# Analysis of large-scale molecular biological data using self-organizing maps

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

**DISSERTATION**

zur Erlangung des akademischen Grades

**DOCTOR RERUM NATURALIUM**
(Dr. rer. nat.)
im Fachgebiet
Informatik

vorgelegt
von Dipl.-Inf. Henry Wirth
geboren am 05.10.1982 in Meerane

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler (Universität Leipzig, Deutschland)
2. Prof. Dr. David Kreil (University of Warwick, England)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 06.12.2012 mit dem Gesamtprädikat summa cum laude.

# Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, die mich während der Anfertigung dieser Arbeit begleitet und unterstützt haben. Dabei gilt mein erster Dank meinem Mentor Hans Binder. Er stand mir stets mit helfender Hand zur Seite, setzte Vertrauen in meine Arbeit und förderte meine persönliche und fachliche Weiterentwicklung.
Ebenfalls für ihre Unterstützung und Förderung während meiner Doktorandenzeit danken möchte ich Martin von Bergen, Markus Löffler und Peter F. Stadler.

Zudem bedanke ich mich bei meiner Familie für ihre immerwährende und liebevolle Unterstützung. Ohne euch wäre ich nie soweit gekommen. Jule und Joey, ihr zeigt mir jeden Tag den richtigen Weg und helft mir ihn zu gehen, auch wenn es mal schwierig erscheint.

Weiterhin möchte ich mich bei meinen tollen Kollegen am IZBI und am Lehrstuhl für Bioinformatik bedanken, allen voran meinen ehemaligen und aktuellen Zimmerkollegen, mit denen ich eine schöne Zeit verbracht habe. Danke euch allen für die unzähligen Gespräche, Diskussionen und kleinen Hilfen.

Danke auch an die immer fleißigen Helfer Corinna, Petra und Jens, die stets organisatorische, bürokratische oder technische Stolpersteine aus dem Weg räumten.

Zuletzt möchte ich besonders bei Lydia, Mario und Volkan bedanken, die sich die Zeit genommen haben diese Arbeit zu lesen und ihr den letzten Schliff zu geben.

Vielen Dank :)

# Abstract

Modern high-throughput technologies such as microarrays, next generation sequencing and mass spectrometry provide huge amounts of data per measurement and challenge traditional analyses. New strategies of data processing, visualization and functional analysis are inevitable. This thesis presents an approach which applies a machine learning technique known as self organizing maps (SOMs). SOMs enable the parallel sample- and feature-centered view of molecular phenotypes combined with strong visualization and second-level analysis capabilities.

We developed a comprehensive analysis and visualization pipeline based on SOMs. The unsupervised SOM mapping projects the initially high number of features, such as gene expression profiles, to meta-feature clusters of similar and hence potentially co-regulated single features. This reduction of dimension is attained by the re-weighting of primary information and does not entail a loss of primary information in contrast to simple filtering approaches. The whole set of single feature profiles remains virtually 'hidden' in the meta-features. The meta-data provided by the SOM algorithm is visualized in terms of intuitive mosaic portraits. Sample-specific and common properties shared between samples emerge as a handful of localized spots in the portraits collecting groups of co-regulated and co-expressed meta-features. This characteristic color patterns reflect the data landscape of each sample and promote immediate identification of (meta-)features of interest. It will be demonstrated that SOM portraits transform large and heterogeneous sets of molecular biological data into an atlas of sample-specific texture maps which can be directly compared in terms of similarities and dissimilarities. Importantly, SOMs preserve the information richness of the original data allowing detailed, multivariate explorative comparisons between meta-features and samples, respectively. Spot-clusters of correlated meta-features can be extracted from the SOM portraits in a subsequent step of aggregation. This spot-clustering effectively enables reduction of the dimensionality of the data to a handful of signature modules in an unsupervised fashion. The SOM method consequently enables compression of the original set of high-dimensional data in two consecutive steps: Firstly, similar profiles of single features are collected in the meta-feature clusters, which reduces the number of relevant features by about one order of magnitude in our applications. Secondly, the spot textures of the obtained SOM portraits are decomposed into a few (typically less than one dozen) spots of similar meta-features.

Furthermore we demonstrate that analysis techniques, which are normally applied at the feature-level, provide enhanced resolution if applied to the meta-features. The improved discrimination power of meta-features in downstream analyses such as hierarchical clustering, independent component analysis or pairwise correlation analysis is ascribed to essentially two facts: Firstly, the set of meta-features better represents the diversity of patterns and modes

inherent in the data and secondly, it also possesses the better signal-to-noise characteristics as a comparable collection of single features.

Additionally to the pattern-driven feature selection in the SOM portraits, we apply statistical measures to detect significantly differential features between sample classes. Implementation of scoring measurements, such as the shrinkage t-score, supplements the basal SOM algorithm. Further, two variants of functional enrichment analyses are introduced which link sample specific patterns of the meta-feature landscape with biological knowledge and support functional interpretation of the data based on the 'guilt by association' principle.

Finally, case studies selected from different 'OMIC' realms are presented in this thesis. In particular, molecular phenotype data derived from expression microarrays (mRNA, miRNA), sequencing (DNA methylation, histone modification patterns) or mass spectrometry (proteome), and also genotype data (SNP-microarrays) is analyzed. It is shown that the SOM analysis pipeline implies strong application capabilities and covers a broad range of potential purposes ranging from time series and treatment-vs.-control experiments to discrimination of samples according to genotypic, phenotypic or taxonomic classifications.

# Contents

# 1 Introduction

## 1.1 General challenges in high-throughput data analysis

In modern molecular biology, high-throughput technologies such as DNA microarrays, next generation sequencing or mass spectrometry allow researchers to assess up to hundreds of thousands of features under up to hundreds of samples or experimental conditions of interest. Not only the progressively increasing data throughput of these methods challenges analysis methods. But also the increasing availability of large data sets in public data repositories such as Gene Expression Omnibus[1] or Array Express[2] requires adequate analysis and meta-analysis strategies. This comprises optimal arrangement and visualization of the huge heaps of data preferably in combined sample- and feature-centered views to capture the global data structure while simultaneously presenting the specifics of each individual sample. Importantly, also appropriate statistics and downstream analyses have to be involved to extract characteristic features, to mine their functional context and to control the error level.

Results are frequently presented in terms of tables and visualized in terms of basic images such as heatmaps or barplots. Such presentations are very popular because they are simple to understand and because they, in most cases, provide an overview about the data which is sufficient to identify characteristic features such as clusters of genes up- or downregulated under selected conditions. On the other hand, important information which is crucial for the understanding of systems behavior might be hidden or even undetectable due to several reasons: complicated multivariate data structure, high connectivity between the features, poor quality of the data or unfavorable presentation. Hence, tasks such as data transformation from measured values into calibrated features, their appropriate evaluation and weighting according to their importance in the biological context and suited support for extraction and interpretation of sought (and unsought) information becomes an extremely puzzling task in modern biology.

A general aim is consequently the provision of comprehensive analysis tools which integrate appropriate methods, data visualization and result presentation. This thesis will present an approach to tackle these challenges utilizing a neural network algorithm called self-organizing maps (SOMs). SOMs combine data processing and dimension reduction with strong visualization capabilities. Especially for large and complex volumes of data, where conventional approaches are revealed to be insufficient, the capability of SOMs will be demonstrated.

[1] www.ncbi.nlm.nih.gov/geo
[2] www.ebi.ac.uk/arrayexpress

## 1.2 Neuronal data perception using machine learning

Despite exponential growth of computational power, information processing capabilities of the human brain are reached by no means so far. Except mathematical calculations in terms of straight analytical solutions and related applications, the brain can solve problems which pose insurmountable obstacles for any computer machine. It appears desirable to make use of the potential of neuronal data processing and decision making and to apply those 'natural' principles 'in silico', i.e. in 'artificial' computer programs. Especially, concepts of neuronal data perception and of low level processing of vast amounts of information occurs as promising attempt to analyze molecular-biological data obtained with new generation high throughput technologies.

One particular method, so-called self-organizing maps (SOM), combines several benefits important in this context namely clustering, dimension reduction, multidimensional scaling and visualization. This machine learning algorithm based on artificial neuronal networks was developed by Kohonen about thirty years ago [1]. It transforms data from the original high-dimensional 'input' space into a low- (usually two-) dimensional 'map' space. Contrary to linear scaling, the multivariate structure of the data is captured in map space because it uses a non-linear transformation. Importantly, the mapped data can be presented in terms of two-dimensional mosaic pictures providing an individual visual identity for each sample. Such 'molecular portraits' highlight relevant intrinsic substructures in the data.

It has been demonstrated that SOM can serve as a powerful tool in large-scale data analysis [2, 3] because, (i) the underlying image-based perception is very intuitive and clearly promotes the discovery of qualitative relationships between the samples in the absence of an existing hypothesis; (ii) it reduces the dimension of the original data and provides new, complex objects for next level analysis; and (iii) it preserves the information richness of the molecular states allowing the detailed, multivariate explorative comparison between samples.

## 1.3 Methodical developments and applications of SOMs in biological data analysis

First approaches applying SOMs to microarray gene expression data were published by Tamayo et al. [4] and Törönen et al. [5] in 1999, emphasizing a gene-centered perspective to cluster gene expression profiles in studies on stem cell and yeast, respectively. Golub et al. [6] published the complementary sample-centered clustering method to discriminate acute myeloid and lymphoblastic leukemia (AML vs. ALL). Covell et al. [7] used the same approach for the classification of human tissues and tumor groups. A series of subsequent microarray studies applied SOM-cluster analyses [8–13] in the fields of stem cell differentiation, cancer dysfunction (leukemia, lymphoma, adenocarcinoma, sarcoma) and toxication of human samples, mice, but also other organisms such as yeast and Caenorhabditis elegans. In the last years, applications of SOM machine learning extended to different modern fields of

bioanalytics beyond gene expression analysis such as proteomics and metabolomics/metabonomics using mass spectrometry [14–16] and NMR spectroscopy [17–19]. Further, clustering of Tyrosine phosphorylation profiles [20] and the webatlas of murine genomic imprinting [21] represent first applications of self-organizing maps in epigenetics.
Note also that SOM are frequently applied in other fields than molecular biology to mine large and complex data, for example to assess epidemiological factors of malaria endemic zones [22] or for textmining and keyword clustering [23]. Also image processing tasks can be solved using SOMs, e.g. to process spectral landscape maps [24].

Other studies address methodical issues, e.g. to further improve the machine learning algorithm in applications to special data types. For instance, customized SOM algorithms such as 'recursive SOM' (RecSOM) or 'SOM for structured data' (SOM-SD) were developed to deal with strongly structured data (see, e.g., [25] for an overview): RecSOM combines the basic SOM learning with a recursive feedback loop, allowing to learn temporal sequences of input data [26]. Another approach was realized by the SOM-SD to map directed acyclic graphs using a recursive learning mechanism [27]. This method was further combined with a hyperbolic map topology as 'SOM for sequences' (SOM-S) [28]. The so-called 'merge SOM' (MSOM) provides a more general extension of SOM-SD without a rigid grid structure suited for the processing of sequence data [29].
Other methodical modifications of the SOM-technique aim at improving data mapping and enabling more flexible learning. A dynamically growing map structure was developed to avoid the problem of fixed - and hence potentially to small - map sizes. The 'growing SOM' (GSOM) automatically adds nodes to the map to better cover dense regions of the input data space [30]. Thus, GSOM automatically adapts size and shape of the SOM. This approach was further improved by automatically adjusting the direction of growth of the SOM ('recursive mean directed growing', RMDG) [31]. Another approach to bypass rigid grid topologies is the 'neural gas' (NG) [32]. Here, the optimal topological structure is iteratively re-determined, leading to versatile node ordering. The concept of NG can be linked with other concepts of SOM-topology and learning to combine the respective advantages [29].
The original SOM method is an unsupervised learning algorithm. However, also supervised modifications were developed to train a SOM with regard to predefined classes. The 'SOM discrimination index' (SOMDI) provides a simple approach to integrate class information into the training data, which has been applied to classify NMR spectra of metabolites [18, 19]. More elaborated methods have been published in the field of mass spectrometry: The 'fuzzy-labeled' SOM (FLSOM) offers a robust semi-supervised classifier, especially suited for uncertain data. Case studies deal with MALDI MS-spectra of bacteria and breast cancer samples [33]. Finally the so-called 'Local Linear Maps' (LLM) were developed to predict intensity amplitudes of peptide peaks in MALDI spectra [14, 34].

A special implementation of the SOM method aims at visualizing the 'landscapes' of large scale molecular data such as ten thousands of gene expression levels in a comprehensive and intuitive fashion. Such data can be presented with the focus to compare the samples in terms of similarity measures or, alternatively, with the focus to extract single characteristic features which discriminate different samples. These alternative sample- and gene-centered views usually require different methods of analysis and visualization, e.g. principal component analysis (PCA [35]) for the former one and significance analysis of microarrays (SAM [36]) for the latter one. The SOM method allows combination of both the sample- and gene-centered perspectives [2, 37, 38]. This specific configuration of the SOM uses the so called component planes of the SOM to decode the expression pattern of the genes within a two-dimensional mosaic pattern. It allows the easy sample-to-sample comparison by direct visual inspection *and* the identification of single features in terms of groups of co-regulated genes. Such SOM portraits have been applied in studies on cell differentiation and development [39–43], organogenesis [44] and tumor progression and classification [3, 45, 46].
Several SOM-based analysis packages were developed as stand-alone or web-based tools [21, 23, 38, 47, 48]. Especially the tool packages 'Gene Expression Dynamics Inspector (GEDI)' [38], 'Grid Analysis of Time series Expression (GATE)' [47] and '*omeSOM' [48] provide extensive and, for many applications, sufficient functionalities. However, these tools are rather inflexible and restricted concerning the challenges of individualized analyses. Especially options in low-level preprocessing and the presentation of specific high-level results are deficient for customized applications in molecular biology.

## 1.4 Objectives and outline

SOM analysis is particularly suited for analysis of large-scale data due to the potent combination of clustering, dimension reduction, multidimensional scaling and visualization capabilities. Further methodical developments continuously improve the method and offer a variety of sophisticated applications. Presumably due to at least two reasons, SOM analyses are still relatively infrequently applied compared to alternative methods such as hierarchical clustering heatmaps or principal component analysis: Firstly, our experience shows that SOM seems quite unaccustomed for many researchers with background in statistics and biology due to the machine learning step and the partly unusual structure of transformed data. Therefore we believe that an improved understanding of the concept of SOM learning and mapping might increase the acceptance and promote application of the method. Secondly, the fundamental SOM algorithm needs to be supplemented with data specific statistical measures, tools for feature extraction and for visualization. Despite the intensive work in developing and applying SOM algorithms, data mining modules for extracting specific information about the systems studied are often missed.
This thesis aims at bridging the gap between the potential of the SOM method and the problems associated with the exploration of the transformed data, and at demonstrating its

strength in selected case studies taken from high-throughput experiments in molecular biology. Under methodical aspects, the advantages of SOM will be evaluated with regard to high-dimensional data analysis and compared with alternative methods, special visualization techniques will be presented to illustrate the meta-data space provided by the SOM and statistical methods for feature extraction will be adapted to the SOM meta-gene structure. Finally, tools for functional analyses, for example enrichment of sets of genes with known biological implication, were applied to SOM clustered data. This thesis hence pursues an interdisciplinary scope: It addresses bioinformaticians and statisticians (methodological aspects and their implementation) as well as biologists (applications).

The contents are organized as follows (see also the workflow shown in Figure 1-1): Chapter 2 addresses primary data analysis: It provides a brief description of the SOM algorithm used and describes associated tasks such as data preprocessing, visualization and extraction of functional modules inherent in the data. Chapters 3, 4 and 5 deal with methodical aspects of 'secondary analysis' of the transformed data delivered by the SOM algorithm. We focus on issues related to data filtering, statistical scores for feature selection and functional enrichment analysis. Chapter 6 presents selected case studies of our SOM analysis which demonstrate particular applications in different OMICs-data. Most of the examples are published, in press or under review. Manuscripts with relevance for this thesis can be downloaded from the author's website [3] to provide details not given in the main text. Additionally, a software package was developed in R [49] including all analysis functionalities described in this thesis. It is available as R-package 'oposSOM' on CRAN repository[4].

[3] http://www.izbi.uni-leipzig.de/izbi/mitarbeiter/wirth.php
[4] http://cran.r-project.org/web/packages/oposSOM

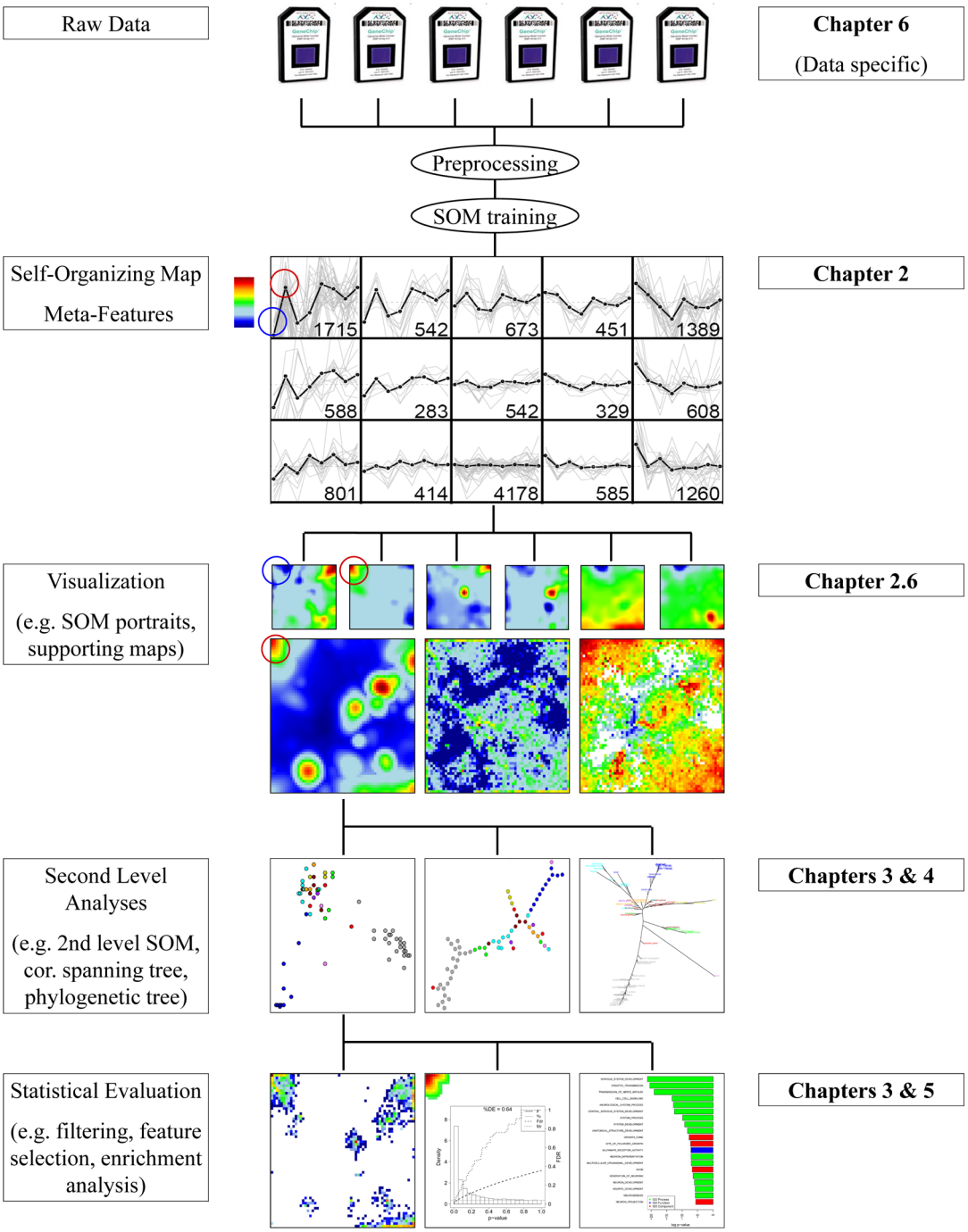


**Figure 1-1: SOM-analysis workflow: Raw experimental data (microarray expression data here) is adequately preprocessed and feed into the SOM-algorithm. It provides meta-expression profiles (thick curves) representing the manifold of expression states observed in the series of samples studied in lower dimensions. Results of the training process (i.e. meta-feature profiles, structural information) are then utilized for direct visualization (SOM portraits, see first row in visualization part) or to create supporting maps (second row) characterizing different aspects of the SOM. Secondary analyses (e.g. component analysis, correlation analysis or clustering) can be applied based on meta-features instead of original data, implying analysis on a higher level of information aggregation. Meta-features can further be used for statistical and functional analysis (e.g. filtering, feature selection, enrichment analysis), complementing the basal SOM algorithm.**
**Detailed aspects of the respective analysis steps are given in the different chapters of this thesis as indicated in the figure.**

# 2 Self-organizing maps

## 2.1 Neural network models

The human brain is very efficient in processing complex information. Consequently there are numerous attempts to understand and to apply principles of natural learning and knowledge processing to enduring tasks in computer science. First studies on artificial intelligence already started about 100 years ago. In 1906 the Nobel Prize was awarded to Camillo Golgi and Ramón y Cajal "in recognition of their work on the structure of the nervous system"[5]. Since then, methods based on artificial neural networks became an important part in the field of machine learning and computer science in general.

An early approach, the so-called McCulloch-Pitts network [50], is capable to learn a requested output for any binary input pattern of length n, for example the Boolean functions 'AND', 'OR' and 'NOT'. In general, each logical function $F:\{0,1\}^n \to \{0,1\}$ can be realized by these networks [51]. The *Perceptron*, a generalized version of this model, applies numerical weights to the connecting edges between neurons, allowing learning by adaption of the weight values [52, 53]. The weights, representing association strengths between the neurons, are assumed as essential ingredient for modeling natural learning by the psychologist Donald Hebb in 1949:

*"Any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other"* [54].

In other words, the connection between simultaneously activated neurons is further strengthened by increased synaptic interaction. Hebb is considered as the discoverer of synaptic plasticity, the basis of learning and memory in nervous systems. The derived 'Hebb's learning rule' for artificial neural network learning consequently describes the adaption of weights between two nodes according to concerted activation:

$$\delta w_{ij} = \eta \cdot a_i \cdot a_j \tag{1}$$

Accordingly, the weight of the edge between nodes i and j, $w_{ij}$, is increased by the increment $\delta w_{ij}$ if both nodes are simultaneously active (i.e. $a_i>0$ and $a_j>0$). The amount of adjustment is controlled by the learning rate $\eta$. This update rule and its adaptations, combined with the Perceptron structure, provide the basis of most machine learning algorithms for artificial neural networks.

Also Kohonen's SOM structure and learning mechanism can be described in terms of a defined network of nodes, interconnected by weighted edges, which in turn are updated according to Hebb's leaning rule [55]. The SOM model will be described more in detail later in this chapter.

[5] http://www.nobelprize.org/nobel_prizes/medicine/laureates/1906/

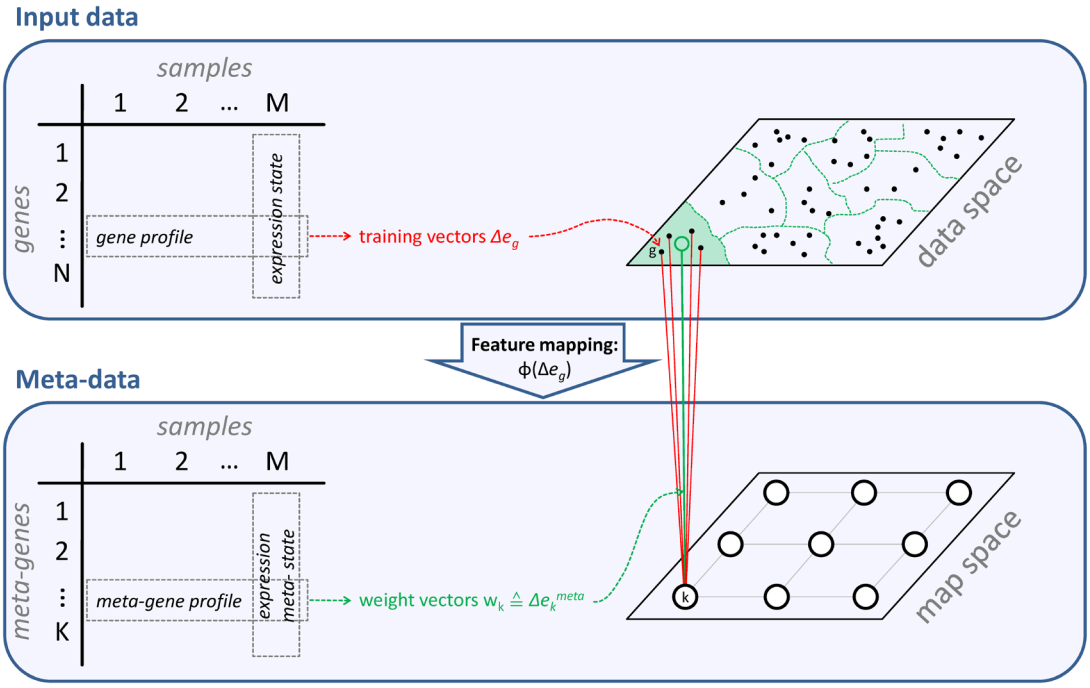


**Figure 2-1: Relation between input data and meta-data of SOM analysis of microarray expression: The input data matrix consists of N rows (representing the 'single' gene expression profiles) and M columns (representing the expression states of the different tissues). Each gene expression profile defines a vector of dimensionality M which is illustrated as one point in the M-dimensional data space. The M-dimensional map space contains K<<N meta-gene expression profiles. SOM machine learning fits the map space to data space such that the meta-gene profiles resemble the single gene profiles. Due to their smaller number, each meta-gene serves as representative of a cluster of single genes as illustrated by the fragmentation of data space. The mapping can be described by the nonlinear feature mapping function Φ.**

## 2.2 Mapping of high throughput data

High throughput screening methods in modern molecular biology such as microarrays, next-generation sequencing, and also mass spectrometry provide a vast amount of data points per measurement. Experimental series on hundreds of samples thus accumulate extensive, large-scale data sets of high-dimensions. In this thesis, SOM machine learning is applied to case studies involving several selected data types to illustrate the benefits and drawbacks of the approach. Table 1 summarizes the different data types used for SOM analysis.

The input data for SOM analysis can be described as data matrices of dimension N x M (for illustration see Figure 2-1, upper panel) where N is the number of features measured per sample and M is the number of samples referring, e.g., to different treatments, time points or individuals. As a convention, each row of the matrices will be termed *profile* of the respective feature (e.g. gene expression profile along the conditions measured). The columns on the other hand will be termed *states* referring to each of the conditions studied (e.g. the expression state of a selected microarray sample).

**Table 1: Different biological data and selected characteristics addressed in this thesis.**

| *'OME' realm* | **Transcriptome (mRNA)** | **Transcriptome (miRNA)** | **Genome** | **Epigenome** | **Methylome** | **Proteome** |
|---|---|---|---|---|---|---|
| ***Chapter; references*** | Chapters 2 - 5 and 6.1.1 - 6.1.3 ; [WIRTH1], [WIRTH3], [HOPP1] | [CAKIR1] | Chapter 6.2 ; [BINDER1] | [STEINER1] | Chapter 6.3 | Chapter 6.4 ; [WIRTH2] |
| ***Technology*** | Expression microarrays, Affymetrix | miRNA expression microarrays, LCSciences | SNP microarrays, Illumina | ChIP-Seq, Illumina | MeDIP sequencing, SOLiD | Mass spectrometry, Bruker |
| ***Features*** | Expression levels of genes | Expression levels of miRNAs | Genotypes of SNPs | Chromatin modification state of the genomic loci | Methylation state of the genomic loci | Peak positions (m/z coordinate) of protein fragments |
| ***Raw data*** | Probe intensities | Probe intensities | Probe intensities | Read counts | Read counts | MS-spectra |
| ***Calibration*** | Hook calibration (also other methods) | Standard software (Array-Pro, Media Cybernetics) | Standard genotyping software (BeadStudio, Illumina) | Library mapping to reference genome | Library mapping to reference genome | Spectral corrections (e.g. baseline subtraction) |
| ***Calibrated features*** | Log differential expression | Log differential expression | Relative allele signals (*RAS*) | Chromatin modification indicator of genomic fragments (>200bp) | Methylation state of binned regions (500bp) | Digitized spectra of peak-regions |
| ***Normalization*** | Quantile normalization | Quantile normalization | Skipped; not necessary | Skipped; not necessary | Quantile normalization | Quantile normalization |
| ***Input data for SOM training*** | Differential expression relative to average over all samples ($\Delta e$) | Differential expression relative to average over all samples ($\Delta e$) | Ternary allele code (*TAC*): major (0), heterozygous (1), minor (2) allele | Binary epigenetic profiles (*EP*): modification absent (0) or present (1) | Reads per million (*RPM*) | Intensity lists of peaks ($I$) |
| ***N( number of features)*** | 20,000 – 50,000 | 200 - 600 | 100,000 – 1,000,000 | 500,000 – 1,000,000 | 10,000 – 100,000 | 1,000 – 2,000 |
| ***Interpretation of the weights*** | Meta-gene expression, $\Delta e^{meta}$ | Meta-miRNA expression, $\Delta e^{meta}$ | Meta-SNP, $TAC^{meta}$ | Meta-EP, $EP^{meta}$ | Meta-reads, $RPM^{meta}$ | Meta-peak amplitudes, $I^{meta}$ |
| ***Typical SOM size; compression (N/K ratio)*** | 50 x 50; N/K ≈10 | 30 x 30; N/K ≈0.5 | 80 x 80; N/K≈80 | 40 x 40; N/K≈500 | 20 x 20; N/K≈100 | 20 x 20; N/K≈2 |

In general, the number of features can range from several thousands to millions, depending on the screening technique. Typically, this number largely exceeds the number of conditions studied, i.e. N>>M. SOM machine learning aims at reducing the number of relevant features by grouping the input data into clusters of appropriate size, and thus to transform the matrix of input data into a matrix of meta data with a reduced number of meta profiles K<<N (Figure 2-1, bottom panel).

Throughout this thesis a microarray gene expression study of a series of human tissues has been chosen to serve as example to describe and to illustrate the SOM method, details of the preprocessing of the input data and different options of downstream analysis of the mapped data. The series of 67 different human tissues[6] is well suited as an illustrative example because the number of different states provides a sufficiently large and diverse data set possessing a relatively complex internal covariance structure [WIRTH1]. Moreover, the samples are well classified into distinct tissues and tissue categories allowing the clear assignment of expression pattern and validation of analysis results, for example in terms of functional enrichment or of similarity relations between the samples.

## 2.3 Preprocessing of microarray data

Preprocessing transforms raw data into input data for SOM training. It aims at removing biases of the detection technology and batch effects due to sample preparation. Preprocessing basically splits into two steps, calibration and normalization. The calibration step rescales the data from detection units (probe intensities in the special case of microarray measurements) into appropriate 'molecular' units which are directly related to the property of interest, e.g. the mRNA-transcript concentration or expression degree, in this application. The normalization step ensures mutual comparability of the series of samples and relates the calibrated data to an appropriate reference level. In general, the preprocessing of different data is specific for each technology and makes use of elaborated methods (see Table 1).
Exemplarily, a microarray data set is considered consisting of the expression levels of N genes in M different samples, each measured in $R_m$ (m=1...M) replicates. The number of genes N is typically in the ten thousands, the number M of experimental conditions is typically in the tens to a few hundreds, and the number of replicates between one and ten. Affymetrix GeneChip 3'-expression microarrays provide typically eleven raw probe intensities per gene constituting one probe set. Raw probe intensity values of each of the M x $R_m$ chips studied are calibrated and summarized into one expression value E per probe set using the hook method [56, 57]. The expression values of all arrays are subsequently quantile-normalized [58] (see Figure 2-2a for illustration).

[6] Gene Expression Omnibus, accession no. GSE7307 :
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7307

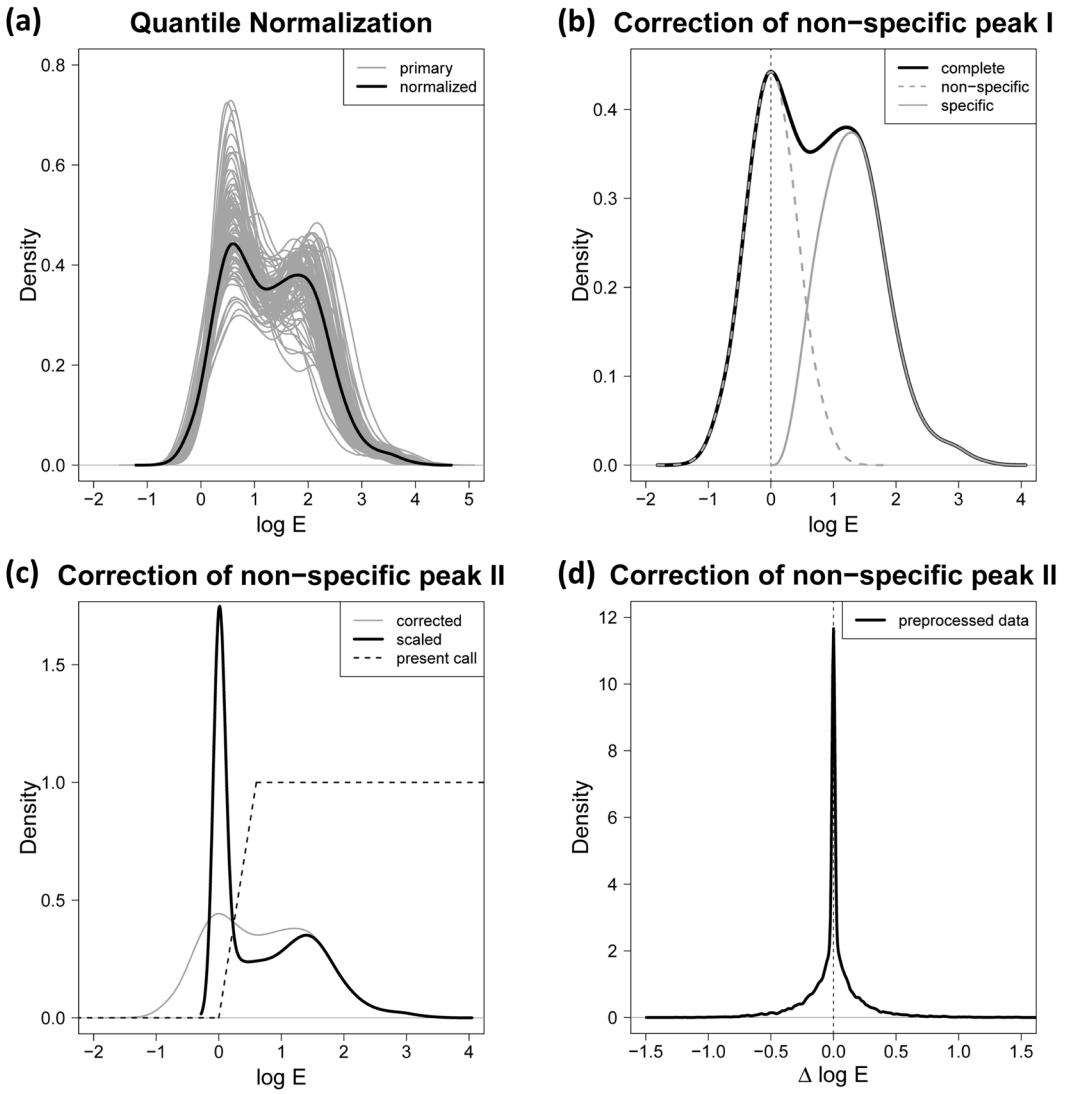


**Figure 2-2: Normalization and adjustment of microarray expression values: The different distributions of hook-calibrated expression values of the samples studied merge into one representative mean distribution after quantile normalization (panel a). Its double peaked shape is decomposed into two single peaked distributions due to non-specific and specific hybridizations at small and larger expression values, respectively (b). The fraction of the specific signal contributing to the total signal density (dashed curve) is used as weighting coefficient of the expression values, e= pc(e')*e', which reshapes the total signal density (c). Finally, the expression values are normalized with respect to the logarithmic mean expression of each gene (d). The large central peak refers to invariant genes under all conditions studied.**

The obtained distribution of expression values shows typically a bimodal shape (Figure 2-2b): It's left peak at smaller expression values and its right peak values were attributed to non-specific and specific hybridization, respectively [BINDER 3]. The peak due to non-specific hybridization is non-informative with respect to the target genes which are therefore called 'absent' because their expression is smaller than the detection threshold of the method. The non-specific peak consequently characterizes the 'chemical' background of the measurement.

The distribution of expression data of each experimental series is then processed as follows: Firstly, the origin of the log-expression axis (log E=0) was positioned to agree with the peak position of the non-specific peak of the distribution. Secondly, both peaks are decomposed as described previously [BINDER 3] assuming mirror symmetry of the left and right flanks of the non-specific peak (Figure 2-2b). Thirdly, we make use of the decomposed distributions to estimate the probability that the specific expression of a selected gene is detected. This 'present-call'-parameter is set to pc=0 and pc=1 for genes with expression values outside the region of overlap of both peaks (see Figure 2-2c). In the range of overlap, the present call is calculated as the fraction of the local density of the specific signal contributing to the total signal distribution. The resulting value of pc roughly linearly scales between zero and one with increasing expression in this range (Figure 2-2c). Fourth, the log-expression of each gene is scaled with its present call, i.e., e= pc(e')*e' where lower case e' define the logarithmic expression values, e'= log E. The used transformation thus considerably narrows the non-specific peak at position e'=0 of the expression axis while leaving the specific signal virtually unaffected. As a consequence, the variability of the signals of absent called and thus of non-informative probes is markedly reduced (Figure 2-2c). This transformation enables noise-reduced conservation of the full set of available genes in the data set used for SOM analysis in contrast to data filtering which removes presumably uninformative probes from the data set prior to downstream analysis.
Expression values of replicates of the same tissue were log-averaged and finally, the log-expression values of each gene were transformed into differential expression values relative to the average expression of each particular gene in the experimental series of tissues considered (Figure 2-2d),

$$\Delta e = e - < e >_{all\_tissues} \qquad (2)$$

Eq. (2) thus defines differential expression in units of the logarithmic fold change, logFC ≡ Δe.

## 2.4 The Kohonen model

The Kohonen model is inspired by our assumptions about the perception of visual information in the brain. Accordingly, optical input stimuli are projected onto the neuronal net in the cortical area. Then, the connections between the neurons adapt to the visual pattern in a learning process [59]. This causes a self-organization of the neuronal network such that it better matches the activation pattern. The self-organizing maps, developed by Teuvo Kohonen in 1982, mimics this input-driven self-organization [1]. This 'standard' SOM consists of a two-dimensional grid of K nodes, each of which is characterized by a representative weight vector of length M, $\vec{w} \equiv \left( w_1, \ldots w_m \ldots, w_M \right)$. In microarray expression analysis the weight vectors have the meaning of *expression profiles of meta-genes*. In general, the meaning of the weight

vectors depends on the particular SOM-applications (see Table 1). The K meta-gene profiles constitute the meta-data matrix of size K x M as illustrated in bottom panel of Figure 2-1. The rows correspond to the meta-gene expression profiles along the M samples studied, and the columns represent their *expression meta-states*.

The relation between the map space and the data space is illustrated in Figure 2-1: The single and meta-gene profile vectors are shown as points in the M-dimensional data and map space, respectively. Each point in the data space is assigned to the closest meta-gene profile using the minimal Euclidean distance as criterion (see green highlighted subspace in Figure 2-1, top right part). Each meta-gene k serves as condensation nucleus for a cluster of $n_k$ 'real' genes with similar expression profiles. Each point in data space $\Delta e_g$ is mapped to the meta-gene of closest distance $\Delta e_k^{meta}$, (see Figure 2-1).

An optimal set of meta-gene profiles captures the range of all individual expression pattern observed in the data space. The task to find this set is accomplished by the SOM machine learning algorithm. It iteratively adjusts the meta-gene profiles to the data space such that they maximally resemble the profiles of the single genes. In general, SOM-training encompasses three basal steps, initialization, training and mapping, shortly described in the following subsections.

### 2.4.1 Initialization

The choice of an appropriate initialization method will affect the quality of the subsequent training process in terms of runtime and data space coverage. Several approaches were introduced to initialize the meta-gene profiles (i.e. weight vectors) of the SOM. A simple approach assigns random values to the meta-gene profiles [60]. This random initialization is suboptimal, because the lack of determinism with respect to the obtained map space after training potentially leads to differing and/or permutated maps [61]. Moreover, randomly initialized maps are prone to *topological defects* representing metastable states which are difficult to overcome and which hamper the optimal coverage of the data space (see Figure 2-5c below). Random initialization will be applied in this context to illustrate the training process (see next subsection and Figure 2-4 and Figure 2-5).

Another method, linear initialization [60, 62], is more suited for our purposes: Here, the initial meta-gene profiles are determined along the linear subspace spanned by the two eigenvectors with largest eigenvalues of the input data. This approach is similar to *principal component analysis (PCA)*, covering the major variability inherent in the data. This initialization technique provides reproducible map topologies for similarly configured training runs and essentially overcomes metastability problems and topological defects. Linear initialization is therefore well suited to train large scale experimental data. It is used throughout this thesis if not stated otherwise.

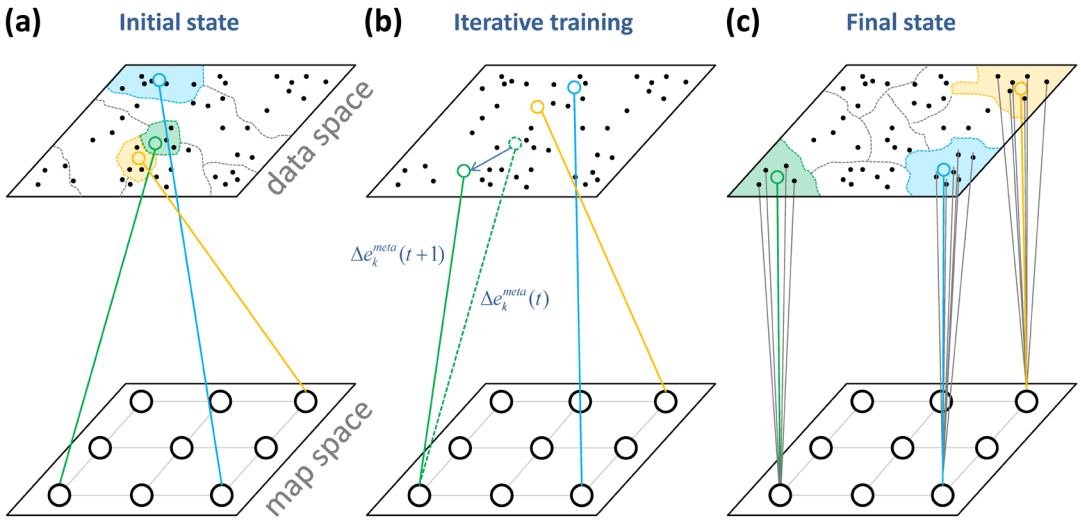


**Figure 2-3: Schematic presentation of SOM machine learning: The arrangement of SOM nodes in a regular lattice illustrates the mosaic used for visualization. Typically each node (i.e. meta-gene) is associated with a cluster of single genes of similar profiles. After initialization, the meta-gene profiles point into data space in a suboptimal fashion (panel a). During training, the meta-gene profiles are adjusted to more closely fit the single gene profiles. Training effectively minimizes the distances between the meta-genes and single genes by iterative adjustment of the meta-gene profiles and reassignment of the genes to the meta-genes in each training step as illustrated by the arrows (panel b). After training, the meta-gene profiles optimally cover the data space. The map becomes 'self-organized' meaning that adjacent meta-genes in map space are more similar than distant meta-genes (panel c).**

### 2.4.2 Training

SOM training iteratively fits the meta-genes to data space. Figure 2-3 illustrates this process: after initialization, the meta-gene profiles point into data space in a suboptimal fashion (Figure 2-3a). During training the meta-gene profiles are iteratively optimized to more closely fit the single gene profiles (Figure 2-3b).

In each step, one gene profile $\Delta e_g$ is selected as training vector. Then, the meta-gene profile of closest similarity is selected using the Euclidean distance. This 'winner' meta-gene meets the condition:

$$BMU(\Delta e_g) = \arg\min_{k=1..K} \left\| \Delta e_g - \Delta e_k^{meta} \right\| \tag{3}$$

It is also called best matching unit (BMU) with the profile $\Delta e_{BMU}^{meta}$. The meta-gene profiles are then adjusted using to the update rule,

$$\Delta e_k^{meta}(t+1) = \Delta e_k^{meta}(t) + \eta \cdot h(BMU, k) \cdot (\Delta e_g - \Delta e_k^{meta}) \tag{4}$$

which is an adaptation of Hebb's learning rule (eq. (1)). Accordingly, for any node k in the SOM and given training vector $\Delta e_g$, the adjustment of the meta-gene profile $\Delta e_k^{meta}$ consists of three terms:

- The leaning rate $\eta$ scales the incremental changes of the meta-gene vector. It decreases with progressive iteration to settle down the adjustment.
- The neighborhood function *h(BMU,k)* controls the distance-dependence in the SOM grid with respect to the BMU (see below).
- The difference term $(\Delta e_g - \Delta e_k^{meta})$ ensures that the meta-gene profiles (and most of all the BMU) are adjusted to better resemble the profile of the training gene.

The amount of adaption is scaled by the *neighborhood function* with respect to the BMU, $h(BMU,k) \in [0,1]$, for each meta-gene k of the SOM. Accordingly, the BMU serves as the central node for adaption, whereas the remaining meta-genes are decreasingly adjusted with increasing distance to the BMU in the node grid of the map. Two options are taken into account: the so-called *bubble* and the *Gaussian neighborhood*. The binary bubble neighborhood function equally affects all nodes within a given radius around the BMU [60]. In contrast, the Gaussian neighborhood continuously decays with increasing distance with respect to the BMU according to a Gaussian bell function. Therefore, it effectively applies to all nodes of the SOM [63].
Due to the joint adjustment of the BMU and its neighbors the algorithm ensures competition between the nodes to be selected as BMU in subsequent steps. It also ensures similarity of adjacent meta-gene profiles and thus self-organization of the whole map.
The adaption to the gene profile selected is iteratively repeated. Convergence of meta-gene profiles with progressive iteration is achieved by their improved fit to the data space. In addition, both learning rate $\eta$ and the range of the neighborhood are progressively decreased to avoid oscillations or instabilities (see [60] for detailed survey). One cycle of iteration steps, which encompasses each of the N genes, is called *epoch*. After a defined number of epochs the training process ends. The final SOM with trained meta-gene profiles is assumed to cover the data space in an organized and close fashion (Figure 2-3c).

Figure 2-4 illustrates the progression of a typical SOM-training in two-dimensional data and map space: The blue dots are synthetic input data generated such that they can be divided into six distinct clusters (200 profiles of the type (x1, x2)). The open circles represent the 100 meta-data of a 10x10 SOM. Adjacent nodes in the rectangular SOM-grid are connected by grey lines. After random initialization of the meta-data the distribution of blue and grey dots in the plot and thus the input- and meta-profiles largely disagree. In the course of training the positions of the meta-data progressively adjust to the input data with increasing number of training-epochs. Moreover, also the network of meta-profiles defined by the nearest neighbors of each SOM-node progressively disentangles and tends to adopt an ordered topology where similar meta-profiles are neighbors (and thus connected by lines). Finally, after 3000 epochs

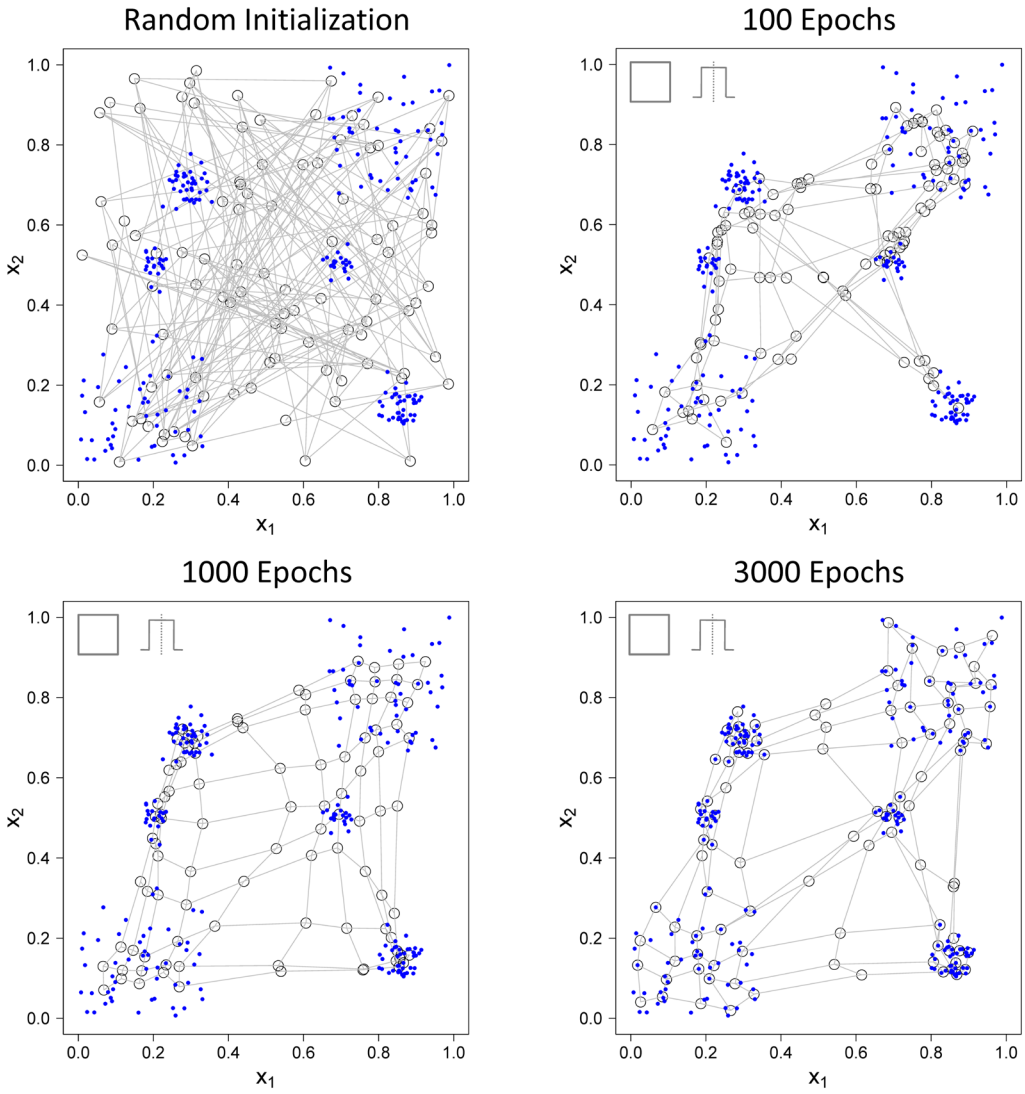


**Figure 2-4: Adjustment of meta-data (open circles) to input data (blue dots) during SOM training. The synthetic input data arrange into six clusters. After random initialization the meta-data progressively adjust to this cluster structure with increasing number of iteration-epochs. Moreover, the network defined by the four nearest neighbors of each meta-data 'disentangle' with progressive training and finally adopts an ordered topology (see grey lines). The icons in the top left corner of the plots indicate the rectangular grid topology and short-range bubble neighborhood.**

the meta-data arranged into a grid-like topology which well matches the cluster structure of the input data.

