Published in final edited form as:

ACM BCB. 2012 October ; 2012: 218–225. doi:10.1145/2382936.2382964.

Biological Interpretation of Morphological Patterns in Histopathological Whole-Slide Images

Sonal Kothari¹, John H. Phan¹, Adeboye O. Osunkoya², and May D. Wang¹

¹Georgia Institute of Technology

²Emory University School of Medicine

Abstract

We propose a framework for studying visual morphological patterns across histopathological whole-slide images (WSIs). Image representation is an important component of computer-aided decision support systems for histopathological cancer diagnosis. Such systems extract hundreds of quantitative image features from digitized tissue biopsy slides and produce models for prediction. The performance of these models depends on the identification of informative features for selection of appropriate regions-of-interest (ROIs) from heterogeneous WSIs and for development of models. However, identification of informative features is hindered by the semantic gap between human interpretation of visual morphological patterns and quantitative image features. We address this challenge by using data mining and information visualization tools to study spatial patterns formed by features extracted from sub-sections of WSIs. Using ovarian serous cystadenocarcinoma (OvCa) WSIs provided by the cancer genome atlas (TCGA), we show that (1) individual and (2) multivariate image features correspond to biologically relevant ROIs, and (3) supervised image feature selection can map histopathology domain knowledge to quantitative image features.

INTRODUCTION

Computer-aided decision support systems provide a fast and quantitative means for processing histopathological images. Such systems convert image samples into lists of quantitative image features and use pre-trained mathematical models to predict various clinical diagnoses. However, in the last decade, only a few computer-aided cancer diagnostic systems were accepted in clinical practice [1–3]. The main challenges for translation of these systems to clinical practice include (1) the semantic gap between image descriptors and histopathology domain knowledge and (2) the noise in histopathological image samples and in subsequent extracted features. With the emergence of whole-slide digital pathology, algorithms for automating the identification of regions-of-interest (ROIs) in large, heterogeneous whole-slide images (WSIs) are essential for reducing noise and improving diagnostic accuracy. We present a visualization and analysis framework that addresses two questions pertaining to WSI analysis: (1) Do common quantitative image features form

*Corresponding Author: Contact information for the corresponding author: maywang@bme.gatech.edu, Phone: 404-385-2954, Fax: 404-894-4243, Address: Suite 4106, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332, USA.

spatial patterns on WSIs? (2) Do these spatial patterns correspond to biologically relevant WSI properties such tissue necrosis or cancer grade? Visualization techniques have previously been applied to clinical histopathological images, but have only recently expanded to include WSIs.

Most existing visualizations focus on linking image features to clinical patient information such as cancer grade, subtype and prognosis. Cruz-Roa et al. applied biclustering to simultaneously cluster image samples and code-words used in a bag-of-features description of the samples. Dendrograms from the biclustering allow users to observe the combination of code-words that represent any concept class [4]. Liu et al. used heat maps with dendrograms (TreeView software) to discover clusters among tissue microarray (TMA) samples [5]. Lessman et al. used self-organizing maps (SOMs) to interpret wavelet-based features [6].

Iglesias-Rozas and Hopf used SOMs to cluster human glioblastoma samples [7]. They discovered four prominent clusters and associated these clusters with the presence of semantic histological features. Lobenhofer et al. used hierarchical clustering to visualize biologically related histopathological samples generated as part of a toxigenomics compendium study [8].

Researchers have proposed interactive information visualization systems to link multivariate bio-image features to clinical factors [9]. Besides visualizing relationships between samples and their diagnoses, some researchers have illustrated the utility of spatial pattern visualization for multivariate bio-images [10, 11]. Researchers have also developed histological image retrieval systems that annotate image blocks with semantic labels such as necrosis, glands, and lymphocytes [12, 13]. These systems were used for analysis of rectangular histological image portions rather than WSIs.

