Laboratory Investigation (2015) 95, 366–376
© 2015 USCAP, Inc All rights reserved 0023-6837/15 $32.00

PATHOBIOLOGY IN FOCUS

Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images

Lee AD Cooper[1,2,3], Jun Kong[1], David A Gutman[1,3], William D Dunn[4], Michael Nalisnik[1] and Daniel J Brat[1,3,5]

Technological advances in computing, imaging, and genomics have created new opportunities for exploring relationships between histology, molecular events, and clinical outcomes using quantitative methods. Slide scanning devices are now capable of rapidly producing massive digital image archives that capture histological details in high resolution. Commensurate advances in computing and image analysis algorithms enable mining of archives to extract descriptions of histology, ranging from basic human annotations to automatic and precisely quantitative morphometric characterization of hundreds of millions of cells. These imaging capabilities represent a new dimension in tissue-based studies, and when combined with genomic and clinical endpoints, can be used to explore biologic characteristics of the tumor microenvironment and to discover new morphologic biomarkers of genetic alterations and patient outcomes. In this paper, we review developments in quantitative imaging technology and illustrate how image features can be integrated with clinical and genomic data to investigate fundamental problems in cancer. Using motivating examples from the study of glioblastomas (GBMs), we demonstrate how public data from The Cancer Genome Atlas (TCGA) can serve as an open platform to conduct *in silico* tissue-based studies that integrate existing data resources. We show how these approaches can be used to explore the relation of the tumor microenvironment to genomic alterations and gene expression patterns and to define nuclear morphometric features that are predictive of genetic alterations and clinical outcomes. Challenges, limitations, and emerging opportunities in the area of quantitative imaging and integrative analyses are also discussed.

Laboratory Investigation (2015) **95,** 366–376; doi:10.1038/labinvest.2014.153; published online 19 January 2015

Visual information embedded in histologic specimens carries prognostic value and reflects the underlying molecular traits of disease. Human evaluation of histology is a time-honored practice and serves as the basis of modern pathology, yet is highly subjective and known for its inter- and intra-observer variations.[1] Human observers are also limited by scale and the need to reduce information into summary categorical descriptions. Diagnostic evaluation of histologic specimens is often performed over a prescribed number of high-power fields, and reasonable reporting cannot possibly capture detailed descriptions of the tissue heterogeneities observed in many diseases.

The digitization of pathologic specimens has advanced with improvements in charge-coupled device sensors, storage, and network performance. Early versions of slide scanning hardware suffered from slow image acquisition, and their practical use was limited by the expense of storage and network limitations that made file transfer and remote viewing difficult. Contemporary slide scanning devices are now capable of digitizing a single slide at 40× objective magnification in two minutes or less, and can produce hundreds of whole-slide images (WSIs) in a single day. With each image occupying hundreds of megabytes to several gigabytes, the recent precipitous decline in storage costs in the past decade has made generation and analysis of large WSI archives more practical. Faster networks and improved software have enabled users to fluidly view and interact with large WSI archives at their desktop by streaming imaging data directly from remote servers. Currently, no universal standards exist for file format or image compression within a WSI, despite some work by the DICOM Working-Group 26, creating significant challenges in the interoperability of

[1]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA; [2]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA; [3]Winship Cancer Institute, Emory University, Atlanta, GA, USA; [4]Department of Neurology, Emory University, Atlanta, GA, USA and [5]Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, USA
Correspondence: Dr Lee AD Cooper, PhD, Department of Biomedical Informatics, Psychology and Interdisciplinary Sciences Building, Emory University, 36 Eagle Row, 5th Floor South Atlanta, Atlanta, GA 30322, USA.
E-mail: lee.cooper@emory.edu

Received 21 July 2014; revised 20 September 2014; accepted 22 September 2014

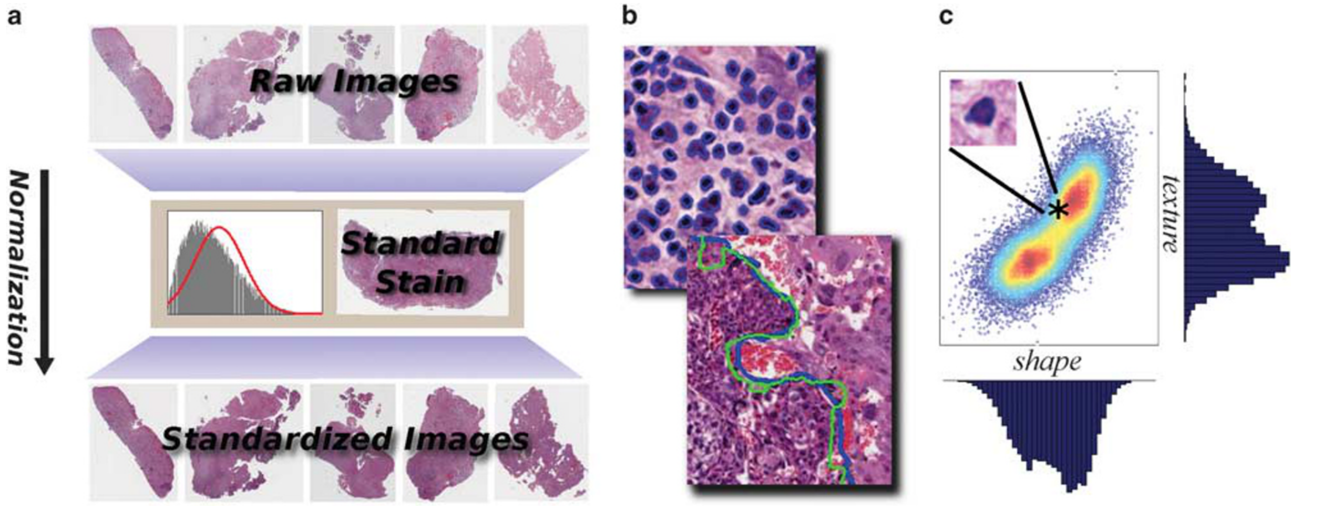

Figure 1 Image analysis enables the reliable and objective characterization of tissues through a process of image normalization, image segmentation, and feature extraction. (**a**) Variations due to staining, sectioning, and scanning can be corrected with algorithms like color normalization that transform the color characteristics of WSIs to a gold standard. (**b**) Image segmentation algorithms can be used to delineate structures in images ranging from cell nuclei to complex tissue interfaces. (**c**) Quantitative features can be calculated to describe the appearance of delineated structures to capture their shape, color, and texture characteristics. These features define each object as a point in high-dimensional space where machine learning algorithms like clustering and classification can be used for further analysis. Improvements in image analysis technology present new opportunities to explore genotype–phenotype relationships by integrating image-derived data with genomic descriptions of tissues.

various hardware and software platforms from different WSI vendors.

