



Research paper

Pan-cancer analysis on microRNA-associated gene activation

Hua Tan^{a,*}, Shan Huang^{b,1}, Zhigang Zhang^a, Xiaohua Qian^c, Peiqing Sun^{b,*}, Xiaobo Zhou^{a,*}^a School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA^b Department of Cancer Biology, Wake Forest Comprehensive Cancer Center, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA^c Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

ARTICLE INFO

Article history:

Received 12 February 2019

Received in revised form 21 March 2019

Accepted 27 March 2019

Available online 5 April 2019

Keywords:

Pan-cancer miRNA

miRNA activation

Intragenic miRNA

Super-enhancer

ABSTRACT

Background: While microRNAs (miRNAs) were widely considered to repress target genes at mRNA and/or protein levels, emerging evidence from *in vitro* experiments has shown that miRNAs can also activate gene expression in particular contexts. However, this counterintuitive observation has rarely been reported or interpreted in *in vivo* conditions.

Methods: We systematically explored the positive correlation between miRNA and gene expressions and its potential implications in tumorigenesis, based on 8375 patient samples across 31 major human cancers from The Cancer Genome Atlas (TCGA).

Findings: We found that positive miRNA-gene correlations are surprisingly prevalent and consistent across cancer types, and show distinct patterns than negative correlations. The top-ranked positive correlations are significantly involved in the immune cell differentiation and cell membrane signaling related processes, and display strong power in stratifying patients in terms of survival rate. Although intragenic miRNAs generally tend to co-express with their host genes, a substantial portion of miRNAs shows no obvious correlation with their host gene plausibly due to non-conservation. A miRNA can upregulate a gene by inhibiting its upstream suppressor, or shares transcription factors with that gene, both leading to positive correlation. The miRNA/gene sites associated with the top-ranked positive correlations are more likely to form super-enhancers compared to randomly chosen pairs. Wet-lab experiments revealed that positive correlations partially remain in *in vitro* condition.

Interpretation: Our study brings new insights into the critical role of miRNA in gene regulation and the complex mechanisms underlying miRNA functions, and reveals both biological and clinical significance of miRNA-associated gene activation.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Cancer is caused by uncontrolled cell growth reflective of multiple established hallmarks [1]. Underlying the aberrant cell proliferation is activation of critical oncogenes and inactivation of tumor suppressor genes resulted from multiple genetic and epigenetic alterations in a cancer tissue-specific manner [2–6]. Among these, microRNAs (miRNAs) are a class of small (~22 nucleotides), non-protein-coding RNAs known as important post-transcriptional regulators of gene expression [7,8]. miRNAs exert regulatory functions by base-pairing with

complementary sequences typically in the 3'-untranslated region (3' UTR) of mRNAs to target them for degradation or prevent their translation [7,9]. It is estimated that >60% of human protein-coding genes are under selective pressure to maintain pairing to miRNAs and over one third of human genes appear to be conserved miRNA targets [10,11]. This indicates that miRNAs can influence almost every critical signaling pathway in a cell [12].

A canonical miRNA is transcribed from miRNA gene by RNA polymerase II (pol II) as a primary miRNA (pri-miRNA). The pri-miRNA transcript is first cleaved in the nucleus by the microprocessor, which contains a nuclear RNase III called Drosha and its cofactor DGCR8, into a hairpin structured precursor miRNA (pre-miRNA). Then, the pre-miRNA is exported to the cytoplasm through the activation of Exportin 5 and RAN-GTP, and further processed by an endonuclease Dicer to generate the miRNA duplex, which contains the miRNA paired to its passenger strand usually called miRNA* (miRNA star). One strand of the miRNA duplex, the mature miRNA, is loaded into an Argonaute protein (AGO) to form a miRNA-induced silencing complex (miRISC), whereas the other strand, the miRNA*, is degraded. Once loaded into the

Abbreviations: miRNA, microRNA; 3'UTR, 3'-Untranslated region; AGO, Argonaute protein; miRISC, miRNA-induced silencing complex; microRNPs, micro-ribonucleoprotein; SE, super-enhancer; GO, gene ontology; KEGG, Kyoto encyclopedia of genes and genomes; TCGA, The Cancer Genome Atlas; ENCODE, encyclopedia of DNA elements; qRT-PCR, quantitative real-time PCR.

* Corresponding authors.

E-mail addresses: Hua.Tan@uth.tmc.edu (H. Tan), psun@wakehealth.edu (P. Sun),

Xiaobo.Zhou@uth.tmc.edu (X. Zhou).

¹ Co-first authors.

Research in context

Evidence before this study

miRNAs have long been well-known to repress target genes at either mRNA or protein level, through which regulating gene expression in a cell to a physiologically favorable balance. Recent evidence from *in vitro* studies indicates that miRNAs can also promote gene expression in particular conditions. However, there is a lack of investigation of this counterintuitive phenomenon in clinical samples. In addition, its implications in tumorigenesis, and the general mechanisms underlying this phenotype have not been systematically explored.

Added value of this study

By integrative analysis of miRNA and gene expression profile of 8375 patient samples across 31 major human cancers from The Cancer Genome Atlas (TCGA), we for the first time showed that miRNA-associated gene activation, rather than suppression, is surprisingly prevalent and conserved across multiple cancer types. We further confirmed the biological and clinical significance of the positive correlations in terms of essential biological processes they participate and cancer hallmarks they present. Additionally, we proposed and explored four potential molecular mechanisms that can well explain the observed gene upregulation associated with miRNA.

Implications of all the available evidence

The present study established a striking phenomenon regarding miRNA-associated gene activation in human cancer samples, corroborated its biological and clinical meaning, and partially explained the underlying molecular basis. Our work sheds new light on the complex miRNA-mRNA interaction and its implications in tumorigenesis.

silencing complex, the miRNA pairs to complementary sites within mRNAs or other transcripts and the Argonaute protein exerts the post-transcriptional repression [13]. Considering that ~2000 miRNA gene have been identified in human genome [14], that a miRNA can target hundreds even thousands of different mRNAs, and an individual mRNA might be influenced by multiple miRNAs [15], the miRNA biogenesis pathway plays an essential role in shaping the gene regulatory networks. Therefore, the miRNA biogenesis is elaborately maintained in a favorable balance under physiological condition, but can be severely impaired in cancer, resulting in differential expression of critical miRNAs compared to that in the normal tissues [16–18].

Based on the canonical miRNA biogenesis pathway and mechanism of function, miRNAs have long been believed to elicit their effects exclusively through mRNA degradation and/or translation inhibition. Recently, evidence has accumulated to suggest that miRNA can also promote gene expression in particular contexts via various mechanisms. For instance, under cell cycle arrest, human miRNA miR-369-3 was reported to activate tumor necrosis factor- α (TNF α) translation by directing micro-ribonucleoproteins (microRNPs) including AGO and FXR1 to a special type of miRNA binding sites called AU-rich elements (AREs) in the 3'UTR of TNF α , while it was also acknowledged that this translational activation can be switched to repression in proliferating cells [19]. In most cases, however, miRNA was reported to activate gene expression by binding to the promoter region followed by recruitment of transcription factors to the miRNA binding sites. Striking examples

include: miR-373 activates E-cadherin (CDH1) and cold-shock domain-containing protein C2 (CSDC2) by binding to their promoter regions [20], miR-205 induces the expression of tumor suppressor genes interleukin 24 (IL24) and IL32 by targeting specific sites in their promoters [21], miR-744/1186/466d-3p induces Ccnb1 expression in mouse cell lines by targeting promoter elements [22], miR-589 binds the promoter RNA and activates cyclooxygenase-2 (COX-2) transcription [23], and let-7i binds to TATA-box of IL2 and activates it at both mRNA and protein level [24]. Recently, a new mechanism regarding miRNA activation has been proposed, showing that miR-24-1 activates FBP1 and FANCC genes by targeting their enhancers [25]. In addition, these studies invariably corroborated that proteins related to miRNA biogenesis or functions, such as AGO and Dicer, or transcription related enzymes are significantly enriched surrounding the binding sites during gene activation. Collectively, these previous findings, mainly observed in *in vitro* conditions, established that miRNA can also up-regulate gene expression by directly binding to the transcriptional regulatory regions.

While the miRNA-mediated gene activation has been well studied in *in vitro* conditions and in animal experiments, it was rarely explored in human samples. To address the gap, we leveraged the Cancer Genome Atlas (TCGA) and conducted an unprecedentedly comprehensive analysis on the miRNA-gene interaction profiles in 8375 patient samples across 31 major human cancer types. We checked the correlations between all 1046 miRNAs and 20,531 genes annotated in TCGA, and found that positive miRNA-gene correlation is surprisingly prevalent and consistent across human cancers, even when the gene bears conserved binding sites for the miRNA. And the positive correlations display disparate patterns compared to the negative correlations. We performed a series of stringent bioinformatics analysis to investigate whether this positive correlation has any biological or clinical implication especially in the context of human oncogenesis. We revealed that the miRNA-gene pairs with positive correlation are extensively involved in many biological processes pertaining to immune response, cell membrane signaling, cell cycle control and other cancer hallmarks. In addition, the top ranked miRNAs and genes can well stratify patients in terms of overall survival rate based on their single or combined expression level. These results warrant the biological and clinical significance of the widespread positive correlations between miRNAs and genes across human cancers.

We further investigated the molecular mechanisms underlying the observed miRNA-gene positive correlation. Most of the positive correlations (~87%) can be explained by one or more of our proposed four indirect-regulation hypotheses, including miRNA-host gene co-expression, gene activation by inhibiting upstream suppressor, co-regulation by shared transcription factors, and co-activation by common histone modifications. Considering that co-expression by shared genetic or epigenetic factors cannot be viewed as causative relationship, our study stresses that although some positive correlations are implicated in tumorigenesis, the expression level of the miRNA and gene are independent on each other. On the other hand, mechanisms related to the miRNA-host gene co-expression and indirect activation of gene by inhibiting its upstream suppressor involve causation, in the sense that the expression level of one will influence that of the other. We further hypothesized that at least part of the remaining positive correlations (~13%) can be explained by direct binding of miRNA to particular transcriptional regulatory regions of the partner gene, as reviewed above. Our wet-lab experiments in corresponding cancer cell line only partially recapitulated the positive correlations observed in human patient samples, implying that the miRNA-gene interaction are dramatically different between *in vitro* and *in vivo* conditions. This in turn highlights the significance of a comprehensive analysis on the miRNA-directed gene activation in human samples.

2. Materials and methods

2.1. TCGA data acquisition, quality control and preprocessing

The miRNA and gene expression data, as well as clinical information of each cancer, were downloaded from TCGA data portal (<https://portal.gdc.cancer.gov/>). Three cancers, including FPPP (FFPE Pilot Phase II), GBM (glioblastoma multiforme) and LAML (acute myeloid leukemia) were excluded due to small sample size or platform inconsistency. The pan-cancer analyses eventually consisted of 31 cancer types: ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, HNSC, KICH, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS and UVM. Sample size ranged from 36 (CHOL) to 1102 (BRCA), see Table S1. A total of 1046 miRNAs and 20,531 genes (protein coding and noncoding) were included in the TCGA IlluminaHiSeq miRNASeq and IlluminaHiSeq RNASeqV2 data, respectively. We used level 3 expression data for both miRNA and gene. The miRNA expression counts the reads aligning with the corresponding precursor, while the gene expression is derived from the reads per kilobase of transcript per million reads mapped (RPKM). The miRNA and gene expression values were logarithmically transformed (base 2) prior to further analysis. Pearson correlation analysis was performed to assess the co-expression between miRNAs and genes across all 31 cancer types. A correlation was deemed significant in a cancer if its absolute correlation coefficient $|R| > 0.1$ ($R > 0.1$ for positive correlation and $R < -0.1$ for negative correlation) and Hochberg adjusted p -value $\text{adj.}P < .05$ unless otherwise stated.

2.2. CCLE data curation and correlation analysis

We downloaded cell line data for gene and miRNA expression from the Cancer Cell Line Encyclopedia (CCLE) database [26] for verification of TCGA data. In summary, the miRNA data involves 654 human miRNAs and 954 cell lines spanning 25 cancer types; and the RNAseq data contains 18,361 genes and 1019 cell lines across 26 cancer types. We calculated the miRNA-gene correlation for each pair in each of the 25 overlapped cancers. Due to relatively small sample size for each cancer type compared to TCGA, at this targeted validation step we imposed a looser restriction on the p -value cut-off ($P < .01$) for significance in CCLE. To compensate this, we used the more conservative Spearman's rank correlation instead of Pearson correlation.

