Analysis of miRNA expression using machine learning

Henry Wirth[1,2*], Volkan Cakir[1*], Lydia Hopp[1,2] and Hans Binder[1,2#]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig, Germany
[2] Leipzig Research Center for Civilization Diseases; Universität Leipzig

* contributed equally
# corresponding author
E-mail: binder@izbi.uni-leipzig.de

Key words: microarrays, self organizing maps, feature selection, mRNA-miRNA coexpression, gene set enrichment analysis

Abstract

The systematic analysis of miRNA expression and of its potential mRNA targets constitutes a basal objective in miRNA research in addition to miRNA gene detection and miRNA target prediction. In this contribution we address methodical issues of miRNA expression analysis using self-organizing maps (SOM), a neural network machine learning algorithm with strong visualization and second-level analysis capabilities widely used to categorize large scale, high-dimensional data. We shortly review selected experimental and theoretical aspects of miRNA expression analysis. Then, the protocol of our SOM-method is outlined with special emphasis on miRNA/mRNA coexpression. The method allows extracting differentially expressed RNA-transcripts, their functional context and also characterizing global properties of expression states and profiles. In addition to the separate study of miRNA and mRNA expression landscapes we propose the combined analysis of both entities using a covariance-SOM.

1. Introduction

A major goal of machine learning research is to recognize complex patterns and to make intelligent decisions based on the data and the particular objective of interest with the help of computers. Usually two research problems involving miRNA are tackled with computational methods, namely detecting miRNA genes and predicting miRNA targets. The use of machine learning methods has been shown to improve the outcome of both miRNA gene detection and -target prediction (see (1) for a review and also (2-4)). These approaches typically make use of sequence information (e.g. of short 6-8 nt miRNA binding motifs), secondary structure (e.g. of stem-loops using thermodynamic modeling) and phylogenetic conservation to classify possible candidates according to their relevance as miRNA genes or targets using methods such as hidden Markov models, Random Forest classifiers or Support Vector Machines.

Recently another machine learning technique, namely self-organizing maps (SOM), was applied for miRNA target prediction based on clustering of short 3'-untranslated regions (5). SOM machine learning was developed by Kohonen about thirty years ago (6). It projects data from the original high dimensional space to reference vectors of lower dimension. First studies applying SOM to microarray gene expression data were published by Tamayo et al. (7) and Törönen et al. (8). These and later applications of the SOM method to expression data emphasized either a gene-centered perspective to cluster genes or a sample-centered mode to map individual samples onto the SOM grid enabling the classification of samples into a small number of diagnostic or prognostic groups (9-10). The SOM method can be configured also in such a way that it combines sample- and gene-centered perspectives (9). In such application SOM analysis seems not only well suited to accomplish downstream analysis tasks specified above but it also offers several advantages that make this method superior compared with other ones.

In this contribution we address methodical issues of miRNA expression analysis with the special focus on the SOM technique. To our best knowledge this method was applied here for the first time to analyze miRNA expression data. The paper is divided into two parts: Firstly, we shortly review selected experimental and theoretical aspects of miRNA expression analysis. Secondly, our SOM-method is presented more in detail with special emphasis on miRNA/mRNA coexpression. In this part details of the method are exemplified using a combined data set of miRNA and mRNA expression in healthy and tumor tissue samples taken from ref. (11).

2. Methods of miRNA expression analysis

2.1. miRNA detection

An overview over different methods of miRNA detection including Northern Blot, in situ detection, bead techniques and quantum dots is given in refs. (12-13). Currently three methods are most commonly applied to measure mRNA and miRNA expression:

1. real-time reverse transcription-PCR (qPCR),
2. microarray hybridization, and
3. massively parallel next-generation sequencing (NGS)

(see (14-16) and references cited therein). Application of these methods to miRNA faces basically two challenges compared to their use in mRNA analysis: the short length of mature miRNA sequences (~20 – 25 nt) and the nearly identical sequences of miRNA of the same family whose members can differ by as little as one nucleotide and nevertheless can exhibit strong differential expression. The specificity required to differentiate between such closely related short RNA fragments surpasses requirements for conventional mRNA detection and raises problems due to the constrained probe design (microarrays) and sequencing errors (NGS). In contrast to mRNA profiling technologies, miRNA profiling must also take into account the difference between mature miRNAs and their precursors.

The technical merits and drawbacks of qPCR, microarrays, and sequencing of miRNAs are similar to their application for mRNA or genomic DNA quantitation. The clear advantage of high-throughput sequencing compared to microarrays is that it is not hindered by the variability of probe affinities and cross hybridization of nearly identical miRNA family members. Moreover, analysis of read patterns allows to identify novel miRNAs (17). On the other hand, RNA ligation and PCR amplification steps (see below) and also library preparation bear inherent biases paralleled, e.g. by systematic preferential representation of the miRNA complement (18). Moreover, NGS of miRNAs can be influenced by sequencing errors and often requires search and removal of adaptor sequences before the miRNA sequence itself can be elucidated.

As in mRNA analysis, microarrays are still a good choice for a standardized genome-wide assay that is amenable to high-throughput applications. The differences between available platforms (e.g. Agilent; Exiqon, Illumina, Ambion, Combimatrix, Invitrogen, Affymetrix) range from surface chemistry and printing technology, through probe design and labeling techniques to the required amount of material for hybridization and costs (see (12, 15-16)). Several attempts have been made in surface chemistry (e.g. with probes containing locked nucleic acid bases (LNA)) and probe structure (e.g. with 'stem-loop' probes) to improve the array sensitivity for discriminating mispaired bases in a better way and to equalize the probe affinities for miRNA target binding, however partly with questionable success (15). miRNA microarrays have high intra-platform repeatability and comparability to quantitative qPCR. However, the current lineup of commercially available miRNA microarray systems fails to show good inter-platform concordance, probably because of severe divergence in stringency of detection call criteria between different platforms (16). Unlike for mRNA gene expression (19), only few attempts have been made so far to establish rigorous parameters for the evaluation of miRNA microarray platforms using standartized quality measures.

Several studies report low degree of overlap in the differentially expressed miRNAs between different detection methods, not easily attributable to the strength or weakness of the different platforms (15, 20). qPCR, often considered a 'gold standard' in the detection and quantification of gene expression, can be used better as a validation rather than as discovery tool because of relatively large number of miRNAs presently known. Even qPCR seems to fail as an infallible validation method of miRNA microarray data (15, 20). Hence despite its descriptive name and the fact that qPCR has been repeatedly used as a validation technique of choice, it is not necessarily appropriate to use qPCR data as an absolute 'gold standard'. The question of a basal standard in miRNA expression awaits further advances in both technology (e.g. deep sequencing) and computation (normalization and downstream analysis algorithms).

2.2. Data preprocessing tasks: Calibration and normalization

qPCR, microarray hybritization as well as NGS methods of miRNA detection face significant introduction of technical and experimental bias. Preprocessing aims at minimizing such systematic errors and thus has significant impact on downstream analysis and, particularly on the detection of differentially expressed miRNA.

