miARma-Seq, a comprehensive pipeline for the simultaneous study and integration of miRNA and mRNA expression data

Eduardo Andrés-León^{1*} and Ana M. Rojas^{2*}

1. Bioinformatics Unit, Institute of Parasitology and Biomedicine "López-Neyra" (IPBLN), Spanish National Research Council (CSIC), 18016 Granada, Spain.

2. Computational Biology and Bioinformatics group. Andalusian Center for Developmental Biology (CABD/CSIC). Crt. Utrera S/N. 41013 Sevilla.

Keywords: miRNA, small RNASeq analysis, RNASeq analysis, miRNA-mRNA target prediction, Integrative analysis.

* Corresponding authors.

E-mail addresses: eandres@ipb.csic.es, a.rojas.m@csic.es

ABSTRACT

Messenger RNAs (mRNAs) fulfil specific biological roles in cells and, thus, their expression may be adapted to suit specific circumstances. This is in part achieved through selective gene transcription and post-transcriptional events, the regulation of which must be tightly integrated and controlled. To comprehensively study the coordinated effects of transcriptional and post-transcriptional regulatory elements, and to obtain coherent results, it is advisable to use different methodologies. Adequately integrating the data derived from these distinct methodologies then becomes critical to elucidating the relationships between the coordinated cellular effects assayed, particularly when applied to normal and disease states. Such integrated studies are likely to be particularly useful to identify markers suitable for early detection of diseases and to devise strategies for therapeutic interventions. Throughout this chapter, we will focus on the methods currently available to analyse mRNA and microRNA (miRNA) expression, paying special attention to the influence of miRNAs on mRNA metabolism. We will introduce miARma-Seq, a comprehensive pipeline that facilitates the simultaneous integration of mRNA and miRNA expression data. For illustrative purposes, we include a case study that incorporates data from RNASeq and small-RNASeq, detailing all the steps necessary to define the differential expression of both mRNA- and miRNA-encoding genes. Finally, we explore the possible regulatory relationships that drive significant and potentially relevant changes in mRNA and miRNA gene expression.

1. Introduction

The transcription of genomic DNA to generate RNAs is a process common to all living organisms. Although the primary sequence of genomic DNA is the same in all cell types of a given organism, the mRNAs expressed may differ according to the cell type and the developmental state of the organism. Differences may be provoked by selective gene transcription and/or post-transcriptional gene regulation, allowing cells to carry out specific biological activities. Not only genetic but also epigenetic mechanisms control gene expression and mRNA production. Known mechanisms include the actions of non-coding RNAs, e.g., long non-coding RNAs (lncRNAs), small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs) and microRNAs (miRNAs) [1-4].

The involvement of miRNAs in disease has particularly stimulated much interest. For instance, different studies have related altered miRNA expression to the initiation and progression of cancer [5, 6], Alzheimer's disease [7] or heart failure [8]. The relevance of miRNAs in the post-transcriptional control of gene expression has also been studied extensively and is largely dependent on biological context. Indeed, miRNAs are thought to regulate the transcripts of minimally 20–30% of human genes, and estimates as high as 92% have been proposed [9, 10].

miRNAs are small non-coding transcripts of around 22 nucleotides that regulate gene expression, mainly provoking the degradation and/or translational repression of mRNAs [11]. Under physiological conditions, miRNAs have been shown to participate in apoptosis, cell growth, cell proliferation and cell differentiation, both in prokaryotes and eukaryotes [12]. The best understood mechanism by which miRNAs elicit gene silencing involves binding of their so-called seed region (an evolutionary-conserved sequence at the 5'-end of the miRNA) to a complementary sequence that resides in the 3'-untranslated region (3'-UTR) of one or more target mRNAs [4, 13]. While miRNA binding to promoter regions has also been implicated [14, 15], we will focus here on the most accepted mechanism, which is direct binding to mRNAs.

The expression of microRNAs must be tightly balanced and controlled by cells given that any subtle alteration to the abundance of a miRNA can modify the expression levels of its target mRNAs, potentially leading to diseases [16, 17]. Depending on the cause of disease, profiling miRNA expression may be potentially more effective for the classification of the stage and/or subtype of some diseases than profiling

mRNA expression [18]. Indeed, models using both miRNA and mRNA expression profiles are likely to better classify disease stages or subtypes.

Recent technological developments have increased the accuracy with which the expression of a particular transcript can be measured [19]. For example, transcriptome-wide RNA sequencing (RNASeq) can generate millions of short cellular RNA sequences at single base resolution, enabling identification and quantitation of many of the expressed RNAs. This development has been accompanied with different computational pipelines that address distinct challenges in comparative data analyses. For example, RNASeq assays quantify relatively long RNAs [20], whereas miRNASeq does the same for miRNAs, piRNAs or siRNAs, and CircleSeq does so for circular RNAs [21].

Initial studies have focused on datasets for specific RNAs [22, 23], providing only part of the “picture” rather than integrating the multiple elements that are mechanistically connected in regulatory networks [24]. Despite the obvious advantages and enormous potential, concomitant analysis of mRNAs and miRNAs in the same sample is still not the norm and remains a challenge mainly because of inaccuracies in miRNA target predictions [25, 26]. Accordingly, it is recommended that miRNA–mRNA associations be predicted using different algorithms in order to minimize such inaccuracies [27].

Many tools have been developed to facilitate the study of miRNA expression profiles and those of their target transcripts. Software like CORNA [28], MMIA [29] and MAGIA [30, 31] aim to identify meaningful associations between mRNAs and miRNAs, accounting for inverse correlations in expression levels. The underlying hypothesis is that because miRNAs silence genes, one prediction is that miRNA overexpression would inhibit target mRNA abundance, assuming there are no other confounding factors.

Here, we review the latest version of miARma-Seq (miRNA-Seq And RNA-Seq Multiprocess Analysis), a comprehensive pipeline analysis suite designed to identify and quantify mRNAs, miRNAs and circRNAs, determining their differential expression while also making miRNA target predictions, and assessing their functional consequences [32]. This pipeline also permits data from miRNASeq and RNASeq to be integrated, making it capable of providing a set of differentially expressed miRNAs from miRNASeq samples and providing a statistical approximation of the regulated mRNAs from RNASeq samples.

Importantly, miARma-Seq integrates results based on miRNA–mRNA target interactions in the miRGate database [33, 34], using inverse correlations of normalized expression levels among mRNAs and miRNAs to filter out spurious relationships. The miARma-Seq pipeline is presented as a stand-alone tool, which is easy to install and flexible to use, bringing together well-established software in a single bundle. The tool is freely available at <https://bitbucket.org/cbbio/miarma>, with full documentation, and examples. Installation details for Unix, Windows and Docker systems can be accessed at <http://miarmaseq.com>.

