Integration of RNA-seq and Ribosome Profiling to Characterize Transcriptional and Translational Estrogen Responses of Genes in Human Breast Cancer

Siew Woh Choo (cwoh@wku.edu.cn)
Department of Biology, College of Science and Technology, Wenzhou-Kean University, 88 Daxue Road, Ouhai, Wenzhou, Zhejiang Province, China 2Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China

Yu Zhong
Xi'an Jiaotong-Liverpool University https://orcid.org/0000-0001-5428-1774

Edward Sendler
Wayne State University

Anton Scott Goustin
Wayne State University

Juan Cai
University of Michigan

Donghong Ju
Wayne State University

James Bentley Brown
Lawrence Berkeley National Laboratory

Mary Ann Kosir
Wayne State University

Leonard Lipovich (llipovich@med.wayne.edu)
Wayne State University

Research

Keywords: breast cancer, estrogen, RNA-Seq, Ribo-Seq, long non-coding RNA (lncRNA), pseudogenes

Posted Date: September 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74119/v1

Abstract

Background

Estrogen is a hormone that is frequently essential in breast cancer to drive key transcriptional programs by interacting with the estrogen receptor alpha that upregulates proliferative and oncogenic genes and represses apoptotic and tumor suppressor genes. Protein-coding targets of estrogen regulation in breast cancer are well-defined. However, long non-coding RNA (lncRNA) genes account for the majority of human gene catalogs. The coding status of these genes – their accidental, or regulated, translation by ribosomes, under the influence of estrogen – remains a controversial topic.

Methods

Here, we performed comprehensive transcriptome analysis using RNA-Seq, as well as ribosome profiling using Ribo-Seq, on the same samples: biological replicates of human estrogen receptor alpha (ERa) positive MCF7 breast cancer cells before and after estrogen treatment. We correlated these two datasets, globally highlighting protein-coding and lncRNA differentially expressed genes and transcripts that were positively as well as negatively responsive to estrogen, separately at the transcriptional level and the translational (as approximated by ribosome binding) level.

Results

Our data showed that some transcripts were more robustly detected in RNA-Seq than in the ribosome-profiling data, and vice versa, suggesting distinct gene-specific estrogen responses at the transcriptional and the translational level, respectively. Certain differentially expressed transcripts may point to the regulation of alternative splicing by estrogen. Several pseudogenes were co- and anti-regulated with their cancer-functional parental genes. Gene ontology analysis highlighted cancer-relevant pathways enriched after estrogen treatment in cells.

Conclusions

Our study represents a significant advance in the estrogen receptor biology, because we demonstrated global effects of estrogen on splicing and translation that are distinct from, and not always correlated with, its effects on transcription, and that differ globally for protein-coding and lncRNA genes. We have also highlighted for the first time the transcriptional and translational response of expressed pseudogenes to estrogen, pointing to new perspectives for biomarker and drug-target development for breast cancer in future.

Background

Breast cancer is the most prevalent cancer for women in the world, with over two million case events and approximately 630,000 deaths in 2018 (Bray et al., 2018). Clinically, several different subtypes of breast cancer have been defined (Russnes et al., 2017): hormone receptor positive (includes estrogen receptor, i.e. ER, positive, and progesterone receptor positive, breast cancer), HER2 positive, and Basal. Some basal breast cancers are triple negative (no ER, PR, or HER2), though some triple negatives are not basal (Rakha et al., 2009; Gazinska et al., 2013)

Approximately 80% of breast cancers are estrogen-receptor (ER) positive, and these are of enormous clinical significance worldwide. Estrogen (17-β-estradiol; abbreviated as E2) activates the estrogen receptor α (ER). ER translocates from the cellular surface to the cytoplasm and then to the nucleus, where it directly regulates its target genes. In ER positive cancer cells, ER transcriptional regulation features massive activation of cell-proliferation genes, including oncogenes, and suppression of pro-apoptotic genes, ultimately stimulating the proliferation of cancer cells.

Moreover, ERα dominates overall ER activity, and is expressed in in most breast cancers, underscoring its significant role in both oncogenesis and progression (Ali & Coombes, 2000). Interrogating the effects of E2-induced ERα-activation on the gene regulatory program in breast cancer cells can improve our understanding of breast cancer, and can point to more effective, less side-effect-ridden, more individualized potential therapies in the future.

High-throughput RNA sequencing (RNA-Seq) provides precise quantification of gene expression in living cells (Conesa et al., 2016) and yields accurate and comprehensive transcriptome profiles of tumors and normal cells, and helps to define and characterize gene-regulatory responses to treatments (Lin et al., 2018). Intriguingly, the correlation of mRNA and protein levels in mammalian cells is imperfect, indicating that the study of gene regulation solely from RNA-Seq data is inherently insufficient to understand the adaptive landscape of cells to stimuli (Haider & Pal, 2013). Therefore, here we deployed distinct but complementary approaches that distinguish transcribed RNAs from those that are bound by ribosomes and translated. This technique provides quantifications of both transcriptional and post-transcriptional responses to estrogen in MCF7 cells, one of the most commonly used human ER-positive breast cancer cell line model systems. Ribo-Seq analysis is based on the deep sequencing exclusively ribosome-protected RNA fragments that are presumably being translated, filling the gap between transcriptome and proteome (mass spectrometry) data (Ingolia et al., 2009), so as to better understand the functions and regulatory responses of specifically the ribosome-bound, presumed-to-be-translated, fraction of the transcriptome (Calviello & Ohler, 2017).

The refinement of the human gene catalogs based on transcriptome data, such as the Gencode effort (www.gencodegenes.org), has revealed that protein-coding genes comprise only less than one-third, of human genes, while the remainder of genes are non-coding RNA genes, which include long non-coding RNA (lncRNA) genes (Derrien et al., 2012). LncRNA genes are as abundant as protein-coding genes in mammalian genomes, but aside from a hundred or so mechanistically well-characterized examples that point to diverse, epigenetic and post-transcriptional, gene-specific and global, nuclear and cytoplasmic as well as extracellular, positive and negative mechanisms that differ drastically between different lncRNAs, their functions remain surprisingly obscure. Classically, lncRNAs were assumed to not be translated, and our prior work found that translation is extremely rare in two human cell line models using deep shotgun proteomics (Banfai et al., 2012). However, many lncRNAs are cytoplasmic, and if ribosome entry sites and open reading frames are present, then ribosomes might not be able to distinguish these non-coding transcripts from conventional messenger RNAs, and might therefore translate proteins from short open reading frames in some lncRNAs. In fact, a commonly used computational definition of lncRNAs stipulates that these transcripts may contain small Open Reading Frames (smORFs) of less than 100 amino acids, provided that such smORFs lack sequence homologies to known proteins (Dinger et al., 2008). Furthermore, all cytoplasmic RNAs, including non-coding RNAs, may be accessed by ribosomes during the “pioneer round of translation,” a key step of the Nonsense-Mediated Decay (NMD) mechanism (Ishigaki et al., 2001). In summary, the possibility of ribosomal translation of lncRNAs, and the potential for regulation driven by response to hormones, cannot be excluded.

Pseudogenes are truncated or full-length copies of functional genes that have been generated through retrotransposition and/or gene duplication mechanisms over evolutionary time (Zheng et al., 2007; Ruan et al., 2007). Formerly thought to be non-functional, pseudogenes are now increasingly recognized as important functional components of the transcriptome, and they can be sequence-specific regulators of their parental genes, such as antisense lncRNA transcription from a pseudogenic locus that regulates the parental gene (Johnsson et al., 2013; Milligan & Lipovich, 2014; Milligan et al., 2016). Global assessments of mammalian transcriptomes in the post-genomic era have revealed that many pseudogenes are not transcriptionally silent, and that some are indeed robustly transcribed. Here, we globally examine the transcriptional and translational response of expressed pseudogenes to estrogen.

We recently showed that estrogen-responsive lncRNAs regulate proliferation, viability, and death in breast cells during estrogen response (Lin et al., 2016). Transcribed pseudogenes are diagnostic as well as prognostic markers in cancer (Poliseno et al., 2015), and the role of pseudogenes in estrogen response merits further and directed study. Therefore, we set out to identify, and study the functions of the differentially expressed genes and transcripts, with an emphasis on alternative splicing, lncRNAs, and pseudogenes, in breast cancer cells before and after estrogen exposure, and to correlate transcriptomics with ribosome-profiling data, to identify mRNA and lncRNA genes with exclusively/primarily transcriptional and exclusively/primarily translational responses to estrogen, and to distinguish them from the majority of genes that had well-correlated transcriptional and translational responses to estrogen stimulation.

Methods

Cell culture and RNA extraction

Human MCF-7 breast cancer cell line (ER positive) was cultured in Dulbecco's modified Eagle's medium (DMEM, Fisher, USA) given with 1% penicillin and 10% FBS under 37 °C in 5% CO_2. E2-starved cells were conducted within 5% charcoal-stripped FBS in phenol-red-free DMEM for 72 h and treated with 10 nM 17-β-estradiol (E2). Control groups were only treated with the ethanol vehicle. Cells were collected after 48 h of experimental treatment and total RNA was harvested by using an RNAeasy Mini kit following the manufacturer's protocols (Qiagen, Valencia, CA). Two biological replicates of RNA-Seq and three biological replicates of Ribo-Seq were sampled both before and after the estrogen treatment.