2.3. Functional characterization of positive correlation associated miRNAs and genes

To explore the biological significance of the top-ranked pairs, we conducted gene ontology (GO) and KEGG [27] signaling pathway enrichment analyses on the genes involved in the top-ranked significant pairs, using the R package clusterProfiler [28]. Enrichment profiles were checked over genes at different pan-levels (PanCan10/15/20/25) respectively, with Fisher exact test p -value $< .05$ deemed significant. To further investigate the clinical relevance of these pairs, we performed survival analysis on different groups based on their miRNA and gene expression profiles. Briefly, patients were first divided into non-overlapped groups based on their miRNA and gene expression in three ways: 1) patients were divided into three groups as high-middle-low (Hi-Mi-Lo) by miRNA expression; 2) patients were divided into three groups as high-middle-low (Hi-Mi-Lo) by gene expression and 3) patients were divided into two groups as high-low (Hi-Lo) by gene:miRNA ratio. Then we adopted log-rank test to compare the overall survival probability of patients from different groups, categorized by single or combined stratification indexes. Therefore, for a particular miRNA-gene pair, the HiHi+Hi group refers to patients from high miRNA expression, high gene expression and high gene:miRNA ratio group. The stratification power in patient survival of miRNAs/genes from different pan-level groups was compared by the Kolmogorov-

Smirnov test (K–S test) performed on the density distribution of their log-rank test p -values.

To further investigate the implications of our detected positive correlations in tumorigenesis, we associated the top-ranked miRNA and genes to the well-known cancer hallmark traits. Briefly, we checked the association of the PanCan10 miRNAs with the well-known 10 cancer hallmarks proposed by Hanahan and Weinberg 2011 [1], as did in a previous study [29]. We also examined the enrichment of the top-ranked genes of different pan-levels in 50 well-established hallmark gene sets related to critical biological processes regarding cellular component, development, DNA damage, immune, metabolic, pathway, proliferation and signaling [30], which are assumedly essentially relevant to cancer initiation and progression.

2.4. Identification of miRNAs with host genes

We downloaded the gene annotations (hg38) of 27,423 genes (including 19,902 protein-coding genes and 7521 long non-coding RNAs – lncRNAs) from GENCODE (GTF v25) [31], and annotations of 1881 human miRNAs from miRBase (hsa.gff3) [14]. We determined whether a miRNA is embedded in another gene (protein-coding or noncoding RNA) according to their genomic coordinates (locations) from the annotation files. A total of 591 miRNAs were found to be located inside a specific gene, termed “host” gene. Of these, 451 pairs were covered in the TCGA data. These 591 host genes consisted of 474 protein-coding genes and 117 lncRNAs, see Table S4. We investigated the preference of a miRNA to co-express with their host gene by hypergeometric test. The conservation information of each miRNA was adopted from TargetScanHuman v7.2 (miR_Family_Info.txt). A positive conservation score (Conservation? = 1, 2) of a miRNA indicates conservation while a negative score (–1) refers to non-conservation, remaining miRNAs with a zero score were ignored (Table S4).

2.5. Detection of double-negative patterns underlying positive correlation

To validate the hypothesis that a miRNA can upregulate a gene by inhibiting its upstream suppressor, we attempted to detect the double-negative patterns. For a significant positive miRNA-gene pair, we first detected all the intermediate (IM) genes that negatively correlate with both the miRNA and gene in the pair across multiple cancers ($R < -0.1$, $\text{adj.}P < .05$, cancer coverage ≥ 5). Then we narrowed down the intermediate genes to real targets of the miRNA based on TargetScanHuman v7.2 [15]. We downloaded the predicted targets (TAR) with context++ and weighted context++ scores, followed by a series of preprocessing steps, including extraction of human species, parsing/trimming miRNA names, removing duplicates, and eventually obtained 198,312 targeting records with high confidence (i.e., with a positive probability of conserved targeting, P_{ct}), which involves 321 different miRNAs and 13,035 genes.

2.6. Detection of double-positive patterns underlying positive correlation

To validate the hypothesis that a positively correlated miRNA-gene pair might be regulated by shared transcription activators, we tried to detect the double-positive patterns. For a significant positive miRNA-gene pair, we first detected all the intermediate (IM) genes that positively correlate with both the miRNA and gene in the pair across multiple cancers ($R > 0.1$, $\text{adj.}P < .05$, cancer coverage ≥ 5). Then we narrowed down the genes by two steps. First, we restricted the IMs into general transcription factors (gTF). We downloaded transcription factors (TF) and their targets with the R data package tftargets (<https://github.com/slowkow/tftargets>). This dataset includes human TF information curated from six published databases: TRED [32], ITFP [33], ENCODE [34], Neph2012 [35], TRRUST [36], Marbach2016 [37]. We mapped the Entrez gene IDs into gene symbols using two R packages: annotate (10.18129/B9.bioc.annotate) and org.Hs.db (10.18129/B9.bioc.org).

Hs.eg.db). After integration and removal of duplicates, we obtained 2705 gTFs with their target genes, based on which we removed intermediate genes of each pair that were not included in the gTF sets. At the second step, we further narrowed down the gTFs into specific transcription factors (sTF) by removing gTFs whose target genes do not include the gene in the miRNA-gene pair under investigation.

2.7. Super-enhancer (SE) identification from ENCODE H3K27ac ChIP-seq data

We downloaded the H3K27ac ChIP-seq data (bed files for narrowPeak) of 15 human tissues including 64 samples from ENCODE (including blood, lung, liver, kidney, brain, large intestine, stomach, pancreas, esophagus, prostate gland, adrenal gland, breast, ovary, thymus and urinary bladder tissues, see Table S5) [34]. To detect the super-enhancer formation profile, we investigated the H3K27ac signal surrounding a miRNA or gene site based on the ROSE (Rank Ordering of Super-enhancers) pipeline [38] with minor modifications. Briefly, we first stitched the detected peaks if they are within a certain distance (called the “region”), and then calculated the average input-subtracted H3K27ac signal intensity within the region by a revised ROSE score:

$$ROSE_{score} = \frac{1}{L} \sum_{i=1}^N \text{width}(peak_i) \times \text{intensity}(peak_i)$$

In this formula, N is the number of peaks detected in the region, and L is the length (in bp) of the region in question. In this study, $L = 100$ K bp, centered at the transcription start site (TSS) of each miRNA or gene. The width and intensity of each peak ($peak_i$) was obtained from its genomic coordinates ($width = \text{chromEnd} - \text{chromStart}$) and signal intensity (signalValue), respectively (see ENCODE narrowPeak format description). We employed the scaled peak intensity (“score” in the bed file) instead of signalValue for better visualization. As did in the original ROSE pipeline, we adopted a promoter exclusion zone of 4 K bp, i.e., if a peak was entirely contained within a window of ± 2 K bp around the TSS, the peak was excluded from the calculation. It should be noted that in contrast to the original ROSE pipeline, we removed a limitation on the maximum distance of 12.5 K bp between two constituent enhancers, to focus on the general H3K27ac intensity surrounding the miRNA/gene TSS site. Under our framework, a region with $ROSE_{score} \geq 10$ in at least 11 out of 15 tissues was considered as a super-enhancer (SE).

2.8. Cell culture

BJ human foreskin fibroblasts were maintained in minimum essential medium supplemented with 10% fetal calf serum, nonessential amino acids, and antibiotics. 293 T, MDA-MB-231 and Huh7 cells were grown in Dulbecco's modified Eagle medium supplemented with 10% fetal calf serum, glutamine, and antibiotics. HeLa, Huh7 and T47D cells were cultured in EMEM, DMEM and RPMI-1640, respectively, all basic medium was supplemented with 10% FCS and 1% antibiotics.

2.9. Plasmids

The expression vectors for miRNAs and 3'UTR dual-luciferase reporter plasmid (pmirGLO) were purchased from Biosettia, Inc. (San Diego, CA). To construct target gene 3'UTR dual-luciferase reporters (pmirGLO-CTLA4-3'UTR, pmirGLO-IGFBP5-3'UTR, pmirGLO-ITK-3'UTR, pmirGLO-PDGFR-3'UTR, pmirGLO-PIK3CG-3'UTR, pmirGLO-TGFBI-3'UTR, pmirGLO-IL7R-3'UTR), target gene 3'UTR exons containing miRNA seed sequences were amplified by PCR from the genomic DNA of 293 T cells. The PCR primers are described in supplemental Materials. All 3'UTR fragments are inserted into pmirGLO by *NheI*-HF and *XhoI*.

The pGL3-basic plasmid was purchased from Promega Corporation. To construct pGL3-IL7R/PIK3CG-3Kb + 3'UTR reporter, a 3 kb promoter sequence from the IL7R and PIK3CG gene was amplified from the genomic DNA of 293 T cells and inserted into pGL3-basic plasmid by *NheI* and *XhoI* (IL7R promoter), or *MluI* and *BglII* (PIK3CG promoter). Then, the PCR product for the 3'UTR of IL7R or PIK3CG (still containing the miRNA seed sequence) was digested by *XbaI* and inserted into pGL3-IL7R/PIK3CG-3Kb reporter, respectively, downstream of luc+. The PCR primers are described in supplemental Materials.

2.10. Lentivirus-based gene transduction

pLV-miR-ctrl, pLV-miR-21, pLV-miR-142, pLV-miR-155, pLV-miR-214 Recombinant lentiviruses were packaged in 293 T cells in the presence of helper plasmids (pMDLg, pRSV-REV, and pVSV-G) using Lipofectamine 2000 (Invitrogen). BJ or MDA-MB-231 cells (1×10^5 /well) were seeded into 6-well plates, grown overnight, infected with 300 μ L virus in 3 mL fresh medium containing 8 μ g/mL polybrene, and spun for 1 h at 1600 to 1800 rpm. Transduced cells were purified with 1.2 μ g/mL of puromycin.

2.11. RNA isolation and quantitative Real-Time PCR

RNA was isolated from cells using TRIzol (Thermo Fisher Scientific) according to manufacturer's protocol. 500 ng of RNA was reverse transcribed to cDNA with iScript™ Reverse Transcription Supermix (Bio-Rad). Quantitative real-time PCR was performed in triplicate with gene-specific primers and SsoAdvance™ SYBR Green Supermix (Bio-Rad) in a Bio-Rad CFX96 REAL TIME SYSTEM following manufacturer's protocols. GAPDH was used as internal control to normalize the mRNA input for each gene. qPCR primers are described in supplemental Table S7.

2.12. Dual-luciferase reporter assay

The target genes 3'UTR activity was analyzed in both 293 T and MDA-MB-231 cells by transient transfection of luciferase reporter constructs. On the 1st day, 6×10^5 /well 293 T or 1.5×10^5 /well MDA-MB-231 cells were seeded into 12-well plates. These cells were transfected with 0.17 μ g pmirGLO reporter vector, 1.43 μ g pLV-miR-ctrl/pLV-miR-142 vector and 4.0 μ L lipo-2000 according to manufacturer's instruction on the next day. 48 h after transfection, cell lysates were collected using Passive Lysis Buffer (E1941, Promega). Firefly and Renilla luciferase activity was detected using Dual-Luciferase Reporter Assay System (E1960, Promega) on GloMax®-Multi+ Microplate Multimode Reader (Promega).

For gene IL7R and PIK3CG 3 kb promoter + 3'UTR reporter analysis, MDA-MB-231 cells were transiently transfected with 0.16 μ g pGL3-IL7R/PIK3CG-3Kb + 3'UTR plasmid, 1.36 μ g pLV-miR-ctrl/pLV-miR-142 vector and 0.08 μ g of the control Rluc vector driven by β -actin, TK or CMV promoter, using 4 μ L lipo-2000 according to manufacturer's instruction. Other procedures are the same with 3'UTR dual-luciferase assay.

3. Results

3.1. Positive miRNA-gene correlations are prevalent and consistent across human cancers

By an integrative miRNA-gene correlation analysis on 1046 miRNAs and 20,531 genes across 31 TCGA cancers (Table S1), we detected a total of 2,842,030 pairs that were significantly positively correlated ($R > 0.1$, $\text{adj.P} < .05$) in at least one TCGA cancer type. We then ranked these positive pairs according to the number of cancer types (called cancer coverage) in which their correlation appear to be positive ($R > 0.1$) and significant ($\text{adj.P} < .05$). In most of the subsequent analysis, we focus on the top ranked pairs with cancer coverage ≥ 10 , totaling 18,996

miRNA-gene pairs, which involves 348 miRNAs and 3074 genes (Table S2). Each of the top 56 significant positive pairs covered at least 27 cancer types (Fig. 1A). Three pairs, including miR-196b-HOXA10, miR-335-MEST and miR-483-IGF2, covered all 31 cancers under investigation. Interestingly, miR-196b and HOXA10 have been reported to co-express and their overexpression characterized poor prognosis in patients with gastric cancer [39]. The IGF2 intronic miR-483 has been

widely recognized as an oncogenic miRNA that transcriptionally upregulates its host gene IGF2 [40,41], while the long non-coding RNA (lncRNA) H19 intragenic miR-675 was shown to be the most highly conserved feature of H19 and serves as the functional regulatory unit of this lncRNA [42]. These previous studies corroborated the significant biological implications of the positive correlations that are prevalent across multiple cancer types.

