Check for updates

RESEARCH ARTICLE

Regulatory landscapes of specific miRNAs are conserved between cell lines and primary tumors [version 1; peer review: 1 approved, 1 approved with reservations]

Hanwen Zhu¹, Boting Ning^{id}²¹YK Pao School, Songjiang District, Shanghai, 201620, China²Boston University School of Medicine, BOSTON, Massachusetts, 02118, USA

V1 First published: 22 Jul 2021, 10:633
<https://doi.org/10.12688/f1000research.52478.1>

Latest published: 22 Jul 2021, 10:633
<https://doi.org/10.12688/f1000research.52478.1>

Abstract

Background: MicroRNAs are essential gene expression regulators and play important roles in various biological processes, such as cancer. They have shown great translational promise as either diagnostic biomarkers or therapeutic targets. While the similarities between transcriptomic profiles from The Cancer Genome Atlas and the Cancer Cell Line Encyclopedia have been thoroughly studied before, less is known on the microRNA side. This project aims to provide critical biological knowledge on the extent of consensus microRNA expression and regulation between cell line models and primary human tumors.

Method: First, we examined the similarity of miRNA expression profiles between CCLE cell lines and TCGA tumor samples for each cancer type. Next, we compared the expression of miRNAs associating the hallmarks of cancer pathways. Finally, we constructed miRNA-mRNA regulatory network for each cancer type and evaluated whether the regulatory role of each miRNA is conserved between cell lines and tumor samples.

Results: Our results indicate that, similar to gene expression, how well cancer cell line microRNA expression would capture the transcriptomic profile of human cancer tissues is greatly affected by the tumor type and purity. The cell-type composition for a cancer type also affects how accurately cancer cell lines could reflect the miRNA expression in tumor tissues. Furthermore, through network analysis, we show that certain microRNAs, not all, regulate the same set of target genes in both the cell line and human cancer tissues.

Conclusions: Through systematically comparing the miRNA expression profile and the regulatory network, our study highlights the biological differences between cell line and tumor samples and provides

Open Peer Review

Approval Status

	1	2
version 1
22 Jul 2021	view	view

1. **Chongzhi Zang** , University of Virginia, Charlottesville, USA

2. **Vinayak C. Palve** , H. Lee Moffitt Cancer Center, Tampa, USA

Any reports and responses or comments on the article can be found at the end of the article.

resources for future miRNA and cancer studies.

Keywords

microrna, Cancer, miRNA-mRNA Network, CCLE, TCGA

This article is included in the **Oncology** gateway.

Corresponding author: Boting Ning (ningb@bu.edu)

Author roles: **Zhu H:** Data Curation, Formal Analysis, Visualization, Writing – Original Draft Preparation; **Ning B:** Conceptualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Zhu H and Ning B. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Zhu H and Ning B. **Regulatory landscapes of specific miRNAs are conserved between cell lines and primary tumors [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2021, 10:633 <https://doi.org/10.12688/f1000research.52478.1>

First published: 22 Jul 2021, 10:633 <https://doi.org/10.12688/f1000research.52478.1>

Introduction

MicroRNAs (miRNAs) are endogenous, small non-coding RNAs that function in the post-transcriptional regulation of gene expression.¹ Using the seed region, the mature miRNAs bind to the 3' untranslated region (UTR) of the mRNAs via complementary base-pairing and suppress the expression of the targeted genes.^{2,3} Recently, miRNA has been found to be heavily dysregulated in cancer cells by functioning as either tumor suppressor or oncogenes.^{4,5} Many investigations suggest that abnormal miRNA expression levels could be used for diagnostic and prognostic biomarkers for lung, prostate, or breast cancers.^{6–8} Also, the usage of miRNAs as potential therapeutic targets has been shown to be promising in many cancer types.^{9–11}

Recent progress in high-throughput sequencing technology has offered researchers unprecedented opportunities to understand the molecular mechanism of cancer. The [Cancer Genome Atlas \(TCGA\)](#) collects molecular profiling of primary tumors of over 11,000 tumor samples and provides real-world cancer patient information.¹² On the other hand, the [Cancer Cell Line Encyclopedia \(CCLE\)](#) characterizes around 1,000 human cancer cell lines for in vivo studies.^{12,13} Both datasets span across several tumor types and contain multi-omics sequencing data. A comparison between them can provide knowledge about the differences between the two sets and capture how well each cell line represents a specific type of primary tumor. K Yu *et al.* have previously conducted a pan-cancer comparison between TCGA and CCLE.¹⁴ Specifically, they used correlation analysis and gene set enrichment analysis (GSEA) to identify differences between cell lines and primary tumors and found that there is a strong correlation between TCGA and CCLE samples for each tumor type and that tumor purity is a main driving factor of cell line and primary tumor differences. They also identified cell lines representative of primary tumor samples in TCGA. While this study offered a comprehensive comparison of the gene expression profiles between TCGA and CCLE, whether miRNA expression differs between the two data sources is still yet unclear. Here, we seek to use similar correlation analysis to evaluate the concordance of miRNA expression profiles between TCGA and CCLE datasets.