Figure 2-5 shows the progress of training after 1000 and 3000 epochs to illustrate the influence of the grid topology and of the neighborhood function. The SOMs share the same size, learning rate and neighborhood radius parameters as the example in Figure 2-4. However, Gaussian (instead of bubble) neighborhood and hexagonal (instead of rectangular) grid topology are applied as indicated by the icons in the top left corner of each plot. The results can be summarized as followed (compare Figure 2-4 and Figure 2-5):

- The longer-range Gaussian neighborhood ensures the 'soft' and more ordered adjustment of meta- to the input data than the short-range bubble neighborhood (compare Figure 2-5a/b and Figure 2-4 at 1000 epochs). Moreover, the Gaussian neighborhood is advantageous because it accelerates convergence of the training algorithm [55] and it is less prone to overfitting.

- Rectangular and hexagonal grid topologies provide almost similar results (see Figure 2-5a and b). The hexagonal grid topology is reported to produce more homogeneous meta-data [10], whereas rectangular grids require slightly less computing and are simpler to visualize. Both topologies are therefore regarded as equivalent options.

- The SOM configuration shown in Figure 2-5c combines random initialization, hexagonal grid topology and bubble neighborhood. This setting is prone to topological defects as indicated by the orange lines in Figure 2-5c [55, 63] reflecting metastability problems which are found in more than 50% of repeated independent training runs.

In our case studies below, rectangular grid topology and a Gaussian neighborhood function were applied as standard. This configuration promotes fast and stable training, and an acceptable trade-off between non-linear but still not overfitted representation of the input data.

### 2.4.3 Final mapping

After training, each of the single gene profiles is associated to the meta-gene profile of minimal Euclidean distance (BMU) giving rise to the segmentation of data space into clusters of genes mapped to each meta-gene (see Figure 2-3c for illustration). These clusters collect genes with highly similar profiles. Adjacent clusters contain genes with more similar profiles than distant ones.

The mapping of each gene to one specific meta-gene (referring to one node in the SOM grid) downscales the data from the M-dimensions of the transformed data into a two-dimensional coordinate system which however preserves the multivariate character of the input data and thus allows their direct visualization in a simple x-y-plot (see below).

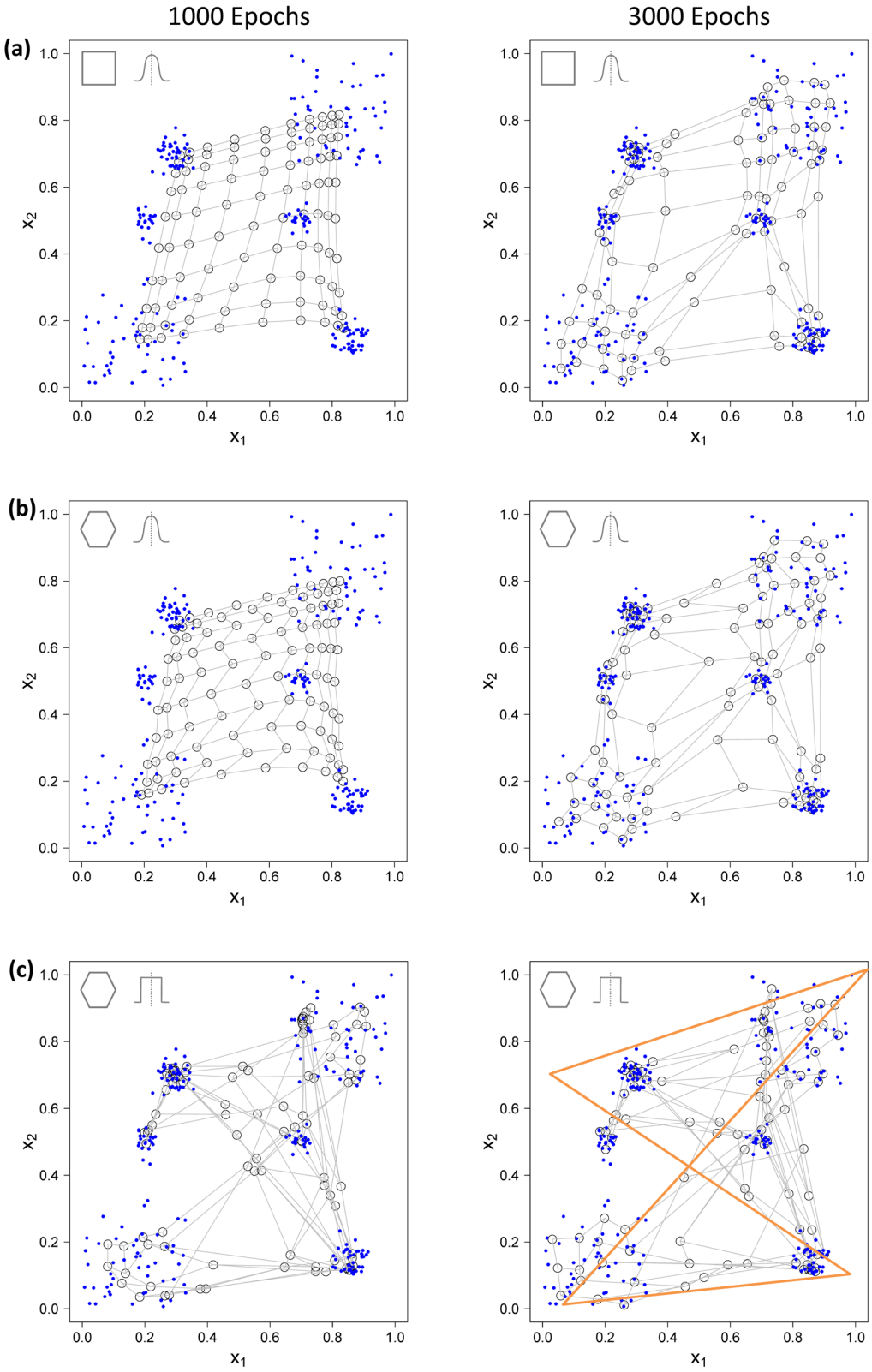


**Figure 2-5: Adjustment of meta-data (open circles) to input data (blue dots) using the same data as in Figure 2-4 but different grid topologies and/or neighborhood functions as indicated by the icons in the top left corners of the plots): a) topology/neighborhood ≙ rectangular/Gaussian; b) hexagonal/Gaussian; c) hexagonal/bubble. The orange lines in panel c indicate a snarled, 'knot-like' topological defect in the network of meta-data.**

### 2.4.4 Summary

The SOM approach provides a powerful combination of clustering, dimension reduction and multidimensional scaling: Mapping of the genes to the meta-genes partitions data space into clusters of genes with similar expression profiles. The meta-gene profile thereby serves as a representative for the respective cluster. These meta-profiles are well suited to be used in downstream analyses instead of utilizing the original data because they potentially provide a higher-level of information with reduced dimension, subsuming the profiles of a set of genes [WIRTH1]. Throughout this thesis, such clusters will appear in different contexts and will be referred to as meta-genes, meta-samples, meta-spectra, meta-genesets etc., depending on the particular application.

## 2.5 Adjusting SOM size and storage capacity

The primary feature of the SOM is extracting and storing information from the training data. The size of the SOM, i.e. the number of nodes and thus number of representative meta-genes, is the limiting parameter for the granularity of this memory. Adjusting this size is consequently an optimization task to obtain a sufficiently resolved map with available computational resources. A too small SOM is unable to capture the diversity of the data, e.g., to distinguish between different expression modes. On the other hand, a too big SOM requires excessive computational resources in terms of CPU-runtime and storage capacity.

The number of distinguishable expression modes inherent in the data set is the crucial issue which governs the required size of the SOM. Preferably, the SOM is capable to locate the major modes in distinct regions of the map to enable their identification and mutual separation. Discretized artificial data sets are generated to deduce an approximate rule relating SOM-size to its resolution, or in other words, its 'information-storage' capacity. Particularly, we used profiles of binary or ternary data, where the former data can adopt the values 0 and 1 whereas the latter data divide into three possible states -1, 0 and 1. The two-state model applies, for example, to data characterizing the presence and absence of gene expression and the three state model to discretized data describing under-, basal (i.e. unchanged) and overexpression levels with respect to a reference.

These binary and ternary artificial data sets are generated for varying number of states M defining the length of the profile vectors. The maximum number of different modes, which can be generated in such data sets, is $2^M$ and $3^M$, respectively. Table 2 lists the number of distinct expression modes for realizations of the binary and ternary approaches with M=2..6 as examples. Then, SOMs[7] of different sizes were trained using each of these data sets to find the minimal dimension of the node grid which is capable to separate all modes inherent in the

[7] Setup: linear initialization, Gaussian neighborhood, number of epochs: 1,000

**Table 2: Characteristics and SOM size of binary and ternary artificial profiles. No minimal SOMs were determined for ternary data and M>3 due to large number of modes.**

| | Binary | | Ternary | |
|---|---|---|---|---|
| **M (states)** | Number of modes | SOM size | Number of modes | SOM size |
| **2** | 4 | 5x5 | 9 | 5x5 |
| **3** | 8 | 8x8 | 27 | 18x18 |
| **4** | 16 | 9x9 | 81 | 50x50 |
| **5** | 32 | 18x18 | 243 | --- |
| **6** | 64 | 40x40 | 729 | --- |

respective input profiles (see Figure 2-6). These modes are required to map to individual SOM nodes well separated by 'empty' nodes, i.e. meta-profiles without associated single profiles.

Figure 2-6a shows an 8x8 SOM trained with 8 expression modes produced by M=3 binary states. As challenged, these 8 modes occupy 8 nodes equidistantly distributed along the edges of the map. This pattern is characteristic for self organization and becomes even more clearly visible for larger numbers of modes (Figure 2-6b). The SOMs trained using ternary data show similar results (Figure 2-6c and d). Figure 2-6a and c assigns the individual meta-profiles for binary (8 modes, M=3) and ternary (9 modes, M=2) artificial data to the respective tiles in the mosaic map, respectively. The eight binary modes virtually arrange along a circle according to the mutual similarities of their profiles: The Hamming distance[8] between all neighboring profiles equals 1. For example, meta-profiles 'A' ("0 0 1", see also Table 3 for assignment of modules and labels) and 'B' ("1 0 1") solely differ in the first position of their profiles. In other words, passing from one mode to the next one changes exactly one value in their profiles.

One of the 9 ternary modes mapped into the SOM shown in Figure 2-6c occupies the central position. It represents the invariant "0 0"-profile (label '*H*'). Such neutral modes usually form an invariant center of the map. The other modes containing at least one non-zero value in their profiles group around the center in a symmetric fashion where increasing profiles are found above the diagonal line and decreasing ones below this line. This symmetry reflects the fact that the SOM tends to arrange mirror-symmetric profiles into opposite regions of the map.

[8] number of positions at which the values in the considered meta-gene profiles are different

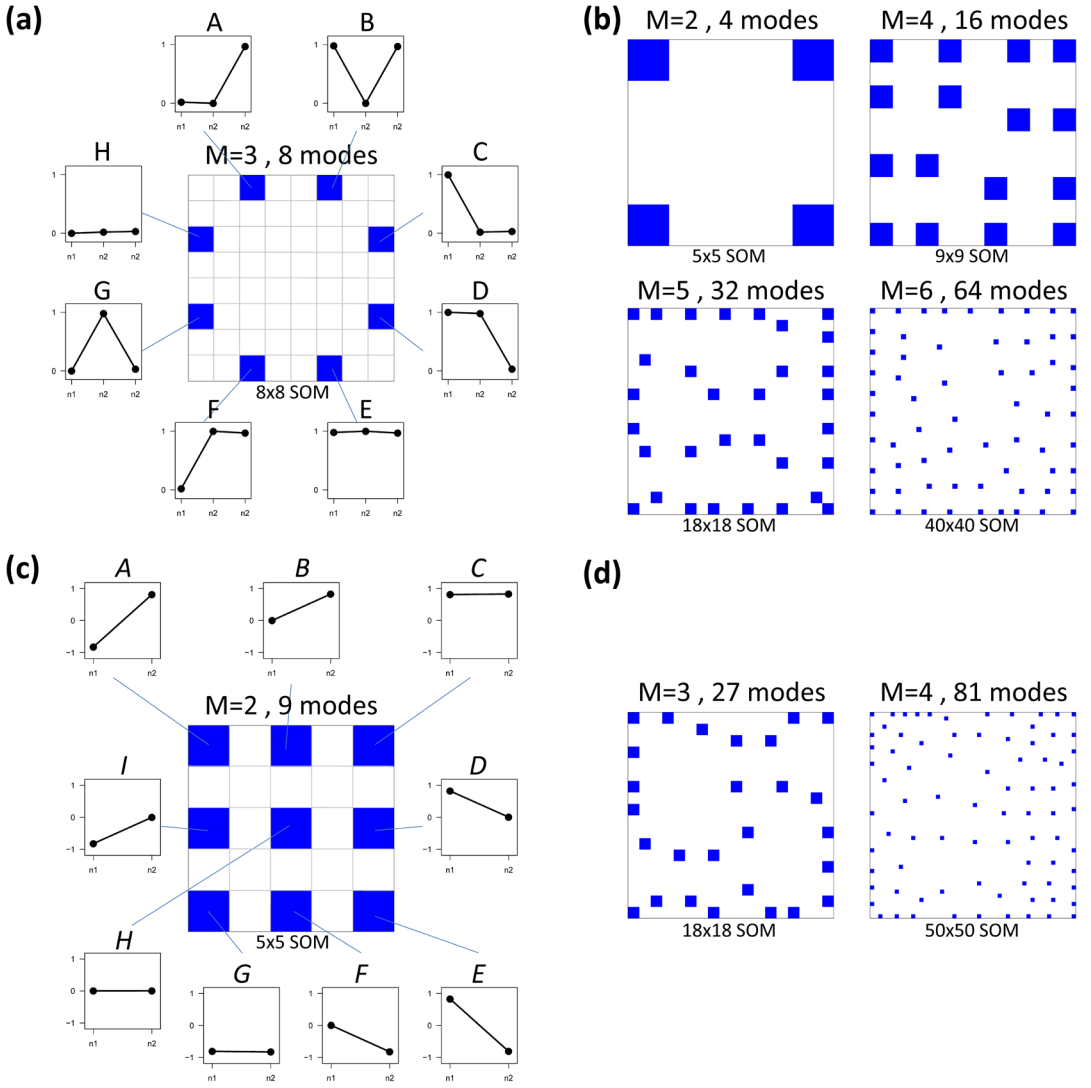


**Figure 2-6: SOM mapping of multimodal expression data: Blue and white tiles in the maps indicate occupied and empty meta-genes, respectively. Both binary (panels a and b) and ternary (panels c and d) profiles are mapped to the grid. The profiles of the respective modes are shown in panel a and c (see also Table 3).**

In summary, SOM learning systematically arranges expression modes according to the following principles:

- Similar profiles are mapped in close position, more different profiles are mapped more distantly.
- Neutral, non-differential and invariant profiles tend to occupy the center of the map.
- Antagonistic modes (i.e. strongly anti-correlated ones) tend to occupy mirror symmetric positions in opposite regions of the SOM whereas orthogonal modes (i.e. mutually independent ones) tend to divide the SOM into different segments each referring to one of the independent modes.

**Table 3: Assignment of modes and labels given in Figure 2-6a and c for binary and ternary data sets, respectively.**

| Binary, M=3 states | | | | Ternary, M = 2 states | | |
|---|---|---|---|---|---|---|
| Module profile | | | Label | Module profile | | Label |
| 0 | 0 | 0 | H | -1 | -1 | *G* |
| 0 | 0 | 1 | A | -1 | 0 | *I* |
| 0 | 1 | 0 | G | -1 | 1 | *A* |
| 0 | 1 | 1 | F | 0 | -1 | *F* |
| 1 | 0 | 0 | C | 0 | 0 | *H* |
| 1 | 0 | 1 | B | 0 | 1 | *B* |
| 1 | 1 | 0 | D | 1 | -1 | *E* |
| 1 | 1 | 1 | E | 1 | 0 | *D* |
| | | | | 1 | 1 | *C* |

The SOM mapping of artificial data has shown that the number of resolved modes roughly scales with the SOM size (see Table 2). Particularly for the binary profiles it was found that increasing the number of states M by one implies to double both the number of modes and the minimal SOM size. For the ternary profiles each additional state requires to triple the minimal SOM size.

Note that the structure of our synthetic data is relatively simple and typically not comparable with real-world examples: The human tissue data set analyzed below consists of M=67 samples. It roughly refers to about $10^{20}$ binary, or $10^{32}$ ternary expression modules. However, the diversity of such continuous expression profiles is potentially much larger. On the other hand, not all possible modes are present in the data owing to correlations between the data. Therefore a basic question addressed by our SOM analysis is about the effective number of distinct modes inherent in real data sets and the characterization of their interrelations.

## 2.6 Visual presentation of SOM data

### 2.6.1 Challenges

The SOM algorithm captures expression modules inherent in the data by sophisticated sampling of the input data space resulting in transformed data given in terms of the meta-gene profiles. Multivariate and multidimensional information is preserved in the meta-data after dimension reduction. This chapter demonstrates how to visualize the meta-data such that most relevant expression modules can be extracted in a simple and intuitive fashion.

Established approaches, such as the 'popular' two-way hierarchical clustering heatmaps (e.g. conditions-versus-genes), are well suited to visualize relatively simple covariance structures in the data as illustrated in Figure 2-7a: This first example refers to four conditions (A-D) where

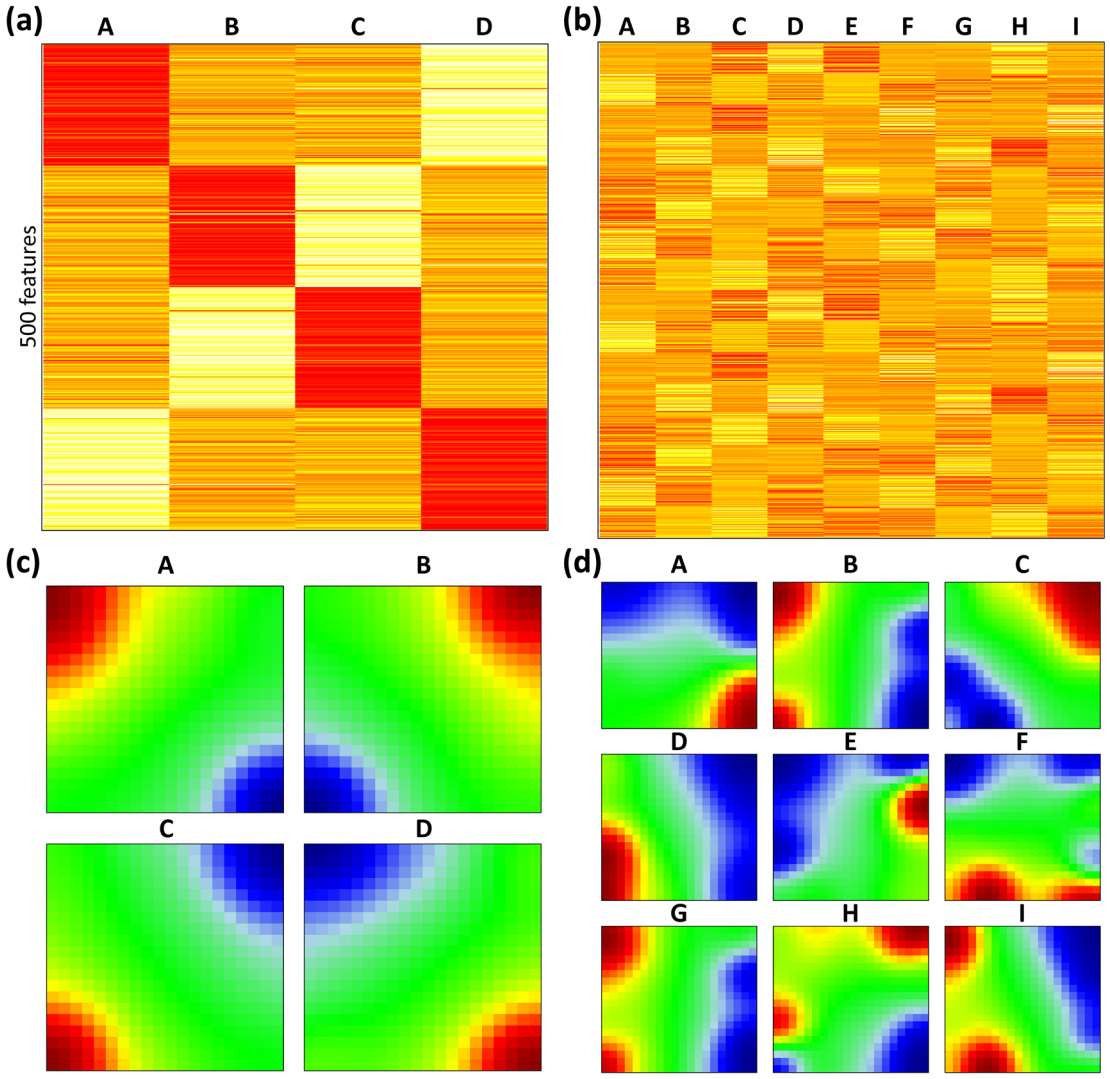


**Figure 2-7: Standard two-way cluster heatmap visualization (panels a and b) and the respective SOM portraits (panels c and d) of an artificial data set, which contains several distinct expression clusters. Evaluation of the heatmap turns into a puzzling task compared with the SOM portraits for the second example. The SOM portraits promote identification of similar patters ('B' and 'G'), for example.**

each is characterized by a unique cluster of overexpressed and a unique cluster of underexpressed genes (see red and yellow squares along the two diagonal lines, respectively). The second situation presented in Figure 2-7b is much more puzzling: It is virtually impossible to extract general relations between the 9 samples and/or about the 16 clusters of co-regulated genes using the heatmap presentation. This simple example illustrates that heatmaps are impractical when utilized to present high-dimensional data with complex intrinsic covariance structures. These problems are related to the 'chessboard'-like texture of the heatmap which becomes confusing for visual perception if the number of clusters exceeds a certain number. Moreover, heatmap presentations are virtually univariate, i.e. multivariate covariance structures become fragmented into 'univariate pieces'.

Contrary, the expression meta-states generated by the SOMs can be visualized in an alternative fashion by transforming them into mosaic portraits of each sample showing a blurry, color texture as illustrated in Figure 2-7c and d. The 'simple' example in Figure 2-7c transforms into four sample-specific portraits each showing one red and one blue spot which contain the genes specifically over- and underexpressed in the respective sample. Hence, in this simple case the well-separated clusters in the heatmap transform into well-separated spots in the SOM portraits. Both visualizations, heatmap and SOM, are virtually equivalent in this respect.

The situation however is different for the second example. The artificial data contains various distinct expression modules, highly expressed in one or several samples. These clusters emerge as red squares in the heatmap (Figure 2-7a) and as red spots the sample portraits (Figure 2-7b). The spot-like texture of the individual SOM portraits enables much better identification of analogies and differences between the samples than the 'chessboard'-pattern of the heatmap. For example, the SOM clearly and immediately reveal that samples 'B' and 'G' are almost identical, slightly vary compared with sample 'I' and completely differ compared with, e.g., 'E', 'C' and 'A'. Modules of high expression in multiple samples emerge as common red spots shared by the respective sample portraits, for example the spot in bottom left corner in samples 'F' and 'I' (Figure 2-7b). Sample specific clusters in turn appear as unique spots evident in the respective portrait only. The 'spot-pattern-like' visualization of the expression meta-states is consequently well suited to display the modularity of the data. Contrary to univariate heatmaps, the multivariate covariance structure of the samples translates into shared spots allowing evaluation of the relations between the samples.

Additionally, SOM portraits combine sample- and gene-centered perspectives: Firstly, the portraits represent the expression state of each sample and thus provide a visual entity for each of the samples. Secondly, the SOM portraits comprise information of all meta-genes, which are in turn representative for the complete set of genes mapped to the SOM. In this sense, SOM portraits allow assessment of individual samples with high resolution that allows identification of specific features essential for differential expression analysis.

### 2.6.2 SOM portraits and profiles

SOM portraits provide the primary way to display the expression (meta-) states of the samples with individual resolution. These portraits directly transform the columns of the meta-data matrix into colored mosaic pictures (see Figure 2-8) [37]. The K meta-genes (i.e. SOM nodes) are arranged in a two-dimensional grid with x and y tiles per dimension. Square SOMs with K=x·x are frequently used, without loss of generality. Thus M sample-specific SOM portraits are generated by color-coding each tile according to the expression value of the meta-gene assigned to this tile in the respective sample m, $\Delta e_{1,m}^{meta}$ … $\Delta e_{K,m}^{meta}$ (k=1…K).

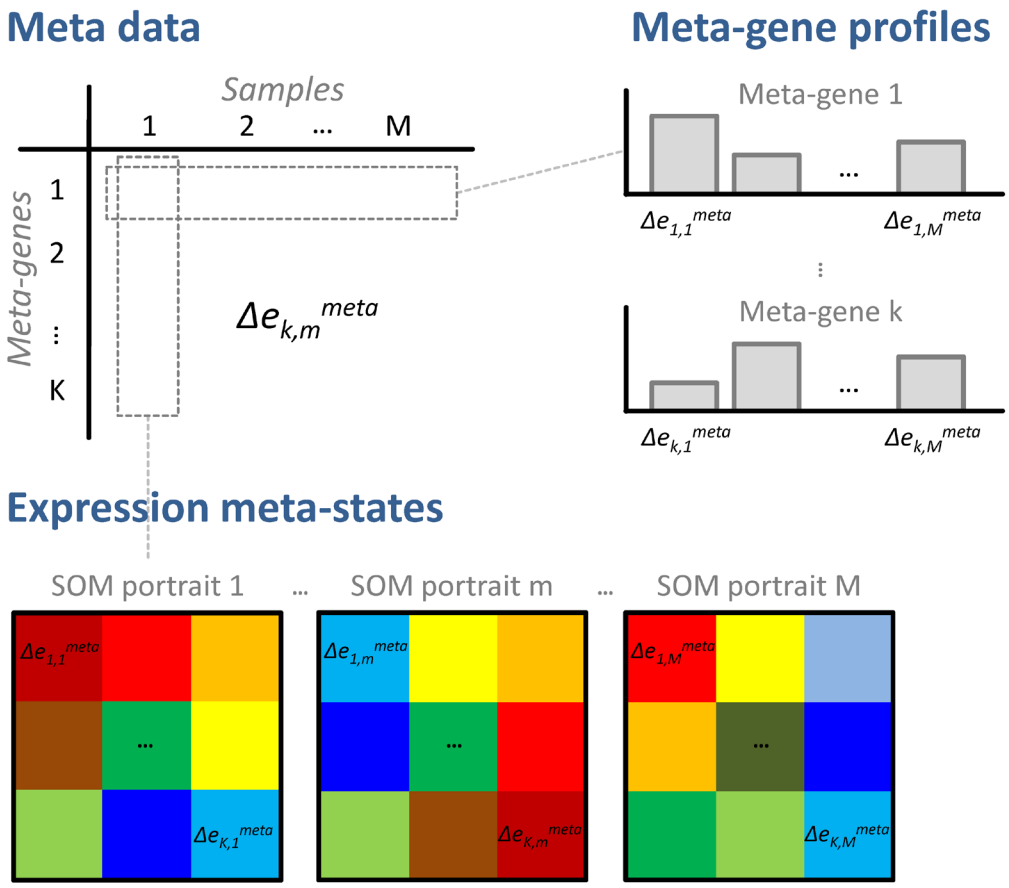


**Figure 2-8: Visualization of the meta data matrix: Meta-gene expression profiles (rows) are shown as barplots. Expression meta-states of the samples (columns) are transformed into mosaic portraits by arrangement into a grid according to the SOM's topology (here K=3x3 nodes with rectangular layout) and application of a suited color code.**

The color gradient of the map was chosen to properly visualize over- or underexpression of the meta-genes: Maroon codes the highest level of gene expression; red, yellow and green indicate intermediate levels and blue corresponds to the lowest level of gene expression. Owing to similarity of adjacent meta-gene profiles, the color patterns emerge as smooth textures rather than noisy pixels. These coherent mosaic patterns are characteristic for each sample and represent a fingerprint of the transcriptional activity in the corresponding expression meta-state [5]. Individual expression modules emerge as spots of similar colored tiles (see Figure 2-7), which correspond to clusters of co-regulated genes. Note that the assignment of genes to meta-genes and of meta-genes to mosaic tiles is identical in all sample portraits. So they can be directly compared to each other allowing immediate identification of unique or ubiquitous expression modules.

The complementary meta-gene profiles are derived from the rows of the meta-data as indicated in Figure 2-8. The meta-gene profiles are representatives of clusters of co-regulated genes. They can be interpreted as expression modules inherent in the data set. The number of meta-genes is markedly smaller than the number of single genes. The number of relevant meta-gene profiles will be further reduced by collecting them into clusters of similar ones using different criteria (see below).

### 2.6.3 Expression portraits of human tissues

To illustrate SOM visualization we generated SOM portraits and meta-gene profiles of the expression landscapes of 67 human tissue samples using gene expression microarray data. The samples are grouped into 10 different tissue categories in accordance with common classifications (e.g. Hornshoj et al. [64]). After preprocessing as described above, a SOM was trained with a resolution of K=60x60=3,600 meta-genes. The created SOM portraits are shown in Figure 2-9. Each tile of the portrait mosaics refers to one of the 3,600 meta-genes characterizing the particular expression level in this tissue. The number of co-regulated single genes per meta-gene typically varies from meta-gene to meta-gene (see population map below).

Most of the samples within one tissue category show similar SOM portraits which are characterized by typical red and blue spots at specific positions due to over- and underexpressed meta-genes as the most evident features. For example, the portraits of adipose tissues (numbered 1-3, first row in Figure 2-9) might be identified by the maroon-red overexpression spot in the bottom right corner and those of nervous tissues (numbers 45-67, last three rows) by a coherent spot in the top left corner. In general, SOM profiles within a tissue category reveal similar pattern, whereas different tissue types show consistently different expression patterns. Such differences can be detected, for example, by simple visual inspection of the mosaic pattern of nervous, immune system and endocrine type tissues. Hence, comparison of the SOM-textures allows the straightforward grouping of the tissues into different categories based on differences of their expression patterns.

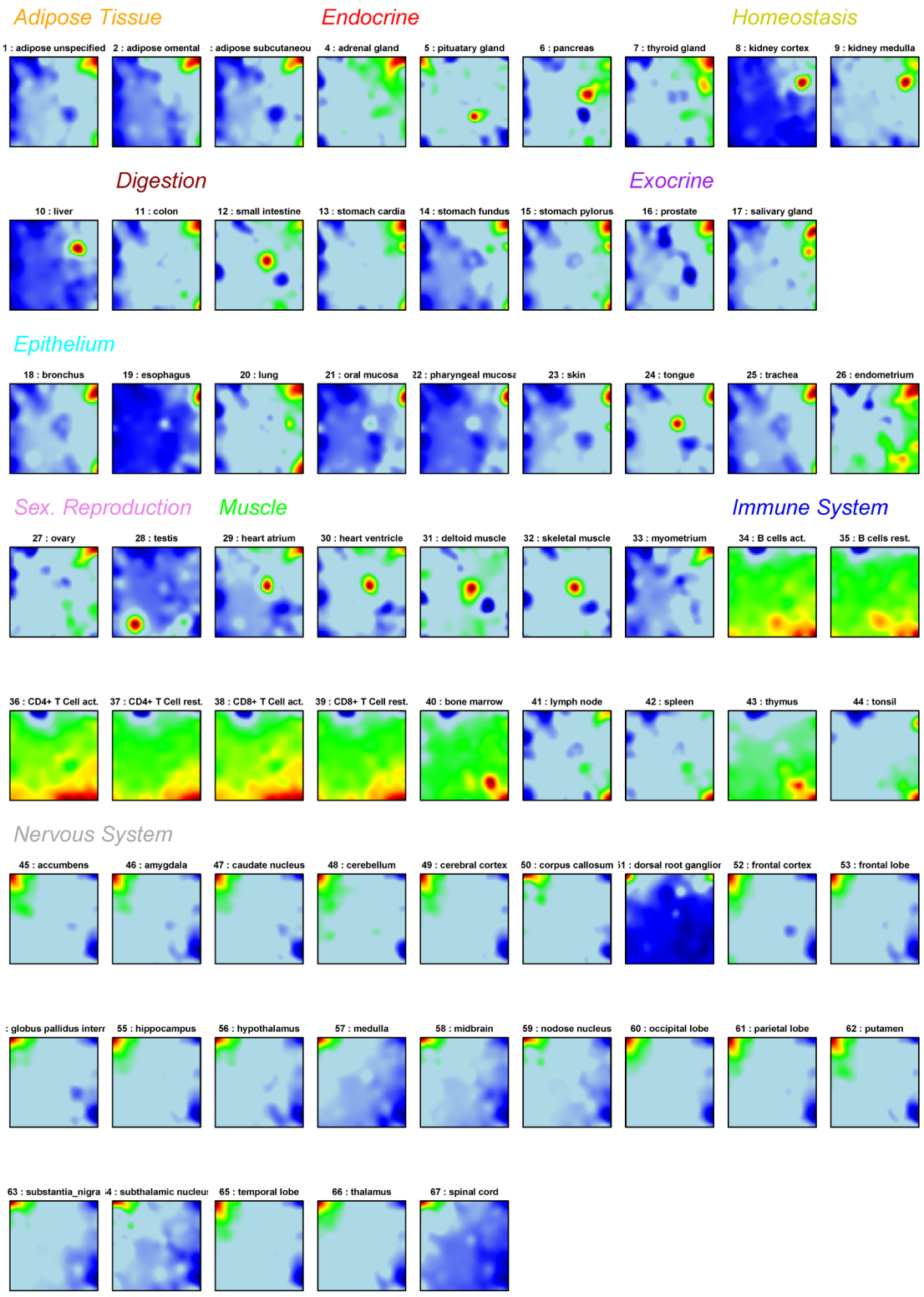


**Figure 2-9: SOM portraits of the tissue transcriptome data set. The tissues are arranged according to tissue categories as indicated by the headlines, whose colors are used throughout this thesis to represent the tissue categories.**

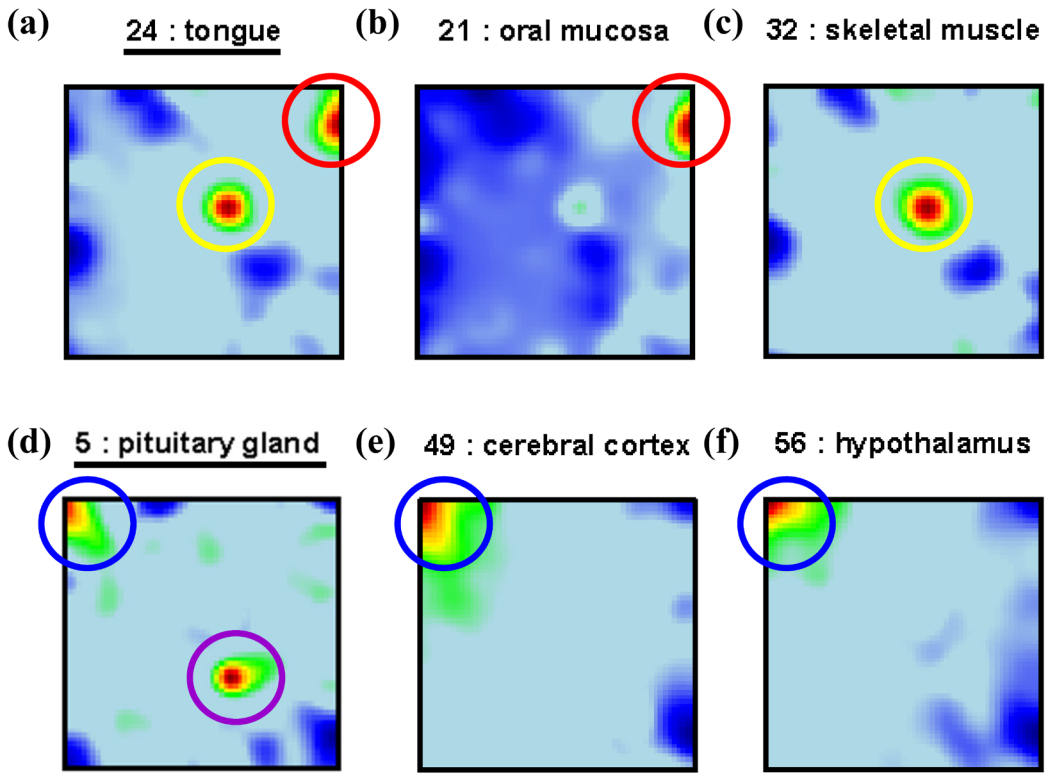


**Figure 2-10: Specific spots in selected sample portraits: The SOM-pattern of tongue (panel a) shows two spots of overexpressed meta-genes. One of them is characteristic for mucosa type tissues (b; red circles) and the other one is found in muscle tissues (c, yellow circles). Pituitary gland (d) shows a specific spot for this particular tissue and one which is characteristic for nervous system tissues (e and f, blue circles) as well.**

Moreover, some tissues combine the characteristic spot pattern of different tissue categories (see Figure 2-10). For example, the sample portrait of tongue (panel a) shows the typical overexpression spot evident in the portraits of other epithelial tissues (e.g. oral mucosa, panel b) but also the spot typically found in muscle tissues (e.g. skeletal muscle, panel c). The physiology of tongue tissue as a 'muscle covered by mucosa' is thus reflected in the SOM portraits. Another example is pituitary gland (panel d), an endocrine gland located near hypothalamus: Its portrait shows the overexpressed spot found also in other nervous tissues (e.g. cerebral cortex or the adjacent hypothalamus, panel d and e, respectively) in the top left corner, as well as a unique spot in the bottom right area not found in the portrait of any other tissue. This spot obviously collects genes which are specifically overexpressed in pituitary gland (see below), whereas the first spot represents a common signature of nervous system samples. Some SOM portraits represent outliers in their tissue category: For example, small intestine (no. 12), classified as digestive tissue, shows the overrepresentation pattern of muscle type tissues. This result does not surprise because small intestine consists of a double layer of smooth muscle. Myometrium (no. 33), the smooth muscle of the uterus, is classified as muscle. Its SOM portrait however closely resembles that of endometrium (no. 26) and also of ovary (no. 27), reflecting the common function of these three organs in female reproduction.

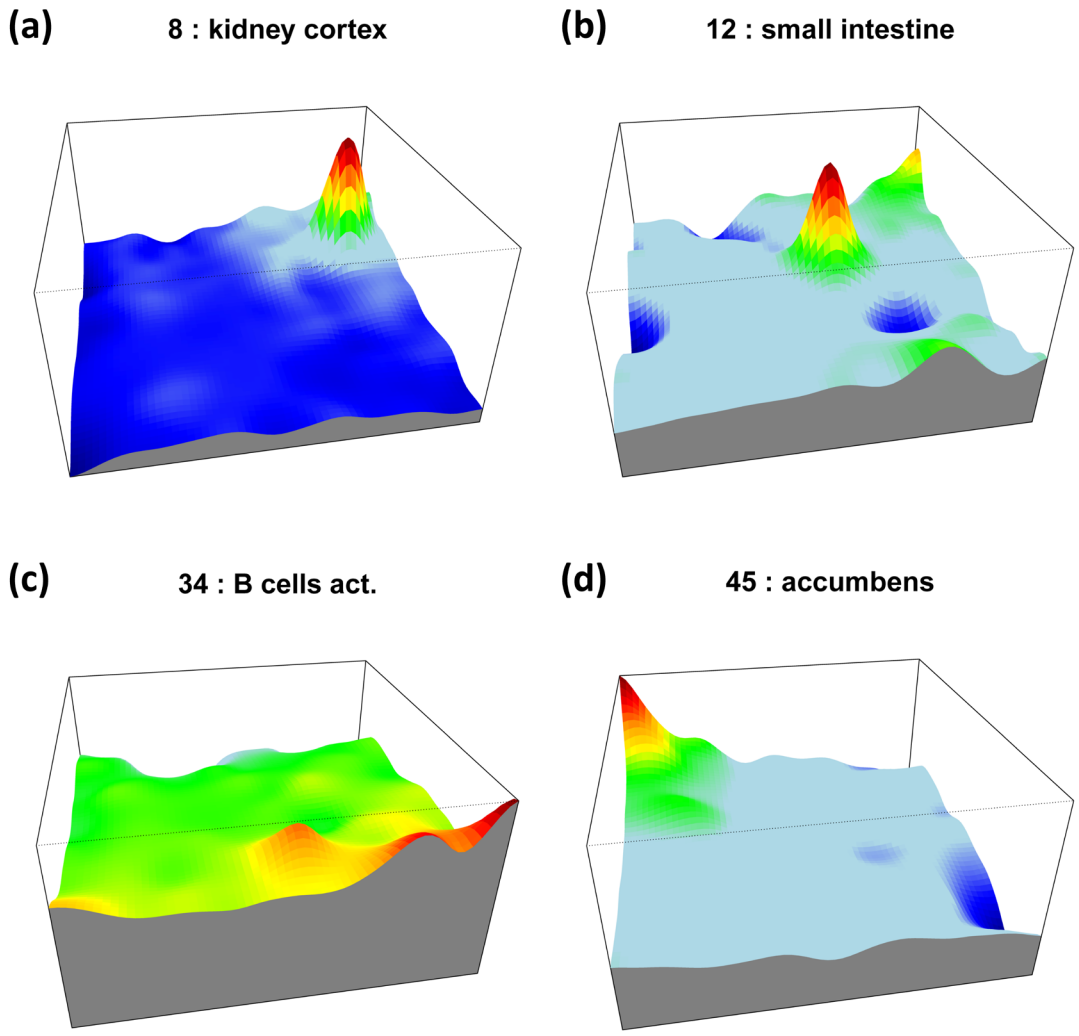


**Figure 2-11: Three-dimensional perspective plots characterize the expression landscapes of selected tissues. The four tissue samples show notably flat 'plain-like' (kidney cortex, panel a), undulating 'hilly' (small intestine, b) and smoothly sloping (B-cells, c, and accumbens, d) meta-gene landscapes, respectively.**

The SOM portraits visualize three-dimensional information where the expression level of the meta-genes is appropriately color coded. In special situations, the three dimensional plot of the expression values further improves visual perception because it explicitly presents the expression landscape of a sample in terms of 'mountains', 'valleys' and 'plains'.

For example, the meta-gene landscape of kidney cortex (Figure 2-11a) is, except for the overexpression peak in top right region, remarkably flat, which indicates basal expression of most of the genes. Contrary small intestine (Figure 2-11b) features a multivariate landscape with diverse 'hills' (regions of overexpressed meta-genes) and 'valleys' (underexpressed meta-genes). This reflects multiple modules of (meta-)genes which are over- and underexpressed in the respective samples in concerted fashion. Finally, B-cell and nucleus accumbens samples

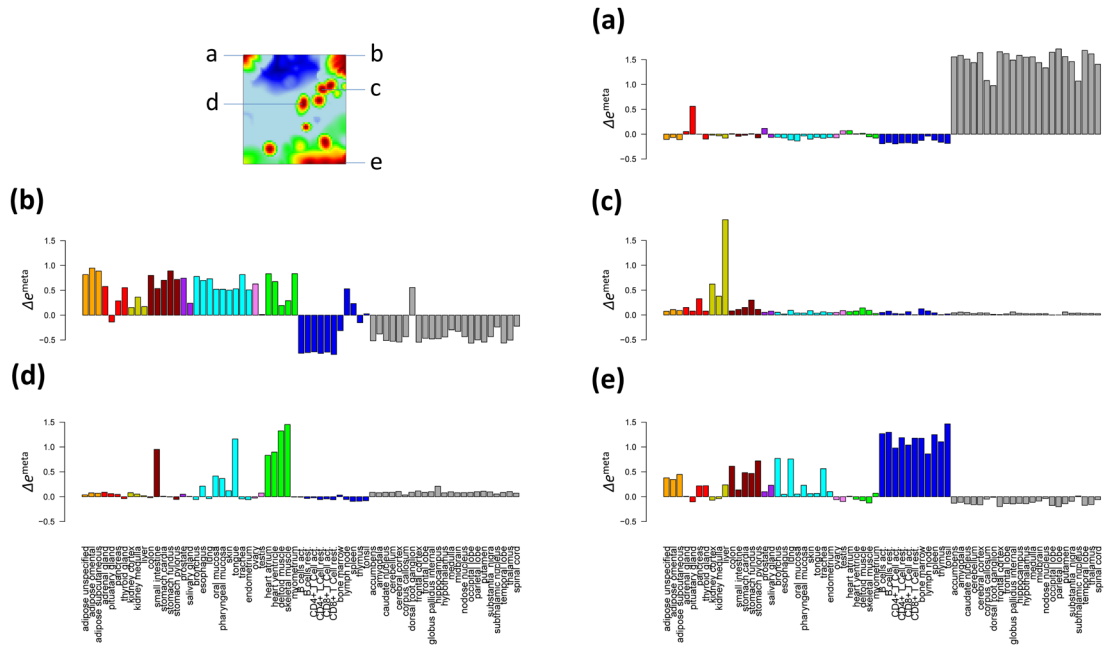


**Figure 2-12: Meta-gene profiles derived from the human tissue transcriptome SOM. The position of the five meta-genes in the SOM grid is indicated in the top left overexpression map (see chapter 2.7.1 for details). The bars are arranged and colored in accordance to the different tissue categories (see Figure 2-9).**

(Figure 2-11c and d, respectively), chosen exemplarily for all immune respectively nervous system samples, show a steady slope along the diagonal from strong overexpression to strong underexpression. Two major groups of meta-genes overexpressed in nervous system (located in the top left corner) and underexpressed in immune system (bottom right corner, see immune system portraits in Figure 2-9) and vice versa form an antagonistic couple of expression modules. This example also confirms the previous finding that the SOM training algorithm tends to distribute meta-genes with anti-correlated expression profiles into opposite corners of the map.

Complementary to the SOM portraits which characterize the expression state of all meta-genes in one sample, we generated expression profiles of selected meta-genes, each characterizing the expression of one meta-gene in all samples studied. Figure 2-12 shows the profiles of five meta-genes selected from different regions in the SOM. The profiles are assigned in the small map in the figure by lower case letters a – e. For example, the profile of meta-gene (a) located in top left corner of the SOM clearly exhibits virtually binary activity of the associated genes which are strongly upregulated in nervous tissues (gray bars) and downregulated in almost all remaining tissues. Contrary meta-gene (b), located in top right corner of the SOM, is overexpressed in a variety of samples assigned to various tissue categories, and underexpressed in immune and nervous tissue samples. Profiles (c), (d) and (e) refer to genes overexpressed in endocrine, muscle and immune system tissues,

respectively. The meta-gene profile plots therefore give insight to the diversity of expression modules captured in the meta-gene clusters.

In summary, the SOM portraits (given either as 2D-projections or 3D-perspective plots) characterize the expression landscapes of the samples in terms of intuitive color-textures, whereas the meta-gene profiles illustrate the expression of selected modules in the series of samples studied.

### 2.6.4 Adjusting contrast in SOM portraits

The standard SOM portraits represent differential expression in units of the logarithmic fold change of the meta-genes, log FC = $\Delta e_{k,m}^{meta}$. The observed spots thus reflect regions of over- and underexpression in the respective meta-gene profiles in logarithmic scale (Figure 2-13a). Alternative scales, such as the double logarithmic log log-FC and the weighted average difference (WAD) score, are applied to vary the contrast of the texture of the SOM portraits in order to highlight different aspects of the expression meta-states:
The WAD-score is calculated for each tile k and sample m according to

$$\mathrm{WAD}_{k,m} = w_{k,m} \cdot \Delta e_{k,m}^{meta} \quad \text{with} \quad w_{k,m} = \frac{\Delta e_{k,m}^{meta} - \min\left(\Delta e_{k,m}^{meta}\right)}{\max\left(\Delta e_{k,m}^{meta}\right) - \min\left(\Delta e_{k,m}^{meta}\right)} \tag{5}$$

The WAD score is a fold change (FC)-based score which 'amplifies' large expression values implementing the observation that 'strong signals are better signals' [65, 66]. Equation (5) adapts the WAD score for meta-gene expression values (compare to WAD score for single genes in chapter 5.2.1). The visualization of the meta-gene WAD-score thus highlights peaks due to overexpression, leading to sharply defined spots with high contrast as shown in Figure 2-13b.
The log log-FC as third option rescales the original log-FC into double-logarithmic units giving rise to a wider distribution in the positive and negative expression ranges, respectively:

$$\text{loglog-FC}_{k,m} = sign(\Delta e_{k,m}^{meta}) \cdot \log\left(1 + \left|\Delta e_{k,m}^{meta}\right|\right) \tag{6}$$

This strongly enhances the discrimination between up- and downregulated meta-genes (Figure 2-13c). The log log-FC scale thus exhibits structured blue and red areas of characteristic shape which clearly emphasizes the borderline between the regions of over- and underexpression. These details are not or only hardly detectable in the log-FC and WAD scales. In contrast, the latter scales express spot-like patterns, which are mostly characteristic for the samples.
The considered options of contrast variation enable accentuation of different ranges of differential meta-gene expression with focus on strong till moderate differential expression

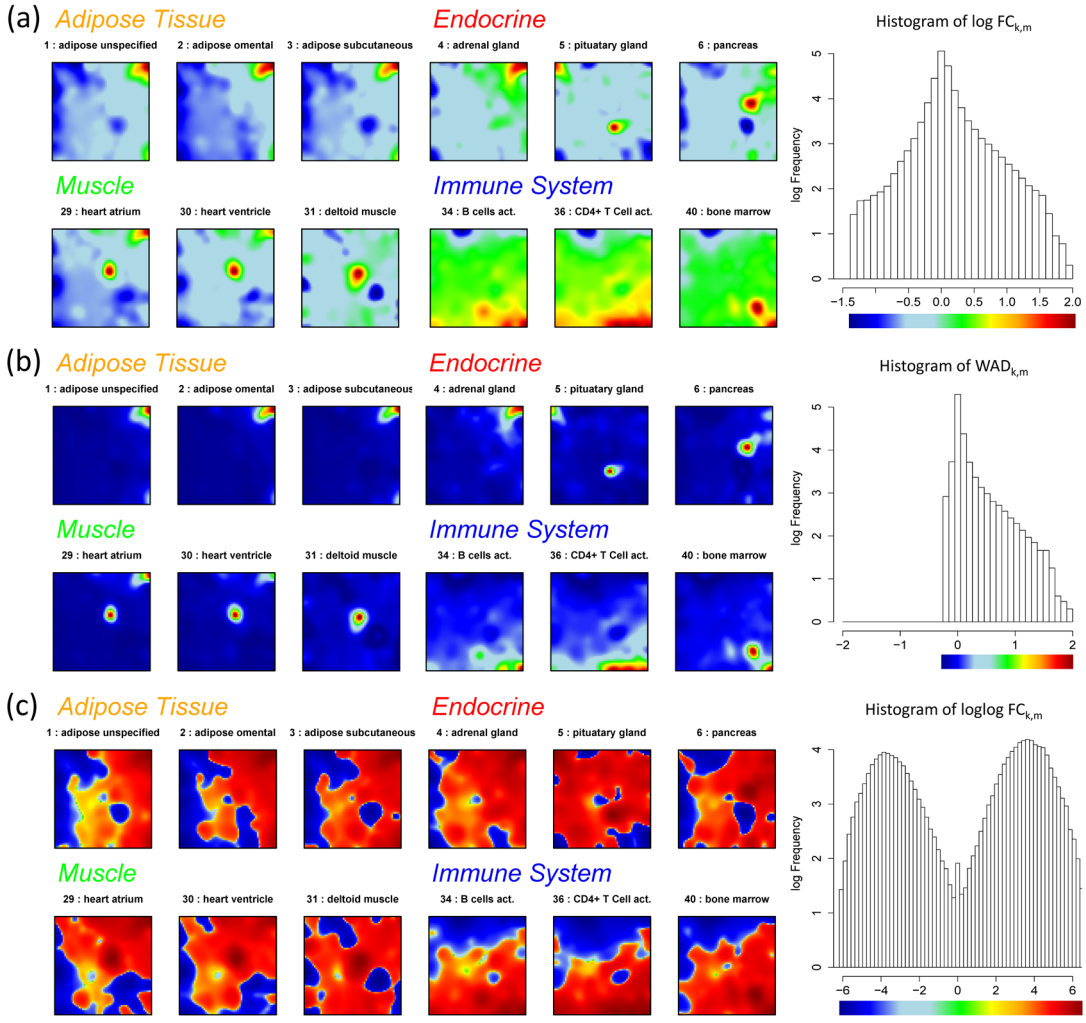


**Figure 2-13: Contrast variation of the SOM portraits using different expression scores in selected tissues: Differential meta-gene expression relative to the mean expression in all samples studied in logarithmic (panel a, Eq. (16)) and double-logarithmic scale (c, Eq. (6)) and using the WAD-score (b, Eq. (5)). The right part of the figure shows the frequency distribution of the scores in logarithmic scale.**

(log-FC), very strong overexpression (WAD) or weak till moderate differential expression (log log-FC). For example, the three adipose tissues show very similar portraits with essentially the same spot of overexpression in the log-FC and WAD scales, whereas the log log-FC map reveals subtle differences between the underexpressed blue regions of 'adipose omental' tissue and the other types of adipose tissues.

