



## OPEN ACCESS

## EDITED BY

Yuriy L. Orlov,  
I. M. Sechenov First Moscow State  
Medical University, Russia

## REVIEWED BY

Teresa Colombo,  
National Research Council (CNR), Italy  
Andrey Zamyatnin,  
I. M. Sechenov First Moscow State  
Medical University, Russia  
Tingming Liang,  
Nanjing Normal University, China

## \*CORRESPONDENCE

Stepan Nersisyan,  
stepan.nersisyan@jefferson.edu

## †PRESENT ADDRESS

Stepan Nersisyan,  
Computational Medicine Center, Sidney  
Kimmel Medical College, Thomas  
Jefferson University, Philadelphia, PA,  
United States

## SPECIALTY SECTION

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

RECEIVED 15 October 2022

ACCEPTED 21 November 2022

PUBLISHED 01 December 2022

## CITATION

Nersisyan S, Zhiyanov A, Engibaryan N,  
Maltseva D and Tonevitsky A (2022), A  
novel approach for a joint analysis of  
isomiR and mRNA expression data  
reveals features of isomiR targeting in  
breast cancer.  
*Front. Genet.* 13:1070528.  
doi: 10.3389/fgene.2022.1070528

## COPYRIGHT

© 2022 Nersisyan, Zhiyanov,  
Engibaryan, Maltseva and Tonevitsky.  
This is an open-access article  
distributed under the terms of the  
[Creative Commons Attribution License  
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original  
publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or  
reproduction is permitted which does  
not comply with these terms.

# A novel approach for a joint analysis of isomiR and mRNA expression data reveals features of isomiR targeting in breast cancer

Stepan Nersisyan<sup>1\*†</sup>, Anton Zhiyanov<sup>2</sup>, Narek Engibaryan<sup>1</sup>,  
Diana Maltseva<sup>1</sup> and Alexander Tonevitsky<sup>1,2,3</sup>

<sup>1</sup>Faculty of Biology and Biotechnology, HSE University, Moscow, Russia, <sup>2</sup>Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia, <sup>3</sup>Art Photonics GmbH, Berlin, Germany

A widely used procedure for selecting significant miRNA-mRNA or isomiR-mRNA pairs out of predicted interactions involves calculating the correlation between expression levels of miRNAs/isomiRs and mRNAs in a series of samples. In this manuscript, we aimed to assess the validity of this procedure by comparing isomiR-mRNA correlation profiles in sets of sequence-based predicted target mRNAs and non-target mRNAs (negative controls). Target prediction was carried out using RNA22 and TargetScan algorithms. Spearman's correlation analysis was conducted using miRNA and mRNA sequencing data of The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) project. Luminal A, luminal B, basal-like breast cancer subtypes, and adjacent normal tissue samples were analyzed separately. Using the sets of putative targets and non-targets, we introduced adjusted isomiR targeting activity (ITA)—the number of negatively correlated potential isomiR targets adjusted by the background (estimated using non-target mRNAs). We found that for most isomiRs a significant negative correlation between isomiR-mRNA expression levels appeared more often in a set of predicted targets compared to the non-targets. This trend was detected for both classical seed region binding types (8mer, 7mer-m8, 7mer-A1, 6mer) predicted by TargetScan and the non-classical ones (G:U wobbles and up to one mismatch or unpaired nucleotide within seed sequence) predicted by RNA22. Adjusted ITA distributions were similar for target sites located in 3'-UTRs and coding mRNA sequences, while 5'-UTRs had much lower scores. Finally, we observed strong cancer subtype-specific patterns of isomiR activity, highlighting the differences between breast cancer molecular subtypes and normal tissues. Surprisingly, our target prediction- and correlation-based estimates of isomiR activities were practically non-correlated with the average isomiR expression levels neither in cancerous nor in normal samples.

## KEYWORDS

isomiR, miRNA, targeting, breast cancer, TCGA

## Introduction

MiRNAs are short non-coding RNAs that post-transcriptionally regulate gene expression. A mature miRNA binds a target mRNA, which results in mRNA degradation or translational repression (Bartel, 2009). Target recognition by a miRNA heavily depends on nucleotides 2–7 at the 5′-end of the miRNA called seed region. Specifically, high complementarity of seed region with target sequence is required for successful miRNA-mRNA binding (Bartel, 2009). Gene expression regulation by miRNA molecules plays a vital role in cancer development, progression, and metastasis (Di Leva et al., 2014; Reddy, 2015).

The application of small RNA sequencing technology led researchers to identify miRNA isoforms (isomiRs): variants of a mature miRNA differing from each other by a few nucleotides at 5′, 3′, or both ends (Tomasello et al., 2021). While the complete pathway of isomiR biogenesis has not yet been discovered, several mechanisms have been shown to contribute to the generation of isomiRs. These mechanisms include but are not limited to heterogeneous cleavage of pri- and pre-miRNA hairpins by Drosha and Dicer enzymes, respectively, post-transcriptional nucleotide addition to the 3′-ends of miRNAs by nucleotidyl transferases (Tomasello et al., 2021). Importantly, length variation at a 5′-end of an isomiR alters the isomiR's targetome since the seed sequence is modified. Experimental validation of non-canonical 5′-isomiR targets was previously conducted for miR-9 (Tan et al., 2014), miR-34/449 (Mercey et al., 2017), miR-101 (Llorens et al., 2013), miR-183 (Telonis et al., 2015), and miR-411 (van der Kwast et al., 2020).

A common strategy for the bioinformatics analysis of the miRNA/isomiR targeting involves a two-step procedure. First, target mRNAs for a given isomiR are predicted based on nucleotide sequences. Several broadly used tools are available for making such predictions, including RNA22 (Miranda et al., 2006), TargetScan (McGeary et al., 2019), miRDB (Chen and Wang, 2020), DIANA-microT (Reczko et al., 2012) and others (Riolo et al., 2020). Then, isomiR and target (mRNA- or protein-level) expression profiles in a set of samples are used to select putative interactions supported by a significant negative correlation. Such a strategy was previously used, e.g., to discover critical miRNA-mRNA interactions in breast cancer (Telonis and Rigoutsos, 2018; Shkurnikov et al., 2019; Nersisyan et al., 2021b), colorectal cancer (Nersisyan et al., 2021a), prostate cancer (Magee et al., 2018) and chronic lymphocytic leukemia (Cimmino et al., 2005).

