PLOS ONE

RESEARCH ARTICLE

Platform-Independent Genome-Wide Pattern of DNA Copy-Number Alterations Predicting Astrocytoma Survival and Response to Treatment Revealed by the GSVD Formulated as a Comparative Spectral Decomposition

Katherine A. Aiello[1,2], **Orly Alter**[1,2,3,4]*

1 Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah, United States of America, **2** Department of Bioengineering, University of Utah, Salt Lake City, Utah, United States of America, **3** Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, United States of America, **4** Department of Human Genetics, University of Utah, Salt Lake City, Utah, United States of America

* orly@sci.utah.edu

OPEN ACCESS

Citation: Aiello KA, Alter O (2016) Platform-Independent Genome-Wide Pattern of DNA Copy-Number Alterations Predicting Astrocytoma Survival and Response to Treatment Revealed by the GSVD Formulated as a Comparative Spectral Decomposition. PLoS ONE 11(10): e0164546. doi:10.1371/journal.pone.0164546

Editor: Shyamal D Peddada, National Institute of Environmental Health Sciences, UNITED STATES

Received: June 6, 2016

Accepted: September 27, 2016

Published: October 31, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Data are also available from http://www.alterlab.org/astrocytoma_prognosis/.

Funding: This research was supported by National Cancer Institute (NCI) U01 Grant CA-202144, http://physics.cancer.gov/network/UniversityofUtah.aspx, and the Utah Science, Technology, and Research (USTAR) Initiative (to OA). This research was also supported by National Center for Advancing Translational Sciences

Abstract

We use the generalized singular value decomposition (GSVD), formulated as a comparative spectral decomposition, to model patient-matched grades III and II, i.e., lower-grade astrocytoma (LGA) brain tumor and normal DNA copy-number profiles. A genome-wide tumor-exclusive pattern of DNA copy-number alterations (CNAs) is revealed, encompassed in that previously uncovered in glioblastoma (GBM), i.e., grade IV astrocytoma, where GBM-specific CNAs encode for enhanced opportunities for transformation and proliferation via growth and developmental signaling pathways in GBM relative to LGA. The GSVD separates the LGA pattern from other sources of biological and experimental variation, common to both, or exclusive to one of the tumor and normal datasets. We find, first, and computationally validate, that the LGA pattern is correlated with a patient's survival and response to treatment. Second, the GBM pattern identifies among the LGA patients a subtype, statistically indistinguishable from that among the GBM patients, where the CNA genotype is correlated with an approximately one-year survival phenotype. Third, cross-platform classification of the Affymetrix-measured LGA and GBM profiles by using the Agilent-derived GBM pattern shows that the GBM pattern is a platform-independent predictor of astrocytoma outcome. Statistically, the pattern is a better predictor (corresponding to greater median survival time difference, proportional hazard ratio, and concordance index) than the patient's age and the tumor's grade, which are the best indicators of astrocytoma currently in clinical use, and laboratory tests. The pattern is also statistically independent of these indicators, and, combined with either one, is an even better predictor of astrocytoma outcome. Recurring DNA CNAs have been observed in astrocytoma tumors' genomes for decades, however, copy-number subtypes that are predictive of patients' outcomes were

(NCATS) UL1 Grant TR-001067. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: OA is a co-founder of and an equity holder in Eigengene, Inc. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

not identified before. This is despite the growing number of datasets recording different aspects of the disease, and due to an existing fundamental need for mathematical frameworks that can simultaneously find similarities and dissimilarities across the datasets. This illustrates the ability of comparative spectral decompositions to find what other methods miss.

Introduction

Recurring DNA copy-number alterations (CNAs) have been recognized as a hallmark of cancer for >100 years [1–3], yet what these alterations imply about a solid tumor's development and progression, and a patient's diagnosis, prognosis, and treatment remains poorly understood. This is despite the growing number of high-dimensional datasets, recording different aspects of a single disease, such as DNA copy-number profiles of two or more cell types from the same set of patients, possibly measured more than once by different platforms. This is due to an existing fundamental need for mathematical frameworks that can create a single coherent model from, i.e., simultaneously find similarities and dissimilarities across such datasets, arranged in two or more tables, of two or possibly more dimensions, i.e., matrices or tensors, of matched columns but independent rows.

A recent comparison of DNA copy-number profiles of primary tumor and normal cells from the same set of ovarian serous cystadenocarcinoma (OV) patients, measured by the same set of platforms, uncovered three tumor-exclusive platform-consistent chromosome arm-wide patterns of DNA CNAs that are correlated with a patient's survival and response to platinum therapy [4]. The datasets had been publicly available in the Cancer Genome Atlas (TCGA) since 2011, and analyzed by using several methods [5]. The patterns, however, remained unknown until the datasets were modeled in 2015 by using a novel comparative spectral decomposition, the tensor generalized singular value decomposition (GSVD). For >30 years prior, statistically the best indicator of OV survival was the tumor's stage at diagnosis [6]. About 25% of primary OV tumors are resistant to platinum therapy, the first-line treatment, yet no diagnostic existed to distinguish resistant from sensitive tumors before the treatment [7].

A previous comparison of copy-number profiles of primary tumor and normal cells from the same set of glioblastoma (GBM) brain cancer patients, uncovered a tumor-exclusive genome-wide pattern of CNAs that is correlated with a patient's survival and response to chemotherapy [8]. The datasets had been publicly available in TCGA since 2008 [9]. The pattern, however, remained unknown until the datasets were modeled in 2012 by using the GSVD [10–16], formulated as a comparative spectral decomposition [17] (see also [18–30]). For >50 years prior, statistically the best indicator of GBM outcome was the patient's age at diagnosis [31–33] (see also [34, 35]). Copy-number subtypes of GBM, i.e., grade IV astrocytoma, which are predictive of survival and response to treatment were not conclusively identified [36, 37].

Results

GSVD Comparison of Patient-Matched LGA Brain Tumor and Normal DNA Copy-Number Profiles

To identify CNAs that might predict grades III and II, i.e., lower-grade astrocytoma (LGA) patients' outcomes, we, therefore, used the GSVD to model TCGA patient-matched LGA

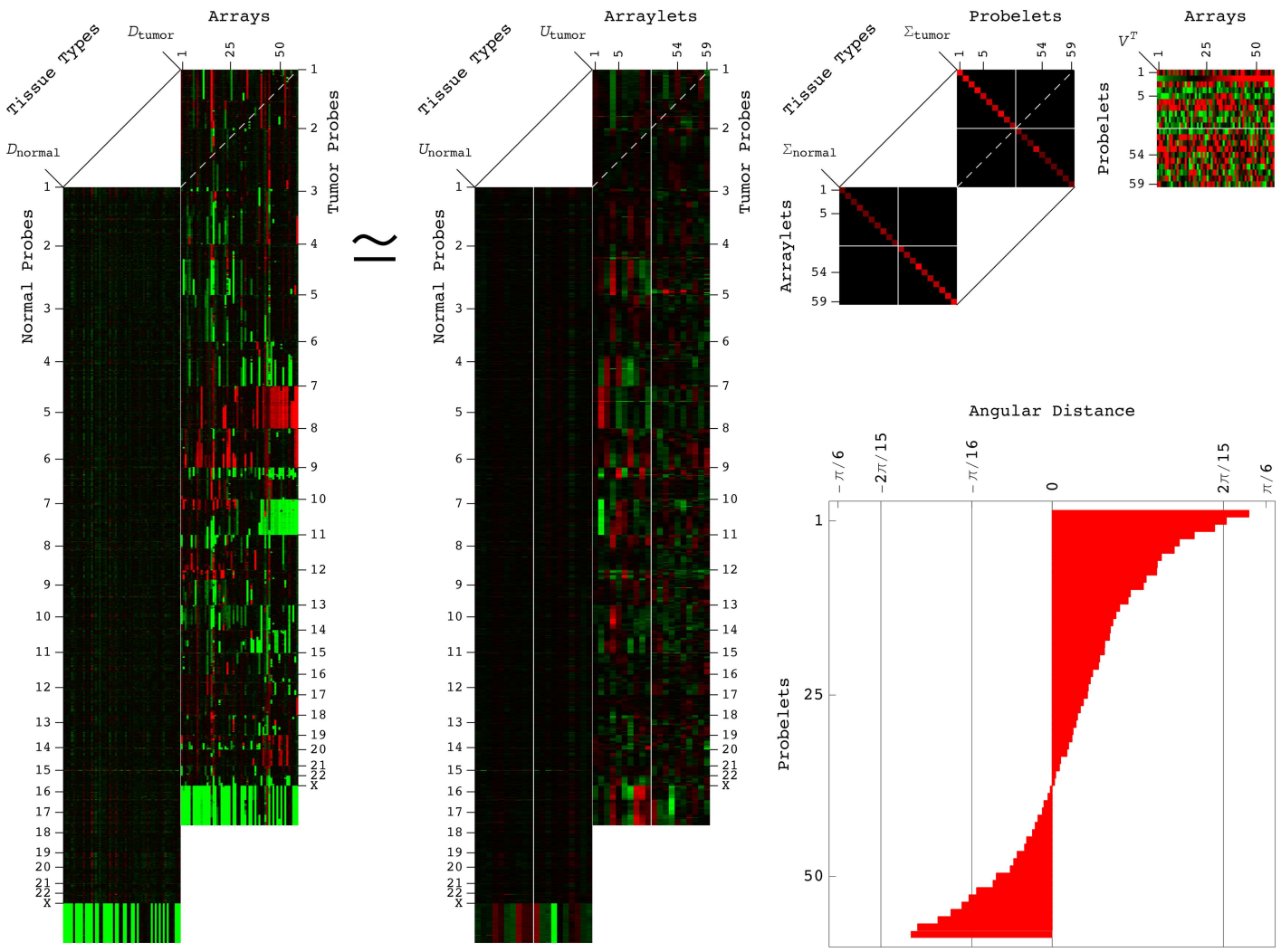

Fig 1. GSVD of the patient-matched LGA tumor and normal DNA copy-number profiles. The structure of the LGA discovery, tumor and normal datasets D_i is that of two matrices of 59 matched columns (i.e., patients), and 933,827, not necessarily matched or equal in numbers, rows (i.e., tumor and normal genomic regions, or Affymetrix probes). The GSVD of Eq (1) simultaneously separates the datasets into a single set of normalized, not necessarily orthogonal probelets V^T (i.e., patterns of variation across the patients), which are identical for both datasets, but correspond to different sets of generalized singular values Σ_i (i.e., weights, or superposition coefficients) and orthonormal arraylets U_i (i.e., patterns of variation across the genome) in each dataset. The GSVD is depicted in a raster display, with relative DNA copy-number gain (red), no change (black), and loss (green), which explicitly shows only the first through the 10th, and the 50th through the 59th probelets and corresponding tumor and normal arraylets, and tumor and normal generalized singular values. The angular distances of Eq (4) define the significance of each probelet in the tumor dataset relative to its significance in the normal dataset in terms of the ratio of the corresponding tumor to normal generalized singular values [17]. The inset bar chart shows that the angular distances largest in magnitude correspond to the first and second probelets, and are > $2\pi/15$, whereas the magnitude of the angular distance that corresponds to the 53rd probelet is < $\pi/16$.

doi:10.1371/journal.pone.0164546.g001

tumor and normal DNA copy-number profiles [38]. We selected patient-matched Affymetrix-measured DNA copy-number profiles of primary LGA tumor and normal tissue samples from a discovery set of 59 patients (Methods and S1 Dataset). The structure of these tumor and normal datasets is that of two full column-rank matrices $D_1 \in \mathbb{R}^{M_1 \times N}$ and $D_2 \in \mathbb{R}^{M_2 \times N}$ of $N = 59$ matched columns (i.e., patients), but independent, i.e., not necessarily matched or equal in numbers M_1, $M_2 = 933{,}827$ rows (i.e., tumor and normal genomic regions, or Affymetrix probes), where $M_1, M_2 \gg N$ (Fig 1).

