Integrated morphologic analysis for the identification and characterization of disease subtypes

Lee A D Cooper,[1] Jun Kong,[1] David A Gutman,[1,2] Fusheng Wang,[1,2] Jingjing Gao,[1] Christina Appin,[3] Sharath Cholleti,[1] Tony Pan,[1] Ashish Sharma,[1,2] Lisa Scarpace,[4] Tom Mikkelsen,[4,5] Tahsin Kurc,[1,2] Carlos S Moreno,[1,2,3,6] Daniel J Brat,[1,2,3,6] Joel H Saltz[1,2]

▸ Additional materials are published online only. To view these files please visit the journal online (http://jamia.bmj.com/content/19/2.toc).

[1]Center for Comprehensive Informatics, Emory University, Atlanta, Georgia, USA
[2]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia, USA
[3]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, Georgia, USA
[4]Department of Neurology, Henry Ford Hospital, Detroit, Michigan, USA
[5]Department of Neurosurgery, Henry Ford Hospital, Detroit, Michigan, USA
[6]Winship Cancer Institute, Atlanta, Georgia, USA

Correspondence to
Dr Lee AD Cooper, Center for Comprehensive Informatics, Emory University, Psychology Building, Suite 566, 36 Eagle Row, Atlanta, GA 30322, USA; lee.cooper@emory.edu

Received 16 November 2011
Accepted 26 December 2011
Published Online First
24 January 2012

ABSTRACT

Background and objective Morphologic variations of disease are often linked to underlying molecular events and patient outcome, suggesting that quantitative morphometric analysis may provide further insight into disease mechanisms. In this paper a methodology for the subclassification of disease is developed using image analysis techniques. Morphologic signatures that represent patient-specific tumor morphology are derived from the analysis of hundreds of millions of cells in digitized whole slide images. Clustering these signatures aggregates tumors into groups with cohesive morphologic characteristics. This methodology is demonstrated with an analysis of glioblastoma, using data from The Cancer Genome Atlas to identify a prognostically significant morphology-driven subclassification, in which clusters are correlated with transcriptional, genetic, and epigenetic events.

Materials and methods Methodology was applied to 162 glioblastomas from The Cancer Genome Atlas to identify morphology-driven clusters and their clinical and molecular correlates. Signatures of patient-specific tumor morphology were generated from analysis of 200 million cells in 462 whole slide images. Morphology-driven clusters were interrogated for associations with patient outcome, response to therapy, molecular classifications, and genetic alterations. An additional layer of deep, genome-wide analysis identified characteristic transcriptional, epigenetic, and copy number variation events.

Results and discussion Analysis of glioblastoma identified three prognostically significant patient clusters (median survival 15.3, 10.7, and 13.0 months, log rank $p=1.4e-3$). Clustering results were validated in a separate dataset. Clusters were characterized by molecular events in nuclear compartment signaling including developmental and cell cycle checkpoint pathways. This analysis demonstrates the potential of high-throughput morphometrics for the subclassification of disease, establishing an approach that complements genomics.

INTRODUCTION

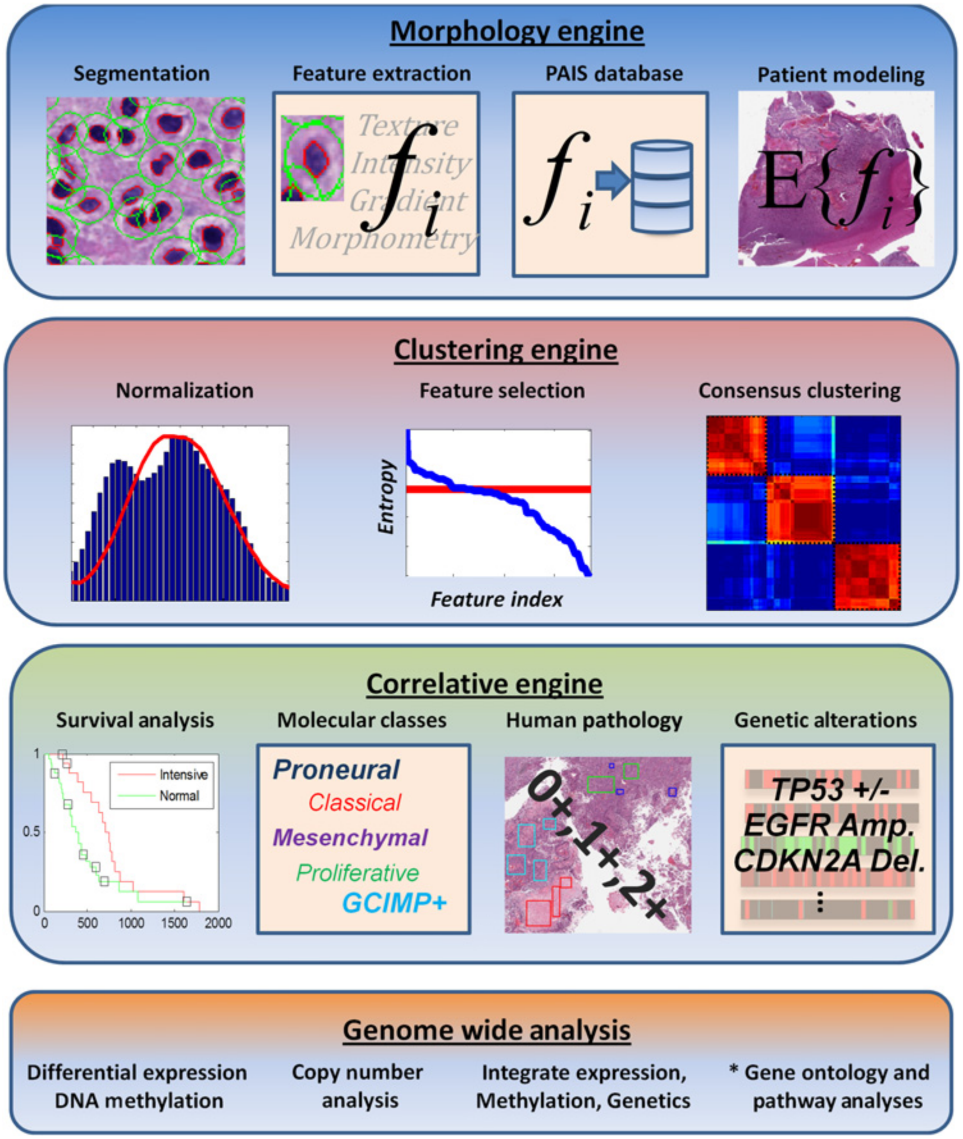
Variations in disease morphology are often linked to underlying molecular events and patient outcomes, suggesting that quantitative morphometric analysis may provide further insight into disease mechanisms. Advances in computing and imaging devices now put large-scale morphometric analysis within reach; whole slide imaging devices are capable of rapidly producing detailed digital representations of an entire glass slide at high magnification, and image analysis algorithms now automate tasks ranging from antibody scoring to the segmentation and classifications of cells.

