Computational staining of pathology images to study tumor microenvironment in lung cancer

Shidan Wang[1], Ruichen Rong[1], Donghan M. Yang[1], Ling Cai[1], Lin Yang[1, 3], Danni Luo[1], Bo Yao[1], Lin Xu[1], Tao Wang[1], Xiaowei Zhan[1], Yang Xie[1, 5, 6], Adi Gazdar[6,7, 8], John Minna[6,8,9] and Guanghua Xiao[1, 5, 6,*]

[1]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, 75390, USA

[3]Department of Pathology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China

[4]Children's Medical Center Research Institute at UT Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX, 75390 USA

[5]Department of Bioinformatics, UT Southwestern Medical Center, Dallas, Texas, 75390, USA

[6]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas, 75390, USA

[7]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas, 75390, USA

[8]Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center, Dallas, Texas, 75390, USA

[9]Departments of Internal Medicine and Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas, 75390, USA

Correspondence to:

Guanghua Xiao, Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Dallas, Texas, 75390, USA. Email: Guanghua.Xiao@utsouthwestern.edu

Tel: +1-214-648-4110

ABSTRACT

The spatial organization of different types of cells in tumor tissues reveals important information about the tumor microenvironment (TME). In order to facilitate the study of cellular spatial organization and interactions, we developed a comprehensive nuclei segmentation and classification tool to characterize the TME from standard Hematoxylin and Eosin (H&E)-stained pathology images. This tool can computationally "stain" different types of cell nuclei in H&E pathology images to facilitate pathologists in analyzing the TME.

A Mask Regional-Convolutional Neural Network (Mask-RCNN) model was developed to segment the nuclei of tumor, stromal, lymphocyte, macrophage, karyorrhexis and red blood cells in lung adenocarcinoma (ADC). Using this tool, we identified and classified cell nuclei and extracted 48 cell spatial organization-related features that characterize the TME. Using these features, we developed a prognostic model from the National Lung Screening Trial dataset, and independently validated the model in The Cancer Genome Atlas (TCGA) lung ADC dataset, in which the predicted high-risk group showed significantly worse survival than the low-risk group (pv= 0.001), with a hazard ratio of 2.23 [1.37-3.65] after adjusting for clinical variables. Furthermore, the image-derived TME features were significantly correlated with the gene expression of biological pathways. For example, transcription activation of both the T-cell receptor (TCR) and Programmed cell death protein 1 (PD1) pathways was positively correlated with the density of detected lymphocytes in tumor tissues, while expression of the extracellular matrix organization pathway was positively correlated with the density of stromal cells.

This study developed a deep learning-based analysis tool to dissect the TME from tumor tissue images. Using this tool, we demonstrated that the spatial organization of different cell types is predictive of patient survival and associated with the gene expression of biological pathways. Although developed from the pathology images of lung ADC, this model can be adapted into other types of cancers.

1 INTRODUCTION

With the advance of technology, Hematoxylin and Eosin (H&E)-stained tissue slide scanning has become a routine clinical procedure, which produces pathology images that capture histological details in high resolution. Pathology images of tumor tissues contain not only essential information for tumor grade and subtype classifications[1], but also information on the tumor microenvironment (TME), such as the spatial organization of different types of cells. Cell spatial organization reveals cell growth patterns and the spatial interactions among different types of cells, which provide important insights into tumor progression and metastasis. A recent study by Cheng et al [2] developed an algorithm to segment the cell nucleus and extract the topological features for TME analysis in renal cell carcinoma (RCC), which improved the understanding of cell spatial organization and patient outcome in RCC. However, this study used an unsupervised approach that assigns an identified cell nucleus to one of the clusters without a clear definition, which hampered interpretation of the results. Recent studies[3, 4] showed that the spatial organization and architecture of tumor-infiltrating lymphocytes (TIL) play important roles in the TME. However, these studies focused solely on recognizing lymphocytes and ignored other types of cells, which greatly limited exploration of the interactions among different types of cells. In this study, we developed a deep learning-based algorithm to examine standard H&E pathology images to automatically segment and classify different types of cell nuclei. This algorithm can be used as a tool to "computationally stain" different types of cell nuclei, in order to facilitate pathologists in examining tissue images and researchers in studying the TME from these standard clinical materials.

The major cell types in a malignant tissue include tumor cells, stromal cells, lymphocytes, and macrophages. Stromal cells are mainly connective tissue cells such as fibroblasts and pericytes. The interactions between tumor cells and stromal cells play an important role in cancer progression[5-7] and metastasis inhibition[8]. Since the cell boundaries of tumor cells and stromal cells are often unclear in standard H&E stained lung cancer pathology images, we segmented and classified cell nuclei instead of whole cells. TIL are mainly white blood cells that have migrated into a tumor region. They are a mix of different types of cells, in which T cells are the most abundant population. The spatial organization of TIL has been associated with patient outcome and molecular profiles in multiple tumor types[9-12]. Macrophages are inflammatory cells, and inflammation in tumor niches has been reported as a prognostic marker and correlated with tumor progression[13-15]. Other tissues and cellular structures existing in the TME include blood vessels and necrosis. In this study, blood cells and karyorrhexis are segmented to represent blood vessels and necrosis, respectively, in order to quantify blood vessels and necrosis and study their interactions with tumor cells, stromal cells, lymphocytes and macrophages.

In this study, we developed a deep learning algorithm based on the Mask Regional Convolutional Neural Network (Mask-RCNN) architecture[16]. We trained this Mask-RCNN model using pathology images of lung adenocarcinoma (ADC) patients from the National Lung Screening Trial (NLST) study with the nuclei of tumor cells, stromal cells, lymphocytes, macrophages, blood cells and karyorrhexis manually labeled by expert pathologists. The model accuracy was validated in a different set of images. From the identified cell types and cell spatial locations, we derived cell spatial organization features to characterize the TME. In our analysis, we found that these TME-related image features were significantly associated with patient overall survival. Based on these image features, a prognostic model for lung ADC patients was developed from the NLST dataset. This model was independently validated in the pathology image data from The Cancer Genome Atlas (TCGA) lung ADC (LUAD) dataset, in which the predicted high-risk group showed significantly worse survival than the low-risk group (pv= 0.001), with a hazard ratio of 2.23 [1.37-3.65] after adjusting for clinical variables.

Furthermore, the image-derived TME features were significantly correlated with the gene expression of biological pathways. For example, transcription activation of both the T-cell receptor (TCR) and Programmed cell death protein 1 (PD1) pathways was positively correlated with the density of detected lymphocytes in tumor tissues, while expression of the extracellular matrix organization pathway was positively correlated with the density of stromal cells.

In this study, we developed a Mask-RCNN algorithm for nuclei segmentation and cell classification as a tool to study the tumor morphological microenvironment using tissue pathology images in lung ADC. In order to facilitate usage of this deep learning tool, a user-friendly web portal has been developed and can be accessed at http://lce.biohpc.swmed.edu/maskrcnn/analysis.php. Although this tool was developed in lung ADC pathology images, our results showed that the Mask-RCNN method can be adapted and applied in head and neck cancer, breast cancer and lung cancer squamous cell carcinoma pathology image datasets. The web portal also provides functions to facilitate researchers in adapting the Mask-RCNN model for other types of cancers.

2 METHODS

2.1 Dataset

Pathology images that support the findings of this study are available online in NLST (https://biometry.nci.nih.gov/cdas/nlst/) and The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD, https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD). mRNA expression data for the TCGA dataset are available online at http://firebrowse.org. The H&E-stained pathology images together with the corresponding clinical data were obtained from the NLST and TCGA Lung ADC cohorts: 208 40X pathology images for 135 lung ADC patients were acquired from the NLST dataset, and 431 40X pathology images for 372 lung ADC patients were acquired from the TCGA LUAD dataset (there could be multiple pathology images for a single patient). A specialized lung cancer pathologist, Dr. Lin Yang, labeled the Region of Interest (ROI) for each of the pathology images (Figure 1). Another lung cancer pathologist, Dr. Adi Gazdar, confirmed the labelling. Clinical characteristics of the patients in this study are summarized in **Supplemental Table 1**.

2.2 Nuclei segmentation using Mask-RCNN

2.2.1 Training, validation, and testing sets preparation

In order to construct the training set for the Mask-RCNN algorithm, 127 image patches (500 × 500 pixels) from 39 pathological ROIs (Figure 1) were extracted from the NLST dataset. In these patches, different types of cell nuclei were labeled. All the pixels with tumor nuclei, stromal nuclei, lymphocyte nuclei, macrophage nuclei, red blood cells, and karyorrhexis were labeled according to their categories and all the remaining pixels were considered “other”. These labels, also collectively called the “mask”, were then used as the ground truth to train the Mask-RCNN model. The labeled images were randomly divided into training, validation, and testing sets. To ensure independence among these datasets, image patches from the same ROI were assigned together. More than 12,000 cell nuclei were included in the training set, while 1227 and 1086 nuclei were included in the validation and testing sets, respectively.

2.2.2 Training process

A deep learning model was developed using the Mask-RCNN architecture. The pre-trained model was fine-tuned on our training dataset from the NLST study. Images were standardized (centered and scaled to have zero mean and unit variance) for each RGB channel. To increase generalizability and avoid bias from different H&E staining conditions, we performed extensive augmentations on the image patches. In particular, random projective transformations were applied to images and their corresponding masks; each image channel was randomly shifted using linear transformation. For the training process, the batch size was set to 2, the learning rate was set to 0.01 and decreased to 0.001 after 500 epochs, the momentum was set to 0.9, and the maximum number of epochs to train was set to 1000. In the validation set, the model trained at the 707th epoch reached the lowest loss. This model was selected and used in the following analysis to avoid overfitting. Python (version 3.5.2) and python libraries (Keras, version 2.1.5; openslide-python, version 1.1.1; tensorflow-gpu, version 1.8.0) were used[17].

2.2.3 Segmentation performance evaluation

Since the Mask-RCNN model simultaneously segments and classifies cell nuclei, three criteria were used to evaluate the segmentation performance in the validation and testing datasets, respectively. First, detection coverage was calculated as the ratio between the detected nuclei and the total ground truth nuclei. Each ground truth nucleus was matched to a segmented nucleus, which generated the maximum Intersection over Union (IoU). If the IoU for a ground truth nuclei was > 0.5, this nuclei was labeled “matched”; otherwise it was labeled “unmatched”. Second, nuclei classification accuracy was determined for the matched nuclei by comparing the predicted nucleus type with the ground truth. Third, segmentation accuracy was evaluated by the IoUs, which were calculated for each detected nucleus and averaged in different nuclei categories.