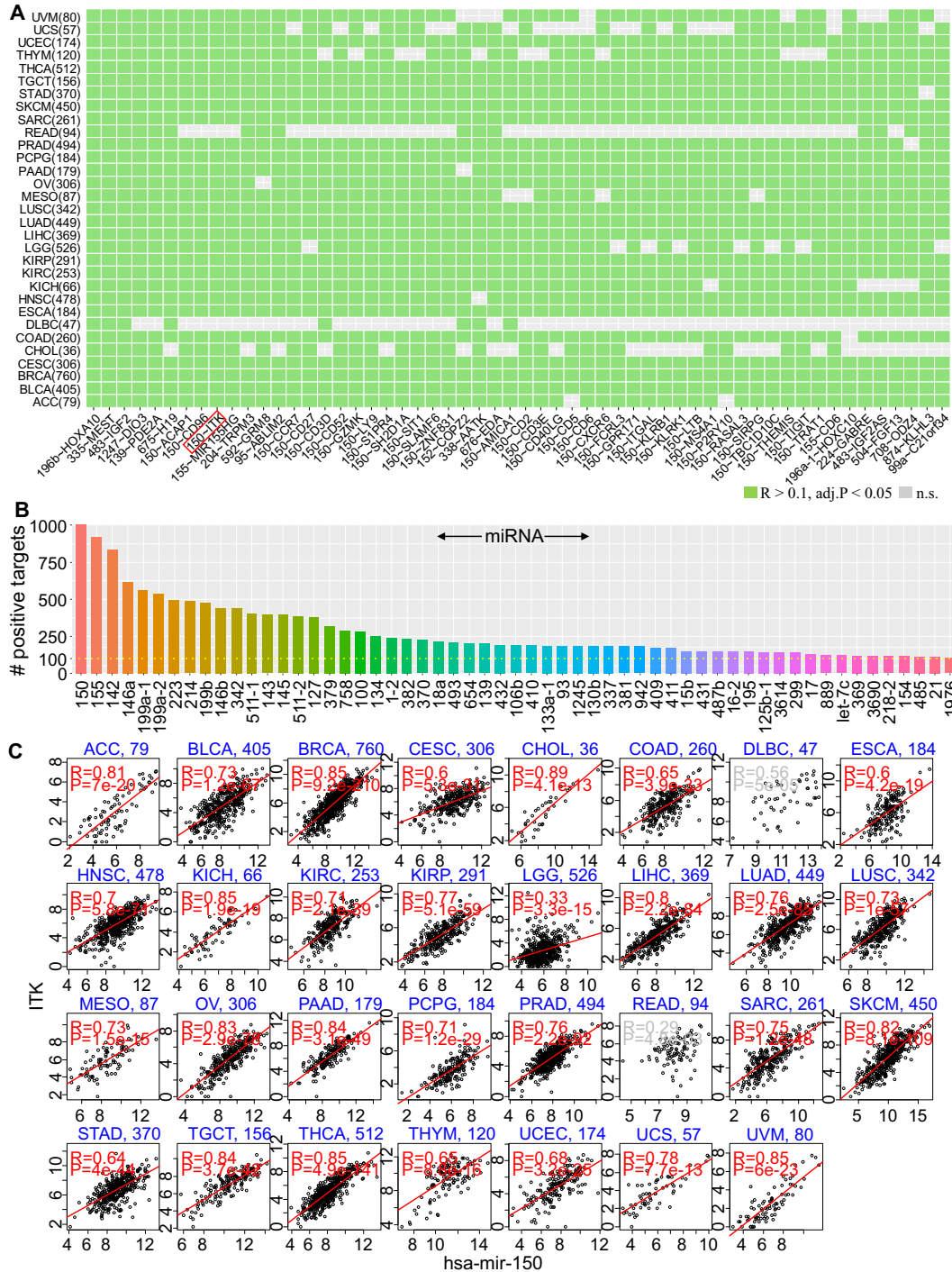
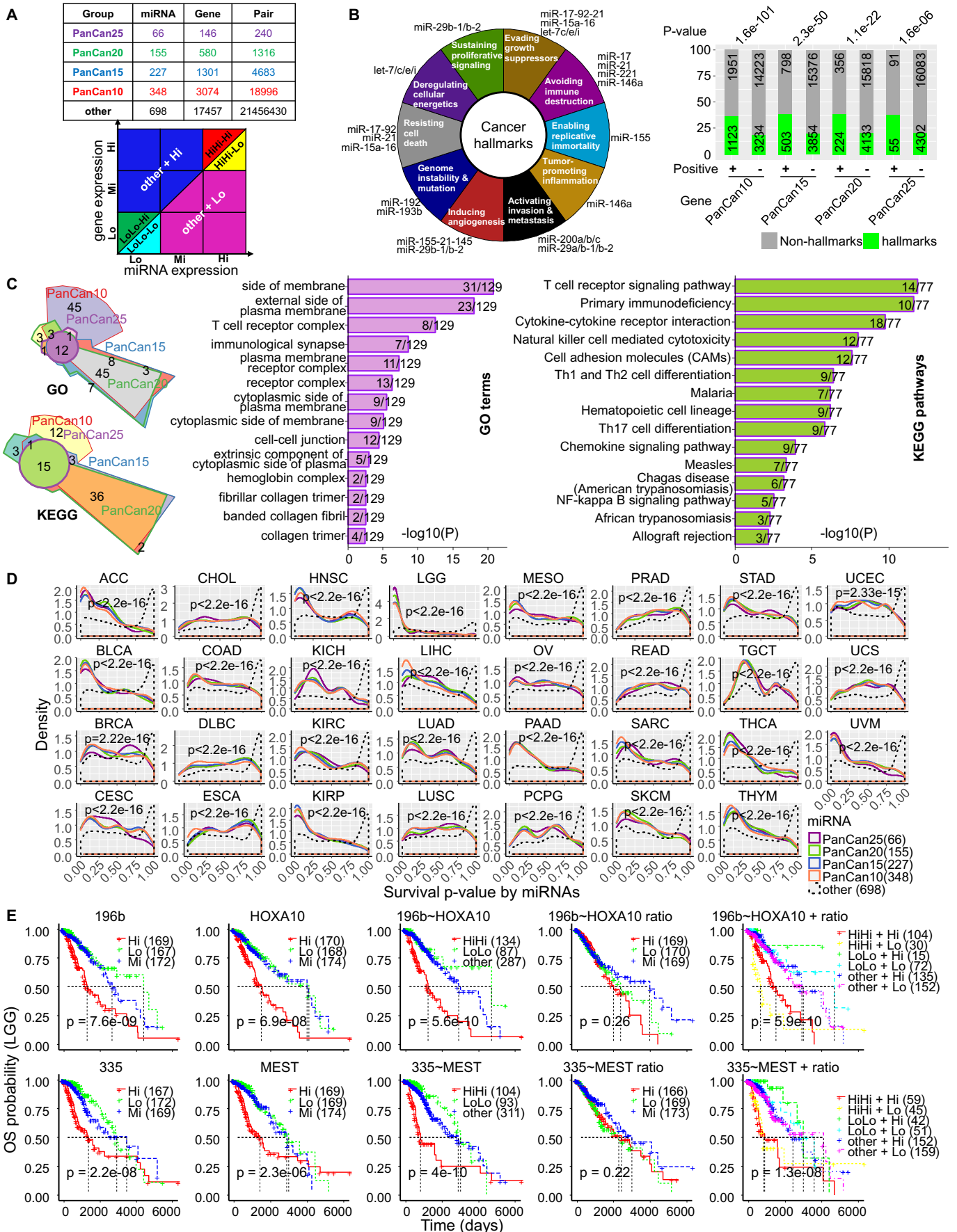


Fig. 1. Overview of miRNA-gene positive correlation landscape in human cancers. (A) Heat map showing top positively correlated miRNA-gene pairs covering ≥27 cancer types. TCGA cancer names and corresponding sample sizes used for correlation calculation is shown on the left. Pairs are shown in the format of miRNA-gene at the bottom, with human miRNA prefix (hsa-mir-) omitted for better visualization. A positive correlation with Pearson correlation coefficient $R > 0.1$ and Hochberg adjusted p -value < 0.05 was deemed significant. n.s.: non-significant. (B) Top miRNAs ranked by the number of its positively correlated genes. For better visualization, only miRNAs with ≥100 targets and across ≥10 cancer types are shown. (C) Detailed correlation profiles of a representative positive pair hsa-mir-15-ITK across 31 TCGA cancers. The sample size of each cancer, Pearson correlation coefficient and p -value are also shown. Red color indicates significant positive correlation. See also Fig. S1–3.



We further extracted the miRNAs and genes from the top-ranked pairs with cancer coverage ≥ 10 , mapped each miRNA to all genes positively correlated with it (positive targets), and ranked the miRNAs according to the number of their positive targets. The top 57 miRNAs were found to positively correlate with at least 100 genes (Fig. 1B). We also explored the genomic spatial distribution of the positive targets of each miRNA, and found that their positive targets tend to distribute along the whole genome evenly (Fig. S2). The top ranked miRNAs can be clustered into several groups based on their positive targets, and miRNAs in the same group tend to share a large number of positively correlated genes (Fig. S3). miR-150, a miRNA located at chromosome 19, turns out to be the most active miRNA, which positively correlates with 1009 different genes each covering at least 10 cancers. It is an important hematopoietic cell (especially B-cell, T-cell and natural killer cell)-specific miRNA and has been widely reported to be differentially expressed in cancers especially in hematopoietic malignancies compared to normal controls [43]. A detailed correlation profile for a representative pair miR-150 ~ ITK (top 9) shows that they are positively correlated in 29 out of 31 TCGA cancers (Fig. 1C). Even in the exceptional scenarios in DLBC and READ cancers, miR-150 shows obvious positive correlation with ITK. It was not detected as significant due to the relatively small sample size of these two cancer types and the stringent adjusted p -value threshold in this study. Collectively, these findings demonstrate that significant positive correlation between miRNAs and genes is highly prevalent and consistent across different cancer types.

3.2. Positive miRNA-gene correlations display distinct patterns than negative correlations

Our study revealed several unprecedented patterns of miRNA-gene interactions in human patient samples. First, in comparison to negative correlations, the miRNA-gene positive correlations are much more significant, prevalent and consistent across cancer types (Fig. 1 and Fig. S1). Using the same thresholds of correlation coefficient and p -value, we detected 18,996 miRNA-gene pairs that are positively correlated in at least ten cancer types (PanCan10), compared to 1514 pairs with negative correlation (Fig. S1B). And the negative correlations generally cover much less cancer types than the positive ones (Fig. S1A, S1D). Interestingly, the constituent miRNAs and genes of these positive and negative pairs are largely overlapped (Fig. S1B), suggesting that there might be a constant and small group of miRNAs and genes taking effect in most biological processes. We also noted that about 17% (3223/18996) of PanCan10 positive pairs are retained in the normal samples (Fig. S1E), indicating that some miRNAs can function by upregulating particular genes under both physiological and oncogenic conditions. Second, a miRNA can significantly correlate with multiple genes and *vice versa*, potentially forming a multiple-to-multiple regulation network, in both positive and negative correlation scenarios. While a miRNA can significantly correlate with an incredibly large number of genes, for example, miR-150 positively correlate with 1009 genes in PanCan10 group and miR-141 negatively correlate with 1140 genes in PanCan7 negative group (Fig. 1B and Fig. S1C), their correlated partners tend to distribute nearly evenly across the whole genome, instead of clustering in particular genomic regions (Fig. S2).

Third, miRNAs in the top-ranked positive correlations can be well clustered according to their correlated partner genes. Particularly, the top three miRNAs miR-150/155/142 share 750 partner genes, an overwhelmingly major part of each individual's partners. Similar pattern exists in the miR-214/199b/199a-2/199a-1 and miR-337/409/431/654 clusters (Fig. S3). Together, these results revealed many previously unrecognized miRNA-gene interaction patterns in human cancer, especially the difference between miRNA-gene positive and negative correlations.

3.3. Positive miRNA-gene correlations demonstrate biological and clinical significance

We defined the cancer coverage of a significant miRNA-gene pair as the number of cancer types in which this pair shows positive correlation. To better focus on the top-ranked significant pairs (cancer coverage ≥ 10 , Table S2), we categorized the pairs into four groups with increasing pan-levels: PanCan10 (covering ≥ 10 TCGA cancer types, $n = 18,996$), PanCan15 ($n = 4683$), PanCan20 ($n = 1316$) and PanCan25 ($n = 240$) (Fig. 2A, top). To explore the biological and clinical significance of the miRNAs and genes involved in the top-ranked positive pairs, we investigated their implications in cancer initiation and progression from three perspectives. First, we examined the associations of the miRNAs with well-known cancer hallmarks, and the enrichment of the genes in established hallmark gene sets of different important biological processes. Second, we explored the major biological processes that the genes participate by GO term and KEGG signaling pathway enrichment analysis. Third, we assessed the stratification power of these miRNAs and genes in characterizing the patient overall survival rate, by dividing patients into different groups based on the expression of the constituent miRNAs and/or genes of each positive pair (Fig. 2A, bottom).

