LINC00520 is induced by Src, STAT3, and PI3K and plays a functional role in breast cancer

Whitney S. Henry[1,*], David G. Hendrickson[2,3,*], Francisco Beca[1], Benjamin Glass[1], Marianne Lindahl-Allen[4], Lizhi He[4], Zhe Ji[4], Kevin Struhl[4], Andrew H. Beck[1], John L. Rinn[2,3] and Alex Toker[1]

[1] Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

[2] Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA

[3] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA

[*] These authors have contributed equally to this work

Correspondence to: *Alex Toker*, *email*: atoker@bidmc.harvard.edu

John L. Rinn, *email*: john_rinn@harvard.edu

ABSTRACT

Long non-coding RNAs (lncRNAs) have been implicated in normal cellular homeostasis as well as pathophysiological conditions, including cancer. Here we performed global gene expression profiling of mammary epithelial cells transformed by oncogenic *v-Src*, and identified a large subset of uncharacterized lncRNAs potentially involved in breast cancer development. Specifically, our analysis revealed a novel lncRNA, LINC00520 that is upregulated upon ectopic expression of oncogenic *v-Src*, in a manner that is dependent on the transcription factor STAT3. Similarly, LINC00520 is also increased in mammary epithelial cells transformed by oncogenic PI3K and its expression is decreased upon knockdown of mutant *PIK3CA.* Additional expression profiling highlight that LINC00520 is elevated in a subset of human breast carcinomas, with preferential enrichment in the basal-like molecular subtype. ShRNA-mediated depletion of LINC00520 results in decreased cell migration and loss of invasive structures in 3D. RNA sequencing analysis uncovers several genes that are differentially expressed upon ectopic expression of LINC00520, a significant subset of which are also induced in *v-Src*-transformed MCF10A cells. Together, these findings characterize LINC00520 as a lncRNA that is regulated by oncogenic Src, *PIK3CA* and STAT3, and which may contribute to the molecular etiology of breast cancer.

INTRODUCTION

Cancer is largely driven by genetic alterations, which lead to the deregulation of gene networks that typically govern normal cellular homeostasis. Recent studies have implicated long non-coding RNAs (lncRNAs) in a diverse array of human cancers [1-4]. Surprisingly, a large number of these non-coding transcripts are found in genomic regions that experience frequent mutation or somatic copy number alterations [5]. In addition, many lncRNAs are transcriptionally regulated by major oncogenes and tumor suppressors including *c-Myc* and *p53* respectively [6, 7].

Gene expression profiling of various disease model systems has proven to be a powerful approach for identifying candidate lncRNAs implicated in cancer. The first cancer-associated lncRNAs to be identified using differential expression profiling of prostate tumors and normal tissue, were prostate cancer associated 3 (*PCA3*, also called *DD3*) which is currently used as a biomarker for prostate cancer [8, 9], and prostate-specific transcript 1 (*PPCGEM1*) which is implicated in androgen receptor transcriptional activation [10, 11]. Differential expression profiling has also led to the discovery of the nuclear lncRNA metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*), as one of the first lncRNAs to be

ascribed a role as a potential prognostic biomarker for lung cancer survival [12].

Collectively, lncRNAs may act in either a tumor suppressive or oncogenic capacity to modulate cellular phenotypes associated with malignancy [13]. One of the best-characterized cancer-associated lncRNAs is *HOTAIR*. This lncRNA acts as a molecular scaffold for Polycomb Repressive Complex 2 (PRC2) and Lysine-Specific Demethylase 1 (LSD1) to facilitate epigenetic silencing of specific gene loci and promotes breast cancer metastasis [1]. Furthermore, expression of *HOTAIR* is also associated with poor survival [1]. *ANRIL* is another lncRNA implicated in cancer. Expression of this antisense non-coding RNA in prostate cancer cells, results in the transcriptional repression of the *INK4n/ARF/INK4a* tumor suppressor genes, which regulate cell cycle progression and senescence [14]. Similarly, in melanoma cells, RNAi-mediated knockdown of the highly expressed lncRNA SPRY4-IT1 results in defects in cell growth and induction of apoptosis [15].

In spite of these examples, less than 1% of the identified human lncRNAs have been characterized [16]. Our understanding of lncRNA biology is far from complete and the identification, regulation and functional characterization of lncRNAs involved in breast cancer pathogenesis may provide novel opportunities for differential diagnoses and therapeutic interventions. Here we identify the novel lncRNA LINC00520 in breast cancer using two independent systems of cellular transformation driven by oncogenic *v-Src* and mutant *PIK3CA*, respectively. We further demonstrate that LINC00520 expression is clinically relevant and is preferentially associated with basal-like breast cancer. We also investigate the transcriptional regulation of LINC00520 and provide evidence for its role in breast cancer development.

RESULTS

Identification and transcriptional regulation of LINC00520 in a model of Src-induced transformation of mammary epithelial cells

In order to assemble a comprehensive list of lncRNAs that are potentially implicated in breast cancer, we systematically surveyed the transcriptome of a well-characterized immortalized mammary epithelial MCF10A cell line model containing a tamoxifen-inducible Src oncoprotein (*v-Src*). Previous studies using this model system have demonstrated that ectopic expression of *Src* results in multiple features associated with cellular transformation, including colony formation in soft agar, increased migration and invasion and tumor formation capability in immunocompromised mice

[17]. Furthermore, Src-induced transformation has been demonstrated to drive an onset of molecular events that involve epigenetic alterations leading to changes in gene expression networks [17].

To explore the transcriptome of MCF10A cells upon Src induction, we collected RNA before (T0) and after Src induction at 4, 12, and 36 hours (T4,T12,T36) and performed RNA-sequencing. Differential expression analysis revealed thousands of protein coding genes and hundreds of differentially regulated non-coding transcripts (Figure 1a). As expected, we observed concordant overlap with the transcriptional signature previously defined in this system [17]. To identify lncRNAs with oncogenic potential we focused on a subset of the ncRNAs whose transcript levels are robustly increased upon *v-Src* induction (Figure 1a).

To pare down the number of potential candidates, we ordered the transformation-induced lncRNAs by fold induction as well as final transcript abundance at 36 hours. We reasoned that a potent oncogenic lncRNA would display both strong induction AND high expression. Topping both criteria was LINC00520, an uncharacterized lncRNA that displayed both striking induction (> 30 fold) and abundance of ~ 80 FPKM at 36 hours (Figure 1b). Consequently, LINC00520 ranked in the ~95 percentile of expressed genes which is at the high end of both reported lncRNA and coding expression regimes. Subsequent analyses on LINC00520 indicates that it resides ~112kb from the kinesin receptor *KTN1* and ~ 321kb from the Pellino E3 ubiquitin ligase family member 2, *PELI2* (Figure 1b). In support of LINC00520 being an independent transcript, we note that LINC00520 is transcribed in the opposite direction to either flanking gene. In addition, transcript structural analysis reveals that LINC00520 undergoes splicing and contains 3-4 exons depending on the isoform type (Figure 1b).

