RADIOGENOMIC CORRELATION FOR PROGNOSIS IN PATIENTS WITH GLIOBLASTOMA MULTIFORME

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Bioinformatics & Medical Informatics

by

Pallavi Machaiah Karnayana

Spring 2013
The Undersigned Faculty Committee Approves the
Thesis of Pallavi Machaiah Karnayana:

Radiogenomic Correlation for Prognosis in Patients with Glioblastoma Multiformae

Usha Sinha, Chair
Department of Physics

Faramarz Valafar
Bioinformatics and Medical Informatics Program

Mauro Tambasco
Department of Physics

4/17/2013
Approval Date
Copyright © 2013
by
Pallavi Machaiah Karnayana
All Rights Reserved
DEDICATION

To my family for their support throughout.
ABSTRACT OF THE THESIS

Radiogenomic Correlation for Prognosis in Patients with Glioblastoma Multiformae
by
Pallavi Machaiah Karnayana
Master of Science in Bioinformatics and Medical Informatics
San Diego State University, 2013

Introduction: Prognostic and predictive models based on pre-operative magnetic resonance (MR) imaging features can be a valuable tool in a clinical setting. The goal of this project is to determine the correlation between a highly prognostic genomic signature and a clinically applicable classification based on a combination of clinical (age, Karnofsky Performance Status) and qualitative imaging features (e.g., tumor volume, edema, non-contrast enhancing tumor (nCET), multifocality).

Methods: Radiogenomic correlation methods were carried out using the caIntegrator tool. Publicly available data known as The Cancer Genomic Atlas (TCGA)-glioblastoma multiformae (GBM) and expert derived qualitative imaging features from The Cancer Imaging Archive (TCIA) website were used. The caIntegrator tool that allows users to access and search across databases was used to perform radiogenomic correlations. The imaging features were categorized by experts reading pre-operative MRI scans and gene signatures were based on gene expression levels. The correlation of gene expression levels to three imaging features (edema, non-contrast enhancing tumor (nCET), satellites) was explored. Further, the correlation of two clinical features (age and KPS) to gene expression levels was also investigated. Cohorts segregated by median value of the top up/down regulated gene were then analyzed for median survival time. The mean value of the top upregulated gene was determined for each GBM subtype to test for significant differences in GBM sub-types.

Results: All three imaging features identified the same top regulated gene Periostin (POSTN). Kaplan-Meier (KM) analysis using POSTN median value showed that subjects in the cohort with greater than median value had a lower mean survival time. The Classical GBM subtype also had significantly elevated values of the POSTN compared to the other sub types.

Discussion: The top upregulated gene identified in this study was the same as that identified by an earlier study using the quantitative index of edema volume. This shows that qualitative index of edema volume (this study) was as effective as the quantitative index of edema volume. The other two imaging features (nCET and satellite) were shown to be associated with edema volume; this may partially explain that the same set of genes were upregulated for all imaging features. Based on the findings in this pilot study, a more comprehensive study of more subjects, qualitative and quantitative imaging features is warranted.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ix</td>
</tr>
</tbody>
</table>

## CHAPTER

1 INTRODUCTION
   1.1 Background and Statement of Problem
   1.2 Overview of the Project
   1.3 Definition of Terms
      1.3.1 Imaging Descriptors
      1.3.2 GBM Subtype Classification

2 REVIEW OF THE LITERATURE

3 METHODS
   3.1 Resources
   3.2 caIntegrator Application
   3.3 Kaplan-Meier Analysis
   3.4 Karnofsky Performance Scoring System
   3.5 Analysis Procedure