However, to the best of our knowledge, there were no attempts at unbiased evaluation of this strategy's validity. Muniategui with coauthors compared different expression correlation-based miRNA target prediction methods using databases of experimentally validated miRNA targets

(Muniategui et al., 2013). While this approach allowed the authors to estimate the sensitivity of the target prediction, the validated miRNA target databases are significantly biased towards well-studied miRNAs and highly complementary interactions in the case of low-throughput experiments (e.g., luciferase reporter assays). While CLIP-seq datasets could serve as an unbiased source of validated interactions, there is a problem related to the absence of gene expression repression for a considerable portion of identified interactions (Chu et al., 2020).

The main objective of the present study was to assess the validity of performing isomiR-mRNA or miRNA-mRNA correlation analysis on the results of sequence-based target prediction. For that, we comprehensively compared miRNA-mRNA expression level correlations between predicted targets and negative controls (briefly, mRNAs that do not contain miRNA seed region binding sites). The analysis was conducted using miRNA and mRNA sequencing data of primary breast cancer (BC) samples and adjacent normal tissues available in The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) project (PAM50 molecular subtypes of BC were analyzed separately). We aimed to perform unbiased and data-driven analysis, so we predicted isomiR targets with two bioinformatics tools based on different ideas, with both classical and non-classical seed region binding types. We also considered targeting within mRNA coding sequences (CDS) and 5′-UTRs.

## Materials and methods

### TCGA data acquisition and processing

Count-level TCGA isomiR expression data ( $n = 11089$  samples) were downloaded from the IsoMiRmap tool (Loher et al., 2021) official website (<https://cm.jefferson.edu/isomirmap/>). Reads exclusively mapped to the miRNA space were selected for further processing. Transcript-level mRNA-seq data for  $n = 10530$  TCGA samples were downloaded from the UCSC Xena portal (Goldman et al., 2020) (<https://xenabrowser.net/>), TOIL RSEM expected\_count dataset built on GENCODE 23 human reference genome was used.

Normalization of both miRNA-seq and mRNA-seq TCGA count-level data was performed using the median of ratios algorithm implemented as the “estimateSizeFactors” function in the DESeq2 R package (Love et al., 2014); “fpm” and “fpkm” functions were used to generate the final normalized expression tables for miRNA-seq and mRNA-seq data, respectively (i.e., mRNA-seq data was additionally normalized by the transcript lengths). Finally, we applied  $\log_2(x + 1)$  transformation for both miRNA-seq and mRNA-seq data.

PAM50 molecular subtyping results of primary breast cancer samples and normal tissues annotation were extracted from the UCSC Xena portal. We set a minimum of 100 sample thresholds for all subtypes to eliminate possible sample size-related biases in the downstream analysis. As a result, four sample groups were considered: luminal A primary tumors ( $n = 561$  samples), luminal B primary tumors ( $n = 210$ ), basal-like primary tumors ( $n = 183$ ), and adjacent normal tissues ( $n = 104$ ). Sample identifiers and corresponding group labels are available in [Supplementary Table S1](#).

Since 3'-end isomiR variations do not alter the isomiR targetome much (seed region not affected), we summed up the expression of isomiRs originating from the same miRNA, which have identical 5'-end sequences (5'-isomiRs). We used our previous notation for 5'-isomiRs (Nersisyan et al., 2022b): a number after the “|” symbol stands for a shift from the canonical 5'-end in the 5'-3' direction. For example, hsa-miR-192-5p|+1 differs from the canonical hsa-miR-192-5p miRNA by the absence of the first nucleotide on its 5'-end.

For further processing, we selected 139 5'-isomiRs with the highest median expression—this was a minimum number of isomiRs that covered 99% of the total median isomiR expression. The used threshold corresponded to approximately 100 DESeq2-normalized FPM. A less strict thresholding was used for the mRNA-seq data: transcripts with zero reads in more than half of the analyzed samples were discarded. The final list of analyzed 5'-isomiRs and their median expression levels is available in [Supplementary Table S2](#).

## IsomiR target prediction

Two tools were used to predict 5'-isomiR targets: RNA22 (Miranda et al., 2006; Loher and Rigoutsos, 2012) and TargetScan 7.2 (Agarwal et al., 2015). Precomputed RNA22 predictions for the canonical miRNAs were downloaded from the official tool's website (<https://cm.jefferson.edu/rna22-full-sets-of-predictions/>; ENSEMBL 96 and miRBase 22 version). For the remaining 30 non-canonical isomiRs, a program that allows one to submit target prediction batch requests was used (<https://cm.jefferson.edu/rna22/Interactive/remoteRNA22v2.zip>, ENSEMBL 96 transcript sequences). The default target predictions settings were used. Since human reference transcriptome versions differed for mRNA-seq reads mapping and isomiR target prediction, we selected transcripts with identical sequences in both versions for further analysis.

The set of predicted isomiR-mRNA interactions was classified into five types based on seed region binding motifs:

8mer, 7mer-m8, 7mer-A1, 6mer (classical seed binding types), and other (including unlimited G:U wobbles and up to one mismatch or unpaired nucleotide in a seed region). Custom TargetScan predictions were done as previously described (Nersisyan et al., 2022a): publicly available Perl scripts and 3'-UTR sequences were downloaded from the official website ([https://www.targetscan.org/cgi-bin/targetscan/data\\_download.vert72.cgi](https://www.targetscan.org/cgi-bin/targetscan/data_download.vert72.cgi)). Assessment of binding with imperfect seed complementarity (3'-compensatory sites) was not possible with TargetScan custom prediction mode. The results of RNA22 and TargetScan predictions are available in figshare repository (<https://doi.org/10.6084/m9.figshare.21579729.v1>). MiRTarBase 9 database (Huang et al., 2022) was used to retrieve experimentally verified miRNA-mRNA interactions (luciferase reporter assays).

