Expression cartography of human tissues using self organizing maps

Henry Wirth[1,2]*, Markus Löffler[1,3,4], Martin von Bergen[2,5], Hans Binder[1,4]*

[1] Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Härtelstr. 16-18
[2] Helmholtz Centre for Environmental Research, Department of Proteomics, D-04318 Leipzig, Permoserstr. 15, Germany
[3] Institute for Medical Informatics, Statistics and Epidemiology, Universität Leipzig, D-4107 Leipzig, Härtelstr. 16-18
[4] Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment (LIFE); Universität Leipzig, D-4103 Leipzig, Philipp-Rosenthalstr. 27, Germany
[5] Helmholtz Centre for Environmental Research, Department of Metabolomics, D-04318 Leipzig, Permoserstr. 15, Germany

* to whom correspondence should be addressed

Abstract

Background: The availability of parallel, high-throughput microarray and sequencing experiments poses a challenge how to best arrange and to analyze the obtained heap of multidimensional data in a concerted way. Self organizing maps (SOM), a machine learning method, enables the parallel sample- and gene-centered view on the data combined with strong visualization and second-level analysis capabilities. The paper addresses aspects of the method with practical impact in the context of expression analysis of complex data sets.

Results: The method was applied to generate a SOM characterizing the whole genome expression profiles of 67 healthy human tissues selected from ten tissue categories (adipose, endocrine, homeostasis, digestion, exocrine, epithelium, sexual reproduction, muscle, immune system and nervous tissues). SOM mapping reduces the dimension of expression data from ten thousands of genes to a few thousands of metagenes where each metagene acts as representative of a minicluster of co-regulated single genes. Tissue-specific and common properties shared between groups of tissues emerge as a handful of localized spots in the tissue maps collecting groups of co-regulated and co-expressed metagenes. The functional context of the spots was discovered using overrepresentation analysis with respect to pre-defined gene sets of known functional impact. We found that tissue related spots typically contain enriched populations of gene sets well corresponding to molecular processes in the respective tissues. Analysis techniques normally used at the gene-level such as two-way hierarchical clustering provide a better signal-to-noise ratio and a better representativeness of the method if applied to the metagenes. Metagene-based clustering analyses aggregate the tissues into essentially three clusters containing nervous, immune system and the remaining tissues.

Conclusions: The global view on the behavior of a few well-defined modules of correlated and differentially expressed genes is more intuitive and more informative than the separate discovery of the expression levels of hundreds or thousands of individual genes. The metagene approach is less sensitive to *a priori* selection of genes. It can detect coordinated expression pattern whose components would not pass single-gene significance thresholds and it is able to extract context-dependent patterns of gene expression in complex data sets.

1. Background

Parallel, high-throughput biological experiments that simultaneously monitor thousands of molecular observables provides an opportunity for investigating cellular behavior at multiple levels of resolution. Especially, DNA microarray and next generation sequencing technologies allow researchers to screen ten thousands of genes for differences in expression between up to hundreds of individuals or experimental conditions of interest. Not only the progressively increasing data throughput of newest array and sequencing technologies challenges data analysis methods but also the increasing availability of large data sets from public data repositories such as gene expression omnibus (www.ncbi.nlm.nih.gov/geo/) or array express (www.ebi.ac.uk/microarray-as/ae/) with to date hundred thousands of different assays implying large-scale meta-analyses.

These resources pose a challenge how to best arrange and to visualize the huge heaps of data in a fashion that enables a combined sample- and a gene-centered view [1]. Gene expression analysis often pursues the latter gene-centered approach. It emphasizes the identification of lists of genes that change their expression with respect to a reference state. This perspective is suitable for discovering details of the biological processes in which individual genes are involved. Contrarily, the alternative sample-centered perspective aims at assigning samples (rather than genes) to groups according to their expression signature. Aggregated displays partly discard relevant information immanent in the full expression profiles of individual gene-related features leading to a significant loss of systems-level information. Contrarily, detailed gene-centered representations in form of gene lists are unsuitable to survey systems characteristics in large series of heterogeneous samples. Therefore, it would be desirable to possess methods which combine sample- and gene-centered views on multidimensional expression data to capture the global picture of groups of samples while simultaneously presenting the specific expression pattern within each individual sample.

Self-organizing map (SOM) machine learning developed by Kohonen [2] projects data vectors from the original high dimensional space to reference vectors in two-dimensions. Recent applications of the SOM method to microarray expression data emphasized either the gene-centered perspective to cluster genes [3] or the sample-centered mode to map individual samples onto the SOM grid enabling, for example, the classification of tumor samples into a small number of diagnostic or prognostic groups [4-6]. However, the SOM method can be configured also in such a way that it combines both, the sample- and gene-centered perspectives [7, 8]. This option was applied in terms of the so-called ‘gene expression dynamics inspector’ (GEDI)[8]. It decodes the expression pattern of ten thousands of genes per sample into a two-dimensional mosaic pattern which allows the sample-to-sample comparison of expression profiles by direct visual inspection.

The SOM-based GEDI program tool has been applied in studies on cell differentiation and development [9-11], organogenesis [12] and tumor differentiation [1]. It has been demonstrated that SOM analysis not only combines the gene-centered with the integrative, sample-centered views in a well-balanced fashion but also visualizes relevant substructures inherent in the data without forcing them into hierarchies and without significant loss of primary information. This intuitive image-based perception clearly promotes the discovery of qualitative relationships between the samples in the absence of an existing hypothesis.

The SOM approach also enables to employ new concepts of data analysis based on, e.g., global entropy estimates and state-space trajectory characteristics [11, 12]. Moreover, the projection of the original ‘real’ gene expression data into reference vectors provides a simple heuristic to adequately reduce the dimensionality of the data while preserving their full structure without loss of essential information. Note that gene expression analysis typically requires data filtering prior to downstream steps to remove weakly expressed and noisy genes. Usually subjective filter criteria are applied without standardized guidelines. This filtering might be dangerous because it potentially modifies the intrinsic data structure leading to serious limitations in the ability to adequately capture the full information content of the original data set which might be hidden, for example, in weakly expressed structures.

Finally, each SOM-reference vector represents one microcluster of coexpressed ‘real’ genes of similar expression profiles in the samples studied. Coexpressed features are likely to be functionally associated because biological processes are governed by coordinated modules of interacting molecules [13]. This so-called ‘guilt-by-association’ principle is widely invoked in functional genomics [14, 15]. Different methods have been developed to identify such groups in terms of sets of correlated genes

[16], ranked coexpression groups [15] or metagenes [17, 18] using test criteria based on multivariate statistics, rank-based grouping or nonnegative matrix factorization recognizing similarities between subportions of the data, respectively. SOM machine learning thus provides another simple heuristic for functional analysis which implies mutual causal associations between the genes of each metagene microcluster.

In the first instance, this paper focuses on selected methodical aspects of the SOM method not or only partly addressed previously: Firstly, we complement the gallery of primary SOM mosaics with a number of summary maps characterizing the data structure after transformation into latent variables. These summary maps allow extraction of so-called spots which comprise clusters of co-expressed metagenes. Secondly, the detected spots are linked with biological knowledge to support functional interpretation of the data using the 'guilt by association' principle. Particularly, we apply gene set overrepresentation analysis to visualization space on two different levels of data compression given by the metagenes and by spots of metagenes, respectively. This grouping of coexpressed genes enables to significantly reduce the dimensionality of expression data from ten thousands of single genes to a handful of representative features. Thirdly, we therefore analyze the capability of the SOM approach for data filtering and dimension reduction in terms of maintaining representativeness and reduction of noisiness of the data. Particularly, we tackle the question whether substitution of single genes by metagenes improves the performance of downstream agglomerative methods such as hierarchical clustering, correlation and independent component analysis. Here we follow partly previous studies on a smaller and less diverse data set [1]. Finally, we applied SOM analysis in a sample-centered second-level version allowing the more detailed assessment of similarity relations between the samples. We developed our own R-program including all analysis functionalities described below for application of the method in a R-environment. Our SOM method includes the calibration of microarray raw intensity data to minimize possible artifacts due to systematic biases caused by improper preprocessing [19].

Microarray expression data of a collection of human tissues were chosen as an illustrative example: Firstly, the selected 67 tissues provide a sufficient large data set of highly diverse expression pattern possessing a complex internal covariance structure. Secondly, the samples are well classified in terms of distinct tissues and tissue categories allowing the clear assignment of expression pattern. Despite these methodical issues the discovery of the human body index data set in this study is also motivated by the argument that tissue-specific RNA expression pattern indicate important clues to the physiological function of the coding genes suitable as a reference for comparison with diseased tissues, as well as a basis for identifying molecular markers of injury to specific organs and tissues. Our analysis thus provides a first step towards a SOM atlas of gene activity in normal human tissues which complements previous work on the diversity of gene expression in human tissues [20-22].

2. Results and discussion

2.1. Expression maps of human tissues

Microarray expression data taken from the human body tissue index data set were input into the SOM machine learning algorithm after calibration and normalization of the raw probe intensities as described in the Methods section below. Our SOM method transformed the whole genome expression pattern of about 22,000 single genes into one mosaic pattern per tissue studied. Figure 1 shows selected SOM-fingerprints of 42 selected tissues using a 60x60 mosaic grid. The collection of SOM of the complete set of 67 tissues is given in Additional file 2. Each tile of the SOM mosaics refers to one of 3,600 metagenes characterizing the expression landscape of the data set. The metagenes act as representatives of miniclusters of single genes with similar expression profiles. Their number varies from metagene to metagene (see below). The color gradient of the map was chosen to visualize over- or underexpression of the metagenes in the particular tissue compared with the mean expression level of each metagene in the pool of all samples studied: Maroon codes the highest level of gene expression; red, yellow and green indicate intermediate levels and blue corresponds to the lowest level of gene expression. Each individual mosaic exhibits characteristic spatial color patterns serving as fingerprint of the transcriptional activity of the respective tissue sample.

The tissues are grouped into ten categories in accordance with the classification used in Hornshoj et al. [23]. Most of these categories show typical SOM-landscapes which are characterized by red and blue spots at specific positions due to over- and underexpressed metagenes as the most evident features. 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.

Some tissues combine the characteristic spot pattern of different tissue categories (see Figure 2). For example, the expression fingerprint of tongue (no. 24) shows the typical overexpression spot evident in the profiles of other epithelial tissues (e.g. 21: oral mucosa) but also the spot typically found in muscle tissues (e.g. 32: skeletal muscle). The physiology of tongue tissue as a ‘mucosa covered muscle’ is thus reflected in the expression profile. Another example is pituatary gland (profile no. 5), an endocrine gland located near hypothalamus: Its SOM landscape shows the upregulated spot found in other nervous system tissues (e.g. cerebral cortex or the adjacent hypothalamus, no. 49 and 56, resp.) in the left upper corner, as well as a unique spot in the right lower area not found in the profiles of other tissues. This spot obviously collects genes which are specifically overexpressed in pituatary gland (see below), whereas the first spot represents a common signature typically found in nervous system samples. Some SOM-fingerprints are outliers in their tissue category: For example, small intestine (no. 12), classified as digestive tissue, shows the overrepresentation pattern of muscle type tissues. This is not surprising as this organ consists of a double layer of smooth muscle. Also myometrium (no. 33), the smooth muscle of the uterus, is classified as muscle. Its SOM expression profile however closely resembles that of endometrium (no. 26) and also of ovary (no. 27), reflecting the common function of these three organs in female reproduction.

Taking together, comparison of the individual SOM fingerprints within each tissue category reveals similar pattern in most cases whereas different tissue types show consistent differences between their landscapes. Such differences can be detected, for example, by simple visual inspection of the mosaic pattern of nervous, immune system and endocrine type tissues. Hence, comparison of the SOM-textures allows the straightforward grouping of the tissues into different categories based on differences of their expression patterns.

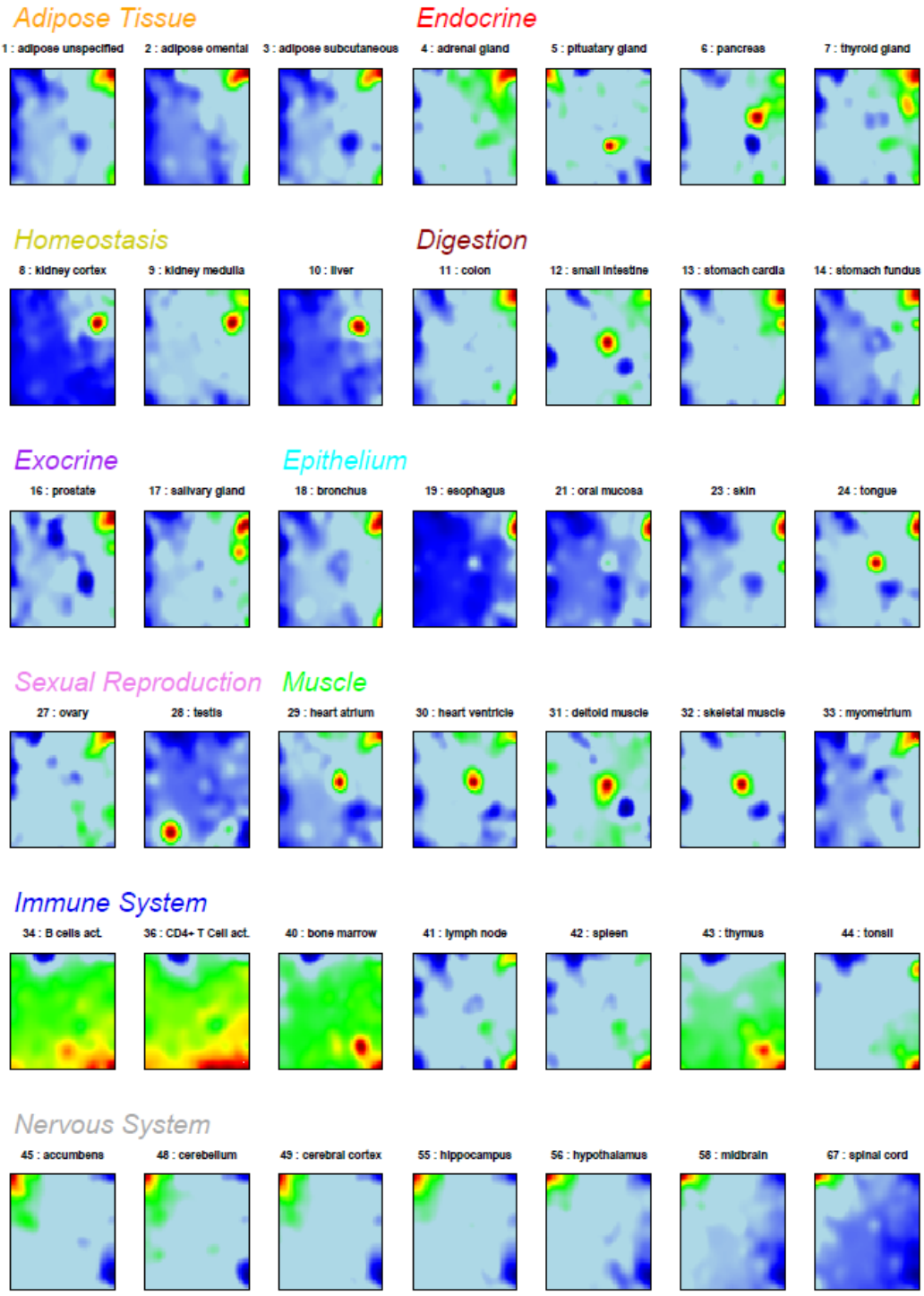

Figure 1: SOM expression profiles of 42 selected tissues. The tissues are sorted according to tissue categories in agreement with the classification used in Hornshøj et al. [23]. The color of the heading of each tissue category and the numbering of tissues are used also in the other figures throughout the paper.

