Published in final edited form as:
IEEE Trans Biomed Eng. 2010 October ; 57(10): 2617–2621. doi:10.1109/TBME.2010.2060338.

An Integrative Approach for In Silico Glioma Research

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

Jun Kong,
Emory University, Atlanta, GA 30322 USA

David A. Gutman,
Emory University, Atlanta, GA 30322 USA

Fusheng Wang,
Emory University, Atlanta, GA 30322 USA

Sharath R. Cholleti,
Emory University, Atlanta, GA 30322 USA

Tony C. Pan,
Emory University, Atlanta, GA 30322 USA

Patrick M. Widener,
Emory University, Atlanta, GA 30322 USA

Ashish Sharma,
Emory University, Atlanta, GA 30322 USA

Tom Mikkelsen,
Department of Neurology and Neurosurgery, Henry Ford Hospital, Detroit, MI 48202 USA

Adam E. Flanders,
Department of Radiology, Thomas Jefferson University, Philadelphia, PA 19107 USA

Daniel L. Rubin,
Stanford University Medical Center, Stanford, CA 94305 USA

Erwin G. Van Meir,
Emory University, Atlanta, GA 30322 USA

Tahsin M. Kurc,
Emory University, Atlanta, GA 30322 USA

Carlos S. Moreno,
Emory University, Atlanta, GA 30322 USA

Daniel J. Brat, and
Emory University, Atlanta, GA 30322 USA

Joel H. Saltz
Emory University, Atlanta, GA 30322 USA

Correspondence to: Lee A. D. Cooper, lee.cooper@emory.edu.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Lee A. D. Cooper: lee.cooper@emory.edu; Jun Kong: jun.kong@emory.edu; David A. Gutman: dgutman@emory.edu; Fusheng Wang: fusheng.wang@emory.edu; Sharath R. Cholleti: sharath.cholleti@emory.edu; Tony C. Pan: tony.pan@emory.edu; Patrick M. Widener: patrick.widener@emory.edu; Ashish Sharma: ashish.sharma@emory.edu; Tom Mikkelsen: nstom@neuro.hfh.edu; Adam E. Flanders: adam.flanders@jefferson.edu; Daniel L. Rubin: rubin@med.stanford.edu; Erwin G. Van Meir: evanmei@emory.edu; Tahsin M. Kurc: tkurc@emory.edu; Carlos S. Moreno: cmoreno@emory.edu; Daniel J. Brat: dbrat@emory.edu; Joel H. Saltz: jhsaltz@emory.edu

Abstract

The integration of imaging and genomic data is critical to forming a better understanding of disease. Large public datasets, such as The Cancer Genome Atlas, present a unique opportunity to integrate these complementary data types for *in silico* scientific research. In this letter, we focus on the aspect of pathology image analysis and illustrate the challenges associated with analyzing and integrating large-scale image datasets with molecular characterizations. We present an example study of diffuse glioma brain tumors, where the morphometric analysis of 81 million nuclei is integrated with clinically relevant transcriptomic and genomic characterizations of glioblastoma tumors. The preliminary results demonstrate the potential of combining morphometric and molecular characterizations for *in silico* research.

Index Terms

Biology; brain tumor; image analysis; *in silico*; microscopy

I. Introduction

The integration of imaging and genomic data is critical to develop a deeper understanding of disease. Projects like The Cancer Genome Atlas1 (TCGA) [1] and the Repository for Molecular Brain Neoplasia (REMBRANDT) [2] are producing extensive multidimensional datasets containing high-resolution pathology imagery, magnetic resonance imaging, and an array of molecular data for the characterization of diseases. These datasets present a unique opportunity to conduct *in silico* scientific research, where image analysis and informatics can converge to shed light on complex biological phenomena. In a project funded by the National Cancer Institute In Silico Research Centers of Excellence program,2 we are conducting an integrative *in silico* study of diffuse glioma brain tumors that leverages clinical, molecular, radiology, and pathology imaging data. Our goals in this project are to achieve a finer granularity in the subtyping of glioma tumors that is predictive of outcome and response to treatment, and to study the mechanisms of progression from low- to high-grade tumors.