We first obtained respective datasets and mapped miRNAs from TCGA to CCLE to provide consistency between cell lines and primary tumors. For both datasets, technical confounders were adjusted, as well as tumor purities, and the distribution of cell lines or tumor samples were visualized using the distribution in each dataset using t-distributed stochastic neighbor embedding (t-SNE). Then, we performed correlation analysis to evaluate the similarity between cancer cell line models and human primary cancer samples based on miRNA expression profiles. We found that the cell-type composition of cancer samples and the nature of cancer greatly affect the correlation strengths. Also, through studying the expression of miRNAs related to the hallmarks of cancer pathways, we found immune-related pathways were among the most differentially expressed pathways between CCLE and TCGA. Lastly, we investigated whether the miRNA–mRNA regulatory networks in the human cancer samples could be accurately captured by the cell line models.

Methods

Data collection and processing

CCLE miRNA count matrix and the cell line annotations were downloaded from the [CCLE Broad Institute data portal](#).¹⁵ The current version of CCLE data contains 954 cell lines across 31 tumor types based on TCGA cancer code, among which 258 were labeled as “NA” or “Unable to classify”. The miRNA expression data, as well as the clinical information of each TCGA sample, were downloaded from the [GDC data portal](#) using R package TCGAAbiolinks, containing 11,082 samples across 32 tumor types.¹⁶ Due to the different sequencing technology, much more unique miRNA could be identified in TCGA comparing to CCLE. Thus, we first unified the miRNAs identified in two datasets by summing all raw miRNA counts in TCGA that correspond to a specific CCLE miRNA ID within each sample. Identical mature miRNA reads from different precursors at different genomic locations (*eg*, hsa-miR-1-1 and hsa-miR-1-2) were also summed in TCGA to match with CCLE (*eg*, hsa-miR-1), if no exact matching was found in CCLE. Also, for miRNA that is characterized as precursor in CCLE (lacking the -3p or -5p information) but as mature miRNA in TCGA, the corresponding mature miRNA reads from different precursors were summed in TCGA to match CCLE. Next, the counts were normalized using the trimmed mean of M values (TMM) and had log2 counts per million computed using EdgeR.¹⁷ MiRNA with an interquartile range equal to zero or a sum across samples equal or less than one was excluded.

Then, since TCGA was sequenced on two different sequencing platforms, namely, 9,411 samples using HiSeq and 1,413 using GA (with 258 unannotated samples dropped), batch correction was done to eliminate differences caused by the difference of platforms using ComBat.¹⁸ Finally, as tumor purity has been proven to be a significant confounder in transcriptomic data of primary tumors by K Yu *et al.*, TCGA miRNA expression profiles were additionally adjusted for their tumor purity scores using limma.¹⁹ Tumor sample purity scores calculated using the ABSOLUTE²⁰ method were obtained from the [TCGA PanCanAtlas website](#). 1,126 batch-corrected samples without tumor purity measurements were discarded, leaving 9,698 samples.

Correlation analysis

TCGA and CCLE expression profiles consisted of the 327 miRNAs that passed the quality control. A Spearman correlation coefficient was calculated between each TCGA sample and CCLE cell line with the 327 miRNAs that passed the quality control in both datasets, resulting in a correlation matrix of 954 cell lines by 9,698 human tumor samples. The correlation coefficients were then Fisher-Z transformed and arranged with a hierarchical clustering by the cancer types of CCLE. Further, the coefficients obtained from correlation comparing CCLE to TCGA expression adjusted for tumor purity and not adjusted for tumor purity were compared using two-group t-test.

In the meantime, the correlation matrices between CCLE and TCGA based on gene expression were obtained from the [github repository](#) from K Yu *et al.*¹⁴ Then, the correlation coefficients between each cell line and TCGA of the same tumor type based on miRNA expression profiles were compared to that from mRNA expression profiles using Spearman correlation.

Hallmarks for cancer pathway analysis

The miRNA set associated with either upregulation or downregulation of these hallmarks for cancer pathways were obtained from A Dhawan, *et al.*⁴ For each pathway, single-sample gene set enrichment analysis (ssGSEA) was used to summarize the miRNA expression in CCLE, and both adjusted and unadjusted TCGA. Afterward, the GSVA scores were compared between CCLE and TCGA, and between tumor purity adjusted and unadjusted TCGA samples using a two-group t-test.

Network analysis

There are very few cell lines included in CCLE for several certain types of cancer. Also, it is less appealing to study the miRNA–mRNA network when the expression level correlation is low between CCLE and TCGA. Thus, to increase the validity of network analysis based on miRNA–mRNA correlation, we first filtered the cancer type and cell lines based on sample size and miRNA expression level correlation strength. Among the 22 overlapping cancer types between CCLE and TCGA, seven tumor types with 30 or more cell lines strongly correlated ($r \geq 0.45$) with TCGA (LUAD, COAD/READ, BRCA, SKCM, STAD, HNSC, SARC) were chosen for miRNA–mRNA network analysis.

