Computational and Structural Biotechnology Journal 19 (2021) 1154–1162

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Integrative approaches for analysis of mRNA and microRNA high-throughput data

Petr V. Nazarov [a,*], Stephanie Kreis [b]

[a] Multiomics Data Science Research Group, Department of Oncology & Quantitative Biology Unit, Luxembourg Institute of Health (LIH), Strassen L-1445, Luxembourg
[b] Signal Transduction Group, Department of Life Sciences and Medicine, University of Luxembourg, Belvaux L-4367, Luxembourg

ARTICLE INFO

Article history:
Received 20 November 2020
Received in revised form 19 January 2021
Accepted 20 January 2021
Available online 26 January 2021

Keywords:
microRNA
Transcriptomics
Data integration
Target prediction
Matrix factorization

ABSTRACT

Advanced sequencing technologies such as RNASeq provide the means for production of massive amounts of data, including transcriptome-wide expression levels of coding RNAs (mRNAs) and non-coding RNAs such as miRNAs, lncRNAs, piRNAs and many other RNA species. *In silico* analysis of datasets, representing only one RNA species is well established and a variety of tools and pipelines are available. However, attaining a more systematic view of how different players come together to regulate the expression of a gene or a group of genes requires a more intricate approach to data analysis. To fully understand complex transcriptional networks, datasets representing different RNA species need to be integrated. In this review, we will focus on miRNAs as key post-transcriptional regulators summarizing current computational approaches for miRNA:target gene prediction as well as new data-driven methods to tackle the problem of comprehensively and accurately dissecting miRNome-targetome interactions.

Contents

Abbreviations: CCA, canonical correlation analysis; CDS, coding sequence; circRNA, circular RNA; CLASH, cross-linking, ligation and sequencing of hybrids; CLIP, cross-linking immunoprecipitation; CNN, convolutional neural network; GO, gene ontology; ICA, independent component analysis; lncRNA, long non-coding RNA; miRNA, microRNA; mRNA, messenger RNA; NGS, next-generation sequencing; NMF, non-negative matrix factorization; PCA, principal component analysis; RNASeq, high-throughput RNA sequencing; TDMD, target RNA-directed miRNA degradation; TF, transcription factors.

* Corresponding author.
E-mail addresses: petr.nazarov@lih.lu (P.V. Nazarov), stephanie.kreis@uni.lu (S. Kreis).

https://doi.org/10.1016/j.csbj.2021.01.029

1. Introduction

An accurately fine-tuned regulation of the expression levels of the ~20 000 protein-coding genes and the far more abundant non-coding RNAs is essential for a healthy functioning of human cells. Gene regulation is achieved at different levels. Epigenetic mechanisms influence, prior to actual gene transcription, which gene is being transcribed in a given cell at a given time. Next, a multitude of intrinsic or extrinsic signals determine which transcription factors are activated downstream of signalling cascades to drive or inhibit transcription of target genes. At the post-transcriptional level, the half-life, stability and other factors affect the available amount of bioactive RNAs, but the most important cellular instrument to adjust expression levels of most classes of RNAs, are small non-coding RNAs called miRNAs (microRNAs).

A recent bioinformatic analysis of some 360 billion sequencing reads, revealed 2300 true human mature miRNAs, roughly half of which are annotated in miRBase V22 [1]. In humans, ~520 high-confidence miRNA canonical genes and ~120 conserved miRNA families (with similar seed sequences) have been identified [2] with each family targeting > 400 conserved mRNAs, altogether resulting in ~60% of all mRNAs being targeted by miRNAs [3].

miRNAs are remarkably stable 22 nt short oligonucleotides that are produced from miRNA-encoding genes in a well understood biogenesis process comprehensively reviewed elsewhere [2]. Mature miRNAs act as guide sequences directing the RNA-induced silencing complex (RISC) to target RNAs, which contain complementary binding sites that allow for a miRNA:target contact. The canonical and biologically most relevant interaction takes place between the seed region (5′ nucleotides 2–7) of a miRNA and the binding site in the 3′ UTR of an mRNA target, resulting in a rapid degradation of the mRNA or less frequently in an inhibition of its translation into protein [4]. It has been estimated that 100 s-1000 s of miRNA molecules are necessary to efficiently repress mRNA levels in a cell [5].

Although canonical interactions are functionally most relevant, the highly abundant non-canonical binding to regions outside the 3′UTR of mRNAs and involving nucleotides beyond the seed sequence within the miRNA as well as miRNA binding to other non-coding RNAs are likely contributing to post-transcriptional gene regulation by competing with canonical binding events and by occasionally leading to down-regulation of the non-canonical target itself [6–8] and own unpublished data [9]. Other non-coding RNAs such as long non-coding RNAs (lncRNAs), competing endogenous RNAs (ceRNAs) and circular RNAs (circRNAs) have also been shown to sequester miRNAs from the biologically active pool within the cell [10,11]. Further contributing to the complexity of post-transcriptional gene regulatory networks is the fact that the expression of miRNAs themselves is also a dynamically regulated process, involving above mentioned pre-transcriptional epigenetic and genetic mechanisms, transcription factors and signaling pathways, which are generally cell type- and disease-specific and have relevance for most physiological processes as well as in complex diseases such as cancer [12,13]. Finally, there is redundancy in miRNA-target gene interactions and different related or unrelated miRNAs might have to cooperate to induce a measurable effect on gene expression. Likewise, for an observable change in cellular phenotypes, expression changes in only one gene are often not sufficient and several genes might have to be regulated at the same time for a functional shift in cellular behavior.

Although target gene prediction has markedly improved over the last decade, we are still far from being able to predict the miRNA targetome with an acceptable quota of false positive and false negative results, although many tools are available for prediction of miRNA target genes [14–16]. The aforementioned properties and the problematic target gene prediction make miRNA-driven gene expression difficult to unravel in all its complexity. Only massive amounts of well-controlled, high quality and preferably dynamic data (collected at various time points) from different cells/tissues from healthy and diseased conditions representing different transcriptional states can improve our understanding of the intricate regulatory circuits involving miRNAs. This being a costly and tedious endeavor might in part explain why the initial enthusiasm on miRNA research some 10–15 years ago has somewhat cooled down. It soon became clear that simply generating mono-phasic miRNA profiling data would not lead to the desired overview of transcriptional regulatory networks.

In the current review, we will provide a brief update on available computational miRNA tools and then focus on approaches of data integration efforts aiming at jointly analysing gene expression datasets representing miRNAs and mRNAs, lncRNAs or other non-coding RNAs. Advancing such integrative *in silico* analyses becomes even more important in view of an ever-growing number of publically available transcriptomic datasets.

