Computational and Structural Biotechnology Journal 21 (2023) 4432–4445

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Delineating highly transcribed noncoding elements landscape in breast cancer

Wenyong Zhu [a,1], Hao Huang [a,1], Wenlong Ming [a], Rongxin Zhang [a], Yu Gu [a], Yunfei Bai [a], Xiaoan Liu [b], Hongde Liu [a], Yun Liu [c], Wanjun Gu [a,d], Xiao Sun [a,*]

[a] State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
[b] Department of Breast Surgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[c] Department of Information, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[d] Collaborative Innovation Center of Jiangsu Province of Cancer Prevention and Treatment of Chinese Medicine, School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, China

ARTICLE INFO

Keywords:
Highly transcribed noncoding elements
Chromatin characterization
Consensus motif analyses
Subtype specific transcriptional processes
Breast cancer oncogenes or tumor suppressors

ABSTRACT

Highly transcribed noncoding elements (HTNEs) are critical noncoding elements with high levels of transcriptional capacity in particular cohorts involved in multiple cellular biological processes. Investigation of HTNEs with persistent aberrant expression in abnormal tissues could be of benefit in exploring their roles in disease occurrence and progression. Breast cancer is a highly heterogeneous disease for which early screening and prognosis are exceedingly crucial. In this study, we developed a HTNE identification framework to systematically investigate HTNE landscapes in breast cancer patients and identified over ten thousand HTNEs. The robustness and rationality of our framework were demonstrated via public datasets. We revealed that HTNEs had significant chromatin characteristics of enhancers and long noncoding RNAs (lncRNAs) and were significantly enriched with RNA-binding proteins as well as targeted by miRNAs. Further, HTNE-associated genes were significantly overexpressed and exhibited strong correlations with breast cancer. Ultimately, we explored the subtype-specific transcriptional processes associated with HTNEs and uncovered the HTNE signatures that could classify breast cancer subtypes based on the properties of hormone receptors. Our results highlight that the identified HTNEs as well as their associated genes play crucial roles in breast cancer progression and correlate with subtype-specific transcriptional processes of breast cancer.

1. Introduction

Breast cancer is the most common malignancy and its high morbidity and mortality rates make it the leading female cancer [1]. Despite tremendous advances in breast cancer research over the last decade, the diagnosis and treatment of this malignancy remain challenging [2]. Recently, numerous studies have highlighted the significance of noncoding elements in cancers [3,4], but a comprehensive knowledge of the molecular mechanisms and functions of noncoding elements in carcinogenesis is still unclear. Therefore, the investigation of novel genomic elements (e.g., noncoding elements) associated with tumorigenesis, invasion and metastasis is of great importance for the understanding of the molecular mechanisms of oncogenesis.

Prior studies have identified various functional noncoding elements, such as enhancer and promoter, that are found to be essential in the regulation of proto-oncogenes or tumor suppressor genes at the transcriptional and post-transcriptional levels [4]. And a category of these noncoding elements exhibits an extraordinary degree of conservation between two or more organisms, known as conserved noncoding element (CNE), which tend to cluster in the vicinity of key developmental regulatory target genes and disruption of them could contribute to cancer [5]. Interestingly, the vast majority of CNEs are found to be transcribed, and their transcripts, which could be aberrantly expressed, exhibit distinct profiles in various human cancers [5,6]. Intensive research into the transcripts of noncoding elements in our genome have revealed that the majority of the human genome is transcribed (at least 76% in humans), yet only 1.2% of these RNAs actually encode proteins [7]. And transcription products tend to serve a certain function,

* Corresponding author.
E-mail address: xsun@seu.edu.cn (X. Sun).
[1] Contributed equally

https://doi.org/10.1016/j.csbj.2023.09.009
Received 3 May 2023; Received in revised form 27 August 2023; Accepted 10 September 2023
Available online 14 September 2023

otherwise, the cost of transcription would be wasted [7]. These noncoding transcripts, which could be highly aberrantly expressed in tumor tissues, typically fulfill diverse biological functions and play different roles in the disease progression [8,9]. For example, enhancer RNA (eRNA) is a kind of functional noncoding transcripts that is transcribed from the active enhancer in a tissue-specific manner and interacts with transcriptional regulators to regulate tumour-promoting genes, accounting for the instability of the cancer genome [10]. A variety of studies have provided insights into intronic noncoding transcription [11–13]. Corces et al. revealed that genetic risk loci for cancer susceptibility are active transposase-accessible DNA elements in cancer, and they lead to gene regulatory interactions underlying cancer immune evasion and are associated with noncoding mutations that affect patient survival [14]. Concurrently, Dong et al. also identified a similar type of DNA regulatory element, named transcribed noncoding elements (TNEs), which was actively transcribed in a merged RNA signal set of 99 human brain samples [15]. A fraction of TNEs was found to be putative enhancers specifically active in dopamine neurons in their study. It was also uncovered that TNEs actively transcribed overrepresented variants associated with diseases and are major cell-autonomous effectors of cis-acting genetic variants. However, their study only focused on TNEs associated with putative enhancers and did not investigate the biological implications of residual TNEs unaffiliated with enhancers. More importantly, RNA signal outliers in individual samples that could interfere with TNE detection posed a challenge to their method and could increase the incidence of false positives. It is conceivable that their approach may not apply to a large dataset of highly complex diseases (e.g., cancer).

To ameliorate the limitations of TNE identification [15] and apply it to our large and heterogeneous breast cancer dataset, we have enhanced their methodology and identified a robust category of noncoding elements that are highly transcribed and highly reliable, which are collectively defined as highly transcribed noncoding elements (HTNEs). It is foreseeable that HTNEs, similar to TNEs, could be closely associated with disease-related genes and function as potential markers and signalling molecules for tumour diagnosis and targeted interventions, which are critical to reveal the cell biological processes of breast cancer development and progression.

In this study, we employed a tailored identification pipeline to delineate the HTNE landscape in a specific cohort of breast cancer patients. To elucidate the critical functions of HTNEs in breast cancer, we characterized the identified HTNEs and their putative target genes by integrative analysis of multi-omics data. And we also investigated the relationship between HTNEs and subtype-specific transcriptional processes, with the goal of revealing the potential biological implications and clinical relevance of HTNEs, which could be beneficial to further dissect the mechanisms of breast cancer progression and facilitate the prediction, diagnosis, treatment, and prognosis of this malignancy.

2. Methods

2.1. Sample collection and RNA sequencing data processing

Frozen tumor tissues were collected from 199 breast cancer samples in the discovery cohort. In light of the manufacturer's protocol, total RNA exclusive of ribosomal RNA (rRNA) was extracted from tumor tissues using the VAHTS Total RNA-seq (H/M/R) Library Prep Kit for Illumina and immediately frozen in liquid nitrogen and stored at − 80 °C. The Ovation Human FFPE RNA-seq Library System (NuGEN Technologies, San Carlos, CA, USA) was used to construct the RNA-seq library and sequenced using 150 bp paired-end runs on the Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA). Raw Illumina sequence reads in FASTQ format was processed in a customized pipeline (Supplementary Table 8). For each sample, FastQC v0.11.9 was first used to quality control the raw reads, and Trimmomatic v0.39 was employed to excise low-quality bases and splice sequences, allowing mismatches at two positions when comparing splice sequences. Then reads were mapped to the human genome (GENCODE GRCh37.p13) using STAR v2.7.1a. Reads mapped to ribosomal RNAs, mitochondrial genome or chromosome Y were excluded from downstream analysis. The Y chromosome was removed because the breast cancer samples were from females and removing mitochondrial genome can reduce interfering or confusing signals [16,17]. Gene expression levels were quantified using featureCounts v2.0.3. BAM files were sorted and formatted using samtools v1.14 and bamCoverage v3.5.1.

2.2. The step-by-step framework for the identification of HTNEs

We developed a HTNE identification framework (Fig. 1 and Supplementary Fig. 1) using the following steps: (1) Genomic domains with reads per million (RPM) higher than transcriptional background levels were screened in each sample. The transcriptional background level was defined as the average read density across the nuclear genome (i.e., the sum of all RPM in a sample divided by the whole number of base pairs constituting the nuclear genome). The boundaries of candidate HTNEs across the genome were defined by the first and last nucleotides that meet the cutoff values. (2) The summit RPM of candidate regions was reserved if achieved a significance level with a local detection $P \leq 0.05$ compared to transcriptional noise. Transcriptional noise was defined by randomly selecting 1,000,000 nucleotide positions without the blocklist and formulating the distribution of their RPM to normal distribution. The blocklist included annotated exons from GENCODE and UCSC with two 500 bp flanking intervals, UTR regions from GENCODE and UCSC, 2000 bp upstream and 1000 bp downstream of the gene defined from GENCODE, FANTOM5 CAGE-defined TSS with two 500 bp flanking intervals, rRNA from UCSC, and genomic gap regions in UCSC Table Browser. Neighboring regions with genomic intervals within 100 bp were merged into novel candidate regions. (3) The regions where candidate HTNEs were located excluded the blocklist mentioned above, and splice junctions were combined from the SJ.out.tab files of STAR outputs. (4) The minimum size of the regions needs to be at least 100 bp. (5) Statistically significant expressions were calculated for candidate regions across all samples. The mean RPM values were calculated for each candidate region and the significance levels were estimated compared to the transcriptional noise observed in the transcriptional background of the samples. P values were calculated by comparing transcript levels to the distribution of transcriptional background levels for each sample. Next, for each candidate region, the number of samples selected at $P \leq 0.05$ among all samples was calculated, and the probability that the sample was selected was calculated using a binomial distribution with the overall probability set to 0.05. Bonferroni correction was applied to the regions that met the above criteria to ensure that were statistically significant. (6) Repetitive experiments were conducted, and the regions acquired for each repeat were overlaid. HTNEs were achieved until the interleaved amount converged to a stable level.

2.3. Validation datasets for HTNEs identification pipeline

The public dataset including 14 paired breast tumor and adjacent samples were derived from NCBI under accession code PRJNA739366. RNA-seq signals were obtained from merged RNA-seq signals of 199 samples of breast cancer. Peaks at high transcript levels in 13 different breast cancer cell lines representing the five major molecular subtypes of breast cancer detected by GRO-seq were from GEO under accession code GSE96859. CAGE-seq data of breast cancer cell line MCF-7 was obtained from GEO under accession code GSM979657. We assembled the alignment results of the breast cancer data to capture potential full-length transcripts for each sample using StringTie v1.3.4d and used proActiv v1.8.0 for alternative promoter start site prediction for our breast cancer cohort.

