MOLECULAR PHENOTYPIC PORTRAITS – EXPLORING THE 'OMES' WITH INDIVIDUAL RESOLUTION

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

[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

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, mRNA and miRNA fingerprints of stem cells, the proteome landscape of algae and genomic relations between humans from different populations. It is further shown that SOM portraiting provides a comprehensive frame to describe development, differentiation and diversity in space and/or time.

1 INTRODUCTION

Molecular biology is presently flooded by masses of high-throughput data generated by newest generation sequencing and microarrays as well as by protein shotgun experiments, for example. This huge amount of data challenges tasks such as dimension reduction, data compression and visual perception to extract reliable biological information. These challenges are still intensified by the fact that the new techniques enable to pursue 'personalized' approaches aiming at resolving and understanding biological variability on the level of individual molecular pheno- and genotypes.

We here portrait the molecular phenotypic landscapes with individual resolution using SOM machine learning. 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) to survey its potency. We performed also second level agglomerative analysis to track the relations between the individual portraits in space and/or time to characterize development, differentiation and diversity.

2 SELF ORGANIZING MAPS

SOM technique has been proven useful 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.

Our SOM algorithm starts with raw experimental data referring to multiple conditions such as different individuals in a patient cohort study, different treatments in a treatment-versus-control investigation or different time points in a time-series experiment. The raw data are subjected to preprocessing which includes calibration and normalization tasks to remove systematic biases from the data and to minimize the scattering between individual samples and to transform them into unique scale chosen typically relative to a suited reference state.

In the next step, the preprocessed data are entered into the unsupervised machine learning program to train a SOM representing information-rich diagrams as illustrated in Figure 1. The SOM method applies a neural network algorithm to project high dimensional data onto a two-dimensional visualization space [4-5]. The algorithm initializes a sufficient number of so-called meta-feature profiles and arranges them in to a two-dimensional grid. They represent vectors of dimensionality given by the number of conditions studied. Then each of the respective vectors of all measured single features is associated with the meta-feature of closest similarity which, in turn is adjusted so that it more closely resembles the profiles of the associated single features. An iterative procedure progressively optimizes the similarity between all meta- and single features where also the meta-features of adjacent tiles are adjusted using a distance dependent weight. The resulting final SOM consists of regions of similar meta-feature profiles. Each of them represents a minicluster of single features with similar profiles. The profiles of the meta-features can be understood as a sort of 'eigenmodes' characterizing the multitude of single profiles inherent in the data.

For each condition studied a mosaic image is constructed by color-coding the tiles according to the amplitude of the respective meta-features. This leads to a coherent texture that is characteristic for each sample. Since the SOMs assign the

978-1-4673-4431-9/11//$31.00 ©2012 IEEE

same meta-feature to the same tile in all samples of a series, they can be directly compared to each other allowing immediate identification of interesting groups of features.

Typically, the number of tiles to 'pixelate' the SOM is about one order of magnitude smaller than the number of single features available giving rise to a smooth spot pattern where each spot clusters metagenes showing local maximum (or minimum) amplitudes of the respective meta-features in the respective sample.