Improvements in computing performance and image analysis algorithms enable large WSI archives to be mined to extract quantitative and objective *imaging features* that describe the visual characteristics of tissue architecture and microanatomy[2–4] (see Figure 1). Image analysis algorithms can now reliably delineate objects across biological scales and provide rich descriptions of specimens ranging from nuclear morphology to characterizations of complex multicellular structures and tissue interfaces.[4–16] The development of methods to correct artifacts and mitigate technical variations in WSI images has been critical in advancing image analysis techniques in this domain.[17–22] Variations in section thickness and exposure to staining agents can alter the color characteristics of WSIs, and the presence of artifacts like tissue folding can obscure important image content. The algorithms used to delineate and measure objects in WSIs are sensitive to these color variations and artifacts, and achieving robust performance across collections of WSIs has been a challenging problem, and a significant barrier to the adoption of image analysis technology in clinical settings. Several groups have developed algorithms for color normalization to normalize pixel data and standardize color characteristics in WSI datasets,[17–19] thus enabling segmentation algorithms to be tuned to a common color standard instead of individually for each WSI (see Figure 1a). Algorithmic approaches have also been developed to deal with tissue folding and tearing that can occur when specimens are physically mounted on glass.[20–22] These pre-processing procedures improve the robustness of segmentation algorithms and result in more uniform features that better reflect biological properties instead of noise introduced by artifacts.

The collection of image segmentation algorithms used to delineate structures in WSIs is wide and varied in their technical approaches and intended purposes[2,3,5] (see Figure 1b). Segmentation of cell nuclei is the most widely studied problem, and has been approached using everything from ordinary thresholding to more sophisticated methods that incorporate prior models of nuclear shape and that rely on advanced network-model or variational methods.[6,15] Segmentation of higher-order structures like stroma, differentiation between regions of tumor and adjacent normal tissue or microvascular proliferation involve more complex methods that model the statistical properties of structures and tissues with textures or other pattern recognition methods,[7–11,13,14,16] and are typically more computationally demanding and difficult to apply to large datasets. With structures delineated by segmentation, quantitative features can be calculated to describe their appearance and capture the shape, color, and texture characteristics. These features define each object as a point in high-dimensional space where they can be further analyzed using clustering and classification machine-learning algorithms (see Figure 1c). In nuclear morphometry applications, features can describe aspects like circularity, nuclear size, hyperchromaticity, and the extent of chromatin clumping.[23] Similar features are used to describe the boundaries of tissue interfaces and their regularity.[24,25] Modern computing hardware like multi-core processors and specialized graphics processors enable these techniques to be scaled to large WSI datasets that can contain millions or billions of such structures. The size in bytes of features extracted from an image can rival that of the image itself, and the management and standardization of image features and their algorithmic provenance is not trivial.

Improved image analysis capabilities have created new opportunities to explore genotype–phenotype relationships by integrating image-derived data with genomic descriptions of tissues. Advanced genomic platforms can generate deep characterizations of the genetic, transcriptional, and epigenetic events within biospecimens. Owing to improvements in cost and clinical adoption, genomic characterizations are increasingly available and gaining relevance through studies like The Cancer Genome Atlas (TCGA).[26] Genomic and quantitative imaging features have complementary strengths that can be leveraged through integration to explore a number of practical problems. Genomics provide extremely high molecular resolution but poor spatial resolution, and the genomic signature of a specimen therefore represents an aggregate measure of heterogeneous molecular profiles within distinct components of the tissue analyzed. Laser capture microdissection provides a way to increase the purity of genomic measurements, but is labor-intensive and difficult to carry out on large cohorts, although image analysis has been used to reduce this burden.[27] An alternative approach is the integration of genomic and imaging features through computational means to deconvolve distinct profiles from the aggregate profile, with the goal of recovering information that is lost when tissue is homogenized for genomic analysis. Histology is also a manifestation of underlying molecular profiles within tissues, so quantitative imaging features can be expected to contain predictive power as biomarkers of genetic alterations and gene expression patterns. By integrating imaging and genomic features into risk models, prognostic variance may be reduced compared with genomics or histopathology alone.

The availability of large de-identified datasets from TCGA has greatly facilitated integrated analyses that use imaging, genomic, and clinical data. This well-characterized and comprehensive data set would be difficult to duplicate at a single institution because of prohibitive cost, privacy concerns, and patient volumes. TCGA is a large public resource that provides comprehensive molecular characterizations of more than 22 cancers types. Although intended primarily as a genomic resource, TCGA contains over 22 000 WSIs from more than 10 000 tumors, in addition to detailed clinical descriptions, and serves as an open platform to perform studies that integrate quantitative histology with molecular and clinical data. The use of these existing resources to conduct *in silico* scientific investigations has enabled researchers in this area to focus effort on developing analysis methods rather than data production, and to scale studies to a number of samples that would be otherwise difficult to achieve (see Figure 2). Although TCGA is an exceptional resource at this point in time, such multifaceted descriptions of tissues will likely become more common place within academic research institutions with increasing clinical adoption of genomics and digital pathology, and as the information management systems that manage these data improve.

In this paper, we present a review of developments in the area of quantitative histology, using examples from glioblastoma (GBM) to illustrate how imaging features can be integrated with genomic and clinical data to improve understanding. The first example explores issues of tumor microenvironment (TME), and how imaging features can illuminate the impact of the TME on the genomic signatures and molecular classifications. In the second example, we present a pipeline for the morphometric characterization of nuclei that is capable of extracting quantitative descriptions of billions of cell nuclei in digital WSI archives. We show how this pipeline can be used along with statistical and statistical learning techniques to define imaging biomarkers of genetic alterations and epigenetic and transcriptional patterns, as well as clinical outcomes. We finish by describing near-term potential opportunities for quantitative imaging and integrated studies, and discuss the limitations and challenges associated with these approaches.

MAIN BODY

The Cancer Genome Atlas

TCGA was established in 2005 to improve understanding of the molecular basis of human cancers through large-scale genomic analysis. With a goal of accruing 500 tumors for each cancer selected for study, TCGA has expanded beyond initial pilot projects in GBM, lung, and ovarian carcinoma to now span more than 22 tumor types. This effort relies on a pipeline of participating institutions that submit frozen tissues and clinical data to a central repository, a set of decentralized genomic analysis centers that produce messenger RNA, micro RNA, DNA exome sequencing, DNA copy number, DNA methylation, and protein expression profiles, and an electronic clearinghouse that then makes these data available to the public (https://tcga-data.nci.nih.gov).

