Function Shapes Content: DNA-Methylation Marker Genes and their Impact for Molecular Mechanisms of Glioma

Lydia Hopp*, Edith Willscher, Henry Löffler-Wirth and Hans Binder*

Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Härtelstr, 16–18, 04107 Leipzig, Germany

Abstract: Glioma is a clinically and biologically diverse disease. It challenges diagnosis and prognosis due to its molecular heterogeneity and diverse regimes of biological dysfunctions which are driven by genetic and epigenetic mechanisms. We discover the functional impact of sets of DNA methylation marker genes in the context of brain cancer subtypes as an exemplary approach how bioinformatics and particularly machine learning using self organizing maps (SOM) complements modern high-throughput genomic technologies. DNA methylation changes in gliomas comprise both, hyper- and hypomethylation in a subtype specific fashion. We compared pediatric (2 subtypes) and adult (4) glioblastoma and non-neoplastic brain. The functional impact of differential methylation marker sets is discovered in terms of gene set analysis which comprises a large collection of markers related to biological processes, literature data on gliomas and also chromatin states of the healthy brain. DNA methylation signature genes from alternative studies well agree with our signatures. SOM mapping of gene sets robustly identifies similarities between different marker sets even under conditions of noisy compositions. Mapping of previous sets of glioma markers reveals high redundancy and mixtures of subtypes in the reference cohorts. Consideration of the regulatory level of DNA methylation is inevitable for understanding cancer genesis and progression. It provides suited markers for diagnosis of glioma subtypes and disentangles tumor heterogeneity.

Keywords: Glioma, molecular subtypes, DNA methylation, gene regulation, bioinformatics.

1. INTRODUCTION

Genomic technologies offer the promise of a comprehensive understanding of cancer and to characterize its molecular determinants in terms of 'marker signatures'. A 'marker signature' can be defined as a concerted alteration of a set of molecular features with specificity in terms of diagnosis, prognosis or prediction of therapeutic response [1]. Marker signatures define cancer subtypes and provide a central tool to disentangle tumor heterogeneity on the molecular level. Marker signatures can be used for classification purposes solely without relevance for cancer biology. On the other hand, the 'guilt by association' principle assumes that the concerted behavior of sets of molecular features reflects underlying common functions [2]. In this publication we discover the functional impact of sets of DNA methylation marker genes in the context of brain cancer subtypes as an exemplary approach how bioinformatics and particularly machine learning complements modern high-throughput genomic technologies.

High-throughput gene expression analysis has revolutionized genetics over the last 15 years since the seminal publication in 1999 [3]. This new technology has been extensively used to find responses to fundamental questions from understanding tumor biology, to prediction of progression, and treatments [1]. Cancer is a process driven by the accumulation of abnormalities in gene function which is initiated by genetic and epigenetic defects. Hence, the expression of genes and also tumor histology are 'only' the phenotypes reflecting the underlying (epi-)genetic alterations in the tumor. With the aim of improving molecular marker signatures in terms of cancer care and cancer biology one must therefore supplement expression signatures with (epi-) genetic information to link causal factors with downstream mechanisms of cancer genesis and progression. Moreover, 'functional' gene sets are expected to outperform purely 'formal' ones because they meet not only statistical but also biological criteria and thus an increased level of evidence (see below).

The arrival of next generation sequencing technologies and also of new types of microarrays and their application to cancer now gives us access to this information in terms of the abundance of ten thousand of mRNA transcripts per sample, millions of mutations, methylation levels of hundred thousand of DNA CpG sites and of histone side chains. Besides genetic defects such as mutations and copy number alterations, genome-wide DNA methylation acts as an epigenetic factor that governs the particular activity state of the genome by assuring the proper regulation of gene expression and stable gene silencing [4]. DNA methylation is associated with histone modifications and the interplay of them is crucial to regulate the

*Address correspondence to these authors at the Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Härtelstr, 16–18, 04107 Leipzig, Germany; Tel: +49-341-9716697; Fax: +49-341-9716669;
E-mails: hopp@izbi.uni-leipzig.de, binder@izbi.uni-leipzig.de

ISSN: 1929-2260 / E-ISSN: 1929-2279/15

functioning of the genome by changing chromatin architecture. Unlike genetic alterations, DNA methylation is heritable and reversible what makes it interesting for therapy approaches. Recent work shows that DNA methylation signatures are robust biomarkers which extend our ability to classify cancer and which predict outcome and therefore represent promising targets [5-10].

Glioblastoma (GBM), the most common primary brain tumor, carries a universally dismal prognosis in children and adults. Affected patients have a uniformly poor prognosis with a median survival of about one year. Thus advances on all scientific and clinical fronts are needed. Moreover, the molecular foundations of lower-grade gliomas (LGGs, WHO grade II and III) remain less well characterized than those of their fully malignant grade IV GBM counterpart. Genome-wide sequencing data show that about 50% of cases have at least one somatic mutation in a gene that is associated with the epigenetic machinery including DNA methylation, histone modifications and/or chromatin remodeling [11]. In an attempt to better understand gliomas, many groups have turned to high dimensional profiling studies based on genetic, transcriptional and DNA methylation markers [8, 12-21]. The mutual relation between the marker sets and their functional impact is not clear in many cases.

The legitimate excitement about the attractiveness of molecular technologies should not overlook adherence to the rules of evidence [22]. A recent study showed that 60% of the published breast cancer outcome signatures were not significantly better outcome predictors than random signatures [23]. Moreover, statistical significance in a training cohort does not demonstrate a specific outcome association [22]. Hence, it is questionable to deduce a mechanism from statistically significant molecular markers solely because also random (false positive) markers can suggest this. Optimally, the statistical significance of markers should by supported by a sound and reliable interpretation in terms of their biological function. Indeed, it has been also shown that random signatures have only weak pathway enrichment and thus functional meaning, whereas non-random ones do have. Thus, a strong cancer-related functional context of marker signatures should make them more relevant and robust.

We developed an innovative analysis pipeline for identifying marker genes from large scale genomic cancer data based on machine learning using self organizing maps (SOM) [24-26]. This method has been recently applied to classify a series of cancer entities into subtypes and to characterize them on molecular level using mainly gene expression data [25, 26]. It enables the 'portrayal' of molecular data landscapes, e.g. in terms of gene expression maps. These maps not only enable evaluation of data based on visual information. They also allow scientists to extract sets of marker genes with high resolution, to evaluate their functional impact and to compare them with alternative markers from independent studies [27]. Here we apply this method to DNA-methylation data of brain cancer taken from ref. [5] in order to extract sets of differentially methylated genes, to demonstrate their functional impact and to discuss their relevance in terms of glioma biology. We show that the intrinsic structure of this methylation data is compatible with a multitude of signature sets extracted from independent cohorts including DNA methylation and gene expression data thus reflecting their common biological background. We show that the specifics of biological functions of different glioma subtypes shape the content of these marker sets. In turn, including not only standard function information according, e.g. to different gene ontology (GO) terms but also about chromatin states of the healthy brain enables to study epigenetic mechanisms of glioma progression and the associated interplay between gene activity and methylation.

2. DATA AND METHODS

2.1. Methylation Data

Microarray-derived DNA methylation data (Illumina HumanMethylation450 BeadChip) of 136 GBM and 6 control samples were taken from ref. [5, 7] (available under GEO Series accession number GSE36278) in terms of the beta values of 485,512 CpG's. The data refer to pediatric and adult GBM and to non-neoplastic cerebellum specimen as controls (Table **1**). GBM samples were classified according to the methylation clusters identified in [5]. Accordingly, the pediatric GBM split into two subtypes carrying mutations of the H3F3A gene which affect two different amino acids of histone H3.3, namely G34 or K27, respectively. The adult GBM were classified into four subtypes labeled according to correlations with genetic defects. These genetic hallmarks constitute mutations of the IDH1 gene ('IDH' subtype) and focal copy number (CN) amplifications of the PDGFRA ('RTKI' subtype) or EGFR ('RTKII' subtype) gene both coding receptor tyrosine kinases (RTK). The RTKII cases are called 'classical' because

Table 1: DNA Methylation Data Set (Sturm *et al*. [5])

subtype	n		genetic hallmark[1]	expression subtype[2]
adult	2	control		
fetus	4	control		
MES(enchymal)	36	adult GBM		mesenchymal
RTKII (classical)	22	adult GBM	CDKN2A (CN loss), EGFR (CN amplification)	classical
RTKI (PDGFRA)	23	adult GBM	PDGFRA (CN amplification)	
IDH	19	adult GBM	IDH1 (mutation)	proneural
G34	18	pediatric GBM	H3F3A/ G34 (mutation)	
K27	18	pediatric GBM	H3F3A/ K27 (mutation)	

[1]see, e.g., [30] for an overview.
[2]according to [31].

they enrich combined gain of CNs at chromosome 7 and loss of CNs at chromosome 10 both representing a hallmark of IDH1 wild type GBM [5]. The 'mesenchymal' subtype shows a lower incidence of GBM typical CN alterations.

Methylation levels were estimated in a gene centric way by averaging the CpG-related beta values over genomic regions of the promoters of each gene ranging from 1500 bp upstream the transcription start site (TSS) to the TSS (Figure **S1a**). Beta values are defined as the relative methylation level which can vary between values of zero (no methylation) and unity (full methylation). For SOM analysis beta values were transformed into M values (M_{gene}= $\log_{10}$ [$beta_{gene}$/(1 – $beta_{gene}$)]) which theoretically cover the range between minus infinity (no methylation) to plus infinity (full methylation). M values are statistically more valid because they avoid heteroscedasticity of differential methylation values for large (beta>0.8) and small (beta<0.2) beta values [28]. In the intermediate beta range (0.2<beta<0.8) beta and M are nearly linearly correlated. For SOM analysis we used either M values (MetSOM) or centralized M values (DmetSOM), ΔM_{gene}= M_{gene} - $\langle M_{gene} \rangle_{samples}$, where the angular brackets denote averaging over all samples studied. DmetSOM attenuates methylation changes independent of the methylation level whereas MetSOM directly considers absolute methylation levels and thus enables to distinguish highly methylated from weakly methylated genes. MetSOM has the advantage to resolve modules of co-methylated genes in more detail with higher granularity [29].

2.2. Gene Expression Data

Three expression data sets were used to establish associations with methylation data (see Table **2**). Microarray expression data of 30 matched samples and 3 unmatched fetal controls were taken from [5]. They comprise the same subtypes as the methylation data. A second set of expression data was taken from [31] and processed and analyzed previously by us [26]. This data comprises healthy brain, mesenchymal,

Table 2: Gene Expression Data of GBM

methylation classes	Sturm *et al*. [5] matched samples	Hopp *et al*. [31] matched classes	Reifenberger *et al*. [20] matched classes
adult		healthy (n=10)	proneural IDH1 wt (n=14)
fetal	fetal (n=3)		
MES	mesenchymal (5)	mesenchymal (50)	mesenchymal (21)
RTKII	RTKII (3)	classical (32)	classical (23)
RTKI	RTKI (6)		
IDH	IDH (7)	proneural (45)	proneural IDH1 mut (12)
G34	G43 (4)		
K27	K27 (5)		

classical, proneural and neural GBM which were matched with the classes of the methylation data. The third data set was taken from [20]. It consists of GBM with mesenchymal, classical, proneural with IDH1/2-mutational and proneural with IDH1/2-wild type characteristics.

2.3. Methylation Portrayal Using Self Organizing Maps

Gene-centric methylation data were clustered using self-organizing map (SOM) machine learning. Only genes on autosomes were considered to avoid gender specific effects [32]. The SOM method translates the gene data matrix into metagene data of reduced dimensionality [32]. Each metagene data is visualized in a sample-specific fashion by arranging the metagenes in a two-dimensional quadratic 40x40 grid and by appropriately color coding of the data values. The mosaic images obtained serve as fingerprint portraits of the methylation landscapes of each sample. Class-specific mean portraits were generated by averaging the metagene landscapes of all cases belonging to one class. SOM size and topology was chosen to allow robust identification of methylation modules inherent in the data in terms of so-called spot clusters as described in our previous publications [24, 27]. Overview spot maps were generated by collecting all hypermethylation spots of individual portraits into one map. Two different SOMs were trained using (i) methylation M-values (MetSOM) and (ii) centralized M-values with respect to the mean M of a gene averaged over all samples (DmetSOM). Expression data were analyzed previously by means of SOM portrayal [20, 26]. For SOM analyses we used the R-package 'oposSOM' which is publically available from the Bioconductor repository [33].

2.4. Marker Set Selection Using Spot Modules

The SOM algorithm arranges similar meta-profiles referring to the same dimension of variation together into neighbored pixels of the map whereas more different ones referring, e.g. to mutually independent dimensions are located at more distant positions. In consequence, neighbored meta-features tend to be colored similarly in each image which shows typically a smooth blurry texture with red and blue spot-like regions representing clusters of high and low meta-feature values, respectively. Meta-features from the same spot can be assumed to be co-regulated (i.e. associated in a functional sense) owing to their similar profiles whereas different, well-separated spots potentially collect meta-data of different regulatory modes. The spot modules detected can be seen as a natural choice to identify context-dependent patterns in complex data sets.

Spot clusters were determined using a hypermethylation percentile threshold applied either to the class-averaged or individual portraits (DmetSOM), or, alternatively a correlation threshold applied to the metagene profiles (MetSOM) [24, 27]. Significance of differential methylation of genes included in the spot clusters was estimated using a shrinked t-test and false discovery rate based multitest adjustment [24, 27] and a multitest adjusted correlation q^2 test as described in [34].

2.5. Gene Set Functional Analysis

For the interpretation of the functional context of spot modules we applied gene set enrichment analysis using the gene set enrichment score (GSZ) [35] or simply calculating mean methylation or expression values averaged over the genes of the set. The GSZ estimates the degree of reliability that a gene set with reference to a certain biological functionality is related to a list of genes with unknown functional impact, e.g., derived from the spot modules. Note that the GSZ-score combines enrichment measures (i.e. the probability that more genes of the set are included in the list as expected by chance) with differential methylation (i.e. the difference of mean methylation between the set and the list) [27]. GSZ analysis was applied also to the full number of genes under study. Then it estimates the degree of conformance of the methylation of the genes of the selected set in a selected sample. High and low GSZ values usually beyond $|GSZ|>5$ reflect concerted hyper-/hypomethylation of the set. GSZ-profiles of all samples studied are complemented by gene set maps which visualize the distribution of genes in the methylation landscape. Their strong accumulation in the spot areas reflects functional impact of the selected set on glioma biology whereas a virtually random spread suggests the lack of biological importance at least in the sample cohort used to train the SOM.

We considered a large collection of gene sets related to biological process (BP), cellular components (CC) or molecular component (MC) taken from the gene ontology classification (GO), standard literature sets taken from the GSEA-repository (see [27] and [36]) together with literature sets related to glioma biology implemented by us (see below). For estimating

the association of DNA-methylation with chromatin states of healthy brain we used lists of genes referring to these states in mid frontal lobe of adult persons taken from the epigenetic roadmap repository (http://www.roadmapepigenomics.org/). In total fifteen states were considered including active, poised and silent promoters and enhancers, heterochromatin and repetitive elements. These states were derived from combinations of histone marks detected in the respective genetic regions using a hidden Markov model [37, 38].

2.6. Diversity Analysis

Further downstream analysis comprises diversity analysis by means of (sample-) pairwise correlation heatmaps and their visualization using a correlation net presentation to extract the mutual similarity relations between the samples (see, e.g., [25, 26]).

For a quantitative estimation of sample clustering we make use of the silhouette score, $s_i = r_x - max\{r_{x'} / x' \neq x\}$ of sample *i* (belonging to subtype *x*). It is defined as the difference between the intra-class and the best inter-class similarity measures, r_x (x=1…X assign the classes) and $r_{x'}$, respectively. As similarity measures we use the *Pearson* correlation coefficient between the metagene methylation landscapes of the SOM portraits of the sample selected and the respective class mean. The silhouette score is positive for samples which best fit into the cluster still chosen whereas the score is negative for samples which better fit to other clusters. The scores were ranked within each class and visualized as 'silhouette plot' together with the cluster of minimum distance for negative scores using a color bar (see below). All downstream methods were described in [24, 27], illustrated in a pilot application [26] and implemented in 'oposSOM' [33].

3. RESULTS

3.1. SOM Portrayal of the Methylation Landscapes in GBM and Healthy Brain

We re-analyzed microarray DNA methylation data published in a previous study on pediatric and adult brain tumors and non-neoplastic controls [5] to get detailed insight into the methylation landscapes of gliomas and their impact for molecular mechanisms of cancer diversity, genesis and progression. On the average, CpG-related beta value reveal a smoothly decaying methylation level upstream of the transcription start site (TSS) of the genes and relatively noisy methylation in their first exon (Figure **S1a**). We averaged CpG-related beta values over the range from -1500 bp to 0 bp with respect to the TSS of each gene to obtain mean gene centric data characterizing the DNA methylation level in the promoter region of each gene. The frequency distribution of gene centric beta values shows a typical bimodal shape with maxima near zero (completely de-methylated CpG sites) and unity (completely methylated CpG sites, see Figure **S1b**). The distribution of the IDH-subtype clearly reveals a trend towards global hypermethylation: The fraction of weakly methylated genes decreases while the fraction of highly methylated genes increases compared with the distributions in the healthy controls. On the other hand the distribution of the G34-subtype shows the opposite effect and thus a trend towards global hypomethylation (see the arrows in Figure **S1b**).

In the next step, SOM data portrayal was applied to the gene-centric methylation data including all glioma samples and the non-neoplastic brain samples serving as reference. The method 'projects' the methylation data onto a two-dimensional grid of 40x40 pixels. Appropriate color-coding then visualized the methylation landscapes of each sample in terms of its individual methylation portrait (not shown). We averaged theses portraits taking into accounts all samples of each class to identify class-specific methylation signatures. Figure **1a** shows the gallery of these mean portraits for all classes studied. Red and blue regions in the images refer to genes with high and low methylation levels of the probed CpG regions, respectively. Hence, the map can be segmented into regions containing genes of high and low methylation levels of their promoters and in regions containing genes with strongly variant and almost invariant methylation levels (Figure **1b**). The regions of variant and of invariant genes thus include regions of high and low mean methylation levels as well.

