Personalized disease phenotypes from massive OMICs data

Hans Binder, Lydia Hopp, Kathrin Lembcke, Henry Wirth

Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany

Abstract

Application of new high-throughput technologies in molecular medicine collects massive data for hundreds to thousands of persons in large cohort studies by characterizing the phenotype of each individual on a personalized basis. It finally aims at increasing our understanding of disease genesis and progression and to improve diagnosis and treatment. New methods are needed to handle such 'big data'. Machine learning enables to recognize and to visualize complex data patterns and to make decisions potentially relevant for diagnosis and treatment. We addressed these tasks by applying the method of self organizing maps and present worked examples from different disease entities of the colon ranging from inflammation to cancer.

Introduction

Application of new high-throughput technologies in molecular medicine such as microarrays and next generation sequencing generates massive amounts of data for each individual patient studied. These methods enable to characterize the genotype and/or molecular phenotype on a personalized basis with the aim to increase our understanding of disease genesis and progression and, in final consequence, to improve diagnosis and treatment options. New methods are needed to handle such 'big data' sets collected for hundreds to thousands of persons in large epidemiological cohort studies, e.g. to accomplish data mining and classification tasks with impact for diagnosis and therapy. From the perspective of bioinformatics and systems biomedicine, 'big data' challenge objectives such as data integration, dimension reduction, data compression and visual perception. To finally achieve a personalized therapy it is necessary to link genetic variations to molecular disease phenotypes, to associate molecular with clinical data, to extract, to filter and to interpret bio-medical information and finally, to translate these discoveries into medical practice.

Machine learning represents one interesting option to tackle these tasks. Particularly, neural network algorithms such as self-organizing maps (SOMs) combine effective data processing and dimension reduction with strong visualization capabilities. These methods provide a suited basis to analyze large and complex data generated by modern bioanalytics.

The present contribution shortly describes the method of 'SOM portraying'. We demonstrate data compression capabilities which reduce the dimension of the relevant (in terms of functional information) data by several orders of magnitude. The strong visualization capabilities of the SOM approach are illustrated. They enable the comprehensive, intuitive and detailed analysis of 'big data' in molecular medicine by mapping them into phenotype and feature space. To illustrate the performance of the method we present a series of representative case studies from different disease entities and OMICs realms related to the human colon.

Background

Big data from high-throughput bioanalytics

Standard medical practice is moving from relatively ad-hoc and subjective decision making to so-called evidence-based healthcare which makes use of complex diagnosis technologies such as comprehensive laboratory analyses and powerful imaging techniques.

Powered by the progress in modern molecular biomedicine the number and granularity of accepted disease types and also the variety of related therapy options steeply increase. This trend is paralleled by increasing volume and complexity of data collected per patient in disease-related cohort studies and also in medical practice. Accordingly, the way of decision making in diagnosis changes, e.g. from evaluating a set of key laboratory markers to information mining in large and potentially 'big' data sets generated by high-throughput technologies. Moreover, the evaluation of currently collected data includes also their comparison with already accumulated knowledge and reference data which itself can constitute a 'big' data challenge.

As generally accepted, big data is characterized by the three (Beyer, 2011), and sometimes four 'V' : big volume, big velocity, big variety and, also, big veracity referring usually to the scale of the data, the handling of streaming data, the manifold and complexity of different forms and values of data and to their uncertainty, respectively. For high-throughput data in molecular medicine these general criteria can be specified: Usually the number of single data items per sample measurement ranges from tens of thousands to several millions and even more depending on the type of data (e.g. proteomics measured by means of mass spectrometry or genomics measured by means of next generation sequencing) and on their level in the processing pipeline starting with raw data and ending with highly (information-) enriched data (see below). In this respect present- and next generation omics-technologies generate massive amounts of data. Velocity in terms of time needed to store and re-store the data and to process them in downstream analysis programs is an important point which however will not be addressed here. Variety is probably the most important aspect in omics-bioinformatics because the assignment of data to the patients on one hand and to relevant biological items such as genes on the other hand, and their covariance structure basically code the useful information which governs biological function. Biostatistics mainly addresses the veracity of biomedical molecular data with the main aim to optimize marker selection tasks by maximizing their significance in terms of sensitivity and specificity by taking into account the uncertainty inherent in the data. Recently more 'V' are added to be essential for big data such as 'variability' (variance in meaning), 'visualization', 'value', 'volatility' and 'validity' (Normandeu, 2013; van Rijmenam, 2013). 'Visualization' is a very important aspect because it makes big data comprehensible in a manner that is easy to understand and read. It is however a difficult but also extremely crucial challenge especially in personalized medicine because it can help medical doctors to evaluate big data based on visual perception without explicitly dealing with numbers.

'One of the biggest new ideas in computing is "big data".' ("The Big Data Conundrum: How to Define It?," 2013). Unfortunately the term 'big' invites quantification and thus overemphasis of the first 'V', synonymous for the volume (see ref. (Ward & Barker, 2013) for a detailed discussion). This makes a definition difficult and susceptible to misinterpretation. Recall that, despite the sudden interest in big data, the concept of the four 'V' is not really new and applies, at least partly, also to conventional data processing. 'Big' implies significance, complexity and challenge. In this sense 'big' can be understood as the challenge to data which are difficult to process using common management tools and/or data processing methods. Hence, 'Big Data can be actually very small and not all large data sets are really big' (Rindler, McLowry, & Hillard, 2013). For biomedical high-throughput data complexity and thus the variety 'V' and not size (usually annotated as 'massive' data) is often the most critical factor if one aims at finding and unlocking interesting patterns and associations to power the advance of stratified medicine.

Interestingly, other definitions link the term ‘big’ data with the technologies machine learning and artificial intelligence requiring significant compute power and focusing on the extraction of information from the data (Microsoft, 2013). However, machine learning and particularly, neural network algorithms such as self-organizing maps (SOMs) are relatively infrequently used in high-throughput bioanalytics possibly because involved bioinformaticians and statisticians are typically trained in ‘classical’ approaches for feature selection, class discovery and classification based on rigorous significance testing of single features. On the other hand, researchers with background in machine learning are often affiliated at engineering and technology departments and not or only peripherally involved into life science problems. Therefore machine learning and particularly SOM are still innovative methods in molecular biology and health science. In consequence application of the concept of SOM learning data transformation and visualization still requires explanation and adaptation. Moreover, the SOM algorithm accomplishes ‘only’ basal sorting and visualization tasks. It needs to be supplemented with add-ons for significance testing and marker extraction, visualization of biological properties inherent in the data and finally for information mining of the biological context to become an attractive application tool in life sciences.

Reducing the dimensionality: Subtyping, filtering and re-weighting

Typical big data in molecular medicine comprises thousands to millions of ‘single’ features related to molecular items which are measured separately in dozens to thousands of patient-related samples. This data can describe the genotype of each individual if it contains heritable genomic information such as mutations or aberrations of the DNA. Other molecular technologies collect data about the molecular phenotype in terms of the abundance of proteins, messenger RNA (mRNA), micro RNA or metabolites. These molecular markers characterize structural and functional building blocks of the organism resulting from the transcription of heritable genomic information under the influence of environmental factors. In consequence the molecular phenotype of each individual is unique.

Formally each individual of the collective under study is represented by its own specific position in the N-dimensional space spanned by the entirety of these phenotypic features (see Figure 1a for illustration). Groups of individuals with similar phenotypes can be stratified into subtypes representing generalized phenotypes. They are characterized by subtype-related ‘mean’ feature values and also by the diversity of values of all individual members of the subtype which can be quantified using probabilistic measures such as frequency distributions and variance measures. The subtyping of large sample collectives into strata is a reasonable strategy in molecular diagnosis because it displays disease subtypes of different molecular origin. Appropriate subtyping is required to select features suited as classification markers with potential impact for diagnosis and therapy.

Usually the highly dimensional feature space is partly redundant because part of the features are covariant or simply uninformative with respect to the phenotypes, e.g., if they lack significant variability. Hence, handling of big data in molecular medicine requires first of all the reduction of dimensionality by removing or down-weighting redundant or uninformative data. This dimension reduction can be applied to the samples by clustering them into a reduced number of groups and using these generalized phenotypes in further analyses. Dimension reduction can be also applied to feature space either by removing uninformative features or by re-weighting the data via clustering.