2. Methodologies for analysis of miRNA-target gene interactions

2.1. Approaches to miRNA target identification

From the first identification of miRNAs in the 1990 s (lin-4 and let-7 in *C.elegans*, [17,18]), tremendous efforts have gone into computational prediction, experimental detection and validation of miRNA target genes. Current experimental methods to analyse the miRNA targetome include: i) profiling expression levels of the entire transcriptome (by RNASeq or microarray) following overexpression or downregulation of a miRNA of interest; ii) measuring selected mRNA levels of predicted targets by qPCR and/or measuring levels of the corresponding proteins following overexpression or downregulation of a miRNA of interest; iii) crosslinking followed by immunoprecipitation of RISC complexes (CLIP and CLASH methods); iv) reporter gene assays with the target site of the miRNA cloned close to a reporter gene and with external delivery of the miRNA v) as well as profiling phenotypic traits following rescue of mutated or deleted miRNAs [4,14,19–22].

Experimental methods can provide direct links between miRNAs and their targets, but they are not error-free, extremely laborious, time consuming and expensive, especially when more than one miRNA is investigated. Some experimental methods were criticized for generating false positive results [21] and this is even more so when analysing data from RNASeq experiments following overexpression of miRNAs. The recent introduction of CLASH [23] provides an interesting addition to the experimental tool box by directly linking miRNAs with the bound mRNAs, lncRNAs or any other potential RNA target. However, also this technology has room for improvement as the identification of miRNA-target hybrids is still rather inefficient [24].

Computational prediction of canonical or non-canonical miRNA targets employ different statistical and machine-learning approaches and generally analyse some of the following criteria: i) degree of Watson-Crick pairing between the miRNA seed region and target site; ii) evolutionary conservation across species; iii) thermodynamic properties; iv) accessibility of target sites; v) sequence composition in the vicinity of seeds and target sites. Many studies and comprehensive reviews have described available tools before [4,8,15,25–28] some of which will be further discussed below.

With the development of high-throughput technologies in experimental genomics, it became feasible to detect complete transcriptomes and miRNomes of cells and tissues under various conditions. Expression profiles of matching miRNomes and tran-

scriptomes from given cells or tissues and CLIP-based next generation sequencing (NGS) provide large datasets, in which the true interaction partners need to be identified and this is not a trivial task. Integration of different datasets generally explores statistical similarities and inverse correlations of mRNA and miRNA expression patterns, which can be suggestive of potential interactions or the presence of co-regulated clusters. Such knowledge would certainly contribute to our understanding of miRNA functions if the number of false positive and negative predictions can be reduced. Below and in Fig. 1, some of the most popular and promising methods for integration of mRNA and miRNA datasets are summarized.

2.2. Databases of experimentally validated miRNA targets

Over the past 10 years, several dozen online resources with predicted or validated miRNA targets have been published [25,26]. However, many of them had quite a short life cycle and are currently either unavailable or outdated. Here we concentrate on databases that are regularly updated, starting with experimentally validated targets, as they are the most valuable source for miRNA-target gene pairs.

DIANA-TarBase is the most complete collection of experimentally supported miRNA targets. The current version 8 of the database contains around 670 k of unique miRNA-mRNA pairs (and over 1 M of entries of which ~800 k have direct experimental support). It has been persistently developed over the past decade and is based on both literature curation (1.2 k) and analysis of results of low- and high-throughput experiments [29]. Importantly, the database can be downloaded, which makes it an interesting source for an automatic computational analysis.

MiRTarBase is the second largest collection of experimentally validated targets [30]. The current version describes over 430 k miRNA-target interactions. It is based on manual curation of around 11 k publications and is also downloadable.

2.3. Tools for prediction of miRNA targets

TargetScan. Among the target gene prediction tools, TargetScan [28] is the most widely used. The tool and corresponding database have been supported and developed since 2005, with the current version 7.2 available since 2018, covering miRNA interactions in 8 mammals (including human, mouse and rat and several other organisms). Predictions by TargetScan are made by searching for conserved 6–8-mer binding sites in mRNAs, matching with miRNA seed regions and taking into account the surrounding sequences (in total, 14 distinct parameters were used for prediction). Prediction databases can be downloaded separately for conserved miRNAs with conserved mRNA targets and for non-conserved miRNAs with conserved and non-conserved mRNA targets (the most complete). Agarwal and colleagues showed that non-canonical sites rarely lead to mRNA modulation despite binding of miRNA, and therefore focused their predictions on the canonical sites [28]. However, given the increasing evidence that non-canonical interactions between miRNAs and their targets also play an important role in some (but not all) gene regulatory networks ([2,6] and unpublished data in [9]), such interactions involving nucleotides outside the seed of the miRNA and within the CDS (rather than the 3′ UTR) represent useful additions to current prediction tools.

DIANA-microT-CDS. DIANA tools, in parallel to its validated target database introduced above also proposes the target-prediction by DIANA-microT-CDS, the 5th version of the microT algorithm. This method uses a machine learning approach that is trained on photoactivatable-ribonucleoside-enhanced cross-linking immunoprecipitation (PAR-CLIP) data and is able to predict miRNA binding sites both in 3′-UTR and in coding sequence (CDS) regions. The tool does not allow the download of predicted targets for all miRNAs, however it allows access to predictions through the Taverna workflow management system.

miRDB is an online resource for miRNA target prediction and functional annotations [31]. The method uses the machine-learning algorithm MirTarget based on a support vector machine, which takes evidences from several other algorithms, including TargetScan [28], PicTar [32], miRanda [33] and combines them with features obtained from miRNA overexpression experiments and CLIP data. The method allows for predicting gene targets with binding sites both in evolutionary conserved and non-conserved regions of 3′-UTR. Current lists of predicted targets are available for human, mouse, rat dog and chicken and can be downloaded from the miRDB server. In total, the database contains 3.5 M miRNA-mRNA target pairs and over 1.6 M human entries.

There are more tools available that build their predictions on combination of the above mentioned databases and algorithms such as the recently updated miRWalk [34]) or including experimental data, miRGator [35], STarMir [36] or miRGate [37]), but these tools seem to be applied less frequently for the moment. Another important property of such databases is the frequency of update. After a 5-year stand-by period, databases lose their attractiveness for the community and deviate too much from the current version of miRBase [38].

Recently, a deep learning approach, which is based on advanced artificial neural networks, has been applied to different tasks in bioinformatics [39]. Although neural networks have been known for decades as universal modelling tools (e.g. [40]) only now have computer power and, more importantly, the size of datasets reached the dimensions where efficient and robust networks can be built. A convolutional neural network (CNN), one application of deep learning models, has been successfully applied to predict binding affinities between miRNAs and 12-mer sequences [8]. The model substantially improved prediction of mRNA repression in cell lines. Considering these promising developments, we should expect to see more works dedicated to application of deep learning in miRNA target predictions. Another example of a deep learning approach is provided by the DeepMirTar tool [41] where authors used stacked de-noising auto-encoders to predict miRNA targets at the site level. Interestingly, the tool showed improved performance in target prediction compared to standard methods, including TargetScan.