This letter focuses on the particular aspect of *pathology image analysis* and the integration of morphometry with clinical and molecular characterizations. Digitized pathology images contain a wealth of information on tissue and microanatomical morphology, and in many cases, these morphologies reflect underlying genetic alterations that are predictive of patient prognosis and response to treatment. Computerized image analysis provides a means for extensive morphometric analysis of microanatomy in large-scale datasets [3], [4]. In this letter, we describe our methodology for morphometric analysis of nuclei in large-scale datasets of diffuse glioma brain tumors, and present preliminary results correlating nuclear morphometry with clinically relevant molecular characterizations. These preliminary results demonstrate the potential of *in silico* research combining morphological analyses with clinical and molecular data.

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

II. Challenges in Microanatomy Characterization

The diffuse gliomas are a broad category of brain tumors that include the astrocytomas, oligodendrogliomas, and oligoastrocytomas [5]. Histopathologic distinction of these lesions requires morphological discrimination between astrocytic and oligodendroglial cell differentiation. Features of cell nuclei morphology are the primary cue in this distinction [6]. In general, the astrocytomas contain an abundance of nuclei that are elongated, irregularly shaped, and contain visible chromatin clumping, resulting in a rough interior texture. In contrast, nuclei in oligodendrogliomas tend to appear smaller, round, and have relatively uniform interior characteristics. Between the endpoints of pure oligodendroglioma and pure astrocytoma tumors, there exists a spectrum of lesions that exhibit mixtures of morphological qualities, as depicted in Fig. 1. Due in part to the qualitative nature of pathological evaluation, the overlap in morphologies significantly confounds diagnosis, resulting in large interob-server variabilities [7]. A limited set of molecular tests is available to aid in diagnosis [5], based on characteristic chromosome deletions, somatic mutations, and gene expression, however, the large majority of morphologically mixed tumors lack definitive genetic markers.

The analysis of large pathology image datasets for *in silico* exploration presents several challenges. Data size, image heterogeneity, validation of algorithms, and management of results are the primary impediments for mining morphological information from large-scale multimodal datasets.

A. Image Size

High-resolution scans of digital slides produce extremely large images, typically with tens of thousands of pixels in each dimension. Typical studies like TCGA may include hundreds of patients, each with multiple associated slides.

B. Heterogeneity

Large collections of tissues spanning multiple diagnoses and individuals inevitably exhibit significant heterogeneity. Variations in slide preparation, scanning, and natural variations between individuals influence the colors, textures, and densities of structures of interest. A fundamental challenge for large-scale *in silico* studies is to develop algorithms that are robust to these variations.

C. Public Datasets

Often features of interest, such as blood vessels can be highlighted using immunohistochemical staining. This option may not be possible when using existing or publicly available datasets that were not designed with image analysis considerations. In the case of TCGA, sections are stained with standard hematoxylin and eosin (H&E), and therefore, structures of interest, such as mitotic cells or blood vessels are not easily distinguishable by stain.

D. Validation

Extensive validation of image analysis algorithms is required to ensure the fidelity of derived scientific conclusions. The analysis of large datasets like TCGA produces morphological information on tens of millions of microanatomical entities, prohibiting even a qualitative, but exhaustive review of results. Acquiring human markup on a sampled subset of results requires careful planning and supporting infrastructure. Sampling must reflect the heterogeneity of tissues to account for regions, where algorithm performance is expected to vary significantly. Additionally, mechanisms must exist for the management and query of algorithm results and the submission of human markup feedback.

III. Methodology

This section presents our methodologies for pathology image analysis for integrative *in silico* study of nuclear morphometry in diffuse gliomas. Our dataset is drawn from TCGA [1]. The digitized slides used in these studies are formalin-fixed paraffin embedded H&E stained sections of tumor resections. Each sample has been characterized with multiple molecular platforms to measure gene expression, micro RNA expression, copy number variation, sequence, and DNA methylation.

