Cell

Supplemental Information

Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma

Michele Ceccarelli, Floris P. Barthel, Tathiane M. Malta, Thais S. Sabedot, Sofie R. Salama, Bradley A. Murray, Olena Morozova, Yulia Newton, Amie Radenbaugh, Stefano M. Pagnotta, Samreen Anjum, Jiguang Wang, Ganiraju Manyam, Pietro Zoppoli, Shiyung Ling, Arjun A. Rao, Mia Grifford, Andrew D. Cherniack, Hailei Zhang, Laila Poisson, Carlos Gilberto Carlotti, Jr., Daniela Pretti da Cunha Tirapelli, Arvind Rao, Tom Mikkelsen, Ching C. Lau, W.K. Alfred Yung, Raul Rabadan, Jason Huse, Daniel J. Brat, Norman L. Lehman, Jill S. Barnholtz-Sloan, Siyuan Zheng, Kenneth Hess, Ganesh Rao, Matthew Meyerson, Rameen Beroukhim, Lee Cooper, Rehan Akbani, Margaret Wrensch, David Haussler, Kenneth D. Aldape, Peter W. Laird, David H. Gutmann, TCGA Research Network, Houtan Noushmehr, Antonio Iavarone, and Roel G.W. Verhaak

Molecular profiling refines the classification of adult diffuse lower- and high-grade glioma

Supplemental Information

Supplemental Information content:

Supplemental Experimental Procedures

Supplemental Experimental Procedures

1. Biospecimens

Authors: Jay Bowen, Kristen M. Leraas, Tara M. Lichtenberg

Correspondence and questions should be directed to: Jay Bowen (jay.Bowen@nationwidechildrens.org)

Biospecimens were collected from patients diagnosed with low grade gliomas (LGG) and glioblastoma multiforme (GBM) undergoing surgical resection.

The case list freeze included 1122 cases comprising 516 LGG and 606 GBM. Samples were from the following 32 tissue source sites: Asterand (n=2); Case Western (n=188); Cedars Sinai (n=34); CHI-Penrose Colorado (n=2); Christiana Healthcare (n=12); Cureline (n=26); Dept of Neurosurgery at University of Heidelberg (n=48); Duke University (n=90); Emory University (n=44); Fondazione-Besta (QH) (n=38); Greenville Health System (n=1); Hartford (n=2); Henry Ford Hospital (n=243); Huntsman Cancer Institute (n=8); International Genomics Consortium (n=2); John Wayne Cancer Center (n=2); Johns Hopkins (n=7); Mayo Clinic (n=39); MD Anderson Cancer Center (n=101); Memorial Sloan Kettering Cancer Center (n=15); Northwestern University (n=2); St. Joseph AZ (n=30); Swedish Neurosciences (n=6); The University of New South Wales (n=19); Thomas Jefferson University (n=44); Toronto Western Hospital (n=14); University of California San Francisco (n=50); University of Florida (n=30); University of Kansas (n=1); University of Miami (n=3); University of North Carolina (n=2); University of Sao Paulo (n=17).

Samples were acquired and processed according to previous descriptions (Brennan et al., 2013; TCGA_Network, 2015).

A detailed list of clinical and molecular data elements is included in Table S1 and reflects the clinical data package frozen on 05/01/2015. Clinical data elements comprise histology, grade, gender, age at diagnosis/surgery, treatments, vital status, overall and progression-free survival. Clinical data available at the BCR was manually curated. Where possible, additional de-identified follow-up data were requested from TSSs through BCR and manually added into the clinical data freeze package.

Overall survival was defined as the time from surgical diagnosis until death. Cases that were still alive at the time of this study have overall survival time censored at the time of last follow-up. Survival curves were estimated and plotted using the Kaplan-Meier method. Log-rank tests were used to compare curves between groups. Single-predictor and multiple-predictor models were fit using Cox regression under the proportional hazards assumption. Hazard ratios and 95% confidence intervals are reported. Nested models were compared using the likelihood ratio test (LRT). Harrell's concordance index (C-index) was used to assess and report model performance

(Harrell et al., 1982). These analyses were conducted in R (v 3.1.2) using the survival package (Therneau, 2014; Therneau and Grambsch, 2000).

2. DNA sequencing

Authors: Floris Barthel, Bradley Murray, Siyuan Zheng, Roel Verhaak

Correspondence and questions should be directed to: Roel Verhaak (rverhaak@mdanderson.org)

2.1 DNA sequencing data production

Whole exome, whole genome and targeted validation and TERT promoter sequencing (including low-pass sequencing) was performed as previously described (Brennan et al., 2013; Cancer Genome Atlas Research, 2015; Verhaak et al., 2010).

Platform	Center	Disease	Exome capture kit	Read length	Paired samples
Illumina HiSeq	BI	GBM	Agilent Sure-Select Hun All Exon v2.0, 44Mb kit	2 x 76 bp	307
Illumina HiSeq	BI	LGG	Agilent Sure-Select Hun All Exon v2.0, 44Mb kit	2 x 76 bp	513
Union					**820**

Whole exome sequencing

Platform	Center	Disease	Libraries	Read length	Paired samples
Illumina HiSeq	BI	GBM	2-59	2 x 101 bp	38
Illumina HiSeq	BI	LGG	3-11	2 x 101 bp	20
Illumina HiSeq	WUGSC	GBM	16-167	100 bp	13
Illumina HiSeq	HMS-RK	LGG	1	2 x 51 bp	52
Union					**123**

Whole genome sequencing (including low-pass)

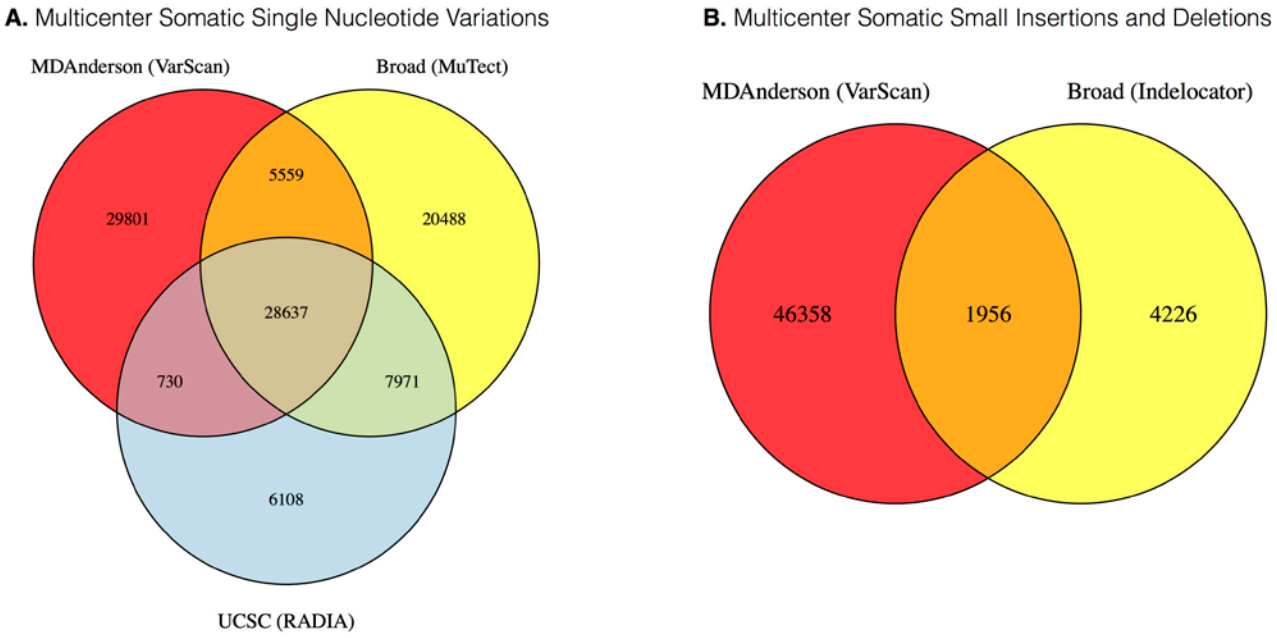
2.2 Identification of somatic mutations

The Broad Institute's Firehose cancer genome analysis pipeline used BAM files for tumor and matched normal samples to perform quality control, local realignment coverage calculations and others on whole exome sequencing (Table 1) as described (Imielinski et al., 2012). For the identification of somatic single nucleotide variations we used a multicenter approach integrating the output of three different somatic mutation algorithms: MuTect (Cibulskis et al., 2013), RADIA (Radenbaugh et al., 2014) and Varscan (Koboldt et al., 2012). MAF files from each mutation calling algorithm were integrated in a unique MAF file considering those mutations that were called at least by two of the three considered methods. The integrated MAF contains 28637 somatic mutation called by all the methods, 5559 called by MuTect and VarScan, 7971 called by MuTect and RADIA

and 730 called by VarScan and RADIA. Similarly, for the detection of somatic insertions and deletions we intersected the calls produced by Indelocator and Varscan algorithms obtaining 1956 high confidence indels.