2.3 Image feature extraction to describe nuclei composition and organization

In order to make the Mask-RCNN model more computationally efficient while retaining a good representation of each ROI, instead of applying the model to the whole slide, 100 image patches (1024 × 1024 pixels) were randomly sampled and analyzed for each pathologist-labeled ROI **(Supplemental Figure 1)**. These 100 image patches provided good coverage of each ROI. Nuclei were then segmented and classified through the Mask-RCNN model developed from this study **(Figure 1)**. In order to characterize the spatial organization of cells using a graph, we calculated the centroids of nuclei and used them as vertices to construct a Delaunay triangle graph for each image patch. The Delaunay triangle graph connects nuclei into a graph, and the number of connections and the average length (i.e. spatial distance) between two types of nuclei summarize the spatial organization of different types of cell. Since 6 nucleus categories were included in this study, the edges of the graph were classified into 21 categories [i.e. $6 \times (6-1)/2 = 21$] according to their vertex pairs. For each image patch, the number of connections (i.e. edges) for different categories was counted (which added up to 21 features), the lengths of the connections were averaged for each edge category (yielding another 21 image features), and the density of each type of nucleus was calculated (yielding 6 image features). In total, 48 image features were extracted. The image features were averaged across the 100 patches for each ROI in the pathology image. When 2 or more pathology slides were available for 1 patient, the features from the slides were averaged for each patient. Thus, in total 48 image features were extracted for each patient, in both the NLST and TCGA datasets.

2.4 Prognostic model development and validation

Overall survival, defined as the date of diagnosis till death or last contact, was used as the response variable for survival analyses. A prognostic model for overall survival was developed in the lung ADC patients in the NLST dataset and independently validated on the TCGA LUAD dataset. Given a set of image-derived TME features for each patient, the prognostic model will assign a risk score for the patient, with a higher risk score indicating worse prognosis. Based on the risk scores, the patients in the TCGA LUAD cohort were dichotomized into predicted high- and low-risk groups using the median risk score as a cutoff. The survival curves of the predicted high- and low-risk groups were estimated using the Kaplan-Meier method. The survival differences between predicted high- and low-risk groups were compared using a log-rank test. A multivariate Cox proportional hazard model was used to determine the prognostic value of predicted risk groups (using image-derived TME features) after adjusting for other clinical characteristics, including age, gender, smoking status, and stage. R software, version 3.4.2, and R packages (survival, version 2.41-3; glmnet, version 2.0-13; spatstat, version 1.55-1) were used[18, 19]. The results were considered significant if the two-tailed p value <0.05.

2.5 Association analysis between image features and gene expression of biological pathways.

Gene expression data of 372 patients from the TCGA LUAD dataset were downloaded and preprocessed. The correlation between mRNA expression levels and image-derived TME features was evaluated using Spearman rank correlation. Gene set enrichment analysis (GSEA) was performed for each TME feature. All gene sets from the Reactome database were used[20]. For multiple testing correction, Benjamini-Hochberg (BH)-adjusted p values were used to detect significantly enriched gene sets. Gene sets with BH-adjusted two-tailed p values < 0.05 were regarded as significantly enriched. R packages Hmisc (version 4.1-1), fgsea (version 1.4.1), and gplots (version 3.0.1) were used[21].

3 RESULTS

3.1 Mask-RCNN simultaneously and accurately classifies and segments cell nuclei

The developed Mask-RCNN model segments and classifies individual nuclei at the same time **(Supplemental Figure 2)**. **Figure 2** demonstrates some of the segmentation and classification results. In total, the segmented cell nuclei were classified into six categories: tumor cell, stromal cell, lymphocyte, macrophage, red blood cell, and karyorrhexis, and all the remaining structures or spaces were considered background. Different nuclei were colored according to the predicted categories **(Figure 2)**. For detected objects, the overall classification accuracy was 85% and 85% in the validation set and the testing set, respectively, while the accuracy for tumor nuclei was 88% in validation and 90% in testing, respectively **(Supplemental Figure 3)**. It is noteworthy that the developed Mask-RCNN model can be applied to the entire digital pathology image to generate a cell spatial organization map across the whole slide, where tumor region and lymphocyte infiltration areas are clearly illustrated **(Figure 3)**.

3.2 Prognostic value of nuclei composition and organization in the TME

A Delaunay triangle graph was constructed for each image patch[2] to extract topological features from nuclei spatial organization to characterize the TME **(Supplemental Figure 4)**. The nuclei and edges were counted and edge lengths averaged to yield 48 image features (see **Methods Section**). **Supplemental Table 2** summarizes the TME features that were significantly correlated with survival outcome in univariate analysis. It shows that higher karyorrhexis density, more karyorrhexis-karyorrhexis connections and more karyorrhexis-red blood cell connections were associated with worse survival outcome, which was expected as these features indicate a higher rate of tumor necrosis. Furthermore, higher stromal nuclei density and more stromal-stromal connections were associated with better survival outcome, which agreed with our current knowledge that more stromal tissues corresponds to better prognosis.

A prognostic model based on the image features was developed in the NLST dataset and then independently validated in the TCGA LUAD dataset. **Figure 4A** shows the survival curves of the predicted high and low-risk groups in the TCGA cohort, where the patients in the predicted high-risk group show significantly worse survival than those in the predicted low-risk group (log-rank test, p value = 0.0011). Furthermore, the risk group defined by the TME features serves as an independent prognostic factor (high- vs. low-risk, Hazard Ratio = 2.23, 95% Confidence Interval = 1.37-3.65, and p value=0.0013), after adjusting for clinical variables, including age, gender, smoking status, and stage **(Figure 4B)**.

3.3 Association between image features and transcriptional activity of biological pathways

GSEA was performed to identify the biological pathways whose mRNA expression profiles were significantly correlated with image-derived TME features in the TCGA dataset. **Figure 5A-D** shows examples of these biological pathways. For example, the transcription activation of both the T-cell receptor (TCR) and Programmed cell death protein 1 (PD1) pathways was positively correlated with lymphocyte density in the tumor tissue **(Figure 5A)**. This observation is consistent with previous reports that genes involved in the TCR and PD1 pathways are expressed in immune cells[22, 23]. In addition, expression of the extracellular matrix organization gene set, for which fibroblasts act as an important source[24], was positively correlated with stromal cell density in tumor tissue **(Figure 5B)**. In a negative control experiment where we randomly shuffled the patient IDs and repeated the same analysis, such correlation was no longer observed **(Supplemental Figure 5)**.

Furthermore, GSEA analysis showed that the cell cycle pathway was significantly enriched with genes whose expression levels were correlated with both the tumor nuclei density **(Figure 5C)** and karyorrhexis density in tumor issue (**Figure 5D)**. To look into the relationship between tumor cell density and the gene expression of the cell cycle pathway, we grouped and sorted the patients in the TCGA LUAD cohort according to their tumor nuclei density. **Figure 5E** shows, for each patient group in the TCGA LUAD dataset, the average expression levels of genes within the cell cycle pathway and whose expression levels are significantly (p value <0.001) correlated with tumor nuclei density. Positive correlations between gene expression and tumor nuclei density can be observed for most of the cell cycle-related genes, except for one gene, POLD4, which showed an inverse trend. Most of the genes in the cell cycle pathway have higher expression in tumors with higher tumor nuclei density (may be a higher grade of tumor), while POLD4 shows the opposite pattern. This pattern of POLD4 compared with other genes in the cell cycle gene set is consistent with a previous study of lung cancer[25]: while most cell cycle genes were upregulated in lung cancer, POLD4 is usually downregulated.

3.4 Webserver for publically accessible pathological image segmentation model

In order to facilitate usage of the Mask-RCNN model developed in this study, we also developed an online tool (http://lce.biohpc.swmed.edu/maskrcnn/analysis.php) for this deep learning-based nuclei segmentation and classification model **(Figure 6)**. This tool requires only a pathology image (or a patch from the image) as the input **(Figure 6A)**. Each uploaded input image will be assigned a job ID

(Figure 6B). The segmentation results will be automatically displayed and the spatial coordinates of each nucleus can be downloaded as an Excel table **(Figure 6C)**. In order to assist researchers in using this tool to study TME-related features for other cancer types, we also provide a function to automatically generate a mask for other cancer types. The newly generated segmentation mask can greatly reduce the manual work of creating the training sets for other cancer types, and thus accelerate the development of applications for pathology image analysis. Results show that the Mask-RCNN method can be adapted and applied in head and neck cancer, breast cancer and lung cancer squamous cell carcinoma pathology image datasets in TCGA (**Supplemental Figure S6, S7 and S8**).

4 DISCUSSION

In this study, we developed a deep learning-based analysis tool to study the TME using standard H&E stained pathology images. This tool successfully visualized and quantified the spatial organization of tumor cells, stromal cells, lymphocytes, inflammatory cells, red blood cells and karyorrhexis in the tumor tissues of lung ADC patients. The topological features of cell spatial organization were used to characterize the TME. Our results showed that these features were associated with patient survival outcome and the gene expression of biological pathways. From these image-derived TME features, we developed a prognostic model for lung ADC patients and independently validated it in another lung ADC patient cohort. The prognostic model predicts patient survival independent of other clinical variables in the validation cohort.

Several previous studies have tried to analyze the TME and discovered prognostic image features. However, these studies involved time-consuming hand-labeling by pathologists[26-28]. In contrast, we developed a fully automated and objective nuclei segmentation and classification strategy. In addition, this deep learning-based method enables segmentation of nuclei within a whole slide image. Since the number of cells in a whole slide image could be tremendous (~2,000,000 on average), manually labelling all of them is impractical. Thus, this deep-learning method empowers quantification of the TME across the whole slide image. Furthermore, although developed in lung ADC, this method can be easily generalized to other cancer types by retraining the model using the tools provided by our web portal.

In pathology image analysis, three-dimensional tissue structures are captured as two-dimensional images, and the cell nuclei may “touch” and “overlap” with each other in the resulting images. This is one of the major challenges for nuclei segmentation in pathology image analysis. In this study, we developed a Mask-RCNN-based deep learning model to segment and classify different types of cell nuclei in order to study the spatial interactions among different cell types and tissue structures. Compared with other image segmentation algorithms, the Mask-RCNN method has several advantages: First, it segments and classifies nuclei at the same time, while traditional nuclei segmentation algorithms relying on color deconvolution cannot classify cell types[2, 29]. Second, by using extensive color augmentation during the training process, it adapts to different staining conditions, which makes the algorithm more robust and allows us to avoid the time-consuming color normalization steps [30]. Third, compared with other popular semantic image segmentation neural networks such as Fully Convolutional Network (FCN) and Deeplab[31, 32] that classify each pixel, Mask-RCNN is intrinsically an instance segmentation algorithm that detects an object bounding box first and assigns pixels as foreground or background within this bounding box[16]. In summary, Mask-RCNN provides a new solution to segmenting closely clustered nuclei in tissue pathology images.

