EPIGENETICS

Portraying breast cancers with long noncoding RNAs

10.1126/sciadv.1600220

Olivier Van Grembergen,[1] Martin Bizet,[1,2,3] Eric J. de Bony,[1] Emilie Calonne,[1] Pascale Putmans,[1] Sylvain Brohée,[4] Catharina Olsen,[2] Mingzhou Guo,[5] Gianluca Bontempi,[2,3] Christos Sotiriou,[4] Matthieu Defrance,[1,3] François Fuks[1]*

Evidence is emerging that long noncoding RNAs (lncRNAs) may play a role in cancer development, but this role is not yet clear. We performed a genome-wide transcriptional survey to explore the lncRNA landscape across 995 breast tissue samples. We identified 215 lncRNAs whose genes are aberrantly expressed in breast tumors, as compared to normal samples. Unsupervised hierarchical clustering of breast tumors on the basis of their lncRNAs revealed four breast cancer subgroups that correlate tightly with PAM50-defined mRNA-based subtypes. Using multivariate analysis, we identified no less than 210 lncRNAs prognostic of clinical outcome. By analyzing the coexpression of lncRNA genes and protein-coding genes, we inferred potential functions of the 215 dysregulated lncRNAs. We then associated subtype-specific lncRNAs with key molecular processes involved in cancer. A correlation was observed, on the one hand, between luminal A–specific lncRNAs and the activation of phosphatidylinositol 3-kinase, fibroblast growth factor, and transforming growth factor–β pathways and, on the other hand, between basal-like–specific lncRNAs and the activation of epidermal growth factor receptor (EGFR)–dependent pathways and of the epithelial-to-mesenchymal transition. Finally, we showed that a specific lncRNA, which we called CYTOR, plays a role in breast cancer. We confirmed its predicted functions, showing that it regulates genes involved in the EGFR/mammalian target of rapamycin pathway and is required for cell proliferation, cell migration, and cytoskeleton organization. Overall, our work provides the most comprehensive analyses for lncRNA in breast cancers. Our findings suggest a wide range of biological functions associated with lncRNAs in breast cancer and provide a foundation for functional investigations that could lead to new therapeutic approaches.

INTRODUCTION

Breast cancer is a major public health issue. According to the most recent worldwide cancer statistics, more than 1,675,000 women are diagnosed with this disease each year and more than 500,000 die of it (*1*). Breast cancer is a heterogeneous disease, and different subtypes have been described (*2*). Beyond the classic grading system (based on tumor cell differentiation status) and the TNM (tumor size, lymph node involvement, and metastasis) classification, breast tumors are also classified on the basis of protein and gene status. Clinically, breast tumors are subclassified into three main subgroups on the basis of estrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) gene expression: ER-positive tumors (the most frequently diagnosed subtype), HER2-positive tumors (harboring an *ERBB2* amplification), and triple-negative breast cancers lacking ER, HER2, and the progesterone receptor (*3*). Additionally, microarray-based coding mRNA expression profiling has identified five "intrinsic" subtypes (*2*). Luminal A (low-grade) and luminal B (high-grade) tumors are predominantly ER^+. The $HER2^+$ subtype mostly comprises tumors overexpressing the $HER2^+$ gene. The basal-like subtype is particularly frequent among triple-negative cancers. The least frequent subtype, called "normal-like," comprises tumors that have an expression pattern similar to that of normal breast tissues. Recently, because of whole-transcriptome sequencing [RNA sequencing (RNA-seq)], this classification has been refined, with the identification of 12 breast tumor subgroups (*4*). Despite these advances, we are far from completely understanding breast cancer heterogeneity because tumors of the same subtype can respond differently to therapy and can have different outcomes (*5*, *6*). Understanding the molecular mechanisms that underlie breast cancer heterogeneity thus remains a major challenge improving diagnosis and therapy, and new approaches are needed to meet this challenge. Recent work has suggested that long noncoding RNAs (lncRNAs) and small noncoding RNAs (such as microRNAs) might play key roles in mammary tumor development (*7*, *8*) and provide new biomarkers and potential targets for future therapies.

lncRNAs are transcripts more than 200 base pairs long that lack an extended open reading frame and thus do not code for proteins (*9*). According to recent studies, the human transcriptome contains up to 16,000 lncRNAs, frequently spliced and polyadenylated, whose genes are mainly transcribed by RNA polymerase II (*9*). Expression of lncRNA genes is lower than that of protein-coding genes, but is tissue-specific (*9*). In recent years, studies have linked lncRNAs to a wide variety of physiological and pathological mechanisms, including pluripotency regulation and cancer development (*10*). Like proteins, lncRNAs may mediate oncogenic or tumor-suppressive effects (*10*, *11*). They can exert various functions in the cytoplasm (for example, as scaffolds between proteins or as microRNA sponges) and the nucleus (*10*). They have emerged as key players in the transcriptional regulation of protein-coding genes, in which case they can act either distally (in trans) or proximally (in cis) (*10*). For instance, the lncRNA HOTAIR (HOX transcript antisense RNA), which is up-regulated in some breast cancer tissues and whose expression is associated with poor prognosis and tumor metastasis, is suggested to silence tumor suppressor genes (*7*, *10*). Other lncRNAs appear as key regulators of pathways underlying carcinogenesis; an example is lincRNA-p21, which mediates global gene repression in the p53 response (*12*). Genome-wide association studies on cancer have

[1]Laboratory of Cancer Epigenetics, Faculty of Medicine, ULB–Cancer Research Center (U-CRC), Université Libre de Bruxelles (ULB), 1070 Brussels, Belgium. [2]Machine Learning Group, Computer Science Department, Université Libre de Bruxelles, 1050 Brussels, Belgium. [3]Interuniversity Institute of Bioinformatics Brussels, Université Libre de Bruxelles–Vrije Universiteit Brussel, 1050 Brussels, Belgium. [4]Breast Cancer Translational Research Laboratory, Jules Bordet Institute, Université Libre de Bruxelles, 1000 Brussels, Belgium. [5]Department of Gastroenterology and Hepatology, Chinese People's Liberation Army General Hospital, Beijing 100853, China. *Corresponding author. Email: ffuks@ulb.ac.be

revealed that more than 80% of cancer-associated single-nucleotide polymorphisms occur in noncoding regions. This suggests that a significant fraction of the genetic etiology of cancer is related to lncRNAs (*13*). Previous studies have documented aberrantly expressed lncRNA genes in breast tumors and have notably established associations between certain lncRNAs and known gene expression–based breast cancer subtypes (*14*, *15*). However, these lncRNAs have not been precisely related to molecular pathways, and their functions have not been investigated.

The goal of the present study was to explore lncRNA landscape in breast cancers and to extract novel biological and clinical information. On the basis of an array-based transcriptional survey of more than 3000 lncRNA genes, we have identified lncRNAs aberrantly expressed across 823 breast tumors, as compared to 172 normal samples. Although gene chips contain less lncRNAs than other technologies for which data are publicly available (for example, RNA-seq), we used microarray data for the following reasons: (i) low technical variation (*16*–*18*), (ii) strand specificity, (iii) larger number of publicly available data, and (iv) long follow-up clinical annotation.

From our in-depth analyses, we have inferred potential functions of dysregulated lncRNAs and demonstrated their relevance to breast cancer classification. We have investigated lncRNAs as survival markers, associating several of them with prognosis. Finally, we have experimentally characterized the functions of one breast cancer–related lncRNA—CYTOR (cytoskeleton regulator). Overall, this work provides the most comprehensive data sets so far for lncRNA in breast cancers. It highlights the influence of lncRNAs in numerous pathways that are dysregulated in tumors and may provide novel approaches to cancer prognosis and treatment.

RESULTS

Breast tumors display lncRNA gene expression profiles that are distinct from those of normal breast tissues

Seven data sets from the Gene Expression Omnibus (GEO) data repository (*19*) were selected and compiled to generate a large cohort of 823 breast tumors and 172 normal breast tissues (Fig. 1A and table S1). The data sets were selected on the basis of their size (more than 50 samples), the presence of breast tumors from each subtype, and extensive clinical annotation, including relapse information. To increase the number of normal samples, we included the GSE10780 series that profiled 143 normal breasts and 42 tumors.

First, we reannotated the entire collection of probe sets of the Affymetrix Human Genome U133 Plus 2.0 Array to analyze levels of 16,951 mRNAs and 3053 lncRNAs (table S2). When the distribution of lncRNA gene expression levels across all breast tissue samples was compared with that of coding gene expression levels, it was confirmed that the former generally show lower level expression than do the latter (fig. S1A) (*9*). On the basis of the observed lncRNA gene expression profiles in tumors and normal breast tissues, we then performed unsupervised consensus clustering (*20*, *21*) and identified three robust clusters, one of which (cluster II) contained almost all of the normal samples (95%) (P value of the association = 2.3×10^{-164}). Cancer and healthy tissues thus appear to have different lncRNA gene expression profiles (Fig. 1B). We therefore sought to identify lncRNAs whose genes are differentially expressed in breast tumors versus normal breast tissue. Because of the heterogeneity of breast cancer, we did not apply the classical t test but adapted a method that allows detection of differentially expressed lncRNAs in a fraction of cancer samples (*22*). We identified 215 lncRNAs whose genes appeared aberrantly expressed in at least 10% of the breast tumors (see Materials and Methods, fig. S1B, and table S3). Among these 215 lncRNAs, 124 appeared up-regulated and 91 appeared down-regulated in breast cancer (Fig. 1, C and D, top). Some of these identified lncRNAs have already been described as involved in breast cancer development, such as HOTAIR, MALAT1, H19, and GAS5 (*10*, *14*). The genes that encode the 215 identified lncRNAs appeared scattered across the genome, without any focal hot spot (fig. S1C).

We next examined whether lncRNA levels, like the levels of protein-coding transcripts, can serve as biomarkers for breast cancer diagnosis. For this, we generated an lncRNA metagene and assessed its ability to discriminate tumor samples from healthy ones. In the discovery data set, significant discrimination of breast tumors from healthy samples was achieved ($P < 0.0001$) with high specificity (0.96) and high sensitivity (0.95). To validate our findings on an independent data set, we used RNA-seq data from The Cancer Genome Atlas (TCGA) consortium reanalyzed by Rahman *et al.* (*23*) that we called "validation set 1A." This processed data set allows assessment of the expression of 1161 lncRNAs from 1052 breast tumors and 113 normal samples. On the basis of lncRNA expression, we performed unsupervised hierarchical clustering and observed three stable clusters (Fig. 1E), one of which contained almost all of the normal samples, suggesting that our results could be extended to independent cohorts profiled by RNA-seq. Of the 215 lncRNAs identified as dysregulated in the discovery cohort, we could assess the expression of 87 lncRNAs present in the RNA-seq data set. Of them, 82 were dysregulated in the RNA-seq data (see examples in Fig. 1D, bottom). In addition, the expression pattern of a metagene that represents the expression of these 87 lncRNAs allowed highly accurate classification of the cancer versus normal samples (specificity, 0.98; sensitivity, 0.93; $P < 0.0001$) (fig. S1D). We also reannotated the custom Agilent 244K microarray used by the TCGA consortium (*4*). Of the 215 dysregulated lncRNAs identified from the discovery cohort, 167 were detectable on the TCGA microarray, which we called "validation set 1B." The expression patterns of the corresponding genes allowed a highly predictive classification of breast cancer and normal breast tissues in the TCGA validation set composed of 524 breast tumors and 63 normal tissues (specificity, 0.97; sensitivity, 0.98; $P < 0.0001$) (fig. S1E). Together, these analyses provide a validated set of lncRNAs that are dysregulated in breast tumors, as compared to normal human samples.

lncRNA expression is associated with ER signaling

On the basis of hierarchical clustering (Fig. 1B), we noticed that breast tumors are separated into two different clusters, and we further searched for their relationship with clinical properties. We observed a significant association with the immunohistochemistry (IHC)–based ER status, that is, marked differential expression of lncRNAs between ER-positive and ER-negative tumors. Cluster I was found to contain 91% ER-positive tumors (P value of the association = 6.7×10^{-61}), whereas cluster III contained 76% ER-negative tumors (P value of the association = 3.6×10^{-63}) (Fig. 2A and table S4). Given this finding, we next used a supervised approach to identify lncRNAs specifically associated with the ER status. From the whole set of lncRNAs (that is, 3053 lncRNAs), our analysis revealed 38 lncRNA genes differentially expressed between ER-positive and ER-negative breast tumors, determined by IHC [false discovery rate (FDR) < 0.05; fold change > 1.5, t test] (see examples in Fig. 2B). Whereas an association between

