Analysis of coding gene expression from small RNA sequencing

Aygun Azadova[1], Anthonia Ekperuoh[1], Greg N. Brooke[*], Antonio Marco[*]

School of Life Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.

[1] Co-first authors

[*] To whom correspondence should be addressed: amarco@essex.ac.uk; gbrooke@essex.ac.uk

Running title: Coding gene expression from sRNAseq

ABSTRACT

The popularity of microRNA expression analyses is reflected by the existence of thousands of works in the literature in which sRNAseq has been performed in myriads of samples, but for which no matched total RNAseq exists. The lack of paired sequencing experiments severely restricts the analysis of microRNA-gene regulatory networks. We therefore explored whether protein-coding gene expression can be quantified from transcript fragments present in sRNAseq experiments. We first considered studies containing matched total RNA and small RNA from four human tissues and recovered transcript fragments from the small RNA experiments. We found that the expression levels of protein-coding gene transcripts from sRNAseq datasets was comparable to those from total RNAseq experiments (R^2 ranging 0.33-0.76). We then analysed the expression of both microRNAs and coding genes from the same sRNAseq experiments and demonstrated that known microRNA-target interactions are, as expected, inversely correlated with the expression profiles of these microRNA-mRNA pairs. To confirm the utility of this approach, we applied our method to investigate microRNAs and their targets in breast cancer patient samples for which only sRNAseq was performed and only microRNAs studied. We again identified a clear inverse correlation between the expression of microRNAs and mRNAs that they are predicted to regulate (i.e. presence of a target site). We also found that upregulation of mir-429 is associated with the downregulation of QKI, an RNA-binding protein, in breast cancer samples. In conclusion, although the analysis of mRNA fragments to study gene expression from sRNAseq experiments may have its limitations, it can be very informative in the study of microRNA-mRNA interactions.

INTRODUCTION

The analysis of gene expression is a cornerstone of functional genomics. Early works in molecular biology on gene expression were limited, since purified RNA is unstable and difficult to work with. However, the discovery (and use in the lab) of reverse transcriptase, permitting the controlled reverse transcription of RNAs into cDNAs (Maniatis et al. 1976), and the development of microarrays first (Schena et al. 1995), and high-throughput sequencing later (Margulies et al. 2005), boosted our capacity to analyse transcriptomes. Nowadays, the most common technique to analyse gene expression is RNA-sequencing, or RNAseq. This technique consists of first isolating RNA from a sample, then reverse transcribing into stable cDNA, and finally the sequencing using (mostly) Illumina technology (Bentley et al. 2008). RNAseq has multiple technical variations, either to identify specific types of transcripts or to characterise other RNA products. For instance, small RNA-sequencing (or sRNAseq) is a specific technique to sequence small RNAs, mostly microRNAs (Grimson et al. 2007; Ruby et al. 2007). To do so, a size selection step is introduced, in which cDNA sequences are selected to be in a particular size range. Some sRNAseq experiments are also coupled with ribosomal RNA (rRNA) depletion to ensure that the samples to be sequenced are enriched in microRNAs and not rRNAs.

MicroRNAs (miRNAs) are short non-coding RNAs around 22 nucleotides in length that play a crucial role in post-transcriptional regulation by targeting messenger RNA (mRNA) for degradation or translational repression (Shang et al. 2023). These small but important molecules are involved in a variety of cellular processes, including development, differentiation, and apoptosis (Ratti et al. 2020). The discovery that microRNAs target transcripts by partial pairwise complementarity, permitted the development of multiple target prediction methods (Lewis et al. 2003; Lai et al. 2003; Enright et al. 2003). In mammals, targeted transcripts are in most cases degraded (Baek et al. 2008). Thus, the joint analysis of microRNA and transcript expression can be used to identify microRNA/transcript interactions in combination with other microRNA target prediction programs (van Dongen et al. 2008).

Since the discovery that a deletion of two intronic microRNAs was associated with chronic lymphocytic leukaemia (Calin et al. 2002), the significance of miRNAs in cancer biology has been increasingly recognized (Vannini et al. 2018). The importance of microRNAs in cancer and other diseases is mirrored by the existence of thousands of publications for which sRNAseq has been performed, either in cells or lab-controlled samples, or in patient-derived material. However, in many instances only small RNAs were analysed, and no matched full transcriptome analysis exists. These studies that utilised precious clinical samples can be reanalysed as better methods and algorithms are developed to study sRNAseq. Unfortunately, changes in gene expression associated to changes in microRNA levels cannot be studied in principle as no matching transcriptomic RNAseq was performed. In this context, we studied whether sRNAseq experiments contained enough fragments from protein-coding transcripts to do a simultaneous analysis of microRNAs and their targets from the same sequencing library. The proposed streamlined approach will not only reduce costs but will also simplify experimental workflows, making the simultaneous analysis of transcriptomes and small RNAs affordable and efficient.

RESULTS

To investigate the potential of sRNAseq for the analysis of protein coding gene expression, we made use of the sRNAseq datasets generated by Meunier and collaborators (Meunier et al. 2013). These datasets contain the microRNA expression levels of multiple healthy tissues in several vertebrates, and we focused exclusively on human samples. The very same group (Kaessmann lab) also generated total RNAseq datasets from (mostly) the same tissue samples (Brawand et al. 2011). Hence, we selected four tissues for which comparable small and total RNAseq datasets are available: brain, cerebellum, heart and kidney. Both total and small RNAseq datasets were processed to identify fragments mapping to annotated transcripts (see Methods). The expression levels between *bona fide* reads from RNAseq and from fragments derived from sRNAseq experiments was comparable for all

four tissues, especially for brain (Figure 1). The correlation between expression levels from RNAseq and sRNAseq libraries was highest in brain (R^2=0.66) and cerebellum (R^2=0.49).

