Profiling of Genetic Switches using Boolean Implications in Expression Data

M.Volkan Çakır[1,*], Hans Binder[1], Henry Wirth[1]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig, Härtelstr. 16 – 18, 04107 Leipzig, Germany

Summary

Correlation analysis assuming coexpression of the genes is a widely used method for gene expression analysis in molecular biology. Yet growing extent, quality and dimensionality of the molecular biological data permits emerging, more sophisticated approaches like Boolean implications.

We present an approach which is a combination of the SOM (self organizing maps) machine learning method and Boolean implication analysis to identify relations between genes, metagenes and similarly behaving metagene groups (spots). Our method provides a way to assign Boolean states to genes/metagenes/spots and offers a functional view over significantly variant elements of gene expression data on these three different levels. While being able to cover relations between weakly correlated entities Boolean implication method also decomposes these relations into six implication classes.

Our method allows one to validate or identify potential relationships between genes and functional modules of interest and to assess their switching behaviour. Furthermore the output of the method renders it possible to construct and study the network of genes. By providing logical implications as updating rules for the network it can also serve to aid modelling approaches.

1 Introduction

Biochemical and molecular biological approaches along with correlation analyses have revealed numerous molecular mechanisms and pairwise relations between elements of the cellular machinery. Analytical efforts so far have mostly concentrated on correlation analysis based on the assumption that co-expression of genes is directly related to concerted regulation [1]. However focusing solely on the correlation measures entails a restriction of the search space by linear relations. Fortunately improving quality, growing volume and extent of expression data allows for further examination via alternative concepts. The complexity and dimensionality of the expression data requires an in-depth and detailed analysis, while imposing the challenge to conceive the big picture. Boolean implication analysis stands out as a straightforward yet robust and promising method to capture linear and non-linear pairwise relations between the features.

Boolean implication is a logical relation between two Boolean variables where a state of one variable implies a state of the other variable. Variables can refer to molecular-biological entities such as protein coding DNA regions, mRNA and proteins transcribed/translated from it. Distinct copy numbers, expression levels or cellular concentration ranges that correspond to certain biologically-relevant activities can be described as their states. In our case study, genes constitute variables while their expression values constitute their states.

[*] To whom correspondence should be addressed. Email: cakir@izbi.uni-leipzig.de

doi:10.2390/biecoll-jib-2014-246

Here we utilize an approach previously proposed by Sahoo et al. [2], [3] which reliably extracts ‘if-then-relationships’ between the states of pairwise combinations of genes from expression data. Such Boolean implications strongly extend the options of pairwise gene activity combinatorics beyond simple correlation measures.

While Sahoo et al. used gene expression data from thousands of microarrays to assess molecular interactions between pairs of genes and their conservation across species, we use a single expression microarray data set to reconstruct the underlying interaction network. On one hand it is hard to have several array studies available for certain experimental setups, e.g. certain cancer types, certain cell types and conditions. On the other hand it might be problematic to analyze expression data obtained from different platforms. In most cases it is thus necessary to stick with the data on-hand and to be able to reconstruct potential implications and the underlying network without additional experiments.

In this paper we carry out Boolean implication analysis for a microarray study on 221 mature aggressive B-cell lymphomas that is publicly available [4]. This heterogeneous disease shows a high variability in both molecular biological and clinical parameters. Burkitt’s lymphoma (BL) is defined and listed in the World Health Organization (WHO) classification of lymphoid tumours as an ‘‘aggressive B-cell non-Hodgkin’s lymphoma’’ (aggressive B-NHL) that is characterized by a high degree of proliferation, malignancy and deregulation of the c-MYC gene [5], [6]. The earliest cases that resembles BL tumour were recorded in 1910 in central sub-Saharan Africa [7]. This fact led to the term African lymphoma and consequently endemic BL. It took several years before MYC oncogene translocation and constitutive activation is detected in lesion of BL [8], [9].

BL is classically diagnosed by the presence of a monotonous infiltrate of medium-sized blastic lymphoid cells with a round nuclei, clumped chromatin and multiple nuclueoli, possessing a high proliferation and low apoptosis rate [10].

MYC, mutations of which constitutes one of the most significant characteristics of BL, is a sequence-specific DNA-binding transcription factor. What makes MYC critical is its acting mechanism especially in B cells: MYC acts as a transcriptional hub that is able to control ~15% of all genes via multiple sub-hubs that are connected to it [11], [12]. An interesting feature of MYC is its nonlinear acting mechanism. By the activity of MYC genes that are already being expressed in the absence of MYC tend to get strongly boosted in the presence of MYC, while genes that have low expression in the absence MYC get only a small boost in the presence of MYC [13]. The processes that are controlled by this hub especially include cell-cycle control, cellular transformation, growth[14], proliferation and apoptosis[11], and tumorigenesis through miRNAs[15], [16]. The role of MYC in these key processes seals dramatic effects of its activation: increased cell growth, proliferation and genomic instability[17] and reduces immunogenicity of the tumour cells[18], [19]. In normal cells MYC also induces apoptosis and this provides a balance between proliferative activity and apoptotic activity. However in BL cases the apoptosis-inducing activity of MYC is reduced.

Addition to MYC, many other chromosomal translocations, genetic and epigenetic alterations are identified in BL cases[9], [20]. Hummel and colleagues was able to introduce a ‘BL similarity index’, by using the expression levels of a set of genes and classify aggressive B-NHL into three classes that we used throughout this paper: molecular BL (mBL), intermediate cases, and non-molecular BL (non-mBL) [4].This collocation enables extraction of differentially interacting features corresponding to functional differences in the two cancer subtypes.

Computation of pairwise relationships for all possible gene combinations is computationally expensive and time consuming. Thus it is the limiting factor for exhaustive analysis of experiments using high-throughput technologies such as microarrays and RNA-sequencing.

Here expression values of several tens of thousands of genes are measured in parallel. Therefore we take advantage of a preceding analysis using self-organizing maps (SOMs) [21]–[24], which were successfully applied in different cancer study evaluations [25]–[29]. SOMs transform the original 'single gene related' data into meta-data of reduced dimensionality. Then mutually independent expression modules can be detected in the SOM-transformed data space. Importantly the two steps of dimension reduction, which reduce the number of relevant features by about four orders of magnitude using the concept of cluster-prototypes in terms of metagenes and so-called spot-modules, respectively, do not entail a loss of primary information. No features are explicitly excluded from analysis and are therefore accessible at any time during SOM analysis [21].

We utilize pairwise relations on both single gene as well as meta-gene level to deduce the logical linkage of the expression modules obtained from SOM analysis of the lymphoma data. The implications are then validated using results of previous descriptions of the underlying molecular mechanisms of the disease [4], [30], [31]. Finally the module level network of Boolean implications is mapped into the SOM space and compared to the module combinatorics observed.

2 Methods

2.1 Data

We employ a microarray study on mature aggressive B-cell lymphomas available under GEO accession number GSE4475 [4]. The study used biopsy specimens from 220 patients measured on *Affymetrix HG-U133A* microarrays (one patient was measured on two microarrays). These arrays measure the expression level of 22,283 gene-related probesets in parallel. For quality control, the samples contain at least 70 percent tumor cells. Hummel et al. then defined a transcriptional signature to distinguish molecular Burkitt's lymphoma (mBL) and non–molecular Burkitt's lymphoma (non-mBL) [4]. Out of the 220 lymphomas, 44 carry the mBL and 128 the non-mBL signature while 48 cases could not be assigned unambiguously to one of the two groups. They form an intermediate group, representing the transition between mBL and non-mBL cases [4], [26].

2.2 Preprocessing

First, raw probe intensity values of each of the 221 arrays are calibrated and summarized into one expression value $E_{i,j}$ per probe set using the hook method [32], [33]. The indices assign the gene number $i = 1 \dots N$ in sample number $j = 1 \dots M$ referring to the different patients. Then the expression values were translated into logarithmic scale: $e_{i,j} = \log_{10} E_{i,j}$. Finally, the expression values of all arrays were quantile-normalized such that they follow one common distribution [34].

Following the convention we use the data as numerical matrix of dimension $N \times M$. Throughout this paper a row of the matrix, $e_{i,j}$ with $i = const$, will be termed 'expression profile' of the respective gene i. The columns, $e_{i,j}$ with $j = const$ on the other hand will be termed 'states' referring to sample j under consideration.

2.3 Boolean relationships on gene level

To detect Boolean relations between two genes their expression profiles have to be transferred into binary data space. There exist different approaches to dichotomize the profiles into the states of 'low' and 'high' expression, respectively. Beside using a global threshold to divide the expression range into the two levels, the 'StepMiner' algorithm uses time course data to calculate an adaptive threshold for each gene individually [35]. However, the lymphoma data set contains independent samples without temporal order. Therefore we implemented an alternative straightforward and robust approach to dichotomize the expression level of each gene:

We first remove invariant genes from the data since they hamper a reliable detection of Boolean implications. Therefore we reject genes whose interquartile expression ranges do not exceed the threshold of 0.5 in $\log_{10}-$expression space. Only genes with a sufficient degree of differential expression are then used for subsequent Boolean implication analysis. The samples in each expression profile i were then ranked with monotonically increasing expression values such that the profile meets the condition $e_{i,j} \leq e_{i,j+1}$ for all $j = 1 \ldots (M - 1)$ (see Figure-1a). Then the threshold for dichotomization of gene i is estimated using

$$t_i = \frac{\sum_{j=1}^{M-1} w_{i,j} \cdot e_{i,j}}{\sum_{j=1}^{M-1} w_{i,j}}$$

Here $w_{i,j} = e_{i,j+1} - e_{i,j}$ is a weighting factor proportional to the increment between subsequent expression values. Hence, we estimate the threshold as the slope-weighted average of expression values. This threshold divides the profile into 'low' and 'high' expression values with $e_{i,j} < t_i$ and $e_{i,j} > t_i$, respectively. To ensure robustness a 'noise zone' of $[t_i - \delta_i,\ t_i + \delta_i]$ where $\delta_i = 0.1t_i$ is declared around the threshold (Figure-1a). Values located in this interval are assumed as 'intermediate' ones and are not further considered in subsequent analysis.

Boolean relations are then identified for each pair of genes, say A and B, as proposed by Sahoo et al. [2]: *(i)* A scatterplot is generated showing the expression values of gene A versus that of gene B in x-and y-coordinates, respectively (see Figure-1a). *(ii)* The thresholds t_A and t_B and the according noise zones distribute the genes over four quadrants which correspond to all the combinations of the expression levels of A and B: A_{low} & B_{high}, A_{low} & B_{low}, A_{high} & B_{low} and A_{high} & B_{high} in counter-clockwise order. *(iii)* The population of each of the quadrants with genes is estimated to identify 'sparse' quadrants with a significantly lower density of data points compared with the other quadrants. The number of expression values is counted for each quadrant, omitting those in the noise zone. The sparseness statistic is then computed as the ratio between the number of observed expression data per quadrant and the number of data expected by chance which depends on the population of the neighboring quadrant's and the total number of expression values. *(iv)* Finally the error of this estimation is calculated as a maximum likelihood estimate implying the observed count and those of the adjacent quadrants. *(v)* A quadrant is finally declared as sparse if the statistic exceeds and the error rate undercuts designated thresholds.