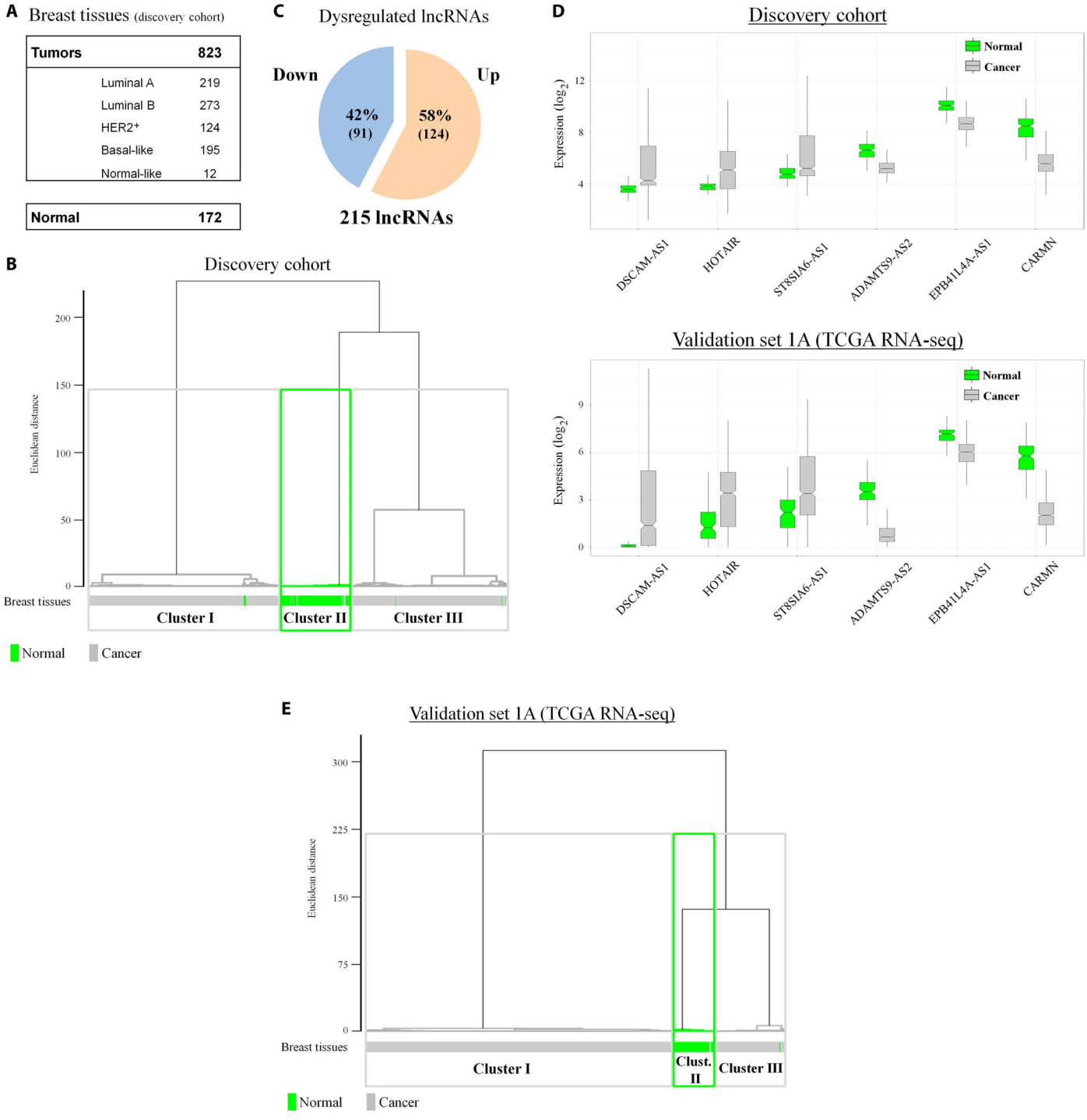


Fig. 1. lncRNA gene expression profiling in breast tissues reveals 215 dysregulated lncRNAs. (**A**) Description of the human breast tissues analyzed in this study. (**B**) Unsupervised consensus clustering of the samples on the basis of lncRNA gene expression. Primary tumors (823) and normal samples (172) were used for hierarchical clustering on the basis of the 500 most variant lncRNAs (based on SD). (**C**) Pie chart showing, among the lncRNAs that are dysregulated in breast tumors versus normal tissues, the numbers of down- and up-regulated ones. (**D**) Box plot for expression levels of the top dysregulated lncRNAs in the discovery cohort (top) and their expression in the TCGA RNA-seq cohort (bottom). Notches are used to compare groups; if the notches of two boxes do not overlap, the medians differ significantly. The whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range of the box. (**E**) Same as in (B) for the TCGA RNA-seq cohort composed of 1052 breast tumors and 113 normal samples.

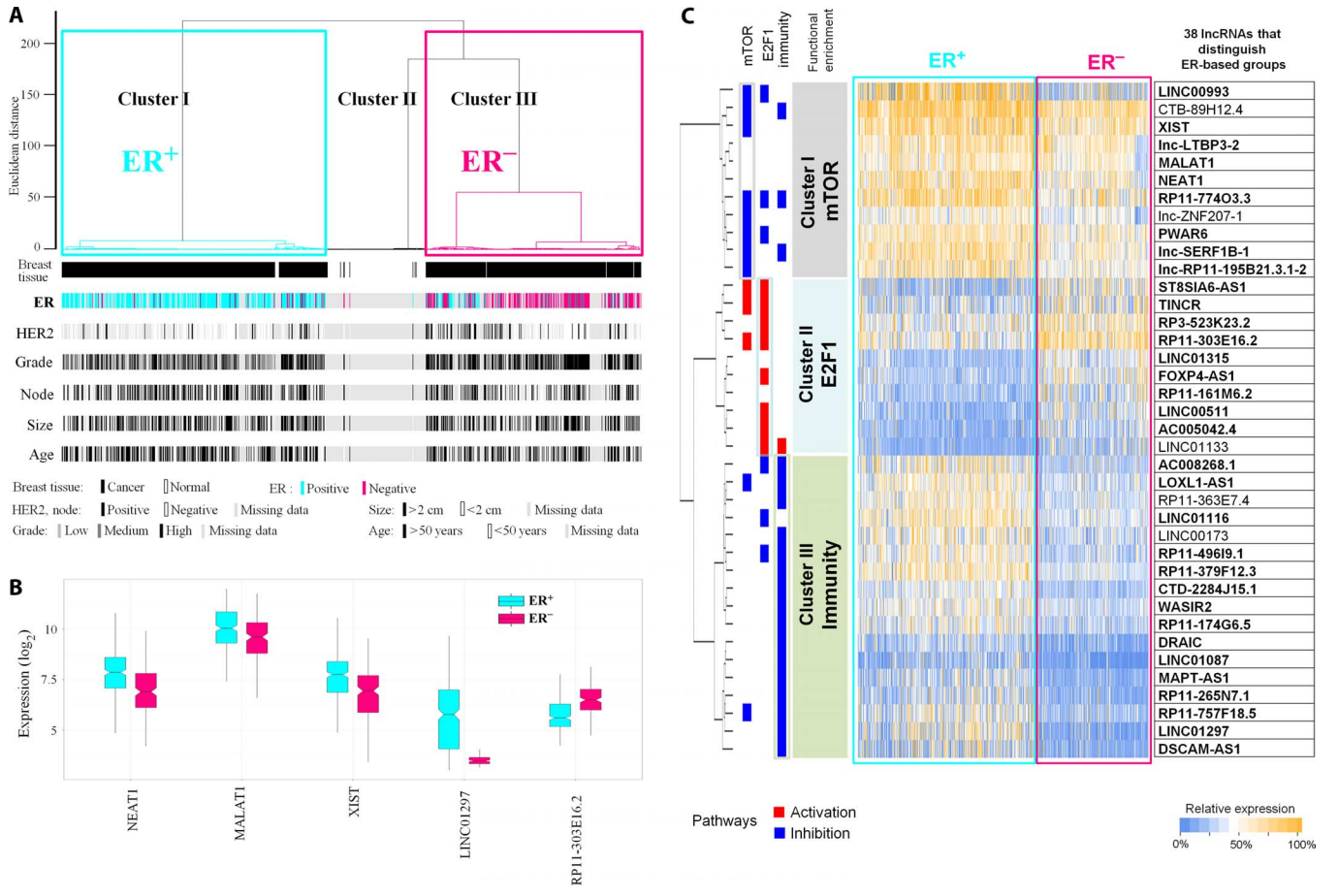


Fig. 2. lncRNA gene expression profiling identifies two main breast tumor categories differing with regard to ER status. (**A**) Dendrogram of 823 primary tumors and 172 normal samples obtained by consensus hierarchical clustering of the samples on the basis of expression of the top 500 most variant lncRNAs. Clusters I and III, encompassing almost all tumors, are related to the ER status. (**B**) Box plot illustrating the expression levels of five lncRNA genes differentially expressed between ER^+ and ER^- tissues. Notches are used to compare groups; if the notches of two boxes do not overlap, the medians differ significantly. The whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range of the box. (**C**) Heat map illustrating the expression of the 38 lncRNA genes (rows) of the ER signature across the breast tumors (columns). The lncRNAs in bold represent lncRNAs dysregulated between breast cancer and normal samples. The color scale of the heat map indicates the relative expression of each lncRNA gene. Hierarchical clustering reveals three clusters of lncRNAs. For each cluster, the most significant functional enrichment term from the guilt-by-association analysis is shown.

the ER status and some of the identified lncRNAs [for example, NEAT1 (*24*), MALAT1 (*25*), and Xist (*26*)] had previously been evidenced, most lncRNAs highlighted here are novel in the context of breast cancer. This is the case for LINC01297, the most significantly up-regulated lncRNA in ER-positive tumors, and RP11-303E16.2, the most significantly down-regulated lncRNA in these tumors (Fig. 2B). Notably, 32 of the 38 ER-associated lncRNAs were also dysregulated in breast tumors, as compared to normal samples.

We next used different data sets [validation set 1A (RNA-seq data from TCGA), validation set 1B (microarray data from TCGA), and validation set 2 (GSE20685)] to assess to what extent one could predict the ER status on the basis of expression levels of the lncRNA genes identified above. In validation set 1A (TCGA RNA-seq), we assessed the expression of 12 of the 38 lncRNAs from the ER-associated lncRNAs and found that their expression allowed highly accurate prediction of the ER status (specificity, 0.93; sensitivity, 0.83; $P < 0.0001$). Validation set 1B (TCGA microarray) was generated on a different microarray platform, which allowed assessment of the expression of 29 of the 38 lncRNA genes identified. The expression patterns of these 29 lncRNA were again highly predictive of the ER status (specificity, 0.89; sensitivity, 0.78; $P < 0.0001$). Finally, in the GSE20685 data set, the expression of the 38 genes can also predict the ER status, with a specificity of 0.86 and a sensitivity of 0.91 ($P < 0.0001$). Prediction of the ER status on the basis of these patterns was thus highly reproducible (fig. S2).

To gain insight into the biological relevance of the 38 lncRNAs, we used the "guilt-by-association" approach (*27*) to investigate their relationship to different pathways (see Materials and Methods). Briefly, we computed the matrix of correlation between lncRNA levels and expression levels of protein-coding genes across all breast cancer samples and then generated hypotheses regarding the function(s) of each dysregulated lncRNA on the basis of the known biological functions and the molecular pathways of protein-coding genes that show a good correlation (table S5). We used hierarchical clustering to subdivide the 38 lncRNAs that constitute the ER signature into three groups on the basis of their levels in the various tumors (Fig. 2C). This analysis revealed enrichment in key breast cancer–related pathways for each group

of lncRNAs. Levels of group 1 lncRNAs, which tended to be up-regulated in ER-positive tumors, appeared to correlate most strongly with mammalian target of rapamycin (mTOR) pathway inhibition. Group 2 lncRNAs appeared most strongly related to E2F1 pathway activation. Group 3 contained lncRNAs that are potentially involved in inhibition of immunity. Overall, our findings suggest that lncRNA levels can distinguish ER^+ from ER^- tumors and that lncRNAs are involved in various processes beyond ER biology. This sheds new light on pathways related to the ER status of breast cancers, such as the mTOR pathway and immunity (*4*, *28*, *29*).