The SOM algorithm clusters genes with similar methylation profiles among the samples together into the spot-like areas appearing in the methylation maps. Accordingly, groups of genes with characteristic methylation profiles can be extracted from the map using a correlation metrics (Figure **1c**). Accordingly, the methylation landscape divides into regions of hyper- and hypomethylated genes in almost all samples and in regions showing differential methylation effects between them as indicated in the figure. We calculated the mean methylation level and its variance separately for each subtype using the individual methylation portraits (Figure **1d**). One sees that IDH, RTKII and, to a less degree mesenchymal tumors are globally

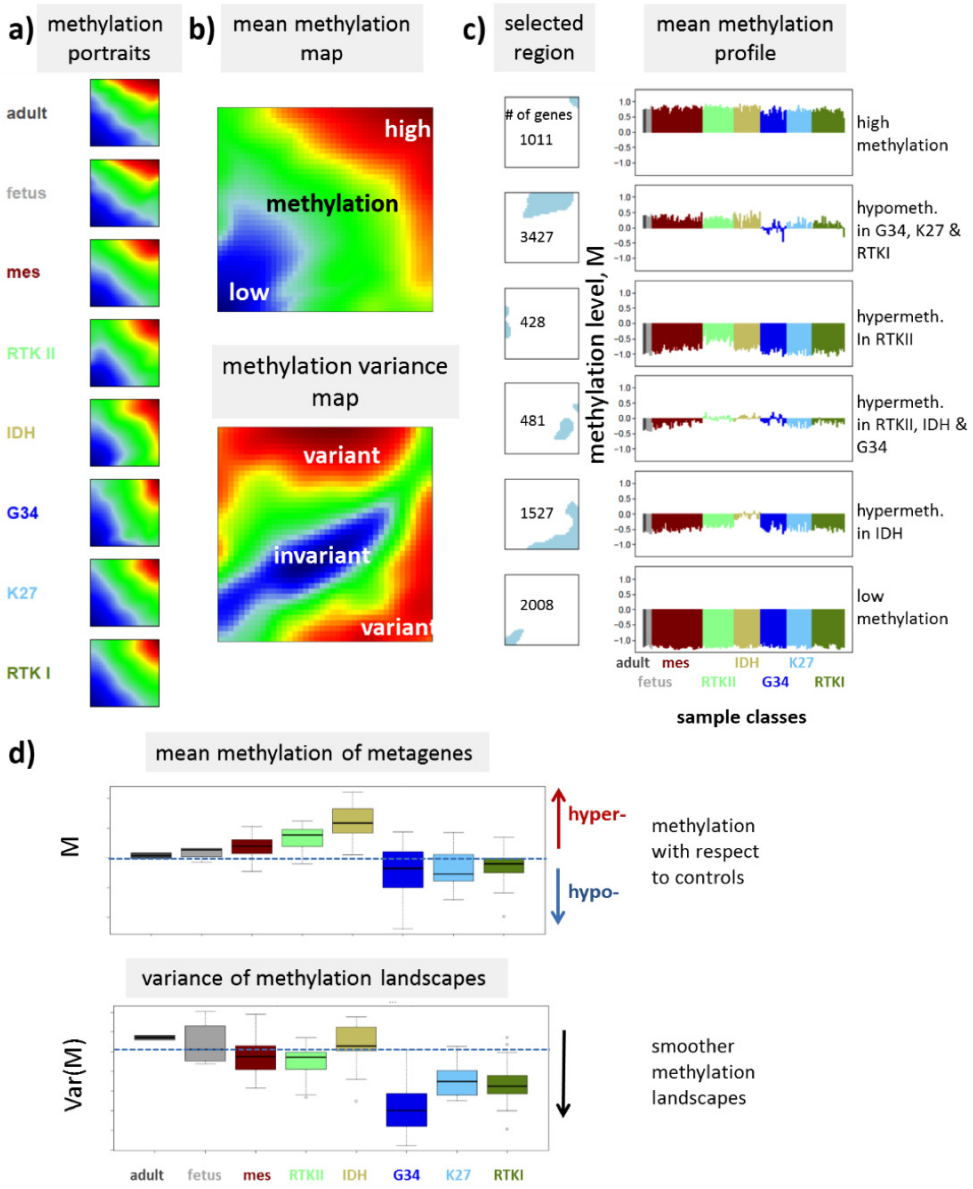


Figure 1: SOM (MetSOM) portrayal of the methylation landscapes of glioma subtypes: **a**) SOM portraits of glioma subtypes and of healthy controls. Red and blue colors assign regions containing genes with high and low methylation levels, respectively. **b**) The methylation overview map visualizes regions of high (red) and low (blue) methylation levels. The methylation variance map identifies regions of genes showing highly variable (red) and almost invariant (blue) methylation. **c**) Selected regions of the map show different methylation profiles among the samples. **d**) Mean methylation level and variance of the classes studied.

hypermethylated with respect to the controls whereas G34, K27 and RTKI are globally hypomethylated. The variance of the methylation level reflects the coarseness of the methylation landscapes of the subtypes. The decreased variance in gliomas compared with the controls reflects smoother landscapes in the tumors with more balanced methylation levels between the genes on the average.

In summary, methylation changes in gliomas comprise both, hyper- and hypomethylation in a subtype specific fashion. SOM mapping identifies genes with different methylation levels and specific alterations of the methylation levels between the subtypes.

3.2. SOM Portrayal of Centralized Methylation Data Improves Resolution (DmetSOM)

In the next step we trained a second SOM using centralized methylation values (DmetSOM) where the mean methylation level of each gene averaged over all samples was subtracted from its actual methylation M-value. Centralization focuses the view on methylation changes between the samples independent of the absolute methylation level of the genes and it improves

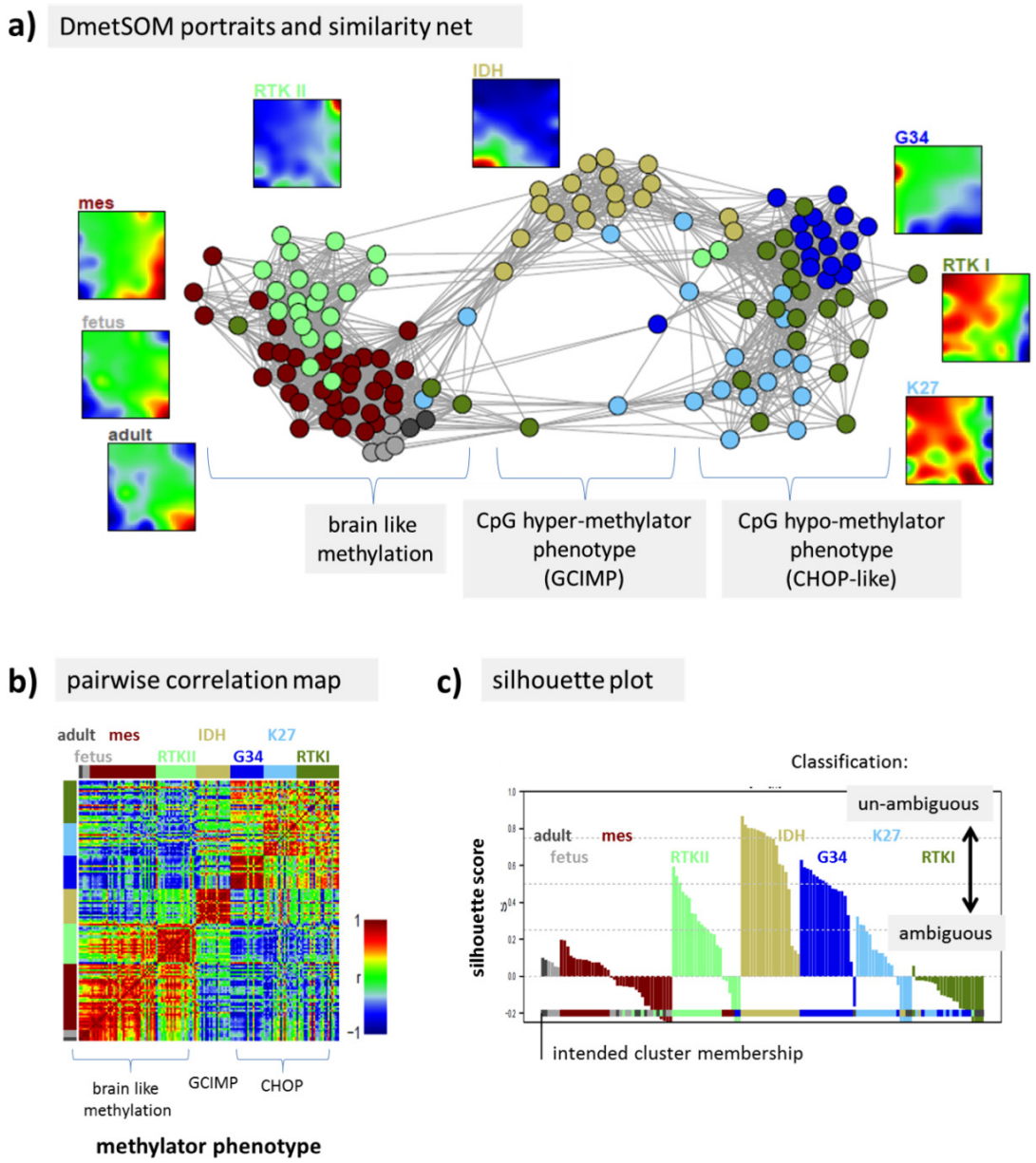


Figure 2: SOM portrayal of centralized methylation data (DmetSOM): **a**) SOM portraits of the GBM subtypes and of the controls and similarity net of the samples studied. Samples with strong mutual correlation coefficients are connected by lines. The sample classes can be divided into three main groups as indicated. **b**) The pairwise correlation heatmap visualizes the mutual correlation coefficient for all pairwise combinations of samples. **c**) The silhouette plot estimates the quality of classification of samples into methylation subtypes. Negative values indicate preference for other subtypes which are assigned as color bar below.

resolution with respect to differential markers that distinguish the different classes [29]. The class-averaged mean DmetSOM portraits shown in Figure **2a** are clearly more diverse then the respective MetSOM portraits shown in Figure **1a**. One clearly identifies similar textures of the maps of non-neoplastic brain (adult and fetal) and mesenchymal GBM and of K27 and RTKI GBM, respectively. The similarity net in Figure **2a** more clearly visualizes the mutual similarities of individual methylation landscapes of the samples based on the mutual (Pearsons) correlation coefficients between them which were color-coded in the heatmap in Figure **2b**. The classes can be roughly grouped into three superclusters which we assign as 'brain-like' because of the only small and moderate methylation changes in GBM; as (hyper-) Glioma CpG methylator phenotype (GCIMP) and as hypomethylator phenotype (CHOP) based on the global methylation drifts in GBM as suggested before in [5]. The brain-like and CHOP (and partly also GCIMP) groups show mainly anti-correlated methylation landscapes meaning that large groups of genes concertedly 'switch' their methylation

levels between these groups (see the blue off-diagonal areas in Figure **2b**).

Note that each class forms its own cloud of samples in the similarity net which still reflects its own specifics within each of the supercluster (see below). On the other hand, one observes a certain degree of fuzziness between the subtypes. For example, the K27 and RTKI sample clouds partly overlap. In the supplement we provide the individual sample portraits sorted for each GBM subtype using hierarchical clustering trees (Figure **S2**). Part of the samples shows methylation landscapes which can be interpreted as mixtures of different subtypes (e.g. of K27, RTKI and G34) or as mixtures with healthy brain methylation characteristics (part of the mesenchymal and RTKII samples). The 'personalized' portrayal of the samples enables the detailed assignment of these mixed characteristics.

The silhouette plot in Figure **2c** evaluates the robustness of class assignment for all samples. It reveals that the IDH, G34, RTKII, partly K27 and the controls form relatively robust classes whereas mesenchymal and especially RTKI are rather unambiguously assigned mainly due to overlapping characteristics with non-neoplastic fetal brain and G34 GBM, respectively (see the color bar in Figure **2c** which annotates the 'best class membership'). Note that our robustness analysis is based on gene-centric whole-genome methylation landscapes and thus it does not contradict the classification proposed in [5] which is based on the 8,000 most variant CpG probes. Our robustness analysis however illustrates the degree of fuzziness of class assignment which reflects the mutual overlap between them and possibly also common biological factors that drive tumorigenesis. We also performed independent component analysis (ICA) to estimate the similarity relations between the samples using an alternative method (Figure **S3**): Especially IDH and G34 systematic deviate in their methylation characteristics whereas the other subtypes are obviously more closely related each to each other.

3.3. Segmentation of the Map into Spot-Sets of Methylation Markers and their Functional Context

The summary map in Figure **3a** colors regions hypermethylated in any of the subtypes in red. After appropriate segmentation (see methods section) we identified twelve spot-clusters containing between nearly two-thousand and sixty single genes. Six of these spot regions labelled A – F show profiles with subtype-specific differential methylation whereas six additional 'satellite' spots reveal more complex profiles (Figure **3b**). For example, the methylation profiles of GBM in spots D and D1 are almost similar whereas methylation of the controls completely changes sign. The methylation profiles of the genes in most of the spots are highly correlated providing significance levels beyond $p<10^{-64}$ using a q^2-test statistics [34, 39]. Note that all spots except E1 and partly B1 are found in regions of highly variable methylation values (Figure **S4**). To estimate the absolute methylation levels of the genes in each of the spots we map them into the MetSOM (Figure **S5**). In general one sees that hypermethylation in G34 means that weakly methylated genes in healthy brain accumulate methylation marks whereas in IDH also genes with intermediate M values are affected. Importantly, two of the satellite spots of less variant genes refer to high (spot B1) or low (E1) methylation levels in all system studied. In the following we will focus on the main spots and the latter two satellite spots.

DmetSOM analysis is based on centralized M values to increase sensitivity to methylation changes relative to the mean M value of each gene. In general one however asks for cancer specific methylation changes relative to the healthy controls. We therefore analyzed difference SOM with respect to the mean methylation map of non-neoplastic brain tissue of adults. The differential methylation landscapes support the superclusters of brain-like, GCIMP and CHOP-like methylation patterns (Figure **S6**). Moreover, one sees that spot A1 is hypomethylated and spot D1 hypermethylated in all GBM compared with the healthy brain.

Extended spot statistics reveals that spots C, E and F are highly sensitive and specific as hypermethylation markers for the IDH, G34 and RTKII subtypes, respectively (Figure **3c**). The respective areas of the map thus can be interpreted as fingerprint regions as indicated in Figure **3a**. The spot number distributions for each of the subtypes show that most of the samples of all classes show only one or two spots (Figure **3d**). However, part of the GBM samples and especially that of the MES- and RTK I- subtypes can express up to five spots in parallel this way reflecting the high degree of fuzziness of these classes on feature level.

Gene set enrichment analysis provides first ideas about the functional context of the genes in the spot modules (Table **3**). Spots D and E are associated with biological processes already found in gene expression analysis on GBM [26] such as immune response and

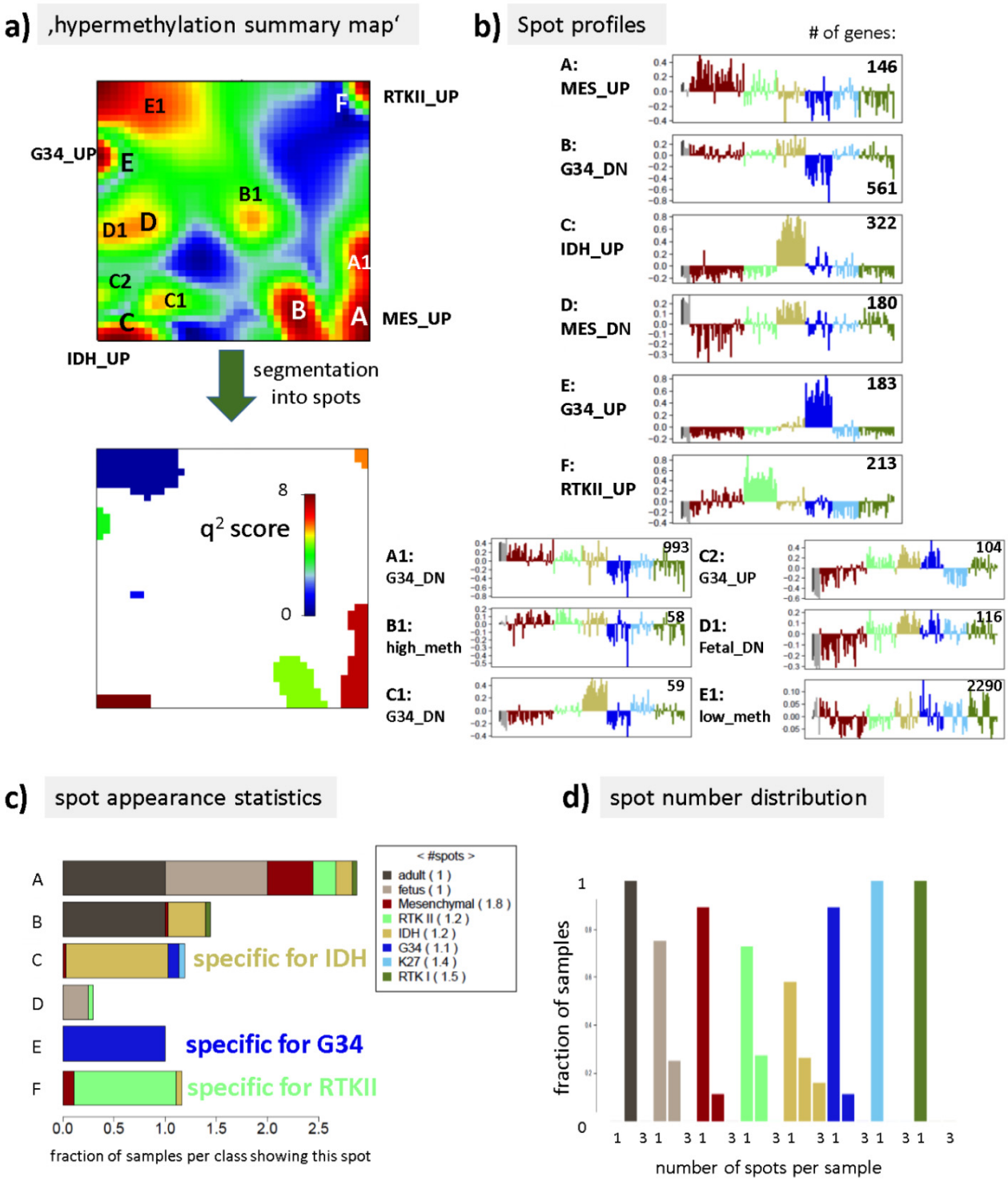


Figure 3: Segmentation of the DmetSOM into spot modules of co-methylated genes: **a**) The hypermethylation summary map indicates regions hypermethylated in any of the classes compared with any other one in red. Each of the 'spot' regions is labeled as indicated. Segmentation of the map provides defined spot regions. Their color codes the q^2 significance score which is minimal for cluster E1. **b**) The methylation spot profiles reveal unique over- or under-methylation of selected classes for the six main spots labeled by capital letters. Six satellite spots show more subtle profiles compared with the respective main spots. Lists of genes in each spot are provided as Table **S1**. **c**) The spot statistics assigns the fraction of samples of each class that shows one of the main spots. A bar length of unity for one subtype means that all samples show this spot. Spots C, E and F are sensitive (nearly each sample of the respective subtype shows this spot) and specific (virtually no other subtypes show this spot) for the IDH, G34 and RTK II subtypes, respectively. **d**) The spot number distributions show that the controls express exclusively one spot. Also most of the GBM samples in each subtype show only one spot. However, also GBM samples with three and even five spots (MES subtype) exist reflecting the increased heterogeneity of their methylation landscapes.