LINC00520 is regulated by STAT3 in Src-transformed cells

Since the transcription factor signal transducer and activator of transcription 3 (STAT3) plays a critical role in Src-induced transcriptional responses during cellular transformation [17], we analyzed published chromatin immunoprecipitation (ChIP) data performed in the MCF10A Src-induced cells to determine whether STAT3 directly binds to the LINC00520 promoter [18]. An enrichment of STAT3 binding to the LINC00520 promoter region is observed as early as 4 hours post Src induction, with a significant increase at 36 hours. This coincides with an increase in LINC00520 transcript levels at this time point (Figure 1c). Moreover, depletion of STAT3 with siRNA abolishes Src-induced upregulation of LINC00520 (Figure 1d). Taken together, these data implicate STAT3 in the transcriptional regulation of LINC00520 during
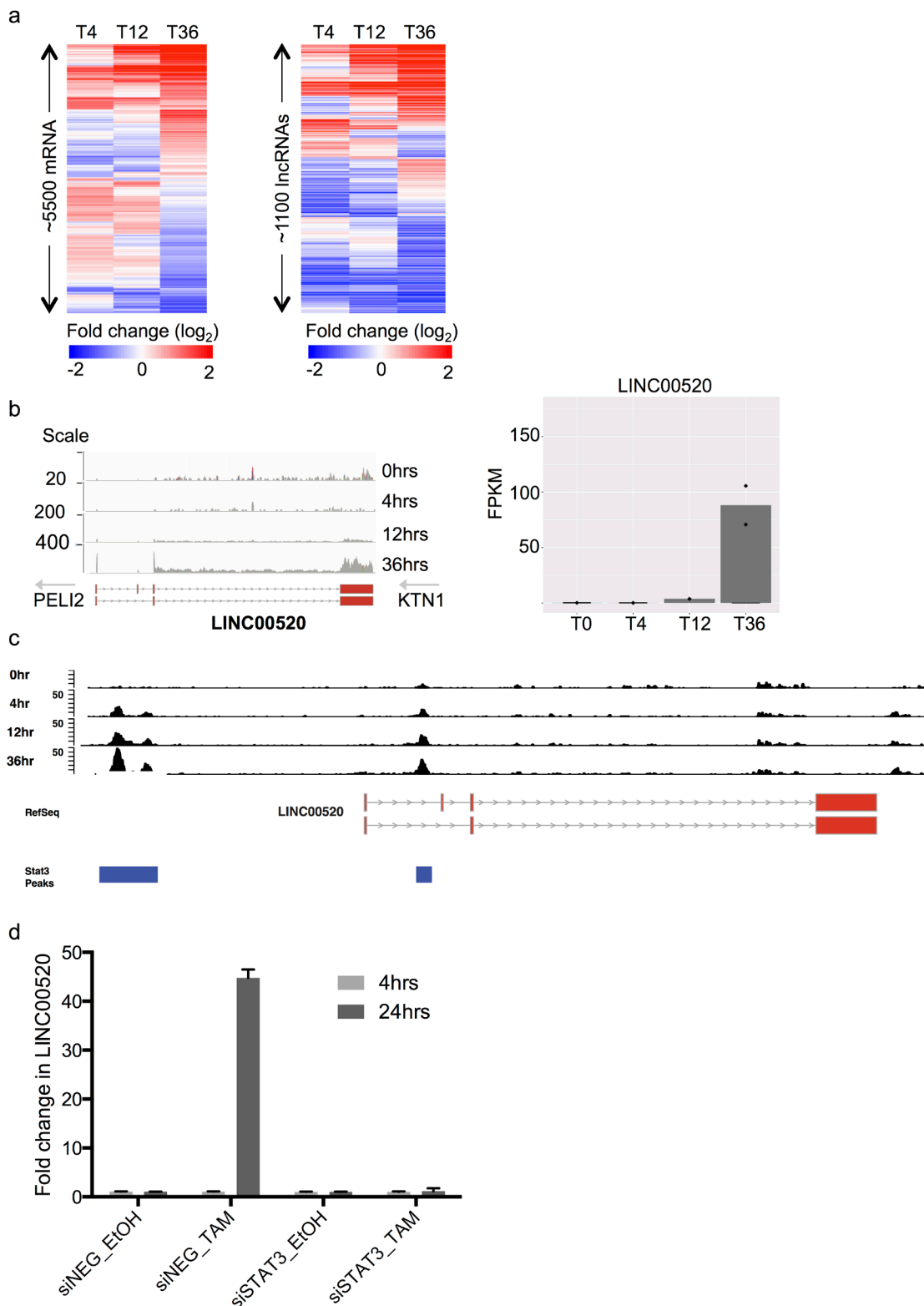
Figure 1: Identification and transcriptional regulation of LINC00520 in a model of Src-induced transformation of mammary epithelial cells. a. Heat maps showing subset of protein coding genes and long non-coding RNAs that are differentially expressed at 4,12 and 36 hours post Src induction, in MCF10A cells. **b.** RNA Sequencing, relative expression of LINC00520 at various time-points post Src induction in immortalized mammary epithelial MCF10A cells. FPKM, fragments per kilobase of transcript per million mapped reads. **c.** STAT3 ChIP enrichment in MCF10A cells post Src induction, at the LINC00520 locus. **d.** Expression of LINC00520 following siRNA-mediated depletion of STAT3 in MCF10A-Src transformed cells. Transcript levels were determined by qRT-PCR and normalized to GAPDH. Values represent the average of three technical triplicates.

cellular transformation of mammary epithelial cells driven by oncogenic Src.

LINC00520 is regulated by the PI3K pathway

To investigate if LINC00520 plays a broader role in transformation we turned to an orthogonal model using MCF10A mammary epithelial cells expressing oncogenic mutants of *PIK3CA*. The PI3K pathway is frequently hyperactivated in breast cancer mainly due to recurrent somatic mutations in *PIK3CA*, the gene that encodes the p110α catalytic subunit of PI3K, or *via* loss of the tumor suppressor Phosphatase and Tensin Homolog (*PTEN)* [19]. As previously reported, oncogenic *PIK3CA* (*H1047R)* induces cellular transformation as indicated by the increase in colony formation in soft agar compared to wild-type *PIK3CA* (Supplementary Figure 1) [20]. We collected gene expression data from *PIK3CA* (*H1047R)* MCF10A cells and compared this to WT p110α using microarrays (Figure 2a). We found that many differentially expressed genes in the *PIK3CA* (*H1047R)*-expressing cells are also differentially expressed in Src-transformed MCF10A cells, including LINC00520 (Figure 2b). We verified LINC00520 expression using qRT-PCR and again found that increased LINC00520 transcript levels are observed in MCF10A cells expressing mutant *PIK3CA* (*H1047R)* relative to wild-type *PIK3CA*-expressing cells (Figure 2c). Likewise, a similar upregulation of LINC00520 is observed in isogenic MDA-MB-231 cells expressing *PIK3CA* (*H1047R)* relative to wild-type (Figure 2d). By contrast, LINC00520 is downregulated upon depletion of *PIK3CA* in SUM159-PT (mutant *PIK3CA H1047L*) breast cancer cells (Figure 2e). Together these data suggest that LINC00520 is regulated downstream of oncogenic PI3K.

LINC00520 is upregulated in human tumors and enriched in basal-like human breast carcinomas

It has been reported that lncRNAs tend to display tissue specificity. Therefore, we next examined the expression of LINC00520 in a panel of breast cancer cell lines with defined genetic alterations and molecular subtypes. Interestingly, we find a preferential upregulation of LINC00520 in basal-like, triple-negative breast cancer cell lines (Figure 3a), many of which display a high metastatic potential and poor prognosis.

