Long intergenic non-coding RNA expression signature in human breast cancer

Yanfeng Zhang1,*,†, Erin K. Wagner2,*, Xingyi Guo3, Isaac May3, Qiuyin Cai1, Wei Zheng3, Chunyan He2,4 & Jirong Long1

Breast cancer is a complex disease, characterized by gene deregulation. There is less systematic investigation of the capacity of long intergenic non-coding RNAs (lincRNAs) as biomarkers associated with breast cancer pathogenesis or several clinicopathological variables including receptor status and patient survival. We designed a two-stage study, including 1,000 breast tumor RNA-seq data from The Cancer Genome Atlas (TCGA) as the discovery stage, and RNA-seq data of matched tumor and adjacent normal tissue from 50 breast cancer patients as well as 23 normal breast tissue from healthy women as the replication stage. We identified 83 lincRNAs showing the significant expression changes in breast tumors with a false discovery rate (FDR) < 1% in the discovery dataset. Thirty-seven out of the 83 were validated in the replication dataset. Integrative genomic analyses suggested that the aberrant expression of these 37 lincRNAs was probably related with the expression alteration of several transcription factors (TFs). We observed a differential co-expression pattern between lincRNAs and their neighboring genes. We found that the expression levels of one lincRNA (RP5-1198O20 with Ensembl ID ENSG00000230615) were associated with breast cancer survival with \( P < 0.05 \). Our study identifies a set of aberrantly expressed lincRNAs in breast cancer.

The advances in high-throughput sequencing technologies provide a systematic landscape in the complexity of transcriptome, especially the discovery of a wealth of long non-coding transcripts1,2. This phenomenon is referred to as pervasive transcription1, albeit some transcripts are probably the byproduct of the transcription noise4. Characterized as transcription from the intergenic regions, lincRNAs are specific types of long non-coding RNAs (lincRNAs), ranging in length from 200 nt to >100 kb5. Biochemically, many lincRNAs harbor mRNA-like features, such as the poly-adenosine (polyA+) tail and splicing6, enabling them to be detectable by polyA+ RNA-sequencing (RNA-seq) method. Functionally, there is mounting evidence that lincRNAs may act as modular scaffolds to recruit other active or repressive regulators7, resulting in widespread roles in imprinting control8, chromosome remodeling9, pluripotency and differentiation10, immune responses11 and cancer progression12. Many lincRNAs are expressed in high tissue or cell specificity10, serving lincRNA molecules as potential disease biomarkers.

Recently, the Cancer Genome Atlas (TCGA) has revealed that the aberrant expression of hundreds of protein-coding genes and tens of microRNAs are associated with breast cancer development, progression, metastasis13-16 and subtypes17,18. One large-scale study also reported hundreds of lincRNAs showing expression changes in breast cancer19, but this study has a lack of replication and comprehensive analyses. To further use lincRNAs as robust biomarkers, we performed an integrative analysis on RNA-seq data from approximately 1,000 breast tumors from the TCGA and an independent RNA-seq dataset of 50 pairs of matched tumors and adjacent normal tissues from breast cancer patients as well as 23 normal breast tissues from healthy women.

1Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37203, USA. 2Department of Epidemiology, Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, IN 46202, USA. 3Bowdoin College, Brunswick, ME, 04011, USA. 4Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, IN 46202, USA. 5Present address: HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.H. (email: chunhe@iu.edu) or J.L. (email: Jirong.Long@vanderbilt.edu)
Results

Data summary. In the discovery stage, RNA-seq data for 85 pairs of breast tumors and adjacent normal tissues were collected, along with 830 primary breast tumors available from TCGA. In the replication stage, RNA-seq data from paired tumor and adjacent normal tissues from 50 breast cancer patients and normal breast tissue from 23 healthy women were included. Overall, the study subjects had similar clinical characteristics between the two stages (Table 1).

Also included in the present study were Genome-wide protein-DNA interaction data using chromatin immunoprecipitation followed by sequencing (ChIP-seq) for 26 transcription factors (TFs) in MCF-7 cell lines from the ENCODE project, and two independent ChIP-seq datasets for transcription factor ERα each with replicated experiments in the MCF-7 cells20,21, as well as RNA-seq data across 14 breast cancer cell lines (Table S1).

Global changes of the lincRNA species in breast cancer. Our transcriptomic analyses focused on the lincRNA species, a specific type of lncRNAs. First, we investigated the lincRNA expression profile in 85 paired samples of primary breast tumor tissue and adjacent normal tissue from TCGA. After stringent filtering (see Methods), we totally identified 2,171 lincRNAs expressed in at least one sample. As expected, most lincRNAs expressed at low levels and sparsely expressed in <10% of all samples (Figure S1). Similar observations were also reported elsewhere22–24. For subsequent analysis, we focused on 584 lincRNAs that expressed in at least 20% of samples. Compared with the matched adjacent normal tissues, we observed that these lincRNAs showed a global decreased expression profile in breast tumors (Figure S2). The similar trend has been reported for small non-coding RNAs in multiple cancer types25,26. Specifically, among these 584 lincRNAs, 83 were significantly differential expression (DE) with a permutation-based FDR <1% (Fig. 1A). Among them, 60 lincRNAs displayed down-regulation in breast tumor tissues.