Table 3: Sets of Methylation Marker Genes and their Functional Context

Spot	UP	DN	Functional context: enriched gene sets[1]	top 10 genes[2]
A	MES		olfactory receptor activity (MF); G-protein coupled receptor signaling pathway (BP), neurological systems process (BP), colon cancer: CIMP_methylation_DN, CIMP_expression_UP [32]	ANGPTL1, BCAN, APOC1, TUT1, FADS1, OR4C46, OR11H6, CDH19, GDF5OS, LAMA4
B		G34	extracellular region (CC); keratin filament (CC); colon cancer: CIMP_methylation_DN [32]	PRR33, VIP, FGF17, EMB, USP44, CCR7, HOXB1, LHX5, PRKCD, C1orf64
C	IDH		hallmark epithelial mesenchymal transition (cancer), GCIMP_signature genes: silenced_by_methylation [8]; colon cancer CIMP_methylation_UP [32]; Christensen_methylated_in_LGG [18]; Benporath_H3K27me3_inES [43]; Meissner_brain_HCP_with-H3K4me3_and_H3K27me3 [44], Verhaak_classical_expression_UP [26], brain development (BP)	MT3, SPATA6L, OSBPL1A, TCEA2, MEOX2, ZNF3, L3MBTL4, KIAA0101, TMEM106A, PLLP
D	controls	MES	immune response (BP), cytokine mediated signaling pathway (BP),	TLR4, RTN4, NR2F2, VIM, TMEM140, NMI, PAXIP1−AS2, DHRS4, CISD2, TM4SF18,
E	G34		EED-targets, SUZ12-targets, PRC2-targets, H3K27me3 [43]; RNA-PolI_opening (reactome); meiosis and telomere maintenance (reactome)	INHBB, MORN3, PCDH10, FGGY, LMCD1, DPYSL3, RASD1, MANF, IGFBP7, NAB2
F	RTKII		EED-targets, SUZ12-targets, PRC2-targets, H3K27me3 [43]; H3K27me3 in HCP [41]; Brain HCP with H3K27me3, with H3K4me3 and H3K27me3 [44], develompmental regulators [45]	PCDHAC1, ZSCAN1, GALNT9, ROBO2, CEP126, POPDC3, EXO5, GRIN3A, HSPA1L, KCNB2
A1	controls, MES	G34	Olfactory receptor activity (MF), neurological system process (BP), keratinization (BP)	RPRD2, DPP10, OR51B4, OR8J3, ACSM1, OR6Y1, SPTA1, STX3, CYB5R2, FBLIM1
B1	high methylation		Hallmark bile acid metabolism, Sensory perception of taste (BP), cell-cell junction (CC)	ANKRD7, COX7A2, RGS21, LINC01588, KRTAP21−3. RNASEH2C, C5 SLC13A4, HRH4, NUPR1L
C1	IDH	G34	SUZ 12 targets, PRC2 targets [43]	PTGER4, PAX7, ACVR1C, OTX1, TTI2, TMEM61, IRX4, SPIN1, MOXD1, SLC6A5
C2	G34	control, MES, K27	Cell adhesion (BP), calcium ion binding (MF), EED targets, PRC2 targets, Suz12 targets [43], ES_WITH_H3K27ME3 [44]	HOXC9, FMN1, ATP8B1, ST6GAL1, EVX2, SFTA3, TBX5, GJA3, GAD2, PAX5
D1	IDH	control , MES	Nervous system development (BP), hemophilic cell adhesion (BP), LINDVALL_IMMORTALIZED_BY_TERT_UP	PAX6, FOXB2, VSX1, MKX, COBL, MTA3, PDGFA, ST8SIA4, SH3BP4, C9orf135
E1	low methylation		KIM_myc-targets [46]	EPM2AIP1, ZNF300, KCNH6, SLC35G1, ZNF580, AUNIP, TSHZ3, ZNF311, MIB1, DLL3

[1]enrichment of predefined gene sets in the spot-lists of genes was calculated as described in [27]. Gene sets were taken from literature or from gene ontology (GO) categories biological process (BP) or cellular component (CC). Only gene sets with GSZ-enrichment $p < 10^{-5}$ were taken into account.
[2]Genes are ranked with decreasing correlation coefficient with the spot profile. Full gene lists together with significance measures (p-values of correlation and differential t-tests and false discovery rates) were given in Table **S1**. The lists contain also genes not included in the functional gene sets.

meiosis, respectively. For example, hypomethylation of genes from spot D in MES GBM is related to immune response. It associates with high expression levels of immune response genes in MES GBM [26] suggesting anticorrelation between DNA methylation and expression. Spots C and E are hypermethylated in IDH

and G34 GBM, respectively. They enrich genes supporting the formation of the polycomb repressive complex (PRC2) and also functionally related genes such as *EED* and *SUZ12* targets which control cellular development and differentiation [40]. These processes correlate with repressive and poised chromatin states defined by H3K27me3 and/or H3K4me3 histone marks in brain tissue and stem cells [41, 42]. Sets of affected genes consequently enrich in these spots C and E, as expected. We also find marker gene sets studied in previous DNA methylation and gene expression studies of GBM: For example, methylation markers for GBM of the GCIMP type [8] strongly enrich in the IDH hypermethylation spot C. Interestingly, genes from these spots are hypermethylated also in other cancers such as colorectal cancer (CIMP-type CRC) and B-cell lymphoma (Table **3**).

In summary, spot-segmentation of the SOM of centralized methylation data provides sets of marker genes which are specifically regulated in different glioma subtypes and which are well characterized in terms of previous knowledge. In the following subsections we will address the latter result more in detail.

3.4. Mapping of Previous Sets of Glioma Markers Reveals High Redundancy and Mixtures of Subtypes in the Reference Cohorts

Previous DNA methylation studies on gliomas have published sets of marker genes for different molecular and histological subtypes [6, 8, 13, 18, 47]. We mapped them into the DmetSOM for analysis in terms of gene set maps and profiles (Figure **4**). The genes extracted in Noushmehr *et al*. [8] as 'hypermethylated and deactivated in GCIMP' indeed show clear hypermethylation in the GCIMP IDH subtype also in our data. However one also finds increased methylation of the G34 subtype suggesting a mixture of mainly IDH but also of G34 signature genes. Mapping of the genes of this set into the DmetSOM indeed reveals two regions of high local densities near the signature spot C (for IDH subtype) and E (for G34 subtype).

Christensen *et al*. [18] published a series of signature genes determined as hypermethylated in different groups of low grade gliomas (LGG) relatively to healthy controls including different WHO gradings (II or III) and histological diagnoses (astrocytoma, oligodendroglioma, oligoastrocytoma). All our maps and profiles in Figure **4** except for one show mainly the IDH signature thus indicating a common methylation patterns in LGG independent of WHO grade and histological assignment. The only exception is the methylation signature of primary GBM which can be interpreted as a mixture of IDH and RTKII cases in the respective data. Other authors found the RTKII–signature for GBM-hypermethylation (see the data of Martinez *et al*. [6] in Figure **4** and also [47]). Hence, the resulting 'hypermethylation signature' obviously strongly depends on the composition of the cohort used for extracting marker gene sets. This result agrees with the fact that the incidence of each of the three subtypes RTKII (classical), MES (mesenchymal) and IDH (proneural) in random adult GBM cohorts is roughly comparable [20, 31]. Without stratification into these subtypes one gets consequently a mixture of the respective signatures as observed. Note in this context that that the signature of the mesenchymal subtype is consistently observed as 'hypomethylated' in GBM in a series of gene sets taken from [13, 18] (Figure **4**). Contrarily, the IDH (proneural) cases typically dominate with usually about 80% of all cases in LGGs [21]. The resulting signatures of different LGG strata are consequently close to that of the IDH subtype as observed. We will further discuss this point below in the context of expression signature genes.

To estimate the similarity of different gene sets one usually counts the number of overlapping genes and represents them in terms of Venn diagrams. Note, however, that, for example, the gene set of Noushmehr *et al*. 'hypermethylated in GBM' overlaps with each of the 'hypermethylated in LGG' sets of Christensen *et al*. by only a few genes. The percentage of overlap refers to less than ten percent of the total number of genes in the Noushmehr *et al*. set. On first sight this result suggests the lack of similarity between these sets. Our analysis using gene set mapping however provides the opposite result. We clearly found similar enrichment profiles and enrichment maps of the different sets. It is an important benefit of our method to detect similarities between different marker sets even in the case of a small overlap between them. Such a small overlap between different but similar sets can be simply rationalized by the application of conservative significance thresholds in the selection algorithms for marker genes. High significance levels for differential expression in the original data however can neglect 'still affected' and thus functionally related genes that can become significant in one but not in alternative studies.

In summary, DNA-methylation signature genes from alternative studies of gliomas well agree with our spot

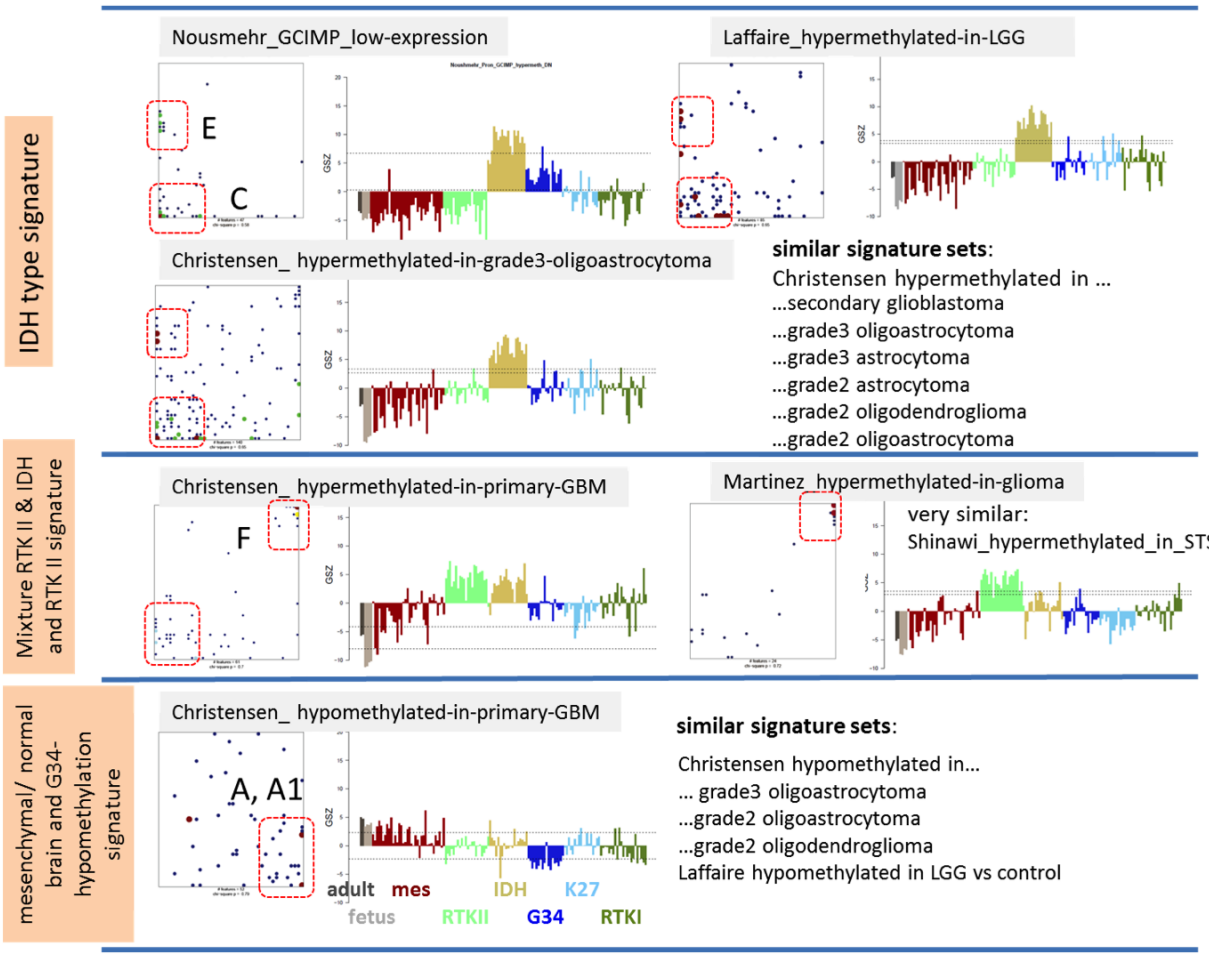


Figure 4: Mapping of methylation marker gene sets for gliomas taken from refs. [6, 8, 13, 18, 47]: The gene set maps show the distribution of marker genes in the DmetSOM. The genes accumulate in different spot areas as indicated by the red dashed frames. The GSZ profiles reveal subtype specific methylation effects. Nearly all sets collecting hypermethylation markers genes show an IDH_UP-signature which partly mixes with the RTKII_UP signature. Sets with very similar signatures are listed without showing the data.

signatures. The IDH (proneuronal) methylation signature dominates in LGG largely independent of WHO grade and histological diagnosis. Especially in GBM the sets reflect mixtures of the subtypes which are present in the cohorts used for extraction of gene sets (typically IDH, classical and mesenchymal). SOM mapping of gene sets robustly identifies similarities between different gene sets even under conditions of noisy compositions. Our approach outperforms overlap-measures as often used in terms of Venn diagrams.

3.5. Marker Sets of B-Cell Lymphomas and Colorectal Cancer Differentiate also between Glioma Classes

We previously found that GCIMP marker genes from glioma studies also differentiate between subtypes of B-cell lymphoma representing a completely different cancer entity [29]. Vice versa, DNA methylation gene sets from previous studies for B cell lymphoma [29] and for colon cancer [32] enrich also in selected spots of the DmetSOM of gliomas studied here (Table **3**). These results motivated us to analyze these sets more in detail using gene set maps and profiles as described in the previous subsection (see Figure **S7**). Genes, hypermethylated in the CIMP-high subtype in CRC and also genes hypermethylated in diffuse large B cell lymphoma (DLBCL) accumulate in spots F and C thus revealing mixed characteristics of the RTKII and IDH subtypes in gliomas. This agreement between different cancers also extends to spots A and B which accumulate genes hypomethylated in G34 gliomas, CRC and also DLBCL compared with Burkitt's lymphoma, another subtype of B cell lymphoma. Hence, the IDH and RTKII subtypes of GBM share similarities with the hyper-methylator phenotypes in CRC and lymphoma. On the other hand the G34 subtype resembles the respective hypomethylator subtypes in lymphoma and CRC.

These striking agreements suggest general mechanisms of aberrant DNA methylation in different cancer entities.

3.6. Gene Expression and Promoter Methylation Change Mostly in Anti-Concert

In the next step we analyzed the association between gene expression and DNA methylation of the spot genes using matched samples taken from [5] and also independent expression data [20, 26] for which we matched the classes with the methylation data studied here (see Figure **5a** and Table **2**). The hypermethylation spots of the MES (spot A), IDH (C) and RTKII (F) subtypes consistently reveal strong anti-correlation between promoter methylation and gene expression in all three analyses. The same result was obtained for the G34 subtype in the matched sample data. The independent GBM expression data do not

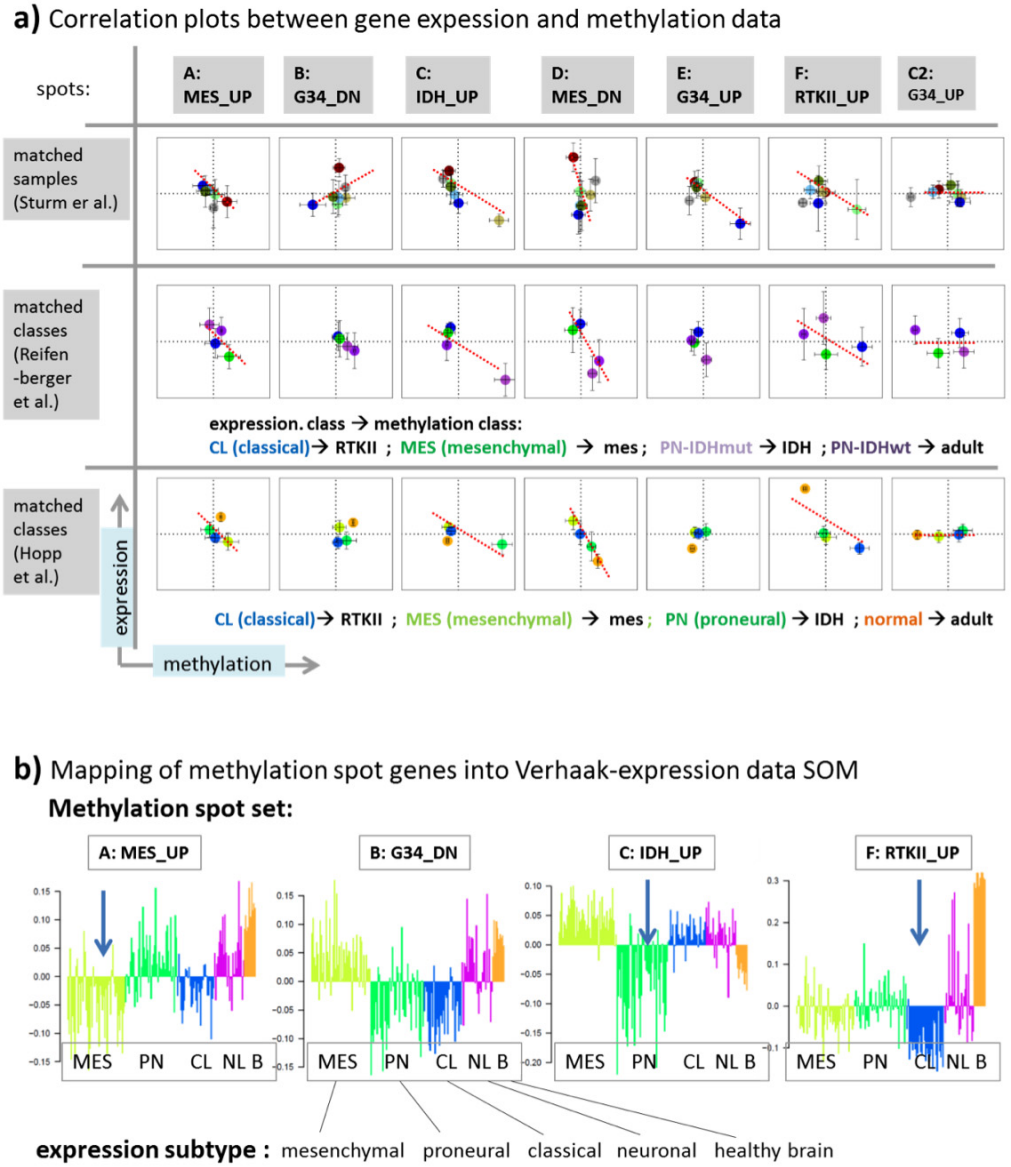


Figure 5: Correlation between DNA methylation and gene expression: **a**) Correlation plots between matched DNA methylation and gene expression data of the spot genes reveal preferentially anti-correlated changes as indicated by the red dotted lines which serve as a guide for the eye. We used matched samples where methylation and expression data are known for the same samples taken from [5] and also matched classes taken from [20] and [26] where the data refer to different samples. The matching rules for the classes are given within the figure. Each full circle provides the mean values for one subtype. The error bars in abscissa and oordinate direction indicate the variance of methylation and expression data for each subtype, respectively. **b**) Gene expression profiles of the methylation spot sets in the GBM expression data analyzed in [26]: Hypermethylation sets (MES_UP, IDH_UP, RTKII_UP) associate with underexpression in the respective subtype as indicated by the arrows. Note that the color code for the GBM subtypes was chosen from the original papers [20, 26].

show this effect because they do not contain pediatric cases. For other spots one observes the absence of systematic expression changes despite marked methylation effects (spot C2), positive correlations (B) and also neither marked expression nor methylation effects for the hyper- and hypomethylation spots B1 and E1, respectively (data not shown).