It is noteworthy that in some cases the expression profile of a gene is dominated by a very few extreme values, i.e. the expression values are not reasonably distributed over the expression range. We identify such outliers using an interquartile range (IQR)-based criterion [36] and discard the respective values from the scatterplots and the corresponding sparseness estimation. More precisely, expression values of gene i are designated as outliers if they meet the condition $e_{i,j} > q_{i,75} + 3IQR_i$ or $e_{i,j} < q_{i,25} - 3IQR_i$, respectively. Here $q_{i,25}$ and $q_{i,75}$ denote the first and third quartiles of gene i's expression values and $IQR_i = q_{i,75} - q_{i,25}$.

The sparsely populated quadrant(s) defines the relationship between the two involved genes (compare Figure-1b). For instance a sparse bottom left (low-low) quadrant reflects the situation that low values of gene A are mostly associated with high values of gene B. In other words, a given low expression value of gene A restricts the possible values of gene B to high ones. On the other hand if A is high then B can be either low or high without restriction. This defines the forward rule: 'If A is low, then B must be high', or as synonym 'A_{low} implies B_{high}' ('$A_{low} \Rightarrow B_{high}$') or in short notation 'LH'. Note that the latter 'LH'-designation applies also to '$B_{low} \Rightarrow A_{high}$' because a sparse low-low quadrant generates a symmetric pattern with respect to genes A and B given that there is no temporal order in the data. These Boolean implications can be interpreted as "if-then-rules" for pairs of genes. Thereby they are asymmetric relationships, as $A_{low} \Rightarrow B_{high}$ holds but $B_{high} \Rightarrow A_{low}$ does not. This inverse 'HL' implication applies if the high-high quadrant is sparse (see Figure-1b). A combination of 'LH' and 'HL' is given if both the low-low and high-high quadrants are sparse. This situation defines the 'Boolean opposite' ('OPP') relation fulfilling the implications '$A_{low} \Rightarrow B_{high}$' AND '$A_{high} \Rightarrow B_{low}$' (Figure-1b, upper part). The LH, HL and OPP relations can be considered as 'opposite-type relations' because low expression of one gene associates with high expression of another gene and vice versa. The second class of 'equivalence-type' relations is defined by 'Boolean equivalence' ('EQV'), LL and HH implications as illustrated in Figure-1b, lower part. In these cases high/low expression of one gene is associated with high/low expression of another gene.

Importantly, expression values can be virtually equally distributed over all four quadrants providing no sparse region, e.g. if the signals are relatively invariant and if their noise level exceeds the systematic changes of the signal values. In such cases the respective pair of genes lacks a Boolean implication.

Boolean implication analysis enables identification of molecular interaction processes [37]. Consider a toy example where genes A and gene B are strong transcription factors of gene C, then one of these transcription factors would be enough to activate transcription of gene C . Which means if A has a high expression value C will also be high. However if A has a low value C can have one of low or high values depending on the expression state of B. In this case A and C will expose the Boolean implication $A_{high} \Rightarrow B_{high}$. In yet another case where A and B are relatively weak, additively acting transcription factors of C, high values of A alone or high values of B alone are not sufficient to induce high expression of C. However if C has a high expression value we can infer that A $(or\ B)$ also has a high value since in this case high expression value of A $(or\ B)$ is necessary which in the end puts forward the relation $C_{high} \Rightarrow A_{high}(or\ C_{high} \Rightarrow B_{high})$. This conformity of Boolean implications to molecular interaction logic manifests a way of understanding biological data while it also offers a way to produce experimentally testable logical hypotheses for a set of genes of interest that constitute possible molecular interactions. However one must be aware of the fact that implications do not necessarily mean a molecular interaction as these relations can be a result of direct molecular interaction or a more complicated indirect relation in the overall network as well.

2.4 Metagenes and expression modules derived from SOM analysis

We implemented an analysis pipeline based on self-organizing maps (SOMs) and successfully applied it to a variety of different case studies referring to tissue systems [21], [38], certain cancer types [26] and stem cells [26]. The SOM analysis was further shown to be applicable also to different data types: Microarray mRNA and miRNA expression [27], [28], proteomic mass spectrometry [39], genomic SNP [26] and ChIP-seq [40] data.

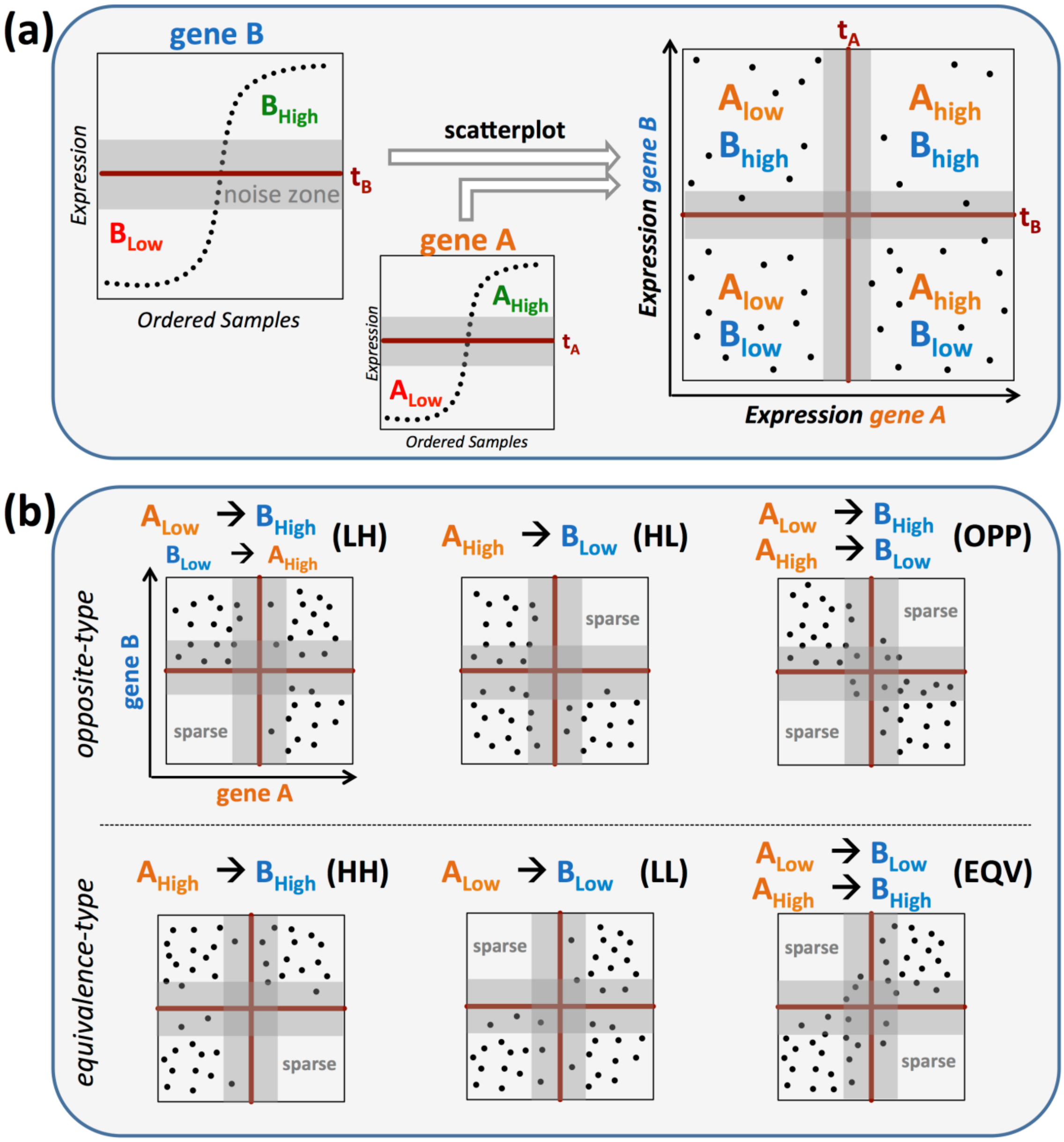

Figure 1: Boolean implications, definition and detection from pairwise gene relations.The ordered expression profiles of a pair of genes are divided into 'low' and 'high' expression levels by corresponding thresholds *t*. A paired scatterplot then yields four quadrants constituted by combination of the two genes and the two expression levels (panel a). Sparse quadrants indicate, depending on the positions, opposite type or equivalence type relationships between two genes (panel b). The abbreviations of the relations LH(Low-High), HL(High-Low), HH(High-High), LL(Low-low), EQV(Equivalence) and OPP(Opposite) are used throughout this manuscript.

The methods and algorithms of the SOM analysis were described in detail previously. In brief: The preprocessed data is gene-wise centered with respect to its mean expression value in all samples representing fold-change values in $\log_{10}$ −scale. A relative log-expression value of zero consequently indicates that a gene is expressed according to its mean expression value, while positive and negative values refer to over- and underexpression in the data set, respectively. These relative expression data are then used to train a self-organising map (SOM) [41]. It translates the high-dimensional input data given as $N \times M$ matrix into a $K \times M$ metadata matrix (K: number of so-called metagenes) of reduced dimensionality $K \ll N$ (N= 22,283 and K=2,500). The metagenes are arranged in a two-dimensional grid of resolution

50×50. During the SOM training algorithm, the metagene profiles are adapted, such that they resemble the real gene profiles. Thereby each metagene serves as a representative prototype of a 'minicluster' of real genes with similar expression profiles. The association of the genes to the metagenes is not fixed and becomes adjusted during the self-organizing process. It arranges the genes such that the degree of similarity between metagenes decreases with increasing distance in the map.

The expression state of each sample is visualized by color-coding the two-dimensional grid of metagenes according to their expression values in the respective sample. In this way individual 'SOM portraits' of each sample are generated by applying an appropriate color gradient (red to blue reflects over- to underexpression).

Owing to similarity of adjacent metagene profiles, the color patterns emerge as smooth textures which are characteristic for each sample and represent a fingerprint of its transcriptional activity. Individual expression patterns emerge as spots of similar colored tiles (see Figure-4a), which correspond to clusters of co-regulated genes. Note that the assignment of genes to metagenes and therefore to tiles of the underlying grid is identical in all sample portraits. So they can be directly compared to each other allowing immediate identification of unique or ubiquitous expression modes. Metagenes in the same spot are co-expressed in the samples studied. Metagenes in different, well-separated spots of a portrait are co-expressed in the particular sample but differently expressed in other samples. Importantly, utilization of metadata instead of the single gene data is advantageous regarding representativeness and noisiness in subsequent downstream analyses [21].

We define so-called spot modules representing clusters of neighboring over- (red) or under- (blue) expressed metagenes detected in at minimum one sample portrait. Such spots are determined by applying a simple 98/2-percentile criterion for over-/underexpression spots which selects the respective fraction of the metagenes showing largest/smallest expression in each sample. All spots detected were transferred into one master map for visualization of the global spot patterns of the sample series studied. Each spot represents an expression mode of a group of metagenes showing concerted expression. Thus, spot clusters provide a simple and intuitive approach for detection of expression modules. Note that it identifies gene clusters in an unsupervised fashion without necessity for prior definition of class prototypes or of a desired number of clusters. For the analyses described in this paper we restrict to overexpression spots because most underexpression spots overlay with them.

The SOM pipeline is publicly available as R-package 'oposSOM' on CRAN repository. The download link can be found on our SOM project page: http://som.izbi.uni-leipzig.de.