### 2.6.5 Supporting maps

We define the following supporting maps to provide additional information about the clusters defined by each meta-gene and the associated real genes. These supporting maps use the same resolution of the two-dimensional mosaic grid as the SOM portraits and appropriate color-scales for direct comparison.

*Population map*

The SOM-algorithm maps the expression profiles of the N input genes to a number of K<<N meta-genes. Each meta-gene thus serves as a sort of condensation nucleus for a cluster of $n_k$ co-regulated 'real' genes. As each gene is mapped to one and only one meta-gene, the sum of all $n_k$ amounts to N: $N = \sum_{k=1..K} n_k$ .

The population map (Figure 2-14a) plots the number of single genes per meta-gene in logarithmic scale, log $n_k$, into the mosaic grid according to the SOM portraits.

*Variance map*

The variance map (Figure 2-14b) illustrates the variability of the expression profile of each meta-gene in the samples studied,

$$\mathrm{var}_k = \frac{1}{M-1}\sum_{m=1}^{M}\left(\Delta e_{k,m}^{meta}\right)^2 \tag{7}$$

This map enables identification of neutral or non-informative (i.e. invariant) meta-genes in the SOM portraits, as well as informative ones representing distinct expression modules.

*Covariance map*

The covariance map (Figure 2-14c) visualizes the degree of concordance between the expression profiles of the single genes and those of the respective meta-genes in each tile of the mosaic portrait in terms of the cross correlation coefficient,

$$r_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathrm{cov}_{k,i} \Big/ \sqrt{\mathrm{var}_k \cdot \mathrm{var}_{k,i}^{single}} \tag{8}$$

$$\text{with} \quad \mathrm{cov}_{k,i} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\Delta e_{k,m}^{meta} \cdot \Delta e_{k,m,i}\right)$$

$$\text{and} \quad \mathrm{var}_{k,i}^{single} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\Delta e_{k,m,i}\right)^2$$

where $\Delta e_{k,m,i}$ denotes the expression value of gene i mapped to meta-gene k under condition m. The measure $r_k$ consequently reflects the average correlation of gene profiles to the associated meta-gene profile for each meta-gene (i.e. portrait tile).

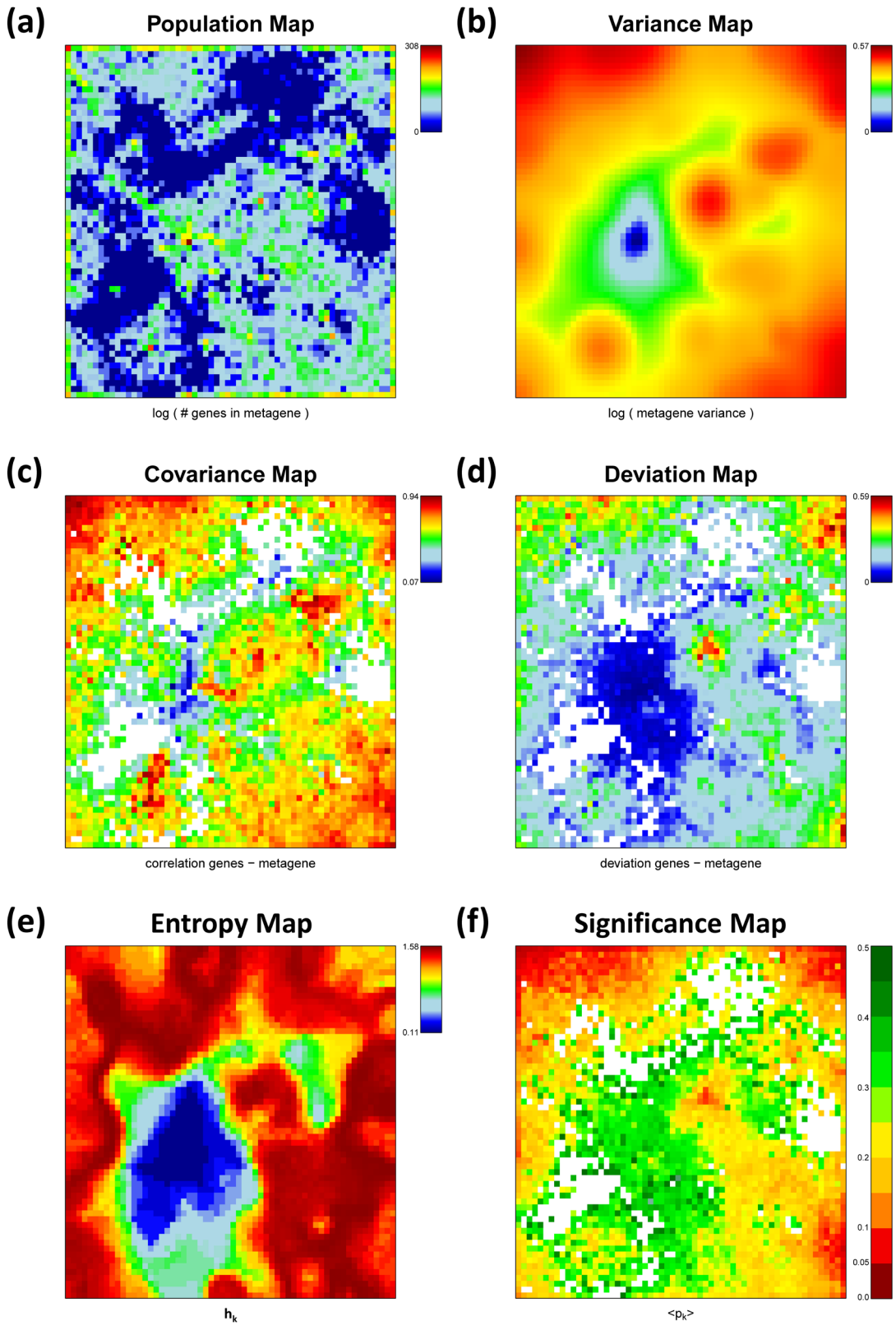


**Figure 2-14: Supporting maps characterizing the meta-genes extracted from the human tissue data set: Population (panel a), variance (b, Eq.(7)), covariance (c, Eq.(8)), deviation (d, Eq.(9)), entropy (e, Eq.(11)) and significance (f, Eq.(12)) maps.**

*Deviation map*

The deviation map (Figure 2-14d) visualizes the degree of concordance between the expression profiles of the single genes in each meta-gene cluster using the quadratic mean of the Euclidean distances between each meta-gene and the respective single gene profiles:

$$d_k = \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k} d_{k,i}^2} \quad \text{with} \quad d_{k,i}^2 = \frac{1}{M-1}\sum_{m=1}^{M}\left(\Delta e_{k,m,i} - \Delta e_{k,m}^{meta}\right)^2 \tag{9}$$

The 'deviation', meta-gene variance and covariance are linked according to ref. [WIRTH1]:

$$r_k = 1 - \frac{d_k^{\,2}}{2\,var_k} \tag{10}$$

Eq. (10) shows that correlation coefficients near unity are obtained for close similarity in terms of the deviation ($d_k \rightarrow 0$) and/or if the meta-gene variance largely exceeds the squared Euclidean distance, $d_k^{\,2} \ll var_k$. Note that the correlation coefficient vanishes for $d_k^{\,2} \approx 2\,var_k$.

*Entropy map*

The entropy map (Figure 2-14e) plots the standard entropy of each meta-gene profile,

$$h_k = -\sum_{i=1}^{3} \rho_{k,i} \cdot \log_2 \rho_{k,i} \tag{11}$$

where $\rho_{k,i}$ is the relative frequency of the three levels of gene expression: overexpression, underexpression and non-differential expression of meta-gene k. Hence meta-genes of a sample are assigned to one of the three levels by application of a defined threshold (here the 25- and 75-percentile of all meta-gene expression values was used). $h_k$ is restricted to values in the interval [0, $\log_2 3$]. An entropy value of 0 represents a perfectly 'ordered' state, where all meta-genes are assigned to only one of the expression levels. Contrary, maximum value of $\log_2 3 \approx 1.58$ is reached when meta-genes uniformly distribute over the three levels.

*Significance map*

The shrinkage t-score links differential gene expression with variance estimates and transforms into a significance measure $p_{g,m}$ for each gene g in sample m (see chapter 5.2). A simple approach of combining significance information for meta-genes is to calculate the mean score log-averaged over the meta-gene members and subsequently averaged over the samples:

$$< p_k >= \frac{1}{M}\sum_{m=1}^{M} \exp\left(\frac{1}{n_k}\sum_{i=1}^{n_k} \log(p_{i,m})\right) \tag{12}$$

The significance map (Figure 2-14f) plots these values for each meta-gene k into the mosaic grid, allowing easy identification of meta-genes that gather real genes with significant differential expression

The population, variance, covariance and deviation maps shown in Figure 2-14 provide information about special properties of the human tissue transcriptome SOM. The population map reveals that the single genes inhomogeneously distribute among the tiles of the mosaic grid. The tile of maximum population ($n_k$=308, see the maroon tile slightly left from the centre of the map in Figure 2-14a) refers to genes with virtually invariant, mostly absent expression in all tissues studied. These invariant genes give rise to the dark blue spot in the central area of the variance map (Figure 2-14b). Against this, spots of highly variant and thus information containing meta-genes are mostly located along the edges of the SOM grid.
The covariance and concordance maps show a similar but more noisy pattern than the variance map due to the fact that they explicitly process single gene profiles (Figure 2-14c and d, respectively). As the three measures variance, covariance and Euclidean distance are related properties (see Eq.(10)), the three maps confirm the concerted changes of real genes together with that of the associated meta-genes in each tile (compare Figure 2-14 b and c). The deviation map more accentuates meta-genes of low variance (blue areas in Figure 2-14d). Recall that the SOM algorithm uses the Euclidean distance between single and meta-gene profiles as similarity criterion to partition the single genes over the tiles of the mosaic. Close similarity in distance scale transforms into correlation coefficients near unity in the areas of relatively large meta-gene variance as predicted by Eq. (10) (see red areas in Figure 2-14 b and c). Contrarily, areas of relatively weak correlations largely agree with the regions of low meta-gene variance (see blue and green areas in Figure 2-14 b and c) which, in turn, lack marked over- and overexpression spots.
The entropy map (Figure 2-14e) reveals minimal entropy in the central part of the SOM, allocated by invariant meta-genes as shown by the variance map. The outer regions of the map with high variant meta-genes contrary imply higher entropy values. Notably, meta-genes of maximum entropy can be identified in the intermediate regions due to balanced over-, basal- and underexpression of the respective meta-genes across the tissues studied.
The significance map (Figure 2-14f) virtually resembles the variance map: meta-genes of high variance mainly show also high significance, and vice versa. Comparison of variance, entropy and significance maps reveals close similarities (compare Figure 2-14 b, e and f) because those measures are direct functions of the differential meta-gene expression (see Eq.(7), Eq.(11) and Eq.(19) below). On the other hand, the entropy map shows a more diverse substructure which allows identification of highly changing meta-genes due to the reasons discussed above.

### 2.6.6 Supporting profiles

*Variance and entropy profiles*

In order to estimate global properties of the expression landscape of every phenotype we calculated the variance of meta-gene expression values in each SOM portrait, $\mathrm{var}_m = \sum_k (\Delta e_{k,m} - < \Delta e_m >)^2 / (K-1)$, and its entropy, $h_m = -\sum_k \rho_{k,m} \cdot \log_2 \rho_{k,m}$ where $\rho_{k,m}$ 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*. This global entropy thus characterizes the information content of each portrait. The variance estimates the variability of the meta-gene expression and the entropy its information content, or in other words, its degree of ordering. Both, the variance and the entropy assess the expression landscape of phenotype *m* as seen by the SOM portrait and hence provide sample-centered information, complementary to tile-based supporting maps providing the respective metagene-centered information.

Supporting variability and entropy profiles of the expression states of human tissues are shown in the barplots in Figure 2-15a and b. The variability profile in Figure 2-15a shows the overall variance of the expression meta-state within each tissue sample. Interestingly, pancreas (endocrine tissues; red bar), liver (homeostasis; dark yellow), testis (sexual reproduction; pink) and T- and B-cells (immune system; blue) reveal large variability of the meta-gene states within their tissue categories. Recently, similar variability measures revealed likewise transitions between stages of organogenesis [44]. The meta-state entropy profile in Figure 2-15b also embraces such transitions from a complementary point of view. Generally, samples exhibit entropy values in the upper range of the potential interval [0, 1.58], indicating balanced distribution of meta-gene expressions to the three levels (see above). Prominent outliers such as pituitary gland (endocrine tissues; red bar), ovary (sexual reproduction; pink) or subthalamic nucleus (nervous system; gray) exhibit exceedingly high fraction of non-differentially expressed meta-genes. The entropy profile thus highlights information-less samples in this application, providing complementary information to the respective variance profiles.

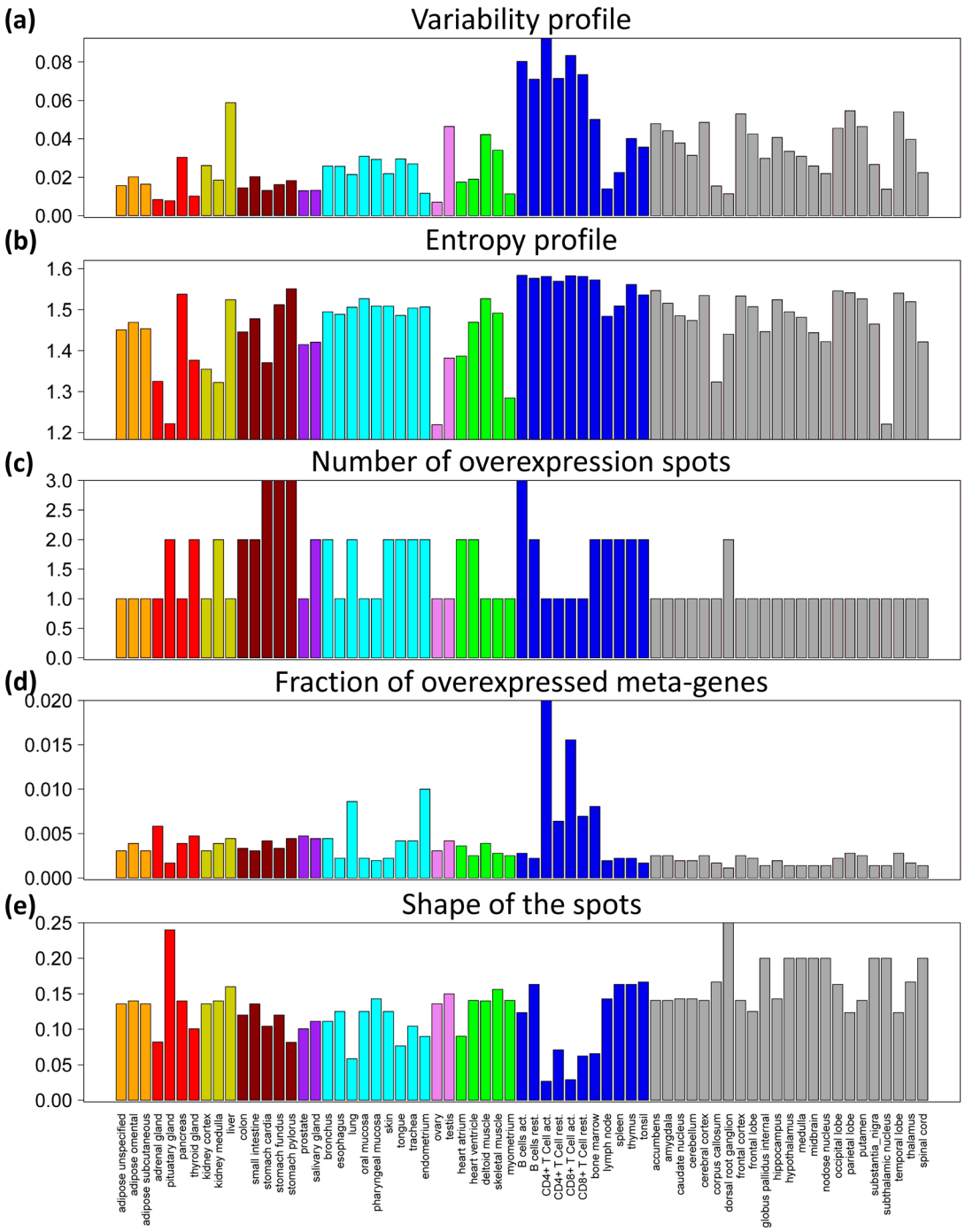


**Figure 2-15: Supporting profiles characterizing the expression meta-states of the human tissue samples and the SOM portraits: Panel (a) shows the variability of expression meta-states of the samples, (b) the corresponding entropy, (c) the number of overexpression spots, (d) the fraction of overexpressed meta-genes and (e) the shape parameter of the spots. The different colors of the bars indicate the different tissue categories.**

*Topological measures profiles*

Three additional supporting profiles are generated to get an overview about basic topological properties of the SOM portraits. Firstly, the number of overexpression spots observed in a sample portrait is determined using the 98-percentile criterion as described below. The barplot in Figure 2-15c gives this spot number for each sample studied. Secondly, the relative number of meta-genes included in all spots observed in one portrait,

$$f_m^{over} = \frac{K_m^{over}}{K} \quad \text{with } K_m^{over} = \sum_{k=1}^{K} \delta(\Delta e_{k,m}^{meta} > \Delta e_{threshold}) \tag{13}$$

describes the relative amount of overexpression in each expression meta-state. It is shown in Figure 2-15d for all samples. Thirdly, the *shape parameter*,

$$shape_m = \frac{K_m^{over}}{(K_m^{border})^2} \tag{14}$$

characterizes the fuzziness of the overexpression spots observed (Figure 2-15e). Here, $K_m^{border}$ denotes the number of tiles along the spot borderlines with at minimum one adjacent tile outside and one tile inside the spot. $K_m^{over}$ and $K_m^{border}$ thus estimate the area occupied by the spots and their limiting contour length, respectively. The shape parameter hence relates the actual area of the spots to an idealized area which is defined by the square of their contour length. For a single spot the shape-value decreases if its shape progressively deviates from a circular one. For *n* non-overlapping spots of identical area and border length, the shape parameter inversely scales with the number of spots, i.e. ~1/n. In general, the shape-value decreases if the number of spots increases [HOPP1].

Figure 2-15c shows the number of overexpressed spots in the tissue SOM portraits, revealing most frequent occurrence of solitary spots. Also a larger number of spots in digestive tissue portraits (see brown bars in Figure 2-15c) is revealed, originating from sets of genes also overexpressed in other tissue categories such as muscle and immune system. The fractions of overexpressed meta-genes (Figure 2-15d) show moderate variability, whereby immune system samples (different T-cell samples, bone marrow; blue bars) and two epithelium samples (lung, endometrium; cyan) strongly surpass the remaining samples. Also the shape coefficient of overexpression spots (Figure 2-15d) shows intermediate values for most of the samples, with few outliers which feature remarkably low (e.g. lymphocyte samples; see blue bars) or high values (pituitary gland and dorsal root ganglion; red and gray bars, respectively). Low shape coefficients originate from unshapely or longish spots, whereas high values are caused by especially round spots. These particular spot shapes are also observable in the SOM portrait gallery (Figure 2-9).

## 2.7 Global meta-gene clusters

The SOM algorithm arranges similar meta-gene profiles in neighbored tiles of the mosaic map whereas different profiles are located more distantly. Neighbored meta-genes thus tend to be colored similarly owing to their similar expression values. In consequence, the obtained mosaic portraits show typically a smooth texture with red and blue spot-like regions referring to sets of over- and underexpressed meta-genes, respectively. Meta-genes from the same spot are co-expressed in the experimental series whereas different, well-separated overexpression spots in the same portrait refer to meta-genes overexpressed in the particular sample but differently expressed in other samples due to different profiles. The sample specific 'local' spots in the SOM portraits consequently combine two characteristics: meta-gene co-regulation and differential expression. Contrary to the local spots, we define 'global' spot clusters which refer to all samples. Later we will present gene set enrichment analyses to assign biological functions to the global clusters, which can therefore be interpreted as 'functional modules' inherent in the data. Below we will also compare the SOM-based clustering approaches with alternative clustering methods applied on the single gene level such as non-negative matrix factorization (NMF, see [67–69]), hierarchical clustering (HC, see [70]) and correlated gene set clustering (CGS, [71, 72]).

### 2.7.1 Spot clusters

For an overview about all local spots observed, two types of integral overview maps are created, characterizing over- and underexpression of the meta-genes in a global view. Firstly, the *meta-gene peak maps* shown in Figure 2-16a and b accentuate the maximum and minimum expression values of the meta-gene profiles, respectively. These maps plot the meta-gene expression profiles into one common scale, representing their maximum and minimum values as color-coded tiles. They allow discrimination between subtle differences of the amplitudes of the maxima and minima considered by amplification of spots referring to local maximum/minimum values in the meta-gene expression profiles. For example, the meta-gene maxima map of human tissues (Figure 2-16a) features differently colored spots along the diagonal line which refer to maxima of different amplitude in the respective SOM portraits (e.g. the amplitude of spot C clearly exceeds that of spot B).

Alternatively, 'overlay maps' are created, which transfer spots of either over- or underexpression observed in the sample portraits into one master map. These *overexpression and underexpression spot maps* are shown in Figure 2-16c and d, respectively. Here, the respective maximum and minimum values observed in one of the samples scale equally showing, for example, equally colored spots along the diagonal line in panel c of Figure 2-16 (e.g. spots B and C are of equal amplitude). Note also that the tissue overexpression spot C decomposes into three subspots which however strongly differ in their amplitude in the original SOM portraits (compare spot C in Figure 2-16 c and a). Both types of overview maps

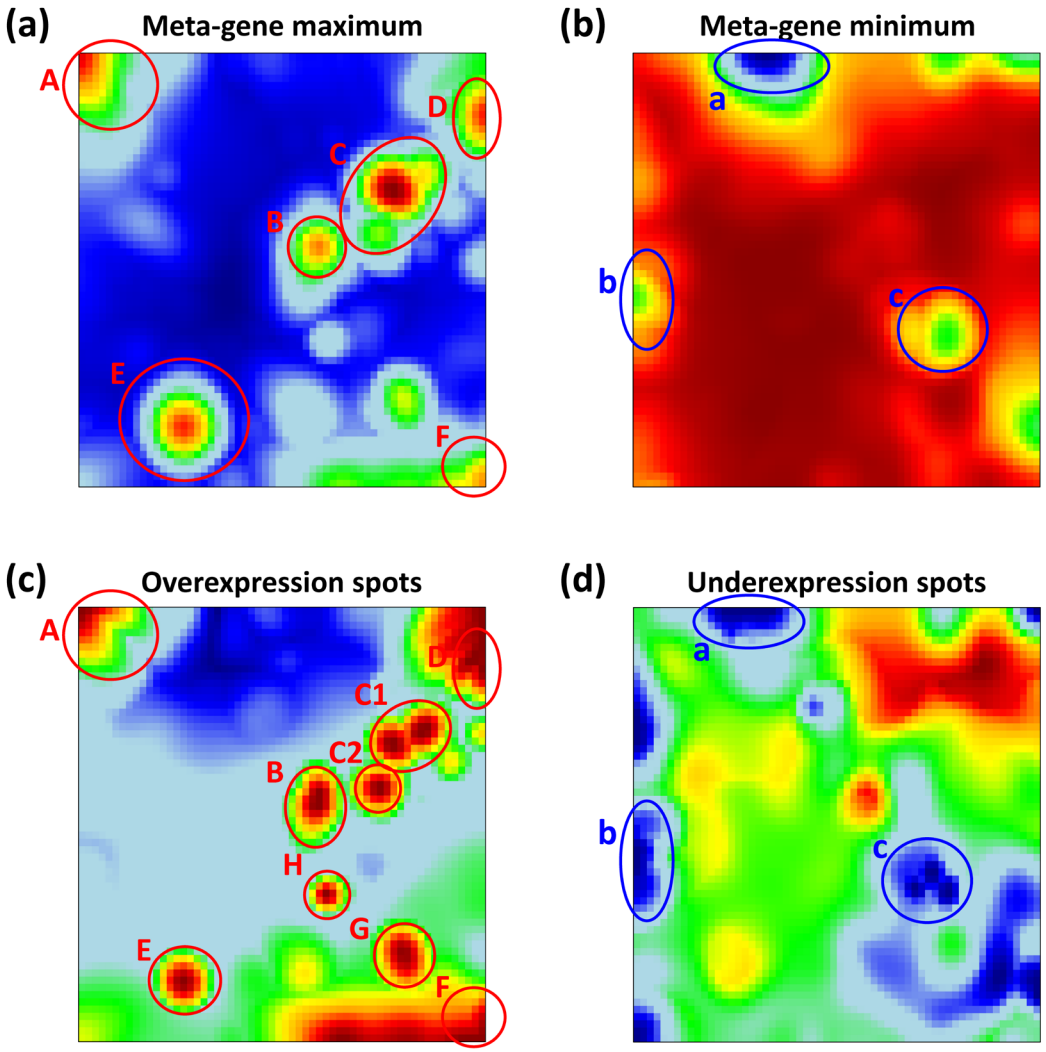


**Figure 2-16: Overview maps of human tissue transcriptome set: Meta-gene maximum (a) and minimum (b) maps and over- (c) and underexpression (d) spot maps. Red/maroon spots mark overexpression/maxima, blue ones underexpression/minima. Selected spots are marked by letters (capital and lower case letters refer to over- and underexpression, respectively). The maximum/minimum maps use a unique scaling for meta-gene expression whereas the over/underexpression maps integrate tissue-specific spots from different scales. As a consequence they show a larger number of spots than the former ones.**

thus reflect similar properties however in a complementary fashion, either with the focus on their absolute amplitude in common scale or on the identification of maxima and minima in the SOM portraits independent of their amplitude.

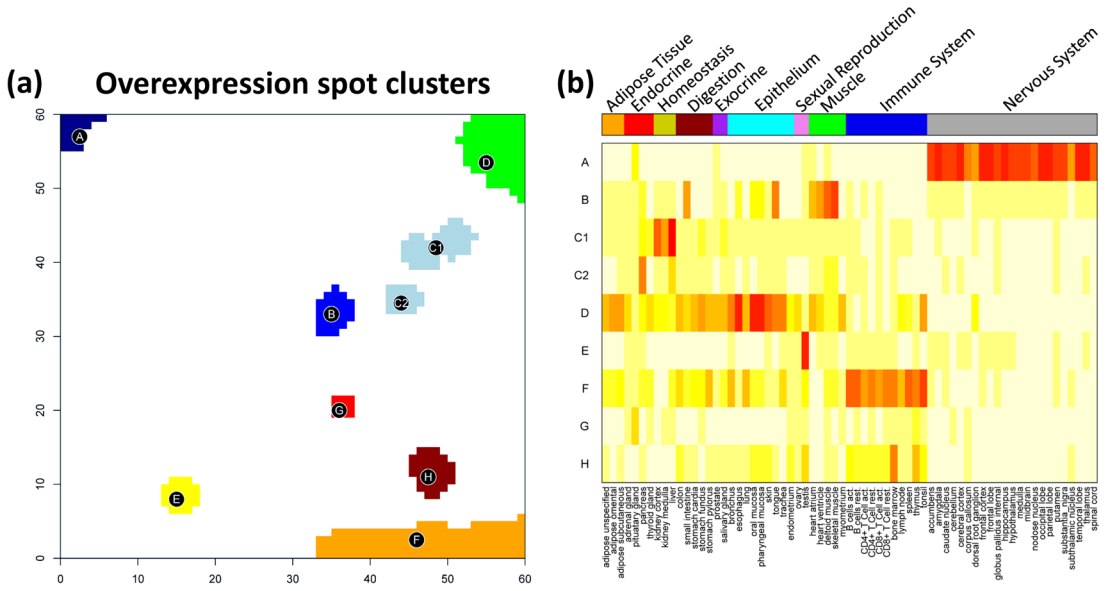


**Figure 2-17: Global overexpression spot clusters identified in the tissue SOM (panel a). The heatmap shows the corresponding expression profiles (panel b). Each cluster refers to one row. The expression scale refers to the mean meta-gene profile averaged over all meta-genes within the respective cluster. The tissues are grouped according to categories in horizontal direction (see the color bar on top of the map; the colors are assigned to the categories in agreement with Figure 2-9).**

Global over- (and also under-) expression spot clusters were defined by applying a simple 98-percentile (and 2-percentile) criterion which selects the respective fraction of the meta-genes showing largest (or smallest) expression in the sample portraits. Figure 2-17a shows the overexpression spots of the human tissue SOM. In total, nine such spot clusters were detected and labeled using capital letters. A representative expression profile was then calculated as the mean over the profiles of all meta-genes of the spot. The heatmap in Figure 2-17b shows these spot profiles in the series of tissues studied. It allows identifying specific expression patterns in each tissue category. For example, spot 'A' is specifically overexpressed in nervous system samples and spot 'B' in the muscle tissues, whereas spot 'G' is more ubiquitous lacking category specific overexpression.

In general, over- and underexpression spot clusters provide a simple and intuitive approach for definition of global meta-gene clusters. It additionally identifies the clusters in unsupervised fashion without necessity of previous definition of class prototypes or desired number of clusters. The obtained overexpression spot profiles carry prominent expression signatures inherent in the data set, which are characteristic for single tissues or tissue categories.

### 2.7.2 Correlation clusters

Alternatively, one can apply a different metric based on the mutual correlation of the meta-genes to cluster co-expressed meta-genes. Particularly, we apply the following algorithm to determine groups of correlated meta-genes :

(i) The Pearson correlation coefficients, $r_{ij}^{meta}$ (i, j= 1…K) are calculated for all pairwise combinations of meta-gene profiles.
(ii) Their maximum value $r_{IJ}=max(r_{ij}^{meta})$ defines a pair of 'source' meta-genes at positions i,j=I,J. They typically refer to neighbored tiles in the SOM.
(iii) Then, the source meta-genes serve as condensation nucleus for the associated group of correlated meta-features which comprises all meta-genes meeting the condition $min(r_{I,x}, r_{J,x}) > r_{threshold}$ where the threshold value for the correlation cluster is typically set to $r_{threshold}= 0.90$.
(iv) The meta-genes of this group were excluded for next iteration which starts again with step (ii) to determine the next group of correlated meta-genes by processing the remaining ones.

Steps (ii) – (iv) are repeated until all meta-genes are clustered into groups of at minimum one member. In total 132 of such highly correlated clusters were identified in the tissue data set. The ten clusters of strongest correlation were then chosen in accordance with the number of overexpression spots discussed in the previous subsection. The correlation map in Figure 2-18a shows the obtained correlation clusters as color-coded regions in the SOM-mosaic. The heatmap in Figure 2-18b illustrates the mean expression profiles of the clusters. Please note that also clusters without pronounced differential expression were selected by this algorithm, for example clusters 'A', 'E' or 'I'. Also very similar profiles are observed, showing specific overexpression for one tissue category: 'F' and 'G' for immune, or 'H' and 'J' for nervous system.

The clustering of correlated meta-genes represents a global approach complementary to the spot clusters. It groups the meta-genes according to most similar expression profiles independent of strong differential expression. The obtained groups form disjunct clusters in the respective correlation cluster map. The clusters of largest mutual correlations are mostly located in the region of largest meta-gene variance (compare to Figure 2-14b). Hence, SOM mapping based on Euclidean distance in the training provides also a characteristic pattern with respect to the correlation metrics.

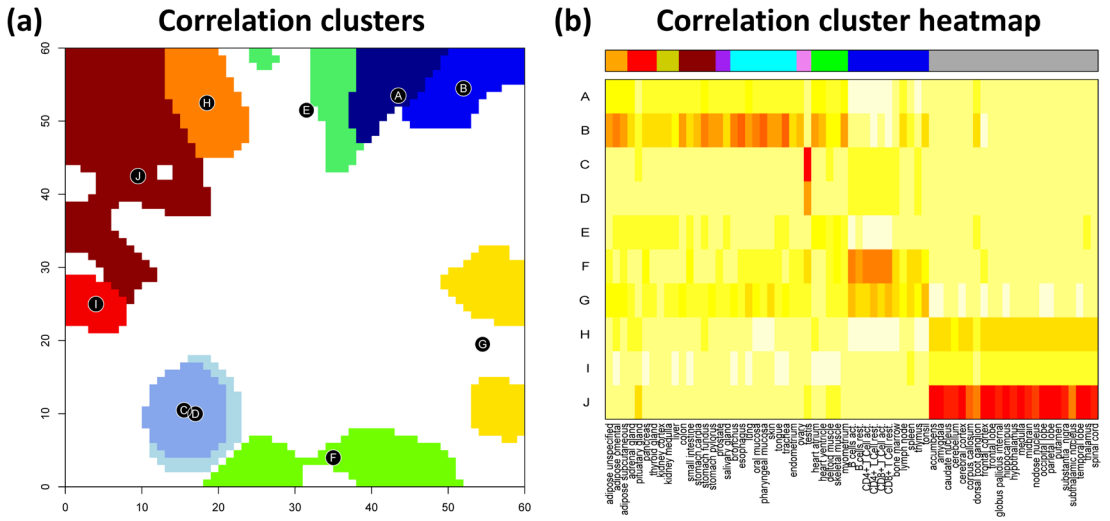


**Figure 2-18: Correlation module map of human tissue SOM (panel a): Each colored area represents a group of meta-genes which strongly correlate with each other with a correlation coefficient of r> $r_{threshold}$=0.90. The ten clusters of strongest intra-correlation were chosen as expression modules. The module expression heatmap shows the corresponding spot expression profiles (panel b). See legend of Figure 2-17 for details.**

### 2.7.3 K-means clusters

As a third option, unsupervised k-means clustering was applied to select groups of co-expressed meta-genes [73]. This approved method divides the profiles of all meta-genes into a predefined number of clusters using the Euclidean distance as similarity criterion. Ten k-means clusters were generated in agreement with the spot and correlation clustering methods used before. Figure 2-19a shows the obtained cluster map. It segments the SOM into ten disjoint regions. Note that no additional meta-gene filtering was applied. Hence the k-means clusters cover the entire map in contrast to the overexpression spot and correlation clusters which cover only part of available meta-genes (compare with Figure 2-17a and Figure 2-18a). Figure 2-19b shows the corresponding mean expression heatmap of the k-means clusters. The expression profiles partly agree with that of the overexpression spots (compare Figure 2-17b and Figure 2-19b), however the k-means clusters are less specific (e.g. ubiquitous clusters 'A' and 'B') and partly redundant (e.g. clusters 'E' and 'F', or 'G' to 'J', respectively).

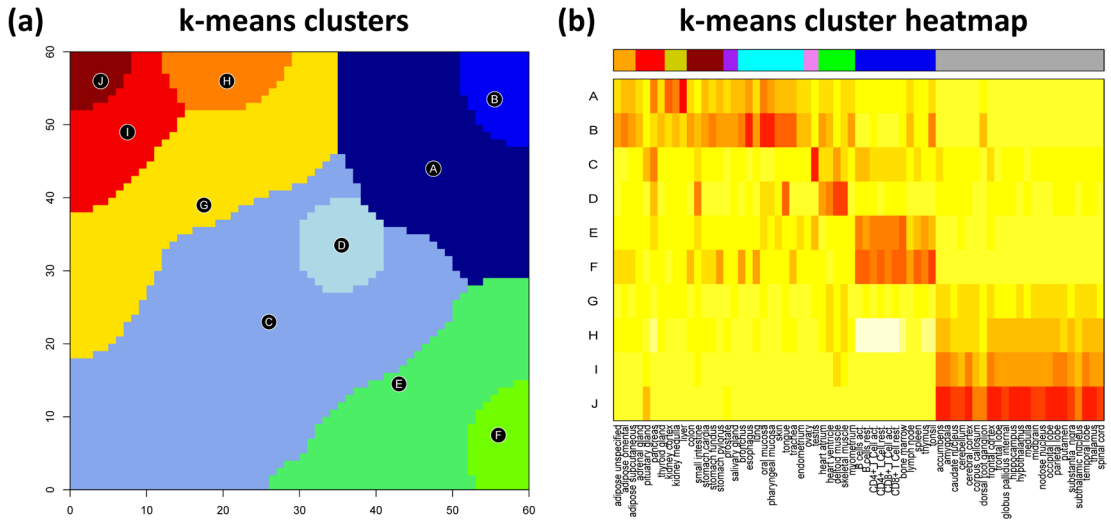


**Figure 2-19: k-means cluster map of human tissue SOM (panel a) and module expression heatmap (panel b). See legend of Figure 2-17 for details.**

### 2.7.4 Alternative methods of gene clustering

One essential feature of the SOM approach is the reduction of dimensionality of the full data set from ten thousands of single gene expression profiles to a few thousand meta-gene profiles. In a second reduction step, the dimensionality is further reduced to a handful of expression modules of co-expressed meta-genes. For comparison with the SOM spot-clusters selected alternative methods of dimension reduction were applied: non-negative matrix factorization (NMF, see [67–69]), hierarchical clustering (HC, see [70]) and correlated gene set clustering (CGS, [71, 72]). These competing clustering methods use different approaches: NMF decomposes each of the expression profiles in original space into an additive set of meta-gene profiles with non-negative expression amplitudes. HC is a heuristic iterative algorithm that separates the original data into hierarchically structured clusters using typically Euclidean distance metrics. CGS uses correlation metrics in combination with stringent significance testing to group the original data into groups of correlated single genes. For NMF- and HC-clustering the R-packages 'NMF' [74] and the basic package 'stats' [49] were used, respectively. CGS-clusters were obtained using an in-house R-program [72].

The number of clusters was set to ten in correspondence with the number of overexpression spots detected in the SOM of human tissues. Figure 2-20 illustrates the distribution of the genes of the five leading clusters in the SOM map for direct comparison with the spot clusters. NMF generates relatively diffuse clusters which spread over wider areas of the SOM. The top two HC clusters also show diffuse patterns whereas the remaining ones localize in relatively small areas of the map. Finally, CGS also generates localized but partly redundant clusters: Three out of five of them occupy the top left corner of the map which is assigned to genes overexpressed in nervous tissues.

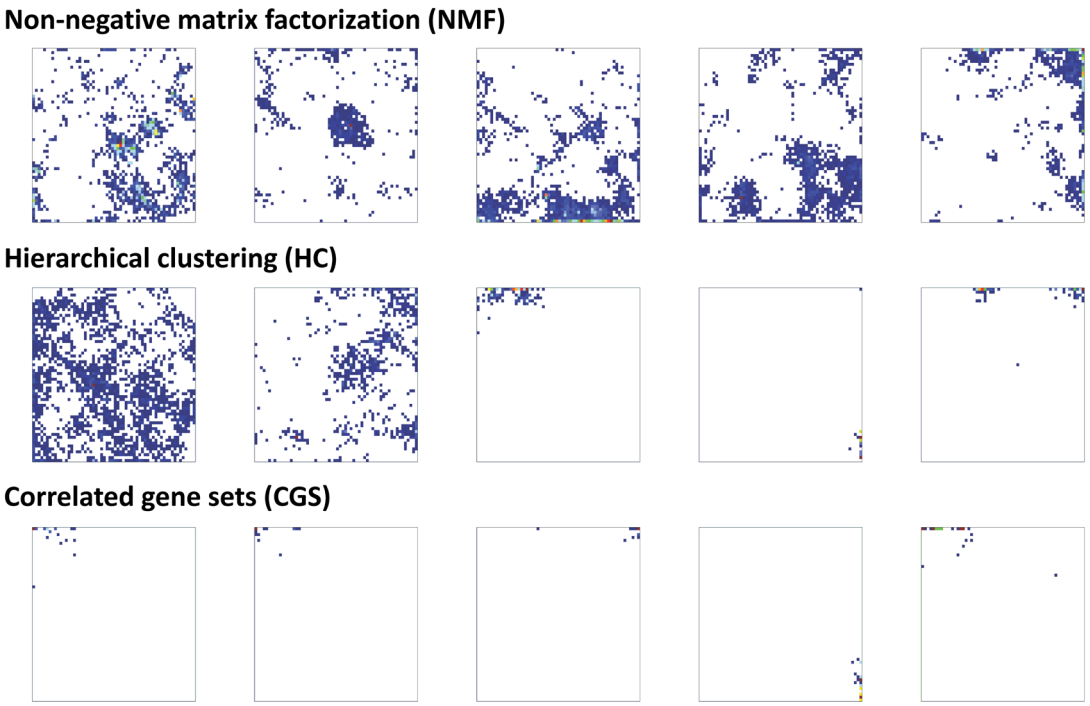


**Figure 2-20: Cluster-specific population maps of the five leading clusters obtained by alternative methods. SOM meta-genes occupied by single genes from the respective clusters are marked by dots colored in blue for few to red for many single genes.**

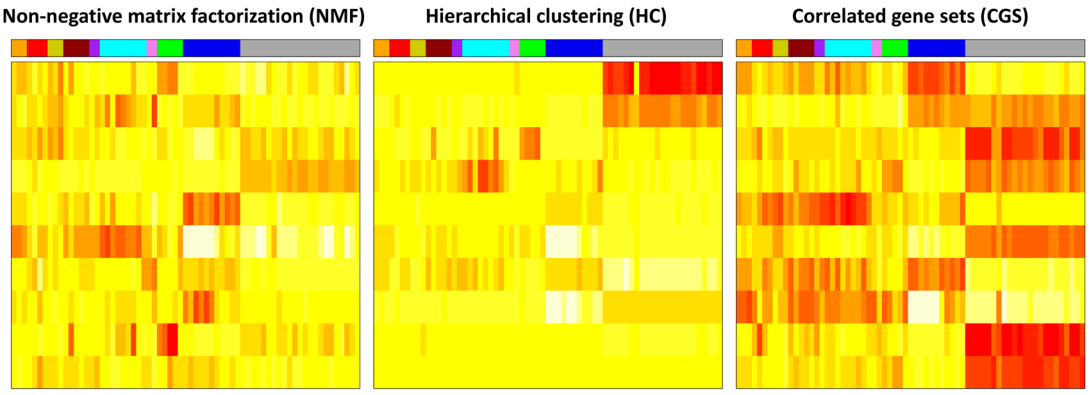


**Figure 2-21: Meta-gene profile heatmaps of ten expression modules obtained by the alternative methods. See legend of Figure 2-17 for details.**

The heatmaps in Figure 2-21 further confirm this observation: The genes which are specifically overexpressed in nervous tissues are captured by at minimum five of the ten CGS-clusters, HC generates two to three of such 'nervous system'-clusters. The SOM contrary provides only one spot which collects virtually all genes overexpressed in nervous tissues. In contrast, the NMF-clusters are clearly not redundant but, on the other hand, most of them are overexpressed in diverse tissue categories and thus unspecific for these tissue groups.

Particularly, NMF decomposes the gene expression patterns as an additive combination of the NMF modules whereas SOM, HC and CGS use a decomposition that insists mutual exclusion of features. In other words, NMF-meta-genes are less specific for single tissues and tissue categories per definition since they imply an alternative context dependency.

### 2.7.5 Benchmarking the clustering methods

It was demonstrated that the global expression landscape of human tissues is characterized by about nine- to - ten overexpression spots (see Figure 2-17a) in the SOM portraits. Additionally, meta-gene correlation and k-means clusters were generated. These SOM clusters are to be compared to the clusters obtained using NMF, HC and CGS dimension reduction with regard to their ability to generate tissue-specific clusters. It is estimated using the entropy [75],

$$\mathscr{H}_m = \frac{-1}{log_2(C)} \sum_{c=1}^{C} \rho_{c,m} \cdot log_2(\rho_{c,m}) \quad \text{with} \quad \rho_{c,m} = e_{c,m} / \sum_{C} e_{c,m} \tag{15}$$

where $e_{c,m}$ is the logarithmic expression of the clusters. It is calculated as mean value over the expression values of its member meta-genes. The entropy is calculated for each tissue sample m = 1...M where the sum runs over all clusters c = 1...C. It has units of bits and ranges from zero for tissues with only one highly expressed cluster to 1 for tissues with uniformly expressed clusters.

Recall that we assumed a number of ten clusters in each of the supervised clustering methods in correspondence with the number of SOM spot clusters identified. Figure 2-22 shows that SOM overexpression spots outperform the alternative methods in terms of specificity of the obtained expression clusters. In other words, spots of overexpressed meta-genes represent the natural choice for identification of major expression modes in the data. The expression signatures obtained from SOM analysis thus feature highest specificity across all methods compared.

## 2.8 SOM analysis of randomized data

The previous subchapters relate to SOMs trained with both artificial and real world data. It was shown, that adjacent meta-genes feature similar expression profiles giving rise to clusters of co-regulated meta-genes. These clusters emerge as spot patterns in the SOM portraits and can be understood as disjunct regulatory modes of gene expression. Co-regulation is thereby often assumed to be caused by the involvement of the genes into common pathway activities according to the 'guilt-by-association'-principle [76]. Alternatively, genes can be ostensibly co-regulated also by chance, for example, in an ensemble of genes with random expression profiles. The probability to find such random 'co-regulation' patterns depends on the number

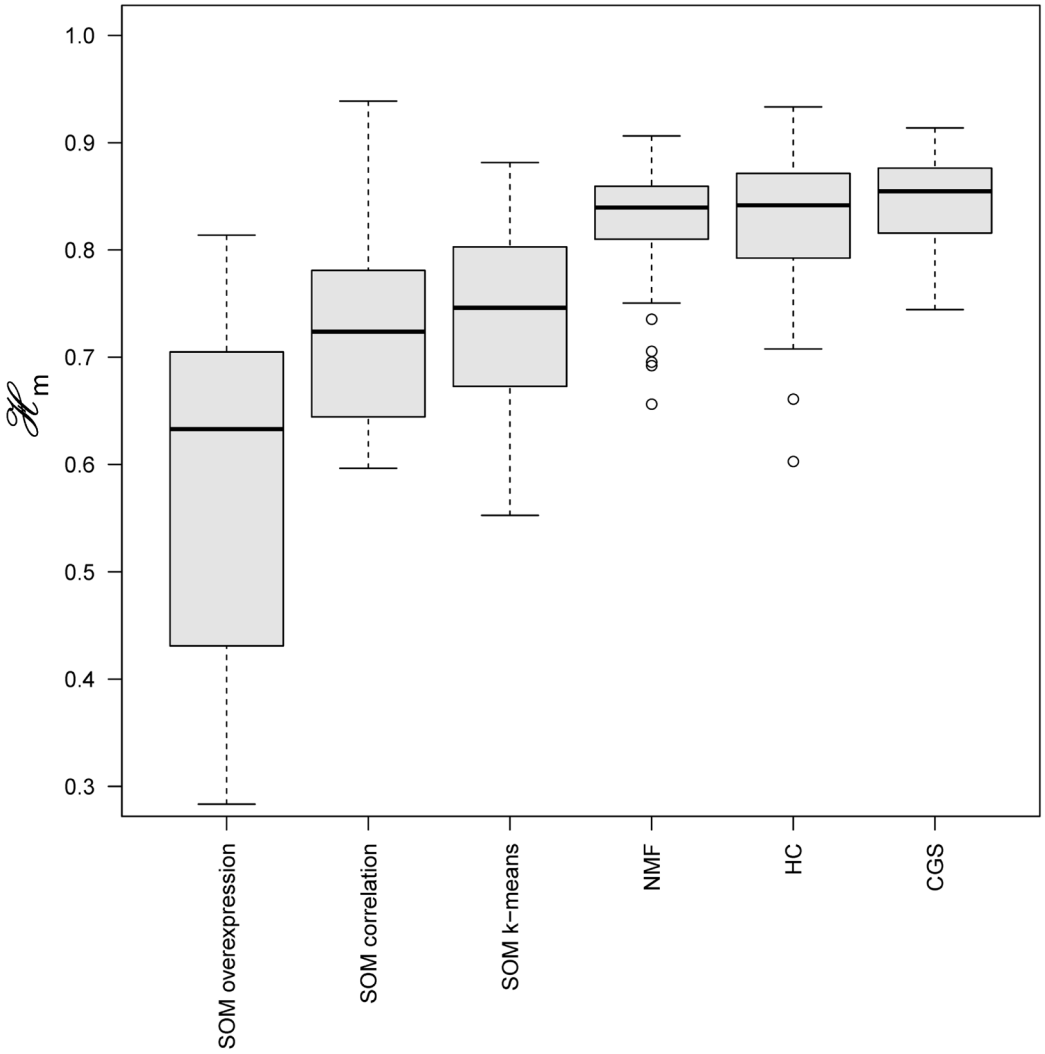


**Figure 2-22: Expression module specificity comparing different methods. The specificity is measured in terms of the entropy (Eq. (15)): Small values refer to tissues which are specifically characterized by only one module of high expression whereas large entropy values refer to tissues with more uniform expression of the module clusters. The boxplot illustrates the distribution of the entropy values for all tissues considered in each method.**

of different conditions studied and on the resolution of the cluster algorithm used. The effect of random expression is studied for the human tissue data simply by permuting the expression values of each gene randomly among the samples. This way the tissue-specificity of each expression profile is virtually destroyed. Then, the randomized data was used to train a SOM utilizing the same SOM-size and grid-topology as used for the unperturbed SOM of human tissues. Finally, both SOMs were compared with regard to the spot clusters and meta-gene characteristics (Table 4).