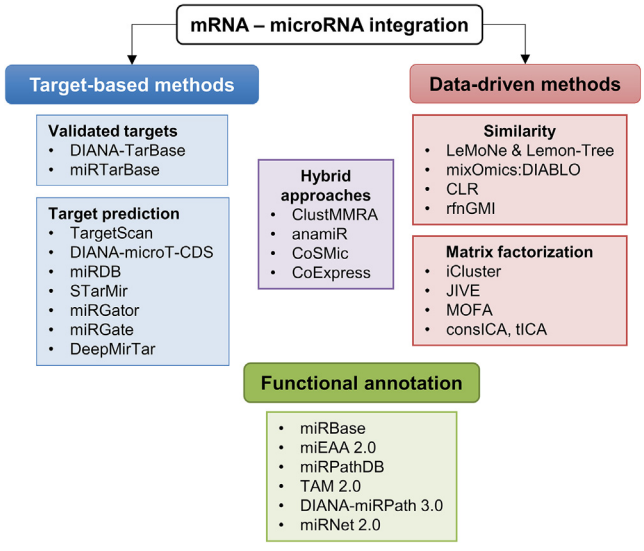


Fig. 1. An overview of the main methods for miRNA:mRNA data integration.

2.4. Functional annotation

Functional annotation of individual miRNAs or related groups or families of miRNAs is important for understanding their biological roles and can also be useful to connect miRNAs to known functional gene sets. For a small number of selected miRNAs, their functions can be determined experimentally. However, high-throughput datasets require bioinformatics analysis with literature curation [42,43] or functional analysis of miRNA targets [44]. The fact that miRNA and regulated mRNAs are linked by a "many-to-many" relationship, significantly increases the complexity of functional miRNA annotation. The most important tools are introduced below.

miRBase – the primary public database for miRNA sequences and nomenclature [38]. The current release 22.1 contains 38,589 entries for 271 organisms. Each entry represents a miRNA precursor sequence with a predicted hairpin of the miRNA transcript, the genomic location, references from literature, the mature miRNA with manually curated gene ontology (GO) terms [42] and other information. Since the first presentation in 2002, miRBase has been widely used as a reference catalogue by other miRNA-related tools. A somewhat adverse effect of miRBase popularity is a requirement of a high robustness for the presented information and, therefore, the lack of pilot tools incorporated. For example, while a single mature miRNA can be linked to GO terms, there is no enrichment analysis and functional annotation for a set or group of miRNAs. Therefore, other tools have to be used in order to obtain such annotations.

One such tool is the 2nd version of the miRNA Enrichment Analysis and Annotation package (miEAA 2.0), which aims at functional analysis of sets of miRNAs [45]. MiEAA is based on GeneTrail, an enrichment analysis tool for gene sets [46] and integrates data from different sources including miRBase, miRWalk, miRTarBase and others. MiEAA can work with lists of precursors and mature miRNAs and performs either over-representation analysis (Fisher exact test) or enrichment analysis (Kolmogorov-Smirnov test). Available categories include for example GO, KEGG pathways, target genes, chromosomal location, diseases, drugs (altogether 130 k categories). In parallel, the same group proposed another tool, miRPathDB. This dictionary, based on over-representation analysis of miRNA targets provides miRNA-centric, gene-centric and pathways-centric views [44].

Another tool, often used for functional annotation of miRNA is TAM 2.0 [43]. The tool is based on a manual curation of 9 k papers. It includes 1238 miRNA sets associated with different diseases, miRNA families, transcription factors and biological functions. With input lists of up- and down-regulated miRNA, TAM 2.0 can analyse deregulated miRNAs in two conditions by cosine similarity. This measure is based on the inner product of two vectors, and is sensitive to the means of these vectors. For centred vectors (zero mean), cosine similarity is equal to the Pearson correlation (see Fig. 2B). Importantly, in both miEAA and TAM, users can provide their own reference (background) set of miRNAs, reducing the bias.

DIANA tools also provide functional annotations – DIANA-miRPath 3.0 [47] works with both single and multiple miRNAs. Categories are represented by KEGG molecular pathways and GO in several organisms. The tool links miRNAs with regulated mRNAs using several target prediction algorithms (DIANA-microT-CDS, TargetScan) and validated targets (DIANA-miRTarBase). An implemented reverse-search module allows identifying miRNAs that control specific pathways. A drawback of the tool is the fact that reference miRNA or mRNA lists cannot be selected or changed.

A visual analytics and integration tool, miRNet 2.0 [48], provides users with the possibility to build networks based on own lists of miRNAs, mRNAs, transcription factors (TF) and other regulatory elements (12 modules in total). In cases where miRNAs are selected as a starting module, the tool builds a network of miRNAs and their targets (based on one of above mentioned databases) and visualizes it. It also allows for functional annotation based on the target list.

2.5. Data-driven methods: similarity-based

Accessibility of large transcriptomic datasets that represent mRNA and miRNA profiles across many samples and conditions facilitate the use of data-driven, hypothesis-free approaches to data integration. The simplest way to integrate different datasets would be the calculation of correlation within and between miRNA and mRNA profiles, which is possible if both data types were collected from the same samples or at least the same conditions (Fig. 2, the heatmaps and networks illustrate the basic idea but are based on real data from [49]). This approach, however, is unable to discriminate between direct and fake miRNA:mRNA interactions, originating from common hidden regulators such as transcription factors. Therefore, additional filters should be used to prioritize meaningful interactions. Such filtering can be done, for example, by considering only negatively correlated miRNA:mRNA pairs, where the mRNA is also predicted as a potential target of the miRNA [13]. Additional layers of complexity were brought in by the fact, that both mRNA and miRNA datasets have a high intrinsic correlation. This could originate from simultaneous activation or repression of genes and miRNAs participating in the same or linked functions. In order to deal with such behaviour, a module-based method, the Learning Module Network LeMoNe was proposed and applied to infer functions and regulated targets of miRNAs [50]. Their two-step algorithm includes a partition of genes into co-expressed clusters followed by inferring a regulatory program for each cluster. This method was further developed in a cross-platform open source Lemon-Tree tool [51].

Another approach is built on canonical correlation analysis (CCA), a classical method to establish linear relations between two sets of correlated observations. The method accepts the fact that experimental data are correlated and builds linear combinations of features for both datasets in a way to maximize correlation between them. One of the most successful example of this approach is DIABLO, an integrative method based on sparse generalized CCA [52]. This method can be used to identify markers and to build links between multi-omics data and patient groups. It was specifically tested on integration of miRNA and mRNA datasets. The method is implemented as a part of a powerful mixOmics R-package, able to integrate results across several omics datasets or derived from different studies [53]. Recently, another useful application of CCA was applied to identify miRNA-disease associations [54].