For each cancer type, the Pearson correlations between each miRNA and mRNA expression levels were calculated for CCLE cell lines or TCGA primary tumor samples. Then, each negative and statistically significant ($FDR < 0.05$) correlation was retained to form an edge in the constructed miRNA–mRNA correlation network. For each cancer type, the miRNA–mRNA network was constructed within CCLE and for TCGA samples separately. The intersection between the resulted correlation network and the TargetScan²¹ predicted target network is taken to construct a regulatory network from miRNA to mRNA.

For each cancer type, the degree distributions of miRNAs of both CCLE and TCGA were calculated. Hub scores of all miRNAs and the correlation between the hub scores from CCLE and TCGA were obtained. Further, the hamming distance between networks of CCLE and TCGA was calculated for each cancer type.

To prioritize miRNA based on the target gene conservation, we compare the overlaps between the connected gene for each miRNA in the CCLE and the TCGA network for each cancer type. Using a one-tailed Fisher's exact test, the significance of the overlap between targets of CCLE and TCGA as FDR was calculated.

Code availability

All code for data processing and analysis is available at <https://github.com/hanwenzhu/mir-tcga-ccle-paper>.²²

Results

MicroRNA expression profiles in CCLE and TCGA

The TCGA miRNA data were sequenced via small-RNA sequencing, while Nanostring platform was used for CCLE. To avoid the confounding caused by the two drastically different platforms, we processed and normalized the two datasets separately and only compared the relative miRNA expression rather than the absolute expression level. MiRNA counts from 954 cell lines across 31 tumor types from CCLE and from 11,082 samples across 32 tumor types from TCGA remained after processing (Figure 1 and Extended Table 1). Note that there were significantly more samples in TCGA than in CCLE. A significant portion of the CCLE lacked specific annotation of tumor type in the format of TCGA disease code ("NA" or "Unable to classify"), which were excluded for the correlation analysis later on. Overall, the number of samples of each tumor type varied by a lot in TCGA, with multiple types from either dataset having less than 10 samples, such as GBM, which had only five samples. Also, the number of BRCA tumors was the highest and was much greater than any other type in TCGA.

Figure 1. The number of samples and cell lines from TCGA and CCLE used in the analysis. The COAD and READ were merged in TCGA to match the classification used in CCLE. (a) The number of samples by tumor type in the TCGA data. (b) The number of cell lines by TCGA tumor type in the CCLE data. BRCA, breast invasive carcinoma; COAD/READ, colorectal adenocarcinoma; KIRC, kidney renal clear cell carcinoma; THCA, thyroid carcinoma; HNSC, head and neck squamous cell carcinoma; UCEC, uterine corpus endometrial carcinoma; LUAD, lung adenocarcinoma; PRAD, prostate adenocarcinoma; LGG, brain lower grade glioma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; STAD, stomach adenocarcinoma; SKCM, skin cutaneous melanoma; BLCA, bladder urothelial carcinoma; LIHC, liver hepatocellular carcinoma; KIRP, kidney renal papillary cell carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; SARC, sarcoma; ESCA, esophageal carcinoma; LAML, acute myeloid leukemia; PCPG, pheochromocytoma and paraganglioma; PAAD, pancreatic adenocarcinoma; TGCT, testicular germ cell tumors; THYM, thymoma; KICH, kidney chromophobe; MESO, mesothelioma; UVM, uveal melanoma; ACC, adrenocortical carcinoma; UCS, uterine carcinosarcoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; CHOL, cholangiocarcinoma; GBM, glioblastoma multiforme; ALL, acute lymphocytic leukemia; NB, neuroblastoma; MM, multiple myeloma; LCML, chronic myelogenous leukemia; MB, medulloblastoma; CLL, chronic lymphocytic leukemia.

While the cell lines of each tumor type in the CCLE did not form perfectly distinct clusters by themselves, we did observe two large separate clusters (Figure 2). The lower cluster consisted of cell lines derived from lymphocyte or myeloid cells, such as those derived from ALL, DLBC, and LAML. The upper cluster contained cell lines derived from solid tumor tissues. This suggests that a large portion of variance within the CCLE miRNA profiles is reflecting the tissue origins of

Figure 2. t-SNE plot of CCLE cell lines. The t-SNE coordinates were calculated based on processed miRNA read counts and the points were colored by tumor types. The “UNABLE TO CLASSIFY” type was merged with NA in the figure above, shown in grey.

the cell lines. On the other hand, the imperfect distinction between clusters also indicates CCLE cell lines of the same tumor type could be very different from each other as a result of the immortalization or lab culturing process.