1. Calibration, the first subtask of preprocessing, aims at removing systematic biases from raw data to get expression estimates which linearly correlate with the 'true' RNA transcript abundance separately in each of the samples. This includes method and platform specific steps, such as e.g., baseline adjustment and threshold setting for qPCR analyses, background and affinity correction for microarray technology, or filtering for small RNA-sequence data (NGS). We refer to special literature addressing such method-specific calibration issues (19, 21-22).

2. The second task, normalization, aims at ensuring comparability of the transcript abundance estimates between the different specimen by adjusting the data to batch effects such as different total RNA concentrations, total read counts (NGS) or residual background levels (microarrays). Normalization is crucial since signal levels may be modulated by the RNA extraction yields and inverse transcription and PCR amplification reaction efficiencies in a sample specific way. In general, there is no consensually best performing normalization method for any of the three miRNA profiling approaches (23). Several normalization techniques are currently applied, some of which are similar to mRNA profiling normalization methods, while others consider the specifics of miRNA data.

Normalization using endogenous control probes represents a simple but powerful strategy widely used in miRNA analysis. It is based on the selection of reference miRNAs or other small non-coding (nc) RNAs (e.g. small nucleolar RNA) as predefined invariant endogenous controls (24). The expression levels of the transcripts measured in each of the samples are then simply scaled by the mean expression levels of the controls (preferentially in log-scale) in the respective samples by assuming that their variations are solely caused by technical and experimental factors. ncRNAs other than miRNA might be problematic because they do not mirror the physico-chemical properties of miRNA and because ncRNA abundance migth not reflect the overall activity of the miRNA processing machinery. It escpecially raises problems if the total miRNA level alters as in comparisons of multiple tissues or cell lines (15). Selection of invariant 'housekeeping' miRNAs identified by different algorithms is superior over small ncRNA based normalization (see (23) and references cited therein).

Another normalization strategy uses global RNA expression measures as intrinsic control. It assumes that the overall transcripton level is constant and one can use, e.g., the median or mean expression of all transcripts measured in each sample as reference scale. Other, more sophisticated methods such as quantile normalization (leveling the expression frequency distributions) (25) or LOESS (26-27) (local regression; leveling the local mean) scale the frequency distribution of expression values of all samples or their local, expression dependent mean, respectively.

Global miRNA expression patterns (and potentially also the expression levels of endogenous controls) however are thought to change dramatically in response to Drosha and histone deacetylase levels, cell division status, neoplastic transformation, developmental stage(s), circadian rhythms, cellular stress, and other factors. Hence the assumptions - common to many mRNA expression profiling experiments - that overall RNA transcription is constant, and that a low percentage of individual transcripts are changed under different test conditions, are mostly not applicable to miRNA studies. The nature of miRNA profiling data which mirrors the distinct biogenesis and physico-chemical nature of miRNAs can challenge conventional normalization methods originally developed for mRNA expression data. Innovative approaches are required in order to reconcile miRNA profiling data analyses with the specifics of miRNA biology. This can make use of combinations of qPCR, microarray and NGS data for mutual validations possibly circumventing the need for external references.

2.3. Downstream analysis tasks

Downstream analysis follows preprocessing.

1 It includes tasks such as differential analysis, also known as marker selection. It is the search for genes that are differentially expressed in distinct phenotypes, treatment conditions, developmental stages etc. One can assess differential expression using different scores such as simple fold change measures or t-test statistics (see below).

2. Another task, supervised learning and class prediction, is the search for a gene expression signature that predicts class (phenotype) membership.

3. Class Discovery (unsupervised learning) is the search for a biologically relevant unknown group identified by a gene expression signature or a biologically relevant set of co-expressed genes. The basic methodology for class discovery is clustering.

4. Finally, functional context analysis is the search for sets of genes differentially expressed in distinct phenotypes using enrichment techniques. We will address these tasks below in the context of SOM machine learning.

2.4. Data

We selected a data set on healthy and tumor tissues to illustrate different aspects of expression analysis using SOM machine learning in form of a case study. This so-called *LU-cancer* set contains miRNA and mRNA measurements from the same samples of seven healthy and tumor tissues (colon, kidney, bladder, prostate, uterus, lung, breast) (11). miRNA were measured using a bead-based profiling method estimating the abundance of 217 miRNAs. mRNA expression was determined using microarrays.

3. Discovering miRNA expression phenotypes using SOM

3.1. Input data

In general, we analyze the expression levels, E_{nmi}, of n=1…N genes measured under m=1…M different conditions such as different sample types (e.g. tissues or cell lines), time points (e.g. in a time series after perturbation), treatments (e.g. using different chemicals) or patients (e.g. from a cohort study). Each condition defines a different molecular expression phenotype which can be measured in i=1…R_m replicates. Replicates might be technical (e.g. by analyzing the RNA extracts several times) or biological (e.g. by extracting RNA from different equally treated specimen) ones. In addition we define phenotype classes as groups of samples of a common functional context such as tissue categories (e.g. nervous or muscle tissues, cancer or healthy samples). The choice of classes depends on the issue studied. For miRNA and mRNA expression studies the number of different transcripts is typically about N=200 – 1,000 and 10,000 – 40,000, respectively. Combined data sets thus contain about 2x10^{6} – 4x10^{7} pairwise combinations of miRNA/mRNA features. Below we will use the termini 'transcript' and 'gene' as synonyms.

3.2. Preprocessing

In the first step, raw expression data of each of the $\sum_{m=1}^{M} R_m$ measurements are calibrated and normalized. For mRNA GeneChip expression data we used hook calibration of the raw probe intensities combined with quantile normalization of the expression values as described previously (10). miRNA expression data were taken from the original publications and then quantile normalized. The replicated expression values are optionally log-averaged over all replicates *i* for each condition, $e_{nm} \equiv \log_{10} E_{nm\bullet} = \frac{1}{R_m}\sum_{i=1}^{R_m} \log_{10} E_{nmi}$. Alternatively one can process each replicate individually to characterize the specifics of its expression. In this case the sample index *m* also runs over the replicates. Finally, the log-expression values of each transcript were centered with respect to their mean expression, $e_{n\bullet}$, averaged over all conditions studied, $\Delta e_{nm} \equiv e_{nm} - e_{n\bullet}$.

