Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Associating spatial diversity features of radiologically defined tumor habitats with epidermal growth factor receptor driver status and 12-month survival in glioblastoma: methods and preliminary investigation

Joonsang Lee
Shivali Narang
Juan J. Martinez
Ganesh Rao
Arvind Rao

SPIE.

Associating spatial diversity features of radiologically defined tumor habitats with epidermal growth factor receptor driver status and 12-month survival in glioblastoma: methods and preliminary investigation

Joonsang Lee,[a] Shivali Narang,[a] Juan J. Martinez,[b] Ganesh Rao,[b] and Arvind Rao[a,*]
[a]University of Texas, MD Anderson Cancer Center, Department of Bioinformatics and Computational Biology, 1515 Holcombe Boulevard, Houston, Texas 77030, United States
[b]University of Texas, MD Anderson Cancer Center, Department of Neurosurgery, 1515 Holcombe Boulevard, Houston, Texas 77030, United States

Abstract. We analyzed the spatial diversity of tumor habitats, regions with distinctly different intensity characteristics of a tumor, using various measurements of habitat diversity within tumor regions. These features were then used for investigating the association with a 12-month survival status in glioblastoma (GBM) patients and for the identification of epidermal growth factor receptor (EGFR)-driven tumors. T1 postcontrast and T2 fluid attenuated inversion recovery images from 65 GBM patients were analyzed in this study. A total of 36 spatial diversity features were obtained based on pixel abundances within regions of interest. Performance in both the classification tasks was assessed using receiver operating characteristic (ROC) analysis. For association with 12-month overall survival, area under the ROC curve was 0.74 with confidence intervals [0.630 to 0.858]. The sensitivity and specificity at the optimal operating point (threshold = 0.5) on the ROC were 0.59 and 0.75, respectively. For the identification of EGFR-driven tumors, the area under the ROC curve (AUC) was 0.85 with confidence intervals [0.750 to 0.945]. The sensitivity and specificity at the optimal operating point (threshold = 0.166) on the ROC were 0.76 and 0.83, respectively. Our findings suggest that these spatial habitat diversity features are associated with these clinical characteristics and could be a useful prognostic tool for magnetic resonance imaging studies of patients with GBM. [DOI: 10.1117/1.JMI.2.4.041006]

Keywords: glioblastoma; radiomics; imaging-genomics; spatial diversity; tumor habitats.

Paper 15041SSRRR received Mar. 1, 2015; accepted for publication Jul. 28, 2015; published online Aug. 25, 2015.

1 Introduction

Glioblastoma (GBM) is the most common primary brain tumor known for its aggressive malignant behavior. Generally, the treatment of GBM involves surgical resection followed by a combination of radiation therapy and temozolomide. Despite multimodality treatment, the median survival time of GBM patients remains poor between 12 and 15 months.[1,2]

Medical image analysis plays an essential role for phenotyping disease and has tremendous applications in clinical decision support. Several computer-based medical image analyses have been studied in GBM.[3–6] Specifically, identifying the image-derived phenotype of the tumor is essential to understand and quantify treatment response and prognosis. For example, gray-level intensity heterogeneity within a tumor is indicative of multiple, potentially distinct subregions within the tumor and analysis of such heterogeneity has the potential to aid treatment of GBM.[7–9] Tumor texture has been investigated as one surrogate for tumor heterogeneity. It has been shown to be associated with the malignancy of the tumor[10] and can provide essential prognostic information.[11–13] Several texture analyses have been investigated in multiple tumor contexts based on imaging methods, such as computed tomography, positron emission tomography, and magnetic resonance imaging (MRI).[14–18]

In an orthogonal approach, researchers have investigated the spatial heterogeneity of tumors using the concept of radiologically defined "tumor habitats," where the tumor regions have distinctly different MRI-derived intensity characteristics.[8] Cancer is considered a disease that involves the clonal evolution of genes associated with cancer risk[19] and the spatial cellular heterogeneity of tumors is clearly evident in the imaging characteristics of tumors.

Many tumors show spatially heterogeneous patterns in contrast to enhancement in their medical images and such spatial patterns of the tumor represent various biological tissue properties based on water content, cellular density, fibrosis, and necrosis. In ecology, methods such as spatial species diversity or biodiversity analysis[20] have been proven to be useful in understanding the distribution and abundance of different species or types in a spatial region. One recent study has suggested that ecologic and evolutionary principles might provide a theoretical framework for linking diversity analysis from clinical imaging with regional variations in blood flow, cell density, and necrosis.[21] Drawing upon these studies, we investigated the

*Address all correspondence to: Arvind Rao, E-mail: aruppore@mdanderson.org

2329-4302/2015/$25.00 © 2015 SPIE

spatial heterogeneity characteristics of the tumor by using habitat diversity analysis, drawing from methodology in ecological statistics literature. The tumor region is treated as an ecological community, and the spatial diversity of multiple tumor habitats [defined radiologically based on T1 postcontrast and T2-fluid attenuated inversion recovery (FLAIR)[22] intensity] is assessed. T1 postcontrast images and T2 FLAIR MR images represent different tissue characteristics within the tumor. For example, characteristics such as perfusion and extravasation of the contrast agent could be determined by T1-weighted MRI sequences, and interstitial edema and cell density could be determined by T2 FLAIR MRI sequences.

In this study, we obtained various ecological diversity indices such as Shannon index, Simpson index, and Fisher's alpha to quantify the habitat diversity of the tumor. These measurements (features) were then used to investigate the association with 12-month overall survival (OS) status as well as driver status of the epidermal growth factor receptor (EGFR) gene (i.e., identifying those tumors driven by EGFR alteration), using a classification framework. Our goal was to determine if such radiographic features could be used as reliable surrogates for OS and tumor molecular (specifically, EGFR driven) status. Driver genes pertain to tumor initiating or maintaining molecular alterations and have been defined based on the combination of mutation events as well as DNA copy number changes (like amplifications or deletions). Several driver genes for GBM have been defined and include genes like phosphatase and tensin homolog (PTEN), EGFR, DNA-damage-inducible transcript, platelet-derived growth factor receptor (PDGFRA), and neurofibromin 1.[23] EGFR is the main regulator of cell function and tumorigenesis in GBM and is one of the most widely recognized driver alterations in this disease. EGFR signaling has been shown to modulate gliomagenesis and constitutes one of the key molecular events underlying the classical subtype.[24] EGFR alterations also provide the basis for therapeutic intervention based on receptor tyrosine kinase (RTK) inhibitor therapies.[25]