To better understand this finding, we analyzed LINC00520 expression in breast carcinoma clinical samples. Using RNA-sequencing data, we found that LINC00520 is upregulated in human breast tumors (Figure 3b) and it is particularly enriched in basal-like and *HER2* intrinsic molecular subtypes (Figure 3c). Furthermore, within the basal-like subtype, we found 166 differentially expressed genes (of 13, 015 transcripts) in samples with high LINC00520 expression (Supplementary Figure

2). Gene set enrichment analysis (GSEA) (based on a pre-ranked gene list using SAM's d-statistic) revealed that the expression of LINC00520 in basal-like breast cancer was significantly associated with upregulation of 26 of the Broad Institute´s Molecular Signatures Database (mSigDB) Hallmark gene sets at a FDR < 5% (Supplementary Figure 3), including gene sets related to activation of the PI3K/AKT/mTOR pathway and epithelial-mesenchymal-transition (EMT) (Figure 3d). Consistent with this finding, functional analysis, indicates that many of the differentially expressed genes in basal-like carcinomas of the breast with high LINC00520 expression were primarily implicated in processes involved in cell adhesion, extracellular matrix remodeling, and wound healing (Figure 3e) and may therefore provide insight into potential roles of LINC00520 in breast cancer.

Depletion of LINC00520 blocks breast cancer cell migration

In order to assess the biological role of LINC00520 in breast cancer, we performed shRNA-mediated loss-of-function studies in basal-like breast cancer cell lines containing endogenous levels of LINC00520. Two distinct shRNA constructs, each targeting a distinct exon of LINC00520, were cloned (Figure 4a). Approximately 70-90% silencing efficiency was observed by qRT-PCR in all cell lines tested including SUM159-PT cells (Figure 4b), MCF10A-Src cells (Figure 4f) and MCF10DCIS cells (Figure 5a). We then investigated the effects of LINC00520 depletion on cell proliferation. We observed no significant effects on the proliferation of SUM159-PT cells upon depletion of LINC00520 (Figure 4c). By contrast, depletion of LINC00520 leads to a decrease in transwell migration of SUM159-PT cells (Figure 4d) and MCF10A-Src transformed cells (Figure 4e). Similarly, a reduction of invasive protrusions from MCF10DCIS spheroids grown in Matrigel is observed upon depletion of LINC00520 (Figure 5b-5c). These data implicate LINC00520 in cell migration and invasion.