The GSVD simultaneously separates the two matrices, or tumor- and normal-specific datasets, into paired weighted sums of outer products, of each normalized, not necessarily orthogonal right basis vector, or "probelet" v_n^T (i.e., a pattern of variation across the patients), which is identical for both datasets, combined with one of the two corresponding orthonormal left basis vectors, or "tumor arraylet" $u_{1,n}$ and "normal arraylet" $u_{2,n}$ (i.e., the tumor- and normal-specific patterns of variation across the genome),

$$D_i = U_i \Sigma_i V^T = \sum_{n=1}^{N} \sigma_{i,n} u_{i,n} \otimes v_n^T, \qquad i = 1, 2. \tag{1}$$

The significance of a probelet v_n^T in either the tumor dataset D_1 or the normal dataset D_2, in terms of the "generalized fraction" of the overall information that it captures in the dataset, is proportional to the corresponding nonnegative generalized singular value $\sigma_{1,n}$ or $\sigma_{2,n}$, respectively,

$$p_{i,n} = \sigma_{i,n}^2 / \sum_{n=1}^{N} \sigma_{i,n}^2, \qquad i = 1, 2. \tag{2}$$

The "generalized normalized Shannon entropy" is defined to measure the complexity of each dataset in terms of the distribution of the overall information in the dataset among the probelets,

$$0 \leq d_i = -(\log N)^{-1} \sum_{n=1}^{N} p_{i,n} \log p_{i,n} \leq 1, \qquad i = 1, 2. \tag{3}$$

An entropy of zero corresponds to an ordered and redundant dataset, in which all the information is captured by a single probelet. An entropy of one corresponds to a disordered and random dataset, in which all probelets are of equal significance.

Following the relation of the GSVD to the cosine-sine (CS) decomposition [14], the significance of a probelet v_n^T in the tumor dataset D_1 relative to its significance in the normal dataset D_2 is defined by the "angular distance" θ_n [17],

$$-\pi/4 \leq \theta_n = \arctan(\sigma_{1,n}/\sigma_{2,n}) - \pi/4 \leq \pi/4. \tag{4}$$

Probelets for which $\theta_n \sim \pm\pi/4$ are exclusive to either the tumor or the normal dataset, respectively, whereas probelets for which $|\theta_n| \sim 0$ are common to both. The probelets are arranged in decreasing order of their angular distances, i.e., their significance in the tumor relative to the normal dataset. The GSVD is unique, except in degenerate subspaces, defined by subsets of equal pairs of generalized singular values $\sigma_{1,n}$ and $\sigma_{2,n}$, and up to phase factors of ±1 of each probelet v_n^T and the corresponding tumor and normal arraylets $u_{1,n}$ and $u_{2,n}$.

We find that the two most tumor-exclusive patterns of variation across the patients, i.e., the first and second probelets, with angular distances $\theta_1, \theta_2 > 2\pi/15$, are also the first and third most significant probelets in the tumor dataset, with >8% and 5% of the information in this dataset, respectively (Fig A in S1 Appendix). The 53rd probelet, which with ∼10% of the information is the most significant probelet in the normal dataset, is approximately common to both datasets with $|\theta_{53}| < \pi/16$.

The GSVD, therefore, creates a single coherent model of the two datasets by simultaneously identifying unique probelets that are significant in, and common to the two datasets, as well as those that are significant in, and exclusive to either one of the datasets. We interpret the model accordingly, in terms of the biological and experimental phenomena that are common to the

LGA tumor and normal profiles, as well as those that are exclusive to the LGA tumor or the normal profiles.

The GSVD Reveals a Genome-Wide LGA Tumor-Exclusive Pattern of CNAs Encompassed in the GBM Pattern. In a previous GSVD comparison of patient-matched Agilent-measured DNA copy-number profiles of primary GBM tumor and normal samples, we found that the second most GBM tumor-exclusive tumor arraylet describes a genome-wide pattern of co-occurring CNAs that is correlated with a GBM patient's outcome [8]. Now, we find that the second LGA tumor arraylet describes a genome-wide pattern of co-occurring CNAs across the Affymetrix probes, which is similar to the GBM pattern (Figs 2 and 3, and Fig B in S1 Appendix). To compare the LGA to the GBM pattern, we assigned to the LGA pattern CNAs in the chromosomes and chromosome arms as well as the genomic segments that were identified in the GBM pattern (S2 Dataset). We find that the LGA pattern is encompassed in the GBM pattern. Chromosomes, chromosome arms, and segments that are amplified or deleted in the LGA pattern are also amplified or deleted in the GBM pattern, respectively, and at a greater magnitude; some of those that show no copy-number change in the LGA pattern are amplified or deleted in the GBM pattern.

Dominant in the LGA pattern, but at a lesser magnitude than in the GBM pattern, are the known, GBM-associated gain of chromosome 7 and loss of chromosome 10 [36, 37]. Also dominant in the LGA pattern, also at a lesser magnitude than in the GBM pattern, are GBM-associated focal CNAs [8] (see also [9, 39]). Among these, we find amplifications and deletions that contribute to decreased activity of the tumor suppressor protein p53. These include gains of segments containing the p53-inactivating protein-encoding *MDM4* (1q32.1) and the p53-degrading protein-encoding *MDM2* (12q15), and losses of segments containing *CDKN2A* and *CDKN2B* (9p21.3), and *PTEN* (10q23.31). The tumor suppressor protein encoded by *PTEN* negatively regulates the Mdm2 protein via the Akt pathway. Of the three known transcript variants of *CDKN2A*, one encodes $p14^{ARF}$, which is a p53-stabilizing, Mdm2-sequestering protein. The other two variants encode isoforms of the tumor suppressor protein $p16^{INK4A}$. *CDKN2B* encodes for the transforming growth factor-β (TGF-β) -induced growth inhibitor $p15^{INK4B}$ [40]. Together with the retinoblastoma (Rb) protein tumor suppressor, and in parallel to p53 and $p14^{ARF}$, $p16^{INK4A}$ and $p15^{INK4B}$ act at a checkpoint for human normal to tumor cell transformation, promoting cell cycle arrest, apoptosis, and senescence in response to rat sarcoma virus (Ras) -mediated hyperactive growth factor signaling [41–44]. Amplifications that are involved in increased growth factor signaling among the GBM-associated LGA-shared CNAs include gains of segments containing the epidermal growth factor receptor *EGFR* (7p11.2), the hepatocyte growth factor receptor *MET* (7q31.2), and the fibroblast growth factor receptor (FGFR) substrate *FRS2* (12q15) [45] (Fig C in S1 Appendix).

Additional LGA- and GBM-shared CNAs contribute to decreased activity of the tumor suppressor protein Ptch1, and increased downstream conversion of the oncogenes Gli1–3 into transcriptional activators by the Hedgehog (Hh) signaling pathway. These include gains of segments containing the Hh ligand-encoding *SHH* (7q36.3) and the Hh signal-transducing protein-encoding *SMO* (7q32.1), and a loss of a segment containing the Hh negative regulator protein-encoding *SUFU* (10q24.32) [46]. Note that reduced Ptch1 activity is also shared by the brain cancer medulloblastoma, where it was shown to contribute to the development of the tumor [47, 48] (Fig D in S1 Appendix).

The GBM pattern consists of additional CNAs that are missing from the LGA pattern, including the GBM-associated loss of the short arm of chromosome 9 (9p), and the long arm of chromosome 22 (22q). Among the GBM-specific CNAs we find amplifications that contribute to decreased Rb activity. These include gains of segments containing the viral protein-binding Rb region-interacting protein-encoding *KDM5A* (12p13.33) [49], the Rb-phosphorylating