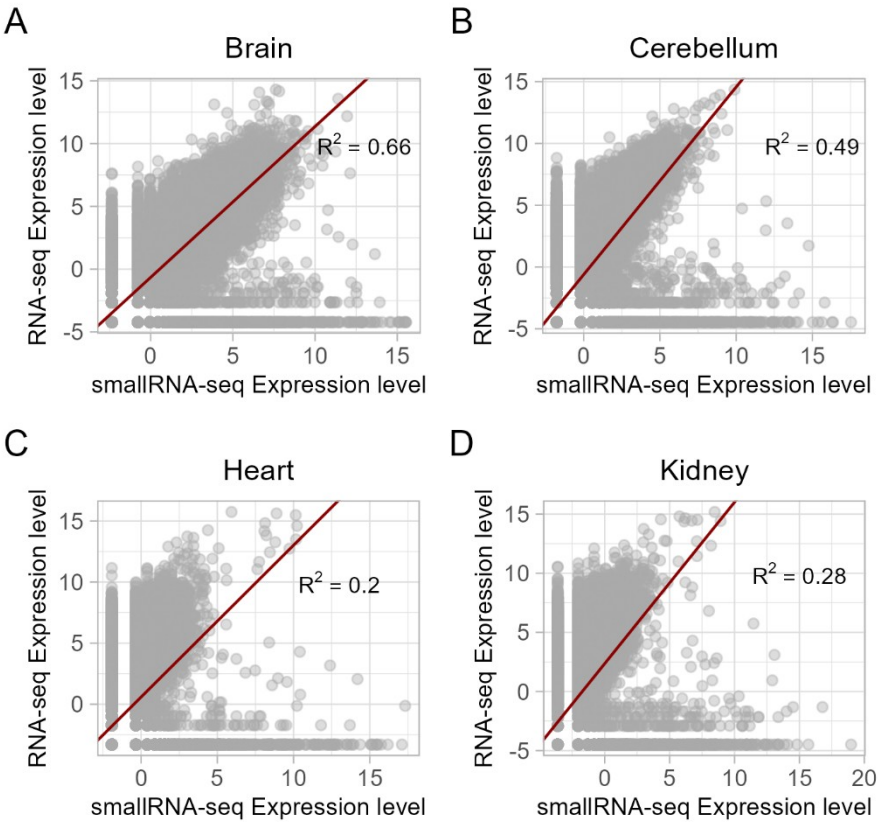


Figure 1. Comparison of the expression levels of gene transcripts from RNAseq and sRNAseq matches datasets. The normalised expression level from matched samples was plotted for brain (A), cerebellum (B), heart (C) and kidney (D) healthy human samples (Brawand et al. 2011; Meunier et al. 2013). A regression line was fitted in all four plots (red line), and the coefficient of determination (R^2) given within the plot.

When comparing the two different sequencing methods, we noticed that some of the biggest discrepancies arose from non-coding RNAs (small RNA genes), which as expected had a high number of reads mapped from the sRNAseq but were virtually absent in the total RNAseq datasets (Figure 1). Thus, for the analysis we excluded the non-coding RNAs and only considered protein-coding genes.

Following exclusion of the non-coding RNAs, the expression similarity between both types of sequencing datasets improved (Figure 2), being again particularly high for brain and cerebellum tissues, with R^2 of 0.76 and 0.62 respectively.

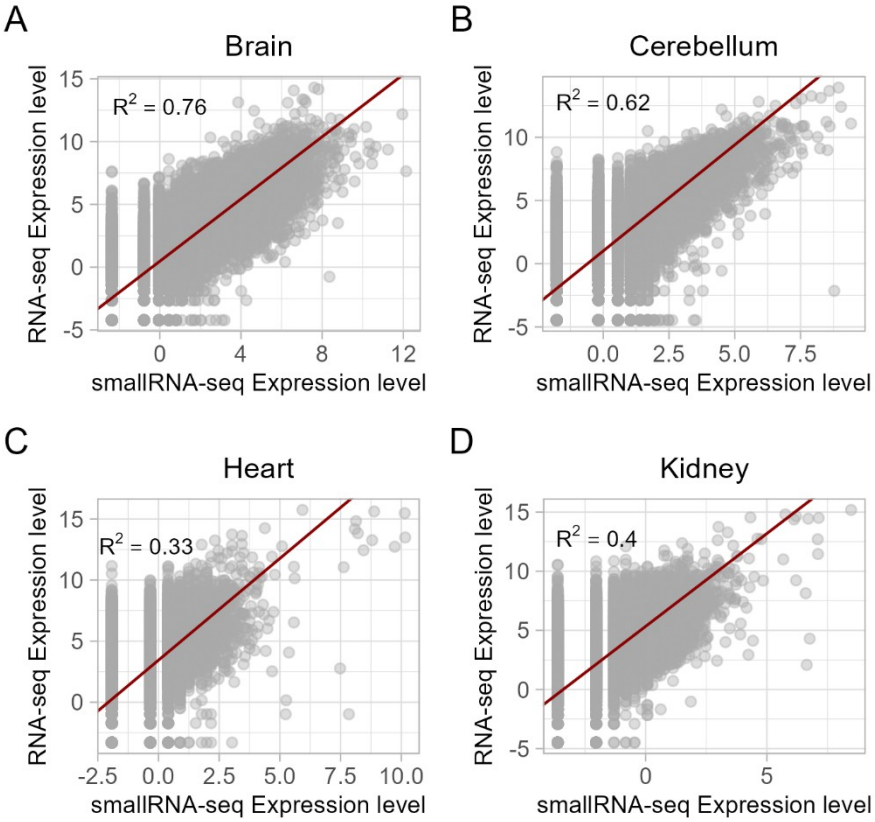


Figure 2. Comparison of the expression levels of protein-coding transcripts from RNAseq and sRNAseq matches datasets. Non-coding RNA were removed from the dataset and only protein-coding genes included. The normalised expression level from matched samples was plotted for brain (A), cerebellum (B), heart (C) and kidney (D) healthy human samples (Brawand et al. 2011; Meunier et al. 2013). A regression line was fitted in all four plots (red line), and the coefficient of determination (R^2) given within the plot.

Given that we successfully measured coding gene expression levels from fragments present in sRNAseq datasets, we also quantified the expression level of microRNAs and compared the expression profile between all microRNA/coding-gene pairs from the same sRNAseq. Then, we

identified those pairs with known interactions previously described in the literature (see Methods). In this analysis, we used all the samples available for small RNAs which include, on top of the four tissues studied already, a sample from testes. By plotting the ratio of known microRNA-target interactions as a function of the expression correlation, we clearly observe that anti-correlated pairs are enriched in target interactions, while highly correlated pairs show a paucity of targets (Figure 3). This indicates that the joint analysis of microRNAs and their potential targets, solely generated from sRNAseq experiments, can be used to study the function of microRNAs in specific tissues.