2. Pipeline for a comprehensive miRNA and mRNA transcriptome analysis

In this section, we will describe the most common analyses performed on RNASeq samples. Emphasis will be placed on the possible workflows available to analyse mRNA samples (RNASeq) and small RNASeq using miARma-Seq. We introduce miRGate [33, 34] as the database that stores miRNA–mRNA associations and is integrated into miARma-Seq. This database predicts the interactions between mRNAs, and thus the gene from which they derive, and miRNAs in human, rat and mouse, using five different prediction algorithms: miRanda [35], Targetscan [36], RNAhybrid [37], Pita [38], and microTar [39]. Finally, we will describe how miARma-Seq integrates all this information in order to eliminate as many spurious or unreliable interactions as possible.

2.1. miARma-Seq

miARma-Seq is a tool designed to analyze most kinds of transcriptomic data. This pipeline aims to minimize difficulties faced by researchers when studying high throughput sequencing data. It involves the simultaneous analysis of both miRNA and mRNA expression data, offering a user-friendly, uncomplicated software that is easy to install, which minimizes the requirement of a third-party that often makes such processes tedious to the less-experienced scientists. The tool imposes no restrictions regarding the reference organism to be used as long as its genome has been sequenced and annotated. Its modular structure makes it flexible, enabling the user to start the analysis at different stages, and it can be executed on any standard laboratory computer or in a massive computer facility (benefiting from its parallelizing characteristics). The tool is very consistent as it implements most of the

standard software available that allows analysis of miRNA-Seq and RNA-Seq data to identify miRNAs, circRNAs or mRNAs. In addition, differential expression analysis, target prediction and functional analyses can be carried out, as well as prediction of novel miRNAs (see Figure 1).

The first step of analysis is quality assessment of the high throughput RNA sequencing data. FastQC [40] software has been included for this purpose, which produces a detailed report of the quality of sequencing reads as well as other valuable information. For instance, it is possible to visualize the adapter sequences that have been included in the output, which can be eliminated later. As such, miARma-Seq includes Cutadapt [41] and Reaper [42] for trimming and quality filtering purposes. In addition, miARma-Seq also includes Minion [42] to perform adapter sequence prediction and an in-house tool to remove a specific number of nucleotides from the 3' or 5' end that often contains poor quality reads.

Sequence alignment to the reference genome is always preceded by removal of low-quality sequencing reads. This is a crucial step since the number of the properly mapped reads for each genomic component (e.g. exons, protein coding sequences, non-coding DNA) is essential for the downstream analysis, and inclusion of low-quality sequences would decrease the accuracy of results. Sequence alignment varies according to the data being analysed: for miRNA expression studies, alignment can be performed in miARma-Seq with Bowtie1 [43], Bowtie2 [44] or both. By contrast, Bowtie1 is recommended for small RNASeq studies that usually involve single-end reads less than 50 nucleotides. In addition, miARma-Seq includes mirDeep2 [45] to predict novel miRNAs. For studies based on the analysis of mRNA samples, three different alignment tools are available: TopHat2 [46], STAR [47], and Hisat2 [48]. In all cases, miARma-Seq offers the option to build the genome indices that make it possible to analyse any sequenced genome.

Once sequences are aligned to the reference genome, they must be quantified. The process of counting reads is called read summarization. To perform this task, miARma-Seq includes FeatureCounts [49] that provides the total number of sequences, which in principle is proportional to the expression of RNA. Read summarization identifies which genes (coding or not) are expressed, and it is interesting to compare two or more samples (e.g., normal and disease samples) in order to obtain a list of genes that are differentially expressed [50, 51]. To perform statistical analysis, miARma-Seq offers two alternatives: edgeR [52] and NOISeq

[53]. In this step, we obtain a comprehensive report that includes boxplots, density plots of the normalized and non-normalized read counts, a multi-dimensional scaling (MDS) plot, a principal component analysis (PCA) plot, heat maps, as well as clustering plots. Finally, a list of genomic components (miRNAs, other transcripts, exons, genes) is deposited in a report file that includes component names, the expression changes under different conditions (fold change), median expression, and statistical values like p-values and false discovery rates (FDRs).

In summary, miARma-Seq performs all the steps necessary for miRNA and mRNA analysis, from introducing the raw read files to obtaining statistically significant results. Users simply have to download the miARma source code (Docker and Virtualbox images are also available) and decompress it. Once these steps are completed, they must fill in a configuration file with the information specific to the experiment, such as the type of sequencing (single-end or paired-end, and stranded or not), and indicate the directories where the sequencing files are stored and where the results are to be saved... etc.

2.2. miRGate

miRGate (<http://mirgate.bioinfo.cnio.es/>) is a database that stores predicted and experimentally confirmed miRNA–mRNA target pairs. To estimate all the possible targets in human, mouse or rat genomes, this resource uses a common and comprehensive sequence dataset rather than downloading pre-calculated data. For each genome, it includes the miRNA sequences available in miRBase [54] and a complete 3-UTR-sequence collection retrieved from Ensembl [55]. In addition, it contains information of experimentally validated targets to assess the consistency of the targets computed, offering valuable information to discriminate weak predictions.

miRGate was designed to circumvent the main obstacle in the field of target prediction, namely, the limited overlap of different target prediction methods (even for experimentally verified targets) due to the poor synchronization of sequences and gene catalogue versions. To the best of our knowledge, miRGate is the only tool available that overcomes this obstacle by using a common dataset.

Although there are many freely available methods that provide miRNA target predictions for standard gene sequences, only a few allow new predictions to be calculated from particular sequences. Therefore, we use the following five methods to identify novel miRNA target interactions: a) miRanda [35], which identifies target

genes on the basis of three parameters : sequence complementarity using a position-weighted local alignment algorithm, the free energy of RNA–RNA duplexes, and conservation of target sites in related genomes; b) TargetScan [36], which scores the interactions of seed matches, binding position and target sequence conservation in different species; c) RNAHybrid [37], which is based on favourable hybridization sites that prevent intramolecular duplexes; d) Pita [38], which recognizes sequences exhibiting fully complementarity to the seed sequence of each miRNA , and that estimates the positive thermodynamic energy of the bound and unbound duplex strand; and e) Microtar [39], which evaluates target sites based on an RNA duplex energy calculation.

To evaluate the putative predictions, miRGate collects information extracted from four different public databases that contain experimentally validated miRNA–mRNA targets: miRTarbase [56] and Tarbase [57], which both rely on text mining techniques to recognize validated targets; miRecords [58], which manually assesses interactions in automatically selected publications; and d) OncomirDB [59], which allocates validated miRNA–mRNA targets by manually curating scientific abstracts.

Information compiled by miRGate, including target sites, energy and conservation scores, miRNA and mRNA genomic coordinates, is available to the user as a web page. In addition, access can be gained through RESTful web services for those who need of a massive and an automatic query system [34].

2.3. Data integration

Biological data integration aims to enhance our knowledge of the relationships that govern cell pathologies, facilitating the development of novel strategies for the early detection and treatment of human diseases [60]. There are several data integration methods that can be applied [61], although “conceptual integration” and “statistical integration” are the two main types used in biomedicine. Conceptual integration refers to situations where multiple data sets are analysed individually and independent conclusions are drawn. While this is a valuable approach, it fails to reveal the relationships that might otherwise be inferred when both data sets are analysed in conjunction. Conversely, statistical integration is the most common procedure used to integrate transcriptomic data, whereby statistically meaningful associations can be extracted from different samples in distinct data sets.