Validation of lincRNA expression in an independent study. To validate the DE lincRNAs identified in TCGA, we generated an independent RNA-seq dataset from 50 pairs of matched tumor and adjacent normal samples from breast cancer patients and normal breast tissues from healthy women in the replication study. Among the 83 DE lincRNAs, 66 lincRNAs (80%) showed consistent expression change direction (Fig. 1B). Specifically, 37 lincRNAs (45%) were replicated with absolute log2-FC ≥1 and P < 0.05 in the 50 pairs of matched tumor and adjacent normal breast tissues in replication stage (Table 2). Over 80% of them also showed significant expression difference in the tumors compared to normal breast tissues in the validation stage (Table 2). When we relaxed the cutoff to absolute log2-FC ≥0.585 and P < 0.05 in the replication dataset (Table S2), 47 DE lincRNAs (57%) could be replicated in our independent study, which is consistent with the previous report19.

We further examined the expression profiling of the 37 replicated lincRNAs in the remaining 830 breast tumor tissues samples without matched normal tissues from the TCGA dataset. The expression patterns of these 37 lincRNAs in the 830 tumor samples were very similar to those in 85 normal-paired breast tumor samples, but very different from those in adjacent normal breast tissue samples in the discovery stage (Fig. 1C). Principal component analysis (PCA) based on these 37 lincRNAs also showed that all breast tumor samples are clustered together, but separated from the adjacent normal tissues (with a few exceptions) (Fig. 1D). In addition, we also checked the expression levels of these 37 lincRNAs across 14 breast cancer cell lines that are commonly used in the lab (Table S3). Most of these 37 lincRNAs were expressed in at least one common breast cancer cell line, making these

|                     | Discovery stage | Replication stage |
|---------------------|-----------------|-------------------|
|                     | Paired tumor-normal | Only tumor tissues\[^a\] | Paired tumor-normal | Healthy Women |
| No. of samples      | 85              | 830               | 50                 | 23            |
| Age (yr \(\pm\) s.d.) | 57.2 \(\pm\) 15.4 | 58.4 \(\pm\) 12.9 | 51.7 \(\pm\) 10.9 | 52.0 \(\pm\) 11.3 |
| Ethnicity           |                 |                   |                   |
| White               | 79 (92.9%)      | 588 (70.8%)       | 39 (78%)          | 18 (78%)      |
| Asian               | 1 (1.2%)        | 56 (6.8%)         | 0                 | 0             |
| Black               | 5 (5.9%)        | 79 (9.5%)         | 11 (22%)          | 5 (22%)       |
| Unknown             | 0               | 107 (12.9%)       | 0                 | 0             |
| ER status           |                 |                   |                   |
| ER+                 | 67 (78.9%)      | 597 (71.9%)       | 40 (80%)          | NA            |
| ER−                 | 15 (17.6%)      | 181 (21.8%)       | 10 (20%)          | NA            |
| Unknown             | 3 (3.5%)        | 52 (6.3%)         | 0                 | NA            |
| Stage               |                 |                   |                   |
| 0-I                 | 16 (18.8%)      | 135 (16.2%)       | 2 (4%)            | NA            |
| II                  | 46 (54.1%)      | 469 (56.5%)       | 14 (28%)          | NA            |
| III                 | 20 (23.5%)      | 179 (21.6%)       | 6 (12%)           | NA            |
| IV                  | 2 (2.4%)        | 13 (1.6%)         | 1 (2%)            | NA            |
| Unknown             | 1 (1.2%)        | 34 (4.1%)         | 27 (54%)          | NA            |

Table 1. Clinical characteristic of study subjects in this study. \[^a\]BRCA (breast invasive carcinoma) with matched adjacent normal tissues. \[^b\]BRCA without matched adjacent normal tissues.
cell lines as possible *in vitro* model to study lincRNAs. Altogether, these results suggested that these 37 lincRNAs showed aberrant expression in breast tumors.