2.3 Identification of IDH mutations

In order to expand the annotation of IDH status in our cohort, previously reported (Cancer Genome



Atlas Research, 2008) mutation calls on Sanger sequenced DNA and exome sequencing of whole genome amplified DNA were used. Sanger sequencing and whole exome sequencing of whole genome amplified DNA was performed as previously described (Brennan et al., 2013; Cancer Genome Atlas Research, 2008; Verhaak et al., 2010). Except for bona fide IDH1/2 mutations, no other mutations were called on these platforms.

Platform	Center	Aliquot	Disease	Paired samples
ABI	WUGSC	DNA	GBM	158
Illumina	BI	WGA	GBM	163
Union*				**174**

Additional data used to determine IDH mutation status.

2.4 Identification of TERT promoter mutations

Targeted sequencing at the TERT promoter region (Chr5:1295150-1295300) was performed on a subset of 287 cases as previously described (Cancer Genome Atlas Research, 2015). Additionally, we evaluated whole genome sequencing (including low-pass) for the presence of somatic variants using GATK pileup. We required a minimum coverage of at least 6 bp and a minimum variant allele faction of 15% for detection of TERT promoter mutations. A total of 328 cases had sufficient coverage to detect a mutation and 162 cases showed a somatic mutation at one of three sites.

Nucleotide change	Site	Paired samples
A161C	Chr5:1295161	2
C228T	Chr5:1295228	121*
C250T	Chr5:1295250	39*

**One case showed mutations in both C250T and C228T*

2.5 Mutation significant analysis

Significantly mutated genes were identified using the MutSigCV algorithm. Analyses were conducted on the entire sample set (n=820) accept a single hypermutator phenotype (TCGA-DU-6392). Intronic mutations were excluded. A mutation blacklist was applied for remove potential technical artifacts (Lawrence et al., 2013b). Genes with a q-value less than 0.1 were considered significant.

2.6 Telomere quantification

Quantification of telomere length was performed using the TelSeq tool (Ding et al., 2014). This tool counts the number of reads containing any (range 0 to *k*) amount of telomeric repeats (n_k), or TTAGGG, and then computes the estimated telomere length in bp *l* further based on the average chromosome length in bp *c* and the total coverage *s*.

1) $l = c \times \frac{n_k}{s}$

The authors recommend a k of 7 based on their experimentally validated results. Given that TelSeq was not designed for cancer, it does not take into account tumor ploidy and purity. We have therefore modified the TelSeq computation to consider tumor purity *p* and ploidy τ:

2) $\frac{n_k}{s} = \frac{l_t \times \tau \times p + l_n \times (1-p)}{\tau \times c \times p + c \times (1-p)}$

Because *p* and τ are given by the ABSOLUTE analysis (Carter et al., 2012), solving l_t is straightforward, whereas l_n can be calculated using 1) above.

The average chromosome length c is calculated as follows:

3) $c = 46/G$

Here *G* is the total genome length and *46* is the expected number of chromosomes. Because GC content is a potential confounding factor, *G* was set to the genome length in bp with GC content between 48% and 52%. The average coverage *s* is adjusted in a similar fashion.

2.7 Whole genome mutation calling

MuTect (Cibulskis et al., 2013) was used to call somatic mutations on 89 matched primary tumor-normal pairs. We required a minimum coverage of 14 in the tumor sample and 8 in the normal sample. Variants known to dbSNP v132 and unknown to COSMIC v54 were filtered resulting in 714,305 variants. Using these samples we used overlapping RNA-seq expression data to form an integrated dataset of 67 pairs (29 GBM, 38 LGG). In order to identify potential promoter sites we used the GENCODE v19 transcript annotation (n= 196,520 transcripts) and used a subset of 24,001 transcripts that have an exact UniProt database match and has been curated according to known clinically relevant protein changes (Ramos et al., 2015). We then reduced the transcripts down to one transcript per gene (n=17,722 transcripts). For each remaining transcript we then took a region spanning from 2,000 bp upstream of the transcription start site and 200 bp into the coding region. We then determined overlapping mutations for each region using the Bioconductor package "GenomicRanges" (Lawrence et al., 2013a). We removed regions with hits from less than 7 unique samples, removed regions which were upstream of genes lacking RNA-seq counts or counts that were lacking any variability, removed regions in which the variants had a median of read count of 1 or more alternate reads in the matching normal. This filtering resulted in 141 mutations across 12 putative promoter regions (Table S2E). For each of the remaining gene promoter regions we then performed a t-test and a mann-whitney-U test comparing the log2 normalized gene expression counts in mutant cases to wild type cases. When we subsequently filtered out promoter regions with a Benjamini-Hochberg adjusted gene expression correlation Q-value < 0.05 only three promoter regions remained including TERT, TRIM28 and CACNG6.

3. DNA copy number analyisis

Authors: Bradley Murray, Floris Barthel, Roel Verhaak

Correspondence and questions should be directed to: Roel Verhaak (rverhaak@mdanderson.org)

3.1 Preprocessing and peak calling

Tumor and normal samples were profiled on Affymetrix SNP6.0 GeneChip arrays and subsequently processed into genome segmentation files (McCarroll et al., 2008). The tool GISTIC 2.0 was then used to identify significantly reoccurring focal and broad copy number changes (Mermel et al.,

2011). Events with a Q-value < 0.10 were considered significant. In order to identify low-frequent subtype specific events, we ran GISTIC both across the entire cohort (n=1084) and smaller subsets within DNA methylation clusters (n=6 groups), RNA expression clusters (n=4 groups) and IDH-codel subtypes (n=3 groups). For each statistically significant peak, GISTIC 2.0 indicates a narrow focal peak and a wider surrounding peak. We intersected all overlapping focal peaks across all GISTIC run and identified 57 disjoint amplified regions and 105 deleted regions. Using this method, while drastically limiting the number of genes compared to using the wide peak boundaries, we were still about to find 80% of genes that were considered as potential tumor drivers in previous studies. Genes previously suggested as tumor drivers not found using this method include IRS2 gain, LSAMP loss and KDR/KIT gain (the neighboring oncogene PDGFRA however was still found). In order to further narrow down the list of genes per peak and to identify potential tumor drivers, we sought to correlate copy number change to gene expression and prioritized genes in which we found significant mutations. Using this method, we were able find evidence for several new tumor drivers including GIGYF2 loss, ERRFI1 loss, ARID2 loss and FGFR2 gain. For the complete list of peaks, genes and their mutation and expression correlates see Table S2B.

3.2 Functional Copy Number (CN) analysis

Authors: Pietro Zoppoli, Antonio Iavarone

Correspondence and questions should be directed to: Pietro Zoppoli (zoppoli@icg.cpmc.columbia.edu)

In order to define the functional copy number (fCN) genes we calculated the spearman's correlation between the copy number and the expression of each gene in the dataset. We selected all the genes with p-value <0.05 and cor >0.5.

In order to highlight the different behavior between the four expression groups, we selected only the differentially expressed (abs (FC>1.5)) and aberrated (abs (**Δ**CN>0.5)) fCN genes obtaining a list of 57 genes (the fCN signature).