3.3. Training the SOM

Self-organizing map (SOM) machine learning was applied to all preprocessed expression data. The expression data are considered as *N* vectors of dimensionality *M* defining the expression profiles of the genes over all phenotypes studied, $\Delta\vec{e}_n \equiv \left\{\Delta e_{n1},...,\Delta e_{nm},...,\Delta e_{nM}\right\}$ (*n*=1…*N*). The algorithm initializes *K* so-called metagene expression profiles also representing vectors of length *M*, $\Delta\vec{e}_k\,(t=0) \equiv \left\{\Delta e_{k1},...,\Delta e_{kM}\right\}$ (*k*=1…*K*, we use the same symbol Δ*e* as above but substitute the gene index *n* by the metagene index *k*; the argument *t*=0…*T* denotes the iteration step). The metagenes are arranged in a two-dimensional grid of rectangular topology $K= K_x{\cdot}K_y$ with K_x=10 - 60 and $K_y{\approx}K_x$ tiles per x- and y-dimension, respectively. Then a single gene *n'* is picked from the list and its profile vector $\Delta\vec{e}_{n'}$ is compared with all metagene profiles using the Euclidean distance ||…|| as similarity measure. The gene picked is associated with the metagene profile of closest similarity, $\min\,\{\left\|\Delta\vec{e}_{n'} - \Delta\vec{e}_k\,\right\|\}$ for k'=k. This 'winner' metagene profile *k'* is then modified, such that it a bit more closely resembles the expression profile of the selected gene. In addition, the neighboring metagene vectors in the two-dimensional grid adjacent to this winning metagene are also modified, so that they also resemble the expression vector a little more closely. The update rule for the metagene vectors at step t+1 can be written as $\Delta\vec{e}_k\,(t+1) = \Delta\vec{e}_k\,(t) + \eta(t)\cdot h_{k'k}\cdot\left(\Delta\vec{e}_{n'} - \Delta\vec{e}_k\right)$ where 0<η(t)<1 is the learning rate decaying with progressive iteration. $h_{k'k}$ denotes the neighborhood kernel around the winner metagene. Different neighborhood kernels such as bubble (only nearest neighbors are considered) or Gaussian (i.e., a smooth decaying neighborhood) can be chosen. This process is applied to all genes and repeated a few hundred thousand times. The radius of considered neighbors decreases with progressing iterations. As a consequence, less metagene vectors are affected by smaller amounts. The metagene vectors therefore asymptotically settle down. The resulting map becomes organized because the similarity of neighboring metagenes decreases with increasing distance in the map. The algorithm ensures that all 'single' genes are assigned to 'their' metagene vector of closest similarity. The training of the SOM also ensures that the obtained metagene profiles cover the manifold of different single gene profiles seen in the experiment.

3.4. Staining the SOM: Individual phenotype portraits

In our application the method sorts the individual genes into K metagene-clusters. Each cluster is characterized by one metagene profile which is used for visualizing the expression pattern of each phenotype. To create an image that portrays the individual expression state of all genes in one of the phenotypes we first normalize the metagene expression data to values between -1 and +1 according to $\Delta e_{km}^{norm} = 2\left(\Delta e_{km} - \Delta e_{\bullet m}^{\min}\right) / \left(\Delta e_{\bullet m}^{\max} - \Delta e_{\bullet m}^{\min}\right) - 1$ where $\Delta e_{\bullet m}^{\max/\min}$ denotes the maximum/minimum values of Δe_{km} for a given m. Then, each tile of the grid is stained according to its Δe_{km}^{norm}-value by applying an appropriate color code. Our standard 'logFC'-color scale linearly transforms the normalized logged fold change, logFC= Δe_{km}^{norm} into the colors green to maroon for increasing $\Delta e_{km} \geq 0$ and green to dark blue for decreasing $\Delta e_{km} \leq 0$. Also other color codes can be applied to highlight, for example, intermediate and low expression degrees (for details see (10, 28)). Although colors are reproduced in grey-scale in this publication we will describe the images assigning the 'true' colors used in the default version of our analysis pipeline.

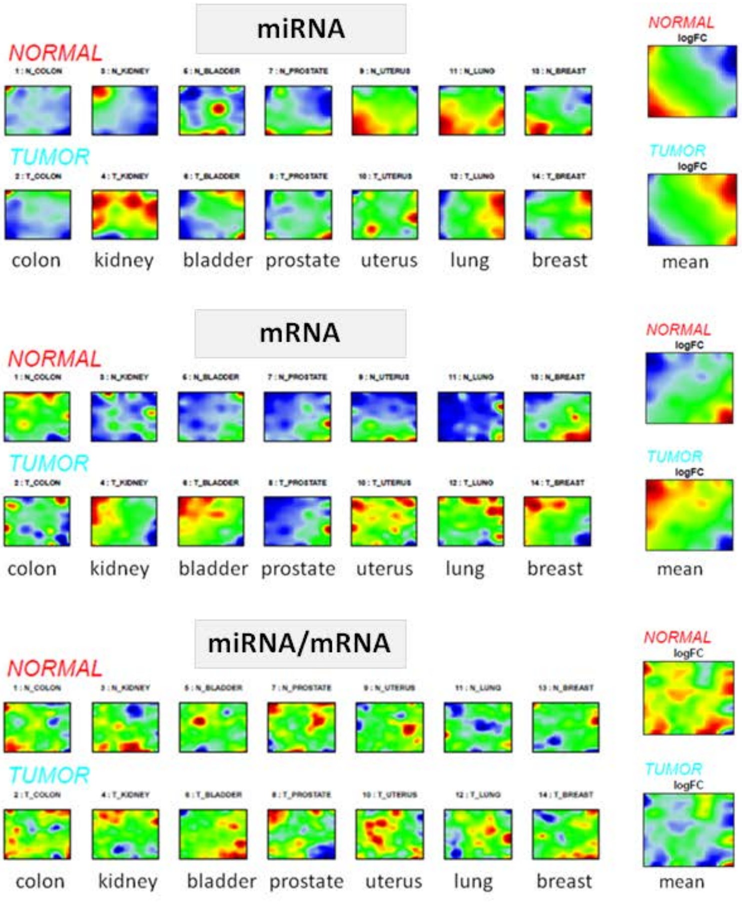

Figure 3.1: Gallery of miRNA, mRNA and combined miRNA/mRNA expression portraits of normal and tumor tissues taken from the LU-data set. The larger portraits on the right are mean portraits averaged over all individual normal and cancer portraits, respectively.

The SOM algorithm arranges similar metagene profiles together into neighbored tiles of the map whereas more different ones are located at more distant positions. In consequence, neighbored metagenes tend to be colored similarly owing to their similar expression values. Therefore, the obtained mosaic portraits show typically a smooth blurry texture with red and blue spot-like regions referring to clusters of over- and underexpressed metagenes. These blurry images portray the expression landscape of each particular phenotype in terms of a visual image. Metagenes from the same spot are co-expressed in the experimental series whereas different, well-separated overexpression spots in the same image refer to metagenes commonly overexpressed in the particular phenotype but differently expressed in other phenotypes.

3.5. Clustering the metagenes: Spots and blurs

Figure 3.1 shows the tissue-specific SOM portraits of miRNA and mRNA expression in normal and cancer samples. The red/blue spots are clusters of metagenes over/underexpressed in the respective samples. In general, the spot patterns of the individual portraits are relatively heterogeneous. Mean subtype-specific portraits are calculated as average value of each metagene expression over all phenotype portraits of one class, $\left\langle \Delta e_{km} \right\rangle_{m \in class}$. The mean portraits (see large mosaics in the right part of Figure 3.1) reveal an antagonistic pattern of up- and downregulated spots in normal and diseased tissues. In the miRNA and mRNA maps one observes virtually only two to three spots upregulated (red) in normal tissues and downregulated (blue) in tumor tissues (and vice versa) which are located preferentially in opposite corners of the map.

