Published in final edited form as:

Clin Cancer Res. 2012 October 15; 18(20): 5595–5605. doi:10.1158/1078-0432.CCR-12-1427.

Genomic estimates of aneuploid content in Glioblastoma Multiforme and improved classification

Bo Li[1], **Yasin Senbabaoglu**[1], **Weiping Peng**[2], **Min-lee Yang**[2], **Jishu Xu**[2], and **Jun Z. Li**[1,2,*]

[1]Bioinformatics Program, University of Michigan, Ann Arbor, MI, USA

[2]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

Abstract

Purpose—Accurate classification of Glioblastoma Multiforme (GBM) is crucial for understanding its biological diversity, and informing diagnosis and treatment. The Cancer Genome Atlas (TCGA) project identified four GBM classes using gene expression data, and separately, identified three classes using methylation data. We sought to integrate multiple data types in GBM classification, understand biological features of the newly defined subtypes, and reconcile with prior studies.

Experimental procedure—We used allele-specific copy number data to estimate the aneuploid content of each tumor, and incorporated this measure of intratumor heterogeneity in class discovery. We estimated the potential cell of origin of individual subtypes and the euploid and aneuploid fractions using reference datasets of known neuronal cell types.

Results—There exists an unexpected correlation between aneuploid content and the observed among-tumor diversity of expression patterns. Joint use of DNA and mRNA data in *ab initio* class discovery revealed a distinct group that resembles the Proneural subtype described in a separate study and the G-CIMP+ class based on methylation data. Three additional subtypes, Classical, Proliferative, and Mesenchymal, were also identified, and revised the assignment for many samples. The revision showed stronger differences in patient outcome and clearer cell type-specific signatures. Mesenchymal GBMs had higher euploid content, potentially contributed by microglia/macrophage infiltration.

Conclusion—We clarified the confusion regarding the "Proneural" subtype that was defined differently in different prior studies. The ability to infer within-tumor heterogeneity improved class discovery, leading to new subtypes that are closer to the fundamental biology of GBM.

Keywords

GBM; classification; aneuploid content; survival time; data integration

INTRODUCTION

Glioblastoma Multiforme (GBM) is an aggressive brain tumor with poor prognosis (1). Recently, genomic profiling studies have provided rich new information for understanding molecular lesions in GBM. For example, the Cancer Genome Atlas (TCGA) project characterized several hundred GBM samples, of which many were analyzed across multiple dimensions, including single nucleotide polymorphism (SNP) genotyping, mRNA and

[*]Corresponding Author: Jun Z. Li, Ph.D., Department of Human Genetics, 5789A Medical Science II, Box 5618, University of Michigan, Ann Arbor, MI 48019, Phone: 734-615-5754, Fax: 734-763-3784, junzli@umich.edu.

The authors declare no conflict of interest.

microRNA (miRNA) profiling, DNA sequencing, and promoter methylation analysis (2). These data highlighted the importance of *ERBB2*, *NF1* and *TP53* genes, and revealed recurrent aberrations in the *RTK/RAS/PI(3)K*, *p53*, and *RB* signaling pathways. Meanwhile, genomewide datasets are also useful for characterizing biological diversity in a tumor collection, as evidenced by numerous reports of molecular subtypes for many cancers based on gene expression cluster analyses (3, 4). In particular, gene expression data for TCGA's first GBM cohort were reported to reveal four subclasses (5): Proneural (PN), Neural (NL), Classical (CL) and Mesenchymal (MES).

However, while the availability of multiple data types in TCGA provides the opportunity for combined analyses, the four-class model was based solely on mRNA expression data. DNA copy number alteration (CNA) patterns were summarized *post hoc*, not incorporated in the initial class discovery. Methylation data were analyzed subsequently (6), and revealed three clusters, which lacked a clear correspondence with the four transcriptome-based classes. Furthermore, the relationship of the four-class model with those previously reported for independent datasets (7–9) was not clarified. While the differences *between* studies could be explained by variations in sample selection criteria, experimental platforms, and analysis methods, the discrepancies among different data types *within* the TCGA's collection remained un-reconciled. The first goal of this work is therefore to combine the CNA and expression data to provide a more integrated view of the molecular diversity in GBM.

Our second goal is to study *within-tumor* heterogeneity. Surgically obtained solid tumor samples (GBM included) often contain both aneuploid cells and euploid cells. We developed a method to leverage the allele-specific CNA data to estimate the fraction of aneuploid cells in each sample, and to incorporate this measure of tumor "purity" in class discovery. We asked if results in GBM were also seen in the ovarian (OV) cancer cohort from TCGA (10). We emphasized between-cohort concordance in deciding the optimal number of clusters, and we annotated the potential cell type of origin of different classes by comparing GBM gene expression data to reference datasets of known cell types. Our results led to a revised framework of GBM classification, and we sought to understand its biological implication and clinical relevance. We validated the between-class difference in survival time in an independent GBM cohort. Finally, we summarized the newly recognized subclasses and associated biomarkers into a hierarchical classification protocol for use in diagnosis and further research.

MATERIALS AND METHODS

Datasets

The sources of the datasets, including (1) gene expression data for TCGA GBM (separately analyzed as 3 batches: GBM1–3), OV (10), Phillips et al. dataset for GBM (GEO accession GSE4271) (8), and the reference dataset by Cahoy et al. for neuronal cells (GEO accession GSE9956) (11), (2) CNA data for GBM1–3, OV, (3) methylation data for GBM1, and (4) clinical data for these studies, are described in Supplemental Information (SI-1.1–1.4).

Inferring Aneuploid Genome Proportion (AGP)

We performed logR and BAF-based segmentation and merged the two series of change points. For each segment we calculated the mean folded BAF and mean logR values, and used a least squared distance-based procedure to scan for the best fitting AGP values across the full range of canonical patterns (described in SI-2.1–2.5). The AGP values and associated confidence measures for each sample were extracted from the model as described in SI-2.6 and presented in Table S1.

Statistical analysis

K-means clustering, quantile normalization, t-tests, principal component analysis (PCA), and heatmap generation were performed using standard functions in R (12). Consensus Clustering was implemented using custom R codes. Survival analysis, including the log-rank test, Kaplan-Meier survival curves, Cox proportional hazard regression, and C-statistic (concordance measure) was performed using the R package *survival*. 3D scatter plots were generated using R package *scatterplot3d*. All scripts and processed data are available from the authors.

RESULTS

Genomic estimates of aneuploid-euploid mixing ratios

Allelic intensity data from SNP genotyping arrays provide quantitative copy number information of the two parental chromosomes: n_A and n_B. In a homogeneous cell population n_A and n_B are both integers, such that the logarithm of total intensity, $logR=log(n_A+n_B)$, and the observed B allele frequency, $BAF=n_B/(n_A+n_B)$, adopt a finite combination of discrete values, which can be shown as "canonical positions" in the BAF-LRR plot (Figure S1A). In a tumor sample, however, the population of aneuploid cells may be mixed with euploid cells, consequently logR and BAF of the former "contract" towards those of the latter; and different mixing ratios result in different degrees of contraction (Figure S1B). An example of such a mixed GBM sample is shown in Figure 1A. Based on this feature we developed an algorithm to quantitatively estimate genomewide mixing ratio from SNP data (see SI-2.1–2.5). By using experimental results for a series of aneuploid-euploid mixtures of known mixing ratios (GEO accession GSE11976) we confirmed that our method provides an unbiased estimate (Figure 1B, SI-2.7).