The emergence of large data repositories such as the Cancer Genome Atlas (TCGA) has shifted the focus of morphological cancer analysis and digital pathology to WSIs. TCGA is a joint project by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that aims to accelerate the understanding of cancer in order to more effectively diagnose, cure and prevent cancer [14]. TCGA provides open access to high-quality genomic, proteomic and imaging data. The imaging data includes whole-slide tissue biopsy samples for several types of cancers. Recently, several researchers have used TCGA data for discovering relevant morphological properties associated with clinical diagnoses. Using ovarian serous carcinoma, Liu et al. studied the relationship between gene expression profiles, morphological properties of histological images, and chemotherapy response [15]. Soslow et al. established morphological properties associated with the BRCA1 and BRCA2 genotypes in ovarian serous carcinoma [16]. Cooper et al. discovered patient clusters in glioblastoma samples based on morphological features that have different prognoses [17]. Kong et al. associated morphological features to genomic subtypes in glioblastoma samples [18]. Chang et al. also established new subtypes in glioblastoma based on morphological features and associated them to differentially expressed genes [19]. They also developed BioSig, a system that allows the user the zoom and pan TCGAs WSIs similar

to Google maps [19]. However, these systems do not have provisions for exploring quantitative image features across WSIs.

We propose a framework for studying visual morphological patterns across whole-slide images in order to address challenges related to biological interpretation of image features and to noise in heterogeneous WSIs that affects diagnostic accuracy. Using this framework, we demonstrate that (1) common histopathological image features are qualitatively associated with biologically interesting regions of WSIs, (2) sets of multiple image features can cluster sections of WSIs into biologically relevant regions, and (3) histopathology domain knowledge pertaining to ROIs can be translated to image features using supervised analysis. Our results indicate that this visualization and analysis framework is useful for discovering informative image features, understanding their biological relevance, and guiding identification of ROIs in WSIs (e.g., regions with cancer cells or regions of necrosis). In the following section, we describe the TCGA image data, image processing and feature extraction, and three visualization-based analysis methods. In the results section, we show how these three visualization-based analysis methods can identify useful biological information or provide useful clinical functionality.

METHODS

Data

We use 1301 hematoxylin and eosin-stained WSI samples of ovarian serous carcinoma (OvCa) from 571 patients provided by TCGA [14]. For each patient, cancer grade, cancer stage and prognosis data is available. For each WSI, information about percent of necrosis, stromal cells, tumor cells and normal cells is available. WSIs from TCGA are available at 4 different resolutions. We use highest and lowest (thumbnail) resolution data for our analysis.

Figure 1 is an example of an OvCa WSI. Typical TCGA WSIs are approximately 20,000×40,000 pixels. Moreover, these WSIs usually contain a large amount of blank space surrounding the tissue specimen as well as artifacts that include (1) dark regions caused by folding of tissue slices when they were placed on glass slides, and (2) pen marks used by pathologists while they were studying the slide. These artifacts must be flagged prior to analysis of the meaningful portions or ROIs of each image.

Image Processing

In order to efficiently identify ROIs in the large TCGA WSIs we crop $H \times W$ -pixel WSIs into a matrix of non-overlapping, 512×512 -pixel tiles such that there are M rows and N columns in the tile matrix, where $M = \lfloor H/512 \rfloor$, and $N = \lfloor W/512 \rfloor$. We propose methods to flag image artifacts, and extract ROIs from each WSI. We then extract image features from each ROI tile.

ROI Selection—We perform ROI selection in HSV-color space. ROI selection is a two step process. First, we identify tiles in the tissue region of a WSI using the following three conditions: (1) Percent blank (white or gray) space is less than 80%, where a pixel (i, j) in a

binary blank mask, w , is given by
$$w(i, j) = \begin{cases} 1 & (I(i, j) \geq 0.1) \cap (S(i, j) \leq 0.1) \\ 0 & \text{o.w.} \end{cases}; (2)$$

Percent pen-mark (bluish-green) is less than 5%, where a pixel (i, j) in a binary pen-mark mask, p , is given by

$$p(i, j) = \begin{cases} 1 & [(0.4 < H(i, j) < 0.7) \cap (S(i, j) > 0.1)] \cup (I(i, j) < 0.1) \\ 0 & \text{o.w.} \end{cases}; (3)$$

