Elucidating the Role of microRNAs in Cancer Through Data Mining Techniques

15

Luciano Cascione, Alfredo Ferro, Rosalba Giugno, Alessandro Laganà, Giuseppe Pigola, Alfredo Pulvirenti, and Dario Veneziano

Abstract

microRNAs (miRNAs) have been shown to play a crucial role in the most important biological processes and their dysregulation has been connected to a variety of diseases, including cancer. The number of computational tools for the analysis of miRNA related data is continuously increasing. They range from simple look-up resources to more sophisticated tools for functional analysis of miRNAs. These systems may help to investigate the role of miRNAs in key biological processes and their involvement in diseases. The ultimate goal is to allow the development of regulatory models describing complex processes and the effects of their dysregulation.

Here we review the most important and recent methods for the analysis of miRNA expression profiles and the tools available on the web for target prediction and functional analysis of miRNAs.

Particular emphasis is given to the integration of heterogeneous data, including target predictions and expression profiles, which can be used to infer miRNA/phenotype associations and for the generation of network models of miRNA function.

L. Cascione • R. Giugno • A. Pulvirenti • D. Veneziano
Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy
e-mail: cascione@dmi.unict.it; giugno@dmi.unict.it; apulvirenti@dmi.unict.it; veneziano@dmi.unict.it

A. Ferro (✉)
Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

Dipartimento di Matematica e Informatica, Università degli Studi di Catania, Città Universitaria – Viale A. Doria, 6, Catania 95125, Italy
e-mail: ferro@dmi.unict.it

A. Laganà
Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA
e-mail: alessandro.lagana@osumc.edu

G. Pigola
Research and Development, IGA Technology Services, Udine, Italy
e-mail: gpigola@igatechnology.com

U. Schmitz et al. (eds.), *MicroRNA Cancer Regulation: Advanced Concepts, Bioinformatics and Systems Biology Tools*, Advances in Experimental Medicine and Biology 774,
DOI 10.1007/978-94-007-5590-1_15,

Keywords
microRNA • Database • Expression profiles • Functional analysis • Network models

15.1 Introduction

In the past decade, many efforts have been spent to demonstrate the crucial role of miRNAs in the most important biological processes, including apoptosis, development and immune response [1–3]. Moreover, the dysregulation of miRNAs has been connected to a variety of diseases, cancer being probably the most extensively studied one [4–6].

The partial complementarity that most miRNAs exhibit to their targets, especially in animals, is the key to their flexibility. Indeed, a single miRNA is usually able to bind to many targets, often in several sites, and a single gene can be targeted by different miRNAs acting cooperatively. This is a clear indication that the simple miRNA/target interactions are actually part of a more complex regulatory system and should be analyzed in the wider context of expression networks.

The initial focus of bioinformatics miRNA research was primarily on the development of tools for the identification of miRNAs and their targets. The prediction of miRNA binding sites on targets still remains a challenge. Indeed, although several studies have uncovered the basic rules of miRNA/target interactions [7], the target prediction tools currently available still produce a significant number of false positives and are not able to identify some experimentally validated miRNA/target pairs [8, 9].

Nevertheless, target prediction tools constitute the essential basis of functional miRNA analysis, allowing to link miRNAs to processes, diseases and pathways, through their targets. Recently, many bioinformatics tools for functional analysis of miRNAs have been developed. Their ultimate goal is the identification of non trivial relationships between miRNAs and other molecular actors, such as genes and transcription factors, and the development of regulatory models describing complex processes and the effects of their dysregulation. These purposes can be fulfilled thanks to the huge amount of data that are produced daily and made publicly available on the internet, among which miRNA/target matches and miRNA expression profiles are mostly predominant.

In this chapter we review the most important and recent methods for the analysis of miRNA expression profiles and the tools available on the web for functional analysis of miRNAs. In particular, in Sect. 15.2 the most used miRNA profiling technologies are described, together with the computational and statistical methods for the analysis of the related data. Emphasis is particularly given to some aspects such as data normalization, the identification of differentially expressed microRNAs, clustering and the role of miRNAs as biomarkers. Section 15.3 is focused on miRNA target prediction. An overview of the general features is given, together with a brief description of the most popular target prediction tools available on the web. Finally, in Sect. 15.4 we present a series of tools for functional analysis of miRNAs. Particular emphasis is given to the integration of heterogeneous data, including target predictions and expression profiles, which can be used to infer miRNA/phenotype associations and for the generation of network models of miRNA function.

15.2 miRNA Profiling: Technologies and Data Analysis

Several methodological approaches for mRNA expression profiling have been applied to profile miRNA expression. Current methods widely used in the study of miRNA expression include northern blotting with radiolabelled probes,

oligonucleotide microarrays, qPCR-based detection of mature miRNAs, single molecule detection in liquid phase, in situ hybridization (ISH) and massively parallel sequencing.

In this section we will review the main technologies used for miRNA profiling as well as the computational and statistical methods used for the normalization and the analysis of the produced data.

15.2.1 Profiling Technologies

In general, all existing profiling methods can be separated into two categories: one that utilizes direct oligo hybridization without sample RNA amplification and the other requiring sample amplification. A caveat to keep in mind is that there are inherent advantages and disadvantages to both approaches.

Nonetheless, three principal methods are currently used more than others to measure the expression levels of miRNAs and genes in general: microarray hybridization [10], real-time reverse transcription- PCR (qPCR) [11, 12] and massively parallel/next-generation sequencing (NGS) [13], all of which face unique challenges compared to their use in mRNA profiling. For example, the existence of miRNA families, the largest encompassing nine variants (hsa-let-7a–i), whose members differ by as little as one nucleotide but nevertheless exhibit differential expression patterns, represents a real challenge in miRNA recognition, regardless of the technology used. Microarray technology is actually based on the Watson–Crick base pairing nature of nucleic acids and thus involves nucleic acid hybridization between target molecules and their corresponding complementary probes. Synthesized antisense probes are spotted and immobilized onto a nylon support platform using a hand held spotting device. This method is relatively low cost and readily available to labs without specialized robotics and equipments dedicated to array fabrication. A disadvantage to this method is its scale. Oligo spots from a hand held device are macroscopic in nature, so the resulting array will be relatively large. About 30 mg of total RNA is commonly used to hybridize an array of this size [14]. To address this issue, automated robots have been employed to spot microscopic oligo dots onto a glass slide [15, 16].

Probes designed to differentiate between mature miRNAs and pre-miRNAs and probes that detect hypothetical miRNAs can all be spotted onto the same array. The isolated microRNAs are labeled with fluorescent dye and then hybridized with the microRNA microarray, resulting in specific binding of the labeled microRNAs to the corresponding probes. The fluorescence emission from labeled microRNAs bound at different positions on the slides can be detected. Consequently, the kinds of microRNAs and their relative quantities in the studied sample can be evaluated by analyzing the fluorescence signal data. The design of the microRNA probes, the preparation of microRNA samples and the labeling of microRNAs are considered the most important procedures in the microRNA microarray platform.

Direct hybridization of miRNA samples onto an oligo array may require a large amount of total RNA; however, some research protocols might have access to a small and limited amount of RNA—such as needle biopsies. A PCR based approach was developed to address this issue. The principle of qPCR is based on the detection, in real-time, of a fluorescent reporter molecule whose signal intensity correlates with amount of DNA present in each cycle of amplification.

In this method, total RNA is isolated as usual. The first step in qPCR of miRNAs is the accurate and complete conversion of RNA into complementary DNA (cDNA) by reverse transcription (RT). The RT reaction first consists of small RNA fractionation, followed by polyadenylation. Then a standard RT protocol is applied where poly(T)s are added to prime the synthesized poly(A) tail so reverse transcriptase can produce cDNAs from the small RNA. Finally, miRNA specific primers will probe for a specific miRNA through PCR amplification [17, 18]. The specificity and sensitivity of qPCR assays are dependent upon primer design. In fact, due to specificity issues and inability to differentiate between mature and pre-miRNA, changes have been made to the

RT step. Instead of a general poly(A) reaction in combination with universal priming through poly(T) adapter molecules, a miRNA specific stem-loop reverse primer is used. This specially designed primer contains a sequence that is antisense to a portion of the 30nt long sequence of the miRNA that is to be amplified. To increase the specificity of the PCR amplification step, the forward primer contains an antisense sequence derived from the mature miRNA, and the reverse primer consists of sequences taken from the stem-loop of the reverse primer. Sensitivity and specificity was found to be dramatically improved. In addition, the nature of specific priming allows this protocol to differentiate between the longer pre-miRNA and the shorter mature active form of the miRNA. Finally, it is claimed that this protocol can discriminate between isoforms of related miRNAs that differ by only one or two base pairs [18, 19].

The major advantages of qPCR over microarrays are (1) the speed and the sensitivity of the qPCR assays, (2) considerably larger dynamic range compared to microarray analysis and (3) a convenient requirement for low amounts of starting material (in the range of nanograms of total RNA).

However, both the RNA ligation [20] and the PCR amplification steps bear inherent biases, the method is laborious and costly, and associated tools for computational analysis are in their infancy. The reliability of miRNA expression profiling depends also on the quality of the total RNA used as input material. Robust, reproducible methods for RNA isolation and estimation of RNA quality should be employed prior to initiating the characterization of miRNA expression levels. The successful outcome of qPCR analysis depends upon a number of interconnected steps that require individual optimization. To perform qPCR that provides meaningful and reproducible results, several parameters such as RNA extraction, RNA integrity control, cDNA synthesis, primer design, amplicon detection, and data normalization must be taken into account.

qPCR is often considered a "gold standard" in the detection and quantitation of gene expression. However, the rapid increase in number of miRNAs renders qPCR inefficient on a genomic scale, and it is probably better used as a validation rather than a discovery tool.

As with genomic DNA and RNA analysis, microarrays are still the best choice for a standardized genome-wide assay that is amenable to high-throughput applications. Whole-genome screening generates a qualitative and quantitative evaluation of how experimental conditions affect miRNA profiles.

