PORTRAITING HIGH-DIMENSIONAL OMICS DATA WITH INDIVIDUAL RESOLUTION

Hans Binder[1,2]*, Mario Fasold[1,2], Lydia Hopp[1], Volkan Cakir[1], Martin von Bergen[3], Henry Wirth[1,3]*

[1] Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Germany
[2] LIFE, Leipzig Research Center for Civilization Diseases; Universität Leipzig, Germany
[3] Helmholtz Centre for Environmental Research, Dept. of Proteomics and Dept. of Metabolomics, Leipzig, Germany
* to whom correspondence should be send: binder@izbi.uni-leipzig.de, wirth@izbi.uni-leipzig.de

ABSTRACT

Self organizing maps (SOMs) portrait molecular phenotypes with individual resolution. We demonstrate the potency of the method in selected applications characterizing the diversity of gene expression in different tissues and cancer subtypes. SOM portraiting provides a comprehensive frame to describe development, differentiation and diversity in space and/or time using concepts of molecular function.

1 Introduction

The huge amount of data produced by high-throughput technologies challenges tasks such as dimension reduction, data compression and visual perception to extract reliable biological information. We here present a machine learning approach which allows to portrait the molecular phenotypic landscape with individual resolution. The method is applied to different levels of organization (cells, tissues, individuals) in different OMICs realms (mRNA and miRNA expression, proteome fingerprinting and SNP genotyping) using data from different technologies (microarrays, mass spectrometry).

2 Self Organizing maps

SOM technique has been proven in visualizing and tracking high-dimensional gene expression data in the context of cell differentiation, organogenesis and classification [1-3]. SOM clusters features by placing those with similar profiles in a series of conditions together into ‘meta-features’ and creates images that serve as molecular portraits of each sample studied. These images show characteristic textures and spot structures which can be treated as new, complex objects for next level data analysis. On the other hand, SOMs preserve the information richness of the original data allowing detailed, multivariate explorative comparisons between samples. SOMs can be generated for all kinds of high dimensional data including mRNA and miRNA expression, SNP- and proteome data obtained from techniques such as microarrays, next generation sequencing and mass spectrometry.

3 Expression portraits of human tissues

Raw data referring to different experiments using microarrays are downloaded from public data repositories such as the Gene expression omnibus (GEO). After preprocessing the expression data are feed into the SOM machine learning algorithm as described previously [4]. Our SOM method transforms the whole genome expression pattern of more than 22,000 single genes into mosaic images. Their colored textures serve as individual portraits of mRNA expression in each sample.

The tissue-specific patterns of mRNA expression can indicate important clues about gene function. Using GeneChip microarray data, we analyzed 67 different tissue types to create a SOM-compendium of gene expression in normal human tissues suitable as a reference for defining basic organ-specific gene activity.

Figure 1a shows SOM-portraits of selected tissues using a 60x60 mosaic grid. Each tile of the SOM mosaics refers to one of 3,600 metagenes characterizing the expression landscape of the tissues. These metagenes act as representatives of miniclusters of co-regulated single genes which number varies from metagene to metagene. The color gradient of the map was chosen to visualize over- and underexpression of the metagenes in the particular tissue

compared with the mean expression level in the pool of all tissues 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.