To the best of our knowledge, spatial diversity analysis of tumor habitats for quantitative imaging data is a novel investigation in neuroradiology as well as in radiomics. The goal of this study was to examine the feasibility of using spatial diversity features obtained from tumor habitats within MR images as reliable surrogates of 12-month OS status as well as tumor driver gene status in patients with GBM. This study was formulated based on reported observations about phenotypic consequences of molecular aberrations in GBM, specifically, changes in proliferation, invasion characteristics of the tumor, as well as associated patient survival. Apart from examining such feasibility, this work attempts to provide a characterization of the spatial variation in radiological habitat abundance across the tumor region, complementing previous studies of habitat abundance.[21,22] In addition, aside from providing a new set of features for radiomic characterization of GBMs, our study of associations with both survival and EGFR status pertains to various ongoing investigations in imaging-genomics analysis as well.[26,27]

2 Materials and Methods

2.1 *Data*

A dataset of 65 patients (21 females and 44 males) with primary, untreated GBM were studied based on the availability of postcontrast T1-weighted and T2-weighted FLAIR image from The Cancer Imaging Archive (TCIA).[28] These patients were selected such that their genomic (specifically EGFR mutation and copy number status), clinical, and companion imaging data were available from either The Cancer Genome Atlas (TCGA) portal or TCIA portal. Clinical data regarding tumor driver gene status and OS were obtained from the cBioPortal.[29] Patient demographics are summarized in Table 1.

Table 1 Patients demographics.

Characteristics	Group 1 (≤12 months)	Group 2 (>12 months)	Total
No. of men	17	27	44
No. of women	9	12	21
Age (years)	60.53 ± 15.9	55.05 ± 14.6	57.2 ± 15.6
Overall survival (months)	6.06 ± 2.96	24.41 ± 12.8	17.1 ± 13.5

The values in age and overall survival are mean ± standard deviation.

2.2 *Genomic and Survival Data Processing*

The first classification task pertains to the assessment of EGFR driver status in the tumor based on image-derived spatial diversity features from radiologically defined tumor habitats. Specifically, we aim to identify EGFR-driven GBMs from non-EGFR driven GBMs based on image-derived spatial diversity characteristics of tumor habitats. Tumor driver gene status was assessed using the combination of gene mutation and copy number change. Specifically, a gene is designated a "driver" if it has both a mutation as well as an amplification or deletion event in a tumor sample. Using prior studies of driver alterations in GBM,[23,30] we assessed the frequency of driver status for each of the 32 genes studied for GBM. We only focused on drivers with a frequency larger than 20% in the dataset to avoid minority-sampling biases during classifier training. Table 2 shows the top five frequencies of each driver gene in the dataset. Only one of the 32 genes met this threshold—EGFR. We designated each of the 65 tumor cases to be EGFR driven or not, based on whether or not EGFR was both mutated and altered (by copy number) in that patient's tumor. This binary designation is subsequently used as the class label in the classification task.

In this study, we also investigated if the image-derived spatial diversity features (specifically, diversity characteristics of radiologically defined tumor habitats) are associated with OS status at the 12-month time point. This choice of cutoff is based on the median survival times of GBM (12 to 15 months) and has been

Table 2 Top five driver events in the dataset (based on mutation and copy number change). Numbers represent the percentage frequency of those driver events in the dataset.

Drivers	EGFR	PDGFRA	DDIT3	PTEN	KIT
Frequencies (%)	36.9	10.8	9.2	9.2	7.7

EGFR, epidermal growth factor receptor; PDGFRA, platelet-derived growth factor receptor; DDIT3, DNA-damage-inducible transcript 3.

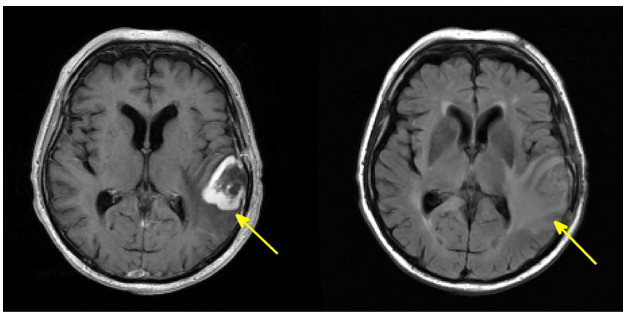


Fig. 1 Example (a) T1 postcontrast image and (b) T2 FLAIR image after preprocessing. An arrow points to the enhanced tumor area in each sequence.

used in other studies as well.[31] For this analysis, the patients were assigned to one of two groups according to their OS at 12 months.[31] One group had 26 patients with OS of 12 months or less, and the other group had 39 patients with OS greater than 12 months.

2.3 *Image Preprocessing*

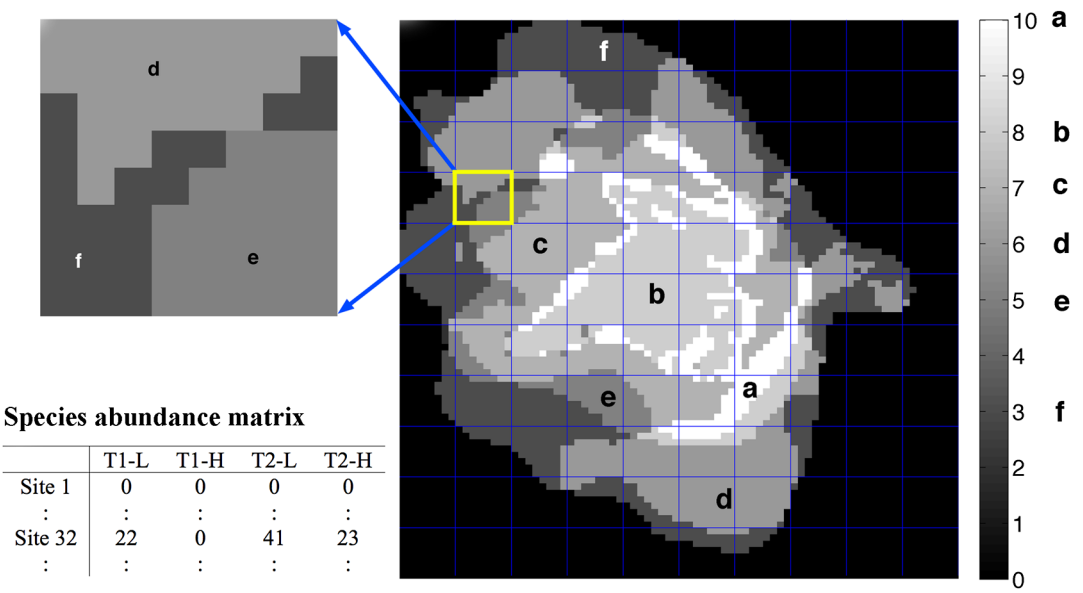
Image registration, nonuniformity correction, reslicing, and intensity normalization were performed as preprocessing procedures before analyzing the data for spatial diversity features. The registration of the T1 postcontrast images and T2 FLAIR images as well as nonuniformity correction for the artifacts in MRI were performed using medical image processing, analysis, and visualization software.[32] Images were subsequently resliced for isotropic pixel resolution using the NIFTI toolbox in MATLAB. Example T1 postcontrast and T2 FLAIR images are shown in Fig. 1.