Percent tissue is greater than 50%, where a pixel (i, j) in binary tissue mask, t , is given by

$$t(i, j) = \begin{cases} 0 & w(i, j) \cup p(i, j) \\ 1 & \text{o.w.} \end{cases}. \text{ Where } H(i, j), S(i, j), \text{ and } I(i, j) \text{ are hue, saturation and intensity of the pixel.}$$

Next, among the selected tissue tiles, we select ROI tiles that contain less than 10% tissue fold artifacts. We detect tissue fold artifacts from the lowest resolution image provided by TCGA Tissue folds regions have high saturation and low intensity [20]. Therefore, we calculate $\Delta = S(i, j) - I(i, j)$ for WSI pixels in each tissue tile and segment the image by identifying regions with $\Delta > 0$ and pixel area greater than five (to avoid noise). All of these regions are included in tissue folds. We then find pixels that are within two pixels of tissue folds (measured using the l_1 norm) and that have $\Delta > -0.3$. Based on visual analysis of the results, we observed that this tissue fold detection method has low false negatives. However, in some WSIs, regions with strongly stained basophilic cytoplasm appear as tissue folds, leading to false positives. In this study, we consider these regions to be folds.

Image Feature Extraction—We extract nine image-feature subsets, including 461 features, from each tile in the ROI of a WSI (Table 1). These feature subsets capture important morphological properties of a histopathological image such as color, texture, shape and topology [21]. We address color batch-effects in samples by color normalizing each tile to a standard H&E reference, and then extract color features [22]. We convert the normalized image into a grayscale image and extract global texture features. We segment each tile image into eosinophilic, basophilic, no-stain and red-blood-cell regions using color segmentation [22]. We then extract shape properties of eosinophilic, basophilic, and no-stain objects. We also extract grayscale texture of eosinophilic and basophilic regions. Basophilic structures often have nuclear clusters with overlapping nuclei [23, 24]. We segment individual nuclei from the basophilic mask and extract shape and topological features.

Visual Analysis

Visualizing Single Feature Variations—We visualize spatial patterns in WSIs formed by single image features in order to determine their biological relevance. By overlaying a heat map on a WSI thumbnail such that the heat map colors correspond to feature values, we can visually associate the feature values with morphological patterns. We map image feature values to heat map colors by discretizing the feature space into 20 equal-sized bins and use two methods for selecting the range of feature values: image-specific and dataset-specific. The image-specific feature range is calculated based on the distribution of feature values across all tiles of the WSI. Let R_n^m be the distribution of a feature, m , across all of the tiles in

a WSI sample, n . Then the dynamic range of R_n^m is given by a lower limit, L_n^m , and an upper limit, U_n^m , which are defined as follows:

$$\begin{aligned} L_n^m &= \max \{ \min (R_n^m), f_{0.25} (R_n^m) - 1.5 * IQD (R_n^m) \} \\ U_n^m &= \min \{ \max (R_n^m), f_{0.75} (R_n^m) + 1.5 * IQD (R_n^m) \} \end{aligned}$$

Where function $f_p(R)$ is the p^{th} percentile of distribution R , and the interquartile distance, IQD , is given by $IQD(R) = f_{0.75}(R) - f_{0.25}(R)$.

Similarly, for the dataset-specific range L_m and U_m for each feature, m , are the lower and upper limits of the feature across all tiles in all WSI samples. Both methods for determining heat map range are useful. The image-specific range is useful for studying spatial patterns within individual image samples while the dataset-specific range is useful for studying spatial patterns in images with respect to the entire dataset. Figure 2 illustrates the variation of a single image feature (median Delaunay triangle area) across a WSI using the image-specific range (Figure 2.A) and dataset-specific range (Figure 2.E). This feature captures the topology/architecture of histopathology images. As such, it is lower in the region of the WSI with cohesive lymphocytes (Figure 2.A). In Figure 2, the image-specific range of the median-Delaunay-area feature is 146–567, which is the upper part of the dataset-specific range (40–454) of this feature. Therefore, the visualization in Figure 2.E is mostly red. Biologically, we observe that this WSI sample has Cribriform/pseudoendometrioid architecture (Figure 2.D). Thus, nuclei in this WSI are farther apart compared to nuclei across the whole OvCa dataset, which tends to exhibit other cohesive architectures such as papillary [16].