Pearson correlation captures only linear dependency between expression of mRNA and miRNA. Changing to Spearman rank correlation could broaden this to a wider range of monotonic dependencies. However, in reality miRNA:mRNA dependency can be more complex, and requires considering two-dimensional distributions between miRNA and mRNA expression. In [55], triangular-shaped patterns between miRNA and its targets were reported. The authors proposed an "antagonism pattern detection algorithm", based on counting and statistical assessment of observations in lower and upper triangles of a scatter plot. Alternatively, mutual information can be used to detect non-random profiles. Mutual information is able to capture different non-linear patterns and is used in the context likelihood of relatedness (CLR) algorithm [56]. As an example of a successful application, this method allowed for identification of miRNA regulatory network in glioblastoma [57].

Several authors addressed the problem of context- or condition-specific relations between miRNAs and their targets [58–60].

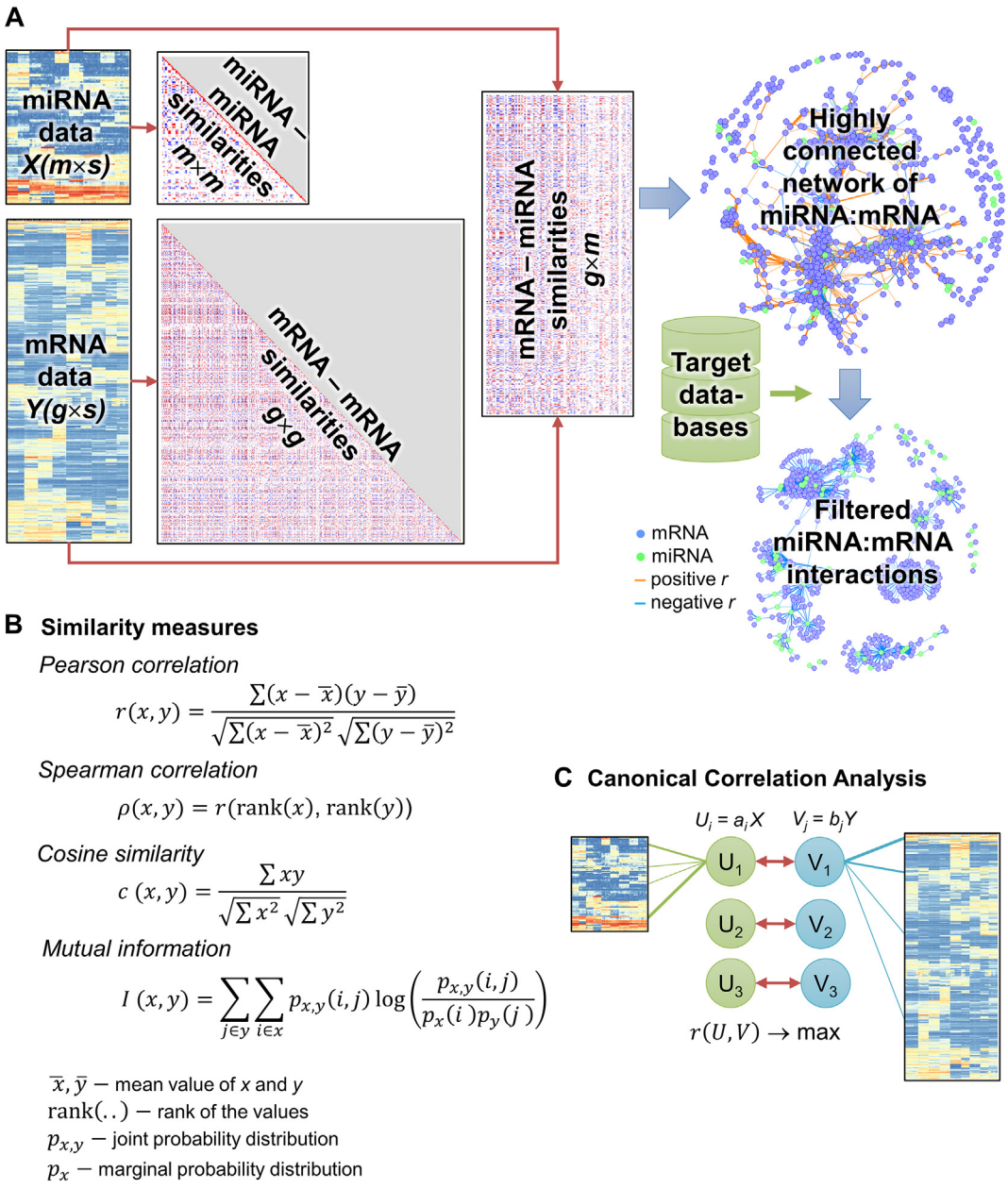


Fig. 2. Similarity-based methods. (A) Correlation or other similarity measures produce a highly-connected network of miRNAs and potentially regulated mRNAs. Due to the high level of correlation between genes, such networks are often redundant and contain many false positives and should be filtered using miRNA target databases. (B) Standard similarity measures used for comparison of miRNA and mRNA profiles. (C) Canonical correlation analysis (CCA) approach: new features *U* and *V* are built as linear combinations of miRNA and mRNA profiles in a way that maximizes correlation between them.

Context-specificity can be taken into account using biclustering, a method to cluster subsets of genes that have similar expression in a subset of conditions [58,59]. For instance, the rectified factor network-based biclustering for genes, miRNAs and interactions (rfnGMI), presented by Su et al. [59], was applied to detect genes and miRNA markers specific to breast cancer.

2.6. Data-driven methods: Matrix factorization

Matrix factorization methods originally represent the expression matrix as a matrix product of lower rank matrices with the original mRNA and miRNA data measured in *m* samples (Fig. 3). We assume that there are some hidden (latent) variables or components that are shared between data types. These variables can for example, explain variability of the data over subgroups of samples (cancer subtypes or patient outcomes). The effect of a latent variable on features (genes, miRNAs) is described by the "metagene" matrix, the effect on samples by the "weight" matrix. Moreover, the weights of the same latent variables affecting mRNA and miRNA data should either be the same, or should at least be correlated. By linking these variables, it is possible to integrate the data, linking miRNAs and mRNA belonging to the same latent variable.

