A Deep Learning Approach for Tissue Spatial Quantification and Genomic Correlations of Histopathological Images

Zixiao Lu[1,#,a], Xiaohui Zhan[2,3,#,b], Yi Wu[3,c], Jun Cheng[2,d], Wei Shao[3,e], Dong Ni[2,f], Zhi Han[3,g], Jie Zhang[4,h], Qianjin Feng[1,*,i], Kun Huang[3,*,j]

[1] *Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China*

[2] *National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 518060, China*

[3] *Regenstrief Institute, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, 46202, USA*

[4] *Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA*

[#]Co-first authors

[*]Corresponding authors

E-mail: fengjq99@fimmu.com (Feng Q), kunhuang@iu.edu (Huang K)

[a] ORCID: 0000-0003-0809-8703

[b] ORCID: 0000-0003-1326-6601

[c] ORCID: 0000-0003-3838-7418

[d] ORCID: 0000-0001-5493-961X

[e] ORCID: 0000-0002-9401-1186

[f] ORCID: 0000-0002-9146-6003

[g] ORCID: 0000-0002-5603-8433

[h] ORCID: 0000-0001-6939-7905

[i] ORCID: 0000-0001-8647-0596

[j] ORCID: 0000-0002-8530-370X

28 **Running title:** *Lu Z et al / Tissue Quantification and Genomic Correlations*

29

30 Total counts of words (from “Introduction” to “Conclusions”): 3613

31 Total counts of references: 36

32 Total counts of figures: 5

33 Total counts of tables: 3

34 Supplementary figures: 1

35 Supplementary tables: 1

Abstract

Epithelial and stromal tissue are components of the tumor microenvironment and play a major role in tumor initiation and progression. Distinguishing stroma from epithelial tissues is critically important for spatial characterization of the tumor microenvironment. We propose an image analysis pipeline based on a Convolutional Neural Network (CNN) model to classify epithelial and stromal regions in whole-slide images. The CNN model was trained using well-annotated breast cancer tissue microarrays and validated with images from The Cancer Genome Atlas (TCGA) project. Our model achieves a classification accuracy of 91.02%, which outperforms other state-of-the-art methods. Using this model, we generated pixel-level epithelial/stromal tissue maps for 1,000 TCGA breast cancer slide images that are paired with gene expression data. We subsequently estimated the epithelial and stromal ratios and performed correlation analysis to model the relationship between gene expression and tissue ratios. Gene Ontology enrichment analyses of genes that were highly correlated with tissue ratios suggest the same tissue was associated with similar biological processes in different breast cancer subtypes, whereas each subtype had its own idiosyncratic biological processes governing the development of these tissues. Taken all together, our approach can lead to new insights in exploring relationships between image-based phenotypes and their underlying genomic data and biological processes for all types of solid tumors.

KEYWORDS: Whole-slide tissue image; Deep learning; Integrative genomics; Breast cancer.

Introduction

Most solid tumors are composed of many tissue types including cancer cells, stroma, and epithelium. The interaction of tissues within such complex neoplasms defines the tumor microenvironment and this variably contributes to cancer initiation, progression, and therapeutic responses. For example, breast cancer epithelial cells of the mammary ducts are commonly the site of tumor initiation, while stromal tissue dynamics drive invasion and metastasis [1]. Tumor-to-stroma ratios of H&E stained images are therefore an important prognostic factor [2,3], and distinguishing stromal from epithelial tissue in histological images constitutes a basic, but crucial, task for cancer pathology. Classification methods (*i.e.* pre-processing, training classifiers with carefully selected features, and patch-level classification) are the most common automated computational methods for tissue segmentation [4,5]. For instance, Bunyak et al. [6] combined traditional feature selection methods and classification methods to perform segmentation of epithelial and stromal tissues on a tissue microarray (TMA) database. While this approach is viable, it can be time-consuming and inefficient given the feature selection process. Convolutional Neural Networks (CNN) models have the potential to improve analysis time and performance. Recently, deep CNN models have greatly boosted the performance of natural image analysis techniques such as image classification [7], object detection [8] and semantic segmentation [9,10], and biomedical image analysis [11–13]. Additionally, Ronneberger et al. [14] proposed implementation of a U-Net architecture to capture context and a symmetric expanding path that enables precise localization in biomedical image segmentation. CNN models have also been combined with traditional approaches to enhance the segmentation performance of epithelial and stromal regions [11,12].

