Published in final edited form as:
IEEE Trans Biomed Eng. 2011 December ; 58(12): 3469–3474. doi:10.1109/TBME.2011.2169256.

Integrative, Multi-modal Analysis of Glioblastoma Using TCGA Molecular Data, Pathology Images and Clinical Outcomes

Jun Kong,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Lee A.D. Cooper,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Fusheng Wang,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

David A. Gutman,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Jingjing Gao,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Candace Chisolm,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Ashish Sharma,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Tony Pan,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Erwin G. Van Meir,
Department of Neurosurgery and Hematology and Medical Oncology, School of Medicine and Winship Cancer Institute, Emory University, Atlanta, GA 30322, USA

Tahsin M. Kurc,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Carlos S. Moreno,
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Joel H. Saltz, and
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Daniel J. Brat
Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA

Jun Kong: jun.kong@emory.edu; Lee A.D. Cooper: lee.cooper@emory.edu; Fusheng Wang: fusheng.wang@emory.edu; David A. Gutman: dgutman@emory.edu; Jingjing Gao: jgao@emory.edu; Candace Chisolm: cchisol@emory.edu; Ashish Sharma: ashish.sharma@emory.edu; Tony Pan: tony.pan@emory.edu; Erwin G. Van Meir: evanmei@emory.edu; Tahsin M. Kurc: tkurc@emory.edu; Carlos S. Moreno: cmoreno@emory.edu; Joel H. Saltz: jhsaltz@emory.edu; Daniel J. Brat: dbrat@emory.edu

Abstract

Multi-modal, multi-scale data synthesis is becoming increasingly critical for successful translational biomedical research. In this paper, we present a large-scale investigative initiative on

Correspondence to: Jun Kong, jun.kong@emory.edu.

glioblastoma, a high-grade brain tumor, with complementary data types using *in silico* approaches. We integrate and analyze data from The Cancer Genome Atlas Project on glioblastoma that includes novel nuclear phenotypic data derived from microscopic slides, genotypic signatures described by transcriptional class and genetic alterations, and clinical outcomes defined by response to therapy and patient survival. Our preliminary results demonstrate numerous clinically and biologically significant correlations across multiple data types, revealing the power of *in silico* multi-modal data integration for cancer research.

Index Terms

Glioblastoma; multi-modal data process; *in silico*; cluster analysis; translational integration

I. Introduction

With rapid technological advances in acquiring data from diverse platforms in cancer research, numerous large scale datasets have become available, providing high resolution views and multi-faceted descriptions of biological systems. Such efforts include those in brain tumor research by The Cancer Genome Atlas (TCGA) [1], and the Repository of Molecular Brain Neoplasia Data (REMBRANDT) [2], which have collected large volumes of multi-modal data from complementary platforms on patients with diffuse glioma. As manual processing of this large-scale data is both error-prone and intractably time-consuming, recent investigations have either primarily focused on *in silico* experiments that interrogate these datasets or use them to generate or corroborate hypotheses.

In the In Silico Brain Tumor Research Center (ISBTRC), one of the six National Cancer Institute (NCI) funded In Silico Research Centers of Excellence[1], we explore novel approaches and develop tools for integrative multi-scale, multi-modal data analysis of diffuse gliomas. Our current research has focused on potential relations across tumor genomic and gene expression profiles, complex nuclear morphometric features, neuro-imaging, and clinical outcomes. By conducting complementary, multi-scale *in silico* experiments, we aim not only to improve prognostic capabilities, but also to develop a better understanding of biological underpinnings that drive the rapid progression of these devastating diseases [3].

As a first step towards seamless data integration for improved diagnosis and stratification, we describe our methodology for correlating nuclear morphometric features derived from digitized microscopic images of glioblastoma with 1) genetic alterations, 2) transcriptional subtypes, and 3) treatment response and patient survival.