Library preparation and high-throughput sequencing

RNA-Seq stranded libraries were prepared with Illumina Truseq kit, with a paired-end sequencing strategy (150-nt). For Ribo-Seq, the stranded libraries were prepared with Illumina Truseq kit, with paired-end (40-nt), based on a previously described robust ribosome profiling technique (Ingolia et al., 2009; Ingolia et al., 2011). All samples were sequenced with Illumina NextSeq sequencer to acquire paired-end sequencing reads. All raw sequence reads can be accessible from the NCBI's Sequence Read Archive SRA database (BioProject ID: PRJNA639213).

Data pre-processing

The quality of all raw reads generated from the transcriptomic and ribosome-profiling libraries were evaluated using a FASTQC v0.11.7 software (Andrews, 2018). Read filtering was performed using FASTP v0.19.7 (Chen et al., 2018), based on the following criteria: Adaptor sequences within the reads were automatically filtered out; reads with global quality score < Q20 for RNA-Seq and 15 for Ribo-Seq were discarded; nucleotides from 5′ or 3′ with quality score < Q20 were trimmed from the read.

Read alignment

The pre-processed reads from the RNA-Seq and Ribo-Seq data were aligned to the reference human genome sequence (hg19 assembly) using HISAT2 v2.1.0 (Kim et al., 2015). For RNA-Seq data, default parameters were used for HISAT, with the exception of setting strandedness to be strand-specific to the first strand. For Ribo-Seq data, the first of paired reads was aligned to reference genome with unstranded-specific option, with the mate read discarded due to the redundant nature derived from two overlapping reads. All SAM files derived from HISAT2 were converted into the BAM files using SAMTOOLS v1.9 for downstream analyses (Li et al., 2009), with multi-mapped reads specifically removed from Ribo-Seq alignments due to the reduced specificity of these short (~ 30nt) reads.

Normalization and expression level measurement

For gene level counts of both RNA-Seq and Ribo-Seq, the pre-processed reads from all samples mapping to human exons within gencode.v19.annotation.gtf were counted using the “summarizeOverlaps” function in R package: GenomicAlignments (Lawrence et al., 2013). The counting mode of “union” was used for both RNA-Seq and Ribo-Seq and un-normalized gene counts were normalized using DEseq2 package (Love et al., 2014).

For transcript level analysis, read counts mapping to individual transcripts were assigned by Cuffdiff2 v2.2.1 (Trapnell et al., 2012). Raw counts were converted into normalized FPKM (fragments per kilobase of exon per million fragments mapped) expression levels, followed by scaling via the median of the geometric means of fragment counts across all samples.

Differential expression analysis

To identify differentially expressed genes (DEGs), DEseq2 was employed, using raw gene counts. Genes with |log2 FC|>1.0 and P_{adj}<0.05 were considered as significantly DEGs. To study differential expression between the estrogen treatment and the control (no estrogen treatment) groups at the transcript level in both RNA-Seq and Ribo-Seq, Cuffdiff2 was used to determine individual transcript abundance based on the assignment of reads to most likely transcript models. Transcripts showing |log2 FC|> 1.0 and q< 0.05 were considered as significantly differentially expressed transcripts (DETs).

Quantitative correlation of RNA-Seq and Ribo-Seq

Lowly expressed genes are automatically filtered out by “independent filtering” function embedded within the DEseq2 package. The presence of discrepancy in sequencing depths between the RNA-Seq and Ribo-Seq could lead to the inaccurate analysis, therefore we performed statistical adjustments on fold changes in terms of the expression level made to two datasets in order to reduce this gap. To compare fold change at the gene level between the transcriptomic and ribosome profiling datasets, we took advantage of SE (standard error) of fold change provided by DEseq2. A differential fold change (between RNA-Seq and Ribo-Seq) was calculated if unadjusted fold change in RNA-Seq is more than in Ribo-Seq, as ($FC_{RNA-Seq}$- 1.0*SE) - ($FC_{Ribo-Seq}$+ 1.0*SE). Genes with a cutoff of differential fold change of 2.0 and P_{adj}< 0.05 were considered to be expressed more in RNA-Seq than in Ribo-SEq. The corresponding genes whose unadjusted fold change in Ribo-Seq is more than in RNA-Seq were calculated in a similar manner.

Differential pseudogenes expression analysis

To identify putative pseudogenes, we used very stringent criteria by only considering uniquely mapped reads in both RNA-Seq and ribosome-profiling datasets. All multi-mapped reads were discarded from the original bam files using SAMTOOLS, followed by the re-counting of read mapping to exons. Differential gene expression analysis was performed using DEseq2 as described above in order to obtain the differentially expressed pseudogenes (DEPs).

To identify the parental genes of each pseudogene, the psiDR database (Pei et al., 2012) was used. The parental gene name of candidate pseudogenes defined by in psiDR database to be extracted for downstream analyses. For those remaining genes whose parental gene could not be defined by this method, NCBI-BLASTN (Altschul et al., 1990) was employed to search for potential parental genes using a e-value cutoff of $10^{-4}$and a sequence similarity > 80%.

Due to the high sequence similarity between the pseudogenes and their parental genes, we also manually examined and visualized the mapped reads to pseudogenes using the UCSC Genome Browser (Kent et al., 2002) in order to ascertain that the reads are accurately mapped with greater confidence to pseudogenes rather than their parental genes. After aligning these reads to genome using BLAT software (Kent, 2002), was used to confirm that these reads

mapped uniquely to pseudogenes or with a higher similarity to the pseudogenes compared to their parental genes (Fig. 5).

Functional and KEGG enrichment analysis

Gene Ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed for up-regulated genes with $\log_2$ (fold change) > 1.0 and P_{adj}< 0.05 and down-regulated genes with $\log_2$(fold change) <-1.0 using Metascape (Zhou et al., 2019). The significantly enriched GO terms and KEGG pathways were identified with P_{adj}< 0.05.

Results

Differential Gene Expression Analysis

To investigate gene regulation during estrogen response, differential gene expression analysis at the gene level was performed on the transcriptome data as well as ribosome-profiling datasets. We identified 1162 transcriptionally up-regulated genes after estrogen treatment, mainly consisting of 989 protein-coding genes and 97 lncRNA genes, in RNA-Seq data, compared to 659 genes with up-regulated ribosome binding after estrogen treatment, mainly consisting of 567 protein-coding genes and 15 lncRNA genes, in Ribo-Seq data (Table 1; **Supplementary Tables 3 & 4**). Similarly, we identified 1286 transcriptionally down-regulated genes (mainly consisting of 856 protein-coding genes and 261 lncRNA genes) in RNA-Seq data, and 793 genes with down-regulated ribosome binding after estrogen treatment (mainly consisting of 626 protein-coding genes and 93 lncRNA genes), in Ribo-Seq data (Table 1; **Supplementary Tables 3 & 4**). As anticipated, the functional-category enrichment analysis of the estrogen-induced protein-coding genes revealed they were significantly enriched in biological processes directly relevant to cancer such as cell division, cell cycle phase transition, microtubule cytoskeleton organization, regulation of cell adhesion, and DNA replication, validating the sufficiently high quality of our cells, treatments, and data (**Supplementary Figs. 3 & 4**).

Table 1
A summary of DEGs found in RNA-Seq and Ribo-Seq.

Gene Category	# of Up-regulated Genes	# of Down-regulated Genes
(a) RNA-Seq		
Protein-coding	989	856
lncRNA	97	261
Pseudogene	46	114
Other	30	55
Total:	1162	1286
(b)Ribo-Seq		
Protein-coding	567	626
lncRNA	15	93
Pseudogene	66	38
Other	11	36
Total:	659	793
Note: Fisher's exact test performed on Protein-coding and lncRNA genes.		

Volcano plots can show genome-wide estrogen regulation of coding and non-coding genes in RNA-Seq and Ribo-Seq data. Our data indicated that protein-coding genes account for the majority of both the differential expression (RNA-Seq) and the differential ribosome occupancy (Ribo-Seq) observed (Figs. 1E & 1G), as the footprints of up- and down-regulated protein-coding genes in both datasets closely parallel the global data footprints (Figs. 1A & 1C). Interestingly, the number of protein-coding genes up-regulated is comparable to those down-regulated (Figs. 1E & 1G), whereas the less proportion of up-regulated lncRNAs in both RNA-Seq and Ribo-Seq suggests that majority of lncRNAs appear to be transcriptionally repressed by E2 (Figs. 1I & 1K), supported by the Fisher's exact test results in the contingency table ($p < 2.2*10^{-16}$ in Table 1a; $p < 1.4*10^{-12}$ in Table 1b).