3.3 Mutations with Common Focal Alterations (MutComFocal)

Authors: Raul Rabadan, Jiguang Wang, Antonio Iavarone

Correspondence and questions should be directed to: Antonio Iavarone (ai2102@cumc.columbia)

By considering both copy number and somatic mutation data of LGG/GBM samples, we applied the algorithm of MutComFocal (Trifonov et al., 2013). Particularly, focality score and recurrence score were calculated based on samples with at least 10 and at most 1,000 copy number segments. The

focality score assigns equal weight to all genes participating in a genomic alteration inversely proportional to the size of that alteration, while recurrence score assigns equal weight to all genes altered in a sample inversely proportional to the total number of gene altered in the sample (Frattini et al., 2013; Trifonov et al., 2013).

4. mRNA Expression

Authors: Michele Ceccarelli, Stefano M. Pagnotta, Antonio Iavarone

Correspondence and questions should be directed to: Michele Ceccarelli (ceccarelli@unisannio.it)

4.1 Data preparation and gene selection

RNA-seq raw counts of 667 cases (513 LLG and 154 GBM) were downloaded, normalized and filtered using the Bioconductor package TCGAbiolinks (Colaprico et al., 2015) using TCGAquey(), TCGAdownload() and TCGAprepare() for both tumor types (“LGG” and “GBM”, level 3, and platform "IlluminaHiSeq_RNASeqV2"). The union of the two matrices was then normalized using within-lane normalization to adjust for GC-content effect on read counts and upper-quantile between-lane normalization for distributional differences between lanes applying the TCGAanalyze_Normalization() function encompassing EDASeq protocol. Gene selected for clustering were chosen by applying two filters, the first was aimed a reducing the batch effect between the two tumor cohorts. We computed differentially expressed genes with TCGAanalyze_DEA() (implementing the EdgeR protocol (Robinson et al., 2010)), and filtered out genes differentially expressed between the two sets (α = 10-10), obtaining 10,389 genes. We then applied variability filters that select genes having a sufficiently high variation (100%) between the mean of top 5% and the mean of the bottom 5% values and having these means respectively above and below the overall median value of the data matrix. The filtering steps resulted in 2,275 genes that were used for the consensus clustering. ConsensusClusterPlus Bioconductor package was used to perform the clustering with hierarchical clustering as inner method and 1000 resampling steps (epsilon=0.8). Number of cluster (n = 4) was used as local maxima of the Calinsky-Harabasz curve. Within cluster analysis was done generating differentially expressed genes between GBM and LGG cohorts (log fold change greater and 1.0 and FDR less than 0.05), lists were then analyzes using DAVID functional annotation tool (Huang et al., 2009) and ClueGO (Bindea et al., 2009).

4.2 Classification of Affymetrix samples

Once the four RNA-seq cluster were obtained, we reclassified 378 GBM samples for which no RNA-seq data were available using their Affymetrix profiles. We used the 151 GBM samples (20 in LGr1,

4 in LGr2, 10 in LGr3 and 117 in LGr4) having both the Affymetrix and RNA-seq profiles as training set of a kNN classifier (k = 3) to assign LGr cluster memberships to the remaining 378 Affymetrix samples. The feature set of the classifier was based on a signature of 327 probesets obtained by selecting up-regulated and down-regulated genes for the training samples in each cluster.

4.3 Tumor Map and Pathway Activity Analysis

Authors: Yulia Newton, Olena Morozova, Sofie Salama

Correspondence and questions should be directed to: Sofie Salama (ssalama@soe.ucsc.edu)

4.3.1 Combining multi-platform multi-tumor datasets

We utilized the ComBat batch effect removal method (Johnson et al., 2007) in order to combine mRNA expression data from the GBM RNA-seq (n=154), GBM Agilent (n=525), LGG RNA-seq (n=513), and LGG Agilent (n=27) datasets. We chose to use data generated using Agilent microarray platform over those generated using Affymetrix because such data were available for both tumor types, while Affymetrix data were only available for GBM samples. We combined the 4 datasets and ran ComBat. We flagged 4 batches, one for each dataset, to be removed by the ComBat method. One hundred and forty nine GBM samples were analyzed using both Agilent and RNA-seq platforms. Twenty seven LGG samples were analyzed using both Agilent and RNA-seq platforms. We utilized these matched samples as biological covariates in the ComBat method. Upon completion of the data transformation, we removed all redundant samples analyzed using the Agilent platform whenever the sample was also analyzed using RNA-seq. This combined mRNA expression dataset (n=1043) was used for Tumor Map analysis.

4.3.2. Tumor Map method (manuscript in preparation)

Tumor Map is a dimensionality reduction and visualization method for high dimensional genomic data. It allows viewing and browsing relationships between high dimensional heterogeneous genomic samples in a two-dimensional map, in a manner much like exploring geo maps in Google Maps web application.

Prior to the analysis, technical and batch effects in the gene expression data were mitigated as a preprocessing step and as described above. We computed sample-by-sample pair-wise similarities. From RNA expression data, we selected 6002 genes whose expression was the most variable based on the variance distribution curve. The 1301 most important methylation probes were selected by manual curation of the probe list as described in the DNA methylation analysis section. We used Spearman rank correlation (Spearman, 1904) on these continuous variable data (mRNA and methylation). To build maps based on a single data type, for each sample the closest

neighborhood of 10 samples is selected. The Tumor Map method represents these local neighborhoods as a graph. The edge weight in this graph is proportional to the magnitude of the similarity metric. Then spring-embedded graph layout (Golbeck and Mutton, 2005) algorithm is applied to the constructed graph. The spring-embedded layout algorithm treats edges as springs and allows the springs to oscillate for a fixed amount of time with the energy inversely proportional to the edge weights. Under these conditions, springs with large weights do not oscillate much, causing those vertices to stay together. However, springs with small weights oscillate more and end up farther away from each other. The method then projects the positions of all the vertices in the resulting graph layout onto a two-dimensional grid. Each cell in the grid allows only one vertex to be placed into it. If multiple vertices contest for the same grid cell, a random vertex selection is made and placed into the cell; and the other competing vertices are placed into the nearest empty cell, snapping around the original cell in a spiral-like manner. Thus, dense clumps of samples are separated so that they can be viewed at approximately the same scale as the distances that separate them. After computing pairwise sample similarities in the gene expression and DNA methylation space separately, the two similarity spaces are combined after standardizing each space was standardized.

4.3.3 Multi-platform maps using Bivariate Standardization similarity space Transformation (BST)

We computed sample pairwise similarities for each data type separately, producing a square samples-by-samples similarity matrix. For each of the similarity matrices, we perform bivariate standardization by transforming each value to be an arithmetic mean of the z-scores of this value within both the row and the column empirical distributions. This method is an adaptation of the approach by Faith et. al (2007). Once each of the similarity matrices is transformed into a z-score space, we combine each available z-scores (from N platforms) for each pair of samples by taking a weighted average of the z-scores, where the weights indicate the importance of each of the N platforms being combined. When genomic data for a given platform is not available for at least one of the samples from the pair, a pairwise similarity for this pair will not be available for this platform. Our method allows such omissions, as it will only combine similarity z-scores from those platforms that are available for any pair of samples. The resulting BST matrix is a square samples-by-samples matrix that contains a union of samples in all the platforms.

4.3.4 Extracting significantly active pathways

We used mRNA expression for samples available through RNA-seq platform only and the CNV data to transform the data into inferred pathway activity levels using PARADIGM (Vaske et al., 2010). We then considered a number of dichotomies, such as LGm1 GBM vs. LGG (see Table S5). Some of the dichotomies we considered have significantly different numbers of samples in each class (see Table S5). In order to make statistically strong inferences about pathway activities we only considered those dichotomies in which both classes are well represented by their members and the variance within the classes is much smaller than the variance between the classes. In other words, we selected those dichotomies where sample scatter is small within the classes and classes are separable in the pathway space. Based on the PARADIGM IPLs (Inferred Pathway Levels) we computed pair-wise Spearman rank correlation for each pair of samples. We then computed within-class and between-class variance of the correlations, first for the first class and then for the second class. We then computed the F-statistic for each of the classes in the dichotomy and the p-value based on the F-distribution. We aggregated the p-value for the dichotomy by computing the mean p-value. We selected those dichotomies that had an aggregated p-value of <= 0.05. Table S5 shows final dichotomies analyzed for the differential pathway activities. For each dichotomy selected, we computed differential activity levels using the linear models for microarrays and RNA-seq data (LIMMA) method (Smyth, 2005). We then applied Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) to the HUGO members of the full differential vector. We extracted only those pathways that had FDR-adjusted q-value of <= 0.1. At the same time, we extracted statistically significant differentials (multiple hypothesis adjusted p-value <= 0.05). We ran PATHMARK (Cancer Genome Atlas Research, 2013) on the statistically significant differential activities obtained from LIMMA to extract connected components of the global PARADIGM regulatory network. An additional filter of 3 standard deviations was applied to the PATHMARK method. This means only those activities that fall outside 3 standard deviations of the empirical distribution of the statistically significant differentials pass through the filter. A network connection is extracted if both vertices connected by that connection pass the filter. For each pathway gene set that passed the GSEA q-value of 0.1 we computed the overlap of the pathway genes and those that survived the PATHMARK filter as well as the over-representation hypergeometric p-value. We then extracted those pathways that passed with the p-value of <= 0.05. Figure S5E shows an overview of the process for extracting significantly active pathway from the glioma data. Figures S5C-D show pathway views of the significant IPLs from Table S5 in which IPLs representing families, complexes, phopho-events and redundant complexes were removed for better visualization.

