Cancer Cell
Article

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak,[1,2,17] Katherine A. Hoadley,[3,4,17] Elizabeth Purdom,[7] Victoria Wang,[8] Yuan Qi,[4,5] Matthew D. Wilkerson,[4,5] C. Ryan Miller,[4,6] Li Ding,[9] Todd Golub,[1,10] Jill P. Mesirov,[1] Gabriele Alexe,[1] Michael Lawrence,[1,2] Michael O'Kelly,[1,2] Pablo Tamayo,[1] Barbara A. Weir,[1,2] Stacey Gabriel,[1] Wendy Winckler,[1,2] Supriya Gupta,[1] Lakshmi Jakkula,[11] Heidi S. Feiler,[11] J. Graeme Hodgson,[12] C. David James,[12] Jann N. Sarkaria,[13] Cameron Brennan,[14] Ari Kahn,[15] Paul T. Spellman,[11] Richard K. Wilson,[9] Terence P. Speed,[7,16] Joe W. Gray,[11] Matthew Meyerson,[1,2] Gad Getz,[1] Charles M. Perou,[3,4,8] D. Neil Hayes,[4,5,*] and The Cancer Genome Atlas Research Network

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[3]Department of Genetics
[4]Lineberger Comprehensive Cancer Center
[5]Department of Internal Medicine, Division of Medical Oncology
[6]Department of Pathology and Laboratory Medicine
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7]Department of Statistics
[8]Group in Biostatistics
University of California, Berkeley, CA 94720, USA
[9]The Genome Center at Washington University, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA
[10]Department of Pediatric Oncology, Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[11]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[12]Department of Neurological Surgery, University of California, San Francisco, CA 94143, USA
[13]Department of Radiation Oncology, Mayo Clinic, Rochester, MN 55905, USA
[14]Department of Neurosurgery, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA
[15]SRA International, Fairfax, VA 22033, USA
[16]Walter and Eliza Hall Institute, Parkville, Victoria 3052, Australia
[17]These authors contributed equally to the work
*Correspondence: hayes@med.unc.edu
DOI 10.1016/j.ccr.2009.12.020

SUMMARY

The Cancer Genome Atlas Network recently cataloged recurrent genomic abnormalities in glioblastoma multiforme (GBM). We describe a robust gene expression-based molecular classification of GBM into Proneural, Neural, Classical, and Mesenchymal subtypes and integrate multidimensional genomic data to establish patterns of somatic mutations and DNA copy number. Aberrations and gene expression of *EGFR*, *NF1*, and *PDGFRA*/*IDH1* each define the Classical, Mesenchymal, and Proneural subtypes, respectively. Gene signatures of normal brain cell types show a strong relationship between subtypes and different neural lineages. Additionally, response to aggressive therapy differs by subtype, with the greatest benefit in the Classical subtype and no benefit in the Proneural subtype. We provide a framework that unifies transcriptomic and genomic dimensions for GBM molecular stratification with important implications for future studies.

SIGNIFICANCE

This work expands on previous glioblastoma classification studies by associating known subtypes with specific alterations in *NF1* and *PDGFRA*/*IDH1* and by identifying two additional subtypes, one of which is characterized by *EGFR* abnormalities and wild-type *p53*. In addition, the subtypes have specific differentiation characteristics that, combined with data from recent mouse studies, suggest a link to alternative cells of origin. Together, these data provide a framework for investigation of targeted therapies. Temozolomide and radiation, a common treatment for glioblastoma, has demonstrated a significant increase in survival. Our analysis illustrates that a survival advantage in heavily treated patients varies by subtype, with Classical or Mesenchymal subtypes having significantly delayed mortality that was not observed in the Proneural subtype.

INTRODUCTION

Glioblastoma multiforme (GBM) is the most common form of malignant brain cancer in adults. Patients with GBM have a uniformly poor prognosis, with a median survival of one year (Ohgaki and Kleihues, 2005); thus, advances on all scientific and clinical fronts are needed. In an attempt to better understand glioblastoma, many groups have turned to high-dimensional profiling studies. Several examples include studies examining copy number alterations (Beroukhim et al., 2007; Ruano et al., 2006) and gene expression profiling studies identifying gene signatures associated with *EGFR* overexpression, clinical features, and survival (Freije et al., 2004; Liang et al., 2005; Mischel et al., 2003; Murat et al., 2008; Nutt et al., 2003; Phillips et al., 2006; Shai et al., 2003; Tso et al., 2006).