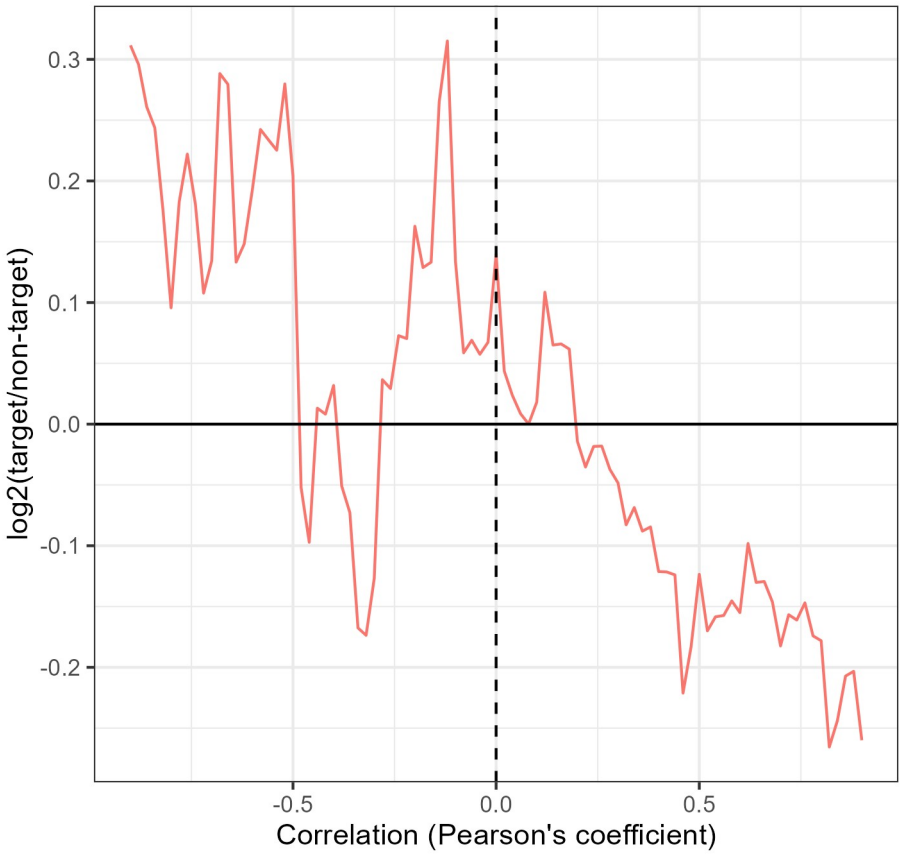


Figure 3. Co-expression of microRNAs and their targets within sRNAseq experiments. For all pairwise comparisons between a microRNA and a protein-coding gene, the x-axis gives the correlation of expression values in the human samples in Figure 2, and the y-axis shows the associate proportion of validated target pairs with respect to the total number of pairs in the bin. Each value in the x-axis is a bin of size +/- 0.2 around the x-axis value in a sliding window analysis with step size of 0.02.

To evaluate whether this methodology can extract useful information from clinical samples, we studied 24 samples from 12 patients with breast cancer sequenced by (Meerson et al. 2019). This study focused on microRNAs and only sRNAseq is provided. However, a clear advantage over many other studies is that samples are paired (tumour/non-tumour) for each individual patient, which can be used as a factor to control for variation between individuals. The expression levels of microRNAs and coding-genes were quantified from the sRNAseq datasets, and we performed differential gene expression analysis (see Methods) for both. The analysis of microRNAs reveals that dozens of microRNAs are differentially expressed between tumour and non-tumour samples correcting for batch (patient). More specifically, for a false discovery rate or 1% and a log2 fold-change difference of at least 1/-1, we identified 54 upregulated and 29 downregulated microRNAs in breast cancer samples. Among microRNAs, the most significant changes are for mir-144 (downregulated in tumours) and mir-429 (upregulated in tumours) (Supplementary Figures 1 and 2).

Importantly, the limited number of reads mapped to coding-genes in the sRNAseq was sufficient to permit differential gene expression analysis, and many coding-genes were found to be up- and downregulated in tumours (Figure 4 and Supplementary Figure 3). The functional annotation of upregulated genes from this analysis reveals an enrichment in functional categories (within the Biological Process domain in Gene Ontology) related with cell proliferation, as expected for cancer samples (Table 1). Also, the annotation to disease related databases (OMIM and Glad4U) consistently shows an enrichment in breast cancer related categories (Table 2). These results confirm that the differentially regulated genes identified from the sRNAseq experiments are consistent with those expected from breast cancer samples.

Table 1. Gene Ontology enrichment analysis for upregulated genes.

Gene Set	Description	Size	Expect	Ratio	q value
GO:0009888	Tissue Development	428	23.033	1.9537	0.008030
GO:0043588	Skin Development	65	3.4980	4.0023	0.008030
GO:0048856	Anatomical Structure Development	1154	62.102	1.4170	0.008030
GO:0042060	Wound Healing	160	8.6104	2.6712	0.008030
GO:0030855	Epithelial Cell Differentiation	151	8.1261	2.7073	0.008030
GO:0009611	Response to Wounding	186	10.010	2.4976	0.008030
GO:0032502	Developmental Process	1240	66.731	1.3787	0.008030
GO:0060429	Epithelium Development	271	14.584	2.1256	0.014430
GO:0007275	Multicellular Organism Development	1052	56.613	1.4131	0.021660
GO:0048731	System Development	961	51.716	1.4309	0.035011

Size: size of the category; Expect: expected number of genes from query in category; Ratio: ratio of observed over expected number of genes.

Table 2. Disease categories enrichment for upregulated genes in OMIM and GLAD4U

Gene Set	Description	Size	Expect	Ratio	q value
114480[1]	Breast Cancer	7	0.026	76.095	0.001574
176807[1]	Prostate Cancer	7	0.026	38.048	0.069562
601626[1]	Leukemia, Acute Myeloid	7	0.026	38.048	0.069562
PA446482[2]	Skin and Connective Tissue Diseases	113	6.396	4.8466	1.382e-11
PA445676[2]	Skin Diseases	121	6.849	4.5262	5.572e-11
PA443560[2]	Breast Neoplasms	140	7.925	3.7857	1.574e-8
PA443559[2]	Breast Diseases	128	7.245	3.7266	2.040e-7
PA445062[2]	Neoplasms	223	12.62	2.7728	1.578e-6
PA447242[2]	Epithelial Cancers	124	7.019	3.5618	1.813e-6
PA446646[2]	Carcinoma, Ductal, Breast	34	1.925	6.7549	2.223e-6
PA165108776[2]	Infiltrating Duct Carcinoma of Breast	34	1.925	6.7549	2.223e-6
PA445058[2]	Neoplasm Metastasis	157	8.887	3.1507	2.399e-6
PA443610[2]	Carcinoma	159	9.000	3.1111	2.901e-6

[1]OMIM; [2]GLAD4U. Headers as in table 1.

To analyse potential microRNA/transcripts interactions, we took all pairs of microRNAs and their targets, as we did with the previous analysis, and grouped them into four categories, depending on whether the transcript and/or the microRNA is up- or downregulated. We observed that, as expected, microRNA/coding-gene pairs with inverse changes in expression are enriched in targets

(Figure 5A). Interestingly, this enrichment is also observed (and even more notable) when we consider predicted targets based on canonical seeds only (Figure 5B). This indicates that the anticorrelation of microRNA-targets from sRNAseq experiments can potentially be used to identify target sites/confirm regulation.