Various methods of building this matrix product and properties of the resulting matrices have been proposed. Here we describe several of the most widespread approaches. Some of them focus on general integration of multi-modal (multi-omics) data, which can also be projected onto miRNA and mRNA data integration. The first attempt to develop matrix factorization for linking sets of mRNA and miRNA was performed within the SNMNMF tool

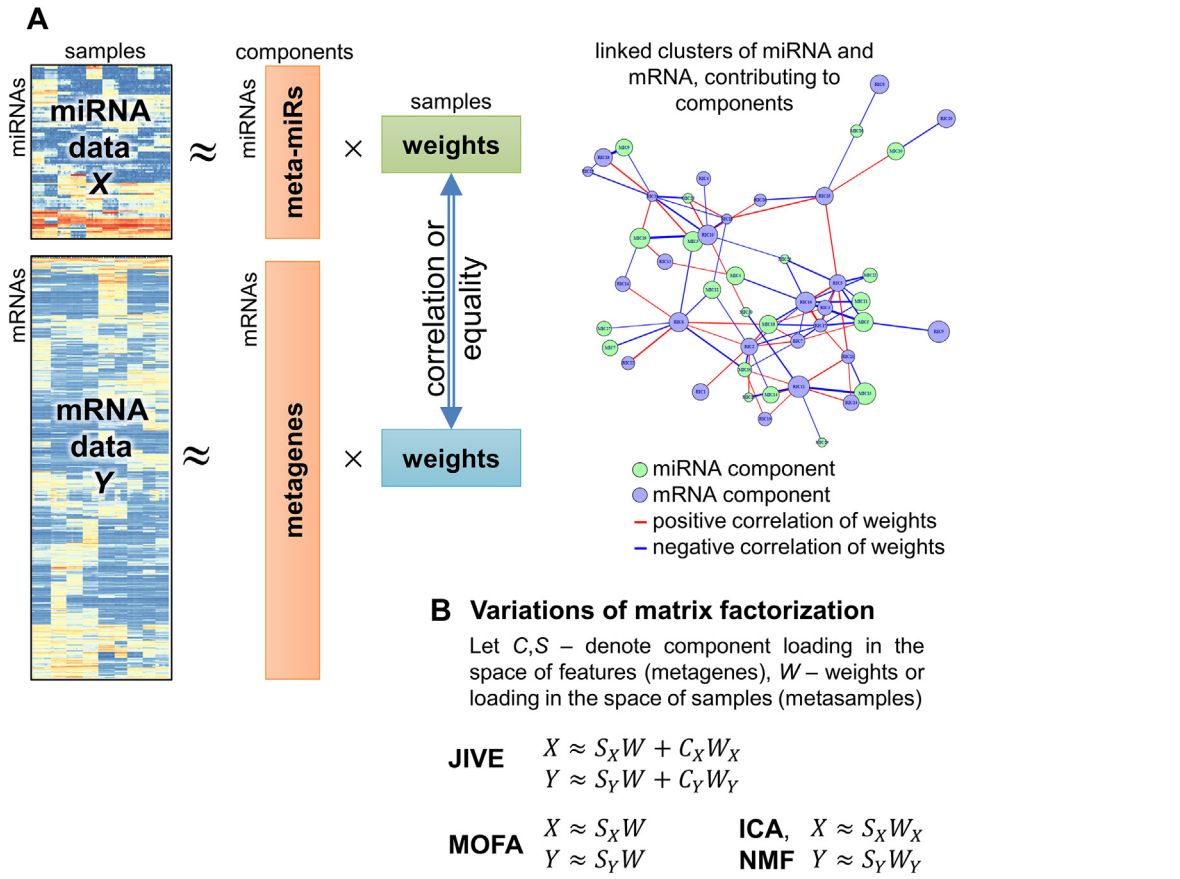


Fig. 3. Matrix factorization methods. (A) Each expression matrix is presented as a product of two lower-rank matrices. Integration can be performed by correlating weight profiles over the samples resulting in a network of linked components. Some methods (e.g. MOFA) use a single weight matrix for both datasets. (B) Variation of the approach in different methods: JIVE, MOFA and ICA. The classical NMF approach has the same mathematical expression as ICA, but requires that both metagene (S) and weight (W) matrices are composed of non-negative elements.

(Sparse Network-Regularized Multiple Non-negative Matrix Factorization) [61], which is, unfortunately, not available anymore. The tool employed non-negative matrix factorization (NMF) [62], which requires matrix elements to be non-negative. Although the method fits well to the physical nature of non-negative RNA quantities, it suffers from ambiguity of matrix decomposition and requires additional restrictions or regularizations.

Integrative clustering of multiple genomic data types (iCluster) used likelihood maximization to build a joint latent variable model, with the "weights" matrix shared between data types. L1-norm penalization was used to limit the loadings [63]. Its further development, iCluster+, now allows a variety of different data types, including binary and categorical data. The potential of the algorithm for miRNA:mRNA integration has also been used in the group structured tight iCluster method (GST-iCluster) [64].

Joint and Individual Variation Explained (JIVE) was proposed as an extension of the Principle Component Analysis (PCA) approach. It decomposes the original data into a sum of two informative parts: a low-rank approximation capturing joint structures between data types, and an approximation capturing the distinct structures for each data type. This was shown to outperform CCA. The method was tested on mRNA and miRNA datasets related to glioblastoma of TCGA and, being supplemented with target prediction tools, identified informative clusters of interacting miRNA and mRNA [65].

One of the most advanced matrix factorization methods is the Multi-Omics Factor Analysis (MOFA) [66]. The method was developed for integration of various levels of omics data and for discovering the main sources of variation in multi-omics datasets. Using a Bayesian approach, MOFA infers a set of hidden or pre-defined factors that capture biological and technical variability in the data. The main paradigm is in line with the representation in Fig. 3 ('components' are now called 'factors'). The specificity of the method is that the weight of matrices is considered to be identical and is estimated for all omics data simultaneously. An advantage of the method is also its ability to work with missing data. If some measurements are not available either for mRNA or for miRNA datasets, they will be imputed. MOFA allows estimating factor importance by assessing the proportion of variance explained by each factor in each dataset. Although this method has not been developed for integration of miRNA and mRNA data, it has a lot of potential for this application.

Independent component analysis (ICA) is another powerful method to integrate multi-omics data. The algorithms decomposes original data into signals that are as statistically independent as possible [67]. Despite the method being completely unsupervised, it usually finds more biologically relevant signals in the data then PCA and helps cleaning signals from technical biases [68,69]. In order to increase reproducibility of the analysis, we recently proposed consensus ICA (consICA) [68] and showed its applicability to integration of miRNA:mRNA data in melanoma. A multidimensional version of the method, tensor ICA (tICA), was shown to outperform CCA, iCluster and JIVE in identifying biological sources of data variation [70].

2.7. Hybrid methods

Methods that combine information about miRNA targets with experimental observations have the highest potential for context-dependent integration of miRNA and mRNA data. For example, we presented a basic user friendly tool CoExpress for the analysis of miRNA:mRNA co-expression that used information from the TargetScan database for filtering potentially linked miRNAs and mRNAs [13].