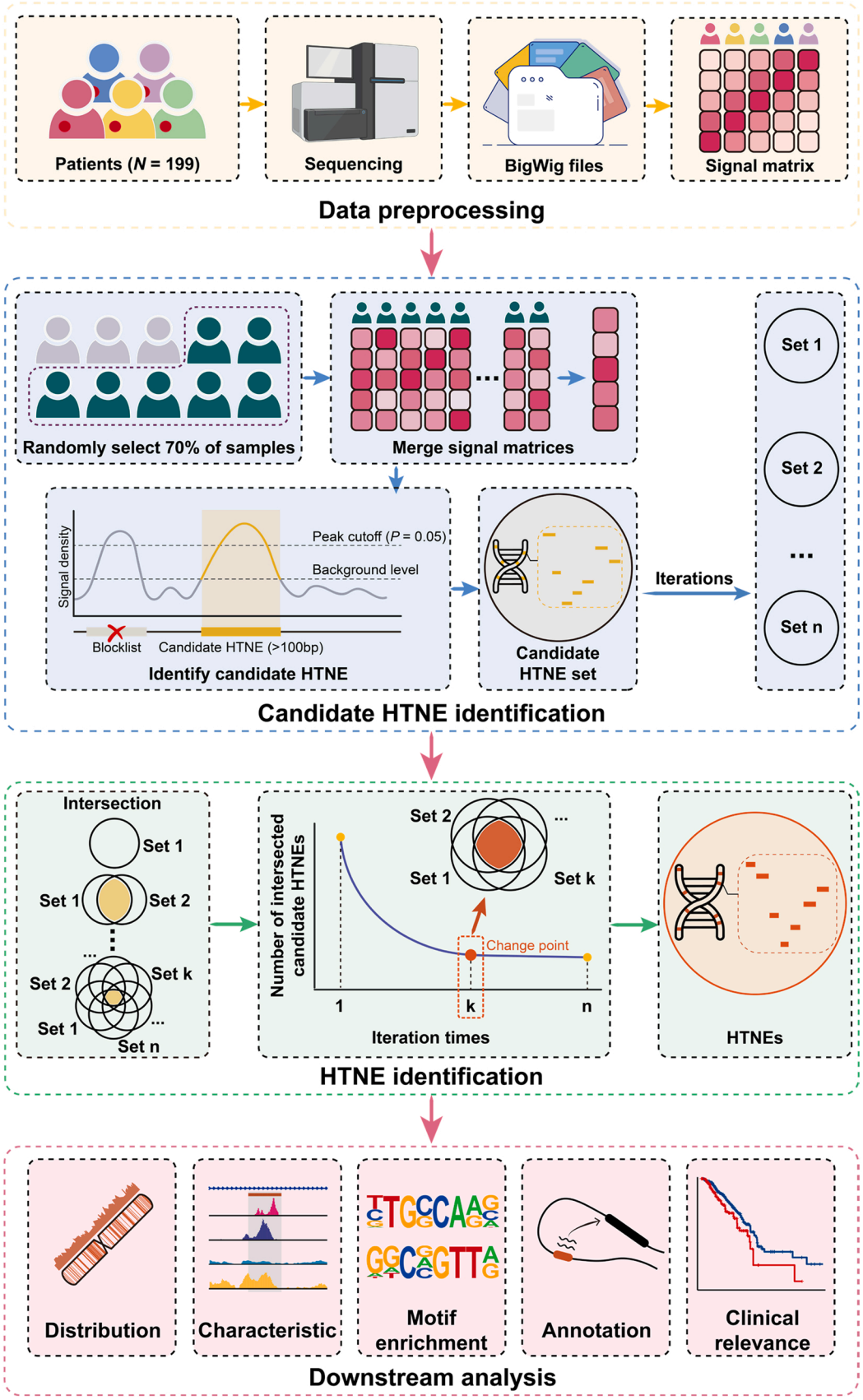

Fig. 1. The overview of our study. The framework includes sample collection and sequencing data preprocessing, HTNE identification, and downstream analysis. The optimal repeat k in HTNE identification section stands for the change point which was described exhaustively in the Methods and Supplementary Figure 1.

2.4. Randomly shuffled sequences and control regions construction procedure

The randomly shuffled sequences were generated using shuffleBed v2.30.0, and all identified HTNEs were used as input. The parameters were set to exclude the overlap of shuffled intervals and each shuffled sequence was on the same chromosome as the corresponding input sequence. For reproducibility of the experiment, the seed was chosen as 123. In the selection of control regions, for each intronic HTNE, all introns within 10 kb upstream and downstream of it were obtained, in which the HTNEs and the blocklist regions mentioned in the identification pipeline of HTNE were excluded, and for HTNEs in intergenic regions, the control regions were restricted to intergenic regions of the same chromosome using the selection approach of randomly shuffled sequences described above.

2.5. Regulatory annotations for validation of chromatin characteristics

To explore the possible roles of HTNEs in gene regulation, we characterized HTNEs with various known regulatory data in human breast or cell lines. We used chromHMM enhancer states in any of the three human breast tissues in the Roadmap Epigenomics Project [18] for histone-defined putative enhancers. Putative enhancers are marked as the E6, E7, or E12 states from the 15-state chromHMM segmentation defined by five core marks. The three breast tissues are breast myoepithelial, breast vHMEC mammary epithelial and HMEC mammary epithelial. We used histone modifications peak called in the above three breast tissues of Roadmap Epigenomics Project [18]. Other regulatory data include CAGE-defined enhancers from FANTOM5 Project [19], global run-on sequencing (GRO-seq) detected enhancers from Enhancer Atlas [20], and transcriptional coactivator P300 binding sites from ENCODE [21]. We also merged ATAC-seq data of breast cancer from TCGA [22] to validate the chromatin accessibility of HTNEs.

2.6. Motif enrichment analysis

We performed motif enrichment using the AME program from the MEME suite [23]. The optimal enrichment of the motifs was performed using one-tailed Fisher's exact test, and the *P* value was adjusted using the Bonferroni correction. RNA-binding motif enrichment analysis was performed using Ray2013 Homo sapiens, where 102 RNA-binding motifs were derived from in vitro experiments using the RNAcompete method. And miRNA motif enrichment analysis was performed using miRbase v22 Homo sapiens miRNA.

2.7. miRNA target genes collection

The target genes for miRNAs were retrieved from the "Target Expression" in miRDB [24], with "Source" being "breast carcinoma" and "expression level"$\geq$ 20.

2.8. Gene enrichment analysis and gene set variation analysis

For gene enrichment analysis, clusterProfiler [25] was used, while msigdbr was used to obtain hallmark gene sets and KEGG subset of curated gene sets from MSigDB [26]. The *P* values were adjusted using the Benjamini-Hochberg procedure. For gene set variation analysis, gene sets were downloaded from MSigDB [26] via keyword indexing. External gene counts for breast tumor and adjacent samples were extracted, filtered, and normalized using the R package TCGAbiolinks v2.24.3, and the scores for each sample were calculated using the R package GSVA v1.44.5.

2.9. The formula for the expression score

The formula for the expression score of each sample is

$$Expression \quad Score = \frac{\sum_{i=1}^{N} \log_2(TPM_i + 1)}{N}$$

where N denotes the number of total genes; i is to imply the order number of gene.

2.10. Classification of breast cancer samples based on HTNEs signatures

Integration of normalized counts of unique HTNEs in each subtype was conducted as signatures for clustering. Pooling of all breast cancer samples was performed using the *t*-distributed stochastic neighbor embedding (*t*-SNE) method, and k-medoid clustering was performed using the partitioning around medoids (PAM) algorithm. The coefficient of variation of normalized counts of all unique HTNEs in each sample was utilized to measure whether there was a significant difference between classes.

2.11. Statistical analysis

Continuous variables were compared by using the Wilcoxon signed-rank test and Categorical variables were compared using the hypergeometric test, permutation test or Fisher's exact test. Overall Survival probabilities were estimated using the Kaplan-Meier method and compared with the Log-rank test. The statistical significance threshold was set at $P < 0.05$. The Bonferroni-Holm (BH) correction was used in multiple hypothesis testing to decrease false positive rates. Statistical analyses were performed with R v4.2.1.

3. Results

3.1. Definition of a landscape for HTNEs in breast cancer

To genome-widely identify and characterize the landscapes of HTNEs in breast cancer, ribosomal RNA-depleted RNA sequencing was performed for 199 breast cancer patients. Beyond traditional mRNA sequencing, ribosomal RNA-depleted RNA sequencing could better capture noncoding RNAs without ploy A tails. Meanwhile, we established a sophisticated framework (see Fig. 1, Supplementary Fig. 1 and Methods for details) to genome-widely identify HTNEs based on ribosomal RNA-depleted mRNA profiles, evolving from an earlier study [15].

10,372 HTNEs were detected in 199 breast cancer samples (Fig. 2A), and the size distribution of HTNEs peaked at 248 bp (Fig. 2B). The substantial majority of HTNEs (8976, accounting for 86.54%) were located in intronic regions, of which 23.73% (2461) were in the first intron of the host genes, and the residual HTNEs were positioned in intergenic regions (Fig. 2C and Supplementary Fig. 2 A). Intronic HTNEs tend to be more abundant in the first half of the host genes than in the second half, with a gradual decrease from the 5′ to the 3′ end (Supplementary Fig. 2B–2C), following a similar distribution pattern to that of intronic TNEs in dopamine neurons of human brain [15]. Dong et al. [15] state that this distribution pattern opposes to that of partial RNA degradation, which preferentially degrades 5′ end and it implies that HTNEs could not only influence the chromatin state and accessibility of target genes, but also determine the gene regulatory activity and expression level of different genes [27]. Additionally, we observed that the length distributions of intronic HTNE and intergenic HTNE were similar (Fig. 1D), and the expression levels of intronic HTNE were significantly higher than those of intergenic HTNE ($P < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test, Fig. 1E).

