Bioinformatics Advance Access published June 10, 2015

Gene expression

oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on Bioconductor

Henry Löffler-Wirth[1,*], Martin Kalcher[1] and Hans Binder[1]

[1]Interdisciplinary Centre for Bioinformatics, Leipzig University, Härtelstraße 16-18, Leipzig 04107, Germany

Associate Editor: Dr. Ziv Bar-Joseph

ABSTRACT

Motivation: Comprehensive analysis of genome-wide molecular data challenges bioinformatics methodology in terms of intuitive visualization with single-sample resolution, biomarker selection, functional information mining and highly granular stratification of sample classes. oposSOM combines those functionalities making use of a comprehensive analysis and visualization strategy based on self-organizing maps (SOM) machine learning which we call 'high-dimensional data portraying'. The method was successfully applied in a series of studies using mostly transcriptome data but also data of other OMICs realms. oposSOM is now publicly available as Bioconductor R package.

1 INTRODUCTION

Bioinformatics tools are needed which allow to statistically, functionally and visually summarise high-dimensional data such as transcriptome studies at different levels of resolution ranging from individual samples and genes to sample classes and expression modules of co-regulated genes. For this purpose, we developed a bioinformatics analysis pipeline based on self-organizing map (SOM) machine learning which facilitates a holistic view on this data (Wirth *et al.*, 2011; Wirth, von Bergen, and Binder, 2012). We termed this technique 'high-dimensional data portraying'. It subsumes the visualization of the data landscape of each individual, a series of downstream bioinformatics and –statistics analysis options and the detailed and comprehensive reporting of the results. We have chosen SOM machine learning as backbone because it combines strong clustering, dimension reduction, multidimensional scaling and visualization capabilities which have been shown to be advantageous compared to alternative methods such as clustering heatmaps and negative matrix factorization when applied to molecular high-throughput data (see (Wirth et al., 2011) and references cited therein). We complemented the basal SOM algorithm with a sophisticated data analysis workflow including visualization of the individual feature landscapes, statistical testing for differential features and biomarker selection, mining of biological function, and also sample diversity analysis to assess classes of samples. oposSOM continues and largely extends the scope of a previous SOM-based expression analysis tool, the 'gene expression dynamic inspector' (GEDI) (Eichler et al., 2003): oposSOM is under steady development, provides a multitude of sample diversity analyses and, most importantly, provides comprehensive functional annotations.

Our portraying-method has been developed in first instance for gene expression data comprising from tens up to thousands of samples (e.g. tumour specimen in patient cohorts, experimental conditions in cell line experiments). The portraying functionality is unique and suited especially for scientists who attach importance to visual control and intuitive perception of complex data. The software was applied in a series of previous studies aiming at discovering the gene expression landscapes of healthy human tissues (Wirth *et al.*, 2011), of cancer subtypes (Hopp, Wirth, *et al.*, 2013; Hopp, Lembcke, *et al.*, 2013; Reifenberger *et al.*, 2014) and of stem cell development (Charbord *et al.*, 2014). Further applications addressed the integrative analysis of mRNA and miRNA expression data (Cakir *et al.*, 2014), the proteome of algae (Wirth, von Bergen, Murugaiyan, *et al.*, 2012), whole genome histone modification patterns (Steiner *et al.*, 2012) and the genomic diversity of human ethnicities (Binder and Wirth, 2015).

2 FUNCTIONALITY

Package usability. The oposSOM package requires the input of gene-centered expression data solely, e.g. as pre-processed microarray intensity data or RNA-seq read counts in log-scale. All other program parameters are optional (see package vignette). An image of the analysis environment is stored upon completion of the oposSOM run.

Workflow. oposSOM comprises a multitude of analysis modules whose functionalities were described in detail in our previous publications. An illustration of the workflow and a complete list of methods implemented in the package can be found in the supplementary material. In brief, the package fulfils the following tasks:

- The SOM space obtained from the training process is characterized by several supporting maps and profiles providing, e.g., the number of genes mapped to each meta-gene.
- Samples are individually portrayed in PDF report sheets allowing the detailed examination of their expression landscapes and especially to identify modules of co-expressed genes.
- Feature maps, reports and lists allow feature selection and evaluation of their statistical significance.
- Gene set enrichment analysis of the expression modules provides their functional context based on a large collection of predefined gene sets.
- Sample diversity analysis and class discovery is performed using multiple algorithms (e.g. hierarchical clustering, correla-

[*]To whom correspondence should be addressed.