In this study we define *Aneuploid Content*, or synonymously, *Aneuploid Genomic Proportion (AGP)*, as the parameter p in a mixture model consisting of two homogeneous populations: (1) aneuploid cells, at the fraction of p, and (2) euploid cells, at (1-p). Euploid cells carry a balanced set of parental chromosomes representing full-integer multiples of the haploid genome, and may include normal stromal cells surrounding the tumors as well as tumor cells without apparent genomic aberrations (e.g., only point mutations). Aneuploid cells, in contrast, carry CNAs at some chromosomes or subchromosomal intervals, resulting in an unbalanced set of genomic segments, each of which still contain an integer combination of parental DNA, e.g., $n_A=2$, and $n_B=1$ in a region of amplification. For many tumors, the two-way mixing model considered here is likely an over-simplification, as multiple subpopulations of tumor cells may exist, each carrying a different integer combination of parental segments. However, a mixture model with three or more subpopulations is computationally intractable using the observed averages of the entire population; and realistically, many tumors may contain a dominant aneuploid population. A two-way mixing model is the simplest scenario that could have generated the observed data regarding varying levels of contraction in different samples. We therefore applied this model for the first-order estimation of within-tumor heterogeneity; and we reported the goodness-of-fit by several quantitative indices (SI-2.6). This approach allowed us to (1) examine the impact of estimated aneuploid content in downstream analyses such as class discovery, (2) assess the adequacy of the model *post hoc*, and (3) refine the analyses by focusing on the subset of samples that were adequately explained by this simple model.

In the first batch (GBM1), seven of 284 tumors had too few CNAs (including copy-neutral loss-of-heterozygosity events) for purity estimation, and were removed. The remaining 277 tumors had > 0.5% of the genome affected by CNAs, with an average Percent Changed (PC) of 37.3%, i.e., > 1/3 of the genome was altered in an "average" GBM. Across the 277 samples, the estimated AGPs ranged from 23% to 99% (mean ± SD: 76% ±17%), indicating significant admixture of euploid cells (average euploid content of 24%). To assess the goodness-of-fit for each sample we quantified the confidence interval (CI, 2.5–97.5%) of AGP and the fraction of CNAs that fall on canonical positions (PoP, Percent-on-Point) in the optimal two-way mixing model (SI-2.6, and Figure S1C–D). PoP values had a median of 92% among 277 GBMs, suggesting that it is indeed adequate to model a single dominant aneuploid cell population in most GBM samples.

Comparison of genomic estimates of aneuploid content with histopathologic reports

Histopathologic assessment of tumor purity provides basic information for clinical diagnosis, and is a key criterion in sample selection for research. In TCGA, for example, only GBM with >80% "tumor nuclei" were studied. We found, however, that aneuploid estimates based on SNP data were only moderately correlated with pathologists' report of "percent tumor cells" (Spearman's $\rho = 0.14$, $P = 0.02$, n=275), not correlated with "percent tumor nuclei" ($\rho = 0.076$, $P = 0.21$, n=275), and were lower than AGP by an average of 7% and 18%, respectively (Figure S2). The difference was not explained by tumors with worse fit in our model, or greater estimation uncertainty (SI-2.8). Our inferred AGP is therefore a novel feature extracted from molecular measurements, and can be complementary to the traditionally observed tumor purity.

Impact of aneuploid content on gene expression patterns

We examined 128 GBM1 samples with both gene expression and CNA data. First, samples of low AGP tend to cluster together in PCA of gene expression data, driving a strong correlation between the first principal component scores (PC1) and AGP (Pearson’s $r = 0.62$, $P = 7.3\times10^{-15}$, n = 128) (Figure 1C). PC2 was also correlated with AGP ($r = 0.48$, $P = 1.1\times10^{-8}$). This pattern suggests that within-tumor heterogeneity is a major driver of gene expression variation, and a factor overlooked in most previous studies. To see if the results for GBM extend to other tumor types, we applied a similar analysis to SNP and expression data for 509 ovarian (OV) tumors from TCGA (10), and observed a similar pattern (Figure S3), with a strong correlation between AGP and PC1 ($r = 0.56$, $P < 2.2\times10^{-16}$, n = 504). In contrast to AGP, clinically recorded purity values showed little correlation with PC1, r was 0.004 ($P = 0.96$) for "tumor nuclei", and 0.14 ($P = 0.10$) for "tumor cells", thus underscoring a key advantage of empirical measures of intra-tumor heterogeneity (13). Similar to mRNA, expression patterns of 504 microRNAs were also correlated with AGP ($r = -0.26$, $P = 3.0\times10^{-3}$ for PC1; $r = -0.56$, $P = 1.7\times10^{-11}$ for PC2, n=125).

Combined use of DNA and mRNA patterns in class discovery

The results above raised the question of whether varying levels of euploid-aneuploid mixing could affect the detection of tumor subtypes. To answer this, we performed a joint classification analysis of DNA and mRNA data. In PCA of DNA copy number data, high-AGP samples had high and low PC1s, flanking low-AGP samples (Figure S4A), and this was mostly due to a split of Proneural samples (colored purple). Interestingly, PC1 for copy number and PC1 for expression data, when plotted together, showed a clear separation of two groups (Figure 1D), which, due to annotation efforts described below, we will call Non-Proneural and Proneural samples (even though the Proneural group defined here only partially overlaps with the previously defined Proneural group (5)). The two groups were not readily separable when either dataset was analyzed by itself. MicroRNA PC2 was highly correlated with PC1 of mRNA data (not shown); thus the joint use of this quantity with copy number PC1 also separated the two groups (Figure S4B). The Proneural class consisted of 20 high-AGP samples (AGP = 0.86 ± 0.11), of which all but one belonged to the Proneural group defined previously (5). Conversely, only 19 out of 38 previously defined Proneural samples (among the 128 analyzed) were Proneural here. Thus, our first revision of GBM classification is that the previously recognized Proneural group splits into two, about half becoming the newly recognized Proneural GBM, another half joining the Non-Proneural class. The Non-Proneural GBMs fell on a continuous distribution that parallels a gradient of AGP (range: 0.23–0.99), and span from the former Mesenchymal samples toward the Classical, Neural, and the rest of the former Proneural samples (Figure 1D).

We sought to validate these findings in the second batch of GBM (GBM2) (SI-3.1), using 154 samples having both DNA and mRNA data. AGP estimates were generated as above, showing a similar distribution of AGP in PCA plots of CNA and gene expression data (Figure S5A–B). Just as in GBM1, combined analysis revealed two well separated classes (Figure S5C), with 15 Proneural samples.

Molecular and clinical features of Proneural GBMs (Proneural/G-CIMP+)

To provide biological annotation of Non-Proneural and Proneural samples, we first note that they carried distinct CNA patterns. Non-Proneural GBMs carried recurrent gains in chromosomes 7, 19, 20, recurrent losses in chromosomes 9p and 10, and a gradient of CNA intensities due to varying AGP (Figure 2A). Proneural tumors, in contrast, lacked most of the Non-Proneural features described above and had high AGP values. They carried a more diverse set of CNAs, including 11p15.2 deletions (n=12 out of 20), 8q24.21 amplification (n=7), and 10p11.23 amplifications (n=14). Two of the Proneural samples showed co-occurrence of chr1p loss and chr19q loss (bottom of Figure 2A), each of which was rarely seen in other samples, yet this co-deletion has been reported as a key feature in anaplastic oligodendrogliomas (14, 15). Proneural GBMs had more *IDH1* mutation, a hallmark of secondary GBM (16–18). They showed higher frequencies of mutations in *TP53*, lower frequencies of mutations in *PTEN*, fewer deletions of *CDKN2A* - these are also signatures of secondary GBM reported previously (17, 19). They also showed fewer amplifications and over-expression of *EGFR*, high expression of *PDGFRA*, and lower expression of *FAS* and *MDM2* (Figure 2B and Table S2).