We hypothesize that digitized pathology images contain rich and as yet untapped biological information trapped in morphologic features that can be resolved by image analysis to provide correlations with genetic alterations and patient prognosis. In this paper, we present results correlating computer-generated nuclear morphometry from large-scale microscopic images with survival, treatment response, and clinically relevant molecular characterizations. The results demonstrate the potential of multi-modal data integration within the setting of large-scale *in silico* research.

[1]https://wiki.nci.nih.gov/display/ISCRE

II. Data set and analysis infrastructure

The overall framework for data analysis, management, integration, and computation infrastructure is illustrated in Fig. 1, where nuclear morphometric features from microscopic images, molecular signatures, clinical outcomes, and neuroimaging annotations from the same cohort of patients are stored in a database for large-scale multi-modal data query, integration, and analysis.

A. Microscopy Imaging Data

Digital microscopy is rapidly emerging as a tool for establishing pathologic diagnosis, evaluating treatment efficacies and performing morphologic research. In distinction to traditional visual review of histological sections, which introduces human bias and remains largely qualitative [4], a computer-based analysis of virtual microscopic slides can be systematic, objective, efficient, and complete [5][6][7]. Moreover, many features in a microscopic image can be identified and analyzed by computer algorithms but not by human observers. Thus, imaging data from histologic slides contains rich phenotypic information that can potentially be exploited to yield clinically meaningful results.

In our research, we have used the microscopic images from TCGA project on glioblastomas (GBMs), which are WHO grade IV astrocytic neoplasms that are rapidly progressive and ultimately fatal. All digitized slides are Haematoxylin and Eosin (H&E) stained permanent sections that were formalin-fixed and paraffin-embedded. In aggregate, 428 whole slides associated with 162 patients are included. All were scanned at 20x magnification with a high-resolution, high-throughput digitized scanner. The overall storage size of the complete image data set for study is about 175Gbytes with JPEG compression ratio of 5.11. The image resolution is up to 63922 × 45753 pixels.

B. “Omics” Data

Phenotypic data derived from digitized images was correlated with TCGA molecular data, providing insight to underlying biological mechanisms and potentially uncovering therapeutic targets within a morphologic class. Each TCGA sample was characterized by multiple molecular platforms including gene (mRNA) and microRNA expression, DNA copy number variation, DNA sequence and DNA methylation.

A recent study of TCGA GBMs defined four transcriptional subtypes: proneural (PN), neural (NR), classical (CL) and mesenchymal (MS) [8]. Each subtype is defined by a characteristic gene expression profile and genetic alterations, including mutations and chromosomal changes (amplification/deletion). For our study, transcriptional subtypes were either obtained from the supplementary information in an earlier work [8] or determined with Prediction Analysis of Microarray (PAM) software version 2.21 using RMA normalized Affymetrix HT-HGU133 mRNA expression platform data. A sample expression average was computed for samples with multiple corresponding arrays. Unlogged expression was filtered to remove probes with a fold change less than 1.5 or an expression range less than 20.

Somatic mutations and chromosome alterations (amplification or deletion) for genes *CDKN2A*, *EGFR*, *IDH1*, *NF1*, *PDGFRA*, *TP53*, and *PTEN*, have been provided by the Memorial Sloan-Kettering Cancer Center (MSKCC)[2]. Mutational status from 205 samples was available. Copy number variation data from the same set consists of a consensus derived from a combination of platforms (Agilent, Affymetrix SNP 6, Illumina) together with

[2]Memorial Sloan-Kettering Cancer Center, http://www.mskcc.org/mskcc/, last access in Mar, 2011

methods (RAE[9], GISTIC[10], GTS[11]) for identifying regions of genomic aberration likely to drive cancer pathogenesis. Copy number alterations are represented by homozygous deletion, hemizygous deletion, neutral change, gain, and high-level amplification.

C. Clinical Outcomes

Clinical data on patient age, chemo- and radiotherapies, and survival was downloaded from the TCGA portal[3].