2.5 Boolean implications on metagene and expression module level

Detection of Boolean implications between pairs of gene clusters, such as metagenes or spot modules, is straightforward to reduce the total number of feature pairs to be assessed and thus to condense the resulting number of relevant relations in the system. It can be performed analogously as described for the genes above using expression profiles of metagenes or spot modules instead of those of single genes (see Figure-2, horizontal arrows). However the question arises if such information aggregation is preferable prior to or after Boolean implication analysis. In other words, we aim at evaluating on which data level information pooling provides the optimal relationships between the expression modules. Figure-2 illustrates the three different options to derive spot module relations from the original single gene expression profiles: Firstly, the SOM method can be utilized to aggregate single genes to metagenes and further to detect spot clusters and the corresponding spot expression profiles. Those are directly used to detect spot level Boolean implications (red arrows in Figure-2). Secondly, the metagenes obtained from SOM are used for pairwise implication analysis and subsequently aggregated to spot-wise relations (blue arrows inFigure-2, see below for details of the aggregation). And thirdly, the single gene profiles are used in Boolean implication analysis and their pairwise relations mapped and aggregated to spot relations (green arrows in Figure-2).

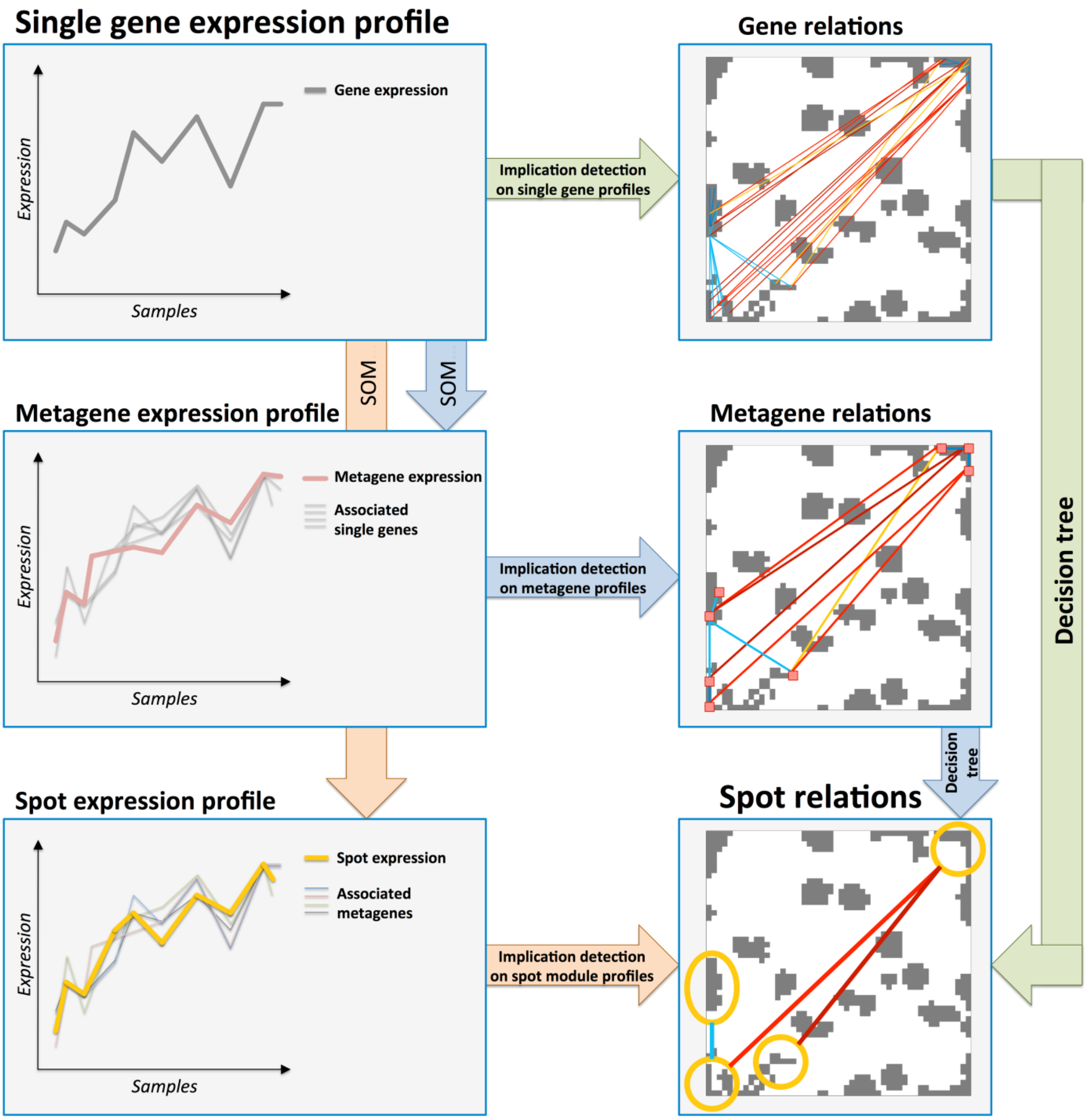

Figure 2: Three levels of expression information used to extract Boolean implications. Single genes, metagenes and spot clusters (from top to bottom): Vertical transitions are accomplished by information aggregation via SOM and the decision tree (left and right parts, respectively), horizontal transition by Boolean implication analysis on the respective data level. There are three alternative options to derive spot relations from the original single gene data; by using gene expression profiles, by using metagene expression profiles and by using spot expression profiles directly (see green, blue and orange horizontal arrows, respectively).

We aim at comparing spot relations obtained from the three competing options. Therefore, beside the direct implication detection using spot module profiles, we implemented an algorithm to derive spot module relations from gene (or analogous from metagene) level implications: First, it computes implications for all possible pairs of genes from a particular couple of spots (Figure-3a). The numbers of each of the six different implication relations are then counted. Finally, module related relations are identified by means of a decision tree using these counts (Figure-3b).

The decision tree first assigns the spot implication to equivalence- or opposite-type according to the majority of single gene implications types and then decides between the respective implications, e.g. OPP, HL or LH, according to the dominating single gene implication. To avoid annotation of pairs of spots based on too little information, we reject those spots whose total number of individual gene implications is smaller than the geometric mean of the numbers of genes contained in the respective spots. We chose the geometric mean as it provides, compared to arithmetic mean, a lower threshold if one of the spots is markedly less populated than the other.

The whole procedure is repeated for all pairwise spot module combinations. By utilizing this algorithm we are able to aggregate single gene implication information and to reliably assign appropriate Boolean implication relations to the spot modules.

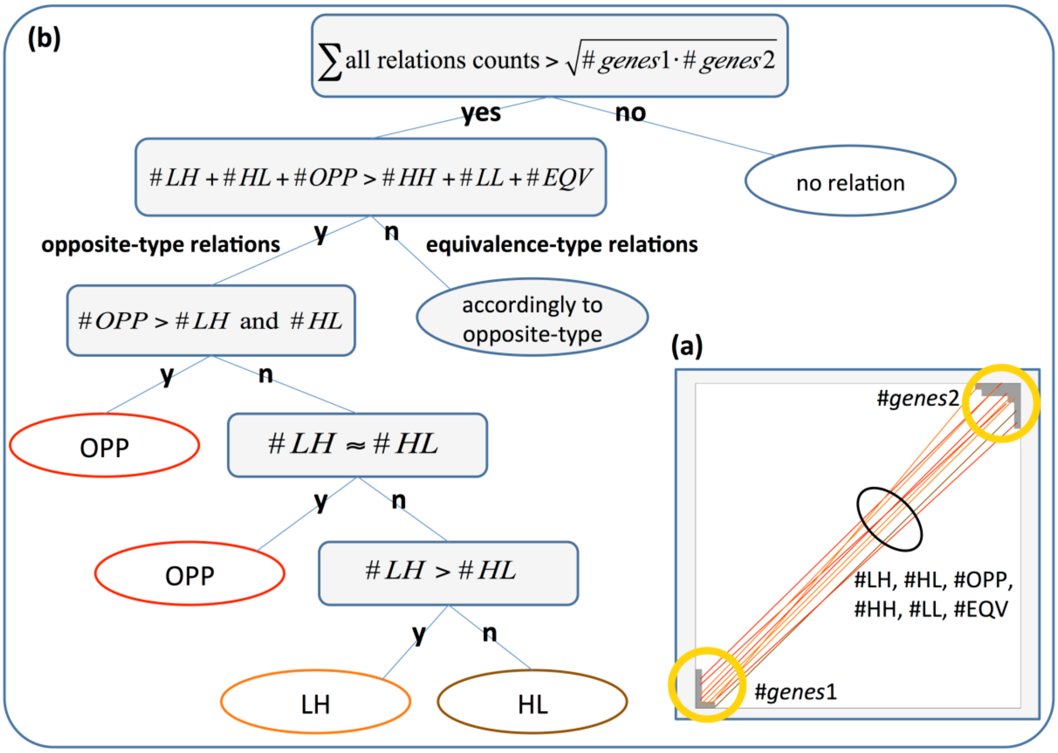
(b)
$\sum$ all relations counts > $\sqrt{\#genes1 \cdot \#genes2}$
yes
no
no relation
$\#LH + \#HL + \#OPP > \#HH + \#LL + \#EQV$
opposite-type relations
y
n
equivalence-type relations
accordingly to opposite-type
$\#OPP > \#LH$ and $\#HL$
y
n
OPP
$\#LH \approx \#HL$
y
n
OPP
$\#LH > \#HL$
y
n
LH
HL
(a)
#genes2
#LH, #HL, #OPP, #HH, #LL, #EQV
#genes1

Figure 3: Decision tree to identify spot module implications from gene level implications. Boolean implications detected and counted for all pairs of genes from two spots are identified first (panel a). Then, the obtained numbers are employed in a decision tree to designate the respective relation between the two spots evaluated (panel b). Note that the part of the tree referring to equivalence-type relations is likewise the part of the opposite-type relations.

3 Results and Discussion

3.1 Module selection using SOM

Expression data was downloaded and preprocessed as described in the methodical section. It contains 220 samples from mature aggressive B-cell lymphomas with about 22,000 genes measured per sample. This data was then used to train a self-organising map (SOM, for details see [21], [38]). It translates the high-dimensional expression data into meta-data of reduced dimensionality. The SOM has a resolution of $50 \times 50 = 2{,}500$ metagenes arranged in a two-dimensional quadratic grid. Each metagene serves as a representative of a cluster of real genes with similar expression profiles, whose number varies from metagene to metagene.