In the meantime, we did observe that the TCGA tumor samples clustered much better than the CCLE cell lines (Figure 3). The clusters of tumor types were more distinct after both batch correction and adjustment for tumor purity. Such observation is similar to the results based on gene expression from a previous study.¹⁴ The lack of clear cancer type clustering in the CCLE t-SNE might suggest that within each tumor type, the CCLE cell lines are more heterogeneous than the TCGA samples, and not all cell lines are good models for miRNA and cancer studies.

Several tumor types with similar pathological features were mixed with each other when being clustered using normalized miRNA counts or batch corrected counts, such as BRCA, HNSC and LUSC (Figure 3a,b). These samples could be distinguished after we adjusted for the tumor purity, as reported by other studies,¹⁴ suggesting that immune infiltration could affect not only the gene but also the miRNA profiles of related tumor types in a similar way (Figure 3c).

Correlation between CCLE and TCGA microRNA expression profiles

Next, we investigated how well CCLE cell lines capture the miRNA expression profiles of TCGA human primary cancer samples. Generally, cell lines show a moderate to high correlation with their matched tumor samples across different cancer types (Figure 4a; see Extended Fig. 1-2 for correlation per each cell line). The median correlation is between 0.5–0.6. However, the correlation coefficient distribution within each tumor type was much larger than that calculated based on gene expression, such as for COAD/READ (0.23–0.83) and SARC (0.19–0.80). This indicates that the miRNA expression might be more prone to the heterogeneity within each tumor type.

Based on the correlation between CCLE and TCGA samples, samples of the same tumor type from cell lines and primary tumors are more correlated (Figure 4b). Also, similar to the observation from K Yu *et al.*¹⁴ cancer types with similar pathology cluster together. The two cancer types on the top of the heatmap are DLBC and LAML, both originating from immune cell types. Such separation indicates that miRNA expression could reflect the difference between hematopoietic and solid tumors. Other cancer types that are biologically related also tend to cluster together, such as LUSC and LUAD, or GBM and LGG.

Given that it is been previously reported that the cellular composition of a tumor sample has strong effects on the correlation analysis, we also compared the correlation calculated using tumor purity adjusted and the unadjusted TCGA

Figure 3. t-SNE plots of TCGA primary tumor samples. The figures above show the t-SNE of TCGA expression data, after normalization (a), batch correction (b), and purity adjustment (c), respectively. The samples formed highly distinct clusters by different tumor types after batch correction and purity adjustment.

miRNA expression profiles (Figure 4a). Most of the cancer types have a significant increase in the correlation coefficient distribution after tumor purity adjustment, such as BRCA, HNSC, LUSC, suggesting purity being a driving factor of this increase, which reflects findings based on mRNA from K Yu *et al.*¹⁴ These cancers, as well as the cell lines, are originated from the epithelial cells, and the amount of immune cell infiltration is expected to have a strong influence on their expression obtained from bulk RNA-sequencing. In the meantime, other hematopoietic cancer types such as LAML and DLBC, have small changes in the correlation coefficient distribution, suggesting their cellular composition is not influenced by immune infiltration.

Finally, we compared miRNA and mRNA in terms of the similarity between CCLE and TCGA expression profiles (Figure 4c). Among the tumor types, the ones with the highest median correlations are PAAD, ESCA, and KIRC, and the lowest ones are HNSC, COAD/READ, and STAD, although all median correlations are positive. Interestingly, the cancer types with a higher correlation between CCLE and TCGA based on either miRNA or mRNA expression profiles are not necessarily the ones with higher concordance between miRNA and mRNA based correlation. For example, HNSC is highly ranked by mRNA and miRNA expression (median = 0.63 and 0.57) but ranked the last based on the correlation between miRNA and mRNA correlation coefficients, while LIHC, although showing lower correlation based on expression (0.50 and 0.49), has strong concordance. Such results show that the expression similarity between cell lines and human primary tumor samples does not directly reflect the concordance in the miRNA–mRNA regulatory relationship between sample types.

Functional characterization of the difference between CCLE and TCGA miRNA expression

To further elucidate the biological difference between CCLE and TCGA miRNA expression profiles, we investigated the expression alteration in the miRNAs that are associated with the hallmarks of cancer pathways. Comparing the GSEA score of miRNAs associated with hallmarks of cancer pathways between CCLE and TCGA (Figure 5a and Extended Figure 3), we saw higher enrichment of pathways related to immune cell infiltration in the TCGA (up-regulation of up hallmark miRNA set and vice versa), such as inflammatory response and IL6-JAK-STAT3 signaling pathways,²³

Figure 4. (a) Two-sided violin plot of the Fisher-transformed Spearman correlation coefficient between CCLE and TCGA before (left) and after (right) adjustment. The bars indicate the median value. Tumor types with $FDR > 0.05$ are labeled with “ns”, $FDR \leq 0.05$ with one star, $FDR \leq 0.01$ with two stars, and $FDR \leq 0.001$ with three stars, and $FDR \leq 0.0001$ with four stars. GBM is not shown since it lacks tumor purity estimates. (b) Heatmap of median Spearman correlation coefficients between each cancer type of TCGA and CCLE. CCLE cancer types are clustered based on their correlation with each TCGA cancer type. (c) Violin plot of the Spearman correlation between miRNA and mRNA data of the correlation coefficients between TCGA and CCLE in each cell line.