Identification and inferred functions of lncRNAs that are associated with molecular subtypes of breast cancers

We next examined the relationship between the expression of lncRNA genes and the intrinsic breast tumor subtypes. lncRNA level–based consensus clustering of the 823 tumors identified four robust clusters (assessed by the ConsensusClusterPlus algorithm). There was also good agreement between lncRNA-based clusters and PAM50-defined mRNA-based subtypes: clusters 1, 2, 3, and 4 were found to contain 96% basal-like (*P* value of the association = 1.9×10^{-142}), 52% $HER2^+$ (*P* value of the association = 2.6×10^{-60}), 84% luminal A (*P* value of the association = 1.6×10^{-50}), and 64% luminal B (*P* value of the association = 1.8×10^{-47}) samples, respectively ($P = 1.2 \times 10^{-242}$, χ^2 test) (table S6). We thus used the PAM50 classification as a reference to identify molecular subtype–specific lncRNAs, which defines a specific lncRNA signature for each breast cancer subtype (see Materials and Methods). These signatures consisted of 42 lncRNAs (36 up-regulated and 6 down-regulated) for luminal A, 9 (8 up-regulated and 1 down-regulated) for luminal B, 14 (8 up-regulated and 6 down-regulated) for $HER2^+$, and 74 (28 up-regulated and 46 down-regulated) for the basal-like subtype (Fig. 3B and table S7). The large size of the basal-like signature distinguishes the basal-like subtype as particularly perturbed at the level of lncRNA gene expression.

Again, we looked at the predictive values of these signatures in the validation data sets used previously [validation set 1A (RNA-seq data from TCGA), validation set 1B (microarray data from TCGA), and validation set 2 (GSE20685)]. We found each subtype-specific lncRNA signature to identify with high-efficiency samples belonging to the considered subtype. As in the case of subtype mRNA signatures (*30*), the best prediction scores were obtained for the basal-like subtype, thus confirming that this subtype has a more distinct profile than the others. Overall, these analyses performed on cohorts independent of our initial 823 tumor samples validated the four subtype-specific signatures (fig. S3).

We next wondered whether these subtype-specific lncRNAs might be globally related to the alteration of different biological functions and molecular pathways. To investigate this, we again used the guilt-by-association approach, relating the functions predicted for subtype-specific

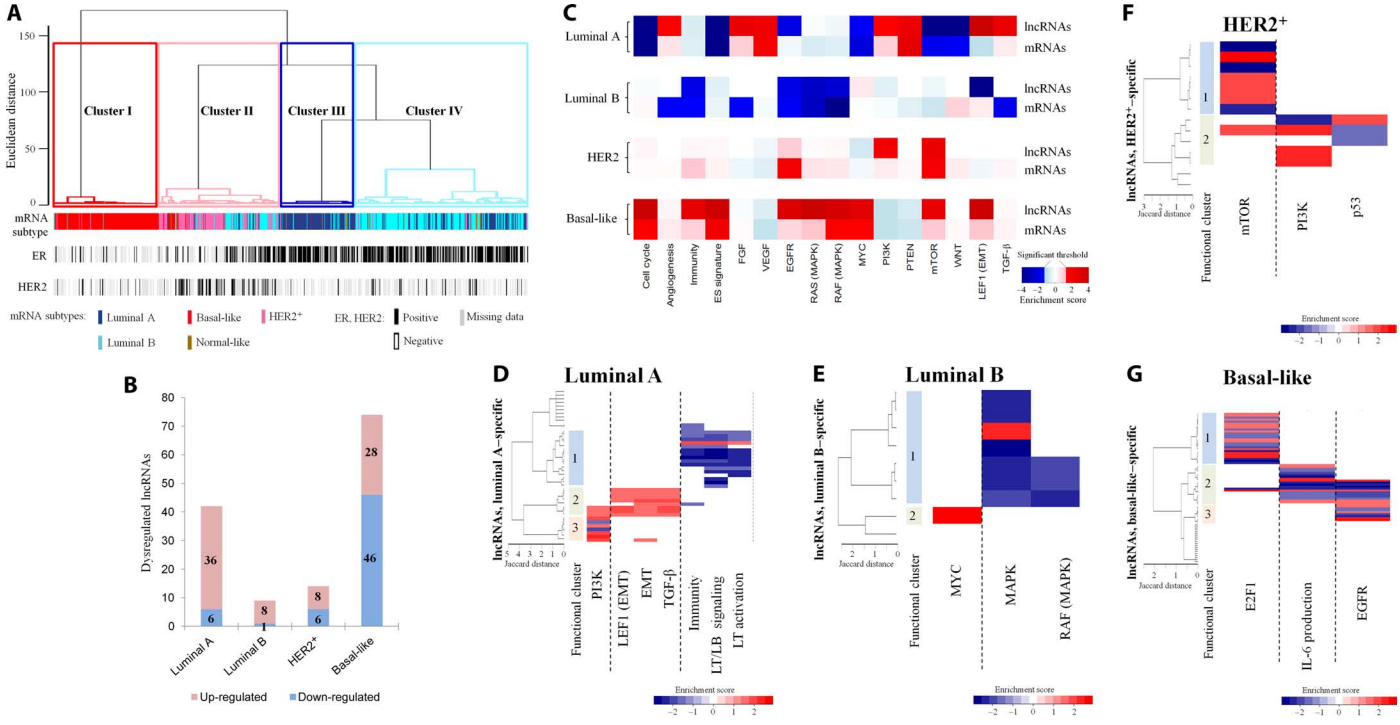


Fig. 3. Identification of four lncRNA-related clusters correlating with the known molecular subtypes and enriched in specific functional terms. (**A**) Dendrogram of the 823 breast tumors obtained by consensus hierarchical clustering according to the levels of the 500 most variant dysregulated lncRNAs, revealing four groups of tumors (clusters I to IV). How these clusters relate to the mRNA-based breast cancer subtype (based on PAM50) is also shown. (**B**) Histogram illustrating the number of specific lncRNAs in each molecular subtype of breast cancer. (**C**) Heat map illustrating the pathways whose activation or inhibition correlates with levels of subtype-specific lncRNAs and mRNAs. To relate subtype-specific lncRNAs to gene sets, an enrichment metascore was computed for each gene set. The *P* value of the metascore was defined as the proportion of random metascores being at least as high (low) as the metascore of the positively (negatively) subtype-specific lncRNAs. Conventional GSEA analysis was used to analyze specific enrichment in mRNAs, comparing one subtype to the three others. The significance score was defined as the log of the *P* value, adjusted by the sign of the enrichment metascore. (**D** to **G**) Heat maps illustrating the enrichment scores of subtype-specific lncRNAs for representative gene sets in the (D) luminal A, (E) luminal B, (F) $HER2^+$, and (G) basal-like subtypes. A positive (negative) score is associated with the activation (repression) of the gene set.

lncRNAs to those predicted for subtype-specific mRNAs (Fig. 3C). We observed good agreement between the two sets of predicted functions. In the basal-like subtype, for example, we found both predicted lncRNAs and mRNAs to be associated with cell cycle activation and the RAF [mitogen-activated protein kinase (MAPK)] and MYC pathways; yet, there were cases where subtype-specific lncRNAs appeared more strongly associated with a function or pathway than subtype-specific mRNAs. This was true for angiogenesis stimulation and activation of phosphatidylinositol 3-kinase (PI3K), fibroblast growth factor (FGF), and transforming growth factor–β (TGF-β) pathways in the luminal A subtype and for activation of epidermal growth factor receptor (EGFR) and lymphoid enhancer binding factor 1 (LEF1) pathways in the basal-like subtype.

We then took a closer look at predicted functions of subtype-specific lncRNAs. Within each subtype-specific lncRNA signature, we first performed hierarchical clustering of the lncRNAs on the basis of their coexpression with coding mRNAs and then examined which biological functions or molecular pathways appeared overrepresented in each group (Fig. 3, D to G). We observed three functional groups for luminal A–specific lncRNAs: lncRNAs related to the PI3K pathway, lncRNAs associated with the epithelial-to-mesenchymal transition (EMT), and lncRNAs related to immunity (Fig. 3D). Hierarchical clustering of the nine lncRNAs that compose the luminal B signature identified one main lncRNA group related to the MAPK pathway, including a group of lncRNAs enriched for the RAF pathway and one additional lncRNA associated with the activation of the MYC pathway (Fig. 3E). We identified two functional groups of $HER2^+$-specific lncRNAs: one related to the mTOR pathway and one related to the PI3K pathway. This second functional cluster encompasses lncRNAs related to the p53 pathway, known to be influenced by the PI3K pathway in the context of $HER2^+$ tumors (Fig. 3F) (*31*). Three groups of basal-like–specific lncRNAs were identified: group 1, which contains lncRNAs whose genes are coexpressed with E2F1 target genes; group 2, which is related to interleukin-6 (IL-6) production and partly to the EGFR pathway [two pathways known to be interconnected (*32*)]; and group 3, which is more specifically related to the EGFR pathway (Fig. 3G).

Together, our analyses highlight important putative functions for subgroups of subtype-specific lncRNA genes whose expression correlates tightly with that of cancer-related coding genes. Moreover, some groups of subtype-specific lncRNAs are involved in common pathways (that is, the PI3K pathway), whereas others appear more specialized (that is, group 3 in the basal-like subtype, which is related to E2F1). Our results show that some lncRNAs seem to be more significantly regulated in many pathways than mRNAs. This illustrates the importance of lncRNAs in breast cancer.

Dysregulated lncRNAs are markers of breast cancer clinical outcome

Like protein-coding RNAs, several lncRNAs have been linked to clinical outcome in different diseases (*33*). We thus used univariate Cox regression models to assess whether expression levels of lncRNAs might correlate with relapse-free survival, using our discovery data set with long follow-up (median follow-up, 6.75 years). No less than 300 lncRNAs emerged as significantly prognostic markers of risk of relapse (table S8), including 41 lncRNAs dysregulated in breast cancer, as compared to healthy tissues. Next, we performed multivariate Cox analysis to examine the possible impact of confounding factors known to affect prognosis (size, node, grade, ER, and HER2). The expression of 210 lncRNAs appeared to be significant risk-of-relapse predictors in this analysis, suggesting that these lncRNAs could be independent prognostic factors (Fig. 4A and table S9). Most of them are novel survival markers in breast cancer, including the newly identified RP11-863K10.2, which has the highest hazard ratio (HR) in our study (HR, 8.5; $P = 0.007$) (Fig. 4, B and C), and LINC00152, which we called CYTOR [following the HUGO Gene Nomenclature Committee guidelines (*34*)] (HR, 1.42; $P = 0.012$) (Fig. 4, D and E), an lncRNA up-regulated in all subtypes of breast cancer and recently revealed as a marker of gastric cancer (*35*). As shown in Fig. 4, Kaplan-Meier curves highlighted significant differences in relapse-free survival between patients whose tumors showed high and low levels of certain lncRNAs.

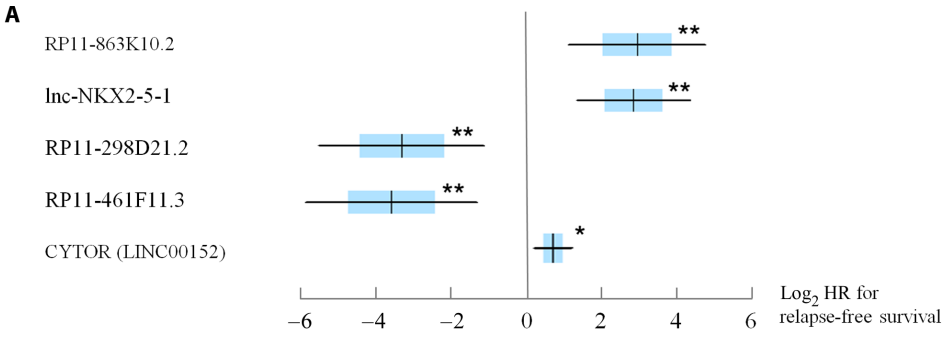
Twenty-seven of the 210 prognostic lncRNAs were also dysregulated in breast cancer, as compared to normal tissues, suggesting their importance in terms of biological relevance. We sought to validate this set of 27 dysregulated and prognostic lncRNAs using a metagene-based approach. As shown in fig. S4, we observed similar Cox HR values for the prognostic lncRNAs in the discovery cohort (HR, 1.23; $P = 1.9 \times 10^{-6}$) and in an independent validation cohort composed of 327 samples with a median follow-up of 9.2 years (HR, 1.22; $P = 2.2 \times 10^{-5}$; validation P value = 0.0029). Together, the above results suggest that a set of 27 dysregulated lncRNAs might be used as novel independent markers reliably predicting the risk of relapse in breast cancer.