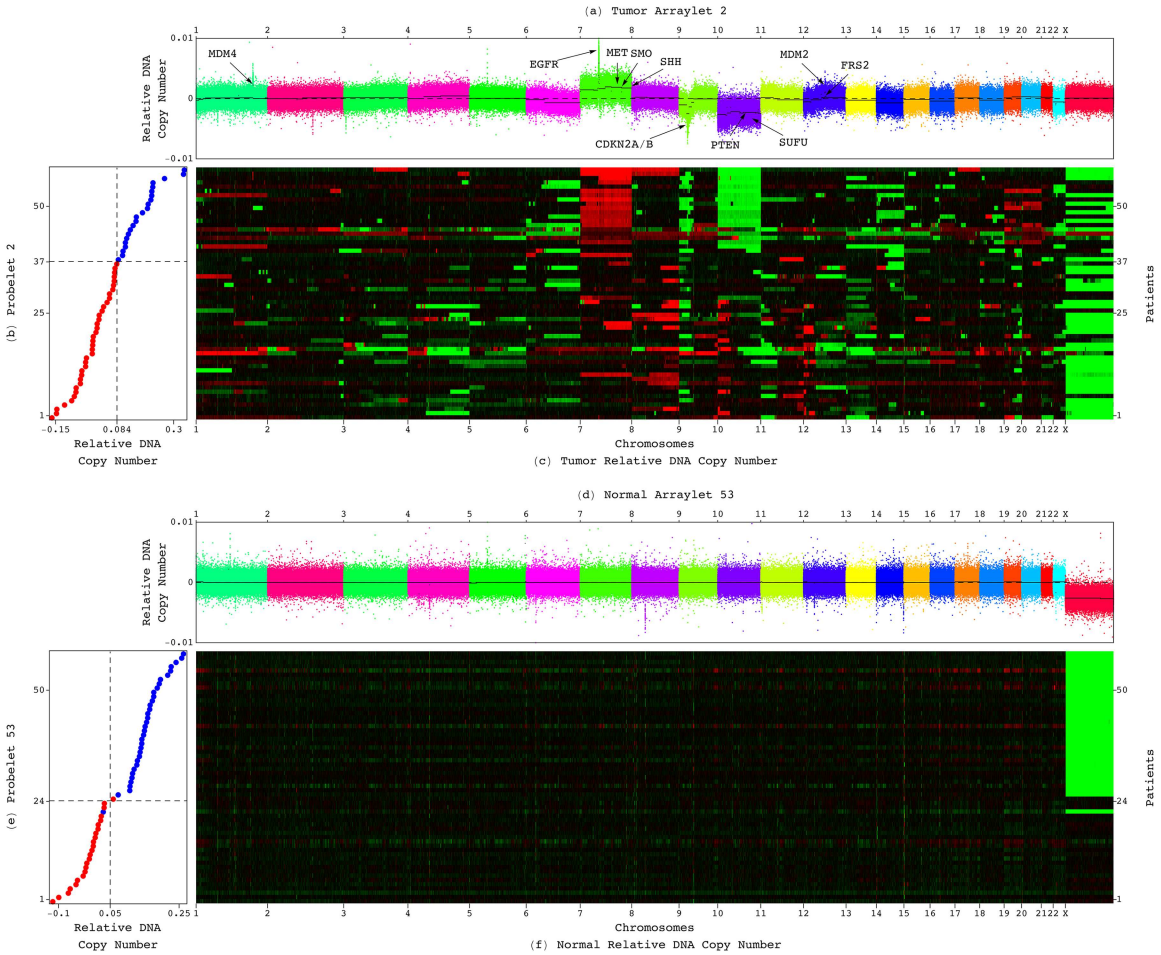

Fig 2. Significant probelets and corresponding tumor and normal arraylets revealed by the GSVD of the LGA discovery datasets. (*a*) Plot of the second most LGA tumor-exclusive tumor arraylet describes a genome-wide pattern of co-occurring CNAs across 933,827 Affymetrix probes. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. This LGA pattern is encompassed in a GBM pattern, which was previously uncovered by the GSVD [8]. Segments (black lines) that were identified in the GBM pattern, and are amplified or deleted in the LGA pattern, are also amplified or deleted in the GBM pattern, respectively, and at a greater magnitude (Fig 3). The GBM-associated LGA-shared focal CNAs (black) include, e.g., a gain of a segment on chromosome 1 containing *MDM4*. (*b*) Plot of the second LGA probelet describes the variation of the weight, or superposition coefficient of the LGA pattern in the tumor profiles of the 59 patients. The second probelet classifies the patients into two groups of low (red) and high (blue) weights, which are of statistically significantly different prognoses (Fig 4). (*c*) Raster display of the tumor dataset shows the correspondence between the tumor profiles and the second LGA probelet and tumor arraylet. (*d*) Plot of the 53rd LGA normal arraylet, which is the most significant in the normal dataset, describes a deletion of the X chromosome. (*e*) Plot of the 53rd LGA probelet, which is approximately common to the tumor and normal datasets, describes a classification of the patients by gender into females (red) and males (blue). The corresponding hypergeometric *P*-value is $<10^{-13}$. (*f*) Raster display of the normal dataset shows the male-specific X chromosome deletion across the normal genomes. This biological variation is conserved in the patient-matched LGA tumor genomes. The GSVD separates this variation from the second LGA tumor arraylet.

doi:10.1371/journal.pone.0164546.g002

protein-encoding *CDK4* (12q14.1), and cyclin E1 *CCNE1* (19q12), which repression by Rb is necessary to prevent replication of senescent cells [50, 51]. Additional GBM-specific gains are of segments containing the oncogenes *AKT3* (1q44) [52] and Harvey Ras-encoding *HRAS* (11p15.5) [53]. We find, therefore, that the GBM-specific amplifications, of *AKT3*, *HRAS*, and genes involved in decreased Rb activity, together with the LGA-shared deletions of *CDKN2A* and *CDKN2B*, and CNAs involved in decreased activity of p53, enhance the opportunity for

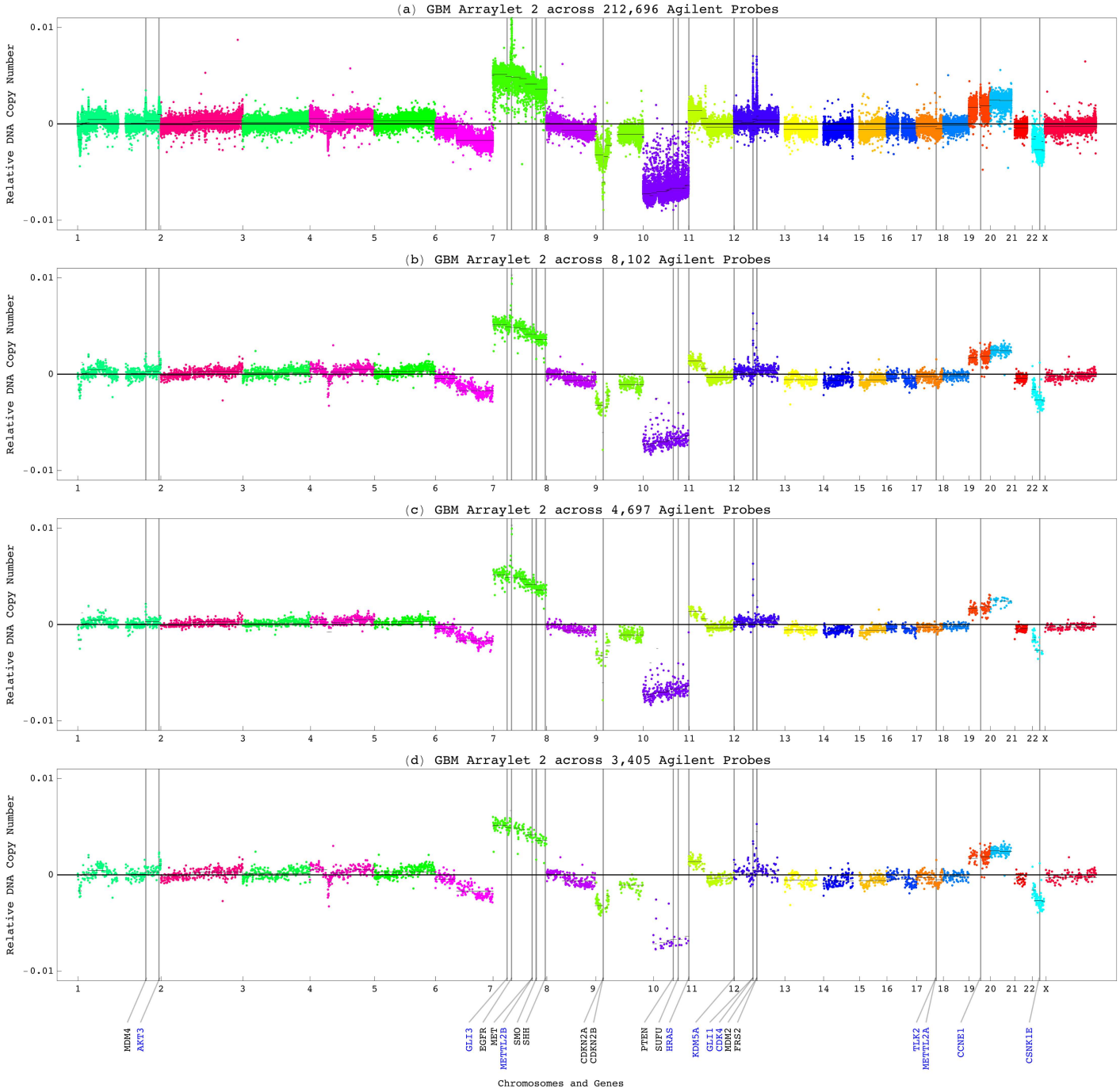

Fig 3. GBM genome-wide pattern of co-occurring CNAs previously uncovered by the GSVD of GBM tumor and normal profiles. (*a*) Plot of the second most GBM tumor-exclusive tumor arraylet, which was previously uncovered by the GSVD [8], describes a genome-wide pattern of co-occurring CNAs across 212,696 Agilent probes. The GBM pattern, which encompasses the LGA pattern (Fig 2), consists of LGA-shared (black) and GBM-specific (blue) CNAs, including, e.g., gains of segments on chromosome 1 containing *MDM4* and *AKT3*, respectively. (*b*) Both LGA-shared and GBM-specific CNAs are visible across the 8,102 Affymetrix-matched Agilent probes, even though these are <4% of the probes that constitute the GBM pattern. (*c*) The LGA-shared CNAs, e.g., in *MDM4*, are visible across the 4,697 Affymetrix-matched consistently-aberrated Agilent probes. (*d*) The GBM-specific CNAs, e.g., in *AKT3*, are visible across the 3,405 remaining probes.

doi:10.1371/journal.pone.0164546.g003

human normal to tumor cell transformation in response to growth factor signaling in GBM relative to LGA.

GBM-specific CNAs that contribute to increased conversion of the Gli oncogenes into transcriptional activators, include gains of segments containing the genes encoding for two of the three Gli proteins, *GLI3* (7p14.1) and *GLI1* (12q13.3), which was first identified in a screen of amplified DNA in a malignant human glioma tumor sample [54]. Also included is a loss of a segment containing the serine/threonine protein kinase-encoding *CSNK1E* (22q13.1). The encoded kinase CKIϵ is one of two members of the casein kinase I (CKI) protein family that in the absence of Hh facilitate the conversion of the Gli proteins into transcriptional repressors [55]. These GBM-specific CNAs that are involved in increased levels of the Gli transcriptional activators, together with the LGA-shared CNAs that are involved in decreased activity of Ptch1, enhance the opportunity for proliferation in response to developmental signals in GBM relative to LGA [56].

Gains of segments containing putative drug targets are also among the GBM-specific CNAs, including the methyltransferases-encoding *METTL2B* (7q32.1) and *METTL2A* (17q23.2), and the serine/threonine kinase-encoding *TLK2* (17q23.2) [8, 57].

To additionally compare the LGA and GBM patterns, we identified 8,102 pairs of one-to-one overlapping Affymetrix and Agilent probes among the 933,827 Affymetrix probes of the LGA pattern and the 212,696 Agilent probes of the GBM pattern. Among these, we identified 4,697 pairs of one-to-one overlapping probes that are consistently aberrated in the LGA and GBM patterns. The LGA-shared CNAs in chromosomes, chromosome arms, and segments are visible in both the LGA and GBM patterns, across the 8,102, and, separately, the 4,697 pairs of probes, even though these are $<1\%$ and 4% of the probes that constitute the LGA and GBM patterns, respectively.

The GSVD Separates the LGA Pattern from CNVs Common to the Normal Human and LGA Tumor Genomes and Tumor-Exclusive Experimental Batch Effects. This is because the second tumor arraylet, which describes the LGA pattern, is mathematically orthogonal to the other tumor arraylets, which describe other sources of biological and experimental variation that compose the tumor dataset.

