Supplementary figures

Machine learning-based analysis of glioma grades reveals co-enrichment

Mateusz Garbulowski[1], Karolina Smolinska[1], Uğur Çabuk[1,2], Sara A. Yones[1], Ludovica Celli[1,3], Esmanur Yaz[1,4], Fredrik Barrenäs[1,5], Klev Diamanti[1,6], Claes Wadelius[6] and Jan Komorowski[1,5,7,8]

[1]Department of Cell and Molecular Biology, Uppsala University, Sweden

[2]Polar Terrestrial Environmental Systems, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam, Germany

[3]Institute of Molecular Genetics Luigi Luca Cavalli-Sforza, National Research Council, Pavia, Italy

[4]Istanbul Medipol University, Biomedical Engineering and Bioinformatics, Istanbul, Turkey

[5]Washington National Primate Research Center, Seattle, WA, USA

[6]Department of Immunology, Genetics and Pathology, Uppsala University, Sweden

[7]Swedish Collegium for Advanced Study, Uppsala, Sweden

[8]Institute of Computer Science, Polish Academy of Sciences, Poland

Correspondence: Jan Komorowski, jan.komorowski@icm.uu.se, Department of Cell and Molecular Biology, Computational Biology and Bioinformatics, Telephone:+4618-471 6692, Mobile phone:+46 70 4250775

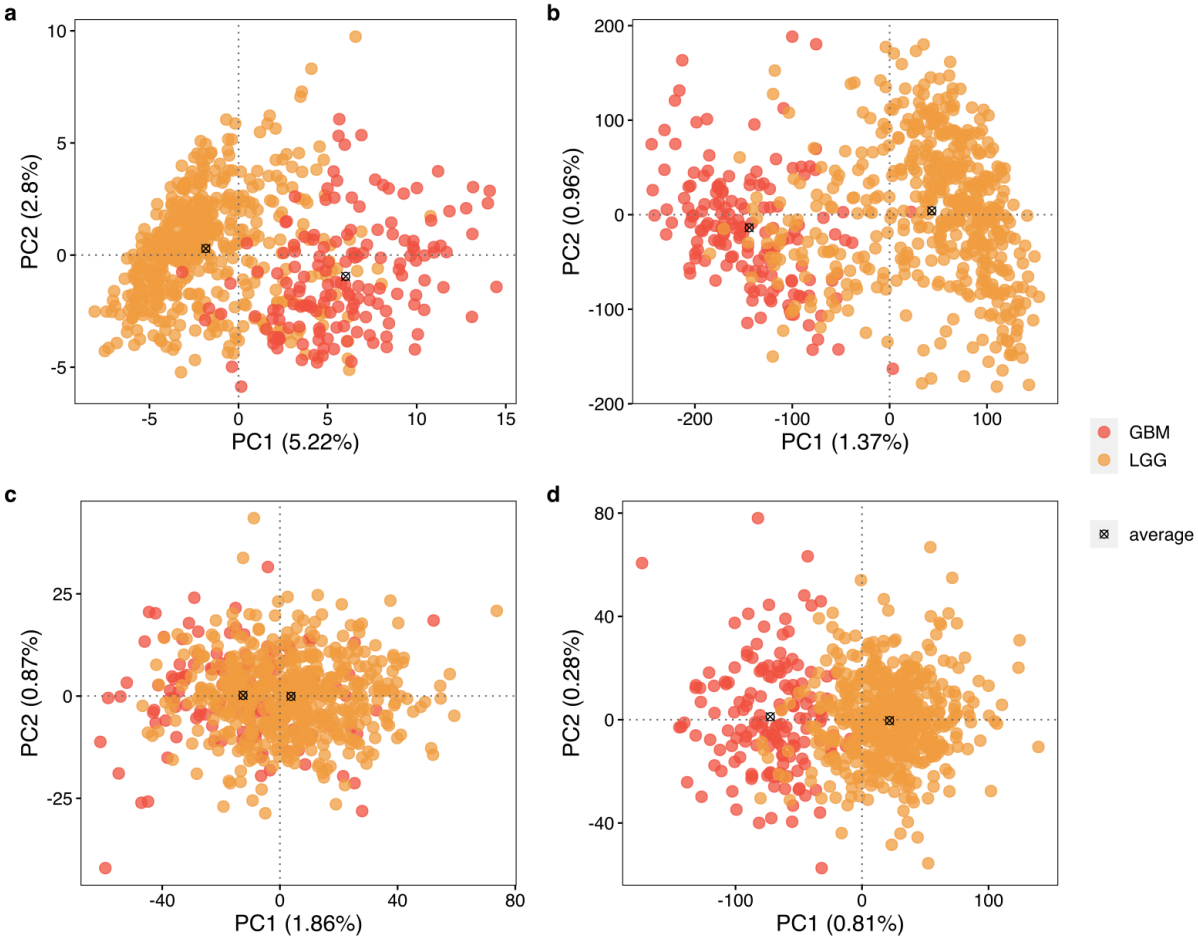


Fig. S1 Principal component analysis (PCA) to investigate the variance for The Cancer Genome Atlas (TCGA) cohorts for Lower Grade Glioma (LGG) and Glioblastoma Multiforme (GBM). **a.** original data from the GDC repository **b.** unified cohorts from UCSC Xena repository **c.** batch effect correction of unified cohorts with leek method with 532 surrogate variables **d.** batch effect correction of unified cohorts with be method with 48 surrogate variables