To investigate the specifics of E2-mediated regulatory events that may affect only differential expression or only differential ribosome occupancy, but not both, we started from the lists of genes that were exclusively detected in only one of RNA-Seq and Ribo-Seq, respectively. Unsurprisingly, the number of genes exclusively detected in RNA-Seq is substantially greater than the number of genes exclusively detected in Ribo-Seq (Figs. 1B & 1D), given that Ribo-Seq, by design, surveys only the actively translated subset of the transcriptome. The numbers for lncRNAs were comparable to those of all-Ribo-Seq-detected vs. exclusively-Ribo-Seq-detected protein-coding gene mRNAs (Figs. 1F, 1H, 1J, & 1L). We further examined the relationship of gene expression profile (before and after E2 treatment) between the transcriptomic and ribosome profiling datasets, the majority of genes that were expressed in RNA-Seq were also detected in Ribo-Seq for both the total gene set (comprised mostly, though not solely, of protein-coding genes) and the protein-coding genes (Figs. 2A & 2B). Most of the differentially expressed protein-coding genes found in Ribo-Seq were also found to be differentially expressed in RNA-Seq, although approximately half of the genes differentially expressed in RNA-Seq were not differentially expressed in Ribo-Seq (Figs. 2D & 2E), implying that some genes may be transcriptionally, but not translationally, regulated by estrogen. Interestingly, approximately 67.6% lncRNA genes are un-detectable in ribosome profiling datasets (Fig. 2C), and most differentially expressed lncRNA genes are only

differentially expressed in RNA-Seq, not in Ribo-Seq (Fig. 2F), implying that many lncRNAs may never encounter a ribosome and that, for those that do, only transcriptional but not translational regulation takes place – consistent with these ncRNAs exerting RNA-mediated functions that do not depend on ribosome binding or translation. Hence, the vast majority lncRNAs appear untranslated in our data, consistent with the previous work by Banfai et al (Banfai, *et al.* 2011).

Quantitative correlation of RNA-Seq and Ribo-Seq

To investigate canonical (transcriptional) and noncanonical (translational) estrogen responses, we compared differential expression and differential ribosome occupancy at the whole-gene level (not for individual transcripts if more than one per gene) for genes that exhibited both types of responses, as measured by RNA-Seq and Ribo-SEq. As anticipated, we observed a strong positive correlation (R = 0.82) between transcriptomic data and ribosome profiling data (Fig. 3A); however, there were 106 genes that were differentially expressed to a far greater extent in RNA-Seq than in Ribo-Seq (transcriptome), and 129 genes that were differentially expressed to a far greater extent in Ribo-Seq than in Ribo-Seq ("translatome"). Those genes potentially represent specific estrogen-regulated differential expression events independent of ribosome occupancy, and specific estrogen-regulated translation differences independent of gene expression levels. This biological finding may help further illuminate the mechanisms of gene regulation in breast cancer and should be explored in future work.

To systemically examine the difference of protein-coding and lncRNA genes in the extent of their correlation between transcriptional and translational estrogen response, we analysed protein-coding and lncRNA genes separately for the above criteria. The strong correlation (R = 0.88) in the protein-coding genes demonstrated that most of these genes have a concordant transcriptional and translational response to estrogen (Fig. 3B). In contrast, we observed a substantially weaker correlation (R = 0.69) between lncRNA genes' transcriptional and translational changes in response to estrogen, indicating that many lncRNA genes do not exhibit consistent transcriptional and translational responses to estrogen, even though they are ribosome-bound (Fig. 3C).

To further investigate the expression level of protein-coding and lncRNA genes with ribosome, we took advantage of RPKM normalization because it takes gene length into account, in order to examine relative expression profiles within and across each expression-level bin. Compared with the protein-coding genes, it is evident that majority of lncRNA genes were not ribosome bound, and that the result was not confined to the lowest-expressed lncRNA genes, especially for those genes with RPKM > 10 (Fig. 4). Together with the weak correlation of expression on gene level for lncRNA genes, our results suggest that many lncRNA genes are not ribosome bound and that – even if they have Open Reading Frames (ORFs) – they are only rarely, if at all, translated into peptides in breast cancer cells, and that the levels of these rare peptides rarely respond to estrogen (Fig. 3).

Differential pseudogene expression analysis

To investigate pseudogene differential transcription and translation within the framework of the estrogen response, we searched all our RNA-Seq and Ribo-Seq datasets for the expression of the pseudogenes, with a high degree of stringency and relying on unambiguous sequencing reads that aligned to the pseudogenes at single loci better than to their putative parental genes and better than to other pseudogene copies. We identified 15 estrogen-induced pseudogenes in transcriptomics data, 31 estrogen-induced pseudogenes in ribosome-profiling data, 22 estrogen-repressed pseudogenes in RNA-Seq data, and 5 estrogen-repressed pseudogenes in ribosome-profiling data (**Supplementary Tables 5 & 6**). As a window into these expressed and ribosome-bound pseudogenes' potential functions, we manually examined the pseudogenes' parental genes for known relevance to cancer by searching the

literature, based on the premise the functions of expressed pseudogenes may be related to or may involve the direct regulation of their parental genes (Johnsson et al., 2013). Interestingly, of the parental genes that gave rise to these differentially expressed pseudogenes, 16 are related to cancer as supported by the evidence from published literature, and 10 of those 16 were shown in previous studies to be specifically functionally relevant to breast cancer (Tables 2 & 3). There were low numbers of overlapping differentially expressed pseudogenes (four pseudogenes) with both differential expression and differential ribosome occupancy, a low number perhaps explainable by the contention that not many expressed-pseudogene transcripts are expected to bind to ribosomes since their ORFs are missing or severely disrupted and hence they may not have ribosome entry sites or start codons, may only bind ribosomes once during the pioneer round of translation, and are generally not be translated into proteins (**Supplementary Table 7**).

Table 2

Top 5 differentially expressed pseudogenes in RNA-SEq. The expression level of the parental gene follows its name, and the names underlined refer to the results of BLASTN human genome searches. Lack of known disease functions of the parental gene is indicated by a "-" symbol.

Gene Symbol	Type of Pseudogene	Log_2FC	P_{adj}Value	Parent Gene	Log_2FC	P_{adj}Value	Risk	References
A. Over-expressed								
RP11-564A8.4	processed	4.74	1.5×10^{-204}	RPL13A	-0.44*	1.6×10^{-44}	breast cancer-related gene	Hellwig et al., 2016
RP11-64B16.2	processed	2.23	4.2×10^{-52}	PDAP1	0.44*	2.9×10^{-35}	cancer-related gene	Weston et al., 2018
RP11-673C5.1	processed	1.18	1.0×10^{-50}	HMGB1	0.98**	5.2×10^{-298}	breast cancer-related gene	Sun et al., 2015
AC005336.4	processed	3.86	1.0×10^{-47}	CYP4F11	3.48***	6.4×10^{-270}	breast cancer-related gene	Cardenas et al., 2012
RP11-424C20.2	processed	1.27	1.4×10^{-21}	UHRF1	1.25***	4.3×10^{-125}	breast cancer-related gene	Gao et al., 2017
B. Under-expressed								
RP1-90G24.5	processed	-2.05	1.3×10^{-26}	CPSF1	0.04	5.3×10^{-1}	-	-
NBPF13P	unprocessed	-3.41	9.3×10^{-19}	NBPF10	-0.13	4.0×10^{-1}	cancer-related gene	Hamdi et al., 2018
RP11-72P19.2	processed	-2.38	8.1×10^{-11}	TATDN2	0.43*	1.1×10^{-08}	-	-
RP11-554D14.4	unprocessed	-1.76	2.0×10^{-10}	-	-	-	-	-
MST1P2	unprocessed	-1.22	4.6×10^{-7}	MST1	-0.71**	5.4×10^{-44}	breast cancer-related gene	Wei et al., 2018

*: |FC| > 1.2; **: |FC| > 1.5; ***: |FC| > 2

Table 3

Top 5 differentially expressed pseudogenes in Ribo-SEq. The expression level of the parental gene follows its name, and the names underlined refer to the results of BLASTN human genome searches. Lack of known disease functions of the parental gene is indicated by a "-" symbol.