Unsupervised Multi-Dimensional Clustering—We study spatial patterns formed by clustering tiles in a multi-dimensional space defined by combinations of feature subsets (listed in Table 1). We use agglomerative hierarchical clustering with Ward’s linkage to cluster both image features and tiles in WSIs [25]. Ward’s linkage minimizes the increase of the within-cluster sum of squares as a result of merging two clusters. The increase in sum of squares by merging clusters k and l is measured by the following distance metric:

$$d(k, l) = \sqrt{\frac{2n_k n_l}{n_k + n_l}} \|\bar{\mathbf{a}}_k - \bar{\mathbf{a}}_l\|_2,$$

where n_k is the number of tiles in cluster k , $\bar{\mathbf{a}}_k$ is the centroid of cluster k , $\|\cdot\|_2$ and is the Euclidian distance.

We visualize clustering by highlighting similar tiles with the same color and generating a heat map to show the grouping of both image features (rows) and tiles (columns) (see Figure 5.B and 5.C). By studying both the WSI with highlighted tiles and the hierarchical clustering heat map, we can associate ROIs in the WSI to image feature groups. The number of clusters for visualization is a variable parameter and we observed that up to six clusters were sufficient for discovering meaningful ROIs.

2.3.3 Supervised Classification—Studying the variations of single image features and the clusters formed by multi-dimensional image features is useful for interpreting the biological relevance of these features. However, it is also important to link histopathology domain knowledge to image features. In order to do this, we start by selecting two biologically different regions in the WSI, and develop a mathematical model using supervised machine learning methods to distinguish these two regions. We then use the predictive model to classify the remaining WSI tiles and visualize the result by highlighting tiles according to the probability of belonging to a selected region (Figure 3). This visualization is useful for (1) identifying image features associated with biologically distinct WSI regions and, based on these features, for (2) identifying image tiles that are similar to these regions. We use the mRMR-q (minimum-redundancy maximum-relevance) feature selection method for ranking features [26]. We then use the top features to develop a classification model using the linear-SVM classifier (LibSVM) [27]. We optimize the number of features used in the classification model using 5-fold, 10-iterations of cross-validation accuracy. The model may be incrementally trained by iteratively adding training samples (from different WSIs) to provide the classifier with more challenging training samples [28]. In Figure 3, we illustrate the selection of tumor (green) and non-tumor (red) regions of a WSI and show the result of classifying the remaining image tiles by visualizing the probability of each tile belonging to the non-tumor region (Figure 3.B, red indicates higher probability).

RESULTS AND DISCUSSION

We studied OvCa WSIs using three visualization methods and found several biologically interpretable morphological patterns. In this section, we discuss an example pattern for each visualization method.

Average Basophilic-Object Eccentricity Differentiates Stroma/Necrosis, Tumor, and Lymphocyte-Infiltrated Tumor Regions

We studied spatial patterns captured using average basophilic-object (i.e., nuclear object) eccentricity (*ec*), a morphological shape property. Using the dataset-specific range of values for the visualization heat map, we observed that image tiles in the stroma/necrosis region have high *ec* while tiles with lymphocytes have low *ec*. This is due to the spindle-like nuclei of stroma/necrosis regions and the circular disc-like structures of lymphocytes [29]. Nuclei in tumor cells have intermediate eccentricity. Figure 4 is an illustration of spatial patterns formed by *ec* for two WSI samples. Both samples are from patients with grade-3 OvCa but have different morphology. The first sample (Figure 4.A–4.E) is reported to have 3% lymphocytes, 15% necrosis, 0% stroma, and 95% tumor, while the second sample (Figure 4.G–4.K) is reported to have 60% lymphocytes, 0% necrosis, 5% stroma, and 95% tumor. We observed that: (1) due to the higher percentage of lymphocytic infiltration, the visualization of the second sample contains more green tiles (Figure 4.K), (2) tiles in stroma/necrosis regions are highlighted with shades of red, (3) high resolution images confirm that tiles with low, medium and high *ec* are from stroma/necrosis, tumor and lymphocytic-infiltrated tumor regions, respectively. Lymphocytic infiltration is considered useful for

predicting patient prognosis [29] and has been correlated to the BRCA1 genotype in OvCa [16].