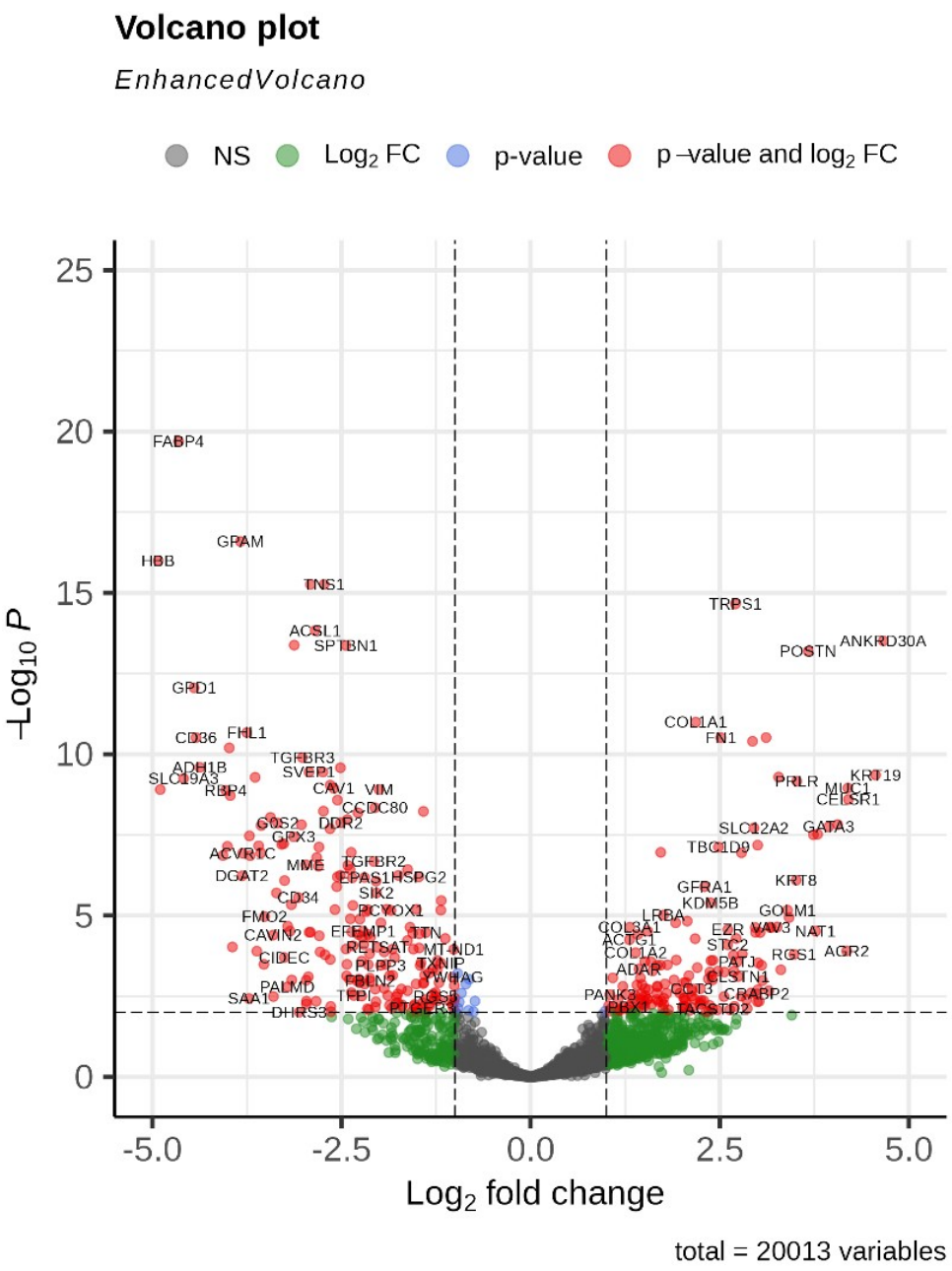


Figure 4. Differential gene expression of protein-coding genes from breast cancer paired sRNAseq experiments. Volcano plot representing the expression fold change on the x-axis of paired breast cancer samples (see main text) against the -log10 of the q value (FDR corrected p value)

generated during the differential gene expression analysis. Identified differentially expressed genes are shown in red.

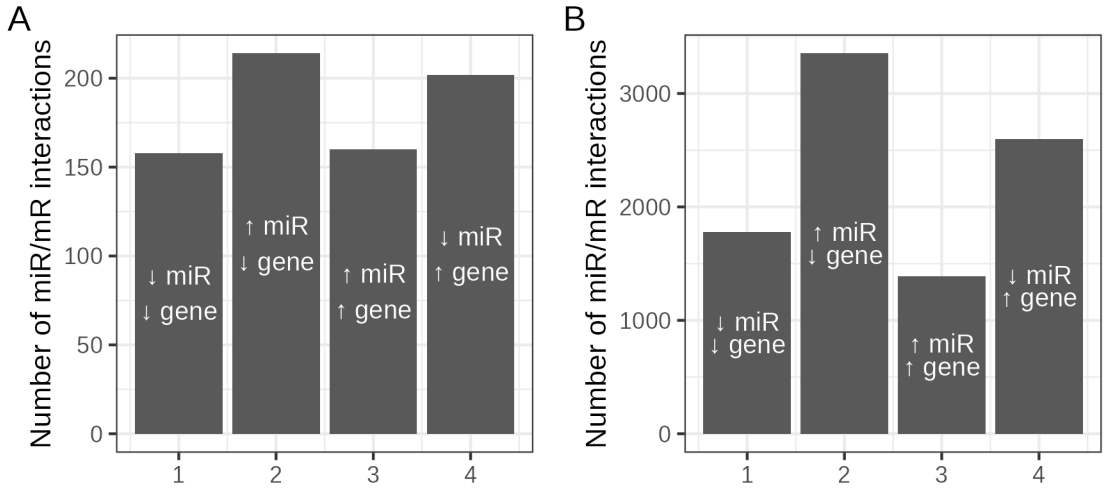


Figure 5 Comparison of the expression levels of gene transcripts from RNAseq and sRNAseq matches datasets. Each microRNA/protein-coding transcript pair was grouped in one of four categories: (i) both microRNA and transcript upregulated in breast cancer samples, (ii) both downregulated, (iii) upregulated microRNA but downregulated transcript, and (iv) the inverse. The bar plot shows how many instances are of each one of the four categories, showing enrichment of interactions in categories with inverse expression profiles between microRNA and transcript. This was performed for (A) validated target interactions and (B) for naïve computational target predictions.

In Figure 6 we specifically show the targets for two differentially expressed microRNAs: miR-429 and miR-144-5p. In the first instance, miR-429 was found to be upregulated in breast cancer samples compared to matched normal tissue, and this microRNA has known functional target sites in COL4A1, DLC1 and QKI (Figure 6A, 6B and 6C respectively). In the plots the relative expression values for the microRNA and the target genes in tumour and normal samples are provided and show that increased levels of miR-429 levels in tumour samples correlates with a decrease in the expression of all three targets. On the contrary, the downregulation of miR-144-5p in tumour samples is

associated with an upregulation of its target RUNX1 (Figure 6D). In conclusion, the analysis of small RNA sequencing experiments from clinical samples can be successfully used to simultaneously analyse changes in expression levels of microRNAs and their associated targets.