The associations between the extracted TME features and patient prognosis were evaluated in this study. Karyorrhexis, a representative of necrosis, has been reported as an aggressive tumor phenotype in lung cancer[33]. Consistently, the density of karyorrhectic cells and numbers of karyorrhexis-karyorrhexis edges were shown as negative prognostic factors in this study. On the other hand, the density of stromal cells and the numbers of stromal cell-stromal cell edges were positive prognostic factors, which is consistent with a recent report on lung ADC patients[8]. These consistencies indicate the validity of this MaskRCNN-based deep neural network and the potentiality of using cell organization features as novel biomarkers for clinical outcomes.

Gene expression patterns have been widely used to study the underlying biological mechanisms of different tumor types and subtypes[35, 36]; moreover, genes with abnormal expression could become potential therapeutic targets of cancers [37, 38]. However, traditional transcriptome profiling is usually done in bulk tumor[35, 39], which contains multiple cell types, such as stromal cells and lymphocytes, in addition to tumor cells. This bulk tumor-based sequencing could blur or diminish the mRNA expression changes arising from a single cell type or from different cell compositions in the TME. Currently, the relationship between the transcription activities of biological pathways and the TME remains unclear. In this study, the image-derived TME features show interesting correlations with the transcriptional activities of biological pathways. For example, gene expression levels of TCR and PD-1 pathways were positively correlated with the density of lymphocytes detected from tumor tissues. This indicates the image-derived TME features may be used to study or predict immunotherapy response, since several promising cancer immunotherapies rely on activation of tumor-infiltrated immune cells and blocking immune checkpoint pathways[23, 34]. In addition, the gene expression extracellular matrix organization pathway is associated with the density of stromal cells in tumor tissues. Since traditional transcriptome sequencing is done in bulk tumor, accurate cell composition derived from pathology images could help to improve the evaluation of gene expression for each individual cell type. Moreover, the correlation between image features and transcriptional patterns of biological pathways hints at the potential usage of image features to study tumor bioprocesses, including cell cycle and metabolism status.

There are some limitations to this Mask-RCNN model. First, information on the individual nuclei, such as nucleus shape and size, was not considered since this study focused on nuclei organization. Morphological and intensity features of nuclei have been reported as prognostic factors, which can be automatically extracted using this nuclei segmentation algorithm[40]. Second, some special structures, such as bronchi and cartilage, were not included in this algorithm. This study handled this problem by avoiding such structures during ROI annotation. However, a more comprehensive training set would be desirable for whole slide analysis.

Code availability

The code is available upon request. We will share the code through Github once this manuscript has been accepted.

Author contributions

S.W. and G.X. designed the study. S.W. and R.R. implemented and trained the neural network model. S.W., L.C. performed genomic analysis. S.W., R.R., D.M.Y., L.C., L.X., T.W., Y.X., A.G., J.M. and G.X. analyzed the data and wrote the manuscript. L.Y. labeled the data. All authors commented on the manuscript.

Acknowledgement

We thank Jessie Norris for helping us to edit this manuscript.

Competing financial interests

The authors declare that they have no competing interests.

Source of funding

This work was partially supported by the National Institutes of Health [5R01CA152301, P50CA70907, 1R01GM115473, and 1R01CA172211], and the Cancer Prevention and Research Institute of Texas [RP190107 and RP180805].

References

1. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, Chirieac LR, Dacic S, Duhig E, Flieder DB, Geisinger K, Hirsch FR, Ishikawa Y, Kerr KM, Noguchi M, Pelosi G, Powell CA, Tsao MS, Wistuba I and Panel WHO. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol*. 2015;10:1243-1260.

2. Cheng J, Mo X, Wang X, Parwani A, Feng Q and Huang K. Identification of topological features in renal tumor microenvironment associated with patient survival. *Bioinformatics*. 2018;34:1024-1030.

3. Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, Rimm DL, Yang M, Romero E, Schalper KA, Velcheti V and Madabhushi A. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. *Clin Cancer Res*. 2019;25:1526-1534.

4. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J, Cancer Genome Atlas Research N, Shmulevich I, Rao AUK, Lazar AJ, Sharma A and Thorsson V. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep*. 2018;23:181-193 e7.

5. Nakamura H, Ichikawa T, Nakasone S, Miyoshi T, Sugano M, Kojima M, Fujii S, Ochiai A, Kuwata T, Aokage K, Suzuki K, Tsuboi M and Ishii G. Abundant tumor promoting stromal cells in lung adenocarcinoma with hypoxic regions. *Lung Cancer*. 2018;115:56-63.

6. Bremnes RM, Donnem T, Al-Saad S, Al-Shibli K, Andersen S, Sirera R, Camps C, Marinez I and Busund LT. The role of tumor stroma in cancer progression and prognosis: emphasis on carcinoma-associated fibroblasts and non-small cell lung cancer. *J Thorac Oncol*. 2011;6:209-17.

7. Pietras K and Ostman A. Hallmarks of cancer: interactions with the tumor stroma. *Exp Cell Res*. 2010;316:1324-31.

8. Ichikawa T, Aokage K, Sugano M, Miyoshi T, Kojima M, Fujii S, Kuwata T, Ochiai A, Suzuki K and Tsuboi M. The ratio of cancer cells to stroma within the invasive area is a histologic prognostic parameter of lung adenocarcinoma. *Lung Cancer*. 2018.

9. Gooden MJ, de Bock GH, Leffers N, Daemen T and Nijman HW. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. *Br J Cancer*. 2011;105:93-103.

10. Miyashita M, Sasano H, Tamaki K, Hirakawa H, Takahashi Y, Nakagawa S, Watanabe G, Tada H, Suzuki A, Ohuchi N and Ishida T. Prognostic significance of tumor-infiltrating CD8+ and FOXP3+ lymphocytes in residual tumors and alterations in these parameters after neoadjuvant chemotherapy in triple-negative breast cancer: a retrospective multicenter study. *Breast Cancer Res*. 2015;17:124.

11. Huh JW, Lee JH and Kim HR. Prognostic significance of tumor-infiltrating lymphocytes for patients with colorectal cancer. *Arch Surg*. 2012;147:366-72.

12. Brambilla E, Le Teuff G, Marguet S, Lantuejoul S, Dunant A, Graziano S, Pirker R, Douillard JY, Le Chevalier T, Filipits M, Rosell R, Kratzke R, Popper H, Soria JC, Shepherd FA, Seymour L and Tsao MS. Prognostic Effect of Tumor Lymphocytic Infiltration in Resectable Non-Small-Cell Lung Cancer. *J Clin Oncol*. 2016;34:1223-30.

13. Proctor MJ, Morrison DS, Talwar D, Balmer SM, O'Reilly DSJ, Foulis AK, Horgan PG and McMillan DC. An inflammation-based prognostic score (mGPS) predicts cancer survival independent of tumour site: a Glasgow Inflammation Outcome Study. *Brit J Cancer*. 2011;104:726-734.

14. Jafri SH, Shi RH and Mills G. Advance lung cancer inflammation index (ALI) at diagnosis is a prognostic marker in patients with metastatic non-small cell lung cancer (NSCLC): a retrospective review. *Bmc Cancer*. 2013;13.

15. Coussens LM and Werb Z. Inflammatory cells and cancer: Think different! *J Exp Med*. 2001;193:F23-F26.

16. He KM, Gkioxari G, Dollar P and Girshick R. Mask R-CNN. *Ieee I Conf Comp Vis*. 2017:2980-2988.

17. Chollet F. Keras. 2015.

18. A Package for Survival Analysis in S [computer program]. 2015.

19. R: A language and environment for statistical computing. [computer program]. R Foundation for Statistical Computing, Vienna, Austria.; 2016.

20. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J,

Williams M, Wu G, Stein L, Hermjakob H and D'Eustachio P. The Reactome pathway Knowledgebase. *Nucleic Acids Res*. 2016;44:D481-7.

21. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*. 2016:060012.

22. Cantrell DA. T cell antigen receptor signal transduction pathways. *Cancer Surv*. 1996;27:165-175.

23. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer*. 2012;12:252-264.

24. Kalluri R and Zeisberg M. Fibroblasts in cancer. *Nat Rev Cancer*. 2006;6:392-401.

25. Huang QM, Tomida S, Masuda Y, Arima C, Cao K, Kasahara TA, Osada H, Yatabe Y, Akashi T, Kamiya K, Takahashi T and Suzuki M. Regulation of DNA Polymerase POLD4 Influences Genomic Instability in Lung Cancer. *Cancer Research*. 2010;70:8407-8416.

26. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, West RB, van de Rijn M and Koller D. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*. 2011;3:108ra113.

27. Wang C, Pecot T, Zynger DL, Machiraju R, Shapiro CL and Huang K. Identifying survival associated morphological features of triple negative breast cancer using multiple datasets. *J Am Med Inform Assoc*. 2013;20:680-7.

28. Yuan Y, Failmezger H, Rueda OM, Ali HR, Graf S, Chin SF, Schwarz RF, Curtis C, Dunning MJ, Bardwell H, Johnson N, Doyle S, Turashvili G, Provenzano E, Aparicio S, Caldas C and Markowetz F. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med*. 2012;4:157ra143.

29. Phoulady HA, Goldgof DB, Hall LO and Mouton PR. Nucleus Segmentation in Histology Images with Hierarchical Multilevel Thresholding. *Medical Imaging 2016: Digital Pathology*. 2016;9791.

30. Alsubaie N, Trahearn N, Raza SEA, Snead D and Rajpoot NM. Stain Deconvolution Using Statistical Analysis of Multi-Resolution Stain Colour Representation. *Plos One*. 2017;12.

31. Shelhamer E, Long J and Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:640-651.

32. Chen LC, Papandreou G, Kokkinos I, Murphy K and Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans Pattern Anal Mach Intell*. 2018;40:834-848.

33. Swinson DEB, Jones JL, Richardson D, Cox G, Edwards JG and O'Byrne KJ. Tumour necrosis is an independent prognostic marker in non-small cell lung cancer: correlation with biological variables. *Lung Cancer*. 2002;37:235-240.

34. Topalian SL, Drake CG and Pardoll DM. Targeting the PD-1/B7-H1(PD-L1) pathway to activate anti-tumor immunity. *Curr Opin Immunol*. 2012;24:207-212.

35. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES and Golub TR. Multiclass cancer diagnosis using tumor gene expression signatures. *P Natl Acad Sci USA*. 2001;98:15149-15154.

36. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE and Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *P Natl Acad Sci USA*. 2001;98:10869-10874.

37. Vermeulen K, Van Bockstaele DR and Berneman ZN. The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif*. 2003;36:131-49.

38. Semenza GL. Targeting HIF-1 for cancer therapy. *Nat Rev Cancer*. 2003;3:721-732.

39. Li HP, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, Wong M, Cima I, Tan MH, Wee LJK, Hillmer AM, Tan IB, Robson P and Prabhakar S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors (vol 50, pg 1754, 2018). *Nat Genet*. 2018;50:1754-1754.

40. Diamond DA, Berry SJ, Umbricht C, Jewett HJ and Coffey DS. Computerized Image-Analysis of Nuclear Shape as a Prognostic Factor for Prostatic-Cancer. *Prostate*. 1982;3:321-332.

Figure Legend

Figure 1. Flow chart of pathology image analysis pipeline in this study.
Mask RCNN: Mask Regional-Convolutional Neural Network.

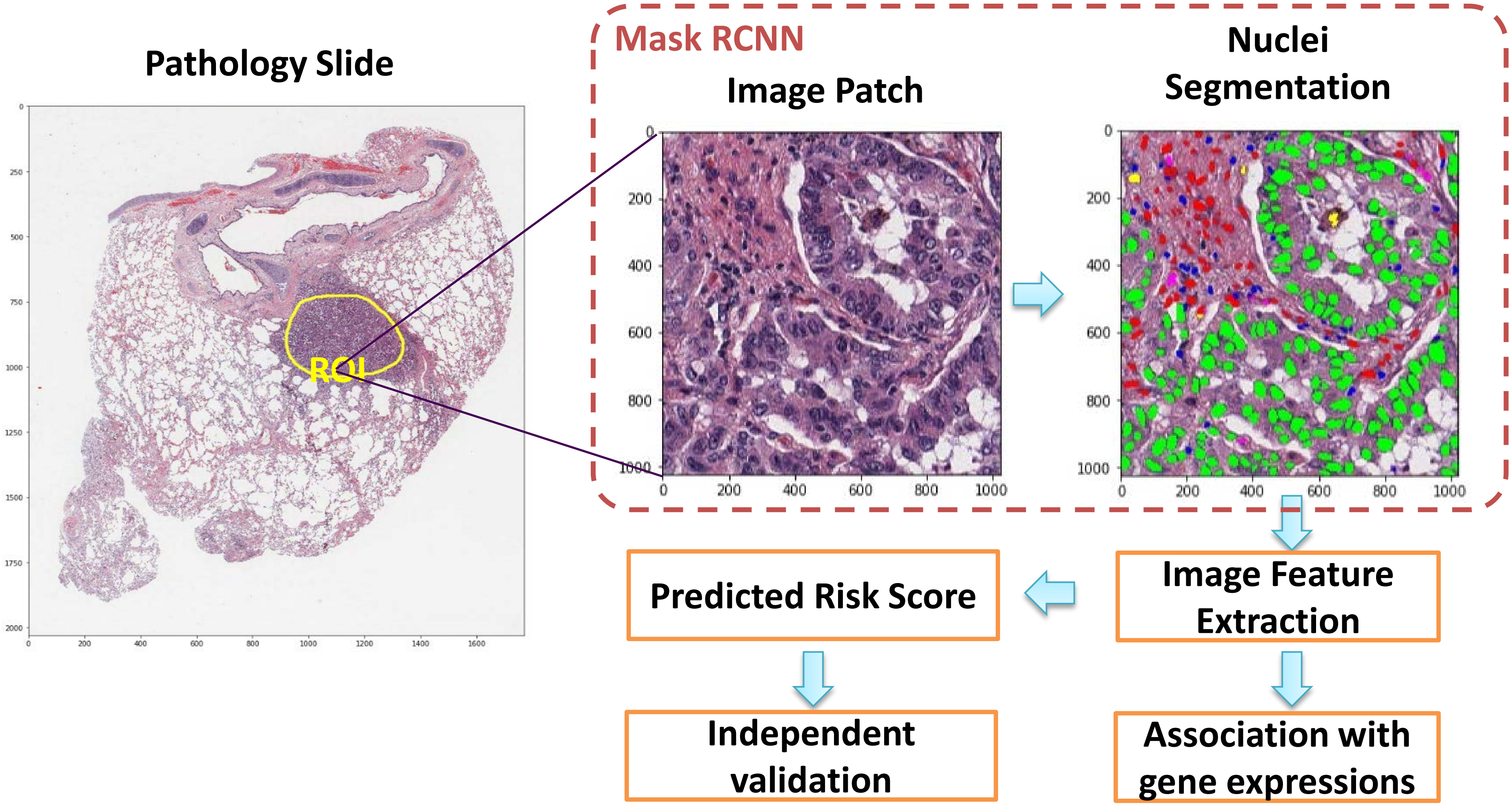
Figure 2. Mask-RCNN-based nuclei segmentation and classification results in lung ADC pathological images.

Figure 3. Nuclei segmentation and classification across whole slide image. Upper panel: Original pathology image. Lower panel: detected and classified nuclei overlay on top of the original pathology image.

Figure 4. Prognostic value of the TME feature-based prognostic model. (A) K-M plot of predicted high- and low-risk groups in the TCGA dataset. Log-rank test, p-value = 0.001. **(B)** Multivariate survival analysis of predicted risk group adjusted by potential confounders. CI, confidence interval; HR, hazard ratio.

Figure 5. Correlation between image features and mRNA expression in tumor. (A-D) Volcano plots of gene set enrichment analysis results correlating mRNA expression level with tumor nuclei density (A), stromal nuclei density (B), lymphocyte density (C), and karyorrhexis density (D) respectively. Thirteen interesting gene sets are highlighted. **(E)** To look into the significantly correlated gene sets, an example heatmap shows that most mRNA expression levels in the cell cycle gene set are positively correlated with tumor nuclei density in tumor issue. Only genes with p value < 0.001 in Spearman rank correlation with tumor cell number are shown. Patients are grouped according to tumor cell density showing on the top row.

Figure 6. An online tool for nuclei segmentation and classification in pathology image. Nuclei segmentation and classification results can be automatically generated from pathology images. The spatial locations and cell types of individual cells can be downloaded as an Excel table.