D. Computational Infrastructure

High resolution digitized pathology images are extremely large, with some occupying several gigabytes even in compressed form. The TCGA dataset include hundreds of pathology images and presents a significant computational challenge for analysis. To expedite processing, we partitioned each whole slide image into non-overlapping regions of 4096 × 4096 pixels to permit parallel analysis. This choice balances between memory requirements and the loss of microanatomy due to tiling. Larger regions have physical memory constraints. Smaller ones place a greater fraction of nuclei on region boundaries resulting in their loss during analysis. To scale up the analysis component of the architecture, we process images with a large-scale, high-performance computation infrastructure where a cluster of computer nodes executes jobs simultaneously. This configuration currently consists of seven Dell 1950 1U rack mount units. Each unit is configured with Dual Xeon E5420 CPUs running with four cores at 2.5Ghz for a total of eight cores per node.

E. Pathology Image Data Representation and Management

Digital microscopy images contain a tremendous array of micro-anatomic structures, which collectively characterize specimens phenotypically. In a study with hundreds or thousands of high-resolution images, millions of nuclear morphometric features need to be represented and curated in a systematic manner such that they can be efficiently queried for correlative investigations. In addition, image analysis using either multiple algorithms or multiple parameter sets can further increase the size of data to be recorded. As a result, information models are needed to organize and represent virtual slide-related image, annotation, mark-up and feature information. To address these challenges, we developed the Pathology Analytical and Imaging Standards (PAIS) model to support flexible, efficient, and semantically enabled data representations for pathology image analysis and characterization[4]. We also implemented a relational database realization of PAIS using IBM DB2 Enterprise Edition 9.7.3 with its spatial extender. The current database runs on PowerEdge T410 Linux server with four quadcore CPUs, 16GB memory, and a 7200 rpm hard drive.

PAIS makes it possible to represent and share data generated from pathology images. More importantly, it is useful tool for scientific discovery through its powerful query support, including those that are metadata-based, spatially based or semantically based [12][13]. Further, we incorporate related molecular data and clinical information into PAIS database to provide integrative queries.

[3]TCGA portal, http://cancergenome.nih.gov/, last access in Dec 2010
[4]PAIS wiki: - https://web.cci.emory.edu/confluence/display/PAIS/

III. Integrated multi-modal data analysis

We next present our methodologies for high throughput microscopic image analysis and multi-modal data integration.

A. Microscopy Imaging Analysis

We developed a suite of image analysis tools for segmenting and characterizing nuclei. To reliably identify nuclei, we applied the fast hybrid grayscale reconstruction algorithm to images for normalizing background regions degraded by artifacts arising from tissue preparation and scanning [18]. This operation substantially separates the foreground from the normalized background and allows recognition of nuclei by simple thresholding. Overlapped nuclei were subsequently separated with the watershed method.

We then extracted a complementary set of features for each identified nucleus to obtain phenotypic signatures of GBMs. These features fall under four primary headings: nuclear morphometry, region texture, intensity and gradient statistics, as summarized in Fig. 2 (a) [14]. Since specific nuclear features have traditionally been used to distinguish types of gliomas, morphometric features (such as the degree of elongation, and size) are included. Nuclear texture information is captured by multiple descriptors, as it varies across nuclei due to the content and clumping of chromatin. Features relevant to nuclear intensity and intensity gradient are included as well. All nuclear features are computed with the grayscale image channel converted from the original color image. Additionally, we applied the same set of texture and gradient features to “cytoplasm” regions surrounding nuclei. Since the true cellular borders of glioma cells cannot be resolved on H&E stained images, cytoplasm refers to a fixed-distance radius surrounding a nucleus. In practice, we dilated the nuclear regions with an eight-pixel margin to identify this space. Fig. 2 (b) presents a small image region where glioma nuclei and cytoplasm regions are depicted. Features derived from cytoplasm are computed with the grayscale image channel as well as the isolated channels for H&E stain signals separated by a color deconvolution algorithm [15]. As the cytoplasm space is obtained by dilating the nuclear regions, its morphologic features are not calculated. Cytoplasm features are then combined with nuclear features for better representation. In aggregate, 74 features extracted from nuclei and proximal cytoplasm describe the morphology and texture characteristics of each nucleus and its neighboring area.