## Composition of isomiR non-target mRNA sets

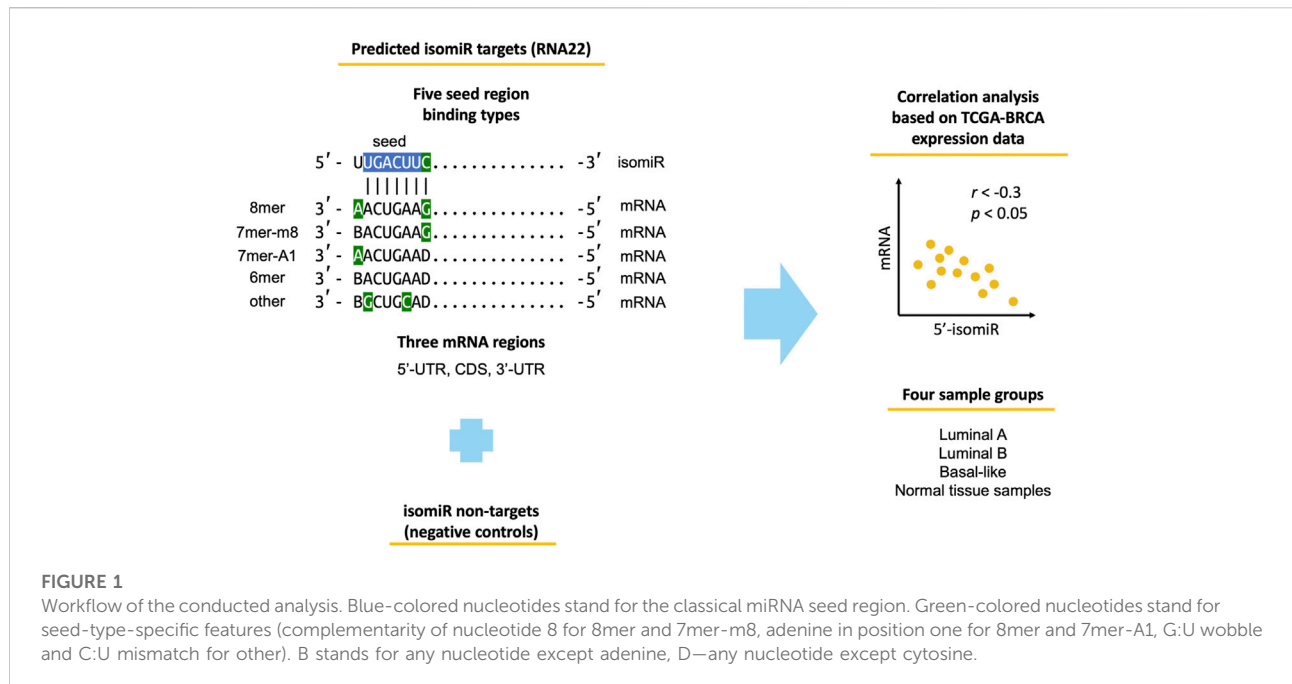
For a given 5'-isomiR, we composed a set of all mRNAs which do not contain the 6mer seed region binding site in mRNA sequences (including an unlimited number of G:U wobbles). The computational approach for solving this problem was based on the composition of the hash table, which mapped all possible 6mers to their positions in all mRNA sequences. The set of negative control interactions is available in [Supplementary Data 1](#).

## Calculating isomiR targeting activity (ITA) and adjusted ITA

We calculated Spearman's correlation coefficients and corresponding  $p$ -values between all 5'-isomiR and mRNA expression levels (separately for the four groups of BC samples). Benjamini-Hochberg procedure was used to adjust  $p$ -values.

For a given 5'-isomiR, we denote the number of predicted targets as  $n_{tar}$  and the number of predicted targets supported by a significant negative correlation of expression levels ( $r < -0.3$ ,  $FDR < 0.05$ ) as  $m_{tar}$  ( $m_{tar} \leq n_{tar}$ ). The analogous terms were used for the isomiR non-targets:  $n_{nontar}$  (number of non-targets predicted) and  $m_{nontar}$  (number of anti-correlated non-targets:  $r < -0.3$ ,  $FDR < 0.05$ ). For the downstream text, we denote  $m_{tar}$  as *isomiR targeting activity (ITA)* and  $m_{tar} - n_{tar} \frac{m_{nontar}}{n_{nontar}}$  as *adjusted ITA*. The statistical significance of an isomiR activity was assessed with the one-sided Fisher's exact test applied to the following contingency

table: 
$$\begin{pmatrix} m_{tar} & n_{tar} - m_{tar} \\ m_{nontar} & n_{nontar} - m_{nontar} \end{pmatrix}$$



Benjamini–Hochberg procedure was used to adjust  $p$ -values.

## Differential expression analysis

We used DESeq2 1.36 (Love et al., 2014) to identify differentially expressed isomiRs for each pair of the considered sample groups (luminal A, luminal B, and basal-like BCs, normal samples). An isomiR was considered as differentially expressed if the fold change was lower than 0.5 or higher than two and the adjusted  $p$ -value was lower than 0.05.

## Statistical analysis and programming

We used SciPy stats 1.9 (Virtanen et al., 2020) for conducting statistical analysis (Spearman's correlation, Fisher's exact test). Pandas 1.3 (McKinney, 2010) and NumPy 1.23 (Harris et al., 2020) were used for miscellaneous computations. Plots were constructed with Seaborn 0.11 (Waskom, 2021) and ggVennDiagram (Gao et al., 2021).