Pathology Slide

Mask RCNN

Image Patch

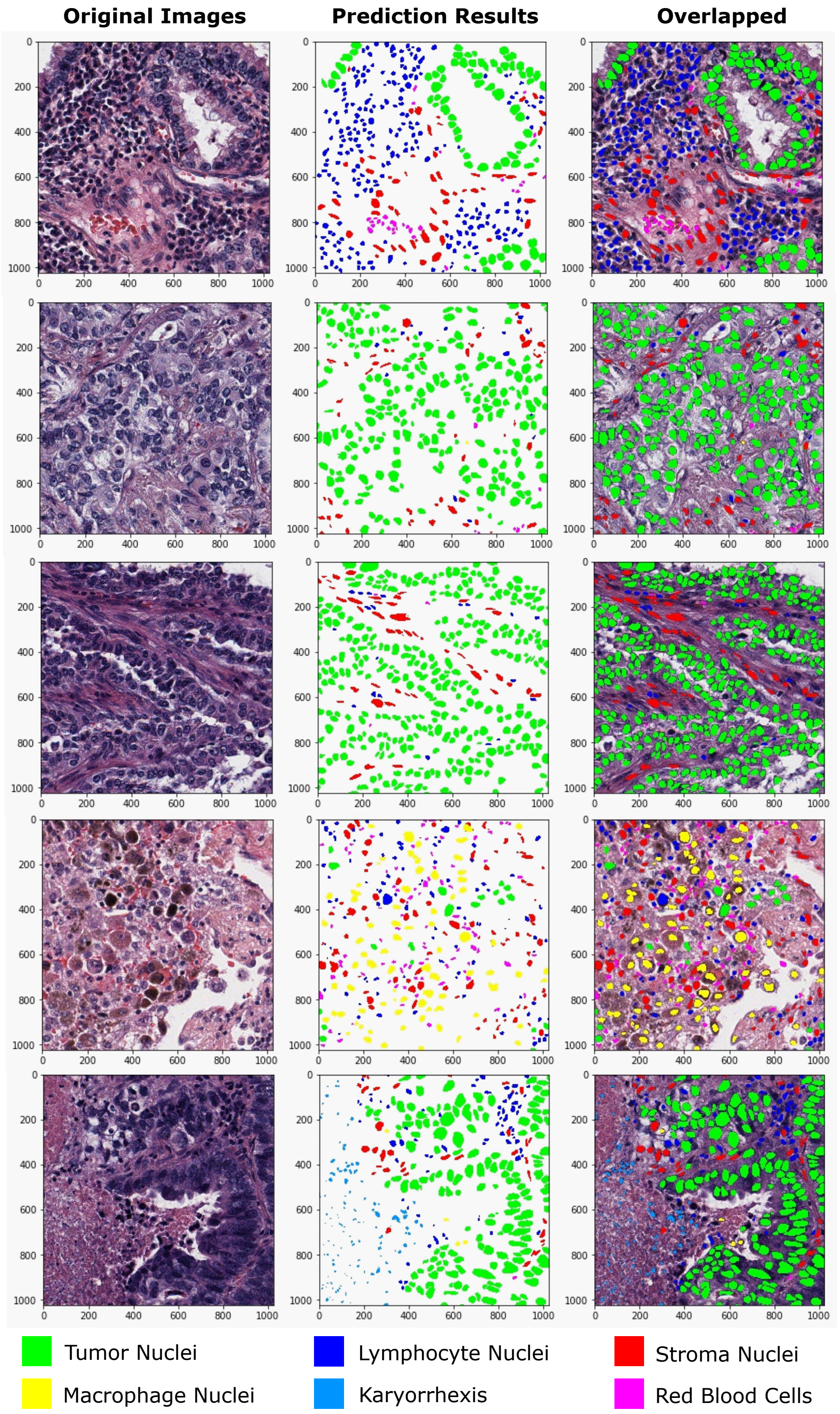
Nuclei Segmentation

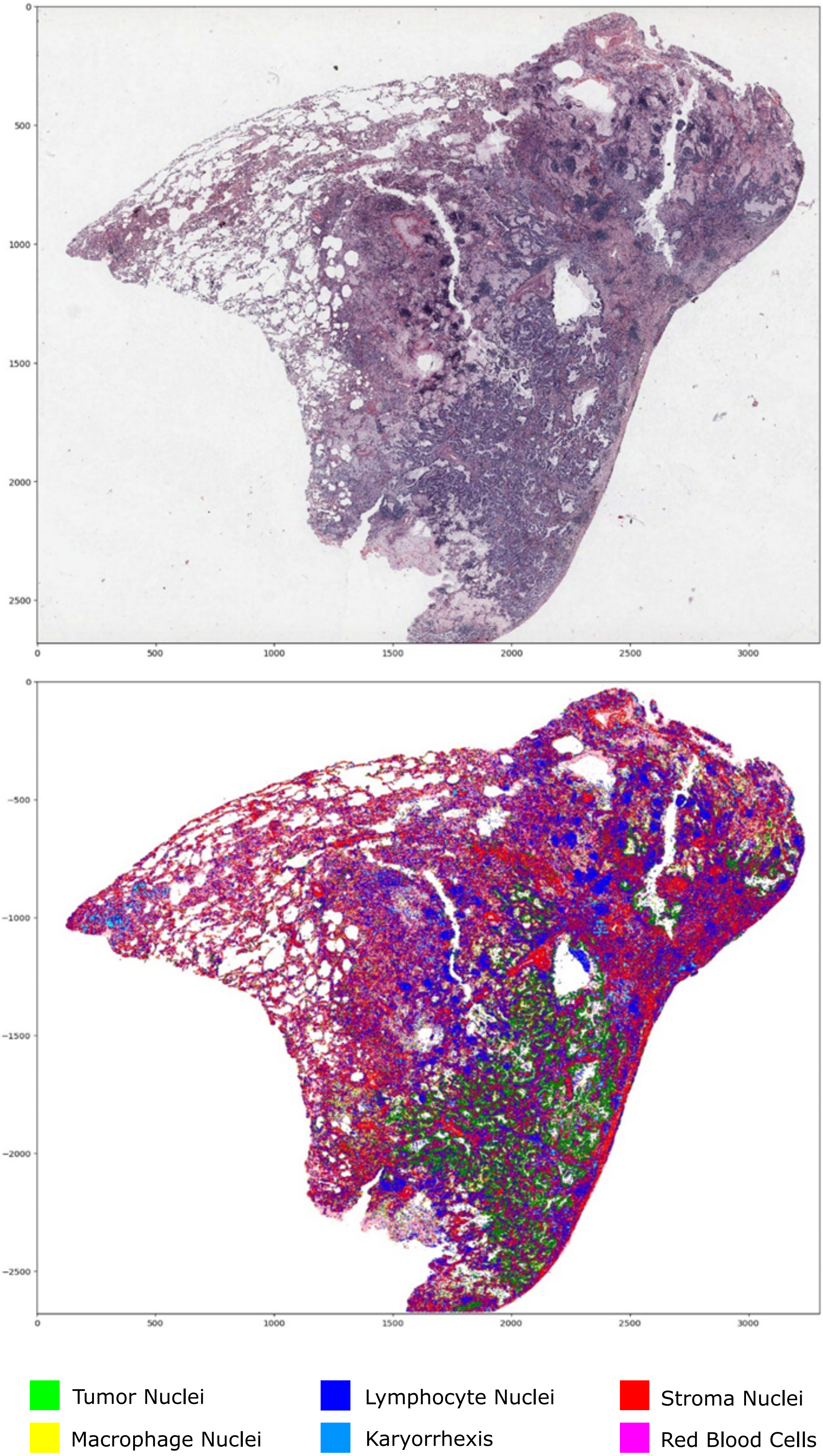
Image Feature Extraction

Predicted Risk Score

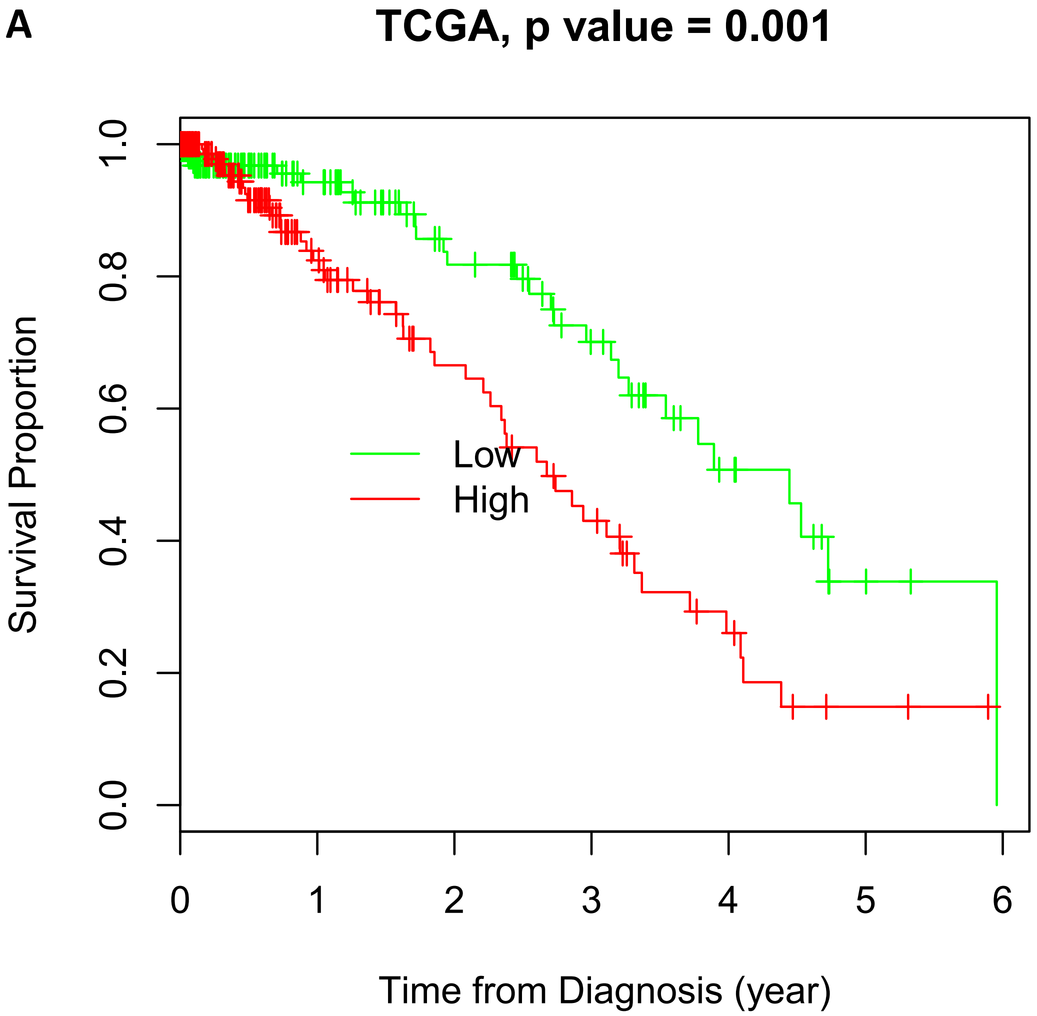
Association with gene expressions

Independent validation


Original Images
Prediction Results
Overlapped
0
200
400
600
800
1000
🟩 Tumor Nuclei
🟦 Lymphocyte Nuclei
🟥 Stroma Nuclei
🟨 Macrophage Nuclei
🟦 Karyorrhexis
🟪 Red Blood Cells

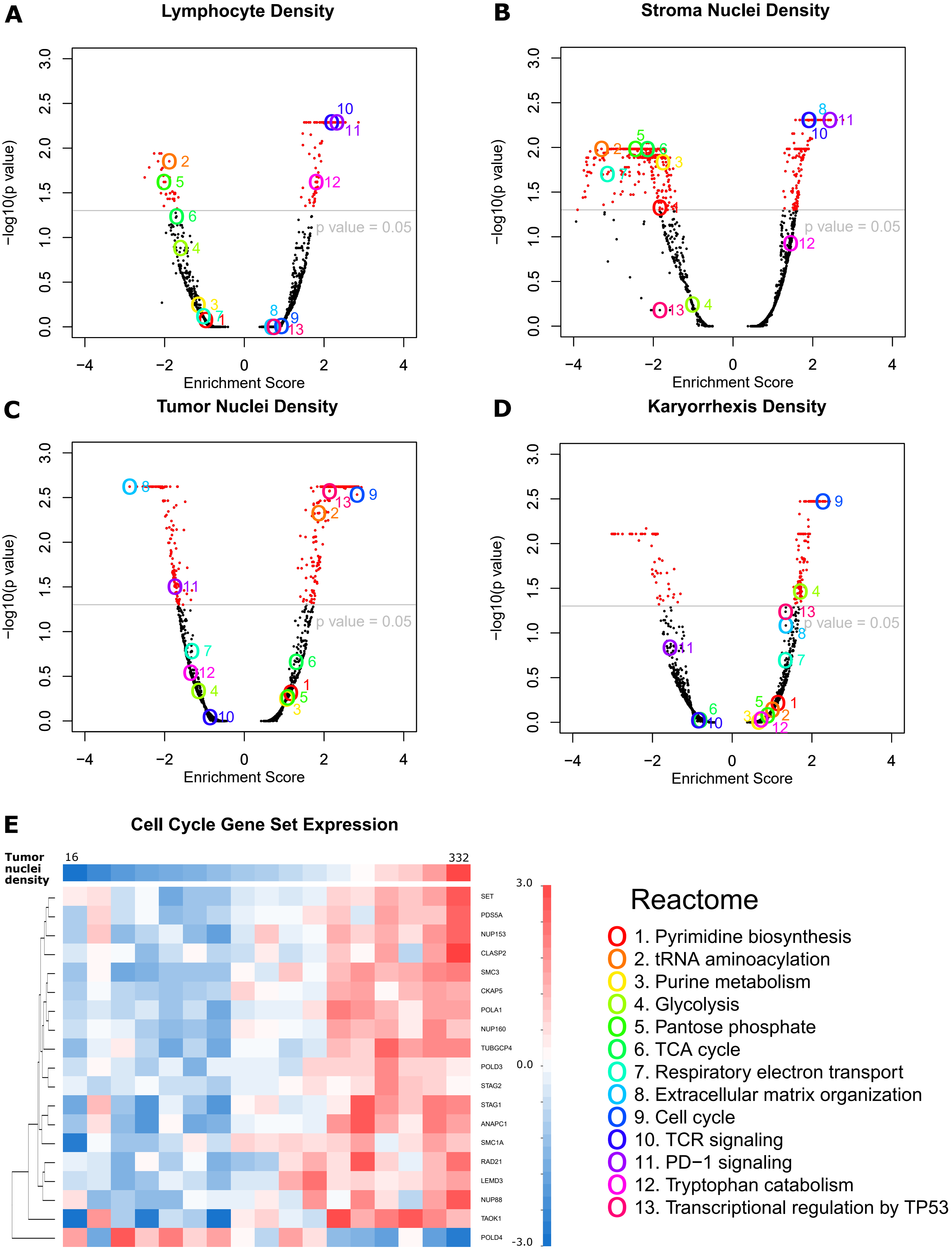

0
500
1000
1500
2000
2500
0
500
1000
1500
2000
2500
3000
0
−500
−1000
−1500
−2000
−2500
0
500
1000
1500
2000
2500
3000
🟩 Tumor Nuclei
🟦 Lymphocyte Nuclei
🟥 Stroma Nuclei
🟨 Macrophage Nuclei
🟦 Karyorrhexis
🟪 Red Blood Cells