Among the different types of statistical analyses available for combined data, we

mainly focus on those based on correlation methods, i.e., approaches that search for statistically meaningful correlations among elements in different datasets. Correlations have frequently been used to explore the relationships between transcriptomic data, metabolomic measurements and proteomic quantifications. There are many ways to assess the correlations between two sets of data, the most common being Pearson and Spearman's correlations for parametric and non-parametric data, respectively.

To summarize, miARma-Seq includes all the tools necessary to integrate and study transcriptomics data, highlighting the genes and miRNAs that are differentially expressed between at least two conditions of interest. In addition, it can search for possible target sites for dysregulated miRNAs through the use of miRGate, combining these results via statistical integration of the expression values. As such, it incorporates both a Pearson linear correlation (which should be suitable for most studies since it assumes normality of the data) and a Spearman's non-parametric approach for monotonic correlations.

3. Method to analyze and integrate transcriptomic data

As an example of how miARma-Seq can be used to analyze '-omics' expression data, we present here an analysis of small-RNASeq and RNASeq data from colon cancer samples obtained from the work by Lu et al [62]. A complete guide to reproduce all the results presented here can be found at: https://github.com/eandresleon/miRNA-mRNA_Integration.

3.1 Case study. Colorectal cancer (CRC) is still the leading cause of cancer-related death worldwide. Cetuximab and panitumumab are typical treatments for CRC that bind to the extracellular domain of the EGF receptor, enhancing its internalization and degradation. Indeed, when these antibodies are combined with chemotherapy, up to 72% response rates have been achieved [63]. Non-coding RNAs mediate cetuximab resistance in colon adenocarcinoma cancer [62], and although *de novo* and acquired drug resistance often arises, little is known about the mechanisms underlying such non-genetic resistance.

3.1.1 Human CRC samples. To enhance our knowledge of the mechanisms underlying resistance, CRC samples from the HCA-7 adenocarcinoma colon cancer cell line were treated with cetuximab for approximately four months to induce

resistance [62]. Three replicates of CRC samples and CRC-cetuximab-resistant (CRC-RC) cells were then analyzed by RNASeq and small-RNASeq, and the results were deposited at the NCBI Gene Expression Omnibus (GEO) repository, accession number GSE82236.

3.1.2 mRNA analysis. RNASeq results of three CRC and CRC-RC replicates were downloaded from GEO and analysed with miARma-Seq. The sequenced reads were inspected with FastQC [40] to identify sequencing errors, bad-quality reads or the possible accumulation of adapter sequences. As a result, an average quality higher than 25 was established as a quality-control threshold. The sequences were aligned using Hisat2 [48], and those with a quality score above 25 were counted and summarized based on their expression values using the GENCODE version 26 annotation file. EdgeR was then used to define genes differentially expressed among CRC and CRC-RC samples. Weakly expressed genes (those with a count-per-million (CPM) value less than 1) were removed from analysis, as recommended previously [64]. Finally, genes with a $|\log_2FC| \geq 1$ ($\log_2$ of Fold Change) and a $FDR \leq 0.01$ were considered as differentially expressed and saved in a tab-format file for subsequent analysis. In addition, miARma-Seq generated several plots to facilitate identifying genes that have undergone major changes in expression, which includes a boxplot comparing the average expression of a given gene in both conditions and a volcano plot highlighting the genes with the highest and most significant change in expression (See Figure 2).

3.1.3 miRNA analysis. Similarly, miRNASeq results of three CRC and CRC-RC replicates were analysed, and single-end sequenced reads were inspected using FastQC [40]. The reads were then pre-processed using Minion [42] to identify and to remove the accumulated adapter sequences, as those sequences were not defined by the authors. Trimmed 18 to 35-nucleotide-long reads that achieved the quality threshold of 25 were aligned using Bowtie1 [43], and quantified using FeatureCounts [49] and the microRNA annotation file from miRBase version 20. Weakly expressed miRNAs ($CPM < 1$) were removed, and differentially expressed microRNAs were acquired using edgeR [52]. As described above, miRNAs with a $|\log_2FC| \geq 1$ and a $FDR \leq 0.01$ were considered as differentially expressed (Figure 3; Volcano plot and the two boxplots indicating the most strongly upregulated and downregulated miRNAs).

3.1.4 Integration. miARma-Seq utilizes the miRGate query API to explore miRNA–mRNA interactions among differentially expressed genes and miRNAs [34]. To include all the expressed genes and miRNAs in the analysis, a threshold FDR of 1 can be specified. All the interactions deposited in miRGate involving miRNAs and genes of interest can be retrieved and stored in a file. In addition, Pearson or Spearman correlation can be carried out for the subsequent integration of the data. As a final result, we obtain a curated dataset that contains differentially expressed genes and miRNAs predicted to enter into a regulatory interaction, where the two elements display a statistically significant correlation.

4. Results and discussion

We have analyzed CRC and CRC-CR samples by miARma-Seq to identify novel, non-genetic mechanisms involved in the acquisition of resistance to cetuximab [62]. The study involved analyzing coding and noncoding transcriptomes, including miRNA and mRNA expression profiles, and obtained a total of 368 and 29 differentially expressed genes and miRNAs, respectively. We identified 165 upregulated and 203 downregulated genes ($|\log_2FC| \geq 1$ and $FDR \leq 0.01$). Consistent with the originally reported analysis [62], we determined that the genes exhibiting the lower fold-change in expression in CRC-CR compared to CRC were Olfactomedin 1 (OLFM1), the stress-associated endoplasmic reticulum protein family member 2 (SERP2), the Cellular retinoic acid-binding protein 1 (CRABP1), the Dickkopf WNT Signalling Pathway Inhibitor 1 (DKK1) and the Zinc Finger Protein 608 (ZNF608). On the contrary, the most upregulated genes were the Mir-100-Let-7a-2-Mir-125b-1 Cluster Host Gene (MIR100HG), the Dual Specificity Phosphatase 4 (DUSP4), the XIAP Associated Factor 1 (XAF1), the Basic Helix-Loop-Helix Family Member E41 (BHLHE41) and the Heparan Sulfate-Glucosamine 3-Sulfotransferase 5 (HS3ST5). Please see Figure 2a-c.

In detail, Olfactomedin 1 (OLFM1) is a member of the family of olfactomedin domain-containing proteins and it plays an important role in the development of neurogenic tissues. OLFM1 has been related to several diseases [65, 66] and dampening its expression appears to be associated with poor prognosis in colon cancer [67]. Indeed, there is a patent filed to study the potential of OLFM1 expression as a prognostic biomarker in colon cancer (ID: CN105779635A). Likewise, the stress-associated endoplasmic reticulum protein family member 2

(SERP2) has been identified as a prognostic indicator in colon cancer [68]. CRABP1 (Cellular retinoic acid-binding protein 1) plays an important role in differentiation and proliferation mediated by retinoic acid, and not only has it been associated with the malignancy of mesenchymal and neuroendocrine tumours [69] but it is also commonly downregulated in colon cancer [70]. Similarly, high DKK1 expression, a potent inhibitor of the Wnt pathway, has been negatively correlated with the progression of colon cancer [71]. However, a relationship between ZNF608 and colon cancer has yet to be established, although this zinc finger protein has been associated with liver and pancreatic cancer [72], and with resistance to cisplatin treatment in ovarian cancer [73].