All nuclear and cytoplasmic features associated with a GBM were then summarized into a single vector to represent each patient. To this end, we calculated the first moment of each feature and the second moments of all possible pairs of features [16]. The first moment represents the average value for a specific feature, whereas the second order statistics define relationships between features regarding 1) nuclear morphology, 2) nuclear morphology and nuclear staining, or 3) nuclear morphology/staining and cytoplasmic staining for each patient. The summarization step produces an $N(N+3)/2$-dimensional feature vector to represent the morphology of each patient in a high dimensional space, where N is equal to 74 in our case. Thus, each patient is represented by a 2849-dimensional imaging signature vector derived by aggregating the features of machine-identified nuclei in the associated microscopic whole-slide images.

This is followed by a consensus clustering procedure to compute the probabilities that signatures of pairwise cases are grouped in the same cluster over 100 independent trials of K-means experiments. This analysis is aimed at uncovering the existence of intrinsic morphological clusters defined by nuclear feature signatures. We set the number of clusters as $K=3$.

B. PAIS Query Support

Segmentation results and features are stored in the PAIS database. To correlate micro-anatomic morphometry with molecular profiles and clinical outcome, summary statistics on image features need to be computed for each patient. This process involves calculating the mean feature vectors and the feature covariance values of all possible feature pairs over all nuclei in images of each patient. The PAIS database is queried to search for feature pairs and retrieve corresponding feature values. The summary statistics for each image are combined in a separate program to create a single-feature vector for a patient. Queries for the mean, standard deviation, and covariance of feature calculations are supported through IBM DB2 Structured Query Language (SQL) queries with DB2's built-in aggregation functions: the AVG, STDDEV, and COVARIANCE functions, respectively. An example of PAIS database query for the mean and covariance of three morphometry features, i.e. area, perimeter, and eccentricity, is shown in Fig. 3 where *calculation_flat*, and *patient* are two tables storing nuclear morphometry features and patient-slide relationships; *pais_uid* is the primary key that joins these two tables.

With the efficient and expressive database query support on morphological signature computation, we are able to correlate nuclear morphometry with clinical outcomes and molecular characterizations and to produce results suggesting a possible relationship across nuclear morphometry, patient survival, and molecular data.

C. Multi-modal Data Correlation

Two methods are used for multi-modal data correlation. The first uses consensus cluster labels to partition patients into three groups and correlate nuclear morphometry signatures with response to treatment and patient survival. This analysis potentially reveals the clinical significance suggested by nuclear morphometry features. The second analysis investigates the relationship of consensus clusters with gene expression subtypes and genetic alterations. The hypergeometric distribution is used to calculate the probability of either a given expression subtype or genetic alteration group being enriched/depleted in a given consensus cluster. This analysis allows us to find those expression subtypes and genetic alteration groups significantly enriched or depleted in a cluster, suggesting a possible relationship between the phenotypic and genomic data of GBMs [16]. We present the hypergeometric probability density function f(x|T,S,K) as in Eq. (1) for x samples of a tumor subtype/genetic alteration group in a consensus cluster with K samples when S out of T samples are expected:

$$f(x|T,S,K)=\begin{pmatrix} K \\ x \end{pmatrix}\begin{pmatrix} T-K \\ S-x \end{pmatrix}/\begin{pmatrix} T \\ S \end{pmatrix} \tag{1}$$

where T is the total population size; S is the number of samples in a given tumor type/genetic alteration group; K is the size of the samples in a given consensus cluster; and x is the number of samples of a given tumor type/genetic alteration group in the given consensus cluster containing K samples. The resulting over- and under-representation p-values can be computed as:

$$O(X,T,S,K)=\sum_{i=X}^{K} f(i|T,S,K), \qquad U(X,T,S,K)=\sum_{i=0}^{X} f(i|T,S,K) \tag{2}$$

where X is the observed number of samples of a given tumor type/genetic alteration group within the given consensus cluster containing K samples.