Different metrics can be applied to select metagene clusters: Firstly, we define over- (and under-) expression spots by applying a simple percentile criterion which selects a certain fraction (usually 2%) of the metagenes showing largest (or smallest) expression in the particular phenotype. The obtained over- and underexpression spots are individual properties depending on the particular metagene expression in each sample. They can change their size from phenotype to phenotype and they can even disappear or transform from an over- into an underexpression spot or vice versa.

Alternatively one can also apply mutual correlation or Euclidean distances between neighbored metagene profiles of the SOM as similarity measures for appropriate clustering (see ref. (10)).

3.6. Adjusting the SOM

SOM machine learning represents an unsupervised clustering algorithm whereby the number of tiles and thus the resolution of the map is predefined by the researcher and therefore constitutes the option of supervised adjustment of the results. Neighboring tiles might cluster into one spot together because they collect genes of similar expression profiles. These spots, their number, shape and size, depend on the intrinsic expression landscape of the phenotypes studied. In this sense SOM spot-clustering is a higher-order unsupervised clustering algorithm which potentially clusters the data into biologically meaningful groups or ‘modes’. This mode-selection is however based on the underlying ‘pixelation’ of the expression landscape which should be chosen such that the SOM algorithm produces a stable and consistent spot pattern.

The SOM can be configured by the number of tiles per image, different topologies (e.g., with rectangular or hexagonal lattices), and different neighborhood kernels describing the range and strength of interactions between the nodes during the training process. For small SOM sizes each metagene will contain a large number of single genes profiles whereas large sizes enable the distribution of the genes over a larger number of metagene clusters which more specifically adapt to details of the expression landscape.

We found that the number of clusters and their assignment converges if the number of tiles exceeds the number of overexpression spot clusters by about two orders of magnitude. This asymptotic behavior indicates that larger SOM sizes essentially do not further improve the information content of the map and that the obtained clusters indeed reflect intrinsic properties of the overall expression pattern. Alternative topologies such as the hexagonal ones and different neighborhood kernels only weakly affect the obtained spot textures (see supplementary text in (10)).

3.7. mRNA/miRNA coexpression

The SOM training described in the previous subsection applies to differential expression of single genes. Hence, mRNA and miRNA data are treated separately providing separate SOMs. One can combine both data sets if measured in the same series of condition by substituting Δe_{nm} by all pairwise products of the expression values of the mRNA and miRNA genes, $\mathrm{cov}_{n1n2m} = \Delta e^{mRNA}_{n1m} \cdot \Delta e^{miRNA}_{n2m}$ (*n1*=1…*N1* and *n2*=1…*N2* are the gene indices of mRNA and miRNA expression values, respectively). The size of the data increases from *N1*~10^4 (mRNA) and N2~10^3

(miRNA) to $N1 \cdot N2$~10^7 (combined miRNA/mRNA) which usually exceeds the maximum data capacity of our software application (~5x10^4) by several orders of magnitude.

One option to handle this problem is to study miRNA/mRNA metagene pairings instead of pairs of single genes, i.e. $\mathrm{cov}_{k1k2m} = \Delta e_{k1m}^{mRNA} \cdot \Delta e_{k2m}^{miRNA}$ (*k1*=1…*K1* and *k2*=1…*K2* are the metagene indices of the mRNA and miRNA expression SOM, respectively). Then, cov_{k1k2m} defines the sample-specific covariance term of both data sets which, in turn, is related to the correlation coefficient of the metagene profiles *k1* and *k2*, $r_{k1k2} = \sum_m \mathrm{cov}_{k1k2m} / \sqrt{\sum_m (\Delta e_{k1m}^{mRNA})^2 \cdot \sum_m (\Delta e_{k2m}^{miRNA})^2}$. Large positive values of cov_{k1k2m} thus indicate concerted up- or down-regulation of mRNA and miRNA expression whereas negative values refer to antagonistic changes of both RNA species in sample *m*. The size of the combined data is $K1 \cdot K2$~ 10^6, which requires further reduction. We filtered miRNA and mRNA metagene profiles for largest variance and population with single genes. Particularly, metagenes are selected whose variance and population exceeds the respective mean value averaged over all metagenes.

SOM training then provides meta profiles for the combined data, cov_{km}, in analogy to the separate data sets as described in the previous subsection. Each meta-covariance feature describes a characteristic profile of the combined expression data observed in the data set. The respective microcluster contains combinations of mRNA and miRNA with similar profiles of their combined expression values as the meta feature, i.e. $\mathrm{cov}_{k1k2m} \propto \mathrm{cov}_{km}$.

The third row of images in Figure 3.1 shows the combined SOM portraits of miRNA/mRNA coexpression. The individual portraits and especially the mean map are more heterogeneous showing more spots than the respective miRNA and mRNA maps. The mean combined portrait shows more than five spots of preferentially concerted (red) changes in normal tissues and anti-concerted (blue) changes in tumor samples. These different trends presumably reflect deregulation of mRNA and miRNA expression in the respective spots: concerted changes between both species in healthy tissues change into anti-concerted changes in the tumor samples.

All subsequent analyses apply to metagenes of single differential expression, Δe_{km}, and to metagenes of combined differential expression, cov_{km}, as well. Special analyses considering the pairwise character of the combined data will be addressed separately.

Another alternative option of studying miRNA/mRNA coexpression makes use of spot-spot correlation coefficients, $r_{s1s2} = \sum_m \left(\Delta e_{s1m}^{mRNA} \cdot \Delta e_{s2m}^{miRNA} \right) / \sqrt{\sum_m (\Delta e_{s1m}^{mRNA})^2 \cdot \sum_m (\Delta e_{s2m}^{miRNA})^2}$, where, e.g., Δe_{s1m}^{mRNA} is the mean expression of spot *s1* averaged over all metagenes included. The cross correlation coefficient is determined for all pairwise combinations of miRNA and mRNA spots taken from the respective overexpression summary maps (see next subsection, *s1* and *s2* are the respective spot indices). This analysis reduces the number of combined features to about 10^2 due to the relatively small number of spots identified in the miRNA and mRNA SOM. An example of such analysis is provided in the accompanying contribution (29).

3.8. Spot overviews

The texture of the SOM visualizes ‘local’ expression properties in terms of spots due to high and low expression levels in the individual phenotypes. For an overview we select all overexpression spots observed in at minimum one of the phenotypes into one overexpression summary map. The respective underexpression summary map provides an overview about all observed underexpression spots.

Figure 3.2 shows the over- and underexpression summary maps of the LU-data set. Note that the spot patterns are more diverse compared with the mean maps shown in Figure 3.1 because alternating over- and underexpression data in the set of samples are not averaged out. The single (mRNA and miRNA) data maps illustrate that regions overexpressed in normal tissues typically become underexpressed in tumor samples and vice versa. Spots showing such antagonistic changes are candidates for tumor-relevant features. The combined data again provides a slightly different picture with mostly different patterns of the over- and underexpression spots.