A simple method for dimension reduction is filtering: It identifies uninformative features and removes them from data space. Note however, that filtering can be dangerous because it might also eliminate valuable information, for example, by removing noisy features which nevertheless carry important biological information. Hence, filtering is an optimization task with the requirement of removing virtually irrelevant data while preserving as much as possible information in the remaining part of the data. We have discussed this issue in terms of the antagonism between ‘representativeness’ and ‘noisiness’ of the filtered data where optimization aims at maximizing representativeness while minimizing noisiness (Wirth, Loeffler, von Bergen, & Binder, 2011).

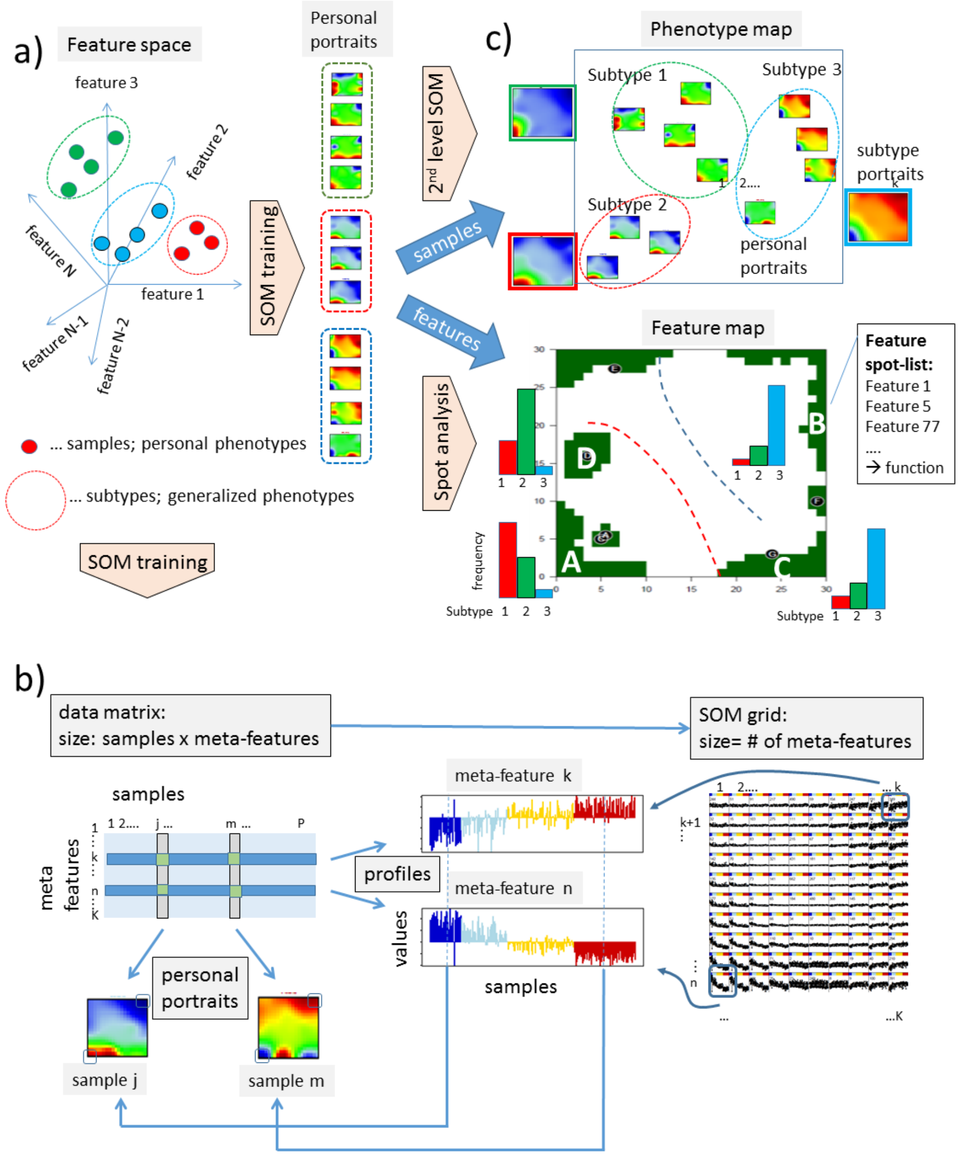


Figure 1: a) Patient-related ‘big’ phenotypic data in molecular medicine are distributed in N-dimensional feature space where each sample occupies one position (colored circles). Similar ‘personal phenotypes’ can be clustered into subtypes (dashed ellipses). b) SOM machine learning clusters the features into meta-features (see also Figure 3 below). They are transferred from a matrix-representation into a quadratic grid for ‘portraying’. The values of the meta-features are visualized as ‘profiles’ (values of one selected meta-feature in all samples) or as ‘personal portrait’ (values of all meta-features in one selected sample) representing a colored image of the meta-feature landscape of a selected sample. c) The multitude of personal portraits included in the study are further analyzed to extract information about sample- and feature-diversity. The phenotype map allows to classify ‘personal’ phenotypes into subtypes. The feature map clusters the features into spots modules (ellipses) where each of them

comprises a list of single features. They can be associated with the phenotypes as illustrated by the bar plots quantifying the fraction of samples showing high feature values in the respective spot in their portrait. Spot-related profiles from opposite regions of the map are usually anticorrelated whereas adjacent spots are often correlated.

Re-weighting is another option to reduce the dimensionality of the data. It can be achieved by clustering the features into appropriate groups where each of them is characterized by a representative meta-feature serving as prototype of the cluster. Further analyses can then be performed based on these prototypic data of reduced size. Each cluster can contain different numbers of single features represented by the respective prototype. Analyses based on prototypic data thus effectively alter the importance of the original data: Single features become effectively down-weighted if the clusters are large. Contrarily, they become effectively up-weighted if the clusters contain only a few single features. Hence, analysis on the level of prototypic data increases representativeness because highly redundant data are down-weighted while rare but important features become up-weighted (Wirth, et al., 2011). In general, reweighting is advantageous compared with filtering because no single feature is removed from the analysis. Instead the single data remain ‘hidden’ behind the prototypic features. Their values can be ‘re-accessed’ and used for analyses if necessary.

Analyzing and visualizing big omics data using Self Organizing Maps

SOM portraying

A large number of clustering methods is available with different advantages and disadvantages depending mostly on the data type, the intrinsic structure of the data, their size but also on the performance and power of the respective algorithms. In this contribution we make use of the method of self organizing maps (SOM), a machine learning technology offering several advantages compared with alternative methods such as non-negative matrix factorization, K-means, hierarchical clustering or correlation clustering when applied to ‘big’ data in molecular medicine (Wirth, et al., 2011).

SOM is a supervised clustering method, i.e. it distributes the features under study over a predefined number of clusters called meta-features where features with similar values in all samples are clustered together. The value of each meta-feature serves as representative (prototypic value) for the respective cluster of single features. The meta-features are arranged in a quadratic grid called SOM space with the objective of visualizing their values separately for each sample (see Figure 1b). This ‘data landscape’ is obtained by coloring each pixel in the grid according to the value of the respective meta-feature using a suited color code. It is an important property of the SOM clustering that it ‘self-organizes’ meta-features with similar profiles together into neighboring pixels within the SOM space. Thus, after training of the data (see below) one obtains an individual mosaic image for each sample. Virtually it portrays the multidimensional data landscape in terms of a colored, blurry texture serving as molecular fingerprint of the respective patient-related sample. These images are therefore called molecular portraits. As a typical pattern they reveal often uniformly colored spot-like areas which represent clusters of co-variant meta-features with relatively high or low values in the respective sample. Further analysis of the spot patterns observed in the molecular portraits enables to identify such ‘spot-‘ cluster of co-variant features. They can be interpreted as intrinsic modes of variability inherent in the data set. Importantly, clustering into spot modules is an unsupervised approach because their number is not defined by the user.

Hence, SOM portraying generates a pixelated mosaic image of the individual feature landscape of each sample where intrinsic modes of co-regulated features appear as colored spots. As a rule of thumb, the number of pixels (i.e. of meta-features) must exceed the number of intrinsic modes by roughly a factor of 20 – 100 to achieve a sufficient resolution which allows to detect all relevant modes. As a simple analogy one can compare SOM portraying with portraying objects on a television screen: TV also displays shapes as pixelated color image. For proper perception, the number of pixels used must largely exceed the number of relevant details.

In summary, SOM portraying combines a two-step compression of the original data with the intuitive imaging of its intrinsic structure. Both aspects will be discussed below.