Overexpression of LINC00520 results in alterations in global gene expression

Based on the above loss-of-function studies, we next determined whether ectopic expression of LINC00520 in parental WT MCF10A cells leads to gene expression alterations in cancer-associated genes. MCF10A cells were transduced with lentivirus made from a modified lentiviral construct designed specifically for the over-expression of non-coding RNAs [21]. This construct was used for the ectopic for expression of either LINC00520 or a control lncRNA (AC006262.1) that is not expressed in either parental MCF10A-Src, or MCF10A *PIK3CA* (*H1047R*) cells. We compared the expression profiles of
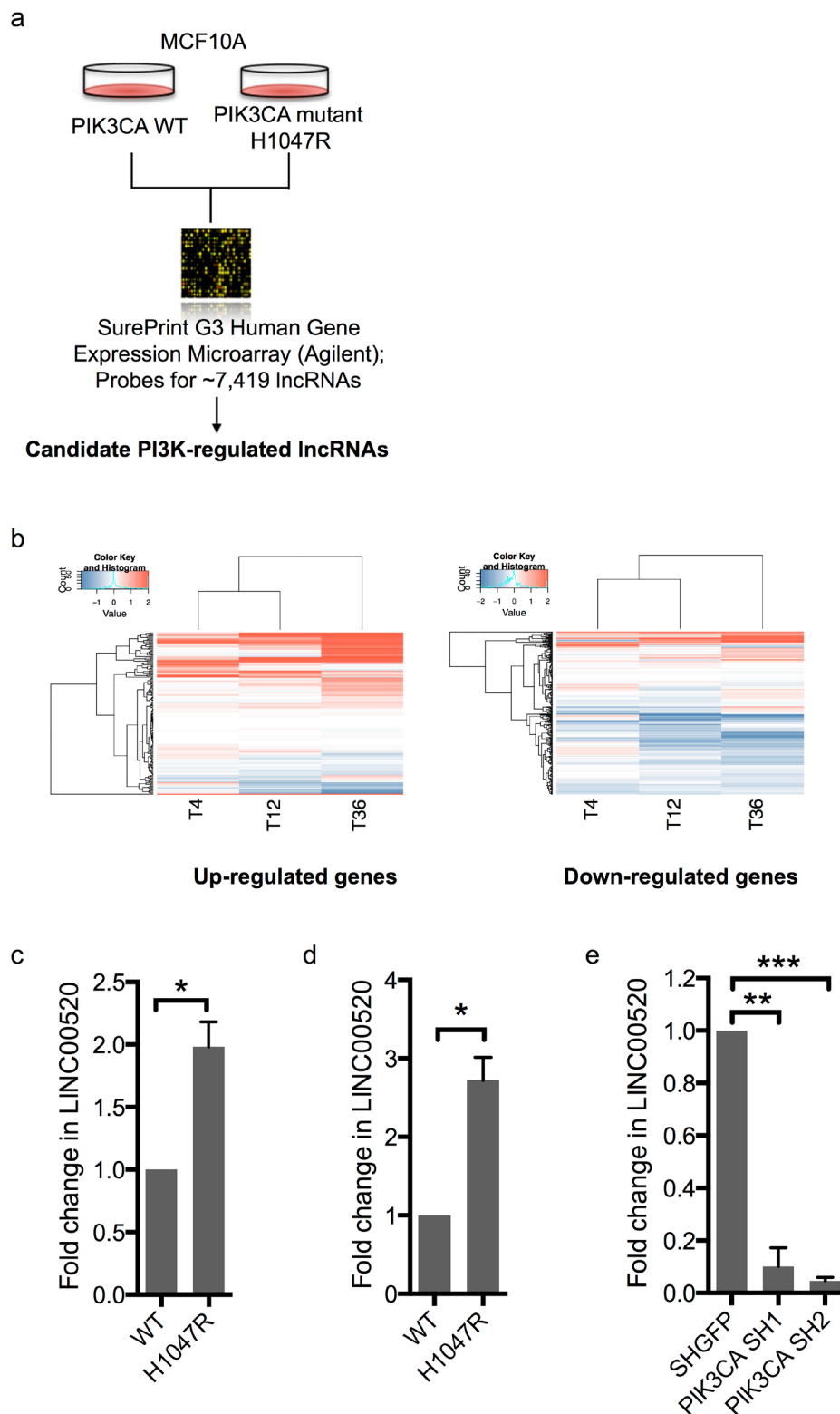
Figure 2: LINC00520 is regulated by the PI3K pathway. a. Schematic of MCF10A system used to identify PI3K regulated genes involved in transformation. **b.** Heat map of gene expression values from Src-transformed MCF10A cells for genes that increased (left) or decreased (right) by two fold in the mutant *PIK3CA*-transformed MCF10A cells. Quantitative RT-PCR, showing relative expression of LINC00520 in: **c.** MCF10A cells expressing wild-type *PIK3CA* (WT) compared to mutant *PIK3CA H1047R*, **d.** Isogenic MDA-MB-231 cells expressing wild-type *PIK3CA* (WT), and mutant *PIK3CA H1047R* respectively and **e.** shRNA-mediated knockdown of *PIK3CA* in SUM159-PT cells. Transcript levels were normalized to GAPDH. Data represents average of three independent experiments ± the SEM. Statistical significance was determined using paired Student's *t*-test, * $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$.
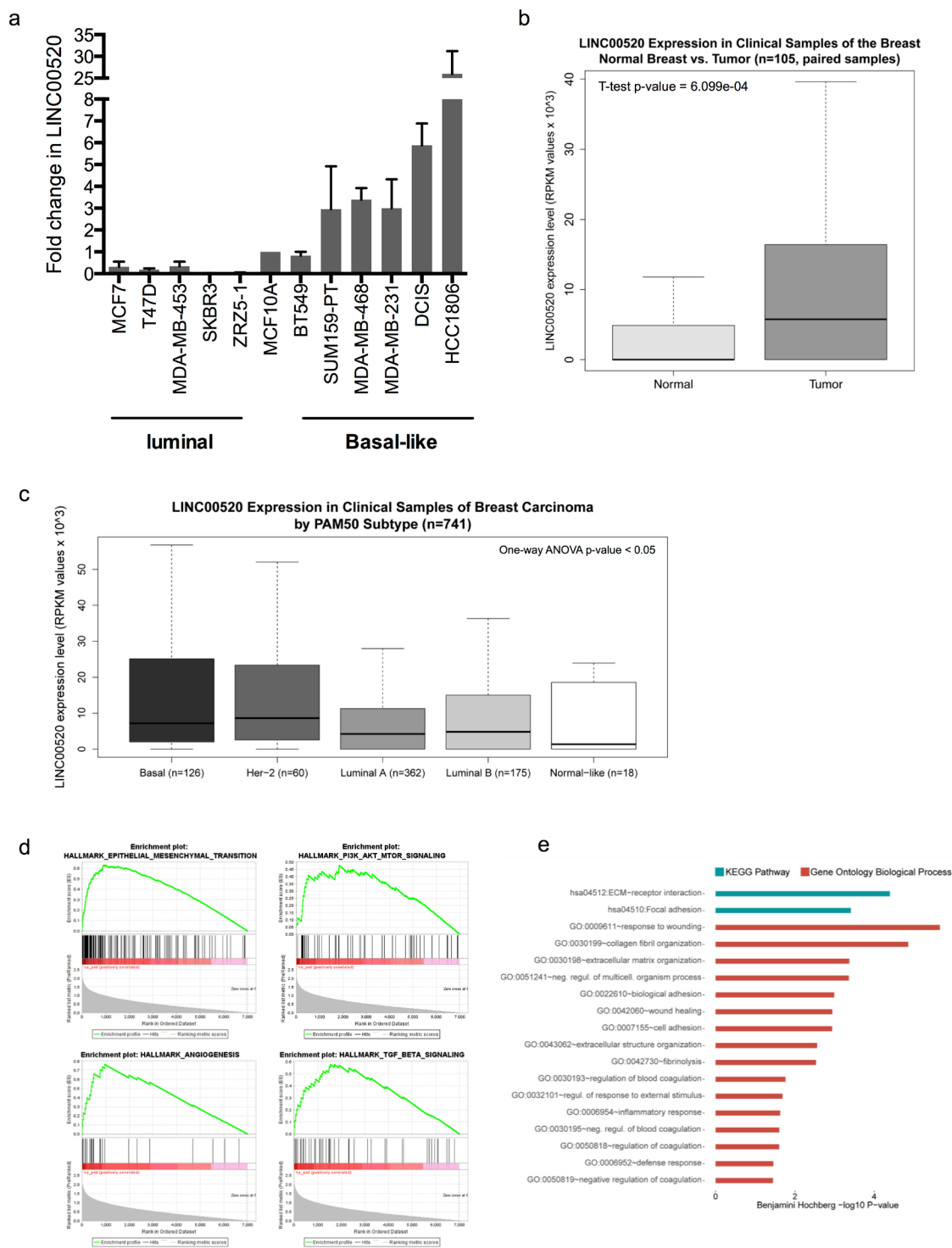
Figure 3: Expression Analysis in Clinical Samples of Human Breast Invasive Carcinoma. a. Relative expression of LINC00520 in a panel of breast cancer cell lines, compared to MCF10A cells. Expression was determined by qRT-PCR and normalized to GAPDH. Values are representative from three independent experiments. **b.** Breast carcinoma clinical samples show higher expression of LINC00520 compared to matched normal breast samples from TCGA dataset (n = 105). **c.** In breast carcinoma clinical samples (n = 741) LINC00520 expression level differs according to the intrinsic molecular subtype. RPKM, reads per kilobase of transcript per million mapped reads. **d.** Gene Set Enrichment analysis (GSEA) of basal carcinomas with high expression of LINC00520. Four of the most associated gene sets, of the 50 Hallmark gene sets of the Broad Institute's MSigDB collection, 11 gene set S were significantly upregulated at an FDR < 1% and 26 at an FDR < 5% using Hallmark gene set collection. **e.** Top-ranked Functional Gene Ontology Biological Process and KEGG Pathway terms using DAVID. All represented terms were associated with the transcriptomic signature of basal breast carcinomas with high LINC00520 expression (Benjamini Hochberg p-value of < 0.05).

Figure 4: Depletion of LINC00520 blocks breast cancer cell migration. a. Schematic showing regions targeted by LINC00520 shRNA#1 and shRNA#2. **b.** Validation of downregulation of LINC00520 in SUM159-PT cells used for proliferation and migration assays. **c.** Proliferation assay using SUM159-PT cells expressing two different LINC00520 shRNA constructs. Data represents average of two biological replicates. **d.** Transwell migration assay of SUM159-PT cells expressing two different LINC00520 shRNA constructs. **e.** Transwell migration assay of MCF10A-Src inducible cells expressing two different LINC00520 shRNA constructs. Cells were induced with vehicle (ethanol) or Tamoxifen for 48hours prior to migration. **f.** Validation of downregulation of LINC00520 in MCF10A-Src inducible cells used for migration assays. Data represents average of three biological replicates ± the SEM. Statistical significance was determined using unpaired Student's t test, * $p < 0.05$, *** $p < 0.001$.