2.4 *Delineation of Tumor Habitats and Their Spatial Point Patterns in the Tumor Region of Interest*

The segmentation of the tumor region was performed by experts (J.M. and G.R.) semiautomatically using the Medical Image Interaction Toolkit.[33] The slice with the largest tumor area in T1 postcontrast image and the corresponding slice in the T2 FLAIR image was selected for analysis. Each pixel in the tumor region from the T1 postcontrast and T2 FLAIR images was assigned to one of two groups according to its intensity, respectively. The threshold between the two intensity groups was determined based on a Gaussian mixture model[34] to assign each tumor pixel to a low-intensity or high-intensity group. These two (T1 postcontrast and T2 FLAIR) regions are combined into a region of interest (ROI) for habitat analysis, as the union of T1 postcontrast and T2 FLAIR tumor regions. The tumor ROI is treated as an ecological community.[20,21] For spatial diversity analysis, the tumor ROI is divided into 8×8 pixel square regions, called "quadrats." Each pixel in each quadrat is designated a "type" (or species) based on the intensity group it belongs to (T1-low, T1-high, FLAIR-low, and FLAIR-high). This creates a spatial point pattern across all the quadrats in the tumor region. Figure 2 illustrates this paradigm.

2.5 *Spatial Diversity Features*

Using the spatial point pattern obtained above, we obtained a range of diversity features over the tumor habitats,[22] based on their relative abundance in the tumor region.[35] First, the number of pixels in each quadrat was counted for each "type" (low or high intensity in T1 and FLAIR images), which gave us the abundance of each point type (or species) within the given quadrat. Subsequently, a species-abundance matrix was obtained. Each row represents a quadrat, and each column represents the abundance of each of the four species (T1-low, T1-high,



Species abundance matrix

	T1-L	T1-H	T2-L	T2-H
Site 1	0	0	0	0
:	:	:	:	:
Site 32	22	0	41	23
:	:	:	:	:

Fig. 2 An example of region of interest (ROI) spatial habitat map combining the low- and high-intensity in T1 postcontrast and T2 FLAIR ROIs (left of the figure). Two-dimensional grid lines were overlaid on each binary mask and were equally spaced at with the distance of 8 pixels. Each square site in the partitioned tumor region is called a "quadrat." The different gray level intensity areas represent radiologically defined combinations of different types. Top-left of the figure shows an example quadrat (32nd site) of this ROI and the species abundance matrix can be constructed from these quadrats by enumerating the pixels belonging to each of the four intensity groups (bottom-left of the figure).

Table 3 36 spatial diversity features.

No.	Diversity index	No.	Diversity index	No.	Diversity index
1	Mean Shannon	13	Mean Fisher alpha	25	Jaccard (nestedness)
2	Std-dev Shannon	14	Std-dev Fisher alpha	26	Jaccard
3	Skewness Shannon	15	Skewness Fisher alpha	27	Kendall index (T1-low)
4	Kurtosis Shannon	16	Kurtosis Fisher alpha	28	Kendall index (T1-high)
5	Mean Simpson	17	Mean Pielou's evenness	29	Kendall index (T2-low)
6	Std-dev Simpson	18	Std Pielou's evenness	30	Kendall index (T2-high)
7	Skewness Simpson	19	Skewness Pielou's evenness	31	α-diversity
8	Kurtosis Simpson	20	Kurtosis Pielou's evenness	32	β-diversity
9	Mean inv. Simpson	21	Sorensen (turnover)	33	γ-diversity
10	Std-dev inv. Simpson	22	Sorensen (nestedness)	34	No. of types (α-div.)
11	Skewness inv. Simpson	23	Sorensen	35	No. of types (β-div.)
12	Kurtosis inv. Simpson	24	Jaccard (turnover)	36	No. of types (γ-div.)

FLAIR-low, FLAIR-high intensity groups) in that quadrat. Next, the various diversity features were calculated from this species-abundance matrix. In this study, 36 diversity features were calculated (across all the quadrats in the tumor ROI) using the R package (vegan),[36] all of which are listed in Table 3.

Shannon, Simpson, inverse Simpson, Fisher indices, and Pielou's evenness are popular diversity indices representing quantitative measures that reflect the abundance of different point types in a spatial region. The definitions of these indices are explained in the Appendix. In addition to the aforementioned indices, we used functions from the "vegan" R-package for nestedness indices, Kendall indices (Kendall coefficient of concordance), and alpha, beta, as well as gamma diversity.[36] Nestedness indices find multiarea dissimilarities and decomposes these into components of turnover and nestedness,[37] and the Kendall index performs a posteriori tests of the contributions of individual types to the concordance of their group.[36] Alpha, beta, and gamma diversity were introduced by Whittaker[38,39] to represent the species richness of an area or the number of species in a habitat, differentiation among sites, and the richness of species present within a large area, respectively.