Gene Symbol	Type of Pseudogene	Log_2FC	P_{adj}Value	Parental Gene	Log_2FC	P_{adj}Value	Risk	References
A. Over-expressed								
RP11-564A8.4	processed	4.29	4.0×10^{-24}	RPL13A	0.43	4.5×10^{-1}	breast cancer-related gene	Hellwig et al., 2016
RP11-83J16.3	processed	1.96	2.4×10^{-5}	TUBA1B	1.19***	4.9×10^{-4}	-	-
AC002069.5	processed	1.65	4.0×10^{-5}	KPNA2	1.36***	2.3×10^{-6}	breast cancer-related gene	Alshareeda et al., 2015
CTD-2184D3.1	processed	1.74	1.4×10^{-4}	RPS13	0.89**	2.2×10^{-2}	cancer-related gene	Shi et al., 2004
RP11-417L14.1	processed	1.21	4.7×10^{-4}	KPNA2	1.36***	2.3×10^{-6}	breast cancer-related gene	Alshareeda et al., 2015
B. Under-expressed								
RP5-1025A1.2	processed	-4.65	1.0×10^{-3}	VTCN1	-5.49***	1.1×10^{-26}	breast cancer-related gene	Tsai et al., 2015
NBPF13P	unprocessed	-3.64	4.5×10^{-3}	NBPF10	-	-	cancer-related gene	Hamdi et al., 2018
MST1P2	unprocessed	-2.45	2.2×10^{-2}	MST1	-1.38	2.9×10^{-1}	cancer-related gene	Wei et al., 2018
RP11-109N23.6	processed	-1.85	3.7×10^{-2}	DCAF4	-0.72**	7.4×10^{-3}	cancer-related gene	Mangino et al., 2015
RP11-458F8.2	unprocessed	-2.21	3.7×10^{-2}	GTF2I	-0.18	7.2×10^{-1}	breast cancer-related gene	Zhou et al., 2019

*: |FC| > 1.2; **: |FC| > 1.5; ***: |FC| > 2

Pseudogene function and breast cancer

Pseudogene sense and antisense transcription can epigenetically regulate the parental genes of the pseudogenes. For instance, PTENP1, the pseudogene of PTEN tumor suppressor gene, expresses an antisense transcript that regulates

PTEN (Poliseno et al., 2010; Johnsson et al., 2013). Correlation of expression levels between the pseudogene andits parental gene may indicate the potential mechanism in gene regulation (Gupta, et al., 2015). Our data reveals significant pairs of potentially co-regulated genes, such as AC005336.4:CYP4F11 and RP11-424C20.2:UHRF1 in transcriptomics data as well as RP11-83J16.3:TUBA1B and RP11-417L14.1:KPNA2 in ribosome-profiling data, that hint at the possibility of specific pseudogene:gene regulatory interaction events after estrogen treatment. Another function of pseudogenes in the regulation of parental-gene mRNA stability is by acting as miRNA molecular sponges, and can render a situation where the pseudogene transcript is “sponging” microRNAs away from the parental gene’s mRNA, stabilizing or upregulating that mRNA (Gupta, et al., 2015). Interestingly, the pairs RP11-564A8.4:RPL13A and RP11-72P19.2:TATDN2 in the transcriptomics data, where the pseudogene RNA and the parental-gene mRNA are discordantly regulated in each pair, are consistent with such a regulatory outcome during estrogen response.

Another interesting example is the KHSRP (KH-type splicing regulatory protein) pseudogene (RP11-473O3.1). KHSRP is oncogenic, and plays an important role in promote tumour growth and metastasis (Yan et al., 201). It also takes part in the maturation of miRNAs in breast cancer (Santarosa et al., 2010; Trabucchi et al., 2009). The expression of this pseudogene was significantly downregulated, whereas its parental gene was significantly upregulated, at the transcript level after estrogen treatment (Fig. 6a). Interestingly, our data showed that the antisense strand of the KHSRP pseudogene was transcribed, therefore it might regulate the expression of its parental gene by binding to the sense strand of the parent gene (Fig. 6a) or by a PTENP1-like mechanism. In other words, the downregulation of the pseudogene might reduce the number of its RNA molecules that can bind and inhibit its parental gene. This might upregulate the expression of the parental gene and make more ribosome-bound mRNA molecules available for protein translation, promoting tumour growth and metastasis.

NONO, a non-POU-domain-containing octamer binding protein, is important for tumorigenesis and progression, including in breast cancer (Zhu et al., 2016; Cheng et al., 2018; Yamamoto et al., 2018). Downregulation of NONO can induce apoptosis and suppress growth and invasion in esophageal squamous cell carcinoma (Cheng et al., 2018). Our data showed that the NONO pseudogene was significantly downregulated, in contrast the expression of the transcripts of its parental gene particularly the ribosome-bound transcripts were significantly up-regulated after estrogen treatment (Fig. 6b). Interestingly, manual examination of the transcript sequence of the pseudogene in the IGV genome browser revealed that both of its sense and antisense strand sequences were transcribed (generally more reads from the antisense strand compared to the sense strand i.e. 16 antisense vs 3 sense reads). We cannot rule out the possibility that the antisense transcripts may regulate the pseudogene or its parental gene, similarly to the PTEN scenario. Therefore, the E2-driven downregulation of the NONO pseudogene might increase the expression level of its parent gene and ribosome-bound mRNAs, promoting E2-dependent breast cancer tumorigenesis and progression.

Discussion

The difference between protein-coding genes (Fig. 1E & 1G) and lncRNA genes (Fig. 1I & 1K), as well as the absence of many expressed lncRNAs from ribosome profiling data (Fig. 2C), led us to conclude that, unlike mRNAs, the vast majority lncRNAs are not ribosome-bound in MCF7 cells, regardless of estrogen treatment. Interestingly, differentially regulated lncRNAs are overwhelming repressed rather than induced, whereas for protein-coding genes about half are induced and half repressed. We identify a small collection of differentially ribosome-bound lncRNAs, (Fig. 1I & 1K), which correlate with differential transcriptional regulation more weakly than their protein-coding counterparts (Fig. 3). These lncRNAs may encode short peptides, and whether they are ectopically translated, or part of a regulatory program, will be an intriguing direction for future study.

These findings challenge numerous reports that most lncRNAs are ribosome associated in human cells (Ingolia et al., 2014), but are consistent with our previous findings using shotgun proteomics (Banfai et al., 2011). The considerable numbers of transcripts not bound by ribosomes suggest at least two hypotheses that future work should distinguish experimentally: (i) the lncRNAs are cytoplasmic, but not bound by ribosomes (outside of pioneer-round-of-proofreading events), or (ii) the lncRNAs are nuclear (or sequestered in other cellular compartments); this could also be true for some of the differentially expressed mRNAs that are Ribo-Seq-negative, as nuclear retention of mRNAs is a stress response mechanism in human cells (Prasanth et al., 2005).

Our data showed only a fairly weak correlation between lncRNA genes' transcriptional and translational changes in response to estrogen, in contrast to protein-codinggenes. These findings directly contradict the still-prevailing view in the field that many lncRNAs are translated (Ji et al., 2015). That emerging de-facto dogma was never previously tested in a controlled experiment in the same human cells before and after a specific nuclear hormone treatment, until our experiments, which suggest that most lncRNAs, including those with robust positive and negative transcriptional responses to estrogen, do not associate with ribosomes and hence are translationally inactive.

Differential transcript analysis revealed that alternative promoter usage and alternative splicing of certain protein-coding genes are responsive to estrogen treatment, and that estrogen may separately regulate the transcription and translation of other genes. The mechanism of this apparent preferential translational regulation by estrogen remains unknown, and these findings need to be validated by independent experiments such as isoform-specific quantitative RT-PCR and targeted mass spectrometry.

Our data showed that most of the differentially expressed protein-coding genes found in Ribo-Seq were also found to be differentially expressed in RNA-Seq, but most differentially expressed lncRNA genes are only differentially expressed in RNA-Seq, not in Ribo-SEq. There are at least three possible explanations for these observations. First, as expected, most protein-coding genes expressed in RNA-Seq were also found in Ribo-Seq– a reasonable finding, as protein-coding transcripts bound on ribosomes would be translated into proteins. Whereas, the majority of lncRNA genes are absent from the ribosome profiling datasets, indicating a likelihood that they are not ribosome-bound – consistent with proteomics evidence (Banfai et al., 2012) as well as, importantly, with other Ribo-Seq efforts (Guttman et al., 2013). Second, because our Ribo-Seq libraries were not as deep as, and (due to the difference in protocols) had shorter reads than, our RNA-Seq libraries, we cannot rule out the possibility that it was generally more difficult to detect both the presence and the differential expression of transcripts expressed at low levels (which are more likely to be lncRNAs than mRNAs [Derrien et al., 2012]) in Ribo-Seq than in RNA-sEq. Hence, some of the "RNA-Seq-only" and "differentially expressed in RNA-Seq but not in Ribo-Seq" hits might be false negatives, although this can be corrected by separating all gene sets into bins – ordered and ranked by gene expression level – and comparing mRNAs and lncRNAs only within same-expression-range bins (i.e. highly expressed mRNAs with equally highly expressed lncRNAs, not with all lncRNAs) (Fig. 4). Third, unlike mRNAs, many lncRNAs are nuclear or otherwise restricted to specific subcellular compartments inaccessible to ribosomes (Lennox & Behlke, 2016; Cabili et al., 2015); this may explain that underrepresentation of lncRNAs in Ribo-Seq-detectable and Ribo-Seq-differentially-expressed datasets. Despite these biological limitations, the results suggest that, if a gene is differentially ribosome-bound upon estrogen treatment, then that gene is likely to also be differentially expressed.

