RESEARCH PAPER

Systems Biomedicine 1:2, 1–23; April/May/June 2013; © 2013 Landes Bioscience

Portraying the expression landscapes of cancer subtypes

A case study of glioblastoma multiforme and prostate cancer

Lydia Hopp,[1,2,†] Henry Wirth[1,2,†] Mario Fasold,[1,2] and Hans Binder[1,2,*]

[1]Interdisciplinary Centre for Bioinformatics; Universität Leipzig; Leipzig, Germany; [2]LIFE–Leipzig Research Center for Civilization Diseases; Universität Leipzig; Leipzig, Germany

[†]These authors contributed equally to this work.

Keywords: gene co-regulation, self-organizing maps, clustering, cancer subtypes

Self-organizing maps (SOM) portray molecular phenotypes with individual resolution. We present an analysis pipeline based on SOM machine learning which allows the comprehensive study of large scale clinical data. The potency of the method is demonstrated in selected applications studying the diversity of gene expression in Glioblastoma Multiforme (GBM) and prostate cancer progression. Our method characterizes relationships between the samples, disentangles the expression patterns into well separated groups of co-regulated genes, extracts their functional contexts using enrichment techniques, and enables the detection of contaminations and outliers in the samples. We found that the four GBM subtypes can be divided into two "localized" and two "intermediate" ones. The localized subtypes are characterized by the antagonistic activation of processes related to immune response and cell division, commonly observed also in other cancers. In contrast, each of the "intermediate" subtypes forms a heterogeneous continuum of expression states linking the "localized" subtypes. Both "intermediate" subtypes are characterized by distinct expression patterns related to translational activity and innate immunity as well as nervous tissue and cell function. We show that SOM portraits provide a comprehensive framework for the description of the diversity of expression landscapes using concepts of molecular function.

Background and Introduction

Critical Assessment of Massive Data Analysis (CAMDA) addresses methodological challenges posed by the huge and ever increasing amount of data produced by new high-throughput technologies. It is expected that the exponential growth of high-throughput data in public databases will continue throughout the coming years.[1,2] The falling costs of data generation are, however, contrasted by the increasing costs for adequate analyses, a development that has been called "the $1,000 genome and the $100,000 analysis" problem.[3] Analysis challenges, in the first instance, arise from elementary hardware requirements such as a need for large data storage facilities, high-capacity network infrastructure, and computing power to handle the huge amounts of data. In the second instance, researchers need faster bioinformatics algorithms, e.g., for sequence analysis and data screening, and appropriate efficient software implementations. Last but not least, tools to exploit the information content of the data in an effective and intelligent way are urgently required in order to extract new insights. This includes tasks such as compressing and filtering of high dimensional data, feature selection, linkage with the biological context using previous knowledge, and visualization. This is particularly important because an intuitive visualization of massive data supports quality control, promotes the discovery of qualitative relationships, and facilitates the development of new hypotheses. A contest study of the Microarray Quality Control Consortium (MAQC) involving 17 teams of researchers showed that differences in proficiency between data analysis teams, especially in their experience levels, could strongly affect the quality of analysis results.[4] This finding emphasizes the need for easy-to-use, intuitive, and easy-to-interpret data analysis approaches.

With this motivation, we here apply self-organizing maps (SOMs), a feature-centered clustering method from the field of machine learning,[5] to a data set of large scale gene expression profiles from Glioblastoma Multiforme and prostate cancer patients. The Glioblastoma Multiforme profiles formed one of the official contest data sets of CAMDA 2011. The prostate cancer profiles extend and complement the application range for which we demonstrate our method. For an unbiased examination of cancer, an integrative approach is required to investigate complex

*Correspondence to: Hans Binder; Email: binder@izbi.uni-leipzig.de
Submitted: XX/XX/XXXX; Revised: XX/XX/XXXX; Accepted: XX/XX/XXXX; Published Online: 04/01/2013
http://dx.doi.org/10.4161/sysb.25897

gene–environment interactions, rather than a testing of individual genes or pathways. Our method thus simultaneously searches for features which are differentially expressed and which are correlated in their profiles across the studied samples.[6] We aim to merge these genes into functional modules that are defined as sets of genes associated with a particular biological process (e.g., inflammation, cell division, etc.), and seek to characterize disease-specific changes in the resulting interaction network.

As a special feature, SOMs can highlight molecular phenotypes at different resolutions. We will demonstrate the power of our method in characterizing the diversity of gene expression profiles across cancer subtypes by providing a modular view of the gene expression patterns. The resulting portraits guide further functional analyses and allow an identification of misclassified samples contributing to new insights and improved quality control in large clinical studies.

Results and Discussion

Hook-calibration of GBM expression data and comparison with RMA

We applied our own microarray data preprocessing pipeline, which includes hook calibration, quantile normalization, and centralization scaling (see **Supp. File 1**). Hook calibration allows identification of batch effects and hybridization biases due to varying scanner settings, washing effects, and different amounts and quality levels of RNA (**Supp. File 1**). Our preprocessing pipeline thus differs in many respects from the standard RMA preprocessing used to generate the Level 2 Glioblastoma Multiforme (GBM) expression data from the TCGA website. The analysis results we present below are therefore largely independent of previous analyses based on RMA expression values which, for example, have been used for classification of GBM subtypes.[7]

As it may be debatable which preprocessing approach performs best, in the absence of an accepted "gold standard," we compare analysis results using either method (**Supp. File 1**). Preprocessing seems to have little of an effect on the most prominent properties of the expression landscape (and thus the classification of cancers into different subtypes) or the biological context of the main expression modules, with some differences for individual genes only found at intermediate and lower expression levels, which are not further investigated here.

First level SOM portraits of tumor samples and subtypes

Our SOM machine learning algorithm transforms the whole genome expression landscape of a patient into one mosaic image, consisting of 40 × 40 tiles for prostate cancer progression (PCP) or 50 × 50 tiles for GBM, with each tile representing one "meta-gene." These meta-genes serve as prototypes of groups ("mini-clusters") of co-regulated genes, the number of which usually varies from meta-gene to meta-gene. **Figure 1** and **Figure 2** display the 1st level SOM portraits of the expression landscapes in GBM and PCP samples, respectively. We sort them into different groups following previous classifications of cancer subtypes or progression stages.[7,8] Each mosaic image exhibits a characteristic texture serving as a fingerprint of the transcriptional activity in the respective cancer sample. These images reveal regions of over- and underexpression ("spots") which characterize the different cancer subtypes in GBM (**Fig. 1**) and stages of PCP (**Fig. 2**). Relatively stable and consistent spot-patterns can dominate (e.g., for MES and PN samples of GBM) or relatively heterogeneous and volatile patterns can be observed (e.g., for CL and NL samples of GBM).

We calculate the mean SOM-portrait of each class (GBM subtype or PCP stage) by averaging patient samples for each meta-gene. On one hand, this averaging may cancel out individual, highly fluctuating features. On the other hand, it amplifies consistent class-specific features. For example, the MES subtype of GBM and normal brain tissue are characterized by two spots in opposite corners of the map, one overexpressed and the other underexpressed in MES samples and vice versa in normal (NOR) samples. These class-specific spots collect highly populated meta-genes with a strong and well delineated signal (see the supporting maps in **Supp. File 1**). The mean portraits of the other three GBM subtypes are more diffuse: the PN, CL, and NL subtypes are characterized by two or three specific spots per subtype.

The mean stage-related PCP portraits show similar properties. Some regions are involved in more than one PCP stage. The spots of subsequent stages tend to overlap, and also spots of the final MET and the initial BPH stages tend to overlap. Overall, the stage-specific spot pattern thus "rotates" along the border of the map in a clockwise direction with progressing cancer.

The log-logFC scale amplifies small expression levels in the individual sample portraits. The mean portraits in log-logFC scale show essentially the same basic features as those in logFC scale. They are, however, richer in details, enabling the identification of more subtle differences between the subtypes. For example, the mean log-logFC portraits of the MES and PN subtypes of GBM are complimentary to each other, i.e., overexpressed red regions in the MES-image largely correspond to underexpressed blue regions in the PN image, indicating strongly anti-correlated expression patterns in the two subtypes.

In summary, SOM images capture the individual sample expression landscape in terms of characteristic color portraits, which enable a visual inspection of subtype-specific features, with spot-like regions representing clusters of differentially expressed and co-regulated genes. In addition, averaging over groups of samples and the considered application of different color scales selects and amplifies subtype- or stage-specific features.

Characterizing the expression phase space: Second level SOM and ICA

The 2nd level SOM analysis visualizes the similarity between individual 1st level SOM portraits, with tiles representing "meta-samples." The four GBM subtypes show a clusters of samples in different, and well separated regions of the map (**Fig. 3A and B**), consistent with earlier subtype specifications.[7] The ten normal samples (subtype NOR) occupy a very compact area in the top right corner of the map. Their SOM portraits most closely resemble those of the neural subtype (NL), namely both subtypes show a common overexpression spot in the bottom left corner in the individual portraits which are not present in the mean portraits of the other GBM subtypes.

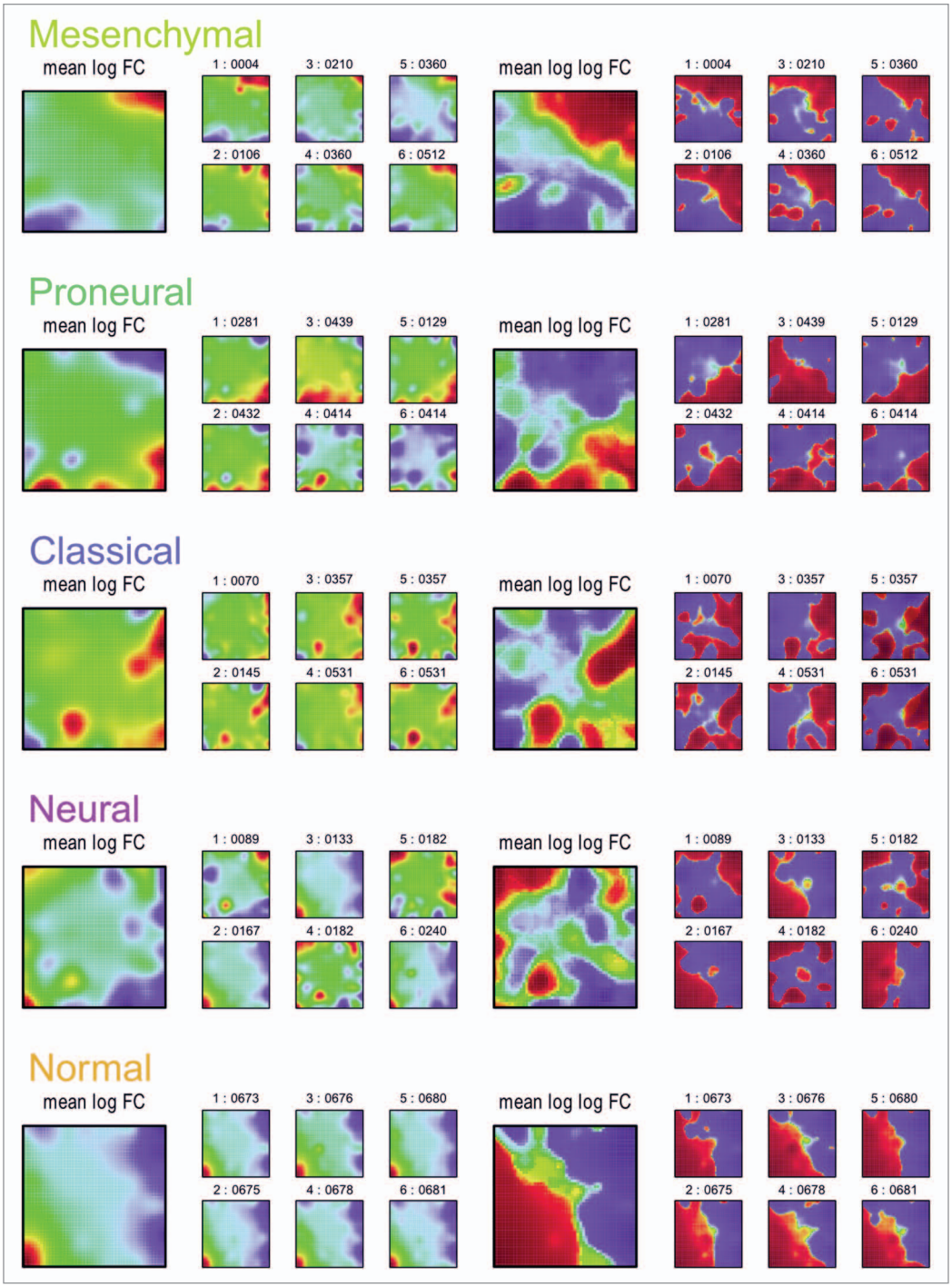

Figure 1. SOM gallery of GBM subtypes. The small mosaic images refer to selected individual tumor samples assigned to four GBM subtypes and normal brain tissue. The large mean images are calculated by averaging over all samples of each subtype (see Materials and Methods section). The images in the left part of the figure use a logFCscale where FC denotes the fold change of the expression of each meta-gene with respect to its mean expression in all samples (maroon to red refers to the 90-percentile and light to dark blue to the 10-percentile thresholds, respectively). This highlights areas of strong differences. The right part uses the smoother log-logFC scale, which increases the contrast in areas of weaker signals. Up and downregulated meta-genes are colored in red and blue, respectively. The SOM here uses a quadratic grid of size 50 × 50 to distribute the expression profiles of the 22.777 genes available on the HT-HU-HGU133A arrays studied. A complete gallery of all individual sample portraits is available in **Supplemental File 2**. The population map showing the distribution of genes is shown in **Supplemental File 1**.

As a complementary method, and for comparison, independent component analysis (ICA) was applied to the 1st level SOM portraits. Three dimensional and two dimensional component plots are shown in **Figure 3C and D** for GBM, with samples similarly separated. Additional information can be extracted from the distribution of the subtypes along the independent

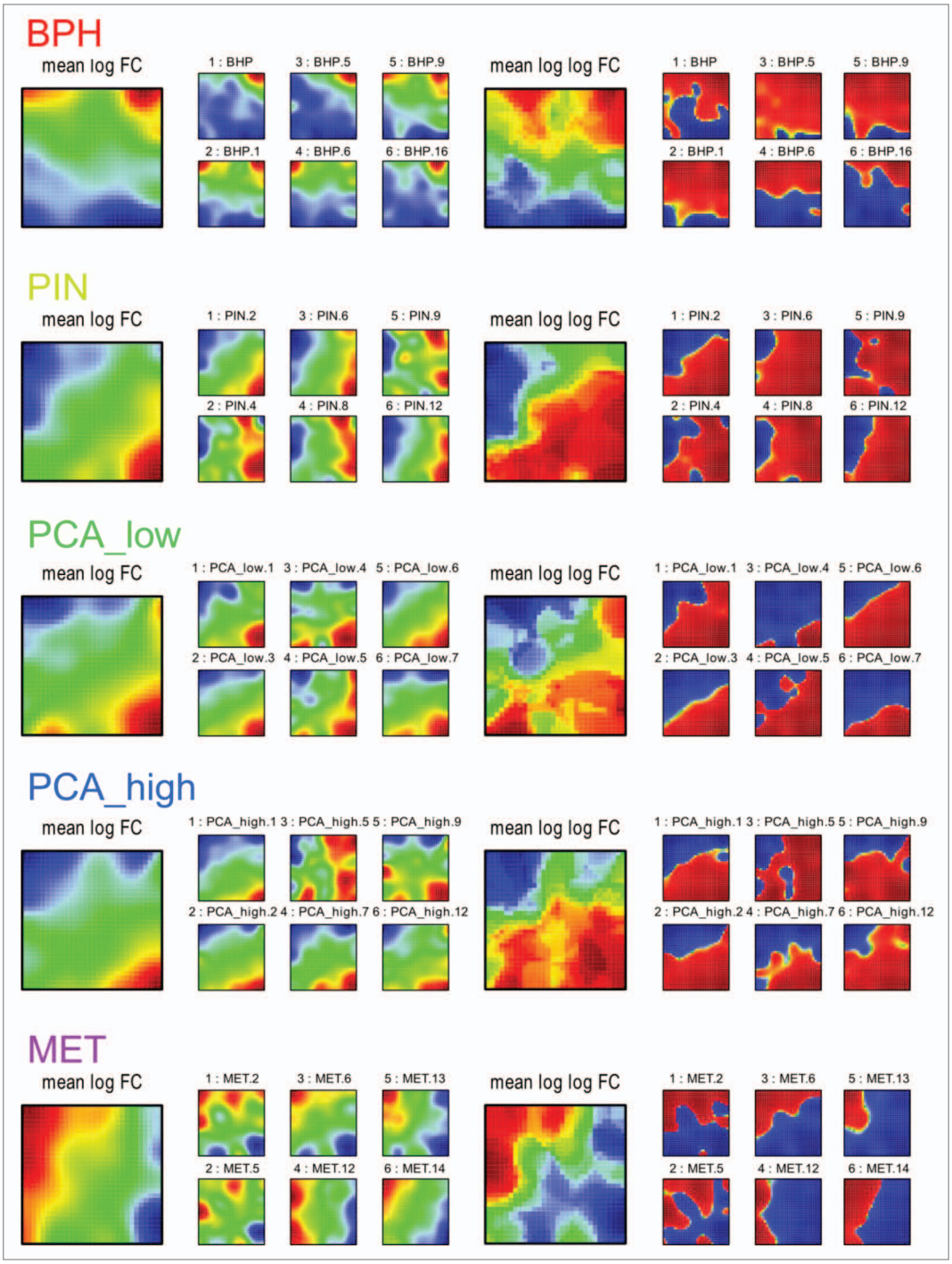

Figure 2. SOM gallery of PCP stages. See legend of **Figure 1**.The SOM uses a quadratic grid of size 40 × 40 to distribute the expression profiles of the 4,181 genes available on the Human 20K Hs6 arrays studied. A complete gallery of all sample portraits is available in **Supplemental File 3**. The population map showing the distribution gene numbers is shown in **Supplemental File 1**.BPH = benign prostatic hyperplasia, PIN = prostatic interepithelial neoplasia, PCA_low = low-grade, PCA_high = high-grade, and MET = metastatic stages of prostate cancer.

component axes IC1, IC2, and IC3. The GBM subtypes mainly separate in the IC1/IC2 plane whereas the normal (NOR) samples separate from most of the cancer samples along the IC3 axis. The MES and PN subtypes systematically differ in their IC2 coordinate whereas the NL and CL subtypes can be distinguished by their IC1 coordinate. Hence, the two groups of subtypes are obviously characterized by two sets of genes that change independently. These, in turn, are mostly independent of those that differentiate between cancer and normal samples along the IC3 axis. Also, the NL samples varying along this component reflect the similarities of expression patterns between NL and normal samples.

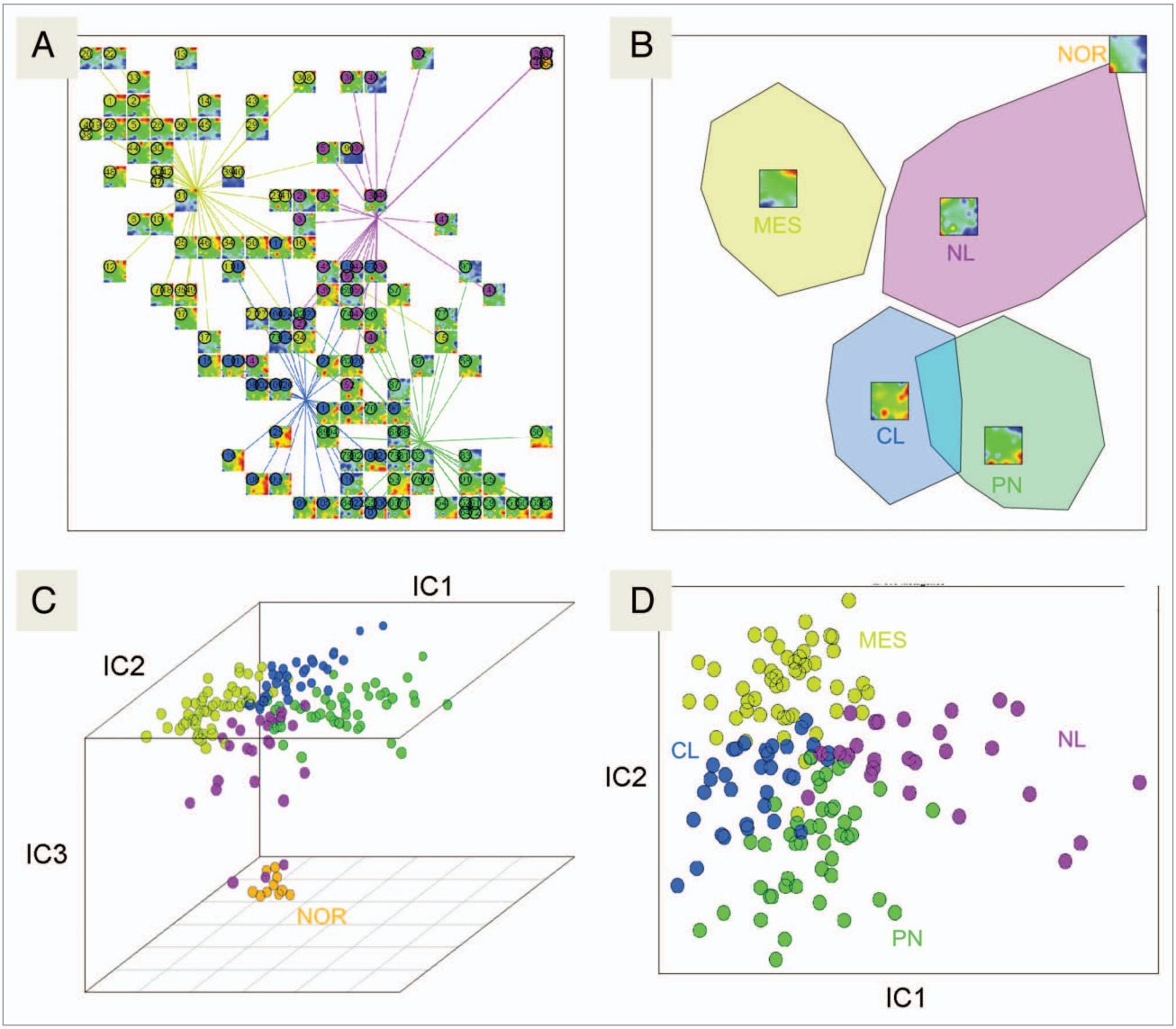

Figure 3. The 2nd level SOM and ICA similarity analysis of GBM cancer subtypes. (**A**) The position of each GBM sample is marked by the respective 1st level SOM image. Samples of the same GBM subtype are connected by lines drawn to the centroid of the respective class. (**B**) The regions occupied by the samples of the four subtypes are represented by the colored polygons. The mean SOM portraits of each GBM subtype are located in the center of the respective polygon. The four GBM subtypes occupy roughly the four quadrants of the map whereas the 10 normal tissue samples aggregate into one tile of the SOM in the top-right corner of the map. (**C**) The three-dimensional distribution of samples in each GBM subtype and normal tissue samples is shown in the space spanned by the three leading independent components IC1 – IC3. (**D**) The projection of the GBM subtypes into the IC1/IC2-plane.

The different PCP stages, in contrast, are located in extended, largely overlapping regions in the 2nd level SOM representation (**Fig. 4A**). The first and final stages (BPH and MET) can be better distinguished, whereas the intermediate stages, PIN, PCA_low, and PCA_high, are found essentially in the same region of the map. The U-shaped "trajectory" of the progression reflects the fact that a significant portion of the genes is similarly expressed in the initial BPH stage and the final MET stage, but differently expressed in the intermediate PIN and PCA stages.

In summary, the 2nd level SOMs visualize the similarity of 1st level SOM spot patterns in terms of partly overlapping regions representing the different subtypes. In general, the symmetry of the spot patterns in the 1st level mean SOM portraits and the arrangement of the subgroups in the 2nd level SOM show similar and partly complementary properties. The complementary ICA analysis allows an estimation of expression change dependencies associated with the different subtypes.

Characterizing pairwise similarities between portraits: Dendrograms and correlation nets

We calculated Pearson correlation coefficients based on the meta-gene states, comparing all pairwise combinations of samples as an alternative approach for studying similarities between the samples (see the pairwise correlation maps in **Supp. File 1**). The obtained covariance structure of the data are visualized using the maximum spanning tree (MST) and the correlation net (CN) representations shown in **Figure 5**. Importantly, all these visualizations

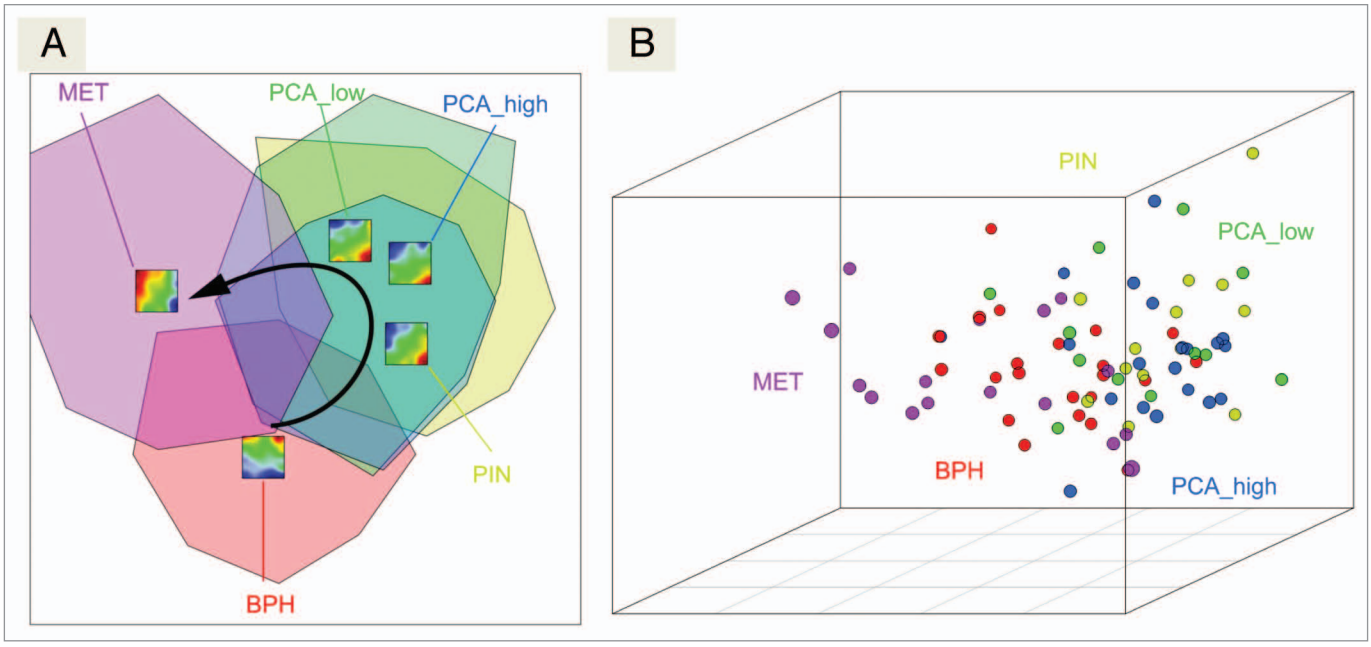

Figure 4. The 2nd level SOM ICA similarity analysis of PCP stages. (**A**) The 2nd level SOM polygon representations of PCP stages. Note that the spot pattern in the mean expression maps of PCP virtually rotates with progressing cancer giving rise to a U-shaped trajectory in the map (see arrows); (**B**) Three dimensional ICA.

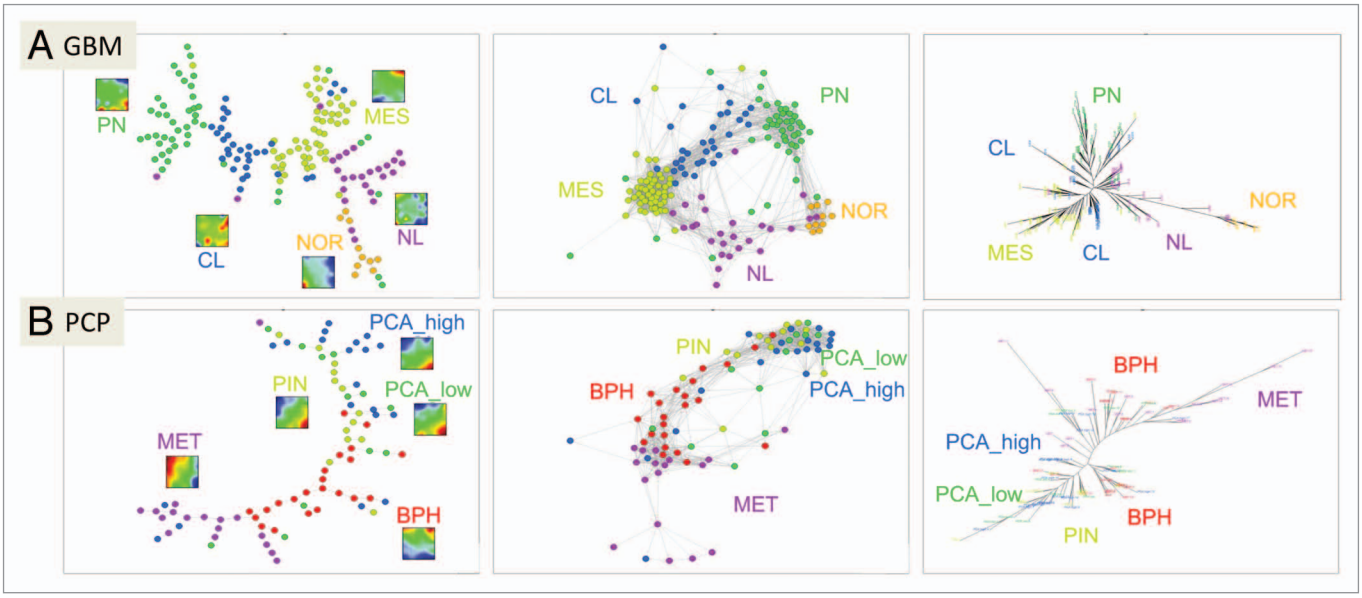

Figure 5. Similarity analysis of the two tumor data sets: (**A**) GBM and (**B**) PCP. Maximum spanning trees (MST) are shown in the left panels together with selected SOM portraits of each subtype. The middle panels show the corresponding correlation nets, while "phylogenetic" cluster trees are provided in the right panels for comparison.

of similarities are based on the meta-genes, providing a better resolution comparison than single gene-based similarity analyses due the lower noise after SOM dimensionality reduction.[6,9]

The MST plot shows a chain-like structure which connects the samples with the strongest mutual correlations. This has the key advantage that it converts multi-dimensional clusters into a relatively simple graph. The CN representation, in turn, transforms the data into a more detailed network of data points. It also considers weaker mutual correlations, shown as lines between the respective samples. The lengths of these lines are approximately inversely proportional to the respective correlation coefficients.