4 RESULTS

5 DISCUSSION

REFERENCES

LIST OF TABLES

PAGE

Table 4.1. Distribution of Subjects Based on Imaging Features ..............................................13
Table 4.2. Subtype Class and Associated Imaging Features ...................................................14
Table 4.3. Top 5 Upregulated Gene for Prominent Imaging Feature ......................................15
Table 4.4. Mean Top Upregulated Gene Value for Each GBM Subtype ................................16
Table 4.5. Distribution for Different Imaging Feature Pairs ...................................................18
Table 5.1. AK Classification ....................................................................................................20
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>(A) Axial T1 weighted image representing necrosis.</td>
<td>4</td>
</tr>
<tr>
<td>3.1</td>
<td>CaIntegrator web interface.</td>
<td>9</td>
</tr>
<tr>
<td>3.2</td>
<td>Response criteria of imaging observations.</td>
<td>10</td>
</tr>
<tr>
<td>4.1</td>
<td>KM plot of subjects partitioned by a POSTN upregulation threshold.</td>
<td>17</td>
</tr>
<tr>
<td>4.2</td>
<td>KM plot for the clinical query of age and KPS score.</td>
<td>17</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

An endeavor over a period can be successful only with the advice and guidance of well-wishers. I take this opportunity to express my deep sense of gratitude and my heartfelt thanks to all those people whose constant support and encouragement has helped me to transform my efforts into completion of this project. I express my deep gratitude and respect to Dr. Faramarz Valafar, Director, Bioinformatics and Medical Informatics (BMI), San Diego State University (SDSU). I would like to express my sincere gratitude and profound regards to Dr. Mauro Tambasco, Physics, SDSU, for his constant motivation and support in every manner for the successful completion of this project. I owe much to my project guide Dr. Usha Sinha, Chair, Physics, SDSU, her expert support, patience, co-operation and guidance in all ways for the successful completion of the project. I would like to thank all the committee members for accommodating my request in spite of their busy schedule and helping me in all ways possible. I extend my thanks to my family and friends for their constant support during the course of the project.
CHAPTER 1

INTRODUCTION

The research study explores gene expression profiles for cohorts separated by imaging features.

1.1 BACKGROUND AND STATEMENT OF PROBLEM

Various genetic factors cause cancer and prognosis is an integral part of clinical evaluation in patients. In recent years, development of high-throughput genomic technologies has driven the cancer genomics field into a rapidly evolving discipline.1

Glioblastoma multiforme (GBM) is the most common malignant primary brain tumor in humans. Malignant gliomas are heterogeneous tumors in both appearance and gene expression. GBM is grade IV brain tumor and is characterized by the presence of necrotizing tissue and hyperplastic blood vessels and differ from astrocytomas (grade III brain tumors). Patients have median survival duration of 15 months with standard-of-care radiation and chemotherapy with temozolomide and patients have median survival duration of 4 1/2 months without treatment.

The discovery of targetable genes responsible for cell spread and invasion can be expected to impact modern therapy and patient survival. In order for personalized medicine to transpire, a cost-effective biomarker that accurately reflects underlying molecular cancer compositions is urgently needed. In this study, we specifically seek to identify and correlate edema GBM radiophenotype to genome signatures. Imaging phenotypes of GBM obtained from routine clinical MRI studies are associated with a specific gene signature; imaging phenotypes serve as non-invasive surrogates for diagnosis, prognosis and optimal treatment.

Whole genome gene and expression micrornucleic acid (microRNA) analysis can potentially be used for cancer characterization, but they are expensive, invasive and localized. Our ultimate goal is quantitative analysis, but our area of focus is on the wide range of options available to perform qualitative analysis on datasets of patients.
1.2 Overview of the Project

The scope of the current project is to explore gene expression profiles for cohorts separated by imaging features. The imaging features are qualitative features based on expert consensus by radiologists, available through caIntegrator, and pertinent imaging features are selected based on a prior study correlating image findings to survival.\textsuperscript{2} We have focused on options available to probe genome data from image data and for qualitative analysis of imaging and genomic findings.

Simple correlations between the 3 most prominent features (edema, nCET, satellite) among the 24 most common magnetic resonance imaging (MRI) features and genomic signatures are analyzed. This study includes identification of different patterns of gene expression, image-based features, and their correlation among TCGA-GBM patients. The four groups that separate GBM patients are Proneural, Neural, Classical and Mesenchymal. The four GBM subtypes differ by the type of genetic abnormalities, clinical characteristics and histological features. The research study is an attempt to aid researchers and clinicians in the individualized treatment of GBM patients.

1.3 Definition of Terms

Significant imaging descriptors which are a measure of T1 and T2 weighted MR images are defined in this subsection.