Despite breakthroughs in the application of CNN models to medical image analysis, automated classification of epithelial and stromal tissues in Whole Slide Tissue Images (WSI) remains challenging due to the large size of WSI. WSI contain billions of pixels, and machine learning methods are limited by the technical hurdles of

working with large datasets [13]. Several solutions based on deep learning for classification of WSI have been proposed. A context-aware stacked CNN was proposed for the classification of breast WSI into multiple categories, such as normal/benign, ductal carcinoma in situ and invasive ductal carcinoma [15]. Saltz et al. presented a patch-based CNN to classify WSI into glioma and non-small-cell lung carcinoma subtypes [16,17].

Additionally, commercial software has been developed to aid in quantitative and objective analyses of tissue WSI. Among them is GENIE (Leica/ Aperio), a tool with proprietary algorithms which incorporate deep learning. While many of its functionalities are designed to handle specific biomarkers using immunohistochemical (IHC) or fluorescent images, for H&E images, tissue segmentation requires user-defined regions of interests (ROI). Similarly, HALO (Indica Labs) and Visiopharm (Hoersholm) provide a toolbox for histopathological image analysis. The toolbox includes unsupervised algorithms for tissue segmentation that require manual configuration of parameters and usually underperform supervised methods. The AQUA system (HistoRx) focuses on estimating tissue scores on TMA based on IHC staining by measuring protein expression within defined ROI. Therefore, reliable systems that enable both fully-automatic tissue segmentation and quantified analysis for H&E whole-slide images are still in great demand.

In this work, we propose a WSI processing pipeline that utilizes deep learning to perform automatic segmentation and quantification of epithelial and stromal tissues for breast cancer WSI from The Cancer Genome Atlas (TCGA). The TCGA data portal provides both clinical information and paired molecular data [18,19]. This offers the opportunity to identify relationships between computational histopathologic image features and the corresponding genomic information, which greatly informs research into the molecular basis of tumor cell and tissue morphology [20–22], as well as important issues such as immune-oncology therapy [17].

We first trained and validated a deep CNN model on annotated H&E stained histologic image patches, then successfully applied the WSI processing pipeline to process 1,000 TCGA breast cancer WSI to segment and quantify epithelial and stromal tissues. Spatial quantification and correlations with genomic data of both tissue types for three breast cancer subtypes (ER-positive, ER-negative and triple negative) were estimated based on the high-resolution global tissue segmentation maps. Gene Ontology (GO) enrichment can indicate when such tissues are associated with similar biological processes in different breast cancer subtypes, whereas each subtype has its own idiosyncratic biological processes governing the development of these tissues. These results are consistent with underlying biological processes for cancer development, which further affirms the robustness of our image processing method.

Spatial characterization of different tissues in histopathological images has shown significant diagnostic and prognostic value, but human assessment of these features is time-consuming and often infeasible for large-scale studies. This study contributes an innovative automated deep-learning analysis pipeline that will enable rapid, accurate quantification of epithelial and stromal tissues from WSI of cancer samples. Such approaches are useful because they may be used for the quantification of tissue-level epithelial/stromal/cancer phenotypes, which in turn may be integrated with other biomedical data. For this reason, we demonstrate how model-generated outputs may be correlated with gene expression and how this may lead to new insights about genetic mechanisms that contribute to tumor microenvironment variability in breast cancer. Additional contributions of this manuscript are that the approach, data, and demonstrated use of the pipeline could be applied to other cancers to improve tissue quantification. To the best of our knowledge, this is the first study to provide pixel-level tissue segmentation maps of TCGA image data.

Method

Datasets

Two breast cancer image sets were used in this study: (1) The Cancer Genome Atlas (TCGA) portal; (2) the Stanford Tissue Microarray Database (sTMA) [2]. The sTMA database consisted of a total of 157 H&E stained rectangular image regions (1128 × 720 pixels) using 20X objective lens, which were acquired from two independent cohorts: 106 samples from Netherlands Cancer Institute (NKI) and 51 samples from Vancouver General Hospital (VGH). Each image of sTMA was manually annotated with epithelial and stromal tissues by pathologists. The TCGA cohort samples include matched H&E stained WSI, gene expression data, and clinical information. Patients with missing expression data or images with cryo-artifacts deemed too severe were excluded, leaving a selected set of 1,000 samples. Since the TCGA clinical information includes subtyping information, we further categorized the selected samples into three breast cancer subtypes for more specific biological analysis: ER-positive, ER-negative and triple negative. Demographic and clinical information for both sTMA and TCGA cohorts are summarized in **Table 1**.

