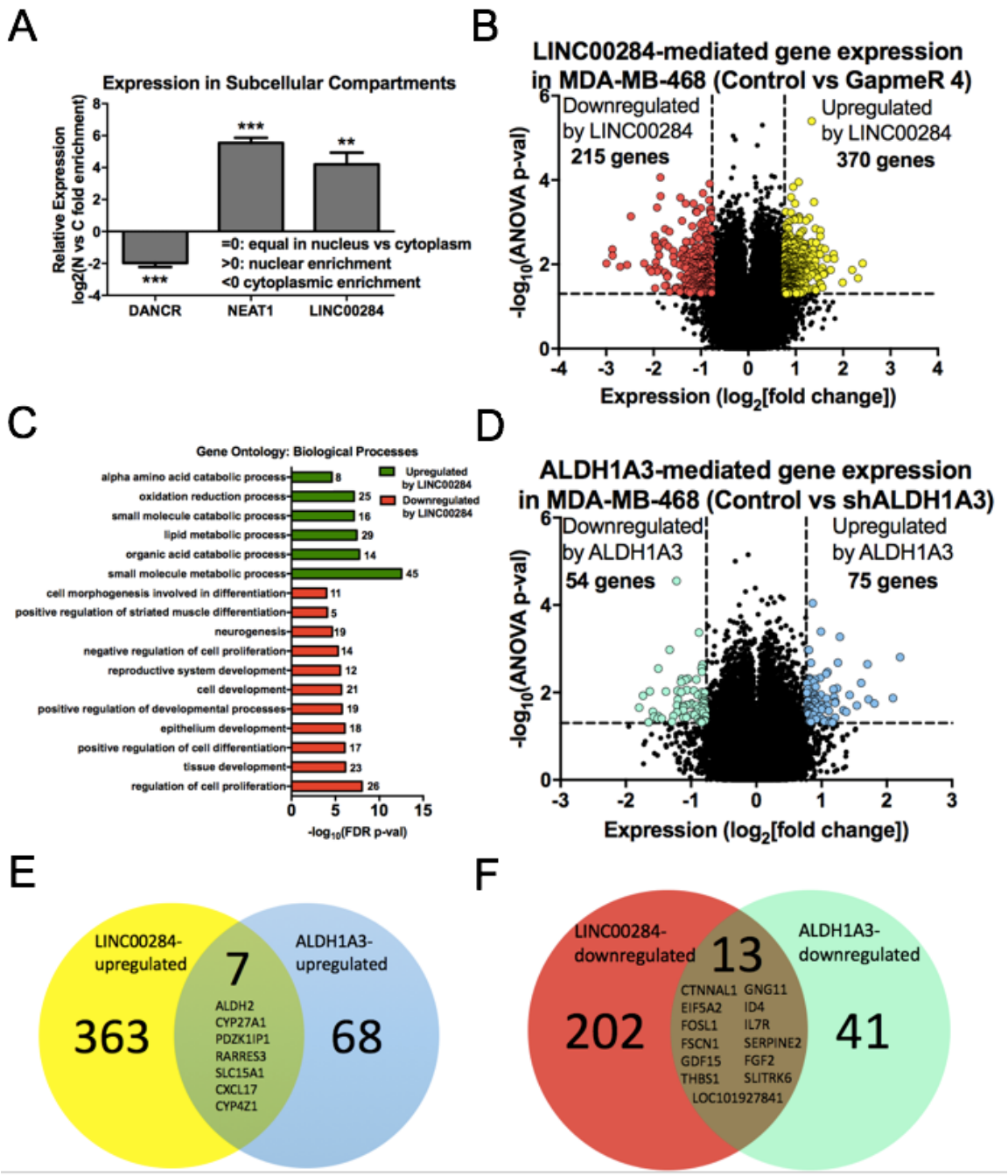

Figure 10. Nuclear LINC00284 and ALDH1A3 co-regulate gene expression in MDA-MB-468 cells. (A) QPCR analysis of nuclear and cytoplasmic fractions of MDA-MB-468 cells using primers against cytoplasmic RNA DANCR, nuclear lncRNA NEAT1, and LINC00284. Relative expression versus GAPDH is shown (n=4). Significance was determined using student's t-test. **(B)** Genome-wide gene expression changes induced by LINC00284 inhibition (control versus GapmeR#4-treated) is quantified in MDA-MB-468 cells using the Affymetrix Human Gene 2.0 ST microarray platform (n=3). The $\log_2$-fold change in expression is plotted versus the $-\log_{10}$(ANOVA p-val) of over 50,000 probes corresponding to 24,838 probesets covering 24,838 RefSeq (Entrez) genes. Only probes

with a >1.70-fold expression change and a p-val of >0.05 are indicated as coloured dots. **(C)** Gene ontology (GO) terms analysis was performed on LINC00284 up- or downregulated genes using GSEA software. The most significant GO terms with high numbers of genes enriched in those pathways are shown. **(D)** Genome-wide gene expression changes induced by ALDH1A3 knockdown (control versus ALDH1A3 knockdown) is quantified in MDA-MB-468 cells using the Affymetrix Human Gene 2.0 ST microarray platform (n=3). The $\log_2$-fold change in expression is plotted versus the $-\log_{10}$(ANOVA p-val) of over 50,000 probes corresponding to 24,838 probesets covering 24,838 RefSeq (Entrez) genes. Only probes with a >1.70-fold expression change and a p-val of >0.05 are indicated as coloured dots. **(E), (F)** The number of genes upregulated (E) or downregulated (F) by LINC00284 and ALDH1A3, with the genes co-regulated by both in the center of the Venn diagrams.

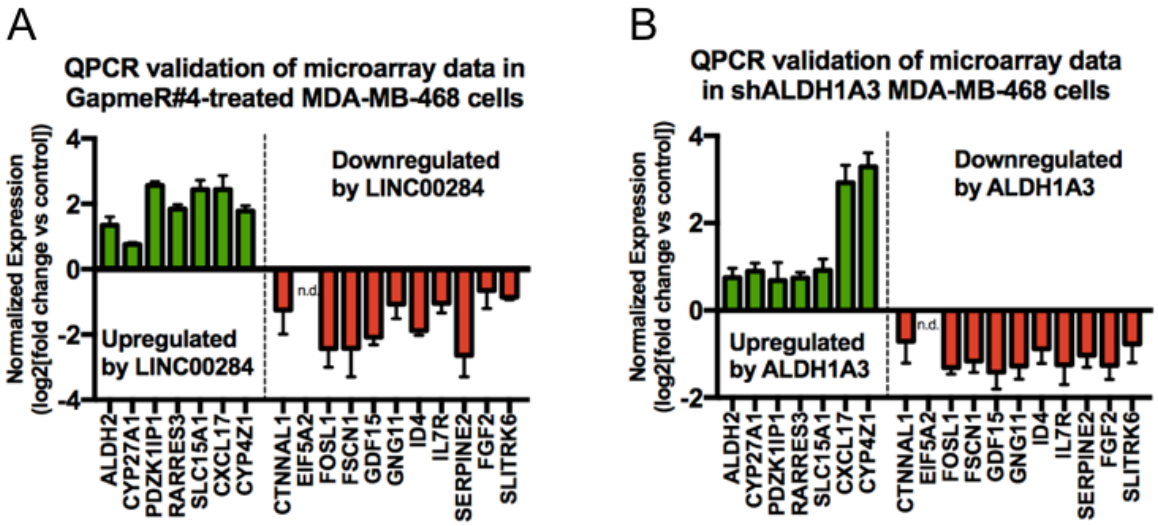

Figure 11. Genes identified as up- or downregulated by LINC00284 and ALDH1A3 via microarray are validated by QPCR. (A) QPCR was used to validate genes identified as co-regulated by LINC00284/ALDH1A3 via microarray, in MDA-MB-468 cells where LINC00284 is inhibited (n=4). **(B)** QPCR was used to validate genes as co-regulated by LINC00284/ALDH1A3 via microarray in MDA-MB-468 cells lacking ALDH1A3. Expression is normalized to reference genes PUM1 and ARF1 and represented as fold change over control cells (n=4). Error bars represent SD (n.d. = not detected).

3.5 Therapeutic inhibition of LINC00284 decreases breast tumor growth

In order to assess the therapeutic potential of LINC00284 inhibition, we implanted SUM149 and MDA-MB-468 tumors into NOD/SCID mice and treated them twice per week with 15 mg/kg control GapmeR or LINC00284-targeting GapmeR#4. In both models, there was a significant decrease in the rate of tumor growth when mice were treated with GapmeR#4 (Fig. 12A,B).