There is currently intense interest in identifying subclassifications of disease with the aim of developing improved personalized therapies that target class-specific mechanisms. Several large-scale disease characterization efforts are underway, all genomic in nature, and most focused on cancers. The Cancer Genome Atlas (TCGA) is the largest, targeting more than 20 cancers for comprehensive genomic characterization. Several efforts have resulted in subclassification of human malignancies, one example being the identification of a molecular classification of glioblastoma (GBM) into four classes defined by gene expression: the proneural, neural, classical, and mesenchymal.[1] These classes have associations with specific genetic alterations, patient outcome, and response to therapy, and have been the subject of considerable study.

Although most tumor classification efforts have focused on molecular data, TCGA is also producing large collections of whole slide images from tissues submitted for molecular analysis. The linkage of pathology images with genomics presents a unique opportunity to study morphology in the context of genetics and patient outcome. For example, GBM, the most common form of primary brain tumor in adults, exhibits tremendous variations in the morphologies of both nuclear and cytoplasmic morphology. What is the underlying molecular basis for this heterogeneity? Is there a spectrum of morphologies or are there discrete morphologic groups? How does morphology associate with survival or response to therapy? High-throughput morphometric analysis has the potential to address these questions.

In this paper we propose a methodology for high-throughput characterization of morphology to identify morphologic subtypes of disease. Signatures of patient-specific tumor morphology are generated by quantitative analysis of hundreds of millions of cells in whole slide images, and are clustered to reveal morphologically cohesive groups. Associations with patient outcome and treatment response as well as genetic associations are analyzed within groups to provide a complete picture of morphology-driven classes. This methodology consists of four layered components (see figure 1).

Figure 1 The integrative morphologic/genomic analysis framework consists of four modules. The morphology engine produces and manages quantitative descriptions of hundreds of millions of cells. The clustering engine normalizes and filters data and identifies morphology-driven patient clusters. The correlative module analyzes clusters for associations with survival, treatment response, human evaluations of pathology, and recognized genetic alterations. Genome-wide analysis performs a deeper investigation of the transcriptional, genetic, and epigenetic associations, and mines these for biological themes and pathway activation. *Gene Ontology and Pathway Analyses are performed offline with separate commercial and public software packages.



Morphometric analysis

This component segments individual nuclei and calculates a quantitative description of each nucleus using a collection of features that describe nuclear shape, and the texture and contrast of nuclear and cytoplasmic staining. The derived segmentation boundaries and cellular features are delivered to a database implementing the Pathological and Analytic Imaging Standards.[2] This database provides query capabilities for the recall and analysis of the hundreds of millions of cells represented in each dataset.

Morphometry-driven patient clustering

The clustering component generates and clusters patient-level summaries of morphology by generating statistical models of the millions of cells found in digitized tissues. This component performs the necessary normalizations required for statistical modeling and clustering, and provides feature selection capability to eliminate redundant, non-informative features to improve clustering performance.

Correlative analysis

The correlative component analyzes morphology-driven patient clusters for significant associations with outcome and well recognized genetic alterations. Associations between morphology clusters and published molecular classes are also examined. A variety of statistical tests are employed, including the log rank test, Kaplan–Meier estimation, and the hypergeometric test for enrichment and depletion.