## Results

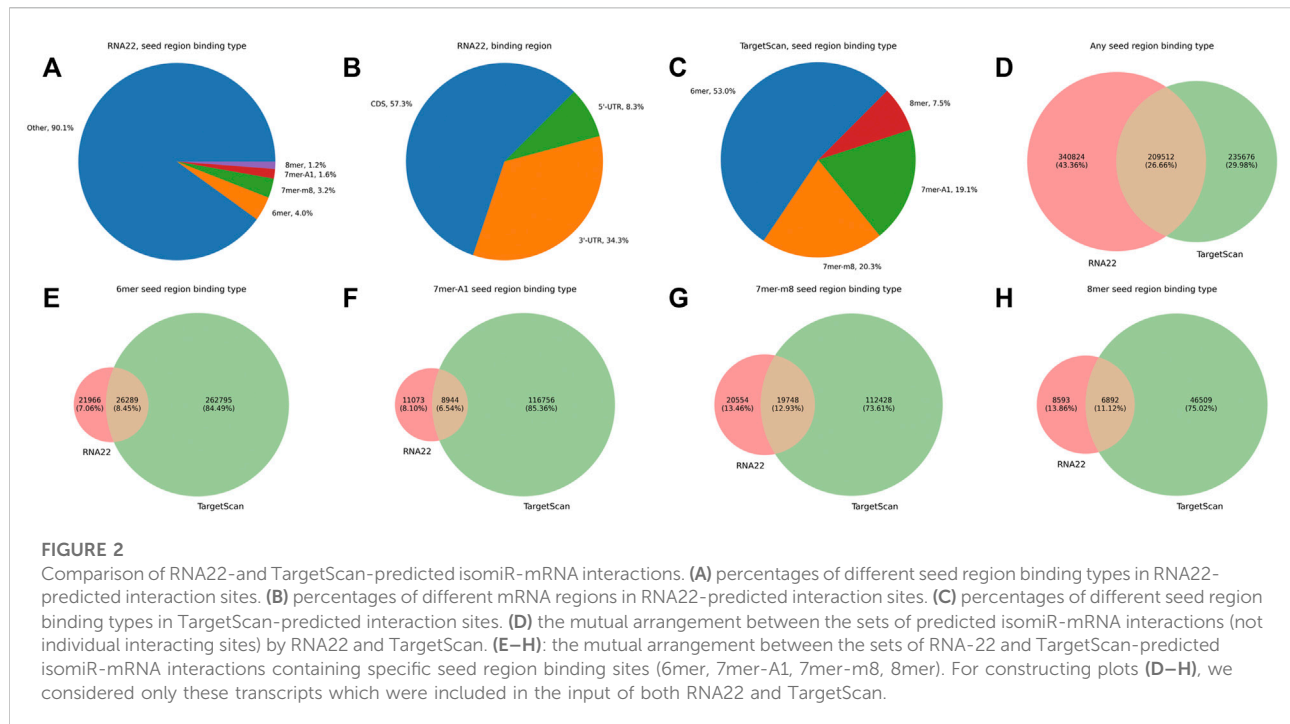
### Bioinformatics analysis workflow

The study workflow is outlined in Figure 1. We used miRNA-seq and mRNA-seq data of TCGA-BRCA primary tumor

samples corresponding to three molecular subtypes (luminal A, luminal B, basal-like) and adjacent normal tissue samples. We identified 139 highly expressed 5'-isomiRs and predicted their target mRNAs using two sequence-based target prediction tools: RNA22 and TargetScan. The tools are based on different ideas. Namely, TargetScan favors evolutionary conserved binding sites, relies on the classical seed region binding types (8mer, 7mer-m8, 7mer-A1, 6mer), and predicts targets only within 3'-UTR. In contrast, RNA22 is a pattern-based method, which does not consider target site conservation, predicts targets within 5'-UTR, CDS, and 3'-UTR, and is significantly more tolerant to mismatches in a seed region. Aside from target predictions, we also composed sets of non-targets (i.e., negative controls) for each 5'-isomiR. Non-targets were defined as mRNAs that do not contain regions complementary to an isomiR's seed region (including possible G:U wobbles). Finally, we calculated Spearman's correlation coefficient between expression levels of all 5'-isomiRs and their targets and non-targets.

### Differences between the sets of RNA22- and TargetScan-predicted targets

Before analyzing the joint expression profiles of isomiRs and their putative targets, we compared the sets of predicted isomiR-target interactions between RNA22 and TargetScan tools. Only 9.9% of RNA22-predicted target sites corresponded to the classical seed region binding (4.0% for 6mer, 3.2% for 7mer-m8, 1.6% for 7mer-A1, and 1.2% for 8mer), while the vast



majority of putative targets had non-classical seed binding (90.1%, Figure 2A). High numbers of target sites were found in 3'-UTR and CDS (57.3% and 34.3%, respectively), while 5'-UTR binding was reported less frequently (8.3%, Figure 2B). TargetScan predictions included only classical seed region binding types within 3'-UTRs, with a prevalence of 6mer sites (53%, Figure 2C). RNA22 predicted 1.2 times more isomiR-mRNA interactions than TargetScan, and the tools shared 26.7% of the putative interactions (Figure 2D). At the same time, RNA22 missed the majority of classical seed region binding sites in 3'-UTRs compared to TargetScan (Figures 2E–H). For the downstream analysis, we used the union of RNA22- and TargetScan-predicted isomiR-mRNA interactions. With such a target prediction strategy, we were able to cover 89.1% out of 1,608 experimentally validated interactions between canonical miRNAs and their target mRNAs. In particular, 49.6% of validated interactions were predicted by both RNA22 and TargetScan, 6.5%—only by RNA22, and 33.0%—only by TargetScan.