We also present a global study of the expressed-pseudogene transcriptome response to a nuclear hormone. We identified several pairs of differentially expressed pseudogenes and their parental genes, some with positive and others with negative correlations between the pseudogene RNA and the parental gene mRNA levels. It is possible that these transcribed pseudogenes regulate their parental genes, instead of being merely differentially co-expressed with them. The estrogen responsive pseudogene transcripts in the positively-correlated pairs may function as ceRNAs

(competing endogenous RNAs), sponging miRNAs that would otherwise bind to, and repress, the co-expressed parental genes' mRNAs.

Most of the protein-coding genes' functional categories significantly up- and down-regulated by estrogen in RNA-Seq are, expectedly, broadly associated with cancer (**Supplementary Figs. 3 & 4**). The estrogen up-regulated genes engaged in cell division and DNA replication confirm cell proliferation after estrogen treatment. Other up-regulated genes involved in microtubule cytoskeleton organization and regulation of cell adhesion facilitate invasion and metastasis, key hallmarks of cancer (Hanahan & Weinberg, 2011). Additionally, the down-regulated genes were engaged in ECM organization, which is usually deregulated during proliferation and metastasis (Walker, et al., 2018). Combined with the regulation of actin filament-based process in cell migration (Yamaguchi & Condeelis, 2007), the gene-ontology findings were fully consistent with estrogen's known role in cancer progression. In KEGG analysis, besides the up-regulated and down-regulated genes involved in similar processes (cell cycle and DNA replication; ECM-receptor interaction), the PPAR signalling pathway was detected and is known to contribute to cell growth in cancer (Tachibana et al., 2008). Activation of cAMP signalling inhibits proliferation of cancer cells (Fajardo et al., 2014), implicating that the down-regulated cAMP signalling may stimulate cancer progression.

Conclusion

Our findings revealed that, unlike protein-coding gene mRNAs, most lncRNAs are not ribosome-bound and therefore are not translated into proteins in human breast cancer cells, whereas for many other lncRNAs, E2 induction affects their expression level, but not their association with ribosomes. Only a small subset of lncRNAs occupied by ribosomes and displaying transcriptional and/or ribosome-occupancy changes was found, while the majority of lncRNAs was not detectably bound by ribosomes, but did respond to estrogen transcriptionally. The results are consistent with mass spectrometric evidence against widespread lncRNA translation, and indicate systematic biological distinctions between the estrogen response of the protein-coding and the non-coding RNA transcriptomes.

Abbreviations

E2
estrogen (17-β-estradiol)
ER
estrogen receptor
DEG
differentially expressed gene
DEP
differentially expressed pseudogene
DET
differentially expressed transcript
lncRNA
long non-coding RNA

Declarations

Ethics approval and consent to participate

Not applicable. Only publicly-available cell lines were used. No patient data or tissue was used.

Consent for publication

All co-authors have reviewed the manuscript and agree to its contents.

Availability of data and material

All datasets will be made available as Supplementary Files on the journal's website immediately upon publication.

Competing interests

None declared.

Funding

This work was supported by the National Institutes of Health (NIH) Director's New Innovator Award (grant number: 1DP2-CA196375) to LL and the high-level talent recruitment programme for academic and research platform construction (Reference Number: 5000105) from Wenzhou-Kean University.

Authors' contributions

SWC and LL designed the project. JC, ASG, and DJ performed cell culture and treatments, RNA-seq and Ribo-seq library construction and sequencing, and all other wet-bench laboratory work. YZ, ES, JBB, MAK, SWC, and LL analyzed the data. ASG and ES co-supervised YZ's internship in LL's laboratory. YZ and ES wrote the first draft of the paper. YZ, SWC, JBB, MAK, and LL wrote and revised the paper.

Acknowledgement

We are grateful to Ms. Pattaraporn Thepsuwan for technical assistance.

References

1. Ali S, Coombes RC. Estrogen receptor alpha in human breast cancer: occurrence and significance. J Mammary Gland Biol Neoplasia. 2000;5(3):271–81.
2. Altschul SF, et al. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
3. Andrews S. *A quality control tool for high throughput sequence data*. 2018; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
4. Banfai B, et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. 2012;22(9):1646–57.
5. Bray F, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
6. Cabili MN, et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. 2015;16:20.
7. Calviello L, Ohler U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. Trends Genet. 2017;33(10):728–44.
8. Chen S, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
9. Cheng R, et al. Downregulation of NONO induces apoptosis, suppressing growth and invasion in esophageal squamous cell carcinoma. Oncol Rep. 2018;39(6):2575–83.

10. Conesa A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
11. Derrien T, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–89.
12. Dinger ME, et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol. 2008;4(11):e1000176.
13. Fajardo AM, Piazza GA, Tinsley HN. The role of cyclic nucleotide signaling pathways in cancer: targets for prevention and treatment. Cancers (Basel). 2014;6(1):436–58.
14. Frasor J, et al. Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. Endocrinology. 2003;144(10):4562–74.
15. Gazinska P, et al. Comparison of basal-like triple-negative breast cancer defined by morphology, immunohistochemistry and transcriptional profiles. Mod Pathol. 2013;26(7):955–66.
16. Gupta A, et al. Differentially-Expressed Pseudogenes in HIV-1 Infection. Viruses. 2015;7(10):5191–205.
17. Guttman M, et al. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154(1):240–51.
18. Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr Genomics. 2013;14(2):91–110.
19. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
20. Ingolia NT, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8(5):1365–79.
21. Ingolia NT, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–23.
22. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789–802.
23. Ishigaki Y, et al. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. Cell. 2001;106(5):607–17.
24. Ji Z, et al. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015;4:e08890.
25. Johnsson P, et al. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat Struct Mol Biol. 2013;20(4):440–6.
26. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
27. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.
28. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
29. Lawrence M, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.
30. Lennox KA, Behlke MA. Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. Nucleic Acids Res. 2016;44(2):863–77.
31. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
32. Li K, et al. Characterization of beta2-microglobulin expression in different types of breast cancer. BMC Cancer. 2014;14:750.

33. Lin CY, et al., *Primate-specific oestrogen-responsive long non-coding RNAs regulate proliferation and viability of human breast cancer cells*. Open Biol, 2016. 6(12).
34. Lin KH, et al. RNA-seq transcriptome analysis of breast cancer cell lines under shikonin treatment. Sci Rep. 2018;8(1):2672.
35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
36. Milligan MJ, et al. Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. Front Genet. 2016;7:26.
37. Milligan MJ, Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet. 2014;5:476.
38. Peck B, et al. Inhibition of fatty acid desaturation is detrimental to cancer cell survival in metabolically compromised environments. Cancer Metab. 2016;4:6.
39. Pei B, et al. The GENCODE pseudogene resource. Genome Biol. 2012;13(9):R51.
40. Poliseno L, Marranci A, Pandolfi PP. Pseudogenes in Human Cancer. Front Med (Lausanne). 2015;2:68.
41. Poliseno L, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465(7301):1033–8.
42. Prasanth KV, et al. Regulating gene expression through RNA nuclear retention. Cell. 2005;123(2):249–63.
43. Rakha EA, et al. Triple-negative breast cancer: distinguishing between basal and nonbasal subtypes. Clin Cancer Res. 2009;15(7):2302–10.
44. Ruan Y, et al. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res. 2007;17(6):828–38.
45. Russnes HG, et al. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. Am J Pathol. 2017;187(10):2152–62.
46. Santarosa M, et al. BRCA1 modulates the expression of hnRNPA2B1 and KHSRP. Cell Cycle. 2010;9(23):4666–73.
47. Tachibana K, et al. The Role of PPARs in Cancer. PPAR Res. 2008;2008:102737.
48. Trabucchi M, et al. The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. Nature. 2009;459(7249):1010–4.
49. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
50. Walker C, Mojares E. and A. Del Rio Hernandez, *Role of Extracellular Matrix in Development and Cancer Progression*. Int J Mol Sci, 2018. 19(10).
51. Yamaguchi H, Condeelis J. Regulation of the actin cytoskeleton in cancer cell migration and invasion. Biochim Biophys Acta. 2007;1773(5):642–52.
52. Yamamoto R, et al. Overexpression of p54(nrb)/NONO induces differential EPHA6 splicing and contributes to castration-resistant prostate cancer growth. Oncotarget. 2018;9(12):10510–24.
53. Yan M, et al. RNA-binding protein KHSRP promotes tumor growth and metastasis in non-small cell lung cancer. J Exp Clin Cancer Res. 2019;38(1):478.
54. Zheng D, et al. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res. 2007;17(6):839–51.
55. Zhou Y, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.

56. Zhu Z, et al., *p54(nrb)/NONO regulates lipid metabolism and breast cancer growth through SREBP-1A.* Oncogene, 2016. **35**(11): p. 1399 – 410.