SOM training

SOM uses an iterative learning algorithm to cluster the data as described. It starts with appropriate initialization of the map space, followed by the training process to adjust the map space to the multivariate covariance structure of the input data, and it ends with the final mapping and visualization of the map space in terms of SOM portraits. The SOM training algorithm iteratively fits the meta-profiles to the profiles of the single features and, in parallel, assigns each single feature to the meta-feature of maximum similarity. Each iteration divides into three sub-steps: Firstly, one feature profile is selected as training vector. Secondly, the meta- profile of closest similarity (also called BMU – best matching unit) is determined using the Euclidean distance as criterion. Thirdly, the profile of the BMU and the profiles of its neighbors are adjusted to better resemble the selected training vector. The amount of adaption is downscaled with increasing distance to the BMU in the two-dimensional grid. With progressive number of iterations the algorithm settles down: the meta-profiles progressively cover the multitude of different profiles of input features, which, in turn, distribute among the meta-features available. SOM training is stopped after a few hundred thousand iterations ensuring convergence.

In consequence, the data space becomes segmented into clusters of single features mapped to each meta-feature after training. Each cluster is characterized by one meta-profile which is used for visualizing the feature's state at each condition studied (see above). For a detailed description of the SOM method we refer to the additional reading section and to (Wirth, von Bergen, & Binder, 2012).

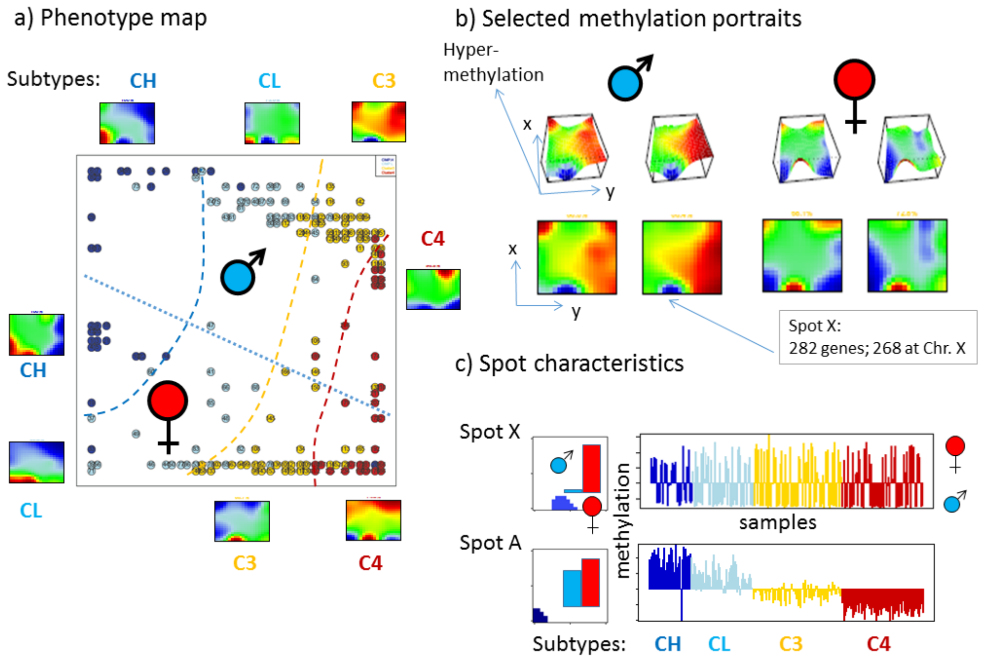
Sampling and visualizing the feature and phenotype space

Our implementation of the SOM method enables to visually portray each sample in terms of a colored two-dimensional image (Wirth, et al., 2011; Wirth, et al., 2012). The portraits of a larger number of samples are analyzed in two principal ways (see Figure 1c):

Firstly, sample similarity analysis describes the feature space covered by the samples. We applied so-called second level SOM analysis to map the distribution of samples from the multidimensional feature space into two dimensions. The obtained phenotype map visualizes the distribution of personal portraits and allows to identify subtypes as clusters of samples with similar feature landscapes. The feature landscape of the generalized phenotypes can then be extracted as a mean image averaged over the personal portraits of each subtype.

Secondly, feature similarity analysis explicitly discovers the diversity of feature values in phenotype space. Features behaving similarly among the samples are clustered together into so-called spot modules in SOM analysis. They are called spots because these clusters appear as red or blue spot-like regions in the individual SOM portraits if the respective feature values are high or low, respectively. Each spot is represented by its prototypic meta-feature profile. Spot diversity analysis then provides statistical measures about the abundance of each of the spots in the different subtypes such as the respective frequency distribution. Moreover, the feature map summarizes the spot landscapes of the personal portraits into a master map which assigns the spots observed to the different subtypes. The feature map complements the phenotype map because it links sample diversity with feature diversity. Importantly, the different spot modules can be interpreted in terms of biological function using previous knowledge about the associated single features under reference conditions such as healthy and well defined diseased states (Wirth, et al., 2012).

Figure 2 shows an example illustrating phenotype and feature mapping: The underlying data are gene-related DNA-methylation data of colon cancer samples taken from 320 individuals collected in the TCGA (The Cancer Genome Atlas) project using the microarray technology (TCGA, 2012). These data estimate the methylation level of the promoter region of about 20,000 human genes in each individual sample. The phenotype map in Figure 2a reveals that the samples split clearly into the two gender-specific phenotypes ‘men’ and ‘women’ in vertical direction. In addition, four gender-independent methylator subtypes were identified (TCGA, 2012) which distribute mostly along the horizontal coordinate of the phenotype map. Visual inspection of the individual portraits indicates that the methylation landscapes decompose into two mutually independent subpatterns which are governed either by gender- or by methylator-phenotype-related traits. The former subpattern is dominated by spot ‘X’ assigned in part b of Figure 2. It contains 282 single genes in total where 268, i.e. virtually all of them are located on the sex-determining chromosome X (allosome). The respective genes are systematically hypermethylated in women (spot X becomes red) and hypomethylated in men (spot X becomes blue). The methylation profile of this spot is shown in panel c where the samples are ordered according to their methylator-phenotypes showing an almost perfect sepearation of samples in a gender-specific manner. The scattered distribution of high and low methylation values reflects the fact that men and women nearly equally distribute over the methylator phenotypes. In contrast, the methylation profile of another spot A strongly associates with the methylator subtypes. It contains genes associating with the methylator-phenotype.


a) Phenotype map

b) Selected methylation portraits

c) Spot characteristics

Spot X:
282 genes; 268 at Chr. X

Figure 2: SOM analysis of gene-related DNA-methylation data of a cohort of colon cancer samples taken from ref. (TCGA, 2012). The data contain two mutually independent kinds of information about gender and cancer subtypes, respectively: (a) The phenotype map separates the samples in both a gender- and a cancer-specific way. It reveals four subtypes taken from the original publication (CH, CL, C3, C4; see below) which split into two gender-specific areas. (b) The methylation landscapes of selected samples show a gender-specific spot. It contains genes which are hypo-methylated in men and hyper-methylated in women. It is visible in the two-dimensional sample portraits as spot X colored in blue (men) or red (women). (c) The methylation profile of spot X strongly depends on the gender of the tumor patients but not on its

subtype. In contrast, the methylation profile of spot A is subtype-specific but it virtually doesn't associate with the gender of the patients.

Gender-specific hyper-methylation of genes on allosome X is well known as X chromosome inactivation (XCI) (Augui, Nora, & Heard, 2011). XCI is a dosage compensation effect leveling the expression of chromosome X genes in female cells. Recall that female cells contain the double set of X-chromosomes (XX) compared with male cells (XY). XCI silences part of the genes on the X chromosome in women by hyper-methylation of their promoter regions to prevent overexpression compared with men. Gender-specific methylation thus considerably expands the phenotype space compared with that occupied by the cancer methylator subtypes solely. Since cancer subtyping studies are primarily interested in subtype-related characteristics it is desirable to confine phenotype space to that of the methylator subtypes. This can be simply achieved by removing the genes on chromosome X from the analysis. The obtained shrunken phenotype space is presented below in the examples section of this chapter.