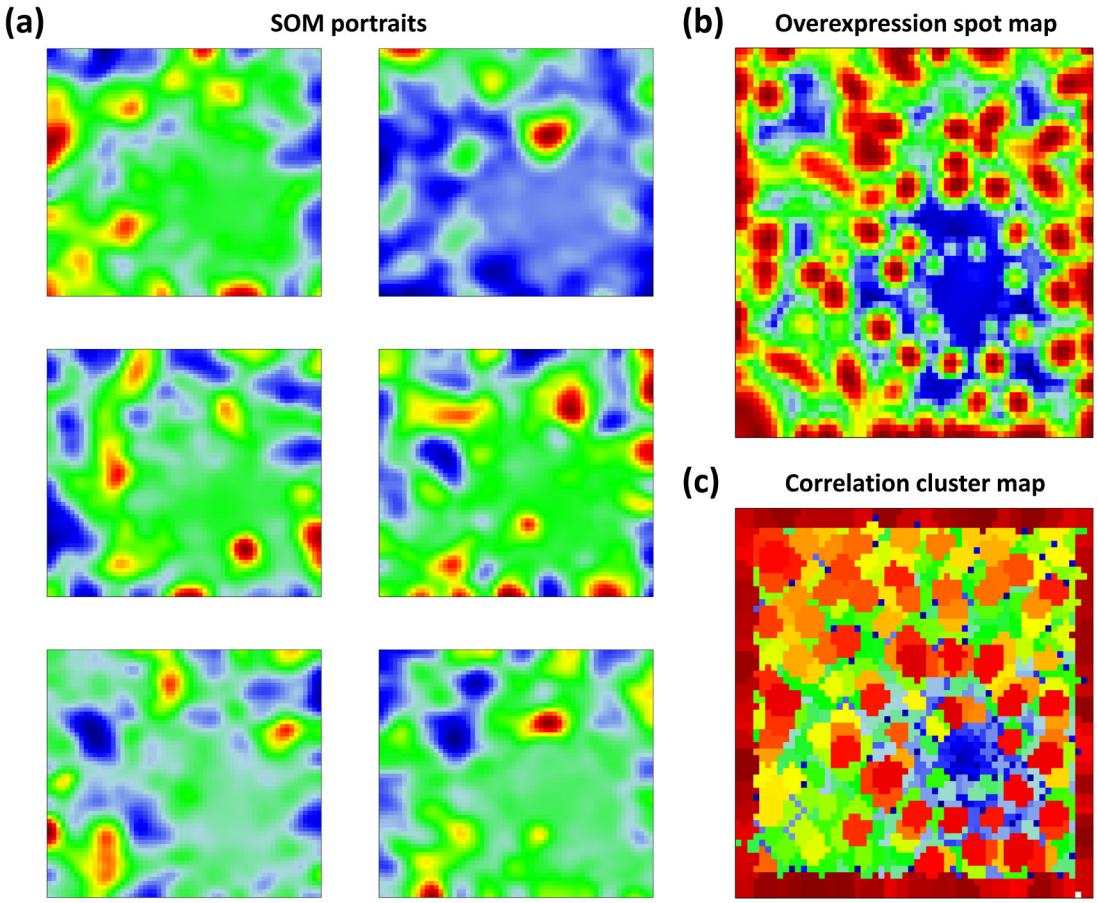


**Figure 2-23: Selected SOM portraits (panel a), overexpression cluster map (b) and correlation cluster map (c) of the 'random SOM' reveal dense arrangement of multitude of random pattern, resulting in large number of over- and underexpressed spots and correlated clusters.**

Figure 2-23a shows six selected sample portraits of the 'random SOM'. They exhibit very diverse pattern with a clearly larger number of over- and underexpression spots compared to the original tissue SOM portraits (Figure 2-23b). On average, the number of overexpression spots increases approximately threefold after randomization (see Table 4).

The observed number of spots in the 'random SOM' monotonously increases with increasing SOM-size whereas that of the 'tissue SOM' levels off to around 10 already for small SOMs (Figure 2-24). In other words, the 'real' expression landscape of human tissues is considerably less fragmented than the respective random one. Hence, the random landscape is characterized by more and SOM-size dependent expression modes without mutual correlations. These are only partly captured by the particular SOM-size used. In consequence, the increase of the SOM-size gives rise to an increasing number of spots. In contrast, the number of expression modes of the 'tissue SOM' asymptotically attains a stable level.

**Table 4: Comparison of the 'tissue SOM' the 'random SOM'.**

| | Tissue SOM | Random SOM |
|---|---|---|
| *#overexpressed spots* [a] | 1.4 | 3.2 |
| *#correlation clusters* [b] | 121 | 549 |
| *Population* [c]: $n_k$ | 4±10; max=308 | 5±6; max=306 |
| *Variance* [d]: $var_k$ | 0.01±0.05; max=0.57 | 0.005±0.003; max=0.01 |
| *Covariance* [d]: $r_k$ | 0.61±0.14; max=0.94 | 0.43±0.05; max=0.70 |
| *Deviation* [d]: $d_k$ | 0.15±0.10; max=0.59 | 0.18±0.07; max=0.53 |
| *Significance* [e]: $<p_k>$ | 0.26±0.09; min=0.02 | 0.47±0.05; min=0.28 |

[a] mean number of overexpression spots per sample portrait (>98% threshold)
[b] number of correlation clusters using the seed algorithm
[c] median number of genes per meta-gene± standard deviation and the maximum occupancy observed
[d] mean, standard deviation and maximum of meta-gene variance, meta-gene - gene covariance and metagene - gene Euclidean distance (deviation) of all meta-gene profiles
[e] mean, standard deviation and minimum of meta-gene significance of all meta-gene profiles

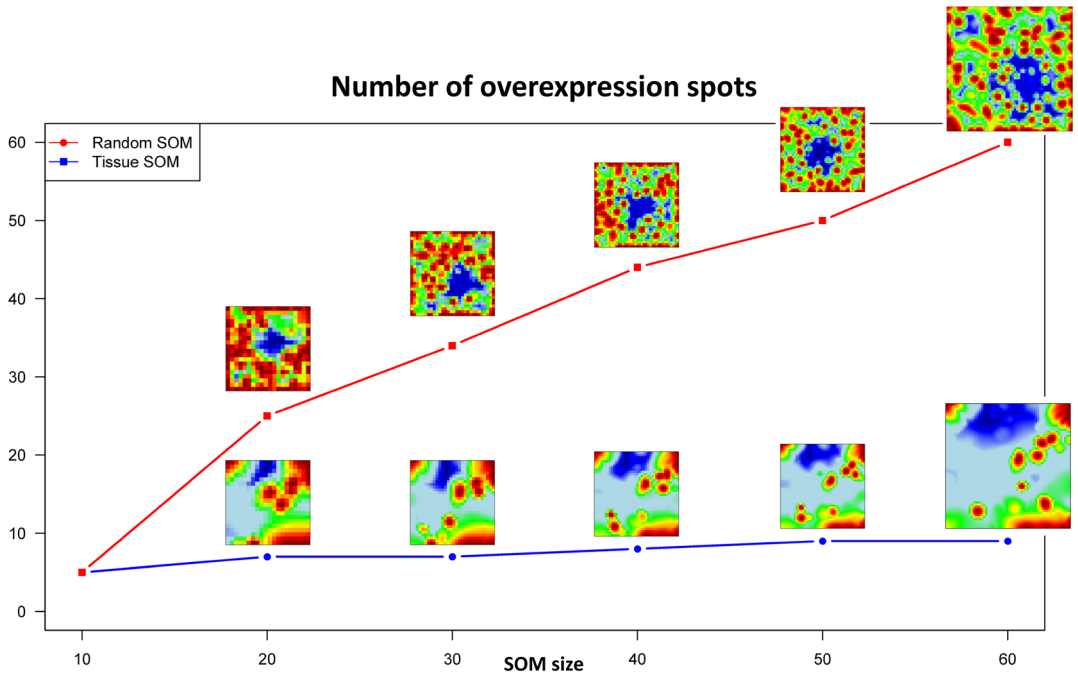


**Figure 2-24: 'Tissue SOM' vs. 'random SOM': Total number of overexpression spots as a function of the SOM-size observed in the SOM portraits of human tissues before (blue curve) and after (red) randomization. The respective overexpression summary maps are shown for SOM-sizes 20x20 to 60x60.**

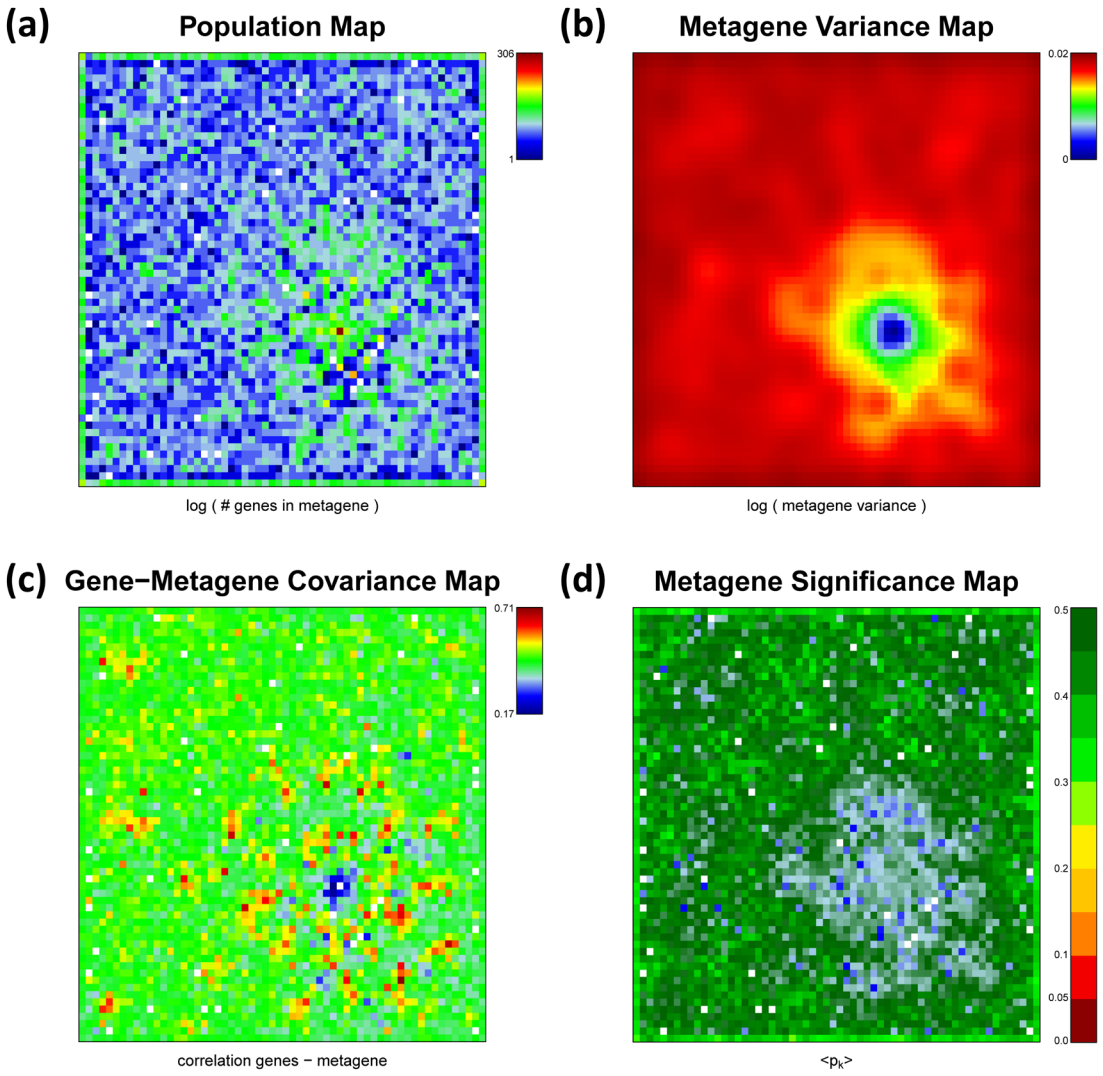


**Figure 2-25: Supporting maps characterizing the 'random SOM': Population (panel a), variance (b), covariance (c) and significance (d) maps.**

The supporting maps of the 'random SOM' allow identification of further properties induced by randomized input data. The population map (Figure 2-25a) reveals clearly more uniform and less structured distribution of the single genes across the meta-genes compared to the 'tissue SOM'. Also the variance map (Figure 2-25b) shows a specific structure: The invariant meta-genes collect within a delimited region of the SOM. They are surrounded by meta-genes of almost constant variance. This homogeneity is caused by the strong overlap of the individual over- and underexpression spots. The covariance map resembles this structure (Figure 2-25c): The invariant meta-genes are characterized by very low covariance to the mapped genes. The remaining meta-genes however feature constantly high covariance values. The decrease of maximum variability of the meta-genes reflect a stronger 'smoothing' effect of the meta-gene profiles due to less concerted single gene profiles in each of the meta-gene

clusters. This effect becomes also obvious in the smaller covariance and increased deviation between the meta-gene and single gene profiles in each of the clusters.
The strongest difference between the ‘tissue SOM’ and the ‘random SOM’ is illustrated by the significance map (compare Figure 2-25d and Figure 2-14f): where the ‘tissue SOM’ shows spot-like regions of significant meta-genes (e.g. $\langle p_k \rangle < 0.05$), the ‘random SOM’ lacks of any significant meta-genes.

In summary, the ‘random SOM’ is characterized by more uniformly populated meta-genes of poor significances and weak concordance to the mapped single genes. Therefore it is clearly possible to distinguish between a SOM trained with structured real world data and a SOM trained with randomized ‘noise’.

# 3 Filtering data using SOM

The use of meta-gene instead of single gene expression data reduces the dimension of the data and potentially leads to an increased discriminating power in downstream analyses. In particular, meta-gene filtering is expected to outperform single gene filtering regarding representativeness and noisiness because the reduced number of meta-genes not only preserves the diversity of single gene profiles but also reduces noise in the expression profiles. In this chapter we analyze the capability of the SOM approach for data filtering and dimension reduction in terms of maintaining representativeness and reducing noisiness of the input data. Additionally, downstream analyses based on either single gene or meta-gene level are compared to verify the benefit of SOM dimension reduction.

## 3.1 Comparing meta-gene and single gene based filtering

The reduction of the size of a data set by removing genes that carry essentially no or low information is common practice with the intention to improve downstream analysis such as two-way hierarchical clustering of genes and samples. Such data reduction has been shown to result in cluster dendrograms which more accurately reflect relationships between the samples with increasing stringency of the filter applied [77]. This improvement can be attributed to the fact that random noise tends to disrupt similarity relations between genes and samples. On the other hand, also systematic errors within the data, e.g. due to batch effects, can cause artificial cluster relations if the bias affects subsets of genes in a concerted fashion. Hence, a favored filter ensures improvement of the data by removing either noisy, biased and/or weakly expressed genes. Nevertheless, extreme filtering is dangerous because it may eliminate valuable information, for example genes of relatively low and thus noisy expression but with important biological impact. Filtering hence is an optimization task with the claim to remove virtually irrelevant data while preserving all information which is important in the context of the particular issue studied. The former property will be further on called ‘noisiness’ of a filter and the latter one ‘representativeness’. Filter optimization thus aims at maximizing representativeness while minimizing noisiness.

SOM analysis facilitates alternative filtering based on the meta-genes as representatives characterizing the expression profiles of clusters of single genes. In other words, the meta-gene profiles themselves serve as a filtered and compressed extract of the original data. In the case of the human tissue data, the SOM assigns the expression profiles of N=22,277 input genes measured to K=3,600 meta-gene clusters. Each meta-gene therefore comprises $N/K=\langle n_k \rangle=6.2$ real genes on the average. Hence, complexity of transcriptome characterization is reduced to about one sixth by utilizing the meta-genes instead of the real genes.

In fact, the local N/K-ratio considerably varies between the different meta-genes with minimum and maximum values of $n_k=0$ (empty meta-genes) and $n_k=308$ as illustrated by the

population map in Figure 2-14a. In consequence, the importance of transcriptome information is effectively reweighted by using meta-genes instead of real genes. For example, the meta-gene of highest population ($n_k$= 308) collects genes of virtually invariant expression profiles. These essentially non-informative features comprise 1.4% (308 out of 22,277) of all single genes but only 0.3% (1 out of 3,600) of all meta-genes. Hence, their contribution is effectively down-scaled by a factor of ~1/5 when using meta-genes instead of real genes. In other words, the SOM algorithm itself embodies a selective compression filter, reducing the number of features by condensing similar single gene profiles into respective meta-gene profiles.

To show characteristics and effects of filtering, top-list selection filters are applied either to the meta-genes or to the single genes. In a first approach, fold change (FC) filtering is used to reduce the number of single genes and meta-genes. Here, the full list of absolute FC-values of all genes ($\Delta e_{g,m}$) respectively all meta-genes ($\Delta e_{k,m}^{meta}$) is ranked and a certain number (e.g. 100, 1,000 and 3,600) of topmost features is selected. Note that lists of equal numbers of meta-genes and of single genes are asymmetric owing to data compression in the meta-gene clusters. Meta-gene lists integrate information of roughly a tenfold larger number of 'real' genes in the example studied. Figure 3-1 compares the areas in the SOM mosaic preserved by FC-lists of different lengths if applied to either meta-genes or single genes. The shorter meta-gene lists cover essentially the same regions of the SOM as the longer single gene lists with considerable overlap of the selected meta- and single genes. The large overlap demonstrates that the meta-gene filter is representative for the associated single genes which are mainly also selected if applying single gene filtering using an approximately ten-times longer list.
Figure 3-1b illustrates that different spot areas are progressively excluded from the list of filtered features with increasing stringency of the filter as expected. For example, the most stringent FC-100 meta-gene filter excludes a few areas selected by the FC-1000 single gene filtering revealing a decreased representativeness. Importantly, the covered SOM regions of gene and meta-gene lists are approximately balanced when using gene lists which are approximately one order of magnitude longer than the respective meta-gene list.
In addition to FC-filtering variance and significance (FDR) filtering were applied which select profiles of largest variance and of highest significance of differential expression, respectively (see [WIRTH1] for details).

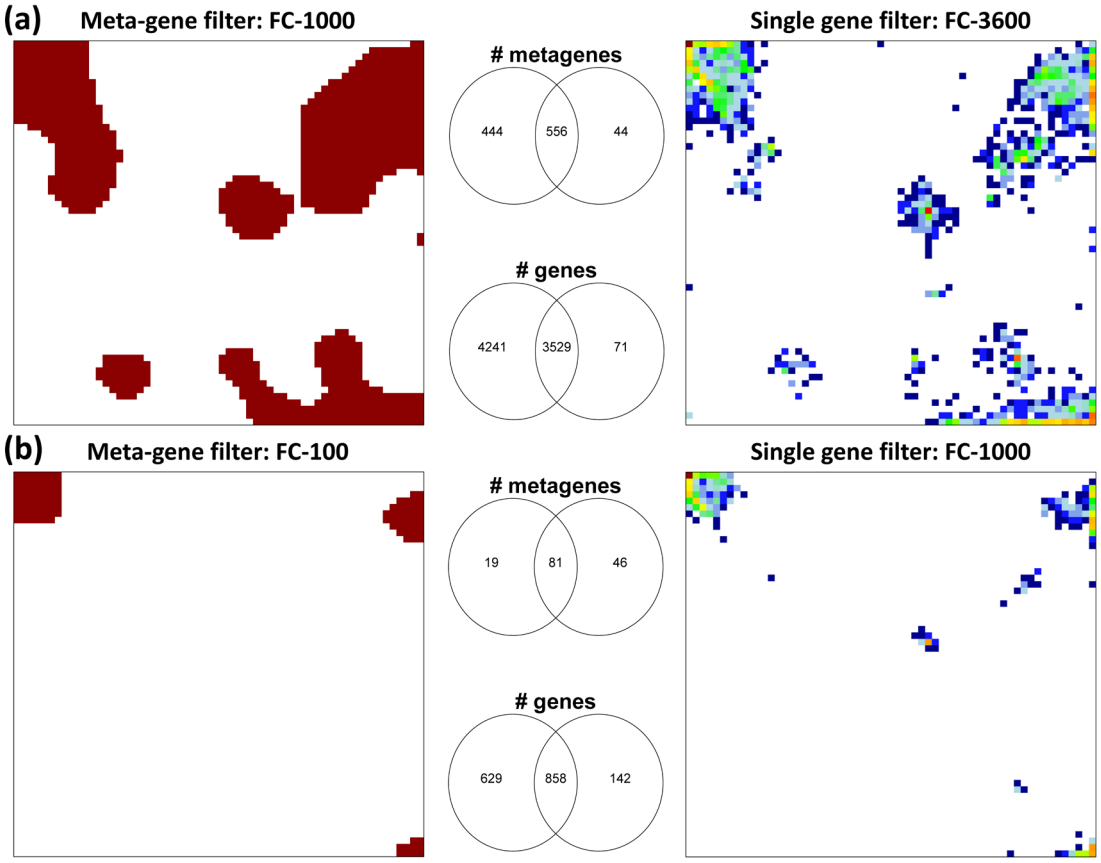


**Figure 3-1: Filtering meta-genes and genes by differential expression: Different numbers of meta-genes (left panels) and single genes (right panels) are selected using the FC-1000/FC-3600 (a) and FC-100/FC-1000 (b) filters to account for the data compression in the meta-gene clusters. The brown areas in the left part show the selected meta-genes and the colored tiles in the right part the number of single genes in the meta-gene clusters analogous to population map in Figure 2-14a. The Venn-diagrams illustrate the degree of overlap between the meta-genes and single genes selected by both filters.**

## 3.2 Meta-gene and single gene based clustering

Hierarchical cluster analysis was applied because this method is often routinely run as a first step of data summarization in microarray data analysis [70]. One way hierarchical cluster trees obtained from single gene and meta-gene FC-lists of length 3600, 1000 and 100 reflect similar properties showing that clustering is relatively robust with respect to the chosen conditions (Figure 3-2a and b). Tissues from categories with homogenous SOM portraits, such as nervous system (grey labels), adipose tissues (orange) and immune system (blue, see also portrait gallery in Figure 2-9), robustly cluster at very low levels of Euclidean distance in the respective branches. Note that the blue cluster of immune system tissues however partly decomposes if using the shortest single gene list (FC-100) owing to the loss of representativeness. On the other hand, the FC-100 meta-gene list of equal length still produces a compact blue cluster reflecting the improved representativeness of the same number of meta-genes. In the case of lowest stringency, i.e. FC-3600 lists, the blue immune

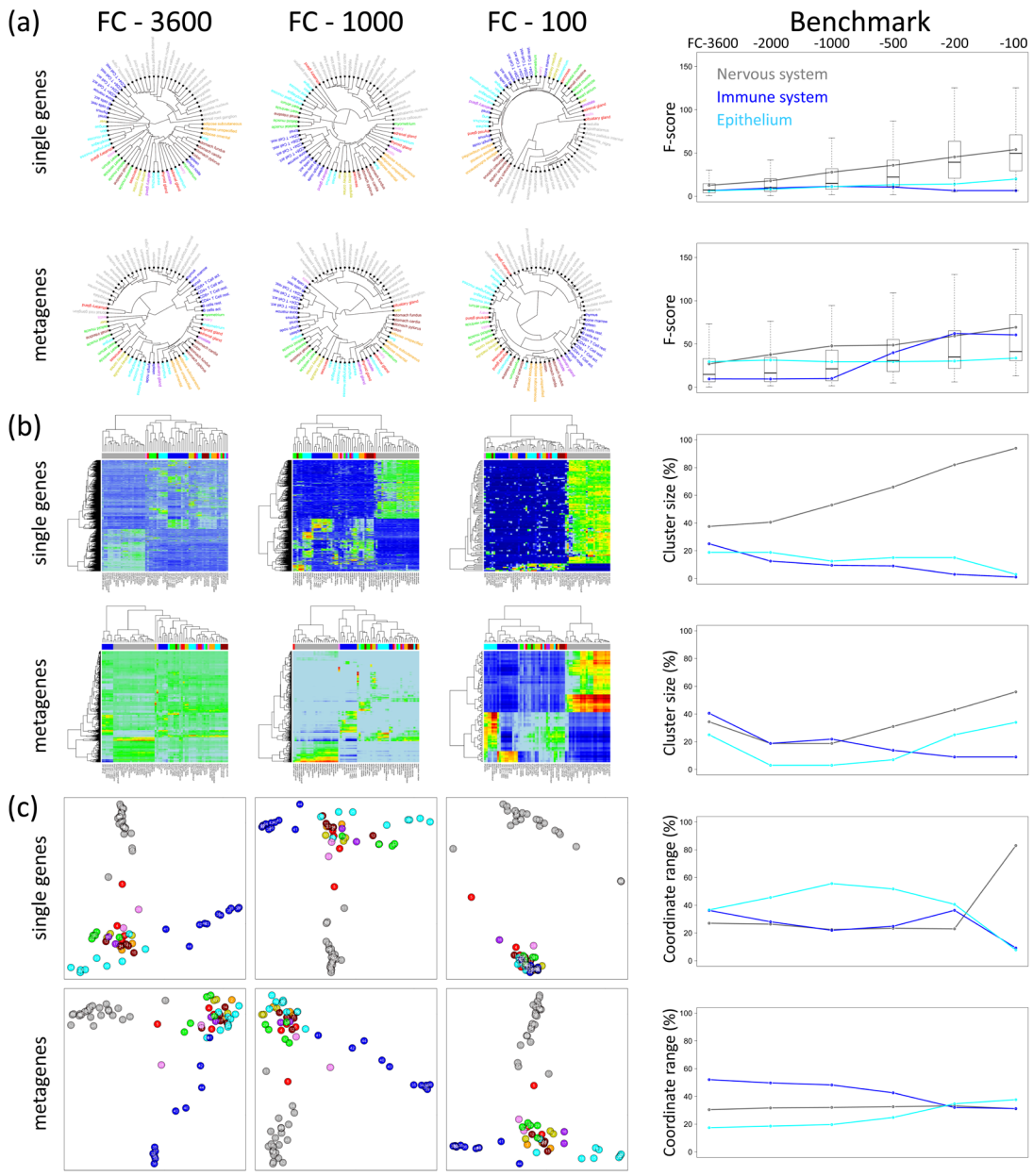


**Figure 3-2: The effect of filtering of single genes and meta-genes on the results of one-way hierarchical clustering trees (part a), two-way hierarchical cluster heatmaps (part b) and independent component analysis (part c) of the 67 human tissues studied. The samples are color-coded according to the classification of tissues introduced in Figure 2-9. Top-list FC filters select the 3600, 1000 and 100 (from left to right) most strongly differentially expressed genes/meta-genes in all samples. Note that the ICA-plots are invariant with respect to mirror and rotational symmetry operations. The right part shows different benchmark criteria for different lengths of the FC-lists ranging from FC-3600 to FC-100 (see top axis). The benchmark criteria were applied to nervous system, immune system and epithelium tissues.**

system cluster splits for both, the single gene and meta-gene filters. Obviously these lists became too long with worse characteristics regarding noisiness: Longer single gene lists reduce the quality of the observed cluster structure due to the progressive inclusion of noisy

genes. Meta-gene lists are contrary more representative and less noisy than single gene lists of equal length in downstream cluster analysis. On the other hand, the length of meta-gene lists is optimal in the intermediate range (e.g., the FC-1000 list in this example): shorter and longer lists are suboptimal in terms of representativeness and noisiness, respectively.

The cluster trees based on single gene and meta-gene lists reveal another important difference (compare the first and second rows in Figure 3-2a): The mean length of the outmost branches is considerably shorter for the meta-gene based trees than for the single gene ones. For the innermost branches, this relation inverts. This systematic difference reflects more compact clusters owing to the decreased noisiness of the meta-gene data: The mean length of the 'outer' branches estimates the mean relative distance between the most similar samples on the lowest level of clustering whereas the mean length of the 'inner' branches estimates the mean mutual distance between the largest clusters. Outer and inner branches are markedly shorter respectively longer for meta-gene cluster trees than for single gene trees. The observed meta-gene clusters are thus more compact in terms of high similarity within and high difference between the clusters.
In the right part of Figure 3-2a the inter-to-intra cluster ratio of the Euclidean distances between the samples (F-score) is shown for the three most prominent tissue categories as a simple measure of the compactness of their clusters. The F-score of the meta-genes systematically exceeds that of the single genes.

Figure 3-2b shows two-way hierarchical cluster heatmaps for meta-gene and single gene FC-filter lists. This representation visualizes similarity relations between the samples in horizontal direction (colored bars indicate tissue categories) and between the filtered (meta-)genes in vertical direction. Clearly observable, the contrast of the heatmaps increases with shorter lists (i.e. from left to right) because more stringent filters certainly select features with strongest over- (red) and underexpression (blue). The heatmaps provide detailed information about the amount of features differentially expressed in the various tissues. This cluster size is explicitly shown in the diagrams in the right part of Figure 3-2b. For example, the percentage of single genes which are overexpressed in nervous system samples and underexpressed in the other tissue categories (see also the green/red areas associated with the grey bars on top of the heatmaps) increases from less than 50% (FC-3600) to a dominating amount of more than 90% (FC-100) whereas the percentage of genes overexpressed in other tissue categories vanishes almost completely. The use of meta-genes instead of single genes effectively re-weights the contribution of tissue-specific genes. Particularly, the percentage of meta-genes which are specific for nervous tissues is markedly smaller in the meta-gene list giving rise to a more balanced distribution of features and enhanced resolution of non-nervous tissue samples.

## 3.3 Meta-gene and single gene based independent component analysis

Hierarchical clustering does not represent the multivariate structure of the data. Such aspects are emphasized by projection of the data onto subspaces of lower dimension spanned by e.g. components of minimum mutual statistical dependence. Independent component analysis (ICA) provides a visual plot in the space spanned by these independent components which are shown to point along the directions of maximum information content in the data [78]. ICA is applied to single gene and meta-gene lists to compare separation among the various tissue groups for these competing data sets.

The ICA-plots of the two leading independent components shown in Figure 3-2c reveal the degree of similarity between the samples as a function of the selected filters. With exception to the stringent FC-100 single gene list, all filters provide three major clusters, nervous (grey circles) and immune system (blue), and the remaining tissues. The FC-100 single gene filter merges the latter two clusters due to its small representativeness with respect to non-nervous tissues (see also the respective heatmap in Figure 3-2b). The relative dimension of the three clusters in the ICA-plot and thus their intrinsic resolution changes from filter to filter, reflecting the subtle interplay between the length of the list and its representativeness and/or noisiness which might overweight one tissue category and underweight another one. For example, the specifics of epithelium tissues (cyan circles) are relatively well resolved using the FC-100 meta-gene or, alternatively, the FC-1000 single gene lists. The diagrams in the right part of Figure 3-2c compare the relative size of the three major clusters in terms of the fraction of encompassed coordinate range. The meta-gene based clusters are less depending on the chosen length of the list and more balanced especially for short lists.

The ICA plots in Figure 3-2c reveal another interesting property inherent in the meta-gene expression states: The points of nervous (grey) and immune systems (blue), but also of epithelium tissues (light blue) form chain-like clusters roughly in parallel with the coordinate axes. This pattern reflects the fact that the transcriptional activity of nervous tissues on one hand and immune system and epithelium tissues on the other hand is defined by different and mutually independent groups of genes. However, this property of the data is partly lost after most stringent single gene filtering (FC-100) whereas essentially all meta-gene lists well reflect the independence of the expression pattern of the different tissue categories.

## 3.4 Meta-gene and single gene based correlation analyses

In addition to cluster and component analyses, pairwise correlation maps (PCM) are generated featuring Pearson correlation coefficients for all mutual combinations of tissue samples. The PCM-heatmaps shown in Figure 3-3a are obtained using the FC-1000 (single genes, left part) and FC-100 (meta-genes, right part) filters representing roughly the same number of genes as discussed above. The meta-genes clearly provide PCM-patterns of higher

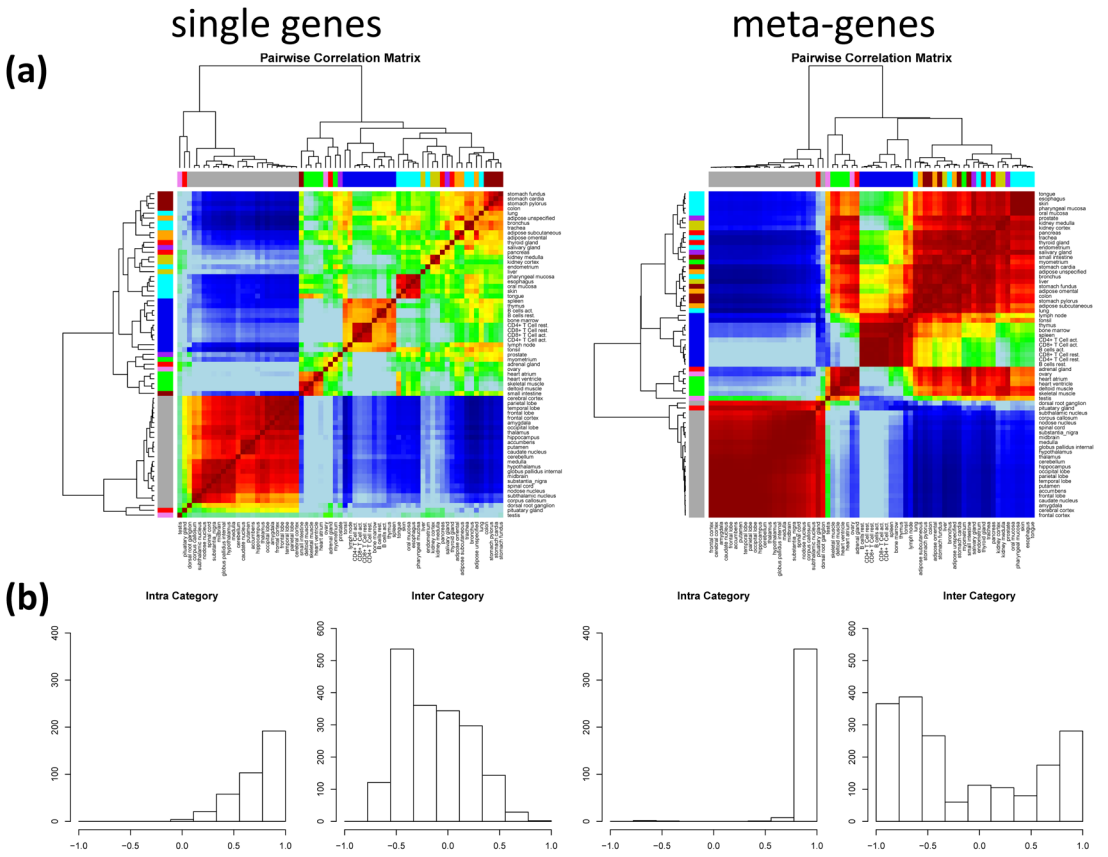


**Figure 3-3: Single gene (left panels) and meta-gene (right panels) based correlation analysis of human tissues using the 1000/100 most strongly regulated genes/meta-genes: (a) Pairwise Correlation Map (PCM); (b) Frequency distributions of correlation coefficients for all intra- and inter-tissue category pairings.**

contrast, reflected in clusters of particularly high (maroon areas) and low (blue areas) correlation coefficients along and offside the diagonal, respectively. They refer to tissue pairings with highly correlated or anti-correlated expression states. The expression states of nervous tissues for example are strongly anti-correlated with essentially all the other tissue categories, i.e. a gene overexpressed in nervous tissues is usually underexpressed in non-nervous tissues and vice versa. The SOM portraits in turn reflect this property in the characteristic spot in the top left corner (see Figure 2-9). Hence, the diagonal and off-diagonal clusters in the meta-gene PCM can be related to spots in the SOM portraits characteristic for different tissue categories.

To illustrate the origin of the contrast differences between the single gene and meta-gene PCM, frequency distributions of the correlation coefficients are shown in Figure 3-3b either for pairings between tissues of one category or between tissues of different categories. Intra-category correlation coefficients are expected to be close to unity because samples of the same categories usually feature similar expression states. Confirming this, meta-gene correlation coefficients are close to unity as expected whereas the respective single gene correlations

however show a markedly broader distribution resulting in smaller correlation values on the average. Inter-category pairings of single genes show a broad distribution centered around zero with a strong component of anti-correlation close to -0.5 reflecting that single genes of different tissue types are either not or anti-correlated. The meta-genes provide a more resolved trimodal distribution with strong components of correlated, anti-correlated and uncorrelated meta-genes near 1.0, -0.7 and 0.0, respectively. The component peaks are clearly sharper and the whole distribution covers a broader range of correlation values. Hence, the meta-genes obviously improve resolution of different subcomponents caused by different tissue types.

## 3.5 Summary

The use of meta-gene data instead of single-gene profiles enhances the discrimination power in downstream analyses such as hierarchical clustering or independent component analysis owing to essentially two facts: Firstly, the set of meta-genes better represents the diversity of expression pattern inherent in the data and secondly, it also possesses the better signal-to-noise characteristics as a comparable collection of single genes. Due to the better representativeness, meta-gene lists are less sensitive to filtering than lists of single genes. Additionally, the meta-genes represent a compression of the feature list by about one order of magnitude, without loss of information.

Single gene and meta-gene based correlation analysis confirmed this improvement in resolution power when using meta-gene expression data. The meta-gene patterns serve as an adequate data filter which appropriately selects representative features characterizing the expression properties of the system studied. Additionally, the findings of Guo et al. [3] were confirmed, who stated that SOM based meta-genes well recapitulate gene expression profiles of the entire gene dataset and capture the real similarity relationships among samples with a high fidelity.

# 4 Discovering similarities between the samples

Sample similarity analysis aims at establishing mutual relations between the phenotypes studied, e.g., to extract a hierarchy of similarities or to estimate mutual distances between the expression states. In our context, similarity analysis compares the expression meta-states as provided by the SOM algorithm. It consequently uses meta-genes instead of single genes as the basal data, which has the advantage of improving the representativeness and resolution of the results as discussed above. We apply multiple approaches additionally to the prior introduced hierarchical clustering, independent component analysis and pairwise correlation maps:

## 4.1 Second level SOM

The second level SOM analysis was proposed by Guo et al. [3] to visualize the similarity relations between the SOM portraits. This SOM maps the sample meta-states and not the genes as in first level SOM analysis. Each node of the second level SOM consequently characterizes the expression state of a representative meta-sample defined by K meta-gene expression values.

The M samples are represented using a SOM grid of size $K_{2SOM}$>M. The meta-samples serve as condensation nuclei of the associated cluster of real samples with similar SOM portraits. The mutual distances between the samples in the map are related to the degree of similarity of their expression meta-states in terms of Euclidean distance. The number of meta-samples usually exceeds the number of real samples. A considerable fraction of tiles of the second level SOM are consequently empty with no sample assigned. Figure 4-1a shows the second level SOM of the human tissue data set with a resolution of $K_{2SOM}$ =40x40=1,600 nodes.

## 4.2 Neighbor-joining tree

Phylogenetic tree reconstruction is an important tool in e.g. evolutionary biology. We apply the neighbor-joining algorithm (NJ) to represent similarity relations based on the Euclidean distances between the samples in terms of similarity trees [79]. The distances between pairs of samples in the tree refer to a common scale. In contrast to other representations, the phylogenetic tree allows to identify 'bush-like' clusters of similar samples and to estimate the degree of mutual dissimilarity between them (see Figure 4-1b).

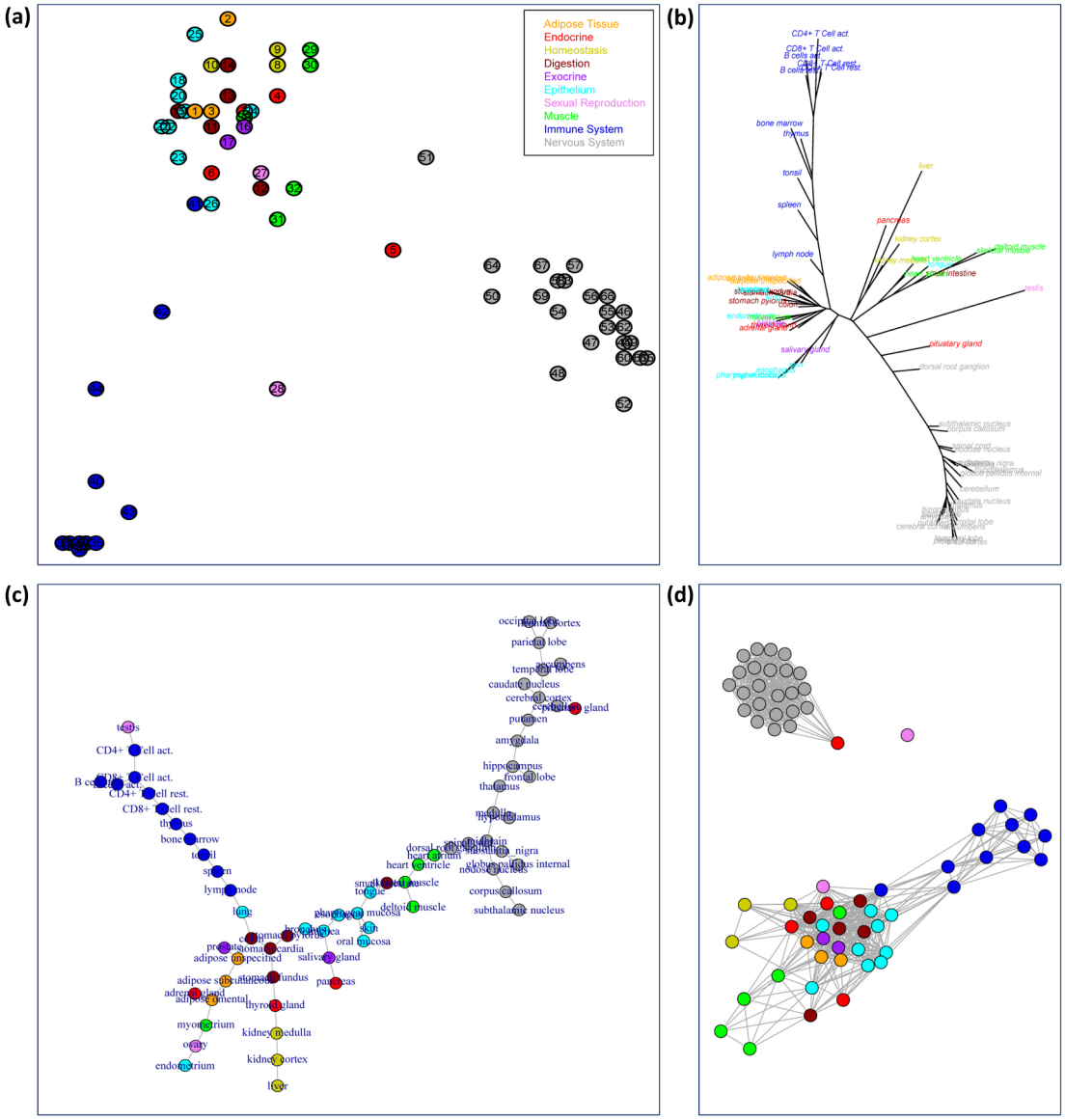


**Figure 4-1: Sample similarity analysis based on expression meta-states: Second level SOM (panel a), neighbor-joining tree (b), correlation spanning tree (c) and correlation net (d). Each tissue is colored according to its tissue category as shown in the legend in panel a.**

## 4.3 Correlation spanning tree

Contrary to the previous approaches, the correlation spanning tree (CST, Figure 4-1c) uses the pairwise sample correlation as basis. The algorithm interprets the correlation matrix as a complete graph in which the edge weights correspond to the distances (here: inverse correlation). The CST is the spanning tree that connects all vertices of that graph with the smallest sum of edge weights. It thus represents effectively the 'shortest' distance between two nodes in the graph. Spanning trees have recently been shown to be useful for clustering and classification of cancer subtypes using microarray data [80]. A major disadvantage of this

method is the lack of ancestral states (inner nodes) in a CST, as opposed to the neighbor-joining or hierarchical clustering trees. On the other hand, CST rigorously converts the multi-dimensional clustering problem to a tree partitioning problem which simplifies the interrelationship between the data without essential loss of information [81].

## 4.4 Correlation cluster net

A second correlation based representation is supplied by the correlation cluster net (CN, Figure 4-1d). This unweighted graph is constructed by connecting the nodes (i.e. the samples), whose pairwise correlation coefficient exceeds a given threshold (here $r_{threshold}=0.5$). This graph supplements the sparse CST with a more detailed and network-like overview about the sample correlation structure. It implies more connections as the CST and thus considers also weaker mutual correlations.

## 4.5 Similarities between the human tissue samples

Figure 4-1 shows the introduced sample similarity analyses for the 67 human tissues studied. The colors represent the tissue categories and are assigned in the legend in Figure 4-1a. Tissues from the same category are mostly consistently grouped in all approaches. Three major clusters are evident in second level SOM and CN: Nervous tissues (grey color), immune system tissues (blue) and the remaining ones. NJ and CST accordingly arrange nervous and immune system tissues into homogeneous groups at opposite branches. This rough classification agrees with the results discussed above.

Outliers with respect to the initial classification of the tissues become directly evident: For example, small intestine (no. 12, brown color), assigned to the category of digestive tissues, shows the same overexpressed meta-genes as the muscle tissues (see Figure 2-9). As a consequence it is located closely to the muscle cluster (green) throughout the four approaches shown. Another outlier may be identified in pituitary gland (no. 5, red color), interfering the dense clusters respectively branches of nervous tissues. However this relation originates from physical location in human brain as well as functional involvement in nervous system of this gland.

Notably, also subtle variations can be observed in the different approaches. For example, the non-linear scale of the second level SOM projects the immune and nervous system categories with a higher resolution relative to the remaining tissues. Consequently, samples belonging to the latter group are only insufficiently resolved in the second level SOM. Another example is the testis sample (no. 28, pink color), which is virtually disparate to all other samples. In NJ and CST, this sample is yet appended to the nervous and immune system branches, respectively. Also in second level SOM, this prominent expression state is not clear. CN provides the most realistic approach in this case, as it arranges the testis sample isolated from all the other ones.

## 4.6 Summary

We presented several methods that are capable to give an overview about structures within a data set and reveal relations in a sample centered view. Second level SOM provides a two-dimensional map presenting sample similarity in non-linear scale allowing separation of even very similar samples as well as investigation of more coarse similarity structures. CST and NJ represent the samples in virtually one-dimensional and hierarchical structures, respectively. These algorithms are therefore especially suited for data in the context of evolutionary processes, cell development or disease progression. CN provides a network representation directly showing sample clusters of strong correlation. Hence, although very similar, the sample similarity analyses visualize partly complementary aspects of the data which can be assessed more in detail using the spot-texture of the individual SOM portraits of the samples studied.

# 5 Selecting differential features and mining the functional context

## 5.1 Challenges

SOM machine learning alone is insufficient for extraction of differential features from the data. The SOM algorithm must therefore be supplemented with appropriate algorithms to assess significance of the features selected. The basal fold-change (FC)-score for example does not provide explicit information about statistical significance for the observed expression changes. The definition of a suited significance measure is closely related to the gene ranking and filtering tasks, which arrange features according to a designated score or remove irrelevant features completely from analysis, respectively. We apply significance analysis using three alternative test statistics based either on FC-measures or on regularized Students t-statistics with special emphasis on the error characteristics of microarray expression data.
Local, spot cluster-related lists of genes are expected to improve identification of sample-specific features with a common functional impact. For this purpose we apply methods of gene set enrichment analysis under special consideration of the meta-gene clusters generated by SOM machine learning. These methods essentially assess the enrichment of a list of differentially expressed genes compared with the total reservoir of genes studied. The members of the set are defined a priori by biological commonality for certain phenotypes. The main advantage of such methods is the direct link between the ranked gene list and biological knowledge. Therefore they provide better functional insight into the cause of the phenotypic differences under study.

## 5.2 Differential expression analysis

### 5.2.1 Scores

Our method transforms expression values in logarithmic scale ($e \equiv \log_{10} E$) into differential expression values relative to the mean expression of the particular gene in the experimental series of samples considered,

$$\Delta e_{g,m} = e_{g,m} - < e >_g \tag{16}$$

where $e_{g,m}$ denotes expression of gene g in sample m, and $<e>_g$ the average expression of g in all samples. Eq. (16) thus defines differential expression in units of the logarithmic fold change, $\log FC \equiv \Delta e$. Please note that the fold change referring to the pooled mean is equivalent to a fold change referring to a control group [82, 83].

Two alternative scores are defined to estimate the differential expression of individual genes:

1. The weighted average difference (WAD)-score,

$$\mathrm{WAD}_{g,m} = w_{g,m} \cdot \Delta e_{g,m} \quad \text{with} \quad w_{g,m} = \frac{\Delta e_{g,m} - \min\left(\Delta e_{g,m}\right)}{\max\left(\Delta e_{g,m}\right) - \min\left(\Delta e_{g,m}\right)} \tag{17}$$

is a fold change (FC)-based score which accentuates large expression values [65, 66]. The main idea of the WAD method is based on the observation that potential marker genes often tend to have high expression levels. Moreover, it intuitively considers the fact that the experimental error of expression values typically inflates at small expression levels in logarithmic scale [84, 85]. Hence, the basic assumption for the WAD-approach is that 'strong signals are better signals' in the gene ranking problem [86–88]. The WAD score therefore 'amplifies' large expression values and 'represses' low ones. It is especially suited for small sample sizes and it partially outperforms popular standard methods for determining differentially expressed genes when sensitivity and specificity are considered simultaneously [65, 66]. Note that the weighting factor in Eq. (17) can be transformed into a function of the absolute expression values as in the original paper of Kadota et al. [65],

$$w_{g,m} = \frac{e_{g,m} - \min\left(e_{g,m}\right)}{\max\left(e_{g,m}\right) - \min\left(e_{g,m}\right)} \tag{18}$$

showing that the weighting factor linearly scales with the expression level of the gene.