The Cancer Genome Atlas (TCGA) Research Network has been established to generate the comprehensive catalog of genomic abnormalities driving tumorigenesis. TCGA provided a detailed view of the genomic changes in a large GBM cohort containing 206 patient samples. Sequence data of 91 patients and 601 genes were used to describe the mutational spectrum of GBM, confirming previously reported *TP53* and *RB1* mutations and identifying GBM-associated mutations in such genes as *PIK3R1*, *NF1*, and *ERBB2*. Projecting copy number and mutation data on the TP53, RB, and receptor tyrosine kinase pathways showed that the majority of GBM tumors harbor abnormalities in all of these pathways, suggesting that this is a core requirement for GBM pathogenesis.

Currently, only a few molecular factors show promise for prognosis or prediction of response to therapy (Curran et al., 1993; Kreth et al., 1999; Scott et al., 1998). An emerging prognostic factor is the methylation status of the *MGMT* promoter (Hegi et al., 2005). The TCGA GBM study (Cancer Genome Atlas Research Network, 2008) suggested that *MGMT* methylation shifts the GBM mutation spectrum in context of alkylating treatment, a finding with potential clinical implications. The inability to define different patient outcomes on the basis of histopathological features illustrates a larger problem in our understanding of the classification of GBM.

In the current study, we leverage the full scope of TCGA data to paint a coherent portrait of the molecular subclasses of GBM.

RESULTS

Consensus Clustering Identifies Four Subtypes of GBM

Factor analysis, a robust method to reduce dimensionality, was used to integrate data from 200 GBM and two normal brain samples assayed on three gene expression platforms (Affymetrix HuEx array, Affymetrix U133A array, and Agilent 244K array) into a single, unified data set. Using the unified data set, we filtered the data to 1740 genes with consistent but highly variable expression across the platforms. Consensus average linkage hierarchical clustering (Monti et al., 2003) of 202 samples and 1740 genes identified four robust clusters, with clustering stability increasing for k = 2 to k = 4, but not for k > 4 (Figures 1A and 1B). Cluster significance was evaluated using SigClust (Liu et al., 2008), and all class boundaries were statistically significant (Figure 1C). Samples most representative of the clusters, hereby called ''core samples'' (n = 173 of 202), were identified on the basis of their positive silhouette width (Rousseeuw, 1987), indicating higher similarity to their own class than to any other class member (Figure 1D). Genes correlated with each subtype were selected using SAM and ROC methods. ClaNC, a nearest centroid-based classifier that balances the number of genes per class, identified signature genes for all four subtypes (Dabney, 2006). An 840 gene signature (210 genes per class) was established from the smallest gene set with the lowest cross-validation (CV) and prediction error. Each of the signatures was highly distinctive (Figure 2A). Signatures and gene lists for all analyses are available at http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/.

These analyses were repeated on the three individual data sets, demonstrating that unifying the data improved CV error rates (see Figures S1A–S1E, available with this article online). Limiting the analysis to core samples reduced the CV error rate from 8.9% to 4.6%, validating their use as most representative of the cluster (Figures S1A and S1B). Importantly, our findings did not correlate with confounding factors well known to interfere with gene expression analysis, such as batch, sample purity, or sample quality (Table 1 and Figure S2). An exception was the sample collection center. However, the collection centers drew from different patient populations, and the relationship to subtype is largely the result of strong clinical differences in their patients, most notably age, as discussed below.

Validation of Subtypes in an Independent Data Set

An independent set of 260 GBM expression profiles was compiled from the public domain to assess subtype reproducibility (Beroukhim et al., 2007; Murat et al., 2008; Phillips et al., 2006; Sun et al., 2006). The subtype of TCGA samples was predicted using ClaNC, and data were visualized using the 840 classifying gene list (Figure 2A). Applying a similar ordering in the validation set clearly recapitulated the gene sample groups (Figure 2B). Importantly, the four subtypes were similarly proportioned in the validation and TCGA data set, as well as in all four individual validation data set cohorts (Figures S2G–S2L). Accounting for differences in sample size and analytic techniques, obvious concordance was seen between our classification and the results from earlier studies (Supplemental Experimental Procedures and Figure S3). To relate tumor subtype to a relevant model system, we obtained gene expression data from a collection of xenografts. The xenografts were established by direct implant of patient surgical specimens in athymic null/null mice (Hodgson et al., 2009). Proneural, Classical, and Mesenchymal subtypes were also reflected in the xenografts (Figure 2C). In contrast, attempts to detect comparable transcriptional subtypes in immortalized cell lines were uninformative (data not shown).

Functional Annotation of Subtypes

Subtype names were chosen on the basis of prior naming and the expression of signature genes: Proneural, Neural, Classical, and Mesenchymal. To get insight into the genomic events differentiating the subtypes, we used copy number data of 170 core samples that were recently described by the Cancer Genome Atlas Research Network (2008). Sequence data were available for 601 genes on 116 core samples; 73 samples were previously described. Fourteen amplifications and seven homozygous or