A. Nuclear Analysis

We have developed an objective system for the quantification of nuclei in diffuse gliomas that is aimed at characterizing the shape and texture of nuclei in whole-slide images. The system consists of three stages, as presented in Fig. 2.

1) Nuclei Segmentation and Characterization—The first stage in nuclear analysis is the identification and segmentation of nuclei. In an effort to solve issues mostly arising from large variations in image intensity, texture and histological shape, we use a computationally efficient method consisting of standard techniques that accommodates the identification of nuclei with distinct characteristics. Image regions exhibiting either nontissue areas or red blood cells are first excluded from analysis by thresholding color channels. The remaining regions are then converted to grayscale prior to applying morphological reconstruction. The reconstruction denoises background regions by removing artifacts due to nonspecific hematoxylin staining and out-of-plane nuclei. Foreground nuclei are then separated from the background by thresholding the reconstruction result. Overlapped nuclei are then separated with a watershed segmentation.

The second stage captures information on the shape and texture of individual nuclei. A collection of features, selected for its ability to represent the differences in oligodendroglial and astrocytic differentiation, is calculated for each nucleus to form a *nuclear feature vector*. These features are drawn from four categories: morphometry, texture, intensity statistics, and gradient statistics. The feature groups are presented in Table I.

2) Data Management and Query Support—Each whole-slide image contains hundreds of thousands of nuclei. Managing the characterizations of these nuclei analyzed under multiple parameter sets is a significant challenge. To address this problem, we have designed and implemented an object-oriented information model, the *Pathology Analytical Imaging Standards* (PAIS), to store pathology image analysis results. This model consists of 62 classes that collectively store segmentation boundaries, annotations/classifications on segmented regions, derived features, human markup and annotation, and provenance information regarding analysis methods and parameters. The major components of this model are shown in Fig. 3. This model supports aggregation, comparison, and metadata-based queries for validation and query of results. For example, one can search for regions that are segmented by human experts, but are not segmented by a computerized algorithm, or find aggregate overlap of intersections of nuclei between two images analyzed by different algorithms. We are in the process of implementing a validation protocol using PAIS to systematically select subsets of images with varying diagnostic and/or molecular characteristics, obtaining pathology expert reviews, markups, and annotations on these images, and storing and comparing computerized and human analysis results.

IV. Results and Discussion

In this section, we demonstrate an example integrative analysis of pathology imaging and molecular data to correlate nuclear morphometry with molecular characterizations, using

publicly available TCGA data for grade IV glioblastomas (GBMs). The TCGA GBM dataset contains extensive molecular characterizations of GBM tissue including multiple gene-expression platforms, comparative genomic hybridization and single nucleotide polymorphism (SNP) arrays, exon tiling arrays, and sequencing analysis [4]. Digitized formalin-fixed paraffin embedded sections are also provided for the same GBM tumors. These images were used for the purpose of diagnosis, and have a rich set of annotations generated by TCGA consortium neuropatholgists.

We obtained 213 20× magnification whole-slide permanent section scans from the TCGA portal, corresponding to 79 distinct patients. A total of 90 million nuclei were segmented in these images, and nuclear features were calculated for each individual nucleus. Sample nuclei segmentations were visually reviewed by two neuropathologists for quality control. We are currently developing a more extensive validation using PAIS as described earlier.

A. Separation of TCGA Molecular Subtypes

A recent study of TCGA GBM data has defined four clinically relevant subtypes of GBM tumors, namely the proneural, neural, mesenchymal, and classical types [8]. These subtypes vary in their response to treatment, with proneural-type patients experiencing a significant survival advantage. These four subtypes were defined through analysis of gene expression and genomic data, and have been demonstrated to exhibit characteristic patterns of gene expression, somatic mutations, and chromosome alterations. A comparison of molecular-subtype gene signatures with signatures of normal brain cell types suggests a link between tumor subtype and neural cell lineages as well.