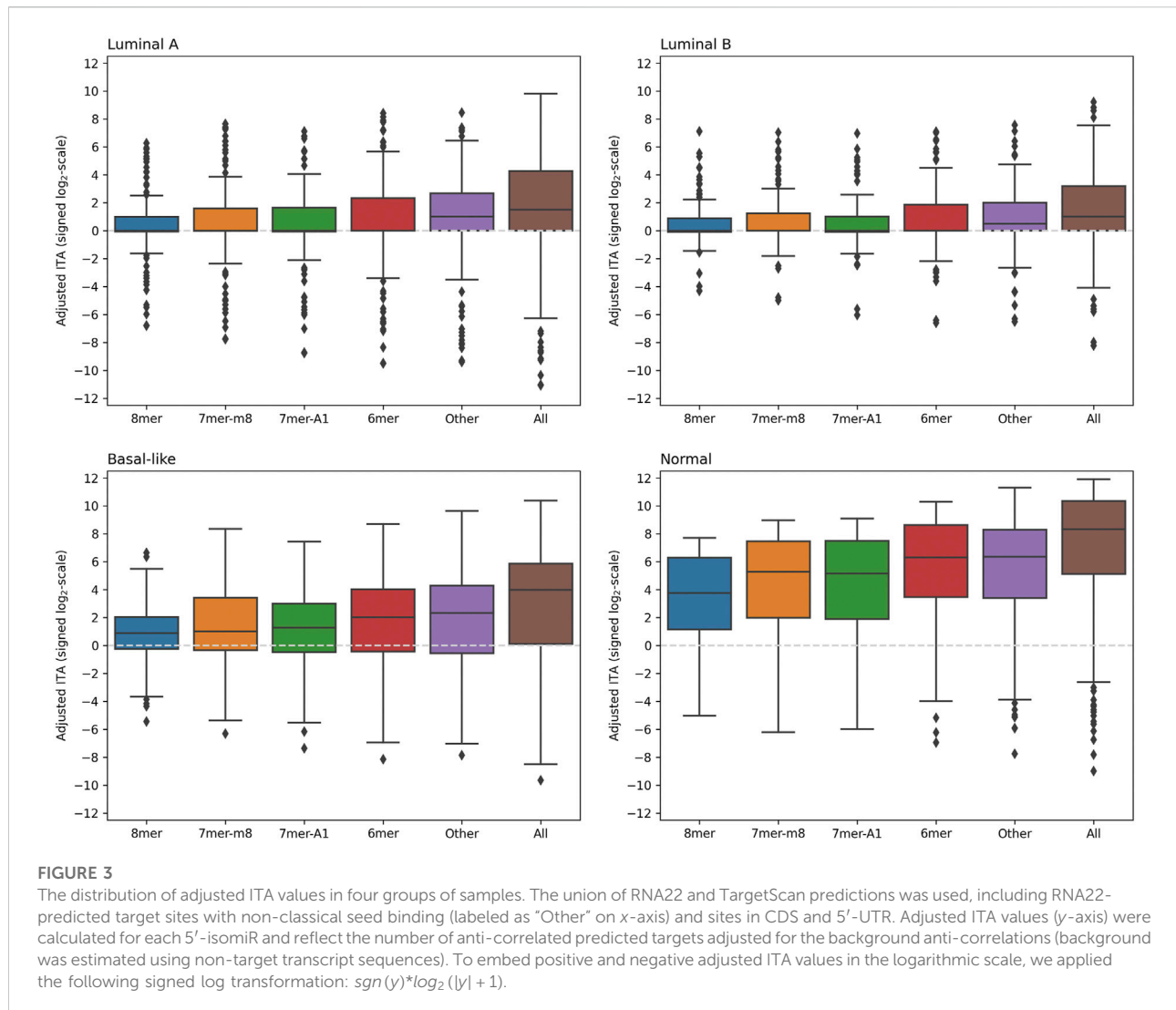
isomiRs tend to stronger anti-correlate with the putative targets rather than with the non-targets.

To systematically compare correlation profiles of 5'-isomiRs with their targets and non-targets, we introduced two measures of isomiR activity. The first one, denoted as *isomiR targeting activity (ITA)*, was calculated as a number of predicted targets, which were additionally supported by the significant negative correlation. However, a negative correlation between expression levels of a 5'-isomiR and its target gene could be explained by non-random side effects, such as opposite-sided regulation of

both molecules by a transcription factor. Therefore, to estimate the number of such false positive interactions (i.e., negatively correlated 5'-isomiRs and their target genes with no direct effect), we used the non-target mRNA sets. Specifically, *background ITA* was calculated as the fraction of non-targets that were significantly anti-correlated with the 5'-isomiR, multiplied by the number of predicted isomiR targets. Finally, *adjusted ITA* was calculated by subtracting background ITA from unadjusted ITA. In other words, adjusted ITA reflects the difference between the number of negatively correlated potential isomiR targets and the expected number of negatively correlated non-targets.

Note that near-zero or negative adjusted ITA does not imply the absence of targeting. For example, consider the synthetic case where the only target of a 5'-isomiR is a hub TF with multiple downstream target genes, which are not direct targets of the isomiR. In this case, unadjusted ITA equals one, while background activity is much greater since the isomiR is anti-correlated with TF targets by transitivity. Thus, the adjusted ITA of the isomiR is negative despite having one functional target.

A comparison of adjusted ITA values calculated using 5'-isomiR and mRNA expression levels for the union of RNA22 and TargetScan predictions (including all seed binding types and mRNA regions) is shown in Figure 3. As can be seen, the distributions were essentially skewed toward the positive direction for each seed region binding type and each sample group. Thus, for the majority of 5'-isomiRs, ITA was higher than the background correlations level. Notably, adjusted ITA values were dramatically higher in the normal tissues compared to the cancerous ones. Thus, we observed global downregulation of

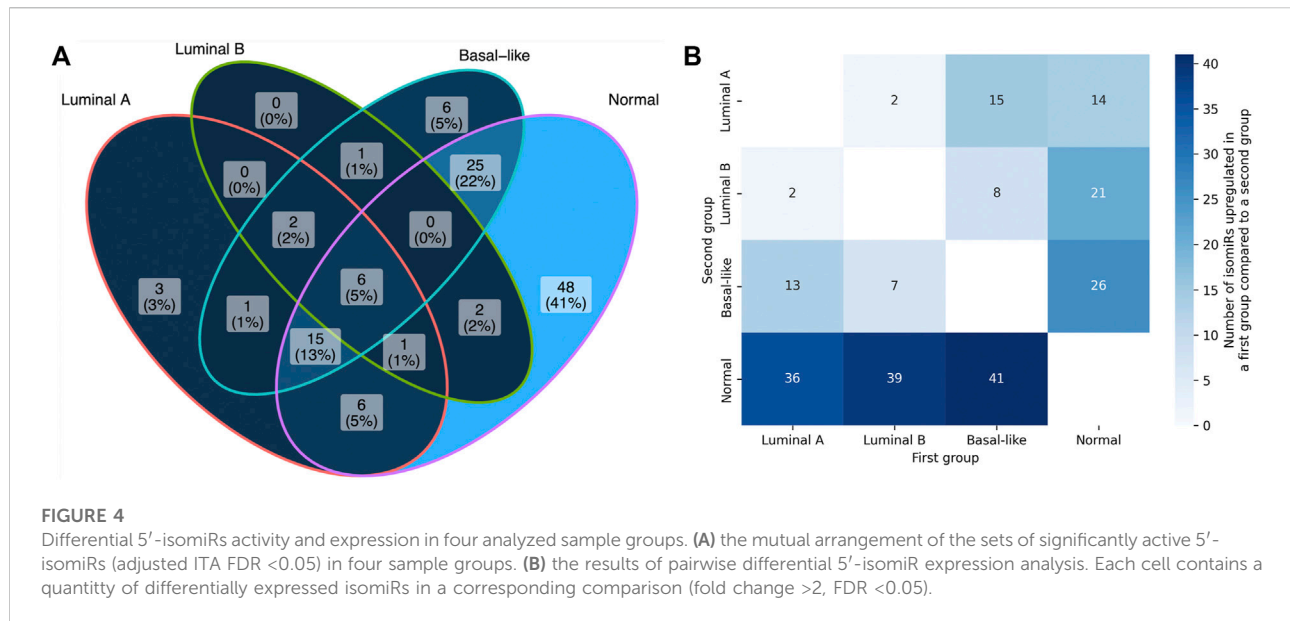


isomiR targeting in breast cancer. The tables with adjusted ITA scores (union of RNA22 and TargetScan) for three BC subtypes and normal samples are available in [Supplementary Table S3-S6](#).