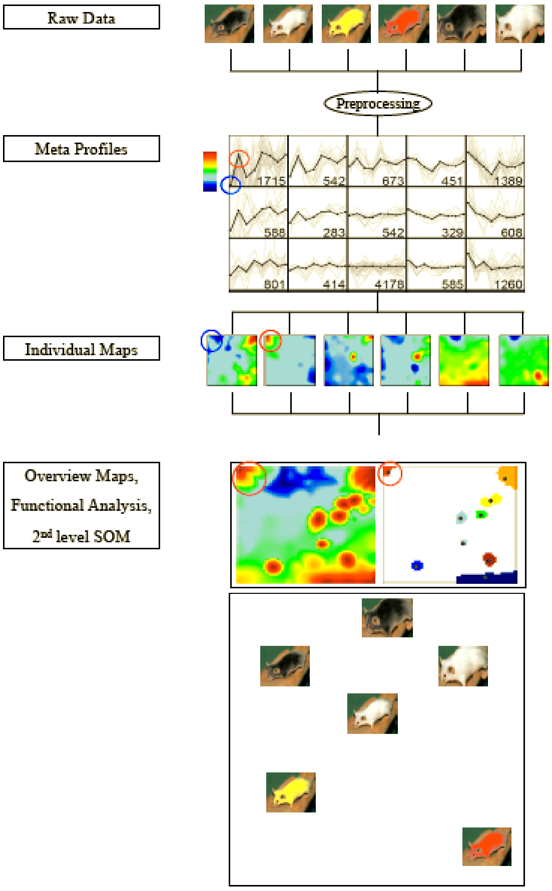

Figure 1: Molecular phenotypic portraits are obtained using SOM machine learning. It projects high dimensional molecular data of a series of different samples into a two-dimensional grid where each tile refers to one characteristic meta feature (thick lines). Different numbers of individual features with similar profiles (thin lines) are assigned to each meta profile (see numbers in the grid). The meta features are then transformed into one individual map per condition. The tiles in these maps are color-coded to indicate high or low amplitudes of the respective meta profile. The parallel evaluation of multiple samples allows to link their overall profile pattern. For example, the metagene of the tile in the left upper corner of the mosaic is underexpressed in sample no. 1 and overexpressed in sample no. 2 as indicated by the red and blue circles and the color-code in the respective mosaics. So-called overview maps can be extracted from the series of individual maps which summarize, e.g., all spots due to high amplitudes of the meta profiles. Similarity relations between the individual maps are shown in the 2nd level SOM.

The method effectively compresses the original high-dimensional data in two respects: Firstly, ten thousands of correlated profiles of single features are collected into a few thousand clusters where each is characterized by one representative meta-feature. Secondly, the textures of the obtained SOM are decomposed into a few (typically less than one dozen) spots of similar amplitudes of the meta-feature. This double compression sequentially applies global and local criteria taking into account the correlated behavior of the features in all samples and their amplitudes in the different samples as well.

For examination of the similarities between the individual maps we used the respective meta features instead of single features which provides better results in terms of sensitivity and specificity [6]. So-called second level SOM analysis aggregates the samples studied into one map which directly visualizes their mutual similarities. 2nd level SOM analysis uses the meta-feature profiles as input and then clusters the samples and not the features as in 1st level SOM analysis. Each tile of the 2nd level SOM mosaic characterizes the profile of a representative meta-sample. In addition, we generate maximum spanning trees (MST) which visualize the mutual correlations between the individual maps considering all pairwise combinations of their meta-features.

3 mRNA EXPRESSION PORTRAITS

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 [6]. 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 (see Figure 1 for a schematic overview).

3.1 Transcriptome atlas of human tissues

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 2a 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.

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.

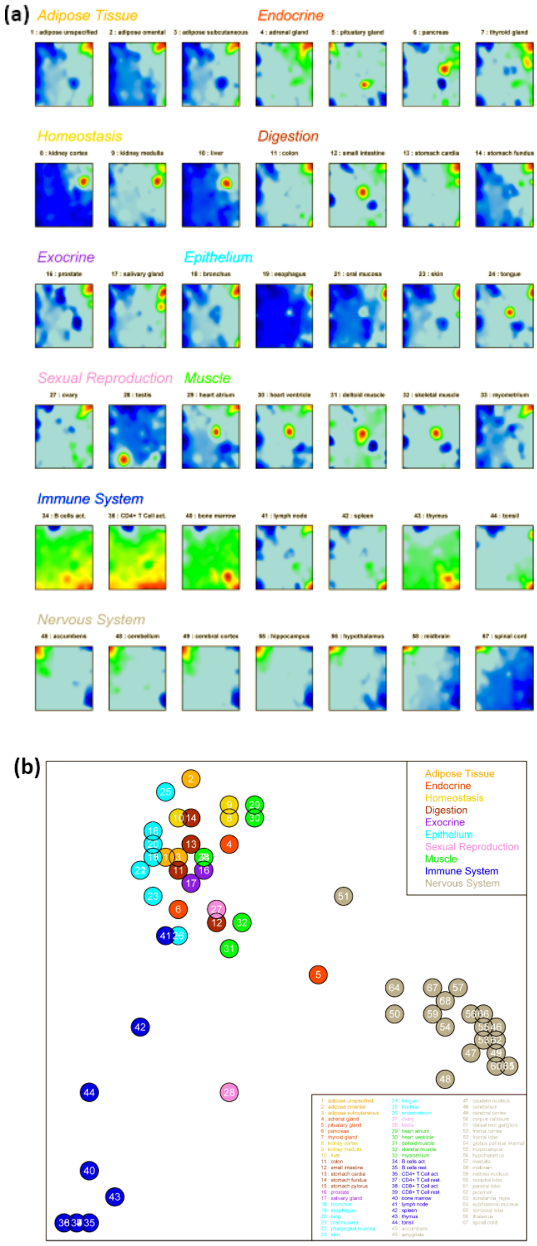

Figure 2: Expression portraits of selected human tissues. The 2nd level SOM shows the similarity relations between the tissues. The dots are colored according to the different tissue categories.