Multi-step information enrichment

The original set of high-dimensional data is presented in form of a matrix of dimension *samples x features*. Our SOM approach compresses the data in several consecutive steps which apply separately to the feature and sample dimensions (Figure 3): (i) SOM training collects similar profiles of single features into typically a few thousand micro clusters called meta-features, which reduces the number of features typically by one-to-two orders of magnitude compared with the original data. The meta-feature landscape of each sample is transformed into one mosaic image where each pixel is assigned to one meta-feature as described above. Further downstream analysis is based on these personal portraits and uses the underlying meta-feature data. (ii) The textures of the obtained SOM-portraits are decomposed into a few (typically about one dozen) spots representing clusters of concerted meta-features. The spot profiles can be understood as a sort of 'eigen-modes' characterizing the multitude of basal feature patterns inherent in the data. In other words, the spot-modules represent a natural choice of context-dependent patterns in complex data sets. Note that spot-clustering further compresses the data and reduces their volume by another one-to-two orders of magnitude compared with the preceding meta-feature level. (iii) Compression so far was applied to the feature dimension of the data. The last 'subtyping' step applies to the sample dimension using appropriate methods of class-discovery. The number of relevant classes is typically much smaller than the number of samples studied giving rise to further reduction of data size by at minimum one order of magnitude.

Taking together, this data compression pipeline reduces the volume of features used by up to six orders of magnitude. The main criterion for applying this pipeline requires that the relevant information content of the original data remains stored in the compressed data. Of course, the 'relevance' of information is related to the particular objective of the study. In the context discussed here, we focus on the main intrinsic structure of the data which decomposes into subtypes of similar samples on one hand and into related sets of features called spot-modules on the other hand.

Importantly, the dimension reduction of the data does not entail the loss of primary information in contrast to simple filtering which irretrievably removes part of the data as discussed above. Instead, the reduction of dimension is attained by re-weighting of primary information in the aggregation step. The whole set of single feature values remains virtually 'hidden' behind the meta-features. This primary information together with the respective annotations of the features can be extracted in later steps of analysis to interpret the observed SOM textures using concepts of biological function.

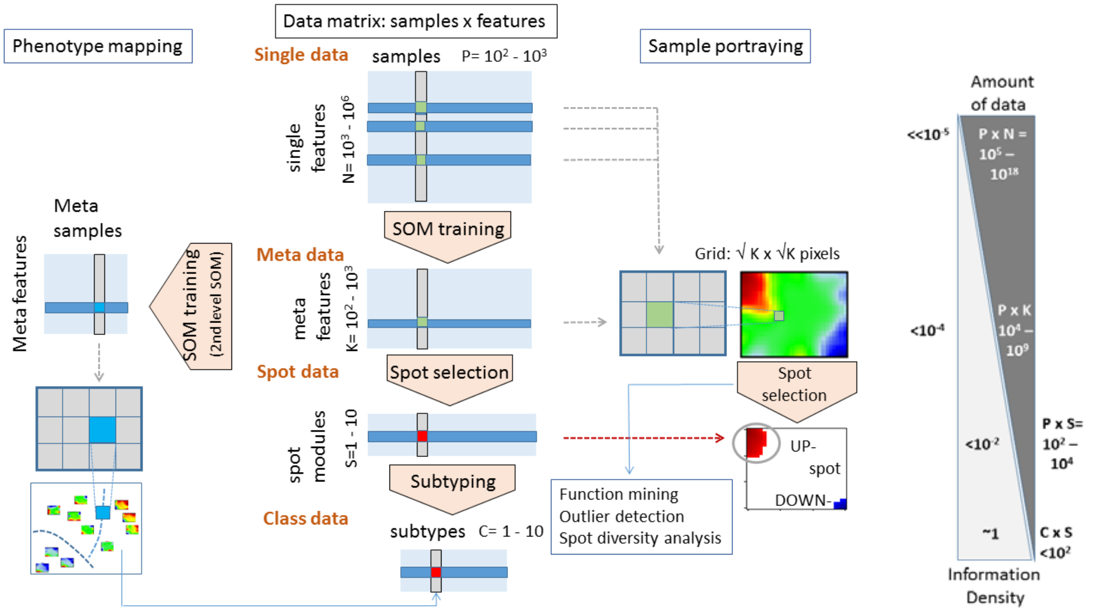
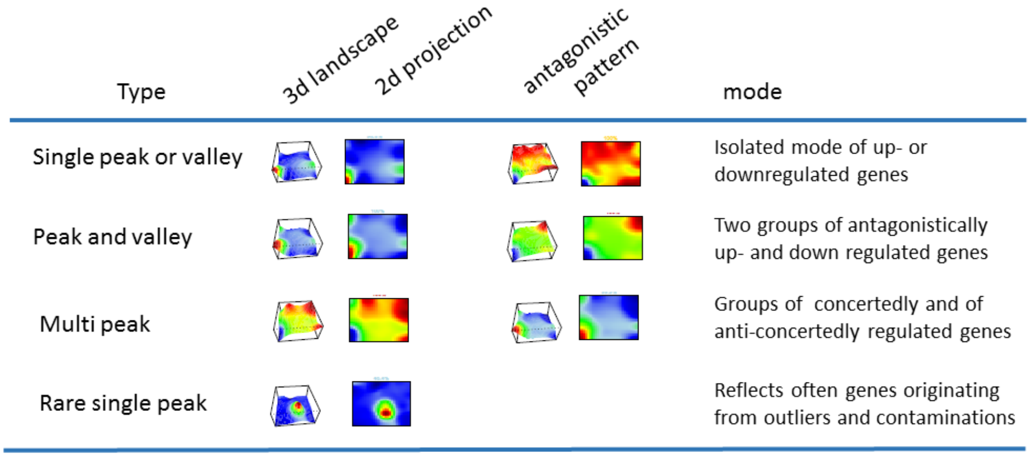


Figure 3: SOM training of big data in molecular medicine is performed on two levels: (1) Feature space of the original items is compressed using 1st level SOM. It provides one mosaic image per sample (so-called sample or personal portrait). It allows to extract clusters of features as spot modules. (2) Sample space is 'scanned' using 2nd level SOM. It maps the sample portraits and allows to classify them into subtypes. Data compression using SOM clustering increases the information density of the data by several orders of magnitude.

Second-level SOM analysis is applied for phenotype mapping as proposed by Guo et al. (Guo, Eichler, Feng, Ingber, & Huang, 2006) to visualize the similarity relations between the individual portraits. Second-level SOM analysis uses the personal portraits of all samples as input. It then clusters the samples and not the features as in first-level SOM analysis. Each tile of the second-level SOM mosaic characterizes the feature landscape of a representative meta-sample. The 2nd level SOM phenotype maps can be used for class discovery in the subtyping step.

Personalized feature landscapes

Each sample's feature landscape is described by the values of the meta-features in the respective column of the metadata matrix (see Figure 3). These values are arranged in the grid of the SOM map and visualized using an appropriate color gradient: Dark red usually reflects high meta-feature values; yellow and green tones indicate intermediate levels; and blue corresponds to low values. The resulting color patterns emerge as smooth textures owing to the self-organizing properties of the training algorithm as described above. The obtained mosaic image visualizes the multidimensional feature landscape of the respective specimen. It thus provides a molecular portrait of each patient-related sample. Please note also that the assignment of the features to meta-features and therefore also their position in the map is identical in all sample portraits trained together. Hence, the coloring at a certain position in the map refers to the same features in all individual portraits. The invariant position of each feature in all portraits allows the direct comparison of its value between the maps.



Type	3d landscape	2d projection	antagonistic pattern		mode
Single peak or valley					Isolated mode of up- or downregulated genes
Peak and valley					Two groups of antagonistically up- and down regulated genes
Multi peak					Groups of concertedly and of anti-concertedly regulated genes
Rare single peak					Reflects often genes originating from outliers and contaminations

Figure 4: Different types of feature landscapes reflect characteristic regulatory modes. Most of the landscapes are observed as mirror symmetric 'positive-negative' twins reflecting antagonistic modes.

Hence, visual inspection and intuitive interpretation play a central role in the initial steps of our portraying method. Figure 4 shows a few examples of characteristic types of feature landscapes observed frequently: The 'single peak'-type is characterized by a single peak protruding in an otherwise flat plain-like landscape. The peak contains single features with high values in the respective samples. The 'single-valley'-type represents the antagonistic landscape with a set of features with low values in the respective samples evident as a hole-like valley in an otherwise flat landscape. The 'peak-and-valley'-type shows a peak *and* a valley at the same time, often in opposite corners of the map. This type reflects two antagonistically-regulated sets of features, e.g. if one is high then the other one is low and vice versa. The 'multi-peak' and/or 'multi-valley' types reflect more complex landscapes which are characterized by the superposition of several modes forming a certain correlation network between the features. The 'rare single peak'-type shows essentially the same landscape as the 'single peak'-type discussed above. The peak however is observed in a relatively small fraction of samples and it is often located in the central part of the map. Moreover, valleys are not observed at the respective positions in any of the samples. Such patterns can be indicative for contaminations with another feature landscape which simply overlays with the first one. We found such patterns e.g. in expression studies of cancer samples which are contaminated with healthy tissue (Hopp, Lembcke, Binder, & Wirth, 2013; Hopp, Wirth, Fasold, & Binder, 2013). We have shown that the landscapes allow identification and interpretation of outlier samples and thus to improve data quality. In summary, intuitive visualization of data landscapes using SOM portraits clearly promotes quality control and the discovery of the intrinsic covariance structure of the data.