2.6 *Statistical Analysis*

A total of 36 diversity features that consist of the mean, standard deviation, skewness, and kurtosis (computed across all the quadrats in the tumor region) of the diversity indices such as the Shannon index, Simpson diversity index, inverse Simpson index, Fisher's alpha, Pielou's evenness index, nestedness and Kendall indices, and spatial measure of richness (alpha, beta, and gamma diversity) were computed from the measurement of abundance from the quadrats of ROIs. For examining association with 12-month survival, we used five diversity features: Kendall index (T1-high), Kendall index (T1-low), mean Fisher's alpha, skewness of the inverse Simpson, and standard deviation of Fisher's alpha. These five features were selected based on the overall coefficient of variation (CoV) across the dataset. These features were used to discriminate OS at the 12-month time point (>12 months or ≤12 months). For classifier modeling, we used a symbolic regression method,[40] with threefold cross validation for assessment of classifier performance. Difference of the classifier's performance [using area under the receiver operating characteristic (ROC) curve, AUC][41] relative to random classification (AUC = 0.5) is assessed via p-value from a Mann–Whitney hypothesis test (using R-package, "verification").[42] We used the Brier score to measure the accuracy of prediction using Eq. (1). The Brier score is a commonly used performance measure for assessing the accuracy of probability predictions, defined as

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{1}$$

where N is the sample size, f_i is the probability that was forecast, and o_i is the actual outcome of the event at instant i. This score ranges from 0 (for a perfect prediction) to 1 (for a prediction that is incorrect on every case). The predictive accuracy of the diversity features, true positive rate (TPR), and true negative rate (TNR) for the survival groups were assessed based on an operating point selected along the ROC to maximize the sum of sensitivity and specificity. The accuracy was calculated using Eq. (2) for the optimal model:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{2}$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative rates, respectively. The same procedure was followed to obtain a classifier to discriminate EGFR-driven tumors from tumors that were not EGFR driven using the top five features based on CoV: Kendall index of the T1-high species, Kendall index of the T1-low species, mean Fisher's alpha, skewness of the inverse Simpson, and standard

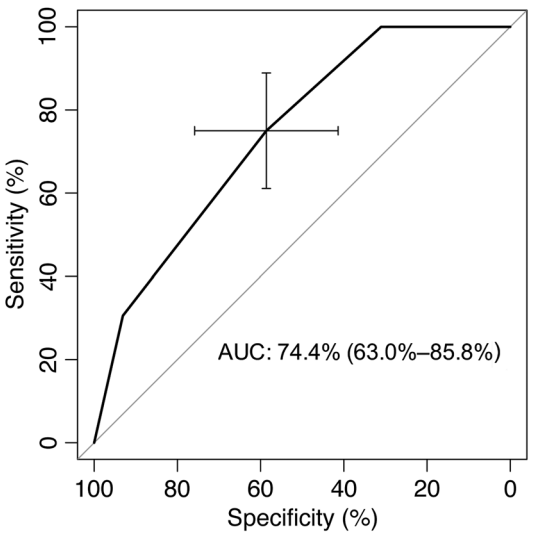


Fig. 3 Receiver operating characteristic (ROC) curve with confidence intervals for association with 12-month survival status (i.e., patient survival at the 12-month time point). The *x*-axis is the true negative rate (TNR) or specificity; the *y*-axis is the true positive rate (TPR) or sensitivity. The area under the ROC curve (AUC) is 0.74 with confidence intervals [0.630 to 0.858]. The vertical and horizontal bars at the optimal operating point (threshold = 0.5, specificity = 0.59, sensitivity = 0.75) indicate confidence intervals on sensitivity and specificity, respectively.

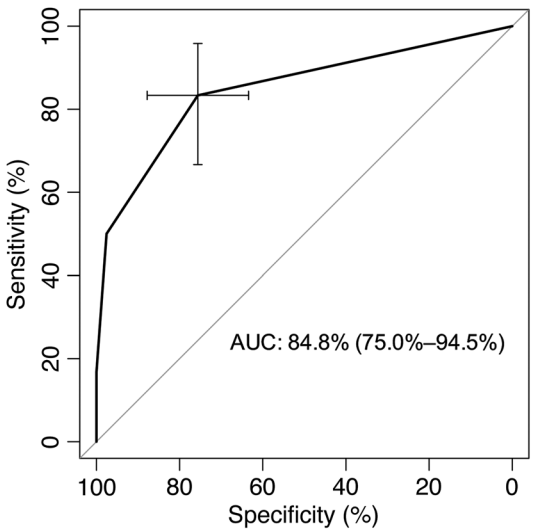


Fig. 4 ROC curve for the identification of epidermal growth factor receptor (EGFR) driver status (i.e., discriminating EGFR-driven tumors from those not driven by EGFR). The *x*-axis is the TNR or specificity; the *y*-axis is the TPR or sensitivity. The area under the ROC curve for EGFR status prediction is 0.848 with confidence intervals [0.750 to 0.945]. The vertical and horizontal bars at an optimal operating point (threshold = 0.166, specificity = 0.76, sensitivity = 0.83) indicate confidence intervals on sensitivity and specificity, respectively.

Table 4 The results for association with 12-month OS (TPR, TNR, and ACC are determined at an operating point that maximizes the sum of sensitivity and specificity).

AUC	TPR	TNR	ACC	p-value	Brier score
0.74	0.59	0.75	0.67	0.00021	0.197

Note: AUC, area under the ROC curve; TPR, true positive rate; TNR, true negative rate; ACC, accuracy.

Table 5 The results for discrimination of EGFR-driven tumors (TPR, TNR, and ACC are measured at an operating point chosen to maximize the sum of sensitivity and specificity along the ROC).

AUC	TPR	TNR	ACC	*p*-value	Brier score
0.85	0.76	0.83	0.79	1.56×10^{-7}	0.147

Note: AUC, area under the ROC curve; TPR, true positive rate; TNR, true negative rate; ACC, accuracy.

deviation of Fisher's alpha, across all quadrats within the tumor.

3 Results

We dichotomized the OS at 12 months,[1,2] yielding a binary label on the cases. This was used to build the classifier using symbolic regression. We computed the *p*-value and Brier score to assess the classifier's prediction of 12-month survival. The AUC was 0.74 and the corresponding *p*-value is 0.00021 indicating that the AUC for association with survival status at the 12-month point (>12 months or ≤12 months) is significantly different from random classification (AUC = 0.5). The Brier score (measures the accuracy of probabilistic predictions) was 0.197 for the survival prediction task. Figure 3 shows the ROC curve with AUC and confidence intervals to illustrate the performance of this binary classifier. At the chosen operating point along the ROC (chosen to maximize the sum of sensitivity and specificity), the sensitivity = 0.75 and specificity = 0.59. The results of the ROC analysis with confidence intervals are shown in Fig. 3 and Table 4.