We found that all previously reported cancer hallmark miRNAs [29] were covered by the PanCan10 miRNAs (Fig. 2B, left). In addition, genes from all pan-level groups were significantly enriched in the pooled hallmark gene sets of essential biological processes [30] including cell development, metabolism, immune and proliferation, which are assumedly critical for cancer development and progression (Fig. 2B, right; Table S3).

The enriched gene ontology terms and signaling pathways for genes at different pan-levels converged to 14 significant GO terms and 15 significant KEGG pathways (Fig. 2C, left; Fig. S4). The enriched GO terms primarily consisted of cytoplasmic membrane and various receptor complexes on the membrane (Fig. 2C, middle), while the enriched KEGG pathways largely covered the immune signaling related processes, including T cell receptor signaling and T helper cell (Th1, Th2, and Th17) differentiation, natural killer cell cytotoxicity and NF-kappa B signaling pathway (Fig. 2C, right).

We used miRNAs from different pan-level groups to stratify patients in terms of overall survival rate, depicted by survival log-rank test p -values. P -value density showed a significant left-shift of p -value in all pan-level miRNA groups across all cancers, compared to the control group – the remaining 698 miRNAs not in any pan-level group (Fig. 2D). This indicates that miRNAs in the top-ranked positive pairs represent much better discriminator of patient survival

Fig. 2. Biological and clinical significance of miRNA-gene positive correlations. (A) Top: Number of miRNAs, genes, and pairs in each pan-level group. The group “other” refers to pairs not covered by PanCan10. Bottom: Patient samples were divided into different groups based on their miRNA and/or gene expression. (B) Left: Association between top-ranked (PanCan10) miRNAs and cancer hallmarks adapted from Hanahna and Weinberg 2011. Right: Enrichment of top-ranked genes of different pan-levels in 50 hallmark gene sets (pooled together) summarized in Liberzon et al. 2015. Numbers inside bars represent the number of genes detected in that group. +/– refers to genes from/not from the top-ranked positive pairs of corresponding pan-levels. P -values were assessed by hypergeometric test. (C) Left: Chow-Ruskey diagrams showing numbers of significant GO terms and KEGG signaling pathways ($P < 0.05$) participated by genes from four pan-level groups as shown in (A). Middle: GO terms shared by genes from the four pan-level groups. Right: KEGG processes shared by genes from the four pan-level groups. (D) Density distributions of p -values of survival analysis (log-rank test) based on expression of miRNAs from indicated pan-level groups. The stratification power in patient survival of miRNAs from different pan-level groups was compared by K–S test. Cancer type and p -value in K–S test are shown on each subgraph. (E) Representative survival analysis results for the top two (ranked by cancer-coverage as in Fig. 1A) positive pairs – miR-196b-HOXA10 and miR-335-MEST in lower grade glioma (LGG) cancers. From left to right panels, patients were stratified into different groups as shown in (A) by miRNA only, gene only, miRNA/gene combination, miRNA/gene ratio (gene:miRNA), miRNA/gene combination and ratio, followed by a log-rank test in each stratification scheme. OS: overall survival. See also Figs. S4–S5.

compared to the other miRNAs. This trend is also applicable to the genes from different pan-level groups (Fig. S5A). In addition, we noticed that in most cancers combined miRNA and gene expression could significantly improve the stratification power of miRNA or gene alone, while further integration of the gene:miRNA ratio showed stratification performance comparable to the optimal method of all other scenarios (Fig. S5B). Take the top two positive pairs miR-196b-HOXA10 and miR-335-MEST in lower grade glioma (LGG) cancer as an example (Fig. 2E), the associated miRNA and gene alone can well stratify the patients individually, while combination of the pair further improved the stratification. Specifically, patients with both higher miRNA and gene expression exemplified much poorer prognosis than those with both lower miRNA and gene expression, or all others absent from these two groups, and this distinction is more significant in comparison with the scenarios of miRNA or gene alone. When the gene:miRNA ratio was introduced and patients were further divided into smaller subgroups, the stratification significance remained comparable.

3.4. Intragenic miRNAs tend to but do not always co-express with their host genes

Some miRNAs are intragenic and hence assumedly to be transcribed together with their host transcription unit, which can be a protein-coding gene or a long non-coding RNA. This hypothesis has been leveraged to facilitate identification of miRNA target genes, based on the correlation between its candidate targets and host gene [44]. We reasoned that a part of the top-ranked positive correlations detected in our study might be ascribed to this intragenic co-expression hypothesis. We checked the genomic locations of all miRNAs and genes from annotation databases, and obtained 591 miRNAs that are embedded in another gene (host gene), 451 of which are covered in the TCGA data. Hypergeometric test results demonstrated a consistently strong tendency of co-expression between intragenic miRNA and its host gene at all pan-levels (Fig. 3A, Table S4). When all pairs covering at least one cancer type were pooled together, the tendency remained significant ($P = 8 \times 10^{-108}$). In addition, conserved miRNAs (conservation adopted from TargetScanHuman, see Methods and Table S4) are more likely to co-express with their host genes compared to their non-conserved counterparts ($P = 5.3 \times 10^{-25}$), consistent well with previous research [45,46].

As an example, the insulin-like growth factor-2 (IGF2) and its intragenic miRNA, miR-483, which is located on intron 2 of IGF2, are positively correlated at high significance level across all 31 cancers (Fig. 3B). Additionally, miR-483 was found to co-express with all IGF2 isoforms, including INS-IGF2 (insulin, isoform 2) and IGF2-AS (IGF2 antisense RNA). The positive correlation is retained in the normal tissue samples, albeit with compromised significance level due to small sample size. As a control, miR-483 shows little correlation with the house-keeping gene GAPDH in all cancers studied.

In another scenario where multiple miRNAs reside within the same gene, we observed two distinct phenotypes. The miRNA-hosting gene GABRE (Gamma-Aminobutyric Acid Type A Receptor Epsilon Subunit), which encodes a ligand-gated ionic channel, gamma-aminobutyric acid (GABA), significantly positively correlates with both of its intragenic miRNAs, miR-224 and miR-452 in most cancers (Fig. 3C). On the other hand, neither of the two embedded miRNAs, miR-1306 and miR-3618, showed significant correlation with their shared host gene DGCR8 (DiGeorge Syndrome Critical Region Gene 8), in any of the studied cancers (Fig. S6A). Not surprisingly, DGCR8 as an essential microprocessor complex component is constantly highly expressed across all cancers studied, and exemplifies negligible variance among patient samples (Fig. S6B). In contrast, its hosted miRNAs, miRNA-1306 and miR-3618, both generally expressed at low levels in all cancers, and showed much higher inter-sample variation, leading to little correlation with their host gene DGCR8. Additional examples phenocopied the

second scenario largely due to negligible expression level of associated miRNAs (Fig. S6D–G). Further analysis excluded the potential impact of relative genomic location of the intragenic miRNA on the positive correlation with its host gene (Fig. S6C).

3.5. miRNAs can activate a gene by inhibiting its upstream suppressors

An observed positive miRNA-gene correlation might also result from indirect regulation. Specifically, a miRNA can target and inhibit an intermediate gene, which in turn serves as an upstream suppressor of a gene positively correlated with this miRNA, forming a double-negative loop. To figure out how many positive pairs were governed by this mechanism, we first identified all intermediate (IM) genes for each positive pair that are significantly negatively correlated with both components (the miRNA and gene) of the pair ($R < -0.1$, $\text{adj.}P < .05$). To ensure consistency, we required that this double-negative correlation holds significant in at least five cancer types. Under this criterion, we detected no intermediate genes in a large portion of the positive pairs, with percentage of non-IM ($N = 0$) ranging from 70.8% in PanCan10 to 56.9% in PanCan20 groups, respectively (Fig. 4A–B, top). When we narrowed down the general IM genes to highly-confident miRNA targets (IM_TAR) of the associated miRNA (predicted by TargetScanHuman with high confidence), the percentage further increased to 90.7% in PanCan15 and 83.7% in PanCan20 groups, respectively (Fig. 4A–B, bottom). These results indicate that an overwhelming part of the top-ranked positive correlations cannot be explained by this double-negative indirect regulation mechanism (Fig. 4C). Interestingly, pairs in high pan-level groups (PanCan20/25) were more likely to bear intermediate genes compared to their counterparts in low pan-level groups (PanCan10/15), in both IM and IM_TAR scenarios.

As a typical example, we detected 912 IM genes and 86 IM_TAR intermediate genes respectively for the positive correlation pair miR-141-C1orf172. Among the 86 IM_TAR genes, a transcription factor gene ZEB1 (zinc finger E-box binding homeobox 1) was found to negatively correlate with both miR-141 and C1orf172 across 14 cancers (Fig. 4D). The miR-141-C1orf172 itself was significantly positively correlated in 10 out of the 14 cancers, while in the remaining four cancers the significance is marginal. We further checked the specific expression profiles of these two components in this double-negative loop in the same 14 TCGA cancer tissues. Results showed that miR-141 and gene C1orf172 were highly expressed while the IM_TAR gene ZEB1 downregulated in tumor samples than in normal controls (Fig. 4E, top). The correlation and differential expression analysis corroborated that the inhibition effect of miR-141 on ZEB1 was potent, assumedly owing to the three highly conserved binding sites of miR-141 on the 3'UTR of ZEB1 (Fig. 4E, bottom). Additional examples regarding positive pairs miR-200c-C1orf172 and miR-429-C1orf172 further confirmed this double-negative regulation mechanism (Fig. S7). These results are well consistent with those from a previous pan-cancer gene expression analysis platform [18] and cell line data from CCLE (Fig. S8). Interestingly, ZEB1 was largely recognized as an oncogene due to its contribution to epithelial-to-mesenchymal transition (EMT), and hence the miR-141/200 family members were considered as tumor suppressors because they target and potently inhibit ZEB1 expression [47,48]. Its general downregulation in tumor samples in various TCGA cancer types further indicates the complexity of gene regulation.

3.6. Positive miRNA-gene correlations can result from co-transcription mechanisms

Analogous to the double-negative regulation, an observed positive miRNA-gene correlation can also result from double-positive correlation. Specifically, a third gene may serve as a transcriptional activator for both the constituent miRNA and gene of a studied pair. To determine how many positive pairs were governed by this mechanism, we first identified all intermediate (IM) genes for each positive pair that are