CYTOR is required for cell proliferation, cell migration, and cytoskeletal organization

We selected CYTOR (previously known as LINC00152) for further functional characterization and experimental validation of the guilt-by-association prediction. CYTOR is an intergenic lncRNA located more than 100 kb away from the nearest protein coding (PLGLB2) in the chromosome 2p11.2. We chose this lncRNA for the following reasons: (i) CYTOR may be a general tumor marker because it is up-regulated in all subtypes of breast cancer (Fig. 5A) and in other types of cancer such as thyroid, stomach, lung, renal, and liver cancer (*36*); (ii) CYTOR is prognostic of relapse in both our discovery and validation cohorts; (iii) our guilt-by-association analysis highlighted an association between lncRNA and key cancer-related pathways such as cell proliferation, cell migration, and EMT and the EGFR, mTOR, and MAPK pathways (Table 1).

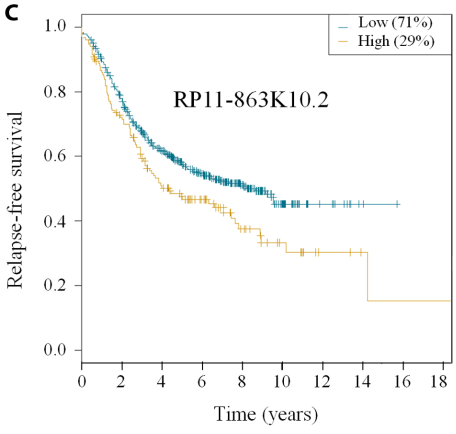
In addition to this, the transcription start site of CYTOR appears associated with H3K27 acetylation, H3K4 trimethylation, and weak H3K4 monomethylation marks in different breast cancer cell lines [MDA-MB-231 (Fig. 5B), HMEC, and MCF-7 (fig. S5A)], suggesting that this intergenic lncRNA is transcribed from promoter-like elements (*37*). The promoter DNA methylation profile of CYTOR in tumors and various breast cell lines suggests that it may be regulated by DNA methylation in breast tumors (fig. S5, B and C).

To assess the function of CYTOR, we used locked nucleic acid (LNA) gapmers to efficiently knock down CYTOR in MDA-MB-231 breast cancer cells (fig. S6A), wherein the basal level of CYTOR is high. Silencing of CYTOR resulted in a significant decrease in cell proliferation, as assessed with xCELLigence technology (Fig. 5C). In 5-bromo-2′-deoxyuridine (BrdU)/7-aminoactinomycin D (7-AAD) flow cytometry experiments, CYTOR knockdown cells accumulated in the G_2/M phase, at the expense of the S phase, confirming cell cycle inhibition (fig. S6, B and C). Thus, CYTOR is required for normal proliferation and cell cycle progression. We then evaluated the effect of CYTOR knockdown on cell migration, another process predicted to be affected by the guilt-by-association analysis. Cell migration kinetics was recorded with the xCELLigence system, using fetal bovine serum (FBS) as a chemoattractant.



B **RP11-863K10.2**

Variable (log_2)	HR	Lower 95% CI	Upper 95% CI	*P* value
Size	0.26	−0.36	0.88	0.414
Node	0.98	0.39	1.57	0.001
Grade	0.51	0.07	0.96	0.022
ER	−0.79	−1.37	−0.21	0.007
HER2	−0.04	−0.72	0.63	0.904
Expression	3.09	0.84	5.33	**0.007**



D **CYTOR**

Variable (log_2)	HR	Lower 95% CI	Upper 95% CI	*P* value
Size	0.38	−0.23	1.01	0.226
Node	0.96	0.37	1.54	0.001
Grade	0.47	0.03	0.91	0.036
ER	−0.75	−1.34	−0.17	0.01
HER2	−0.13	−0.81	0.54	0.692
Expression	0.51	0.11	0.91	**0.012**

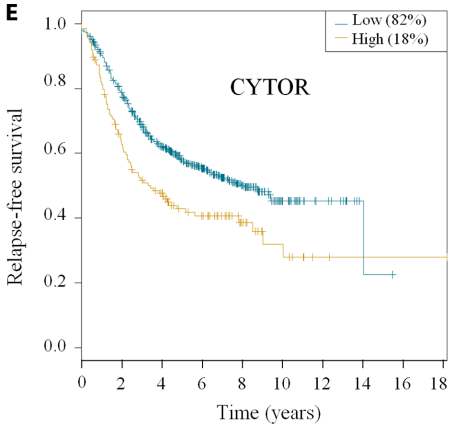


Fig. 4. lncRNAs in breast cancer have prognostic value. (**A**) Forest plot showing the log_2 HR with the SD (blue boxes) and the 95% confidence interval (bars) of the relapse-free survival analysis (multivariate Cox analysis). A negative HR reveals that a high lncRNA level is associated with a good outcome, and conversely. For example, five lncRNAs significantly related to relapse are shown. *P < 0.05, **P < 0.01. (**B**) Multivariate analysis with all the classical markers used clinically. RP11-863K10.2 is used as an example (see table S9 for the complete analysis). CI, confidence interval. (**C**) Exemplative Kaplan-Meier curves for RP11-863K10.2 (P = 0.01, log-rank test). (**D**) Same as in (B) for CYTOR. (**E**) Same as in (C) for CYTOR (P = 0.000776, log-rank test).

Without FBS in the lower chamber medium, cells did not migrate. With FBS, control cells migrated, whereas cells depleted for CYTOR did not (Fig. 5D).

We next examined the possible role of CYTOR in controlling gene expression. In gene microarray experiments on control and CYTOR-knockdown MDA-MB-231 cells, we found 2 genes that are up-regulated and 79 genes that are down-regulated in the latter (fold change > 1.5; FDR < 0.05) (Fig. 5E), suggesting a rather positive role for this lncRNA in gene regulation. Some of the identified targets are involved in breast cancer [*KIF14* (*38*) and *GOLPH3* (*39*)], some are involved in key signaling pathways (for example, *mTOR*, *PRKCA*, and *IGFR2*), and some are involved in cytoskeleton remodeling [for example, tubulin tyrosine ligase (*TTL*), Rho guanosine triphosphatase (GTPase) (*Rhobtb3*), and plakophilin-4 (*PKP4*) (*40*)]. Note that none of the genes in the large neighborhood (±1 Mb) of *CYTOR* was affected by the knockdown, suggesting that this lncRNA acts in trans. We selected five CYTOR target genes (*mTOR*, *GOLPH3*, *KIF14*, *PRKCA*, and *SMYD3*) for validation by reverse transcription quantitative polymerase chain reaction (RT-qPCR)

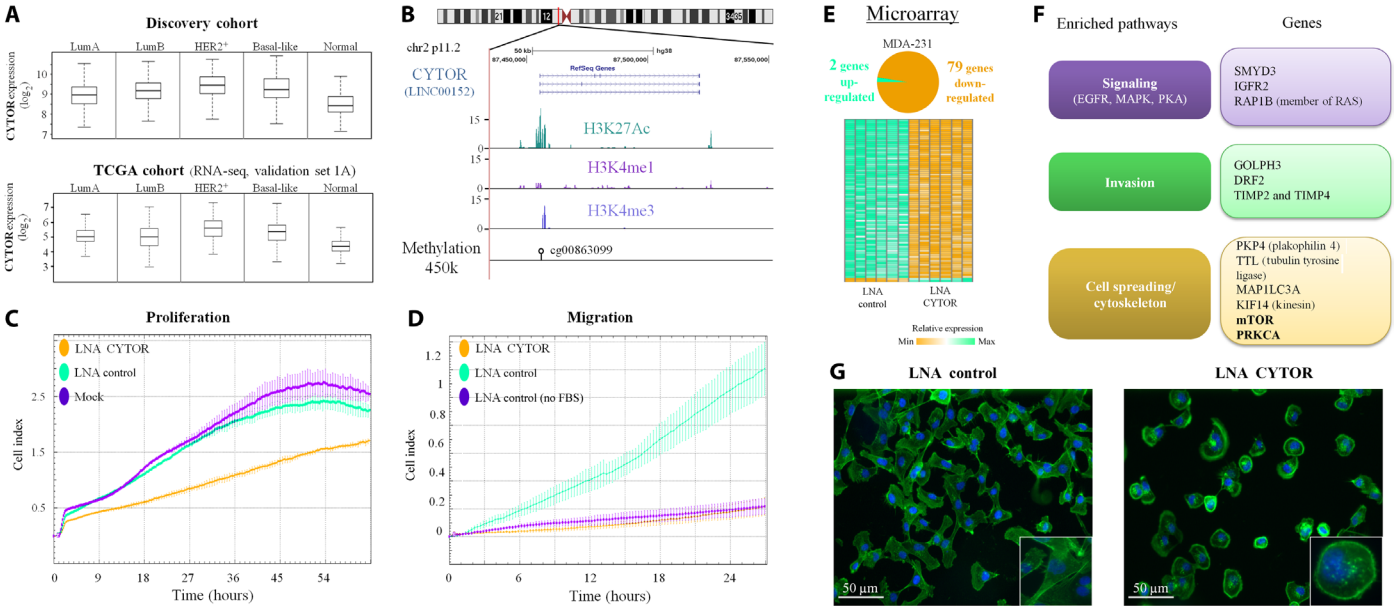


Fig. 5. CYTOR controls cell proliferation, cell migration, and cytoskeleton organization. (**A**) Expression of the gene encoding CYTOR in the discovery data set (823 tumors and 172 normal breast samples) and the TCGA RNA-seq data set GSE62944 (971 tumors and 103 normal breast samples). (**B**) UCSC (University of California, Santa Cruz) genome browser view of chromosome locus 2p11.2, which contains CYTOR. The magnified view depicts CYTOR with its associated histone marks (H3K27 acetylation, H3K4 trimethylation, and H3K4 monomethylation, from GSE49651) and methylation marks in MDA-MB-231 cells. The unfilled lollipop represents unmethylated CG dinucleotides from the Infinium HumanMethylation450. (**C**) Proliferation curve of MDA-MB-231 cells with and without LNA gapmer–mediated knockdown of CYTOR. A real-time cell analyzer (RTCA) software representative trace of a triplicate experiment is shown. (**D**) Migration kinetics of MDA-MB-231 cells assessed by continuous monitoring for approximately 24 hours. FBS (10%) in the lower chamber was used as chemoattractant, except for the control curve (purple), which represents untransfected cells with serum-free medium in the lower chamber. An RTCA software representative trace of a triplicate experiment is shown. (**E**) Pie chart and heat map showing the distribution of differentially expressed genes after CYTOR knockdown in MDA-MB-231 cells. (**F**) Enriched pathways (left) from Ingenuity Pathway Analysis in CYTOR knockdown cells. Representative differentially expressed genes of enriched gene sets are shown on the right (see also fig. S6E.) (**G**) Staining of F-actin with Acti-stain 488 fluorescent phalloidin (green) and of DNA with 4′,6-diamidino-2-phenylindole (blue) reveals reorganization of the actin cytoskeleton in MDA-MB-231 cells transfected with an LNA gapmer against CYTOR (right), as compared to cells transfected with a control LNA gapmer (left). Images captured with a 40× objective are shown.

after CYTOR knockdown. The results confirmed changes in their transcript-level expression (fig. S6D). Furthermore, Ingenuity Pathway Analysis of the targets revealed significant overrepresentation of pathways related to EGFR and MAPK signaling (Fig. 5F and fig. S6E), as predicted by our guilt-by-association analysis (Table 1). The most overrepresented term was cell spreading, a process driven by actin polymerization and promoted by the Rho family of GTPase and the mTORC2 complex (*41*, *42*). Therefore, we stained the filamentous actin (F-actin) cytoskeleton with Acti-stain 488 fluorescent phalloidin in cells treated with the LNA gapmer targeting CYTOR, or with LNA control cells. The former cells appeared smaller and rounder than the latter, and global reorganization of actin fibers was observed in these cells, with fewer stress fibers and thick actin fibers present mainly on the cell cortex (Fig. 5G).

Overall, these results indicate that CYTOR, which correlates with poor outcome, is required for breast cancer cell growth, migration, and normal morphology. They also show that CYTOR may act in trans to control genes involved in the mTOR pathway. It may thus be a good candidate target for new therapeutic approaches.