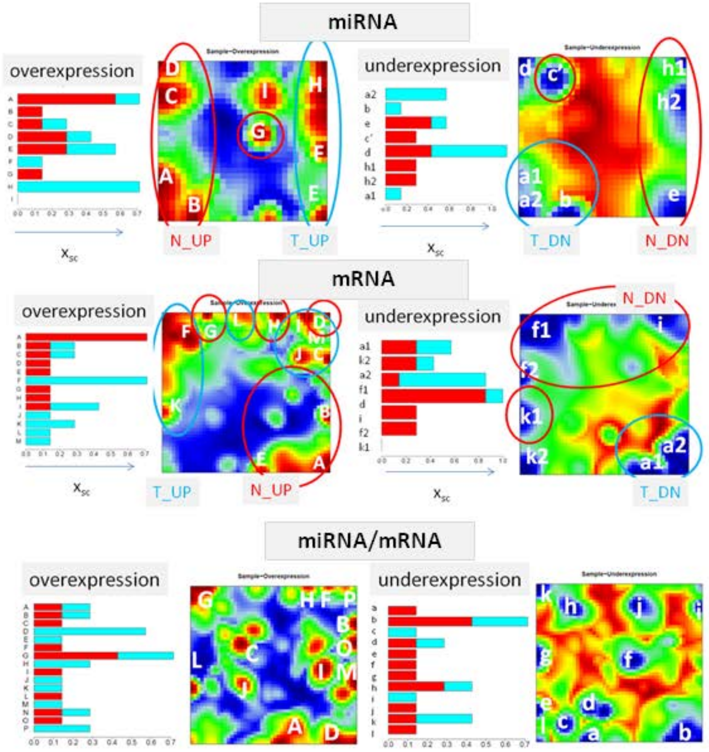

Figure 3.2: Over- and underexpression spot summary maps of the miRNA, mRNA and combined miRNA/mRNA expression portraits shown in Figure 3.1. Spots are annotated using capital letters (overexpression) or lower case letters (underexpression) where the latter ones are chosen to agree with respective overexpression spot annotation for spots having a big overlap of metagenes in both summary maps. Regions up- or downregulated in normal (N_up, N_down) and tumor (T_up, T_down) samples are indicated by the ellipses. The barplots provide the fractional abundance of each spot in healthy (red) and tumor (cyan) tissues.

The abundance of each spot (*s*=A,B,…) in the individual portraits is calculated as its relative frequency of appearance in the samples of each class (*c*=N, T), $x_{sc} = n_{sc} / N_c$, where the numerator and denominator define the number of sample portraits n_{cs} showing a particular spot and the total number of samples per class N_c, respectively. The spot abundances are represented as stacked bar for each spot. The integral abundance, $X_s = \sum_c x_{sc}$, can be interpreted as the mean number of classes seen by the respective spot. Its maximum possible value equals the number of classes considered.

The respective barplots in Figure 3.2 show the clear preference of the selected spots to overexpress metagenes in normal tissues (N_UP) and to underexpress these metagenes in tumor tissues (T_DN) and vice versa (i.e. N_DN and T_UP). These class-specific spots tend to accumulate in regions detected as differentially expressed in the mean portraits (see red and blue ellipses in Figure 3.2 referring to N_UP and T_UP spots, respectively). On the other hand, T_UP and N_UP mix in the right upper corner of the mRNA-overexpression map reflecting more complex patterns than the mean maps in Figure 3.1.

The spots observed in the combined data sets are more or less unique for the individual samples, i.e. only a few of them are found in more than one sample. ‘Coexpression’ spots are more abundant in tumor samples whereas ‘antiexpression’ spots are more abundant in normal samples. The non-equivalence of ‘over-‘ and ‘underexpression’ spots reflects the fact that concerted modes dominate in tumor samples whereas anti-concerted modes dominate in healthy tissues.

3.9. Supporting maps

The individual SOM portraits partly mask information about the single gene level. For example, it remains unclear how many single genes are associated with one particular metagene, and thus how importantly it contributes to the overall expression landscape. We therefore defined a series of supporting maps which provide additional information about selected properties of the metagene-miniclusters:

The population map plots the number of real genes per metagene in logarithmic scale, log n_k. The variance map illustrates the variability of the metagene profiles, $\mathrm{var}_k = \sum_m \left(\Delta e_{km} - \Delta e_{k\bullet}\right)^2 / (M-1)$, where $\Delta e_{k\bullet} = 0$ is the respective mean expression averaged over the profile. The entropy map plots the standard entropy of each metagene profile, $h_k = -\sum_m p_{km} \log_2 p_{km}$ where p_{km} is the relative frequency of the three levels of gene expression: overexpression, underexpression and non-differential expression of metagene k. Therefore expression values of the metagene profiles of each sample are assigned to one of the three levels by application of a defined threshold (here the 25- and 75-percentile of all metagene expression values was used). h_m is restricted to values in the interval [0, $\log_2$ 3]. An entropy value of 0 represents a perfectly 'ordered' state, where all metagenes are assigned to only one of the expression levels. Contrary, maximum value of $\log_2 3 \approx 1.58$ is reached when metagenes uniformly distribute over the three levels.

The 217 miRNAs studied in the LU-data set distribute over K=30x30 metagenes giving rise to a relatively sparse populated map with a series of empty metagenes (Figure 3.3, left part). The mean number of single genes per metagene is G/M~ 0.24, with G being the number of genes per metagene. The mRNA map of size K=50x50 contains 15,500 single genes. It is much denser populated (G/M~ 6.2) with less empty metagenes which however accumulate into larger empty areas. These empty regions usually separate different types of profiles defining different modes of regulation. Such modes appear as separate spots of highly variant (red) metagenes in the variance map (Figure 3.3, middle part). These maps show that the (Euclidean distance-based) SOM algorithm not only clusters correlated expression profiles together in different regions of the SOM but also genes of virtually invariant profiles. These two groups of profiles tend to occupy different regions either along the edges or in the central area of the mosaic image, respectively.

The combined map of size K=50x50 collects 6,100 miRNA/mRNA features (G/M~ 2.4). It is much more fragmented into regulatory modes as indicated by the spot pattern of the variance map. The blue spots refer to relatively invariant meta-profiles forming another kind of separators between different regulatory modes.

The entropy maps (Figure 3.3, right part) closely resemble the respective variance maps. The color of each pixel visualizes the information content of each metagene profile. Invariant profiles contain no sample-specific information (low entropy, blue areas) whereas variant profiles are more informative (red areas). Interestingly, the entropy map is more structured than the variance map.