The MST and, especially, the CN plots of the GBM data set reveal similarities between the GBM subtypes which are less evident in the 2nd level SOMs (compare **Fig. 5A** and **Fig. 3B**): For

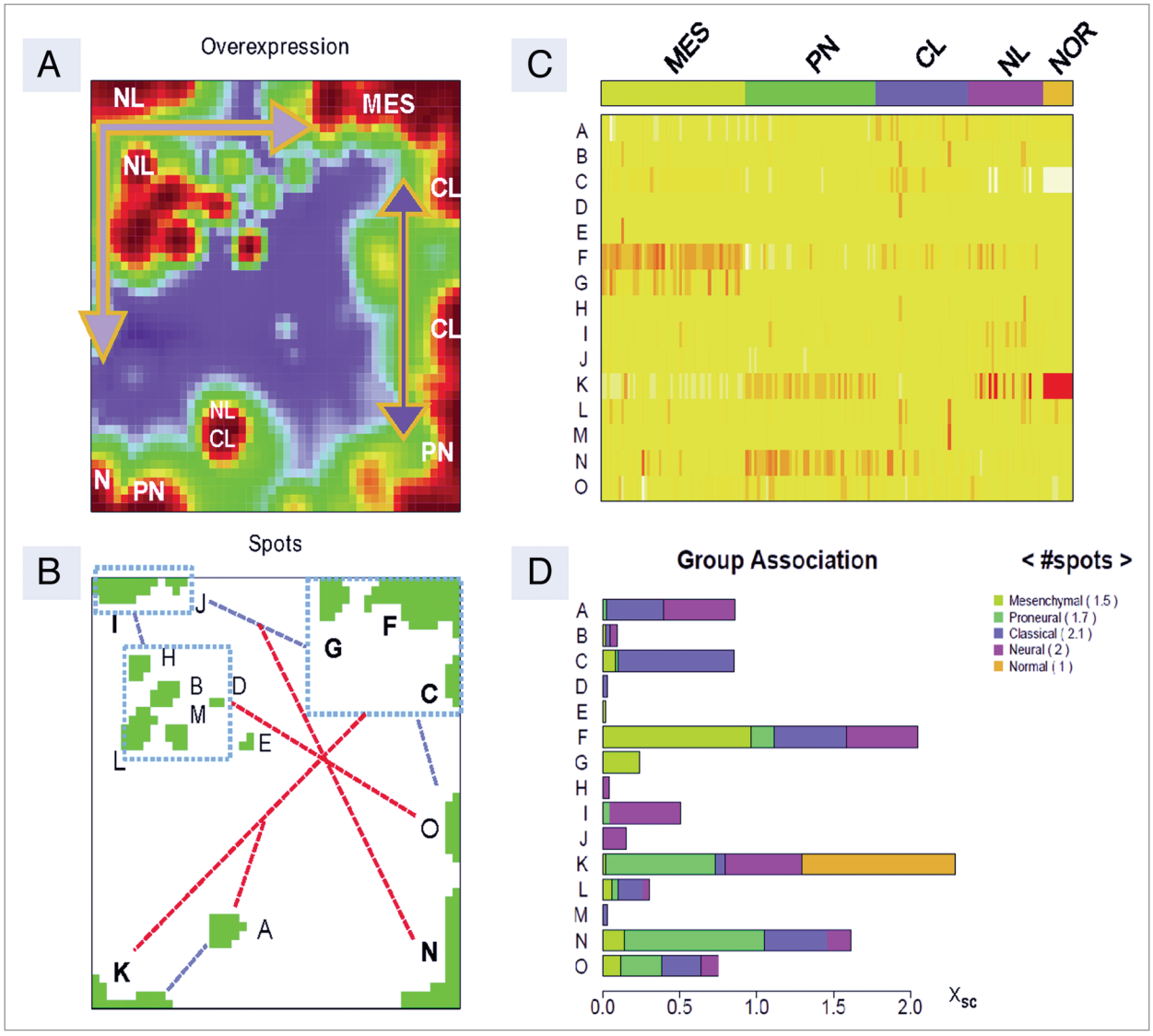

Figure 6. Overexpression spot characteristics of GBM. (**A**) The overexpression summary map collects all spots with overexpression observed in the individual profiles into one map. GBM subtypes associated with particular spots are indicated in the map. (**B**) Construction of the overexpression spot map defining the spots used for further analysis. Spots are labeled by capital letters. The blue rectangles include highly correlated spots ($r > 0.7$). The blue and red dashed lines connect correlated ($0.4 < r < 0.7$) and anti-correlated ($r < -0.6$) spots respectively. C) The heatmap shows the mean meta-gene expression for each spot A..O.. The samples are sorted according to their subtype membership as indicated along the top of the figure . (**D**) The bar plot shows the fraction of samples of each subtype which exhibit a given spot. The total bar length represents overall frequency, while colors indicate the frequency by subtype. The average numbers of spots in the portraits of each subtype are given in parantheses in the top right legend (MES = Mesenchymal, PN = Proneural, CL = Classical, NL = Neural, NOR = Normal).

example, the CN plot suggests that the NL and PN subtypes share more similarities with the NOR reference samples than the CL and MES subtypes. Note also that the PN and MES samples accumulate within compact clusters whereas the CL and NL clusters are fuzzier. The CL subtype forms a continuum between the MES and PN samples, which distribute along two separate branches. The CN plot forms a "donut-like" structure composed of alternating compact and fuzzy clusters. The intermediate NL and CL subtype samples in the fuzzy clusters link the compact MES and PN subtype clusters.

These similarity relationships could be transformed into star-like dendrograms similar to phylogenetic trees using the neighbor-joining algorithm with Euclidean distance metrics (**Fig. 5**, right panels). The more localized MES and PN subtypes form clearly separate branches whereas the intermediate NL and CL subtypes occupy several more central branches. The dendrogram for GBM finally reveals that the NL samples group along a separate branch together with the normal samples (NOR).

The CN and dendrogram plots of the PCP samples show a slightly different, backbone-like structure, reflecting the temporal progression of prostate cancer. As a rule of thumb, the mutual distance increases with progressing stages (BPH to PCA_low, to PCA_high). The final stage MET samples, however, are again found near initial stage BPH samples, consistent with the

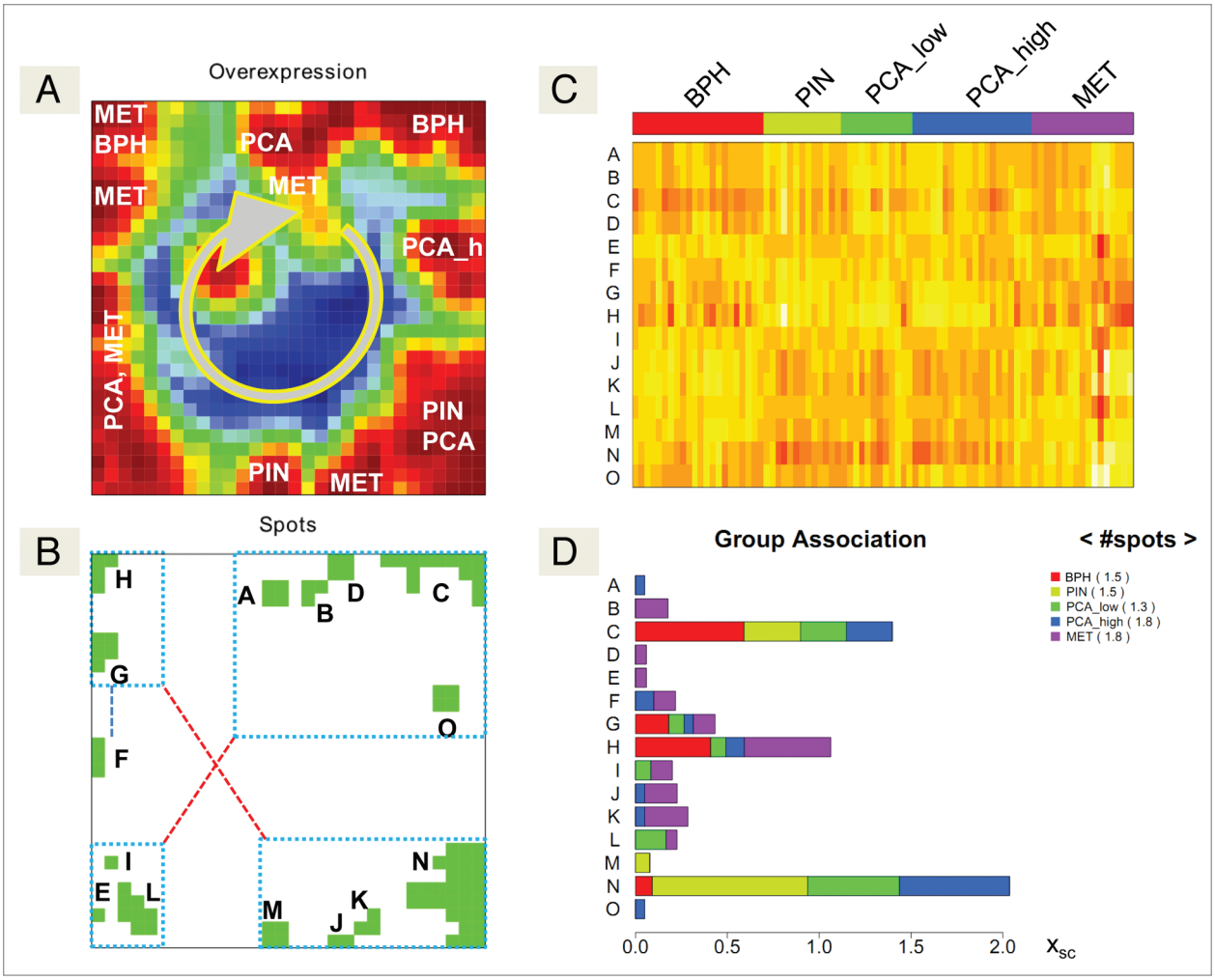

Figure 7. Overexpression spot characteristics of PCP. See legend of **Figure 6** for details. The arrow in (**A**) represents the appearance of overexpression spots with cancer progression.

U-shaped arrangement of the PCP stages in the 2nd level SOMs (see **Fig. 4A**), suggesting a greater similarity between the first and final stages than between the first and intermediate PCA_low and PCA_high stages.

Describing the expression landscapes: Spot analysis

In the next step, we analyze the spot patterns on SOM portraits to identify differences and common properties shared between the cancer subtypes. Unique or more common spots can provide information about the functional impact of gene activities specific to cancer subtype. **Fig. 6A** shows the overexpression summary map of GBM, which collects all spots with overexpression observed in the individual GBM portraits into one master map (see also[6]). Each distinct region of meta-genes in the portraits exceeding a certain overexpression threshold (typically the 98-percentile in at least one sample) defines a spot on the overexpression map, labeled by capital letters in **Figure 6B**. In total, we identified 15 such spots, "A" to "O," for GBM. Lists of genes contained in the spots are given in **Supplemental File 8**. **Figure 6C** visualizes the mean expression level across the meta-genes of each spot for all samples. This heatmap thus provides an overview over the subtype-specific expression activity in each spot. For example, spot "G" and partly also spot "F" are selectively overexpressed in the MES subtype, and spot "I" in the NL subtype, whereas spots "M" and "O" show sample specific activity, not specific to any subtype.

Our spot selection algorithm thus identifies both rare and frequent spot patterns. We next assess the relative frequency x_{sc} of each spot (see **Equ. 7** in Materials and Methods). As shown in **Figure 6D**, the most abundant spots (K, F, and N) are found in about 30% to 50% of the samples of all five subtypes. They are, however, relatively unspecific for tumor subtypes. Spot "F," for example, is found in almost every MES sample, in about 40% of the PN and NL samples, and some PN samples. Spot "K" appears in 40% of the NL samples, about 80% of the PN samples, and in nearly in every normal sample. In contrast, other spots are more specific to particular subtypes: Particularly, spot "C" is largely occurs in the CL subtype, spot "G" is unique to MES, and spot "I" largely occurs in the NL subtype. Other spots such as "D," "E," and "H" are very rare with relative frequencies less than 0.1.

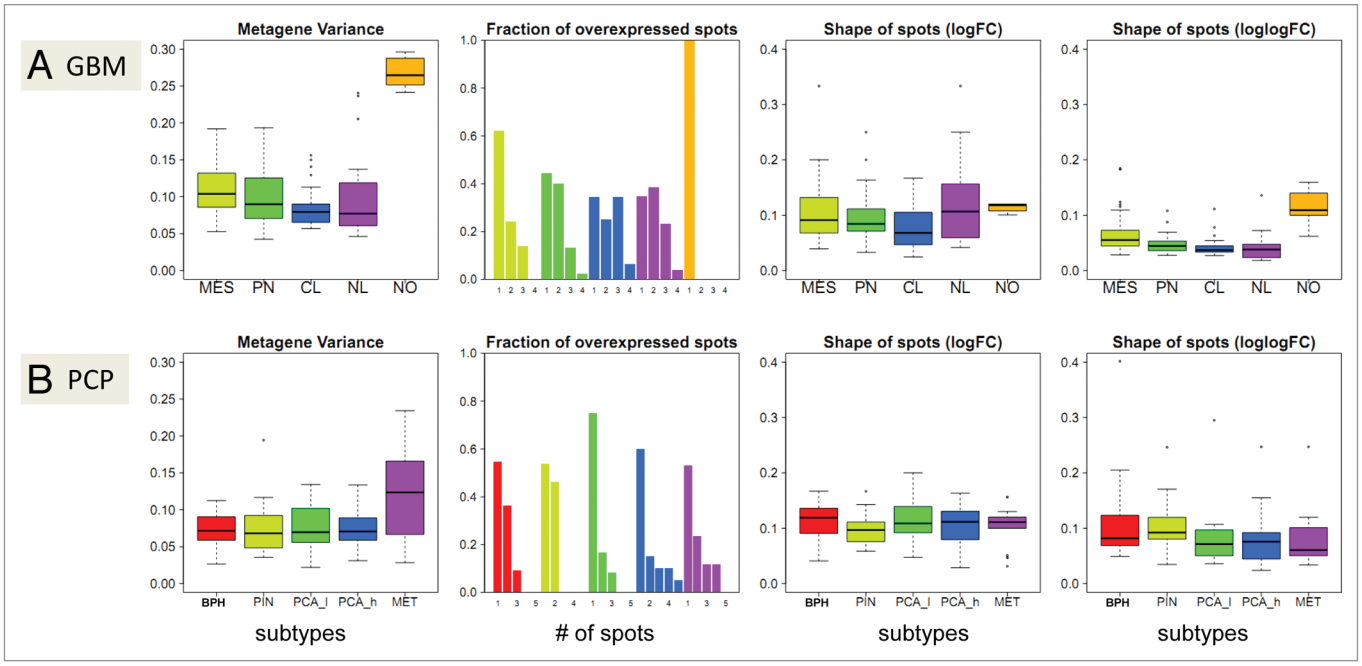

Figure 8. Global characteristics of the expression landscapes of cancer subtypes: (**A**) GBM and (**B**) PCP. From the left to right panels show meta-gene variance, **Equation 3**, the probability distribution of the portraits to show a certain fraction of spots with overexpression, and the spot shape parameter, **Equation 6** of spots from portraits in logFC and log-logFC scales. The spot shape parameter reflects circularity, i.e., the fuzzier a spot, the smaller the shape parameter, **Equation 6**. The global characteristics of spots with underexpression are shown in **Supplemental File 1**.

In order to discover covariance between the meta-gene expression profiles in different spots, we calculated pairwise correlation maps and maximum spanning trees exploring relationships between spots (see **Supp. File 1**). As a rule of thumb, neighboring spots are strongly positively correlated and spots located in opposite corners of the map are often strongly anti-correlated. For example, spots "C," "G," and "F" are highly correlated (**Fig. 6B**, blue dashed lines), whereas the spots "I" and "N" are anti-correlated (**Fig. 6B**, red dashed lines).

Spot analysis of PCP (**Fig. 7**) detects 15 spots with overexpression, and only 6 are relatively frequent ($x_{sc} > 0.2$). Ten spots are observed in MET samples, reflecting that the expression patterns of the metastatic cancer samples are highly diverse with spots located in nearly all regions of the map. In contrast, the PIN and BPH samples show only 3 and 4 spots with overexpression, respectively.

Analogous results from analyzing spots with underexpression are in line with these observations (**Supp. File 1** and see below).

Global characteristics of the expression landscapes

In the next step, we characterized the global properties of the expression landscapes, such as the extent of variability, typical spot numbers and shapes for each data set. Meta-gene variance and the circularity of spot shapes in the log-logFC portraits change similarly across GBM-subtypes. Moreover, the frequency of spots with overexpression decreases as spot circularity and the variance of meta-gene expression increase (**Fig. 8A**). Hence, the variability of meta-gene expression, the number of spots with overexpression, and their circularity/fuzziness are obviously closely related properties. Highly variant meta-gene landscapes observed in NOR samples are associated with less and more compact spots than those with less variant meta-genes typically observed for the intermediate NL and CL subtypes. The shape of the spots in logFC and log-logFC scales reflect different aspects of the data: For example, the mean spot circularity of NL is the smallest among the GBM subtypes in log-logFC scale portraits, whereas it has more variable and larger spot circularity in logFC scale portraits.

The MES, PN, and NOR subtypes have more stable expression landscapes than the NL and CL subtypes (**Fig. 8A**, fewer and less fuzzy spots) and they assemble into more compact clusters in analyses of sample similarity (e.g., the PN samples in **Fig. 5**). The higher stability of the class specific spot patterns in the individual sample portraits thus seems related to compact clusters in similarity analyses and vice versa. This illustrates the relation between global properties of the expression landscape of each subtype and the degree of similarity between its individual samples. While for GBM more stable patterns (fewer and less fuzzy spots) are associated with larger differences in individual expression landscapes (meta-gene variance),for PCP one finds the opposite relationship, i.e., the increase of the metagene variance upon progressing cancer is associated with an increasing fuzziness of the spots especially in log-logFC scale (**Fig. 8B**).

The global characteristics of overexpression describe landscapes mostly in the range of intermediate and high expression levels. We also analyzed the landscapes in the range of low expression values by similarly studying underexpression. Results for over- and underexpression largely match, and differences are discussed in **Supplemental File 1**.

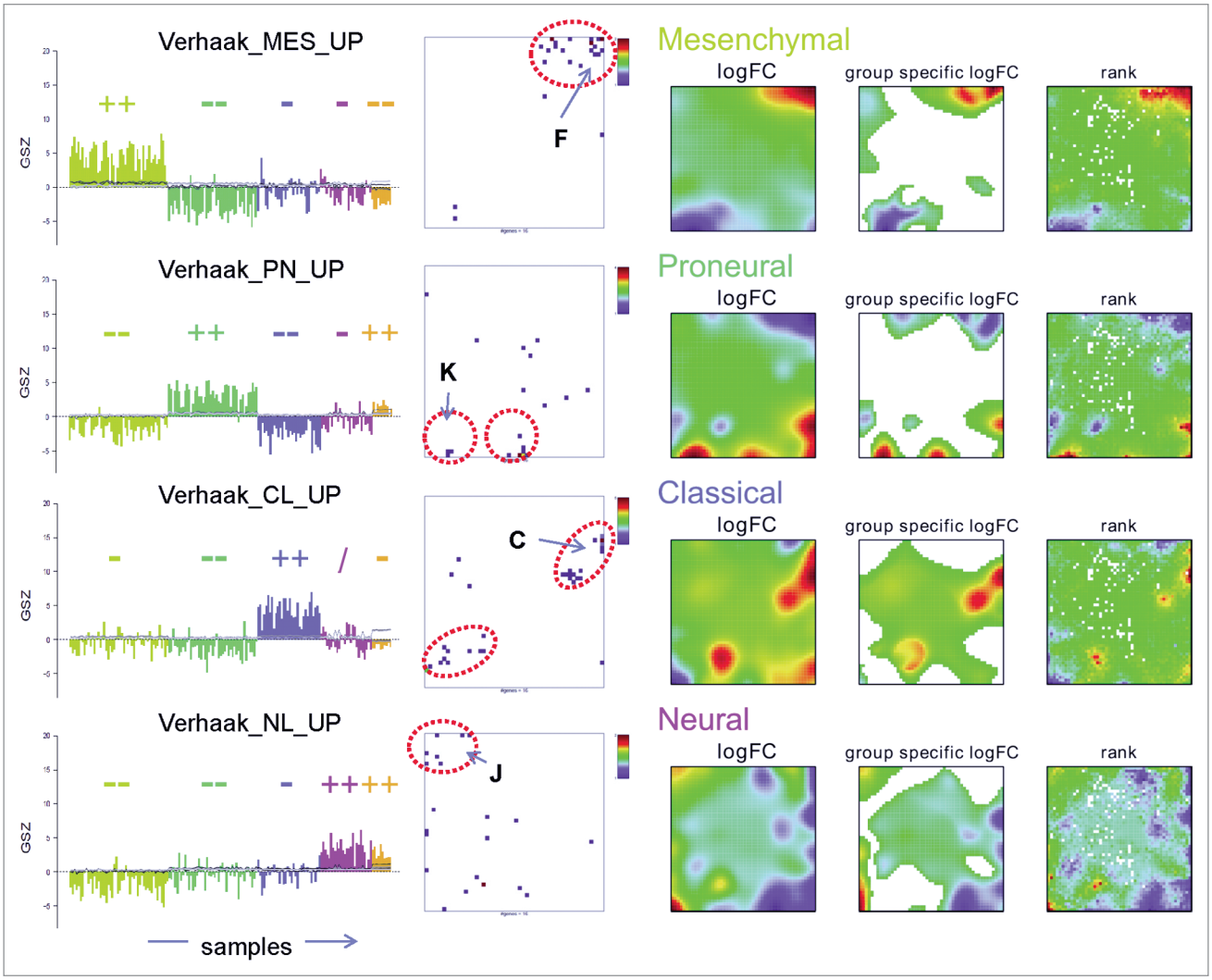

Figure 9. Subtype specific genes in GBM. Panels, left to the right: GSZ-score profiles for subtype specific gene sets reported earlier,[7] **Equation 8**; the corresponding gene set population maps; the subtype mean in logFC scale; difference portraits, **Equation 1**; and rank maps. In the GSZ-profiles, each bar represents one sample, color coded according to subtype as before (e.g., **Fig. 1**). The ± signs above the profiles indicate over- and underexpression. In the gene set population maps, the number of genes from a gene set in each meta-gene is color coded from white (no gene) to maroon (maximum number observed). The red dashed ellipses indicate gene sets accumulating in distinct regions of the map, which to a good approximation agree with the subtype specific spots in the average and difference portraits.

Mapping subtype-specific differential expression

To extract unique, subtype-specific spot patterns, we calculated Difference Maps; see Materials and Methods, **Equation 1**. These maps select meta-genes over- and underexpressed specifically in only one of the GBM subtypes (**Fig. 9**) or PCP stages (**Fig. 10**). A meta-gene is represented in red (blue) if its expression value in the chosen subtype is higher (lower) than in all the other subtypes. The difference maps reveal subtype-specific over- and underexpression spots which largely agree with the features seen in the mean portraits of the respective subtypes. Non-specific features, however, (such as the spot "N," found in several subtypes, **Fig. 6B and D**) disappear by applying **Equation 1**, as expected.

In addition, we calculated mean rank maps as described previously.[10] In short, a ranked list of differentially expressed genes is calculated for each sample using a regularized *t*-score as ranking criterion. The *t*-score takes into account the noise-level of expression values which is neglected in the logFC scale. The effective rank of each meta-gene is then calculated as logged mean rank of the associated genes (**Figs. 9 and 10**, rightmost columns).

Genes that are significantly overexpressed in a GBM subtype or a PCP stage relative to all the others have been determined using SAM significance analysis of microarrays[11] before.[7] We calculated the gene set enrichment scores GSZ, **Equation 8**, for the gene sets obtained there . The GSZ profiles in **Figures 9 and 10** show that indeed each of these gene sets is specifically overexpressed in the respective GBM subtype or PCP stage and underexpressed in the remaining subtypes or stages. Interestingly, the original NL-specific GBM signature also shows overexpression in the healthy brain tissue, which had not been considered in the original study.[7] For PCP samples, we compared stage specific significant

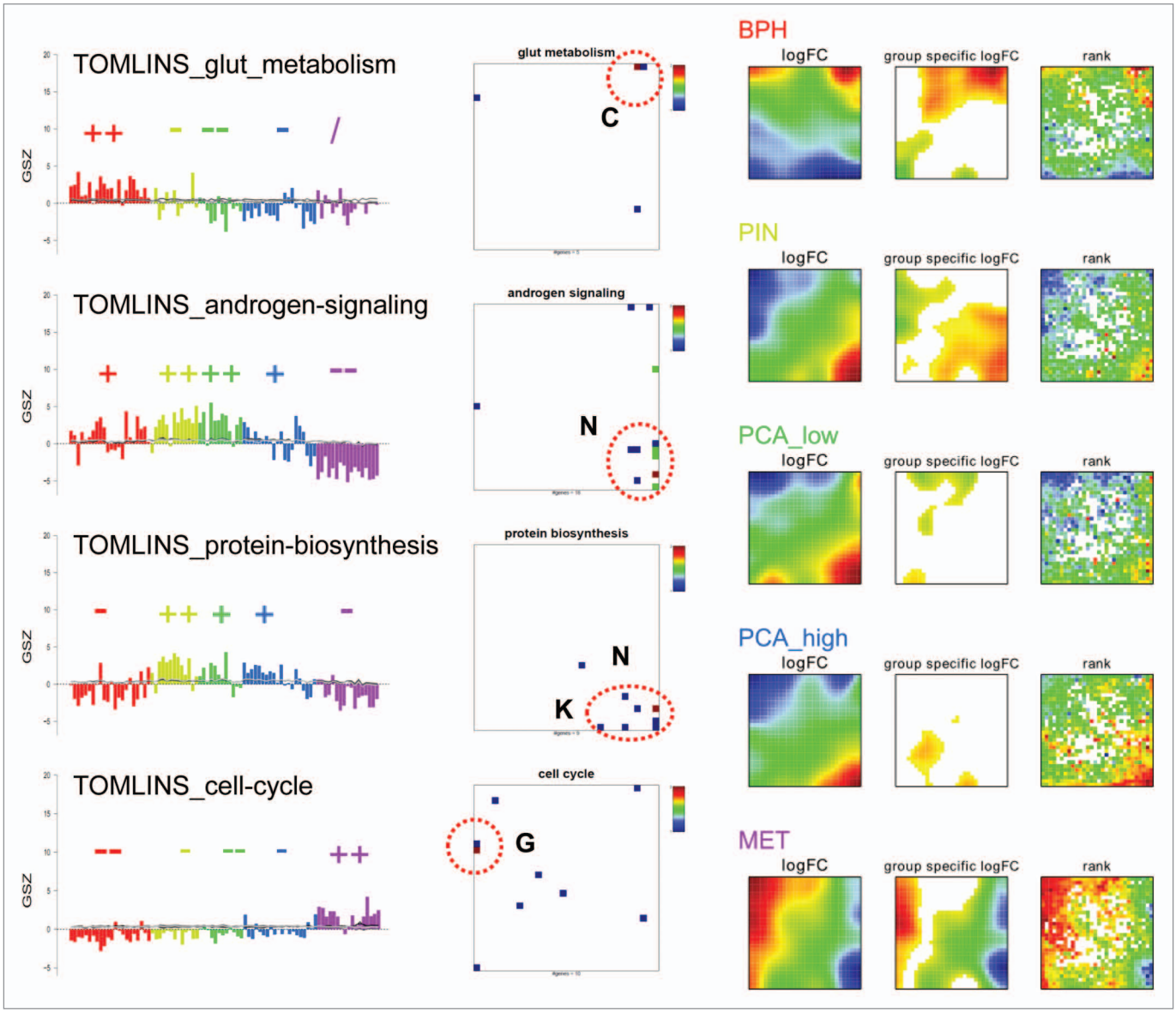

Figure 10. Stage specific genes of PCP, testing gene sets reported as overexpressed in different PCP stages.[8] The gene sets marked with red dashed ellipses in the population map agree with the stage specific spots in the average and difference portraits. See legend of **Figure 9** for details.

genes with gene sets taken from the original publications,[8] confirming enrichment of expected biological pathways (**Fig. 10**).

Mapping global differential expression

A recent study[12] identified 1,236 genes significantly differentially expressed between GBM and normal brain samples in the TCGA repository, not considering GBM subtypes. Among these 1,236 genes, we identified three sets: 425 strongly downregulated genes, 426 moderately downregulated, and 376 upregulated. The strongly or moderately downregulated genes largely accumulate in SOM spot "K," whereas the downregulated genes mostly fall into the "C" and "N" SOM spots (**Fig. 11**, scatter plots). This pattern agrees well with the strongest over- and underexpression spots observed comparing the mean portraits of the normal (NOR) and the GBM samples (**Fig. 11**, heatmaps). The expression levels of the mean NOR and GBM heatmaps are strongly anticorrelated, i.e, strongly positive spots in the NOR are strongly negative in the GBM samples and vice versa, reflecting the fact that the expression amplitudes of NOR samples largely exceed those of the GBM samples. The extracted gene sets in this comparison thus cover only a small part of the expression modules detected in our differentiated SOM analysis. Spots of weak differential expression but of potential high relevance for discriminating the different GBM subtypes remain undetected. GSZ-profiles (**Fig. 11**, left panels) reveal that the three gene sets only weakly differentiate between the MES, PN, and CL subtypes (yellow, green, blue). The profiles also show that the NL subtype (magenta) partly follows the expression patterns of normal brain (NOR, orange).

Results illustrate the benefits of our approach: It provides a detailed view on the compartmentalization of the expression

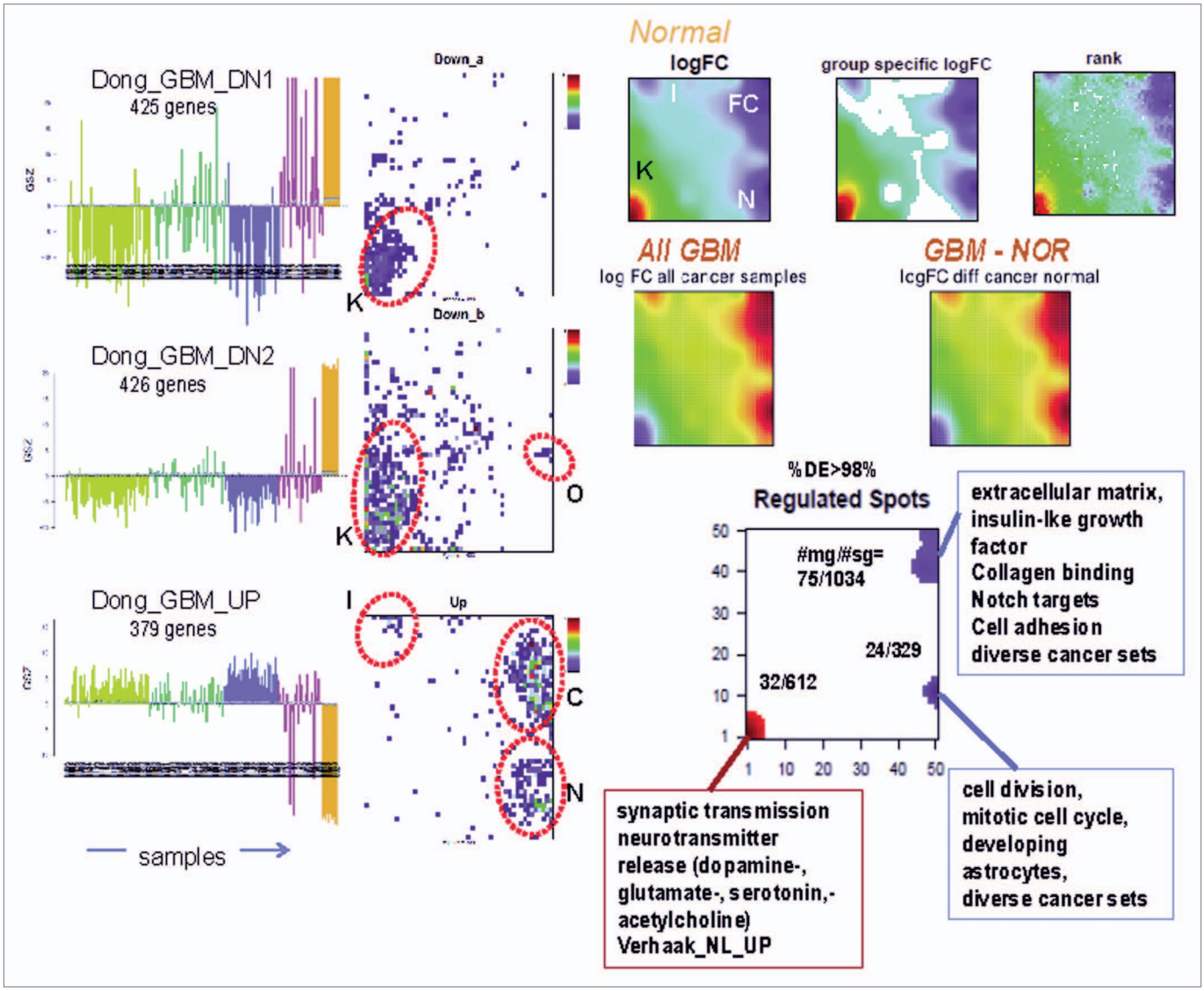

Figure 11. Differentially expressed genes from a study comparing GBM vs. normal samples.[12] Among the differentially expressed genes reported, we identified three gene sets of comparable sizes: strongly downregulated (DN1), moderately downregulated (DN2), and upregulated (UP). The left part of the figure shows the GSZ-profiles and population maps of the three gene sets (color code as before, e.g., **Fig. 1**). In the right part, the first row shows the mean logFC, group specific difference, and mean rank maps of normal brain samples from our full SOM analysis. The second row shows the mean logFC map of all GBM samples and the difference of the mean GBM and NOR maps. The map in the bottom right panel shows the selected spots and their functional contexts.

landscape, which allows separate analyses of individual modules in terms of biological context. These can identify functional details that are easily missed in a simple differential disease vs. normal approach.

Discovering the functional context: Gene set profiles and population maps

Each spot in the SOM portraits represents a cluster of co-regulated genes. We applied gene set over-representation analysis to each spot-cluster using a collection of about 6000 predefined gene sets for specific GO categories, pathways, diseases, human tissues, and cell types (see Material and Methods section). For each of the spots detected in GBM and PCP, gene sets with $P \leq 0.0001$ are listed in **Supplemental File 1**. Based on the functional context of the over-represented gene sets obtained, we assign a short label to each detected spot (see **Fig. 12A** for GBM, **Fig. 13A** for PCP, and **Supp. File 1**).