**Association of LincRNAs with transcription factors.** We then integrated 26 known TFs ChIP-seq data in MCF-7 cells, an ER+ breast cancer cell line from the ENCODE project (Table S4), to investigate whether and to what extent the 37 differentially expressed lincRNAs are functionally related to TFs. We constructed a lincRNA-TF bipartite network (see Methods), which consisted of 45 nodes and 101 edges (Fig. 2A). We found

---

**Figure 1. Expression differences of lincRNAs between human breast cancer and normal tissues.** (A) Heatmap of 83 differentially expressed lincRNAs in 85 pair-matched breast cancer samples. Red and green bars denote 85 breast cancer specimens and adjacent normal tissues from TCGA, respectively. Colors ranging from blue to red represent the relative expression levels (Z-score) of lincRNAs. (B) Dot plot of the Log2-transformed FC for 83 lincRNAs between 50 paired breast tumor and adjacent normal tissues (Tumor−Adjacent), and between 50 breast tumors and 23 normal tissues from healthy women (Tumor−Normal) as validation (X-axis) versus that from TCGA breast cancer patients (Y-axis). (C) Heatmap of 37 differentially expressed and replicated lincRNAs in 830 unmatched tumors (blue bar) in combined with 85 pairs of tumors (red bar) and adjacent normal tissues (green bar). Values range in color from blue to red, meaning the relative expression levels. (D) PCA of 85 paired breast cancer and adjacent normal samples (85 BRCA and 85 Adjacent) and 830 unpaired breast cancer samples (830 BRCA) based on 37 differentially expressed lincRNAs.
Several studies have reported that 20 (54.1%) lincRNAs were bound by at least one TF, and 25 of the 26 TFs regulated at least one lincRNA. Especially, each of the eight key TFs, including GATA3, RAD21, MYC, CTCF, MAX, E2F1, CEBPB and RNA polymerase II, binds ≥ 5 lincRNAs. Among them, 11 TFs exhibited a significant differential expression (absolute log2-FC ≥ 0.585, BH-adjusted P ≤ 0.01) in breast cancer tissues compared with adjacent normal tissues (Fig. 2B). We also found that several over-expressed lincRNAs may be related to these key TFs. For example, two lincRNA genes (RP11-279F6 with Ensemble ID ENSG00000245750 and GATA3-AS1 with Ensemble ID ENSG00000197308) had DNA binding peaks by 16 and 20 TFs, respectively. Specifically, all three over-expressed key regulators in breast tumors, GATA3, E2F1 and RAD21, showed connection with these two lincRNAs (Fig. 2C,D). These results suggest that the expression alteration of lincRNAs in breast cancer is probably functionally associated by one or multiple TFs in a complex way.

**Co-expression of lincRNAs with nearby protein-coding genes.** Several studies have reported that lincRNAs show an expression correlation with their neighboring protein-coding genes in normal tissues/cells. We examined whether 37 DE lincRNAs could also exhibit the differential co-expression profile with the adjacent genes in breast cancer development. We classified these 37 lincRNA/mRNA pairs into three types: head-head

| Ensembl ID | Symbol | Discovery stagea | Replication Stageb | Replication Stagec |
|------------|--------|-----------------|--------------------|--------------------|
| ENSG00000261039 | RP11-417E7 | 7.04 | 3.49 | 7.11 |
| ENSG000002303838 | AC093580 | 5.84 | 9.72 | 15.03 |
| ENSG000002245777 | AC017048 | 3.26 | 2.17 | 1.03 |
| ENSG000002533646 | RP11-731F5 | 2.52 | 9.31 | 13.17 |
| ENSG00000197308 | GATA3-AS1 | 2.49 | 8.27 | 5.58 |
| ENSG00000259093 | RP11-111210 | 2.20 | 2.02 | 2.78 |
| ENSG00000230615 | RP5-1198C20 | 1.72 | 6.53 | 7.58 |
| ENSG00000253161 | RP11-150C12 | 1.68 | 4.61 | 5.54 |
| ENSG00000245750 | RP11-279F6 | 1.57 | 4.96 | 6.07 |
| ENSG00000214548 | MEG3 | -1.17 | -2.35 | -0.75 |
| ENSG00000237897 | LINC00312 | -1.18 | -3.10 | 0.41 |
| ENSG00000267519 | CTD-3252C9 | -1.48 | -1.01 | 1.33 |
| ENSG00000267272 | RP5-105235 | -1.56 | -3.02 | -2.37 |
| ENSG00000255455 | RP11-890B15 | -1.62 | -1.59 | -1.35 |
| ENSG00000234076 | TPRG1-AS1 | -1.74 | -1.93 | -3.37 |
| ENSG00000262728 | AC123768 | -1.82 | -1.58 | -3.01 |
| ENSG00000269930 | RP11-932C9 | -1.85 | 1.82 | -1.74 |
| ENSG00000260025 | RP11-490M8 | -2.15 | -1.99 | -2.78 |
| ENSG00000260807 | RP11-161N6 | -2.24 | -1.96 | -5.65 |
| ENSG00000231246 | RPS-966F6 | -2.32 | -3.42 | -2.94 |
| ENSG00000266176 | RP11-855A2 | -2.39 | 4.16 | 0.16 |
| ENSG00000260993 | AC026150 | -2.44 | -1.81 | -1.19 |
| ENSG00000223477 | LINC00842 | -2.46 | -2.93 | -5.50 |
| ENSG00000182021 | RP11-38107 | -2.57 | -3.26 | -3.67 |
| ENSG00000258498 | DIO3OS | -2.68 | -3.34 | -3.76 |
| ENSG00000269936 | MIR145 | -2.70 | -1.34 | -1.61 |
| ENSG00000229645 | LINC00341 | -2.77 | -3.89 | -2.82 |
| ENSG00000261064 | RP11-100B6 | -2.79 | -1.33 | -1.77 |
| ENSG00000229108 | AC005550 | -2.81 | -4.31 | -6.34 |
| ENSG00000267194 | RP11-193H18 | -2.81 | -2.32 | -3.05 |
| ENSG00000260124 | RP4-791K14 | -2.84 | -1.84 | -2.32 |
| ENSG00000224307 | RP11-341B5 | -3.10 | -3.07 | -1.88 |
| ENSG00000178947 | LINC00806 | -3.23 | -1.59 | -2.66 |
| ENSG00000259070 | LINC00639 | -3.45 | -2.30 | -3.48 |
| ENSG00000271239 | RP11-238F2 | -3.67 | -3.19 | -3.86 |
| ENSG00000262179 | RP1-302G2 | -4.26 | -5.21 | -4.74 |
| ENSG00000246430 | RP11-161M8 | -5.25 | -6.36 | -3.19 |