To improve the clinical validity of our *in vivo* studies, we also implanted mice with 2 cm^3 chunks of PDX 7482 tumors and then treated them as above. We noted a significant decrease in the rates of tumor growth when mice were treated with GapmeR#4 (Fig. 12C). Once the experiment had reached a humane endpoint, we sacrificed the mice, extracted their tumors, and assessed levels of 7AAD positivity within the tumor. The levels of 7AAD+ cells in the GapmeR#4-treated tumor was significantly increased compared to the control (Fig. 12D). We also wondered if LINC00284 inhibition in the PDX 7482 tumors would decrease the Aldefluor^{high} portion of cells within the tumors. Using FACS, we observed a modest yet non-significant decrease in Aldefluor positivity following LINC00284 knockdown, suggesting that inhibition of LINC00284 may somewhat alter the course of stem cell-mediated tumor growth (Fig. 12E). Importantly, when we seeded equal numbers of live cells extracted from the control GapmeR or GapmeR#4-treated PDX 7482 tumors into a mammosphere assay, we noted a dramatic hindrance of the GapmeR#4-treated tumors to form spheres (Fig. 12F). Together, these results show that LINC00284 inhibition *in vivo* does seem to impede the viability of the stem-like population of cells, thus causing decreased tumor growth.

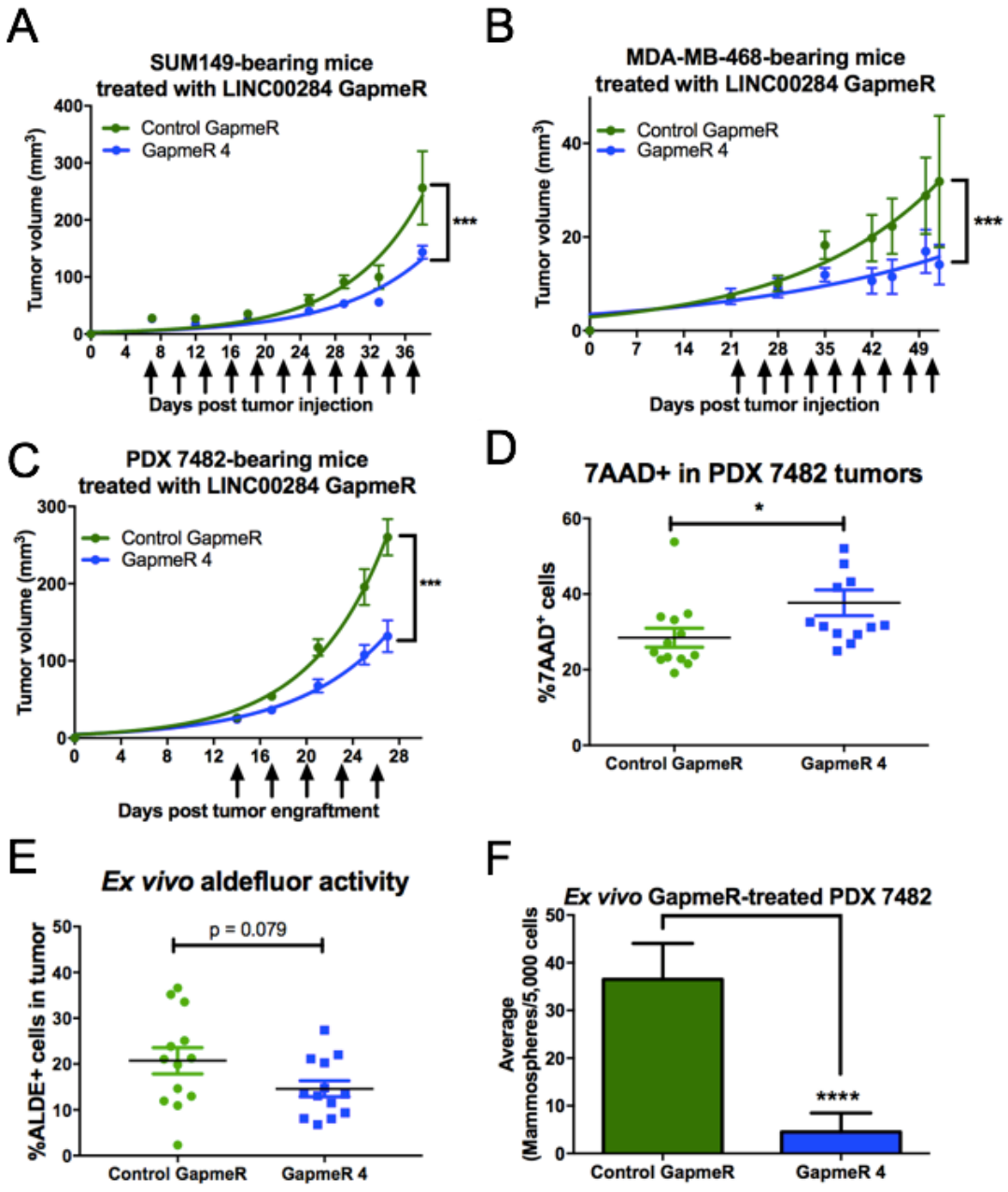

Figure 12. LINC00284 inhibition *in vivo* slows tumor growth and abrogates CSC populations. NOD/SCID mice bearing palpable SUM149 (n=6)**(A),** MDA-MB-468 (n=6) **(B),** or PDX 7482 (n=13) **(C)** tumors were treated by subcutaneous injection of either control GapmeR or anti-LINC00284 GapmeR#4 at a concentration of 15mg/kg, and tumor volumes were measured. Following sacrifice of mice bearing PDX 7482 tumors from **(C)**, single cell suspensions of the tumor were generated and flow cytometry performed to total 7AAD staining **(D)** and Aldefluor levels **(E)** in the tumor (n=6). The remaining cells were counted, dead cells excluded, and equal numbers of live cells seeded into MammoCult media in ultra-low adherence plates. Spheroids >50 uM were counted (n=6). Tumor volume significance was modelled using exponential regression. Significance in (F) determined using student's t-test. Error bars represent standard error.

3.6 LINC00284 binds to SND1, a basal-enriched protein

In order to determine the mechanism behind LINC00284 function in breast cancer, we performed RNA pulldown experiment. To do so, an *in vitro*-transcribed LINC00284 transcript was incubated in nuclear MDA-MB-468 lysates and the resulting LINC00284-bound proteins were analyzed with mass spectroscopy. The analyses found that one of the top LINC00284-bound proteins was Staphylococcal nuclease domain-containing protein 1 (SND1, also called Tudor-SN and p100; Fig 13A). In contrast, the antisense form of LINC00284 did not bind to SND1, but instead to non-specific cellular debris (Fig 13A), likely contaminants present in the pulldown. To further evaluate SND1 binding to LINC00284, we performed RNA immunoprecipitation (RIP) experiments, first confirming that our RNA immunoprecipitation pulled down endogenous SND1 by immunoblot (Fig. 13B). Using LINC00284-specific primers and QPCR, we quantified levels of LINC00284 bound to the SND1 pulldown and found a significant enrichment versus IgG, confirming that LINC00284 is indeed bound to endogenous SND1 (Fig. 13C). Using TCGA data, we also observed that like LINC00284, SND1 is significantly enriched in basal-like breast cancer, suggesting that LINC00284 and SND1 are co-enriched (Fig. 13D). Together, these results suggest that LINC00284 may mediate some of its effects through binding to SND1.