Overview of the workflow

Figure 1 outlines our workflow for both image processing and biological analysis. **Figure 1A** shows the detailed structure of our deep CNN model for tissue segmentation. **Figure 1B** is the whole-slide image processing pipeline. **Figure 1C** shows an overview of the biological analysis of gene expression data and image features. Details of each part are described in the following subsection.

CNN model for tissue segmentation

Given an RGB image of height H, width W, and color channels C, the goal of segmentation is to predict a label map with size $H \times W$ where each pixel is labeled with a category. CNN-based framework for segmentation fundamentally consists of encoding and decoding counterparts.

The encoding block is derived from classification models which perform down-sampling operators to capture global information from input images. Max-pooling is the most commonly adopted operations in encoding, which integrates

neighbouring pixels to learn invariance from local image transformation. More recently, dilated convolution was proposed to control spatial resolution, and thus enable dense feature extraction. Given a 1-D input signal $x[i]$ with a filter $w[k]$ of length K, the output of dilated convolution is defined as:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k] \quad (1)$$

where r is the stride in the sampling input signal, referred to as $rate$. By filling zeros between pixels in the filter, dilated convolution can enlarge receptive fields without substantially increasing computational cost.

We carefully constructed our deep hierarchical segmentation model using specific strategies in both encoder and decoder, as shown in **Figure 1A**. The ResNet-101 structure [7], which contains 101 convolution layers, was adopted as the backbone of our proposed model. Since dilated convolution inserts zeros between pixels in the filter, it can enlarge receptive fields without substantially increasing computational cost. The encoder of our model inherited the first three blocks of ResNet-101, while the rest were modified into six dilated convolution blocks, each of which further contained four ResUnits with different dilation rates. This configuration was inspired by the success of the atrous spatial pyramid pooling (DeepLab-ASPP) approach from Chen et al. [10], which captures objects as well as image context at multiple scales, and thus robustly improves the segmentation performance. In our work, the modification of convolution layers was conducted to ensure that our encoder learned both tissue structures and contextual information for the next phase of processing. In the decoding step, we adopted a multi-channel convolution approach to generate high-resolution segmentation maps. Given a feature map of dimension $h \times w \times c$, multi-channel convolution first generated features of $h \times w \times (r^2 \times c)$, where r is the upsampling rate. Then the features were reshaped to obtain upsampled features of $H' \times W' \times c, where\ H' = h \times r, W' = w \times r.$ To this end, we stretched each individual pixel in the small feature map to the channel of $r^2 \times c$ so that it corresponded to a fixed area $(r \times r)$ in the upsampled output map. We applied four parallel dilated

multi-channel convolutions with a range of dilation rates and added all of their outputs pixel by pixel in order to further exploit multi-scale contextual information from the encoding feature map.

We next used sTMA to train our CNN model in a five-folder-cross-validation. The proposed model was implemented using MXNet toolbox. Parameters in the encoder were initialized with pre-trained weights from Deep-Lab V2 [10], while the decoder layers were randomly initialized by Xavier method. Due to GPU memory limitations (8 GB for GeForce GTX 1080), we randomly cropped 600×600 patches from the raw images and performed random mirror and random crop as data augmentation in the training stage.

WSI processing pipeline

During biopsy slide examination, pathologists search for a region-of-interest (ROI) that contains cancer cells and conduct diagnostic assessment. Inspired by these human analysis steps, we built an automatic pipeline to perform tissue segmentation on WSI, as shown in **Figure 1B**. Our WSI processing pipeline consists of two parts: 1) automatic identification of ROI, and 2) epithelial and stromal tissue segmentation on the ROI. Given a WSI I, we first downsampled I into $I^{'}$ at a factor of 16 in both horizontal and vertical directions. Then we converted $I^{'}$ from RGB color space to CIELAB color space ($L^*a^*b^*$), denoted as $I^{'}_{lab}$. Since the L^* channel in $L^*a^*b^*$ color space represents the brightness, we extracted the a^* and b^* values representing color components in $I^{'}_{lab}$ and obtained a new image $I^{'}_{ab}$. Each pixel in $I^{'}_{ab}$ is then a 2-dimentional vector. Next, we applied K-means clustering algorithm (K=2) to divide the pixels of $I^{'}_{ab}$ into two groups. Considering that corners of pathology images are usually unstained, we classified pixels in the same cluster as the upper-left pixel in $I^{'}_{ab}$ as background, while the other pixels were classified as foreground. In this way, we generated a binary mask M^1, where 0 and 1 in M^1 correspond to background and foreground pixels in $I^{'}_{ab}$, respectively. Denoting the smallest rectangle region that contains the largest connected component in M^1 as

F_m, we identified the ROI F_I by mapping the coordinates of F_m onto I. Finally, F_I was cropped from I for downstream processing.