The expression state of each sample can then be visualized in terms of a mosaic image by color-coding the grid of metagenes according to their expression values in the respective sample. Figure-4a shows examples of such SOM expression portraits assigned to each of the molecular cancer subtypes classified independently [4]. These portraits serve as transcriptional fingerprints of the respective samples. They show smooth patterns with red and blue regions referring to over- and underexpressed genes, respectively. Green colored regions represent genes expressed on intermediate levels. The portraits can be directly compared each with another because each gene occupies the same position in each of the mosaic portraits. Visual inspection of the portraits reveals relatively homogeneous patterns for the main subtypes mBL and non-mBL, but also very heterogeneous patterns in the intermediate group. These observations are addressed in a previous publication [42]. It is also revealed that the portraits expose a significant amount of variation in their patterns and partly resemble both mBL and non-mBL characteristics. Please note that we adopted the approved classification of Hummel et al.[4], who utilized a linear score for rigorous differentiation of mBL and non-mBL cases. The lymphoma expression landscape however is continuous and hampers a reliable classification of border cases, such as the rightmost portrait of non-mBL in Figure 4a. Border cases are found to show patterns in their portraits which cannot be clearly attributed to either the mBL or non-mBL subtype or which can be attributed to both subtypes. We will address the diversity of lymphoma expression landscapes in another publication. Subtype-specific portraits are calculated as the mean images averaged over the metagene values in all samples of the respective subtype (Figure-4b). The mBL and non-mBL subtypes reveal a relatively simple texture with essentially one over- and one underexpression spot in two opposite corners of the map. This ‘binary’ spot pattern indicates that genes overexpressed in mBL become underexpressed in non-mBL and vice versa. Gene set enrichment analysis shows that genes related to the GO-terms ‘cell-cycle’ and ‘DNA-repair’ accumulate in the mBL overexpression spot in the top right corner whereas genes related to ‘cell adhesion’ and ‘inflammation/immune response’ dominate in the non-mBL overexpression spot in the opposite corner [42]. Note that also the individual sample portraits show relatively homogenous spot patterns with small deviation from the respective mean portrait.

The genes located in these two corners refer to antagonistic expression modes. In the SOM portraits such modules emerge as spots of concertedly over- or underexpressed genes on segregated positions. These modules can be identified using a simple threshold criterion which selects metagenes with expression values beyond the 98 and 2 percentiles in each portrait, respectively [21]. Figure-4c shows the overexpression spot modules detected in the lymphoma SOM. The expression profiles of each module are calculated as mean expression values averaged over all metagene values of each spot in each of the samples. The profiles of the spots in the top-right and bottom-left corners again reveal the antagonistic character of the respective expression modules (Figure-4d). In contrast, spots in the bottom-right and top-left corners however are virtually unspecific with respect to the subtypes.

In summary, disjunct modes of strongly over- or underexpressed genes emerge as spots in the SOM portraits. According to their particular expression profiles and assigned biological function, these modules reflect major expression changes in the data set in agreement with their classification into two main molecular subtypes. In the next step, module-related switching rules can then be derived from the data.

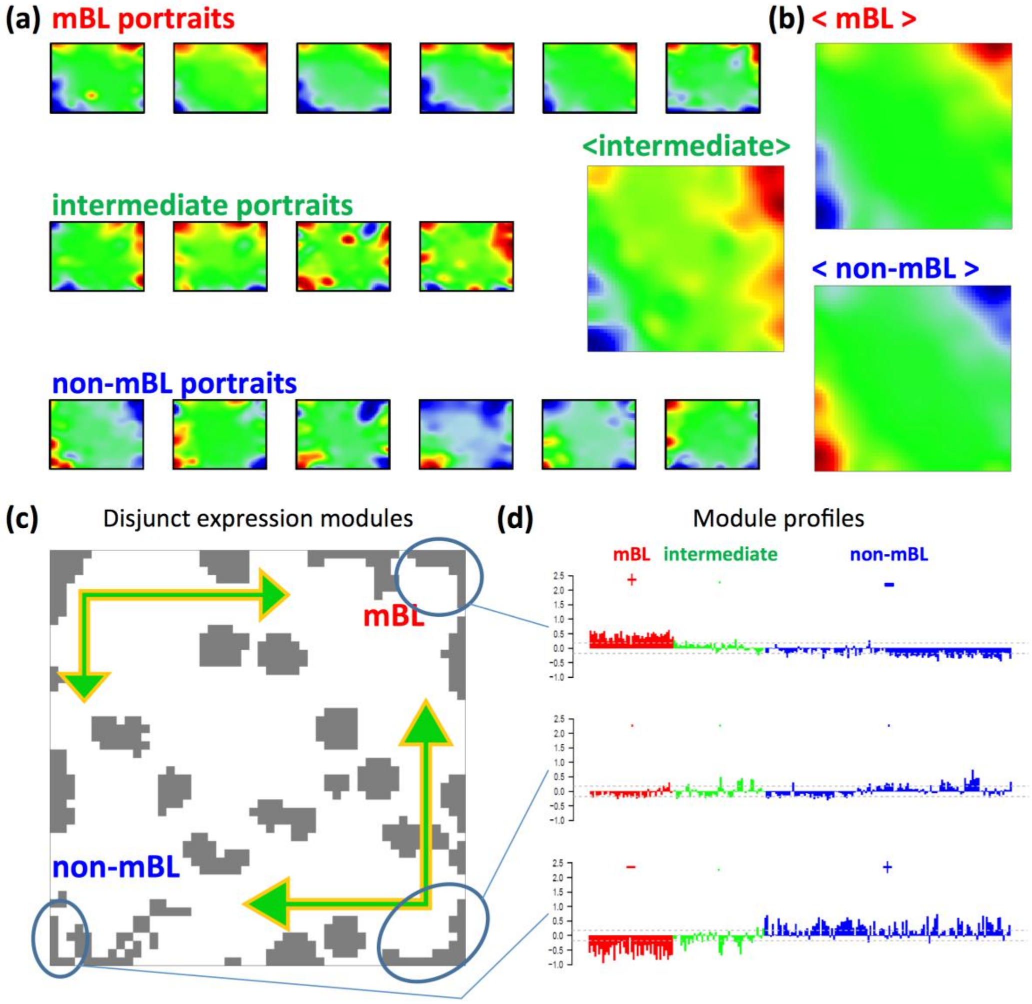

Figure 4: Results of the SOM analysis of the lymphoma data set. Expression portraits of selected samples are shown for each of the three subtypes (panel a). Red and blue colors indicate over- and underexpression in the samples, respectively, green color indicates intermediate levels of gene expression. Average subtype specific expression portraits reveal antagonistic expression modes of the mBL and non-mBL subtypes (panel b). Distinct spot-modules are identified (panel c), whose expression profiles change in a subtype-specific fashion (panel d).

3.2 Boolean implications between single genes

Boolean implication analysis is a suited approach for identifying relations between genes beyond simple correlation measures. To demonstrate the method we focus on selected mBL/non-mBL signature genes published by Hummel et al. [4] which localize in the antagonistic spot modules in the top-right and bottom-left corners of the SOM, respectively (see small maps in Figure-5a and b. Package KernSmooth is used for contour lines in figures).

The ranked expression values of the genes CD44 and BACH2 upregulated in non-mBL and mBL samples are shown in Figure-5a and b, respectively. The threshold t and the surrounding 'noise zone' divide the data into low and high expression values (see the Methods section for details). These data are redrawn as pairwise scatterplot, distributing the expression values of BACH2 along the x- and those of CD44 along the y-coordinates (Figure-5c). Here, the thresholds divide the plot into four quadrants and subsequent population analysis determines their occupation level. The number and position of the sparsely populated quadrants then defines the type of Boolean implication as described in the Methods section. For this example we identified a low-high-relation between BACH2 and CD44 ($BACH2_{low} \Rightarrow CD44_{high}$, LH) since CD44 always shows high expression when BACH2 is low, and vice versa, BACH2 is always high when CD44 is low.

PRDM10, another gene upregulated in mBL can be assigned to the opposite relation with respect to CD44 (Figure-5d). Note that CD44 is located in the mBL_up spot whereas PRDM10 and BACH2 are located in the antagonistic non-mBL_UP spot. Combinations of genes from these antagonistic spots are assigned to opposite-type Boolean implications, as expected. On the other hand, genes taken from the same spot are assigned to implications of the equivalence-type as illustrated in Figure-5e and f for genes selected from the non-mBL_UP spot. The type of relations for the genes in certain spots can be predicted depending on the relative positioning of the spots in SOM map while exposed implications by these genes can differ: CD44 exposes equivalence-type relations with NFKIBA and MDFIC, however it switches according to the EQV- and LL-implications respectively. In correlation analysis two types of relations, namely positive correlation and negative correlation, resolve into six different type of relations in implication analysis. Two of these six types, equivalent and opposite implications, often have high correlation coefficients whereas other four types, high-high, low-low, high-low and low-high implications, can have relatively low correlations coefficients (Figure-7d). This property gives Boolean implication analysis the ability to capture a bigger space of genewise relations while resolving relations in six qualitative classes.

3.3 Boolean implications between metagenes and between spot modules

The algorithm to detect Boolean implications from pairs of gene expression profiles can be easily applied also to metagene or spot module expression profiles. Using the respective metadata instead of the single gene data reduces the number of all pairwise combinations of features by four to six orders of magnitude.

Boolean implications on metagene level are detected using the same algorithm and thresholds as for gene level implications (Figure-6a-d, see Materials and Methods for details). Note that on the metagene level, a detected implication corresponds to the relation between two sets of genes that are represented by the respective metagene profiles. Analogously, implications between spots represent the relations between the mean expression profiles of two sets of metagenes (see Figure-6e and 6f).

However using sole mean expression profiles of metagenes in spots is a crude way to extract implications between spots since each spot encloses a range from a handful of genes to thousands of genes. Therefore we implemented a decision tree for extracting spot level implications from gene/metagene level implications (Figure-3, see next section, see Materials and Methods for details). Nevertheless positive and negative relations are well conserved on both metagene and spot level even after the two steps of dimensional reduction. Metagenes enclose similarly behaving genes while spots enclose similarly behaving metagenes, i.e. expression profiles of genes are organized in discrete sets of similarly behaving interrelated groups. Thus these similarity sets bring forth specific types of Boolean implications amongst them that in turn constitute the underlying interaction network.