Note that this image-based perception and decision making of multidimensional high-throughput data has something similar with conventional methods in pathology which evaluate and classify microscopic images of tissue sections using a set of key features taken from previous knowledge. Also these images represent surrogates of the underlying complex molecular patterns allowing precise diagnosis in many cases.

Case study: Gene expression and -methylation phenotypes of the healthy and diseased human colon

For illustration of the SOM portraying method we selected a series of examples addressing the transcriptome and methylome of colon mucosa in the healthy and diseased states. We demonstrate that different diseases and molecular parameters provide data of varying complexity. We also show that the information content of these large data sets can be easily extracted by means of SOM analysis. For direct comparison we use the same presentations of all examples.

Ulcerative colitis before and after onset of inflammation

Specimen were sampled by endoscopic mucosal biopsies from ulcerative colitis patients with and without microscopic signs of inflammation to investigate the effect of inflammation on gene expression (see (Olsen et al., 2009) for details). We transformed the expression data of all samples into personal portraits grouped into the not-inflamed (Ni) and inflamed subtypes (If) (Figure 5a). The different textures of both subtypes clearly provide first indications of different gene activation patterns. Note that the whole data set comprises the expression levels of 22,000 genes measured in 24 samples, i.e. in total about half a million single items. The mean portraits were obtained by averaging all personal portraits of each subtype (Figure 5b). They reveal relatively simple landscapes of the ‘peak and valley’ type (see Figure 4 above) where two groups of genes are antagonistically up- or down-regulated in either of the subtypes. The minimum content of relevant information stored in the data is approximately one bit only.

More detailed inspection of the personal portraits and especially of that of the not-inflamed subtype reveals however a certain diversity of spot patterns. It is averaged out in the mean portraits (Figure 5b). Figure 5c estimates the diversity of spot patterns in each of the subtypes: It shows frequency histograms of the number of red overexpression spots observed in the personal portraits. Whereas the images of the inflamed cases mostly contain only one spot, lack of inflammation gives rise to at minimum one more spot in many of the personal portraits. However spots can appear at different positions. The spot/feature map in Figure 5d summarizes all relevant spots observed in any sample into one master map to provide an overview over all relevant expression modes. In total we identified seven spots denoted with capital letters A – G. The spot association histogram in Figure 5f depicts the faction of personal portraits of each subtype which show the respective spot. It consequently visualizes spot abundances. One sees that spot G is specific for the inflamed state: It is found in almost all inflamed samples. This result is in correspondence with the mean subtype portrait essentially showing this overexpression spot solely. In contrast, the remaining six spots are predominantly found in samples of the not-inflamed subtype however with different frequencies: The rarest one, spot F, is found only in one sample whereas the most frequent one D is found in 80% of the not-inflamed samples.

The information about spot-abundance allows to divide the feature map into areas which contain spots overexpressed in a certain subtype: The area of spots up-regulated in the inflamed state essentially restricts to the left upper corner of the map whereas the area of spots up-regulated in the not-inflamed state occupy a much larger area. This relation reflects the high heterogeneity of not-inflamed states compared with the relatively homogeneous landscapes of inflamed states. Interestingly, the phenotype map in Figure 5e transports essentially the same information: The inflamed samples occupy a relative small area in the map near the left-lower corner whereas the not-inflamed samples spread over a much larger area reflecting the larger diversity of their expression landscapes. Despite this heterogeneity, both inflamed and not-inflamed subtypes remain still separated without overlap of the respective areas. The variance profiles of the expression landscapes are given as boxplots in Figure 5g: Expression values are more variable in the not-inflamed samples. The dumbbell-like lines in the feature map Figure 5d connect anti-correlated spots: The strongest effect refers to spot-pairs identified as ‘peak and valley’-like landscapes (spots G and D). Another pair of spots (B and C) however reveals a weaker anti-concerted mode appearing in the not-inflamed samples.

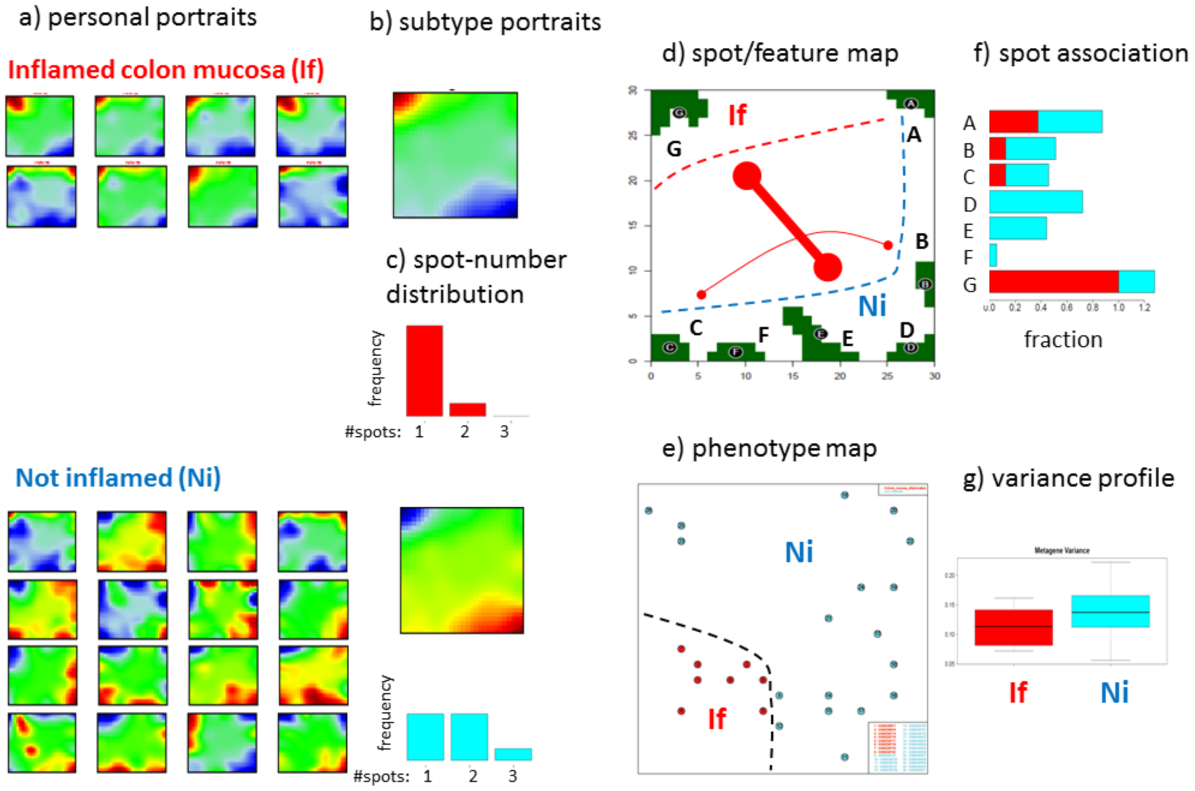


Figure 5: SOM portraying of gene expression landscapes of samples of inflamed and not-inflamed colon mucosa (data and sample assignments were taken from (Olsen, et al., 2009)). See text.

Finally, researchers are mostly interested in the biological meaning of the observed landscapes. Using gene set enrichment analysis we assign characteristic biological functions to the genes collected in the spot clusters (see ref. (Wirth, et al., 2012) for details). In the inflamed state genes related to processes like ‘cell adhesion’, ‘chemokine activity’ and ‘blood coagulation’ (spot G) become-up regulated whereas ‘metabolism’ and ‘mitochondrion’ (spot D) become down-regulated. In addition to these mostly expected results one finds that the diversity of not-inflamed states can be attributed to varying activities of processes related to ‘transcription’ (E), ‘RNA-processing’ (C), and ‘cytosol’ (B). The antagonism between spots C and B can be also attributed to the activity of the BCR-gene.