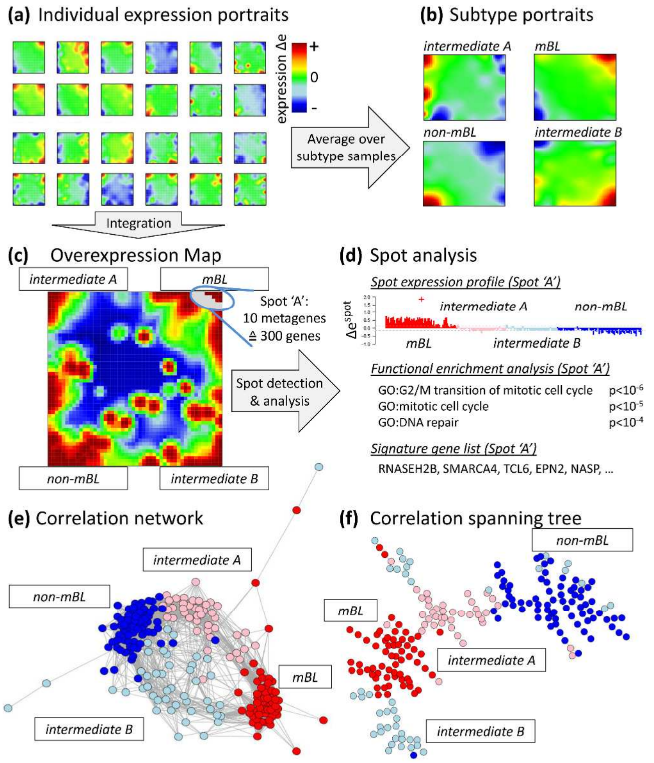


Fig. 1. oposSOM analysis of a cohort of 231 mature B-cell lymphoma cases (see text).

tion spanning tree) and different metrics (Euclidean distance, Pearson's correlation coefficient).

Results. oposSOM stores the results in a defined folder structure. These results comprise a variety of PDF documents, which provide extensive information about the systems studied (for example plots and images of the input data, supplementary descriptions of the SOM generated and associated metadata, the sample diversity landscape and also functional annotations). The PDF reports are complemented by CSV spreadsheets, which render the complete information accessible. Detailed descriptions of the algorithms and visualizations were given in our previous publications (Wirth *et al.*, 2011; Wirth, von Bergen, and Binder, 2012; Wirth, 2012; Hopp, Wirth, *et al.*, 2013; Hopp, Lembcke, *et al.*, 2013). HTML files are generated to provide easy access to the analysis results via an intuitive and descriptive interface. A *Summary.html* can be found in the results folder created by oposSOM. We recommend new users to browse the results using this interface.

3 USE CASE: PORTRAYING OF CANCER SUBTYPES

We applied oposSOM to patient expression data of mature aggressive B-cell lymphomas to characterize their genome wide expression landscapes in terms of four distinct molecular subtypes which associate with differing clinical phenotypes and survival prognosis (Hopp, Lembcke, *et al.*, 2013).

Fig. **1** provides an overview of the analysis steps: The expression portraits visualize the expression landscape of each individual sample (Fig. **1**a) and of each subtype (Fig. **1**b). Red and blue 'spots' in the portraits can be assigned to modules of co-expressed genes up- and down-regulated in the respective sample/subtype, respectively. The subtype portraits in Fig. **1**b immediately reveal distinct and subtype-specifically over-expressed expression modules emerging as red spots located near the corners of the respective portrait.

All expression modules detected are summarized in the spot-overview map (Fig. **1**c). Each module is characterized in terms of the list of genes included, their mean expression profile in all samples studied and a list of enriched gene sets enabling functional interpretation (Fig. **1**d). Sample diversity plots, e.g. based on correlation network and correlation spanning tree algorithms visualize multivariate similarity relations between the samples (Fig. **1**e & f). They support our definition of the molecular subtypes by forming well separated sample clusters.

