Supplementary Appendix

Supplement to: The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 2015;372:2481-98. DOI: 10.1056/NEJMoa1402121

SUPPLEMENTARY INFORMATION

for

Comprehensive, Integrative Genomic Analysis of Diffuse Lower Grade Gliomas

by

The Cancer Genome Atlas Research Network*

Address correspondence to:
Daniel J. Brat, MD, PhD
Department of Pathology and Laboratory Medicine
Emory University Hospital, G-167
1364 Clifton Rd. NE
Atlanta, GA 30322
e-mail: dbrat@emory.edu

Author Contributions

The Cancer Genome Atlas Research Network

Analysis and Disease Working Groups: Daniel J Brat[1], Roel GW Verhaak[2], Sofie R. Salama[3], Lee Cooper[1], Kenneth D Aldape[2], W.K Alfred Yung[2], Rehan Akbani[2], Jill S. Barnholtz-Sloan[11], Mitchel S. Berger[12], Cameron Brennan[9], Christopher A. Bristow[2] ,Andrew D. Cherniack[4], Giovanni Ciriello[9], Rivka Colen[2], Howard Colman[13], Adam E. Flanders[14], Caterina Giannini[15], Mia Grifford[3], Jason T. Huse[9], Antonio Iavarone[16], Rajan Jain[17], Isaac Joseph[18], Jaegil Kim[4], Katayoon Kasaian[6], Peter W. Laird[8], Tom Mikkelsen[10], C. Ryan Miller[5], Olena Morozova[3], Bradley A. Murray[4], Houtan Noushmehr[7], Brian Patrick O'Neill[15], Lior Pachter[18], Donald W. Parsons[19], Laila M. Poisson[10], Esther Rheinbay[4], A. Gordon Robertson[6], Carrie Sougnez[4], Erik P. Sulman[2], Scott R. Vandenberg[20], Erwin G. Van Meir[1], Mark Vitucci[5], Andreas von Deimling[21], Kosuke Yoshihara[2], Hailei Zhang[4] ,Siyuan Zheng[2]

International Genomics Consortium Biospecimen Core Resource: Daniel Crain[22], Kevin Lau[22], David Mallery[22], Scott Morris[22], Joseph Paulauskis[22], Robert Penny[22], Troy Shelton[22], Mark Sherman[22], Peggy Yena[22]

Nationwide Children's Hospital Biospecimen Core Resource: Aaron Black[23], Jay Bowen[23], Katie Dicostanzo[23], Julie Gastier-Foster[23], Kristen M. Leraas[23], Tara M. Lichtenberg[23], Christopher R. Pierson[23], Nilsa C. Ramirez[23], Cynthia Taylor[23], Stephanie Weaver[23], Lisa Wise[23], Erik Zmuda[23]

TCGA Project Team: Tanja Davidsen[24], John A. Demchok[24], Greg Eley[24], Martin L. Ferguson[24], Carolyn M. Hutter[25], Kenna R. Mills Shaw[2], Bradley A. Ozenberger[25], Margi Sheth[25], Heidi J. Sofia[25], Roy Tarnuzzer[24], Zhining Wang[24], Liming Yang[24], Jean Claude Zenklusen[24]

Data Coordinating Center: Brenda Ayala[26], Julien Baboud[26], Sudha Chudamani[27], Mark A. Jensen[26], Jia Liu[27], Todd Pihl[26], Rohini Raman[26], Yunhu Wan[26], Ye Wu[27]

Genome characterization/sequencing centers - Adrian Ally[6], J.Todd Auman[5], Miruna Balasundaram[6], Saianand Balu[5], Stephen B. Baylin[28], Rameen Beroukhim[4], Moiz S. Bootwalla[8], Reanne Bowlby[6], Christopher A. Bristow[2], Denise Brooks[6], Yaron Butterfield[6], Rebecca Carlsen[6], Scott Carter[4], Andrew D. Cherniack[4], Lynda Chin[2], Andy Chu[6], Eric Chuah[6], Kristian Cibulskis[4], Amanda Clarke[6], Simon G. Coetzee[7], Noreen Dhalla[6], Tim Fennell[4], Sheila Fisher[4], Stacey Gabriel[4], Gad Getz[4], Richard Gibbs[19], Jonna L. Grimsby[4], Ranabir Guin[6], Angela Hadjipanayis[29], D. Neil Hayes[5], Toshinori Hinoue[8], Katherine Hoadley[5], Robert A. Holt[6], Alan P. Hoyle[5], Franklin Huang[4], Stuart R. Jefferys[5], Steven Jones[6], Corbin D. Jones[5], Katayoon Kasaian[6], Raju Kucherlapati[29], Phillip H. Lai[8], Peter W. Laird[8], Eric Lander[4], Semin Lee[29], Lee Lichtenstein[4], Yussanne Ma[6], Dennis T. Maglinte[8], Harshad S. Mahadeshwar[2], Marco A. Marra[6], Michael Mayo[6], Shaowu Meng[5], Matthew L. Meyerson[4], Piotr A. Mieczkowski[5], Richard A. Moore[6], Lisle E. Mose[5], Andrew J. Mungall[6], Bradley A. Murray4, Houtan Noushmehr[7], Angeliki Pantazi[29], Michael Parfenov[29], Peter J. Park[29], Joel S. Parker[5], Charles M. Perou[5], Alexei Protopopov[2], Xiaojia Ren[29], Esther Rheinbay[4], Jeffrey Roach[5], A. Gordon Robertson[6], Mara Rosenberg[4], Thaís S. Sabedot[7], Sara Sadeghi[6], Jacqueline Schein[6], Steven E. Schumacher[4], Jonathan G. Seidman[29], Sahil Seth[2], Hui Shen[8], Janae V. Simons[5], Payal Sipahimalani[6], Matthew G. Soloway[5], Xingzhi Song[2], Carrie Sougnez[4], Chip Stewart[4], Huandong Sun[2], Barbara Tabak[4], Angela Tam[6], Donghui Tan[5], Jiabin Tang[2], Nina Thiessen[6], Timothy Triche, Jr.[8], David J. Van Den Berg[8], Umadevi Veluvolu[5], Scot Waring[5], Daniel J. Weisenberger[8], Matthew D. Wilkerson[5], Tina Wong[6], Junyuan Wu[5], Liu Xi[19], Andrew W. Xu[29], Lixing Yang[29], Travis I. Zack[4], Jianhua Zhang[2]

Genome Data Analysis Centers: B. Arman Aksoy[9], Harindra Arachchi[4], Chris Benz[3], Brady Bernard[30], Daniel Carlin[3], Lynda Chin[2], Juok Cho[4], Daniel DiCara[4], Scott Frazer[4], Gregory N. Fuller[2], JianJiong Gao[9], Nils Gehlenborg[4], Gad Getz[4], David Haussler[3], David I. Heiman[4], Lisa Iype[30], Anders Jacobsen[9], Zhenlin Ju[2], Sol Katzman[3], Jaegil Kim[4], Hoon Kim[2], Theo Knijnenburg[30], Richard Bailey Kreisberg[30], Michael S. Lawrence[4], William Lee[9], Kalle Leinonen[30], Pei Lin[4], Shiyun Ling[2], Wenbin Liu[2], Yingchun Liu[4], Yuexin Liu[2], Yiling Lu[2], Gordon Mills[2], Sam Ng[3], Michael S. Noble[4], Evan Paull[3], Amie J. Radenbaugh[3], Arvind Rao[2], Sheila Reynolds[30], Gordon Saksena[4], Zack Sanborn[31], Chris Sander[9], Nikolaus Schultz[9], Yasin Senbabaoglu[9], Ronglai Shen[9], Ilya Shmulevich[30], Rileen Sinha[9], Josh Stuart[3], S. Onur Sumer[9], Yichao Sun[9], Natalie Tasman[30], Barry S. Taylor[12], Roel GW Verhaak[2], Doug Voet[4], Nils Weinhold[9], John N. Weinstein[2], Da Yang[2], Kosuke Yoshihara[2], Hailei Zhang[4], Wei Zhang[2], Siyuan Zheng[2], Lihua Zou[4]

Neuropathology review: Ty Abel[32], Daniel J. Brat[1], Mark L. Cohen[11], Jenny Eschbacher[33], Caterina Giannini[15], Eyas M. Hattab[34], Christopher R. Pierson[23], Aditya Raghunathan[10], Matthew J. Schniederjan[35]

Tissue Source Sites: Dina Aziz[36], Gene Barnett[37], Jill S. Barnholtz-Sloan[11], Wendi Barrett[11], Darell D. Bigner[38], Lori Boice[5], Cathy Brewer[37], Chiara Calatozzolo[39], Carlos Gilberto Carlotti Jr[7], Timothy A. Chan[9], Lucia Cuppini[39], Erin Curley[22], Stefania Cuzzubbo[39], Karen Devine[11], Francesco DiMeco[39], Rebecca Duell[36], J Bradley Elder[36], Ashley Fehrenbach[40], Gaetano Finocchiaro[39], William Friedman[41], Jordonna Fulop[11], Johanna Gardner[22], Beth Hermes[33], Ady Kendler[40], Norman L. Lehman[36], Eric Lipp[38], Ouida Liu[42], Randy Mandt[36], Mary McGraw[37], Roger McLendon[38], Christopher McPherson[40], Luciano Neder[7], Phuong Nguyen[36], Ardene Noss[33], Raffaele Nunziata[39], Brian Patrick O'Neill[15], Quinn T. Ostrom[11], Cheryl Palmer[13], Alessandro Perin[39], Bianca Pollo[39], Alexander Potapov[43], Olga Potapova[42], W. Kimryn Rathmell[5], Daniil Rotin[43], Lisa Scarpace[10], Cathy Schilero[37], Kelly Senecal[36], Kristen Shimmel[11], Vsevolod Shurkhay[43], Suzanne Sifri[40], Rosy Singh[33], Andrew E. Sloan[11], Kathy Smolenski[37], Susan M. Staugaitis[37], Ruth Steele[40], Leigh Thorne[5], Daniela P.C. Tirapelli[7], Mahitha Vallurupalli[9], Yun Wang[10], Ronald Warnick[40], Felicia Williams[37], Yingli Wolinsky[11], Jianan Zhang[44], Sue Bell[36], Benito Campos[45], Christel Herold-Mende[45], Christine Jungk[45], Andreas Unterberg[45].

AUTHOR AFFILIATIONS:

1. Emory University School of Medicine, Winship Cancer Institute, Atlanta GA 30322
2. University of Texas MD Anderson Cancer Center, Houston, TX 77030
3. University of California Santa Cruz, Santa Cruz, CA 95064
4. The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142
5. University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
6. Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC CAN V5Z 1L3
7. Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, BR 14049-900
8. University of Southern California Epigenome Center, Los Angeles, CA 90089
9. Memorial Sloan-Kettering Cancer Center, New York, NY 10065
10. Hermelin Brain Tumor Center, Henry Ford Health Systems, Detroit, MI 48202
11. Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH 44106
12. University of California, San Francisco, San Francisco, CA 94143
13. Huntsman Cancer Center, University of Utah, Salt Lake City, UT 84112

14. Thomas Jefferson University Hospital, Philadelphia, PA 19107
15. Mayo Clinic, Rochester, MN 55905
16. Columbia University Medical Center, New York, NY 10032
17. New York University School of Medicine, New York, NY 10003
18. University of California Berkeley, Berkeley, CA 94720
19. Baylor College of Medicine Houston, TX 77030
20. University of California, San Diego, San Diego, CA 92103
21. Department of Neuropathology, University Hospital Heidelberg, German Cancer Research Center (DKFZ) and DKTK, Heidelberg, DE 69120
22. International Genomics Consortium (IGC), Phoenix, AZ 85004
23. The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205
24. National Cancer Institute, National Institutes of Health, Bethesda, MD 20892
25. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892
26. SRA International, Fairfax VA 22033
27. Leidos Biomedical Research Inc, Frederick MD 21702
28. Johns Hopkins University School of Medicine, Baltimore, MD 21287
29. Harvard Medical School, Boston, MA 02115
30. Institute for Systems Biology, Seattle, WA 98109
31. Five3 Genomics, LLC, Santa Cruz, CA 95060
32. Vanderbilt University School of Medicine, Nashville, TN 37232
33. Barrow Neurological Institute, St. Joseph's Hospital and Medical Center, Phoenix, AZ 85013
34. Indiana University School of Medicine, Indianapolis, IN 46202
35. Children’s Healthcare of Atlanta, Emory University, Atlanta, GA 30342
36. The Ohio State University Medical Center, Columbus OH 43210
37. Cleveland Clinic Foundation, Cleveland, OH 44195
38. The Preston Robert Tisch Brain Tumor Center, Duke University Medical Center Durham, NC 27710
39. Fondazione IRCCS Istituto Neurologico Carlo Besta, Milano, Italy 20133
40. University of Cincinnati College of Medicine and Health University Hospital, Cincinnati, OH 45267
41. University of Florida, Gainesville, FL 32611
42. Cureline, Inc., South San Francisco, CA 94080
43. NN Burdenko Scientific Research Institute of Neurosurgery, Moscow, Russia 125047
44. Fred Hutchinson Cancer Research Center, Seattle, WA 98109
45. Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, DE 69120

Contents

1 Biospecimens

To the section on biospecimens was contributed by: Troy Shelton, Jay Bowen, David Mallery, Tara Lichtenberg, Kristen Leraas, Scott Morris, Christopher R. Pierson, Joseph Paulauskis, Erin Curley, Mark Sherman, Kevin Lau, Robert Penny.

Correspondence and questions should be directed to: Troy Shelton (tshelton@intgen.org)

To the section on clinical data was contributed by: Laila Poisson, Jill Barnholtz-Sloan, Jay Bowen, Johanna Gardner, Caterina Giannini, Haley Gittleman, Mia Griffod, Tara Lichtenberg, Brian P. O'Neill, Sofie Salama, Lisa Scarpace, Candace Shelton, Yun Wang, Peggy Yena.

Correspondence and questions should be directed to: Laila Poisson (laila.poisson@hfhs.org)

a) Biospecimen acquisition and quality control

Biospecimens were collected from patients diagnosed with diffuse glioma (grade II or III) undergoing surgical resection and had received no prior treatment for their disease (chemotherapy or radiotherapy). The pathologic diagnosis of astrocytoma, oligodendroglioma, oligoastrocytoma of grade II or III was established by the neuropathologist at the tissue source site. Institutional review boards at each tissue source site reviewed protocols and consent documentation and approved submission of cases to TCGA. Each frozen primary tumor specimen had a companion blood/blood components specimen (including DNA extracted at the tissue source site). Tumor samples were required to be 200 mg. Recurrent tumor specimens for 9 of the cases in the final data freeze list were submitted for molecular characterization. Two of these cases had two separate recurrences analyzed. Specimens were shipped overnight from 27 tissue source site using a cryoport that maintained an average temperature of less than -180°C. Each tumor was embedded in optimal cutting temperature (OCT) medium and a histologic section was obtained for review. Each H&E stained case was reviewed by a board-certified TCGA neuropathologist for quality assurance and to confirm that the tumor specimen was histologically consistent with diffuse glioma (astrocytoma, oligoastrocytoma or oligodendroglioma of WHO grade II or III). The tumor sections were required to contain an average of 60% tumor cell nuclei with less than 50% necrosis for inclusion in the study per TCGA protocol requirements.

The case list freeze included 293 cases in batches 78, 112, 146, 163, 189, 219, 245, 282, 292, 295, and 306. Samples were from the following 13 tissue source sites: Thomas Jefferson

University (CS); Mayo Clinic (DB); University of Florida (DH); Henry Ford Hospital (DU); Duke University (E1); University of North Carolina (EZ); Case Western (FG); International Genomics Consortium (FN); St. Joseph AZ (HT); Memorial Sloan Kettering Cancer Center (HW); Christiana Care Health Services, Inc. (IK); Cureline, Inc. (P5); Fondazione-Besta (QH).

b) Sample processing

RNA and DNA were extracted from tumor specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). Flow-through from the Qiagen DNA column was processed using a mirVana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen).

Each specimen was quantified by measuring Abs260 with a UV spectrophotometer or by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. Only specimens yielding a minimum of 6.9 µg of tumor DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN ≥ 7.0 were included in this study. Overall, the BCR received 827 lower-grade glioma cases of which 572 (69%) passed quality control.

c) Clinical files and analysis

LGG clinical files were downloaded from the TCGA Data portal (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm, August 25, 2014). All available XML files for records through batch 306 were included in the download according the specifications of the data freeze (Batches 78, 112, 146, 163, 189, 219, 245, 282, 292, 295, 306). The case-level XML files were converted to rectangular files, cleaned, and integrated using SAS v9.2. Supplemental radiotherapy and pharmacotherapy receipt biotab files were queried to update missing information regarding primary therapy receipt, i.e., therapy received as part of the initial treatment of disease. We classify primary therapy as treatment having started within 90 days of diagnosis and prior to the first reported tumor event. Cases with supplemental therapy data reported, but without treatments meeting the primary therapy criteria, are listed as not receiving primary therapy. The working data file was cross-referenced against the list of samples contained in the data freeze. This file with variable key can be found in the supplemental materials (Table S1). For cases without available clinical records, the histology type was obtained from the biospecimen XML files.

Overall survival is defined as the years from initial diagnosis until death. Cases still alive at the time of this study have overall survival time censored at the time of last follow-up. Progression-free survival is defined as the years from initial diagnosis until either the time of the first new tumor event (progression or recurrence) or death. Persons who were alive and without progression or recurrence at the time of the study had progression-free survival time censored at the time of last follow-up. For analysis including GBM, data were obtained from the recent TCGA publication (Brennan et al., 2013).Only primary GBM samples were included.

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. Area under the receiver operator characteristic curve (AUC) is estimated for prediction of the survival outcomes accounting for observations censored due to loss to follow-up. Prediction error curves provide estimates of error across time based on the Brier score, which compares the predicted probability of the event to the observed event status. The marginal model was used for censoring probability and bootstrap resampling (0.632 weight) was used to enhance generalizability of the error estimates. A Brier score of 0 is perfect prediction and 0.33 is equivalent to random prediction. The reference curve is based on a model with no covariates. These analyses were conducted in R (v 3.02) using survival and risksetROC (Heagerty and Zheng, 2005) and pec (Mogensen et al., 2012) packages.

Clinical Data: Summary of Results. There are 289 cases (98.6%) with clinical data in the data freeze (Table S1). Three cases with clinical data were excluded from the freeze because they were found to be recurrent (TCGA-CS-6670, TCGA-DU-7011) or not in accordance with the TCGA protocol (TCGA-DB-5270). Survival by grade and histology was consistent with prior studies (Figure S22A-E). IDH-1p19q codel status is known for 282 cases. *IDH*wt tumors have no mutation in either *IDH1* or *IDH2*. Tumors with a mutation in either *IDH1* or *IDH2* are further subdivided by whether there is co-deletion of 1p19q (*IDH*mut-codel) or not (*IDH*mut-non-codel).; see Table S2A.i. Histological type is associated with *IDH*-1p19q codel status (FET p<0.0001) with 82% of *IDH*mut-codel cases classified as oligodendroglioma and 56% of *IDH*wt cases classified as astrocytoma. The *IDH*wt group was predominantly WHO Grade III at diagnosis (76%) compared to the *IDH*mut classes (45%, FET p<0.0001). Age differs between *IDH*-codel groups (ANOVA p<0.0001) with *IDH*wt cases tending to be oldest at diagnosis (mean 49.9 years) and *IDH*mut-non-codel cases tending to be youngest (mean 38.1 years), while *IDH*mut-codel cases have intermediate age (mean 45.4 years). *IDH*wt tumors are more likely to report family history of cancer (62%) compared to *IDH*mut tumors (40%, FET p=0.0228). Though this group tends to be older, *IDH*wt tumors tend to be more likely to have a family history of cancer after adjusting for age (adjusted odds ratio 2.05, p=0.0770). There is no evidence of association with familial history of primary brain tumor specifically. *IDH*mut tumors tend to present in the frontal lobe (68%) compared to *IDH*wt tumors (36%, FET p<0.0001). *IDH*wt present more frequently in the temporal lobe (45% versus 22% *IDH*mut, FET p=0.0010). There was no evidence of differences in first presenting symptom, extent of resection, or the use of pre-operative medication by *IDH*-1p19q codel status. Though *IDH* status is not necessarily known at the time of diagnosis, the receipt of radiotherapy differs between the groups with *IDH*wt tumors most likely to receive radiotherapy as part of primary therapy (88%), and *IDH*mut-codel least likely (56%, p=0.0129). There is no evidence of a difference in pharmacotherapy receipt by *IDH*-codel group (FET p=0.6390).

Histological type was defined as astrocytoma, oligoastrocytoma, and oligodendroglioma and WHO grade was listed as grade II or grade III (Table S2A.ii). Fisher's exact test (FET) was used to compare categorical variable distributions between the six type-grade groups. A simulated FET p-value was obtained from 10^7 cycles for large tables. Sub analyses collapsing across clinical groups or across tumor type and grade were also considered. Two-way Analysis of Variance (ANOVA) was used to compare continuous variables between type and grade. Few of these features are statistically associated with histological type and WHO grade. The few non-US cases clustering in the WHO grade II, particularly astrocytoma, result in a marginally significant association for origin of case (FET, p=0.0688). For first presenting symptom, there is evidence of an association by tumor group for those with seizure versus all others (FET p=0.0086), but the association is not specifically due to histology (FET p=0.2088) or grade (FET p=0.1057). Overall, 54% of LGGs present first with seizure, whereas 25% present first with headache. Performing a two-way ANOVA on age at diagnosis, we found that grade is the significant predictor (p<0.0001), with grade III presenting in persons 6.7 years older than grade II, on average. Consistent with current clinical practice, grade III glioma cases are more likely to report receipt of radiotherapy (88%) and/or pharmacotherapy (76%) compared to grade II cases (51%, FET p<0.0001 and 42%, FET p<0.0001, respectively).

Survival Modeling: Summary of Results. There are 289 observations included in this study. Among these, 60 cases were deceased at the time of analysis with median overall survival (OS) estimated at 6.51 years (95% CI 5.24, 9.50). The median follow-up time is estimated at 1.45 years (95% CI 1.28, 1.60), with 13% of cases with follow-up of at least 5 years. Additionally, 77 cases experienced at least one progression or recurrence resulting in 88 events for the progression-free survival models (PFS, n=250). The median time to progression is estimated at 3.30 years (95% CI 2.97, 4.46). Extent of resection is unknown for 10 cases. Models including extent of resection are based on 279 observations with 59 deaths for OS and 241 observations and 85 events for PFS. *IDH*-codel status was unknown for 11 cases. Models including IDH-codel status are based on 278 observations with 58 deaths for OS and 241 observations and 86 events for PFS. Models including both extent of resection and *IDH*-codel status are based on 268 observations with 57 deaths and 85 events.

Kaplan-Meier plots of estimated survival, overall and progression free, are presented in Figures S22. Age at diagnosis differs significantly between the three LGG molecular groups and two GBM *IDH* groups used in the comparison of outcomes (one-way ANOVA p<0.0001); see Figure 6B. In post-hoc t-tests the GBMwt group is significantly older than the other groups (mean±SD: 61.0±12.7 years, p<0.001). Also the *IDH*wt group is older (49.9±15.3 years) than the *IDH*mut-non-codel (38.1±10.9 years) (p<0.001) and nearly so for GBMmut (42.0±13.6 years, p=0.083). The *IDH*mut-codel is significantly older (45.4±13.2 years) than the *IDH*mut-non-codel (p<0.001). Given that age at diagnosis is associated with overall survival and time to progression, we examined Cox regression models accounting for age and extent of resection. Table S2B presents estimates from single predictor models as well as multiple-predictor models (Models I – III) of overall survival. Similarly, Table S2C presents the estimates from single predictor models as well as multiple-predictor models (Models I – III) of progression-free survival. Table S2D presents estimates for single predictor an age adjusted models of overall survival associated with 5A.

In the analysis of overall survival, the risk of death is significantly increased with increasing age (HR 1.38 per 5 years) and marginally with a less than gross total resection (HR 1.71) (Table S2B). A histological diagnosis of astrocytoma has a significantly increased risk (HR 2.26) compared to oligodendroglioma, but the mixed oligoastrocytoma is not a significantly increased risk (HR 1.13). WHO grade III is associated with increased risk of death (HR 3.36) relative to WHO grade II. Wild-type *IDH* is associated with increased risk (HR 9.22) relative to *IDH*mut-codel tumors, but *IDH*mut-non-codel tumors do not show significantly increased risk (HR 1.32). Prediction error curves for overall survival show an advantage for the *IDH*/Codel grouping relative to the Histology grouping (Figure S22i). The error curves are truncated at 5 years because we have less than 10% of the samples followed beyond that time. These trends are consistent after adjusting for age at diagnosis.

In multivariable Model I, adjusting for age balances the effect of histological type with astrocytoma (HR 1.80) and oligoastrocytoma (HR 1.88) having increased risk relative to the oligodendroglioma type, though the risk is not significantly increased (Table S2B). The effect of a WHO grade III classification relative to grade II is retained (HR 3.60). In Model II, the *IDH*mut-non-codel group is estimated to have increased risk compared to the *IDH*mut-codel group (HR 1.80), though it is not statistically significant. The *IDH*wt group has substantially increased risk (HR 11.22). In Model III, increasing age (HR 1.37 per 5 years), a grade III diagnosis (HR 2.79), and an IDHwt molecular class (HR 6.67) remain statistically significant predictors of increased risk of death. The AUC for prediction of death by one year is improved from Model I to Model II, and the multi-predictor models are better than any of the single-predictor models. The AUC is slightly improved further with Model III (0.87).

In the analysis of progression-free survival, the risk of an event is significantly increased with increasing age (HR 1.14 per 5 years) and marginally so with a less than gross total resection (HR 1.25; Table S2C). A histological diagnosis of astrocytoma has an increased risk (HF 1.60) compared to oligodendroglioma. WHO grade III is associated with increased risk (HR 1.58) relative to WHO grade II. Wild-type *IDH* is associated with increased risk (HR 8.89) relative to *IDH*mut-codel tumors, but *IDH*mut-non-codel tumors do not show significantly increased risk (HR 1.48). Prediction error curves for progression-free survival show a consistent advantage for the IDH/codel grouping relative to the Histology grouping, which tends to follow the reference curve (Figure S22j). This trend is consistent after adjusting for age at diagnosis.

In multivariable Model I, adjusting for age impacts the effect of histological type, with astrocytoma (HR 1.77) having increase risk relative to the oligodendroglioma (Table S2C). Oligoastrocytoma has some evidence of an increased risk (HR 1.70) compared to oligodendroglioma, but not statistically significant. The effect of a WHO grade III relative to grade II is retained (HR 1.69). In Model II, the *IDH*mut-non-codel group is estimated to have increased risk compared to the *IDH*mut-codel group (HR 2.02) and the *IDH*wt group has substantially increased risk (HR 9.17). In Model III, increasing age (HR 1.10 per 5 years) and *IDH*wt molecular class (HR 8.30) retain significance as increasing the risk of an event. The AUC is improved between models I and II but nor further by model III.

For models including GBM cases, there were 675 samples between the two datasets with complete information on age at diagnosis, overall survival, and *IDH* mutation. Of these 397 are GBM with 6% (n=24) having a mutation in the *IDH1* gene. There were 345 deaths observed during follow-up for 49% censoring. Data on extent of resection are not available, so adjusted models only include age.

Table S2D provides estimates of the hazard ratios from a Cox regression model using the five classes (*IDH*mut-codel, *IDH*mut-non-codel, *IDH*wt, GBMmut, GBMwt) as predictors alone (left column) and adjusting for age at diagnosis (right column). The *IDH*mut-codel and *IDH*mut-non-codel do not have significantly different hazard rates, nor do the *IDH*wt and GBMmut groups. Otherwise all other pairs of groups differ in overall survival risk. The same trends hold when adjusting for age, though the risk is muted in some, especially for comparisons involving GBMwt cases, which have the oldest age at diagnosis. In particular, the increased risk for GBMwt tumors compared to *IDH*wt tumors becomes non-significant after accounting for age.

Comparison of Clinical Subtypes to Molecular Clustering Solutions

Clustering solutions for LGG were defined for each molecular platform used in this project. Summary solutions such as the “cluster-of-clusters” (CoC) and Oncosign clustering were also constructed (see Figure 1, 3). Here we quantify how well specific clustering solutions relate to the *IDH*/codel status. We chose to use the adjusted Rand index (ARI) which quantifies how often a pair of samples is segregated into the same cluster by the *IDH*/codel classification and the clustering solution. The ARI adjusts for the likelihood of random concordance. If two classification methods are not associated beyond random chance, the ARI has an expected value of 0. If two classification methods result in identical segregation, the ARI is 1. Negative values are possible if classification methods are more discordant than expected by chance. It is not required that the classification methods compared have the same number of classes.

Methods. Molecular clustering solutions, *IDH*/codel classification, and histology/grade assignment can be found in Table S1. The estimates were made using the adjusted Rand Index (mclust) function in R. Only the 209 samples with clustering groups defined in each of the 7 molecular classes, *IDH*/codel classes and with histology/grade information are considered so that the measures are more directly comparable. Bootstrap sampling was used to construct 95% confidence intervals for the ARI estimates (m=1000 samples).

Summary of Results. In Table S2E, ARI values are provided for each classification and clustering scheme against the *IDH/*codel classification. ARI values are also given for similarities with the three histology groups and the six histology and WHO grade groups. The summary solutions resulting from the “cluster-of-clusters” (CoC) and Oncosign show relatively strong concordance with ARI of 0.79 and 0.83, respectively. Moderate to strong concordance is also shown with some of the single-molecular type classifications. Notably, there is little similarity (scores near zero) for the histology/grade classes with respect to the molecular clustering solutions.

Prognostic Ability of IDH/Codel Subtypes and Molecular Clustering Solutions

There are 257 samples with clinical data (including extent of resection), *IDH*/Codel status, CoC assignment, Oncosign assignment, and survival follow-up (time>0). Of these 57 experienced a death during their follow-up.

A base model of overall survival using age at diagnosis and extent of resection as predictors shows that age is a significant predictor of OS (HR 1.41, p<0.0001) but extent (<GTR vs GTR) is not (HR: 1.19, p=0.56). For this model, the AUC for prediction of survival is 0.74 at one year. Since GTR is a standard predictor in overall survival it will be retained in the following models.

The addition of *IDH*/codel status to the base model is significant (likelihood ratio test (LRT), p<0.0001). *IDH*wt increased risk of death (HR 11.4, 95%CI: 5.0, 26.3) relative to *IDH*mut-codel. There is not significant evidence of a difference between the *IDH*mut-codel and *IDH*mut-non-codel groups (HR 1.8, 95% CI: 0.9, 3.7). The AUC for prediction of survival is 0.85 at one year.

Similarly, adding CoC group to the base model is significant (LRT, p<0.0001). CoC2 increased risk of death relative to CoC3 (HR 9.2, 95%CI: 4.2, 20.0). This is as expected since 82% of the tumors in CoC2 are *IDH*wt and 94% of the COC3 tumors are *IDH*mut-codel. Similarly, since CoC1 is primarily composed of *IDH*mut-non-codel tumors (96%), there is not significant evidence of a difference between CoC1 and CoC3 (HR 1.7, 95% CI: 0.8, 3.6). The AUC for prediction of survival is 0.84 at one year.

Adding Oncosign group to the base model is also significant (LRT, p<0.0001). OSC4 has increased risk of death relative to OSC1 (HR 6.9, 95%CI: 3.0, 15.8). Again, this is as expected since 100% of the tumors in OSC4 are *IDH*wt and 97% of the OSC1 tumors are *IDH*mut-non-codel tumors. Similarly, since OSC2 and OSC3 are primarily composed of *IDH*mut-codel tumors (100%, 67%, respectively), there is not significant evidence of a difference in OSC2 (HR 0.5, 95% CI: 0.2, 1.1) or OSC3 (HR 0.6, 95% CI: 0.1, 2.6), relative to OSC1. The AUC for prediction of survival is 0.84 at one year.

The survival model using *IDH*/codel status captures the information found in either of the more complex molecular clusterings (Cluster of Clusters and OncoSign). Predictive accuracy is minimally improved when we can add WHO grade to our model with *IDH*/codel status (LRT p=0.0005). Though, grade is a significant predictor in this model, with grade III having increased risk of death (HR 3.1, 95%CI: 1.6, 6.1), the AUC for prediction of survival is only 0.86 at one year with this model.

In contrast, the addition of histology diagnosis to the base model is significant (LRT, p=0.0013) and WHO grade is additionally significant (LRT, p<0.0001). In the model with histology and grade, we see that astrocytoma has increased risk of death (HR 2.4, 95%CI: 1.2, 4.5) relative to oligodendroglioma. There is not significant evidence of a difference between the oligoastrocytoma and oligodendroglioma diagnoses (HR 1.9, 95% CI: 0.9, 4.1). There is a significant increase of risk for grade III tumors relative to grade II (HR 3.8, 95% CI: 1.9, 7.4). The AUC for prediction of survival is 0.81 at one year. If grade were excluded, the AUC for prediction of survival based on histology is 0.76 at one year.

2 Whole genome and exome sequencing

To this section on mutation calling was contributed by: Esther Rheinbay, David Haussler, Katayoon Kasaian, Jaegil Kim, Amie Radenbaugh, Mara Rosenberg, Sofie Salama, Carrie Sougnez, Chip Stewart, Hailei Zhang.

Correspondence and questions should be directed to: Esther Rheinbay (esther@broadinstitute.org)

To this section on structural variants was contributed by: Mia Grifford, Zack Sanborn, Siyuan Zheng, Roel G.W. Verhaak, Sofie Salama and David Haussler

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

a) Whole exome and whole genome sequencing data production

Whole exome sequencing of 0.5-3 micrograms of DNA from tumor and normal blood samples was performed as previously described (Cancer Genome Atlas Research, 2012). Exome capture was performed using the Agilent Sure-Select Human All Exon v2.0, 44Mb kit, followed by 2 x 76 bp paired-end sequencing on the Illumina HiSeq platform. For whole genome sequencing, 2 x 101 bp reads were sequenced on the same platform. Read alignment and processing were performed using BWA and the Picard and Firehose pipelines at the Broad Institute as previously described (Institute, 2014).

b) Identification of somatic mutations (Mutect/Broad Institute)

Alignments were first subjected to quality control to avoid mix-ups between tumor and normal samples, as well as cross-contamination between tumor samples using ContEst (Cibulskis et al., 2011). Only samples with less than 4% of estimated cross-contamination were further analyzed. We used MuTect algorithm version 1.1.6 (Cibulskis et al., 2013) to generate somatic mutation calls, which were subsequently filtered to remove any spurious calls due to shearing-induced generation of 8-oxoguanine (Costello et al., 2013). Indels were identified using the indel locator algorithm as previously described (Chapman et al., 2011). Details and tools are available at www.broadinstitute.org/cancer/cga.

c) Identification of somatic mutations (RADIA/University of California Santa Cruz)

Single nucleotide somatic mutations were identified by RADIA (RNA AND DNA Integrated Analysis), a method that combines the patient matched normal and tumor DNA whole exome sequencing (DNA-WES) with the tumor RNA sequencing (RNA-Seq) for somatic mutation detection(Radenbaugh AJ et al., 2014). The inclusion of the RNA increases the power to detect somatic mutations, especially at low DNA allelic frequencies. By integrating the DNA and RNA, mutations that would be missed by traditional mutation calling algorithms that only examine the DNA can be rescued back. RADIA classifies somatic mutations into 3 categories depending on the read support from the DNA and RNA: 1) DNA mutations – mutations that had high support in the DNA, 2) RNA Confirmation calls – mutations that had high support in both the DNA and RNA, 3) RNA Rescue calls – mutations that had high support in the RNA and weak support in the DNA. Here our analysis pipeline identified 13,720 DNA mutations, 3,926 RNA Confirmation calls, and 1,015 RNA Rescue calls.

d) Identification of somatic mutations (Baylor College of Medicine)

Methods were used to identify mutations as previously described (Cancer Genome Atlas Research, 2013a).

e) Identification of somatic indels (Strelka/BC Cancer Agency)

Strelka (Saunders et al., 2012) (v1.0.6) was used to identify somatic single nucleotide variants, and short insertions and deletions from the TCGA LGG exome dataset. All parameters were set to defaults, with the exception of "isSkipDepthFilters", which was set to 1 in order to skip depth filtration given the higher coverage in exome datasets. 307 pairs of libraries were analyzed. When a blood sample was available, it served as the matched normal specimen; otherwise, the matched normal tissue was used. The variants were subsequently annotated using SnpEff (Cingolani et al., 2012), and the COSMIC (v61)(Forbes et al., 2010) and dbSNP (v137) (Forbes et al., 2010) databases.

f) Generation of exome "consensus" mutation set

Based on the mutation tables from the three centers, a "consensus" set was generated using custom scripts according to the following criteria: 1) For SSNV, a mutation was included if it was called by at least two out of the three calling centers. Identical chromosomal position as well as nucleotide change were required. 2) Indel calls were provided by two centers (Baylor College of Medicine and Broad Institute). An indel was included in the consensus mutation table if it was called by both centers, allowing a maximum distance of 10 nucleotides between sites. Functional annotation of mutations was performed with Oncotator (http://www.broadinstitute.org/cancer/cga/oncotator) using Gencode V18. Somatic mutation calls were filtered to remove common artifacts and germline variation.

g) Mutation annotation, validation and TERT promoter assay

Functional annotation of mutations was performed with Oncotator (http://www.broadinstitute.org/cancer/cga/oncotator) using Gencode V18.

Validation of somatic mutations. Targeted resequencing of selected mutations for validation was performed using a microfluidic device (Fluidigm). A total of 19 mutations in the genes identified by MutSig analysis (*ARID1A, ATRX, CIC, EGFR, FUBP1, IDH1, IDH2, NF1, NOTCH1, PIK3CA, PIK3R1, PLCG1, PTEN, PTPN11, SMARCA4, TCF12, TP53, ZBTB20, ZCCHC12*) as significantly mutated in LGG as well as additional genes with known roles in glioma (*CDKN2A, PDGFRA, BRAF, RB1*) were selected for targeted resequencing. PCR primers were designed around sites of interest with desired amplicon size of 200 bp. 20 – 50 ng of each DNA sample was mixed with oligonucleotides containing Illumina adapter sequences, a sample-specific molecular barcode and a sequence complementary to the primer tails. This mixture was used as the PCR template for each sample amplified on the Fluidigm access array. This method allowed the use of universal PCR primers while ensuring all amplicons for a given sample received the same sequencing barcode. PCR was performed on the Fluidigm access array according to manufacturers' instructions. Barcoded libraries were recovered for each sample in a single collection well on the Fluidigm access array, quantified using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA) and concentrations normalized across libraries. Libraries were loaded on the Illumina MiSEQ instrument and sequenced using paired end 150bp sequencing reads. Mutations were categorized according to base substitutions, with C>T transitions having the highest frequency (Figure S12).

In addition to targeted resequencing (available for 260 patients), mRNA-sequencing (available for 280 patients), and whole-genome sequencing data (available for 19 patients) were included as independent validation data sources.

Validation status of a given powered site was determined as previously described[8]. In brief, a site was considered "powered" if at least two alternate alleles were present in the tumor validation data, and the detection power was greater than or equal to 0.95. 96.3 percent of mutations in significantly mutated genes validated in mRNA-sequencing, 99.4% in targeted resequencing, and 100% in WGS data. For indels, validation rates were 81.2% for mRNA-seq, 96.4% for targeted resequencing and 100% for WGS. Because of detection issues of large indels in targeted resequencing, we manually reviewed these mutations to exclude false negative calls. Based on negative validation results, two *EGFR* mutations were removed from the final analysis set. One *IDH1* mutation was added manually to the mutation table as it was not called by any of the three centers but was observed in RNA and targeted sequencing data.

TERT promoter mutation assay. Targeted assay of the *TERT* promoter region was performed as previously described (The Cancer Genome Atlas Research Network (2014) Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell* (in press)). Paired-end sequencing with 150 bp read length was performed of PCR amplicons of length 273 bp to ultra-high depth (median coverage at site 1295228: 91430; range 242-234201) on a Illumina MiSeq instrument. Sites in the interval ch5:1295150-1295300 were considered for mutation calling, however, six sites were blacklisted due to particularly low average base quality. Samtools mpileup was used to extract base calls and original base qualities. We then applied a custom script to call mutations at non-blacklisted sites in the calling interval, considering only bases with quality >=25. Only sites with alternative allele fraction above 5% were classified as mutated.

h) Mutation significance analysis

Driver mutations were identified with the MutSig2CV algorithm (Lawrence et al., 2014) applied to the consensus mutation call set. Genes with q-values above 0.1 were considered significantly recurrently mutated (Supplemental Table 4A). After this discovery step, we examined mutation calls made by each individual center to exclude false negative decisions on driver genes. Single-center mutation calls were added to the final mutation table (Figure 2) if they had additional support in the validation experiment from at least one independent data type (RNA-seq, WGS or targeted sequencing). MutSig2CV analysis was also conducted separately on the three molecular subtypes, and results for these runs are presented in Supplemental Tables 4B-D.

i) Exome analysis to identify candidate double minute chromosomes

Samples likely to contain circular amplicons corresponding to double minute chromosomes (DM) or homologously staining regions (HSR) were identified as in Sanborn et al. (Sanborn et al., 2013). Briefly, tumor and matched normal exomes were processed by BamBam to compute relative coverage and identify somatic rearrangements. Tumors with multiple high copy number peaks (5-fold increased relative coverage versus their matched normal) were manually analyzed to discover any oncogenes within peaks, associate peaks with nearby somatic rearrangements, and determine if a sample exhibits multiple peaks with similar copy number levels. Samples were scored as having possible DM/HSR if they either contained multiple distinct high copy number peaks with similar copy number levels and at least one peak contained an oncogene, or a single distinct peak that spanned approximately 1 Mb, contained an oncogene and had an associated rearrangement. Results are summarized in Table S5 (tab “DM_HSR”).