An important, yet underappreciated aspect of acquiring clinical data from tissue source sites includes the collection of digitized WSIs of submitted tumors. Frozen sections are produced from the top and bottom of tissue samples that are submitted for genomic analysis, and are used for quality control to evaluate the percentage of tumor, the presence of necrosis, and other factors that will influence the quality of genomic results. These images are a valuable resource because they are immediately adjacent to tissues used for genomics, and provide the most faithful representation of genomic-annotated tissues. Diagnostic permanent section slides are also solicited from participating institutions. The higher quality of these images and lack of freezing artifacts makes them more suitable for algorithmic analysis, particularly at high magnification. Expert pathology committees that are selected by disease area review these permanent sections to ensure correct diagnosis and to evaluate the presence of important pathologic criteria. Examples from the GBM project include the categorical scoring (0, 1+, 2+) of

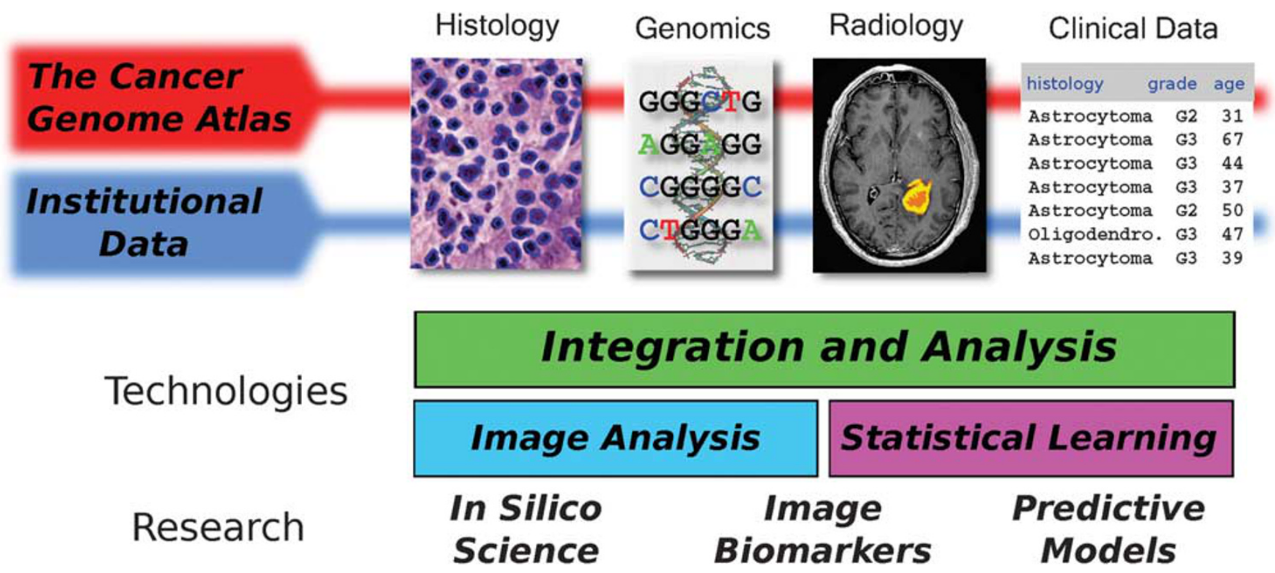

Figure 2 Integration of quantitative histology with multifaceted clinical and genomic data. Image analysis algorithms can extract features that describe the histology in digital whole-slide image datasets. This information can be combined with genomic, clinical, and radiology data to identify image biomarkers of genetic alterations, to build predictive models of clinical outcomes, and to better understand tumor biology. Public data provided by The Cancer Genome Atlas (TCGA) make it possible to explore these topics in large cohorts of more than 22 types of cancers. Discoveries made in analysis of public TCGA data can be validated in smaller institutional datasets.

qualities like microvascular proliferation, pseudopalisading necrosis, and lymphocytic infiltration. All permanent and frozen sections are digitized at × 20 or × 40 objective magnification and made publicly available for download.

Transcriptional Classification of GBMs and the TME

One of the main outcomes from TCGA has been the discovery of previously unrecognized genomic sub-classifications for a broad set of cancers.[28–35] Using clustering analysis of gene expression and other molecular platforms, the goal of these analyses is to define cohesive sub-classes of tumors with distinct molecular signatures that may benefit from class-specific targeted therapies. In GBM, two studies using TCGA data have identified tumor sub-classifications based on gene expression and DNA methylation.[36,37] The initial TCGA analysis of GBMs identified four gene expression classes (GECs): proneural, neural, classic, and mesenchymal. These classes exhibit clear and distinct patterns of gene expression, and are highly correlated with genetic alterations in *EGFR*, *IDH1*, *NF1*, *PDGFRA*, and *TP53*. A subsequent analysis of DNA methylation data revealed that proneural GBMs are further subdivided into two groups—those with *IDH* mutations that have significant hypermethylation of CpG islands (GCIMP) and are typically secondary GBMs afflicting younger patients, and *IDH* wild-type tumors that do not exhibit DNA hypermethylation patterns.[37]

One of the first goals of our *in silico* research was to investigate the relationships between gene expression classifications and the TME in GBMs.[33,38] Most tissue-based transcriptional classification studies of tumors are subjective, in that neoplasms are highly heterogeneous, and gene expression measurements can vary significantly among different samples from the same tumor. GBMs are no exception, being spatially complex tumors that harbor a variety of non-neoplastic cell types and microenvironmental elements that can significantly impact gene expression measurements. Pseudopalisading necrosis and microvascular proliferation are perhaps the most notable elements, being part of the diagnostic criteria that distinguish GBMs from lower-grade gliomas, and indicators of poor prognosis.[39] The development of necrosis and microvascular proliferation can be focal at first, but then expands, and signals severe underlying hypoxia with resultant profound transcriptional changes.

Having access to both frozen sections and comprehensive molecular profiles from adjacent tissues from TCGA, we sought to measure the impact of necrosis and angiogenesis on gene expression patterns used to classify GBMs. We hypothesized the extent of necrosis and angiogenesis in a histologic section are tightly associated with the presence of hypoxia, which could play an important role in establishing GEC expression signatures by activation of hypoxia-inducible transcription factors. With the degree of hypoxia varying spatially throughout a tumor, multiple GES classes could possibly co-exist within the same tumor, and so classification by gene expression could be subject to random effects in tissue sampling. Intra-tumoral variations in GEC classification would have significant implications in using these classes as platforms for the development of targeted therapies.

Using the Imagescope human-computer interface (Leica Biosystems, Nussloch, Germany), we manually annotated 177 digitized frozen section images to define the boundaries of necrosis and angiogenesis for 99 tumors. The lumens within