In Figure **5b** we explicitly show the expression profiles of the genes from selected methylation spot sets in the Verhaak-reference data set as analyzed in [26]. One clearly sees that hypermethylation of the promoters of the selected genes in a selected subtype accompanies strong downregulation of gene expression of these genes. Importantly, the expression profiles respond in a subtype-specific fashion. This result reflects the important fact that the methylation classes show also class-specific expression effects and thus a close mutual relation between gene expression and DNA methylation.

To further proof this relation we mapped gene expression marker sets for LGGs (WHO grade II and III) and GBM (grade IV) into the DmetSOM to estimate their DNA methylation status (Figure **S8**). In general, we found strong subtype-specific effects thus confirming the close relation between expression and methylation. For example, LGGs with a co-deletion on chromosomes 1 and 19 as a hallmark of oligodendroglioma show the RTKII hypermethylation signature (Figure **S8a**). Grade II and III LGGs differ in the methylation level of RTKII and IDH signature genes on one hand and of G34 signature genes on the other hand. Hence, we again found a mixing between different methylation classes in the subcohorts selected. The expression classes proposed by Gorovets *et al*. [19] for LGGs can be assigned to a brain-like_UP methylation signature (neuroblastic LGG), a mixed RTKII and IDH signature (early progenitor LGGs) and an IDH_UP signature (pre-glioblastoma, PG; see Figure **S8a**). Note that genes hypermethylated in IDH tumors (IDH_UP) are on low expression level in IDH1 mutated tumors but on high level in IDH1 wild type tumors such as PG. Hence, hypermethylation signatures of IDH1 mutated tumors correspond to overexpression signatures of IDH1 wild type tumors and vice versa due to the anti-correlation between expression and methylation effects.

This anti-concerted assignment of methylation and expression signatures is evident also in the expression signatures of GBM (Figure **S8b**): Genes, overexpressed in IDH1 wild type tumors of the mesenchymal and/or classical subtypes are mostly hypermethylated in IDH1 mutated,-proneural tumors. A sketchy use of terms like 'IDH_UP-signature' can imply incorrect associations because 'over'-methylation in the IDH subtype associates with 'over'-expression of another one, namely in IDH1 wild type mesenchymal and/or classical subtypes.

Please note also, that deactivation of gene expression by DNA methylation of gene promoters represents only one possible mechanism how DNA methylation affects transcription. Alternative mechanisms are discussed which, for example, explain also correlated changes between gene expression and DNA methylation. For example, a methylated DNA sequence motif can take on a new function by creating a novel DNA binding site for transcriptional activators that could not be predicted from sequence information alone. Such mechanisms expand the functional role of DNA methylation in gene regulation, being capable to regulate active and repressive gene states in a site-specific manner [48].

3.7. Methylation of Glioma Subtypes Associates with Cellular Programs and their (de-)Activation by Chromatin Remodeling

Functional analysis of the spot lists of genes revealed specific functional modes and states of gene activity which associate with the different sets of markers and thus also with the methylation subtype (Table **3**). To study the biological context more in detail we generated one-way clustered heatmaps of gene sets referring to the GO-category 'biological process' (Figure **6a**), to chromatin states of brain tissue (Figure **6b**), to regulators in poorly differentiated cells [43], to repressive, poised and active histone methylation states [41] (see Figure **S9**) and also special gene sets with notably profiles (Figure **S10**).

Firstly, one finds two 'limiting' profiles characterized by (i) high methylation of the brain-like classes and low methylation of CHOP-like classes and (ii) by the respective antagonistic CHOP-like_UP/brain-like_DN profile. The former profile comprises functions like 'neurological systems process', 'immune response' (Figure **6a**), transcription factors (TF) associated with low expression levels in mammalian cells in general [49], fatty acid metabolism (Figure **S10**) and partly transcriptional active chromatin states (Figure **6b**). These profiles are characterized by strong hypomethylation of G34, K27 and RTKI GBM compared with the other GBM subtypes and also

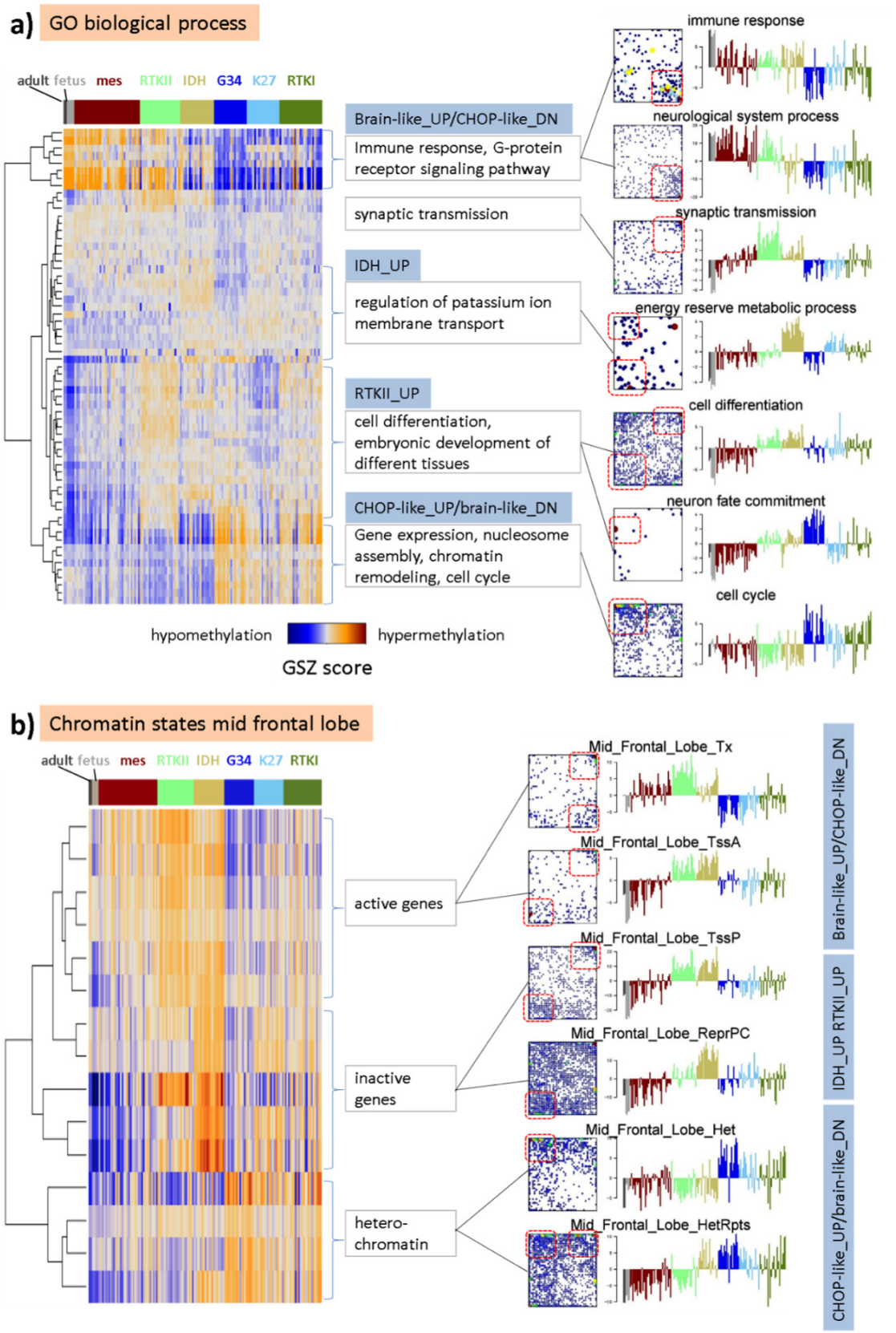


Figure 6: Methylation heatmap of genes referring to the GO term biological process (part **a**), and genes assigned to different chromatin states in healthy brain (mid frontal lobe) (part **b**). Colors maroon to blue indicate high to small methylation levels, respectively. Chromatin states were grouped into active ones (e.g. Tx, Txn, TSSA), inactive (ReprPC, Quies, TSSP) and closed/heterochromatin (Het, HetRpts, ZNF) roughly agreeing with the clustering of methylation patterns shown in the right part of the figure.

healthy brain. The second types of profiles (ii) are associated with high methylation levels of 'cell cycle' (Figure **6a**), ribosomal, mitochondrial genes, high transcription TFs, hypoxia, DNA repair and ageing, partly EZH2 targets (Figure **S10**), MYC-, NOTCH- and SOX2-targtes (Figure **S9a**) and heterochromatin states (Figure **6b**) in brain-like classes and low methylation in CHOP. These two groups (i) and (ii) of antagonistic DNA-methylation are mainly responsible for the two superclusters established in the similarity plots (Figure **2a** and **b**).

Note that group (ii) associates with highly methylated genes that enrich in and near spot E1. Recall that high methylation levels correlate mostly with low gene activities. Hence, high transcription TFs are repressed by DNA methylation in group (ii) and packed into closed chromatin states whereas lower methylation levels associate with active chromatin states. The situation reverses in group (i) where low transcription TFs and active chromatin states become repressed by high methylation levels.

In between these two 'limiting' states one finds a third type of profiles (iii) with uniquely high methylation in RTKII, IDH or G34 and also mixtures of them. These states enrich inactive chromatin states with repressed and/or poised promoters, developmental and tissue-differentiation genes and targets of the polycomb repressive complex 2 (PRC2) and related genes: targets of *EED*, *SUZ12* (Figure **S9a**) and *EZH2*, the catalytic subunit of a H3K27 methyltransferase (Figure **S10**). These results are supported by histone modification data which show that type (iii) profiles associate with repressive H3K27me3 and bivalent H3K27me3 and H3K4me3 marks (Figure **S9b**). These data also show that so-called high CpG promoters are mainly involved in repression of these genes whereas repressed low CpG promoters associate partly with type (i) brain-like_UP methylation profiles.

Interestingly, G34 tumors associate with strong hypermethylation of genes related to promoter opening and telomere end packing (Figure **S10**). Pediatric GBM and especially G34 tumors show alternative lengthening of telomeres (ALT) mediated by homologous recombination and supported by mutations of the *ATRX* gene which mediate histone assembly in subtelomeric regions [30]. We found strong hypermethylation of genes coding histone clusters 1 and partly also clusters 2 and 3 thus suggesting aberrant expression and in final consequence aberrant nucleosome assembly and aberrant telomere maintenance function.

Hence, IDH, G34 and also RTKII are characterized by DNA methylation and thus transcriptional repression of genes which obviously suppress tumorigenesis in healthy brain. Hypermethalytion of PRC2 repressed targets and of poised promotors is a molecular hallmark of many cancer types [43] including B-cell lymphomas [7, 29] and colorectal cancer. This ubiquitous property partly explains the similar signatures of high CpG methylator phenotypes in gliomas, colon cancer and lymphomas. This agreement is further supported by overlapping chromatin states in the healthy tissues: Especially genes with poised promoter states (TSSP) agree to about 50% (of about 3000 genes) between brain tissue and colon and brain and lymphoblastoid cells as well.

These results show that methylation effects associate with different chromatin states which, in turn, enable different modes of gene activity in terms of transcriptional programs. Global hypermethylation of the brain-like and IDH subtypes and global hypomethylation of the CHOP-like subtypes associates with open chromatin states which are either transcriptional active in the RTKII and also mesenchymal subtypes or inactive in the IDH and partly RTKII and G34 subtypes. Methylation of closed chromatin counteracts the global net methylation tendencies, i.e. it is associated with reduced methylation in the brain-like and IDH subtypes and increased methylation in the CHOP-like subtypes. Note that the assignment of chromatin states is based on healthy brain data (mid frontal lobe) which presumably only partly can be applied to the diseased brain. Hence, methylation effects associate with changes of the chromatin states, for example if highly methylated nominal active promoter states transform into inactive ones or even into heterochromatin.

4. DISCUSSION

4.1. SOM Portrayal of Marker Sets Resolve Heterogeneity of DNA Methylation Across Glioma Subtypes, Cancer Entities and Different Cohorts

Our study focused on DNA methylation data stratified with respect to molecular subtypes of adult and pediatric GBM and healthy brain controls. We applied SOM machine learning to the data, a powerful technique to 'organize' complex, multivariate data.

Using centralized methylation data we identified clusters of co-methylated genes among the samples studied which we call 'spot-modules' because of their spot-like appearance in the SOM-portraits. The DmetSOM disentangles genes systematically hyper- and hypomethylated in gliomas compared with healthy brain and it extracts systematic methylation differences between the glioma subtypes. SOM portrayal was shown to serve as an effective 'sorting machine' to extract different modes of DNA methylation in gliomas. To assign the functional meaning to the spot modules we applied enrichment analysis using a multitude of pre-defined gene sets related to categories such as biological function (e.g. inflammation, cell development and ageing), targets of different transcription factors (e.g. *MYC*, *NANOG*, high and low expression TFs) and epigenetic modulators (e.g. *EED*, *SUZ12*; PRC2*, EZH2*), different chromatin states in reference mid frontal lobe tissue, genes differently methylated in other cancers (e.g. CIMP in colorectal cancer and methylation subtypes of B-cell lymphoma) and also of marker gene sets for differential methylation and expression between glioma subtypes obtained in independent studies.

Interestingly, we found pronounced subtype-specific methylation signatures of gene sets from different glioma studies which indicate a common scheme of aberrant gene regulation in LGGs and adult and pediatric GBM. The GCIMP signature is found across most of the glioma studies as a basal hallmark of IDH1-mutated tumors. However, our analysis finds also mixed methylation signatures in many cases especially for histological classes which often represent mixtures of different molecular subtypes. Hence, methylation signatures enable the further 'de-mixing' of histological classes according to molecular variants. This result supports recent studies showing that DNA-based molecular profiling of gliomas distinguishes biologically distinct tumor groups and provides prognostically relevant information beyond histological classification [21]. On the other hand, molecular profiling is hardly suited for reliable distinction of tumor grades due to grade-independent mechanisms.

We also found pronounced correlation between gene expression and methylation signatures of gliomas. It reflects coupled mechanisms of methylation and gene activity. Whether DNA methylation profiling provides a more robust and clinically useful platform for GBM subgrouping remains to be tested. The enrichment of DNA methylation signatures of other cancer entities in gliomas suggests general oncogenic methylation mechanisms.

4.2. Methylation Marker Sets Reveal Molecular Mechanisms of Gliomas

DNA methylation acts as an epigenetic modification in vertebrate DNA. It has become clear that the DNA and histone lysine methylation systems are highly interrelated and rely mechanistically on each other for normal chromatin function [50]. Controlling the timing and placement of DNA methylation in the genome is essential for normal cellular function and its dysfunction de-regulates cell activities. Figure **7** summarizes the main results of our study by relating methylation profiles, glioma subtypes, biological functions and chromatin states each to another.

The global methylation profile, namely hypermethylation in brain-like and GCIMP tumors and hypomethylation in pediatric GBM and RTKI, associates with the biological processes immune response and fatty acid metabolism. This mode is counterbalanced by antagonistic methylation changes which can be assigned to cell cycle activity and energy metabolism. In healthy brain these modes accumulate genes from different chromatin states, namely transcribed states and silent heterochromatin, respectively. This result suggests that transcribed states in healthy brain become suppressed by DNA hyper-methylation in brain-like and GCIMP subtypes whereas silent heterochromatin becomes possibly activated due to hypomethylation of the affected genes in these tumors. In contrast, methylation levels in the CHOP-like pediatric GBM and RTKI correspond to the chromatin states assigned in the healthy brain. These results suggest chromatin remodeling between brain-like and GCIMP on one hand and CHOP-like tumors on the other hand. In other words, global methylation effects seem to associate with a different chromatin organization in the methylation superclusters.

These global changes were further modulated by a series of methylation effects which refer to only a few or even single subtypes and thus define their specificity. We found hypermethylation of genes normally activated in stem cells, combined with preferential repression of polycomb-regulated genes (PRC2-, *EED*- and *SUZ12*-targets) in RTKII, IDH and also G34 tumors. These genes are enriched in chromatin states assigned to repressed and bivalent promoters with H3K27me3 or H3K27me3 and

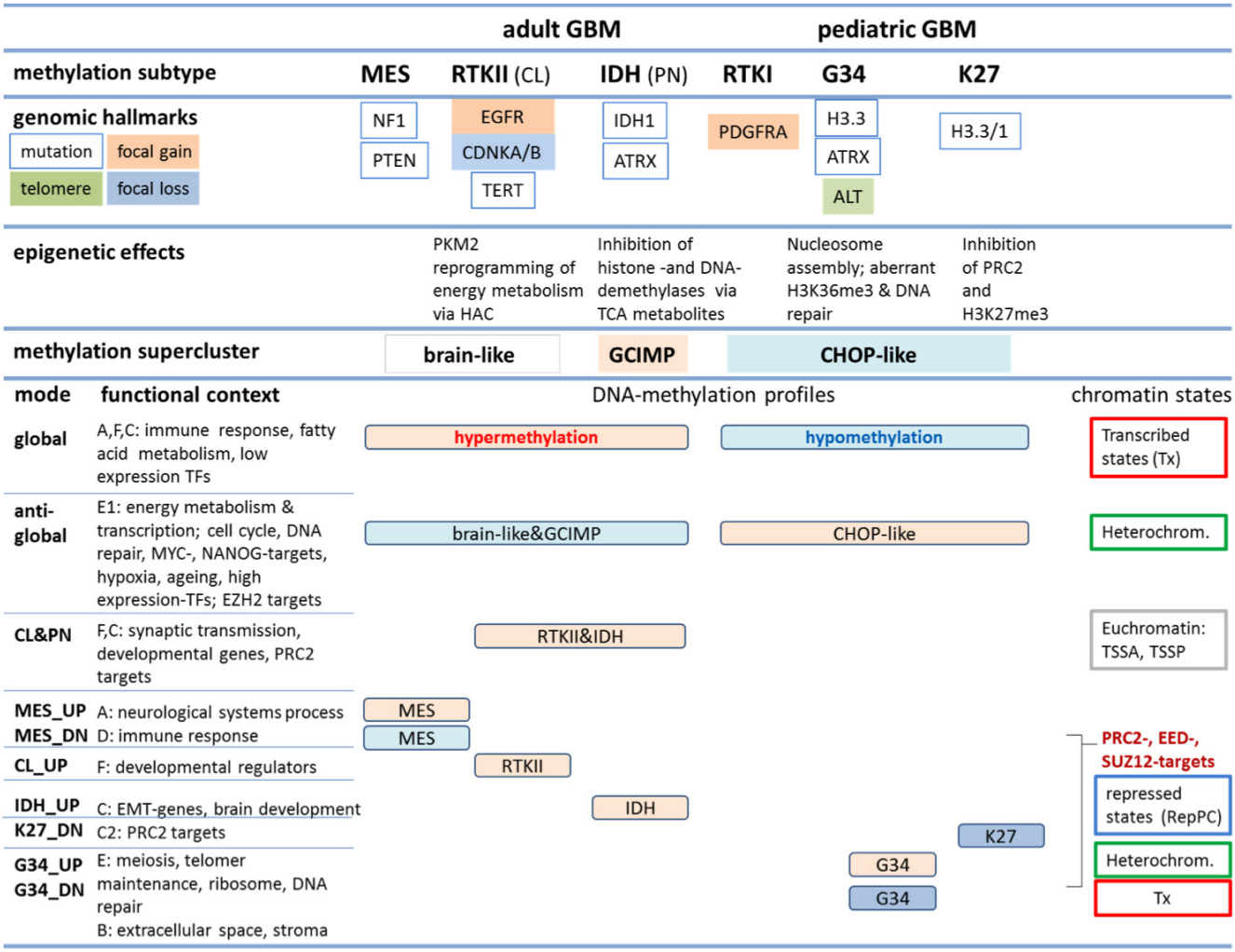


Figure 7: Overview scheme summarizing genomic hallmarks of adult and pediatric glioma subtypes and epigenetic mechanisms and regulatory modes of promotor methylation and gene activity extracted from our analysis. The functional context associates with the spot clusters of genes obtained from DmetSOM analysis. The chromatin states refer to healthy mid frontal lobe tissue. Their assignment to the regulatory modes suggests specific targets for DNA methylation: For example, transcribed states in healthy brain are prone to global hypermethylation in brain-like and GCIMP tumors and prone to global hypomethylation in CHOP-like states. The antagonistic mode of methylation affects mainly heterochromatin in healthy brain. Note that promoter methylation mostly anticorrelates with gene activity: e.g., energy metabolism becomes upregulated in brain-like and GCIMP tumors compared with CHOP-like ones. Targets of the PRC2 complex and of its cofactors (*EED* and *SUZ12*) are hypermethylated and thus transcriptionally repressed in RTKII, IDH and G34 but activated in MES and K27 tumors by hypomethylation.