2. The shrinkage t-score,

$$t_{g,m} = \sqrt{R_m}\,\frac{\Delta e_{g,m}}{\sigma^{\mathrm{shr}}_{g,m}} \tag{19}$$

integrates the standard error of gene expression values in replicated measurements. The shrinkage statistic in Eq. (19) was defined in analogy with previous approaches [36, 89, 90]. Here $\sigma^{shr}_{g,m}$ denotes the standard deviation of differential expression of gene g measured under condition m. It is estimated using the shrinkage approach which considers two components: firstly, the individual standard deviation of the expression values is calculated using the $R_m$ available replicates, $\sigma_{g,m} \equiv \sqrt{\left\langle \left(e_{r,g,m} - e_{g,m}\right)^2 \right\rangle_r}$. Secondly, the locally pooled error (LPE) robustly estimates the mean standard deviation as a function of the expression, $\sigma_{LPE}(e_{g,m})$. To obtain this LPE function the values of individual standard deviation $\sigma_{g,m}$ are plotted for each sample as a function of the logarithmic expression, $e_{g,m}$, and locally pooled over a moving average window of a few hundred neighboring values. Figure 5-1 shows these

plots for selected tissue examples, where dots represent individual genes with coordinates $e_{g,m}$ and $\sigma_{g,m}$. The green curves indicate the respective LPE function.
Finally, the individual standard deviation and LPE measure of each gene are combined to provide the shrinkage error estimate used in Eq. (19):

$$\sigma_{g,m}^{shr} = \sqrt{\lambda \cdot \sigma_{g,m}{}^2 + (1-\lambda) \cdot \sigma_{LPE}(e_{g,m})^2} \tag{20}$$

The parameter $\lambda$ ($0 \leq \lambda \leq 1$) scales the degree of shrinking $\sigma_{g,m}$ towards $\sigma_{LPE}$.

The shrinkage t-statistics was developed in the framework of James-Stein analytic shrinkage and applied in different modifications in gene expression analysis (see [89] and references cited therein). The basic idea behind Eq. (20) implies that the error estimate based on $\sigma_{g,m}$ alone might be very imprecise, e.g. if only few replicates are available. The resulting large 'error of the error' leads to highly uncertain naive t-scores associated with large false positives rates. Additionally, it has been previously suggested that estimates of the variance for individual genes is questionable [91, 92]. Yet accurate estimation of variability of gene expression is essential for correct identification of differentially expressed genes. Additional information may be gained by involving variance estimates across all or part of the experiment. Such information borrowing methods that exploit this information are able to improve the results [87, 91]. Particularly, local-pooled-error (LPE) estimates for evaluating significance of each gene's differential expression have been shown to effectively identify significant differential expression patterns with a small number of replicated arrays [92]. Eq. (20) therefore realizes the shrinkage approach, combining the pooled and the gene-specific error to consider both, individual and common factors. Shrinkage t-score consistently leads to an accurate gene ranking which might outperform simple t-statistics or FC-scores [89].

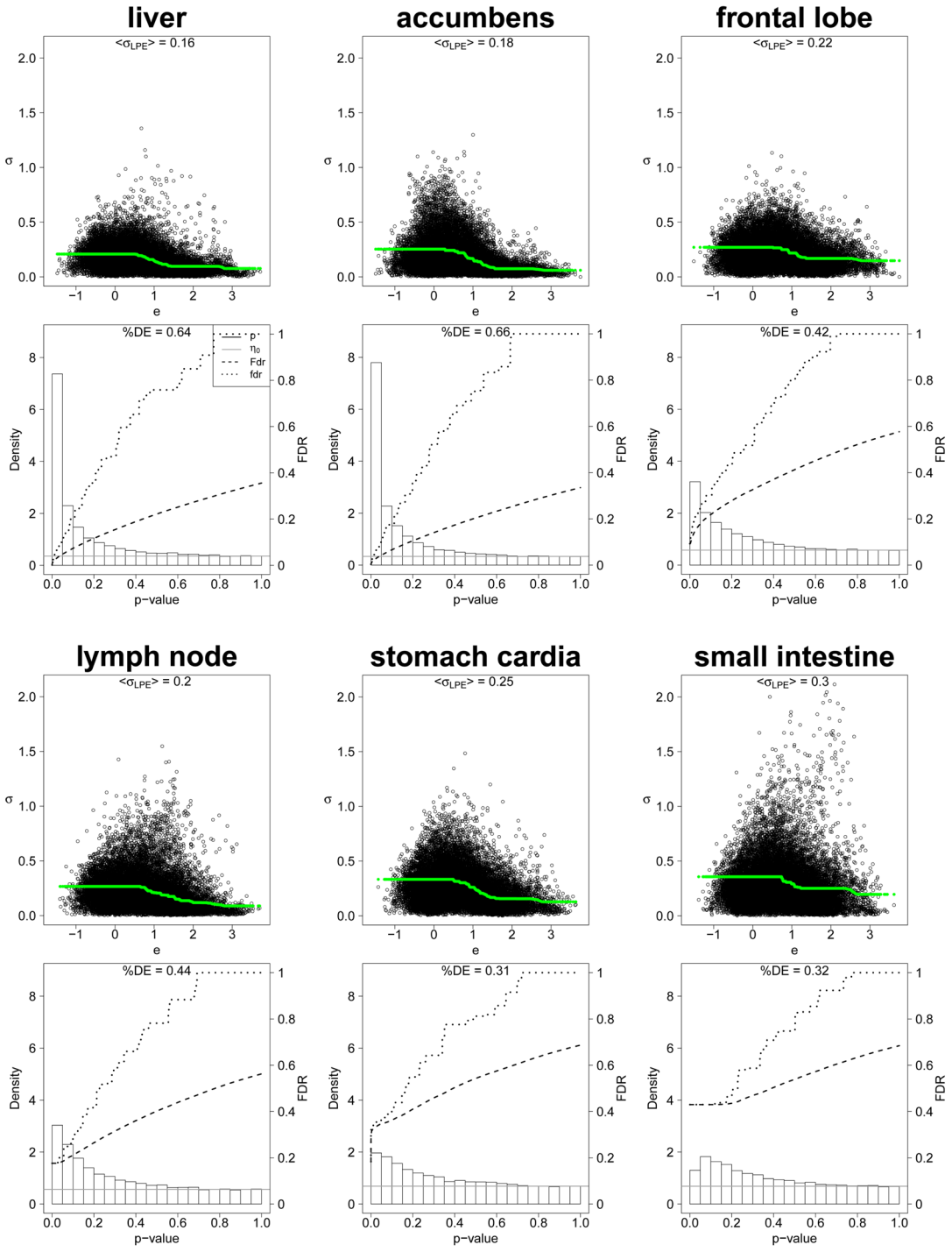


**Figure 5-1: Error and significance characteristics of selected tissue examples: The first and third row of figures show error distributions (dots) and locally pooled estimates (green curves) as a function of the logarithmic expression, e. The LPE-curves are calculated as moving average over 500 single probe values under the condition of non-positive slope which ensures that the LPE is maximal at small expression values. The second and fourth rows of figures show the respective p-value density distributions (bar histograms) together with the local FDR (dotted curves) and tail area-based FDR (dashed curves) obtained from the shrinkage t-statistics (see Eqs. (19) and (23)). The density-levels of null-genes, $\eta_0$, are indicated by fine horizontal lines. The examples shown are ordered with fraction of differentially expressed genes %DE (Eq. (22)).**

### 5.2.2 p-values and false discovery rate

p-values can be derived from the shrinkage t-statistics (Eq. (19)) to characterize the significance of differential expression for each gene assuming Student's t-distribution. The obtained density distribution for the p-values of all genes in one sample, $\rho$(p), meets the normalization condition $\int_0^1 \rho(p) \cdot dp = 1$. P-value distributions are shown for selected tissues of different mean error level in Figure 5-1. Under the null hypothesis a uniform distribution $\rho_0$(p) = 1 is expected, whereas the alternative hypothesis will produce a skewed distribution, $\rho_{DE}$(p), decaying with increasing p because differentially expressed genes tend to gather close to p=0 [93]. In general, the observed distribution can be interpreted as the superposition of two components due to differentially and not-differentially expressed genes,

$$\rho(\mathrm{p}) = \rho_{\mathrm{DE}}(\mathrm{p})\ (1-\eta_0)\ + \rho_0(\mathrm{p})\eta_0 \tag{21}$$

where $\eta_0$ is the fraction of non-informative 'null'-genes among all genes considered [93, 94]. It was derived using "fdrtool" [95] under the assumption of vanishing differential expression at p=1: $\rho_{DE}(1) = 0$, giving rise to $\rho(1) = \eta_0$ [96].

The total fraction of differentially expressed and thus informative genes per sample can be estimated using the background level of the p-value distribution, $\eta_0$:

$$\%DE = 1-\eta_0 \tag{22}$$

"fdrtool" was further used to calculate false discovery rates (FDR) to control the number of false discoveries:

$$\mathrm{fdr(p)} = \frac{\eta_0}{\rho(\mathrm{p})} \quad \text{and} \quad \mathrm{Fdr(p)} = \frac{\eta_0 \cdot \mathrm{p}}{\int_0^{\mathrm{p}} \rho(\mathrm{p}) \cdot \mathrm{dp}} \tag{23}$$

Here fdr and FDR denote the local and tail area-based FDR estimates, respectively. The Fdr(p)-values provide a cumulative estimate of FDR referring to all genes on top of a list with p-values p'≤p whereas fdr(p) estimates the FDR of a selected gene with p'=p [97]. For a monotonically decaying total density $\rho$(p) both, fdr(p) and Fdr(p), are increasing functions which well correlate in the intermediate p range. The local FDR-estimate thereby systematically exceeds the tail-based one, fdr(p)≥ Fdr(p) (see the examples shown in Figure 5-1). Their limiting values at p=0 and 1 are given by the equations Fdr(0)= fdr(0), Fdr(1)= $\eta_0$ and fdr(1)= 1, respectively.

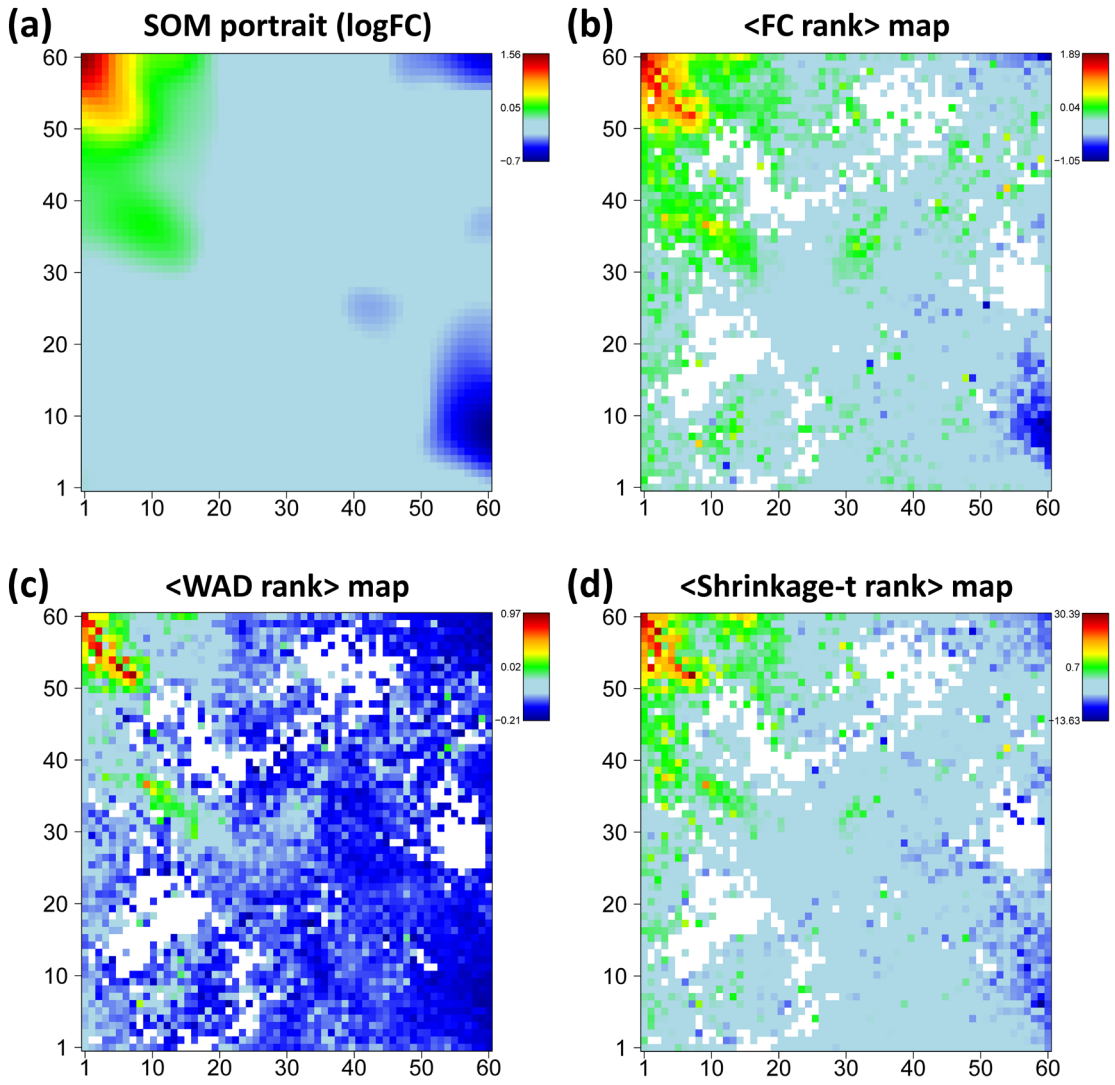


**Figure 5-2: SOM portrait of *nucleus accumbens* (standard FC-portrait, panel a) and the average-rank maps for FC, WAD and shrinkage t-score statistic (b-d). White areas indicate empty meta-genes.**

### 5.2.3 Rank maps

The SOM approach processes information about differential gene expression ($\Delta e_{g,m}$ , see above) and features this information in a compressed fashion in terms of meta-gene expression values ($\Delta e_{k,m}^{meta}$, k and m denote a particular meta-gene and sample, respectively). SOM portraits consequently visualize differential (fold change) expression pattern. Alternatively one can map other measures onto the SOM grid, such as the rank of the genes taken from their ranked list of differential expression.

Figure 5-2 shows the SOM portrait of one particular tissue example, nucleus accumbens, taken from the category of nervous system in log FC units (panel a), together with the respective average-rank maps for the three different scores defined: FC-, WAD- and shrinkage

t-score (panels b-d, respectively). The rankings of genes refer to the total gene lists which contain all genes studied. The maps color-code the mean rank for each meta-gene, calculated as the arithmetic average over the individual ranks of the associated single genes in the total list.

The three alternative scores provide very similar pattern, however with subtle differences: The contrast, i.e. the gradient between areas of under- and overexpression is largest for the WAD-ranking and similar for FC-ranking and t-shrinkage. In general, genes on top of the three score rankings accumulate in the red overexpression spot of the standard SOM portrait. Additionally, the rank maps reveal hidden details within the SOM spots such as the chain-like cluster of meta-genes of small rank within the overexpression spot (compare panel a with b-d in Figure 5-2). The analysis of such fine-structures might help to refine the subsequent selection of relevant genes and meta-genes within the spots.

Summarizing, standard fold change based SOM portraits provide reliable characterization of the samples in terms of particular over- and underexpressed meta-genes. To this, rank maps reveal details potentially important in particular problems.

## 5.3 Mining the functional context: Gene set enrichment analysis

The SOM assigns meta-gene clusters of single genes with similar, mostly highly correlated expression profiles. The correlation and thus coexpression of the single gene profiles can be utilized with regard to putative gene function because biological processes are usually governed by coordinated modules of interacting molecules [98]. This 'guilt-by-association' principle assumes, that co-expressed genes are likely to be co-regulated and thus functionally associated [76, 99].

Gene set analysis requires knowledge of predefined gene sets and the corresponding biological meanings to study their enrichment in gene lists obtained from independent differential expression analysis (see [100] for a critical review and references cited therein). For example, a large and diverse collection of such sets can be downloaded from the 'gene-set-enrichment-analysis'-website[9]. Particularly, 1454 gene sets were included into our analysis according to the GO terms 'biological process' (825 sets), 'molecular function' (396 sets) and 'cellular component' (233 sets). These sets may partly overlap in component genes, and some gene sets are subsets of others due to the hierarchical nature of the GO-systematics [101]. To maximize the functional annotation conveyed by the gene sets, all these sets are considered.

Previous SOM analyses have shown that functionally related genes indeed cluster in the SOM portraits [10]. Here, three potential approaches are described combining the meta-gene concept and gene set enrichment analysis:

[9] http://www.broadinstitute.org/gsea

1. Meta-gene as clusters of single genes are individually analyzed for overrepresentation of genes defined in a certain gene set.
2. Spots of (e.g. simultaneous overexpressed) meta-genes are identified in the overexpression maps and associated genes evaluated in terms of overrepresentation.
3. Spots of meta-genes are identified in the samples' SOM portraits, giving rise to inclusion of the sample specific expression values. This enables combined overrepresentation and overexpression analysis.

The term *overrepresentation* is hereby used to assign the probability to find members of a set in a given gene list, compared with their random appearance. This method is therefore independent of the respective gene expression values or scores. Contrarily, *overexpression* terms deviation between the mean expression score taken from all set-members in a list, compared with the mean score of all list members. The term *enrichment* will be finally used for estimates which combine overrepresentation and overexpression.

### 5.3.1 Gene set overrepresentation maps

Gene set overrepresentation analysis classifies each gene studied according to two memberships leading to a 2×2 contingency table for further testing (Table 5): firstly, its membership in the particular set of functionally annotated genes of length $N_{set}$ and, secondly, its membership in the respective list of selected genes of length $N_{list}$. The intersection of the 'set' and the 'list' defines the number of '*positive*' genes, $N_+$. Then, overrepresentation of these positive genes is estimated using the hypergeometric distribution. It allows to estimate the cumulative probability that there is more overlap between the 'list' and the 'set' than would be expected by chance [102–104],

$$p = P(n > N_+) = \sum_{n=N_+ +1}^{N_{set}} p_{HG}(n) \quad \text{with} \quad p_{HG}(n) = \frac{\binom{N_{set}}{n}\binom{N - N_{set}}{N_{list} - n}}{\binom{N}{N_{list}}} \qquad (24)$$

The gene set overrepresentation approach thus considers the joint membership of a gene in a gene set and an independent list of genes, without taking into account the particular expression values or scores of the genes in the list. For example, it ignores whether a 'positive' features strong or weak differential expression.

**Table 5: 2x2 contingency table specifying the numbers of genes in different classes concerning gene set overrepresentation in a list of selected genes**

| # of genes | in list | not in list | total |
|---|---|---|---|
| **in set** | $\underline{N_+}$ | $N_{set}- N_+$ | $\underline{N_{set}}$ |
| **not in set** | $N_{list}- N_+$ | $N- (N_{list}+ N_{set})+ N_+$ | $N- N_{set}$ |
| **total** | $\underline{N_{list}}$ | $N- N_{list}$ | $\underline{N}$ |

For each of the clusters defined by the meta-genes the degree of overrepresentation is estimated with respect to each pre-defined gene sets using the hypergeometric (HG-) test. It provides a p-value for each meta-gene and each gene set considered. The p-values of a certain gene set are visualized using a two-dimensional mosaic analogous to the SOM portraits and appropriate color-coding. These overrepresentation maps allow identification of meta-genes containing a considerable fraction of genes for a selected gene set, e.g. by simple visual inspection. Note that these maps apply to the SOM itself rather than to individual samples, because mapping of the genes to the meta-gene clusters is a property of the whole series of samples studied.

Figure 5-3 shows global overrepresentation patterns in the SOM of human tissues for selected gene sets. Overexpression is observed in different regions of the map, for example in the bottom right and top left corner for genes related to 'immune response' and to 'nervous system development', respectively (see red circles in Figure 5-3). The examples also show that overrepresentation is either strongly localized in one region of the map (e.g. for 'nervous system' or, to a less degree, for 'RNA repair' and 'immune system process') or it spreads over different and disjunct regions of the SOM (e.g. for 'apoptosis').

### 5.3.2 Spot-related overrepresentation

Overrepresentation analysis is not restricted to single meta-genes. It is applied to spots of over- (or under-) expressed meta-genes detected in the SOM portraits. Such spots of co-expressed meta-genes are potentially co-regulated and thus they might carry important functional information. This approach links overrepresentation with overexpression by combining spot selection with overrepresentation analysis using the HG-test, as described above.

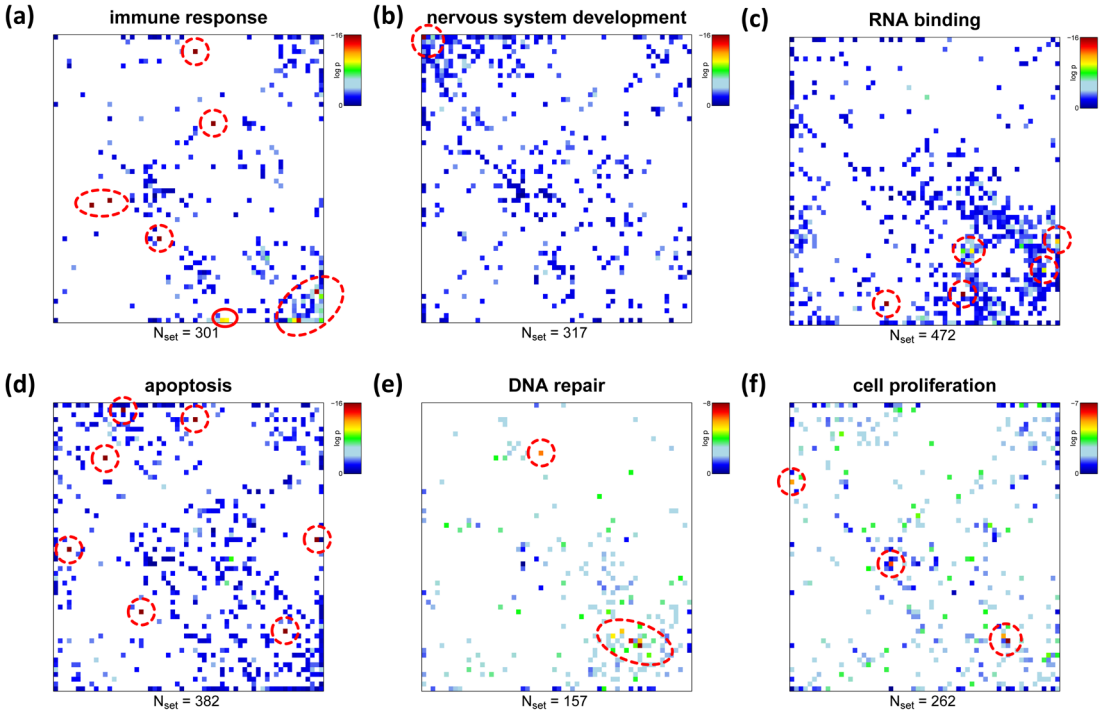


**Figure 5-3: Overrepresentation maps of six selected gene sets containing between $N_{set}$= 157 and 472 genes. Overrepresentation in each tile of the mosaic is calculated in units of log(p) using the hypergeometric distribution. Red circles indicate areas of strongest enrichment, white tiles indicate meta-genes not containing genes from the respective set.**

Nine essential overexpression spots are identified in the SOM of human tissues, using the 98-percentile criterion of maximum expression (see Figure 2-17a). Figure 5-4a shows the nine spots of strongly overexpressed meta-genes, along with a legend assigning the two leading overrepresented gene sets for each of the spots to get a first idea about their possible biological context. For example, spot 'A' in the top left corner of the map is clearly related to molecular processes in nervous tissues according to the leading gene sets obtained.

The overexpression spot heatmap in Figure 5-4b provides a direct link between HG-overrepresentation and overexpression in a tissue- and spot-specific way. It visualizes the average meta-gene expression in each of the spots in the series of tissues. This representation reveals that nervous, muscle and homeostasis tissues are characterized by essentially only one overexpression spot (spot 'A', 'B' and 'C1', respectively) with clearly assigned molecular function. Some of the tissue-specific spots are also overexpressed in other tissues. For example, the muscle-specific spot 'B' shows overexpression also in tongue and small intestine: Both organs partly contain also muscle tissues.

The enriched areas in the overrepresentation maps of the gene sets 'nervous system development' and 'immune response' (see Figure 5-3) largely agree with the overexpression spots in the SOM portraits of nervous and immune system tissues, respectively. A non-

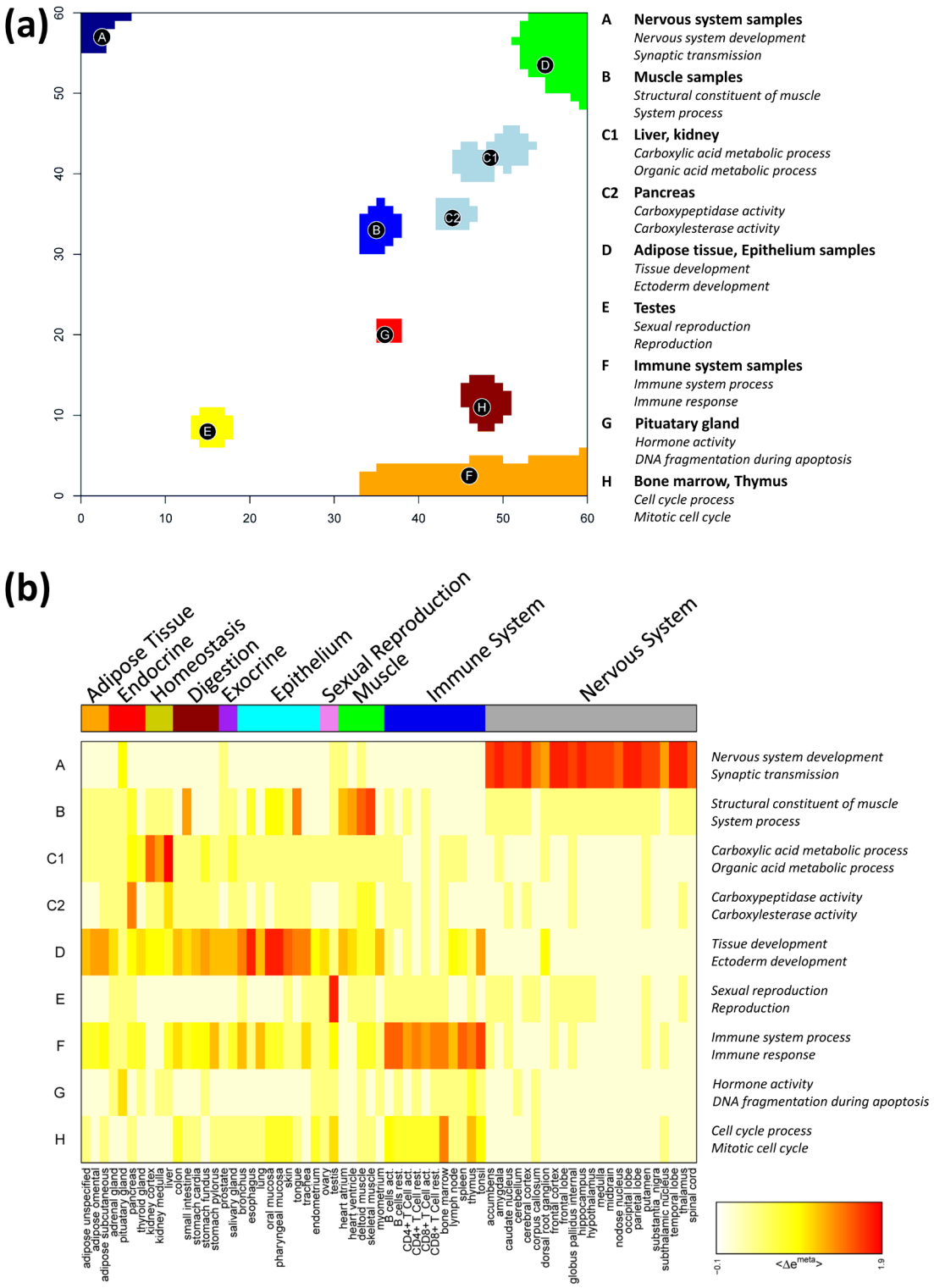


**Figure 5-4: The spot map (panel a) shows nine spots which are identified as strongly overexpressed in different tissues. Overrepresentation of a collection of 1454 gene sets is estimated for each spot using the HG-test. The right legend assigns tissue categories and the two most significantly overrepresented gene sets to the respective spots. Overexpression spot heatmap (panel b) directly links gene set HG-analysis and expression values: Each spot refers to one row, containing meta-gene expressions averaged over the respective spot for each sample.**

negligible number of genes from these sets are however located in other regions of the map which are assigned to alternative molecular functions. For example, genes from the gene set 'immune response' also accumulate in spot 'D' (top right corner of the overrepresentation map in Figure 5-3a), which is however assigned to 'tissue development'. This spot is overexpressed in a larger number of tissues such as epithelium and adipose tissues which are not explicitly assigned to the category immune system tissues. Moreover, subgroups of genes from the gene sets shown are located in the central area of the map which accumulates virtually invariant and weakly expressed genes (compare with variance map in Figure 2-14b). Possibly part of the genes in these sets are incorrectly specified and/or possess a more complex activation pattern 'beyond' the input patterns used to train the SOM. Hence, combination of gene set overrepresentation analysis with SOM-expression profiling allows verification and further refinement of existing gene sets.

In summary, gene set overrepresentation maps link selected gene sets and different regions of the SOM portraits with single-tile resolution. Regions of the SOM, in turn, can be grouped into over- or underexpression spots in different tissues. Overrepresentation analysis then provides lists of significantly overrepresented gene sets which characterize the respective spot in a functional context.

Both, the meta-gene-wise overrepresentation maps and the spot-wise overrepresentation analysis constitute a link between characteristic expression pattern and concepts of molecular function for the associated genes. These orthogonal views complement each other: The former one judges the homogeneity of a selected set with respect to different meta-gene expression profiles. The latter one assigns selected expression profiles to their tentative molecular function.

### 5.3.3 Gene set enrichment score

The hypergeometric test applies a binary 'included – or - not included' criterion to assess the positive membership of the genes from a gene set in a selected list, e.g. taken from meta-gene clusters or spots as described above. Contrary, the so-called gene set overexpression approach compares the gene set statistics with the null hypothesis given by the ensemble of all genes studied (see refs. [100] and [105] for a review). In this case however no overrepresentation of a set in a sub-ensemble of a gene list is taken into account.

The gene-set-Z (GSZ)-score provides a combination of overrepresentation and overexpression which explicitly considers the individual expression values of the genes included in the list [105]: The GSZ measure estimates enrichment of a gene set in a list using its score statistics, for example $S_g = t_{g,m}$ utilizing shrinkage t-score of gene g in sample m. It is designed in such a way that top-ranked members of the gene list with high scores more intensively contribute to the GSZ than members with lower values down the list. In a first step, the total sum of the score function over the complete gene list is decomposed into two components, containing members and non-members of the set,

$$S_{list} = \sum_{all\ g \in list} S_g = S_{list}^{+} + S_{list}^{-} \qquad \text{with } S_{list}^{+} = \sum_{g \in list\ AND\ g \in set} S_g \quad \text{and} \quad S_{list}^{-} = \sum_{g \in list\ AND\ g \notin set} S_g \tag{25}$$

Secondly, the regularized Z-score of the differential score, $\Delta S_{list} = S_{list}^{+} - S_{list}^{-}$ is defined as

$$GSZ = \frac{\Delta S_{list} - E(\Delta S_{list})}{\sqrt{\lambda \cdot var(\Delta S_{list})^2 + (1-\lambda) \cdot var_0{}^2}} \tag{26}$$

(see [WIRTH3] and [105] for details).

Here, $E(\Delta S_{list})$ and $var(\Delta S_{list})$ denote the expected mean and the variance of $\Delta S_{list}$, respectively. $var_o$ and $\lambda$ denote the regularization constant and a scaling factor ($0 \leq \lambda \leq 1$) which were chosen to stabilize the variance in the denominator of Eq. (26) especially for short lists [WIRTH3].
The differential score $\Delta S_{list}$ reflects the summarized score of the members in the list compared to the non-members integral score. This implies strong effect of the numbers of these two fractions, which is considered in the expectancy value $E(\Delta S_{list})$:

$$E(\Delta S_{list}) = \langle S \rangle_{list} \cdot \left( \langle N_+ \rangle_{HG} - \langle N_- \rangle_{HG} \right) \tag{27}$$

where $\langle S \rangle_{list} = S_{list} / N_{list}$ describes the mean value of the expression score in the gene list. Additionally, the second factor in eq. (27) reflects the difference of expected number of members and non-members of the set, given by expectancy value of the hypergeometric distribution:

$$\langle N_+ \rangle_{HG} = N_{set} \frac{N_{list}}{N} \quad \text{and} \quad \langle N_- \rangle_{HG} = N_{list} - \langle N_+ \rangle_{HG} \tag{28}$$

The variance of $\Delta S_{list}$ is calculated according to

$$var(\Delta S_{list})^2 = 4 \left( \frac{var(S_{list})}{N_{list} - 1} \left( \langle N_+ \rangle_{HG} \cdot \left( N_{list} - \langle N_+ \rangle_{HG} \right) - var(N_+) \right) + \langle S \rangle_{list}{}^2 \cdot var(N_+) \right) \tag{29}$$

which combines the variance of the score statistics, $var(S_{list}) = \frac{1}{N_{list}} \sum_{g \in list} \left( S_g - \langle S \rangle_{list} \right)^2$ , and the variance of the hypergeometric distribution $var(N_+) = \langle N_+ \rangle_{HG} \cdot \left( 1 - \frac{N_{set}}{N} \right) \left( \frac{N - N_{list}}{N - 1} \right)$.

Finally, the obtained GSZ-values were transformed into p-values using a permutation approach which generates the respective null distribution by random rearrangement of genes in the collection of predefined gene sets. One and two tailed tests were applied to assess over- or underexpression and differential expression (i.e., under- *and* overexpression), respectively.

Two special cases of the GSZ-score can be derived referring to overexpression and overrepresentation, respectively. Firstly, the GSZ-score can be calculated for the whole gene list, i.e. $N_{list}=N$. For this special case, differential score can be rewritten as $\Delta S_{list}\big|_{N_{list}=N} = \left(2\left\langle S^{+}\right\rangle_{list}\cdot N_{set} - \left\langle S\right\rangle_{list}\cdot N\right)$ where $\left\langle S^{+}\right\rangle_{list} = S^{+}_{list}/N_{set}$ is the mean expression score averaged over all members of the gene set, and the according expectancy value as $E(\Delta S_{list})\big|_{N_{list}=N} = \left\langle S\right\rangle_{list}\cdot\left(2\cdot N_{set} - N\right)$ . Combined with the error estimator $SE(\Delta S_{list})^2\Big|_{N_{list}=N} \approx 4\cdot N_{set}\cdot \mathrm{var}(S_{list})$ , eq. (26) provides the GSZ-score of the full gene list

$$GSZ\big|_{N_{list}=N} = \frac{\left\langle S^{+}\right\rangle_{list} - \left\langle S\right\rangle_{list}}{\sqrt{\mathrm{var}(S_{list})/N_{set}}} \tag{30}$$

assuming $\lambda=1$ without loss of generality. It represents a Z-statistics estimating the overexpression averaged over the gene set compared to average expression of the total gene list. The standard error here is estimated using the variance of S for sample size $N_{set}$. This approach is used to obtain a GSZ-score for the total list of gene expression scores of a sample, reflecting the global tendencies of functional involvement of a sample.

The second special case assumes an identical value of the expression score for all genes, $S_g=1$, after ranking. The difference score thus simply counts the difference of members and non-members of the set in the list, $\Delta S_{list}\big|_{S=1} = N_{+} - N_{-}$. The expected mean and the variance of the difference score are given by $\langle S\rangle_{list}=1$ and $\mathrm{var}(S_{list})=0$, respectively. Insertion into Eq. (26) provides the GSZ-score for $\lambda=1$

$$GSZ\big|_{S=1} = \frac{\left(N_{+} - \left\langle N_{+}\right\rangle_{HG}\right)}{\sqrt{\mathrm{var}(N_{+})}} \tag{31}$$

It represents a Z-statistics estimating the overrepresentation in terms of the deviation of the actual number of positive members from the expected mean according to the hypergeometric distribution and the respective variance.

Equations (30) and (31) illustrate that the GSZ-score in its general formulation in Eq. (26) estimates enrichment in terms of a combination of overexpression and overrepresentation Z-scores. It has been shown in ref. [105] that the GSZ-score is related to alternative scores, namely the Random Sets [106] and the max-mean gene set statistics [107] representing a unification between these relevant scoring functions. Another comparative study on different gene set enrichment methods showed that removing incoherent pathways prior to analysis improves specificity [108]. The GSZ-score implicitly accounts for coherency because inconsistent genes with positive and negative contributions to the sum in Eq. (25) virtually compensate each other.

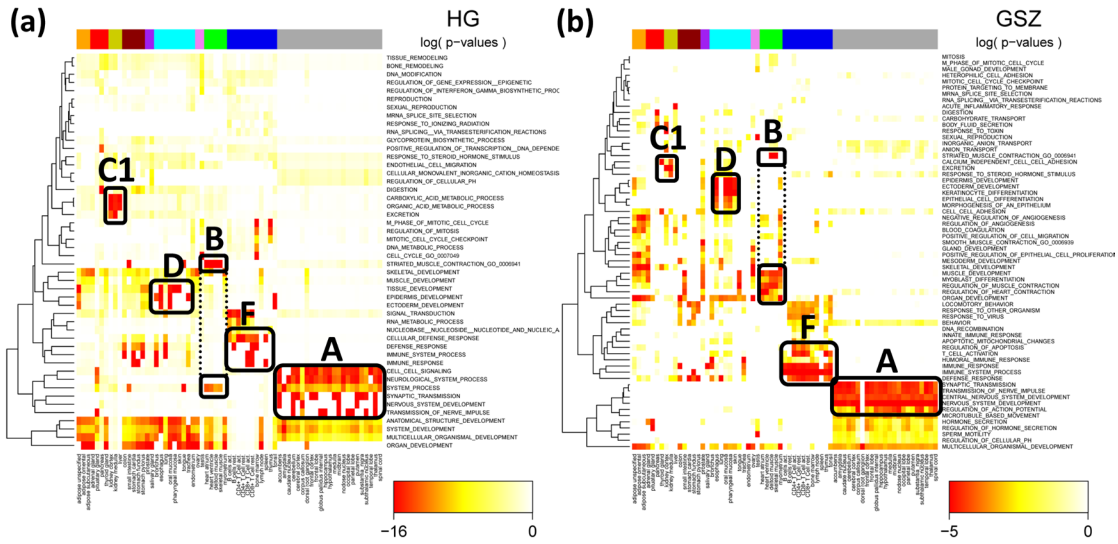


**Figure 5-5: One-way hierarchical clustering heatmaps of significantly enriched gene sets (rows) versus tissues (columns) using the HG- (panel a) and the GSZ- (panel b) statistics. The three-top gene sets per overexpression spot are selected to contribute to the according map. The heatmap color-codes the p-values of the respective score in log-scale (see the legends in the figure). The tissue categories are color-coded in the bar above the heatmap according to assignment in Figure 2-9. The capital letters approximately assign clusters of enriched gene sets in correspondence with the spots selected in integral overexpression map (see Figure 5-4).**

### 5.3.4 Spot-related GSZ-analysis

The algorithm of spot-based GSZ-enrichment analysis is mainly identical with that of spot-related HG-overrepresentation analysis introduced above. Tissue specific spots of overexpressed meta-genes are identified in the SOM portraits using, for example, the 98-percentile criterion. The obtained list of associated genes together with respective expression scores in the sample are then analyzed for gene set enrichment using the GSZ-score.

To compare GSZ- and HG-analysis, both statistics are applied to overexpression spots identified in the sample's SOM portraits, either with (GSZ) or without (HG) explicit consideration of the expression values. The top-three gene sets per spot in each tissue are selected and collected in a list of most enriched gene sets in all spots. This comprehensive list was used to generate heatmaps, visualizing the sample specific enrichment in terms of GSZ- and HG-based p-values (Figure 5-5). Additionally, hierarchical clustering was applied to group similarly expressed gene sets in vertical direction. The HG-heatmap in Figure 5-5a reveals five to six clusters of gene sets, which can be clearly assigned to respective spots of overexpression (see spot labels in Figure 5-5 and Figure 5-4): a group of about six gene sets associated with the nervous tissues represents overexpression spot 'A' in a tissue-specific fashion. Other groups of enriched gene sets can be associated with immune systems tissues ('F'), muscle tissues ('B'), epithelial ('D') and homeostasis tissues ('C1'). Figure 5-5b shows the respective GSZ-enrichment heatmap, featuring essentially the same clusters of gene sets as its HG counterpart. Table 6 lists the HG- and GSZ-enriched gene sets associated with the

overexpression spots. The obtained number of 64 gene sets however exceeds the 48 gene sets in the HG-enrichment map indicating the increased diversity of the GSZ approach.
The standard algorithm applies the 'top-three' criterion, i.e. it selects the three top gene sets of each local spot list, to characterize the functional context of gene expression in the different samples. This approach equally weights each spot in terms of the number of selected gene sets and thus ensures that each spot is equally represented in the heatmap. Alternatively, gene sets are selected according to their significance of enrichment in each of the tissues. The obtained enrichment lists are very similar compared with those obtained using the 'top-three' selection criterion [WIRTH3]. In summary, HG- and GSZ-enrichment maps based on the 'top-three' selection criterion provide a suited overview of the gene sets most important in the experimental series studied. For a more detailed analysis, full lists of gene sets for each spot are generated, whereas enrichment heatmaps provide information in summarized fashion.

**Table 6: Molecular characteristics of selected overexpression spots as obtained by HG- and GSZ-enrichment analysis** [a]

| spot | GSZ | HG |
|---|---|---|
| **A** | **Synaptic Transmission** | Cell-Cell Signaling |
| | **Transmission of Nerve Impulse** | Neurological System Process |
| | Central Nervous System Development | **Synaptic Transmission** |
| | **Nervous System Development** | **Transmission of Nerve Impulse** |
| | Regulation of Action Potential | **Nervous System Development** |
| **B** | Muscle Development | **Striated Muscle Contraction** |
| | Myoblast Differentiation | System Process |
| | Regulation of Muscle Contraction | |
| | Regulation of Heart Contraction | |
| | **Striated Muscle Contraction** | |
| **C1** | Carboxylic Acid Metabolic Process | Calcium Independent Cell-Cell Adhesion |
| | Organic Acid Metabolic Process | **Excretion** |
| | **Excretion** | Response to Steroid Hormone Stimulus |
| **D** | **Epidermis Development** | Tissue Development |
| | **Ectodermis Development** | **Epidermis Development** |
| | Keratinocyte Differentiation | **Ectodermis Development** |
| | Epithelial Cell Differentiation | |
| | Morphogenesis of an Epithelium | |
| **F** | Regulation of Apoptosis | Cellular Defense Response |
| | T-Cell Activation | **Defense Response** |
| | Humoral Immune Resonse | **Immune System Process** |
| | **Immune System Process** | **Immune Response** |
| | **Immune Response** | |
| | **Defense Response** | |

[a] Gene sets enriched in both approaches are printed in bold letters.

### 5.3.5 Gene set SOM

A complementary approach of sample profiling is provided by the so-called *gene set SOM* which relates expression measures to gene sets instead of single genes. So-far, single gene expression data was used as input for the SOM. In an additional step of aggregation, these single features can be pooled prior to SOM training according to a higher level of information. For that it is necessary to access previously defined sets of usually functionally related features, for example GO gene sets as discussed above.

For illustration we use the GSZ overexpression scores of $N_{GS}$=1,454 GO sets in the 67 human tissue samples to train a 60x60 SOM which can be directly compared with the original 'single-gene' SOM. More concretely: The gene-level expression profiles of N single genes measured in M samples were substituted by the GSZ-expression profiles of the set=1…$N_{GS}$ gene sets. In this case the GSZ-scores refer to the full gene lists, i.e. to the special case $N_{list}$=N given in Eq. (30) with S=Δe,

$$GSZ_{set,m} = \frac{\left\langle \Delta e_{g,m} \right\rangle_{g \in set} - \left\langle \Delta e_{g,m} \right\rangle_{g \in N}}{\sqrt{var(\Delta e_{g,m})_{g \in set} / N_{set}}} \qquad (32)$$

The SOM is then trained with the GSZ profiles. It consequently provides K *meta-gene set profiles*. The resulting occupancy of the meta-genesets is less than one individual gene set per node (<$n_k$>=1,454/3,600=0.4). In this particular application, the SOM algorithm clusters gene sets of similar profiles together using the Euclidean distance as similarity measure. Gene sets of related functionality are likely to behave similarly in terms of their GSZ-scores and thus they are expected to be mapped to the same or neighbored meta-genesets.

Figure 5-6 shows the gene set SOM portraits of 42 samples selected from the human tissue data. First inspection of these portraits reveals consistent pattern for most of the categories, agreeing with the original 'single gene' SOM portraits (compare with Figure 2-9). However, the spots of overexpressed meta-genesets appear better resolved with less overlapping regions in most cases. Most samples show one category-specific spot. In addition to these characteristic spots, individual samples show further spots which are either unique for the respective tissue, or emerge in other samples too. For example, CD4+ T-cells (no.36) shows, beside the immune-specific spot in the top left corner, a unique spot in the center of the left edge. Bone marrow and thymus (no. 40 and 43, respectively) share a common spot on the left edge with ovary and testis sample (no. 27 and 28, see discussion below).

On the other hand, regions of underexpressed meta-genesets are widespread and mostly without pronounced spot-like structure. Consequently, gene set based SOM portraits well characterize the human tissue data in terms of gene sets, which are overexpressed in specific tissue samples, but poorly in terms of underexpressed gene sets.

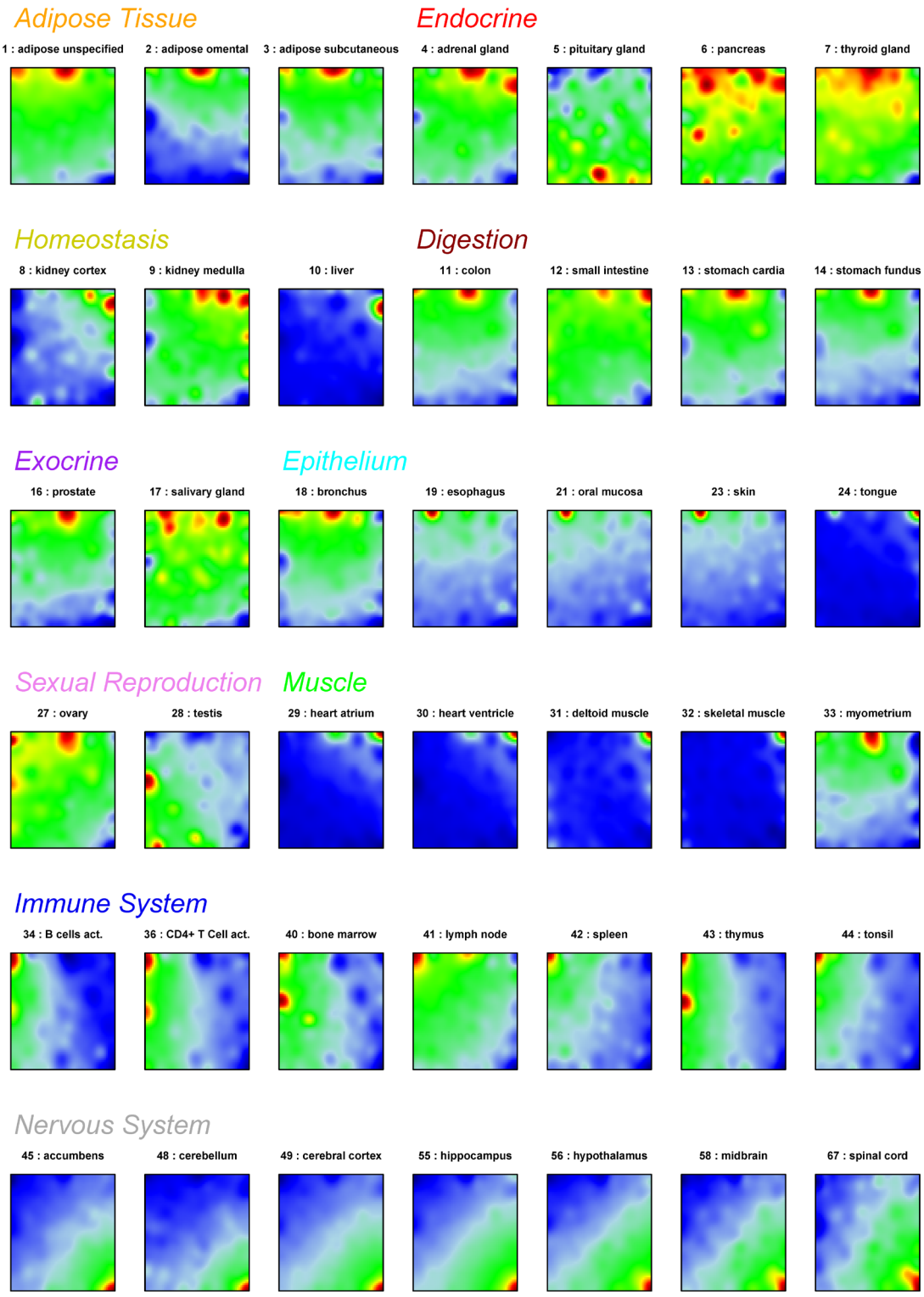


**Figure 5-6: Gene set SOM profiles of 42 human tissues selected from the tissue data set. Instead of single gene expression values, GSZ overexpression scores of 1,454 GO sets were used as input data.**

Figure 5-7a shows the overexpression summary map of the gene set SOM. It collects nine spots which were identified using the 98-percentile criterion. The legend on the right hand side of Figure 5-7 lists the top-five overexpressed gene sets in each of the spots together with tissues and categories showing this spot. As mentioned above, thymus, bone marrow, testis and ovary, classified as immune system and sexual reproduction samples, respectively, share one particular spot at the left edge (spot 'F' in Figure 5-7). Gene sets located in this spot are almost exclusively related to replication of cells, which in turn is the major physiological 'task' of these tissues: Leucocytes proliferate in thymus and bone marrow, spermatozoa and ovocytes in testis and ovary, respectively. Spots with consistent functional annotation can be also found for epithelium (spot 'C'), muscle ('D'), immune ('G') and nervous system ('H').

The spot heatmap shown in Figure 5-7b exhibits well defined overexpression patterns across the samples: For example spots 'G' and 'H' are exclusively overexpressed in immune and nervous system, respectively. Also non-specific spots without clear overexpression pattern are evident in the GSZ spot heatmap, for example spots 'A' and 'I'. These spots can be observed in only a few samples.