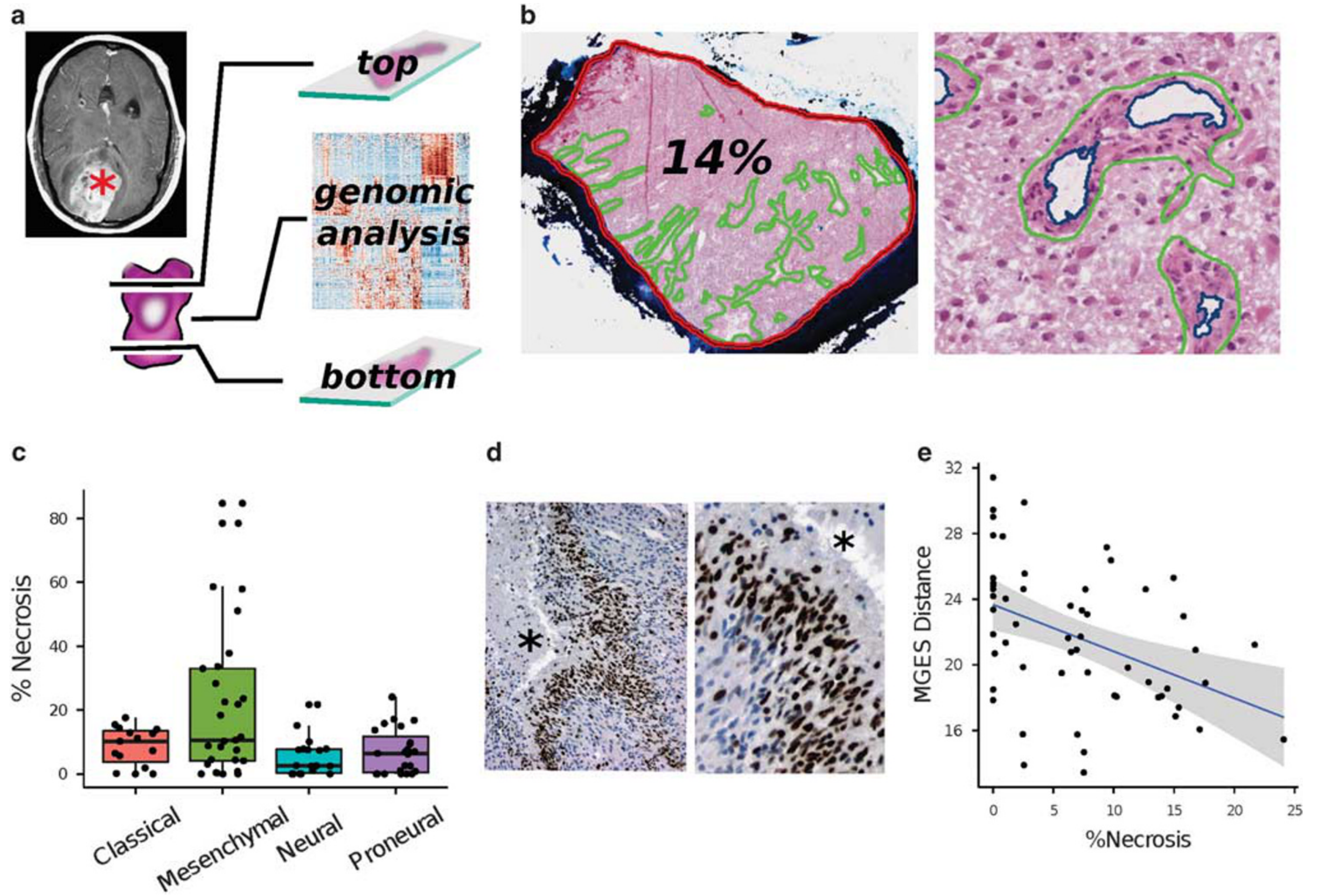

Figure 3 Tumor microenvironment studies integrating histology and genomics from TCGA. (**a**) TCGA specimens are sections from the top and bottom to produce slides, and the middle portion is submitted for genomic analysis. (**b**) Digitized images from top/bottom sections were annotated to calculate the percentage of necrosis and angiogenesis for each tumor. (**c**) Tumors from the mesenchymal gene expression class are significantly enriched with necrosis. (**d**) *CEBPB* expression is increased in the hypoxic perinecrotic pseudopalisading cells immediately surrounding necrosis (asterisk). (**e**) As the amount of necrosis increases in non-mesenchymal GBMs, gene expression patterns shift towards a mesenchymal expression signature.

angiogenic regions were subtracted using an image analysis segmentation algorithm. The sections in these images are immediately adjacent to those used for genomic analysis, and these annotations therefore provide the most faithful representation of the microenvironmental conditions in tissues with genomic annotations (Figure 3a). The extent of necrosis and angiogenesis was calculated as a percentage of the total tissue area, and these quantitative features were linked to gene expression and other genomic measurements from the same tissue (Figure 3b).

We first examined the abundance of necrosis and angiogenesis in tumors organized by TCGA transcriptional class. Tumors with a mesenchymal GES were clearly enriched with higher amounts of necrosis (one-way ANOVA $P = 8.7e\text{-}4$, see Figure 3c), suggesting a strong association between the mesenchymal gene expression signatures, necrosis, and hypoxia. All tumors with greater than 22% necrosis were members of the mesenchymal class. The relationship between angiogenesis and transcriptional class was less clear. The percentage of angiogenesis ranged from 0 to 4% for the large majority of tumors. There were only four outliers with much higher levels of angiogenesis and three were from the mesenchymal GES and one was from the proneural GES.[38] Although we would expect the presence of angiogenesis to influence gene expression, the ability to detect this association may be limited in part by the relatively small contribution of angiogenic regions to the RNA extracted for analysis.

Next, we performed a genome-wide analysis of transcriptional data to discover the impact of necrosis on gene expression. A normalized linear regression coefficient was calculated for each transcript to measure the strength of relationship between extent of necrosis and gene expression for more than 22 000 transcripts measured with Affymetrix arrays. Significance analysis of microarrays correction was applied to obtain multiple-test corrected *P*-values for each gene.[40] This analysis identified 2422 genes that are significantly correlated with extent of necrosis at 5%

false-discovery rate or below, suggesting that necrosis has tremendous influence on gene expression in GBMs. Among the genes most significantly correlated with necrosis were a set of transcription factors known as *mesenchymal master regulators*: *CEBPB*, *FOSL2*, *CEBPD*, *STAT3*, *BHLHE40* (ranked 4th, 10th, 60th, 213th, and 221st respectively). These transcription factors have been shown to form a small module regulating a much broader gene expression network that is responsible for mesenchymal tumor phenotype in GBMs.[41] At the top of this regulatory module are the transcription factors *CEBPB*/*CEBPD* and *STAT3*, whose coexpression is necessary and sufficient for activating the mesenchymal expression network. To explore the expression of these regulators in tissues, we performed immunohistochemistry on archived surgically resected GBMs from our own institution. We observed that *CEBPB*/*CEBPD* expression was strongly and specifically expressed in the hypoxic pseudopalisading cells surrounding areas of necrosis (Figure 3d). *CEBPB* was strongly expressed in nuclei of the first two to five cell layers immediately surrounding necrosis, and *CEBPD* expression was found in both nuclear and cytoplasmic regions of perinecrotic cells but extending slightly farther beyond *CEBPD.* In regions between foci of necrosis, only a small portion of cells expressed either *CEBPB* or *CEBPD*. *STAT3* did not show a specific perinecrotic pattern of expression.[38]

Gene expression classifications are made by comparing a tumor's expression profile to a set of signatures representing each GEC in *gene expression space.* A tumor is assigned to the class with the most similar GEC signature, where similarity can be measured using metrics like Pearson correlation distance or Euclidean distance. Although a given tissue could potentially contain individual cells/regions with diverging gene expression profiles, these classifications force a selection of a single GEC that best describes the aggregate measure of the entire sample. To explore how the formation of necrosis influences the expression patterns of non-mesenchymal GBMs, we examined the relationship between extent of necrosis and distance to the mesenchymal GEC within this cohort. We observed a clear trend—the more necrosis that a sample contains, the more its expression profile resembles the mesenchymal GEC signature (Figure 3e). This finding further suggests that mesenchymal gene expression is strongly impacted by hypoxia and that expression signatures are strongly impacted by regionally varying elements of the microenvironment.