Figures

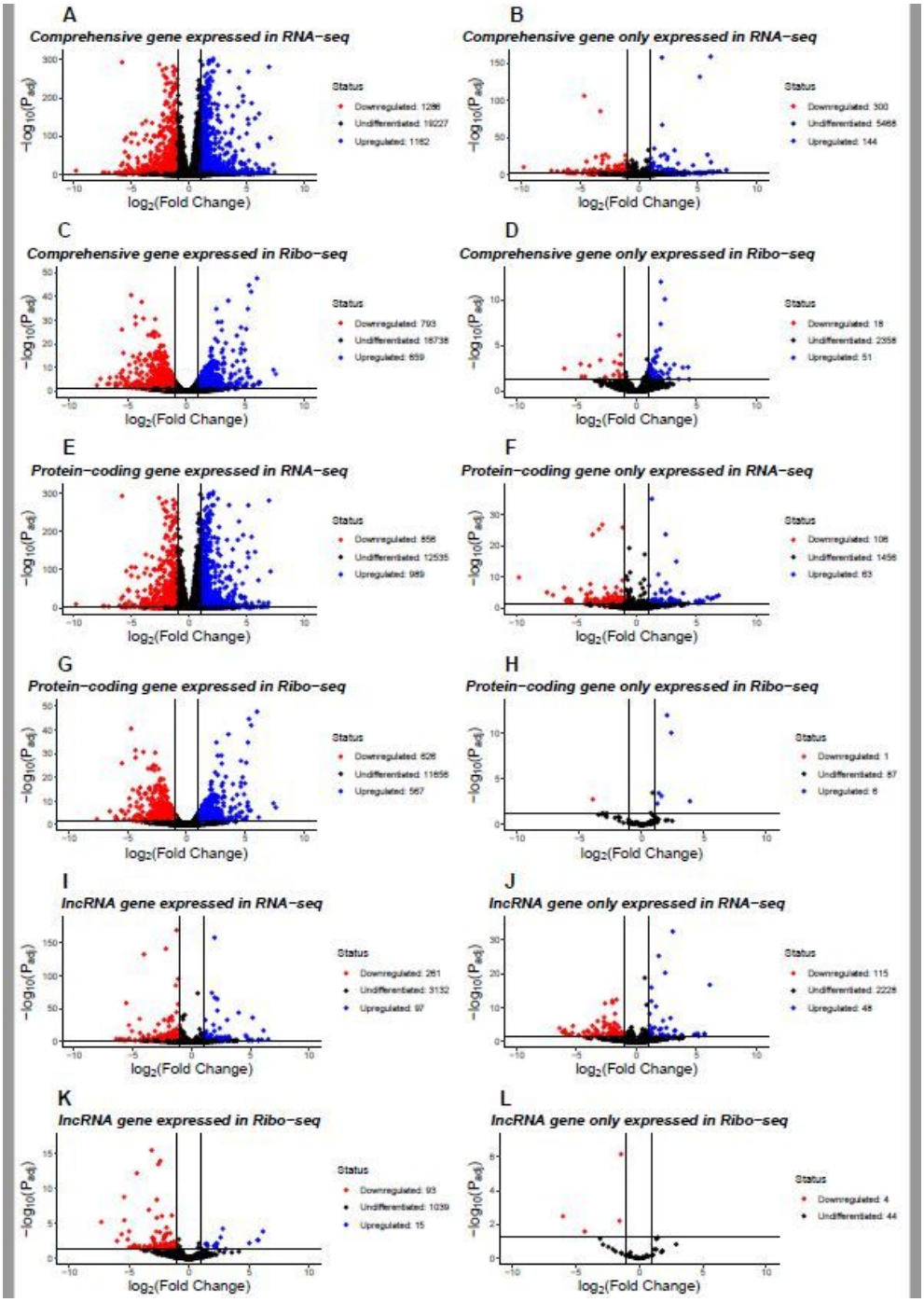

Figure 1

Volcano plots in RNA-Seq and Ribo-Seq. (A) ~ (L) evaluate the differential expression and differential ribosome occupancy changes for genes as its title indicates. The x-axis corresponds to the log2 (fold change) in the estrogen treatment versus control group. The y-axis represents the negative log10 of Padj value. A cutoff with Padj value of 0.05 and a fold change of 2 are addressed by dashed lines within the plots. The number of upregulated,

downregulated, and “undifferentiated” (detected and expressed, but not differentially expressed) genes is indicated next to each plot.

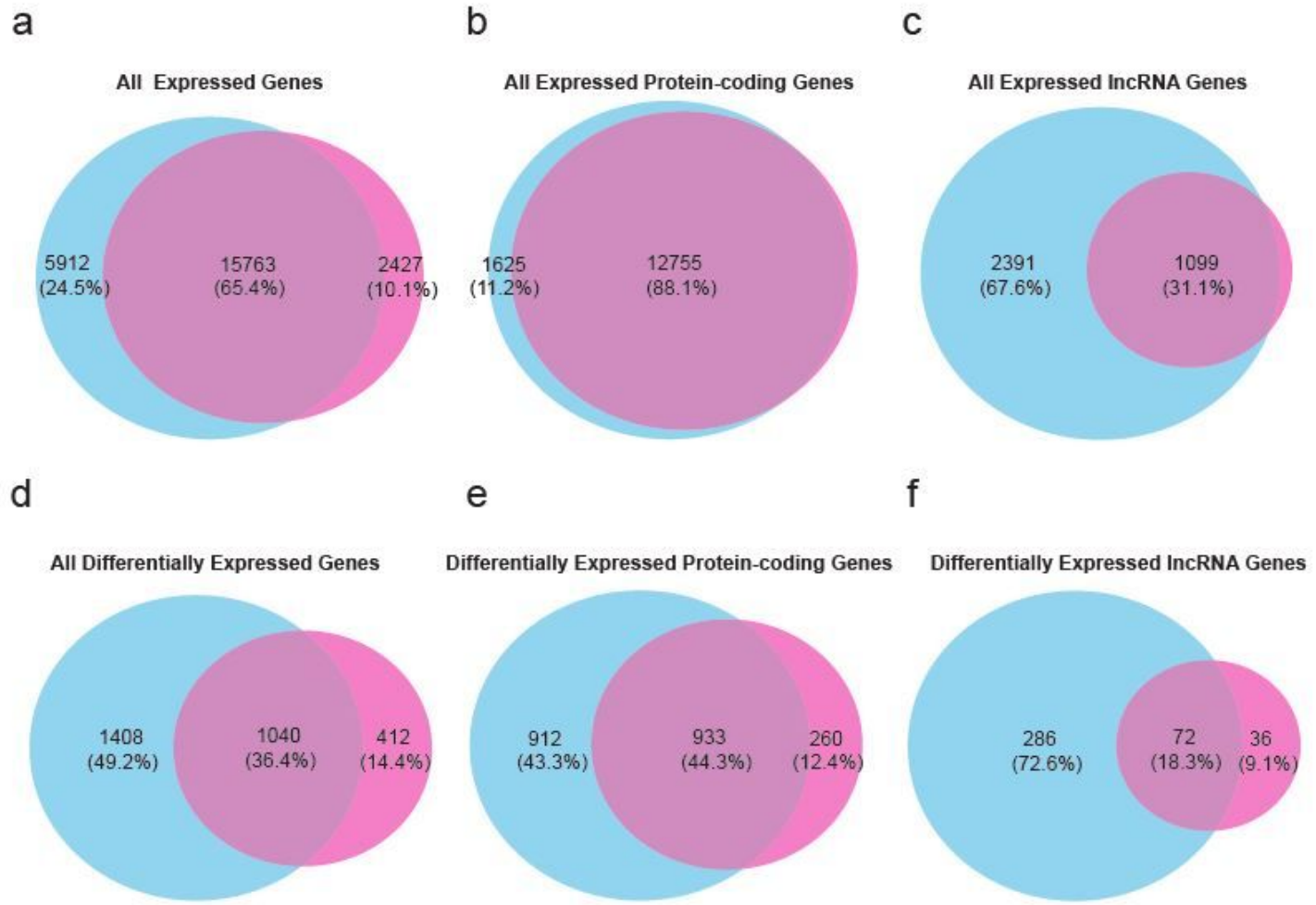

Figure 2

Venn diagrams of comparisons between RNA-Seq and Ribo-Seq. (A) ~ (F) illustrate the numbers of genes common to both, or exclusively found in, RNA-Seq (blue, left) or in Ribo-Seq (purple, right).