untransduced to transduced cells 72 hours post infection using RNA-seq. Using Cuffdiff2 for differential gene expression analysis we first established that LINC00520 was induced upon transduction (~500 fold) and then found that 1,898 genes changed significantly in response to transduction with LINC00520. To determine if LINC00520 might have a regulatory role in transformation we next asked if the upregulated genes during MCF10A-Src transformation were enriched in the genes induced by LINC00520 transduction (976 out of the 1,898). We tested

a

b

c
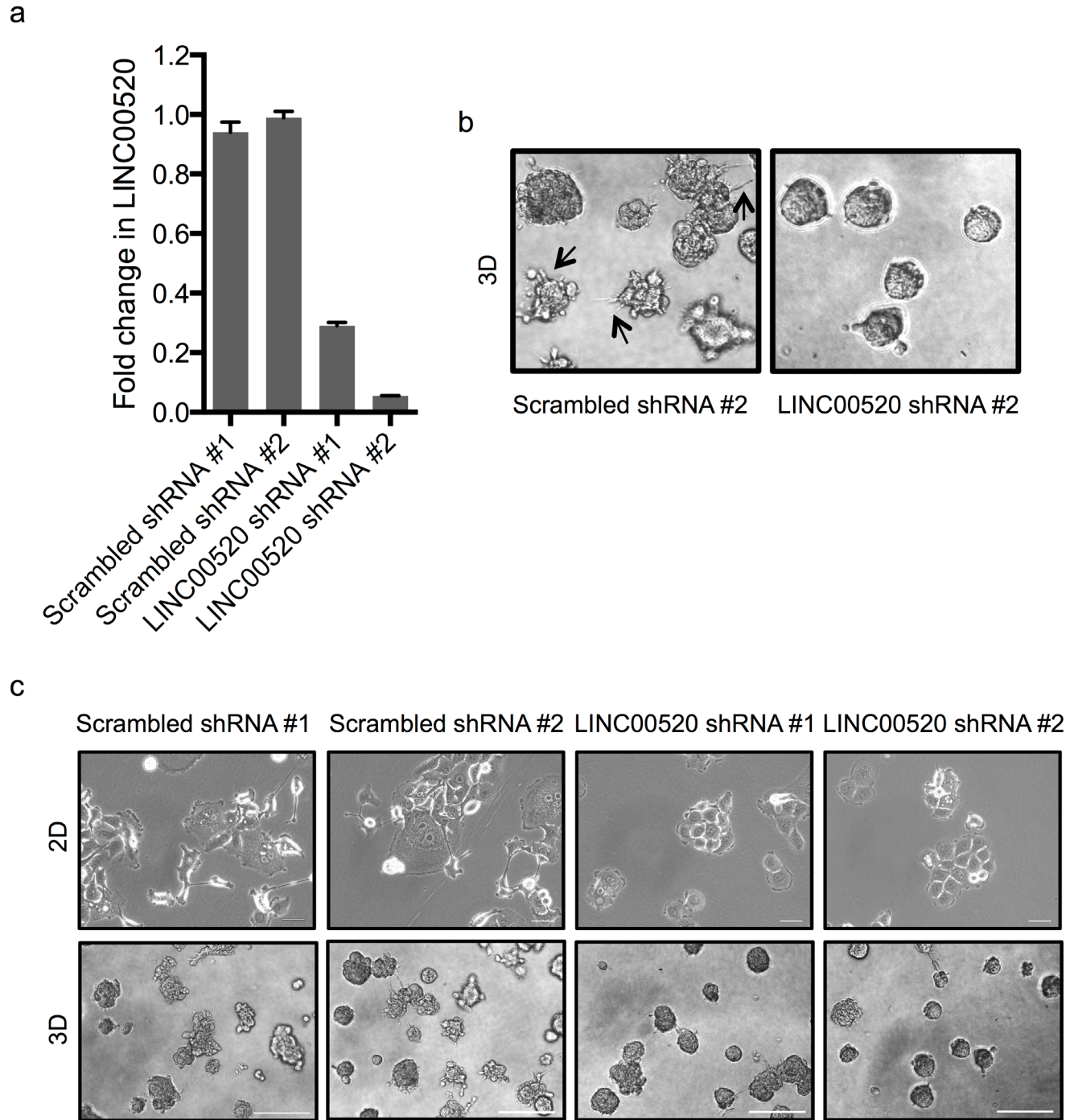
Figure 5: Depletion of LINC00520 affects the integrity of invasive protrusions. a. Quantitative RT-PCR validation of silencing of LINC00520 in MCF10DCIS cells. Transcript levels were normalized to GAPDH and fold change was calculated relative to scrambled shRNA #1. **b.** Representative phase-contrast images of MCF10DCIS acini infected with lentiviral scrambled or LINC00520 shRNA grown in 3D Matrigel. Images displayed were taken on day 5. **c.** Morphological effects of silencing of LINC00520 in MCF10DCIS cells grown in (Top) 2D, scale bar = 100μm and (Bottom) in 3D Matrigel cultures. For 3D, images shown were taken at day 5, scale bar = 500μm.

for the enrichment of upregulated genes from each time point using the hypergeometric test (4,12, and 36 hours) and found significant overlap with LINC00520 induction signature for all time points (at 4 hours $p < 1e\text{-}20$, at 12 hours $p < 8e\text{-}38$ and at 36 hours $p < 0.008$). The most significant overlap was found not at the time point with the highest LINC00520 expression (36 hours) but at 12 hours where LINC00520 expression begins to ramp up in
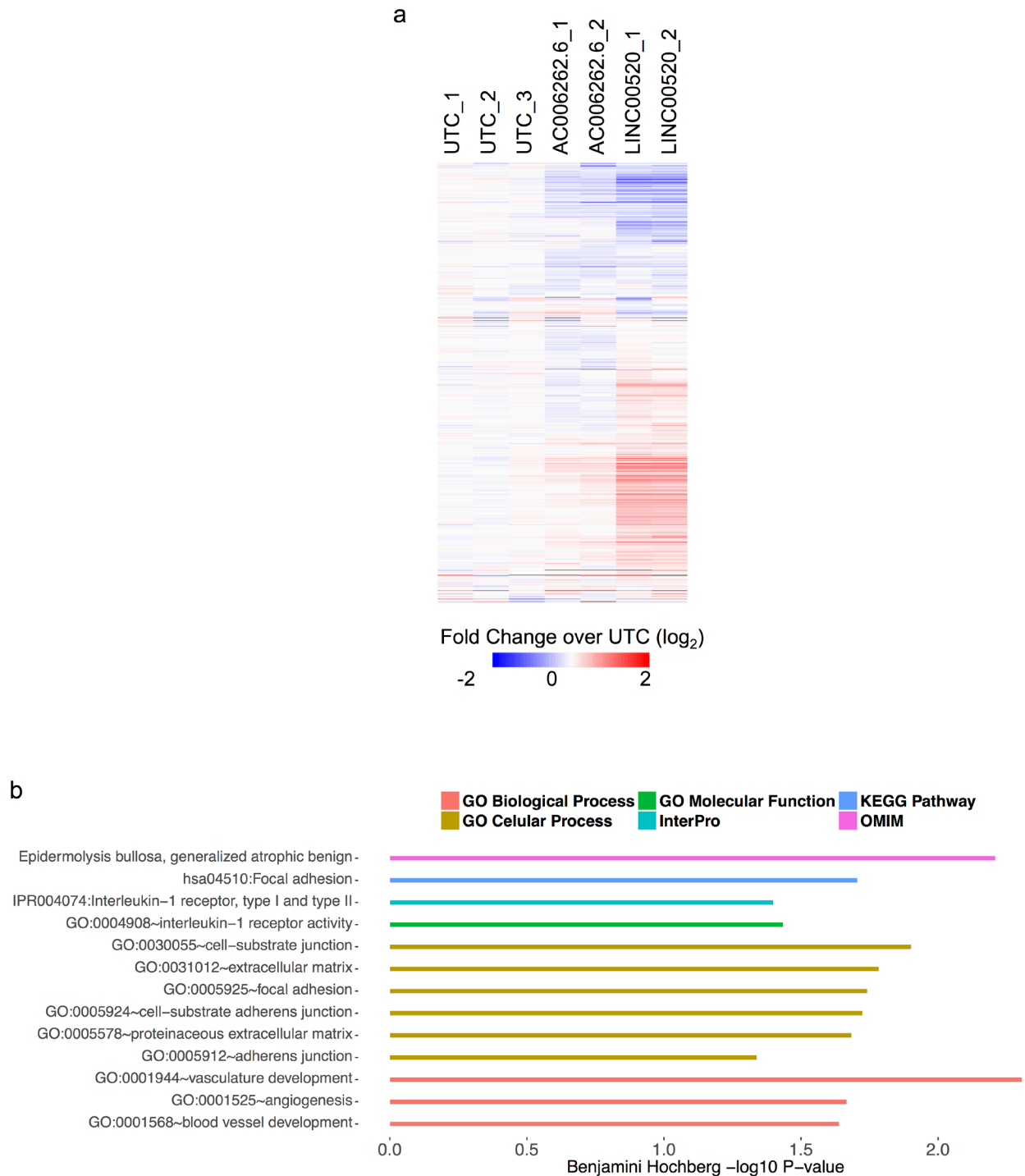