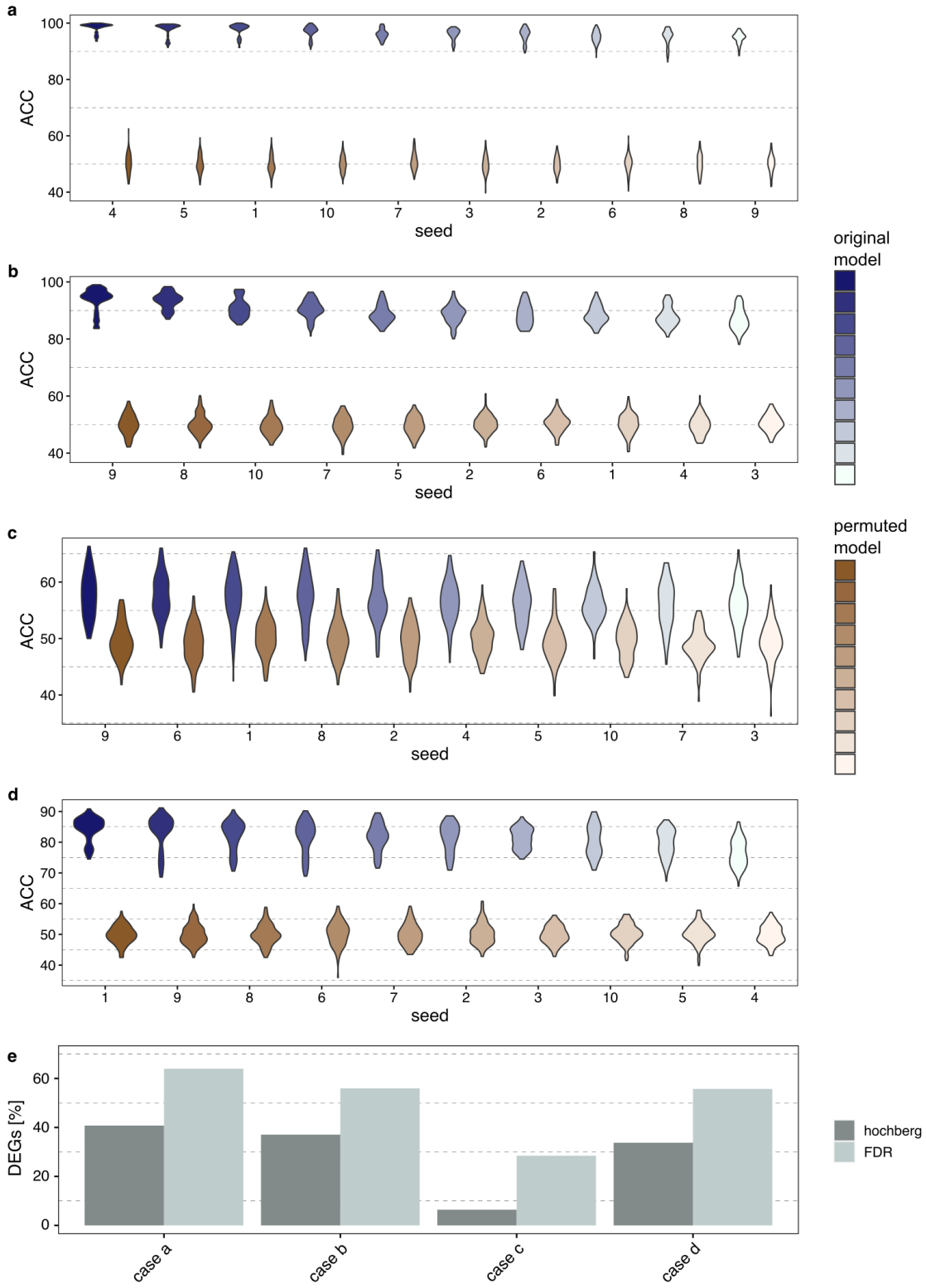


Fig. S2 Machine learning analysis performed on 100 randomly selected genes with 10 various seed values to investigate the bias for TCGA cohorts of LGGs and GBMs. **a.** original data from the GDC repository **b.** unified cohorts from UCSC Xena repository **c.** batch effect correction of unified cohorts with Leek method with 532 surrogate variables **d.** batch effect correction of unified cohorts with be method with 48 surrogate variables **e.** Student's t-test results for a particular cases a-d corresponding to subplots a-d.

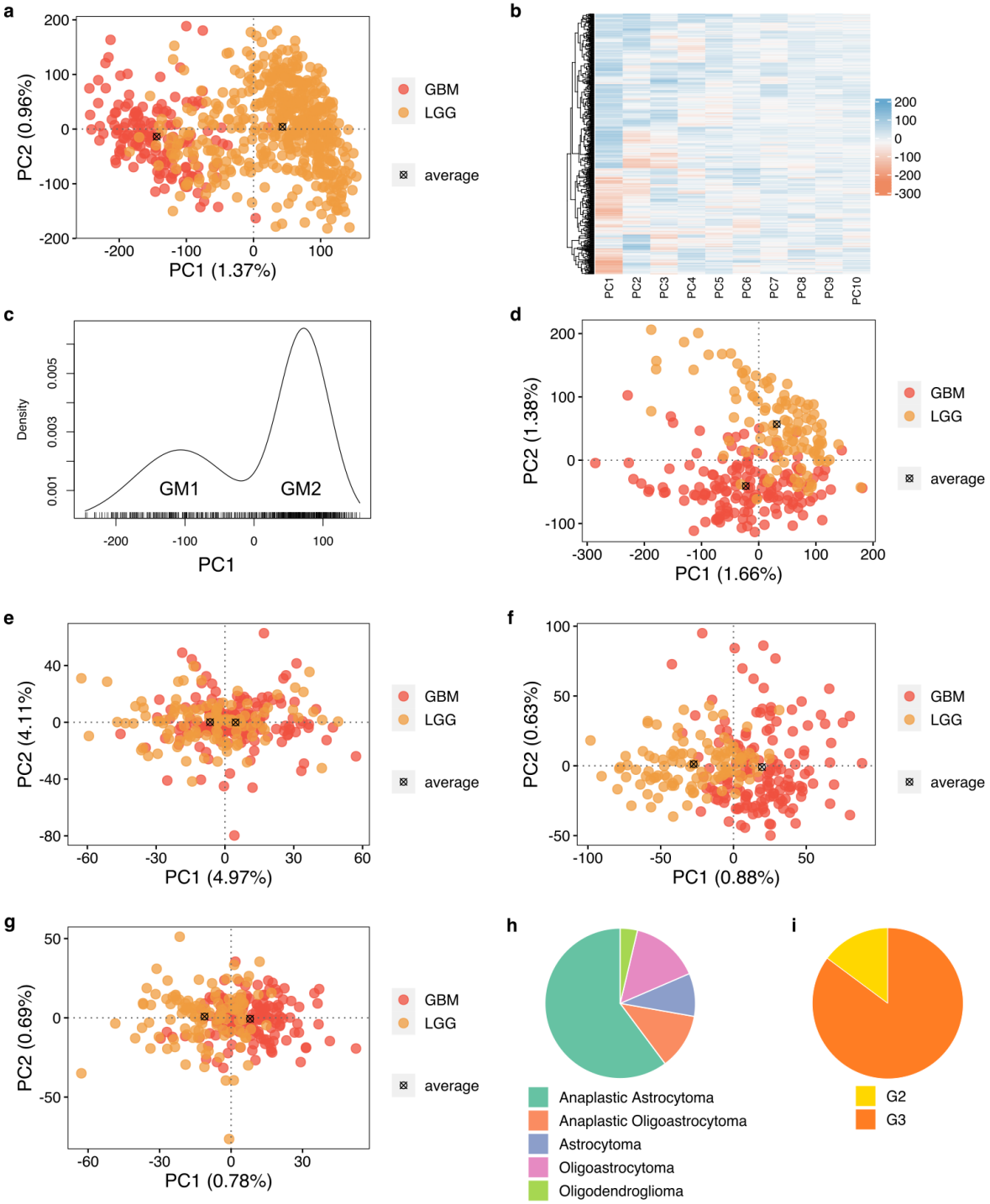


Fig. S3 Data evaluation to localize and remove bias from the glioma cohorts **a.** PCA for unified cohorts from UCSC Xena repository **b.** Clustering of 1-10 PCs of unified cohorts from UCSC Xena repository **c.** Gaussian mixtures (GM) detected from PC1. In GM2-based sample sets, two GBMs were included in the mixture and excluded in further analyses. **d.** PCA performed on LGG and GBM samples selected based on GM1 **e.** batch effect correction performed on LGG and GBM samples selected based on GM1 with the Leek method for 232 surrogate variables **f.** batch effect correction with the be method for 30 surrogate variables performed on samples selected based on GM1 **g.** the final data set based on GM1 after batch effect correction and filtration for protein coding genes (**h, i)** Evaluation of LGG samples included in GM1 based data set.