A

Unique LINC00284-antisense pulldown proteins	Peptide Spectral Matches
U2 small nucleolar ribonucleoprotein (SNRPB2)	6
Histone H1x (H1FX)	7
14-3-3 protein epsilon (YWAE)	5
Unique LINC00284-sense pulldown proteins	
Staphlyococcal nuclease domain-containing protein 1 (SND1)	22
DNA-dependent protein kinase catalytic subunit (PRKDC)	19
High density lipoprotein binding protein (HDLBP)	2
DEAD-box helicase 3, X-linked (DDX3X)	7
TAR DNA-binding protein (TARDBP)	2

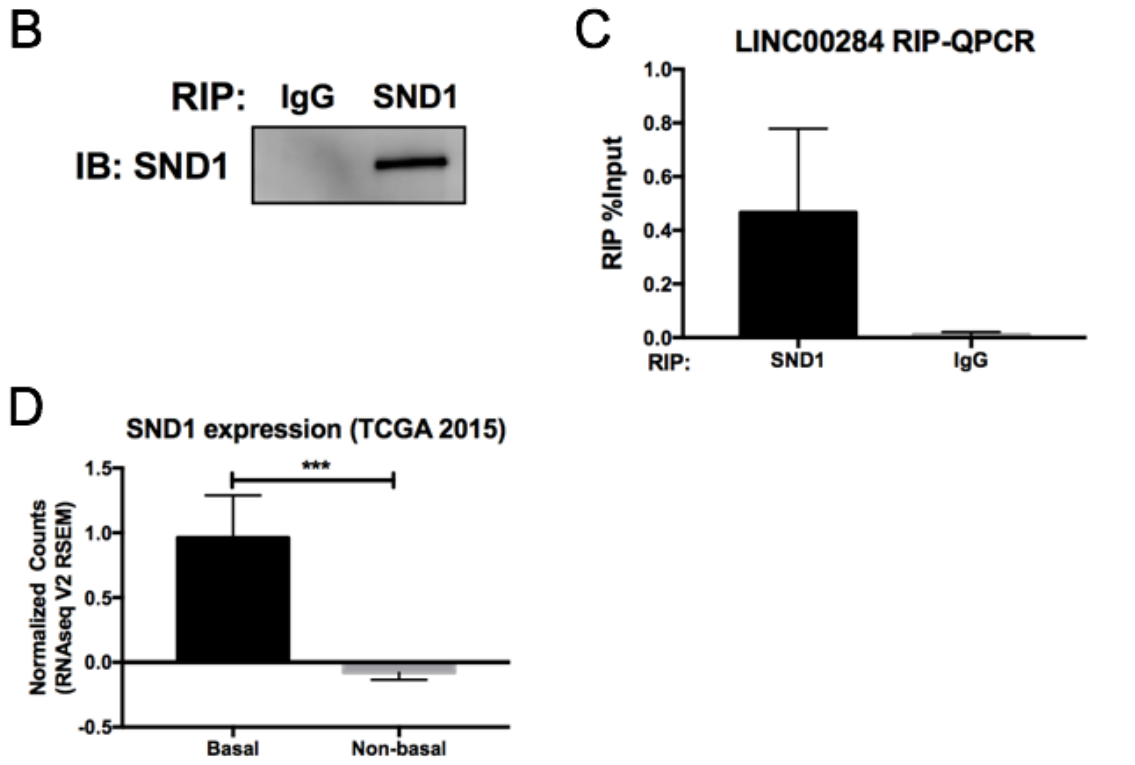

Figure 13. LINC00284 binds to basal-enriched protein SND1. (A) RNA pulldown using either biotinylated sense-LINC00284 or anti-sense LINC00284 as a negative control was performed, and mass spectroscopy identified LINC00284-bound proteins. Unqiue proteins with high numbers of peptide spectral matches are shown. RNA immunoprecipitation (RIP) was performed using antibodies confirmed by western blot to pulldown SND1 **(B),** and relative levels of LINC00284 transcript bound to SND1 versus non-specific IgG over total input was identified using QPCR (n=3) **(C).** RNA-seq expression of SND1 in basal-like versus non-basal breast cancer patients (TCGA Cell 2015) was analyzed and data extracted using cBioportal. Significance was determined using student's t-test. Error bars represent SD in (C), and standard error in (D).

3.7 LINC00284 displays high levels of genomic occupancy

As a nuclear lncRNA, it is likely LINC00284 regulates gene expression either directly through RNA-chromatin interactions or indirectly through protein complexes. In order to further study the role of LINC00284 in the genome, we performed ChIRP-seq in MDA-MB-468 cells to retrieve genomic DNA bound to LINC00284, which was then sequenced and analyzed (Fig. 14A). We confirmed that the even or odd-pool of biotinylated probes used to isolate LINC00284 were successful in enriching LINC00284 above input compared to enriched a non-specific transcript, GAPDH (Fig. 14B). Our analyses showed that LINC00284 is present at high levels throughout the genome; notably, thousands of regions throughout gene bodies (i.e. intron, exon, 3'UTR, 5'UTR, promoter, ncRNA, and transcription termination sites) are occupied by LINC00284 (Fig. 14C). While our previous microarray analysis showed that LINC00284 regulates a total of 585 genes, our ChIRP-seq indicated that LINC00284 is bound to genic regions of 3519 genes (Fig. 14D). This represents approximately 10% of directly LINC00284-regulated gene expression that can be attributed to LINC00284 genomic occupancy, suggesting that most of the genome-bound LINC00284 is bound "weakly", which is not enough to regulate gene expression. Furthermore, most of the overlapping genes had a large number of peaks associated with them, suggesting that increased LINC00284 occupancy within a gene may lead to a higher level of expression regulation by LINC00284 (Fig. 14E), supporting the notion that LINC00284-dependent gene expression requires high genic binding.

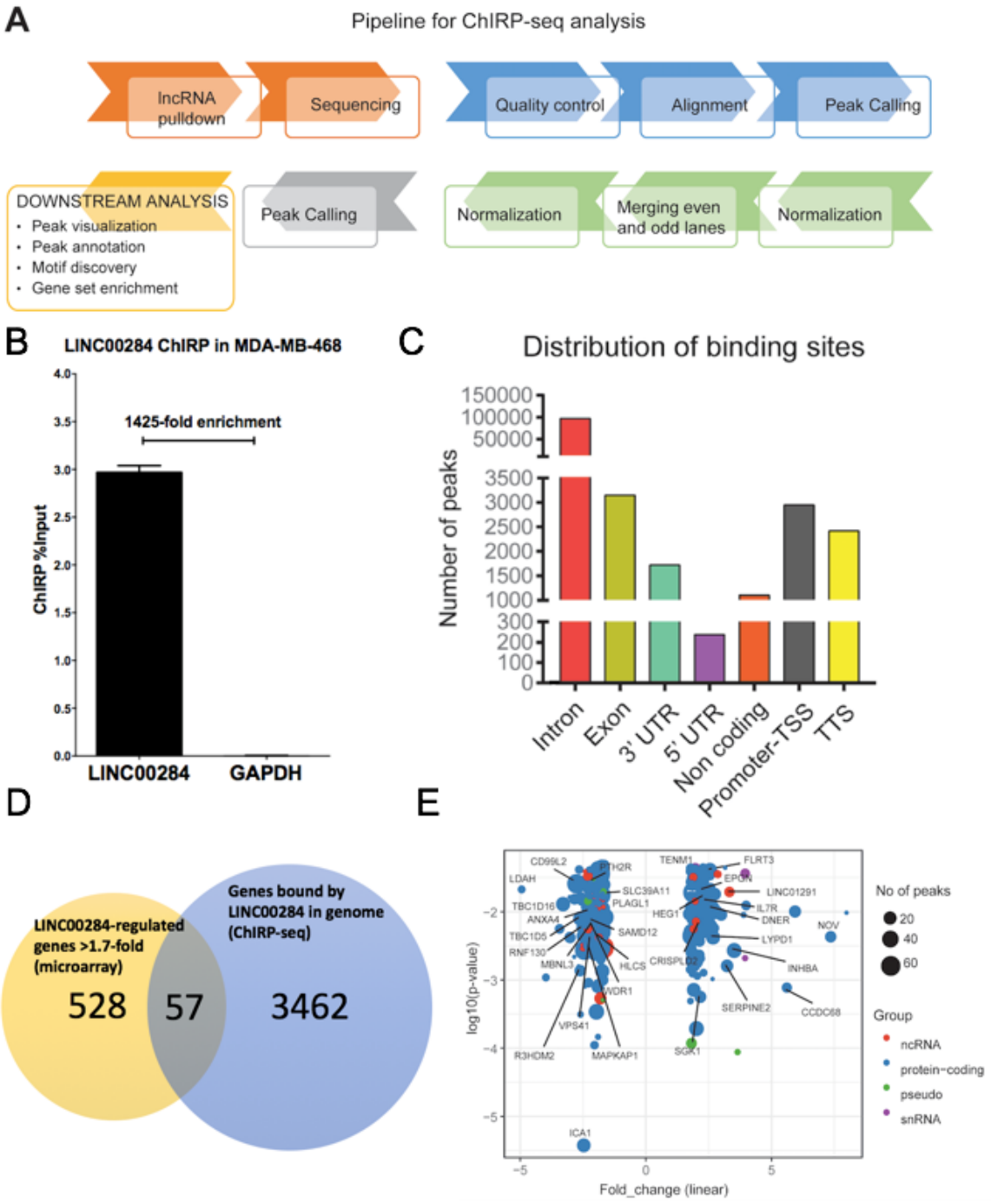

Figure 14. LINC00284 is bound to a large number of genic regions in MDA-MB-468 cells. (A) Schematic diagram showing the workflow of our ChIRP-seq experiment. **(B)** QPCR on the biotinylated probe-bound fraction confirms that the ChIRP assay successfully enriches LINC00284 RNA transcripts over input, versus a GAPDH negative control. Error bars represent SD. **(C)** ChIRP-seq analysis shows the distribution of LINC00284-bound peaks relative to genic regions of interest. **(D)** A venn diagram shows the overlapping genes (not peaks) regulated by LINC00284 in the microarray gene

expression experiment performed in Fig. 10B and the ChIRP-seq experiment. **(E)** The common genes identified in (D) (57) plotted according to the $-\log_{10}$(ANOVA p-val) verus the fold-change in expression following LINC00284 knockdown. The number of LINC00284-bound peaks associated with that gene (circle size) and the gene group (color) is also shown.