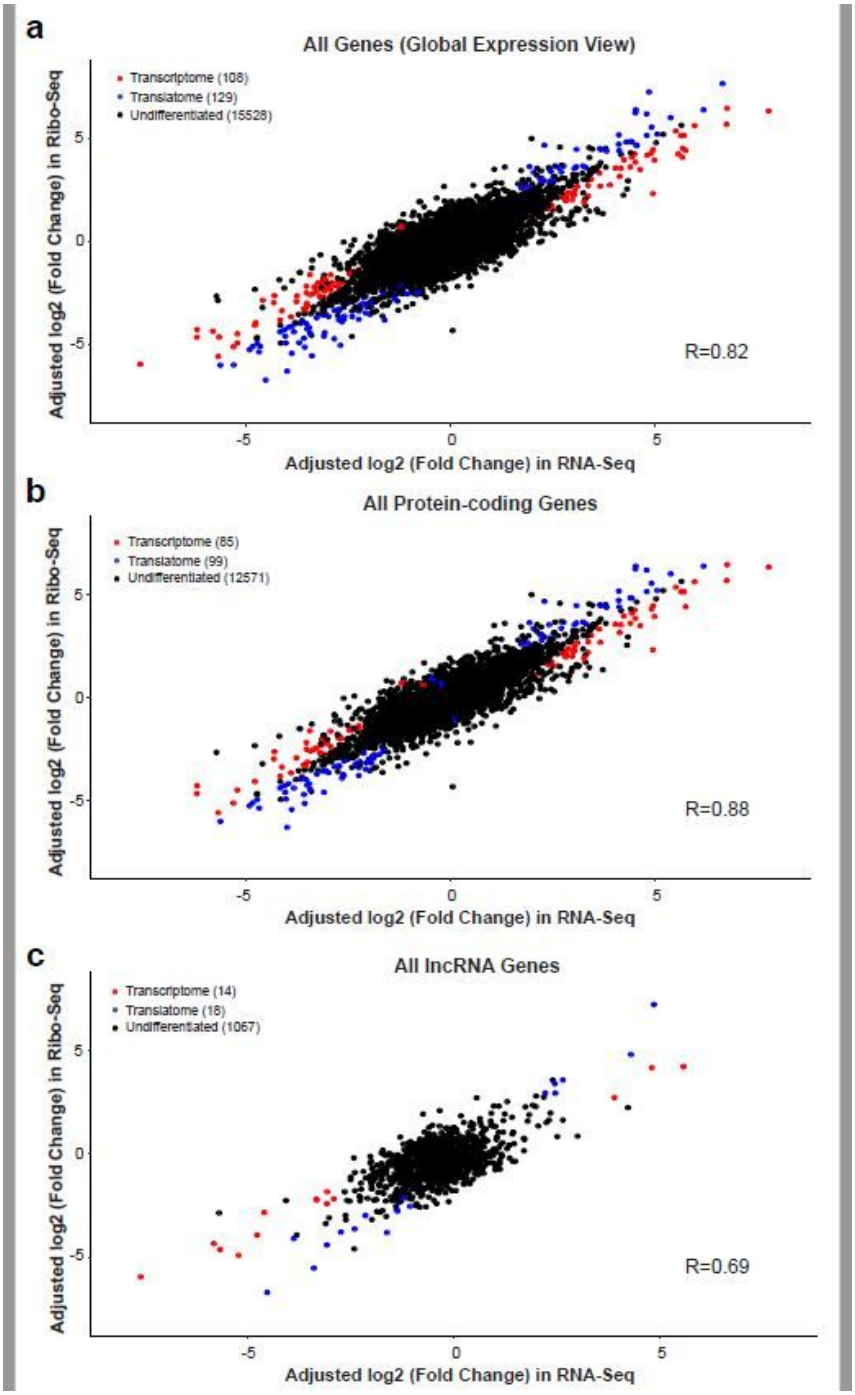

Figure 3

Correlation of gene expression between transcriptomic and ribosome profiling data. The y-axis corresponds to thelog2 of adjusted fold change value in Ribo-Seq and the x-axis corresponds to the log2 of adjusted fold change value in RNA-Seq.The gene whose discrepancy in adjusted fold change between in both datasets > 2 or < -2 or between -2 and 2, are marked by “Transcriptome” - the gene is disproportionately more differentially expressed in RNA-Seq, “Translatome” - the gene is disproportionately more differentially expressed in Ribo-Seq, “Undifferentiated” – no evident discrepancy in expression level between the two datasets. Disproportionately greater differential expression refers to the magnitude of absolute-value fold change and can be up or down; we only considered genes that are up in both RNA-Seq and Ribo-Seq after estrogen treatment, and genes that are down in both RNA-Seq and Ribo-Seq after estrogen treatment, in this analysis.

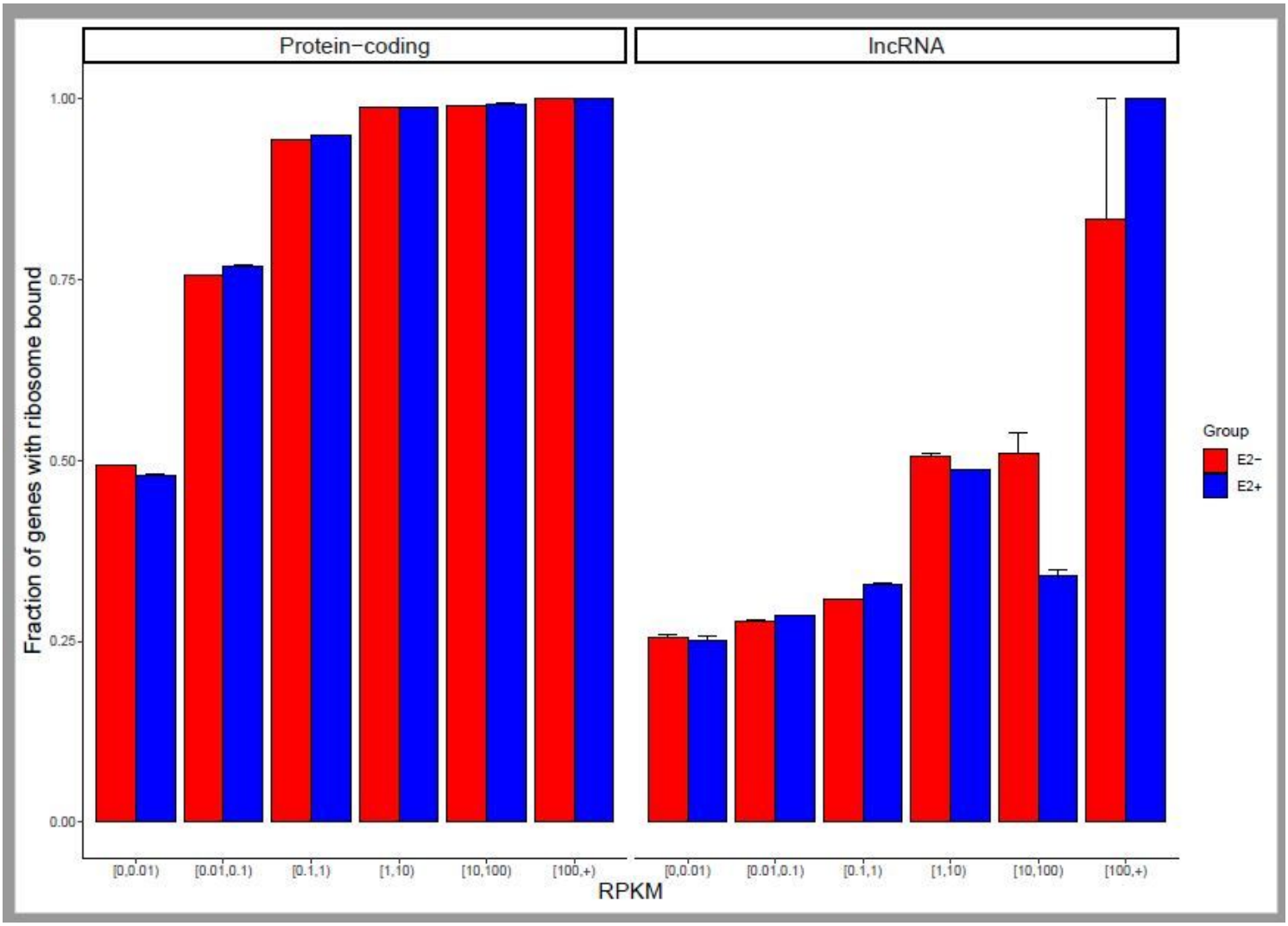

Figure 4

Expression-level bins of protein-coding and lncRNA genes with ribosome binding. The y-axis corresponds to the fraction of genes (in RNA-Seq) with ribosome binding, in the total number of genes detectable in RNA-Seq within the same expression-level-range (RPKM) bin. The x-axis corresponds to the scale normalized by RPKM method. “E2-”: Control group; “E2+”: Estrogen treatment.