For example, the first tumor arraylet, which is mathematically the most significant arraylet in the tumor dataset, describes mostly unsegmented chromosomes [58, 59], each with a copy-number distribution that is approximately centered at the autosomal genome with a relatively large, chromosome-invariant width (Fig E in S1 Appendix and S3 Dataset). The first probelet, which is mathematically the most tumor-exclusive probelet, is correlated with a tumor-exclusive experimental variation in the hybridization plate of the LGA tumor samples, with both hypergeometric [60] and Mann-Whitney-Wilcoxon *P*-values $<10^{-2}$ (Fig F in S1 Appendix). Together, the first probelet and tumor arraylet describe a tumor-exclusive experimental batch effect.

The 53rd normal arraylet, which is mathematically the most significant arraylet in the normal dataset, and the 53rd LGA tumor arraylet (Fig G in S1 Appendix), both describe a deletion of the X chromosome relative to the autosomal genome. Consistently, the 53rd probelet, which is mathematically approximately common to the tumor and normal datasets, classifies the patients by gender, with both hypergeometric and Mann-Whitney-Wilcoxon *P*-values $<10^{-9}$. Together, the 53rd probelet and arraylets describe a male-specific X chromosome deletion, a CNV across the normal genomes that is conserved in the patient-matched LGA tumor genomes.

Note that although the male-specific X chromosome deletion is conserved in the tumor genomes, the LGA pattern, which is described by the second tumor arraylet, exhibits an unsegmented X chromosome copy-number distribution that is approximately centered at the autosomal genome with a relatively small, invariant width. This illustrates the separation of the LGA tumor-exclusive pattern from the male-specific X chromosome deletion that is common to the tumor and normal profiles.

This GSVD separation of the LGA tumor and normal datasets into probelets, and tumor and normal arraylets, is blind, that is, without a-priori knowledge of the sources of variation that compose the datasets. The TCGA annotations that describe the patients (e.g., gender), and the corresponding tumor and normal samples (e.g., the hybridization plate of the tumor vs. the normal samples), are used only to interpret the patterns of variation across the patients, and the tumor and normal genomes, which were uncovered by the GSVD.

The LGA Pattern is Correlated with LGA Outcome. To examine the correlation of the LGA pattern with an LGA patient's survival, we classified the discovery set of patients based upon the weight of the pattern, that is, the superposition coefficient of the second LGA tumor arraylet, in each patient's tumor profile. These coefficients are linearly proportional to the relative copy numbers listed in the second LGA probelet. For the cutoff to be consistent with that previously established for the GBM pattern [8], we scaled the second GBM arraylet correlation cutoff of 0.15 by the Euclidean-, i.e., 2-norm of the Pearson correlations of the discovery tumor profiles with the second LGA tumor arraylet. The second probelet classifies the discovery set of patients into two groups of statistically significantly different prognoses (Fig 4). The univariate Cox [61] proportional hazard ratio is >9. This means that a high weight of the LGA pattern in an LGA tumor's profile confers >9 times the hazard of a low weight.

To examine the correlation of the pattern with response to treatment, we classified the discovery set of patients by the GSVD and, in addition, by chemotherapy or radiation. Among the patients who were treated by either chemotherapy or radiation, the Kaplan-Meier (KM) [62] median survival time of the groups of patients with low coefficients is $\sim$3.5 times, and $\sim$4 years greater than the median survival time of the groups of patients with high coefficients. A low weight of the LGA pattern in an LGA tumor's profile is, therefore, correlated with a significantly longer survival time, also in response to chemotherapy or radiation.

To computationally validate that the LGA pattern is correlated with LGA outcome, we classified the Affymetrix-measured primary LGA tumor profiles of a validation set of 74 TCGA patients, mutually exclusive of the discovery set (S4 Dataset). The classification is based upon the correlation of the second LGA tumor arraylet with each patient's tumor profile across the 933,827 Affymetrix probes. We find that the results of the survival analyses of the LGA validation set are consistent with those of the LGA discovery set. Note also that in classifying the tumor profiles, the 8,102 Agilent-matched Affymetrix probes and, separately, the 4,697 consistently-aberrated probes among them, give qualitatively the same and quantitatively similar results as the 933,827 Affymetrix probes.

The GBM Pattern Identifies among the LGA Patients a Subtype, Similar to that among the GBM Patients, where the CNA Genotype is Correlated with an Approximately One-Year Survival Phenotype

Because the GBM pattern encompasses the LGA pattern, we also examined the correlation of the GBM pattern with an LGA patient's survival. To start, we used the GBM pattern to classify the primary GBM tumor profiles of a set of 364 TCGA patients (S5 Dataset). We find that the GBM pattern is a platform-independent predictor of GBM survival. Classifying the GBM patients based upon the Affymetrix-measured tumor profiles, and across just the 4,697 probes (Fig 5), gives qualitatively the same and quantitatively similar results as the previous classification based upon the Agilent-measured profiles, across the 212,696 Agilent probes [8]. As in the previous classification, the KM median survival time of the group of patients with low correlations is >2.5 times, and >1.5 years greater than the approximately one-year median survival time of the group of patients with high correlations.

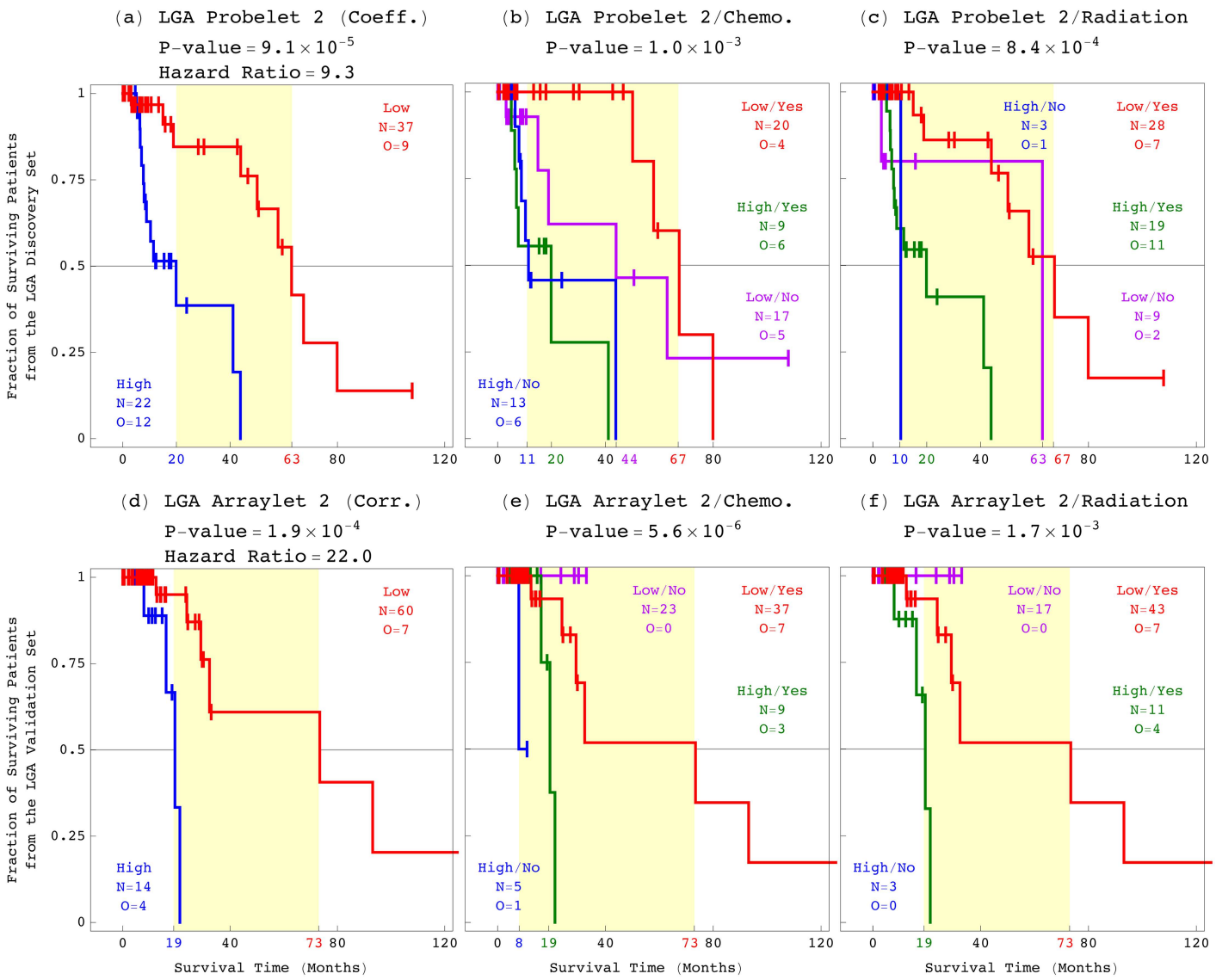

Fig 4. Survival analyses of the LGA patients classified by the LGA pattern and by treatment. (*a*) KM curves of the discovery set of 59 patients classified by the weights, or superposition coefficients of the LGA pattern in their tumor profiles, as listed in the second probelet (Fig 3). The 63-month KM median survival time of the group of patients with low coefficients is >3 times greater than that of the group of patients with high coefficients, with the corresponding log-rank test *P*-value $<10^{-4}$. The univariate Cox proportional hazard ratio is >9. (*b*) Among the 29 patients in the discovery set treated by chemotherapy, the median survival time of the patients with low coefficients is $\sim$3.5 times greater than that of the patients with high coefficients. (*c*) Among the patients treated by radiation, the median survival times of patients with low and high coefficients are the same as among the chemotherapy-treated patients. (*d*) KM curves of the validation set of 74 patients classified by the Pearson correlation of the LGA pattern with their tumor profiles. The 73-month median survival time of the patients with low correlations is >3.5 times greater than that of the patients with high correlations, consistent with the median survival times of the patients in the discovery set. (*e*) The median survival times of the 46 chemotherapy-treated validation patients with low and high correlations are the same as those of the 74 validation patients, and consistent with those of the 27 chemotherapy-treated discovery patients. (*f*) The median survival times of the radiation-treated validation patients are the same as those of the validation patients, and consistent with those of the radiation-treated discovery patients.

doi:10.1371/journal.pone.0164546.g004

Next, we used the GBM pattern to classify the Affymetrix-measured tumor profiles of the 133 TCGA patients in the LGA discovery and validation sets. The survival analysis results are consistent with those based upon the correlation with the Affymetrix-derived LGA pattern across the 933,827 Affymetrix probes.