CHAPTER 4 DISCUSSION

4.10 Preamble

TNBC/basal-like breast cancer tumors lack robust therapeutic targets[16,23]. A major driver of tumorigenesis is the presence of CSCs, which are present in higher numbers in TNBC/basal-like tumors[54–62]. Therapeutic strategies are in development to inhibit CSC-specific markers and pathways, with limited success. A major focus has been placed on inhibitors of the products of protein-coding genes, such as Wnt[64], CD44[71–73], or ALDH1A3[75,77,79] inhibitors, despite protein-coding genes only being responsible for 2% of genomic transcription[86]. Therefore, it is likely that ncRNAs, such as lncRNAs, may play a vital, yet undiscovered role in CSC-mediated tumor progression. Thus, the identification and therapeutic inhibition of lncRNAs important in these CSC-specific pathways and enriched in CSCs may be a valid strategy for targeted therapies. Here, we report the existence of a previously unstudied CSC-enriched lncRNA, LINC00284, and for the first time, demonstrate that inhibition of a CSC-enriched lncRNA *in vivo* has effects on tumor progression. Furthermore, we characterize the likely molecular mechanism behind LINC00284 function in basal-like/TNBC.

4.11 A clinically-guided RNA-seq screen identifies novel CSC-enriched lncRNA LINC00284

The search for lncRNA targets is often a top-down approach rather than a bottom-up approach. However, the discovery of novel, unstudied targets is often dependent on RNA-seq data (i.e. a bottom-up approach) which is able to capture the entire transcriptome. For this reason, we sought to use TCGA RNA-seq data of breast cancer

patients to identify the most enriched lncRNAs in TNBC. A recently published study used this approach to identify the top 50 most TNBC-enriched lncRNA, and they discovered novel lncRNA LINP1 as being vital in regulating DNA repair processes in TNBC. Importantly, they successfully demonstrated that shRNA-mediated inhibition of LINP1 sensitized TNBC cells to radiotherapy *in vivo*[103], suggesting that this approach could be used to identify novel TNBC-enriched targets. We made use of their data identifying the top 50 most enriched lncRNAs in TNBC, and screened those lncRNA in TNBC CSCs to potentially identify CSC-enriched lncRNA targets. Our screen revealed 10 lncRNAs across two CSC models as highly enriched – one of which, LINC00284, could also predict patient survival, thus fulfilling our prioritization strategy.

In our screen, we were only able to detect expression of 37 of the 50 lncRNA, which may be due to multiple reasons. First, lncRNA transcripts (and RNA transcripts in general) are somewhat inaccurate at the 5' ends, due to constraints in transcript assembly during RNA-seq data processing[87,132]. With the ambiguity in isoform and transcript assembly, primer pairs generated to detect specific lncRNAs may have been unable to bind to the true transcripts. Furthermore, lncRNAs are generally expressed in a context-specific manner[102] – while some transcripts may have been enriched in TCGA TNBC patients, they were not enriched in two specific TNBC models used to assess expression initially. The exclusion of these 13 transcripts is therefore the first filter in our screen. An ideal lncRNA target must be enriched in all or most TNBC tumors across different patients. Since we were unable to detect them at all, they are therefore poor targets.

To control for the sensitivity and specificity of our CSC models and approach, we included QPCR analyses to see if previously studied breast CSC lncRNAs or the most

expressed lncRNAs in breast cancer would be enriched in our models. We observed that all of the previously cited breast CSC-involved lncRNAs were indeed enriched in our CSC models, and contrarily, none of the indiscriminately most expressed lncRNAs in breast cancer were enriched in our CSC models (Fig. 3A,B). We were therefore confident in the robustness of our approach in identifying lncRNAs that may play a role in CSC biology. LINC00284 was the only ALDH1A3-regulated lncRNA that could demonstrably predict patient survival and was therefore the best target to further study. However, the other 9 CSC-enriched lncRNA may also be important in CSC function, and should be studied further as well.

4.12 LINC00284 is a novel downstream factor of the ALDH1A3 pathway

The exact role of ALDH1A3 in breast cancer biology is poorly understood. Some have hypothesized that ALDH1A3 regulates CSC biology through RA production[133], which induces gene expression programs that may be related to cellular differentiation. However, the direct link between ALDH1A3-mediated RA signaling and CSC biology is still not clear. Our recent work has shown that ALDH1A3-downstream signaling diverges in different contexts, inducing very different effects in different breast cancer models[76]. Furthermore, ALDH1A3 and RA produce quite distinct gene expression programs, with little overlap[80]. Thus, the ALDH1A3$^+$ CSC response to RA is disparate in different tumor contexts, indicating ALDH1A3-mediated CSC function may not be dependent on RA. Instead, there are likely unknown, additional factors that play a role in ALDH1A3-mediated CSC function.

Here, we have identified LINC00284 as a CSC-enriched lncRNA that is also regulated by ALDH1A3 across different breast cancer contexts (Fig. 5A,B). Extending our analyses beyond cell lines, we observed that LINC00284 and ALDH1A3 were co-expressed in many patient tumors and multiple cell lines, indicating that LINC00284 is likely a downstream factor of the ALDH1A3 pathway. With the knowledge that LINC00284 is enriched in ALDH1A3^{+} breast CSCs (Fig. 2A,B), it suggests that LINC00284 is likely a factor in the ALDH1A3 pathway. Our gene expression analyses support this finding, with 10% and 25% of ALDH1A3-upregulated genes and downregulated genes controlled by LINC00284, respectively (Fig. 10E,F). Therefore, for the first time, we report a novel lncRNA that may be important in ALDH1A3-mediated signaling in CSCs.

Despite our gene expression analyses, it is unclear whether these gene expression changes are conferred directly upon CSCs. The evidence we have shown here, however, supports the notion that LINC00284 may play a role in CSC biology. In TNBC cells, LINC00284 upregulated genes involved in metabolic processes, such as alpha-amino acid metabolism and lipid metabolism (Fig. 10C). Intriguingly, these pathways play hallmark roles in stem cell maintenance[134], and their importance in CSC biology is just beginning to be understood[135]. For example, inhibition of fatty acid oxidation (FAO) by etomoxir, an FAO inhibitor, significantly decreases breast CSC mammosphere formation[136]. Additionally, alpha-amino acid metabolism has vital roles in CSC function in pancreatic CSCs, where glutamine deprivation was shown to significantly inhibit self-renewal, stemness, and CSC survival[137]. Unfortunately, the importance of ALDH1A3 in CSC

metabolism is still unclear; however, it is involved in some of these metabolic processes in the normal context[84].

Conversely, LINC00284 suppresses genes that positively regulate muscle and cell differentiation and cellular development (Fig. 10C). This is consistent with a proposed role in preventing stem-like cells from committing to cellular lineages; which may be why ALDE+ cells express high levels of LINC00284. Unfortunately, however, our data cannot suggest a direct relationship between LINC00284 and the differentiation process in breast CSCs.

4.13 LINC00284 confers oncogenic benefits to breast cancer cells

A contentious issue amongst ncRNA researchers is whether or not ncRNA genes, such as lncRNA, actually encode functional products or if their products are just a result of "accidental" transcription[138,139]. In fact, many advocate that the null hypothesis is that ncRNAs are non-functional junk[138]. If a lncRNA confers a phenotype upon cells, then it likely has some importance in maintaining that cellular phenotype. Here, we have shown that inhibition of LINC00284 with GapmeRs dramatically decreases cellular growth, and increases cell apoptosis (Fig. 7). We therefore report that LINC00284 is an oncogenic functional lncRNA.

