Mining SOM expression portraits: Feature selection and integrating concepts of molecular function

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

[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] Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment (LIFE); Universität Leipzig, D-4103 Leipzig, Philipp-Rosenthalstr. 27, Germany
[4] Helmholtz Centre for Environmental Research, Department of Metabolomics, D-04318 Leipzig, Permoserstr. 15, Germany

* to whom correspondence should be addressed

Abstract

Background: Self organizing maps (SOM) enable the straightforward portraying of high-dimensional data of large sample collections in terms of sample-specific images. The analysis of their texture provides so-called spot-clusters of co-expressed genes which require subsequent significance filtering and functional interpretation. We address feature selection in terms of the gene ranking problem and the interpretation of the obtained spot-related lists using concepts of molecular function.

Results: Different expression scores based either on simple fold change-measures or on regularized Students t-statistics are applied to spot-related gene lists and compared with special emphasis on the error characteristics of microarray expression data. The spot-clusters are analyzed using different methods of gene set enrichment analysis with the focus on overexpression and/or overrepresentation of predefined sets of genes. Metagene-related overrepresentation of selected gene sets was mapped into the SOM images to assign gene function to different regions. Alternatively we estimated set-related overexpression profiles over all samples studied using a gene set enrichment score. It was also applied to the spot-clusters to generate lists of enriched gene sets. We used the tissue body index data set, a collection of expression data of human tissues as an illustrative example. We found that tissue related spots typically contain enriched populations of gene sets well corresponding to molecular processes in the respective tissues. In addition, we display special sets of housekeeping and of consistently weak and high expressed genes using SOM data filtering.

Conclusions: The presented methods allow the comprehensive downstream analysis of SOM-transformed expression data in terms of cluster-related gene lists and enriched gene sets for functional interpretation. SOM clustering implies the ability to define either new gene sets using selected SOM spots or to verify and/or to amend existing ones.

1. Introduction

High-throughput genome-scale sequencing and microarray technologies generate huge amounts of data which challenge tasks such as dimension reduction, data compression, visual perception, data integration and extraction of biological information. A natural basis for organizing gene expression data is to group together genes with similar patterns of expression, e.g. of highly correlated expression values. A series of different similarity measures and clustering algorithms have been developed in the last decade for this purpose. Another important task in extracting reliable information is to examine the extremes, e.g., genes with significant differential expression in two individual samples or in a series of measurements and to judge the degree of significance. Finally, gene set enrichment methods have been established to link previous biological knowledge about groups of functionally related genes with the results of new investigations.

This study addresses the question how to properly combine self organizing maps (SOM) machine learning with differential expression and gene set enrichment analysis. SOMs describe a family of nonlinear, topology preserving mapping methods with attributes of clustering and strong visualization. They are generally used in many fields like bioinformatics for dimension reduction and the grouping of high dimensional data. SOMs are very intuitive and easy to understand and therefore used in decision-making. SOMs were devised by Kohonen [1], and first used by Tamayo et al. [2] and Törönen et al. [3] to analyze gene expression data. Our approach follows that of Nikkilä et al. [4] and of Eichler et al. [5] who configured the SOM method in such a way that it combines sample- and gene-centered perspectives. Particularly it transforms large and heterogeneous sets of expression data into a gallery of sample-specific ‘portraits’ which can be directly compared in terms of similarities and dissimilarities of their textures.

The ‘portraits’ represent mosaic-images where each tile represents a microcluster of single-gene of similar expression profiles which is characterized by one metagene reflecting the mean expression profile of the associated single genes. Due to the specifics of the machine learning algorithm metagenes of similar profiles cluster together into so-called spots in the SOM-images which can be easily identified by visual inspection and used as unsupervised gene clusters in downstream analysis. This SOM-spot clustering combines the criteria of coexpression (i.e. of similar profiles in the series of samples studied) and of over-/underexpression (in a subset of the samples studied). Our previous publication addresses methodical aspects of the machine learning step, spot selection and compares the transformed metagene-profiles with that of the original single gene profiles [6].

SOM machine learning (and these methodical aspects) alone is however insufficient to extract important features and biological information from the data. The obtained spot-clusters need further filtering and association with previous knowledge for this purpose. Here we address these data mining tasks with special emphasis on the structure of SOM-transformed data to enable their downstream analysis and biological interpretation.

The first focus of this publication addresses the gene ranking problem in SOM-transformed data. SOM training typically uses a simple fold-change (FC) scale with respect to the mean expression of each gene in the pool of all samples to detect genes of interest. The FC-score however does not provide explicit information about statistical significance for the observed expression changes and thus it might have disadvantages in generating false signals, e.g., if large expression changes are paralleled with high uncertainties of the respective signals or, vice versa, if relatively small changes refer to accurate signals. SOM mapping must therefore be supplemented with appropriate algorithms to assess significance of the features selected. In this publication we apply significance analysis to the spot-clusters of genes identified by the SOM method using three alternative test statistics based either on FC-measures or on regularized Students t-statistics with special emphasis on the error characteristics of microarray expression data. Such local, cluster-related lists of genes are expected to improve the resolution of the method to identify sample-specific features with a common functional impact.

The second focus of this publication addresses gene set enrichment analysis under special consideration of the spot-clusters generated by SOM machine learning. It is based on the fact that the importance of genes in terms of their relation to a particular molecular function is not necessarily associated with strongest or most significant changes of expression provided by their rank in the obtained lists. Instead, it can also involve weak but consistent alterations of transcript abundance. Therefore gene set based methods have been developed to investigate phenotypic changes at the level of biological function considering, for example, the involvement of genes into signalling pathways, their relation to cellular components or their chromosome location [14-20]. These methods essentially assess the enrichment of a set of several genes in the list of differentially expressed genes compared with the total reservoir of genes studied. The members of the set are defined a priori by some biological commonality for certain phenotypes. The main advantage of such methods over single gene based methods is that they directly link the ranked gene list with biological knowledge and therefore provide better functional insight into the cause of the phenotypic differences under study.

Our work thus aims at refining the avenues for feature mapping and data reduction offered by SOM machine learning. We use the microarray expression data of a series of 67 different human tissues taken from ten tissue categories such as nervous, immune system, epithelial and muscle tissues as an illustrative example to demonstrate the strengths of the SOM method in disentangling large heterogeneous data sets. The paper is organized as follows: In the Results-section we present and discuss our approach of significance and enrichment analysis of SOM-transformed data if applied to the tissue body-index data set. In the methodical part we provide details of the applied methods and algorithms and of relevant characteristics of microarray data. In the additional material we address aspects of SOM data mining which supplement our main results. Finally, we complemented our R-package ‘oposSOM’ [6] with appropriate add-on functions enabling the differential expression and geneset enrichment analysis of SOM-transformed microarray data.

2. Results

2.1. SOM-portraits and rank maps

Genome-wide gene expression data of 67 selected tissues taken from 10 tissue categories were pre-processed and subsequently used to train a SOM as described previously [6]. Figure 1 shows the obtained SOM-portraits of selected tissues using a 60x60 mosaic grid. The method identifies coherent tissue-specific texture patterns of gene expression readily discernable in the obtained gallery of SOM images. Particularly, our SOM machine learning method partitions the more than twenty thousand 'single' genes probed by each microarray into 3600 miniclusters arranged in a two-dimensional mosaic map which visualizes the specific expression pattern of each sample in terms of a color-coded texture indicating regions of over- and underexpression by red and blue spots, respectively. Most of the spots are tissue specific features which are found only in one or a very few tissue categories such as nervous, immune system or muscle tissues.

Each tile of the SOM mosaics thus refers to one metagene which, in turn, is associated with a microcluster of single genes the number of which varies from tile to tile. The expression profiles of each metagene serves as representative (or prototype) of the respective cluster of co-regulated single genes. The color gradient of the map was chosen to visualize over- and underexpression of the metagenes compared with the mean expression level in the pool of all tissues studied.

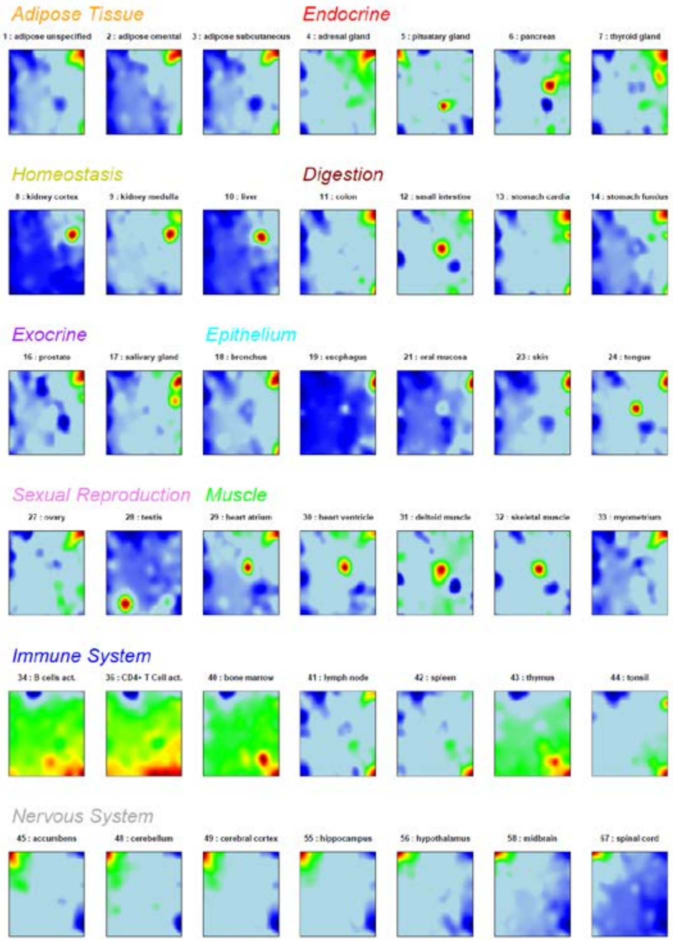


Figure 1: Gallery of SOM portraits of 42 selected tissues of different tissue categories (see ref. [6] for details).

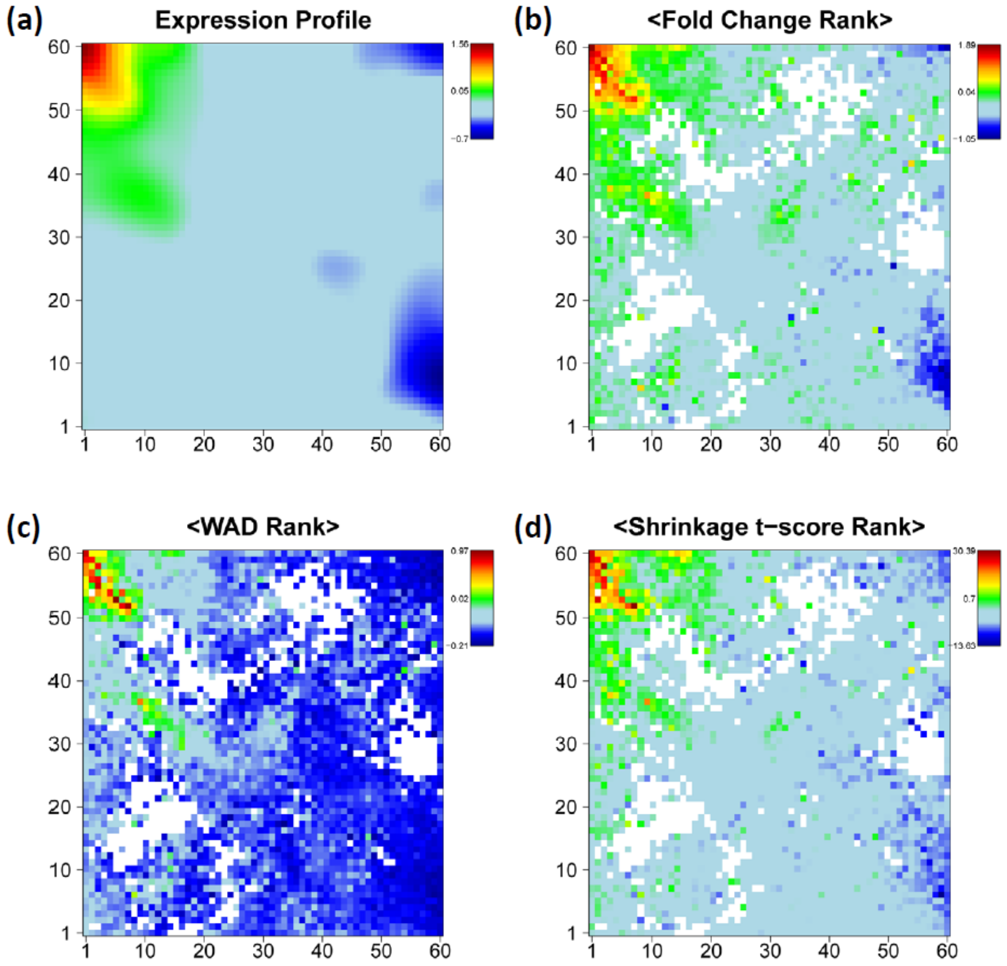


Figure 2: Expression image of *nucleus accumbens* ('standard' SOM profile, panel a) and the average-rank maps for FC, WAD and shrinkage t-score statistic (b-d). The numberings of the tiles k=1…60 are given at the vertical and horizontal borders of the SOM. White areas indicate empty metagenes.

Figure 2 shows the SOM expression image of one particular tissue example, nucleus accumbens, taken from the category of nervous tissues in log FC units (panel a) together with the respective average-rank maps for the three different scores used (panels b-d), namely the FC-, weighted average difference (WAD)- and shrinkage t-score, respectively (see (Eqs. (1) and (2) in the methodical part). The rankings of genes refer to total gene lists which contain all genes studied. These maps color-code the mean rank of each metagene which was calculated as the arithmetic average over the individual rankings of the associated single genes in the total list. In general, genes on top of the list accumulate in the red overexpression spot of the standard SOM-profile however with a few exceptions, e.g. in the range of the green spot below the red one. The three alternative scores provide very similar pattern, however with subtle differences: The contrast, i.e. the gradient between areas of under- and overexpression is largest for the WAD-ranking and smallest for FC-ranking with t-shrinkage in-between. Similar trends are observed for the SOM expression profiles which are color-coded according to the FC- and WAD-scores of their metagenes [6]. Note also that the rank maps reveal subtle details within the SOM-spots such as the chain-like cluster of metagenes of small rank within the overexpression spot (compare panel a with b-d in Figure 2). The analysis of such fine-structures might help to refine the subsequent selection of relevant genes within the spots.

The examples shown in Figure 3 further support this result: The t-shrinkage rank-map of small intestine, T-cells and lymph node show a partly better resolved fine structure of highly ranked genes in different regions of the map than the standard SOM mosaics which use the log FC expression scale. On the other hand, the rank map of colon is dominated by blue areas which reveal an average level of relatively low rankings. This effect presumably reflects the relatively small expression level of the genes in the overexpression spot in the top right corner of the map which give rise to relatively large rank numbers. The whole atlas of the rank maps of all tissues studied is shown in Additional file 1.

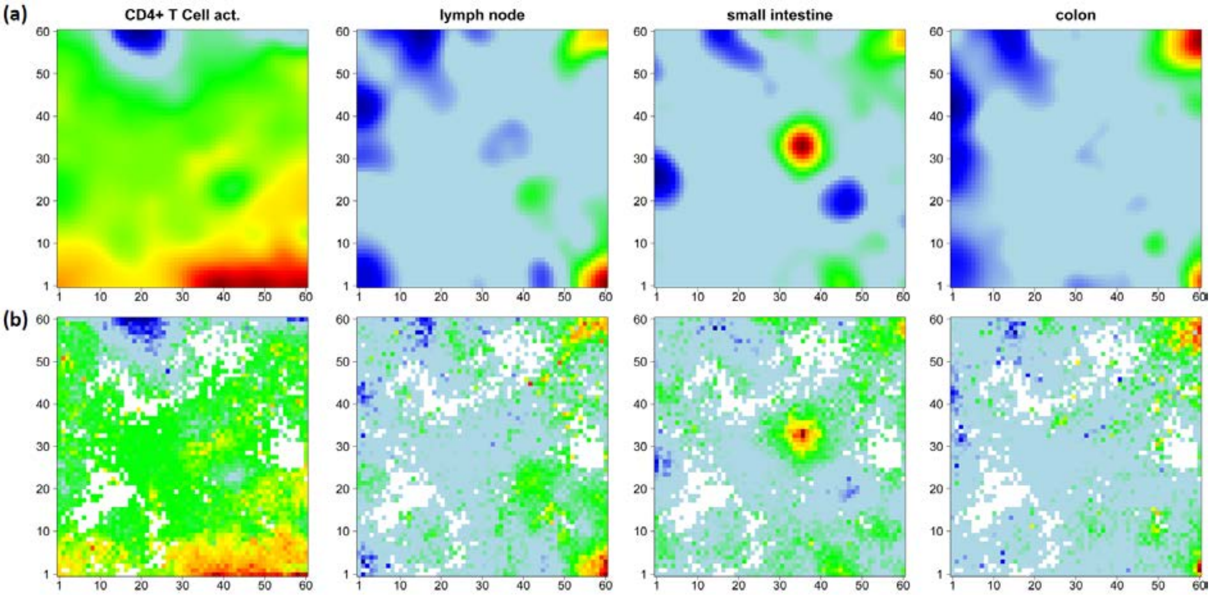


Figure 3: Comparison of standard SOM in log FC-scale with rank maps based on global gene lists according to the t-shrinkage statistics. Metagenes of high overexpression and of small average rank of the associated single genes are coded in red. Both options show essentially similar textures. The rank maps partly reveal more detailed spot pattern or a low overall rank level (blue, e.g. colon). The atlas of rank maps of all tissues studied is shown in Additional file 1.

2.2. Total gene lists

The alternative scores generate ordered total lists of genes for each tissue with characteristic differences between the methods as illustrated in the rank-map shown in Figure 2. The WAD-score, for example, strongly weights highly expressed genes which concentrate in a few metagene-tiles in the top left corner of the map. As a consequence, these metagenes occupy smaller ranks in the WAD-list than in the respective FC- or shrinkage-t lists with consequences for the textures of the respective rank maps. The present study does not aim at comparing the performance of different expression scores in absolute units, an objective which is problematic in the absence of a suited gold standard. Previous work makes use either of synthetic simulation data, of correlation measures in real-world chip applications or of special calibration data sets to judge the quality of different expression scores [24-29]. It turned out that t-shrinkage and different FC-based scores such as the WAD-score are generally suited measures to generate lists of regulated genes. Here we apply the three scores as three complementary alternatives with a specific focus on different expression properties: Particularly, WAD-lists heavily weight strongly expressed genes. In consequence, subtle expression changes of weakly expressed genes potentially get lost in WAD-lists. FC-lists directly rank the genes according to their differential expression and thus represent a simple and intuitive measure related to the change of mRNA abundance. FC-lists are however prone to generate false positives because the FC-score equally weights strongly and weakly expressed genes with usually smaller and larger noise levels, respectively. The t-shrinkage score explicitly considers the noise level of the genes which however might raise problems due to the uncertainty of the error estimates as discussed in the methodical section. Because of their specific advantages and disadvantages we consider the different scores rather as complementary measures than as competitive ones providing information which mutually supplement each other.

Figure 4a shows the p-value distribution of differential expression of nucleus accumbens based on the t-shrinkage score (the atlas of the p-value distributions of all tissues studied is given in Additional file 2). It well separates into a constant noise floor and the left-skewed subpopulation of differentially expressed genes constituting a percentage of about 66% of all genes available. We compare the global lists ranked with increasing t-shrinkage, FC- and WAD-scores using four plots, namely (i) the rank comparison (RC), (ii) the correspondence at the top (CAT)- ,(iii) the p-CAT and (iv) the Δp-CAT plots (Figure 4b, see Methods section for description). The RC-plot compares the individual positions on top of the lists by appropriate color-coding. It reveals moderate disordering between the three lists where most ranks agree within ±20 positions up to rank r=50 (see green symbols). The CAT-plot presents the cumulative fraction of common genes on top of the list for positions below a running threshold. In our example it shows that best agreement is achieved in FC/WAD-comparisons for ranks r ~ 10…100. However, also the other combinations provide acceptable agreement between the lists with CAT(r)≥ 0.5 for positions r<100, meaning that at minimum 50% of the same genes are included in pairs of lists up to rank one hundred.

The p-CAT plot estimates the agreement between the lists in units of the cumulative log p-value of the t-shrinkage statistics. It enables to differentiate whether a given CAT-value refers to more similar or very different p-values and thus it estimates the importance of rank differences. The respective Δp-CAT plot shows the difference between the p-CAT value of the FC- or WAD-score and that of the t-shrinkage statistics which provides the lower margin per definition. The Δp-CAT values of the global lists of the FC- and WAD-scores initially increase for ranks below 5-20 indicating that the different rankings are associated with clearly different p-values. For positions r> 20 the Δp-CAT values remain virtually constant indicating that the alternative lists provide consistent results where rank differences reflect rather the noise inherent in the data than systematic biases between the scores used.

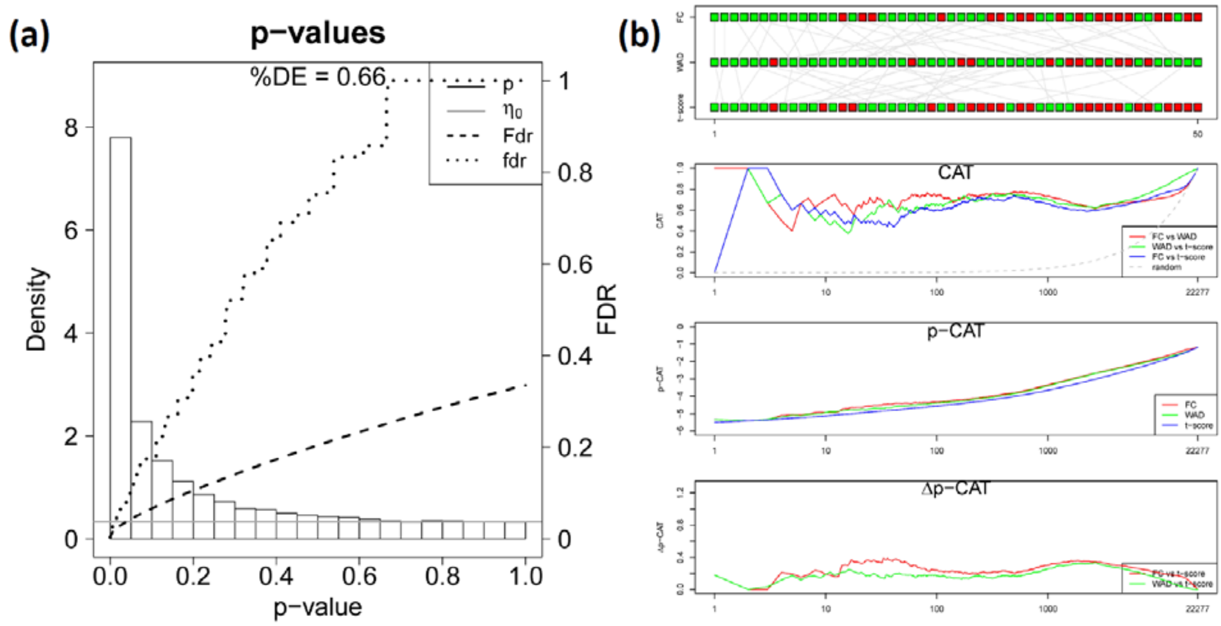


Figure 4: Global significance analysis of accumbens sample: p-value distribution and fdr- and FDR-curves of the t-shrinkage statistics (panel a) and comparison of gene rankings for FC-, WAD and t-shrinkage scores using the RC-, CAT-, p- and Δp-CAT plots (panel b).

2.3. Local, spot-related gene lists

The spot-texture of the SOM portraits of individual tissues implies to generate spot-related gene lists by taking into account only the single genes which are associated with the metagenes forming a particular spot. Recall that a spot clusters genes of similar and thus co-variant expression profiles in the series of samples studied. Our spot-based significance analysis therefore shares similarity with methods which exploit the correlation between genes in significance testing of differential expression [30], [31] because it selectively applies to sub-ensembles of genes of highly correlated expression profiles.

In the next step we therefore analyzed the p-value distribution and the mutual list characteristics for three selected spots referring to over- (spot I), under- (spot II) and virtually mean (spot III) expression (see Figure 5) which contain different numbers of single genes (I: 980, II: 745, III: 1,947). Spots of regulated metagenes are detected for each tissue using the 98% / 2% quantile criterion for over- / underexpressed metagenes, respectively. The fraction of differentially expressed genes in the spots either markedly exceeds (I,II: %DE=0.95) or falls below (III, %DE=0.53) the global value (%DE=0.66). The ranking characteristics of the overexpression spot I closely resembles that of the global lists indicating that this spot contains most of the 'leading' genes of the global list (compare Figure 5 and Figure 4). Note that the overexpression spot selects strongly differentially expressed genes. Therefore the level of agreement between the alternative lists is slightly better especially for FC/WAD-comparison (CAT(r<100) ~ 0.6) compared with the respective comparisons between the global lists. Note that the spot-filtering effectively combines the scoring of differential expression with the selection of co-expressed and correlated genes. It has been previously shown that 'correlation-sharing' for the detection of differentially expressed genes improves the performance of the analysis in terms of the false discovery rate [30]. For spot I we indeed obtain a much smaller total cumulative FDR value of Fdr(p=1)≈0.05 (Figure 4a) compared with the total list (Fdr(p=1)≈0.35; Figure 5).

Contrarily, the alternative gene lists taken from the underexpression spot II largely diverge revealing the lack of agreement among the top 10 – 50 features. The CAT-plot shows best agreement for FC/WAD-comparisons with CAT(r≈100)< 0.6 and worst for FC/t-shrinkage (CAT(r≈100)<0.2). These rank comparisons are paralleled by relatively large differences of the p-CAT and Δp-CAT characteristics revealing systematic and significant rank differences due to the specific biases of the used scores. Particularly, FC/t-shrinkage comparisons shows largest dissimilarity in the CAT- and p-CAT-plots for r<50 followed by WAD/t-shrinkage comparisons. These discrepancies can be rationalized by the large uncertainty of low expression genes which accumulate in the underexpression spot. Note that the different rank-maps clearly express the discrepancy between the rankings in the regions of underexpression which largely lose their structured texture especially in the t-shrinkage rank-map.

Interestingly, also spot III contains a large fraction of differentially expressed genes (%DE=0.53) despite the fact that the metagene expression is virtually on the moderate level. The comparisons between the alternative lists provide less agreement when compared with spot I but almost similar trends. The spot of 'mean expression' obviously still contains residual amounts of significantly differently expressed genes which appear as green and grey tiles in the region of spot III in the rank-map (Figure 2).

To generalize these results we calculated mean global and local CAT(r) and Δp-CAT(r) values for lists of length r=10 and 100 of all tissue samples studied considering either all genes or the genes of the strongest overexpression spot, respectively (see Additional file 3 for details). The results of these global and local rank comparisons confirm the trends discussed above: global FC- and WAD-lists of length r=10 – 100 agree to about 70% on the average whereas global FC/t-shrinkage and WAD/t-shrinkage lists are identical to about 50%. Local lists are slightly more similar by a few percent than global ones due to the pre-filtering of the genes in the SOM-spots. The respective Δp-CAT values reveal that the significance level of the alternative scores is virtually identical for all considered lists.

In summary, the different scoring methods typically provide similar and virtually equivalent gene lists for overexpression spots but diverging lists for underexpression spots. The rank-map of the respective methods clearly express this difference: Whereas the regions of overexpression are essentially similar in the different rank-maps (see red areas in Figure 2) the regions of underexpression appear either as relatively localized spots in the FC-rank and, to a less degree, in the WAD-rank maps or they 'smear' over larger regions in the shrinkage-t rank map due to the large uncertainty of low expression values. In conclusion, overexpression rankings provide robust lists of differentially expressed genes which are relatively independent of the scoring method used thus allowing the quantitative analysis in terms of the obtained rank and expression level. In contrast, underexpression lists are highly uncertain providing essentially qualitative information, namely that the respective genes are weakly expressed. Discrimination analysis between the different samples and especially GO-enrichment analysis to identify overrepresented gene sets should therefore focus on overexpression spots. The t-shrinkage score will be applied as the default criterion for gene ranking problem in the remainder of this study.

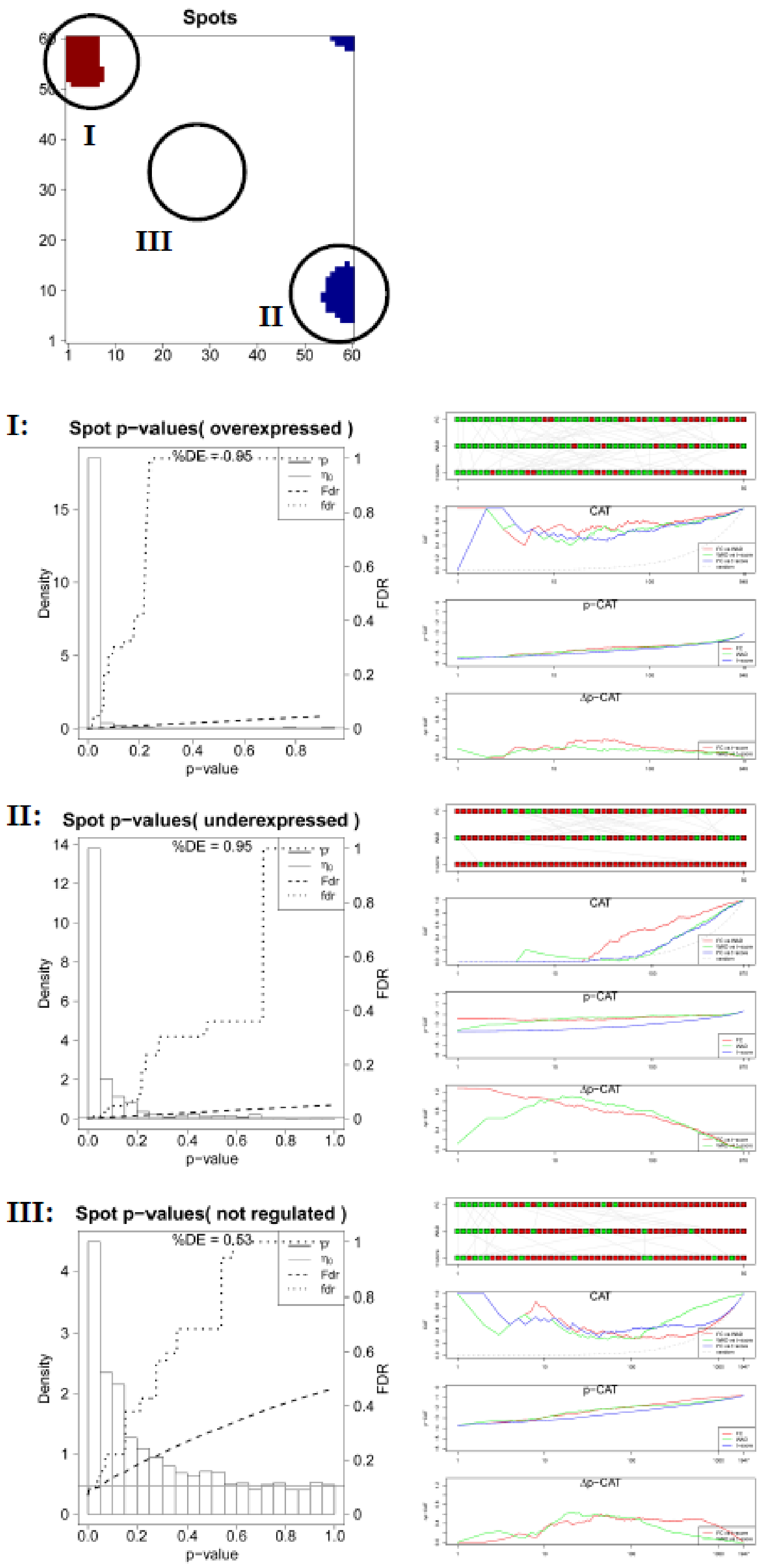


Figure 5: Local significance analysis of selected spots of the accumbens sample (see part above): p-value distribution and fdr- and FDR-curves of the t-shrinkage statistics (left part) and comparison of gene rankings for FC-, WAD and t-shrinkage scores using the RC-, CAT-, p- and Δp-CAT plots (right part).

2.4. Global overrepresentation analysis

The correlation and coexpression of the gene profiles in each spot can be utilized as a simple heuristic with implications for tentative gene function because biological processes are governed by coordinated modules of interacting molecules [21]. Application of gene set enrichment analysis to the series of about one dozen stable over- and underexpression spots detected in the SOM of human tissues will make explicitly use of this 'guilt-by-association' principle which assumes that co-expressed genes are likely to be functionally associated [22][23]. Enrichment analysis is expected to assign putative gene function(s) to the selected spots. Below we compare several options of enrichment analysis estimating either 'overrepresentation' of the members of a priori functional gene sets in the spot list, their 'overexpression' in terms of differences of the average expression levels in the set and the list and the combination of both options.

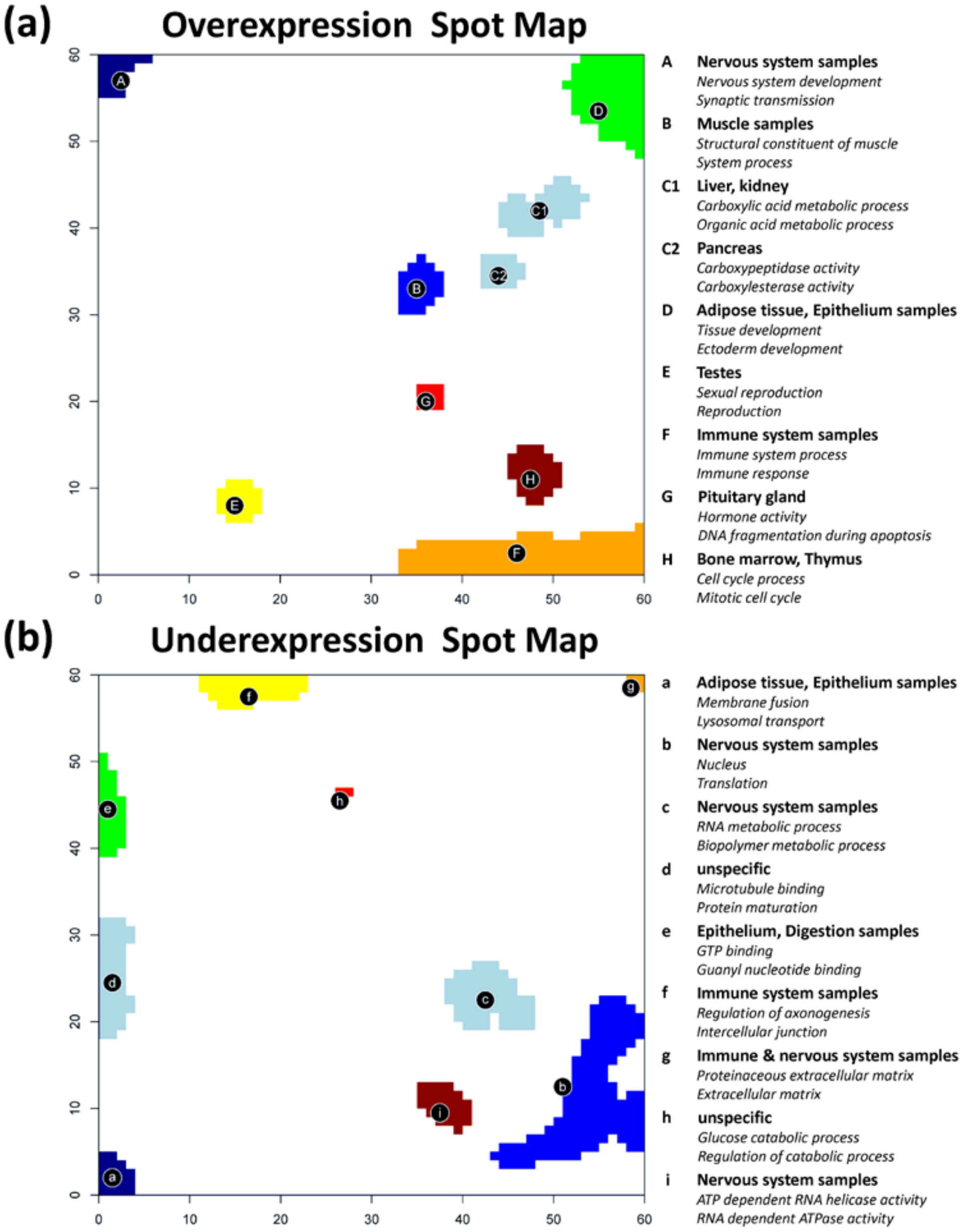


Figure 6: The over- and underexpression summary spot maps show nine spots each which are strongly over-/underexpressed in different tissues (part a and b, respectively). Overrepresentation of a collection of 1454 gene sets is estimated for each spot using the hypergeometrical distribution. The right legend assigns the two most significantly overrepresented gene sets to the respective spots.

Essentially nine overexpression spots are identified in the SOM-images of all tissues studied using the 98-percentile criterion of maximum expression. These spots are collected into one, so-called overexpression summary map as described in [6]. Subsequently GO-gene set overrepresentation analysis using the hypergeometrical (HG-) test is applied to the lists of genes contained in each of the overexpression spots (see the Methods section below). 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, biological process and cellular component. The HG-test then provides an ordered list of gene sets ranked with decreasing significance of overrepresentation with respect to the random appearance of genes from the set in each of the spots.

Figure 6a shows the overexpression summary map with the nine spots of strongly overexpressed metagenes. The legend assigns the two leading overrepresented gene sets in the list of each of the spots to get a first idea about the possible biological context of the genes in the spots. For example, spot A in the top left corner of the SOM is clearly related to molecular processes in nervous cells according to the two leading gene sets. The more detailed inspection of the lists reveals that ten out of the top-twenty gene sets of spot A are related to nervous system (see Additional file 3). Also other tissue-specific spots can be associated with distinct molecular functions such as immune system processes (immune systems samples, spot F), sexual reproduction (testis, spot E) or muscle contraction (muscle tissues, spot B). Hence, the functional context of the different spots according to previous knowledge is clearly related to the tissues showing the respective overexpression spot.

The analogous overrepresentation analysis was performed for the underexpression spots related to local minima of the metagene expression profiles (Figure 6b). The functional context of these spots thus refers to genes which are strongly underexpressed in the tissues showing this spot (see also the respective spot expression heatmap shown in Additional file 3). For example, spot b, c and g related to processes in the nucleus, RNA processing and the extracellular region, respectively, are underexpressed in most nervous tissues. Spot g and also spot f (related to neurogenesis) are underexpressed in immune system tissues. The latter spot, in turn, shows clear overexpression in nervous tissues, which is however not detected in the overexpression map selecting only the regions of strongest overexpression. Thus, overrepresentation analysis of both, over- and underexpression spots provide complementary information: On one hand, they allow to assign antagonistic gene activities in the same tissue and in different tissues. On the other hand, parts of the underexpression spots occupy different regions of the map than the overexpression spots. In consequence, combination of both maps extends the range of relevant gene sets and thus also the functional context studied. For example, spots a and d related to biopolymer metabolism and microtubules, respectively, are not detected in the overexpression map. Spots e and f are both overexpressed in nervous tissues. They occupy regions near the spot A also overexpressed in nervous tissues. The respective functional context of all three different spots allows to disentangle subtle details of gene activity in nervous tissues. A similar relation exists for overexpression spot F and underexpression spot b, where the former one overrepresents gene sets related to cell cycle and the latter one gene sets related to nucleus activity.

2.5. Alternative spots selections

In the previous subsection we have shown that over- and underexpression spots partly occupy different regions of the map with complementary information about their functional context. One can apply also alternative methods of spot selection using hierarchical clustering of the metagenes based on the Euclidian distance between them or determining correlation cluster based on Pearson correlation coefficients between the metagenes [6]. The former method provides an area-filling fragmentation of the map into different spots which typically occupy larger areas than the spots from the over-/underexpression summary maps. In the Additional file 3 we demonstrate that the cluster-spots detect, for example, different groups of genes related to the functioning of nervous tissues.

The correlation clusters provide almost similar results however also with subtle specifics of their functional context (Additional file 3). This method preferentially selects areas of highly variable metagenes along the border of the map with subtle differences between the functional context of adjacent clusters. In summary, different spot selection algorithms and criteria fragment the expression landscape of the map in partly different ways with complementary information about the functional context of the associated genes. The suitability of the different methods depends on the particular aims of the issues studied and is not in the focus of this methodical publication. In the remainder of the paper we will use the overexpression spots to extract further functional information from the maps. Note however that over- and underexpression spot selection can be applied to the individual portraits of each sample and thus they provide specific enrichment characteristics as described below. In contrast, the k-means and the correlation clusters are based on the similarities between the metagene profiles and thus they refer to all samples in terms of the global overrepresentation of the associated genes. Application of the GSZ-score allows however to study also sample-specific enrichment of the respective genes (see below).

2.6. HG-enrichment analysis

Gene set overrepresentation analysis as described in the previous subsection applies to global spots of adjacent metagenes taken from the overexpression summary map. The real genes associated with each spot are the same in all tissues studied because the overexpression spot map summarizes the maximum size of each spot sizes observed in any of the tissues and thus it neglects sample-specific alteration of the spot size. This global approach applies to the whole series of tissue samples. It consequently lacks sample-specificity. Thus, overrepresentation of a selected gene set is independent of the individual expression level of the genes in the different samples. In the following we present and discuss two approaches to take into account sample-specific gene expression. We will use the term gene set overexpression analysis if the mean expression of the set-members is compared with the mean expression of all genes in the list without considering the number of set members in the list in contrast to gene set overrepresentation analysis which is based solely on the latter criterion. The term enrichment analysis will be used if both criteria, overrepresentation and overexpression, are combined which enables the refinement of gene set analysis in terms of sample-specificity.

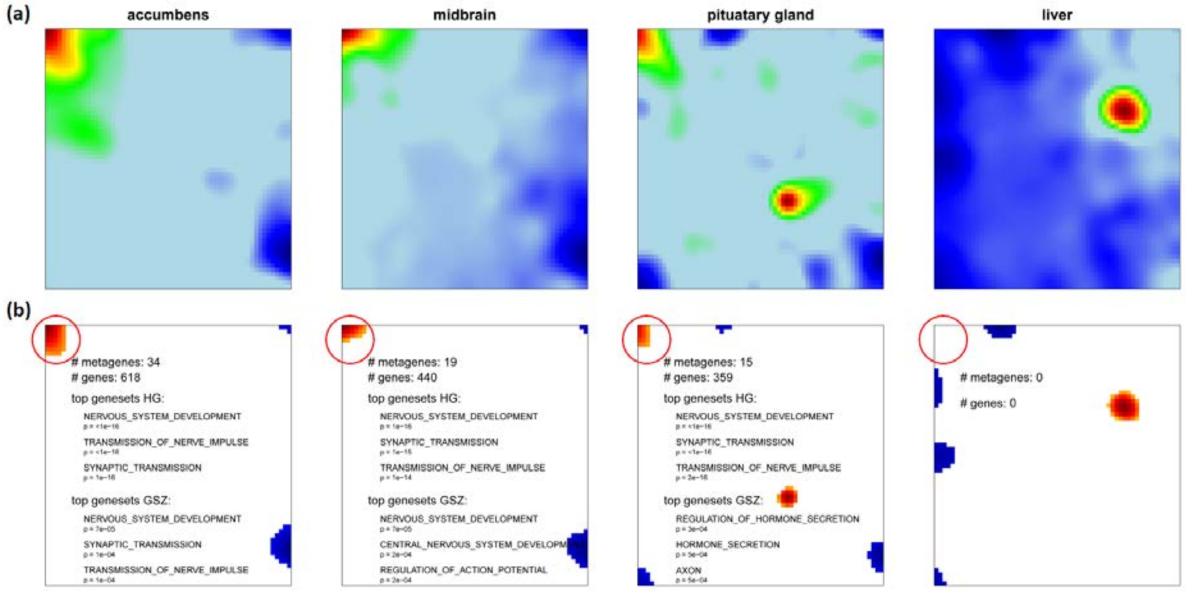


Figure 7: Local spot characteristics of the 'nervous' spot A in different tissues. Panel a shows the original expression profile of selected tissues and panel b the selected overexpression spot(s) by applying the 98% quantile criterion to the metagenes (red color). Note that the spot size (# of metagenes) and consequently also the number of associated genes with spot A (red circle) changes from tissue to tissue affecting the results of enrichment analysis using either the HG- or the GSZ-scores: The top three gene sets are given for each of the examples.

The first option of HG-enrichment analysis simply substitutes the global spots by tissue specific ones. These local spots are determined individually for each tissue-specific SOM by applying the 98-percentile threshold. The size of one particular spot usually varies from tissue to tissue and it can even disappear if the expression values of the respective metagenes do not meet the threshold criterion as illustrated in Figure 7 for the 'nervous tissue'-spot A (see also [6] for the full set of SOM images). In consequence, the spot-related lists of single genes and the derived list of overrepresented gene sets vary between the different samples. Subsequent application of overrepresentation analysis based on the HG-distribution (Eq. (8)) to these local spots provides tissue-specific p-values and thus one list of overrepresented gene sets for each of the spots in each of the samples.

We selected the top-three gene sets per spot in each tissue and merged them into one global list of most enriched gene sets in all spots. Finally, this global list was converted into the HG-enrichment heatmap shown in Figure 8a. We applied hierarchical clustering to group similarly expressed gene sets in vertical direction. It reveals five to six gene sets associated with the 'nervous tissue'-spot A in a tissue-specific fashion. Other groups of enriched gene sets can be associated with immune systems tissues (F), muscle tissues (B), epithelial (D) and homeostasis tissues (C1). The selected gene sets are listed in Table 1. Please note that we chose the same capital letters as labels as for the spot assignments discussed above for sake of comparison (see Figure 6a).

2.7. GSZ-enrichment analysis

HG-enrichment analysis applies a binary 'included-or-not included' criterion to assess the positive membership of the genes from a gene set in a selected spot-cluster. The gene set Z (GSZ)-score (Eq. (10), see the Methods section below) provides an alternative, second option for enrichment analysis which explicitly considers the individual expression values of the genes included in the list. The algorithm of GSZ-enrichment analysis is largely identical with that of HG-enrichment analysis; namely it starts with the tissue-specific identification of overexpression spots in the respective SOM-images followed by the identification of spot- and tissue-specific lists of gene sets and their aggregation into one global lists using the top-three gene sets from each individual list. The only difference refers to the expression-dependent GSZ-score (Eq. (10)) which is used instead of the expression-independent HG-score (Eq. (8)).

Figure 8b shows the GSZ-enrichment heatmap obtained from the aggregated list of all relevant spots. The obtained number of 64 gene sets exceeds the 48 gene sets in the HG-enrichment map in Figure 8a indicating the increased diversity of the GSZ approach. It can be adjusted by using stricter or more lax thresholds in the GSZ- and/or HG-mappings for the number of selected top-gene sets per spot, respectively. Both heatmaps reveal clusters of molecular characteristics which can be clearly assigned to selected tissue types, e.g. nervous processes to nervous tissues (cluster A in Figure 8) and muscle-related function to muscle tissues (cluster B). Table 1 lists the HG- and GSZ-enriched gene sets associated with the main spots.

In Additional file 3 we further disentangle the obtained GSZ-lists for the three spots selected in the bar plots in Figure 6b to illustrate the specifics of GSZ-enrichment analysis. Our standard algorithm applies the 'top-three' criterion, i.e. it selects the three top gene sets of each local spot list and merges them into the global list of gene sets which is further used to characterize the functional context of gene expression in the different tissues. This approach equally weights each spot in terms of the number of selected gene sets and thus it ensures that each spot-feature is equally represented in the resulting global list. Alternatively one can generate a global list of gene sets ranked according to their significance of enrichment in each of the tissues and cut this list using appropriate criteria. Results of this approach are presented in Additional file 3. The enrichment lists are very similar compared with those obtained from the 'top-three' selection criterion.

In summary, HG- and GSZ-enrichment maps based on the 'top-three' selection criterion provide a suited overview of the gene sets most important in the experimental series studied. For the more detailed analysis we recommend using full lists of gene sets for each spot which are provided as additional material in the spot-reports as described below.

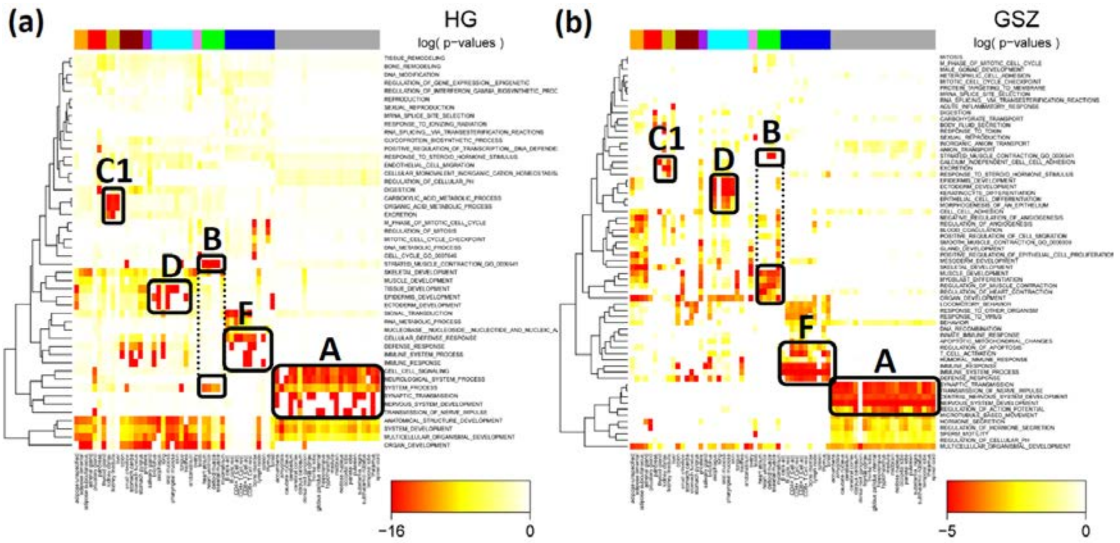


Figure 8: One-way hierarchical clustering heatmap of significantly enriched gene sets (rows) versus tissues (columns) using the HG- (a) and the GSZ- (b) statistics. The three-top gen sets per overexpression spot are selected in each of the maps. The heatmap color-codes the p-values of the respective score in log-scale (see the legends in the figure). The tissue categories are color-coded in the bar above the heatmap according to the assignments given in [6]. The gene sets are clustered in vertical direction. The capital letters approximately assign clusters of enriched gene sets in correspondence with the spots selected in Figure 6a and Table 1. The GSZ-score provides a larger number of gene sets (factor 1.8) and thus a more diverse pattern.

Table 1: Molecular characteristics of selected overexpression spots as obtained by HG- and GSZ-enrichment analysis [a]

spot	GSZ	HG
A	**Synaptic Transmission** **Transmission of Nerve Impulse** Central Nervous System Development **Nervous System Development** Regulation of Action Potential	Cell-Cell Signaling Neurological System Process **Synaptic Transmission** **Transmission of Nerve Impulse** **Nervous System Development**
B	Muscle Development Myoblast Differentiation Regulation of Muscle Contraction Regulation of Heart Contraction **Striated Muscle Contraction**	**Striated Muscle Contraction** System Process
C1	Carboxylic Acid Metabolic Process Organic Acid Metabolic Process **Excretion**	Calcium Independent Cell-Cell Adhesion **Excretion** Response to Steroid Hormone Stimulus
D	**Epidermis Development** **Ectodermis Development** Keratinocyte Differentiation Epithelial Cell Differentiation Morphogenesis of an Epithelium	Tissue Development **Epidermis Development** **Ectodermis Development**
F	Regulation of Apoptosis T-Cell Activation Humoral Immune Resonse **Immune System Process** **Immune Response** **Defense Response**	Cellular Defense Response **Defense Response** **Immune System Process** **Immune Response**

[a] Gene sets enriched in both approaches are printed in bold letters.

2.8. Overexpression maps and profiles of selected gene sets

In the previous subsections we applied ‘spot-centered’ gene set enrichment analysis to extract the most relevant functional gene sets in each tissue sample. One can also pursue a ‘gene set-centered’ approach and map the overrepresentation of one selected gene set in each tissue-specific mosaic image. Particularly, we estimate the degree of overrepresentation of this gene set in each metagene minicluster using the hypergeometrical (HG-) distribution. It provides an overrepresentation p-value for each metagene and each gene set considered. Then the distribution of p-values is visualized in the same two-dimensional mosaic which was used for the original expression images. Figure 9 shows overrepresentation maps of gene sets selected from each spot in Table 1. Overrepresentation is observed in different regions of the map, for example in the top left and bottom right corner for genes related to ‘synaptic transmission’ and to ‘immune system process’, respectively. The examples also show that overrepresentation is either strongly localized in one region of the map (e.g. for ‘striated muscle contraction’ or, to a less degree, for ‘synaptic transmission’ and ‘immune system process’) or it spreads over wider areas of the SOM (e.g. for ‘transmission of nerve impulse’). Note that this overrepresentation map applies to all samples studied owing to the fixed gene composition of the metagene clusters.

One can also apply an orthogonal approach to characterize the ‘enrichment’ profile of a selected gene set in all tissues studied. Our approach makes use of the full list of genes and calculates the GSZ-score for the gene set of interest in all tissues. In this special case the GSZ-score estimates overexpression in terms of the normalized difference between the mean expression averaged either over the gene set of interest and over the full list of genes (see Eq. (15)). The bar plots in Figure 10 show overexpression profiles of the selected gene sets. The gene sets are strongly and consistently overexpressed in different tissue categories. For example, the profiles of ‘synaptic transmission’ and ‘transmission of nerve impulse’ are strongly overexpressed in nervous tissues and underexpressed in virtually all non-nervous tissues. Contrarily, ‘immune system process’-genes show a more heterogeneous expression pattern in the non-nervous tissues with ‘local’ over- (especially in immune systems tissues) and underexpression characteristics while remaining strongly underexpressed in the nervous tissues. Genes related to muscle contraction are naturally overexpressed in muscle tissues but also in tongue which also contains muscle tissue. Note also that the gene set ‘epidermis development’ is overexpressed in epidermal tissues and in tonsil assigned to tissues of the immune system.

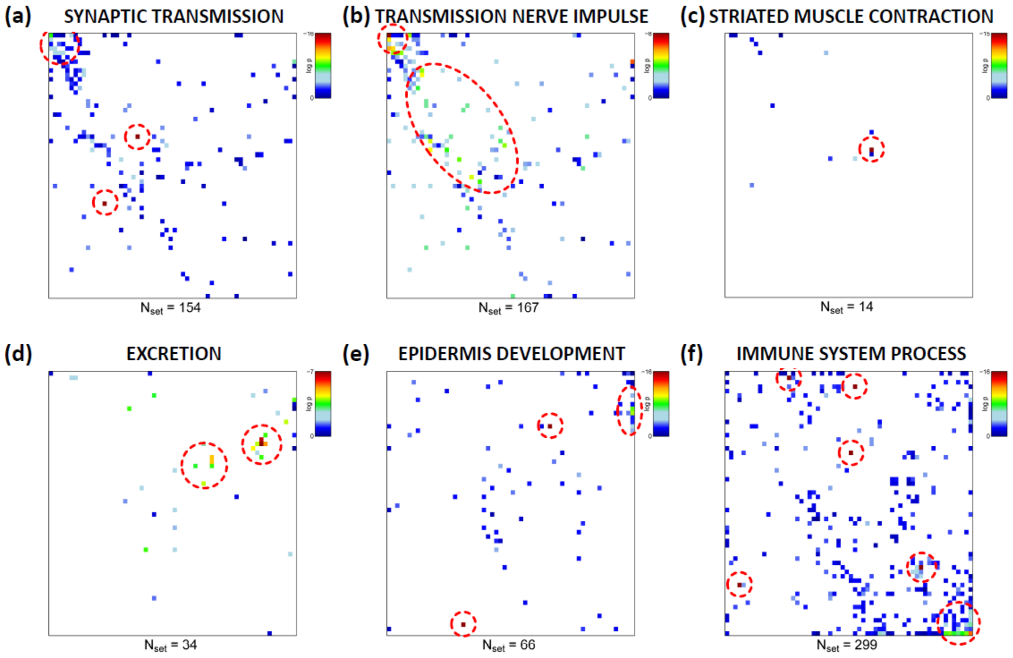


Figure 9: Overrepresentation maps of six selected gene sets containing between N_{set}= 157 and 472 genes. Overrepresentation in each tile of the mosaic is calculated in units of $\log(p_{HG})$ using the hypergeometrical distribution and color-coded (maroon>red>yellow>green>blue). White areas indicate 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).

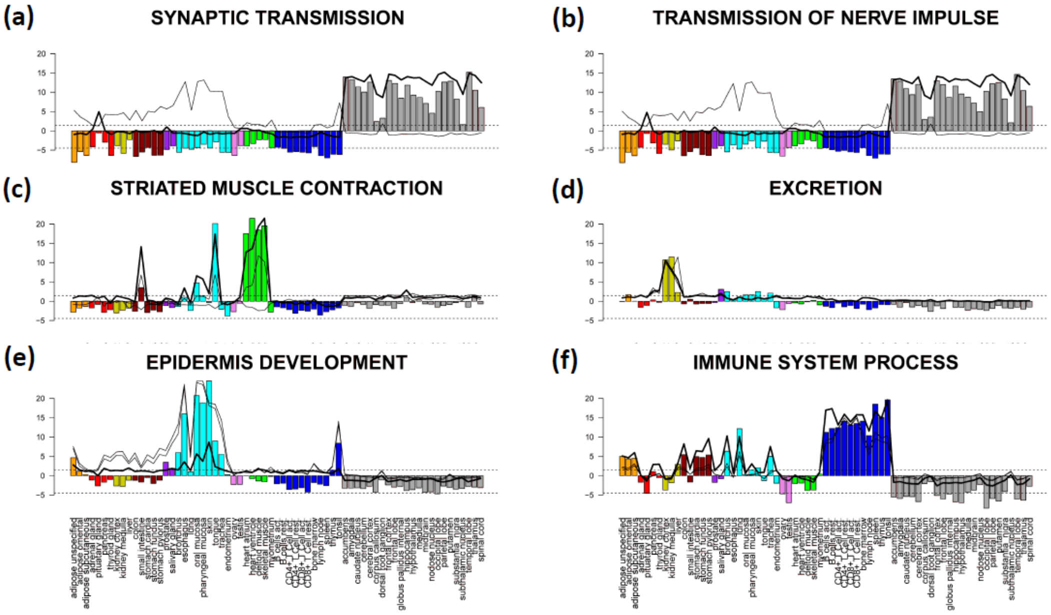


Figure 10: Overexpression profiles of selected gene sets (bar plots, compare with Figure 9). The bars are colored in accordance to the color-codes of the tissue categories introduces in ref. [6]. They are scaled in units of the GSZ-score (left axis). The horizontal dotted lines mark the fdr=0.2 significance threshold estimated from the p-value distribution of the GSZ-score. The inserted curves show the logged FC-expression profiles of the top-three metagenes of strongest enrichment of the respective gene set.

The curve plots inserted in all panels of Figure 10 show the expression profiles of the topmost three enriched metagenes containing the respective gene set. Most of these metagene expression profiles are very similar compared with the respective GSZ-overexpression profiles. Hence, representative profiles of the selected metagene miniclusters of co-regulated real genes well agree with the expression profiles of functionally related sets of genes which have been collected independently. This result supports the ‘guilt-by-association’ principle which states that coexpressed genes are likely to be functionally associated because biological processes are governed by coordinated modules of interacting molecules [21].

The ‘guilt-by-association’ principle, in turn, implies the ability to define either new gene sets using selected metagene-miniclusters or to verify and/or to amend existing ones. Such verification can address the distribution of the single genes of a selected gene set over different regions of the SOM (see, e.g. Figure 9) to prove their set membership by independent methods. On the other hand, spot-members not assigned to any gene set constitute potential new candidates for those gene sets which are highly enriched in the respective spot. For example, the tissue specific spots A (nervous system tissues), B (muscle tissues) and F (immune system tissues) contain about 30% - 40% genes which are not assigned to any of the gene sets tested and about 50% genes which are members of gene sets not listed at the top of the list (Table 2). These genes constitute potential candidates for further verification of their functional context.

Based on our spot analysis we define tissue-specific gene sets. Spots are selected which can be clearly assigned to selected tissue categories. The single genes of each spot are filtered using a correlation threshold for mutual correlations between the single gene and metagene profiles: Only genes are considered with Pearson correlation coefficient larger than 0.8. The defined gene sets are available in Additional file 4.

Table 2: Assignment of genes in selected spots to functional gene sets

Spot [a]	Total [b]	Primary gene sets [b]	Other gene sets [b]	Not in gene sets [b]
	# of genes	% in enriched sets	% in other sets	% not assigned
A nervous system	445	22%	40%	38%
B muscle	229	22%	49%	29%
F immune system	1558	13%	52%	35%.

[a] Spots are assigned in correspondence with Figure 6a.

[b] total number of genes in the respective spots which decompose into genes with membership in the top-three HG-enriched gene sets (see also Figure 6a), genes with membership in at least one of the remaining gene sets tested and genes without membership in any of the gene sets

2.9. Zoom-in analysis

We applied so-called ‚zoom-in' SOM analysis to study the expression profiles of subgroups of samples such as nervous and immune system tissues with enlarged resolution as described previously [6]. The zoom-in maps were trained using reduced sets of tissue samples but the same number of tiles of the SOM-mosaic. They show 'new' textures of characteristic over- and underexpression spots which reflect the expression profiles of the tissues of interest more in detail than the original SOM. In the supplementary material (Additional file 3) we present the results of global overrepresentation and of local GSZ-enrichment analysis applied to the respective subgroups of tissues. The zoom-in analysis of nervous tissues, for example, provides clusters of genes related to signal transduction and replication which are not clearly detected in the original maps. Both approaches, global overrepresentation and local GSZ-enrichment analysis, provide consistent results. In the addional material we provide also overrepresentation maps and overexpression profiles of the same gene sets shown in Figure 9 and Figure 10, respectively, to illustrate re-distribution of gene sets after zoom-in.

2.10. SOM-mapping of strongly expressed, absent and housekeeping genes

The gene sets studied in the previous subsections are chosen from GO-categories. They are subsequently processed to estimate their enrichment in overexpressed spot-clusters of co-regulated genes taken from the SOM mosaics. Gene sets can also be collected by applying alternative criteria such as the consistent high or weak expression of the selected genes in all samples. The population mapping of these sets into the SOM mosaic then specifies the activity of the respective genes in different areas of the map. Gene function of these sets can be specified using GO-overrepresentation analysis as described above. However, such global expression criteria itself lend to define groups of genes related to specific functions such as housekeeping gene activity. Housekeeping genes are thought to be by nature significantly expressed in all somatic cells under all circumstances because their gene products are required for the maintenance of basal cellular function (see, e.g., [32], [33] and references cited therein). In addition to housekeepers we select special sets of highly expressed (using differential expression and ranking criteria) and of absent (i.e. consistently not or weakly expressed genes) to obtain information about additional aspects of genomewide transcriptional activity which complements the functional analysis of tissue-specific overexpressed and co-regulated gene sets discussed above (see Table 3 for an overview; the genes of these sets are given in Additional file 5).

We analyze the SOM population patterns, the tissue-wide overexpression profiles and also GO-set overrepresentation of these special gene sets. Figure 11 and Figure 12 show the population maps of these gene sets and their GSZ-overexpression profiles, respectively. Highly expressed genes were selected by taking the top-10% genes either from the global overexpression list (panel a) or from the global rank product list (b, see Additional file 3 for details and also [34]). These criteria select genes either from a larger number of overexpression spots (e.g. spots A, C, D, H; compare Figure 11a and b and Figure 6) or from only a few ones (Figure 11b). Note that only about one fourth of the genes in each of the sets are commonly found in both sets due to the different criteria which select either maximum expressed genes or consistently top ranked genes. The overexpression profiles in Figure 12 (panel a and b) reveal that the rank criterion (b) more strongly weights highly expressed genes from nervous tissues than the alternative high expression criterion (a). On top of the HG-overrepresentation lists one finds gene sets related to homeostasis for high expression (a) and to morphogenesis and cell migration for consistently highly ranked genes (b) (see Table 3).

The expression of 'absent' genes per definition falls below the detection threshold for specifically hybridized probes in the microarray measurement. One can detect the respective genes using two different but closely related criteria (see rows c and d in Table 3). The first one extracts these genes directly after single-array intensity calibration using the hook method [35], [36] whereas the second one is based on the present-call parameter of each gene which was obtained after applying background correction and chip-to-chip normalization to all arrays of the series (see the methods section in [6] for details). The latter criterion selects about twice as much genes as the former one with only moderate overlap between both groups (Table 3 and Figure 11). Both criteria however provide very similar characteristics of absent and weakly expressed genes despite these differences (see panels c and d in Figure 11 and Figure 12): the genes selected strongly accumulate within one localized area near the centre of the SOM which has been assigned to virtually invariant genes in the respective summary map (see also the variance map in [6]). The GSZ-profiles support this result: They show relatively constant profiles for these sets which contain enriched populations from GO-sets related to receptor activity and signal transduction (Table 3).

Table 3: Special gene sets

	Gene set [a]	Selection criterion	# of genes	Top three overrepresented GO-sets [b]
a	Highly expressed	Top ranked expression in the global overexpression list	2,227 (10%)	Cation homeostasis, chemical homeostasis, multicellular organismal development
b	Highly ranked	Top ranked in the global rank product list [c]	2,227 (10%)	Anatomical structure morphogenesis, axonogenesis, cell migration
c	Inactive (consistently not or weakly expressed)	Member of the N-range of the hook curve, absent in all tissues	688	Receptor activity, signal transduction, plasma membrane
d		Present call parameter pc=0 in all tissues	1,156	Receptor-protein signaling pathway, neurological system process, signal transduction
e	Housekeepers (consistently expressed)	Not member of the N-range of the hook curve, present in all tissues	3,561	Anti-apoptosis, apoptosis, cell development, RNA processing, DNA/RNA binding, DNA metabolic process, metabolic process, transcription, translation [d]....
f		Present call parameter pc=1 in all tissues	3,167	see e
g		Top ranked in mean expression list averaged over all tissues	2,227 (10%)	Macromolecular complex assembly, nucleic acid metabolic process, regulation of cellular metabolic process
h		Taken from ref. [32], criterion analogous to d	852	Cellular macromolecule metabolic process, cellular protein metabolic process, protein metabolic process

[a] gene lists are given in Additional file 5

[b] HG-enrichment, lists are given in Additional file 5

[c] details are given in Additional file 3

[d] about 150 gene sets (see Additional file 5 and Table 4)

The criteria e and f (Table 3) essentially invert the previous selection of absent genes. They select genes which are significantly expressed in all tissues studied. These genes widely distribute over different regions of the SOM mosaics forming several highly populated 'hot spots' (see panel e and f in Figure 11). Spots of high tissue specificity are virtually not selected by these criteria as expected (compare with Figure 6). Interestingly, these consistently present genes are overexpressed in immune system tissues and underexpressed in nervous tissues, a pattern which basically inverts the respective profiles of the highly expressed genes in these two tissue categories (compare e and f with a and b in Figure 12).

Criteria e and f essentially meet the conditions for housekeeping genes (see above). We applied an alternative criterion which chooses 10% of the genes of highest mean expression log-averaged over all tissues. Most of the genes selected are common members also in the sets e and f. These three sets consequently possess very similar characteristics (see Figure 11 and Figure 12). For comparison we included a list of housekeepers taken from a previous microarray study [32]. The respective selection condition essentially agrees with our criterion d. However it was applied to an alternative tissue data set which was studied using a previous generation of HGU95a- GeneChip arrays [37], [38]. We reanalyzed this data set and found that it contains a much higher fraction of absent genes in most of the tissues (data not shown). This difference presumably explains the relatively small number of housekeepers detected in this data set. Despite this difference it reveals a similar overexpression profile compared with our alternative sets.

HG-overrepresentation analysis of the housekeepers provides functional gene sets related to basal cell activity such as ‘metabolic process’, ‘transcription’, ‘translation’ and ‘RNA processing’. Note that the housekeepers distribute over several separated spot-like areas in the SOM mosaic which partly contain enriched fractions of the same gene sets such as ‘cytoplasm’ found on top of gene set lists in the spots h1-3, 5, 11 (see Table 4). Other gene sets accumulate in single or only a few spots only, for example ‘nucleus’ in h3, h9 and h10; ‘mitochondrion’ in h4 and ‘lipid binding’ in h5. The SOM approach thus enables to further disentangle larger groups of genes such as housekeepers into subgroups of more specific function. For example, housekeepers related to nucleic acid processing accumulate in spots h7, h9 and h10 whereas genes related to actin functioning in h2. Note also that the spots of housekeepers discussed are still located in regions of relatively highly variable and thus specific metagene profiles (compare with the variability map given in [6]).

In conclusion, global expression criteria represent an alternative option for selecting metagenes and spots of metagenes with functional impact. These criteria complement the overexpression criteria discussed above. Note for completeness that both options can be combined, for example, to mask absent genes in the overexpression SOM to exclude noisy and thus presumably irrelevant genes.

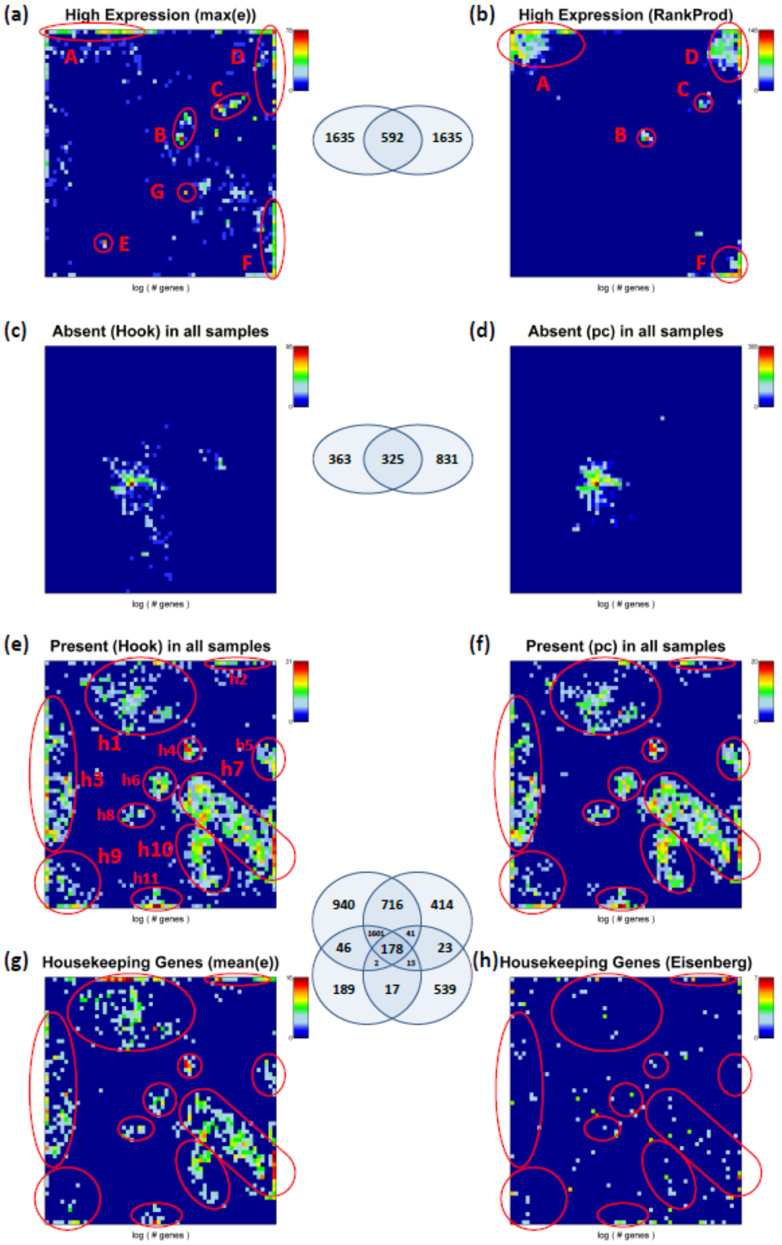


Figure 11: Population maps of special gene sets: Genes of highest expression (top 10%) preferentially accumulate in a few metagenes in spots A – F (spots are assigned in agreement with Figure 9) whereas the consistently absent genes (~3-5% of all genes) are found in the area of minimum variability (see variability map in [6]). Housekeeping genes selected as consistently present in all tissues (not-absent, ~15% of all genes) and as the top 10% most stable expressed genes are compared with the set of housekeeping genes taken from ref. [32]. The gene sets enriched in selected highly populated spots (h1 – h11) are given in Table 4. The Venn diagrams show the overlap between different gene sets as illustrated.

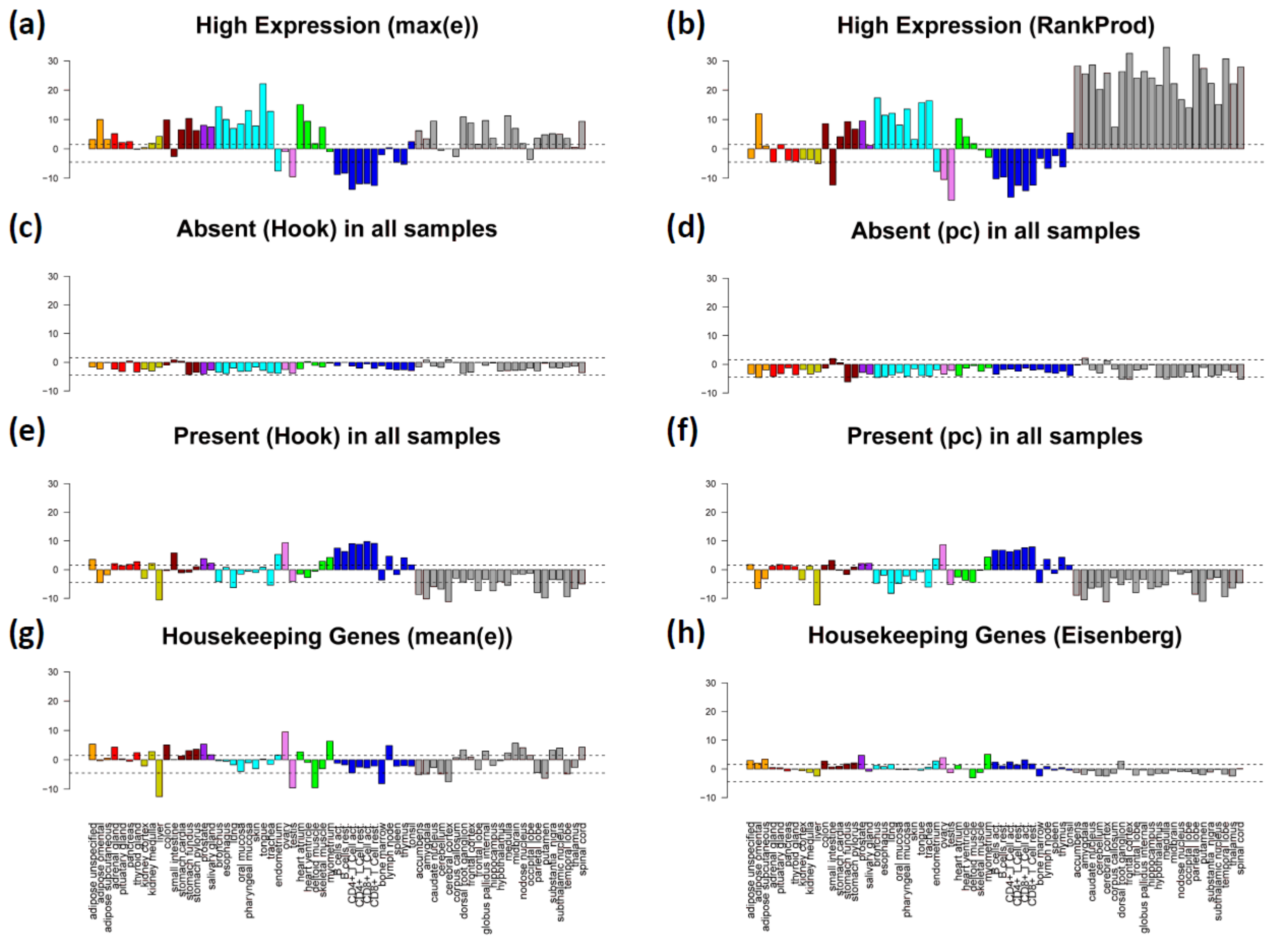


Figure 12: GSZ-overexpression profiles of the special gene sets defined in Table 3.

Table 4: GO-overrepresented gene sets in SOM-spots of highly populated housekeeping metagenes

spot[a]	# of genes	Top overrepresented gene sets
h1	333	Cytoplasm, enzyme regulator activity, vesicle mediated transport, establishment of localization
h2	74	Cytoplasm, oxidoreductase activity, actin binding, endoplasmic reticulum, cytosol
h3	418	Cytoplasm, macromolecular complex, nucleus, protein metabolic process, protein complex
h4	89	Oxidoreductase activity, cytoplasm, mitochondrion, envelope, organelle
h5	91	Cytoplasm, Golgi apparatus, cofactor catabolic process, lipid binding, microsome
h6	101	Protein complex, macromolecular complex, cytoplasm, protein catabolic process
h7	775	Biopolymer metabolic process, biosynthetic process, nucleic acid, RNA processing
h8	50	Protein metabolic process, endosome, cellular metabolic process, phosphatase activity
h9	176	Nucleus, biopolymer metabolic process, nucleic acid / RNA metabolic process
h10	253	Biopolymer metabolic process, mRNA metabolic process, RNA processing, nucleus
h11	118	Cytoplasm, proteasome complex, cellular protein metabolic process, protein metabolic process

[a] spots are defined in Figure 11e

2.11. Reports

Our SOM approach enables views from different perspectives on large sets of high dimensional data. They include overview characteristics which address similarity relations between different samples and the detailed description of the expression pattern in each of the samples studied as well. Moreover, differential expression analysis identifies ordered lists of over- and underexpressed genes taken either from the full ensemble of all genes available or from subensembles selected from metagene spots of co-regulated genes. Information about the functional context is extracted by applying enrichment analysis to the different gene lists.

We presented the basal frame of SOM-based data mining in this and the previous paper [6]. Results are discussed with the focus on methodical issues such as different options of data presentation and analysis. Selected examples taken from the tissue data set are studied to illustrate the capabilities of different aspects of the approach. The methods are implemented in the R-program 'oposSOM' available as CRAN package via http://cran.r-project.org/.

We applied these methods to the full set of 67 tissues to obtain the systematic, comprehensive and detailed characterization of the transcriptome of human tissues as seen by GeneChip microarrays. Our study produced an extensive collection of results such as various SOM expression profiles, global and spot-related gene lists, GO-gene set enrichment data and metagene-based cluster plots obtained from agglomerative analyses. Only selected results are presented here as illustrative examples to explain and to illustrate different aspects of the method.

We designed a set of standard PDF-reports which allows the systematic browsing in the full set of results. The whole report is organized into several main topics each of them contains a series of documents for download from our website (http://som.izbi.uni-leipzig.de):

(i) Maps (experiment atlas)

These reports show the collection of first level SOM of all tissue samples, supporting maps and the second level SOM as described in ref. [6]. First level SOM profiles are shown with different contrast (log FC-, WAD- and double log-scale) and also as rank-maps using the different scores as described above.

(ii) Metagene and enrichment analysis

Several agglomerative methods based either on distance or on correlation metrics are applied to the samples using filtered subsets of metagenes. The reports show the respective two-way hierarchical clustering heatmaps, pairwise correlation maps, minimum spanning trees and the results of independent component analysis [6]. In addition, GSZ- and HG-enrichment clustering heatmaps are available to associate the most relevant functional gene sets with the different samples. This information is supplemented by the respective p-value distributions to assess the quality of the data.

(iii) Spot summaries

These analyses apply different criteria of spot selection such as overexpression, underexpression, maximum and minimum of metagene expression and mutual correlations between the metagenes as described in ref. [6]. GO-enrichment analysis provides the three leading genes in the respective HG-enrichment list of each of the spots considered. Spot-related heatmaps characterize the expression profiles of the selected features in the series of samples. Single spot summary sheets provide detailed information about each of the spots such as the ranked list of samples which overexpress this feature according to the mean t-shrinkage statistics of the spot and the ranked list of the top-twenty HG-overrepresented gene sets together with the histogram of the respective p-value distribution (Figure 13a).

(iv) Sample summaries

For each sample we generated one PDF-report which summarizes the most relevant information using the global (i.e. sample-centered) and local (i.e. spot-centered) perspective as well. The global summary shows the ranked list of differentially expressed genes together with the associated significance characteristics for the whole sample, the ranked list of over- and underexpressed gene sets after GSZ-overexpression analysis and the respective p-value distributions (Figure 13b). The local summary sheets present the analogous information for each single spot which is selected using the 98%-quantile criterion. The two maps in the left part of the sheet show the respective first level SOM and the selected spot, respectively. The full global and local lists can be downloaded in excel format for detailed inspection and further processing. We also present the locally pooled error (LPE) characteristics for each sample to judge its quality.

(v) Cross-tissue enrichment and metagene expression profiles

Enrichment maps and profiles of individual GO gene sets are shown as bar plots and provided as excel-files for download (Additional file 6). These cross-tissue characteristics are supplemented by the log FC-expression profiles of the leading metagenes of the respective gene set.

This compendium of gene expression characteristics in human tissues supplements previous data collections obtained with alternative arrays, sample sets and methods of analysis [37-39].

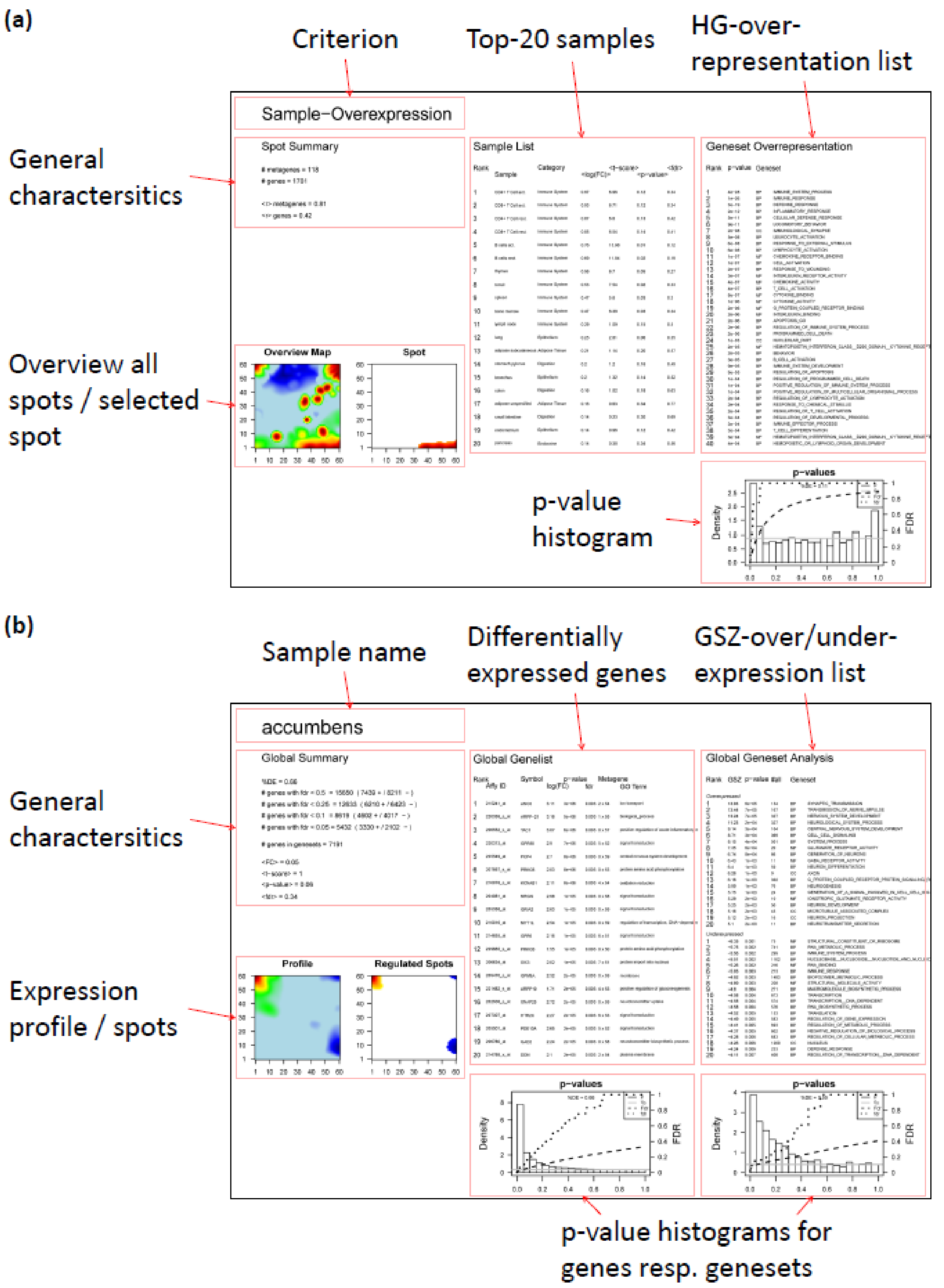


Figure 13: Single spot (panel a) and single sample global (panel b) summary report sheets (see also Additional file 3 for more details).

3. Summary and conclusions

SOM machine learning transforms large and heterogeneous sets of expression data into mosaic images which visualize tissue-specific over- and underexpression in terms of characteristic textures. This view is very intuitive to identify modules of correlated and differentially expressed genes in terms of well defined colored spots. Thus, SOM analysis basically rearranges and classifies the primary information of single gene expression in a series of samples without filtering. It preserves the whole information content of the original data set despite the dimension reduction used to visualize the most essential expression profiles inherent in the data.

This primary information together with the respective gene annotations is further processed in differential expression analysis using three alternative scores which place emphasis either exclusively on the fold change of gene expression or, in addition, on the precision of the measurement. It is taken into account either by the simple down-weighting of the impact of low expression values or by applying a regularized t-score. It explicitly considers the standard error of the expression values as the combination of individual and locally pooled error estimates. The latter error-pooling approach confirms the inverse relation between the magnitude of the logged expression and its significance. Significance of ranked gene lists is controlled by the local false discovery rate using standard methods.

SOM analysis provides special advantage to generate local lists of genes taken from selected spots of the map. Thus, the impact of differential expression can be studied not only in a sample-specific fashion but also for selected subgroups of co-regulated genes. The alternative scores studied provide slightly different but at the end consistent rankings for lists containing a few dozen or more genes. The FC-, WAD- and shrinkage t-scores are judged rather as complementary measures than as competitive ones providing information which mutually supplements each other because of their specific advantages and disadvantages.

To extract the functional context of spot and metagene related lists of single genes we applied overrepresentation- and overexpression analysis, and a combination of both with respect to pre-defined gene sets of basically known functional impact. Overreprepresentation analysis combines the criterion membership in a gene set with that of co-(i.e. correlated-) expression in a series of samples whereas overexpression analysis compares the mean expression of genes from the set with that of all genes. 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. Alternatively, one can screen the degree of overrepresentation of a large number of gene sets in a selected metagene spot to discover its potential functional context. Both complementary views provide a link between the tiles and/or spots of the SOM mosaic and their potential molecular function. It applies to all samples of the series due to the fixed distribution of single genes in the mosaics.

Overexpression analysis of a selected gene set, on the other hand, profiles a selected molecular function across the different samples studied, for example, to identify tissues with highly active or inactive genes from the set of interest. The gene set enrichment approach combines both overrepresentation and –expresssion analysis. It was applied to discover the functional context of the metagene overexpression spots in a sample specific fashion by estimating significance using either the hypergeometrical statistics or the gene set enrichment Z-score with similar results in both cases. GSZ-enrichment however tends to select more diverse lists of gene sets because it explicit takes into account the expression profile of the associated genes in the different samples. The use of multiple options of ranking scores for differential expression and for gene set functional analysis enable to test the robustness of single gene and gene set rankings with potential consequences for their biological interpretation.

The tissue related spots of the SOM typically contain enriched populations of gene sets corresponding to molecular processes in the respective tissues in most cases. The representative expression profiles of the leading metagenes of the spots well agree with the expression profiles of gene sets functionally related to the respective tissues. This result strongly supports the ‘guilt-by-association’ principle that co-expressed 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 refine existing ones. This objective requires further study to judge the significance of these spot- or metagene-related sets using suited correlation or mutual information metrics. In addition to overexpression criteria for selecting SOM spots (given in units of expression differences) we study absolute ones (given in units of expression values) which allow identification of alternative sets of housekeeping genes and of consistently-high or -low expressed genes. Finally, we considered molecular function of the overexpression spots of the SOM after zooming–in.

Application of SOM-based analysis to the full set of 67 tissues provides the comprehensive and detailed characterization of the transcriptome of human tissues as seen by GeneChip microarrays. Our study produced an extensive collection of results which are provided as supplementary reports to illustrate the potency of the method and also as data base for further studies in the context of gene regulation in different tissues and its dysfunction. The methods of differential gene expression and enrichment analysis are implemented in the R-program ‘oposSOM’ available as CRAN package.

4. Data and Methods

4.1. Microarray data and SOM-cartography

The raw microarray data and their primary and secondary analysis in terms of calibration, normalization and SOM-cartography was described in [6]. In short: Gene expression profiles were downloaded from Gene Expression Omnibus under accession number GSE7307 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7307). The data set consists of 677 human tissue samples measured with the Affymetrix HG-U133 plus 2.0 array. We selected 187 of these samples derived from 67 different tissues for further analysis.

Microarray intensities were transformed into expression values, $E_{g,m,r}$, using hook calibration [35], [36] and quantile normalization. The indices assign the gene (g=1…N), the tissue (m=1…M) and the replicate (r=1…R_m) where the number of replicates can vary between the tissues. The logged expression values of each gene, $e_{g,m,r} \equiv \log_{10} E_{g,m,r}$, are averaged over the replicates, $e_{g,m} \equiv \left\langle e_{r,g,m} \right\rangle_r$ (angular brackets denote arithmetic averaging), and transformed into differential expression values, $\Delta e_{g,m} \equiv e_{g,m} - e_g$, with respect to the mean expression of each gene averaged over all tissues studied, $e_g \equiv \left\langle e_{g,m} \right\rangle_m$.

Subsequently self organizing maps (SOM) machine learning was applied to all differential expression data. The algorithm initializes K so-called metagene expression profiles. These profiles represent vectors of dimensionality M given by the number of conditions studied. Then a gene is picked from the gene list and its vector of differential expression $\Delta e_{g,m}$ is compared with the metagene profiles using the Euclidian distance as similarity measure. The metagene profile of closest similarity is then modified, so that it more closely resembles the expression profile of the selected gene. In addition, the neighboring metagene vectors in the two-dimensional grid closest to this metagene are also modified, so that they also resemble the gene's expression vector a little more closely. This process is applied to all genes and repeated about 250,000 times. The radius of considered neighbors is decreased with progressive iteration which modifies fewer metagene vectors by smaller amounts, so that the metagene vectors asymptotically settle down. The resulting map becomes organized because the similarity of neighboring metagenes decreases with increasing distance in the map. Each ‘single’ gene is assigned to the metagene vector of closest similarity.

The final SOM thus consists of regions of similar metagene profiles each of them represents a minicluster of single genes with similar expression profiles. The distance similarity metrics and the training algorithm used gives rise to a characteristic metagene spot pattern where spots of high-variable metagene profiles arrange near the edges of the map about a central region of less variable metagenes. The spots of highly variable metagenes differ by the particular sample in which they are over- and underexpressed. These sample-specific over- and underexpression spots are selected among all metagenes using the 98% and 2% quantile criterion, respectively. These spots collect sets of genes with highly correlated expression profiles.

In our particular application the method sorts the individual genes into K=60x60 miniclusters. Each minicluster is characterized by one metagene profile which is used for visualizing the expression pattern of each tissue in terms of an individual mosaic picture of characteristic texture showing distinct over- and underexpression spots.

4.2. Differential expression scores

A large multitude of various methods have been developed in the last decade to assess statistical significance of differential expression in microarray data analysis (see, e.g., the overview given in [7] and the references cited therein). Most statistical methods aim at generating ranked lists of single genes which are differentially expressed according to a certain level of significance. Microarray data are very noisy and prone to systematic errors [8-13]. The proper estimation of the level of precision constitutes therefore one basal problem in significance analysis, especially if only a few replicates are available. Another problem is raised by the highly multivariate character of the data which requires suited concepts to control significance in multiple testing.

In this study we estimated differential expression of individual genes using three alternative scores:

1. The fold change (FC) simply estimates the expression change in logarithmic scale, log $\mathrm{FC}_{g,m} \equiv \Delta e_{g,m}$.

2. The weighted average difference (WAD)-score,

$$\mathrm{WAD}_{g,m} = w_{g,m} \cdot \Delta e_{g,m} \quad \text{with} \quad w_{g,m} = \frac{\Delta e_{g,m} - \min\left(\Delta e_{g,m}\right)}{\max\left(\Delta e_{g,m}\right) - \min\left(\Delta e_{g,m}\right)}, \tag{1}$$

is a fold-change based measure well performing in differential expression analysis [24], [25]. The main idea behind the WAD method assumes that relevant marker genes tend to have high expression levels, i.e. ‘strong signals are better signals’ in the gene ranking problem [8], [26], [40]. This assumption accounts for the fact that the experimental error of expression values inflates at small expression levels in logarithmic scale [41-43]. Note that the weighting factor in Eq. (1) can be expressed as a function of the absolute expression values as in the original paper of Kadota et al. [24], $w_{g,m} = \left(e_{g,m} - \min\left(e_{g,m}\right)\right) / \left(\max\left(e_{g,m}\right) - \min\left(e_{g,m}\right)\right)$, showing that the weighting factor linearly scales with the expression level of the gene.

3. The shrinkage t-score,

$$t_{g,m} = \frac{\Delta e_{g,m}}{\mathrm{SE}^{\mathrm{diff}}_{g,m}} \quad \text{with} \quad \mathrm{SE}^{\mathrm{diff}}_{g,m} = \sqrt{\frac{\left(\sigma^{\mathrm{shr}}_{g,m}\right)^2}{R_m} + \frac{\left\langle \left(\sigma^{\mathrm{shr}}_{g,m}\right)^2 \right\rangle_m}{\sum_{m=1}^{M} R_m}} \approx \frac{\sigma^{\mathrm{shr}}_{g,m}}{\sqrt{R_m}}, \tag{2}$$

accounts for the standard error of the expression values of each gene in replicated measurements. Our shrinkage statistics was defined in Eq. (2) in analogy with previous approaches [44-46]. Here $\mathrm{SE}_{g,m}{}^{\mathrm{diff}}$ denotes the standard error of differential expression of gene g measured under condition m. To estimate the standard error in Eq. (2) we first calculate the standard deviation of the log-expression values using the available replicates, $\sigma_{g,m} \equiv \sqrt{\left\langle \left(e_{r,g,m} - e_{g,m}\right)^2 \right\rangle_r}$. These values are then plotted for each sample as a function of the logged expression degree, $e_{g,m}$, and locally pooled over a moving window of a few hundred neighboring values. The obtained locally pooled error (LPE) estimates the mean standard deviation as a function of the expression, $\sigma_{\mathrm{LPE}}(e_{g,m})$. It is combined with the individual standard deviation for each gene to provide the shrinkage error estimate used in Eq. (2)

$$\sigma^{\mathrm{shr}}_{g,m} = \sqrt{\lambda \cdot \sigma_{g,m}{}^2 + (1-\lambda) \cdot \sigma_{\mathrm{LPE}}(e_{g,m})^2}. \tag{3}$$

The parameter λ ($0 \le \lambda \le 1$) scales the degree of shrinking $\sigma_{g,m}$ towards σ_{LPE}.

The shrinkage t-statistics was developed in the framework of James-Stein analytic shrinkage and applied in different modifications in gene expression analysis (see [44] and references cited therein). The basic idea behind Eq. (3) assumes that the error estimate based on $\sigma_{g,m}$ alone might be very imprecise, e.g. if only a few replicates are available. The resulting large ‘error of the error’ leads to highly uncertain naive t-scores associated with large false positives rates (see Eqs. (2) and (3) with λ=1).

It has been suggested previously that estimates of the variance from individual genes is questionable [26][39][47-50]. Yet accurately estimating variability of gene expression is essential for correctly identifying differentially expressed genes. Additional information may be gained by combining variance estimates across all or part of the experiment. Such information borrowing methods that exploit this information are able to improve the results [26], [47], [49]. Particularly, local-pooled-error (LPE) estimates for evaluating significance of each gene's differential expression have been shown to effectively identify significant differential expression patterns with a small number of replicated arrays [49].

To get more precise error estimates, the shrinkage t-score makes therefore use of the fact that the variability of microarray expression values is governed by methodical factors which allow to express the measurement error as a function of the expression level [43], [51]. This error can be estimated with high precision using the LPE averaging approach. Finally, Eq. (3) combines the pooled and the gene-specific error to take into account both, individual and common factors. Shrinkage t-scores consistently lead to accurate gene rankings which might outperform simple t-statistics or FC-scores [44].

4.3. Locally pooled error functions of single tissues

The LPE approach pools genes with similar expression values to estimate their variance with improved precision. It is justified by the observation that the experimental error of microarray expression values is governed by systematic factors caused by the physico-chemical principle of probe intensity detection. Particularly, the technical error of the measurement is a function of the expression degree [8], [26], [40], [47], [49] which can be derived using error propagation of the underlying hybridization isotherm [8], [43]. It predicts that the uncertainty of determining expression estimates inflates towards small expression levels due to the increasing contribution of the non-specific background and it progressively decreases towards large expression degrees due to saturation effects.

Indeed, the locally pooled error of the array data studied typically decreases with increasing expression degree (see Figure 14, Additional file 3 and Additional file 2 which shows the error functions of all tissues). The significance level of strongly expressed genes is consequently larger than that of weakly expressed genes for identical fold changes (see Eqs. (2) and (3) with $\lambda<1$). In addition, the mean error level averaged over all probes can markedly vary between the different samples. We proved two summary measures to quantify the mean error level of each sample:

a) the mean standard deviation averaged over all probes,

$$\langle\sigma\rangle = \sqrt{\left\langle\left(\sigma_{g,m}\right)^2\right\rangle_{\text{all probes}}} \quad \text{with} \quad \left\langle\left(\sigma_{g,m}\right)^2\right\rangle \equiv \frac{1}{N\cdot M}\sum_{\text{all } g,m}\left(\sigma_{g,m}\right)^2 \approx \int_{e_{min}}^{e_{max}} P(e)\cdot\left(\sigma_{LPE}(e)\right)^2\cdot de/\left(e_{max}-e_{min}\right),$$

and

b) the mean LPE-error,

$$\left\langle\sigma_{LPE}\right\rangle = \sqrt{\left\langle\left(\sigma_{LPE}(e)\right)^2\right\rangle} \quad \text{with} \quad \left\langle\left(\sigma_{LPE}(e)\right)^2\right\rangle \equiv \int_{e_{min}}^{e_{max}}\left(\sigma_{LPE}(e)\right)^2\cdot de/\left(e_{max}-e_{min}\right),$$

where e_{max} and e_{min} are appropriate integration limits of maximum and minimum expression values. Note that also $<\sigma>$ can be obtained by integrating over the LPE-error function to a good approximation (see the right part of the equation above), where P(e) is the normalized probability density to find a probe with expression e. Accordingly, the mean LPE-error equally weights the error function whereas the mean standard deviation in addition considers the population of the expression values. In consequence, the value of $<\sigma>$ is closer to the standard deviation of strongly populated background probes whereas $<\sigma_{LPE}>$ better reflects the error of specifically hybridized, more strongly expressed but less populated probes.

Both error measures strongly correlate with r=0.99 (Figure 14, panel a). In the remainder of this study we will use $<\sigma_{LPE}>$ as a characteristic measure of the mean level of scattering of the expression values between replicated samples. Group-averaging over the tissue categories (see [6] for details) reveals significant differences of their mean error level (Figure 14, panel b). For example, adipose tissues and tissues related to digestion show nearly twice as large gene-related error levels than tissues of sexual reproduction, of exocrine function and partly of homeostasis.

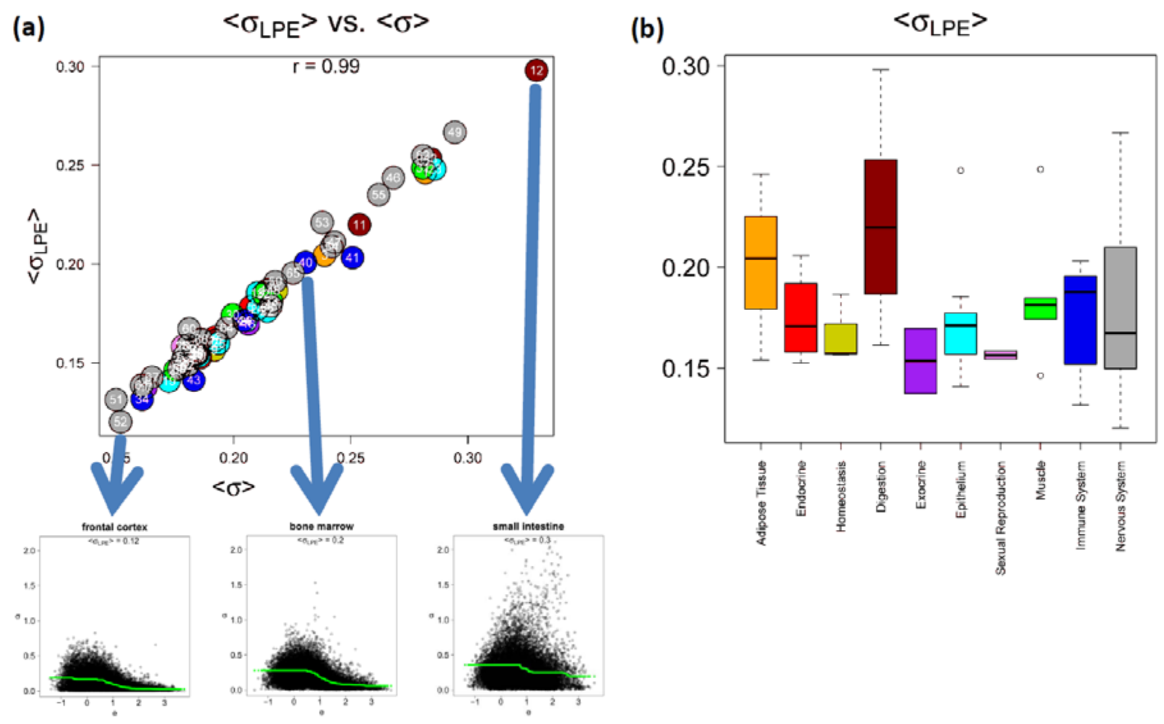


Figure 14: Tissue specific error levels of microarray expression data: Panel a) Comparison of <σ> and <σ_{LPE}> for all tissues studied reveals strong correlation with r=0.99. LPE error functions are shown for selected tissues referring to low, intermediate and high error levels: 'frontal cortex' (no.52, tissue numberings are assigned in [6]), 'bone marrow' (no.40) and 'small intestine' (no.12). Panel b) Boxplot of <σ_{LPE}> for different tissue categories defined in [6]. Note the different error levels. The color code is used also in the scatter plot (panel a). The error plots and <σ_{LPE}> values of all tissues are given in Additional file 2.

4.4. Significance analysis

The shrinkage t-statistics (Eq. (2)) transforms into p-values characterizing the significance of differential expression for each gene assuming Student's t-distribution. The obtained density distribution for the p-values of all genes in one selected tissue, ρ(p), meets the normalization condition $\int_0^1 \rho(p)\cdot dp = 1$. Examples for selected tissues of different mean error level are shown in Additional file 3. Under the null hypothesis one expects a uniform distribution, ρ_0(p)= 1, whereas the alternative hypothesis will produce a skewed distribution, ρ_{DE}(p), decaying with increasing p because differentially expressed genes tend to cluster closer to p=0 [52]. In the general case, the observed distribution can be interpreted as the superposition of two components due to differentially and not-differentially expressed genes, $\rho(p)= \rho_{DE}(p)\ (1-\eta_0) + \rho_0(p)\ \eta_0$, where η_0 is the fraction of non-informative 'null'-genes among all genes considered [52], [53]. It was derived using the "fdrtool" R-package [54] under the assumption of vanishing differential expression at p=1, $\rho_{DE}(1)=0$, giving rise to $\rho(1)=\eta_0$ [55]. "fdrtool" was further used to calculate false discovery rates (FDR) to control the number of false discoveries:

$$\mathrm{fdr}(p)=\frac{\eta_0}{\rho(p)} \quad \text{and} \quad \mathrm{Fdr}(p)=\frac{\eta_0\cdot p}{\int_0^p \rho(p)\cdot dp} \qquad . \tag{4}$$

Here fdr and FDR denote the local and tail area-based FDR estimates, respectively. The latter Fdr(p)-values provide a cumulative estimate of FDR referring to all genes on top of a list with p-values p'≤p whereas fdr(p) estimates the FDR of a selected gene with p'=p [56]. For a monotonically decaying total density ρ(p) both, fdr(p) and Fdr(p), are increasing functions which well correlate in the intermediate p range. The local FDR-estimate however systematically exceeds the tail-based one, fdr(p)≥ Fdr(p), at intermediate and large values of argument (see the examples shown in Additional file 3). Their limiting values at p=0 and 1 are given by the equations Fdr(0)= fdr(0), Fdr(1)= η_0 and fdr(1)= 1, respectively.

4.5. Non-informative and absent-called genes

The total fraction of differentially expressed and thus informative genes per sample can be estimated using the background level of the respective p-value distribution,

$$\%\mathrm{DE} = 1 - \eta_0 \quad . \qquad (5)$$

%DE decreases with increasing mean error level $<\sigma_{LPE}>$ and with increasing FDR at a selected p-value (p=const; see Additional file 3). In analogy to %DE we define %fdr (and %FDR) as the fraction of genes, the FDR-value of which falls below a given threshold, e.g. $fdr(p)<fdr_{threshold}$ for the local FDR-value. We arbitrarily chose $fdr_{threshold}=0.5 > Fdr_{threshold}=0.2$ where the latter relation ensures similar values of %fdr and %Fdr (see previous subsection). Both, %fdr and %Fdr, strongly correlate with each other (r=0.97; data not shown) and with %DE (see the %fdr-vs-%DE plot in Figure 15, r=0.98). The latter result indicates that %fdr is largely determined by the noise floor of non-informative probes whereas the slope of the decay of the p-value distribution near its left boundary has, if at all, an almost tiny effect. Note that both factors, the non-informative noise floor and the particular shape of the distribution of informative probes, can affect %fdr.

%DE (and %fdr) negatively correlates with the mean error level $<\sigma_{LPE}>$ (r= -0.79, Figure 15), i.e., a higher uncertainty of the expression measures is accompanied by a smaller number of differentially expressed genes on the average. This result reflects the fact that a larger uncertainty of the expression estimates effectively increases the fraction of non-informative probes which contribute to the null distribution only. Note that %DE more than halves from values about 0.7 to 0.3 if $<\sigma_{LPE}>$ increases from ~0.1 to 0.3.

On the other hand, $<\sigma_{LPE}>$ is related to absolute expression values whereas %DE refers to differential expression relative to a reference level. %DE of a particular tissue is consequently affected by its expression profile and by the respective noise floor. The expression level is governed by biological factors, e.g. by the tissue specifics of gene activity, whereas the noise level mainly depends on the precision of the measurement, which is affected by biological and methodical effects as well. Hence, the obtained correlation between %DE and $<\sigma_{LPE}>$ indicates that the precision of the expression measurement largely affects the number of detected differentially expressed features.

To further analyze the noise-level inherent in the data we included the fraction of absent-called genes (%N) in our correlation plot where %N is defined as the fraction of genes the expression of which falls below the detection threshold of the microarray measurement. It is determined separately for each chip in the calibration step [6], [35], [36]. Interestingly, %N does virtually not correlate with %DE (r=-0.04), however it moderately correlates with $<\sigma_{LPE}>$ (r=0.38), which, in turn, correlates with %DE (r=-0.79, see previous paragraph). The quantile normalization and scaling algorithms used transform the individual sample-specific density distributions of expression values into one common average distribution [6]. As a result the potential relation between %N and %DE gets mostly lost in this step presumably because also absent-called genes can differentially express in different samples. In consequence, %N essentially does not affect the differential expression estimates whereas it is directly related to the mean error level of each array. In turn, $<\sigma_{LPE}>$ affects %DE because it determines the significance of the differential expression values and thus %DE.

This somewhat puzzling relation between the error measures considered shows that data transformation after preprocessing and normalization can mask mutual relations. Most importantly, the number of differently expressed genes meeting a given significance criterion is governed by the error level of the expression measures which, in turn, systematically varies between the different tissues and tissue types.

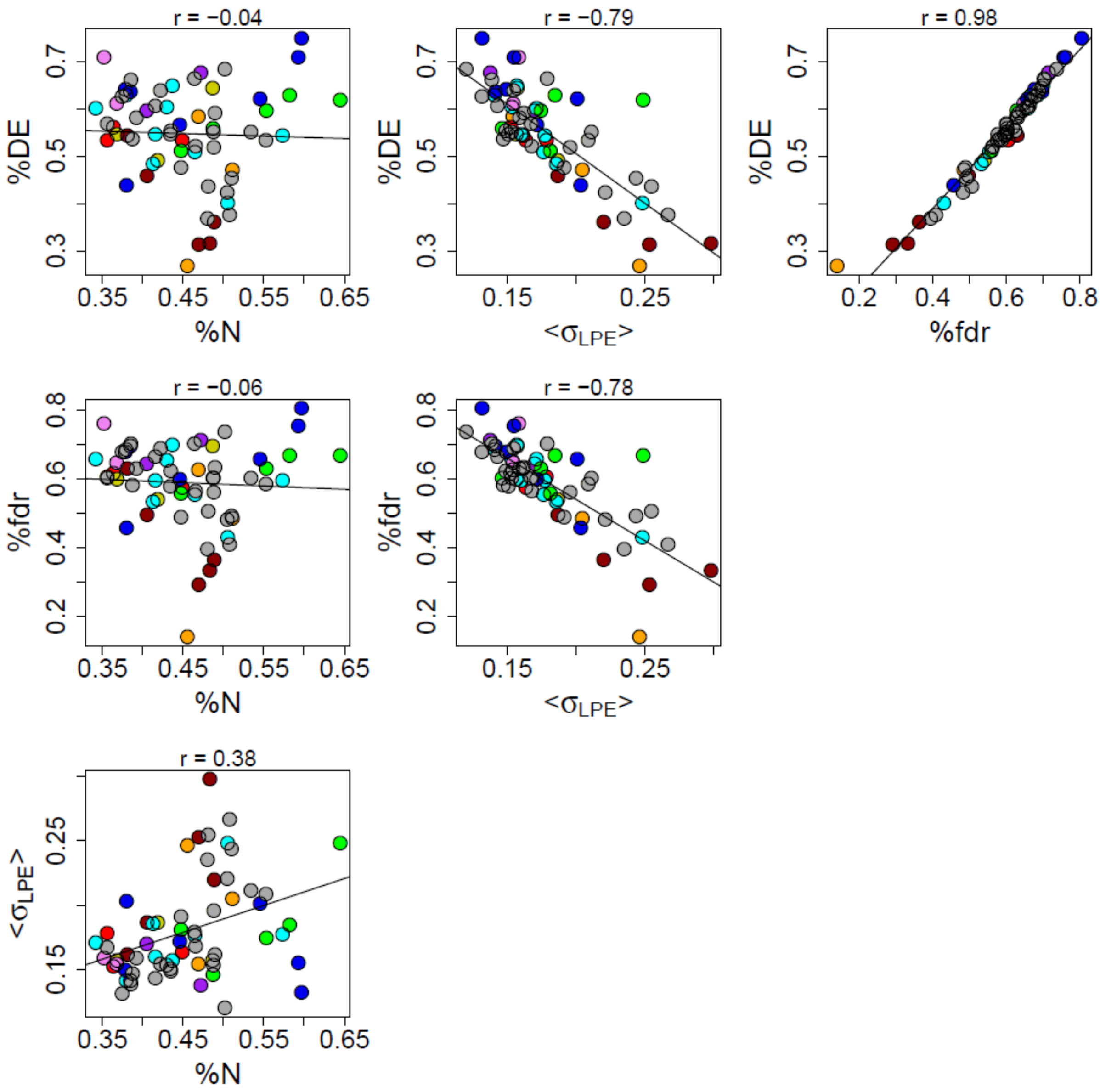


Figure 15: Correlation plots of different error estimates, the fraction of differentially expressed genes (%DE), of genes meeting a minimum FDR-criterion (%fdr), absent-called genes (%N) and the mean error level ($<\sigma_{LPE}>$), of the tissues studied. Regresssion lines and the respective regression coefficients are given within the figure.

4.6. Comparing alternative gene lists

Each of the alternative scores of differential expression provides an ordered list of differentially expressed genes per tissue which are ranked, for example, with decreasing absolute value of the score. The similarity between two lists of length r can be described using the ‘correspondence at the top’ (CAT(r)) plot. It shows the fraction of genes commonly found at the top of both lists up to rank r [57]. Note that ‘null-correspondence’ for randomly ranked genes can be estimated using the hypergeometrical distribution and Eqs. (8) and (12) (see below). The respective CAT(r) value is given by the probability that a selection of N_{set}=r genes is found among the top N_{list}=r positions of a total list of length N, p_{HG}= r/N (see below).

The CAT-plot thus estimates the agreement between two lists irrespective of the particular score values of the genes in the lists. For example, two lists can agree with CAT=0.5 but differ with respect to the significance level of the remaining 50% of genes. To assess this aspect of pairwise list comparisons we define the p-CAT(r) value as the cumulative logged p-values of the t-shrinkage score of the r genes at the top of the list obtained from the t-shrinkage or from the alternative scores. The p-CAT value of the t-shrinkage score provides the lower limit because it per definition is ranked with

increasing p value. The corresponding p-CAT value of an alternative score such as the WAD-statistics consequently judges the degree of discordance with respect to the t-shrinkage statistics. It is given as the difference Δp-CAT = p-CAT(r)$_{\text{alternative score}}$ – p-CAT(r)$_{\text{t-shrinkage}}$.
Finally, the rank-correspondence (RC) plot illustrates the agreement between two lists by color-coding each position either in red or in green: green symbols assign ranks which agree with ±20 positions in the alternative list whereas red ranks do not.

4.7. Differential expression of metagenes

SOM machine learning identifies k=1…K metagenes where each of them is representative for a minicluster of n_k real genes of correlated expression profiles. A simple natural approach of combining significance information for a group of genes is to calculate the mean characteristics averaged over the group members. Accordingly, we calculate the mean p- and fdr- (Fdr) values for each metagene via arithmetic averaging,

$$\left\langle S \right\rangle_{k,m} = \frac{1}{n_k} \sum_{g=1}^{n_k} S_{g \in k,m} \, , \qquad (6)$$

where $S_{g,m}= t_{g,m}$, $\log(p_{g,m})$, $fdr_{g,m}$ are the single gene significance characteristics of gene g in metagene k and tissue m. Ranking of the averaged characteristics provides ordered lists of metagenes according to their differential expression.
In [6] we defined spots of adjacent metagenes by applying different criteria, such as the mutual correlations between the metagene profiles or their differential expression beyond an appropriately chosen threshold value. For example, metagenes are classified as over- (or under-) expressed, if their expression value exceeds the 98% (or falls below the 2%) quantile-level of the expression range of all metagenes in the particular tissue studied. These spots are characterized by their mean significance characteristics as averages over all genes of the respective spot in analogy with Eq. (6)

$$\left\langle \mathrm{S_m} \right\rangle_{\mathrm{spot}} = \frac{1}{\sum\limits_{\mathrm{k \in spot}} \mathrm{n_k}} \sum_{\mathrm{k \in spot}} \sum_{\mathrm{g=1}}^{\mathrm{n_k}} \mathrm{S_{g \in k,m}} \; . \qquad (7)$$

4.8. Gene set overrepresentation analysis: integrating concepts of molecular function

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 [14], [15] 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). These sets can partly overlap in component genes, and some gene sets are subsets of others due to the hierarchical nature of the GO-systematics [39]. Rather than merge these sets we kept them all in order to maximize the functional annotation conveyed by the gene set names.
We will 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. Contrarily, the term ‘overexpression’ will be used to characterize deviations between the mean expression score averaged over the set-members in a list compared with the mean score of all list members independent of their overrepresentation. The term ‘enrichment’ will be used for estimates which combine overrepresentation and overexpression (see below).
Particularly, in gene set overrepresentation analysis, each gene studied is classified according to two memberships leading to a 2×2 contingency table for further testing (Table 5): firstly, its membership in the particular set of functionally related genes of length N_{set} and, secondly, its membership in the respective list of differentially expressed genes of length N_{list}. The intersection of the set and the list is given by the number of ‘positive’ genes, N_+. Then, one can estimate overrepresentation of these positive genes using the hypergeometric distribution by calculating the cumulative probability that there is more overlap between the list and the set than would be expected by chance [58-60],

$$p = P(n > N_+) = \sum_{n=N_+ +1}^{N_{set}} p_{HG}(n) \quad \text{with} \quad p_{HG}(n) = \frac{\begin{pmatrix} N_{set} \\ n \end{pmatrix}\begin{pmatrix} N - N_{set} \\ N_{list} - n \end{pmatrix}}{\begin{pmatrix} N \\ N_{list} \end{pmatrix}} \quad . \tag{8}$$

The obtained p-value estimates the probability to find a stronger overlap between the list and the set by chance than actually detected.

The gene set overrepresentation approach thus considers the joint membership of a gene in a gene set and in an independent list of genes without taking into account the rank and the particular values of the respective test statistics of the genes in the list. For example, it ignores whether a positive gene is found on top or on bottom of the list or whether a gene is strongly or weakly differentially expressed. In contrast, the so-called gene set overexpression approach compares the gene set statistics with the null given by the ensemble of all genes studied (see refs. [15] and [17] for a review). In this case however no enrichment of a set in a sub-ensemble of a gene list is taken into account.

Table 5: 2x2 contingency table of the number of genes in different classes for gene set overrepresentation in a list of differentially expressed genes

# of genes	in list	not in list	total
in set	N_+	N_{set}- N_+	N_{set}
not in set	N_{list}- N_+	N- (N_{list}+ N_{set})+ N_+	N- N_{set}
total	N_{list}	N- N_{list}	N

4.9. Gene set enrichment analysis: the GSZ-score

The so-called gene set Z-score (GSZ) merges both options provided by the gene set overrepresentation and the gene set overexpression approaches [17]. Namely, the GSZ method estimates overrepresentation of a gene set in a list using its score statistics, for example, $S_{g\in list} = t_{g\in list}$. It is designed in such a way that members of the list with high values on top of the list more heavily contribute than members with lower values down the list. Particularly, one first transforms the total sum of the score function over the gene list into two components containing members and non-members of the set,

$$S_{list} = \sum_{\text{all } g\in list} S_g = S_{list}^+ + S_{list}^- \qquad \text{with } S_{list}^+ = \sum_{g\in list \text{ AND } g\in set} S_g \quad \text{and} \quad S_{list}^- = \sum_{g\in list \text{ AND } g\notin set} S_g \quad . \tag{9}$$

Secondly, one defines the regularized Z-value of the differential score, $\Delta S_{list} = S_{list}^+ - S_{list}^-$, of the form (see [17] for details)

$$GSZ = \frac{\Delta S_{list} - E(\Delta S_{list})}{\sqrt{\lambda \cdot SE(\Delta S_{list})^2 + (1-\lambda)\cdot SE_0{}^2}} \quad . \tag{10}$$

Here,

$$\begin{aligned} &E(\Delta S_{list}) = \langle S\rangle_{list} \cdot \left(\langle N_+\rangle_{HG} - \langle N_-\rangle_{HG}\right) = \langle S\rangle_{list} \cdot \left(2\langle N_+\rangle_{HG} - N_{list}\right) \quad \text{and} \\ &SE(\Delta S_{list})^2 = 4\left(\frac{var(S)_{list}}{N_{list}-1}\left(\langle N_+\rangle_{HG} \cdot \left(N_{list} - \langle N_+\rangle_{HG}\right) - var(N_+)\right) + \langle S\rangle_{list}{}^2 \cdot var(N_+)\right) \end{aligned} \tag{11}$$

are the expected mean and the standard error of ΔS_{list} for the selected list under the null hypothesis. $\langle S\rangle_{list} = S_{list} / N_{list}$ and $var(S)_{list} = \frac{1}{N_{list}} \sum_{g\in list} \left(S_g - \langle S\rangle_{list}\right)^2$ are the mean and the variance of the expression score in the list, respectively. SE_0 and λ denote the regularization constant and a scaling factor (1 ≤λ ≤1) which were chosen to stabilize the variance in the denominator of Eq. (10) especially for short lists (see below).

The mean and the variance of positive members of the hypergeometrical distribution are

$$\langle N_+\rangle_{HG} = N_{set}\frac{N_{list}}{N} \quad \text{and} \quad var(N_+) = \langle N_+\rangle_{HG} \cdot \left(1 - \frac{N_{set}}{N}\right)\left(\frac{N - N_{list}}{N-1}\right) , \tag{12}$$

respectively. The respective mean number of negative members is $\langle N_{-}\rangle_{HG} = N_{list} - \langle N_{+}\rangle_{HG}$. One gets after inserting Eq. (12) into Eq. (11) for the special case $N, N_{list} >> 1$

$$E(\Delta S_{list}) = \langle S\rangle_{list}\cdot N_{list}\cdot\left(\frac{2\cdot N_{set}}{N}-1\right) \quad \text{and}$$
$$SE(\Delta S_{list})^2 \approx 4\cdot N_{set}\frac{N_{list}}{N}\cdot\left(\mathrm{var}(S)_{list}\cdot(1-\frac{N_{set}}{N})+\langle S\rangle_{list}^{\;2}\cdot(1-\frac{N_{set}}{N})\cdot(1-\frac{N_{list}}{N})\right) \quad . \qquad (13)$$

Eq. (13) indicates that the standard error in Eq. (10) vanishes for small sets and/or short lists (compared with the total number of genes , i.e. $N_{list}/N<<1$) giving rise to instable estimates of the GSZ-score [17]. Making use of approximation Eq. (13) we chose the regularization constant according to

$$SE_0^{\;2} \approx 4\cdot N_{list}^{min}\frac{N_{set}^{min}}{N}\cdot\left(\mathrm{var}(S)_{list}\cdot(1-\frac{N_{set}^{min}}{N})+\langle S\rangle_{list}^{\;2}\cdot(1-\frac{N_{set}^{min}}{N})\cdot(1-\frac{N_{list}^{min}}{N})\right)$$
$$\text{and} \quad \lambda = 1-\min\left(1,\sqrt{\frac{N_{list}^{min}}{N_{list}}\cdot\frac{N_{set}^{min}}{N_{set}}}\right) \qquad (14)$$

to penalize small lists and sets. N_{list}^{min} and N_{set}^{min} are minimum settings (typically 5-10) and $\langle S\rangle_{list}$ and $\mathrm{var}(S)_{list}$ are the mean and the variance of the significance score in the ensemble of all genes of the list. The ad-hoc estimate of the scaling factor λ ensures that SE_0 progressively increases with decreasing number of genes in the list and/or set. Obtained GSZ-values were transformed into p-values using a permutation approach which generates the respective null distribution by random rearrangement of genes in the collection of predefined gene sets. One and two tailed tests were applied to assess over- or underexpression and differential expression (i.e., under- *and* overexpression), respectively.

In the following we consider two special cases of the GSZ-score referring to overexpression and overrepresentation, respectively.

Firstly, the GSZ-score can be calculated for the whole gene list with $N_{list}=N$. Eq. (13) provides for this special case $E(\Delta S_{list})\big|_{N_{list}=N} = \langle S\rangle_{list}\cdot(2\cdot N_{set}-N)$ and $SE(\Delta S_{list})^2\big|_{N_{list}=N} \approx 4\cdot N_{set}\cdot\mathrm{var}(S)_{list}$. The difference score becomes $\Delta S_{list}\big|_{N_{list}=N} = \left(2\langle S^{+}\rangle_{list}\cdot N_{set} - \langle S\ \rangle_{list}\cdot N\right)$ where $\langle S^{+}\rangle_{list} = S_{list}^{+}/N_{set}$ is the mean expression score averaged over all members of the gene set. Insertion into Eq. (10) for the special case λ=1 provides the GSZ-score of the full list

$$GSZ\big|_{N_{list}=N} = \frac{\langle S^{+}\rangle_{list} - \langle S\ \rangle_{list}}{\sqrt{\mathrm{var}(S)_{list}/N_{set}}} \quad . \qquad (15)$$

It represents a Z-statistics estimating the overexpression in terms of the deviation of the set average of the expression score from its total average over the whole gene list where the standard error is estimated using the variance of S for sample size N_{set}. The respective shrinkage statistics is obtained with the substitution $\mathrm{var}(S) \to \mathrm{var}(S)\cdot(\lambda + (1-\lambda)\cdot N_{set}^{min}) \approx \mathrm{var}(S)\cdot N_{set}^{min}$ in the denominator of Eq. (15).

The second special case assumes an identical value of the expression score for all genes, $S_g=1$, after ranking. The difference score thus simply counts the difference of members and non-members of the set in the list, $\Delta S_{list}\big|_{S=1} = N_{+} - N_{-} = 2N_{+} - N_{list}$. The expected mean and the variance in Eqs. (11) and (13) are given by $\langle S\rangle_{list}=1$ and $\mathrm{var}(S)_{list}=0$, respectively. Insertion into Eq. (10) provides the GSZ-score with λ=1

$$GSZ\big|_{S=1} = \frac{\left(N_{+} - \langle N_{+}\rangle_{HG}\right)}{\sqrt{\mathrm{var}(N_{+})}} \approx \frac{\left(N_{+} - \langle N_{+}\rangle_{HG}\right)}{\sqrt{\frac{N_{set}\cdot N_{list}}{N}\cdot\left((1-\frac{N_{list}}{N})\cdot(1-\frac{N_{set}}{N})\right)}} \quad , \qquad (16)$$

where the right hand approximation assumes $N, N_{list} >> 1$. It represents a Z-statistics estimating the overrepresentation in terms of the deviation of the actual number of positive members from the expected mean according to the hypergeometrical distribution and the respective variance. Eq. (16) further simplifies for short lists and sets, $N_{list},N_{set}<<N$, into:

$$GSZ\big|_{S=1} \approx \frac{\left(N_{list}^{+} - \langle N_{+}\rangle_{HG}\right)}{\sqrt{\dfrac{N_{set}\cdot N_{list}}{N}}} \quad . \qquad (17)$$

The denominator substitutes for the shrinkage statistics with $\frac{N_{set}\cdot N_{list}}{N} \rightarrow \frac{\max(N_{list}\cdot N_{set}, N_{list}^{min}\cdot N_{set}^{min})}{N}$.

Eqs. (15) and (16) thus illustrate that the GSZ-score in its general formulation in Eq. (10) estimates enrichment in terms of a combination of overexpression and overrepresentation Z-scores. It has been shown in ref. [17] that the GSZ-score is related to alternative scores, namely the Random Sets [61] and the max-mean gene set statistics [62] representing a unification between these relevant scoring functions. Another comparative study on different gene set enrichment methods showed that removing incoherent pathways prior to analysis improves specificity [39]. The GSZ-score implicitly accounts for coherency because inconsistent genes with positive and negative contributions to the sum in Eq. (9) virtually compensate each other.

4.10. SOM-based metagene and spot enrichment

SOM analysis provides two-dimensional contour maps visualizing the expression pattern of k=1…K metagenes in a series of m=1…M tissues. Each tile of the SOM refers to a minicluster of n_k genes associated with the respective metagene. The overrepresentation and/or overexpression of a gene set can be estimated for these metagene-related lists of genes using the methods presented in the previous subsection. Importantly, the list of length $N_{list}=n_k$ per tile is invariant in all SOMs independently of the chosen tissue sample. In consequence, overrepresentation analysis in terms of the hypergeometric distribution (Eq. (8)) provides p-values for each gene-set s and metagene, $p_{s,k}$, which apply to all particular SOMs of the series of tissues studied. In other words, metagene-related overrepresentation is independent of the particular sample considered. We estimated overrepresentation of the whole collection of 1454 gene sets in terms of a ranked list of p-values to identify the most relevant gene sets for each metagene.

One can also pursue an orthogonal approach which calculates the significance of one selected gene sets in all metagenes to identify those of them which contain an enriched population of the genes from the chosen set. The results are visualized in terms of the so-called overrepresentation map. It color-codes the p-values of a particular gene-set in the two-dimensional mosaic of the SOM. The overrepresentation map also allows to link overrepresentation of a particular gene set with overexpression of the respective metagene by comparison with the sample-specific SOM. Particularly, overrepresented and overexpressed genes can be simply identified if overrepresentation and overexpression spots overlap in both maps. Note that the metagenes are located at the same positions in both maps.

In contrast to these sample-independent overrepresentation maps based on the hypergeometrical distribution one can use the GSZ-score (Eq. (10)) to study metagene-related gene set enrichment in a sample-specific fashion. Also in this case we calculated p-values for all 1454 gene sets as default. The null distribution of the GSZ-score was calculated for each list using randomly composed gene sets of equal length.

Gene set overrepresentation and enrichment analysis was also applied to gene lists which are extracted from spots of adjacent metagenes. In this case, the respective length of the list is given by the sum of the number of real genes belonging to all metagenes forming the spot, $N_{list} = \sum_{k\in spot} n_k$. Spot-related overrepresentation analysis based on the HG-distribution is characterized by one p-value per gene set and spot. It is independent of the selected sample if the spot is invariant in all samples. We applied this approach by using the global spots taken from the overexpression summary map which apply to all samples of the series. In addition, sample-specific spots are determined using a common overexpression threshold criterion to the SOM of different tissues. In this case one gets sample-specific overrepresentation lists because the size and position of each spot can vary from sample to sample and it can even disappear if the expression of the metagene strongly drops in a particular tissue. The GSZ-score delivers sample specific lists of gene sets for global and local spots as well because it explicitly processed the expression values of the genes in each spot.

Additional Material

Additional file 1: Atlas of the ranking maps of all tissues studied
Additional file 2: Atlas of errors and p-value distributions of all tissues studied
Additional file 3: The supplementary text addresses the error characteristics in different tissues, the gene ranking of single genes, gene set overrepresentation and alternative spot selection, GSZ-enrichment of selected spots in selected tissues and the selection of gene sets using global lists and gene set. Furthermore, the results of gene set analysis of subsets of tissues (zoom-in) and examples of summary reports are provided.
Additional file 4: Tissue specific gene sets
Additional file 5: Special gene sets of highly and weakly expressed and of housekeeping genes
Additional file 6: Results of gene set averaging approach

Complete sets of results for full tissue dataset as well as zooming-in analysis can be found on our website: http://som.izbi.uni-leipzig.de

Competing interests

The authors declare that they have no competing interests.

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 Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE)[63].

5. References

[1] T. Kohonen, “Self-organizing formation of topologically correct feature maps,” *Biological Cypernetics*, vol. 43, pp. 59-69, 1982.

[2] P. Tamayo 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*, vol. 96, no. 6, pp. 2907-12, Mar. 1999.

[3] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, “Analysis of gene expression data using self-organizing maps.,” *FEBS letters*, vol. 451, no. 2, pp. 142-6, May. 1999.

[4] J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong, “Analysis and visualization of gene expression data using self-organizing maps.,” *Neural networks : the official journal of the International Neural Network Society*, vol. 15, no. 8-9, pp. 953-66, 2002.

[5] G. S. Eichler, S. Huang, and D. E. Ingber, “Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles.,” *Bioinformatics (Oxford, England)*, vol. 19, no. 17, pp. 2321-2, Nov. 2003.

[6] H. Wirth, M. Loffler, M. von Bergen, and H. Binder, “Expression cartography of human tissues using self organizing maps,” *BMC Bioinformatics*, vol. 12, no. 1, p. 306, 2011.

[7] M. Dondrup, A. T. Hüser, D. Mertens, and A. Goesmann, “An evaluation framework for statistical tests on microarray data.,” *Journal of biotechnology*, vol. 140, no. 1-2, pp. 18-26, Mar. 2009.

[8] H. Binder, T. Kirsten, M. Löffler, and P. F. Stadler, “Sensitivity of microarray oligonucleotide probes: variability and effect of base composition,” *The Journal of Physical …*, 2004.

[9] H. Binder and S. Preibisch, “GeneChip microarrays—signal intensities, RNA concentrations and probe sequences,” *Journal of Physics: Condensed Matter*, 2006.

[10] H. Binder, J. Brücker, and C. J. Burden, “Nonspecific hybridization scaling of microarray expression estimates: a physicochemical approach for chip-to-chip normalization.,” *The journal of physical chemistry. B*, vol. 113, no. 9, pp. 2874-95, Mar. 2009.

[11] H. Binder, K. Krohn, and C. J. Burden, “Washing scaling of GeneChip microarray expression.,” *BMC bioinformatics*, vol. 11, p. 291, Jan. 2010.

[12] C. J. Burden and H. Binder, “Physico-chemical modelling of target depletion during hybridization on oligonulceotide microarrays.,” *Physical biology*, vol. 7, no. 1, p. 016004, Mar. 2010.

[13] M. Fasold, P. F. Stadler, and H. Binder, “G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration.,” *BMC bioinformatics*, vol. 11, p. 207, Jan. 2010.

[14] J. J. Goeman and P. Bühlmann, “Analyzing gene expression data in terms of gene sets: methodological issues.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 8, pp. 980-7, Apr. 2007.

[15] M. Ackermann and K. Strimmer, “A general modular framework for gene set enrichment analysis.,” *BMC bioinformatics*, vol. 10, p. 47, 2009.

[16] Z. Jiang and R. Gentleman, “Extensions to gene set enrichment.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 3, pp. 306-13, Feb. 2007.

[17] P. Törönen, P. J. Ojala, P. Marttinen, and L. Holm, “Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function.,” *BMC bioinformatics*, vol. 10, p. 307, Jan. 2009.

[18] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, “Discovering statistically significant pathways in expression profiling studies.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544-9, Sep. 2005.

[19] A. Subramanian et al., “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545-50, Oct. 2005.

[20] D. W. Huang, B. Sherman, and R. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nature protocols*, 2008.

[21] S. K. Sieberts and E. E. Schadt, “Moving toward a system genetics view of disease.,” *Mammalian genome : official journal of the International Mammalian Genome Society*, vol. 18, no. 6-7, pp. 389-401, Jul. 2007.

[22] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, “A gene-coexpression network for global discovery of conserved genetic modules.,” *Science (New York, N.Y.)*, vol. 302, no. 5643, pp. 249-55, Oct. 2003.

[23] J. Quackenbush, “Genomics. Microarrays--guilt by association.,” *Science (New York, N.Y.)*, vol. 302, no. 5643, pp. 240-1, Oct. 2003.

[24] K. Kadota, Y. Nakai, and K. Shimizu, “A weighted average difference method for detecting differentially expressed genes from microarray data.,” *Algorithms for molecular biology : AMB*, vol. 3, p. 8, Jan. 2008.

[25] K. Kadota, Y. Nakai, and K. Shimizu, “Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity.,” *Algorithms for molecular biology : AMB*, vol. 4, p. 7, Jan. 2009.

[26] M. A. Sartor, C. R. Tomlinson, S. C. Wesselkamper, S. Sivaganesan, G. D. Leikauf, and M. Medvedovic, “Intensity-based hierarchical Bayes method improves testing for

differentially expressed genes in microarray experiments.," *BMC bioinformatics*, vol. 7, p. 538, Jan. 2006.

[27] L. Shi, R. G. Perkins, H. Fang, and W. Tong, "Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential.," *Current opinion in biotechnology*, vol. 19, no. 1, pp. 10-8, Feb. 2008.

[28] C. Murie, O. Woody, A. Y. Lee, and R. Nadon, "Comparison of small n statistical tests of differential expression applied to microarrays.," *BMC bioinformatics*, vol. 10, p. 45, Jan. 2009.

[29] B. De Hertogh et al., "A benchmark for statistical microarray data analysis that preserves actual biological and technical variance.," *BMC bioinformatics*, vol. 11, p. 17, Jan. 2010.

[30] R. Tibshirani and L. Wasserman, "Correlation-sharing for detection of differential gene expression," *Arxiv preprint math/0608061*, 2006.

[31] J. Läuter, F. Horn, M. Rosołowski, and E. Glimm, "High-dimensional data analysis: selection of variables, data compression and graphics--application to gene expression.," *Biometrical journal. Biometrische Zeitschrift*, vol. 51, no. 2, pp. 235-51, Apr. 2009.

[32] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes are compact.," *Trends in genetics : TIG*, vol. 19, no. 7, pp. 362-5, Jul. 2003.

[33] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, "Promoter features related to tissue specificity as measured by Shannon entropy.," *Genome biology*, vol. 6, no. 4, p. R33, Jan. 2005.

[34] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.," *FEBS letters*, vol. 573, no. 1-3, pp. 83-92, Aug. 2004.

[35] H. Binder, K. Krohn, and S. Preibisch, "'Hook'-calibration of GeneChip-microarrays: chip characteristics and expression measures.," *Algorithms for molecular biology : AMB*, vol. 3, p. 11, Jan. 2008.

[36] H. Binder and S. Preibisch, "'Hook'-calibration of GeneChip-microarrays: theory and algorithm.," *Algorithms for molecular biology : AMB*, vol. 3, p. 12, Jan. 2008.

[37] A. I. Su et al., "Large-scale analysis of the human and mouse transcriptomes.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 7, pp. 4465-70, Apr. 2002.

[38] A. I. Su et al., "A gene atlas of the mouse and human protein-encoding transcriptomes.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 16, pp. 6062-7, Apr. 2004.

[39] D. M. Levine et al., “Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways.,” *Genome biology*, vol. 7, no. 10, p. R93, Jan. 2006.

[40] A. Zeisel, A. Amir, W. J. Köstler, and E. Domany, “Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes.,” *BMC bioinformatics*, vol. 11, p. 400, Jan. 2010.

[41] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, “A variance-stabilizing transformation for gene-expression microarray data.,” *Bioinformatics (Oxford, England)*, vol. 18 Suppl 1, pp. S105-10, Jan. 2002.

[42] D. Abdueva, D. Skvortsov, and S. Tavaré, “Non-linear analysis of GeneChip arrays.,” *Nucleic acids research*, vol. 34, no. 15, p. e105, Jan. 2006.

[43] H. Binder, S. Preibisch, and H. Berger, “Calibration of microarray gene-expression data.,” *Methods in molecular biology (Clifton, N.J.)*, vol. 576, pp. 375-407, Jan. 2010.

[44] R. Opgen-Rhein and K. Strimmer, “Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach.,” *Statistical applications in genetics and molecular biology*, vol. 6, p. Article9, Jan. 2007.

[45] G. K. Smyth, “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.,” *Statistical applications in genetics and molecular biology*, vol. 3, p. Article3, Jan. 2004.

[46] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116-21, Apr. 2001.

[47] A.-M. K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green, “BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data.,” *Biostatistics (Oxford, England)*, vol. 6, no. 3, pp. 349-73, Jul. 2005.

[48] A. A. Fodor, T. L. Tickle, and C. Richardson, “Towards the uniform distribution of null P values on Affymetrix microarrays.,” *Genome biology*, vol. 8, no. 5, p. R69, Jan. 2007.

[49] N. Jain, J. Thatte, T. Braciale, K. Ley, M. O’Connell, and J. K. Lee, “Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.,” *Bioinformatics (Oxford, England)*, vol. 19, no. 15, pp. 1945-51, Oct. 2003.

[50] H. H. Yang, Y. Hu, K. H. Buetow, and M. P. Lee, “A computational approach to measuring coherence of gene expression in pathways.,” *Genomics*, vol. 84, no. 1, pp. 211-7, Jul. 2004.

[51] H. R. Ueda et al., “Universality and flexibility in gene expression from bacteria to human.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3765-9, Mar. 2004.

[52] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, “Microarray data analysis: from disarray to consolidation and consensus.,” *Nature reviews. Genetics*, vol. 7, no. 1, pp. 55-65, Jan. 2006.

[53] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440-5, Aug. 2003.

[54] K. Strimmer, “fdrtool: a versatile R package for estimating local and tail area-based false discovery rates.,” *Bioinformatics (Oxford, England)*, vol. 24, no. 12, pp. 1461-2, Jun. 2008.

[55] K. Strimmer, “A unified approach to false discovery rate estimation.,” *BMC bioinformatics*, vol. 9, p. 303, Jan. 2008.

[56] J. Aubert, A. Bar-Hen, J. J. Daudin, and S. Robin, “Determination of the differentially expressed genes in microarray experiments using local FDR.,” *BMC bioinformatics*, vol. 5, p. 125, Sep. 2004.

[57] R. A. Irizarry et al., “Multiple-laboratory comparison of microarray platforms.,” *Nature methods*, vol. 2, no. 5, pp. 345-50, May. 2005.

[58] D. A. Hosack, G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki, “Identifying biological themes within lists of genes with EASE.,” *Genome biology*, vol. 4, no. 10, p. R70, Jan. 2003.

[59] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy, “GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.,” *BMC bioinformatics*, vol. 5, p. 16, Feb. 2004.

[60] R. Z. N. Vêncio and I. Shmulevich, “ProbCD: enrichment analysis accounting for categorization uncertainty.,” *BMC bioinformatics*, vol. 8, p. 383, Jan. 2007.

[61] M. Newton and F. Quintana, “Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis,” *The Annals of …*, 2007.

[62] B. Efron and R. Tibshirani, “On testing the significance of sets of genes,” pp. 1-31, 2006.

[63] V. Bissinger and O. Kolditz, “Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE),” *GAIA-Ecological Perspectives for Science*, 2008.

Supplementary Text

Mining SOM expression portraits: Feature selection and integrating concepts of molecular function

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

[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] Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes and Environment (LIFE); Universität Leipzig, D-4103 Leipzig, Philipp-Rosenthalstr. 27, Germany
[4] Helmholtz Centre for Environmental Research, Department of Metabolomics, D-04318 Leipzig, Permoserstr. 15, Germany

* to whom correspondence should be addressed

1. Error characteristics of different tissues

We calculated the standard deviation of the expression of each probed gene for each tissue using the replicate samples available. The level of variability markedly decreases as a function of the expression value (Figure S 1). The locally pooled error function (LPE, see green curves in Figure S 1) was combined with the gene-specific error into the regularized t-shrinkage values which provide specific p-value distributions for each tissue (Figure S 1). These distributions are used to estimate the false discovery rate (local and tail-based ones, fdr and FDR, respectively) and the fraction of differentially expressed genes, %DE. Details of the method are given in the Methods section of the main paper.

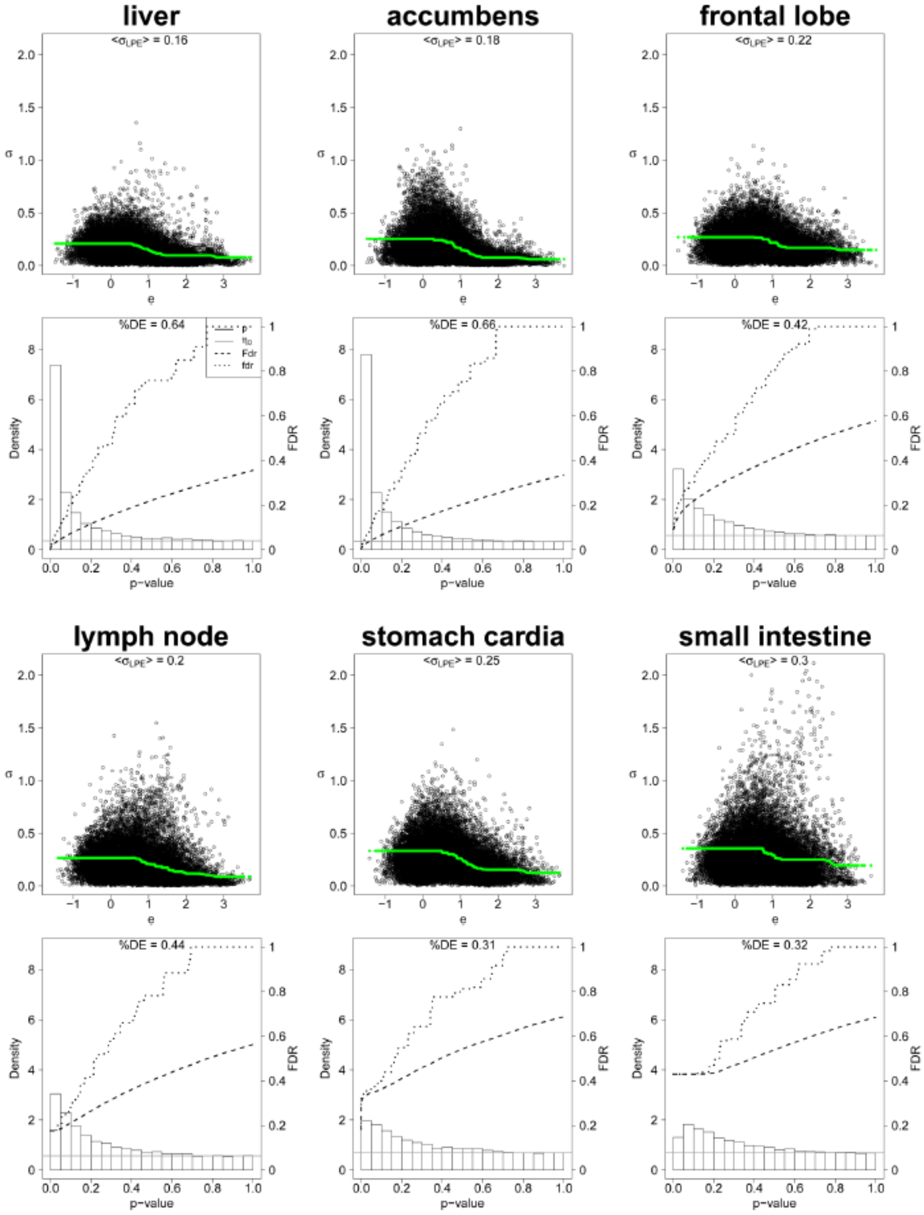


Figure S 1: Error characteristics of selected tissues: The first and third row of figures show error distributions (dots) and locally pooled estimates (green curves) of selected tissues as a function of the logged expression, e. The LPE-curves are calculated as moving average over 500 single probe values under the condition of non-positive slope which ensures that the LPE is maximal at small expression values. The second and fourth rows of figures show the respective p-value density distributions (bar histograms) together with the local FDR (dotted curves) and tail area-based FDR (dashed curves) obtained from the shrinkage t-statistics. The density-levels of null-genes, η_0, are shown by horizontal thin lines. The examples shown are ordered with increasing fraction of differentially expressed genes %DE.

2. Single gene ranking characteristics

We calculated mean global and local CAT(r) and Δp-CAT(r) values for lists of length r=10 and 100 of all tissue samples studied considering either all genes or the genes taken from the strongest overexpression spot, respectively. Figure S 2 shows boxplots of the data. The results of these global and local rank comparisons at both rank positions consistently show that similarities between the different lists are maximum for FC/WAD, and worse but virtually similar for WAD/t-shrinkage and FC/t-shrinkage pairings. Lists agree to about 70% (FC/WAD) and 50% (WAD/t-shrinkage and FC/t-shrinkage) on both, the global and local level and for both considered lengths (r=10 and 100) on the average. Local lists are slightly more similar by a few percent than global ones due to the pre-filtering of the genes in the SOM-spots. The averaged Δp-CAT values show that the penalty of the WAD- and FC-lists in units of the cumulative p-value of the t-shrinkage statistics is very similar for the global lists at rank r=10 and 100 and for the local lists of the overexpression spots at r=10. The former results indicate that the global lists are virtually equivalent at r>10 for all scores applied. Interestingly, the penalty of the Δp-CAT score of the local list almost completely disappears at r=100. This result can be rationalized by the fact that genes which penalize the p-CAT score at r<10 are simply shift to ranks 10<r<100 in the alternative lists where they compensate the penalty on top of the list. This effect of compensation is not observed for the global lists. Overexpressed genes are obviously preselected within the respective overexpression spots by the SOM machine learning algorithm making the local rankings more stable in the considered r-range.

The spot-filtering effectively combines the scoring of differential expression with the selection of co-expressed and correlated genes. It has been previously shown that ‘correlation-sharing’ for the detection of differentially expressed genes improves the performance of the analysis in terms of the false discovery rate [1].

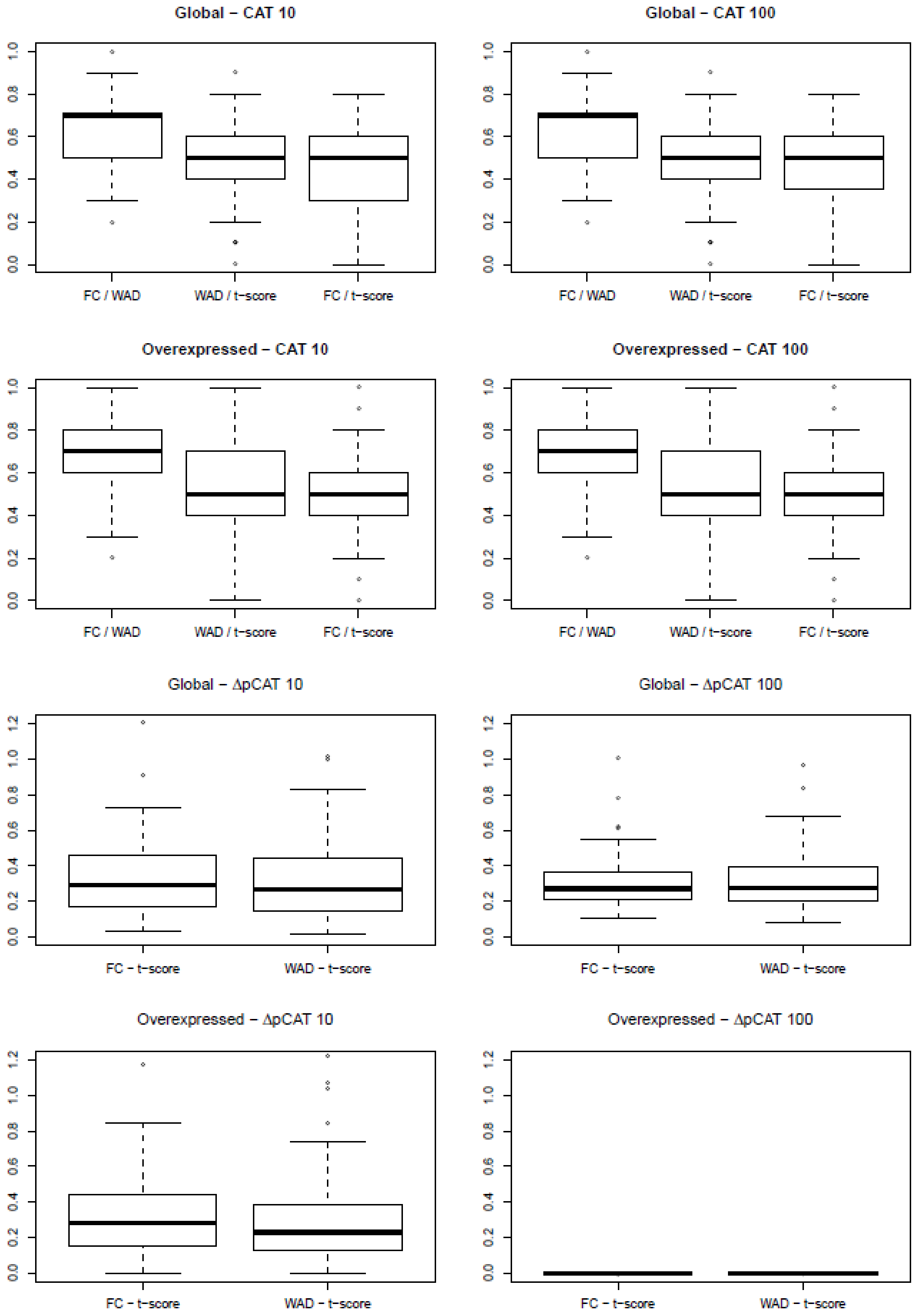


Figure S 2: Boxplots of rank-differences of ordered lists obtained from the different significance scores used at position r=10 (left) and r=100 (right) of all tissues studied: The differences are estimated using the CAT- and Δp-CAT scores for global and local lists considering all genes or genes of the strongest overexpression spot, respectively. The Δp-CAT(r) values are given as difference with respect to respective p-CAT(r) value of the t-shrinkage statistics.

3. HG- and GSZ-enrichment of selected spots in selected tissues

Figure S 3 shows bar plots of the top-twenty HG-overrepresented gene sets in the three spots A, B and F. Ten out of the top-twenty gene sets of spot A are related to nervous system and virtually all twenty gene sets overrepresented in spot B to muscle. Spot F overrepresents sets related to inflammation, leukocyte function etc. as expected for immune systems tissues. The annotation of the overrepresented gene sets clearly agrees with the tissues overexpressing the respective spot.

GSZ-enrichment analysis takes into account overrepresentation and overexpression of the genes of each set. It consequently provides sample-specific enrichment lists for constant spots due to the changing expression values of each gene in contrast to HG-overrepresentation which is sample independent. Figure S 4 shows bar plots of the top-ten GSZ-scored gene sets which are over- and underexpressed in the three spots A, B and F in three selected tissues (frontal lobe, skeletal muscle, lymph node).

The three arrows indicate the same three gene sets enriched in each of the spots for comparison. The GSZ-ranking provides very similar positions on top of the enrichment list in the tissues which overexpress a given set of genes in the respective spot (compare with Figure S 3): for example gene sets related to nervous processes are overexpressed in spot A of nervous tissue taken from the frontal lobe; gene sets related to muscle contraction are overexpressed in the muscle-related spot B of skeletal muscle tissue and gene sets related to immune system processes are overexpressed in the ‘immune system spot’ F of lymph node tissue. The expression of these spots can drop drastically in the other tissues considered. In consequence, part of the discussed gene sets occupy even leading position in the respective underexpression lists: for example, the gene sets addressing nervous processes (spot A) and muscle processes (spot B) are on leading positions in the underexpression list of lymph node tissue and the gene sets addressing immune system processes (spot F) and muscle processes (spot B) are found on top of the underexpression list in frontal lobe tissue. This result illustrates the property of the GSZ-score to combine gene set overrepresentation with over- (and under-) expression of the associated genes.

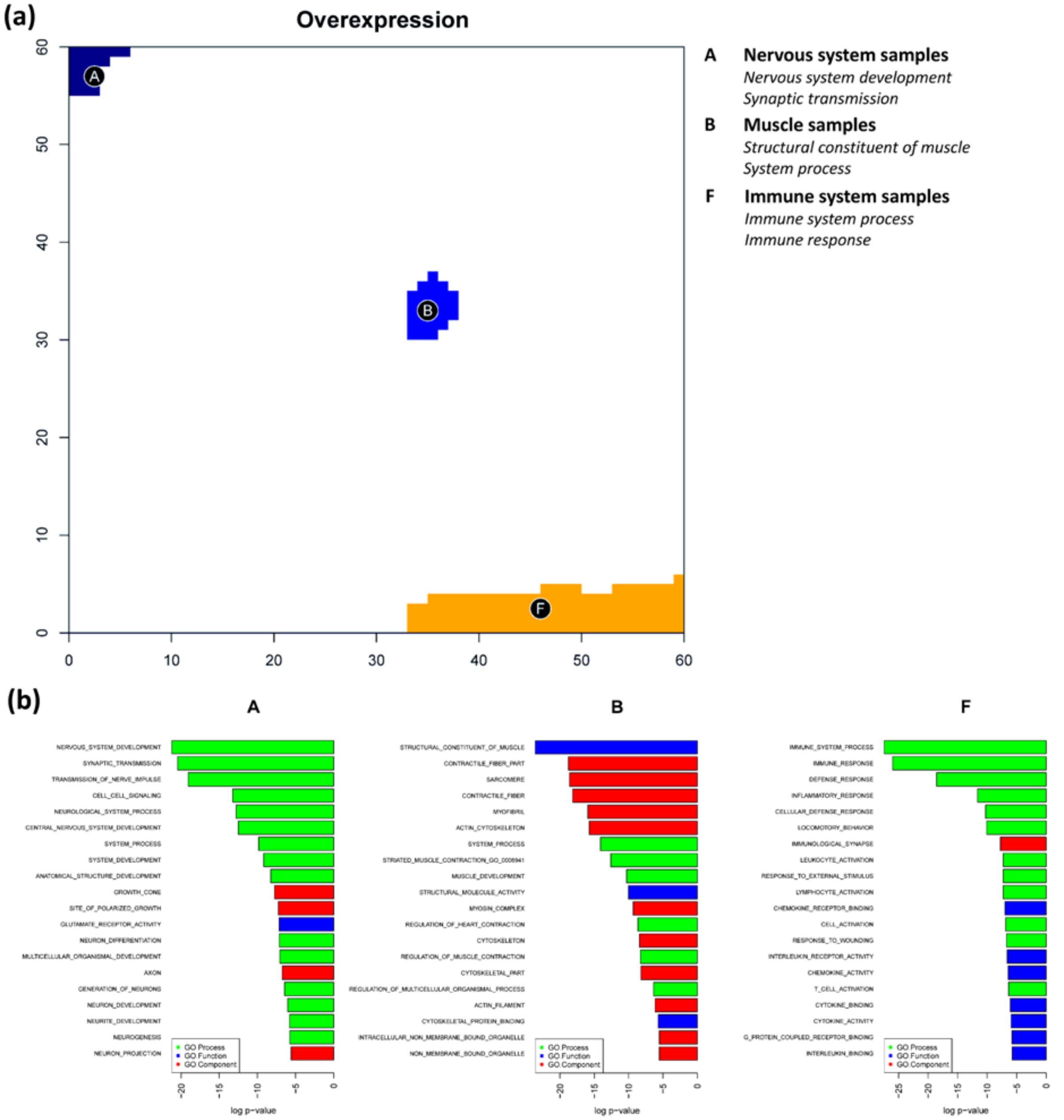


Figure S 3: The top-twenty gene sets of three selected spots. The length of the bars scales with the logged overrepresentation p-value of the sets. The color assigns the category of the gene sets according to the GO terms ‘molecular process’ (green), ‘molecular component’ (red) and ‘molecular process’ (blue).

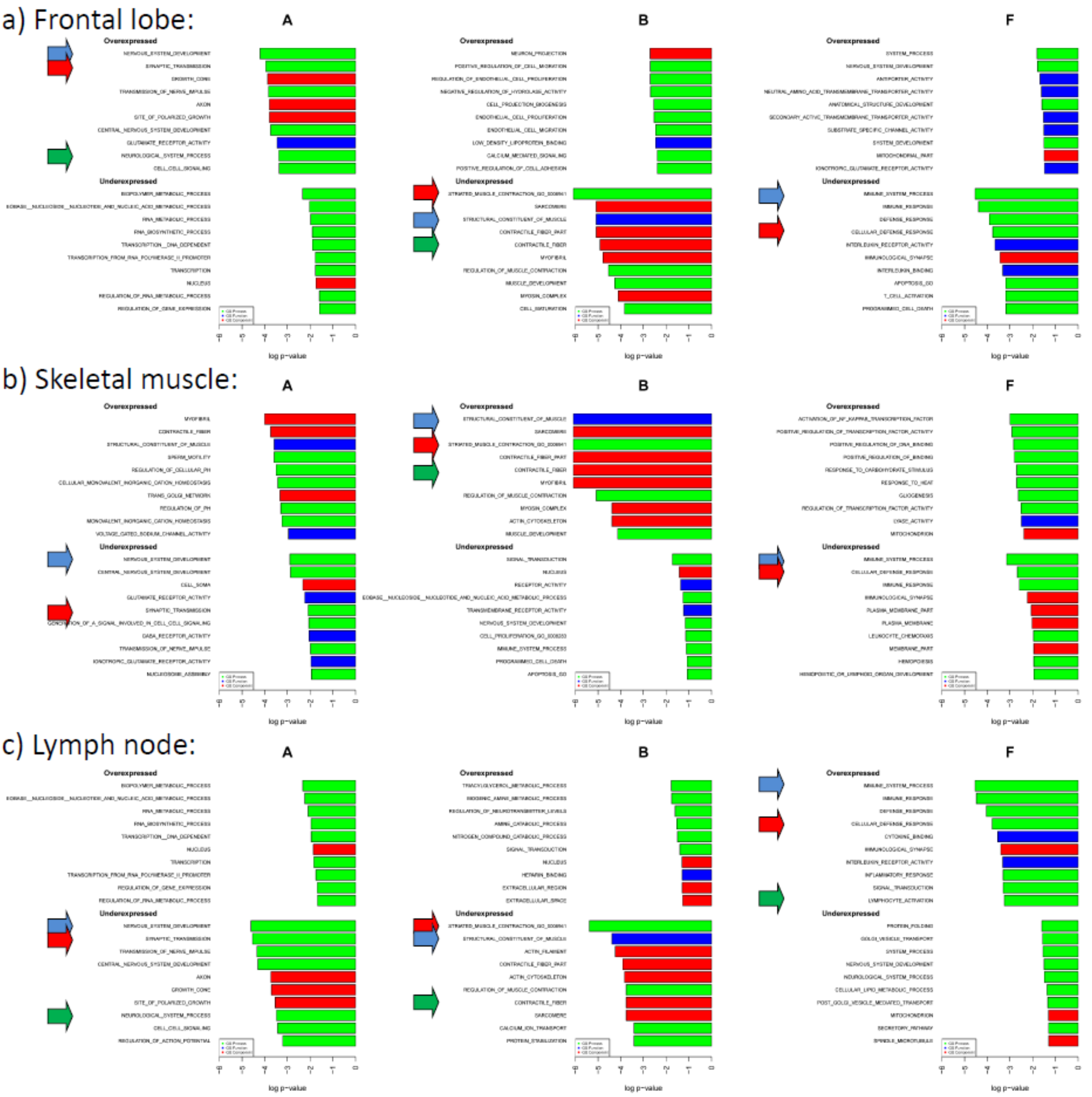


Figure S 4: Top-ten GSZ-over- and -underexpression lists of gene sets in the three spots for three tissues (the spot assignments are given in the main paper and in ref. [2]). The arrows indicate the same sets in each spot for direct comparison.

4. Functional context of over- and underexpression spots

The heatmaps in Figure S 5 show the mean expression of the overexpression and underexpression spots selected (see also the over- and underexpression maps in the right part of the figure for assignments of the spots). The three top overrepresented gene sets given in the figures allow to assign the functional context of each of the spots. Note that the terminus ‘over-/under-expression’ spot refers to the criterion of spot detection. Both types of spots show usually high expression in one and low expression in other tissues. A few of the over- and underexpression spots occupy the same (e.g. spots D and g) but mostly different positions in the maps. They consequently carry complementary information of high- and low-expression genes. For example, overexpression spot F can be assigned to ‘immune response’ whereas the nearby located underexpression spot b refers mainly to the translation machinery in the nucleus. The underexpression map detects also spots in regions without strongly overexpressed genes: For example, underexpression spot ‘a’ which can be assigned to endocytosis and membrane-related transport. Note also that the underexpression landscape is less sharp compared with the overexpression landscape.

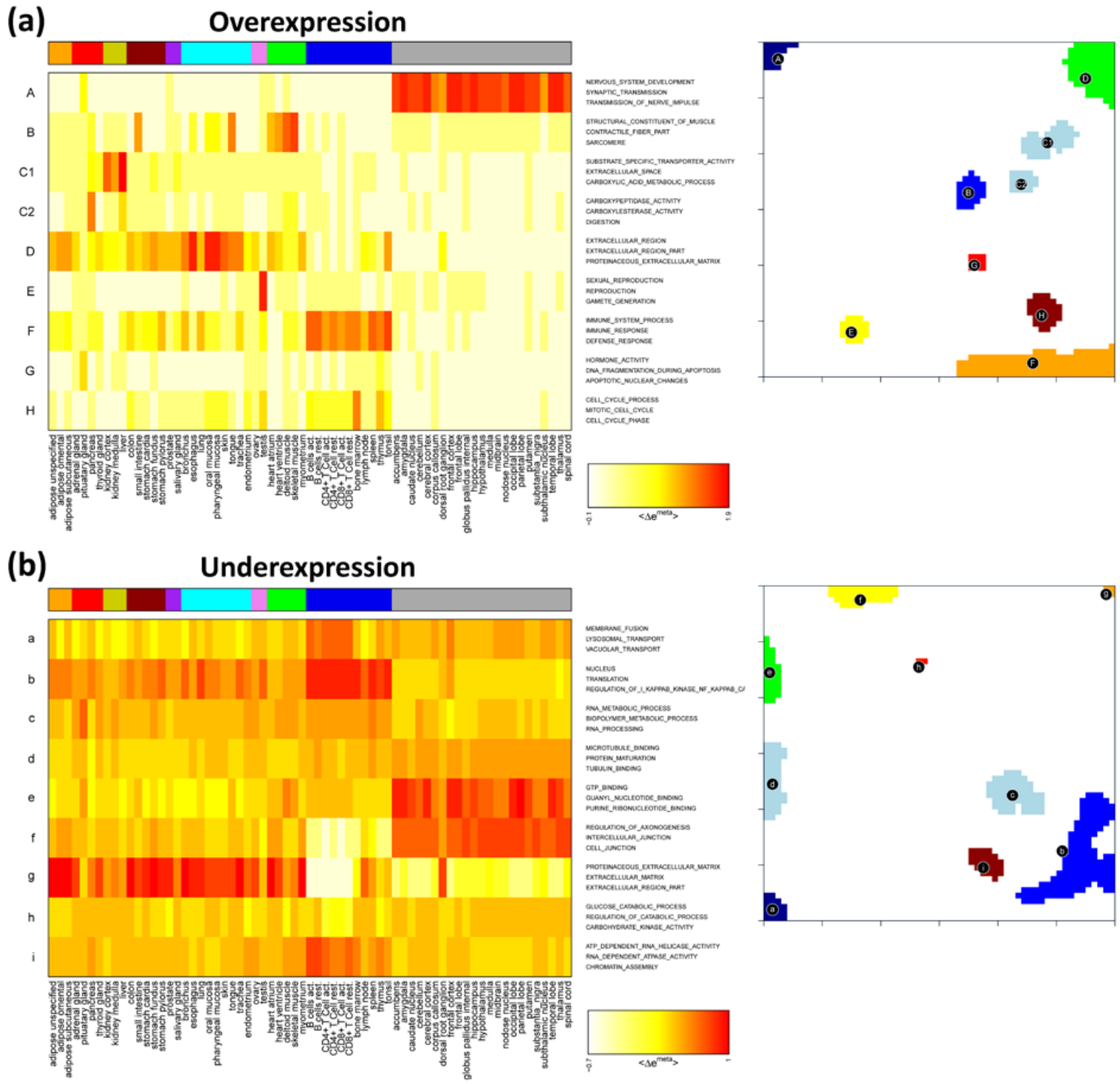


Figure S 5: Overrepresentation analysis of overexpression (panel a) and underexpression (panel b) spots. The heatmaps show the mean expression of the selected spots in all tissues studied. The top three overrepresented gene sets in each spot are given for each spot. The respective spot maps are redrawn from the main paper for direct assignment of the respective spot positions in the map.

5. Alternative spot selections

We applied alternative methods of spot selection partly described previously [2]. K-means clustering of the metagene profiles provides an area-filling spot pattern (Figure S 6). Here we arbitrarily set the number of clusters to fifteen to distribute the metagenes over a similar numbers of clusters as detected in the unsupervised spot selection based on the over- or underexpression of the metagegenes. Partly the position and size of the obtained spots agree with that of the overexpression and/or underexpression maps (e.g. the clustering spot C with overexpression spot A, and also H with F and J with f). Other spots occupy different areas of the map not selected by the over- or underexpression criteria (e.g. M, K). Moreover, most of the cluster-spots are larger than the typical over-/underexpression spots giving rise to a more coarse fragmentation of the map. On the other hand, the clustering spots enable the gapless sorting of genes into cluster-spots. Note that four of these cluster-spots are specifically overexpressed in nervous tissues (C, G, I, J, see Figure S 6b) with subtle differences in their functional context: Whereas spot C and G both overrepresent genes related to synaptic transmission, spots I and J collect genes associated with the pernuclear region and axiogenesis, respectively. The former ones are strongly underexpressed in most of immune system tissues.

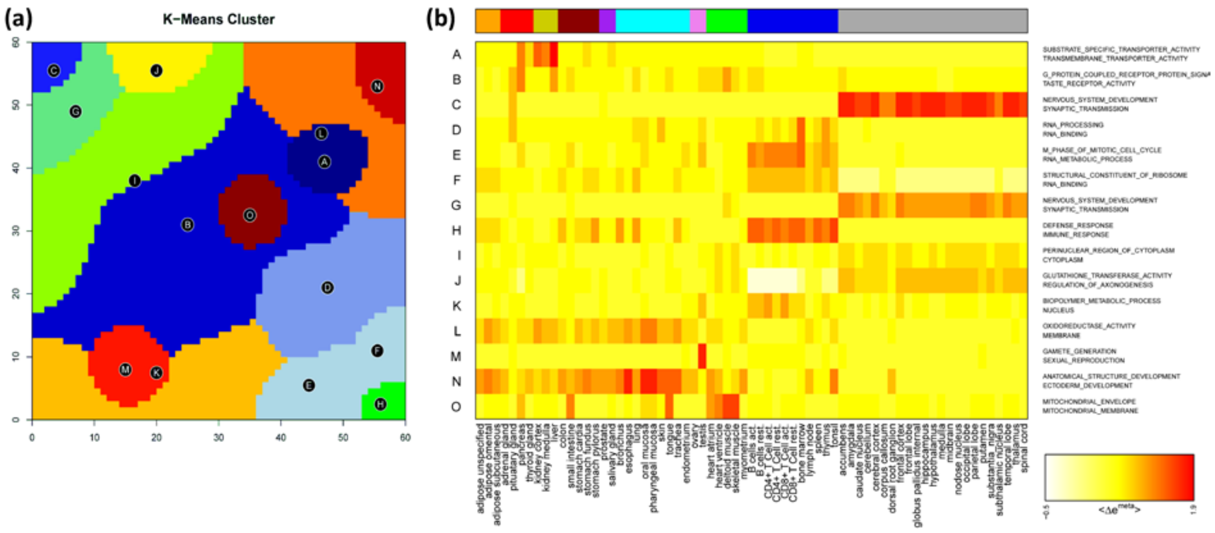


Figure S 6: Spot map based on k-means clustering of the metagene profiles (panel a). The heatmap shows the mean expression of the spots detected in all tissues studied. Spots are assigned using capital letters.

As an additonal option we selected spots of highly correlated metagenes (see ref. [2] for details of the spot selection algorithm). The obtained spot areas are again partly different compared with the spots obtained by the other methods discussed so far. The correlation spots tend to fragment the regions along the border of the map which refer to the metagenes of strongest variability of their expression profiles ([2]). The functional context reveals further details of SOM-mapping: For example, spots C, O, D, I, J and K are related to different aspects of nucleus function such as sexual reproduction (D), the ubiquitin ligase complex (O) and RNA metabolism (I).

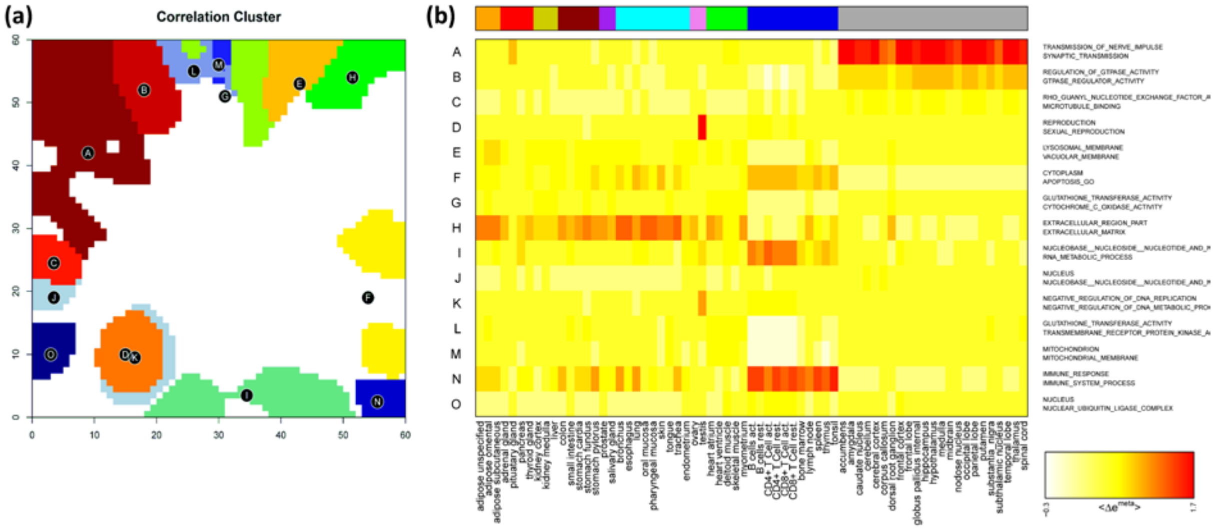


Figure S 7: Spot map based on Perssons correlation coefficient between adjacent metagenes (panel a; see [2] for details). The heatmap shows the mean expression of the spots detected in all tissues studied. Spots are assigned using capital letters.

6. Selecting gene sets from global lists

Our basic algorithm applies the 'top-three' criterion to the local lists of gene sets and extracts the selected sets into one global list. Particularly, it selects the three top gene sets per spot and merges them into the global list of gene sets which is further used to characterize gene expression in the different tissues in a functional context. This approach equally weights each spot in terms of the number of selected gene sets. This way, it ensures that each spot-feature is equally represented in the resulting global list.

Alternatively one can merge the full local spot lists of gene sets (i.e. without selecting the top-three sets) into one global one, rank them with increasing p-value and finally cut the list either at a suited significance threshold or after a certain number of positions. In this case the spots contribute with different numbers of gene sets depending on the respective degree of enrichment. We applied this approach using the p-values of the hypergeometrical distribution and of the GSZ-score which was calculated separately for over- and underexpression spots.

Figure S 8b shows the respective density distribution of the p-values (from the left to the right). These p-value distributions provide the total fraction of significantly enriched gene sets, %DE, and the respective local (fdr) and tail based (FDR) false discovery rates in analogy with the single gene analysis described in the methodical part of the main paper. The FDR- and fdr-functions increase much more steeply for the GSZ-lists compared with the HG-list. In consequence, application of a constant significance level (e.g. FDR<0.1) selects much less features from the GSZ-list than from the HG-list. Recall that the respective null distributions are given either analytically by the hypergeometrical distribution or they are estimated empirically for the GSZ-distributions using random permutations. These different approaches presumably produce the different FDR-levels of both approaches. Note that the null distribution of a test statistic under permutation is not necessarily the same for equally and differentially expressed genes. Previously it was suggested to use suited subsets of the data to more accurately estimate true nulls and to substantially increase the power of significance testing [3-4]. Moreover, methodical problems with the proper definition of the null hypothesis and the proper calculation of p-values which arise in the context of gene set enrichment analysis have been identified [5].

To compare the results of both approaches we select similar numbers of gene sets in each of the global lists referring to HG-enrichment (145 gene sets with fdr<0.0001), GSZ-overexpression (169 gene sets with fdr<0.1) and GSZ-underexpression (72 gene sets with fdr<0.2). The obtained gene set enrichment heatmaps in Figure S 8a reveal very similar spot pattern with essentially the same lists of enriched genes sets (Table S 1). Note that the underexpression GSZ-heatmap collects sets of virtually inactive genes whereas the overexpression heatmaps (HG and GSZ) refer to strongly overexpressed gene sets. The obtained sets refer consequently to different functions.

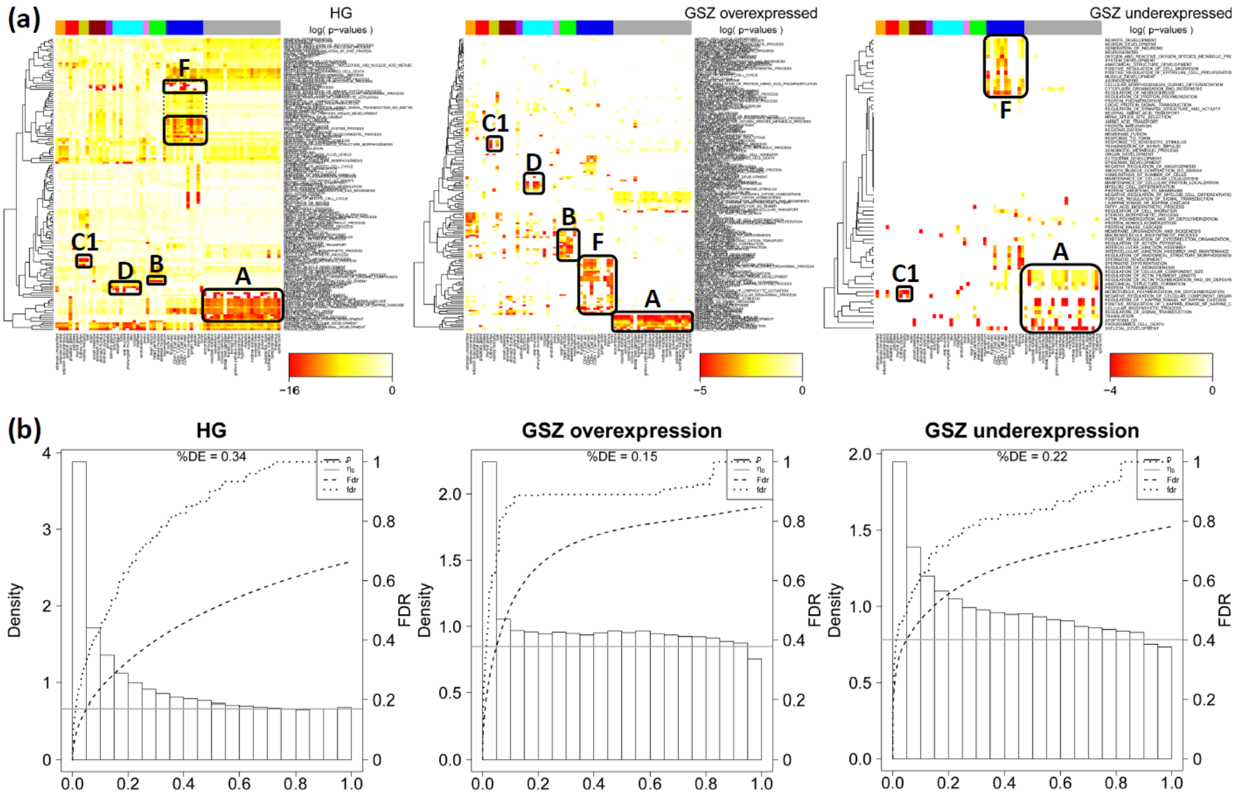


Figure S 8: One-way hierarchical clustering heatmap of significantly enriched gene sets (rows) versus tissues (columns) using the HG- and the GSZ-statistics using a fdr-threshold (panel a). The GSZ-statistics was applied separately to over- and underexpression spots. The heatmap color codes the p-values of the respective score in log-scale (see the legends in the figure). The tissue categories are color coded in the bar above the heatmap according to the assignments given in [2]. The gene sets are clustered in vertical direction. Panel b shows the respective p-value density distributions together with the local (dotted curves) and tail based (dashed curve) false discovery rates (see right ordinates).

Table S 1: Top gene sets from selected spots of the heatmaps shown in Figure S 8

spot	HG	GSZ-overexpression	GSZ-underexpression [a]
A	Cell-cell signaling	Cell-cell signaling	Skeletal development
	Neurological system process	Transmission of nerve impulse	Regulation of I-κB cascade
	Synaptic transmission	Synaptic transmission	Translation
	Nervous system development	Nervous system development	Apoptosis
F	Lymphocyte activation	Lymphocyte differentiation	Regulation of neurogenesis
	Regulation of immune system	Immune system development	Cytoplasm organization
	Immune response	Immune response	Axonogenesis
	Defense response	Defense response	Neuron development
C1	Organic acid metabolic process	Organic acid metabolic process	Microtubule polymerization
	Carboxylic acid metabolic process	Carboxylic acid metabolic process	Negative regulation of cellular component organization
	Excretion	Glutamine family metabolic process	

[a] assignment of underexpression spots to capital letter is arbitrary

7. Zoom-in: Nervous tissues

We applied a ‚zoom-in‘ step of SOM analysis to study the expression profiles of the subgroup of nervous immune systems and the remaining ‘diverse’ tissues with enlarged resolution as described in ref. [2]. They show ‘new’ textures of characteristic over- and underexpression spots which reflect the expression profiles of the tissues of interest more in detail than the original SOM. Figure S 9a shows the obtained overexpression spots and the three leading overexpressed gene sets after global overexpression analysis of nervous tissues. Spot H collects processes directly related to nervous system whereas spots G and H refer to nucleus-related and cell membrane-related processes, respectively. The zoom-in map amplifies subtle details of the expression profile of these genes in the reduced subset selected for zoom-in analysis. Also that the GSZ-overexpression profile of the gene set ‘nervous system development’ shows a heterogeneous fine structure which reflects modulation of the expression of this set in the nervous tissues. The GSZ-enrichment heatmap after zoom-in is shown in Figure S 9b. It provides a detailed picture of the gene set enrichment in nervous tissues.

Figure S 10 and Figure S 11 provide overrepresentation maps and overexpression profiles of the same gene sets selected in the respective plots in the main paper. The genes of the sets widely distribute over the maps. The gene sets related to synaptic transmission and to the transmission of nerve impulse indeed accumulate in the region of spot H and the gene set ‘immune systems process’ in the region of spot H, as expected. The overexpression level however can strongly vary in the different nervous systems tissues: For example, the former two gene sets are clearly underexpressed in corpus callosum and subthalamic nucleus, which, on the other hand, show relative overexpression of the gene set immune systems process. The overexpression of the remaining three gene sets considered is mostly invariant in nervous tissues.

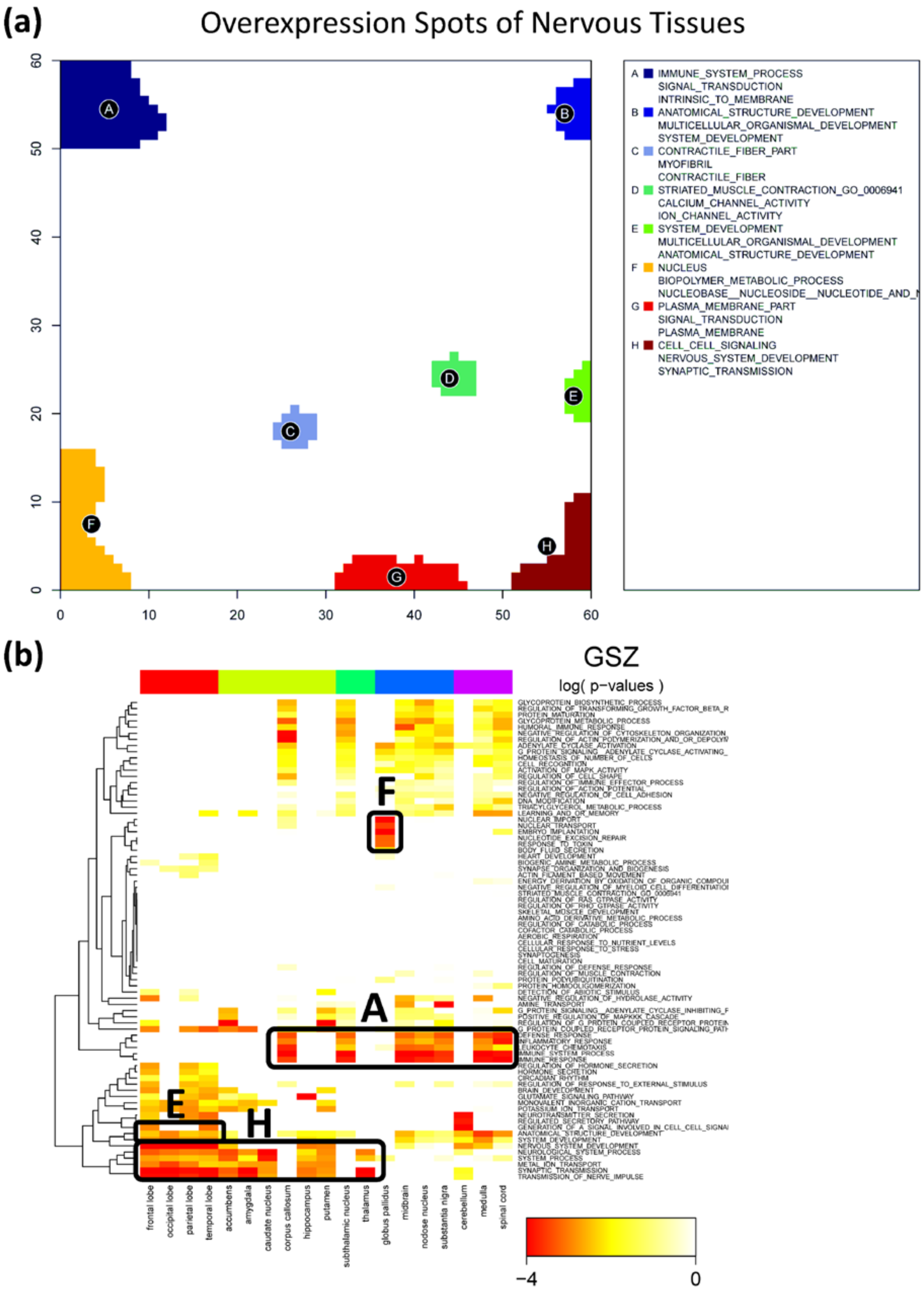


Figure S 9: The overexpression summary map of nervous tissues shows eight spots (A – H) which are strongly overexpressed in at least one of the 19 nervous tissues studied. Global overrepresentation analysis is estimated for each spot using the hypergeometrical distribution. The right legend assigns the two most significantly overrepresented gene sets in each spot. Expression heatmaps of the spots are shown in the supplementary material of ref. [2].

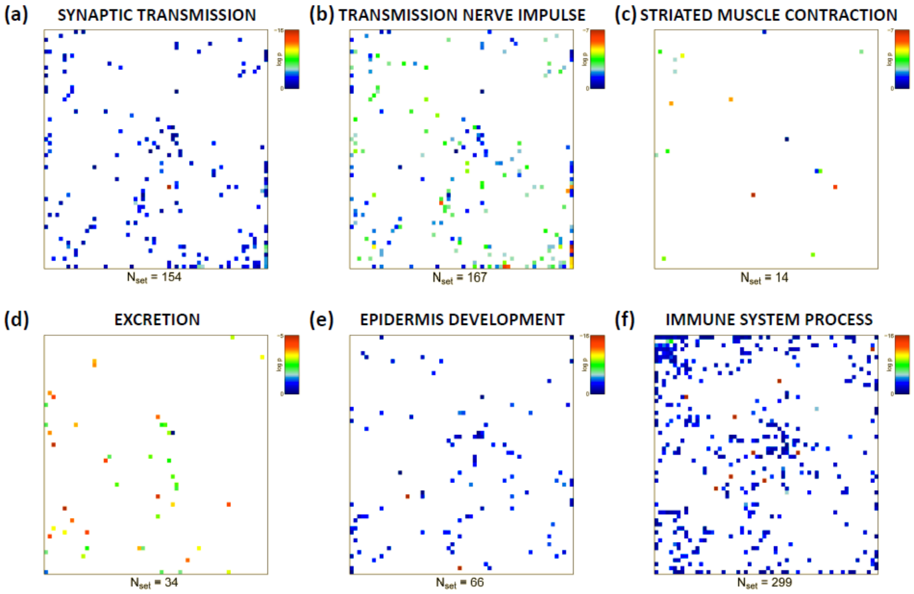


Figure S 10: Overrepresentation maps of six selected gene sets for the zoom-in SOM of nervous tissues. Overrepresentation in each tile of the mosaic is calculated in units of log(p_{HG}) using the hypergeometrical distribution and color-coded (maroon>red>yellow>green>blue). White areas indicate metagenes not containing genes from the respective set).

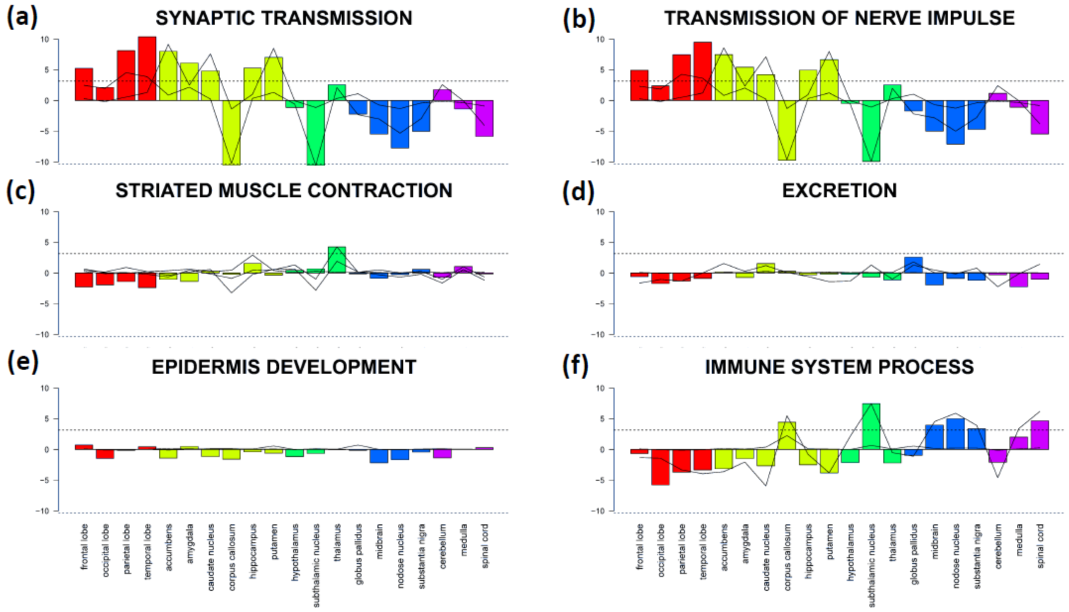


Figure S 11: Overexpression profiles of selected gene sets in nervous tissues. The bars are colored in accordance to the color-codes of the tissue categories. They are scaled in units of the GSZ-score (left axis). The horizontal dotted lines mark the fdr=0.2 significance threshold estimated from the p-value distribution of the GSZ-score. The inserted curves show the logged FC-expression profiles of the top-three metagenes of strongest enrichment of the respective gene set.

8. Zoom-in: Immune systems tissues

Figure S 12 shows the global spot overexpression maps after zoom-in of immune system tissues, and Figure S 13 below the respective GSZ-enrichment heatmap. Both approaches of gene set analysis provide consistent results where local GSZ-enrichment analysis shows a slightly more diverse pattern than global overrepresentation analysis. Note however, that the overexpression spot maps list only the three leading gene sets. Extended lists are available in the detailed reports described below.

The respective overrepresentation (Figure S 14) and overexpression profiles (Figure S 15) show that gene sets which are obviously not related to these tissue categories (e.g. ‘synaptic transmission’) are virtually invariant and accumulate around the centre of the map. On the other hand, gene sets related to selected tissues accumulate in special regions of the maps and show heterogeneous overexpression (see, for example, the gene sets ‘striated muscle contraction’ and ‘epidermis development’ in Figure S 18).

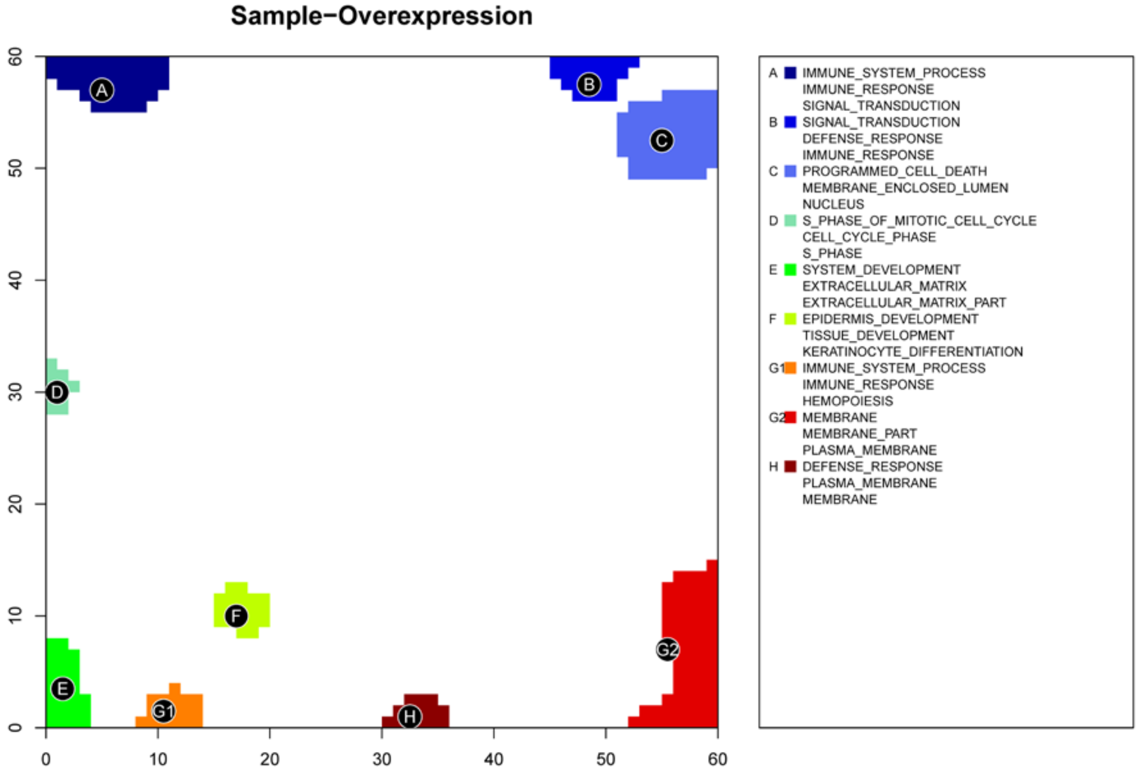


Figure S 12: The overexpression summary map of immune systems tissues shows nine spots (A – H) which are strongly overexpressed in at least of the 11 immune systems tissues studied. Global overrepresentation analysis is estimated for each spot using the hypergeometrical distribution. The right legend assigns the most significantly overrepresented gene sets in each spot. Expression heatmaps of the spots are shown in the supplementary material of ref. [2]

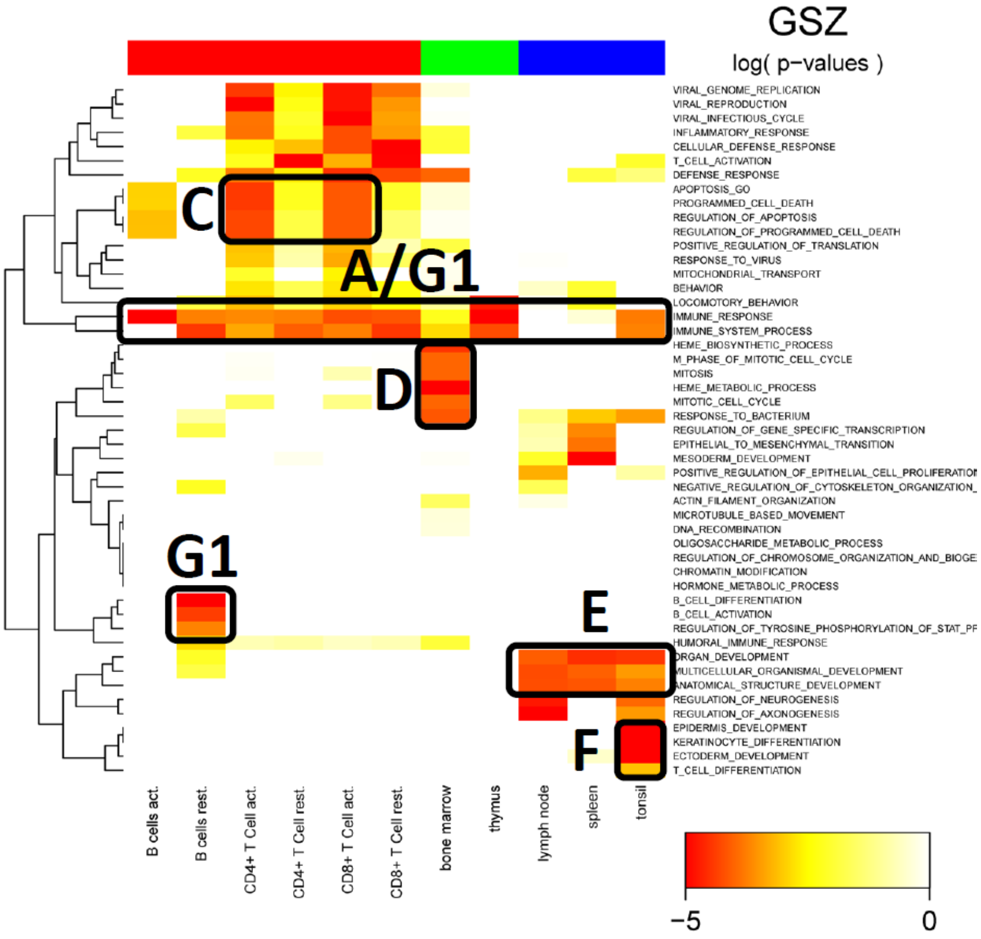


Figure S 13: One-way hierarchical clustering heatmap of significantly enriched gene sets (rows) in immune systems tissues (columns) using the GSZ-statistics. The top-three gen sets per overexpression spot are selected. The heatmap color codes the p-values of the respective score in log-scale (see the legends in the figure). The tissue categories are color coded in the bar above the heatmap according to the assignments given in [2]. The gene sets are clustered in vertical direction. The capital letters assign clusters of enriched gene sets in correspondence with the spots shown in Figure S 12.

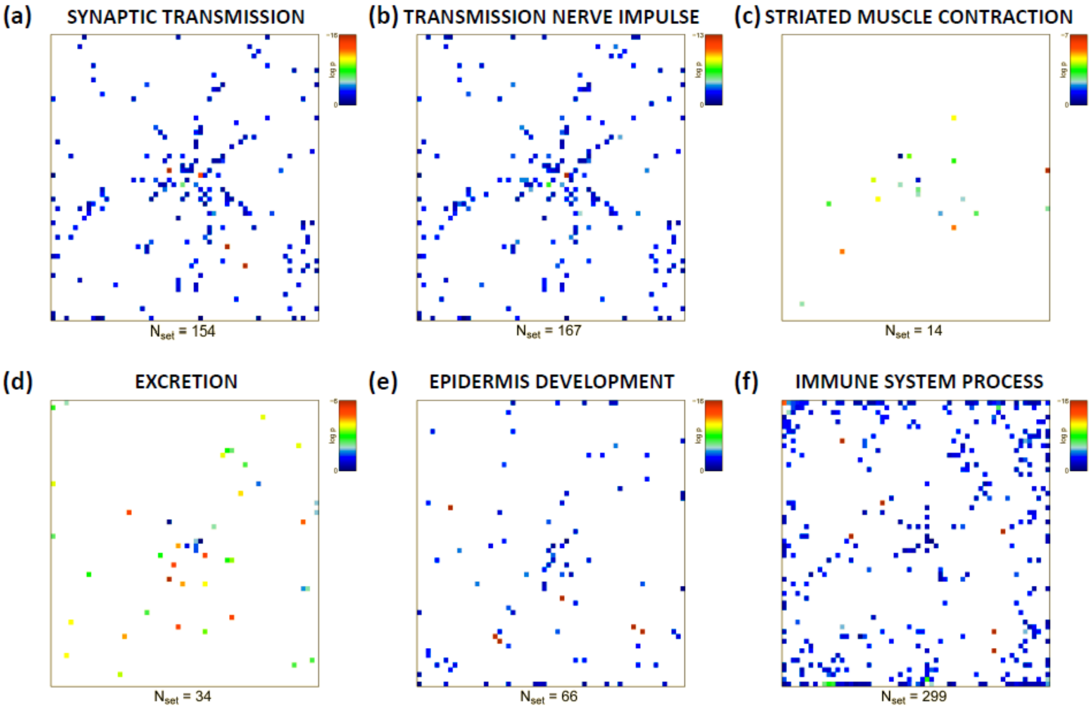


Figure S 14: Overrepresentation maps of six selected gene sets for the zoom-in SOM of immune sytem tissues.

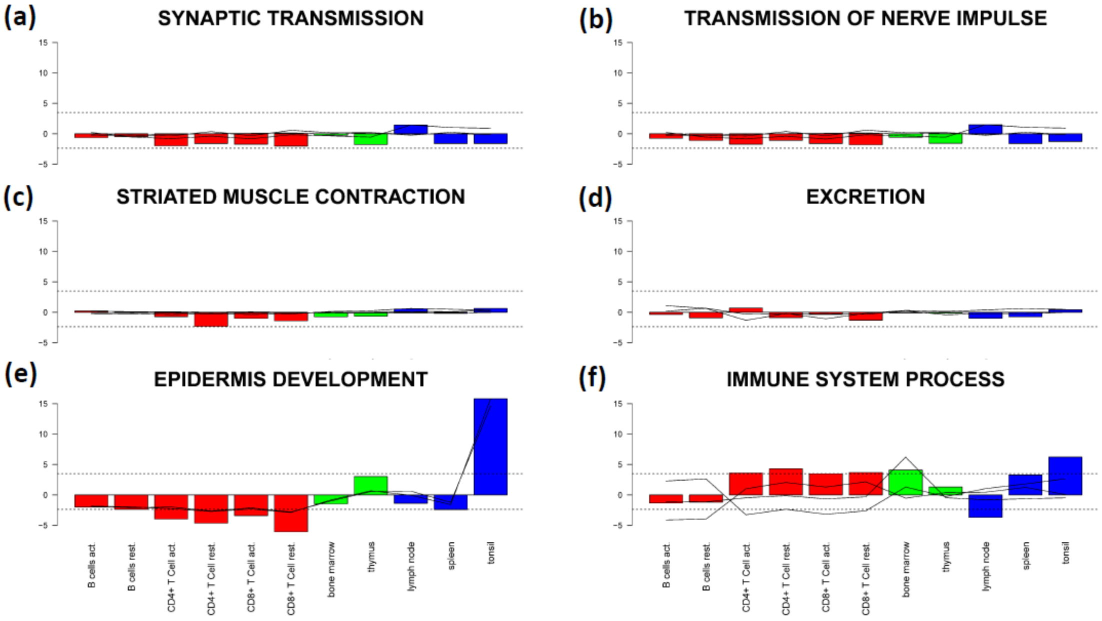


Figure S 15: Overexpression profiles of selected gene sets in immune system tissues.

9. Zoom-in: Diverse tissues

The collection of ‘diverse’ tissues subsumes the categories adipose, endocrine, homeostasis, digestion, exocrine, epithelium and muscle tissues which cluster relatively tightly together in the agglomerative analyses provided previously. Figure S 16 shows the global spot overexpression maps after zoom-in, and Figure S 17 below the respective GSZ-enrichment heatmap. Both approaches of gene set analysis show consistent results where local GSZ-enrichment analysis provides a slightly more diverse pattern than global overrepresentation analysis.

The respective overrepresentation (Figure S 18) and overexpression profiles (Figure S 19) show that gene sets which are obviously not related to these tissue categories (e.g. ‘synaptic transmission’) are virtually invariant and accumulate around the centre of the map. On the other hand, gene sets related to selected tissues accumulate in special regions of the maps and show heterogeneous overexpression (see, for example, the gene sets ‘striated muscle contraction’ and ‘epidermis development’ in Figure S 18 and Figure S 19).

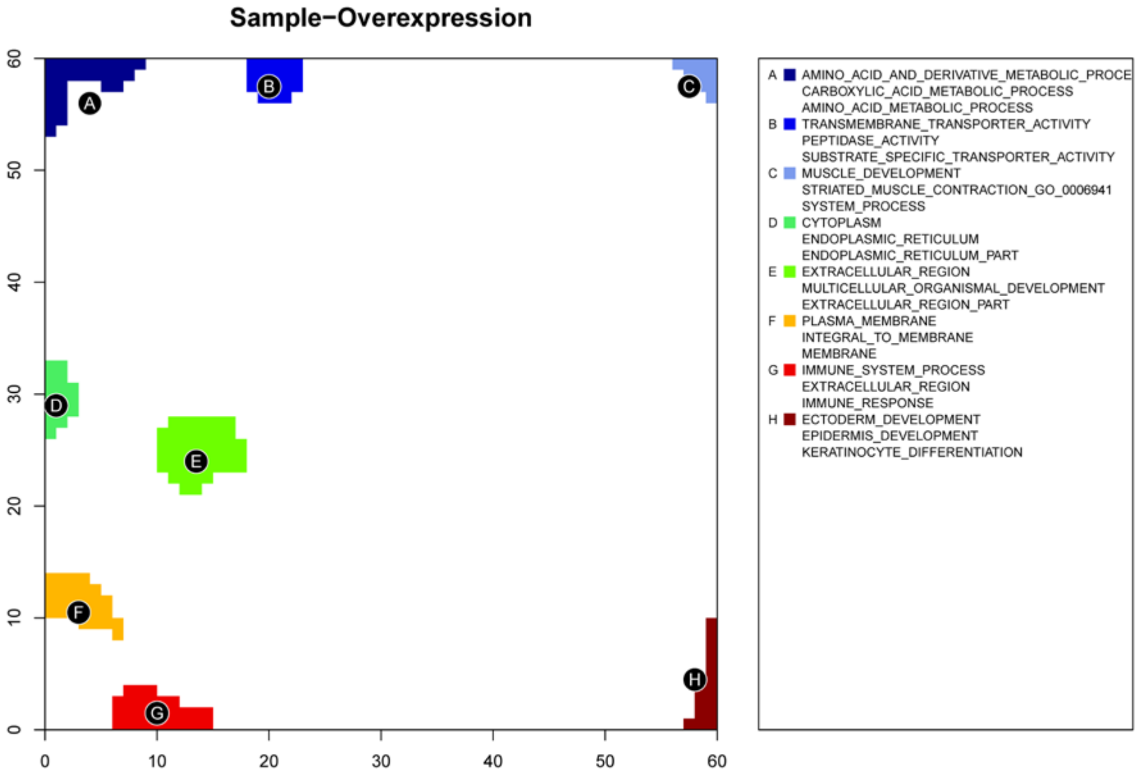


Figure S 16: The overexpression summary map of the group of ‘diverse’ tissues shows ten spots (A – J). The right legend assigns the two most significantly overrepresented gene sets in each spot. Expression heatmaps of the spots are shown in the supplementary material of ref. [2]

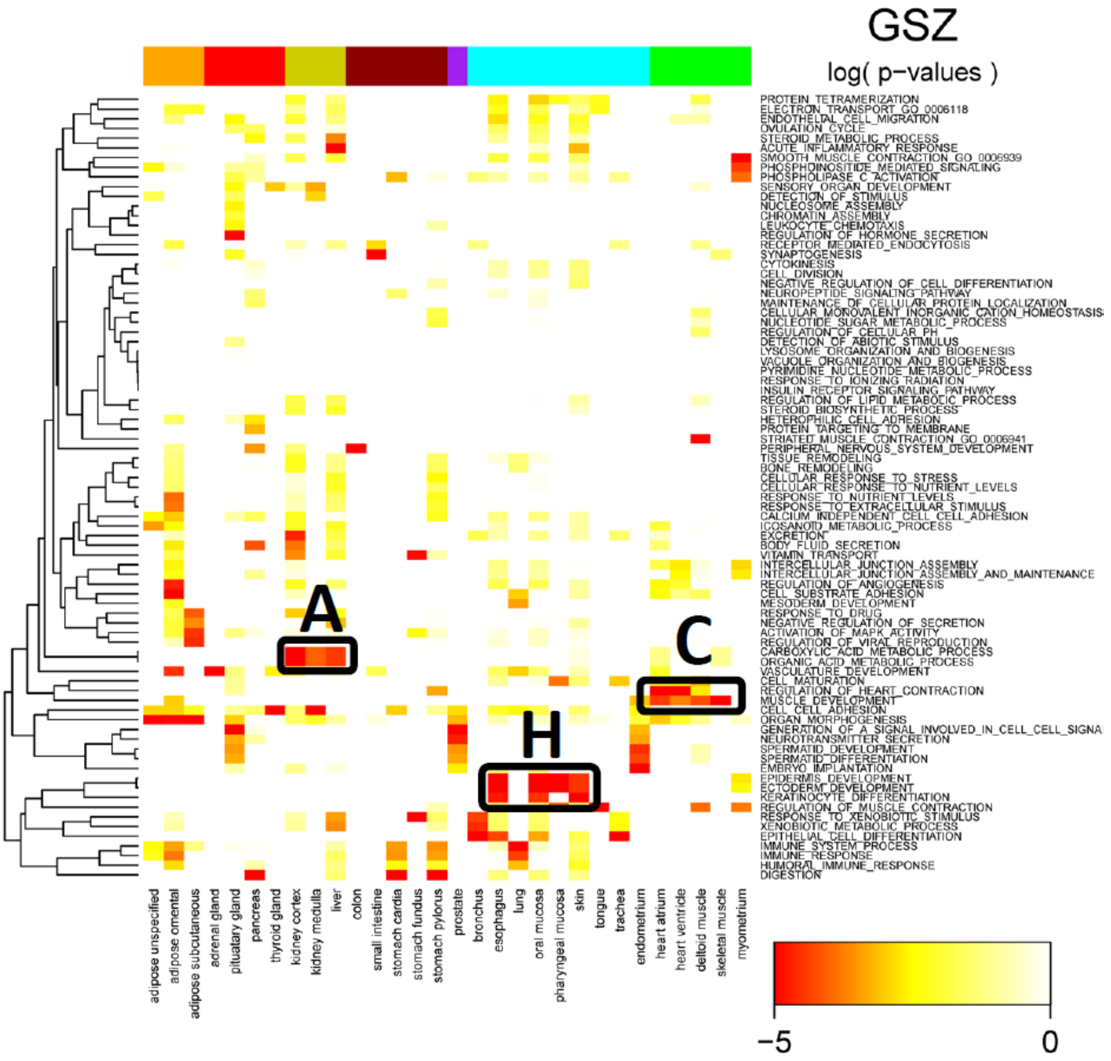


Figure S 17: One-way hierarchical clustering heatmap of significantly enriched gene sets (rows) in the selection of diverse tissues (columns) using the GSZ-statistics. The top-three gen sets per overexpression spot are selected. The heatmap color codes the p-values of the respective score in log-scale (see the legends in the figure). The tissue categories are color coded in the bar above the heatmap according to the assignments given in [2]. The gene sets are clustered in vertical direction.

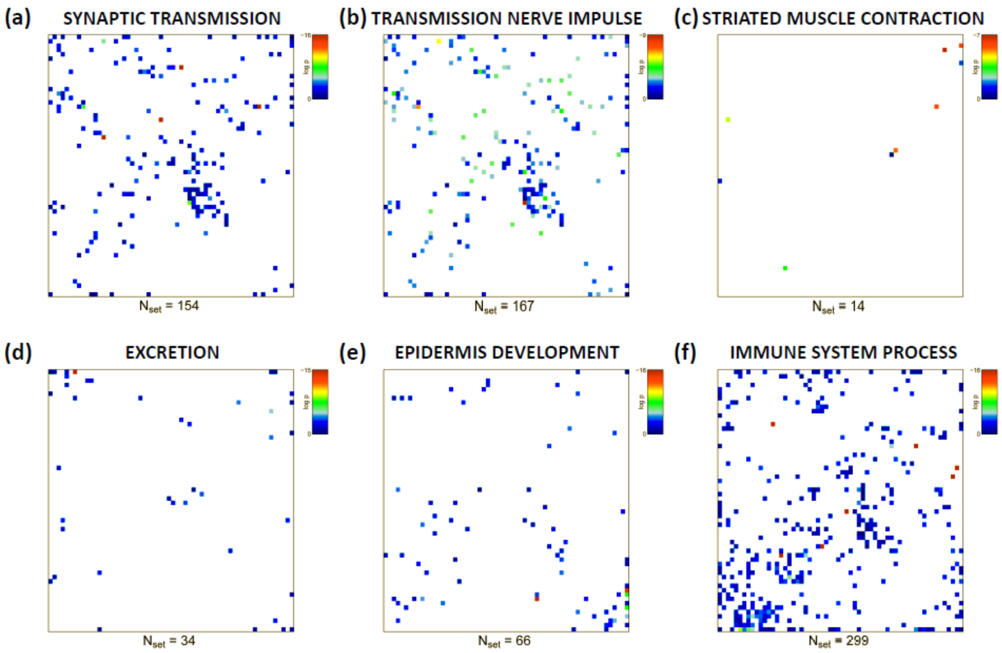


Figure S 18: Overrepresentation maps of six selected gene sets for the zoom-in SOM of the group of diverse tissues.

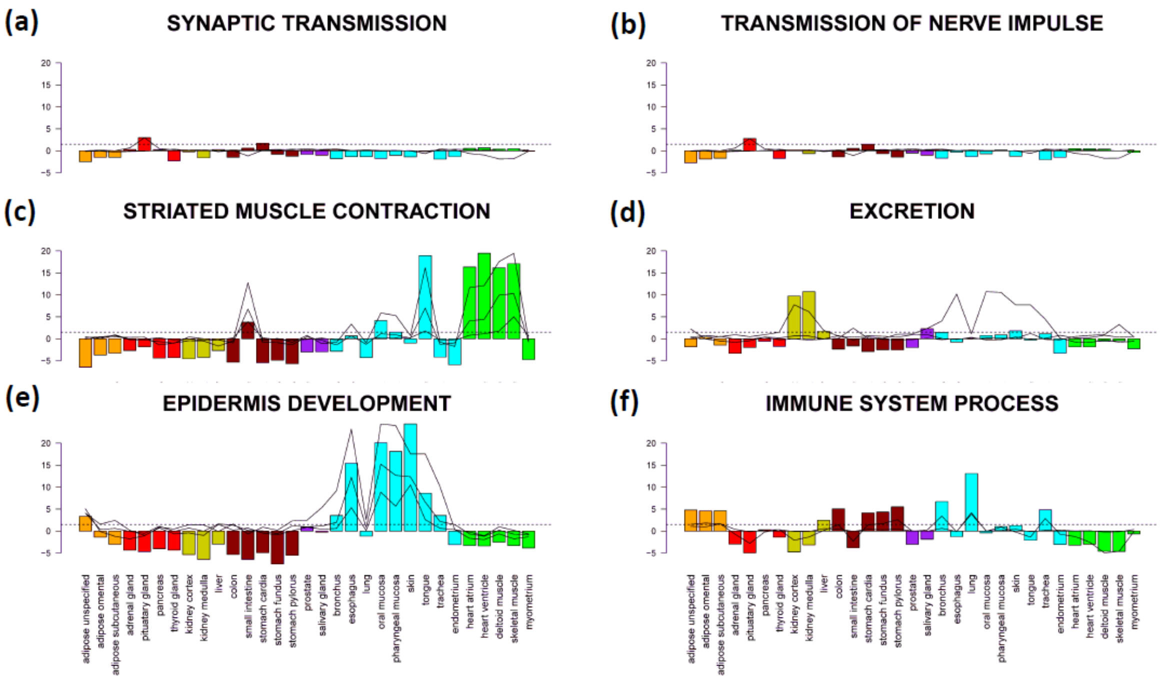


Figure S 19: Overexpression profiles of selected gene sets in the group of diverse tissues.

10. Selecting special sets of genes using ranking and expression criteria

We applied different criteria to select genes which are consistently expressed in all tissues studied. The first method uses the rank product approach [6]: The genes are ranked with decreasing expression score in each tissue. Then, the genes are re-ordered according the geometric mean of their ranks averaged over all tissue lists (i.e. by calculating the product of the tissue-specific ranks of each gene). Panel a of Figure S 20 shows the obtained logged average rank, $\log_{10} < r >= \frac{\log_{10} \mathrm{RankProduct}_r}{M}$ as function of rank number r for the three alternative scores, FC, WAD and t-shrinkage. The initial part of the curves steeply increases. It collects the genes which are consistently ranked on top of the individual tissue lists with small rank numbers. The slope of the curves markedly drops and virtually levels off for rank numbers greater than 3,000 revealing a background floor of weakly expressed genes with almost high rankings in the individual tissue lists. The three alternative scores provide very similar curves showing a transition between both classes of consistently high and weak expressed genes near r= 1000- 3000. We arbitrarily select 10% of the genes on top of the lists as consistently expressed (2,227). The slightly smaller value of the mean rank of the FC-scores in the transition range indicates the slightly better consistency of the FC-score at ranks smaller than 3000 compared with the alternative scores (red curve in Figure S 20a).

Panel b of Figure S 20 compares the gene lists obtained from the different scores using the respective CAT-plots of all three pairwise combinations of the respective lists. For the FC/WAD- and WAD/t-pairings of lists we found overlap of about 70% of the genes for rankings r< 2000 whereas FC/t-pairings are common to only 50% in this range. This result is slightly different if compared with the CAT-values of the tissue specific lists (see Figure S 2 and the main paper). In these comparisons the FC/WAD-lists best agree to about 70% whereas WAD/t- and FC/t- lists overlap to 50% only. Hence, rank-averaging over all tissues slightly modifies the overlap between the different lists. Nevertheless, the observed differences are relatively small confirming our conclusion that all scores considered provide reliable and partly complementary results.

In addition to the rank product approach we applied another one which makes use of the present call parameter, 0≤pc≤1, estimated in the normalization step of gene expression data (see ref. [2] for details). Figure S 21 shows the distribution of the number of genes significantly expressed in a certain number of tissues. The histogram ranges from genes which are consistently absent in all tissues (with present calls pc<1) to genes consistently present (pc=1) in all tissues. Interestingly, the distribution shows maxima at their left and right borders. This result reveals that many genes tend to be expressed either in most of the tissues or in only a few ones. A number of about 200 – 300 genes forms a sort of constant background level which characterizes the incremental cumulative tissue specificity of gene expression. Eisenberg and Lavanon [7] obtained a similar histogram using an alternative tissue data set. We collect the genes which are strictly absent or strictly present in all tissues into two groups for further analysis (see also Figure S 21).

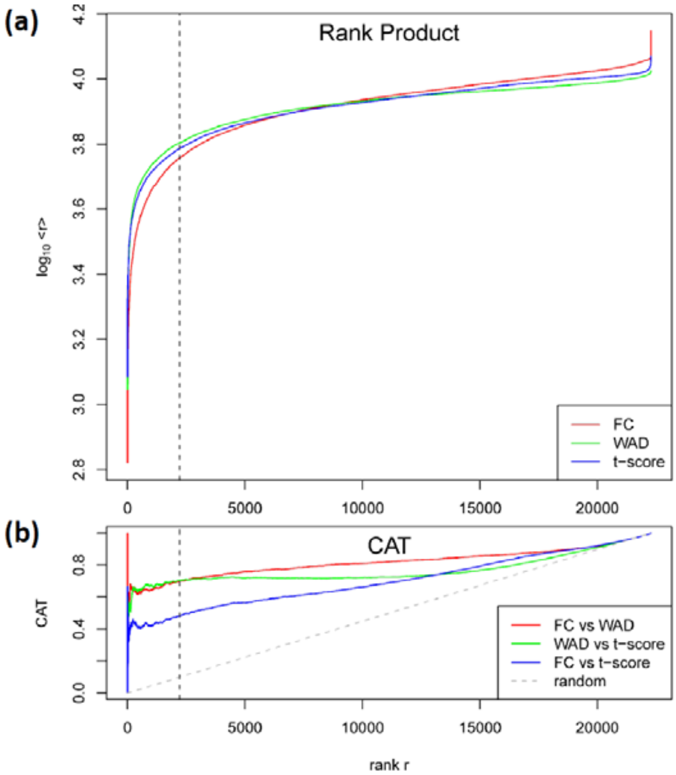


Figure S 20: Logged rank product of ranked gene lists of all 67 tissues as a function of the gene index. 2,227 genes are selected on top of the list over the range of steep slope as ’consistently high ranked’ (see the vertical dashed line).

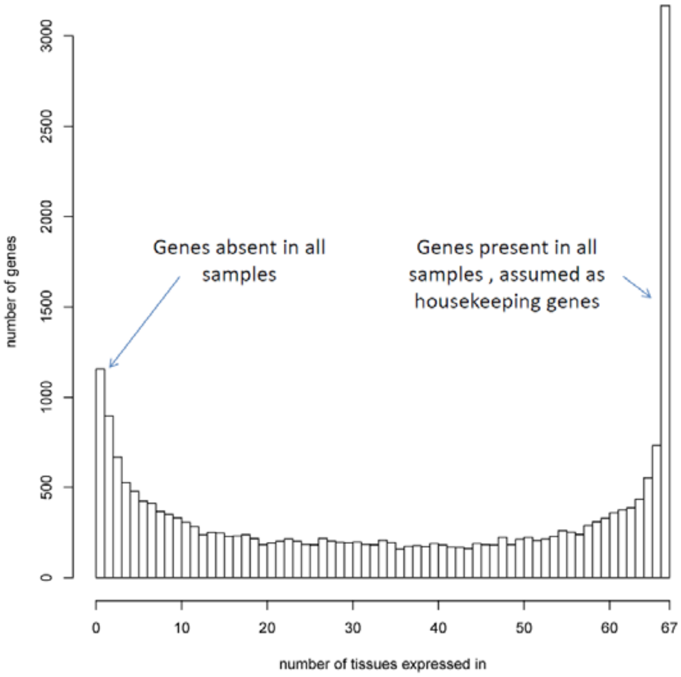


Figure S 21: Histogram of the number of genes expressed in different numbers of tissues using the present-call criterion. Absent genes (not expressed in any tissue) and housekeeping genes (expressed in all tissues) are found at the left and right positions, respectively.

11. Summary reports

Our SOM analysis of human tissues produces a series of additional reports allowing extraction of details not explicitly presented in the publication. Here we describe the reports characterizing the different tissue samples (sample summaries) and the different spots collecting groups of correlated and coexpressed genes (spot summaries).

11.1. Global and local tissue characteristics (sample summaries)

The ‘global summary’ sheet resumes the results of differential expression analysis of all genes studied whereas each ‘local summary’ sheet resumes the results of differential expression analysis of genes from one selected metagene spot. The collection of all spots detected in the respective tissue is depicted in the right small map shown in the ‘Global summary’ sheet. The small map shown in the ‘Local summary’ sheet depicts the spot selected for analysis. Figure S 22 presents the PDF-report for one particular tissue (accumbens). Table S 2 and Table S 3 provide glossaries of the data given in the sheets.

Table S 2: Glossary of data given in the global summary sheet

	Description
1	Sample name
2	General characteristics: number of differentially expressed genes (“#DE”); numbers of genes below particular fdr thresholds (“#genes with fdr < …”); number of genes covered by GO genesets (“#genes in genesets”); average scores (“<FC>”, “<shrinkage-t>”, “<p-value>”, “<fdr>”) over all genes
3	Expression profile (SOM) & map showing all spots selected using the 98%-quantile criterion
4	Ranking of differentially expressed genes: Affymetrix gene id (“Affy-ID”), gene symbol (“Symbol”), $\log_{10}$ fold-change (“log(FC)”), p-value(“p-value”), fdr (“fdr”), x-y-position of associated metagene in expression profile (“Metagene”, x-y are the tile numbers in horizontal and vertical directions, respectively) and GO-term (“GO-Term”)
5	Distribution of p-values from gene list (4) along with FDR-analysis: η_0 level is shown as horizontal line, fdr and Fdr are shown as dotted and dashed lines, respectively
6	GSZ enrichment list of top-20 over- and underexpressed genesets: GSZ score (“GSZ”) and p-value (“p-value”), number of genes in particular set (“#in”), GO-category and -term (“Geneset”)
7	Distribution of p-values from GSZ enrichment analysis (6), analogous to (5)

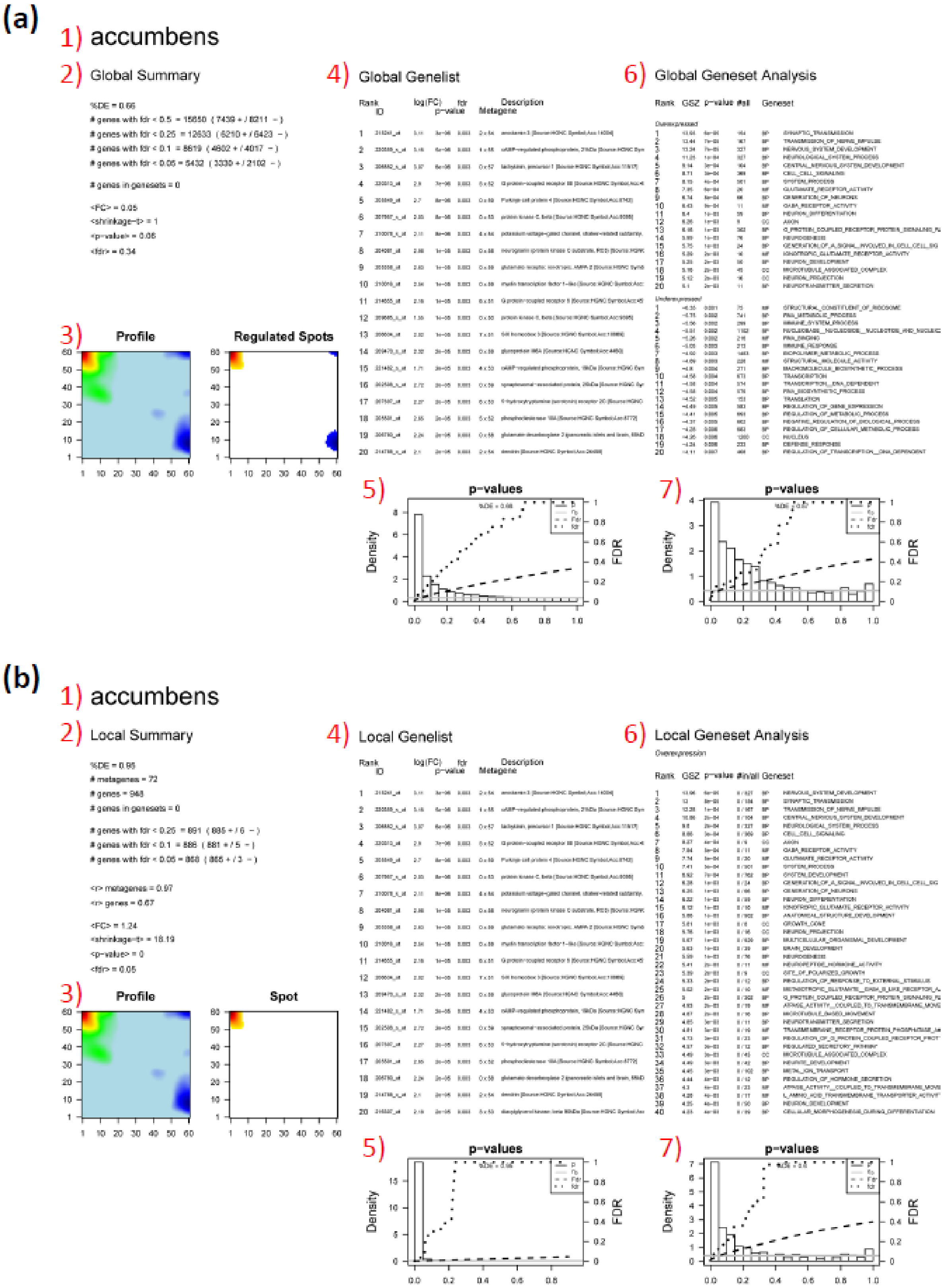


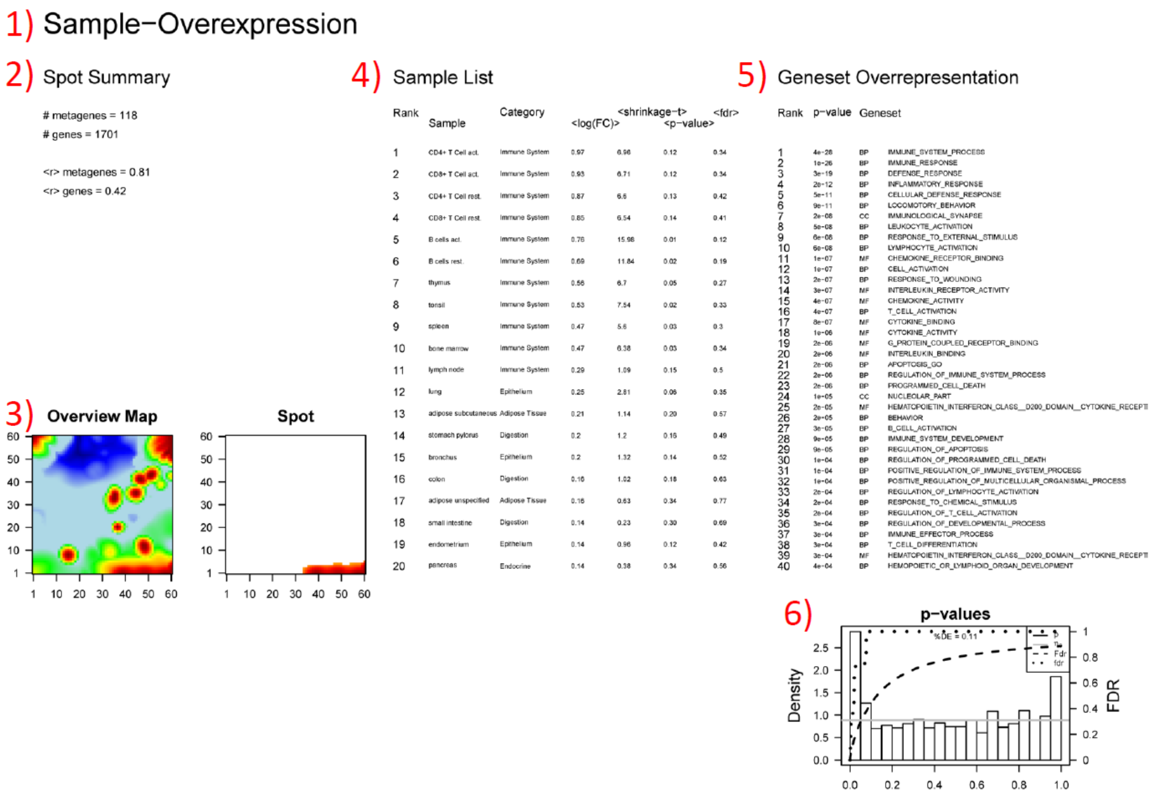
Figure S 22: Global (panel a) and local summary sheet (panel b) of accumbens. Note that only one out of three local summary sheets of this particular tissue is shown as example.

Table S 3: Glossary of data given in the local summary sheet

	Description
1	Sample name
2	General characteristics: number of differentially expressed genes (#DE); numbers of genes & metagenes in current spot; numbers of genes below particular fdr thresholds ("#genes with fdr < …"); number of genes covered by GO genesets ("#genes in genesets"); average scores (correlation coefficient r ("<r>"), "<FC>", "<shrinkage-t>", "<p-value>", "<fdr>") over spot genes
3	Expression profile (SOM map) & the local spot analyzed
4	Ranking of differentially expressed genes: Affymetrix gene id ("Affy-ID"), gene symbol ("Symbol"), $\log_{10}$ fold-change ("log(FC)"), p-value("p-value"), fdr ("fdr"), x-y-position of associated metagene in expression profile ("Metagene", x-y are the tile numbers in horizontal and vertical directions, respectively) and GO-term ("GO-Term")
5	Distribution of p-values from gene list (4) along with FDR-analysis: η_0 level is shown as horizontal line, fdr and Fdr are shown as dotted and dashed lines, respectively
6	GSZ enrichment list of top-40 over- resp. underexpressed genesets: GSZ score ("GSZ") and p-value ("p-value"), numbers of genes in particular geneset (#all) and associated genes found in current spot (#in), GO-category and -term ("Geneset")
7	Distribution of p-values from GSZ enrichment analysis (6), analogous to (5)

11.2. Spot summary reports

The sample summary reports collect analysis data for each sample either considering all genes or genes taken from a selected spot as described in the previous subsection. The spot summary reports pursue the orthogonal view: Each relevant over-/underexpression spot is characterized across all samples. Figure S 23 and Table S 4 show one example and the glossary of data listed in the sheet, respectively.



1) Sample−Overexpression

2) Spot Summary

metagenes = 118
genes = 1701

<r> metagenes = 0.81
<r> genes = 0.42

4) Sample List

Rank	Sample	Category	<log(FC)>	<shrinkage−t>	<p−value>	<fdr>
1	CD4+ T Cell act.	Immune System	0.97	6.96	0.12	0.34
2	CD8+ T Cell act.	Immune System	0.93	6.71	0.12	0.34
3	CD4+ T Cell rest.	Immune System	0.87	6.6	0.13	0.42
4	CD8+ T Cell rest.	Immune System	0.85	6.54	0.14	0.41
5	B cells act.	Immune System	0.76	15.98	0.01	0.12
6	B cells rest.	Immune System	0.69	11.84	0.02	0.19
7	thymus	Immune System	0.56	6.7	0.05	0.27
8	tonsil	Immune System	0.53	7.54	0.02	0.33
9	spleen	Immune System	0.47	5.6	0.03	0.3
10	bone marrow	Immune System	0.47	6.38	0.03	0.34
11	lymph node	Immune System	0.29	1.09	0.15	0.5
12	lung	Epithelium	0.25	2.81	0.06	0.35
13	adipose subcutaneous	Adipose Tissue	0.21	1.14	0.20	0.57
14	stomach pylorus	Digestion	0.2	1.2	0.16	0.49
15	bronchus	Epithelium	0.2	1.32	0.14	0.52
16	colon	Digestion	0.16	1.02	0.18	0.63
17	adipose unspecified	Adipose Tissue	0.16	0.63	0.34	0.77
18	small intestine	Digestion	0.14	0.23	0.30	0.69
19	endometrium	Epithelium	0.14	0.96	0.12	0.42
20	pancreas	Endocrine	0.14	0.38	0.34	0.56

5) Geneset Overrepresentation

Rank	p−value	Geneset	
1	4e−28	BP	IMMUNE_SYSTEM_PROCESS
2	1e−26	BP	IMMUNE_RESPONSE
3	3e−19	BP	DEFENSE_RESPONSE
4	2e−12	BP	INFLAMMATORY_RESPONSE
5	5e−11	BP	CELLULAR_DEFENSE_RESPONSE
6	9e−11	BP	LOCOMOTORY_BEHAVIOR
7	2e−08	CC	IMMUNOLOGICAL_SYNAPSE
8	5e−08	BP	LEUKOCYTE_ACTIVATION
9	6e−08	BP	RESPONSE_TO_EXTERNAL_STIMULUS
10	6e−08	BP	LYMPHOCYTE_ACTIVATION
11	1e−07	MF	CHEMOKINE_RECEPTOR_BINDING
12	1e−07	BP	CELL_ACTIVATION
13	2e−07	BP	RESPONSE_TO_WOUNDING
14	3e−07	MF	INTERLEUKIN_RECEPTOR_ACTIVITY
15	4e−07	MF	CHEMOKINE_ACTIVITY
16	4e−07	BP	T_CELL_ACTIVATION
17	8e−07	MF	CYTOKINE_BINDING
18	1e−06	MF	CYTOKINE_ACTIVITY
19	2e−06	MF	G_PROTEIN_COUPLED_RECEPTOR_BINDING
20	2e−06	MF	INTERLEUKIN_BINDING
21	2e−06	BP	APOPTOSIS_GO
22	2e−06	BP	REGULATION_OF_IMMUNE_SYSTEM_PROCESS
23	2e−06	BP	PROGRAMMED_CELL_DEATH
24	1e−05	CC	NUCLEOLAR_PART
25	2e−05	MF	HEMATOPOIETIN_INTERFERON_CLASS__D200_DOMAIN__CYTOKINE_RECEPT
26	2e−05	BP	BEHAVIOR
27	3e−05	BP	B_CELL_ACTIVATION
28	9e−05	BP	IMMUNE_SYSTEM_DEVELOPMENT
29	9e−05	BP	REGULATION_OF_APOPTOSIS
30	1e−04	BP	REGULATION_OF_PROGRAMMED_CELL_DEATH
31	1e−04	BP	POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
32	1e−04	BP	POSITIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS
33	2e−04	BP	REGULATION_OF_LYMPHOCYTE_ACTIVATION
34	2e−04	BP	RESPONSE_TO_CHEMICAL_STIMULUS
35	2e−04	BP	REGULATION_OF_T_CELL_ACTIVATION
36	3e−04	BP	REGULATION_OF_DEVELOPMENTAL_PROCESS
37	3e−04	BP	IMMUNE_EFFECTOR_PROCESS
38	3e−04	BP	T_CELL_DIFFERENTIATION
39	3e−04	MF	HEMATOPOIETIN_INTERFERON_CLASS__D200_DOMAIN__CYTOKINE_RECEPT
40	4e−04	BP	HEMOPOIETIC_OR_LYMPHOID_ORGAN_DEVELOPMENT

Figure S 23: Spot summary sheet for the ‘immune systems’ spot F.

Table S 4: Glossary of data given in the spot summary sheet

	Description
1	Summarization criterion
2	General characteristics: numbers of genes (“#genes”) & metagenes (“#metagenes”) in current spot; average correlation coefficient (“<r>”) among spots genes & metagenes
3	Summary map & mal showing the analyzed spot
4	Ranking of samples according to average fold change of genes within current spot: sample name (“Sample”); average scores (“<FC>”, “<shrinkage-t>”, “<p-value>”, “<fdr>”) of respective sample
5	HG-overrepresentation list: p-value (“p-value”), GO-category and –term (“Geneset”) are given for top-40 genesets
6	Distribution of p-values from HG-overrepresentation list (5) along with FDR-analysis: η_0 level is shown as horizontal line, fdr and Fdr are shown as dotted and dashed lines, respectively

12. References

1. Tibshirani R, Wasserman L: **Correlation-sharing for detection of differential gene expression**. *arXiv:math/0608061v1* 2006, **[math.ST]**.
2. Wirth H, Loeffler M, von Bergen M, Binder H: **Expression cartography of human tissues using self organizing maps**. *BMC Bioinformatics* 2011, **12**:306.
3. Yang H, Churchill G: **Estimating p-values in small microarray experiments**. *Bioinformatics* 2007, **23**(1):38-43.
4. Xie Y, Pan W, Khodursky AB: **A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data**. *Bioinformatics* 2005, **21**(23):4280-4288.
5. Goeman JJ, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues**. *Bioinformatics* 2007, **23**(8):980-987.
6. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments**. *FEBS Letters* 2004, **573**(1-3):83-92.
7. Eisenberg E, Levanon EY: **Human housekeeping genes are compact**. *Trends in Genetics* 2003, **19**(7):362-365.