To further validate the reliability of our identified HTNEs in breast cancer samples, we verified the enrichment of HTNEs with separate transcriptional signals using cap analysis of gene expression and deep sequencing (CAGE-seq), which can be used to identify all transcription

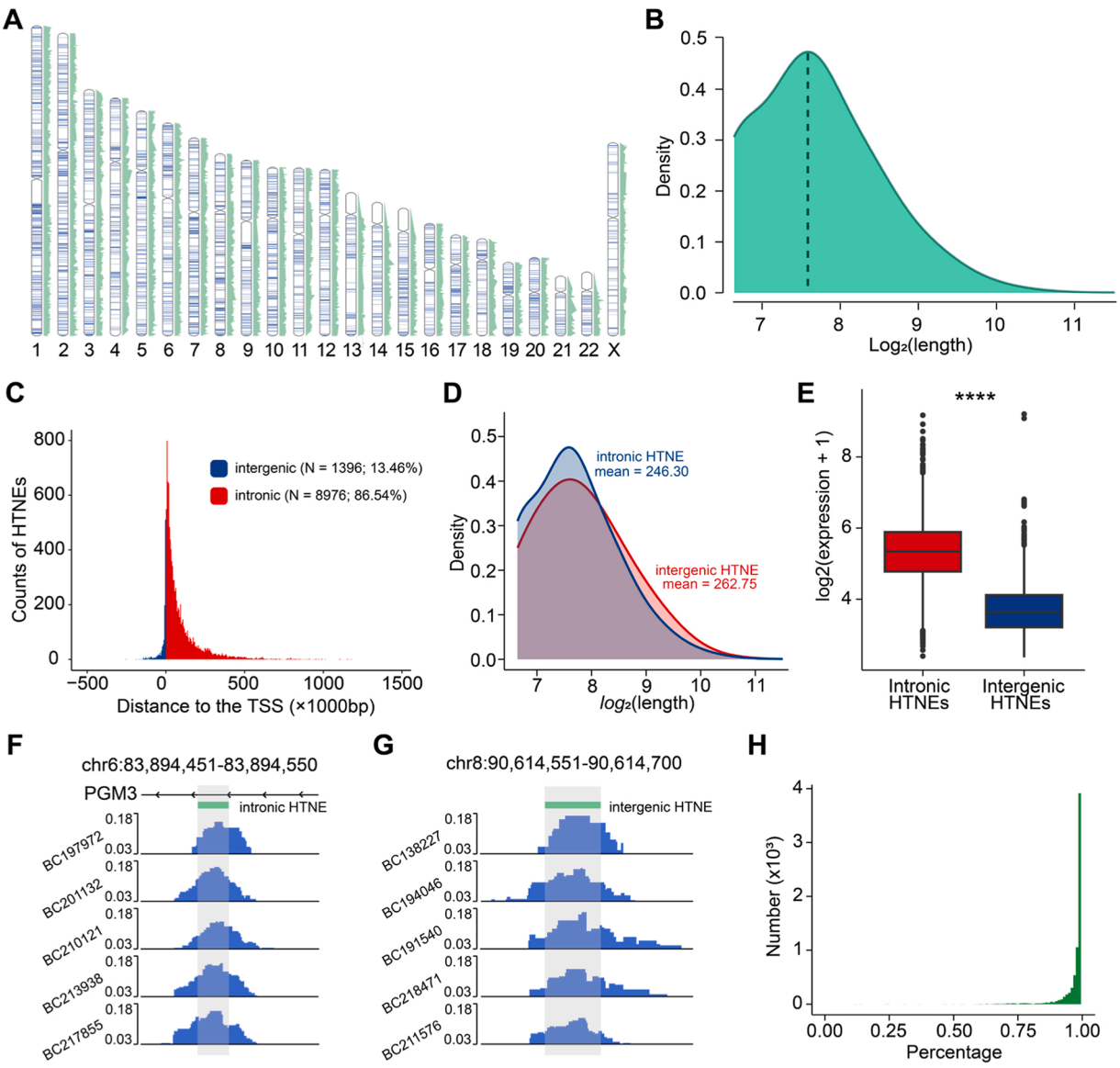

Fig. 2. Genome-wide identification of HTNEs in breast cancer. (A) Genome-wide overview of the location (blue) and the transcription levels of HTNEs (green) in 199 breast cancer samples; (B) Size distribution of HTNEs, where the peak is marked with a blue dashed line and the expression level is depicted in green; (C) Distribution of distances from the HTNEs to the host gene TSS (for intronic HTNEs) or to the nearest gene TSS (for intergenic HTNEs). Distances are marked positive (red) for intronic HTNEs and negative for intergenic HTNEs (blue); (D) Distribution of length for intronic HTNEs and intergenic HTNEs; (E) Comparison of expression levels between intronic HTNEs and intergenic HTNEs; (F-G) RNA-seq signals of HTNEs in five breast cancer samples; (H) Distribution of the percentage of samples in which HTNE was detected to be expressed among our breast cancer cohort, and the x-axis indicated the percentage of samples in our dataset where each HTNE was detected as transcribed.

start sites (TSS) in mRNA by identifying the cap site (Supplementary Fig. 2D), suggesting HTNEs are independent transcription units. Meanwhile, we also explored the correlation of the expression level between intronic HTNEs and their host genes and found no significant correlation ($r = 0.184$, $P < 2.2 \times 10^{-16}$, Pearson correlation analysis, Supplementary Fig. 2E). Besides, global run-on sequencing (GRO-seq) is a straightforward transcriptional measure, offering advantages over traditional bulk RNA-seq, that can derive the location and orientation of all actively transcribing RNA polymerases across the genome, facilitating the comprehensive identification of transcriptional functional elements in the transcriptomics of breast cancer cells [3]. Thus, we extracted GRO-seq data from 13 different breast cancer cell lines (GSE96859) which represent the five major molecular subtypes of breast cancer to validate the reliability of HTNEs. It was observed that up to 7368 (71.04%) of HTNEs exhibited elevated transcription levels in these breast cancer cell lines by comparing HTNEs with the peaks that exceeded the average transcription levels (Supplementary Fig. 2H). Furthermore, we compared transcription levels of these HTNEs between cancer and normal states based on a public dataset, including 14 breast cancer and adjacent samples [28]. 7237 out of 1,0372 HTNEs were observed to be expressed significantly higher in breast cancer compared to adjacent tissues, accounting for 69.77% of the total and it was evident that the discrepancy between breast cancer and adjacent tissues in the identified regions was significant ($P = 2.14 \times 10^{-25}$, Wilcoxon signed-rank test, Supplementary Fig. 2 F). Additionally, HTNEs did not overlap with any structural RNAs such as ribosomal RNA (rRNA), transfer RNA (tRNA), and small nuclear RNA (snRNA), small nucleolar RNA (snoRNA) and miscellaneous RNA (miscRNA) which have been found to be enriched in RNA-induced silencing complexes [29]. We also found that all HTNEs were not overlapping with alternative promoters [30] in our breast cancer cohort and 10,072 (97.11%) HTNEs were located entirely within potential full-length transcripts without isoforms

of protein-encoding genes. Ultimately, we also confirmed that HTNEs are indeed noncoding regions that are highly transcribed in our breast cancer samples (Fig. 2F-2 H and Supplementary Fig. 3A–3L). Taken together, these results validate our approach to identify HTNEs in breast cancer as independent highly reliable and highly expressed transcriptional regulatory elements.

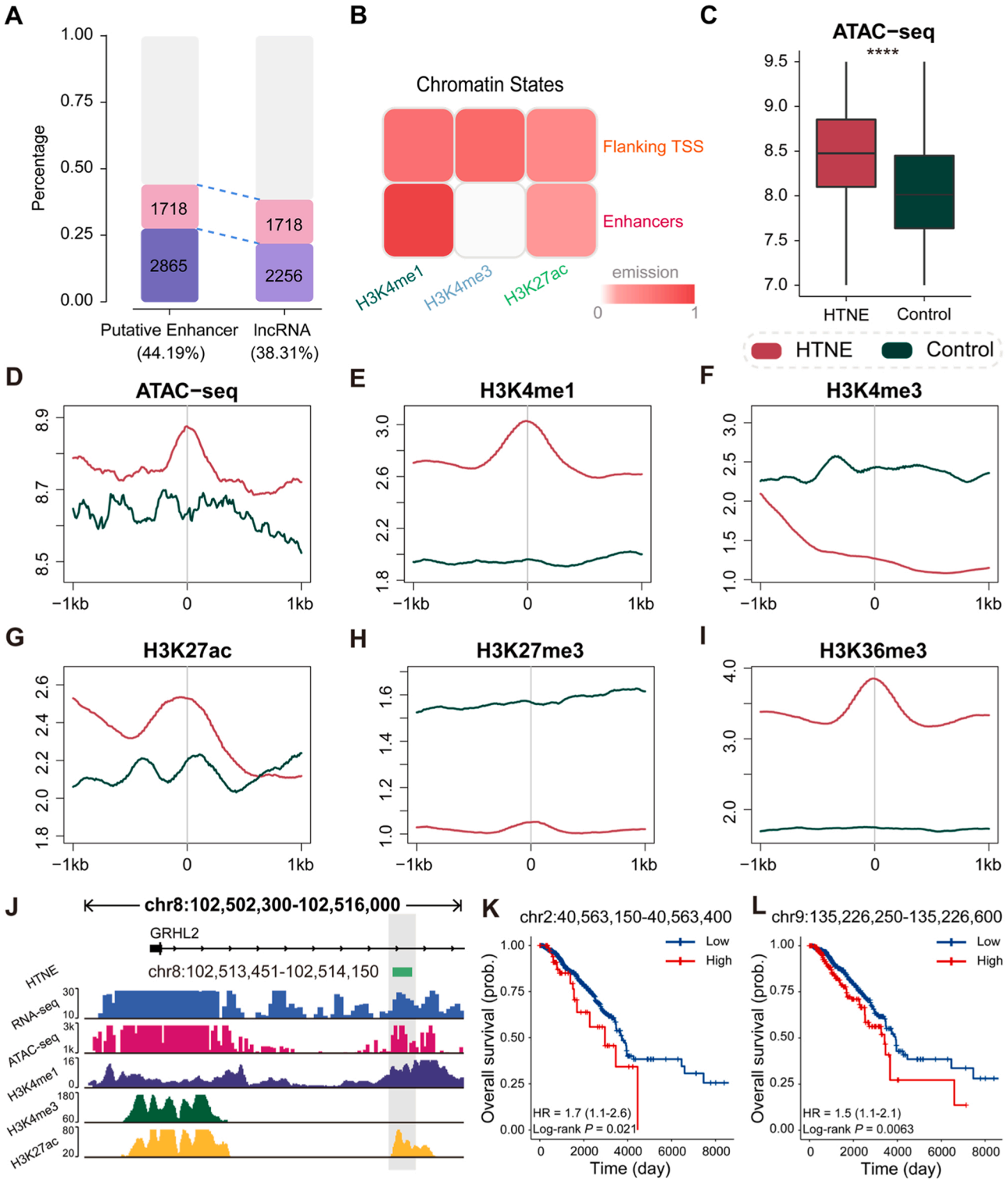

Fig. 3. Characterization of HTNEs based on epigenomic features. (A) Percentage of HTNEs that overlap with known putative enhancers and lncRNAs, and 1718 HTNEs between the dashed lines were overlapping with / identified as both enhancer and lncRNA; (B) Chromatin states of HTNEs, and the shade of red represents the mark probabilities of the corresponding chromatin states; (C-I) Comparison of chromatin accessibility and different histone modification marks between HTNEs and control regions at a distance of 1000 bp each upstream and downstream; (J) An example of the HTNEs located within GRHL2. GRHL2-HTNE with high enrichment of RNA-seq signals, ATAC-seq signals, H3K4me1 and H3K27ac signals, and depleted enrichment of H3K4me3 signals; (K-L) Kaplan-Meier survival plots show the prognostic relevance of two HTNEs (chr2:40,563,150–40,563,400 and chr9:135,226,250–135,226,600), respectively.