A second use case addressing the expression landscapes of human tissues can be found in the supplement. It illustrates advantages of oposSOM data portraying compared to a 'traditional' two-way clustering heatmap.

4 CONCLUSION

oposSOM bundles a series of sophisticated analysis methods with intuitive visualization options to study high-dimensional data with the special focus on gene-centered expression data. It is designed for a broad user community ranging from bioinformaticians with demands for comprehensive analyses in a sophisticated workflow to application-oriented experimenters with needs in intuitive visualization options for their data.

ACKNOWLEDGEMENTS

This publication is supported by the Federal Ministry of Education and Research (BMBF), project grant No. FKZ 031 6166 (MMML-MYC-SYS) and FKZ 031 6065A (HNPCC-SYS).

REFERENCES

Binder,H. and Wirth,H. (2015) Analysis of Large-Scale OMIC Data Using Self Organizing Maps. In, Encyclopedia of Information Science and Technology, Third Edition, M. Khosrow-Pour, Editor. 2014, IGI global. p. 1642-1654.

Cakir,M.V. et al. (2014) MicroRNA Expression Landscapes in Stem Cells, Tissues, and Cancer. Methods Mol. Biol., 1107, 279–302.

Charbord,P. et al. (2014) A Systems Biology Approach for Defining the Molecular Framework of the Hematopoietic Stem Cell Niche. Cell Stem Cell, 15, 376–391.

Eichler,G.S. et al. (2003) Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles. Bioinformatics, 19, 2321–2.

Hopp,L., Lembcke,K., et al. (2013) Portraying the Expression Landscapes of B-Cell Lymphoma - Intuitive Detection of Outlier Samples and of Molecular Subtypes. Biology (Basel)., 2, 1411–1437.

Hopp,L., Wirth,H., et al. (2013) Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. Syst. Biomed., 1, 1–23.

Reifenberger,G. et al. (2014) Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. Int. J. Cancer, 135, 1822–31.

Steiner,L. et al. (2012) A global genome segmentation method for exploration of epigenetic patterns. PLoS One, 7.

Wirth,H. (2012) Analysis of large-scale molecular biological data using self-organizing maps. Dissertation thesis, University of Leipzig

Wirth,H. et al. (2011) Expression cartography of human tissues using self organizing maps. BMC Bioinformatics, 12, 306–352.

Wirth,H., von Bergen,M., Murugaiyan,J., et al. (2012) MALDI-typing of infectious algae of the genus Prototheca using SOM portraits. J. Microbiol. Methods, 88, 83–97.

Wirth,H., von Bergen,M., and Binder,H. (2012) Mining SOM expression portraits: feature selection and integrating concepts of molecular function. BioData Min., 5, 18–63.

oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on Bioconductor

Henry Löffler-Wirth[1,*], Martin Kalcher[1] and Hans Binder[1]

[1]Interdisciplinary Centre for Bioinformatics, Leipzig University, Härtelstraße 16-18, Leipzig 04107, Germany

Supplementary text

1. High-dimensional data portraying of healthy human tissues

In a pilot publication, we analyzed microarray expression data of 67 healthy human tissue specimen (Wirth *et al.*, 2011). Our method transformed the whole genome expression pattern of about 22,000 genes into a SOM coordinate system, which allowed intuitive visualization of transcriptional activity of each sample in terms of mosaic portraits. They exhibit characteristic spatial color patterns serving as fingerprint of the transcriptional activity of the respective tissue sample (see Figure S 1a and (Wirth *et al.*, 2011) for details), and allow for direct comparison of the expression of individual samples in a simple and intuitive way: In particular, each tile of the portraits refers to one metagene. The metagenes act as representative of disjoint clusters of single genes with similar expression profiles. The color gradient was chosen to visualize over- or underexpression of the metagenes in the particular sample compared with the mean expression level of each metagene in the pool of all samples studied: 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. The emerging spot patterns enables identification of clusters of signature genes, called expression modules, which are activated or deactivated in a sample specific fashion (Wirth, von Bergen, and Binder, 2012).