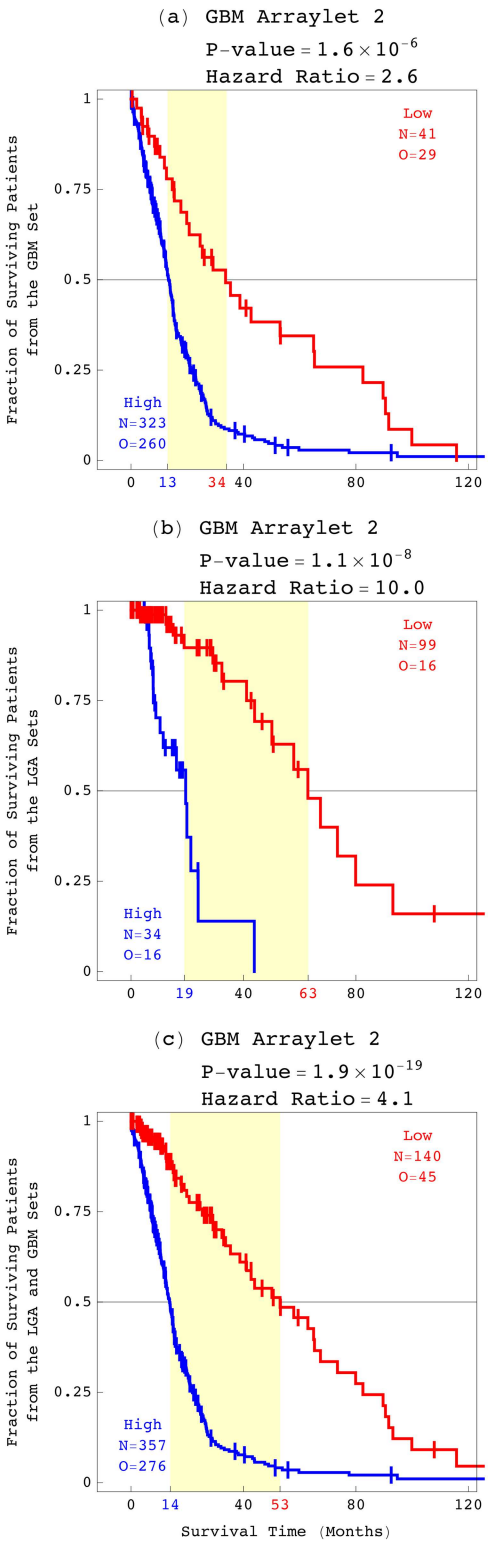

Fig 5. Survival analyses of the LGA and GBM patients classified by the GBM pattern. KM curves, log-rank test *P*-values, and Cox proportional hazard ratios of (*a*) the GBM set of 364 patients, (*b*) the LGA discovery and validation sets of 133 patients, and (*c*) the LGA and GBM sets of 497 patients.

doi:10.1371/journal.pone.0164546.g005

Because a high weight of the GBM pattern in either an LGA or a GBM tumor's profile confers a greater hazard and a shorter survival time, we compared the survival of the groups of LGA and GBM patients that are identified by the GBM pattern. We find that the KM curves for these two groups overlap, with the corresponding log-rank test P-value >0.05, which means that the two groups are statistically indistinguishable based upon survival.

Classifying the 133 LGA and 364 GBM, i.e., 497 astrocytoma patients, based upon the weight of the GBM pattern in each patient's tumor profile, we find that the GBM pattern is a predictor of survival among the general primary astrocytoma population, independent of grade, where the CNA genotype that the GBM pattern describes is correlated with an approximately one-year survival phenotype. We also assessed the distribution of several TCGA annotations of intratumor heterogeneity in each astrocytoma grade, including the tumor sample's volume, the slide's percents of tumor cells and nuclei, the portion's weight, and the analyte's and aliquot's native, unamplified DNA quantities. We find that at the TCGA ranges for these annotations, the GBM pattern is independent of intratumor heterogeneity.

The GBM Pattern is a Platform-Independent Predictor of Astrocytoma Outcome, Statistically Better Than, and Independent of Age, Grade, and Existing Laboratory Tests

To examine the correlation of the GBM pattern with an astrocytoma patient's response to treatment, we classified the 497 patients by chemotherapy or radiation and, in addition, by the GBM pattern (Fig 6). These classifications give bivariate Cox hazard ratios which are close to, and within the 95% confidence intervals of the corresponding univariate ratios (Table A in S1 Appendix). This means that the GBM pattern is a predictor of a patient's survival independent of treatment, and, therefore, also a predictor of the patient's response to treatment.

Next, we examined the correlation of the GBM pattern with a patient's age and a tumor's grade (Fig 7) [31–38]. We find that the log-rank test P-value, which corresponds to the classification by the GBM pattern, is less than the P-values which correspond to the classifications by age and grade. The univariate hazard ratio and the concordance index, which correspond to the GBM pattern, are greater than those that correspond to age and grade. These mean that the GBM pattern is statistically a better predictor of astrocytoma outcome than age or grade. Classifying the patients by the GBM pattern in addition to age or grade, we find that the GBM pattern is also statistically independent of age and grade.

Combined with either age or grade, therefore, the GBM pattern is statistically an even better predictor of astrocytoma outcome. For example, the >4-year survival difference among the patients classified by both the GBM pattern and age, is >3 times, and >2.5 years greater than the difference between the patients classified by age alone. The >3.5-year difference among the grades III and IV astrocytoma patients classified by the GBM pattern and grade, is >1.5 times, and 1.5 years greater than the difference between these patients classified by grade alone.

We also compared the GBM pattern to the existing pathology laboratory tests for astrocytoma. Silencing of a tumor's *MGMT* gene by hypermethylation of its DNA promoter region was associated with a GBM and, recently, also an LGA patient's longer survival in response to temozolomide chemotherapy treatment [63, 64]. Mutation of the gene *IDH1* was associated with a patient's longer survival [65], and linked with patterns of mRNA expression and DNA methylation across several hundred genes and genomic regions, respectively, in the tumor's genome [66–68].

We find that the genome-wide GBM pattern of CNAs is statistically a better predictor of astrocytoma outcome, corresponding to greater median survival time difference, proportional

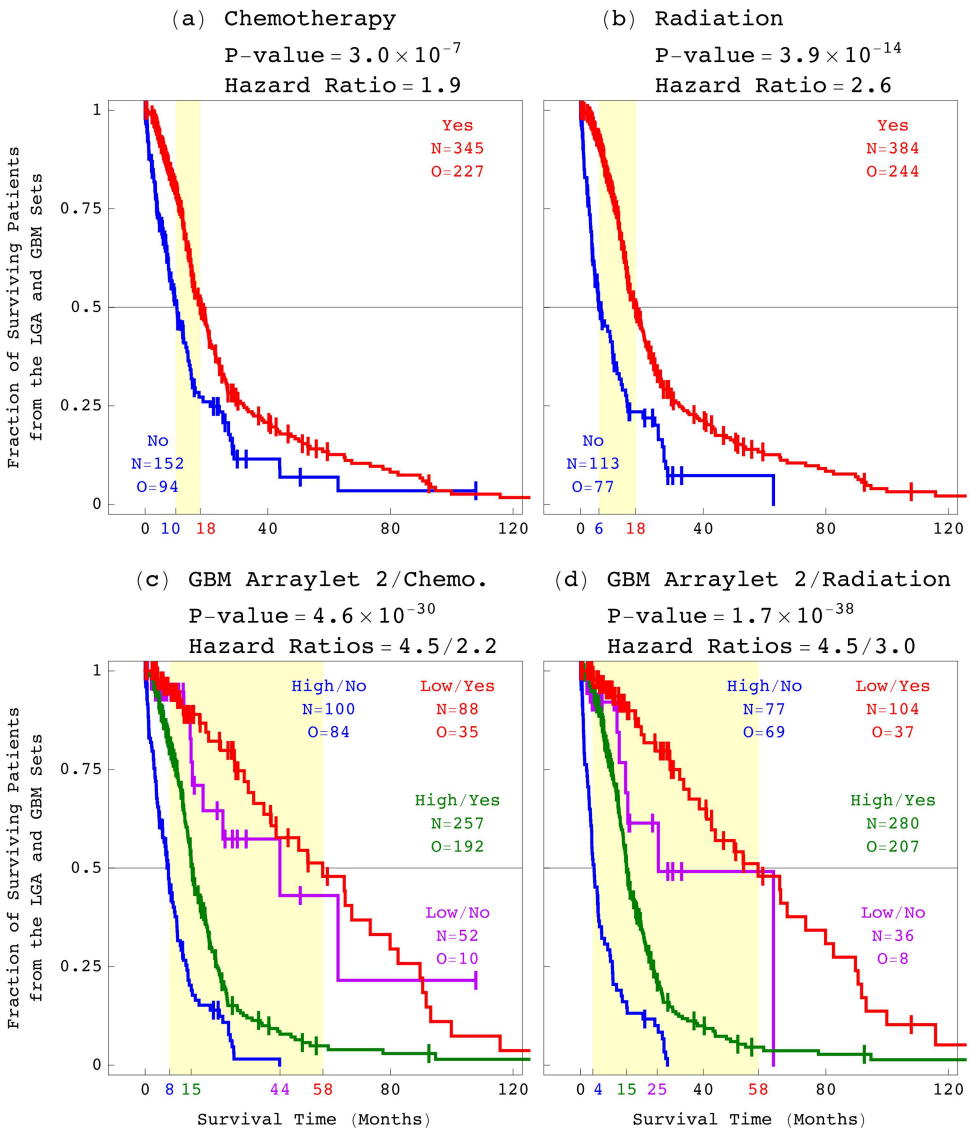

Fig 6. Survival analyses of the astrocytoma patients classified by treatment and by the GBM pattern. KM curves, log-rank test *P*-values, and Cox proportional hazard ratios of the 497 astrocytoma patients classified by (*a*) chemotherapy, (*b*) radiation, (*c*) the GBM pattern combined with chemotherapy, and (*d*) the GBM pattern combined with radiation.

doi:10.1371/journal.pone.0164546.g006

hazard ratio, and concordance index, than *MGMT* promoter methylation and *IDH1* mutation (Fig 8). The GBM pattern additionally classifies the patients with either a methylated or an unmethylated *MGMT* promoter, or a mutated or an unmutated *IDH1*, into two groups each, with an approximately one-year to >4-year survival differences, which means that it is independent of both. Combined with either existing pathology laboratory test, therefore, the GBM pattern is an even better predictor of astrocytoma.