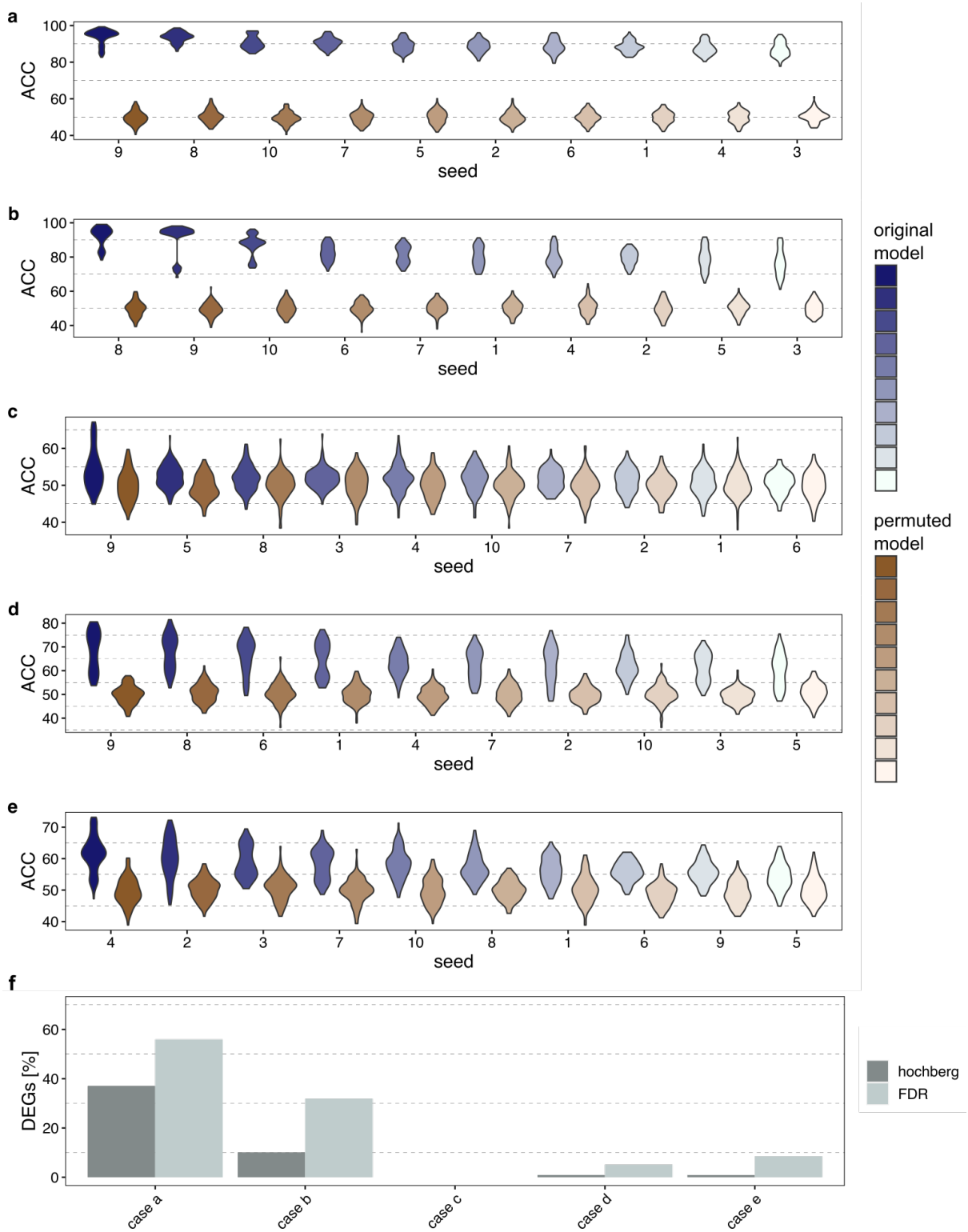


Fig. S4 Machine learning analysis performed on 100 randomly selected genes with 10 various seed values to investigate the bias for TCGA cohorts of LGGs and GBMs. **a.** unified cohorts from UCSC Xena repository **b.** GM2 samples from unified cohorts from UCSC Xena repository **c.** batch effect correction of 259 GM2-based samples with leek method with 232 surrogate variables **d.** batch effect correction of 259 GM2-based samples with be method with 30 surrogate variables **e.** the final data set based on GM1 after batch effect correction and filtration for protein coding genes **f.** Student's t-test results for a particular cases a-e corresponding to subplots a-e.

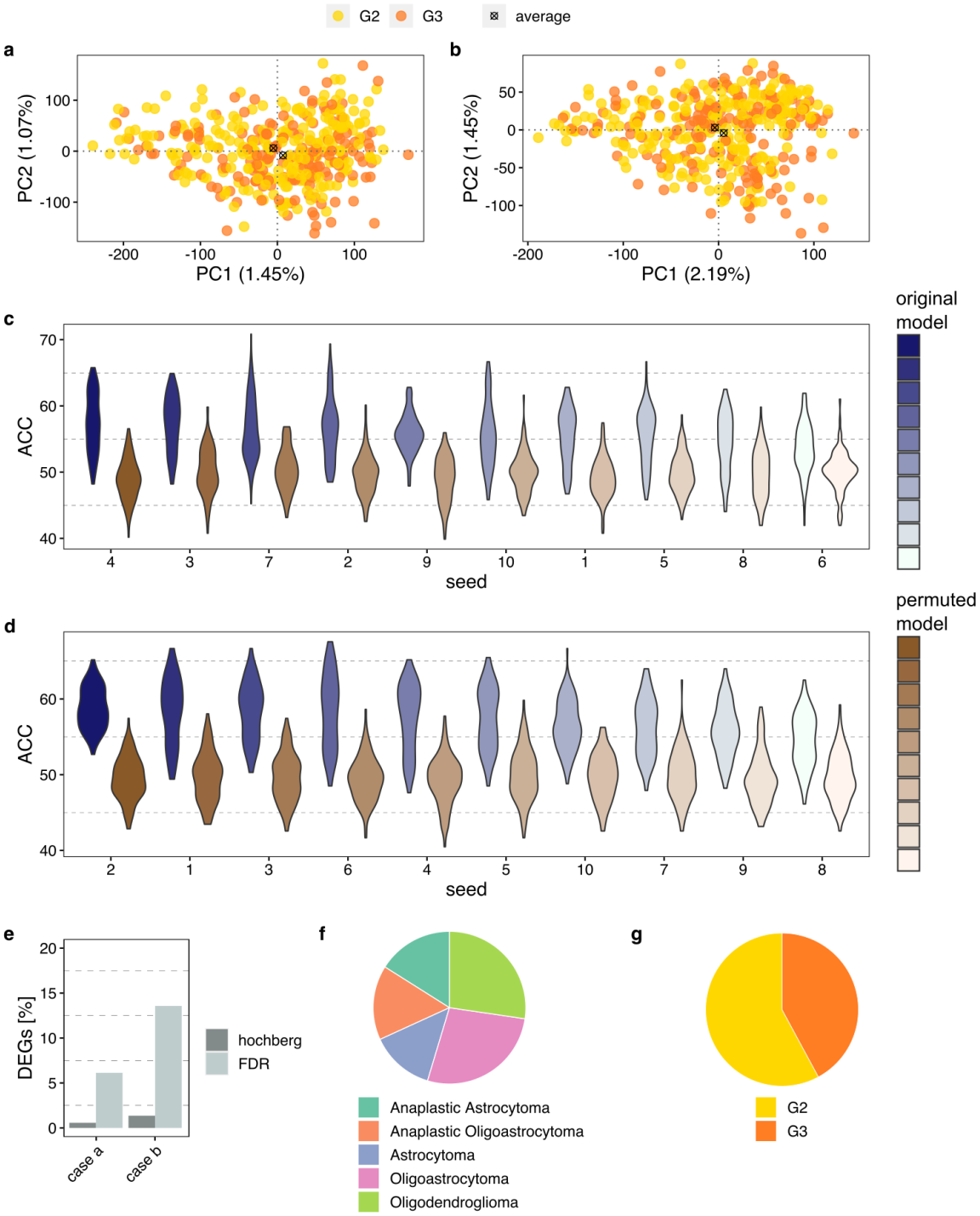


Fig. S5 The evaluation of samples in GM2 consisting of vast majority of LGGs. **a.** PCA for LGGs based on GM2 samples **b.** PCA for LGGs based on GM2 samples for protein coding genes only **c.** Machine learning evaluation of GM2 samples for discerning between GII and GIII. **d.** Machine learning evaluation of GM2 samples for discerning GII and GIII for protein coding genes only. **e.** Evaluation of fraction of statistically significant genes discerning between GII and GIII with all genes (case a) and protein coding genes only (case b) **f.** Proportion of grades within a GM2-based data set for LGGs.