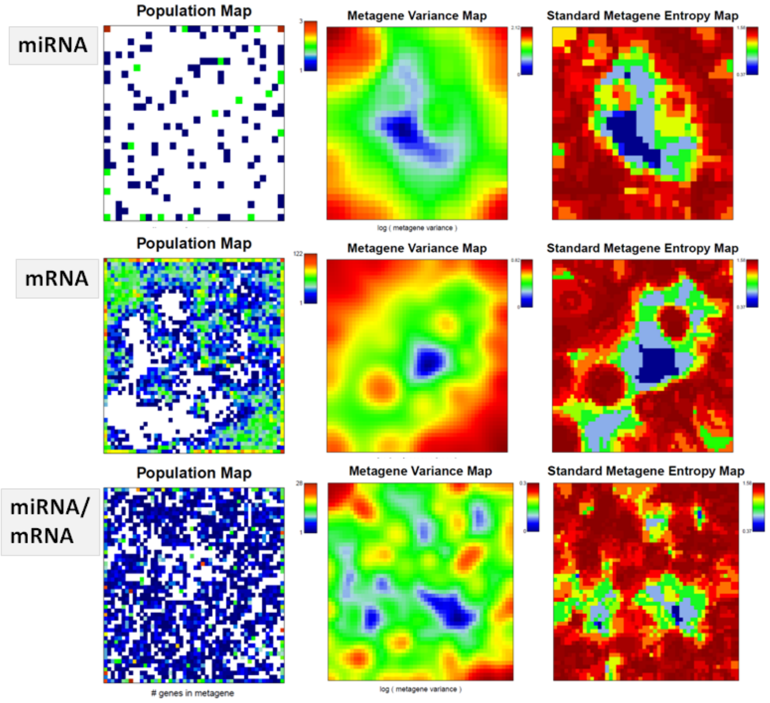

Figure 3.3: Supporting maps miRNA, mRNA and combined miRNA/mRNA expression data of the LU-data set.

3.10. Global portrait characteristics

In order to estimate global properties of the expression landscape of every phenotype we calculated the variance of metagene expression values in each SOM image, $\mathrm{var}_m = \sum_k \left(\Delta e_{km} - \Delta e_{\bullet m}\right)^2 / (K-1)$, and its entropy, $h_m = -\sum_k p_{km} \log_2 p_{km}$ where p_{km} is the relative frequency of expression as described above for the supporting maps. Here, the relative frequency refers to the expression state *m* and not to the expression profile *k*. The global entropy thus characterizes the information content of each portrait. Both, the variance and the entropy assess the expression landscape of phenotype *m* as seen by the SOM portrait. The variance estimates the variability of the metagene expression and the entropy its information content, or in other words, its degree of ordering.

The bar plots in Figure 3.4 illustrate that entropy and variance of the expression states mutually correlate. Entropies accentuate highly ordered states (low entropy) whereas variances accentuate strongly variable ones. Entropies and variances of miRNA and mRNA expression states are almost comparable in their order of magnitude reflecting similar complexities of the respective expression landscapes. The combined miRNA/mRNA covariance-patterns slightly lose information (i.e., they become more evenly distributed) for part of the cancer tissues (for bladder, prostate, uterus, lung and breast cancer). This trend might reflect the partial loss of mutual co-regulation between miRNA and mRNA expression in the diseased tissues.

Other global properties of the expression landscapes are the mean spot number detected per class, the mean 'shape' of the spots and the fraction of overexpressed metagenes. The shape is defined as, $shape_m = A_m / L_m^{\ 2}$ where A_m denotes the number of tiles included in all spots observed and L_m is the number of tiles forming the border of the spots with at minimum one adjacent tile outside the spots. It judges the fuzziness of the observed spots in the portraits. One finds that the number of spots per mRNA-portrait (and thus the number of distinct regulatory modes) slightly exceeds that of the miRNA-portraits. Interestingly, the number of overexpressed miRNA-metagenes is smaller in most of the cancer tissues than in the respective healthy tissues. This relation reverses for the number of overexpressed mRNA-metagenes: i.e. one observes an increased number of overexpressed metagenes in cancer. This trend presumably reflects the antagonistic effect of miRNA and mRNA expression expected: Particularly, global downregulation of miRNA expression in cancer (compared with healthy tissues) associates with global upregulation of mRNA expression. The global down regulation of miRNA expression in tumors was reported also in the original paper (11). It has been hypothesized that that global miRNA expression reflects the state of cellular differentiation and that the abrogation of which is a hallmark of cancers.

This trend also corresponds to the loss of information of the covariance landscape discussed above. Note also that the number of spots observed in the cancer covariance landscapes in most cases exceeds that observed for normal tissues. This difference can be interpreted in terms of a loss of concerted expression of both, miRNA and mRNA. The larger number of spots in the cancer samples is paralleled by a decreased *shape*-parameter which simply reflects the increase of fuzziness with increasing spot number.

The global characteristics of the expression landscapes of miRNA, mRNA and of their combination thus describe distortions of the gene regulation patterns due to disease.

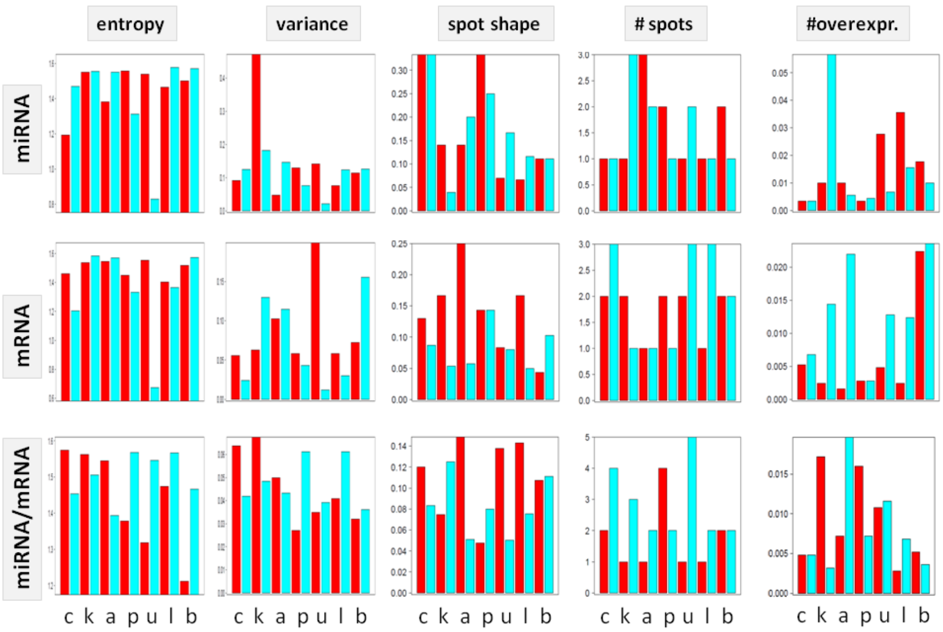

Figure 3.4: Sample related metagene-entropy, -variance, spot shape, spot number and area of overexpressed regions of the sample portraits of the LU-data set of healthy (red) and tumor (cyan) tissues (c: colon; k: kidney; a: bladder; p: prostate; u: uterus; l: lung; b: breast).

3.11. Phenotype similarity analysis

This task aims at establishing the mutual relations between the phenotypes studied to group them into different classes. Particularly, we analyze the hierarchy of similarities and estimate the mutual distances between the expression states. Similarity analysis compares the expression states as seen by the SOM portraits. It consequently uses the metagenes instead of single genes as the basal data set. Using meta- instead of single genes is advantageous because it improves the representativeness and resolution of the results (10).

We applied second-level SOM analysis as proposed by Guo et al. (30) as a first option to visualize the similarity relations between the individual SOM-metagene expression patterns. The method clusters the samples and not the genes as in first-level SOM analysis. In addition, we characterize similarities using the neighbor-joining algorithm based on the Euclidean distances in terms of similarity trees (31).