Molecular and Clinical Associations Revealed Through Quantitative Nuclear Morphometry

The morphologic characteristics of cell nuclei convey important clinical information in many types of neoplasms. Besides determining histologic classification and subtype, nuclear qualities including shape, texture, and spatial arrangement can be indicative of more specific molecular alterations and patient prognosis.[42–45] Gains, losses, and rearrangements of DNA along with epigenetic modifications affecting chromatin structure can manifest in observable changes within nuclei of neoplastic cells. In the diffuse gliomas, nuclear features are of particular importance, as their classification of oligodendroglioma or astocytoma is based in large part on nuclear morphology. However, histopathologic classification based on human review is subjective and prone to substantial inter-observer variation. Understanding the relationships between nuclear morphology, tumor genetics and clinical outcomes will provide a better understanding of tumor biology and further improve the precision of clinical predictions.

Our studies of TME used human markups and annotations to generate quantitative features from WSIs. The limitations of human annotations are apparent when dealing with nuclear morphology—nuclei can number in the hundreds of millions in even a modestly sized set of images, and qualities of interest like nuclear texture are difficult to accurately characterize objectively by human observers. To address these challenges, we have developed a computational system for the study of nuclear morphometry in large archives of WSIs (Figure 4a). This system uses image analysis algorithms to delineate individual cell nuclei, and to calculate a set of objective *nuclear features* for each nucleus to describe its shape and texture. High performance and parallel computing approaches are used to scale this approach to hundreds of millions of cells. This system presents opportunities to define quantitative morphologic biomarkers of molecular and clinical endpoints by enabling the extraction of objective, repeatable measurements from WSI archives.

Our initial morphometric study focused on the quantitative characterization of oligodendroglial differentiation in GBMs.[23] Although GBM is defined as a grade IV astrocytoma, an important subset exhibits varying degrees of oligodendroglial differentiation in addition to the dominant astrocytic component.[39,43,46,47] Neoplasms with pure oligodendroglial differentiation typically have slower growth and better survivals when compared with astrocytomas of the same grades. The morphologic characteristics of oligodendrogliomas distinguish them from astrocytomas: oligodendroglial nuclei tend to be smaller, round, and hyperchromatic with a lack of detailed texture, in contrast with astrocytoma nuclei that are larger, irregularly shaped, typically elongated, and unevenly textured. In most instances, GBMs contain a heterogeneous mixture of neoplastic cells with wide variations in nuclear characteristics, many of which are not clearly astrocytic or oligodendroglial. The volume and heterogeneity of cells present in GBMs combined with subtle differences in morphologic diversity make them an ideal candidate for computational morphometric approaches.

Using our computational pipeline for nuclear morphometry, we analyzed 200 million nuclei from digitized images of diagnostic slides corresponding to 117 TCGA GBMs.[23] Twenty-three quantitative features from four categories

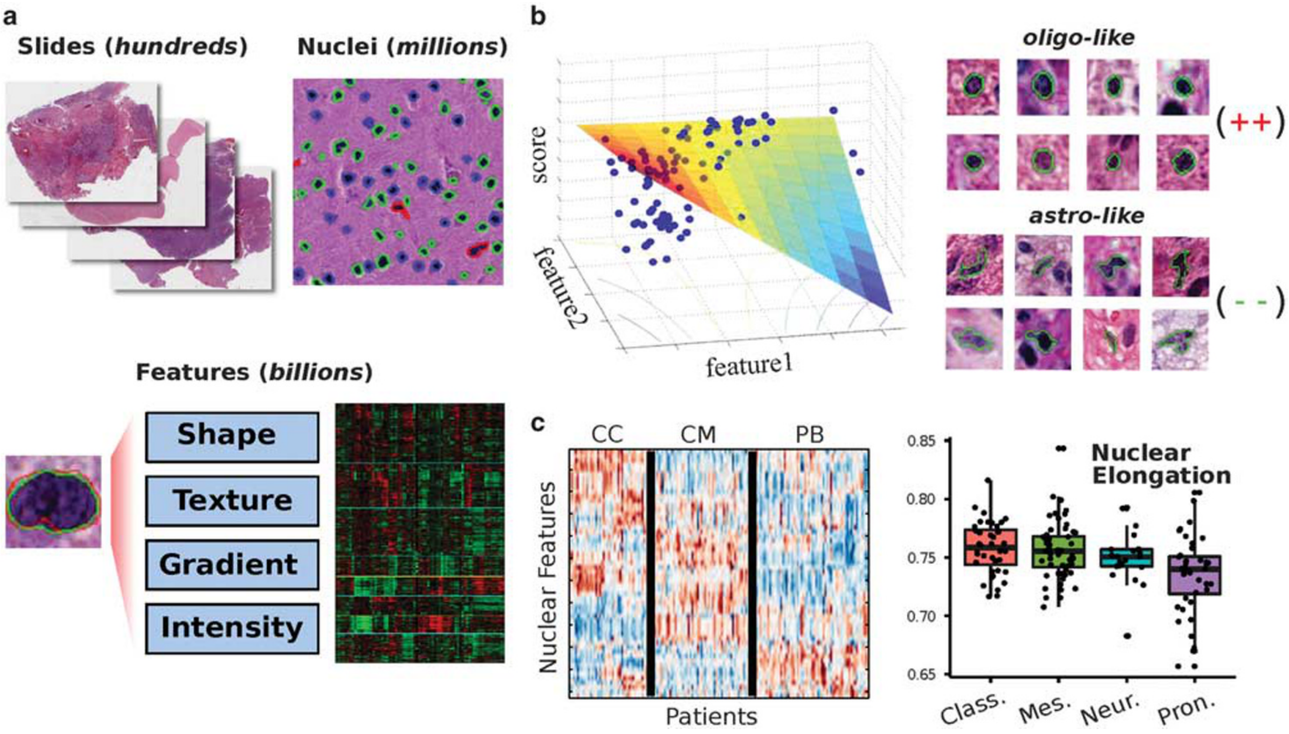

Figure 4 Quantitative nuclear morphometry. (**a**) Image analysis algorithms are used to delineate nuclei in whole-slide images. A set of features is calculated to describe the shape and texture of each nucleus.[48] This system is capable of processing thousands of slides and hundreds of millions of nuclei. (**b**) We developed a model-based system to score nuclei based on oligodendroglial differentiation. This model was validated by correlation of nuclear scores and gene expression data. (**c**) Model-free approaches were used to explore the clinical and genomic associations of nuclear features. Clustering of patient morphological signatures revealed three distinct patient clusters, *cell cycle* (CC), *chromatin modification* (CM) and *protein biosynthesis* (PB), named for the biologic functions associated with their genomic correlates. Unsupervised analysis of features revealed that proneural tumors are associated with more round, regular nuclei.

(shape, intensity, texture, and gradient) were calculated to describe each nucleus. To represent the differentiation of each nucleus along the oligodendroglial/astrocytic spectrum, we built a regression model that uses the nuclear feature values to calculate a score for each nucleus representing its degree of oligodendroglial appearance (Figure 4b). Combining the 200 million scores obtained from our pipeline with gene expression, copy number, DNA sequence, and methylation data from the same TCGA tumors, we were able to clearly separate a set of tumor enriched with oligodendroglial-like cells that had strong associations with *PDGFRA* amplification, proneural transcriptional class, and expression of the oligodendrocyte signature genes *MBP*, *HOXD1*, *PLP1*, *MOBP*, and *PDGFRA*. These results provide molecular validation that the quantitative features extracted by our software pipeline can capture the morphologic variations of nuclei encountered in gliomas.

