Genomic Estimates of Aneuploid Content in Glioblastoma Multiforme and Improved Classification

Bo Li¹, Yasin Senbabaoglu¹, Weiping Peng², Min-Lee Yang², Jishu Xu², and Jun Z. Li¹,²

Abstract

**Purpose:** Accurate classification of glioblastoma multiforme (GBM) is crucial for understanding its biologic 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 biologic features of the newly defined subtypes, and reconcile with prior studies.

**Experimental Design:** 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 glioma-CpG island methylator phenotype (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 about 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. *Clin Cancer Res; 1–11.* ©2012 AACR.

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 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*/PI3K*, *p53*, and *RB* signaling pathways. Meanwhile, genome-wide datasets are also useful for characterizing biologic 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 4 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 4-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 3 clusters, which lacked a clear correspondence with the 4 transcriptome-based classes. Furthermore, the relationship of the 4-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 unreconciled. 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...
Translational Relevance
Accurate definition of glioblastoma multiforme (GBM) subtypes can highlight relevant biologic 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 \textit{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 about 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-/radiotherapies; 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.

Materials and Methods
Datasets
The sources of the datasets, including (1) gene expression data for TCGA GBM (separately analyzed as 3 batches: GBM1–3), ovarian (10), Phillips and colleagues dataset for GBM [Gene Expression Omnibus (GEO) accession GSE4271; ref. 8], and the reference dataset by Cahoy and colleagues for neuronal cells (GEO accession GSE9956; ref. 11), (2) DNA data for GBM1–3, ovarian, (3) methylation data for GBM1, and (4) clinical data for these studies, are described in Supplementary Information (SI, 2.1–2.5). Inferring aneuploid genome proportion (AGP). We conducted logR and B allele frequency (BAF)–based segmentation and merged the 2 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 Supplementary Table S1.

Statistical analysis
K-means clustering, quantile normalization, \(t\) tests, principal component analysis (PCA), and heatmap generation were conducted 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 conducted using the R package \textit{survival}. 3D scatter plots were generated using R package \textit{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 2 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, \(\log R = \log(n_A + n_B)\), and the observed 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 (Supplementary Fig. 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” toward those of the latter; and different mixing ratios result in different degrees of contraction (Supplementary Fig. S1B). An example of such a mixed GBM sample is shown in Fig. 1A. On the basis of this feature, we developed an algorithm to quantitatively estimate genome-wide 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 (Fig. 1B; SI, 2.7).

In this study, we define aneuploid content, or synonymously, AGP, as the parameter \(p\) in a mixture model consisting of 2 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, for example, \(n_A = 2\) and \(n_B = 1\) in a region of amplification. For many tumors, the 2-way mixing model considered here is likely an oversimplification, as multiple subpopulations of tumor cells may exist, each carrying a different integer combination of parental segments. However, a mixture model with 3 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 2-way mixing model is the simplest scenario that could have generated the observed data about 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), 7 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 percentage changed (PC) of 37.3%, that is, greater than one third 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 2-way mixing model (SI, 2.6 and

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 BAF between tumor and matched normal samples, at heterozygous SNP loci in the normal sample; y-axis is the logR ratio (LRR), between tumor and normal samples. Canonical positions representing integer combinations of (NB, NA + NB) are marked with red stars, with red dashed lines indicating the contraction paths when AGP < 1 (see also Supplementary Fig. S1). Most CNAs, shown as bubbles, fell on canonical positions. The size of the bubble shows CNA length. PC (percentage of genome changed) = 0.20 for this sample. Inferred AGP is 0.82. PoP (percentage 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% CI 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 2 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 (C)); Non-Proneural and Proneural/G-CIMP+ GBM samples were indicated by filled and open symbols, respectively.
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 more than 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 \( \rho = 0.14, P = 0.02, n = 275 \)), 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 (Supplementary Fig. 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 \( r = 0.62, P < 10^{-15}, n = 128 \), Fig. 1C). PC2 was also correlated with AGP (\( r = 0.48, P = 1.1 \times 10^{-8} \)). 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 tumors from TCGA (10), and observed a similar pattern (Supplementary Fig. S5A). 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 earlier, showing a similar distribution of AGP in PCA plots of CNA and gene expression data (Supplementary Fig. S5A and S5B). Just as in GBM1, combined analysis revealed 2 well-separated classes (Supplementary Fig. SSC) with 15 Proneural samples.

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

To provide biologic 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, and 20, recurrent losses in chromosomes 9p and 10, and a gradient of CNA intensities due to varying AGP (Fig. 2A). Proneural tumors, in contrast, lacked most of the Non-Proneural features described earlier 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 Fig. 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 overexpression of EGFR, high expression of PDGFRα, and lower expression of FAS and MDM2 (Fig. 2B and Supplementary Table S2).
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 nonsilent mutations in IDH1, TP53, and PTEN as reported in ref. 2; G-CIMP+, a methylation signature described in ref. 6; total copy number in CDKN2A and EGFR; expression levels of NF1, PDGFRα, FAS, MDM2, and EGFR, as described in ref. 2. The 4 classes defined in Verhaak and colleagues, and the 3 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 2 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.
We also compared clinical outcome between the 2 groups. As compared with Non-Proneural GBM, patients with Proneural GBM were younger at diagnosis (Fig. 2C) and had longer survival time (Fig. 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 2 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$^+$ (Fig. 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 (Supplementary Fig. S6). Proneural samples also match the Proneural GBMs defined in Phillips and colleagues (8; Sl. 3.2 and Supplementary Table S3). As the term "Proneural" was applied differently in Verhaak and colleagues and Phillips and colleagues, 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 4 samples in TCGA have been designated as primary (3 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 2 main groups may not be most informative for the within-group analyses, and could confound the latter. We applied a 2-step method that emphasizes GBM1–GBM2 mutual validation (SI, 3.1 and Supplementary Figs. S7 and S8) and discovered 3 Non-Proneural classes. By comparing with the class assignments reported in Verhaak and colleagues and the annotated features in Phillips and colleagues we named the 3 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 Supplementary Fig. S9); thus, 35% of GBM1 samples received revised assignments. We similarly analyzed the 46 Non-Proneural samples in Phillips and colleagues (Supplementary Figs. S10 and S11) 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.

Because any new method could lead to a different classification, we pursued an important question: are the biologic 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 Supplementary Table S4). In CNA patterns (Fig. 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: $P = 2.6 \times 10^{-3}$ and $3.1 \times 10^{-3}$, for chr14 and chr15, respectively), whereas Classical samples carried more amplifications in chr19 ($P = 1.1 \times 10^{-6}$) and chr20 ($P = 3.6 \times 10^{-6}$) 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 ($P = 9.6 \times 10^{-11}$). For Proliferative and MES samples, chr14q and 15 deletions tended to be mutually exclusive (mean Pearson $r = -0.23$ for Prolif and $-0.26$ for MES); whereas for Classical samples, chr19 gains tended to cooccur with chr20 gains (mean Pearson $r = 0.46$). These results showed that in addition to the CNA differences between Non-Proneural and PN/G-CIMP$^+$ (Fig. 2A), the 3 Non-Proneural classes carried different patterns of genomic aberration, possibly reflecting their differences in cell lineage, transcriptional patterns, and patient outcome.

The 3 Non-Proneural classes also showed significant differences in survival time in a 3-way comparison in GBM1 (Fig. 3B, log-rank test $P = 0.011$). This is in contrast to the previous class assignments (5), for which the 3-way comparison was not significant (Fig. 3C). For individual pairs of classes, 5 out of 6 pairwise comparisons were significant in the revised system, whereas only 1 of 6 was significant in the previous system (Supplementary Fig. S12). The revised classes for Phillips’ dataset also had significant survival differences in the 3-way comparison ($P = 0.033$, log-rank test) and in the 4-way comparison that included the PN/G-CIMP$^+$ group ($P = 0.014$).

To directly compare the relative hazard across the 4 GBM subtypes and incorporate relevant patient characteristics, we conducted a Cox proportional hazard regression analysis using our 4-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 3 Non-Proneural subtypes had higher hazard ratios (HR) in the revised system (Supplementary Fig. S13A) than in the previous system (Supplementary Fig. S13B). Second, when we focused only on the 3 Non-Proneural subtypes, using Classical as the reference, the 108 samples in the revised system (Supplementary Fig. S13C) showed higher HRs than the 98 samples...
in the previous system (Supplementary Fig. 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 earlier 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 3 Non-Proneural GBM classes (Supplementary 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 5 out of 6 pairwise comparisons showing significant differences (Table 1, upper). To compare with the previous system, we used the 840 markers suggested by Verhaak and colleagues to classify the GBM3 samples and found that only 1 of 6 pairwise comparisons was significant (Table 1, lower).

Inference of cell-type composition of GBM classes

We attempted to deduce the possible cell-type composition of the 4 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 4 main cell types in the central nervous system: acutely isolated astrocytes, neurons, oligodendrocytes, and cultured astroglia. The 38 samples formed 4 well-separated clusters, in agreement with their known identity (Supplementary Fig. 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; Fig. 4A–C). Both PN/G-CIMP+ and Proliferative samples showed high correlations with neurons and oligodendrocytes, suggesting that they both resemble oligodendroglomas. The Classical samples were similar to the astrocytes, suggesting that they may be related to astrocytomas. Finally, 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 was 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 (Fig. 4D), suggesting that the euploid 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/macrophage isolated from freshly excised brain tumors (“TI. Microglia," in GEO accession GSE25289; ref. 22) and for microglia fraction from postoperative GBM tissue (“G. microglia," in GSE16119; ref. 21). The correlation of these cells with MES tumors showed negative correlations with AGP (Fig. 4D), suggesting that expression signatures of MES euploid cells are similar to microglia/macrophage. Moreover, 2 microglia/macrophage-specific transcripts, integrin alpha M (ITGAM; ref. 23) and allograft inflammatory factor-1 (AIF1; ref. 24), were negatively correlated with AGP (r = −0.58, P = 6.3 × 10−4 for ITGAM; r = −0.53, P = 5.3 × 10−5 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 (Fig. 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, hypermethylation 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 noncancerous cells, containing aneuploid reactive astrocyte-like cells intermingled with cells such as microglia/macrophage that lack CNAs. Finally, Classical and Proliferative samples can be distinguished by gene expression patterns that resemble different neural cell types. Known markers highlighted by previous studies (Supplementary Table S4), such as PCNA and TOP2A overexpression in

### Table 1. Pairwise survival time comparisons in GBM3 for classes assigned by using the most informative markers from our study (upper) or those from Verhaak and colleagues (lower)

<table>
<thead>
<tr>
<th></th>
<th>PN/G-CIMP+</th>
<th>Prolif</th>
<th>Classical</th>
<th>MES</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN/G-CIMP+</td>
<td>—</td>
<td>1.5e−4</td>
<td>0.018</td>
<td>0.0015</td>
</tr>
<tr>
<td>Prolif</td>
<td>—</td>
<td>—</td>
<td>0.010</td>
<td>0.040</td>
</tr>
<tr>
<td>Classical</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.33</td>
</tr>
<tr>
<td>MES</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PN</td>
<td>—</td>
<td>0.059</td>
<td>0.046</td>
<td>0.091</td>
</tr>
<tr>
<td>NL</td>
<td>—</td>
<td>—</td>
<td>0.60</td>
<td>0.85</td>
</tr>
<tr>
<td>CL</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.92</td>
</tr>
<tr>
<td>MES</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
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+). This subtype was described separately in a study of methylation data (6), but our approach reached the conclusions from 2 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 nonmalignant 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 downregulation 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 conduct class discovery in 1 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 2 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.

![Figure 4. Inference of cell-type composition of revised Non-Proneural classes. A to 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 4 GBM classes or for the 4 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 4 neuronal cell types in Cahoy and colleagues (GEO accession GSE9956) and 2 datasets for microglia (GSE25289 and GSE16119).](Image)
and has a short median survival time (18 months); the ability to delineate subclasses with differing survival time, even by 2 to 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 4 classes could be sequentially related. Philipp and colleagues 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 3 classes upon acquiring further genetic abnormalities. In contrast, Nourshemehr and colleagues 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 4 classes; Proneural/G-CIMP+ samples lacked chr7 gains and chr10 losses that were present in the 3 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 TP53 mutations might be sufficient primary drivers of GBM in these individuals. Patients not carrying IDH1 or TP53 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.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors' Contributions
Conception and design: B. Li, J.Z. Li
Development of methodology: B. Li, J.Z. Li
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): B. Li, Y. Senbabaoglu, J.Z. Li
Acknowledgments
The authors thank members of TCGA Research Network for sharing data and for stimulating discussions, and Drs. Yuan Zhu, Terry Speed, and Kerby Shedden for helpful comments on our approach. The authors also thank Drs. Richard Myers and Devin Absher for their support and advice.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 9, 2012; revised July 15, 2012; accepted August 2, 2012; published OnlineFirst August 21, 2012.

References


Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): B. Li, Y. Senbabaoglu, M.-L. Yang, J. Xu, J.Z. Li
Writing, review, and/or revision of the manuscript: B. Li, W. Peng, J.Z. Li
Study supervision: J.Z. Li

www.aacrjournals.org Clin Cancer Res; 2012 OF11

Published OnlineFirst August 21, 2012; DOI: 10.1158/1078-0432.CCR-12-1427

Downloaded from clinicanereres.aacrjournals.org on July 16, 2017. © 2012 American Association for Cancer Research.
Clinical Cancer Research

Genomic Estimates of Aneuploid Content in Glioblastoma Multiforme and Improved Classification

Bo Li, Yasin Senbabaoglu, Weiping Peng, et al.

Clin Cancer Res  Published OnlineFirst August 21, 2012.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-12-1427

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2012/08/21/1078-0432.CCR-12-1427.DC1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.