We applied 2nd level SOM analysis to establish similarity relations between the individual 1st level SOM portraits (Figure 2b). 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 [7]. Figure 3 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 3).

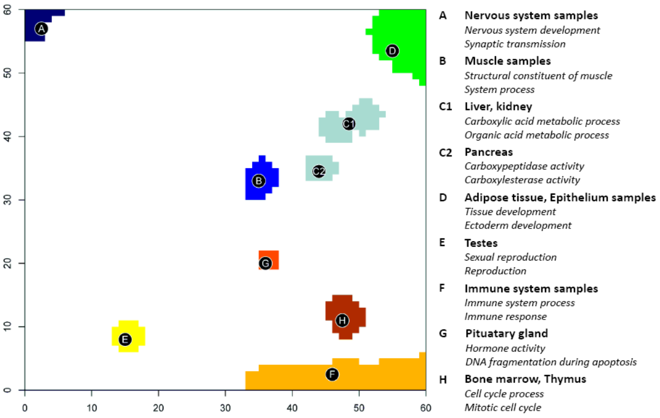

Figure 3: The overexpression summary map shows nine spots representing metagenes which are strongly overexpressed in different tissues. Enrichment of a collection of 1454 gene sets is estimated for each spot using the hypergeometric distribution. The right legend assigns the two topmost enriched gene sets to the respective spots together with the tissues which overexpress this particular spot.

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.

3.2 Disentangling 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 [8].

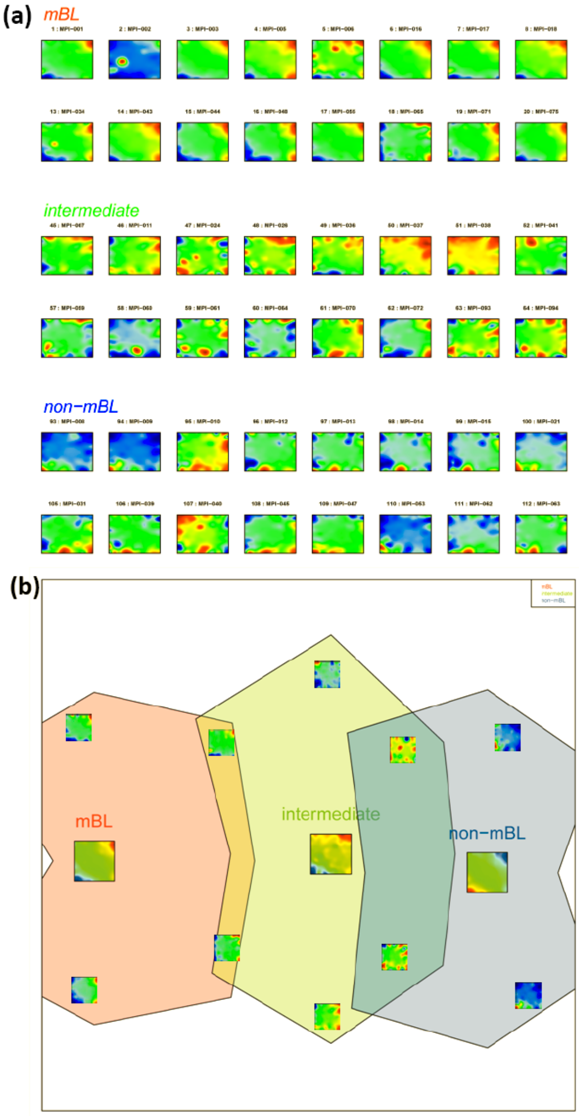

Figure 4: Expression portraits of B-cell lymphoma cancer subtypes. The 2nd level SOM in the part below reveals a virtually univariate expression signature giving rise to the one-dimensional arrangement of the three groups in horizontal direction.

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 4a shows individual SOM-portraits of all three groups illustrating the heterogeneity of their expression patterns. These individual portraits occupy three distinct, partly overlapping areas in the 2nd level SOM (Figure 4b). 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.

The maximum spanning tree of the B-cell lymphoma samples provides an alternative view on the heterogeneity reported by the individual expression portraits (Figure 5): The mBL- and non-mBL groups aggregate into clearly separated clusters. Contrarily, samples of the intermediate type form a sort of outer layer in many branches of the mBL- and non-mBL clusters indicating that they share some of the expression characteristics with the compact groups, however in a relatively diffuse fashion.