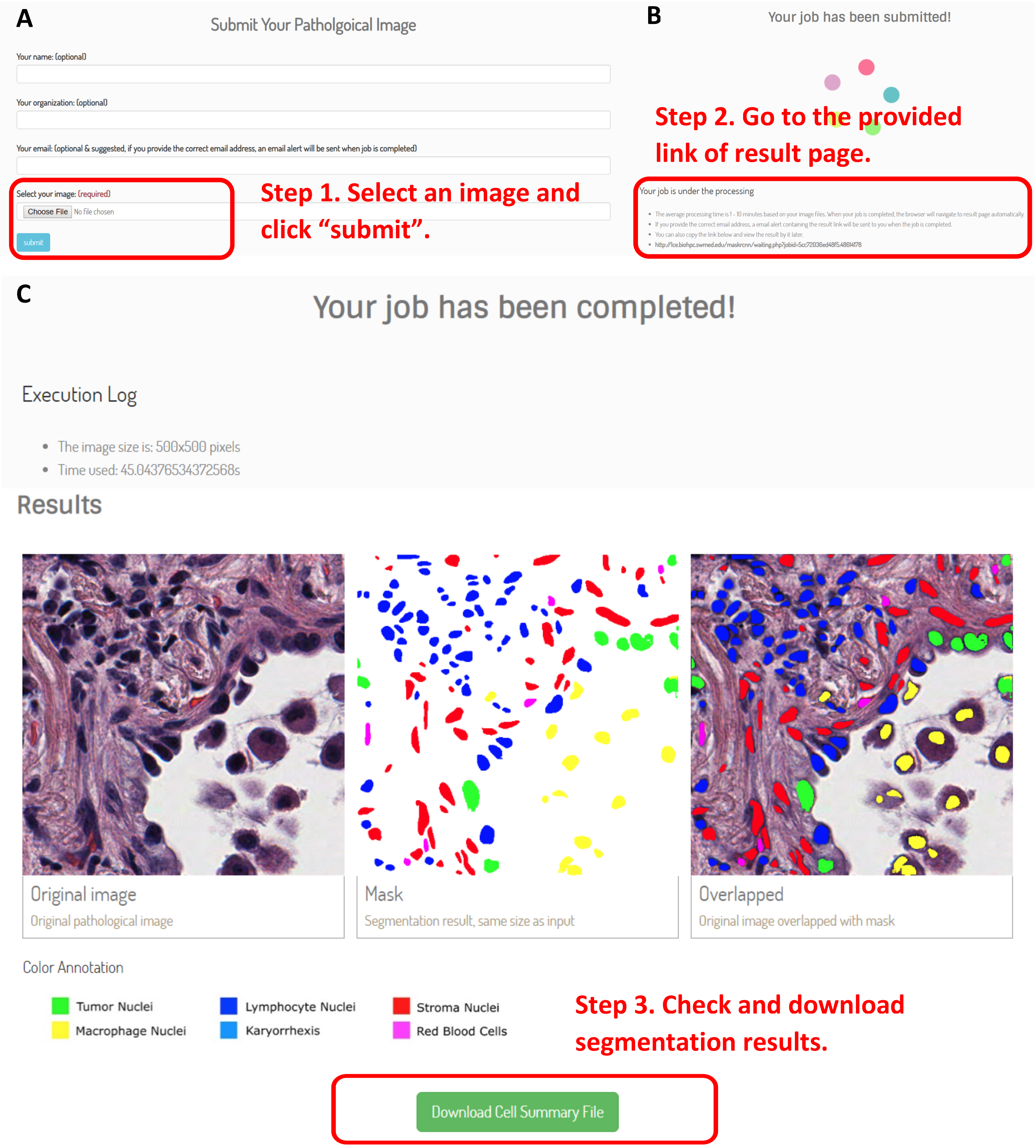
A



B

TCGA dataset (n=371)	HR (95% CI)	p value
High- vs. low-risk	2.23 (1.37 – 3.65)	0.0013
Age (year)	1.02 (0.99 – 1.05)	0.097
Male vs. female	0.90 (0.55 – 1.49)	0.70
Smoker vs. non-smoker	1.09 (0.66 – 1.81)	0.73
Stage		
Stage I	ref	-
Stage II	2.36 (1.28 – 5.30)	0.0029
Stage III	4.59 (2.44 – 8.63)	<0.001
Stage IV	3.30 (1.30 – 8.41)	0.012


A
Lymphocyte Density
B
Stroma Nuclei Density
C
Tumor Nuclei Density
D
Karyorrhexis Density
−log10(p value)
Enrichment Score
p value = 0.05
−4
−2
0
2
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
E
Cell Cycle Gene Set Expression
Tumor nuclei density
16
332
3.0
0.0
-3.0
SET
PDS5A
NUP153
CLASP2
SMC3
CKAP5
POLA1
NUP160
TUBGCP4
POLD3
STAG2
STAG1
ANAPC1
SMC1A
RAD21
LEMD3
NUP88
TAOK1
POLD4
Reactome
1. Pyrimidine biosynthesis
2. tRNA aminoacylation
3. Purine metabolism
4. Glycolysis
5. Pantose phosphate
6. TCA cycle
7. Respiratory electron transport
8. Extracellular matrix organization
9. Cell cycle
10. TCR signaling
11. PD−1 signaling
12. Tryptophan catabolism
13. Transcriptional regulation by TP53



Submit Your Patholgoical Image

Your name: (optional)

Your organization: (optional)

Your email: (optional & suggested, if you provide the correct email address, an email alert will be sent when job is completed)

Select your image: (required)

No file chosen

Step 1. Select an image and click "submit".

Your job has been submitted!

Step 2. Go to the provided link of result page.

Your job is under the processing

- The average processing time is 1 - 10 minutes based on your image files. When your job is completed, the browser will navigate to result page automatically.
- If you provide the correct email address, a email alert containing the result link will be sent to you when the job is completed.
- You can also copy the link below and view the result by it later.
- http://lce.biohpc.swmed.edu/maskrcnn/waiting.php?jobid=5cc72036ed48f5.48614178

Your job has been completed!

Execution Log

- The image size is: 500x500 pixels
- Time used: 45.04376534372568s

Results

Original image
Original pathological image

Mask
Segmentation result, same size as input

Overlapped
Original image overlapped with mask

Color Annotation

🟩 Tumor Nuclei
🟨 Macrophage Nuclei
🟦 Lymphocyte Nuclei
🟦 Karyorrhexis
🟥 Stroma Nuclei
🟪 Red Blood Cells

Step 3. Check and download segmentation results.

Supplemental Material

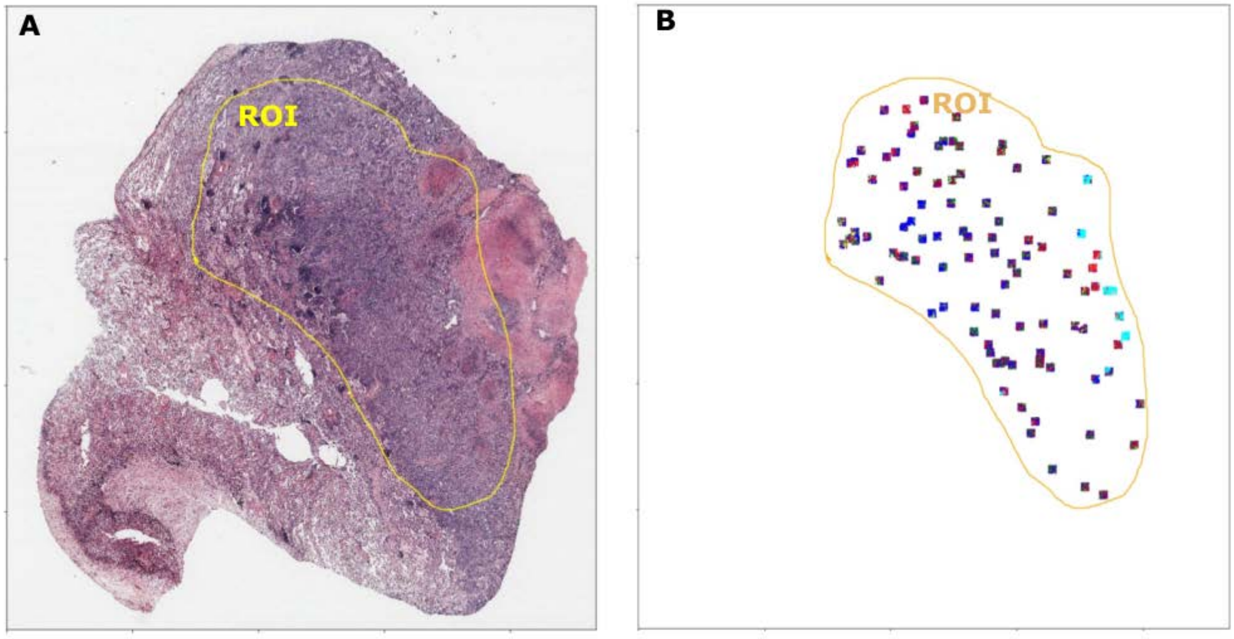
Supplemental Table 1. Patient characteristics for the National Lung Screening Trial (NLST) and The Cancer Genome Atlas (TCGA) datasets.

		NLST	**TCGA**
Number of patients		135	372
Number of pathology slides		208	431
Age at diagnosis (years, median (min - max))		64 (55 - 74)	66 (33 - 88)
Vital status (%)	Alive	94 (69.6)	297 (79.8)
	Deceased	41 (30.4)	75 (20.2)
Gender (%)	M	77 (57.0)	208 (55.9)
	F	58 (43.0)	164 (44.1)
Cancer stage (%)	I	90 (66.7)	208 (55.9)
	II	12 (8.9)	95 (25.5)
	III	22 (16.3)	47 (12.6)
	IV	10 (7.4)	21 (5.6)
	NA	0 (0.0)	1 (0.3)
Smoking status (%)	Smoker	74 (54.8)	254 (68.3)
	Non-smoker	61 (45.2)	118 (31.7)

Supplemental Table 2. Image features that are significantly associated with patient overall survival outcome in univariate survival analysis in the National Lung Screening Trial (NLST) dataset. Features are dichotomized by the median value. P values are calculated using Ward test. CI, confidence interval; HR, hazard ratio.