We also compared clinical outcome between the two groups. Compared to Non-Proneural GBM, patients with Proneural GBM were younger at diagnosis (Figure 2C) and had longer survival time (Figure 2D). Notably, while the Proneural group defined here has a better outcome, the other half of the former Proneural group (which we assigned to non-Proneural), is significantly worse than the rest of the Non-Proneural group (P=0.0059). Thus, lumping the two dissimilar types of GBM in the previously defined Proneural class would have missed a clinically relevant distinction.

A recent study of methylation patterns in TCGA samples revealed a subclass of GBM with glioma-CpG island methylator phenotype (G-CIMP+), an epigenetic signature associated with secondary or recurrent GBM and with IDH1 mutations (6). Of the 20 Proneural samples we identified, 15 were G-CIMP+ (Figure 2B); whereas of the 108 Non-Proneural samples none was G-CIMP+, strongly supporting Proneural GBM as a biologically distinct subtype. Indeed, 3-way analysis of CNA, gene expression, and DNA methylation data revealed consistent separation between Proneural and Non-Proneural GBMs (Figure S6). Proneural samples also match the Proneural GBMs defined in Phillips et al. (8) (SI-3.2, Table S3). As the term "Proneural" was applied differently in Verhaak et al. and Phillips et al. we renamed the Proneural group as Proneural/G-CIMP+ (or PN/G-CIMP+). PN/G-CIMP+ samples carry signatures resembling those of secondary GBM or low-grade gliomas (16), despite the fact that all but four samples in TCGA have been designated as primary (three of these were PN/G-CIMP+). These results suggest that a fraction (20 of 128 analyzed, ~16%) of the apparently primary GBM cases recruited in TCGA may in fact be latent secondary cases.

Three subclasses within Non-Proneural GBMs: Molecular and clinical signatures

After Proneural/G-CIMP+ GBMs were recognized, we sought to identify subclasses within the remaining, Non-Proneural GBMs. The reason for removing an already recognized group (i.e., Proneural/G-CIMP+) when studying the fine structure inside another (Non-Proneural) is that the markers distinguishing the two main groups may not be most informative for the within-group analyses, and could confound the latter. We applied a two-step method that emphasizes GBM1-GBM2 mutual validation (SI-3.1 and Figure S7, S8) and discovered three Non-Proneural classes. By comparing with the class assignments reported in Verhaak et al. and the annotated features in Phillips et al. we named the three classes as Classical, Proliferative, and Mesenchymal (we chose to apply similar terminology as in previous work even though many samples were reassigned, and the revised subtypes took on new characteristics). Of the 108 samples, 70 (65%) had one-to-one mapping to the previous NL, CL, and MES classes (SI-3.2 and Figure S9); thus 35% of GBM1 samples received revised assignments. We similarly analyzed the 46 Non-Proneural samples in Phillips et al. (Figure S10–11), and found that the former Proliferative group was split into the new Proliferative and Classical groups, and 11 (24%) were reassigned into or out of the MES group.

Since any new method could lead to a different classification, we pursued an important question: are the biological features of the new classes more robust than in the old system? Many marker genes highlighted in previous studies were consistently observed (SI-3.3 and Table S4). In CNA patterns (Figure 3A), while Non-Proneural samples shared the chr7 gains and chr10 losses, Proliferative samples carried additional deletions in chr14 and chr15 rarely seen in Classical samples (Student t test for chromosome-wide averaged copy number: $P=2.6\times10^{-3}$ and 3.1×10^{-4}, for chr14 and chr15, respectively), whereas Classical samples carried more amplifications in chr19 ($P=1.1\times10^{-6}$) and chr20 ($P=3.6\times10^{-6}$) than in Proliferative samples. Interestingly, many MES samples carried both the chr14–15 deletions and the chr19–20 gains, although with varying intensities due to lower aneuploid content, and with significantly more chr13 deletions compared with non-MES tumors ($P=9.6\times10^{-3}$). For Proliferative and MES samples, chr14q and 15 deletions tended to be mutually exclusive (mean Pearson’s $r = -0.23$ for Prolif and -0.26 for MES); whereas for Classical samples, chr19 gains tended to co-occur with chr20 gains (mean Pearson’s $r = 0.46$). These results showed that in addition to the CNA differences between Non-Proneural and PN/G-CIMP+ (Figure 2A), the three Non-Proneural classes carried different patterns of genomic aberration, possibly reflecting their differences in cell lineage, transcriptome patterns, and patient outcome.

The three Non-Proneural classes also showed significant differences in survival time in a three-way comparison in GBM1 (Figure 3B, log-rank test P=0.011). This is in contrast to the previous class assignments (5), for which the three-way comparison was not significant (Figure 3C). For individual pairs of classes, five out of six pairwise comparisons were significant in the revised system, while only one of six was significant in the previous system (Figure S12). The revised classes for Phillips' dataset also had significant survival differences in the three-way comparison (P = 0.033, log-rank test) and in the four-way comparison that included the PN/G-CIMP+ group (P = 0.014).

To directly compare the relative hazard across the four GBM subtypes and incorporate relevant patient characteristics, we performed a Cox proportional hazard regression analysis using our four-class assignments as explanatory covariates, and including patient age and the Karnofsky Performance Status (KPS) scores. First, for the entire set of 128 GBM1 samples, with the PN/G-CIMP+ subtype used as the reference category, the three non-Proneural subtypes had higher hazard ratios in the revised system (Figure S13A) than in the previous system (Figure S13B). Second, when we focused only on the three non-Proneural subtypes, using Classical as the reference, the 108 samples in the revised system (Figure S13C) showed higher hazard ratios than the 98 samples in the previous system (Figure S13D). To compare concordance between tumor classification and patient outcomes, we computed the C-statistics (20) for the 128 GBM1 dataset using the Cox regression model with age, KPS and subtypes as covariates. Revised classification had a concordance score of 0.668, higher than using age and KPS alone (0.643) by 2.5%, whereas the previous system had a concordance of 0.651, higher than using age and KPS alone (0.643) by only 0.8%, indicating that the revised system had improved predictive power for patient outcome.

Validation of survival time differences in an independent cohort

The Non-Proneural classes described above were defined by mutual validation of GBM1 and GBM2, thus having used information from both cohorts. To validate the survival time differences in a new, independent dataset, we analyzed a third batch of 144 TCGA samples (GBM3). As before, we identified 26 PN/G-CIMP+ samples using expression data and CNA data. After "locking down" the class assignment in GBM1 and GBM2 we selected 651 genes as the most informative predictors of the three Non-Proneural GBM classes (Table S5), and applied them in *supervised* class assignment of Non-Proneural samples in GBM3 (SI-4). Survival time differences were indeed validated in GBM3, with five out of six pairwise comparisons showing significant differences (Table 1A). To compare with the previous system, we used the 840 markers suggested by Verhaak et al. to classify the GBM3 samples and found that only one of six pairwise comparisons was significant (Table 1B).