IV. EXPERIMENTAL RESULTS

With the consensus clustering process, we grouped patients based on the patient-level nuclear morphometry signatures into three clusters, consisting of 70, 10, and 82 patients, respectively. More than 22 million neoplastic nuclei in 428 whole slides from 162 patients were analyzed with the aforementioned image-processing pipeline. We excluded nuclei crossing the tile borders from further analysis, as the number of such nuclei is so small when compared with the enormous number of nuclei completely contained by partitioned regions.

A. Response to Therapy and Survival Analysis

Summary nuclear feature vectors are computed with slides grouped by patients. These patients are further grouped into three consensus clusters based on nuclear morphometry. In Table 1, we present the p-values of the Log-Rank test [17] comparing patient survival to the three consensus clusters. The Log-Rank test between cluster two and three yields statistically significant difference in survival, with longer survivals for patients in cluster three. Additionally, the Kaplan-Meier plot for the three clusters of patients is shown in Fig. 4, where Area Under Curve (AUC) for cluster two (AUC = 296.06) is much smaller than that for clusters one (AUC = 2441.81), and three (AUC = 1302.79). This suggests that patients in cluster two have worse prognosis than those in cluster one and three, although this observation needs to be further validated with a larger number of samples. In Fig. 5, we present the Kaplan-Meier plots of three clusters of patients showing surivals of those treated with either standard and aggressive therapy. The resulting p-values of the Log-Rank tests with patient survivals with regard to response to therapy from cluster one, two and three are 0.00705, 0.158, and 0.000640, respectively. The results suggest that patients in cluster one and three show significantly favorable response to aggressive therapy compared to standard therapy. Cluster two contains a small number of patients and conclusions regarding response to therapy are limited.

B. Correlation with Phenotypic and Genotypic Data

We also investigated whether any of the morphometric clusters was characterized either by a specific gene expression subtype or genetic alteration. We therefore studied the enrichment/depletion relationship between phenotypic and genotypic data. After computing the p-values for over- and under-representations with tumor subtypes in the three consensus clusters, we find that mesenchymal samples are enriched in cluster one with an over-representation p-value of 0.0372. In Fig. 6, the genetic alteration profiles of samples in three clusters are presented for genes of interest for GBMs. With copy number variations, we observe that cluster one is enriched with *EGFR* amplification (p-value 0.0211) and *CDKN2A* deletion samples (p-value 0.00586). Cluster two is enrched with *PTEN* deletion samples with p-value of 0.0244. Additionally, *CDKN2A* deletion samples are depleted in cluster three with p-value of 0.00958. However, no specific mutations are found to be significantly correlated with the nuclear morphometry clusters.

V. Conclusions

In this letter, we present a large-scale multimodal data correlation study of GBM. Morphological characteristics derived from whole-slide microscopic images are correlated with clinical and molecular data. Results from these analyses revealed a significant survival difference between GBM patients based on the nuclear morphometry cluster of their tumor. This observation suggests a potential for predicting patient outcome based on nuclear morphometry. Our results also suggest that patients within specific nuclear morphometry clusters demonstrate differential therapeutic responses, as the patients in clusters 1 and 3 showed favorable response to aggressive therapy. In a future work, we plan to investigate

morphometric features that are most predictive of molecular subtype and clinical behavior. These phenotypic features could then be incorporated into clinical diagnostics.

Acknowledgments

This work was supported by the NCI Contract HHSN261200800001E; TCGA Contract 29 x 55193; NIH 5R01LM009239-04; NHLBI R24 HL085343; and by the Clinical and Translational Science Awards program under PHS Grant UL1RR025008.