Spots in GBM are clearly related to biological processes associated with cancer physiology, such as inflammation (**Fig. 12C**, spot "F") and cell division (**Fig. 12D**, spot "N"), as expected. These GSZ-profiles reflect the fact that the respective biological processes are selectively activated/de-activated in a subtype-specific fashion, namely inflammatory response in the MES and cell division in the PN subtypes of GBM. The respective gene set population maps (**Fig. 12C and D**) reveal that the associated genes accumulate in the regions of spots overexpressed in the maps of the different subtypes (cf. **Fig. 9**).

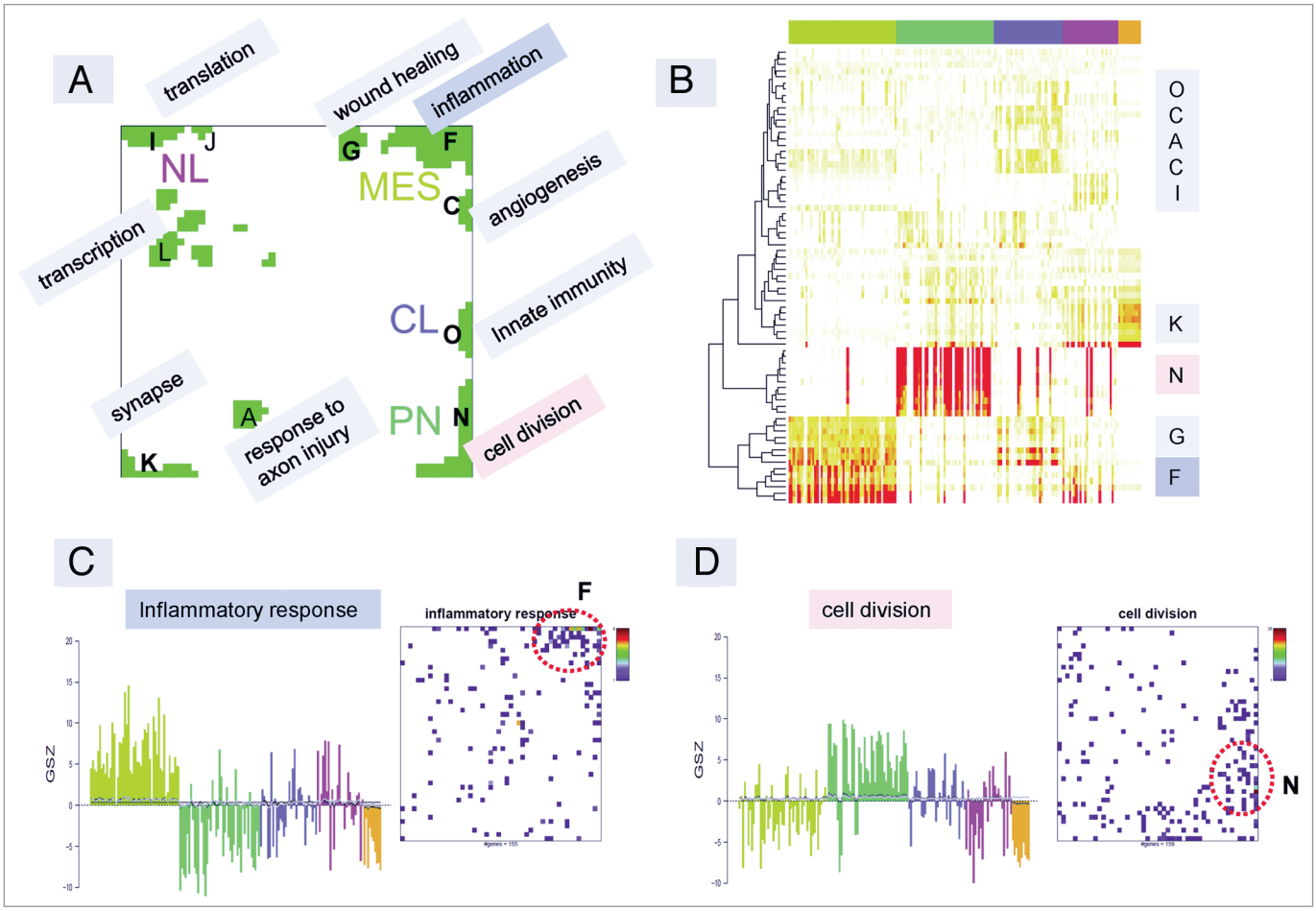

Figure 12. Gene set enrichment analysis of GBM. (**A**) The spot summary map shows the functional context of the most abundant spots (boxed labels) together with the associated subtypes (NL, MES, CL, PN). (**B**) The over-representation heat map of gene sets for the GO term "biological process" provides an overview. The different clusters correspond to the spots in the SOM portraits. The letters on the right refer to the spots identified in (**A**). The color bars on the top represent the GBM subtypes. (**C and D**) Overexpression profile and map of the "inflammatory response" and "cell division" gene sets, respectively. The red dotted ellipses in the map indicate the spots of strongest enrichment. The full list of enriched gene sets, over-representation heatmaps of different gene set categories, and a gallery of the overexpression profiles and maps are given in **Supplemental File 1**.

"Inflammatory response" and "cell division" are not among the leading gene sets of any of the spots in PCP (**Supp. File 1**). The respective GSZ-profiles, however, show that "inflammatory response" is selectively activated in the BPH and MET stages, whereas "cell division" genes are overexpressed in the MET stage only (**Fig. 13C and D**). The population maps of these gene sets indicate that the respective genes accumulate in the regions of more than one overexpression spot. For example, larger concentrations of genes related to cell division are found in spots, for which the leading biological processes/cellular components are "RNAPII activity" (spot "G") and "ribosome" (spot "N"), whereas genes related to inflammation accumulate in spots assigned to "mitochondrion" (spots "J" and "K") and "nucleosome" (spots "A" and "B") (**Fig. 13A**).

Categorizing the gene sets: GO terms, cancer, and cell type related genes

Neighboring spots of strongly correlated meta-gene expression profiles can be assigned to related biological processes: As shown in **Figure 12**, the "inflammation" spot "F" in GBM is close to spots assigned to "wound healing" and "angiogenesis"; the "cell division" spot "N" is close to spot "O" labeled "innate immunety," where "stress activated signalling" was the most strongly over-represented gene set. Note that, although related, these neighboring spots are usually characterized by subtle differences in their expression profiles and presumably also by fine differences in the functional context of the over-represented gene sets. In **Figure 14** we provide GSZ-profiles and population maps of a series of gene sets selected from the GO-terms "biological process" (BP), "cellular component" (CC) and "molecular function" (MF) which change in concert with "inflammation" and "cell division." The population maps clearly reveal these subtle differences: For example, both GSZ profiles of "immune response" and "wound healing" gene sets change together with inflammation and accumulate in adjacent but different regions of the maps of GBM (see **Supp. File 1**). The population maps of the "angiogenesis" gene set and, to a lesser degree, of the "wound healing" gene set, give rise to the overexpression of the respective GSZ-profiles

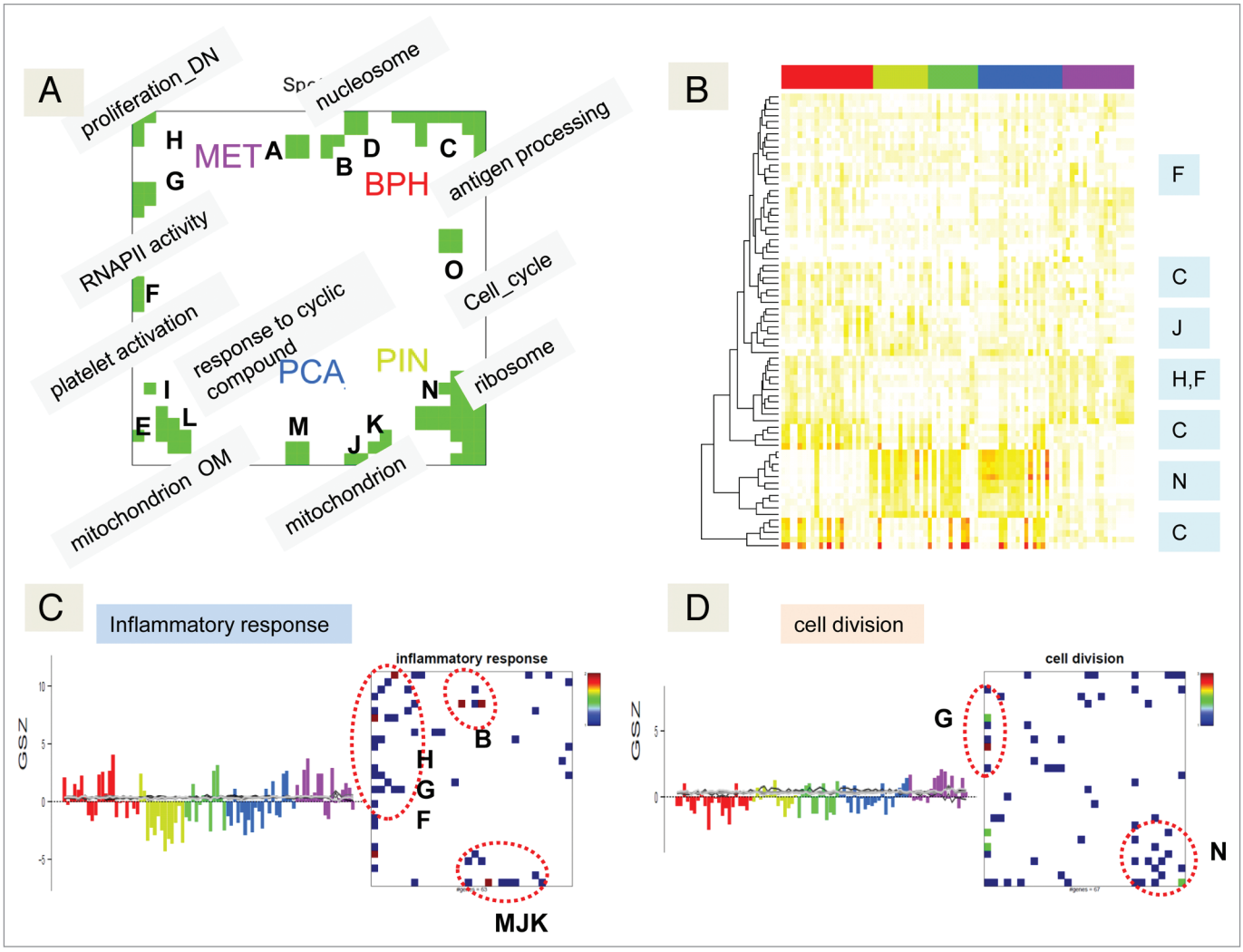

Figure 13. Gene set enrichment analysis of PCP. (**A**) The spot summary map shows the functional context of the most abundant spots (boxed labels) together with the associated stages (MET, BPH, PIN, PCA). (**B**) The over-representation heat map of gene sets for the GO term "biological process" provides an overview. The different clusters correspond to the spots in the SOM portraits. The letters on the right refer to the spots identified in (**A**). The color bars on the top represent the PCP stages. (**C and D**) Overexpression profile and map of the "inflammatory response" and "cell division" gene sets, respectively. The red dotted ellipses in the map indicate the spots of strongest enrichment. The full list of enriched gene sets, over-representation heatmaps of different gene set categories, and a gallery of the overexpression profiles and maps are given in **Supplemental File 1**.

in the CL subtype, while underexpression was observed for other "inflammation"-like gene sets.

The results so far show that GBM splits into subtypes differing by the antagonistic activation of biological processes related to "inflammation" and "immune response" vs. processes related to "cell division" and "transcriptional and translational machinery" (*viz*. MES *vs* PN). We have observed a similar separation of subtypes related to "inflammation" and "cell division" in B-cell lymphoma (BL, unpublished results). In order to evaluate the degrees of similarity between both GBM and BL cancer entities in this respect, we studied the enrichment of signature gene sets from BL in GBM (**Supp. File 1**): It turned out that the two signature gene sets up- and downregulated in the BL subtypes strongly accumulated in spots F and N (**Fig. 12**), which are overexpressed in the MES and the PN subtypes, respectively. This result suggests a more generic nature of the underlying processes related to "inflammation" and "cell division" in cancers.

We extended this comparative view by extracting low abundance "rare" transcript sets deregulated in hepatocellular carcinoma, breast carcinoma, and nasopharyngeal carcinoma.[13] The resulting three gene sets were of poor prognosic value for "metastatic cancer" related to the c-Myc oncogene[14] and as a universal transcriptional profile essential in neoplastic transformation and commonly activated in many cancers.[15] All three gene sets show GSZ profiles and population maps similar to the "cell division"-like sets in GBM, i.e., mostly overexpressed in the PN subtype and underexpressed in the MES subtype (**Supp. File 1**). Two gene sets of "myc-poor-prognosis" (8 genes) and of "undifferentiated_cancer" (16 genes) essentially occupy the same regions of the map as spot "O" in **Figure 12**, despite sharing just a single gene. Finally, the "common_cancer_gene" set is found in spot

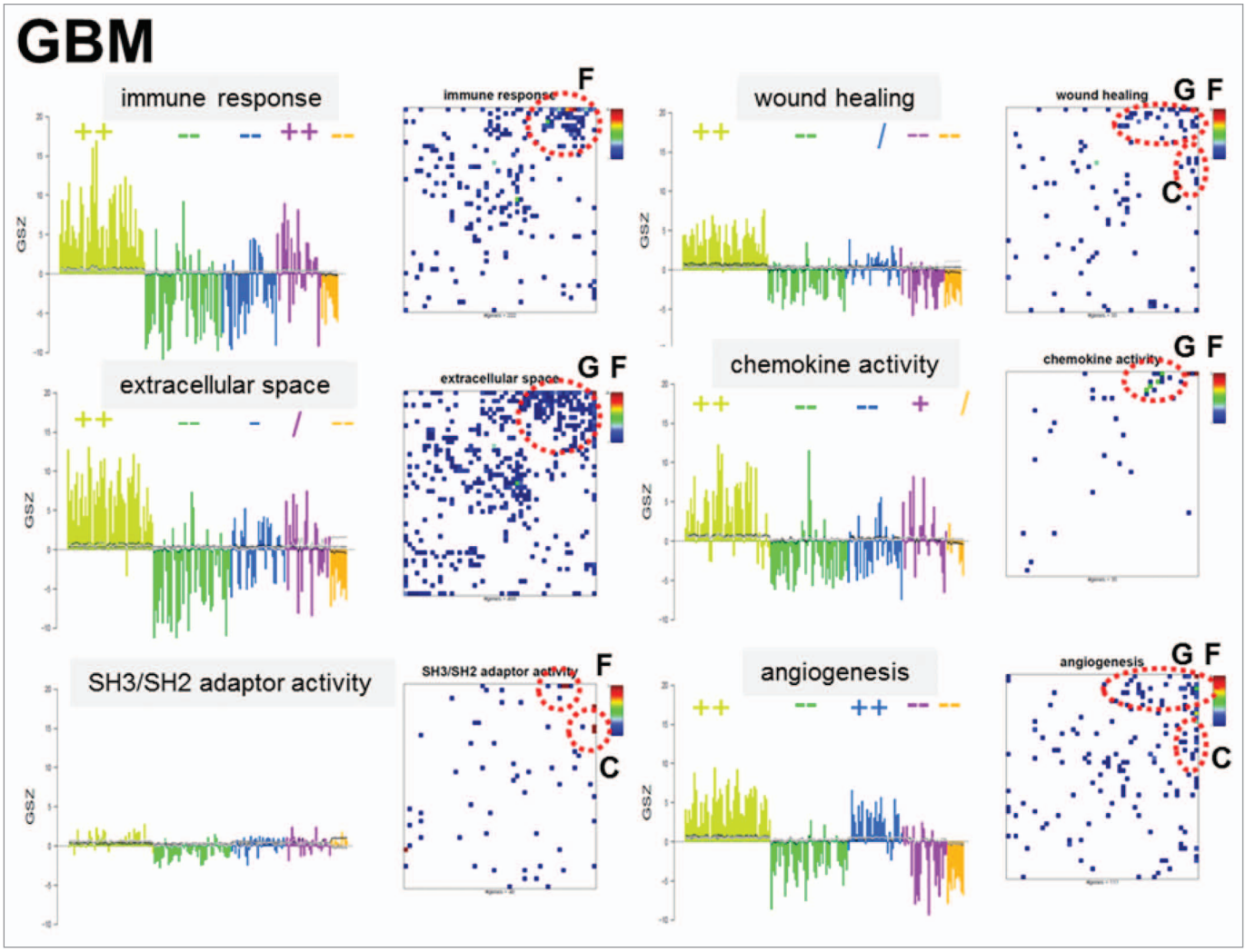

Figure 14. Selected profiles and population maps of "inflammatory response" gene sets for GBM. Regions of over-representation in the population maps are indicated by red-dotted ellipses. The letters refer to the corresponding spots in the SOM portraits.

"N." This suggests that the low abundance "rare" transcripts might provide new signatures in addition to those from high abundance genes.

Gene sets related to "innate immunity" are found overrepresented in spot "O" of the GBM map which is overexpressed in the CL- and PN- and partly also in MES-subtypes. The intermediate CL- and NL-subtypes of GBM are characterized by spots "A" and "F," "C" (CL-subtype) and "I," "J," and partly "K" (NL-subtype, see **Fig. 9**). The latter spot "K" is characteristic also for healthy brain tissue. It is therefore not surprisingly that it contains overrepresented populations of gene sets related to "nervous processes" such as "synaptic transmission" and "neurotransmitter secretion." Population maps and profiles are shown in **Supplemental File 1**. The NL-subtype however differs from the healthy brain tissue mainly by the appearance of spots "I" and "J" (see **Fig. 9**) which contain overrepresented gene sets related to "translation," such as "ribosome" and "mitochondrion." The CL-subtype-specific spot "C" related to "angiogenesis," and thus reflects a common cancer process.

We further analyzed the relationship between gene sets and the cell type or tissue specificity in order to understand the biological meaning of the GBM subtypes. We collected the gene set enrichment level from the brain transcriptome database[16] as proposed before.[7] Mature cell types such as neurons, oligodendrocytes, astrocytes, and cultured astroglial cells may be of interest for their primary associations with tumor subtypes and as inherent signatures retained from progenitor cells. In agreement with earlier studies,[7] we found subtype specific enrichment of signatures: "oligodendrocytic" (in PN and NL), "astrocytic" (in CL and NL), "neuronal" (in NL), and "cultured astroglia" (in MES and partly CL); see **Supplemental File 1**. We also tested signatures for "developing astrocytes" (enriched in PN and partly NL) and "nervous tissue" (enriched in NL and NOR).

Our SOM mapping and profiling of the different signature sets, however, in addition provides a finer assignment to the different GBM subtypes: The "oligodendrocytic," "neuronal" and "nervous system" genes accumulate preferentially in spot "K," which is overexpressed in normal brain tissue. Its key

property in GBM is the antagonistic upregulation in NL and downregulation in CL subtypes. In contrast, the "astrocytic" signature genes accumulate in spot "A," with the corresponding upregulation in NL and CL subtypes and downregulation in MES and PN subtypes. Hence, the co-located spots "A" and "K" can be associated with different regulation patterns especially in NL and CL subtypes, while these can be associated with different cell types. Interestingly, the "astrocyte" signature strongly resembles that of the "aging brain_DN" set of genes which reduce their activity in the aging cortex[17] and the GO term "negative regulation of cell death" (see **Supp. File 1**). The signatures of "nervous tissue" and "developing astrocytes" are enriched in spots "K" and "O," respectively. They similarly respond in an antagonistic up-*vs*-down fashion in PN and MES subtypes. Note that spot "O" was associated also with biological processes related to "cell division" such as "mitosis," "DNA-repair," and undifferentiated cancer. Finally, while the signature genes of "cultured astroglia" also accumulate in spot "O," they are found primarily in spots "G" and "F," showing upregulation in the MES subtype and antagonistic downregulation in normal brain tissue.

Gene set overview maps

For a more general overview of over-represented gene sets, we generated gene set enrichment heatmaps to survey a larger collection of biological functions potentially contributing to the expression landscape. These heatmaps collect gene sets significantly over-represented in the SOM portrait spots in a sample-specific fashion, and cluster them according to their degree of over-representation. **Figure 12B** and **Figure 13B** shows the heatmap for gene sets associated with the GO term "biological process" and enriched in spots of the GBM and PCP SOM portraits, respectively. The one-way clustering separates the gene sets in agreement with their spot associations: For example in GBM, spot "F" mainly collects gene sets overexpressed in the MES and also the CL and NL subtypes, whereas the adjacent spot "G" contains gene sets overexpressed in the MES subtype. The heatmap also shows that gene sets from the spot "K" tend to be overexpressed in normal brain tissue as well as in the NL and PN subtypes. It further associates gene sets overexpressed in the PN subtype with spot "N." Complete heatmaps with detailed named gene set categories are given in **Supplemental File 6** (GBM) and **Supplemental File 7** (PCP).

Detailed inspection of the GBM heatmap reveals that the "cell division" spot "N" contains additional related gene sets such as "mitosis," "DNA replication," "spindle organization," and "cell cycle checkpoint." These sets refer to different levels in the GO hierarchy, partly giving rise to overlapping groups of genes which, in consequence, trivially link similar expression patterns.[18] Here we neglect any interdependency due to such an overlap in gene sets, which may also arise across different GO categories and the curated gene sets from the literature. This redundancy might, however, highlight alternateve aspects of annotated gene function: For example, the PN specific spot "N" resembles the expression characteristics of the "cerebellum" tissue gene set, spot "G" is associated with "epithelium" and "primary lymphoic organs," and "F" with "immune systems tissues" and "mucosa."

In addition to one-way clustering heatmaps, we also performed two-way clustering of gene sets and samples to detect inconsistencies in the class labeling of the samples. The resulting heatmaps for the literature gene sets (GSEA2) reveal that the cancer related gene sets essentially form two clusters with strong enrichment, respectively, in spots highly overexpressed in the MES subtype (spot "F" and "G") and the PN subtype (spot "N"). This seems to reflect common gene activation patterns present in different tumors associated with either "inflammation" (for MES) or "cell division" (for PN). See **Supplemental File 6** for supporting results and complementary analyses, including by cell-type.

Outliers, misclassified samples, and mixed subtypes

Large tumor sample collections are prone to different effects not, or not directly related to the disease such as varying tissue compositions, RNA quality, lab protocols, and high biological patient-to-patient variance. The SOM portraits introduced here offer a simple and direct approach to checking the whole-transcriptome expression landscapes of the individual samples by visual inspection for deviations from the majority of samples assigned to the same class.

In **Figure 15** we show the CN similarity plot for GBM together with selected individual portraits of samples which are located either outside the main clusters or which seem to be misclustered. For example, samples 326 (MES subtype) and 156 (PN subtype) are found near the PN and MES-clusters, respectively. Comparison of the portrait of sample 156 with the mean portraits of MES and PN subtypes shows that its expression landscape represents a combination of both subtype signatures, where the MES-signature more heavily contributes to the mixture than the PN-signature, in contradiction to the original class assignment.[7] Another heterogeneous group of samples (290, 152, and 358) form a set of outliers near the CL cluster. Inspection of the respective portraits reveals that a few overexpression spots ("L," "B," and "D") are responsible, as they are not observed in the majority of the remaining CL samples. Other outlier groups are samples 326, 84, and 87, showing strong expression of spot "n1."

Outliers are mostly with different subtypes, and are relatively rare (see the abundance bar plot for spots "L" and "D" in **Fig. 6**). This suggests that they are presumably caused by contamination with non-tumor cells. Treatment effects may also generate outliers: For example, gene set analysis shows that spot "B" contains an enriched number of genes related to "xenobiotics" and "drug metabolism" (**Supp. File 1**).

These examples demonstrate that our portraying approach cannot only detect potential outliers and misclassified samples but also helps researchers generate hypotheses about the origin of these effects and follow up, for example, by applying spot-related functional analysis.

Weighted topological overlap (wTO) correlation network analysis: Modular gene regulation the subtypes

Weighted topological overlap (wTO) network analysis allows a visualization of correlations between spots that include the effects of indirect interactions mediated by other meta-genes. The resulting wTO network of GBM (**Fig. 16A**) shows antagonistic relationships between spot "K," which is upregulated in normal

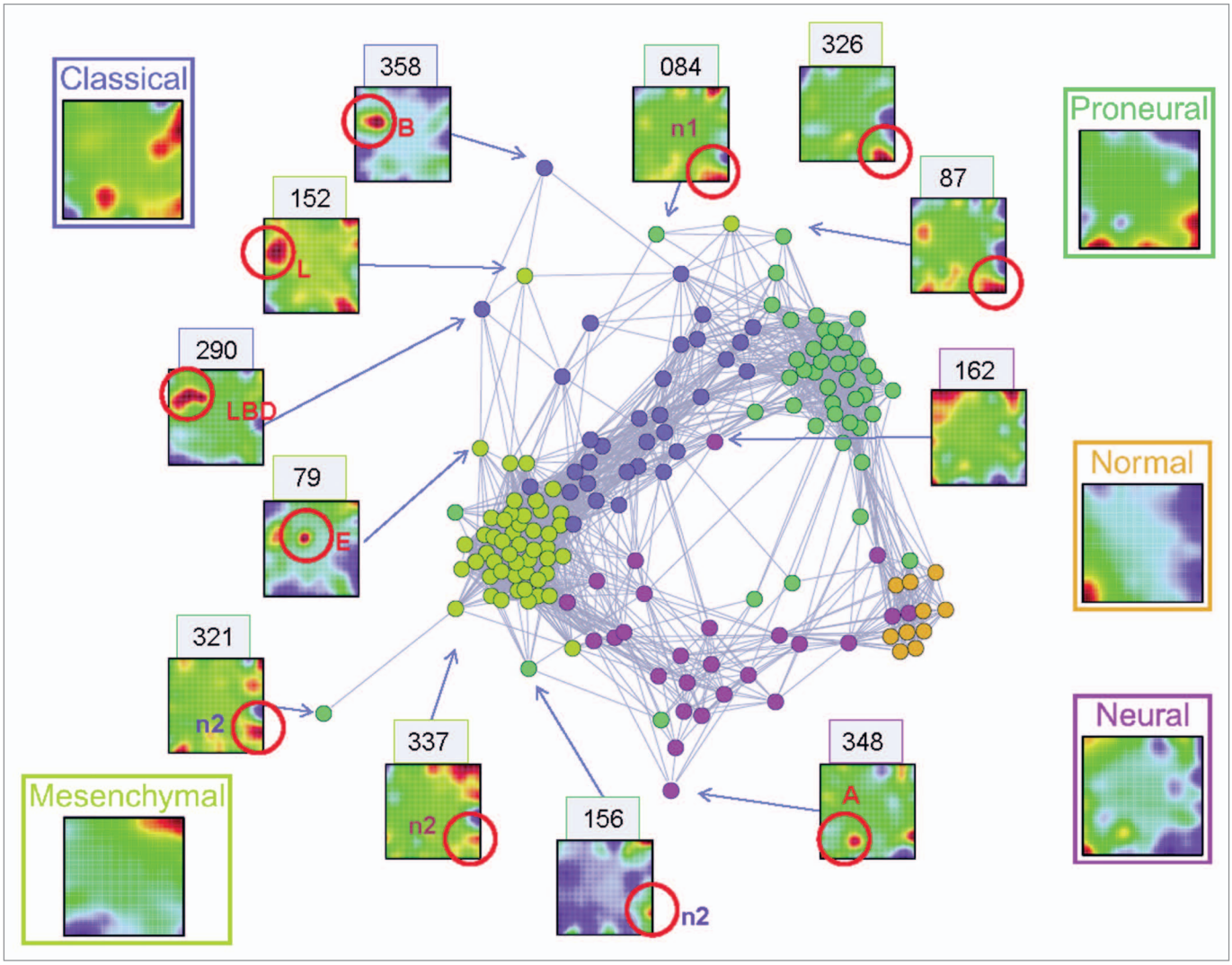

Figure 15. Outliers and misclassified samples in GBM in the CN similarity plot. The subtype-averaged mean portraits are shown for comparison at the left and right to the network. The red circles and letters identify the spots responseble for the deviations.

brain tissue, and a first layer of spots upregulated in the subtypes MES ("G," "F"), CL ("O," "C"), and NL ("J" and indirectly "I"), while these spots are partly co-regulated with one another. A second layer of spots appearing in the subtypes PN ("N"), NL, and CL ("A") changes in an anti-correlated fashion with respect to this first layer of spots. Another, well separated cluster seems to be formed by spots that presumably reflect contamination or other effects not related to cancer. The most abundant spot in this cluster ("L") is positively correlated with the "nervous processes" in spot "K." Hence, GBM features a group of expression modules anti-correlated with genes specifically upregulated in normal brain tissue (*cf.* "regulating synaptic activity" gene signature). These modules, in turn, are associated with functions common in cancer, such as "inflammation," "angiogenesis," "cell division," "translation," and "mitochondrial activity." They vary specifically in the different subtypes, either in a correlated or anti-correlated fashion, with considerable mixing especially for the intermediate subtypes (NL and CL), which form rather a continuum of expression states than a distinct entity. These two intermediate subtypes on one hand and the two "separated" subtypes (MES and PN) on the other hand are governed by the antagonistic expression of independent gene activities, as revealed by sample similarity analysis using ICA and correlation nets (CN). Hence, "inflammation" and "cell cycle" activity for MES- and PN subtypes on one hand and "translation" and "angiogenesis"/"innate immunity" for NL- and CL subtypes on the other hand change in an antagonistic fashion but are characterized by independent gene sets.

Interestingly, spot "A," which is related to "axon injury," is upregulated in the intermediate subtypes NL and CL, and is positively correlated with normal brain tissue activity ("K" in NOR). It is strongly downregulated in PN and MES subtypes, accompanied by changes of the gene sets "synaptic transmission" (down), "mitochondrion" (down), "inflammation" (up in MES), and "DNA-repair" (up in PN). Cell-type specific functional analysis suggests that spot "A" is associated with astrocyte function and

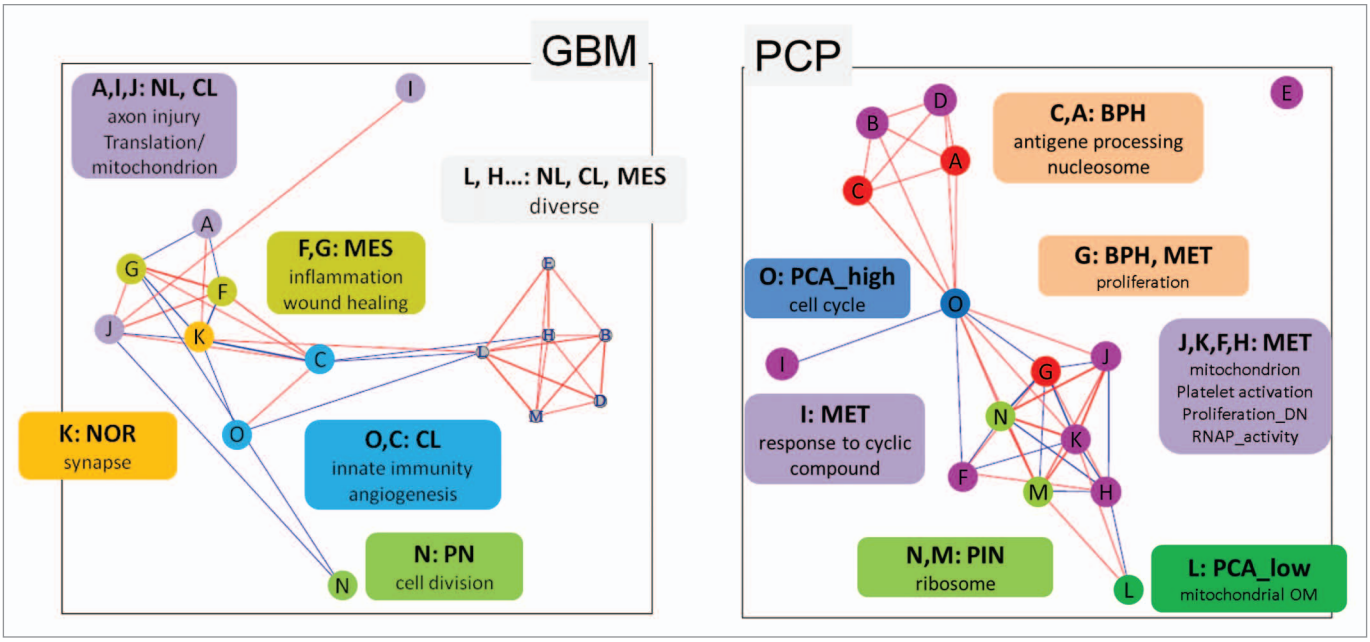

Figure 16. Weighted topological overlap (wTO) networks of the expression spot-modules of GBM and PCP. The nodes represent the spots from **Figures 6 and 7**. Red and blue lines indicate positive and negative correlations, respectively. The threshold for plotting overlaps is $\omega > 0.35$. The text boxes list the dominant cancer subtype or stage, as well as the leading functional context obtained from gene set enrichment analysis.

enriched in genes which progressively lose activity in the aging brain[17] (see **Supp. File 1**).