Inference of cell type composition of GBM classes

We attempted to deduce the possible cell type composition of the four GBM classes to shed light on the cellular origins of this heterogeneous cancer. To do so, we compared GBM expression data with a reference dataset, GEO accession GSE9566 (11), for 38 samples that represent four main cell types in the central nervous system: acutely isolated astrocytes, neurons, oligodendrocytes, and cultured astroglia. The 38 samples formed four well-separated clusters, in agreement with their known identity (Figure S14). Cross-correlations of Non-Proneural GBM samples with the 38 reference samples, when grouped by class (for GBM) and cell type (for reference samples), showed recognizable mapping of GBM classes to known neural cell types, for GBM1 (N=128), GBM2 (N=154), and Phillips' dataset (N=56) (Figure 4A–C). Both PN/G-CIMP+ and Proliferative samples showed high correlations with neurons and oligodendrocytes, suggesting that they both resemble oligodendrogliomas. The Classical samples were similar to the astrocytes, suggesting that they may be related to astrocytomas. Lastly, the Mesenchymal samples showed high similarities with the cultured astroglia samples, which had an "immature or reactive phenotype" (11), consistent with the MES signatures of angiogenesis and inflammatory infiltration (5, 8, 21). The observed resemblance to known cell types were generally consistent with what was reported previously (5), but with important differences (SI-3.4).

As most of the low-AGP samples fell in the Mesenchymal group, we attempted to clarify the cell lineage of the aneuploid and euploid populations. If the aneuploid cells were derived from one of the reference cell types, there should be a positive correlation between (1) the correlation between samples of that particular cell type and individual MES tumors and (2) the MES tumors' AGP values, which measure how much aneuploid cells they contain. We calculated the correlation coefficients r, for each of the 38 reference samples, between its correlation coefficients with the MES samples and the AGP values of the MES samples, and found consistent and positive r values for Cultured Astrocytes (Figure 4D), suggesting that the aneuploid cells in MES share gene expression features, and possible common lineage, with reactive astrocytes (11).

As no other cell type in the reference set showed negative correlations, the identity of the euploid cells in MES remained unexplained. MES tumors carry angiogenic and inflammatory signatures, and some microglia markers are highly expressed in MES samples (5). We therefore hypothesize that the euploid fraction may be related to microglia/macrophage infiltration. To test this hypothesis, we searched public databases for gene expression data for microglia samples (SI-1.1.2), and found data for tumor-infiltrating microglia/macrophage isolated from freshly excised brain tumors ("TI. microglia", in GEO accession GSE25289) (22) and for microglia fraction from postoperative GBM tissue ("G. microglia", in GSE16119) (21). The correlation of these cells with MES tumors showed negative correlations with AGP (Figure 4D), suggesting that expression signatures of MES euploid cells are similar to microglia/macrophage. Moreover, two microglia/macrophage-specific transcripts, integrin alpha M (*ITGAM*) (23) and allograft inflammatory factor-1 (*AIF1*) (24), were negatively correlated with AGP ($r = -0.58$, $P = 6.3\times10^{-6}$ for *ITGAM*; $r = -0.53$, $P = 5.3\times10^{-5}$ for *AIF1*), further supporting microglia/macrophage as the probable source of euploid population in MES.

Hierarchical classification of GBM

The new understanding of GBM genomic landscape led to our proposal of a cohesive stepwise classification procedure (Figure 5). First, Proneural/G-CIMP+ GBMs can be identified with joint analyses of copy number and mRNA profiles, along with clinical data such as patient age. Even if a case was recorded as primary GBM due to the apparent lack of antecedent tumors, it could be recognized as Proneural/G-CIMP+ by features such as younger age, *IDH1* mutations, lack of *PTEN* mutations, hyper-methylation patterns, and lack of chr7 gains and chr10 losses. Among the remaining, Non-Proneural samples, MES samples can be separated from the Classical and Proliferative samples by lower AGP values, necrosis signatures, higher expression of *FAS* and *CHI3L1*, etc. These tumors experienced more infiltration of non-cancerous cells, containing aneuploid reactive astrocyte-like cells intermingled with cells such as microglia/macrophage that lack CNAs. Lastly, Classical and Proliferative samples can be distinguished by gene expression patterns that resemble different neural cell types. Known markers highlighted by previous studies (Table S4), such as *PCNA* and *TOPA2A* overexpression in Proliferative samples, can also be incorporated in this step.

DISCUSSION

Discoveries of GBM subtypes have so far relied on single data types. The work reported here combined DNA genotyping data and gene expression data, and revealed a novel GBM subtype (Proneural/G-CIMP+) that carried distinct molecular, clinical, and demographic features. While this subtype was described separately in a study of methylation data (6), our approach reached the conclusions from two other, independent data types, and suggests that such a combined approach will be useful in genomic analysis of other cancers.

We refrained from equating AGP with "tumor purity", because some bona fide tumor cells may be euploid (but may carry point mutations in key "driver" genes), and non-malignant cells may carry high levels of genomic aberration. Importantly, AGP values correlated poorly with histopathologic report of tumor content, yet unlike the latter, were strongly correlated with CNA and gene expression data. This underscores the limitation of traditional concepts such as up- and down-regulation of transcripts in samples of homogeneous cellular composition or uniform character. Similar results were seen in the second GBM cohort and the ovarian cancer data, suggesting that heterogeneity estimates from molecular data should be considered as a covariate in tumor classification and fundamental parameter in clinical diagnosis. More generally, the ability to infer AGP informs sample selection for further characterization such as sequencing, and provides a key variable for understanding tumor progression.

Many previous studies perform class discovery in one cohort and validate only the key results (such as the most discriminant genes) in a new cohort. One of the strengths of our approach is in relying on mutual validation between two cohorts (GBM1 and GBM2 in our case) in initial class discovery. Although the classification was conducted without using outcome data, the groups defined here showed stronger differences in survival time than those reported previously. And we were able to validate the differences in a third, independent cohort. GBM is exceedingly aggressive and has a short median survival time (18 months); the ability to delineate subclasses with differing survival time, even by 2–3 months, is clinically relevant, especially when the molecular markers can be used to prospectively identify patients of different prognosis. Another advance in this work is the apparent mapping of individual GBM classes to known cell types: Cultured Astroglia (for MES), Astrocyte (Classical) and Oligodendrocyte-Neuron (Proliferative), as well as the euploid (microglia/macrophage) and aneuploid (astrocytes) components of the MES group. It is important to emphasize, however, that cell lineage and cell differentiation are extremely complex in brain tumors, and the results presented here are based on indirect, correlative analyses of reference datasets. The new hypotheses thus generated remain to be tested in future experimental work.