Figure 6: Heat map showing expression profile of MCF10A cells expressing LINC00520. a. All genes (1693 genes) depicted in the heat map are upregulated in MCF10A-Src transformed cells. Abbreviations: UTC - Untransformed MCF10A cells; AC006262.6 is a lncRNAs that is not involved in transformation (negative control). Scale bar shows log2 row median centered expression values. b. GO analysis of a subset of genes that are upregulated upon ectopic expression of LINC00520 in MCF10A cells (grey line in heat map denotes top 291 genes used for analysis).

the MCF10A-Src time course (Figure 6a). A minority of the 12-hour MCF10A-Src up-regulated gene set is down regulated in response to LINC00520 overexpression. However most of the changes in either direction with respect to the untransfected control are not observed in the control lncRNA transduction. Interestingly, a subset of these LINC00520-upregulated genes have been linked to cell migration and invasion among other cellular functions (Figure 6b, Supplementary Figure 4). This analysis is consistent with the findings that LINC00520 modulates cell migration in breast cancer cell lines, a phenotype that is recapitulated in *v-Src*-transformed cells.

DISCUSSION

Here we identify and describe a novel lncRNA, LINC00520, and provide evidence for a potential role in breast cancer. LINC00520 is transcriptionally upregulated in immortalized mammary epithelial cells transformed by oncogenic Src or oncogenic PI3K. We also show that the transcription factor STAT3 is critically required for the regulation of LINC00520 expression. Notably, expression of LINC00520 is increased in basal-like breast cancer cells, which also show a preferential increase in STAT3 activity. Thousands of lncRNAs have been annotated to date yet the biological significance of the majority of these lncRNAs remains unclear. In this study, we provide evidence for a role for LINC00520 in breast cancer. Both loss-of-function and gain-of-function approaches point to a role for LINC00520 in cell migration and cell invasiveness, arguing against the notion that LINC00520 represents a trivial, promiscuous transcript.

To date, no orthologs of LINC00520 have been described in other mammalian species. This is not unexpected, since most lncRNAs appear to be poorly conserved and tend to undergo rapid evolution. For instance, only approximately 12% of human and mouse lncRNAs have orthologs in other species [22, 23] and less than 6% of zebrafish lncRNAs share homologous sequences in humans or mice [24]. While this lack of conservation may suggest lncRNAs are merely cryptic noise in the genome, the characterization of Xist challenges this notion. Although Xist is critical for dosage compensation, most of its sequence is poorly conserved and deletion of exon 4, which shows the most convincing evidence for conservation, has no functional consequence [25]. A more comprehensive catalog of lncRNAs from several vertebrate species along with advancements in computational assessment tools may improve comparative lncRNA studies.

Our study also demonstrates that LINC00520 affects global gene expression. Importantly, ectopic expression of LINC00520 results in a transcriptional profile that shows significant overlap with that generated by Src-transformed cells. LINC00520 might mediate its function by interacting with various chromatin modifiers and/or

transcriptional regulators, as observed for lncRNAs such as HOTAIR, HOTTIP, MISTRAL and Xist [24, 26-28]. Despite these examples, it is still unclear whether these RNA-protein interactions are direct. Furthermore, the factors that dictate the specificity of these interactions have not been clearly defined. Regardless, future work will determine the mechanistic action of LINC00520. While there are well-studied examples of lncRNAs that are enriched in the nucleus [29-31], many lncRNAs appear to be cytoplasmic [32-34]. It will be interesting to determine the cellular localization of LINC00520, whether it undergoes any re-localization to various intracellular compartments and how it evades RNA decay.

Despite the definition of lncRNAs as noncoding, nearly half of the expressed lncRNAs encode peptides that are actually translated [35]. Ribosome profiling in ER-Src cells [35] reveals that non-overlapping regions of LINC00520 RNA are translated into 4 peptides (26, 32, 32, and 40 amino acids) that are conserved in monkeys but not in more evolutionary distant species. The stabilities and hence steady-state levels of these peptides are unknown, and whether one or more of these peptides are biologically functional remains to be determined.

In summary, this study supports the accumulating evidence that lncRNAs may function to modulate human cancer pathogenesis. It points to a role for lncRNAs in the mechanism of action of critical oncogenes, namely Src, PI3K and STAT3. To our knowledge, this is the first study that investigates the regulation and biological function of LINC00520. It also implicates the first lncRNA identified as a downstream effector of the PI3K pathway. Future studies will dissect the entire complement of LINC00520 biology and its significance in processes critical for breast cancer initiation and progression.

MATERIALS AND METHODS

Antibodies, plasmids and reagents

pBabe puro HA PIK3CA WT (Addgene plasmid #12522) and pBabe puro HA PIK3CA H1047R (Addgene plasmid #12524) was a gift from Dr. Jean Zhao. pBABE-puro was a gift from Hartmut Land, Jay Morgenstern and Robert Weinberg (Addgene plasmid #1764). Isogenic MDA-MB-231 cells expressing PIK3CA WT, PIK3CA E545K and PIK3CA H1047R were a kind gift from Dr. Jonathan Backer, and have been described [36].

Retroviral expression of LINC00520

LINC00520 was cloned as described previously [21].

RNA interference

For LINC00520 and PIK3CA shRNAs, single-stranded oligonucleotides sense and antisense pairs, encoding the indicated target sequences were synthesized and cloned into PLKO.1 vector. Stable cell lines were maintained in 2μg/ml puromycin.