The basal structure of the wTO network of PCP (**Fig. 16**, right panel) shows a rather different picture: A series of spots upregulated in different progression stages are co-regulated. They form a "backbone" with anti-correlated links to a second layer of modules. The backbone is formed by modules associated with "antigene processing," "cell cycle," "mitochondrion," "ribosome," and "response to cyclic compound." Note that most of the spots are rather heterogeneous and usually show increased activity in different PCP stages (**Fig. 7**). These mutually correlated backbone spots also reflect time delays between the different biological processes as illustrated in the GSZ-profiles of **Figure 10**. Spots "M" and "N" are mostly associated with the PIN stage enriched sets, related to "protein biosynthesis" and "androgen signaling" (**Supp. File 1**). These are the key processes defining cancer progression at early stages and particularly the transition from localized to hormone-refractory metastatic prostate cancer[8] (**Fig. 10**).

Interestingly, the "cell cycle" spot "O" forms a bottleneck between two separate clusters of spots. The cluster in the upper part of the graph is mostly associated with the early BPH stage, showing activated "antigene processing" and "response to progesterone stimulus" gene sets. The other cluster is associated with later stages which, in general, are characterized by activated mitochondrial, transcriptional, and translational machineries, as well as proliferation. Some of these modules are anti-correlated: for example, the backbone modules "N" (ribosome) and "J" (mitochondrion) anti-correlate with "G" (proliferation), reflecting mutually antagonistic regulatory modes of the progressing cancer. Spot I ("response to cyclic compound") appears relatively isolated from the rest of the network suggesting that it might be caused by contaminations or processes not related to cancer .

The wTO-network representation thus illustrates the mutual relationships between all the identified spots. The separation of positive and negative correlations extracts concerted and antagonist pairings of expression modules, which form a basal "skeleton" picture of the gene expression network of the cancer. Each node itself subsumes a sub-network of typically hundreds to thousands of genes with a defined functional context. The particular expression state of the nodes of the basal skeleton is characteristic for each subtype or stage, with possible relevance for classification and functional interpretation.

Outlook

SOM machine learning enables a kaleidoscopic and intuitive view of high-dimensional data without a loss of the primary information. It provides a general framework for analytic tasks such as feature selection, integrating concepts of molecular function and systems tracking with a resolution of individual samples. The method extracts abstract features such as meta-genes and spots/modules expressing basal modes of systems behavior important for higher-level, holistic analysis.

We applied SOM to expression profiles of glioblastoma multiforme (GBM) and prostate cancer (PCP) to characterize the specifics of the genome wide expression landscapes in different subtypes or stages of cancer. Our method simultaneously detects features which are differentially expressed and correlated in their profiles in the set of samples studied. Functionally related genes often merge into larger aggregates which can then be interpreted as functional modules. They characterize disease-specific

changes in the resulting interaction network. Characteristic differences between subtypes and disease stages can be clearly identified and further analyzed using meta-gene profiles representing the intrinsic correlation groups. Our case study has demonstrated that analyzing gene expression landscapes in the context of a compendium of molecular concepts is useful in understanding cancer biology.

Ongoing tasks that build on this framework also address issues such as "interOMICs" integration, as well as the extension of the method to next generation sequencing and other data types.

Materials and Methods

Expression data

GlioblastomaMultiforme(GBM)

Microarray data are available on "The Cancer Genome Atlas" (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/). We downloaded level 1 (raw intensity) data of 153 GBM and 10 normal brain tissue specimen hybridized on Affymetrix HT-HG-U133A arrays. We used the classification of tumor subtypes given in reference 7. The samples were assigned to Mesenchymal (MES, 50 samples), Proneural (PN, 45), Neural (NL, 26), Classical (CL, 32) GBM-subtypes and to normal healthy brain (11) for comparison. The latter specimens were taken from adjacent brain tissue of GBM patients. In addition we also downloaded level 2 (RMA preprocessed) data of the same patients for comparison with our hook-preprocessed data.

Prostate cancer progression (PCP)

Microarray data are available under GEO accession number GSE6099 (104 non-commercial spotted Human 20K Hs6 arrays). The original study[8] addresses the molecular mechanisms associated with gene expression changes in the course of prostate cancer progression using laser capture microdissection by means of 84 samples from 44 individuals. The samples used are assigned to five stages of cancer progression ranging from benign prostatic hyperplasia (BPH, 22 samples) and prostatic interepithelial neoplasia (PIN, 13) to low-grade (PCA_low, Gleason score 3, 12 samples), high-grade (PCA_high, Gleason score 4–5, 20 samples), and metastatic (MET, 17) prostate cancer.

Calibration and normalization

Raw probe intensity values of Affymetrix arrays for GBM were calibrated and summarized into one expression value per probe set using the hook method,[19,20] quantile-normalized[21] and corrected for background-noise as described in reference 6 and **Supplemental File 1**. Expression values of the custom PCP sample arrays were quantile-normalized. Then, the expression value of each gene was transformed into log10-scale and centered with respect to the mean value averaged over all samples considered in the respective series. A relative log-expression value of zero consequently means that the gene is expressed according to its mean expression value while positive and negative values refer to over- and underexpression, respectively. We use the term "expression" for these relative expression values if not stated otherwise.

SOM training

The preprocessed expression values of each cancer data set are used to train an SOM. It translates the high-dimensional expression data given as N × M matrix (N: number of genes, M: number of samples) into a K × M matrix (K: number of meta-genes) of reduced dimensionality K < < N (N ~ 10^4 and K ~ 10^3). The meta-gene profiles are obtained via iterative machine learning. First, the meta-genes are arranged in a two-dimensional quadratic grid of K tiles where each tile refers to one meta-gene. After linear initialization[22] of the meta-gene profiles, a single gene is picked from the gene list and its expression profile is compared with that of the meta-genes using the Euclidean distance as similarity measure. The meta-gene profile of the closest similarity and its nearest neighbors are then modified, so that they more closely resemble the expression profile of the selected gene. This process is applied to all genes and repeated 250,000 times. The radius of considered neighbors is decreased with the progressive iteration which modifies fewer meta-gene vectors by smaller amounts, so that the meta-gene vectors asymptotically stabilize. The resulting map becomes organized because the similarity of neighboring meta-genes decreases with increasing distance in the map. The final SOM consists of regions of similar meta-gene profiles. In turn, each meta-gene serves as a representative prototype of a "microcluster" of genes with similar expression profiles. The differential expression of the prototypic meta-gene is defined as $\Delta e_{km} = e_{km} - e_{k\bullet}$, where e_{km} is the logged expression of meta-gene k in sample m and $e_{k\bullet}$ is the respective profile mean.

SOM staining

Each expression state is visualized by color coding the two-dimensional mosaic of meta-genes according to their expression values in a sample. We first normalize the meta-gene expression data in each state to the range, $-1 \le \Delta e_{km}^{norm} \le +1$, and then color code Δe_{km}^{norm} according to two alternative scales: i) The "logFC" scale linearly transforms the normalized logged fold change, $\log FC = \Delta e_{km}^{norm}$, into green to maroon for $\Delta e_{km}^{norm} \ge 0$ and green to dark blue for $\Delta e_{km}^{norm} \le 0$; ii) The alternative "log-logFC"scale uses the double logarithmic scale log-logFC = $sign\left(\Delta e_{km}^{norm}\right)\log|\Delta e_{km}^{norm}|$ with the same color code as logFC. Note that log-logFC is steeper near $\Delta e_{km} \approx 0$ which strongly condenses green regions and expands red ($\Delta e_{km} > 0$) and blue ($\Delta e_{km} < 0$) regions compared with logFC.

Average subtypespecific portraits are calculated as the mean value of each meta-gene expression over all phenotype portraits of one subtype,

$$\Delta e_{kc} \equiv \left\langle \Delta e_{km} \right\rangle_{m \in class}$$

(c is the class index of each subtype) followed by normalization and coloring in logFC and log-logFC scales.To extract subtype-specific differential expression landscapes, we calculated difference maps, representing each meta-gene k in the mean SOM portrait of each subtype c according to:

$$diff_{kc} = \Delta e_{kc} - sign\left(\Delta e_{kc}\right) \bullet \min\left(\max\left(|\Delta e_{kc'}|\right)_{c' \ne c}, |\Delta e_{kc}|\right). \quad (1)$$

Equation 1 selects specifically over- and underexpressed meta-genes in a subtype. Particularly, $diff_{kc} > 0$ (or $diff_{kc} < 0$) means the expression of subtype c in meta-gene k exceeds (or falls below) the respective meta-gene expression in all other subtypes considered.

$diff_{kc} = 0$ is obtained if the relative expression of the meta-gene selected is unspecific for subtype *c*.

Supporting maps and meta-gene variability

Additional information such as the population of the meta-genes, the variance of meta-gene expression profiles, and the mean Euclidean distance with their nearest neighbors can be visualized using the same mosaic structure as in the expression portraits. The additional information is then color coded using proper scales. For example the variance map visualizes the variance of the meta-genes in each of the tiles,

$$\mathrm{var}_k = \frac{1}{M-1}\sum_m (\Delta e_{km} - \Delta e_{k\bullet})^2 = \frac{1}{M-1}\sum_m \Delta e_{km}{}^2, \quad (2)$$

with $\Delta e_{k\bullet} = 0$. We also calculate the orthogonal variability of the meta-gene expression landscape of each SOM image,

$$\mathrm{var}_m = \frac{1}{K-1}\sum_k (\Delta e_{km} - \Delta e_{\bullet m})^2, \quad (3)$$

where $\Delta e_{\bullet m}$ is the mean differential expression averaged over all meta-genes of sample *m*.

Detecting expression modules: Spot selection

The SOM algorithm arranges similar meta-gene profiles in neighboring tiles of the map, whereas more different profiles are located more distantly. As a result, neighboring meta-genes tend to be colored similarly. Therefore, the obtained mosaic portraits show typically a smooth texture with red/blue spot-like regions referring to clusters of over/underexpressed meta-genes. These blurry images represent the expression landscape of a particular sample. Meta-genes from the same spot mean that they co-expressed in the experimental series. Different, well-separated overexpression spots in the same image refer to meta-genes over-expressed in a particular sample but differently expressed in other samples because of their different profiles. Each spot can consequently be interpreted as an expression module of a group of meta-genes (and of associated genes) with concerted expression profiles.

We define over/under expression spots by applying a simple 98/2 percentile) criterion which selects the respective fraction of meta-genes with the top/bottom expression in each sample. Hence, the obtained over and underexpression spots are individual properties depending on the particular meta-gene expression in each sample. They can change their size from phenotype to phenotype and they can even disappear or transform from an over into an underexpression spot or vice versa.

Weighted topological overlap network of the spot modules

Networks provide a straightforward representation of interactions between expression modules. We applied the weighted topological overlap network (wTO) approach to the meta-genes which considers not only direct interactions between all pairwise combinations of meta-genes but also "mediated" ones acting via all possible third meta-genes in the map.[23] This "tree body" approach defines the topological overlap. The overlap ensures that strongly overlapping interactions (i.e., if both meta-genes strongly interact with the third one) contribute more than weak ones (e.g., if at least one of the meta-genes weakly interacts with the third one).

First, one determines the adjacency matrix between all meta-genes *i,j* with their mutual Pearson correlation coefficient, $a_{i,j} \in [-1,+1]$, where self correlations were neglected, $a_{i,i} = 0$. Then, the weighted topological overlap (wTO) matrix was calculated according to (see ref. 24 and refs. cited therein),

$$\omega_{i,j} = \frac{sign(a_{ij})\sum_{u\neq i,j} |a_{iu}a_{uj}| + a_{ij}}{\min(k_i,k_j)+1-|a_{ij}|}, \quad (4)$$

where

$$k_i = \sum_{u\neq i} |a_{iu}|$$

and

$$k_j = \sum_{u\neq j} |a_{ju}|$$

define the connectivity of the considered meta-genes *i* and *j*.

Equation 4 considers both positive and negative correlations. The topological overlap takes into account direct and indirect adjacencies between pairs of meta-genes "mediated" via all third party meta-genes *u*. We finally reduced the meta-gene pair topological overlap matrix to a spot-spot matrix after taking averages over all meta-gene pairs included in the spots s_1 and s_2),

$$\omega_{s1,s2} = \frac{1}{K_{s1}K_{s2}}\sum_{i\in s1, j\in s2}\omega_{ij} \quad (5)$$

where K_{s1} and K_{s2} are the numbers of meta-genes included in the spots s_1 and s_2 respectively.

The spot wTO-matrix is visualized using the R package "igraph" by applying a threshold $|\omega| > 0.35$.

Global spot characteristics

The spot characteristic analyses aim at characterizing global properties of the expression landscapes as seen by the over and underexpression spots. We calculated the mean spot number detected per subtype, the spot number distributions, the spot shape, their fractional abundance and a spot tree which visualizes the concerted expression changes between the spots.

The distribution of spot numbers was simply obtained as the fraction of sample portraits observed with one, two, etc. spots per subtype. The spot shape parameter, proportional to the classical definition of circularity,

$$shape_m = A_m/L_m{}^2, \quad (6)$$

characterizes the fuzziness of the observed over/under expression spots in each sample portrait. A_m denotes the number of tiles included in all spots observed in the image and L_m is the number of tiles along their inner borderlines with at minimum one adjacent tile outside and one tile inside the spots. A_m and L_m thus

estimate the area occupied by the spots and their limiting contour length respectively. The shape parameter hence relates the actual area of the spots to an theoretical area defined by the square of their contour length. For a single spot the shape parameter value decreases if its shape progressively deviates from a circular one. For n non-overlapping spots of identical area (a_m) and shape (l_m is their contour length), the total area and the total contour length scale with n and n^2, respectively, namely, $A_m = n \cdot a_m$ and $L_m = n^2 \cdot l_m$. The obtained shape parameter is inversely proportional to the number of spots. We calculated the shape parameter for the images in logFC and log-logFC scales independently to characterize the spot landscapes at high and intermediate expression levels.

The abundance of each spot is calculated as the relative frequency of appearance of each spot in the samples of each cancer subtype,

$$x_{sc} = \frac{m_{sc}}{M_c} \tag{7}$$

where M_c is the total number of samples in the subtype c and m_{sc} is the number of portraits showing a particular spot s among those samples. The spot abundances are represented as stacked bar plot for each spot. The integral abundance,

$$X_s = \sum_c x_{sc},$$

can be interpreted as the average number of classes showing a particular spot. Its maximum value equals the number of classes considered, $X^{max} = 5$ for both GBM and PCP.

Gene set overexpression profiles and population maps

Co-expressed genes of each expression module can be assumed to be functionally related according to the "guilt-by-association" principle.[25] Gene set analysis aims at identifying the functional context of these expression modules. This method estimates the enrichment of groups of predefined gene sets which are obtained independently, for example from SOM spot analysis (see ref. 26 for a critical review and references cited therein). Enriched gene sets suggest association between their functional context and the system studied. A large and diverse collection of such sets can be derived from GO[27] using the biomaRt interface.[28] Particularly, we included 5730 gene sets for GBM and 4349 for PCP in our analysis taken from the following categories (the different numbers are caused by the different chip types with different genes): i) GO gene sets (2192 GBM, 1054 PCP), composed of "biological process" (1394 GBM, 643 PCP), "molecular function" (488 GBM, 230 PCP) and "cellular component" (310 GBM, 181 PCP); ii) canonical pathways (880 GBM, 812 PCP),compiled from Biocarta (217 GBM, 214 PCP), KEGG (186 GBM, 183 PCP) and Reactome (430 GBM, 415 PCP); iii) curated gene sets taken from the literature on chemical and genetic perturbations ("literature sets," 2392 GBM); iv) tissue specific gene sets (25 GBM) determined previously[10]; and v) "special" gene sets taken from the literature on the cancer types addressed in this study (see below).

The "enrichment analysis" includes "over-representation" analysis, "overexpression" analysis, and their combination.[10,29] Over-representation estimates the probability to find members of a given set in a list, e.g., the genes included in a spot cluster, compared with their random appearance independent of their expression scores. For any gene set, right-tail modified Fisher exact test was used to determine whether the number of genes within this set is overrepresented in a particular list of genes included in a spot-cluster. The hypergeometric distribution then provides a P-value for each set and spot which estimates the probability to find a stronger overlap between the genes in a spot cluster and the set than expected by chance given a certain total number of genes studied.[37,38] We considered over-represented sets with $P \leq 0.0001$ which ensures reasonable adjustment for false positives in the multiple testing problem.

Contrarily, the term "overexpression" defines the deviation between the mean expression value averaged over the set-members included in a spot-cluster and the mean expression value of genes independent of their over-representation. The gene set Z-score (GSZ) merges both gene set overrepresentation and overexpression approaches.[10,30] In particular, the GSZ-score for all the genes studied is given by,[10]

$$GSZ = \frac{\langle \Delta e \rangle_{set} - \langle \Delta e \rangle_{all\_genes}}{\sqrt{\mathrm{var}(\Delta e) / N_{set}}}, \tag{8}$$

The denominator defines the respective standard error. The GSZ-score defined by **Equation 8** thus estimates the degree of significance of concerted changes of the expression of groups of genes in a particular sample relative to the mean expression of all genes. We use the GSZ-score to profile overexpression of a selected gene set.

In addition to GSZ-profiles we generated gene set population maps to visualize the distribution of the genes of a selected set in the SOM portraits. This population map color codes the number of genes taken from the set in each of the tiles of the mosaic image. It ranges from white (no gene) to maroon (maximum number per tile observed for the particular gene set).

Finally we generated gene set over-representation heatmaps using an algorithm described previously.[10] We merged the top three gene sets per spot in a sample. Redundant gene sets were removed and represented by their minimum P-value. The resulted non-redundant global list of gene sets was converted into the HG or GSZ enrichment heatmaps by applying either one-way (only gene sets) or two-way (gene sets and samples) hierarchical clustering.

Sample similarity analysis

Similarity analysis compares the expression states in SOM portraits. It uses meta-genes instead of single genes as the basal data, which has the advantage of improving the representativeness and resolution of the results.[6,31]

We applied second-level SOM analysis as proposed by Guo et al.[9] to visualize the similarity between the individual SOM meta-gene expression patterns.

Independent component analysis (ICA)[32] was applied to the SOM meta-genes using the Rpackage "fastICA." It distributes

the samples in the space spanned by the components of minimum mutual statistical dependence. These components point along the directions of maximum information content in the data which is estimated by their deviation from a (non-informative) normal distribution.[33] ICA was based on covariance matrix calculated in terms of Pearson correlation coefficients between all meta-genes from any two samples. The correlation matrix was visualized using pairwise correlation maps (PCM), minimum spanning tree (MST) and correlation cluster net (CN)representations.

MST"s have been shown to be useful for clustering and classification of cancer subtypes using microarray data.[34] For the MST calculation we use the *spantree* function of the R package "igraph." A major disadvantage of this method is the lack of ancestral states (inner nodes) in an MST, as opposed to phylogenetic trees where subtypes are leaves in the tree and other nodes are created as ancestral states. On the other hand, MST rigorously converts the multi-dimensional clustering problem to a tree partitioning problem which simplifies the interrelationship between the data without essential loss of information.[35]

CN constructs an unweighted graph by connecting the nodes (samples) whose pairwise correlation coefficient exceeds a given threshold (0.5 here). This graph supplements the sparse MST with a more detailed and network-like overview about the sample correlation structure. It implies more connections than MST and thus considers also weaker mutual correlations.

Finally, we also applied the neighbor-joining algorithm (Rpackage "ape") to represent similarity relations based on the Euclidean distances between the samples in terms of similarity trees.[36] The distances between pairs of samples in the tree are in scale. In contrast to MST-representation the phylogenetic tree allows to identify "bush-like" clusters of similar clusters and to estimate the degree of mutual dissimilarity between them.

Program

Each cancer data set was analyzed in a separate training run. We used our R program "oposSOM" for SOM training and downstream analysis.[6] It is available as R package on CRAN repository (http://cran.r-project.org/). Complete reports of our analyses are available on our website (http://som.izbi.uni-leipzig.de).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

HB conceived and designed most parts of this study, performed data analysis and wrote the manuscript; LH and HW wrote the programs and performed the calculations, contributed to the manuscript. MF performed hook-preprocessing.

We thank Katja Nowick for introducing us into wTO networks. We thank also Edith Willscher for performing part of the calculations on GBM. This publication is supported by LIFE–Leipzig Research Center for Civilization Diseases, Universität Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF), and by the Free State of Saxony within the framework of the Excellence Initiative.

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/systemsbiomedicine/article/25897

References

1. Guazzaroni M-E, Beloqui A, Golyshin P, Ferrer M. Metagenomics as a new technological tool to gain scientific knowledge. World J Microbiol Biotechnol 2009; 25:945-54; http://dx.doi.org/10.1007/s11274-009-9971-z
2. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol 2011; 12:125; PMID:21867570; http://dx.doi.org/10.1186/gb-2011-12-8-125
3. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med 2010; 2:84; PMID:21114804; http://dx.doi.org/10.1186/gm205
4. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, et al.; MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 2010; 28:827-38; PMID:20676074; http://dx.doi.org/10.1038/nbt.1665
5. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982; 43:59-69; http://dx.doi.org/10.1007/BF00337288
6. Wirth H, Löffler M, von Bergen M, Binder H. Expression cartography of human tissues using self organizing maps. BMC Bioinformatics 2011; 12:306; PMID:21794127; http://dx.doi.org/10.1186/1471-2105-12-306
7. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al.; Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17:98-110; PMID:20129251; http://dx.doi.org/10.1016/j.ccr.2009.12.020
8. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, et al. Integrative molecular concept modeling of prostate cancer progression. Nat Genet 2007; 39:41-51; PMID:17173048; http://dx.doi.org/10.1038/ng1935
9. Guo Y, Eichler GS, Feng Y, Ingber DE, Huang S. Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers. J Biomed Biotechnol 2006; 2006:69141; PMID:17489018; http://dx.doi.org/10.1155/JBB/2006/69141
10. Wirth H, von Bergen M, Binder H. Mining SOM expression portraits: Feature selection and integrating concepts of molecular function. 2011; submitted: http://precedings.nature.com/documents/6666/version/1
11. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98:5116-21; PMID:11309499; http://dx.doi.org/10.1073/pnas.091062498
12. Dong H, Siu H, Luo L, Fang X, Jin L, Xiong M. Investigation gene and microRNA expression in glioblastoma. BMC Genomics 2010; 11(Suppl 3):S16; PMID:21143783; http://dx.doi.org/10.1186/1471-2164-11-S3-S16
13. Liu BH, Goh CHK, Ooi LLPJ, Hui KM. Identification of unique and common low abundance tumour-specific transcripts by suppression subtractive hybridization and oligonucleotide probe array analysis. Oncogene 2008; 27:4128-36; PMID:18332864; http://dx.doi.org/10.1038/onc.2008.50
14. Wolfer A, Wittner BS, Irimia D, Flavin RJ, Lupien M, Gunawardane RN, Meyer CA, Lightcap ES, Tamayo P, Mesirov JP, et al. MYC regulation of a "poor-prognosis" metastatic cancer cell state. Proc Natl Acad Sci U S A 2010; 107:3698-703; PMID:20133671; http://dx.doi.org/10.1073/pnas.0914203107
15. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 2004; 101:9309-14; PMID:15184677; http://dx.doi.org/10.1073/pnas.0401994101
16. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci 2008; 28:264-78; PMID:18171944; http://dx.doi.org/10.1523/JNEUROSCI.4178-07.2008
17. Lu T, Pan Y, Kao S-Y, Li C, Kohane I, Chan J, Yankner BA. Gene regulation and DNA damage in the ageing human brain. Nature 2004; 429:883-91; PMID:15190254; http://dx.doi.org/10.1038/nature02661

18. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006; 22:1600-7; PMID:16606683; http://dx.doi.org/10.1093/bioinformatics/btl140
19. Binder H, Preibisch S. "Hook"-calibration of GeneChip-microarrays: theory and algorithm. Algorithms Mol Biol 2008; 3:12; PMID:18759985; http://dx.doi.org/10.1186/1748-7188-3-12
20. Binder H, Krohn K, Preibisch S. "Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures. Algorithms Mol Biol 2008; 3:11; PMID:18759984; http://dx.doi.org/10.1186/1748-7188-3-11
21. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19:185-93; PMID:12538238; http://dx.doi.org/10.1093/bioinformatics/19.2.185
22. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Trans Neural Netw 2000; 11:586-600; PMID:18249787; http://dx.doi.org/10.1109/72.846731
23. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005; 4:e17; PMID:16646834; http://dx.doi.org/10.2202/1544-6115.1128
24. Nowick K, Gernat T, Almaas E, Stubbs L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. Proc Natl Acad Sci U S A 2009; 106:22358-63; PMID:20007773; http://dx.doi.org/10.1073/pnas.0911376106
25. Quackenbush J. Genomics. Microarrays--guilt by association. Science 2003; 302:240-1; PMID:14551426; http://dx.doi.org/10.1126/science.1090887
26. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007; 23:980-7; PMID:17303618; http://dx.doi.org/10.1093/bioinformatics/btm051
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.; The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000; 25:25-9; PMID:10802651; http://dx.doi.org/10.1038/75556
28. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A. BioMart Central Portal--unified access to biological data. Nucleic Acids Res 2009; 37(suppl 2):W23-7; PMID:19420058; http://dx.doi.org/10.1093/nar/gkp265
29. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics 2009; 10:47; PMID:19192285; http://dx.doi.org/10.1186/1471-2105-10-47
30. Törönen P, Ojala PJ, Marttinen P, Holm L. Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. BMC Bioinformatics 2009; 10:307; PMID:19775443; http://dx.doi.org/10.1186/1471-2105-10-307
31. Wirth H, von Bergen M, Murugaiyan J, Rösler U, Stokowy T, Binder H. MALDI-typing of infectious algae of the genus Prototheca using SOM portraits. Journal of Microbiological Methods 2011; 88:83-97; PMID:22062088; http://dx.doi.org/10.1016/j.mimet.2011.10.013
32. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw 2000; 13:411-30; PMID:10946390; http://dx.doi.org/10.1016/S0893-6080(00)00026-5
33. Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics 2002; 18:51-60; PMID:11836211; http://dx.doi.org/10.1093/bioinformatics/18.1.51
34. Riester M, Stephan-Otto Attolini C, Downey RJ, Singer S, Michor F. A differentiation-based phylogeny of cancer subtypes. PLoS Comput Biol 2010; 6:e1000777; PMID:20463876; http://dx.doi.org/10.1371/journal.pcbi.1000777
35. Xu Y, Olman V, Xu D. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. Bioinformatics 2002; 18:536-45; PMID:12016051; http://dx.doi.org/10.1093/bioinformatics/18.4.536
36. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4:406-25; PMID:3447015
37. Vêncio RZ, Shmulevich I. ProbCD: enrichment analysis accounting for categorization uncertainty. BMC Bioinformatics 2007; 8:383; PMID:17935624; http://dx.doi.org/10.1186/1471-2105-8-383
38. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005; 33(suppl 2):W741-8; PMID:15980575; http://dx.doi.org/10.1093/nar/gki475

Supplementary Text

Portraying the expression landscapes of cancer subtypes: a glioblastoma multiforme and prostate cancer case study

Lydia Hopp[1,2]*, Henry Wirth[1,2]*, Mario Fasold[1,2] and Hans Binder[1,2#]

[1] Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Germany
[2] LIFE, Leipzig Research Center for Civilization Diseases; Universität Leipzig, Germany

to whom correspondence should be send: binder@izbi.uni-leipzig.de, wirth@izbi.uni-leipzig.de
* contributed equally

1 Materials and Methods

1.1 Hook quality control and preprocessing of microarray intensities

Quality control and calibration of microarray data account for detection and correction of technical variation. Our hook approach generates chip-specific metrics using the raw intensity data of each particular GeneChip independently [1-3]. It generates a series of chip characteristics suited for quality control and the assessment of the global expression degree.

In short, the hook method applies to microarrays of the GeneChip-type containing pairs of perfect match (PM) and mismatch (MM) probes. It independently analyzes the intensity data of each GeneChip microarray using the two-species Langmuir hybridization isotherm which assumes competitive binding of specific and 'representative' non-specific transcripts to each probe. The method processes the PM and MM probe intensities (I^{PM} and I^{MM}, respectively) using the transformation, $\Delta = \log I^{PM} - \log I^{MM}$ *and* $\Sigma = \frac{1}{2}\left\langle \log I^{PM} + \log I^{MM} \right\rangle_{set}$, where $<\ldots>_{set}$ denotes averaging over each probe set of usually 11 PM/MM probe pairs addressing one transcript. Smoothing of the delta-versus-sigma plot provides the hook curve which enables decomposition of the probe signals into contributions due to specific and non-specific hybridization by simple graphical analysis and subsequent correction of the intensities for sequence specific effects using the positional-dependent nearest neighbour model as standard [4-6]. The corrected signals are re-plotted into delta-versus-sigma coordinates and again smoothed to obtain the corrected version of the hook curve. The curve divides into five hybridization regimes: non-specific (N), mixed, specific (S), partial (sat) and complete (as) saturation (left part of Figure S 1).

Analysis of the hook curve in terms of the two-species Langmuir binding model provides the two parameter couples (Σ^{start}= log N, Δ^{start}) and (β= log M – log N, α) characterize the position and the geometrical dimensions of the hook curve in terms of the coordinates of their starting point and their width and height, respectively (Figure S 1, upper part). They are related to well-defined hybridization characteristics of the selected chip, namely the background intensity level due to non-specific hybridization ('start' coordinates), the saturation level of the probes (end coordinates), the non-specific background in dimensionless units of the logged binding strength (width) and the difference of the logged binding strengths between the PM and MM probes (height). These parameters change in a characteristic fashion owing to different effects which implies theirs application in quality control tasks of large scale microarray studies. Particularly, the horizontal shift of the whole hook curve reflects alterations of the intensity scaling of the measurements, e.g. after changes of the scanner settings as schematically illustrated in Figure S 1a. The widening of the curve after the left-shift of its increasing branch can be attributed to dilution effects, e.g. if one reduces the amount of RNA used for hybridization (Figure S 1b, see ref. [7] for details). The change of the vertical dimension of the hook curve typically reflects either alterations of the MM-design or alterations of the washing efficiency (Figure S 1c and d, see ref. [8] for details). In addition, the hook method estimates the percentage of absent probes (%N) showing intensities below the detection threshold of a particular hybridization (log N, see ref. [2]) and the hybridization dependent degradation index which inversely scales with the RNA quality in terms of the mean transcript length (see [2, 9]).

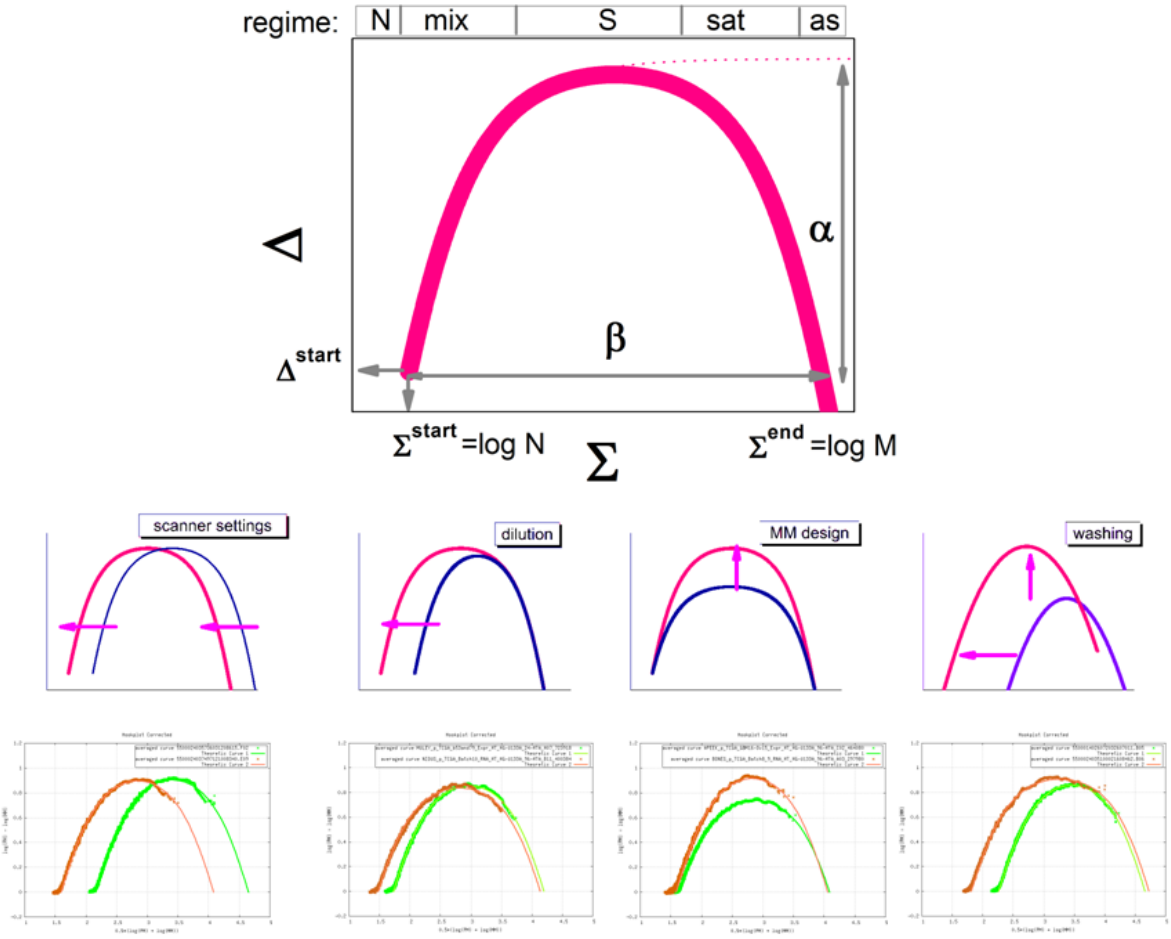

Figure S 1: Schematic illustration of the hook dimensions and their modifications due to typical experimental effects: The hook curve is defined by its height (α), width (β), and 'start- and end-coordinates' (part above). The bottom part illustrates the effects of a) optical scaling of the intensity owing to changes of the scanner settings or the labeling which equally shift the start AND end points in horizontal direction; b) alterations of the non-specific background level owing to changes of the amount of RNA and/or of its composition which shift ONLY the start point in horizontal direction giving rise to the narrowing of the hook for larger background contributions; c) modifications of the mismatch design change the vertical dimensions of the hook, e.g. a smaller PM/MM gain decreases its height; and d) alterations of the washing efficiency mainly affect the height and width of the hook curve. The panels below compare pairs of hook curves taken from different batches of the GBM-series referring to the effects schematically illustrated in panels a) – d), respectively. The thick curves are experimental data and the thin curves are theoretical hook curves calculated according to the competitive Langmuir binding model [1].