The overexpression of MIR100HG, DUSP4, XAF1, BHLHE41 and HS3ST5 seems to underlie the initiation of numerous diseases, and the development and poor prognosis of colon cancer. For example, the miRNA-host gene MIR100HG has been linked to metastasis in this type of tumour [74], and DUSP4 (also known as mitogen-activated protein kinase phosphatase-2) is responsible for microsatellite instability and enhanced cell proliferation in colorectal cancer [75]. Similarly, the XIAP-associated factor 1 (XAF1) mediates apoptosis through an extracellular signal-regulated kinase pathway and is deregulated in colon cancer cells [76]. In tumour cells, BHLHE41 shows circadian expression and it inhibits the epithelial–mesenchymal transition (EMT), apoptosis and metastasis of sarcoma and hepatocellular carcinoma [77]. Significantly, normal tissues adjacent to colon carcinomas show elevated expression of BHLHE41 [78]. HS3ST5, which are heparan sulphate (HS) proteoglycans, are key components of the cell microenvironment, and the fine structure of their polysaccharide HS chains influences cell-cell interactions, adhesion, migration and signalling [79, 80]. Indeed, a relationship between HS3ST5 and metastasis was recently established in colon tumours [81].

By contrast, we obtained seven upregulated and 22 downregulated miRNAs from our small RNA analysis, similar to the seven and 24 miRNAs identified previously [62]. miR-100-5p, miR-125b-1-3p, miR-125b-5p and let-7a-2-3p were the most upregulated miRNAs in CRC-CR (Figure 3a). Conversely, miR-99a-5p, miR-34b-5p, miR-125b-2-3p, and the tumour suppressors miR-1, miR-145-5p and miR-143-3p were the most downregulated (Figures 3b, c). All the miRNAs we identified to be dysregulated in CRC-CR samples have been reported to be implicated in the

development and progression of tumours. For instance, the dysregulation of miR-100, miR125, let-7, miR-99a or miR-34b has already been demonstrated in colon cancer [82-86]. It is important to note that most of the reported results were obtained by comparing samples derived from healthy individual and colorectal cancer patients. Therefore, it is surprising to obtain similar results by comparing CRC and CRC-RC cancer cell lines.

Once the 29 miRNAs and the 368 differentially expressed genes had been identified, miARma-Seq carried out a statistical integration of these elements. Of a possible total of 10,672 correlations, 6,287 were statistically significant according to the Pearson correlation method ($p\text{-value} \leq 0.05$). Among the most upregulated genes in CRC-RC compared to CRC, a positive correlation with several of the most upregulated microRNAs stood out (see Table 1). As such, MIR100HG (a gene that shows 10.6 fold upregulation in CRC-RC samples) has an average positive R coefficient above 0.97 and a $p\text{-value} < 0.05$ for the four most upregulated microRNAs: miR-100-5p, let-7a-2-3p, miR-125b-5p and miR-125b-1-3p. This is due to the fact that miR100HG is a polycistronic long non-coding RNA that encodes for the four miRNAs seen to be upregulated: miR-100-5p, let-7a-2-3p, miR-125b-5p, and miR-125b-1-3p. Another interesting result is that the change in expression of MIR100HG negatively correlates with that of the three miRNAs whose expression was most strongly repressed (miR-99a-5p, miR-125b-2-3p and miR-34b- 5p), possibly indicating a second level of regulation (Table 1). To determine how a regulatory interaction between a microRNA and a gene could explain such negative correlations in expression, our tool assesses all the possible miRNA-mRNA interactions stored in the miRGate database. All the relevant information obtained is saved in an excel compatible file to be later studied in detail, including information regarding the prediction method, target site, etc.

Finally, to miARma-Seq performed an integration analysis between the differential expression values, the miRNA-mRNA target pairs, and the statistical correlation to corroborate the obtained results. The tool combined all the relevant information into an excel file, which contains the predicted pairs among the differentially expressed miRNAs and mRNAs that generated significant correlations in expression levels (Table 2, showing the results for the five most upregulated and downregulated genes in our data set). The report defines the putative targets of every miRNA and the correlation/anticorrelation values between the expression level of each pair

(Figure 4).

5. Conclusions

The integration of data from different sources helps improve our knowledge about the mechanisms that might underlie the pathophysiology of disease, allowing new strategies to be developed for the early detection and treatment of human diseases. An example of this is the recent article published by the consortium of the cancer genome atlas (TCGA) that integrated information from about 11,000 samples of 33 tumour types, using up to 4 ‘-omics’ technologies: exome sequencing, miRNA and mRNA transcriptome sequencing, and DNA methylation. This combination of data has resulted in an unprecedented molecular grouping of tumours [87].

The methodology necessary for such data integration includes the management of a large number of programs, making it a difficult task to perform. Therefore, it is advisable to create tools to conduct these analyses and minimize the technical challenges. In this work, we have described the miARma-Seq pipeline that can be used to combine transcriptome information from RNASeq and miRNASeq analyses. We have also described the statistical framework used in the context of previous studies that aimed to identify novel miRNA-mRNA interactions conserved in different tumour types [6]. This tool aims to help extract meaningful information from biological samples that help define the regulatory mRNA-miRNA interactions.

Acknowledgements

The authors would like to thank Lourdes Ortiz from the Institute of Parasitology and Biomedicine “López-Neyra” (IPBLN-CSIC) for her interesting comments and helpful suggestions.