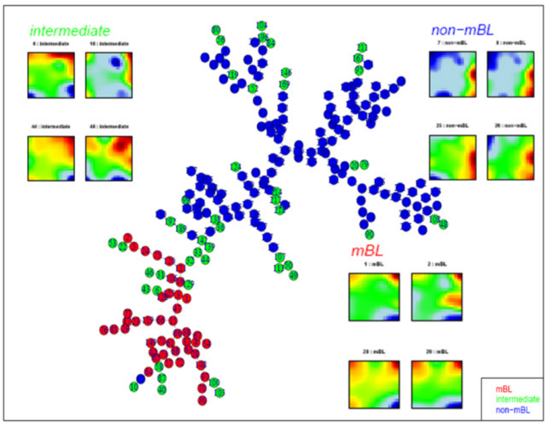

Figure 5: Maximum spanning tree of 220 B-cell lymphoma samples assigned to mBL-, non-mBL and intermediate groups.

3.3 Trajectory of 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 [9] 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.

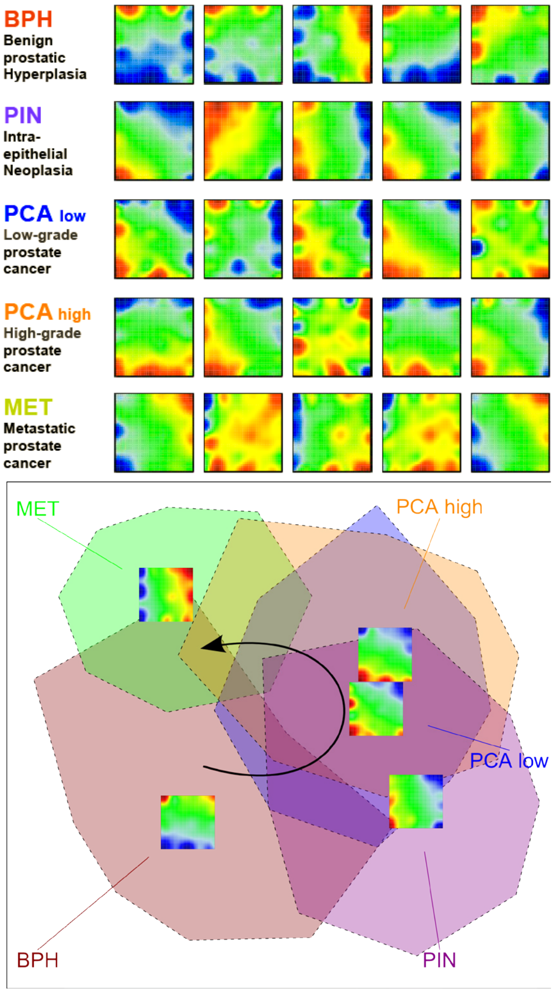

Figure 6: Expression portraits of progressing prostate cancer: The part above shows selected SOM of individual laser dissected samples. In total we included the following sample sizes: 22 (BPH), 13 (PIN), 12 (PCAlow), 20 (PCAhigh), 17 (MET). They occupy wide regions in the 2nd level SOM as illustrated 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).

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 6). 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 6). 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 [9] (see also [10] for details).

3.4 Stem cells in question

Induced pluripotent stem cells (IPS) are stem cells artificially derived from adult somatic cells by inducing a ‘forced’ expression of specific genes. IPS are similar to natural pluripotent stem cells, such as embryonic stem cells (ESC), in many respects, such as the expression of certain stem cell genes and the potency and differentiability, but the full extent of their relation to ESC is still being assessed. The opportunity of reprogramming somatic cells into IPS suggests better disease modelling in vitro and potential clinical applications.

Gene expression profiling provides an important basis for revealing the molecular mechanisms involved in pluripotency and initial differentiation events that involve embryonic stem cell populations. Utilization of microarray technology allows potential opportunities for comparison of datasets from different experiments and different stem cell lines.