We applied hook calibration to the whole GBM data set of more than 500 cel files downloaded from the 'The Cancer Genome Atlas' (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/). The series divides into 16 distinct groups of sample identifiers which refer to different batches. Figure S 2 depicts the hook parameters of all samples studied. Batch effects become evident by stepwise changes of the parameter values between the batches.
Comparison with the characteristic alterations of the hook curves illustrated in Figure S 1 allows interpreting the observed hook parameters: For example, batches 5 and 6 differ by start and end points of the hook, log N and log M, respectively, whereas the dimensions of the hook (height and width) remain virtually unaffected. This situation refers to the shift of the whole hook curve illustrated in Figure S 1a. Comparison with Figure S 1d showing examples from the two batches 5 and 6 confirm this expectation. Hence, the two batches discussed presumably differ in the scanner settings used, where batch 6 is measured using a lower intensity level.
Another comparison of batches 10 and 15 reveals, that the former one is slightly diluted compared with the latter one as indicated by the decreased width of batch 15 (compare also Figure S 1b). Batches 6 and 7, on the other hand, differ solely by the heights of their hook curves (compare also Figure S 1c). The percentage of absent probes shows

only weak batch effects. It fluctuates however strongly in selected batches (e.g., batch 2). The degradation parameter indicates poor RNA-quality of batches 7, 12 and 13.
To judge the batch effect on the expression values obtained after hook calibration we defined stratified the genes into 10 groups with increasing expression level. Figure S 3a shows the logged mean values averaged over each group as a function of the sample index. Especially the alteration of the scanner settings between batches 5 and 6 systematically changes the expression levels especially for larger expression values. This bias remains largely uncorrected by the hook calibration applied. We therefore used quantile normalization in the next step which largely removes the bias from the data (Figure S 3b). Note that our approach normalizes expression data in contrast to standard normalization methods such as vsn and RMA which normalize raw probe intensities which might be problematic due to the up-down effect [7].

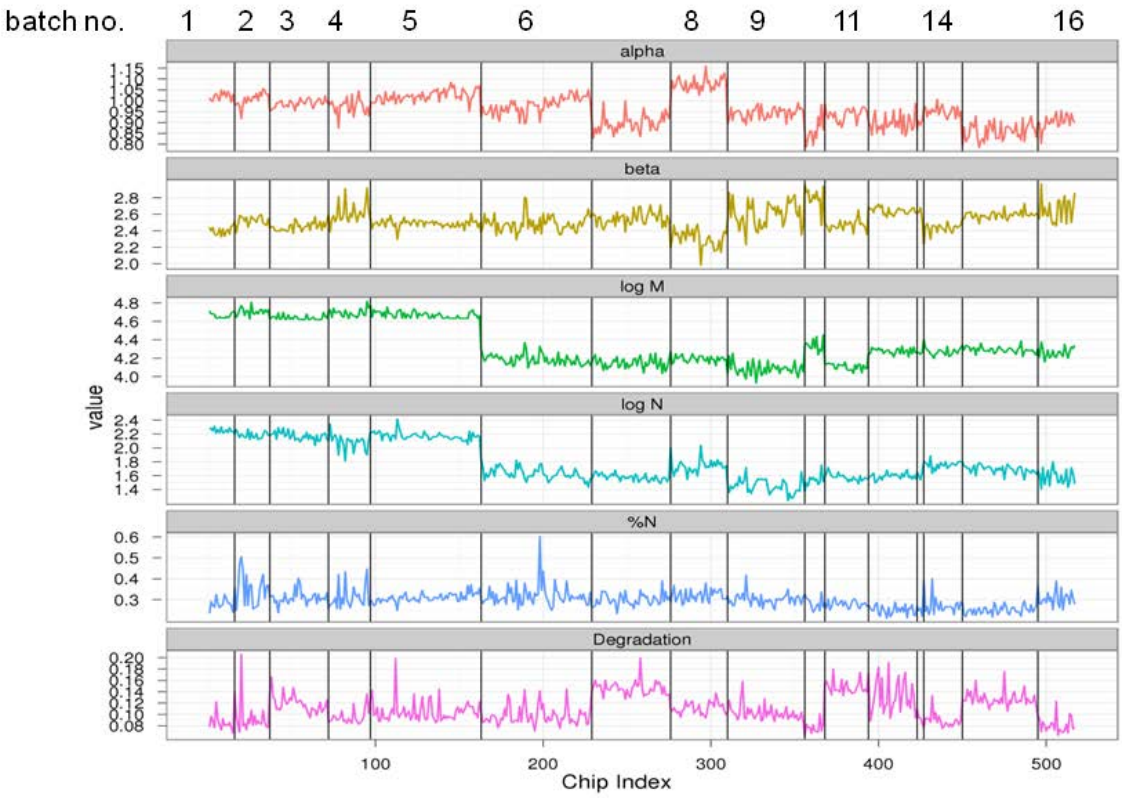

Figure S 2: Hook analysis of the GBM-series: The figure shows the hook parameters (from top to bottom) alpha (height), beta (width), log M (mean logged saturation intensity), log N (mean logged intensity of the non-specific background), %N (percentage of absent probes) and the degradation index for each of the more than 500 samples of the series. The data set divides into 16 batches (see vertical lines). Typically the hook-parameters stepwisely change between the batches.

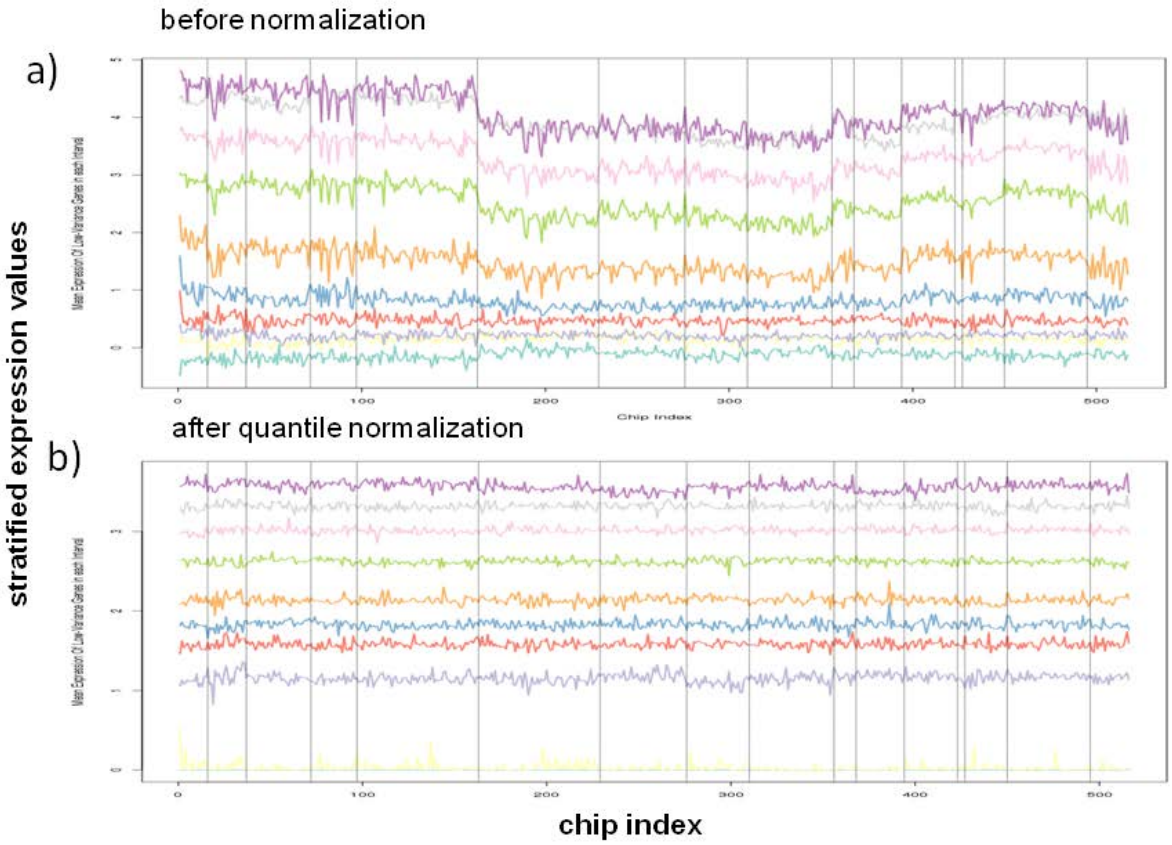

Figure S 3: Stratification of expression values into 10 groups of genes differing in their expression values. The figure shows the log-mean averaged over all genes of each group in all samples studied before (panel a) and after quantile normalization of the expression values (panel b).

Figure S 4a shows the distributions of the expression values before and after quantile normalization. Each of these distributions are characterized by a bimodal shape: It's left peak at smaller expression can be attributed to 'absent' and thus to virtually inactive genes whereas its right peak values refers to 'present' and thus to active genes [10, 11]. The 'absent-peak' due to non-specific hybridization is non-informative with respect to the target genes because their expression is smaller than the detection threshold of the method. The logged expression values of absent genes are therefore arbitrarily set to zero. Expression values of genes in the overlap region of both peaks are scaled with a factor $0<pc<1$ which estimates the relative contribution of specific hybridization. Expression values of present genes are used without further scaling. In the last step the logged expression values of each gene are centered with respect to the mean value averaged over all samples considered in the series of samples. A relative log-expression value of zero consequently means that the gene is expressed according to its mean expression value and positive and negative values refer to over and under-expression, respectively.

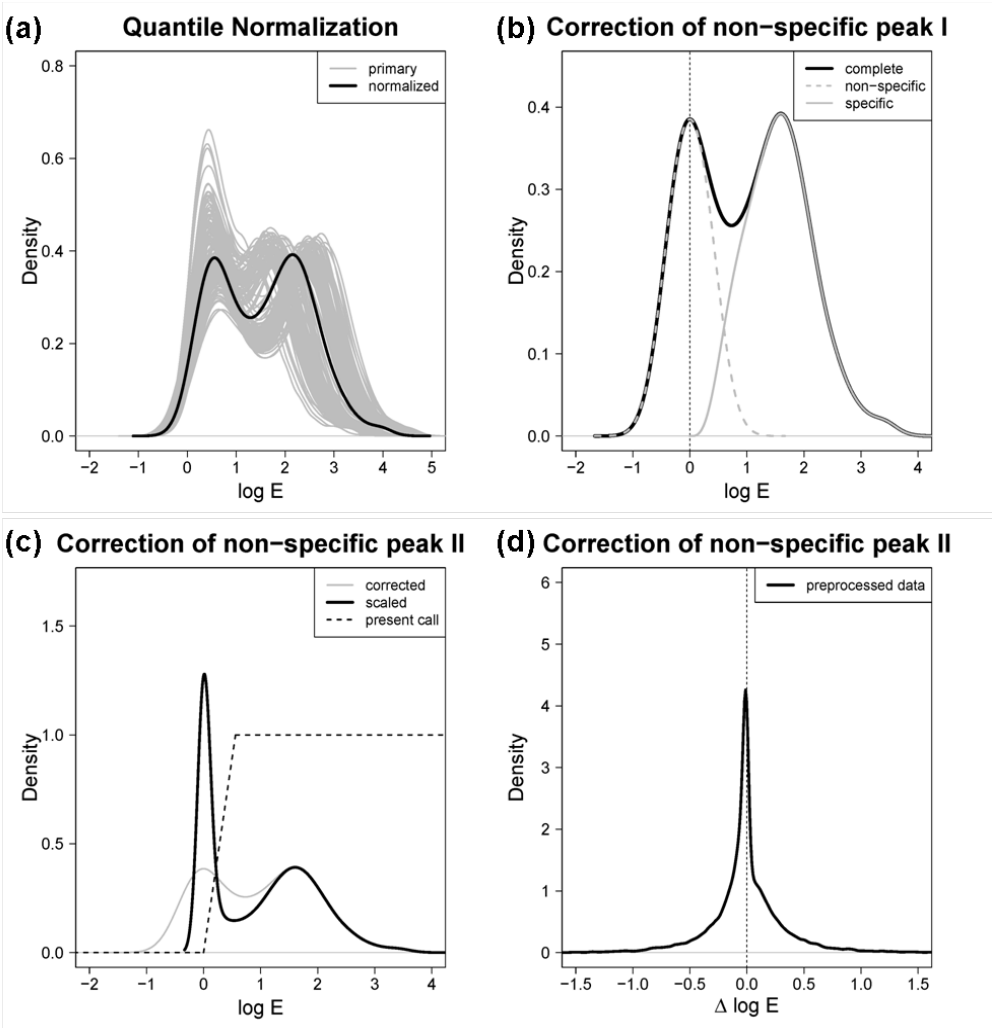

Figure S 4: Normalization and adjustment of expression values: The distributions of hook-calibrated expression values of the samples studied merge into one representative mean distribution after quantile normalization (panel a). Its double peaked shape is decomposed into two single peaked distributions due to non-specific and specific hybridization at small and larger expression values, respectively (b). The fraction of the specific signal contributing to the total signal density (`present call`, dashed curve) is used as weighting coefficient of the logged expression values, which narrows the left peak of the total signal density (c). Finally, the expression values are normalized with respect to the logged mean expression of each gene (d). The intense central peak refers to invariant genes under all conditions studied.

1.2 Error characteristics of cancer subtypes

The standard error of the expression of each gene was estimated using a modified locally pooled error (LPE) approach [12]: All gene-specific expression values of one subtype are treated as pseudo-replicates which provide the standard deviation for each gene for each subtype. These data are then plotted as a function of their logged expression value and smoothed using a sliding window of appropriate width (Figure S 5). The obtained LPE-function decays with increasing expression. A combination of the LPE-value and of the individual standard deviation of each gene are used to estimate differential expression using a regularized t-score [12].
The mean over the LPE-function provides the average standard deviation of each cancer subtype. Figure S 5 shows that the error level slightly varies between the subtypes being minimal in normal brain samples and being maximal in PN-GBM samples.

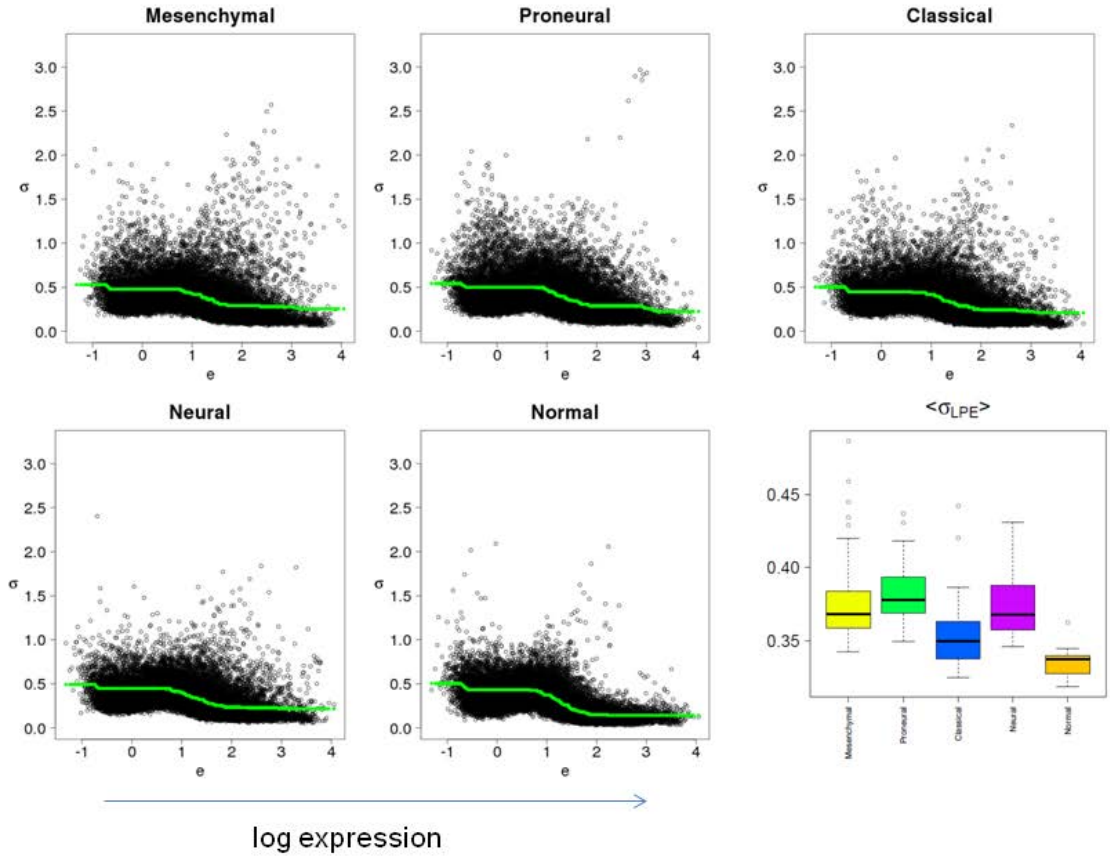

Figure S 5: Error characteristics of GBM subtypes: The figures show error distributions (dots) and locally pooled estimates (green curves) of GBM-subtypes as a function of the logged expression, e. The LPE-curves are calculated as moving average over 500 single probe values. The bottom-right box plot depicts the mean errors for each subtype.

1.3 Supporting maps: Population, variance and distance maps

The population, variance and distance (U-matrix) maps shown in Figure S 6 provide information about the number of single genes per metagene minicluster, about the variability of the metagene profiles and about their mutual Euclidian distances using appropriate color coding. SOM-machine learning scales the difference between the expression profiles of adjacent metagenes inversely to their population, i.e., adjacent metagene profiles become more similar for highly populated metagenes. This way the method tends to distribute the single genes over as much as possible tiles. The population map reveals that the single genes inhomogeneously distribute among the tiles of the mosaic. Highly populated metagenes (see yellow and red tiles) predominantly group along the edges of the map whereas only a few highly populated tiles are found in its central area. A zone of 'empty' metagenes not containing real genes (G/M=0, see dark blue tiles) clusters around the highly populated central area of the map. The tiles of maximum population in the central area refer to genes with virtually invariant, mostly absent specific expression in all samples studied. These invariant genes give rise to the dark blue spot in the central area of the variance map. The variance map also reveals that other nearly invariant metagenes cluster around this tile in the central area of the map (see blue and green areas). Both, invariant and empty metagenes carry essentially no specific information as classification markers in transcriptional profiling. Hence, the tiles occupied by empty and invariant genes form regions not suited for differential expression analysis between the cancer subtypes studied.
The more variant and higher populated metagenes reveal an underlying spot like pattern preferentially along the boundaries of the map (red areas), which agrees with the over- and underexpression spots detected in the SOM mosaics of individual samples. The distance map color codes the distances between adjacent metagenes [13]. Dark coloring corresponds to a large distance and thus a gap between the features whereas a light coloring signifies that the metagene profiles are close to each other. Light areas thus can be thought as clusters and dark areas as cluster separators.

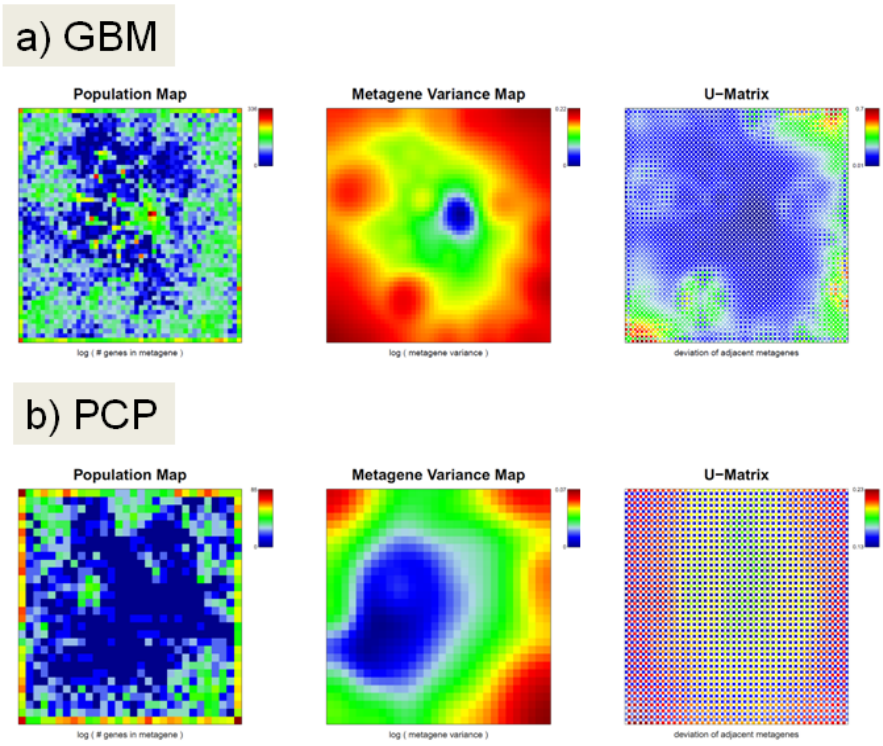

Figure S 6: Supporting maps characterizing the SOM trained for each cancer set: The population map visualizes the number of single genes per metagene. Highly populated metagenes accumulate along the edges and partly also in the centre of the maps (red tiles). The metagene variance map color codes the variance of the metagene profiles. Virtually invariant metagene profiles form the central blue spot whereas highly variant ones are found in the peripheral regions of the map. The U-matrix color-codes the distances between neighboring metagene profiles: Dark regions refer to larger and light to closer distances.

1.4 Alternative methods of module selection

Expression modules are selected using the percentile over- and underexpression criteria as standard. Alternative criteria are mutual correlations between the metagenes using a seed algorithm or distance based K-means clustering [14]. Each method selects clusters of specific areas and sizes in the map (see Figure S 7) which, in turn, can be adjusted by appropriate thresholds separating the clusters. The mapping of the obtained clusters into the SOM images provides an intuitive view on the consequences of different clustering algorithms. For example, correlation clustering arranges the clusters roughly in form of three annuluses (blue, green, red) referring to decreasing values of the correlation coefficients for mutual correlations of the metagegenes in the cluster. The respective K-means clusters and overexpression spots mostly correspond to the outer annulus, however the different similarity metrics produce different shapes of the clusters (see below).

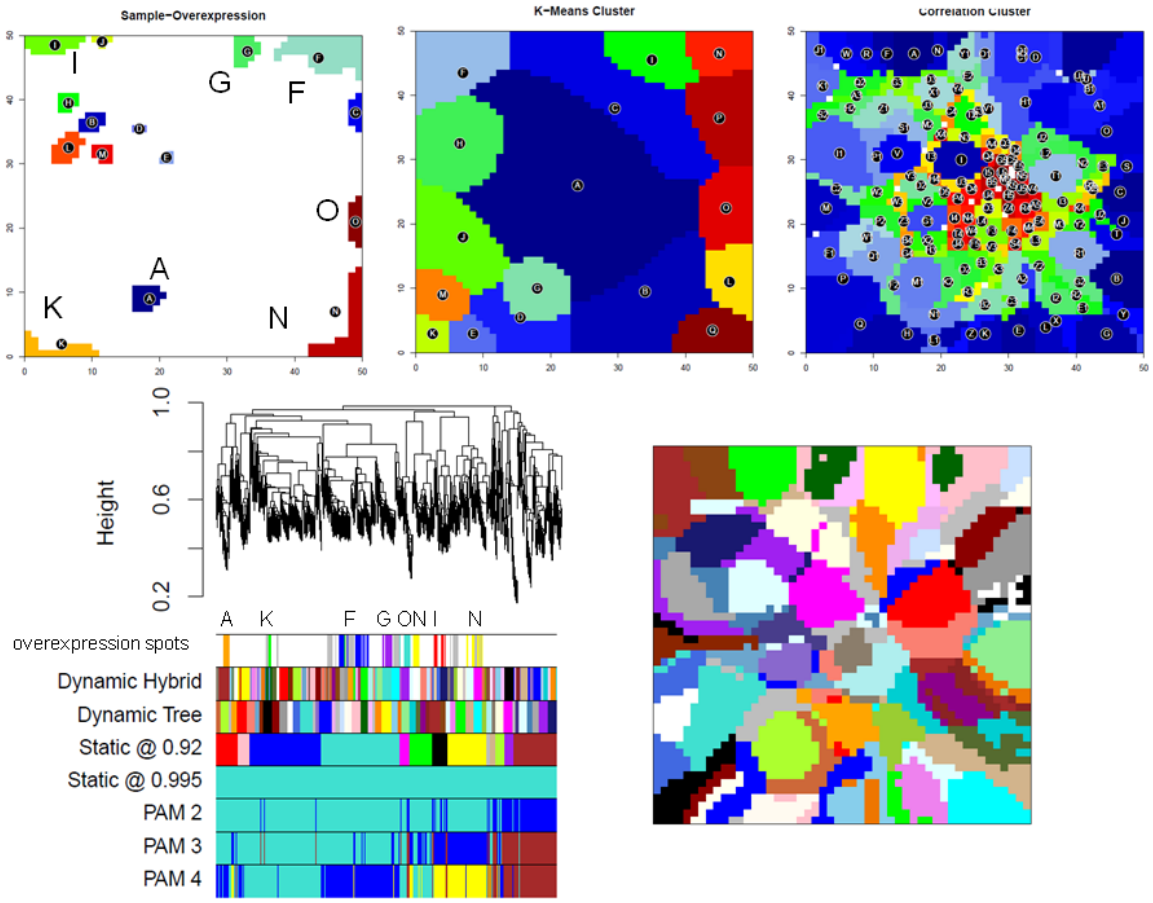

Figure S 7: Alternatively to spot selection based on differential over- and underexpression one can apply K-means clustering or correlation clustering based on Euclidian distances or Pearsons correlation coefficients between the metagenes (upper row of figures). As an additional method we applied Dynamic Tree Cut (which also includes alternative tree-cut methods) and map the obtained clusters into the SOM image (row below). Vice versa, the overexpresssion spots are visualized as color bar (each color refers to one overexpression spot) in the Dynamic Tree Cut algorithm.

As an additional option we applied 'Dynamic Tree Cut', a hierarchical clustering algorithm based on correlation metrics with dynamic branch cutting [15], to the metagenes and mapped the obtained clusters into the SOM (Figure S 7). The report produced by the program 'Dynamic Tree Cut' assigns clusters obtained by different cutting algorithms as color bars below the clustering tree (Figure S 7). We add a bar that visualizes the overexpression spots in this presentation. It shows that the overexpression spots refer to disjunct outer branches of the tree. The projection of the Dynamic Tree Cut-clusters into the SOM mosaic reveals similarities of the cluster patterns with our correlation seed algorithm which reflect the common correlation metric used in both cases. Note that these correlation

clusters tend to extend in radial direction whereas the overexpression spot clusters extend often along the borders of the map. K-means clusters partly more resemble the overexpression spots but they occupy more extended areas mainly due to the particular threshold. The different cluster selection algorithms thus provide different options of expression module selection. We applied the intuitive and easy-to-interpret differential expression criterion for module selection as standard. It selects relatively small and localized expression modules. On the other hand, wider areas of metagenes are not included into the clusters. The impact of the alternative methods on the functional interpretation will be addressed elsewhere.

1.5 The effect of alternative preprocessing methods on downstream analysis: hook versus RMA

We compared the effect of preprocessing on downstream analysis results after SOM training. Particularly, we used either our hook calibrated and quantile normalized expression data or RMA-preprocessed data which are downloaded from the TCGA-website (level 2 data) of the same sample set. Figure S 8 shows correlation net (CN) similarity plots and the SOM portraits of selected samples.

The CN of both data sets strongly agree showing only small differences in the overall similarity patterns of subtypes. A few samples shift their position slightly, e.g. sample no. 321 (Figure S 8). The spot patterns of the individual portraits differ but they show essentially the same trends, e.g. of global underxpression in sample no. 156 and the appearance of more than two spots in sample no. 084. Global spot analysis shows that hook calibration provides a slightly more diverse pattern of differential expression (Figure S 9) and a slightly larger number of distinct spots (data not shown).

Figure S 10 directly compares the spots patterns seen by both methods in terms of the overexpression summary maps. Hook and RMA data produce very similar patterns which strong correspondence between hook- and RMA-spots in terms of their localization and ordering in the map (compare their ordering in clockwise direction along the border of the map) and also in terms of size and of the number of metagenes and single genes included in the spots. The hook-map in total detects a few more spots (e.g., spot J).

In the next step we compare spot lists (of genes in the particular spots clusters; Figure S 11) and total lists (of all genes studied; Figure S 12) in selected samples overexpressing the respective spots. The ranked lists are generated using three scores (see [12] for details): the fold change (FC) considers only the differential expression of the genes; the regularized t-score also takes into account the error of differential expression; and the WAD score is similar to FC but it more heavily weights large expression values. The rank comparisons plots in Figure S 11 clearly shows that the regularized t-score produces better agreement between hook and RMA spot lists presumably because it explicitly considers the error which removes less significant genes from the top of the list. FC and WAD lists are virtually identical (data not shown). Nevertheless, also t-scored hook and RMA lists mostly rank genes in different orders.

Comparison of the total lists using the correspondence at the top (CAT) plots in Figure S 12 shows that at the top-50 of the lists FC and WAD rankings better agree between the methods presumably because these genes refer to large expression values and thus to reliable data. At ranks larger than 50 the lists agree mostly to 60% - 80% of the genes considered.

Finally we compared the functional context of the spots as seen by hook and RMA by applying gene set enrichment analysis of gene sets of the category biological process (BP, see Table S 1). Especially ‘leading’ spots collecting strongly overexpressed genes largely agree in the sets enriched (e.g. spot K, and partly E, F and D). Note that especially these spots govern the classification of the samples into different subtypes. Other spots agree to a less degree in their functional context (e.g. spots A and G). Note that spots A and G are related to one of the subtypes (NL) of GBM and thus they are important to interpret its functional context. A third group of spots partly found in the central area of the map diverge in the gene sets detected.

In summary, the choice of the preprocessing method seems to have only a small effect on the most prominent properties of the expression landscape which govern the similarities between the samples (and thus their classification into different subtypes) and the functional context of the leading expression modules. On the other hand, gene lists partly differ especially for genes differentially expressed at intermediate and lower levels. The significance and impact of such differences can not been judged here and must be addressed separately.