H3K4me3 marked histones, respectively. This methylation signature is generally found in poorly differentiated tumors [43] and, for example, also in B cell lymphoma [7, 29] indicating 'suppression of tumor suppressors' associated with tissue-specific cell differentiation [51, 41, 52]. Interestingly, targets of TFs involved in development and differentiation (*OCT4*, *NANOG*, *SOX2*) and also *MYC* are antagonistically methylated compared with the PRC2 targets thus suggesting different regulatory modes for more repressed and more active genes, respectively. A similar dualism was previously suggested in terms of high and low transcription TFs in metazoan which associate with high and low gene expression levels of their targets, respectively [49]. The split between both types of TFs was recently established also in lymphomas [29]. The high transcription TFs show generally a low DNA methylation level in the brain-like and GCIMP tumors. Contrarily, low transcription TFs are associated with high methylation levels reflecting the expected anticorrelated activation pattern between methylation and gene expression. In CHOP-like tumors this relation however reverses showing hypermethylation of high expression TFs and hypomethylation of low expression TFs and thus apparently improper expression levels in these tumors which possibly reflects chromatin remodeling as discussed above.

Also K27 tumors show enrichment of PRC2 target genes becoming however hypomethylated in terms of DNA methylation and low levels of repressive H3K27me3 histone marks as well. We conclude that these genes are transcriptionally more active in K27 gliomas compared with the other subtypes in agreement with [53, 54]. Hence, the Lys27 mutation of H3.3 associates with the global reduction of repressive histone marks of the H3K27me3 type, activation of gene expression and DNA de-methylation.

Aberrant hypermethylation in IDH tumors of the GCIMP type is induced mostly by the IDH1 mutation leading to inhibition of histone-lysine- and DNA de-methylases carrying the Jumonji-domain *via* intermediate metabolites of the citrate cycle which act as their coenzymes [55]. In tumors of the RTK-types epigenetic dysregulation associates also with metabolic reprogramming, namely with aberrant activation of the pyruvate kinase M2 (*PKM2*) isoform, a glycolytic enzyme involved in ATP generation and pyruvate production which plays an essential role in tumor metabolism and growth. It also functions as a protein kinase that phosphorylates and/or acetylates histones during transcription and chromatin remodeling with consequences for CpG methylation [56]. RTKI tumors together with K27 and G34 show hypermethylation of genes related to pyruvate metabolism, ATP binding, mitochondrion and ribosome cellular components suggesting transcriptional down regulation of the energy metabolism and protein synthesis. Interestingly, targets of *EZH2*, a compound of PRC2 catalyzing the formation of H3K27me3, show similar methylation profiles which also resemble those of genes up-regulated upon ageing and under hypoxia. Subtle differences between the methylation profiles 'pyruvate metabolism' and 'ATP binding'/ 'mitochondrion' in IDH gliomas on one hand and G34 and RTKI on the other hand however suggest different mechanisms of metabolic control in these subtypes.

G34 tumors show specific hypermethylation of genes associated with telomere length maintenance (Reactome sets 'packaging of telomere lengths' and 'pol I promoter opening'), histone assembly and DNA repair suggesting increased genomic instability of this subtype. G34 tumors display an ALT (alternative lengthening of telomeres) phenotype presumably mediated by homologous recombination and caused by the mutation of the *ATRX* gene and possibly also by the G34 mutation of H3.3 itself [5, 57]. Aberrant DNA repair functionality in G34 is possibly associated with DNA hypermethylation of the respective genes and aberrant methylation markings of H3K36me3 required for proper recruitment of the DNA-repair machinery [30, 58, 59]. Interestingly we find also strong hypermethylation of genes referring to ribosome and mitochondrial functions in G34 suggesting a deactivation of transcriptional and energy-metabolic processes in this subtype.

Taken together, these findings illustrate a widespread functional role of DNA methylation in gene regulation in gliomas essentially contributing to the heterogeneity of glioma subtypes and strongly affecting the underlying molecular mechanisms of cell function.

5. CONCLUSIONS

Sets of differential methylation genes in gliomas represent surrogate markers of molecular mechanisms governing (epi)genomic dysregulation. DNA methylation phenomena are complex ensuring complex tuning of gene function. Consideration of this regulatory level is inevitable for understanding cancer genesis and progression. It provides suited markers for diagnosis of glioma subtypes and disentangles tumor heterogeneity.

ACKNOWLEDGEMENTS

This work is supported by the Federal Ministry of Education and Research (BMBF), project grant No. FKZ 031 6166 (MMML-MYC-SYS), FKZ 031 6065A (HNPCCSys), SysGlio and the DKTK joint funding project "Next generation molecular diagnostics of malignant gliomas" and the German Glioma Network.

AUTHOR CONTRIBUTIONS

LH and HB conceived this study, performed all analyses and wrote the paper. HLW contributed the program for SOM analyses and EW performed part of gene set analyses.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

The supplementary material can be downloaded from the journal website along with the article.

REFERENCES

[1] Chibon F. Cancer gene expression signatures – The rise and fall? European Journal of Cancer 2013; 49: 2000-2009. http://dx.doi.org/10.1016/j.ejca.2013.02.021

[2] Quackenbush J. Microarrays--Guilt by Association. Science 2003; 302: 240-241. http://dx.doi.org/10.1126/science.1090887

[3] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Lander ES. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; 286: 531-537. http://dx.doi.org/10.1126/science.286.5439.531

[4] Kulis M, Esteller M. 2 - DNA Methylation and Cancer. In Advances in Genetics; Zdenko, H, Toshikazu, U, Eds, Academic Press 2010; Vol. 70: pp. 27-56.

[5] Sturm D, Witt H, Hovestadt V, Khuong-Quang D-A, Jones David TW, Konermann C, Pfister Stefan M. Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. Cancer Cell 2012; 22: 425-437. http://dx.doi.org/10.1016/j.ccr.2012.08.024

[6] Martinez R, Martin-Subero JI, Rohde V, Kirsch M, Alaminos M, Fernández AF, Esteller M. A microarray-based DNA methylation study of glioblastoma multiforme. Epigenetics 2009; 4: 255-264. http://dx.doi.org/10.4161/epi.9130

[7] Martin-Subero JI, Ammerpohl O, Bibikova M, Wickham-Garcia E, Agirre X, Alvarez S, Siebert R. A Comprehensive Microarray-Based DNA Methylation Study of 367 Hematological Neoplasms. PLOS One 2009; 4: e6986. http://dx.doi.org/10.1371/journal.pone.0006986

[8] Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Aldape K. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. Cancer Cell 2010; 17: 510-522. http://dx.doi.org/10.1016/j.ccr.2010.03.017

[9] Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa J-PJ. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA 1999; 96: 8681-8686. http://dx.doi.org/10.1073/pnas.96.15.8681

[10] Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, Melnick A. DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia. Cancer Cell 2010; 17: 13-27. http://dx.doi.org/10.1016/j.ccr.2009.11.020

[11] Brennan, Cameron W, Verhaak Roel GW, McKenna A, Campos B, Noushmehr H, Salama Sofie R, McLendon R. The Somatic Genomic Landscape of Glioblastoma. Cell 155: 462-477. http://dx.doi.org/10.1016/j.cell.2013.09.034

[12] Colman H, Zhang L, Sulman EP, McDonald JM, Shooshtari NL, Rivera A, Aldape K. A multigene predictor of outcome in glioblastoma. Neuro-Oncology 2010; 12: 49-57. http://dx.doi.org/10.1093/neuonc/nop007

[13] Laffaire J, Everhard S, Idbaih A, Crinière E, Marie Y, de Reyniès A, Ducray F. Methylation profiling identifies 2 groups of gliomas according to their tumorigenesis. Neuro-Oncology 2011; 13: 84-98. http://dx.doi.org/10.1093/neuonc/noq110

[14] Kim Y-W, Koul D, Kim SH, Lucio-Eterovic AK, Freire PR, Yao J, Yung WKA. Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. Neuro-Oncology 2013; 15: 829-839. http://dx.doi.org/10.1093/neuonc/not024

[15] Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Louis DN. Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification. Cancer Research 2003; 63: 1602-1607.

[16] Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, Aldape K. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 2006; 9: 157-173. http://dx.doi.org/10.1016/j.ccr.2006.02.019

[17] Dang L, Jin S, Su SM. IDH mutations in glioma and acute myeloid leukemia. Trends in Molecular Medicine 2010; 16: 387-397. http://dx.doi.org/10.1016/j.molmed.2010.07.002

[18] Christensen BC, Smith AA, Zheng S, Koestler DC, Houseman EA, Marsit CJ, Wiencke JK. DNA Methylation, Isocitrate Dehydrogenase Mutation, and Survival in Glioma. Journal of the National Cancer Institute 2011; 103: 143-153. http://dx.doi.org/10.1093/jnci/djq497

[19] Gorovets D, Kannan K, Shen R, Kastenhuber ER, Islamdoust N, Campos C, Huse JT. IDH Mutation and Neuroglial Developmental Features Define Clinically Distinct Subclasses of Lower Grade Diffuse Astrocytic Glioma. Clinical Cancer Research 2012; 18: 2490-2501. http://dx.doi.org/10.1158/1078-0432.CCR-11-2977

[20] Reifenberger G, Weber RG, Riehmer V, Kaulich K, Willscher E, Wirth H, Glioma N. Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. International Journal of Cancer 2014; 135: 1822-1831. http://dx.doi.org/10.1002/ijc.28836

[21] Weller M, Weber R, Willscher E, Riehmer V, Hentschel B, Kreuz M, Reifenberger G. Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome- and transcriptome-wide profiling improves stratification of prognostically distinct patient groups. Acta Neuropathologica 2015; 1-15.

[22] Brulard C, Chibon F. Robust gene expression signature is not merely a significant P value. European Journal of Cancer 2013; 49: 2771-2773. http://dx.doi.org/10.1016/j.ejca.2013.03.033

[23] Venet D, Dumont JE, Detours V. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 2011; 7: e1002240. http://dx.doi.org/10.1371/journal.pcbi.1002240

[24] Wirth H, Loeffler M, von Bergen M, Binder H. Expression cartography of human tissues using self organizing maps. BMC Bioinformatics 2011; 12: 306. http://dx.doi.org/10.1186/1471-2105-12-306

[25] Hopp L, Lembcke K, Binder H, Wirth H. Portraying the Expression Landscapes of B-Cell Lymphoma- Intuitive Detection of Outlier Samples and of Molecular Subtypes. Biology 2013; 2: 1411-1437. http://dx.doi.org/10.3390/biology2041411

[26] Hopp L, Wirth H, Fasold M, Binder H. Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. Systems Biomedicine 2013; 1. http://dx.doi.org/10.4161/sysb.25897

[27] Wirth H, von Bergen M, Binder H. Mining SOM expression portraits: Feature selection and integrating concepts of molecular function. BioData Mining 2012; 5: 18. http://dx.doi.org/10.1186/1756-0381-5-18

[28] Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010; 11: 587-587. http://dx.doi.org/10.1186/1471-2105-11-587

[29] Hopp L, Wirth-Loeffler H, Binder H. Epigenetic heterogeneity of B-cell lymphoma: DNA-methylation, gene expression and chromatin states. Genes 2015; in press.

[30] Sturm D, Bender S, Jones DTW, Lichter P, Grill J, Becher O, Pfister SM. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. Nat Rev Cancer 2014; 14: 92-107. http://dx.doi.org/10.1038/nrc3655

[31] Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Hayes DN. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17: 98-110. http://dx.doi.org/10.1016/j.ccr.2009.12.020

[32] Binder H, Hopp L, Lembcke K, Wirth H. Personalized Disease Phenotypes from Massive OMICs Data. In Big Data Analytics in Bioinformatics and Healthcare; Baoying, W, Ruowang, L, William, P, Eds, IGI Global: Hershey, PA, USA, 2015; pp. 359-378. http://dx.doi.org/10.4018/978-1-4666-6611-5.ch015

[33] Wirth-Loeffler H, Kalcher M, Binder H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on Bioconductor. Bioinformatics 2015; in revision.

[34] Binder H, Wirth H, Arakelyan A, Lembcke K, Tiys ES, Ivanishenko V, Larina IM. Time-course human urine proteomics in space-flight simulation experiments. BMC Genomics 2014; 15: S2. http://dx.doi.org/10.1186/1471-2164-15-S12-S2

[35] Toronen P, Ojala P, 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. http://dx.doi.org/10.1186/1471-2105-10-307

[36] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005; 102: 15545-15550. http://dx.doi.org/10.1073/pnas.0506580102

[37] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotech 2010; 28: 817-825. http://dx.doi.org/10.1038/nbt.1662

[38] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011; 473: 43-49. http://dx.doi.org/10.1038/nature09906

[39] Läuter J, Glimm E, Eszlinger M. Search for relevant sets of variables in a high-dimensional setup keeping the familywise error rate. Statistica Neerlandica 2005; 59: 298-312. http://dx.doi.org/10.1111/j.1467-9574.2005.00290.x

[40] Walker E, Manias JL, Chang WY, Stanford WL. PCL2 modulates gene regulatory networks controlling self-renewal and commitment in embryonic stem cells. Cell Cycle 2011; 10: 45-51. http://dx.doi.org/10.4161/cc.10.1.14389

[41] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007; 448: 553-560. http://dx.doi.org/10.1038/nature06008

[42] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Lander ES. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008; 454: 766-770.

[43] Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 2008; 40: 499-507. http://dx.doi.org/10.1038/ng.127

[44] Meissner A. Epigenetic modifications in pluripotent and differentiated cells. Nat Biotech 2010; 28: 1079-1088. http://dx.doi.org/10.1038/nbt.1684

[45] Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Young RA. Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. Cell 2006; 125: 301-313. http://dx.doi.org/10.1016/j.cell.2006.02.043

[46] Kim YH, Girard L, Giacomini CP, Wang P, Hernandez-Boussard T, Tibshirani R, Pollack JR. Combined microarray analysis of small cell lung cancer reveals altered apoptotic balance and distinct expression signatures of MYC family gene amplification. Oncogene 2005; 25: 130-138.

[47] Shinawi T, Hill VK, Krex D, Schackert G, Gentle D, Morris MR, Latif F. DNA methylation profiles of long- and short-term glioblastoma survivors. Epigenetics 2013; 8: 149-156. http://dx.doi.org/10.4161/epi.23398

[48] Rothbart SB, Strahl BD. Interpreting the language of histone and DNA modifications. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 2014; 1839: 627-643. http://dx.doi.org/10.1016/j.bbagrm.2014.03.001

[49] 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.

[50] Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 2014; 1839: 1362-1372. http://dx.doi.org/10.1016/j.bbagrm.2014.02.007

[51] Li G, Warden C, Zou Z, Neman J, Krueger JS, Jain A, Chen M. Altered Expression of Polycomb Group Genes in Glioblastoma Multiforme. PLOS One 2013; 8: e80970. http://dx.doi.org/10.1371/journal.pone.0080970

[52] Watson CT, Disanto G, Sandve GK, Breden F, Giovannoni G, Ramagopalan SV. Age-Associated Hyper-Methylated Regions in the Human Brain Overlap with Bivalent Chromatin Domains. PLOS One 2012; 7: e43840. http://dx.doi.org/10.1371/journal.pone.0043840

[53] Voigt P, Reinberg D. Putting a halt on PRC2 in pediatric glioblastoma. Nat Genet 2013; 45: 587-589. http://dx.doi.org/10.1038/ng.2647

[54] Epigenetic Dysregulation Promotes Gene Activation in Pediatric Glioma. Cancer Discovery 2013; 3: OF15.

[55] Xiao M, Yang H, Xu W, Ma S, Lin H, Zhu H, Guan K-L. Inhibition of α-KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors. Genes & Development 2012; 26: 1326-1338. http://dx.doi.org/10.1101/gad.191056.112

[56] Chen L, Shi Y, Liu S, Cao Y, Wang X, Tao Y. PKM2: The Thread Linking Energy Metabolism Reprogramming with Epigenetics in Cancer. International Journal of Molecular Sciences 2014; 15: 11435-11445. http://dx.doi.org/10.3390/ijms150711435

[57] Kannan K, Inagaki A, Silber J, Gorovets D, Zhang J, Kastenhuber ER, Huse JT. Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma 2012; Vol. 3.

[58] Pfister Sophia X, Ahrabi S, Zalmas L-P, Sarkar S, Aymard F, Bachrati Csanád Z, Humphrey Timothy C. SETD2-Dependent Histone H3K36 Trimethylation Is Required for Homologous Recombination Repair and Genome Stability. Cell Reports 2014; 7: 2006-2018. http://dx.doi.org/10.1016/j.celrep.2014.05.026

[59] Pai C-C, Deegan RS, Subramanian L, Gal C, Sarkar S, Blaikley EJ, Humphrey TC. A histone H3K36 chromatin switch coordinates DNA double-strand break repair pathway choice. Nat Commun 2014; 5.