Here we investigate the expression signature of induced stem cells by comparing their SOM portraits with that of their somatic progenitor cells and of ESC. Figure 7 shows the SOM gallery of differentiated somatic, derived IPS and ESC cells taken from five experiments and of B-cells for comparison. Particularly, we ask whether the expression profiles of the IPS obtained in the experiments E2-E4 resemble that of the respective ESC or not (see also [11] for the detailed discussion of the objective). Visual comparison of the SOM portraits depicted in Figure 7 provides a clear answer, namely that experiments E3 and E4 succeeded to derive IPS-like expression pattern but E2 does not. Note also that typical stemness genes such as OCT4, SOX2, LIN28 and NANOG are located in the spot overexpressed in ESC and underexpressed in somatic cells near the left lower corner of the map. The 2nd level SOM shown in Figure 7 confirms these results: The expression portraits of differentiated cells accumulate in the right part of the map whereas the ESC are found exclusively in the left part despite the scattering of the individual sample points due to different cell types and experiments performed independently in different laboratories. The IPS obtained in experiments E3 and E4 are located closely to the respective ESC whereas ‘adult germline stem’ cells (haGSC) obtained in E2 are clustered with differentiated cells in the right part of the map. Hence, the SOM portraits clearly show that haGSC were fibroblastic but not pluripotent in their gene expression profile in agreement with the results presented in [11] but in contrast to the claim given in [12].

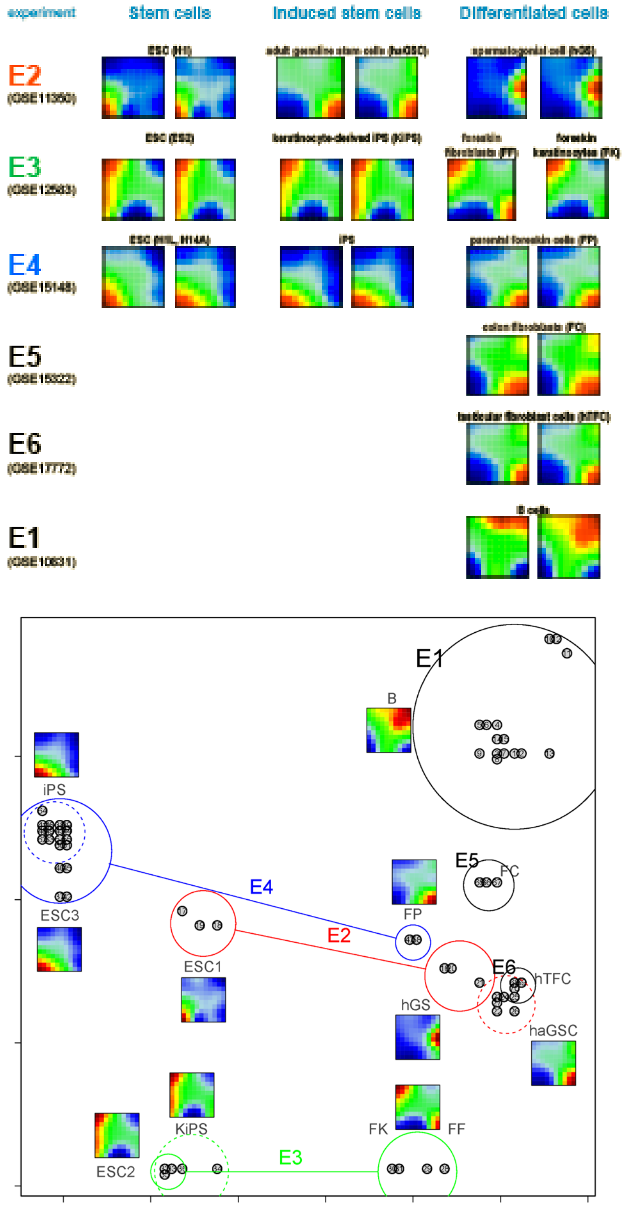

Figure 7: Expression portraits of stem cells and of differentiated cells: Data were taken from six experiments (E1…E6) which provided the SOM portraits of differentiated cells, embryonic stem cells (ESC) and induced pluripotent stem cells (IPS, part above, see also [11] and references cited therein). The 2nd level SOM illustrates the similarity relations between the different cell portraits (small filled circles): Differentiated cells and ESC of each experiment are marked by large circles connected by a line (E2 – E4). The induced stem cells obtained in the respective experiment are marked by dashed circles. Their expression portraits in E3 and E4 closely resemble that of the respective ESC whereas that in E2 does not. Selected 1st level SOM are shown for illustration.

Thus SOM transcriptional portraits of cells allow to directly evaluate their stemness and to uncover molecular mechanisms involved in differentiation and the maintenance of the undifferentiated state.