We split F_I into patches of 1128 × 720 pixels to fully utilize the proposed CNN model for tissue segmentation. Patches with more than 80% background were discarded. The retained patches were then fed into the CNN model and all the patch-level predictions were combined to generate a global tissue mask M^2 for F_I.

Tissue quantification and biological analysis

We applied our WSI processing pipeline on 1,000 TCGA breast cancer WSI for further biological analysis, as shown in **Figure 1C**. For each WSI I, we performed tissue spatial quantification based on its tissue mask M^2 derived from our method. The two tissue ratios, $Ratio_{epi}$ and $Ratio_{stro}$, that characterize the ratio of epithelial tissue areas and stromal tissue areas to overall tissue areas were estimated as:

$$Ratio_{epi} = {\sum_i^N E_i}\Big/{\sum_i^N T_i}, \; Ratio_{stro} = {\sum_i^N S_i}\Big/{\sum_i^N T_i} \quad (2)$$

where T_i, E_i and S_i represent the number of pixels classified as foreground, epithelial and stromal in the ith valid patch in F_I respectively, and N represents the total number of valid patches in F_I.

To explore the relationships between gene expression data and tissue ratios in different breast cancer subtypes, we divided all TCGA samples into three types: ER-positive, ER-negative, and triple negative, as seen in **Table 1**. Then, we computed the Spearman correlation coefficients between gene expression data and the two tissue ratios $Ratio_{epi}$ and $Ratio_{stro}$ for each breast cancer subtype. Next, we sorted all the Spearman coefficients and selected the gene symbols which were in the top 1% of correlation coefficients with $Ratio_{epi}$ and $Ratio_{stro}$ for each breast cancer subtype. For the selected gene symbols, we performed Gene Ontology (GO) enrichment analysis on them using WebGestalt [23]. Meanwhile, the Overrepresentation Enrichment Analysis (ORA) with Bonferroni adjustment methods was also used to

253 determine statistical significance of the enrichment. Genes presented by the
254 "Genome" platform were used as the reference gene. Finally, the top 10 enriched
255 biological process categories were selected to reveal the biological process underlying
256 the development of epithelial and stromal tissues for each breast cancer subtype.

257 ## Results

258 ### Validation of CNN model

259 We evaluated the effectiveness of our proposed deep CNN model on segmentation of
260 epithelial and stromal tissues by testing and comparing our model with several
261 state-of-the-art methods [11,12,24,25]. Our model outperformed all of these methods
262 based on a comparison of classification accuracies and achieved an average accuracy
263 of 91.02% on the whole sTMA dataset (NKI + VGH), as shown in **Table 2** and **Table**
264 **3**. Visual segmentation results also demonstrated that our model could accurately
265 classify epithelial and stromal tissues (**Figure 2**). Note that in the ground truth data,
266 some areas belonging to epithelia have been overlooked and incorrectly annotated as
267 background (an example is shown in the third row of **Figure 2**). However, our model
268 still yielded correct predictions on this area (marked by a black circle in **Figure 2**).
269 This indicates that our model is robust enough to make the right judgment, even under
270 misleading supervision. We believe this is valuable for future work in biomedical
271 image tasks with only partial or inaccurate annotations.

272 ### Tissue segmentation and quantification on WSI

273 We validated the trained CNN model on 171 image patches each from the TCGA
274 breast cancer slide images annotated with epithelial/stromal tissues by two domain
275 experts. The validation results indicated that our model was robust enough to predict
276 credible tissue mask for the TCGA dataset (Table T1 and Figure S1). We then applied
277 the trained CNN model to the tissue segmentation of 1,000 whole-slide images from
278 three TCGA breast cancer subtypes. Visual results showed that our pipeline could
279 robustly identify epithelial/stromal tissues in whole-slide images (**Figure 3)**.