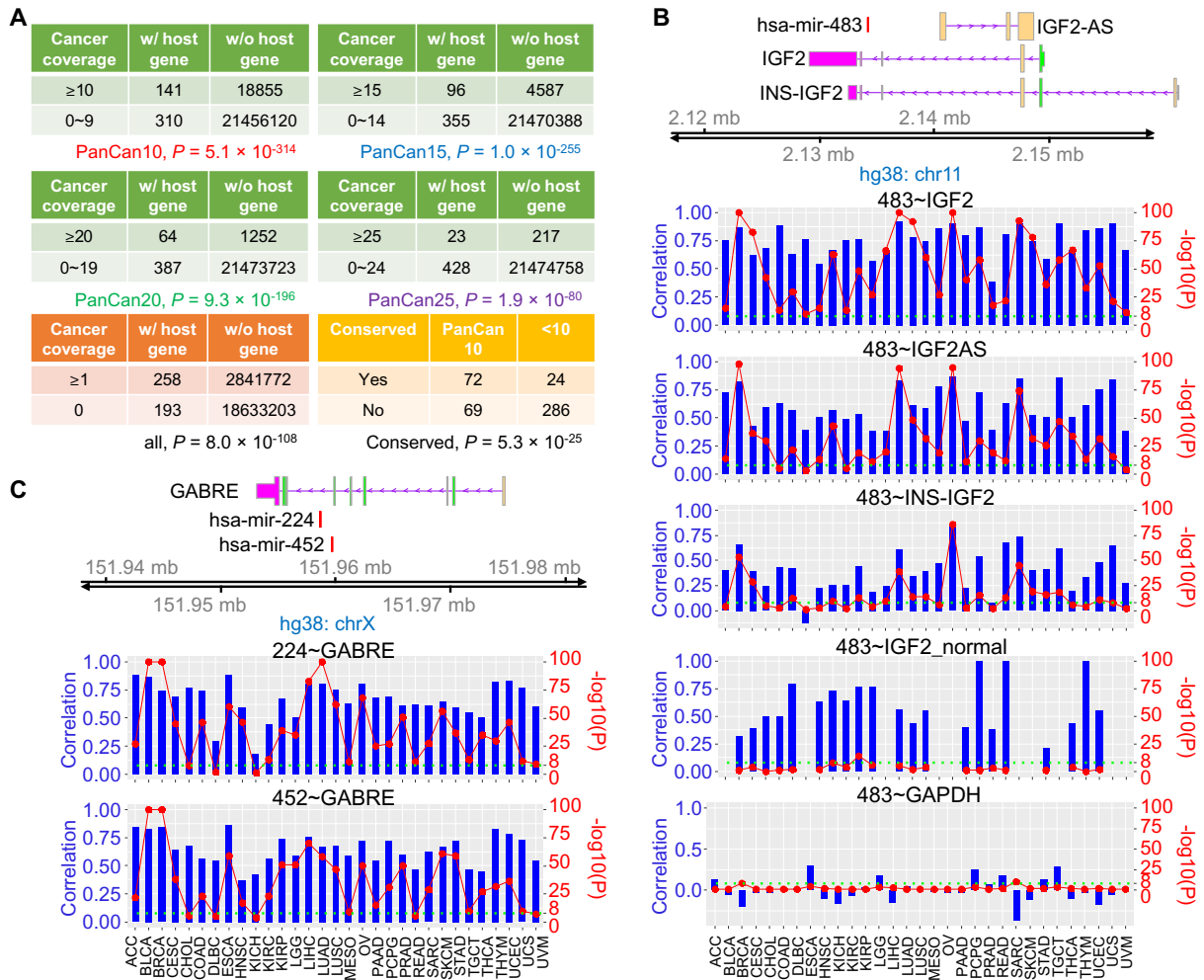


Fig. 3. miRNAs tend to co-express with their host genes. (A) Overview of intragenic miRNA–host gene co-expression patterns. The first five tables show the tendency of a miRNA–host gene pair to be positively correlated at each pan-level respectively. Cancer coverage refers to the number of cancer types in which a specific pair is positively correlated. The last table shows the tendency of an intragenic miRNA co-expressing with its host gene to be conserved across species. Conservation was predicted by TargetScanHuman. P-values were assessed by hypergeometric test. (B) Correlation between hsa-mir-483 and its host gene IGF2, isoform of the host gene, INS-IGF2, and IGF2 antisense, IGF2-AS. The correlation between hsa-mir-483 and IGF2 in normal tissues, as well as correlation between hsa-mir-483 and housekeeping gene GAPDH are also shown for comparison and as control. Pearson correlation coefficient and significance level ($-\log_{10}(P)$) are illustrated by blue bars and red solid circles respectively. A threshold $-\log_{10}(P) = 8$ is marked by a green dotted line. Normal tissues which are not available in TCGA are left blank. (C) Correlation between gene GABRE and its two embedding miRNAs: hsa-mir-224 and hsa-mir-452. Shape and color code are same as (B). See also Fig. S6 and Table S4.

significantly positively correlated with both the constituent miRNA and gene ($R > 0.1$, $\text{adj.}P < .05$). To ensure consistency, we required that this double-positive correlation holds significant in at least five cancer types, as in the double-negative scenario. Under this criterion, we found only a negligible portion of pairs having no intermediate genes, with percentage of non-IM ($N = 0$) ranging from 0.9% in PanCan10 to 2.1% in PanCan25 groups (Fig. 5A, top). Surprisingly, a considerable portion of pairs presented >1000 intermediate genes in all pan-level groups, ranging from 17.4% in PanCan10 to 82% in PanCan20 groups. These patterns are largely distinct from those observed in the double-negative scenario. When we narrowed down the general intermediate genes to general transcription factors (IM_gTF), *i.e.*, those have been reported as a transcription factor in at least one of six previous studies (see Methods), the number of intermediate genes decreased dramatically. The $N > 1000$ group disappeared, while the $101 \leq N \leq 1000$ group replaced it to dominate, and the $N = 0$ group increased a little (Fig. 5A, middle). When we further narrowed down the IM_gTF genes to those TFs that have been reported as a specific TF (IM_sTF) of the constituent gene in the pair, the number of intermediate genes further significantly decreased. About 20–30% of pairs lost all intermediate genes, while all the remaining pairs presented <100 intermediate genes

(Fig. 5A, bottom). In contrast to the double-negative scenario, which showed monotonically decreasing number of positive pairs over number of intermediate gene N , the number of positive pairs in double-positive scenario displayed a normal-like distribution over number of intermediate genes, especially in the IM and IM_gTF cases (Fig. 5B). These results indicate that a large part (70–80%) of the top-ranked positive correlations might be explained by this double-positive co-transcription mechanism (Fig. 5C).

We then focused on the specific transcription factors that served as the intermediate double-positive genes (IM_sTF). We noted that TFs tended to be shared by multiple positive pairs, and each of the top 45 specific TFs covered at least 800 pairs (Fig. 5D, top), implying that some transcription factors are highly active and tend to function as common activators of many target genes. Striking examples include the friend leukemia integration 1 (FLI1) transcription factor, which presented as an IM_sTF of 4357 pairs spanning 40 miRNAs and 800 genes, and the cell division cycle protein 20 (CDC20), which appeared in 1145 positive pairs involving 44 miRNAs and 162 genes (Fig. 5D, bottom). Looking into the correlation details regarding the detected double-positive loop miR-18a–CDC20–CDCA5, we found that they were positively correlated in 16 TCGA cancers in a pairwise manner

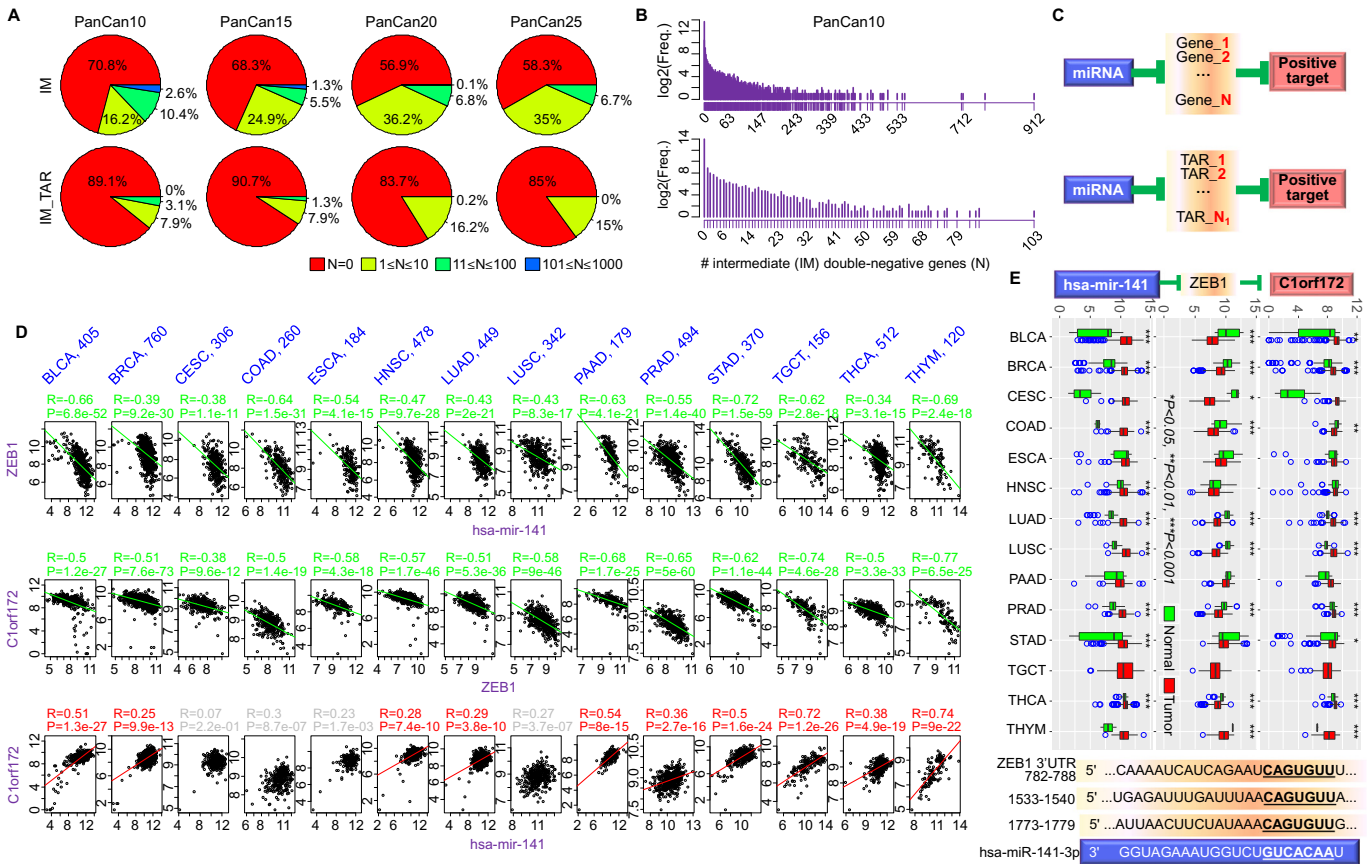


Fig. 4. miRNAs can activate a gene by inhibiting its upstream suppressors. (A) Top: Percentage of positive pairs with N ($0 \leq N \leq 1000$) detected intermediate genes (IM) that are negatively correlated with both (double-negative) the miRNA and gene in the pair at different pan-levels. Bottom: Percentage of positive pairs with N ($0 \leq N \leq 1000$) IMs predicted to be targets of the associated miRNA in TargetScanHuman (IM_TAR). The number (N) of intermediate genes for each pair is color-coded in the legend. (B) Frequency (\log_2 transformed) distribution of number of intermediate double-negative genes for PanCan10 positive pairs for IM (top) and IM_TAR (bottom) scenarios as in (A). (C) Schematic illustration of double-negative regulation for IM (top) and IM_TAR (bottom) scenarios as in (A). (D) A representative example (hsa-mir-141–ZEB1–C1orf172) of the double-negative regulation mechanism for positive correlation (IM_TAR scenario). Top: Correlation between hsa-mir-141 and one of its target genes ZEB1 (predicted by TargetScanHuman with high confidence). Middle: Correlation between ZEB1 and C1orf172, one of those genes positively correlated with hsa-mir-141. Bottom: Correlation between hsa-mir-141 and C1orf172. Correlations were assessed by Pearson correlation coefficient and p -value across 14 TCGA cancers, with sample sizes indicated above corresponding columns. Green: significant negative correlation; Red: significant positive correlation; Gray: non-significant correlation. (E) Top: Expression level of hsa-mir-141, ZEB1 and C1orf172 in the same 14 cancer tissues as in (D), and in comparison with their normal controls if available. The middle line in the box is the median, the bottom and top of the box are the first and third quartiles, the whiskers extend to 1.5 IQR of the lower quartile and the upper quartile respectively, and the blue hollow circles represents outliers. Expression difference significance (p -value) was obtained from two-sided student t -test. Bottom: Presence of binding sites of hsa-mir-141 on the 3' untranslated region (3'UTR) of ZEB1 predicted by TargetScanHuman. See also Fig. S7–8.

(Fig. 5E), which is retained in 4 CCLE tissue types including skin ($n = 49$), esophagus ($n = 25$), hematopoietic and lymphoid tissue ($n = 164$) and fibroblast ($n = 36$) (data not shown). FLI1 has been recognized as a prototype oncogene, whose aberrant expression is associated with chromosome abnormalities leading to an oncogenic fusion gene EWSR1/FLI1 with strong transforming capabilities [49,50]. Additional examples regarding FLI1's correlation in miR-139~PDE2A and miR-142~ITK pairs further confirmed its critical role in gene regulation (Fig. S9). CDC20 is a regulatory factor interacting with other cell cycle related proteins and has recently been claimed as a potential novel therapeutic target for cancer treatment [51]. These results documented the biological significance of the detected intermediate genes as specific transcription activators, and consolidated the potential co-transcription mechanism.

3.7. miRNAs/genes with positive correlations tend to form super-enhancer-like regions

Super-enhancers are an emerging subclass of regulatory regions that have been suggested to drive the processing and production of master miRNAs crucial for cell identity and oncogenesis (i.e. associated with cancer hallmarks) [29]. Super-enhancers are described as a type of

enhancer domains that are densely occupied by the master regulators (Oct4, Sox2 and Nanog) and Mediator [52]. In the established ROSE (Rank Ordering of Super-Enhancers) pipeline [53], super-enhancers were identified based on the ChIP-seq signal of a typical enhancer marker histone H3 lysine 27 acetylation (H3K27ac). We analyzed the H3K27ac ChIP-seq data of 64 samples across 15 human tissues from ENCODE, based on a modified ROSE pipeline (see Methods). We detected significant enrichment of H3K27ac signal surrounding the miRNA/gene sites associated with the top-ranked positive pairs. The top 64 miRNAs ranked by their average ROSE score over 15 tissues displayed discernible patterns in different tissue types (Fig. 6A). The top 64 miRNAs themselves can also be clustered into several groups with clear group-specific signatures. These results verified previous findings that super-enhancers are able to mark master miRNAs crucial for cell identity.