3.2. Chromatin characterization of HTNEs in breast cancer

To comprehensively characterize HTNEs in breast cancer, multi-omics profiles were employed to explore the regulatory roles that are attributed to HTNEs. 6839 (65.94%) of the 10,372 active HTNEs in breast cancer were coincided with preceding identified putative enhancers or long noncoding RNAs (lncRNAs) defined by one or more epigenomic features (Fig. 3A). Validation by comparing HTNEs with randomly shuffled regions revealed that HTNEs superimposed on regions with putative enhancer or lncRNAs were significantly higher than expected by chance alone ($P < 1 \times 10^{-9}$ by 1000,000,000 permutation test). Of the 10,372 identified HTNEs, 4583 (44.19%) were consistent with well-known characteristics of putative enhancers (Fig. 3A). These features incorporate genome-wide chromatin accessibility (such as ATAC-seq), characteristic histone modifications (such as H3K4me1, H3K27ac, and H3K4me3), CAGE defined putative enhancers, GRO-seq detected enhancers, transcription factor binding, and transcriptional coactivator P300 binding sites that were derived from Roadmap Epigenomics Project [18], FANTOM5 Project [19], ENCODE [21] and Enhancer Atlas [20]. Simultaneously, all HTNEs were compared with lncRNAs emerging from the high confidence lncRNA set (putative protein-coding genes are excluded) of LNCipedia (version 5.2) and the human lncRNA set of NOCODE (version 6.0). Of the 10,372 active HTNEs in breast cancer, 3974 (38.31%) were associated with known lncRNAs (Fig. 3A). Overall, 1718 HTNEs were both overlapped with putative enhancers and associated with lncRNAs (Fig. 3A).

We further explored the chromatin states of HTNEs and collected three histone modifications in three tissue and cell type groups mentioned above for the peaks of the imputed signal data. It is revealed that a large number of HTNEs overlapped with flanking TSS (Fig. 3B). Moreover, HTNEs overlapped with earlier identified putative enhancers were enriched with high levels of H3K4me1 and H3K27ac, as well as low levels of H3K4me3 (Fig. 3B). And ncRNAs can be identified and classified based on chromatin features, e.g., eRNAs are transcribed from activated enhancers with H3K4me1/H3K27ac marks and most of the mRNA-like long intergenic noncoding RNAs (lincRNAs) are generated from genomic regions with H3K36me3 marks [29]. Hence, we integrated chromatin accessibility and multiple histone modification marks to compare HTNEs and control regions for exploration of the prominent functions of HTNEs (see Methods for selection of control regions). As expected, HTNEs were enriched with higher level of chromatin accessibility than control regions, indicating that HTNEs are in transcriptionally activated states ($P < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test, Fig. 3C-3D). And HTNEs were enriched with the enhancer mark H3K4me1 and the active marker H3K27ac and exhibited low levels of H3K4me3, a mark of promoter, which were consistent with histone modifications in mammary epithelial cells from ENCODE (Fig. 3E-3 G). It was suggested that a group of HTNEs might be putative active enhancers that were consistent with the chromatin states. Moreover, transcription activation mark H3K36me3 and suppressive mark H3K27me3, which were formerly reported to characterize lncRNA in breast cancer [29,31,32], were also represented in identified HTNEs, implying that a fraction of HTNEs could be putative lncRNAs (Fig. 3H-3I). Simultaneously, comparing HTNEs with randomly shuffled sequences that differ from the control regions (see Methods) for ATAC-seq signals as well as various histone modification signals showed similar results (Supplementary Fig. 4A–4F, see Methods for details). The above various signals between all HTNEs, intronic HTNEs and intergenic HTNEs also showed similar patterns, respectively (Supplementary Fig. ^{4}G–^{4}L). To further explore whether HTNEs are associated with lncRNAs that contribute to breast cancer, we screened 197 experimentally validated breast cancer-related lncRNAs from LncRNADisease 2.0 [33] and found that nine HTNEs could overlap with functional domains of the lncRNA PVT1 which could promote breast cancer proliferation and metastasis [34].

To concretely clarify the function of HTNEs in breast cancer, five example regions were shown to highlight the chromatin characterization of HTNEs (Fig. 3J and Supplementary Fig. 5A–5D). For example, we spotlighted one of the HTNEs referred to as GRHL2-HTNE (chr8:102,513,451–102,514,150), located in the first intron of the human gene GRHL2 (Fig. 3) and its higher expression correlates positively with poorer survival in breast cancer patients ($P = 4.0 \times 10^{-4}$, log-rank test) [35]. GRHL2-HTNE harbors high levels of H3K4me1 and H3K27ac signals, low levels of H3K4me3 signal, and was predicted as a putative enhancer in Roadmap Epigenomics Project.

Since a part of HTNEs was formerly confirmed to be putative enhancers, the transcripts of these HTNEs are expected to be eRNAs. Following the analysis of the 'expression levels' of identified typical enhancers (eRNA expression/transcription levels) that overlap with HTNEs and clinical information from breast cancer samples in the TCGA-BRCA cohort [36], it revealed that breast cancer patients with high expression of these putative enhancer related HTNEs have a worse overall survival rate than the group with low expression of HTNEs (Fig. 3K-3 L, Supplementary Table 1 and Supplementary Fig. 6A–6G). Additionally, one of the clinically relevant HTNEs (chr5:137,868,250–137,868,450) has a host gene ETF1 that is also survival associated (Supplementary Fig. 6H). It suggests that HTNEs are related to the progression of breast cancer and potentially associated with breast cancer tumorigenesis, proliferation and migration of cancer cells, which could be used as potential prognostic markers or even prospective therapeutic targets for breast cancer.

3.3. Consensus motif analyses of HTNEs in breast cancer

HTNE transcripts could serve as decoys for various RNA-binding proteins involved in the control of gene expression and participate in the construction of regulatory networks of organisms in cancer biology [37]. To further investigate the roles of HTNEs derived from breast cancer on gene regulation, the enrichment of known motifs in HTNE sequences was analyzed. Consequently, 34 RNA-binding motifs significantly enriched in HTNEs, corresponding to 25 RNA-binding proteins (Supplementary Table 2). Of the 10,372 HTNEs identified, 10,345 (99.71%) were enriched with at least one RNA-binding motif and the majority were enriched up to 7 motifs (Fig. 4A). Among the 34 RNA-binding motifs significantly enriched, the motif termed in Ray2013 [38] as PCBP2 was ranking first with the most significant adjusted P value ($P = 1.12 \times 10^{-28}$, Fisher's exact test) and enriched by 4822 (46.49%) HTNEs. The survival analysis of PCBP2 expression status was performed using GEPIA2 [39] and a Kaplan-Meier survival curve was generated based on the data from TCGA and GTEx (Fig. 4B). It is shown that breast cancer patients with high PCBP2 expression have a worse overall survival rate than the low PCBP2 group ($P = 1.3 \times 10^{-2}$, log-rank test). And it was found that these 4822 HTNEs enriched by PCBP2 in our breast cancer cohort showed higher expression levels in high PCBP2 group than in low PCBP2 group (Fig. 4C, $P = 3.67 \times 10^{-3}$, Wilcoxon signed-rank test). Meanwhile, PTBP1 was enriched by the maximum number of HTNEs (4947, accounting for 47.70%), which were significantly upregulated in breast cancer samples compared to adjacent samples from TCGA ($P = 1.09 \times 10^{-34}$, Wilcoxon signed-rank test, Fig. 4D). It has been demonstrated that PTBP1 is a potential biomarker and molecular therapeutic target for breast cancer and overexpression of PTBP1 promotes the growth of breast cancer cells through the PTEN/Akt pathway and autophagy, thereby affecting the proliferation and migration of cancer cells [40,41].

Furthermore, noncoding elements could potentially function as competitive endogenous RNAs (ceRNAs) that regulate genes by competitively binding miRNAs [42]. ceRNAs can bind to miRNAs through miRNA response elements and thus affect miRNA-induced genes silencing, which play vital roles in pathological aspects associated with abnormal transcriptome changes (e.g., tumors). Of the 2656 motifs converted from Homo sapiens miRNAs in miRBase v22, the