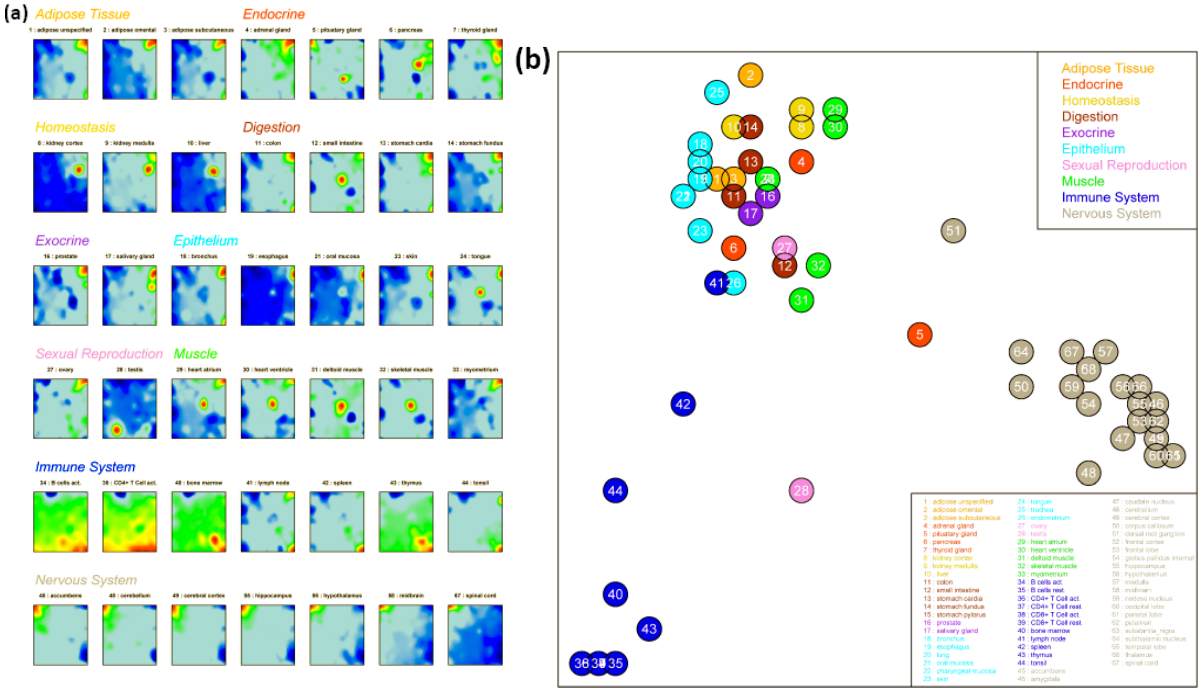

Figure 1: Expression portraits of selected human tissues (a). The 2nd level SOM (b) shows the similarity relations between the tissues. The dots are colored according to the different tissue categories.

Each mosaic exhibits characteristic spatial patterns serving as fingerprint of the transcriptional activity of the respective tissue. These expression portraits reveal a series of about one dozen stable over- and underexpression spots which selectively characterize different tissue categories such as nervous, immune system, muscle, exocrine, epithelial or adipose tissues. For example, the profiles of adipose tissues might be identified by the maroon-red overexpression spot in the right upper corner and those of nervous tissues by a similar spot in the left upper corner. Single tissues of mixed characteristics such as tongue (composed of expression spots found in muscle and epithelial tissues) can be easily identified. Some of the patterns reveal strong anticorrelation, e.g. the spot which shows overexpression in nervous tissues but underexpression in the other tissues and vice versa.

We applied 2nd level SOM analysis to establish similarity relations between the individual 1st level SOM portraits (Figure 1b). Each tissue is represented by small circles filled with the color of its previously assigned tissue category. This map offers an option to visualize similarities and differences between the samples with direct relation to the original SOM pattern. Essentially one distinguishes three main clusters namely that of nervous tissues (grey), immune system tissues (blue) and the remaining ones confirming the hypothesis that the mosaic textures also portrait tissue function.

To further consolidate this result we applied gene set enrichment analysis to the most pronounced overexpression spots. Figure 2 shows the overexpression summary map which integrates nine spots showing strong overexpression in any of the tissues. The genes associated with each spot are analyzed for enrichment of genes taken from a collection of 1454 gene sets pre-selected according to the GO-categories molecular function, molecular process and molecular component. Enrichment of the genes from each set was estimated for each of the spots using the hypergeometric distribution which provides an ordered list of gene sets ranked with decreasing significance of overrepresentation. Hence, each spot is assigned to tissues strongly overexpressing the respective metagenes and to the GO-categories of the most enriched gene sets (see the right legend in Figure 2).

This combination of SOM-spots with concepts of molecular function enables identification of subsets of tissue specific genes that potentially define key biological processes characterizing each organ. For example, spot A in the left upper corner of the SOM is clearly related to molecular processes in nervous cells according to the leading gene sets. Also other spots can be associated with distinct molecular functions such as immune system processes (spot F), sexual reproduction (spot E) or muscle contraction (spot B).

These results illustrate the general utility of the SOM-approach by constructing a map of function-related gene sets for large, heterogeneous sets of gene level expression data. This map is consistent with known tissue-specific pathways and enables verification and amendment of function-related gene sets.