For one potential DM/HSR sample, TCGA-CS-5395, whole genome sequencing data was available, allowing us to precisely reconstruct a circular amplicon according to published methods (Sanborn et al., 2013)(Figure S17).

j) DNA rearrangement analysis

We used BamBam (Sanborn et al., 2013) to identify genomic rearrangements and compute relative copy numbers between pairs of tumor and matched normal coordinate-sorted BAM files. Briefly, BamBam uses a dynamic windowing approach that aggregates read counts in both tumor and normal data sets within genomic windows defined such that each window contains a roughly equivalent number of reads from the tumor and/or normal data sets, thereby producing large windows in regions of low coverage to improve signal-to-noise ratio and small windows in highly amplified regions to increase the resolution of amplicon boundaries. This method identified both local and chromosome-level deletions and amplifications in these samples. Potential intra- and interchromosomal rearrangements were discovered using discordant paired reads, where each read in the discordant pair map to disparate regions of the reference sequence. The discordant paired-end reads from both tumor and normal data sets were clustered according to their genomic location to define an approximate genomic region of the putative breakpoint.

Where possible, the breakpoints discovered from paired-end clustering were refined using split reads anchored near the breakpoints. These split reads were discovered from unaligned reads anchored near the breakpoint by a mapped mate. Breakpoints were filtered using a minimum criterion of 4 discordant reads supporting the rearrangement and average mapping quality of 20 for the reads spanning the rearrangement break point. BamBam rearrangements reported in Figure 5B and Table S5 correspond to those found with whole genome sequencing, low pass whole genome sequencing or exome sequencing that had at least 10 discordant reads supporting the breakpoint, spanning read mapping quality of at least 30 and split read evidence directly overlapping the putative breakpoint's junction with a split read score of at least 50.

3 Low pass genome wide DNA sequencing

To this section was contributed by: Christopher A. Bristow, Lixing Yang, Xingzhi Song, Semin Lee, Sahil Seth, Jianhua Zhang, Lynda Chin, Peter J. Park, Raju Kucherlapati, Alexei Protopopov.

Correspondence and questions should be directed to: Christopher A. Bristow (cabristow@mdanderson.org) and Alexei Protopopov (aprotopopov@mdanderson.org)

a) Library construction and sequencing

Illumina pair-end libraries were generated from genomic DNA according to the manufacturer's protocol (Illumina Inc.) with modification. Briefly, 500 ng to 700 ng of genomic DNA was sheared into ~200 base-pair fragments using the Covaris plate with E220 system (Covaris, Inc. Woburn, MA). Libraries were generated using KAPA Bio kits with the Caliper robotic NGS Suite according to the manufactures' protocols (PerkinElmer). Libraries for 52 tumor-normal pairs were sequenced in paired-end mode, with the 51 bp read length, using the Illumina HiSeq 2000. Each lane was loaded with a single sample, with the tumor and normal typically run on the same flowcell. The average sequence coverage, read quality, and percentage of mapped reads was: 5.2x, 39.5, and 96.5%, respectively. Raw sequencing reads were converted to FASTQ formed data using CASAVA (Illumina) and then mapped to the human genome with the Burrows-Wheeler Aligner (Li and Durbin, 2009) to generate .bam files that are stored and accessible on CGHub.

b) Identification of copy number variants

To characterize somatic copy number alterations in the tumor genome, we applied BIC-seq (Xi et al., 2011), a previously developed algorithm. Briefly, we first counted uniquely aligned reads in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumor and matched normal genomes, BIC-seq attempts to iteratively combine neighboring bins with similar copy numbers. Whether the two neighboring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both the fit and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Consecutive segments with copy ratio difference smaller than 0.1 (log2 scale) were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts. BICseq DNA copy number profiles can be obtained from http://gdac.broadinstitute.org/runs/stddata__2014_04_16/data/LGG/20140416/gdac.broadinstitute.org_LGG.Merge_cna__illuminahiseq_dnaseqc__hms_harvard_edu__Level_3__segmentation__seg.Level_3.2014041600.0.0.tar.gz.

c) Translocation discovery with BreakDancer and MEERKAT

Structural Variation detection is performed with the program BreakDancer (Chen et al., 2009) on a .bam file constructed from HiSeq sequencing of each tumor pair. The first step requires a configuration file of each bam file for each tumor pair with the bam2cfg.pl perl module of the program. After the configuration file, the perl module BreakDancerMax.pl is run on the configuration file in order to call structural variants in the tumor and control files. Each tumor structural variant file is filtered with its matched normal to remove any false positives. Structural variations are also detected by Meerkat (Yang et al., 2013) which require at least two discordant read pairs supporting one event and at least one read covering the breakpoint junction. Each variant detected from tumor genome is filtered with all normal genomes to remove germline events. The structural variants are filtered out if both breakpoints fall into simple repeats or satellite repeats.

4 DNA copy number analysis

To this section was contributed by: Andrew D. Cherniack, Hailei Zhang, Bradley A. Murray, Gordon Saksena, and Carrie Sougnez

Correspondence and questions should be directed to: Andrew D. Cherniack (achernia@broadinstitute.org)

To interrogate genome wide DNA copy number levels, all tumor and normal samples were analyzed using Affymetrix SNP6.0 GeneChip arrays, as described previously (Cancer Genome Atlas Research, 2008, 2011, 2012, 2013a, b, 2014; Cancer Genome Atlas Research et al., 2013). Somatic copy number alterations (SCNAs) in 285 low grade gliomas were determined by SNP 6.0 analysis. Within each histological class, heterogeneous patterns of SCNAs were observed (Figure S8A). Similar to GBMs, chromosomal 7 gain with chromosome 9p and chromosome 10 loss were seen in all histological types. Additional arm level changes were seen in astrocytomas and oligoastrocytomas, but 1p loss and 19q loss were specific to oligodendrogliomas. GISTIC 2.0 analyses identified significantly reoccurring focal amplifications containing the oncogenes *MDM4*, *EGFR* and *CDK4* and deletion of the tumor suppressor *CDKN2A* in all histological groups (Table S3). Reoccurring 4q12 amplifications containing *PDGFRA*, *KDR* and *KIT* were found in astrocytomas and oligodendrogliomas but not oligoastrocytomas.

Distinct patterns of SCNAs become apparent when tumors were analyzed by molecular subtypes. In addition, unsupervised hierarchical clustering of arm level alterations produced three major clusters in which 87% of the tumors clustered with other members of their molecular subtype (Figure S8B). Tumors without *IDH* mutations (*IDH*wt) have patterns of both broad and focal SCNAs that are highly reminiscent of what is observed in GBM and dissimilar to tumors with *IDH* mutations (*IDH*mut) (Brennan et al., 2013). In particular, co-occurrence of both 7 gain and 10 loss occurs in 50% of tumors in this subgroup, yet this event is absent in *IDH*mut tumors. Reoccurring focal amplifications containing *EGFR*, *MDM4* and *SOX2* are also seen exclusively in *IDH*wt LGGs (Figure S14). The *IDH*mut-non-codel subgroup is different from the other two groups in that many tumors have either broad or focal amplification in 8q, possibly implicating *MYC* in this tumor type. Also unique to *IDH*mut-non-codel tumors are 10p amplifications in 27%, reoccurring focal amplifications containing *MYCN* and focal deletions containing *ATRX*. The *IDH*mut-codel subgroup has few reoccurring broad or focal copy number alterations outside of 1p and 19q deletions. These include reoccurring 4q12 focal amplifications and 1p focal deletions including *CDKN2C* that were present in a few. The *IDH*mut-codel group is also unlike the others in that it lacks reoccurring homozygous focal deletions of *CDKN2A*.

5 RNA sequencing

To the section on expression clustering was contributed by: Mark Vitucci, Siyuan Zheng, Olena Morozova, Da Yang, Youting Sun, Yuexin Liu, Brady Bernard, Sheila Reynolds, W.K. Alfred Yung, Greg Fuller, Wei Zhang, Mia Gifford, David Haussler, Ilya Shumlevich, Sofie Salama, Kenneth D. Aldape, Roel G.W. Verhaak, Ryan Miller

Correspondence and questions should be directed to: Ryan Miller (Ryan_Miller@med.unc.edu), Roel G.W. Verhaak (rverhaak@mdanderson.org)

To the section on gene fusions was contributed by: Olena Morozova, Siyuan Zheng, Isaac Joseph, Sol Katzman, Arjun Rao, David Haussler, Lior Pachter, Sofie Salama, Roel G.W. Verhaak

Correspondence and questions should be directed to: Sofie Salama (ssalama@soe.ucsc.edu), Roel G.W. Verhaak (rverhaak@mdanderson.org)

a) RNA sequencing data generation

RNA was extracted, prepared into mRNA libraries, and sequenced by Illumina HiSeq resulting in paired 50nt reads, and subjected to quality control as previously described (Cancer Genome Atlas Research, 2012). RNA reads were aligned to the hg19 genome assembly using Mapsplice (Wang et al., 2010). RNA fusion events were automatically detected by MapSplice as previously described (Cancer Genome Atlas Research, 2012). Note that these MapSplice fusion transcript annotations are available through the TCGA portal, but MapSplice fusion events were not used in the figures of this manuscript. Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1 using RSEM (Li and Dewey, 2011) and normalized within-sample to a fixed upper quartile. For further details on this processing, refer to Description file at the DCC data portal under the V2_MapSpliceRSEM workflow. Gene-level data was restricted to genes expressed in at least 70% of samples. Data were Log2 transformed and median centered across samples prior to further analysis.

b) Expression clustering and subtype election

The most variable genes were selected as the 1500 genes with the highest median absolute deviation. Consensus clustering was performed using ConsensusClusterPlus (Wilkerson and Hayes, 2010) (1000 iterations, resample rate of 80%, and Pearson correlation) and at k=6 a local maxima is reached. 254 out of 266 gliomas are in one of four subtypes. These four subtypes were further restricted to 239 "core" members based on positive silhouette width values (Rousseeuw, 1987). Single sample gene set enrichment was performed as previously described (Barbie et al., 2009) using astrocytoma (Gorovets et al., 2012) and neural ontology (Cahoy et al., 2008) signatures. For neural ontology signatures, the 500 most highly expressed genes per subtype were used. Subtype-specific genes used in Figure S6 and S7 were the 100 most highly expressed genes per class as defined by a 1-versus rest 2-class SAM (Tusher et al., 2001) (Table S7).

c) Gene expression clustering results

To identify mRNA expression subtypes, we performed unsupervised hierarchical clustering analysis on 266 gliomas using the 1500 most variably expressed genes across samples (Figure S6). Core members in four well-defined subtypes were identified and found to be distinctly enriched for previously defined astrocytoma subtype and neural ontology signatures and correlated with specific genomic events (Figure S7). Notably, expression subtype R2 was mostly composed of G3 tumors (77%), were mostly astrocytoma histology (68%), was enriched for methylation subtype M2 (62%) and *IDH*wt (67%) tumors, and correlated with GBM-related events such as *PTEN* mutation, chromosome 10 loss, and *EGFR* mutation and amplification. These were also enriched for the pre-glioblastoma expression signature previously identified in poor-surviving and *IDH*wt astrocytomas (Gorovets et al., 2012). Accordingly, R2 showed significantly worse survival as compared to the other RNA subtypes, which did not significantly differ from one another (Figure S6). The other RNA subtypes (R1, R3, R4) were populated with

IDH-mutant gliomas. R1 lacked 1p/19q codeletion, was comprised of two methylation subtypes M5 (70%) and M3 (30%), and the vast majority had *TP53* and *ATRX* mutations. Conversely, R3 was entirely composed of *IDH*mut-codel gliomas, and was equally distributed across methylation subtypes M2 and M3. It was also enriched for oligodendrogliomas (85%), mutations in *NOTCH1, FUBP1, CIC*, and oligodendrocyte progenitor-specific expression. Both R1 and R3 highly expressed an early progenitor-like astrocytoma gene signature. In contrast to the other expression subtypes, R4 was not overwhelmingly associated with a single molecular or methylation subtype, grade, histopathology, mutation, or genomic copy number event, but was heterogeneous throughout. However, R4 did highly express a neuron-specific signature and a neuroblastic astrocytoma signature (Figure S7B). Genes that characterize each of the expression subtypes (using a 2-class SAM algorithm) are shown in Table S7. In addition, gene ontology analyses using DAVID are shown in Table S7 to elucidate biologic characteristics for each of the expression subclasses. To address whether R4 tumor purity was lower than other subtypes, and may contain a higher ratio of native neurons to tumor cells, which could lead to enrichment in neuron-related signatures, we compared purity measurements across subtypes. We found average purity of R4 gliomas was 67.5%, lower than the next least pure subtype R2 at 71.3% (*P*=0.06), and significantly lower than R1 and R3 (P $\geq$ 0.009). In GBM, the neural subtype also showed high expression of this neuron-specific signature and those neural GBM were not less pure than the other GBM subtypes (Verhaak et al., 2010). Therefore, the explicit genotypic and phenotypic causes for variable transcription across glioma subtypes remains undefined, but gene signature enrichment patterns in Figure S7 suggest that transcriptomal subtypes of non-GBM gliomas are influenced by the lineage and differentiation state of the initial tumor cell of origin.

d) Fusion transcript detection using deFuse

RNA-Seq reads were analyzed using deFuse package version 0.6.0 (McPherson et al., 2011). A selection of deFuse predictions was computationally verified by aligning reads against the predicted transcript structure using TopHat (Trapnell et al., 2009) and visualizing the resulting alignments in the UCSC Genome Browser (Morozova, Grifford et al, in preparation). A Support Vector Machine (SVM) classifier available as part of the deFuse package was trained on 13 events verified as described above as well as 8 events predicted by both PRADA and defuse (21 events total). Following SVM analysis, candidate fusions were filtered based on the following deFuse parameters:

Splitr_count > 3

Span_count > splitr_count

SVM score > 0.65

Read_through ~ "N"

Splitr_span_pvalue > 0.1

Repeat_proportion1 < 0.78

Repeat_proportion2 < 0.78

Genome_breakseqs_percident < 0.1

Span_coverage_max > 1.2

All fusions that passed this filtering process are provided in Table S6 (events marked "deFuse" in "PRADA&deFusewithDNA" worksheet). Gene fusions affecting protein kinases and significantly mutated genes were further manually curated by tiled BLAT alignment analysis (Kent, 2002) of the fusion breakpoints using the UCSC Genome Browser (Karolchik et al., 2014) and examining the resulting alignment patterns (Table S6, "deFuse_reviewed" worksheet). The predicted protein sequences of gene fusions affecting *EGFR* and *FGFR3* are provided in Table S6 "Protein_sequences_RTK_fusions" worksheet.

e) Fusion transcript detection using PRADA

Transcript fusions were detected in 272 LGG samples by analyzing RNA sequencing data using the Pipeline for RNAseq Data Analysis (PRADA) (http://sourceforge.net/projects/prada/) for gene fusions(Torres-Garcia et al., 2014). PRADA aligned RNAseq reads to a composite reference database composed of whole genome sequences (hg19) and transcriptome sequences (Ensembl64). Two lines of evidence were required for identification of a gene fusion: 1) a minimum of two discordant read pairs mapping to a candidate gene pair; 2) a minimum of one junction spanning read mapping to a junction that connected exons between the candidate gene pair, with its pair mate mapping to the either of the two genes. Several filters were applied to remove false positives and artifacts, of which the most prominent is based on significant sequence similarity between the two fusion genes (using BLASTN, Expect value = 0.01). Specific details of the PRADA pipeline are described elsewhere (Torres-Garcia et al., 2014). PRADA predictions satisfying these criteria are listed in Table S6 "PRADA&deFusewithDNA"

worksheet marked “PRADA”. A summary of deFuse and PRADA predictions supported by DNA evidence or manual review are listed in Table S6 “Fusions_with_evidence” worksheet.

f) Quantification of fusion transcript isoform expression

For samples from the two patients with *FGFR3-TACC3* and *EGFR-SEPT14* fusions, respectively, we constructed plausible fusion transcript splice isoforms (FTSIs) based on predicted genomic fusion breakpoints from deFuse and Ensembl (Flicek et al., 2014) transcriptome assembly GRCh37 (hg19) database version 72. Genomic fusion breakpoints were filtered according to methods described in the RNA-Seq fusion analysis section of the supplementary material. Plausible FTSIs were added to the normal transcriptome to create a unique FTSI-appended transcriptome for each patient.

Plausible FTSIs were constructed by the following procedure: for each genomic fusion breakpoint, we enumerate a list of transcript splice isoforms crossing the upstream and downstream genomic breakpoint locations. Then, for each pairwise combination of upstream and downstream breakpoint-crossing isoforms, we append the appropriate regions from each transcript to create a plausible FTSI.

After appending each transcriptome with plausible FTSIs, we aligned RNA-Seq reads to the appended transcriptome using bowtie2 v 2.1.0 (Roberts et al., 2013), allowing for all possible alignments of each read, maximum fragment length of 800, and length of seed substring 25 (bowtie2 -a -X 800 -L 25 …). We then ran eXpress (Roberts and Pachter, 2013) v1.4.0 with default parameters on the resultant aligned reads.

Finally, in order to increase the accuracy of the expression estimates for the FTSIs and their constituent transcripts, we modified reXpress (Roberts et al., 2013). The modified version runs extra expectation-maximization (EM) rounds of eXpress on only specified transcripts for computational efficiency purposes; each EM round increases the likelihood of the probabilistic model for read generation by updating abundance and other parameters. The modified version will be released shortly at http://bio.math.berkeley.edu/ReXpress/, and this behavior is known as ‘focus’ mode. We used parameters for 50 maximum alignments for each read, and partitioning of the transcript graph via the greedy algorithm (reXpress --greedy --max-alignments=50 --focus …). The resultant expression estimates allow for comparison of the expression of the FTSIs with all transcripts in the reference transcriptome, including the constituents of the fusion transcripts.

6 MicroRNA sequencing

To this section was contributed by: A. Gordon Robertson, Brady Bernard, Reanne Bowlby, Denise Brooks, Andy Chu

Correspondence and questions should be directed to: A. Gordon Robertson (grobertson@bcgsc.ca)

a) Methods

microRNA sequence (miRNA-seq) data were generated for 293 tumors using previously described methods (Cancer Genome Atlas, 2012). We used unsupervised NMF consensus clustering of reads-per-million (RPM) data to identify groups of samples with similar abundance profiles; as previously described (Cancer Genome Atlas, 2012), the input was the ~300 (25%) most-variant 5p or 3p miRBase v16 mature strands. We generated a log2, median-centered heatmap for the discriminatory miRNAs that had the top of scores (e.g. 5%) in each of the five NMF metagenes (Gaujoux and Seoighe, 2010). For the heatmap, columns (samples) in a RPM-normalized abundance matrix were ordered to match the NMF result. Rows of this matrix (i.e. miRNA abundances) were log2-transformed, median-centered, and then hierarchically clustered using an absolute centered correlation distance metric and average linkage(de Hoon et al., 2004). 5p and 3p strand names were assigned using miRBase v20.

Contingency table association P-values were generated with a Fisher exact test using R v3.0.x. Kaplan-Meier curves were calculated using the survival package v2.37-7 in R 3.0.3. Tumor purity was estimated using ABSOLUTE (Carter et al., 2012).

We used SAMseq's (samr v2.0, R 3.0.2) two-class unpaired analyses with a read count input matrix and an FDR threshold of 0.05 to identify miRs that were differentially expressed (see Overview worksheet in Table S8). Each run generated a pair of files: genes 'up' and 'down'. We filtered each file by removing miRs with median expression less than 50 RPM in either of the input sample groups (Tay et al., 2014), and miRs for which the Wilcoxon adjusted P-value was greater than 0.05; then ranked the filtered results by a median-based fold change.

We calculated anticorrelations between normalized abundance for miRNA mature strands (RPM) and RPPA data for 189 antibodies with MatrixEQTL v2.1.1 (Shabalin, 2012), after transforming miR and antibody abundances into integer ranks. We worked with the 255 of 293 miRNA tumor samples that were present in both miRNA and RPPA data. Restricting the miRs to the 481 mature strands with a mean RPM above the 60th percentile (0.83 RPM), we calculated correlations with a P-value threshold of 0.05, and filtered the resulting anticorrelations by FDR<0.05. After associating gene symbols with antibodies, we then extracted records that corresponded to miR-gene (i.e. miR-antibody) pairs that were supported by functional validation publications reported by MiRTarBase v4.5 (Hsu et al., 2014), for stronger (e.g. luciferase reporter, Western blot) vs. weaker (e.g. microarray, sequencing, immunoblot, qRT-PCR, proteomics, CLASH) experimental evidence types. We also extracted records that corresponded to miR-gene (i.e. miR-antibody) pairs in TargetScan v6.2 (Grimson et al., 2007) conserved and nonconserved predicted targeting relationships.

To generate functional insight from the miR-antibody anticorrelations, we considered only miRs that were differentially abundant (FDR<0.05) between each molecular subtype and all other samples. For these miRs, we extracted the miR-antibody anticorrelation records that were supported by functional validation publications with strong evidence types. This generated a list of miRs and associated antibodies, for which we generated plots of normalized abundance across the molecular subtypes. We identified the subset of these antibodies whose abundance varied significantly between molecular subtypes by using R v3.1.2 to calculate BH-corrected P-values for Wilcoxon and KS tests, for all samples in each subtype, against all other tumor samples (n=255). Finally, we manually assessed antibodies with significant differences in abundance between subtypes, miR abundance distributions, and significant miR-antibody anticorrelations, identifying miRs and antibodies for which the changes in RPM abundance distributions of potentially targeting miRs were complementary to the protein abundance changes across subtypes.

b) Results

Unsupervised clustering For RPM abundance profiles for 293 tumor samples, rank survey profiles for cophenetic correlation coefficient and average silhouette width suggested 4- and 7-group (cluster) solutions (Figure S9). Here, we show the 4-group solution, since covariate associations did not strongly recommend the 7-group solution over the 4 (data not shown), and because this solution had a number of significant associations with clinical parameters or other molecular results. For example, groups 3 and 4 were enriched in astrocytoma samples (Fisher exact P=3.3e-5, Figure S1c), *IDH*wt samples (P=1.2e-12), mRNAseq cluster 2 (P=5.6e-17), DNA methylation cluster K4 (P=3.1e-11), and sCNA cluster 3 (P=4.8e-9). Cluster 3 was enriched in RPPA cluster 3 (P=5.2e-9).

Group 2 was large (n=184), and many of its discriminatory miRs were relatively less abundant. Group 1 was intermediate in size (n=69), and many of its discriminatory miRs were relatively abundant. Groups 3 and 4 were small (n=17, 23 respectively). Visually, the normalized abundance heatmap suggested a reasonably high level of sample heterogeneity. Tumor purity was lowest in group 1, intermediate in group 3, and highest in groups 2 and 4 (Figure S9).

Kaplan-Meier plots showed statistically significant overall differences in overall survival (OS) and progression-free survival (PFS) (log-rank P-values were 4.3e-6 and 0.0054 respectively). Of the four groups, group 2 had the most favorable OS and PFS, group 1 was intermediate, and groups 3 and 4 had the least favorable outcomes (Figure S9e).

The heatmap (Figure S9b) shows the 37 miRNA 5p and 3p strands that had relatively high cores in the NMF metagene matrix; i.e. miRs that were more discriminatory in unsupervised clustering (Gaujoux and Seoighe, 2010). A number of these have been reported for glioma or GBM and have been the subject of reviews (Brower et al., 2014; Karsy et al., 2012; Nikaki et al., 2012; Palumbo et al., 2014): miR-9 (proliferation, self-renewal), miR-10a/b (apoptosis, autophagy, chemoresistance, invasion, prognosis, proliferation, senescence, tumor growth), miR-21 (apoptosis, chemoresistance, invasion, proliferation, tumor growth), miR-101 (invasion, neoangiogenesis, proliferation, tumor growth), miR-128 (proliferation, self-renewal, tumor growth), miR-181a (apoptosis, colony formation, invasion, proliferation, radiosensitivity) and let-7a (migration, proliferation).

As noted, groups 3 and 4 were enriched in *IDH*wt samples (Fisher exact P=1.2e-12). Groups 1, 3 and 4 were depleted in *IDH*mut-codel, while group 2 was enriched in this subtype (P=5.2e-4). The *IDH*mut-non-codel subtype was marginally enriched in group 2 (P=0.048).

miRs and molecular subtypes We identified miRs that were differentially abundant (DA) between the three molecular subtypes, using two-class analyses for each subtype vs. all other samples, and then for pairs of subtypes (Figure S10b-g, Table S8). We noted that many of these miRs are known to be either relatively abundant, or depleted, in gliomas and/or glioblastoma (GBM), relative to noncancerous tissues or cell lines, and to be involved in glioma oncogenesis or progression (Brower et al., 2014; Karsy et al., 2012; Nikaki et al., 2012; Palumbo et al., 2014). Such miRs may be influential in differences between molecular subtypes.

Considering only the top 15 positive and negative fold changes in Figure S10b-g, the following miRs were differentially abundant between subtypes, and are in described in the above reviews as relatively abundant in gliomas and/or GBM.

- Both miR-9-5p and -3p were less abundant in *IDH*wt than in either *IDH*mut-codel or *IDH*mut-non-codel samples (Figure S10e,f).
- The 5p mature strands of miR-10a and 10b were more abundant in *IDH*wt than in *IDH*mut-codel (Figure S10f). miR-10a-5p was strongly more abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e**)** or the two *IDH*mut groups combined (Figure S10b), and less abundant in *IDH*mut-non-codel than in *IDH*wt combined with *IDH*mut-codel (Figure S10c). miR-10b-5p was more abundant in *IDH*mut-non-codel than in *IDH*mut-codel (Figure S10g).
- miR-17-5p, 19b-3p and 20a-5p from the miR-17~92a genomic cluster (Tan et al., 2014) were less abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e). miR-20a-5p was more abundant in *IDH*mut-codel than in *IDH*wt (Figure S10f). miR-17-5p may be influential in targeting p21/CDKN1A (below).
- miR-21-5p and 3p were more abundant in *IDH*wt than in *IDH*mut-codel, or the two *IDH*mut subtypes combined (Figure S10f,b), and both strands were more abundant in *IDH*mut-non-codel than in *IDH*mut-codel (Figure S10g). miR-21-5p, with miR-23a-3p, may be influential in targeting PTEN (below).
- miR-155-5p (Ling et al., 2013; Liu et al., 2014; Sun et al., 2014; Zhou et al., 2013) was more abundant in *IDH*wt than in either *IDH*mut subtype (Figure S10e,f).
- The 5p strands of miR-182 (Song et al., 2012; Tang et al., 2014)and miR-183 from the miR-183/96/182 genomic cluster were less abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e) or *IDH*mut-codel samples (Figure S10f).
- miR-196b-5p was strongly more abundant in *IDH*wt than in the two *IDH*mut subtypes combined (Figure S10b).
- miR-221-3p (Quintavalle et al., 2013) was more abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e) or *IDH*mut-codel (Figure S10f). miR-221-3p may be influential in targeting HER3 and ER-alpha (below).

In contrast, the following differentially abundant miRNAs are described in reviews as relatively depleted in gliomas and GBM.

- miR-124-3p (An et al., 2013; Chen et al., 2014; Lv and Yang, 2013) was more abundant in *IDH*mut-codel than in *IDH*mut-non-codel (Figure S10g), and less abundant in *IDH*mut-non-codel than in *IDH*wt + *IDH*mut-non-codel combined (Figure S10c).
- miR-146b-5p (Li et al., 2013) and 3p were more abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e) or *IDH*mut-codel (Figure S10f).
- Two miR-181 family members were differentially abundant. miR-181a-2-3p was less abundant in *IDH*wt than in I*DH*mut-codel (Figure S10f) or the two *IDH*mut groups combined (Figure S10b). miR-181c-3p was more abundant in *IDH*mut-non-codel than in the other two subtypes combined (Figure S10c). Both miR-181a-2-3p and 181c-5p were less abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e).

We also noted a number of differentially abundant miRs that were not listed in the above reviews, but have been reported for glioma or GBM.

- miR-23a-3p (Hu et al., 2013; Koshkin et al., 2014; Tan et al., 2012) was more abundant in *IDH*wt than in the two *IDH*mut groups combined (Figure S10b), and more abundant in *IDH*mut-non-codel than in *IDH*mut-codel (Figure S10g). miR-23a-3p, with miR-21-5p, may be influential in targeting PTEN (below).
- miR-99a-5p (Chakrabarti et al., 2013; Parker et al., 2013) was more abundant in *IDH*mut-non-codel than in *IDH*mut-codel (Figure S10g) or *IDH*mut-codel and *IDH*wt combined (Figure S10c).
- The glioma tumor suppressor miR-143-3p (Wang et al., 2014) was more abundant in *IDH*wt (Figure S10e) and in *IDH*mut-codel (Figure S10g) than in *IDH*mut-non-codel.
- miR-148a-3p is relatively abundant in human glioblastoma specimens, cell lines, and stem cells (Kim et al., 2014). It was more abundant in *IDH*wt than in I*DH*mut-codel (Figure S10f) or the two *IDH*mut subtypes combined (Figure S10b), and less abundant in *IDH*mut-codel than in the other two subtypes combined (Figure S10d).
- miR-204-5p (Mao et al., 2014; Ying et al., 2013) was more abundant in *IDH*wt than in *IDH*mut-non-codel (Figure S10e) or *IDH*mut-codel (Figure S10f), and was strongly less abundant in *IDH*mut-codel than in the other two subtypes combined (Figure S10d).

miR targeting and the proteome We assessed potential miR targeting through Spearman correlations between normalized abundance profiles for miRNA mature strands and RPPA protein data for 189 antibodies, for 255 of the 293 tumor samples. Of 17277 significant anti-correlations (FDR<0.05, Table S9, tab 1), 77 miR-antibody pairs corresponding to 36 antibodies were supported by functional validation publications with stronger evidence types (Methods, Figure S10h), and 128 miR-antibody pairs corresponding to 70 antibodies were supported by functional validation publications with weaker evidence types (Table S9, tabs 2 and 3). Target predictions from TargetScan v6.2 returned 155 conserved and 920 nonconserved miR-antibody pairs that corresponded to 52 and 102 antibodies respectively (Table S9, tabs 4 and 5).

To generate functional insight from these miR-protein anticorrelations, we assessed miRs that were differentially abundant (FDR<0.05) between the molecular subtypes, and that were anticorrelated to antibodies (FDR<0.05) in potential targeting relationships that had been functionally validated with stronger evidence types. We anticipated that a subset of these relationships could identify miRs that may influence differences in protein levels between the molecular subtypes. This process identified 26 miRs and 22 antibodies (Figures S10i and j, Table S9, tabs 6 to 8**)**. A number of the antibodies had statistically significant differences in abundance between molecular subtypes (Figure S10i, Table S9, tab 9), and we noted a number of miR-antibody pairs for which the changes in abundance distributions of the potentially targeting miRs were complementary to the protein abundance changes across the subtypes (Figure S10k). For example, PTEN levels, which are progressively lower as we consider *IDH*mut-codel, *IDH*mut-non-codel and then *IDH*wt samples, may be influenced by miR-21-5p and miR-23a-3p (Figure S10k-i). p21 (CDKN1A) levels may be influenced by miR-17-5p, which is expressed from the oncogenic miR-17~92a genomic cluster (Tan et al., 2014) (Figure S10k-ii). Low levels of HER3 and ER-alpha in *IDH*wt samples may be influenced by levels of miR-22-3p and miR-221-3p (Figure S10k-iii). We noted above that miR-21-5p, miR-23a-3p, miR-17-5p and miR-221-3p are known to be involved in gliomas (see above reviews); in contrast, while miR-22 is associated with diverse cancers (Wan et al., 2014), to our knowledge it has not been reported as associated with glioma.

7 Genome wide DNA methylation profiling

To this section was contributed by: Houtan Noushmehr, Thais S. Sabedot, Simon G. Coetzee, Daniel J. Weisenberger, Toshi Hinoue, Hui Shen, Peter W. Laird

Correspondence and questions should be directed to: Houtan Noushmehr (houtan@usp.br), Peter W. Laird (plaird@usc.edu)

a) DNA methylation analysis

DNA Methylation cluster identification was performed as previously described (Cancer Genome Atlas Research, 2014). In short, we used the Illumina Infinium HumanMethylation450 (HM450) platform (Illumina, San Diego, CA) to obtain DNA methylation profiles of 289 TCGA lower grade glioma samples. We also obtained 77 non-tumor brain samples from Gene Expression Omnibus (GEO)(Guintivano et al., 2013), profiled using HM450. This platform covers 99% of RefSeq genes with multiple probes per gene, 96% of CpG islands from the UCSC database and their flanking regions. 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, as detailed below. Further details on the Illumina Infinium HM450 DNA methylation assay technology has been described previously. The assay probe sequences and information on each interrogated CpG/CpH site on the Infinium HM450 BeadChip are available from Illumina (www.illumina.com).

The data levels and the files contained in each data level package are described below and are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/) and the publication page (https://tcga-data.nci.nih.gov/docs/publications/lgg_2015/).

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. Level 1: Level 1 data contain raw IDAT files (two per sample) as produced by the iScan system. Level 2: Level 2 data contain background-corrected methylated (M) and unmethylated (U) summary intensities as extracted by the methylumi R package. Non-detection probabilities (P values) were computed as the minimum of the two values (one per allele) for the empirical cumulative density function of the negative control probes in the appropriate color channel. Background correction is performed via normal-exponential deconvolution (currently not stratified by probe sequence). Multiple-batch archives have the intensities in each of the two channels multiplicatively scaled to match a reference sample (sample with R/G ratio closest to 1.0). Level 3: Level 3 data contain β-value calculations with HGNC gene symbol, chromosome

(UCSC hg19, Feb 2009), and genomic coordinate (UCSC hg19, Feb 2009) for each targeted CpG/CpH site on the array. Probes having a common SNP (MAF > 0.01, per dbSNP build 135 via the UCSC snp135common track) within 10bp of the interrogated CpG site or having 15bp from the interrogated CpG site overlap with a repetitive element (as defined by RepeatMasker and Tandem Repeat Finder Masks based on UCSC hg19, Feb 2009) are masked as “NA” across all samples, and probes with a non-detection probability (P value) greater than 0.01 in a given sample are masked as “NA” on that chip. Probes that are mapped to multiple sites on hg19 are annotated as “NA” for chromosome and 0 for CpG/CpH coordinate.

b) Unsupervised clustering analysis of DNA methylation data

Methods were used as recently described (Cancer Genome Atlas Research, 2014) 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. To capture cancer-specific DNA hypermethylation events, we further eliminated sites that were methylated (mean β-value ≥0.2) in histologically normal brain tissues. 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 consensus clustering on 11,977 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 evaluating the reports generated by consensus clustering. The probes are arranged based on the order of unsupervised hierarchal clustering of the DNA methylation β-value data. We performed Fisher’s exact tests to quantify associations between mutations and clustering assignments results. To identify probes that show significant DNA methylation differences between the hypermethylated subgroup (M1; n= 12) and all the other groups we performed Wilcoxon rank-sum test on β-values across all loci after “NA”-masked and sex-linked probes are eliminated (n = 380,836). The resulting P values were corrected using the Benjamini-Hochberg procedure.

Others Statistical analysis and data visualization were carried out using the R/Biocoductor software packages (http://www.bioconductor.org).

8 Protein expression using Reverse Phase Protein Arrays

To this section was contributed by: Rehan Akbani, Wenbin Liu, Zhenlin Ju, Yiling Lu, Gordon B.Mills

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

a) RPPA experiments and 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 (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 189 validated primary antibodies 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 CanoScan 9000F. Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI (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 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 (Gonzalez-Angulo et al., 2011; Hu et al., 2007) using median centering across antibodies (level 3 data). In total, 189 antibodies and 255 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.

b) 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.

c) Hierarchical clustering

We used bootstrap to resample (N=3000) the proteins to estimate the number of sample clusters. Pearson correlation was used as distance metric and Ward was used as a linkage algorithm in the unsupervised hierarchical clustering analysis. This method clustered the samples and counted how frequently two samples were in the same cluster (Figure S20A).The bootstrap resampling analysis identified four robust sample clusters (Figure S20B). The four clusters and their protein expression patterns can be viewed through the next generation clustered heat map (NG-CHM) pipeline developed at the University of Texas MD Anderson Cancer Center. A total of 255 samples and 189 antibodies were used in the analysis. The analysis showed an interesting cluster (third from the right in Figure S20B) that had depletions in *IDH* mutations along with enrichment in *EGFR* and *PTEN* mutations. The *PTEN* mutations corresponded with decreased PTEN protein levels, whereas *EGFR* mutations tended to increase protein expression (Figure S20C) creating possibilities for clinically targeting EGFR. Interestingly, the same cluster also had high levels of HER2, further increasing the chances for targeted therapy. The first cluster from the right, on the other hand, had high PTEN, BRAF, and MEK1 levels. The second cluster had high Beta Catenin, ERK2, PKC alpha, phosphoPKC alpha and PKC delta protein levels. The last cluster showed depletion of all of those proteins. Clusters 1, 2 and 4 all showed enrichment of *IDH*, *TP53*, and *ATRX* mutants and had many grade II tumors. Cluster 3 was depleted in all of those features, and had poor overall survival and progression free survival (Figure S20D).

9 Miscellaneous Analysis

To this section was contributed by: Mia Grifford, Olena Morozova, Sam Ng, Dan Carlin, Josh Stuart, Sofie Salama, David Haussler

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

a) Cluster of clusters analysis

Cluster of clusters analysis was performed as previously described (Cancer Genome Atlas, 2012). Briefly, subtype calls from each of the following 4 platforms: mRNA, miRNA, methylation, and copy number were used to identify relationships between the different data type's classifications. Subtypes defined from each platform were coded into a series of indicator variables, resulting in a matrix of 1s and 0s. Hierarchical clustering of this matrix was performed using the ConsensusClusterPlus R-package (Monti et al., 2003; Wilkerson and Hayes, 2010). Parameters for ConsensusClusterPlus were 80% sample resampling with 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric. Cluster assignments were determined for different total numbers of clusters ranging from 2 to 7. The best total number of clusters was determined to be 3 based on the cumulative distribution functions, cophenetic correlation coefficients and average silhouette widths (Figure S23).

b) Comparative analysis of IDHwt LGGs and GBMs

Figure 5A was generated using the UCSC Cancer Genomics Browser (Goldman et al., 2013) chromosome view of the TCGA LGG GISTIC thresholded copy number dataset. Samples were grouped based on the *IDH* mutation and 1p/19q co-deletion status into 3 subtypes (*IDH*wt, *IDH*mut-codel and *IDH*mut-non-codel) and then sub-grouped based on tumor grade. Percentages of copy number events in LGGs are based on the CNV arm-level calls (See Copy Number Analysis Methods). Copy number calls indicate predicted hemizygous amplifications or deletions. The percentages of corresponding arm-level CNV alterations in GBMs are based on GISTIC thresholded arm-level calls (Brennan et al., 2013). Percentages of double minute chromosomes/homogenously staining regions are based on predictions made from exome data (See exome analysis to identify candidate double minute chromosomes in methods).

Mutation calls in Figure 5B were obtained from the cross-center combined MAF file, which included mutations in the 285 samples analyzed using whole exome sequencing. Copy number alterations in Figure 5B were obtained from the GISTIC thresholded gene-level analysis, and included gene-level copy number events in the 266 samples with copy number data. A gene was considered to be amplified if the GISTIC thresholded value was 2 and deleted if it was -2. High confidence DNA rearrangements identified using BamBam (Sanborn et al., 2013) and BreakDancer (Chen et al., 2009) and listed in Table S5 were also included in Figure 5B and denoted “SV”. Manually reviewed fusion predictions (Table S6) affecting genes in Figure 5B were also included. Samples were divided into 3 subtypes based on the IDH mutation and 1p/19q co-deletion status, as described above. Mutation, amplification, deletion, structural variant (SV) and fusion frequencies in TCGA GBM samples were as previously published (Brennan et al., 2013). *IDH* mutation status in TCGA GBM samples was available for 423 patients, determined either by whole-exome Illumina sequencing or Sanger-based sequencing (Cibulskis et al., 2013).

c) Analysis of NOTCH1 mutation and fusion signaling consequences

As *NOTCH1* appears to be a highly recurrent mutation identified in the *IDH*mut-codel subgroup we investigated the pathway evidence for loss-of-function, gain-of-function, or neutrality of specific mutations in this gene across the LGG cohort. To do this, we employed the PARADIGM-SHIFT algorithm (Ng et al., 2012). PARADIGM-SHIFT assigns a pathway impact score for each mutation. Along with mutations in *NOTCH1*, we also analyzed one homozygous deletion, one genomic rearrangement, and two fusion events in *NOTCH1*. Thus, pathway signatures identified by PARADIGM-SHIFT shared with the *NOTCH1* mutations could provide clues about the mechanism of action of these genomic events. There were 246 samples within the *IDH* mutant and wild-type subtypes and 1 sample that was not assigned a subtype (*NOTCH1* fusion) with available genomic data to run PARADIGM-SHIFT analysis, with 29

NOTCH1 mutations, one homozygous deletion, one genomic rearrangement, and two fusions in this set. PARADIGM parameters were trained on the complete cohort of samples with available copy number and expression data with a total of 272 samples. The *NOTCH1* mutation neighborhoods were selected in a supervised fashion by selecting features based on Fisher score. PARADIGM-SHIFT (P-Shift) scores for *NOTCH1*, which reflect the discrepancy in upstream versus downstream pathway signals, were calculated as the difference in inferred activity between the two runs of PARADIGM; one where only the connections with the upstream regulators are retained (R-run) and one where only the connections with the downstream targets are retained (T-run). When the distribution of P-Shift scores for samples with alterations in *NOTCH1* are compared to samples without either of these alterations, an enrichment of negative P-Shift scores was identified indicating loss-of-function (LOF) on average through the *NOTCH1* signaling pathway. The significance of this aggregated LOF score was determined by running a background model in which the selected network topology is fixed, but the data is permuted, thus assigning random genes to the surrounding network neighborhood of the *NOTCH1* protein. Under this background model, the LOF aggregated score was found to have a p-value of 0.02. Altogether, these findings suggest that the signaling consequences of *NOTCH1* genomic events in LGG lead to LOF based on the discrepancy of up- vs. down- stream activity signals. PARADIGM-SHIFT was run on the complete cohort to determine the functional impact of alterations on the network and its network was viewed with a CircleMap display (Figure S16)(Wong et al., 2013). The pattern of expression for many of the downstream targets of *NOTCH1* mirrors the profile of P-Shift score concordant with *NOTCH1* pathway deactivation in the samples with alterations. Low P-Shift scores are also observed in many of the other samples within the IDHmut-codel subtype, which suggests there may be additional mechanisms of NOTCH signaling pathway deactivation not considered in this analysis.