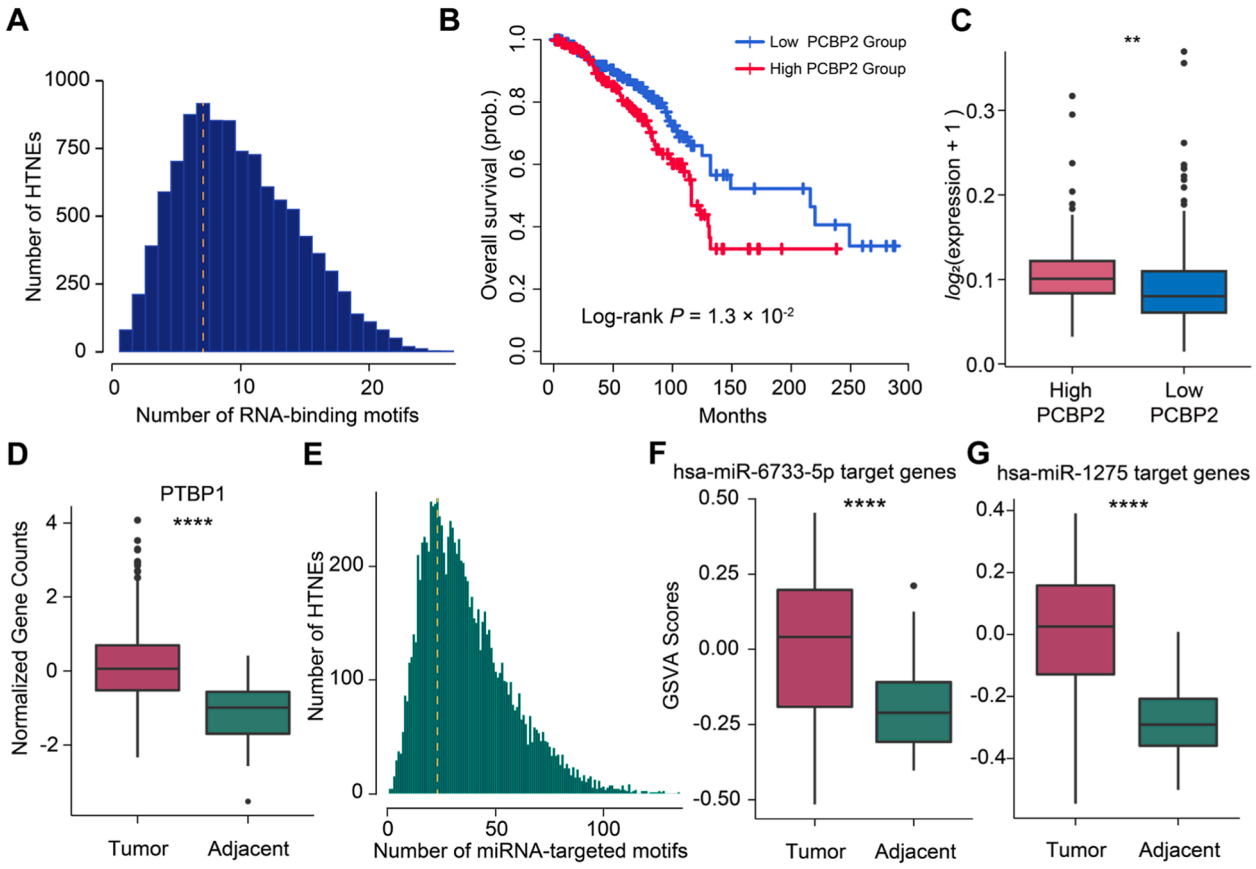

Fig. 4. Identification of consensus motifs in HTNEs. (A) Distribution of HTNEs enriched with the corresponding quantity of RNA-binding motifs. The maximum number of HTNEs enriched with 7 RNA-binding motifs is marked with a dashed line; (B) Kaplan-Meier survival plot of PCBP2 in TCGA breast cancer and adjacent samples; (C) Differences in expression level for high/low PCBP2 group; (D) Differences in normalized gene counts of PTBP1 between breast cancer and adjacent samples from TCGA breast cancer cohort; (E) Distribution of HTNEs enriched with the corresponding quantity of miRNA-targeted motifs. The maximum number of HTNEs enriched with 23 miRNA-targeted motifs is marked with a dashed line; (F-G) Differences in GSVA scores for the genes regulated by hsa-miR-6733-5p or hsa-miR-1275 in TCGA breast tumor and adjacent samples.

identified HTNEs were enriched for a total of 375 miRNA-targeted motifs, corresponding to 375 miRNAs (Supplementary Table 3). Evaluation of the miRNA-targeted motif enrichment analysis indicated that each HTNE was enriched with at least one motif, and the maximum number of HTNEs reaching 23 motifs (Fig. 4E). Of all 375 miRNA-targeted motifs enriched, the motif termed hsa-miR-6733-5p, a potential prognostic biomarker in breast cancer patients [43], was the most significant one ($P = 1.27 \times 10^{-18}$, Fisher's exact test) and enriched by 1805 (17.40%) HTNEs. 146 genes were regulated by hsa-miR-6733-5p in breast cancer from miRDB [24] (Methods) and found to be aberrantly overexpressed in the breast cancer samples ($P = 2.87 \times 10^{-19}$, Wilcoxon signed-rank test, Fig. 4F). In addition, hsa-miR-1275 was attracted by the most (2396, accounting for 23.10%) HTNEs ($P = 5.14 \times 10^{-9}$, Fisher's exact test). To demonstrate that HTNEs could potentially influence miRNA to regulate genes, we contrasted the expression levels of gene sets regulated by miRNAs recruited by HTNEs between breast cancer and adjacent samples. Then, 155 genes regulated by hsa-miR-1275 in breast cancer were retrieved from miRDB [24] (Methods). In comparison of gene set variation analysis (GSVA) scores between TCGA breast cancer and adjacent samples, genes regulated by hsa-miR-1275 were also abnormally overexpressed in breast cancer samples ($P = 5.39 \times 10^{-43}$, Wilcoxon signed-rank test, Fig. 4G). And hsa-miR-1275 is a biomarker for breast cancer and downregulation of hsa-miR-1275 is intimately relevant to biological mechanisms of breast cancer, including proliferation, invasion and metastasis [44].

3.4. Assignment of HTNEs to associated genes in breast cancer

To further investigate whether HTNEs are involved in the transcriptional regulation of genes in breast cancer, we assigned HTNEs to putative associated genes based on a baseline approach of target gene prediction using genomic distances. We procured 3734 HTNE-associated genes, including 3114 genes localized by intronic HTNEs and 620 genes nearest to intergenic HTNEs. And 1860 genes have more than one HTNE and each gene is mapped with approximately three HTNEs on average (Fig. 5A). And the expression scores of these genes were calculated for each sample (see Methods for details). Meanwhile, a group of 3734 genes excluded HTNE-associated genes was randomly selected and the expression scores were also calculated for comparison. As shown in Fig. 5B, the expression scores of HTNEs-associated genes were significantly higher than randomly selected genes ($P = 5.13 \times 10^{-67}$, Wilcoxon signed-rank test). Furthermore, the expression levels of genes associated with intronic HTNEs might be biased due to high expression levels of HTNEs. For intergenic HTNE associated genes, expression scores were also significantly higher than those of the randomly selected genes ($P = 1.09 \times 10^{-62}$, Wilcoxon signed-rank test, Fig. 5C). Via analyzing gene expression of tumor and normal tissues and clinical information of the samples from TCGA, there were 1420 genes associated with survival. Among them, 116 host genes were found to be associated with HTNEs ($P < 2.2 \times 10^{-16}$, Fisher's exact test), of which 42 host genes could be considered as independent prognostic factors separately from other genes (Supplementary Table 4).

eRNA is one of the markers of active enhancers, and it can also play a

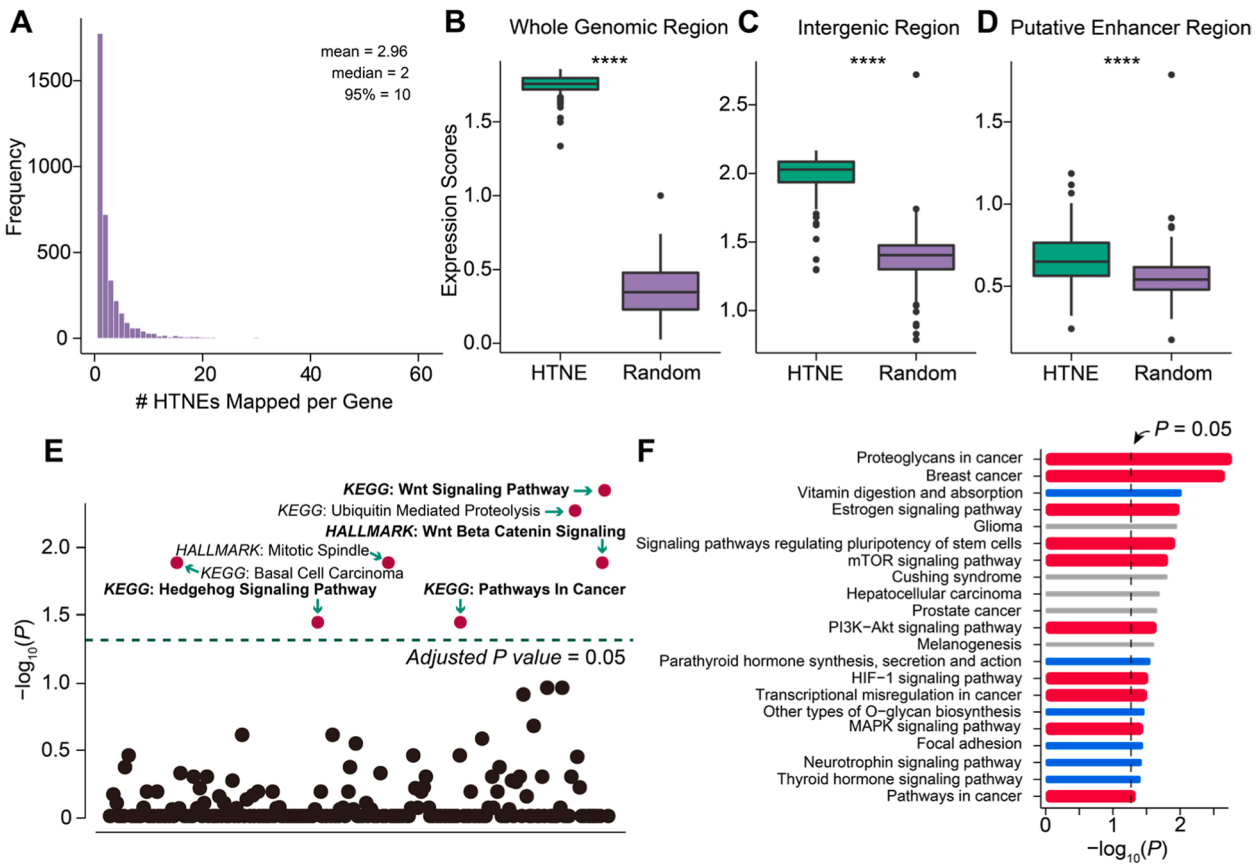

Fig. 5. Assigning HTNE associated genes in breast cancer. (A) Distribution of the number of HTNEs mapped per gene in breast cancer cohort; Comparison of expression scores between (B) all associated genes, (C) intergenic HTNE associated genes, (D) HTNE associated eRNA target genes and other randomly selected genes; (E) Enrichment analysis of HTNEs associated eRNAs targeted genes. (F) Enrichment analysis of HTNE associated genes set harbored SNPs in GWAS catalog. The red bars represent pathways associated with breast cancer, the blues for the potential pathways involved in breast cancer, and the greys for the pathways not yet documented as relevant to breast cancer to our best knowledge.

role in regulating cellular processes with the similar effect as lncRNA [45]. As mentioned before, a group of HTNEs was considered as putative enhancers and exhibited the evident characteristics of enhancers. To further explore the roles of eRNAs in the transcription products of HTNEs, the eRNA expression profile as well as eRNA target genes of breast cancer were extracted from the eRic database [46] and further compared with HTNEs. Consequently, 3099 eRNA target genes were detected, where these eRNAs intersected with identified HTNEs. These HTNEs associated eRNA target genes showed higher expression scores compared to randomly selected genes of the same scale ($P = 1.44 \times 10^{-15}$, Wilcoxon signed-rank test, Fig. 5D), suggesting the *cis*-regulatory functions of HTNEs in gene transcription. Subsequently, functional enrichment analysis was performed on 3099 HTNE associated eRNA target genes. Among all the pathways enriched by these genes, seven of them were statistically significant (Fig. 5E) and extensive studies have suggested that the majority of these pathways are associated with breast cancer progression and metastasis, including Wnt signaling pathway, ubiquitin mediated proteolysis, hedgehog signaling pathway and pathways in cancer [47–49] (Supplementary Table 5). For example, Wnt signaling pathway plays a principal role in controlling cancer progression and aberrant activation of Wnt signaling is observed from the onset of breast tumors to distant metastases [48]. Hedgehog signaling pathway has been implicated in tumorigenesis and progression of many cancer types [49].