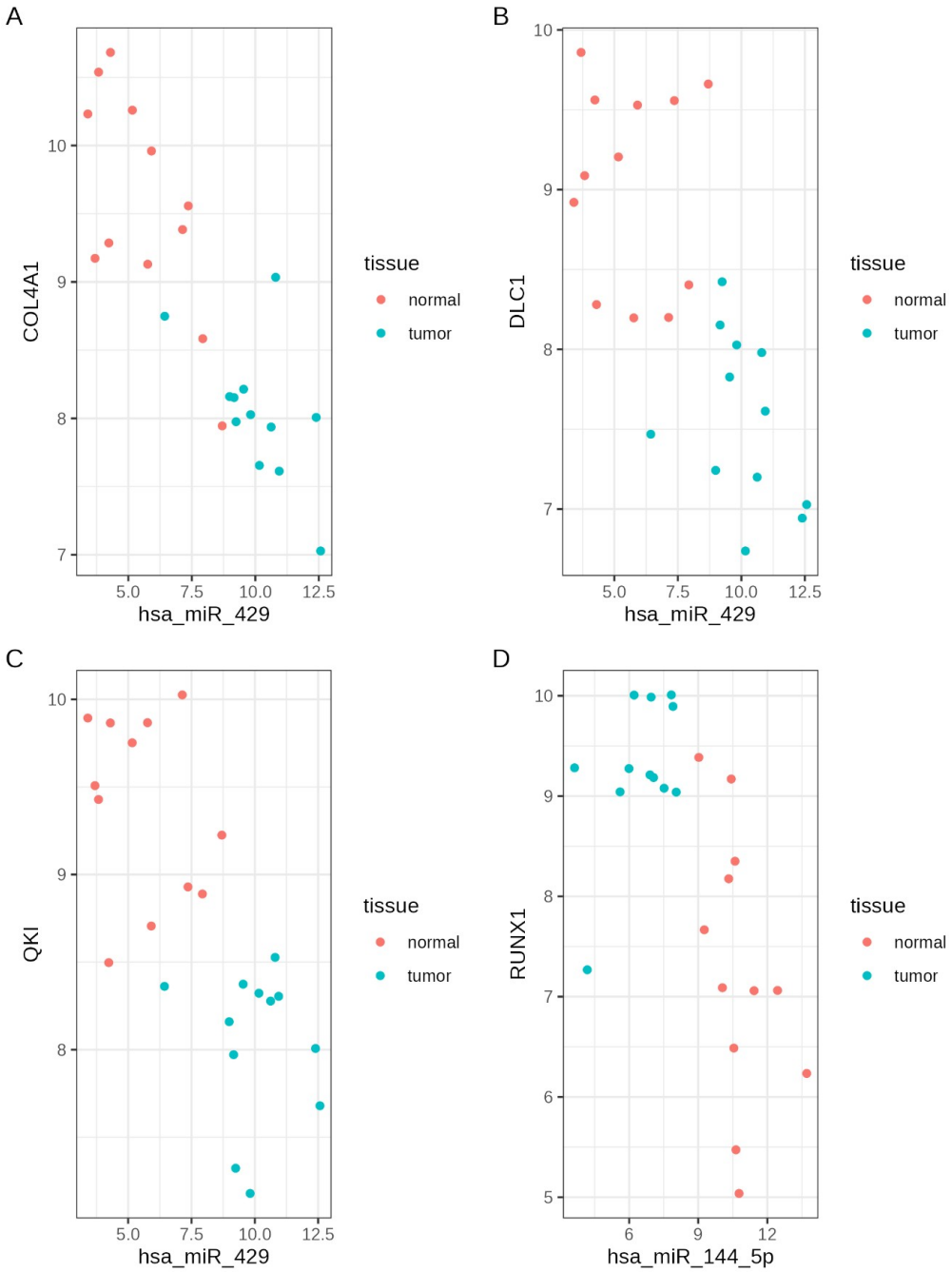


Figure 6. Comparison of microRNAs expression levels and their targets identified in breast cancer. The normalised expression level of a microRNA (x-axis) was plotted against the expression value of its target (y-axis). Normal tissue levels are in red and tumour levels in blue. In panels A to C

the expression levels or miR-429 is compared to those of (A) COL4A1, (B) DLC1 and (C) QKI. Panel (D) shows the comparison of miR-144-3p and RUNX1.

DISCUSSION

In this work we first compared the expression levels of protein-coding sequences from matched RNAseq and sRNAseq experiments, showing that sRNAseq experiments provide a sufficient amount of transcript fragments to be analysed and provide meaningful data. This allowed us to consider if the simultaneous analysis of small and long transcripts is possible solely from sRNAseq experiments. However, the correlation between the published gene expression data and the data that we generated from the sRNAseq between tissues. Specifically, the correlation was highest for the brain and cerebellum datasets while it was relatively low in kidney and heart; it is not clear to us why these discrepancies exist. One explanation could be that the brain and cerebellum samples from the commercial supplier were more homogeneous than those for the heart and kidney samples. Indeed, the brain samples were specifically extracted from the frontal cortex (Brawand et al. 2011). In any case, the expression levels of genes identified using both RNAseq and matched sRNAseq were consistently correlated.

We also analysed paired breast –cancer and normal samples for which only small RNA sequencing data was available, and we successfully identified differentially expressed microRNAs and coding-genes, as well as identifying potential microRNA-transcript interactions. The proposed workflow will not necessarily work with all sRNAseq datasets, as several factors may have an impact: library quality, sequencing depth or even the preservation of the source sample. These are considerations that researchers should take into account when designing future experiments. However, as a proof-of-principle, this work demonstrates that the simultaneous analysis of microRNAs and their targets from the same sRNAseq experiments is possible.

Although the computational prediction of microRNA targets has been relatively successful (see Introduction), experimental techniques have improved our ability to detect/confirm bona fide

targets. These techniques are diverse and include variations of immunoprecipitation and expression analysis (reviewed in Thomson et al. 2011). The joint analysis of gene expression of both microRNAs and their potential targets has been successfully used in the past, starting with the pioneering work by Huang et al. (2007). In that article, the authors describe a method that builds regulatory networks by combining the expression profile of matched miRNA/mRNA microarray experiments. Ever since, other studies using RNAseq/sRNAseq matched experiments have been used to identify microRNA targets (e.g. Jacobsen et al. 2013). A step further, in part to avoid unwanted effects due to differences in sample/library preparation, is the use of the same high-throughput expression experiment to study simultaneously both microRNAs and their potential targets. Two works from the Banfi lab exploited this idea in two different ways. First, considering that many intronic microRNAs have their expression linked to that of their host gene (Baskerville and Bartel 2005), they use gene expression microarrays to identify genes with anti-correlated expression with the host gene, as a proxy of intronic microRNA/target interaction (Gennarino et al. 2009). Second, they considered, again using microarray experiments, that targets of the same microRNA are co-expressed, and use this to identify microRNA targets (Gennarino et al. 2012). In this work we go a step further and study, as far as we are aware, for the first time, simultaneously the expression level of microRNAs and their potential targets from the same sRNAseq experiments.