We then compared the H3K27ac enrichment profiles between positive pairs and pairs randomly chosen from the complementary set of PanCan10 group. The percentage of pairs covering >10 out of 15 ENCODE ($11 \leq N \leq 15$) tissues steadily increased with pan-levels (from 21% in PanCan10 to 38.3% in PanCan25) in the positive pair group, while the randomly chosen pairs remained at much lower percentage across pan-levels (Fig. 6B). Frequency distribution analysis of the ROSE

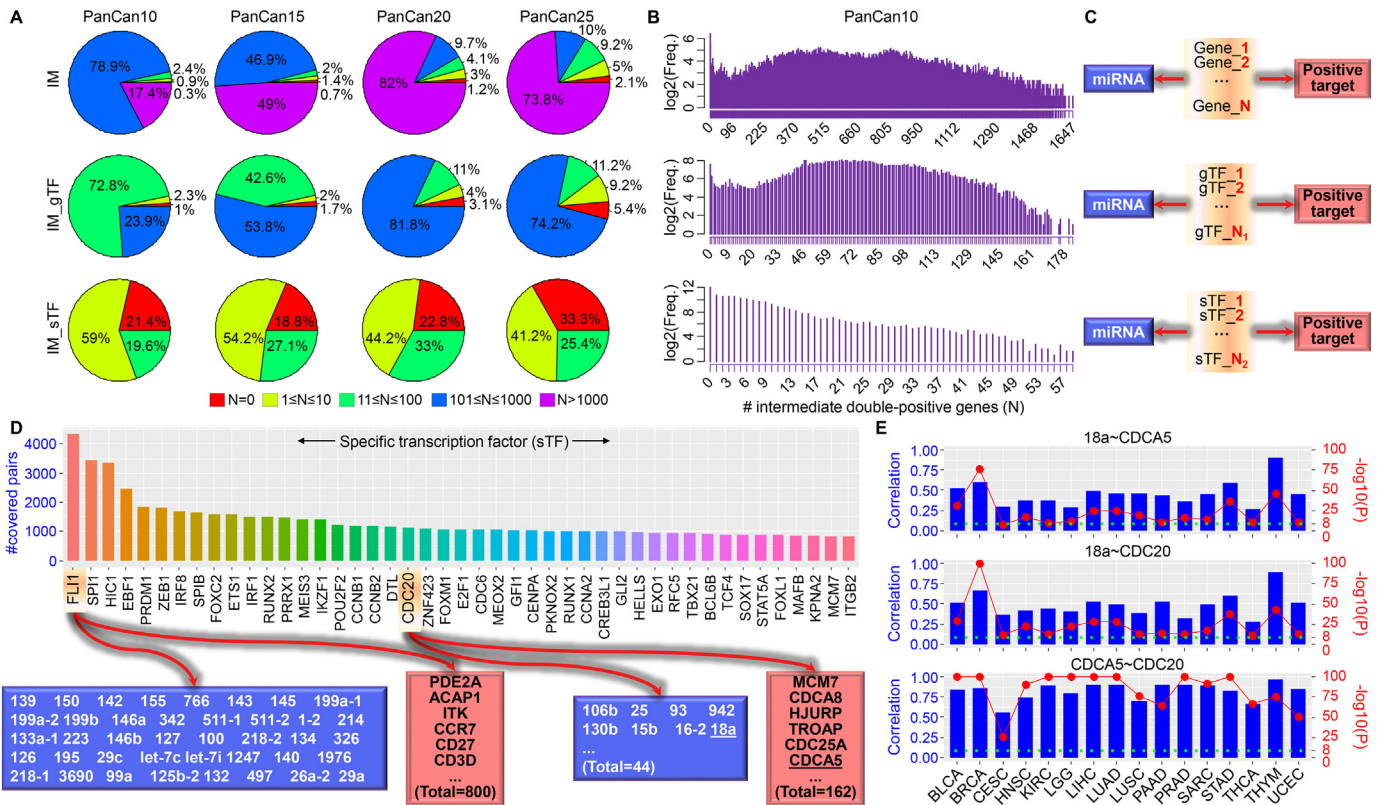


Fig. 5. Positive miRNA-gene correlations can result from co-transcription mechanisms. (A) Top: Percentage of positive pairs with $N (N \geq 0)$ detected intermediate genes (IM) that are positively correlated with both (double-positive) the miRNA and gene in the pair at different pan-levels. Middle: Percentage of positive pairs with $N (N \geq 0)$ IMs previously reported as a general transcription factor (IM_gTF). Bottom: Percentage of positive pairs with $N (N \geq 0)$ IMs previously reported as a specific transcription factor targeting on the gene in the positive pair (IM_sTF). The number (N) of intermediate genes for each pair is color-coded in the legend. (B) Frequency (log₂ transformed) distribution of number of intermediate double-positive genes for PanCan10 positive pairs for IM (top), IM_gTF (middle) and IM_sTF (bottom). (C) Schematic illustration of the double-positive regulation (co-transcription) for IM (top), IM_gTF (middle) and IM_sTF (bottom) scenarios. (D) Specific transcription factors (sTF) appearing in multiple positive pairs. The sTFs are ranked by the number of positive pairs they covered (top), with two representative examples for TFs FLI1 (involved in 4357 pairs) and CDC20 (involved in 1145 pairs) detailed (bottom). (E) A representative example showing transcription factor CDC20 was significantly positively correlated with both the miRNA and gene in a pair across 16 TCGA cancer types. Pairwise correlation among the miRNA hsa-mir-18a, the gene CDCA5 and the associated TF CDC20 are shown. Pearson correlation coefficient and significance level ($-\log_{10}(P)$) are illustrated as blue bars and red solid circles respectively. A threshold $-\log_{10}(P) = 8$ is marked by a green dotted line. See also Fig. S9.

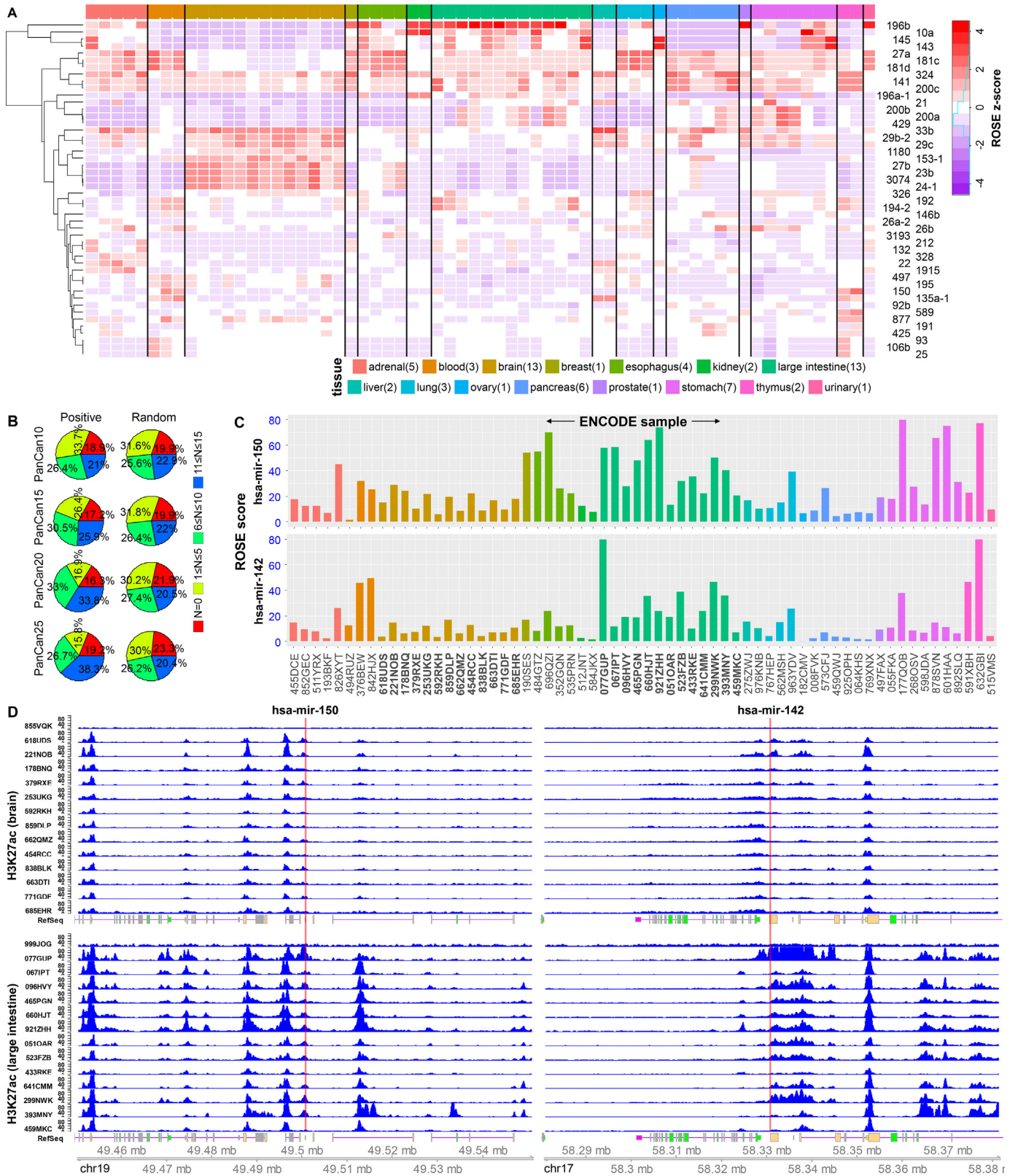
score of the PanCan25 positive pairs and randomly chosen pairs showed a right shift of the PanCan25 group (Fig. S10), indicating that positive pairs tend to yield a higher ROSE score than random ones. We then went into the detail of two representative miRNAs miR-150 and miR-142, which rank the top 1st and 3rd based on number of positive targets (Fig. 1B) respectively, and detected high intensity of H3K27ac signal surrounding their TSS sites across all 15 ENCODE tissues studied (Fig. 6C). Furthermore, while inter-tissue variation exists, samples of same tissue type showed considerable consistency in H3K27ac enrichment at the miR-150 and miR-142 loci, as exemplified by the two representative tissues brain and large intestine (Fig. 6D). By checking the top 10 pairs enriching for H3K27ac signal in various ENCODE tissues, we found miR-196b/150/142/155 and the HOX family genes are involved most frequently (Fig. S10–11). Together, these results revealed that the positive correlation associated miRNA/gene sites are more likely to enrich H3K27ac histone modification in a special manner that mimics the established super-enhancer pattern.

3.8. Positive miRNA-gene correlations are partially retained *in vitro* conditions

To check whether the positive miRNA-gene correlations detected in patient samples are also retained in the *in vitro* condition, we conducted a series of wet-lab experiments in several cancer cell lines as well as BJ primary human fibroblast cells. We selected as prototype 8 miRNA-gene pairs that are discernably detectable and positively correlated across

TCGA cancers, including miR-21-TGFB1, miR-142-CTLA4/IL7R/PIK3CG, miR-155-CTLA4/ITK, and miR-214-PDGFR/IGFBP5 (Fig. S12A–B). Among these 8 pairs, miR-21-TGFB1 and miR-214-IGFBP5 are not in the PanCan10 group. All of the 8 pairs except miR-214-PDGFR possess miRNA binding sites in the 3'UTR of the corresponding partner gene (Fig. S12C). To select proper cell lines for experiments, we checked the gene expression in several candidate cell lines derived from human cancers, including breast cancer (MDA-MB-231 and T47D), liver cancer (Huh7) and cervical cancer (HeLa), as well as primary human fibroblast cells (BJ). We found that these genes are expressed in a highly cell line-specific manner (Fig. 7A, bottom and Fig. S12D). However, these genes are generally highly expressed in all TCGA cancer types (Fig. 7A, top and Fig. S12B), which demonstrates difference between *in vitro* and *in vivo* conditions.

We further found that the positive regulation was retained in only two pairs miRNA-142 ~ (IL7R, PIK3CG) in MDA-MB-231 breast cancer cells, while in the other pairs, there was either no positive regulation at all or positive regulations that were consistent but not statistically significant, in the breast cancer or BJ cells (Fig. 7B bottom), in contrast to all positive correlations in TCGA samples (Fig. 7B top). Then we investigated the mechanisms underlying this positive regulation of IL7R and PIK3CG by miR-142 in the subsequent experiments. We did 3'UTR reporter assay in two cell lines, 293 T (Fig. 7C left) and MDA-MB-231 (Fig. 7C right), and found that co-transfection of miR-142 did not increase the 3'UTR reporters of IL7R and PIK3CG, suggesting that the positive regulation is not mediated by 3'UTR, which is different from the



mechanism for the negative regulation of genes by miRNA. Next, we combined promoter and 3'UTR of each gene in the reporter, and miRNA still has no positive effects on these reporters (Fig. 7D), implying that the regulation is not achieved *via* the promoters or the combination of the promoters and 3'UTRs either. Taken these results together, we conclude that the positive regulations observed in tumors are only partially retained in cancer cell cultures, and that the mechanism underlying the positive regulation *in vitro* is not mediated *via* promoters or 3'UTRs of the genes, even though the promoters and/or 3'UTRs contain the miRNA binding sites. It is possible that the mechanisms for the positive regulations of genes by miRNAs *in vitro* differ from those *in vivo*. Alternatively, the positive correlation might be mediated by certain epigenetic modifications, which cannot be detected on transiently transfected reporters that are not integrated into the original locus of the genes.