10 Batch Effects

To this section was contributed by: Rehan Akbani, Shiyun Ling, John N. Weinstein

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

a) Methods

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the low grade glioma data sets. Four different data sets were analyzed: miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium HM450 microarray), mRNA sequencing (Illumina HiSeq), and protein expression (reverse-phase protein arrays). All of the data sets were at TCGA level 3, since this represents the level on which most analyses are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS). Detailed results and batch effects analysis of those and other TCGA data sets can be found at: http://bioinformatics.mdanderson.org/tcgabatcheffects.

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To visually assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines.Centroids were computed by taking the mean across all samples in the batch. This produced a visual representation of the relationships among batch centroids and the scatter within batches. Results for the four data sets follow.

b) miRNA (Illumina HiSeq)

Figures S24A-C show clustering and PCA plots for miRNA seq data. miRNAs with zero values were removed and the read counts were log2-transformed before generating the figures. Figure S24A shows some batch effects in batch numbers 78 and 112. However, as seen from Figure S24B, the magnitude of batch effects was not substantial and did not warrant batch effects correction for the type of analyses. Batch effects correction algorithms may lead to loss of important biological variation in the data, along with the technical variation.

c) DNA Methylation (Infinium HM450 microarray)

Figures S24D-F show clustering and PCA plots for the Infinium DNA methylation platform. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

d) RNASeqV2 (RNA-Seq Illumina HiSeq)

Figures S24G-I show clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the values were log2-transformed before generating the figures. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

e) Protein expression (Reverse-Phase Protein Array – RPPA)

Figures S24J-K show clustering and PCA plots for protein expression data using the RPPA platform. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

f) Conclusions

Batch effects were analyzed in four different data sets. miRNA data showed a small batch effect in batch numbers 78 and 112. These weren’t considered strong enough to warrant algorithmic batch effect correction, since that often removes useful biology along with batch effects. DNA methylation, mRNA and protein expression data didn’t show major batch effects.

11 References

An, L., Liu, Y., Wu, A., and Guan, Y. (2013). microRNA-124 inhibits migration and invasion by down-regulating ROCK1 in glioma. PLoS One *8*, e69478.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C.*, et al.* (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108-112.

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.

Brower, J.V., Clark, P.A., Lyon, W., and Kuo, J.S. (2014). MicroRNAs in cancer: glioblastoma and glioblastoma cancer stem cells. Neurochem Int *77*, 68-77.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A.*, et al.* (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. The Journal of neuroscience : the official journal of the Society for Neuroscience *28*, 264-278.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

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. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Cancer Genome Atlas Research, N. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

Cancer Genome Atlas Research, N. (2013a). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43-49.

Cancer Genome Atlas Research, N. (2013b). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England journal of medicine *368*, 2059-2074.

Cancer Genome Atlas Research, N. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315-322.

Cancer Genome Atlas Research, N., Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R.*, et al.* (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67-73.

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. Nature biotechnology *30*, 413-421.

Chakrabarti, M., Banik, N.L., and Ray, S.K. (2013). Photofrin based photodynamic therapy and miR-99a transfection inhibited FGFR3 and PI3K/Akt signaling mechanisms to control growth of human glioblastoma In vitro and in vivo. PLoS One *8*, e55652.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M.*, et al.* (2011). Initial genome sequencing and analysis of multiple myeloma. Nature *471*, 467-472.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P.*, et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods *6*, 677-681.

Chen, Q., Lu, G., Cai, Y., Li, Y., Xu, R., Ke, Y., and Zhang, S. (2014). MiR-124-5p inhibits the growth of high-grade gliomas through posttranscriptional regulation of LAMB1. Neuro Oncol *16*, 637-651.

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.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601-2602.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80-92.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D*., et al.* (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic acids research *41*, e67.

de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453-1454.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S*., et al.* (2014). Ensembl 2014. Nucleic acids research *42*, D749-755.

Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A*., et al.* (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic acids research *38*, D652-657.

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC bioinformatics *11*, 367.

Goldman, M., Craft, B., Swatloski, T., Ellrott, K., Cline, M., Diekhans, M., Ma, S., Wilks, C., Stuart, J., Haussler, D*., et al.* (2013). The UCSC Cancer Genomics Browser: update 2013. Nucleic acids research *41*, D949-954.

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. Clinical proteomics *8*, 11.

Gorovets, D., Kannan, K., Shen, R., Kastenhuber, E.R., Islamdoust, N., Campos, C., Pentsova, E., Heguy, A., Jhanwar, S.C., Mellinghoff, I.K*., et al.* (2012). IDH mutation and neuroglial developmental features define clinically distinct subclasses of lower grade diffuse astrocytic glioma. Clin Cancer Res *18*, 2490-2501.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

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 : official journal of the DNA Methylation Society *8*, 290-302.

Heagerty, P.J., and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. Biometrics *61*, 92-105.

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. Clinical proteomics *6*, 129-151.

Hennessy, B.T., Lu, Y., Poradosu, E., Yu, Q., Yu, S., Hall, H., Carey, M.S., Ravoori, M., Gonzalez-Angulo, A.M., Birch, R*., et al.* (2007). Pharmacodynamic markers of perifosine efficacy. Clinical cancer research : an official journal of the American Association for Cancer Research *13*, 7421-7431.

Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y.*, et al.* (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res *42*, D78-85.

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.

Hu, X., Chen, D., Cui, Y., Li, Z., and Huang, J. (2013). Targeting microRNA-23a to inhibit glioma cell invasion via HOXD10. Sci Rep *3*, 3423.

Institute, B. (2014). Broad Institute Firehose Portal.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2014). The UCSC Genome Browser database: 2014 update. Nucleic acids research *42*, D764-770.

Karsy, M., Arslan, E., and Moy, F. (2012). Current Progress on Understanding MicroRNAs in Glioblastoma Multiforme. Genes Cancer *3*, 3-15.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome research *12*, 656-664.

Kim, J., Zhang, Y., Skalski, M., Hayes, J., Kefas, B., Schiff, D., Purow, B., Parsons, S., Lawler, S., and Abounader, R. (2014). microRNA-148a is a prognostic oncomiR that targets MIG6 and BIM to regulate EGFR and apoptosis in glioblastoma. Cancer Res *74*, 1541-1553.

Koshkin, P.A., Chistiakov, D.A., Nikitin, A.G., Konovalov, A.N., Potapov, A.A., Usachev, D.Y., Pitskhelauri, D.I., Kobyakov, G.L., Shishkina, L.V., and Chekhonin, V.P. (2014). Analysis of expression of microRNAs and genes involved in the control of key signaling mechanisms that support or inhibit development of brain tumors of different grades. Clin Chim Acta *430*, 55-62.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics *12*, 323.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, Y., Wang, Y., Yu, L., Sun, C., Cheng, D., Yu, S., Wang, Q., Yan, Y., Kang, C., Jin, S.*, et al.* (2013). miR-146b-5p inhibits glioma migration and invasion by targeting MMP16. Cancer Lett *339*, 260-269.

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. Nature cell biology *9*, 218-224.

Ling, N., Gu, J., Lei, Z., Li, M., Zhao, J., Zhang, H.T., and Li, X. (2013). microRNA-155 regulates cell proliferation and invasion by targeting FOXO3a in glioma. Oncol Rep *30*, 2111-2118.

Liu, S., Yin, F., Zhang, J., Wicha, M.S., Chang, A.E., Fan, W., Chen, L., Fan, M., and Li, Q. (2014). Regulatory roles of miRNA in the human neural stem cell transformation to glioma stem cells. J Cell Biochem *115*, 1368-1380.

Lv, Z., and Yang, L. (2013). MiR-124 inhibits the growth of glioblastoma through the downregulation of SOS1. Mol Med Rep *8*, 345-349.

Mao, J., Zhang, M., Zhong, M., Zhang, Y., and Lv, K. (2014). MicroRNA-204, a direct negative regulator of ezrin gene expression, inhibits glioma cell migration and invasion. Mol Cell Biochem *396*, 117-128.

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.

Mogensen, U.B., Ishwaran, H., and Gerds, T.A. (2012). Evaluating Random Forests for Survival Analysis using Prediction Error Curves. J Stat Softw *50*, 1-23.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning *52*, 91-118.

Ng, S., Collisson, E.A., Sokolov, A., Goldstein, T., Gonzalez-Perez, A., Lopez-Bigas, N., Benz, C., Haussler, D., and Stuart, J.M. (2012). PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics *28*, i640-i646.

Nikaki, A., Piperi, C., and Papavassiliou, A.G. (2012). Role of microRNAs in gliomagenesis: targeting miRNAs in glioblastoma multiforme therapy. Expert Opin Investig Drugs *21*, 1475-1488.

Palumbo, S., Miracco, C., Pirtoli, L., and Comincini, S. (2014). Emerging roles of microRNA in modulating cell-death processes in malignant glioma. J Cell Physiol *229*, 277-286.

Parker, B.C., Annala, M.J., Cogdell, D.E., Granberg, K.J., Sun, Y., Ji, P., Li, X., Gumin, J., Zheng, H., Hu, L.*, et al.* (2013). The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma. J Clin Invest *123*, 855-865.

Quintavalle, C., Mangani, D., Roscigno, G., Romano, G., Diaz-Lagares, A., Iaboni, M., Donnarumma, E., Fiore, D., De Marinis, P., Soini, Y.*, et al.* (2013). MiR-221/222 target the DNA methyltransferase MGMT in glioma cells. PLoS One *8*, e74466.

Radenbaugh AJ, Ma S, Ewing A, Stuart J, Collisson EA, Zhu J, and D, H. (2014). RADIA: RNA and DNA Integrated Analysis for Somatic Mutation Detection. PLOS ONE *doi: 10.1371/journal.pone.0111516*.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nature methods *10*, 71-73.

Roberts, A., Schaeffer, L., and Pachter, L. (2013). Updating RNA-Seq analyses after re-annotation. Bioinformatics *29*, 1631-1637.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math *20*, 53-65.

Sanborn, J.Z., Salama, S.R., Grifford, M., Brennan, C.W., Mikkelsen, T., Jhanwar, S., Katzman, S., Chin, L., and Haussler, D. (2013). Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. Cancer research *73*, 6036-6045.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811-1817.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353-1358.

Shaw, E.G., Scheithauer, B.W., and O'Fallon, J.R. (1997). Supratentorial gliomas: a comparative study by grade and histologic type. J Neurooncol *31*, 273-278.

Song, L., Liu, L., Wu, Z., Li, Y., Ying, Z., Lin, C., Wu, J., Hu, B., Cheng, S.Y., Li, M.*, et al.* (2012). TGF-beta induces miR-182 to sustain NF-kappaB activation in glioma subsets. J Clin Invest *122*, 3563-3578.

Sun, J., Shi, H., Lai, N., Liao, K., Zhang, S., and Lu, X. (2014). Overexpression of microRNA-155 predicts poor prognosis in glioma patients. Med Oncol *31*, 911.

Tan, W., Li, Y., Lim, S.G., and Tan, T.M. (2014). miR-106b-25/miR-17-92 clusters: polycistrons with oncogenic roles in hepatocellular carcinoma. World J Gastroenterol *20*, 5962-5972.

Tan, X., Wang, S., Zhu, L., Wu, C., Yin, B., Zhao, J., Yuan, J., Qiang, B., and Peng, X. (2012). cAMP response element-binding protein promotes gliomagenesis by modulating the expression of oncogenic microRNA-23a. Proc Natl Acad Sci U S A *109*, 15805-15810.

Tang, H., Wang, Z., Liu, Q., Liu, X., Wu, M., and Li, G. (2014). Disturbing miR-182 and -381 inhibits BRD7 transcription and glioma growth by directly targeting LRRC4. PLoS One *9*, e84146.

Tay, Y., Rinn, J., and Pandolfi, P.P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344-352.

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. Molecular cancer therapeutics *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.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

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.

Wan, W.N., Zhang, Y.Q., Wang, X.M., Liu, Y.J., Zhang, Y.X., Que, Y.H., Zhao, W.J., and Li, P. (2014). Down-regulated miR-22 as predictive biomarkers for prognosis of epithelial ovarian cancer. Diagn Pathol *9*, 178.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M.*, et al.* (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research *38*, e178.

Wang, L., Shi, Z.M., Jiang, C.F., Liu, X., Chen, Q.D., Qian, X., Li, D.M., Ge, X., Wang, X.F., Liu, L.Z.*, et al.* (2014). MiR-143 acts as a tumor suppressor by targeting N-RAS and enhances temozolomide-induced apoptosis in glioma. Oncotarget *5*, 5416-5427.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572-1573.

Wong, C.K., Vaske, C.J., Ng, S., Sanborn, J.Z., Benz, S.C., Haussler, D., and Stuart, J.M. (2013). The UCSC Interaction Browser: multidimensional data views in pathway context. Nucleic acids research *41*, W218-224.

Xi, R., Hadjipanayis, A.G., Luquette, L.J., Kim, T.M., Lee, E., Zhang, J., Johnson, M.D., Muzny, D.M., Wheeler, D.A., Gibbs, R.A.*, et al.* (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proceedings of the National Academy of Sciences of the United States of America *108*, E1128-1136.

Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L.*, et al.* (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell *153*, 919-929.

Ying, Z., Li, Y., Wu, J., Zhu, X., Yang, Y., Tian, H., Li, W., Hu, B., Cheng, S.Y., and Li, M. (2013). Loss of miR-204 expression enhances glioma migration and stem cell-like phenotype. Cancer Res *73*, 990-999.

Zhou, J., Wang, W., Gao, Z., Peng, X., Chen, X., Chen, W., Xu, W., Xu, H., Lin, M.C., and Jiang, S. (2013). MicroRNA-155 promotes glioma cell proliferation via the regulation of MXI1. PLoS One *8*, e83055.

Table of Contents

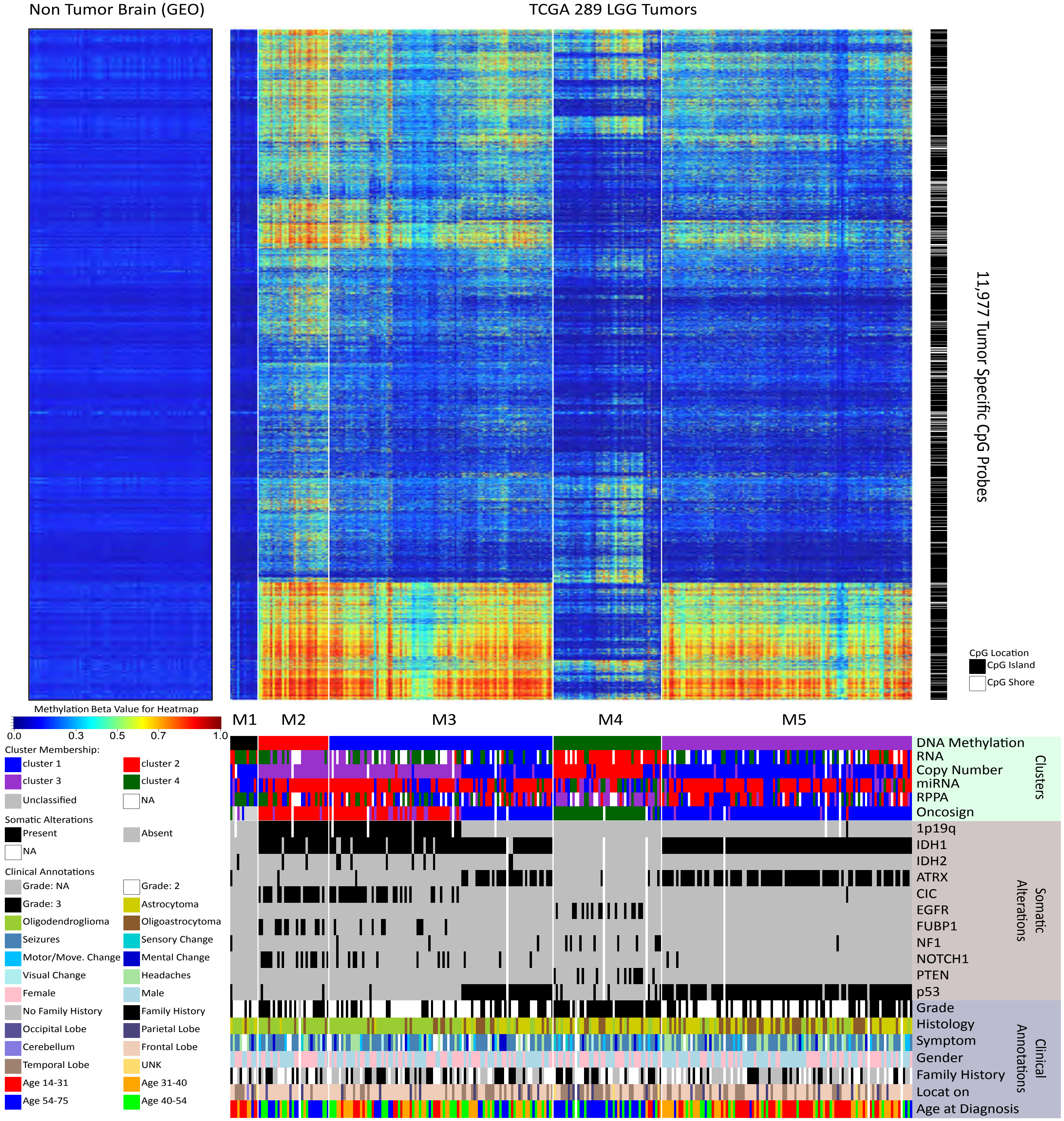


Figure S1A. Unsupervised consensus clustering of DNA methylation TCGA LGG data. Unsupervised consensus clustering was performed using 11,977 Infinium HM450 DNA methylation probes whose DNA methylation beta values were defined as tumor specific. DNA Methylation clusters are distinguished with a color code at the bottom of the panel. Each sample within each DNA Methylation cluster are color-labeled as described in the key for its gene expression, copy number, miRNA, RPPA and OncoSign cluster. Somatic mutation status of genes enriched within one or more clusters are indicated by black squares; gray squares indicate absence of mutations; and white squares indicate that the gene was not screened in the sample. Each row represents a probe; each column represents a sample. The level of DNA methylation (beta value) for each probe, in each sample, is represented by using a color scale as shown in the legend; white indicates missing data. Non-tumor brain samples profiled using the same platform were downloaded and used in this study (n=77). The non-tumor brain samples did not contribute to the unsupervised clustering.

Back to Table of Contents

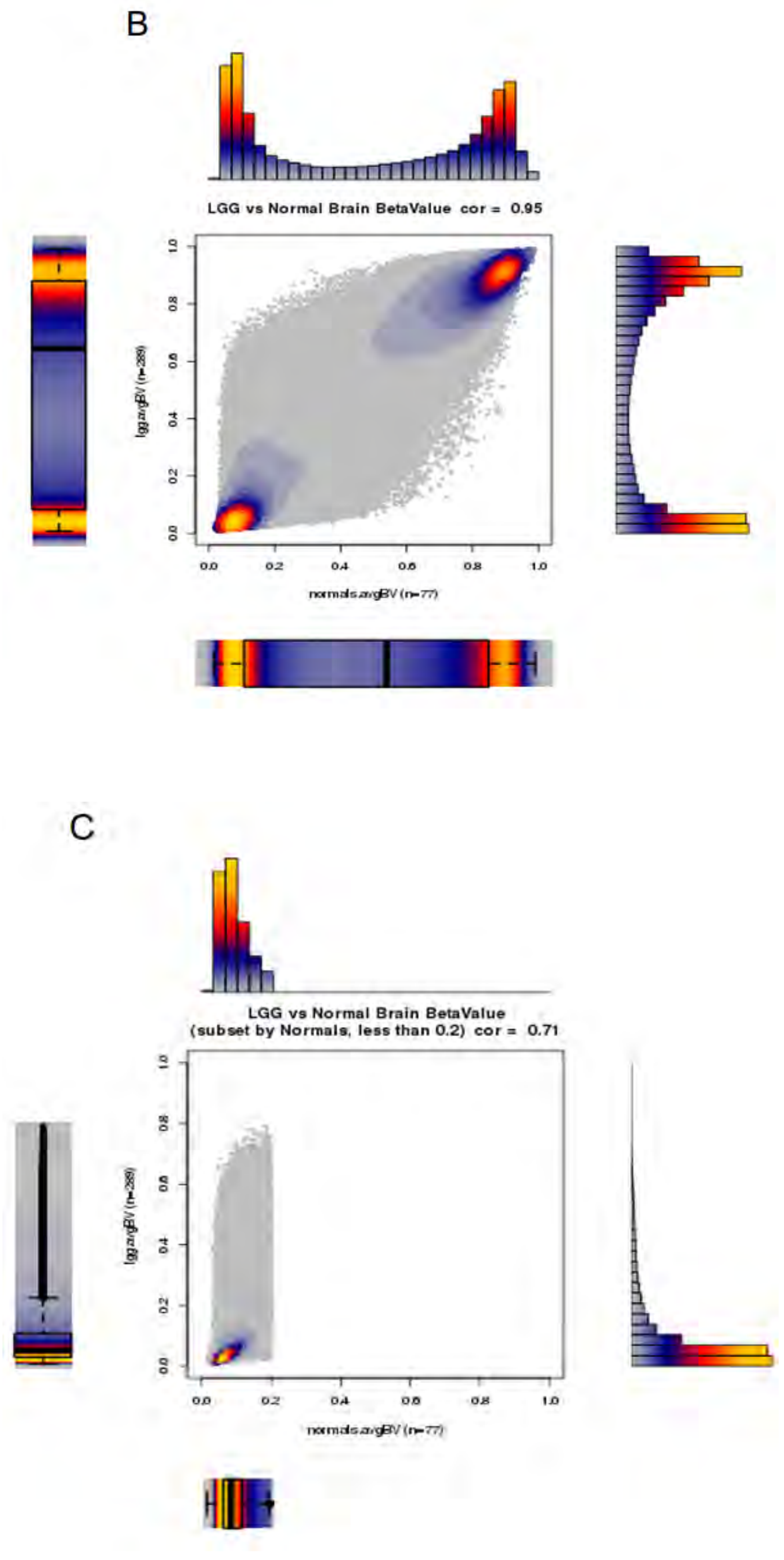


Figure S1B/C. Scatter plots of HM450 probes between the average beta-values for all 289 TCGA LGG samples (y-axis) vs 77 non-tumor brain samples (x-axis). Boxplots and frequency distribution are shown on each axes to illustrate the overall distribution of the data. Colors are shown to highlight the distribution of the data. Grey indicates low density, while orange indicates high density. B. Shows all the data. C. filtered data, showing only the tumor specific probes (non-tumor brain beta values <=0.2).

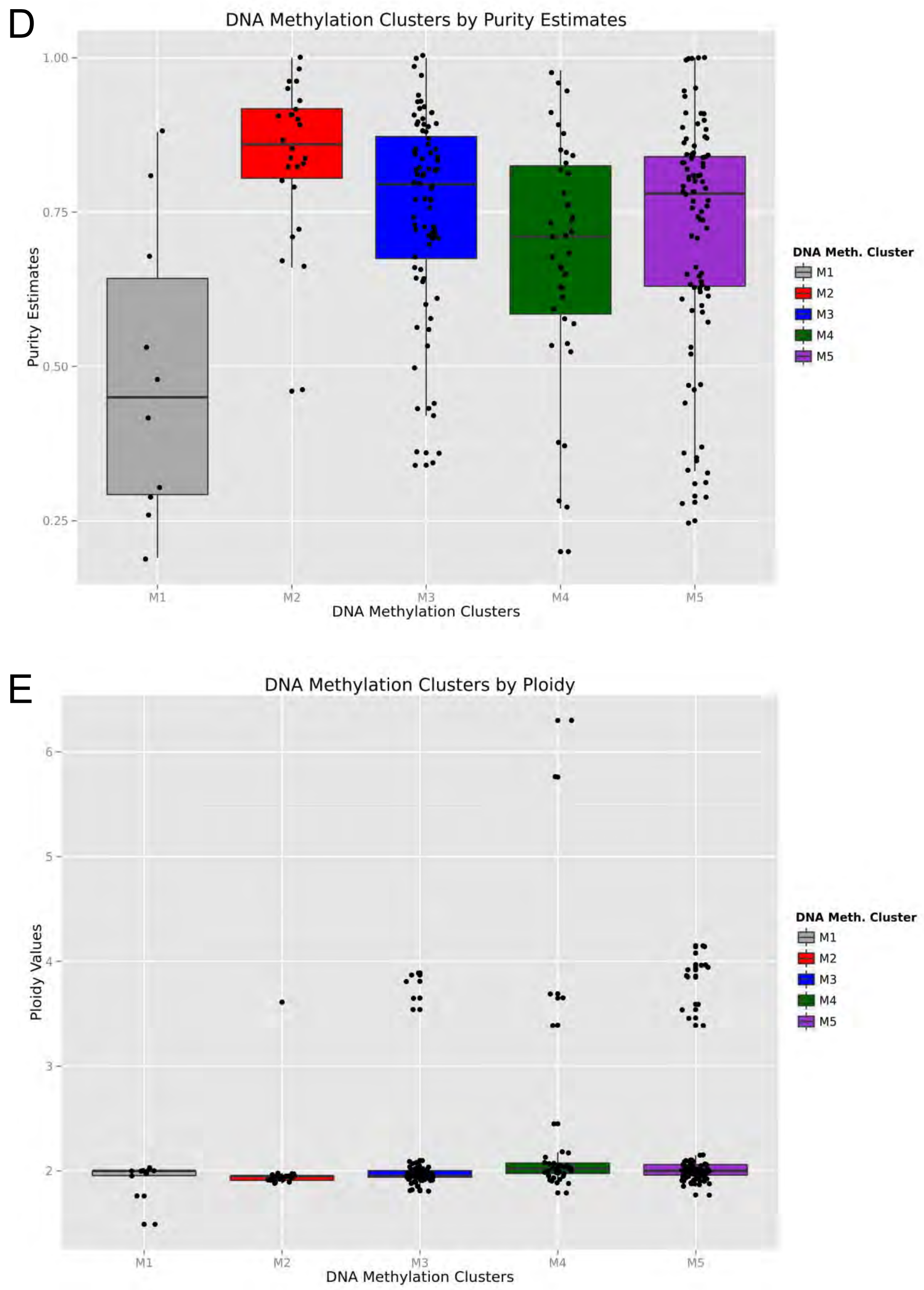


Figure S1D/E. Distribution of purity and ploidy counts by DNA methylation clusters. D. distribution of purity values by clusters. E. Distribution of ploidy by clusters.

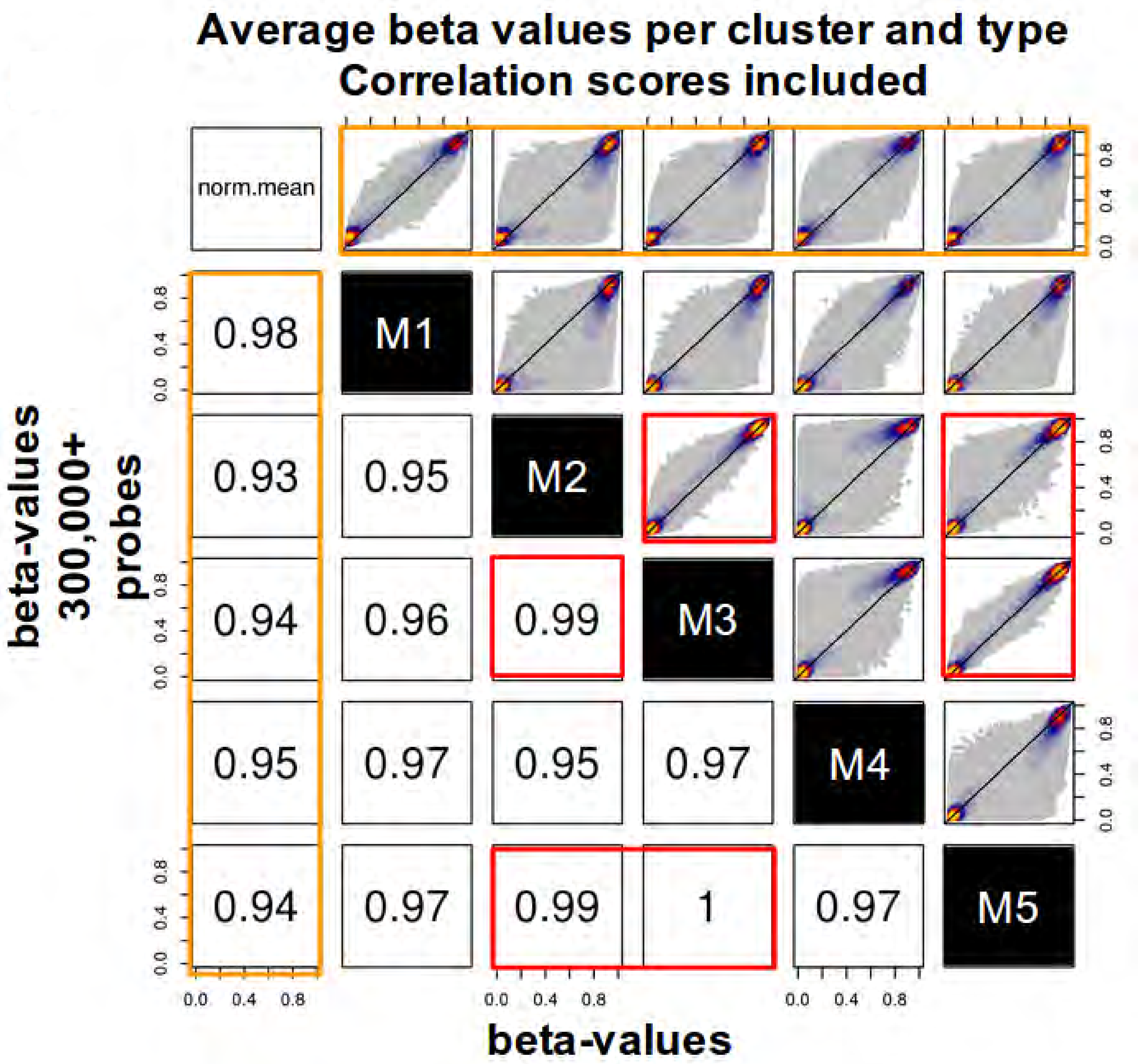


Figure S2A. Correlation plots showing beta values for more than 300,000 probes averaged for non-tumor, cluster 1-5. The top right panel shows the scatter plots while the lower left panel shows the correlation values by rho.

Back to Table of Contents

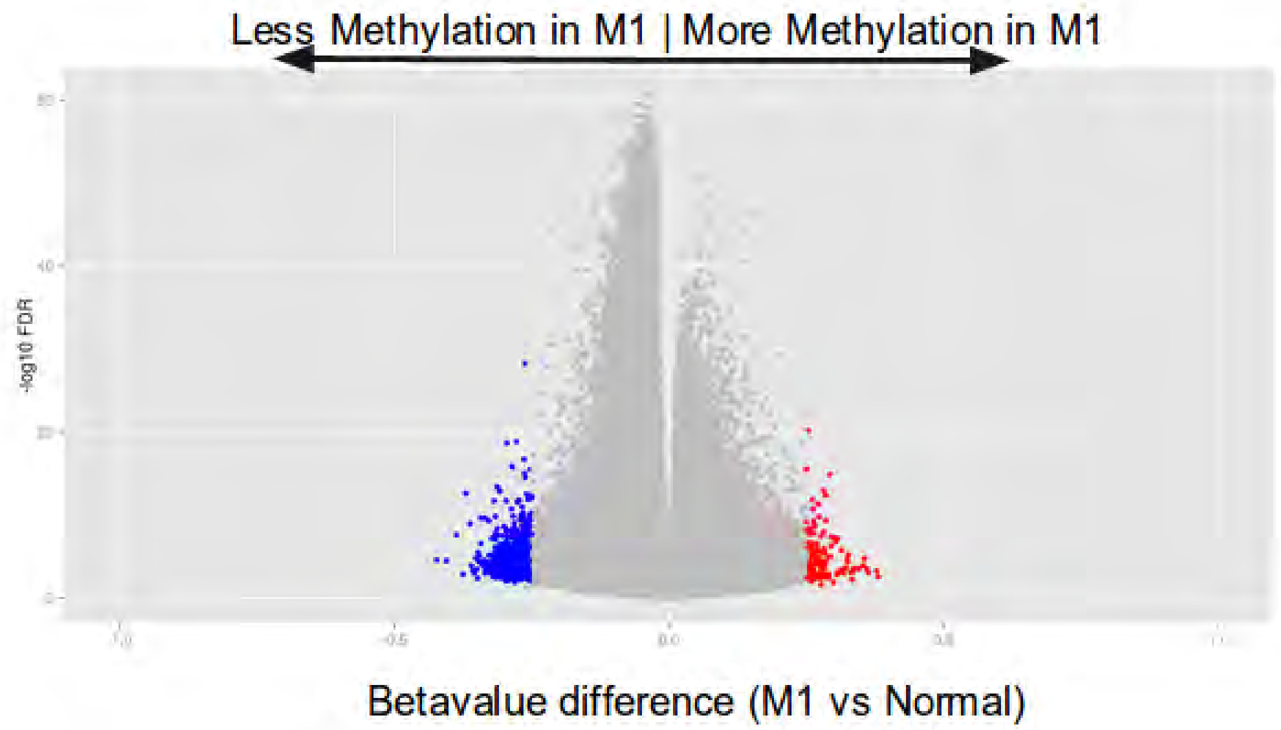


Figure S2B. **Volcano plot between cluster 1 and non-tumor brain samples**. X-axis is the beta value difference. Y-axis is –log10 of the FDR values. Colors indicate significantly different ($p<0.05$; |beta value difference| > 0.2).

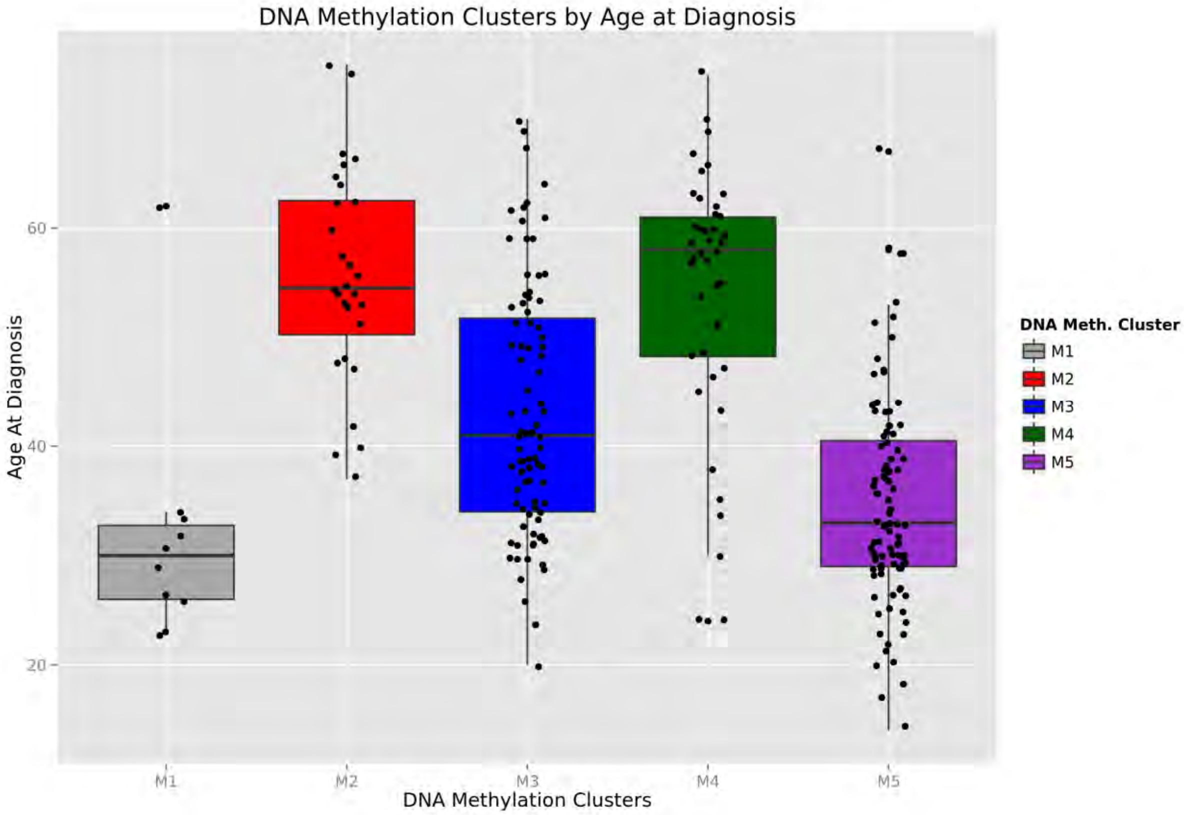


Figure S3A. **Clinical features describing each clusters**. Age at diagnosis distribution by clusters. Y-axis indicates age at diagnosis. Clusters are colored and each spot indicates a sample.

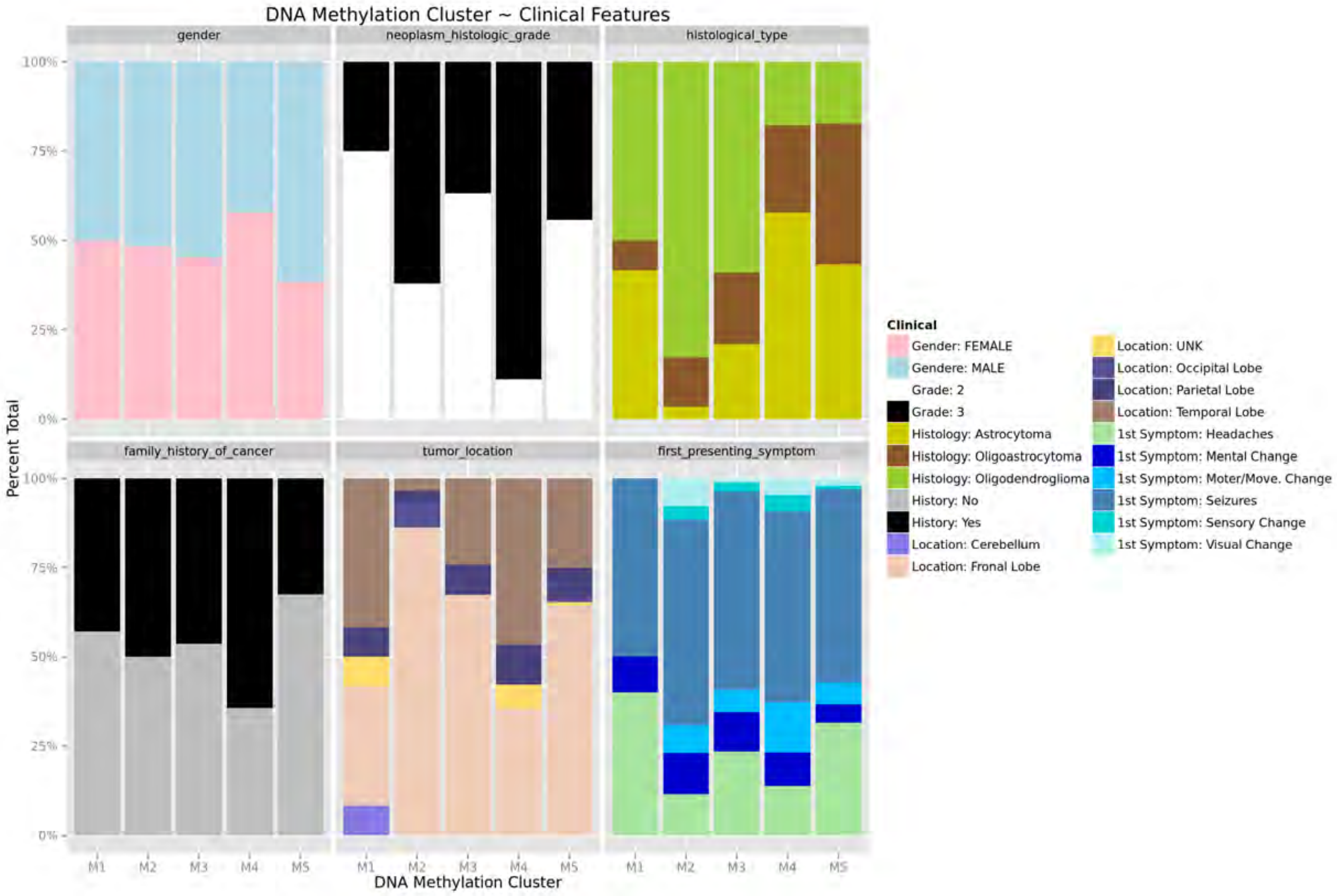


Figure S3B. Clinical features describing each clusters. Histogram of 6 different clinical features found to be enriched for one or more cluster. Legend on the right is provided to aid in the distribution. Y-axis is percent total, x-axis is by cluster.