To examine the relationship between molecularly defined tumor subtypes and nuclear morphology, the subtypes for the 213 image dataset were obtained from [8] using TCGA sample codes. Of the 213 images, 183 have available molecular-subtype classifications. Among this set, 48 are proneural type, 33 are neural, 61 classical, and 41 mesenchymal.

For each subtype-labeled image, we calculated the mean feature vector and the feature covariance over all nuclei in the image as a summary statistic. These summary statistics were combined into a single-feature vector to represent each image as a point in the summary statistic feature space. We then performed pairwise classifications between the four subtypes using simple linear support vector machines (SVM) to examine the linear separability of the subtypes-based purely on nuclear morpholgy. A linear SVM was chosen both to avoid overfitting and to preserve the feature space structure, as the transformations induced by kernels can complicate biological interpretation of results. A tenfold cross validation with stratified sampling was used to maintain the proportionality of subtypes in training data. The validation was averaged over 1000 trials with randomized folds. The significance of classification accuracy was also examined using a permutation test with 50 000 trials [9]. By randomly permuting the sample labels in each trial, we obtain an estimate of classifier accuracy distribution under the null hypothesis that morphometry and subtype are not associated.

The averaged classification accuracies are presented in Table II. Many subtype pairs are mutually well separated, at 80% or greater classification accuracy. The permutation test results indicate significance bounded by $p \leq (2e–5)$ for all subtype pairs except the classical/mesenchymal. These results suggest a possible link between nuclear morphology and clinically relevant subtypes defined by molecular analysis.

The aim of this integrated analysis is not to develop morphometry-based classifiers of tumor subtypes, but rather to gain insight into the possible underlying biological mechanisms by determining which morphological features best distinguish the subtypes. To further illustrate

this point, we have tested the binary classification power of individual summary statistics for the proneural and classical subtypes (the two subtypes receiving least and most benefit from aggressive therapy, respectively). Treating each statistic independently, a two sample *t*-test was used to calculate *p*-values, which are then sorted to rank prediction power. Table III contains the top five distinguishing statistics from the proneural/classical comparison, all of which are covariance statistics. These covariance statistics have morphological interpretations, for example, larger covariance between axis length and intensity suggests an increased correlation between staining and nuclei size or elongation. Visualizations of the top-ranked summary statistic for the proneural/classical comparison are presented in Fig. 4. Sets of nuclei from multiple images for the proneural and classical subtypes are presented in Fig. 4(b). These nuclei were sampled using a search criterion to identify candidate nuclei, where the product of max intensity and major-axis length falls within a small interval centered at the proneural or classical covariance, respectively.

V. Conclusion

The public datasets produced by large-scale efforts, such as TCGA, provide unique opportunities to integrate complementary data sources and conduct scientific research *in silico*. The pathology images in these datasets contain a wealth of morphological information that can be correlated with genomic characterizations. In this letter, we present our vision for the role of pathology image analysis in integrative *in silico* research and provide a motivating example that correlates nuclear morphometry with clinically relevant molecular GBM tumor subtypes. Our analysis of TCGA GBM data suggests a possible relationship between nuclear morphometry and the established subtypes defined by the analysis of Verhaak *et al.* [8].

In future work, we plan to further investigate the connections between morphometry and molecular characterization in the TCGA and REMBRANDT datasets. Additionally, we are planning a similar investigation of the morphology of blood vessels in angiogenesis within the context of tumor progression.