In consistence with the previous section, RNA22 and TargetScan had different contributions to the correlation analysis. Namely, in the case of RNA22 predictions, non-classical seed binding types (referred to as “Other” in [Figure 3](#)) had significantly higher adjusted ITA scores compared to the four classical seed types ([Supplementary Figure S1](#)). Moreover, in the case of luminal A and luminal B BC subtypes, the whole distribution of adjusted ITA values (except outliers) was concentrated near zero, which was not the case for the “Other” binding type. Near-equal adjusted ITA scores were found for the 3'-UTR and CDS binding, while adjusted ITA in 5'-UTRs was systematically lower. In contrast to the RNA22 case, an essential mass of the TargetScan-based adjusted ITA distribution was in the

positive zone for each classical seed binding type ([Supplementary Figure S2](#)). Thus, combining two target prediction algorithms ultimately allowed us to make valid target predictions both for classical and non-classical seed binding sites. It is also worth noting that adjusted ITA scores calculated separately using RNA22 and TargetScan were highly correlated: Spearman's  $r = 0.70$ ,  $p = 5.25 \times 10^{-22}$  for Luminal A,  $r = 0.51$ ,  $p = 1.73 \times 10^{-10}$  for Luminal B,  $r = 0.88$ ,  $p = 3.54 \times 10^{-47}$  for Basal-like, and  $r = 0.85$ ,  $p = 1.31 \times 10^{-39}$  for Normal samples.

We were not able to identify any further subgroups within the predicted targets, which would have significantly different adjusted ITA scores. The tested groupings included partitioning of RNA22-predicted interactions by ten quantiles of binding energy, separating RNA22-predicted non-classical target sites by a number of G:U wobbles and mismatches, and partitioning of TargetScan-predicted interactions by ten quantiles of weighted context++ score.



## IsomiR activities markedly vary across BC subtypes and normal tissues

Comparison of the numbers of anti-correlated isomiR-mRNA pairs between the sets of predicted targets and non-targets with Fisher's exact test allowed us to identify 5'-isomiRs with a statistically significant activity over the background. The mutual arrangement of the significant isomiR sets for four sample groups (luminal A, luminal B, basal-like, normal) is presented in Figure 4A. As can be seen, isomiR activity showed clear subtype-specific patterns. Consistently with the previous section, the maximum number of significantly active isomiRs was detected in normal breast tissue samples: 103 out of 139 considered isomiRs. Normal tissues were followed by the basal-like (56 isomiRs), luminal A (34 isomiRs), and luminal B (12 isomiRs) BCs. Only eight isomiRs were significantly active in all BC subtypes: hsa-miR-17-3p|0, hsa-miR-30b-5p|0, hsa-miR-93-5p|0, hsa-miR-101-3p|-1, hsa-miR-106b-5p|0, hsa-miR-182-5p|0, hsa-miR-200b-3p|0, and hsa-miR-210-3p|0.

We then wondered whether the results of differential isomiR activity are in line with expression levels of these isomiRs in four considered sample groups. First, we noticed that the unsupervised clustering of samples based on the isomiR expression profiles was similar enough to the four considered sample groups (importantly, normal samples clustered separately from the cancerous ones, Supplementary Figure S3). Then, we conducted differential isomiR expression analysis for each pair of sample groups (Figure 4B). In contrast to the higher number of active isomiRs in normal tissues, we saw a higher number of isomiR upregulation events in cancer than downregulation events (this held for each BC subtype). The other interesting observation could be made while comparing luminal A and

luminal B subtypes: only four isomiRs passed the differential expression significance thresholds (fold change and adjusted *p*-value), while the Jaccard index for the sets of the active isomiRs (9 isomiRs in the intersection divided by 37 isomiRs in the union) was equal to just 0.24. The detailed analysis of intersections between the sets of active/not active and upregulated/downregulated isomiRs for each pair of sample groups also did not allow us to find any strong associations between an isomiR's expression level and significance of adjusted ITA score (Supplementary Figure S4). Finally, it is worth noting that we did not find a correlation between adjusted ITA scores and isomiR median expression levels neither in cancerous nor in normal samples (Spearman's  $r < 0.18$ ). Thus, the analyzed data told us that the variation of isomiR expression levels and correlations between isomiRs and putative targets constitute two independent dimensions.

## Discussion

Correlation analysis applied to expression data of miRNAs and mRNAs in a set of samples is widely used to select important miRNA-mRNA interactions. In this manuscript, we assessed the validity of this approach by performing a comprehensive comparison of correlations between the sets of predicted isomiR targets and non-targets. Ultimately, we found that for most isomiRs, the number of anti-correlating mRNAs was higher among predicted targets rather than non-targets. These observations support the validity of the commonly used bioinformatics approach, consisting in searching for negative correlations among sequence-based predicted isomiR-mRNA interactions.