Table 2. Expression alteration of 37 lincRNAs in discovery and replication stages. aThe discovery stage from TCGA. b,cThe replication stage from an independent study; bcomparison between paired tumor and adjacent normal breast tissue from cancer patients; ccomparison between tumor and normal breast tissue from healthy controls. dLog2-FC denotes log2-transformed fold change between breast tumor and non-tumor samples.
(H2H), head-tail (H2T) and tail-tail (T2T) based on the paired transcriptional orientation (Fig. 3A). We observed a coordinated expression change ($r^2 = 0.25, P = 1.0 \times 10^{-4}$) between lincRNAs and their neighbor mRNAs via a comparison of 85 paired breast tumor and adjacent normal tissues (Fig. 3A and B). We detected 14 (37.8%) lincRNA/mRNA pairs with the H2H type, which shows marginally significant over-representation.
Taking the H2H pair of GATA3-AS1 (Ensembl (p = 0.06, one-sided Binomial test) relative to the global distribution. Similar results were observed when we analyzed 83 initially detected DE lincRNAs (Figure S3). Taking the H2H pair of GATA3-AS1 (Ensembl...
α-lincRNAs may be regulated by ERα. Also observed in the other two lincRNAs (Fig. 4C,D). These results indicated that these three ER status associated region using chromatin states in human mammary epithelial cells (HMECs). Similarly, this enhancer state was on the analysis of ERα ChIP-seq data (Fig. 4B). This binding region was further annotated as the active enhancer on the example, an ERα associated lincRNA (RP5-1198O20) showed the up-regulation in both stages associated lincRNA RP5-1198O20 (ENSG00000230615) and AC093850 (Ensembl ID ENSG00000224577). For example, two over-expressed lincRNAs (RP11-417E7 with Ensembl ID ENSG00000261039 and AC093850 with Ensembl ID ENSG00000230615) showed over-expression in both the discovery and the replication stage. These two lincRNAs may be involved in cell adhesion, TGFβ signaling pathway and others (Fig. 3C–F). The TGFβ signaling pathway is well documented with a promoter of tumor progression and invasion. Together, these two lincRNAs probably participate in the pathogenesis of breast cancer.

LincRNAs associated with breast cancer subtypes. We also investigated whether these 37 DE lincRNAs exhibited expression difference across different breast cancer subtypes. We found three lincRNAs: GATA3-AS1 (ENSG00000197308), RP11-279F6 (ENSG00000245750) and AC017048 (ENSG00000224577), showed specifically high expression levels in ER-positive (ER+) cancers, compared to ER-negative (ER-) cancers and normal breast tissue samples (Fig. 4A). The specific expression alteration of these three lincRNAs in ER+ subtype was also validated in the replication stage. We also used the data of differentially expressed lincRNAs across breast cancer subtypes (Luminal A, Luminal B, Her2, and Basal-like) reported in Su et al. study28 to evaluate the subtype associations for these 37 lincRNAs. We found that the 22 (59.5%) lincRNAs (8 up-regulated lincRNAs and 14 down-regulated lincRNAs) were differentially expressed in at least one breast cancer subtype (Table S6). Three above reported lincRNAs over-expressed in the ER+ subtype (Fig. 4A) showed much higher expression levels in Luminal A and Luminal B subtypes (enriched for ER+), relative to Her2, and Basal-like subtypes.

Further integrating two independent estrogen receptor alpha (ERα) ChIP-seq dataset (each with two replicates) in MCF-7 cancer cells, we observed the ERα binding sites in all three lincRNA loci (Fig. 4B–D). For example, an ERα DNA binding site was found near the transcriptional terminal region of the GATA3-AS1 based on the analysis of ERα ChIP-seq data (Fig. 4B). This binding region was further annotated as the active enhancer region using chromatin states in human mammary epithelial cells (HMECs). Similarly, this enhancer state was also observed in the other two lincRNAs (Fig. 4C,D). These results indicated that these three ER status associated lincRNAs may be regulated by ERα.