Finally, we proposed a hierarchical classification scheme for GBM that integrates diverse molecular and clinical observations. While previous studies aimed to discover mutually exclusive classes that *divide* the data, our hierarchical system raised the question of whether the four classes could be sequentially related. Philipps et al. studied 26 pairs of matched primary and recurrent astrocytomas from the same patients, and found that upon recurrence some Proneural or Proliferative tumors shifted toward the MES phenotype. This suggests that MES might be a late-stage class that could be reached by one (or any) of the other three classes upon acquiring further genetic abnormalities. In contrast, Noushmehr et al. (6) studied 15 pairs of primary and recurrent GBM, and did not observe any class switching between G-CIMP+ and G-CIMP- types. Future studies will be needed to elucidate the molecular mechanisms that underlie the initial neoplastic transformation leading to different classes of GBMs. Still, the CNA data revealed an interesting pattern of incremental change across the four classes: Proneural/G-CIMP+ samples lacked chr7 gains and chr10 losses that were present in the three Non-Proneural classes. Many Proliferative and Classical samples acquired chr13/14/15 deletions and chr19/20 amplifications, respectively, but not both. And finally, MES samples carried both chr13/14/15 deletions and chr19/20 amplifications, but with varying levels of mixing with euploid cells. It is tempting to speculate that this pattern of progressive acquisition of genomic aberration is consistent with a model in which Proneural/G-CIMP+ GBM may develop in younger patients without chr7 and chr10 CNAs, as *IDH1* mutations and/or *P53* mutations might be sufficient primary drivers of GBM in these individuals. Patients not carrying *IDH1* or *P53* mutations may acquire chr7 gains and chr10 losses before developing GBM, an event that is accompanied by additional aberrations in either chr13/14/15 in neurons or oligodendrocytes, or chr19/20 in astrocytes. Depending on the cell lineage the tumor may arise as either Proliferative or Classical. Finally, upon hypoxia, necrosis, and angiogenesis, as well as possible further differentiation, Mesenchymal samples emerge from these "earlier" classes and carry both chr13/14/15 and chr19/20 abnormalities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of TCGA Research Network for sharing data and for stimulation discussions, and Drs. Yuan Zhu, Terry Speed, and Kerby Shedden for helpful comments on our approach. We also thank Drs. Richard Myers and Devin Absher for their support and advice.

References

1. Adamson C, Kanu OO, Mehta AI, Di C, Lin N, Mattox AK, et al. Glioblastoma multiforme: a review of where we have been and where we are going. Expert Opin Investig Drugs. 2009; 18:1061–1083.
2. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403:503–511. [PubMed: 10676951]
4. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406:747–752. [PubMed: 10963602]
5. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]
6. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. Cancer Cell. 2010; 17:510–522. [PubMed: 20399149]
7. Murat A, Migliavacca E, Gorlia T, Lambiv WL, Shay T, Hamou MF, et al. Stem cell-related "Self-Renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. J Clin Oncol. 2008; 26:3015–3024. [PubMed: 18565887]
8. Phillips HS, Kharbanda S, Chen RH, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006; 9:157–173. [PubMed: 16530701]
9. Sun LX, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, Pastorino S, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell. 2006; 9:287–300. [PubMed: 16616334]
10. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]
11. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci. 2008; 28:264–278. [PubMed: 18171944]
12. R Development Core Team R. A language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria. 2010
13. Shirahata M, Iwao-Koizumi K, Saito S, Ueno N, Oda M, Hashimoto N, et al. Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. Clin Cancer Res. 2007; 13:7341–7356. [PubMed: 18094416]
14. Cairncross JG, Ueki K, Zlatescu MC, Lisle DK, Finkelstein DM, Hammond RR, et al. Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. J Natl Cancer Inst. 1998; 90:1473–1479. [PubMed: 9776413]
15. Ducray F, Idbaih A, de Reynies A, Bieche I, Thillet J, Mokhtari K, et al. Anaplastic oligodendrogliomas with 1p19q codeletion have a proneural gene expression profile. Mol Cancer. 2008; 7:41. [PubMed: 18492260]
16. Cooper LA, Gutman DA, Long Q, Johnson BA, Cholleti SR, Kurc T, et al. The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas. PLoS One. 2010; 5:e12548. [PubMed: 20838435]
17. Kleihues P, Ohgaki H. Primary and secondary glioblastomas: from concept to clinical diagnosis. Neuro Oncol. 1999; 1:44–51. [PubMed: 11550301]

18. Nobusawa S, Watanabe T, Kleihues P, Ohgaki H. IDH1 Mutations as Molecular Signature and Predictive Factor of Secondary Glioblastomas. Clin Cancer Res. 2009; 15:6002–6007. [PubMed: 19755387]
19. Ohgaki H, Kleihues P. Genetic pathways to primary and secondary glioblastoma. Am J Pathol. 2007; 170:1445–1453. [PubMed: 17456751]
20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15:361–387. [PubMed: 8668867]
21. Murat A, Migliavacca E, Hussain SF, Heimberger AB, Desbaillets I, Hamou MF, et al. Modulation of Angiogenic and Inflammatory Response in Glioblastoma by Hypoxia. PLoS One. 2009; 4
22. Mora R, Dokic I, Kees T, Huber CM, Keitel D, Geibig R, et al. Sphingolipid Rheostat Alterations Related to Transformation Can Be Exploited for Specific Induction of Lysosomal Cell Death in Murine and Human Glioma. Glia. 2010; 58:1364–1383. [PubMed: 20607862]
23. Guillemin GJ, Brew BJ. Microglia, macrophages, perivascular macrophages, and pericytes: a review of function and identification. J Leukocyte Biol. 2004; 75:388–397. [PubMed: 14612429]
24. Schwab JM, Frei E, Klusman I, Schnell L, Schwab ME, Schluesener HJ. AIF-1 expression defines a proliferating and alert microglial/macrophage phenotype following spinal cord injury in rats. J Neuroimmunol. 2001; 119:214–222. [PubMed: 11585624]

STATEMENT OF TRANSLATIONAL RELEVANCE

Accurate definition of GBM subtypes can highlight relevant biological pathways to inform diagnosis and treatment. However, current classification has relied mainly on gene expression data. We used allelic DNA copy number data to estimate the aneuploid content in each GBM as an intrinsic measure of its heterogeneity. Joint use of DNA and mRNA data in *ab initio* class discovery led to a revised classification scheme, with improved between-class differences in survival time and clearer relationships to known cell types. Our stepwise framework also clarified a long-standing confusion regarding the Proneural group, and identified the microglia/macrophage as the likely euploid source for the Mesenchymal subtype. The Proneural/G-CIMP+ group carries signatures of secondary GBMs and is resistant to chemo-/radio-therapies; thus its accurate diagnosis is clinically relevant. The ability to infer within-tumor heterogeneity opens new ground for studying clonal evolution, the role of stromal cells in tumor growth and metastasis, and subtype-specific treatment.