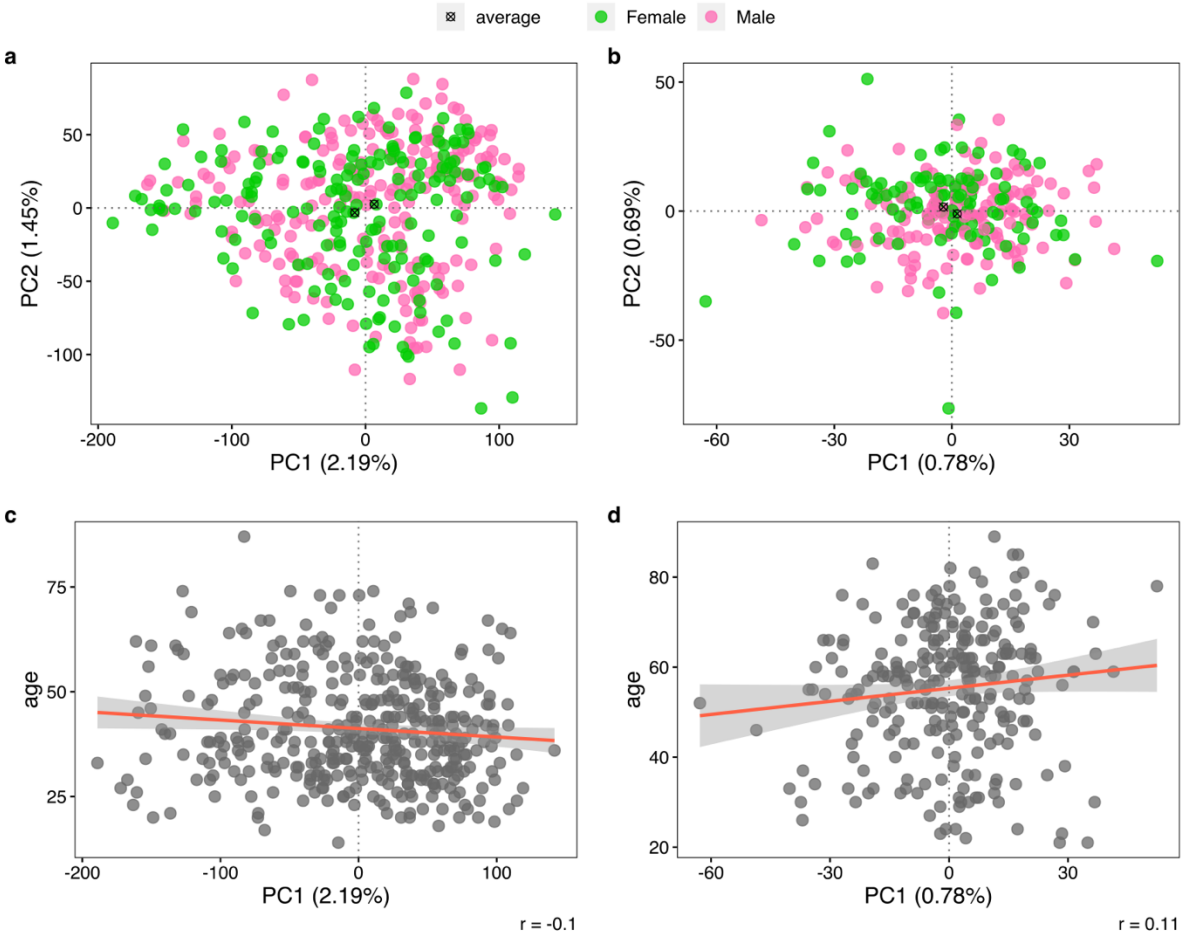


Fig. S6 The evaluation of other clinical factors within GM1- and GM2-based data sets **a.** PCA for GM1 for LGGs and their sex information **b.** PCA for GM2 for LGGs and GBMs and their sex information **c.** the variation of age of LGG samples based on GM1. Pearson correlation (r) value is marked in the plot caption. **d.** the variation of age of LGG and GBM samples based on GM2. Pearson correlation (r) value is marked in the plot caption.

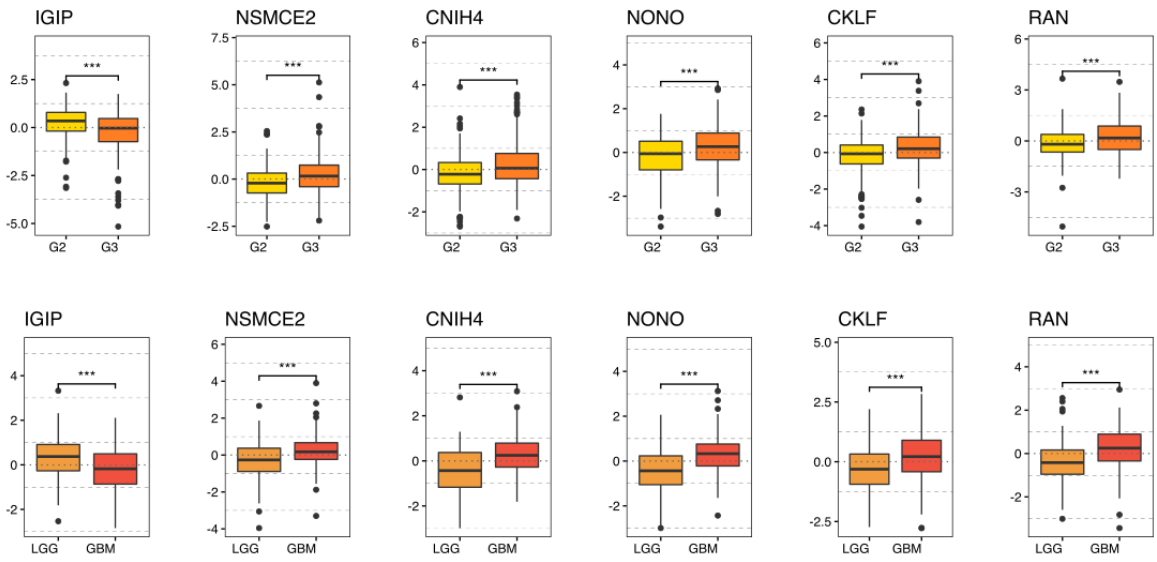


Fig. S7 Expression profiles for six common differentially expressed genes (DEGs) selected based on intersection of DEGs list between GII vs GIII and LGG vs GBM. Gene expression profiles were generated based on TCGA cohorts. *P* values were marked on boxplots as ns ($p > 0.05$), *($p \leq 0.05$), **($p \leq 0.01$), and ***($p \leq 0.001$).