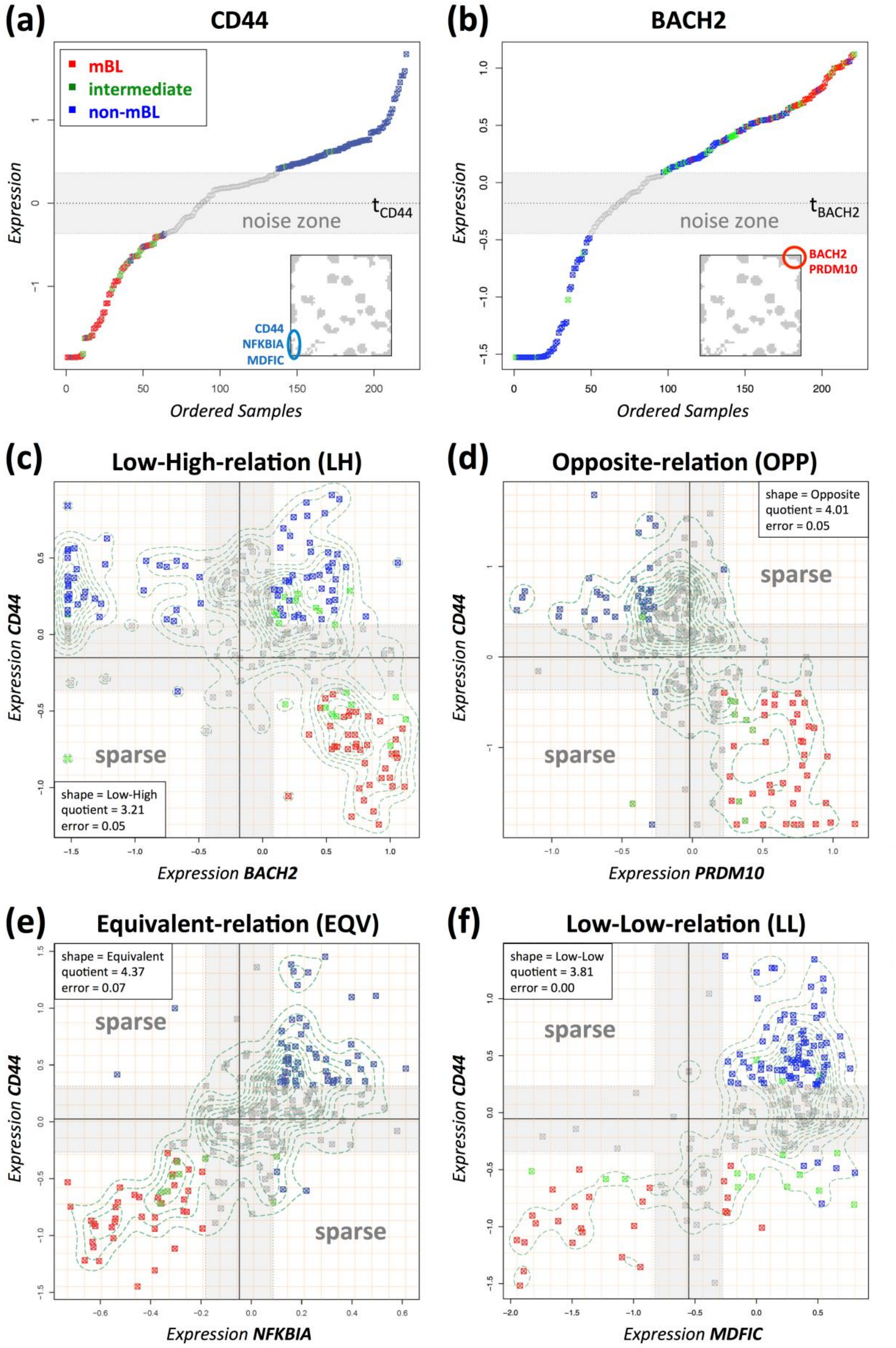

Figure 5: Boolean implication analysis for selected mBL- and non-mBL signature genes. The icons show that these genes are located within the regions of the spot-modules mBL-UP and non-mBL-UP (see small maps within panel b and a, respectively and see Figure-4b and c). Ordered expression profiles of non-mBL-UP gene CD44 (panel a) and of the mBL-UP gene BACH2 (panel b). Pairwise scatterplots of genes from different (panels c and d) and from the same signature (panels e and f) can be assigned to Boolean implications of the opposite- and of the equivalence-type, respectively. (R package KernSmooth version 2.23-7 is used for the contour lines in Figure-5 and 6 [43]).

The ranked expression values of the metagenes '1×1' and '50×50' are shown in Figure-6a and b. It is clearly seen in Figure 6a that low values of metagene '1×1' are enriched by mBL-type lymphoma (see red dots) while high values are enriched by non-mBL-type lymphoma (blue dots). On the contrary low values of the antagonistic metagene '50×50' are enriched by non-mBL lymphoma while high values are enriched by mBL lymphoma (Figure 6b). Not surprisingly these metagenes are located on the opposite corners of the SOM map and constitute an opposite type Boolean implication (see inserted maps in Figure-6a, b and c). Similarly metagenes that are in a close proximity in SOM map constitute equivalent type Boolean implications (demonstrated in Figure-6d). Importantly, we found that the behavior of the metagenenes is well preserved also within spot level analysis (Figure-6e and f).

3.4 Comparing data aggregation before and after implication analysis

Using exclusively the mean expression profiles of the metagenes or spots reduces the diversity of implications seen on the single gene level and thus it possibly represents a suboptimal way to extract metagene-metagene and spot-spot implications. We therefore implemented a decision tree algorithm which primarily uses information about implications from the next lower level, i.e. between the single genes for metagene-level implications and between the metagenes for spot-level implications (Figure-3, see next section and Materials and Methods for details). Effectively, the decision tree votes for the majority of implications observed at the respective lower level.

Panel a and b of Figure-7 and Figure-8 show the derived spot module implications of the opposite and equivalence type derived from gene level and metagene level implication analysis, respectively. Opposite type relations are mainly found between more distant spots located in opposite corners of the map whereas equivalence type relations are found between neighboring and closely located spots owing to the the SOM training algorithm, which tends to cluster similar profiles together [21], [41]. This fact is reflected in the frequency distribution plot of the number of distances bridged by the six different implications (Figure-7c and Figure-8c). It shows a clear bimodal shape separating opposite and equivalence type relations. Each of the main peaks includes implications between features from the same spot (EQU-type) or from the main antagonistic spot pairs (OPP-type) and is flanked by smaller peaks caused by implications referring to adjacent spots.

The distributions of the Pearson correlation coefficients, computed from the respective spot module profiles also show this bimodal shape (Figure-7d and Figure-8d): Here, opposite and equivalence type relations characterized by negative and positive correlation coefficients, respectively. Note that the symmetric OPP and EQV relations possess larger absolute values compared with the asymmetric LH, HL, LL and HH relations. Hence, the explicit consideration of the asymmetric relations diversifies the co-expression options and, particularly, extends the range of correlation coefficients considered to lower values.

When using metagene level Boolean implications to deduce the spot module relations, one finds a distinctly lower number of mutual connections compared with the single gene approach (compare Figure-8 with Figure-7, a and b, respectively). The correlation coefficients with regard to the metagene profiles shift to higher absolute values than observed for the gene profiles (compare Figure-7d and Figure-8d). This shift is caused by the increased contrast in the metadata correlations as shown previously [21].

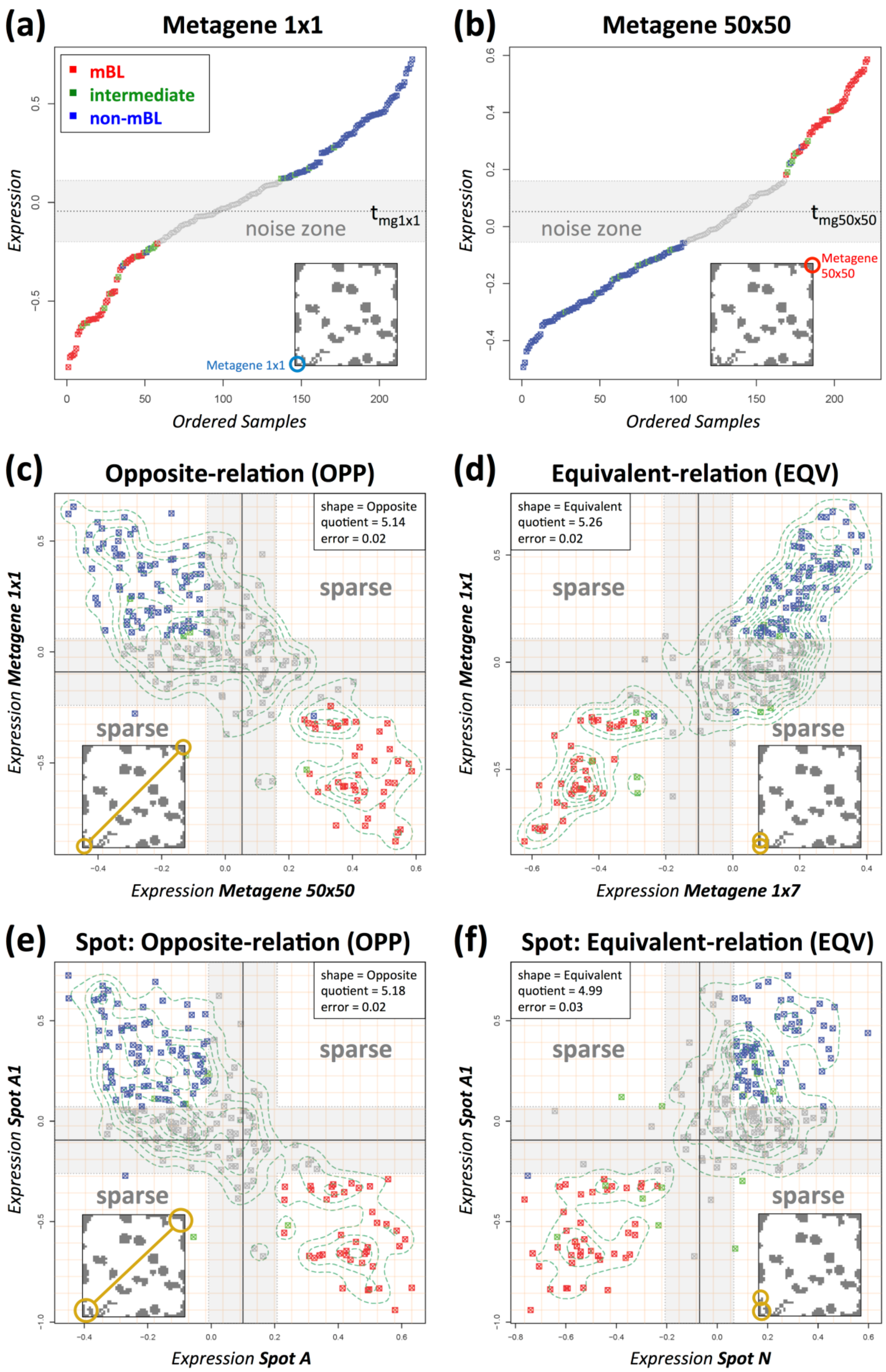

Figure 6: Boolean implication analysis on metagene and spot levels. Ordered expression profiles and scatterplots based on metagene profiles analogous to Figure-5 (panels a-d) and based on spot expression profiles (panels e and f). Positions of the metagenes and spots in the SOM map are shown in the icons in each of the panels.

Finally, Figure-9 shows the implications directly identified using pairs of spot module expression profiles. Obviously this option looses a great amount of information during the data aggregation from single gene profiles to module profiles, as only very few relations can be detected here. When comparing the total number of implications, one finds a strong decrease from 21 spot relations detected via gene level implications to 14 relations from metagene profiles to only 7 relations from spot module profiles (Figure-10a). Using low level gene data to detect implications and afterwards aggregate them to spot relations consequently conserves most information. In our particular case we only miss two (opposite-type) relations on the single gene level, which can be captured by using metagene level implications sole (Figure-10b). Please note that spot pairs are assigned to identical implications on all three levels with the exception of three cases ($OPP \leftrightarrow HL$, $EQV \leftrightarrow LL$ and $EQV \leftrightarrow HH$).