suggesting the absence of immune infiltration in CCLE pure cancerous cell compositions might be the major difference between two sample types. In the meantime, CCLE is more enriched in hallmarks related to tumor cells, such as more G2M checkpoint, p53 pathway, and epithelial-mesenchymal transformation pathways, revealing a higher portion of cancerous cells composing the cell lines. These results are consistent with GSVA scores of mRNA from K Yu *et al.*¹⁴ suggesting miRNA is similar to mRNA in terms of the difference in composition and hallmark enrichment between TCGA and CCLE.

Although both purity-adjusted and unadjusted TCGA compare similarly to CCLE, there is a consistent difference across cancer types due to purity adjustment (Figure 5b). Inflammatory response and TNF- α signaling are significantly more enriched in TCGA before purity adjustment, and other hallmarks from cancer cells such as DNA repair and oxidative phosphorylation are more enriched in TCGA after adjustment. This demonstrates that the purity adjustment of TCGA effectively reduces the impurity of the gene expression data.

miRNA–mRNA regulatory network comparison

To further characterize the difference between cell line models and human primary tumor tissues in terms of transcriptomic profiles, we constructed the miRNA–mRNA network in CCLE and TCGA for each cancer type separately and compared the network metrics. The degree distribution shows the distribution of the number of targets of all 230 miRNAs

Figure 5. GSVA score comparison between the cancer hallmark associated miRNA expressions in TCGA and CCLE. The downregulated hallmarks are compared between CCLE and tumor purity-adjusted TCGA (a) and between TCGA before and after purity adjustment (b) using a t-test and visualized as heatmaps. It is possible to observe that immune infiltrate hallmarks are more present in TCGA, especially before adjustment.

for each cancer type, reflecting the connectivity of the target network. The results reveal differences in the global connectivity between CCLE and TCGA regulatory networks. CCLE miRNAs have, on average, fewer obvious targets than TCGA, with miRNA out-degrees concentrated at lower values (Figure 6a). Both CCLE and TCGA contain some outlying miRNA hubs with significantly more targets than average. However, since TCGA contains more samples than CCLE (1,030 vs 50 for BRCA in Figure 6b; Extended Fig. 4), there are expected to be more significant correlated miRNA–mRNA pairs in the TCGA data so the degree distribution comparison may be inconclusive.

The hamming distance offers a metric of the difference between target networks that also takes the edges to mRNA targets into account rather than merely the out-degree of each miRNA. The hamming distances between CCLE and TCGA miRNA–mRNA networks in Figure 6c measure the difference of miRNA regulation between CCLE and TCGA of each cancer type, and a small hamming distance indicates that the two networks are more similar to each other. Cancer types with more similar networks are COAD/READ (distance = 18,114) and SKCM (distance = 20,872), while more different ones are BRCA (distance = 34,677) and HNSC (distance = 28,047). The hamming distance, however, failed to adjust to the greater total number of edges in TCGA target networks than in CCLE, so an analysis focused on miRNA hub scores and without the influence of network size was conducted.

While network metrics as the degree distribution and hamming distance provide a global view on the network similarity between CCLE and TCGA, it is unclear whether individual miRNA targets the same set of genes in both cell lines and human tumor samples. To further explore each miRNA, the Pearson correlations of 230 miRNA hub scores between CCLE and TCGA were calculated (Figure 6d). They reflect whether the network topology, especially the central regulatory miRNA, is conserved between CCLE and TCGA. All coefficients are positive, meaning the regulatory roles between CCLE and TCGA are largely similar, although the correlation strength is very varied in different cancer types. Cancer types with more pronounced similarity of miRNA centrality are STAD ($r = 0.57$) and HNSC ($r = 0.53$), while less similar ones are SKCM ($r = 0.01$) and SARC ($r = 0.12$).

To prioritize hub miRNAs based on the conservation of regulatory potential, networks of hub miRNAs with the most significant overlap of their targets between CCLE and TCGA were obtained (Extended Fig. 5). Taking BRCA as an example, the miRNAs of BRCA with the most significant overlap of mRNA targets between CCLE and TCGA are hsa-miR-200c, hsa-miR-141, hsa-miR-29c, hsa-miR-200b, and hsa-miR-221 (FDR < 0.0001; Figure 6e). They are miRNA

Figure 6. (a) Median and mean number of mRNA targets of CCLE and TCGA miRNA. (b) The miRNA out-degree distribution of BRCA target networks. (c) Hamming distance between CCLE and TCGA target networks for each cancer type. (d) Pearson correlation coefficient of hub scores of all miRNAs between CCLE and TCGA for each cancer type. (e) Common targets of the five miRNA hubs with the most significant overlap of targets between CCLE and TCGA networks of BRCA. Each triangle represents a miRNA hub and each circle node represents a target. The FDR values of the one-tailed Fisher exact test of the overlap are labeled below the miRNA names.