[60] 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: 883-891. http://dx.doi.org/10.1038/nature02661

[61] Lee C-K, Klopp RG, Weindruch R, Prolla TA. Gene Expression Profile of Aging and Its Retardation by Caloric Restriction. Science 1999; 285: 1390-1393. http://dx.doi.org/10.1126/science.285.5432.1390

[62] Winter SC, Buffa FM, Silva P, Miller C, Valentine HR, Turley H, Harris AL. Relation of a Hypoxia Metagene Derived from Head and Neck Cancer to Prognosis of Multiple Cancers. Cancer Research 2007; 67: 3441-3449. http://dx.doi.org/10.1158/0008-5472.CAN-06-3322

Received on 18-08-2015 Accepted on 25-08-2015 Published on 26-11-2015

DOI: http://dx.doi.org/10.6000/1929-2279.2015.04.04.1

SUPPLEMENTARY MATERIAL

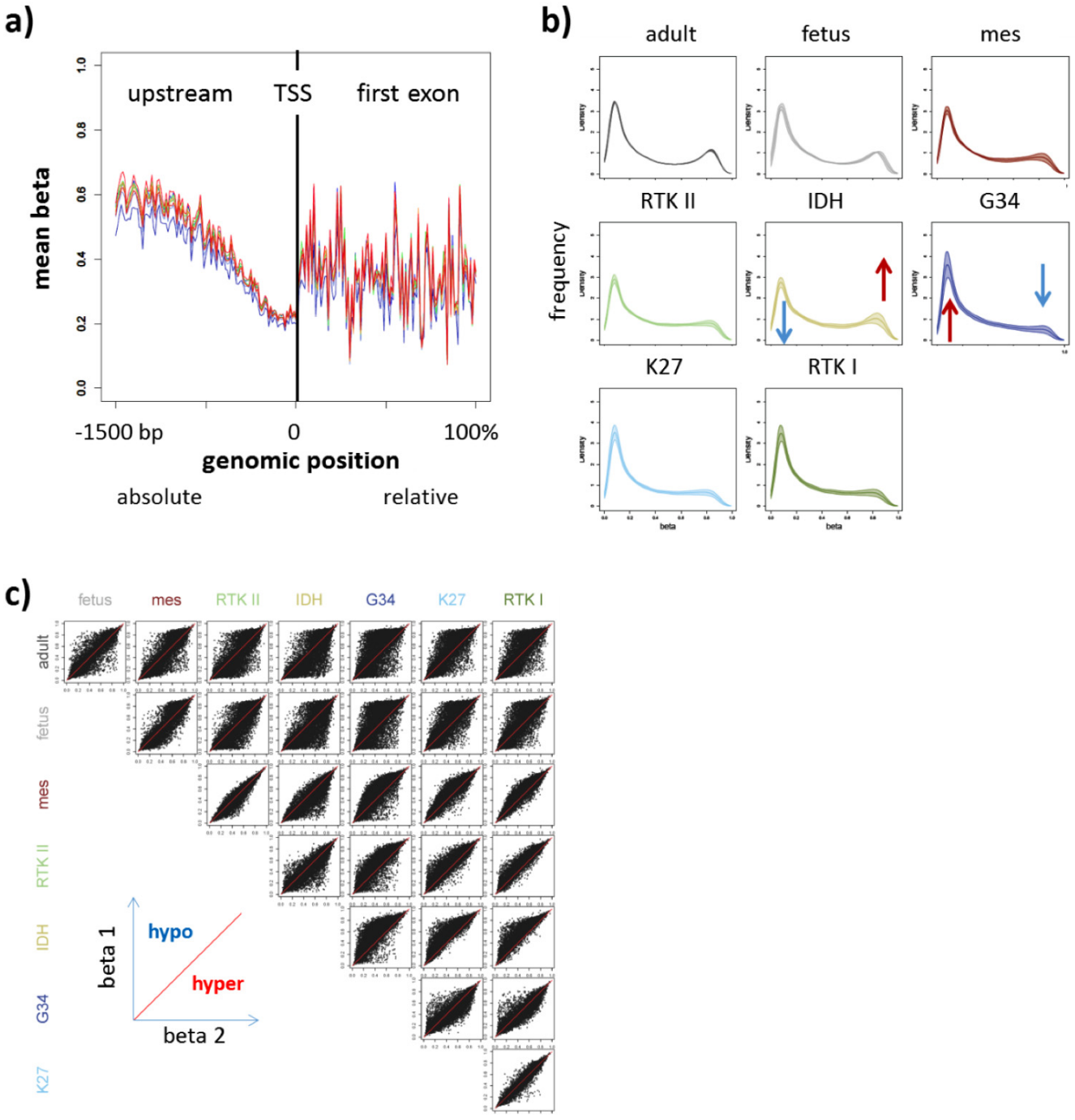
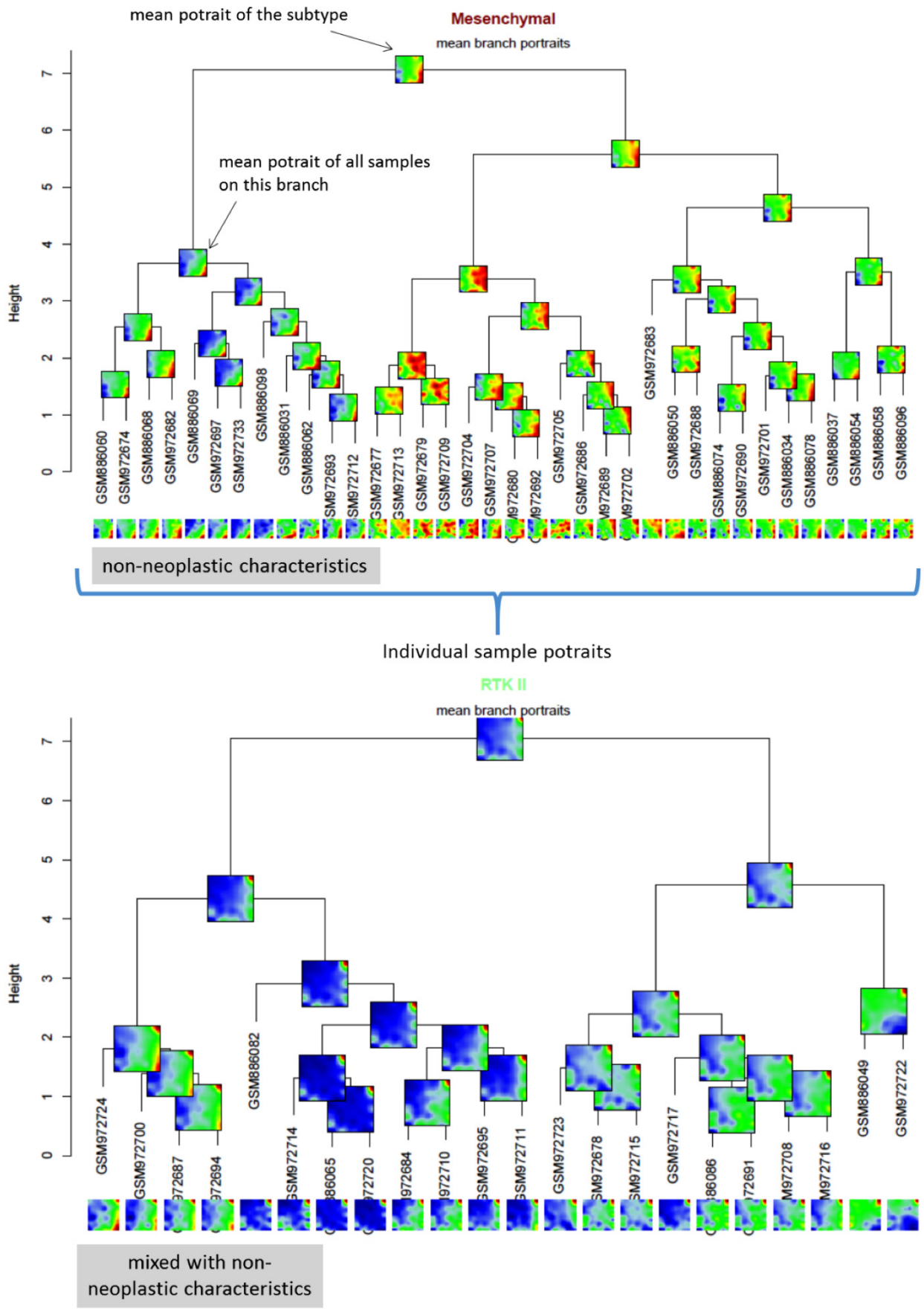
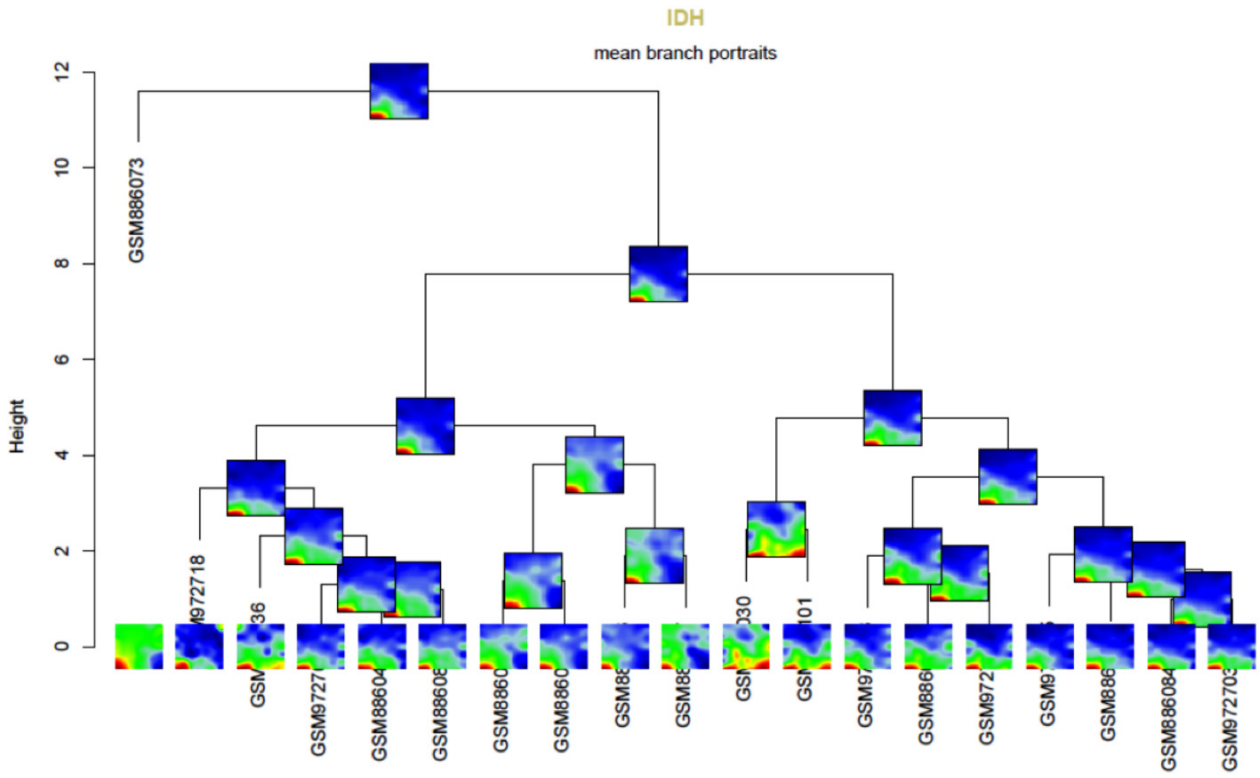
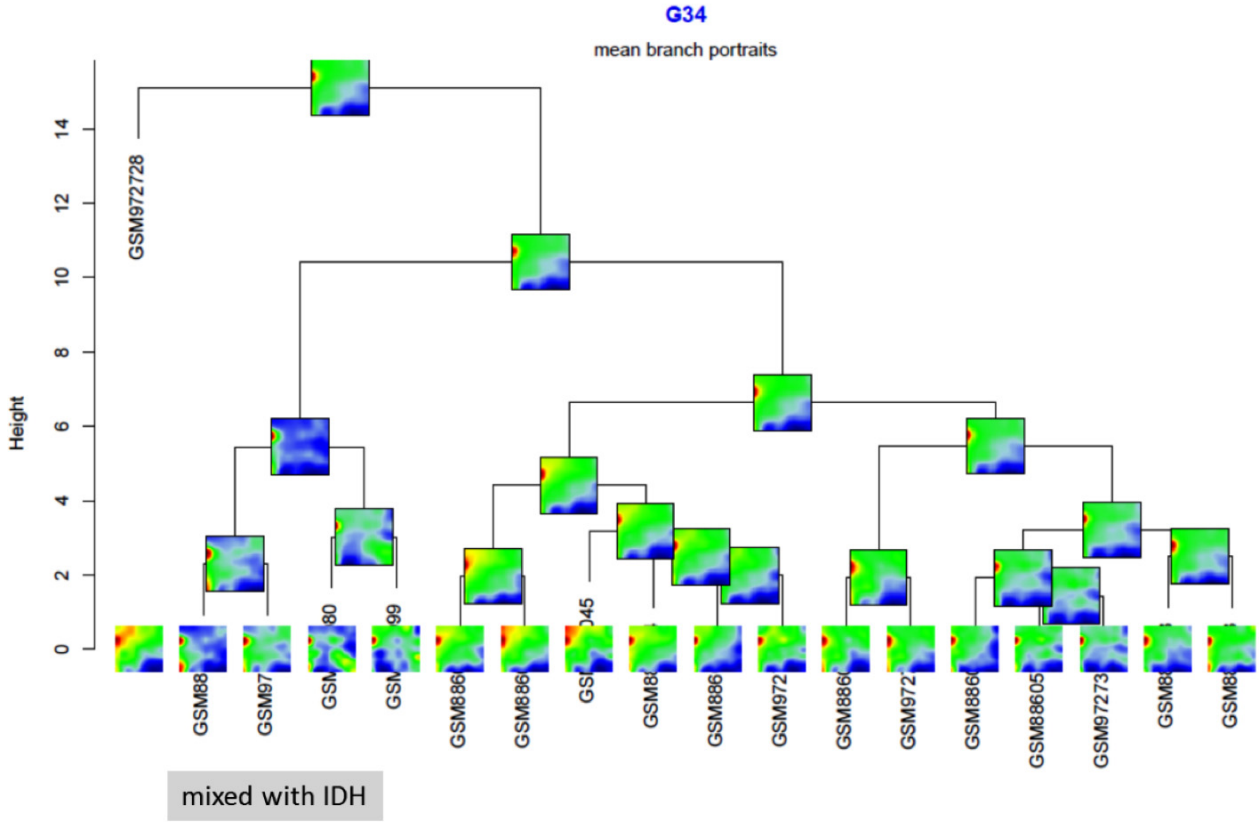


Figure S1: Global beta methylation characteristics: **a**) Mean methylation level as a function of the genomic position relative to the transcription start side (TSS). CpG-beta values were averaged over all genes for each subtype (the colors were assigned in **b**); **b**) frequency distribution of beta values for the GBM subtypes and controls. The arrows serve as a guide for the eye to indicate methylation changes leading to global hyper- or hypomethylation in IDH- and G34-type GBM compared with healthy controls; and **c**) mutual correlations of beta values between the classes.


mean potrait of the subtype
Mesenchymal
mean branch portraits
mean potrait of all samples on this branch
Height
non-neoplastic characteristics
Individual sample potraits
RTK II
mean branch portraits
Height
mixed with non-neoplastic characteristics


IDH
mean branch portraits
Height
0
2
4
6
8
10
12
GSM886073
GSM972718
GSM972727
GSM886042
GSM886083
GSM886084
GSM972703


G34
mean branch portraits
Height
0
2
4
6
8
10
12
14
GSM972728
GSM886055
GSM886056
GSM972731
GSM886057
GSM886058
GSM886059
GSM972732
GSM886060
GSM972733
GSM886061
GSM886062
GSM972734
GSM886063
GSM886064
GSM886065
GSM972735
GSM886066
GSM886067
GSM886068
GSM886069
GSM886070
GSM886071
GSM886072
mixed with IDH

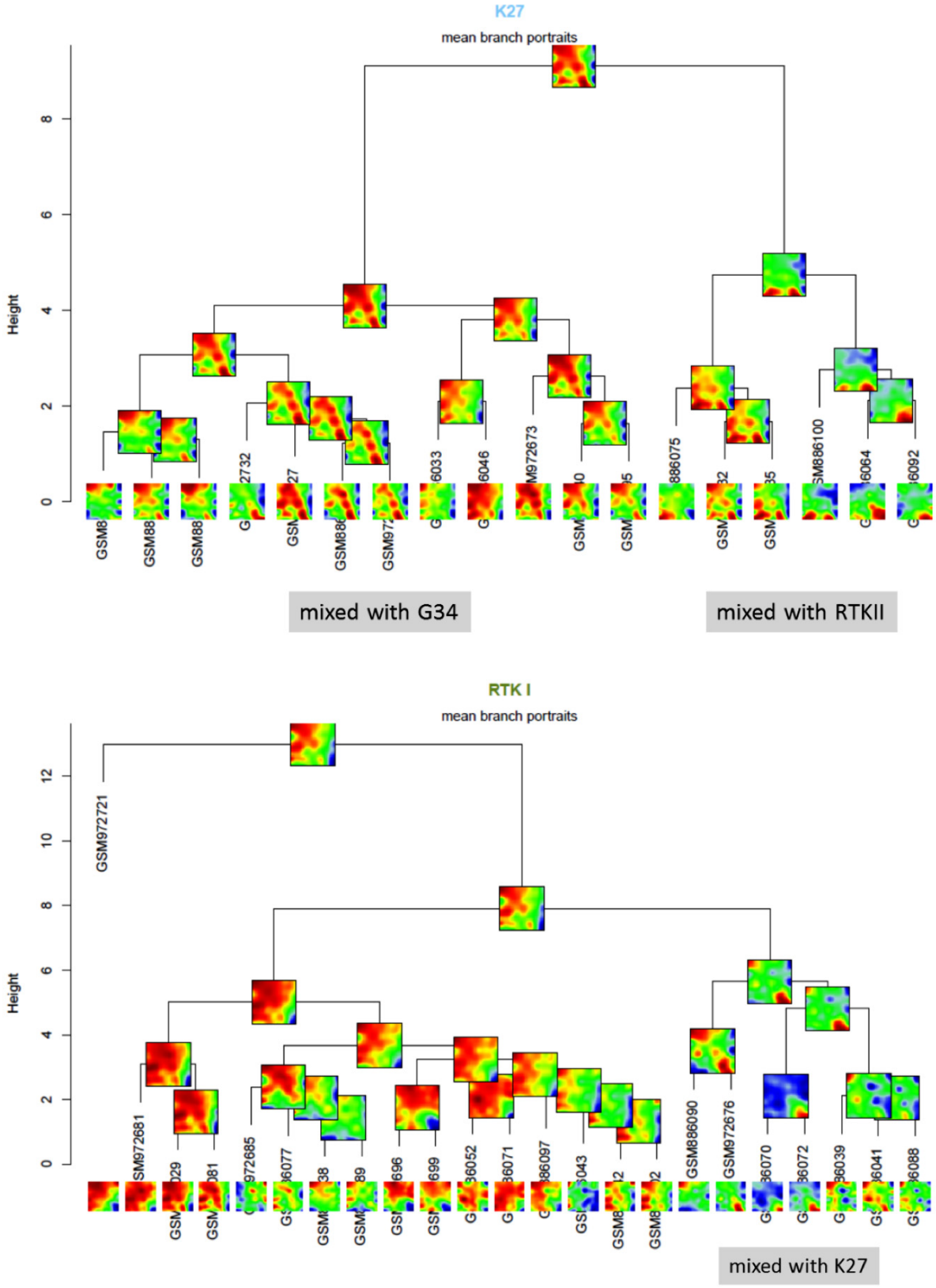


Figure S2: Hierarchical cluster trees of the sample SOM portraits for the GBM subtypes identify partial mixing between class characteristics.