Discussion

To date, statistically the best indicators of astrocytoma outcome in clinical use remain the patient's age at diagnosis and the tumor's grade [31–35, 38]. High-throughput molecular

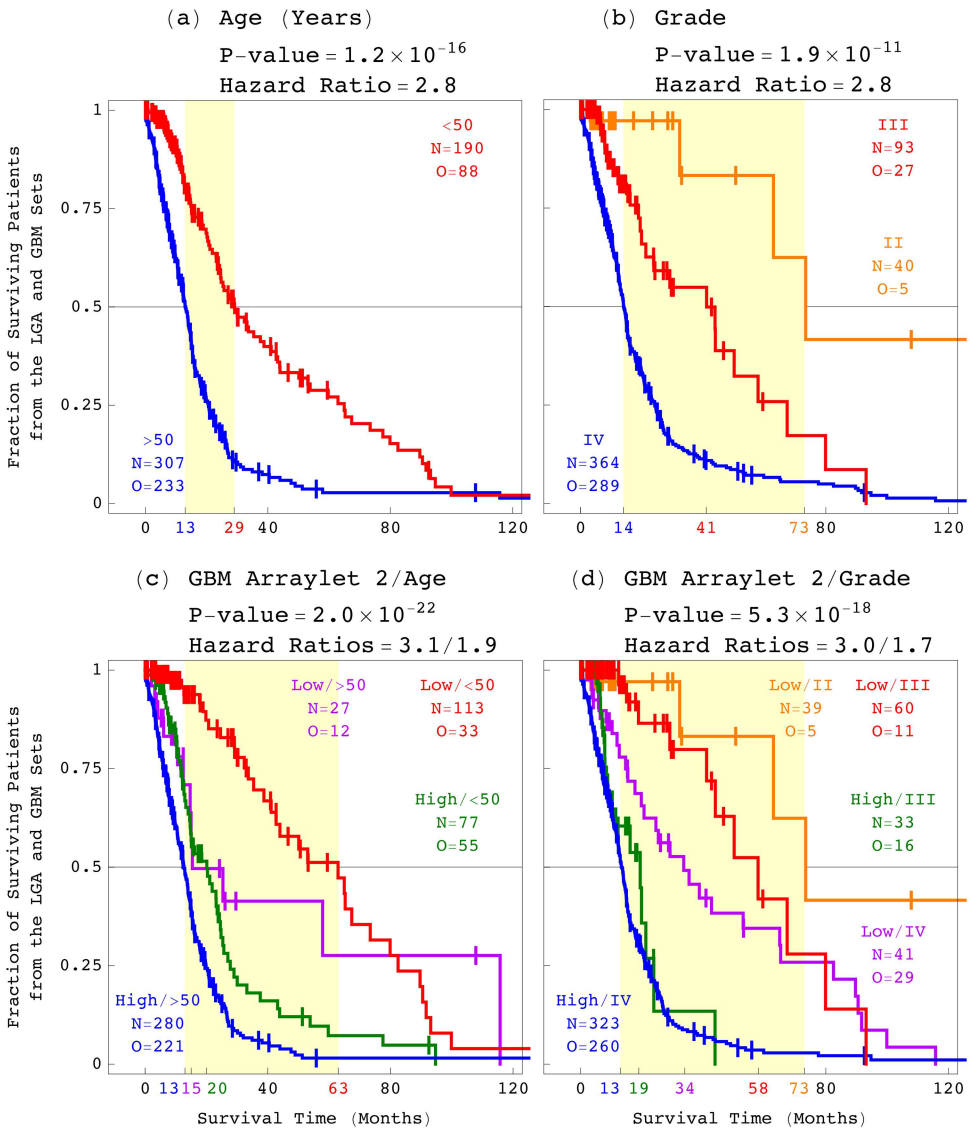

Fig 7. Survival analyses of the astrocytoma patients classified by the patient's age at diagnosis and the tumor's grade, and by the GBM pattern. The 497 astrocytoma patients classified by (*a*) the patient's age, (*b*) the tumor's grade, (*c*) the GBM pattern combined with age, and (*d*) the GBM pattern combined with grade.

doi:10.1371/journal.pone.0164546.g007

profiling efforts identified two indicative genetic loci that were translated into pathology laboratory tests, one locus of DNA hypermethylation, and the other of DNA mutation linked with mRNA expression and DNA methylation subtypes of astrocytoma [39, 63–68]. Recurring DNA CNAs have been observed in astrocytoma tumors' genomes for decades, however, copy-number subtypes that are predictive of astrocytoma patients' outcomes were not identified [36, 37].

Here, we showed that a genome-wide pattern of CNAs in a primary astrocytoma tumor's DNA copy-number profile is a predictor of the patient's survival and response to chemotherapy and radiation, statistically better than, and independent of the patient's age, the tumor's grade, and the existing laboratory tests. We showed that the pattern is correlated with an approximately one-year survival phenotype among the astrocytoma patients. The pattern is a platform-independent predictor, and, therefore, it can be translated into a laboratory test by

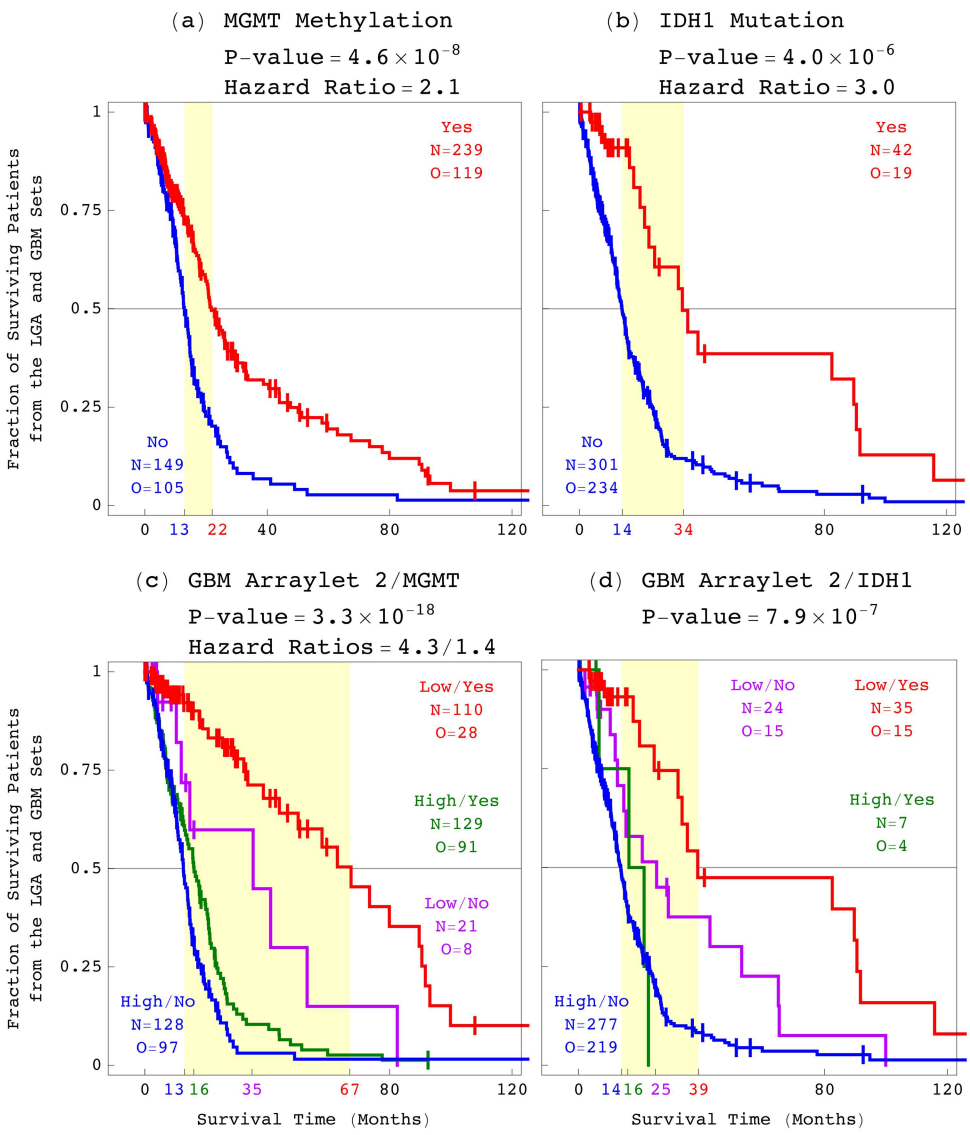

Fig 8. Survival analyses of the astrocytoma patients classified by the existing laboratory tests and by the GBM pattern. The 497 astrocytoma patients classified by (*a*) *MGMT* promoter methylation, (*b*) *IDH1* mutation, (*c*) the GBM pattern combined with *MGMT*, and (*d*) the GBM pattern combined with *IDH1*.

doi:10.1371/journal.pone.0164546.g008

using non-disease-specific FDA-approved platforms, such as next-generation sequencing (NGS) [69].

The genome-wide pattern of CNAs was previously uncovered by using the GSVD to model patient-matched copy-number profiles of GBM tumor and normal samples [8]. Here, a GSVD comparison of patient-matched profiles of LGA tumor and normal samples, revealed a tumor-exclusive genome-wide pattern of CNAs. We showed, and computationally validated, that this LGA pattern is correlated with an LGA patient's outcome. The GSVD separated this pattern from other sources of experimental and biological variation, common to the tumor and normal profiles, or exclusive to the tumor or the normal profiles, without a-priori knowledge of these variations. We also showed that the LGA pattern is encompassed in the GBM pattern, where GBM-specific CNAs encode for enhanced opportunities for transformation and proliferation

via growth and developmental signaling pathways in GBM relative to LGA. The LGA datasets had been publicly available in TCGA since 2015, and analyzed by using several methods. The pattern, however, remained unknown until the datasets were modeled by using the GSVD. This illustrates the ability of comparative spectral decompositions in general, and the GSVD in particular to find what other methods miss.

Note that in a GSVD comparison between two datasets, the only assumption is that the structure of the datasets is that of two full column-rank matrices of matched columns. It is, therefore, not limited to profiles of human cells, DNA copy-number profiles, or profiles measured by DNA microarray platforms, nor is it limited to molecular biological datasets. The GSVD was first formulated as a comparative spectral decomposition to model cell cycle phase-matched mRNA expression profiles of synchronized cells from human and yeast [17]. The model predicted a genome-wide causal coordination between DNA replication and mRNA expression [27, 28], which was then experimentally verified [70]. This demonstrated that the GSVD can be used to correctly predict previously unknown cellular mechanisms. Since then, the GSVD has been used to separate the similar from the dissimilar between different species, as well as between different types of molecular biological profiles, mostly large-scale (e.g., mRNA and protein expression in addition to DNA copy-number profiles), and different profiling technologies (e.g., NGS and quantitative real-time PCR in addition to DNA microarray platforms) [18–23] (see also [24–26]).

Methods

LGA Discovery Datasets Construction. We selected an LGA discovery set of 59 TCGA patients of consistent survival annotations. The 59 patients were diagnosed with World Health Organization (WHO) grades III or II astrocytoma. The patient-matched primary LGA tumor and normal tissue samples were obtained from US tissue source sites. Each tumor or normal profile lists median-centered $\log_2$ TCGA raw level 2 of the Affymetrix Genome-Wide Human SNP Array 6.0-measured DNA copy numbers. The profiles are organized in one tumor and one normal dataset, of M_1, M_2 = 933,827 autosomal and X chromosome nonpolymorphic copy-number probes, with valid data in all N = 59 patient-matched pairs of tumor and normal profiles, respectively.