hubs with conserved connectivity between CCLE and TCGA and are ideal regulatory hubs to be studied using cell line models. In fact, previous research involving cell line models has shown that hsa-miR-200 promotes the proliferation of breast luminal progenitor cells and facilitates the growth and metastasis of breast cancer, while hsa-miR-221 regulates breast cancer development and progression and serves as a promising biomarker.^{24,25} These findings highlight the value of our results for identifying miRNAs with conserved regulatory roles in both cell lines and human tumor samples and could guide cell-line-based research to model how miRNA regulates mRNA in primary tumors.

Conclusions

While previous research has characterized how CCLE cell lines reflect the gene expression profiles of human primary tumors in TCGA, our work extends the analysis to miRNA expression profiles. We show that cell lines of the related cancer type cluster closer based on miRNA expression profiles, similar to the mRNA based analysis. In the meantime, there seems to be a larger variance in the correlation distribution within a cancer type, indicating miRNA affecting by the heterogeneity of the samples. There are two potential technical reasons for such observation: 1) different platforms were used to profile miRNA between CCLE and TCGA; 2) there are smaller number of unique miRNA species comparing the mRNA and could be more affected by the noise. Yet, the larger variance between cell lines and tumor samples based on miRNA profiles indicates that choosing the cell line most similar to the cancer type of interests is even more important when studying miRNAs.

We have also shown that tumor purity affects the correlation between CCLE and TCGA miRNA expressions. With the adjustment of tumor purity in the TCGA miRNA expression profiles, we observe an increase in the correlation between miRNA expression profiles of cell lines and human primary tumor samples. Also, through functional characterization of miRNA related to hallmarks of cancer pathways, we find that TCGA samples have higher expression of immune pathways and lower expression of cell-cycle and cell growth pathways. The discrepancy is particularly stronger for solid tumor comparing to blood cancer, such as AML or DLBC. The observation could be the result of immortalization or simply cells being removed from the original microenvironment. The loss of cancer cell communication with surrounding

structural cells and particularly the lack of immune infiltration within cell lines could greatly affect the transcriptomic profiling. We believe extra cautions should be made when cell lines are being used to tackle question involving cell-to-cell interaction.

More importantly, we compared the miRNA–mRNA regulatory networks between cell lines and human tumor samples and observed a very different trend than comparing the expression profiles alone. This result indicates that the similarity in miRNA or mRNA expression profiles between cell line model and human tumor tissues does not guarantee a conserved miRNA regulatory pattern. To comprehensively examine the miRNA–mRNA network, we further evaluated whether each miRNA regulates the same set of genes in CCLE and TCGA samples. The top miRNAs from our analysis function as hub regulators of gene expression and have been previously shown to be biologically essential to cancer development. Such miRNAs with strong conservation in regulatory functions from tumor to cell lines are ideal candidates for *in vitro* analysis.

There are several limitations to our analysis. Firstly, the CCLE and TCGA miRNA expressions were measured using very different techniques. We tried to overcome such discrepancy by applying a stringent filter during data preprocessing and avoided direct expression comparison. Given the consistent observations between our results and previous work, we believed our analysis is valid. Second, we used CCLE cell lines of the same cancer type to calculate the correlation network even though there might be underlying differences. Given that there is only one replicate for each cell line in CCLE, there was no easy way to circumvent this issue. We filtered for cell lines with a high correlation to TCGA samples and cancer types with enough number of cell lines to ensure the reliability of the analysis. Finally, we understand miRNA–mRNA correlation based on normalized count data should not be interpreted as causal relationship. In a more rigorous setting, miRNA regulatory roles should be examined based on perturbation assays, in which we would observe whether target gene expression levels are altered following over-expression or knock-out of a miRNA. However, such task is beyond the scope of this project. Here, we only aim to objectively explore the differences between cell lines and human tumors samples from two public databases.

In summary, we conducted systematic comparison between CCLE and TCGA in terms of miRNA expression profiles and the regulatory landscapes. Our results highlight the importance of choosing appropriate cell lines to study miRNA in cancer research. The cellular composition heterogeneity, particularly for the solid tissue tumors, greatly affect whether the cell lines can accurately capture the miRNA expression profiles of the tumor. Certain miRNAs, not all, have preserved target gene regulatory roles in the cell lines and may be more suitable for *in vitro* investigation. We believe our results can provide valuable resources for selecting cell lines to study how a particular miRNA regulates cancer-related gene expression.

Acknowledgements

We would like to graciously acknowledge the Pioneers China Program (PCP) program which supported this research.

Additional information

Author endorsement

Prof. Marc Lenburg confirms that the authors have an appropriate level of expertise to conduct this research and confirms that the submission is of an acceptable scientific standard. Prof. Lenburg declares they have no competing interests. Affiliation: Boston University School of Medicine.