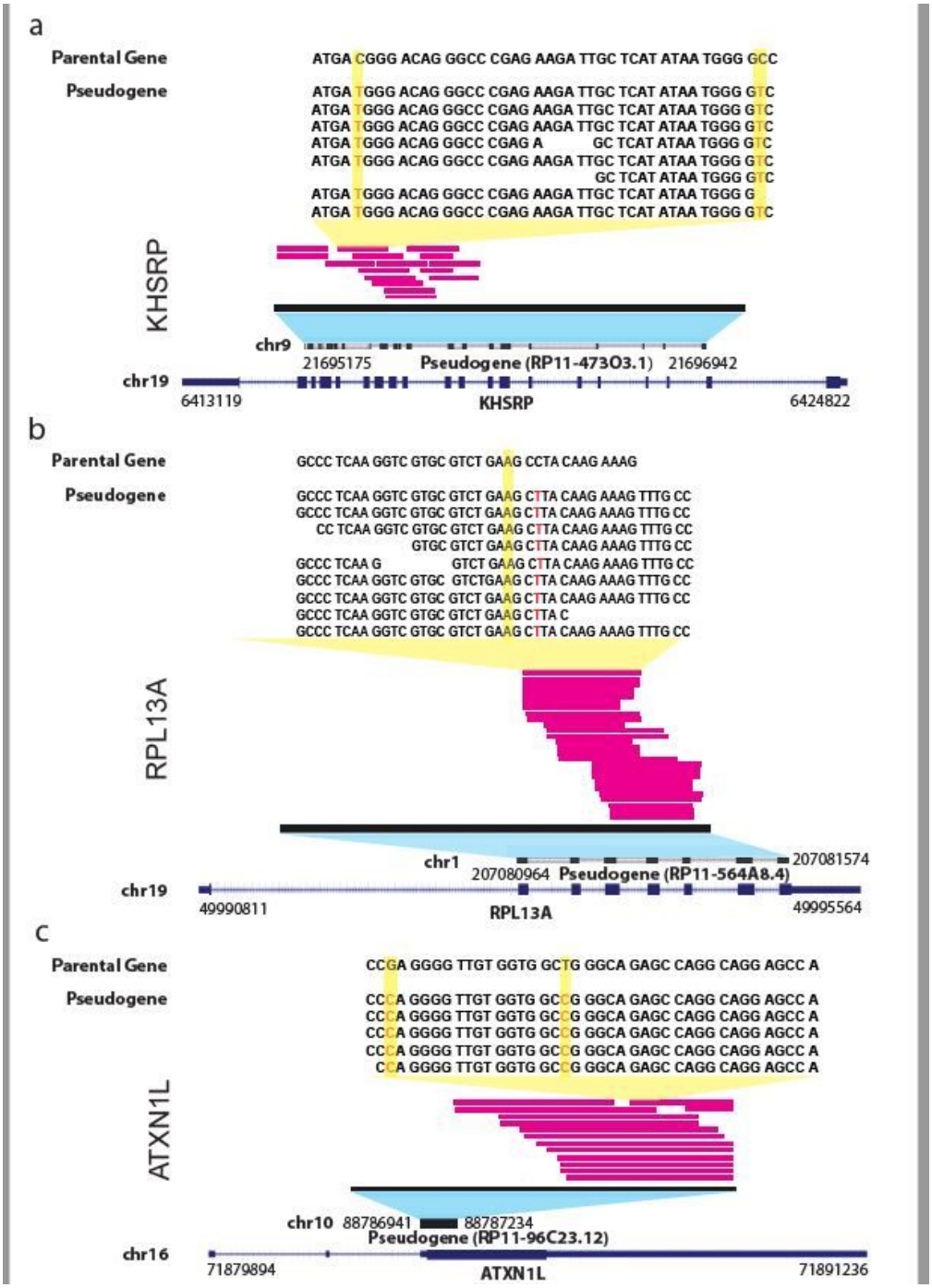

Figure 5

Three examples of pseudogene RNA-Seq transcript-to-genome alignment validations. These differentially expressed pseudogenes were indeed expressed supported by the mapped pseudogene-specific reads, instead of the mapping artefacts. Pseudogene reads were indicated with pink lines and only some reads were shown in this figure. (A) RP11-473O3.1 and KHSRP. (B) RP11-564A8.4 and RPL13A. (C) RP11-96C23.12 and ATXN1L. Sequence differences that definitively distinguish between the pseudogene and the parental gene, mirrored in strand-specific RNA-Seq reads aligning perfectly to the pseudogene but not to the parental gene, are highlighted in red. The pseudogene transcripts are represented as black bars. Mapping details for pseudogenes and parental genes are shown along with transcript representation. These schematic representations are overlaid with multiple reads used to aggregate clusters (in pink), followed by overlapping sequences of pseudogenes in a zoomed-in view.

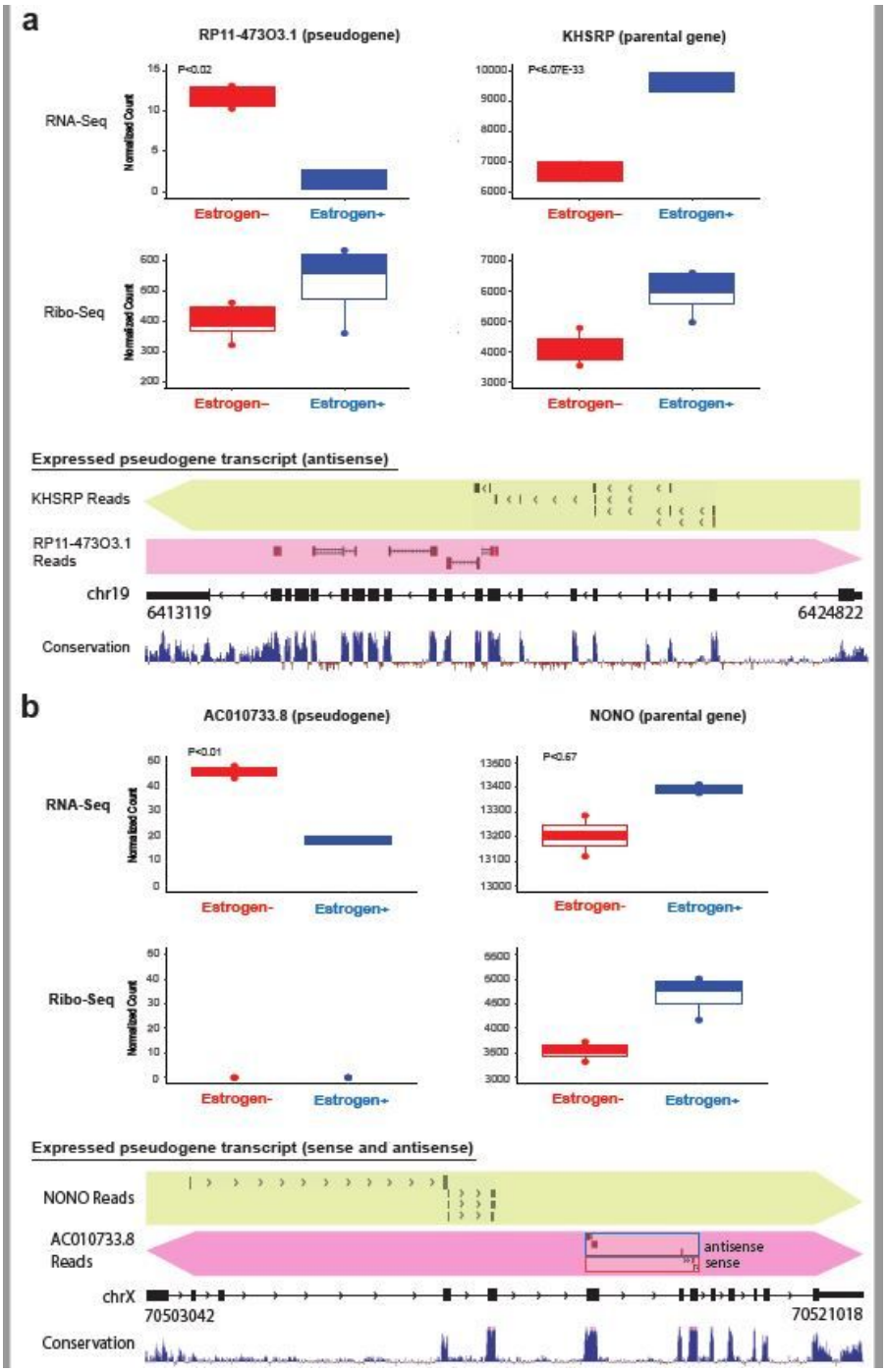

Figure 6

Distinct and opposite differential expression patterns of parental genes and their pseudogenes in RNA-Seq and Ribo-Seq. (a) Left panel shows KHSRP pseudogene (RP11-473O3.1), whereas the right panel shows its parental gene. Bottom: The orientations of the transcripts expressed by pseudogenes (purple arrowhead) and their parent genes (green arrowhead). The selected pseudogene (imperfect mapping) and parental gene (perfect-match mapping) reads were mapped, strand-specifically, only to the parental-gene locus in this diagram. The antisense strand of the KHSRP pseudogene was transcribed and hence yielded RNA theoretically capable of binding the sense-stand mRNA of its parental gene. (b) Top: Left panel shows the NONO pseudogene (RP11-473O3.1), whereas the right panel shows its parental gene; The orientations of the transcripts expressed by pseudogenes (purple arrowhead) and their parent genes (green arrowhead). Only selected pseudogene and parental gene reads were mapped, strand-specifically, only to the parental-gene locus in this diagram. Both sense and antisense strands of the NONO pseudogene were transcribed.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable4.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable3.xlsx
- SupplementaryData.pdf
- SupplementaryData.pdf