Genome-wide analysis

The genome-wide analysis component goes beyond well recognized genetic alterations to perform a deeper analysis that identifies significant transcriptional, epigenetic, and copy number events. Transcriptional events are mapped to copy number and epigenetic events to determine causation of transcriptional differences. Lists of significant genes are further analyzed to identify enrichment of concepts defined by gene ontology (GO) and to determine which signaling pathways are enriched in the clusters. Significant gene lists are also filtered to identify recognized cancer-related genes for abbreviated reporting.

We demonstrate this methodology using an analysis of GBM data from TCGA to identify three prognostically significant morphologic clusters. Correlative and genome wide analysis shows that these clusters are characterized by nuclear compartment signaling networks related to development, DNA transcription regulation, and cell cycle (CC) checkpoint/DNA repair. This analysis demonstrates the power of morphologic analysis and the potential of high throughput morphometrics as a platform that complements genomics in ongoing disease characterization efforts.

BACKGROUND

Microscopy image analysis

A comprehensive set of techniques has been developed to address fundamental problems in pathology image analysis over the last decade. Algorithms are available to automate tasks from the segmentation of cells and classification of multicellular regions to the automated grading of entire images. An in-depth review of these methods has been produced by Gurcan *et al*.[3] We limit the scope here to focus on complex systems that incorporate multiple components. These systems tend to focus on computer aided evaluation of diagnostic criteria for detection and grading of disease.

A computer based grading system for neuroblastoma that distinguishes grades of differentiation and stromal development was developed by Gurcan *et al*.[4–9] This system uses a texture based analysis to classify stroma and degree of cellular differentiation to support diagnosis. Other CAD techniques have also been developed for lymphoma,[10–13] and breast[14] and prostate cancers.[14–16] For prostate cancer, Madabhushi *et al* developed an innovative system to predict recurrence risk following prostatectomy using the fusion of image derived and proteomics data.[14] Recently, Beck *et al* published an analysis of image derived breast epithelium and stromal features to generate a system that assigns an image-based prognostic score independent of clinical, molecular, and pathologic factors.[17]

Molecular subtypes of glioblastoma

Genomic analyses have illustrated molecular heterogeneity in many cancers, including GBM, resulting in robust molecular classification.[1 18–21] An analysis of TCGA GBM gene expression data by Verhaak *et al* identified four classes, named for the genes that compose each class's signature: proneural, neural, classical, and mesenchymal.[1] While clustering was driven by gene expression patterns, subclasses had strong associations with frequent mutations and copy number alterations. These same gene expression classifications have also been observed in lower grade gliomas.[22]

Subtypes defined by epigenetic criteria have also been identified for GBM. An analysis of DNA methylation in TCGA samples identified a CpG island methylator phenotype (GCIMP) of GBM that is associated almost exclusively with proneural tumors and secondary GBMs with somatic mutations of IDH1.[23] The non-GCIMP proneural patients have significantly worse outcome than GCIMP + proneural subjects.

Morphologic analysis of TCGA data

Morphologic analysis of TCGA data has been limited when compared to the overall scope of TCGA. A texture-based classification of GBM image regions into normal, necrotic, apoptotic, and tumor was performed using combined color and texture features.[24] We have previously published an investigation of GBM tumor morphology that defined a morphology-driven clustering of patients using patient-level morphologic signatures.[25] This analysis investigated both the top-down power of morphologic signatures to predict the molecular classifications of Verhaak *et al*, and associations between molecular classifications and a patient grouping defined by a bottom-up clustering analysis of morphologic signatures. In this paper we extend the preliminary work in bottom-up clustering analysis to present an end-to-end pipeline for cluster analysis that examines the survival, molecular, and pathologic associations of morphologically defined clusters. In this work we demonstrate the existence of prognostically significant clusters in glioblastoma, and validate their existence in a separate dataset. We show that these clusters do not recapitulate previously identified molecular subtypes, and are not explained by traditional morphologic analysis performed by human pathologists. We also illustrate the gene expression, genetic, and epigenetic associations of these clusters, showing that activities in cancer-related pathways significantly distinguish each morphology cluster.

MATERIALS AND METHODS

TCGA glioblastoma dataset

The proposed methodology was tested using data from the TCGA GBM project. Digitized formalin fixed paraffin-embedded H&E slides were obtained from the TCGA portal.[26] Slides were manually curated to remove images that were poorly focused, or that contain digitization or preparation artifacts. Survival, chemotherapy, and radiotherapy data were also obtained from the TCGA portal. Survival was taken as 'days to death' for non-right-censored patients, and 'days to last follow-up' for right-censored patients. Intensive therapy was defined as three or more cycles of chemotherapy, or concurrent radiation and chemotherapy, as in Verhaak *et al*.[1]

Transcriptional class labels for the Verhaak and Phillips classifications were obtained from the TCGA Advanced Working Group.[21] The updated Verhaak labeling extends the original labeled set presented in Verhaak *et al*[1] by using the originally labeled samples along with Affymetrix HT_HG-U133A data to label previously unclassified samples. The original Phillips classifications were derived from different samples and a similar but different platform (Affymetrix HG-U133A), and so this scheme was translated to the TCGA samples and platform. Samples with a negative silhouette width were discarded from analysis. Methylation phenotype CIMP status was calculated using level 2 Illumina OMA002 and OMA003 cancer panels as described previously.[23] Genetic alterations of recognized genes were obtained from the Sloan Kettering CBIO Cancer Genetics portal (http://cbio.mskcc.org/cancergenomics/).