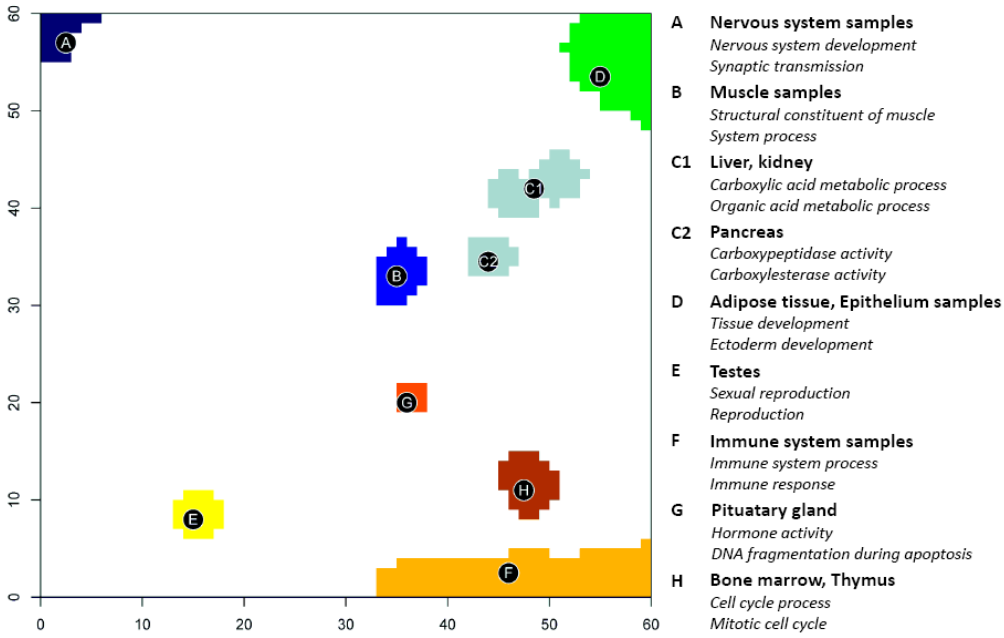

Figure 2: The overexpression summary map shows nine spots representing metagenes which are strongly overexpressed in different tissues. Enrichment of a collection of 1454 gene sets is estimated for each spot using the hypergeometric distribution [4]. The right legend assigns the two topmost enriched gene sets to the respective spots together with the tissues which overexpress this particular spot.

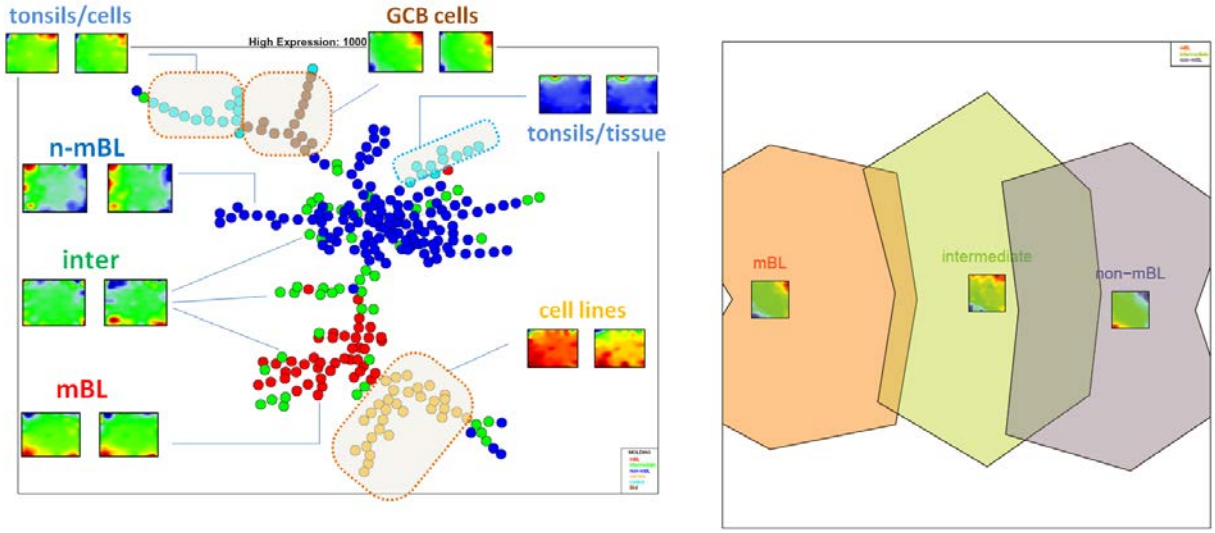

Figure 3: Maximum spanning tree of 220 B-cell lymphoma samples assigned to mBL-, non-mBL and intermediate subtypes and to cancer cell lines and controls (tonsil tissue, germinal center B-cells derived from tonsils and blood). Characteristic SOM images are shown in the figure. The 2nd level SOM in the right part reveals a virtually univariate expression signature giving rise to the one-dimensional arrangement of the three cancer subtypes in horizontal direction.