PIK3CA shRNA#1, sense, 5'-CCGGGCACAATCCATGAACAGCATTCGAGAATGCTGTTCATGGATTGTGCTTTTTTG-3';

PIK3CA shRNA#1, antisense, 5'-AATTCAAAAACACAATCCATGAACAGCATTCTCGAGAATGCTGTTCATGGATTGTG-3'

PIK3CA shRNA#2, sense, 5'-CCGGGCATTAGAATTTACAGCAAGACTCGAGTCTTGCTGTAAATTCTAATGCTTTTTTG-3';

PIK3CA shRNA#2, antisense, 5'-AATTCAAAAAGCATTAGAATTTACAGCAAGACTCGAGTCTTGCTGTAAATTCTAATGC-3'

LINC00520 shRNA#1, sense, 5'-ccggAAGAGAAAAGCTGAAGGACACctcgagGTGTCCTTCAGCTTTTCTCTTtttttg

LINC00520 shRNA#1, antisense, 5'-ccggAAGAGAAAAGCTGAGAAGGACACctcgagGTGTCCTTCAGCTTTTCTCTT

LINC00520 shRNA#2, sense, 5'-ccggACCTCAAATCTTTCGAGAACActcgagTGTTCTCGAAAGATTTGAGGTttttg

LINC00520 shRNA#2, antisense, 5'-AATTCAAAAAACCTCAAATCTTTCGAGAACActcgagTGTTCTCGAAAGATTTGAGGT

Cell culture and immunoblotting

MCF10A and MCF10DCIS.com [37] cells were cultured in Ham's F12/DMEM (Cellgro) supplemented with 5% equine serum (Cellgro), 500ng/ml hydrocortisone (Sigma-Aldrich), 100ng/ml cholera toxin (List Biological Laboratories), 20ng/ml EGF (R&D Systems) and 10ug/ml insulin. SKBR3 cells were maintained in 10% FBS/McCoy's (Cellgro). MCF7, and MDA-MB-468, MDA-MB-453, MDA-MB-231 cells were maintained in Dulbecco's modified Eagle medium (DMEM; Cellgro) supplemented with 10% Fetal Bovine Serum (FBS; Cyclone). T47D and BT549 cells were cultured in 10% FBS/DMEM, supplemented with 1mg/ml insulin (Sigma-Aldrich). SUM159-PT cells were grown in Ham's F12 medium (Cellgro) supplemented with 5% FBS, 1ug/ml hydrocortisone (Sigma-Aldrich) and 5ug/ml insulin (Sigma-Aldrich). ZR75-1 and HCC1806 were maintained in RPMI 1640 supplemented with 10% FBS (Cellgro). MCF10A-Src inducible cell lines were cultured as previously described [17]. For all western blotting, cells were lysed in RIPA buffer with protease and phosphatase inhibitors.

3D Morphogenesis assay

MCF10DCIS cells were grown in three dimensional Matrigel cultures as described [38]. Briefly, $3x10^3$ cells were suspended in modified growth medium containing 2% growth factor-reduced Matrigel, 2% Horse serum (Cellgro) and 5ng/ml EGF (R&D systems). Cell mixture was plated on top of a solidified layer of growth factor-reduced Matrigel. Cells were fed every 4 days. Phase contrast images were acquired using the Nikon Eclipse Ti microscope.

Quantitative real-time RT PCR

Total RNA was isolated using the RNeasy kit following the manufacturer's instructions (Qiagen). Reverse transcription was performed using Quantitect Reverse transcription kit according to the manufacturer's instructions (Qiagen). Quantitative real-time RT-PCR was performed using SYBR Green PCR Master Mix (BioRad) and the ABI Prism 7900 sequence detector (Applied Biosystems). Relative mRNA expression was calculated by the ΔΔCT method with GAPDH as reference. Primer sequences:

LINC00520 sense: 5'- GTGTACATTTCTGGGTAGCTT

LINC00520 antisense: 5' - AAAGGAAAACAATACAGGCTTG

GAPDH sense: 5' - GCAAATTCCATGGCACCGT

GAPDH antisense: 5'- TCGCCCCACTTGATTTTGG

Transwell migration assay

Transwell migration assay using SUM159-PT cells was performed as previously described [39]. Cells were allowed to migrate through an 8μm-pore transwell (Corning) for 16 hours and NIH3T3 conditioned media was used as a chemoattractant in the lower chamber. MCF10A-Src-Hygro inducible cells containing LINC00520 shRNA constructs were first treated with 1μM Tamoxifen or ethanol (vehicle) for 48 hours prior to migration. Cells were then serum starved and then allowed to migrate for 16 hours using MCF10A growth media as the chemoattractant. MCF10A-Src-Hygro inducible cell lines were generated using a retroviral pLHCX vector (Clontech) containing an ER-Src inducible gene. Stable cell lines were established after selection with 150μg/ml hygromycin for 5 days and maintained thereafter in 50μg/ml.

RNA-Seq analysis and ChIP-Seq analysis

RNA whole transcriptome sequencing (RNA-seq) was carried out at 0 (1 replicate), 4 (2 replicates), 12 (1 replicate) and 36hrs (2 replicates) post tamoxifen treatment on polyA selected RNA using Illumina tru-seq library construction. RNA-seq was also carried out for the over-expression experiments in parental MCF10A cells that were untransduced (3 replicates), transduced with LINC00520 (2 replicates) or transduced with a control lincRNA AC006262.6 (2 replicates) on polyA selected RNA using Illumina tru-seq library construction harvested 72 hours post transduction.

Transcript levels were quantified and differentially expressed genes were called using cuffdiff2 [40]. Relative transcript levels are expressed as a "Fragments Per Kilobase of transcript per Million mapped" (FPKM) which corrects for the length of the transcript and the depth of the libraries. Raw and processed ChIP-Seq data was retrieved from [18].

Expression analysis in clinical samples of human breast invasive carcinoma

Genomic alterations associated with LINC00520 were identified by querying genomic data from Breast Invasive Carcinomas dataset of TCGA. Data from 850 breast cancer cases were retrieved from the TCGA database (http://cancergenome.nih.gov/), including RNAseq gene expression (Illumina HiSeq RNASeqV2 Level 3.1.9.0) and lncRNA expression from TANRIC database (http://bioinformatics.mdanderson.org/main/TANRIC:Overview). Intrinsic molecular subtype data was obtained by applying the PAM50 algorithm to the RNA-seq data.

LINC00520 expression level was compared in normal-tumor paired samples ($n = 105$) using a T-test for paired samples and compared using One-way ANOVA (with unequal variance) in carcinoma samples ($n = 741$) by intrinsic molecular subtypes. Afterwards, LINC00520 expression levels were dichotomized into low and high expression categories using univariate clustering based on finite normal mixture modeling (mclust 5.1 package, R 3.2.2). Differential gene expression was performed using two-class unpaired significance of microarray analysis (SAM 2.0 package, R 3.2.2) [41] in basal carcinomas previously classified with low ($n = 96$) and high ($n = 30$) expression of LINC00520. Next, we performed a pre-ranked GSEA (Gene-Set Enrichment Analysis) using software provided by the Broad Institute (http://www.broadinstitute.org/gsea/msigdb/annotate.jsp) on a gene list ranked based on the d-statistic computed from the LINC00520 differential expression analysis, and we assessed enrichment using the Broad Institute´s Molecular Signatures Database (mSigDB) Hallmark gene sets collection ($n = 50$) (http://www.broadinstitute.org/gsea/msigdb/). Finally, functional annotation for the transcriptomic signature of basal-like high LINC00520 expression carcinomas was performed using Database for Annotation, Visualization and Integrated Discovery (DAVID) 6.7 [42]. Gene set enrichments were assessed using Gene Ontology Biological Pathway and KEGG Pathways summarized version with 13,015 genes (total genes in RNASeqV2 data) as background list and gene sets containing a minimum of 15 genes. Statistical significance was considered when p-value was < 0.05 or false discovery rate (FDR) was < 0.05.