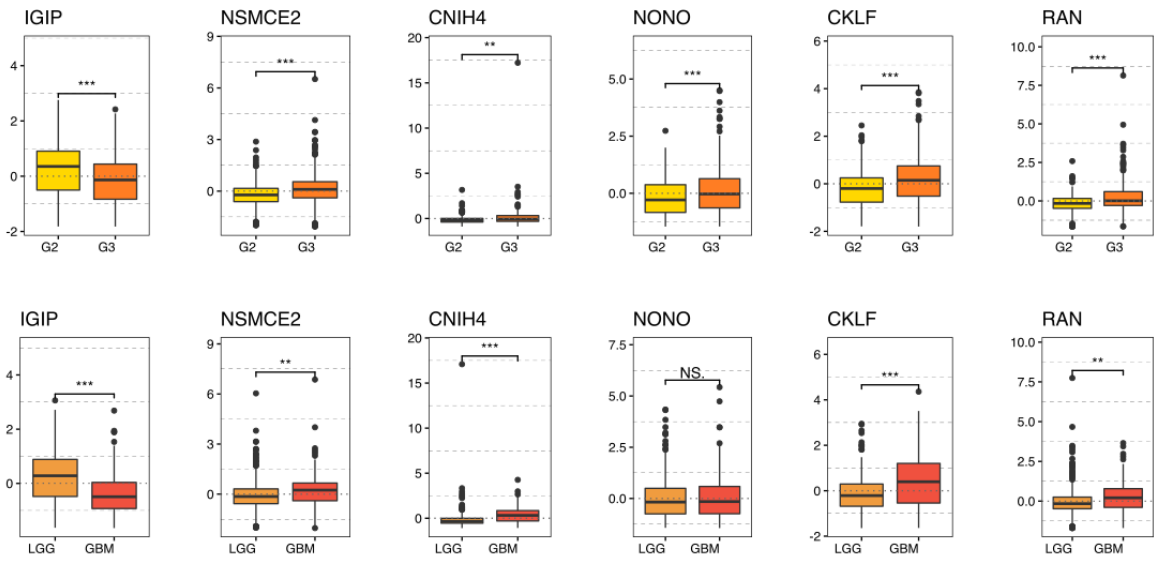


Fig. S8 Expression profiles for six common DEGs selected based on intersection of DEGs list between GII vs GIII and LGG vs GBM. Gene expression profiles were generated based on Chinese Glioma Genome Atlas (CGGA) batch 1 cohort. *P* values were marked on boxplots as ns ($p > 0.05$), *($p \leq 0.05$), **($p \leq 0.01$), and ***($p \leq 0.001$).

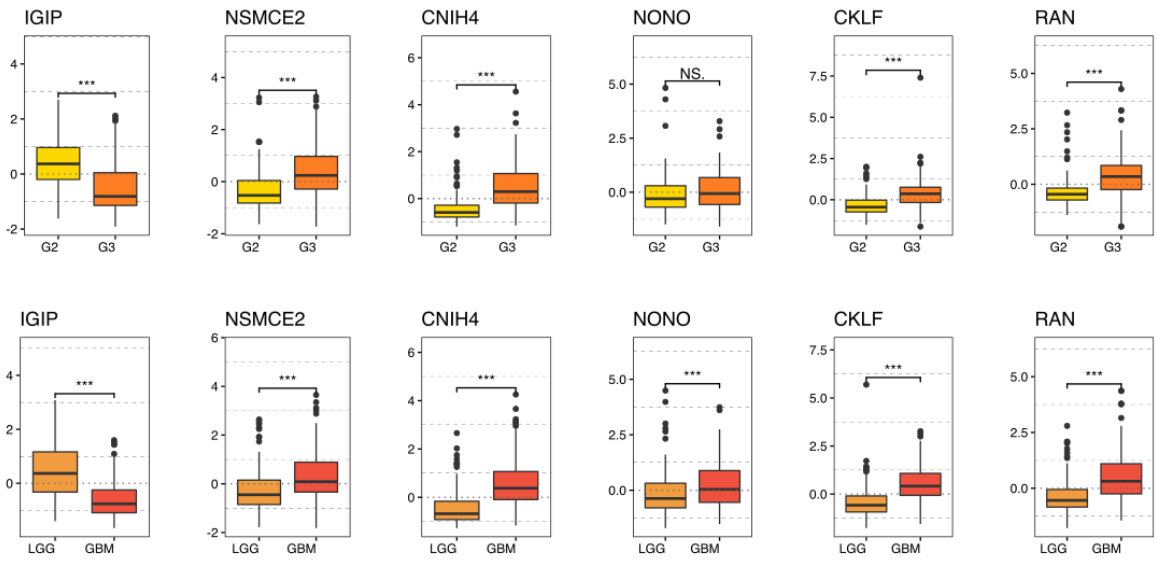


Fig. S9 Expression profiles for six common DEGs selected based on intersection of DEGs list between GII vs GIII and LGG vs GBM. Gene expression profiles were generated based on CGGA batch 2 cohort. *P* values were marked on boxplots as ns ($p > 0.05$), *($p \leq 0.05$), **($p \leq 0.01$), and ***($p \leq 0.001$).

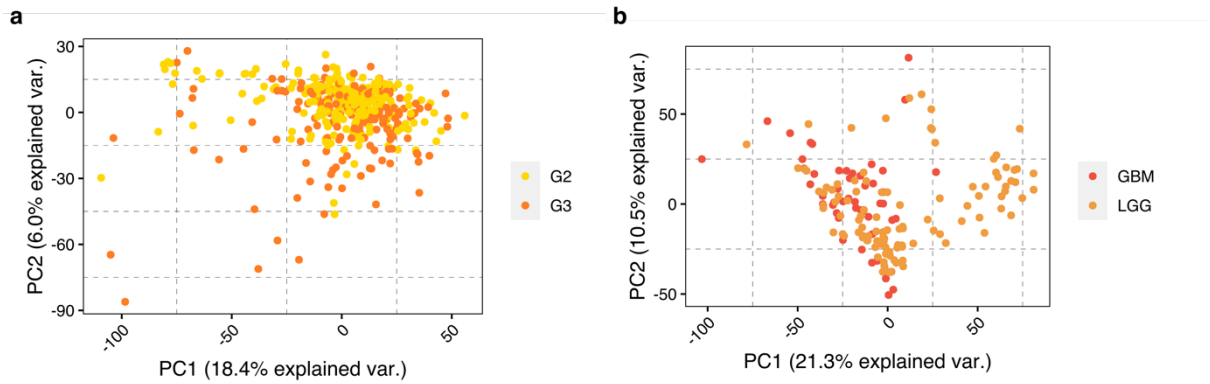


Fig. S10 Variability in DNA methylation data from TCGA based on **a**. GM2 and **b**. GM1 samples.

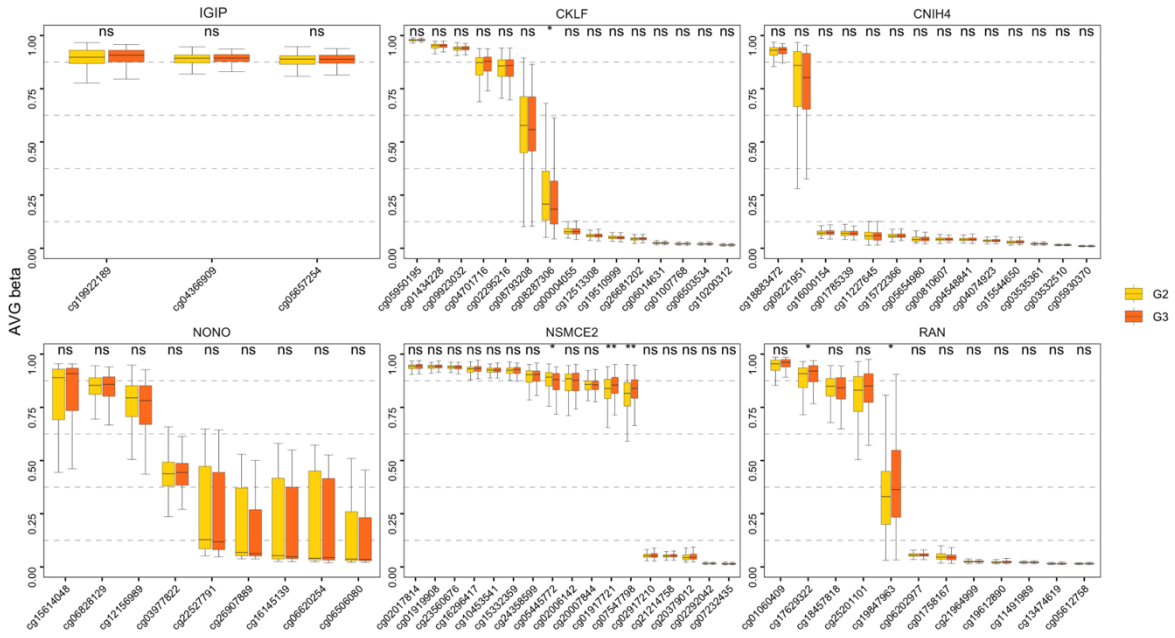


Fig. S11 DNA methylation status of six common DEGs for GII vs GIII. *P* values were marked on boxplots as ns ($p > 0.05$), *($p \leq 0.05$), **($p \leq 0.01$), and ***($p \leq 0.001$).