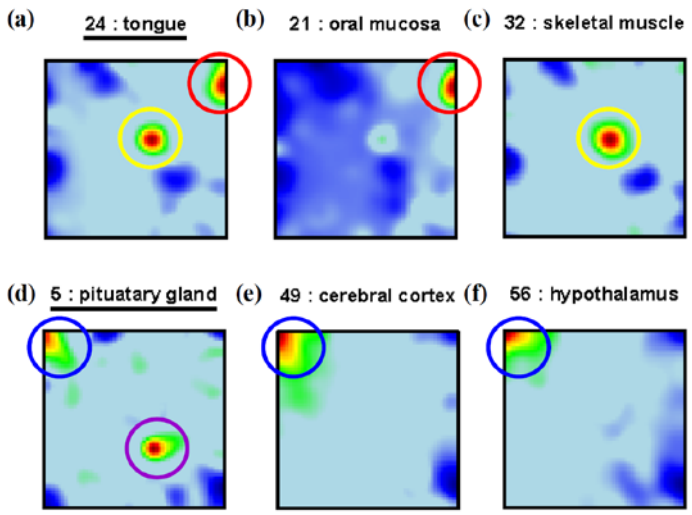

Figure 2: Specific spots in selected expression profiles: The SOM-pattern of tongue (a) shows two spots of upregulated metagenes. One of them is characteristic for mucosa type tissues (b; red circles) and the other one is found in muscle tissues (c, yellow circles). Pituatary gland (d) shows a specific spot for this particular tissue and one which is characteristic for nervous system tissues (e and f, blue circles) as well.

2.2. Metagene characteristics and overexpression spots

The metagene expression profiling map in Figure 3a illustrates the systematic character of the alterations of metagene expression within the SOM with strongest similarities between adjacent metagenes. It also shows that more distant metagenes largely differ in their profiles: In the centre one finds virtually invariant metagenes whereas the profiles along the borders of the map more strongly vary thus collecting specific information for selected tissues. This distribution of the profiles reflects the fact that SOM machine learning tends to maximally segregate different modes of variable profiles on one hand while maximizing their distance with respect to virtually invariant profiles on the other hand. Note also, that the number of real genes per metagene strongly varies as indicated by the numbers given in each tile of the metagene expression profiling map.

The metagene expression profiling map uses a smaller number of tiles and thus a coarse grained latticing of the mosaic. The population and variance maps shown in panel b and c of Figure 3 provide information about the number of single genes per metagene minicluster and the variability of the metagene profiles via appropriate color coding using the finer granularity of the individual tissue profile. SOM-machine learning scales the difference between the expression profiles of adjacent metagenes inversely to their population, i.e., adjacent metagene profiles become more similar for highly populated metagenes. This way the method tends to distribute the single genes over as much as possible tiles. The population map reveals that the real genes nevertheless inhomogeneously distribute among the tiles of the mosaic (Figure 3b). Highly populated metagenes (n_k>20, see yellow and red tiles) predominantly group along the edges of the map whereas only a few highly populated tiles are found in its central area. A zone of ‘empty’ metagenes lacking real genes (n_k=0, see dark blue tiles) clusters in four regions halfway between the centre and the edges of the map. The tile of maximum population (n_k=308, see the dark brown tile slightly left from the centre of the map) refers to genes with virtually invariant, mostly absent specific expression in all tissues studied. These genes form the strong peak in the distribution of differential expression shown in the methods section below (see also Figure 11c and d below).

These invariant genes give rise to the dark blue spot in the central area of the variance map (Figure 3c). The variance map also reveals that other nearly invariant metagenes cluster around this tile in the central area of the map (see blue and green areas in Figure 3c). Both, invariant and empty metagenes carry essentially no specific information as classification markers in transcriptional profiling. Hence, the tiles occupied by empty and invariant genes form regions not suited for differential expression analysis between the tissues studied.

The more variant and higher populated metagenes reveals an underlying spot like pattern in direction towards and at the boundaries of the map (red areas in Figure 3c) which largely agrees with the over- and underexpression spots detected in the SOM mosaics of individual tissues. For an overview about

all observed spots we generate two types of integral maps characterizing over- and underexpression, respectively (Figure 3d and e). They transfer either the over- or the underexpression spots observed in the individual profile into one master map. The profiles of selected metagenes reveal marked under- and overexpression for distinct tissue types which transform into a characteristic spot patter (see Figure 1 and Figure 3). For example, the metagenes in the left upper corner show overexpression for nervous system and underexpression for immune system tissues whereas the metagenes in the right lower corner are, in turn, characterized by overexpression in immune system tissues. Table 1 assigns the different spots to the tissues mosaics in which they appear.

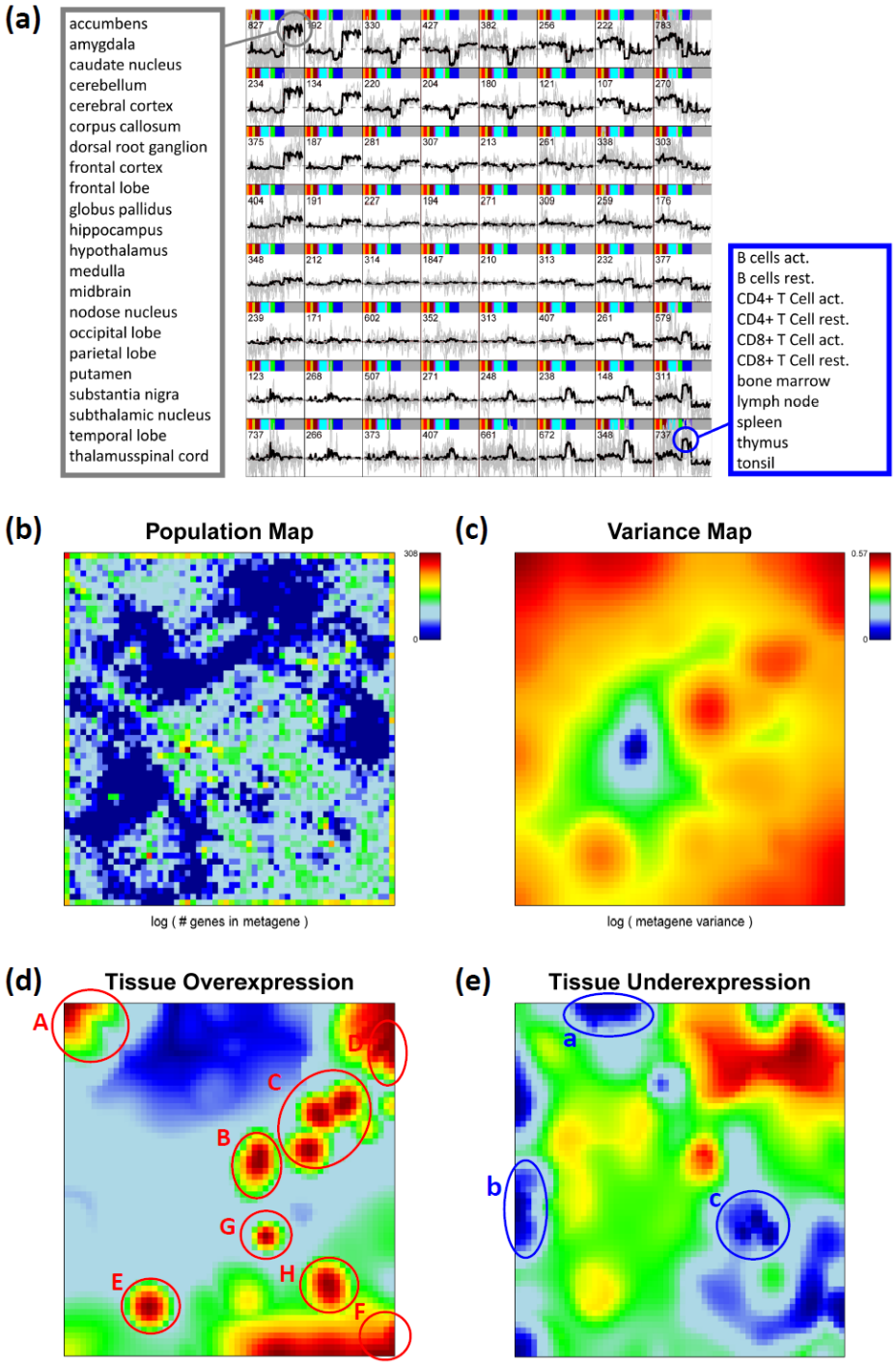

Figure 3: Metagene characteristics: Metagene expression profiling map of the 67 tissues studied (panel a), population (b), variability (c, Eq. (2)), metagene over- (d) and underexpression (e) maps. Panel a): Metagene profiles are shown by thick curves whereas thin grey ones show the profiles of associated real genes. The vertical

axis is the logged expression change relatively to the mean expression of the selected gene averaged over all tissues. All tiles use the same vertical scale. The number in each tile gives the population of the respective metagene cluster with real genes. The bars color-code the tissue samples (compare with headings in Figure 1). The circles indicate over- and under-expression in selected tissues listed in the boxes (see text). One sees that the metagenes in the left upper and the right lower corner cluster genes strongly overexpressed in nervous (grey circle) and immune system (blue circle) tissues, respectively. Panel d) and e): Red/maroon spots mark overexpression, blue ones underexpression. Selected spots are marked by letters (capital and lower case letters refer to maxima and minima, respectively). They are assigned to different tissues in Table 1.

Table 1: Functional assignment of tissue specific over- and underexpression spots using the GO-terms biological process/molecular function (see also Figure 3d and e).

Spot [a]	Over-/underexpressed in tissue [a]	Biological process / Molecular function (overrepresented genes set) [b]
A	Nervous system samples (45-67), pituatary gland(5)	Nervous system development Synaptic transmission Transmission of nerve impuls
B	Muscle related: small intestine (12), tongue (24), heart atrium&ventricle (29, 30), muscle (31, 32)	Structural constituent of muscle System process Striated muscle contraction
C1	Liver (10), kidney cortex&medulla (8,9)	Substrate specific transporter activity Carboxylic acid metabolic process Organic acid metabolic process
C2	Pancreas (6)	Carboxypeptidase activity Carboxylesterase activity Digestion
D	Adipose tissue (1-3), epithelium tissue (18-26), ovary (27)	Tissue development Organ development Ectoderm development
E	Male reproduction: testis (28)	Sexual reproduction Reproduction Gamete generation
F	Immune system samples (34-44)	Immune system process Immune response Defense response
G	Pituatary gland(5)	Hormone activity DNA fragmentation during apoptosis Apoptotic nuclear changes
H	Bone marrow (40), thymus (43)	Cell cycle process Mitotic cell cycle Cell cycle phase
a	Immune system (34-44)	Regulation of axonogenesis Regulation of structural morphogenesis Regulation of neurogenesis
b	Various samples without clear assignment, e.g., sexual reproduction and muscle	Microtubule binding Protein maturation Tubulin binding
c	Epithelium and muscle tissues	RNA metabolic process Biopolymer metabolic process RNA processing

[a] Spots are assigned in Figure 3d and e

[b] Top-three overrepresented gene sets

2.3. Gene set overrepresentation

The SOM assigns mini-clusters of real genes to each metagene represented by a tile in the two-dimensional mosaic pattern. For each of these mini-clusters one can estimate the degree of overrepresentation with respect to genes taken from a pre-defined gene set using the hypergeometrical (HG-) distribution distribution. It provides one p-value per metagene which can be visualized in the same two-dimensional mosaic as the original SOM via appropriate color-coding. This so-called overrepresentation map allows identification of metagenes containing an overrepresented fraction of genes from the particular gene set by visual inspection. This map applies to all samples studied because the genes in each of the mini-clusters are invariant for all samples used to train the SOM. The overrepresentation map thus reflects the global overrepresentation pattern in the experimental series studied.

Figure 4 shows overrepresentation maps for selected gene sets. Their overrepresentation is usually observed in different regions of the map, for example in the right lower and left upper corner for genes related to 'immune system process' and to the 'transmission of nerve impulse', respectively. The examples also show that overrepresentation is either strongly localized in one region of the map (e.g. for 'nervous system' or, to a less degree, for 'RNA repair' and 'immune system process') or it spreads over wider areas of the SOM (e.g. for 'apoptosis').

Overrepresentation analysis is not restricted to single tiles but it can be applied also to the over- and/or underexpression spots of adjacent tiles discussed in the previous subsection. Hence, overrepresentation of selected gene sets can be linked with alternative properties of the expression profiles such as overexpression simply by a combining spot selection and subsequent overrepresentation analysis. Particularly, the genes associated with each spot are analyzed for overrepresentation of genes taken from the collection of 1454 gene sets downloaded from the GSEA-homepage according to the GO-categories molecular function, molecular process and molecular component (see methods section). The hypergeometrical distribution then provides an ordered list of gene sets ranked with decreasing significance of overrepresentation for each of the spots.

Figure 5a shows nine spots of strongly overexpressed metagenes and the three leading gene sets in the list to get a first idea about the possible molecular function of the genes in the spot. The most significant, top-twenty gene sets for three selected spots are presented as bar plots in Figure 5b. For example, spot A in the left upper corner of the SOM is clearly related to molecular processes in nervous cells according to the three leading gene sets (see also Table 1). In addition, ten out of the top-twenty gene sets are also related to nervous system (Figure 5b). 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).

The heat map in Figure 6 links overrepresentation of the three topmost gene sets with differential expression in a tissue- and spot-specific fashion. It clearly reveals that essentially only one spot with clearly assigned molecular function is overexpressed in nervous (see grey bar on top of the heatmap for assignment), muscle (green) and homeostasis (ocher) tissues. Some of these tissue-specific gene sets are also overexpressed in other tissues. For example, the muscle-specific gene sets are highly expressed also in tongue and small intestine which partly contain muscle tissues (see above).

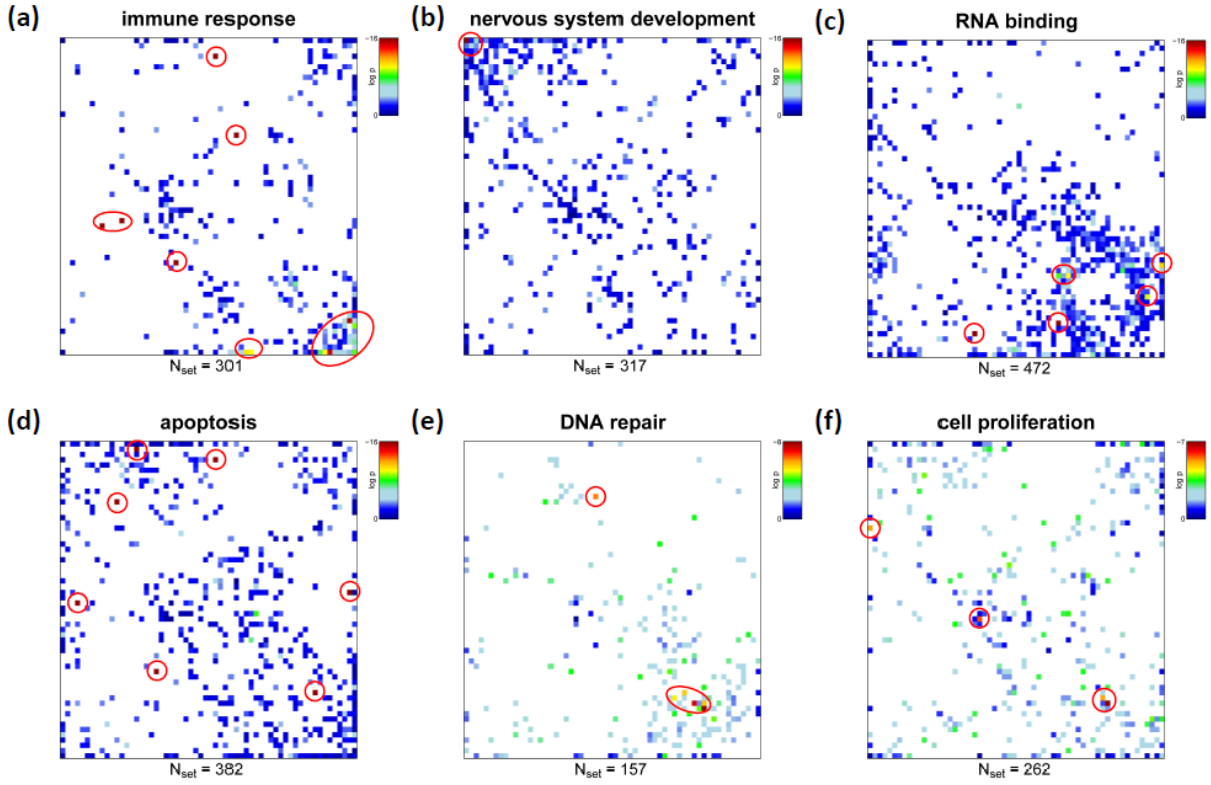

Figure 4: Overrepresentation maps of six selected gene sets containing between N_{set}= 157 and 472 genes. Overrepresentation in each tile of the mosaic is calculated in units of log(p_{HG}) using the hypergeometrical distribution and color-coded (maroon>red>yellow>green>blue). White areas indicate metagenes not containing genes from the respective set). Strongest overrepresentation of the different gene sets is found in different regions of the SOM (see red circles). Overrepresentation can be concentrated within one or a few adjacent metagenes (e.g. nervous system, panel b) or spread over different disjunct regions of the map (apoptosis, panel d).