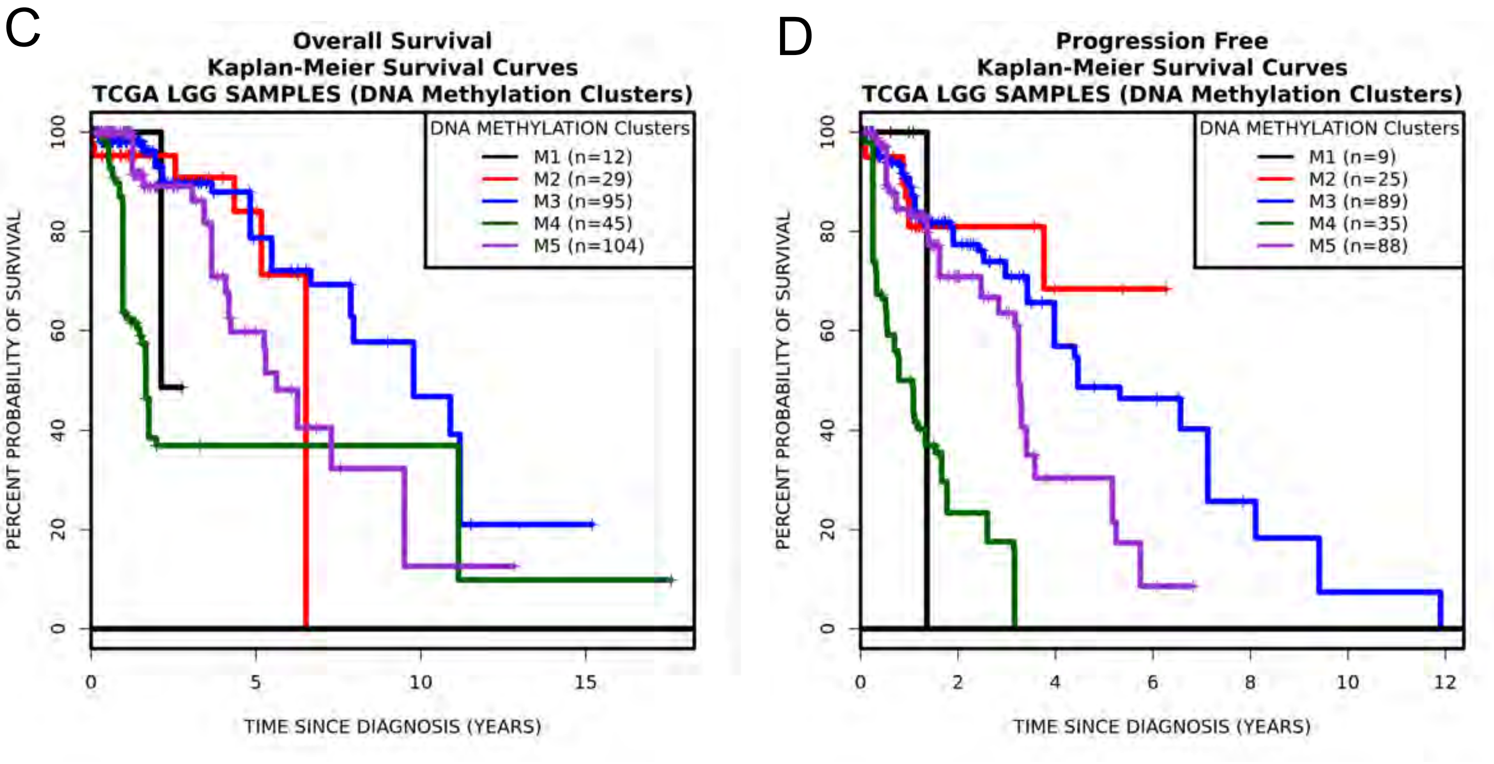


Figure S3C/D. Kaplan-meier survival curves by cluster distribution. Y-axis is percent probability of survival; x-axis is years (time since diagnosis). C, Overall survival. D, Progression free survival.

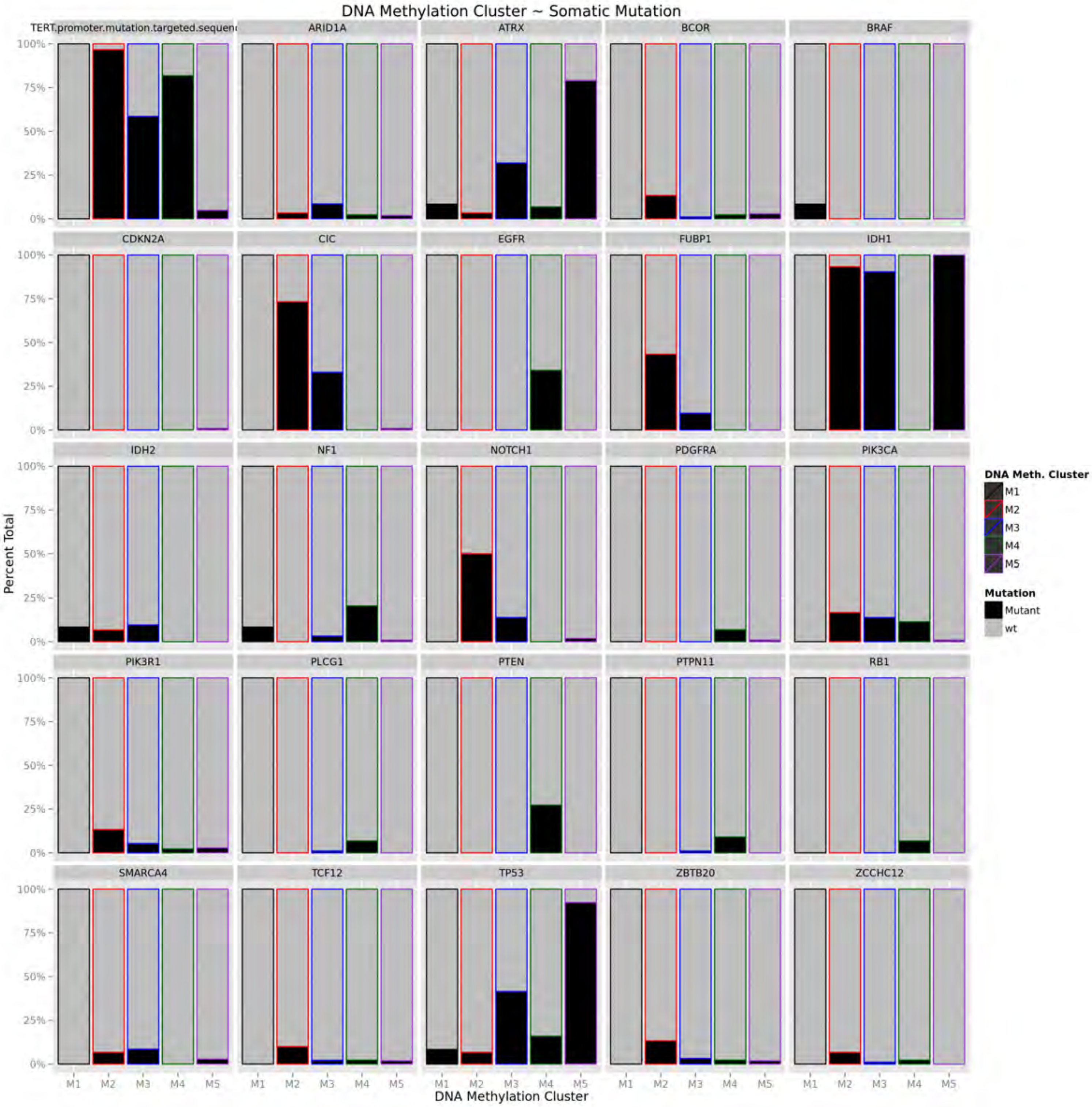


Figure S4. Somatic alterations associated with one or more DNA methylation cluster. Black indicates mutation and grey indicates wildtype. Y-axis is percent total and x-axis is cluster membership. Each panel highlights a specific gene mutation. Panels are labeled by gene name.

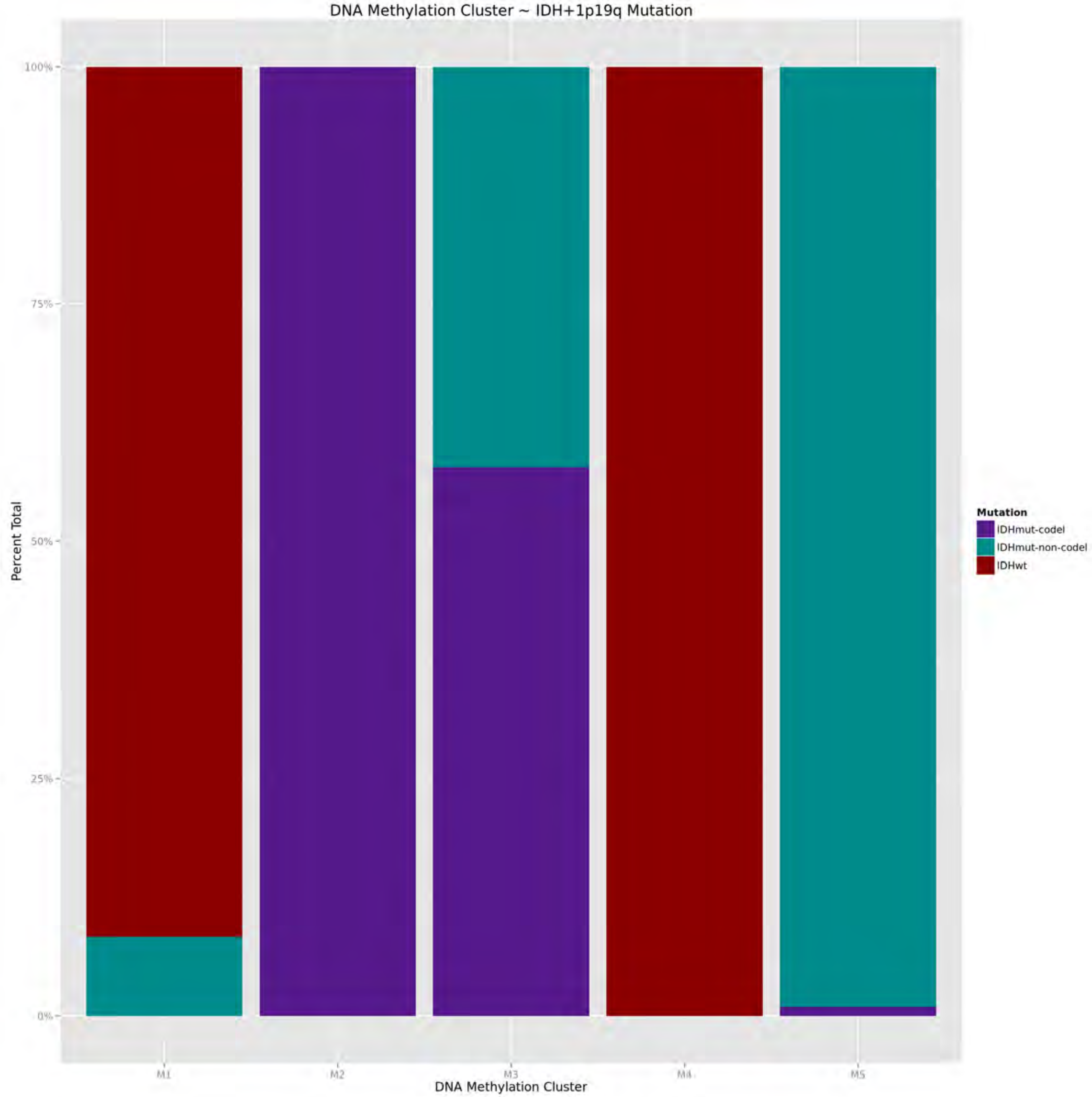


Figure S5. Histogram distribution of IDH1/2 mutation and 1p19q deletion status by clusters. Y- axis is percent total. X-axis is cluster membership.

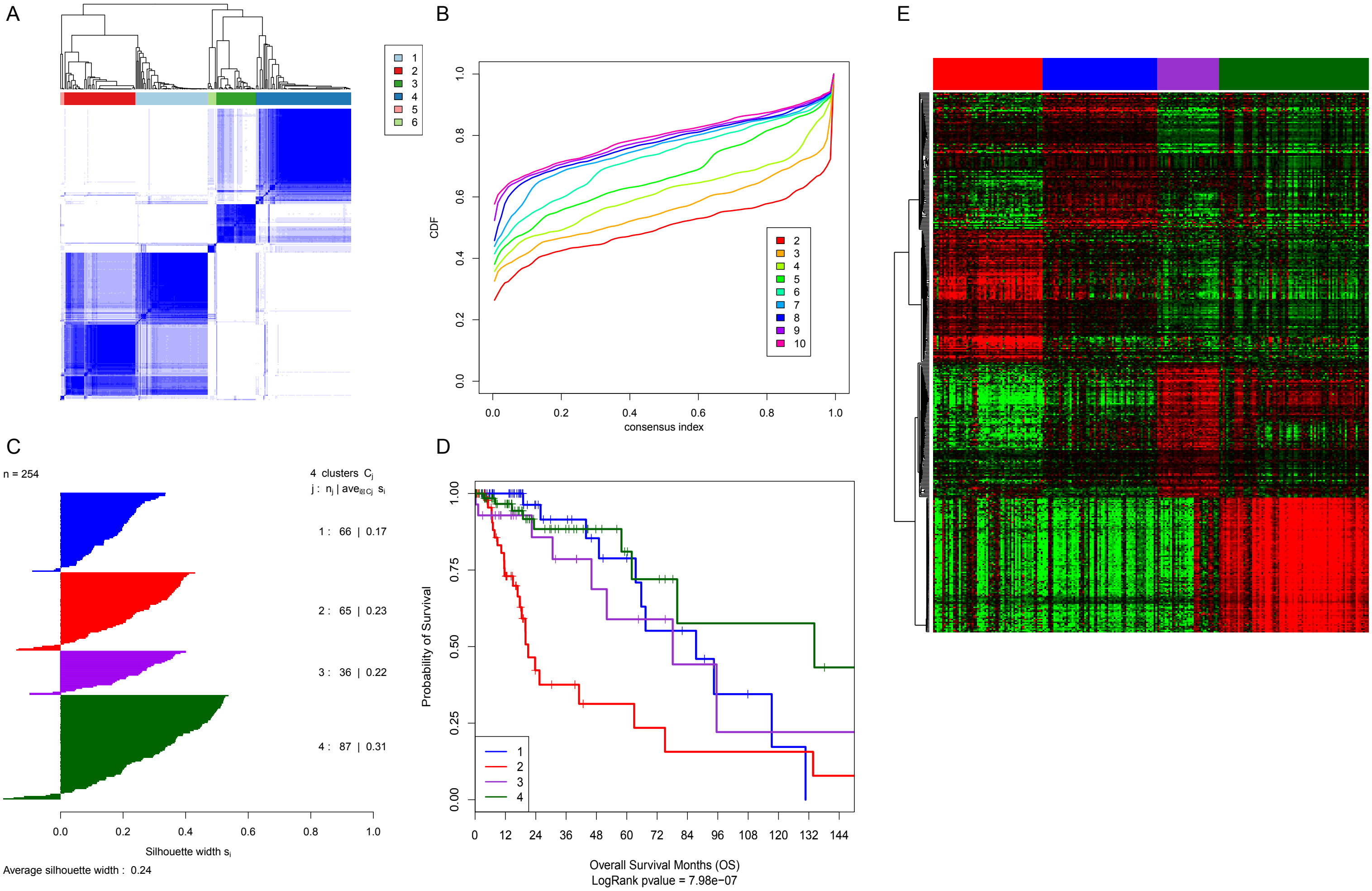


Figure S6. mRNA expression based clustering. A) Consensus clustering at K=6 using 1500 genes. B) CDF plot for K2-K10. C) Silhouette widths of the 4 RNA subtypes. D) Overall Survival for “core” RNA subtype members. E) Heatmap of the 100 most highly expressed genes per subtype. Samples are ordered according to the clustering in Figure S7.

Back to Table of Contents

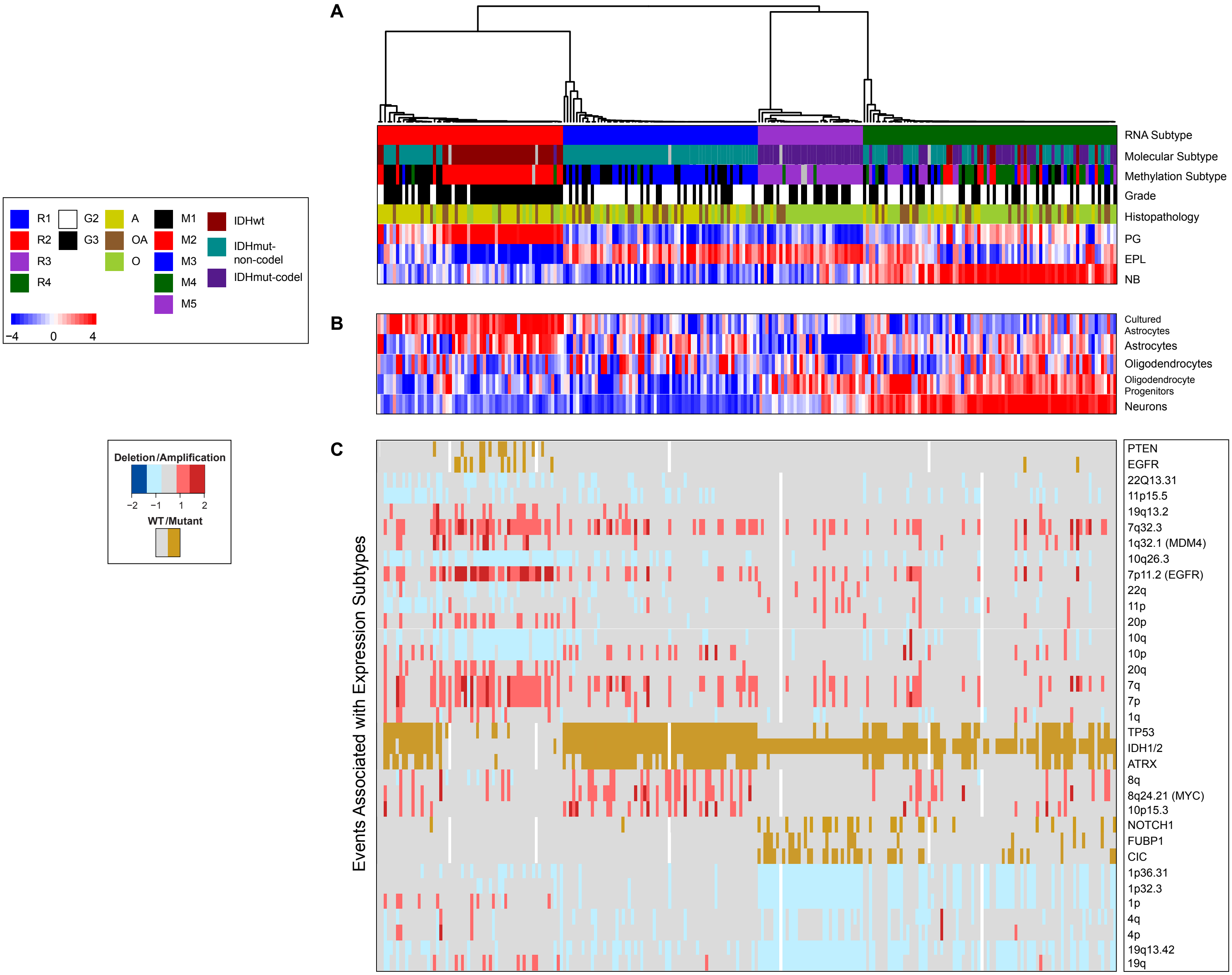


Figure S7. Genomic associations of mRNA expression clusters. A) 239 “core” RNA-seq samples were clustered using the 1500 most variably expressed genes. Single sample gene set enrichment (ssGSEA) was performed using the astrocytoma subtype signatures previously defined (Gorovets et al., 2012). B) ssGSEA using the neural ontology signatures previously defined (Cahoy et al., 2008). C) GBM transcriptional subtypes have been previously described (Verhaak et al, 2010). D) Genomic events that were significantly associated with RNA subtype (see CustomEvents comparison at : http://gdac.broadinstitute.org/runs/awg_lgg__2013_10_28/reports/). Abbreviations: A, astrocytoma; OA, oligoastrocytoma; O, oligodendroglioma; PN, proneural; N, neural; CL, classical MES, mesenchymal; PG, preglioblastoma; NB, neuroblastic; EPL, early progenitor-like.

Back to Table of Contents

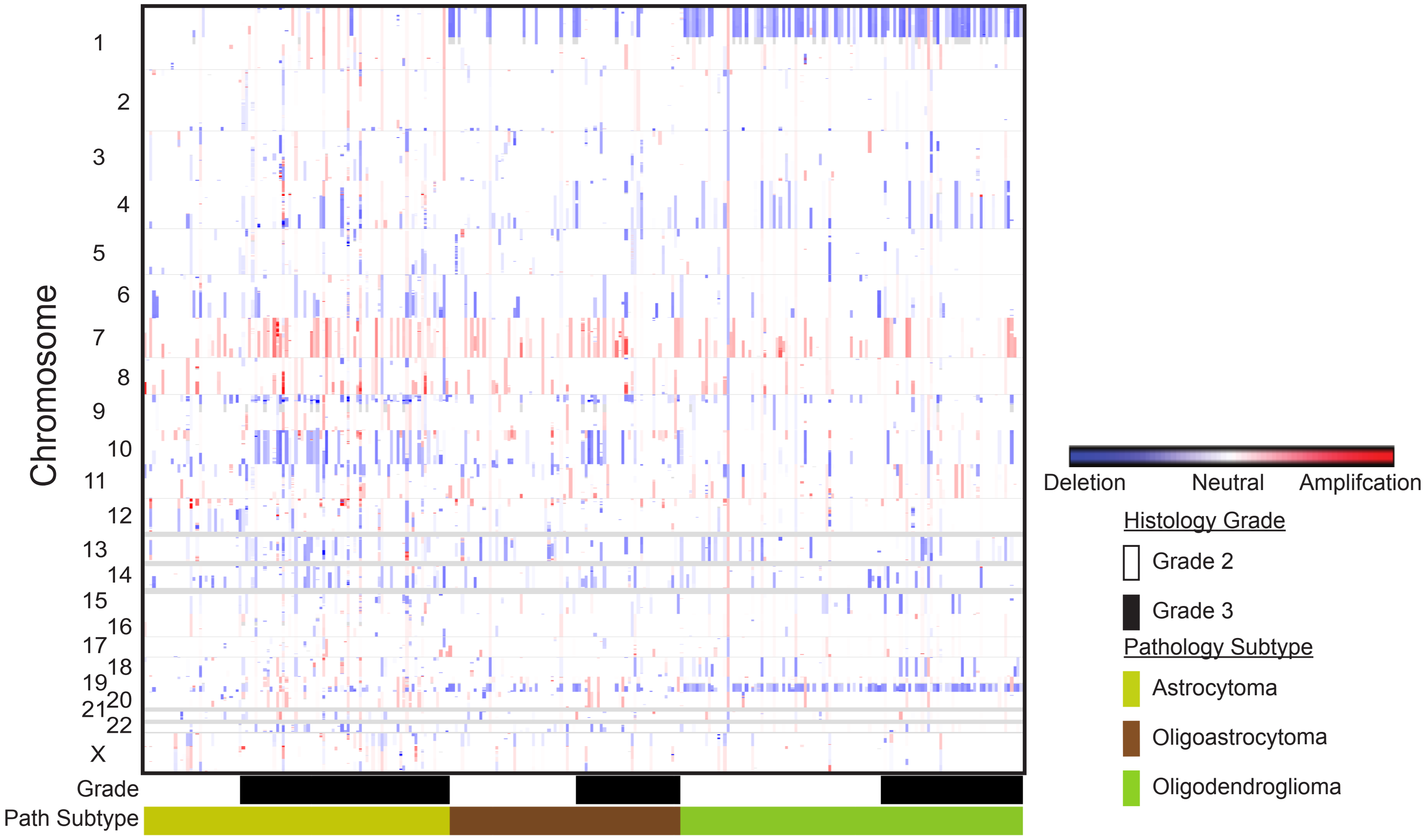


Figure S8A. DNA copy number profiles by histology and grade. The heatmap shows DNA copy number levels in tumors (horizontal axis) plotted by chromosomal location (vertical axis). Amplified regions are colored red, while deleted regions are colored blue. Horizontal side bars show pathological subtypes and tumor grades.

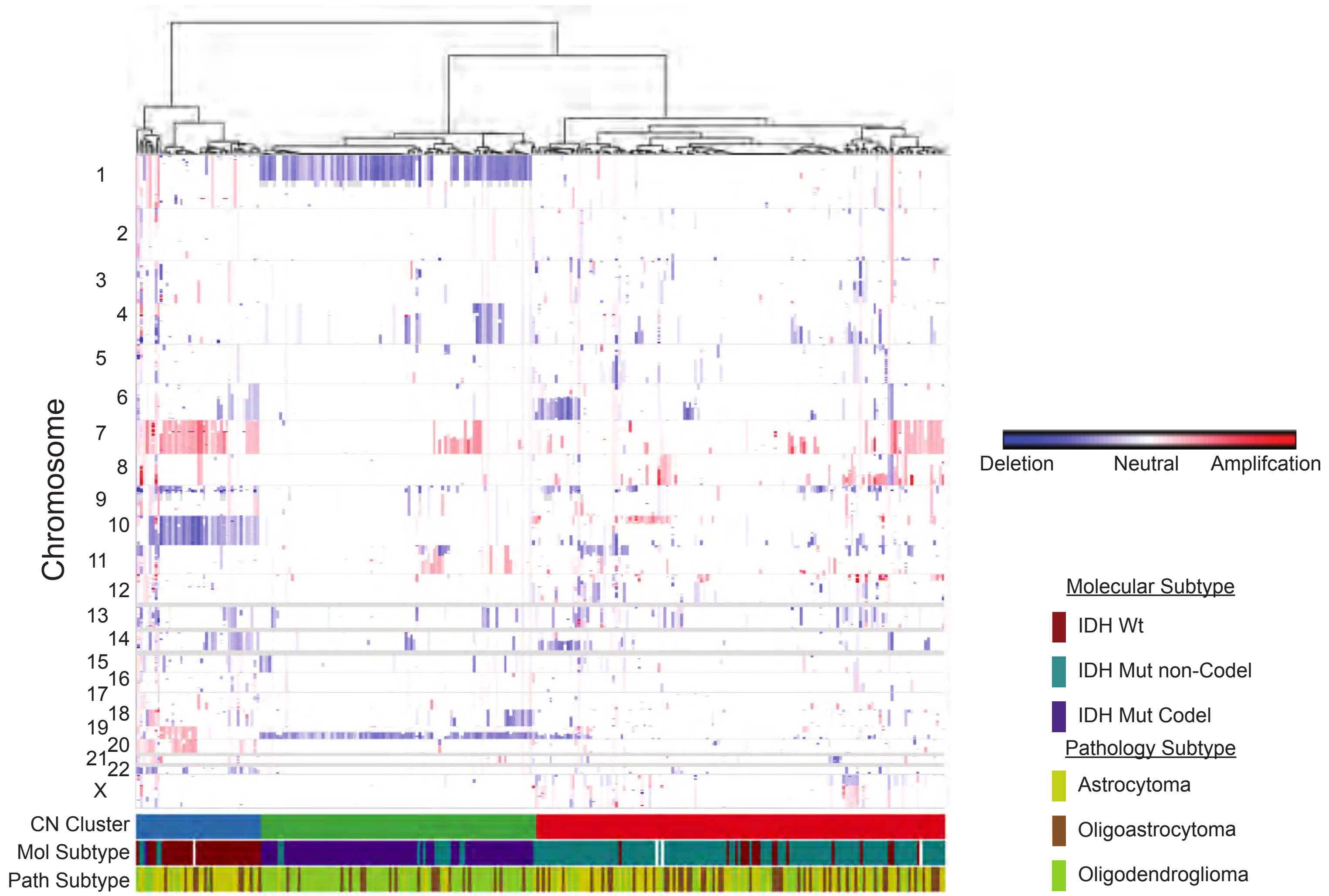


Figure S8B. Copy number based clustering of LGG tumors. Tumors were hierarchical clustered based on chromosomal arm level alterations. In the heatmap, SCNAs in tumors (horizontal axis) are plotted by chromosomal location (vertical axis). Amplified regions are colored red, while deleted regions are colored blue. Horizontal side bars show major copy number cluster groups, molecular subtypes and pathological subtypes.

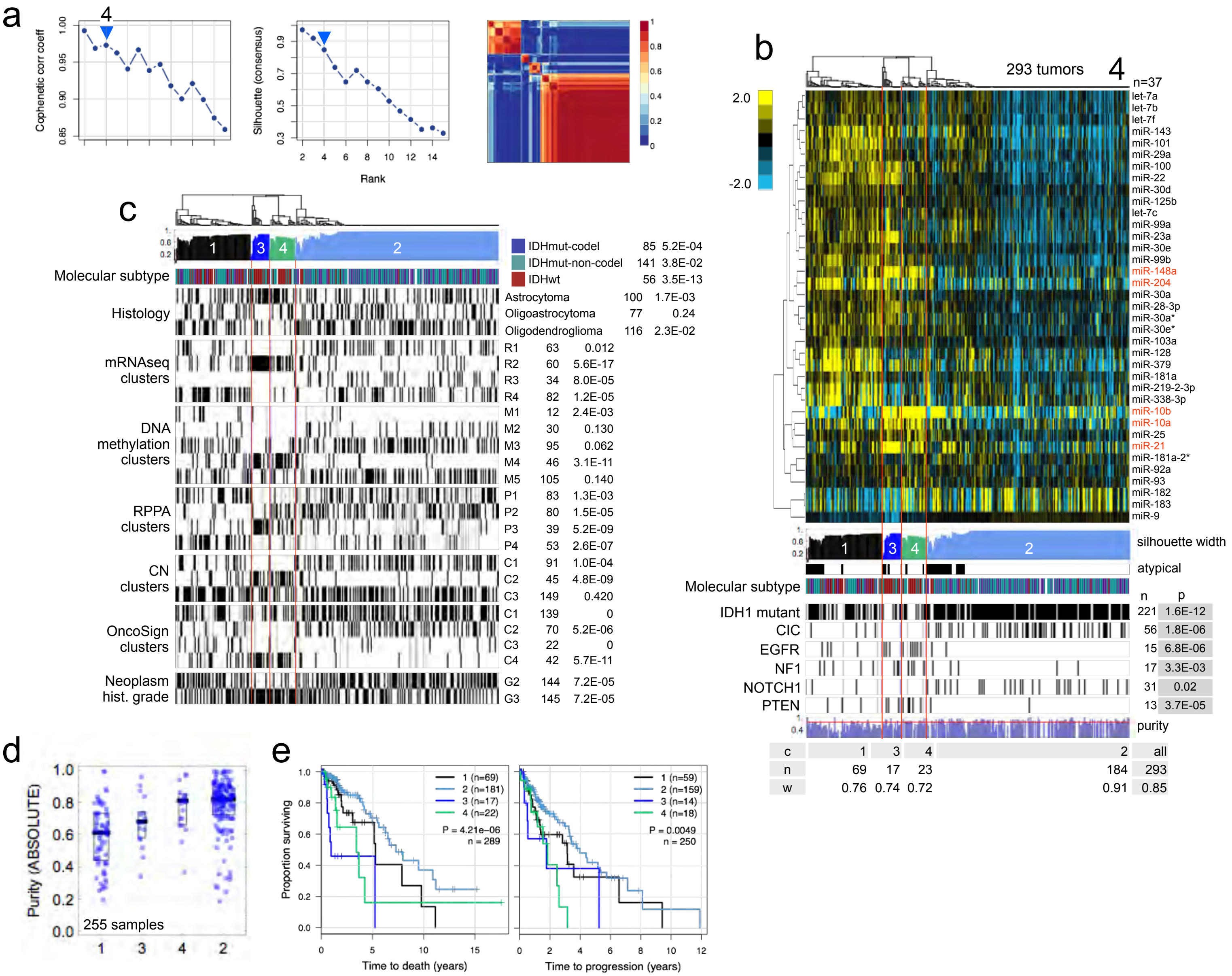


Figure S9. Unsupervised clustering of miRNA-seq 5p/3p data for 293 tumor samples. a) From the NMF consensus clustering rank survey (Gaujoux and Seoighe 2010), the cophenetic correlation coefficient and average silhouette width profiles suggest four- and seven-group solutions. The blue/red image shows the consensus membership heatmap for the four-group solution, with yellow-white indicating samples that are less ‘typical’ cluster or group members. b) A four-group NMF consensus clustering solution. miR names are from miRBase v16. Top to bottom: a normalized abundance heatmap for the 36 5p or 3p strands that were most discriminatory for NMF (i.e. miRs with the top 5% of scores in each metagene), with red text highlighting a more discriminatory subset (i.e. miRs with relatively high scores in the metagene W matrix); silhouette width profile; ‘atypical’ group members, which are samples whose width is below 0.9 of the maximum in a group; covariates with Fisher exact P-values; and a summary table of cluster number, number of samples in each cluster, and the overall average silhouette width. The scale bar shows log2 normalized abundances (i.e. reads-per-million, RPM) that are median-centred for each miR (row) in the heatmap. c) Covariates, as in (b). d) Per-group distributions of tumor sample purity as calculated by ABSOLUTE. e) Kaplan-Meier plots for overall survival and progression-free survival, with log-rank P-values.

Back to Table of Contents

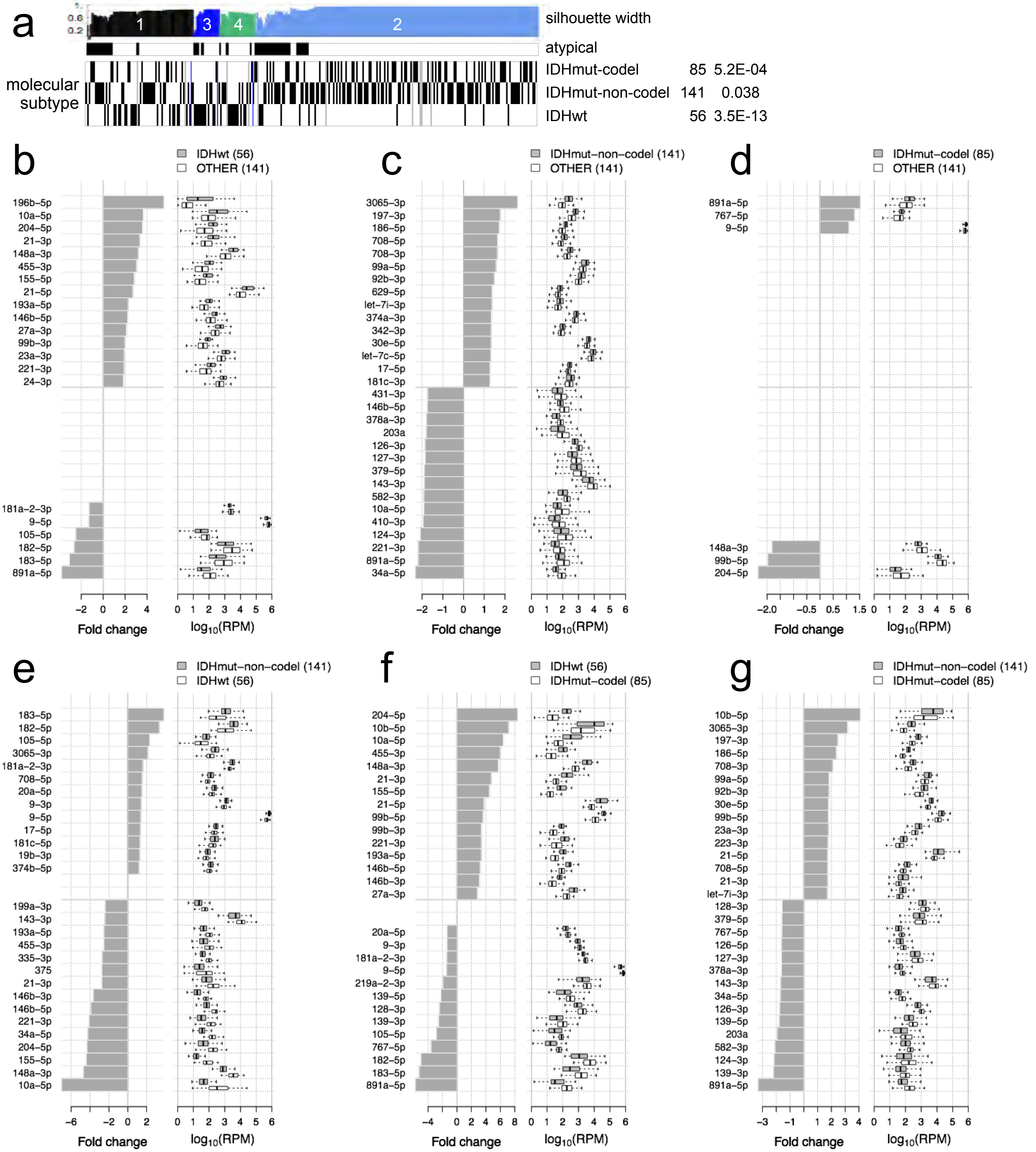


Figure S10. Differentially abundant miRs, miR-protein anticorrelations, and miRs that may influence protein abundance between subtypes. a-g) miRs that are differentially abundant between subtypes. a) Silhouette width profile and molecular subtype covariate tracks from Fig. S1. b,c,d) Comparisons are between one molecular subtype and all other samples: b) IDH wild type, c) IDH mutant/non-codeleted, d) IDH mutant with a 1p/19q co-deletion. e,f,g) Comparisons between pairs of molecular subtypes. Left: median-based fold change, linear scale. Right: Boxplots showing distributions of normalized (RPM) abundance, with black vertical lines indicating medians. Up to 15 of the largest fold changes in each direction are shown; FDR ≤ 0.05. Because miRs with higher abundance are likely more influential (Tay et al. 2014), only miRs with a mean abundance of at least 50 RPM are shown (see Tables S8,S9).

Back to Table of Contents

(255 tumours): 77 functionally validated miR-RPPA anticorrelations, FDR<0.05

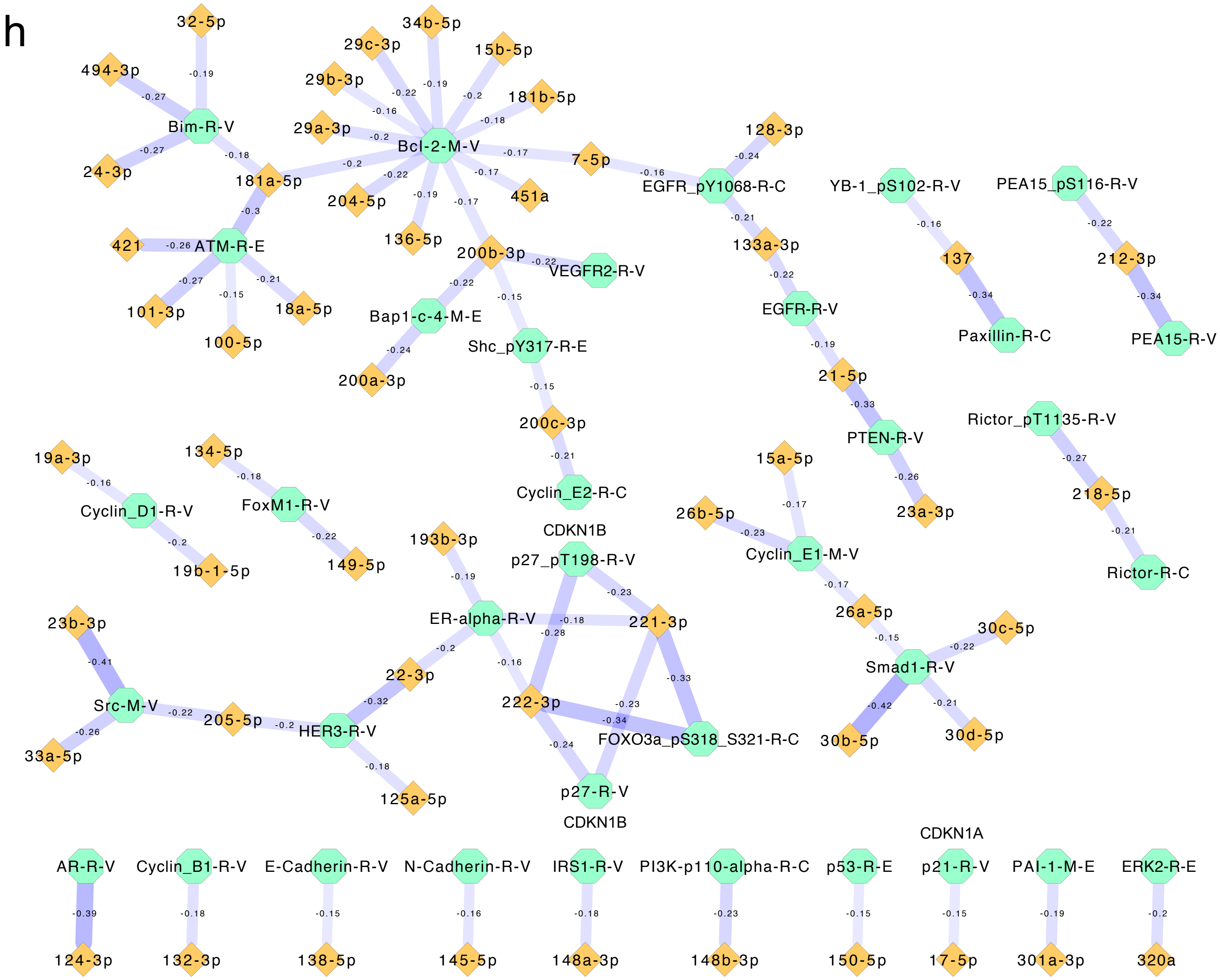


Figure S10h. Functionally validated miR-antibody anticorrelations. The network shows 77 miR-antibody inverse correlations (FDR<0.05) that are supported by functional validation publications. The correlations were calculated for 255 tumor samples from normalized abundance data for miRNA mature strands and 189 RPPA antibodies.