4. Discussion

While miRNAs have recently been reported to activate genes in particular contexts based on *in vitro* cell culture experiments, the phenomenon and its clinical relevance have not been established in *in vivo* conditions. By a comprehensive investigation on the expression correlation between all 1046 miRNAs and 20,531 genes in 8375 patient samples across 31 major human cancers in The Cancer Genome Atlas (TCGA), we detected compelling positive miRNA-gene correlation patterns that are highly consistent (conserved) across cancer types. We further confirmed that these conserved positive correlations bear biological significance and clinical relevance. In addition, we identified potential mechanisms underlying the upregulation of genes by miRNAs, including the intragenic miRNA-host gene co-expression, miRNA inhibition of the upstream suppressor of the gene (double-negative), co-transcription by shared transcription factors (double-positive), super-enhancer mediated miRNA-gene co-expression, and direct binding of miRNA to the regulatory regions of the partner gene (Fig. 8). Therefore, the present study established a striking phenomenon regarding miRNA-directed gene activation in human cancer samples, confirmed its biological and clinical significance, and partially explained the potential mechanisms underlying this activation, which deserve further experimental exploration.

Besides revelation of these unusual and seemingly counterintuitive profiles of miRNA-gene correlations, we also confirmed their biological and clinical significance through a series of stringent analysis. The constituent miRNAs in our detected PanCan10 positive pairs covered all previously reported cancer hallmark miRNAs, while the constituent genes were significantly enriched in the established 50 hallmark processes pertaining to cell division and proliferation, demonstrating their implications in oncogenesis. Moreover, with increasing pan-levels, the constituent genes converge into a smaller set of biological signaling pathways and cellular processes regarding immune cell differentiation and membrane signaling cascades, indicating that the positive correlations largely participate in the immune response and extracellular message stimulation and subsequent intracellular transduction (Fig. 2 and Fig. S4). Most strikingly, both miRNAs and genes across all pan-levels exhibit dramatically higher capability in stratifying patients in terms of overall survival rate, compared to the remaining miRNAs/genes. In most cases, combined miRNA and gene stratifiers perform

better than individual ones (Fig. 2 and Fig. S5). These findings argue that the miRNAs and genes with positive correlation and the positive correlations themselves are clinically significant, rather than random events.

Although intragenic miRNAs tend to positively correlate with their host genes in expression as expected, we found a considerable number of counterexamples (Fig. S6). Our further analysis verified the previous hypothesis that conserved miRNAs (Methods) are more likely to co-express with their host genes. However, this hypothesis still could not account for all observations. We checked whether the relative location of a miRNA in its host gene plays an import role in determining their co-expression but found little association (Fig. S6C). We proposed two possible reasons for the exceptions. First, a housekeeping or likewise gene tends not to show any correlation with its embedding miRNAs, since this type of genes are critical for cell vitality and always express at high levels. Examples include DGCR8, which is an essential member of the microprocessor machinery complex and shows little correlation with its intragenic miRNAs miR-3618/1306 (Fig. S6A-B); and CDC37, an important partner of HSP90 participating the regulation of protein kinases, displays poor correlation with its intragenic miRNA miR-1181 (Fig. S6E). Second, some miRNAs do not express at all or express at very low level across all cancer tissues, and hence exemplify little if any correlation with their host genes. Examples include miR-1182~FAM89A and miR-1183~SP4 (Fig. S6F-G). These miRNAs might be just putatively derived from conserved hairpin sequences, rather than well-established and well-annotated miRNAs. Several previous studies [54,55] and the FANTOM project [56] have suggested that an intronic miRNA is not necessarily co-transcribed with its host gene by having its own independent intronic promoter, plausibly explained the mentioned discrepancy in expression. Another possibility is that there is additional layer of regulation at the level of miRNA biogenesis, due to the multiple and delicate steps required in miRNA biogenesis as described in Introduction.

A straightforward question pertaining to the miRNA-gene positive correlation is whether this correlation is direct or indirect; and a more fundamental question is whether this correlation reflects real causation or is just an association of two unrelated events in a statistics sense. We found that most of the observed positive correlations can be well explained by the indirect regulation mechanism, including double-negative (29.2%) and/or double-positive (78.6%) regulation routines (Fig. 4-5, 8). In the double-negative scenario, a miRNA inhibits its target genes by direct binding to their 3'UTR, and those target genes in turn suppress their downstream genes in a same signaling pathway. This will lead to an observation of positive correlation between the miRNA and these downstream genes. In addition, such double-negative regulation might be repeated multiple times in a pathway resulting in multiple intermediate layers between the miRNA and its positively correlated genes. It is worth noting that although we only kept the miRNA targets with high confidence (measured by probability of conserved targeting, see Methods), false positive targets still exist, and hence the eventual double-negative loops might not be biologically true. In addition, the expression level of miRNAs/genes, as well as their interaction profiles are usually tissue-specific [57–59], which potentially further increases the false positive risk. Therefore, further work should take advantage of experimentally validated miRNA targets [60] to reduce the false positive rate.

Fig. 6. Positive correlation associated miRNAs/genes tend to form super-enhancers. (A) Scaled ROSE score of H3K27ac signal surrounding sites of the top 64 miRNAs from top positive pairs covering ≥ 10 cancers (PanCan10). ROSE scores were centered and scaled across miRNAs for each ENCODE sample before clustering over miRNAs. Names of 64 miRNAs and 15 ENCODE tissues are shown on the right and at the bottom of the heat map, respectively. (B) Percentage of significantly positive (Positive, left) or randomly chosen (Random, right) miRNA-gene pairs with high ROSE score (≥ 10) on either the miRNA or gene site in each pair at different pan-levels. The Random group consists of same number of miRNA-gene pairs as the Positive group at each pan-level, with constituent miRNAs and genes randomly chosen from those not involved in the PanCan10 pairs. The number (N , $0 \leq N \leq 15$) of ENCODE tissues covered by a pair is shown in the legend. (C) A representative example showing ROSE score of H3K27ac signal for hsa-mir-150 (top) and hsa-mir-142 (bottom) in 64 ENCODE samples across 15 human tissues. The color code for each tissue type is same with (A). (D) Detailed illustration of H3K27ac signal intensity surrounding ± 50 K bp of hsa-mir-150 (left) and hsa-mir-142 (right) in ENCODE brain (top) and large intestine (bottom) tissues, respectively. Genomic coordinates for each region is shown at the bottom. The TSS of each miRNA is marked by a red solid line. Peak regions were binned to windows of 200 bp for visualization of the H3K27ac ChIP-seq signal. See also Fig. S10–11 and Table S5.

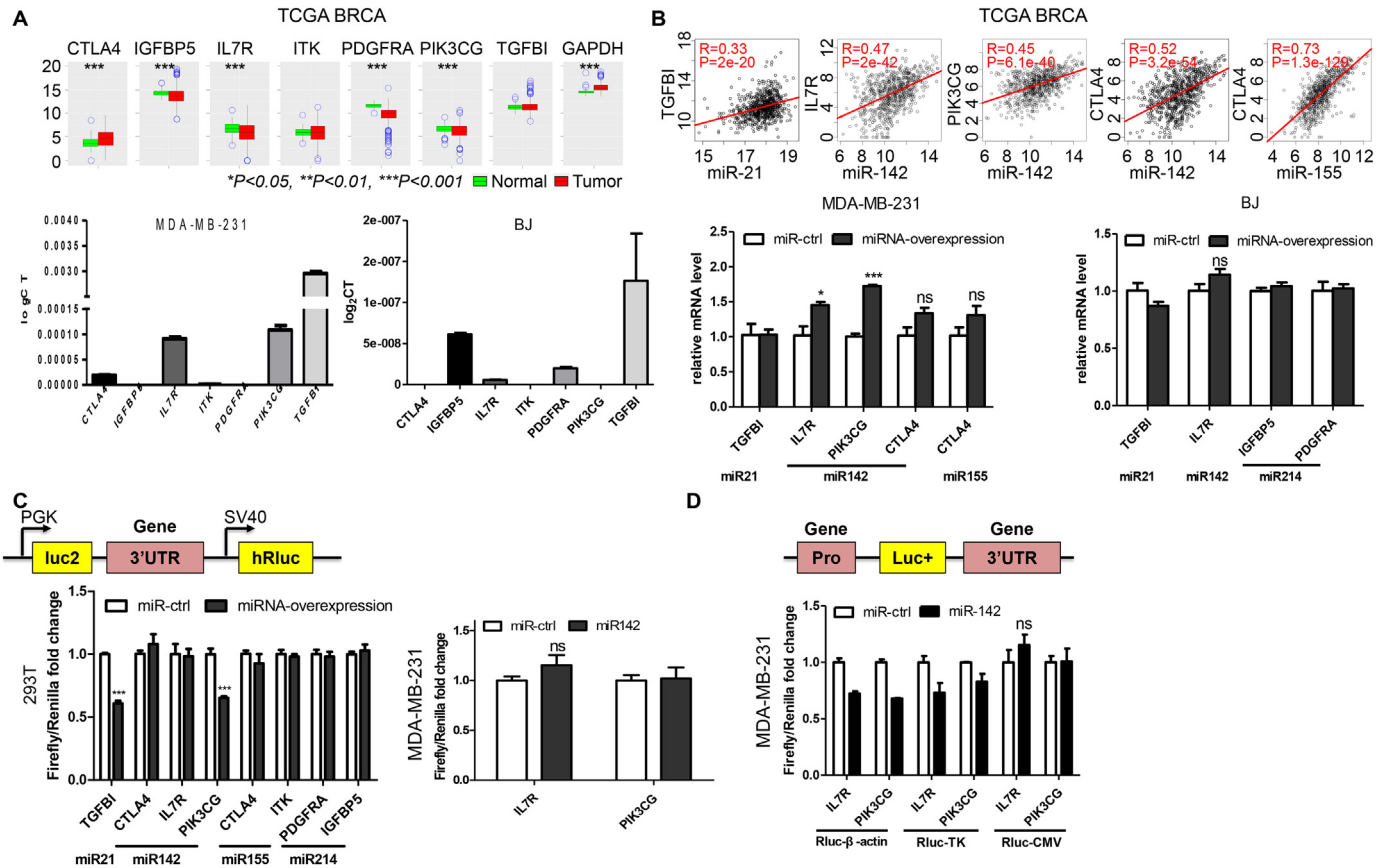


Fig. 7. The positive regulation of gene expression by miRNAs detected in patient samples is partially recapitulated in cell lines. (A) Top: Boxplot illustrating gene expression in TCGA BRCA cancer samples and associated normal controls for genes in the selected 8 miRNA-gene pairs. Housekeeping gene GAPDH is also shown for control. Comparison was performed by two-sided student *t*-test. Color and shape code for boxplot is as in Fig. 4E. Bottom: Endogenous expression level of the same genes, detected by qRT-PCR, in MDA-MB-231 and BJ cell lines as indicated. (B) Top: Pearson correlation between indicated miRNA-gene pairs in TCGA BRCA cancer samples. These pairs are selected since only these associated genes were endogenously expressed at detectable levels in the MDA-MB-231 breast cancer cell line. Bottom: qPCR results showing endogenous gene expression levels with or without stable overexpression of the constituent miRNAs in the selected positive pairs in MDA-MB-231 cells (left) and BJ cells (right). (C) Left: 3'UTR reporter assays showing the activity of the reporters with or without miRNA overexpression for the 8 selected positive pairs in 293 T cells. Right: 3'UTR reporter assays showing the activity of the 3'UTR reporters of the IL7R and PIK3CG genes with or without miR-142 in MDA-MB-231 cells. Correlations of these pairs in all TCGA cancers are illustrated in Fig. S12A. There is no miR-142 binding sites on its partner gene PDGFRA; for the other 7 pairs, there is at least one conserved miRNA binding site on its partner gene, see Fig. S12C. (D) Dual-luciferase assay results of reporters containing both a 3-kb promoter and the 3'UTR carrying the miR-142 binding sites of the IL7R or PIK3CG genes in MDA-MB-231 cells. The promoter-3'UTR Firefly luciferase reporters were co-transfected into MDA-MB-231 cells with Renilla Luciferase and corresponding miR-142 or vector control (Ctrl). Luciferase activity was determined 48 h after transfection. Firefly luciferase activity was normalized to the Renilla luciferase activity. The experiment was done for 3 times, with β -actin, TK, and CMV driven renilla as control, respectively, and result was displayed separately. In all wet-lab experiments, values are mean \pm SD for triplicates. See also Fig. S12 and S8.