In the ROC analysis for EGFR-driven tumor identification, the AUC is 0.845 and the corresponding *p*-value is 1.56×10^{-7} indicating that this AUC is also significantly different from random classification (AUC = 0.5). At an operating point (sensitivity = 0.83, 1 − specificity = 0.24) determined by maximizing the sum of sensitivity and specificity, the TPR and TNR are thus 0.77 and 0.83, respectively. The corresponding accuracy is 0.79. The Brier score, for the task of predicting driver gene status of EGFR, was 0.147, again suggesting good classifier performance. Figure 4 shows the ROC curve with confidence intervals for identification of EGFR-driven tumors, and the results are summarized in Table 5.

4 Discussion

In this work, we present a methodology to determine the ability of spatial habitat diversity features from radiologically defined tumor habitats to investigate the association with the 12-month survival of patients with GBM as well as the driver status of the EGFR gene. The case IDs for the 65 glioblastoma patients are listed in Table 6. In this study, we defined four distinct groups based on tumor intensity obtained from different MR sequences. These groups are considered as different species within an ROI; subsequently, we performed spatial diversity analysis using various measures of species distribution within the tumor. Our findings show that diversity features obtained from MR images are associated with 12-month OS and EGFR driver status in GBM

Table 6 Listing of the case IDs for the 65 glioblastoma patients.

#	Case ID	#	Case ID	#	Case ID	#	Case ID	#	Case ID
1	TCGA-02-0011	14	TCGA-02-0086	27	TCGA-06-0158	40	TCGA-06-0187	53	TCGA-08-0385
2	TCGA-02-0027	15	TCGA-02-0087	28	TCGA-06-0162	41	TCGA-06-0189	54	TCGA-08-0390
3	TCGA-02-0033	16	TCGA-02-0102	29	TCGA-06-0164	42	TCGA-06-0190	55	TCGA-08-0392
4	TCGA-02-0034	17	TCGA-02-0106	30	TCGA-06-0166	43	TCGA-06-0210	56	TCGA-08-0509
5	TCGA-02-0046	18	TCGA-06-0122	31	TCGA-06-0168	44	TCGA-06-0237	57	TCGA-08-0510
6	TCGA-02-0047	19	TCGA-06-0127	32	TCGA-06-0171	45	TCGA-06-0238	58	TCGA-08-0511
7	TCGA-02-0060	20	TCGA-06-0129	33	TCGA-06-0173	46	TCGA-06-0241	59	TCGA-08-0512
8	TCGA-02-0064	21	TCGA-06-0133	34	TCGA-06-0174	47	TCGA-06-0644	60	TCGA-08-0518
9	TCGA-02-0068	22	TCGA-06-0137	35	TCGA-06-0175	48	TCGA-08-0350	61	TCGA-08-0520
10	TCGA-02-0069	23	TCGA-06-0145	36	TCGA-06-0176	49	TCGA-08-0353	62	TCGA-08-0521
11	TCGA-02-0070	24	TCGA-06-0147	37	TCGA-06-0177	50	TCGA-08-0357	63	TCGA-08-0522
12	TCGA-02-0075	25	TCGA-06-0149	38	TCGA-06-0179	51	TCGA-08-0358	64	TCGA-08-0524
13	TCGA-02-0085	26	TCGA-06-0154	39	TCGA-06-0185	52	TCGA-08-0360	65	TCGA-08-0529

patients. In this study, we used the top five features based on CoV, which are Kendall index (from the T1-high group), Kendall index (T1-low group), mean Fisher's alpha, standard deviation of Fisher's alpha, and skewness of the inverse Simpson, across the tumor's quadrats. These spatial diversity features are based on the following diversity indices: the Kendall coefficient of concordance is a nonparametric statistic and is a measure of agreement or association among species. Fisher's alpha is a parametric diversity index and assumes that species abundance follows logarithmic distribution, which can predict the number of types at different levels of individual points. Inverse Simpson diversity index is a reciprocal Simpson's index and Simpson's index measures the probability that two individuals (points) randomly selected from a sample will belong to the same type. In our case, the Kendall index evaluates the degree of similarity of individual species to the overall concordance of their groups. The mean and standard deviation of Fisher's alpha indices indicate that the average and the

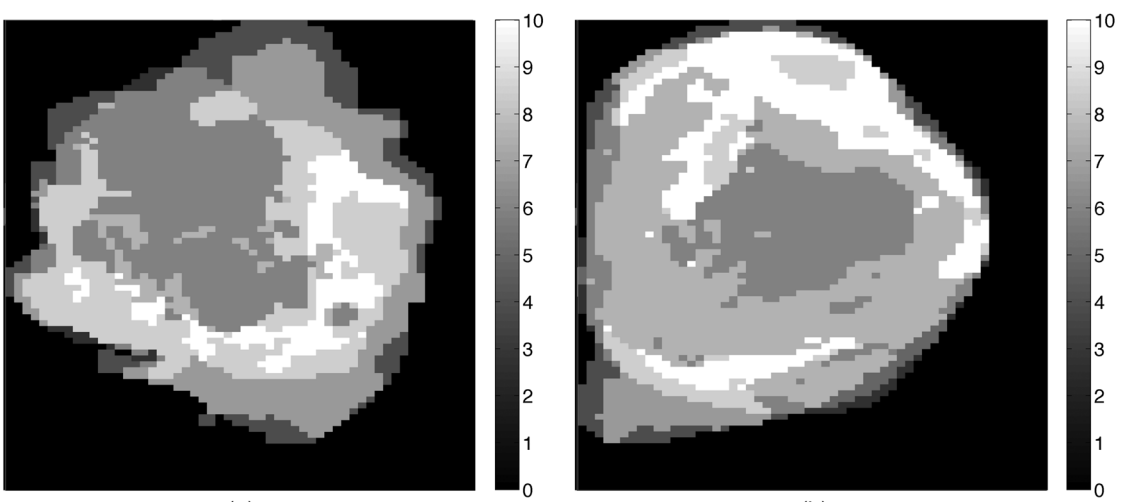


Fig. 5 Examples of ROI spatial habitat map combining the low- and high-intensity groups in T1 post-contrast and T2 FLAIR ROIs in (a) a low survival patient (4.8 months) and (b) a high survival patient (57.8 months). The values of the five spatial diversity features such as Kendall index of T1-high, Kendall index of the T1-low, mean Fisher's alpha, skewness of the inverse Simpson, and standard deviation of Fisher's alpha are 0.004, 0.004, 3.2×10^7, 0.48, and 2.3×10^8 in the low survival patient, and 0.14, 0.75, 5.3×10^7, -0.075, and 4.0×10^8 in the high survival patient, respectively. Also, patient (a) represents a patient with EGFR-driven glioblastoma, whereas patient (b) is not EGFR-driven.

amount of variation of Fisher's alpha across all the quadrats in the tumor region. Skewness of the inverse Simpson indicates a measure of the asymmetry of the distribution of the inverse Simpson indices across all the quadrats in the tumor region.