Next, we compare the gene sets overrepresented in the spots of the original gene-based SOM and the gene sets accumulated in the spots of the gene set SOM. Recall that these gene sets are determined by the hypergeometric test in the former case (see Figure 5-4 and Table 6). In the latter case however they are determined by the respective gene set clusters within the overexpression spots (Figure 5-7). The top-most gene sets in corresponding spots well agree for most categories. Solely the spot expressed in the testis sample reveals a difference between the two SOM approaches: On the one hand, the spot in the original SOM is associated with reproduction and related processes. In the gene set SOM, these sets are located in the two marginal spots on left side of the bottom edge as observed in the SOM portrait of testis (Figure 5-6). Spot 'F' in the gene set SOM on the other hand relates to cell differentiation and is expressed also by thymus, bone marrow and ovary samples as discussed above. In this sense, the testis sample is characterized by a spot ('F') assigning functions as differentiation, and additional specific spots containing special functions related to reproduction and spermatogenesis, for example.

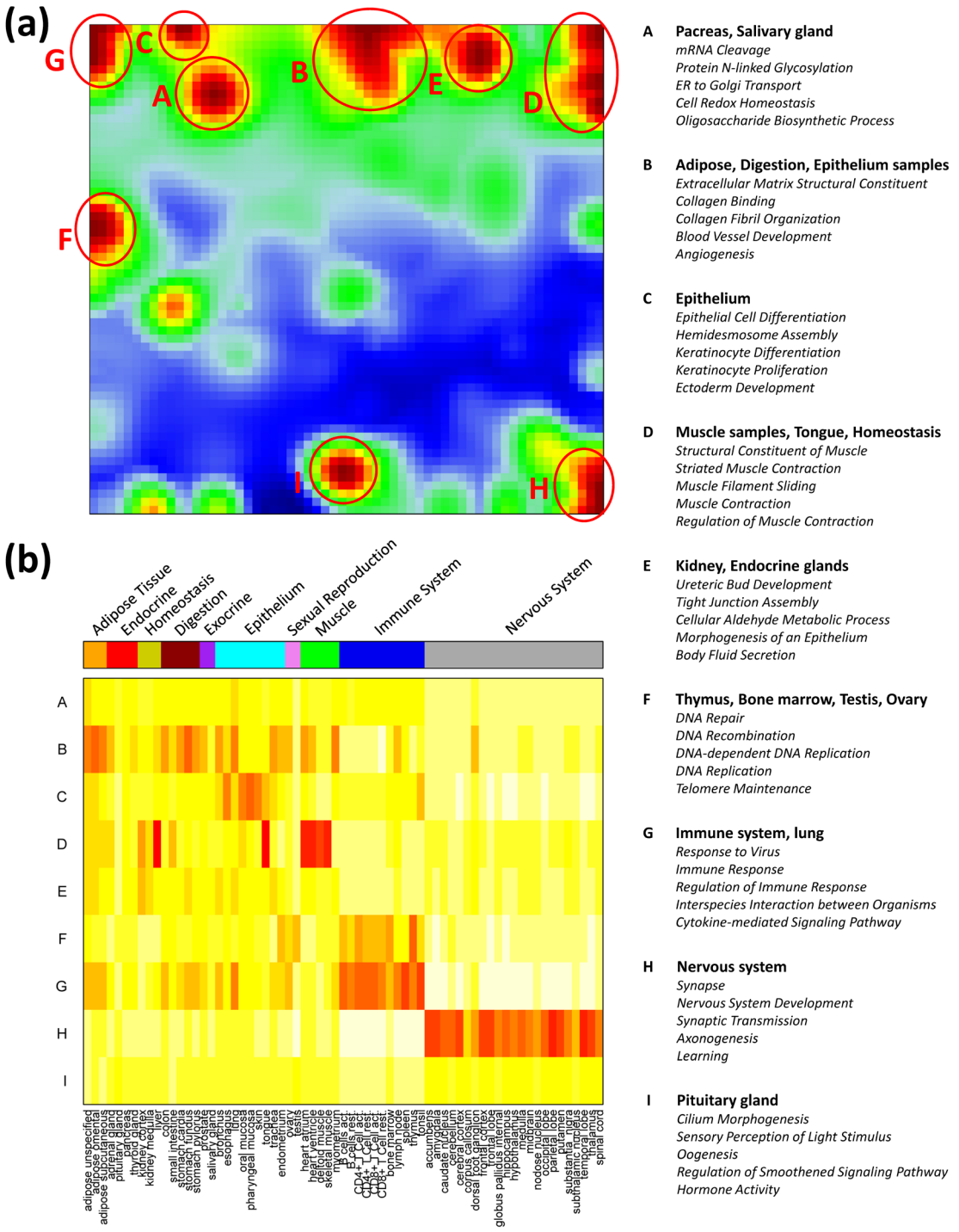


**Figure 5-7: The overexpression summary map (panel a) shows spots of overexpressed meta-genesets. Nine spots were identified using the 98-percentile criterion. Average GSZ scores of these spots for each sample are shown in the spot heatmap (panel b). The legend on the right assigns corresponding samples or categories to the spots. Additionally, the top-five overexpressed gene sets are given for each spot.**

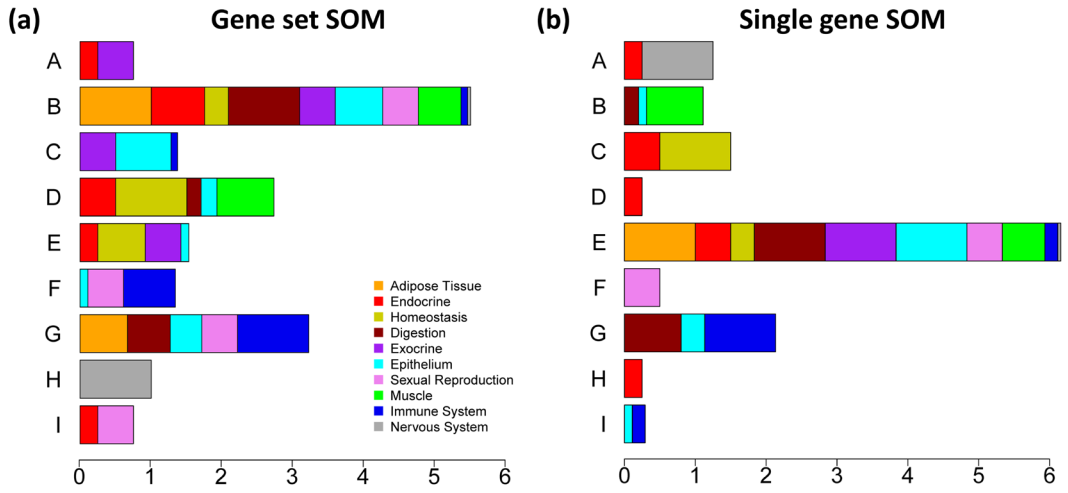


**Figure 5-8: The spot-abundance bar plots for gene set SOM (panel a) and single gene SOM (panel b) show the fraction of samples of each tissue category which exhibit a given spot. The total length of the horizontal bars characterizes the total abundance of the spots and the length of each colored region of the bars the abundance of this spot in one of the categories (see the legend for assignment).**

These findings indicate that the gene set SOM tends to generate more spots in each sample portrait than the respective gene based SOM.

The abundance bar plots in Figure 5-8 supports this observation. They visualize the relative frequency of appearance of each spot s='A', 'B',... in the samples of each tissue category c='adipose', 'endocrine',...:

$$x_{sc} = \frac{n_{sc}}{N_c} \tag{33}$$

where the numerator and denominator define the number of sample portraits $n_{cs}$ showing a particular spot and the total number of samples per tissue category $N_c$, respectively.

The stacked bar plots in Figure 5-8 give a first impression about the distribution of spot abundances. In both, gene set and single gene SOM, one spot is exhibited by all the tissue categories (spot 'B' and 'E', respectively). Also strongly category-associated spots are present, for example the nervous system spot 'H' and 'A', respectively. The individual spots in the gene set SOM are observed, however, in more tissue categories (see Table 7). For example, the immune system related spot 'G' is present in samples of 5 different categories in the gene set SOM, but only in 3 categories in the original SOM. In turn, also the number of different spots observed in the sample portraits of a particular tissue category is larger in many cases (Table 7): Epithelium samples for instance express six spots in gene set SOM ('B', 'C', 'D', 'E', 'F' and 'G', compare Figure 5-8), but only 4 in the original one ('B', 'E', 'G', 'I').

In summary, the gene set SOM exhibits a similar number of spots as the original gene based SOM. The spots in the gene set SOM are however more widespread, occurring in more

**Table 7: Spot and category abundances in gene set SOM and single gene SOM, respectively.**

| | Gene set SOM | Single gene SOM |
|---|---|---|
| **Nervous system** [a] | 2 spots | 2 spots |
| **Muscle** | 2 spots | 2 spots |
| **Epithelium** | 6 spots | 4 spots |
| **Testis** | 3 spots | 1 spot |
| **Immune system** | 4 spots | 3 spots |
| **Average spot occupancy** [b] | 3.8 categories per spot | 2.7 categories spot |

[a] Number of different spots observed in the respective SOM portraits.
[b] Average number of stacked segments per bar in Figure 5-8.

categories, each of them characterized by more individual spots. This potentially represents a sort of additive functional description of the samples, rather than the orthogonal expression modules identified in the gene based SOM (see chapter 2.7.1).

We generated the second level SOM of the gene set SOM to evaluate its discrimination power. This projection of the samples onto the two-dimensional SOM grid is shown in Figure 5-9. It well separates not only the tissue categories immune and nervous system, but also adipose tissues in the top left corner, and digestion and epithelium in the more central part of the map. Direct comparison with the original gene-based second level SOM (Figure 4-1a) reveals that the gene set-based second level SOM provides essentially the same discrimination with respect to the different tissue categories. The samples however cover a broader range in the map, leading to improved resolution of the formerly dense clusters.

In general, set-wise aggregation of single gene expression data into GSZ-scores summarizes expression values of functionally related genes. This supervised filtering step effectively removes genes without functional annotation. The gene set SOM thus provides enhanced classification capability compared to SOM based on single gene expressions. On the other hand, loss of information caused by the removal of not annotated genes might bias analysis results.
The principle of the gene set SOM approach can be transferred to other types of data while maintaining the discussed advantages. For example, sets of related features can be built according to chromosomal location and applied to transcriptome studies as well as next-generation sequencing data. Likewise, aggregation of proteins according to classes (e.g. hormones, toxins, enzymes) is another possibility in the context of proteome data.

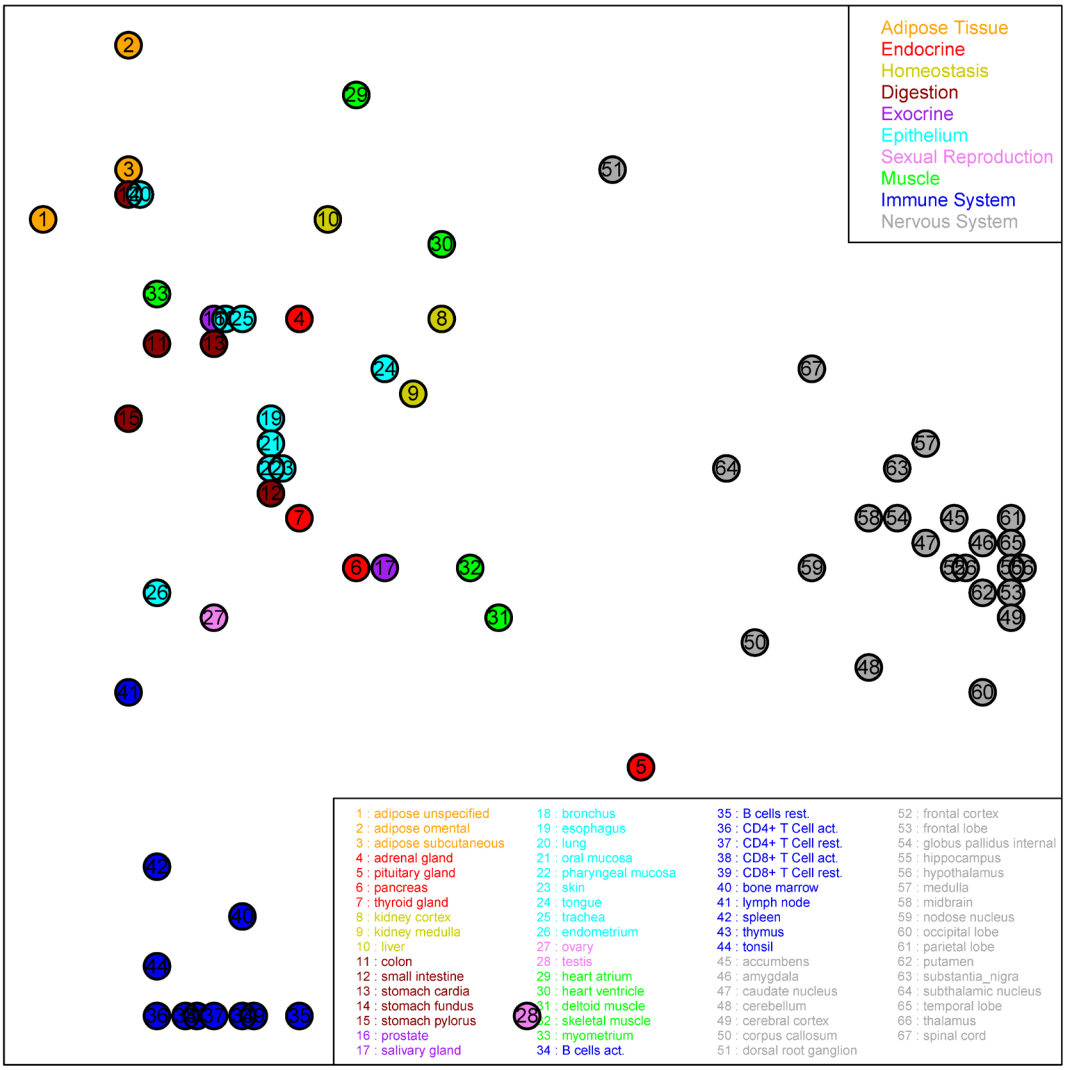


**Figure 5-9: Second level SOM of the meta-geneset expression states of all 67 samples with a resolution of 40x40 nodes.**

### 5.3.6 Summary

To extract the functional context of spot and meta-gene related lists of single genes we applied overrepresentation- and overexpression analysis, and a combination of both with respect to pre-defined gene sets of basically known functional impact. The mapping of overrepresentation of a selected gene set to the SOM mosaic provides a 'functional' map showing areas which are potentially relevant for this function. Alternatively, one can screen the degree of overrepresentation of a large number of gene sets in a selected meta-gene spot to discover its potential functional context. Both views provide a link between the tiles and/or spots of the SOM mosaic and their potential molecular function. Notably, they apply to all samples of the study due to the fixed mapping of single genes to the meta-genes. The gene set

enrichment approach combines both overrepresentation and overexpression analysis. It was applied to discover the functional context of the meta-gene overexpression spots in a sample specific fashion.

The tissue related spots of the SOM portraits typically contain enriched populations of gene sets corresponding to molecular processes in the respective tissues in most cases. This result supports the 'guilt-by-association' principle which states that co-expressed genes are likely to be functionally associated. It, in turn, implies the ability to define either new gene sets using selected SOM spots, or to verify and refine existing ones [WIRTH3].

The gene set SOM finally provides a complementary option to gene-based SOM analysis. Especially the use of GO sets entails the need for algorithms to handle the redundancy of the gene sets [109]. Both, high number of annotated sets, as well as strongly overlapping members are implied by the hierarchical structure of this ontology. The gene set SOM here represents an appropriate tool to deal with these difficulties and rearranges the gene sets for straightforward interpretation and for detection of overlapping functional themes.

# 6 Case studies

In this chapter, we present SOM analyses for different types of molecular biological data in form of case studies to illustrate strengths of the method in the respective applications. The examples are selected from different 'OMIC' realms, such as transcriptome, genome, methylome and proteome, to show the broad and flexible range of applications of machine learning in this context. Importantly, our SOM based analyses divide into method-specific and virtually method-unspecific tasks. The latter task comprises machine learning and, partly, clustering and similarity analysis which can be applied usually without special emphasis on the data type used. In contrast, the former method-specific tasks include data preprocessing and downstream analysis in terms of feature selection and functional analysis. Particularly, preprocessing requires special consideration of the particular method of measurement (e.g. mass spectrometry, microarray probe intensities or next-generation sequencing library preparation) to minimize the associated systematic biases in the data. The downstream analysis tasks address first of all functional interpretation of the observed single features and clusters of meta-features. Our examples were also selected with special emphasis to different and more general issues such as time series, class characterization and discrimination tasks. Table 8 provides an overview of the data sets studied.

## 6.1 Transcriptome data

### 6.1.1 Time series experiments: mining the yeast metabolic cycle

The yeast metabolic cycle (YMC) is one of the best studied model systems to discover basal rules of genomic regulation. Taking advantage of this knowledge, the YMC transcriptome is utilized to evaluate the SOM method with regard to extraction of information about dynamics of gene expression in time series experiments. Microarray data was obtained from Gene Expression Omnibus, accession number GSE9302. This dataset consists of 48 samples assessed with the Affymetrix Yeast Genome 2.0 arrays, measuring the expression of 5,900 genes of Saccharomyces cerevisiae (budding yeast). The data set comprises 48 measurements taken in intervals of 4 minutes (see [110] for details). This sampling covers four complete periods of the ~40-min continuous respiratory-reductive synchrony cycle of budding yeast. Our examination includes two independent analyses based on either the subset of the control cycle, consisting of the first 11 samples, or on the complete set of 48 samples covering four cycles: one control cycle and three subsequent cycles after treatment with phenelzine. This oxidase inhibitor is known to double the reductive phase of the YMC whereas the length of the oxidative phase is unchanged. It was chosen to study the regulatory mechanisms which lead to the increased period of the circadian clock [110].

**Table 8: Summary of data and SOM properties for the case studies presented.**

| *'OME' realm* | **Transcriptome (mRNA)** | | | | **Genome** | **Methylome** | **Proteome** |
|---|---|---|---|---|---|---|---|
| ***Experiment*** | Microarray study of human tissues in triplicates | Microarray study of the yeast metabolic cycle, no replicates | Toxication with DMSO and BaP, triplicated microarrays | Tumor subtype microarray studies: B-cell lymphoma; Glioblastoma; Prostate cancer | SNP microarray study of humans across the world | Tumor-vs.-Control study of human prostate cancer patients | Typing of algae *Prototheca* based on mass spectrometry |
| ***Own publications*** | [WIRTH1], [WIRTH3] | | | [HOPP1], [BINDER2] | [BINDER1] | | [WIRTH2] |
| ***Data source*** | 'Gene Expression Omnibus GEO', accession no. GSE7307, www.ncbi.nlm.nih.gov/geo/ | 'GEO', GSE9302 | Department for Environmental Immunology, Helmholtz-Centre for Environmental Research, Leipzig, Germany | 'GEO', GSE4475; 'The cancer genome atlas', http://tcga-data.nci.nih.gov; 'GEO', GSE4475 | 'Human Genome Diversity Project': http://hagsc.org/hgdp/files.html | Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany | Department for Proteomics, Helmholtz-Centre for Environmental Research, Leipzig, Germany |
| ***Data dimension (M x N)*** | 67 x 22,277 | 11 x 5,900 | 12 x 21,800 | 221 x 22,283<br>164 x 22,277<br>84 x 9,936 | 1,043 x 50,000 | 104 x 20,000 | 324 x 1,406 |
| ***Sample structure*** | 10 tissue categories | 11 individual samples | 3 treatments measured at 4 time points each | Individual samples | 7 geographical regions | 2 classes: normal, tumor | 6 genotypes |
| ***SOM size*** | 60 x 60 | 30 x 30 | 40 x 40 | 50 x 50<br>50 x 50<br>40 x 40 | 80 x 80 | 20 x 20 | 20 x 20 |
| $\langle n_k \rangle$ | 6.2 | 6.5 | 13.6 | 8.9<br>8.9<br>6.2 | 7.8 | 50.0 | 3.5 |
| ***Processing time (standard PC)*** | ~ 45 min | ~ 5 min | ~ 15 min | ~ 45 min<br>~ 30 min<br>~ 15 min | ~ 1 week | ~ 30 min | ~ 3 min |

Raw microarray data of the YMC was preprocessed as described for the human tissue data set (see chapter 2.3). The machine learning algorithm then assigns the 5,900 single genes to 30x30 = 900 meta-genes. Figure 6-1a shows the SOM portraits of the first 11 samples illustrating one 40min oscillation in YMC. The portraits are arranged in a circular way corresponding to the fashion of a metabolic cycle: the state of transcriptional activity is expected to be very similar in first ('t 1') and eleventh sample ('t 11'). Each of the portraits shows spots of overexpressed meta-genes revealing close relations of consecutive samples: The transition of overexpressed meta-genes is reflected in the trace of the red spots in the SOM portraits along the edges in counter clockwise direction. For example, the first sample 't 1' features three overexpressed spots with the predominant one located in the top left corner. Meta-genes, and thus associated single genes, located in this region reach their expression maximum in the first sample. This spot is also featured by the adjacent samples 't 11' and 't 2', but in less pronounced manner. The other two overexpressed spots are located in the top right and bottom left corner. The former is the residue of the larger red spot in sample 't 11' and disappears in 't 2', whereas the latter one is not present in 't 11' but grows to a major spot in 't 2'. In general, spots of overexpression shift along the edges and mostly take about three subsequent samples to emerge, reach the maximum and then disappear. Such sequential patterns reveal the intersection of gene expression modes, whereas the counter clockwise manner of transitions reflects the cyclic nature of gene expression in the YMC.
Inspection of all four cycles reveals an increased period in the oscillation after perturbation by phenelzine (Figure 6-1b-d): The first samples of the respective new cycle are clearly identified in 't 1', 't 12', 't 23', 't 35' and 't 48', showing virtually identical portraits. In original literature, the latter sample 't 48' is assigned to the fourth cycle [110]. The respective SOM portrait suggests a fifth cycle with this sample as starting point. However, the fourth cycle covers 3 (original literature: 4) samples more than the control cycle, implying elongation of the cycle after treatment.

The spot heatmap in Figure 6-2 shows the mean expression level of the six major spots identified for each time point in the control cycle. It shows the transition of (meta-) gene expression modes from the perspective of spot pattern: each spot features increased expression levels for at least five time points comprising increasing, maximum and decreasing parts of profiles. Furthermore each spot disappears with time and is followed by a new spot characteristic for a set of subsequent samples. The sequence of neighbored spots well agrees with their chronological order of appearance.

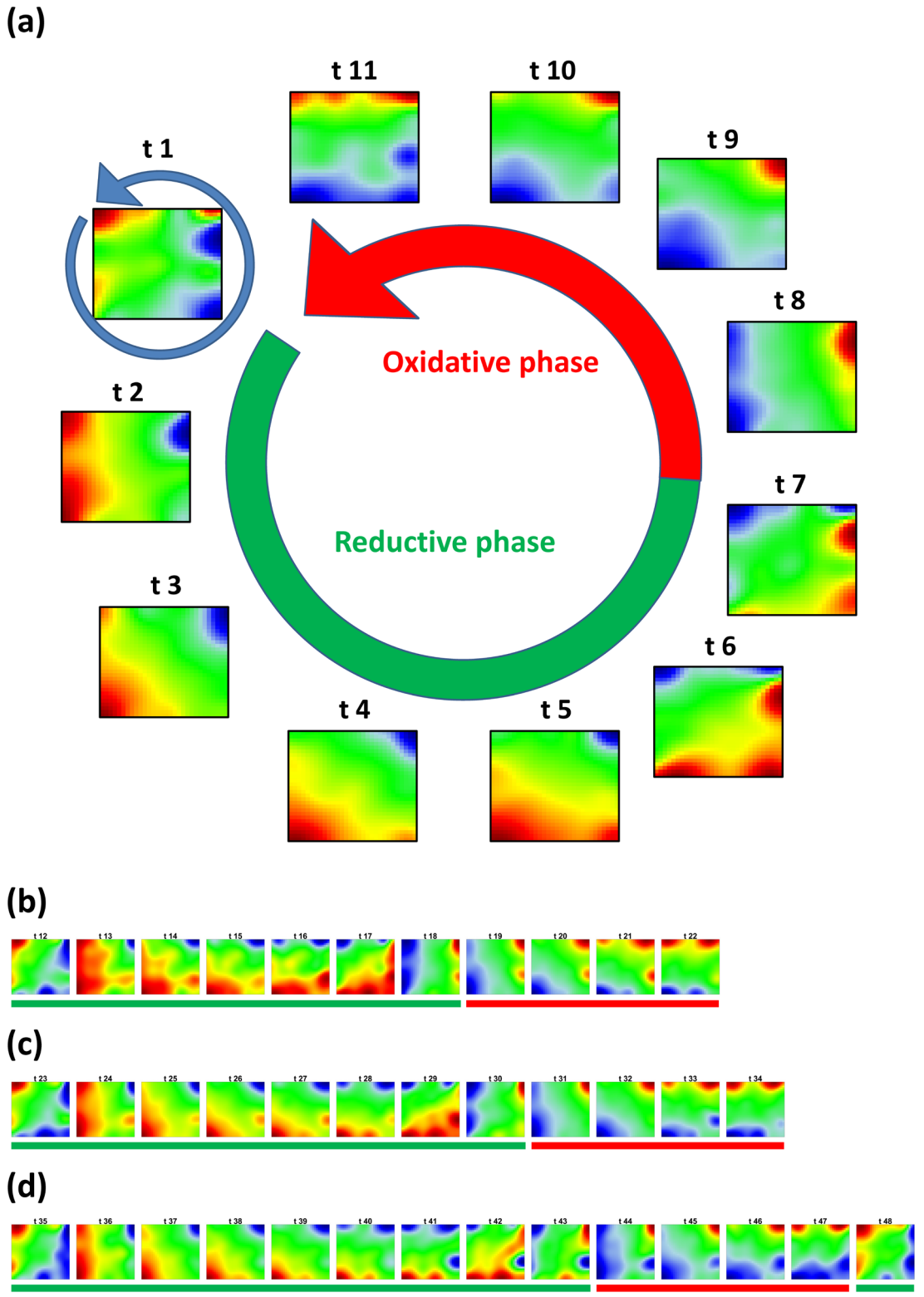


**Figure 6-1: SOM portraits of the YMC. The portraits of the control cycle (panel a) are arranged in circular grid according to reductive phase (samples 't 1' to 't 7') and oxidative phase (samples 't 8' to 't 11'). Portraits of treatment cycles (panels b-d) are arranged chronologically and labeled as reductive and oxidative phases by green and red bars, respectively. The spots of overexpressed metagenes shift in counter clockwise fashion along the edges of the portraits as indicated by the blue arrow at the first sample in panel a.**

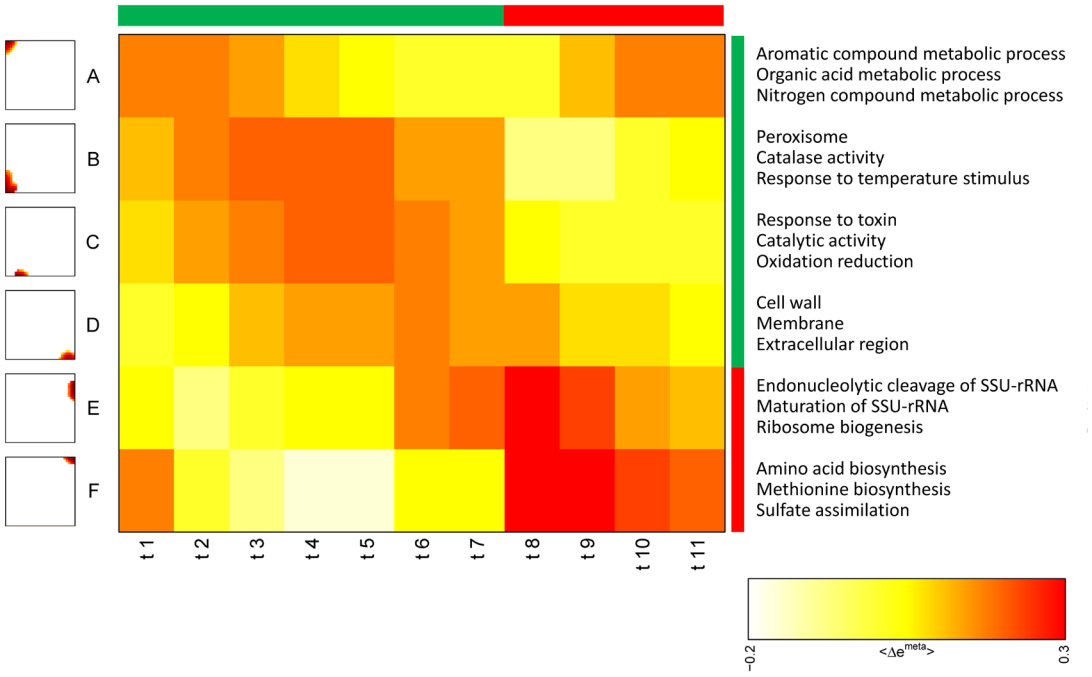


**Figure 6-2: Overexpression spot heatmap of selected spots. Each column refers to one sample of the control cycle. Spots are depicted by small mosaics in the left, and associated with overrepresented gene sets in the right legend.**

The genes associated with each spot shown in Figure 6-2 are separately analyzed for overrepresentation using the HG-test. The top-three overrepresented gene sets are given in the right legend in Figure 6-2. In accordance with Tu et al. [111], we found gene sets characteristic for the reductive phase: 'peroxisome', 'response to temperature stimulus' (spot 'B'), as well as gene sets related to transport of sugars (spot 'B'), metabolic process, chromosome (spot 'C'), and cell wall, wall assembly and membrane (spot 'D'). For the oxidative phase we found the gene sets ribosome (spot 'E'), sulfate assimilation (spot 'F'), and amino acid biosynthesis / metabolism and related sets (spots 'F' and 'A'). Note the specificity of spot 'E' for targeting RNA cleavage, maturation and processing.

The oscillatory character of the YMC data set is not only reflected in cyclic meta-gene expression profiles as described above. A complementary view is provided by overexpression analysis of selected gene sets, known to be activated in the reductive (Figure 6-3, panel a) and oxidative phases (panel b). The GSZ scores were calculated using total lists of differential gene expression in the samples. Obviously, activation and deactivation of gene sets and thus the related biological functions, do not follow an abrupt ON/OFF-switching process but rather a smooth transition passing cyclic increase, maximum and decrease of the respective GSZ-scores. For example, genes associated to 'peroxisome' reach their maximum of activity in sample 't 2' and the minimum in 't 8'. Intermediate time points of measurement reflect a

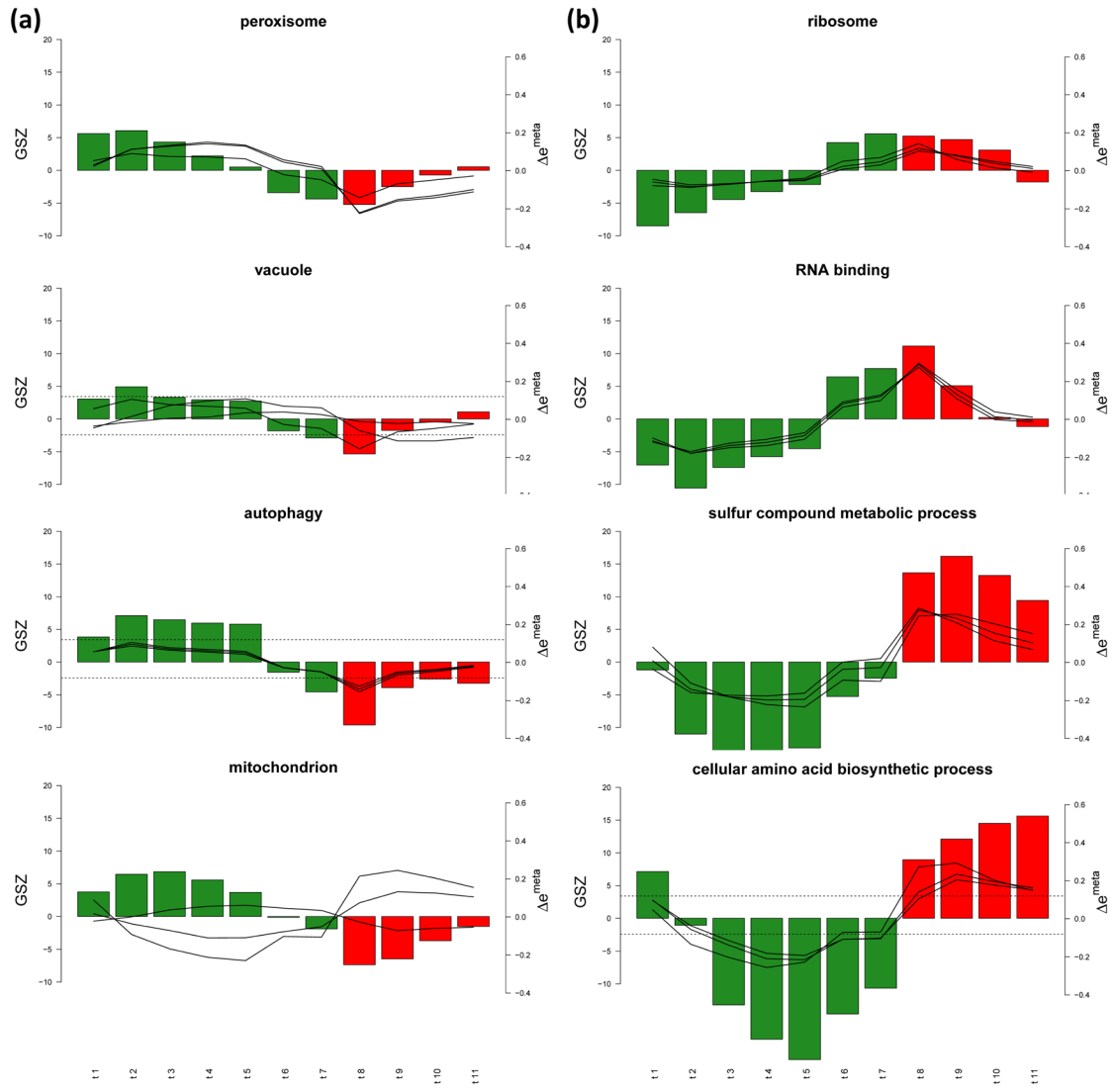


**Figure 6-3: Profile of GSZ scores for gene sets reported to be activated in reductive (panel a) and oxidative phase (panel b) [111]. The inserted curves show the expression profiles of the top-three meta-genes with strongest enrichment of the respective gene set.**

series of smooth transition states between the two extremes. Further, also a shift of the maximum position of the GSZ score is observed for the eight gene sets examined where the phase shift covers the complete oscillation cycle.

Taking together, the oscillating characteristics of the YMC expression data could be easily verified with special regard to either meta-genes, spots of meta-genes or functional gene sets.

For a sample centered view we generated second level SOMs for the control cycle and all four cycles, respectively. Figure 6-4a shows the second level SOM of the first 11 samples of YMC with a resolution of 7x7 nodes. In analogy to the circular patterns in the SOM portraits, second level SOM arranges the samples along a circle in clockwise direction. Second level SOM of the

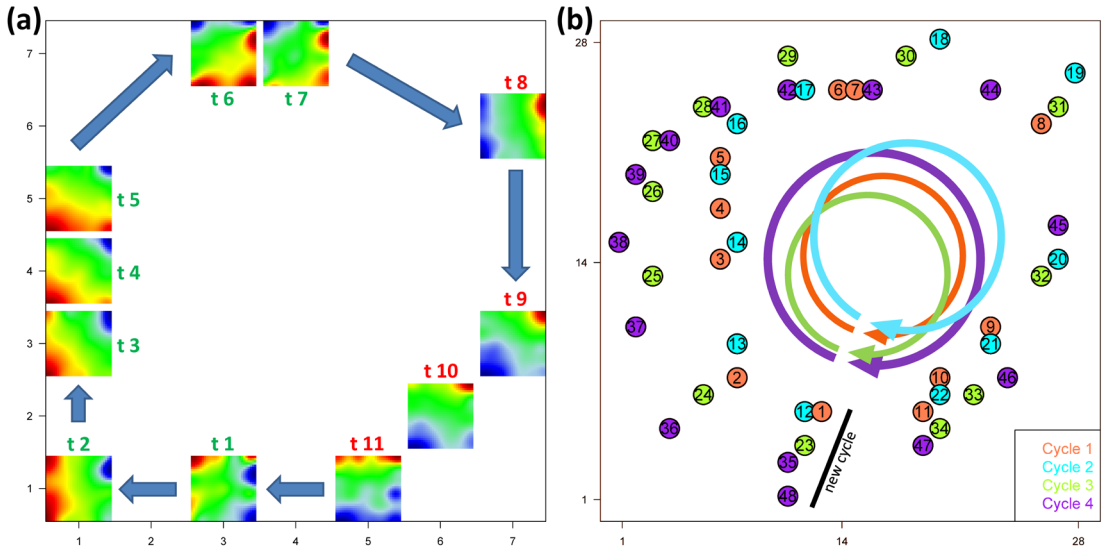


**Figure 6-4: Second level SOM of the meta-gene expression states: samples of the first control cycle (panel a) arrange in circular order analogous to Figure 6-1a. Mapping of all 48 samples (panel b) shows arrangement of the four cycles along concentric circles as indicated by the arrows.**

complete set of four cycles studied (Figure 6-4b, 30x30 nodes) reveals a series of concentric circles, in agreement with SOM portraits as discussed above. Further, elongation of YMC period transforms into slightly increased diameters of the circles especially in the third and fourth cycle. This sample representation provides an elegant way to visualize sample development in terms of trajectories in time series or cell development experiments in general.

### 6.1.2 Discovering time and dose effects: gene expression after exposure to toxins

Simultaneous evaluation of dose-dependent treatments in parallel time series experiments is a popular experimental design. In the present study, effect of toxication of murine hepatocytes was analyzed at four time points, 2h, 4h, 12h and 24h after exposure to dimethyl sulfoxide (DMSO) and benzo-a-pyrene (BaP). The cytotoxin BaP was applied in relatively high (5µM) and low (0.05µM) concentrations. This cyclic compound is an environmental contaminant, mainly arising from combustion of organic substances. It is found, e.g., in cigarette smoke or motor vehicle emissions, but it can also be detected in grilled foods. BaP is known to act with high toxic, mutagenic and carcinogenic activity [112]. It has been studied recently [113, 114].
Varying phenotypic responses of the cells are observed after treatment with BaP and DMSO [112, 115, 116]. Whereas DMSO causes small phenotypic effects, BaP treated cells observably suffer from toxication. Whereas cells regenerate after treatment with low BaP dose, high BaP dose causes death of most of the treated cells. To evaluate effects of exposure to BaP on transcriptional level, the samples were assessed using Affymetrix Mouse 1.0 ST arrays,

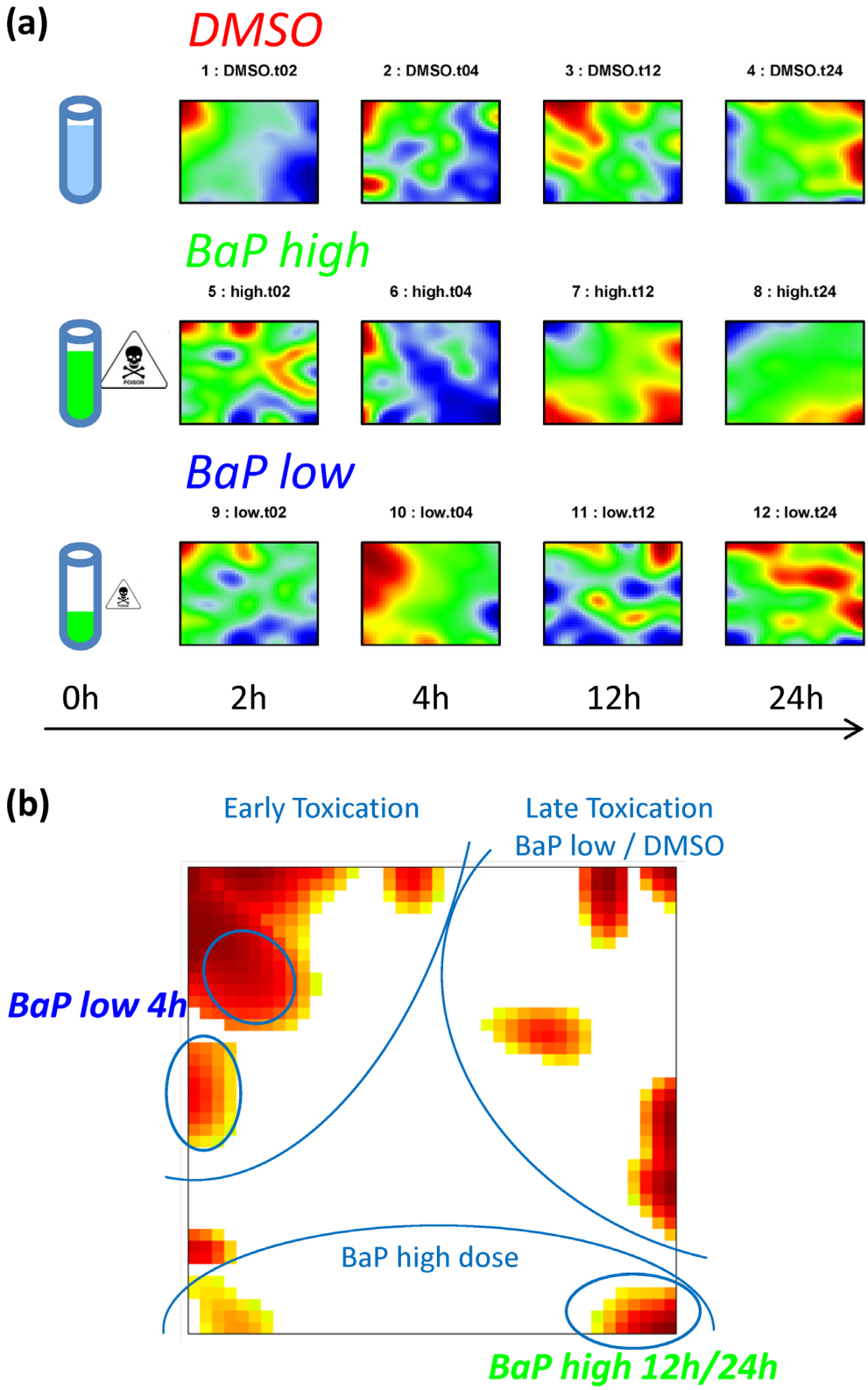


**Figure 6-5: SOM portraits of the 'DMSO-', 'BaP high'- and 'BaP low'-series, arranged according to treatment and elapsed time (panel a). Spots identified in the integral overexpression map (panel b) can be divided into three major groups comprising meta-genes specific for early toxication, late toxication ('BaP low' and 'DMSO'), and 'BaP high', respectively.**

measuring expression levels of 21,799 genes. This data set was preprocessed using RMA summarization and normalization [117] and subsequently transformed into logFC expression values as described above. These data are used to generate a SOM with resolution of 40x40 nodes. The obtained SOM portraits are shown in Figure 6-5a. At the first point of measurement, all portraits of the three treatments share one common overexpression spot in the top left corner, indicating a joint treatment independent mechanism. The portraits referring to BaP-treatment show additional spots, one in shared manner. SOM portraits of 'DMSO' and 'BaP high' samples slightly changed 4h after treatment, whereas the 'BaP low-portrait' shows a larger spot overexpressed in this sample. In the late stage of toxication after 12h and 24h, SOM portraits diverge in a dose-dependent fashion. Note that the 'BaP low' portrait shares similarities with 'DMSO' 12h and 24h after treatment.
Figure 6-5b shows the overexpression spot map displaying spots after applying the 98-percentile criterion. The overexpression spots can be roughly classified into three major groups: spots observed in 2h- and 4h-portraits in top left range, spots of the 'DMSO' and 'BaP low' portraits in the top right range and spots associated with the 'BaP high' portraits along bottom edge.
Two spots highlighted in Figure 6-5b are of major importance to understand molecular mechanisms caused by BaP: the 'BaP low 4h' specific spots in upper left part of the SOM, as well as the spot in bottom right corner, overexpressed in 'BaP high' at 12h and 24h after treatment. Genes associated to the latter one are supposed to support the necrosis in a direct or indirect way. The cells typically die off in the respective stages. Contrarily, genes associated to the former spot putatively cause an answer to BaP low dose treatment and initiate regeneration of the cells. The time point 4h after exposure seems crucial for regeneration of the cells in the 'BaP low' series. Note that cells from 'BaP high' series start to die at this time point.

Another interesting observation can be extracted from the variance data of the expression meta-states shown in Figure 6-6a. The four 'DMSO' samples are the less variant among the different treatments. 'BaP high' reveals strongly increased variance of the expression states at 12h and 24h whereas 'BaP low' shows a maximum at 4h'. The variability of the expression states seems to be related to necrosis and regeneration processes in the cells, respectively.

The second level SOM in Figure 6-6b shows the trajectories of the time series in the two-dimensional map. Early time points (2h and 4h) mainly gather in the center of the map. 'BaP low' reveals a specific short-time response (4h) before this trajectory turns into the same direction as DMSO, presumably due to regeneration of the cells. In contrast, 'BaP high' seems to respond with a longer delay but then turns into another direction compared with 'BaP low' and 'DMSO'. The mutually orthogonal direction of these trajectories suggests independently regulated sets of genes: Genes activated to regenerate the cells from low and moderate

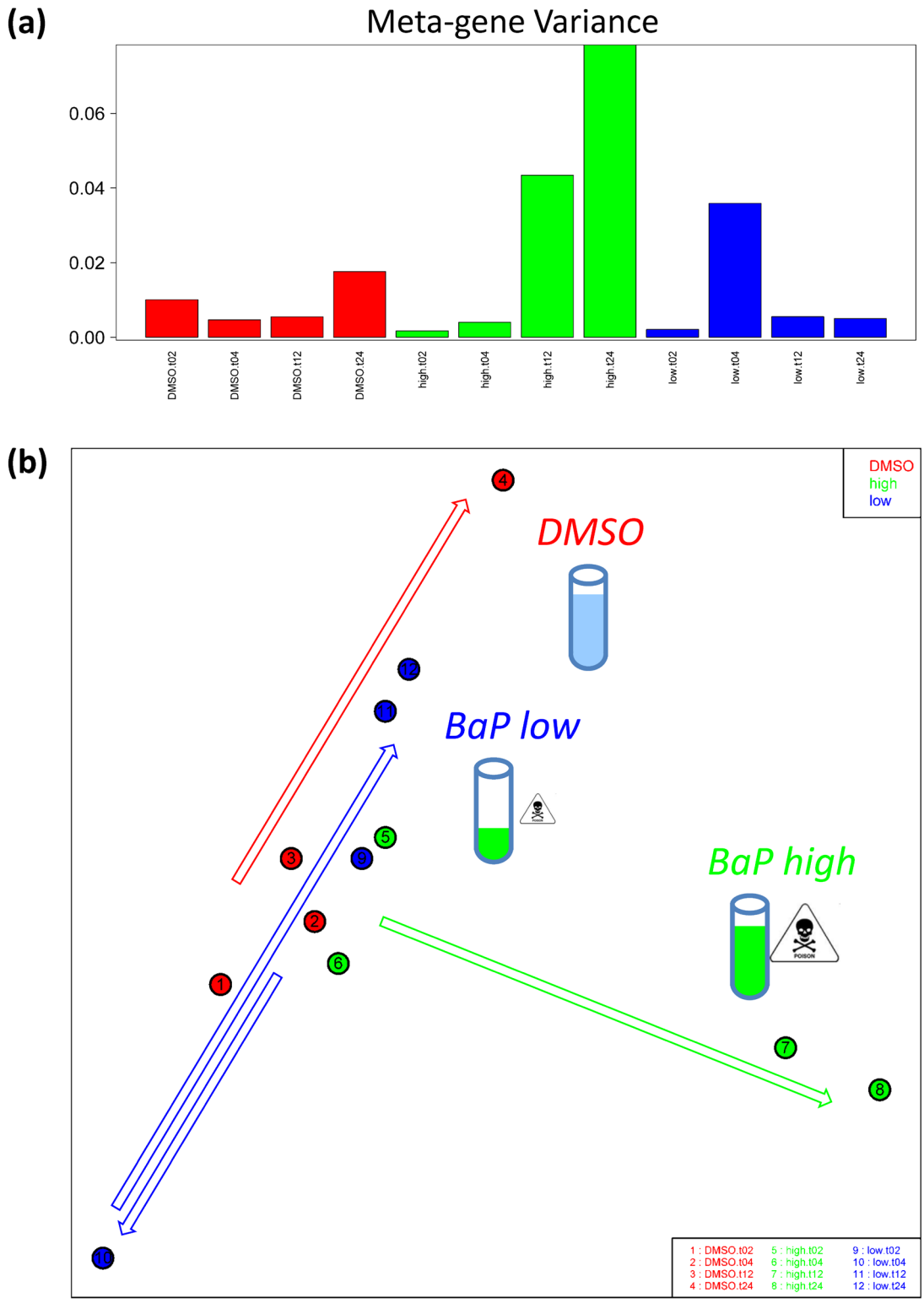


**Figure 6-6: Meta-gene state based analyses of the BaP study: Variance plot (panel a) reveals augmented activity in 'BaP low 4h' and 'BaP high 12h/24h'. The second level SOM (panel b) illustrates trajectories of the treatment progressions. Different treatment series are highlighted using arrows and according pictograms.**

**Table 9: Enriched GO gene sets in 'Bap low 4h' and 'BaP high 12h/24h' samples. The gene sets are arranged according to different functional groups. Key processes are indicated with bold letters.**

| BaP low 4h | BaP high 12h/24h |
|---|---|
| **DNA repair** | **Response to organic cyclic compound, activation of cyclase** |
| **Cell cycle, cell division, mitosis**, DNA replication, nucleotide binding, chromatin modification | **Glutathione metabolic process, glutathione transferase**, dopamine binding, response to cAMP, response to cytokine stimulus |
| Protein binding, ATP binding, metal ion binding | Response to calcium ion, sodium ion transport, response to stimulus |

toxication levels ('DMSO' and 'BaP low'), and genes activated (or repressed) in necrosis. Hence, the second level SOM supports previous findings: Firstly, it confirms the crucial role of the 'BaP 4h' sample, representing a sort of turning point from pollution to regeneration. Secondly, it reflects the irreversible effect of cells after contamination in the 'BaP high' treatment series.

In a final analysis step, the functional context of the observed expression changes is evaluated using GSZ-statistic. Gene set enrichment analysis was applied utilizing a collection of 1,933 GO gene sets[10]. It turned out that hepatocytes treated with low BaP dose exhibit activation of 'DNA repair pathway' accompanied by intensive proliferation ('cell cycle', 'cell division', 'mitosis' etc., see Table 9) which reflect regeneration of the liver cells. Contrarily, in late stage of 'BaP high' treatment, oxidative stress (glutathione related processes) in combination with strongly activated metabolism ('response to stimulus' and related pathways) accompany cell death. Additionally, pathways 'response to cyclic compound' and 'activation of cyclase' reflect the attempt to degrade BaP in the cells. These findings directly link the transcriptional activity to observable phenotypic effects, namely regeneration and cell death in 'BaP low' and 'BaP high' samples, respectively.
In summary, the SOM method provides a suited framework for analysis of time series data under varying treatment. In a first step, SOM portraits of the toxication study allow to identify samples (and spots) with major impact for the behavior of the cells. These finding are verified and further supported in secondary similarity and functional analyses.