1.3.1 Imaging Descriptors

MRI provides non-invasive comprehensive evaluation of complex and heterogeneous appearance of high grade astrocytomas and glioblastomas. Cancer-associated gene alterations exploit many different molecular mechanisms that disrupt cellular pathways and result in uncontrolled cell proliferation. Volumes of radiophenotype such as edema/tumor infiltration, enhancing tumor and necrosis are currently correlated with the genomic findings.

Cysts are well defined, rounded, eccentric regions of very bright T2W and low T1W signal. Necrosis is defined as region within the tumor that does not enhance or shows markedly diminished enhancement, is high on T2W and proton density images, is low on T1W images, and has an irregular border. Multifocal is defined as having at least one region of tumor, either enhancing or non-enhancing, which is not contiguous with the dominant
lesion and is outside the region of signal abnormality (edema) surrounding the dominant mass. Some of the imaging observations we have considered in this study are:

- **nCET**: nCET are regions of T2W hyper intensity that are associated with mass effect and architectural distortion, including blurring of the gray-white interface.

- **Edema**: Edema are regions of very bright T2W signal, that are not associated with mass effect and architectural distortion and white-matter tracts extending significantly beyond (>1 cm) the margins of the tumor.

- **Satellite**: A satellite lesion is within the region of signal abnormality surrounding the dominant lesion but not contiguous in any part with the major tumor mass.

### 1.3.2 GBM Subtype Classification

Four distinct subtypes of GBM are distinguished by gene expression patterns and clinical characteristics.

1. **Classical**: has extra copies of the epidermal growth factor receptor (EGFR) gene and higher than normal expression of EGFR. Clinically, the Classical group survived the longest of the subgroups in response to aggressive treatment.

2. **Mesenchymal**: has high rates of mutations or other alterations in neurofibromatosis type 1 (NF1). Clinically, the Mesenchymal group had significantly reduced mortality in response to aggressive treatment.

3. **Neural**: has the expression of neuron markers such as neurofilament (NEFL), gamma-aminobutyric acid 1 (GABRA1), Synaptotagmin 1 (SYT1) and solute carrier family 12 (SLC12A5). Clinically, the Neural group had efficacy suggested in response to aggressive treatment.

4. **Proneural**: has high rates of alterations in tumor protein 53 (TP53), platelet-derived growth factor receptor (PDGFRA), Isocitrate dehydrogenase 1 IDH1. Clinically, the Proneural group had younger patients overrepresented and aggressive treatment did not alter survival. MR images of patients with GBM is shown in Figure 1.1.
Figure 1.1. (A) Axial T1 weighted image representing necrosis. (B) Axial post T2-weighted image for vasogenic edema. (C) Axial T1-weighted image. (D) Axial T2-weighted image for non-enhancing tumor.
CHAPTER 2

REVIEW OF THE LITERATURE

Gene signatures are characteristic of a medical condition and a number of techniques on how to probe genome data using image data have been suggested in the papers referenced. The TCGA web portal has publicly available results on GBM subtypes based on histology. Radiogenomics has experienced a boom in the recent years and has great potential in linking gene expression profiles with MRI phenotypes.

Zinn et al. have published quantitative MR imaging analysis based on edema tumor volume calculated from Fluid attenuated inversion recovery (FLAIR) imaging sequences. A cutoff based on the edema volume (state the values for the high/low volumes) was used to select subjects in two cohorts. The genome analysis of these two cohorts identified periostin (POSTN) and miRNA-219 as the top upregulated gene and microRNA respectively. A median value of differential fold regulation among high versus low groups by Comparative Marker selection was chosen to divide all the GBM subjects in TCGA database into two cohorts. Kaplan Meier (KM) analysis demonstrated that above median expression of POSTN results in significantly decreased survival and shorter time of disease progression as also shown in our KM plot (log rank p-value 0.001=probability of survival 1). High POSTN was significantly associated with the classical and mesenchymal GBM subtype (P<0.0001). In this study, significant upregulated and downregulated gene is associated with mesenchymal GBM subtype, suggesting a regulatory network promoting mesenchymal transition and cellular invasion in GBM.