References

1. TCGA Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. Sep.2008 455:1061–1068. [PubMed: 18772890]
2. Madhavan S, Zenklusen JC, Kotilarov Y, Sahni H, Fine HA, Buetow K. Rembrandt: Helping personalized medicine become a reality through integrative translational research. Mol Cancer Res. Feb; 2009 2(7):157–167. [PubMed: 19208739]
3. Van Meir EG, Hadjipanayis CG, Norden AD, Shu HK, Wen PY, Olson JJ. Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma. CA Cancer J Clin. 2010; 60:166–193. [PubMed: 20445000]
4. Aldape K, Burger PC, Perry A. Clinicopathologic Aspects of 1p/19q Loss and the Diagnosis of Oligodendroglioma. Archives of Pathology & Laboratory Medicine. 2007; 131(2):242–251. [PubMed: 17284109]
5. Wang W, Ozolek JA, Rohde GK. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. Cytometry. 2010; 77A:485–494. [PubMed: 20099247]
6. Wang W, Ozolek JA, Slepcev D, Lee AB, Chen C, Rohde GK. An optimal transportation approach for nuclear structure-based pathology. IEEE Transactions on Medical Imaging. 2011; 30:621–631. [PubMed: 20977984]
7. Rohde GK, Ribeiro AJS, Dahl KN, Murphy RF. Deformation-based nuclear morphometry: Capturing nuclear shape variation in HeLa cells. Cytometry. 2008; 73A:341–350. [PubMed: 18163487]
8. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17(1):98–110. [PubMed: 20129251]
9. Taylor BS, Barretina J, Socci ND, DeCarolis P, Ladanyi M, Meyerson M, Singer S, Sander C. Functional copy-number alterations in cancer. PLoS ONE. 2008; 3(9):1–16. e3179.
10. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proc Nat Acad Sci. 2007; 104(50):20007–20012. [PubMed: 18077431]
11. Wiedemeyer R, Brennan C, Heffernan TP, Xiao Y, Mahoney J, Protopopov A, Zheng H, Bignell G, Furnari F, Cavenee WK, Hahn WC, Ichimura K, Collins VP, Chu GC, Stratton MR, Ligon KL, Futreal PA, Chin L. Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. Cancer Cell. 2008; 13:355–364. [PubMed: 18394558]
12. Foran DJ, Yang L, Chen W, Hu J, Goodell LA, Reiss M, Wang F, Kurc T, Pan T, Sharma A, Saltz JH. ImageMiner: a software system for comparative analysis of tissuemicroarrays using content-based image retrieval, high-performance computing, and grid technology. J Amer Med Infom Assoc. 2011; 18(4):403–415.

13. Wang, F. Developing Data Model Standard for Pathology Analytical Imaging. Presentation at HL7 Anatomic Pathology Work Group Meeting; 2011.
14. Cooper LA, Kong J, Gutman DA, Wang F, Cholleti SR, Pan TC, Widener PM, Sharma A, Mikkelsen T, Flanders AE, Rubin DL, Van Meir EG, Kurc TM, Moreno CS, Brat DJ, Saltz JH. An integrative approach for in silico glioma research. IEEE Trans Biomed Eng. Oct; 2010 57(10): 2617–2621. [PubMed: 20656651]
15. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Ana Quant Cyt and Hist. 2001; 23(4):291–9.
16. Cooper, L.; Kong, J.; Wang, FS.; Kurc, T.; Moreno, C.; Brat, D.; Saltz, J. Morphological Signatures and Genomic Correlates in Glioblastoma. International Symposium on Biomedical Imaging; 2011. p. 1624-1627.
17. Harrington, D. Encyclopedia of Biostatistics. Hoboken, NJ: Wiley Interscience; 2005. Linear rank tests in survival analysis.
18. Vincent L. Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms. IEEE Transactions on Image Processing. 1993; 2(2):176–201. [PubMed: 18296207]