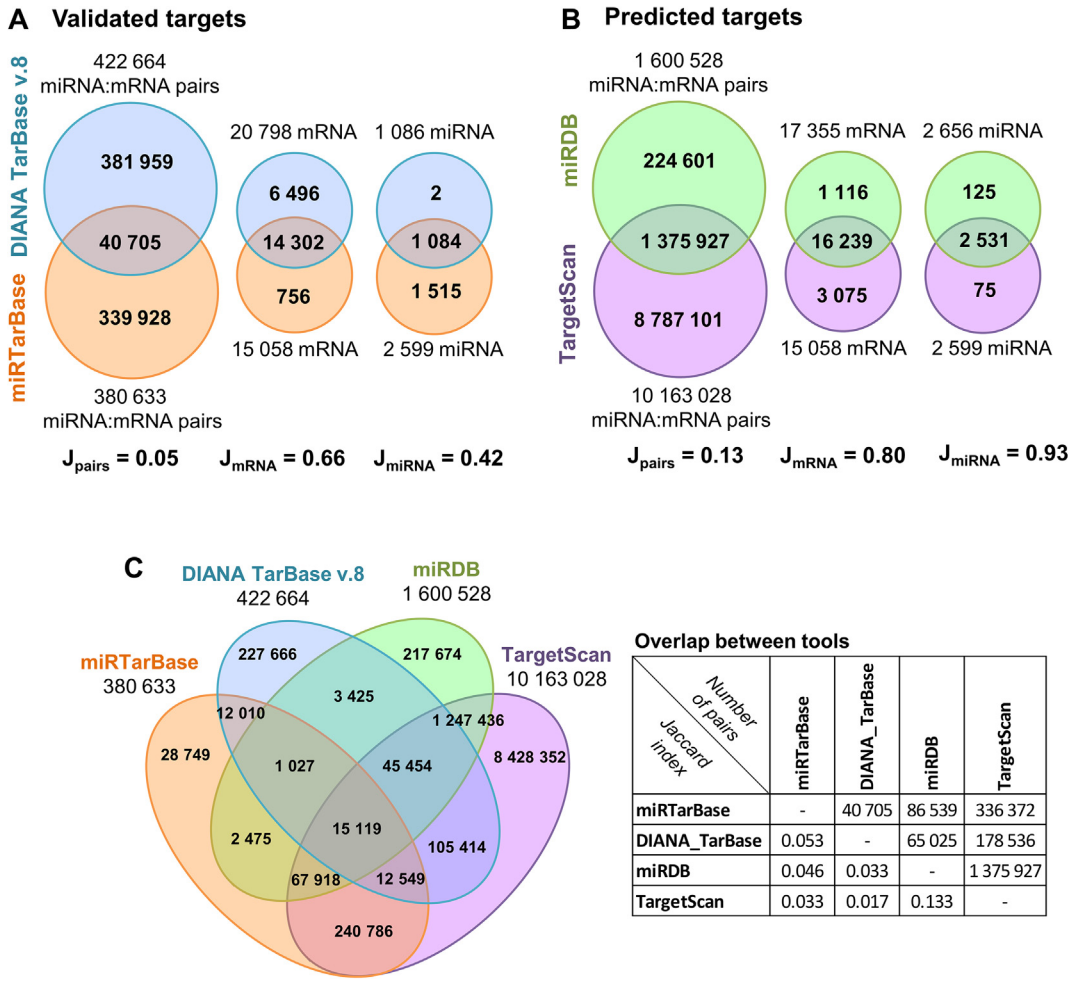
Context-specific interactions between miRNA and mRNA were also investigated in context-specific microRNA analysis (CoSMic) [58]. This algorithm combines sequence-based predictions by TargetScan and other tools with miRNA and mRNA expression data (Spearman correlation) and focuses only on miRNAs that are active in a specific subgroup of samples. The authors demonstrated in a well-controlled cell line experiment that their method efficiently filters out false positive interactions and helps identifying context-specific targets.

Another example is the miRNA master regulator analysis (MMRA), which starts from a differential analysis of miRNA and mRNA expression and then looks for miRNAs whose targets are enriched among differentially expressed mRNAs. After this, a network is built around each miRNA. miRNAs with the highest potential to explain subtype-specific mRNAs are selected (a stepwise linear regression is used to predict mRNA expression by miRNA) [60]. The approach was further developed in clustered miRNA master regulator analysis (ClustMMRA), and tested on epithelial to mesenchymal transition in triple negative breast cancer cells [71]. Now, instead of analysis of each specific miRNA, genomic clusters of miRNAs are considered. Targets of miRNAs were predicted using a combination of tools, including TargetScan and miRTarBase.

A similar paradigm was realized in the anamiR R/Bioconductor package. Experimental data undergo differential expression analysis, correlation and intersection with databases of predicted or validated targets and functional annotation of both miRNA and mRNA data [72]. The authors developed a function-driven analysis workflow to identify miRNA-gene interaction pairs among those participating in significant pathways.

3. Summary

Here we provide an overview of the most common and promising approaches that were developed over the past decade to integrate miRNA and mRNA expression data in order to gain a deeper understanding of the gene regulatory fine-tuning events in cellular processes. Historically, a main emphasis has been placed on miRNA-target prediction methods. However, despite recent experimental advances with regards to next generation sequencing



Overlap between tools

Number of pairs / Jaccard index	miRTarBase	DIANA_TarBase	miRDB	TargetScan
miRTarBase	-	40 705	86 539	336 372
DIANA_TarBase	0.053	-	65 025	178 536
miRDB	0.046	0.033	-	1 375 927
TargetScan	0.033	0.017	0.133	-

Fig. 4. Limited overlap of validated (A) and predicted (B) miRNA targets. For TargetScan, all available miRNA:mRNA pairs including conserved and non-conserved sites and miRNAs were used. To avoid biases due to the selection of miRNAs and mRNAs, we separately intersected lists of miRNAs and mRNAs considered in the tools. The Jaccard index is reported beneath each Venn diagram. (C) Overview of all 4 tools. The number of overlapping miRNA:mRNA pairs as well as the corresponding Jaccard indexes are shown in the table.

analysis of cellular transcriptomes, the main issue of *in silico* target prediction remains the limited accuracy and low agreement between the tools. Even databases with experimentally validated targets and targets based on curated literature mining suffer from little congruence and not much has changed since an initial assessment of precision and sensitivity of available tools was published by Alexiou et al. [73].