In our analysis of breast cancer paired samples (tumour/non-tumour from the same patient) we identify (among others) two microRNAs differentially expressed whose targets have consistent changes in their expression level. Of special interest is the interaction between miR-429 and QKI. QK1 has been described as a tumour-suppressor in various cancers; low expression has been reported in lung (Zong et al. 2014; de Miguel et al. 2016), gastric (Bian et al. 2012) and breast (Cao et al. 2021) cancers. Our analysis supports that the downregulation of QKI that may influence the development of breast cancer, as our analysis demonstrated that it is expressed 2.4-fold lower in breast cancer samples, and this down-regulation appears to be a response to miR-429 overexpression. This interaction has been described for another evolutionarily related microRNA, miR-200b-3p, which shares the targeting properties of miR-429, with QKI has been described (Kim et al. 2019b).

Further work is necessary to better understand the factors that will allow for the systematic analysis of both microRNAs and their targets from the same sRNAseq experiments, but from the outcomes of this work, this approach is valid and will provide useful data to better understand microRNA-target gene regulatory networks.

METHODS

Datasets and databases.

The total human tissue RNA expression datasets were those from (Brawand et al. 2011) available at https://www.ebi.ac.uk/ena with accession PRJNA143627. The small RNA human tissue expression datasets are from (Meunier et al. 2013) with accession number PRJNA174234. The small RNA and total RNA experiments were performed by the same group. Breast cancer (patients) data was retrieved from (Meerson et al. 2019) with accession PRJNA494326. Experimentally validated microRNA targets are from miRTarBase [v 9.0] (Chou et al. 2018), and canonical microRNA targets were predicted using seedVicious [v1.3] (Marco 2018).

Analysis of coding gene and microRNA expression.

Adaptors were removed from reads with cutadapt [v3.7] (Martin 2011) and reads were mapped to the human genome hg38 with HISAT2 [v2.2.1] (Kim et al. 2019a). We then used featureCounts v2.0.2 (Liao et al. 2014) to count the number of reads in each feature. For human transcripts we used the annotation in Gencode [v43] (Frankish et al. 2019) and for microRNAs we used miRBase 22.1 (Kozomara et al. 2019). When comparing the expression profile of RNAseq and sRNAseq experiments for the same tissues/samples, we first used the Voom transformation on read counts (Law et al. 2014) using limma 3.52.2 (Ritchie et al. 2015). Differential gene expression analysis was done with DESeq2 (v1.36.0) (Love et al. 2014), including a patient term in the model (*expression ~ patient + tumour*). All statistical analyses and figures were done using R 4.2.1 (R Development Core Team

2004). Functional annotation was performed with WebGestalt 2024 (Elizarraras et al. 2024) setting a minimum number of elements per category to 5 and using the list of genes that passed the DESeq2 default filtering as the background list. The categories evaluated were Biological Process in Gene Ontology (Ashburner et al. 2000), OMIM (Hamosh et al. 2005) and GLAD4U (Jourquin et al. 2012). Volcano plots were drawn with the EnhancedVolcano R package (Blighe et al. 2024).

DATA ACCESS

All results generated and the scripts are available from GitHub at https://github.com/antoniomarco/CDS_from_sRNAseq

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

AA was funded by the Ministry of Science and Education, Republic of Azerbaijan. The authors acknowledge the use of the High-Performance Computing Facility (Ceres) and its associated support services at the University of Essex in the completion of this work.

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–9.

Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. 2008. The impact of microRNAs on protein output. *Nature* **455**: 64–71.

Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**: 241–247.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Bian Y, Wang L, Lu H, Yang G, Zhang Z, Fu H, Lu X, Wei M, Sun J, Zhao Q, et al. 2012. Downregulation of tumor suppressor QKI in gastric cancer and its implication in cancer prognosis. *Biochemical and Biophysical Research Communications* **422**: 187–193.

Blighe K, Rana S, Lewis M. 2024. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. *R package version 1220*.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, et al. 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci USA* **99**: 15524–15529.

Cao Y, Chu C, Li X, Gu S, Zou Q, Jin Y. 2021. RNA-binding protein QKI suppresses breast cancer via RASA1/MAPK signaling pathway. *Ann Transl Med* **9**: 104.

Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H, et al. 2018. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* **46**: D296–D302.

de Miguel FJ, Pajares MJ, Martínez-Terroba E, Ajona D, Morales X, Sharma RD, Pardo FJ, Rouzaut A, Rubio A, Montuenga LM, et al. 2016. A large-scale analysis of alternative splicing reveals a key role of QKI in lung cancer. *Molecular Oncology* **10**: 1437–1449.

Elizarraras JM, Liao Y, Shi Z, Zhu Q, Pico AR, Zhang B. 2024. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Research* gkae456.

Enright A, John B, Gaul U, Tuschl T, Sander C, Marks D. 2003. MicroRNA targets in Drosophila. *Genome Biology* **5**: R1.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Gennarino VA, D'Angelo G, Dharmalingam G, Fernandez S, Russolillo G, Sanges R, Mutarelli M, Belcastro V, Ballabio A, Verde P, et al. 2012. Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res* **22**: 1163–1172.

Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, Cutillo L, Ballabio A, Banfi S. 2009. MicroRNA target prediction by expression analysis of host genes. *Genome research* **19**: 490, 481.

Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**: 91–105.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**: D514–D517.

Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD. 2007. Using expression profiling data to identify human microRNA targets. *Nat Meth* **4**: 1045–1049.

Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C. 2013. Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* **20**: 1325–1332.

Jourquin J, Duncan D, Shi Z, Zhang B. 2012. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics* **13 Suppl 8**: S20.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019a. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Kim EJ, Kim JS, Lee S, Lee H, Yoon J-S, Hong JH, Chun SH, Sun DS, Won HS, Hong SA, et al. 2019b. QKI, a miR-200 target gene, suppresses epithelial-to-mesenchymal transition and tumor growth. *International Journal of Cancer* **145**: 1585–1595.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Research* **47**: D155–D162.

Lai EC, Tomancak P, Williams RW, Rubin GM. 2003. Computational identification of Drosophila microRNA genes. *Genome Biol* **4**: R42.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**: R29.

Lewis BP, Shih I, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of mammalian microRNA targets. *Cell* **115**: 787–798.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Maniatis T, Kee SG, Efstratiadis A, Kafatos FC. 1976. Amplification and characterization of a beta-globin gene synthesized in vitro. *Cell* **8**: 163–182.

Marco A. 2018. SeedVicious: Analysis of microRNA target and near-target sites. *PLOS ONE* **13**: e0195532.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Meerson A, Eliraz Y, Yehuda H, Knight B, Crundwell M, Ferguson D, Lee BP, Harries LW. 2019. Obesity impacts the regulation of miR-10b and its targets in primary breast tumors. *BMC Cancer* **19**: 86.

Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. *Genome Res* **23**: 34–45.

R Development Core Team. 2004. R: A language and environment for statistical computing. http://www.R-project.org (Accessed May 1, 2013).

Ratti M, Lampis A, Ghidini M, Salati M, Mirchev MB, Valeri N, Hahne JC. 2020. MicroRNAs (miRNAs) and Long Non-Coding RNAs (lncRNAs) as New Tools for Cancer Therapy: First Steps from Bench to Bedside. *Target Oncol* **15**: 261–278.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**: e47.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* **17**: 1850–1864.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–470.

Shang R, Lee S, Senavirathne G, Lai EC. 2023. microRNAs in action: biogenesis, function and regulation. *Nat Rev Genet* **24**: 816–833.

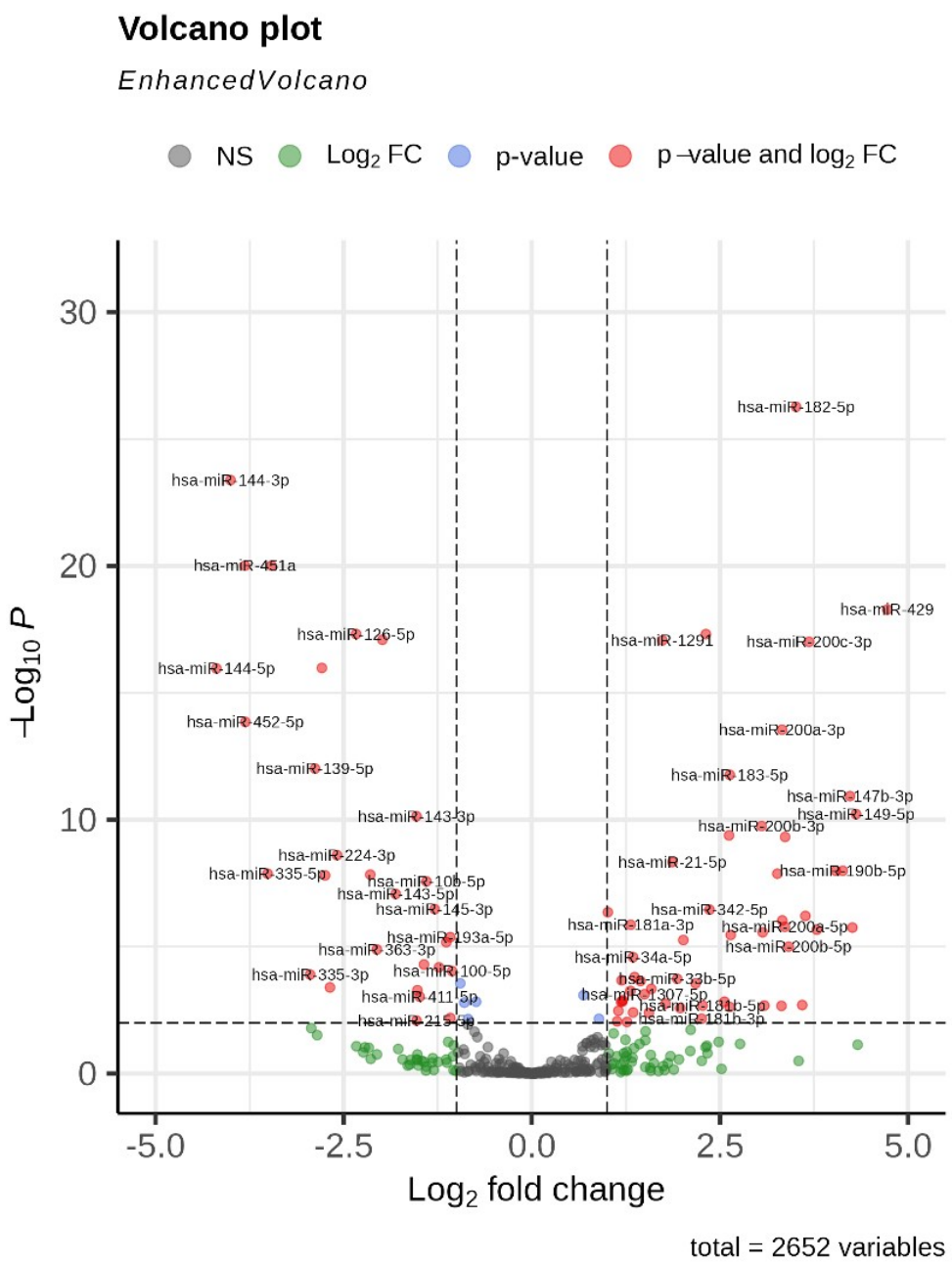
Thomson DW, Bracken CP, Goodall GJ. 2011. Experimental strategies for microRNA target identification. *Nucleic Acids Res* **39**: 6845–6853.

van Dongen S, Abreu-Goodger C, Enright AJ. 2008. Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods* **5**: 1023–1025.

Vannini I, Fanini F, Fabbri M. 2018. Emerging roles of microRNAs in cancer. *Curr Opin Genet Dev* **48**: 128–133.