Color and Basophilic-Object Shape Features Reveal Three Distinct Clusters

We studied spatial patterns formed by the morphological color and basophilic-object shape feature subsets that include 124 features. Using Ward's linkage with agglomerative clustering (as described in section 2.3.2), we illustrate the spatial patterns formed by three clusters of tiles in a grade-3 sample reported to have 15% necrosis, 15% stroma, and 70% tumor (Figure 5.A and 5.B). Figures 5.D–5.F are high resolution tiles from each of the three clusters, capturing tumor (yellow), necrosis (magenta), and stroma (cyan) regions, respectively. Out of 1174 tiles in the sample, 603, 316, and 255 tiles were clustered into each of these clusters (C1, C2, and C3), respectively. The heatmap (Figure 5.C) shows that there are several correlated features in these subsets that separate these clusters. We use the mRMR-q feature selection method to find the top five most informative, uncorrelated features in these subsets that will separate C1, C2 and C3. In Table 2, we report these features (in the order of preference) with their average value and standard deviation for each cluster. These features can be easily visually associated with morphological properties of tumor, necrosis, and stroma. Due to the red appearance of stroma/necrosis regions compared to tumor regions, color features are important. Moreover, as discussed in the previous section, stroma/necrosis regions have spindle-like nuclei. Therefore, C2 and C3 have larger standard deviation in major-axis length and larger maximum eccentricity of basophilic structures. Due to the more elliptical nuclei in tumor regions, the mean boundary-fractal dimension is higher than that of stroma and necrosis. Therefore, using this visualization we can discover morphological patterns in WSI based on multiple features.

Identification of Tumor and Non-Tumor ROIs in WSIs

We developed a classification model that separates tumor and non-tumor (including stroma and necrosis) in OvCa WSIs. We selected this model because pathologists generally focus on tumor regions, excluding necrotic and stroma regions, while making histological diagnoses. Therefore, pathologists can use this model to facilitate selection of ROIs in WSIs. For this model, we performed training using tumor and non-tumor regions from 15 WSIs. On each WSI, a pathologist marked tiles in both non-tumor and tumor regions (Figure 3.A). In total, 2098 tiles were used for training, among which 638 and 1460 tiles were from non-tumor and tumor regions, respectively. We performed cross-validation to select an appropriate feature size (from 1 to 20 top features out of 461 features). Based on average classification accuracy, we selected 17 as the optimal feature size. Table 3 lists the top 5 features. Similar to Table 2, there are some color and basophilic-object shape properties that were useful for classifying tumor and non-tumor regions. In addition, energy of Gabor filter response is higher for tumor regions than non-tumor regions. Gabor filter response is high if there are edges in the image in the $\theta + \pi/2$ direction. Higher Gabor filter response in the tumor region is mostly due to stronger edges formed in the grayscale image by basophilic-structures compared to eosinophilic-structures. Also, mean intensity of eosin stained pixels is slightly higher (brighter/whiter) in non-tumor compared to tumor.

We validate this classification model on a separate set of 100 OvCa WSIs. Again, tiles in the tumor and non-tumor regions of these images were marked by a pathologist. In total, there were 29424 tiles in the testing dataset with 17181 tumor and 12243 non-tumor tiles. Overall classification accuracy on the testing dataset was 90%. In Table 4, we have provided a confusion matrix for the classification model. Most of the error is due to false negatives, i.e., our system classifies borderline cases as non-tumor rather than tumor. From the point of view of highlighting useful regions for diagnosis, false negatives are more favorable than false positives.