DISCUSSION

Although previous studies have demonstrated the involvement of lncRNAs in breast cancer pathology (*14*, *15*), the functions they exert in breast cancer development remain poorly understood. The microarray approach we used here is limited to detecting lncRNAs that are known and present on the Affymetrix U133 Plus design. However, while representing a fraction of lncRNAs encoded by the human genome, our integrative analysis also brings significant insights and advances over previous studies because it provides the most comprehensive data sets so far for breast cancers, a resource of clinically relevant lncRNAs, and a potential lncRNA function in the breast cancer context and uncovers their utility in prognosis. It is worth adding that the clinical information available from the array-based expression profiles are more extensive than publicly available data from the TCGA RNA-seq experiment, especially in terms of median follow-up [median follow-up of 6.75 years for the discovery cohort and 9.2 years for the validation cohort, in comparison to 1.08 years for the biggest RNA-seq cohort publicly available from the TCGA consortium (*4*)]. Our work provides an important foundation for the potential function and clinical relevance of lncRNA in breast cancers, and future work would certainly be valuable to extend our results to the entire set of lncRNAs, notably by RNA-seq.

Here, we show that 215 lncRNAs are dysregulated in breast cancers. Using (i) already processed TCGA RNA-seq data, hence easily usable, and (ii) two independent microarray cohorts, we could validate our discovery signature on data profiled by two different technologies. We also demonstrated an association between lncRNAs and clinical features and relapse. In terms of function, these lncRNAs appear to be associated with

Table 1. Predicted functions of CYTOR. A selected gene set significantly associated with CYTOR by the guilt-by-association analysis (see also table S5).

Gene set	Score
Epithelial-mesenchymal transition (HALLMARK_EMT)	2.75
Proliferation (BENPORATH_PROLIFERATION)	2.44
EGFR signaling (EGFR_UP.V1)	2.33
MAPK signaling (MAPK CASCADE.POS)	2.25
RAS (MAPK) signaling (RAS.POS)	2.14
mTORC signaling (HALLMARK_MTORC_SIGNAL.)	2.12
RAF (MAPK) signaling (RAF_UP.V1.POS)	2.12
mTOR signaling (MTOR_UP.N4.V1.POS)	1.99
Migration (CELL_MIGRATION.POS)	1.92

numerous key molecular processes, including the EGFR, PI3K, MAPK, and E2F1 pathways. In addition, we find that CYTOR, an lncRNA that is dysregulated in breast cancer and is associated with bad outcome, is essential to the proliferation, migration, and normal morphology of breast cancer cells.

Our in-depth transcriptomic analysis has revealed that lncRNA profiling of breast tumors distinguishes ER^+ from ER^- tumors and allows stratification into different molecular subtypes, in agreement with findings of recent studies (*14*, *15*). By RNA-seq in breast cancer cell lines, two recent studies identified lncRNAs regulated by the ER in the presence or absence of an estrogen agonist (*43*, *44*). However, we found only two lncRNAs (DSCAM-AS1 and RP11-161M6.2) that overlap with our ER-related signature and also between these two studies themselves, suggesting differential mechanism between cell lines and tissues and/or specificities related to the method/platform used.

In addition, we provide validated ER status– and subtype-specific lncRNA signatures and highlight possible roles of lncRNAs in ER- and subtype-related pathways. We report that many dysregulated lncRNAs are related to processes or pathways that play key roles in breast cancer development, such as immunity and the MAPK, PI3K, and mTOR pathways. A few of the functions predicted here have been validated by previous studies, suggesting that our guilt-by-association approach is effective. For example, *PVT1* and *MINCR*, two lncRNAs predicted here to play a role in the MYC pathway, were recently found to control MYC activity in mammary tumor and in Burkitt's lymphoma, respectively (*45*, *46*).

Notably, we show that the lncRNAs that compose our ER-related signature can be subdivided into three clusters, which predicted to influence various biological functions previously described in ER signaling. For instance, the cluster I lncRNAs appear associated with the mTOR pathway, which drives cell growth and promotes survival. Hyperactivation of this pathway is involved in the development of ER^+ breast cancer and in resistance to endocrine therapy (*47*). Further characterization of the lncRNAs involved in this pathway may be of interest in the search for potential new therapies. Here, we demonstrate an association of ER-related cluster II lncRNAs, most of which are up-regulated in ER-negative tumors, with activation of E2F1 signaling. Regulation of E2F1 by ER signaling is reported to mediate resistance to hormone therapy (*48*). Moreover, the *E2F1* gene appears more highly expressed in ER tissues, and high levels of E2F1 transcript correlate with an unfavorable outcome (*49*). In cluster III, we highlight lncRNAs that may play roles related to the immune response. Although the link between ER and the immune response in breast cancer remains unclear, some studies suggest that ER plays an immunosuppressive role (*50*, *51*). Furthermore, 9 of the 17 lncRNAs in cluster III are associated with the production of IL-6, a potential regulator of normal and tumor stem cell self-renewal. Abnormally high IL-6 levels seen in basal-like breast tumors (ER-negative) are associated with EMT and with poor clinical outcome (*52*, *53*). Our present results reveal a potential link between lncRNAs, IL-6 production, and immunity in ER-negative tumors.

We have taken a closer look at the identified molecular subtype–specific lncRNAs, some of which have already been described in breast cancer. For example, H19, the first identified imprinted lncRNA (*54*), is up-regulated in luminal A samples in both our discovery and validation cohorts. These results agree with in situ hybridization data that show overexpression of the H19 gene in ER-positive tumors and with the finding that estradiol transcriptionally regulates H19 (*55*). Although H19 and its involvement in breast cancer have been extensively studied, its function remains unclear because it is reported to both promote and suppress metastasis (*56*). The results of our guilt-by-association study highlight an association of this lncRNA with the TGF-β pathway and with the EMT, as already described in several studies (*57*, *58*). Our analysis also links H19 to epigenetic proteins such as SETD7, KDM1, and EZH1 and provides a new context for further characterization of H19. In conjunction with Chen *et al.* (*59*), we observed that most of our basal-like–specific lncRNAs are down-regulated and also identified LINC00993 as the most down-regulated lncRNA in the basal-like subtype. Our last example involved HOTAIR; in agreement with a recent study that suggests that HOTAIR may be related to the $HER2^+$ subtype (*14*), we find it to be the most up-regulated lncRNA of the $HER2^+$ signature. In support of a link between HER2 and HOTAIR, the latter has recently been reported as an $HER2^+$ regulator in gastric cancer (*60*). We also predicted a positive association between HOTAIR and the PI3K pathway; this relation was recently reported in other cancers (*61*, *62*), but the mechanisms involved are still poorly understood. Therefore, further studies are needed to better understand HOTAIR/PI3K signaling transduction. Note that, in addition to linking HOTAIR to the $HER2^+$ and PI3K pathways, our analysis also links HOTAIR to endocrine therapy resistance and to other processes, such as histone modification by EHMT1 (euchromatic histone lysine *N*-methyltransferase 1) or CREBP (cAMP response element–binding protein). These links should be further explored.

Also noteworthy is our observation that subtype-specific lncRNAs and subtype-specific mRNAs are globally predicted to activate or inhibit the same pathways. However, some specificities appear. For example, lncRNAs specific to the basal-like subtype seem particularly involved in the activation of the EGFR pathway. Amplification of the EGFR gene is common in basal-like tumors (*4*). Here, we provide evidence for an interesting correlation between EGFR gene amplification and lncRNAs associated with the EGFR pathway. For instance, the gene *TPT1-AS1*, which is associated with EGFR pathway activation, is more expressed in EGFR-amplified tumors, whereas *DRAIC*, which is associated with EGFR pathway inhibition, is less expressed in such tumors. This suggests that these EGFR signaling–associated lncRNAs may be regulated by somatic copy number alteration of the EGFR. Overall, we show that our prediction method provides results in keeping with

known functions of lncRNAs. This lends weight to the hypotheses it has enabled us to generate regarding new functions of dysregulated lncRNAs in breast cancer. Confirming these hypotheses may provide a strong basis for further functional studies.

To test the reliability of the guilt-by-association approach, we selected CYTOR for validation. The results of our functional assays confirm the predicted involvement of CYTOR in cell proliferation and migration and its link to the EGFR and mTOR pathways. These findings support recent studies that show that CYTOR promotes proliferation (i) in gastric cancer, through the EGFR-dependent pathway (*63*), and (ii) in hepatocellular carcinoma, through the mTOR signaling pathway (*64*). Moreover, we reveal a drastic reorganization of F-actin upon CYTOR knockdown. We propose a model wherein CYTOR ultimately influences F-actin organization by regulating *GOLPH3* expression, which in turn regulates cell size (*65*), and affects the mTORC2 complex by controlling *mTOR* and *PRKCA* expression. Alternatively, CYTOR could bind directly to the mTORC2 complex or to modulators of the cytoskeleton. Overall, our data indicate that CYTOR may be involved in breast cancer development by playing an essential role in cell proliferation, migration, and morphology. It thus appears as a potential target for future therapies. Further research is required to better understand the mechanisms underlying the involvement of CYTOR in breast cancer.

Besides providing a basis for studying functions of dysregulated lncRNAs in breast cancer, we also identified and validated a set of 27 lncRNAs predictive of relapse, using multivariate Cox analysis. Whereas most of these markers are novel, some have already been shown in other studies to have prognostic value. Regarding HOTAIR, which was reported to be a significant predictor of metastasis and death (*7*), this lncRNA does not reach statistical relevance in our data set, suggesting that other lncRNAs identified here could be more informative with regard to the probability of relapse. Furthermore, a recent paper demonstrates that HOTAIR has prognostic value in ER-negative patients only (*66*). This may explain why HOTAIR does not emerge as related to relapse from our analysis. Globally, these 27 lncRNAs might have clinical use as molecular diagnostic markers for identifying patients with low risk of relapse and who do not need aggressive therapy. We further compared our list of relapse-associated lncRNAs with recently published survey of prognostic lncRNAs in breast tumor. Sun *et al.* (*67*) identified nine lncRNAs associated with metastasis in breast cancer, of which four are present in the Affymetrix U133 Plus 2.0 array. However, none of them were significantly associated with relapse in our uni- or multivariate Cox analysis. Another study identified a set of 45 lncRNAs prognostic of metastasis in lymph node–negative breast cancer (*68*); this said, we found only a small overlap with our results (five and two, respectively, in common with our uni- and multivariate analysis). This suggests that additional efforts should be done to robustly identify sets of prognostic lncRNAs.

Overall, our study provides an in-depth analysis of the lncRNA transcriptome in breast cancer and provides numerous new lncRNA markers associated with ER status, tumor subtype, and clinical outcome. We have inferred functions of these dysregulated lncRNAs, and we demonstrate for the first time that lncRNAs might contribute to the dysregulation of nearly every known breast cancer pathway. These data lay the ground for future studies that address the biological mechanisms involving these lncRNAs and their use as diagnostic markers and therapeutic targets. We have experimentally confirmed the predicted function of one such dysregulated lncRNA, concluding that our integrative approach is effective. These findings should contribute to a better understanding of the mechanisms of action of lncRNAs in breast cancer.

MATERIALS AND METHODS

Experimental design

The goal of this study was to analyze the transcriptome of breast cancer in order to identify aberrantly expressed lncRNAs and to infered their functions. We began by downloading publicly available microarray data and extracted information on lncRNA expression. We used a specific method to identify lncRNAs that were dysregulated in small subset of breast tumors and infer their functions through a guilt-by-association approach. We have investigated lncRNAs as survival markers. Finally, we have characterized the functions of one breast cancer–related lncRNA, CYTOR, by means of in vitro experiment. The detailed procedure is described below.

Breast cancer gene expression data and reannotation of the Affymetrix microarray

To obtain a genome-wide view of lncRNA expression in breast cancer, we reannotated the entire collection of probe sets of the Affymetrix Human Genome U133 Plus 2.0 array. We downloaded the microarray (U133 Plus 2.0 Affymetrix) gene expression data sets GSE9195, GSE10780, GSE10810, GSE12276, GSE19615, GSE20711, and GSE21653 from GEO (http://ncbi.nlm.nih.gov/geo/). The raw CEL files were frozen robust multiarray analysis (fRMA)–normalized in the R environment using the limma and fRMA libraries to obtain $\log_2$-normalized expression signals for each probe set. We then applied the ComBat algorithm of the sva library with default parameters to adjust the data for batch effects, using tumor versus normal tissue as covariate. Hybridization probe sets were locally mapped by sequence alignment (National Center for Biotechnology Information BLAST 2.2.29+) against a reference transcriptome in LNCipedia database version 2.1 (*69*), a database dedicated to lncRNAs and Ensembl 84 transcriptome. We required that at least 80% of a probe set should hit a given transcript sequence. For the lncRNAs, we kept probes that target lncRNAs present in the LNCipedia database. We then discarded probes that were discordant between LNCipedia and Ensembl, in terms of transcript biotype. We also excluded probe sets that target multiple genes, except if the target was a duplicated lncRNA (that is, corresponding to a duplicated region of the genome). To identify these duplicated lncRNAs, we blasted all lncRNA transcripts against the LNCipedia database, where transcripts were defined as duplicated if the smallest transcript shared at least 95% of its sequence with the other. Because these duplicated lncRNAs could not be distinguished from each other, they were analyzed as arising from a single gene and the tag "multi" was added to the name of one of them. Alternative transcripts were considered to be from the same gene. When multiple probe sets mapped to the same gene (corresponding to an lncRNA or an mRNA), the one with the highest variance was selected. We computed a full annotation table for the 3053 lncRNAs, comprising their corresponding names in the Ensembl database (if available), their genomic location, their category, and the probe sets that matched each lncRNA (table S2).