4.4 Combining GBM Agilent G4502A mRNA data with LGG Illumina Hi-Seq RNA-seq data

Authors: Shiyun Ling, Rehan Akbani

Correspondence and questions should be directed to: Rehan Akbani (rakbani@mdanderson.org)

Approximately 15,700 genes were common between the two platforms and a total of 185 pairs of GBM and LGG sample replicates were run on both platforms. Initial tests by combining the GBM and LGG replicates and clustering them showed two clusters based entirely on platform differences and the replicates didn't merge with each other. To remove the platform effect, we developed a novel algorithm that randomly divided the 185 replicate pairs into training, testing and validation sets. The training set was used to train an Empirical Bayes (Johnson et al., 2007) based model, which was then applied to the testing set. The testing set was used to figure out which genes didn't merge well by using a *t*-test to find the genes with the most differences between the platforms. The process was repeated 1000 times by using a bootstrapping approach for the training set. The top 3000 genes that were consistently found to be the most variable in the testing set were removed from the data set. The resulting model was then applied to the validation set, after removing those 3000 genes, to evaluate the algorithm. The evaluations showed that all 43 of the replicate pairs in the validation set clustered in matched pairs. The median of Pearson's correlations between the matched pairs was 0.23 before adjustment and 0.93 after adjustment, indicating very successful merging. We then applied the model to the full GBM and LGG dataset to perform overall merging, and then removed duplicates by randomly keeping one sample from the pairs. The final dataset had 1032 samples and 12,717 genes.

4.5 RNA Fusion analysis

Authors: Olena Morozova, Floris Barthel, Sofie Salama, Roel Verhaak

Correspondence and questions should be directed to: Roel Verhaak (rverhaak@mdanderson.org)

4.5.1 Fusion transcript detection using PRADA

Transcript fusions were detected in 665 samples using the Pipeline for RNA-seq Data Analysis (PRADA) fusion detection tool (Torres-Garcia et al., 2014). We classified fusions to one of four tiers based on the number of junction spanning reads and discordant read pairs, gene partner uniqueness, gene homology, whether the fusion preserves the open reading frame, transcript allele fraction and DNA breakpoints in SNP6 array data, as previously described (Yoshihara et al., 2014). Briefly, tier one fusions are the highest confidence fusions and tier four fusions are the lowest

confidence ones. For the purpose of this analysis we chose to include tiers one and two. A summary of included fusions can be found in Table S2C.

4.5.2 Fusion transcript detection using deFuse

RNA-seq reads were analyzed using deFuse package version 0.6.0 (McPherson et al., 2011). Fusions involving receptor tyrosine kinase genes were manually reviewed using blat analysis (Kent, 2002) of the breakpoint sequence in the UCSC Genome Browser (Kent et al., 2002). Candidate fusions were filtered based on the following deFuse parameters:

- Splitr_count > = 5 (5 or more split reads supporting the fusion)
- Span_count > = 10 (10 or more spanning reads supporting the fusion)
- Read_through ~ “N” (fusion is not a readthrough)
- Adjacent ~ “N” (fusion does not involve adjacent genes)
- Altsplice ~ “N” (fusion cannot be explained by alternative splicing)
- Min_map_count = 1 (at least one spanning read supporting the fusion is uniquely mapped)
- ORF ~ “Y” (fusion preserves the open reading frame)

deFuse and PRADA fusion predictions were combined to generate a list of 204 events identified by both methods (Table S2C).

4.6 Identification of Transcriptional Regulatory Factors underlying IDH wild type and IDH mutant phenotypes in Glioma

Authors: Ganiraji Manyam, Arvind Rao, Ganesh Rao

Correspondence and questions should be directed to: Ganesh Rao (grao@mdanderson.org)

Batch-corrected expression data from Agilent Microarray and Illumina Hiseq RNA-seq platforms using MBatch was used for differential expression and transcription factor analysis. Linear regression was used to find the genes that are differentially expressed between IDH wild type and IDH-mutant groups after adjusting for the effect of expression platform (RNA-seq or microarray) in the model. The p-values are adjusted for multiple testing using the Bonferroni method. Genes with adjusted p-value less than 0.01 are considered significant.

Transcription Factor (TFs) Analysis was performed using the Match Algorithm of Biobase (TRANSFAC) system to identify TFs enriched in promoters of genes differentially expressed between IDH wild type and mutant groups. This algorithm compares the number of TF binding sites found in a query sequence set against a background set and identifies factors whose frequencies are enriched in the query compared to the background. Genes significantly upregulated in the IDH

mutant group are considered as the background for TF analysis of genes upregulated in IDH wild type group and vice-versa. The TFs enriched with p-value less than 0.05 are considered significant.
Differential expression analysis was used to assess the expression differences of the enriched TFs themselves between the two groups (IDHmut vs wt). The transcription factors with Bonferroni-adjusted p-value less than 0.05 are defined as significant candidates (Excel file).
Ingenuity Pathway Analysis (IPA) was used to generate downstream networks for the top ranking transcription factors. Rank of the transcription factor is defined based on fold change between the two groups and the number of transcription factor binding sites in the promoter region of the target genes. Twelve transcription factor families were found to have log fold change of >1 between the IDH mut and IDHwt groups. The ones with the highest number of target genes are NKX2-5, PAX8, ETV7, CEBPD, ETV4, ELF4, and NFE2L3. Several of these TFs have been shown to be important for carcinogenesis. For example, Pax8 has been shown to be minimally expressed in LGG and normal brain but highly expressed in glioblastoma (Hung et al., 2014) and plays a role in telomerase regulation (Chen et al., 2008). Similarly, enrichment of the pro-proliferative TF ETV4 in 1p/19q codeleted gliomas has been demonstrated (Gleize et al., 2015).

5. DNA methylation profiling

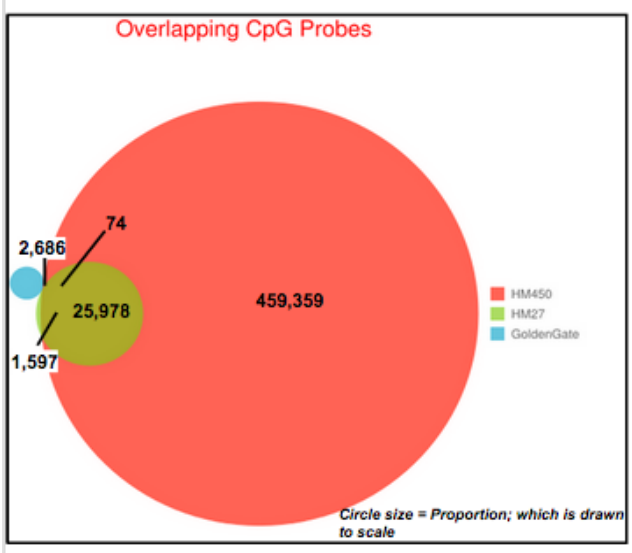
Authors: Thais S. Sabedot, Tathiane M. Malta, Simon G. Coetzee, Peter W. Laird, Houtan Noushmehr