To illustrate this point, the intersection between different tools is visualized in Fig. 4. First, the overlap of experimentally validated targets from DIANA TarBase v8 and miRTarBase is shown. Both tools are widely used in the scientific community, but not much attention is given to the discrepancy between these two databases. From our observations for human miRNAs, only ~10% of the recorded miRNA targets are common between these databases (Jaccard index J_{pairs} = 0.05). At the same time, Jaccard indexes for the considered miRNA and mRNA lists are 0.42 and 0.66, respectively, and cannot explain such low overlap of the pairs. Interestingly, prediction algorithms have a slightly higher accordance (Fig. 4B) but this is most probably linked to the extremely high number of reported miRNA:mRNA pairs. Indeed, TargetScan, in its most unrestricted configuration, which includes both conserved and non-conserved sites as well as miRNAs, reported ~26% of all possible miRNA:mRNA pairs as potentially possible. An intersection of all four databases (Fig. 4C), resulted in a somewhat higher concordance (J_{pairs} = 0.046) between miRTarBase and MirDB when predicted and validated targets were combined. The overall very low agreement between different methods and databases can be explained, to some extent, by the notion of context-specific interactions between miRNAs and mRNAs. Indeed, as one miRNA can regulate many mRNAs, the concentration of an active miRNA strongly depends on stoichiometric flux balance with all its interactors. In addition to mRNAs, miRNAs can also bind to lncRNAs or circRNAs that may act as a miRNA sponges [74] or miRNAs may be degraded via target RNA-directed miRNA degradation (TDMD) [75]. Considering the complexity of cellular gene regulatory models, which involve many players with unknown binding affinities and interaction potential, we reckon that a combination of validated experimental data with prediction algorithms trained on more and better high throughout datasets will eventually lead to a more accurate forecast of miRNA targets.

It is a widely accepted simplification that miRNA/mRNA interactions are generally representing a negative correlation, an observation backed by most experimental data. However, several factors can reduce observable negative correlations: (i) co-existence of miRNAs and their targets in specific cell types can lead to a positive correlation in an experiment where several cell types are considered; (ii) as previously shown, miRNA responses to a stimulus may be delayed compared to mRNA responses and thus can only be captures in a time-course experiment [13]; (iii) finally, a scatter plot of miRNA and its targets should have a shape of a triangular area rather than of a simple linear dependency (an absence of miR should not lead to an increase of its target mRNA, if it was not produced beforehand) [55].

Moreover, it is important to consider the number of samples required for the different types of data integration. If methods based on differential expression analysis and target databases require only few samples (enough to find differences between two conditions), correlation-based approaches already need around 10 independent conditions to ensure a bell-shaped correlation distribution under the null hypothesis, i.e. independence of miRNA and mRNA. Semi-supervised matrix factorization methods, such as MOFA, may work when the number of samples is twice higher than the number of estimated factors. Finally, data-driven ICA is sensitive to the number of samples. From our experience, at least 4 samples are needed per one independent component, with 40–50 components required for a good interpretation and associating of the components to the clinical information [67]. Thus, hundreds of samples are needed for such an approach.

Acquisition of high quality datasets of both miRNA and mRNA expression profiles opens the door to apply these advanced deconvolution methods (MOFA, JIVA, ICA) on one side and modern deep-learning models that predict binding affinities on the other side. In future, we may expect to see inclusion of target predictions into matrix factorization algorithms, for example as a regularization factor taken into account during building of the metagene matrices. At the same time, deep learning models can take into account the context of miRNA:mRNA interaction, if enough data are available. Together, these approaches will certainly help to provide improved miRNA targetome predictions in the near future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Luxembourg National Research Fund (C17/BM/11664971/DEMICS) to PN and by a Fondation Cancer (Luxembourg) grant (SecMelPro) to SK.

References

[1] Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, et al. An estimate of the total number of true human miRNAs. Nucleic Acids Res 2019;47(7):3353–64.
[2] Bartel DP. Metazoan MicroRNAs. Cell 2018;173(1):20–51.
[3] Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19(1):92–105.
[4] Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet 2012;13(4):271–82.
[5] Denzler R, McGeary SE, Title AC, Agarwal V, Bartel DP, Stoffel M. Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. Mol Cell 2016;64(3):565–79.
[6] Chipman LB, Pasquinelli AE. miRNA Targeting: Growing beyond the Seed. Trends Genet 2019;35(3):215–22.
[7] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013;153(3):654–65.
[8] McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science 2019;366(6472):eaav1741.
[9] Kozar I. From drug resistance mechanisms to microRNA function in melanoma. Luxembourg: University of Luxembourg; 2020.
[10] Panda AC. Circular RNAs Act as miRNA Sponges. Adv Exp Med Biol 2018;1087:67–79.
[11] Lou W, Ding B, Fu P. Pseudogene-Derived lncRNAs and Their miRNA Sponging Mechanism in Human Cancer. Front Cell Dev Biol 2020;8:85.
[12] Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C. Analysis of microRNA-target interactions across diverse cancer types. Nat Struct Mol Biol 2013;20(11):1325–32.
[13] Nazarov PV, Reinsbach SE, Muller A, Nicot N, Philippidou D, Vallar L, et al. Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function. Nucleic Acids Res 2013;41(5):2817–31.
[14] Rojo Arias JE, Busskamp V. Challenges in microRNAs' targetome prediction and validation. Neural Regen Res 2019;14(10):1672–7.
[15] Riffo-Campos AL, Riquelme I, Brebi-Mieville P. Tools for Sequence-Based miRNA Target Prediction: What to Choose?. Int J Mol Sci 2016;17(12).
[16] Chen L, Heikkinen L, Wang C, Yang Y, Sun H, Wong G. Trends in the development of miRNA bioinformatics tools. Brief Bioinform 2019;20(5):1836–52.
[17] Ambros V. microRNAs: tiny regulators with great potential. Cell 2001;107(7):823–6.
[18] Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75(5):843–54.
[19] Mittal N, Zavolan M. Seq and CLIP through the miRNA world. Genome Biol 2014;15(1):202.
[20] Li J, Zhang Y. Current experimental strategies for intracellular target identification of microRNA. ExRNA 2019;1(1):6.