In the ROC analysis, we determined an operating point based on maximizing the sum of sensitivity and specificity. For association with 12-month survival, the TPR (0.59), TNR (0.75), accuracy (0.67) at this optimal point, and AUC of 0.74 were relatively high, indicating that species diversity features can discriminate the survival classes of GBM. In addition to assessing the predictive ability for discriminating survival status of patients with GBM, we also assessed performance in predicting the driver status of the EGFR gene in these tumors. Mutation in the EGFR gene has been associated with a number of cancers including GBM. Our findings indicate that EGFR-driven GBMs can be classified with high AUC of 0.85, as well as high TPR (0.76), high TNR (0.83), and high accuracy of 0.79 (at the chosen operating point along the ROC) based on the tumor-derived spatial diversity features. This suggests a potential relationship between habitat diversity and driver gene status of the EGFR gene, with potential value for the prioritization of appropriate candidate therapies (e.g., RTK inhibitors targeted to EGFR alteration).[25] In this study, EGFR was picked only because in the dataset of the 65 patients, this was the only gene with more than 20% frequency of driver-event occurrence within the dataset. There are multiple known drivers for GBM (e.g., PTEN, PDGFRA, etc.), and indeed, it would be very interesting to study their status as a function of spatial diversity. However, their occurrence in the dataset was low and thus, it was infeasible to build a classifier to predict driver status reliably. For future work, other investigations could include noninvasive assessment of pathway activity (rather than single gene entities). This would permit the assessment of groups of genes participating in tumorogenesis, rather than individual pathway components, perhaps being more relevant to the systems biology of the disease.

In this study, we have shown that radiologically defined habitat features are potential surrogates of both OS and EGFR status and can be used as prognostic tools as well as for noninvasive assessment of EGFR-driven tumors (this could have value for determining eligibility for EGFR-targeted therapies). However, there are some limitations. As with almost any retrospective analysis of multisite radiology data, one potential limitation in our study was the variation in scanning and acquisition protocols across MRI systems within the publicly available TCIA database. Although we performed intensity normalization to account for some of this variation, the impact of such variation in image resolution on spatial diversity features needs to be examined more systematically. Further, variation across cancer treatment regimens, such as surgery, radiation, and chemotherapy, may have an effect on the survival rates of the patients as well. Also, the assessment of the predictive utility of these spatial diversity features in an independent validation cohort with matched clinical characteristics is essential to assess their prognostic reliability and robustness. Finally, since habitat abundances have been shown to be associated with survival,[22] the role of these diversity features in the context of clinical variables like age, Karnofsky score, and habitat abundance will be useful to understand the added predictive value of these spatial diversity features.

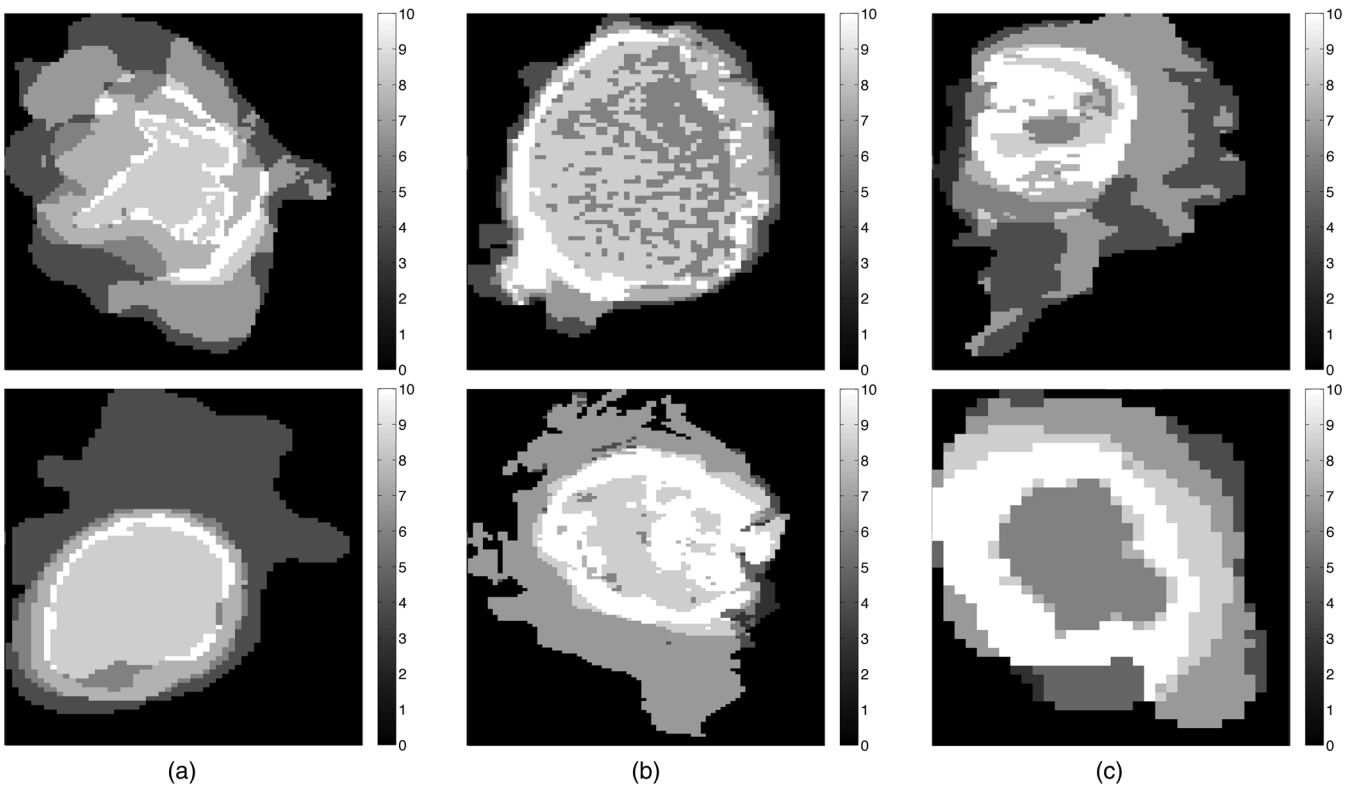


Fig. 6 Examples of different ROI spatial habitat maps combining the low- and high-intensity groups in T1 postcontrast and T2 FLAIR ROIs for (a) mean Fisher's alpha, (b) skewness of the inverse Simpson, and (c) standard deviation of Fisher's alpha between low survival patients (<12 months) with EGFR-driven (upper panel in each column) and high survival patients (>12 months) with non-EGFR driven (bottom panel in each column).