It is noteworthy that the enriched areas in the overrepresentation maps of 'nervous system development' and of 'immune response' gene sets (see Figure 4) largely agree with the overexpression spots in the expression profiles of nervous and immune system tissues, respectively. A non-negligible number of members of these gene sets are however located also in other regions of the overrepresentation map which are partly assigned to alternative functions. For example, genes from the 'immune response' set also accumulate in spot D assigned to tissue development. It is overexpressed in a larger number of tissues not explicitly assigned to the category of immune system tissues. Moreover, subgroups of genes from these gene sets are located in the central area of the map which accumulates virtually invariant and weakly expressed genes (compare with Figure 3). Possibly part of the genes in these sets are incorrectly specified and/or possess a more complex activation pattern 'beyond' the similarity metrics used to train the SOM. We suggest that combination of gene set overrepresentation analysis with SOM-expression profiling allows verification and further refinement of existing gene sets.

In summary, gene set overrepresentation analysis links selected gene sets and different regions of the SOM with single-tile resolution. These regions, in turn, can be collected into over- or underexpression spots in different tissues. Overrepresentation analysis then provides lists of significantly overrepresented gene sets which characterize the respective spot in a functional context. Some of the spots can be assigned to specific molecular characteristics such as 'nervous processes', 'muscle contraction' and 'immune response'. Both, the single-set SOM-wide and the multi-set spot-wise overrepresentation analysis constitute a link between characteristic expression pattern and concepts of molecular function of the associated genes. These orthogonal views complement each other: The former one judges the homogeneity of a selected set with respect to different metagene expression profiles. The latter one assigns selected expression profiles to their tentative molecular function.

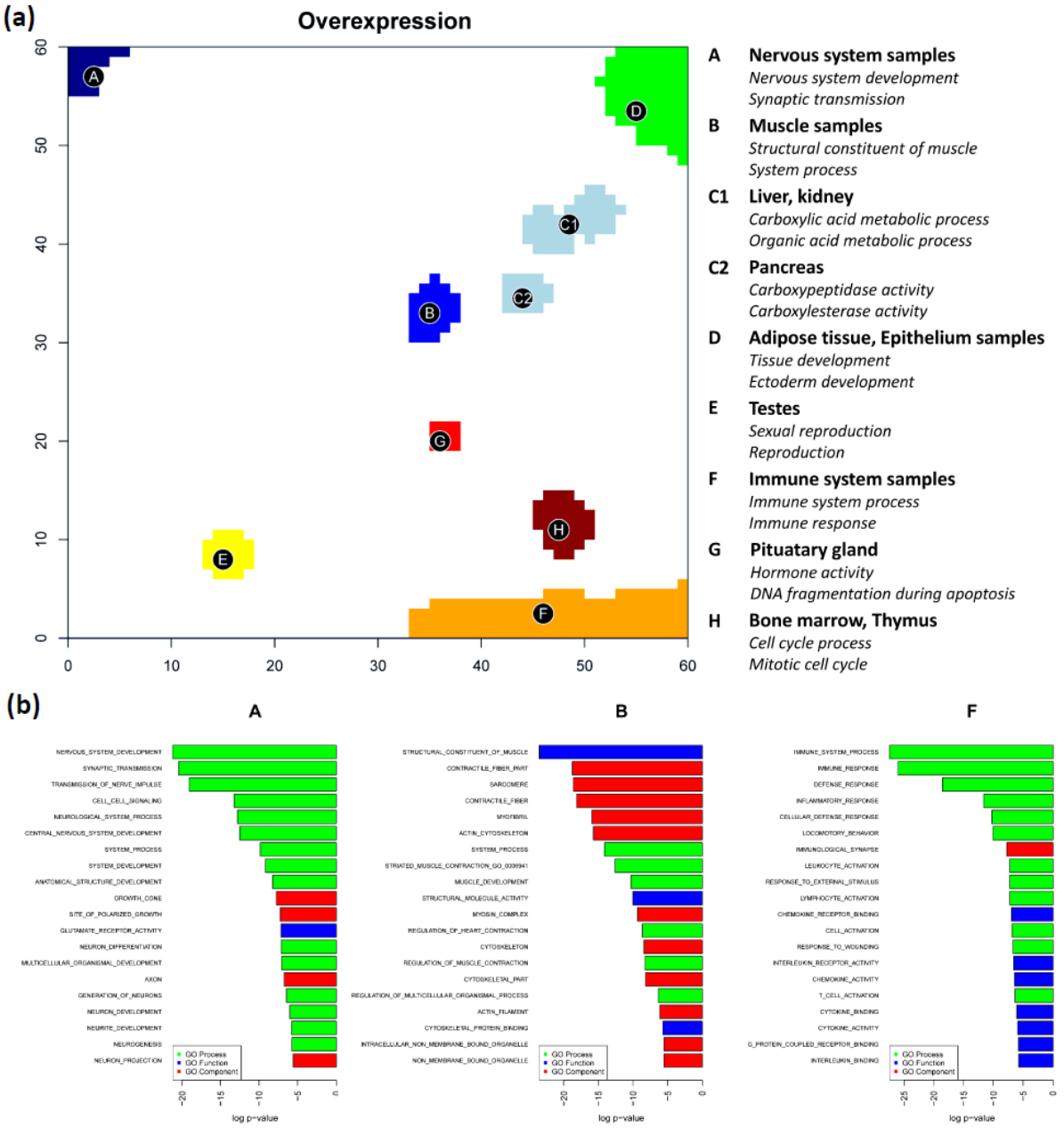

Figure 5: The overexpression summary map shows nine spots which are strongly overexpressed in different tissues (part a). Overrepresentation of a collection of 1454 gene sets is estimated for each spot using the hypergeometrical distribution. The right legend assigns the three most significantly overrepresented gene sets to the respective spots. The top-twenty gene sets of the ranked list are shown in part b for three selected spots. The length of the bars scales with the logged overrepresentation p-value of the sets. The color assigns the category of the gene sets according to the GO terms ‘molecular process’ (green), ‘molecular component’ (red) and ‘molecular process’ (blue).

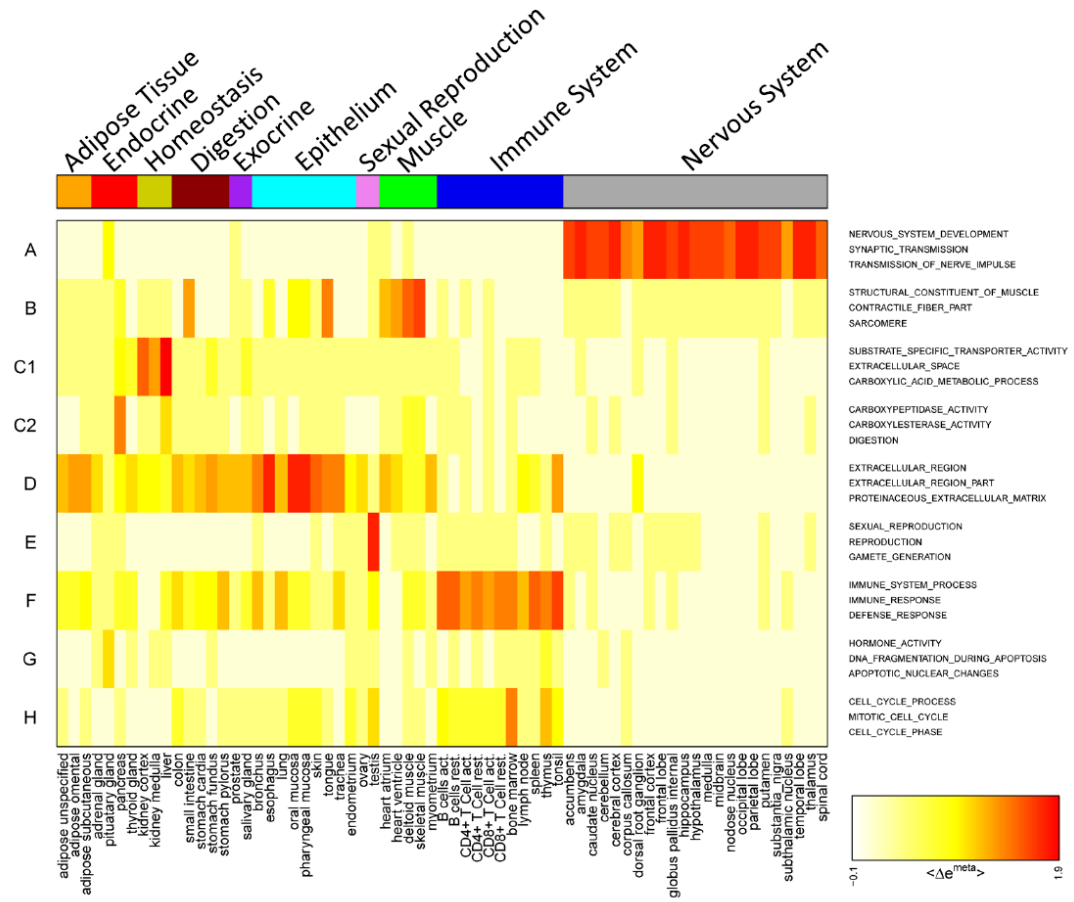

Figure 6: Overexpression summary heatmap of selected global spots (A – H, see Figure 5 and Table 1) in all tissues studied. They are grouped in horizontal direction according to their tissue categories (see color bar on top of the map where the colors are assigned to the categories in agreement with Figure 1). Each spot refers to one row. It is associated with the top-three overrepresented gene sets on the right axis of the map. The expression scale refers to the maximum metagene expression in the respective spot.

2.4. Filtering metagenes and single genes

The reduction of the size of the data set by removing genes that carry essentially no or low information is common practice to improve downstream analysis such as two-way hierarchical clustering of genes and samples. Such data reduction has been shown to result in dendrograms which more accurately reflect relationships between the samples with increasing stringency of the filter applied [24]. This improvement can be rationalized by the fact that random noise tends to disrupt similarity relations between genes and samples. On the other hand, also the opposite trend is possible: systematic errors in the data, e.g. due to batch effects, can cause artificial clustering if the bias affects subsets of genes in a coordinated fashion. Hence, a particular filter aims at improving data by removing either noisy, biased and/or weakly expressed genes. On the other hand, extreme filtering is dangerous because it may eliminate valuable information, for example, about genes of relatively low and thus noisy expression but with important biological impact. Hence, filtering is an optimization task with the requirement of removing virtually irrelevant data while preserving all information in the remaining part of the data which is important in the context of the particular issue studied. We will shortly call the latter property as the ‘representativeness’ of a filter and the former one as its ‘noisiness’, i.e. the mean noise-to-signal ratio of the data included. Optimization thus aims at maximizing representativeness while minimizing noisiness.

‘Top-list selection’ is probably the simplest method of filtering: One first defines a ranking criterion such as differential expression or variability (see below), then one ranks the data accordingly and finally selects a certain number of features on top of the list for further analysis. The length of the list can be cut by applying different criteria such as a fixed number of features or a significance threshold.

SOM analysis enables alternative filtering based on the metagenes as representative features characterizing the expression profiles of microclusters of single genes. In other words, the metagene profiles itself can serve as a filtered and compressed extract of the original data. Our SOM-method assigns the expression profiles of the N=22,277 input genes measured in 67 tissues to 3,600 metagene clusters. Each metagene cluster consequently contains G/M=<n_k>=6.2 real genes on the average. Hence, complexity of transcriptome characterization is reduced to about one sixth by utilizing the metagenes instead of the 'real' genes.

Moreover, the local G/M-ratio considerably varies between the different metagene clusters with minimum and maximum values of n_k=0 (empty metagenes) and n_k=308 (see Figure 3b). Thus each metagene can be representative for a very different number of real genes. In consequence, the importance of transcriptome information is effectively reweighted by using metagenes instead of real genes. For example, the metagene of highest population (n_k= 308) clusters genes of virtually invariant expression profiles. These essentially not-informative features comprise 1.4% (308/22,277x100%) of all single genes but only 0.3% (1/3,600x100%) of all metagenes. Hence, their contribution is effectively down-scaled by nearly a factor of ~1/5 if one uses the metagenes instead of real genes. In other words, SOM clustering itself can be viewed as a sort of selective compression filter reducing the number of features considered by condensing larger numbers of similar single gene profiles into one metagene profile with a profile-specific compression factor, $F_k^{compression}= (n_k \cdot K/N)^{-1}$ (K and N are the total numbers of metagenes and of single genes).

Metagene filtering is expected to outperform single gene filtering in terms of representativeness and noisiness because the reduced number of metagenes not only preserves the diversity of the different single gene profiles but also amends the resolution of downstream analysis due to the reduced noise of the metagene profiles. With the objective of proving this expectation we compare two options for data filtering by applying top-list selection either to the metagenes or to the single 'real' genes. We used three types of filters to reduce the number of single genes and metagenes, namely fold change(FC)-expression, variance and significance (FDR-) filtering (see Additional file 3 and the methodical section). In the first case the full set of absolute FC-values of all genes (real genes and metagenes) under all conditions studied are ranked and a certain number of topmost features is considered for further analysis.

Note that lists of equal numbers of metagenes and single genes are asymmetric owing to data compression of the metagene miniclusters discussed above. The different sample sizes selected by both options of filtering are given in detail Additional file 3. Metagene lists integrate roughly a tenfold larger list of 'real' genes in our particular SOM settings. Figure 7 compares therefore the selected areas in the SOM mosaic filtered by FC-lists of different lengths if applied either to metagenes or to single genes. The shorter metagene lists cover essentially the same regions of the SOM as the longer single gene lists with considerable overlap of the selected meta- and single genes. The large overlap demonstrates that the metagene filter is representative for the metagene-associated single genes which to a large fraction are also selected if one applies single gene filtering using a much longer list. For example, 3,529 out of the 3,600 single genes are shared by the FC-3600 single gene and the FC-1000 metagene lists ('FC-1000' denotes the 'fold change top-1000' criterion, see Figure 7a). However, 444 out of the top-1000 metagenes do not contain the genes from the single gene list which, on the other hand, contains 71 single genes in 44 metagenes not selected by the metagene list. Hence, the metagene filter covers a wider range of expression profiles than the single gene filter which selects only a few additional features. Figure 7b illustrates that different spot areas are progressively excluded from the list of filtered features with increasing stringency of the filter as expected.

In addition to FC-filtering we applied variance and significance filtering which select profiles of largest variance or of highest significance of differential expression. The former filter type possesses similar properties as the FC-filters. In contrast, significance filters select more diverse collections of features spread over partly different areas in the respective mosaic representations (see Additional file 3). Below we apply FC-filtering in the more detailed analysis to judge the consequences of both filters for selected downstream characteristics.

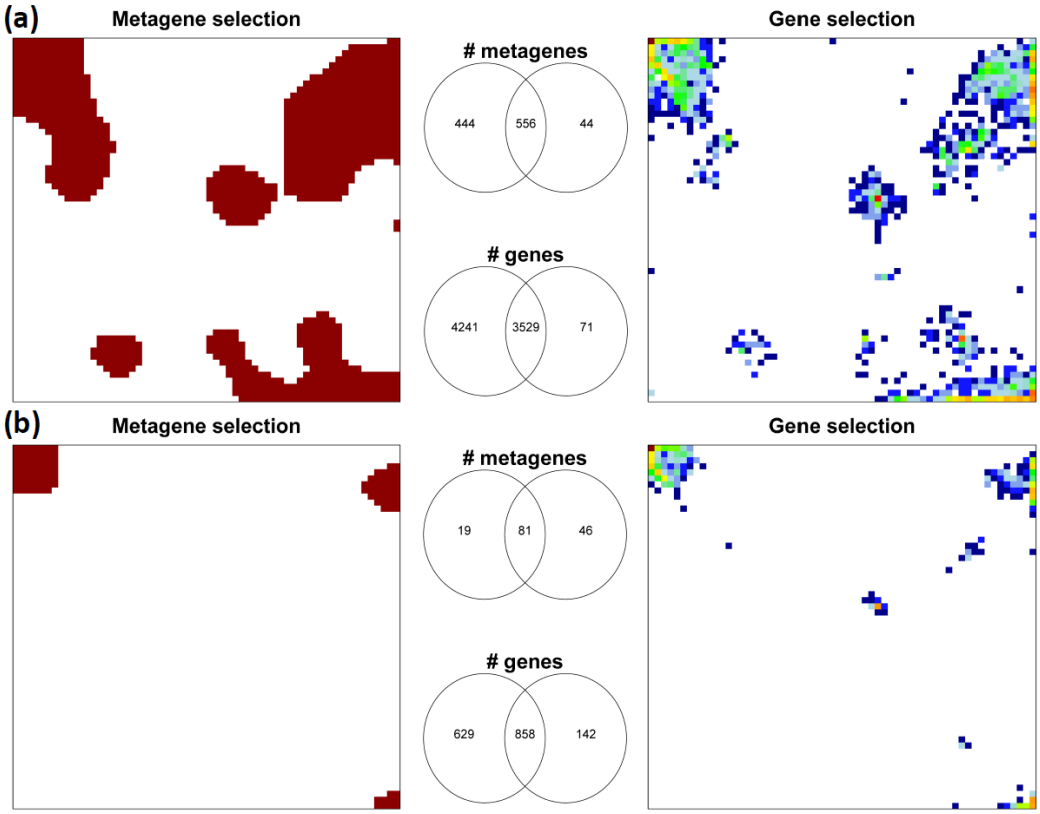

Figure 7: Filtering genes or metagenes by differential expression: Different numbers of metagenes (left mosaics) and single genes (right mosaics) are selected using the FC-1000/FC-3600 (a) and FC-100/FC-1000 (b) filters to account for the data compression in the metagene clusters. The brown areas in the left part show the selected metagenes and the colored tiles in the right part the fraction of single genes in the metagene miniclusters (maroon to blue codes high to low fractions). The Venn-diagrams illustrate the degree of overlap between the metagenes and single genes selected by both filters.