Single nucleotide polymorphisms (SNPs), known as potential markers of carcinogenesis, are important genetic markers for the research of breast cancer characteristics [50,51]. Through integrative analysis of the SNPs in GWAS Catalog and the identified HTNEs, 252 diseases/traits associated SNPs localizing at HTNEs were obtained and 147 SNPs were retrieved from these diseases/traits, which are collectively associated with 124 genes. To explore the functional relevance of these genes, we performed functional enrichment analysis and found that all pathways were associated with cancers, and a majority of them were intimately associated with breast cancer (Fig. 5F). In particular, the pathway termed breast cancer was ranked second ($P = 2.16 \times 10^{-3}$) and the most significant pathway was proteoglycans in cancer ($P = 1.70 \times 10^{-3}$). It has been revealed that proteoglycans could activate essential cellular signaling pathways and drive proliferation, invasion and metastasis of cancer [52].

3.5. Subtype-specific transcriptional processes associated with HTNEs

PAM50 is a molecular typing criterion that is extensively applied to classify intrinsic subtypes of breast cancer based on gene expression of 50 genes [53]. According to PAM50, breast cancer can be classified into five molecular intrinsic subtypes: Luminal A, Luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, Basal-like, and Normal-like. Each of the five molecular subtypes varies depending on its biological characteristics and prognosis. To further determine if the HTNEs and their associated genes might be relevant to the biology of different subtypes of breast cancer, we employed PAM50 to classify 199 breast cancer samples into five categories (Fig. 6A). Due to the relatively small number of breast cancer samples belonging to Normal-like subtype, these 19 samples were excluded for downstream analysis to avoid excessive false positive results. The HTNE identification pipeline was re-performed on 180 retained samples covering four different subtypes.

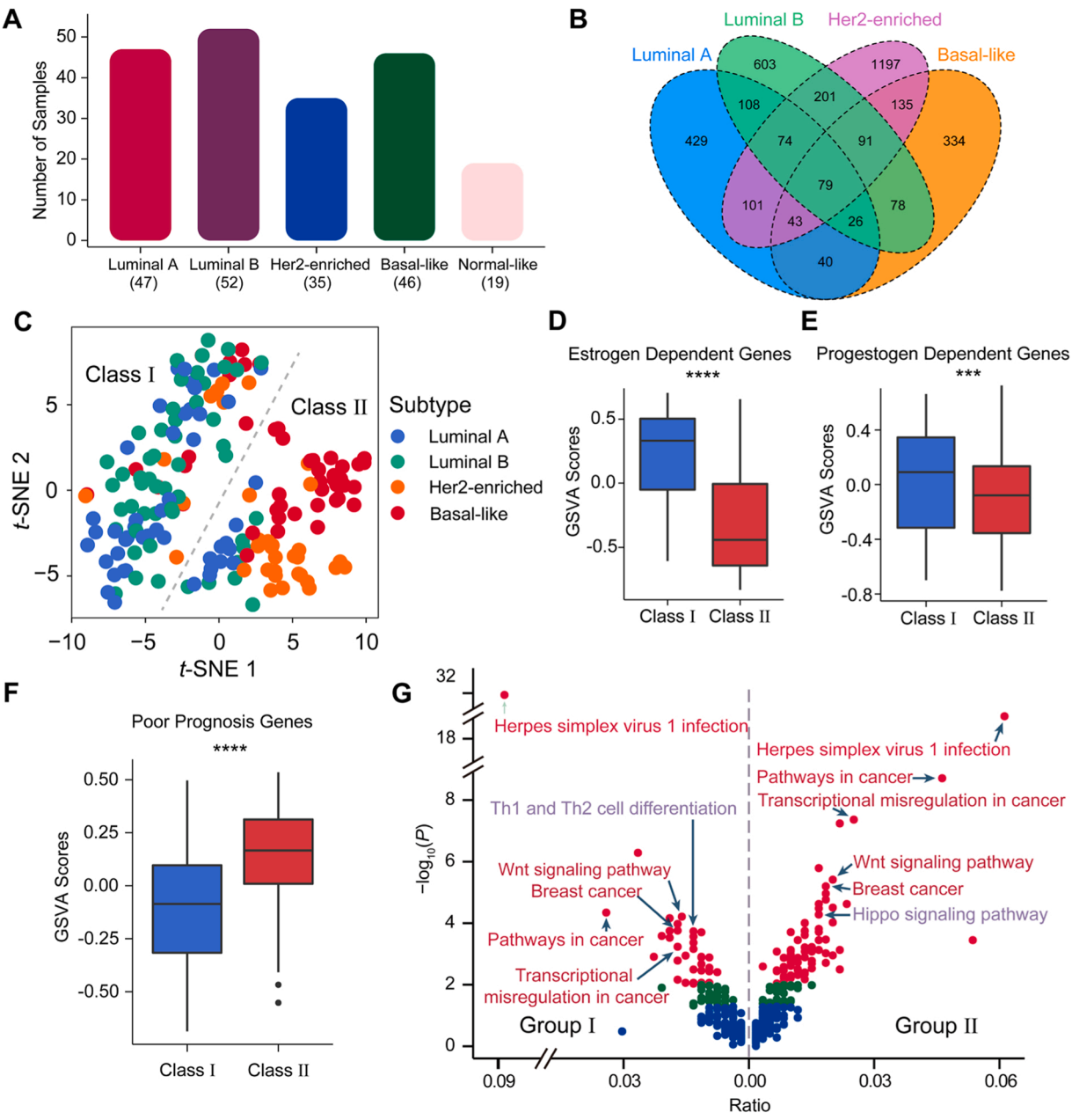

Fig. 6. Subtype specific transcriptional processes associated with HTNEs. (A) Number of samples for each subtype based on PAM50 typing criteria; (B) Comparison of HTNEs between four subtypes; (C) Classification of breast tumor tissues based on normalized counts of unique HTNEs for each subtype; (D-F) Differences of GSVA scores for the two classes in the estrogen-dependent gene set (M19439), the progestin-dependent gene set (M42918), and the poor prognosis gene set (M14693) from MSigDB database; (G) Enrichment analysis results for unique HTNEs associated eRNA target genes across two groups. The flanking pathways (purple) are specific to each group, and the middle pathways (red) are shared by both groups.

By comparing the identified HTNEs in each subtype, we detected 79 common HTNEs shared by four subtypes, of which 76 (96.20%) were recorded in the HTNE identification results of 199 breast cancer samples (Fig. 6B).

To evaluate the heterogeneity of HTNEs across four subtypes, the breast cancer samples were clustered based on the expression levels of 2563 unique HTNEs as to investigate whether these unique HTNEs of each subtype were subtype specific. It is depicted that patient samples were roughly divided into two classes (Class I chiefly included Luminal A and Luminal B, Class II mostly included HER2-enriched and Basal-like) ($P = 1.02 \times 10^{-61}$, Wilcoxon signed-rank test, Fig. 6C and Supplementary Fig. 7 A). To demonstrate that HTNEs are able to broadly distinguish the two classes mentioned above, we generated a ROC curve based on the expression pattern of unique HTNEs in both classes with an AUC equals to 0.7637 (Supplementary Fig. 7B), which indicated that HTNEs would be clinically relevant in distinguishing the various breast cancer subtypes. Information from preexisting researches suggested that Class I is a hormone receptor positive subtype of breast cancer (with positive estrogen receptor and progesterone receptor), whereas Class II is a hormone receptor negative subtype of breast cancer (with negative estrogen receptor and progesterone receptor) and associated with a high risk of metastasis and invasion [54,55].

Estrogen receptor alpha (ERα), also known as NR3A1, is one of the two main types of estrogen receptors, which are nuclear receptors activated by estrogen, and ERα is encoded by estrogen receptor 1 (ESR1) [56,57]. The gain and amplification of ESR1 correlate with ERα expression in breast cancer [58]. The comparison of the expression scores across the two classes revealed that ESR1 was significantly

overexpressed at higher levels in Class I than Class II ($P = 1.97 \times 10^{-16}$, Wilcoxon signed-rank test, Supplementary Fig. 7 C). Furthermore, phosphoserine aminotransferase 1 (PSAT1) is a protein-coding gene that catalyzes the serine synthesis pathway and independent of estrogen or progesterone. It is overexpressed in aggressive tumor types and is clinically associated with inferior distant metastasis free survival and overall survival in breast cancer patients [59,60]. As compared to the expression scores in Class II, the hormone independent PSAT1 gene was significantly under-expressed in Class I ($P = 1.13 \times 10^{-10}$, Wilcoxon signed-rank test, Supplementary Fig. 7D). Furthermore, to verify the differences between the two classes from a comprehensive perspective, GSVA was performed on the gene expression of two classes. And 21 pathways associated with breast cancer in MSigDB [26] were retrieved to compare whether the two classes were associated with hormone dependence or with cancer invasion and metastasis. It is evident that Class I is hormone-dependent, while Class II is associated with invasive and metastatic, where hormone-related genes are significantly down-regulated (Supplementary Table 6). For example, Class I has significantly higher GSVA scores in the estrogen receptor ($P = 2.75 \times 10^{-13}$, Wilcoxon signed-rank test, Fig. 6D) and progesterone receptor associated pathways than Class II ($P = 9.00 \times 10^{-3}$, Wilcoxon signed-rank test, Fig. 6E). Comparatively, Class II has significantly higher GSVA scores than Class I in the poor prognosis related pathway ($P = 3.21 \times 10^{-7}$, Wilcoxon signed-rank test, Fig. 6F).

Meanwhile, to further verify that the 2563 unique HTNEs exhibit subtype-specific features, we classified the samples into Group I (Luminal A and Luminal B) and Group II (HER2-enriched and Basal-like) based on PAM50 for comparison. Enrichment analysis of HTNE-associated genes unique in two groups revealed that Group I was significantly enriched in hormone dependent pathways, while Group II, in contrast, was absent from apparent hormone dependent pathways, consistent with PAM50 (Supplementary Fig. 7E). Moreover, KEGG enrichment analysis of unique HTNEs associated eRNA target genes in both groups (Fig. 6G, Supplementary Table 7) revealed that almost all pathways associated with breast cancer were observed in both groups. In addition, hormone dependent pathways were enriched in Group I (e.g., Th1 and Th2 cell differentiation) [61], whereas pathways in Group II were rarely relevant to hormone and preferred to be associated with cancer invasion and metastasis (e.g., hippo signaling pathway) [62,63].