Ratios of epithelial and stromal tissue areas to overall tissue areas were estimated based on the WSI segmentation results. Wide differences in tissue ratios were seen among different breast cancer subtypes (**Figure 4**). ER-positive images were predominantly enriched with stromal tissues with a mean stromal ratio of 72.8%, while triple negative images were abundant in epithelial tissues with a mean epithelial ratio of 63.56%. Epithelial and stromal tissues were nearly equivalent for ER-negative images with mean ratios of 49.35% and 50.65%, respectively.

Tissue-specific functional analysis

We explored which genes contributed to the development of different tissues in various subtypes of breast cancers by computing pairwise Spearman correlation coefficients between gene expression data and both tissue ratios. Genes in the top 1% of correlation with tissue ratios in each subtype of breast cancer were selected for further analysis. We then performed functional Gene Ontology (GO) analysis for the selected gene-sets. Genes correlated with the epithelial tissues were enriched in biological processes during the cell cycle, among which sister chromatid segregation, nuclear division, and mitotic cell cycle are the most commonly enriched GO terms shared by the three breast cancer subtypes. However, we also observed specifically enriched GO terms and genes for each subtype that correspond to different cell cycle stages. The Growth phase related genes including G1 phase and G2 phase were specifically enriched for the ER-positive subtype, whereas Mitotic phase genes were specifically enriched for the triple negative subtype, and S phase related genes were specific for the ER-negative subtype.

Similarly, such patterns of shared high-level biological processes with specific functions were also observed for the stromal tissues. For the stromal tissue, the most significantly enriched GO biological process terms were all related to the development of the tumor microenvironment, including vasculature development, cellular component movement, and growth factor stimuli-related GO functions which were shared among the three breast cancer subtypes. For the ER-positive subtype,

angiogenesis-related genes were specifically enriched, while for the triple negative subtype, muscle structure genes (especially the ones related to actin fibers and cytoskeleton) were specifically enriched. In addition, for the ER-negative subtype, growth factor genes were enriched. Altogether, our results (**Figure 5**) suggest that even though the same tissue was associated with similar biological processes in different subtypes, each subtype still had its idiosyncratic biological processes governing the development of these tissues.

Other applications

Our WSI processing pipeline can be easily applied to histological images of other types of cancers. The global tissue segmentation maps we have presented could also be used for other more specific computational analysis. For example, global morphological features of different tissues could be estimated for better survival prediction [22,26], and lymphocytes in different tissues could be distinguished for observation of more detailed immune response. Imaging data resources have not been exploited to the degree of the other TCGA molecular and clinical outcome resources, likely because automatic image annotation is still impeded by data volume challenges. In this manuscript we presented global tissue maps of all the TCGA breast cancer WSI, and it is our aspiration that they will facilitate further exploration and utilization of these imaging data for various cancers.

Conclusions

Epithelial and stromal regions of tumors, as well as their spatial characterizations in histopathology images, play a very important role in cancer diagnosis, prognosis, and treatment. Recently, some research studies have focused on developing systems for automatically analyzing H&E stained histological images from tissue microarrays in order to predict prognosis [26,27]. In contrast, our approach is aimed at whole slide images (WSI) rather than manually extracted regions since WSI provide much more comprehensive information, including heterogeneity. Mackie et al. [28] summarized the research progress and challenges facing the application of big data quantitative

imaging to cancer treatment, focusing on 3D imaging modalities including CT, PET, and MRI. Our quantitative analysis of histopathology images complements and extends this work in terms of data modality and size, application areas, and computational challenges.

Based on our global tissue quantification, distinct differences were observed in the enriched GO terms for epithelial and stromal tissues [29]. At the same time, highly overlapping biological properties were observed in the same tissue across different subtypes, all tied to cancer progression in one way or another. For example, in epithelial tissue, genes from cell cycle-related processes were significantly enriched. Previous studies have addressed that sustaining proliferative signaling is one of the hallmarks of cancer, during which cell cycle plays quite an important role [30]. In addition, *CDK4/6* inhibitors (such as Palbociclib and ribociclib) target this biological process [31,32]. For stromal tissue, genes related to the tumor microenvironment were significantly enriched (e.g., vasculature and locomotion). Vasculature is vital for inducing angiogenesis, which is another important hallmark of cancer.