Our data revealed that breast cancer cells are indiscriminately at a disadvantage when LINC00284 is inhibited (Fig. 7). While we were successful in targeting TNBC cells, we also observed that LINC00284 inhibition significantly killed MCF7 cells, which are not TNBC, but ER^+, suggesting that LINC00284 inhibition may not be specific for TNBC. It is likely that in order for LINC00284 inhibition to successfully kill cells, it

must be expressed at moderate levels (Fig. 6). Our GTEx results showed that LINC00284 was lowly present in normal tissues (Fig. 5), suggesting that there would be few, if any, toxicities in body organs if LINC00284 was inhibited *in vivo*. In contrast, others have suggested inhibition of oncogenic lncRNAs such as MALAT1 for the treatment of cancer[119], without considering that MALAT1 function is paramount in nuclear paraspeckle dynamics in normal cells[140]. Inhibition of MALAT1 in mammary epithelial cells caused aberrant mitotic changes, including G2/M arrest and increased cell death[140]. Therefore, the successful targeting of LINC00284 is aided by the fact that it is lowly expressed in normal tissue – its function is likely prevalent at moderate expression levels.

While LINC00284 inhibition significantly increased death of heterogeneous populations of cells, we also aimed to assess whether it could prevent CSC propagation and maintenance. Using a mammosphere assay with our two CSC models from our screen, we showed that LINC00284 inhibition significantly decreased mammosphere-forming capability (Fig. 9). If LINC00284 function was not important for CSC function, we would have observed no change in the mammosphere assays following LINC00284 inhibition. Thus, LINC00284 inhibition is able to abrogate CSC function.

4.14 LINC00284 can be targeted in vivo to slow tumor growth and possibly abrogate CSC pools

Very few studies have successfully utilized LNA GapmeRs to inhibit lncRNAs *in vivo*. Here, we used multiple breast cancer models to generalize our results unto multiple different TNBC tumors. We observed significant decreases in the tumor growth rates of SUM149 and MDA-MB-468 tumors when treated with LINC00284-targeting GapmeRs

subcutaneously (Fig. 12A-C). Importantly, in PDX 7482, which is much more reflective of true patient tumor biology[141], we observed a dramatic decrease in the rate of tumor growth (Fig. 12C). The ability of LINC00284 inhibition to cause macroscopic changes in tumor size is supported by the fact that LINC00284 inhibition *in vitro* is indiscriminate in causing cancer cell death, as previously discussed

It was necessary to identify whether CSC populations were affected by LINC00284 inhibition in order to decrease CSC pools and prevent recurrence. We noted a non-significant but modest decrease in the Aldefluor^{high} percentage of cells within the tumor, indicating some successful CSC targeting (Fig. 12E). Since LINC00284 inhibition kills both CSCs and non-CSCs (Fig. 7, Fig. 9), it is more difficult to tease apart differences in Aldefluor activity following LINC00284 inhibition. Importantly, however, the inhibition of CSC maintenance was confirmed by showing that GapmeR#4-treated tumors were significantly impaired at forming mammospheres compared to the control-treated mice (Fig 12F). Therefore, for the first time in the literature, we show that a CSC-enriched lncRNA can be inhibited *in vivo* to successfully target CSC populations.

4.15 LINC00284 binds to cancer stem cell factor SND1

In order to learn how LINC00284 functions in basal-like/TNBC, we performed extensive molecular characterization analyses of LINC00284. Our RNA pulldown experiments and RIP experiments revealed that in basal-like/TNBC, LINC00284 is bound to SND1 within nuclei (Fig. 13A-C). Interestingly, we observed that like LINC00284, SND1 was also enriched in basal-like breast cancer (Fig. 13D). This

corroborates a previous report that SND1 tended to be highly expressed in TNBC cell lines compared to non-TNBC cell lines[142].

The SND1 protein is a highly multifunctional protein, with reported roles in gene transcription[143], RNA editing[144], and regulation of apoptosis[145]. The function of SND1 in breast cancer has been actively studied over the past decade. There is a large body of evidence showing that SND1 promotes breast cancer metastasis and poor outcomes[142,146–148], likely through Smad signaling and induction of metastasis-promoting factor TGFβ. The role of SND1 in breast CSCs has also recently been elucidated, where it is stabilized by binding to metadherin (MTDH) – together, the stabilized MTDH-SND1 complex is vital for CSC expansion and tumorigenesis[147]. In TNBC PDX cells with shRNA-mediated SND1 knockdown, mammosphere formation was significantly impaired independent of MTDH levels, indicating that SND1 is a key effector behind CSC expansion[147]. Given that SND1 is a highly multifunctional protein, the exact mechanisms through which LINC00284 works with SND1 will still need to be elucidated.

4.16 LINC00284 may mediate gene expression regulation through increasing genomic occupancy

Our ChIRP experiments revealed that LINC00284 is bound to a large number of genomic sites (Fig. 14C), which is consistent with the potential role LINC00284 in regulating gene expression. However, we observed that only 10% of the genes significantly regulated by LINC00284 in our microarray data were bound by LINC00284 at the chromatin level (Fig. 14D). This may suggest that in order for LINC00284 to change levels of gene transcription, it must be have high levels of genic coverage – thus,

genes with few LINC00284 peaks would not change in expression upon LINC00284 inhibition, and would not reflect as up- or downregulated in our microarray.

It is also likely that other factors contribute to both LINC00284-mediated gene expression as observed in our microarray – indirect effects, or downstream signaling pathways, which may account for the rest (i.e. 90%) of LINC00284-mediated gene expression. For example, LINC00284 inhibition could cause an increase in hypothetical protein X expression, which then itself prevents gene expression of numerous other genes, which would be reflected in our microarray but not in our ChIRP.

The high numbers of LINC00284-bound genes may be mediated by numerous pieces of chromatin machinery. Of note, SND1, which contains an RNA-binding domain, plays a key role in gene expression regulation by acting as a transcription factor bridge where it promiscuously binds to dozens of different transcription factors which have motifs throughout the genome[149]. If SND1-bound LINC00284 is co-bound to at least some of those regions, then it may explain a large portion of the peaks associated with LINC00284. Furthermore, it is possible that other proteins bound to LINC00284 identified in our RNA pulldown may also contribute to LINC00284 binding (such as PRKDC; Fig. 13A).

4.17 Limitations

As a proof-of-principle, we have shown that a CSC-enriched lncRNA can be inhibited *in vivo.* However, as suggested by our *in vitro* results (Fig. 6, Fig. 7), LINC00284 inhibition may be dependent on LINC00284 expression. Thus, not all tumors may respond. Patient tumors expressing low levels of LINC00284 would likely be

resistant to LINC00284-inhibiting therapies. High lncRNA expression would have to be a prerequisite of LINC00284 inhibition. Furthermore, extensive pre-clinical *in vivo* studies will be necessary to reveal if any long-term toxicities exist from LINC00284 inhibition.

We cited the importance of certain metabolic processes in stem cell biology[135], and our gene expression experiments revealed these pathways as being upregulated by LINC00284. However, these experiments were performed *in vitro* under conditions where growth conditions were curated, and may not accurately reflect an *in vivo* cellular environment[150]. Therefore, while our studies reveal interesting metabolic pathways regulated by LINC00284 *in vitro*, care should be taken when generalizing these results *in vivo.*

While most of LINC00284 was expressed in the nucleus and, from our evidence, plays a role in gene expression and transcription, there may still be other ways in which LINC00284 functions to alter gene expression which we have not explored. For example, it could function as a miRNA sponge to regulate gene expression in indirect ways. To elucidate these alternative functions, further molecular interrogation is required.

4.18 Future directions

Our clinically-guided RNA-seq screen identified 10 highly enriched lncRNAs in basal-like/TNBC and two different CSC models. While we focused on LINC00284, mainly because it also predicted patient outcomes, any of the other 9 lncRNAs still represent valuable lncRNAs for further study. Uncovering their roles in basal-like/TNBC or CSC biology would still elucidate new information behind these novel factors and how

they drive tumor growth and CSC persistence, and may reveal new pathways important in cancer.