References

- [1] M. Simonelig, piRNAs, master regulators of gene expression, *Cell Res* 24(7) (2014) 779-80.
- [2] C. Zhang, Y. Xu, H. Yang, Y. Xu, Q. Dong, S. Liu, T. Wu, Y. Zhang, Integrating gene and lncRNA expression to infer subpathway activity for tumor analyses, *Oncotarget* 8(67) (2017) 111433-111443.
- [3] Y. Abel, M. Rederstorff, Gene expression knockdown by transfection of siRNAs into mammalian cells, *Methods Mol Biol* 1296 (2015) 199-202.
- [4] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell* 116(2) (2004) 281-97.
- [5] N. Matamala, M.T. Vargas, R. Gonzalez-Campora, J.I. Arias, P. Menendez, E. Andres-Leon, K. Yanowsky, A. Llana-Folgueras, R. Minambres, B. Martinez-Delgado, J. Benitez, MicroRNA deregulation in triple negative breast cancer reveals a role of miR-498 in regulating BRCA1 expression, *Oncotarget* 7(15) (2016) 20068-79.
- [6] E. Andres-Leon, I. Cases, S. Alonso, A.M. Rojas, Novel miRNA-mRNA interactions conserved in essential cancer pathways, *Sci Rep* 7 (2017) 46101.
- [7] S.I. Negoita, D. Sandesc, A.F. Rogobete, M. Dutu, O.H. Bedreag, M. Papurica, M.F. Ercisli, S.E. Popovici, R. Dumache, M. Sandesc, A. Dinu, A.M. Sas, D. Serban, D. Corneci, miRNAs Expressions and Interaction with Biological Systems in Patients with Alzheimer`s Disease. Using miRNAs as a Diagnosis and Prognosis Biomarker, *Clin Lab* 63(9) (2017) 1315-1321.
- [8] T. Thum, P. Galuppo, C. Wolf, J. Fiedler, S. Kneitz, L.W. van Laake, P.A. Doevendans, C.L. Mummery, J. Borlak, A. Haverich, C. Gross, S. Engelhardt, G. Ertl, J. Bauersachs, MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure, *Circulation* 116(3) (2007) 258-67.
- [9] L.P. Lim, N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, J.M. Johnson, Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs, *Nature* 433(7027) (2005) 769-73.
- [10] B.P. Lewis, C.B. Burge, D.P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *Cell* 120(1) (2005) 15-20.

- [11] N. Morozova, A. Zinovyev, N. Nonne, L.L. Pritchard, A.N. Gorban, A. Harel-Bellan, Kinetic signatures of microRNA modes of action, *RNA* 18(9) (2012) 1635-55.
- [12] S.A. Shabalina, E.V. Koonin, Origins and evolution of eukaryotic RNA interference, *Trends Ecol Evol* 23(10) (2008) 578-87.
- [13] A. Grimson, K.K. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, D.P. Bartel, MicroRNA targeting specificity in mammals: determinants beyond seed pairing, *Mol Cell* 27(1) (2007) 91-105.
- [14] R.F. Place, L.C. Li, D. Pookot, E.J. Noonan, R. Dahiya, MicroRNA-373 induces expression of genes with complementary promoter sequences, *Proc Natl Acad Sci U S A* 105(5) (2008) 1608-13.
- [15] J.R. Lytle, T.A. Yario, J.A. Steitz, 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) (2007) 9667-72.
- [16] C.D. Johnson, A. Esquela-Kerscher, G. Stefani, M. Byrom, K. Kelnar, D. Ovcharenko, M. Wilson, X. Wang, J. Shelton, J. Shingara, L. Chin, D. Brown, F.J. Slack, The let-7 microRNA represses cell proliferation pathways in human cells, *Cancer Res* 67(16) (2007) 7713-22.
- [17] M.L. Si, S. Zhu, H. Wu, Z. Lu, F. Wu, Y.Y. Mo, miR-21-mediated tumor growth, *Oncogene* 26(19) (2007) 2799-803.
- [18] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H.R. Horvitz, T.R. Golub, MicroRNA expression profiles classify human cancers, *Nature* 435(7043) (2005) 834-8.
- [19] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat Methods* 5(7) (2008) 621-8.
- [20] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev Genet* 10(1) (2009) 57-63.
- [21] W.R. Jeck, N.E. Sharpless, Detecting and characterizing circular RNAs, *Nat Biotechnol* 32(5) (2014) 453-61.
- [22] M. Tanic, E. Andres, S.M. Rodriguez-Pinilla, I. Marquez-Rodas, M. Cebollero-Presmanes, V. Fernandez, A. Osorio, J. Benitez, B. Martinez-Delgado, MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours, *Br J Cancer* 109(10) (2013) 2724-34.

- [23] C. Sotiriou, M.J. Piccart, Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?, *Nat Rev Cancer* 7(7) (2007) 545-53.
- [24] R. Cavill, D. Jennen, J. Kleinjans, J.J. Briede, Transcriptomic and metabolomic data integration, *Brief Bioinform* 17(5) (2016) 891-901.
- [25] G. Wilk, R. Braun, Integrative analysis reveals disrupted pathways regulated by microRNAs in cancer, *Nucleic Acids Res* 46(3) (2018) 1089-1101.
- [26] C.I. Oduor, Y. Kaymaz, K. Chelimo, J.A. Otieno, J.M. Ong'echa, A.M. Moormann, J.A. Bailey, Integrative microRNA and mRNA deep-sequencing expression profiling in endemic Burkitt lymphoma, *BMC Cancer* 17(1) (2017) 761.
- [27] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szczesniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, *Genome Biol* 17 (2016) 13.
- [28] X. Wu, M. Watson, CORNA: testing gene lists for regulation by microRNAs, *Bioinformatics* 25(6) (2009) 832-3.
- [29] S. Nam, M. Li, K. Choi, C. Balch, S. Kim, K.P. Nephew, MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression, *Nucleic Acids Res* 37(Web Server issue) (2009) W356-62.
- [30] G. Sales, A. Coppe, A. Bisognin, M. Biasiolo, S. Bortoluzzi, C. Romualdi, MAGIA, a web-based tool for miRNA and Genes Integrated Analysis, *Nucleic Acids Res* 38(Web Server issue) (2010) W352-9.
- [31] A. Bisognin, G. Sales, A. Coppe, S. Bortoluzzi, C. Romualdi, MAGIA(2): from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update), *Nucleic Acids Res* 40(Web Server issue) (2012) W13-21.
- [32] E. Andres-Leon, R. Nunez-Torres, A.M. Rojas, miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis, *Sci Rep* 6 (2016) 25749.
- [33] E. Andres-Leon, D. Gonzalez Pena, G. Gomez-Lopez, D.G. Pisano, miRGate: a curated database of human, mouse and rat miRNA-mRNA targets, *Database* (Oxford) 2015 (2015) bav035.
- [34] E. Andres-Leon, G. Gomez-Lopez, D.G. Pisano, Prediction of miRNA-mRNA Interactions Using miRGate, *Methods Mol Biol* 1580 (2017) 225-237.

- [35] D. Betel, A. Koppal, P. Agius, C. Sander, C. Leslie, Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites, *Genome Biol* 11(8) (2010) R90.
- [36] R.C. Friedman, K.K. Farh, C.B. Burge, D.P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs, *Genome Res* 19(1) (2009) 92-105.
- [37] J. Kruger, M. Rehmsmeier, RNAhybrid: microRNA target prediction easy, fast and flexible, *Nucleic Acids Res* 34(Web Server issue) (2006) W451-4.
- [38] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, E. Segal, The role of site accessibility in microRNA target recognition, *Nat Genet* 39(10) (2007) 1278-84.
- [39] R. Thadani, M.T. Tammi, MicroTar: predicting microRNA targets from RNA duplexes, *BMC Bioinformatics* 7 Suppl 5 (2006) S20.
- [40] S. Anders, FastQC: a quality control tool for high throughput sequence data, Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).
- [41] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, 2011 17(1) (2011).
- [42] M.P. Davis, S. van Dongen, C. Abreu-Goodger, N. Bartonicek, A.J. Enright, Kraken: a set of tools for quality control and analysis of high-throughput sequence data, *Methods* 63(1) (2013) 41-9.
- [43] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol* 10(3) (2009) R25.
- [44] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, *Nat Methods* 9(4) (2012) 357-9.
- [45] M.R. Friedlander, S.D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, *Nucleic Acids Res* 40(1) (2012) 37-52.
- [46] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biol* 14(4) (2013) R36.
- [47] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics* 29(1) (2013) 15-21.