TCGA consortium neuropathologists evaluated 112 GBMs for the presence of 18 pathologic criteria including necrosis, microvascular hyperplasia, or inflammatory cell infiltration. Three pathologists rated each case as absent (0+), present (1+), or abundant (2+) for each criterion, with a single pathologist serving as the adjudicator. Ratings were obtained to analyze associations between human-derived pathologic criteria and machine based morphologic clustering, serving as a valuable resource for corroborating computer-based results.

Genome-wide analysis of molecular data was performed on level 2 normalized data including the Affymetrix HT_HG-U133A platform, the Agilent Human miRNA 8x15K platform, and the Illumina OMA002 and OMA003 cancer panels. Patient-reduced profiles were calculated for each platform when multiple patient arrays were encountered.

An independent set of GBM slides were obtained from the Henry Ford health system to validate morphology clustering and associations with patient outcome. Slides were similarly curated to remove images unsuitable for analysis.

Segmentation and feature extraction

The first stage of image analysis identifies the boundaries of cell nuclei and extracts features to describe nuclear and cytoplasmic morphology. Color images were first thresholded to identify and remove blood spills. Remaining areas were processed with a fast hybrid algorithm for grayscale reconstruction. This stage permits local discrimination between whole-nuclei to be retained for analysis and faint out-of-plane nuclei and other non-nuclear background hematoxylin stain.[27] Tightly packed clumps of nuclei were then separated using watershed segmentation. A region of high-confidence cytoplasm was identified for each nucleus in the absence of a membrane marker by dilating the nuclear boundary by eight pixels.

Each nucleus/cytoplasmic region is then described by a set of 74 features representing shape and staining characteristics. A complete list of these features and their definitions is available in Cooper *et al*.[28] The features describing nuclei are taken from four broad categories: morphology, intensity, texture, and gradient. Morphology features describe nuclear shape including size and boundary irregularity. Intensity, texture, and gradient features describe the distributions and spatial patterns of intensity values to reflect differences in staining. The corresponding cytoplasmic region is similarly described, however morphology is not calculated since cytoplasmic region boundary is strictly derived from the nucleus. The cytoplasmic region is first decomposed into separate hematoxylin and eosin signals using a color deconvolution algorithm.[29] Features are extracted separately for both the hematoxylin and eosin signals as depicted in Cooper *et al*.[30]

Clustering morphology signatures

The clustering module uses measurements of typical nuclei to self-aggregate patients into morphology-driven clusters. For each patient, the average cell appearance is calculated over all cells to define a 74-dimensional morphologic signature. These signatures are then quantile normalized across patients to provide comparability between the different scales of mixed feature types.[31] The normalized signature features are then subjected to an entropy-scoring feature selection that eliminates redundant features in order to identify a core set that accounts for most of the variations across patients.[32] Clustering of signatures was performed using the consensus clustering method to robustly identify structure over many clustering realizations.[33] Multiple clustering hypotheses were considered, from two clusters up to seven. Maximization of total silhouette area was used as feedback to select clustering outcome that best describes the structure of the dataset.

Correlative and genome-wide analyses

The correlative analysis module evaluates the morphology clusters for associations with patient outcome and treatment response, related molecular classifications, recognized genetic alterations, and human-derived evaluations of pathological criteria. Differences in outcome and treatment response were analyzed using the log rank test to compare the outcome of each cluster against all others, and differences in outcome within each cluster for patients receiving intensive treatment.[34] Associations between morphology clusters and transcriptional subtypes, genetic alterations, and human pathology data were evaluated using the hypergeometric distribution to determine if a property appears more or less often than by chance within each morphology cluster, as described in the online appendix.

RESULTS AND DISCUSSION

An analysis of 200 million nuclei associated with TCGA GBMs revealed three prognostically significant morphology-driven clusters. For clarity, we name these clusters after the biological functions of cluster-associated genes: the CC cluster, the chromatin modification (CM) cluster, and the protein biosynthesis (PB) cluster. Outcome and genetic associations of clusters are presented in table 1. A complete listing of patients and their characteristics is presented in online appendix table 1, cluster characteristics and correlates in online appendix tables 2 and 3–6.

A separate analysis was also carried out on the validation tissue to confirm existence of these morphology clusters and their associations with survival.

Analysis of glioblastoma identifies clusters of patient morphology

Means-based signatures were calculated to represent average cell appearance for each patient. The top 75% of informative features were selected using entropy-contribution ranking, and the signatures were quantile normalized. Consensus clustering was applied to the normalized signatures, varying the number of clusters from two to seven, with silhouette area used as a clustering quality measure. The three-cluster analysis produced the best separation, with the maximum silhouette area of 48.6 (online appendix figure 1). The structure formed by these clusters is visible in the heatmap of figure 2.