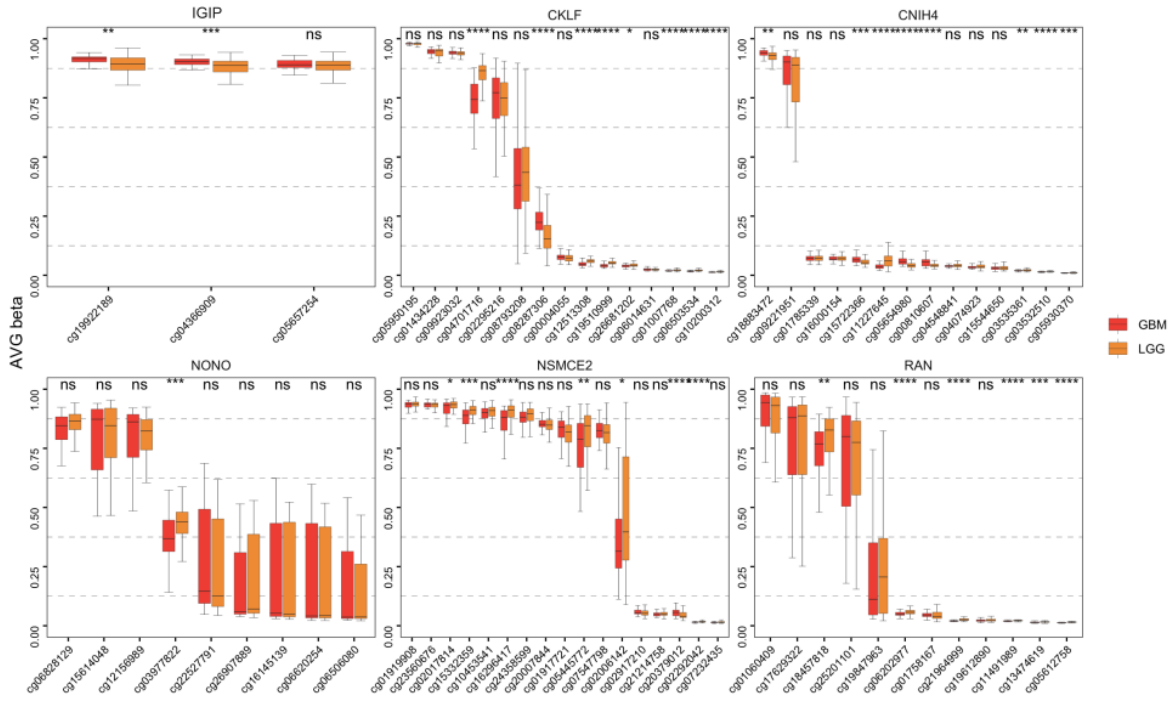


Fig. S12 DNA methylation status of six common DEGs for LGG vs GBM. *P* values were marked on boxplots as ns ($p > 0.05$), *($p \leq 0.05$), **($p \leq 0.01$), and ***($p \leq 0.001$).

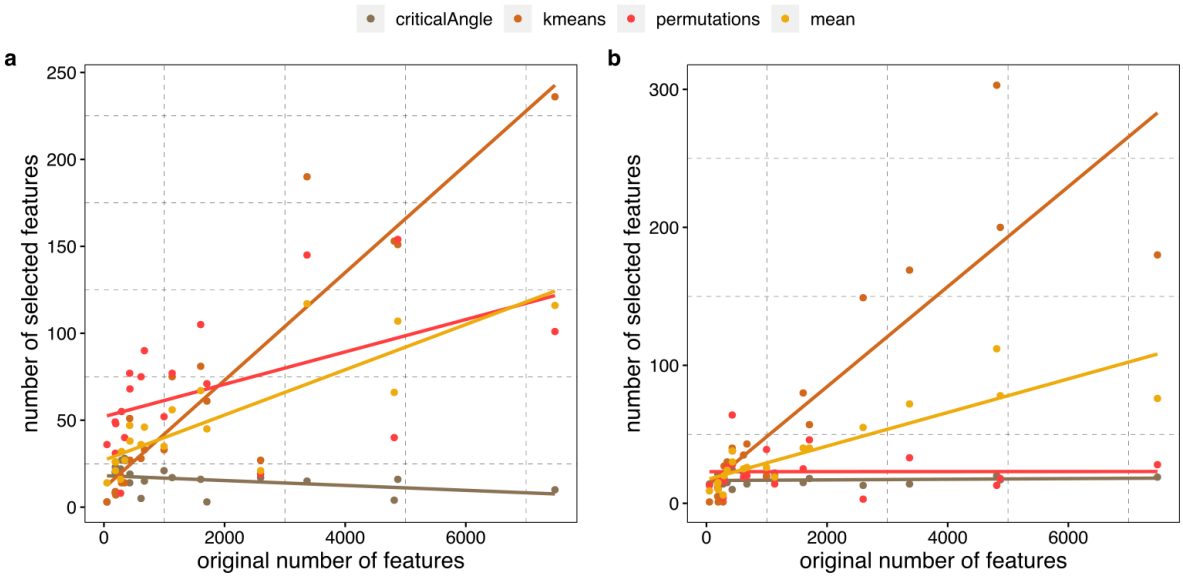


Fig. S13 The Monte Carlo feature selection (MCFS) evaluation of thresholds used for selecting top features. **a.** all 20 Molecular Signatures Database (MSigDB) collections for GII vs GIII **b.** all 20 MSigDB collections for LGG vs GBM.