4. Discussion

DNA methylation and mutations are early indicators of molecular abnormalities during carcinogenesis and potential biological markers for the early diagnosis of tumors, which are significantly associated with high transcription levels of genomic regulatory elements [64,65]. As a group of transcribed genomic elements without the ability to encode proteins, HTNEs exhibit particularly high rates of transcription in certain individuals. The massive transcription of HTNEs is functional and beneficial, since the expenditures in terms of transcription being wasteful are unreasonable [7]. Genomic elements with high transcript abundance have been demonstrated to be consistently and aberrantly highly expressed in the assays of diseases [66–68].

In this study, we collected and sequenced 199 breast cancer samples and developed a systematic and comprehensive framework to be the first time to identify HTNEs genome-widely in cancer samples. Given the heterogeneity of cancer, the methodology of Dong et al. [15] didn't consider that exception RNA-seq signals in a few samples could interfere with identifying TNEs, leading to an increase in false positives. In our study, the identification pipeline of HTNE will identify the candidate HTNEs multiple times and the intersection of candidate HTNEs will be defined as HTNEs, which could be considered as highly transcribed and highly reliable TNEs, as well as eliminating randomly transcribed regions (e.g., cryptic exons, uneven/noisy signals). The robustness and reasonability of the framework were validated, and it suggested that HTNEs are separate transcribed elements using CAGE-seq data, in particular that the expression levels of intronic HTNEs are weakly correlated with their host genes. Meanwhile, we discovered that more than seventy percent of the identified HTNEs overlapped with the peaks detected by GRO-seq of breast cancer cell lines, reflecting that a majority of the HTNEs we identified were truly transcribed elements, rather than introns that would be degraded in alternative splicing. And nearly seventy percent of the HTNEs we identified were significantly overexpressed in breast cancer compared to adjacent tissues, suggesting that most of identified HTNEs exhibited high expression levels in independent datasets as well. Finally, to further validate the reliability of HTNEs, we also ascertained that HTNEs are indeed highly transcribed noncoding regions in our breast cancer samples by examining the expression of HTNEs in breast cancer samples. As we expected, our method is reliable despite the differences arising from the strong heterogeneity of breast cancer.

We delivered mechanistic evidence that above two-thirds of the identified HTNEs were discovered to function as putative enhancers or lncRNAs in breast cancer based on multiple evidence tracks. We discovered that the identified HTNEs had obvious chromatin accessibility and exhibited the chromatin characteristics of activated enhancers (high H3K4me1 and H3K27ac signals, and low H3K4me3 signal). Furthermore, transcription activation signal H3K36me3 and suppressive signal H3K27me3, which were previously reported to characterize lncRNA in breast cancer [29,31,32], were also distinctly observed in the identified HTNEs. GRHL2-HTNE, one of the HTNEs, was distinctly marked as an active enhancer and has been demonstrated to be overexpressed in breast cancer and correlated with metastasis and poor prognosis in patients [69]. Following analysis of the expression levels of identified typical enhancers that overlap with HTNEs, it showed that the expression of HTNEs associated with putative enhancers was positively correlated with overall survival in breast cancer patients. These corroborate the view that a majority of HTNEs are either acting as putative enhancers or designated as lncRNAs that are specifically activated in breast cancer and may play valuable roles in the progression of breast cancer.

The transcript of HTNEs could recruit RNA-binding proteins which involved in the control of gene expression to participate in the formation of regulatory networks in organisms [37]. In our study, the sequences of HTNEs were discovered to be significantly enriched with 34 RNA binding motifs compared to shuffled sequences. More specifically, the top ranked motif in terms of significance was PCBP2 and the most frequently enriched motif was PTBP1. Clinical relevance analysis revealed that there is a negative correlation between high expression levels of RNA-binding motifs and poor overall survival. Earlier investigations have illustrated that these RNA-binding motifs are overexpressed in breast cancer and are associated with progression and worse prognosis [40,70]. More interestingly, we found HTNEs could potentially function as ceRNAs that regulate genes by competitively binding miRNAs. We discovered 375 miRNA-targeted motifs were enriched in the identified HTNEs, all HTNEs were enriched for at least one motif and most HTNEs were enriched for up to 23 miRNA-targeted motifs. In this regard, hsa-miR-6733-5p targeted motif was the most significant one of all motifs and hsa-miR-1275 targeted motif was the most frequently enriched one. By comparing gene expression levels of breast cancer and adjacent samples in the TCGA cohort, genes regulated by hsa-miR-6733-5p or hsa-miR-1275 were also aberrantly overexpressed in breast cancer samples. Antecedent studies have revealed that in terms of cell biological mechanisms down-regulation of hsa-miR-1275 leads to proliferation, invasion and metastasis of breast cancer [71,72]. Our findings suggested that HTNEs could competitively bind these miRNAs and inhibit their binding to target mRNAs, resulting in overexpression of miRNA-targeted genes relevant to the progression of breast cancer.

We found that there was a significant association between HTNEs and aberrant expression of oncogenes or tumor suppressors in breast cancer. Further, to exclude the case where high expression of intronic

HTNEs could influence expression levels of associated genes, we deliberately selected intergenic HTNE-associated genes, and as expected, those genes remained highly expressed. Moreover, as a subset of HTNE transcripts, eRNAs are the transcriptional products of active enhancers and can be used as markers of enhancer activity in particular cell types [73]. It is foreseen that target genes of eRNAs originating from HTNEs could also have statistically significant overall elevated expression levels. In the functional enrichment analysis, all pathways enriched by HTNE associated eRNA target genes were closely relevant to breast cancer tumorigenesis and metastasis. Wnt signaling pathway plays a principal role in controlling cancer progression and aberrant activation of Wnt signaling is observed from the onset of breast tumors to distant metastases [48]. Another enriched pathway, the hedgehog signaling pathway, has been implicated in tumorigenesis and progression of many cancer types [49]. We also integrated GWAS data to investigate the roles of genes regulated by HTNEs localized with SNPs relevant to the diseases/traits. In the functional enrichment analysis, the pathway breast cancer was ranked second in the enrichment analysis based on GWAS data, after the pathway termed proteoglycans in cancer. Prior studies have revealed that proteoglycans are heterogeneous glycoproteins and as a part of the extracellular matrix and cell surface, proteoglycans are simultaneously expressed in cells of the tumor microenvironment and on tumor cells [52,74]. Owing to interactions with other extracellular matrix proteins, growth factors and receptors, proteoglycans can activate essential cell signaling pathways (such as MAPK, Wnt, Hedgehog, TNF, TGF-β, etc.) and their targets are related to proliferation, angiogenesis and cell motility [52,74]. Specifically, as described in earlier studies, aberrant proteoglycans expression affects signaling pathways in breast cancer cells that drive proliferation and growth, insensitivity to anti-growth signals, evasion of apoptotic processes, unlimited replicative potential, tissue invasion, and metastasis [71,72].

Functional genomic elements tend to manifest strong heterogeneity across tumors and specificity in individual tumor subtypes, little is known about HTNEs in breast cancer [75,76]. In our study, HTNEs were also associated with breast cancer subtype-specific transcriptional processes and could cluster breast cancer samples significantly into two classes aggregated each with statistical biological significance, which could be used to stratify breast cancer patients into various clinical subtypes. Class I with positive hormone receptor principally comprises Luminal A and Luminal B samples, and the corresponding Class II, which is hormone receptor negative, chiefly contains HER2-enriched and Basal-like samples. We found that there were significant differences between these two classes in either individual gene or multiple pathways that were positively or negatively correlated with hormones receptor of breast cancer. Complementary to this, we also discovered that HTNEs could reflect breast cancer subtype-specific transcriptional processes consistent with PAM50 when we performed functional enrichment analysis of HTNE-associated genes in Luminal A and Luminal B samples as well as HER2-enriched and Basal-like samples based on PAM50. We discovered distinct pathways in the functional enrichment results of the two classes, including Th1 and Th2 cell differentiation pathway unique to Class I and hippo signaling pathway specific in Class II. Numerous clinical reports have indicated that estrogen induces the shift between Th1 and Th2 in the pathway named Th1 and Th2 cell differentiation [61]. Correspondingly, accumulating evidence suggests that the hippo signaling pathway could regulate the growth, metastasis, and drug resistance of breast tumor [62,63] and has not been reported as hormone dependent. In addition, we also found that the two classes shared a large number of pathways that are highly relevant to breast cancer, such as breast cancer, pathways in cancer, transcriptional misregulation in cancer and Wnt signaling pathway.

Although we performed bioinformatics analyses as comprehensively as possible, further experiments are still needed in the future to validate our conclusions. For example, to assign the target genes of HTNEs with more accuracy, further integration of Hi-C, HiChIP or Capture Hi-C is required to capture regions interacting with HTNEs. Additionally, there were 3533 (34.06%) HTNEs that did not intersect with known putative enhancers or lncRNAs, requiring more experiments to explore their potential biological implications. Moreover, the collected cohort was entirely Chinese breast cancer samples, lacking a large quantity of samples from other countries or regions for further supplementation. The small number of samples and the lack of control adjacent samples are also among the factors limiting the analysis in this study.

5. Conclusions

In conclusion, this study clarified that HTNEs identified in breast cancer samples are critical regulators in breast cancer progression. HTNEs are noncoding elements that are separately transcribed with highly reliable and highly transcribed features. Most HTNEs intersect with putative enhancers or lncRNAs with significant chromatin accessibility as well as histone modification characteristics. Besides, HTNEs can recruit RNA binding proteins or competitively bind miRNAs to participate in the control of gene expression and the formation of organismal regulatory networks, and they have a significant correlation with the aberrant expression of breast cancer oncogenes or tumor suppressors. HTNEs also showed clinical relevance in distinguishing between various breast cancer subtypes owing to their association with cancer subtype-specific transcriptional processes. Therefore, the investigation of HTNEs, a functional element with significantly high abundance in specific cohorts, will facilitate the dissection of mechanisms of breast cancer development, further facilitating the prediction, diagnosis and treatment of breast cancer. It is clear to foresee that its application in molecular diagnosis, disease phenotypic analysis and prognostic assessment would benefit breast cancer patients worldwide.