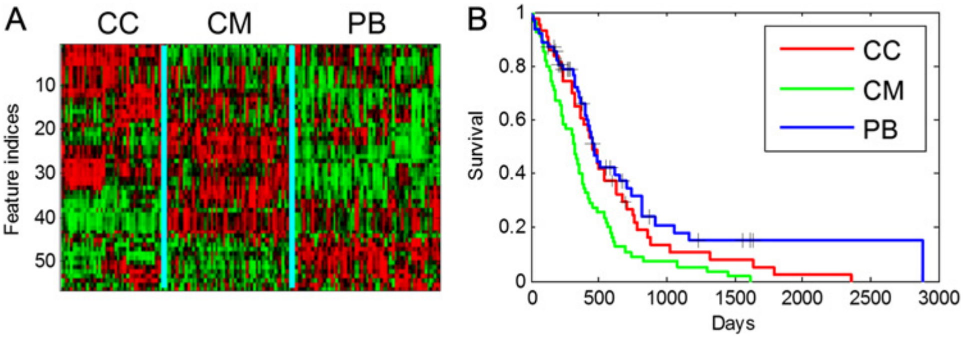
Representative nuclei from each patient are presented in figure 3, selected as the segmented cells that most resemble each tumor's morphologic signature. Nuclei in the CC cluster

Table 1 Survival, genomic, and pathology correlates of morphology clusters

	CC cluster	CM cluster	PB cluster
Prognosis	Average	Poor	Better
Subtype associations	Neural depleted	Neural enriched Proneural depleted	GCIMP depleted
Pathology	Small cells enriched	Lymphocytes enriched	Inflammation depleted
Genetics	*NF1* mutant depleted *TP53* mutant depleted	None	None
Pathways	*TP53* signaling *Wnt* signaling *β-catenin*	*Wnt* signaling *β-catenin*	*NF-κB* signaling *ATM* checkpoint
Differential Expression	2740/663 genes up/down 97/100 miRNAs up/down	200/463 genes up/down 121/81 miRNAs up/down	0/188 genes up/down 15/5 miRNAs up/down
Differential methylation	69 Genes hypermethylated	244 Genes hypermethylated	45 Genes hypomethylated
Copy number	1068 deletions 38 amplifications	301 deletions 5 amplifications	399 deletions 7 amplifications
Expression mapping	23 mapped to methylated sites 595 mapped to CNV sites	8 mapped to methylated sites 27 mapped to CNV sites	1 mapped to methylated sites 19 mapped to CNV sites

CC, cell cycle; CM, chromatin modification; CNV, copy number variation; GCIMP, CpG island methylator phenotype; PB, protein biosynthesis.

Figure 2 Glioblastoma (GBM) clusters, survival, and relationship to molecular subtypes. (A) Means-based analysis of GBM morphology reveals three patient clusters. (B) Survival differences between these clusters are statistically significant. CC, cell cycle; CM, chromatin modification; PB, protein biosynthesis.



appeared to be the most hyperchromatic and slightly larger than in the CM and PB clusters; the cytoplasm of these cells was the most basophilic. The CM cluster had the smallest and least intensely staining nuclei, and also the most eosinophilic cytoplasm. The nuclear and cytoplasmic staining and texture of the PB cluster were intermediate. Trends in nuclear shape were not obvious to the human eye. Analysis of the most distinguishing feature types confirms these observations. The 56 selected signature features were ranked for their power to separate the clusters in online appendix table 2, where the t-statistic was used to rank features for each class. We observe that the top 10 features for the PB and CC clusters all describe cytoplasmic texture and color. In the CM cluster, 8 of 10 features also describe cytoplasm, but nuclear area and perimeter also appear.

Morphology clusters are prognostically significant

Figure 1B shows the Kaplan–Meier estimation of survival of TCGA GBM patients, organized cluster membership. The PB cluster has the most favorable overall clinical outcome, followed by the CC and CM clusters, and contains a subset of patients with prolonged survival. Log rank tests show that the CM ($p=4.5e-4$) and PB ($p=1.0e-2$) clusters have significantly different outcomes compared with other clusters, but that the CC cluster is indistinguishable ($p=0.54$). Each cluster was also analyzed to determine the effectiveness of intensive therapy (see online appendix figure 2). Intensively treated patients have significantly better outcome in the CM cluster ($p=1.4e-2$), but not in the CC and PB clusters. A Cox proportional hazard model incorporating both morphology cluster labels and Verhaak subtypes further shows that the morphology cluster label is a significant predictor of survival ($p=5.0e-3$) where the Verhaak subtype is not ($p=0.58$) (SAS PHREG V.9.2).