High-throughput sequencing of miRNAs, though, is coming into wider use and is unmatched for the discovery and experimental validation of novel or predicted miRNAs. The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once [21, 22]. These technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods. In particular, next generation sequencing (NGS) technologies provide a digital expression profiling readout that is fundamentally different than analog measurement systems like microarrays. A variety of different approaches are being used. They generally involve the amplification of DNA templates by PCR and the physical binding of template DNA to a solid surface or to tiny beads called microbeads. These techniques are often referred to as massively parallel DNA sequencing, because thousands or millions of sequencing reactions are run at once to greatly speed up the process. All next generation sequencing systems use clonal cluster sequencing. The process, which begins with a single target molecule, involves creation of a clonal target during an intermediate amplification step. Multiple identical copies are required to produce a high signal-to-noise-ratio.

Finally, the Nanostring technology can be used to detect any type of nucleic acid in solution and could be modified with appropriate recognition probes to detect other biological molecules as well.

Nanostring utilizes a digital technology that is based on direct multiplexed measurement of gene

expression that is capable of high level precision and sensitivity at less than one transcript copy per cell [23]. The technology uses molecular "barcodes" and single-molecule imaging to detect and count hundreds of unique transcripts in a single reaction [24]. Each color-coded barcode is attached to a single target-specific probe corresponding to a gene of interest. Mixed together with controls, they form a multiplexed assay. The degree of multiplexing is in the hundreds, which is less than that of microarrays. However, the Nanostring technology has higher throughput, accuracy and sensitivity than microarrays, which makes it preferable for low-multiplex applications, such as biomarker validation or molecular diagnostics [25].

Cancer research and biomarker validation are two of the areas where Nanostring has been most rapidly adopted. Advantages over existing platforms include direct measurement of mRNA expression levels without enzymatic reactions or bias, sensitivity coupled with high multiplex capability, and digital readout. Comparison of the Nanostring gene expression system with microarrays and TaqMan PCR demonstrated that the Nanostring system is more sensitive than microarrays and similar in sensitivity to real-time PCR [24].

Although each of these methods has their own unique advantages, they have not been perfected yet. However, at present, the method chosen for miRNA detection should best fit experience, the experimental conditions in the laboratory, and the goal of research.

15.2.2 miRNA Profiling-Normalization

The signal intensities of miRNA microarray experiments may be biased by differences in sample RNA preparation, dye labelling, hybridization and washing efficiency, peculiarities of print tip, spatial or hybridization specific effects or pre-amplification of extracted RNA. For these reasons normalization is an essential aspect of data processing.

It can minimize systematic, technical or experimental variation and thus has significant impact on the detection of differentially expressed miRNAs between two or more conditions.

Several studies pointed out that the selection of the data pre-processing method can have great impact on the resulting data outcome [26–30].

Inappropriate normalization of the data can lead to incorrect conclusions. Rigorous normalization of miRNA data may even be more critical than that of other RNA functional classes since relatively small changes in miRNA expression may be biologically and clinically significant [31, 32].

At present, there is no consensus normalization method for the three miRNA profiling approaches cited above. Several normalization techniques are similar to mRNA profiling normalization methods while others are specifically modified or developed for miRNA data. Indeed, miRNAs have some unique signatures such as their small total number and short length.

Prior to normalization, data pre-processing of miRNA profiling experiments includes platform and vendor specific steps, such as baseline adjustment and threshold setting for RT-qPCR analyses, background correction for microarray technology, or filtering for small RNA-sequence data. Following these very first steps of raw data pre-processing, one needs to choose the optimal normalization strategy to correct for systematic and technical variation enabling a better estimation of the biological variation.

15.2.2.1 Normalization Approaches for microRNA RT-PCR

RT-PCR is generally accepted as gold standard for miRNA expression measurement and normalized microRNA RT-PCR profiling data is used for the evaluation of the goodness of miRNA microarray normalization methods [27, 33].

Normalization of RT-qPCR miRNA profiling data is needed because signal intensities may depend on reverse transcription and PCR reaction efficiencies.

There are two types of sources of variation in RT-qPCR experiments. The first one is technical: there may be differences in sample procurement, stabilization, RNA extraction, reverse transcription and PCR reaction efficiencies. The second one is biological, there may be sample-to-sample

inconsistencies in cellular subpopulations or even differences in bulk transcriptional activity. For these reasons normalization of RT-qPCR miRNA profiling data is needed.

The common normalization methods for microRNA RT-PCR profiling are based on predefined invariant endogenous controls, reference miRNAs [31] or other small non-coding RNAs such as small nuclear and small nucleolar RNA [28, 34, 35].

However, in [36] the authors argued that it is best to normalize genes with reference genes belonging to the same RNA class because the use of small non-coding RNAs other than miRNAs does not mirror the physicochemical properties of miRNA molecules.

Using non-miRNA reference genes for qPCR normalization is not advisable when the overall abundance of miRNA varies, e.g., in experiments affecting the miRNA processing machinery, or in comparisons involving multiple tissues or combinations of tissues and cell lines [37].

Selection of invariant miRNAs identified by algorithms specifically developed for reference gene evaluation and selection was superior over small non-coding RNA based normalization [31, 35]. These algorithms are based on reference gene ranking and stepwise elimination of the least stable gene [36], repeated pairwise correlation and regression analysis [38] or statistical linear mixed-effects modelling [39] of the respective experimental data.

Moreover invariant miRNAs can be selected based on a distinguishable low standard deviation and high-mean population as suggested by Pradervand et al. [28] for miRNA microarray preprocessing and this approach is applicable for RT-qPCR profiling experiments as well. Basically, the use of more than one reference gene increases the accuracy of quantification compared to the use of a single reference gene [36, 39].

Commonly used methods for miRNA raw data processing use median or mean value of the raw readings as normalization factor. However, many miRNAs may not be expressed in a biological sample, and thus median or mean value may be skewed towards the assay readings for lowly expressed miRNAs, which tend to be more variable compared with the readings for more abundantly expressed miRNAs. A scaling method suggested by Wang et al. [40], uses the average expression values of eight selected miRNAs with relatively high expression from a descending sorted list.

For large scale microRNA expression profiling studies the mean expression value normalization outperformed the current normalization strategy that makes use of stable small RNA controls, such as snoRNAs proposed by manufacturers, in terms of better reduction of technical variation [35].

However, the selection of a limited number of miRNAs or small RNA controls that resemble the mean expression value can be successfully used for normalization in follow-up studies where only a limited number of miRNA molecules are profiled to allow a more accurate assessment of relevant biological variation from a miRNA RT-qPCR profiling experiment [32, 35].

15.2.2.2 Normalization Methods for miRNA Microarray Experiments

Different normalization methods have been used on miRNA microarray expression profiling data sets, but there is currently no clear consensus about their relative performances [28].

Some have even chosen to omit normalization [41–43] but comparative studies on the relative performance of different normalization methods within a miRNA microarray platform have emphasized the need for evaluating and identifying appropriate normalization methods [27, 28, 44]. miRNA microarrays can be single-color or dual-color systems calling for different normalization approaches. Single-color miRNA microarrays have been predominately used, while dual-color hybridization systems are less frequently prevalent [44].

Both can be observed with respect to intra-array normalization for the correction of dye effects and inter-array approaches for the balance of the distribution differences among experiments [45].

The first normalization methods to be used with miRNA array data employed centring to median values [46–48] or scaling based on total array intensities [49, 50].

Certain methodologies currently used for large-scale genome arrays have been adapted to and modified for miRNA arrays such as Quantile [51] and LOESS (Locally Weighted Regression and Smoothing Scatterplots) [52]. Various assumptions are often taken by several normalization methods. Scaling, LOESS and Quantile [26, 27] are based on two assumptions, (i) only a small portion of spots is differentially expressed, and (ii) differentially expressed spots are homogeneously distributed with respect to both, over- and under-expressed miRNAs [29].

However, these assumptions could fail for miRNA platforms as they are printed with a relatively small number of selected sequences [27, 29]. Moreover, the number of expressed miRNAs in a miRNA microarray profiling is small (typically in the order of hundreds) compared with a few thousands of genes [30]. Hence, among the expressed microRNAs the proportion of those that are differentially expressed is much larger than that observed in mRNA expression profiling [30]. Experiments with most miRNAs differentially expressed predominantly in one direction, that is only up- or down-regulated, are not unusual.

Thus, it must be verified whether these assumptions hold true for the respective datasets and one should choose a normalization method that makes only minimal assumption about the presence of a set of constant miRNAs, like invariant-based normalization [28]. Alternatively, a normalization method free of assumption, the majority of algorithms for variance stabilization normalization [53] or even an assumption free approach [54] can be utilized instead.

Quantile normalization is a transformation method originally proposed by Bolstad et al. [51] for oligonucleotide arrays. It is now widely used for one-color miRNA microarrays as well and was confirmed as one of the most robust methods [27, 28, 44, 55]. It is an inter-array approach and equalizes the distributions of expression intensities across arrays. Thus, quantile normalization assumes that the overall distribution of signal intensity does not change. While this assumption likely holds true for the comparison of p53 overexpressing versus control cells [28] or even for brain–heart comparisons according to Rao et al. [44] where only 5% of miRNAs were differentially expressed, it may not hold true in case large numbers of miRNAs are differentially expressed in only one direction.

Such cases may be, for example, the knockout of proteins essential for miRNA biogenesis, which lead to a dramatic reduction in steady state miRNA levels by blocking production of mature miRNAs [44].

Rao et al. [44] compare the performance of several normalization methods on miRNA single channel microarray profiling, showing the better performance of quantile normalization.

Quantile normalization can be applied to dual-labeled array data if red and green channels are treated as two independent single-labeled array data. On the contrary two single-labeled array data can be considered as a dual-labeled data and LOESS normalization may be used in this case [56].

For the two colors microarray data, normalization is usually applied to the log-ratios of green channel signal (Cy3) and red channel (Cy5) signal, which will be written as M and A .