Another study Zinn et al. is based on a combination of clinical and imaging features. In the study a 3-point scoring system based on tumor volume, patient age and KPS were combined to form the VAK classification. VAK-A is significantly associated with TP53 activation, while VAK-B shows significant TP53 inhibition. A median cut-off for the differential fold regulated gene was used to divide GBM patients into two cohorts through Comparative Marker selection. In this study, disease progression P<0.001 corresponds to probability of survival ranging from 0 to 1 from KM analysis. Some datasets/studies do not provide date fields for survival data but do provide disease progression last follow-up or
death durations (i.e. months to death from enrollment date etc.). In these cases either we have to ask the user to make up (calculate) dates based on a fixed enrollment date to make the KM plot work. This calculation should be done by the software and column choices should not be limited to dates types only.

The study by Ovaska et al. based on the integration of multi-dimensional data including clinical parameters, imaging observations and genomic knowledge from bio-databases are an essential part of diagnosis, treatment and prevention strategies of diseases.\(^5\) This requires computational tools designed for translating large-scale genome data. In order to enable users who do not have skills in computer science and/or in bio-informatics, there have been concerted efforts to provide publicly available large-scale data integration framework and tools. One such project is the Anduril framework which can analyze heterogeneous and multidimensional genomic data, wherein genes are sorted according to the survival P-value on the exon platform (exon array values, transcript levels). The significant feature of Anduril analysis is integration of patient survival information with both expression and SNP data, thereby allowing the user to sort genomic alterations with respect to clinical matrix. One outcome of analysis on the Anduril platform is the identification of novel genetic alteration linkage to GBM progression, where the GBM associated Moesin gene has a strong survival effect. Some discoveries such as molecular inhibition of Edema/cellular invasion is proposed to bring improvement in therapy and patient survival in GBM. In this case analysis, identification of genetic loci and genes that have significant survival effect is the central idea among large-scale molecular data.

The study by Verhaak et al. focused on integrated genomic analysis on the clinically relevant subtypes of GBM.\(^6\) Aberrations and expression patterns of EGFR, NF1, and PDGFRA/IDH1 define the four subtypes of GBM- Proneural, Neural, Classical and Mesenchymal. A highly variable expression across the platforms is filtered and consensus average linking hierarchical clustering identifies four clusters and cluster significance is evaluated. Genes correlated with each subtype are selected using Significance Analysis of Microarrays (SAM) and Receiver Operating Characteristic (ROC) methods. It is summarized in the study that four tumors of GBM are associated with imaging features and are independent of the percentage of tumor nuclei and Mesenchymal subjects are found to have higher fraction of necrosis.
The UCSC cancer genomics browser is open-source software that comprises web-based tools to visualize and analyze cancer genomics and clinical data.\textsuperscript{1} The repository also has affiliations with the TCGA and has provisions for several novel tumor image and genomic viewers (hgMicroscope, hgSignature). Data interpretation in the context of biological pathways, pathway recognition algorithm using data integration on genomic models (PARADIGM) integrates functional genomic data in patient samples.

In the study by Marko et al., clustering algorithms including support vector machine (SVM) is used to group tumors into genotypic and corresponding phenotypic subtypes and classification algorithms are used to train unknown GBMs from trained expression data.\textsuperscript{7} Differential fold regulated genes are found through classification methods and are cross-checked and verified for differential fold regulation of a subset of the same genes using reverse transcriptase polymerase chain reaction (RT-PCR) fingerprint analysis.

The case study illustrates use of TCGA, microarray data from caArray, and imaging data from National Biomedical Imaging Archive (NBIA) and AIM via caGrid. Subjects with specific genomic criteria, clinical characteristic, or imaging observations are searched.\textsuperscript{8} Kaplan-Meier (KM) plot analysis is performed to determine how patients with hemorrhagic tumors fare in comparison to those without hemorrhage. Gene expression is evaluated as a function of fold change for the genomic control samples uploaded in the study. The user can analyze the data in real time for genes with their expression levels and up/down regulation. Gene pattern analysis is performed using principal component analysis (PCA), comparative marker selection (CMS).
CHAPTER 3

METHODS

The method section provides a detailed explanation on the analysis tools used for assessing radiogenomic correlation on GBM subjects.