Correspondence and questions should be directed to: Houtan Noushmehr (houtan@usp.br)

5.1 Preprocessing and clustering

For data acquisition, we used the the Bioconductor package TCGAbiolinks (Colaprico et al., 2015). First, TGCAquery() was used to search the samples of “GBM” and “LGG” tumors in TCGA repository using the following parameters: data level = 3, platform type = “HumanMethylation450” and “HumanMethylation27”, version 12 for LGG and version 5 GBM samples. Second, TCGAdownload() was used to download the data; and, finally, TCGAprepare() was used to read the data into a dataframe. A total of 932 TCGA glioma samples assessed for DNA methylation, including 516 LGG and 416 GBM samples, profiled using 2 different Illumina platforms, were included. During the initial phase of the TCGA project, 287 GBM samples (batches 1 to 9) were profiled using the Illumina HumanMethylation 27 platform (HM27), which interrogates 27,578 CpG probes. As a new platform became available, the TCGA LGG (batches 1 to 16) and 129 GBM (batches 1 to 12) samples were transitioning into the larger more comprehensive Illumina platform known as the HumanMethylation450 (HM450), which interrogates 485,421 CpG sites. The DNA methylation score for each locus is presented as a beta (β) value (β = (M/(M+U)) in which M and U indicate the mean

methylated and unmethylated signal intensities for each locus, respectively. β-values range from zero to one, with scores of zero indicating no DNA methylation and scores of one indicating complete DNA methylation. A detection p-value also accompanies each data point and compares the signal intensity difference between the analytical probes and a set of negative control probes on the array. Any data point with a corresponding p-value greater than 0.01 is deemed not to be statistically significantly different from background and is thus masked as “NA” in TCGA level 3 data packages. The data levels and the files contained in each data level package are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). Please note that as continuing updates of genomic databases and data archive revisions frequently become available, the data packages on TCGA Data Portal are updated accordingly. Data of the two platforms (HM450 and HM27) were merged as previously described (Brennan et al., 2013) and we ended with 25,978 probes that match both 27k and 450k platforms, as illustrated in the following Venn diagram. Duplicated samples and secondary tumors were excluded. The 932 sample IDs used for DNA methylation analysis are listed in Table S1.



5.2 Unsupervised clustering analysis of DNA methylation data

Methods to capture tumor-specific DNA methylation probes were used as recently described (Cancer Genome Atlas Research, 2014b) and is provided here as reference, with slight modifications to the total numbers. We used the Level 3 DNA methylation data contained in the packages listed above for analyses. We first removed probes which had any “NA”-masked data points and probes that were designed for sequences on X and Y chromosomes. We selected CpG sites that are located in high CpG density regions (top 25% of the sites with the highest observed/expected CpG ratio around their 3kb regions spanning from 1,500 bp upstream to 1,500

bp downstream of the transcription start sites) and CpGs associated with CpG islands extracted from the UCSC Genome Browser (http://genome.ucsc.edu). To capture cancer-specific DNA hypermethylation events, we further eliminated sites that were methylated (mean β-value ≥0.3) in histologically non-tumor brain tissues (Guintivano et al., 2013). This selection method reduced the initial 25,978 probes to 1,300 glioma-specific CpG probes, which corresponded to 6.5% of the full available data. However, a clustering analysis can be strongly confounded by the purity of tumor samples. To alleviate the potential influence of variable levels of tumor purity in our sample set on our clustering result, we dichotomized the data using a β-value of >0.3 as a threshold for positive DNA methylation. We then performed unsupervised hierarchical clustering on 1,300 CpG sites with this threshold that are methylated in at least 10% of the tumors using a binary distance metric for clustering and Ward's method for linkage. The cluster assignments were generated by cutting the resulting dendrogram. The probes are arranged based on the order of unsupervised hierarchical clustering of the dichotomous data using a binary distance metric and Ward's linkage method. We identified six groups (LGm1-LGm6) shown in Figure 2A generated based on the original β-values to visualize 1,300 CpG sites used in the clustering.

The approach described above to capture tumor-specific DNA methylation probes was used to select glioma-specific CpG probes and perform unsupervised clustering separated by IDH status. We identified 1,308 tumor specific CpG probes for IDH-mutant analysis and identified three IDH-mutant-specific clusters (Figure S3A). Likewise, we identifed 914 tumor specific CpG probes for for IDH-wild type samples and identified three IDH-wildtype-specific clusters (Figure S4A).

In order to classify the newly acquired TCGA samples (not included in the previous studies; LGG = 227; GBM = 20) into the context of previously published DNA methylation clusters (Brennan et al., 2013; TCGA_Network, 2015), we randomly selected a set of 80% of TCGA samples to train a random forest machine-learning. We then evaluated the performance on the remaining 20% of samples and got an accuracy of more than 88% on average. We then tested the new TCGA samples and classified them into the previously DNA methylation clusters.

5.3 Supervised analysis of DNA methylation

We used Wilcoxon test followed by multiple testing using the Benjamini and Hochberg (BH) method for false discovery rate estimation (Benjamini and Hochberg, 1995) to identify differentially DNA methylated probes between two groups of interest.

The 131 probes presented in Figure 3A were defined comparing samples from IDHmut-K1 (n=53) to IDHmut-K2 (n=221), using the following criteria: FDR < 10e-15, absolute difference in mean methylation beta-value > 0.27.

The 90 probes presented in Figure 3H were identified comparing samples from G-CIMP-low (n=25) to G-CIMP-high (n=249), in order to identify probes defining the G-CIMP-low group, using the following criteria: FDR < 10e-13, difference in mean methylation beta-value > 0.3 and < -0.4.
The 149 probes presented in Figure 3H were a combination of the 90 probes described above with 73 probes identified from the comparison between non-codels (from LGm2, n=210) and codels (from LGm3, n=120), using the following criteria: FDR < 10e-30, absolute difference in mean methylation beta-value > 0.25, removing probes with NA values. All probeset lists are provided on the publication portal accompanying this publication (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/).

5.4 Identification of Epigenetically Regulated Genes

To increase our statistical power, we decided to evaluate epigenetically regulated genes using the Pan-glioma subtypes, which allowed us to use the entire TCGA glioma cohort. We selected tumor samples that have both DNA methylation and RNA-sequencing based gene expression data to do this analysis, resulting in 636 samples (513 LGG and 123 GBM). We also randomly selected 110 non-tumor TCGA samples from 11 different tissues (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/), profiled using the same platforms. Each DNA methylation probe was mapped to the nearest UCSC gene, and after merging the DNA methylation and gene expression data, we retained a total of 19,530 pairs of DNA methylation and gene expression probes. We organized the tumor samples as either methylated (β >= 0.3) or unmethylated (β < 0.3) for each probe. We selected the pair of DNA methylation and gene expression probes for which the mean expression in the methylated group was lower than 1.28 standard deviation (bottom 10%) of the mean expression in the unmethylated group, and in which >80% of the samples in the methylated group have expression levels lower than the mean expression in the unmethylated group. We labeled each tumor sample as epigenetically silenced for a specific probe/gene pair if: it belonged to the methylated group and the gene expression level was lower than the mean of the unmethylated group silenced (Cancer Genome Atlas Research, 2014a), resulting in 3,806 probes/genes identified as epigenetically regulated. A Fisher test was used to detect if these 3,806 pairs were enriched in a DNA methylation cluster. For each probe, tumor samples labeled as methylated and downregulated by cluster, while non-tumor samples labeled as unmethylated and upregulated, were counted and arranged into a contingency table for a Fisher test, using 50% as a cutoff. p-value was calculated for each probe/gene pair and then was adjusted for multiple testing using the BH method for false discovery rate estimation (Benjamini and Hochberg, 1995). This analysis identified 3 Epigenetically Regulated groups (EReg): EReg2 with 233 genes enriched in LGm2 (resembling G-CIMP high), EReg3 with 15 genes enriched in LGm3

(resembling Codels) and EReg4 with 14 genes enriched in LGm4 (resembling Classic-like) and 1 gene enriched in LGm5 (resembling Mesenchymal-like). Since LGm1 (enriched for G-CIMP-low) and LGm6 (comprising LGm6-GBM and PA-like) are heterogeneous clusters, we applied a different approach in order to identify epigenetically regulated genes for these groups. For EReg1, we compared the DNA methylation and gene expression levels for G-CIMP-low samples (n=25) with G-CIMP-high samples (randomly selected 50 samples out of 249) and those probes/genes with Wilcoxon BH adjusted p-value less than 1e-10, methylation difference greater than 0,25 and RNA expression log Fold Change greater than 0,85 were selected, resulting in 15 epigenetically regulated genes enriched in G-CIMP-low. For EReg5, we compared the DNA methylation levels for LGm6 samples (n=77) with a subset of randomly selected samples from the 855 remaining TCGA glioma samples (n=140) and those probes with Wilcoxon BH adjusted p-value less than 1e-21 and methylation difference greater than 0,33 were selected, resulting in 12 epigenetically regulated genes enriched in LGm6.