LincRNAs associated with breast cancer survival. We assessed the association of those 37 lincRNAs expression levels with breast cancer survival and found one candidate (RP5-1198O20 with Ensembl ID ENSG00000230615, Fig. 5). Patients with expression levels of lincRNA ENSG00000230615 higher than the median value (median FPKM value is 2.8) had worse survival rates than those women with expression levels less than the median value (median FPKM value is 0.5) (Fig. 5).

Discussion
In this study, we analyzed lincRNA transcriptome in over 1000 breast tissue samples. Two independent sequencing datasets consistently identified a set of lincRNAs deregulated in breast carcinogenesis. Importantly, the protein-coding genes neighboring these deregulated lincRNA loci also showed expression alternation in breast cancer tissues, implying the transcriptional concordance between lincRNAs loci and neighboring genomic regions in cancer development. The expression aberration of lincRNAs in breast cancer may be associated with the expression alteration of multiple TFs. Our work substantially extends the biological understanding of the lincRNA repertoire in the pathogenesis of breast cancer.

It should be noted that the aberrant expression of several lincRNAs previously identified in multiple cancers showed deregulation in the present study. For example, in both discovery and validation stages, expression levels of the MEG3 decreased in breast tumor samples, which is consistent with the proposed tumor suppressor role for MEG3. Another un-regulated lincRNA, GATA3-AS1 (Ensembl ID ENSG00000197308) was also identified in Ding et al. study33. Our in silico functional prediction (Table S5) indicated the GATA3-AS1 probably performs an immune response associated role in breast cancer progression. This result was supported by previous two studies26,34 showing that the GATA3-AS1 is highly expressed in T helper subsets. Spurlock et al. also speculated that the GATA3-AS1 might play a role in allergic or asthmatic responses34. In addition, this lincRNA was also reported to have a decreased expression level in brain, bladder and prostate cancers. Another breast cancer survival associated lincRNA RP5-1198O20 (Ensembl ID ENSG00000230615) showed the up-regulation in both stages
of this study. Re-examining Gibb et al. integrated SAGE-seq findings22, we confirmed that this lincRNA also showed an increased expression in breast cancer. Another study about transcriptome analysis of aging identified the down-regulation of this lincRNA35. Further functional investigation of the lincRNA RP5-1198O20 in either carcinogenesis or aging would be interesting. In another example, our data showed the down-regulation of the miRNA-145 host gene (MIR145 with Ensembl ID ENSG00000269936) in breast cancer, consistent with previous reports in breast cancer36 and other cancer types37,38. However, several well-characterized lincRNAs12,22,23,39,40, including HOTAIR, H19, GAS5, PCA3, PVT1, were not investigated in this study. Those lncRNAs are not transcribed from intergenic regions, belonging to either anti-sense or other types of long non-coding transcripts, according to the GENCODE (version17) annotation. That means these lncRNAs are difficult to distinguish from their host protein-coding genes using the typical RNA-seq technology. There are several lncRNAs with no polyA tails41,42 which are unmeasured in the current study as well. Therefore, more sophisticated methods, such as strand-specific and non-poly(A) tail RNA-seq technology, are required to distinguish anti-sense transcripts from protein-coding genes, and to comprehensively capture the lncRNA transcriptome.

Several lincRNAs show expression alteration in other cancers, including NEAT1 (Ensembl ID ENSG00000245532) down-regulated in retinoblastoma22, MALAT1 (Ensembl ID ENSG00000251562)
up-regulated in lung and colorectal cancer\textsuperscript{43,44}. However, these lincRNAs do not exhibit expression aberration in breast cancer. Likewise, the \textit{PCAT1} (Ensembl ID ENSG00000253438), a prostate-specific lincRNA regulating cell proliferation, shows over-expression in a subset of prostate cancers\textsuperscript{45}. In our study, we almost did not detect the expression level of the \textit{PCAT1} in either breast normal or breast tumor samples. These results partially support the observation that many lincRNAs are expressed in a tissue- and cancer-type restricted manner, making them useful as prognostic markers\textsuperscript{12}.

However, some limitations remain in this study. Firstly, no functional validations on these DE lincRNAs prevent us draw the further conclusion about how the aberrant expression of lincRNAs contribute to tumorigenesis. Secondly, the relative small sample size and the unavailability of breast cancer subtype data in the replication stage also prevent us replicate the subtype-specific lincRNAs. Finally, due to the tumor heterogeneity and the cell mixture in tumor tissues, global comparison of gene expression profiles among breast tumor, adjacent and normal tissues is insufficient, particularly for lincRNAs whose expression patterns broadly show the cell- or tissue-type specificity. Thus, single-cell or sorted cell population based transcriptomic analysis\textsuperscript{46,47} will be favorable to determine these lincRNAs as robust biomarkers.