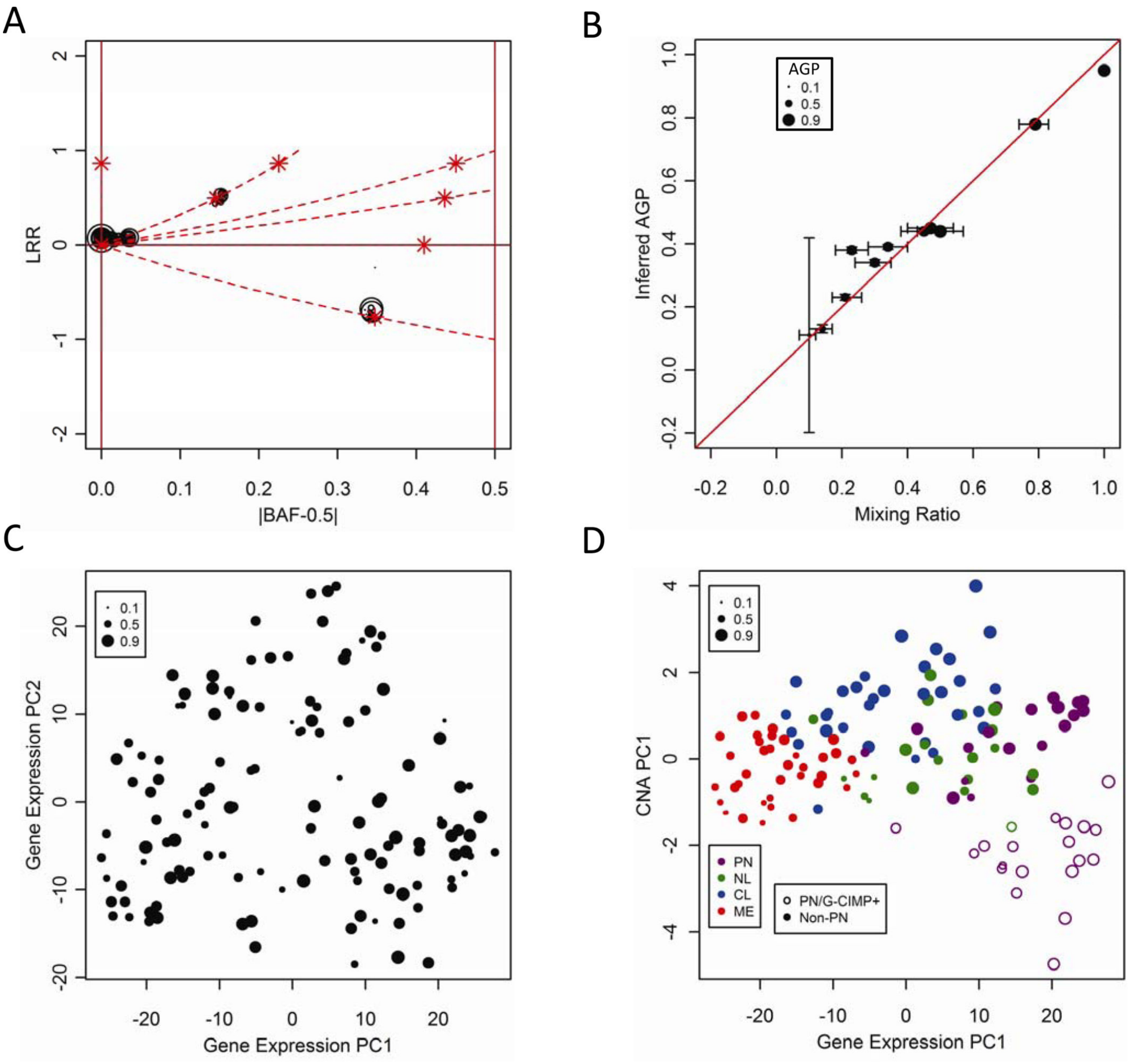

Figure 1. AGP and relationship to gene expression patterns

A. BAF-LRR plot of sample TCGA-02-0038 as an example of using allelic intensity data to estimate AGP. The x axis shows |BAF-0.5|, the absolute deviation of B allele frequency (BAF) between tumor and matched normal samples, at heterozygous SNP loci in the normal sample; y axis is the LRR, logR Ratio between tumor and normal samples. Canonical positions representing integer combinations of (N_B, N_A+N_B) are marked with red stars, with red dashed lines indicating the contraction paths when AGP < 1 (see also Figure S1). Most CNAs, shown as "bubbles", fell on canonical positions. The size of the bubble shows CNA length. PC (percent of genome changed) = 0.20 for this sample. Inferred AGP is 0.82. PoP (percent of changed genome on canonical points) = 0.99. **B.** Validation of AGP inference algorithm, using reference dataset GSE11976, for DNA pools of a breast cancer cell line mixed with a lymphoblastoid cell line at known ratios. Error bars show the 95% confidence intervals from the experimental procedures (horizontal) and from our bootstrap method (vertical). The red line has a slope of 1 and intercept of 0. **C.** Scatter plot of PC1-PC2 (the first two principal component scores) of GBM1 gene expression data. Symbol size is proportional to AGP as indicated in the legend. **D.** Scatter plot of PC1 of CNA (also shown on the x-axis in S4A) versus PC1 of gene expression data (shown on the x-axis in **1C**); Non-Proneural and Proneural/G-CIMP+ GBM samples were indicated by filled and open symbols, respectively.

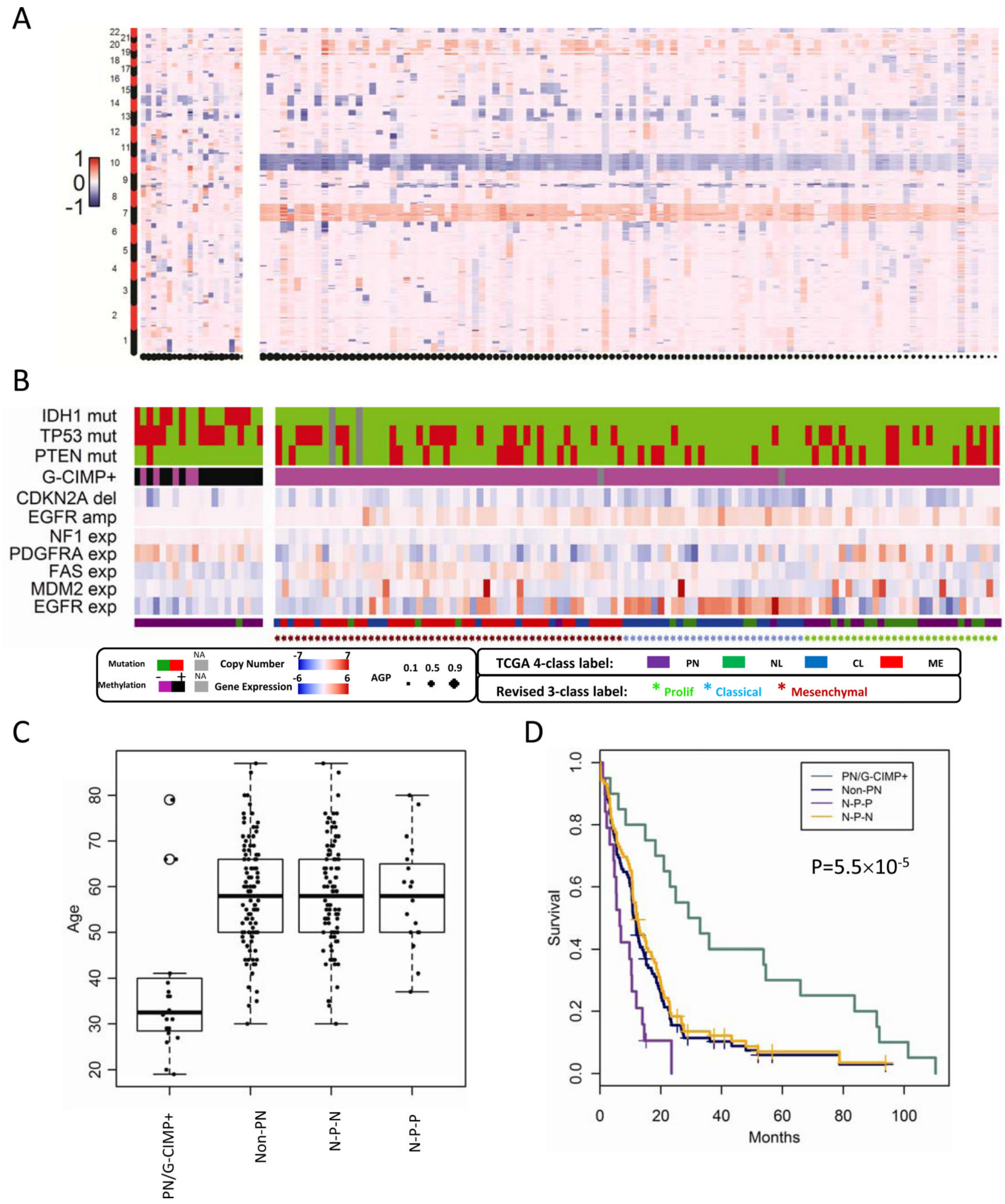

Figure 2. Molecular and clinical features of Proneural/G-CIMP+ GBM

A. Heatmap of per-cytoband total copy number in Non-Proneural and Proneural/G-CIMP+ samples, with Chr1–22 arranged from bottom to top. Non-Proneural samples were ordered from left to right by decreasing AGP, and showed characteristic features, such as chr7 amplifications (shown in red) and chr10 deletions (in blue), across most samples, albeit with a gradient of magnitude. **B.** Selected molecular features, including, from top to bottom, presence or absence of non-silent mutations in *IDH1*, *TP53* and *PTEN* as reported by (2); G-CIMP+, a methylation signature described in (6); total copy number in *CDKN2A* and *EGFR*; expression levels of *NF1*, *PDGRFA*, *FAS*, *MDM2* and *EGFR*, as described in (2).