3.1 RESOURCES

The databases used in the study are:

- **TCGA** is a catalogue of genetic mutations responsible for cancer, using genome sequencing and bioinformatics. TCGA is supervised by the National Cancer Institute and the National Human Genome Research Institute. The overarching goal of TCGA is to improve our ability to diagnose, treat and prevent cancer. The TCGA data portal provides a platform for researchers to search, download and analyze data sets generated by TCGA. The portal contains all TCGA data sets pertaining to clinical information, genomic characterization, and high-throughput sequencing analysis of the tumor genomes.

  Data currently available through the TCGA portal include molecular characterization data sets for ovarian cancer, and molecular characterization, high-throughput sequencing, and clinical data for glioblastoma multiforme (GBM) tumors. New data are derived on an ongoing basis from TCGA analyses and are deposited into databases for download in the portal.

- **TCIA** is publicly available and images are organized as "Collections". TCIA is supervised by the Cancer Imaging Program of the National Cancer Institute. The image data in TCIA is organized into purpose-built Collections. A collection of studies typically includes studies from several subjects (patients) followed over time. The subjects typically have in common a particular disease and/or particular anatomical site (lung, brain, etc.). TCIA submissions are organized in the following hierarchy: Collection > Patient (Subject) > Study > Series > Images. TCGA-GBM collection used in our project has primarily MR with a few CT studies.

  Computer and informatics expertise is required to analyze the vast amount of information (DNA sequence variants, copy-number alterations (CNA), epigenetic and transcriptomic changes) stored in the genome databases such as TCGA. This had led to the development of web based tools that enables users to query these databases and conduct their
own studies. The tools in the browser used in this study, caIntegrator, allow one to query the clinical, genomic and imaging information.

- caIntegrator is a web-based software application that allows researchers to set up custom, Cancer Biomedical Informatics Grid (caBIG)-compatible web portals to conduct integrative research. These portals brought together heterogeneous clinical, microarray and medical imaging data and is depicted in Figure 3.1.

- caArray guides the annotation and exchange of array data using a federated model of local installations whose results were shareable across caBIG. Array Providers like Affymetrix, Agilent, GenePix, Illumina, ScanArray. caArray is open source software operating via caBIG and also provides an Annotate and Image mark-up tool for quantitative prospects.

![Figure 3.1. CaIntegrator web interface.](image)

### 3.2 CaIntegrator Application

Query and Browse Study Data: The tool allows one to browse and filter lists of study objects, such as patients, image series, and genes. Specific features can be selected for each study object and query results can be saved or exported to spreadsheets.

Deploy Study: The Study Manager was used to define the sources of data for a new study. The data was then accessed over the Grid or uploaded from files. The Study Manager defined relationships between patients and samples so that data could be compared across the
data sources. The Study Manager also defined mapping of uploaded annotation fields to cancer Data Standards Repository (caDSR) data types.

Perform Analysis: Data from calIntegrator2 was sent to other software packages for more sophisticated analysis. GenePattern and GeWorkbench were available for clustering and classification analysis and several other analysis tools were built in, including Kaplan-Meier survival plots.

Each criterion in a study does not have to have the same annotations. A search can be filtered using operators and automatically populates a set of permissible values. Criteria can be clinical, imaging or gene expression. Descriptions of the imaging features and the appropriate response criteria are listed in Figure 3.2:

![Figure 3.2. Response criteria of imaging observations.](image)

### 3.3 Kaplan-Meier Analysis

Kaplan-Meier analysis allows estimation of survival over time, even when patients drop out or are studied for different lengths of time. It estimates a population survival curve from a sample. For each interval, survival probability is calculated as the number of patients surviving divided by the number of patients at risk.
Interpretation of Kaplan-Meier Curves: Vertical axis represents estimated probability of survival for a hypothetical cohort, not actual % surviving. Precision of estimates depend on the number of observations. The KM plots were used to compare survival statistics among comparative groups.