Additionally, we observed differences in biological processes between different subtypes resulting from tumor heterogeneity. Specific biological process features for each subtype were also identified among the same tissue. For epithelial tissue, genes associated with different stages of the cell cycle were specifically enriched for different subtype. For ER-positive breast epithelia, we found that G1 and G2 phase-related GO terms were enriched, among which G2/M transition is an important element. Wang et al. [27] have highlighted the importance of G2/M transition in ER-positive breast cancer. For the triple negative subtype, we found that M phase related GO terms were enriched, during which chromosome segregation plays a key role. Witkiewicet et al. [33] have shown the close relationship between chromosome segregation (*PLK1*) with triple negative Breast Cancer. Similarly, angiogenesis related biological processes were significantly associated with the stroma of the ER-positive subtype. Previous studies have indicated that vasculature is one of the

364 important components for tumor stroma [34], as stromal cells can build blood vessels
365 to supply oxygen and nutrients [35].
366 While the correlation analysis of this study reveals clear pairwise relationships
367 between morphological and genomic features, there are two major limitations to our
368 approach. First, correlation cannot reveal highly nonlinear relationships or
369 multivariate complication relationships. For instance, Wang et al. [36] demonstrated
370 that complicated morphological features might need to be modeled using multiple
371 genomic features, implying contributions from multiple genetic factors. Similarly,
372 with our data, more sophisticated analysis such as nonlinear correlation analysis can
373 be applied to reveal deeper relationships. Secondly, correlation is not causation. The
374 genes that are strongly correlated with the stromal or epithelial content may not be the
375 underlying driver genes for the development of the tissues. Identification of such key
376 genes requires further incorporation of biological knowledge, as well as future
377 experimental validation.
378 In summary, our framework provides not only fully automatic and detailed analysis
379 for large H&E stained images based on a state-of-the-art deep learning model, but
380 also integrated analysis of image features and molecular data. The proposed
381 framework enables us to effectively explore the underlying relationships between
382 gene expression and tissue quantification, free from the extensive labelling and
383 annotation that is laborious even to skilled pathologists.

The details about code and data in this manuscript is provided on Github with the link at *https://github.com/Serian1992/ImgBio.*

Authors' contributions

LZ carried out the pathology image processing, participated in the genetic studies and drafted the manuscript. ZX carried out the enrichment analysis and helped to draft the manuscript. WY participated in the development of methodology. CJ participated in the acquisition of data. SW participated in the development of methodology. HZ participated in the acquisition of data and the development of methodology. ZJ and DN participated in the review and revision of the manuscript. FQ participated in the development of methodology and helped to review and revise the manuscript. HK conceived of the study, and participated in its design and coordination, reviewed and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

This work was supported by Indiana University Precision Health Initiative to HK and ZJ, the NSFC-Guangdong United Found of China (No. U1501256) to FQ, and Shenzhen Peacock Plan (No. KQTD2016053112051497) to ZX and DN. We thank Dr. Natalie Lambert, Dr. Bryan Helm and Ms. Megan Metzger for their tremendous help in the discussion and editing of the manuscript.

Reference

[1] Arendt LM, Rudnick JA, Keller PJ, Kuperwasser C. Stroma in breast development and disease. Semin Cell Dev Biol 2010;21:11–8.

[2] de Kruijf EM, van Nes JGH, van de Velde CJH, Putter H, Smit VTHBM, Liefers GJ, et al. Tumor--stroma ratio in the primary tumor is a prognostic factor in early breast cancer patients, especially in triple-negative carcinoma patients. Breast Cancer Res Treat 2011;125:687–96.

[3] Toss MS, Miligy I, Al-Kawaz A, Alsleem M, Khout H, Rida PC, et al. Prognostic significance of tumor-infiltrating lymphocytes in ductal carcinoma in situ of the breast. Mod Pathol 2018;31(8):1226.

[4] Fouad S, Randell D, Galton A, Mehanna H, Landini G. Epithelium and Stroma Identification in Histopathological Images Using Unsupervised and Semi-Supervised Superpixel-Based Segmentation. J Imaging 2017;3.

[5] Haridas A, Bunyak F, Palaniappan K. Interactive Segmentation Relabeling for Classification of Whole-Slide Histopathology Imagery. 2015 IEEE 28th Int. Symp. Comput. Med. Syst., 2015, p. 84–7.

[6] Bunyak F, Hafiane A, Al-Milaji Z, Ersoy I, Haridas A, Palaniappan K. A segmentation-based multi-scale framework for the classification of epithelial and stromal tissues in H E images. 2015 IEEE Int. Conf. Bioinforma. Biomed., 2015, p. 450–3.

[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. pattern Recognit., 2016, 39(7): 1476-81.

[8] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Adv. Neural Inf. Process. Syst. 28, Curran Associates, Inc.; 2015, p. 91–9.