Conclusions
We identified a signature of lincRNA expression profile for breast cancer. Further functional surveys of these lincRNAs will be warranted to discover the biological mechanisms of lincRNA in breast cancer development and progression.

Materials and Methods

\textbf{Discovery dataset.} After approval, raw RNA-seq data (Level 1) of breast tumor tissues (\(N = 915\)) and adjacent normal breast tissues (\(N = 85\)) were acquired from the TCGA. Clinical data (Biotab format) for these 915 breast invasive carcinoma (BRCA) were also acquired from the TCGA.

\textbf{Replication dataset.} Our replication study consisted of 50 breast cancer cases and 23 healthy controls. The patients were pathologically confirmed for primary breast cancer diagnosed at one of three hospitals in Indianapolis, Indiana, between 1998 and 2009: Indiana University Hospital, Wishard Hospital (now Eskenazi Hospital), and the Indiana University Simon Cancer Center (IUSCC). Controls were randomly selected from a pool of healthy women who donated normal breast tissues to the Susan G. Komen Tissue Bank between 2005 and 2009, and were free of breast cancer at the time of donation. The participants completed a questionnaire on medical histories and health-related exposures at the time of donation. Controls were frequency matched to cases based on self-reported ancestry and age (within 5 years).

Breast tissue (untreated tumor or normal) biospecimens were collected from each case and control under standard operating procedures at IUSCC. All breast tissue samples were snap-frozen immediately after removal and stored in liquid nitrogen until processed, and were determined to be of high quality through histological and molecular quality control tests. Tumor samples were pathologically verified for high tumor content. Information concerning demographics, clinical data, and personal risk factors, including age and race, were either extracted from medical records (for cases) or obtained through the questionnaires completed by the participants (for controls).
Total RNA was extracted from freshly frozen breast tissue samples using the Qiagen® miRNeasy Mini Kit. Construction of cDNA libraries and subsequent RNA sequencing of paired-end reads (2 × 50 nt reads) were performed according to the standard Illumina protocol using the HiSeq2000 sequencing systems. The raw sequencing output was 25–35 million reads per sample.

**Other dataset.** The annotation data for known lncRNAs (n = 6,020) were extracted from the gencode.v17. long_noncoding_RNAs.gtf.gz file from the GENCODE database. Other data used in this study include: human Refseq genes obtained from the NCBI database; protein-DNA interactions from ChIP-seq data in MCF-7 breast cancer cell lines downloaded from ENCODE project⁴⁸; DNA binding by ERα using ChiP-seq in MCF-7 cells, and RNA-seq data across 14 breast cancer cell lines⁹,50 downloaded from the Gene Expression Omnibus (GEO) database³¹.

**RNA-seq data processing and lincRNA annotation.** In the discovery stage, for the 85 paired match breast tumor and normal tissues, mapped reads in BAM format (Level 1) were assembled with Cufflinks (v2.1.1)⁵². The lincRNA annotation was conducted by the following procedures: 1) we retained assembled transcripts whose genomic loci are overlapped or embedded with known annotated lincRNAs; 2) we removed transcripts in length of <200 nt; 3) we eliminated transcripts showing coding potential (score <0.5) predicted with iSeeRNA program (version 1.2)⁵³; 4) if two or more transcripts (isoforms) mapped to a lincRNA locus, we assigned the mean coverage and expression value to that lincRNA. We applied the same procedures for the remaining RNA-seq data, except for publicly accessible RNA-seq data that were processed as described elsewhere³⁵. To reduce the noise caused by the lincRNA expression variability between samples, we plotted the calling rate (occurrence of lincRNA transcription based on Cufflinks FPKM ≥ 0.3) versus lincRNAs ranked by missing rate across 170 samples. As Ramskold et al. proposed, this cutoff (FPKM ≥ 0.3) is an optimized threshold for detectable expression above background. Then, we set the calling rate ≥ 20% as a threshold to ensure high-confidence lincRNAs for further analyses.

For RNA-seq data in the replication stage, quality control (QC) filtering was first performed on raw RNA-seq data to remove adapter sequences and poor quality bases using the FastqMCF clipping algorithm. RNA-seq reads were then mapped by Bowtie v1.0.0, to GENCODE IncRNA reference (release 17) for IncRNA (including lincRNA) annotations. Transcript abundances were quantified using NGSUitls. Samples were further filtered based on percentage of genes detected (less than 50%) and percentage of reads mapped to the reference (less than 25%). Extreme outliers were further identified and filtered from the dataset using principal component analysis (PCA). After these steps, a total of 7,450 lncRNAs retained and were used in further analyses.

**ChiP-seq data processing.** For ChiP-seq data from the ENCODE project, we directly downloaded files with called peaks for subsequent analyses. For two independent ERα ChiP-seq data whose processed peaks were unavailable, we downloaded the raw data in FASTQ format and conducted the peak calling as follows. Reads were mapped to the human reference genome (hg19) using Bowtie2 program⁵⁶ in the default parameters. Aligned data were processed and converted into BAM files using SAMtools program⁵⁷. Then, we used the MACS14 (version 1.4.2) program⁵⁸ to call peaks in 20 bp resolution. We visualized the results in the UCSC Genome Browser.