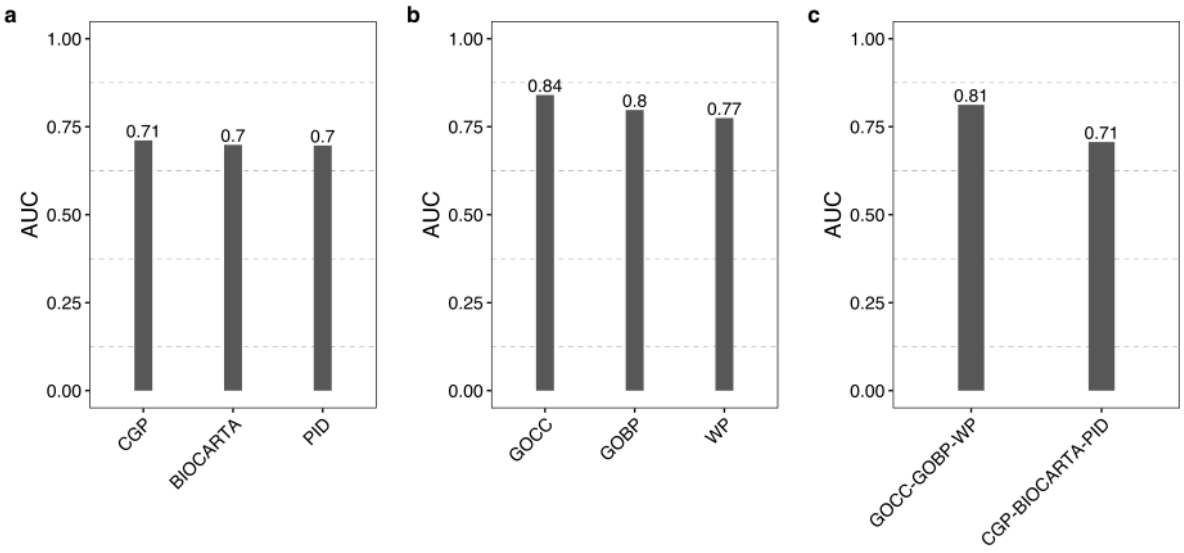


Fig. S14 Evaluation of MSigDB collections with R.ROSETTA rule-based learning for discerning glioma grades using single-sample Gene Set Enrichment Analysis (ssGSEA) scores based on TCGA cohorts. **a.** top three MSigDB collections for discerning between GII and GIII. **b.** top three MSigDB collections for discerning between LGG and GBM **c.** merged top three MSigDB collections for two discerning between GII and GIII (right bar), and between LGG and GBM (left bar).

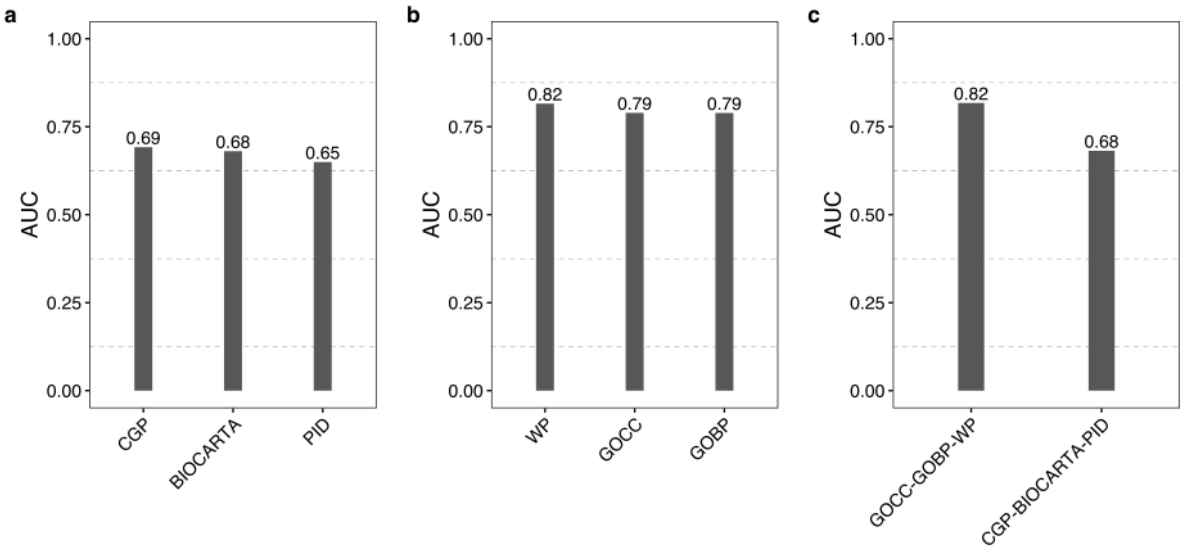


Fig. S15 Evaluation of MSigDB collections with R.ROSETTA rule-based learning for discerning glioma grades using ssGSEA scores based on CGGA batch 1. **a.** top three MSigDB collections for discerning between GII and GIII. **b.** top three MSigDB collections for discerning between LGG and GBM **c.** merged top three MSigDB collections for two discerning between GII and GIII (right bar), and between LGG and GBM (left bar).

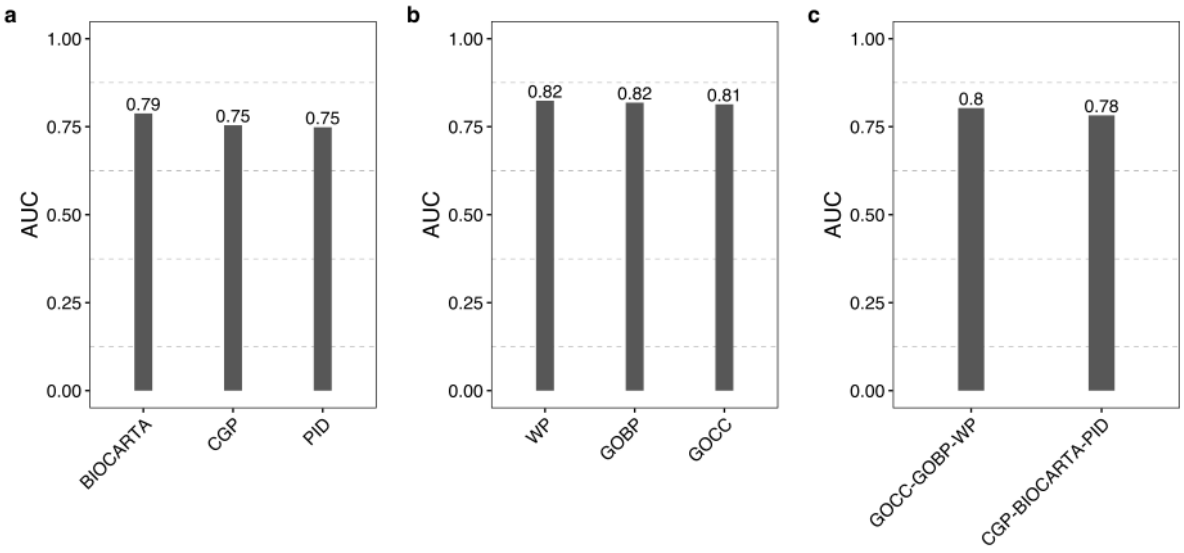


Fig. S16 Evaluation of MSigDB collections with R.ROSETTA rule-based learning for discerning glioma grades using ssGSEA scores based on CGGA batch 2. **a.** top three MSigDB collections for discerning between GII and GIII. **b.** top three MSigDB collections for discerning between LGG and GBM **c.** merged top three MSigDB collections for two discerning between GII and GIII (right bar), and between LGG and GBM (left bar).