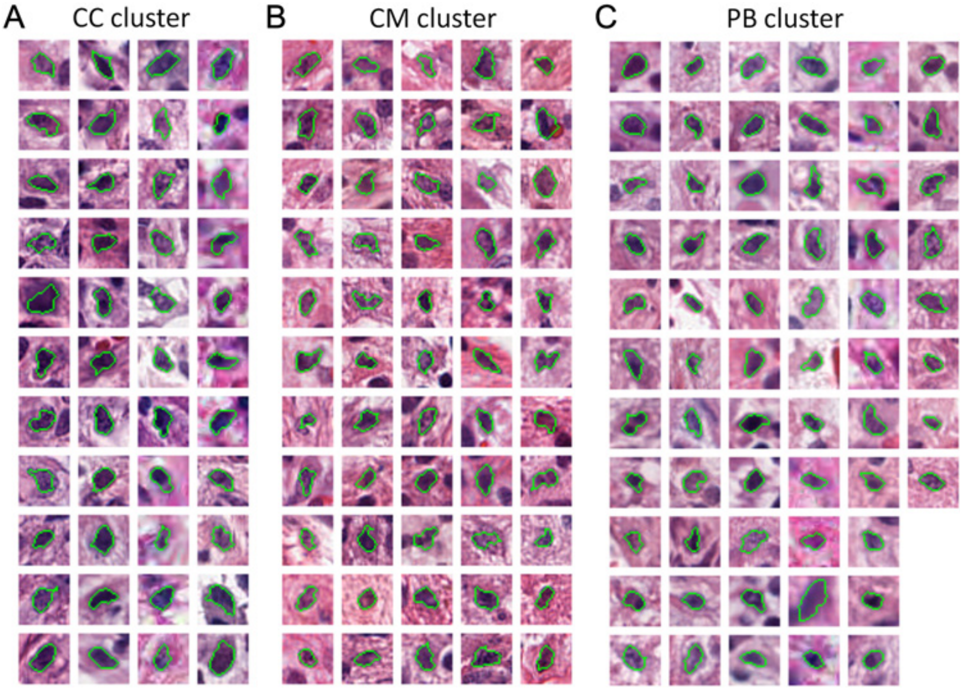
Validation

We performed a de novo clustering on 57 million nuclei from 84 GBM samples within the Henry Ford dataset, using the same set of selected features from the TCGA analysis. Using the In Group Proportion method (ClusterRepro, version 1.1) we validated the CC ($p=7.2e-3$) and CM ($p=0.013$) clusters in the Henry Ford dataset.[35] The survival trends for these clusters remain consistent with observations from the TCGA dataset (online appendix figure 3). The PB cluster was not observed in the Henry Ford dataset ($p=0.63$).

Morphology clusters are not strongly associated with traditional neuropathology classification or recognized genetic alterations

We did not observe significant associations between the morphology-driven clusters and the classification by expert

Figure 3 Signature nuclei for: (A) cell cycle (CC), (B) chromatin modification (CM), and (C) protein biosynthesis (PB) clusters. Cluster morphology is visualized by selecting the most representative nucleus from each patient. The selection is defined by the cell with the shortest distance in feature space to the patient's morphology signature.



neuropathologists based on tradition morphologic features. Categorical neuropathologist ratings were analyzed for associations with morphology clusters. Using a threshold of abundant or greater, no abundant inflammation was observed within any case in the PB cluster (hypergeometric p=4.7e-2); there was a slight enrichment of small cells in the CC cluster (p=1.3e-2), and a slight enrichment of lymphocytes in the CM cluster (p=1.3e-2).

We also did not observe a strong association between the morphology clusters and the gene expression subtypes of Verhaak *et al*. Neural samples are conspicuously absent from the CC cluster (p=4.7e-3) and are consequently enriched within the CM cluster (p=1.7e-3). The proneural subtype is also depleted within the CM cluster (p=4.7e-3), and the proneural-GCIMP phenotype is absent from the PB cluster (p=4.0e-2). Morphology clusters were also analyzed for associations with mutations and copy number variations recognized in the GBM literature. The CC cluster lacks any *NF1* mutations, despite being composed of 34% mesenchymal samples (hypergeometric p=7.8e-3). *TP53* mutations are also slightly underrepresented in the CC cluster (p=2.0e-2).

Genome-wide analysis

Genome-wide analysis of gene expression, DNA methylation, and copy number variation identified characteristic genes for each morphology cluster. These gene sets were analyzed by GO and pathway analysis tools to identify cancer-related pathways and biological functions enriched within each cluster. Detailed descriptions of these events are presented in online appendix tables 3–6. Pathways identified by ingenuity pathway analysis are shown in online appendix figure 5.

Analysis of the genes differentially expressed across the morphology clusters using the DAVID database determined that the most significant annotation for each cluster was in fact nuclear lumen localization cellular component (GO:0031981, Benjamini FDR=1.08E-15, 2.8E-36, 2.17E-19 for clusters PB, CC, and CM respectively). Comparative analysis of gene ontology enrichment using DAVID analysis indicated that genes involved in RNA splicing, and transcriptional regulation were enriched in all three clusters, where genes associated with PB, CC, and CM were differentially regulated across clusters. Genes involved in DNA damage and repair were enriched in the CC and PB clusters (online appendix table 6). These data support our hypothesis that quantitative analysis of nuclear morphology can detect subtle differences in gene expression. Moreover, when the sets of differentially expressed genes were subjected to ingenuity pathway analysis, several cancer-related pathways were differentially enriched among the morphology clusters, including the *ATM* and *TP53* DNA damage checkpoints, the *NFκB* pathway, and the *Wnt* signaling and *PTEN-AKT* pathways. A full description of differences is presented in the online appendix.