The LOESS normalization and its variants [27, 29, 44] are the most used transformation based methods. They use local regression via locally weighted scatter plot smoothing. M is defined as the log transformation of Cy3/Cy5 and A as the log transformation of the squared root of Cy3*Cy5 (as used in the MA-plot). It is advisable to introduce weights that penalize outliers because outlier values can strongly influence the local regression curve (LOWESS). However, Lowess and Loess are treated as synonyms. Local regression via LOESS uses a quadratic polynomial weighted regression function with Tukey's biweight function [52] of the log ratios of the Cy3 and Cy5 signals on overall spot intensity of the two signals (Cy3*Cy5).

In addition to intensity-dependent variation in log ratios, spatial bias could also be a significant source of systematic error (print-tip effect). It is possible to correct for both print-tip and intensity-dependent bias by performing a within-print-tip-group normalization using LOESS.

Print-tip LOESS normalizes each M value by subtracting from it the corresponding value on

the tip-group LOESS curve [27]. Finally, the normalized log-ratios (N) are:

$$N = M - \text{loess}_i(A)$$

where $\text{loess}_i(A)$ is the loess curve as a function of A for the *ith* tip group.

However, Sarkar et al. [30] did not find significant differences between print-tip LOESS and other normalizations.

Hua et al. [27] compared 15 normalization methods using microarray data and RT-PCR data. It was found that microRNA normalized data by print-tip LOESS method were most consistent with the RT-PCR results.

A variant of LOESS normalization called LOESSM was proposed by Risso et al. [29]. This non-parametric normalization scales the expression data on the global median expression rather than on zero. This modification relaxes the assumption of symmetry among up- and down-regulated genes and it was shown that LOESSM, in case of absence of channel-effect, has better performance. In addition, LOESS combined with Generalized Procrustes Analysis (GPA), an assumption free inter-array normalization [54], improved its results and outperformed the other normalizations in terms of sensitivity and specificity [29].

LOESS normalizations and its variants emerged as being robust in the reduction of non-biological bias.

Variance stabilization normalization (VSN), an inter-array transformation method, is widely used for microRNA microarray data [28, 30]. It was developed for mRNA arrays and is based on a parameterized arcsinh transformation instead of a logarithmic transformation that calibrates sample-to-sample variations and renders variance approximately independent of the mean intensity [53].

Spike-in VSN normalization restricts the model fit to spike-in spots. These spots recognize specific RNA transcripts that can be added as internal controls in the experiments. Normalization intensities for all miRNAs are then obtained by applying the resulting transformation to all spots of interest on the array [30].

One limitation of this approach is that reliable results can only be obtained for intensities within the range covered by the spike-in used and that excludes targets that are not expressed.

Pradervand et al. [30] proposed a linear regression method to select a set of miRNAs with constant expression (invariants) and used these invariants to calculate VSN parameter (VSN-INV). The invariant probes are those that have medium-high mean intensity and low variance across samples. VSN used with default parameter settings assumes that most genes are not differentially expressed whereas the invariant-based regression only assumes that a subpopulation of expressed genes does not change. So, VSN-INV is appropriate only if a significant fraction of miRNAs is expected to be differentially expressed.

Based on theirs comparisons, Pradervand et al. found that VSN-INV and quantile normalization were the most robust normalization methods compared to VSN with default parameter or scaling. In general, one should note that VSN strongly affects the distribution of the large fraction of miRNAs whose expression is near or at background, resulting in the large increase of variability for those microRNAs.

15.2.2.3 Scaling Normalization

The first normalization methods for mRNA microarray were based on the selections of predefined and stably expressed housekeeping genes, as described by Garzon et al. and Perkins at al. [57, 58] that uses all probes. These methods have been applied to one- or two-channel miRNA microarray profiling. Most commercially available miRNA microarrays do not have controls for endogenous RNAs that have been shown to be robustly invariant between various different tissue samples or conditions [44]. To date, there is no consensus on the existence and reliability of reference gene miRNAs. The selection of reference genes to normalize miRNA levels depends on the bioinformatics analysis of the respective data (as shown for mRNA in [36, 39]) and is otherwise still rather empirical due to the lack of robust reference miRNAs [34], although a universal reference miRNA reagent set has been proposed [30].

Bargaje et al. [55] identified constitutively expressed miRNAs across tissues. A mean of expression levels of a set of 16 microRNAs showing

minimum variability, was reasonably successful as a normalization factor for comparing datasets generated by the same platforms. However, normalization using constitutive microRNAs was ineffective when comparing bead-based and microarray-based datasets. In these cases quantile and Z-score normalization were both successful in transforming the data sets generating comparable means and scale.

The scaling methods like Z-score, mean, median, or 75th percentile assume that different sets of intensities differ by a constant global factor and all raw intensity values are multiplied with one common (i.e., global) scaling factor [26, 27, 55]. The Z-score provides a mean-centered rank for the expression level in units of standard deviation. Z-scores thus provide an index of the expression level of the miRNA with respect to the cellular pool of miRNA. Unlike other normalization methods, Z-scores are not influenced by the addition of new datasets allowing flexible cross-platform validation of miRNA microarray profiling experiments [55].

Recently, Wang et al. [40] suggested the pre-evaluation of the overall miRNA expression pattern by a panel of miRNAs using RT-qPCR assays to build a logistic regression model based on these results. The personalized logistic regression model based on 29 miRNAs efficiently calibrated the variance across arrays and improved miRNA microarray discovery accuracy compared with different scaling methods, LOESS or quantile normalization [40].

15.2.3 Identification of Differentially Expressed Genes and miRNA

Several methods have been applied to the identification of differentially expressed genes and microRNA in microarray data.

The simplest method is to evaluate the log ratio between two conditions (or the average of ratios when there are replicates) and consider all the genes that differ by more than an arbitrary cut-off value to be differentially expressed. This is not a statistical test, and there is no associated value that can indicate the level of confidence in the designation of genes as differentially or not differentially expressed.

It is considered to be unreliable [59] because statistical variability is not taken into account and is susceptible to outliers.

More sophisticated statistical methods have been proposed. The classification success is affected by the choice of the method, the number of genes in the genelist, the number of cases (samples) and the noise in the dataset.

Different methods produce dissimilar gene lists, which can produce dramatically different discrimination performance when trained as gene classifiers.

The gene lists produced by the feature selection methods can be grouped broadly according to the manner in which they treat gene variance.

15.2.3.1 t-Statistic

The simplest statistical method for detecting differential expression is *t* test. It can be used to compare two conditions when there is replication of samples. With more than two conditions, analysis of variance (ANOVA) can be used.

The t-test calculates the observed t-statistic for each gene. The idea is to compare between-group difference and within-group difference and then to calculate the probability value (p-value) of t-statistic for each gene from t-distribution.

The output of the analysis is a p-value for each gene. It represents the chance of getting the t-statistic as large as, or larger than the observed one, under the hypothesis of no differential expression (null hypothesis). A small p-value indicates that the hypothesis of no differential expression is not true and the gene is differentially expressed.

15.2.3.2 SAM

Several modified t-statistics have been proposed to address this problem. SAM [60] is one of the most popular. It performs moderately well except when applied to data with low sample size and to the noisy datasets.

SAM uses a moderated t-statistic, whereby a constant is added to the denominator of the t-statistic. The addition of this constant reduces the chance of detecting genes which have a low standard deviation by chance. The constant is estimated

from the sum of the global standard error of the genes [61–63].

15.2.3.3 Empirical Bayes Method (Limma)

The empirical bayes method provides a more complex model of the gene variance. The gene standard error is estimated as a representative value of the variance of the genes at the same level of expression as the gene of interest [64]. In training sets with a large number of cases, the empirical bayes method performed comparably with ANOVA. Importantly, unlike most other methods, the empirical bayes t-statistic proved equally robust with low numbers of cases. The Bayesian statistic also provides p-values and has the advantage that it can be expanded to deal with datasets that have more then two classes.

Limma provides advanced statistical methods for linear modelling of microarray data and for identifying differentially expressed genes. It fits a linear model to the data and uses an empirical Bayes method for assessing differential expression [65]. One or two experiment definition matrices need to be specified during the analysis: a *design matrix* defining the RNA samples and a *contrast matrix* (optional for simple experiments) defining the comparisons to be performed.

When there are more than two conditions in an experiment, a more general concept of relative expression is needed. One approach that can be applied to cDNA microarray data from any experimental design is to use an analysis of variance model (ANOVA) to obtain estimates of the relative expression (*VG*) for each gene in each sample [66, 67]. In the ANOVA model, the expression level of a gene in a given sample is computed relative to the weighted average expression of that gene over all samples in the experiment.

The microarray ANOVA model is not based on ratios but it is applied directly to intensity data; the difference between two relative expression values can be interpreted as the mean log ratio for comparing two samples (as $\log A - \log B = \log(A/B)$, where $\log A$ and $\log B$ are two relative expression values). Alternatively, if each sample is compared with a common reference sample, one can use normalized ratios directly. This is an intuitive but less efficient approach to obtain relative expression values than using the ANOVA estimates. Direct estimates of relative expression can also be obtained from single-color expression assays [68].

The set of estimated relative expression values, one for each gene in each RNA sample, is a derived data set that can be subjected to a second level of analysis. There should be one relative expression value for each gene in each independent sample. The distinction between technical replication and biological replication should be kept in mind when interpreting results from the analysis of a derived data. If inference is being made on the basis of biological replicates and there is also technical replication in the experiment, the technical replicates should be averaged to yield a single value for each independent biological unit. The derived data can be analyzed on a gene-by-gene basis using standard ANOVA methods to test for differences among conditions.

15.2.3.4 ROC

Classifiers built using gene lists from the ROC method outperform all other methods when applied to large datasets. High RCI scores are observed even when only a few of the most highly ranked genes are examined. These high RCI scores are maintained when the number of genes examined is increased. It is possible to obtain p-values using this method [69]. ROC, like the t-statistic methods, loses power when the number of samples is reduced. It ranks a gene based on its power to discriminate between the groups given a threshold false positive rate. This means that it ignores the level of expression of the gene in the two groups. Therefore as the training size decreases, the likelihood of a gene with low variance and no biological meaning being a good discriminator by chance increases. ROC is an unsuitable method when the sample size is below 30 (class size of 15).