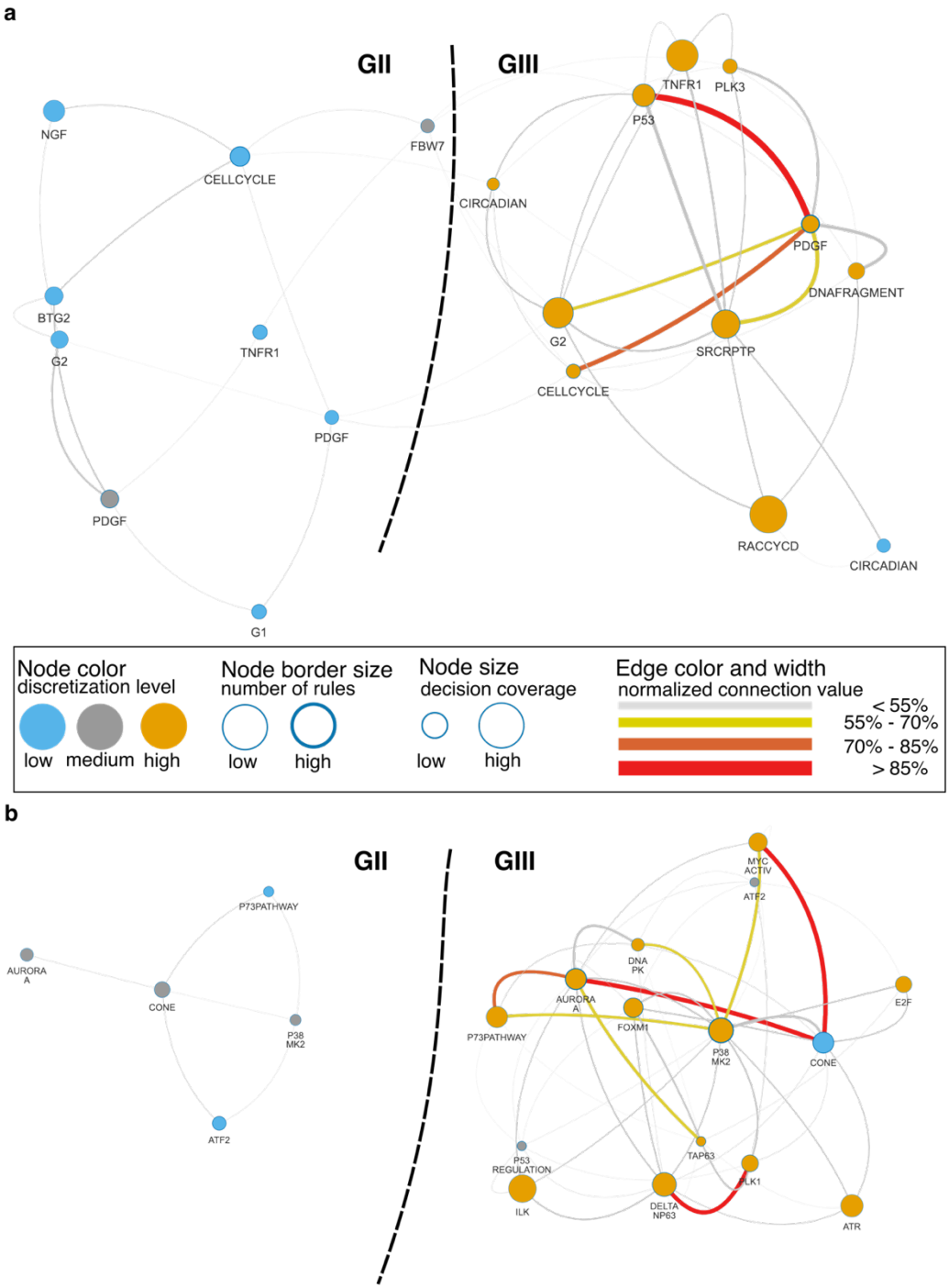


Fig. S17 Rule-based network displaying the most relevant co-enrichments of annotations obtained from **a.** BioCarta and **b.** PID collections for GII vs GIII models. The networks show 20 most connected nodes obtained from top 10% based on the rule connection from a set of significant rules (FDR-adjusted *P* value < 0.01). Connection values of nodes and edges represent a strength of co-enrichment from the classifier. Black dashed lines divide decision-related subnetworks.

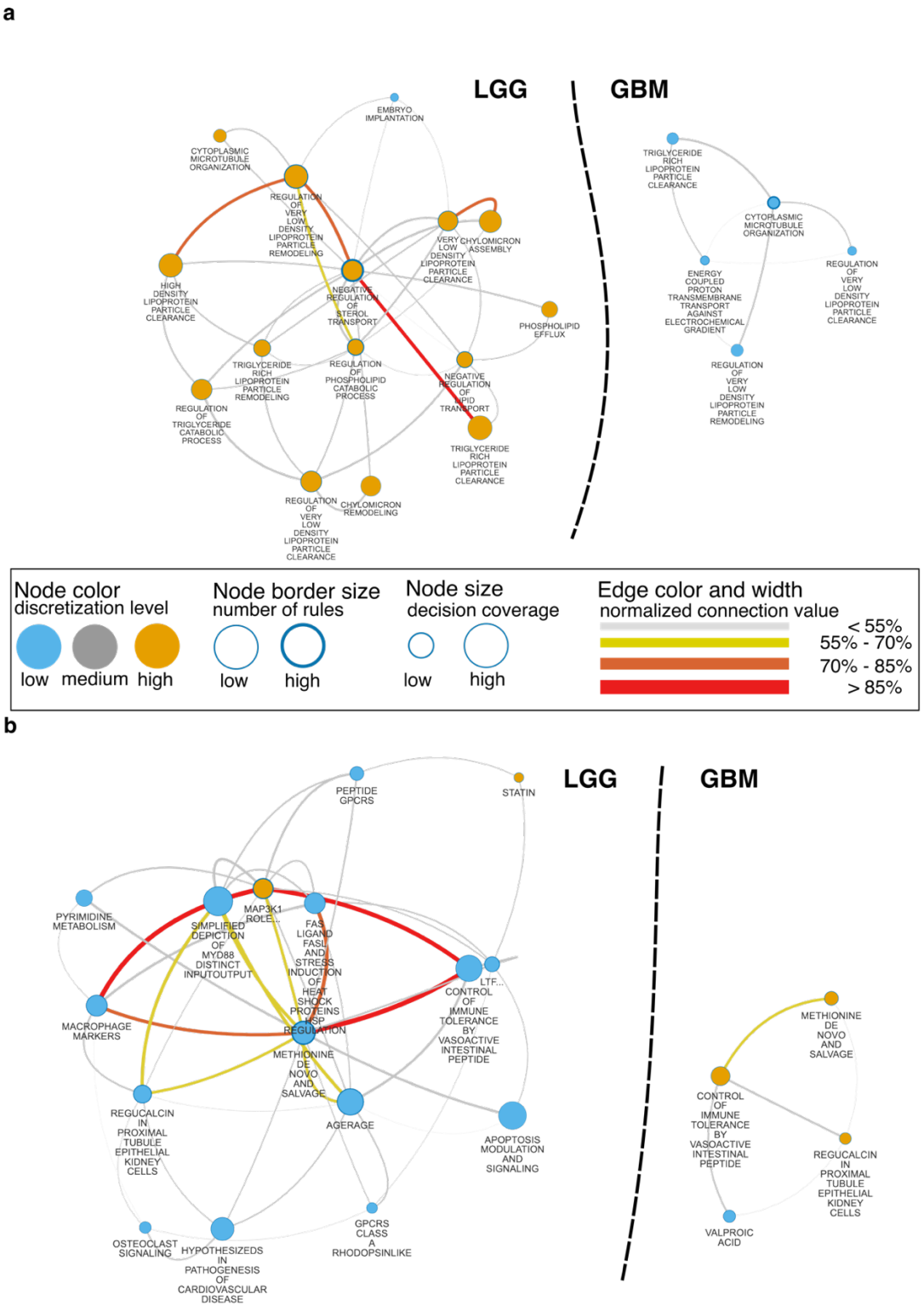


Fig. S18 Rule-based network displaying the most relevant co-enrichments of annotations obtained from **a.** GOBP and **b.** WP collections for LGG vs GBM models. The networks show 20 most connected nodes obtained from top 10% based on the rule connection from a set of significant rules (FDR-adjusted *P* value < 0.01). Connection values of nodes and edges represent a strength of co-enrichment from the classifier. Black dashed lines divide decision-related subnetworks. The following abbreviations were used for MAP3K1 ROLE… - MAP3K1 ROLE IN PROMOTING AND BLOCKING GONADAL DETERMINATION and LTF… - LTF DANGER SIGNAL RESPONSE