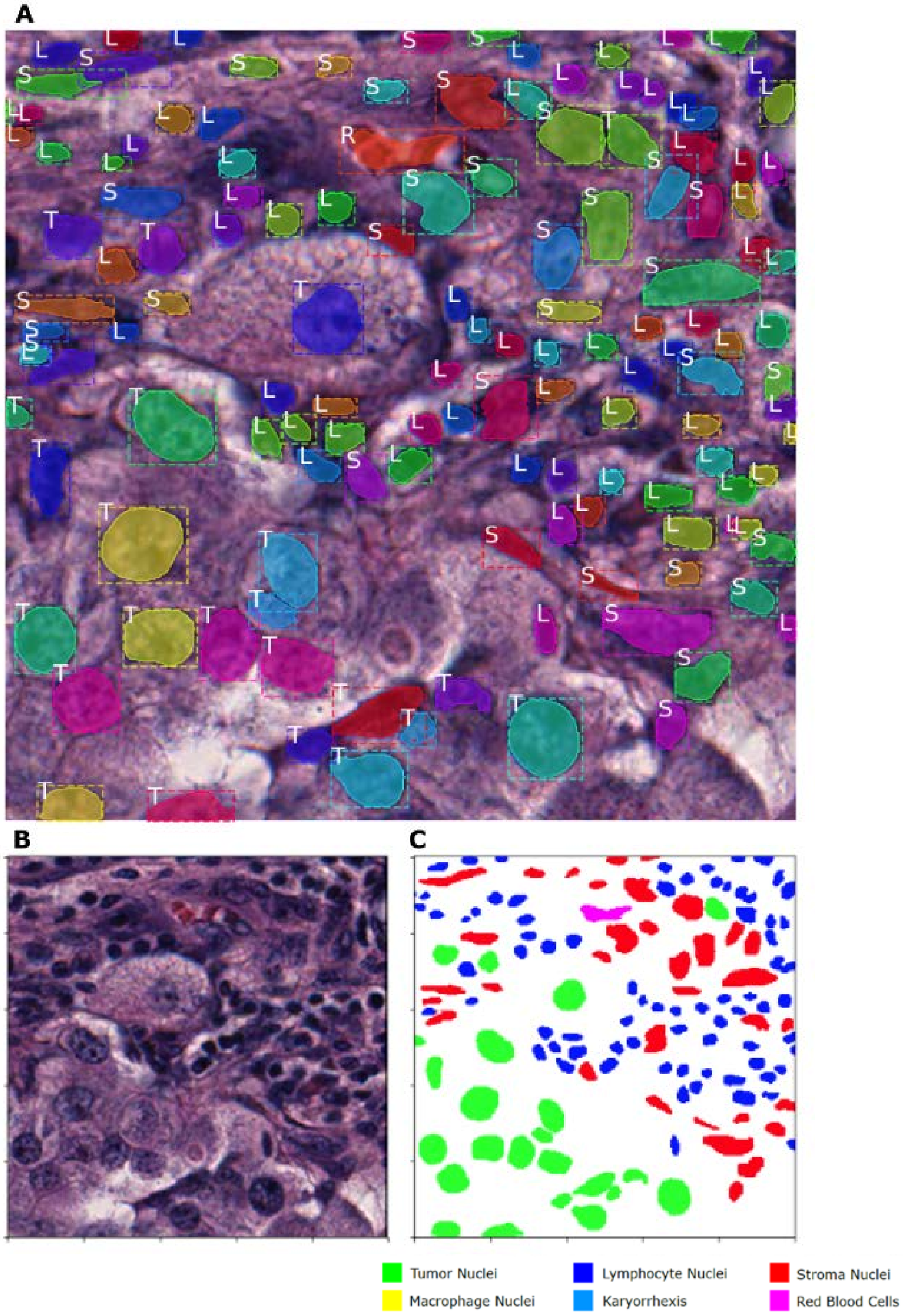
Features (high vs. low)	**HR (95% CI)**	**p value**
Number of stromal cell – stromal cell connections	0.32 (0.16 – 0.63)	0.0011
Number of karyorrhexis – karyorrhexis connections	2.54 (1.32 – 4.91)	0.0055
Stroma nuclei density	0.41 (0.21 – 0.80)	0.0088
Karyorrhexis density	2.27 (1.19 – 4.33)	0.013
Number of red blood cell-karyorrhexis connections	1.98 (1.05 – 3.74)	0.035

Supplemental Figure 1. Illustration of 100 patches randomly sampled from the Region of Interest (ROI). **(A)** A pathology image with ROI labelled by pathologists (yellow line). **(B)** 100 patches are randomly sampled within the ROI. Each square represents a 1024 × 1024 pixel image patch.

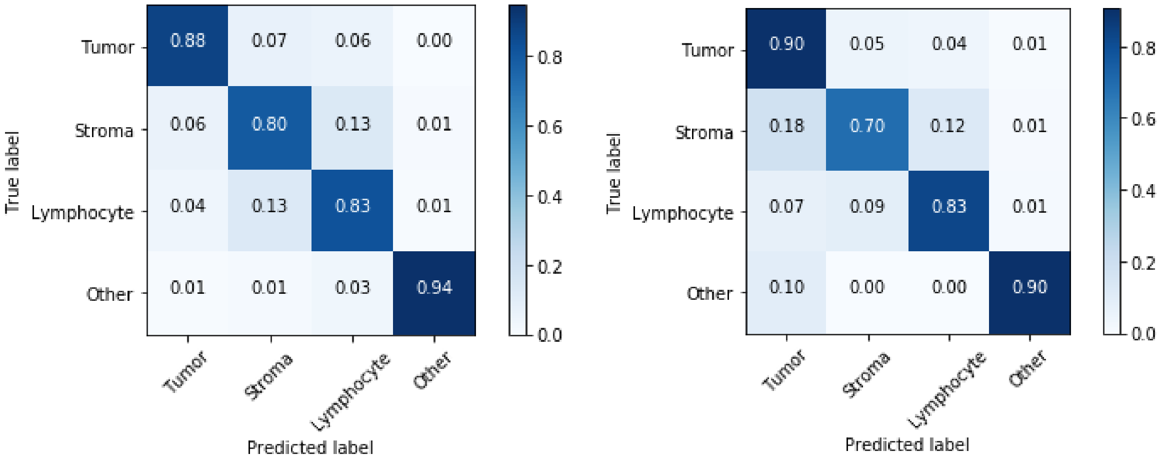


Supplemental Figure 2 Illustration of instance segmentation using the developed Mask-RCNN model.

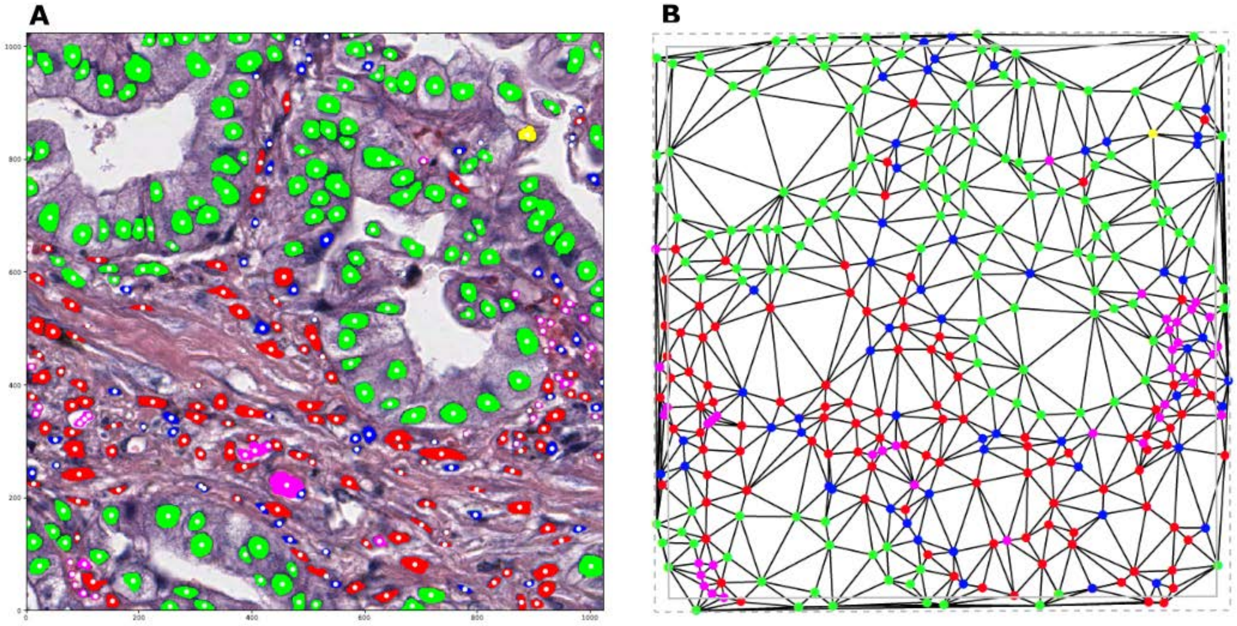
(A) Each nucleus has a bounding box (dashed lines). Nucleus segmentation is made within this bounding box. The class of the nucleus is also predicted at the same time as the segmentation, and the nuclei class is labeled on the top left of its bound box. (T: tumor nuclei; S: stromal nuclei; L: lymphocyte nuclei; R: red blood cell; M: macrophage nuclei and K: karyorrhexis) **(B)** Original image patch (500 × 500 pixels). **(C)** For better viewing, the nuclei are colored according to their categories.