To validate the EReg genes in order to confirm the existence of these signatures in an independent, non-TCGA data, we downloaded 4 different and publicly available datasets (Lambert et al., 2013; Mur et al., 2013; Sturm et al., 2012; Turcan et al., 2012), comprising 324 samples with distinct histology and clinical attributes. These samples included adult, pediatric gliomas of both low and high grade, reported with codel, IDH status and G-CIMP status. Our independent data set included a pool of 61 pilocytic astrocytomas defined as grade I gliomas (Lambert et al., 2013). In order to classify the additional non-TCGA gliomas into our LGm clusters, we selected a random set of 80% TCGA samples to train a random forest machine-learning model and evaluated the performance on the remaining 20%. Given the high specificity and sensitivity of our model (accuracy > 88% on average), we, then, tested the LGm cluster prediction model on the additional non-TCGA samples using the random forest method. Data were visualized using the same 45 pairs of CpG probes/genes that define the epigenetically regulated genes for IDH mutant samples (Figure 3F) and the same 27 pairs of CpG probes/genes that define the epigenetically regulated genes for IDH wild type samples (Figure 4D). Applying a similar ordering in the validation set and accounting for differences in sample size, we recapitulated the five EReg groups both for IDH mutant samples (Figure 3G) and IDH wild type samples (Figure 4E) in molecular level. The list of epigenetically regulated genes can be found at https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/ .

The same random forest machine learning model approach was used for the IDH-mutant samples (using the 1,308 IDH-mutant tumor specific CpG probes) and for the IDH-wildtype samples (using the 914 IDH-wildtype tumor specific CpG probes), separately. We then tested the models in the IDH-

mutant and IDH-wildtype samples from the validation set (Lambert et al., 2013; Mur et al., 2013; Sturm et al., 2012; Turcan et al., 2012) (Figure S4B).

5.5 Classification of new glioma samples based on DNA methylation glioma subtypes

New glioma samples can be classified into one of our glioma subtypes using our CpG probe methylation signatures provided on the publication portal accompanying this publication (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/).

First, all glioma samples should be divided by their known IDH status, separated into either IDH-mutant and IDH-wildtype. IDH-mutant is defined as those samples harboring any type of known IDH1 or IDH2 mutation as described recently (TCGA_Network, 2015). IDH-wildtype refers to those samples with an intact IDH1 or IDH2. Samples as either IDH-mutant or IDH-wildtype are then further classified accordingly:

IDH-mutant:

In order to define newly diagnosed glioma samples into one of the 3 glioma subtypes within IDH-mutants, we recommend applying Random Forest in a two-step process. 1) using the 1,308 tumor specific CpG probes which defines the IDHmut specific clusters (Fig S3A) and 2) using the 163 CpG probes which defines each TCGA IDH-mutant glioma subtype (Fig S3C).

1. If the sample was classified as IDHmut-K1 or IDHmut-K2 using the 1,308 tumor specific CpG probes for IDH-mutant and as G-CIMP-low using the 163 CpG probes defined by a supervised analysis across IDH-mutant subgroups, we classify the sample as G-CIMP-low;
2. If the sample was classified as IDHmut-K1 or IDHmut-K2 using the 1,308 tumor specific CpG probes for IDH-mutant and as G-CIMP-high using the 163 CpG probes defined by a supervised analysis across IDH-mutant subgroups, we classify the sample as G-CIMP-high;
3. If the sample was classified as IDHmut-K3 using the 1,308 tumor specific CpG probes for IDH-mutant, we classify the sample as Codel.

IDH-wildtype:

Likewise, IDH-wildtype can be classified using a single random forest machine-learning model applied with a signature defined by the 914 tumor specific CpG probes for IDH-wildtype (Figures S4A-B). Samples following into IDHwt-K3 (aka LGm6), we recommend subdividing this group based on grade, resulting in either LGm6-GBM and PA-like (LGG).

5.6 Patient centric table (DNA methylation)

To generate DNA methylation calls for each sample per gene per overlapping platforms (HM27, HM450), we began by first collapsing multiple CpGs to one representative gene. Using the associated gene expression data (organized as one gene - one expression value per sample), we merged the samples and CpG probes with gene expression data for each platform. We next calculated the spearman correlation (ρ) across all samples for all CpG probes for each gene to one gene expression value. For multiple CpGs for each annotated gene promoter, we selected one CpG probe with the lowest correlation rho value to the associated gene expression profile to capture the most biologically representative event (epigenetic silencing). This effectively reduced the number of CpG probes from N:1 to 1:1. Our data set was then reduced down to 636 samples x 19,486 CpG:Gene.

Next, we assigned discrete categories based on the spearman correlation rho value according to the following criteria:

1. Strongly negatively correlated (SNC) when ρ value is less than 0.5;
2. Weakly negatively correlated (WNC) when ρ value is between 0.5 and 0.25;
3. No negative correlation (NNC) when ρ value is greater than 0.25.

Next, we assigned samples to either the 10th (T10 or N10) or 90th (T90 or N90) percentile based on the observed beta-value across tumor samples (T) and normal samples (N). For the normal samples, we used 110 non-tumor TCGA samples from 11 different tissues previously described. We assigned labels for each gene per platform per tissue type (tumor and normal) according to the following rules:

1. If percentile 90 < 0.25, we assign it as CUN or CUT (constitutively unmethylated in normal or tumor);
2. If percentile 10 > 0.75, we assign it as CMN or CMT (constitutively methylated in normal or tumor);
3. If percentile 10 > 0.25 and percentile 90 < 0.75, we assign it as IMN or IMT (intermediate methylated in normal or tumor);
4. If it doesn't fall in any of the above categories, it is assign VMN or VMT (variably methylated in normal or tumor).

Next we assigned a 'call' and a confidence 'score' for each possible combinations (48) [3 (SNC, WNC, NNC) x 4 (CUN, CMN, VMN, IMN) x 4 (CUT, CMT, VMT, IMT)]. We created the following relationship for each call and score based on our interpretation of the most informative epigenetic event (e.g. promoter DNA hypermethylation and low expression). Users should understand that the selection and criteria performed were done to the best of our knowledge at the time. We felt most

confident with calling epigenetically silenced events and this is reflected in the confidence score. The methylation calls are as follows:

MG: Methylation gain compared to normal

ML: Methylation loss compared to normal

MT: Methylated in tumor

UT: Unmethylated in tumor

ES: Epigenetically silenced

UC: Unable to make call

Methylation class confidence scores vary from 0 (no call) to 4 (high confidence). Patient centric table can be accessed at https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/.

5.7 Homer de novo motif searches

De novo Motif discovery was performed using HOMER (script v4.4 (8-25-2014)), an algorithm previously described (Heinz et al., 2010). Briefly, differentially methylated probes were classified according to genomic location into CpG island, CpG shores, and open seas as follow: CpG islands were defined based on UCSC annotation and as per the criteria previously described (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). Coverage of CpG island regions was further enhanced by including the 2 kb regions flanking CpG island, referred to here as CpG shores. CpGs isolated in the genome were defined as open seas. Probes mapped to each region were used to performed de novo motif analysis using HOMER (HOMER perl script 'findMotifsGenome.pl'). To increase sensitivity of the method, up to two mismatches were allowed in each oligonucleotide sequence and distributions of CpG content in 'target' and 'background' sequences were selectively weighted to equalize the distributions of CpG content in both sets. Raw outputs from HOMER can be found at https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/ .