We also apply this model to all tumor samples in the OvCa TCGA dataset (including the training samples) that have reported percentage of tumor cells (in total 1094 WSIs). Then we calculated the approximate percentage of tumor cells in the WSI using percent of tiles in the tumor region. We included tiles in the tissue fold regions of WSIs on the assumption that reported percentages of tumor cells are with respect to the entire samples. We found that the predicted percentage of tumor cells was correlated with TCGA-reported tumor cell percentage (Pearson's correlation coefficient = 0.4293, p-value = 2.8134e-50). Figure 6 is a scatter plot of reported percentage of tumor cells vs. predicted percentage of tiles in the tumor region. Data points are colored based on the density of points in the region. A large number of points are on the diagonal (high correlation). The points that are not correlated are mostly below the diagonal, indicating that our system tends to classify more tiles into the non-tumor region compared to what is reported by TCGA.

CONCLUSION

We have proposed a visualization framework for studying morphological patterns across histopathological whole-slide images (WSIs). This visualization highlights variations across regions in WSIs based on one or more quantitative image features. The visualization can be used to study patterns formed by features (single feature variations and unsupervised multi-dimensional clustering) or it can be used to identify features that capture histopathology domain knowledge (supervised classification). Our results indicate that these visualization techniques are useful for semantic interpretation of quantitative image features. These visualizations can address the challenge of translating computer-aided decision support systems to clinical practice by reducing the semantic gap between quantitative image features and histopathology domain knowledge. Although we have used ovarian serous cystadenocarcinoma (OvCa) WSIs provided by TCGA, our methods can be applied to any WSI dataset.

Acknowledgments

This research was supported by grants from the National Institutes of Health (R01CA108468, P20GM072069, and U54CA119338) to M.D.W., Georgia Cancer Coalition (Distinguished Cancer Scholar Award to M.D.W.), Hewlett Packard, and Microsoft Research.

References

1. Liu T, Peng H, Zhou X. Imaging informatics for personalised medicine: applications and challenges. *Int J Funct Inform Personal Med*. 2009; 2(2):125–135. [PubMed: 19862353]

2. Gabril MY, Yousef GM. Informatics for practicing anatomical pathologists: marking a new era in pathology practice. *Mod Pathol*. 2010; 23(3):349–358. [PubMed: 20081805]
3. Rocha R, Vassallo J, Soares F, et al. Digital slides: present status of a tool for consultation, teaching, and quality control in pathology. *Pathol Res Pract*. 2009; 205(11):735–741. [PubMed: 19501988]
4. Cruz-Roa A, Caicedo JC, González FA. Visual pattern mining in histology image collections using bag of features. *Artif Intell Med*. 2011; 52(2):91. [PubMed: 21664806]
5. Liu CL, Prapong W, Natkunam Y, et al. Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays. *American Journal of Pathology*. 2002; 161(5):1557. [PubMed: 12414504]
6. Lessmann B, Nattkemper TW, Hans VH, et al. A method for linking computed image features to histological semantics in neuropathology. *J Biomed Inform*. 2007; 40(6):631–641. [PubMed: 17698418]
7. Iglesias-Rozas JR, Hopf N. Histological heterogeneity of human glioblastomas investigated with an unsupervised neural network (SOM). *Histol Histopathol*. 2005; 20(2):351–356. [PubMed: 15736037]
8. Lobenhofer EK, Boorman GA, Phillips KL, et al. Application of visualization tools to the analysis of histopathological data enhances biological insight and interpretation. *Toxicol Pathol*. 2006; 34(7):921–928. [PubMed: 17178692]
9. Herold J, Loyek C, Nattkemper Tim W. Multivariate image mining. *WIREs Data Mining Knowl Discov*. 2011; 1(1):2.
10. Loyek C, Rajpoot N, Khan M, et al. BioIMAX: A Web 2.0 approach for easy exploratory and collaborative access to multivariate bioimage data. *BMC Bioinformatics*. 2011; 12(1):297. [PubMed: 21777450]
11. Kölling J, Langenkämper D, Abouna S, et al. WHITE—a web tool for visual data mining colocation patterns in multivariate bioimages. *Bioinformatics*. Apr 15; 2012 28(8):1143–1150. 2012. [PubMed: 22390938]
12. Tang HL, Hanka R, Ip HHS. Histological image retrieval based on semantic content analysis. *Information Technology in Biomedicine, IEEE Transactions on*. 2003; 7(1):26–36.
13. Yu F, Ip Horace HS. Semantic content analysis and annotation of histological images. *Computers in Biology and Medicine*. 2008; 38(6):635–649. [PubMed: 18455157]
14. McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008; 455(7216):1061–1068. [PubMed: 18772890]
15. Liu Y, Sun Y, Broaddus R, et al. Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma. *PLoS One*. 2012; 7(5):e36383. [PubMed: 22590536]
16. Soslow RA, Han G, Park KJ, et al. Morphologic patterns associated with BRCA1 and BRCA2 genotype in ovarian carcinoma. *Mod Pathol*. 2012; 25(4):625–636. [PubMed: 22193042]
17. 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(2):317–323. [PubMed: 22278382]
18. Kong J, Cooper D, A L, Wang F, et al. Integrative, Multimodal Analysis of Glioblastoma Using TCGA Molecular Data, Pathology Images, and Clinical Outcomes. *Biomedical Engineering, IEEE Transactions on*. 2011; 58(12):3469–3474.
19. Chang H, Fontenay GV, Han J, et al. Morphometric Analysis of TCGA Glioblastoma Multiforme. *BMC Bioinformatics*. 2011; 12(1):484. [PubMed: 22185703]
20. Palokangas S, Selinummi J, Yli-Harja O. Segmentation of folds in tissue section images. :5641–5644.
21. Kothari S, Phan JH, Young AN, et al. Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer. :422–425.
22. Kothari S, Phan JH, Moffitt RA, et al. Automatic batch-invariant color segmentation of histological cancer images. :657–660.
23. Kothari S, Chaudry Q, Wang MD. Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques. :795–798.