Normal breast tissues are composed of a mixture of different cell types, mainly including epithelial cells and adipocytes, whereas breast tumors are composed mostly of epithelial cells. Because we focused on lncRNA genes differentially expressed between normal and cancerous

breast epithelial cells, we excluded six lncRNA genes whose expression correlated strongly (Pearson $R > 0.6$) with adipose markers (*70*) in normal breast tissues, suggesting that they might be expressed in adipocytes rather than in epithelial cells.

To assess potential remaining batch effects in the breast cancer data sets, we performed unsupervised hierarchical clustering of the top 500 most variable genes (coding and noncoding). None of the breast samples clustered according to the data set they came from, indicating that no strong batch effect was present.

For the validation steps, the fRMA-normalized data set GSE20685 was downloaded from InSilico DB (www.insilicodb.com). We also acquired raw data from the TCGA consortium. GSE20685 contains no normal samples and was therefore not used to validate the dysregulated lncRNAs in breast cancer.

Clinical data and molecular subtype prediction

The clinical data were downloaded from GEO and were merged (table S1). In the original studies, the ER and HER2 statuses were determined by IHC. Intrinsic mRNA-based breast cancer subtypes were determined with the 50-gene PAM50 predictor (*71*).

TCGA RNA-seq analysis

Because the raw RNA-seq data (level 1) from TCGA are challenging to analyze, we chose to use the TCGA data reprocessed by Rahman *et al.* (*23*) (GSE62944) that allowed to assess the expression of 1241 lncRNAs, of which 804 were present on the Affymetrix microarray. We first selected data from breast tissues and then applied the voom transformation from the limma package (*72*, *73*) on the transcripts per million expression values to obtain $\log_2$ expression data.

TCGA microarray reannotation

TCGA raw data were processed as previously described (*4*). Probes of the TCGA microarray were mapped to the LNCipedia database using the TCGA annotation file. Briefly, coordinates targeted by TCGA microarray probes were first extracted from the annotation file "AnotAgilentG4502A_07_3.adf" available on the TCGA Web site and converted to hg19 genome build, using liftOver (UCSC). We selected probes where at least 58 base pairs (bp) of the targeted region overlapped, in a strand-specific way, with exons of lncRNA transcripts of the LNCipedia version 2.1 database. Because the boundaries of exons are not always clearly defined, we added 5 bp on both sides of each lncRNA exon. In parallel, probes targeting protein-coding regions were extracted similarly with Ensembl version 72 and RefSeq version 58. Finally, probes targeting lncRNAs were only kept for further analysis.

Clustering

For hierarchical clustering, we used the ward.D aggregation method of the hclust algorithm in R, with the Euclidean distance as the dissimilarity measure. To maximize the robustness of the clustering toward overfitting, we ran the ConsensusClusterPlus algorithm using 1000 repetitions with subsets obtained by sampling 80% of the samples [proportion of items (pI), 0.8] and keeping all the features [proportion of features (pF), 1]. Both inner and final clustering included in this method were realized with ward.D linkage and Euclidean distance.

Identification of dysregulated lncRNAs

Because breast tumors are highly heterogeneous, classical *t* tests cannot identify dysregulated lncRNAs in a small subset of breast tumors. Therefore, the method was adapted to identify lncRNAs that were dysregulated in at least 10% of the breast tumors (*22*). This ensured selection of a reasonable number of dysregulated lncRNAs with a potentially informative variance related to the heterogeneity of breast cancer. To find significantly up-regulated genes, we explored the upper tail of the expression distribution by selecting the 10% of samples (normal or cancerous) with the highest expression. We then used a nonparametric Mann-Whitney test to compare expression levels between the normal samples and tumor samples. We extended this approach to a growing proportion of samples (15, 20, 25, 30, 35, 40, and 45%), and all lncRNA genes additionally found to be differentially expressed were retained. In parallel, we identified down-regulated lncRNAs by applying the same methodology to the lower tail of the expression distribution. The 215 lncRNA genes identified as dysregulated had an FDR of <0.05 and a fold change of >1.5 in at least 10% of the samples.

Prediction of lncRNA functions by the guilt-by-association approach

This approach is based on establishing correlations between the expression of lncRNA genes and that of protein-coding genes known to be involved in particular functions (gene sets). It enabled us to generate hypotheses regarding the function(s) of a given lncRNA. Publicly available gene sets were selected from both the KEGG (Kyoto Encyclopedia of Genes and Genomes) gene set [Molecular Signatures Database (MsigDB)] and the bdfunc.enrichment.human database of the sRAP library. Following the guidelines of the GSEA (Gene Set Enrichment Analysis) software, we grouped gene sets that contain redundant genes as follows: (i) we computed a between-gene-set distance matrix using the overlap distance (defined as the number of common genes/number of genes that compose the smallest gene set); (ii) we performed hierarchical clustering on the basis of this matrix (complete linkage); and (iii) we used a threshold level of 0.5 to cut the tree and group gene sets that belong to the same cluster.

We chose to focus only on the potential functions of dysregulated, ER- and subtype-specific lncRNAs because the guilt-by-association approach is computation-consuming. First, we randomly divided our breast tumor expression data into two data sets, one with 411 samples and one with 412 samples. For each data set, we computed a Pearson matrix of correlation between each lncRNA and each coding gene to produce two matrices of 236 lncRNAs × 16,951 mRNAs. In every matrix, the protein-coding genes were ranked for each lncRNA on the basis of the correlation coefficient. The GSEA software (parameters, 1000 permutations on gene sets; minimum size, 15; maximum size, 500) was used to calculate a running sum statistic, corresponding to the enrichment score, on the basis of the ranks of the investigated gene set members, relative to those of nonmembers. We thus obtained two matrices containing an enrichment score and a statistical family-wise error rate (FWER) value for each lncRNA/gene set pair (236 lncRNAs × 422 gene sets). To obtain high-confidence associations of lncRNAs with functions, we finally selected gene sets that were statistically (FWER < 0.05) associated with an lncRNA in both matrices and computed the mean of their enrichment scores.

To relate clusters of lncRNAs to gene sets, we started by computing, for each cluster, a metascore of enrichment for a gene set. This metascore was defined as the weighted sum of the enrichment scores obtained for the members of the lncRNA cluster, with the weight defined as −1 if the lncRNA was repressed in the condition of interest and as 1 if otherwise. Then, 10,000 random groups of lncRNAs of the same

size as the lncRNA cluster of interest were generated by random selection. For each group, a metascore was generated using the same weighted sum approach. Finally, the *P* value of the metascore was defined as the proportion of random metascores being at least as high (low) as the metascore of the positively (negatively) associated cluster.

Signature validation

The different signatures were validated on three cohorts using the following protocol. First, one should note that for the validation with the TCGA data, the analysis was restricted to lncRNAs common to the two platforms. This could lead to a reduction of signature size. In this case, the size of this "between-platform signature" is specified in the main text. The data were scaled to make the two data sets more comparable. For each signature, a metagene was then defined for the discovery cohort, as the first component of a principal components analysis (PCA). A receiver operating characteristic (ROC) curve was generated from that metagene, and the threshold that maximizes the Youden index was selected. A metagene was then generated from the validation data using the first component eigenvector from the discovery cohort. For visualization purposes, a ROC curve was generated using this metagene. The threshold selected for class prediction was the one obtained for the discovery cohort to avoid any bias. The prediction accuracy of the metagene was assessed using the balanced error rate (BER) metric. Finally, to evaluate the significance of signature performance, we generated 10,000 random signatures of the same size as the real between-platform signature by randomly sampling lncRNAs. For each random signature, the same process that was used for the real signature was applied. A *P* value was defined as the proportion of random signatures that show a BER lower than or equal to the real signature, and signatures with a *P* value lower than 0.05 were considered significant.

Determination of subtype-specific lncRNA signatures

For each subtype, we first selected lncRNA genes that show significant differential expression (FDR < 0.05; fold change > 1.5, *t* test) in one particular subtype versus the three others. We then filtered out lncRNAs that were up-regulated (or down-regulated) in more than one subtype to obtain a list of lncRNAs characteristic of each subtype.

Relapse-free survival analysis

The analysis was computed in R using the "survival" library. The prognostic value of individual lncRNAs was estimated by univariate Cox regression. In parallel, multivariate Cox regression was used to test the independent prognostic values of lncRNAs, using clinical properties as covariates. Univariate analysis was used to select covariates that were prognostic. The proportional hazard assumption was verified with the "cox.zph" function (threshold, 0.01). For all analyses, $P < 0.05$ was the criterion of statistical significance. For visualization purposes, we also generated Kaplan-Meier curves using the "survfit" function. To define a high- and low-level group, we assessed all possible thresholds that lead to groups that represent at least 10% of the data set using the log-rank test. The threshold that leads to the lowest *P* value was selected as the final cutoff for group definition.

A set of 27 lncRNAs associated to relapse-free survival has been validated using a metagene-based approach: a metagene was defined for the discovery cohort as the first component of a PCA. The metagene was then generated from the validation data, using the PCA eigenvector from the discovery cohort. The association of this metagene with relapse-free survival was assessed on the validation cohort using a Cox model. The significance of this association can be visualized using forest plot representation. Because random signatures can be associated to survival (*74*), we assessed whether the association of our metagene on the validation cohort is significantly better than metagenes obtained from randomly selected lncRNAs. Therefore, we generated 10,000 random signatures of the same size as the real relapse-free associated signature by randomly sampling lncRNAs. For each random signature, the same process of computing a Cox model from the first eigenvector from the discovery cohort was applied. Finally, we called "validation *P* value" the proportion of random signatures that show a Cox *P* value lower than or equal to the real signature. The signature was considered as significant if the validation *P* value was lower than 0.05.

Culture of breast cell lines and silencing of target lncRNAs

MDA-MB-231 cells were grown in Dulbecco's modified Eagle's medium (Gibco) supplemented with 10% FBS (Gibco). They were maintained at 37°C in 5% CO_2. To silence target lncRNAs, cells were treated with LNA GapmeRs (Exiqon), according to the manufacturer's instructions. Briefly, 300,000 cells were transfected in six-well plates with 30 nM LNA GapmeRs and 5 μl of Lipofectamine 2000 in 2-ml total volume and were incubated for 24 hours. Staining of F-actin was performed following the manufacturer's protocol (http://cytoskeleton.com).

RNA purification and RT-qPCR

RNA purification was performed with the RNeasy kit (Qiagen) according to the manufacturer's instructions. Deoxyribonuclease (DNase) treatment was performed with a DNA-free DNase kit (Ambion) according to the manufacturer's protocol. qPCRs were performed with SYBR Green dye (Eurogentec) in LightCycler 480 (Roche). Briefly, complementary DNA was obtained by reverse transcription of 1 μg of RNA, with random hexamers (Amersham/Pharmacia Biotech) and SuperScript II Reverse Transcriptase (Life Technologies Inc.). The results were normalized against the following housekeeping genes: SDHA, GAPDH, and ACTIN.

The following LNA gapmer sequences were used: CYTOR LNA, 5′-TCATAGAGCTTCCTGT-3′.