6. Reverse phase protein array (RPPA)

Authors: Rehan Akbani, Zhenlin Ju, Yiling Lu, Gordon Mills

Correspondence and questions should be directed to: (rakbani@mdanderson.org)

6.1 Data Processing

Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 mmol/L Hepes (pH 7.4), 150 mmol/L NaCl, 1.5 mmol/L MgCl2, 1 mmol/L EGTA, 100 mmol/L NaF, 10 mmol/L NaPPi, 10% glycerol, 1 mmol/L phenylmethylsulfonyl fluoride, 1 mmol/L Na3VO4, and aprotinin 10 ug/mL) from human tumors and RPPA was performed as described previously (Coombes, 2011; Hennessy et al., 2007; Hu et al., 2007; Liang et al., 2007; Tibes et al., 2006). Lysis buffer was used to lyse frozen

tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually serial diluted in two-fold of 5 dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 196 validated primary antibodies (Cancer Genome Atlas Research, 2015) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction. Slides were scanned in a CanoScan 9000F. Spot intensities were analyzed and quantified using Array-Pro Analyzer (Media Cybernetics Washington DC) to generate spot signal intensities (Level 1 data). The software SuperCurveGUI (Coombes, 2011; Hu et al., 2007), available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model (Tibes et al., 2006). During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric (Coombes, 2011) was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described (Coombes, 2011; Gonzalez-Angulo et al., 2011; Hu et al., 2007) using median centering across antibodies (level 3 data). In total, 196 antibodies and 473 samples were used. Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described (Hennessy et al., 2010). These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described (Hennessy et al., 2010).

Two RPPA arrays were quantitated and processed (including normalization and load controlling) as described previously, using MicroVigene (VigeneTech, Inc., Carlisle, MA) and the R package SuperCurve (version-1.3), available at http://bioinformatics.mdanderson.org/OOMPA (Hu et al., 2007; Tibes et al., 2006). Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

6.2 Data normalization

We performed median centering across all the antibodies for each sample to correct for sample loading differences. Those differences arise because protein concentrations are not uniformly distributed per unit volume. That may be due to several factors, such as differences in protein concentrations of large and small cells, differences in the amount of proteins per cell, or heterogeneity of the cells comprising the samples. By observing the expression levels across many different proteins in a sample, we can estimate differences in the total amount of protein in that sample vs. other samples. Subtracting the median protein expression level forces the median value to become zero, allowing us to compare protein expressions across samples.

Surprisingly, processing similar sets of samples on different slides of the same antibody may result in datasets that have very different means and variances. Neely et al. (2009) processed clinically similar ALL samples in two batches and observed differences in their protein data distributions. There were additive and multiplicative effects in the data that could not be accounted by biological or sample loading differences. We observed similar effects when we compared the two batches of GBM and LGG tumor protein expression data. A new algorithm, replicates-based normalization (RBN), was therefore developed using replicate samples run across multiple batches to adjust the data for batch effects. The underlying hypothesis is that any observed variation between replicates in different batches is primarily due to linear batch effects plus a component due to random noise. Given a sufficiently large number of replicates, the random noise is expected to cancel out (mean=zero by definition). Remaining differences are treated as systematic batch effects. We can compute those effects for each antibody and subtract them out. Many samples were run in both batches. One batch was arbitrarily designated the “anchor” batch and was to remain unchanged. We then computed the means and standard deviations of the common samples in the anchor batch, as well as the other batch. The difference between the means of each antibody in the two batches and the ratio of the standard deviations provided an estimate of the systematic effects between the batches for that antibody (both location-wise and scale-wise). Each data point in the non-anchor batch was adjusted by subtracting the difference in means and multiplying by the inverse ratio of the standard deviations to cancel out those systematic differences. Our normalization procedure significantly reduced technical effects, thereby allowing us to merge the datasets from different batches.

6.3 Clustering

We used consensus clustering to cluster the samples in an unsupervised way, with Pearson correlation as the distance metric and Ward as the linkage algorithm. A total of 473 samples and

196 antibodies were used in the analysis. Two clusters were observed that largely corresponded with tumor type (Figure S3E), however, there were a few notable exceptions. Whereas only one GBM sample clustered with the LGG samples, twenty-six LGG samples were found to cluster with the GBM samples. Seventeen of those twenty-six samples had no mutations in IDH1/2, similar to the GBM samples. Furthermore, compared to the LGG-like cluster, the GBM-like cluster had elevated expression of IGFBP2, fibronectin, PAI1, HSP70, EGFR, phosphoEGFR, phosphoAKT, Cyclin B1, Caveolin, Collagen VI, Annexin1 and ASNS, whereas it had low expression of PKC (alpha, beta and delta), PTEN, BRAF, and phosphoP70S6K.

7. Regulome Explorer

Authors: Geetika Sethi, Brady Bernard, Vesteinn Thorsson, Sheila Reynolds, Lisa Lype, Ilya Shmulevich

Correspondence and questions should be directed to: ilya.shmulevich@systemsbiology.org

7.1. Feature Matrix

Associations among the diverse clinical and molecular data are identified through construction of a “feature matrix” (FM) by integrating values from all data types. Each column in the FM represents one of the 1123 tumor samples. Each row in the FM represents a single clinical, sample or molecular data element (mRNA expression levels, microRNA expression levels, protein levels (RPPA), copy number alterations, DNA methylation levels and somatic mutations), and the individual data values may be numerical (continuous or discrete) or categorical, as appropriate. Missing values are indicated within the FM by “NA”, and the number of non-NA data values varies significantly across the different data types (rows). Data were retrieved from the DCC on November 18, 2015 and further processed as follows. Clinical and sample data (633 features): DCC clinical and sample data were processed into a matrix. Cluster assignments: Cluster memberships resulting from unsupervised clustering for each of the individual molecular data types: SCNA (Supplement 3), RNAseq (Supplement 4), DNA methylation (Supplement 5), and RPPA (Supplement 6) were incorporated into the FM. Mutation

rates and categories (Supplement 2) were included in the FM as well. Molecular datasets include Gene expression (15,401 features): Gene level RSEM values from RNA-seq (Supplement SA) were log2 transformed, and filtered to remove low-variability genes (bottom 25% removed, based on interdecile range). MicroRNA expression (692 features): The summed and normalized microRNA quantification files were log2 transformed, and filtered to remove low-variability microRNAs (bottom 25% based on zero-count). Somatic copy number alterations: Copy number and focal copy number changes were obtained for peaks identified by GISTIC as described above (Supplement 3, 6318 features). DNA methylation (19,727 features): Probe-specific level-3 β-values were obtained as described above (Supplement). We started with the probes common between the two methylation platforms, and then removed the bottom 25% based on interdecile range. Somatic mutations (2842): The Mutations Annotation Format file (Supplement 1), was used to generate a binary indicator vector indicating whether a particular non-silent mutation is present in a specific sample. Mutation features found in fewer than two tumor samples were removed. Overall, the gbm_lgg feature matrix has 45839 features (inclusive of the above mentioned analysis platforms) for all the 1122 patients (data freeze list) resulting in 51477197 matrix elements (48501 x 38), with approximately 89% non-NA elements (197478 out of 1843038).

The Synapse platform by Sage Bionetworks (www.sagebase.org, [1]) was used during the development of this project for distributing versioned data to project researchers and as a staging area for assembling files into the Feature Matrix.

7.2. All-by-all Pairwise Associations

Statistical association among the diverse data elements in this study was evaluated by comparing pairs of columns in the feature matrix. Hypothesis testing was performed by testing

against null models for absence of association, yielding a p-value. P-values for the association between and among clinical and molecular data elements were computed according to the nature of the data levels for each pair: discrete vs. discrete (Fisher's exact test); discrete vs. continuous (ANOVA F- test, equivalently t-test for binary vs. continuous) or continuous vs. continuous (F-test). Ranked data values were used in each case. To account for multiple-testing bias, the p-value was adjusted using the Bonferroni correction. Exploring potentially interesting genomic relationships have been of interest to researchers previously [2]. In order to allow researchers to further explore genomic associations in TCGA gbm_lgg dataset, including primary data, the statistically significant pairs of associations were loaded into the Regulome Explorer web application, which is designed to enable researchers to explore associations among multiple data types in cancer genomics. Prior to loading, a p-value threshold was chosen specific to each pair of data types in such a way as to strike a balance between making potentially interesting associations available to queries by the tool, while still allowing the tool to be responsive, since the number of loaded graph edges (each corresponding to a statistically significant relationship) is in the millions. All identified pairwise relationships, including those described in this manuscript can be found at http://explorer.cancerregulome.org.