**Differential expression of lincRNAs.** The fragments per kilobase of exon per million fragments mapped (FPKM) values were calculated from the Cufflinks program to represent lincRNA expression levels. The differential expression of each lincRNA between breast tumor and adjacent normal tissues was defined as fold change (FC) ≥ 2 with Benjamini-Hochberg (BH) adjusted P < 0.01 based on non-parametric Wilcoxon rank sum paired test.

Following the permutation-based method developed by Xie et al.⁵⁹, we estimated the FDR in identifying lincRNAs with differential expression. Briefly, FDR is estimated as $\overline{\text{FDR}}(d) = \frac{\overline{FP}(d)}{\overline{TP}(d)}$ where $d$ is a fixed threshold, $Z_i$ is a test statistic and $B = 1000$ permutations. $\overline{TP}(d) = \# \{i: |Z_i| > d\}$ and $\overline{FP}(d) = \frac{1}{B} \sum_{i=1}^{B} I(|Z_i| > d)$. The 1000 permutations were achieved by shuffling the tumor/normal samples in the lincRNA-sample matrix (584 lincRNAs and 170 samples).

For RNA-seq data in the replication stage, differential expression (DE) analyses were performed using edgeR v2.6.12, implemented in the Bioconductor package to identify differentially expressed lincRNAs between tumor and adjacent normal, as well as tumor and normal breast tissue. The trimmed mean of M-values (TMM) method was used to normalize the raw read counts. Biological coefficients of variation between the samples were estimated using an empirical Bayes approach under the assumption that the data follows a negative binomial distribution. Differential expression between tumor and adjacent normal or normal breast tissue was analyzed using a generalized linear model to regress lincRNA (expression on tissue type, adjusting for age, race, and sequencing batch). Statistical significance was defined as FDR p-value < 0.05 and a two-fold change (FC) of expression level between comparison of tumor and adjacent normal or normal breast tissue.

**Construction of transcription factor and lincRNA network.** We downloaded ChiP-seq peak files of 26 transcription factors (TFs) in the MCF-7 breast cancer cells from the ENCODE project. A lincRNA bound by a TF was determined, if the TF binding peak is located within the region of 2 kb upstream of its transcription start site (TSS) to transcription end site (TES) of the lincRNA locus. Then, a lincRNA-TF bipartite network was constructed using the Cytoscape (v2.8.3) software⁶⁰, where each node denotes a lincRNA or TF and each edge represents a lincRNA-TF interaction. The degree of each node in this bipartite network was calculated. We regarded the nodes representing TFs having ≥ 5 degrees as key TFs.
Co-expression analysis. We used similar methods for lincRNA annotation to annotate Refseq protein-coding genes and calculate their FPKM values. The neighboring protein-coding genes were defined as the ones with the closest physical distance to known lincRNAs. We coupled lincRNA-encoding locus and neighboring protein-coding genes, and referred to it as a lincRNAs-mRNA pair. We further classified lincRNAs-mRNA pairs into three types, according to their transcriptional orientation: head-to-head (H2H), head-to-tail (H2T) and tail-to-tail (T2T). The linear relationship of log2-transformed fold change in a comparison of 85 paired samples for lincRNA-mRNA pairs was used to evaluate the coordinated changes.

We defined the non-neighboring genes as the ones over a 1 Mb physical distance from lincRNAs on both strands. We coupled them as non-neighboring lincRNAs-mRNA pairs. Similarly, we determined the coordinated changes for non-neighboring lincRNA/mRNA pairs. We repeated the analysis using randomly selected 1,000 lincRNA-mRNA pairs (regardless of the distance) from the entire transcriptome data.

Functional prediction of lincRNAs. To predict biological functions for these DE lincRNAs, we calculated the correlation coefficients between DE lincRNAs and all Refseq protein-coding genes using the Spearman rank correlation analysis. We regarded the Spearman rank correlation coefficients calculated from randomly shuffling lincRNA/mRNA pairs (1000 times) as null distribution. Compared with the null distribution, we set a threshold for the Spearman rank correlation coefficient ≥0.4 (or ≤-0.4) (Figure S6) to reflect the strong co-expression between lincRNAs and Refseq genes in high confidence. On this basis, a set of Refseq genes passing this threshold were regarded as the functional association to lincRNAs, and used for functional enrichment analysis using DAVID annotation61. GO terms with BH-adjusted P ≤ 0.05 served as functional enrichment for lincRNAs.

Survival analyses. Excluding participants with unknown survival information (n = 16), the remaining 899 subjects were retained in survival analyses. We split patients into two groups (higher and lower expressions of lincRNA) based on the median level of lincRNA expression. The Kaplan–Meier curve and hazard ratio (HR) of higher versus lower expressed groups were generated in R (versions 2.15.0) using the survival package.