The four classes defined in Verhaak et al., and the three classes defined in this work, are indicated as colored symbols in the bottom row. **C**. Distribution of age-of-diagnosis in Non-Proneural (n=110) and Proneural/G-CIMP+ (n=20) samples. Also shown are two subgroups of Non-Proneural GBM: Proneural (N-P-P) and non-Proneural (N-P-N). **D.** Kaplan-Meier survival curves for Non-Proneural and Proneural/G-CIMP+ groups, with the latter showing better outcome (log rank test p-value=7.5E-7). The Non-Proneural group was further split into the former Proneural (N-P-P) and non-Proneural (N-P-N) samples.

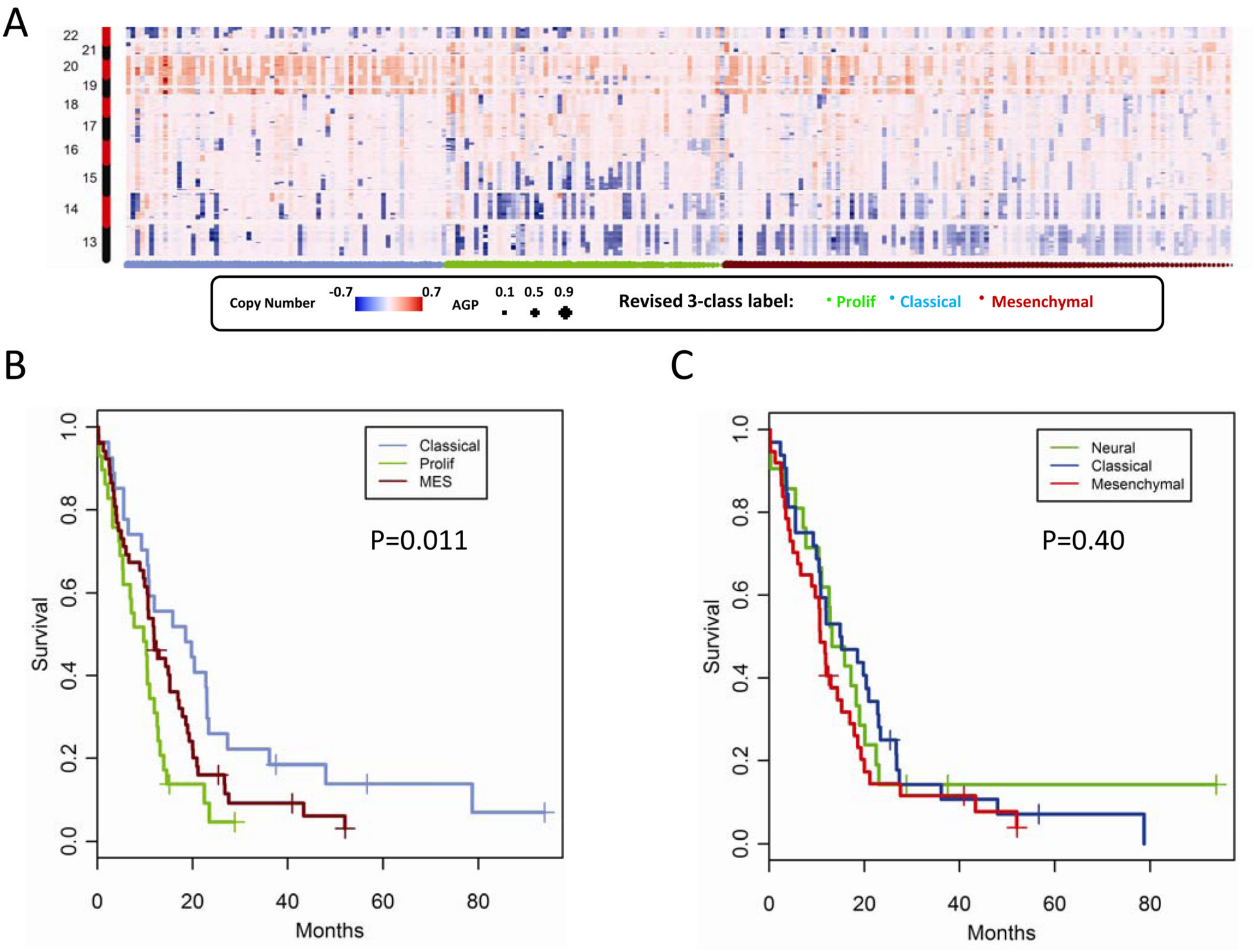

Figure 3. Molecular and clinical signatures for Non-Proneural GBM classes

A. Chr13–22 CNA patterns in the revised Non-Proneural GBM classes. Class assignments were indicated by the stars at the bottom, with size proportional to AGP. **B–C**. Kaplan-Meier curves for GBM1 according to the revised classes (**B**) and the previous classes (**C**). The overall log-rank test for the three classes was significant in **B** (P=0.011), but not in C.