- [48] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, *Nat Methods* 12(4) (2015) 357-60.
- [49] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics* 30(7) (2014) 923-30.
- [50] V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G.H. Eiriksdomtir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K.P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H.G. Gislason, T. Stefansson, B.G. Leifsson, U. Thorsteinsdottir, J.R. Lamb, J.R. Gulcher, M.L. Reitman, A. Kong, E.E. Schadt, K. Stefansson, Genetics of gene expression and its effect on disease, *Nature* 452(7186) (2008) 423-8.
- [51] J.W. Ho, M. Stefani, C.G. dos Remedios, M.A. Charleston, Differential variability analysis of gene expression and its application to human diseases, *Bioinformatics* 24(13) (2008) i390-8.
- [52] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics* 26(1) (2010) 139-40.
- [53] S. Tarazona, F. Garcia-Alcalde, J. Dopazo, A. Ferrer, A. Conesa, Differential expression in RNA-seq: a matter of depth, *Genome Res* 21(12) (2011) 2213-23.
- [54] A. Kozomara, S. Griffiths-Jones, miRBase: annotating high confidence microRNAs using deep sequencing data, *Nucleic Acids Res* 42(Database issue) (2014) D68-73.
- [55] A. Yates, W. Akanni, M.R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C.G. Giron, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F.J. Martin, T. Maurel, W. McLaren, D.N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H.S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S.P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S.J. Trevanion, F. Cunningham, B.L. Aken, D.R. Zerbino, P. Flicek, Ensembl 2016, *Nucleic Acids Res* 44(D1) (2016) D710-6.
- [56] C.H. Chou, N.W. Chang, S. Shrestha, S.D. Hsu, Y.L. Lin, W.H. Lee, C.D. Yang, H.C. Hong, T.Y. Wei, S.J. Tu, T.R. Tsai, S.Y. Ho, T.Y. Jian, H.Y. Wu, P.R. Chen,

N.C. Lin, H.T. Huang, T.L. Yang, C.Y. Pai, C.S. Tai, W.L. Chen, C.Y. Huang, C.C. Liu, S.L. Weng, K.W. Liao, W.L. Hsu, H.D. Huang, miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database, *Nucleic Acids Res* 44(D1) (2016) D239-47.

[57] T. Vergoulis, I.S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, A.G. Hatzigeorgiou, TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support, *Nucleic Acids Res* 40(Database issue) (2012) D222-9.

[58] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, T. Li, miRecords: an integrated resource for microRNA-target interactions, *Nucleic Acids Res* 37(Database issue) (2009) D105-10.

[59] D. Wang, J. Gu, T. Wang, Z. Ding, OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs, *Bioinformatics* 30(15) (2014) 2237-8.

[60] Y.V. Sun, Y.J. Hu, Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases, *Adv Genet* 93 (2016) 147-90.

[61] T. Ebbels, R. Cavill, *Bioinformatic methods in NMR-based metabolic profiling*, 2009.

[62] Y. Lu, X. Zhao, Q. Liu, C. Li, R. Graves-Deal, Z. Cao, B. Singh, J.L. Franklin, J. Wang, H. Hu, T. Wei, M. Yang, T.J. Yeatman, E. Lee, K. Saito-Diaz, S. Hinger, J.G. Patton, C.H. Chung, S. Emmrich, J.H. Klusmann, D. Fan, R.J. Coffey, lncRNA MIR100HG-derived miR-100 and miR-125b mediate cetuximab resistance via Wnt/beta-catenin signaling, *Nat Med* 23(11) (2017) 1331-1341.

[63] S. Stintzing, D.P. Modest, L. Rossius, M.M. Lerch, L.F. von Weikersthal, T. Decker, A. Kiani, U. Vehling-Kaiser, S.E. Al-Batran, T. Heintges, C. Lerchenmuller, C. Kahl, G. Seipelt, F. Kullmann, M. Stauch, W. Scheithauer, S. Held, C. Giessen-Jung, M. Moehler, A. Jagenburg, T. Kirchner, A. Jung, V. Heinemann, F.-. investigators, FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab for metastatic colorectal cancer (FIRE-3): a post-hoc analysis of tumour dynamics in the final RAS wild-type subgroup of this randomised open-label phase 3 trial, *Lancet Oncol* 17(10) (2016) 1426-1434.

[64] S. Anders, D.J. McCarthy, Y. Chen, M. Okoniewski, G.K. Smyth, W. Huber, M.D. Robinson, Count-based differential expression analysis of RNA sequencing data using R and Bioconductor, *Nature protocols* 8(9) (2013) 1765-86.

- [65] R.R. Anholt, Olfactomedin proteins: central players in development and disease, *Front Cell Dev Biol* 2 (2014) 6.
- [66] S.P. Kodithuwakku, R.T. Pang, E.H. Ng, A.N. Cheung, A.W. Horne, P.C. Ho, W.S. Yeung, K.F. Lee, Wnt activation downregulates olfactomedin-1 in Fallopian tubal epithelial cells: a microenvironment predisposed to tubal ectopic pregnancy, *Lab Invest* 92(2) (2012) 256-64.
- [67] W. Shi, Z. Ye, L. Zhuang, Y. Li, W. Shuai, Z. Zuo, X. Mao, R. Liu, J. Wu, S. Chen, W. Huang, Olfactomedin 1 negatively regulates NF-kappaB signalling and suppresses the growth and metastasis of colorectal cancer cells, *J Pathol* 240(3) (2016) 352-365.
- [68] X. Zhai, Q. Xue, Q. Liu, Y. Guo, Z. Chen, Colon cancer recurrence associated genes revealed by WGCNA coexpression network analysis, *Mol Med Rep* 16(5) (2017) 6499-6505.
- [69] Y. Kainov, I. Favorskaya, V. Delektorskaya, G. Chemeris, A. Komelkov, A. Zhuravskaya, L. Trukhanova, E. Zueva, B. Tavitian, N. Dyakova, I. Zborovskaya, E. Tchevkina, CRABP1 provides high malignancy of transformed mesenchymal cells and contributes to the pathogenesis of mesenchymal and neuroendocrine tumors, *Cell Cycle* 13(10) (2014) 1530-9.
- [70] G.E. Lind, K. Kleivi, G.I. Meling, M.R. Teixeira, E. Thiis-Evensen, T.O. Rognum, R.A. Lothe, ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis, *Cell Oncol* 28(5-6) (2006) 259-72.
- [71] L. Qi, B. Sun, Z. Liu, H. Li, J. Gao, X. Leng, Dickkopf-1 inhibits epithelial-mesenchymal transition of colon cancer cells and contributes to colon cancer suppression, *Cancer Sci* 103(4) (2012) 828-35.
- [72] K.A. O'Donnell, V.W. Keng, B. York, E.L. Reineke, D. Seo, D. Fan, K.A. Silverstein, C.T. Schrum, W.R. Xie, L. Mularoni, S.J. Wheelan, M.S. Torbenson, B.W. O'Malley, D.A. Largaespada, J.D. Boeke, A Sleeping Beauty mutagenesis screen reveals a tumor suppressor role for Ncoa2/Src-2 in liver cancer, *Proc Natl Acad Sci U S A* 109(21) (2012) E1377-86.
- [73] P.M. Baez-Vega, I.M. Echevarria Vargas, F. Valiyeva, J. Encarnacion-Rosado, A. Roman, J. Flores, M.J. Marcos-Martinez, P.E. Vivas-Mejia, Targeting miR-21-3p inhibits proliferation and invasion of ovarian cancer cells, *Oncotarget* 7(24) (2016) 36321-36337.