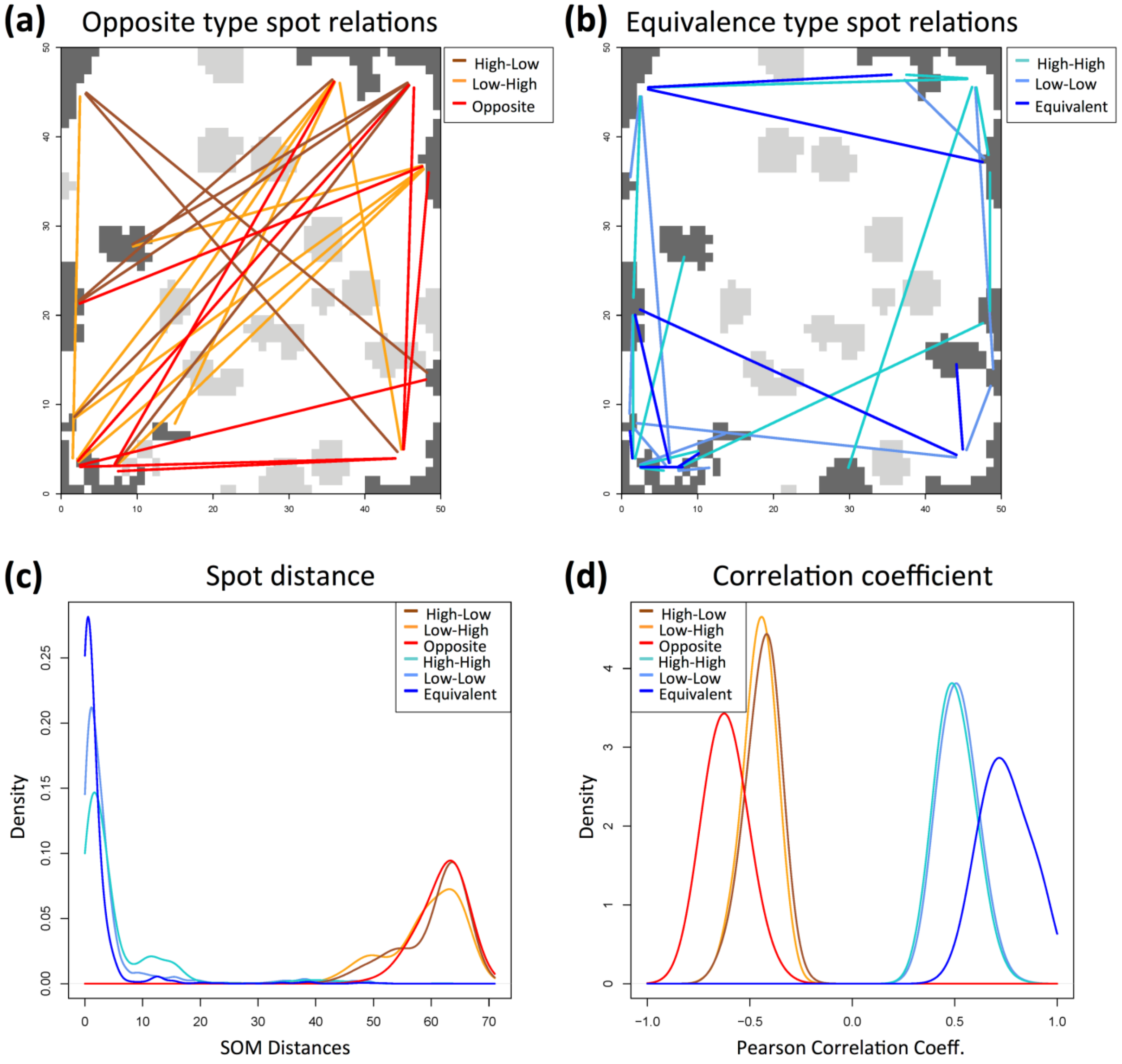

Figure 7: Spot module relations derived from single gene implications. Mapping of the opposite and equivalence type relations of the spot modules into the SOM space (panels a and b, respectively). Distribution plots of the distances bridged by the six implication types and of the according Pearson correlation coefficients of the spot module profiles (panels c and d, respectively). Note the greater spot distances and lower correlation coefficients of LH, HL, HL and HH implications relative to OPP and EQV implications.

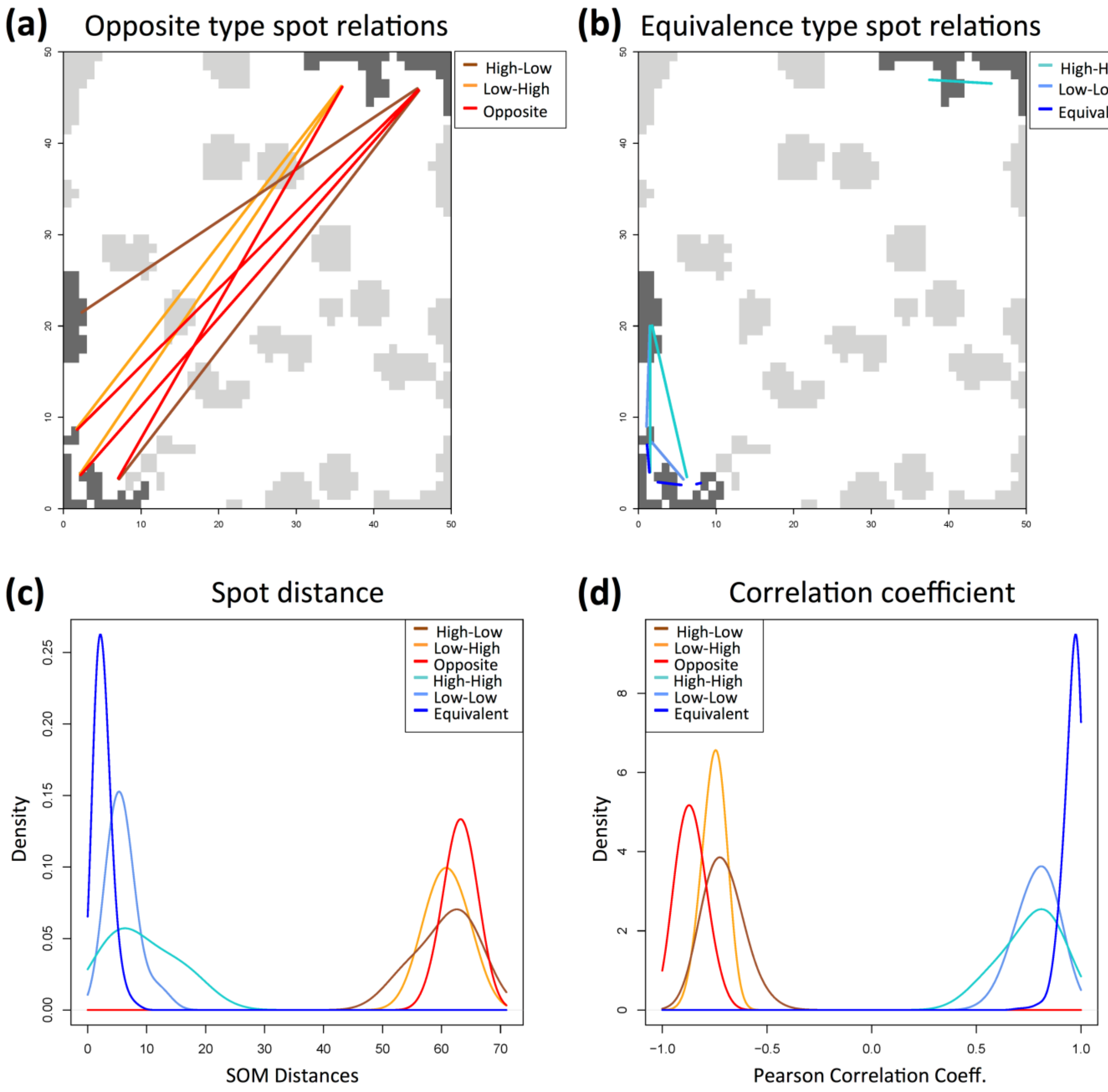

Figure 8: Spot module relations derived from metagene implications. See description of Figure-7 for further details.

3.5 Spot module relations using alternative correlation metrics

Spot module expression profiles can be also exploited in alternative approaches. Firstly we compare the relations obtained by Boolean implication analysis (Figure-11a) with those obtained by standard Pearson correlation. Therefore we compute all pairwise correlation coefficients of the spot profiles and discard those whose absolute value is below a certain threshold. The resulting spot relations are mapped into the SOM space and shown in Figure-11b. As a second option, we calculate the weighted topological overlap (wTO) between spot profile pairs, which is a correlation based measure that additionally considers indirect relations between two profiles via third ones [44]. The respective spot maps are shown in Figure-11c for varying thresholds. As a first result one can see that the overall patterns and also the total number of relations detected mainly concurs when column-wise comparing Boolean implications with correlation and wTO relations. For example in the first column, implications derived from gene level analysis, correlation with $|r| > 0.5$ and wTO with $|\omega| > 0.35$ mainly imply the same spot modules from the lymphoma SOM, whereas the wTO relations accumulate in the mBL- and non-mBL- dominated (the top right and bottom left, respectively) corners only. In general, opposite type relations and negative correlation or wTO coefficients correspond to diagonal spot pairs. Equivalence type relations and positive coefficients in contrast link closely located spots which are mostly found in the corners of the SOM.

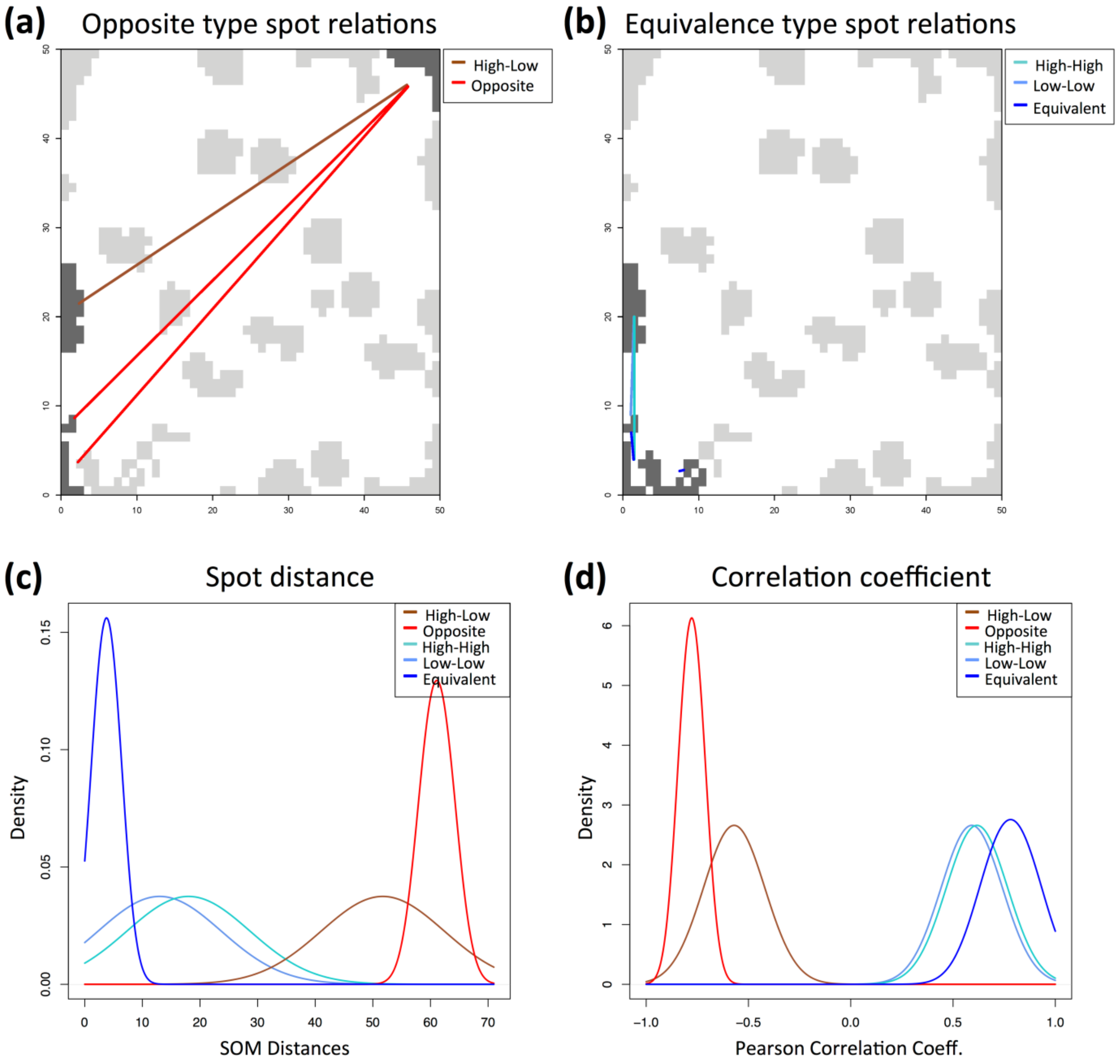

Figure 9: Spot module relations based on spot expression profiles. See description of Figure-7 for further details.