Colorectal adenoma

The second example compares the expression landscapes of healthy colon mucosa (He) with that of colorectal adenoma (Ad), a precancerous lesion of colon mucosa. The original data set was taken from (Sabates-Bellver et al., 2007) and comprises 32 samples of each class and 54,000 transcripts measured per sample. Figure 6 shows the characteristics for this data analogous to Figure 5. Both classes of samples were very similar with respect to the heterogeneity of their expression landscapes as estimated by the spot number distribution and variability of their expression landscapes. Accordingly both feature and phenotype maps divide into two nearly equal regions where each of them can be assigned to one of the subtypes. Importantly, these regions assigned to either healthy colon mucosa or adenoma are well separated without overlap. The mean subtype portraits are again of the 'peak-and-valley' type reflecting that the basal information content of this data is again one bit. Adenoma are characterized by up-regulation of processes like 'DNA-repair' and 'cell cycle' (spot E) and by down-regulation of 'T-cell activation' (A) and 'extracellular space' (B).

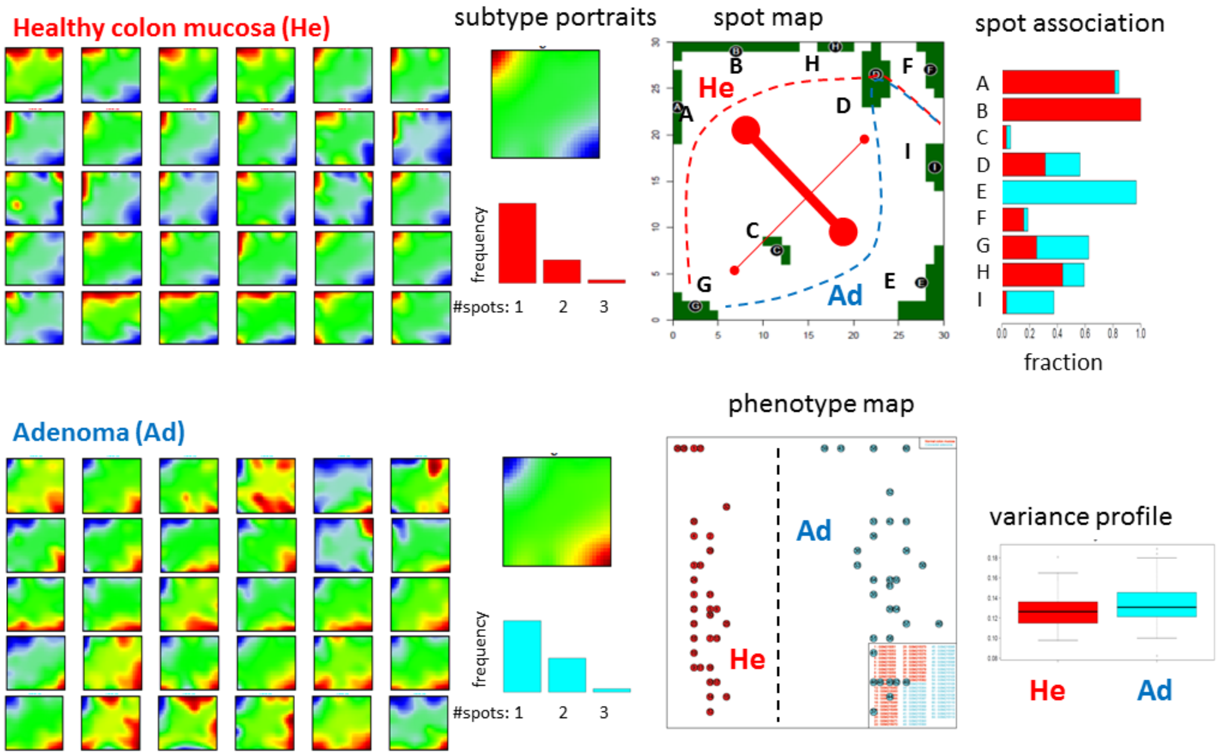


Figure 6: SOM portraying of gene expression landscapes of healthy colon mucosa and of colorectal adenoma (data and sample assignments were taken from (Sabates-Bellver, et al., 2007)). Only 30 samples portraits per subtype are shown. See text.

Methylator phenotypes of colorectal cancer

Colorectal cancer (CRC) as most of other cancer entities has predominantly been considered a genetic disease. It is characterized by sequential accumulation of genetic alterations such as mutations and chromosomal instabilities leading to selective and progressive dysfunctions of genomic regulation. Epigenetic alterations however add an additional layer of complexity to the pathogenesis of CRC, and characterize a subgroup of CRC with a distinct etiology and prognosis (see, e.g. (van Engeland, Derks, Smits, Meijer, & Herman, 2011)). The most extensively characterized epigenetic alteration in CRC is hypermethylation, which occurs at CpG dinucleotide dense regions, called CpG islands, present at the 5' region of approximately 60% of the genes. Most CpG islands lack methylation in normal colon mucosa, independent of the transcriptional status of the gene. Hypermethylation of promoter CpG islands however has been observed for numerous tumor suppressor- and DNA repair-genes with consequences for gene activity equivalent to mutations.

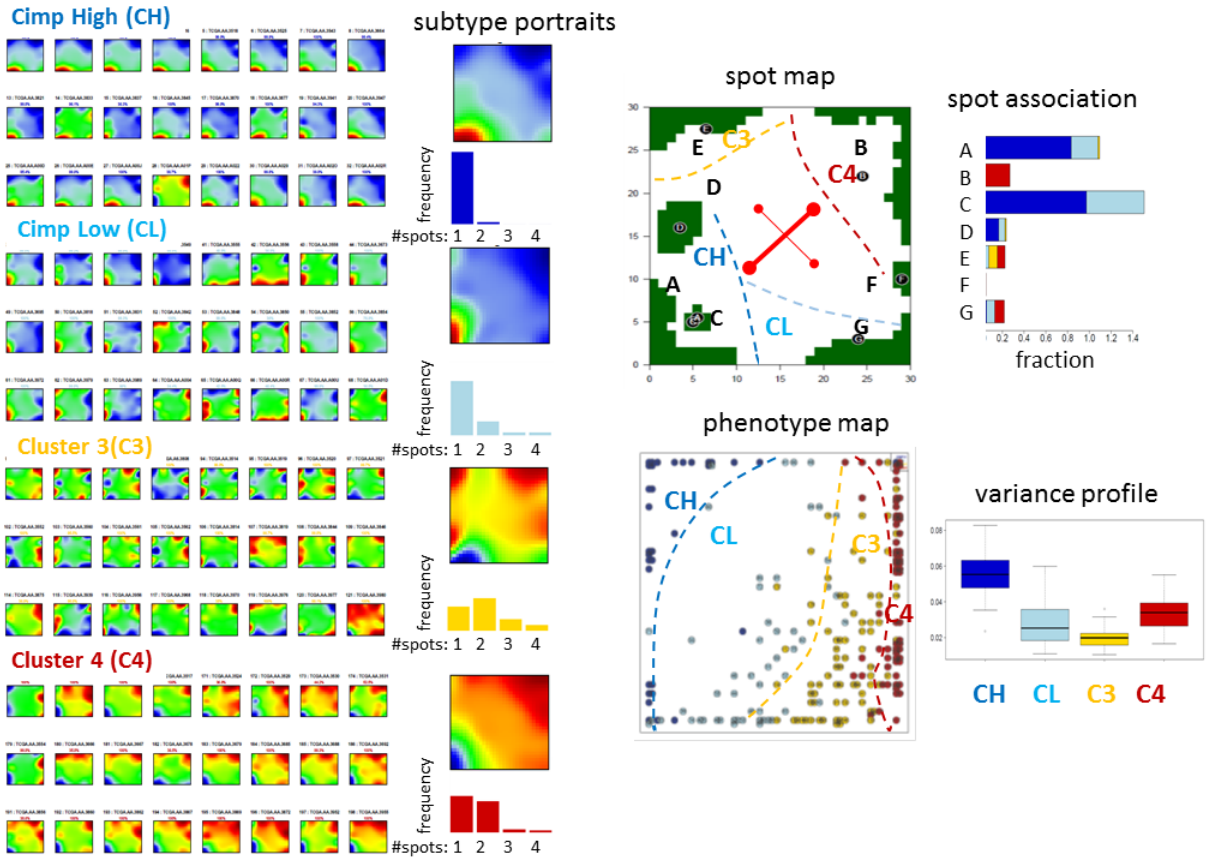


Figure 7: SOM portraying of gene promoter methylation data of colorectal cancer (data and assignment of the samples to the four methylator phenotypes were taken from (TCGA, 2012)).