Sulforhodamine B (SRB) proliferation assay

For proliferation assay 0.01 x 10^6 cells were seeded in complete growth media. Relative number of adherent cells at day 0 (~16hrs post-seeding), days 1, 3, and 4, were assessed using sulforhodamine B assay as previously described [43]. Briefly, adherent cells were fixed with 12.5% (w/v) trichloroacetic acid for 1 hour at 4°C. Cells were then rinsed three times with water and stained with a solution of 0.5% (w/v) SRB in 1% acetic acid for at least 30 minutes at room temperature. Cells were then washed three times with 1% acetic acid and allowed to dry. SRB was dissolved in 10 mmol/l Tris (pH 10.5). Absorbance of solubilized SRB was measured at 510 nm.

Accesion number

The accession number for the raw RNA-Seq data reported in this manuscript is GEO: GSE86159.

REFERENCES

1. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071-1076.

2. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature biotechnology. 2011; 29:742-749.

3. Lin R, Maeda S, Liu C, Karin M and Edgington TS. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene. 2007; 26:851-858.

4. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE, Shimizu M, Tili E, Rossi S, Taccioli C, Pichiorri F, Liu X, et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer cell. 2007; 12:215-229.

5. Huarte M. The emerging role of lncRNAs in cancer. Nat Med. 2015; 21:1253-1261.

6. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T and Rinn JL. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010; 142:409-419.

7. Zheng GX, Do BT, Webster DE, Khavari PA and Chang HY. Dicer-microRNA-Myc circuit promotes transcription of hundreds of long noncoding RNAs. Nature structural & molecular biology. 2014; 21:585-590.

8. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N and Isaacs WB. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer research. 1999; 59:5975-5979.

9. Hessels D, Klein Gunnewiek JM, van Oort I, Karthaus HF, van Leenders GJ, van Balken B, Kiemeney LA, Witjes JA and Schalken JA. DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. European urology. 2003; 44:8-15; discussion 15-16.

10. Srikantan V, Zou Z, Petrovics G, Xu L, Augustus M, Davis L, Livezey JR, Connell T, Sesterhenn IA, Yoshino K, Buzard GS, Mostofi FK, McLeod DG, Moul JW

and Srivastava S. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97:12216-12221.

11. Yang L, Lin C, Jin C, Yang JC, Tanasa B, Li W, Merkurjev D, Ohgi KA, Meng D, Zhang J, Evans CP and Rosenfeld MG. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature. 2013; 500:598-602.

12. Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H and Muller-Tidow C. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene. 2003; 22:8031-8041.

13. Huarte M and Rinn JL. Large non-coding RNAs: missing links in cancer? Human molecular genetics. 2010; 19(R2):R152-161.

14. Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ and Zhou MM. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Molecular cell. 2010; 38:662-674.

15. Khaitan D, Dinger ME, Mazar J, Crawford J, Smith MA, Mattick JS and Perera RJ. The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. Cancer research. 2011; 71:3852-3862.

16. Chisholm KM, Wan Y, Li R, Montgomery KD, Chang HY and West RB. Detection of long non-coding RNA in archival tissue: correlation with polycomb protein expression in primary and metastatic breast carcinoma. PLoS One. 2012; 7:e47998.

17. Iliopoulos D, Hirsch HA and Struhl K. An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation. Cell. 2009; 139:693-706.

18. Fleming JD, Giresi PG, Lindahl-Allen M, Krall EB, Lieb JD and Struhl K. STAT3 acts through pre-existing nucleosome-depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer. Epigenetics & chromatin. 2015; 8:7.

19. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, Yan H, Gazdar A, Powell SM, Riggins GJ, Willson JK, Markowitz S, Kinzler KW, Vogelstein B and Velculescu VE. High frequency of mutations of the PIK3CA gene in human cancers. Science. 2004; 304:554.

20. Isakoff SJ, Engelman JA, Irie HY, Luo J, Brachmann SM, Pearline RV, Cantley LC and Brugge JS. Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. Cancer research. 2005; 65:10992-11000.

21. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, Morse M, Engreitz J, Lander ES, Guttman M, Lodish HF, Flavell R, et al. Topological

organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nature structural & molecular biology. 2014; 21:198-206.

22. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques AC, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009; 7:e1000112.

23. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A and Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development. 2011; 25:1915-1927.

24. Penny GD, Kay GF, Sheardown SA, Rastan S and Brockdorff N. Requirement for Xist in X chromosome inactivation. Nature. 1996; 379:131-137.

25. Caparros ML, Alexiou M, Webster Z and Brockdorff N. Functional analysis of the highly conserved exon IV of XIST RNA. Cytogenetic and genome research. 2002; 99:99-105.

26. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E and Chang HY. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010; 329:689-693.

27. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA and Chang HY. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011; 472:120-124.

28. Bertani S, Sauer S, Bolotin E and Sauer F. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Molecular cell. 2011; 43:1040-1046.

29. Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J and Willard HF. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell. 1992; 71:527-542.

30. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB and Chess A. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC genomics. 2007; 8:39.

31. Sone M, Hayashi T, Tarui H, Agata K, Takeichi M and Nakagawa S. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. Journal of cell science. 2007; 120:2498-2506.

32. Coccia EM, Cicala C, Charlesworth A, Ciccarelli C, Rossi GB, Philipson L and Sorrentino V. Regulation and expression of a growth arrest-specific gene (gas5) during growth, differentiation, and development. Molecular and cellular biology. 1992; 12:3514-3521.

33. Kino T, Hurt DE, Ichijo T, Nader N and Chrousos GP. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci Signal. 2010; 3:ra8.

34. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, Huarte M, Zhan M, Becker KG and Gorospe M. LincRNA-p21 suppresses target mRNA translation. Molecular cell. 2012; 47:648-655.

35. Ji Z, Song R, Regev A and Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife. 2015; 4.

36. Pang H, Flinn R, Patsialou A, Wyckoff J, Roussos ET, Wu H, Pozzuto M, Goswami S, Condeelis JS, Bresnick AR, Segall JE and Backer JM. Differential enhancement of breast cancer cell motility and metastasis by helical and kinase domain mutations of class IA phosphoinositide 3-kinase. Cancer research. 2009; 69:8868-8876.

37. Miller FR, Santner SJ, Tait L and Dawson PJ. MCF10DCIS.com xenograft model of human comedo ductal carcinoma in situ. Journal of the National Cancer Institute. 2000; 92:1185-1186.

38. Debnath J, Muthuswamy SK and Brugge JS. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods. 2003; 30:256-268.

39. Chin YR and Toker A. The Actin-Bundling Protein Palladin Is an Akt1-Specific Substrate that Regulates Breast Cancer Cell Migration. Molecular cell. 2010; 38:333-344.

40. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL and Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013; 31:46-53.

41. Tusher VG, Tibshirani R and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:5116-5121.

42. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology. 2003; 4:P3.

43. Brown KK, Montaser-Kouhsari L, Beck AH and Toker A. MERIT40 Is an Akt Substrate that Promotes Resolution of DNA Damage Induced by Chemotherapy. Cell reports. 2015; 11:1358-1366.