[10] GO annotation derived from Ensembl data base [137]

### 6.1.3 Disentangling and characterizing subtypes of human cancer

In the last years, large-scale studies were undertaken with the intention to extract reliable molecular profiles of cancer cells and to derive underlying regulatory mechanisms. This ambition is hampered by the large biological variability of the tumor cells, but also by ambiguous and partly unknown subclasses of the cancer types. Here one can take advantage of the SOM [HOPP1], which enables characterization of the expression landscapes on the individual level of patient samples. In this subchapter we demonstrate the application of the SOM pipeline to characterize cancer subtypes. It will be shown, that each of the obtained expression modules can be interpreted in terms of distinct biological processes, either utilizing the GO annotation to derive functional context of the subtypes, or utilizing sets of genes published in recent assessments of cancer samples. Three publicly available data sets were chosen as examples:

*B-cell lymphoma (BL):* Microarray data are available under GEO accession number GSE4475 (220 Affymetrix HG-U133 arrays). This study used biopsy specimens of mature aggressive B-cell lymphoma in which at least 70 percent of all cells were tumor cells. The classification of lymphoma subtypes and sample assignments are used as given in Hummel et al. [118]: Of all 220 lymphomas, 44 were assigned to the mBL (molecular Burkitt's lymphoma) signature and 128 to non-mBL signature. 48 cases form an intermediate group, representing the transition zone between the mBL and non-mBL groups.

*Glioblastoma multiforme (GBM)*: Raw intensity data were downloaded from 'The Cancer Genome Atlas (TCGA)' portal[11]. The study comprises 153 tumor and 11 normal brain tissue specimen hybridized on Affymetrix HT-HG-U133A arrays. The samples were assigned to the GBM-subtypes 'mesenchymal' (MES, 50 samples), 'proneural' (PN, 45), 'neural' (NL, 26) and 'classical' (CL, 32) according to Verhaak et al. [119], and to normal healthy brain (N, 11 samples) for comparison. The latter specimens were taken from adjacent brain tissue of glioblastoma patients.

*Prostate cancer progression (PCP)*: Microarray data are available under GEO accession number GSE 6099 (84 non-commercial spotted Chinnaiyan Human 20K Hs6 arrays). The original evaluation by Tomlins et al. [120] addresses the molecular mechanisms associated with gene expression changes in the course of prostate cancer progression using laser-capture microdissection of 84 specific cell populations taken from 44 individuals. Five stages of cancer progression are captured in this study, ranging from benign prostatic hyperplasia (BPH, 22 samples) and prostatic interepithelial neoplasia (PIN, 13) to low-grade (PCA_low, Gleason

[11] http://tcga-data.nci.nih.gov

score 3, 12 samples), high-grade (PCA_high; Gleason score 4-5, 20 samples) and metastatic (MET, 17 samples) prostate cancer.

Raw probe intensity values of Affymetrix arrays (BL and GBM) were calibrated and summarized into one expression value per probe set using the hook method [56, 57]. For the customized arrays (PCP), preprocessed expression data were downloaded. Subsequent preprocessing was performed as described for the human tissue study in chapter 2.3, comprising quantile normalization, transformation into log10-scale and centering with respect to the mean expression level of each gene (differential expression). A SOM was then generated for each cancer data set in independent training runs. The SOMs for BL and GBM consist of K=50x50 = 2,500 meta-genes, for prostate cancer K=40x40=1,600 meta-genes.
Panel a of Figure 6-7 to Figure 6-9 portray the meta-gene expression landscape of lymphoma (BL), glioblastoma multiforme (GBM) and of prostate cancer (PCP), respectively. The shown mean SOM portraits of each class are calculated by averaging the expression values of each meta-gene over all class members. This averaging cancels out individual, highly fluctuating features on one hand. On the other hand, it amplifies consistent and class-specific features. The SOM portraits are arranged according to the previously published classifications into subtypes or progression stages [118–120] and shown in log-FC and loglog-FC color scale.
The expression portraits in log-FC scale reveal a handful of over- and underexpression spots which mostly characterize different cancer subtypes and stages in specific fashion. For example, the mBL and non-mBL subtypes (Figure 6-7a) are characterized by two spots in opposite corners of the map where one is overexpressed and the other one is underexpressed in mBL and vice versa in non-mBL subtype. These subtype-specific spots collect highly populated, variable and resolved meta-genes (see [HOPP1]). The mean SOM portraits of the four glioblastoma-subtypes (Figure 6-8a) are however more diverse: Only the portraits of the MES-subtype and of the N-reference show one specific overexpression spot whereas the PN-, CL- and NL-subtypes are characterized by two or three specific spots per subtype. The stage-related portraits of prostate cancer progression (Figure 6-9a) show analogous properties. Parts of the spots are observed in more than one PCP-stage. As a rule of thumb the spots of subsequent stages and also of the final MET- and of the initial BHP-stages tend to overlap. In consequence, the stage-specific spot pattern ‘rotates’ along the border of the map in clockwise direction with progressing cancer.
The loglog-FC-scale portraits feature more detailed information, enabling to identify finer, more subtle differences between the subtypes. For, example the mean loglog-FC maps of the MES- and PN-subtypes of GBM resemble each other like film positives and negatives, i.e. overexpressed red regions in the MES-portrait largely convert into underexpressed blue regions in the PN-portrait indicating a strongly anti-correlated expression pattern in both subtypes.

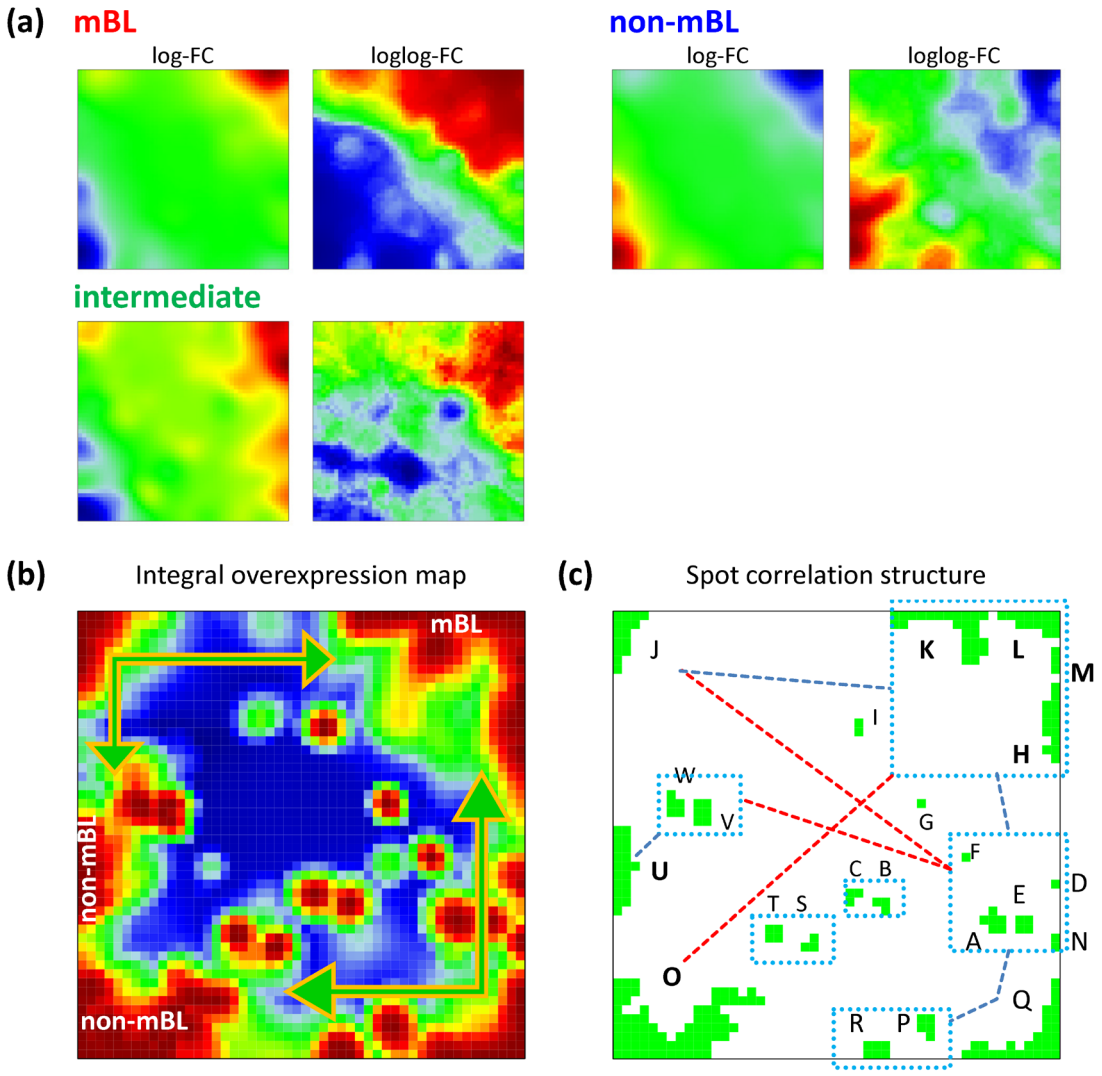


**Figure 6-7: SOM gallery of Burkitt's lymphoma: Representative SOM portraits of the three subtypes are calculated as mean meta-gene states averaged over all samples of each class and shown in standard log-FC and smooth log log-FC scale (panel a). The overexpression map (panel b) links opposite spots at bottom left and top right corners to the non-mBL and mBL subclasses, respectively. Individual spots are defined by the 98-percentile criterion and assigned by capital letters (panel c). The blue rectangles include highly correlated spots (r>0.7). The blue and red dashed lines connect correlated (0.4< r<0.7) and anti-correlated (r<-0.6) spots, respectively.**

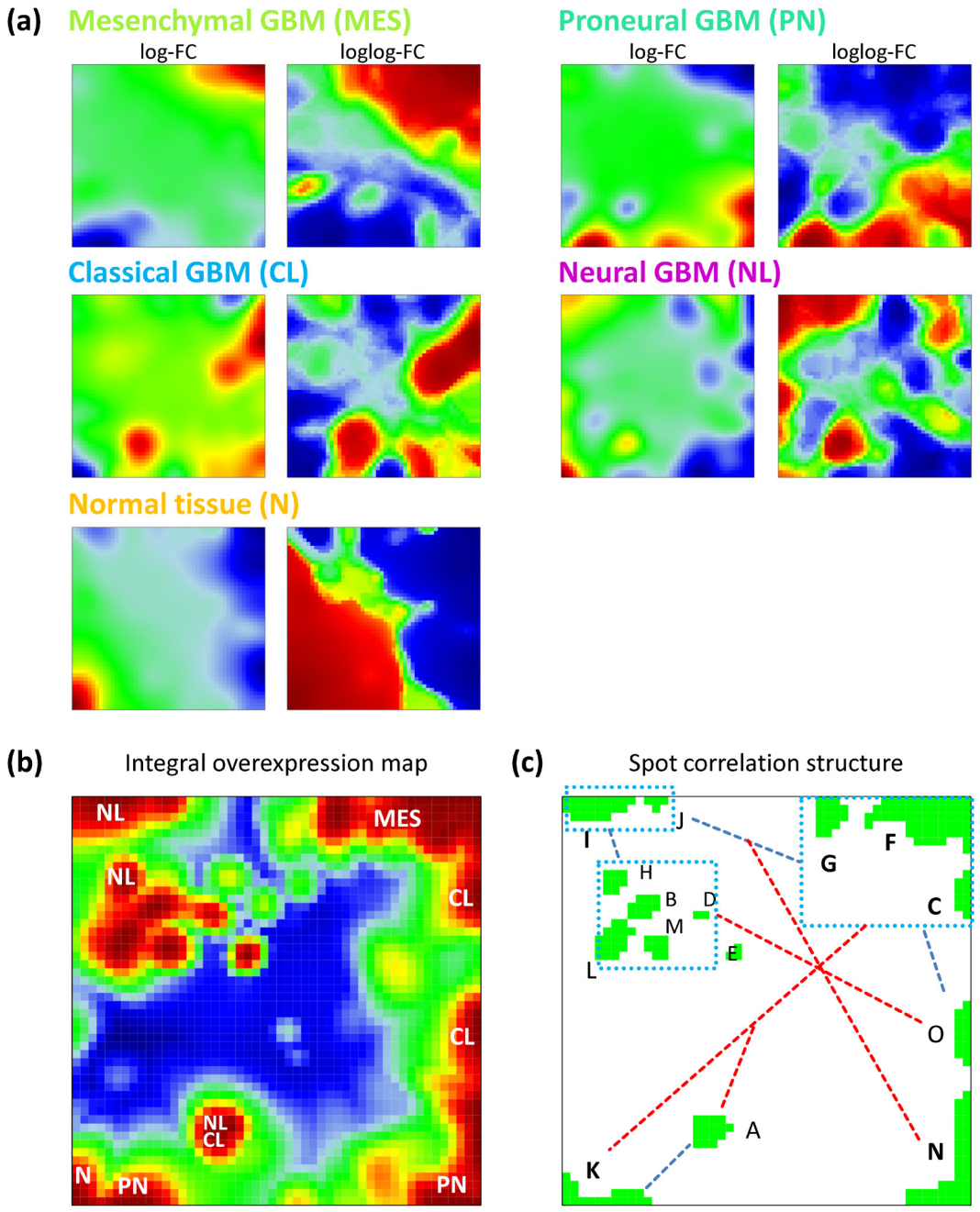


**Figure 6-8: SOM gallery of Glioblastoma multiforme. See legend of Figure 6-7 for details.**

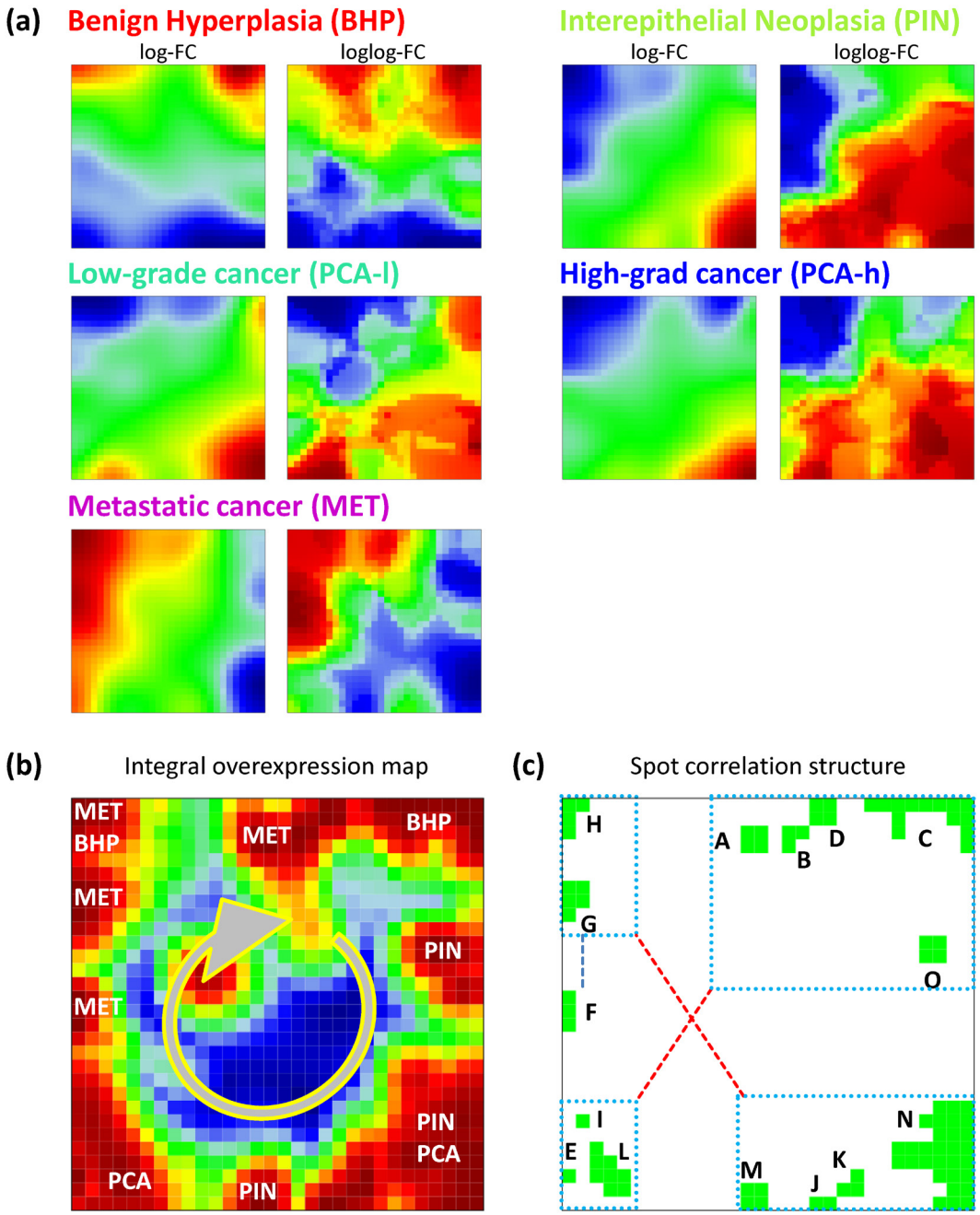


Figure 6-9: SOM gallery of prostate cancer. See legend of Figure 6-7 for details.

A global analysis of sample- or subtype-similarities might miss subtle effects due to individual properties of small groups of genes. Since such details are captured in the SOM portrait patterns, especially spots of overexpressed meta-genes are capable to resolve both cancer subtypes (see spot assignments in panel b of Figure 6-7 to Figure 6-9) but also effects in individual samples (see [HOPP1] for detailed spot discussions).
Panels c of Figure 6-7 to Figure 6-9 visualize the pairwise correlation strengths between the individual spots detected. Highly correlated spots are included into dotted rectangles or connected by blue dotted lines whereas red dotted lines indicate anti-correlation. For example, overexpression spots 'H', 'K', 'L' and 'M' are typical for the mBL subtype and feature strongly correlated expression profiles (Figure 6-7c). Those spots however are strongly anti-correlated to spot 'O', located in opposite corner and characteristic for the antagonistic non-mBL subtype. A similar correlation structure can be observed for the GBM-SOM (Figure 6-8c). The spots in the SOM of PCP shown in Figure 6-9c feature the most pronounced and unambiguous correlation pattern. The four corners of the map are occupied by each strongly correlated groups of spots, which in turn are strongly mutually anti-correlated.
Analogous analysis of spot assignment and correlations was also performed for underexpression spots with similar results [HOPP1]. Position and size of most of the detected underexpression spots agree with the position and size of the overexpression spots, indicating overexpression of the respective meta-genes in part of the samples changes into underexpression in other samples.

Gene set overrepresentation analysis was performed to evaluate enrichment of GO gene sets in the overexpression spots. Based on the functional context of the overrepresented sets obtained, a short notation was assigned to each of the spots (see left part of Figure 6-10). Selected spots the cancer SOMs are related to processes generally associated with cancer physiology such as inflammation (BL spot 'O'; GBM spot 'F') and cell division (BL: 'K'; GBM: 'N'). The right part of Figure 6-10 depicts the GSZ-profiles and the overrepresentation maps of the two gene sets 'inflammatory response' and 'cell division'. The profiles clearly reflect the fact that the respective processes are selectively activated and de-activated in a subtype-specific fashion. For example, inflammatory response is activated in the non-mBL and MES-GBM subtypes. The respective gene set population maps reveal that the associated genes accumulate in the regions of spots overexpressed in the different subtypes.
'Inflammatory response' and 'cell division' are not among the leading gene sets of any of the spots in PCP (Figure 6-10c). The respective GSZ-profiles however show that 'inflammatory response' is selectively activated in the BHP- and MET-stages whereas 'cell division'-genes are overexpressed in MET-samples only. The overrepresentation maps of these sets indicate that the respective genes accumulate in the regions of multiple spots.

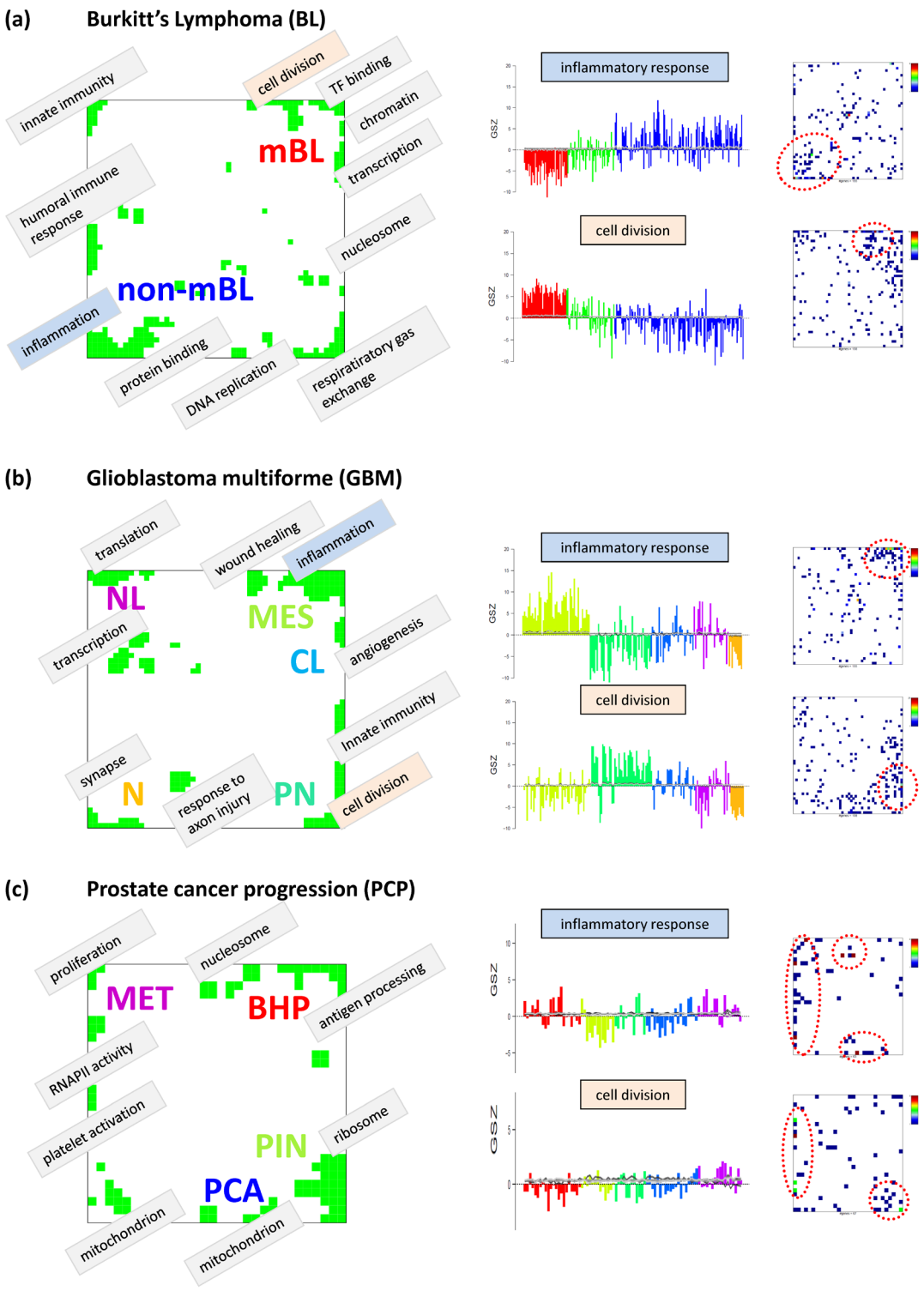


**Figure 6-10: Gene set enrichment analysis of BL, GBM and PCP (panels a, b and c, respectively). Left part: The overexpression map assigns the functional context of the most abundant spots, the subtypes are labeled beside specific spots. Right part: GSZ-profile and overrepresentation map of the gene sets 'inflammatory response' and 'cell division'. The red dotted ellipses in the maps indicate strongest enrichment.**

Special sets of genes reported to be regulated in designated subtypes or stages are evaluated for coincidence with the overexpression spots. Published signature gene sets of the subtypes can thus be directly compared with the spots of the cancer SOMs. In particular, sets of genes up- and downregulated in mBL were taken from Hummel et al. [118], four GBM subtype specific sets from Verhaak et al. [119] and four PCP stage related sets from Tomlins et al. [120]. For these gene sets, GSZ-profiles and overrepresentation maps were generated. Figure 6-11a shows the 'mBL up' and 'mBL down' sets, which clearly show a bimodal behavior in the mBL and non-mBL types. The intermediate BL subtype however remains unresolved. Notably the mapping of 'mBL up' and 'mBL down' genes in the SOM resembles the overexpression spot patterns of the mean mBL- and non-mBL-portraits.
The signature sets of GBM subtypes (Figure 6-11b) also confirm specific overexpression in the respective subtype and underexpression in the remaining three subtypes of GBM. The NL-specific signature shows overexpression also in the healthy brain tissue which was not taken into account while extracting specific signature genes [119]. Again, overrepresentation maps of the signature sets reveal that genes of each of the sets accumulate in the spots of subtype-specific overexpression identified in the mean SOM portraits. The signature genes of the PN- and CL-subtypes yet distribute over more than one overexpression spot. They obviously belong to different functional modules of co-expressed genes.
Finally, PCP signature are associated with different functional concepts such as 'glutathione metabolism' (specifically overexpressed in BHP), 'androgen signaling' (overexpressed in PIN and PCA_low), 'protein biosynthesis' (overexpressed in PIN and PCA) and 'cell cycle' (overexpressed in MET) [120]. Genes from these sets feature the expected GSZ-profiles and accumulate within the subtype-specific overexpression spots.

Summarizing, self-organizing maps were used to process expression data of B-cell lymphoma, glioblastoma multiforme and prostate cancer. The cancer subtypes were characterized in terms of about a dozen of overexpression spots, which can be easily assigned to their functional context using gene set enrichment analyses. This enables data driven generation of hypotheses, but also validation of subtype classifications and corresponding signature gene sets. In the cases presented, GSZ-profiles and gene set maps confirm the class-specific over- and underexpression modules defined by independent statistical analyses in the original papers.

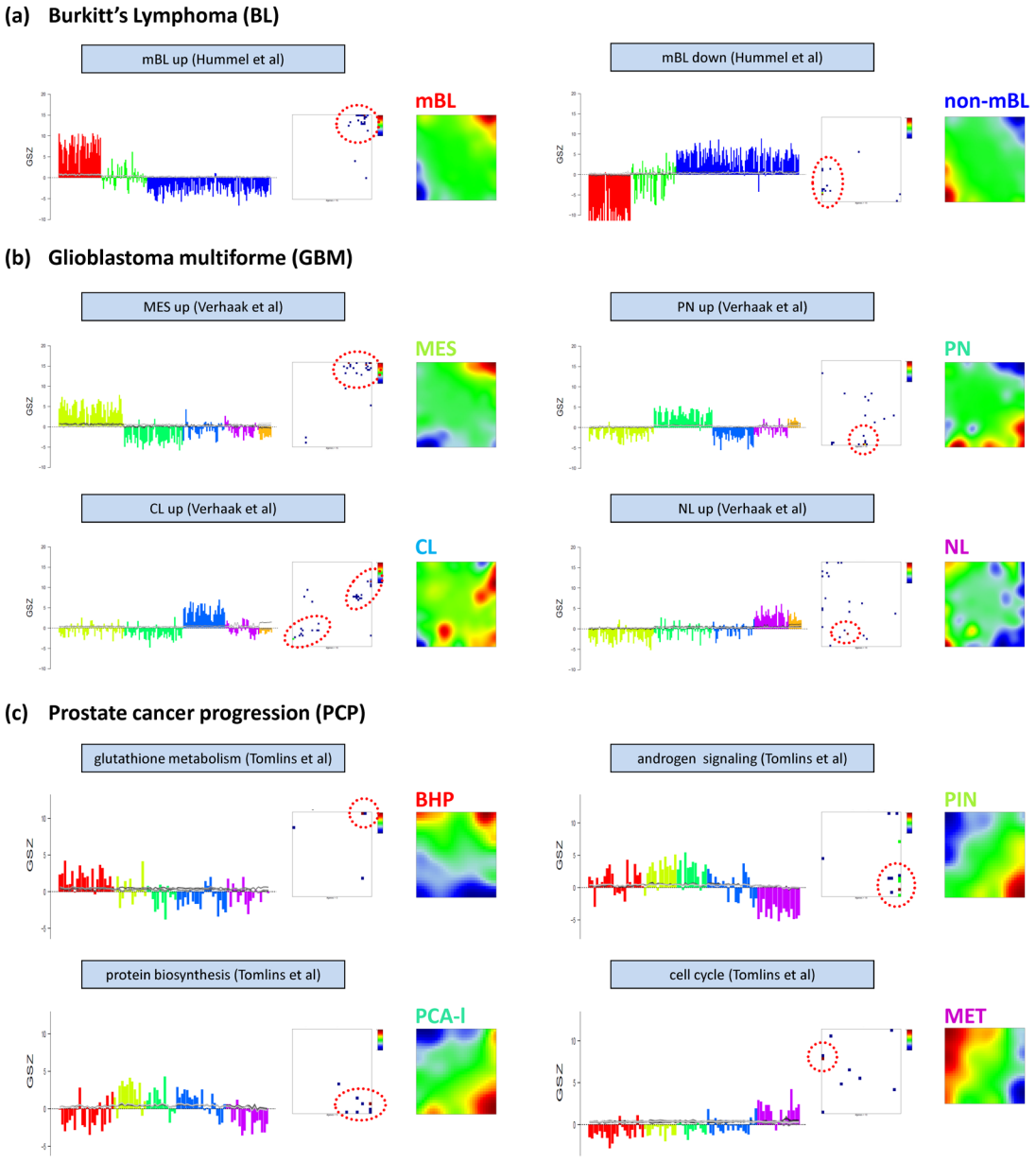


**Figure 6-11: Subtype specific genes of BL, GBM and PCP (panels a, b and c, respectively) taken from literature. Each gene set is depicted as GSZ-expression profile and population map. Additionally, mean SOM portraits of the corresponding subtypes are shown.**

## 6.2 SNP arrays: Atlas of human genome diversity

Human genetic diversity is shaped by both demographic and biological factors and has fundamental implications for understanding the genetic basis of diseases. Array-based genome-wide scans have been applied to worldwide populations, resulting in new insights into the genetic structure and relationships of human populations. Genotype data is available for nearly thousand individuals from the Human Genome Diversity Project, measuring approximately 660,000 SNPs (single polynucleotide polymorphisms) with Illumina 650Y arrays [121]. In particular, 1,043 individuals were analyzed, covering 57 ethnic groups assigned to 7 geographical regions.

Preprocessed data was downloaded from Human Genome Diversity Project. It contains genotype calls of both DNA strands for each loci and individual. For SOM analysis, these calls had to be transformed into numerical values. Therefore, each allele is classified as major (most frequent) homozygous allele, heterozygous allele or minor homozygous allele for each loci considered. Ternary values are used to encode these classes: '0' represents major allele, '1' heterozygous and '2' minor allele.

We selected the 50,000 most variant alleles among all individuals in the data. Note that normalization and standardization, as applied for gene expression data, is not necessary in this application. This data was used to train a SOM with resolution of 80x80 nodes, aggregating the 50,000 single alleles to 6,400 meta-alleles. The corresponding SNP meta-states of the samples are visualized in terms of SOM portraits. Figure 6-12 shows a gallery of 48 individuals selected out of 16 ethnicities. According to the ternary allelic code, blue and red colors in the SOM portraits refer to major and minor alleles, respectively. Green color represents heterozygous alleles. The portraits reveal a high diversity of patterns reflecting areas of major-, heterozygous- and minor-allelic genotypes. The SOM portraits are typically very similar for individuals from the same geographic region. For individuals originating from different regions, the portraits however progressively diverge with increasing geographic distance in most cases. In general, minor- and major-allelic regions in the portraits feature clockwise rotation in accordance with increasing migration distance from presumed human origin in Africa. For example, portraits of African individuals exhibit major homozygous alleles along top edge (see blue region in respective portraits), shifting to right edge in Middle-East and Europe, and further to bottom edge in Asia and particularly to bottom left corner in portraits of individuals from east Asia. This smooth conversion of the portraits indicates steady modifications in the genome due to early human migration. It also promotes the alleged route from Africa to Middle East (and Europe), further to Central Asia and via East Asia to America. The portraits of Oceanic individuals show a more speckle-like structures with spots arranged along all four edges. These individuals thus share allele characteristics with all the other regions, especially those of African and East Asian peoples. Possibly, this supports the theory of parallel human expansion across continents and via seafaring to Oceania,

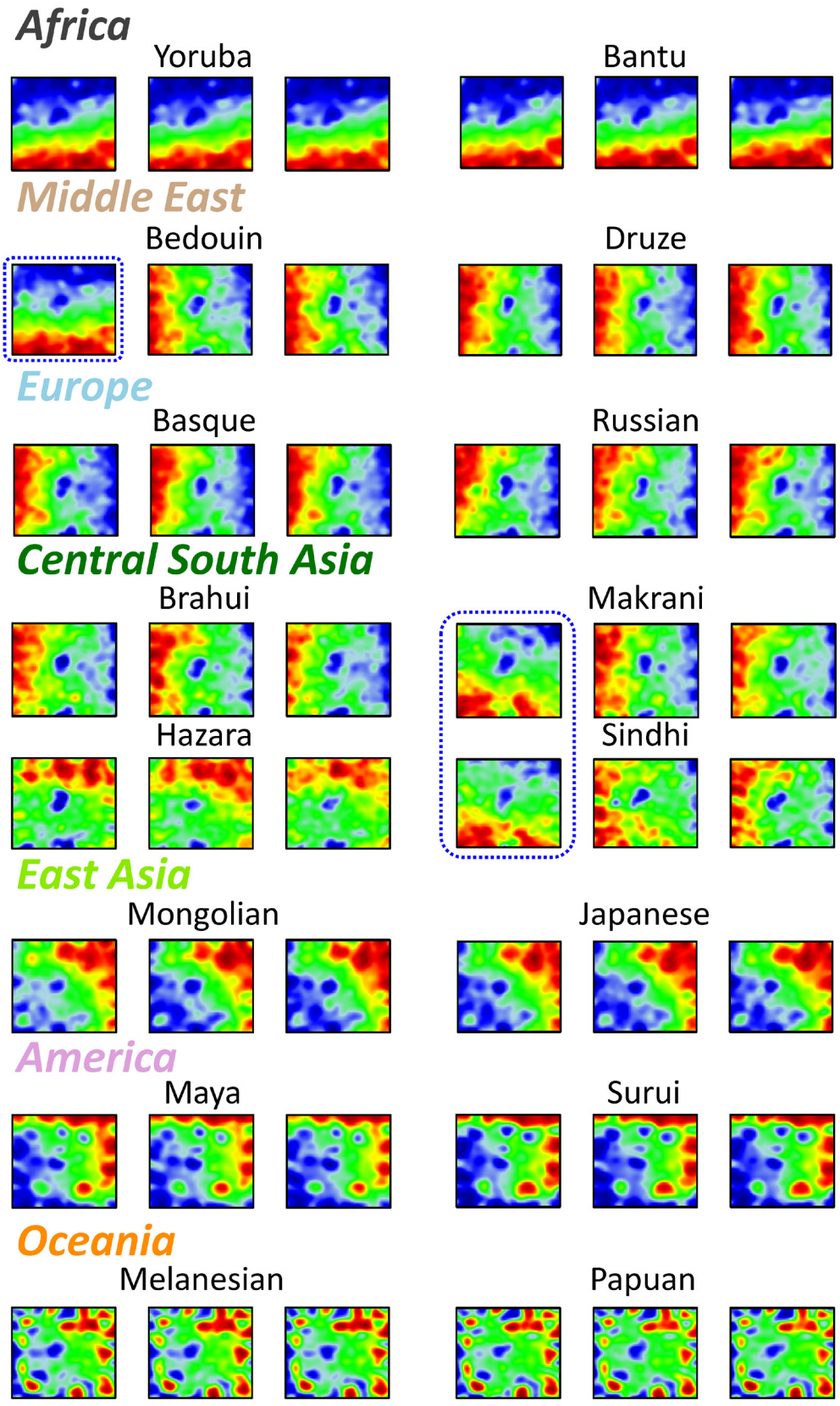


**Figure 6-12: Worldwide SNP-genotype portraits of human peoples: SOM portraits of 48 individuals selected from different regions of the world. Red, green and blue regions refer to minor-homozygous, heterozygous and major-homozygous allelic genotypes, respectively.**

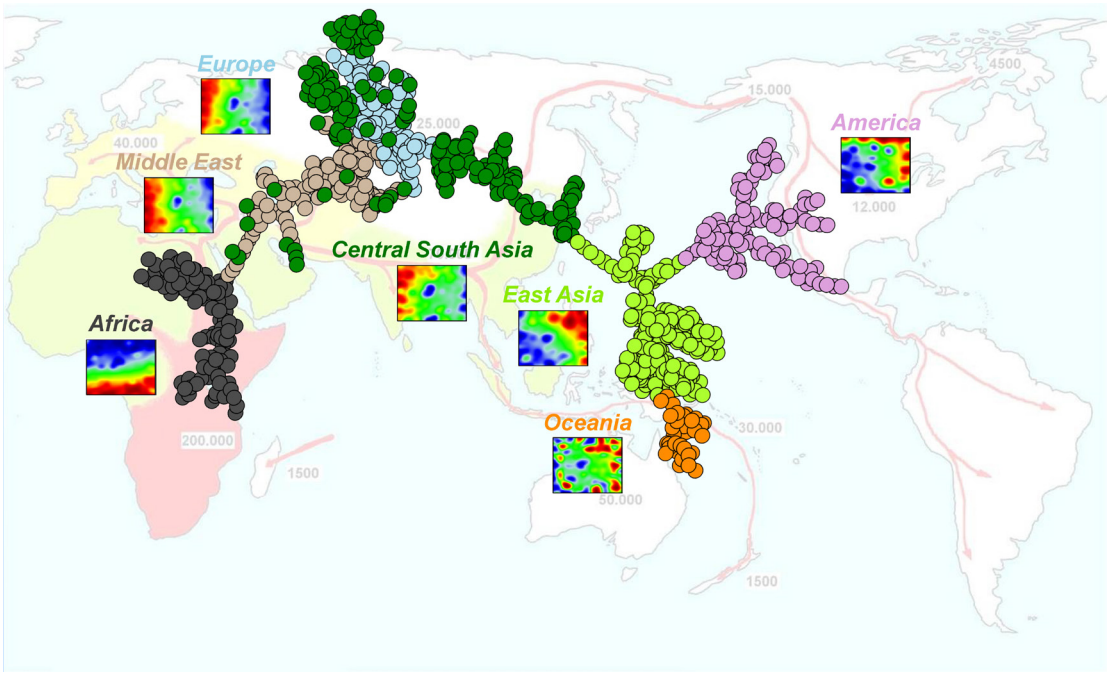


**Figure 6-13: The correlation spanning tree based on the meta-alleles illustrates similarity relations between the 1,043 individuals. The shown SOM portraits refer to average SNP characteristics of each region.[12]**

conserving genomic patterns of ancient African people, which are subsequently merged with those of neighboring East Asian ethnicities.

Notably, the SNP-SOM portraits of a few individuals are easy to identify as outliers in their ethnic group. For example, selected Makrani and Sindhi people are found to exhibit very similar portraits compared to the African group (see blue framing in Figure 6-12). Makrani are descendants of black Africans brought as slaves to Balochistan in medieval times. The portraits not only intuitively reflect this fact but also the circumstance that Makrani individuals feature close similarity to other groups from this region, such as Brahui or Sindhi, due to intermixing between the different ethnic groups. Also the SNP-portrait of one of the Bedouin individuals shows clearly the characteristics of black Africans, indicating ancestors from this region. The SNP-portraits of Hazara, another group from central Asia, reveal considerable similarity with the East Asian population presumably due to its partly Mongolian ancestry as descents of Mongolian military forces entering this region 500-700 years ago.

We generated a correlation spanning tree to analyze the similarity relations between the 1,043 individuals studied (Figure 6-13). Interestingly, the tree roughly resembles the geographic

[12] Background picture: http://en.wikipedia.org/wiki/File:Spreading_homo_sapiens.jpg

distribution of the populations, which, in turn, reflects the migration history among geographic regions. Hence, the tree reflects the fact that the mutual similarities between the SNP meta-states decreases with increasing 'decoupling' between the respective populations (see ref. [121–124] for a detailed discussion). Note that Principal Component Analysis (PCA) represents a widespread tool in population genetics for producing maps to summarize human genetic variation across continental regions since nearly 30 years [125]. However, the behavior of PCA for genetic data showing continuous spatial variation shows gradients and waves representing sinusoidal mathematical artifacts. Those arise generally when PCA is applied to spatial data, implying that the patterns do not necessarily reflect specific migration events [126].

Our examples illustrate the capability of SOM machine learning to map a large number of genotypes with individual resolution, and to judge relationships between populations and individuals in a simple and intuitive fashion. The question whether SOM mapping better reflects geographic migration than PCA-analysis requires further attempts presently under way.

## 6.3 Clustering of methylome Seq-data of prostate cancer

In this case study we demonstrate the capabilities of our SOM pipeline to analysze sequencing data of DNA-methylation's epigenetic modifications. The data was supplied by an immunoprecipitation-based approach combined with next generation sequencing (MeDIP-Seq). This technique allows to detect changes in the DNA-methylation state. It is often applied in research of cancer development, where cytosine DNA methylation is one of the initial processes on molecular level [127]. Cancer epigenomes are reported to be hypomethylated with specific hypermethylations [128].

The study was performed to survey the difference between healthy and prostate cancer tissue. It compares 53 control and 51 tumor samples [BÖRNO1]. Prostate cancer is one of the most common causes of male cancer deaths but however a curable disease when diagnosed at early stage. Reliable identification of tumor samples is therefore of great importance.

The data was preprocessed as follows: After preparation according to MeDIP assay and SOLiD sequencing, reads were mapped to the human genome HG19 using Applied Biosystems Bioscope software[13]. The reference genome was then split into bins of length 500bp, and the number of reads per bin was counted. Subsequently, obtained read number data was quantile-normalized, implying normalization of the total read count for each patient and ensuring

[13] http://www.appliedbiosystems.com

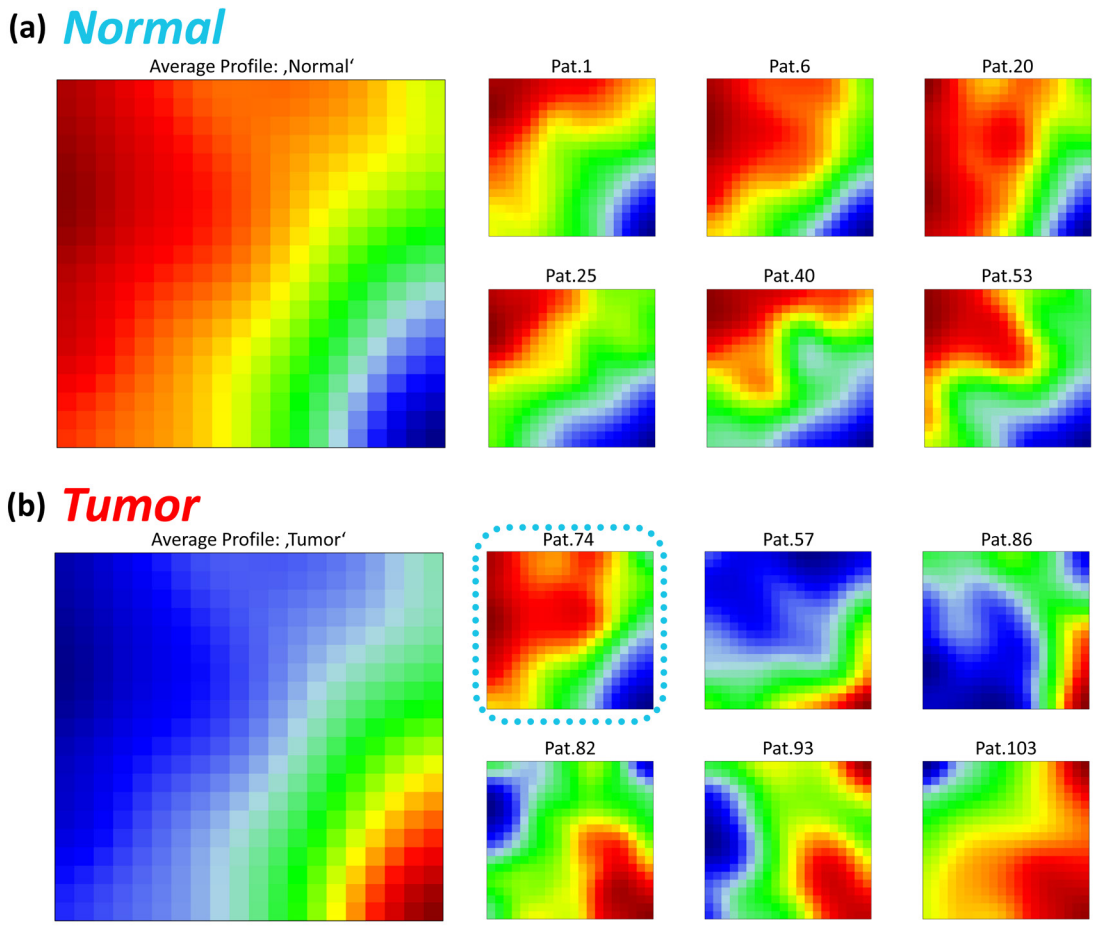


**Figure 6-14: Methylation SOM portraits of normal (panel a) and tumor (panel b) samples. The left panels show the mean portraits of the 53 normal or 51 tumor samples, respectively. The small portraits on the right show individual SOM portraits selected representatively. Portrait of patient '74' can be easily identified as outlier in the tumor sample class.**

comparability between the patients. Finally, the read counts were transformed into differential count values relative to the mean count of the particular loci. This is analogous to the transformation of gene expression values to differential expression and provides the data set with regard to differential methylation.

Out of 368,647 loci matched, most variant 20,000 were used as input for the SOM machine learning. It assigns the differential read count profiles of the input loci to K=20x20=400 meta-loci profiles. The corresponding SOM portraits directly represent the differential methylation in the samples as blue (hypomethylation) and red (hypermethylation) areas in the mosaic portraits. The left panels in Figure 6-14 show the mean methylation SOM portraits of the 53 normal and 51 tumor samples, respectively. These two classes feature virtually inverse portraits: Meta-loci located in top left corner of the SOM reveal hypermethylation in normal samples (see red colored regions in Figure 6-14) and hypomethylation in tumor samples (blue regions). Meta-loci in the bottom left corner show the opposite characteristics. On average, the

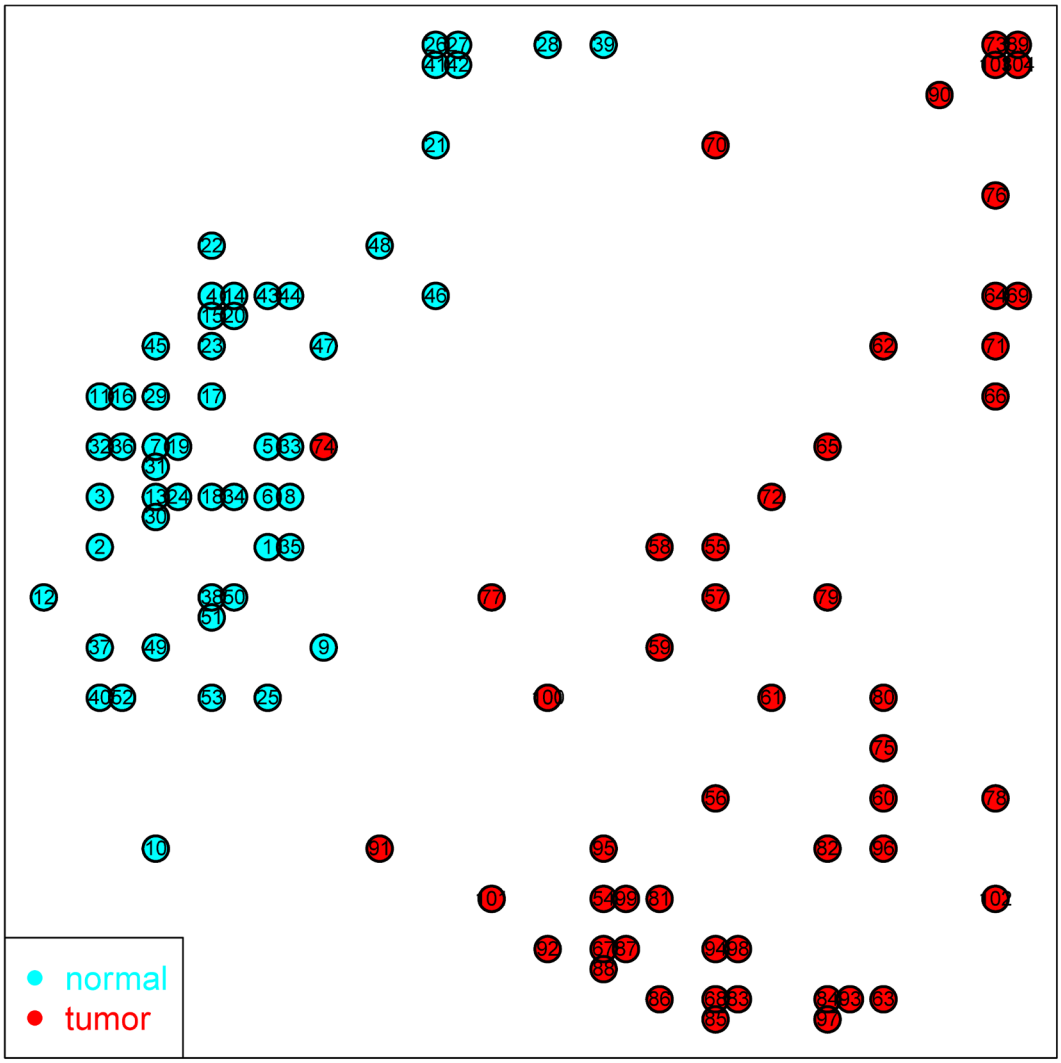


**Figure 6-15: Second level SOM of the meta-loci states separates normal and tumor samples (cyan and red points, respectively). Outlier tumor sample '74' clearly belongs to the cluster of normal samples.**

tumor SOM portraits exhibit a broad range of hypomethylated and only few hypermethylated meta-loci. This proportion agrees with previous findings, that prostate cancer epigenome is predominantly hypomethylated with few promoter-specific hypermethylations [128].
Right part of Figure 6-14 shows selected methylation SOM portraits of six normal and six tumor patient samples. Notice that sample '74', labeled as tumor tissue, reveals a methylation pattern which clearly resembles that of the normal samples.