24. Kothari S, Chaudry Q, Wang MD. Extraction of informative cell features by segmentation of densely clustered tissue images. :6706–6709.
25. Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*. 2001; 17(suppl 1):S22–S29. [PubMed: 11472989]
26. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005; 27:1226–1238. [PubMed: 16119262]
27. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2011; 2(3):27.
28. Doyle, S., Madabhushi, A. *Pattern Recognition in Bioinformatics*, Lecture Notes in Computer Science, pp. Springer Berlin; Heidelberg: 2010. *Consensus of Ambiguity: Theory and Application of Active Learning for Biomedical Image Analysis*; p. 313-324.
29. Busam KJ, Antonescu CR, Marghoob AA, et al. Histologic classification of tumor-infiltrating lymphocytes in primary cutaneous malignant melanoma. A study of interobserver agreement. *Am J Clin Pathol*. 2001; 115(6):856–860. [PubMed: 11392882]

Figure 1. Region-of-interest (ROI) selection in a TCGA WSI. Squares (any color) indicate image regions with significant tissue. Red squares indicate tissue fold artifacts and green squares indicate ROI tiles.

Figure 2.

Visualization of variations in a single image feature (median Delaunay area, i.e. nuclear-topology feature) across the tiles of a WSI. (A) Variations in the feature across the WSI using the image-specific range (146–577). (B) High resolution tile from the green region of the visualization in (A), indicating a low value for the feature that corresponds to more cohesive nuclear structures due to partial necrosis and lymphocytes. (C) Colormap for the visualization in (A). (D) High resolution tile from the red region of the visualization in (A). (E) Variations in the feature across the WSI using the dataset-specific range (40–454).

Figure 3. Visualization of Supervised Classification. (A) Supervised selection of non-tumor (red) and tumor (green) regions from a WSI sample. (B) Based on a prediction model derived from the supervised tile selection, remaining tiles are highlighted with the probability of being in the non-tumor class. (C) Color map for the visualization in (B).

Figure 4.