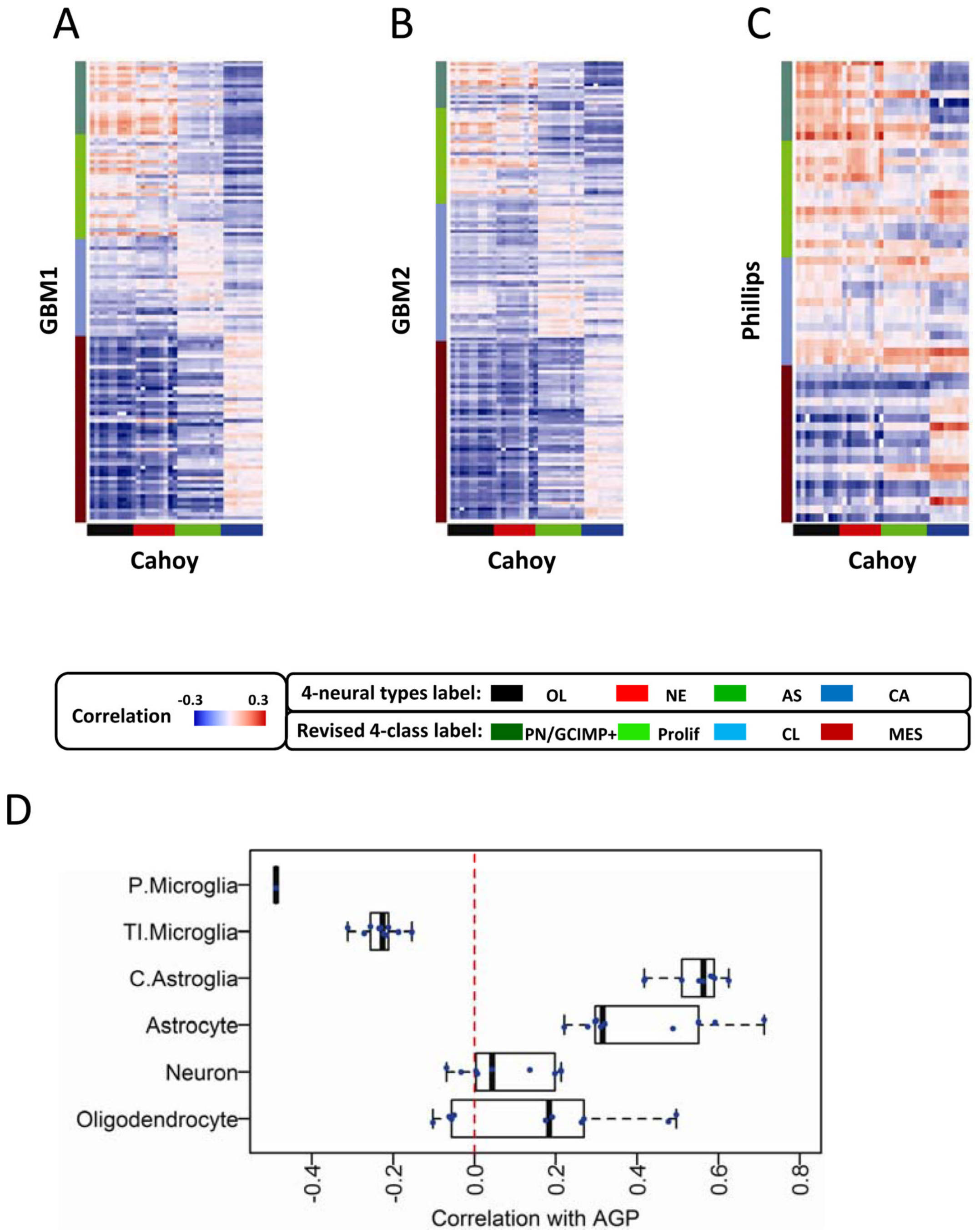

Figure 4. Inference of cell type composition of revised Non-Proneural classes

A–C. Heatmaps of the cross-correlation coefficients between the reference dataset of 38 samples of known neural cell types and samples in GBM1 (**A**), GBM2 (**B**) and Phillips' study (**C**). Colored segments in sidebars indicate sample assignments for four GBM classes or for the four neural cell types. **D.** Distribution of the correlation coefficients between (1) AGP values of MES samples and (2) correlations with individual reference samples, for the four neuronal cell types in Cahoy, et al. (GEO accession GSE9956) and two datasets for microglia (GSE25289, GSE16119).

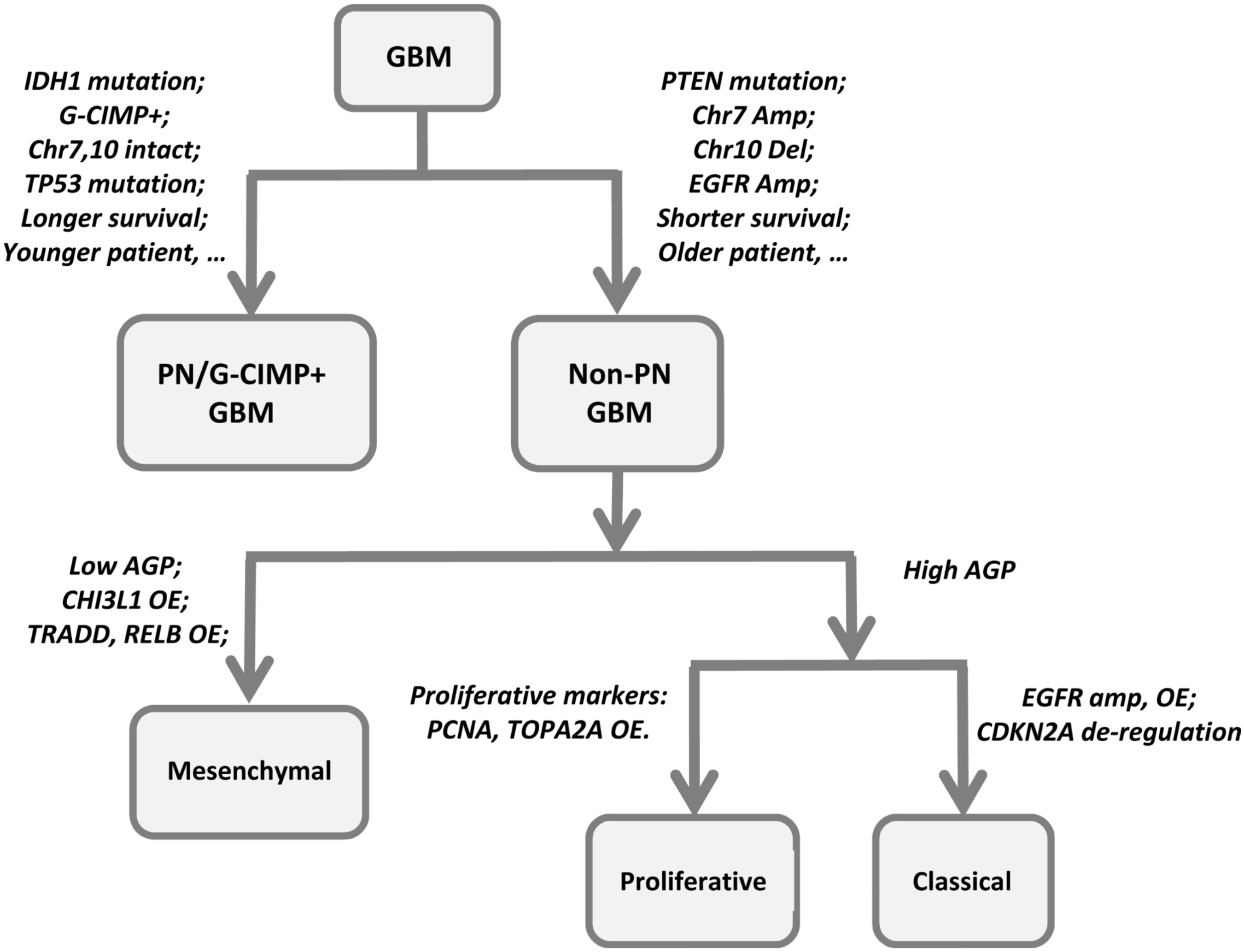

Figure 5. A proposed hierarchical classification scheme for GBM

Joint use of DNA and mRNA data, along with patient age and outcome data, separates Proneural/G-CIMP+ GBMs from Non-Proneural GBMs in the first step of the decision tree. The two subsequent two-way decisions define the three Non-Proneural classes, using features indicated in the diagram and the most informative transcripts in Table S4.

Table 1

Pairwise survival time comparisons in GBM3 for classes assigned by using the most informative markers from our study (**A**) or those from Verhaak et al. (**B**).

A

	PN/G-CIMP+	**Prolif**	**Classical**	**MES**
PN/G-CIMP+	-	**1.5e-4**	**0.018**	**0.0015**
Prolif		-	**0.010**	**0.040**
Classical			-	0.33
MES				-

B

	PN	**NL**	**CL**	**MES**
PN	-	0.059	**0.046**	0.091
NL		-	0.60	0.85
CL			-	0.92
MES				-