Back to Table of Contents

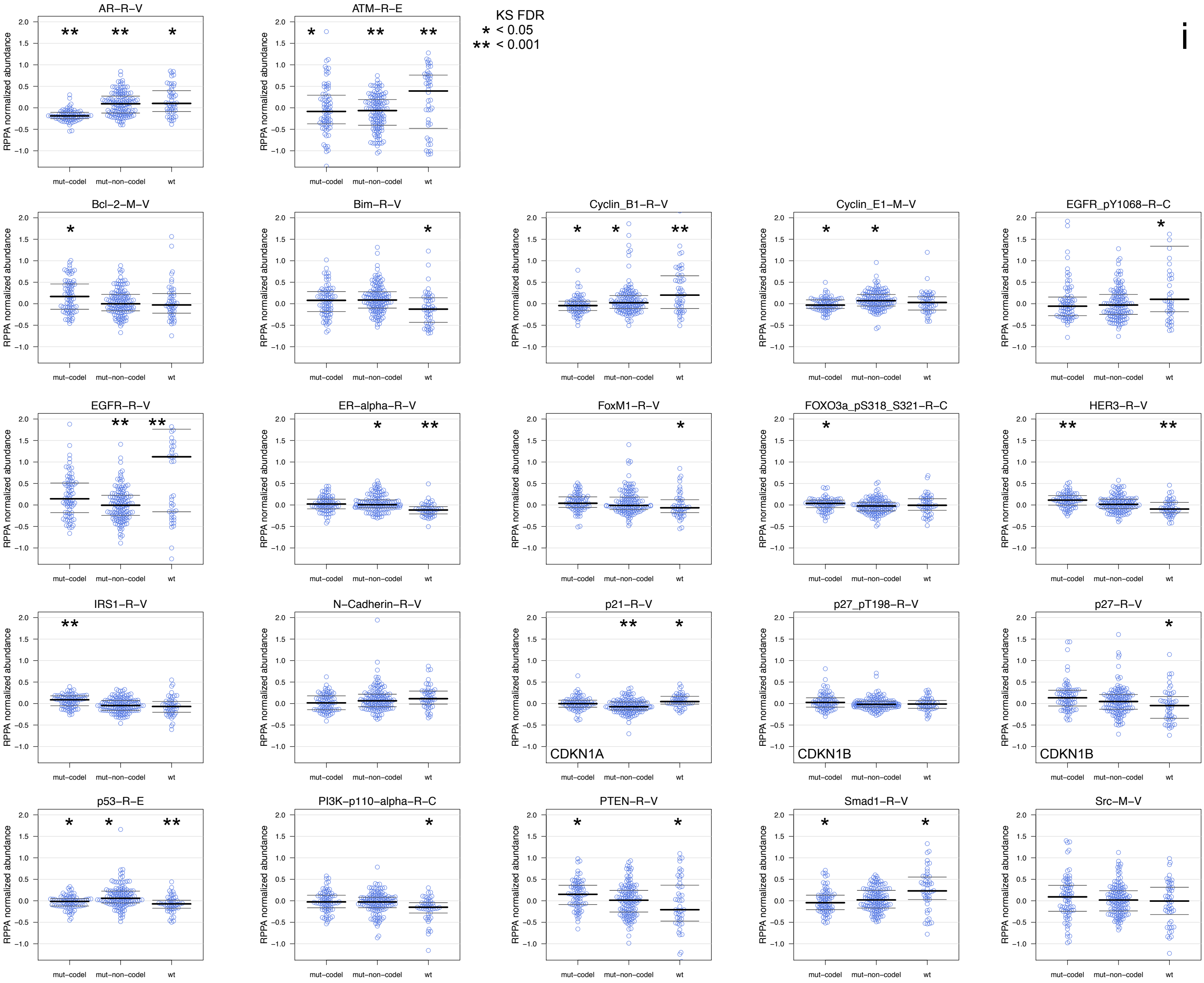


Figure S10i. Distribution between molecular subtypes of normalized abundance for 22 antibodies. These antibodies were anticorrelated to miRs that were differentially abundant between the three molecular substypes (Fig. S10b-g), and had targeting relationships with miRNAs that were supported by published functional validations (Fig. S0h). Asterisks indicate statistical significance from BH-corrected KS tests for all samples in each subtype, against all other tumor samples (n=255). For example, p21's abundance distribution was statistically different for IDHmut-non-codel samples (**, i.e. BH-corrected P-value <0.001) and for IDHwt samples (*, i.e. BH-P<0.05), but was not statistically different for IDHmut-codel.

Back to Table of Contents

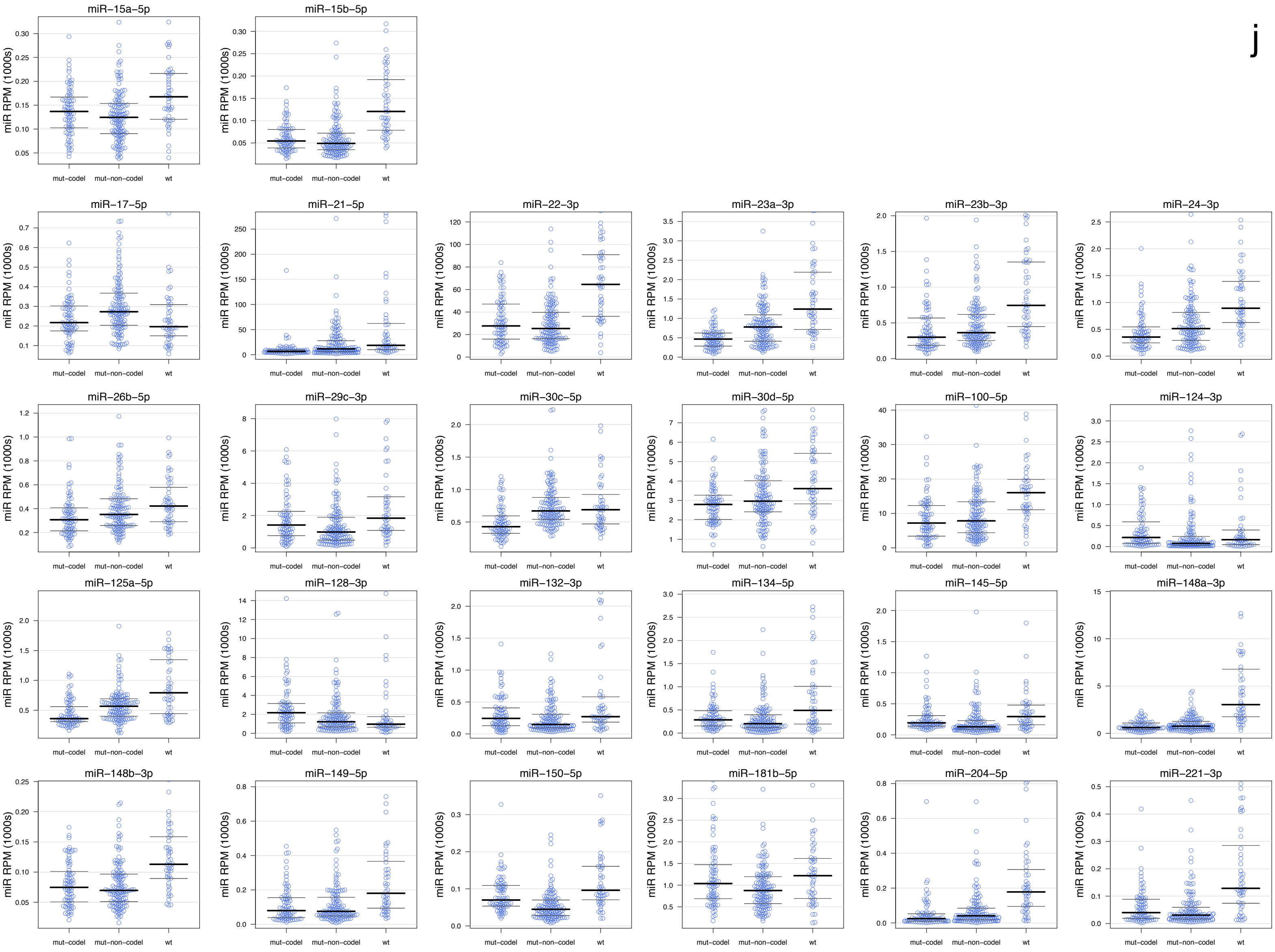


Figure S10j. Distribution between molecular subtypes of normalized (RPM) abundance for 26 miRNA mature strands. These miRs were differentially abundant between the three molecular subtypes (Fig. S10b-g), and had targeting relationships with antibodies that were supported by published functional validations (Fig. S10h).

Back to Table of Contents

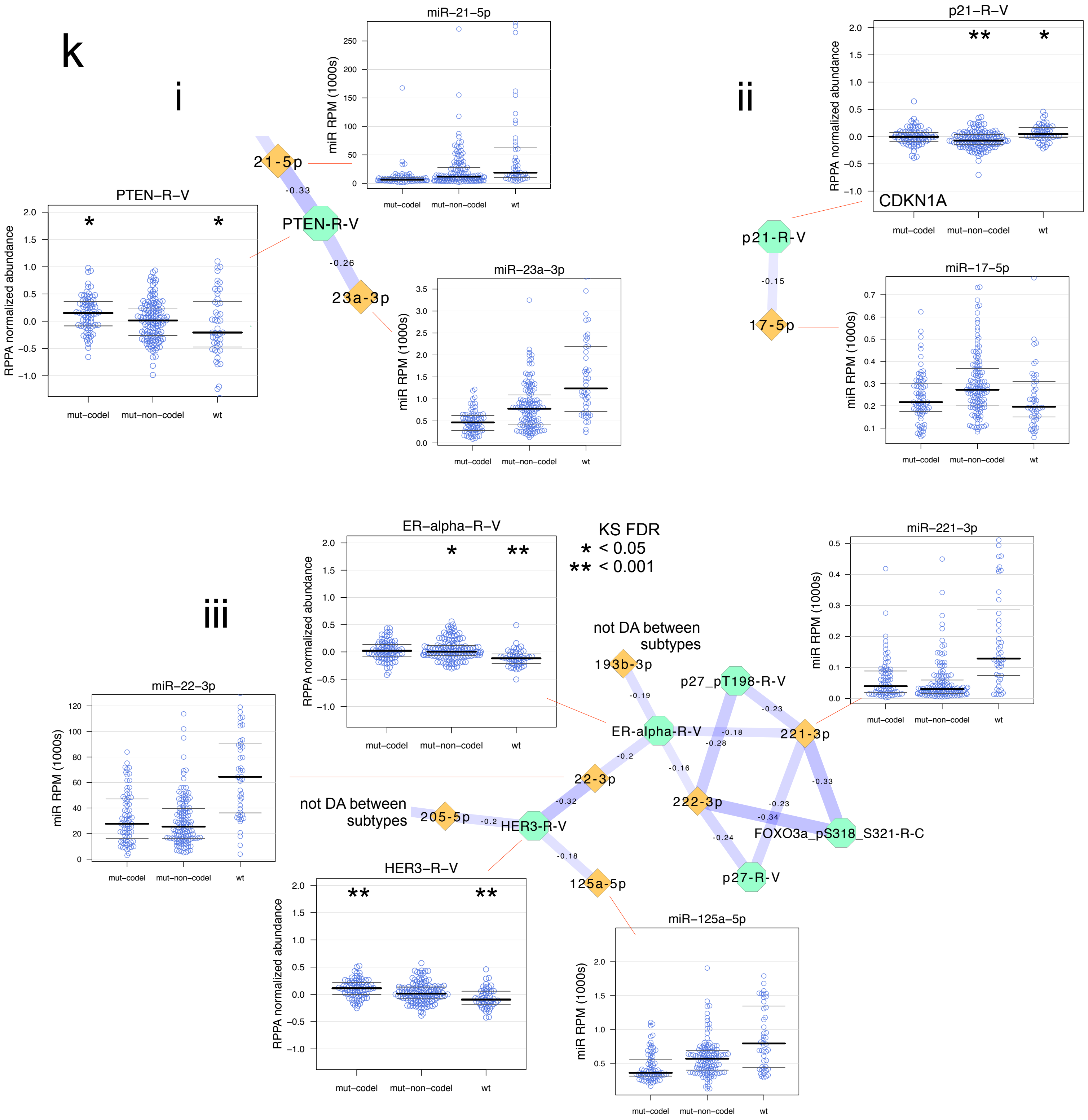


Figure S10k. miRs that may influence protein abundance across molecular subtypes. Examples from the 22 antibodies and 26 miRs from Figures S10h-j. These miRs were differentially abundant between the three molecular subtypes, and had targeting relationships with RPPA antibodies that were supported by published functional validations. i) PTEN, miR-21-5p and miR- 23a-3p. ii) p21 (CDKN1A) and miR-17-5p. iii) HER3, miR-22-3p and 125a-5p, and ER-alpha, miR- 22-3p and 221-3p.

Back to Table of Contents

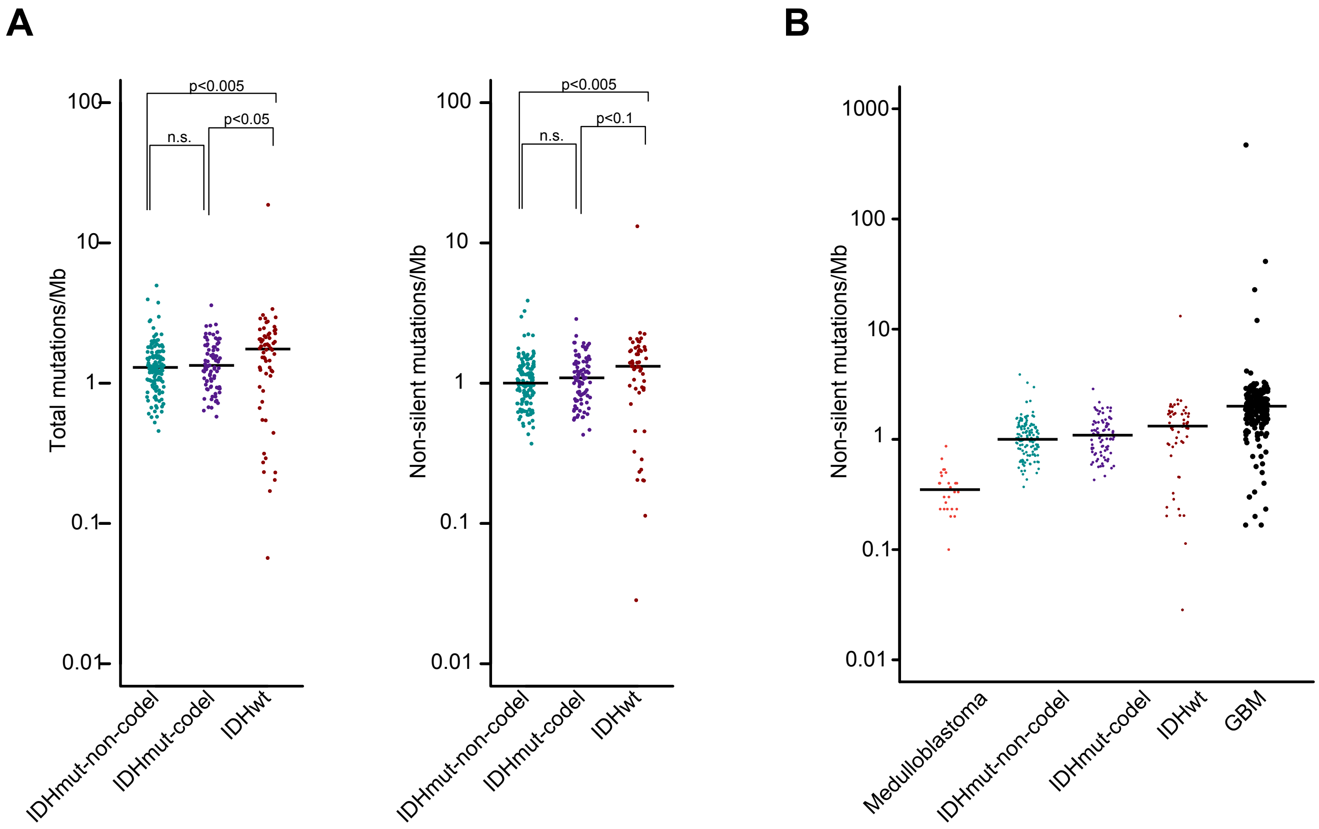


Figure S11A. Distribution of total and non-silent mutation rates stratified by molecular subtype based on calls made by MuTect (Broad Institute). Each dot represents the mutation rate for a single sample. The median for each set is indicated by a black horizontal bar. Mutation rates in LGG are lower in samples with IDH mutation (total rate median,1.3 mutation per Mb for codel and non-codel samples; non-silent rate median, 1.1 and 1 per Mb for codel and non-codel samples, respectively) than in IDHwt samples (median total rate, 1.75 per Mb; non-silent rate, 1.3 per Mb). B. LGG rates are higher than medulloblastoma (median rate, 0.35 per Mb) but lower than GBM (median rate, 2 per Mb) and intermediate in the spectrum of mutation frequencies for reported TCGA malignancies (median ranges, 0.17 – 13.2 per Mb) (data from Lawrence et al, 2013). All p-values calculated with the Wilcoxon rank-sum test.

Back to Table of Contents

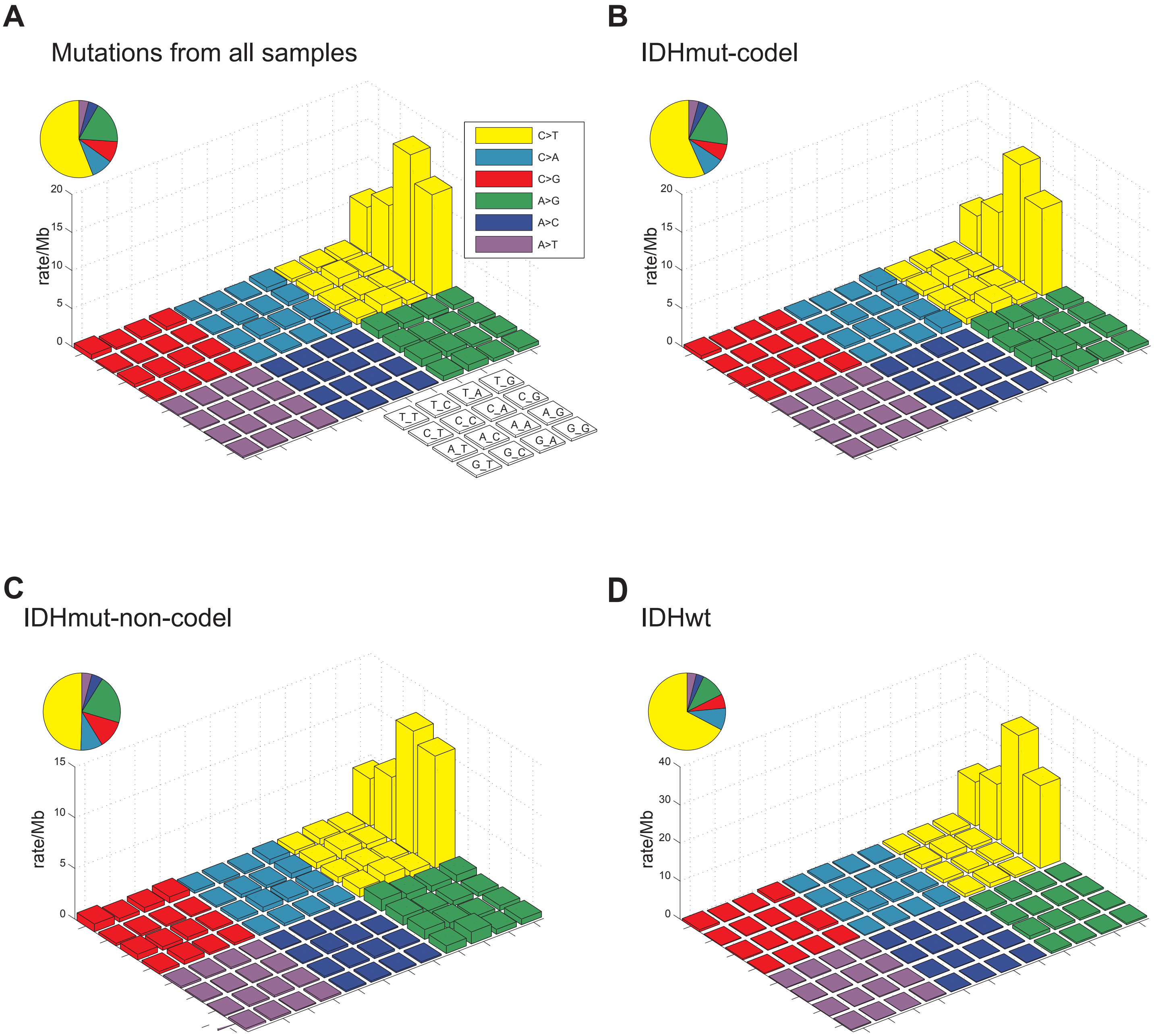


Figure S12. Mutation spectra for all samples and differentiated by molecular subtype based on MuTect calls (Broad Institute). Each column represents the mutation rate per Mb of a specific mutation in all samples (A) or the three molecular subtypes (B-D). Substitution types are separated by color, and position within the grid corresponds to mutation context 5'-base - mutated base - 3' - base as indicated in the legend in panel A.

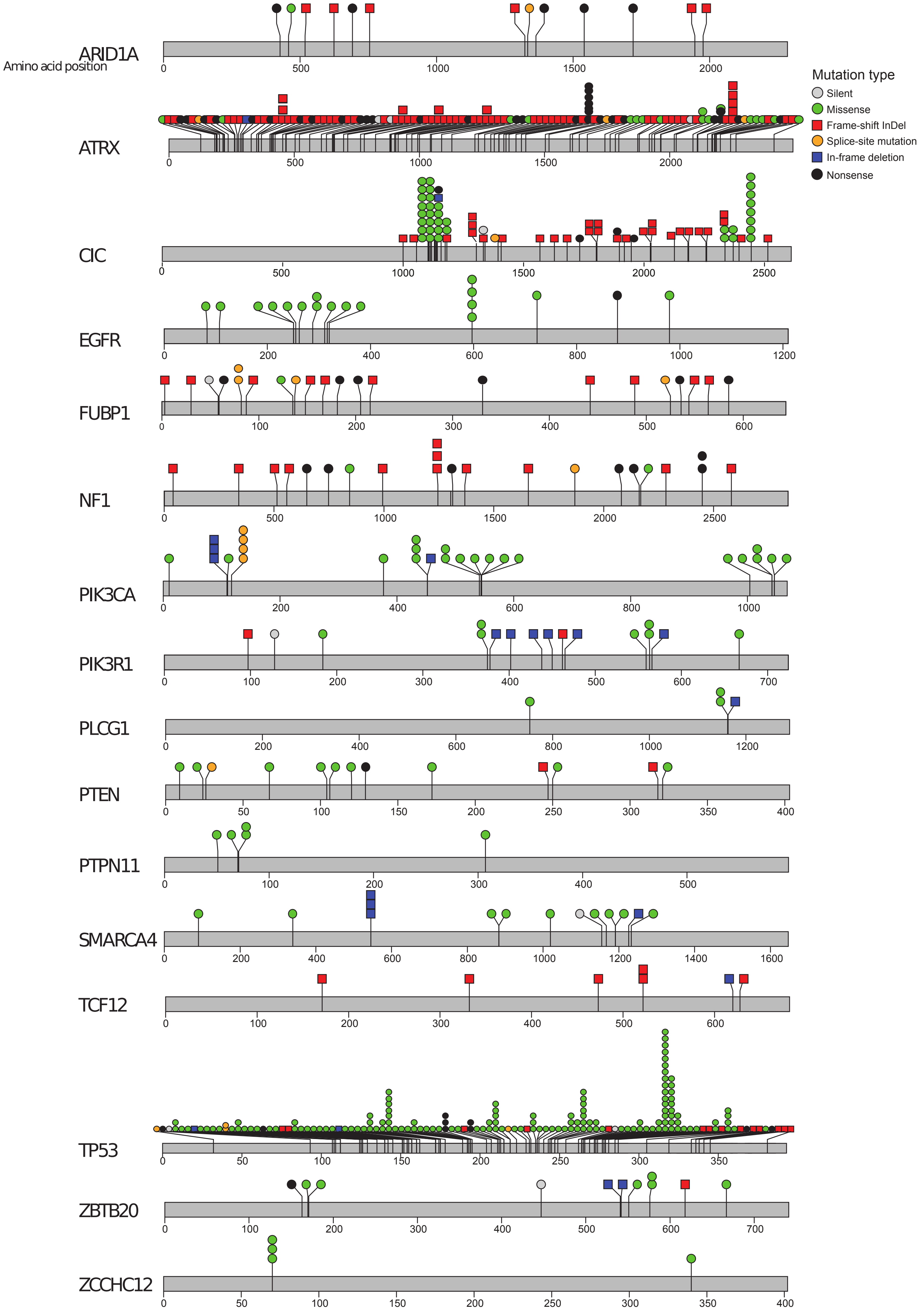


Figure S13. Distribution of mutations in protein products of significantly mutated genes. Each mutation is represented by a filled circle or square. Types of mutations are color-coded as indicated in the legend.

Back to Table of Contents

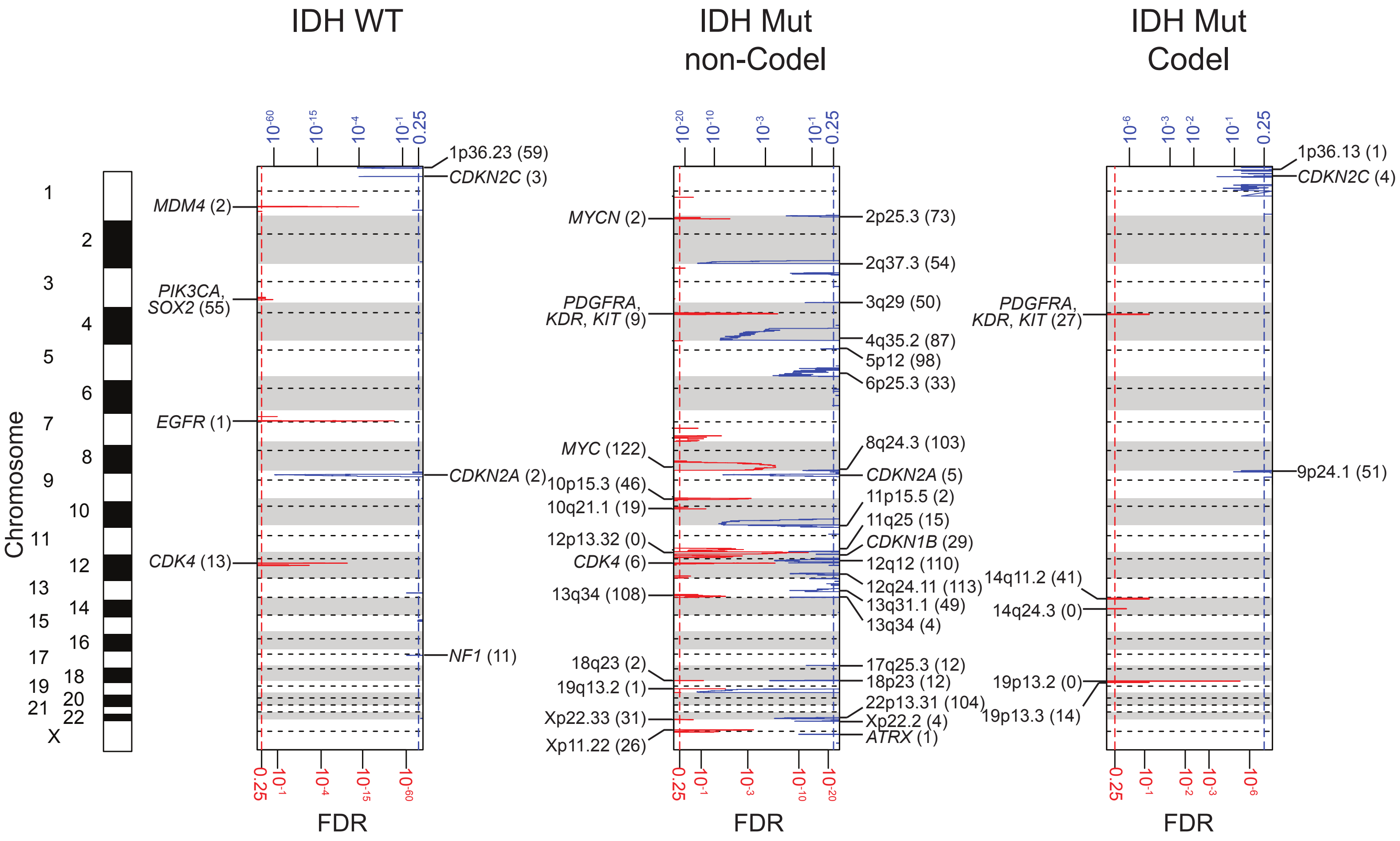


Figure S14A. GISTIC 2.0 analyses of three molecular subtypes. Chromosomal locations of peaks of significantly recurring focal amplifications (red) and deletions (blue) are plotted by false discovery rates. Annotated peaks have a FDR < .2 and encompass 125 or less genes. Peaks are annotated with candidate driver oncogenes or tumor suppressors or by cytoband. The number of genes within each peak are shown next driver gene or cytobands.

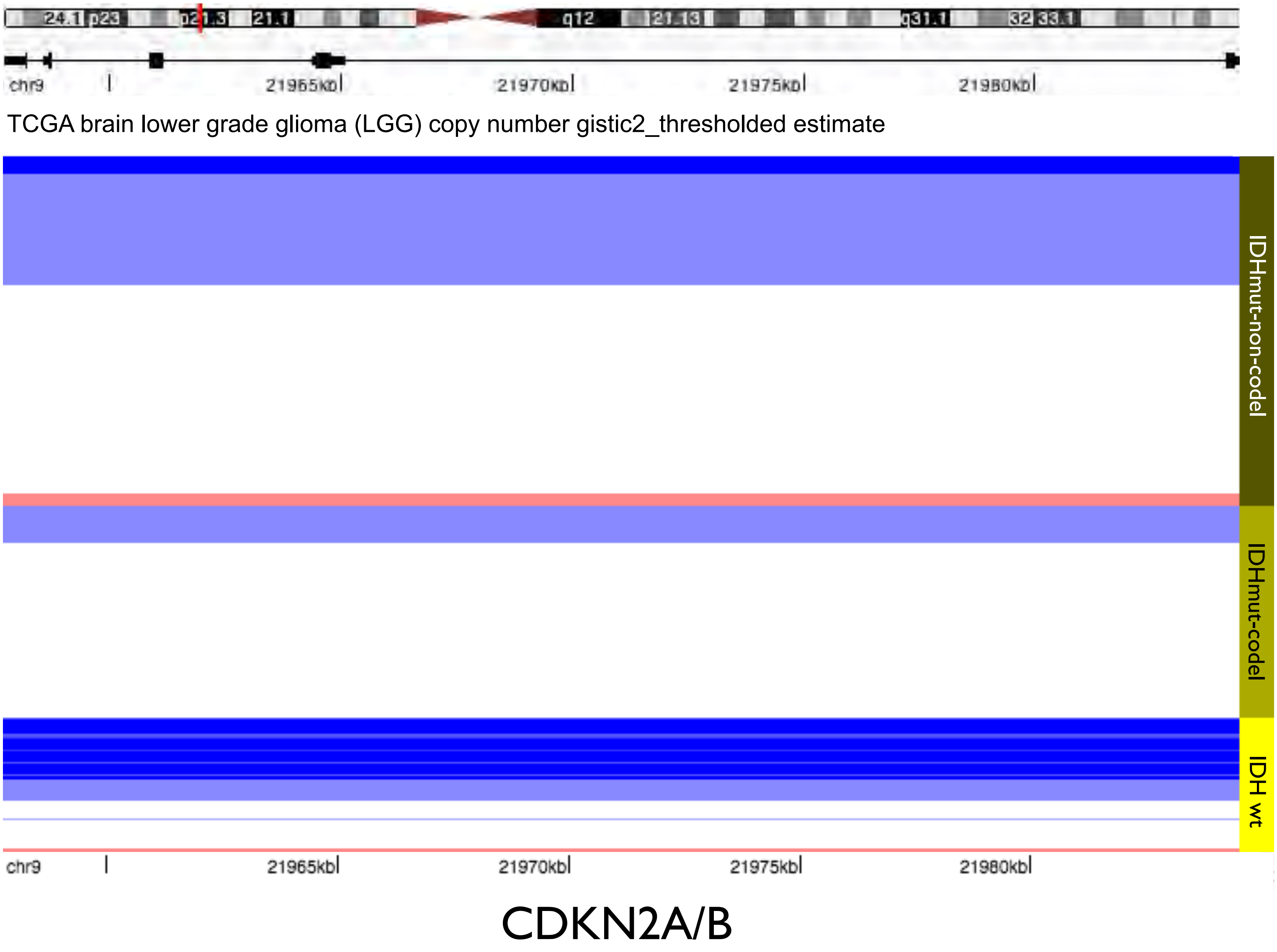


Figure S14B. The UCSC Cancer Genomics Browser view of GISTIC thresholded copy number calls for the region of chromosome 9 containing CDKN2A/B. Samples are sorted based on molecular subtype. About 30% of IDHmut-non-codel samples have deletions in CDKN2A/B, similar to the 29% of samples with chr 9p deletions shown in figure 3a, while only 4% passed the cutoff (-2) used for gene-level calls made in figure 3b.

Back to Table of Contents

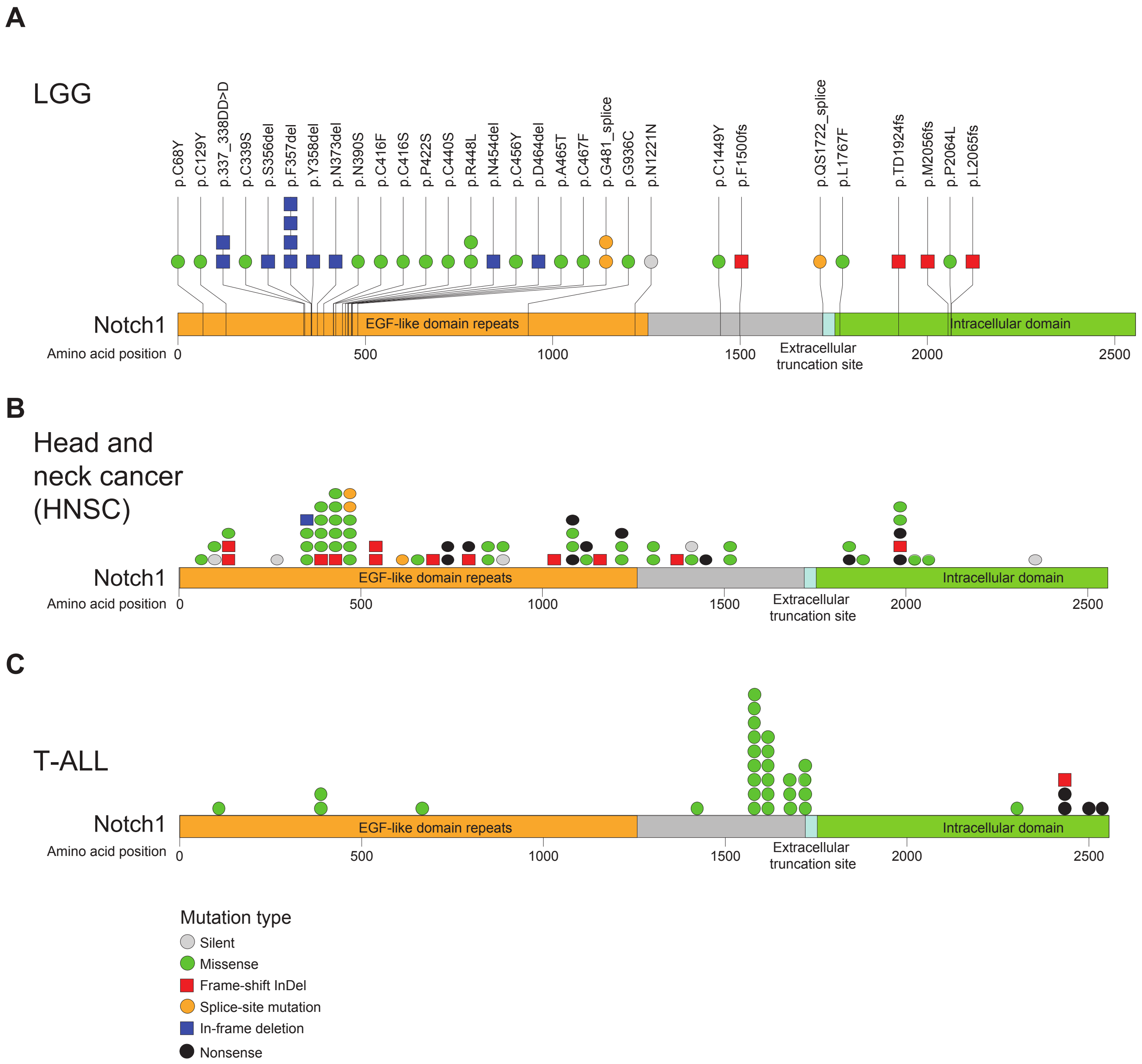


Figure S15. Similar distribution of somatic mutation events in Notch1 in LGG and head and neck cancer but not in T-ALL patients and cell lines suggests that NOTCH1 mutations in LGG are inactivating. HNSC mutation calls were obtained from Lawrence et al, 2014 through download from http://cancergenome.broadinstitute.org/. T-ALL mutations are from Keersmaecker et al, 2013.

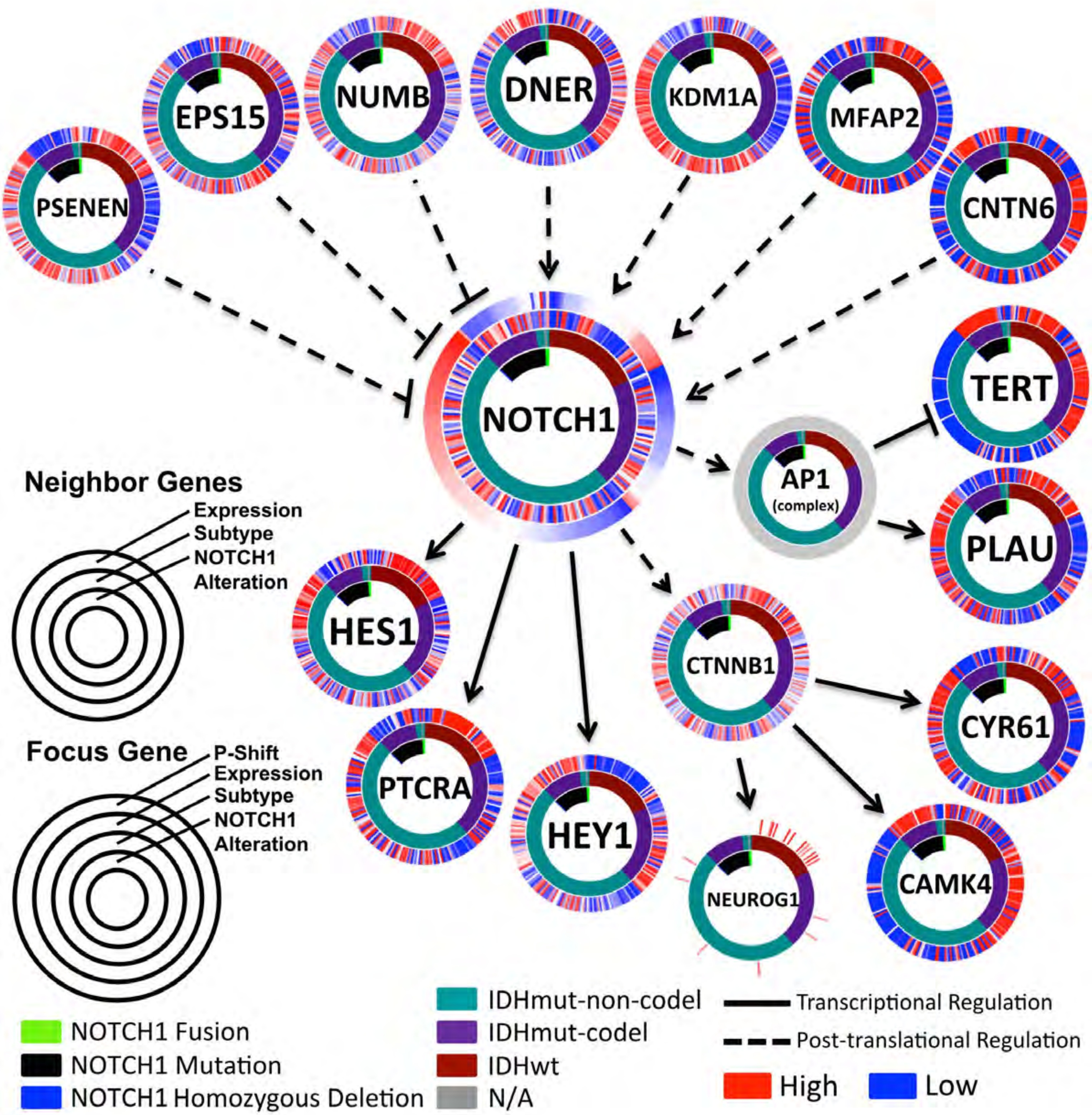


Figure S16. NOTCH1 mutations are loss of function in LGGs according to PARADIGM-SHIFT analysis. PARADIGM-SHIFT analysis of NOTCH1 pathway alterations, NOTCH1 mutations, fusions, and deletions. Circlemap display of mutation neighborhood selected for NOTCH1. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation. Samples were sorted first by the NOTCH1 alteration status (Green: NOTCH1 fusion, Black: NOTCH1 mutation, Blue: Homozygous Deletion), then by subtype, and finally by P-Shift score.