Variations in average nuclear eccentricity (ec) across two grade-3 ovarian serous cystadenocarcinoma WSI samples (B & K). High ec (red) corresponds to stroma or necrotic regions with elongated nuclei, low ec (green) corresponds to lymphocyte-infiltrated tumor regions with circular lymphocytes, and medium ec (yellow) corresponds to tumor regions. (A) WSI sample with 3% lymphocytes, 15% necrosis, 0% stroma and 95% tumor. (B) WSI sample with 60% lymphocytes, 0% necrosis, 5% stroma and 95% tumor. (F) Colormap corresponding to ec values. (C–I) High-resolution image tiles from sample A (C–E) and sample B (G–I). Numbers above the high resolution tiles indicate the feature value. These tiles are also outlined with a color corresponding to the highlighted tile in (B) and (K).

Figure 5.

Three distinct tile clusters (B) in a grade-3 ovarian serous cystadenocarcinoma WSI sample (A) with 15% necrosis, 15% stroma and 70% tumor. Clusters were found in the feature space defined by color and nuclear-object-shape features (a total of 124 features). (C) The heatmap illustrates the variations in features (row) across various image tiles (column). Red and green values in the heatmap correspond to values above and below the mean feature value. Out of 1174 tiles in the sample, 603, 316, and 255 tiles were distributed into clusters C1, C2, and C3, respectively. (D–F) High-resolution image tiles belonging to the three clusters. Clusters C1, C2 and C3 correspond to regions of tumor, necrosis and stroma, respectively.

Figure 6.

Correlation between reported percentage of tumor cells and predicted percentage of tiles in tumor region using a tumor vs. non-tumor classification model (same as Table 4). This plot includes all WSI TCGA samples that have percentage of tumor cell information available (1094 images). Tissue fold artifacts were included when calculating the percentage of tumor regions. Pearson's correlation coefficient = 0.4293 (p-value = 2.8134×10^{-50}). The color represents the density of data points per unit of area.

Table 1

Morphological image feature subsets.

Feature Subset	# of Features	Description
Color	73	RGB histograms, histogram statistics, and stain co-occurrence
Global texture	138	Haralick, gray-level histogram statistics, fractal, GHM multiwavelet, and Gabor
Eosinophilic-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Eosinophilic-region texture	18	Haralick and gray-level histogram statistics
No-stain-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Basophilic-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Basophilic-region texture	18	Haralick and gray-level histogram statistics
Nuclear shape	26	Count, elliptical area, major-minor axes lengths, eccentricity, and cluster size
Nuclear topology	35	Delaunay triangle, Voronoi diagram, minimum spanning tree, and closeness

Table 2

Top five features that separate the clusters C1, C2, and C3

Feature Name	C1	C2	C3
Percent of pixels in R histogram bin (80–96)	0.025 ± 0.004	0.017 ± 0.004	0.015 ± 0.004
Percent of pixels in G histogram bin (127–143)	0.052 ± 0.006	0.040 ± 0.006	0.058 ± 0.007
Std in major-axis length of basophilic structures	6.669 ± 2.049	8.048 ± 2.402	9.174 ± 2.505
Maximum eccentricity of basophilic structures	0.975 ± 0.009	0.971 ± 0.011	0.987 ± 0.007
Mean boundary-fractal dimension of basophilic structures	1.505 ± 0.010	1.486 ± 0.020	1.458 ± 0.019

Table 3

Top 5 features in tumor/non-tumor classification model

Feature Name	Tumor	Non-tumor
Adjacency of basophilic stained pixels per unit area	0.400 ± 0.080	0.277 ± 0.091
Energy of Gabor filter response with frequency, $f = \sqrt{2}/64$ cycles per pixel, and orientation, $\theta = \pi/2$ radians.	57.103 ± 7.564	51.662 ± 10.845
Percent of pixels in B histogram bin (191–207)	0.059 ± 0.013	0.053 ± 0.023
Eosin stained pixels' mean intensity	0.523 ± 0.016	0.536 ± 0.021
Mean boundary fractal of basophilic structures	1.491 ± 0.014	1.472 ± 0.018

Table 4

Confusion matrix for tumor/non-tumor model

		Predicted label		
		Tumor	Non-tumor	
Actual Label	Tumor	87%	13%	Accuracy= 90%, Number of tumor tiles=17181, Number of non-tumor tiles=12243
	Non-tumor	6%	94%	