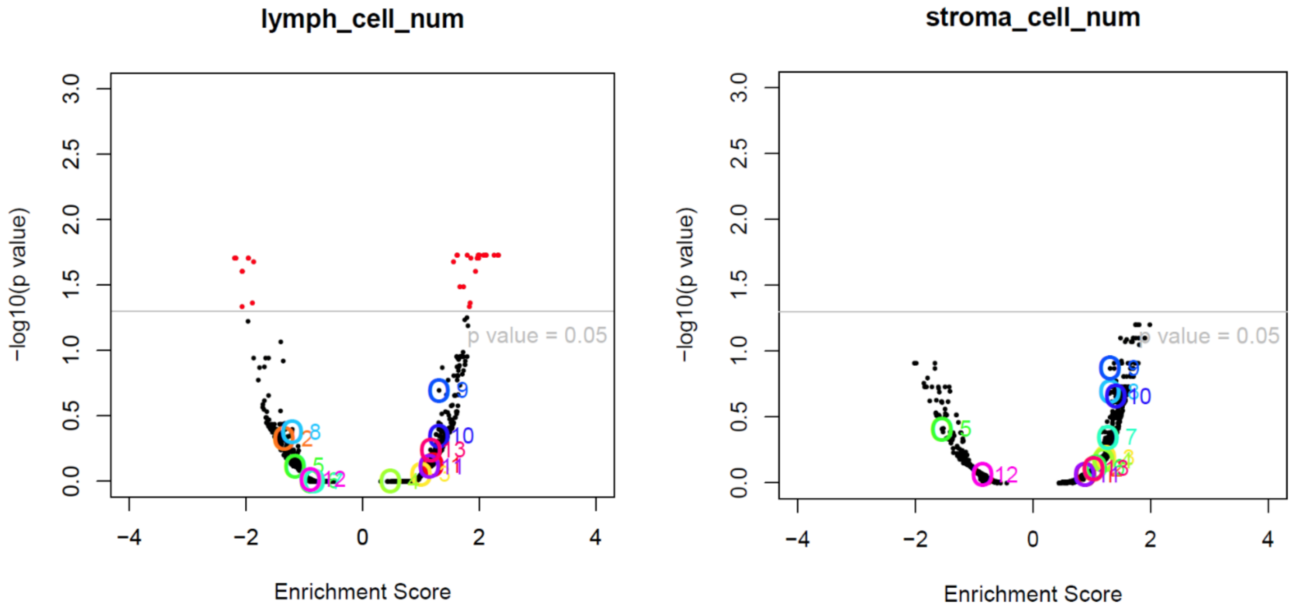
Supplemental Figure 3. Confusion matrix between true labels and predicted labels, in the validation set (left) and testing set (right) respectively.



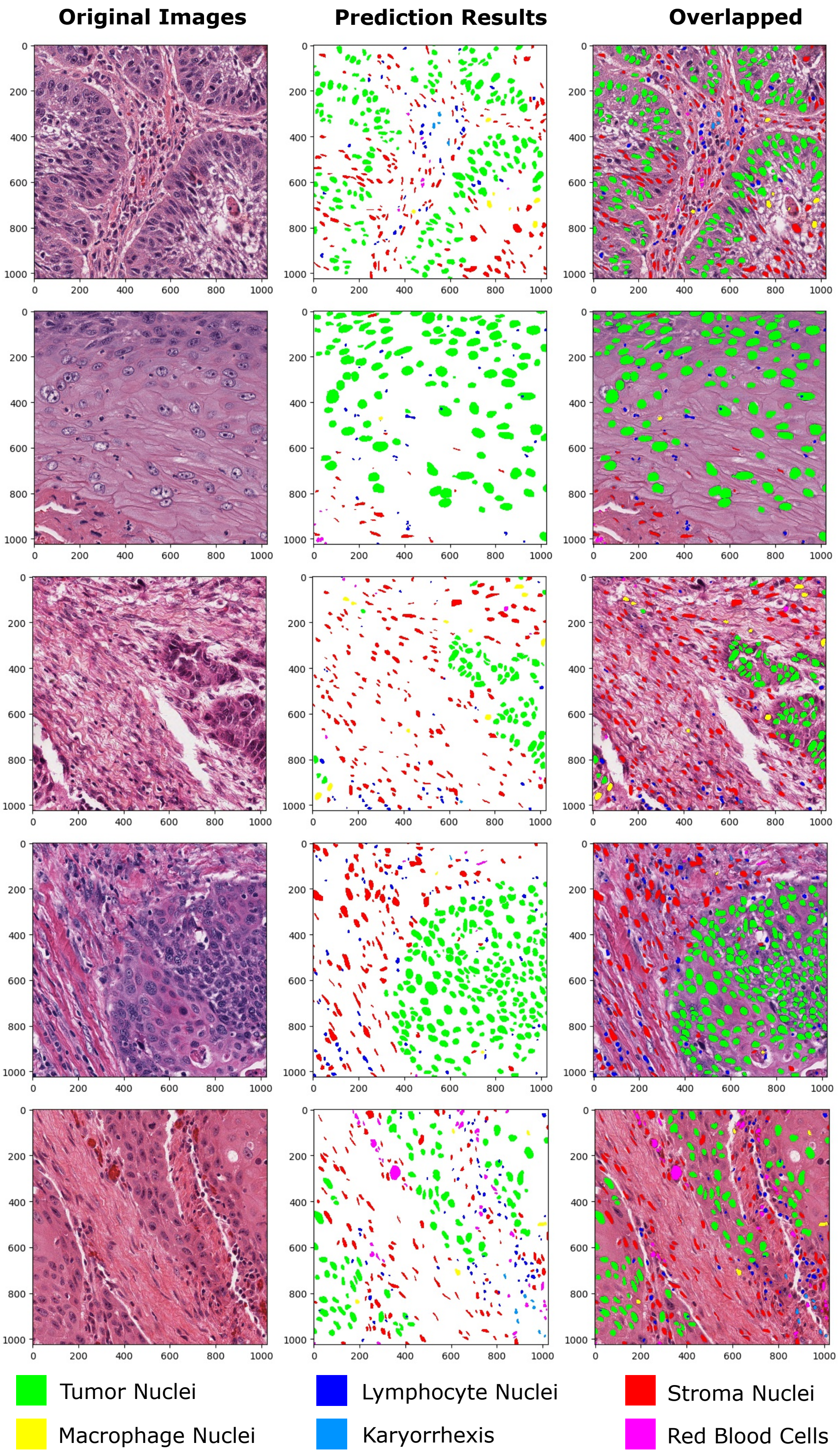
Supplemental Figure 4. Illustration of Delaunay triangle construction. **(A)** Nuclei segmentation results using deep learning. Centroids of detected nuclei are marked in white. **(B)** A Delaunay triangle graph is constructed using nuclei centroids as vertices. To remove edge effect, when extracting graph properties, only edges with both ends within the solid gray square (976 × 976 pixels) are counted. The dotted gray square (1024 × 1024 pixels) plots the boundary of the input image patch. Green, tumor nuclei; red, stroma nuclei; blue, lymphocyte nuclei; yellow, macrophage.



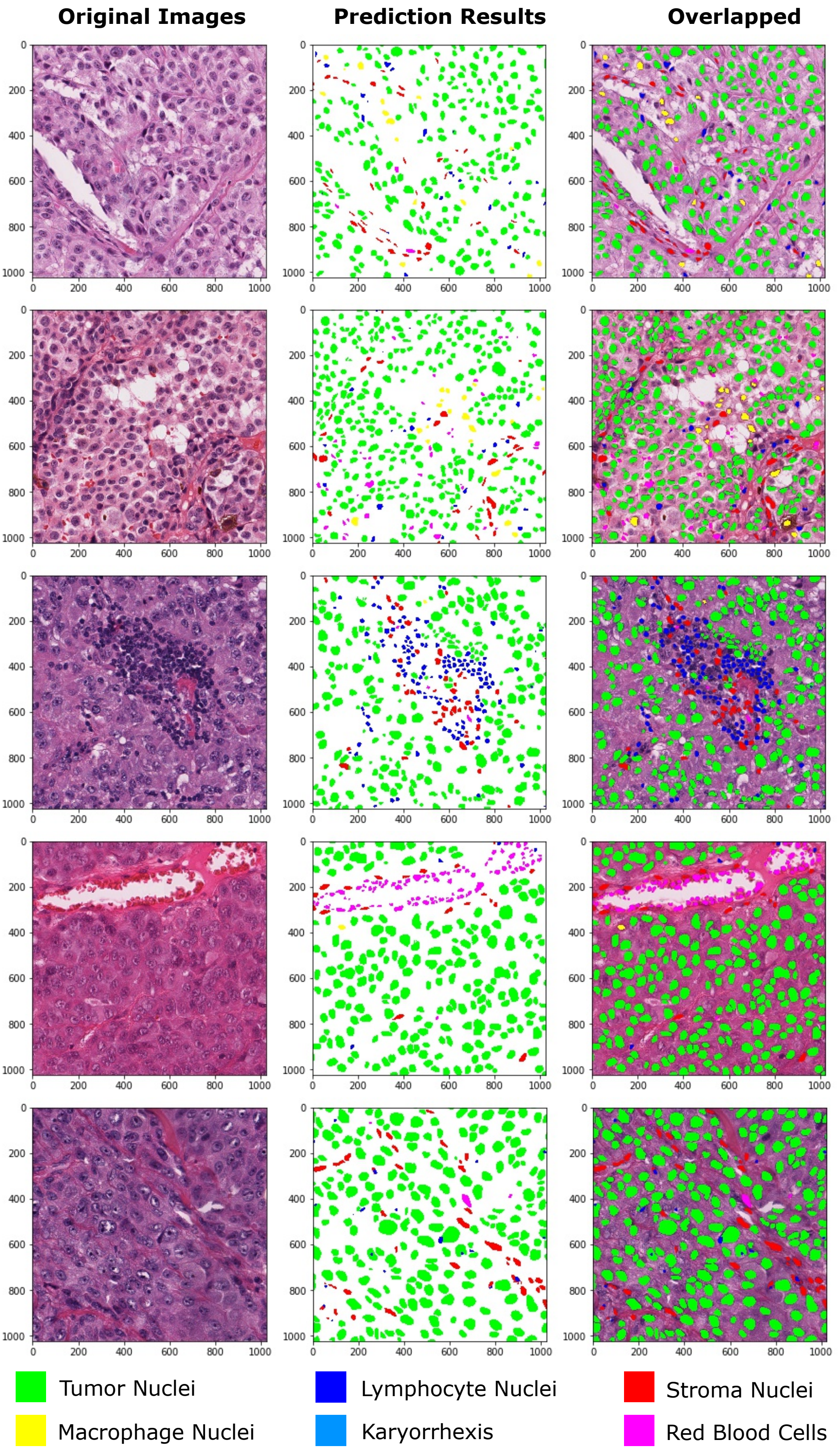
Supplemental Figure 5. Example of gene set enrichment analysis (GSEA) results correlating mRNA expression with image-derived stromal cell density and lymphocyte density while patient IDs are randomly shuffled as a negative control experiment.



Supplemental Figure 6. Deep learning based nuclei segmentation results in TCGA head and neck cancer pathology images.



Supplemental Figure 7. Deep learning based nuclei segmentation results in TCGA breast cancer pathology images.



Supplemental Figure 8. Deep learning based nuclei segmentation results in TCGA lung squamous cell carcinoma pathology images.