The recognition that a distinct subset of colorectal adenomas and CRCs display significantly more promoter methylation than others has led to the introduction of the concept of CpG island methylator phenotype (CIMP) (Toyota et al., 1999). The existence of four different CIMP subtypes, each with specific molecular characteristics, has recently been postulated based on high-throughput methylation data (TCGA, 2012). This data is presented above using methylation values of genes taken from all chromosomes including chromosome X (Figure 2). The same data set was again analyzed considering however only genes located on autosomes (Figure 7). The phenotype map obtained clearly distinguishes the four different methylator subtypes without the gender-specific split seen in Figure 2. The data set comprises 235 samples where only randomly selected examples from each subtype are shown in Figure 7 as portraits. The mean portraits of two subtypes (CIMP_high (CH) and to a less degree CIMP_low (CL)) are of the 'single-peak' types. The red peak-spot contains about 1,000 single genes strongly hyper-methylated in these CIMP subtypes. The personal CIMP-portraits are relatively homogeneous showing mostly only this single peak. The landscapes of the other two subtypes, 'cluster 3' (C3) and 'cluster 4' (C4), are characterized by a broader spot number distribution reflecting a more heterogeneous diversity of methylation landscapes. The mean subtype portraits of C3 and partly of C4 can be attributed to the 'multi-peak'-type revealing a more complex substructure of three to four groups of differentially methylated genes. The respective peak landscapes of the C3 and partly of the C4 types are however more smooth compared with the steeper landscapes of the CIMP high subtypes (see the variance profile in Figure 7). In consequence a smaller number of peaks is detected in C3 and C4 (see spot associations in Figure 7). Note however that the basal information content of the data is multivariate and splits into two-to four different modes.

The spot hyper-methylated in the CIMP subtypes (spot A) contains a high number of genes related to 'nervous systems function and development' whereas the spot hyper-methylated in C4 (spot B) enriches genes related to 'DNA transcription' and 'cell division'.

The expression landscape of colorectal cancer is governed by its methylator phenotype

Expression of hyper-methylated DNA is silenced via recruitment and binding of methyl-CpG binding proteins and associated co-repressors such as HDACs, which create a repressive chromatin structure. Hence, methylation is expected to modulate the gene expression landscapes of CRC. Expression data of the CRC-cases analyzed in the previous subsection are available and analyzed using our SOM portraying pipeline (Figure 8). Interestingly, the expression portraits of the subtypes clearly reveal an anti-matching characteristics when compared with the respective methylation landscapes. For example, the strong red hyper-methylation peak in the methylation landscapes of the CIMP subtypes is paralleled by a sharp expression minimum in their expression landscapes whereas the hypo-methylation minimum of the C3 and C4 subtypes is paralleled by an expression peak in the expression landscapes (compare with Figure 7). This result reflects the basal inhibitory effect of methylation on expression, namely that increased promoter methylation reduces the expression level of the respective gene and vice versa. Note that the position of the maxima and minima in both types of maps disagrees because both SOM are trained independently giving rise to different distributions of the genes in the maps.

Note also that the phenotype and feature maps of the expression data clearly reflect the grouping of samples into methylator subtypes. In other words, the expression landscapes are clearly affected by the underlying methylation landscapes which govern activity of many genes. The variance profile of expression landscapes reflect this gene-dosis effect: large methylation changes in the CIMP_H subtype are associated with large changes of the expression level. The gene expression landscapes are however affected also by other factors than DNA methylation. For example, many genes up-regulated in the CIMP-subtypes (spot D) can be attributed to immune function and inflammation which constitute cancer hallmarks activated by other mechanisms than methylation.

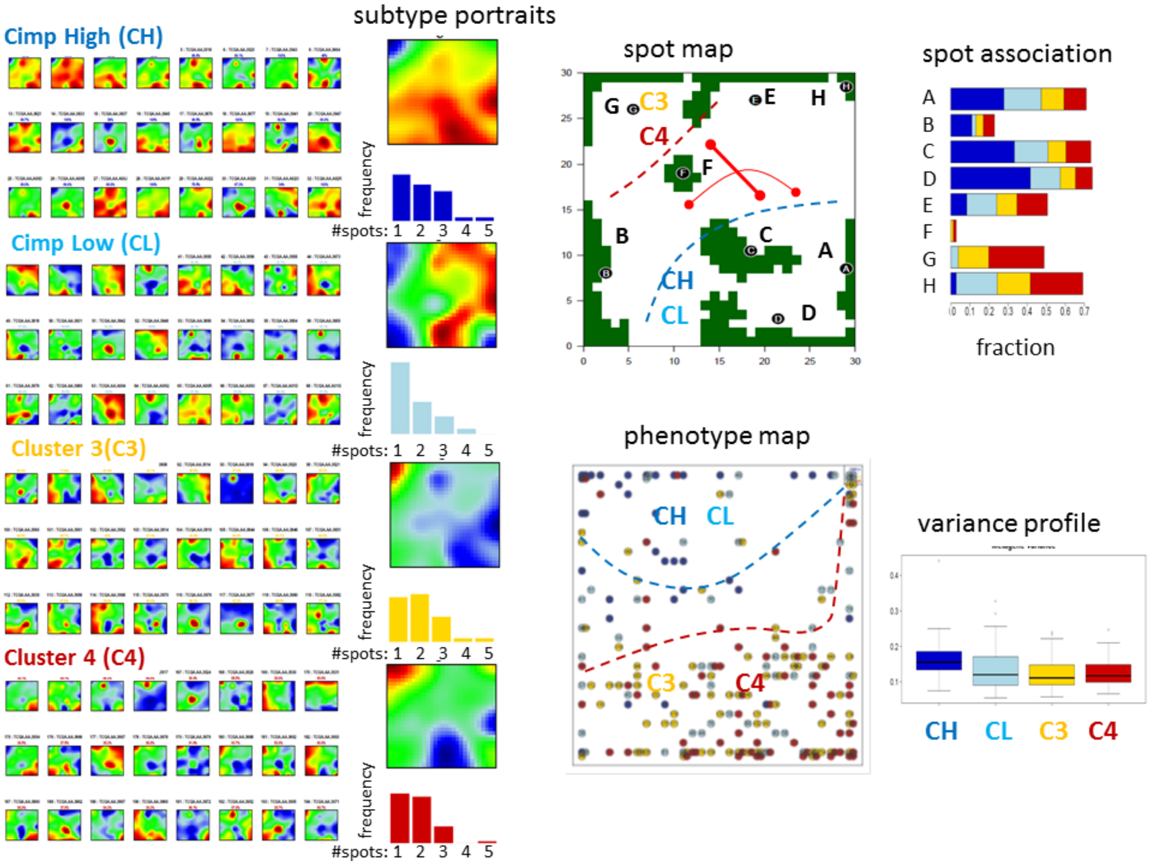


Figure 8: SOM portraying of gene expression data of colorectal cancer (samples and assignment of the samples were taken from (TCGA, 2012)).

Future research directions

Portraying the epigenome: Growing recognition of the importance of epigenetic processes in development and disease has fueled an insatiable thirst for new technologies to detect epigenetic modifications on a genome wide scale. The complex nature of epigenetic modifications and their manifold places many demands on analytical tools and data management. This comprises DNA-methylation and the variety of different histone modifications and their combinatorial patterns. We recently developed a method that combines global epigenome segmentation with SOM machine learning (Steiner et al., 2012). It provides intuitive maps of epigenetic patterns across multiple levels of organization, e.g. of the co-occurrence of different epigenetic marks in different cell types. Another recent application analyzes complex epigenome data using SOM mapping ("An integrated encyclopedia of DNA elements in the human genome," 2012).

OMICs integration and association: Our examples address applications of the portraying method to single OMICs realms such as transcriptome and methylome. There is however increasing need in the emerging field of joint analysis of disparate OMIC data from genomics, transcriptomics, proteomics, etc. in order to better understand key biological processes on the systems level, and particularly, to extract associations and causalities between different OMICs levels, e.g. between mutations and gene activity. Our examples demonstrate that expression and methylation landscapes of colon cancer samples strongly affect each other. Fist attempts are made to combine mRNA and miRNA expression levels using SOM machine learning (Çakir, Wirth, Hopp, & Binder, 2014; Wirth, Çakir, Hopp, & Binder, 2014). Recently a multi-OMICs clustering framework using hidden variables was proposed (Mo et al., 2013). Further conceptual and methodological developments of these tools are needed for the integration of various data types across the multiple levels of OMICs-organization.