Ethical consent. Utilization of data was conducted in accordance with TCGA data access policies. Signed informed consent was obtained from each subject in the replication study. The study was approved by Indiana University institutional review board.

References
1. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science 322, 1845–1848 (2008).
2. van Bakel, H., Niskov, C., Blencowe, B. J. & Hughes, T. R. Most Dark Matter Transcripts Are Associated With Known Genes. PLoS Biol 8, e1000371 (2010).
3. Costa, F. F. Non-coding RNAs: Meet thy masters. BioEssays 32, 599–608 (2010).
4. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Research 17, 556–565 (2007).
5. Uiltsky, I. & Bartel, D.-P. lincRNAs: Genomics, Evolution, or Mechanisms? Cell 154, 26–46 (2013).
6. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915–1927, doi: 10.1101/gad.17466111 (2011).
7. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. Nature 482, 339–346 (2012).
8. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. Molecular and Cellular Biology 10, 28–36 (1990).
9. Khalil, A. M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106, 11667–11672, doi: 10.1073/pnas.0904715106 (2009).
10. Sigova, A. A. et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci USA 110, 2876–2881, doi: 10.1073/pnas.1221944110 (2013).
11. Carpenter, S. et al. A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. Science 341, 789–792 (2013).
12. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the genetics of cancer. Br J Cancer 108, 2419–2425, doi: Minireview (2013).
13. Andorfer, C. A., Neele, B. M., Thompson, E. A. & Perez, E. A. MicroRNA signatures: clinical biomarkers for the diagnosis and treatment of breast cancer. Trends in Molecular Medicine 17, 313–319 (2011).
14. Gianfrocca, M. & Gradishar, W. New Molecular Classifications of Breast Cancer. CA: A Cancer Journal for Clinicians 59, 303–313 (2009).
15. Volinia, S. et al. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci USA 109, 3024–3029, doi: 10.1073/pnas.1200010109 (2012).
16. Weigelt, B., Peterse, J. L. & van’ t Veer, L. J. Breast cancer metastasis: markers and models. Nat Rev Cancer 5, 591–602, doi: 10.1038/nrc1760 (2005).
17. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352, doi: 10.1038/nature10983 (2012).
18. Perou, C. M. et al. Molecular portraits of human breast tumours. Nature 406, 747–752, doi: 10.1038/35021093 (2000).
19. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47, 199–208, doi: 10.1038/ng.3192 (2015).
20. Lai, C. F. et al. Co-regulated gene expression by oestrogen receptor a and liver receptor homolog-1 is a feature of the oestrogen response in breast cancer cells. Nucleic Acids Research 41, 10228–10240 (2013).
21. Li, W. et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498, 516–520 (2013).
22. Gibb, E. A. et al. Human Cancer Long Non-Coding RNA Transcripts. PLoS ONE 6, e25915, doi: 10.1371/journal.pone.0025915 (2011).
23. Kim, K. et al. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene 32, 1616–1625 (2013).
24. Marques, A. & Ponting, C. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biology 10, R124, doi: 10.1186/gb-2009-10-11-r124 (2009).
25. Lu, J. et al. MicroRNA expression profiles classify human cancers. Nature 435, 834–838, doi: 10.1038/nature03702 (2005).
Acknowledgements
We thank Jing He and Yan Guo for data processing and analyses, and Marshal S Younger for clerical support in the preparation of this manuscript. The results here are in part based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/). Data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. We thank the technical staff at Q Squared Solutions Expression Analysis LLC for their collaboration and scientific support in performing the RNA-seqencing for the replication study. We acknowledge Indiana CTSI Specimen Storage Facility (SSF), which is supported by NCRR Clinical and Translational Sciences Award (U54-RR025761) and NCRR construction award (C06-RR02128-01). We also thank the Indiana University Simon Cancer Center at Indiana University School of Medicine for the use of the Tissue Procurement & Distribution Core, which provided breast tumor specimens. Samples from the Susan G. Komen Tissue Bank at the Indiana University Simon Cancer Center were used in this study. We thank contributors, including Indiana University who collected samples used in this study, as well as donors and their families, whose help and participation made this work possible. This work was supported in
part by US National Institutes of Health grants R01CA137013 to JL and NIH R01 grant R01CA194030 and Susan G. Komen Career Catalyst Research grant CCR15333233 to CH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Author Contributions**
Y.Z., J.L. and C.H. conceived and designed the study; Y.Z., E.W., X.G., I.M. carried out data analysis; Y.Z., E.W., J.L., C.H. wrote the manuscript. C.H., Q.C. and W.Z. directed the data management. All authors read and approved the final manuscript.

**Additional Information**
Supplementary information accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Zhang, Y. *et al.* Long intergenic non-coding RNA expression signature in human breast cancer. *Sci. Rep.* 6, 37821; doi: 10.1038/srep37821 (2016).

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2016