Our differentiation study used a supervised approach to build a quantitative model of the oligodendroglial/astrocytic spectrum in gliomas. Model-based approaches are a powerful way to incorporate prior knowledge into morphologic analyses, and to use quantitative measures of recognized morphologic patterns to explore their molecular correlates. Because model-based approaches are built on prior knowledge, their ability to reveal previously unrecognized or unknown morphologic patterns is limited. To address this limitation, we have developed several *unsupervised* or *model-free* approaches that do not impose established constructs in the morphological analysis of WSI data. Instead, these approaches let data speak for itself, using clustering analysis and other statistical learning techniques to reveal natural structure within the feature data in a bottom-up manner.

Our first study with unsupervised methods investigated *patient clustering* of GBMs into morphologically defined subtypes.[48,49] Using nuclear features, we sought to determine whether there are clear and distinct groups of tumors that emerge from clustering analysis, similar to gene expression studies where transcriptional profiles are clustered to reveal molecular tumor subtypes. Taking the nuclear features from the TCGA cohort, a morphologic signature was calculated for each tumor to represent the morphologic properties of its average nuclei. These signatures were analyzed using a consensus-clustering algorithm to find natural groups within the data and to measure their robustness. Three clear clusters emerged from this analysis and we named them for themes observed in their molecular correlates: cell cycle, protein biosynthesis (PB) and chromatin modification (see Figure 4c). We observed that these clusters had significant differences in patient survival (logrank $P = 1.4e\text{-}3$), with the PB cluster containing patients with relatively better outcomes and the chromatin modification cluster relatively worse. These clusters were also observed in an independent dataset of 84 GBMs where the relative differences in outcomes between the clusters were also confirmed. To explore the meaning of these clusters, we used the various genomic platforms made available by TCGA including gene expression, DNA methylation, copy number, and DNA sequencing. A pathway analysis using DAVID[50] and Ingenuity Pathway Analysis software (Qiagen, Redwood City, CA, USA) found that the clusters varied in the extent of *TP53 WNT*, and *NFKB* signaling, and had variations in the extent of total DNA methylation. An analysis of the pathologic features in these clusters using categorical human annotations (0-absent, 1+ - present, 2+ - abundant) generated by a consortium of TCGA neuropathologists found that tumors in the chromatin modification cluster were enriched with lymphocytes, and that PB cluster tumors conspicuously lacked inflammation.

To further explore model-free associations of nuclear morphometry in GBM and clinical and genomic endpoints, we took a more direct approach of correlating raw nuclear features with genomic and clinical endpoints.[51] For each patient, we calculated the mean and standard deviation of each feature as metrics and correlating them directly with patient survival using Cox proportional hazards analysis using significance analysis of microarrays. Notably, the mean circularity was significantly associated with longer patient survival, an observation consistent with prolonged clinical outcomes in gliomas with oligodendroglial differentiation. Other features that were significantly associated with outcome include major axis length, with longer nuclei associated with a shorter survival, and min nuclear pixel intensity, with higher values associated with longer survival. The fact that these features emerged from a more data-driven approach provides some level of confidence in our analysis workflow. To correlate these features with genomic measurements, we performed a one-way ANOVA for each feature metric across transcriptional classifications, somatic mutations, and DNA copy number alterations. Features distinguishing transcriptional classes include nuclear eccentricity ($P = 3.81e\text{-}4$), minor axis length ($P = 8.87e\text{-}3$), and nuclear extent ($P = 3.2e\text{-}2$). The greatest morphology differences were observed between the proneural and mesenchymal tumors (Figure 4c). Those hypermethylated (GCIMP) tumors within the proneural group had greater variation of pixel intensities within their nuclei (*nuclear energy*, $P = 2.28e\text{-}5$), and greater variation in nuclear size. Genetic events having significant differences in nuclear morphometry included *PTEN* and *TP53* mutations, and PDGFRA amplification. *PTEN* and *TP53* mutant tumors were both associated with less circular nuclei ($P = 9.68e\text{-}3$, 3.77e-2, respectively). *PDGFRA*-amplified tumors were associated with greater circularity ($P = 2.31e\text{-}2$), consistent with *PDGFRA* amplifications being associated with oligodendroglial differentiation. Other genetic alterations with significant associations included *EGFR* amplification, which was associated with greater nuclear eccentricity and canny, and *MDM2* amplifications, which were associated with greater minor axis length, area, and circularity.

DISCUSSION

Emerging Challenges and Opportunities

Advances in whole-slide imaging and computing hardware have made it possible to approach increasingly difficult image analysis problems in pathology. At the same time, the increasing availability of rich genomic data have made pathology image analysis studies more interesting by allowing histologic features to be linked to comprehensive molecular descriptions. Within the last decade, the goals of pathology image analysis have shifted from implementing computer-aided diagnostic procedures that recapitulate existing classifications, like tumor grade, to more creative analyses that explore genotype–phenotype associations and that define novel prognostic measures. The emerging goal is to go beyond replicating tasks performed by human pathologists to develop novel techniques that extract latent content from images that has unrecognized clinical and scientific value. This convergence of image analysis and bioinformatics has produced some exciting results in several different areas beyond the work presented in the previous sections. In GBMs, morphometry-driven tumor subtypes were identified on the basis of nuclear morphometry and cellularity measurements and found to be predictive of clinical outcomes and pathway activation.[6,52] In these studies, nuclei were segmented and characterized with a set of features designed to capture characteristics of nuclear morphology like shape, texture and intensity, and spatial organization. An intermediate level of analysis was performed by dividing each whole-slide image into 500-micron tile regions, and then generating summary statistics of the nuclei within each block. These block-level profiles were then clustered to identify regional types that resemble necrotic, normal brain, medium-density, and high-density tumor regions. These block-level profiles were used to derive compositional profiles of the tissues by mapping each block to these characteristic types. These compositional profiles were then compared with clinical outcomes and gene expression data to identify hazard ratios and the molecular correlates of tumor composition. In breast cancer, morphologic features describing stromal/tumor interface were found to be predictive of overall survival independent of other clinical, pathologic, and molecular factors in two independent cohorts.[24] This analysis used a *superpixel* segmentation that first breaks the image data into compact regions of several microns that share common color characteristics. Nuclei were then identified within these superpixels using a second segmentation step. These superpixels were used to develop a tissue classifier of epithelial *vs* stroma to partition the slide into higher-level regions. A set of 6642 features was then calculated in each specimen to capture the higher-level organization of the specimen, representing the relationships between epithelial and stromal regions and characteristics of stromal and epithelial cell nuclei and cytoplasm. These features were used to train a 5-year survival model that validated in an independent cohort, with the surprising finding that features describing stromal patterns were highly prognostic. Another analysis of breast cancer focused on the correction of copy number profiles by analysis and classification of cell nuclei.[53] Nuclear segmentation and feature extraction were used to characterize cell nuclei and to develop a classifier for neoplastic, stromal, and lymphocytic cell types. This classifier was used to derive improved tumor purity measurements that were then used to enhance copy number profiling of the same tissues. Image analysis and molecular features were then combined with clinical variables like node-size and grade to develop a prognostic model for estrogen-receptor-negative patients that was validated in an independent cohort. This modeling exercise showed that the integrated prognostic model outperformed single-mode models based on only genomics, imaging or clinical variables, and again found that spatial stromal patterns are among the most valuable prognostic variables. In another analysis of triple-negative breast cancers, image analysis was used to develop a prognostic model based entirely on imaging features.[54] Morphometric features of segmented cell nuclei were combined with higher-order superpixel analysis to build features that capture the organization of tumor and stroma that define the TME. A univariate survival analysis of these morphologic features revealed four prognostically significant features including variation in nuclear size that were verified through cross validation of the training data. Recognizing that the availability of datasets with both genomics and histological data are limited, this analysis used a creative approach to define a gene signature that serves as a surrogate for these prognostic morphologic features, and used this signature to validate the prognostic value of these features in datasets where WSIs were not available.