- [74] M. Rokavec, D. Horst, H. Hermeking, Cellular Model of Colon Cancer Progression Reveals Signatures of mRNAs, miRNA, lncRNAs, and Epigenetic Modifications Associated with Metastasis, *Cancer Res* 77(8) (2017) 1854-1867.
- [75] B. Groschl, M. Bettstetter, C. Giedl, M. Woenckhaus, T. Edmonston, F. Hofstadter, W. Dietmaier, Expression of the MAP kinase phosphatase DUSP4 is associated with microsatellite instability in colorectal cancer (CRC) and causes increased cell proliferation, *Int J Cancer* 132(7) (2013) 1537-46.
- [76] Y. Sun, L. Qiao, H.H. Xia, M.C. Lin, B. Zou, Y. Yuan, S. Zhu, Q. Gu, T.K. Cheung, H.F. Kung, M.F. Yuen, A.O. Chan, B.C. Wong, Regulation of XAF1 expression in human colon cancer cell by interferon beta: activation by the transcription regulator STAT1, *Cancer Lett* 260(1-2) (2008) 62-71.
- [77] F. Sato, U.K. Bhawal, T. Yoshimura, Y. Muragaki, DEC1 and DEC2 Crosstalk between Circadian Rhythm and Tumor Progression, *J Cancer* 7(2) (2016) 153-9.
- [78] K. Yamada, K. Miyamoto, Basic helix-loop-helix transcription factors, BHLHB2 and BHLHB3; their gene expressions are regulated by multiple extracellular stimuli, *Front Biosci* 10 (2005) 3151-71.
- [79] A.V. Suhovskih, N.V. Domanitskaya, A.Y. Tsidulko, T.Y. Prudnikova, V.I. Kashuba, E.V. Grigorieva, Tissue-specificity of heparan sulfate biosynthetic machinery in cancer, *Cell Adh Migr* 9(6) (2015) 452-9.
- [80] J.L. Dreyfuss, C.V. Regatieri, T.R. Jarrouge, R.P. Cavalheiro, L.O. Sampaio, H.B. Nader, Heparan sulfate proteoglycans: structure, protein interactions and cell signaling, *An Acad Bras Cienc* 81(3) (2009) 409-29.
- [81] I. Fernandez-Vega, O. Garcia-Suarez, B. Garcia, A. Crespo, A. Astudillo, L.M. Quiros, Heparan sulfate proteoglycans undergo differential expression alterations in right sided colorectal cancer, depending on their metastatic character, *BMC Cancer* 15 (2015) 742.
- [82] Y. Akao, Y. Nakagawa, T. Naoe, let-7 microRNA functions as a potential growth suppressor in human colon cancer cells, *Biol Pharm Bull* 29(5) (2006) 903-6.
- [83] S. Molina-Pinelo, A. Carnero, F. Rivera, P. Estevez-Garcia, J.M. Bozada, M.L. Limon, M. Benavent, J. Gomez, M.D. Pastor, M. Chaves, R. Suarez, L. Paz-Ares, F. de la Portilla, A. Carranza-Carranza, I. Sevilla, L. Vicioso, R. Garcia-Carbonero, MiR-107 and miR-99a-3p predict chemotherapy response in patients with advanced colorectal cancer, *BMC Cancer* 14 (2014) 656.

- [84] N. Nishida, T. Yokobori, K. Mimori, T. Sudo, F. Tanaka, K. Shibata, H. Ishii, Y. Doki, H. Kuwano, M. Mori, MicroRNA miR-125b is a prognostic marker in human colorectal cancer, *Int J Oncol* 38(5) (2011) 1437-43.
- [85] S. Roy, E. Levi, A.P. Majumdar, F.H. Sarkar, Expression of miR-34 is lost in colon cancer which can be re-expressed by a novel agent CDF, *J Hematol Oncol* 5 (2012) 58.
- [86] M.K. Zhou, X.J. Liu, Z.G. Zhao, Y.M. Cheng, MicroRNA-100 functions as a tumor suppressor by inhibiting Lgr5 expression in colon cancer cells, *Mol Med Rep* 11(4) (2015) 2947-52.
- [87] K.A. Hoadley, C. Yau, T. Hinoue, D.M. Wolf, A.J. Lazar, E. Drill, R. Shen, A.M. Taylor, A.D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C.K. Wong, M. Wiznerowicz, F. Sanchez-Vega, A.G. Robertson, B.G. Schneider, M.S. Lawrence, H. Noushmehr, T.M. Malta, N. Cancer Genome Atlas, J.M. Stuart, C.C. Benz, P.W. Laird, Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer, *Cell* 173(2) (2018) 291-304 e6.

Tables and figures

Table 1

List of the most highly positive and negative correlated miRNAs for the five most upregulated (left) and downregulated (right) genes. Pearson's R coefficient values and P-values obtained from genes and miRNAs expression values are given.