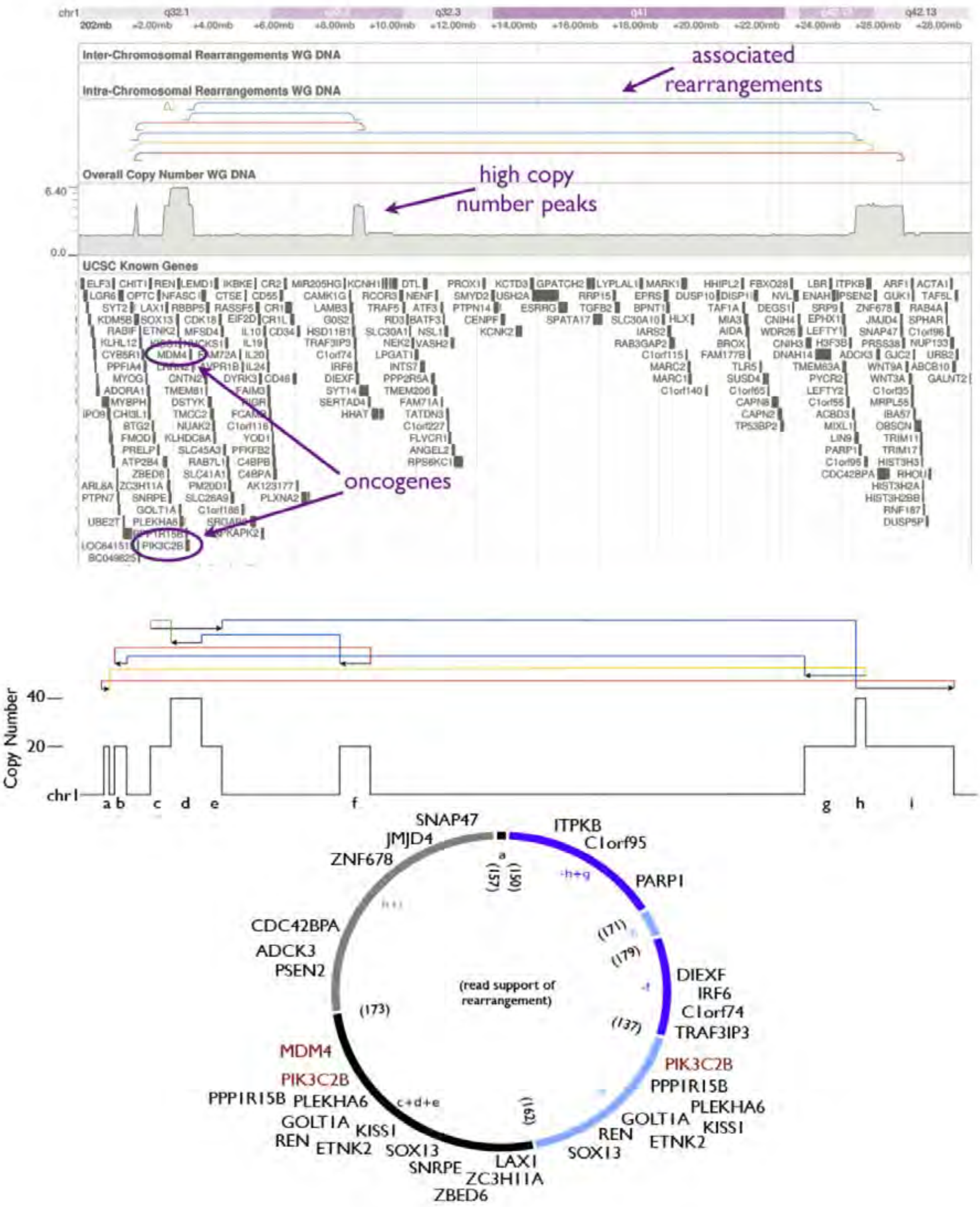


Figure S17. Reconstruction of TCGA-CS-5395 circular amplicon. Panel A shows a tumor browser view of the region of chromosome 1 containing the DM/HSR. Common features of DMs/HSRs are labeled in purple. The bottom track shows protein coding genes within the region. The track above shows relative copy number of the tumor DNA compared with normal, showing distinct blocks of elevated copy number with similar total copy number between blocks. The top two tracks show intra- and interchromosomal rearrangement breakpoints. All rearrangements shown are supported by at least 100 discordant reads. The type of rearrangement is indicated by the color of the line: duplication (red), deletion (blue) and inversion (yellow and green). Panel B shows a diagram of the amplified segments and structural variants identified on chromosome 1. Walking through this diagram results in a circular solution, suggesting the double minute chromosome diagrammed in panel C, where segments inverted relative to their orientation in the reference genome are indicated by (–) and colored blue. The letters inside the circle correspond to the segments in B. The numbers inside the circle indicate the number of sequencing reads supporting each breakpoint. Oncogenes are colored red.

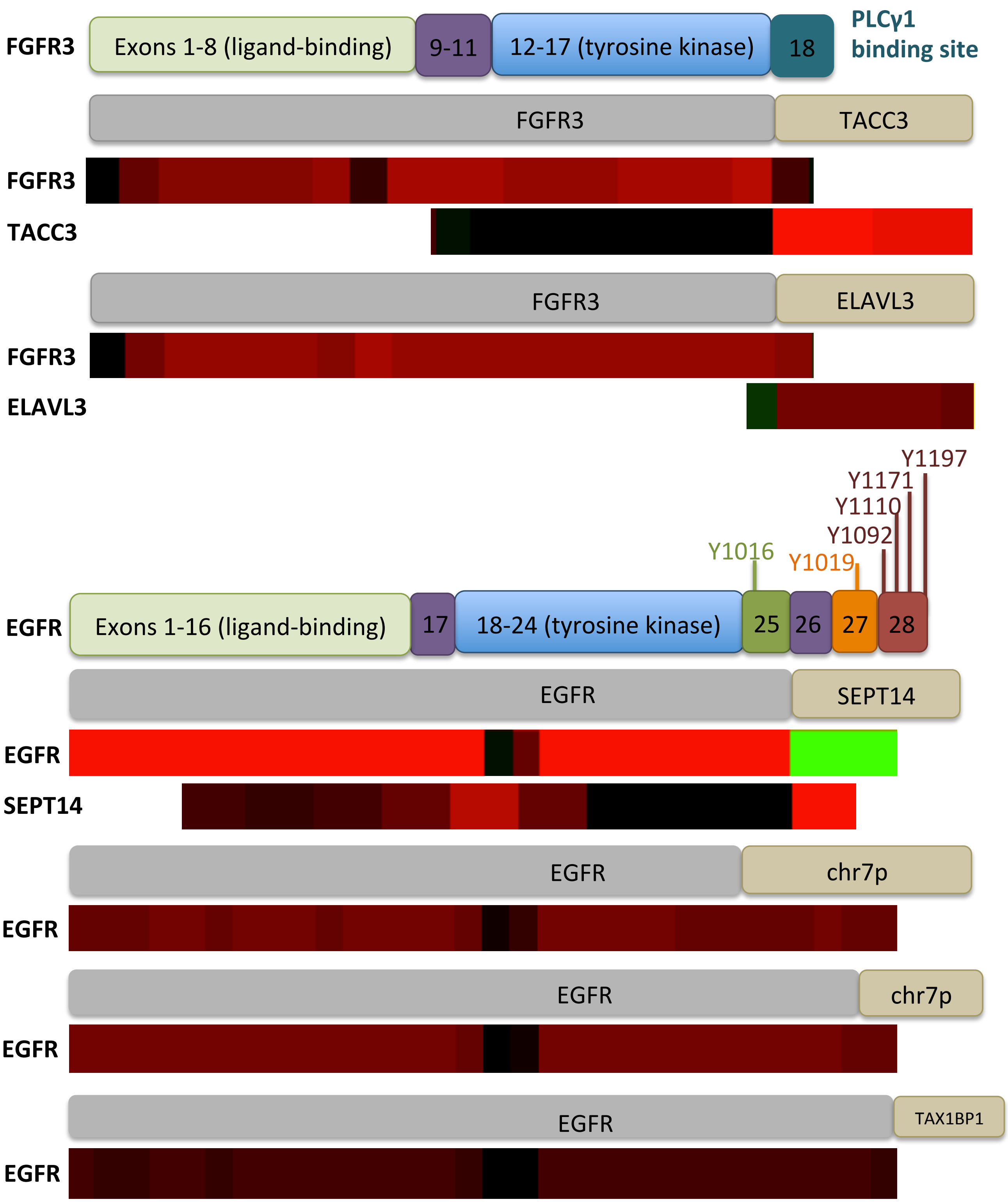


Figure S18A. Schematic representations of activating receptor tyrosine kinase fusions involving *EGFR* and *FGFR3* that were identified in IDHwt LGG tumors. The UCSC Cancer Genomics Browser view of exon-level expression for the fusion partners is displayed for each fusion-positive patient. The exon-level expression tracks are lined up with the fusion breakpoints in each case.

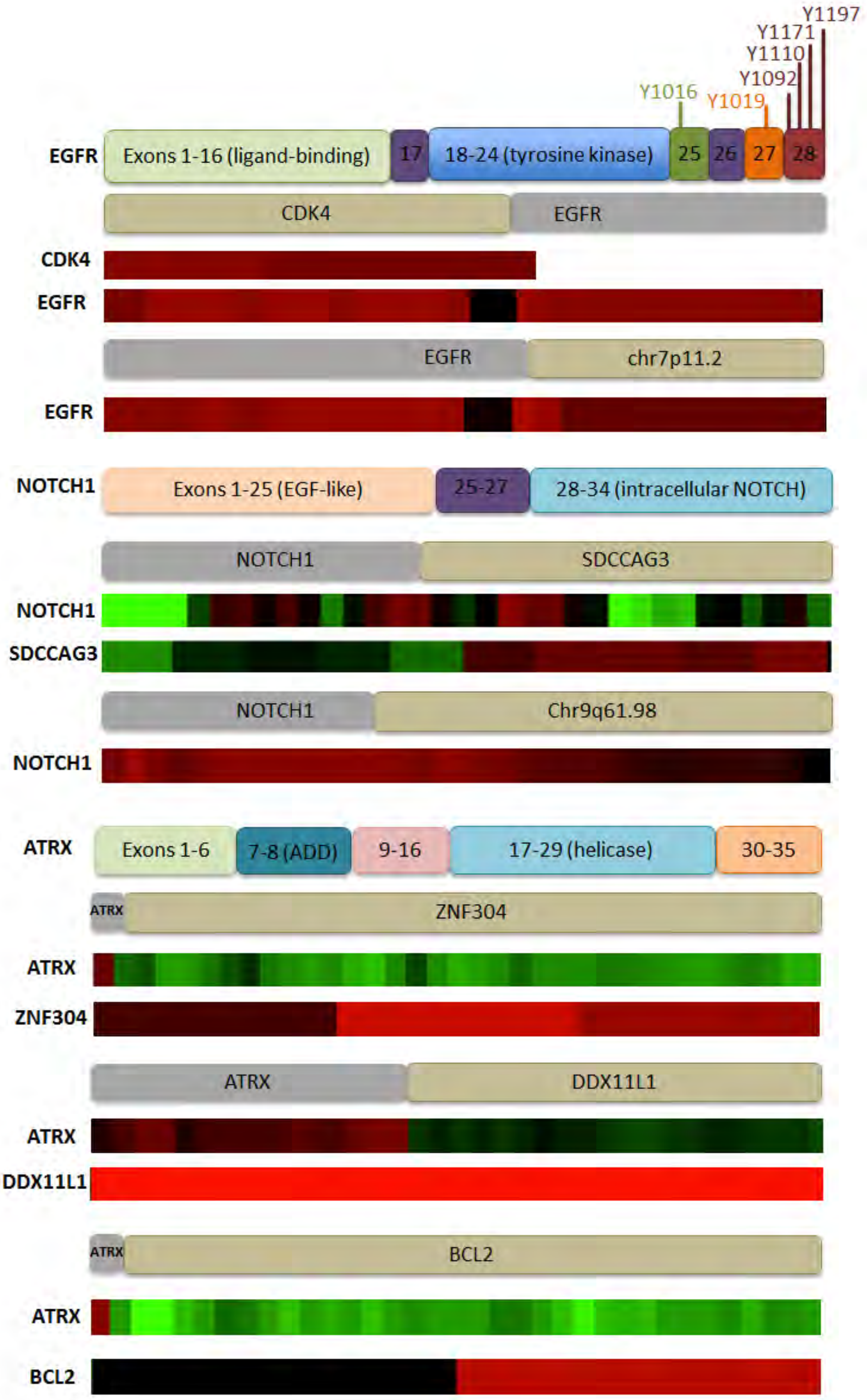


Figure S18B. Candidate loss of function fusions in ATRX, NOTCH1 and EGFR. The UCSC Cancer Genomics Browser view of exon-level expression for the fusion partners is displayed for each fusionpositive patient. The exon-level expression tracks are lined up with the fusion breakpoints in each case.

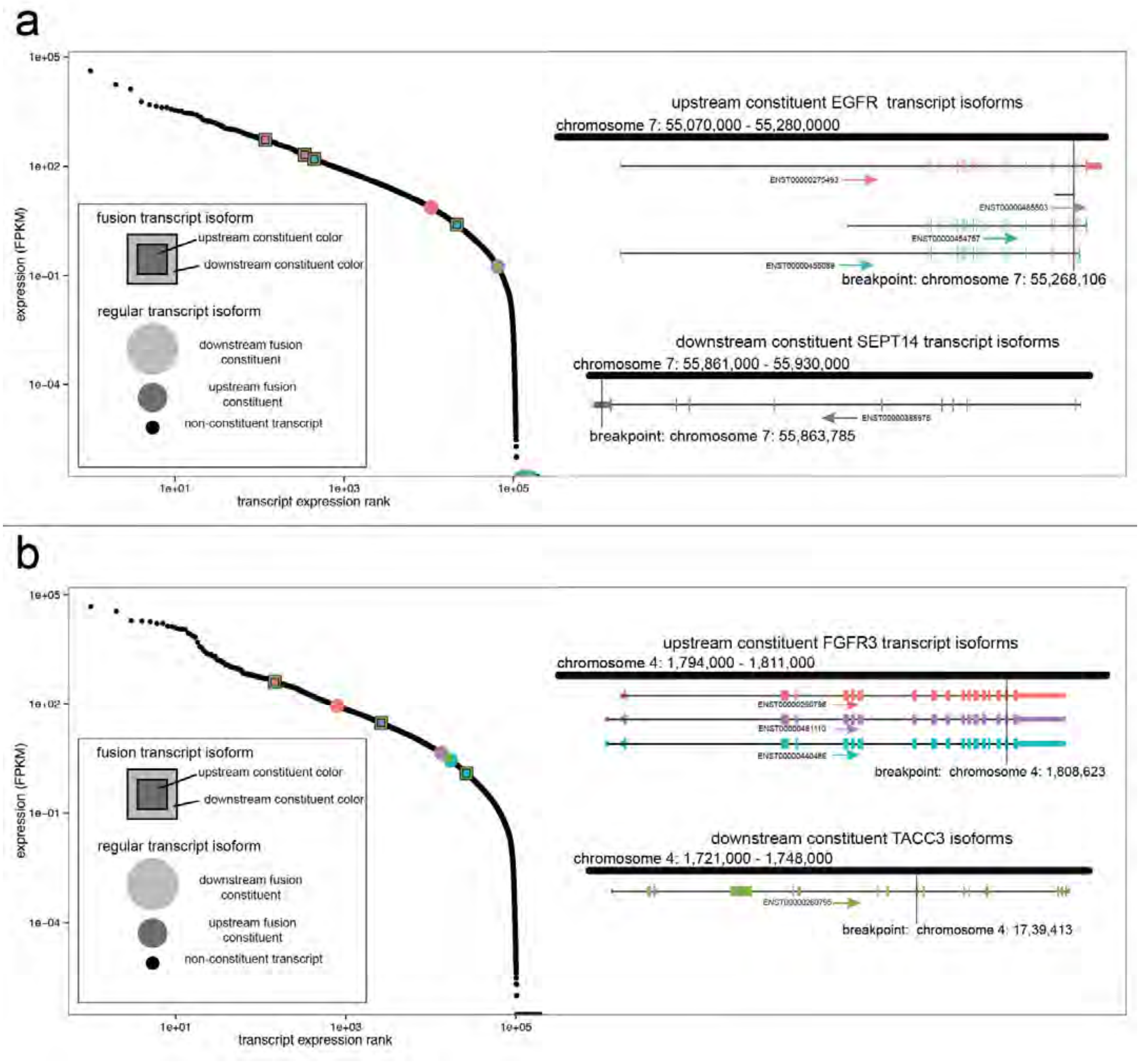


Figure S19. A. Full transcriptome profile of LGG with *EGFR-SEPT14* fusion showing expression of all transcripts, including predicted EGFR-SEPT14 fusion transcript isoforms (square points) and their constituents (larger circular points). Note that predicted fusion transcripts are among the highest expressed transcripts in the transcriptome, and that each fusion transcript is expressed much more highly than its constituent transcripts. B. Similar to a. for LGG *with FGFR3-TACC3* fusion. Here, the top two most expressed fusions (green/red square and green/purple square, respectively) are expressed much higher than constituents.

Back to Table of Contents

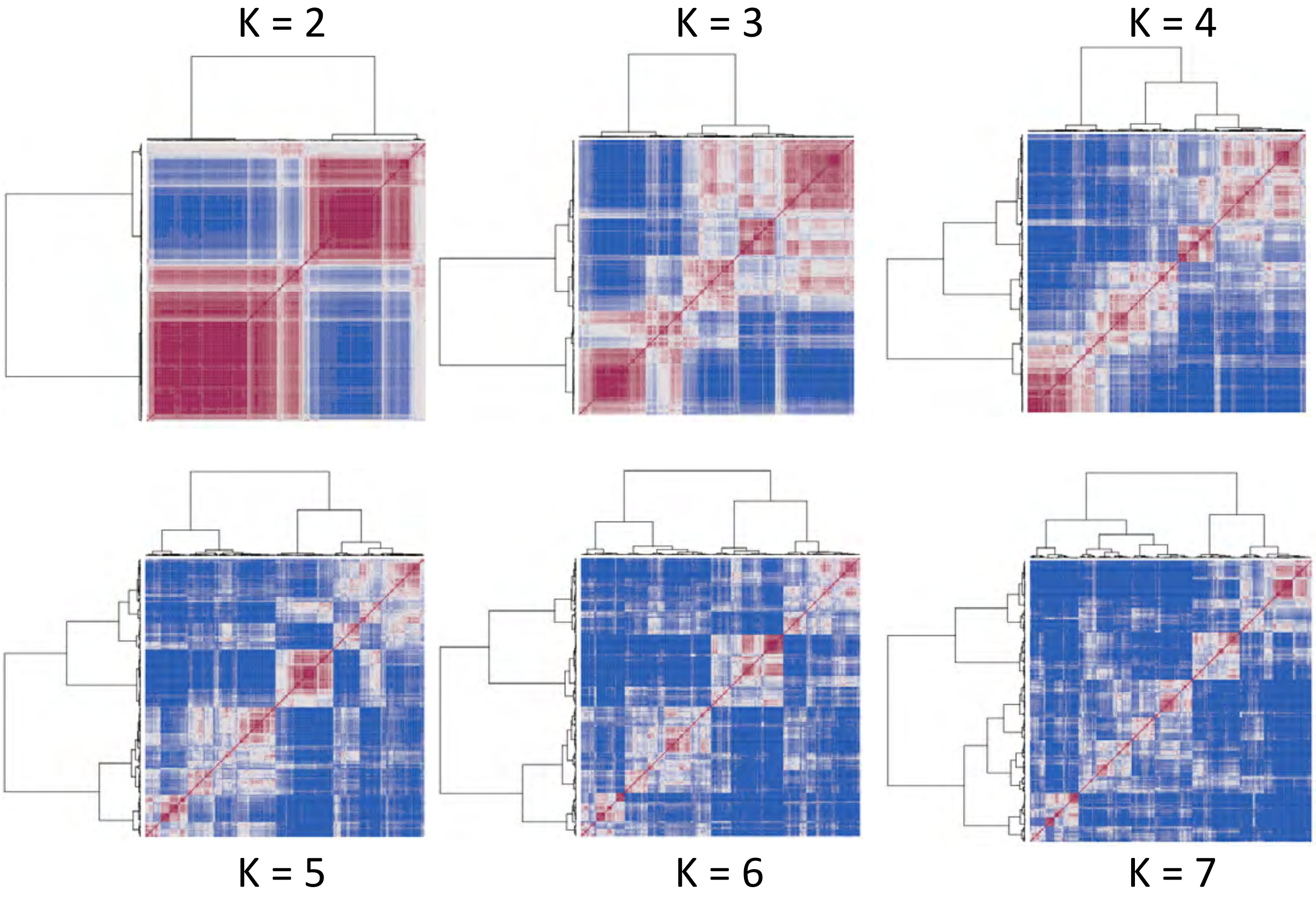


Figure S20A. Consensus clustering of RPPA data from 255 LGG samples with different values of the number of clusters, K (2-7). K=4 was chosen as the most robust number. The third cluster in K=4 is quite robust, whereas the others are less robust.

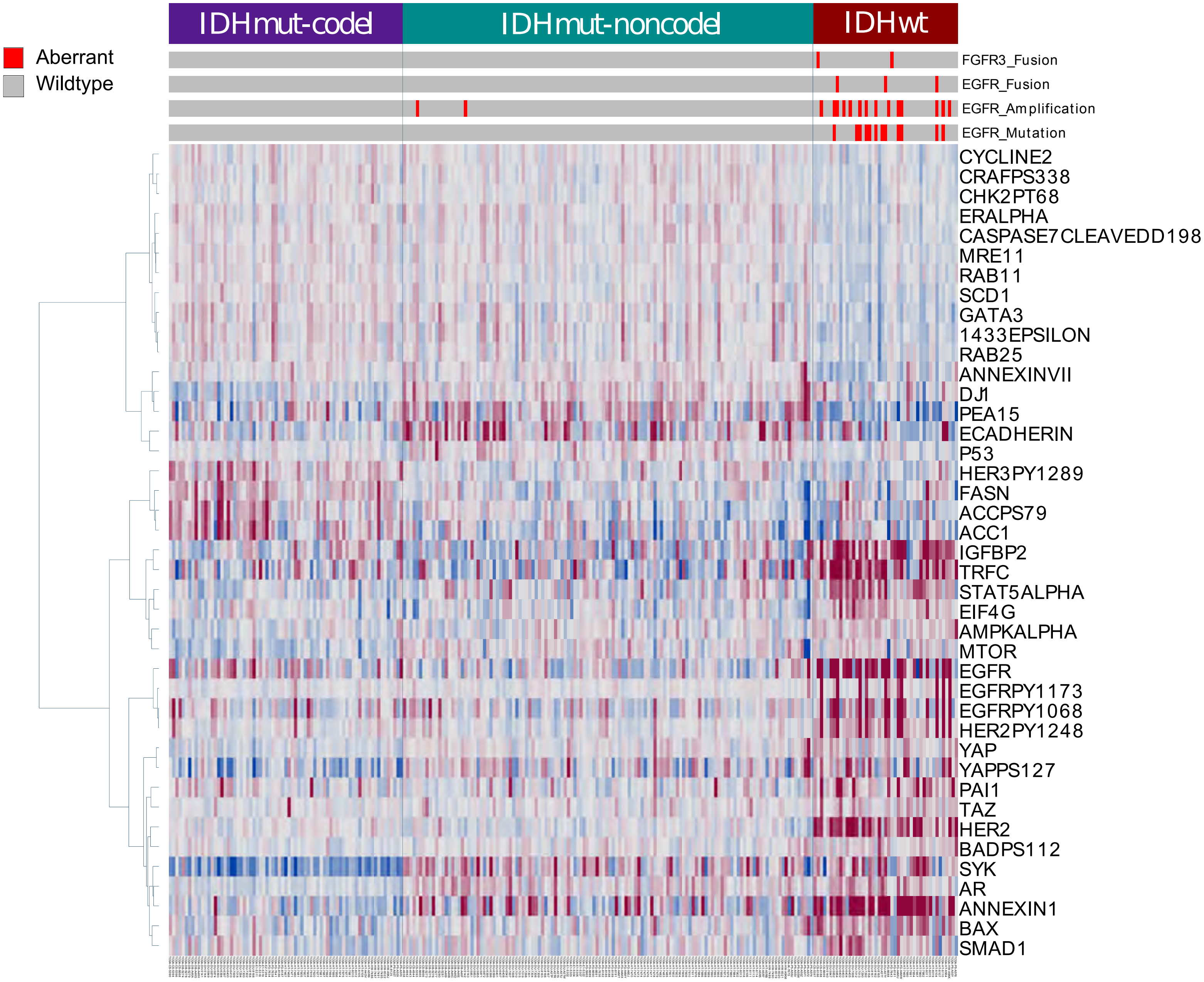


Figure S20B. The relative total and phosphoprotein expression levels from reverse phase protein array (RPPA) analysis. From a cohort of 189 proteins, the 40 proteins shown here were found to be differentially expressed (BH-adjusted p-value < 1e-5) among the three molecular subtypes. Expression levels of most of the receptor tyrosine kinase pathway proteins (highlighted) except phosphoHER3 are significantly elevated in IDHwt LGG compared to other subtypes. All of the samples with EGFR amplification or fusion, and most of the samples with EGFR mutation showed elevated EGFR expression levels.

Back to Table of Contents

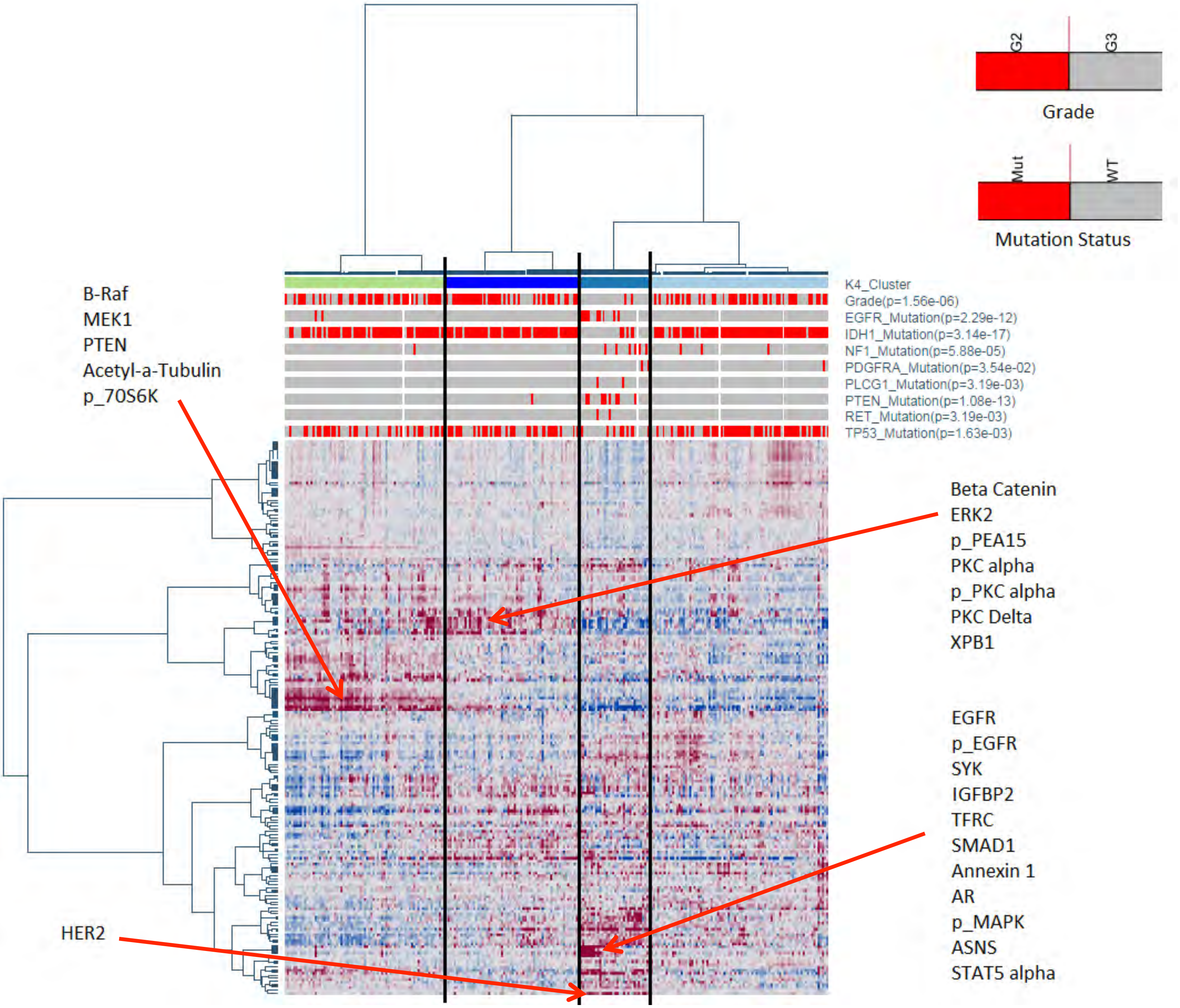
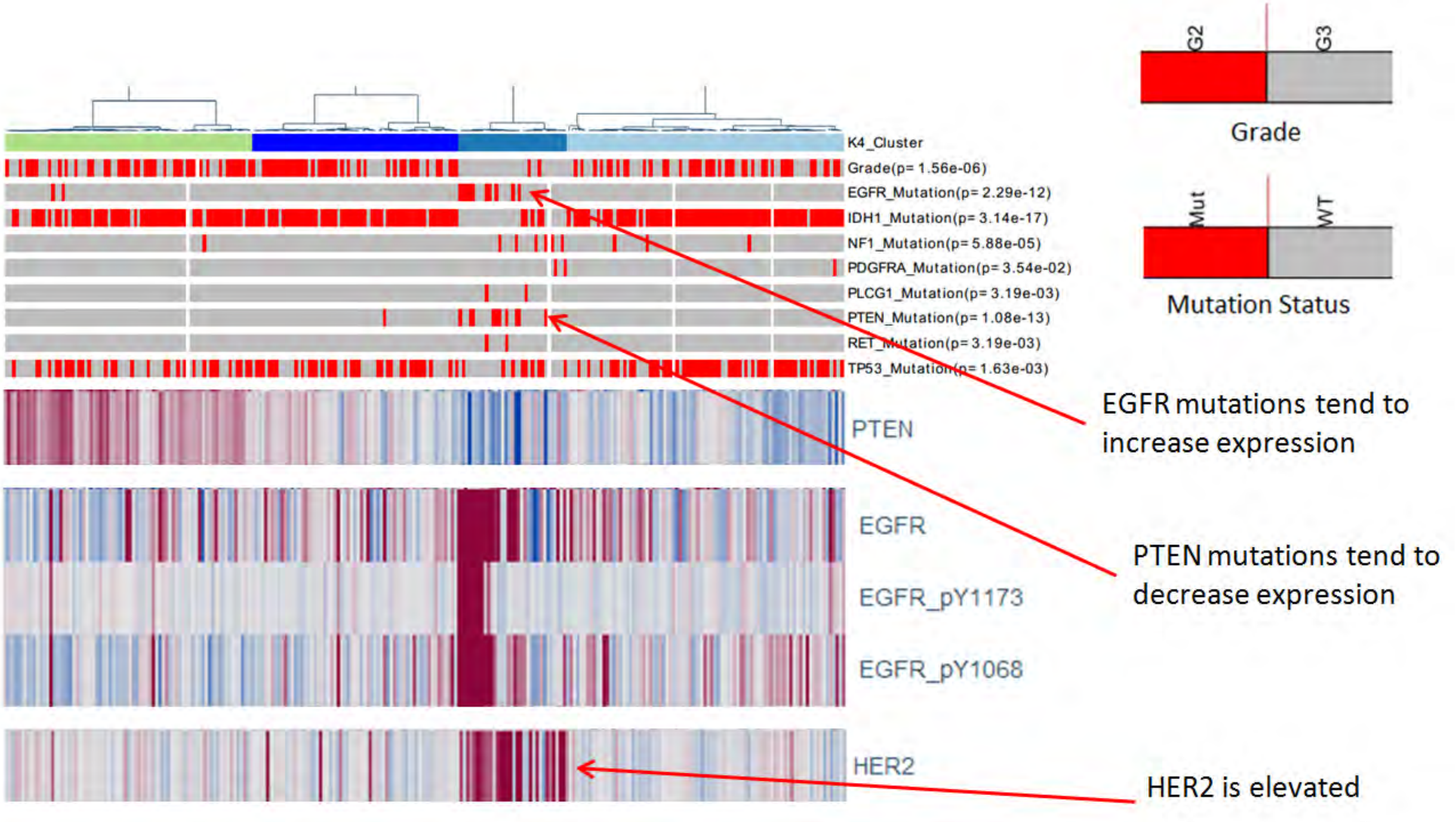


Figure S20C. Unsupervised clustering of 255 LGG samples and 189 proteins shows 4 clusters. Red color indicates high protein level, white indicates medium level, and blue indicates low level. The third cluster from the left is depleted in IDH mutants and has worse prognosis. Markers for each cluster are identified. The last cluster is depleted in markers for the other three. The association of clusters with mutations and grade is shown on top of the heatmap with annotation bars. Chi-squared P-values were computed for each variable versus the clusters. A zoomed in view can be seen in Figure S20D.

Back to Table of Contents

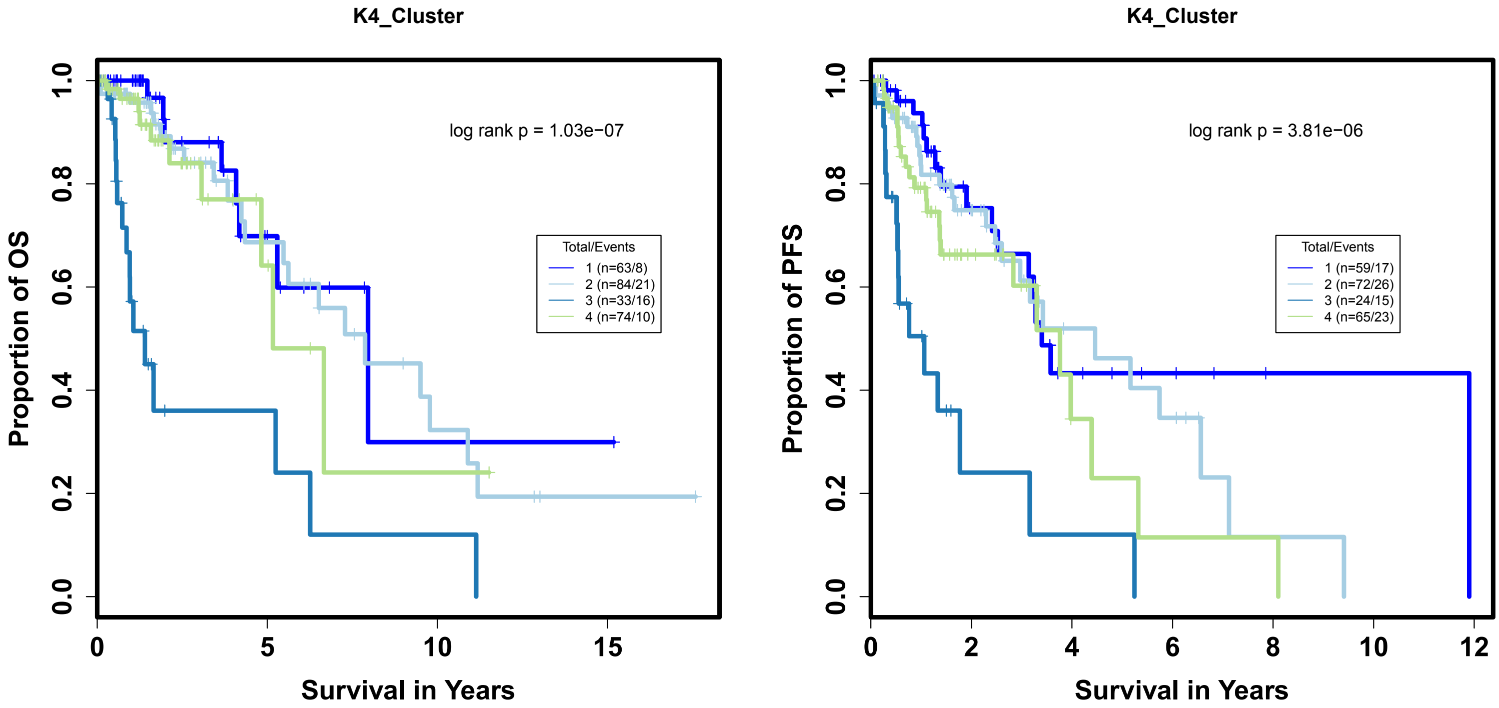


EGFR mutations tend to increase expression

PTEN mutations tend to decrease expression

HER2 is elevated

Figure S20D. Zoomed in view of the heatmap shown in Figure S20C. The third cluster has EGFR mutations that correlate with high EGFR and phosphoEGFR levels. PTEN mutations correlate with low PTEN expression. The cluster also has high HER2 levels, raising possibilities for targeted therapies. The P-values in parentheses are based on Student’s t-test.



Overall Survival

Progression Free Survival

Figure S20E. Survival analysis based on RPPA cluster. RPPA Cluster 3 has statistically significant shorter overall and progression free survival.

Back to Table of Contents

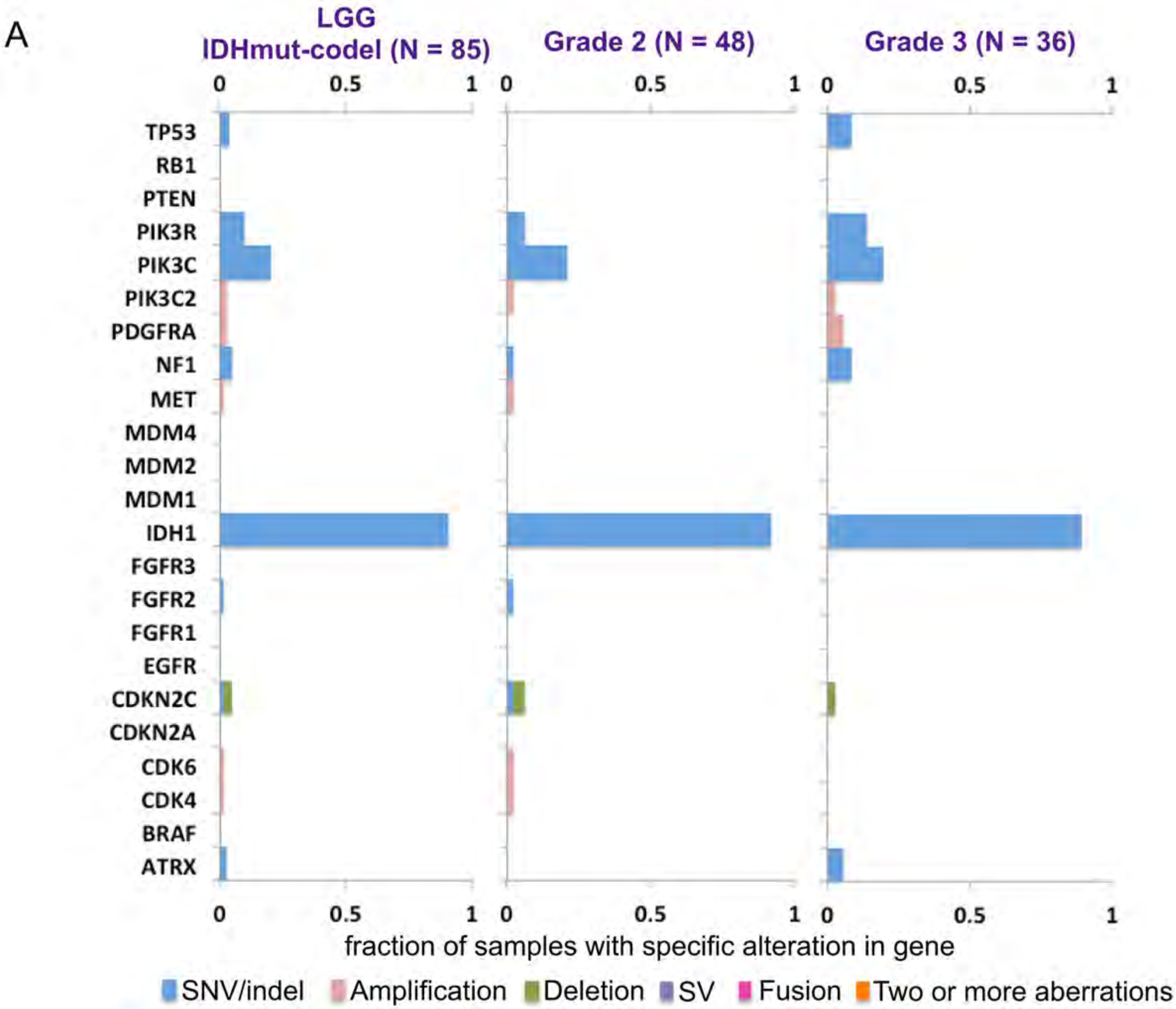


Figure S21A: Frequencies of mutational events commonly found in IDHwt GBM in grade 2 and grade 3 IDHmut-codel LGG. There was no grade information for one IDHmut-codel LGG.

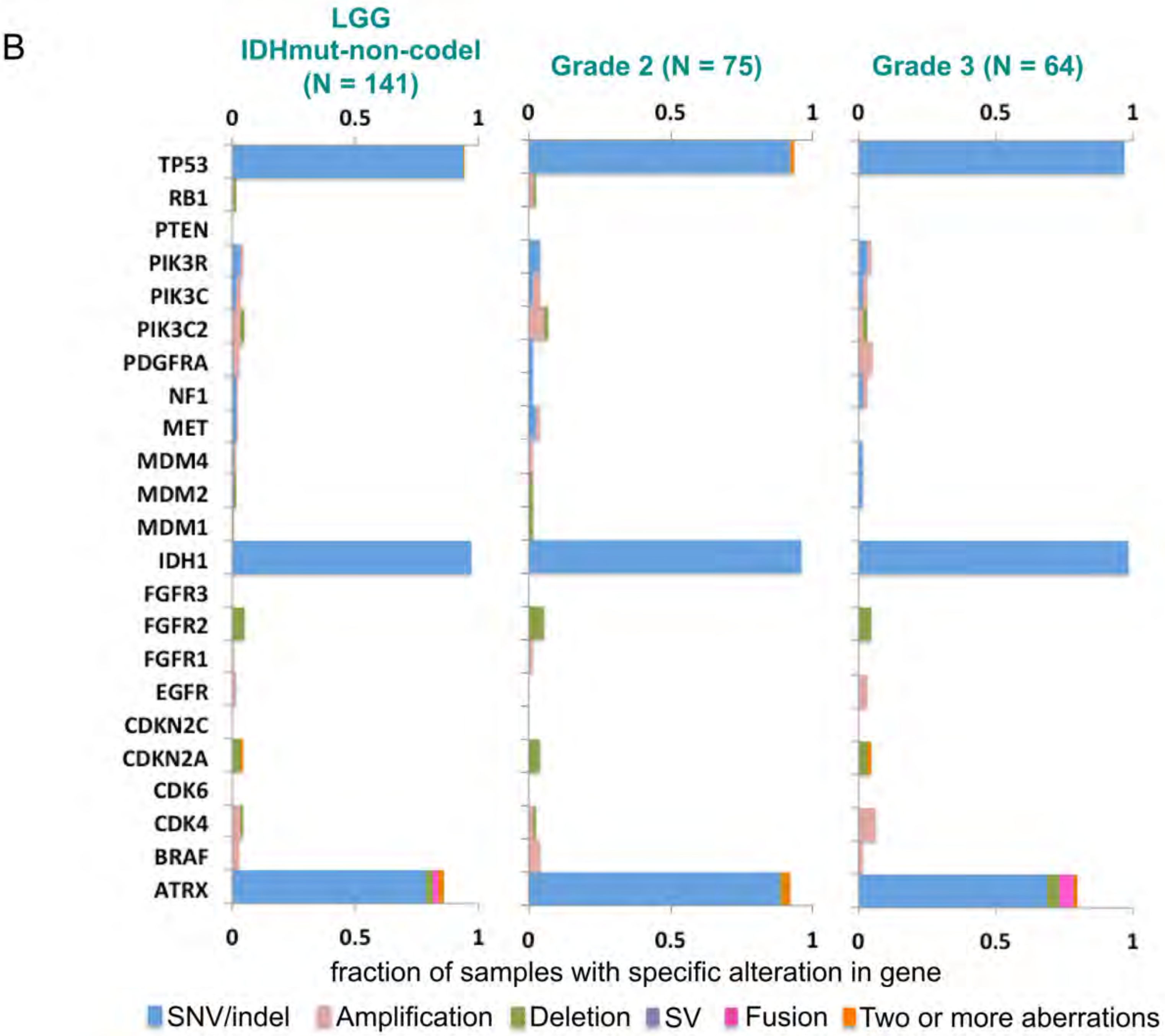


Figure S21B: Frequencies of mutational events commonly found in IDHwt GBM in grade 2 and grade 3 IDHmut-non-codel LGG. There was no grade information for two IDHmut-non-codel LGGs.