Acknowledgments

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This work was supported by Federal funds from the National Cancer Institute, National Institutes of Health (NIH) under Contract HHSN261200800001E, Contract 94995NBS23, Contract N01-CO-12400, and Contract 85983CBS43; by TCGA Contract 29X55193; by National Heart, Lung, and Blood Institute under Grant R24HL085343; by NIH under Grant U54 CA113001, Grant R01 CA86335, and Grant R01 CA116804; and NIH Public Health Service under Grant UL1 RR025008, Grant KL2 RR025009, or Grant TL1 RR025010 from the Clinical and Translational Science Awards program of National Center for Research Resources; by National Library of Medicine under Grant R01LM009239; and by Biomedical Information Science and Technology Initiative under Grant P20 EB000591.

References

1. TCGA Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Sep..vol. 455:1061–1068.
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. 2009 Feb.; vol. 2(no. 7):157–167. [PubMed: 19208739]
3. Pincus Z, Theriot JA. Comparison of quantitative methods for cell-shape analysis. J. Microscopy. 2007 Mar.; vol. 227(no. 2):140–156.

4. Boland MV, Markey MK, Murphy RF. Automated recognition of patterns characteristics of subcellular structures in fluorescence microscopy images. Cytometry. 1998; vol. 33:711–720.
5. Brat DJ, Prayson RA, Ryken TC, Olson JJ. Diagnosis of malignant glioma: Role of neuropathology. J. Neuro-Oncol. 2008 Sep.; vol. 3(no. 89):287–311.
6. Gupta M, Djalilvand A, Brat DJ. Clarifying the diffuse gliomas: An update on the morphologic features that discriminate oligodendroglioma from astrocytoma. Amer. J. Clin. Pathol. 2005 Nov.; vol. 5(no. 124):755–768. [PubMed: 16203285]
7. Coons SW, Johnson PC, Scheithauer BW, Yates AJ, Pearl DK. Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. Cancer. 1997 Apr.; vol. 7(no. 79):1381–1393. [PubMed: 9083161]
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. and The 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 Jan.; vol. 17(no. 1):98–110. [PubMed: 20129251]
9. Fisher, RA. The Design of Experiment. New York: Hafner; 1935.

Fig. 1.
Spectrum of nuclear morphologies in glioma tumors varies between the pure morphologies of oligodendroglial and astrocytic nuclei.

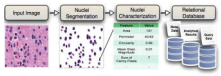

Fig. 2.
Overview of the nuclear analysis workflow is presented. Each nucleus is characterized by a set of feature descriptors that are stored in a relational database for further analysis.