The novel discovery of the function behind an unstudied lncRNA often occurs in only one context. Here, we report that LINC00284 is important in CSC functioning, possibly by binding to CSC factor SND1. However, we did not directly explore the relationship between ALDH1A3^{+} CSCs, SND1, and LINC00284. Furthermore, LINC00284 may be performing other, unexplored roles in CSCs, independent of SND1. For example, we did not explore the role of LINC00284 in altering the CSC metabolome, even though our gene expression data suggests that LINC00284 may have a role in regulating some metabolic pathways. Combining our ChIRP data with full-scale metabolomics approaches, one could characterize the effect of LINC00284 inhibition on genes involved in altering the metabolome of CSCs, to alter their function.

We also performed RNA pulldown experiments which revealed SND1 as a major LINC00284-binding protein. Ideally, further functional characterization of the role of SND1 and LINC00284 in breast CSCs is required. Chromatin immunoprecipitation (ChIP) experiments of SND1 in cells where LINC00284 is inhibited compared to control cells could reveal LINC00284-dependent SND1 functions. The combination of this data with our ChIRP-seq data may reveal interesting mechanisms behind LINC00284-driven gene expression changes. Additionally, our RNA pulldown analyses revealed other proteins, such as DNA-dependent protein kinase catalytic subunit (PRKDC) and high density lipoprotein binding protein (HDLBP) as LINC00284-binding proteins. The role of these proteins binding to LINC00284 in breast cancer may also be important. For example, PRKDC is known to regulate chemoresistance in breast cancer, and is

associated with high tumor grades and poor prognosis[151]. Therefore, further interrogation of the other LINC00284-bound proteins is required.

We ultimately explored the role of LINC00284 in breast cancer. It is possible, and likely, that others will identify its role in other cancers, in normal cells, and possibly in the developmental context. Functional approaches similar to the those presented here (i.e. ChIRP, RIP, RNA pulldown) in those contexts will reveal other mechanisms behind LINC00284 function.

4.19 Conclusion

The emerging field of lncRNA research has shown that many lncRNAs play key roles in cancer[102]. Yet thousands of lncRNAs remain functionally uncharacterized and unknown. Here, we utilize a clinically-guided RNA-seq screen to identify novel LINC00284 as an ALDH1A3-regulated lncRNA that is enriched in TNBC and CSCs. We determine that LINC00284 functions by modulating gene expression programs. For the first time, we also successfully inhibit a CSC-enriched lncRNA *in vivo* to abrogate tumor growth.

\

BIBLIOGRAPHY

1. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10,** 789–799 (2004).

2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).

3. Collado, M. & Serrano, M. Senescence in tumours: evidence from mice and humans. *Nat. Rev. Cancer* **10,** 51–57 (2010).

4. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88,** 593–602 (1997).

5. Ventura, A. *et al.* Restoration of p53 function leads to tumour regression in vivo. *Nature* **445,** 661–665 (2007).

6. Sherr, C. J. Principles of tumor suppression. *Cell* **116,** 235–246 (2004).

7. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148,** 59–71 (2012).

8. Milholland, B. *et al.* Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8,** 15183 (2017).

9. Sobhian, B. *et al.* RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites. *Science* **316,** 1198–1202 (2007).

10. Ohta, T., Sato, K. & Wu, W. The BRCA1 ubiquitin ligase and homologous recombination repair. *FEBS Lett.* **585,** 2836–2844 (2011).

11. Heerboth, S. *et al.* EMT and tumor metastasis. *Clin. Transl. Med.* **4,** (2015).

12. Breast cancer statistics - Canadian Cancer Society. *www.cancer.ca* Available at: http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on. (Accessed: 7th June 2018)

13. Langlands, F. *et al.* Contralateral breast cancer: incidence according to ductal or lobular phenotype of the primary. *Clin. Radiol.* **71,** 159–163 (2016).

14. Sinn, H.-P. & Kreipe, H. A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. *Breast Care* **8,** 149–154 (2013).

15. Carey, L. A. *et al.* Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* **295,** 2492–2502 (2006).

16. Schneider, B. P. *et al.* Triple-negative breast cancer: risk factors to potential targets. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **14,** 8010–8018 (2008).

17. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

18. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* **8,** R76 (2007).

19. Nguyen, P. L. *et al.* Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **26,** 2373–2378 (2008).

20. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12,** R68 (2010).

21. Peddi, P. F., Ellis, M. J. & Ma, C. Molecular Basis of Triple Negative Breast Cancer and Implications for Therapy. *Int. J. Breast Cancer* **2012,** (2012).

22. Shipitsin, M. *et al.* Molecular definition of breast tumor heterogeneity. *Cancer Cell* **11,** 259–273 (2007).

23. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121,** 2750–2767 (2011).

24. Kim, G. *et al.* FDA Approval Summary: Olaparib Monotherapy in Patients with Deleterious Germline BRCA-Mutated Advanced Ovarian Cancer Treated with Three or More Lines of Chemotherapy. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **21,** 4257–4261 (2015).

25. Dalerba, P., Cho, R. W. & Clarke, M. F. Cancer stem cells: models and concepts. *Annu. Rev. Med.* **58,** 267–284 (2007).

26. Marcato, P., Dean, C. A., Giacomantonio, C. A. & Lee, P. W. If cancer stem cells are resistant to current therapies, what's next? *Future Oncol.* **5,** 747–750 (2009).

27. Rich, J. N. & Bao, S. Chemotherapy and cancer stem cells. *Cell Stem Cell* **1,** 353–355 (2007).

28. Sultan, M. *et al.* Epigenetic Silencing of TAP1 in Aldefluor+ Breast Cancer Stem Cells Contributes to Their Enhanced Immune Evasion. *STEM CELLS* **36,** 641–654.

29. Sultan, M. *et al.* Hide-and-seek: the interplay between cancer stem cells and the immune system. *Carcinogenesis* **38,** 107–118 (2017).

30. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3,** 730–737 (1997).

31. Chaffer, C. L. & Weinberg, R. A. How does multistep tumorigenesis really proceed? *Cancer Discov.* **5,** 22–24 (2015).

32. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 3983–3988 (2003).

33. Ginestier, C. *et al.* ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell* **1,** 555–567 (2007).

34. Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumors. *Cancer Res.* **63,** 5821–5828 (2003).

35. Bapat, S. A., Mali, A. M., Koppikar, C. B. & Kurrey, N. K. Stem and progenitor-like cells contribute to the aggressive behavior of human epithelial ovarian cancer. *Cancer Res.* **65,** 3025–3029 (2005).

36. Yang, Z. F. *et al.* Significance of CD90+ cancer stem cells in human liver cancer. *Cancer Cell* **13,** 153–166 (2008).

37. Eramo, A. *et al.* Identification and expansion of the tumorigenic lung cancer stem cell population. *Cell Death Differ.* **15,** 504–514 (2008).

38. Dalerba, P. *et al.* Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 10158–10163 (2007).

39. Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell Stem Cell* **1,** 313–323 (2007).

40. Mitsutake, N. *et al.* Characterization of side population in thyroid cancer cell lines: cancer stem-like cells are enriched partly but not exclusively. *Endocrinology* **148,** 1797–1803 (2007).

41. Ben-Porath, I. *et al.* An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40,** 499–507 (2008).

42. Seymour, T., Twigger, A.-J. & Kakulas, F. Pluripotency Genes and Their Functions in the Normal and Aberrant Breast and Brain. *Int. J. Mol. Sci.* **16,** 27288–27301 (2015).

43. Martinez-Cruzado, L. *et al.* Aldh1 Expression and Activity Increase During Tumor Evolution in Sarcoma Cancer Stem Cell Populations. *Sci. Rep.* **6,** 27878 (2016).

44. Nogami, T. *et al.* Expression of ALDH1 in axillary lymph node metastases is a prognostic factor of poor clinical outcome in breast cancer patients with 1–3 lymph node metastases. *Breast Cancer* **21,** 58–65 (2014).

45. Wei, W. *et al. Relationship of CD44 + CD24 -/low breast cancer stem cells and axillary lymph node metastasis*. S6 (BioMed Central, 2012). doi:10.1186/1479-5876-10-S1-S6

46. Marcato, P. *et al.* Aldehyde dehydrogenase activity of breast cancer stem cells is primarily due to isoform ALDH1A3 and its expression is predictive of metastasis. *Stem Cells Dayt. Ohio* **29,** 32–45 (2011).

47. Zhong, Y. *et al.* ALDH1 is a better clinical indicator for relapse of invasive ductal breast cancer than the CD44+/CD24- phenotype. *Med. Oncol. Northwood Lond. Engl.* **31,** 864 (2014).