2.5. Metagene- and single genes-based clustering analysis

In the next step we subjected the lists of filtered genes and metagenes to secondary standard analysis methods to assess the particular effect of filtering. We performed one- and two-way hierarchical clustering and independent component analysis (ICA) using either the expression values of a list of real genes or of a list of metagenes of selected lengths. Hierarchical cluster analysis was applied because this method is often routinely run as a first step of data summary in microarray data analysis [25].

One way hierarchical cluster trees obtained from single gene and metagene FC-lists of length 3600, 1000 and 100 reflect similar properties showing that clustering is relatively robust with respect to the chosen conditions (Figure 8a). Tissues from categories with homogenous SOM-pattern such as nervous (grey), adipose (orange) and immune system (blue) tissues (see also Figure 1) robustly cluster together at nearly all conditions studied. Note that the blue cluster of immune system tissues however partly decomposes if one uses the shortest single gene list (FC-100) owing to the loss of representativeness. On the other hand, the FC-100 metagene list of equal length still produces a compact blue cluster reflecting the improved representativeness of the same number of metagenes.

The blue immune system tissue cluster splits for both, the single gene and metagene filters in the opposite limit of low stringency using FC-3600 lists. These lists obviously become too long with worse noisiness characteristics. Note, that the FC-3600 metagene list considers all available metagenes whereas the FC-3600 single gene list is still limited to only 16% of all available single genes. Longer single gene lists even more downgrade the observed cluster structure due to the progressive inclusion of noisy genes (data not shown). In summary, metagene lists are more representative and less noisy than single gene lists of equal length in downstream cluster analysis. On the other hand, also the length

of metagene lists is optimum in the intermediate range (e.g., the FC-1000 list in our study): shorter and longer lists are suboptimal in terms of representativeness and noisiness, respectively.
The cluster trees generated on the basis of single gene and metagene lists reveal another interesting difference (compare the first and second rows in Figure 8a): The mean length of the outmost branches between the periphery of the circles and the first split point is considerably shorter for the metagene-based trees than for the single gene-based trees. This relation reverses for the innermost branches. This systematic difference reveals that metagene clusters are more compact than the respective single gene clusters (an illustrative explanation for this difference is given in Additional file 3) which, in turn, reflects the decreased noisiness of the metagene data. In the right part of Figure 8a we compare the inter-to-intra cluster ratio of the Euclidian distances between the samples (F-score) for three tissue categories as a simple measure of the compactness of their clusters. The F-score of the metagenes systematically exceeds that of the single genes.
Figure 8b shows two-way hierarchical cluster heatmaps after FC-filtering of metagenes and single genes. This type of representation visualizes similarity relations between the samples in horizontal direction (see the color bars which assign the tissue categories) and between the filtered genes in vertical direction. One immediately observes that the contrast of the heatmaps increases from the left to the right because more stringent filters trivially accentuate larger differences between over- (red) and under (blue) expressed features. The loss of contrast for the longer FC-3600 and FC-1000 lists compared with the FC-100 list is stronger for the metagenes because data compression includes a larger fraction of features of small differential expression (green and light blue areas) then the respective single gene lists. On the other hand, the short FC-100 list of metagenes produces the heatmap of strongest contrast illustrating the favorable signal-to-noise characteristics of the metagenes.
The heatmaps express detailed information about the amount of genes differentially expressed in the various tissues (cluster size, see the right part of Figure 8b). For example, the percentage of single genes consistently overexpressed in the nervous tissues and underexpressed in the other tissue categories (see also the green/maroon area associated with the grey bar on top of the heatmaps) increases from values of less than 50% (FC-3600) to a dominating amount of more than 90% (FC-100) whereas the percentage of genes overexpressed in other tissue categories nearly vanishes. Hence, the relative contribution of genes collected into clusters characterizing a selected tissue clearly depends on the length of the list. The use of metagenes instead of single genes effectively re-weights the contribution of tissue-specific genes. Particularly, the percentage of metagenes which are specific for nervous tissues is markedly smaller in the metagene list giving rise to a more balanced distribution of features.

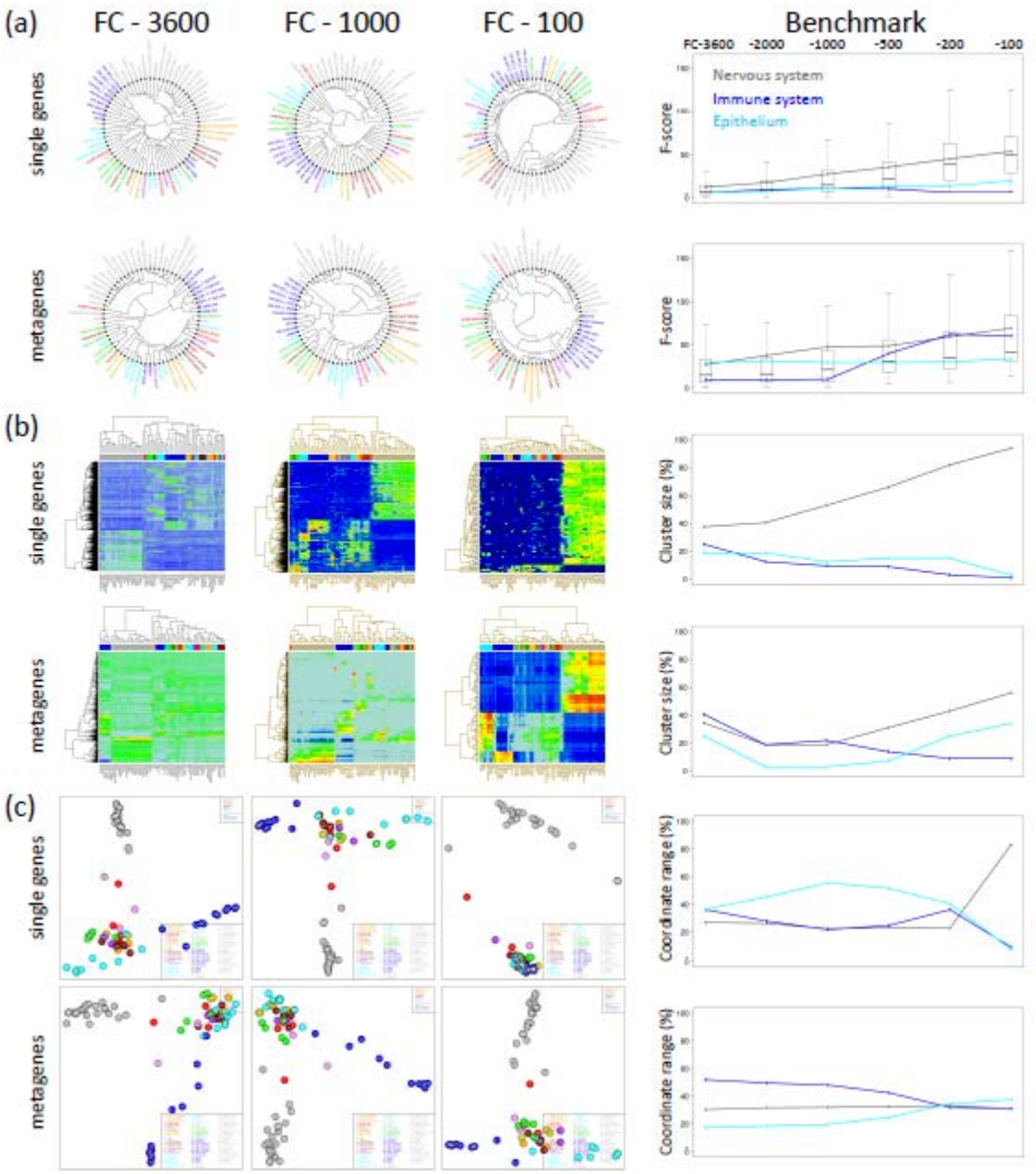

Figure 8: The effect of filtering of single genes and metagenes on the results of one-way hierarchical clustering trees (part a), two-way hierarchical cluster heatmaps (part b) and independent component analysis (part c) of the 67 tissues studied. The samples are color-coded according to the classification of tissues introduced in Figure 1. Top-list FC filters select the 3600, 1000 and 100 (from left to right) most strongly differentially expressed genes/metagenes in all samples. Note that the ICA-plots are invariant with respect to reversing the direction(s) of the coordinate axe(s) and thus to mirror and rotational symmetry operations. The right part shows different benchmark criteria for different lengths of the FC-lists ranging from FC-3600 to FC-100 (see top axis). The benchmark criteria were applied to nervous system, immune system and epithelium tissues (see text and Methods section).

2.6. Metagene- and single genes-based ICA analysis

While clustering may identify groups of samples which share genes/metagenes of similar expression pattern it does not represent the multivariate structure of the data. Such aspects become highlighted by projecting the data to subspaces of lower dimension spanned by interesting modes such as the components of minimum mutual statistical dependence. ICA provides a visual plot in the space spanned by these independent components which are shown to point along the directions of maximum information content in the data or, equivalently, of non-normal distribution of the data [26]. We

applied ICA to single and metagene lists to see which of the alternative data sets offers the better separation among the various tissue groups.
The ICA-plots of the two leading independent components shown in Figure 8c illustrate the degree of similarity between the samples in dependence on the selected filters. All filters except one provide virtually three clusters, namely that of nervous (grey circles), immune system (blue) and the remaining tissues. The FC-100 single gene filter merges the latter two clusters due to its small representativeness with respect to non-nervous tissues (see also the respective heatmap in Figure 8b). Note also that the relative dimension of the three clusters in the ICA-plot and thus also their intrinsic resolution changes from filter to filter. These trends reflect the subtle interplay between the length of the list and its representativeness and/or noisiness which might overweight one tissue category and underweight another one. For example, the specifics of epithelium tissues (cyan circles) become relatively well resolved using the FC-100 metagene or, alternatively, the FC-1000 single gene lists. The respective heatmaps in Figure 8b confirm that this tissue category is well represented by a reasonable number of specifically over- and underexpressed genes/metagenes in these lists. The fraction of these genes however clearly decreases in the other filtering lists giving rise to the suboptimal resolution of the cluster of cyan circles in the ICA plots. The right part of Figure 8c compares the relative size of three clusters in terms of the fraction of the covered coordinate region. The metagene-based clusters are less dependent on the chosen length of the list and more balanced especially for short lists.
The ICA plots in Figure 8c reveal another interesting property inherent in the expression profiles: The points especially of nervous (grey) and immune systems (blue) but also of epithelium (light blue) tissues form chain-like clusters which point roughly along the coordinate axes. This pattern reflects the fact that the transcriptional activity of nervous tissues on one hand side and immune system and epithelium tissues on the other hand side are governed by different and mutually independent groups of genes. We will discuss this point below more in detail in the context of the SOM mosaics. In the context of the filter lists it should be noticed that this property of the data gets partly lost after most stringent single gene filtering (FC-100) whereas essentially all metagene lists well reflect the independence of the expression pattern of the different tissue categories.
In summary, ICA analysis illustrates the robustness and the discrimination power inherent in the metagene lists. The use of metagenes allows compressing the length of the list by about one order of magnitude without loss of information. The filtering conditions govern the resolution between different tissue categories in the ICA plot in a subtle way. Short and intermediate metagene lists provide best results in this respect. Notably, consideration of the full metagene information without filtering (FC-3600) provides still quite reasonable resolved clusters in the ICA-plot. In conclusion, metagenes are more robust with respect to the quality of secondary analysis than single gene lists owing to their better representativeness. Hence, the reduction of dimensionality provided by SOM analysis improves the performance of downstream hierarchical clustering and ICA analysis. The number of considered features can be reduced by about one order of magnitude without loss of information if one uses metagenes instead of real genes. Clustering and ICA characteristics obtained for the metagene and single gene lists after variance and FDR filtering virtually agree with the results of FC-filtering (see Additional file 4).

2.7. Metagene- and single gene-based correlation analyses

In the next step we calculated pairwise correlation maps (PCM) illustrating Pearson correlation coefficients for all mutual combinations between the tissues. The PCM-heatmaps shown in Figure 9a are obtained using the FC-1000 (single genes, left part) and FC-100 (metagenes, right part) filters representing both roughly the same number of genes (see discussion above). The metagenes clearly provide PCM-patterns of higher contrast which becomes emergent as diagonal and off-diagonal dark red/maroon and blue clusters. They refer to tissue pairings with highly correlated and anti-correlated expression profiles, respectively. Both, the single gene and the metagene PCM reveal essentially four groups of tissues which consist mainly of nervous (see the grey bar at the margins), immune system (blue bar), muscle (green bar) tissues and also of a mix of diverse tissue categories.
The expression profiles of nervous tissues strongly anti-correlate with essentially all the other tissue categories, i.e. a gene overexpressed in nervous tissues usually becomes underexpressed in non-nervous tissues and vice versa. The original expression SOM always reflect this property showing one

characteristic overexpression spot in the left upper corner (see spot A in Figure 3 and Table 1) and otherwise a blue and light blue background due to underexpressed genes/metagenes (Figure 1). Muscle tissues show strong off-diagonal correlation with the group of diverse tissues but not with the immune system tissue group. This property can be mainly attributed to spot D in the right upper corner in the SOM of these tissues whereas the diagonal correlation component mainly originates from the muscle-specific spot B (see Figure 3 and Table 1). The cluster of immune system tissues along the diagonal of the PCM can be attributed to spot F in their SOM. Hence, the diagonal and off-diagonal clusters in the metagene PCM can be related to different spots in the original expression SOM of the different tissue categories.

To get further insights into the origin of the contrast differences between the single gene and metagene PCM we calculated frequency distributions of the pairwise correlation coefficients either between tissues of one category or between tissues of different categories (Figure 9b). Intra-category correlation coefficients are expected to be close to unity because samples of the same categories show usually similar expression profiles. Indeed, these metagene correlation coefficients are close to unity as expected whereas the respective single gene correlations show a markedly broader distribution resulting in smaller correlation values on the average. Inter-category pairings of single genes show a broad distribution centered about zero with a strong component of anti-correlation near -0.5 revealing that single genes of different tissue types are either not or anti-correlated. The metagenes produce a more resolved trimodal distribution with strong components of correlated, anti-correlated and uncorrelated metagenes near 1.0, -0.7 and 0.0, respectively. The component peaks are clearly sharper and the whole distribution covers a wider range of correlation values. Hence, the metagenes obviously enable the better resolution of different subcomponents produced by different tissue types.

The PCMs reveal that anti-correlated metagene expression profiles are especially found between nervous tissues and the other tissues. We therefore calculated a second set of frequency distributions restricting the intra-tissue correlations to nervous tissues only and the inter-tissue correlations to that between nervous and all the other tissues (Figure 9c). The latter histograms reveal that the degree of anti-correlation is much stronger for the metagenes than for the single-genes again showing that metagenes more sharply express the correlation pattern of gene expression. Note that this anti-correlation is evident already in the textures of the original tissue SOM: Large blue areas in the SOM of nervous tissues reveal under-expression of the respective metagenes which become selectively overexpressed in the SOM of other, non-nervous tissues (Figure 1). The inter-nervous tissue correlation histogram also shows a strong correlation peak near unity which is caused by the metagenes commonly overexpressed in nervous tissues and pituatary gland (endocrine tissue, no. 5) as discussed above.