4 miRNA PORTRAITS OF STEMNESS

MicroRNAs are small noncoding RNAs that play important posttranscriptional regulatory roles by targeting mRNAs for cleavage or translational repression. Owing to their ability to regulate numerous genes, often in common pathways, miRNAs may be regulators of cellular processes, akin to transcription factors that control entire programs of cellular differentiation and organogenesis. These miRNA signatures therefore represent another layer of regulatory control for cell fate decisions in addition to histone modifications, promoter methylation, transcription factors, and other regulator elements. This level of miRNA regulation is important for self-renewal, pluripotency and differentiation of ESC but also for reprogramming of somatic cells into IPS [13].

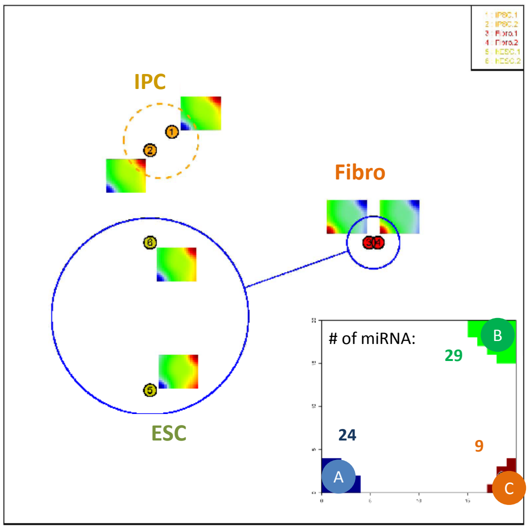

Figure 8: Similarity relations (2nd level SOM) between miRNA expression portraits of stem cells (ESC and IPS) and of differentiated cells (fibroblasts). The two mosaics per cell system refer to biological replicates. The insertion assigns the spots in the 1st level SOM and the respective number of associated miRNA.

We generated SOM-portraits of the expression of 697 miRNA in ESC, fibroblasts and derived IPS obtained in a recent microarray study [13]. The miRNA SOMs complement the respective mRNA profiles presented in the previous section. Note that SOM training of miRNA expression data uses more than ten times less features than the analogous study on mRNA expression. The obtained miRNA portraits of the IPS closely resemble that of ESC while differing strongly with respect to the precursor fibroblast (Figure 8). As in the case of mRNA one finds essentially one pair of spots (A and B) referring to features overexpressed in IPS and ESC but underexpressed in fibroblasts and vice versa. The overexpression spot in ESC and IPS provides tentative ‘stemness’ miRNA such as the mir-302 and mir-17-92 groups where the former is known to regulate switching between embryonic and mature phenotypes. The SOM portraits also reveal that spot C characterizes dissimilar miRNA expression in ESC and IPS. It contains, e.g., mir-371,372 and 373.

These results together with the mRNA profiling presented above lead to a general question regarding the fundamental difference between miRNA and mRNA activity with respect to regulatory mechanisms of differentiation and also, how both data sets can be combined to generate inter-OMICs maps allowing to identify significant associations between mRNA and miRNA expression pattern (see [14] for details).

5 PROTEOME PORTRAITS OF ALGAE

In the previous sections we applied SOM machine learning to microarray expression data. Figure 9 depicts a series of SOM mosaics obtained from another high throughput method namely MALDI-TOF mass spectrometry. It was applied to extracts of green algae from the genus Prototheca which are often overseen or mistaken for yeast in clinical diagnosis. These algae from the Chlorella family are the only known plants that cause infections in humans and animals. To overcome this diagnostic gap, a MS-method was developed for fast and reliable identification of Prototheca [15].
Most of the MS-peaks were found in the range from 4 to 20 kDa due to high abundant proteins such as ribosomal proteins and ubiquitin. The peaks showed a high reproducibility in their peak positions but high variability in the peak amplitudes. The SOM was trained using MS-peaklists of 324 samples referring to five Prototheca species where one of them splits into two genotypes. Each peaklist contains the amplitudes of 1406 peaks.