Some tissues combine the characteristic spot patterns of other tissues (Figure S 1a). For example, the expression portrait of tongue shows the typical over-expression spot evident in the portraits of epithelial tissues (e.g. oral mucosa) but also the over-expression spot typically found in muscle tissues (e.g. skeletal muscle) as well, thus directly identifying tongue as a ‘mucosa covered muscle’.

The two-way clustering heatmap of the tissue data is shown in Figure S 1b for comparison. It has its strengths in summarizing expression patterns of a large number of samples, however the individualized examination of specific sample characteristics is less intuitive compared with SOM portraying. In particular, the modular combination of expression patterns of muscle and mucosa are not evident in the heatmap column representing the tongue sample (Figure S 1b).

Further, the use of metagene instead of single gene expression data leads to an increased discriminating power in downstream agglomerative analyses such as hierarchical clustering and independent component analysis (Wirth *et al.*, 2011). In consequence, metagenes can be seen as a natural choice to detect context-dependent patterns of gene expression in complex data sets.

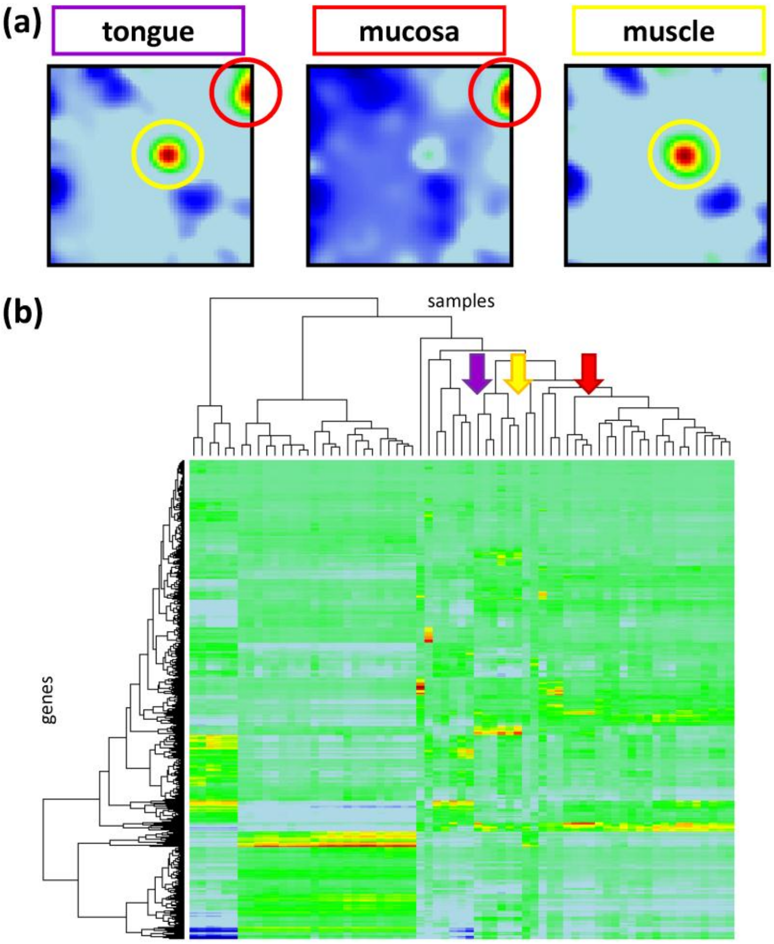


Figure S 1: (a) Tissue specific expression patterns combine in selected expression portraits: The portrait of tongue shows two spots of up-regulated metagenes. One of them is characteristic for mucosa type tissues (red circles) and the other one for muscle tissues (yellow circles). (b) Two-way clustering heatmap of the tissue data masks this combinatorics. The tongue, mucosa and muscle samples are indicated using arrows colored according to the frames in panel a.