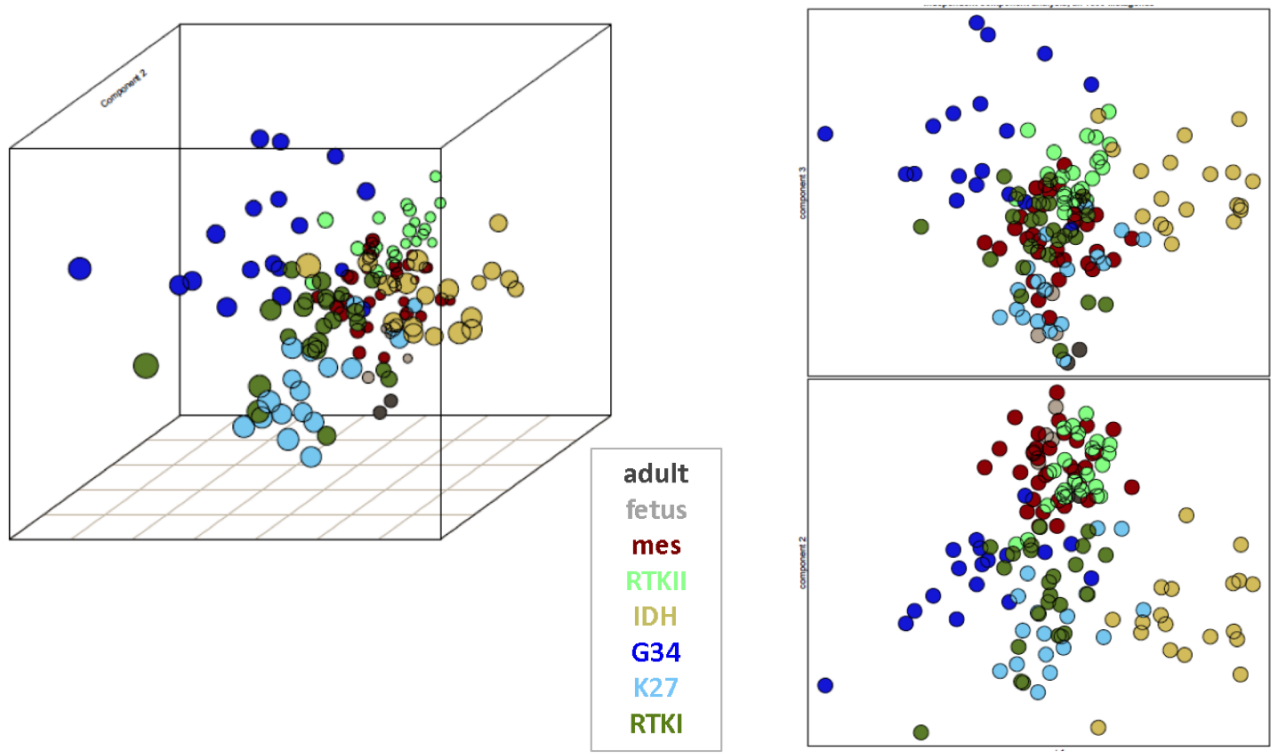


Figure S3: Independent component analysis of methylation data: IDH and G34 separate along the first independent component (IC1), the superclusters 'brain-like' and CHOP (except G34) along the IC2 and partly IC3.

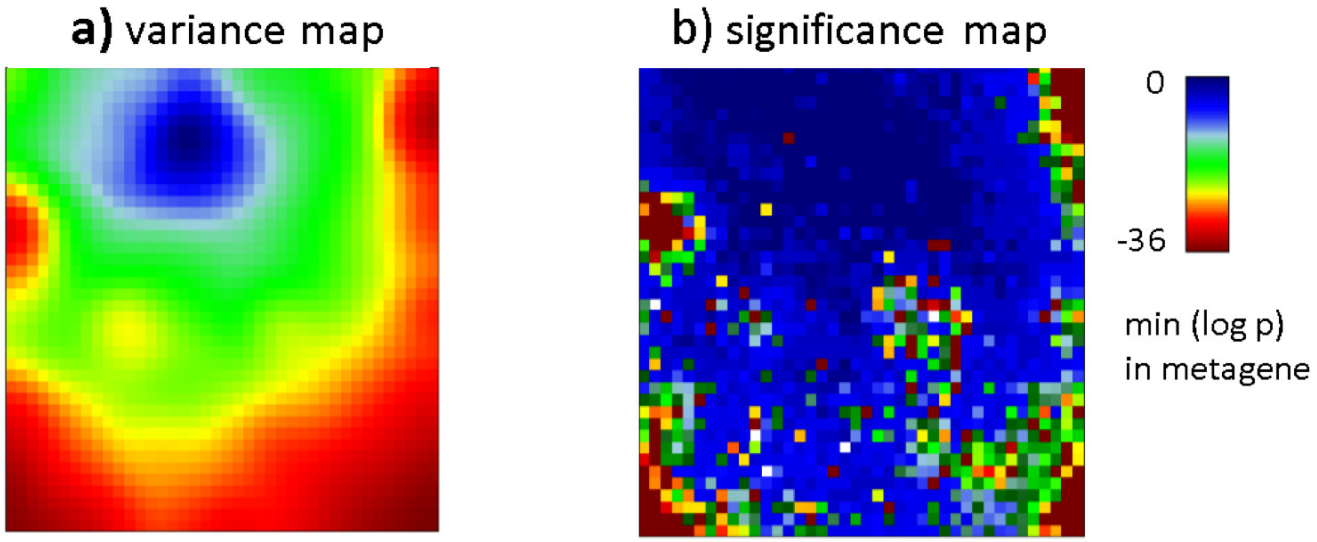


Figure S4: Additional DmetSOM information: **a**) The variance map shows that spot E1 (in the left upper corner) lacks variability. **b**) The p-value significance map color codes minimum log p-values found in each pixel. Maroon areas are highly significant differently methylated. They agree with the positions of the main spots. p values are adjusted using a shrinkage t-test [1].

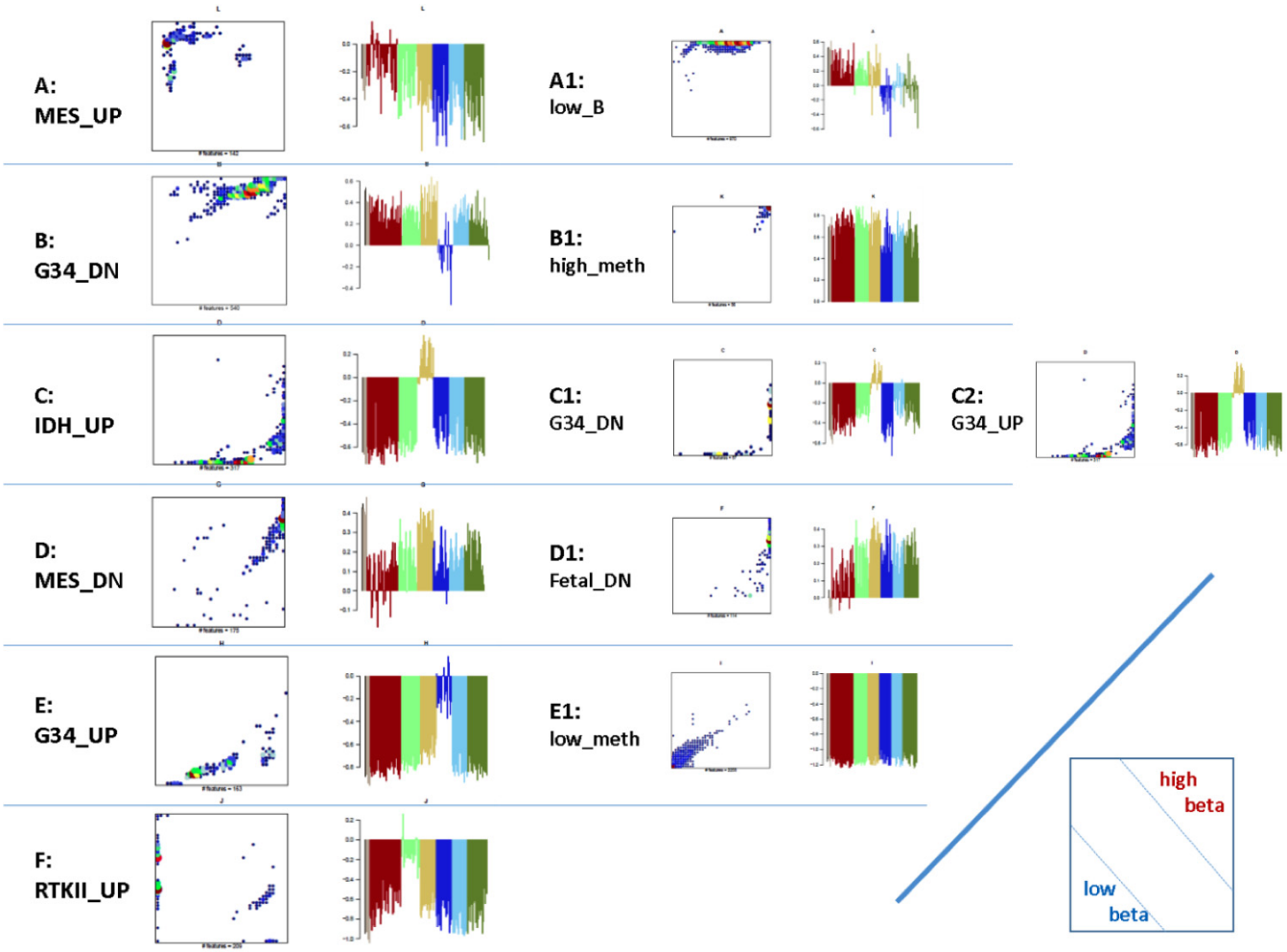
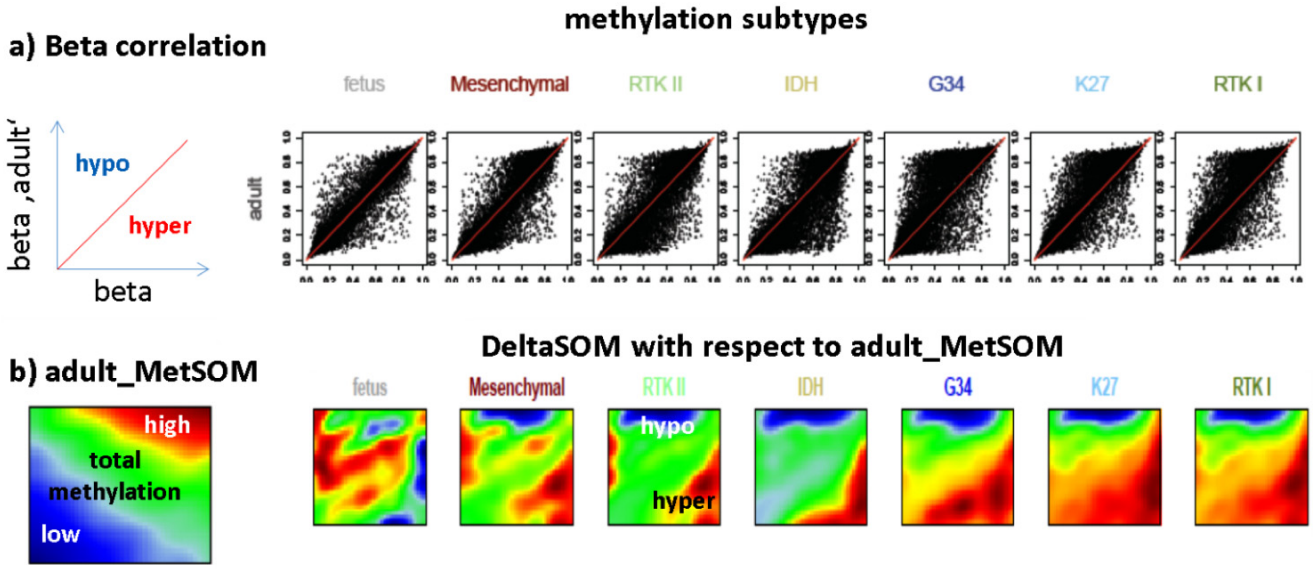


Figure S5: Mapping of the DmetSOM-spots into the MetSOM: All DmetSOM-spots accumulate in well limited areas indicating that DmetSOM and MetSOM are organized in a similar way. Some of the satellite spots indicate that they collect highly or weakly methylated genes (see also the legend in the right below part of the figure). Hence, differential methylation is obviously the main factor that organizes the genes. Nevertheless, absolute methylations levels slightly modulate the spot structure.



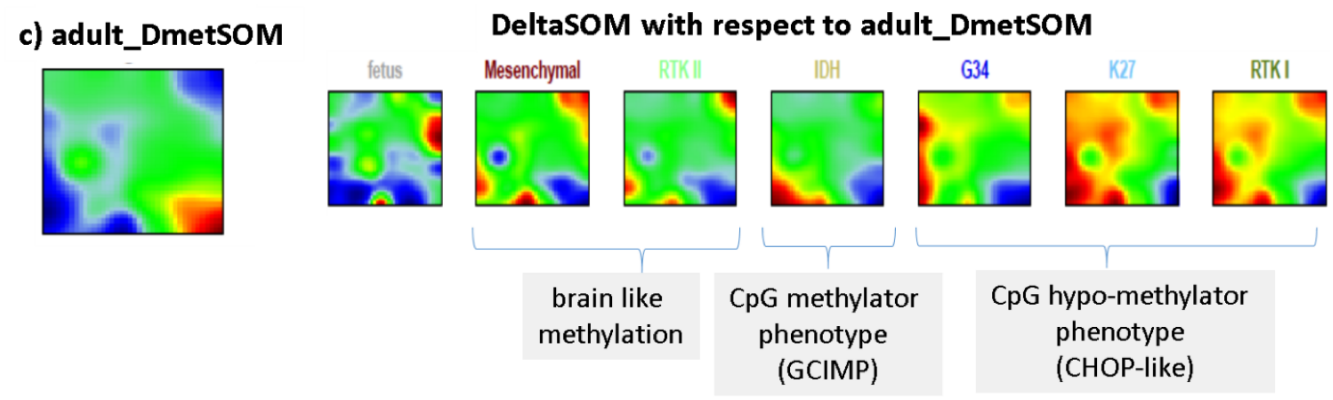


Figure S6: Differential methylation analysis with respect to adult healthy brain. **a**) The beta-correlation plot reveals global hyper- and hypomethylation for IDH and G34 GBM. However, a considerable number of genes shows the opposite trend in each of the subtypes. **b**) Difference MetSOM portraits (metagene-methylation data are subtractedpixelwise) show hypermethylation in GBM in the region of spots A1 and partly B and hypermethylation in GBM in the region of spots C, F and partly D. **c**) Difference DmetSOM portraits reveal global hyper- and hypomethylation in spots F and A1, respectively.

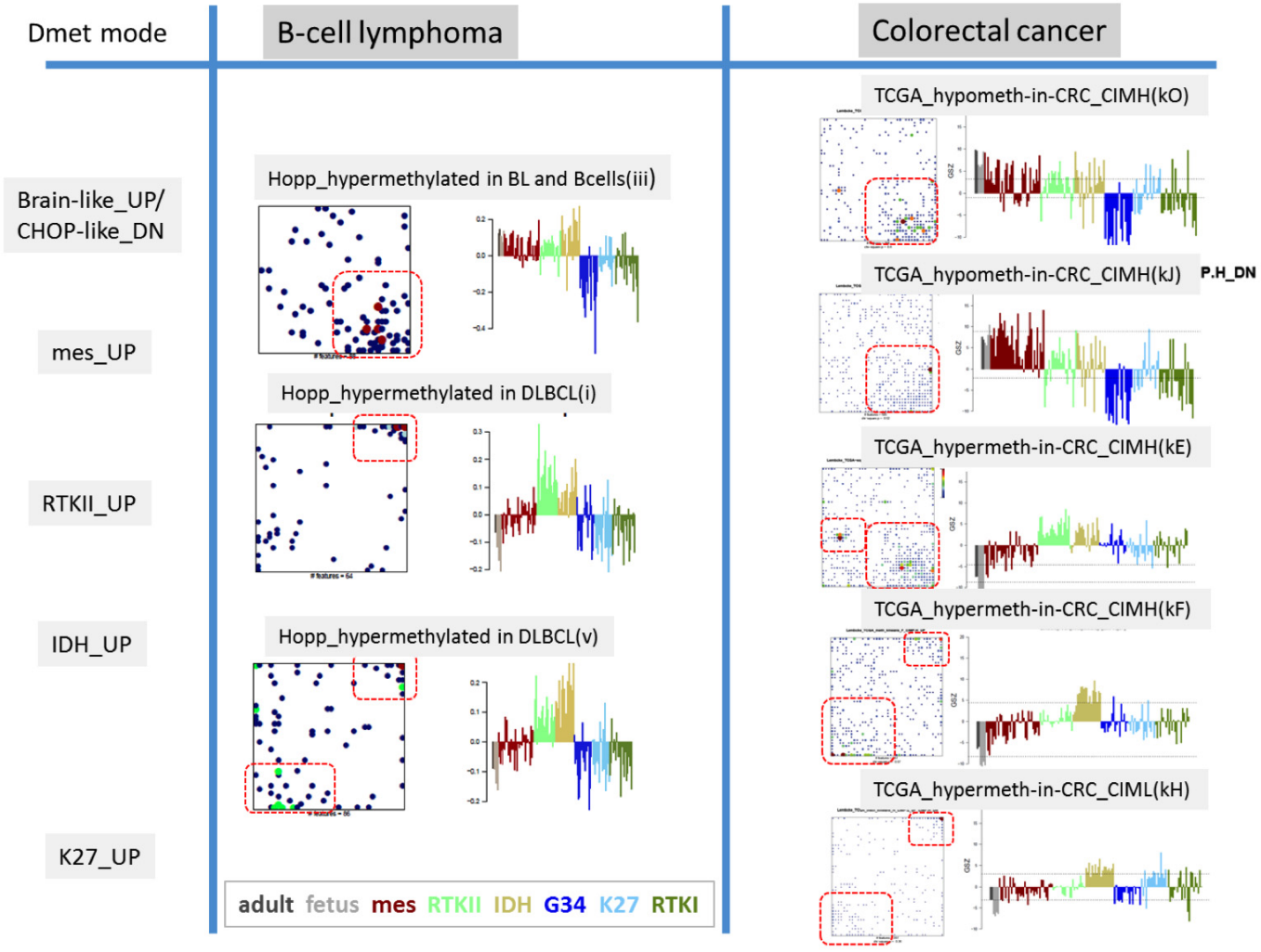


Figure S7: Mapping of methylation-signature gene sets of B-cell lymphoma and of colorectal cancer into the DmetSOM of glioma. The gene sets were determined using SOM spot analysis in recent studies on DNA methylation data in [2] and [3], respectively. The red frames indicate regions of increased local densities of genes. The profiles indicate subtype specific hyper- (and hypomethylation) in glioma. List of gene are available from the authors upon request.