3.4 Karnofsky Performance Scoring System

Karnofsky performance scoring (KPS) system was named after Dr. David A. Karnofsky where the survival rate divides the subjects into high and low survival population. Performance status is an attempt to quantify cancer patients' general well-being and activities of daily life. The scoring system is used to determine whether they could receive chemotherapy, whether dose adjustment is necessary, and as a measure of the required intensity of palliative care. The Karnofsky score runs from 100 to 0, where 100 is "perfect" health and 0 is death.

3.5 Analysis Procedure

The pilot study was performed on TCGA GBM data and the caIntegrator was used to conduct the analysis. 68 patients with unique subject identifiers had imaging features associated with them on caIntegrator. The patient cohorts A and B were formed based on the following imaging/clinical four criteria:

1. AKA/AKB based on clinical descriptors of age and KPS score (AK). For cohort A, age >=60 years and KPS<100 and for cohort B, age <60 years and KPS>100 was used.

2. EdemaA/EdemaB based on high volume of edema (A) and low volume of edema (B)

3. nCET A/nCET B based on presence (A) or absence (B) of non contrast enhancing tumor

4. Satellite A/Satellite B based on presence (A) or absence (B) of satellites.

The Comparative Marker Selection (CMS) module: Given two classes of samples, CMS finds expression values that correlate with the difference between those two classes. For each cohort pair, CMS analysis was performed where 22627 probe ids and corresponding fold changes were recorded. CMS module outputs were open document format (ODF) file that contained 18 Header/attribute values, gene cluster text (GCT) file that contained a column for each sample, a row for each gene, and an expression value for each gene in each sample and categorical class (CLS) file that contained the total number of samples, classes.
The top upregulated gene was extracted and the median value of the top upregulated gene was used to divide all TCGA GBM subjects into two cohorts for Kaplan Meier analysis.

Genomic signatures or gene expressions were represented as an algebraic expression over a set of genes. Once a signature was defined, the probe values for each sample were substituted for gene names and the algebraic expression was evaluated. A very limited set of predefined signatures were currently available.
CHAPTER 4

RESULTS

The distribution of patients with each of the imaging features used in the present study is summarized in Table 4.1.

<table>
<thead>
<tr>
<th>Imaging Observations</th>
<th>Y/Low/A</th>
<th>N/High/B</th>
</tr>
</thead>
<tbody>
<tr>
<td>#subjects</td>
<td></td>
<td>#subjects</td>
</tr>
<tr>
<td>nCET</td>
<td>14</td>
<td>55</td>
</tr>
<tr>
<td>Edema</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Satellite</td>
<td>14</td>
<td>55</td>
</tr>
<tr>
<td>Age,KPS</td>
<td>119</td>
<td>34</td>
</tr>
</tbody>
</table>

The publicly available study- TCGA GBM on caIntegrator has the number of subjects divided based on histopathological features and lists the imaging features associated with each subtype. MRI features that distinguish GBM tumor subtypes are listed in Table 4.2.

The Comparative Marker Selection analysis tool in caIntegrator was used to divide GBM subjects into 2 cohorts. The median value of differential fold regulation among high/low Edema, no/yes nCET and no/yes Satellite was found. The same gene was identified as the top upregulated gene associated with each of the imaging features (nCET, Satellite and edema): POSTN. The 3 imaging features were significant prognostic indicators.
Table 4.2. Subtype Class and Associated Imaging Features

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Subjects</th>
<th>Imaging Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>20</td>
<td>Cysts No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Calvarial Remodeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multifocal</td>
</tr>
<tr>
<td>Mesenchymal</td>
<td>41</td>
<td>Gliomatosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multifocal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multicentric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Calvarial Remodeling</td>
</tr>
<tr>
<td>Neural</td>
<td>17</td>
<td>Ependymal Extension</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cysts</td>
</tr>
<tr>
<td>Proneural</td>
<td>45</td>
<td>Calvarial Remodeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cysts</td>
</tr>
</tbody>
</table>
The top 5 upregulated genes for each imaging feature and their fold increase is shown in Table 4.3.