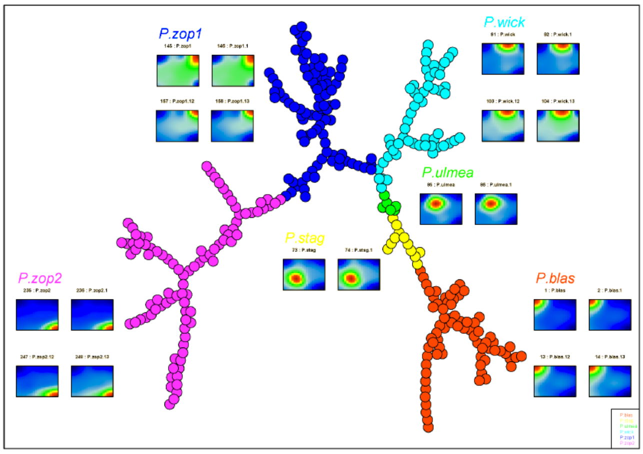

Figure 9: Maximum spanning tree (MST) of SOM-proteome portraits of 58 samples referring to six groups of algae of the genus Prototheca.

The individual SOM mosaics typically show only one overexpression spot the position of which however varies in a species-specific fashion (Figure 9). This property means that each species is characterized by a set of peaks showing high amplitudes only for this particular species but small amplitudes for all other ones. This highly species-specific pattern gives rise to the perfect separation of the six groups in the maximum spanning tree (MST) directly connecting the samples of strongest mutual correlation between their metafeatures. The correct identification of Prototheca species is of considerable importance in clinical microbiological laboratories because of its epidemiological impact and because of the broadly occurring resistance of pathogenic Prototheca isolates against antimycotics (see [16] for details).

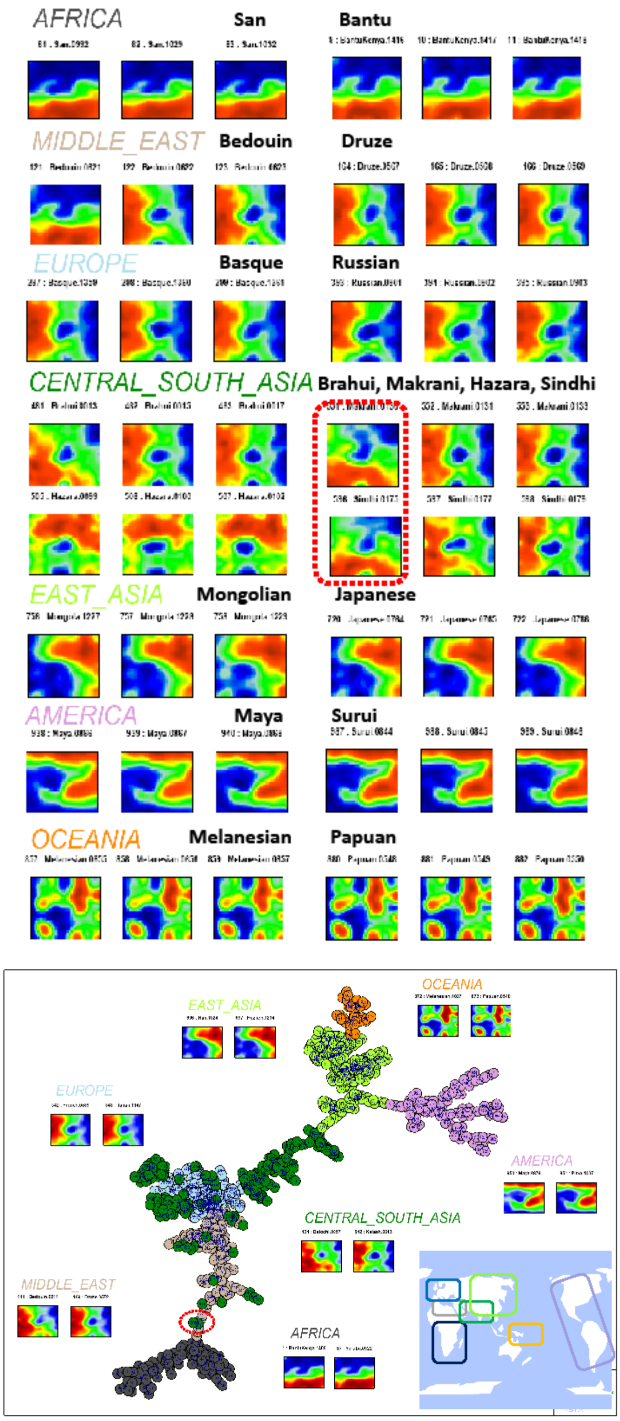

Figure 10: Worldwide SNP-genotype portraits of humans: SOM-portraits of 48 individuals from different regions of the world (part above). Red, green and blue regions refer to minor-homozygous, heterozygous and major-homozygous allelic genotypes, respectively. The MST illustrates similarity relations between the 1st level SOMs of 995 individuals belonging to 52 ethnic groups from 7 geographic regions. The inserted SOM refer to arbitrarily selected individuals from each region. The dotted circles in both parts of the figure mark the same two individuals from Makrani and Sindhi populations showing partly African genotypic characteristics.