Funding

This work was funded by the Leading Technology Program of Jiangsu Province (BK20222008), the Natural Science Foundation of Jiangsu Province (BK20220823), and the National Natural Science Foundation of China (81830053, 62202098).

CRediT authorship contribution statement

W. Zhu. and H. Huang. contributed equally to this work and should be considered as co-first authors. X. Sun. supervised the conception of the work and revised it critically. X. Liu and Y. Liu. collected samples, and Y. Bai. and W. Gu. completed the RNA sequencing. W. Zhu. developed the protocol and performed the data analysis. H. Huang., W. Ming., R. Zhang., Y. Gu. and H. Liu. advised and helped in data analysis. W. Zhu. and H. Huang. developed the draft manuscript. All authors were involved in writing, reviewing and editing the manuscript, approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics statement

The studies involving human participants were reviewed and approved by the ethical committee of the First Affiliated Hospital of Nanjing Medical University. The participants provided their written informed consent to participate in this study.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability

The RNA-seq data of the cohort are available at https://ngdc.cncb.ac.cn/bioproject/ (BioProject number: PRJCA005965, GSA-Human number: HRA001100. The pipeline of HTNE identification was available via GitHub (https://github.com/weylz/HTNEseeker) and it includes full details of requirements and usage, as well as a demo example for guidance.

Acknowledgments

We thank professor Xianjun Dong from Harvard Medical School and Brigham & Women's Hospital, Boston, MA, USA, for his help in the identification of HTNEs.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.09.009.

References

[1] Cancer IA f R o. IARC Biennial Report 2020-2021. Lyon: International Agency for Research on Cancer; 2021.
[2] Xu S, Kong D, Chen Q, Ping Y, Pang D. Oncogenic long noncoding RNA landscape in breast cancer. Mol Cancer 2017;16:1–15.
[3] Franco HL, et al. Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. Genome Res 2018;28:159–70.
[4] Zhang X, Meyerson M. Illuminating the noncoding genome in cancer. Nat Cancer 2020;1:864–72.
[5] Polychronopoulos D, King JW, Nash AJ, Tan G, Lenhard B. Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Res 2017;45:12611–24.
[6] Braconi C, et al. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. Proc Natl Acad Sci 2011;108:786–91.
[7] Lee H, Zhang Z, Krause HM. Long noncoding RNAs and repetitive elements: junk or intimate evolutionary partners? TRENDS Genet 2019;35:892–902.
[8] Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nat Rev Cancer 2018;18:5–18.
[9] Slack FJ, Chinnaiyan AM. The role of non-coding RNAs in oncology. Cell 2019;179:1033–55.
[10] Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol 2020;27:521–8.
[11] Nakaya HI, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 2007;8:1–25.
[12] Calin GA, et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell 2007;12:215–29.
[13] Djebali S, et al. Landscape of transcription in human cells. Nature 2012;489:101–8.
[14] Corces MR, et al. The chromatin accessibility landscape of primary human cancers. eaav1898 Science 2018;362. eaav1898.
[15] Dong X, et al. Enhancers active in dopamine neurons are a primary link between genetic variation and neuropsychiatric disease. Nat Neurosci 2018;21:1482–92.
[16] Barshad G, Marom S, Cohen T, Mishmar D. Mitochondrial DNA transcription and its regulation: an evolutionary perspective. Trends Genet 2018;34:682–92.
[17] Scarpulla RC. Transcriptional paradigms in mammalian mitochondrial biogenesis and function. Physiol Rev 2008;88:611–38.
[18] Kundaje A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30.
[19] Lizio M, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 2015;16:1–14.
[20] Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res 2020;48:D58–64.
[21] Frankish A, et al. GENCODE 2021. Nucleic Acids Res 2021;49:D916–23.
[22] Chen H, et al. A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. e312 Cell 2018;173:386–99. e312.
[23] Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res 2015;43:W39–49.
[24] Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res 2020;48:D127–31.
[25] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a J Integr Biol 2012;16:284–7.
[26] Liberzon A, et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739–40.
[27] Sahu B, et al. Sequence determinants of human gene regulatory elements. Nat Genet 2022;54:283–94.
[28] Xu J, et al. TEM8 marks neovasculogenic tumor-initiating cells in triple-negative breast cancer. Nat Commun 2021;12:1–15.
[29] Sun Y-M, Chen Y-Q. Principles and innovative technologies for decrypting noncoding RNAs: from discovery and functional prediction to clinical application. J Hematol Oncol 2020;13:1–27.
[30] Nepal C, Andersen JB. Alternative promoters in CpG depleted regions are prevalently associated with epigenetic misregulation of liver cancer transcriptomes. Nat Commun 2023;14:2712.
[31] Kim JH, et al. Modulation of mRNA and lncRNA expression dynamics by the Set2–Rpd3S pathway. Nat Commun 2016;7:1–11.
[32] Wu SC, Kallin EM, Zhang Y. Role of H3K27 methylation in the regulation of lncRNA expression. Cell Res 2010;20:1109–16.
[33] Bao Z, et al. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res 2019;47:D1034–7.
[34] Cho SW, et al. Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. e1322 Cell 2018;173:1398–412. e1322.
[35] Uhlen M, et al. A pathology atlas of the human cancer transcriptome. Science 2017;357:eaan2507.
[36] Chen H, Liang H. A high-resolution map of human enhancer RNA loci characterizes super-enhancer activities in cancer. e705 Cancer Cell 2020;38:701–15. e705.
[37] Li L, et al. Multidimensional crosstalk between RNA-binding proteins and noncoding RNAs in cancer biology. Seminars in Cancer Biology 2021;75:84–96.
[38] Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013;499:172–7.
[39] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 2019;47:W556–60.
[40] Wang X, et al. PTBP1 promotes the growth of breast cancer cells through the PTEN/Akt pathway and autophagy. J Cell Physiol 2018;233:8930–9.
[41] He X, et al. Involvement of polypyrimidine tract-binding protein (PTBP1) in maintaining breast cancer cell growth and malignant properties. e84-e84 Oncogenesis 2014;3. e84-e84.
[42] Karreth FA, Pandolfi PP. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. Cancer Discov 2013;3:1113–21.
[43] Tang J, et al. Identification of miRNA-based signature as a novel potential prognostic biomarker in patients with breast cancer. Dis Markers 2019;2019.
[44] Majed SO, Mustafa SA. MACE-Seq-based coding RNA and TrueQuant-based small RNA profile in breast cancer: tumor-suppressive miRNA-1275 identified as a novel marker. BMC Cancer 2021;21:1–13.
[45] Kim T-K, Hemberg M, Gray JM. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. Cold Spring Harb Perspect Biol 2015;7:a018622.
[46] Zhang Z, et al. Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. Nat Commun 2019;10:1–12.
[47] Wang X, Yin Y, Fu Z. Abstract P2-03-20: Clinical profiling and comprehensive analysis of candidate genes related to breast cancer estrogen receptor intratumour heterogeneity. P2-03-20-P02-03-20 Cancer Res 2023;83. P2-03-20-P02-03-20.
[48] Xu X, Zhang M, Xu F, Jiang S. Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities. Mol Cancer 2020;19:1–35.
[49] Riobo-Del Galdo NA, Lara Montero A, Wertheimer EV. Role of Hedgehog signaling in breast cancer: pathogenesis and therapeutics. Cells 2019;8:375.
[50] Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Ann Oncol 2015;26:1291–9.
[51] Gao C, et al. SNP mutation-related genes in breast cancer for monitoring and prognosis of patients: a study based on the TCGA database. Cancer Med 2019;8:2303–12.
[52] Espinoza-Sánchez NA, Götte M. Role of cell surface proteoglycans in cancer immunotherapy. Semin Cancer Biol 2020;62:48–67.
[53] Parker JS, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160.
[54] Harbeck N, Gnant M. Breast cancer. Lancet 2017;389:1134–50.
[55] Waks AG, Winer EP. Breast cancer treatment: a review. Jama 2019;321:288–300.
[56] Dustin D, Gu G, Fuqua SA. ESR1 mutations in breast cancer. Cancer 2019;125:3714–28.
[57] Reis-Filho JS, et al. ESR1 gene amplification in breast cancer: a common phenomenon? Nat Genet 2008;40:809–10.
[58] Robinson DR, et al. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat Genet 2013;45:1446–51.
[59] Metcalf S, et al. The role of PSAT1 in triple negative breast cancer metastasis. 2845-2845 Cancer Res 2019;79. 2845-2845.
[60] Gao S, et al. PSAT1 is regulated by ATF4 and enhances cell proliferation via the GSK3β/β-catenin/cyclin D1 signaling pathway in ER-negative breast cancer. J Exp Clin Cancer Res 2017;36:1–13.
[61] Hong C-C, et al. Pretreatment levels of circulating Th1 and Th2 cytokines, and their ratios, are associated with ER-negative and triple negative breast cancers. Breast Cancer Res Treat 2013;139:477–88.
[62] Wang Y, et al. Comprehensive molecular characterization of the hippo signaling pathway in cancer. e1305 Cell Rep 2018;25:1304–17. e1305.
[63] Wang S, et al. Exosomes secreted by mesenchymal stromal/stem cell-derived adipocytes promote breast cancer cell growth via activation of Hippo signaling pathway. Stem Cell Res Ther 2019;10:1–12.
[64] Greenberg MV, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol 2019;20:590–607.
[65] Rheinbay E, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 2020;578:102–11.
[66] Martin MM, et al. Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res 2011;21:1822–32.

[67] Gemmell P, Hein J, Katzourakis A. The exaptation of HERV-H: evolutionary analyses reveal the genomic features of highly transcribed elements. Front Immunol 2019:1339.
[68] Boivin V, et al. Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. Rna 2018;24:950–65.
[69] Kumegawa K, et al. GRHL2 motif is associated with intratumor heterogeneity of cis-regulatory elements in luminal breast cancer. NPJ Breast Cancer 2022;8:70.
[70] Wang X, et al. PCBP2 posttranscriptional modifications induce breast cancer progression via upregulation of UFD1 and NT5EPCBP2 alternative splicing and polyadenylation in BrCa. Mol Cancer Res 2021;19:86–98.
[71] Cox TR. The matrix in cancer. Nat Rev Cancer 2021;21:217–38.
[72] Insua-Rodríguez J, Oskarsson T. The extracellular matrix in breast cancer. Adv Drug Deliv Rev 2016;97:41–55.
[73] Core LJ, et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet 2014;46:1311–20.
[74] Theocharis AD, et al. Insights into the key roles of proteoglycans in breast cancer biology and translational medicine. Biochim Et Biophys Acta (BBA)-Rev Cancer 2015;1855:276–300.
[75] Zhang H, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet 2020;52:572–81.
[76] Nik-Zainal S, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 2016;534:47–54.