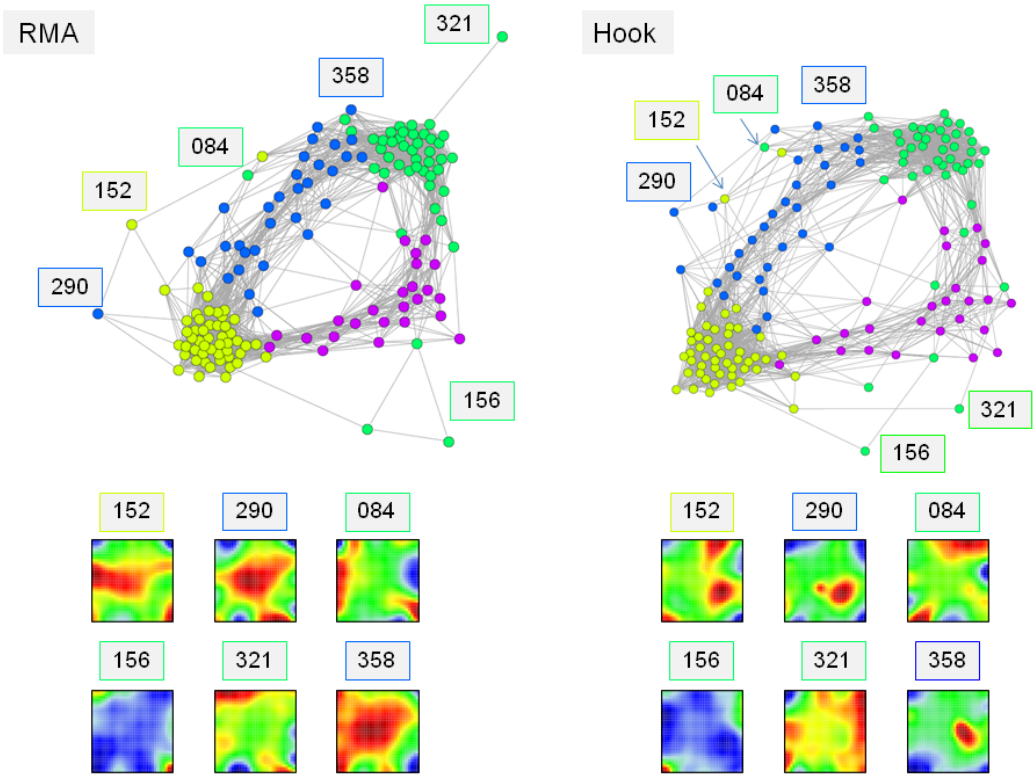

Figure S 8: The effect of preprocessing using either RMA or hook methods on downstream similarity analysis using correlation nets and SOM portraits of selected samples (the position of the samples in the net are indicated by their ID's).

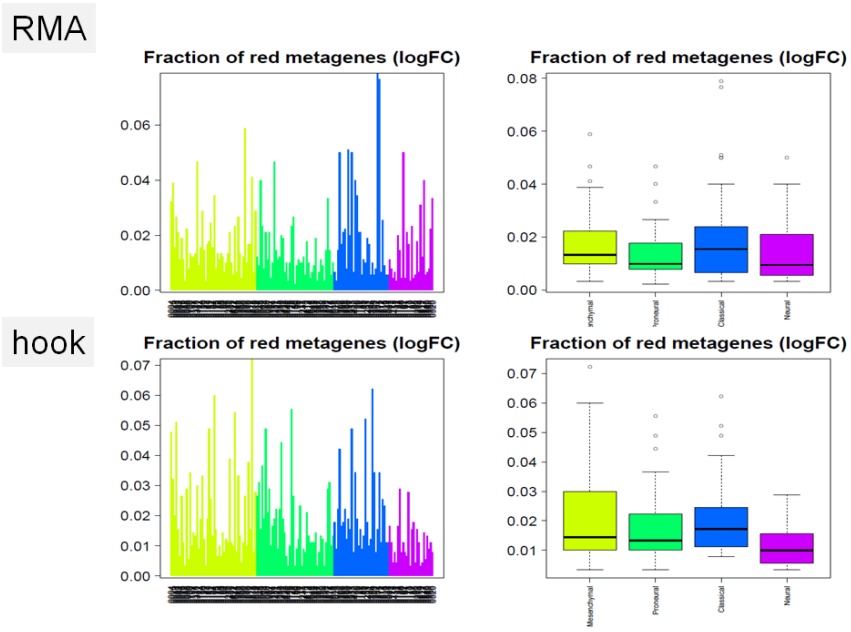

Figure S 9: Fraction of metagenes in the overexpression spots after RMA and hook preprocessing.

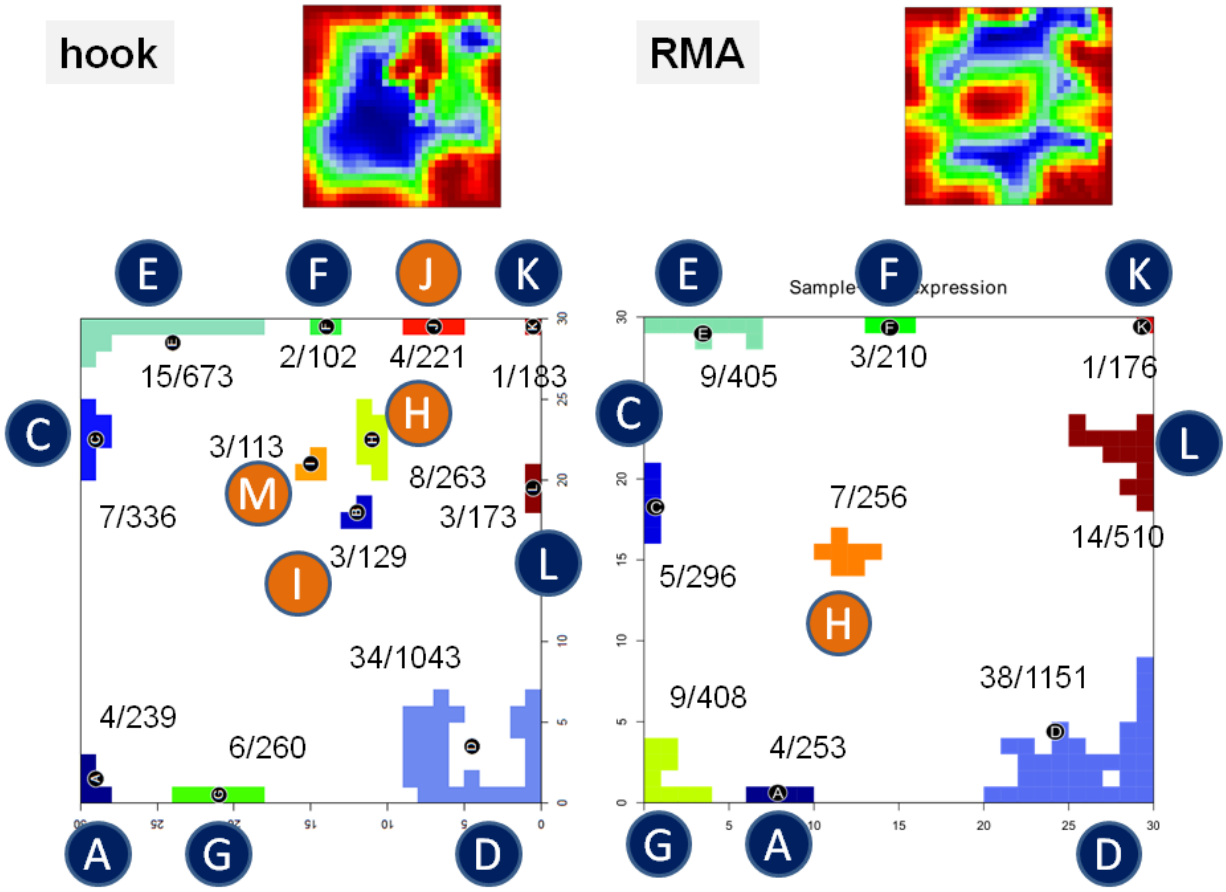

Figure S 10: Spot comparisons as seen by hook and RMA: Overexpression summary maps (small images in the part above) and selected spots (below) are labeled by capital letters spots and blue circles for twin-spots with similar functional context and red circles for spots without clear correspondence. The numbers give the ratio of metagenes/genes in the respective spots. Note that both spot patterns are very similar showing virtually the same spot order along the border of the map (spots A and G exchange their order and spot J is additionally detected by hook). Most of the corresponding spots also agree in their size and the number of metagenes/single genes.

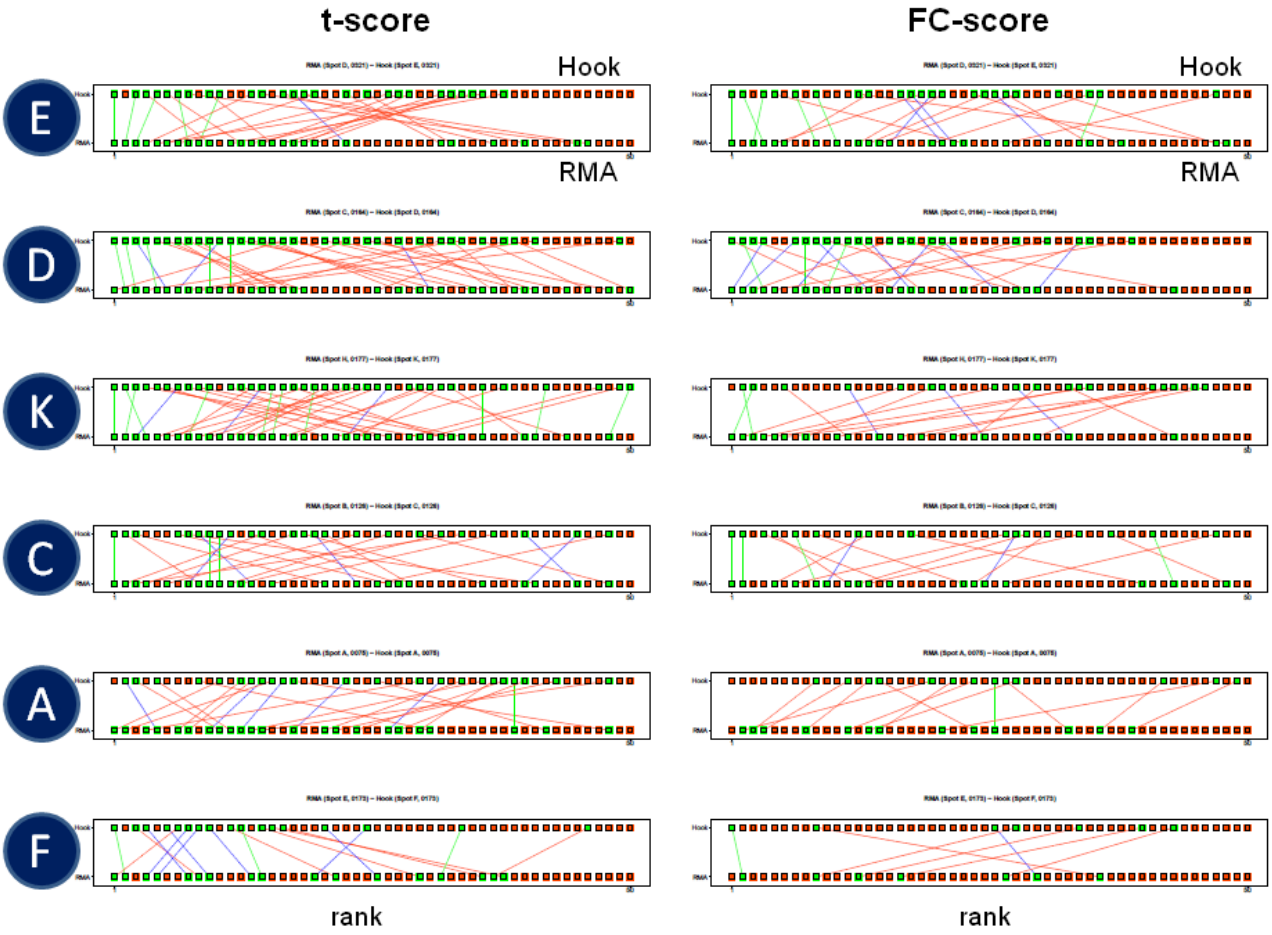

Figure S 11: Comparison of selected top-50 spot lists of genes obtained by hook and RMA using the regularized t-score or the FC-score in terms of rank comparison (RC) plots. The letters assign the spots (see Figure S 10). Each gene in each top-50 hook-list is represented by a small circle colored green if it is also found in the respective top-50 RMA-list. Such overlap genes are connected by green/blue/red lines if the rank difference is small/intermediate/large. Non-overlap genes are indicated by red circles. Hence, the more lines connect both lists the more similar they are. Green lines overweight blue ones and blue ones overweight red ones. Each comparison refers to one selected sample strongly expressing the respective spot.

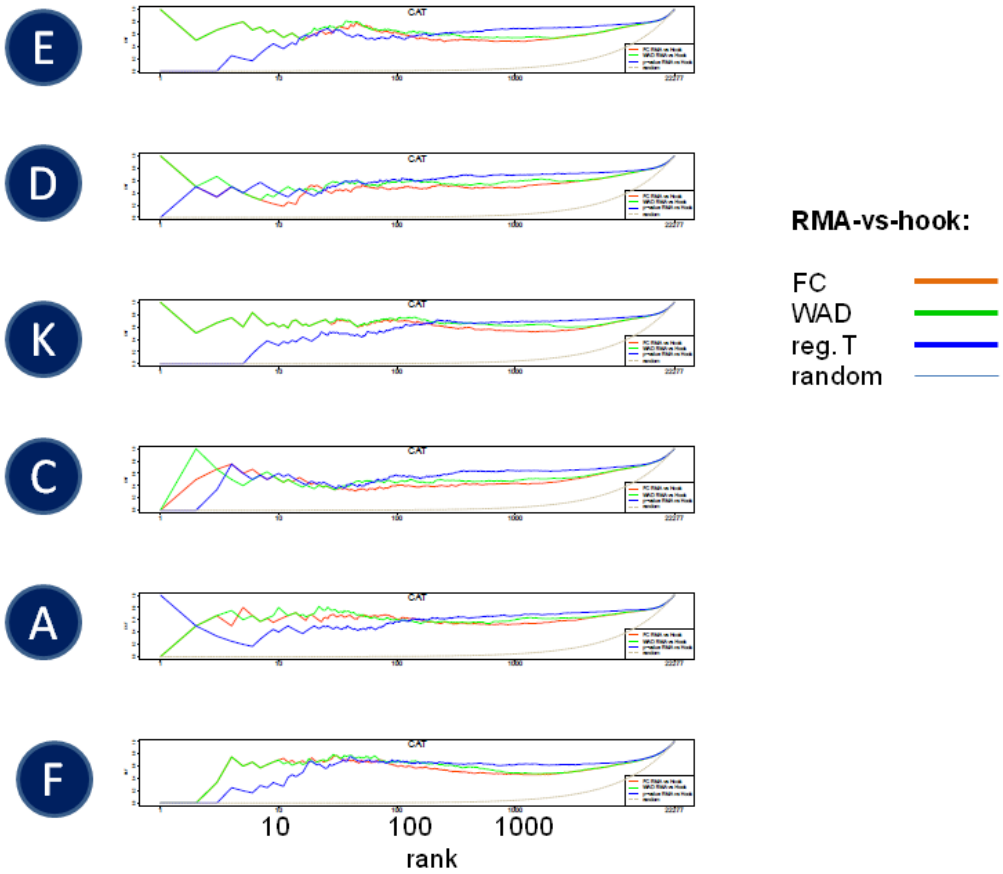

Figure S 12: Comparison of selected total list of genes obtained by hook and RMA using the regularized t-score, the WAD-score or the FC-score. Lists are compared using the correspondence-at-the-top (CAT) plots showing the fraction of overlap genes in pairs of lists as a function of the gene rank. The letters refer to samples which are chosen for spot comparisons in Figure S 11.

Table S 1: Comparison of spot enrichment as seen by hook and RMA

Spots[a]	Hook		RMA	
	#M/#G[b]	top 5 gene sets (BP)[c]	#M/#G[b]	top 5 gene sets (BP)[c]
A	4/239	protein transport (-9), stress−activated MAPK cascade (-9), GTP catabolic process (-8), small GTPase mediated signal transduction (-8), toll−like receptor 3 signaling pathway (-8), *intracellular protein transport (-5)*	4/253	RNA splicing (-7), regulation of protein localization (-6), intracellular protein transport (-6), GTP catabolic process (-5), protein destabilization (-5), *small GTPase mediated signal (-4)*
G	6/260	regulation of G−protein coupled receptor protein signaling pathway (-7), negative regulation of cell death (-6), cell adhesion (-6), cell fate commitment (-5), negative regulation of neuron differentiation (-5)	9/408	negative regulation of neuron differentiation (-6), cell adhesion (-6), ion transport (-5), transmembrane transport (-5), *negative regulation of cell death (-4)*
C	7/336	RNA splicing (-5), G2/M transition of mitotic cell cycle (-4), DNA recombination (-4), mRNA processing (-4), nuclear mRNA splicing, via spliceosome (-4), *RNA processing (-3)*	5/296	RNA splicing (-9), nuclear mRNA splicing, via spliceosome (-7), RNA processing (-7), transcription elongation from RNA polymerase I promoter (-6), mRNA processing (-6), *G2/M transition of mitotic cell cycle (-3)*
D	34/1043	cell adhesion (-16), immune response (-16), inflammatory response (-16), chemotaxis (-16), signal transduction (-13)	38/1151	cytokine−mediated signaling pathway (-16), immune response (-16), inflammatory response (-16), type I interferon−mediated signaling pathway (-15), cell adhesion (-14), *signal transduction (-11)*
E	15/673	cell division (-16), M phase of mitotic cell cycle (-16), mitotic cell cycle (-16), mitosis (-16), mitotic prometaphase (-16), *DNA replication (-12)*	9/405	M phase of mitotic cell cycle (-16), cell division (-16), mitotic cell cycle (-16), mitotic prometaphase (-16), DNA replication (-12), *mitosis (-12)*
F	2/102	anterior/posterior pattern formation (-14), proximal/distal pattern (-11), formation(-5), embryonic skeletal system morphogenesis (-7), embryonic limb morphogenesis (-7), regulation of transcription, **DNA−dependent (**-7)	3/210	anterior/posterior pattern formation (-10), M phase of mitotic cell cycle (-10), proximal/distal pattern formation (-10), mitotic prometaphase (-10), regulation of transcription, **DNA−dependent (**-9), *embryonic skeletal system morphogenesis (-4)*
J	4/221	water transport (-7), neurotransmitter secretion (-6) axonogenesis (-5)		
K	1/183	synaptic transmission (-12), neurotransmitter secretion (-8), central nervous system development (-6), myelination (-6), glutamate secretion (-6)	1/176	synaptic transmission (-13), neurotransmitter secretion (-8), central nervous system development (-6), myelination (-6), glutamate secretion (-6)

H	7/256	associative learning (-5), negative regulation of gene−specific transcription from RNA polymerase II promoter (-5), regulation of transcription from RNA polymerase II promoter by nuclear hormone (-5), steroid hormone mediated signaling pathway (-4), regulation of transcription, DNA−dependen (-4)	8/263	embryonic forelimb morphogenesis (-5), calcium−independent cell−cell adhesion (-4), induction of positive chemotaxis (-4), acute−phase response (-4), nucleosome assembly (-4)
I		response to light stimulus (-4), muscle contraction (-3), androgen metabolic process (-3)		
M	3/113	regulation of transcription, DNA−dependent (-5), cellular response to lithium ion (-5), positive regulation of transcription from RNA polymerase II promoter (-5)		

[a] for spot assignments see Figure S 10

[b] number of metagenes/number of genes ratio of the respective spot

[c] top-five enriched gene sets of each spot taken from the category biological process (BP); overlap sets are shown in bold letters; italic letters indicate highly ranked non-overlap sets among the top-five.

2 Results

2.1 Pairwise correlation maps

We generated pairwise correlation maps (PCM) which visualize the Pearson correlation coefficients between the metagene expression landscapes in all pairwise combinations of sample portraits. Maroon and red colored tiles assign strong correlations and thus pairwise combinations of similar portraits and blue colored tile anticorrelated portraits where usually overexpressed regions have switched into underexpressed ones. The samples of the same tumor subtype are grouped together to visualize the intra- and inter-class similarity of the samples (see color bars along the edges of the map). The covariance structure of the data is visualized using the maximum spanning tree (MST) and the correlation net (CN) representations shown in the main paper.
The PCM of GBM for example clearly shows that the expression landscapes of the MES and PN subtypes are strongly anticorrelated whereas that of MES and CL are partly correlated. Anticorrelated portraits are also observed for MET and BHP on one hand and PIN and PCA on the other hand for PCP.

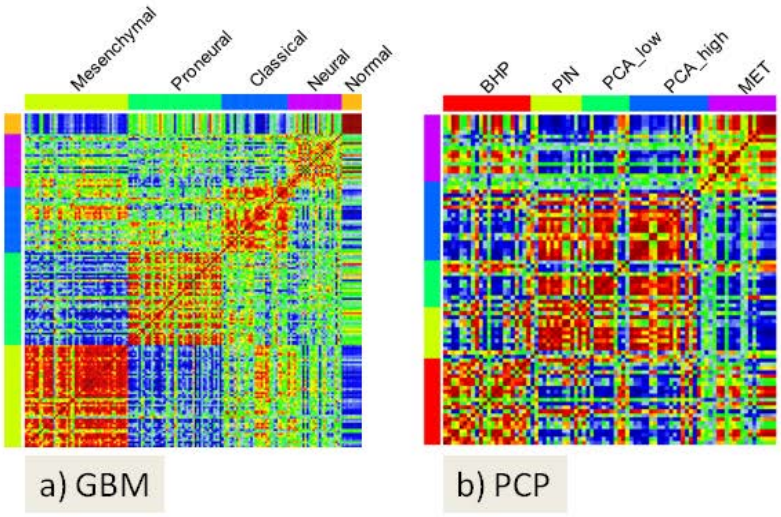

Figure S 13: Pairwise correlation maps visualizing the Pearsson correlation coefficient of all pairwise combinations of sample portraits. The samples are grouped according to their class membership (see the color bars along the edges of the map). Each subtype is characterized by a more or less pronounced brown-to-red square along the diagonal line which reflects self-similarity of samples of the same type. Off-diagonal brown and blue regions refer to correlated and anti-correlated SOM-spot pattern. For example, GBM-MES samples are predominantly anticorrelated with GBM-PN samples but partly correlated with GBM-CL and GBM-NL samples (see also the respective anticorrelated or correlated spot pattern of the mean portraits per class).

2.2 Underexpression spot characteristics

We analyzed the underexpression spots visualized as blue regions expression in the SOM-portraits using the same characteristics which were applied to the overexpression spots. Figure S 14 - Figure S 15 show the underexpression spot characteristics of GBM and PCP, respectively, using the underexpression summary map, the spot expression heatmap and the spot abundance plot. The respective overexpression characteristics are given in the main paper.
Position and size of most of the detected underexpression spots agree with the position and size of one of the overexpression spots. We use the respective lower case letters for assignment of the underexpression spots to express this pairwise correspondence, e.g. underexpression spot 'a' roughly corresponds to overexpression spot 'A'. The detection of an over- and underexpression spot at the same position in the SOM-portraits of different samples simply reflects marked oscillations of the expression amplitude of the respective metagenes: for example, spot K is overexpressed in PN- and spot k, at nearly the same position, is underexpressed in MES-samples. In some cases an overexpression spot splits into two or three underexpression spots or vice versa due to subtle differences of the local expression patterns affecting spot selection. We use the annotations N→n1, n2,.. and F+C→(fc), respectively.

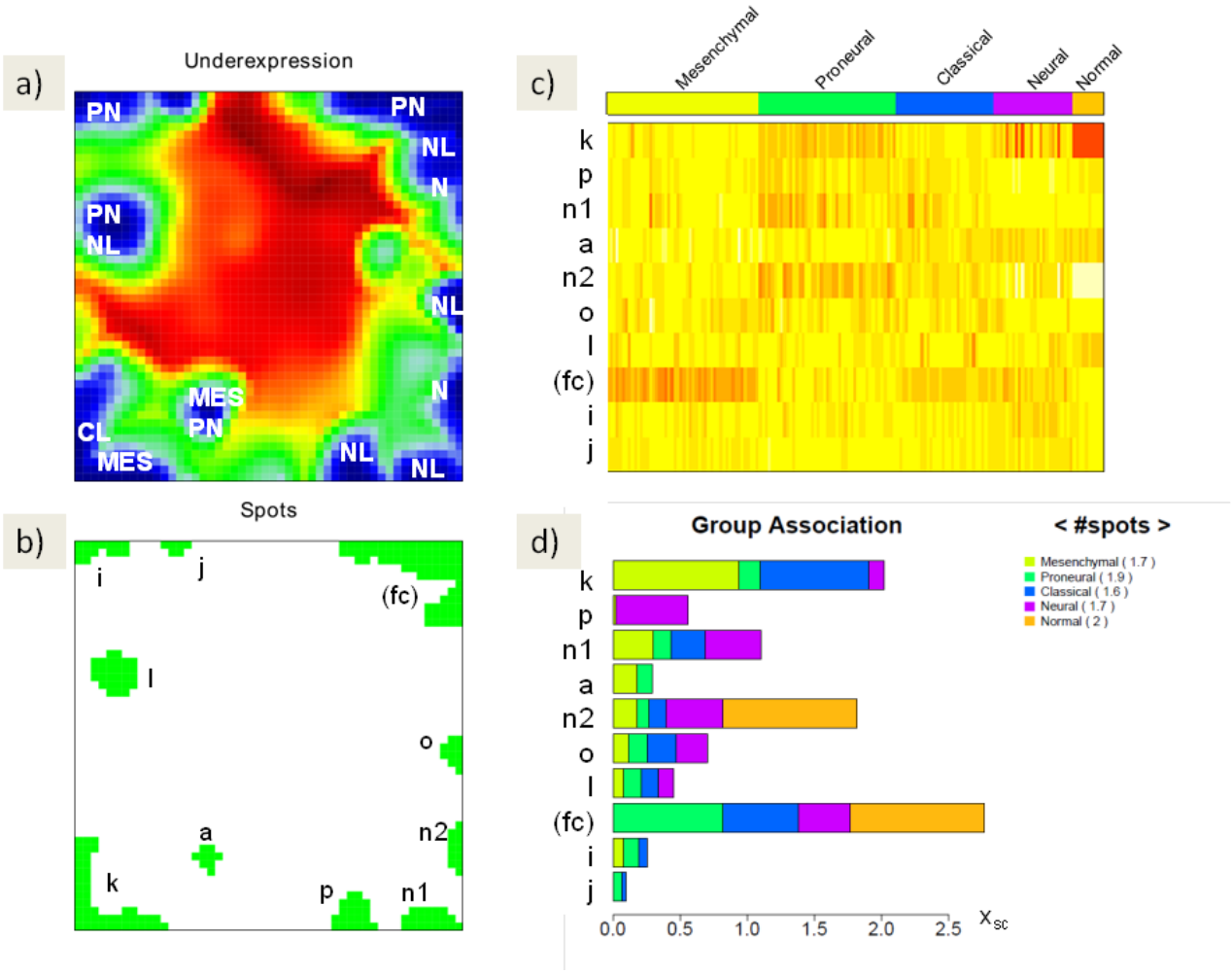

Figure S 14: Underexpression spot characteristics of GBM. Underexpression spots represent regions below the 2-percentile threshold of expression values. The spots are assigned by lower case letters. We use the same letter as used for assigning the overexpression spots (e.g. underexpression spot 'a' is located in the same region of the map as overexpression spot 'A'). If the region of one of these overexpression spots splits into two or more underexpression spots a number was added to the letter (e.g. d1, d2). In the opposite situation, i.e., if more than one overexpression spots merge into one underexpression spot we merge the respective letters (e.g. spots E, K and J merge into underexpression spot 'ekj'.

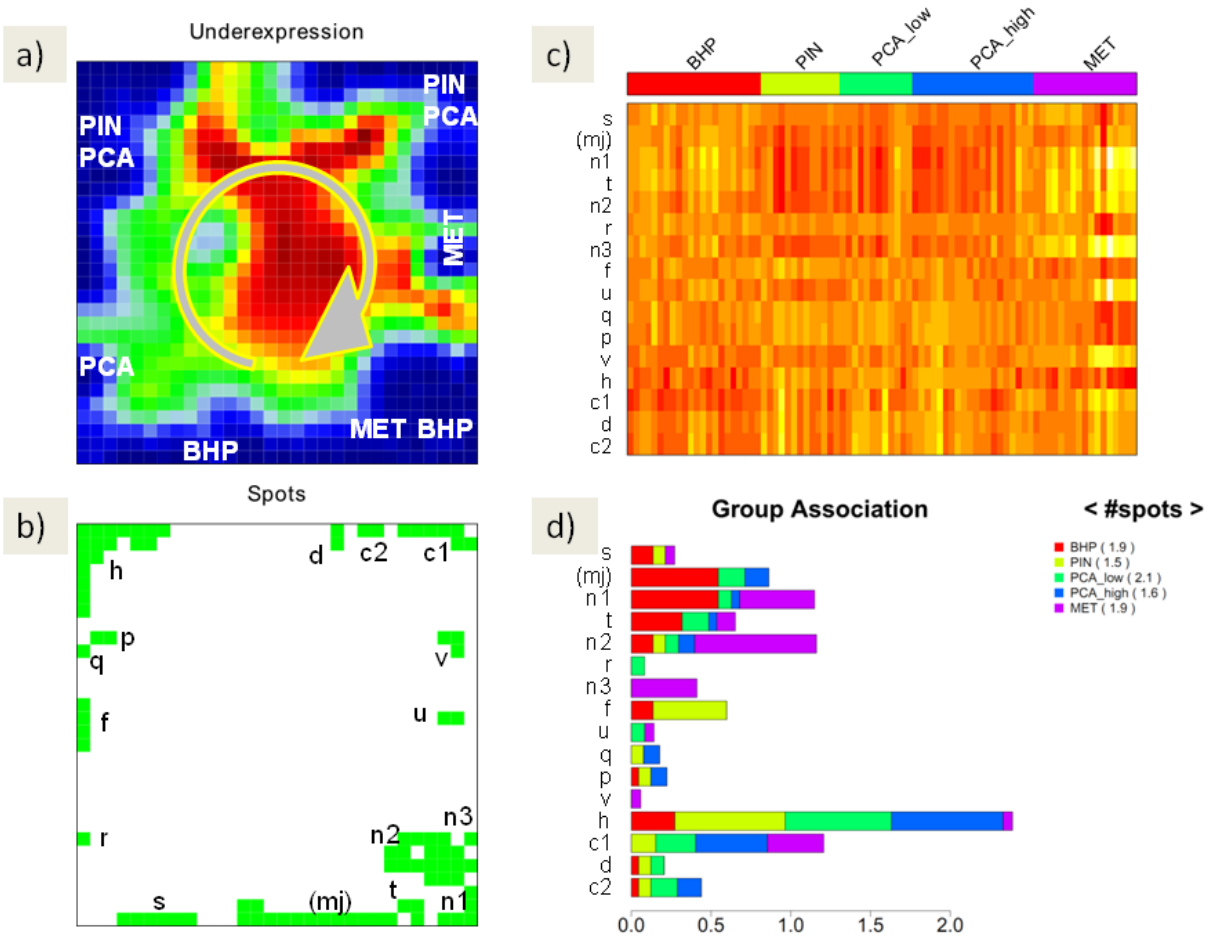

Figure S 15: Underexpression spot characteristics of PCP. See legend Figure S 14 for details. The arrow illustrates the appearance of spots with progressing cancer.

2.3 Spot correlations

To explore similarity relations between the spot patterns of the samples we calculate the 'spot expression matrix'. In this matrix, the spot expression state of each sample is characterized by one column containing the mean logged expression values of each spot averaged over all metagenes of this spot. Each row then provides the expression profile of one particular spot in all samples. The spot tree is calculated as the respective maximum spanning tree (MST, see the main paper) connecting spots of strongest correlation in all pairwise combinations of their spot profiles (i.e. the row-vectors of the spot matrix). The spot tree consequently characterizes similarities between the spots in contrast to the sample-similarity MST which characterize similarities between the samples. The nodes of the spot-MST are complemented with pie charts illustrating the fraction of the sample classes expressing this particular spot.
We calculated MSTs separately for over- and underexpression spots based on the spot-spot correlation matrix. They visualize diagonal clusters of correlated (marroon and red) and off-diagonal anti-correlated spot combinations in GBM (Figure S 16) and PCP (Figure S 17). The respective spot MSTs divide into closely located spots along the tree showing concerted expression changes and more distant anticorrelated spots. Over- and underexpression spot trees reveal partly similar structures reflecting antagonistic swicthing behavior of the associated genes: For example, overexpression spots K and F+C are anticorrelated in GBM. The same relation is observed for the underexpression spots k and (fc).