6 HUMAN GENOTYPE PORTRAITS

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. Genotypes are available for nearly thousand individuals from the Human Genome Diversity Project, analyzed for approximately 650,000 SNPs (single polynucleotide polymorphism) with Illumina 650Y arrays [17].

To illustrate the potency of SOM-portraiting of genotype data we trained a SOM using the 50,000 most variant alleles among all individuals in this data set. Each of the considered alleles provides a trinary profile among the cohort with the values 0 (major allele), 1 (heterozygous) and 2 (minor allele). The gallery of individual maps shown in Figure 10 reveals a high diversity of textures reflecting areas of major-, heterozygous- and minor-allelic genotypes of the underlying 'meta-alleles' color-coded in blue, green and red, respectively. Most of these portraits are 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. The MST-presentation in Figure 10 clarifies this trend: the tree roughly resembles the actual geographic distribution of the populations, which, in turn, reflects the variation in population dynamics among geographic regions.

Interestingly, the SNP-portraits of a few individuals are located away from their expected geographic neighbourhood. For example, one Makrani and one Sindhi people are found close to the African group in the MST (see dotted circle in Figure 10). Detailed inspection of the respective SNP-maps reveals that the spot patterns more strongly overlap with the typical texture of the African population than with the maps of the remainder individuals of Makrani and Sindhi groups studied. Makrani are descendants of black Africans brought as slaves to Balochistan in medieval times (see, e.g., Wikipedia). The SNP-portraits not only intuitively reflect this fact but also the circumstance that other individuals from this ethnic group show SNP-portraits closely resembling that of other groups from this region such as Brahui or Sindhi. 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. Finally, also the SNP-portrait of one of the Bedouin individuals shows clearly the characteristics of black Africans. These examples illustrate the applicability of SOM machine learning to portrait genotypes with individual resolution and to judge relationships between populations and individuals in a simple and intuitive fashion.

7 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. First examples will be given in the talk.

Acknowledgements: The project LIFE is financially supported by the European EFRE fund and the State of Saxony. LH and MF were kindly supported by the European Social Fund.

8 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] Bishop CM, Svensén M *et al*: **GTM: The Generative Topographic Mapping**. *Neural Computation* 1998, **10**(1):215-234.

[5] Kohonen T: **Self-organized formation of topologically correct feature maps**. *Biological Cybernetics* 1982, **43**(1):59-69.

[6] Wirth H, Loeffler M *et al*: **Expression cartography of human tissues using self organizing maps**. *BMC Bioinformatics* 2011, **12**:306.

[7] Wirth H, von Bergen M *et al*: **Mining SOM expression portraits: Feature selection and integrating concepts of molecular function** *BioData Mining* 2012, **in press**:see preprint: http://precedings.nature.com/documents/6666/version/6661.

[8] 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.

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

[10] Hopp L, Wirth H *et al*: **Portraying the expression landscapes of cancer subtypes: a glioblastoma multiforme and prostate cancer case study**. *Nature Preceedings* 2012, **preprint**.

[11] Ko K, Arauzo-Bravo MJ *et al*: **Human adult germline stem cells in question**. *Nature* 2010, **465**(7301):E1-E1.

[12] Conrad S, Renninger M *et al*: **Generation of pluripotent stem cells from adult human testis**. *Nature* 2008, **456**(7220):344-349.

[13] Wilson KD, Venkatasubrahmanyam S *et al*: **MicroRNA Profiling of Human-Induced Pluripotent Stem Cells**. *Stem Cells and Development* 2009, **18**(5):749-757.

[14] Cakir V, Wirth H *et al*: **Portraying miRNA expression landscapes using machine learning**. *Methods of Molecular Biology* 2012, **in press**.

[15] Bergen Mv, Eidner A *et al*: **Identification of harmless and pathogenic algae of the genus Prototheca by MALDI-MS**. *PROTEOMICS - CLINICAL APPLICATIONS* 2009, **3**(7):774-784.

[16] Wirth H, von Bergen M *et al*: **MALDI-typing of infectious algae of the genus Prototheca using SOM portraits**. *Journal of Microbiological Methods* 2012, **88**(1):83-97.

[17] Li JZ, Absher DM *et al*: **Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation**. *Science* 2008, **319**(5866):1100-1104.