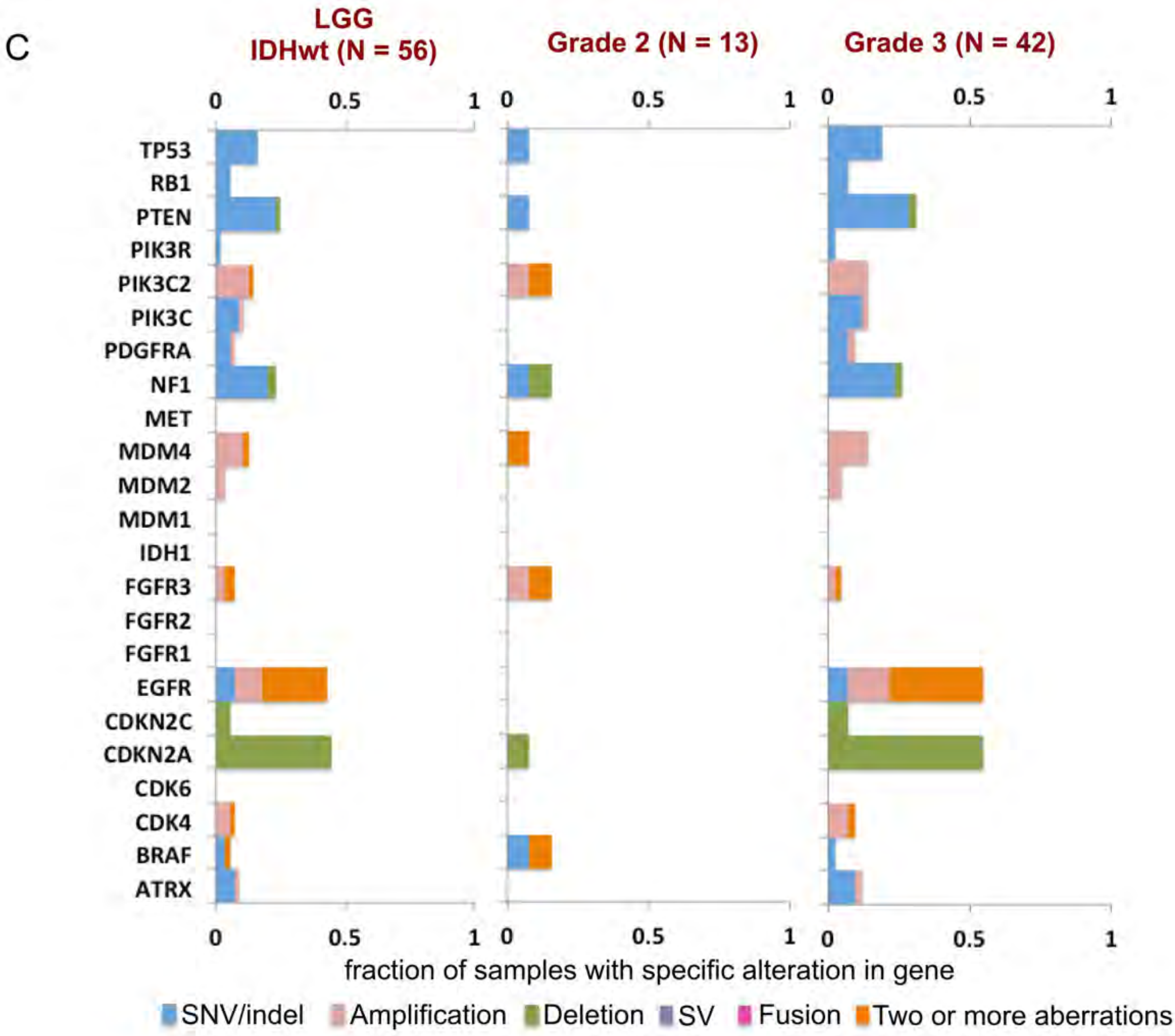


Figure S21C: Frequencies of mutational events commonly found in IDHwt GBM in grade 2 and grade 3 IDHwt LGG. There was no grade information for one IDHwt LGG.

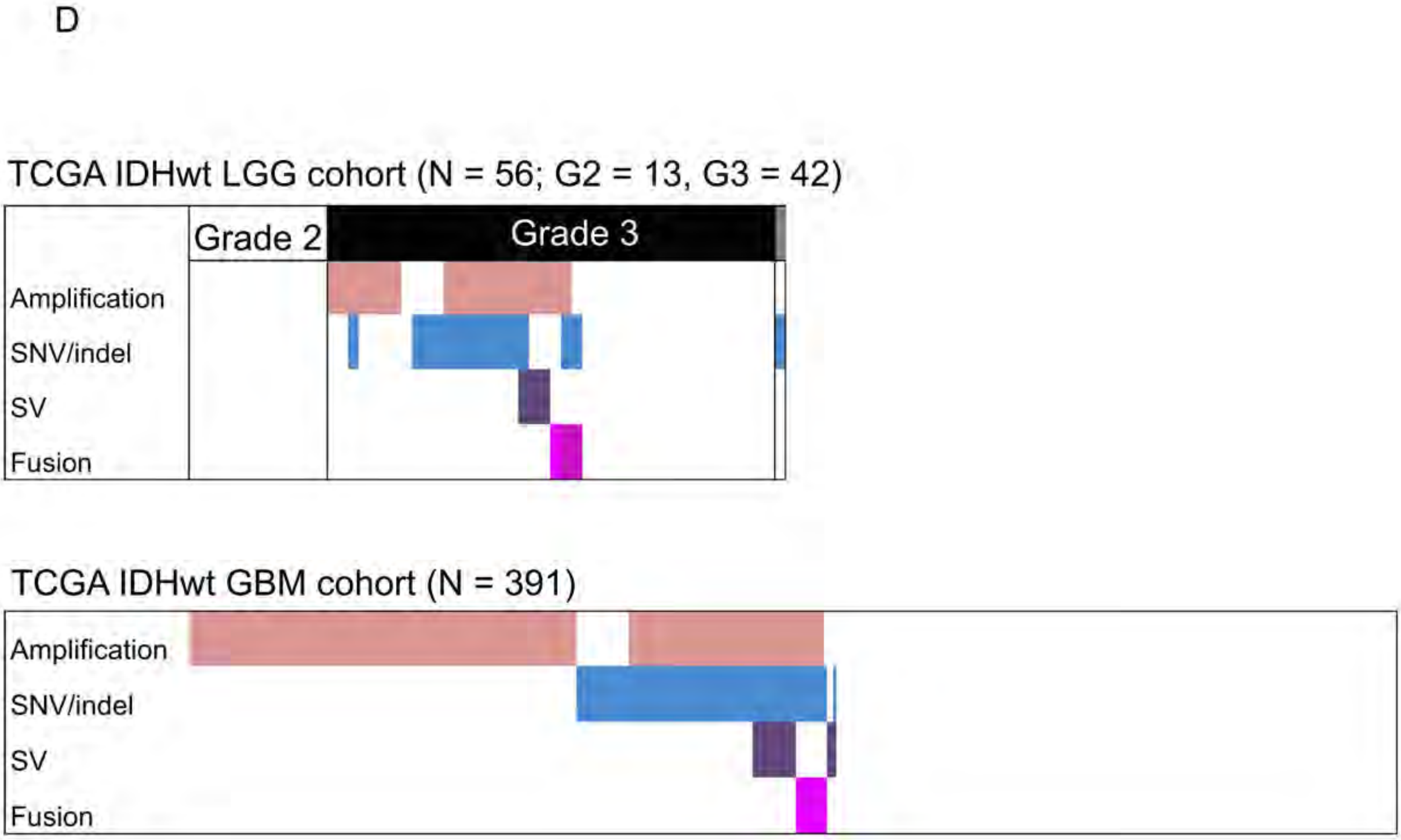


Figure S21D. Spectrum of EGFR alterations in IDHwt LGG and GBMs. Mutation calls in EGFR were obtained from the cross-center combined MAF file. Copy number alterations in EGFR were obtained from GISTIC thresholded gene-level analysis, and included gene-level copy number. A gene was considered amplified if the GISTIC thresholded value was 2 and deleted if it was -2. High confidence DNA rearrangements identified using BamBam and BreakDancer and listed in Table S5 were included and denoted “SV”. Manually reviewed fusion predictions (Figure 3C, Table S6) affecting EGFR were also shown.

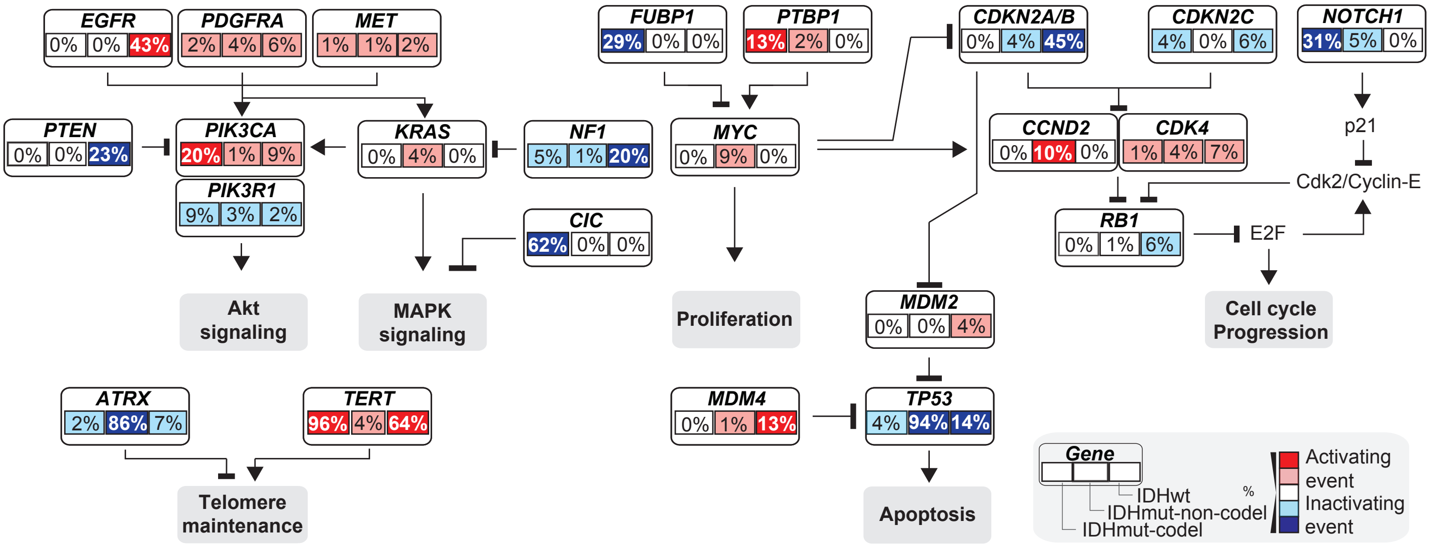


Figure S21E. Pathway alterations of LGG as a schematic that summarizes genomic alterations (mutations, focal amplifications, homozygous deletions and fusions) in LGG across molecular subtypes. Canonical RTK/PI3K/MAPK, RB and p53 regulatory pathways are frequently altered, as is telomere maintenance. Alterations with specific pathway components depend strongly on the LGG subtype.

Back to Table of Contents

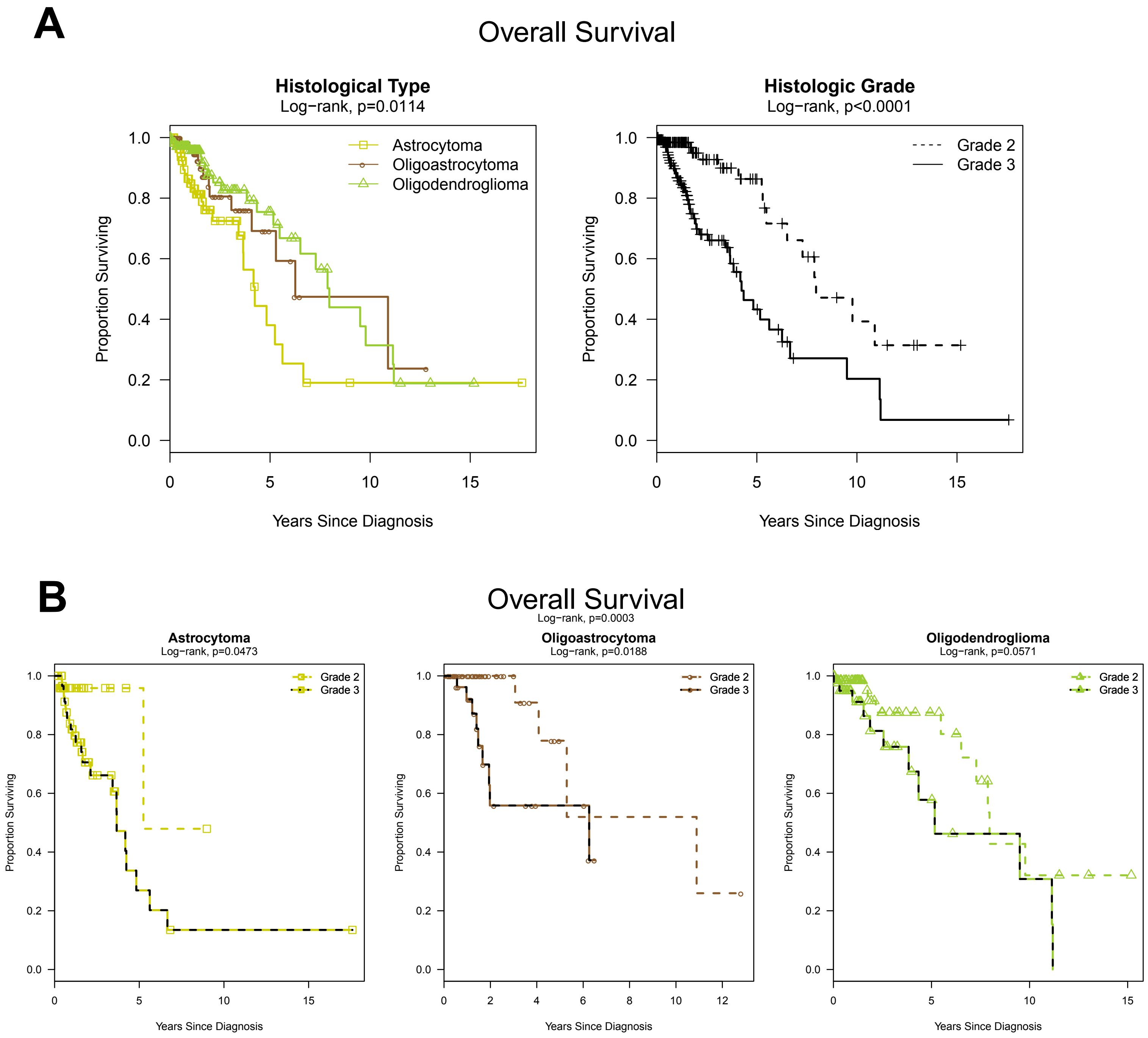


Figure S22A,B. Estimated survival curves. A. Kaplan-Meier survival curves are drawn for overall survival by histopathological type (left) and by WHO grade (right). B. Kaplan-Meier survival curves of overall survival by type-grade combinations. Comparisons of the survival curves are tested by log-rank test for each plot and across type-grade combinations.

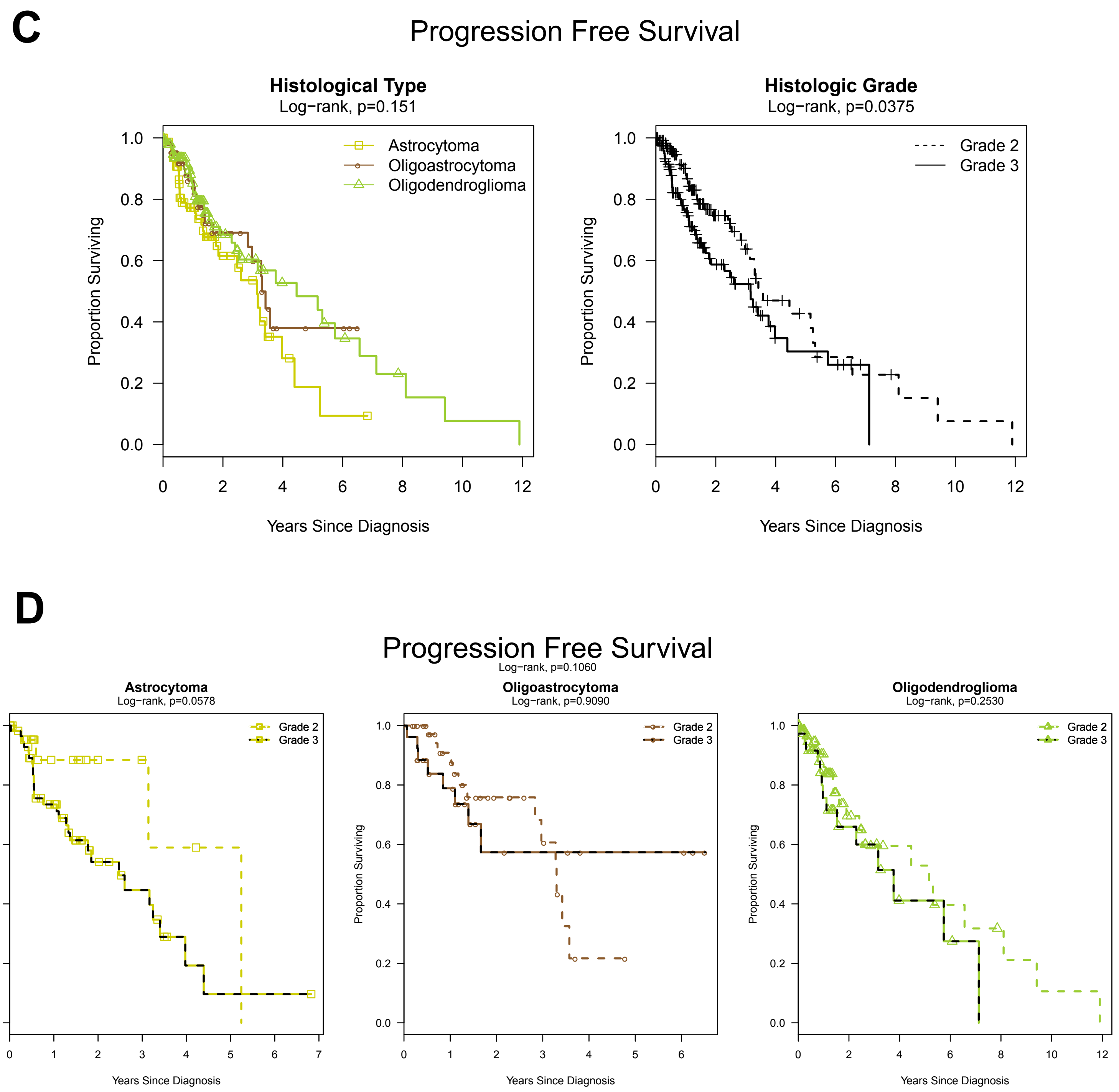


Figure S22C,D. Estimated survival curves. C. Kaplan-Meier survival curves are drawn for progression free survival by histopathological type (left) and by WHO grade (right, N=250). D. Kaplan-Meier survival curves of progression free survival by type-grade combinations. Comparisons of the survival curves are tested by log-rank test for each plot and across type-grade combinations.

Back to Table of Contents

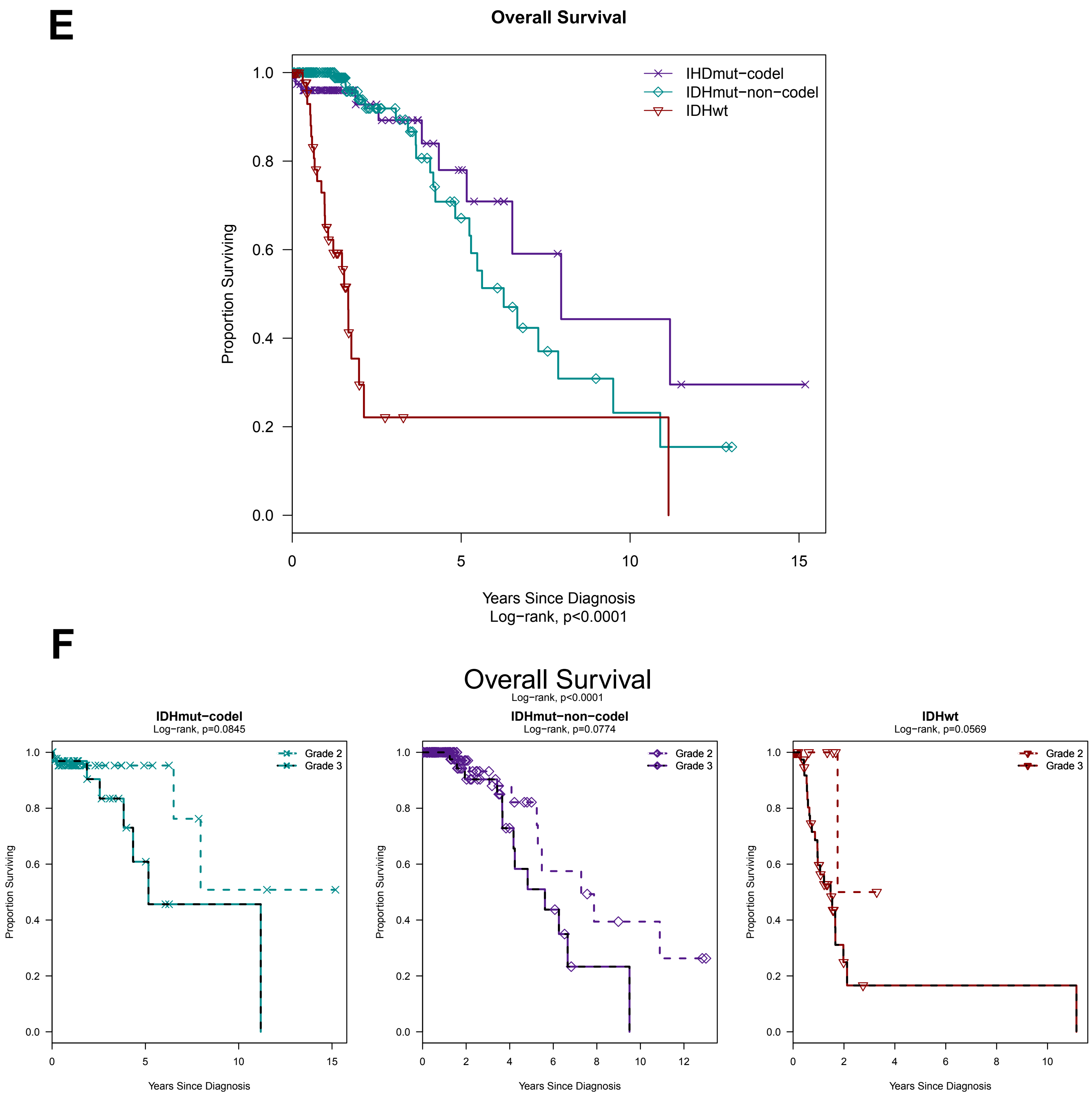


Figure S22E,F. Estimated survival curves. E. Kaplan-Meier survival curves are drawn for overall survival by IDH/Codel group (N=278). F. Kaplan-Meier survival curves of overall survival by IDH/Codel-grade combinations. Comparisons of the survival curves are tested by log-rank test for each plot and across IDH/Codel-grade combinations.

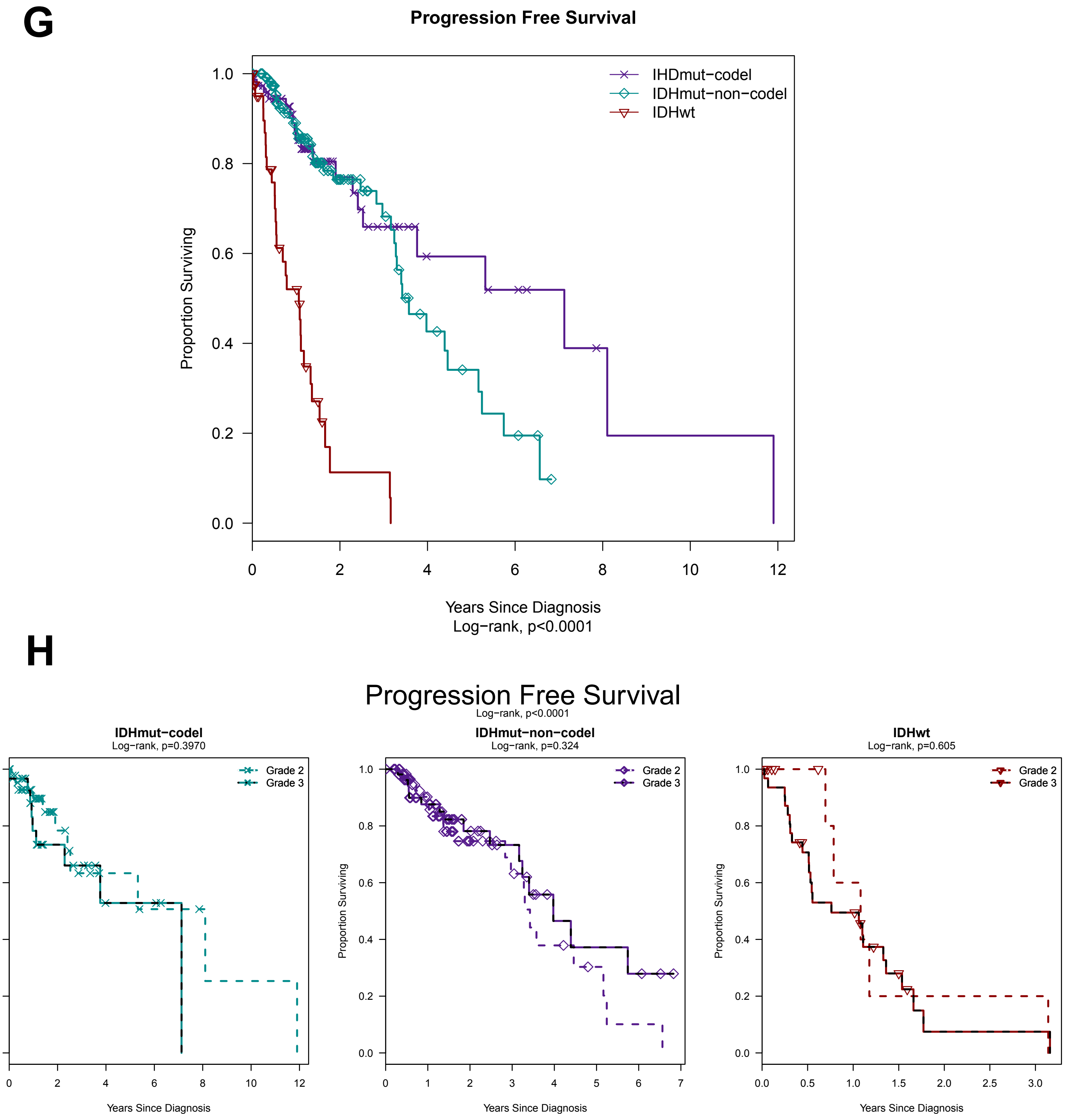


Figure S22G,H. Estimated survival curves. G. Kaplan-Meier survival curves are drawn for progression-free survival by IDH/Codel group (N=241). F. Kaplan-Meier survival curves of progression-free survival by IDH/Codel-grade combinations. Comparisons of the survival curves are tested by log-rank test for each plot and across IDH/Codel-grade combinations

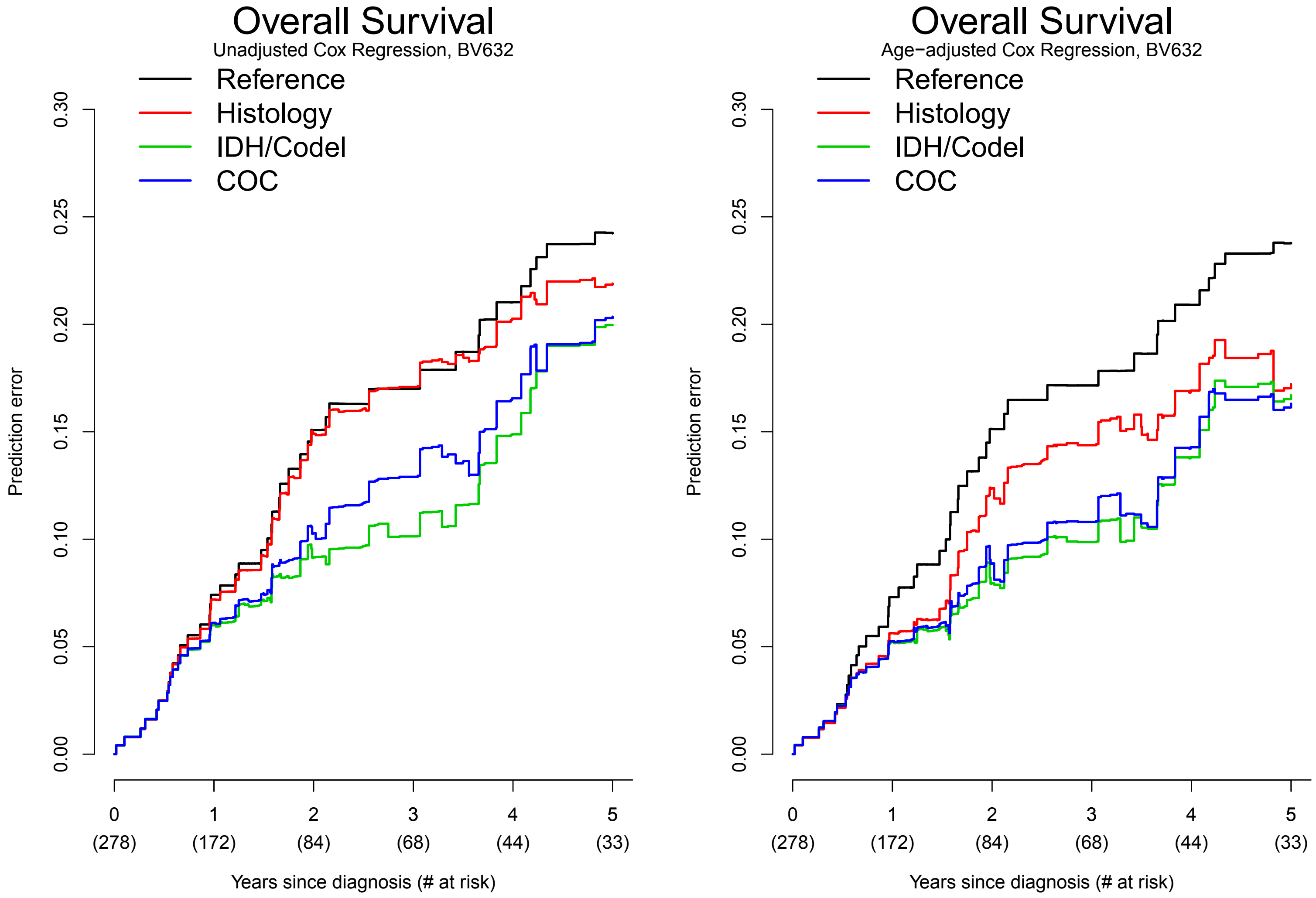


Figure S22I. Prediction error curves are plotted for Cox regression models of overall survival using the three histology classes (red), the three IDH/Codel classes (green), or the three clusterof- cluster classes (blue) as predictors with (right) and without (left) adjustment for age at diagnosis. Error for a model with no predictors is given as reference (black). The number of samples at risk at the start of each year are given in the horizontal axis labels. Bootstrap resampling (B=100) was used to temper optimistic estimation of the apparent error according to the 0.632 weighting rule.

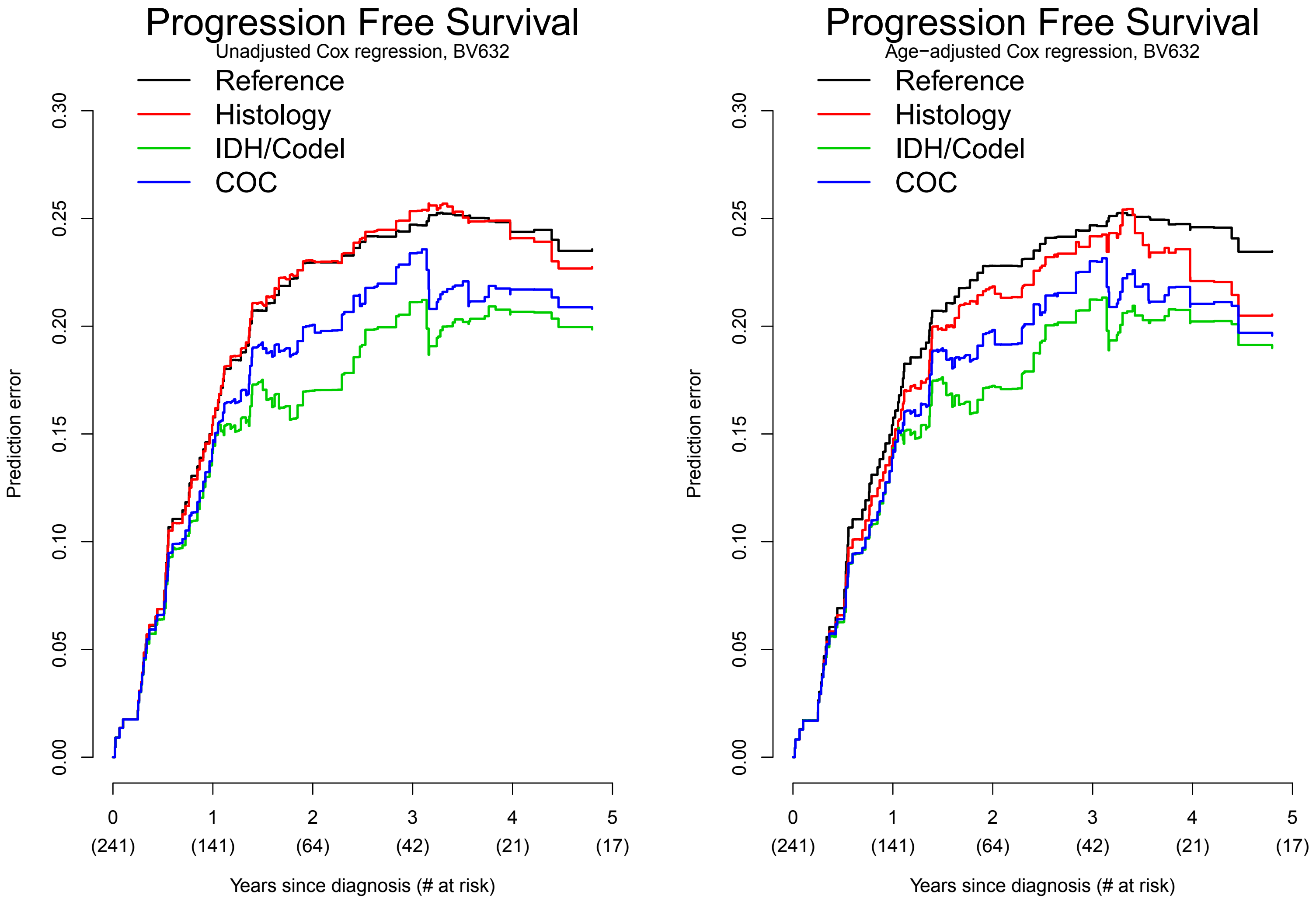


Figure S22J. Prediction error curves are plotted for Cox regression models of progression-free survival using the three histology classes (red), the three IDH/Codel classes (green), or the three cluster-of-cluster classes (blue) as predictors with (right) and without (left) adjustment for age at diagnosis. Error for a model with no predictors is given as reference (black). The number of samples at risk at the start of each year are given in the horizontal axis labels. Bootstrap resampling (B=100) was used to temper optimistic estimation of the apparent error according to the 0.632 weighting rule.

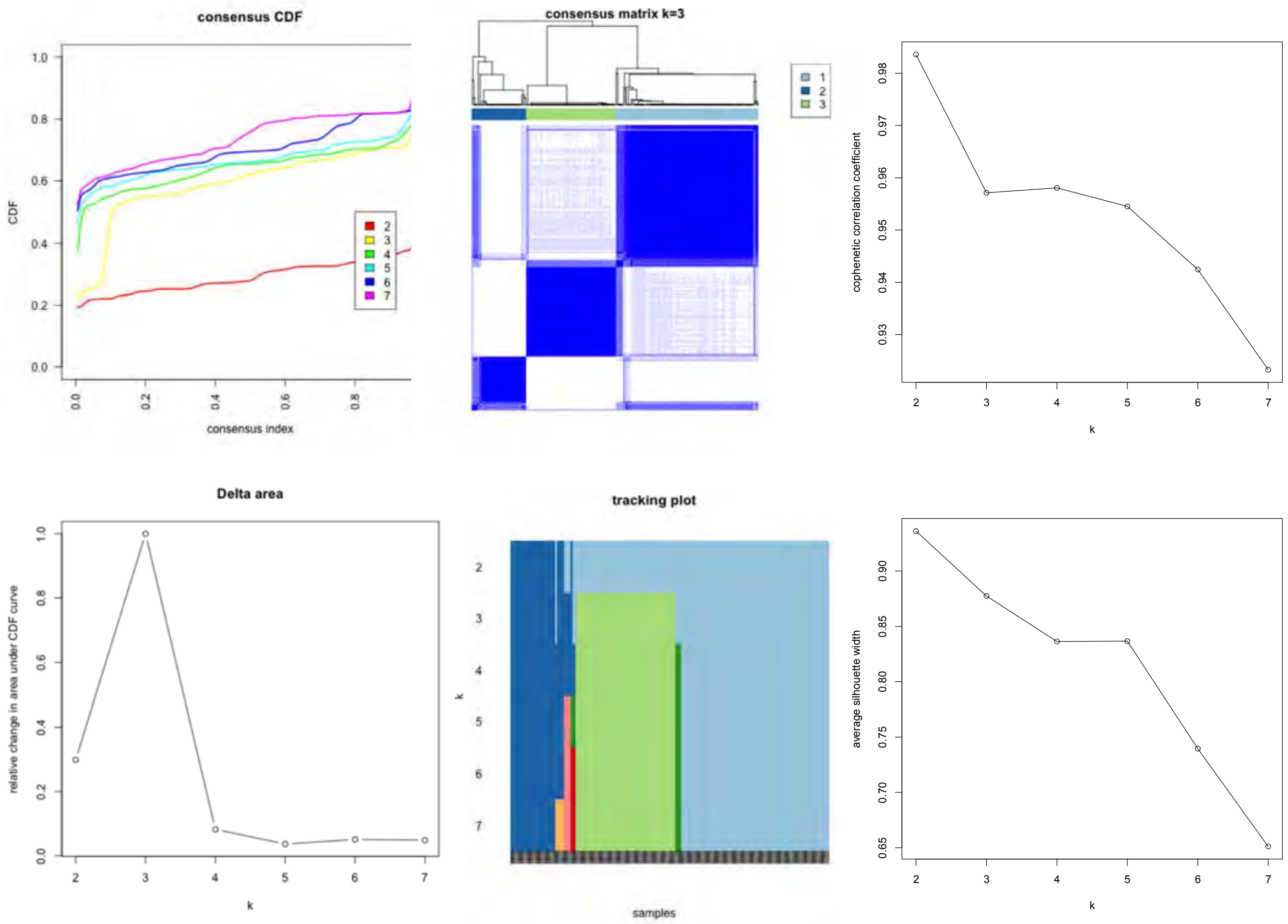


Figure S23. **Metrics for cluster of clusters model selection.** From top-left: consensus cumulative distribution function (CDF), consensus matrix for k=3, cophenetic correlation coefficient, relative change in area under the CDF curve, sample tracking plot, average silhouette width.