In summary, our extended dataset of human tissues confirms the results of Guo et al. [1] who found that SOM based metagenes well recapitulate gene expression profiles of the entire gene dataset despite dimension reduction and that the visual patterns capture the real similarity relationships among samples with a high fidelity. Moreover, using metagenes instead of real genes one can improve the resolution power of popular standard analyses based on two-way hierarchical clustering or pairwise correlation heatmaps. The SOM metagene pattern serves as an adequate data filter which appropriately selects representative features characterizing the expression properties of the system studied.

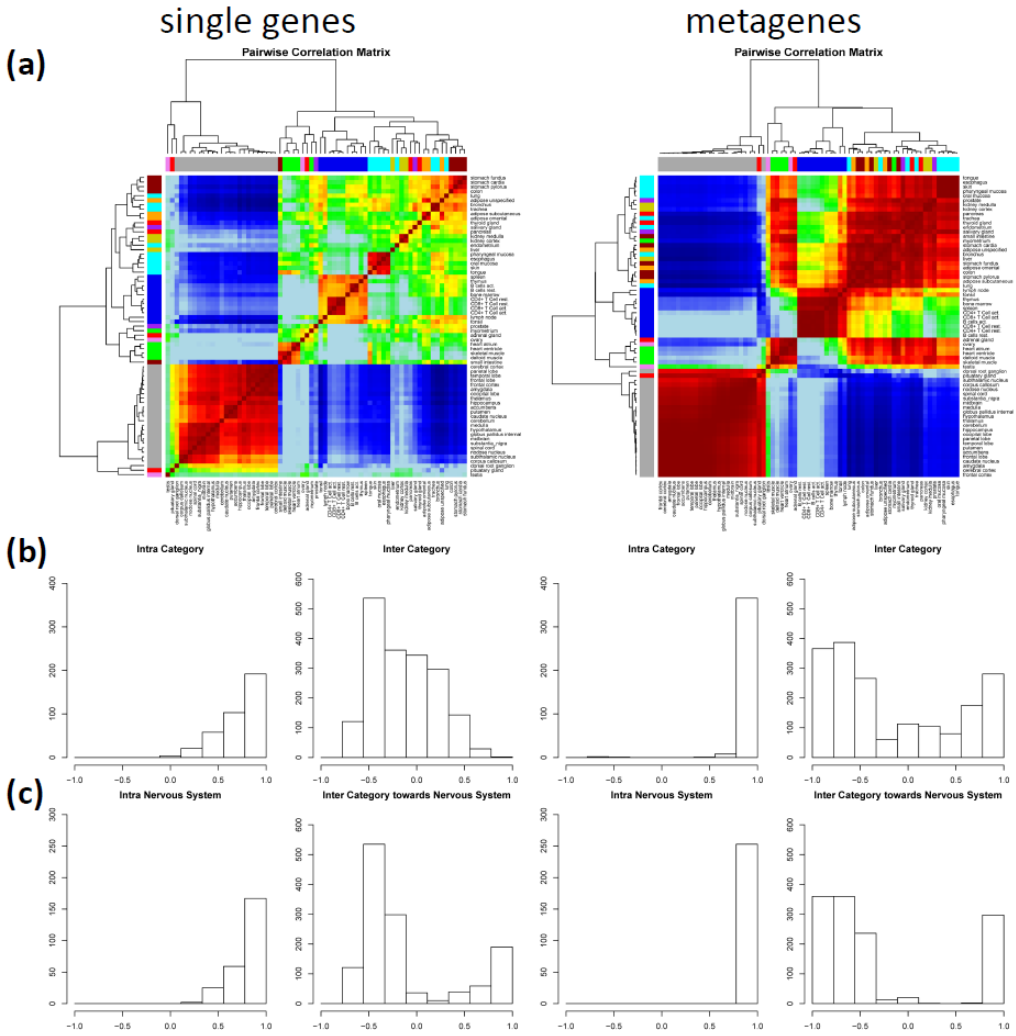

Figure 9: Single gene (left panels) and metagene (right panels) correlation analysis of human tissues using the 1000/100 most strongly regulated genes/metagenes: (a) Pairwise Correlation Matrix (PCM); (b) Frequency distributions of correlation coefficients for all intra- and inter-tissue category pairings and (c) for pairings of intra-nervous tissue pairings and for pairings between nervous and all other tissues. Note that the metagenes produce the stronger contrast of the PCM clusters due to the sharper and better resolved distributions.

2.8. Sample cartography: Second level SOM

Guo et al. proposed an alternative second-level SOM analysis step [1]. It maps all samples together into one two-dimensional mosaic pattern to visualize the degree of similarity between their expression profiles. The second-level SOM algorithm uses the metagene expression of each sample as input. After training, each tile of the mosaic is characterized by the expression profile of one 'metasample' which serves as the condensation nucleus of the associated minicluster of real samples possessing similar SOM pattern. The mutual distances between the samples in the map are related to the degree of similarity of their SOM expression pattern. Typically, second level SOM use a resolution where the number of mosaic tiles exceed the number of samples. In consequence most tiles remain empty. Figure 10 shows the second level SOM of all 67 human tissues studied using a 9x9 grid: Each tissue is represented by its tissue number and the color of its previously assigned tissue category (see the circles and also Figure 1). In addition, representative first-level SOMs are shown in each of the not-empty tiles representing the respective metasample.

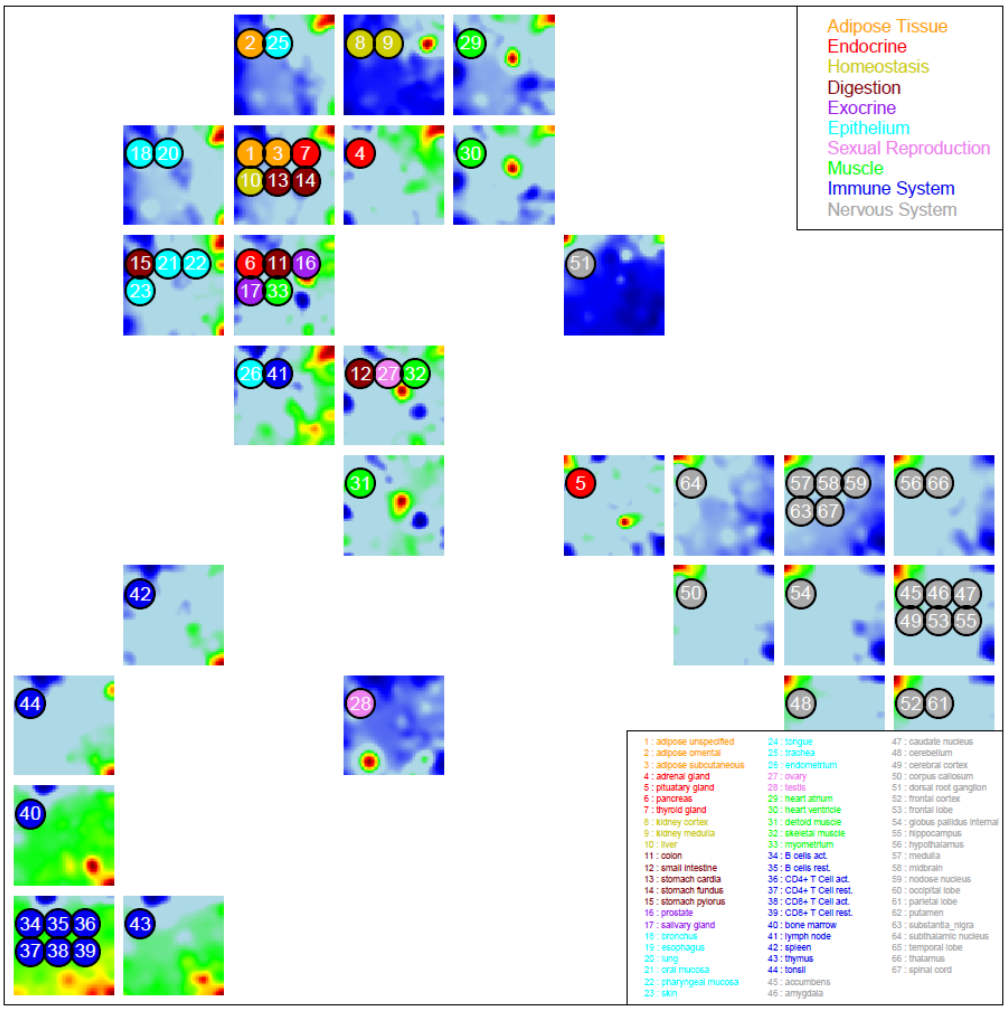

Figure 10: Second level SOM of the metagene expression profiles of all 67 samples: Each tissue is color-coded by the circles according to its tissue category and assigned by its number. The small mosaics show the relevant first level SOM pattern of the not-empty metasamples which might be occupied by up to six real samples.

Essentially one distinguishes the same three main clusters which were detected in the ICA-plots shown in Figure 8c; namely that of nervous tissues (grey), immune system tissues (blue) and the remaining ones. In general, ICA and 2nd level SOM provide a similar view on the samples, however with subtle differences. For example, the ICA algorithm distributes the sample points continuously in the coordinate system spanned by the two leading principal components of maximum information content. The mutual separation between the points linearly scales with their distance in units of these components. In contrast, SOM machine learning non-linearly distributes the positions of the sample points in the discrete space defined by the mosaic grid of metasamples which enables to display differences between the samples with improved resolution. In consequence, the individual tissues spread over a larger area in the SOM mosaics than in the respective ICA.

As noticed above, most of the samples group into linear clusters which orient along one of the coordinate axes in the two dimensional ICA plots. The orthogonal orientation of most of these clusters indicates that each of them is characterized by genes which vary mutually independently. In the original ICA this property applies to nervous tissues on one hand and all other tissues on the other hand. Nervous tissues are characterized by their specific spot A (see Figure 3a) containing genes which vary virtually independent of that of the other tissue spots. We also generated three dimensional ICA-plot to assess the third main component of variability (see Additional file 3). This plot reveals that the characteristic pattern of orthogonal linear clusters of selected tissue categories extends into the third dimension (see, e.g. the clusters of nervous system, immune system and epithelium tissues in the 3D-ICA of all tissues). Hence, the metagene-based ICA plots in two and three dimensions allow to disentangle tissue categories of virtually independent expression profiles. The responsible groups of

genes can be identified using the spot pattern of the original SOM where they typically aggregate into metagene spots specifically overexpressed in the respective tissue category.
The second-level SOM similarly, but not identically arrange the samples as discussed above. The non-linear scaling in the SOM partly disturbs the arrangement of samples according to the mutual independence of their expression profiles as in the ICA-plots. For example, the linear ICA-clusters of the nervous tissues (grey circles) transform into slightly more compact clusters in the 2nd level SOM. Hence, although very similar, 2nd level SOM and ICA visualize partly complementary aspects of the data which can be studied more in detail using the spot-texture of the individual SOM of the samples studied.

3. Summary and Conclusions

The microarray expression data of 67 human tissues was used as an illustrative example to demonstrate the strengths of the SOM method in disentangling large sets of heterogeneous data. After suited preprocessing and training, the SOM method decomposes the original data into metagene expression profiles representing clusters of correlated single genes. Metagene expression values in the individual samples provide mosaic pictures visualizing tissue-specific over- and underexpression in terms of characteristic color-coded textures. They enable the direct comparison of the expression of individual samples in a simple and intuitive way.
Particularly, the tissue-specific patterns of gene expression were readily discernable in the obtained gallery of individual tissue maps. They reveal a series of about one handful stable over- and underexpression spots which selectively characterize different tissue categories such as nervous, immune system, muscle, exocrine, epithelial or adipose tissues. Single tissues of mixed characteristics such as tongue (composed of expressions spots found in muscle or epithelial tissues) can be easily identified. Also anti-correlated expression spots are detected which, for example, are overexpressed in nervous tissues but underexpressed in the other tissues and vice versa.
To extract the functional context of spot and metagene related lists of single genes we applied overrepresentation analysis with respect to pre-defined gene sets of basically known functional impact. The mapping of overrepresentation of a selected gene set into the SOM mosaic provides a ‘functional’ map showing areas which are potentially relevant for this function. Application to the SOM atlas of human tissues shows that the tissue related spots typically contain enriched populations of function-related gene sets well corresponding to molecular processes in the respective tissues. The representative expression profiles of the leading metagenes of the spots well agree with the expression profiles of leading functionally related gene sets. This result strongly supports the ‘guilt-by-association’ principle that coexpressed genes are likely to be functionally associated. It, in turn, implies the ability to define either new gene sets using selected SOM spots or to verify and/or to amend existing ones.
The SOM method compresses the original set of high-dimensional data in two consecutive steps: Firstly, similar expression profiles of single genes are collected into metagene clusters, which reduces the number of relevant features nearly by one order of magnitude in our application. These metagene profiles can be understood as a sort of ‘eigen-modes’ characterizing the multitude of expression pattern inherent in the data. Secondly, the textures of the obtained SOM are decomposed into a few (typically less than one dozen) spots of similarly (over- or under-) expressed metagenes. This ‘double compression’ sequentially applies global (similar profiles) and local (over-/underexpression in part of the samples) criteria.
The use of metagene instead of single gene expression reduces the dimension of the data and leads to an increased discriminating power in downstream agglomerative analysis such as hierarchical clustering and independent component analysis owing to essentially two facts: Firstly, the set of metagenes better represents the diversity of expression pattern inherent in the data and secondly, it also possesses the better signal-to-noise characteristics as a comparable collection of single genes. Due to the better representativeness, metagene lists are less sensitive to downstream filtering than lists of single genes. Metagenes can be seen as a natural choice to detect context-dependent patterns of gene expression in complex data sets.
Our example shows that SOM cartography transforms large and heterogeneous sets of expression data into an atlas of sample-specific texture maps which can be directly compared in terms of similarities

and dissimilarities. This global view on the behavior of defined modules of correlated and differentially expressed genes is more intuitive than ranked lists of hundreds or thousands of individual genes. Importantly, the dimension reduction of the data does not entail the loss of primary information in contrast to simple filtering approaches which irretrievably removes part of the data. Instead, the reduction of dimension is attained by the re-weighting of primary information in the aggregation step. The whole set of single gene expression profiles remains virtually 'hidden' behind the metagenes. This primary information together with the respective gene annotations can be extracted in later steps of analysis to interpret the observed SOM textures using concepts of molecular biological function.

4. Data and Methods

4.1. Microarray Data

Microarray raw intensity data (*.cel files, Affymetrix HG-U133 plus 2 array) of M=67 tissues each measured in R_m=1, 2… (m=1…M) replicates were downloaded from the Gene Expression Omnibus repository as the 'human body index - transcriptional profiling' - data set (http://www.ncbi.nlm.nih.gov/geo, GEO accession no. GSE7307; see Additional file 1 for the detailed list of samples used). The used HG-U133p2 array essentially merges the probes printed on two previous GeneChip arrays (HG-U133A and B). In our analyses we masked the probes referring to the HG-U133 B array to assure comparability with array studies which use the popular HG-U133A array.

4.2. Preprocessing of microarray intensities

We consider a data set consisting of the expression levels of N genes in M different sample categories such as different tissues each measured in R_m (m=1…M) replicates. For gene expression studies the number of genes N is typically in the ten thousands, the number M of experimental conditions is typically in the tens to hundreds and the number of replicates between one and ten.

The used GeneChip microarrays provide typically eleven raw probe intensities per gene constituting one probe set. Raw probe intensity values of each of the MxR_m chips are calibrated and summarized into one expression value E per probe set using the hook method [27, 28]. The expression values from all arrays are subsequently quantile-normalized [29] (see Figure 11a for illustration).

The obtained distribution of expression values typically shows a bimodal shape: It's left and right peaks at smaller and larger expression values were attributed to non-specific and specific hybridization, respectively [30]. The peak due to non-specific hybridization is non-informative with respect to the target genes which are therefore called 'absent' because their expression is smaller than the detection threshold of the method. The non-specific peak consequently characterizes the 'chemical' background of the measurement.