The following qPCR assay primer sequences were used: CYTOR, 5′-CTGGATGGTCGCTGCTTTTT-3′ (forward) and 5′-GATCTGAAGACAGGCACGGG-3′ (reverse); SMYD3, 5′-TTACTGCGAGCAGTCCGAGACA-3′ (forward) and 5′-TTGTCCTGGGTTTGGCAACGGA-3′ (reverse); GOLPH3, 5′-CTAGAGGCTTGTGGAATGAGACG-3′ (forward) and 5′-GACCGTTTCTGGAGGCTGAGTT-3′ (reverse); KIF14, 5′-GCACTTTCGGAACAAGCAAACCA-3′ (forward) and 5′-ATGTTGCTGGCAGCGGGACTAA-3′ (reverse); mTOR, 5′-AGCATCGGATGCTTAGGAGTGG-3′ (forward) and 5′-CAGCCAGTCATCTTTGGAGACC-3′ (reverse); PRKCA, 5′-GCCTATGGCGTCCTGTTGTATG-3′ (forward) and 5′-GAAACAGCCTCCTTGGACAAGG-3′ (reverse).

Cell proliferation/migration

To evaluate breast cancer cell proliferation, MDA-MB-231 cells (10,000 per well) transfected with LNA gapmers were seeded into the xCELLigence E-Plate 16 (Roche) 24 hours after transfection, according to the manufacturer's instructions. In this system, the electrical impedance is used to derive a cell index, which gives a real-time representation of cell growth when continuously monitored. Measurements were automatically collected by the RTCA Dual Plate analyzer every 30 min for up to 3 days. The data were analyzed with the RTCA software.

To examine breast cancer cell migration, MDA-MB-231 cells transfected with LNA gapmers were seeded into the xCELLigence CIM-Plate 16 (Roche) 24 hours after transfection. Briefly, a 165-µl volume of fresh medium containing 10% FBS (chemoattractant) or serum-free medium (control) was added to the lower chambers of the CIM-Plate 16. The upper chambers were filled with serum-free medium (30 µl per well), and the plate was incubated in 5% CO_2 for 1 hour at 37°C. Cells (20,000 per well) were then added to each well of the upper chamber, and cell migration was assessed at 30-min intervals for 18 hours at 37°C in 5% CO_2. Upon migration, cells adhere to the surface of the filter electrode and increase the impedance.

Cell cycle analysis by flow cytometry

The distribution of cells through the various phases of the cell cycle was determined by measuring the DNA content with BD BrdU Flow kits (catalog no. 552598) according to the manufacturer's instructions. Briefly, cells were incubated with 10 µM BrdU for 2 hours. They were then fixed with paraformaldehyde, permeabilized with saponin, and treated with DNase to expose the BrdU epitope. BrdU was stained with fluorescent anti-BrdU antibodies. The total DNA level was assessed by staining with 7-AAD. Data were analyzed with the Kaluza Analysis Software (Beckman Coulter).

Illumina Expression HT12 arrays

Total RNA (200 ng) was amplified with the Illumina TotalPrep RNA Amplification Kit (Ambion) and then hybridized with the array according to the Whole-Genome Gene Expression Direct Hybridization Assay (Illumina). Chips were scanned with the HiScan Reader (Illumina).

The raw data were normalized using the quantile normalization method from the lumi package (*75*), and batch effect was corrected with the ComBat algorithm of the sva library with default parameters, using the CYTOR status (LNA-silenced versus control) as covariate. Then, to identify probes that are differentially expressed, a *t* test was applied, and the *P* values obtained were corrected for multitesting using the Benjamini-Hochberg method. The probes that simultaneously show an FDR of <0.05 and an absolute fold change between the median of each group higher than 1.5 were reported as differentially expressed if the median expression value in each group was significantly higher than the background level (detection *P* value < 0.05). The raw data have been uploaded in GEO database and are accessible under accession no. GSE77491.

Statistical analysis

Data values were expressed as means ± SD of at least two independent experiments and evaluated using Student's *t* test for unpaired samples, or otherwise specified. Mean differences were considered significant when $P < 0.05$, or otherwise specified.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/2/9/e1600220/DC1

fig. S1. Dysregulated lncRNAs in breast tumors.
fig. S2. Validation of the ER-related lncRNA signature in three data sets.
fig. S3. Validation of the subtype-specific lncRNA signature in three data sets.
fig. S4. Validation of the set of 27 lncRNAs predictive of relapse and dysregulated in breast cancer.
fig. S5. Methylation of the CYTOR gene in relation to its expression.
fig. S6. Impact of CYTOR depletion on cell cycle and gene expression.
table S1. Clinical annotation of the 823 primary tumors and the 172 normal tissue samples from the discovery cohort, including age, size, lymph node status, ER status, HER2 status, grade, PAM50-associated subtype, and relapse information.
table S2. Reannotation of the Affymetrix Human Genome U133 Plus 2.0 array.
table S3. Description of the 215 dysregulated lncRNAs in breast cancer.
table S4. Contingency table of the lncRNA-related clusters that correlate with ER status.
table S5. Significant enrichment scores from the guilt-by-association analysis.
table S6. Contingency table of the lncRNA-related clusters that correlate with the known molecular subtypes.
table S7. lncRNAs signatures of the known molecular subtypes of breast cancer.
table S8. Survival analysis: Univariate results.
table S9. Survival analysis: Multivariate results.

REFERENCES AND NOTES

1. J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, F. Bray, *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11* (International Agency for Research on Cancer, Lyon, 2012).
2. C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A.-L. Børresen-Dale, P. O. Brown, D. Botstein, Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
3. E. Senkus, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, E. Rutgers, S. Zackrisson, F. Cardoso, Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26** (Suppl. 5), v8–v30 (2015).
4. The. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
5. J. S. Reis-Filho, L. Pusztai, Gene expression profiling in breast cancer: Classification, prognostication, and prediction. *Lancet* **378**, 1812–1823 (2011).
6. L. Pusztai, R. Rouzier, W. F. Symmans, *CCR* 20th anniversary commentary: Divide and conquer—Breast cancer subtypes and response to therapy. *Clin. Cancer Res.* **21**, 3575–3577 (2015).
7. R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, H. Y. Chang, Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
8. J. M. Silva, N. J. Boczek, M. W. Berres, X. Ma, D. I. Smith, LSINCT5 is over expressed in breast and ovarian cancer and affects cellular proliferation. *RNA Biol.* **8**, 496–505 (2011).
9. T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, R. Guigó, The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
10. T. Gutschner, S. Diederichs, The hallmarks of cancer: A long non-coding RNA point of view. *RNA Biol.* **9**, 703–719 (2012).
11. M. Esteller, Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12**, 861–874 (2011).
12. M. Huarte, M. Guttman, D. Feldser, M. Garber, M. J. Koziol, D. Kenzelmann-Broz, A. M. Khalil, O. Zuk, I. Amit, M. Rabani, L. D. Attardi, A. Regev, E. S. Lander, T. Jacks, J. L. Rinn, A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409–419 (2010).
13. S. W. Cheetham, F. Gruhl, J. S. Mattick, M. E. Dinger, Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* **108**, 2419–2425 (2013).
14. X. Su, G. G. Malouf, Y. Chen, J. Zhang, H. Yao, V. Valero, J. N. Weinstein, J.-P. Spano, F. Meric-Bernstam, D. Khayat, F. J. Esteva, Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget* **5**, 9864–9876 (2014).
15. W. Zhao, J. Luo, S. Jiao, Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci. Rep.* **4**, 6591 (2014).
16. L. M. McIntyre, K. K. Lopiano, A. M. Morse, V. Amin, A. L. Oberg, L. J. Young, S. V. Nuzhdin, RNA-seq: Technical variability and sampling. *BMC Genomics* **12**, 293 (2011).
17. N. Raghavachari, J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, C. J. O'Donnell, P. J. Munson, G. J. Kato, A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med. Genomics* **5**, 28 (2012).
18. L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, B. Oliver, Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
19. R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
20. S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
21. M. D. Wilkerson, D. N. Hayes, ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).