We used two commonly applied sequence-based miRNA target prediction tools to conduct bioinformatics analysis: RNA22 and TargetScan. TargetScan focuses on four classical seed region binding types (8mer, 7mer-m8, 7mer-A1, 6mer) and 3'-UTR targeting. In contrast, RNA22 mainly predicted sites with non-classical seed binding and many sites within mRNA CDS and 5'-UTR. Indeed, four classical sequence types show striking enrichment in large-scale studies involving miRNA overexpression and transcriptome/proteome profiling. Specifically, Lim with co-authors overexpressed miR-1 and miR-124 in HeLa cells and found the most enriched nucleotide motifs in downregulated transcripts detected by microarray analysis—the answer was 6mer seed region for both miRNAs (Lim et al., 2005). Similar results were further derived for five miRNAs using a proteomic approach (Selbach et al., 2008), and for 25 miRNAs using RNA sequencing (Liu and Wang, 2019). Indeed, such experimental setups do not provide evidence of direct miRNA-mRNA interactions. Several Ago-CLIP methods were developed to close this gap, including HITS-CLIP (Chi et al., 2009) and PAR-CLIP (Hafner et al., 2010). Concordantly, motif search in the datasets generated by both methods led researchers to the miRNA seed sequences. However, all the mentioned experiments provided sufficient evidence of efficient targeting despite unpaired nucleotides/bulges in a seed region, and successful targeting in mRNA coding sequences. Specifically, G-bulged sites were present in more than 15% of identified interactions by HITS-CLIP in a mouse brain and were shown to be conserved (Chi et al., 2012); Hafner et al. reported about 7% of non-classical seed sites in PAR-CLIP data (Hafner et al., 2010). Both HITS-CLIP and PAR-CLIP datasets supported extensive miRNA binding within coding sequences of mRNAs. Ribosome profiling data upon miRNA transfection suggested functional differences in CDS and 3'-UTR miRNA binding: CDS sites were more associated with inhibiting translation, while 3'-UTR sites were better at initiating mRNA degradation (Hausser et al., 2013). In a recent study, McGeary et al. used their novel  $K_d$ -based mathematical model to conclude that CDS-located sites have 5.5-folds lower affinity compared to the 3'-UTR sites (McGeary et al., 2019). Target repression through 5'-UTR binding was also reported (Lytle et al., 2007).

Importantly, our analysis was in agreement with both models: the adjusted ITA scores were skewed towards the positive direction in cases of both RNA22-and TargetScan-predicted targets (though 5'-UTR targeting was predicted to be much less efficient compared to 3'-UTR and CDS). However, one limitation of the conducted analysis should be noted. Namely, a single mRNA could contain more than one binding site for a given isomiR, and these sites could be localized in different parts of mRNA or have different seed region binding types. In this case, the same isomiR-mRNA pair was considered in all corresponding designs.

With the use of the developed ITA scores, we were able to select isomiRs that were anti-correlated with a statistically significant number of predicted targets over the background. It turned out that the sets of the most active isomiRs poorly overlapped between the three subtypes of BC and the normal tissue samples. Only eight common 5'-isomiRs for three cancer subtypes were found. Based on these results, we emphasize the importance of considering sample heterogeneity while performing isomiR-mRNA correlation analysis for cancer data. Another striking observation was the absence of correlation between the adjusted ITA scores and the median isomiR expression levels in all considered sample groups. These results could mean that the expression level of a 5'-isomiR is not the main determinant of its functional activity (at least when the set of highly expressed isomiRs is considered). Alternative mechanisms controlling the miRNA targeting were reported. For example, Kim with co-authors recently reported that RNA-binding proteins could significantly enhance miRNA targeting efficiency by making a secondary structure of a target site accessible to the miRNA-Argonaute complex (Kim et al., 2021). Further experiments are warranted to understand the nature of correlation-based isomiR activity in BC samples.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

## Author contributions

SN, DM, and AT contributed to the conceptualization and methodology. SN, AZ, NE contributed to the formal analysis. SN contributed to the writing—original draft. SN, AZ, NE, DM, and AT contributed to the writing—review and editing. All authors contributed to the article and approved the submitted version.

## Funding

The research was performed within the framework of the Basic Research Program at HSE University.

## Acknowledgments

The authors thank Maxim Shkurnikov for his valuable comments and discussions.



## Conflict of interest

AT is employed by Art Photonics GmbH.

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

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their

affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.1070528/full#supplementary-material>