Towards personalized medicine: Traditional clinical diagnosis focuses on the individual patient's clinical signs and symptoms. Decreasing costs and increasing technological prospects of modern molecular bioanalytics open the perspective for applying also high-throughput bioanalytics for diagnostic and prognostic tasks: Laboratory technologies however need to be complemented with suited data mining and imaging tools. Here, SOM portraying constitutes one option for an individual view of complex data as one requirement of personalized diagnostics.

Conclusions

Gathering and maintaining large collections of data is one thing, but extracting useful information from these collections is often more challenging. Big Data not only changes the tools one can use for predictive analytics, it also changes our way of thinking about knowledge extraction and interpretation. Ironically, availability of more data at present can lead to fewer options in constructing predictive models, because tools allow for processing large datasets in a reasonable amount of time are often not available. Machine learning constitutes a clever alternative to overcome those problems at the edge of statistics, computer science and emerging applications in systems biomedicine.

SOM machine learning enables recognizing complex patterns in large-scale data generated by high-throughput omics technologies. The method allows portraying molecular phenotypes by generating individualized, easy to interpret images of the particular phenotypic data landscape in combination with phenotype and feature mappings and other analysis options. SOM machine learning reduces dimension of big data and enriches their information content using the concept of stepwise data compression in feature space (single features, meta-features, spot-modules) and the concept of subtyping in sample space.

In future healthcare every patient will be represented by a large dossier of data including massive data from omics diagnostics. Personal portraying of these data as feature landscapes and their projection into phenotype and feature maps previously ‘learned’ from large collectives of patients would provide one way to evaluate quantitatively the risk for the respective patient. Image-based, reductionist machine learning methods thus provide one interesting perspective how to deal with massive molecular omics data in future diagnostics of complex diseases.

References

Augui, S., Nora, E. P., & Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. [10.1038/nrg2987]. *Nat Rev Genet, 12*(6), 429-442.

Beyer, M. (2011). Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. *Gartner*, http://www.gartner.com/newsroom/id/1731916.

The Big Data Conundrum: How to Define It? (2013). *MIT Technology Review, http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/*.

Çakir, M., Wirth, H., Hopp, L., & Binder, H. (2014). MicroRNA Expression Landscapes in Stem Cells, Tissues, and Cancer. In M. Yousef & J. Allmer (Eds.), *miRNomics: MicroRNA Biology and Computational Analysis* (Vol. 1107, pp. 279-302): Humana Press.

Guo, Y., Eichler, G. S., Feng, Y., Ingber, D. E., & Huang, S. (2006). Towards a Holistic, Yet Gene-Centered Analysis of Gene Expression Profiles: A Case Study of Human Lung Cancers. *Journal of Biomedicine and Biotechnology, 2006,* Article ID 69141.

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

Hopp, L., Wirth, H., Fasold, M., & Binder, H. (2013). Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. *Systems Biomedicine, 1*(2).

An integrated encyclopedia of DNA elements in the human genome. (2012). [10.1038/nature11247]. *Nature, 489*(7414), 57-74.

Microsoft. (2013). The Big Bang: How the Big Data Explosion Is Changing the World. *Microsoft UK Enterprise Insights Blog, http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/04/15/big-bang-how-the-big-data-explosion-is-changing-the-world.aspx*.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences, 110*(11), 4245-4250.

Normandeu, K. (2013). Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. *Inside Big Data, http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/*.

Olsen, J., Gerds, T. A., Seidelin, J. B., Csillag, C., Bjerrum, J. T., Troelsen, J. T., et al. (2009). Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflammatory Bowel Diseases, 15*(7), 1032-1038 1010.1002/ibd.20879.

Rindler, A., McLowry, S., & Hillard, R. (2013). Big Data Definition. *MIKE2.0, the open source methodology for Information Development, http://mike2.openmethodology.org/wiki/Big_Data_Definition*.

Sabates-Bellver, J., Van der Flier, L. G., de Palo, M., Cattaneo, E., Maake, C., Rehrauer, H., et al. (2007). Transcriptome Profile of Human Colorectal Adenomas. *Molecular Cancer Research, 5*(12), 1263-1275.

Steiner, L., Hopp, L., Wirth, H., Galle, J., Binder, H., Prohaska, S. J., et al. (2012). A Global Genome Segmentation Method for Exploration of Epigenetic Patterns. [doi:10.1371/journal.pone.0046811]. *PLOS one, 7*(10), e46811.

TCGA. (2012). Comprehensive molecular characterization of human colon and rectal cancer. [10.1038/nature11252]. *Nature, 487*(7407), 330-337.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., & Issa, J.-P. J. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences, 96*(15), 8681-8686.

van Engeland, M., Derks, S., Smits, K. M., Meijer, G. A., & Herman, J. G. (2011). Colorectal Cancer Epigenetics: Complex Simplicity. *Journal of Clinical Oncology, 29*(10), 1382-1391.

van Rijmenam, M. (2013). Why The 3V's Are Not Sufficient To Describe Big Data. *Big data startup, http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/*.

Ward, J. S., & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions. *eprint arXiv:1309.5821*.

Wirth, H., Çakir, M., Hopp, L., & Binder, H. (2014). Analysis of MicroRNA Expression Using Machine Learning. In M. Yousef & J. Allmer (Eds.), *miRNomics: MicroRNA Biology and Computational Analysis* (Vol. 1107, pp. 257-278): Humana Press.

Wirth, H., Loeffler, M., von Bergen, M., & Binder, H. (2011). Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics, 12*, 306.

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

ADDITIONAL READING SECTION

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proceedings Of The National Academy Of Sciences Of The United States Of America, 101(12), 4164-4169.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. [10.1038/nrg2005]. Nat Rev Genet, 8(4), 286-298.

Kim, P. M., & Tidor, B. (2003). Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data. Genome Research, 13(7), 1706-1718.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. [10.1007/BF00337288]. Biological Cybernetics, 43(1), 59-69.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.

KEY TERMS & DEFINITIONS

Gene is the molecular unit of heredity of an organism. Its structural basis is some region of deoxyribonucleic acids (DNA) that code for a polypeptide or for an RNA chain that has a function in the organism.

Genotype is the entirety of hereditary information coded in the genes of an organism.

Phenotype is the entirety of an organism’s observable characteristics resulting from both, the expression of the genes and environmental factors. The molecular phenotype consequently comprises the entirety of molecular building blocks providing the basis of organism’s structure and function.

Feature is an observable characterizing the genotype and/or phenotype of an organism.

Omics is a useful concept in biology aiming at the collective characterization and quantification of pools of biological molecules that translate into the structure, dynamics and function of an organism. Accordingly ‘genomics’ deals with the entirety of an organism's hereditary information coded in its DNA (also called genome); ‘transcriptomics’ deals with the entirety of RNA transcribed from the DNA (transcriptome), ‘proteomics’ deals with the entirety of proteins translated from the mRNA (proteome) and ‘epigenomics’ addresses factors and mechanisms affecting the accessibility of genomic information by modifications of its structure, e.g. via DNA-methylation or chemical modifications of the histones serving as DNA-packing proteins (epigenome).

Chromosome is an elementary unit of packed DNA in the cell nucleus. It represents a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences but also DNA-bound proteins, which serve to package the DNA and control its functions. Chromosomal DNA encodes organism's genetic information. Chromosomes appear as autosomes equally found in males and females as well or as allosomes (sex chromosomes) found either only in males (Y-chromosome) or in different quantities in males (one X-chromosome) and females (two X-chromosomes). Humans contain 22 autosomes and one allosome.

Gene expression is the process by which information from certain functional regions of the DNA defined as genes is transcribed into RNA-products which can be further used in the synthesis of functional products. These products can be proteins (in this case the RNA-product is messenger RNA) or non-protein coding functional RNA.

DNA-methylation substitutes hydrogens at the DNA-nucleotides cytosine or adenine by methyl groups. This chemical modification strongly affects the structure of the DNA and, as a consequence, the expression of the associated genes.

Acknowledgments
H.W., K.L. and L.H. were supported by the Federal Ministry of Education and Research (BMBF), project grant Nos. FKZ 031 6065 (HNPCC-SYS) and FKZ 031 6166 (MMML-MYC-SYS), and by the European Regional Development Fund (ERDF) within the PhD groups ‘Systems Medicine’ and ‘Cancer Microenvironment and Metastasis’.