One barrier to progress in this area is the dissemination of algorithms and image features beyond image analysis experts to the broader research community. Making software and feature data publicly accessible will facilitate advances in this field by more fully engaging the pathology community and providing opportunities for comparative studies of prognostic models. Establishing the computational resources needed to execute image analysis algorithms on the primary images is difficult, and the sharing of derived feature data is limited by a lack of standardization. To begin to address these issues, we have developed web-based interfaces and data standards to support the visualization, federation, and analysis of pathology image data. The Cancer Digital Slide Archive (CDSA, http://cancer.digitalslidearchive.net/) is a web-based resource that was originally developed to facilitate the visualization and analysis of pathology imaging and clinical data from TCGA.[55] The CDSA currently hosts over 22 000 images and associated clinical data from over 22 different cancers represented in TCGA. This interface provides access to pathology and clinical data through a simple web-browser interface. Although currently focused on serving primary images for visualization, the CDSA and other similar resources could naturally serve as clearing-houses that allow a broader set of users to interact with image analysis algorithms and feature sets. Cloud-based services could be established that enable primary image data, derived features and computational tools to reside in a common computing environment, avoiding the need for costly transfer of massive amounts of image and feature data. Users could then perform end-to-end integrated analyses of pathology imaging online without the need to establish local computing resources or shepherding primary image or feature data between systems across the Internet. In cases where feature data and algorithms are exchanged, we have also developed the Pathology Analytic Imaging Standards (PAIS) to support the standardization of image analysis algorithms and image features.[56,57] PAIS provides data standards that enable users to capture software and algorithm parameter provenance, and a common file format that enables results to be stored in a database for search and exchange.

The validation of findings remains another significant barrier in advancing integrative pathology imaging studies. This issue is particularly important in studies that identify image biomarkers of clinical outcomes or genomic features. One of the risks in using image features to predict outcome or genomic measurements is *overfitting*—are the algorithms learning meaningful relationships that will generalize to new unseen datasets, or simply generating predictions that are specific to the noise and artifacts of the data? Validating findings in external datasets, when available, or by proper cross-validation of a single dataset is important for distinguishing true findings from algorithmic artifacts. Although many of the studies described above have validated their findings in independent datasets, the availability of multiple large WSI datasets that have genomic and clinical annotations is limited and lack of data remains as a severe constraint in broadening studies to other diseases. The predictive morphologic features that are identified by integrated studies are not always visually apparent to the human eye, whereas the validation of these features in independent datasets suggests that their predictive power is very real. In the cases where visualization is not possible or does not produce any obvious visible distinction, it is difficult to interpret the meaning of predictive features or to link them to existing knowledge about the histology of disease. Greater availability of benchmark datasets containing WSIs, and genomic and clinical data could help to establish reproducibility and improve confidence in the relationships defined through computational analysis.

Another area for growth is the integration of radiology imaging with pathology, genomics, and clinical data. Image analysis methods for radiology data are more mature than for pathology, and are able to extract meaningful features from MR, PET, CT, and other medical imaging modalities. The global perspective of tumor provided by medical imaging is complementary to the tissue and molecular scale measurements provided by pathology and genomics, and a number of studies have already explored the relationships between quantitative radiology imaging features, genomic profiles, and clinical outcomes.[58–61] Integrating complementary features across biological scales into prognostic models is a promising avenue to improve the precision of clinical predictions and risk stratification of patients.

ACKNOWLEDGMENTS

This work was supported by US Public Health Service National Institutes of Health (NIH) grants K22LM011576 (LADC), K25CA181503 (JK), R01CA176659 (DJB) and the Georgia Research Alliance (DJB).

DISCLOSURE/CONFLICT OF INTEREST

The authors declare no conflict of interest.