Data availability

Underlying data

CCLE miRNA counts were downloaded from Broad Institute data portal: <https://portals.broadinstitute.org/ccle/data>. TCGA miRNA counts were downloaded from the GDC data portal <https://portal.gdc.cancer.gov/> using the R package TCGAabioblinks.¹⁶ Correlation matrices between transcriptomic data of TCGA and CCLE from the work by K. Yu were released on GitHub: https://github.com/katharineyu/TCGA_CCLE_test. Tumor purity estimates of TCGA using the ABSOLUTE method are available on the TCGA PanCanAtlas publications website: <https://gdc.cancer.gov/about-data/publications/pancanatlas>. The hallmark miRNA gene sets from A. Dhawan were available on GitHub: https://github.com/andrewdhawan/miRNA_hallmarks_of_cancer/. TargetScan miRNA target predictions were downloaded from the TargetScan website: http://www.targetscan.org/vert_72/.

Extended data

Github: Regulatory landscapes of specific miRNAs are conserved between cell lines and primary tumors, <https://github.com/hanwenzhu/mir-tcga-ccle-paper>.²²

This project contains the following extended data:

- Extended Table 1.pdf (TCGA and CCLE tumor type abbreviations).
- Extended Figure 1.pdf

Distribution of Spearman correlation coefficients between CCLE and miRNA expression profiles for each cancer type.

For each cell line of CCLE, Spearman correlation was calculated with TCGA tumor samples of the same tumor type based on miRNA expression (see Methods). The distribution of correlation coefficients for each cell line is plotted as a boxplot, where the center line shows the median and the upper and lower hinges show the first and the third quantiles. A violin plot is also plotted to represent the shape of the distribution.

- Extended Figure 2.pdf

Distribution of Spearman correlation coefficients between CCLE and tumor purity adjusted TCGA miRNA expression profiles for each cancer type.

Similar to Extended Fig. S1, except that the correlation was calculated between CCLE cell line and the tumor purity adjusted TCGA miRNA expression profiles.

- Extended Figure 3.pdf

Cancer hallmark GSVA score for the miRNA expression of TCGA and CCLE.

The upregulated hallmarks GSVA scores were compared between CCLE and tumor purity adjusted TCGA (a) and between TCGA before and after purity adjustment (b) using a t-test and visualized as heatmaps.

- Extended Figure 4.pdf

The miRNA out-degree distribution in target networks of selected cancer types.

The degree distribution is plotted for CCLE and TCGA separately. The selection criteria for cancer types and methods for constructing miRNA–mRNA networks are described in the Methods section.

- Extended Figure 5.pdf

The miRNA–mRNA networks for miRNAs with conserved target genes between CCLE cell lines and TCGA tumor samples.

For each miRNA within each cancer type, a Fisher exact test was performed to find the significance of the overlap of targets between CCLE and TCGA networks. The top five most significant miRNA hubs are shown in the figure for each cancer types. miRNAs are shown as triangles and target genes as circles. Edges represent significant negative correlation between each miRNA and its predicted targets in both CCLE and TCGA. The Fisher exact test FDR values are shown below each miRNA name.

Software availability

The source code for this project can be found at: Zenodo: hanwenzhu/mir-tcga-ccle-paper. <http://doi.org/10.5281/zenodo.4726328>.²²

This project contains the following files;

- README.md (Description for the github repository)
- compare. Rmd (Rmarkdown file including scripts for analysis done in this manuscript, including visualization, correlation analysis, gene set variation analysis and network analysis)
- download. R (R script for downloading TCGA miRNA expression data using TCGAbiolinks¹⁶)
- preprocessing. Rmd (Rmarkdown file including scripts for processing and normalizing data used in this manuscript)
- LICENSE (MIT license for the codes in the github repository)