**Table 4.3. Top 5 Upregulated Gene for Prominent Imaging Feature**

<table>
<thead>
<tr>
<th>Genes</th>
<th>Mean Fold change</th>
<th>Edema low/high</th>
<th>Satellite yes/no</th>
<th>nCET yes/no</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSTN</td>
<td>11</td>
<td>8 up</td>
<td>10 up</td>
<td>11 up</td>
</tr>
<tr>
<td>PPA2</td>
<td>10</td>
<td>6 up</td>
<td>9 up</td>
<td>10 up</td>
</tr>
<tr>
<td>KPNA2</td>
<td>9</td>
<td>6 up</td>
<td>8 up</td>
<td>9 up</td>
</tr>
<tr>
<td>PPAP2A</td>
<td>6</td>
<td>4 up</td>
<td>5 up</td>
<td>6 up</td>
</tr>
<tr>
<td>PAGE4</td>
<td>3</td>
<td>1 up</td>
<td>2 up</td>
<td>3 up</td>
</tr>
</tbody>
</table>

Significant observations from the CMS analysis were:

- The next upregulated genes were PyroPhosphatase 2 (PPA2), karyopherin alpha 2 (KPNA2), phosphate phosphohydrolase 1 (PPAP2A), Prostate-Associated 4 (PAGE4), Peroxiredoxin-4 (PRDX4), palmitoyl-protein hydrolase 1 (PPT1), kininogen 1 (KNG1).
- The mean of expression value is 2125 with a change of 11-fold highest upregulation (POSTN).
- The median of the GBM population for POSTN was determined and cohorts were formed based on the mean value as the threshold. Table 4.4 shows the mean values of the top upregulated gene-POSTN for all GBM subtypes and Mesenchymal subtype is found to be significantly different.\(^6\)
- KM plot with approximately 1-fold change (1.15773) was plotted and cohorts were made with values above and below the threshold. A one fold change is defined as gene
Table 4.4. Mean Top Upregulated Gene Value for Each GBM Subtype

<table>
<thead>
<tr>
<th>GBM Subtype</th>
<th>Mean value upregulated gene (POSTN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>2911.236</td>
</tr>
<tr>
<td>Mesenchymal</td>
<td>3692.682</td>
</tr>
<tr>
<td>Neural</td>
<td>586.1877</td>
</tr>
<tr>
<td>Proneural</td>
<td>554.5659</td>
</tr>
</tbody>
</table>


expression level change from 100 to 150 corresponding to upregulation. The probability of survival ranges from 0 to 1 from KM analysis and is related to disease progression $P < 0.001$. The Kaplan-Meier plot in Figure 4.1 depicts the GBM subjects partitioned by the median value of POSTN. Since all imaging features yielded the same top regulated gene, they all resulted in the same KM plot.

AK-A and AK-B classes according to our analysis shows significant median survival difference of (P-value 0.21 probability of survival 1.23). KM plot of AK-A and AK-B classes, based on clinical descriptors of age and KPS score (AK) is shown in Figure 4.2. The top upregulated gene for AK-A and AK-B is enabled homolog (ENAH) and Ran GTPase-activating protein 1 (RANGAP1).

We found that the AK-A classification is associated with TP53 activation and AK-B classification is associated with TP53 inhibition. The 68 subjects with imaging findings were further analyzed for co-occurrence of imaging features. Table 4.5 shows the distribution for different pairs of imaging features.
Figure 4.1. KM plot of subjects partitioned by a POSTN upregulation threshold.

Figure 4.2. KM plot for the clinical query of age and KPS score.
Table 4.5. Distribution for Different Imaging Feature Pairs

<table>
<thead>
<tr>
<th>Feature combination</th>
<th>#Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCET-no Edema-High</td>
<td>4</td>
</tr>
<tr>
<td>nCET-Y Edema-Low</td>
<td>6</td>
</tr>
<tr>
<td>nCET-Y Edema-High</td>
<td>0</td>
</tr>
<tr>
<td>nCET-no Edema-Low</td>
<td>17</td>
</tr>
<tr>
<td>Sat-No Edema-High</td>
<td>5</td>
</tr>
<tr>
<td>Sat-No Edema-Low</td>
<td>17</td>
</tr>
<tr>
<td>Sat-Y Edema-High</td>
<td>0</td>
</tr>
<tr>
<td>Sat-Y Edema-Low</td>
<td>5</td>
</tr>
<tr>
<td>nCET-no Sat-No</td>
<td>18</td>
</tr>
<tr>
<td>nCET-Y Sat-No</td>
<td>6</td>
</tr>
<tr>
<td>nCET-Y Sat-Y</td>
<td>6</td>
</tr>
<tr>
<td>nCET-no Sat-Y</td>
<td>10</td>
</tr>
</tbody>
</table>
CHAPTER 5