4 Molecular subtypes of B-cell Lymphoma

Aggressive B-cell lymphoma is a heterogeneous disease with recognized variability in clinical outcome, genetic features, and cells of origin. To date, transcriptional profiling has been used to highlight similarities between tumor cells and normal B-cell subtypes and to associate genes and pathways with unfavorable outcome. Transcriptional

profiling has been recently used to define B-cell lymphoma more precisely and to distinguish subgroups assigned to the molecular (mBL) and non–molecular (non-mBL) Burkitt's lymphoma signatures [5].
This study used biopsy specimens of 220 mature aggressive B-cell lymphomas in which at least 70 percent of all cells were tumor cells. Of all lymphomas, 44 were assigned to the mBL signature and 128 to non-mBL signature. 48 cases could not be assigned unambiguously to either of the two groups. They form an intermediate group, representing the transition zone between the mBL and non-mBL groups. Microarray data are available under GEO accession number GSE4475.
Figure 3 shows similarity relations between the SOM images of the cancer samples in terms of maximum spanning tree (MST, left part) and 2nd level SOM (right part) together with characteristic SOM portraits of the different cancer subtypes. The MST also shows cancer cell lines, and different controls referring to tissue (tonsils) and cell lines (Germinal center B-cells derived from tonsils or blood). The controls group into distinct regions near the non-mBL specimens whereas the cancer cell lines show strong similarity with the mBL subtype. The portraits of the cancer subtypes occupy three distinct, partly overlapping areas in the 2nd level SOM. Importantly, the three groups arrange virtually along a line in the horizontal direction whereas the vertical dimension essentially covers the intra-group variability of the data. The small mosaics depicted in the center of each of the three areas are mean expression profiles averaged over all individual pattern of each group. These mean SOM of the mBL and non-mBL groups reveal a relatively unstructured texture with one over- and one underexpression spot in two opposite corners of the map. This 'binary' spot pattern indicates that genes overexpressed in mBL become underexpressed in non-mBL and vice versa. Hence, both groups can be distinguished using an essentially univariate signature which, in turn, explains the one-dimensional arrangement of the three groups in the 2nd level SOM. Gene set enrichment analysis shows that genes related to the GO-terms 'cell-cycle' and 'DNA-repair' accumulate in the mBL overexpression spot in the right upper corner whereas genes related to 'cell adhesion' and 'inflammation/immune response' dominate in the non-mBL overexpression spot.