[21] Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. Nucleic Acids Res 2011;39(16):6845–53.
[22] Clement T, Salone V, Rederstorff M. Dual luciferase gene reporter assays to study miRNA function. Methods Mol Biol 2015;1296:187–98.
[23] Helwak A, Tollervey D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). Nat Protoc 2014;9(3):711–28.
[24] Sethuraman S, Thomas M, Gay LA, Renne R. Computational analysis of ribonomics datasets identifies long non-coding RNA targets of gamma-herpesviral miRNAs. Nucleic Acids Res 2018;46(16):8574–89.
[25] Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. Bioinformatic tools for microRNA dissection. Nucleic Acids Res 2016;44(1):24–44.
[26] Shukla V, Varghese VK, Kabekkodu SP, Mallya S, Satyamoorthy K. A compilation of Web-based research tools for miRNA analysis. Brief Funct Genomics 2017;16(5):249–73.
[27] Singh NK. miRNAs target databases: developmental methods and target identification techniques with functional annotations. Cell Mol Life Sci 2017;74(12):2239–61.
[28] Agarwal V, Bell GW, Nam JW, Bartel DP: Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015, 4.
[29] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res 2018;46(D1):D239–45.
[30] Huang HY, Lin YC, Li J, Huang KY, Shrestha S, Hong HC, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res 2020;48(D1):D148–54.
[31] Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res 2020;48(D1):D127–31.
[32] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet 2005;37(5):495–500.
[33] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol 2004;2(11):e363.
[34] Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: An online resource for prediction of microRNA binding sites. PLoS ONE 2018;13(10):e0206239.
[35] Cho S, Jang I, Jun Y, Yoon S, Ko M, Kwon Y, et al. Lee B *et al.*: MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. Nucleic Acids Res 2013;41(Database.
[36] Rennie W, Liu C, Carmack CS, Wolenc A, Kanoria S, Lu J, Long D, Ding Y: STarMir: a web server for prediction of microRNA binding sites. Nucleic Acids Res 2014, 42(Web Server issue):W114-118.
[37] Andres-Leon E, Gonzalez Pena D, Gomez-Lopez G, Pisano DG. miRGate: a curated database of human, mouse and rat miRNA-mRNA targets. Database (Oxford) 2015. 2015:bav035.
[38] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47(D1):D155–62.
[39] Tang B, Pan Z, Yin K, Khateeb A. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Front Genet 2019;10:214.
[40] Nazarov PV, Apanasovich VV, Lutkovski VM, Yatskou MM, Koehorst RB, Hemminga MA. Artificial neural network modification of simulation-based fitting: application to a protein-lipid system. J Chem Inf Comput Sci 2004;44(2):568–74.
[41] Wen M, Cong P, Zhang Z, Lu H, Li T. DeepMirTar: a deep-learning approach for predicting human miRNA targets. Bioinformatics 2018;34(22):3781–7.
[42] Huntley RP, Kramarz B, Sawford T, Umrao Z, Kalea A, Acquaah V, et al. Expanding the horizons of microRNA bioinformatics. RNA 2018;24(8):1005–17.
[43] Li J, Han X, Wan Y, Zhang S, Zhao Y, Fan R, et al. TAM 2.0: tool for MicroRNA set analysis. Nucleic Acids Res 2018;46(W1):W180–5.
[44] Kehl T, Kern F, Backes C, Fehlmann T, Stockel D, Meese E, et al. miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. Nucleic Acids Res 2020;48(D1):D142–7.
[45] Kern F, Fehlmann T, Solomon J, Schwed L, Grammes N, Backes C, et al. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. Nucleic Acids Res 2020;48(W1):W521–8.
[46] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof HP: GeneTrail–advanced gene set enrichment analysis. Nucleic Acids Res 2007, 35(Web Server issue):W186-192.
[47] Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic Acids Res 2015;43(W1):W460–6.
[48] Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res 2020;48(W1):W244–51.
[49] Kirchmeyer M, Servais F, Ginolhac A, Nazarov PV, Margue C, Philippidou D, et al. Systematic Transcriptional Profiling of Responses to STAT1- and STAT3-Activating Cytokines in Different Cancer Types. J Mol Biol 2020;432(22):5902–19.
[50] Bonnet E, Tatari M, Joshi A, Michoel T, Marchal K, Berx G, et al. Module network inference from a cancer gene expression data set identifies microRNA regulated modules. PLoS ONE 2010;5(4):e10162.
[51] Bonnet E, Calzone L, Michoel T. Integrative multi-omics module network inference with Lemon-Tree. PLoS Comput Biol 2015;11(2):e1003983.
[52] Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 2019;35(17):3055–62.
[53] Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 2017;13(11):e1005752.
[54] Chen H, Zhang Z, Feng D. Prediction and interpretation of miRNA-disease associations based on miRNA target genes using canonical correlation analysis. BMC Bioinf 2019;20(1):404.
[55] Martignetti L, Laud-Duval K, Tirode F, Pierron G, Reynaud S, Barillot E, et al. Antagonism pattern detection between microRNA and target expression in Ewing's sarcoma. PLoS ONE 2012;7(7):e41770.
[56] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 2007;5(1):e8.
[57] Genovese G, Ergun A, Shukla SA, Campos B, Hanna J, Ghosh P, et al. microRNA regulatory network inference identifies miR-34a as a novel regulator of TGF-beta signaling in glioblastoma. Cancer Discov 2012;2(8):736–49.
[58] Bossel Ben-Moshe N, Avraham R, Kedmi M, Zeisel A, Yitzhaky A, Yarden Y, et al. Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets. Nucleic Acids Res 2012;40(21):10614–27.
[59] Su L, Liu G, Wang J, Xu D. A rectified factor network based biclustering method for detecting cancer-related coding genes and miRNAs, and their interactions. Methods 2019;166:22–30.
[60] Cantini L, Isella C, Petti C, Picco G, Chiola S, Ficarra E, et al. MicroRNA-mRNA interactions underlying colorectal cancer molecular subtypes. Nat Commun 2015;6:8878.
[61] Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics 2011;27(13):i401–9.
[62] Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 2004;101(12):4164–9.
[63] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 2009;25(22):2906–12.
[64] Kim S, Oesterreich S, Kim S, Park Y, Tseng GC. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. Biostatistics 2017;18(1):165–79.
[65] Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and Individual Variation Explained (Jive) for Integrated Analysis of Multiple Data Types. Ann Appl Stat 2013;7(1):523–42.
[66] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 2018;14(6):e8124.
[67] Sompairac N, Nazarov PV, Czerwinska U, Cantini L, Biton A, Molkenov A, et al. Gorban A *et al.*: Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int J Mol Sci 2019;20(18).
[68] Nazarov PV, Wienecke-Baldacchino AK, Zinovyev A, Czerwinska U, Muller A, Nashan D, et al. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. BMC Med Genomics 2019;12(1):132.
[69] Scherer M, Nazarov PV, Toth R, Sahay S, Kaoma T, Maurer V, et al. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline. MeDeCom and FactorViz. *Nat Protoc* 2020;15(10):3240–63.
[70] Teschendorff AE, Jing H, Paul DS, Virta J, Nordhausen K. Tensorial blind source separation for improved analysis of multi-omic data. Genome Biol 2018;19(1):76.
[71] Cantini L, Bertoli G, Cava C, Dubois T, Zinovyev A, Caselle M, et al. Identification of microRNA clusters cooperatively acting on epithelial to mesenchymal transition in triple negative breast cancer. Nucleic Acids Res 2019;47(5):2205–15.
[72] Wang TT, Lee CY, Lai LC, Tsai MH, Lu TP, Chuang EY. anamiR: integrated analysis of MicroRNA and gene expression profiling. BMC Bioinf 2019;20(1):239.
[73] Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics 2009;25(23):3049–55.
[74] Lopez-Urrutia E, Bustamante Montes LP. Ladron de Guevara Cervantes D, Perez-Plasencia C, Campos-Parra AD: Crosstalk Between Long Non-coding RNAs, Micro-RNAs and mRNAs: Deciphering Molecular Mechanisms of Master Regulators in Cancer. Front Oncol 2019;9:669.
[75] Bitetti A, Mallory AC, Golini E, Carrieri C, Carreno Gutierrez H, Perlas E, et al. MicroRNA degradation by a conserved target RNA regulates animal behavior. Nat Struct Mol Biol 2018;25(3):244–51.