DISCUSSION

This report focuses on exploring the correlation between qualitative imaging features and gene expression data. Three imaging features were chosen based on prior studies that identified these features as predictors of survival time and time to progression. Surprisingly, each imaging feature identified the same set of genes as the top upregulated. One of these genes, POSTN, has also been identified in an earlier study that used the quantitative index of edema volume to segregate the cohorts into low and high volume groups. The same gene was also identified as the top upregulated gene when using the qualitative feature for the edema volume. This shows there is good consistency in the expert classification of edema into high/medium and low volume. It is also encouraging to find that qualitative imaging features can be used in place of quantitative features; the latter features (such as volume of edema) are more difficult to extract as fully automated segmentation techniques are not available. A more comprehensive analysis has to be performed to see what other qualitative features (such as tumor size and necrosis) can be substituted for quantitative features. It may be that a combination of imaging features (quantitative and qualitative) may ultimately provide the best correlation to genomic signatures and serve as prognostic indices.

The analysis also showed that the same set of genes was upregulated for the other imaging features selected in this study (nCET, satellites). This may arise if these imaging features were not independent—it is likely that presence of low volume of edema is associated with presence of nCET, or low volume of edema is associated with absence of satellites. In order to test this hypothesis, the number of co-occurrences of each imaging feature pair was tabulated (Table 4.5).

It can be seen from the Table 4.5 that there are 18 Satellite no-edema low, no nCET(Y)-edema-high co-occurrence. Thus, the same subjects (or close to the same) would have been identified in the cohorts segregated by edema(L) or satellite(N). This is also true for subjects in the cohorts selected by nCET(Y) and edema(H). As the cohorts segregated by each imaging feature contained almost the same set of patients in each, this is likely to be the
reason that the same gene sets were identified as the top up regulated ones. It is also worth investigating if POSTN is responsible for the tissue imaging features: i.e., POSTN upregulation may cause a high volume of edema, absence of nCET as well as the presence of satellites. A pathway analysis is required to identify these links. Further, a more detailed analysis may yield a set of independent imaging features that may be better prognostic predictors as well as classifiers of GBM sub types. AK Classification system based on age and KPS score is tabulated in Table 5.1.

**Table 5.1. AK Classification**

<table>
<thead>
<tr>
<th>Indices KM Plot</th>
<th>AK-A</th>
<th>AK-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>57</td>
<td>26</td>
</tr>
<tr>
<td>Mean survival time</td>
<td>439.4386</td>
<td>647.1923</td>
</tr>
<tr>
<td>Log-rank P-value for significance of difference in survival between groups</td>
<td>AK-A vs AK-B</td>
<td>0.27</td>
</tr>
</tbody>
</table>

The current report is a very preliminary study to establish the feasibility of radiogenomic analysis. This pilot report confirms that analysis can be performed using the publicly available tools such as caIntegrator. However, the drawback is that the image descriptors (based on expert consensus reports) are available on a small subset of the total number of patients in the TCGA database. The next steps are to (i) complete the qualitative image descriptors for all the imaging studies available in TCIA by downloading the datasets and having an expert neuroradiology panel annotate the studies, (ii) extract quantitative descriptors such as volume of tumor/necrosis/edema, 3D shape descriptors of these volumes, tumor edge characteristics, intensity histogram descriptors and 3D texture features.

The inclusion of more subjects with a complete set of qualitative and quantitative features will provide a larger set for deriving correlations to genomic signatures and to develop prognostic models. Furthermore, pathway analysis of upregulated genes will aid in the understanding of the patho-physiology basis of phenotype-genotype correlations. These ideas will be part of future work.
REFERENCES