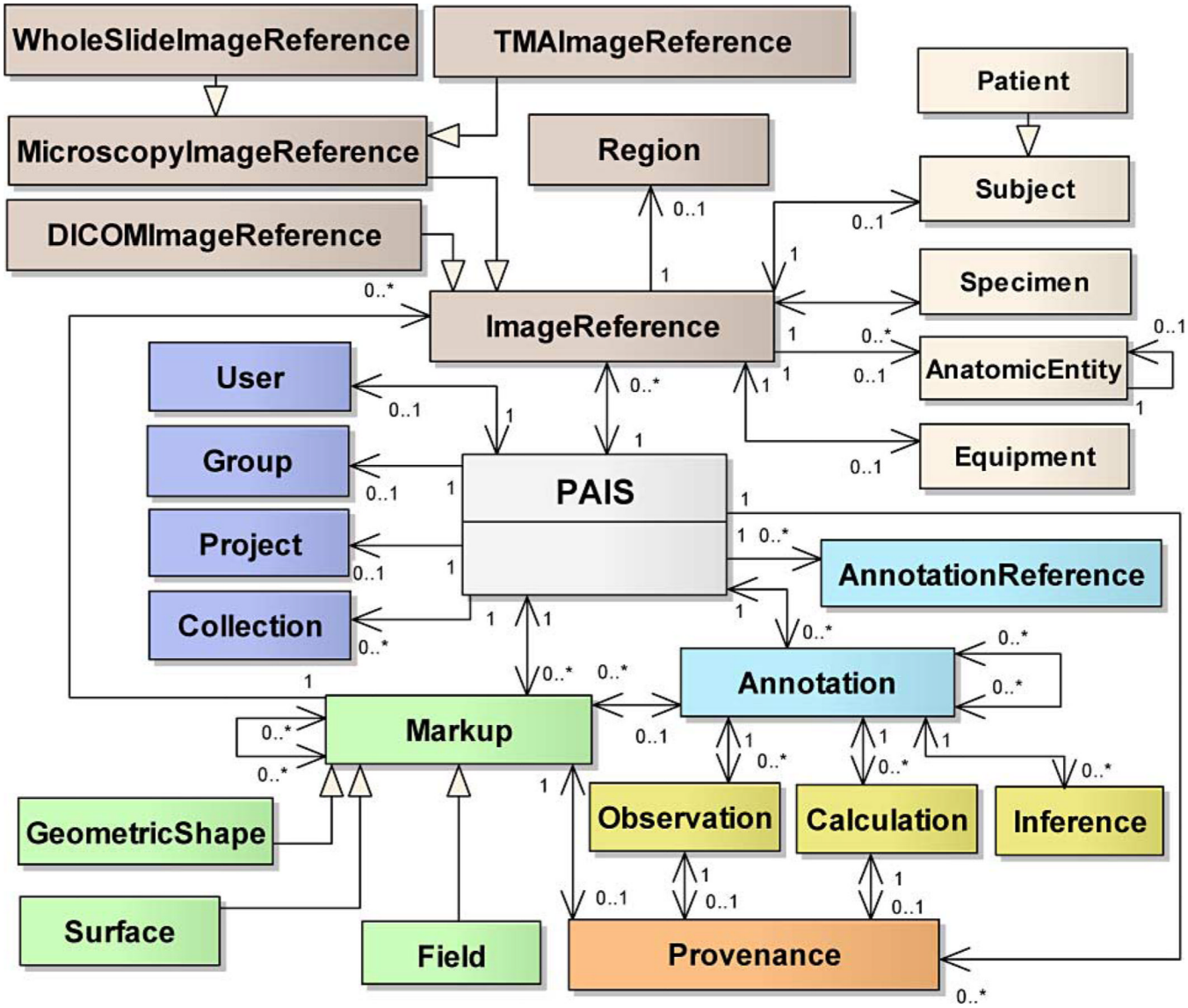

Fig. 3.
Pathology analytical imaging standards schema supports storage and retrieval of human markup and annotation as well as algorithmic results for pathology images. Numbers indicate.