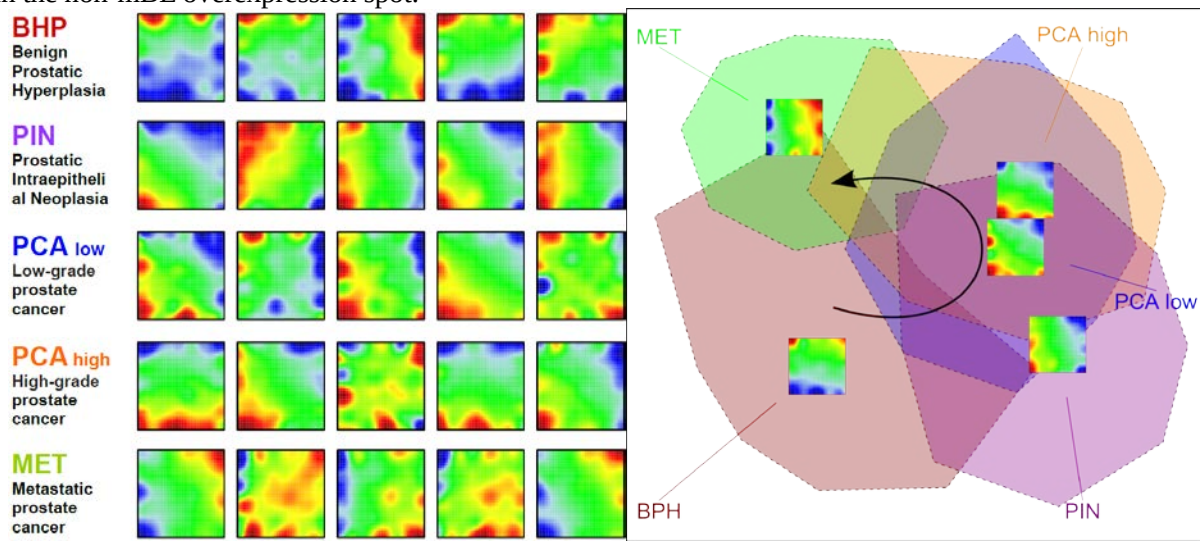

Figure 4: Expression portraits of progressing prostate cancer: The left part shows selected SOM of individual laser dissected samples. In total we included the following sample sizes: 22 (BPH), 13 (PIN), 12 (PCAlow), 20 (PCAhigh), 17 (MET). They occupy wide regions in the 2nd level SOM as illustrated by the colored polygons. The mean SOM portraits per stage are located in the center of the respective polygon. Note that the spot pattern in theses maps virtually rotates with progressing cancer giving rise to a U-shaped trajectory in the map (see arrow) in contrast to the virtually one-dimensional distribution of the Lymphoma subtypes.

5 Prostate cancer progression

Despite efforts to profile prostate cancer, the genetic alterations and biological processes that correlate with the observed histological progression are largely unclear. Prostate cancer is most commonly graded using the Gleason grading system, which relies entirely on the architectural pattern of cancerous glands. The underlying expression signatures and the processes driving the different architectural patterns are mostly unknown. A recent microarray study [6] addresses the molecular mechanisms associated with gene expression changes in the course of prostate cancer progression using laser-capture microdissection to isolate 101 specific cell populations from 44 individuals.

The samples are assigned to five stages of cancer progression ranging from benign prostatic hyperplasia (BPH) and prostatic interepithelial neoplasia (PIN) to low-grade (Gleason score 3), high-grade (4-5, PCA) and metastatic (MET) prostate cancer.
We transformed the gene expression data (available under GSE 6099) into SOM portraits revealing a relatively diverse texture landscape even within the sample groups assigned to the different stages of progression (see Figure 4). In the 2nd level SOM representation these groups occupy extended regions of strong mutual overlap. Despite their fuzziness the stage related areas order along a U-shaped path with progressing cancer. To get further insights into this trend we calculated mean SOM mosaics averaged over all individual samples of each group (Figure 4). These mean portraits of each stage reveal that the areas of over- (red) and under- (blue) expression rotate in counterclock direction along the edges of the maps. This result clearly shows that the different groups indeed form an ordered developmental series with partly overlapping microscopic states in consecutive stages. Moreover, the partly circular character of the trajectory reflects the fact that a significant part of the genes are similarly expressed in the final MET-stage and in the initial BPH-stage, but differently expressed in the intermediate PIN- and PCA-stages. The detailed gene-level analysis reveals that genes related to protein biosynthesis and ETS (E26 transformation specific) target genes show these properties and, moreover, demarcate critical transitions in cancer progression [6].

6 Conclusions and outlook

SOM machine learning enables the kaleidoscopic and intuitive view on high-dimensional data without loss of primary information. It provides a general frame for analytic tasks such as feature selection, integrating concepts of molecular function and systems tracking with individual resolution. The method extracts meta-features such as meta-genes, -peaks and –alleles expressing basal modes of systems behaviour important for higher-level, holistic analysis. Ongoing tasks also address issues such as ‘interOMICs’ integration and associations and the extension of the method to next generation sequencing and other data types. Examples will be given in the talk.

Acknowledgements: The project LIFE is financially supported by the European EFRE fund and the State of Saxony. HW and MF were kindly supported by the HIGRADE Graduate school and by the European Social Fund, respectively.

7 References

[1] Tsigelny IF, Kouznetsova VL *et al*: **Analysis of Metagene Portraits Reveals Distinct Transitions During Kidney Organogenesis**. *Sci Signal* 2008, **1**(49):ra16-.

[2] Huang S, Eichler G *et al*: **Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network**. *Phys Rev Lett J1 - PRL* 2005, **94**(12):128701.

[3] Mar JC, Quackenbush J: **Decomposition of Gene Expression State Space Trajectories**. *PLoS Comput Biol* 2009, **5**(12):e1000626.

[4] Wirth H, Loeffler M *et al*: **Expression cartography of human tissues using self organizing maps**. *BMC Bioinformatics* 2011, **in review**:see preprint http://precedings.nature.com/documents/5825/version/5821.

[5] Hummel M, Bentink S *et al*: **A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling**. *N Engl J Med* 2006, **354**(23):2419-2430.

[6] Tomlins SA, Mehra R *et al*: **Integrative molecular concept modeling of prostate cancer progression**. *Nat Genet* 2007, **39**(1):41-51.