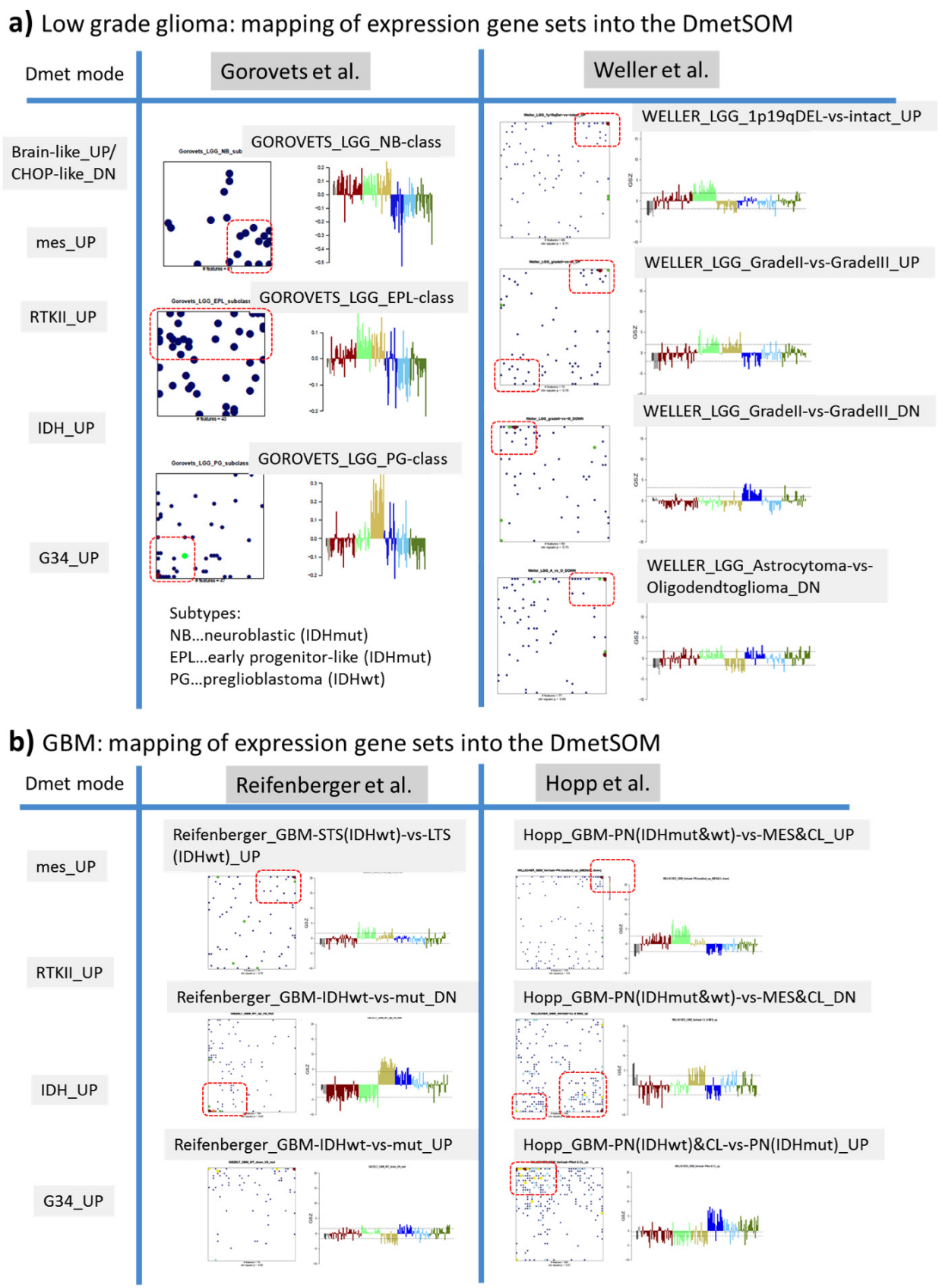


Figure S8: Mapping of gene expression marker sets for low (part **a**) and high (**b**) grade gliomas into the DmetSOM. Gene sets were taken from refs. [4, 5] (LGG) and [6, 7] (GBM). The profiles reveal subtype specific methylation patterns which mostly agree between LGGs and GBM. The mutational status of the IDH1 gene has strong effect of the profiles observed.

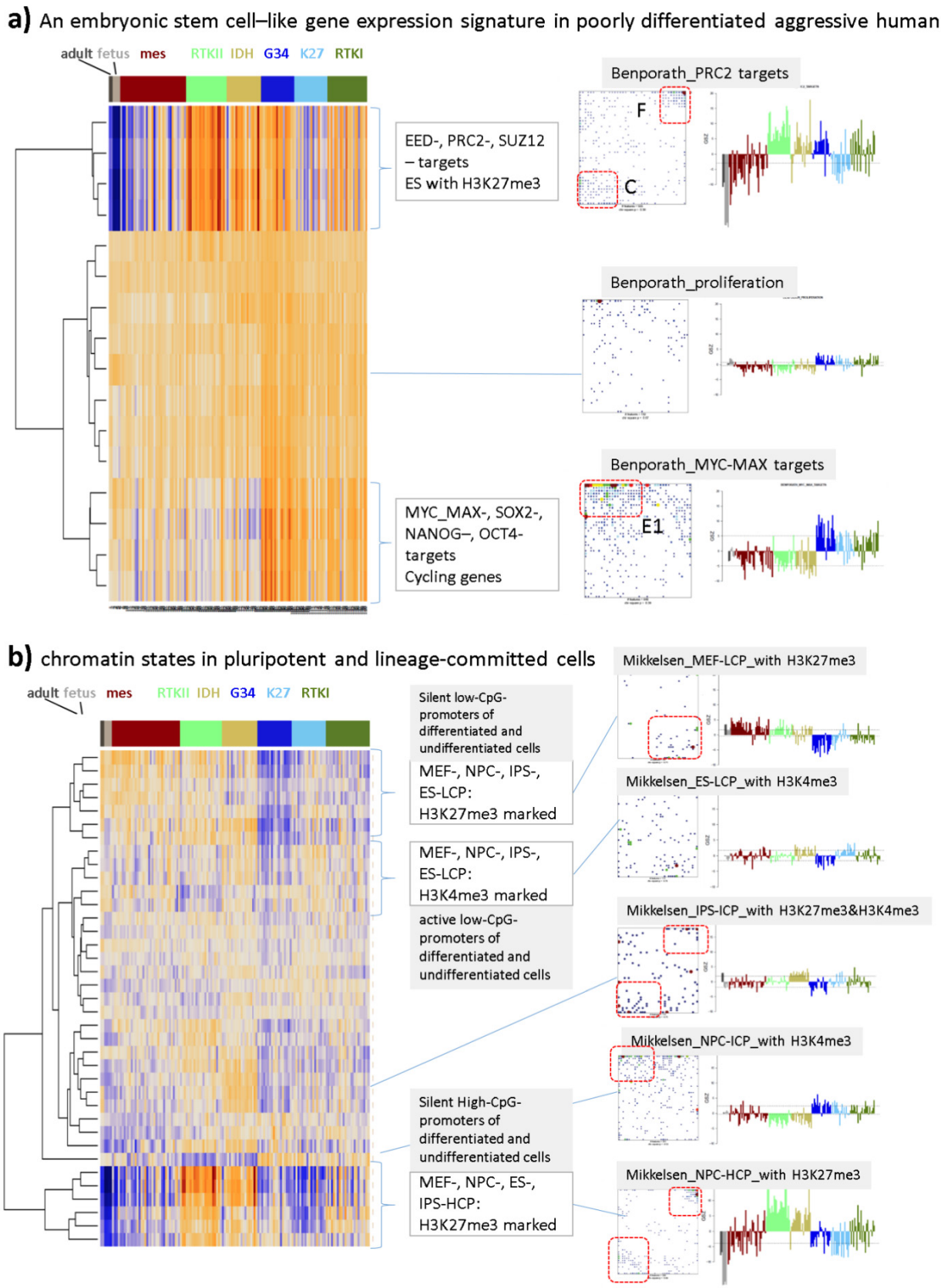
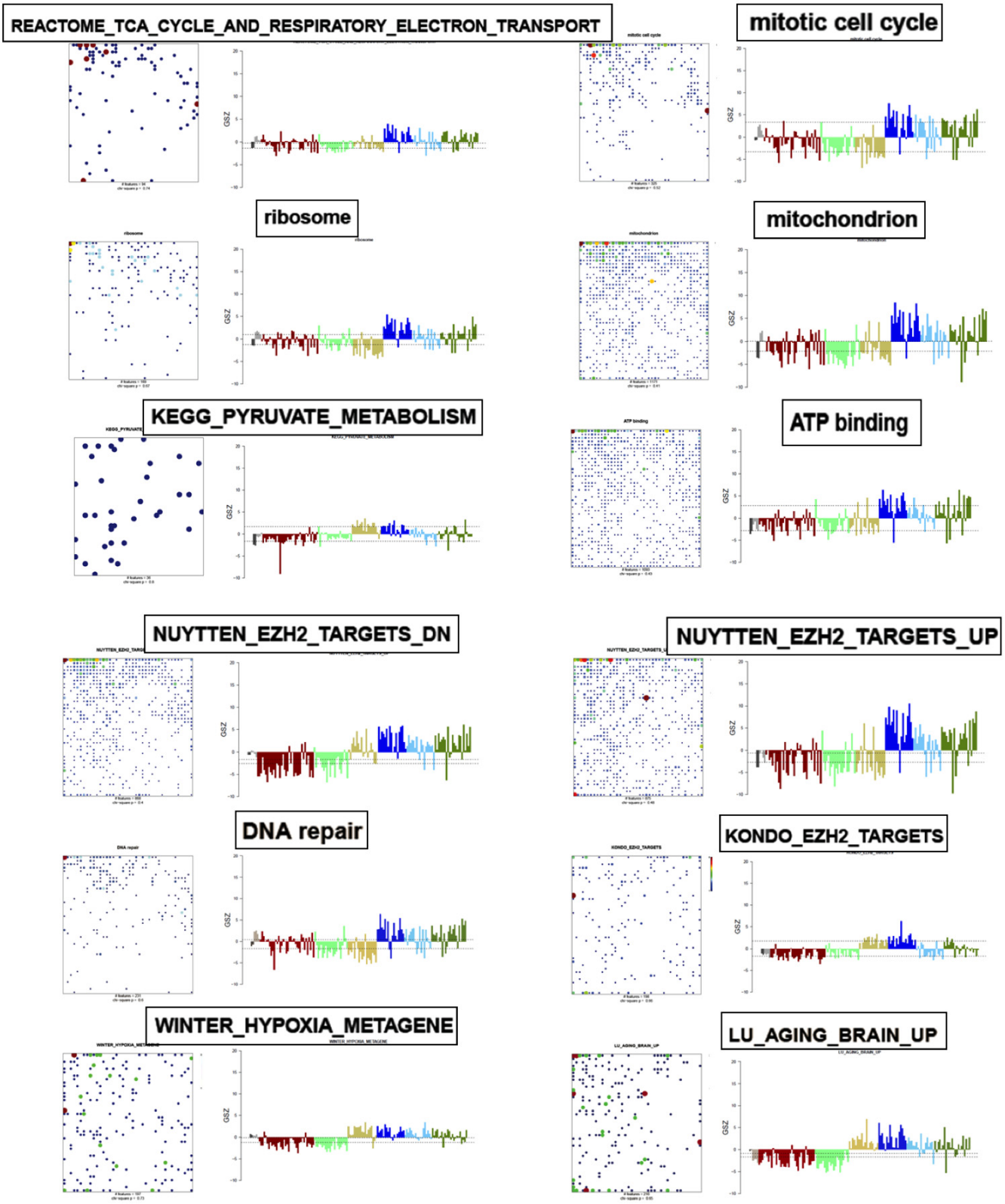


Figure S9: Enrichment heatmaps of gene sets referring to stemness-related genes including PRC2 and MYC targets (part **a**, [8]) and to chromatin states in pluripotent and committed cells (part **b**, [9]).

a) CHOP-like_UP/brain-like_DN profiles



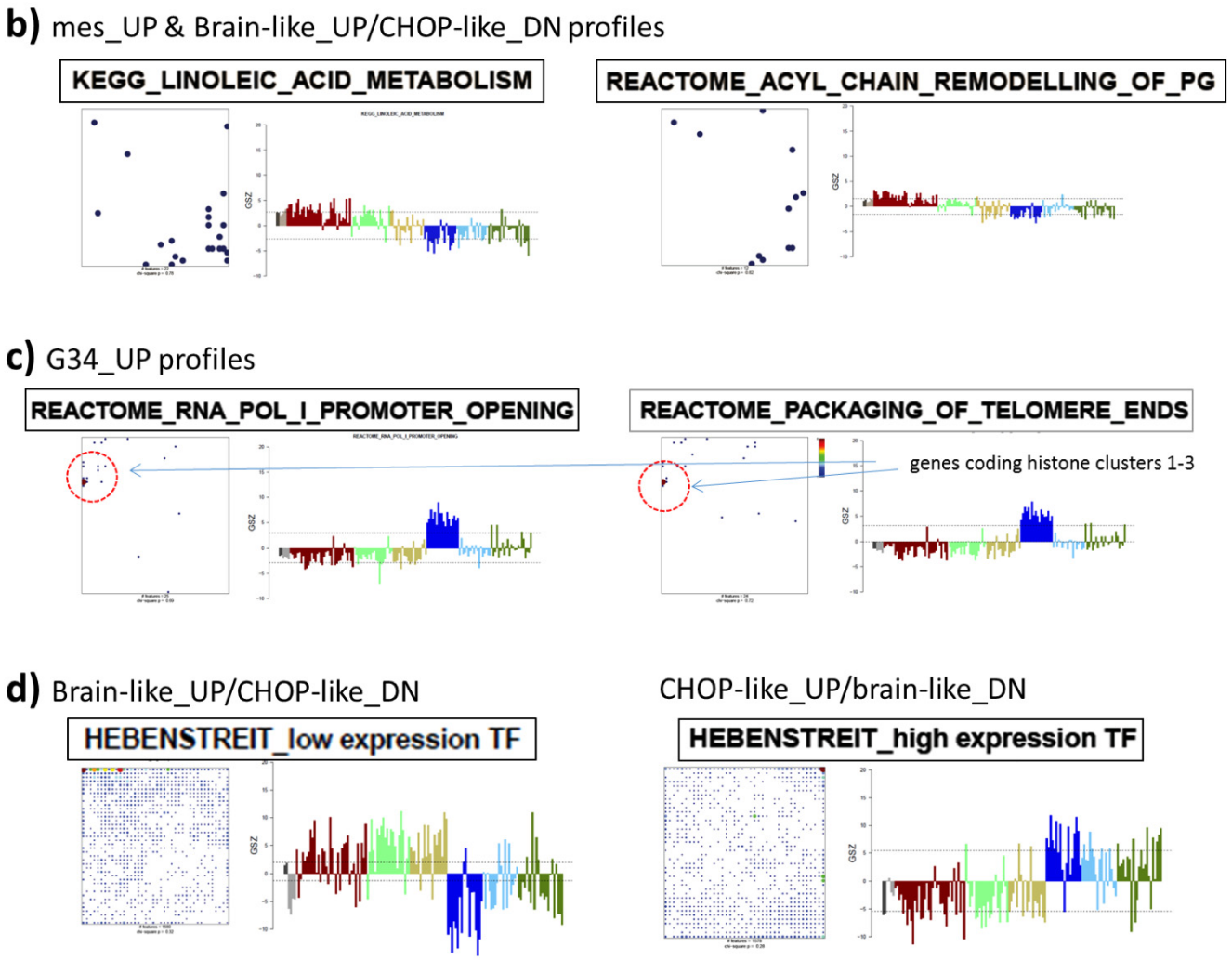


Figure S10: Selected gene set characteristics taken from different GO-terms, KEGG, Reactome and literature [10-13].

Table S1: Lists of Genes Included in the Methylation Spot-Clusters A – F and A1 – E1 (Figure 3)

The supplementary table can be downloaded from the link <Table S1>.

REFERENCES

[1] Wirth H, von Bergen M, Binder H. Mining SOM expression portraits: Feature selection and integrating concepts of molecular function BioData Mining 2012; 5: 18. http://dx.doi.org/10.1186/1756-0381-5-18

[2] Hopp L, Wirth-Loeffler H, Binder H. Epigenetic heterogeneity of B-cell lymphoma: DNA-methylation, gene expression and chromatin states. Genes 2015; in press.

[3] Binder H, Hopp L, Lembcke K, Wirth H. Personalized Disease Phenotypes from Massive OMICs Data. In Big Data Analytics in Bioinformatics and Healthcare; Baoying W, Ruowang L, William P, Eds, IGI Global: Hershey, PA, USA 2015; pp. 359-378. http://dx.doi.org/10.4018/978-1-4666-6611-5.ch015

[4] Gorovets D, Kannan K, Shen R, Kastenhuber ER, Islamdoust N, Campos C, Huse JT. IDH Mutation and Neuroglial Developmental Features Define Clinically Distinct Subclasses of Lower Grade Diffuse Astrocytic Glioma. Clinical Cancer Research 2012; 18: 2490-2501. http://dx.doi.org/10.1158/1078-0432.CCR-11-2977

[5] Weller M, Weber R, Willscher E, Riehmer V, Hentschel B, Kreuz M, Reifenberger G. Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome- and transcriptome-wide profiling improves stratification of prognostically distinct patient groups. Acta Neuropathologica 2015; 1-15.

[6] Reifenberger G, Weber RG, Riehmer V, Kaulich K, Willscher E, Wirth H, for the German Glioma N. Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. International Journal of Cancer 2014; 135: 1822-1831. http://dx.doi.org/10.1002/ijc.28836

[7] Hopp L, Wirth H, Fasold M, Binder H. Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate

cancer case study. Systems Biomedicine 2013; 1.
http://dx.doi.org/10.4161/sysb.25897

[8] Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 2008; 40: 499-507.
http://dx.doi.org/10.1038/ng.127

[9] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007; 448: 553-560.
http://dx.doi.org/10.1038/nature06008

[10] 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: 883-891.
http://dx.doi.org/10.1038/nature02661

[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] Lee C-K, Klopp RG, Weindruch R, Prolla TA. Gene Expression Profile of Aging and Its Retardation by Caloric Restriction. Science 1999; 285: 1390-1393.
http://dx.doi.org/10.1126/science.285.5432.1390

[13] Winter SC, Buffa FM, Silva P, Miller C, Valentine HR, Turley H, Harris AL. Relation of a Hypoxia Metagene Derived from Head and Neck Cancer to Prognosis of Multiple Cancers. Cancer Research 2007; 67: 3441-3449.
http://dx.doi.org/10.1158/0008-5472.CAN-06-3322