The distribution of expression data of each experimental series is then processed as follows: Firstly, the origin of the log-expression axis at log E=0 was positioned to agree with the peak position of the non-specific peak of the distribution. Secondly, both peaks are decomposed as described previously [30] assuming mirror symmetry of the left and right flanks of the non-specific peak (Figure 11b). Thirdly, we make use of the decomposed distributions to estimates the probability that the specific expression of a selected gene is detected. This 'present-call'-parameter is set to pc=0 and pc=1 for genes with expression values outside the region of overlap of both peaks (see Figure 11c). The present call is calculated as the fraction of the local density of the specific signal contributing to the total signal distribution in the range of overlap. The resulting value of pc roughly linearly scales between zero and one with increasing expression in this range (Figure 11c). Fourth, the log-expression of each gene is scaled with its present call, i.e., e= pc(e')*e' where lower case e define the logged expression values, e'= log E. The used transformation thus considerably narrows the non-specific peak at position e'=0 of the expression axis while leaving the specific signal virtually unaffected. As a consequence, the variability of the signals of absent called and thus of non-informative probes is markedly reduced (Figure 11c). This transformation enables to conserve the full set of available genes in the data set used for SOM analysis in contrast to data filtering which removes presumably uninformative probes from the data set prior to downstream analysis.

Expression values of replicates of the same tissue were log-averaged and finally, the logged expression values of each gene were transformed into differential expression values relative to the mean expression of the particular gene in the experimental series of tissues considered (Figure 11d),

$$\Delta e = e - \langle e \rangle_{all\_tissues} \quad . \qquad (1)$$

Eq. (1) thus defines differential expression in units of the logged fold change, logFC≡ Δe.

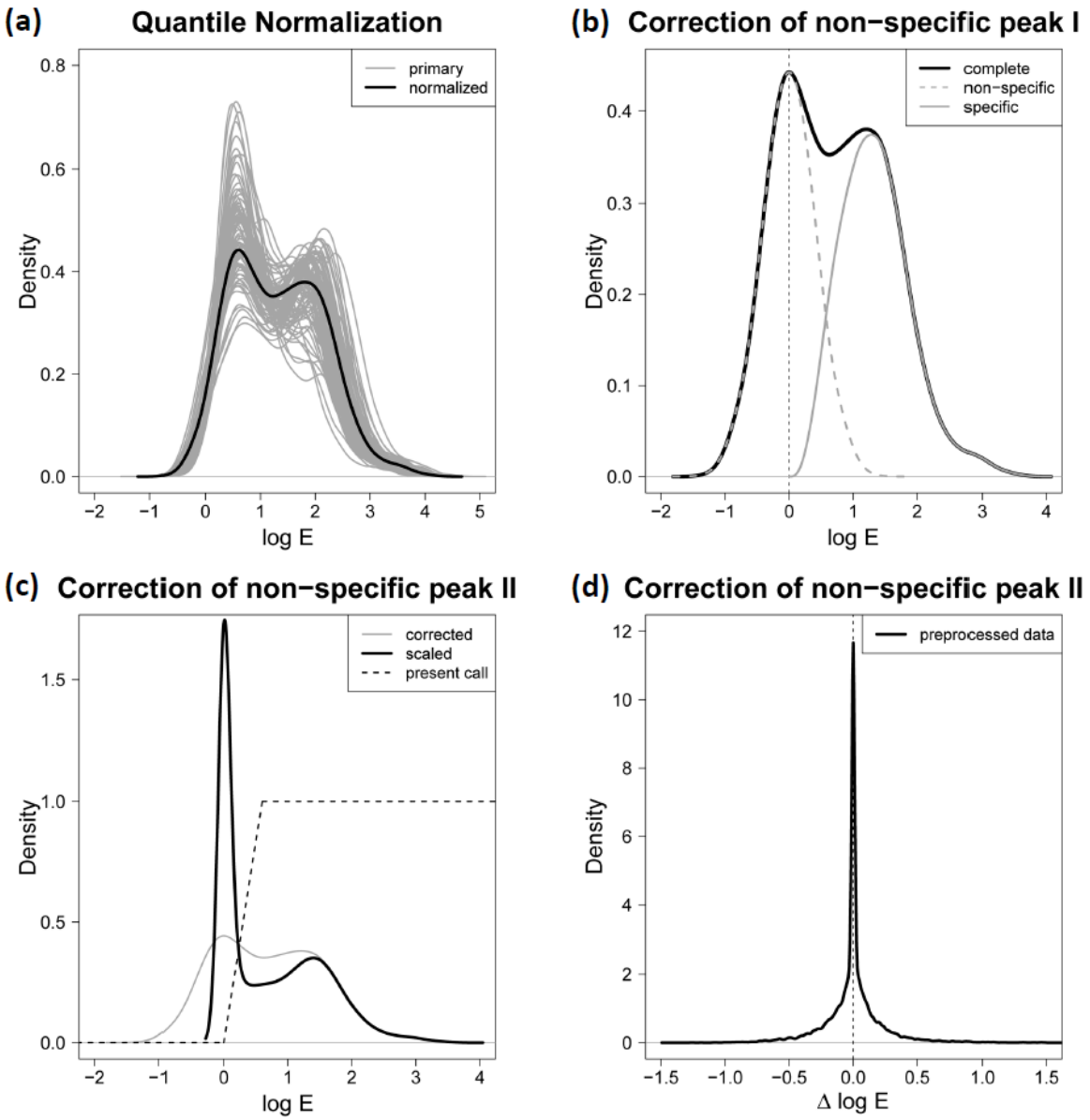

Figure 11: Normalization and adjustment of expression values: The different distributions of hook-calibrated expression values of the samples studied merge into one representative mean distribution after quantile normalization (panel a). Its double peaked shape is decomposed into two single peaked distributions due to non-specific and specific hybridizations at small and larger expression values, respectively (b). The fraction of the specific signal contributing to the total signal density (dashed curve) is used as weighting coefficient of the expression values, e= pc(e')*e', which reshapes the total signal density (c). Finally, the expression values are normalized with respect to the logged mean expression of each gene (d). The large central peak refers to invariant genes under all conditions studied.

4.3. SOM-mapping of gene expression profiles

In the next step, the preprocessed differential expression values of the series of tissue samples, Δe, are feed into the unsupervised machine learning program to train a self organizing map (SOM) representing information-rich diagrams as illustrated in Figure 12. The SOM method applies a neural network algorithm to project high dimensional data onto a two-dimensional visualization space [2, 31]. SOMs have a strong visualization capability by presenting each individual sample as an entity allowing, for example, its identification in a series of samples. At the same time each SOM still keeps high-resolution information about the co-expression pattern of the genes in the samples studied.

Particularly, we apply a home-made R-program which uses the CRAN package 'som'. Our program was designed referring to the Gene Expression Dynamics Inspector (GEDI) (www.chip.org/~ge/gedihome.html), a freeware MatLab-program, which translates high-dimensional data into a two-dimensional mosaic pattern (Eichler et al. [8]). This SOM-algorithm assigns the expression profiles of the N input genes measured under M conditions to a number of K < N rectangular 'tiles' (so-called SOM nodes), each of which is characterized by one representative profile of metagene expression given by a vector of length M, $\Delta e_k^{meta}=(\Delta e_{k,1}^{meta}, \Delta e_{k,2}^{meta},\ldots, \Delta e_{k,M}^{meta})$ (k=1…K). It is trained such that the profiles of the metagenes capture the range of all individual expression pattern observed. Each individual expression profile of a 'real' gene is assigned to the metagene pattern of closest similarity using the minimum Euclidian distance as criterion. Each metagene thus serves as a sort of condensation nucleus for a minicluster of n_k 'real' genes with similar expression profiles, $\Delta e_{k,i}=(\Delta e_{k,1,i}, \Delta e_{k,2,i},\ldots, \Delta e_{k,M,i})$, with i=1…$n_k$ and $N = \sum_{k=1\ldots K} n_k$.

The metagenes are arranged in a two-dimensional x*y grid with K=x·y where most similar expression profiles of metagenes are located adjacent each to another. The correlation between metagene expression decreases with the mutual distance between the tiles on the mosaic. The degree of similarity between adjacent metagenes depends on the number of genes assigned to the respective metagenes being closer for larger populated metagenes and vice versa. For each measuring condition m=1…M a SOM mosaic pattern is constructed by color-coding the tiles k=1…K according to its metagene expression, $\Delta e_{k,m}^{meta}$. This way one obtains a coherent mosaic pattern that is characteristic for each sample owing to the similarity of adjacent metagenes. Since the SOMs assign the same metagene to the same tile in all samples, they can be directly compared to each other allowing immediate identification of biologically interesting groups of genes.

The expression profiles of individual genes and of the respective metagenes are shown in Figure 12 for a segment of 5x3 tiles. Typically, the number of tiles to 'pixelate' the expression profiles is considerably larger, K=10x10 – 100x100= 10^2 – 10^4 with, on the average, n_k=5 – 100 genes per metagene. The obtained mosaic pattern is usually more homogeneous than typical gene clustering heatmaps containing typically about 10^2 clusters. This finer granularity of SOM-maps is associated with a fewer number of genes per unit (cluster/metagene) which in consequence gives rise to a more detailed expression pattern.

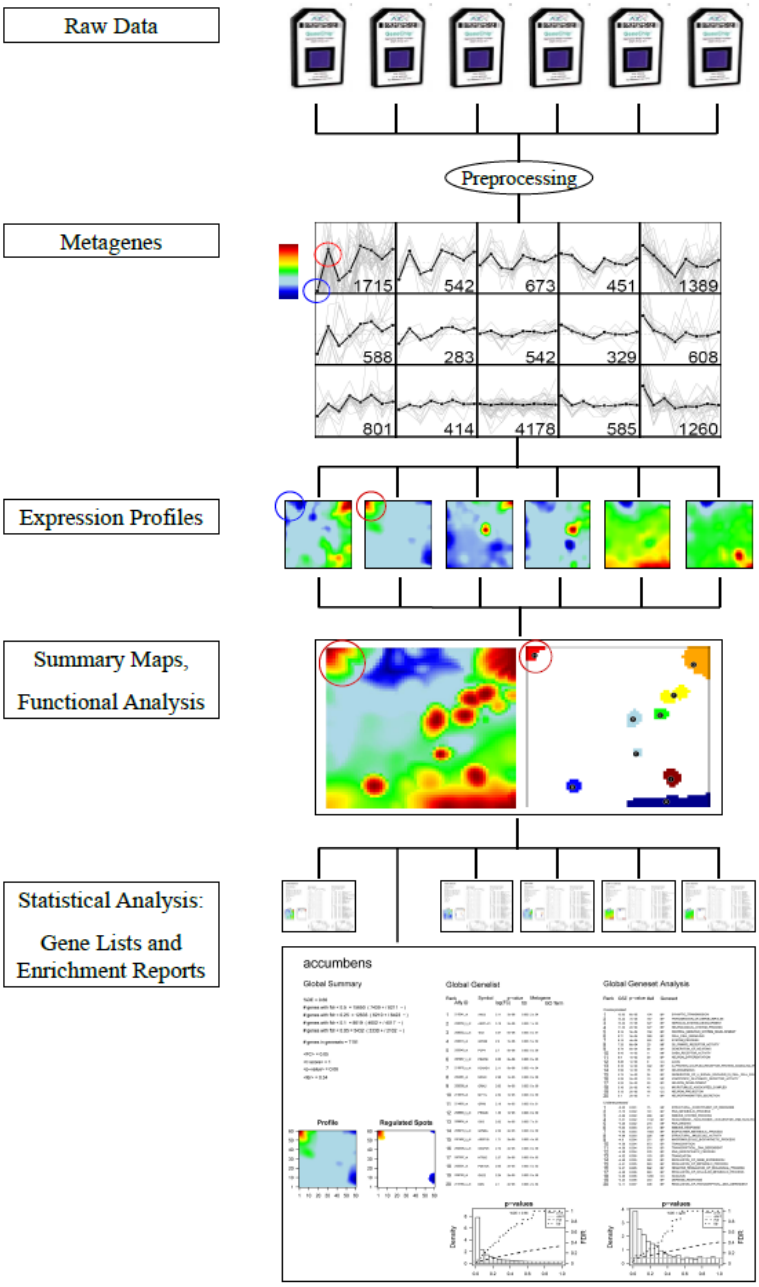

Figure 12: Expression profiling using self organizing maps (SOMs): Raw microarray probe intensity data referring to an experimental series of different conditions was preprocessed including calibration, normalization and adjustment. The obtained expression values are then feed into the SOM-algorithm. It translates the high-dimensional expression data into a two-dimensional grid of expression profiles. Each tile represents a cluster of individual genes (thin lines in the graphs, numbers of genes are given for each cluster) characterized by the expression profile of a representative metagene (thick lines). The expression profile is given by the expression values in the samples studied. The expression profiles of the metagenes are then transformed into one mosaic picture per condition which are shown in the row „expression profiles“ below. The tiles in these maps are color-coded to represent overexpression or underexpression of each metagene in the respective sample to map the underlying gene expression pattern. The parallel evaluation of multiple samples allows linking 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 respective color-code in the respective pictures. Summary maps characterize different aspects of the individual SOM such as the population of metagenes or the summary of all overexpression peaks. Metagene expression can further be used for statistical and functional analysis as will be described elsewhere. In the last step, summary reports for each sample are generated providing lists of differentially expressed genes, enriched gene sets, error statistics and further information (see http://som.izbi.uni-leipzig.de).

4.4. Supporting maps

We define the following supporting maps which provide additional information about the miniclusters defined by each metagene and the associated real genes:
(i) The metagene expression profiling map uses a coarse grained mosaic to provide an overview of the courses of the metagene profiles. For visualization purposes we use a coarse grained (e.g., 8x8) mosaic with considerably less tiles than the mosaic grid applied for the SOMs (60x60). The metagene profiles might be plotted together with the associated single gene profiles.

(ii) The population map plots the number of real genes per metagene in logarithmic scale, log n_k.

(iii) The variance map illustrates the variability of the expression profile of each metagene in the samples studied,

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

(iv) The integral over-/under-expression summary maps collect all over-/underexpression spots observed in the individual sample SOMs into one master map.

4.5. Gene set overrepresentation analysis

Gene set analysis requires the knowledge of predefined gene sets to study their enrichment in gene lists which are obtained from independent differential expression analysis (see [32] for a critical review and references cited therein). A large and diverse collection of such sets can be downloaded from the 'gene-set-enrichment-analysis'-website (http://www.broadinstitute.org/gsea). Particularly, we included in total 1454 gene sets in our analysis according to the GO terms 'biological process' (825 sets), 'molecular function' (396 sets) and 'cellular component' (233 sets). We use the term 'overrepresentation' to assign the probability to find members of a given set in a list compared with their random appearance independent of the values of their expression scores. We use the hypergeometric distribution to characterize overrepresentation in terms of a p-value which estimates the probability to find a stronger overlap between the list and the set by chance than actually detected [33, 34].

4.6. Grouping samples: Second level SOM cartography

We applied second-level SOM analysis as proposed by Guo et al. [1] to visualize the similarity relations between the individual SOM-metagene expression pattern. Second-level SOM analysis uses the K metagene expression profiles of the M samples as input and then cluster the samples and not the genes as in first-level SOM analysis. Each tile of the second-level SOM mosaic characterizes the expression profile of a representative metasample defined by K metagene expression values. The M samples were presented using a mosaic grid of size $K_{2SOM}>M$. Note that the number of metasamples usually exceeds the number of real samples whereas in first order SOM the number of metagenes is usually much smaller than the number of real genes. A considerable fraction of tiles of the second order SOM are consequently empty with no sample assigned.

4.7. Estimating similarities: Clustering-, tree- and independent component-analysis

One- and two-way hierarchical clustering [25] and independent component analysis [35] were applied in two versions using either the profiles of the SOM-metagenes (metagene analysis) or the profiles of individual 'real' genes (single gene analysis) using the R-packages 'stats' and 'fastICA' for clustering and ICA, respectively. Hierarchical clustering uses Euclidian distances between the genes/metagenes as similarity measure, whereas ICA is based on covariance. In addition to two-way hierchical clustering heatmaps, we generate pairwise correlation maps (PCM) which visualize the Pearson correlation coefficients between the gene expression profiles (metagenes or 'real' genes) in all pairwise combinations of samples.

4.8. Filtering genes and metagenes

Optionally, the number of real genes and/or metagenes used in the analyses is reduced by applying three types of filters to exclude genes/metagenes of weak or of virtually invariant differential expression from downstream analysis: (i) FC-filtering: the genes/metagenes are ranked with decreasing absolute value of the fold change (FC) for each sample and a certain number (e.g., 100, 1000 and 3600) of the top-most features is selected; (ii) Variance filtering: the genes/metagenes are ranked with decreasing variance of their expression profiles and a certain number of top-most features is selected; (iii) FDR-filtering: only genes/metagenes with a local false discovery rate (FDR) smaller than a certain threshold (0.005, 0.01, 0.05) were selected. The local FDR estimates the probability of false positives in a list genes/metagenes. We used a shrinkage t-score statistics to assign p-values to each single gene the distribution of which then provides its FDR-values. The FDR of the metagenes is simply calculated as log-average of the single gene FDR of the respective metagene cluster. Details of the method are given elsewhere.