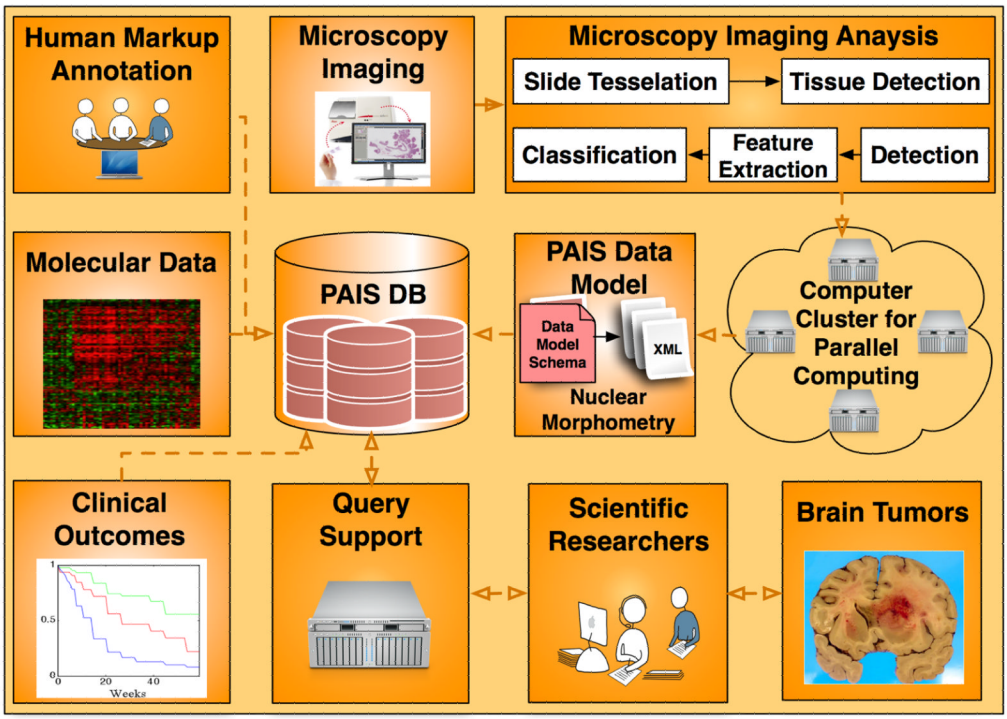
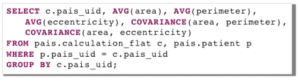


Fig. 1.

The overall architecture for the *In Silico* brain tumor research is presented. All data sources, i.e. microscopy imaging features, molecular data, and clinical outcomes are captured by PAIS database with which scientific researchers can query for multi-data integration results.

Fig. 2.
(a) Features computed from nuclear and “cytoplasm” regions are summarized; (b) A typical image region is overlaid with nuclear (red) and “cytoplasm” (green) boundaries identified by computer algorithms.



```
SELECT c.pais_uid, AVG(area), AVG(perimeter),
    AVG(eccentricity), COVARIANCE(area, perimeter),
    COVARIANCE(area, eccentricity)
FROM pais.calculation_flat c, pais.patient p
WHERE p.pais_uid = c.pais_uid
GROUP BY c.pais_uid;
```

Fig. 3.

A SQL example is presented to query for mean and covariance nuclear feature vector for each patient in PAIS database.