The separate miRNA and mRNA data well separate healthy and tumor samples in both plots in most cases (Figure 3.5). This result reflects the tumor-specific over- and underexpression of selected spots discussed above (see Figure 3.2). Note that the normal and tumor samples of colon (miRNA) and prostate (mRNA) are found at the same branch of the respective trees reflecting the close similarities of the respective SOM portraits (see Figure 3.1). The question about the origin of this result is beyond our study. They possibly reflect weak tumor effects or strong contamination of the tumor samples with healthy tissue. The combined miRNA/mRNA data allows virtually no differentiation between tumor and healthy tissues based on a common set of features. Essentially each sample obeys its own specifics. This result reflects the lack of clearly resolved tumor-specific spots in the sample portraits (Figure 3.1).

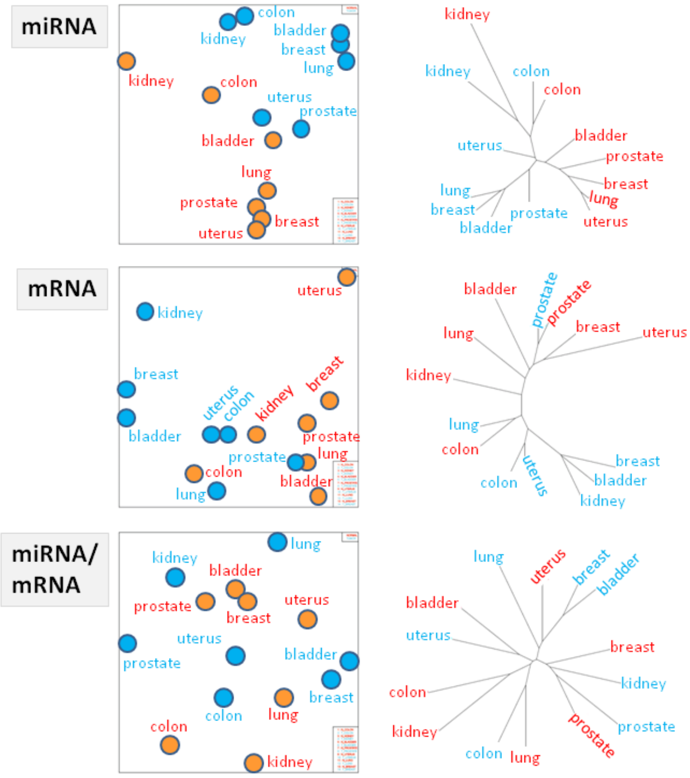

Figure 3.5: Similarity analysis using 2nd level SOM (left part) and nearest neighbor joining trees (right part) of miRNA, mRNA and the combined cov-expression landscapes of healthy (red) and tumor (cyan) samples.

3.12. Differential expression and concordance analysis

The tasks described above characterize the metagene expression landscapes. Usually researchers are interested to select single genes associated with each of the spot cluster detected. Two different analyses are implemented in our method. Concordance analysis estimates the similarity between each metagene profile and the profiles of the associated single genes and ranks them with decreasing agreement using either correlation- or distance-based significance scores for each spot cluster. Concordance analysis thus refers to the whole profiles of all different expression states studied. In contrast, differential expression analysis estimates the most prominently up- and/or downregulated single genes in each spot cluster using scores based on fold change or t-statistics.

Correlation-based concordance simply uses the correlation coefficient between each single gene and the respective metagene, $r_{nk} = \sum_m (\Delta e_{nm} \cdot \Delta e_{km}) / \sqrt{\sum_m \Delta e_{nm}^2 \cdot \sum_m \Delta e_{km}^2}$, and then estimates the p-value for each gene using the t-statistics, $t_{nk} = r_{nk} / \sqrt{(1 - r_{nk}{}^2)/(M-2)}$. Distance based concordance uses the sum of squared normalized residual expression, $d_{nk}^2 = \sum_m \left((\Delta e_{nm} - \Delta e_{km})^2 / SD_{nm}{}^2 \right) / (M-1)$ (SD is the regularized standard deviation of the gene expression in state *m*, see below). Significance is then estimated using χ^2 statistics. The latter distance-based measure allows to identify similarities between virtually invariant profiles whereas correlation-based measures preferentially select highly variant profiles.

For differential expression analysis a large multitude of various methods are available to assess statistical significance. As the standard method we apply a regularized t-score on the single gene level, $t_{nm} = \Delta e_{nm} / (\sqrt{SD_{nm}} / R_m)$. The regularized standard deviation, $SD_{nm} \approx \sqrt{\lambda \sigma_{nm}{}^2 + (1-\lambda)\sigma^{LPE}(e_{nm})^2}$, is calculated as weighted mean of the 'individual' standard deviation of the expression of each gene (σ_{nm}) and of a locally pooled error value ($\sigma^{LPE}(e_{nm})$). Both values are combined using the empirical scaling factor λ= 0.5. Such regularized t-scores consistently lead to relatively accurate gene rankings which might outperform simple t-statistics or FC-scores (32). The regularized t-statistics transforms into p-values characterizing the significance of differential expression for each gene assuming Student's t-distribution. Consideration of the density distribution of the p-values of all genes in each phenotype allows to transform the p-values into false discovery rates to control the number of false discoveries in the multiple testing problem (33). Differential features extracted from the LU-dataset are given in the accompanying paper (29).

3.13. Gene set enrichment analysis

Gene set analysis aims at evaluating the relevance of selected predefined sets of genes in the expression landscapes of the phenotypes studied. A gene set usually collects genes of common functional context together. This context is given by independent knowledge such as the Gene Ontology (GO) classification (e.g. according to selected GO-terms such as 'biological process', 'molecular function' or 'cellular component'), chromosome location, involvement into biochemical pathways or independent gene expression studies on diseases or toxic effects. For miRNA/mRNA coexpression studies mRNA-targets for one selected miRNA provide specific mRNA-target sets or vice versa, sets of miRNA affecting the same mRNA are collected together into miRNA-sets.

Basically, gene set analysis estimates the enrichment of genes of the set within a list of genes which is obtained independently. Our SOM method provides a natural choice of gene lists in terms of the genes contained in the spot clusters defined above. Particularly, each gene studied is classified according to two memberships leading to a 2×2 contingency table for further testing: firstly, its membership in the set of functionally related genes of length N_{set} (N_+ 'positive' genes in list *and* set and (N_{set} - N_+) 'negative' genes in set but *not* in list) and, secondly, its membership in the respective list of length N_{list} ((N_{list}- N_+) genes in list but *not* in set; (N- (N_{set}+N_{list})+N_+) genes not in list and set as well). The intersection of the set and the list is given by the number of 'positive' genes, N_+. For any gene set, right-tailed modified Fisher's exact test was used to determine whether the number of genes with this set is overrepresented in a particular list of genes included in a spot-cluster. The hypergeometric distribution then provides a p-value for each set and spot which estimates the cumulative probability to find a stronger overlap between the genes in a spot cluster and the set than expected by chance given a certain total number N of genes studied (28). This 'overrepresentation'-analysis assesses the probability to find more members of the set in the list compared with their random appearance.