15.2.3.5 Rank Product

The Rank Product [70] package contains functions for the identification of differentially expressed

genes using the rank product non-parametric method described in [63]. It generates a list of up- or down-regulated genes based on the estimated percentage of false positive predictions (pfp), which is also known as false discovery rate (FDR). The attractiveness of this method is its ability to analyse data sets from different origins (e.g. laboratories) or variable environments.

Rank product assumes constant variance across all samples. It compares the product of the ranks of genes in a class with the product of the ranks of genes in the second class. For each gene in the dataset, rank products sorts the genes according to the likelihood of observing their ranked positions on the lists of differentially expressed genes just by chance.

15.2.4 Clustering

Clustering algorithms are widely used in the analysis of microRNA profiling data. In clinical studies, they are not only used to cluster microRNA into groups of co-regulated miRNA, but also for clustering patients, and thereby defying novel disease entities based on miRNA expression profiles.

A reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer.

Current methods for classifying human malignancies rely on a variety of morphological, clinical, and molecular variables. In spite of recent progress, there are still uncertainties in diagnosis. Also, it is likely that the existing classes are heterogeneous and comprise diseases which are molecularly distinct and follow different clinical courses. microRNA microarray datasets have been used to characterize the molecular variations among tumors by monitoring microRNA expression profiles on a genomic scale. This led to more reliable classification of tumors and to the identification of marker miRNA that distinguish among these classes. Eventual clinical implications include an improved ability to understand and predict cancer survival. However, there are three main types of statistical problems associated with tumor classification:

- The identification of new tumor classes using microRNA expression profiles – unsupervised learning;
- The classification of malignancies into known classes – supervised learning
- The identification of marker microRNA that characterize the different tumor classes – feature selection.

Clustering can answer these problems. It is possible to cluster rows, columns or both. Rows (miRNA) clustering can identify groups of co-regulated miRNA, spatial or temporal expression patterns, reduce redundancy (cf. feature selection) in prediction, and detect experimental artefacts. On the other hand columns clustering allows to identify new classes of biological samples, new tumor classes or new cell types. Moreover, it allows to detect experimental artefacts.

In order to perform clustering, a way to measure how similar or dissimilar two objects are is needed. The feature data are often transformed to an $n \times n$ distance or similarity matrix, $D = (dij)$, for the n objects to be clustered. Features correspond to expression levels of different microRNAs and possible classes include tumor types or clinical outcomes (survival, non-survival). Other information such as age and sex may also be important and can be included in the analysis. The most popular distances are Euclidean distance and Manhattan distance. Hamming distance is used for ordinal, binary or categorical data.

Clustering procedures can be divided into three categories: Hierarchical, Partitioning (K-means K-medoids/partitioning around medoids) and Model based approaches. The first one is either divisive or agglomerative and provides a hierarchy of clusters, from the smallest, where all objects are in one cluster, through to the largest set, where each observation is in its own cluster. One must often also define a distance measure between clusters or groups of miRNA and the linkage methods used are single, complete, average, distance between centroids and Ward Linkage. Hierarchical clustering methods produce a tree or dendrogram. The partitions are obtained from cutting the tree at different levels. The tree can be built in two distinct ways bottom-up (agglomerative clustering)

or top-down (divisive clustering). Examples of Hierarchical clustering methods are Self-Organizing Tree Algorithm – SOTA [71] and DIvisive ANAlysis – DIANA [72].

Partitioning methods require the specification of the number of clusters. A mechanism for apportioning objects to clusters must be determined, then data is portioned into a prespecified number K of mutually exclusive and exhaustive groups and iteratively reallocated to clusters until some criterion is met, e.g., minimize within-cluster sums-of-squares. Examples of partitioning methods are k -means and its extension to fuzzy k -means, Partitioning Around Medoids – PAM [72], – Self-Organizing Maps – SOM [73] and model-based clustering, e.g., Gaussian mixtures in [74–76] and McLachlan et al. [77, 78].

An important feature of partitioning methods consists in satisfying an optimality criterion (approximately), however they need an initial K and long computation time. Hierarchical methods are computationally fast (for agglomerative clustering) but rigid, since they cannot later correct for earlier erroneous decisions.

Most methods used in practice are agglomerative hierarchical methods. In large part, this is due to the availability of efficient exact algorithms that implement them.

Model based approaches assume that data are 'generated' from a mixture of K distribution. They try to fit a model to the data and try to get the best fit. A classic example is a mixture of Gaussians (mixture of normals). They take advantage of probability theory and well-defined distributions in statistics.

In microarray experiments is also useful to detect the presence of outliers. Outlier detection is an important step since they can greatly affect the between-cluster distances. Simple tests for outliers should be identifying observations that are responsible for a disproportionate amount of the within-cluster sum-of-squares.

Most features in high dimensional datasets will be uninformative, examples are unexpressed genes, housekeeping genes, 'passenger alterations'. Clustering (and classification) has a much higher chance of success if uninformative features are removed. Simple approaches to feature selection are: selecting intrinsically variable genes or requiring a minimum level of expression in a proportion of samples.

Clustering can be also employed for quality control purposes. The clusters that are obtain from clustering samples/microRNA should be compared with different experimental conditions such as batch or production order of the arrays, batch of reagents, microRNA amplification procedure, technician, plate origin of clones, and so on. Any relationships observed should be considered as a potentially serious source of bias.

15.2.5 miRNA as Biomarkers

miRNAs have a very important role in cancer. Their expression is often dysregulated in malignant cells. Some miRNAs that are temporarily over-expressed in early development and shut off in the normal differentiated state may re-express in cancer, causing a persistent stem cell–like dedifferentiated state. Many miRNAs may act like oncogenes by promoting proliferation and/or repressing apoptosis. Other miRNAs play the role of tumor-suppressors. They have a regulatory function in normal tissues but when they are down-regulated in cancer, they abrogate their tumor-suppressor activity.

Over-expression or lack of expression of specific miRNAs appears to correlate with clinically aggressive or metastatic phenotypes [79, 80].

miRNA expression has tissue specificity and has been used for identifying the tissue in which cancers of unknown primary origin arose [81]. Rosenfeld and colleagues constructed a miRNA-based tissue classifier by measuring miRNA expression levels using a microarray platform in 336 primary and metastatic tumors representing 22 different cancer types. They built and tested a classifier for 48 miRNAs that accurately predicted tissue type in 86% of the test set, including 77% of the metastatic samples. Moreover, the classifier predicted tissue type with 100% accuracy for six of the ten tumor types in the metastatic test set. The authors proposed that their classification system could be

applied to cancer of unknown primary origin, defined as histologically confirmed metastatic cancer for which no primary site of disease can be identified.

Cancer classifications previously determined by mRNA expression profiling are now being investigated with miRNAs. One study has directly compared mRNA and miRNA microarray expression data and shown that known molecular subtypes of breast cancer can also be identified using miRNAs, and that expression of processing enzymes and proteins involved in miRNA biogenesis are down-regulated in the more aggressive subtypes [82]. Clinical trials are underway that test different therapies in different breast cancer molecular subtypes as defined by mRNA expression.

Mitchel et al. [83] discovered microRNAs in healthy human plasma that can be traced back to specific tissue (miRNA-15b, miRNA-16, and miRNA-24). In addition, they found that serum is more readily available than plasma and the stability of miRNA compared to the plasma is strongly positively correlated.

They found the baseline levels of miRNA expression in healthy individuals and detected the levels of prostate cancer-expressed miRNAs (miRNA-100, miRNA-125b, miRNA-141, miRNA-143, miRNA-205, and miRNA-296) in serum. miRNA-141 level is specifically elevated in prostate cancer in serum and several experiments illustrated that miRNA-141 is expressed by several common human cancers. They established that tumor-derived miRNAs can be detected in plasma or serum and serve as an effective circulating biomarker of common human cancer types.

A lot of benefits will come from using miRNA to diagnose cancer. miRNA is 97.6% accurate for sensitivity as a biomarker for cancer and 96.3% accurate as a biomarker for the classification of cancer [84]. It means less false positive or false negative cases. It will decrease the delay in diagnosis of cancer because a blood test with miRNA assay or electrophoresis is much cheaper and sufficient for diagnosis.

It will avoid invasive, expensive and/or unnecessary tests to find out if a patient has cancer and what type of cancer. All this will get patients less stressed.

Significant progress has been made on the relationship between miRNAs and cancers and the important function of miRNAs in a variety of cancers has been reviewed by several research groups.

In fact as shown by Lu et al., the miRNA expression profile based on the expression of only 200 miRNA genes successfully classified poorly differentiated tumors confirming in the majority of cases the clinical diagnosis whereas mRNA profiling, based on the expression of about 16,000 protein coding genes, failed to do so [85].

Visone et al. found out that miR-181b is a unique biomarker for CLL since its expression can be monitored throughout the disease course of a patient and this change in the leukemic cells correlate with the overexpression of four genes with great significance in CLL and other cancers (i.e. MCL1, TCL1, BCL2 and AID). Collectively, this information together with the analysis of stable prognostic markers (e.g. ZAP-70 and IGHV mutation status) specify disease progression in chronic lymphocytic leukemia and is associated with clinical outcome [86].

Finally, a significant justification for using miRNAs as biomarkers is that miRNAs have an unusually high stability in formalin-fixed tissues, which means that the miRNA can be stored and extracted with minute degradation. Short miRNAs from older tumors preserved as formalin-fixed paraffin-embedded tissue are less susceptible to chemical modification and degradation over time and have proven satisfactory for miRNA analysis.