In this study, we used spatial diversity analysis of radiological habitats to investigate spatial heterogeneity characteristics of the tumor for their association with 12-month survival and EGFR driver gene status of patients with GBM. This type of diversity analysis is, to the best of our knowledge, a novel way to analyze multiparametric MRI data. As alluded to earlier, such an investigation is pertinent to both radiomics and radiogenomic analysis paradigms. Specifically, we have focused on a new radiomic characterization of tumor diversity, based on radiologically defined habitats. Further, these features have been used to assess relationships with genomic events in the tumor, namely, driver status of EGFR. This radiogenomics or imaging-genomic analysis reveals that image-derived features could serve as potential noninvasive surrogates of tumor biology. Thus, the spatial diversity of the habitats within the tumor might have information associated with the biology of the tumor. Figure 5 shows examples of ROI spatial habitat maps in a low survival patient (4.8 months) and a high survival patient (57.8 months). Also, Fig. 6 shows examples of different spatial habitat maps for three different diversity indices between low survival patients with EGFR-driven and high survival patients with non-EGFR driven tumors. The EGFR driver event, though an early event, could potentially initiate a phenotypic evolution of the tumor (that manifests itself as distinct spatial distributions of tumor habitats when assessed by MRI). It has been reported[43,44] that the EGFR pathway regulates multiple key phenotypes such as cell proliferation, angiogenesis, invasion, and metastasis. These phenotypes have distinctly different characteristics in MRI. Our results suggest that the spatial diversity of radiologically observed habitats within the tumor region could act as a surrogate for the altered EGFR status. A mechanistic relationship can only be reliably inferred via *in-vivo* experiments and could be an interesting avenue for follow-up investigation. Such spatial diversity analysis of the tumor habitats[21] might provide an additional characterization of the tumor ecological landscape, complementing previous work on habitat abundance within tumors.[21,22]

Our studies in this cohort have shown that several habitat diversity features are associated with survival and EGFR driver gene status with ROC prediction accuracies of 0.67 for 12-month survival and 0.79 for EGFR driver gene status. However, we note that these results remain to be confirmed in an independent cohort of patients with GBM. Nonetheless, these results indicate that such tumor habitat features could potentially be a useful clinical prognostic tool in radiology studies, in addition to providing a noninvasive surrogate of tumor biology (via inference of underlying gene driver status). Further, though this study has been done using only two sequences, T1 postcontrast and T2 FLAIR, there is no conceptual barrier to doing this kind of analysis with more sequences in the multiparametric MRI context. Also, a principled study of driver status inference using radiology habitat features for all other GBM drivers[23] is a topic of future study, subject to the identification of a suitable clinical cohort with sufficient samples in both the driver and nondriver groups.

Appendix

The Shannon index is a measure for diversity in ecology and takes into account both the abundance and evenness of point types present in a region and is defined as

$$H = -\sum_{i=1}^{S} p_i \log\ p_i, \tag{3}$$

where p_i is the proportional abundance of type (species) i and S is the number of types in an area.

The Simpson diversity index is a measurement that accounts for the abundance and the proportion of each species (type) within a region. More specifically, the Simpson diversity index represents the probability that two randomly selected individual points in a region belong to different types and is defined as

$$D_1 = 1 - \sum_{i=1}^{S} p_i^2. \tag{4}$$

The inverse Simpson index represents the number of equally common types that will produce the observed probability that two randomly selected individual points in the region belong to the same "type" and is defined as

$$D_2 = \frac{1}{\sum_{i=1}^{S} p_i^2}. \tag{5}$$

The maximum value will be the number of types in the region, with a high value of the inverse Simpson index representing a high degree of diversity.

Fisher's alpha, also known as the log series, is a diversity index that is used to measure abundance within a spatial region, and it assumes that species abundance follows a logarithmic distribution. This index is defined as

$$S = \alpha \ln\left(1 + \frac{N}{\alpha}\right), \tag{6}$$

where S is the number of species in the region, N is the number of individuals sampled, and α is a Fisher's constant derived from the sample data. Also, the expected number of types with n individuals can be calculated in Fisher's logarithmic series:

$$S_n = \frac{\alpha x^n}{n}, \tag{7}$$

where S_n is the number of types with an abundance of n.

Pielou's evenness is a measurement representing the species (type) evenness within a region and is defined as

$$J = H/\log(k), \tag{8}$$

where k is the number of point types.

Acknowledgments

The authors acknowledge the support of NCI P30 CA016672, a UTMDACC Institution Research Grant and a Career Development Award from the Brain Tumor SPORE (to A.R.), NIH award K08NS070928 (to G.R.) and start-up funding (to A. R.) from MD Anderson Cancer Center for this research. We would also like to thank Sarah Bronson, scientific editor, for her help with manuscript editing and suggestions.

References

1. K. R. Porter et al., "Prevalence estimates for primary brain tumors in the United States by age, gender, behavior, and histology," *Neuro Oncol.* **12**(6), 520–527 (2010).