miRNA	Gene	R	P	miRNA	Gene	R	P
miR-302d-3p	BHLHE41	-0.975	0.001	miR-4286	CRABP1	-0.903	0.014
miR-125b-2-3p	BHLHE41	-0.967	0.002	miR-152-3p	DKK1	0.930	0.007
let-7a-2-3p	BHLHE41	0.978	0.001	miR-4286	OLFM1	-0.973	0.001
miR-34b-5p	DUSP4	-0.972	0.001	miR-100-5p	OLFM1	-0.948	0.004
miR-34c-5p	DUSP4	-0.922	0.009	miR-125b-5p	OLFM1	-0.905	0.013
miR-125b-5p	DUSP4	0.977	0.001	miR-152-3p	OLFM1	0.948	0.004
miR-302d-3p	HS3ST5	-0.939	0.005	miR-34b-5p	OLFM1	0.952	0.003
miR-302b-3p	HS3ST5	-0.925	0.008	miR-99a-5p	OLFM1	0.963	0.002
miR-302a-3p	HS3ST5	-0.915	0.010	miR-100-5p	SERP2	-0.963	0.002
miR-125b-1-3p	MIR100HG	0.938	0.006	let-7a-2-3p	SERP2	-0.958	0.003
miR-125b-5p	MIR100HG	0.957	0.003	miR-375	SERP2	-0.942	0.005
let-7a-2-3p	MIR100HG	0.978	0.001	miR-194-5p	SERP2	0.973	0.001
miR-100-5p	MIR100HG	0.992	9.8E-5	miR-152-3p	SERP2	0.991	1.1E-4
miR-125b-2-3p	MIR100HG	-0.939	0.005	miR-302a-3p	ZNF608	0.911	0.012
miR-34b-5p	MIR100HG	-0.983	0.000	miR-302b-3p	ZNF608	0.916	0.010
miR-99a-5p	MIR100HG	-0.916	0.019	miR-34b-5p	ZNF608	0.949	0.004
miR-302b-3p	XAF1	-0.985	0.000	miR-125b-2-3p	ZNF608	0.968	0.002
miR-302d-3p	XAF1	-0.975	0.001	miR-302d-3p	ZNF608	0.972	0.001
miR-3065-5p	XAF1	-0.954	0.003	miR-152-3p	ZNF608	0.973	0.001

Table 2

Output of miARma-Seq analysis that combines the differential expression values, the miRNA–mRNA predicted pairs, prediction method name, predicted target site and the statistical correlation. Results are shown for the most upregulated and downregulated genes, and the Pearson’s R coefficient values and P-values are given.

Gene	miRNA	Gene logFC	miRNA logFC	Method	Target Site	R	P
MIR100HG	miR-99a-5p	10.641	-3.753	RNAHybrid	7-mer-m8	-0.916	0.019
BHLHE41	miR-125b-2-3p	3.656	-6.627	Miranda	7mer-m8	-0.967	0.002
BHLHE41	miR-152-3p	3.656	-1.230	Miranda	8mer	-0.933	0.006
BHLHE41	miR-194-5p	3.656	-1.492	Miranda	7mer-m8	-0.956	0.003
BHLHE41	miR-302d-3p	3.656	-2.322	Miranda	6mer	-0.975	0.001
BHLHE41	miR-3065-5p	3.656	-1.436	Miranda	7mer-m8	-0.838	0.037
XAF1	miR-302a-3p	3.583	-1.508	Miranda	8mer	-0.837	0.038
XAF1	miR-302a-5p	3.583	-1.788	Miranda, Microtar	7mer-m8	-0.894	0.016
XAF1	miR-302b-3p	3.583	-1.366	Miranda	6mer	-0.985	3.5E-4
XAF1	miR-302d-3p	3.583	-2.322	Miranda	7mer-m8	-0.975	0.001
XAF1	miR-3065-5p	3.583	-1.436	Miranda	7mer-m8	-0.954	0.003
XAF1	miR-34c-5p	3.583	-1.524	Miranda, RNAHybrid	8mer	-0.886	0.019
HS3ST5	miR-302a-3p	3.543	-1.508	Miranda	7mer-m8	-0.915	0.010
HS3ST5	miR-302b-3p	3.543	-1.366	Miranda	7mer-m8	-0.925	0.008
HS3ST5	miR-302d-3p	3.543	-2.322	Miranda, Microtar	7mer-m8	-0.939	0.005
HS3ST5	miR-3065-5p	3.543	-1.436	Miranda	3-8 6mer	-0.862	0.027
HS3ST5	miR-34c-3p	3.543	-2.290	Microtar, Miranda	7mer-m8	-0.824	0.044
DUSP4	miR-125b-2-3p	3.533	-6.627	Miranda	7mer-m8	-0.859	0.028
DUSP4	miR-205-5p	3.533	-1.444	Miranda	7mer-m8	-0.900	0.015
DUSP4	miR-302b-3p	3.533	-1.366	Miranda	7mer-m8	-0.815	0.048
DUSP4	miR-34b-5p	3.533	-3.320	Miranda	3-8 6mer	-0.972	0.001
DUSP4	miR-34c-5p	3.533	-1.524	Miranda	8mer	-0.922	0.009
OLFM1	miR-100-5p	-7.852	6.455	RNAHybrid	8mer	-0.948	0.004
OLFM1	miR-125b-5p	-7.852	4.528	RNAHybrid	1-7 8mer	-0.905	0.013
OLFM1	miR-4286	-7.852	3.695	RNAHybrid	1-7 8mer	-0.973	0.001
CRABP1	miR-4286	-6.396	3.695	RNAHybrid	1-7 7mer-m8	-0.903	0.014
ZNF608	miR-7704	-6.221	1.639	RNAHybrid, Miranda	8mer	-0.828	0.042
DKK1	miR-125b-1-3p	-6.181	7.406	Miranda	3-8 6mer	-0.864	0.026

Figure 1. Diagram representing all possible workflows in miARma-Seq used to analyzed miRNASeq or RNASeq data. It includes: a quality module to analyse the reads in order to detect sequencing errors; miRDeep to predict novel miRNAs; Reaper, cutadapt and minion to process adapter sequences; and Bowtie1, Bowtie2, STAR, Hisat2 and Tophat to align the reads. Finally, it is possible to carry out differential expression analysis, target prediction, functional analysis, and the integration of miRNAs and genes.

Figure 2. The most upregulated and downregulated genes in cetuximab resistant (CRC-CR) compared to non-resistant (CRC) samples derived from HCA-7 cell line. A) A volcano plot showing the five most downregulated (OLFM1, SERP2, CRABP1, DKK1 and ZNF608) and upregulated (MIR100HG, DUSP4, XAF1, BHLHE41 and HS3ST5) genes. The log2Fold Change and False Discovery Rate (-log10 FDR) values are shown. B) Boxplot showing the expression level in CRC-RC and CRC samples of the two most upregulated genes: MIR100HG and BHLHE41. C) As in B except showing the results of the two most repressed genes: OLFM1 and SERP2.

Figure 3. The most upregulated and downregulated miRNAs in CRC-CR relative to CRC samples. A) A volcano plot showing the five most upregulated (miR-100-5p, miR-125b-1-3p, miR-125b-5p, let-7a-2-3p and miR-4286) and downregulated (miR-1, miR-99a-5p, miR-125b-2-3p, miR-145-5p and miR-143-3p) miRNAs. The log2 Fold Change and False Discovery Rate ($-\log_{10}$ FDR) values are shown. B) Boxplot showing the expression level in CRC-RC and CRC samples of the two most upregulated miRNAs: miR-100-5p and miR-125b-1-3p. C) As in B except showing the results of the two most downregulated miRNAs: miR-1 and miR-125b-2-3p.

Figure 4. Diagram of predicted miRNA-mRNA interactions. Interactions are shown among the five most upregulated and downregulated genes in CRC-RC relative to CRC samples and miRNAs that i) are differentially expressed among CRC-RC and CRC samples, and ii) have significant negative expression correlations (R Pearson's coefficient ≥ -0.9 ; p-value ≤ 0.05) with the studied genes.