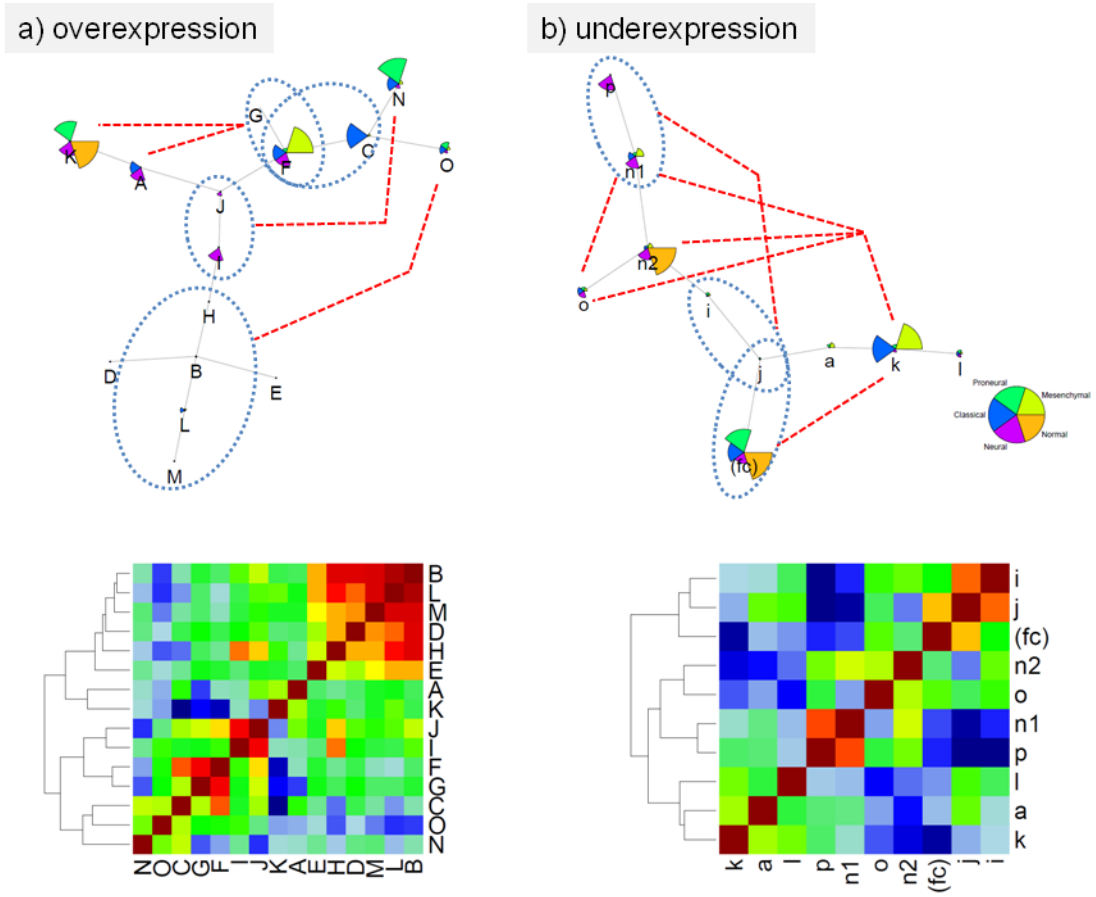

Figure S 16: Spot correlation analysis of GBM: The part above shows the maximum spanning trees of over- and underexpression spots. Spots are assigned by the same letters as in the main paper. The spot abundances of the subtypes are illustrated using pie diagrams: The radius of the segments scales with the percentage of spots per class. Strong positive correlations between spots are illustrated by dotted ellipses including the respective spots. Red dotted lines connect strongly anticorrelated spots. The part below shows the respective pairwise correlation maps of the spots. Strong positive and negative correlations are colored in maroon/red and dark/light blue, respectively.

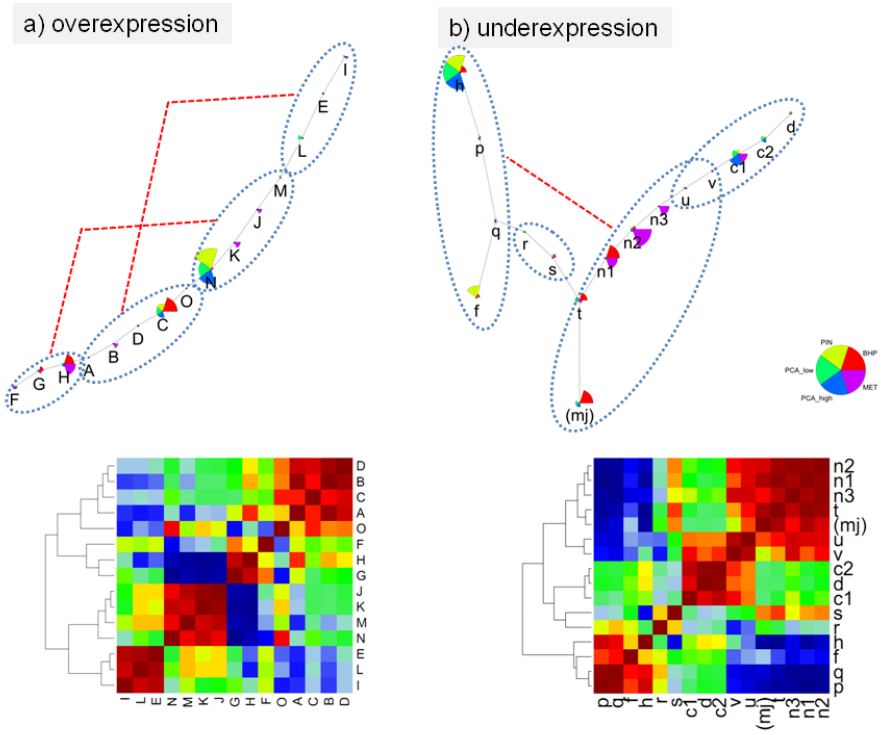

Figure S 17: Spot correlation analysis of PCP. See legend of Figure S 16.

2.4 Global underexpression characteristics

We analyzed the landscapes in the range of low expression values by calculating the number distribution and the shape of the underexpression spots. For PCP the obtained underexpression characteristics are almost symmetric showing similar properties compared with the respective overexpression landscapes (see the main paper): For example, over- and underexpression spots are both either more (e.g. in the BHP-subtype) or less (e.g. in MET-samples) compact. Note that the number distribution of underexpression spots is slightly shifted towards smaller values due to the, on average, smaller number of underexpression spots observed.
In contrast, for the GBM-subtypes one finds a partly anti-symmetric behavior of the respective parameters: Whereas the number distribution of the underexpression spots of the PN-, NL- and CL-subtypes is narrower compared with that of the MES- and PN-subtypes the relation reverses for the overexpression spots. The narrow shape of the underexpression spots is paralleled by a more compact shape of the spots especially in logFC scale. Hence, the subtypes with a more fuzzy overexpression landscape seem to show more sharp underexpression patterns. This asymmetry can be explained by the fact that the metagenes expression distributes differently between over- and underexpression in the different subtypes. Dominating overexpression is associated with the more compact spots in the MES-images whereas dominating underexpression gives rise to more compact underexpression spots in CL- and NL-subtypes.

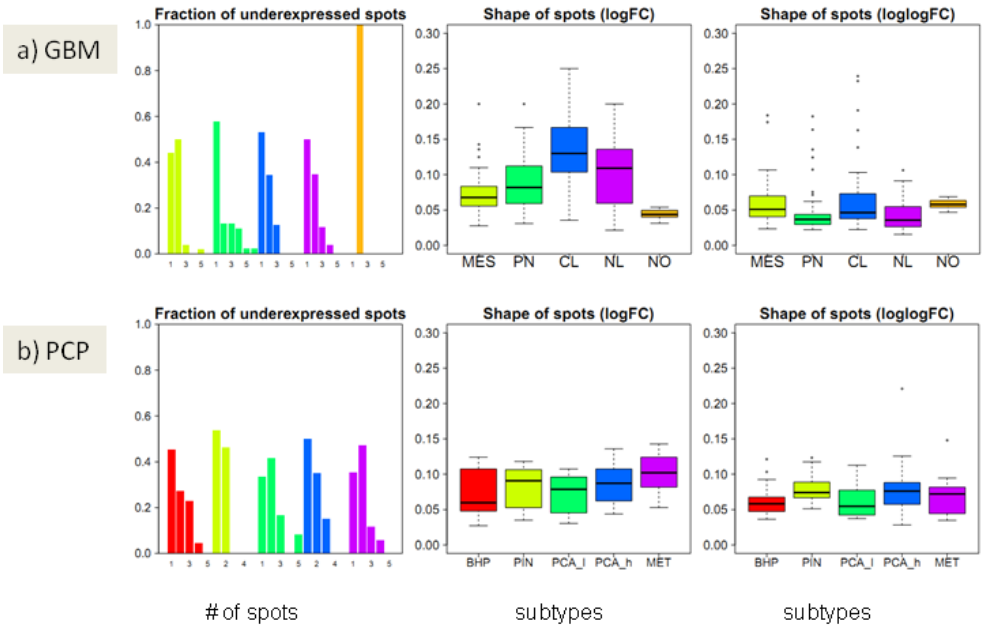

Figure S 18: Fractional distribution of the numbers of underexpression spots, and the underexpression spot shape in log FC and log logFC scales (from the left to the right).

2.5 Spot enrichment analysis: Tables of enriched gene sets in GBM and PCP

The tables below collect enriched gene sets in different spots of the GBM and PCP SOM. Spots are assigned according to our overexpression (capital letters) and underexpression (lower case letters) signature.

Table S 2: Top-enriched gene sets in spots detected in GBM

Spot[a]	Short name	up[b]	down[b]	GO and pathway sets [c]	Tissue and literature sets [d]
G	wound healing	MES	PN	Plasma membrane (CC,-6); wound healing (BP,-5); fibronectin binding (-5)	KOBAYASHI_EGFR_SIGNALING_6HR_DN (LS,-10); Wu_Cell-migration (LS,-7); Lee_Mestasis (LS, -7), Boyault_Liver-Cancer (LS, -6); Charafe_Breast-Cancer (LS,-6); cultured_astroglia (CS,-6)
F	Inflammation	MES, NL,CL	PN	Extracellular_space (CC,-18); inflammatory response (BP,-16); chemotaxis (BP,-15); cytokine_pathway (BP,-14)	Mucosa (TS,-13); cultured_astroglia (CS,-15); Verhaak_MES (GBM,-9); Hummel_BL-DN (LS, -8); Farmer_Breast-Cancer (LS,-11)
E	neurotransmitter	MES		neurotransmitter:sodium symporter activity (MF,-5); neurotransmitter transport (BP,-4); acetylcholine receptor activity (BP,-4)	MILICIC_FAMILIAL_ADENOMATOUS_POLYPOSIS_UP (LS,-4)
N	Cell division	PN	NOR, NL	Cell_cycle (BP, 0); cell_division (BP,0); mitosis (BP, 0); nucleus (CC, 0) ; chromosome (CC,-12); DNA_binding (MF,-14);	Developing_astrocytes (CS,0); Farmer_breast-cancer (LS,0); Liang_silenced_by_methylation (LS,0); OPC (CS,-6)
p	chromatin	PN	NL	Nucleus (CC,-14); nucleic_acid_binding (MF,-13); chromatin_modification (BP,-13); DNA_binding (MF,-10); histone_h3_acetylation (BP,-7)	REACTOME_PECAM1_INTERACTIONS (RE,-5); NIKOLSKY_BREAST_CANCER_12Q24_AMPLICON (LS,-4); BIOCARTA_PITX2_PATHWAY (BC,-4); Globus pallidus (TS,-4)
n1	Cell division	PN	NL	Chromosome (CC,0); nucleus (CC,0); cell_division (BP,-16); mitosis (BP, -16)	SEMBA_FHIT_TARGETS_DN (LS,0); FINETTI_BREAST_CANCER_KINOME_RED (LS,0); LIANG_SILENCED_BY_METHYLATION_DN (LS,0); LY_AGING_MIDDLE_DN (LS,0); developing astrocytes (CS,-15)
n2	RNA splicing	PN	NOR	RNA splicing (BP,-7); midbrain development (BP,-5); RNA processing (BP,-5); ribonucleoprotein complex (CC,-5)	SPIRA_SMOKERS_LUNG_CANCER_DN (LS,-7); LIU_COMMON_CANCER_GENES (LS,-4)
C	Angiogenesis	CL, MES	NOR, NL	Angiogenesis (BP,-6); blood vessel morphogenesis (BP,-7); basement_membrane(CC,-7); extracellular_matrix_binding (MF, -5)	Verhaak_CL (GBM,-11); TCGA_GLIOBLASTOMA_MUTATED (LS,-6); CHEN_HOXA5_TARGETS_6HR_DN (LS,-5); LEE_LIVER_CANCER_HEPATOBLAST (LS,-5)
O	Innate immunity	CL,NL, PN	NL	Protein_binding (MF,-12); nucleus (MF,-12); nucleotide_binding (MF,-10); cytosol (CC,-8); HCMV_pathway (BC,-7); downstream_signal_cascade (RE,-6); stress_activated_MAPK_cascade (BC,-7); Toll signaling pathway (BP-5);	B-cells (TS,-6); DING_LUNG_CANCER_EXPRESSION_BY_COPY_NUMBER (LS,-5); UZONYI_RESPONSE_TO_LEUKOTRIENE_AND_THROMBIN (LS,-5)
I	Mitochondrion/ translation	NL	PN	Mitochondrion (CC,-14); ribosome (CC,-10); translation (BP,-8); C−terminal protein lipidation (BP,-6); PACKAGING _OF_TELOMERE_ENDS (RE,-5)	OUELLET_CULTURED_OVARIAN_CANCER_INVASIVE_VS_LMP_UP (LS; -6); BARRIER_CANCER_RELAPSE_TUMOR_SAMPLE_UP (LS,-5); MOOTHA_TCA (LS,-5)
J	Mitochondrion	NL		Mitochondrion (CC,-5); mitochondrial_small_ribosome_unit (CC,-5); CELL_DEATH_SIGNALLING_VIA_NRAGE_NRIF_AND_NAD (RE,-4); nucleotide metabolic process (BP,-4)	SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN (LS,-4); Verhaak_NL (GBM,-5); DODD_NASOPHARYNGEAL_CARCINOMA_UP (LS,-4);
H	Cell-cell adhesion	NL		heterophilic cell−cell adhesion (BP,-5); TIGHT_JUNCTION (KG,-5); extrinsic to internal side of plasma membrane (CC,-4)	CROMER_TUMORIGENESIS_DN (LS,-5);

B	xenobiotics	NL		RETINOL_METABOLISM (KG,-5); CARM1_PATHWAY (BC,-5); XENOBIOTICS (RE,-5); positive regulation of stress fiber assembly (BP,-5); promoter binding (MF,-5); DRUG_METABOLISM_OTHER_ENZYMES (KG,-5); transcription factor complex (CC,-4)	BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_DN (LS,-5)
A	response to axon injury	NL, CL		Cilium_axoneme (CC,-7); negative_regulation_of _cell_death (BP,-6); response_to_axon_injury (-6,BP); fibroblast growth factor receptor signaling pathway (BP,-5)	Astrocytes_glio (CS,-16); Lu_aging_brain-DN (LS, -8);
L	Transcription	NL		DNA_dependent_transcription (BP,-5); sequence−specific enhancer binding RNA polymerase II transcription factor (MF,-5);	NIKOLSKY_BREAST_CANCER_21Q22_AMPLICON (LS,-5); Turjanski_Mapk8+9_targets (LS,-4)
M	development	NL		pancreas development (BP,-5); skin development (BP,-5); androgen metabolic process (BP,-5);	
I	Transcription	NL		DNA_dependent_transcription (BP,-5)	Turjanski_Mapk8+9_targets (LS,-4)
K	synapse	NOR, PN	NL, MES	Synaptic_transmission (BP,-16); synapse (CC,-12); nervous_system_development (BP,-11); Glutamate (RE, -7); Serotonin_Neurotransmitter (RE,-6); Dopamin_Neurotransmitter (RE,-6)	Azgharzadeh_Neuroblastoma (LS,-7); Nakayama_soft-tissue-tumor (LS,-6); Nervous_system (TS,-6); in_vivo_astrocytes (CS,-8); neurons_glio (CS,-6)

[a] sets are assigned using the letter-nomenclature introduced in the main paper

[b] each spots is shortly named according to a biological context derived from the enriched gene sets

[c] Cancer subtypes showing up- or downregulation of the respective spot.

[d] Top enriched gene sets from the categories biological process (BP), molecular function (MF), cellular component (CC), Reactome (RE), BioCarta (BC), KEGG (KG). Enrichment is estimated using the p-value of the right-tail Fishers exact test based on the hypergeometrical distribution. The table lists the name of the gene set and the set category and the log p-value in the brackets.

[e] Top enriched genesets from the categories 'literature sets' (LS), cell systems (CS), tissue sets (TS)

Table S 3: Top-enriched genesets in spots detected in PCP (see legend of Table S 2 for assignments)

Spot[a]	Short name	up[b]	down[b]	GO and pathway sets [c]	Tissue and disease sets [d]
C	antigen processing	BHP	MET	antigen processing and presentation (BP,-9), response to progesterone stimulus (BP,-7); cytokine−mediated signaling pathway (BP,-7); extracellular region (CC,-10); insulin−like growth factor binding (MF,-6); integrin binding (MF,-5); MHC class II protein complex (CC,-5); SMOOTH_MUSCLE_CONTRACTION (RE,-7)	CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_UP (LS,-6); cultured astroglia vs. in vivo astrocytes (CS,-4); Sec. lymphoid organs (TS,-4); HUMMEL_BURKITTS_LYMPHOMA_DN (LS, -5);
G	proliferation	BHP, MET		negative regulation of epithelial cell proliferation (BP,-4)	FIRESTEIN_PROLIFERATION (LS,-4);
M	chromatin remodeling	PIN	BHP, MET	chromatin assembly or disassembly (BP,-4); chromatin remodeling (BP,-4); mRNA transport (BP,-3); condensed nuclear chromosome (CC,-4); nuclear inner membrane (CC,-4);	ZHAN_V1_LATE_DIFFERENTIATION_GENES_DN (LS,-5); LINDGREN_BLADDER_CANCER_CLUSTER_2A_DN (LS,-5); BARIS_THYROID_CANCER_DN (LS,-5); SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP (LS,-4)
N	ribosome	PIN, PCA_low	MET	structural constituent of ribosome (MF,-5), endoplasmic reticulum (CC,-5); cellular protein metabolic process (BP,-8); translational termination (BP,-6); post−translational protein modification (BP,-5); SIGNALING_BY_EGFR (RE,-4)	TOMLINS_METASTASIS_DN (LS,-6); TOMLINS_PROSTATE_CANCER_UP (LS,-4); DAVICIONI_RHABDOMYOSARCOMA_PAX_FOXO1_FUSION_UP (LS,-5); AMIT_EGF_RESPONSE_60_MCF10A (LS,-4); androgen signaling (LS,-9)
L	cell adhesion	PCA_low		homophilic cell adhesion (BP,-4); response to insulin stimulus (BP,-3); visual learning (BP,-3); caveola (CC,-3); actin filament (CC,-3);	OKAWA_NEUROBLASTOMA_1P36_31_DELETION (LS,-16); WHITE_NEUROBLASTOMA_WITH_1P36.3_DELETION (LS,-5); SPIELMAN_LYMPHOBLAST_EUROPEAN_VS_ASIAN_2FC_UP (LS,-4); Muscle (TS,-4); CHUNG_BLISTER_CYTOTOXICITY_DN (LS,-4)
A		PCA_high		intracellular protein kinase cascade (BP,-3); negative regulation of cell cycle (BP,-3); induction of apoptosis by extracellular signals (BP,-3); hydrolase activity, acting on ester bonds (MF,-3); insulin−like growth factor binding (MF,-3)	YE_METASTATIC_LIVER_CANCER (LS,-5); SMID_BREAST_CANCER_RELAPSE_IN_LIVER_DN (LS,-5); EHLERS_ANEUPLOIDY_DN (LS,-5); HUMMEL_BURKITTS_LYMPHOMA_DN (LS,-4); myc (LS,-4);
O	cell cycle	PCA_high	PCA_low, MET	CELL_CYCLE (KG,-7); protein targeting (BP,-5); transcription elongation from RNA polymerase II promoter (BP,-4); endoplasmic reticulum unfolded protein response (BP,-4); general RNA polymerase II transcription factor activity (MF,-6); PDZ domain binding (MF,-5)	REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES (RE,-16); MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_16 (LS,-8); HSIAO_HOUSEKEEPING_GENES (LS,-7); LEE_LIVER_CANCER_SURVIVAL_DN (LS,-7)
F	platelet activation	MET, PCA_high	BHP, PIN	platelet activation (BP,-6), extracellular matrix (CC,-8); extracellular matrix structural constituent (MF,-7); blood vessel remodeling (BP,-5); calcium ion binding (MF,-6); INTEGRIN_CELL_SURFACE_INTERACTIONS ((RE, -7)	LEE_NEURAL_CREST_STEM_CELL_UP (LS,-5); FARMER_BREAST_CANCER_CLUSTER_5 (LS,-5)
H	RNAP II activity	MET, BHP	PIN	cholesterol homeostasis (BP,-4); specific RNA polymerase II transcription factor activity (MF,-5); HISTIDINE_METABOLISM (KG,-5)	MISHRA_CARCINOMA_ASSOCIATED_FIBROBLAST_UP (LS,-5); MULLIGHAN_MLL_SIGNATURE_1_DN (LS,-4)
B	nucleosome	MET		nucleosome assembly (BP,-5); nucleosome (CC,-7)	PENG_GLUTAMINE_DEPRIVATION_UP (LS,-16); BIOCARTA_NO2IL12_PATHWAY (BC,-16); GRAHAM_CML_DIVIDING_VS_NORMAL_DIVIDING_UP (LS,-7); GNATENKO_PLATELET_SIGNATURE (LS,-7)
D	development	MET		post−embryonic development (BP,-4); pyridoxal phosphate binding (MF,-4);	CHIANG_LIVER_CANCER_SUBCLASS_INTERFERON_UP (LS,-5)

E	mitochon-drion outer membrane	MET		regulation of insulin secretion (BP,-5); mitochondrial outer membrane (CC,-4);	CYTOTOXIC_PATHWAY (BC,-16); THELPER_PATHWAY (BC,-16); SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN (LS,-5); Muscle (TS,-4)
I	response to cyclic com-pound	MET		cellular response to organic cyclic com-pound (BP,-5); endoplasmic reticulum lumen (CC, -5)	KEGG_FOLATE_BIOSYNTHESIS (KG,-7); MARTINELLI_IMMATURE_NEUTROPHIL_DN (LS,-6); Ben−Porath_UP (LS,-6)
J	mitochon-drion	MET		sphingolipid metabolic process (BP,-5); ATP hydrolysis coupled proton transport (BP,-4); cytochrome−c oxidase activity (MF,-4); mitochondrion (CC,-4)	GAL_LEUKEMIC_STEM_CELL_DN (LS,-6); WU_CELL_MIGRATION (LS,-5); WALLACE_PROSTATE_CANCER_UP (LS,-4);
K	respiratory chain	MET		extracellular matrix organization (BP,-4); respiratory chain (CC,-5);	HADDAD_T_LYMPHOCYTE_AND_NK_PROGENITOR_UP (LS,-5)

2.6 Gene sets in concert with inflammation

Below we show galleries illustrating the behavior of selected gene sets using the respective GSZ-profiles and gene set population maps. All gene sets selected in this subsection mostly accumulate in the right upper corner of the map for GBM and strongly upregulate in the MES- and downregulate in the PN-subtypes. Note however that slightly different accumulation patterns for, e.g. 'chemokine activity' and 'angionesis' gives rise to different profiles for the CL (blue), NL (pink) and NOR (ocher) samples. These gene sets also enrich in the PCP-maps however in a slightly less localized manner.

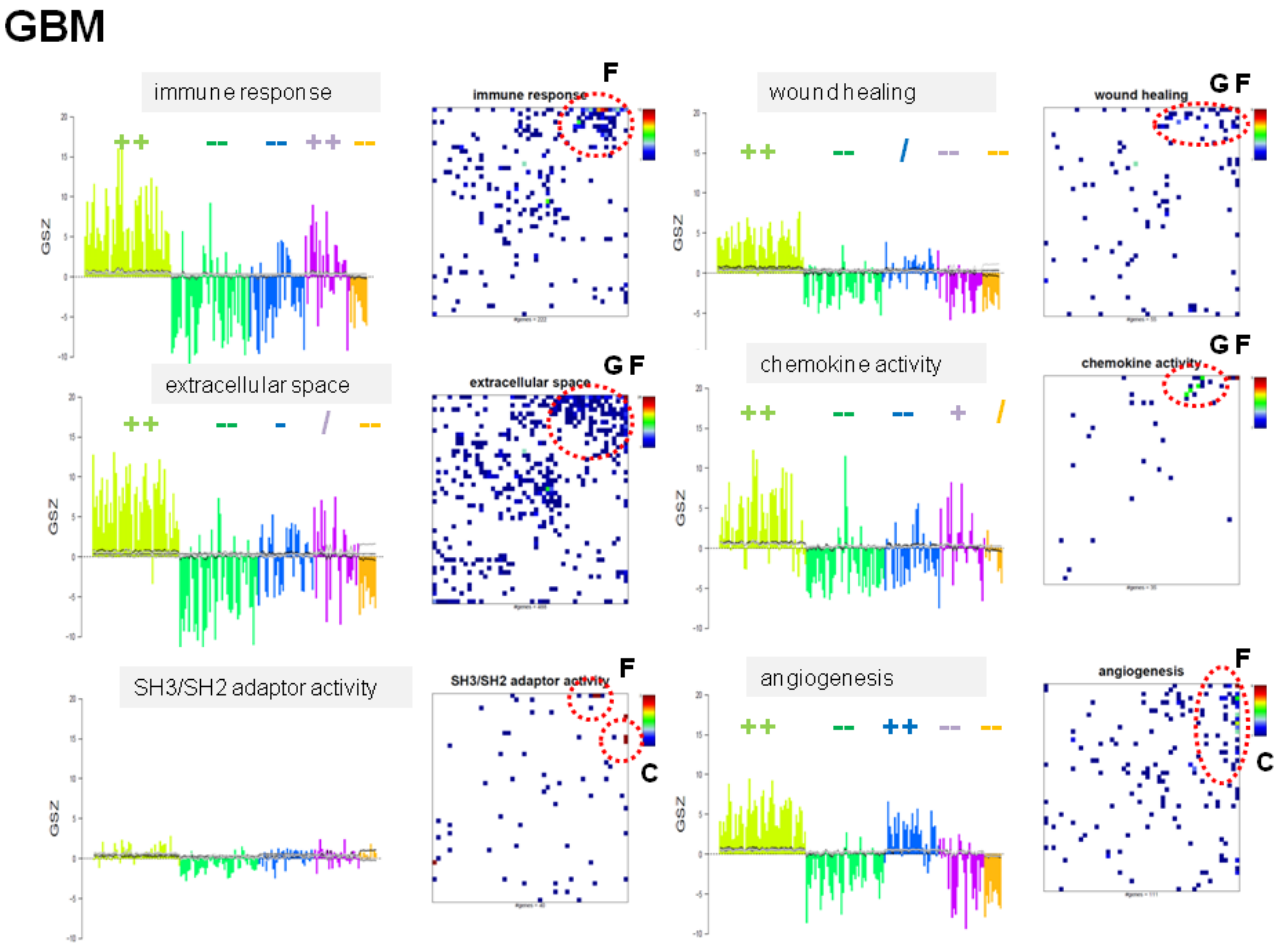

Figure S 19: Selected profiles and population maps of gene sets 'in concert with inflammatory response' in GBM. Regions of overrepresentation in the maps are indicated by red-dotted ellipses. The letters refer to the respective overexpression spots.

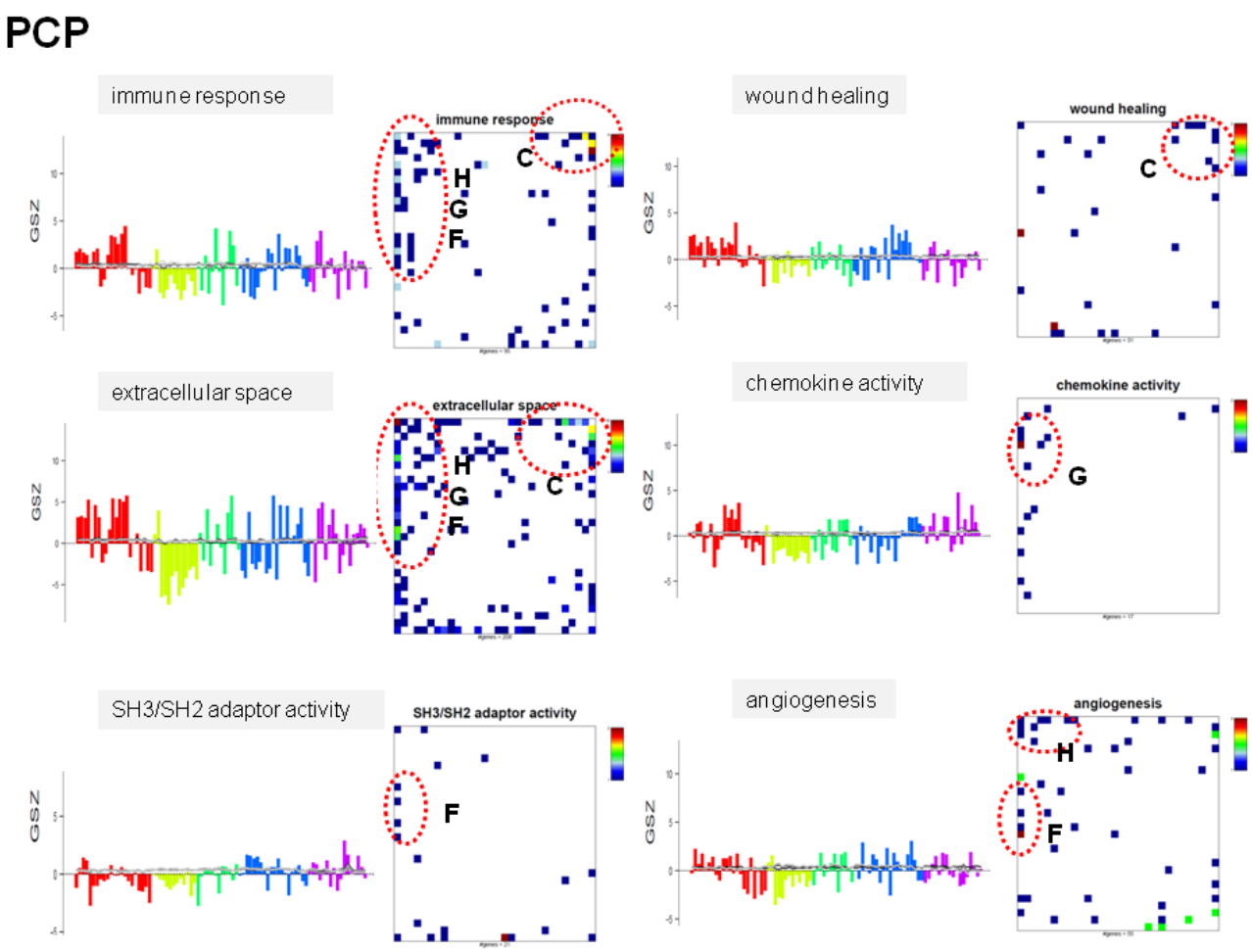

Figure S 20: Selected profiles and population maps of gene sets ‘in concert with inflammatory response’ in PCP. Regions of overrepresentation in the maps are indicated by red-dotted ellipses. The letters refer to the respective overexpression spots. Note that we chose the same sets as in Figure S 19 in GBM.

2.7 Gene sets in concert with cell division

Gene sets which change in concert with cell division act partly antagonistic in the PN and MEs subtypes compared with the gene sets changing in concert with inflammation analyzed in the previous subsection. Also here, modifications of the distributions of the genes in the map give rise to different profiles for the CL, NL and NOR samples.