2. D. R. Johnson and B. P. O'Neill, "Glioblastoma survival in the United States before and during the temozolomide era," *J. Neuro-Oncol.* **107**(2), 359–364 (2012).
3. C. S. McArdle et al., "Prospective study of colorectal cancer in the west of Scotland: 10-year follow-up," *Br. J. Surg.* **77**(3), 280–282 (1990).
4. F. Ng et al., "Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival," *Radiology* **266**(1), 177–184 (2013).
5. F. Tixier et al., "Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer," *J. Nucl. Med.* **52**(3), 369–378 (2011).
6. S. Chicklore et al., "Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis," *Eur. J. Nucl. Med. Mol. Imaging* **40**(1), 133–140 (2013).
7. P. L. Bedard et al., "Tumour heterogeneity in the clinic," *Nat. Biotechnol.* **501**(7467), 355–364 (2013).
8. M. Zhou et al., "Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results," *Trans. Oncol.* **7**(1), 5–13 (2014).
9. M. D. Inda et al., "Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma," *Genes Dev.* **24**(16), 1731–1745 (2010).
10. A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences," *Biochim. Biophys. Acta* **1805**(1), 105–117 (2010).
11. W. Chen et al., "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," *Med. Phys.* **33**(8), 2878–2887 (2006).
12. S. C. Agner et al., "Textural kinetics: a novel dynamic contrast-enhanced (DCE)-MRI feature for breast lesion classification," *J. Digital Imaging* **24**(3), 446–463 (2011).
13. S. Herlidou-Meme et al., "MRI texture analysis on texture test objects, normal brain and intracranial tumors," *Mag. Reson. Imaging* **21**(9), 989–993 (2003).
14. N. M. Cheng et al., "Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma," *J. Nucl. Med.* **54**(10), 1703–1709 (2013).
15. C. C. Chen, J. S. Daponte, and M. D. Fox, "Fractal feature analysis and classification in medical imaging," *IEEE Trans. Med. Imaging* **8**(2), 133–142 (1989).
16. K. Held et al., "Markov random field segmentation of brain MR images," *IEEE Trans. Med. Imaging* **16**(6), 878–886 (1997).
17. L. Tesar et al., "Medical image analysis of 3D CT images based on extension of Haralick texture features," *Comput. Med. Imaging Graph.* **32**(6), 513–520 (2008).
18. K. A. Miles et al., "Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival," *Radiology* **250**(2), 444–452 (2009).
19. B. Crespi and K. Summers, "Evolutionary biology of cancer," *Trends Ecol. Evol.* **20**(10), 545–552 (2005).
20. J. Oksanen et al., *Package 'vegan', Community Ecology Package Version 2*, R foundation, Vienna, Austria (2013).
21. R. A. Gatenby, O. Grove, and R. J. Gillies, "Quantitative imaging in cancer evolution and ecology," *Radiology* **269**(1), 8–15 (2013).
22. M. Zhou et al., "Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results," *Trans. Oncol.* **7**(1), 5–13 (2014).
23. V. Frattini et al., "The integrated landscape of driver genomic alterations in glioblastoma," *Nat. Genet.* **45**(10), 1141–1149 (2013).
24. H. Ying et al., "Mig-6 controls EGFR trafficking and suppresses gliomagenesis," *Proc. Natl. Acad. Sci. U. S. A.* **107**(15), 6912–6917 (2010).
25. E. Padfield, H. P. Ellis, and K. M. Kurian, "Current therapeutic advances targeting EGFR and EGFRvIII in glioblastoma," *Front. Oncol.* **5**(5), 1–8 (2015).
26. W. B. Pope, "Genomics of brain tumor imaging," *Neuroimaging Clin. North Am.* **25**(1), 105–119 (2015).
27. V. Kumar et al., "Radiomics: the process and the challenges," *Mag. Reson. Imaging* **30**(9), 1234–1248 (2012).
28. http://cancerimagingarchive.net.
29. http://www.cbioportal.org.
30. P. J. Stephens et al., "The landscape of cancer genes and mutational processes in breast cancer," *Nat. Biotechnol.* **486**(7403), 400–404 (2012).
31. M. A. Mazurowski, A. Desjardins, and J. M. Malof, "Imaging descriptors improve the predictive power of survival models for glioblastoma patients," *Neuro Oncol.* **15**(10), 1389–1394 (2013).
32. www.mipav.cit.nih.gov.
33. www.mitk.org.
34. D. Reynolds, "Gaussian mixture models," *Encycl. Biom.* 659–663 (2009).
35. A. E. Magurran, *Measuring Biological Diversity*, John Wiley & Sons (2013).
36. J. Oksanen et al., "The vegan package," *Community Ecol. Package* (2007).
37. A. Baselga, "Partitioning the turnover and nestedness components of beta diversity," *Global Ecol. Biogeogr.* **19**(1), 134–143 (2010).
38. R. H. Whittaker, "Vegetation of the Siskiyou mountains, Oregon and California," *Ecol. Monogr.* **30**(3), 279–338 (1960).
39. R. H. Whittaker, "Evolution and measurement of species diversity," *Taxon* **21**, 213–251 (1972).
40. M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Sci. Adv.* **324**(5923), 81–85 (2009).
41. S. J. Mason and N. E. Graham, "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation," *Q. J. R. Meteor. Soc.* **128**(584), 2145–2166 (2002).
42. NCAR–Research Applications Laboratory, "Weather forecast verification utilities," 10 July 2015, http://CRAN.R-project.org/package=verification.
43. K. M. Talasila et al., "EGFR wild-type amplification and activation promote invasion and development of glioblastoma independent of angiogenesis," *Acta Neuropathol.* **125**(5), 683–698 (2013).
44. H. W. Lo, "EGFR-targeted therapy in malignant glioma: novel aspects and mechanisms of drug resistance," *Curr. Mol. Pharmacol.* **3**(1), 37–52 (2010).

Joonsang Lee is a postdoctoral fellow in the Department of Bioinformatics and Computational Biology at the University of Texas MD Anderson Cancer Center. He received his PhD in the Department of Physics at the University of Georgia. His research focuses primarily on image processing on brain tumor images with various statistical techniques, such as machine learning, classification, and clustering algorithms.

Shivali Narang is a research assistant 1 in the Department of Bioinformatics and Computational Biology at the University of Texas MD Anderson Cancer Center. She obtained her bachelor's degree in biomedical engineering from the University of Houston, Texas, in 2014. Her work focuses on linking imaging data with genomics data using image-processing and data mining concepts.

Juan J. Martinez holds both a bachelor's degree in electrical engineering from Monterrey Institute of Technology and a master's degree in bioengineering from Rice University. During his graduate studies, he investigated the construction of novel imaging systems to enable early cancer detection through *in vivo* confocal microscopy and spectroscopy. He is currently a clinical specialist at Brainlab, where he provides on-site consulting to neurosurgeons and other medical personnel about cancer treatment solutions based on image-guided surgery techniques.

Ganesh Rao received his undergraduate degrees in chemistry and microbiology and his medical degree from the University of Arizona. He completed a residency in neurological surgery at the University of Utah. He is currently an associate professor of neurosurgery at the University of Texas, MD Anderson Cancer Center. His laboratory and clinical research interests include understanding the process of malignant progression in brain tumors.

Arvind Rao is an assistant professor in the Department of Bioinformatics and Computational Biology at the University of Texas, MD Anderson Cancer Center. He obtained his PhD from the University of Michigan, Ann Arbor. His work focuses on building decision algorithms that integrate imaging and genetics data in the context of cancer prognosis and treatment.