With this regard, Boolean implications are an alternative approach to the established correlation based methods. They do not only resemble the outcome of Pearson correlation and wTO analysis as shown above but also implications preserve more information, as we can distinguish six individual types instead of positive and negative correlation relations only.

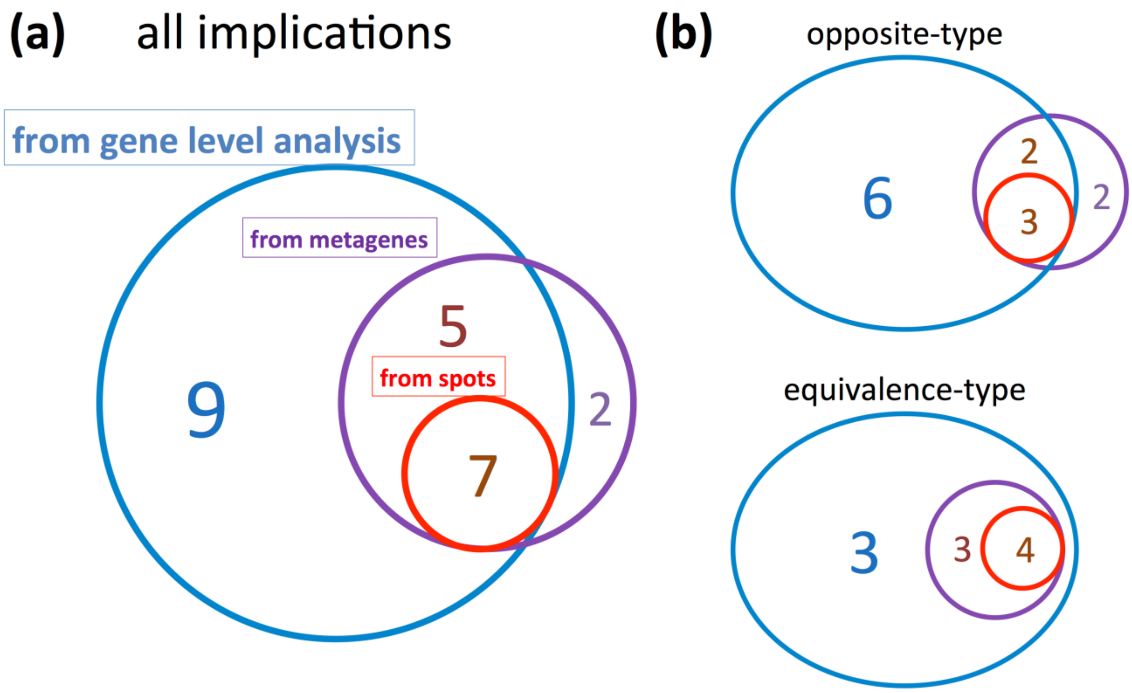

Figure 10: Venn diagrams of spot module relations. Spot module relations derived from single gene profile implication analysis (blue circles), from metagene profiles (purple circles) and spot module profiles (red circles), respectively. Total numbers for all relations are given (panel a) as well as numbers seperated for opposite and equivalence types, respectively (panel b). The number of implications markedly decreases if one uses higher levels of data integration for extracting implications.

4 Conclusions

It is possible to capture logical implications between genes by applying implication analysis over all variant gene pairs. Combination of Boolean implication analysis with SOM metadata entails the possibility to recover logical relations not only on gene level but also on metagene level (implications between sets of genes) and spot module level (implications between sets of metagenes) which eventually provides a more general and functional module oriented view over the data.

We proposed an efficient way of combining Boolean implication logic with SOM machine learning algorithm. In many cases the number of genes does not render single gene analysis feasible. In a data set of n genes there are $\binom{n}{2} = \frac{n(n-1)}{2}$ possible implications to consider. 10,000 to more than 50,000 genes (and splice variants) measured on modern microarrays or by high-throughput sequencing result in a number of gene pairs to consider in the order of 10^7 to 10^9. Since the number of genewise relations grows with n^2 reducing the number of pairs assessed by identifying “strong” candidates to focus on is worthwhile if not inevitable. By incorporating SOM pipeline we consider only *variant* genes, i.e. genes that change their expression state from considerably low values to high values or vice versa. As stated in SOM analysis, such genes accumulate in the spot-like regions mostly located in the corners and along the edges of the map. Contrary, invariant genes do not change their expression values and are consequently undetectable in terms of correlation- or implication-analyses. Furthermore metagene and spot profiles presented by SOM provides an overview of the data on different levels while increasing efficiency by reducing the dimension of the data. The SOM analysis provides a suitable framework to filter uninteresting elements and to improve Boolean implication detection.

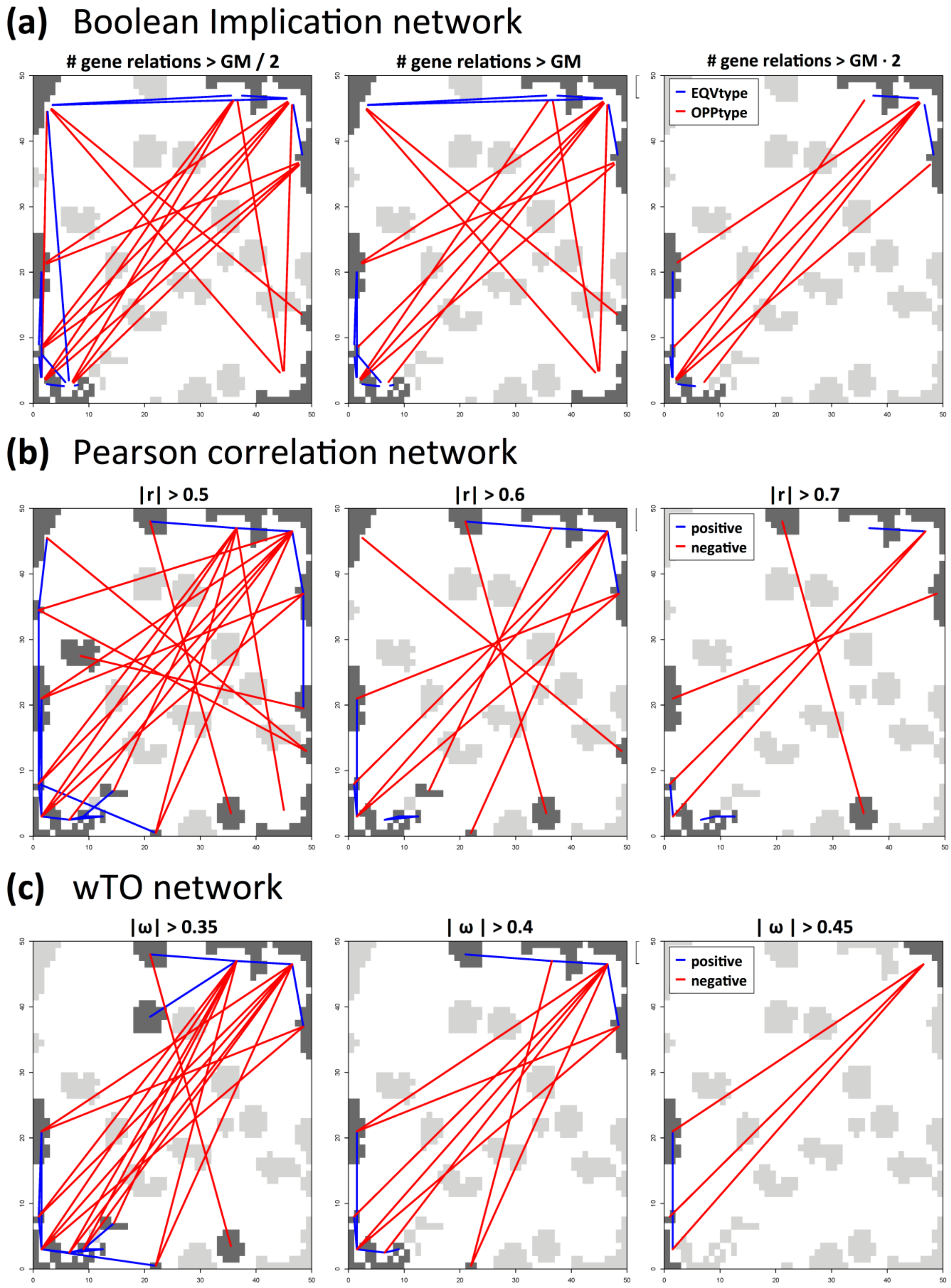

Figure 11: Spot module relations provided by different approaches. Boolean implications derived from gene-level relations with differing filter threshold; GM denotes the geometric mean number of genes in the respective spot pairs (panel a); spot profiles exceeding varying thresholds in Pearson correlation analysis (panel b) and weighted topological (wTO) analysis (panel c), respectively.

The output of Boolean implication analysis are logical relations of pairs that in turn provides a network of implications where genes/metagenes/spots constitute the nodes and edges stand for the relations between genes/metagenes/spots. Essentially resultant implication networks have a different structure than correlation networks because of their six different edge types. Besides, given a temporal ordering in the feature set of the data an implication network is a directed network i.e. an implication is not necessarily a two-way logical relation.

5 Outlook

Boolean implications can be interpreted as updating rules for the Boolean states of the network nodes and can provide a basis for modeling approaches. Particularly when the network in hand is well defined identified implications can be used to investigate the state space of the system. Also by using the biological functions associated to the genes in the network Boolean implication analysis can be used to adress biological questions regarding the functional outputs of network. In the case of Burkitt's lymphoma there are well known key players like MYC, ID3 and TCF3. Furthermore the information regarding interactions of these key players and their corresponding global biological functions are identified by numerous experiments and available. By probing Boolean implications of the genes that take part in lymphoma in a given data set, and investigating the attractors in the state space of the network in hand one not only can estimate the biological fate of the system but also identify processes to interfere with the network and manipulate it to desired fate. Thus our approach can be valuable both to researchers in the field of cancer biology, systems biology and to clinical applications. Such an analyse will be addressed in our forthcoming work.

Authors' contributions

MVÇ: Conceived and designed this study; performed boolean implication analysis, implemented the program. MVÇ, HB, HW: wrote the manuscript. HW: implemented the SOM program.

Acknowledgements

MVÇ is mainly supported by the Turkish Ministry of Education.

H.W. is supported by the Federal Ministry of Education and Research (BMBF), project grant No. FKZ 031 6166 (MMML-MYC-SYS) and by the European Regional Development Fund (ERDF) within the PhD group 'Cancer Microenvironment and Metastasis'.