22. L. Yao, H. Shen, P. W. Laird, P. J. Farnham, B. P. Berman, Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* **16**, 105 (2015).
23. M. Rahman, L. K. Jackson, W. E. Johnson, D. Y. Li, A. H. Bild, S. R. Piccolo, Alternative preprocessing of RNA-sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **31**, 3666–3672 (2015).
24. D. Chakravarty, A. Sboner, S. S. Nair, E. Giannopoulou, R. Li, S. Hennig, J. M. Mosquera, J. Pauwels, K. Park, M. Kossai, T. Y. MacDonald, J. Fontugne, N. Erho, I. A. Vergara, M. Ghadessi, E. Davicioni, R. B. Jenkins, N. Palanisamy, Z. Chen, S. Nakagawa, T. Hirose, N. H. Bander, H. Beltran, A. H. Fox, O. Elemento, M. A. Rubin, The oestrogen receptor α-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat. Commun.* **5**, 5383 (2014).
25. H. Hansji, E. Y. Leung, B. C. Baguley, G. J. Finlay, M. E. Askarian-Amiri, Keeping abreast with long non-coding RNAs in mammary gland development and breast cancer. *Front. Genet.* **5**, 379 (2014).
26. N. C. Turner, J. S. Reis-Filho, A. M. Russell, R. J. Springall, K. Ryder, D. Steele, K. Savage, C. E. Gillett, F. C. Schmitt, A. Ashworth, A. N. Tutt, BRCA1 dysfunction in sporadic basal-like breast cancer. *Oncogene* **26**, 2126–2132 (2007).
27. J. L. Rinn, H. Y. Chang, Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
28. M. F. Montenegro, M. d. M. Collado-González, M. P. Fernández-Pérez, M. B. Hammouda, L. Tolordava, M. Gamkrelidze, J. N. Rodríguez-López, Promoting E2F1-mediated apoptosis in oestrogen receptor-α-negative breast cancer cells. *BMC Cancer* **14**, 539 (2014).
29. X. Jiang, D. J. Shapiro, The immune system and inflammation in breast cancer. *Mol. Cell. Endocrinol.* **382**, 673–682 (2014).
30. C. Sweeney, P. S. Bernard, R. E. Factor, M. L. Kwan, L. A. Habel, C. P. Quesenberry, K. Shakespear, E. K. Weltzien, I. J. Stijleman, C. A. Davis, M. T. W. Ebbert, A. Castillo, L. H. Kushi, B. J. Caan, Intrinsic subtypes from PAM50 gene expression assay in a population-based breast cancer cohort: Differences by age, race, and tumor characteristics. *Cancer Epidemiol. Biomarkers Prev.* **23**, 714–724 (2014).
31. L. Zheng, J. Q. Ren, H. Li, Z. L. Kong, H. G. Zhu, Downregulation of wild-type p53 protein by HER-2/neu mediated PI3K pathway activation in human breast cancer cells: Its effect on cell proliferation and implication for therapy. *Cell Res.* **14**, 497–506 (2004).
32. S. Grivennikov, M. Karin, Autocrine IL-6 signaling: A key event in tumorigenesis? *Cancer Cell* **13**, 7–9 (2008).
33. Z. Du, T. Fei, R. G. W. Verhaak, Z. Su, Y. Zhang, M. Brown, Y. Chen, X. S. Liu, Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* **20**, 908–913 (2013).
34. M. W. Wright, A short guide to long non-coding RNA gene nomenclature. *Hum. Genomics* **8**, 7 (2014).
35. Q. Pang, J. Ge, Y. Shao, W. Sun, H. Song, T. Xia, B. Xiao, J. Guo, Increased expression of long intergenic non-coding RNA LINC00152 in gastric cancer and its clinical significance. *Tumour Biol.* **35**, 5441–5447 (2014).
36. M. K. Iyer, Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, A. Poliakov, X. Cao, S. M. Dhanasekaran, Y.-M. Wu, D. R. Robinson, D. G. Beer, F. Y. Feng, H. K. Iyer, A. M. Chinnaiyan, The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
37. A. C. Marques, J. Hughes, B. Graham, M. S. Kowalczyk, D. R. Higgs, C. P. Ponting, Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol.* **14**, R131 (2013).
38. S. M. Singel, C. Cornelius, E. Zaganjor, K. Batten, V. R. Sarode, D. L. Buckley, Y. Peng, G. B. John, H. C. Li, N. Sadeghi, W. E. Wright, L. Lum, T. W. Corson, J. W. Shay, KIF14 promotes AKT phosphorylation and contributes to chemoresistance in triple-negative breast cancer. *Neoplasia* **16**, 247–256.e2 (2014).
39. Z. Zeng, H. Lin, X. Zhao, G. Liu, X. Wang, R. Xu, K. Chen, J. Li, L. Song, Overexpression of GOLPH3 promotes proliferation and tumorigenicity in breast cancer via suppression of the FOXO1 transcription factor. *Clin. Cancer Res.* **18**, 4059–4069 (2012).
40. R. Keil, J. Schulz, M. Hatzfeld, p0071/PKP4, a multifunctional protein coordinating cell adhesion with cytoskeletal organization. *Biol. Chem.* **394**, 1005–1017 (2013).
41. T. Wakatsuki, R. B. Wysolmerski, E. L. Elson, Mechanics of cell spreading: Role of myosin II. *J. Cell Sci.* **116**, 1617–1625 (2003).
42. D. D. Sarbassov, S. M. Ali, D.-H. Kim, D. A. Guertin, R. R. Latek, H. Erdjument-Bromage, P. Tempst, D. M. Sabatini, Rictor, a novel binding partner of mTOR, defines a rapamycin-insensitive and raptor-independent pathway that regulates the cytoskeleton. *Curr. Biol.* **14**, 1296–1302 (2004).
43. P. Jonsson, C. Coarfa, F. Mesmar, T. Raz, K. Rajapakshe, J. F. Thompson, P. H. Gunaratne, C. Williams, Single-molecule sequencing reveals estrogen-regulated clinically relevant lncRNAs in breast cancer. *Mol. Endocrinol.* **29**, 1634–1645 (2015).
44. V. Miano, G. Ferrero, S. Reineri, L. Caizzi, L. Annaratone, L. Ricci, S. Cutrupi, I. Castellano, F. Cordero, M. De Bortoli, Luminal long non-coding RNAs regulated by estrogen receptor α in a ligand-independent manner show functional roles in breast cancer. *Oncotarget* **7**, 3201–3216 (2016).
45. Y.-Y. Tseng, B. S. Moriarity, W. Gong, R. Akiyama, A. Tiwari, H. Kawakami, P. Ronning, B. Reuland, K. Guenther, T. C. Beadnell, J. Essig, G. M. Otto, M. G. O'Sullivan, D. A. Largaespada, K. L. Schwertfeger, Y. Marahrens, Y. Kawakami, A. Bagchi, *PVT1* dependence in cancer with *MYC* copy-number increase. *Nature* **512**, 82–86 (2014).
46. G. Doose, A. Haake, S. H. Bernhart, C. López, S. Duggimpudi, F. Wojciech, A. K. Bergmann, A. Borkhardt, B. Burkhardt, A. Claviez, L. Dimitrova, S. Haas, J. I. Hoell, M. Hummel, D. Karsch, W. Klapper, K. Kleo, H. Kretzmer, M. Kreuz, R. Küppers, C. Lawerenz, D. Lenze, M. Loeffler, L. Mantovani-Löffler, P. Möller, G. Ott, J. Richter, M. Rohde, P. Rosenstiel, A. Rosenwald, M. Schilhabel, M. Schneider, I. Scholz, S. Stilgenbauer, H. G. Stunnenberg, M. Szczepanowski, L. Trümper, M. A. Weniger; ICGC MMML-Seq Consortium, S. Hoffmann, R. Siebert, I. Iaccarino, MINCR is a MYC-induced lncRNA able to modulate MYC's transcriptional network in Burkitt lymphoma cells. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5261–E5270 (2015).
47. E. M. Ciruelos Gil, Targeting the PI3K/AKT/mTOR pathway in estrogen receptor-positive breast cancer. *Cancer Treat. Rev.* **40**, 862–871 (2014).
48. M. C. Louie, A. McClellan, C. Siewit, L. Kawabata, Estrogen receptor regulates E2F1 expression to mediate tamoxifen resistance. *Mol. Cancer Res.* **8**, 343–352 (2010).
49. V. Vuaroqueaux, P. Urban, M. Labuhn, M. Delorenzi, P. Wirapati, C. C. Benz, R. Flury, H. Dieterich, F. Spyratos, U. Eppenberger, S. Eppenberger-Castori, Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome. *Breast Cancer Res.* **9**, R33 (2007).
50. E. M. Curran, B. M. Judy, N. A. Duru, H.-Q. Wang, L. A. Vergara, D. B. Lubahn, D. M. Estes, Estrogenic regulation of host immunity against an estrogen receptor–negative human breast cancer. *Clin. Cancer Res.* **12**, 5641–5647 (2006).
51. A. A. Mostafa, D. Codner, K. Hirasawa, Y. Komatsu, M. N. Young, V. Steimle, S. Drover, Activation of ERα signaling differentially modulates IFN-γ induced HLA-class II expression in breast cancer cells. *PLOS One* **9**, e87377 (2014).
52. N. J. Sullivan, A. K. Sasser, A. E. Axel, F. Vesuna, V. Raman, N. Ramirez, T. M. Oberyszyn, B. M. Hall, Interleukin-6 induces an epithelial–mesenchymal transition phenotype in human breast cancer cells. *Oncogene* **28**, 2940–2947 (2009).
53. Z. T. Schafer, J. S. Brugge, IL-6 involvement in epithelial cancers. *J. Clin. Invest.* **117**, 3660–3663 (2007).
54. N. Berteaux, N. Aptel, G. Cathala, C. Genton, J. Coll, A. Daccache, N. Spruyt, H. Hondermarck, T. Dugimont, J.-J. Curgy, T. Forné, E. Adriaenssens, A novel *H19* antisense RNA overexpressed in breast cancer contributes to paternal *IGF2* expression. *Mol. Cell. Biol.* **28**, 6731–6745 (2008).
55. E. Adriaenssens, L. Dumont, S. Lottin, D. Bolle, A. Leprêtre, A. Delobelle, F. Bouali, T. Dugimont, J. Coll, J.-J. Curgy, *H19* overexpression in breast adenocarcinoma stromal cells is associated with tumor values and steroid receptor status but independent of p53 and Ki-67 expression. *Am. J. Pathol.* **153**, 1597–1607 (1998).
56. L. Zhang, F. Yang, J.-h. Yuan, S.-x. Yuan, W.-p. Zhou, X.-s. Huo, D. Xu, H.-s. Bi, F. Wang, S.-h. Sun, Epigenetic activation of the MiR-200 family contributes to H19-mediated metastasis suppression in hepatocellular carcinoma. *Carcinogenesis* **34**, 577–586 (2013).
57. R. Bergström, K. Savary, A. Morén, S. Guibert, C.-H. Heldin, R. Ohlsson, A. Moustakas, Transforming growth factor β promotes complexes between Smad proteins and the CCCTC-binding factor on the *H19* imprinting control region chromatin. *J. Biol. Chem.* **285**, 19727–19737 (2010).
58. I. J. Matouk, E. Raveh, R. Abu-lail, S. Mezan, M. Gilon, E. Gershtain, T. Birman, J. Gallula, T. Schneider, M. Barkali, C. Richler, Y. Fellig, V. Sorin, A. Hubert, A. Hochberg, A. Czerniak, Oncofetal H19 RNA promotes tumor metastasis. *Biochim. Biophys. Acta* **1843**, 1414–1426 (2014).
59. C. Chen, Z. Li, Y. Yang, T. Xiang, W. Song, S. Liu, Microarray expression profiling of dysregulated long non-coding RNAs in triple-negative breast cancer. *Cancer Biol. Ther.* **16**, 856–865 (2015).
60. X.-H. Liu, M. Sun, F.-q. Nie, Y.-b. Ge, E.-b. Zhang, D.-d. Yin, R. Kong, R. Xia, K.-h. Lu, J.-h. Li, W. De, K.-m. Wang, Z.-x. Wang, Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Mol. Cancer* **13**, 92 (2014).
61. G. Yang, S. Zhang, F. Gao, Z. Liu, M. Lu, S. Peng, T. Zhang, F. Zhang, Osteopontin enhances the expression of HOTAIR in cancer cells via IRF1. *Biochim. Biophys. Acta* **1839**, 837–848 (2014).
62. Z. Hui, M. Xianglin, Association of HOTAIR expression with PI3K/Akt pathway activation in adenocarcinoma of esophagogastric junction. *Open Med.* **11**, 36–40 (2016).
63. J. Zhou, X. Zhi, L. Wang, W. Wang, Z. Li, J. Tang, J. Wang, Q. Zhang, Z. Xu, Linc00152 promotes proliferation in gastric cancer through the EGFR-dependent pathway. *J. Exp. Clin. Cancer Res.* **34**, 135 (2015).
64. J. Ji, J. Tang, L. Deng, Y. Xie, R. Jiang, G. Li, B. Sun, LINC00152 promotes proliferation in hepatocellular carcinoma by targeting EpCAM via the mTOR signaling pathway. *Oncotarget* **6**, 42813–42824 (2015).
65. K. L. Scott, O. Kabbarah, M.-C. Liang, E. Ivanova, V. Anagnostou, J. Wu, S. Dhakal, M. Wu, S. Chen, T. Feinberg, J. Huang, A. Saci, H. R. Widlund, D. E. Fisher, Y. Xiao, D. L. Rimm,

A. Protopopov, K.-K. Wong, L. Chin, GOLPH3 modulates mTOR signalling and rapamycin sensitivity in cancer. *Nature* **459**, 1085–1090 (2009).
66. Y. Gökmen-Polar, I. T. Vladislav, Y. Neelamraju, S. C. Janga, S. Badve, Prognostic impact of HOTAIR expression is restricted to ER-negative breast cancers. *Sci. Rep.* **5**, 8765 (2015).
67. J. Sun, X. Chen, Z. Wang, M. Guo, H. Shi, X. Wang, L. Cheng, M. Zhou, A potential prognostic long non-coding RNA signature to predict metastasis-free survival of breast cancer patients. *Sci. Rep.* **5**, 16553 (2015).
68. K. P. Sørensen, M. Thomassen, Q. Tan, M. Bak, S. Cold, M. Burton, M. J. Larsen, T. A. Kruse, Long non-coding RNA expression profiles predict metastasis in lymph node-negative breast cancer independently of traditional prognostic markers. *Breast Cancer Res.* **17**, 55 (2015).
69. P.-J. Volders, K. Helsens, X. Wang, B. Menten, L. Martens, K. Gevaert, J. Vandesompele, P. Mestdagh, LNCipedia: A database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* **41**, D246–D251 (2013).
70. P. Stępniak, M. Maycock, K. Wojdan, M. Markowska, S. Perun, A. Srivastava, L. S. Wyrwicz, K. Świrski, Microarray Inspector: Tissue cross contamination detection tool for microarray data. *Acta Biochim. Pol.* **60**, 647–655 (2013).
71. J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, P. S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
72. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
73. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
74. D. Venet, J. E. Dumont, V. Detours, Most random gene expression signatures are significantly associated with breast cancer outcome. *PLOS Comput. Biol.* **7**, e1002240 (2011).
75. P. Du, W. A. Kibbe, S. M. Lin, *lumi*: A pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).

Acknowledgments: We thank J. Vandesompele, P. Mestdag, H. Duvillier, and S. Garaud for their helpful advice. **Funding:** O.V.G. was supported by the Télévie. M.B. was supported by the Télévie. M.D. was supported by the Walloon Region (WB Health grant, CanDx). E.J.d.B. was supported by the Belgian F.R.I.A. This work was funded by grants from the Fonds de la Recherche Scientifique and Télévie, as well as by grants from the IUAP P7/03 program, the Action de Recherche Concerté (AUWB-2010-2015 ULB-No 7), the Belgian "Foundation against Cancer," the WB Health program, and the Fonds Gaston Ithier. **Author contributions:** O.V.G. and F.F. designed the experiments and interpreted the data. O.V.G., M.B., E.J.d.B., C.O., S.B., and M.D. performed the bioinformatics analysis. O.V.G., E.C., and P.P. performed the qPCR and proliferation/migration experiments. M.G., G.B., and C.S. provided technical support and advices. F.F. directed the study, and O.V.G. and F.F. wrote the article. **Competing interests:** G.B., C.S., and F.F. are Université Libre de Bruxelles professors. The other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 4 February 2016
Accepted 5 August 2016
Published 2 September 2016
10.1126/sciadv.1600220

Citation: O. Van Grembergen, M. Bizet, E. J. de Bony, E. Calonne, P. Putmans, S. Brohée, C. Olsen, M. Guo, G. Bontempi, C. Sotiriou, M. Defrance, F. Fuks, Portraying breast cancers with long noncoding RNAs. *Sci. Adv.* **2**, e1600220 (2016).

Science Advances

Portraying breast cancers with long noncoding RNAs

Olivier Van Grembergen, Martin Bizet, Eric J. de Bony, Emilie Calonne, Pascale Putmans, Sylvain Brohée, Catharina Olsen, Mingzhou Guo, Gianluca Bontempi, Christos Sotiriou, Matthieu Defrance and François Fuks

Sci Adv **2** (9), e1600220.
DOI: 10.1126/sciadv.1600220