Limitations and future directions

The tumor microenvironment contains a complex mixture of cell types, each having a specific role in sustaining the tumor and promoting growth. Developing enhanced models of patient morphology to represent these heterogeneous cell populations is the focus of our work moving forward. Capturing this heterogeneity within mixture-type models will permit finer morphologic distinctions between patients and also improve interpretability and reveal the specific cellular populations associated with prognosis and molecular events. The robust estimation and comparison of multi-modal models is a challenging problem for large high-dimensional datasets. Approaches to estimation are largely iterative and require many realizations to achieve satisfactory convergence and to avoid poorly fitted models that correspond to suboptimal local minima. Comparison of estimated models also requires intensive calculations that span high dimensional spaces and methods that robustly establish the correspondences of modes across multiple patients.

We are already using this methodology to study other tumor types included in TCGA. Each tissue presents unique challenges for segmentation and feature extraction, but the framework presented here represents a generalizable approach to performing correlative studies. We also anticipate that this approach can naturally extend to study diseases other than cancer. As genomics becomes more affordable, and whole slide imaging devices more common, we expect that efforts similar to TCGA will describe other diseases and produce datasets that link pathology, genomics, and patient outcome.

CONCLUSION

This paper presents a methodology for identifying morphologic subtypes of disease and an end-to-end framework for the analysis of survival and genomic subtype correlates. We demonstrate our methodology with an analysis of TCGA GBM data, showing that tumors self-aggregate into three prognostically significant clusters, characterized by variations in pathology and genetics, and the activation of cancer-related pathways in the nuclear compartment. While the proposed system is demonstrated on GBM, these techniques can be applied to any of the more than 20 tumor types now included in TCGA. We are currently translating our system to study lung adenocarcinomas, a disease that is also rich with morphologic variation.

The ability to quantitatively characterize disease at multiple biological scales has the potential to improve personalization of preventive strategies and treatments. Advances in pathology imaging devices and computing have added quantitative morphology to the collection of high-throughput technologies available to describe genetics, biological function, and now structure. These high-resolution, high-throughput capabilities are being employed not only in research, but also increasingly in healthcare settings. We predict that advances in information technology will distill volumes of multidimensional imaging and genomic data into the information that will drive discovery and development of novel mechanisms for preventing and treating disease.

Molecular tests and human interpretations of anatomic pathology are currently used to guide diagnosis and treatment. Researchers in the image analysis community have demonstrated that in some cases algorithms can reproduce the process of pattern recognition used to produce render diagnoses. The results presented here take this process beyond recapitulation of established diagnoses, and represent what we think will become an increasingly common practice where computational methods are used to define previously unrecognized categories of disease.

Funding This work was supported in part by PHS Grant UL 1RR025008 from the Clinical and Translational Science Award Program, National Institutes of Health, Grant numbers R01LM009239 and R01LM011119 from the National Library of Medicine, Contract No. HHSN261200800001E from the National Cancer Institute, National Institutes of Health, and the Georgia Cancer Coalition.

Competing interests None.

Ethics approval Ethics approval was provided by the Institutional Review Board.

Contributors All authors listed are justifiably credited with authorship. In detail: LADC, conception, design, analysis, and interpretation of data, drafting of manuscript. JK, design, and analysis of imaging data and methods. DAG, acquisition and interpretation of imaging data. FW, design of imaging database. JG, Cox hazard analysis and interpretation. CA, interpretation of pathology data. SC, analysis of imaging data. TCP, design of imaging database. AS, design of imaging database. LS and TM, provision of validation data. TK, conception and manuscript editing. CSM, conception, analysis, and interpretation of molecular data, drafting of manuscript. DJB, conception, design, interpretation, drafting of manuscript. JHS, conception, design, and interpretation, drafting of manuscript.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data used in this study which is not presented in the supplementary tables is available from the authors on request.