48. Ma, F. *et al.* Aldehyde dehydrogenase 1 (ALDH1) expression is an independent prognostic factor in triple negative breast cancer (TNBC). *Medicine (Baltimore)* **96,** (2017).

49. Miyoshi, Y. *et al.* Differences in expression of the cancer stem cell marker aldehyde dehydrogenase 1 among estrogen receptor-positive/human epidermal growth factor receptor type 2-negative breast cancer cases with early, late, and no recurrence. *Breast Cancer Res.* **18,** 73 (2016).

50. Kida, K. *et al.* Effect of ALDH1 on prognosis and chemoresistance by breast cancer subtype. *Breast Cancer Res. Treat.* **156,** 261–269 (2016).

51. Croker, A. K. & Allan, A. L. Inhibition of aldehyde dehydrogenase (ALDH) activity reduces chemotherapy and radiation resistance of stem-like ALDHhiCD44^{+} human breast cancer cells. *Breast Cancer Res. Treat.* **133,** 75–87 (2012).

52. Doherty, M. R., Smigiel, J. M., Junk, D. J. & Jackson, M. W. Cancer Stem Cell Plasticity Drives Therapeutic Resistance. *Cancers* **8,** (2016).

53. Gong, C. *et al.* Markers of Tumor-Initiating Cells Predict Chemoresistance in Breast Cancer. *PLOS ONE* **5,** e15630 (2010).

54. Charafe-Jauffret, E. *et al.* Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer Res.* **69,** 1302–1313 (2009).

55. Ma, F. *et al.* Enriched CD44(+)/CD24(-) population drives the aggressive phenotypes presented in triple-negative breast cancer (TNBC). *Cancer Lett.* **353,** 153–159 (2014).

56. Li, H. *et al.* Stem cell marker aldehyde dehydrogenase 1 (ALDH1)-expressing cells are enriched in triple-negative breast cancer. *Int. J. Biol. Markers* **28,** e357-364 (2013).

57. Giatromanolaki, A., Sivridis, E., Fiska, A. & Koukourakis, M. I. The CD44+/CD24- phenotype relates to 'triple-negative' state and unfavorable prognosis in breast cancer patients. *Med. Oncol. Northwood Lond. Engl.* **28,** 745–752 (2011).

58. Wu, Y., Sarkissyan, M., Elshimali, Y. & Vadgama, J. V. Triple negative breast tumors in African-American and Hispanic/Latina women are high in CD44+, low in CD24+, and have loss of PTEN. *PloS One* **8,** e78259 (2013).

59. Tsang, J. Y. S. *et al.* Cancer stem cell markers are associated with adverse biomarker profiles and molecular subtypes of breast cancer. *Breast Cancer Res. Treat.* **136,** 407–417 (2012).

60. Perrone, G. *et al.* In Situ Identification of CD44+/CD24− Cancer Cells in Primary Human Breast Carcinomas. *PLoS ONE* **7,** (2012).

61. Ricardo, S. *et al.* Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype. *J. Clin. Pathol.* **64,** 937–946 (2011).

62. Idowu, M. O. *et al.* CD44(+)/CD24(-/low) cancer stem/progenitor cells are more abundant in triple-negative invasive breast carcinoma phenotype and are associated with poor outcome. *Hum. Pathol.* **43,** 364–373 (2012).

63. Komiya, Y. & Habas, R. Wnt signal transduction pathways. *Organogenesis* **4,** 68–75 (2008).

64. Zhao, P., Xia, G., Dong, S., Jiang, Z.-X. & Chen, M. An iTEP-salinomycin nanoparticle that specifically and effectively inhibits metastases of 4T1 orthotopic breast tumors. *Biomaterials* **93,** 1–9 (2016).

65. Jangamreddy, J. R. *et al.* Salinomycin induces activation of autophagy, mitophagy and affects mitochondrial polarity: Differences between primary and cancer cells. *Biochim. Biophys. Acta BBA - Mol. Cell Res.* **1833,** 2057–2069 (2013).

66. Aleman, M., Magdesian, K. G., Peterson, T. S. & Galey, F. D. Salinomycin toxicosis in horses. *J. Am. Vet. Med. Assoc.* **230,** 1822–1826 (2007).

67. Boehmerle, W. & Endres, M. Salinomycin induces calpain and cytochrome c-mediated neuronal cell death. *Cell Death Dis.* **2,** e168 (2011).

68. Van Assen, E. J. A case of salinomycin intoxication in turkeys. *Can. Vet. J.* **47,** 256–258 (2006).

69. Plumlee, K. H., Johnson, B. & Galey, F. D. Acute salinomycin toxicosis of pigs. *J. Vet. Diagn. Investig. Off. Publ. Am. Assoc. Vet. Lab. Diagn. Inc* **7,** 419–420 (1995).

70. Bhavanasi, D. & Klein, P. S. Wnt Signaling in Normal and Malignant Stem Cells. *Curr. Stem Cell Rep.* **2,** 379–387 (2016).

71. Jin, J., Krishnamachary, B., Mironchik, Y., Kobayashi, H. & Bhujwalla, Z. M. Phototheranostics of CD44-positive cell populations in triple negative breast cancer. *Sci. Rep.* **6,** 27871 (2016).

72. Iida, J. *et al.* DNA Aptamers against Exon v10 of CD44 Inhibit Breast Cancer Cell Migration. *PLOS ONE* **9,** e88712 (2014).

73. Muntimadugu, E., Kumar, R., Saladi, S., Rafeeqi, T. A. & Khan, W. CD44 targeted chemotherapy for co-eradication of breast cancer stem cells and cancer cells using polymeric nanoparticles of salinomycin and paclitaxel. *Colloids Surf. B Biointerfaces* **143,** 532–546 (2016).

74. Marchitti, S. A., Brocker, C., Stagos, D. & Vasiliou, V. Non-P450 aldehyde oxidizing enzymes: the aldehyde dehydrogenase superfamily. *Expert Opin. Drug Metab. Toxicol.* **4,** 697–720 (2008).

75. Thomas, M. L. *et al.* Citral reduces breast tumor growth by inhibiting the cancer stem cell marker ALDH1A3. *Mol. Oncol.* **10,** 1485–1496 (2016).

76. Marcato, P. *et al.* Aldehyde dehydrogenase 1A3 influences breast cancer progression via differential retinoic acid signaling. *Mol. Oncol.* **9,** 17–31 (2015).

77. Kim, R.-J. *et al.* High aldehyde dehydrogenase activity enhances stem cell features in breast cancer cells by activating hypoxia-inducible factor-2α. *Cancer Lett.* **333,** 18–31 (2013).

78. Morgan, C. A., Parajuli, B., Buchman, C. D., Dria, K. & Hurley, T. D. N,N-diethylaminobenzaldehyde (DEAB) as a substrate and mechanism-based inhibitor for human ALDH isoenzymes. *Chem. Biol. Interact.* **234,** 18–28 (2015).

79. Liu, P. *et al.* Disulfiram targets cancer stem-like cells and reverses resistance and cross-resistance in acquired paclitaxel-resistant triple-negative breast cancer cells. *Br. J. Cancer* **109,** 1876–1885 (2013).

80. Coyle, K. M., Maxwell, S., Thomas, M. L. & Marcato, P. Profiling of the transcriptional response to all-trans retinoic acid in breast cancer cells reveals RARE-independent mechanisms of gene expression. *Sci. Rep.* **7,** 16684 (2017).

81. Coyle, K. M. *et al.* Breast cancer subtype dictates DNA methylation and ALDH1A3-mediated expression of tumor suppressor RARRES1. *Oncotarget* **7,** 44096–44112 (2016).

82. Zhao, D. *et al.* Molecular identification of a major retinoic-acid-synthesizing enzyme, a retinaldehyde-specific dehydrogenase. *Eur. J. Biochem.* **240,** 15–22 (1996).

83. Coyle, K. M., Sultan, M., Thomas, M., Vagar-Kashani, A. & Marcato, P. Retinoid Signaling in Cancer and Its Promise for Therapy. *J. Carcinog. Mutagen.* **0,** 1–15 (2013).

84. Duan, J.-J., Cai, J., Guo, Y.-F., Bian, X.-W. & Yu, S.-C. ALDH1A3, a metabolic target for cancer diagnosis and therapy. *Int. J. Cancer* **139,** 965–975 (2016).

85. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193,** 651–669 (2013).

86. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).

87. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543,** 199–204 (2017).

88. Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* **41,** D246–D251 (2013).