2. Example session of the oposSOM package

As example, we make use of human tissue data downloaded from Gene Expression Omnibus repository under GEO accession no. GSE7307. It contains about 20,000 genes in more than 650 tissue samples. A subset of 12 selected tissues is provided in oposSOM. The following R-code creates the runtime environment and launches the analysis pipeline for this human tissue expression data:

```
library(oposSOM)
data(opossom.tissues)

env <- opossom.new( list(
      dataset.name = "Tissues",
      dim.1stLvlSom = 20 ) )

env$indata <- opossom.tissues
```

```
env$group.labels <- c(rep("Homeostasis", 2),
      "Endocrine",
      "Digestion",
      "Exocrine",
      "Epithelium",
      "Reproduction",
      "Muscle",
      rep("Immune System", 2),
      rep("Nervous System", 2) )

env$group.colors <- c(rep("gold", 2),
      "red2",
      "brown",
      "purple",
      "cyan",
      "pink",
      "green2",
      rep("blue2", 2),
      rep("gray", 2) )

opossom.run(env)
```

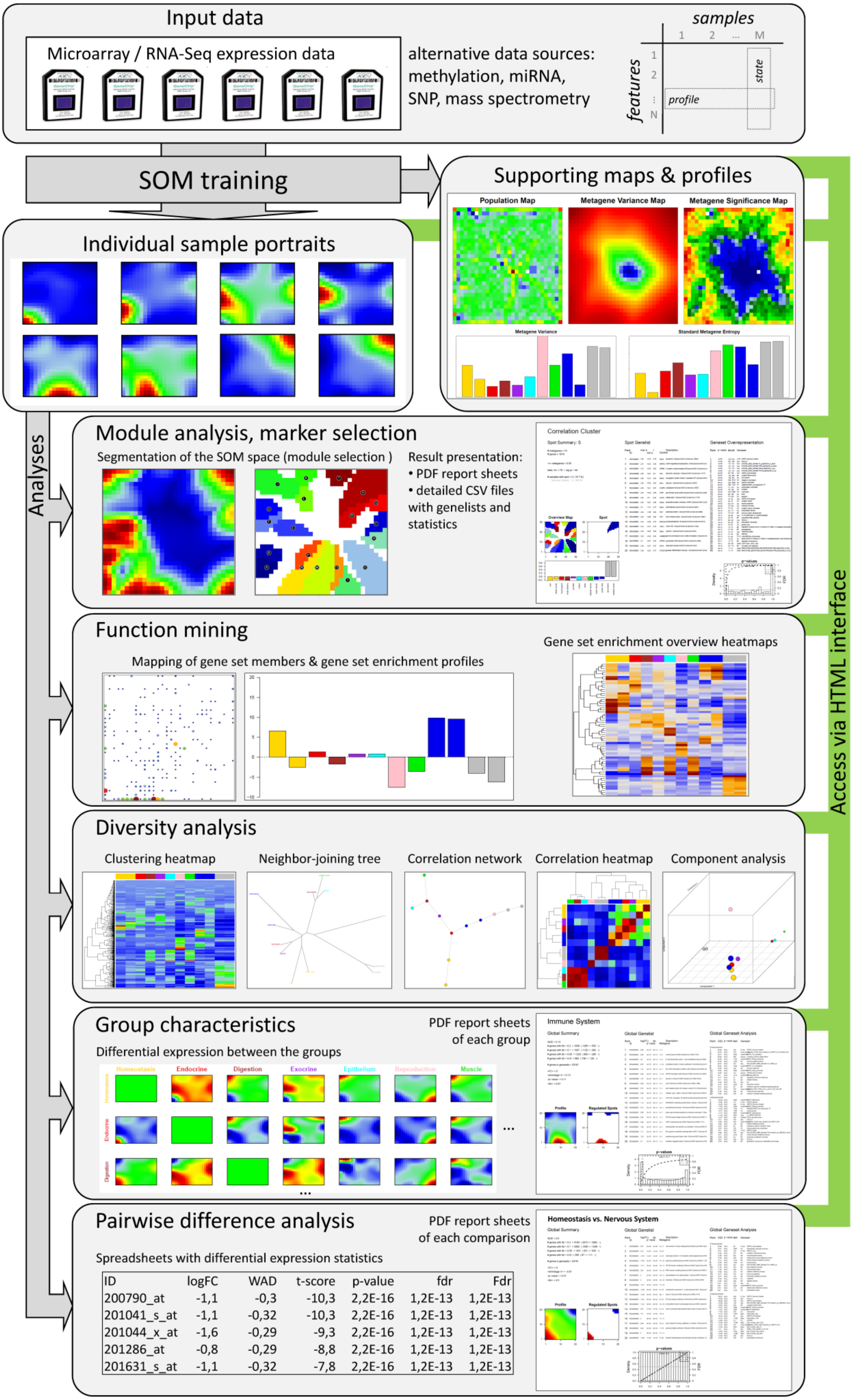
oposSOM will run through all analysis modules without further input. The tissue example will take approximately 30 minutes to finish, depending on users' hardware. Please note that subsidiary parameters are omitted here. A detailed description can be found in the package vignette.