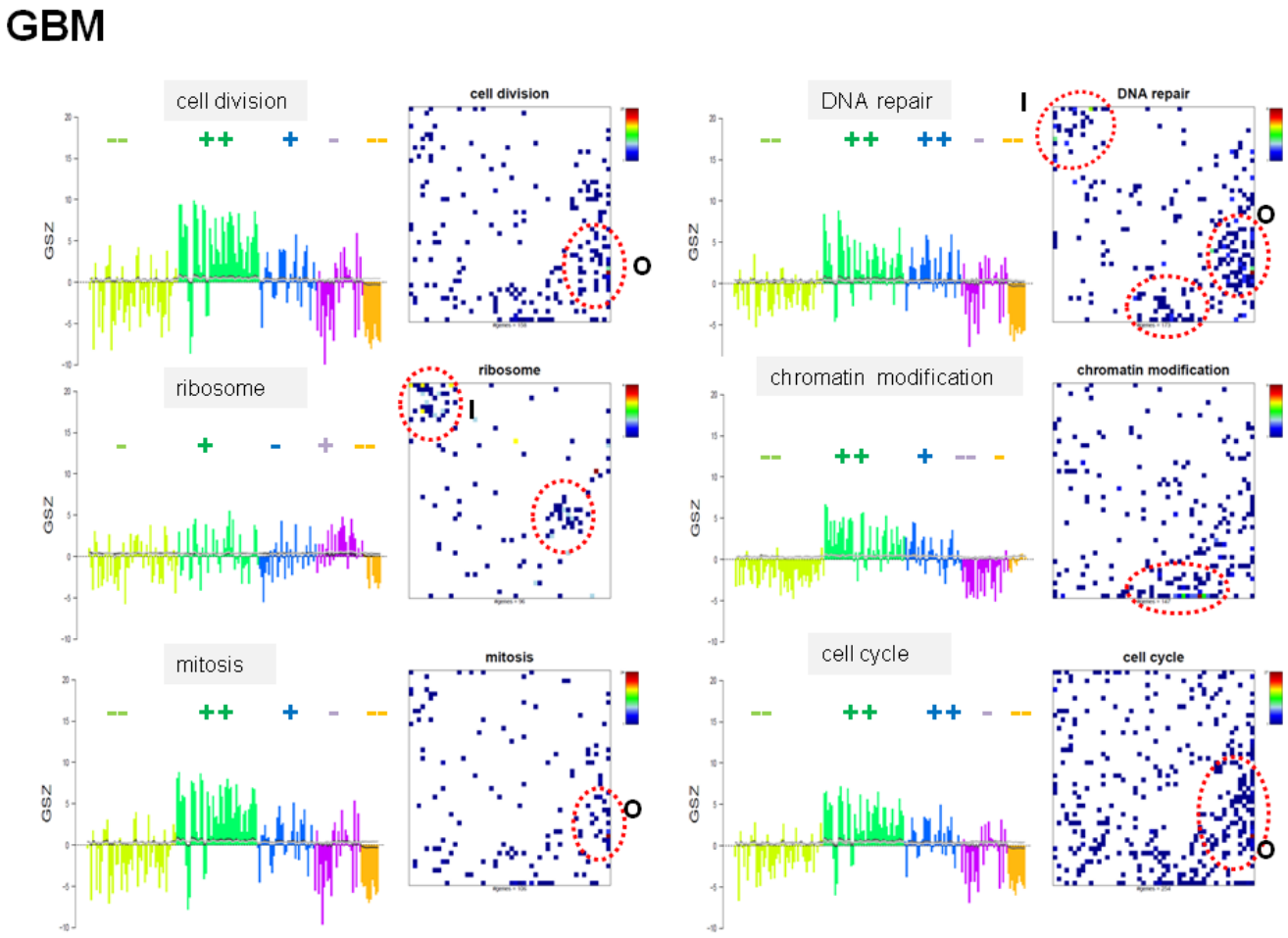

Figure S 21: Selected profiles and population maps of gene sets 'in concert with cell division' in GBM. Regions of overrepresentation in the maps are indicated by red-dotted ellipses. The letters refer to the respective overexpression spots.

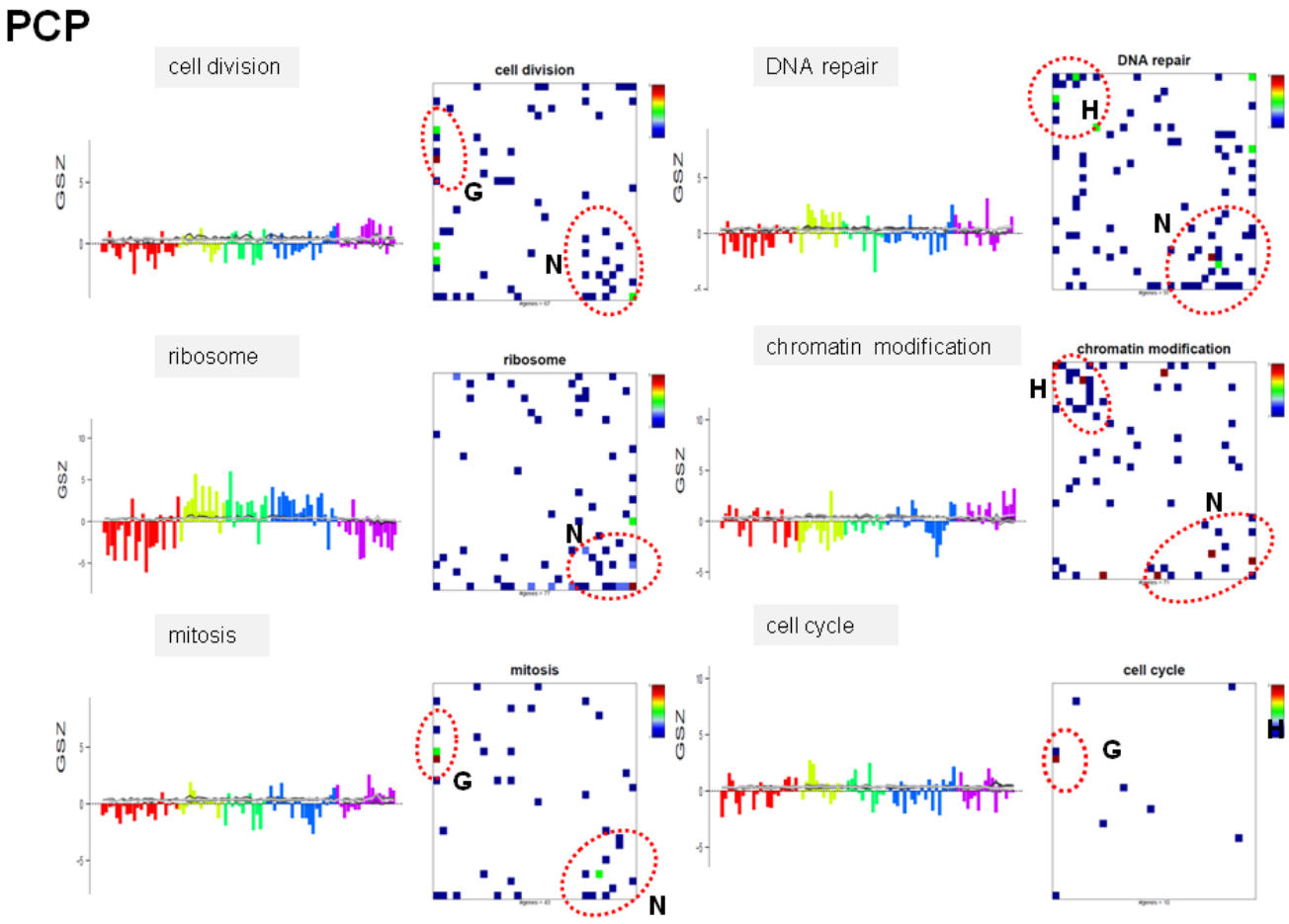

Figure S 22: Selected profiles and population maps of gene sets 'in concert with cell division' in PCP. Regions of overrepresentation in the maps are indicated by red-dotted ellipses. The letters refer to the respective overexpression spots. Note that we chose the same sets as in Figure S 21 for GBM.

2.8 Cancer gene sets in GBM

In this subsection we characterize gene sets extracted from other cancer studies either as highly specific for lymphoma subtypes and/or as signature sets for poor prognosis or common cancer genes. Note that these latter gene sets are typically upregulated in PN, and to a less degree, CL samples of GBM. They accumulate mainly in the ranges of spots N and/or O.

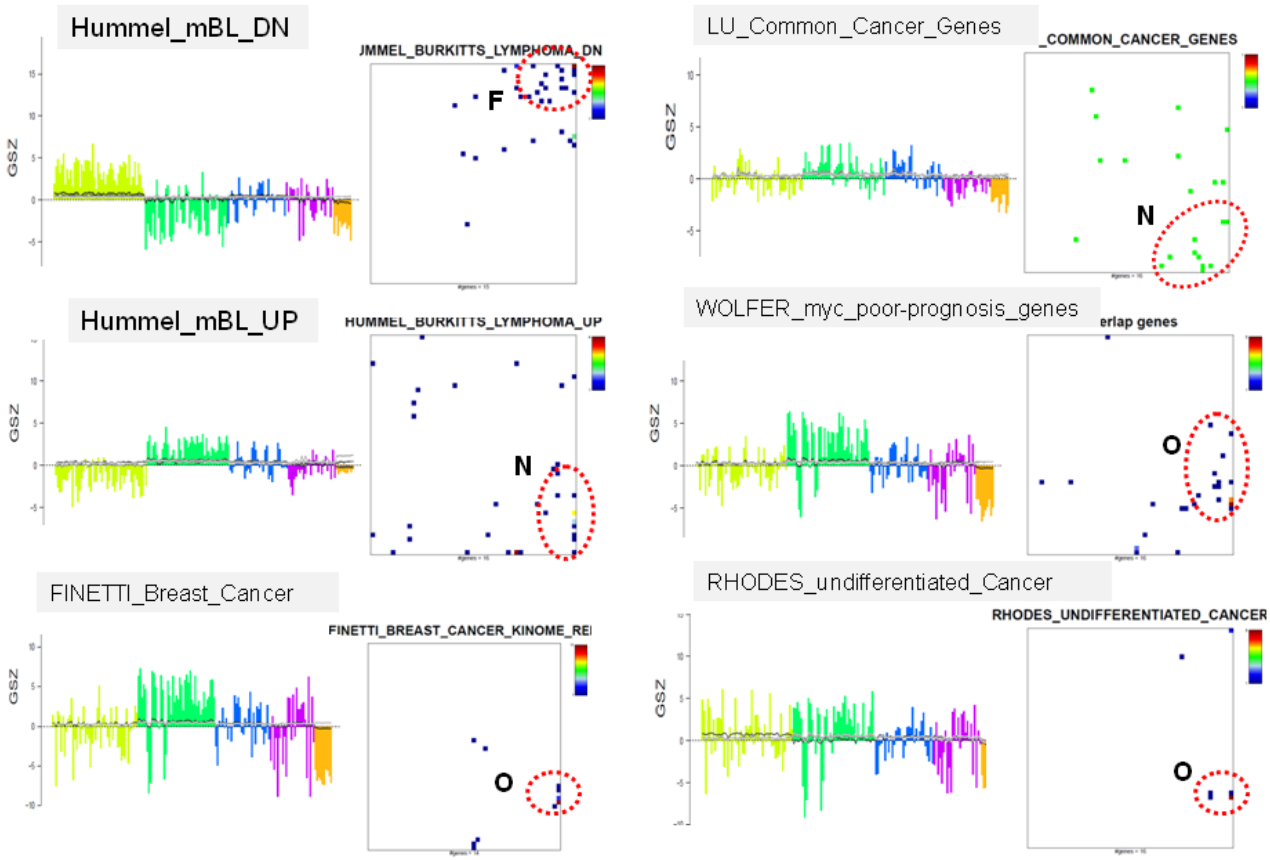

Figure S 23: Enrichment of cancer sets in GBM: Sets up- and down-regulated in Burkitts lymphoma (left part, [16]) and common cancer gene sets taken from refs. [17-19].

2.9 Gene sets related to different spot-modules of GBM

In this subsection we collect gene sets showing more diverse enrichment patterns and which are mainly located near spots which are upregulated in the intermediate GBM-subtypes NL or CL and in healthy brain (Figure S 24). For example, sets related to nervous function accumulate in spot K and upregulate in NOR- and NL-samples.
Figure S 25 shows a collection of diverse gene sets. 'Mitochondral activity' is related to spots I and J with specific overexpression in the NL-subtype and underexpression in the CL-subtype. In contrast, spot A can be assigned to 'aging brain_DN' and 'axon injury' with moderate upregulation in CL- and NL-subtypes. Spot A might be related to DNA damage as indicated by the population map of the gene set referring to DNA damage after UV radiation.

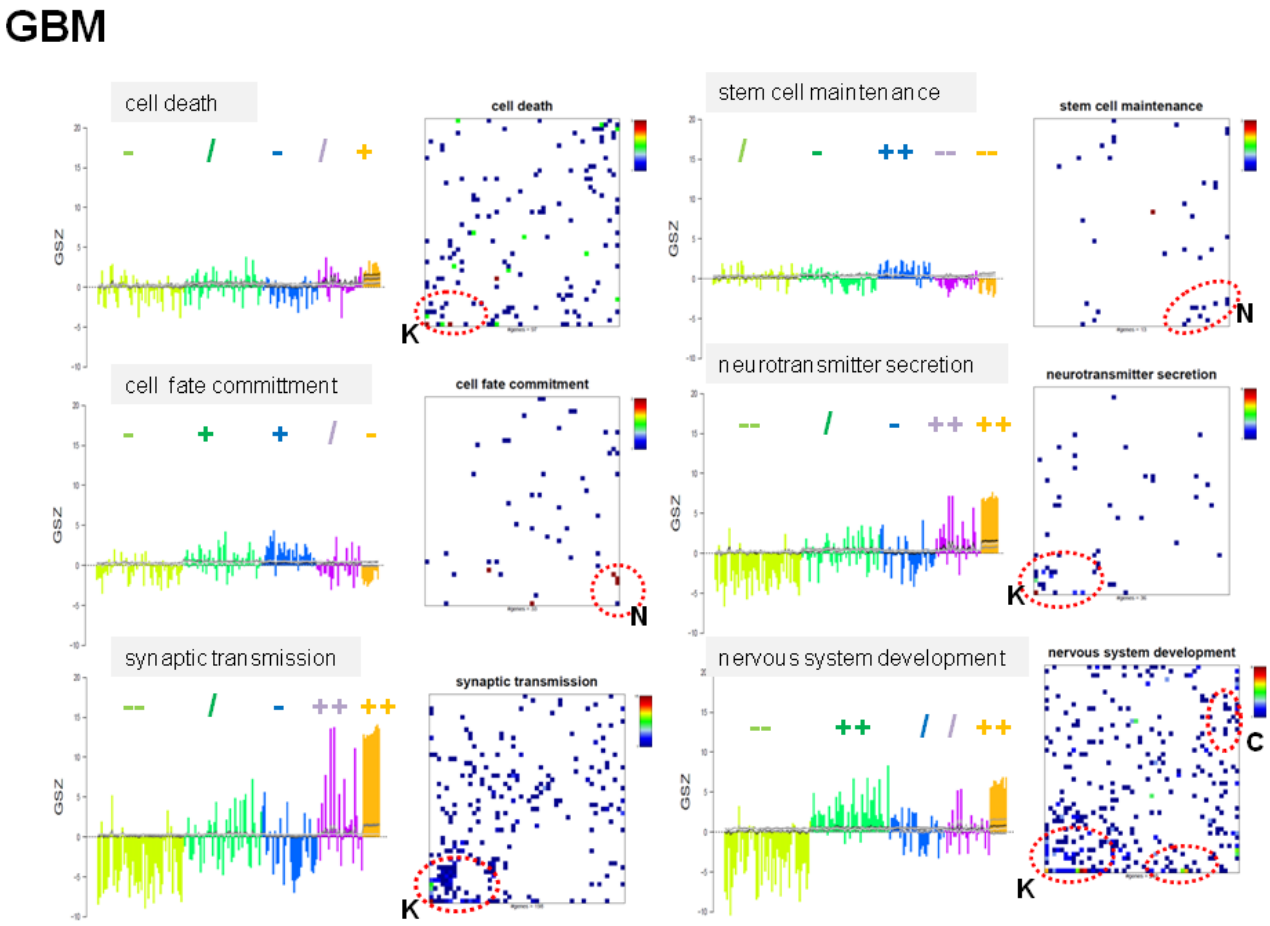

Figure S 24: Selected profiles and population maps of gene sets overexpressed in the NL- or CL-subtypes. Regions of overrepresentation in the maps are indicated by red-dotted ellipses. The letters refer to the respective overexpression spots.

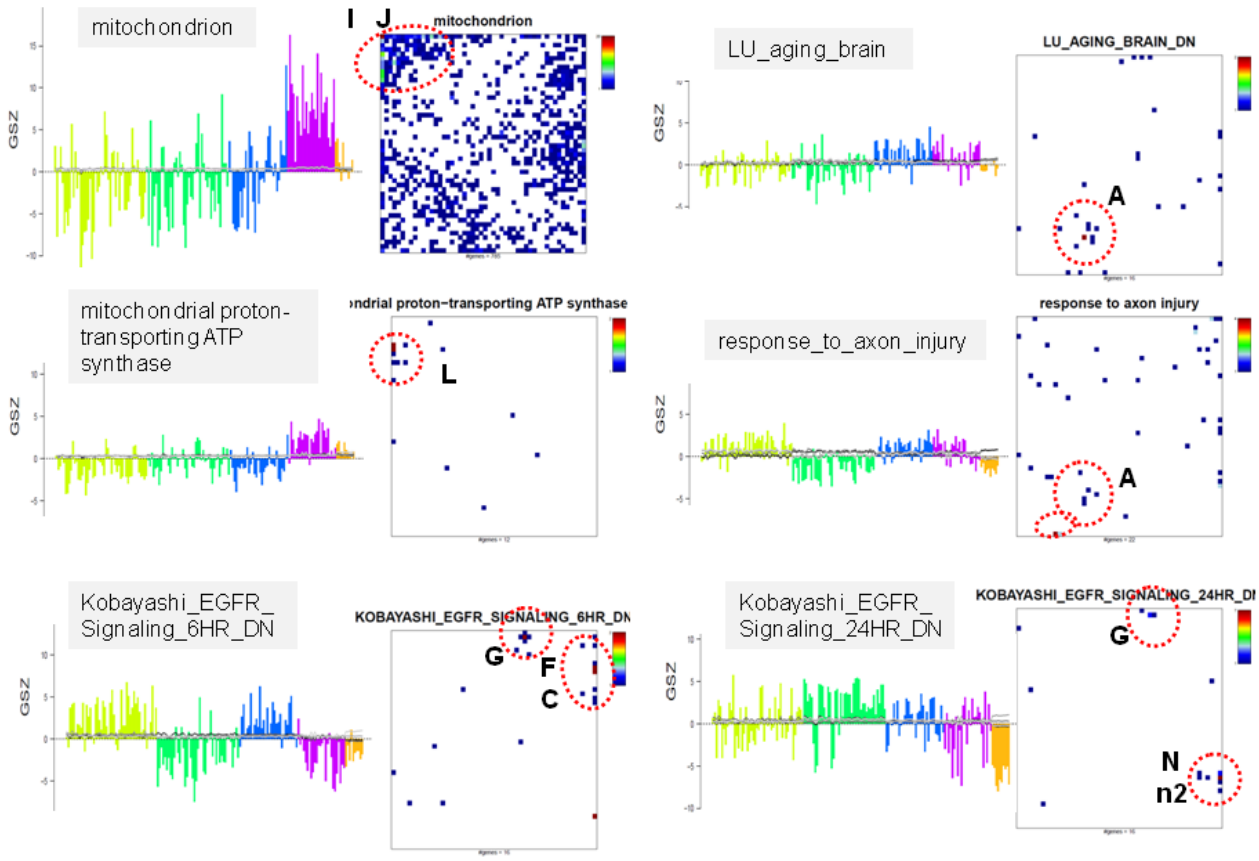

Figure S 25: Gene sets related to diverse processes.

2.10 Cell type and tissue sets in GBM

In this subsection we analyze gene sets derived from different brain tissues and cell lines of nervous tissues. Note for example, that developing astrocytes show a similar signature as genes sets changing in concert with 'cell division' genes (see Figure S 17) and partly, common cancer genes (Figure S 23). The astrocytic signature, on the other hand, accumulates in spot A with joint upregulation in NL and CL subtypes and joint downregulation in MES and PN subtypes. On the other hand, the oligodendrocytic, neuronal and healthy brain signatures are clearly antagonistic in CL and NL samples. However, they differ in the regulation profiles in the MES and PN samples.

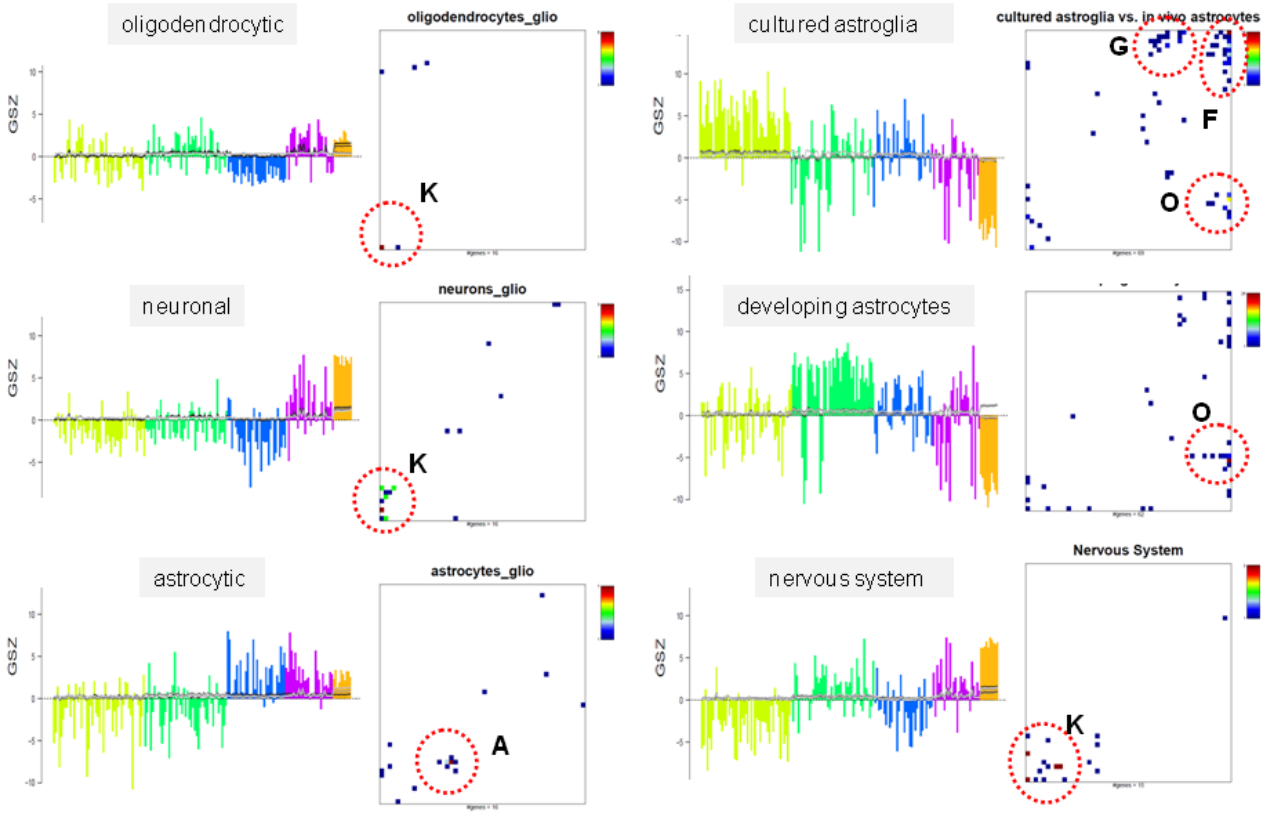

Figure S 26: Selected profiles and population maps of gene sets related to different nervous cells and tissue. Sets are taken from the brain transcriptome data base [20] and tissue profiling study (nervous system) [12].

2.11 Contaminations, outliers and misclassified samples

Large tumor sample collections are prone to different effects not (or not directly) related to the expression profiles of the diseased tissue such as contaminations with healthy tissue (brain, blood etc.), different levels of RNA quality after extraction and wet lab preparation, technical biases due to day-to-day variations of hybridizations and data recordings. Moreover, biological patient-to-patient variance is typically high and can be caused by other factors than the disease under study. The noisy character of the GSZ-profiles and also the scatter of the global expression characteristics manifest this variability of the data. The development, selection and qualified application of suited methods of quality control aiming at identifying, understanding and possibly also removing such effects represent a separate complex topic not addressed here in detail. However, our portraying approach offers a simple and direct option to check the whole-genome expression landscapes of the individual samples by visual inspection of their molecular 'faces'. Particularly one searches for conspicuous spot patterns that clearly deviate from that of the majority of samples assigned to the same class.

In Figure S 27 we re-plotted the CN similarity plot of GBM together with selected individual portraits of samples which are located either outside of the main clusters and/or within an apparently 'false cluster'. For example, samples no. 326 and 156 originally assigned to the MES- and PN-subtypes are found within the 'wrong' area of the net near the green PN- and yellow MES-cluster, respectively. Comparison of the portrait of sample 156 (and partly 321) with the mean portraits of the MES-and PN-subtypes reveals that its expression landscape obviously represents a combination of both expression signatures where the MES-signature more heavily contributes to the mixture than the PN-signature in contradiction to the original class assignment taken from ref. [21]. Another heterogeneous group of samples (e.g. no. 290, 152, 358) form a set of outliers near the blue CL-cluster. Inspection of the respective portraits reveals that a few overexpression spots (e.g., 'L', 'B' and 'D') are obviously responsible for this behavior: They are not observed in the majority of the remaining CL-samples. A similar argumentation applies to outlier samples no. 326, 84 and 87 showing strong expression of spot 'n1'. Note also that these groups of outliers are mostly heterogeneous, i.e. they contain samples assigned to different subtypes.

These 'outlier'-spots are mostly relatively rare and unspecific for one of the GBM-subtypes (see e.g. the abundance bar plot for spots 'L' and 'D'). This result suggests that these features are presumably caused by contaminations of non-tumor cells or by treatment effects and thus they are not or not directly related to GBM. Gene set analysis shows, that spot 'B', for example, contains an enriched number of genes related to 'xenobiotics' and 'drug metabolism'.

Hence, misclassifications of samples can be caused by the mixing of different subtypes and also by outlier features which are presumably not related to cancer, but which make samples of different subtypes similar. These examples demonstrate that our portraying approach not only detects potential outliers and misclassified samples but in addition helps researchers to generate hypotheses about the origin of these effects and also to extract more detailed information from the data, for example, by applying spot-related functional analysis.

Figure S 28 illustrates how the expression portraits in log FC and log log FC systematically change throughout the samples. One detects samples of the PN and MES subtypes which are obviously misclassified showing similar portraits as the majority of samples of the respective antagonistic subtype. On the other hand, the portraits of the intermediate subtypes CL and NL continuously vary between the portraits of the PN and MES subtypes.

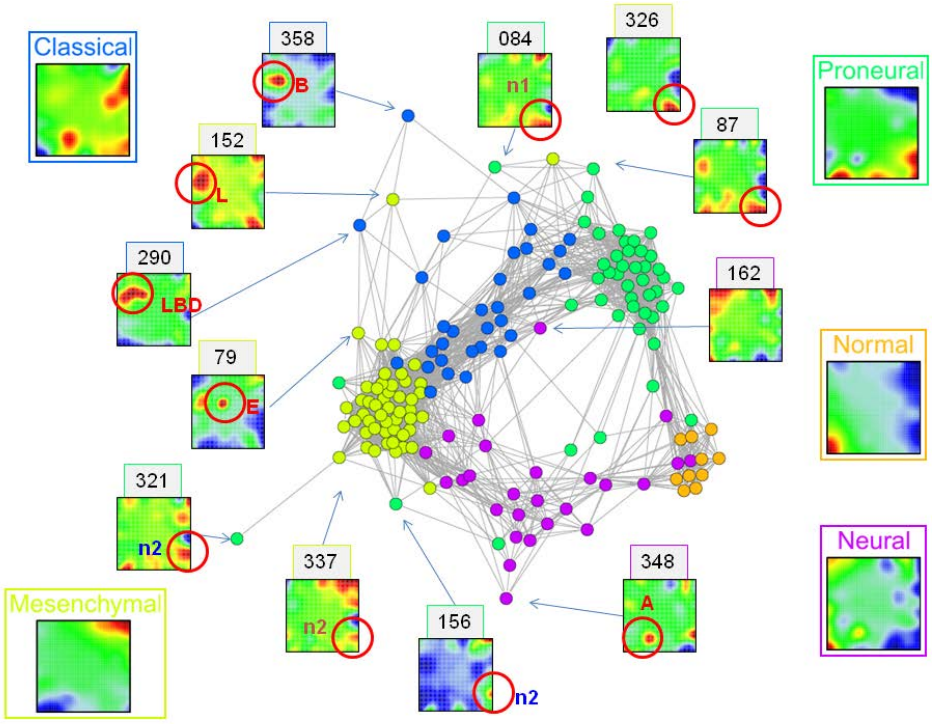

Figure S 27: Outliers and misclassified samples in GBM are indicated in the CN-similarity plot by arrows together with the respective sample portraits. The subtype-averaged mean portraits are shown for comparison at the left and right margins of the figure. The red circles and the letters assign the spots causing the partly atypical properties of the samples.

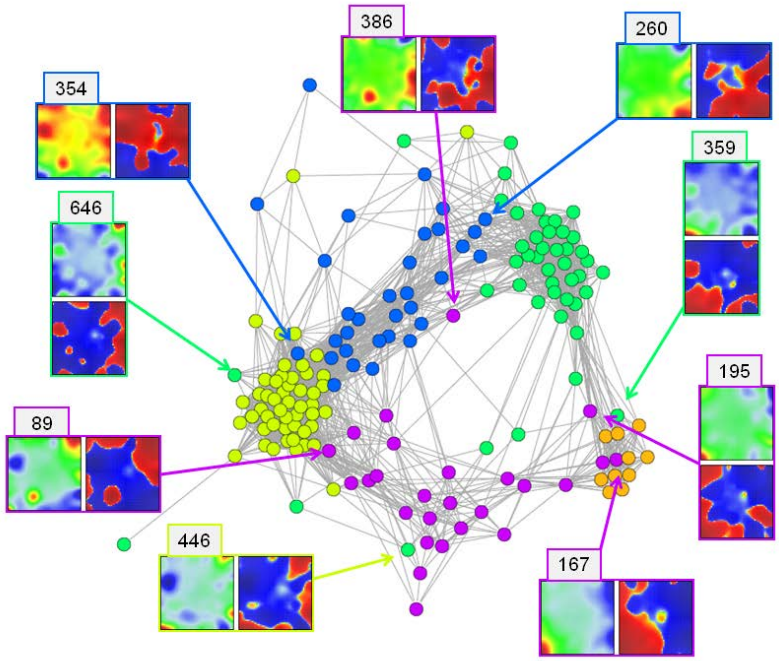

Figure S 28: Samples suspect for misclassification are indicated in the CN-plot of GBM together with their portraits in log FC- and log log FC-scales.

3 References

1. Binder H, Preibisch S. "Hook" calibration of GeneChip-microarrays: Theory and algorithm. Algorithms for Molecular Biology 2008; 3:12.
2. Binder H, Krohn K, Preibisch S. "Hook" calibration of GeneChip-microarrays: chip characteristics and expression measures. Algorithms for Molecular Biology 2008; 3:11.
3. Binder H, Preibisch S, Berger H. Calibration of microarray gene-expression data. In: Grützmann R, Pilarski C, eds. Methods in Molecular Biology. New York: Humana Press, 2009:376-407.
4. Binder H, Preibisch S, Kirsten T. Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. Langmuir 2005; 21:9287-302.
5. Binder H, Preibisch S. GeneChip microarrays - signal intensities, RNA concentrations and probe sequences. J Phys Cond Mat 2006; 18:S537-S66.
6. Fasold M, Stadler PF, Binder H. G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration. BMC Bioinformatics 2010; 11:207.
7. Binder H, Bruecker J, Burden CJ. Non-specific hybridization scaling of microarray expression estimates - a physico-chemical approach for chip-to-chip normalization. J Phys Chem B 2009; 113:2874-95.
8. Binder H, Krohn K, Burden C. Washing scaling of microarray expression. BMC Bioinformatics 2010; 11:291.
9. Fasold M, Binder H. RNA-quality scaling of GeneChip microarray expression. submitted 2012.
10. Binder H, Wirth H, Galle J. Gene expression density profiles characterize modes of genomic regulation – theory and experiment. Journal of Biotechnology 2010; 149:98-114.
11. Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst Biol 2011; 7.
12. Wirth H, von Bergen M, Binder H. Mining SOM expression portraits: Feature selection and integrating concepts of molecular function BioData Mining 2012; 5:18.
13. Ultsch A, Siemon HP. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proceedings of International Neural Networks Conference (INNC): Kluwer Academic Press, 1990:305-8.
14. Wirth H, Loeffler M, von Bergen M, Binder H. Expression cartography of human tissues using self organizing maps. BMC Bioinformatics 2011; 12:306.
15. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 2008; 24:719-20.
16. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TFE, et al. A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling. N Engl J Med 2006; 354:2419-30.
17. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, et al. The Prognostic Role of a Gene Signature from Tumorigenic Breast-Cancer Cells. New England Journal of Medicine 2007; 356:217-26.
18. Wolfer A, Wittner BS, Irimia D, Flavin RJ, Lupien M, Gunawardane RN, et al. MYC regulation of a “poor-prognosis” metastatic cancer cell state. Proc Natl Acad Sci USA 2010; 107:3698-703.
19. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings Of The National Academy Of Sciences Of The United States Of America 2004; 101:9309-14.
20. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. The Journal of Neuroscience 2008; 28:264-78.
21. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17:98-110.