4.9. Filtering benchmarks

The performance of metagene and single gene filters was compared using the following benchmarks (see also Figure 8):
Hierarchical clustering: The ratio of the inter-class and intra-class variance of the Euclidian distances between the respective expression data (F-score) was used to estimate the quality of the clusters.
Two-way hierarchical clustering: The percentage of genes/metagenes attributed to tissue-specific clusters for three tissue categories (nervous, immune systems and epithelium) was used to estimate the representativeness of the list.
ICA: The percentage of the variance of the independent components IC1 and IC2 of one tissue category, $\%= (varIC1+varIC2)_{one\_category}/(varIC1+varIC2)_{three\_categories}$, was used to judge the relative size of the respective cluster.

Additional material

Additional file 1: Table of samples studied
Additional file 2: Whole set of 67 SOM expression profiles of human tissues
Additional file 3: The additional text addresses the filtering of metagenes/single genes and the interpretation of cluster trees. Further details of zooming-in of two tissue subgroups are given together with the 3D-ICA plots of the tissues studied.
Additional file 4: Agglomerative cluster analyses after single gene and metagene filtering using FDR and variance criteria

The complete set of results of our SOM analysis of the human tissue dataset can be found on our website: http://som.izbi.uni-leipzig.de

Authors contributions

HW and HB: Conceived and designed this study, performed data analysis and wrote the manuscript. HW: Wrote the R-programs and performed the calculations. All authors read and approved the final manuscript.

Acknowledgements

The project LIFE is financially supported by the European Fonds for Regional Development (EFRE) and the State of Saxony (Ministry for Science and the Arts). HW was kindly supported by the Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE)[36].

5. References

1. Guo Y, Eichler GS, Feng Y, Ingber DE, Huang S: **Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers.** *Journal of biomedicine & biotechnology* 2006, **2006**:69141.

2. Kohonen T: **Self-organizing formation of topologically correct feature maps**. *Biological Cypernetics* 1982, **43**:59-69.

3. Tamayo P, Slonim D, Mesirov J, et al. **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**:2907-12.

4. Golub TR, Slonim DK, Tamayo P, et al. **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science (New York, N.Y.)* 1999, **286**:531-7.

5. Covell DG, Wallqvist A, Rabow AA, Thanki N: **Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data.** *Molecular cancer therapeutics* 2003, **2**:317-32.

6. Buckhaults P, Zhang Z, Chen Y-C, et al. **Identifying tumor origin using a gene expression-based classification map.** *Cancer research* 2003, **63**:4144-9.

7. Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O: **Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study.** *BMC bioinformatics* 2002, **3**:36.

8. Eichler GS, Huang S, Ingber DE: **Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles.** *Bioinformatics (Oxford, England)* 2003, **19**:2321-2.

9. Camphausen K, Purow B, Sproull M, et al. **Influence of in vivo growth on human glioma cell line gene expression: convergent profiles under orthotopic conditions.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:8287-92.

10. Huang S, Eichler G, Bar-Yam Y, Ingber DE: **Cell fates as high-dimensional attractor states of a complex gene regulatory network.** *Physical review letters* 2005, **94**:128701.

11. Mar JC, Quackenbush J: **Decomposition of gene expression state space trajectories.** *PLoS computational biology* 2009, **5**:e1000626.

12. Tsigelny IF, Kouznetsova VL, Sweeney DE, et al. **Analysis of metagene portraits reveals distinct transitions during kidney organogenesis.** *Science signaling* 2008, **1**:ra16.

13. Sieberts SK, Schadt EE: **Moving toward a system genetics view of disease.** *Mammalian genome : official journal of the International Mammalian Genome Society* 2007, **18**:389-401.

14. Wolfe CJ, Kohane IS, Butte AJ: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** *BMC bioinformatics* 2005, **6**:227.

15. Miozzi L, Piro RM, Rosa F, et al. **Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data.** *PloS one* 2008, **3**:e2439.

16. Läuter J, Horn F, Rosołowski M, Glimm E: **High-dimensional data analysis: selection of variables, data compression and graphics--application to gene expression.** *Biometrical journal. Biometrische Zeitschrift* 2009, **51**:235-51.

17. Brunet J-P, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:4164-9.

18. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data.** *Genome research* 2003, **13**:1706-18.

19. Binder H, Brücker J, Burden CJ: **Nonspecific hybridization scaling of microarray expression estimates: a physicochemical approach for chip-to-chip normalization.** *The journal of physical chemistry. B* 2009, **113**:2874-95.

20. Hsiao LL, Dangond F, Yoshida T, et al. **A compendium of gene expression in normal human tissues.** *Physiological genomics* 2001, **7**:97-104.

21. Shyamsundar R, Kim YH, Higgins JP, et al. **A DNA microarray survey of gene expression in normal human tissues.** *Genome biology* 2005, **6**:R22.

22. Levine DM, Haynor DR, Castle JC, et al. **Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways.** *Genome biology* 2006, **7**:R93.

23. Hornshøj H, Conley LN, Hedegaard J, et al. **Microarray expression profiles of 20.000 genes across 23 healthy porcine tissues.** *PloS one* 2007, **2**:e1203.

24. Eklund AC, Szallasi Z: **Correction of technical bias in clinical microarray data improves concordance with known biological information.** *Genome biology* 2008, **9**:R26.

25. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863-8.

26. Liebermeister W: **Linear modes of gene expression determined by independent component analysis.** *Bioinformatics (Oxford, England)* 2002, **18**:51-60.

27. Binder H, Krohn K, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures.** *Algorithms for molecular biology : AMB* 2008, **3**:11.

28. Binder H, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: theory and algorithm.** *Algorithms for molecular biology : AMB* 2008, **3**:12.

29. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics (Oxford, England)* 2003, **19**:185-93.

30. Binder H, Wirth H, Galle J: **Gene expression density profiles characterize modes of genomic regulation-theory and experiment.** *Journal of biotechnology* 2010.

31. Bishop C, Svensén M, Williams C: **GTM: The generative topographic mapping**. *Neural computation* 1998.

32. Goeman JJ, Bühlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics (Oxford, England)* 2007, **23**:980-7.

33. Vêncio RZN, Shmulevich I: **ProbCD: enrichment analysis accounting for categorization uncertainty.** *BMC bioinformatics* 2007, **8**:383.

34. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic acids research* 2005, **33**:W741-8.

35. Hyvärinen A, Oja E: **Independent component analysis: algorithms and applications.** *Neural networks : the official journal of the International Neural Network Society* 2000, **13**:411-30.

36. Bissinger V, Kolditz O: **Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE)**. *GAIA-Ecological Perspectives for Science* 2008.

Supplementary text

Expression cartography of human tissues using self organizing maps

Henry Wirth[1,2]*, Markus Löffler[1,3,4], Martin von Bergen[2,5], Hans Binder[1,4]*

[1] Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Härtelstr. 16-18
[2] Helmholtz Centre for Environmental Research, Department of Proteomics, D-04318 Leipzig, Permoserstr. 15, Germany
[3] Institute for Medical Informatics, Statistics and Epidemiology, Universität Leipzig, D-4107 Leipzig, Härtelstr. 16-18
[4] Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment (LIFE); Universität Leipzig, D-4103 Leipzig, Philipp-Rosenthalstr. 27, Germany
[5] Helmholtz Centre for Environmental Research, Department of Metabolomics, D-04318 Leipzig, Permoserstr. 15, Germany

* to whom correspondence should be addressed

1. Filtering metagenes and single genes

We analyzed the tissue data set using three types of filters to reduce the number of single genes and metagenes, namely FC-expression, variance and significance (FDR-) filtering (see Supplementary Table 1 and the methodical section). In the first case of expression filtering the full set of absolute differential expression values of all genes (real genes and metagenes) under all measured condition are ranked and a certain number of topmost genes in the list is considered for further analysis. In variance filtering the ranked list is generated using the variance of the expression profiles of the genes. These filterings improve the sensitivity of downstream discriminant analyses because they remove non- and less-informative weakly expressed, 'noisy' and/or virtually invariant genes from the data set. In the case of significance filtering the false discovery rate (FDR) of the features (metagenes or real genes) is used as filter criterion. Details of the statistical analysis of differential expression using SOM clustering will be presented elsewhere.

Supplementary Table 1 shows that the filter criteria when applied to the metagenes gives rise to a gene-to-metagene ratios between G/M= 3 and 19 where more stringent filters increase the G/M-values. For example, the 100 selected metagenes (FC-filtering) are representative for 1,487 single genes (G/M=14.9). In turn, selection of real genes roughly maintains this relation: The filtered 100 'real' genes distribute over 8 metagenes only (G/M=12.5) which are all enclosed in the 100 members of the metagene list. Hence, both subsets of metagenes after metagene and single gene filtering completely intersect each other reflecting the high degree of correlation between the metagenes and the associated 'real' genes. Figure S 1 shows the areas in the SOM mosaics covered by the filtered features and their mutual overlap after metagene and single gene filtering in terms of Venn diagrams. The left/right part of the figure highlights the selected metagenes/genes per tile of the SOM-mosaic. For example, the FC-3600 filter selects 100% of the metagenes but only 16% of the real genes. These genes accumulate essentially in the same areas of the SOM-mosaic as the metagenes, however whenselected using the more stringent FC-1000 filter, which selects 28% of the metagenes only. The FC-1000 single gene filter, in turn, delivers genes which preferentially accumulate in the metagenes which are mostly selected by the more stringent FC-100 metagene filter (compare the right mosaics in Figure S 1 with the left one in the respective row below).

Hence, equal numbers of 'real' genes and of metagenes selected by the respective filters reflect effectively different sample sizes owing to the G/M-compression which reduces the length of the metagene list. It integrates the properties of roughly a tenfold larger list of 'real' genes and vice versa in our particular SOM settings.

With increasing stringency of filtering, whole spot areas and thus also the respective expression profiles are progressively excluded from the list of filtered features. For example, the most stringent FC-100 metagene filter excludes a few areas selected by the FC-1000 single gene filtering thus revealing a decreased representativeness. Variance-filtering essentially provides similar relations between metagenes and real genes as FC-filtering (see Supplementary Table 1).

As a third option, we applied filtering using equal significance levels estimated in terms of the false discovery rate (FDR) to adjust the sample size of gene and metagene filters. The FDR-value defines the probability that each of the selected features is a differentially 'null' and thus a false positive one [1]. It applies to single genes and to the metagenes as well. In the latter filtering the FDR-threshold applies to the mean FDR of the single genes associated with each metagene.

The number of selected real genes after FDR-filtering is similar if the filter is applied to real or to metagenes (for example, 670 versus 387 for FDR<0.2; Figure S 2 and Supplementary Table 1) in contrast to the previously applied FC- and variance filters. Note however that the single genes after real gene filtering spread over a much larger number of metagenes than the metagenes which are directly selected after filtering metagenes (116 versus 14 for FDR<0.2, Figure S 2). This difference simply reflects the fact that genes selected by the single gene filter might be associated with metagenes which are not selected by the metagene filter as illustrated by the mosaics shown in Figure S 2. Hence, the FDR-filter if applied to single genes provides a similar numbers of real genes compared with the respective metagene filter. These genes however spread over a markedly larger number of metagenes and suggest an increased representativeness. In other words, significance filtering is roughly symmetric with respect to sample size but asymmetric with respect to 'representativeness' of the

selected features. Figure S 3 illustrates the consequences of shifted FDR-significance criteria which increases the number and representativeness of the features selected by metagene filters.

Supplementary Table 1: Filtering metagenes and real genes

filter		applied to metagenes			applied to real genes		
	threshold	#metagenes	#real genes	G/M [a]	#metagenes	#real genes	G/M* [a]
fold change	100	100	1,487	14.9	8 (8/0)[c]	100 (100/0)[c]	12.5
(FC)[b]	1,000	1,000	7,770	7.8	127 (127/0)	1,000 (1,000/0)	7.9
	3,600	3,600	22,277	6.2	600 (600/0)	3,600 (3,600/0)	6.0
variance	100	100	1,889	18.9	20 (19/1)	100 (97/3)	5,0
(Var)[d]	1,000	1,000	9,924	9.9	126 (124/2)	1,000 (995/5)	7,9
false	0.2	14	387	27.6	116 (14/102)	670 (317/353)	5.7
discovery rate	0.4	666	6,576	9.8	1,390 (666/724)	7,088 (5,587/1,501)	5.1
(fdr)[e]	0.5	1,751	13,692	7.8	2,332 (1,751/581)	13,063 (11,812/1,251)	5.6

[a] G/M, G/M*: ratio #real genes/#metagenes. All genes of the filtered metagenes are considered in the first case (G/M). In the second case (G/M*) the metagenes containing the filtered single genes are considered. Consequently not all single genes of the respective metagenes are taken into account and one gets on the averaged G/M>G/M* for the same criterion.

[b] Toplist FC-expression filter: Metagenes/genes are ranked with decreasing FC-value. The number of items indicated on top of the list are selected.

[c] (#in/#out): #in denotes the intersection between the number of metagenes/real genes sampled by filtering the metagenes and real genes. #out is the respective number of genes not sampled by the metagene filter

[d] Toplist variance filter: Metagenes/genes are ranked with decreasing variance of their expression profile. The number of items indicated on top of the list are selected.

[e] False discovery rate (fdr) significance filter, i.e. all metagenes/genes with smaller fdr-values than the indicated threshold are included in the list

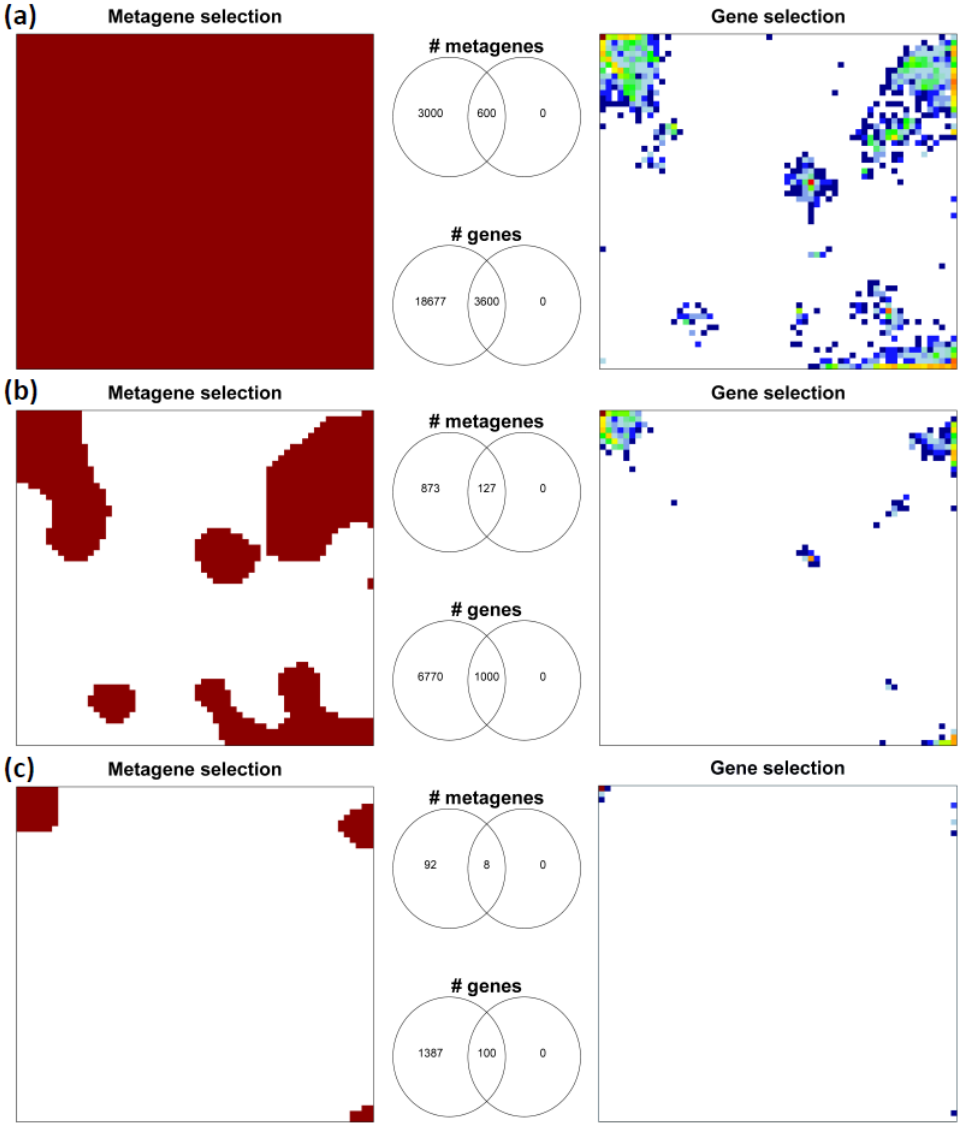

Figure S 1: Filtering genes or metagenes by differential expression: Equal numbers of metagenes (left mosaics) and single genes (right mosaics) are selected using the FC-3600 (a), FC-1000 (b) and FC-100 (c) filters. The brown areas in the left part show the selected metagenes and the colored tiles in the right part the density of single genes (maroon to blue codes high to low densities). The Venn-diagrams illustrate the degree of overlap between the metagenes and genes after metagene and single gene filtering. Note that the FC-3600 filter if applied to single genes (right mosaic in panel a) selects features in the same areas of the mosaic as the FC-1000 filter if applied to metagenes (left mosaic in panel b). The similar result was found for FC-100 and FC-1000 filters if applied to metagenes and single genes, respectively.