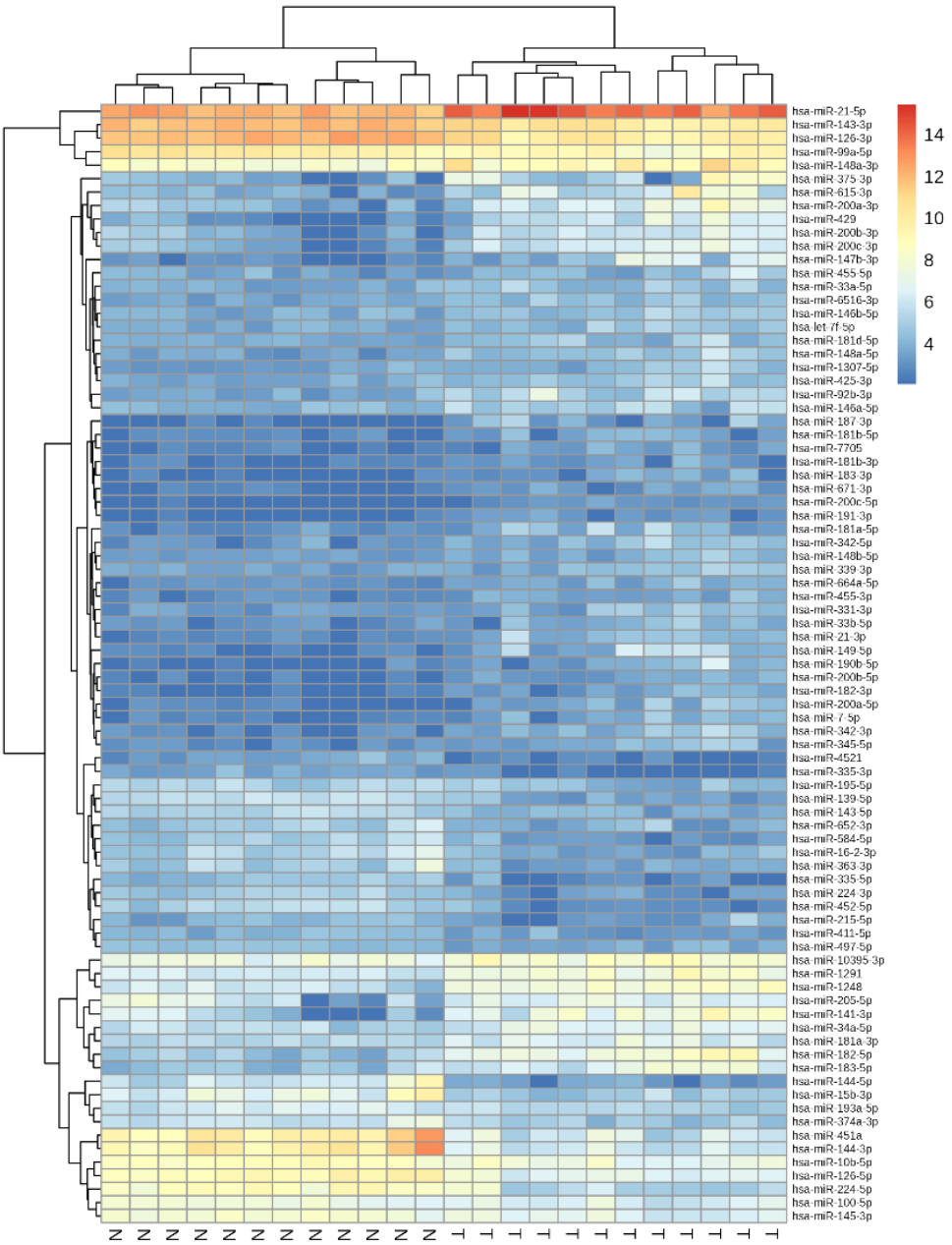
Zong F-Y, Fu X, Wei W-J, Luo Y-G, Heiner M, Cao L-J, Fang Z, Fang R, Lu D, Ji H, et al. 2014. The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *PLOS Genetics* **10**: e1004289.

Supplementary Figure 1



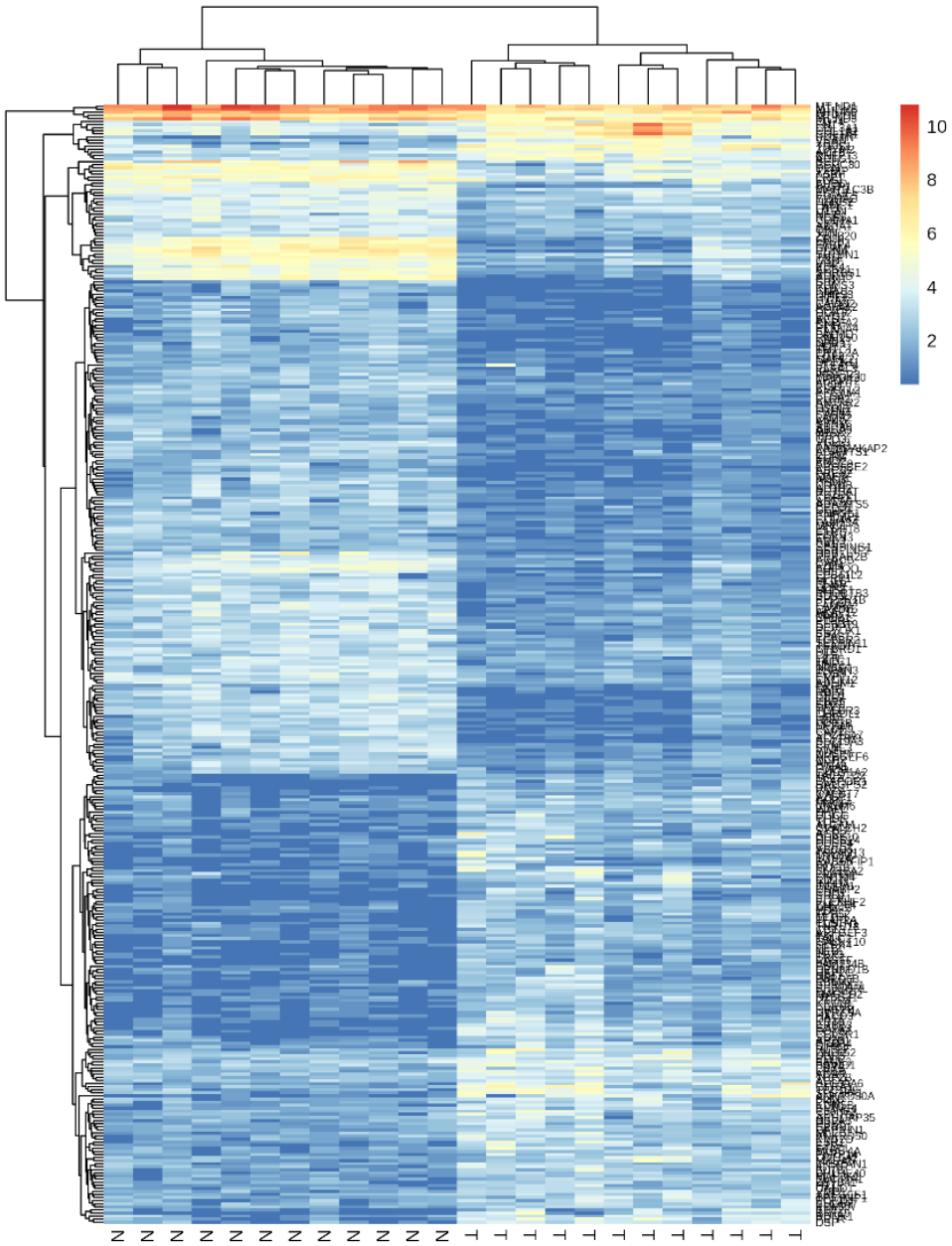
Supplementary Figure 1. Differential gene expression of microRNAs from breast cancer paired sRNAseq experiments. Volcano plot representing the expression fold change in the x-axis of paired breast cancer samples (see main text) against the -log10 of the q value (FDR corrected p value) generated during the differential gene expression analysis. Identified differentially expressed genes are shown in red.

Supplementary Figure 2



Supplementary Figure 2. Gene expression profile of microRNAs from breast cancer. Heatmap showing the expression levels of microRNAs from small RNA sequencing experiments in paired breast cancer samples.

Supplementary Figure 3



Supplementary Figure 3. Gene expression profile of protein-coding transcripts from breast cancer. Heatmap showing the expression levels of protein-coding genes from small RNA sequencing experiments in paired breast cancer samples.