To gain a more comprehensive overview about the relations between the samples, the meta-loci data was used to train a second level SOM with a resolution of 20x20 nodes. It is shown in Figure 6-15 and underlines the strong class structure, differentiating between normal samples in the left and tumor samples in the right part of the map. According to its 'normal-like' SOM

portrait, Sample '74' occurs as obvious outlier from the tumor tissue categorie. Reevaluation of sample '74' revealed insufficient tumor cell content and it was consequently removed from further analysis. After that, 7 specific loci could be identified as strongly differentially methylated. They provide a classifier that enables 100% correct classification of normal and tumor samples [BÖRNO1].
In this case study, capability of the SOM pipeline in context of next generation sequencing data was evaluated. SOM portraits here allow simple visual inspection of the quality of samples, including detection of misclassified or corrupt samples. Secondary analysis methods as second level SOM and classification algorithms are applied on meta-loci level and provide reliable differentiation between the tumor and the control samples.

## 6.4 MALDI-typing of infectious algae of the genus Prototheca

Beside microarrays and high-throughput sequencing, mass spectrometry is another emerging technique in molecular biology and led to an enormous increase in high content data in the fields of metabolomics and proteomics. A widely used approach is the combination of 'matrix-assisted laser desorption/ionization' and 'time-of-flight mass spectrometry' (MALDI-ToF MS). One unique feature of MALDI-ToF is the parallel assessment of all masses in a wide mass range. Thereby it inherently provides information of a wide range of proteins which can be used for protein identification by 'peptide mass fingerprinting' (PMF). The so-called MALDI-typing however employs the entire spectra to classify samples on proteome level without the need for detailed knowledge of the composition of single proteins. This method was developed for the rapid identification of bacterial samples [129] and subsequently extended to diverse phyla, ranging from microorganisms as bacteria [130] towards small invertebrates [131] and vertebrates [132]. We applied MALDI-typing to extracts of green algae from the genus *Prototheca* which are often overseen or mistaken for yeast in clinical diagnosis [vBERGEN1]. These algae from the Chlorella family are the only known plants that cause infections in humans and animals. To promote identification of those pathogens, the SOM-method was applied for fast and reliable distinction of *Prototheca* species [WIRTH2].
The study comprises 324 *Prototheca* samples referring to five species with one of them differentiated into two genotypes. They were extracted and prepared using a standard protocol [133]. The mass spectra were then recorded in MALDI-ToF-MS with a mass range from about 2,000 to 20,000Da. Peaks were detected from the raw mass spectra after baseline subtraction using the centroid algorithm implemented in the standard Bruker Daltonics software[14]. Subsequently, the MS-Screener 1.0.1 software extracts discrete supporting points along the m/z-axis which meet the condition of non-zero intensity amplitude in at minimum one sample spectrum of the series [134]. Those supporting points characterize the continuous spectra in

[14] FlexAnalysis 2.4 (Bruker Daltonics, Bremen, Germany)

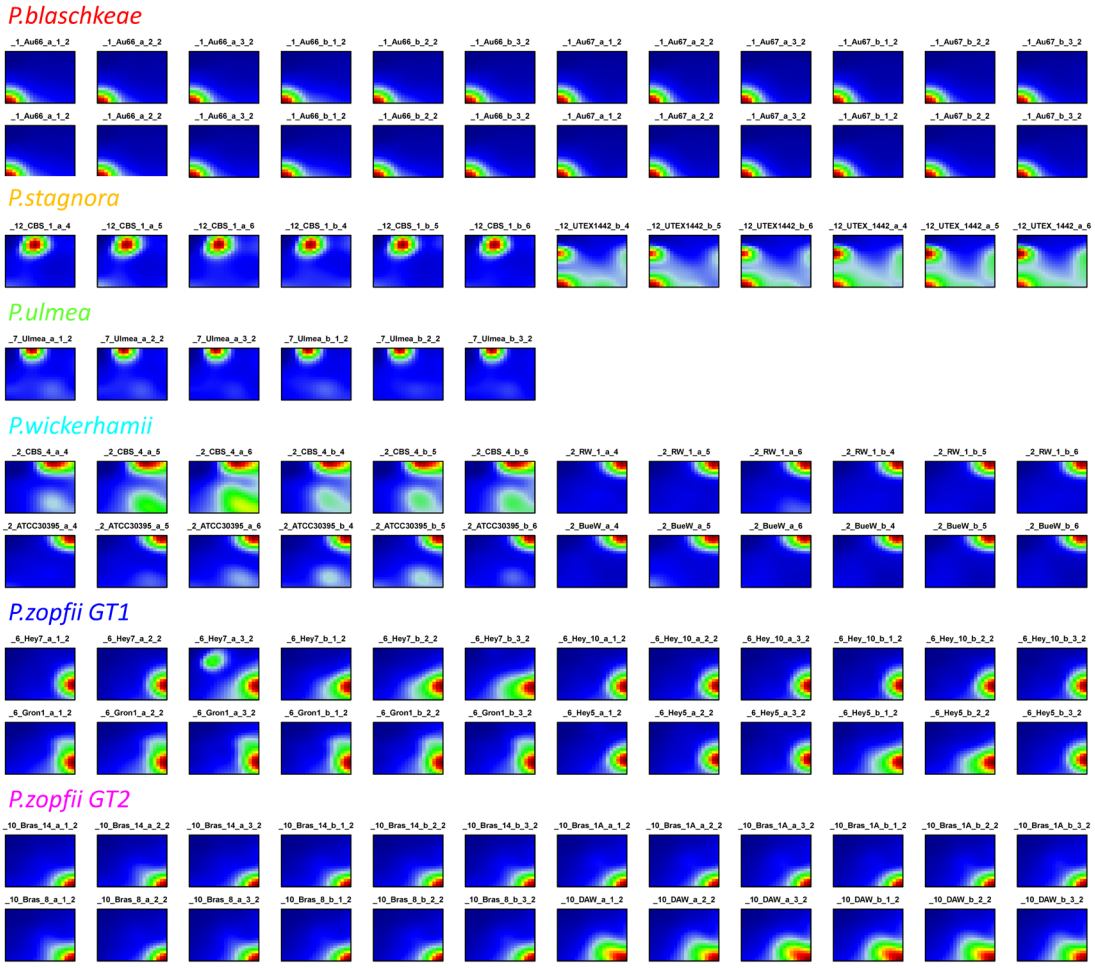


**Figure 6-16: SOM portraits of 114 selected *Prototheca* samples. The portraits are arranged according to their taxonomic categories.**

terms of designated positions along the m/z-axis of the spectra and will be further on referred as *peaklist*. The peaklist derived from *Prototheca* spectra contains 1,406 intensity amplitudes and covers the range from 4,135 to 16,954Da [vBERGEN1]. The peaklists were quantile-normalized to ensure comparability between the samples. Notably standardization to the mean value does not apply to the peak intensities. We used a SOM to map the 1,406 peak intensity profiles to K=20x20=400 meta-peak clusters (see [WIRTH2] for details).

The respective SOM portraits are shown in Figure 6-16, reflecting the underlying MS-pattern. Each of those exhibits characteristic spatial and color patterns, serving as MS-fingerprint of the *Prototheca* samples studied: the portraits typically feature one characteristic red spot, referring to peaks of high amplitude. The position of these spots varies in a species-specific fashion. Each species is characterized by a set of peaks showing high amplitudes only for this particular species, and small amplitudes for all other ones. Comparison of the portrait-textures therefore enables the straightforward classification of the samples according to their taxonomic membership.

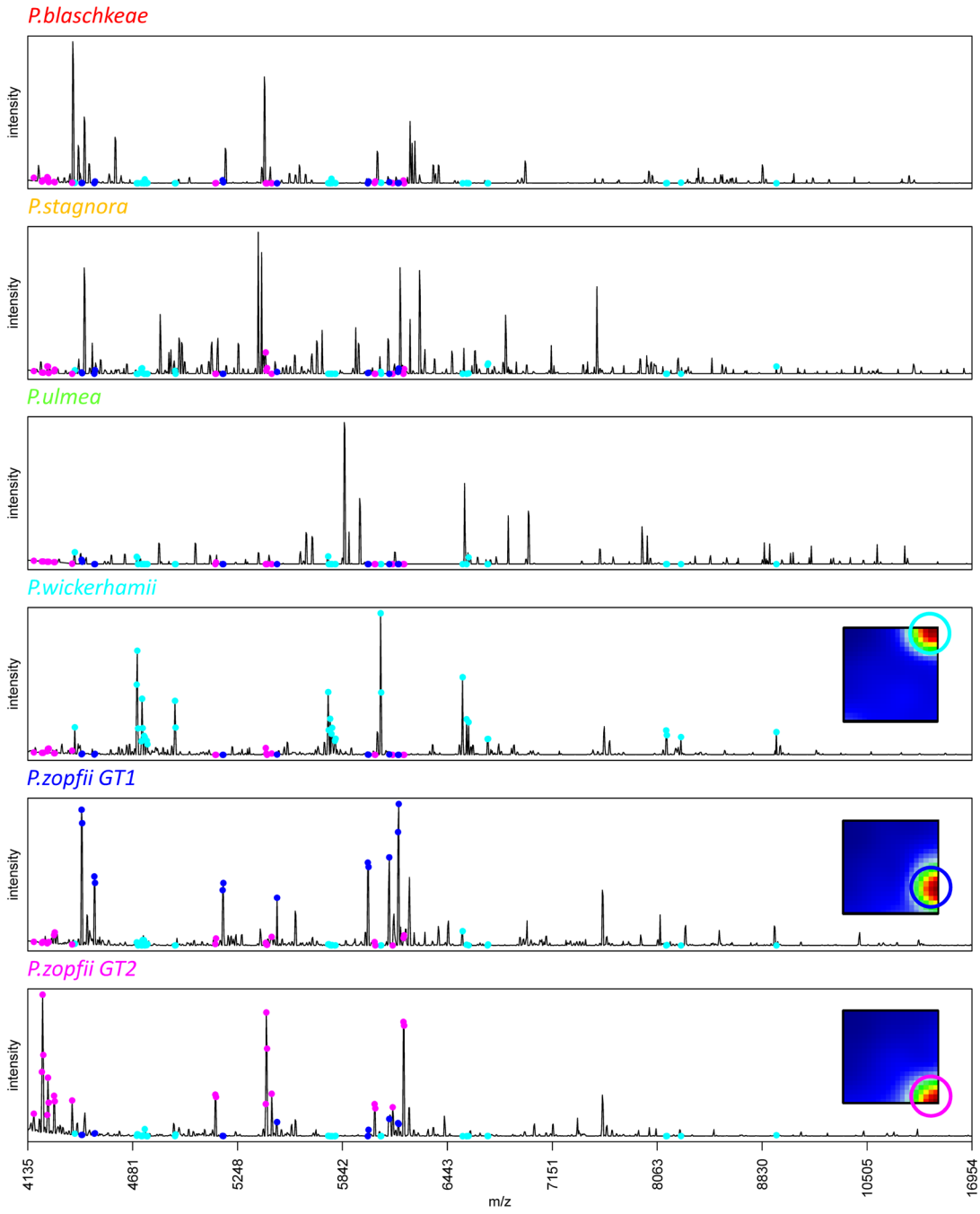


**Figure 6-17: MALDI-ToF spectra of different *Prototheca* species. Peaks indicated by cyan, blue or magenta dots refer to the respective red 'high-amplitude' spots in SOM portraits of *P. wickerhamii*, *P. zopfii GT1* or *P. zopfii GT2*, respectively.**

Each tile of the mosaic portraits refers to one of 400 meta-peak profiles, serving as representatives for clusters of similar single peak profiles. Figure 6-17 links representative mass spectra of all species studied with meta-peaks marked in the SOM portraits of *P. wickerhamii* and the two *P. zopfii* genotypes. Peaks of the spectra colored in cyan, blue and magenta are associated to the meta-peaks of high amplitude in the *P. wickerhamii*, *P.zopfii*

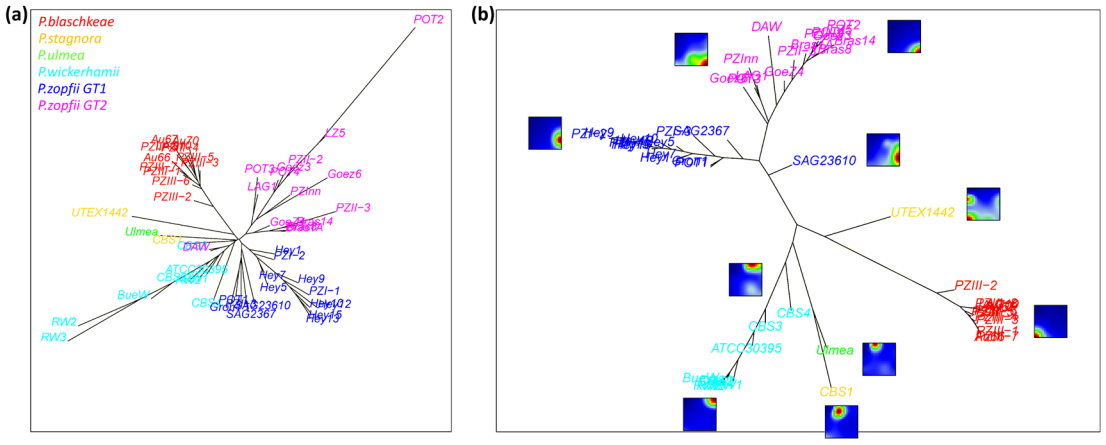


**Figure 6-18: Phylogenetic trees based on original spectral data (panel a) feature less discrimination power than the meta-spectra based tree (panel b). SOM portraits are shown for selected branches in panel b.**

*GT1* and *GT2* portraits, respectively. This representation clearly shows that the selected peaks form a characteristic set with high amplitudes in the spectra of one of the species sole.

To verify the improvement of secondary analysis methods in the context of MS data, phylogenetic cluster trees were generated using the neighbor-joining algorithm [79]. Figure 6-18 shows those trees based on either single peak or meta-peak data. Here, the leaves represent the samples and the lengths of the branches are directly related to the distances between them. Both trees cluster the different *Prototheca* species into different branches reflecting strong classification power of both data sets. The single peak based tree in Figure 6-18a is however more compact than the meta-peak based one in Figure 6-18b. Detailed evaluation reveals that the increased compactness of the former tree results from the small distances between the branches of different species. This consequently reveals enhanced separation of the different species in the meta-peak based tree. Additionally, subtle substructures not clearly evident in the single peak tree are resolved: For example, the *P. zopfii GT1* (blue color) sample 'SAG23610' is characterized by slight, but systematic differences in the SOM portraits compared to those of the other portraits of *P. zopfii GT1*. On the other hand, sample 'POT2' protrudes as outlier among the *P. zopfii GT2* samples (magenta color) in the single peak tree primarily due to an extraordinarily strong intensity of the MS-peaks at 4234.2 and 4237.2Da. They are however are averaged out in the meta-peak profiles. As consequence, the 'POT2' sample is clearly better integrated in the cluster of *P. zopfii GT2* samples in the meta-peak tree.

In summary, SOM portraits reflect characteristic pattern for each of the *Prototheca* species, but also allows identification and examination of outliers. Furthermore, improvement of downstream analysis was verified using phylogenetic trees as example.

## 6.5 Comparison of SOM analyses customized for different 'OMEs'

SOM was applied to microarray-, sequencing- and MS-data sets. An overview was given about capabilities of SOM based analysis regarding multiple tasks as comparison of different treatments, data monitoring and classification. Table 8 outlines key aspects of the case studies presented here. Application of the SOM pipeline splits into data-specific and common tasks. Preprocessing of the raw data is naturally specific and typically varies from study to study (see Table 8). Mostly quantile normalization is applied to improve the comparability of the samples by assuming identical distributions of the data for each sample. Note that quantile normalization is not applicable in case of SNP-array data, where values represent ternary allelic states.

Using appropriately preprocessed data allows utilization of SOM machine learning. This common task is virtually independent of original data source. SOM then extracts major effects (e.g. expression modes) inherent in the data set, where the resolution depends on dimension of the node grid. Comparison of data dimension and utilized SOM size (see Table 8) reveals heterogeneous requirements: on the one hand, large cancer transcriptome and SNP data sets (M=221; N=22,283 and M=1,034; N=50,000, respectively) require large SOMs for the purpose of a widespread overview of the samples with individual resolution (50x50 and 80x80 nodes, respectively). On the other hand, smaller data sets (e.g. YMC transcriptome with M=11 and N=5,900) are sufficiently captured even in smaller SOMs. Also separation of disjunct classes (e.g. in the prostate cancer methylome or *Prototheca* proteome studies) is adequately supplied by small 20x20 node SOMs. Please note that SOM size in combination with dimension of input data set determines the processing time of the SOM training. It ranges from few minutes for small data sets and SOMs to several days for high-dimensional data and high-resolution SOMs (see Table 8 for details). However the SOM learning algorithm can be parallelized (e.g. [135]), taking advance of high-performing multi-core computers which significantly reduces the processing time. Also memory requirements can be reduced by application of batch algorithms. They divide the data to disjoint batches and perform SOM analysis on those dimension-reduced subsets [60, 136].

According to individual character of the considered data type, the SOM portraits require OME-specific interpretations. The common color gradient was chosen for sake of optimal visual perception of different value levels. Red and blue colored tiles refer to over- and underexpressed meta-genes in transcriptome applications, whereas intermediate colors indicate invariant or information-less meta-genes. In case of sequencing data, red and blue refers to particularly high and low read numbers (e.g. hyper- and hypomethylation) of the meta-loci. SNP-array data uses a digitized coding of the alleles, resulting in principally ternary SOM portraits encoding minor, heterozygous and major alleles in red, green and blue, respectively. Finally, SOM portraits derived from MS-spectra exhibit an asymmetric character, as red tiles imply meta-peaks of high intensity, which are of exclusive interest. Blue (low

intensity) meta-peaks carry essentially no information. Proper interpretation of the SOM portraits therefore requires consideration of the specifics of the data.

Second level analyses apply to all data types in the same fashion and take advantage of better representativeness and reduced noisiness of the meta-features. Selection of appropriate methods is thereby again data-specific. For example, studies in the field of evolutionary biology prefer hierarchical structures as represented by cluster dendrograms, phylogenetic trees or correlation spanning trees. Those methods are able to capture incremental transitions and to depict progressive developments. On the other hand, studies with complex sample structures require second level analyses that do not force the sample relations into a hierarchical or mutually correlated structure. For such studies second level SOMs and correlation networks provide efficient tools to capture and visualize the multivariate sample similarity structure.

Additionally to the examples presented in this thesis, the SOM pipeline was applied to several further data sets, for example stem cell development, comparison of human and chimp organs, MS-based proteome of *Drosophila* and miRNA surveys of murine and human tissues. The SOM pipeline provides excellent results for all these applications.

# 7 Summary

We developed and presented a SOM-based analysis workflow for high-dimensional molecular-biological data which splits into a series of modular tasks. The first task is data preprocessing and normalization to transform the raw data into appropriate input data for SOM training. These high-dimensional data are afterwards processed in the SOM machine learning algorithm. It condenses the full data information into meta-feature clusters of similar and hence potentially co-regulated single features. Importantly, this dimension reduction does not entail a loss of primary information in contrast to simple filtering approaches which irretrievably remove parts of the data. Instead, the reduction of dimension is attained by the re-weighting of primary information in the aggregation step. The whole set of single feature profiles remains virtually 'hidden' in the meta-features. The meta-data provided by the SOM algorithm is then visualized in terms of sample specific mosaic portraits. They provide an intuitive way of visualization with strong capabilities in immediate identification of (meta-)features of interest.

The case studies demonstrated that SOM portraits transform large and heterogeneous sets of molecular biological data into an atlas of sample-specific texture maps which can be directly compared in terms of similarities and dissimilarities. The use of SOM portraits as primary visualization method is therefore straightforward. A number of supporting maps, supporting profiles and summary maps characterize selected properties of the meta-data.

Spot-clusters of correlated meta-features are extracted from the SOM portraits in a subsequent step of aggregation. This spot-clustering effectively enables reduction of the dimensionality of the data to a handful of signature modules in an unsupervised fashion. The SOM method consequently compresses the original set of high-dimensional data in two consecutive steps: Firstly, similar profiles of single features are collected in the meta-feature clusters, which reduces the number of relevant features by about one order of magnitude in our applications. Secondly, the spot textures of the obtained SOM portraits are decomposed into a few (typically less than one dozen) spots of similar meta-features. This 'double compression' sequentially applies global (similar profiles) and local (e.g. over-/underexpression in part of the samples) criteria.

An optional filtering step is applied to remove noisy or non-informative meta-features after SOM training. Recall that these features were involved in the training process, which is necessary to obtain a holistic characterization of the data set represented by the meta-features. Utilization of variance and significance based filters reveal similar filtering characteristics, whereas single feature lists are expected to be one order of magnitude longer than the comparable meta-feature lists. Different levels of feature and meta-feature filtering were applied and assessed in terms of maintaining representativeness and reducing noisiness of the data in downstream hierarchical clustering, independent component analysis and pairwise correlation analysis. The improved discrimination power of meta-features in such analyses

can be ascribed to essentially two facts: Firstly, the set of meta-features better represents the diversity of patterns and modes inherent in the data and secondly, it also possesses the better signal-to-noise characteristics as a comparable collection of single features. Due to the better representativeness, meta-feature lists are less sensitive to downstream filtering than lists of single feature. Meta-features can thus be seen as a natural choice to detect context-dependent patterns in complex data sets.

Additionally to the pattern-driven feature selection in the SOM portraits, statistical measures are applied to detect significantly differential features between sample classes. Implementation of scoring measurements, such as the shrinkage t-score, supplements the SOM analysis. Further, two variants of functional enrichment analyses were introduced, linking meta-features and spot-clusters with biological knowledge and support functional interpretation of the data based on the 'guilt by association' principle. They provide efficient tools for functional interpretation of the meta-features and of sample-specific patterns.

Selected case studies were presented in this thesis. In particular, molecular phenotype data derived from expression microarrays (mRNA, miRNA), sequencing (DNA methylation, histone modification patterns) or mass spectrometry (proteome), and also genotype data (SNP-microarrays) was analyzed. It was shown that the SOM analysis pipeline implies strong application capabilities and covers a broad range of potential purposes ranging from time series and treatment-vs.-control experiments to discrimination of samples according to genotypic, phenotypic or taxonomic classifications.

All analyses described in this work were carried out by our homemade software package. It was implemented in the common R-language [49] and published as open-source CRAN package 'oposSOM'[15]. To account for the challenges given by the diverse studies, the software provides a variety of visualizations, report sheets, downstream analyses and, for detailed and accurate descriptions, the complete statistical assessment summarized in spreadsheets.

[15] http://cran.r-project.org/web/packages/oposSOM

# 8 Conclusion

The methods presented in this thesis aimed at bridging the gap between the potency of SOM-based machine learning on the one hand and its relatively infrequent application in molecular biology on the other hand. Methodical aspects of the SOM framework were presented, aiming at disentangling large-scale data sets by clustering of related features. It was shown that the SOM algorithm is especially suited for application in large and high-dimensional data sets due to the combination of clustering, dimension reduction, multidimensional scaling and strong visualization capabilities. Alternative methods usually facilitate one of these components sole. It was shown that the SOM approach outperforms pure clustering approaches in terms of extraction of characteristic expression modules. Additionally, individual sample visualization as mosaic portraits is highly sophisticated and surpasses competing approaches such as heatmaps. The SOM portraits serve as unmistakable fingerprints of the molecular phenotypes. Together with the supporting maps and profiles, they help to understand the structure of the transformed data and hence to convey SOM application to a broader field of researchers. Additional software modules provide measures for differential expression and functional enrichment. They complement the SOM machine learning with statistical components which the basal algorithm lacks of. It was shown that the comprehensive analysis package is capable to meet all the challenges of the different applications presented.

# List of Figures

# List of Tables

# References


1. Kohonen T: **Self-organizing formation of topologically correct feature maps**. *Biological Cypernetics* 1982, **43**:59-69.
2. Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O: **Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study.** *BMC bioinformatics* 2002, **3**:36.
3. Guo Y, Eichler GS, Feng Y, Ingber DE, Huang S: **Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers.** *Journal of biomedicine & biotechnology* 2006, **2006**:69141.
4. Tamayo P, Slonim D, Mesirov J, et al.: **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* 1999, **96**:2907-12.
5. Törönen P, Kolehmainen M, Wong G, Castrén E: **Analysis of gene expression data using self-organizing maps.** *FEBS letters* 1999, **451**:142-6.
6. Golub TR, Slonim DK, Tamayo P, et al.: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science (New York, N.Y.)* 1999, **286**:531-7.
7. Covell DG, Wallqvist A, Rabow AA, Thanki N: **Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data.** *Molecular cancer therapeutics* 2003, **2**:317-32.
8. Bivort B de, Huang S, Bar-Yam Y: **Dynamics of cellular level function and regulation derived from murine expression array data.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:17687-92.
9. Bivort B de, Huang S, Bar-Yam Y: **Empirical multiscale networks of cellular regulation.** *PLoS computational biology* 2007, **3**:1968-78.
10. Nikkilä J, Törönen P, Kaski S, et al.: **Analysis and visualization of gene expression data using self-organizing maps.** *Neural networks : the official journal of the International Neural Network Society* 2002, **15**:953-66.
11. Zollanvari A, Cunningham MJ, Braga-Neto U, Dougherty ER: **Analysis and modeling of time-course gene-expression profiles from nanomaterial-exposed primary human epidermal keratinocytes.** *BMC bioinformatics* 2009, **10 Suppl 1**:S10.
12. Hur K, Lee H-J, Woo JH, Kim JH, Yang H-K: **Gene expression profiling of human gastrointestinal stromal tumors according to its malignant potential.** *Digestive diseases and sciences* 2010, **55**:2561-7.
13. Spencer WC, Zeller G, Watson JD, et al.: **A Spatial and Temporal Map of C. elegans Gene Expression.** *Genome research* 2010, **21**:325-41.
14. Scherbart A, Timm W, Böcker S, Nattkemper TW: **Som-based peptide prototyping for mass spectrometry peak intensity prediction**. *WSOM'07* 2007.
15. Meinicke P, Lingner T, Kaever A, et al.: **Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps.** *Algorithms for molecular biology : AMB* 2008, **3**:9.
16. Han EC, Lee Y-S, Liao W-S, et al.: **Direct tissue analysis by MALDI-TOF mass spectrometry in human hepatocellular carcinoma.** *Clinica chimica acta; international journal of clinical chemistry* 2011, **412**:230-9.
17. Suna T, Salminen A, Soininen P, et al.: **1H NMR metabonomics of plasma lipoprotein subclasses: elucidation of metabolic clustering by self-organising maps.** *NMR in biomedicine* 2007, **20**:658-72.
18. Lloyd G, Wongravee K: **Self Organising Maps for variable selection: Application to human saliva analysed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral healthcare product**. *Chemometrics and Intelligent Laboratory Systems* 2009.
19. Wongravee K, Lloyd GR, Silwood CJ, Grootveld M, Brereton RG: **Supervised self organizing maps for classification and determination of potentially discriminatory variables: illustrated by application to nuclear magnetic resonance metabolomic profiling.** *Analytical chemistry* 2010, **82**:628-38.

20. Zhang Y, Wolf-Yadlin A, Ross PL, et al.: **Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules.** *Molecular & cellular proteomics : MCP* 2005, **4**:1240-50.
21. Schulz R, Woodfine K, Menheniott TR, et al.: **WAMIDEX: a web atlas of murine genomic imprinting and differential expression.** *Epigenetics : official journal of the DNA Methylation Society* 2008, **3**:89-96.
22. Murty US, Srinivasa Rao M, Misra S: **Prioritization of malaria endemic zones using self-organizing maps in the Manipur state of India.** *Informatics for health & social care* 2008, **33**:170-8.
23. Lagus K, Kaski S, Kohonen T: **Mining massive document collections by the WEBSOM method**. *Information Sciences* 2004, **163**:135–156.
24. Villmann T, Merényi E, Hammer B: **Neural maps in remote sensing image analysis.** *Neural networks : the official journal of the International Neural Network Society* 2003, **16**:389-403.
25. Hammer B, Micheli A, Sperduti A, Strickert M: **Recursive self-organizing network models.** *Neural networks : the official journal of the International Neural Network Society* 2004, **17**:1061-85.
26. Voegtlin T: **Recursive self-organizing maps.** *Neural networks : the official journal of the International Neural Network Society* 2002, **15**:979-91.
27. Hagenbuchner M, Sperduti A, Tsoi AC: **A self-organizing map for adaptive processing of structured data.** *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 2003, **14**:491-505.
28. Strickert M, Hammer B, Blohm S: **Unsupervised recursive sequence processing**. *Neurocomputing* 2005, **63**:69-97.
29. Strickert M, Hammer B: **Neural gas for sequences**. *Proceedings of the Workshop on Self-Organizing Maps* 2003, **1**.
30. Bauer HU, Villmann T: **Growing a hypercubical output space in a self-organizing feature map.** *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 1997, **8**:218-26.
31. Hsu AL, Halgamuge SK: **Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation**. *International Journal of Approximate Reasoning* 2003, **32**:259–279.
32. Martinetz TM, Berkovich SG, Schulten KJ: **Neural-gas network for vector quantization and its application to time-series prediction.** *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 1993, **4**:558-69.
33. Villmann T, Schleif F-M, Kostrzewa M, Walch A, Hammer B: **Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods.** *Briefings in bioinformatics* 2008, **9**:129-43.
34. Timm W, Scherbart A, Böcker S, Kohlbacher O, Nattkemper TW: **Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics.** *BMC bioinformatics* 2008, **9**:443.
35. Wold S, Esbensen K, Geladi P: **Principal component analysis**. *Chemometrics and Intelligent Laboratory Systems* 1987, **2**:37-52.
36. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:5116-21.
37. Vesanto J: **SOM-based data visualization methods**. *Intelligent data analysis* 1999, **3**:111–126.
38. Eichler GS, Huang S, Ingber DE: **Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles.** *Bioinformatics (Oxford, England)* 2003, **19**:2321-2.
39. Björkbacka H, Fitzgerald KA, Huet F, et al.: **The induction of macrophage gene expression by LPS predominantly utilizes Myd88-independent signaling cascades.** *Physiological genomics* 2004, **19**:319-30.
40. Huang S, Eichler G, Bar-Yam Y, Ingber DE: **Cell fates as high-dimensional attractor states of a complex gene regulatory network.** *Physical review letters* 2005, **94**:128701.
41. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S: **Transcriptome-wide noise controls lineage choice in mammalian progenitor cells.** *Nature* 2008, **453**:544-7.
42. Mar JC, Quackenbush J: **Decomposition of gene expression state space trajectories.** *PLoS computational biology* 2009, **5**:e1000626.

43. Huang S, Ernberg I, Kauffman S: **Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective.** *Seminars in cell & developmental biology* 2009, **20**:869-76.
44. Tsigelny IF, Kouznetsova VL, Sweeney DE, et al.: **Analysis of metagene portraits reveals distinct transitions during kidney organogenesis.** *Science signaling* 2008, **1**:ra16.
45. Buckhaults P, Zhang Z, Chen Y-C, et al.: **Identifying tumor origin using a gene expression-based classification map.** *Cancer research* 2003, **63**:4144-9.
46. Camphausen K, Purow B, Sproull M, et al.: **Influence of in vivo growth on human glioma cell line gene expression: convergent profiles under orthotopic conditions.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:8287-92.
47. MacArthur BD, Lachmann A, Lemischka IR, Ma'ayan A: **GATE: software for the analysis and visualization of high-dimensional time series expression data.** *Bioinformatics (Oxford, England)* 2010, **26**:143-4.
48. Milone DH, Stegmayer GS, Kamenetzky L, et al.: ***omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants.** *BMC bioinformatics* 2010, **11**:438.
49. Development Core Team R: **R: A Language and Environment for Statistical Computing**. 2011.
50. McCulloch W, Pitts W: **A logical calculus of the ideas immanent in nervous activity**. *Bulletin of Mathematical Biology* 1943.
51. Rojas R: **Neural networks: a systematic introduction**. 1996.
52. Rosenblatt F: **The perceptron: a probabilistic model for information storage and organization in the brain.** *Psychological review* 1958, **65**:386-408.
53. Minsky M, Papert S: *Perceptrons: An introduction to computational geometry*. MIT Press; 1969.
54. Hebb D: *The Organization of Behavior. A Neuropsychological Theory*. Wiley, New York; 1949.
55. Haykin S: **Neural Networks and Learning Machines**. *Pearson Prentice Hall, New Jersey, USA, 936 p.[Links]* 2008.
56. Binder H, Krohn K, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures.** *Algorithms for molecular biology : AMB* 2008, **3**:11.
57. Binder H, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: theory and algorithm.** *Algorithms for molecular biology : AMB* 2008, **3**:12.
58. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics (Oxford, England)* 2003, **19**:185-93.
59. Miikkulainen R, Bednar J, Choe Y, Sirosh J: *Computational Maps in the Visual Cortex*. Springer; 2005.
60. Kohonen T: **Self Organizing Maps**. *Springer, Berlin, Heidelberg, New York* 1995.
61. Tan H, George S: **Investigating learning parameters in a standard 2-D SOM model to select good maps and avoid poor ones**. *AI 2004: Advances in Artificial Intelligence* 2005.
62. Vesanto J, Alhoniemi E: **Clustering of the self-organizing map.** *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 2000, **11**:586-600.
63. Erwin E, Obermayer K, Schulten K: **Self-organizing maps: stationary states, metastability and convergence rate.** *Biological cybernetics* 1992, **67**:35-45.
64. Hornshøj H, Conley LN, Hedegaard J, et al.: **Microarray expression profiles of 20.000 genes across 23 healthy porcine tissues.** *PloS one* 2007, **2**:e1203.
65. Kadota K, Nakai Y, Shimizu K: **A weighted average difference method for detecting differentially expressed genes from microarray data.** *Algorithms for molecular biology : AMB* 2008, **3**:8.
66. Kadota K, Nakai Y, Shimizu K: **Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity.** *Algorithms for molecular biology : AMB* 2009, **4**:7.
67. Lee DD, Seung HS: **Learning the parts of objects by non-negative matrix factorization.** *Nature* 1999, **401**:788-91.
68. Brunet J-P, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:4164-9.
69. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data.** *Genome research* 2003, **13**:1706-18.

70. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863-8.
71. Läuter J, Glimm E, Eszlinger M: **Search for relevant sets of variables in a high dimensional setup keeping the familywise error rate**. *Statistica Neerlandica* 2005.
72. Läuter J, Horn F, Rosołowski M, Glimm E: **High-dimensional data analysis: selection of variables, data compression and graphics--application to gene expression.** *Biometrical journal. Biometrische Zeitschrift* 2009, **51**:235-51.
73. MacQueen JB: **Some methods for classification and analysis of multivariate observations**. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. edited by Cam LML, Neyman J University of California Press; 1967, **1**:281-297.
74. Gaujoux R, Seoighe C: **A flexible R package for nonnegative matrix factorization**. *BMC bioinformatics* 2010.
75. Schug J, Schuller W-P, Kappen C, et al.: **Promoter features related to tissue specificity as measured by Shannon entropy.** *Genome biology* 2005, **6**:R33.
76. Quackenbush J: **Genomics. Microarrays--guilt by association.** *Science (New York, N.Y.)* 2003, **302**:240-1.
77. Eklund AC, Szallasi Z: **Correction of technical bias in clinical microarray data improves concordance with known biological information.** *Genome biology* 2008, **9**:R26.
78. Liebermeister W: **Linear modes of gene expression determined by independent component analysis.** *Bioinformatics (Oxford, England)* 2002, **18**:51-60.
79. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular biology and evolution* 1987, **4**:406-25.
80. Riester M, Stephan-Otto Attolini C, Downey RJ, Singer S, Michor F: **A Differentiation-Based Phylogeny of Cancer Subtypes**. *PLoS Computational Biology* 2010, **6**:e1000777.
81. Xu Y, Olman V, Xu D: **Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees.** *Bioinformatics (Oxford, England)* 2002, **18**:536-45.
82. Ahdesmäki M, Strimmer K: **Feature selection in omics prediction problems using cat scores and false non-discovery rate control**. *The Annals of Applied Statistics* 2010, **4**:503-519.
83. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:6567-72.
84. Binder H, Preibisch S, Berger H: **Calibration of microarray gene-expression data.** *Methods in molecular biology (Clifton, N.J.)* 2010, **576**:375-407.
85. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics (Oxford, England)* 2002, **18 Suppl 1**:S105-10.
86. Binder H, Kirsten T, Löffler M, Stadler PF: **Sensitivity of microarray oligonucleotide probes: variability and effect of base composition**. *The journal of physical chemistry. B* 2004, **108**:18003-18014.
87. Sartor MA, Tomlinson CR, Wesselkamper SC, et al.: **Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments.** *BMC bioinformatics* 2006, **7**:538.
88. Zeisel A, Amir A, Köstler WJ, Domany E: **Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes.** *BMC bioinformatics* 2010, **11**:400.
89. Opgen-Rhein R, Strimmer K: **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach.** *Statistical applications in genetics and molecular biology* 2007, **6**:Article9.
90. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.
91. Fodor AA, Tickle TL, Richardson C: **Towards the uniform distribution of null P values on Affymetrix microarrays.** *Genome biology* 2007, **8**:R69.
92. Jain N, Thatte J, Braciale T, et al.: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics (Oxford, England)* 2003, **19**:1945-51.
93. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nature reviews. Genetics* 2006, **7**:55-65.

94. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:9440-5.
95. Strimmer K: **fdrtool: a versatile R package for estimating local and tail area-based false discovery rates.** *Bioinformatics (Oxford, England)* 2008, **24**:1461-2.
96. Strimmer K: **A unified approach to false discovery rate estimation.** *BMC bioinformatics* 2008, **9**:303.
97. Aubert J, Bar-Hen A, Daudin JJ, Robin S: **Determination of the differentially expressed genes in microarray experiments using local FDR.** *BMC bioinformatics* 2004, **5**:125.
98. Sieberts SK, Schadt EE: **Moving toward a system genetics view of disease.** *Mammalian genome : official journal of the International Mammalian Genome Society* 2007, **18**:389-401.
99. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science (New York, N.Y.)* 2003, **302**:249-55.
100. Ackermann M, Strimmer K: **A general modular framework for gene set enrichment analysis.** *BMC bioinformatics* 2009, **10**:47.
101. Ashburner M, Ball CA, Blake JA, et al.: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**:25-9.
102. Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome biology* 2003, **4**:R70.
103. Zhang B, Schmoyer D, Kirov S, Snoddy J: **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.** *BMC bioinformatics* 2004, **5**:16.
104. Vêncio RZN, Shmulevich I: **ProbCD: enrichment analysis accounting for categorization uncertainty.** *BMC bioinformatics* 2007, **8**:383.
105. Törönen P, Ojala PJ, Marttinen P, Holm L: **Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function.** *BMC bioinformatics* 2009, **10**:307.
106. Newton M, Quintana F: **Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis**. *The Annals of ...* 2007.
107. Efron B, Tibshirani R: **On testing the significance of sets of genes**. 2006:1-31.
108. Levine DM, Haynor DR, Castle JC, et al.: **Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways.** *Genome biology* 2006, **7**:R93.
109. Alexa A, Rahnenführer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics (Oxford, England)* 2006, **22**:1600-7.
110. Li CM, Klevecz RR: **A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:16254-9.
111. Tu BP, Kudlicki A, Rowicka M, McKnight SL: **Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes.** *Science (New York, N.Y.)* 2005, **310**:1152-8.
112. **Polynuclear aromatic compounds. General remarks on the substances considered.** *IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans* 1983, **32**:33-91.
113. Dautel F, Kalkhof S, Trump S, et al.: **DIGE-based protein expression analysis of B[a]P-exposed hepatoma cells reveals a complex stress response including alterations in oxidative stress, cell cycle control, and cytoskeleton motility at toxic and subacute concentrations.** *Journal of proteome research* 2011, **10**:379-93.
114. Michaelson JJ, Trump S, Rudzok S, et al.: **Transcriptional signatures of regulatory and toxic responses to benzo-[a]-pyrene exposure.** *BMC genomics* 2011, **12**:502.
115. Miller KP, Ramos KS: **Impact of cellular metabolism on the biological effects of benzo[a]pyrene and related hydrocarbons.** *Drug metabolism reviews* 2001, **33**:1-35.
116. Brobyn RD: **The human toxicology of dimethyl sulfoxide.** *Annals of the New York Academy of Sciences* 1975, **243**:497-506.
117. Irizarry RA, Bolstad BM, Collin F, et al.: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic acids research* 2003, **31**:e15.
118. Hummel M, Bentink S, Berger H, et al.: **A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling.** *The New England journal of medicine* 2006, **354**:2419-30.
119. Verhaak RGW, Hoadley KA, Purdom E, et al.: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer cell* 2010, **17**:98-110.

120. Tomlins SA, Mehra R, Rhodes DR, et al.: **Integrative molecular concept modeling of prostate cancer progression.** *Nature genetics* 2007, **39**:41-51.
121. Li JZ, Absher DM, Tang H, et al.: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science (New York, N.Y.)* 2008, **319**:1100-4.
122. Novembre J, Johnson T, Bryc K, et al.: **Genes mirror geography within Europe.** *Nature* 2008, **456**:98-101.
123. Pickrell JK, Coop G, Novembre J, et al.: **Signals of recent positive selection in a worldwide sample of human populations.** *Genome research* 2009, **19**:826-37.
124. Novembre J, Rienzo A Di: **Spatial patterns of variation due to natural selection in humans.** *Nature reviews. Genetics* 2009, **10**:745-55.
125. Menozzi P, Piazza A, Cavalli-Sforza L: **Synthetic maps of human gene frequencies in Europeans.** *Science (New York, N.Y.)* 1978, **201**:786-92.
126. Novembre J, Stephens M: **Interpreting principal component analyses of spatial population genetic variation.** *Nature genetics* 2008, **40**:646-9.
127. Sharma S, Kelly TK, Jones PA: **Epigenetics in cancer.** *Carcinogenesis* 2010, **31**:27-36.
128. Irizarry RA, Ladd-Acosta C, Wen B, et al.: **The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.** *Nature genetics* 2009, **41**:178-86.
129. Bright JJ, Claydon MA, Soufian M, Gordon DB: **Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software.** *Journal of microbiological methods* 2002, **48**:127-38.
130. Jehmlich N, Schmidt F, Taubert M, et al.: **Comparison of methods for simultaneous identification of bacterial species and determination of metabolic activity by protein-based stable isotope probing (Protein-SIP) experiments.** *Rapid communications in mass spectrometry : RCM* 2009, **23**:1871-8.
131. Feltens R, Görner R, Kalkhof S, Gröger-Arndt H, Bergen M von: **Discrimination of different species from the genus Drosophila by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry.** *BMC evolutionary biology* 2010, **10**:95.
132. Mazzeo MF, Giulio BD, Guerriero G, et al.: **Fish authentication by MALDI-TOF mass spectrometry.** *Journal of agricultural and food chemistry* 2008, **56**:11071-6.
133. Maier T, Kostrzewa M: **Fast and reliable MALDI-TOF MS-based microorganism identification**. *Chemistry Today* 2007, **25**:68-71.
134. Schmidt F, Schmid M, Jungblut P, et al.: **Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis**. *Journal of the American Society for Mass Spectrometry* 2003, **14**:943-956.
135. Guan H, Cheung T: **Parallel design and implementation of SOM neural computing model in PVM environment of a distributed system**. *APDC* 1997.
136. Hammer B, Hasenfuss A, Schwenker F, El Gayar N: **Artificial Neural Networks in Pattern Recognition**. In *Artificial Neural Networks in Pattern Recognition*. edited by Schwenker F, Gayar N Berlin, Heidelberg: Springer Berlin Heidelberg; 2010, **5998**:259-273.
137. Birney E, Andrews TD, Bevan P, et al.: **An overview of Ensembl.** *Genome research* 2004, **14**:925-8.

# Curriculum vitae

## Personal Information

| | |
|---|---|
| Name | Dipl.-Inf. Henry Wirth |
| Date of birth | 05.10.1982 |
| Place of birth | Meerane, Germany |

## Education

| | | |
|---|---|---|
| 1999 - 2001 | Secondary School: Glauchau, Germany | A-level, grade “A” |
| 2002 - 2008 | University: Chemnitz, Germany | Diploma, grade “A” |
| since 2008 | University: Leipzig, Germany | PhD student |

## Studies

| | |
|---|---|
| 2002 - 2008 | Technical University of Chemnitz |
| | Diploma in “Computer Science”, main course “Artificial Intelligence” |
| | Diploma thesis: “Modeling of a cellular automaton based on dendritic cells and application in pattern recognition” |
| 2008 - 2011 | Interdisciplinary Centre for Bioinformatics - IZBI, University of Leipzig<br>Helmholtz Centre for Environmental Research – UFZ, Leipzig |
| | PhD student, scholarship holder |
| since 2011 | Interdisciplinary Centre for Bioinformatics - IZBI, University of Leipzig<br>Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment - LIFE, University of Leipzig |
| | PhD student, research fellow |

## Conferences / Summer schools

24th TBI Winterseminar 2009, Bled, Slovenia:
**Introduction to Statistical Analysis of Metabolomics** (talk)

Wissenschaftliches Jahreskolloquium des IMISE 2009, Leipzig, Germany:
**Integrative analysis of metabolomics and proteomics data in the context of obesity** (talk)

Saxon Biotechnology Symposium 2009, Leipzig, Germany:
**Integrative Analysis of Metabolomic and Proteomic Data** (poster presentation)

German Conference on Bioinformatics 2009, Halle, Germany:
**Combining the ‚OMICS': Integrative Data Analysis** (poster presentation)

25th TBI Winterseminar 2010, Bled, Slovenia:
**Gene Expression Analysis by staring at colorful pictures** (talk)

Wissenschaftliches Jahreskolloquium des IMISE 2010, Leipzig, Germany:
**Kartierung mittels hochdimensionaler Daten: Der Anatomieatlas des Menschen** (talk)

Lipari School on Bioinformatics and Computational Biology 2010, Lipari, Italy:
**Statistical and Machine Learning Methods in Computational Biology** (summer school)

INRIA-IZBI Workshop: From genes to tissues 2010, Leipzig, Germany:
**Whole genome expression profiling using self organizing maps** (talk)

Wissenschaftliches Jahreskolloquium des IMISE 2011, Leipzig, Germany:
**Profiling the ‚OMES': Molecular phenotypic portraits using self organizing maps** (talk)

Research Festival 2011, Leipzig, Germany:
**Self-organizing maps: Portraying the OMEs with individual resolution** (poster presentation)

Wissenschaftliches Jahreskolloquium des IMISE 2012, Leipzig, Germany:
**Improved MALDI-typing of *Prototheca* using self organizing maps** (talk)

# Publications

*: Contains results presented in this dissertation

1st: Contributed equally

## Proceedings

*[BINDER 1] Binder, H., Fasold, M., Hopp, L., Cakir, V., Bergen, M.v. & Wirth, H.: **Molecular phenotypic portraits - exploring the 'OMEs' with individual resolution.** HIBIT conference 2011 proceedings

*[BINDER 2] Binder, H., Fasold, M., Hopp, L., Cakir, V., Bergen, M.v. & Wirth, H.: **Portraying high-dimensional OMICs data with individual resolution.** CAMDA conference 2011 proceedings

## Journal Publications

[VOIGT 1] Voigt, D., Wirth, H., & Dilger, W.: **A Computational Model for the Cognitive Immune System Theory Based on Learning Classifier Systems.** ICARIS conference 2007 proceedings

[vBERGEN 1] Bergen, M.v., Eidner, A., Schmidt, F., Murugaiyan, J., Wirth, H., Binder, H., Maier, T. & Rösler, U.: **Identification of harmless and pathogenic algae of the genus Prototheca by MALDI-MS.** Proteomics. Clinical Applications 2009

[MÖRBT 1] Mörbt, N., Tomm, J., Feltens, R., Mögel, I., Kalkhof, S., Murugesan, K., Wirth, H., Vogt, C., Binder, H., Lehmann, I. & Bergen, M.v.: **Chlorinated Benzenes Cause Concomitantly Oxidative Stress and Induction of Apoptotic Markers in Lung Epithelial Cells (A549) at Nonacute Toxic Concentrations.** Journal of Proteome Research 2010

[BINDER 3] Binder, H., Wirth, H. & Galle, J.: **Gene expression density profiles characterize modes of genomic regulation: theory and experiment.** Journal of Biotechnology 2010

[OBERBACH 1] Oberbach, A., Blüher, M., Wirth, H., Till, H., Kovacs, P., Kullnick, Y., Schlichting, N., Tomm, J., Rolle-Kampczyk, U., Murugaiyan, J., Binder, H., Dietrich, A. & Bergen, M.v.: **Integrated serum proteomic and metabolomic profiling reveals association of the complement system with obesity and identifies novel markers of body fat mass changes.** Journal of Proteome Research 2011

[ARNOLD 1] Arnold, A., Naaldijk, Y., Fabian, C., Wirth, H., Binder, H., Nikkhah, G., Armstrong, L. & Stolzing, A.: **Reprogramming of human Huntington fibroblasts using mRNA.** ISRN Cell Biology 2011

*[WIRTH 1] Wirth, H., Löffler, M., Bergen, M. v., & Binder, H.: **Expression cartography of human tissues using self-organizing maps.** BMC Bioinformatics 2011

*[WIRTH 2] Wirth, H., Bergen, M.v., Murugaiyan, J., Rösler, U., Stokowy, T. & Binder, H.: **MALDI-typing of infectious algae of the genus Prototheca using SOM portraits.** Journal of Microbiological Methods 2012

*[HOPP 1] Hopp, L. 1st, Wirth, H. 1st, Fasold, M. & Binder, H.: **Portraying the expression landscapes of cancer subtypes: a glioblastoma multiforme and prostate cancer case study.** BMC Proceedings CAMDA 2011, in press

*[WIRTH 3] Wirth, H., Bergen, M. V., & Binder, H.: **Mining SOM expression portraits: Feature selection and integrating concepts of molecular function.** Under review

*[STEINER 1] Steiner, L. 1st, Hopp, L. 1st, Wirth, H., Galle, J., Binder, H., Prohaska, S. & Rohlf, T.: **A global genome segmentation method for exploration of epigenetic patterns.** Submitted

*[BÖRNO 1] Börno, S., Wirth, H. et al.: **Analysis of genome-wide DNA-methylation patterns in prostate cancer allows for detection of a highly predictive marker set for PCa diagnosis.** Submitted

## Book chapters

*[CAKIR 1] Cakir, V. 1st, Wirth, H. 1st, Hopp, L. & Binder, H.: **Portraying miRNA expression landscapes using machine learning.** Methods in Molecular Biology 2012, in press

# Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 22.05.2012

Henry Wirth