3. oposSOM workflow

Figure S 2 shows a brief overview of the oposSOM workflow:

- Input data are given as numerical matrix with rows and columns representing genes and samples, respectively. The samples are usually quantile normalized and the genes centralized with respect to each gene's mean expression level. In general, different sources of molecular-biological data can be processed by oposSOM (Wirth, von Bergen, Murugaiyan, *et al.*, 2012; Binder and Wirth, 2015; Steiner *et al.*, 2012).
- A self-organizing map (SOM) is trained using the input data (Kohonen, 1995). Parameters of SOM training were systematically evaluated and adjusted in our previous publications (Wirth *et al.*, 2011; Wirth, von Bergen, and Binder, 2012; Wirth, 2012; Binder and Wirth, 2015). We have shown that a SOM of size between K=40x40 and 60x60 metagenes with rectangular topology and Gaussian neighborhood function provides optimal results in many applications on large-scale molecular-biological data.
- Supporting maps and profiles are generated to provide additional information about the structure of the SOM space obtained after training, and about the metagenes and associated 'single' genes (Wirth *et al.*, 2011; Wirth, 2012). They comprise:
    - The population map, presenting the number of genes mapped to each individual metagene
    - Maps of metagene variance & entropy and significance of differential expression
    - Profiles of different sample entropy measures
    - Profiles of topological characteristics of the samples' expression portraits
- Visualization of the samples in terms of expression portraits exhibits characteristic spatial color patterns and serves as fingerprint of the transcriptional activity (Wirth *et al.*, 2011). Additional portraying options are given by:

    - Alternative color scales such as WAD and loglog-fold-change (Wirth *et al.*, 2011)
    - Rank maps according to differential feature analyses (fold-change, shrinkage-t test and WAD-score) (Wirth, von Bergen, and Binder, 2012)
- oposSOM applies different algorithms to segment the SOM space into distinct expression modules of co-expressed genes (Wirth *et al.*, 2011; Wirth, 2012), which are subsequently characterized using PDF reports and CSV spreadsheets.
    - Integration of all over- and underexpression spots into one summary map, respectively
    - k-means clustering of the SOM metadata space
    - Clustering of highly correlated metagenes using iterative pooling according to a threshold criterion
- Function mining of the samples and expression modules is achieved by gene set enrichment analysis (Wirth, von Bergen, and Binder, 2012). It includes more than 6,000 sets of genes with known biological background derived, e.g., from gene ontology and literature. The results are provided in terms of comprehensive spreadsheets and in several report sheets:
    - Sample and module reports
    - Overview heatmaps summarizing enrichment of a large number of gene sets
    - Enrichment profiles for the individual gene sets
    - Mapping of members of each gene set into SOM space
    - Cancer hallmark enrichment analyses
    - Enrichment analyses for genes sets relating to chromosomal positions
- Sample diversity analysis applies different algorithms and distance metrics to discover the class structure of the data (Wirth *et al.*, 2011; Wirth, 2012; Hopp *et al.*, 2013):
    - Hierarchical clustering heatmaps
    - Neighbor joining clustering trees
    - Graph-based algorithms: correlation spanning tree and correlation network approaches
    - Independent component analysis (ICA)
- Group centered analyses allow for evaluation of specific characteristics of the groups defined:
    - Group specific expression portraits in different color scales and pairwise differential expression maps (fold-change, significance, fdr) directly compare expression landscapes of the groups
    - Specific functional characteristics are given within PDF repots and spreadsheets
    - Stability of the groups is estimated using correlation silhouette methods
- Differential expression analyses can be applied for pairs of samples or groups of samples. PDF report sheets and spreadsheets are generated consisting of:
    - Statistical evaluation of differential expression (Wirth, von Bergen, and Binder, 2012)
    - Pairwise differential expression portraits
    - Functional characterization using gene set enrichment analysis
- A HTML interface provides easy access to all analysis results accompanied by brief descriptions.