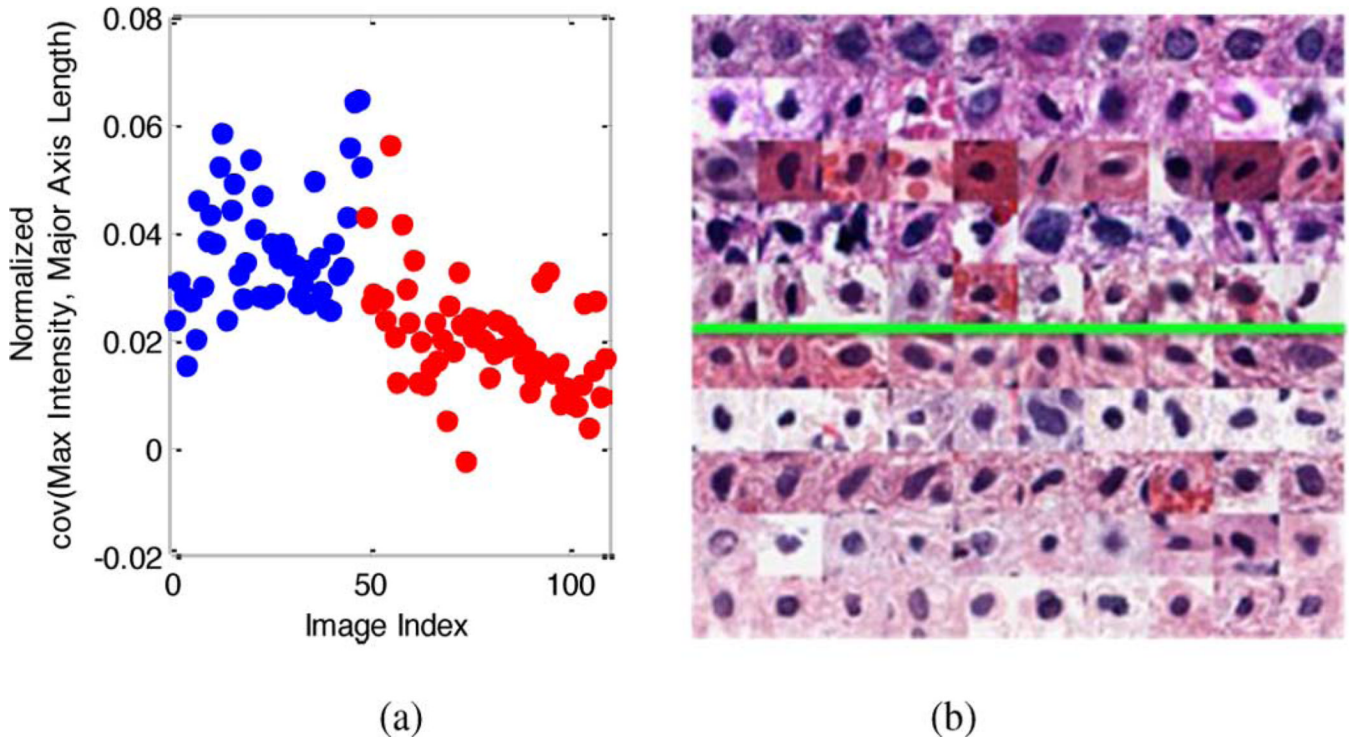

Fig. 4.

Separation of proneural (blue) and classical (red) tumor subtypes. (a) Individual summary statistics indicate potential morphological distinction between proneural and classical tumor nuclei populations. (b) “Nuclei microarray” composed of nuclei from different tumor subtypes aids in interpretation of results. The green line separates nuclei from (top) proneural tumors and (bottom) classical tumors.

TABLE I

Nuclear Features

Category	Features
Morphometry	Area, Perimeter, Eccentricity, Circularity, Major Axis Length, Minor Axis Length, Extent Ratio
Intensity Statistics	Mean Intensity, Max Intensity, Min Intensity, Std. Dev. Intensity
Texture	Entropy, Energy, Skewness, Kurtosis
Gradient Statistics	Mean Grad. Magnitude, Std. Dev. Gradient Magnitude, Entropy Gradient Magnitude, Energy Gradient Magnitude, Skewness Gradient Magnitude, Kurtosis Gradient Magnitude, Sum Canny Pixels, Mean Canny Pixels

Note: Set of 23 features for characterization of nuclei fall into four broad categories.

TABLE II

Classification Accuracy of TCGA Subtypes Using Nuclear Morphometry

	Neural	Classical	Mesenchymal
Proneural	76.3±3.0%	82.0±2.1%	76.6±2.6%
Neural		80.3±2.5%	81.7±2.8%
Classical			70.0±2.5%

Note: Pairwise classification accuracies of tumor subtypes based on nuclear morphometry summary statistics. The indicated intervals are one standard deviation of 1000 trials of tenfold cross validation. The results suggest a possible link between nuclear morphology and the clinically relevant subtypes derived from genomic and transcriptomic analysis.

TABLE III

Feature Ranking for TCGA-Subtype Classifications

Subtype Pair	Top Five Features
Proneural/Classical	covariance(Max Intensity, Major Axis Length)
	covariance(Area, Max Intensity)
	covariance(Perimeter, Max Intensity)
	covariance(Max Intensity, Sum Canny Pixels)
	covariance(Max Intensity, Entropy)

Note: Ranking features according to separation power provides cues on which features best separate and define tumor classes.