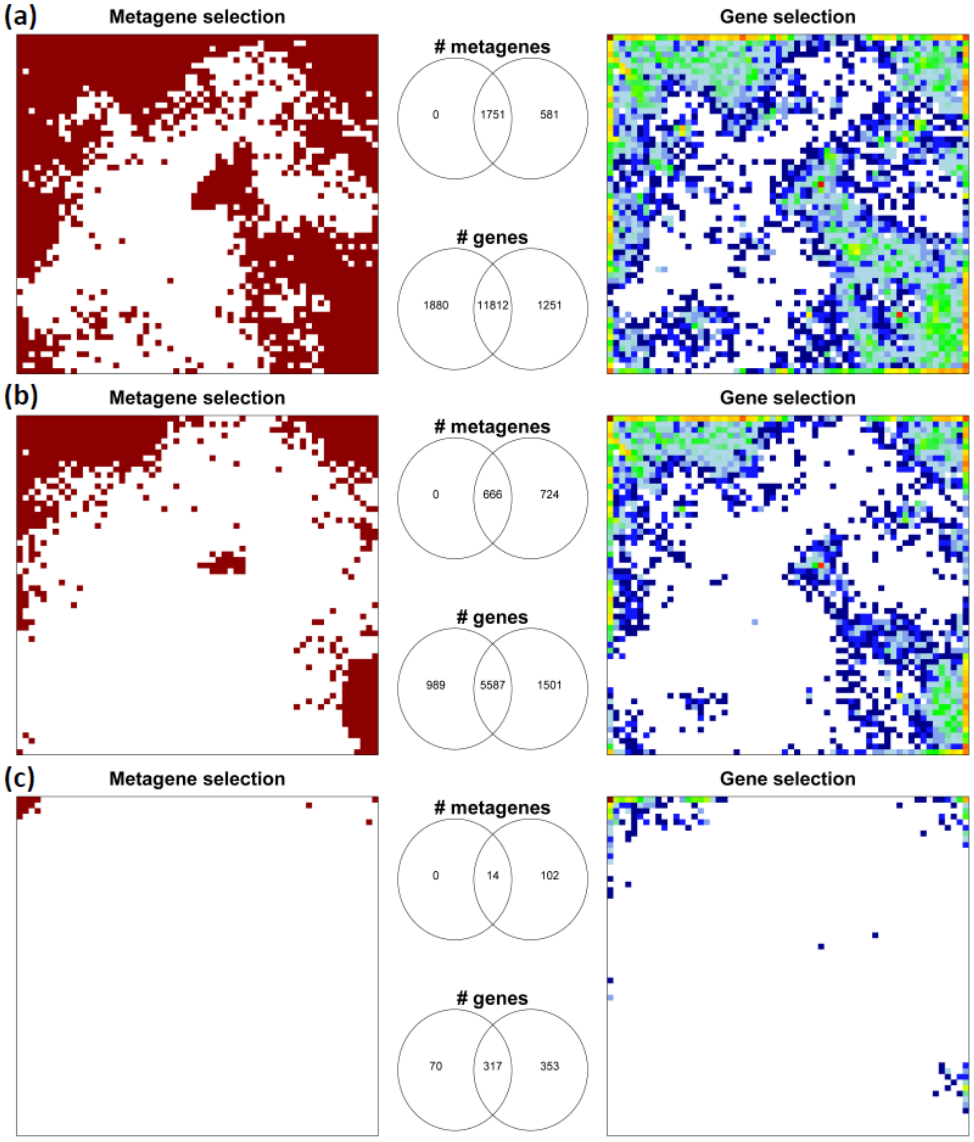

Figure S 2: Filtering genes and metagenes by significance: Equal FDR-thresholds are applied to metagene (left mosaics) and single gene (right mosaics) lists selected using FDR< 0.5 (panel a), FDR<0.4 (panel b) and FDR<0.2 (panel c) filters. The brown areas in the left part show the selected metagenes and the colored tiles in the right part the density of single genes (maroon to blue codes high to low densities) selected by filtering metagene and single gene lists, respectively. The Venn-diagrams illustrate the degree of overlap between the metagenes and genes after metagene and single gene filtering. The single gene filter selects consistently a roughly twice as large number of metagenes and a slightly larger number of single genes than the respective metagene filters.

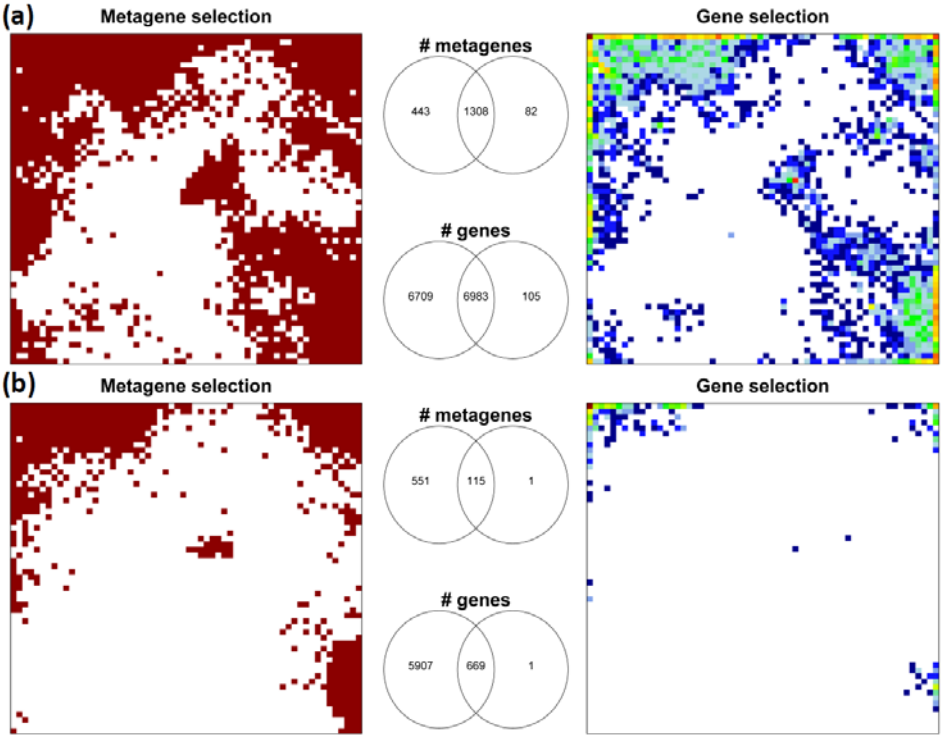

Figure S 3: Analogous to Figure S 2: Comparison of metagene/single gene filters using different FDR thresholds, 0.5/0.4 (panel a), 0.4/0.2 (b). The Venn diagrams indicate that the less stringent metagene filter shifts the number of metagenes and single genes selected towards the metagene filter.

2. Clustering metagenes and single genes

Figure S 4 shows two simple cluster trees with different relative distances between their branching points, L_1 and L_2. The left one characterizes more compact clusters than the right one. It qualitatively explains the difference between the cluster trees obtained from single gene (left below) and metagene (right below) lists. In the chosen radial representation the cluster trees are projected to unit circles, which is equivalent with the normalization of the mean Euclidian distance between all samples to a unique constant. The length of a particular branch in this plot consequently estimates its relative length defined as the ratio of its Euclidian distance divided by the mean value. The mean length of the 'outer' branches, $\langle L_1 \rangle$, thus estimates the mean relative distance between most similar samples on the lowest level of clustering whereas the mean length of the 'inner' branches estimates the mean mutual distance between the largest clusters. This distance of closest approach is markedly smaller for metagene gene cluster trees than for single genes meaning that the observed metagene clusters are more compact as illustrated schematically by the sketch in Figure S 4.

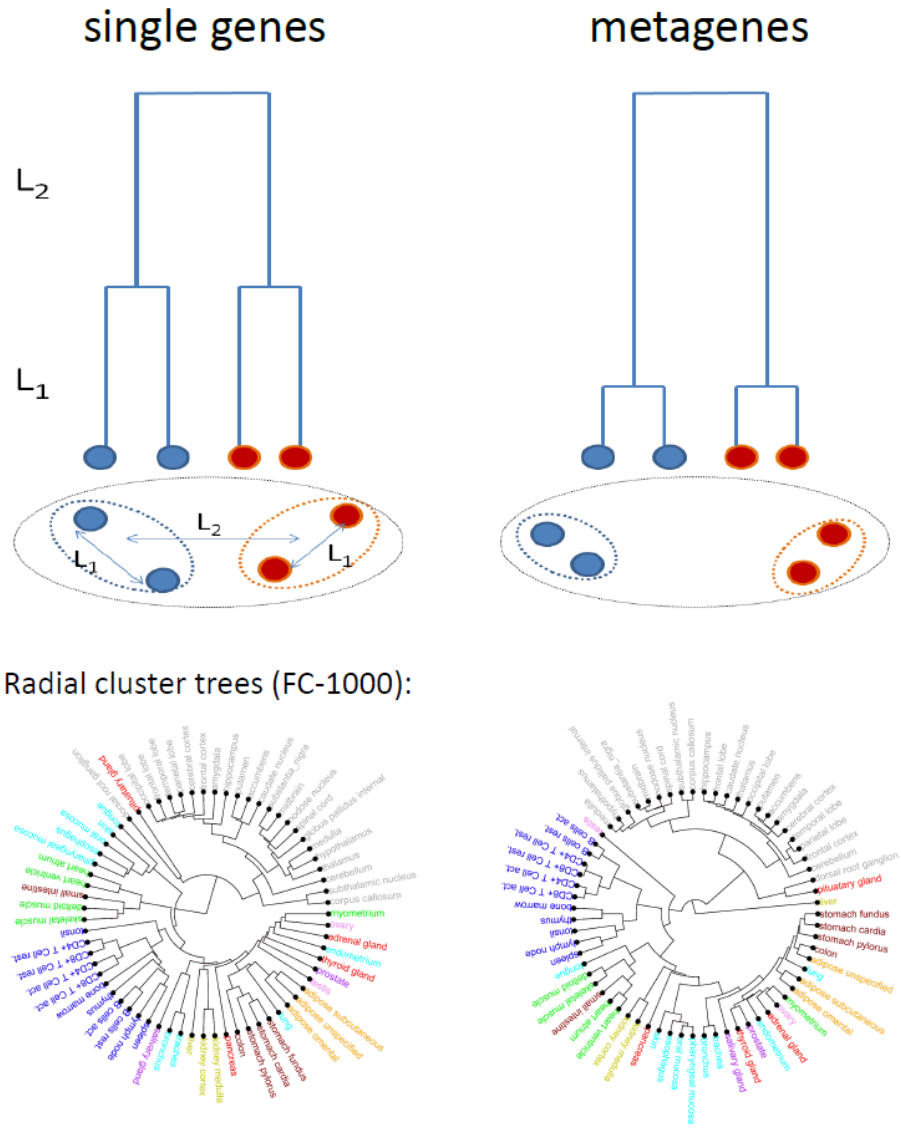

Figure S 4: Schematic illustration how the relative length of the branches in the tree transforms into the compactness of clusters: This distance of closest approach is markedly smaller for metagene gene cluster trees than for single genes meaning that the observed metagene clusters are more compact as illustrated schematically by the sketch in the part above.

3. 3D-ICA map

We generated three dimensional ICA versions of the 2D ICA plots shown in the main paper. They clearly show that typically one of the linear clusters of one of the tissue categories points along the third main component of independent variation. The smaller ICA on the right enlarge the clusters formed by three selected tissue categories. The respective ICA-plots were calculated separately.

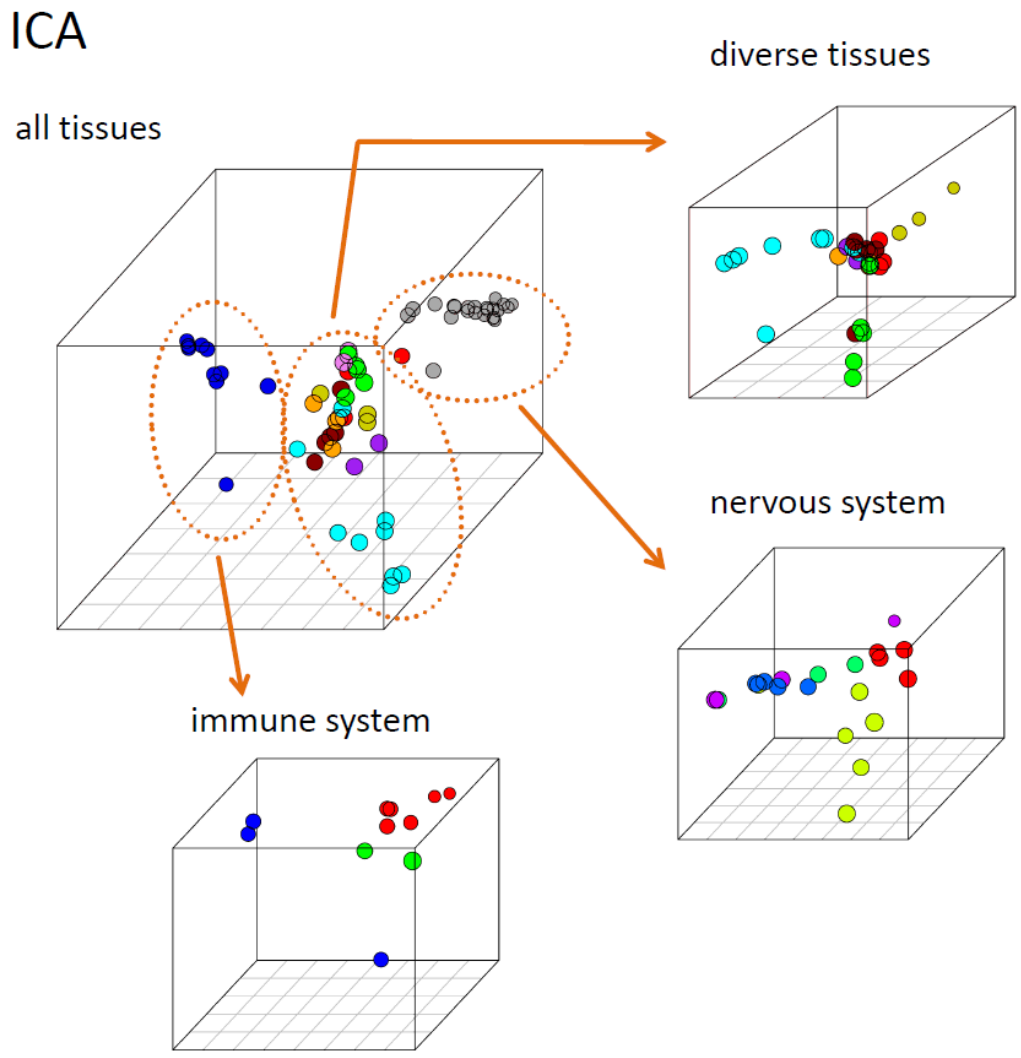

Figure S 5: 3D ICA plots of the tissues studied.

4. Reference

[1] Strimmer K: **A unified approach to false discovery rate estimation**. *BMC Bioinformatics* 2008, **9**(1):303.