3.14. Program and availability

The SOM method is implemented as R program 'oposSOM' available on CRAN (Comprehensive R Archive Network, http://cran.r-project.org/) and on our website http://som.izbi.uni-leipzig.de.

4. Notes, conclusions and outlook

SOM machine learning enables analysis of miRNA expression landscapes. The method extracts differentially expressed single features, their functional context and also characterizes global properties of expression states and profiles. Despite the relatively small number of miRNAs, their expression landscapes are of similar heterogeneity as that of the much more numerous mRNAs in the systems studied. Application of SOM portraying to miRNA expression was performed for the first time to our best knowledge. Also the combined analysis of miRNA and mRNA differential expression using covariance terms is novel. It provides a very detailed resolution of the coexpression landscape which however awaits for further discovery and interpretation. SOM clusters features (miRNA- and mRNA-expression or miRNA/mRNA covariance) of similar profiles together. Enrichment techniques allow association of functional themes with each of the clusters. The spectrum of available gene sets can be extended in future studies by taking into account sets of mRNA targets of single miRNA, sets of miRNA targeting the same mRNA or sets of miRNA regulated in the same functional context. The sets might be collected using computational as well as experimental methods. Of special interest in this context are new experimental techniques such as high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) that directly decode mRNA-miRNA interactions (34).

Acknowledgements:
This publication is supported by LIFE –Center for Civilization Diseases, Universität Leipzig. LIFE is funded by the European ERDF fund and by the Free State of Saxony.

5. References

1. Mendes, N. D., Freitas, A. T., and Sagot, M.-F. (2009), Current tools for the identification of miRNA genes and their targets, *Nucleic Acids Research* **37,** 2419-33.
2. Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J., and Zhang, B.-T. (2006), miTarget: microRNA target gene prediction using a support vector machine, *BMC Bioinformatics* **7,** 411.
3. Wang, X., and El Naqa, I. M. (2008), Prediction of both conserved and nonconserved microRNA targets in animals, *Bioinformatics* **24,** 325-32.
4. Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C., and Showe, M. K. (2007), Naïve Bayes for microRNA target predictions—machine learning for microRNA targets, *Bioinformatics* **23,** 2987-92.
5. Heikkinen, L., Kolehmainen, M., and Wong, G. (2011), Prediction of microRNA targets in Caenorhabditis elegans using a Self-Organizing Map, *Bioinformatics*.
6. Kohonen, T. (1982), Self-organized formation of topologically correct feature maps, *Biological Cybernetics* **43,** 59-69.
7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999), Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proceedings Of The National Academy Of Sciences Of The United States Of America* **96,** 2907-12.
8. Törönen, P., Kolehmainen, M., Wong, G., and Castrén, E. (1999), Analysis of gene expression data using self-organizing maps, *FEBS Letters* **451,** 142-46.
9. Eichler, G. S., Huang, S., and Ingber, D. E. (2003), Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles, *Bioinformatics* **19,** 2321-22.
10. Wirth, H., Loeffler, M., von Bergen, M., and Binder, H. (2011), Expression cartography of human tissues using self organizing maps, *BMC Bioinformatics* **12,** 306.
11. Lu, J., et al. (2005), MicroRNA expression profiles classify human cancers, *Nature* **435,** 834-38.
12. Yin, J. Q., Zhao, R. C., and Morris, K. V. (2008), Profiling microRNA expression with microarrays, *Trends in Biotechnology* **26,** 70-76.
13. Wang, Z., and Yang, B. (Eds.) (2010) MicroRNA Expression Detection Methods, Springer, Springer, Heidelberg Dordrecht London New York.
14. Kong, W., Zhao, J.-J., He, L., and Cheng, J. Q. (2009), Strategies for profiling MicroRNA expression, *Journal of Cellular Physiology* **218,** 22-25.
15. Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P., and 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.
16. Sato, F., Tsuchiya, S., Terasawa, K., and Tsujimoto, G. (2009), Intra-Platform Repeatability and Inter-Platform Comparability of MicroRNA Microarray Technology, *PLOS one* **4,** e5540.
17. Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., and Hoffmann, S. (2011), DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments, *Nucleic Acids Research* **39,** W112-W17.

18. Linsen, S. E. V., et al. (2009), Limitations and possibilities of small RNA digital gene expression profiling, *Nat Meth* **6,** 474-76.

19. Binder, H., Preibisch, S., and Berger, H. (2009) *in* "Methods in Molecular Biology" (Grützmann, R., and Pilarski, C., Eds.), Vol. 575, pp. 376-407, Humana Press, New York.

20. Nelson, P. T., Wang, W.-X., Wilfred, B. R., and Tang, G. (2008), Technical variables in high-throughput miRNA expression profiling: Much work remains to be done, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1779,** 758-65.

21. Yuan, J., Reed, A., Chen, F., and Stewart, C. N. (2006), Statistical analysis of real-time PCR data, *BMC Bioinformatics* **7,** 85.

22. Meacham, F., Boffelli, D., Dhahbi, J., Martin, D., Singer, M., and Pachter, L. (2011), Identification and correction of systematic error in high-throughput sequence data, *BMC Bioinformatics* **12,** 451.

23. Meyer, S., Pfaffl, M., and Ulbrich, S. (2010), Normalization strategies for microRNA profiling experiments: a ‘normal’ way to a hidden layer of complexity?, *Biotechnology Letters* **32,** 1777-88.

24. Chang, K., Mestdagh, P., Vandesompele, J., Kerin, M., and Miller, N. (2010), MicroRNA expression profiling to identify and validate reference genes for relative quantification in colorectal cancer, *BMC Cancer* **10,** 173.

25. Bolstad, B. (2002), Probe Level Quantile Normalization of High Density Oligonucleotide Array Data, *preprint***,** (8).

26. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002), Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments, *Stat. Sin.* **12,** 111-39.

27. Smyth, G., and Speed, T. (2003), Normalization of cDNA Microarray Data, *Methods* **31,** 265-73.

28. Wirth, H., von Bergen, M., and Binder, H. (2012), Mining SOM expression portraits: Feature selection and integrating concepts of molecular function *BioData Mining* **5:18**.

29. Cakir, V., Wirth, H., Hopp, L., and Binder, H. (2013), miRNA expression landscapes in stem cells, tissues and cancer *Methods of Molecular Biology* **accompanying chapter in this issue**.

30. Guo, Y., Eichler, G. S., Feng, Y., Ingber, D. E., and Huang, S. (2006), Towards a Holistic, Yet Gene-Centered Analysis of Gene Expression Profiles: A Case Study of Human Lung Cancers, *Journal of Biomedicine and Biotechnology* **2006,** Article ID 69141.

31. Saitou, N., and Nei, M. (1987), The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution* **4,** 406-25.

32. Opgen-Rhein, R., and Strimmer, K. (2007), Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach, *Statist. Appl. Genet. Mol. Biol.* **6**.

33. Strimmer, K. (2008), A unified approach to false discovery rate estimation, *BMC Bioinformatics* **9,** 303.

34. Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009), Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps, *Nature* **460,** 479-86.