15.3 miRNA Target Prediction

In order to determine miRNA functions it is fundamental to find their targets. While miRNA target prediction in plants is rather simple, due to the perfect complementarity that plant miRNAs usually exhibit to their targets, the prediction of miRNA binding sites in animals is much more challenging. In fact, perfect complementarity in

animals is usually limited to the 5′ end of the miRNA, which is usually referred to as the seed (~6–8 nt long) [87]. The target sites are usually located in the 3′ UTR sequences of mRNAs. In order to significantly reduce the number of false positives, other determinants are needed due to the fact that the short length of the miRNA seeds raises the probability of finding random matches that don't correspond to functional sites [7]. Such determinants or rules should be primarily inferred from experimentally verified targets, thus showing how important it is to have good sources of data as the basic step in the development of prediction tools. A significant amount of miRNA/target interactions data, usually coming from the literature, is publicly available on web databases, such as Tarbase [88] and miRecords [89]. Information on the binding sites of miRNAs in their verified targets is usually provided by this data. Moreover, high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) has recently identified functional RISC interaction sites on mRNAs, allowing the creation of libraries of reliable miRNA binding sites [90]. Data Mining analysis of these sequences could help identifying important discriminant features for the prediction of new binding sites.

In predicting functional targets, miRNA/target interaction rules are generally not sufficient due to the high number of false positives that derive from random matches of the short seed region of miRNAs to false targets. Consequently, other kinds of data are needed to improve prediction algorithms. For instance, target conservation is widely used as a valid additional criterion. High sequence conservation is indeed revealed by the alignment of miRNAs in different species, especially in the seed regions, which often corresponds to high conservation of their targets. Therefore, an help in detecting functional sites could come from the identification of conserved regions in the 3′ UTR of a gene, even though this approach is not useful in the case of non-conserved miRNAs [91]. Several prediction methods exploit thermodynamics properties. Free energy (ΔG) can be used to evaluate the stability of the predicted duplexes. Low values of free energy, usually below −20 kcal/mol, characterize indeed all validated miRNA/target pairs [92]. A low energy value, however, is a necessary but not sufficient condition. Not all energetically favourable miRNA/target duplexes, in fact, are functional. Structural accessibility of the target molecule is another thermodynamic feature used by computational methods. miRNA binding sites shouldn't be involved in any intra-molecular base pairing, and any existing secondary structure should be disrupted in order to make the site accessible to the miRNA [93]. This very complex problem mostly relies on secondary structure prediction computation, which is still one of the challenges of computational biology [94].

Nucleotide composition surrounding the binding sites and the position of the sites in the UTR, as well as the presence of multiple sites on the same UTR, are additional features used by prediction tools. In fact, it is proven that a single miRNA can have more binding sites on the same target and that a target can have multiple sites for different miRNAs [95].

15.3.1 Tools for the Prediction of miRNA Targets

Several computational tools for the prediction of miRNA targets are currently available on the web [96].

In this subsection we will review the basic concepts behind the most popular ones: TargetScan, miRanda, Pictar, Diana-microT, RNA22, RNAHybrid, StarMir and PITA.

TargetScan is one of the most popular tools for miRNA targets prediction. It's a sophisticated algorithm based on both conservation and base pairing rules [91, 95] that searches for miRNA seed matches on UTRs, considering different kinds of seeds and making use also of secondary structure prediction in order to calculate the free energy of the predicted duplexes. In addition, it considers the presence of multiple sites for the same miRNA on a target as positive contribution to the score of the prediction. Through sequence alignment, TargetScan also takes into account the conservation on different species for the

identification of the most probable targets. All the predictions, computed for different species like human, mouse and rat, are available on the TargetScan website.

miRanda is another web tool that performs predictions and it's based on an alignment algorithm which uses a weighted matrix aimed at promoting the binding of the seed of the miRNA rather than its 3′ end. It also uses the free energy of predicted duplexes and the conservation criteria to select the most probable targets. Its website allows predictions on human, mouse and rat [97, 98].

Another popular tool for the prediction of miRNA targets on vertebrates, nematodes and flies is PicTar [99]. Its algorithm is trained to identify binding sites for a single miRNA and multiple sites regulated by different miRNAs acting cooperatively. It makes use of a pairwise alignment algorithm in order to find sites conserved in many species (7 Drosophila species and 8 vertebrate species), considering also the clustering and co-expression of miRNAs together with ontological information, such as the time and tissue specificity of miRNAs and their potential targets, to enhance its predictions.

Diana-MicroT implements an algorithm that is trained to identify targets with a single binding site for a miRNA [100]. Its sequence alignment algorithm focuses on the search for miRNA/target duplexes characterized by central bulges and paired 5′ and 3′ ends.

A different approach is instead adopted by the web tool RNA22. It performs the analysis of miRNA sequences to find intra- and inter-species patterns of conserved sequence features [101]. The algorithm generates the reverse complement of the most significant patterns and searches for their instances in the UTRs in order to identify the target islands supported by a minimum number of pattern hits. A target island is defined as any hot spot where the reverse complement of mature miRNA patterns aggregate. It then computes the pairing of each target island with each candidate miRNA and evaluates the thermodynamic stability of the duplex obtained.

The miRNA target prediction tool RNAHybrid is conceived as an extension of the RNA secondary structure prediction algorithm by Zuker and Stiegler to two sequences [92]. Hybridization of the miRNA to the target is considered through an energetically optimal criterion, i.e. yielding the Minimum Free Energy (MFE), but absolutely avoiding intra-molecular base pairing and multi-loops. The algorithm used adopts dynamic programming, forcing the perfect match of the seed. Bulges and internal loops are restricted to a constant maximum length in either sequence.

The computation of the structural accessibility of the targets is instead the main feature of the tools StarMir and PITA. StarMir is based on the target's secondary structure as predicted by the tool Sfold [102]. The miRNA/target interaction is modelled as a two-step hybridization reaction: the nucleation at an accessible site and the hybrid elongation to disrupt the local secondary structure of the target and form the complete duplex. PITA is based on a slightly different model which computes the difference between the free energy gained from the formation of the miRNA/target duplex and the energetic cost of unpairing the target to make it accessible to the miRNA [103].

In spite of the rather successful predictions of effective miRNA targets performed by the tools mentioned above, the problem still remains a big challenge. The high number of false positives and the use of conservation criteria clearly show our partial knowledge in the targeting mechanisms. Combining Data Mining, Pattern Discovery and Machine Learning techniques together with thermodynamics and the availability of more reliable experimental data, will allow the improvement of predictions and enhance our knowledge and understanding of RNAi.

15.4 Functional Annotation of miRNAs

As discussed in the previous section, our knowledge about the molecular rules that underlie miRNA targeting is still incomplete, hence the huge number of false positives that target prediction tools can produce. Functional analysis of miRNAs may help to identify the most probable targets and to uncover non trivial relationships between miRNAs and other molecular actors,

such as genes and transcription factors, allowing the development of regulatory models describing complex processes and the effects of their dysregulation. There are several tools available online which collect and integrate miRNA-related data retrieved from different sources in order to infer miRNA functions.

In this section we are going to describe the most popular tools for functional analysis of miRNAs, that we divided in three categories: tools for miRNA/phenotype associations, tools integrating target prediction with expression data and tools for the generation and the analysis of network models of miRNA function.

15.4.1 miRNA/Phenotype Associations

Several tools provide users with manually curated information about the involvement of miRNAs in diseases and biological processes. Some of them also make use of computational predictions, statistics and data mining features in order to filter the data and infer new knowledge.

miR2Disease and the Human microRNA Disease Database are manually curated databases based on experimental data. They aim at providing a comprehensive resource of miRNA deregulation in various human diseases [104, 105]. These web based tools offer user friendly interfaces to query the information on miRNA/disease relationships. miR2Disease also allows researchers to contribute to the data contents through a submission page.

The authors of the Human microRNA Disease Database performed some analysis on their dataset and found that there is a negative correlation between the tissue-specificity of a microRNA and the number of diseases associated to it. They also found that miRNAs that are close in the genome, like members of the same clusters, are often associated with the same diseases. This suggests that neighboring miRNAs might be regulated by common regulators, and that they might regulate different genes involved in the same pathways. Finally, the analysis revealed that miRNAs which are conserved in other species, tend to be significantly associated with diseases with a higher probability.

miReg is also a manually curated miRNA Regulation Resource that provides users with regulatory relationships among validated upstream regulators like transcription factors or drugs, downstream targets, associated biological processes, experimental conditions or disease states and dysregulation of the miRNA in those conditions [106]. All the collected data is described in the literature and the corresponding references are provided together with other useful links about the studied miRNAs. The website has a user-friendly interface browseable through different options.

A further step in the integration of heterogeneous information about miRNA is miRo', a web environment that provides users with miRNA–phenotype associations in humans [107]. It integrates data from various online sources, such as databases of miRNAs and targets, Gene Ontology terms and diseases into a unified database equipped with a flexible query interface and data mining facilities. miRo' allows both simple and advanced queries and introduces a new layer of associations between genes and phenotypes inferred based on miRNAs annotations.

miRNAs are connected to diseases, GO processes and functions through their validated and predicted targets (miRecords, miRanda, PicTar, TargetScan) [89, 91, 97, 99, 108].

The simple search allows the selection of a single miRNA, process, function, disease or tissue and quickly displays the corresponding information. By selecting a miRNA, for example, the user can obtain the list of diseases, processes and functions in which the miRNA is potentially involved through its targets. Moreover, a list of tissues expressing the miRNA and the corresponding expression values is given. These are obtained from the Mammalian microRNA Atlas [109].

Similarly, by selecting a process, a disease, a function or a tissue, the user obtains a list of miRNAs related to the selected item. In all cases, detailed information about the miRNAs, the targets and the source of predictions are given, together with links to the original data sources.