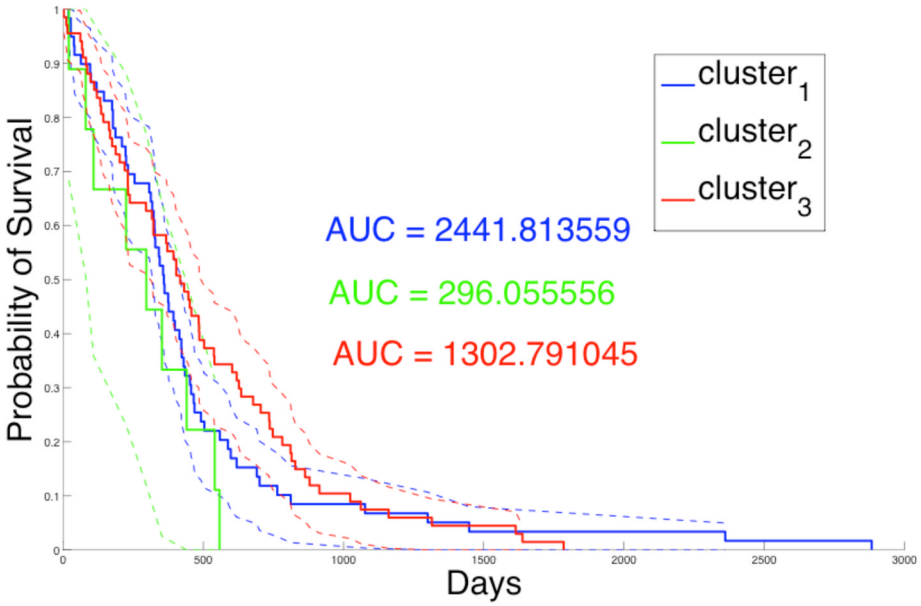


Fig. 4.

Kaplan-Meier plot for patients of three consensus clusters is presented with AUC values and 95% lower and upper confidence bounds.

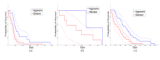


Fig. 5.
Kaplan-Meier plots for patients treated with aggressive (blue) and standard (red) therapy from (a) cluster 1, (b) cluster 2, and (c) cluster 3, are presented with 95% lower and upper confidence bounds.

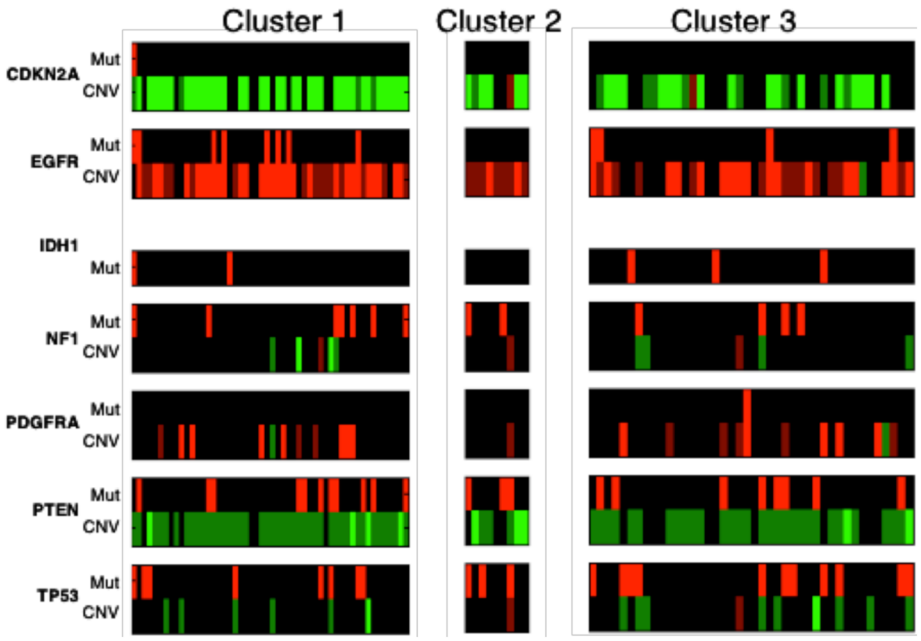


Fig. 6.

The genetic alteration profiles for three consensus clusters of samples are presented. For copy number data (denoted as CNV), light green, dark green, black, dark red, and light red represent homozygous deletion, hemizygous deletion, neutral change, gain, and high-level amplification. For somatic mutation data (denoted as Mut), red represents mutant.

TABLE I

We present P-values from log-rank test with survival data from patients of different nuclear morphometry clusters

Consensus Cluster	Consensus Cluster(s)	P-value of Log-rank Test
1	(2, 3)	0.322
2	(1, 3)	0.0719
3	(1, 2)	0.131
1	2	0.156
1	3	0.222
2	3	**0.0437**