Back to Table of Contents

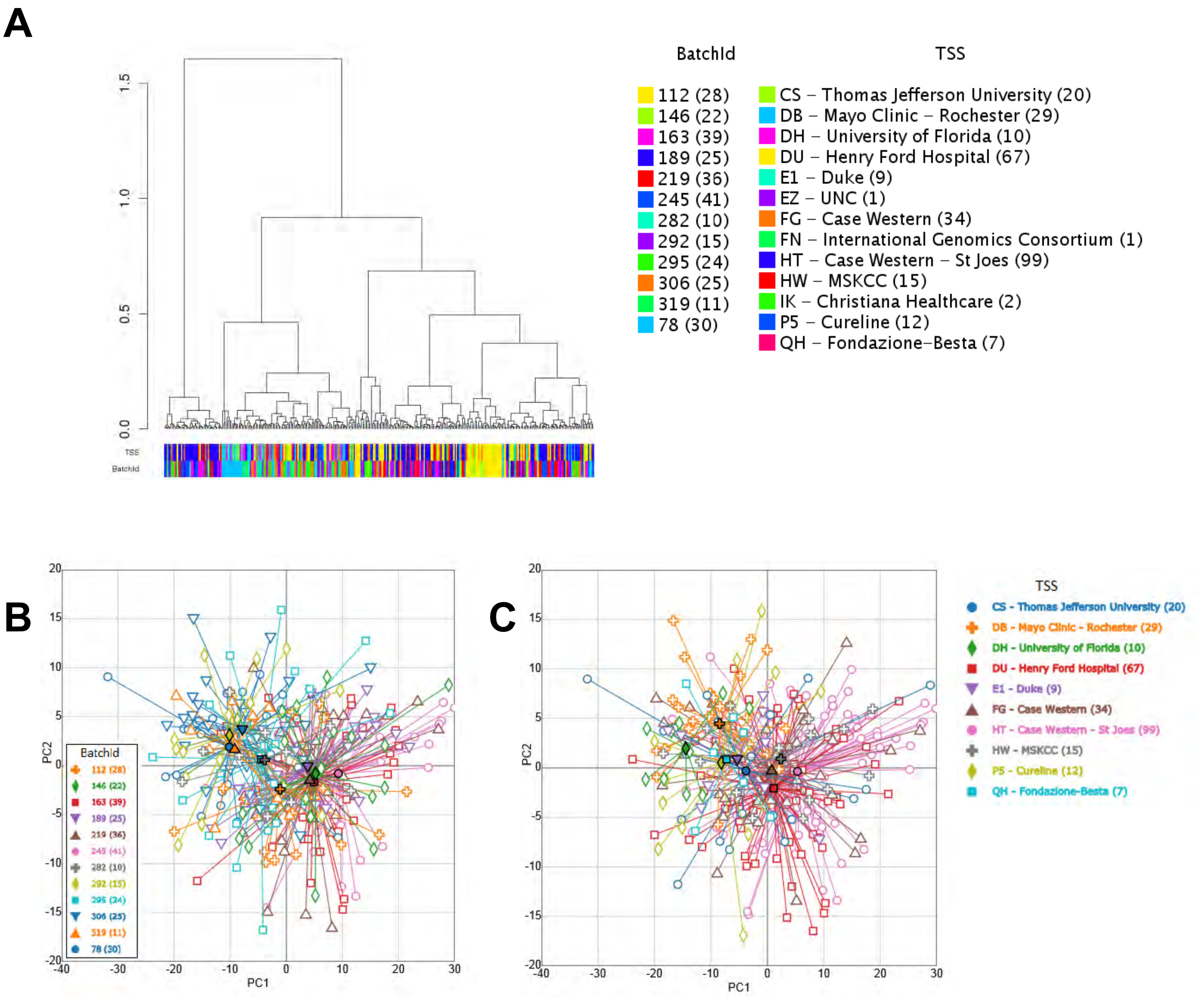


Figure S24. **Batch effects.** A. Hierarchical clustering for miRNA expression from miRNA-seq data. B. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to batch ID. C. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to TSS.

Back to Table of Contents

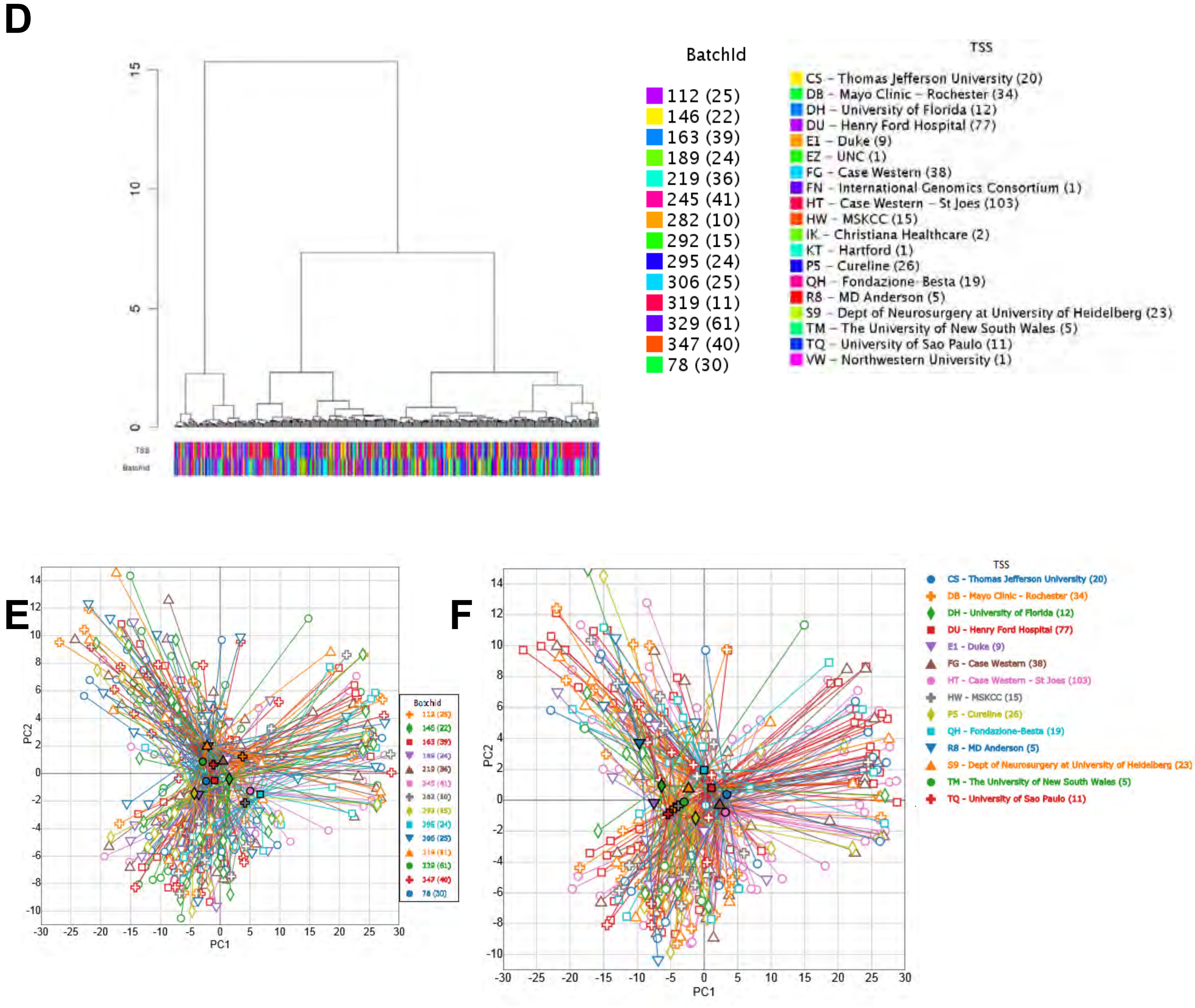


Figure S24. D. Hierarchical clustering plot for DNA methylation data. E. PCA for DNA methylation, with samples connected by centroids according to batch ID.Fig. F. PCA for DNA methylation, with samples connected by centroids according to TSS.

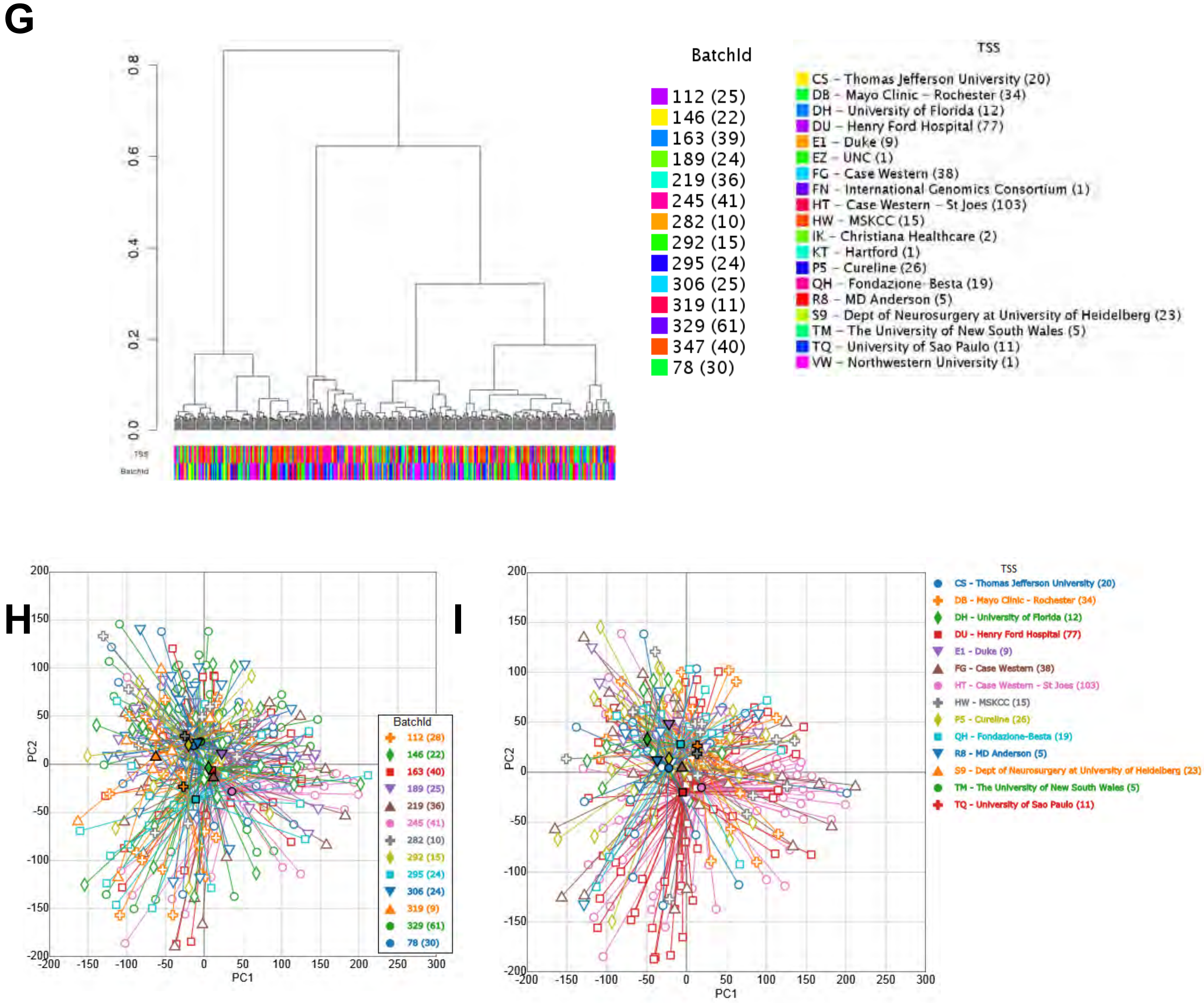


Figure S24. G. Hierarchical clustering for mRNA expression from RNA-seq data. H. PCA: First two principal components for RNA-seq, with samples connected by centroids according to batch ID. I. PCA: First two principal components for RNA-seq, with samples connected by centroids according to TSS.

Back to Table of Contents

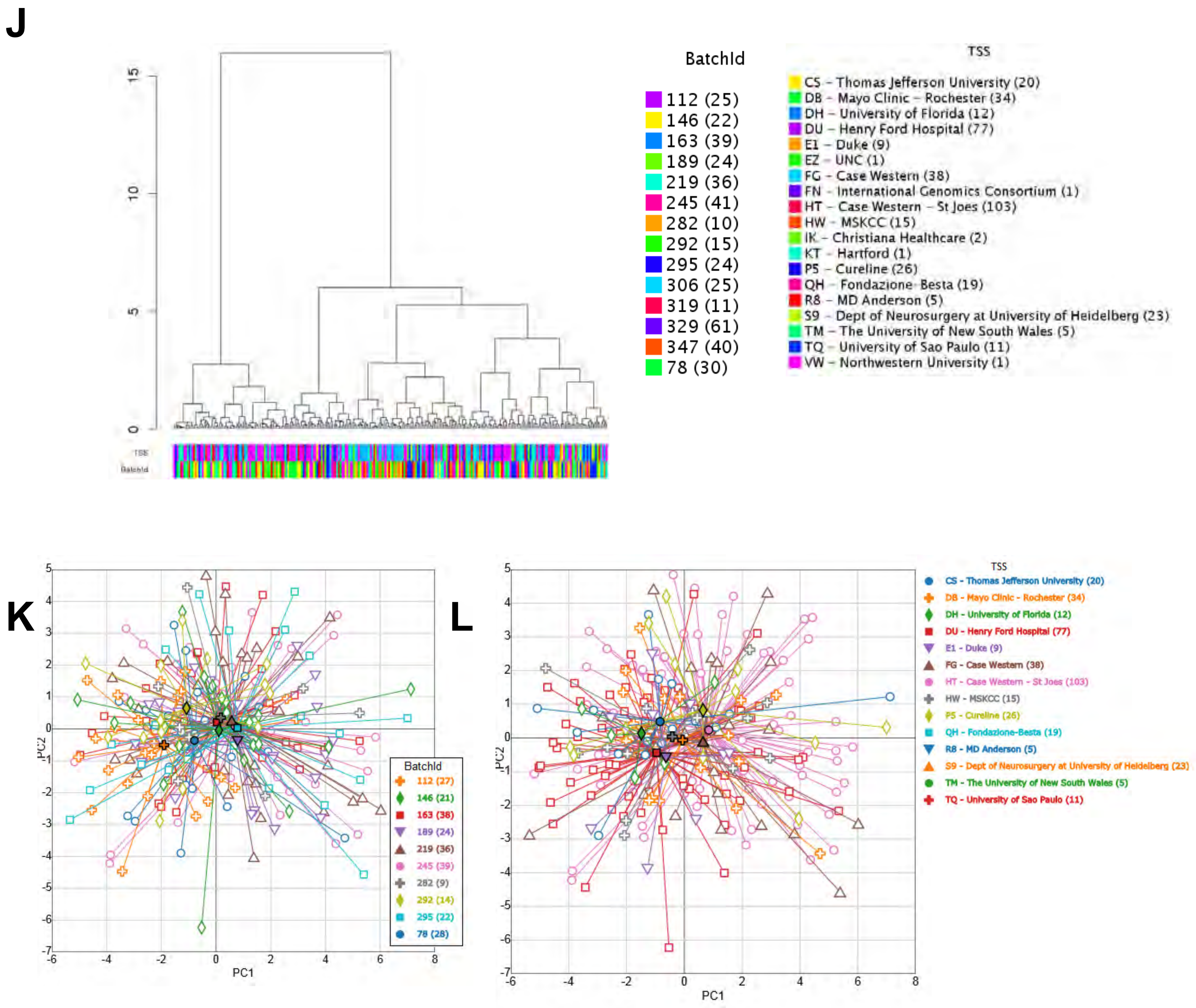


Figure S24. J. Hierarchical clustering of RPPA data. K. PCA: First two principal components for RPPA, with samples connected by centroids according to batch ID. L. PCA: First two principal components for RPPA, with samples connected by centroids according to TSS.

Back to Table of Contents

Supplemental Tables*

Table	#
LGG Master Table*	**S1**
Clinical associations	**S2**
GISTIC regions*	**S3**
Significantly mutated genes	**S4**
Double minutes*	**S5**
Transcript Fusions*	**S6**
mRNA cluster annotations*	**S7**
miRNA associations*	**S8**
miRNA-RPPA anti-correlations*	**S9**

*See online links to Excel spreadsheets for Tables S1, S3, S5-9.

Supplemental Table Legends
(see online links to Excel spreadsheets for Tables S1, S3, S5-9)

Table S1. Patient summary table. Sheet A contains patient clustering assignments for messenger RNA, DNA methylation, micro RNA, DNA copy number, reverse-phase protein array, Oncosign and cluster-of-clusters. Sheet B contains genetic information including *IDH*/1p19q molecular subtype, and mutation status for the *TERT* promoter and genes selected by MutSig analysis. Sheet C summarizes the availability of genomic platforms by patient. Sheet D contains clinical and demographic data. Sheet E provides a key describing the clinical and demographic fields used in Sheet D (see online link to spreadsheet).

Table S2. Clinical data, including clinical characteristics of the sample set (**S2A**), overall survival models (**S2B**), progression free survival models (**S2C**), hazard ratios for LGG and GBM (**S2D**), and adjusted Rand index scores for comparison of *IDH*/codel status, histology, and histology with grade to molecular platform clusters (**S2E**).

Table S3. Results of GISTIC by molecular subtype. Amplifications and deletions are shown for all tumors and for molecular subclasses based on *IDH* and 1p/19q status (see online link to spreadsheet).

Table S4A: Significantly mutated genes with q<0.1 in all LGG identified with the MutSig2CV algorithm (Lawrence et al, 2014).
Table S4B: Significantly mutated genes with q<0.1 in *IDH*wt LGG identified with the MutSig2CV algorithm (Lawrence et al, 2014).
Table S4C: Significantly mutated genes with q<0.1 in *IDH*mut-codel LGG identified with the MutSig2CV algorithm (Lawrence et al, 2014).
Table S4D: Significantly mutated genes with q<0.1 in *IDH*mut-non-codel LGG identified with the MutSig2CV algorithm (Lawrence et al, 2014).

Table S5. Results from BamBam and BreakDancer analysis pipelines to identify tumor-specific DNA rearrangements and evidence for complex amplicons associated with highly amplified tumor DNA segments, including features of potential double minute chromosomes/homogeneously staining regions (see online link to spreadsheet)..

Table S6. Overview of transcript fusions (see online link to spreadsheet).

Table S7. Annotation of RNA clusters. Lists of differentially expressed genes for each subtype were determined using 2-class SAM (11309499) in 1 versus rest comparisons. Functional Annotation Clustering was determined using DAVID Bioinformatics Resources 6.7 (19033363 and 19131956) where the inputs for each RNA-seq expression subtype were the 2000 most highly differentially expressed genes as determined by the SAM outputs (see online link to spreadsheet).

Table S8. miRs that are differentially abundant between molecular subtypes. These results expand on Figure S10. The spreadsheet gives detailed results for miRs (i.e. 5p and 3p mature strands) that are differentially abundant between pairs of molecular subtypes (worksheets 1 to 6), or between samples in one subtype and samples in the other two subtypes (worksheets 7 to 12). The Contents worksheet lists the six comparisons. Each comparison has two worksheets: miRs that are more abundant (UP), and less abundant (DOWN), in one of the two sample groups. 1,2) *IDH*mut-codel vs *IDH*mut-non-codel; 3,4) *IDH*mut-codel vs. *IDH*wt; 5,6) *IDH*wt vs.

*IDH*mut-non-codel; 7,8) *IDH*mut-codel vs. other tumor samples; 9,10) *IDH*mut-non-codel vs. other tumor samples; 11,12) *IDH*wt vs. other tumor samples. The README worksheet describes the method used, and the columns in each worksheet. To allow a user to choose either group as the reference, each result worksheet has 'fold_change' and '-1/fold_change' columns (see online link to spreadsheet).

Table S9. miR-protein anticorrelations and differentially abundant miRs. a) Five tabs give all significant miR-antibody anticorrelations (FDR<0.05), then anticorrelations filtered by functional validation publications using miRTarBase v4.5, then anticorrelations filtered by TargetScan v6.2 conserved and nonconserved target predictions. b) For miRs that are differentially abundant for each molecular subtype vs. all other tumour samples, three additional tabs report miR-antibody anticorrelations that are supported by miRTarBase publications with strong evidence types. c) For the 22 antibodies from (b), one tab reports BH-corrected P-values from Wilcoxon and KS tests for an antibody being differentially abundant in each molecular subtype (see online link to spreadsheet).

TABLE S2

Table S2A.i: Clinical characteristics of the sample set. Clinical characteristics are presented in total and by IDH mutant status and 1p/19q co-deletion status within the *IDH* mutant group. The number of samples with known information is given when complete data are not available. Count, and percentage within group, is given for categorical variables and distributions are compared by Fisher's exact test. Mean, SD, and range are given for age at diagnosis and ANOVA was used to compare groups. Significance is noted as ‡ 0.10<p<0.05, # p<0.05, ## p<0.01.

		Total (n=278£)	***IDH*mut codel (N=84)**	***IDH*mut non-codel (N=139)**	***IDH*wt (N=55)**
Histological Type ## and Grade ##	Oligodendroglioma II	65 (23%)	**38 (45%)**	21 (15%)	6 (11%)
	Oligodendroglioma III	44 (16%)	**31 (37%)**	6 (4%)	**7 (13%)**
	Oligoastrocytoma II	41 (15%)	9 (11%)	30 (22%)	2 (4%)
	Oligoastrocytoma III	33 (12%)	4 (5%)	**20 (14%)**	**9 (16%)**
	Astrocytoma II	30 (11%)	1 (1%)	24 (17%)	5 (9%)
	Astrocytoma III	65 (23%)	1 (1%)	38 (27%)	**26 (47%)**
Age (Yrs) at Diagnosis ##	Mean (SD)	42.6 (13.5)	**45.4 (13.2)**	**38.1 (10.9)**	**49.9 (15.3)**
	Min, Max	14, 75	17, 75	14, 70	21, 74
Gender	Male	155 (56%)	45 (54%)	84 (60%)	26 (47%)
Race (self-report, n=274)	White	261 (95%)	79 (98%)	131 (95%)	51 (93%)
Ethnicity (self-report, n=261)	Hispanic/Latino	14 (5%)	5 (6%)	6 (5%)	3 (6%)
Treating Country	United States*	265 (95%)	81 (96%)	131 (94%)	53 (96%)
Year of Diagnosis	Before 2005	38 (14%)	10 (12%)	18 (13%)	10 (18%)
	2005-2009	88 (32%)	30 (36%)	44 (32%)	14 (25%)
	2010-2013	152 (55%)	44 (52%)	77 (55%)	31 (56%)
Family History of Cancer (n=190**)#	No History	108 (56%)	30 (52%)	64 (65%)	13 (38%)
	Primary Brain	11 (6%)	2 (3%)	7 (7%)	**2 (6%)**
	Other Cancers	72 (38%)	26 (45%)	27 (28%)	**19 (56%)**
Extent of resection (n=268)	Open Biopsy	6 (2%)	1 (1%)	4 (3%)	1 (2%)
	Subtotal Resection	98 (37%)	31 (38%)	45 (34%)	22 (40%)
	Gross Total Resection	164 (61%)	49 (60%)	83 (63%)	32 (58%)
Tumor Location##	Frontal Lobe	172 (62%)	**68 (81%)**	**84 (60%)**	20 (36%)
	Parietal Lobe	23 (8%)	5 (6%)	13 (9%)	5 (9%)
	Temporal Lobe	74 (27%)	9 (11%)	40 (29%)	**25 (45%)**
	Other+	9 (3%)	2 (2%)	2 (1%)	5 (9%)
Laterality (n=276)	Left	133 (48%)	37 (44%)	69 (50%)	27 (49%)
	Midline	5 (2%)	2 (2%)	2 (1%)	1 (2%)
	Right	138 (50%)	45 (54%)	66 (48%)	27 (49%)
White vs. Grey Matter (n=144)	White Matter	74 (51%)	26 (54%)	37 (51%)	11 (46%)
First Presenting Symptom (n=252)	Headache	64 (25%)	15 (21%)	39 (30%)	10 (20%)
	Mental Status	22 (9%)	7 (10%)	10 (8%)	5 (10%)
	Motor/Movement	18 (7%)	6 (8%)	7 (5%)	5 (10%)
	Seizure	135 (54%)	38 (53%)	70 (54%)	27 (53%)
	Sensory	6 (2%)	3 (4%)	1 (1%)	2 (4%)
	Visual	7 (3%)	3 (4%)	2 (2%)	2 (4%)
Primary Radiotherapy (n=166)#	Yes	122 (73%)	**27 (59%)**	**66 (76%)**	**29 (88%)**
Primary Pharmacotherapy (n=153)	Yes	95 (62%)	27 (61%)	49 (60%)	19 (70%)
Preop. Antiseizure Med. (n=202)	Yes	148 (73%)	48 (75%)	73 (72%)	27 (75%)
Preop. Corticosteroids (n=207)	Yes	90 (43%)	29 (43%)	46 (44%)	15 (43%)

£ Eleven cases with clinical information do not have IDH/Codel status determined

* Twelve cases were submitted from Russia (3 IDH$_{\text{mut-codel}}$, 7 IDH$_{\text{mut-non-codel}}$, 2 IDH$_{\text{wt}}$), and three case from Italy (1 IDH$_{\text{mut-non-codel}}$, 2 unknown IDH/codel)

** Cases with response to both questions of family history of any cancer (n=192) and of family history of primary brain cancer (n=197)

+ One case (IDH$_{\text{wt}}$) was in the cerebellum, three cases were in the occipital lobe (2 IDH$_{\text{mut-codel}}$, 1 IDH$_{\text{mut-non-codel}}$) and five cases were listed as "supratentorial, not otherwise specified" (1 IDH$_{\text{mut-non-codel}}$, 4 IDH$_{\text{wt}}$)

Table S2A.ii: Clinical characteristics of the sample set. Clinical characteristics are presented in total and by histological type and WHO grade. The number of samples with known information is given when complete data are not available. Count, and percentage within group, is given for categorical variables and distributions are compared by Fisher's exact test. Mean, standard deviation, and range are given for age at diagnosis and ANOVA was used to compare groups. Significance is noted as ǂ 0.10<p<0.05, # p<0.05, ## p<0.01.

			Astrocytoma		**Oligoastrocytoma**		**Oligdendroglioma**	
		Total	**II**	**III**	**II**	**III**	**II**	**III**
		(N=289)	**(N=31)**	**(N=67)**	**(N=43)**	**(N=34)**	**(N=70)**	**(N=44)**
IDH/Codel Group (n=278)$^{\#\#}$	*IDH*mut-codel	84 (30%)	1 (3%)	1 (2%)	9 (22%)	4 (12%)	**38 (58%)**	**31 (70%)**
	*IDH*mut-non-codel	138 (50%)	**24 (80%)**	38 (58%)	**30 (73%)**	**20 (61%)**	21 (32%)	6 (14%)
	*IDH*wt	56 (20%)	5 (17%)	**26 (40%)**	2 (5%)	9 (27%)	6 (9%)	7 (16%)
Age (Yrs) at Diagnosis$^{\#\#}$	Mean (SD)	42.7 (13.5)	36.8 (11.6)	**45.0 (12.8)**	38.4 (12.3)	**44.2 (14.6)**	41.7 (13.1)	**47.9 (14.1)**
	Min, Max	14, 75	20, 62	22, 74	14, 69	23, 67	17, 70	23, 75
Gender	Male	160 (55%)	17 (55%)	37 (55%)	22 (51%)	21 (62%)	37 (53%)	26 (59%)
Race (self-report, n=285)	White	272 (95%)	30 (97%)	61 (92%)	40 (95%)	34 (100%)	63 (93%)	44 (100%)
Ethnicity (self-report, n=272)	Hispanic/Latino	14 (5%)	3 (11%)	1 (2%)	2 (5%)	1 (3%)	4 (6%)	3 (7%)
Treating Country$^{\#}$	United States[*]	274 (95%)	**26 (84%)**	65 (97%)	42 (98%)	34 (100%)	65 (93%)	42 (95%)
Year of Diagnosis	Before 2005	41 (14%)	2 (6%)	15 (22%)	3 (7%)	3 (9%)	12 (17%)	6 (14%)
	2005-2009	90 (31%)	8 (26%)	20 (340)	14 (33%)	14 (41%)	19 (27%)	15 (34%)
	2010-2013	158 (55%)	21 (68%)	32 (48%)	26 (60%)	17 (50%)	39 (56%)	23 (52%)
Family History of Cancer (n=196[**])	No History	110 (55%)	12 (67%)	18 (46%)	22 (59%)	16 (73%)	27 (57%)	15 (45%)
	Primary Brain	11 (6%)	1 (6%)	4 (10%)	3 (8%)	0 (0%)	1 (2%)	2 (6%)
	Other Cancers	75 (39%)	5 (28%)	17 (44%)	12 (32%)	6 (27%)	19 (40%)	16 (48%)
Extent of resection (n=255) ǂ	Open Biopsy	6 (2%)	0 (0%)	1 (2%)	3 (7%)	**1 (3%)**	1 (1%)	0 (0%)
	Subtotal Resection	104 (37%)	7 (23%)	24 (37%)	12 (29%)	**17 (57%)**	28 (41%)	16 (36%)
	Gross Total Resection	169 (61%)	23 (77%)	40 (62%)	27 (64%)	12 (40%)	39 (57%)	28 (64%)
Tumor Location ǂ (between histological types ǂ)	Frontal Lobe	178 (62%)	16 (52%)	35 (52%)	27 (63%)	19 (56%)	47 (67%)	34 (77%)
	Parietal Lobe	24 (8%)	2 (6%)	11 (16%)	3 (7%)	1 (3%)	6 (9%)	1 (2%)
	Temporal Lobe	78 (27%)	10 (32%)	20 (30%)	12 (28%)	13 (38%)	16 (23%)	7 (16%)
	Other[+]	9 (3%)	3 (10%)	1 (1%)	1 (2%)	1 (3%)	1 (1%)	2 (5%)
Laterality (n=286)	Left	136 (48%)	15 (48%)	35 (53%)	20 (48%)	16 (47%)	33 (48%)	17 (39%)
	Midline	5 (2%)	1 (3%)	1 (2%)	0 (0%)	1 (3%)	0 (0%)	2 (5%)
	Right	145 (51%)	15 (48%)	30 (45%)	22 (52%)	17 (50%)	36 (52%)	25 (57%)
White vs. Grey Matter (n=151)	White Matter	79 (52%)	12 (67%)	12 (36%)	11 (46%)	8 (42%)	22 (61%)	14 (67%)
First Presenting Symptom (n=261) (seizure vs other$^{\#\#}$)	Headache	64 (25%)	10 (38%)	13 (21%)	4 (11%)	12 (36%)	13 (21%)	12 (29%)
	Mental Status	22 (8%)	3 (12%)	8 (13%)	2 (5%)	1 (3%)	4 (6%)	4 (10%)
	Motor/Movement	20 (8%)	3 (12%)	5 (8%)	2 (5%)	2 (6%)	3 (5%)	5 (12%)
	Seizure	**142 (54%)**	**9 (35%)**	**34 (55%)**	**28 (76%)**	**16 (48%)**	**38 (61%)**	**17 (41%)**
	Sensory	6 (2%)	0 (0%)	1 (2%)	0 (0%)	1 (3%)	3 (5%)	1 (2%)
	Visual	7 (3%)	1 (4%)	1 (2%)	1 (3%)	1 (3%)	1 (2%)	2 (5%)
Primary Radiotherapy (n=174) by grade$^{\#\#}$	Yes	126 (72%)	8 (57%)	**41 (87%)**	13 (46%)	**25 (96%)**	16 (52%)	**23 (82%)**
Primary Pharmacotherapy (n=161) by grade$^{\#\#}$	Yes	101 (63%)	5 (45%)	**31 (70%)**	6 (25%)	**21 (88%)**	16 (55%)	**22 (76%)**
Preop. Antiseizure Med. (n=210)	Yes	154 (73%)	10 (53%)	32 (70%)	28 (82%)	20 (74%)	40 (80%)	24 (71%)
Preop. Corticosteriods (n=215)	Yes	95 (44%)	6 (27%)	24 (51%)	13 (39%)	14 (52%)	24 (45%)	14 (42%)

* Twelve cases were submitted from Russia and three cases were submitted from Italy

** Cases with response to both questions of family history of any cancer (n=198) and of family history of primary brain cancer (n=204)

+ One case (Astrocytoma-II) was in the cerebellum, three cases were in the occipital lobe, and five cases were listed as "supratentorial, not otherwise specified"

Table S2B. Overall Survival Models Cox regression models of overall survival considering age at diagnosis, extent of resection, histological type, WHO grade, and IDH/Codel group as single predictors and in combination in multiple-predictor models. Area under the survival ROC is provided as a measure of predictive ability for each model at 1 year past diagnosis. Bold denotes hazard ratios significantly different from 1.0 (no difference).

		Single Predictor Models		**Multi-predictor Models**		
				Model I (AUC=0.80)	**Model II (AUC=0.85)**	**Model III (AUC=0.87)**
Predictor	**Levels**	**HR (95% CI)**	**AUC**	**HR (95% CI)**	**HR (95% CI)**	**HR (95% CI)**
Age at Diagnosis	(per 5 yrs)	**1.38 (1.24, 1.53)**	0.72	**1.42 (1.27, 1.60)**	**1.36 (1.20, 1.53)**	**1.37 (1.20, 1.55)**
Extent of Resection	Gross Total	1.0 (ref)	0.57	1.0 (ref)	1.0 (ref)	1.0 (ref)
	< Gross Total	1.71 (0.97, 3.01)		1.09 (0.60, 1.98)	1.64 (0.90, 3.00)	1.62 (0.87, 3.02)
Histological Type	Astrocytoma	**2.26 (1.26, 4.08)**	0.60	1.80 (0.96, 3.36)	--	1.79 (0.80, 4.03)
	Oligoastrocytoma	1.13 (0.56, 2.26)		1.88 (0.91 3.86)	--	1.45 (0.62, 3.35)
	Oligodendroglioma	1.0 (ref)		1.0 (ref)	--	1.0 (ref)
WHO Grade	II	1.0 (ref)	0.64	1.0 (ref)	--	1.0 (ref)
	III	**3.36 (1.89, 5.98)**		**3.60 (1.87, 6.90)**	--	**2.79 (1.40, 5.58)**
IDH/Codel group	IDHmut-codel	1.0 (ref)	0.72	--	1.0 (ref)	1.0 (ref)
	IDHmut-non-codel	1.32 (0.64, 2.71)		--	1.80 (0.87, 3.73)	1.47 (0.60, 3.63)
	IDHwt	**9.22 (4.36, 19.52)**		--	**11.22 (4.86, 25.90)**	**6.67 (2.66, 16.72)**

Table S2C. Progression Free Survival Models Cox regression models of progression-free survival considering age at diagnosis, extent of resection, histological type, WHO grade, and IDH/Codel group as single predictors and in combination in multiple-predictor models. Area under the survival ROC is provided as a measure of predictive ability for each model at 1 year past diagnosis. Bold denotes hazard ratios significantly different from 1.0 (no difference).

		Single Predictor Models		**Multi-predictor Models**		
				Model I (AUC=0.65)	**Model II (AUC=0.74)**	**Model III (AUC=0.74)**
Predictor	**Levels**	**HR (95% CI)**	**AUC**	**HR (95% CI)**	**HR (95% CI)**	**HR (95% CI)**
Age at Diagnosis	(per 5 yrs)	**1.14 (1.05, 1.24)**	0.59	**1.15 (1.05, 1.25)**	1.10 (1.00, 1.20)	1.10 (1.00, 1.20)
Extent of Resection	Gross Total	1.0 (ref)	0.53	1.0 (ref)	1.0 (ref)	1.0 (ref)
(n=240, 85events)	< Gross Total	1.25 (0.80,1.97)		1.14 (0.72, 1.82)	1.24 (0.78, 1.99)	1.24 (0.77, 2.00)
Histological Type	Astrocytoma	1.60 (0.97, 2.64)	0.55	**1.77 (1.01, 3.11)**	--	1.06 (0.57, 1.98)
	Oligoastrocytoma	1.09 (0.62, 1.90)		1.70 (0.94, 3.08)	--	1.33 (0.71, 2.48)
	Oligodendroglioma	1.0 (ref)		1.0 (ref)	--	1.0 (ref)
WHO Grade	II	1.0 (ref)	0.56	1.0 (ref)	--	1.0 (ref)
	III	**1.58 (1.02, 244)**		**1.69 (1.04, 2.74)**	--	1.27 (0.77, 2.09)
IDH/Codel group	IDHmut-codel	1.0 (ref)	0.70	--	1.0 (ref)	1.0 (ref)
(n=240, 86evt)	IDHmut-non-codel	1.48 (0.83, 2.63)		--	**2.02 (1.08, 3.76)**	1.90 (0.95, 3.80)
	IDHwt	**8.89 (4.66, 16.97)**		--	**9.17 (4.56, 18.45)**	**8.30 (3.89, 17.68)**

Table S2D. Hazard Ratios for LGG and GBM Hazard ratios from Cox Proportional Hazards models with and without adjusting for age at diagnosis. Hazard ratios and 95% confidence intervals are given for each pair of groups. Bold denotes hazard ratios significantly different from 1.0 (no difference).

	Hazard Ratio (95% CI)	**Age Adjusted Hazard Ratio (95% CI)**
IDHmut-non-codel vs IDHmut-codel	1.29 (0.63, 2.66)	1.53 (0.74, 3.14)
IDHwt vs IDHmut-codel	**7.38 (3.57, 15.24)**	**7.02 (3.39, 14.55)**
GBMmut vs IDHmut-codel	**4.91 (2.11, 11.44)**	**3.96 (1.69, 9.29)**
GBMwt vs IDHmut-codel	**13.64 (7.33, 25.38)**	**9.11 (4.85, 17.13)**
IDHwt vs IDHmut-non-codel	**5.71 (3.20, 10.19)**	**4.60 (2.56, 8.27)**
GBMmut vs IDHmut-non-codel	**3.80 (1.84, 7.82)**	**2.60 (1.25, 5.42)**
GBMwt vs IDHmut-non-codel	**10.55 (6.81, 16.34)**	**5.97 (3.76, 9.49)**
GBMmut vs IDHwt	0.67 (0.32, 1.36)	0.56 (0.27, 1.17)
GBMwt vs IDHwt	**1.85 (1.21, 2.83)**	1.30 (0.85, 1.99)
GBMwt vs GBMmut	**2.78 (1.52, 5.09)**	**2.30 (1.24, 4.27)**

Table S2E. Adjusted Rand Index Scores. Comparisons were made between the IDH/Codel classification and each of the molecular clustering solutions as well as with histology, and histology-grade groups, using the Adjusted Rand Index. Identical classification will have a score of 1.0 whereas no similarity beyond chance has an expected score of 0. A 95% CI for the ARI estimate is given as derived from 1000 bootstrap samples. To allow direct comparison,, only samples with classification in all 10 schemes are retained in this analysis (N=209). Bold denotes ARI of 0.5 or greater.

Adjusted Rand Index (ARI)	**IDH/Codel**	**Histology**	**Histology/Grade**
vs. IDH/Codel	**1 (1, 1)**	0.2 (0.12, 0.29)	0.12 (0.08, 0.18)
vs. RNA Clusters	0.30 (0.23, 0.40)	0.09 (0.05, 0.16)	0.09 (0.06, 0.16)
vs.DNA Methylation Clusters	**0.52 (0.42, 0.60)**	0.11 (0.07, 0.18)	0.09 (0.06, 0.15)
vs. miRNA Clusters	0.14 (0.08, 0.20)	0.03 (0, 0.08)	0.02 (0, 0.06)
vs. DNA CN Clusters	**0.73 (0.63, 0.83)**	0.20 (0.12, 0.30)	0.11 (0.07, 0.18)
vs. RPPA Clusters	0.09 (0.06, 0.16)	0.04 (0.01, 0.09)	0.03 (0.02, 0.08)
vs. Cluster of Clusters	**0.79 (0.69, 0.88)**	0.19 (0.12, 0.29)	0.12 (0.09, 0.19)
vs. OncoSign Clusters	**0.83 (0.75, 0.89)**	0.15 (0.09, 0.23)	0.09 (0.06, 0.15)
vs. Histology	0.20 (0.12, 0.29)	**1 (1, 1)**	**0.60 (0.57, 0.64)**
vs. WHO Grade	0.04 (0.01, 0.10)	0.04 (0.01, 0.10)	0.36 (0.34, 0.41)

TABLE S4

Gene	Number of patients with mutation	Number of mutated sites	p-value	q-value
TP53	146	91	1.00E-16	4.56E-13
CIC	49	43	1.00E-16	4.56E-13
NOTCH1	22	22	1.00E-16	4.56E-13
IDH2	12	3	1.00E-16	4.56E-13
IDH1	221	2	5.55E-16	2.03E-12
ATRX	112	106	2.44E-15	7.43E-12
FUBP1	22	22	2.03E-14	5.30E-11
NF1	16	20	3.46E-14	7.90E-11
PTEN	12	12	5.47E-13	1.11E-09
PIK3R1	11	10	7.00E-13	1.28E-09
PIK3CA	23	15	3.63E-11	6.02E-08
EGFR	16	12	2.23E-10	3.40E-07
ARID1A	9	9	3.35E-08	4.70E-05
TCF12	6	6	6.17E-08	8.05E-05
SMARCA4	12	10	8.50E-08	1.03E-04
ZBTB20	10	9	1.01E-06	1.16E-03
PLCG1	4	3	1.47E-06	1.58E-03
PTPN11	5	4	5.42E-05	5.50E-02
ZCCHC12	4	2	6.03E-05	5.80E-02

Table S4A: Significantly mutated genes with q<0.1 in **all LGG** identified with the MutSig2CV algorithm (Lawrence et al, 2014). Discovery analysis based on consensus MAF (see Methods for details).

Gene	Number of patients with mutation	Number of mutated sites	p-value	q-value
PTEN	12	12	3.57E-14	6.53E-10
TP53	7	12	1.06E-12	9.68E-09
EGFR	16	12	2.34E-12	1.42E-08
NF1	10	14	1.72E-09	7.85E-06
PLCG1	3	2	3.07E-07	1.12E-03
PTPN11	4	3	2.17E-06	6.59E-03
PIK3CA	5	5	1.39E-05	3.62E-02

Table S4B: Significantly mutated genes with q<0.1 in **IDHwt LGG** identified with the MutSig2CV algorithm (Lawrence et al, 2014). Discovery analysis based on consensus MAF (see Methods for details).

Gene	Number of patients with mutation	Number of mutated sites	p-value	q-value
IDH1	77	1	1.00E-16	6.75E-13
CIC	46	41	1.00E-16	6.75E-13
IDH2	8	2	1.11E-16	6.75E-13
NOTCH1	18	19	4.44E-16	2.03E-12
FUBP1	22	22	3.77E-15	1.38E-11
PIK3CA	16	11	2.17E-10	6.60E-07
PIK3R1	6	6	6.42E-10	1.67E-06
ZBTB20	7	6	5.96E-08	1.36E-04
TCF12	4	5	1.06E-07	2.15E-04
ZCCHC12	3	1	7.10E-06	1.30E-02
ARID1A	5	5	9.12E-06	1.51E-02
TP53	3	3	4.60E-05	6.90E-02

Table S4C: Significantly mutated genes with q<0.1 in **IDHmut-codel LGG** identified with the MutSig2CV algorithm (Lawrence et al, 2014). Discovery analysis based on consensus MAF (see Methods for details).

Gene	Number of patients with mutation	Number of mutated sites	p-value	q-value
TP53	133	82	1.00E-16	6.75E-13
ATRX	107	100	1.00E-16	6.75E-13
IDH1	137	2	1.11E-16	6.75E-13
SMARCA4	8	6	2.64E-09	1.21E-05
IDH2	4	3	6.27E-08	2.29E-04
EIF1AX	3	2	3.25E-05	9.90E-02

Table S4D: Significantly mutated genes with q<0.1 in **IDHmut-non-codel LGG** identified with the MutSig2CV algorithm (Lawrence et al, 2014). Discovery analysis based on consensus MAF (see Methods for details).