## References

- Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4. doi:10.7554/eLife.05005
- Bartel, D. P. (2009). MicroRNAs: Target recognition and regulatory functions. *Cell* 136 (2), 215–233. doi:10.1016/j.cell.2009.01.002
- Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48 (D1), D127–D131. doi:10.1093/nar/gkz757
- Chi, S. W., Hannon, G. J., and Darnell, R. B. (2012). An alternative mode of microRNA target recognition. *Nat. Struct. Mol. Biol.* 19 (3), 321–327. doi:10.1038/nsmb.2230
- Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460 (7254), 479–486. doi:10.1038/nature08170
- Chu, Y., Kilikevicius, A., Liu, J., Johnson, K. C., Yokota, S., and Corey, D. R. (2020). Argonaute binding within 3'-untranslated regions poorly predicts gene repression. *Nucleic Acids Res.* 48 (13), 7439–7453. doi:10.1093/nar/gkaa478
- Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., et al. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl. Acad. Sci. U. S. A.* 102 (39), 13944–13949. doi:10.1073/pnas.0506654102
- Di Leva, G., Garofalo, M., and Croce, C. M. (2014). MicroRNAs in cancer. *Annu. Rev. Pathol.* 9, 287–314. doi:10.1146/annurev-pathol-012513-104715
- Gao, C. H., Yu, G., and Cai, P. (2021). ggVennDiagram: An intuitive, easy-to-use, and highly customizable R package to generate venn diagram. *Front. Genet.* 12, 706907. doi:10.3389/fgene.2021.706907
- Goldman, M. J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* 38 (6), 675–678. doi:10.1038/s41587-020-0546-8
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141 (1), 129–141. doi:10.1016/j.cell.2010.03.009
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585 (7825), 357–362. doi:10.1038/s41586-020-2649-2
- Hausser, J., Syed, A. P., Bilen, B., and Zavolan, M. (2013). Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* 23 (4), 604–615. doi:10.1101/gr.139758.112
- Huang, H. Y., Lin, Y. C., Cui, S., Huang, Y., Tang, Y., Xu, J., et al. (2022). miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* 50 (D1), D222–D230. doi:10.1093/nar/gkab1079
- Kim, S., Kim, S., Chang, H. R., Kim, D., Park, J., Son, N., et al. (2021). The regulatory impact of RNA-binding proteins on microRNA targeting. *Nat. Commun.* 12 (1), 5057. doi:10.1038/s41467-021-25078-5
- Lim, L. P., Lau, N. C., Garrett-Engle, P., Grimson, A., Schelter, J. M., Castle, J., et al. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433 (7027), 769–773. doi:10.1038/nature03315
- Liu, W., and Wang, X. (2019). Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* 20 (1), 18. doi:10.1186/s13059-019-1629-z
- Llorens, F., Banez-Coronel, M., Pantano, L., del Rio, J. A., Ferrer, I., Estivill, X., et al. (2013). A highly expressed miR-101 isomiR is a functional silencing small RNA. *BMC Genomics* 14, 104. doi:10.1186/1471-2164-14-104
- Loher, P., Karathanasis, N., Londin, E., Bray, P., Pliatsika, V., Telonis, A. G., et al. (2021). IsoMiRmap-fast, deterministic, and exhaustive mining of isomiRs from short RNA-seq datasets. *Bioinformatics* 37, 1828–1838. doi:10.1093/bioinformatics/btab016
- Loher, P., and Rigoutsos, I. (2012). Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics* 28 (24), 3322–3323. doi:10.1093/bioinformatics/bts615
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8
- Lytle, J. R., Yario, T. A., and Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc. Natl. Acad. Sci. U. S. A.* 104 (23), 9667–9672. doi:10.1073/pnas.0703820104
- Magee, R. G., Telonis, A. G., Loher, P., Londin, E., and Rigoutsos, I. (2018). Profiles of miRNA isoforms and tRNA fragments in prostate cancer. *Sci. Rep.* 8 (1), 5314. doi:10.1038/s41598-018-22488-2
- McGeary, S. E., Lin, K. S., Shi, C. Y., Pham, T. M., Bisaria, N., Kelley, G. M., et al. (2019). The biochemical basis of microRNA targeting efficacy. *Science* 366 (6472), eaav1741. doi:10.1126/science.aav1741
- McKinney, W. (2010). "Data structures for statistical computing in Python," in *Proceedings of the 9th Python in science conference*. Editor S. v. d. W. a. J. Millman, 56–61.
- Mercey, O., Popa, A., Cavard, A., Paquet, A., Chevalier, B., Pons, N., et al. (2017). Characterizing isomiR variants within the microRNA-34/449 family. *FEBS Lett.* 591 (5), 693–705. doi:10.1002/1873-3468.12595
- Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., et al. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126 (6), 1203–1217. doi:10.1016/j.cell.2006.07.031
- Muniategui, A., Pey, J., Planes, F. J., and Rubio, A. (2013). Joint analysis of miRNA and mRNA expression data. *Brief. Bioinform.* 14 (3), 263–278. doi:10.1093/bib/bbs028
- Nersisyan, S., Galatenko, A., Chekova, M., and Tonevitsky, A. (2021a). Hypoxia-Induced miR-148a downregulation contributes to poor survival in colorectal cancer. *Front. Genet.* 12, 662468. doi:10.3389/fgene.2021.662468
- Nersisyan, S., Galatenko, A., Galatenko, V., Shkurnikov, M., and Tonevitsky, A. (2021b). miRGTF-net: Integrative miRNA-gene-TF network analysis reveals key drivers of breast cancer recurrence. *PLoS One* 16 (4), e0249424. doi:10.1371/journal.pone.0249424
- Nersisyan, S., Gorboson, A., Makhonin, A., Zhiyanov, A., Shkurnikov, M., and Tonevitsky, A. (2022a). isomiRTar: a comprehensive portal of pan-cancer 5'-isomiR targeting. *PeerJ* 10, e14205. doi:10.7717/peerj.14205
- Nersisyan, S., Novosad, V., Galatenko, A., Sokolov, A., Bokov, G., Kononov, A., et al. (2022b). ExhauFS: Exhaustive search-based feature selection for classification and survival regression. *PeerJ* 10, e13200. doi:10.7717/peerj.13200
- Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., and Hatzigeorgiou, A. G. (2012). Functional microRNA targets in protein coding sequences. *Bioinformatics* 28 (6), 771–776. doi:10.1093/bioinformatics/bts043
- Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. *Cancer Cell Int.* 15, 38. doi:10.1186/s12935-015-0185-1

- Riolo, G., Cantara, S., Marzocchi, C., and Ricci, C. (2020). miRNA targets: From prediction tools to experimental validation. *Methods Protoc.* 4 (1), 1. doi:10.3390/mps4010001
- Selbach, M., Schwanhauser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature* 455 (7209), 58–63. doi:10.1038/nature07228
- Shkurnikov, M., Nikulin, S., Nersisyan, S., Poloznikov, A., Zaidi, S., Baranova, A., et al. (2019). LAMA4-Regulating miR-4274 and its host gene SORCS2 play a role in IGFBP6-dependent effects on phenotype of basal-like breast cancer. *Front. Mol. Biosci.* 6, 122. doi:10.3389/fmolb.2019.00122
- Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I. M., et al. (2014). 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.* 42 (14), 9424–9435. doi:10.1093/nar/gku656
- Telonis, A. G., Loher, P., Jing, Y., Londin, E., and Rigoutsos, I. (2015). Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Res.* 43 (19), 9158–9175. doi:10.1093/nar/gkv922
- Telonis, A. G., and Rigoutsos, I. (2018). Race disparities in the contribution of miRNA isoforms and tRNA-derived fragments to triple-negative breast cancer. *Cancer Res.* 78 (5), 1140–1154. doi:10.1158/0008-5472.CAN-17-1947
- Tomasello, L., Distefano, R., Nigita, G., and Croce, C. M. (2021). The MicroRNA family gets wider: The IsomiRs classification and role. *Front. Cell Dev. Biol.* 9, 668648. doi:10.3389/fcell.2021.668648
- van der Kwast, R., Woudenberg, T., Quax, P. H. A., and Nossent, A. Y. (2020). MicroRNA-411 and its 5'-IsomiR have distinct targets and functions and are differentially regulated in the vasculature under ischemia. *Mol. Ther.* 28 (1), 157–170. doi:10.1016/j.ymthe.2019.10.002
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* 17 (3), 261–272. doi:10.1038/s41592-019-0686-2
- Waskom, M. L. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* 6 (60), 3021. doi:10.21105/joss.03021