CNAs in the LGA Pattern. To compare the Affymetrix-derived LGA pattern to the Agilent-derived GBM pattern, we mapped the 933,827 Affymetrix probes that constitute the LGA pattern onto the National Center for Biotechnology Information (NCBI) human genome sequence build 36 at the University of California at Santa Cruz (UCSC) human genome browser [58]. Previously, we also mapped the 212,696 probes of the Agilent Human Genome CGH 244A microarray platform that constitute the GBM pattern onto the same sequence. We then assigned to the LGA pattern CNAs in the chromosomes and chromosome arms, as well as the 111 of the 130 genomic segments that were previously identified in the GBM pattern by using the circular binary segmentation (CBS) [59], which are of $\geq$5 Agilent probes in length.

The LGA pattern was assigned a gain or a loss in a chromosome or a chromosome arm if the deviation of the mean copy number of the chromosome or the arm from the genomic mean is greater than twice the genomic standard deviation. The genomic mean and standard deviation are calculated for the autosomal genome, excluding the outlying chromosomes 7 and 10, and chromosome arm 9p [8]. A gain or a loss in a segment were assigned if the deviation of the segment mean copy number from the genomic mean is greater than twice the genomic standard deviation, or if the deviation from the chromosomal mean is greater than the chromosomal standard deviation, when this deviation is consistent with the deviation from the genomic mean.

Cross-Platform Probe Matching. We matched pairs of one Agilent and one Affymetrix probe that overlap by at least one nucleotide. When multiple Affymetrix or Agilent probes overlapped a single Agilent or Affymetrix probe, the Affymetrix or Agilent probe closest to the genomic end or start coordinate of the Agilent or Affymetrix probe was selected, respectively, to maintain a one-to-one matching between the platforms. This identified 8,102 pairs of one-to-one overlapping Affymetrix and Agilent probes.

To identify the 4,697 pairs of one-to-one overlapping probes that are consistently aberrated in the LGA and GBM patterns, we assigned to the patterns CNAs in the 8,102 Affymetrix and Agilent probes, respectively. A gain or a loss in a probe were assigned if the deviation of the probe copy number from the genomic mean is greater than twice the genomic standard deviation, or if the deviation from the chromosomal mean is greater than the chromosomal standard deviation, when this deviation is consistent with the deviation from the genomic mean.

Arraylet Visualization. To visualize the first tumor arraylet and 53rd normal and tumor arraylets, we segmented each arraylet by using the CBS [59].

Probelet Interpretation. To biologically or experimentally interpret the first and 53rd probelets, which are the most significant probelets in the tumor and normal datasets, respectively, we assessed the subsets of patients that are of high or low relative copy numbers in each probelet for enrichment in any one of the multiple TCGA annotations that describe the patients (e.g., gender), and the corresponding tumor and normal tissue samples (e.g., the hybridization plate of the tumor vs. the normal samples). The *P*-value of each enrichment was calculated assuming a hypergeometric probability distribution of the K annotations among the N patients of the discovery set, and of the subset of $k \subseteq K$ observed annotations among the subset of n patients that are of high or low copy numbers in each probelet [60], $P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{K}{i} \binom{N-K}{n-i}$.

In each probelet, we also assessed the distribution of the copy numbers among the different groups of each TCGA annotation by using boxplots, and calculating the corresponding Mann-Whitney-Wilcoxon *P*-values.

LGA Validation Dataset Construction. We selected an LGA validation set of 74 TCGA patients, which is mutually exclusive of the discovery set. Missing data among the 933,827 Affymetrix probes of the LGA pattern in any of the corresponding tumor profiles were not estimated. The corresponding probes were excluded from the calculations of this profile's median copy number as well as the profile's Pearson correlations with the LGA and GBM patterns.

GBM Dataset Construction. We selected a GBM set of 364 patients from the previous GBM discovery and validation sets [8]. For patients with more than one primary tumor profile, medians of the profiles were taken. Missing data among the 933,827 Affymetrix probes of the LGA pattern in any of the corresponding tumor profiles were not estimated. The corresponding probes were excluded from the calculations of this profile's median copy number as well as the profile's correlations with the GBM pattern.

***MGMT* Promoter Methylation and *IDH1* Mutation Annotations.** To estimate the *MGMT* promoter methylation status of a tumor, we used the TCGA raw level 1 of the Illumina Infinium Human Methylation 27 or 450 BeadChip-measured DNA methylation levels [64].

The *IDH1* mutation status of the LGA and GBM tumors is from TCGA [38, 68].

Supporting Information

S1 Appendix. Figs A–G and Table A. The Mathematica Notebook is available at http://www.alterlab.org/astrocytoma_prognosis/.
(PDF)

S1 Dataset. LGA Discovery Set of Patients. The corresponding Affymetrix-measured LGA tumor and normal profiles are at http://www.alterlab.org/astrocytoma_prognosis/.
(TXT)

S2 Dataset. GBM Segments. Segments previously identified by the CBS in the GBM pattern [8].
(TXT)

S3 Dataset. LGA Segments. Segments identified by the CBS in significant tumor and normal arraylets revealed by the GSVD of the LGA discovery datasets.
(TXT)

S4 Dataset. LGA Validation Set of Patients. The corresponding tumor profiles are at http://www.alterlab.org/astrocytoma_prognosis/.
(TXT)

S5 Dataset. GBM Set of Patients. The corresponding tumor profiles are at http://www.alterlab.org/astrocytoma_prognosis/.
(TXT)

Acknowledgments

We thank RA Horn for thoughtful discussions of matrix analysis, RL Jensen and CA Palmer for useful notes on astrocytoma intratumor heterogeneity and pathology, and MP Scott and RA Weinberg for helpful comments on the Hedgehog (Hh) and the rat sarcoma virus (Ras) signaling pathways. We also thank TE Schomay for technical assistance.

Author Contributions

Conceptualization: KAA OA.

Data curation: KAA OA.

Formal analysis: KAA OA.

Funding acquisition: OA.

Investigation: KAA OA.

Methodology: KAA OA.

Project administration: OA.

Resources: KAA OA.

Software: KAA OA.

Supervision: OA.

Validation: KAA OA.

Visualization: KAA OA.

Writing – original draft: KAA OA.

Writing – review & editing: KAA OA.

References

1. Boveri T. Concerning the origin of malignant tumours. Jena, Germany: Gustav Fischer Verlag; 1914. Translated and annotated by Harris, H. J Cell Sci. 2008; 121 (Suppl 1): 1–84. doi: 10.1242/jcs.025742 PMID: 18089652
2. Heim S. Boveri at 100: Boveri, chromosomes and cancer. J Pathol. 2014; 234 (2): 138–141. PMID: 25043504 doi: 10.1002/path.4406
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144 (5): 646–674. doi: 10.1016/j.cell.2011.02.013 PMID: 21376230
4. Sankaranarayanan P, Schomay TE, Aiello KA, Alter O. Tensor GSVD of patient-and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. PLoS One. 2015; 10 (4): e0121396. doi: 10.1371/journal.pone.0121396 PMID: 25875127
5. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474 (7353): 609–615. doi: 10.1038/nature10166 PMID: 21720365
6. Prisco MG, Zannoni GF, De Stefano I, Vellone VG, Tortorella L, Fagotti A, et al. Prognostic role of metastasis tumor antigen 1 in patients with ovarian cancer: a clinical study. Hum Pathol. 2012; 43 (2): 282–288. doi: 10.1016/j.humpath.2011.05.002 PMID: 21835429
7. Harries M, Gore M. Chemotherapy for epithelial ovarian cancer–treatment at first diagnosis. Lancet Oncol. 2002; 3 (9): 529–536. doi: 10.1016/S1470-2045(02)00846-X PMID: 12217790
8. Lee CH, Alpert BO, Sankaranarayanan P, Alter O. GSVD comparison of patient-matched normal and tumor aCGH profiles reveals global copy-number alterations predicting glioblastoma multiforme survival. PLoS One. 2012; 7 (1): e30098. doi: 10.1371/journal.pone.0030098 PMID: 22291905
9. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455 (7216): 1061–1068. doi: 10.1038/nature07385 PMID: 18772890
10. Golub GH, Van Loan CF. Matrix Computations. 4th ed. Baltimore, MD: Johns Hopkins University Press; 2012.
11. Horn RA, Johnson CR. Matrix Analysis. 2nd ed. Cambridge, UK: Cambridge University Press; 2012. doi: 10.1017/CBO9781139020411
12. Van Loan CF. Generalizing the singular value decomposition. SIAM J Numer Anal. 1976; 13 (1): 76–83. doi: 10.1137/0713009
13. Paige CC, Saunders MA. Towards a generalized singular value decomposition. SIAM J Numer Anal. 1981; 18 (3): 398–405. doi: 10.1137/0718026
14. Van Loan CF. Computing the CS and the generalized singular value decompositions. Numer Math. 1985; 46 (4): 479–491. doi: 10.1007/BF01389653
15. Bai Z, Demmel JW. Computing the generalized singular value decomposition. SIAM J Sci Comput. 1993; 14 (6): 1464–1486. doi: 10.1137/0914085
16. Friedland S. A new approach to generalized singular value decomposition. SIAM J Matrix Anal Appl. 2005; 27 (2): 434–444. doi: 10.1137/S0895479804439791
17. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci USA. 2003; 100 (6): 3351–3356. doi: 10.1073/pnas.0530258100 PMID: 12631705
18. Berger JA, Hautaniemi S, Mitra SK, Astola J. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. IEEE/ACM Trans Comput Biol Bioinform. 2006; 3 (1): 2–16. doi: 10.1109/TCBB.2006.10 PMID: 17048389
19. Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, et al. Conservation of the metabolomic response to starvation across two divergent microbes. Proc Natl Acad Sci USA. 2006; 103 (51): 19302–19307. doi: 10.1073/pnas.0609508103 PMID: 17159141
20. Schreiber AW, Shirley NJ, Burton RA, Fincher GB. Combining transcriptional datasets using the generalized singular value decomposition. BMC Bioinformatics. 2008; 9: 335. doi: 10.1186/1471-2105-9-335 PMID: 18687147
21. Sun Y, Li H, Liu Y, Mattson MP, Rao MS, Zhan M. Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation. PLoS One. 2008; 3 (10): e3406. doi: 10.1371/journal.pone.0003406 PMID: 18923680
22. Xiao X, Dawson N, MacIntyre L, Morris BJ, Pratt JA, Watson DG, et al. Exploring metabolic pathway disruption in the subchronic phencyclidine model of schizophrenia with the generalized singular value decomposition. BMC Syst Biol. 2011; 5: 72. doi: 10.1186/1752-0509-5-72 PMID: 21575198

23. Tomescu OA, Mattanovich D, Thallinger GG. Integrative omics analysis. A study based on *Plasmodium falciparum* mRNA and protein data. BMC Syst Biol. 2014; 8 (Suppl 2): S4. doi: 10.1186/1752-0509-8-S2-S4 PMID: 25033389

24. Ponnapalli SP, Golub GH, Alter O. A novel higher-order generalized singular value decomposition for comparative analysis of multiple genome-scale datasets. Workshop on Algorithms for Modern Massive Datasets (MMDS). Stanford, CA: Stanford University and Yahoo! Research; June 21–24, 2006.