REFERENCES

1. **Verhaak RG,** Hoadley KA, Purdom E, *et al*; Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 2010;**17**:98—110.
2. **Wang F,** Kong J, Cooper L, *et al*. A data model and database for high-resolution pathology analytical image informatics. *J Pathol Inform* 2011;**2**:32.
3. **Gurcan MN,** Boucheron L, Can A, *et al*. Histopathological image analysis: a review. *IEEE Rev Biomed Eng* 2009;**2**:147—71.
4. **Gurcan MN,** Kong J, Sertel O, *et al*. Computerized pathological image analysis for neuroblastoma prognosis. *AMIA Annu Symp Proc* 2007:304—8.
5. **Gurcan MN,** Pan T, Shimada H, *et al*. Image analysis for neuroblastoma classification: segmentation of cell nuclei. *Conf Proc IEEE Eng Med Biol Soc* 2006;**1**:4844—7.
6. **Kong J,** Sertel O, Boyer KL, *et al*. Computer-assisted grading of neuroblastic differentiation. *Arch Pathol Lab Med* 2008;**132**:903—4; author reply 904.
7. **Kong J,** Sertel O, Shimada H, *et al*, eds. *A Multi-resolution Image Analysis System for Computer-assisted Grading of Neuroblastoma Differentiation*. California, San Diego: SPIE Medical Imaging, 2008.
8. **Kong J,** Sertel O, Shimada H, *et al*. Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation. *Pattern Recognit* 2009;**42**:1080—92.
9. **Sertel O,** Kong J, Shimada H, *et al*. Computer-aided prognosis of neuroblastoma on whole-slide images: classification of stromal development. *Pattern Recognit* 2009;**42**:1093—103.
10. **Belkacem-Boussaid K,** Pennell M, Lozanski G, *et al*. Computer-aided classification of centroblast cells in follicular lymphoma. *Anal Quant Cytol Histol* 2010;**32**:254—60.
11. **Cooper L,** Sertel O, Kong J, *et al*. Feature-based registration of histopathology images with different stains: an application for computerized follicular lymphoma prognosis. *Comput Methods Programs Biomed* 2009;**96**:182—92.
12. **Samsi S,** Lozanski G, Shana'ah A, *et al*. Detection of follicles from IHC-stained slides of follicular lymphoma using iterative watershed. *IEEE Trans Biomed Eng* 2010;**57**:2609—12.
13. **Sertel O,** Lozanski G, Shana'ah A, *et al*. Computer-aided detection of centroblasts for follicular lymphoma grading using adaptive likelihood-based cell segmentation. *IEEE Trans Biomed Eng* 2010;**57**:2613—16.
14. **Madabhushi A,** Agner S, Basavanhally A, *et al*. Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Comput Med Imaging Graph* 2011;**35**:506—14.
15. **Khurd P,** Bahlmann C, Maday P, *et al*, eds. *Computer-aided Gleason Grading of Prostate Cancer Histopathological Images Using Texton Forests. 2010 IEEE International Symposium on*. Rotterdam, the Netherlands: Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2010.
16. **Tabesh A,** Teverovskiy M, Pang HY, *et al*. Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans Med Imaging* 2007;**26**:1366—78.
17. **Beck AH,** Sangoi AR, Leung S, *et al*. Systematic analysis of breast cancer morphology Uncovers stromal features associated with survival. *Sci Transl Med* 2011;**3**:108ra13.
18. **Cancer Genome Atlas Research Network.** Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**:609—15.
19. **Virtaneva K,** Wright FA, Tanner SM, *et al*. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci U S A* 2001;**98**:1124—9.
20. **Phillips HS,** Kharbanda S, Chen R, *et al*. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 2006;**9**:157—73.
21. **Huse JT,** Phillips HS, Brennan CW. Molecular subclassification of diffuse gliomas: seeing order in the chaos. *Glia* 2011;**59**:1190—9.
22. **Cooper LA,** Gutman DA, Long Q, *et al*. The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas. *PLoS One* 2010;**5**:e12548.
23. **Noushmehr H,** Weisenberger DJ, Diefes K, *et al*; Cancer Genome Atlas Research Network. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 2010;**17**:510—22.
24. **Han J,** Chang H, Loss L, *et al*, eds. *Comparison of Sparse Coding and Kernel Methods for Histopathological Classification of Glioblastoma Multiforme. International Symposium on Biomedical Imaging*. Chicago: IEEE, 2011.
25. **Cooper LAD,** Kong J, Wang F, *et al*. *Morphological Signatures and Genomic Correlates in Glioblastoma. International Symposium on Biomedical Engineering*. Chicago: IEEE, 2011.
26. **NCI.** *The Cancer Genome Atlas*. 2011. http://cancergenome.nih.gov/ (accessed 1 Sep 2011).
27. **Vincent L.** Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Trans Image Process* 1993;**2**:176—201.
28. **Cooper LA,** Kong J, Gutman DA, *et al*. An integrative approach for in silico glioma research. *IEEE Trans Biomed Eng* 2010;**57**:2617—21.
29. **Ruifrok AC,** Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol Histol* 2001;**23**:291—9.
30. **Cooper LAD,** Jun K, Fusheng W, *et al*, eds. *Morphological Signatures and Genomic Correlates in Glioblastoma. 2011 IEEE International Symposium on*. Chicago, Illinois, USA: Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2011.
31. **Bolstad BM,** Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**:185—93.
32. **Varshavsky R,** Gottlieb A, Linial M, *et al*. Novel unsupervised feature filtering of biological data. *Bioinformatics* 2006;**22**:e507—13.
33. **Monti S,** Tamayo P, Mesirov J, *et al*. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* 2003;**52**:91—118.
34. **Mantel N.** Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 1966;**50**:163—70.
35. **Kapp AV,** Tibshirani R. Are clusters found in one dataset present in another dataset? *Biostatistics* 2007;**8**:9—31.