1. van den Bent MJ. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathol 2010;120:297–304.
2. Gurcan MN, Boucheron LE, Can A, *et al.* Histopathological image analysis: a review. IEEE Rev Biomed Eng 2009;2:147–171.
3. Cooper LAD, Carter AB, Farris AB, *et al.* Digital Pathology: Data-Intensive Frontier in Medical Imaging: Health-information sharing, specifically of digital pathology, is the subject of this paper which discusses how sharing the rich images in pathology can stretch the capabilities of all otherwise well-practiced disciplines. Proc IEEE Inst Electr Electron Eng 2012;100:991–1003.
4. Kothari S, Phan JH, Stokes TH, *et al.* Pathology imaging informatics for quantitative analysis of whole-slide images. J Am Med Inform Assoc 2013;20:1099–1108.
5. Hsu W, Markey MK, Wang MD. Biomedical imaging informatics in the era of precision medicine: progress, challenges, and opportunities. J Am Med Inform Assoc 2013;20:1010–1013.
6. Chang H, Han J, Borowsky A, *et al.* Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. IEEE Trans Med Imaging 2013;32:670–682.
7. Chang H, Nayak N, Spellman PT, *et al.* Characterization of tissue histopathology via predictive sparse decomposition and spatial pyramid matching. Med Image Comput Comput Assist Interv 2013;16(Pt 2):91–98.
8. Hipp J, Smith SC, Cheng J, Tomlins SA, *et al.* Optimization of complex cancer morphology detection using the SIVQ pattern recognition algorithm. Anal Cell Pathol (Amst) 2012;35:41–50.
9. Cheng J, Hipp J, Monaco J, *et al.* Automated vector selection of SIVQ and parallel computing integration MATLAB: Innovations supporting large-scale and high-throughput image analysis studies. J Pathol Inform 2011;2:37.
10. Janowczyk A, Chandran S, Madabhushi A. Quantifying local heterogeneity via morphologic scale: Distinguishing tumoral from stromal regions. J Pathol Inform 2013;4(Suppl):S8.
11. Janowczyk A, Chandran S, Singh R, *et al.* High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts. IEEE Trans Biomed Eng 2012;59:1240–1252.
12. Song Y, Treanor D, Bulpitt AJ, *et al.* Unsupervised content classification based nonrigid registration of differently stained histology images. IEEE Trans Biomed Eng 2014;61(1):96–108.
13. Mosaliganti K, Janoos F, Irfanoglu O, *et al.* Tensor classification of N-point correlation function features for histology tissue segmentation. Med Image Anal 2009;13:156–166.
14. Cooper L, Saltz J, Machiraju R, *et al.* Two-point correlation as a feature for histology images: feature space structure and correlation updating. Conf Comput Vis Pattern Recognit Workshops 201079–86.
15. Qi X, Xing F, Foran DJ, *et al.* Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. IEEE Trans Biomed Eng 2012;59:754–765.
16. Samsi S, Krishnamurthy AK, Gurcan MN. An efficient computational framework for the analysis of whole slide images: application to follicular lymphoma immunohistochemistry. J Comput Sci 2012;3: 269–279.
17. Khan AM, Rajpoot N, Treanor D, *et al.* A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 2014;61:1729–1738.
18. Murakami Y, Abe T, Hashiguchi A, *et al.* Color correction for automatic fibrosis quantification in liver biopsy specimens. J Pathol Inform 2013;4:36.
19. Bautista PA, Hashimoto N, Yagi Y. Color standardization in whole slide imaging using a color calibration slide. J Pathol Inform 2014;5:4.
20. Bautista PA, Yagi Y.. Improving the visualization and detection of tissue folds in whole slide images through color enhancement. J Pathol Inform 2010;1:25.
21. Chappelow J, Tomaszewski JE, Feldman M, *et al.* HistoStitcher((c)): an interactive program for accurate and rapid reconstruction of digitized whole histological sections from tissue fragments. Comput Med Imaging Graph 2011;35:557–567.
22. Kothari S, Phan JH, Wang MD. Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J Pathol Inform 2013;4:22.
23. Kong J, Cooper LA, Wang F, *et al.* Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. PLoS One 2013;8:e81049.
24. Beck AH, Sangoi AR, Leung S, *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3108ra13.
25. Mosaliganti K, Pan T, Ridgway R, *et al.* An imaging workflow for characterizing phenotypical change in large histological mouse model datasets. J Biomed Inform 2008;41:863–873.
26. Network TCGAR. The Cancer Genome Atlas.
27. Roy Chowdhuri S, Hanson J, Cheng J, *et al.* Semiautomated laser capture microdissection of lung adenocarcinoma cytology samples. Acta Cytol 2012;56:622–631.
28. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
29. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–337.
30. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499:43–49.
31. Cancer Genome Atlas Research NKandoth C, Schultz N, Cherniack AD, *et al.* Integrated genomic characterization of endometrial carcinoma. Nature 2013;497:67–73.
32. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519–525.
33. Rutledge WC, Kong J, Gao J, *et al.* Tumor-infiltrating lymphocytes in glioblastoma are associated with specific genomic alterations and related to transcriptional class. Clin Cancer Res 2013;19:4951–4960.
34. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609–615.
35. Cancer Genome Atlas Research N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 2014;507: 315–322.
36. Verhaak RG, Hoadley KA, Purdom E, *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17:98–110.
37. Noushmehr H, Weisenberger DJ, Diefes K, *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 2010;17:510–522.
38. Cooper LA, Gutman DA, Chisolm C, *et al.* The tumor microenvironment strongly impacts master transcriptional regulators and gene expression class of glioblastoma. Am J Pathol 2012;180:2108–2119.
39. Louis DN, Ohgaki H, Wiestler OD, *et al.* The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol 207;114: 97–109.

40. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–5121.
41. Carro MS, Lim WK, Alvarez MJ, *et al*. The transcriptional network for mesenchymal transformation of brain tumours. Nature 2010;463: 318–325.
42. Aldape K, Burger PC, Perry A. Clinicopathologic aspects of 1p/19q loss and the diagnosis of oligodendroglioma. Arch Pathol Lab Med 2007;131:242–251.
43. Gupta M, Djalilvand A, Brat DJ. Clarifying the diffuse gliomas: an update on the morphologic features and markers that discriminate oligodendroglioma from astrocytoma. Am J Clin Pathol 2005;124: 755–768.
44. Peraud A, Watanabe K, Schwechheimer K, *et al*. Genetic profile of the giant cell glioblastoma. Lab Invest 1999;79:123–129.
45. Perry A, Aldape KD, George DH, *et al*. Small cell astrocytoma: an aggressive variant that is clinicopathologically and genetically distinct from anaplastic oligodendroglioma. Cancer 2004;101: 2318–2326.
46. Hegi ME, Janzer RC, Lambiv WL, *et al*. Presence of an oligodendroglioma-like component in newly diagnosed glioblastoma identifies a pathogenetically heterogeneous subgroup and lacks prognostic value: central pathology review of the EORTC_26981/NCIC_CE.3 trial. Acta Neuropathol 2012;123:841–852.
47. Appin CL, Gao J, Chisolm C, *et al*. Glioblastoma with oligodendroglioma component (GBM-O): molecular genetic and clinical characteristics. Brain Pathol 2013;23:454–461.
48. Cooper LA, Kong J, Gutman DA, *et al*. Integrated morphologic analysis for the identification and characterization of disease subtypes. J Am Med Inform Assoc 2012;19:317–323.
49. Cooper LA, Kong J, Wang F, *et al*. Morphological Signatures and Genomic Correlates in Glioblastoma. Proc IEEE Int Symp Biomed Imaging 20111624–1627.
50. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
51. Jun K, Fusheng W, Teodoro G, *et al*. High-performance computational analysis of glioblastoma pathology images with database support identifies molecular and survival correlates. Bioinformatics and Biomedicine (BIBM). Proceedings (IEEE Int Conf Bioinformatics Biomed) 2013229–236.
52. Ju H, Hang C, Fontenay GV, *et al*. Molecular bases of morphometric composition in Glioblastoma multiforme.. Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on; 2012 2-5 May 2012.
53. Yuan Y, Failmezger H, Rueda OM, *et al*. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci Transl Med 2012;4157ra43.
54. Wang C, Pecot T, Zynger DL, *et al*. Identifying survival associated morphological features of triple negative breast cancer using multiple datasets. J Am Med Inform Assoc 2013;20:680–687.
55. Gutman DA, Cobb J, Somanna D, *et al*. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. J Am Med Inform Assoc 2013;20:1091–1098.
56. 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.
57. Wang F, Kong J, Gao J, *et al*. A high-performance spatial database based approach for pathology imaging algorithm evaluation. J Pathol Inform 2013;4:5.
58. Gutman DA, Cooper LA, Hwang SN, *et al*. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. Radiology 2013;267:560–569.
59. Zinn PO, Mahajan B, Sathyan P, *et al*. Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. PLoS One 2011;6:e25451.
60. Jain R, Poisson L, Narang J, *et al*. Genomic mapping and survival prediction in glioblastoma: molecular subclassification strengthened by hemodynamic imaging biomarkers. Radiology 2013;267:212–220.
61. Gevaert O, Xu J, Hoang CD, *et al*. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results. Radiology 2012;264:387–396.