License: MIT

References

1. Bartel DP: **MicroRNAs: Target Recognition and Regulatory Functions.** *Cell.* 2009; **136**: 215–233.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Jonas S, Izaurralde E: **Towards a molecular understanding of microRNA-mediated gene silencing.** *Nat Rev Genetics.* 2015; **16**: 421–433.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Ha M, Kim VN: **Regulation of microRNA biogenesis.** *Nat Rev Mol Cell Biol.* 2014; **15**: 509–524.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Dhawan A, Scott JG, Harris AL, et al.: **Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors.** *Nat. Commun.* 2018; **9**: 1–13.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Lin S, Gregory RI: **MicroRNA biogenesis pathways in cancer.** *Nat. Rev. Cancer.* 2015; **15**: 321–333.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Shin VY, Siu JM, Cheuk I, et al.: **Circulating cell-free miRNAs as biomarker for triple-negative breast cancer.** *Br. J. Cancer.* 2015; **112**: 1751–1759.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Montani F, et al.: **MiR-test: A blood test for lung cancer early detection.** *J. Natl. Cancer Inst.* 2015; **107**.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Singh PK, et al.: **Serum microRNA expression patterns that predict early treatment failure in prostate cancer patients.** *Oncotarget.* 2014; **5**: 824–840.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Wiggins JF, et al.: **Development of a lung cancer therapeutic based on the tumor suppressor microRNA-34.** *Cancer Res.* 2010; **70**: 5923–5930.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Li J, et al.: **Registered report: The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44.** *Elife.* 2015; **4**.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Trang P, et al.: **Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice.** *Mol. Ther.* 2011; **19**: 1116–1122.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Network, The Cancer Genome Atlas (TCGA) Research. (Accessed: 18th December 2019).
[Reference Source](#)
13. Ghandi M, et al.: **Next-generation characterization of the Cancer Cell Line Encyclopedia.** *Nature.* 2019.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Yu K, et al.: **Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types.** *Nat. Commun.* 2019; **10**: 3574.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Broad Institute Cancer Cell Line Encyclopedia (CCLE): (Accessed: 10th February 2020).
[Reference Source](#)
16. Colaprico A, et al.: **TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.** *Nucleic Acids Res.* 2016; **44**: e71–e71.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.** *Nucleic Acids Res.* 2012; **40**: 4288–4297.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Leek JT, Johnson WE, Parker HS, et al.: **sva: Surrogate Variable Analysis. R package version 3.32.1.** 2019. (Accessed: 5th June 2019).
[Reference Source](#)
19. Ritchie ME, et al.: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** 2015; **43**.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Carter SL, et al.: **Absolute quantification of somatic DNA alterations in human cancer.** *Nat. Biotechnol.* 2012; **30**: 413–421.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Friedman RC, Farh KKH, Burge CB, et al.: **Most mammalian mRNAs are conserved targets of microRNAs.** *Genome Res.* 2009; **19**: 92–105.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Zhu H: *hanwenzhu/mir-tcga-ccle-paper.* 2021.
[Publisher Full Text](#)
23. Yu H, Pardoll D, Jove R: **STATs in cancer inflammation and immunity: A leading role for STAT3.** *Nat Rev Cancer.* 2009;
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Sánchez-Cid L, et al.: **MicroRNA-200, associated with metastatic breast cancer, promotes traits of mammary luminal progenitor cells.** *Oncotarget.* 2017; **8**: 83384–83406.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Chen WX, et al.: **MIR-221/222: Promising biomarkers for breast cancer.** *Tumor Biol.* 2013; **34**: 1361–1370.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 04 February 2022

<https://doi.org/10.5256/f1000research.55761.r118670>

© 2022 Palve V. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vinayak C. Palve

Department of Drug Discovery, H. Lee Moffitt Cancer Center, Tampa, FL, USA

In the present manuscript the authors aim to provide critical biological knowledge on the extent of consensus microRNA expression and regulation between cell line models and primary human tumours. The authors have utilised the CCLE cell lines and TCG tumour samples databased to compare the expression of miRNAs associating the hallmarks of cancer pathways a regulatory network. The authors have nicely shown the regulatory network and correlation between the miRNA expression between cell lines and tumor tissues. However, I believe there are some major points that needs to be addressed.

Major:

The present manuscript only concludes on the basis of co-relation of expressions from datasets to make conclusions. The authors should show some experimental evidence to validate their findings. The authors can discuss giving certain examples like targeting XYZ miRNA using RNAi or CRISPR studies in certain cancer cell lines to show validation of their network, etc. Adding this would strengthen the overall impact and quality of the manuscript.

Minor:

Most of the figures (Fig 2, 3 , 5 & 6) in the manuscript needs major updates as the resolution is not enough. Kindly change the figures to at least 300dpi.

I believe the authors have utilised databases to compare the expression of miRNAs associating the hallmarks of cancer pathways a regulatory network but to improve the overall quality and impact of the manuscript the suggested experiments and data would be highly crucial. Without these changes the conclusions drawn are not adequately supported by the results.

Thank you.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Drug discovery, gene expression, signaling mechanism, prognostic biomarkers, proteomics studies

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 29 July 2021

<https://doi.org/10.5256/f1000research.55761.r90434>

© 2021 Zang C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chongzhi Zang

University of Virginia, Charlottesville, VA, USA

In this manuscript, Zhu and Ning performed a systematic and comparative analysis of transcriptomic profiling data from TCGA and CCLE across 22 cancer types to characterize the miRNA regulatory network. The results demonstrate high variances between primary tumor samples and cancer cell line models, reveal heterogeneity of cancer cells, and emphasize the importance of selecting the right cell line model when studying cancer gene regulation. The manuscript is written clearly, including a thorough discussion on the scope and limitations of the work. The computational methods are appropriate, and the authors provide sufficient details on the methods as well as source code and data, which not only ensure reproducibility but also can be useful resource to both cancer research and gene regulation research community.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: computational biology, cancer genomics and epigenetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research