[9] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640–51.

[10] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 2018;40:834–48.

[11] Al-Milaji Z, Ersoy I, Hafiane A, Palaniappan K, Bunyak F. Integrating segmentation with deep learning for enhanced classification of epithelial and stromal tissues in H&E images. Pattern Recognit Lett 2019;119:214–21.

[12] Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016;191:214–23.

[13] Farahani N, Parwani A V, Pantanowitz L. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol Lab Med Int 2015;7:23–33.

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2015;9351:234–41.

[15] Bejnordi BE, Zuidhof GCA, Balkenhol M, Hermsen M, Bult P, van Ginneken B, et al. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. CoRR 2017;abs/1705.0.

[16] Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, p. 2424–33.

[17] Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep 2018;23:181.

[18] Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. Nat Methods 2013;10:1046.

[19] Akbani R, Ng PKS, Werner HMJ, Shahmoradgoli M, Zhang F, Ju Z, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat Commun 2014;5:3887.

[20] Cheng J, Mo X, Wang X, Parwani A, Feng Q, Huang K. Identification of topological features in renal tumor microenvironment associated with patient survival. Bioinformatics 2017;34:1024–30.

[21] Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer Res 2017;77:e91--e100.

[22] Shao W, Cheng J, Sun L, Han Z, Feng Q, Zhang D, et al. Ordinal Multi-modal Feature Selection for Survival Analysis of Early-Stage Renal Cancer: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II, 2018, p. 648–56.

[23] Wang J, Vasaikar S V, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res., 2017;45(W1):W130-W137.

[24] Du Y, Zhang R, Zargari A, Thai TC, Gunderson CC, Moxley KM, et al. A performance comparison of low-and high-level features learned by deep convolutional neural networks in epithelium and stroma classification. Med. Imaging 2018 Digit. Pathol., vol. 10581, 2018, p. 1058116.

[25] Vu QD, Kwak JT. A Dense Multi-Path Decoder for Tissue Segmentation in Histopathology Images. Comput Methods Programs Biomed 2019;173:119--129.

[26] Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. Sci Transl Med 2011;3:108ra113-108ra113.

[27] Wang C, Pécot T, Zynger DL, Machiraju R, Shapiro CL, Huang K. Identifying survival associated morphological features of triple negative breast cancer

using multiple datasets. J Am Med Informatics Assoc 2013;20:680–7.

[28] Mackie TR, Jackson EF, Giger M. Opportunities and challenges to utilization of quantitative imaging: Report of the AAPM practical big data workshop. Med Phys 2018;45:e820–8.

[29] Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 2018;173:291–304.e6.

[30] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell 2011;144:646–74.

[31] Rocca A, Farolfi A, Bravaccini S, Schirone A, Amadori D. Palbociclib (PD 0332991): targeting the cell cycle machinery in breast cancer. Expert Opin Pharmacother 2014;15:407–20.

[32] Murphy CG, Dickler MN. The role of CDK4/6 inhibition in breast cancer. Oncologist 2015;20:483–90.

[33] Witkiewicz AK, Chung S, Brough R, Vail P, Franco J, Lord CJ, et al. Targeting the Vulnerability of RB Tumor Suppressor Loss in Triple-Negative Breast Cancer. Cell Rep 2018;22:1185–99.

[34] Bremnes RM, D\onnem T, Al-Saad S, Al-Shibli KI, Andersen S, Sirera R, et al. 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 1:209–17.

[35] Ghesquière B, Wong BW, Kuchnio A, Carmeliet P. Metabolism of stromal and immune cells in health and disease. Nature 2014;511:167–76.

[36] Wang C, Su H, Yang L, Huang K. Integrative analysis for lung adenocarcinoma predicts morphological features associated with genetic variations. PACIFIC Symp. Biocomput. 2017, 2017, p. 82–93.

Figure legends

Figure 1 Workflow for image processing and biological analysis

A. Detailed structure of our deep CNN model for segmentation. **B.** Whole-slide image processing pipeline. **C.** Overview of biological analysis of gene expression data and image features.

Figure 2 Segmentation results on TMA

Column (a) are raw images; column (b) are annotations by pathologists; column (c) are predictions of the proposed model. Red, green and black areas in column (b) and (c) represent epithelial, stromal and background regions in raw images, respectively. Note that in the last row, the overlooked tumor area (Marked with black circle) is still well recognized by our model.