The advanced search allows users to perform more complex queries through the introduction of specific constraints that data must satisfy. For

example, it is possible to search for all the miRNAs which are involved in a group of diseases and processes or for all the diseases related to a group of miRNAs, genes, processes and functions. The results are given in a table with details about the miRNA/target predictions. Furthermore, this advanced query tool allows to identify new potential associations between diseases, processes and functions inferred based on miRNA annotations. For example, a disease d and a process p which are not linked through any common gene might be associated through a miRNA which regulates a gene gd, involved in d, and a gene gp, involved in p.

miRo' is also equipped with a special Data Mining module which allows clustering of miRNAs that are associated to the same set of terms. Chosen a set of up to five miRNAs and an association criteria (i.e. process or disease), the system will find all the subsets of the selected miRNAs which are closely associated to groups of processes or diseases. This feature may help to identify a set of miRNAs acting cooperatively to carry out certain biological functions. Moreover, a specificity score allows to evaluate the relationships between the miRNAs and their annotation terms.

In a similar way, the tool FAME uses computational target predictions in order to automatically infer the processes affected by human miRNAs [110]. The website provides a simple menu for retrieving offline computed data. By choosing a miRNA from the list, the user obtains two tables reporting the most significantly associated Gene Ontology processes and KEGG Pathways, respectively.

For each miRNA-process/miRNA-pathway association, a score, a p-value, a q-value and an enrichment factor are given, together with the list of target genes involved in the process/pathway.

In the paper, the authors used their method to identify 68 miRNA families and 27 genomic clusters regulating 21 gene co-expression clusters in diverse human stem cell lines. They found out that clusters enriched with the targets of a specific miRNA tend to be anti-correlated with the miRNA expression, whereas clusters depleted of miRNA targets are co-expressed with it.

15.4.2 miRNA Target Prediction Consensus and Gene Expression Data Integration

Most of the available tools for miRNA functional analysis make use of heterogeneous information, and their classification into categories, based on their purposes and the kind of data that they use, is not an easy task. However, there is a well distinct class of tools which make use of miRNA and gene expression data, either retrieved from public sources or provided by users. As discussed in Sects. 15.2 and 15.3, miRNA and gene expression profiling is an important source of information in the study of miRNA functions. In this section we introduce miRonTop, MAGIA and Diana-miRExTra, three tools that combine target prediction with expression data.

miRonTop is an online application allowing the detection of miRNAs that significantly affect gene expression at a large scale [111]. It is a java web tool that integrates DNA microarrays or high-throughput sequencing data with target predictions in order to identify the potential implication of miRNAs on a specific biological system.

Users have to provide a table summarizing a large-scale gene expression study in a tab-delimited file, and select the prediction software to be used, among miRbase, miRanda, TargetScan, PicTar or the exact seed (7-mer/8-mer) match.

The program then performs an enrichment analysis of the predicted targets, for each miRNA considered in the expression table, according to the selected prediction tool across the DOWN and the UP gene sets. The significance is evaluated using the hypergeometric distribution.

MAGIA is a web-based tool which allows to retrieve and browse miRNA target predictions for human miRNAs, based on a number of different algorithms (PITA, miRanda and TargetScan), setting cuttoffs on prediction scores, with the possibility of combining them with Boolean operators [112]. The query output is a table including the list of predicted target genes or transcripts with different prediction scores according to the methods chosen by the user. For each prediction several external links are provided.

The tool also includes an analysis framework. Given as input miRNA and gene expression profiles (MATCHED or UNMATCHED expression data) it provides different statistical measures of profiles relatedness and algorithms for expression profiles combination.

For unmatched expression data, MAGIA employs a meta-analysis approach based on a p-value combination, while one of four different measures of relatedness (Spearman and Pearson correlation, mutual information, and a variational Bayesian model) can be adopted for the analysis of matched profiles.

The results are reported in a web page containing different sections. For the top 250 most probable functional miRNA–mRNA interactions according to the association measure selected by the user, the interactive bipartite regulatory network obtained through the analysis is reported along with the corresponding browsable table of relationships.

Finally, Diana-miRExTra is a web-based tool that allows the detection of overrepresented motifs (hexamers) on the 3′ UTRs of deregulated genes, in order to identify miRNAs responsible for such deregulation [113].

The input consists of two lists: a list of changed genes and a list of unchanged genes (background). Moreover, the web server offers the option to use evolutionary information in order to refine results.

Instead of a gene list the user may provide a list of genes with associated fold change values (or any other metric used in high-throughput experiments). Optionally, the user may provide a list of miRNAs of interest to calculate results only for hexamers corresponding to these miRNAs.

The tool compares the distributions of all possible hexamers on the 3′UTR sequences between changed and unchanged genes. A one-sided Wilcoxon Rank Sum test is used in order to identify hexamers that are present significantly more often in the set of changed genes compared to the background of unchanged genes. A p-value for each motif is calculated signifying the probability that the changed and unchanged sets are produced by the same distribution and the differences between them are due to chance alone. DIANA-mirExTra provides a combinatorial hexamer score that takes into account the whole active region of the 8 first nucleotides of the miRNA.

15.4.3 miRNA, Gene Expression and Networks

The third class of miRNA functional analysis tools that we consider provides users with network oriented data. Networks constitute an effective tool for modelling complex biological systems and since miRNAs play a central role in many processes and pathways, it is important to have tools able to integrate miRNA related data into networks. In this subsection we briefly introduce four different tools which combine miRNA related data with other information such as transcription factors or gene expression in order to create interaction networks which model and describe the molecular systems involving miRNA regulation. Most of these tools also offer computational facilities for the visualization and the analysis of such networks.

The first tool that we describe is strictly connected to Diana-miRExTra, introduced in the previous section, and is called Diana-miRPath. It is a web-based computational tool developed to identify molecular pathways potentially altered by the expression of single or multiple microRNAs [114]. The user can select either a single miRNA or multiple miRNAs and specify the tools used for the prediction of targets, among Diana-MicroT, PicTar and TargetScan [115]. The software then performs an enrichment analysis of the predicted miRNA targets comparing them to all known KEGG pathways. The output consists of a list of pathways in which the miRNA is potentially involved through its target genes. For each association an enrichment p-value is given. When working on multiple miRNAs, the algorithm also performs an enrichment analysis of the Union and Intersection target sets. The graphical output of the program provides an overview of the parts of the pathway modulated by microRNAs, facilitating the interpretation and

presentation of the analysis results. A direct link to the Diana-miRPath analysis is also provided in Diana-miRExTra for the targets of each miRNA belonging to the set of 'changed' genes.

MIR@NT@N is a tool which predicts regulatory networks and sub-networks including conserved motifs, feedback loops (FBL) and feed-forward loops (FFL) [116]. It integrates Transcription Factors, miRNAs and genes into a unified model and allows the identification and the analysis of molecular interaction networks within a given biological context.

The MIR@NT@N database integrates information from multiple available databases: PAZAR, JASPAR and oPOSSUM for TF regulations, miRBase, MicroCosm and microRNA.org for miRNA target predictions, UniHI for protein-protein interactions and Ensembl for gene annotations [117–124]. The tool is based on a meta-regulation network model that illustrates interactions between the considered three biological entities, transcription factors, microRNAs and protein-coding genes.

The tool allows to perform two types of query. The first type allows to search for novel key actors in a biological context. This query includes three sections. The first one is called Transcription Factor regulation which statistically predicts potential TFs regulating a list of miRNAs, or conversely miRNAs regulated by a list of TFs.

The second section is called miRNA regulation and allows the prediction of significant targets of a list of miRNAs or the miRNAs targeting a list of genes. The third section is called Regulation Network. It allows to reconstitute meta-regulation networks together with the detection of regulatory motifs such as FBL or FFL, by combining both TF and miRNA regulation predictions.

Users can also provide a list of miRNA-gene interactions experimentally inferred from microarray data combining genes and miRNA expression, or a list of published TF-miRNA interactions.

The second type of query provides an overview on any TF, gene or miRNA, including their interactions. It has two types of search called Quick Search and Quick Network.

The first one rapidly retrieves information on any actor, its regulators and/or targets, while Quick Network generates regulation networks from a list of actors presumed to be involved in a particular biological context, and also allows the extraction of sub-networks including regulatory motifs. The output is an exportable interaction graph recapitulating all predicted interactions and which is linked to external resources.

Based on these predictions, the user can generate networks and further analyze them to identify sub-networks, including motifs such as FBL and FFL. In addition, networks can be built from lists of molecular actors in a given biological process to predict novel and unanticipated interactions.

miRConnX is a web tool for the identification of gene network motifs involving transcription factors and miRNAs [125]. Users have to provide a document with a gene expression profile. Optionally, a document with a miRNA expression profile can be provided.

The output consists of the graphic visualization of networks involving miRNAs, transcription factors and miRNA-regulated genes.

Details about the miRNA/gene and the transcription factor/miRNA interactions are provided in tables, reporting the effect (activation/repression), the identified FFL motifs, if any, the strength of the interaction and several links to other resources about the corresponding miRNA and genes, like Gene Ontology, miRo' and miR2Disease.

The tool uses a pre-compiled network, which is derived from transcription factors binding predictions, miRNA target predictions and literature evidences. All the connections in this network correspond to direct, predicted or verified interactions. Another network based on the input expression data is then created by using a statistical association measure. This network connects transcription factors and miRNAs and doesn't discriminate between direct and indirect interactions. The two networks are superimposed via an integration function. The result is a directed network, which is a smaller version of the pre-compiled network, refined by the user provided expression data. Since the expression profiles can be related to a certain disease or phenotype, the

resulting network is representative of the condition of interest.

Finally, we describe miRScape, a Cytoscape plugin for annotating networks with miRNAs. Cytoscape is a software environment for the visualization and analysis of biological networks [126–129]. It has a basic set of features for data integration and visualization, while additional features are available as plug-ins. miRScape is the first Cytoscape plug-in allowing the mining of biological networks annotated with miRNAs. The data is retrieved from miRò, thus miRScape represents a bridge connecting miRò and Cytoscape. Given a network, previously loaded into Cytoscape, miRScape allows to identify relationships among genes, processes, functions and diseases at the miRNA level and annotating them as attributes of each network node. These annotated networks may be further analyzed by using mining features available as plug-ins on Cytoscape allowing to find for examples hubs, interesting motifs and so on.

mirScape is equipped with two modules, available on two different panels. The first panel allows users to perform a "Search by Gene" query. Once a set of nodes in the network have been selected, users can choose the kind of data to be retrieved from miRò, which can be processes, functions, and diseases in which the selected genes are involved and the miRNAs regulating them. The result is the annotation of the network with the obtained information. The "Search by miRNA" panel allows the selection of a set of miRNAs and the source of miRNA target information, which can be TargetScan, PicTar, miRanda and miRecords (validated interactions). Moreover, it is possible to choose to annotate the nodes with information about the related diseases, processes and functions.

Once the information has been retrieved from miRò, the new miRNA nodes are added to the network and connected to target gene nodes, if they are present in the network. The annotation function allows to store such acquired data as network attributes. However, the annotation function can be used as a stand-alone tool, storing in the network all the information retrievable from miRò.

15.5 Conclusions

MiRNA and, more in general, ncRNA research is in its golden age. It is clear that miRNAs are involved in a variety of fundamental processes and that their dysregulation can be related to cancer and many other diseases. Evidence shows that they don't act as single actors but cooperate among themselves and together with other molecules, like transcription factors, to regulate gene expression and, indirectly, carry out specific functions.

The number of computational tools for the analysis of miRNA related data is continuously increasing. They range from simple look-up resources to more sophisticated analysis tools. Some of them are based on manually curated information but the vast majority makes also use of computationally predicted data. Although miRNA profiling is a valuable diagnostic and prognostic tool itself, allowing the classification of samples and the identification of biomarkers, the central data in the analysis pipeline is the target gene, through which the miRNA is connected to all the other data. Indeed, miRNAs exert their functions by directly regulating the expression of their target genes and most genes are well annotated with the processes, diseases and pathways in which they are involved. Thus, miRNAs inherit these annotations, but this only represents a first step in their functional analysis. Much effort is needed to uncover the real role of miRNAs in the great number of processes and diseases in which they are potentially involved and this is the ultimate goal of most of the computational tools reviewed in this chapter.

Some of them are focused on specific kinds of data, while others try to provide a complete view of the environment in which miRNAs operate and offer modules for the analysis of the complex relationships in that they intertwine with the other molecular actors.

The increase of precision in the data produced by the use of new technologies for the measurement of gene expression and high-throughput sequencing, involves the need for more sophisticated software tools for the analysis of this data.

As in a bottom-up schema, the collected raw data constitutes a first layer. The upper layers consist of tools for the annotation of this data, often focused on specific aspects. This annotated data constitutes the input for the top layer tools, whose aim is the integration of heterogeneous information in order to produce general models of miRNA functions in the context of complex processes. These tools must be equipped with powerful analysis facilities, helping researchers to formulate concrete functional hypotheses and guiding them to design the correct experiments to perform hypotheses validation. Then, the data produced with these experiments represents a feedback for the refinement of the analysis pipeline.

The final key point is the integration of public data with user data, and this is already partly fulfilled by some of the reviewed tools. In fact, many users typically get original data from their experiments, thus it is important to have tools able to combine this data with the other information stored in databases, in order to produce more reliable models specific to user needs.

References

1. Jovanovic M, Hengartner MO (2006) miRNAs and apoptosis: RNAs to die for. Oncogene 25:6176–6187
2. Wienholds E, Plasterk RHA (2005) MicroRNA function in animal development. FEBS Lett 579:5911–5922
3. Xiao C, Rajewsky K (2009) MicroRNA control in the immune system: basic principles. Cell 136(1):26–36
4. Small EM, Olson EN (2011) Pervasive roles of microRNAs in cardiovascular biology. Nature 469(7330):336–342
5. Lau P, de Strooper B (2010) Dysregulated microRNAs in neurodegenerative disorders. Semin Cell Dev Biol 21(7):768–773
6. Iorio MV, Croce CM (2009) MicroRNAs in cancer: small molecules with a huge impact. J Clin Oncol 27(34):5848–5856
7. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136(2):215–233
8. Didiano D, Hobert O (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol 13(9):849–851
9. Didiano D, Hobert O (2008) Molecular architecture of a miRNA-regulated 3′ UTR. RNA 14(7):1297–1317
10. Yin JQ, Zhao RC, Morris KV (2008) Profiling microRNA expression with microarrays. Trends Biotechnol 26:70–76
11. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR et al (2005) Real-time quantification of microRNAs by stem–loop RT-PCR. Nucleic Acids Res 33:e179. doi:10.1093/nar/gni178
12. Shi R, Chiang VL (2005) Facile means for quantifying microRNA expression by real-time PCR. Biotechniques 39:519–525
13. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods 44:3–12
14. Krichevsky AM, King KS, Donahue CP, Khrapko K, Kosik KS (2003) A microRNA array reveals extensive regulation of microRNAs during brain development. RNA 9:1274–1281
15. Liu CG, Calin GA, Meloon B, Gamliel N, Sevignani C, Ferracin M, Dumitru CD, Shimizu M, Zupo S, Dono M, Alder H, Bullrich F, Negrini M, Croce CM (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. Proc Natl Acad Sci USA 101:9740–9744
16. Zhao JJ, Hua YJ, Sun DG, Meng XX, Xiao HS, Ma X (2006) Genome-wide microRNA profiling in human fetal nervous tissues by oligonucleotide microarray. Childs Nerv Syst 22:1419–1425
17. Schmittgen TD, Jiang J, Liu Q, Yang L (2004) A high-throughput method to monitor the expression of microRNA precursors. Nucleic Acids Res 32:e43
18. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33:e179
19. Schmittgen TD, Lee EJ, Jiang J, Sarkar A, Yang L, Elton TS, Chen C (2008) Real-time PCR quantification of precursor and mature microRNA. Methods 44:31–38
20. Bissels U, Wild S, Tomiuk S, Holste A, Hafner M, Tuschl T, Bosio A (2009) Absolute quantification of microRNAs by using a universal reference. RNA 15:2375–2384
21. Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 210(Pt 9):1518–1525
22. Church GM (2006) Genomes for all. Sci Am 294(1):46–54
23. Zak DE, Aderem A (2009) A systems view of host defense. Nat Biotechnol 27(11):999–1001
24. Geiss GK et al (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 26:317–325
25. Kulkarni MM (2011) Digital multiplexed gene expression analysis using the NanoString nCounter system. Curr Protoc Mol Biol Chapter 25:Unit25B.10

26. Lopez-Romero P, Gonzalez MA, Callejas S, Dopazo A, Irizarry RA (2010) Processing of Agilent microRNA array data. BMC Res Notes 3:18
27. Hua YJ, Tu K, Tang ZY, Li YX, Xiao HS (2008) Comparison of normalization methods with microRNA microarray. Genomics 92:122–128
28. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, Harshman K (2009) Impact of normalization on miRNA microarray expression profiling. RNA 15:493–501
29. Risso D, Massa MS, Chiogna M, Romualdi C (2009) A modified LOESS normalization applied to microRNA arrays: a comparative evaluation. Bioinformatics 25:2685–2691
30. Sarkar D, Parkin R, Wyman S, Bendoraite A, Sather C, Delrow J, Godwin AK, Drescher C, Huber W, Gentleman R, Tewari M (2009) Quality assessment and data analysis for microRNA expression arrays. Nucleic Acids Res 37:e17
31. Peltier HJ, Latham GJ (2008) Normalization of microRNA expression levels in quantitative RTPCR assays: identification of suitable reference RNA targets in normal and cancerous human solid tissues. RNA 14:844–852
32. Chang KH, Mestdagh P, Vandesompele J, Kerin MJ, Miller N (2010) MicroRNA expression profiling to identify and validate reference genes for relative quantification in colorectal cancer. BMC Cancer 10:173
33. Sato F, Tsuchiya S, Terasawa K, Tsujimoto G (2009) Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. PLoS One 4:e5540
34. Benes V, Castoldi M (2010) Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available. Methods 50:244–249
35. Mestdagh P, Van VP, De WA, Muth D, Westermann F, Speleman F, Vandesompele J (2009) A novel and universal method for microRNA RT-qPCR data normalization. Genome Biol 10:R64
36. Vandesompele J, De PK, Pattyn F, Poppe B, Van RN, De PA, Speleman F (2003) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034
37. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA 16:991–1006
38. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–excel-based tool using pair-wise correlations. Biotechnol Lett 26:509–515
39. Andersen CL, Jensen JL, Orntoft TF (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res 64:5245–5250
40. Wang B, Wang XF, Howell P, Qian X, Huang K, Riker AI, Ju J, Xi Y (2010) A personalized microRNA microarray normalization method using a logistic regression model. Bioinformatics 26:228–234
41. Baskerville S, Bartel DP (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 11:241–247
42. Liang RQ, Li W, Li Y, Tan CY, Li JX, Jin YX, Ruan KC (2005) An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe. Nucleic Acids Res 33:e17. doi:10.1093/nar/gni019
43. Wang H, Ach RA, Curry B (2007) Direct and sensitive miRNA profiling from low-input total RNA. RNA 13:151–159
44. Rao Y, Lee Y, Jarjoura D, Ruppert AS, Liu CG, Hsu JC, Hagan JP (2008) A comparison of normalization techniques for microRNA microarray data. Stat Appl Genet Mol Biol 7: Article22
45. Chiogna M, Massa MS, Risso D, Romualdi C (2009) A comparison on effects of normalisations in the detection of differentially expressed genes. BMC Bioinform 10:61
46. Sun Y, Koo S, White N, Peralta E, Esau C, Dean NM, Perera RJ (2004) Development of a microarray to detect human and mouse microRNAs and characterization of expression in human organs. Nucleic Acids Res 32:e188. doi:10.1093/nar/gnh186
47. Castoldi M, Schmidt S, Benes V, Hentze MW, Muckenthaler MU (2008) miChip: an array-based method for microRNA expression profiling using locked nucleic acid capture probes. Nat Protoc 3:321–329
48. Garzon R, Volinia S, Liu CG, Fernandez-Cymering C, Palumbo T, Pichiorri F, Fabbri M, Coombes K, Alder H, Nakamura T, Flomenberg N, Marcucci G, Calin GA, Kornblau SM, Kantarjian H, Bloomfield CD, Andreeff M, Croce CM (2008) MicroRNA signatures associated with cytogenetics and prognosis in acute myeloid leukemia. Blood 111(6): 3183–3189
49. Miska EA, Alvarez-Saavedra E, Townsend M, Yoshii A, Sestan N, Rakic P, Constantine-Paton M, Horvitz HR (2004) Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biol 5:R68. doi:10.1186/gb-2004-5-9-r68
50. Tian Z, Greene AS, Pietrusz JL, Matus IR, Liang M (2008) MicroRNA-target pairs in the rat kidney identified by microRNA microarray, proteomic, and bioinformatic analysis. Genome Res 18:404–411
51. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193
52. Steinhoff C, Vingron M (2006) Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform 7:166–177

53. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(Suppl 1):S96–S104
54. Xiong H, Zhang D, Martyniuk CJ, Trudeau VL, Xia X (2008) Using generalized procrustes analysis (GPA) for normalization of cDNA microarray data. BMC Bioinformatics 9:25
55. Bargaje R, Hariharan M, Scaria V, Pillai B (2010) Consensus miRNA expression profiles derived from interplatform normalization of microarray data. RNA 16:16–25
56. Do JH, Choi DK (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. Mol Cells 22:254–261
57. Garzon R, Garofalo M, Martelli MP, Briesewitz R, Wang L, Fernandez-Cymering C, Volinia S, Liu CG, Schnittger S, Haferlach T, Liso A, Diverio D, Mancini M, Meloni G, Foa R, Martelli MF, Marcucci C, Croce CM, Falini B (2008) Distinctive microRNA signature of acute myeloid leukemia bearing cytoplasmic mutated nucleophosmin. Proc Natl Acad Sci USA 105(10):3945–3950
58. Perkins DO, Jeffries CD, Jarskog LF, Thomson JM, Woods K, Newman MA, Parker JS, Jin J, Hammond SM (2007) MicroRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. Genome Biol 8:R27. doi:10.1186/gb-2007-8-2-r27
59. Chen Y, Dougherty E, Bittner ML (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Opt 2:364–374
60. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98(9):5116–5121
61. Mukherjee S, Roberts SJ, van der Laan M (2005) Data-adaptive test statistics for microarray data. In: The ninth annual international conference on research in computational molecular biology, Cambridge, MA, pp 237–238
62. Martin DE, Demougin P, Hall MN, Bellis M (2004) Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. BMC Bioinform 5(1):148
63. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573(1–3):83–92
64. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3(1):Article 3
65. Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7:819–837
66. Lee ML, Lu W, Whitmore GA, Beier D (2002) Models for microarray gene expression data. J Biopharm Stat 12:1–19
67. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2:research0049.1–0049.12
68. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264
69. Tsai CA, Chen JJ (2004) Significance analysis of ROC indices for comparing diagnostic markers: applications to gene microarray data. J Biopharm Stat 14(4):985–1003
70. http://bioconductor.org/packages/release/bioc/html/RankProd.html
71. Dopazo J, Carazo JM (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J Mol Evol 44:226–233
72. Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
73. Kohonen T (1995) Self-organizing maps, vol 30, Springer series in information sciences. Springer, Berlin
74. Fraley C, Raftery AE (1998) How many clusters? Which clustering methods? Answers via model-based cluster analysis. Comput J 41:578–588
75. Fraley C, Raftery AE (1999) MCLUST: software for model-based cluster analysis. J Classif 16:297–306
76. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631
77. McLachlan GJ, Bean RW, Ben-Tovim JL, Zhu JX (2005) Using mixture models to detect differentially expressed genes. Aust J Exp Agric 45:859–866
78. McLachlan GJ, Bean RW, Ben-Tovim JL (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics 22:1608–1615
79. Ma L, Teruya-Feldstein J, Weinberg RA (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature 449:682–688
80. Tavazoie SF, Alarcon C, Oskarsson T, Padua D, Wang Q, Bos PD, Gerald WL, Massague J (2007) Endogenous human microRNAs that suppress breast cancer metastasis. Nature 451:147–152
81. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, Lebanony D, Goren Y, Silberschein E, Targan N, Ben Ari A, Gilad S, Sion-Vardy N, Tobar A, Feinmesser M, Kharenko O, Nativ O, Nass D, Perelman M, Yosepovich A, Shalmon B, Polak-Charcon S, Fridman E, Avniel A, Bentwich I, Bentwich Z, Cohen D, Chajut A, Barshack I (2008) MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol 26:462–469
82. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa-Morais NL, Teschendorff AE, Green AR, Ellis IO et al (2007)

MicroRNA expression profiling of human breast cancer identifies new markers of tumour subtype. Genome Biol 8:R214
83. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, Knudsen BS, Stirewalt DL, Gentleman R, Vessella RL, Nelson PS, Martin DB, Tewari M (2008) Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci USA 105:10513–10518
84. Lange J (2010) microRNA profiling on automated biochip platform reveals biomarker signatures from blood samples. Nat Methods 7. doi:10.1038/nmeth.f.281
85. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838
86. Visone R, Veronese A, Rassenti LZ, Balatti V, Pearl DK, Acunzo M, Volinia S, Taccioli C, Kipps TJ, Croce CM (2011) miR-181b is a biomarker of disease progression in chronic lymphocytic leukemia. Blood 118(11):3072–3079
87. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297
88. Sethupathy P et al (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. RNA 12:192–197
89. Xiao F et al (2009) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37(Database issue):D105–D110
90. Chi SW et al (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 460(7254):479–486
91. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120(1):15–20
92. Rehmsmeier M et al (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517
93. Hofacker IL (2007) How microRNAs choose their targets. Nat Genet 39(10):1191–1192
94. Mathews DH (2006) Revolutions in RNA secondary structure prediction. J Mol Biol 359(3):526–532
95. Grimson A et al (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27:91–105
96. Mazière P, Enright AJ (2007) Prediction of microRNA targets. Drug Discov Today 12(11–12):452–458
97. John B et al (2004) Human MicroRNA targets. PLoS Biol 2(11):1862–1879
98. Betel D et al (2007) The microRNA.org resource: targets and expression. Nucleic Acids Res 36(Database issue):D149–D153
99. Krek A et al (2005) Combinatorial microRNA target predictions. Nat Genet 37(5):495–500
100. Kiriakidou M et al (2004) A combined computational-experimental approach predicts human microRNA targets. Genes Dev 18(10):1165–1178
101. Miranda KC et al (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 126:1203–1217
102. Long D et al (2007) Potent effect of target structure on microRNA function. Nat Struct Mol Biol 14(4):287–294
103. Kertesz M et al (2007) The role of site accessibilità in microRNA target recognition. Nat Genet 39(10):1278–1284
104. Jiang Q, Wang Y, Juan L et al (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37(Database issue):D98–D104
105. Lu M, Zhang Q, Deng M et al (2008) An analysis of human microRNA and disease associations. PLoS One 3(10):e3420
106. Barh D, Bhat D, Viero C (2010) miReg: a resource for microRNA regulation. J Integr Bioinform 7(1).
107. Laganà A, Forte S, Giudice A et al (2009) miRò: a miRNA knowledge base. Database (Oxford) 2009:bap008. doi:10.1093/database/bap008
108. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
109. Landgraf P, Rusu M, Sheridan R et al (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 129:1401–1414
110. Ulitsky I, Laurent LC, Shamir R (2010) Towards computational prediction of microRNA function and activity. Nucleic Acids Res 38(15):e160
111. Le Brigand K, Robbe-Sermesant K, Mari B et al (2010) MiRonTop: mining microRNAs targets across large scale gene expression studies. Bioinformatics 26(24):3131–3132
112. Sales G, Coppe A, Bisognin A et al (2010) MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. Nucleic Acids Res 38(Web Server issue):W352–W359
113. Alexiou P, Maragkakis M, Papadopoulos GL et al (2010) The DIANA-mirExTra web server: from gene expression data to microRNA function. PLoS One 5(2):e9171
114. Papadopoulos GL, Alexiou P, Maragkakis M et al (2009) DIANA-mirPath: integrating human and mouse microRNAs in pathways. Bioinformatics 25(15):1991–1993
115. Maragkakis M, Alexiou P, Papadopoulos GL et al (2009) Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinform 10:295
116. Le Béchec A, Portales-Casamar E, Vetter G et al (2011) MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model. BMC Bioinform 12:67

117. Portales-Casamar E, Arenillas D, Lim J et al (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. Nucleic Acids Res 37(Database issue):D54–D60
118. Portales-Casamar E, Thongjuea S, Kwon AT et al (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 38(Database issue):D105–D110
119. Ho Sui SJ, Mortimer JR, Arenillas DJ et al (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res 33(10):3154–3164
120. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39(Database Issue):D152–D157
121. Griffiths-Jones S, Saini HK, van Dongen S et al (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36(Database Issue):D154–D158
122. Griffiths-Jones S, Grocock RJ, van Dongen S et al (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34(Database Issue):D140–D144
123. Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32(Database Issue):D109–D111
124. Chaurasia G, Malhotra S, Russ J et al (2009) UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome. Nucleic Acids Res 37(Database issue):D657–D660
125. Huang GT, Athanassiou C, Benos PV (2011) mirConnX: condition-specific mRNA-microRNA network integrator. Nucleic Acids Res 39(Web Server issue):W416–W423
126. Ferro A, Giugno R, Laganà A et al (2009) miRScape: a Cytoscape plugin to annotate biological networks with microRNAs. NETTAB 2009 conference, Catania, 10–12 June 2009
127. Smoot ME, Ono K, Ruscheinski J et al (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27(3):431–432
128. Cline MS, Smoot M, Cerami E et al (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2(10):2366–2382
129. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504