ID	logFC	WAD	t-score	p-value	fdr	Fdr
200790_at	-1,1	-0,3	-10,3	2,2E-16	1,2E-13	1,2E-13
201041_s_at	-1,1	-0,32	-10,3	2,2E-16	1,2E-13	1,2E-13
201044_x_at	-1,6	-0,29	-9,3	2,2E-16	1,2E-13	1,2E-13
201286_at	-0,8	-0,29	-8,8	2,2E-16	1,2E-13	1,2E-13
201631_s_at	-1,1	-0,32	-7,8	2,2E-16	1,2E-13	1,2E-13

Figure S 2: Overview of the oposSOM workflow and analysis modules. Example illustrations are shown within each module.

4. New functionalities introduced with oposSOM 1.0 on Bioconductor

The oposSOM-package release on Bioconductor is highly superior to the version released on CRAN in 2011:

- Structure of the source code was thoroughly revised to meet the requirements of Bioconductor.
- Organization and presentation of the results output was improved, accompanied with an extended HTML interface to access all results.
- A package vignette was introduced.
- New analysis modules were implemented:
  - Metagene entropy and portrait topology analyses
  - Neighbor-joining clustering of the samples
  - Correlation Network analysis of the samples
  - Enrichment profiles for the individual gene sets
  - Overview heatmaps summarizing enrichment of a large number of gene sets
  - Cancer hallmark enrichment analyses
  - Enrichment analyses for genes sets relating to chromosomal positions
  - Spot report sheets and spot correlation (wTO) networks
  - Expression portraits, differential expression analyses and functional characteristics summarized for the groups defined
  - Stability analyses of the groups using correlation silhouette methods
  - Differential expression analyses for pairs of samples or groups of samples, including differential expression portraits and functional characterization
- Primary input data can be given as Bioconductor ‘ExpressionSet’ object.

5. References

Binder,H. and Wirth,H. (2015) Analysis of Large-Scale OMIC Data Using Self Organizing Maps. In, *Encyclopedia of Information Science and Technology*, *Third Edition*, M. Khosrow-Pour, Editor. 2014, IGI global. p. 1642-1654.

Hopp,L. *et al.* (2013) Portraying the Expression Landscapes of B-Cell Lymphoma - Intuitive Detection of Outlier Samples and of Molecular Subtypes. *Biology (Basel).*, **2**, 1411–1437.

Kohonen,T. (1995) Self Organizing Maps. *Springer, Berlin, Heidelberg, New York*.

Steiner,L. *et al.* (2012) A global genome segmentation method for exploration of epigenetic patterns. *PLoS One*, **7**.

Wirth,H. (2012) Analysis of large-scale molecular biological data using self-organizing maps. Dissertation thesis, University of Leipzig, Available online: http://www.qucosa.de/fileadmin/data/qucosa/documents/10129/Dissertation%20Henry%20Wirth.pdf (accessed on 13 April 2015)

Wirth,H. *et al.* (2011) Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics*, **12**, 306-352.

Wirth,H., von Bergen,M., Murugaiyan,J., *et al.* (2012) MALDI-typing of infectious algae of the genus Prototheca using SOM portraits. *J. Microbiol. Methods*, **88**, 83–97.

Wirth,H., von Bergen,M., and Binder,H. (2012) Mining SOM expression portraits: feature selection and integrating concepts of molecular function. *BioData Min.*, **5**, 18-63.