The authors acknowledge support from the German Research Foundation (DFG) and Leipzig University within the program of Open Access Publishing.

List of abbreviations

SOM: self organizing maps, mBL: molecular Burkitt's lymphoma, IQR: interquartile range, OPP: opposite, EQV: equivalent, HL: High-Low, LH: Low-High, HH: High-High, LL: Low-Low, wTO: weighted topological overlap

References

[1] C. J. Wolfe, I. S. Kohane, and A. J. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks., *BMC Bioinformatics*, 6:227, 2005.

[2] D. Sahoo, D. L. Dill, A. J. Gentles, R. Tibshirani, and S. K. Plevritis. Boolean implication networks derived from large scale, whole genome microarray datasets., *Genome Biol.*, 9(10):R157, 2008.

[3] D. Sahoo. The power of boolean implication networks., *Front. Physiol.*, 3:276, 2012.

[4] M. Hummel, S. Bentink, H. Berger, W. Klapper, S. Wessendorf, T. F. E. Barth, H.-W. Bernd, S. B. Cogliatti, J. Dierlamm, A. C. Feller, M.-L. Hansmann, E. Haralambieva, L. Harder, D. Hasenclever, M. Kühn, D. Lenze, P. Lichter, J. I. Martin-Subero, P. Möller, H.-K. Müller-Hermelink, G. Ott, R. M. Parwaresch, C. Pott, A. Rosenwald, M. Rosolowski, C. Schwaenen, B. Stürzenhofecker, M. Szczepanowski, H. Trautmann, H.-H. Wacker, R. Spang, M. Loeffler, L. Trümper, H. Stein, and R. Siebert. A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling., *N. Engl. J. Med.*, 354(23):2419–2430, 2006.

[5] S. S. Dave, K. Fu, G. W. Wright, L. T. Lam, P. Kluin, E.-J. Boerma, T. C. Greiner, D. D. Weisenburger, A. Rosenwald, G. Ott, H.-K. Müller-Hermelink, R. D. Gascoyne, J. Delabie, L. M. Rimsza, R. M. Braziel, T. M. Grogan, E. Campo, E. S. Jaffe, B. J. Dave, W. Sanger, M. Bast, J. M. Vose, J. O. Armitage, J. M. Connors, E. B. Smeland, S. Kvaloy, H. Holte, R. I. Fisher, T. P. Miller, E. Montserrat, W. H. Wilson, M. Bahl, H. Zhao, L. Yang, J. Powell, R. Simon, W. C. Chan, and L. M. Staudt. Molecular Diagnosis of Burkitt's Lymphoma, *N. Engl. J. Med.*, 354(23):2431–2442, 2006.

[6] E. S. Campo Steven H. Harris, Nancy L. Pileri, Stefano Stein, Harald Jaffe, Elaine S. The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications, *Blood*, 117(19):5019–5032, 2011.

[7] G. J. P. Tyg and O. Region. Cancer in an African Community , I897-I956 * An Analysis of the Records of Mengo Hospital , Kampala , Uganda : Part 2 Cancer at Special Sites, *Br. Med. J.*, 1(5379):336–341, 1964.

[8] G. MANOLOV and Y. MANOLOVA. Marker Band in One Chromosome 14 from Burkitt Lymphomas, *Nature*, 237(5349):33–34, 1972.

[9] G. Klein. Specific chromosomal translocations and the genesis of B-cell-derived tumors in mice and men, *Cell*, 32(2):311–315, 1983.

[10] C. Bellan, L. Stefano, D. F. Giulia, E. A. Rogena, and L. Lorenzo. Burkitt lymphoma versus diffuse large B-cell lymphoma: a practical approach, *Hematol. Oncol.*, 27(4):182–185, 2009.

[11] C. V Dang, K. A. O'Donnell, K. I. Zeller, T. Nguyen, R. C. Osthus, and F. Li. The c-Myc target gene network, *Semin. Cancer Biol.*, 16(4):253–264, 2006.

[12] J. Gearhart, E. E. Pashos, and M. K. Prasad. Pluripotency Redux — Advances in Stem-Cell Research, *N. Engl. J. Med.*, 357(15):1469–1472, 2007.

[13] Z. Nie, G. Hu, G. Wei, K. Cui, A. Yamane, W. Resch, R. Wang, D. R. Green, L. Tessarollo, R. Casellas, K. Zhao, and D. Levens. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells, *Cell*, 151(1):68–79, 2012.

[14] K. I. Zeller, X. Zhao, C. W. H. Lee, K. P. Chiu, F. Yao, J. T. Yustein, H. S. Ooi, Y. L. Orlov, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, V. A. Kuznetsov, W.-K. Sung, Y. Ruan, C. V Dang, and C.-L. Wei. Global mapping of c-Myc binding sites and target gene networks in human B cells, *Proc. Natl. Acad. Sci.*, 103(47):17834–17839, 2006.

[15] T.-C. Chang, D. Yu, Y.-S. Lee, E. A. Wentzel, D. E. Arking, K. M. West, C. V Dang, A. Thomas-Tikhonenko, and J. T. Mendell. Widespread microRNA repression by Myc contributes to tumorigenesis, *Nat Genet*, 40(1):43–50, 2008.

[16] K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, C. V Dang, and J. T. Mendell. c-Myc-regulated microRNAs modulate E2F1 expression, *Nature*, 435(7043):839–843, 2005.

[17] M. Wade and G. M. Wahl. c-Myc, Genome Instability, and Tumorigenesis: The Devil Is in the Details, in *The Myc/Max/Mad Transcription Factor Network SE - 7*, vol. 302, R. N. Eisenman, Ed. Springer Berlin Heidelberg, 2006, pp. 169–203.

[18] M. Schlee, M. Hölzel, S. Bernard, R. Mailhammer, M. Schuhmacher, J. Reschke, D. Eick, D. Marinkovic, T. Wirth, A. Rosenwald, L. M. Staudt, M. Eilers, F. Baran-Marszak, R. Fagard, J. Feuillard, G. Laux, and G. W. Bornkamm. c-MYC activation impairs the NF-κB and the interferon response: Implications for the pathogenesis of Burkitt's lymphoma, *Int. J. Cancer*, 120(7):1387–1395, 2007.

[19] M. S. Staege, S. P. Lee, T. Frisan, J. Mautner, S. Scholz, A. Pajic, A. B. Rickinson, M. G. Masucci, A. Polack, and G. W. Bornkamm. MYC overexpression imposes a nonimmunogenic phenotype on Epstein–Barr virus-infected B cells, *Proc. Natl. Acad. Sci.*, 99(7):4550–4555, 2002.

[20] J. Hecht and J. Aster. Molecular biology of Burkitt's lymphoma, *J. Clin. Oncol.*, 18:3707–3721, 2000.

[21] H. Wirth, M. Loffler, M. von Bergen, and H. Binder. Expression cartography of human tissues using self organizing maps, *BMC Bioinformatics*, 12(1):306, 2011.

[22] J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong. Analysis and visualization of gene expression data using Self-Organizing Maps, *Neural Networks*, 15(8–9):953–966, 2002.

[23] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik. An improved algorithm for clustering gene expression data, *Bioinforma.*, 23(21):2859–2865, 2007.

[24] K. Torkkola, R. Mike Gardner, T. Kaysser-Kranich, and C. Ma. Self-organizing maps in mining gene expression data, *Inf. Sci. (Ny).*, 139(1–2):79–96, 2001.

[25] L. Hopp, H. Wirth, M. Fasold, and H. Binder. Portraying the expression landscapes of cancer subtypes: a glioblastoma multiforme and prostate cancer case study., *Syst. Biomed.*, 2013.

[26] H. Binder, L. Hopp, V. Cakir, M. Fasold, M. von Bergen, and H. Wirth. Molecular phenotypic portraits - Exploring the 'OMES' with individual resolution, in *Health Informatics and Bioinformatics (HIBIT), 2011 6th International Symposium*, J. Allmer, Ed. IEEE Xplore, 2012, pp. 99–107.

[27] H. Wirth, M. Çakir, L. Hopp, and H. Binder. Analysis of MicroRNA Expression Using Machine Learning, *Methods Mol. Biol.*, vol. 1107, no. miRNomics: MicroRNA Biology and Computational Analysis, pp. 257–278, 2014.

[28] V. Cakir, H. Wirth, L. Hopp, and H. Binder. miRNA expression landscapes in stem cells, tissues and cancer., *Methods Mol. Biol.*, vol. 1107, no. miRNomics: MicroRNA Biology and Computational Analysis, pp. 279–302, 2014.

[29] D. G. Covell, A. Wallqvist, A. A. Rabow, and N. Thanki. Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data1, *Mol. Cancer Ther.*, 2(3):317–332, 2003.

[30] J. Läuter, F. Horn, M. Rosołowski, and E. Glimm. High-dimensional data analysis: selection of variables, data compression and graphics--application to gene expression., *Biom. J.*, 51(2):235–251, 2009.

[31] G. Wright, B. Tan, A. Rosenwald, E. H. Hurt, A. Wiestner, and L. M. Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma., *Proc. Natl. Acad. Sci. U. S. A.*, 100(17): 9991–9996, 2003.

[32] H. Binder and S. Preibisch. "Hook"-calibration of GeneChip-microarrays: theory and algorithm., *Algorithms Mol. Biol.*, 3:12, 2008.

[33] H. Binder, K. Krohn, and S. Preibisch. "Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures., *Algorithms Mol. Biol.*, 3:11, 2008.

[34] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias., *Bioinformatics*, 19(2):185–193, 2003.

[35] D. Sahoo, D. L. Dill, R. Tibshirani, and S. K. Plevritis. Extracting binary signals from microarray time-course data., *Nucleic Acids Res.*, 35(11):3705–3712, 2007.

[36] R. R. Sokal and F. J. Rohlf. *Introduction to Biostatistics*, 2nd editio. W.H. Freeman & Company, 1987, p. 363.

[37] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic., *Proc. Natl. Acad. Sci. U. S. A.*, 100(9):5136–5141, 2003.

[38] H. Wirth, M. von Bergen, and H. Binder. Mining SOM expression portraits: feature selection and integrating concepts of molecular function., *BioData Min.*, 5(1):18, 2012.

[39] H. Wirth, M. von Bergen, J. Murugaiyan, U. Rösler, T. Stokowy, and H. Binder. MALDI-typing of infectious algae of the genus Prototheca using SOM portraits., *J. Microbiol. Methods*, 88(1):83–97, 2012.

[40] L. Steiner, L. Hopp, H. Wirth, J. Galle, H. Binder, S. Prohaska, and T. Rohlf. A global genome segmentation method for exploration of epigenetic patterns, *PLoS One*, 7(10): e46811, 2012.

[41] T. Kohonen. Self Organizing Maps, *Springer, Berlin, Heidelberg, New York*, 1995.

[42] L. Hopp, K. Lembcke, H. Binder, and H. Wirth. Portraying the Expression Landscapes of B-Cell Lymphoma - Intuitive Detection of Outlier Samples and of Molecular Subtypes., *Biology (Basel).*, 2(4):1411–1437, 2013.

[43] W. Matt. KernSmooth: Functions for kernel smoothing for Wand & Jones (1995). 2011.

[44] K. Nowick, T. Gernat, E. Almaas, and L. Stubbs. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain, *Proc. Natl. Acad. Sci. U. S. A.*, 106(52):22358–22363, 2009.