While the double-inhibition mechanism is indirect, it is causative. In the double-positive scenario, on the other hand, components of a positively correlated pair are regulated by shared regulatory factors and seem not causally linked. However, this type of positive correlation has also been reported to be implicated in tumorigenesis [39]. This is sensible since not all miRNA-gene pairs sharing regulatory factors exemplify positive correlation due to many complex regulation networks, and cancer cells take advantage of this switch to control miRNA-gene co-expression. For example, a genetic variation in a regulatory region can impact the binding affinity of miRNAs or transcription factors, leading to abnormal gene abundance [61,62]. It should be noted that it is possible that a shared TF may repress both the miRNA and gene in a positive pair, or repress one and simultaneously activate the other. We reasoned that besides co-activation, the co-repression mechanism may also result in a positive correlation. Since transcriptional repression is relatively less common than transcriptional activation, we did not go into the depth in the present study. Notably, the other two potential mechanisms, miRNA-host gene co-expression, and super-enhancer mediated miRNA-gene co-expression, might also be ascribed to a mechanism similar to the dual-positive regulation. Striking examples include the FL1 regulated miR-139 and PDE2A co-upregulation (Fig. S9B), and H3K27ac enrichment associated miR-196b and HOXA10 co-activation

(Fig. S10–11). We also noticed a significant overlap between different potential mechanisms as elucidated above (Fig. 8). In particular, four pairs, including 152-COP22, 106b-MCM7, 25-MCM7 and 93-MCM7, can be explained by either of the four proposed mechanisms (Table S6). We conclude that in a more general sense, all of the four hypotheses (Fig. 8B–E) are indirect regulation that accumulatively accounts for the majority (87%) of the PanCan10 positive pairs.

The more exciting scenario for miRNA-directed gene upregulation is through direct binding and activation. Previous studies confirmed that miRNA can activate a gene by direct binding to its regulatory elements at the RNA or DNA level including 3'UTR [19], 5'UTR [40], promoter [20–22,24], or enhancer [25] regions, as abovementioned. In our study, a small portion of miRNA-gene positive correlations (13%) that could not be accounted for by any of our proposed four indirect regulation hypotheses might be explained by the direct binding and activation hypothesis, provided the gene bears binding sites for its partner miRNAs. It should be realized that this hypothesis necessitates further experimental validation. Actually, even the indirect regulation hypotheses as discussed above only provide candidate intermediate regulators to be narrowed down by elaborately designed experiments. Our *in vitro* experiments partially verified that miR-142 overexpression can activate IL7R and PIK3CG in human breast cancer cells line MDA-

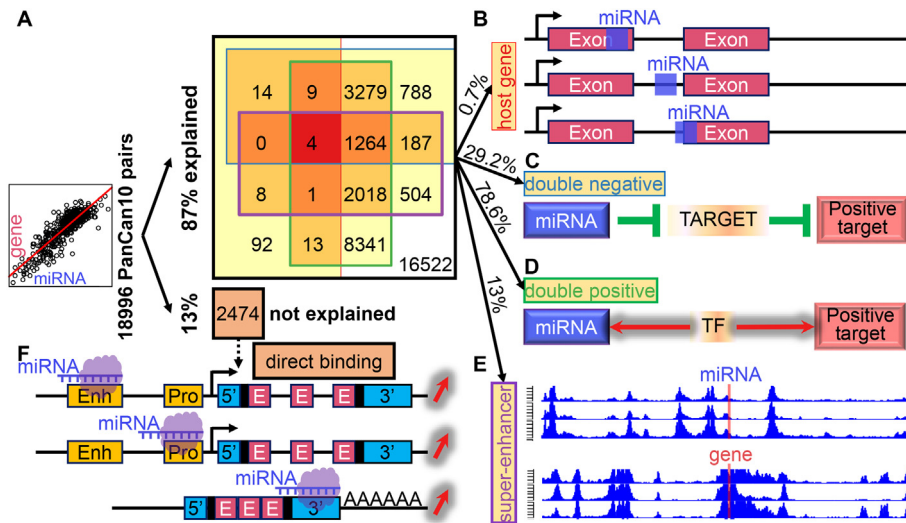


Fig. 8. A summary of putative mechanisms underlying miRNA-associated gene activation. (A) Numbers and percentages of positive pairs (PanCan10, $n = 18,996$) explained by various potential mechanisms. (B) Schematic illustration of intragenic (exonic, intronic, and mixed) miRNA and its host gene. (C) Schematic illustration of how miRNAs activate a gene by inhibiting its upstream suppressor. (D) Schematic showing miRNA and gene are positively correlated due to co-regulation by shared transcription factor (TF). (E) Schematic depicting either the miRNA or gene in a pair is enriched for super-enhancer marker H3K27ac. (F) Schematic illustration of how miRNAs activate a gene by directly binding to gene on its enhancer (Enh), promoter (Pro), or 3'UTR region. See also Table S6.

MB-231 (Fig. 7B). However, in other cases, our wet-lab experiment results showed distinct miRNA-gene interaction patterns in comparison with those observed in corresponding TCGA cancer samples. This discrepancy argues the importance of the critical impact of various *in vivo* microenvironment factors, e.g., the intercellular communications [63] and the immune responses, on the miRNA-gene interactions in cancer. And this argument is further consolidated by our enrichment analysis revealing that cell cycle related processes, cell membrane signaling and immune cell differentiation are involved in the top-ranked positive pairs (Fig. 2). Therefore, our comprehensive analysis based on large sets of patient samples is essential for the revelation of phenomenon, biological significance and potential mechanisms underlying the miRNA-associated gene activation.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ebiom.2019.03.082>.

Funding sources

This work was partially funded by NIH grants U01AR069395 (X.Z.) and R01CA172115 and R01CA131231 (P.S.). The funders had no role in study design, data collection, data analysis, interpretation, writing of the report, or decision to submit the paper for publication.

Declaration of interests

The authors declare that they have no competing interests.

Author contributions

H.T., X.Z. and P.S. conceptualized the study. H.T. performed data analysis and visualization, and wrote the manuscript. S.H. conducted the wet-lab experiments. P.S. and X.Z. revised the manuscript. Z.Z. and X.Q. provided technical assistance. All authors read and approved the final manuscript.

References

- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144:646–74.
- Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell* 2012;150:12–27.
- Tan H, Bao J, Zhou X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Sci Rep* 2015;5:12566.
- Tan H, Bao J, Zhou X. A novel missense-mutation-related feature extraction scheme for 'driver' mutation identification. *Bioinformatics* 2012;28:2948–55.
- Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. *Science* 2013;339:1546–58.
- Tan H, Zhou X. Detection of combinatorial mutational patterns in human Cancer genomes by exclusivity analysis. *Methods Mol Biol* 2018;1711:3–11.
- Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;116:281–97.
- He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* 2004;5:522–31.
- Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;136:215–33.
- Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2009;19:92–105.
- Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;120:15–20.
- Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 2008;9:102–14.
- Bartel DP. Metazoan MicroRNAs. *Cell* 2018;173:20–51.
- Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;42:D68–73.
- Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;4.
- Croce CM. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* 2009;10:704–14.
- Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, et al. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A* 2006;103:2257–61.
- Hu Y, Dingerdissen H, Gupta S, Kahsay R, Shanker V, Wan Q, et al. Identification of key differentially expressed MicroRNAs in cancer patients through pan-cancer analysis. *Comput Biol Med* 2018;103:183–97.
- Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. *Science* 2007;318:1931–4.
- Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R. MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci U S A* 2008;105:1608–13.
- Majid S, Dar AA, Saini S, Yamamura S, Hirata H, Tanaka Y, et al. MicroRNA-205-directed transcriptional activation of tumor suppressor genes in prostate cancer. *Cancer* 2010;116:5637–49.
- Huang V, Place RF, Portnoy V, Wang J, Qi Z, Jia Z, et al. Upregulation of Cyclin B1 by miRNA and its implications in cancer. *Nucleic Acids Res* 2012;40:1695–707.
- Matsui M, Chu Y, Zhang H, Gagnon KT, Shaikh S, Kuchimanchi S, et al. Promoter RNA links transcriptional regulation of inflammatory pathway genes. *Nucleic Acids Res* 2013;41:10086–109.
- Zhang Y, Fan M, Zhang X, Huang F, Wu K, Zhang J, et al. Cellular microRNAs up-regulate transcription via interaction with promoter TATA-box motifs. *RNA* 2014;20:1878–89.
- Xiao M, Li J, Li W, Wang Y, Wu F, Xi Y, et al. MicroRNAs activate gene transcription epigenetically as an enhancer trigger. *RNA Biol* 2017;14:1326–34.

- [26] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;483:603–7.
- [27] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–61.
- [28] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16:284–7.
- [29] Suzuki HI, Young RA, Sharp PA. Super-enhancer-mediated RNA processing revealed by integrative MicroRNA network analysis. *Cell* 2017;168:1000–14 [e1015].
- [30] Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;1:417–25.
- [31] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res* 2012;22:1760–74.
- [32] Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 2007;35:D137–40.
- [33] Zheng G, Tu K, Yang Q, Xiong Y, Wei C, Xie L, et al. ITPP: an integrated platform of mammalian transcription factors. *Bioinformatics* 2008;24:2416–7.
- [34] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
- [35] Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyanopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 2012;150:1274–86.
- [36] Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, et al. TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep* 2015;5:11432.
- [37] Marbach D, Lamparter D, Quon G, Kellis M, Kutralik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods* 2016;13:366–70.
- [38] Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* 2013;153:320–34.
- [39] Lim JY, Yoon SO, Seol SY, Hong SW, Kim JW, Choi SH, et al. Overexpression of miR-196b and HOXA10 characterize a poor-prognosis gastric cancer subtype. *World J Gastroenterol* 2013;19:7078–88.
- [40] Liu M, Roth A, Yu M, Morris R, Bersani F, Rivera MN, et al. The IGF2 intronic miR-483 selectively enhances transcription from IGF2 fetal promoters and enhances tumorigenesis. *Genes Dev* 2013;27:2543–8.
- [41] Veronese A, Lupini L, Consiglio J, Visone R, Ferracin M, Fornari F, et al. Oncogenic role of miR-483-3p at the IGF2/483 locus. *Cancer Res* 2010;70:3140–9.
- [42] Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, et al. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nat Cell Biol* 2012;14:659–65.
- [43] He Y, Jiang X, Chen J. The role of miR-150 in normal and malignant hematopoiesis. *Oncogene* 2014;33:3887–93.
- [44] Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, et al. MicroRNA target prediction by expression analysis of host genes. *Genome Res* 2009;19:481–90.
- [45] He C, Li Z, Chen P, Huang H, Hurst LD, Chen J. Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene co-evolution. *Nucleic Acids Res* 2012;40:4002–12.
- [46] Franca GS, Vibanovski MD, Galante PA. Host gene constraints and genomic context impact the expression and evolution of human microRNAs. *Nat Commun* 2016;7:11438.
- [47] Cursons J, Pillman KA, Scheer KG, Gregory PA, Foroutan M, Hediyyeh-Zadeh S, et al. Combinatorial targeting by MicroRNAs co-ordinates post-transcriptional control of EMT. *Cell Syst* 2018;7:77–91 [e77].
- [48] Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* 2008;10:593–601.
- [49] Ohno T, Rao VN, Reddy ES. EWS/Fli-1 chimeric protein is a transcriptional activator. *Cancer Res* 1993;53:5859–63.
- [50] Rao VN, Ohno T, Prasad DD, Bhattacharya G, Reddy ES. Analysis of the DNA-binding and transcriptional activation functions of human Fli-1 protein. *Oncogene* 1993;8:2167–73.
- [51] Wang Z, Wan L, Zhong J, Inuzuka H, Liu P, Sarkar FH, et al. Cdc20: a potential novel therapeutic target for cancer treatment. *Curr Pharm Des* 2013;19:3210–4.
- [52] Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;153:307–19.
- [53] Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. *Cell* 2013;155:934–47.
- [54] Monteyts AM, Spengler RM, Wan J, Tecedor L, Lennox KA, Xing Y, et al. Structure and activity of putative intronic miRNA promoters. *RNA* 2010;16:495–505.
- [55] Oszolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, et al. Chromatin structure analyses identify miRNA promoters. *Genes Dev* 2008;22:3172–83.
- [56] de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat Biotechnol* 2017;35:872–8.
- [57] Chiu HS, Somvanshi S, Patel E, Chen TW, Singh VP, Zorman B, et al. Cancer genome atlas research N, et al: Pan-Cancer analysis of lincRNA regulation supports their targeting of Cancer genes in each tumor context. *Cell Rep* 2018;23:297–312 [e212].
- [58] Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C. Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* 2013;20:1325–32.
- [59] Dhawan A, Scott JG, Harris AL, Buffa FM. Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. *Nat Commun* 2018;9:5228.
- [60] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013;153:654–65.
- [61] Tan H. The association between gene SNPs and cancer predisposition: correlation or causality? *EBioMedicine* 2017;16:8–9.
- [62] Tan H. On the protective effects of gene SNPs against human Cancer. *EBioMedicine* 2018;33:4–5.
- [63] Tan H, Li F, Singh J, Xia X, Criderbring D, Yang J, et al. A 3-Dimensional multiscale model to simulate tumor progression in response to interactions between cancer stem cells and tumor microenvironmental factors. *IEEE 6th International Conference on Systems Biology (ISB)*; 2012. p. 297–303.