Regulome explorer references

1. Omberg, L., et al., *Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas.* Nat Genet, 2013. **45**(10): p. 1121-6.

2. Sethi, G., et al., *An RNA interference lethality screen of the human druggable genome to identify molecular vulnerabilities in epithelial ovarian cancer.* PLoS One, 2012. **7**(10): p. e47086.

8. Supplemental References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pages, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091-1093.
Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H.*, et al.* (2013). The somatic genomic landscape of glioblastoma. Cell *155*, 462-477.
Cancer Genome Atlas Research, N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068.
Cancer Genome Atlas Research, N. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43-49.
Cancer Genome Atlas Research, N. (2014a). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202-209.
Cancer Genome Atlas Research, N. (2014b). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315-322.
Cancer Genome Atlas Research, N. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med.
Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A.*, et al.* (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol *30*, 413-421.
Chen, Y.J., Campbell, H.G., Wiles, A.K., Eccles, M.R., Reddel, R.R., Braithwaite, A.W., and Royds, J.A. (2008). PAX8 regulates telomerase reverse transcriptase and telomerase RNA component in glioma. Cancer research *68*, 5724-5732.
Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology *31*, 213-219.
Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T., Pagnotta, S.M., Castiglioni, I.*, et al.* (2015). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. doi: 10.1093/nar/gkv1507.
Coombes, K.N., S.; Joy, C.; Hu, J.; Baggerly, K.; et al. (2011). SuperCurve Package. R package version 1.4.1.
Ding, Z., Mangino, M., Aviv, A., Spector, T., Durbin, R., and Consortium, U.K. (2014). Estimating telomere length from whole genome sequence data. Nucleic acids research *42*, e75.
Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. Plos Biol *5*, 54-66.
Frattini, V., Trifonov, V., Chan, J.M., Castano, A., Lia, M., Abate, F., Keir, S.T., Ji, A.X., Zoppoli, P., Niola, F.*, et al.* (2013). The integrated landscape of driver genomic alterations in glioblastoma. Nature genetics *45*, 1141-1149.
Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol *196*, 261-282.
Gleize, V., Alentorn, A., Connen de Kerillis, L., Labussiere, M., Nadaradjane, A., Mundwiller, E., Ottolenghi, C., Mangesius, S., Rahimian, A., Ducray, F.*, et al.* (2015). CIC inactivating mutations identify aggressive subset of 1p19q codeleted gliomas. Annals of neurology.
Golbeck, J., and Mutton, P. (2005). Spring-Embedded graphs for semantic visualization. Visualizing the Semantic Web, 172-182.
Gonzalez-Angulo, A.M., Hennessy, B.T., Meric-Bernstam, F., Sahin, A., Liu, W., Ju, Z., Carey, M.S., Myhre, S., Speers, C., Deng, L.*, et al.* (2011). Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin Proteomics *8*, 11.

Guintivano, J., Aryee, M.J., and Kaminsky, Z.A. (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics *8*, 290-302.
Harrell, F.E., Jr., Califf, R.M., Pryor, D.B., Lee, K.L., and Rosati, R.A. (1982). Evaluating the yield of medical tests. JAMA *247*, 2543-2546.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.
Hennessy, B.T., Lu, Y., Gonzalez-Angulo, A.M., Carey, M.S., Myhre, S., Ju, Z., Davies, M.A., Liu, W., Coombes, K., Meric-Bernstam, F*., et al.* (2010). A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. Clin Proteomics *6*, 129-151.
Hennessy, B.T., Lu, Y.L., Poradosu, E., Yu, Q.H., Yu, S.X., Hall, H., Carey, M.S., Ravoori, M., Gonzalez-Angulo, A.M., Birch, R*., et al.* (2007). Pharmacodynamic markers of perifosine efficacy. Clin Cancer Res *13*, 7421-7431.
Hu, J., He, X., Baggerly, K.A., Coombes, K.R., Hennessy, B.T., and Mills, G.B. (2007). Non-parametric quantification of protein lysate arrays. Bioinformatics *23*, 1986-1994.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.
Hung, N., Chen, Y.J., Taha, A., Olivecrona, M., Boet, R., Wiles, A., Warr, T., Shaw, A., Eiholzer, R., Baguley, B.C*., et al.* (2014). Increased paired box transcription factor 8 has a survival function in glioma. BMC cancer *14*, 159.
Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A*., et al.* (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell *150*, 1107-1120.
Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.
Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.
Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research *22*, 568-576.
Lambert, S.R., Witt, H., Hovestadt, V., Zucknick, M., Kool, M., Pearson, D.M., Korshunov, A., Ryzhova, M., Ichimura, K., Jabado, N*., et al.* (2013). Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. Acta neuropathologica *126*, 291-301.
Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013a). Software for computing and annotating genomic ranges. PLoS Comput Biol *9*, e1003118.
Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A*., et al.* (2013b). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.
Liang, J., Shao, S.H., Xu, Z.X., Hennessy, B., Ding, Z., Larrea, M., Kondo, S., Dumont, D.J., Gutterman, J.U., Walker, C.L*., et al.* (2007). The energy sensing LKB1-AMPK pathway regulates p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis. Nat Cell Biol *9*, 218-224.
McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A*., et al.* (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature genetics *40*, 1166-1174.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M.G., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N.*, et al.* (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS computational biology *7*, e1001138.
Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology *12*, R41.
Mur, P., Mollejo, M., Ruano, Y., de Lope, A.R., Fiano, C., Garcia, J.F., Castresana, J.S., Hernandez-Lain, A., Rey, J.A., and Melendez, B. (2013). Codeletion of 1p and 19q determines distinct gene methylation and expression profiles in IDH-mutated oligodendroglial tumors. Acta neuropathologica *126*, 277-289.
Neeley, E.S., Kornblau, S.M., Coombes, K.R., and Baggerly, K.A. (2009). Variable slope normalization of reverse phase protein arrays. Bioinformatics *25*, 1384-1389.
Radenbaugh, A.J., Ma, S., Ewing, A., Stuart, J.M., Collisson, E.A., Zhu, J., and Haussler, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. PloS one *9*, e111516.
Ramos, A.H., Lichtenstein, L., Gupta, M., Lawrence, M.S., Pugh, T.J., Saksena, G., Meyerson, M., and Getz, G. (2015). Oncotator: cancer variant annotation tool. Hum Mutat *36*, E2423-2429.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Smyth, G.K. (2005). Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor, 397-420.
Spearman, C. (1904). The proof and measurement of association between two things. Am J Psychol *15*, 72-101.
Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D.A., Jones, D.T., Konermann, C., Pfaff, E., Tonjes, M., Sill, M., Bender, S.*, et al.* (2012). Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer cell *22*, 425-437.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.
Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A *99*, 3740-3745.
TCGA_Network (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower Grade Gliomas. New England Journal of Medicine, in press.
Therneau, T.M. (2014). A package for survival analysis in S. version 2.37-7. http://CRANR-projectorg/package=survival.
Therneau, T.M., and Grambsch, P.M. (2000). Modeling survival data : extending the Cox model (New York: Springer).
Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G.B., and Kornblau, S.M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther *5*, 2512-2521.
Torres-Garcia, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M.F., Weinstein, J.N., Getz, G., and Verhaak, R.G. (2014). PRADA: pipeline for RNA sequencing data analysis. Bioinformatics *30*, 2224-2226.
Trifonov, V., Pasqualucci, L., Dalla Favera, R., and Rabadan, R. (2013). MutComFocal: an integrative approach to identifying recurrent and focal genomic alterations in tumor samples. BMC Syst Biol *7*, 25.
Turcan, S., Rohle, D., Goenka, A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W., Lu, C., Ward, P.S.*, et al.* (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature *483*, 479-483.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J.C., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237-i245.
Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P.*, et al.* (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell *17*, 98-110.
Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R.G. (2014). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene.