Supplemental Information

Genomic estimates of aneuploid content in Glioblastoma Multiforme and proposal of a new integrated classification

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

¹Bioinformatics Program, ²Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA
Table of Contents

1. Data sources

1.1. Glioblastoma Multiforme (GBM) 3
1.2. Ovarian Cancer (OV) 4
1.3. Phillips et al. dataset 5
1.4. Cahoy et al. dataset 5
1.5. Data for microglia/macrophage 5

2. Inferring Aneuploid Genome Proportion

2.1. Data processing, DNA segmentation, and merging 5
2.2. Per-segment summary of LRR and BAF 6
2.3. LRR scale-normalization 6
2.4. BAF-LRR plot: canonical points and tracks 7
2.5. Inference of Aneuploid Genome Proportion 7
2.6. Genomic features and QC measures. 9
2.7. Validation of AGP algorithm 10
2.8. Impact of goodness-of-fit on comparisons with histopathologic reports and expression PCs 11

3. Classification of Non-Proneural GBM

3.1. A two-step procedure that relies on GBM1-GBM2 mutual validation 11
3.2. Comparison with previous studies 13
3.3. Marker genes for Non-Proneural subgroups 13
3.4. Comparison with previous assignment in terms of mapping to known cell types.

4. Supervised class assignment in GBM3 14

References 15
1. Data sources:

1.1. Glioblastoma Multiforme (GBM)

This study covered three cohorts of GBM samples. GBM1 is the cohort analyzed by the TCGA pilot study [1, 2]. A second cohort was subsequently available and was called GBM2 [2]. For validating the survival time differences we selected additional samples that became available by early 2012, and called it GBM3.

1.1.1 DNA copy number data

For GBM1-2, Allele-specific copy number data for Illumina HumanHap550K arrays were downloaded from the Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/) on 4/14/2010. We queried the Data Access Matrix by choosing

- Disease: GBM;
- Data Type: SNP;
- Data Level: 2 and 3;
- Platform: HAIB (HumanHap550).

This query yielded tumor-normal logR Ratio (LRR) data for 284 paired samples, and B allele frequency (BAF) data for 347 tumor samples, of which 284 had matched normal samples. The overlapping set of 284 paired samples, 130 in GBM1 and 154 in GBM2, was selected for further analysis. The dataset contains 561,468 autosomal SNPs.

Allele-specific copy number data for GBM3 came from TCGA batches 26, 38, 62, 79, 111, and 130. A total of 156 samples had both Affymetrix SNP 6.0 genotyping data and Affymetrix gene expression data available, and these were downloaded in bulk on 1/31/2012. The copy number data for Affymetrix SNP arrays were Birdsuite output files and were converted to logR and BAF. Ten samples were apparent outliers on the gene expression PCA plot (not shown), and were removed. Of the remaining 146 samples, two female patients (TCGA-12-3644 and TCGA-12-3646) had exceptionally longer survival time (62 and 44 months respectively), and were removed before classification analysis and survival time comparisons. The genome coordinates for the 811 autosomal cytoband were from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cytoBand.txt.gz.

1.1.2 Gene expression data

Gene expression data were downloaded from http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/. Most of our analyses were based on "unifiedScaledFiltered.txt", which contains processed data for 1,740 most variable genes for 202 GBM samples. The data processing procedure was reported previously (Verhaak et al, 2010). We also analyzed the full dataset in "unifiedScaled.txt", containing 11,861 genes before filtering. In
GBM1, a subset of 130 samples (out of 202) had both gene expression and DNA copy number data. In GBM2, all 154 samples had both gene expression and DNA copy number data.

Expression data for GBM3 were downloaded on 1/9/2012, from TCGA Data Portal (http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm), by choosing Platform= BI_HT_HG-U133A. GBM3 contains 170 samples from TCGA batches 26, 38, 62, 19 and 111. We quantile normalized these data before running downstream analysis. The 840 genes selected in Verhaak et al. for distinguishing the four former subtypes were provided at (http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/ClaNC840_centroids.xls) and accessed on 04/12/2011.

1.1.3 MicroRNA data

MicroRNA data were downloaded on 1/7/2012 from TCGA portal by choosing

Data Type: Expression miRNA;
Batch: All;
Level: 3;
Platform: UNC_miRNA_8×15K,

This returned a dataset for 534 miRNAs in 506 samples, of which 125 overlapped with the 202 GBM1 tumors. We quantile normalized these data before running downstream analysis.

1.1.4 Clinical information

Clinical data for individual patients and samples were downloaded from TCGA data-access site: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/bcr/intgen.org/minbiotab/c lin on 1/18/2011. We extracted information regarding age of diagnosis, survival time, tumor cell, and tumor nuclei. An updated version, containing information for GBM3, was accessed on 1/5/2012. The Karnofsky performance scores were extracted from the clinical data accessed on 6/22/2012.

1.2. Ovarian Cancer (OV)

1.2.1 DNA copy number data

Copy number data for Illumina 1M-Duo arrays were downloaded from TCGA Bulk Download site http://tcga-data.nci.nih.gov/tcga/findArchives.htm on 10/01/2010. We queried the Data Access Matrix by choosing

Disease: Ovarian Cancer;
Data Type: SNP;
Batch number: all;
Data Level: 2 and 3;
Platform: HAIB (Human1M Duo).

The BAF files contain 1,199,189 SNP markers and 516 paired samples. The LRR files contained 530 samples, of which 509 were paired. The overlapping set of 509 paired samples was selected for further analysis.

1.2.2 Gene expression data

Gene expression data for OV were obtained from TCGA data analysis working group. The file "TCGA_batch9-15_17-19_21-22_24.UE.txt" dated 05/05/2010 contains 11,864 genes and 524 samples, of which 504 overlap with the DNA copy number dataset.

1.3. Phillips et al. dataset

Phillips et al. data were accessed from GEO dataset GSE4271. The processed gene expression data for 56 samples were obtained at http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/.

1.4. Cahoy et al. dataset

Cahoy et al. data were accessed from GEO dataset GSE9956.

1.5. Data for microglia/macrophage

We queried of Gene Expression Omnibus dataset (http://www.ncbi.nlm.nih.gov/gds) [3] to identify gene expression profiles for microglia/macrophage cells. In January 5-7th, 2012 we searched for keywords "microglia" AND "human" and found 23 independent datasets. Among these, we selected experiments for tumor cells, and this resulted in 2 datasets, GSE25289 and GSE16119, which we used to infer the likely cell types contributing to the euploid population in MES tumors.

2. Inferring Aneuploid Genome Proportion

2.1. Data processing, DNA segmentation, and merging

Throughout this study we focused on somatic events, defined by the differences between tumor-normal pairs, thus ignoring inherited aneuploidy.

Seven GBM samples, TCGA-06-0139, TCGA-06-0160, TCGA-06-0165, TCGA-06-0167, TCGA-06-0189, TCGA-06-0240, and TCGA-06-0881 bear few copy number alterations (CNAs), and were excluded from further analysis.

As homozygous loci are uninformative for detecting changes in BAF patterns we focused on BAF data at heterozygous loci. For each tumor-normal pair, loci with BAF value >=0.9 or <=0.1 in the normal sample
were designated as homozygous. Altering the stringency of this definition did not make a major impact on AGP inference, as AGP will be driven by large aneuploid events, for which having more or fewer heterozygous loci would not substantially change the estimate of "contraction" (see below). For the heterozygous loci thus defined, we extracted tumor BAF data and generated the "folded BAF", defined as the absolute value of (BAF-0.5), for segmentation. For both GBM and OV, we performed segmentation on folded BAF using the Circular Binary Segmentation (CBS) algorithm, implemented in the R package DNAcopy with default parameters, except that "minimal markers required" was set to 5. The series of BAF change points were merged with the corresponding LRR change points, which were generated by Dr. Devin Absher at the HudsonAlpha Institute of Biotechnology using CBS [4], and were made publicly accessible as TCGA Level 3 data. As the BAF segments and LRR segments sometimes captured the same event, we merged the combined change points as follows: if a BAF change point was within 5 markers of a LRR change point, either upstream or downstream, it was removed, i.e. only the LRR breakpoint was kept, under the assumption that the two change points captured the same event, but the BAF change point was less accurately placed due to the constraint of using only heterozygous markers. After merging, small segments, defined as containing fewer than 10 BAF markers were merged with adjacent segments by removing the flanking change points. These steps resulted in a final set of CNA segments for each tumor sample.

2.2. Per-segment summary of LRR and BAF

For each segment in the final CNA call set, we re-calculated the median LRR and mean folded BAF to update these values within each merged segment. For segments with balanced parental chromosomes, BAF values at heterozygous loci are distributed as one track near 0.5, but it may not be centered exactly at 0.5. Likewise for segments with unbalanced parental chromosomes, BAF values split to two tracks, which may not fall symmetrically around 0.5. To increase the accuracy of BAF estimation we fit each segment's distribution of heterozygous BAF values as either one Gaussian distribution or the summation of two Gaussian distributions. When there were in fact two tracks but the separation between tracks was small, the summed distribution might resemble a single Gaussian distribution. We used the baseline variance of BAF as the criterion to distinguish the two cases: segments with folded BAF standard variation >=0.1 were considered as two-track segments, and fit with two Gaussians. For segments with one track, we obtained the best fitting distribution as $N(\mu, \sigma)$, and defined the folded BAF value as 0. For segments with two tracks, the best fitting distribution is $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, and the folded BAF value is $|\mu_1 - \mu_2|/2$. If the distribution cannot be fit in R or if the segment had fewer than 100 markers, the folded BAF value is taken as the mean absolute deviation around the mean: $\text{mean}(|x_i - \text{mean}(x_i)|)$, where $x_i$ is the BAF value at the i-th marker.

2.3. LRR scale-normalization:
The primary goal of Illumina's data processing algorithm is to find clusters that represent discrete genotypes. As a result, the LRR values are not linearly scaled with the true copy number changes, e.g., when the true DNA copy number drops from 2 to 1, LRR drops less than 1 unit (2-fold). Moreover, the severity of this "saturation effect" is different between amplifications and deletions. For Illumina 550K arrays, the correction factors are 0.572 for deletions and 0.553 for amplifications [5]. We re-scaled LRR segmental means by these ratios before downstream analysis.

2.4. BAF-LRR plot: canonical points and tracks

We used the BAF-LRR plot to depict the bivariate data of allele-specific copy numbers. In this plot, the folded BAF values are shown on the x-axis, and the normalized LRR values are shown on the y-axis. Each segment is plotted as a point in the BAF-LRR space, with the symbol size indicating segment length. Amplifications, deletions, and copy-neutral LOH segments are uniquely placed in the plot.

Canonical points, representing integer combinations of A and B alleles, were placed as follows. For a pair of integers \((n_B, n_T)\), where \(n_T\) denotes the copy number for the B allele, and \(n_T\) denotes the total copy number of both alleles, its x and y coordinates are:

\[
x = \left| \frac{n_B}{n_T} - 0.5 \right|, \quad y = \log_2 \left( \frac{n_T}{2} - y^0_{pl} \right)
\]

where \(y^0_{pl}\) is an adjustable offset of LRR level to reflect (1) the average ploidy of the aneuploid population, which can be a non-integer, and (2) potential alternative ploidy of the euploid population. In some tumors, the euploid portion might be \(n_T=4\) (or \(n_T=6\)) rather than \(n_T=2\), yet the normalization procedure of each sample tended to center its genomic average LRR to 0, thus an offset is needed to adjust the y-positions of the canonical points to achieve a maximal fit. We will separately fit \(n_T=2, 4,\) and 6 for the euploid population when searching for the optimal AGP (see below).

Tumor samples that contain a mixture of euploid cells and aneuploid cells will show a contraction of canonical points from its original position toward the origin, where the euploid segments reside. The paths of the contraction when AGP decreases from 1 to 0 are called canonical tracks (FigS1A and S1B). For a given \(p\), canonical points on the BAF-LRR plot can be organized into a 2D lattice, in which the near-vertical gridlines connect points of equal \(n_B\). The first line, located at the right, contains all LOH points with \((n_B, n_T)=(0,1), (0,2), (0,3), \) etc. The second line, to the left, contains \((n_B, n_T)=(1,2), (1,3), (1,4), \) etc. And the third line contains \((n_B, n_T)=(2,4), (2,5), (2,6)\), etc. They are orthogonal to the canonical tracks, which describe the movement of canonical points toward the origin \(((n_B, n_T) = (1,2))\) under shrinking values of AGP. The relative
positions of the canonical points contain information for distinguishing the alternative ploidy of the euploid genome, which define the origin of contraction for the aneuploid segments.

2.5. Inference of Aneuploid Genome Proportion:

A. Definition of Euploid Segments: On a BAF-LRR plot, euploid regions tend to land near the point \((x,y) = (0,0)\). But due to random noise and various technical artifacts some segments may lie slightly off \((0,0)\). Precise assignment of the near-(0,0) segments into the euploid cluster is important because it affects the relative distances to other canonical points and the AGP estimates. To anchor its position, we first ran k-means clustering 10 times on the observed BAF-LRR values for all segments. For each run, we identified the segments that belong to the cluster nearest to \((0,0)\), and tagged them as euploid. Segments that were tagged more than 6 times out of 10 were used to define the seed position, located at the cluster mean \((x_s, y_s)\) of the tagged segments, weighted by segment size. Second, we examined each non-seed segment to see if its coordinates \((x, y)\) were sufficiently close to the seed location. If \(|x-x_s| \leq \sigma_{BAF}\) and \(|y-y_s| \leq \sigma_{LRR}\), this segment was "pulled" into the euploid cluster, where \(\sigma_{BAF} = 0.04\), the empirically estimated standard variation of BAF, and \(\sigma_{LRR} = 0.16\), the empirical standard variation of LRR. This step was iterated, with more segments joining the euploid cluster until the cluster was no longer updated. The final coordinate of the weighted center of the euploid cluster is denoted as \((x_0^f, y_0^f)\).

B. Canonical Points under admixture: Consider the mixture containing a population of cells carrying an aneuploid segment \(nB, nT\), and a second population of cells carrying an euploid segment \(n_{pl}, 2n_{pl}\), \(n_{pl} = 1, 2, \) or 3, and that the euploid portion makes up \(1-p\) of the total (i.e., AGP = \(p\)). The coordinates for the mixed population are given by:

\[
\begin{align*}
x &= \left| \frac{pxB + (1-p)x_{pl}}{pxB + 2(1-p)x_{pl}} - 0.5 \right| + x_0^f \\
y &= \log_2 \left( \frac{pxB + 2nx_{pl}}{2} \right) - y_{pl}^0 + y_0^f
\end{align*}
\]  

(2)

C. Aneuploid genome Proportion: For each sample, after the euploid cluster was defined, we searched for the best fitting \(p\) and \(n_{pl}\) by screening the parameter space of \(p \in (0,1)\), and \(n_{pl} \in \{1,2,3,4\}\). We did not include the canonical point for homozygous deletions because their BAF or LRR values are not determined.

For each \((p, n_{pl})\) combination being considered, the canonical points were calculated and the penalized sum of squared distance \(SSD_{pl}^p(p)\) was calculated as:
\[ SSD'_{pl}(p) = \sum_{i=1}^{\Omega} d_{\text{min},pl}^i(p) + D_{pl}(p) \]  

where \( i \) is the segment index and \( \Omega \) represents all segments in this sample (excluding those in the euploid cluster), \( pl \) stands for ploidy \( n_{pl} \), \( d_{\text{min},pl}^i \) is the squared distance of the segment to the nearest canonical point.

\( D_{pl}(p) \) is the penalty score for applying a larger \( n_{pl} \), as increasing \( n_{pl} \) results in a larger number of available canonical points to fit with, and consequently a smaller sum of squared distances. Applying this penalty will avoid making excessively high euploid baseline assignments. \( D_{pl}(p) \) is linearly correlated with the approximate distance between adjacent canonical points such as \((2,4)\) and \((1,4)\). We defined

\[ D_{pl}(p) = p_n \times (n_{pl} - 1) \times \text{distance between canonical points } (2,4) \text{ and } (1,4) \]

Penalty \( P_n \) was manually chosen as \( p_n = 200 \) as it generated the most reasonable assignments.

Best fitting AGP value was determined by the smallest \( SSD'_{pl}(p) \). The scanning of the parameters was carried out in two steps to increase computation speed: a coarse scan of \( p \in (0.05, 0.95) \) at an interval of 0.05 was performed, with a best fitting value \( p^* \) determined. Then, in the second step, a finer scan of \( p \in (p^*-0.1, p^*+0.1) \) at an interval of 0.02 was performed to refine the final score. The model also yielded the optimal ploidy value, resulting in 135 diploid, 127 tetraploid and 22 hexaploid samples for GBM1 and GBM2. For OV, we identified 23 diploid, 64 triploid, 273 tetraploid, 127 hexaploid and 22 octoploid samples.

2.6. Genomic features and QC measures.

We extracted multiple genomic measures for each tumor sample, including percent of genome changed (PC) and percent of genome on canonical points (PoP). Let \( P_1 \) denote the proportion of genome in the euploid cluster, thus \( 1 - P_1 \) of genome has been alerted either in copy number or in the B allele frequency. We define:

\[ PC = 1 - P_1 \]

Extremely low PC indicates that there is insufficient amount of CNAs to inform model parameters and should be considered as having yielded low-quality AGP estimates.

Actual segments on the BAF-LRR plot may fall near or far from a canonical point for a given AGP. We quantify these deviations as measures of goodness-of-fit by the two-way mixing model with optimal AGP. If there is more than one dominant aneuploid population mixed with the euploid population, some segments would have a different mixing ratio than some other segments, and as a consequence, the fit at a single AGP would not be suitable for all segments, and this can be reflected by a low rate of "Percent-on-Point", defined as the
proportion of segments falling within $\sigma_{BAF} = 0.04$ and $\sigma_{LRR} = 0.16$ of a canonical point. If this proportion is $P_2$ of the genome, we define

$$PoP = \frac{P_2}{1 - P_1}$$

As aneuploid cells carry variable copy numbers at different segments, it is no longer sufficient to define an integer ploidy as a genomewide attribute of a tumor. However we define average aneuploid ploidy as the genomewide mean copy number for the aneuploid cells of the tumor, and average overall ploidy as the weighted average of euploid and aneuploid populations. Specifically, as we assign ploidy status for every segment in the aneuploid genome, the average aneuploid ploidy, $\Psi_{\text{tumor}}$, can be defined as the length-weighted means of segmental ploidy. The average overall ploidy of the sample, containing $p$ of aneuploid genome and $1-p$ of euploid genome, is

$$\Psi_{\text{overall}} = \Psi_{\text{tumor}} \times p + 2 \times n_{p} \times (1-p)$$

Other tumor genomic features, including percentage of genome amplified (%amp), deleted (%del), percentage of hemizygous deletion (%del.loh), and percent of genome underwent loss of heterozygosity (%LOH), were also extracted.

We use a bootstrap method to estimate the confidence intervals of AGP. A weighted resampling was performed for each sample, such that each segment was chosen with the probability proportional to its size. Permutation was done 100 times for each sample, and for each run, 80% segments were resampled and AGP recalculated. The standard deviation, and the 2.5%, 50%, and 97.5% quantiles of AGP, were extracted and included in Table S1. The 2.5-97.5% confidence interval (CI) can be calculated from these results. We also calculated the relative confidence interval (rCI) as the ratio of CI to the median of AGP. Eighty-eight percent of samples had rCI less than 100%.

PoPs were negatively correlated with AGPs (Figure S1C, Spearman's $\rho = -0.40$, $P = 3.3 \times 10^{-12}$), suggesting that samples better accounted for by the model (i.e., higher PoP) tend to have lower AGP estimates, thus our method may have over-estimated AGP for poorly fit samples. The CIs, however, were positively correlated with AGPs (Figure S1D, $\rho = 0.13$, $P = 0.03$).

2.7. Validation of AGP algorithm

The validation dataset, GSE11976, was downloaded from the Gene Expression Omnibus (GEO) [6]. It contained 11 samples of DNA from the human breast carcinoma cell line CRL2324 mixed with DNA from the lymphoblastoid cell line HCC1395BL with known mixing ratios. Samples were measured across 370,404
SNP loci by using the Illumina HumanCNV370-Duov1 BeadChips. Known CNVs in HCC1395BL were removed so that HCC1395BL DNA represents the euploid portion of the mixture. AGP value for each sample was calculated using our algorithm, and compared with the mixing percentage. Pearson’s correlation coefficient $r = 0.979$, confirming that our method accurately estimated aneuploid content.

2.8. Impact of goodness-of-fit on comparisons with histopathologic reports and expression PCs.

The difference with histopathologic reports was not explained by tumors with worse fit in our model, or greater estimation uncertainty. Among 243 GBMs with PoP > 0.6 (i.e., better fit), AGPs were lower by 9% and 20% than "tumor nuclei" and "tumor cells", respectively. Similarly, among 131 GBMs with CI < 20% (better fit), AGPs were lower by 5% and 16%, respectively, and the correlation remained low for "tumor nuclei" ($\rho = 0.06$, $P = 0.52$) and "tumor cells" ($\rho = 0.13$, $P = 0.13$).

Samples with better fit in the two-way mixing model showed the same level of correlation with PC1 as all samples: among 109 GBMs with PoP > 0.7, $r = 0.63$ ($P = 3.1 \times 10^{-13}$); and among 355 OVs with PoP $\geq$ 0.7, $r = 0.65$ ($P < 2.2 \times 10^{-16}$).

3. Classification of Non-Proneural GBM

3.1. A two-step procedure that relies on GBM1-GBM2 mutual validation

In PCA, Non-Proneural GBMs described a nearly continuous distribution (Figure 1C), in which the low-AGP samples aggregated to the left, and there were no clearly separated sub-groups in this type of plot. For practical reasons it is often useful to partition seemingly continuously varying samples into discrete classes in order to draw broad biological conclusions, and to aid clinical decision-making. With high-dimension data, however, even samples from a homogeneous distribution can be divided into pre-specified numbers of clusters; but the result can be unstable, and be sensitive to samples included, or the statistical algorithms applied. Self-aggregating algorithms such as hierarchical clustering or k-means clustering will always produce a desired number of clusters; and Consensus Clustering is prone to exaggerate cluster stability (Senbabaoglu et al., manuscript under preparation). In CC, class assignment can be sensitive to outlier samples, chance occurrence of tightly clustered samples, and the markers used. In addition, as gene-gene correlation is ubiquitously observed, and if groups of highly correlated genes appeared in both the test cohort and validation cohort, it is easy to find that the most discriminating genes in one cohort are "validated" in the second cohort by observing similar clustering patterns.

To address these methodological challenges, we placed major emphases on mutual validation between the GBM1 and GBM2 cohorts rather than first selecting the most informative genes in one and testing them in
another. We also focused on the Non-Proneural samples. We ran K-means-based CC on quantile-normalized gene expression data for GBM1, and separately for GBM2, recording the class assignments for K = 2, 3, 4 (K is the number of clusters) for both cohorts. To assess classification concordance between GBM1 and GBM2, we calculated the cross-correlation matrix between every sample in GBM1 and every sample in GBM2, and displayed the resulting matrix where samples were grouped by class assignments independently obtained for the two cohorts (Figure S7A-B). If samples of a given class in GBM1 showed high correlation coefficients (r) with those of a particular class in GBM2, and showed low r values with other GBM2 classes, the class discovery was considered mutually validated. Conversely, if the classes did not show a one-to-one correspondence between the two independent cohorts, we considered the class definition poorly replicated. Figure S7A showed the GBM1-GBM2 cross-correlation matrix for K=2, where the two classes defined in GBM1 could be matched, one-to-one, to the two classes independently defined in GBM2. In comparison, K=3 or 4 yielded substantially worse matching (Figure S8A-B).

At K=2, one of the two classes for GBM1 contained all the 37 samples in the Mesenchymal group defined previously (Figure S7A). We therefore named it the Mesenchymal (MES) group even though it now also contained 4 former Neural/Proneural samples and 11 former Classical samples. The other class showed hints of finer structure in Figure S7A; and this was explored by repeating the analysis described above within this class. This led to a further split into 2 subclasses (n=27 and 29, Figure S7B), with K=2 being better than k=3 or 4 (Figure S8C-D). One of the subclasses was dominated by the previously defined Classical samples, and was thus named the Classical group even though it also contained 5 Neural and 1 Proneural samples. The other subclass, with a mixture of previously defined Proneural (but recognized as non-Neuronal in this work) and Neural tumors, was named Proliferative for its similarity with the Proliferative samples identified by Phillips et al. [7]. Attempts to identify further subtypes within the Proliferative group were not supported by mutual validation between GBM1 and GBM2 (not shown). This led us to conclude that the G-CIMP-minus (G-CIMP-) subset of previously defined Proneural samples did not form a distinct group. In other words, there wasn't a second, self-contained Proneural group in the current GBM dataset, although it is possible that a larger sample size in future studies could have the power to reveal finer splits. In all, we identified three subclasses for Non-Proneural GBM through a two-stage, stepwise clustering procedure, with optimal K=2 at both stages, and supported by concordance between GBM1 and GBM2. The resulting assignments were different (by 12% of samples) from those assigned by a one-stage, K=3 approach. We consider the two-stage approach more appropriate because the finer division in the second stage is not affected by the more divergent profiles of the two main classes identified in the first stage. The three newly identified Non-Proneural GBM classes are visually coherent on the gene expression PCA plot, for both GBM1 (Figure S7C) and GBM2 (Figure S7D).
3.2. Comparison with previous studies

Phillips et al. [7] proposed a three-class system for GBM: Proneural, Proliferative and Mesenchymal. Verhaak et al. [2] reported four classes for TCGA samples: Proneural, Neural, Classical, and Mesenchymal; and in this work we described a revised four-class system for the same dataset as in Verhaak et al.. In order to summarize how these systems have evolved (i.e., how different classes correspond to each other), we first reanalyzed the Phillips’ data, which were made publicly available and a subset of 56 samples were subsequently processed to combine two technical platforms [2]. Among the 56 samples we first observed that the Proneural samples in Phillips’ study showed high similarities to our PN/G-CIMP+ GBMs in terms of patient age, survival time, and patterns of CNAs (not shown). For the remaining 46 samples, which were designated Non-Proneural GBMs here, we followed the procedure of Phillips et al. to select 584 genes most highly correlated with patients’ survival time (out of 1,740 most variable genes) and performed k-means clustering, using cross-correlation with TCGA’s GBM1 to find the optimal number of classes. Again, K=2 yielded the best match for both steps in a two-step procedure (Figure S10), leading to the recognition of 19 Mesenchymal, 14 Proliferative and 13 Classical samples. This new three-way classification of 46 Non-Proneural samples showed better cohesion on the PCA plot (Figure S11) than the original classification, and this could be explained by noting that the latter was based on a different set of (and much fewer) genes.

By tracking the class reassignments between the two datasets and between the original classification and our revised classification (Figure S9A-B), we documented the commonalities and differences among different classification systems (Figure S9C, Table S3). The MES class was reproducibly identified in both datasets and in both the original and the revised schemes. The original Proneural group for TCGA was split into (1) the PN/G-CIMP+ group, which is equivalent to Phillips’ Proneural group, and (2) Non-Proneural-Proneural (N-P-P), which was merged with the original Neural samples to form the revised Proliferative group, which closely resembles Phillips’ Proliferative group. However, some of Phillips’ Proliferative samples split and formed the revised Classical group, which closely resembles the original Classical group for TCGA samples. In sum, a major revision of the Verhaak et al. classification is in recognizing that the Proneural group contains two distinct subgroups, one of which, PN/G-CIMP+, is well separated from the other three classes by CNA patterns, IDH1 mutations, patient age, and outcome. For the Philipps' dataset, a major revision is in separating the original Proliferative group into the revised Proliferative and Classical groups.

The primary reason that the Proneural/G-CIMP+ class was previously mis-grouped with some Neural samples is that their gene expression signatures, when viewed without other genomic data, were not sufficiently distinctive, because Proneural samples share a cell type-specific signature with the Neural samples (renamed as the Proliferative samples in our system). It was only by integrating the CNA data (this work) or by using the
methylation data [8] that the Proneural/G-CIMP+ group became evident. The Phillips' study did not miss this group because the authors selected genes strongly correlated with survival time rather than those showing the largest variation. Since patients in the Proneural/G-CIMP+ group survived longer, genes that were most informative for recognizing this group were used in that study.

3.3. Marker genes for Non-Proneural subgroups

We individually examined marker genes highlighted in previous studies. EGFR overexpression and amplification, and CDKN2A down-regulation, are hallmarks for Classical GBMs, and this was true in our classification (Table S4). The Proliferative group showed higher expression for two well-known proliferative markers: proliferating cell nuclear antigen (PCNA) and topoisomerase II α (TOP2A). Finally, in our MES group we observed higher expression of the Mesenchymal marker, CHI3L1, and genes involved in tumor necrosis factor superfamily pathway and NF-κB pathway, including TRADD, RELB and TNFRSF1A.

3.4. Comparison with previous assignment in terms of mapping to known cell types.

The observed resemblance to known cell types were generally consistent with what was reported previously [2], but with important differences. First, the former Neural group did not show clear mapping to any cell type. Second, the mapping to reference cell types is much stronger with the new system: the difference (D) of the mean correlation coefficients between the mapped diagonal blocks and the off-diagonal blocks of the correlation matrix (Figure 4A-C) was 0.562 for the new classes, much higher than in the previously reported classes (D = 0.247) even when we counted the best mapped blocks for the latter.

4. Supervised class assignment in GBM3

We selected the most informative gene expression markers for the Non-Proneural GBM classes using GBM1-2 data, and performed class assignments on independent data in GBM3. We first "locked down" the class assignment for GBM1 and GBM2. For each of the three pairwise comparisons among the three Non-Proneural GBM groups we extracted the 300 most significantly different genes by the student-t test on the 1,740 genes. The union of the three gene sets, including 651 unique genes, was used to calculate the distance of each GBM3 sample to the centroids of the three Non-Proneural classes as defined by GBM1-2. The class assignment of GBM3 samples was determined by the centroid with the shortest distance.

In parallel, we performed class assignment for GBM3 according to the previous system. This was done by using the four-class centroids involving 840 genes selected in Verhaak et al.
References