89. Quek, X. C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **43,** D168-173 (2015).

90. Batista, P. J. & Chang, H. Y. Long noncoding RNAs: Cellular address codes in development and disease. *Cell* **152,** 1298–1307 (2013).

91. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** (2012).

92. Rashid, F., Shah, A. & Shan, G. Long Non-coding RNAs in the Cytoplasm. *Genomics Proteomics Bioinformatics* **14,** 73–80 (2016).

93. Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71,** 515–526 (1992).

94. Sheardown, S. A. *et al.* Stabilization of Xist RNA mediates initiation of X chromosome inactivation. *Cell* **91,** 99–107 (1997).

95. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161,** 404–416 (2015).

96. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).

97. Yan, X. *et al.* Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell* **28,** 529–540 (2015).

98. Chiu, H.-S. *et al.* Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Rep.* **23,** 297-312.e12 (2018).

99. Lanzós, A. *et al.* Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Sci. Rep.* **7,** (2017).

100. Gupta, R. A. *et al.* Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).

101. Leucci, E. *et al.* Melanoma addiction to the long non-coding RNA *SAMMSON*. *Nature* **531,** 518–522 (2016).

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

103. Zhang, Y. *et al.* Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat. Struct. Mol. Biol.* **23,** 522–530 (2016).

104. Deng, J. *et al.* Long Non-Coding RNA HOTAIR Regulates the Proliferation, Self-Renewal Capacity, Tumor Formation and Migration of the Cancer Stem-Like Cell (CSC) Subpopulation Enriched from Breast Cancer Cells. *PLOS ONE* **12,** e0170860 (2017).

105. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42,** 1113–1117 (2010).

106. Hou, P. *et al.* LincRNA-ROR induces epithelial-to-mesenchymal transition and contributes to breast cancer tumorigenesis and metastasis. *Cell Death Dis.* **5,** e1287 (2014).

107. Lin, N. *et al.* An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol. Cell* **53,** 1005–1019 (2014).

108. Li, H. *et al.* Long noncoding RNA linc00617 exhibits oncogenic activity in breast cancer. *Mol. Carcinog.* **56,** 3–17.

109. Zhou, M. *et al.* LncRNA-Hh Strengthen Cancer Stem Cells Generation in Twist-Positive Breast Cancer via Activation of Hedgehog Signaling Pathway. *Stem Cells Dayt. Ohio* **34,** 55–66 (2016).

110. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* **10,** 28–36 (1990).

111. Shima, H. *et al.* Lnc RNA H19 is associated with poor prognosis in breast cancer patients and promotes cancer stemness. *Breast Cancer Res. Treat.* (2018). doi:10.1007/s10549-018-4793-z

112. Peng, F. *et al.* H19/let-7/LIN28 reciprocal negative regulatory circuit promotes breast cancer stem cell maintenance. *Cell Death Dis.* **8,** e2569 (2017).

113. Lennox, K. A. & Behlke, M. A. Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. *Nucleic Acids Res.* **44,** 863–877 (2016).

114. Sakamoto, N. *et al.* Non-coding RNAs are promising targets for stem cell-based cancer therapy. *Non-Coding RNA Res.* **2,** 83–87 (2017).

115. Barata, P., Sood, A. K. & Hong, D. S. RNA-targeted therapeutics in cancer clinical trials: Current status and future directions. *Cancer Treat. Rev.* **50,** 35–47 (2016).

116. Zamecnik, P. C. & Stephenson, M. L. Inhibition of Rous sarcoma virus replication and cell transformation by a specific oligodeoxynucleotide. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 280–284 (1978).

117. Eder, P. S., DeVine, R. J., Dagle, J. M. & Walder, J. A. Substrate specificity and kinetics of degradation of antisense oligonucleotides by a 3' exonuclease in plasma. *Antisense Res. Dev.* **1,** 141–151 (1991).

118. Braasch, D. A. & Corey, D. R. Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. *Chem. Biol.* **8,** 1–7 (2001).

119. Gutschner, T. *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* **73,** 1180–1189 (2013).

120. Xing, Z. *et al.* lncRNA directs cooperative epigenetic regulation downstream of chemokine signals. *Cell* **159,** 1110–1125 (2014).

121. Chu, C., Quinn, J. & Chang, H. Y. Chromatin isolation by RNA purification (ChIRP). *J. Vis. Exp. JoVE* (2012). doi:10.3791/3912.

122. Marín-Béjar, O. & Huarte, M. RNA pulldown protocol for in vitro detection and identification of RNA-associated proteins. *Methods Mol. Biol. Clifton NJ* **1206,** 87–95 (2015).

123. Győrffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* **123,** 725–731 (2010).

124. Li, J. *et al.* TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer. *Cancer Res.* **75,** 3728–3737 (2015).

125. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2,** 401–404 (2012).

126. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).

127. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).

128. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–273 (2003).

129. Park, S. Y. *et al.* Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **16,** 876 (2010).

130. Lánczky, A. *et al.* miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res. Treat.* **160,** 439–446 (2016).

131. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

132. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–1184 (2013).

133. Ginestier, C. *et al.* Retinoid signaling regulates breast cancer stem cell differentiation. *Cell Cycle Georget. Tex* **8,** 3297–3302 (2009).

134. Ito, K. & Suda, T. Metabolic requirements for the maintenance of self-renewing stem cells. *Nat. Rev. Mol. Cell Biol.* **15,** 243–256 (2014).

135. Peiris-Pagès, M., Martinez-Outschoorn, U. E., Pestell, R. G., Sotgia, F. & Lisanti, M. P. Cancer stem cell metabolism. *Breast Cancer Res. BCR* **18,** (2016).

136. De Francesco, E. M., Maggiolini, M., Tanowitz, H. B., Sotgia, F. & Lisanti, M. P. Targeting hypoxic cancer stem cells (CSCs) with Doxycycline: Implications for optimizing anti-angiogenic therapy. *Oncotarget* **8,** 56126–56142 (2017).

137. Li, D. *et al.* Inhibition of glutamine metabolism counteracts pancreatic cancer stem cell features and sensitizes cells to radiotherapy. *Oncotarget* **6,** 31151–31163 (2015).

138. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front. Genet.* **6,** (2015).

139. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 5294–5300 (2013).

140. Tripathi, V. *et al.* The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol. Cell* **39,** 925–938 (2010).

141. Zhang, X. *et al.* A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.* **73,** 4885–4897 (2013).

142. Yu, L. *et al.* SND1 Acts Downstream of TGFβ1 and Upstream of Smurf1 to Promote Breast Cancer Metastasis. *Cancer Res.* **75,** 1275–1286 (2015).

143. Yang, J. *et al.* Identification of p100 as a coactivator for STAT6 that bridges STAT6 with RNA polymerase II. *EMBO J.* **21,** 4950–4958 (2002).

144. Yang, W. *et al.* Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat. Struct. Mol. Biol.* **13,** 13–21 (2006).

145. Sundström, J. F. *et al.* Tudor staphylococcal nuclease is an evolutionarily conserved component of the programmed cell death degradome. *Nat. Cell Biol.* **11,** 1347–1354 (2009).

146. Yu, L. *et al.* SND1 acts as a novel gene transcription activator recognizing the conserved Motif domains of Smad promoters, inducing TGFβ1 response and breast cancer metastasis. *Oncogene* **36,** 3903–3914 (2017).

147. Wan, L. *et al.* MTDH-SND1 Interaction Is Crucial for Expansion and Activity of Tumor-Initiating Cells in Diverse Oncogene- and Carcinogen-Induced Mammary Tumors. *Cancer Cell* **26,** 92–105 (2014).

148. Blanco, M. A. *et al.* Identification of staphylococcal nuclease domain-containing 1 (SND1) as a Metadherin-interacting protein with metastasis-promoting functions. *J. Biol. Chem.* **286,** 19982–19992 (2011).

149. Arretxe, E. *et al.* Profiling of promoter occupancy by the SND1 transcriptional coactivator identifies downstream glycerolipid metabolic genes involved in TNFα response in human hepatoma cells. *Nucleic Acids Res.* **43,** 10673–10688 (2015).

150. Čuperlović-Culf, M., Barnett, D. A., Culf, A. S. & Chute, I. Cell culture metabolomics: applications and future directions. *Drug Discov. Today* **15,** 610–621 (2010).

151. Sun, G. *et al.* PRKDC regulates chemosensitivity and is a potential prognostic and predictive marker of response to adjuvant chemotherapy in breast cancer patients. *Oncol. Rep.* **37,** 3536–3542 (2017).