25. Ponnapalli SP, Saunders MA, Van Loan CF, Alter O. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PLoS One. 2011; 6 (12): e28072. doi: 10.1371/journal.pone.0028072 PMID: 22216090

26. Xiao X, Moreno-Moral A, Rotival M, Bottolo L, Petretto E. Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. PLoS Genet. 2014; 10 (1): e1004006. doi: 10.1371/journal.pgen.1004006 PMID: 24391511

27. Alter O, Golub GH. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. Proc Natl Acad Sci USA. 2004; 101 (47): 16577–16582. doi: 10.1073/pnas.0406767101 PMID: 15545604

28. Alter O, Golub GH, Brown PO, Botstein D. Novel genome-scale correlation between DNA replication and RNA transcription during the cell cycle in yeast is predicted by data-driven models. Miami Nature Biotechnology Winter Symposium: Cell Cycle, Chromosomes and Cancer. Miami Beach, FL: University of Miami School of Medicine, vol. 15; January 31–February 4, 2004.

29. De Clercq W, Vergult A, Vanrumste B, Van Paesschen W, Van Huffel S. Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. IEEE Trans Biomed Eng. 2006; 53 (12): 2583–2587. doi: 10.1109/TBME.2006.879459 PMID: 17153216

30. Acar E, Bro R, Smilde AK. Data fusion in metabolomics using coupled matrix and tensor factorizations. Proc IEEE. 2015; 103 (9): 1602–1620. doi: 10.1109/JPROC.2015.2438719

31. Netsky MG, August B, Fowler W. The longevity of patients with glioblastoma multiforme. J Neurosurg. 1950; 7 (3): 261–269. doi: 10.3171/jns.1950.7.3.0261 PMID: 15415784

32. Curran WJ Jr, Scott CB, Horton J, Nelson JS, Weinstein AS, Fischbach AJ, et al. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. J Natl Cancer Inst. 1993; 85 (9): 704–710. doi: 10.1093/jnci/85.9.704 PMID: 8478956

33. Gorlia T, van den Bent MJ, Hegi ME, Mirimanoff RO, Weller M, Cairncross JG, et al. Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. Lancet Oncol. 2008; 9 (1): 29–38. doi: 10.1016/S1470-2045(07)70384-4 PMID: 18082451

34. Daumas-Duport C, Scheithauer B, O'Fallon J, Kelly P. Grading of astrocytomas. A simple and reproducible method. Cancer. 1988; 62 (10): 2152–2165. doi: 10.1002/1097-0142(19881115)62:10%3C2152::AID-CNCR2820621015%3E3.0.CO;2-T PMID: 3179928

35. Van Veelen MLC, Avezaat CJJ, Kros JM, van Putten W, Vecht CH. Supratentorial low grade astrocytoma: prognostic factors, dedifferentiation, and the issue of early versus late surgery. J Neurol Neurosurg Psychiatry. 1998; 64 (5): 581–587. doi: 10.1136/jnnp.64.5.581 PMID: 9598670

36. Wiltshire RN, Rasheed BKA, Friedman HS, Friedman AH, Bigner SH. Comparative genetic patterns of glioblastoma multiforme: potential diagnostic tool for tumor classification. Neuro Oncol. 2000; 2 (3): 164–173. doi: 10.1093/neuonc/2.3.164 PMID: 11302337

37. Misra A, Pellarin M, Nigro J, Smirnov I, Moore D, Lamborn KR, et al. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. Clin Cancer Res. 2005; 11 (8): 2907–2918. doi: 10.1158/1078-0432.CCR-04-0708 PMID: 15837741

38. Cancer Genome Atlas Research Network, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med. 2015; 372 (26): 2481–2498. doi: 10.1056/NEJMoa1402121 PMID: 26061751

39. Mischel PS, Cloughesy TF, Nelson SF. DNA-microarray analysis of brain cancer: molecular classification for therapy. Nat Rev Neurosci. 2004; 5 (10): 782–792. doi: 10.1038/nrn1518 PMID: 15378038

40. Hannon GJ, Beach D. $p15^{INK4B}$ is a potential effector of TGF-β-induced cell cycle arrest. Nature. 1994; 371 (6494): 257–261. doi: 10.1038/371257a0 PMID: 8078588

41. Karnoub AE, Weinberg RA. Ras oncogenes: split personalities. Nat Rev Mol Cell Biol. 2008; 9 (7): 517–531. doi: 10.1038/nrm2438 PMID: 18568040

42. Sherr CJ, McCormick F. The RB and p53 pathways in cancer. Cancer Cell. 2002; 2 (2): 103–112. doi: 10.1016/S1535-6108(02)00102-2 PMID: 12204530

43. Hahn WC, Counter CM, Lundberg AS, Beijersbergen RL, Brooks MW, Weinberg RA. Creation of human tumour cells with defined genetic elements. Nature. 1999; 400 (6743): 464–468. doi: 10.1038/22780 PMID: 10440377

44. Serrano M, Lin AW, McCurrach ME, Beach D, Lowe SW. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and $p16^{INK4A}$. Cell. 1997; 88 (5): 593–602. doi: 10.1016/S0092-8674(00)81902-9 PMID: 9054499

45. Fischer U, Keller A, Leidinger P, Deutscher S, Heisel S, Urbschat S, et al. A different view on DNA amplifications indicates frequent, highly complex, and stable amplicons on 12q13-21 in glioma. Mol Cancer Res. 2008; 6 (4): 576–584. doi: 10.1158/1541-7786.MCR-07-0283 PMID: 18403636

46. Rohatgi R, Scott MP. Patching the gaps in Hedgehog signalling. Nat Cell Biol. 2007; 9 (9): 1005–1009. doi: 10.1038/ncb435 PMID: 17762891

47. Wechsler-Reya R, Scott MP. The developmental biology of brain tumors. Annu Rev Neurosci. 2001; 24: 385–428. doi: 10.1146/annurev.neuro.24.1.385 PMID: 11283316

48. Kool M, Jones DT, Jäger N, Northcott PA, Pugh TJ, Hovestadt V, et al. Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell. 2014; 25 (3): 393–405. doi: 10.1016/j.ccr.2014.02.004 PMID: 24651015

49. Defeo-Jones D, Huang PS, Jones RE, Haskell KM, Vuocolo GA, Hanobik MG, et al. Cloning of cDNAs for cellular proteins that bind to the retinoblastoma gene product. Nature. 1991; 352 (6332): 251–254. doi: 10.1038/352251a0 PMID: 1857421

50. Chicas A, Wang X, Zhang C, McCurrach M, Zhao Z, Mert O, et al. Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. Cancer Cell. 2010; 17 (4): 376–387. doi: 10.1016/j.ccr.2010.01.023 PMID: 20385362

51. Etemadmoghadam D, George J, Cowin PA, Cullinane C, Kansara M, Australian Ovarian Cancer Study Group, et al. Amplicon-dependent *CCNE1* expression is critical for clonogenic survival after cisplatin treatment and is correlated with 20q11 gain in ovarian cancer. PLoS One. 2010; 5 (11): e15498. doi: 10.1371/journal.pone.0015498 PMID: 21103391

52. Turner KM, Sun Y, Ji P, Granberg KJ, Bernard B, Hu L, et al. Genomically amplified Akt3 activates DNA repair pathway and promotes glioma progression. Proc Natl Acad Sci USA. 2015; 112 (11): 3421–3426. doi: 10.1073/pnas.1414573112 PMID: 25737557

53. Reilly KM, Loisel DA, Bronson RT, McLaughlin ME, Jacks T. *Nf1*;*Trp53* mutant mice develop glioblastoma with evidence of strain-specific effects. Nat Genet. 2000; 26 (1): 109–113. doi: 10.1038/79075 PMID: 10973261

54. Kinzler KW, Bigner SH, Bigner DD, Trent JM, Law ML, O'Brien SJ, et al. Identification of an amplified, highly expressed gene in a human glioma. Science. 1987; 236 (4797): 70–73. doi: 10.1126/science.3563490 PMID: 3563490

55. Jia J, Zhang L, Zhang Q, Tong C, Wang B, Hou F, et al. Phosphorylation by double-time/CKIϵ and CKIα targets Cubitus interruptus for Slimb/β-TRCP-mediated proteolytic processing. Dev Cell. 2005; 9 (6): 819–830. doi: 10.1016/j.devcel.2005.10.006 PMID: 16326393

56. Regl G, Neill GW, Eichberger T, Kasper M, Ikram MS, Koller J, et al. Human GLI2 and GLI1 are part of a positive feedback mechanism in basal cell carcinoma. Oncogene. 2002; 21 (36): 5529–5539. doi: 10.1038/sj.onc.1205748 PMID: 12165851

57. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002; 1 (9): 727–730. doi: 10.1038/nrd892 PMID: 12209152

58. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002; 12 (6): 996–1006. doi: 10.1101/gr.229102 PMID: 12045153

59. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004; 5 (4): 557–572. doi: 10.1093/biostatistics/kxh008 PMID: 15475419

60. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. *GOrilla*: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009; 10: 48. doi: 10.1186/1471-2105-10-48 PMID: 19192299

61. Cox DR. Regression models and life-tables. J Roy Statist Soc B. 1972; 34 (2): 187–220.

62. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Amer Statist Assn. 1958; 53 (282): 457–481. doi: 10.1080/01621459.1958.10501452

63. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, et al. *MGMT* gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005; 352 (10): 997–1003. doi: 10.1056/NEJMoa043331 PMID: 15758010

64. Bady P, Sciuscio D, Diserens AC, Bloch J, van den Bent MJ, Marosi C, et al. *MGMT* methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions

associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. Acta Neuropathol. 2012; 124 (4): 547–560. doi: 10.1007/s00401-012-1016-2 PMID: 22810491

65. Purow B, Schiff D. Advances in the genetics of glioblastoma: are we reaching critical mass? Nat Rev Neurol. 2009; 5 (8): 419–426. doi: 10.1038/nrneurol.2009.96 PMID: 19597514

66. 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 (1): 98–110. doi: 10.1016/j.ccr.2009.12.020 PMID: 20129251

67. 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 (5): 510–522. doi: 10.1016/j.ccr.2010.03.017 PMID: 20399149

68. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013; 155 (2): 462–477. doi: 10.1016/j.cell.2013.09.034 PMID: 24120142

69. Collins FS, Hamburg MA. First FDA authorization for next-generation sequencer. N Engl J Med. 2013; 369 (25): 2369–2371. doi: 10.1056/NEJMp1314561 PMID: 24251383

70. Omberg L, Meyerson JR, Kobayashi K, Drury LS, Diffley JF, Alter O. Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. Mol Syst Biol. 2009; 5: 312. doi: 10.1038/msb.2009.70 PMID: 19888207