Figure 3 Segmentation results on TCGA WSIs

For each TCGA whole-slide image **A, B, C: Step 1** represents the WSI; **Step 2** represents the background map of WSI; **Step 3** represents the region of interest (ROI) in the WSI of raw resolution; **Step 4** represents the tissue segmentation result of ROI. Red, green and black areas in **Step 4** represent the predicted epithelial, stromal and background regions, respectively.

Figure 4 Tissue distribution on different breast cancer subtypes

The Variable Epithelial_ratio, Stromal_ratio represent the ratios of epithelial tissue areas and stromal tissue areas to overall tissue areas, respectively.

Figure 5 Results of GO enrichment analysis

Dots represent most significantly enriched Biological Process term for each cancer subtype with color coding: purple indicates high enrichment, red indicates low enrichment. Sizes of dots represent the ratio of enrichment (GO category). FDR is the method used for multiple comparison correction.

Tables

Table 1 Demographic and clinical characteristics

Note: For TCGA cohort, samples in Triple Negative subgroup also belong to ER-negative subgroup.

Table 2 Evaluation of CNN model on NKI and VGH

Note: From the third to the last column are the ten evaluations metrics. Value in bold represents the best result under each metric among different models.

* TPR (True positive rate) = TP / (TP + FN); TNR (True negative rate) = TN / (FP + TN); PPV (Positive Predictive Value) = TP / (TP + FP); NPV (Negative Predictive Value) = TN / (FN + TN); FPR (False positive rate) = FP / (FP + TN); FDR (False Discovery Rate) = 1 - TP / (TP + FP); FNR(False Negative Rate) = FN / (FN + TP); ACC (Accuracy) = (TP + TN) / (TP + FP + TN + FN); F1_score = 2∗TP / (2∗TP + FP + FN); MCC (Mattews Correlation Coefficient) = $(TP * TN - FP * FN) / \sqrt{TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$. TP, FP, TN and FN represent the true positive, false positive, true negative and false negative, respectively.

Table 3 Quantitative evaluation on the whole TMA dataset

Supplementary material

Figure S1 Qualitative segmentation results on TCGA dataset

The first column are the raw TCGA images; the second column are annotations by pathologists; the third column are predictions of the proposed model. Red, green and black areas in the annotations and predictions represent epithelial, stromal and background regions in raw images, respectively.

Table T1 Quantitative evaluation on TCGA dataset

* TPR (True positive rate) = TP / (TP + FN); TNR (True negative rate) = TN / (FP + TN); FPR (False positive rate) = FP / (FP + TN); FNR(False Negative Rate) = FN / (FN + TP); ACC (Accuracy) = (TP + TN) / (TP + FP + TN + FN); F1_score = 2∗TP / (2∗TP + FP + FN). TP, FP, TN and FN represent the true positive, false positive, true negative and false negative, respectively.

Table 1 Demographic and clinical characteristics

Cohort	SubGroup	Image Type	Image Number	Total
TMA	NKI	H&E stained image region (1128 * 720)	106	157
	VGH		51	
TCGA	ER-positive	Whole-slide image	773	1000
	ER-negative		227	
	Triple negative		112	

Note: For TCGA cohort, samples in Triple Negative subgroup also belong to ER-negative subgroup.

Table 2 Evaluation of CNN model on NKI and VGH

Datasets	Models	TPR	TNR	PPV	NPV	FPR	FDR	FNR	ACC	F1	MCC
NKI	Xu.et [12]	86.31	82.15	84.11	84.60	17.85	15.89	13.66	84.34	85.21	68.60
	CNN only [11]	81.34	82.89	84.11	80.05	17.11	15.89	18.57	81.69	82.75	64.24
	CNN+HFCM [11]	89.48	85.96	85.94	89.50	14.04	14.06	10.52	87.19	87.68	75.44
	Our model	**90.71**	**89.83**	**90.81**	**89.72**	**10.17**	**9.19**	**9.29**	**90.29**	**90.76**	**80.54**
VGH	Xu.et [12]	88.29	88.40	89.93	86.55	11.60	10.07	11.71	88.34	89.10	76.59
	CNN only [11]	90.32	88.15	92.98	83.97	11.85	7.02	9.68	89.14	91.63	77.70
	CNN+HFCM [11]	**91.96**	**92.21**	**95.45**	86.59	**7.79**	**4.55**	**8.04**	91.04	**93.67**	**83.10**
	Our model	91.37	91.49	92.37	**90.38**	8.51	7.63	8.63	**91.42**	91.87	82.80

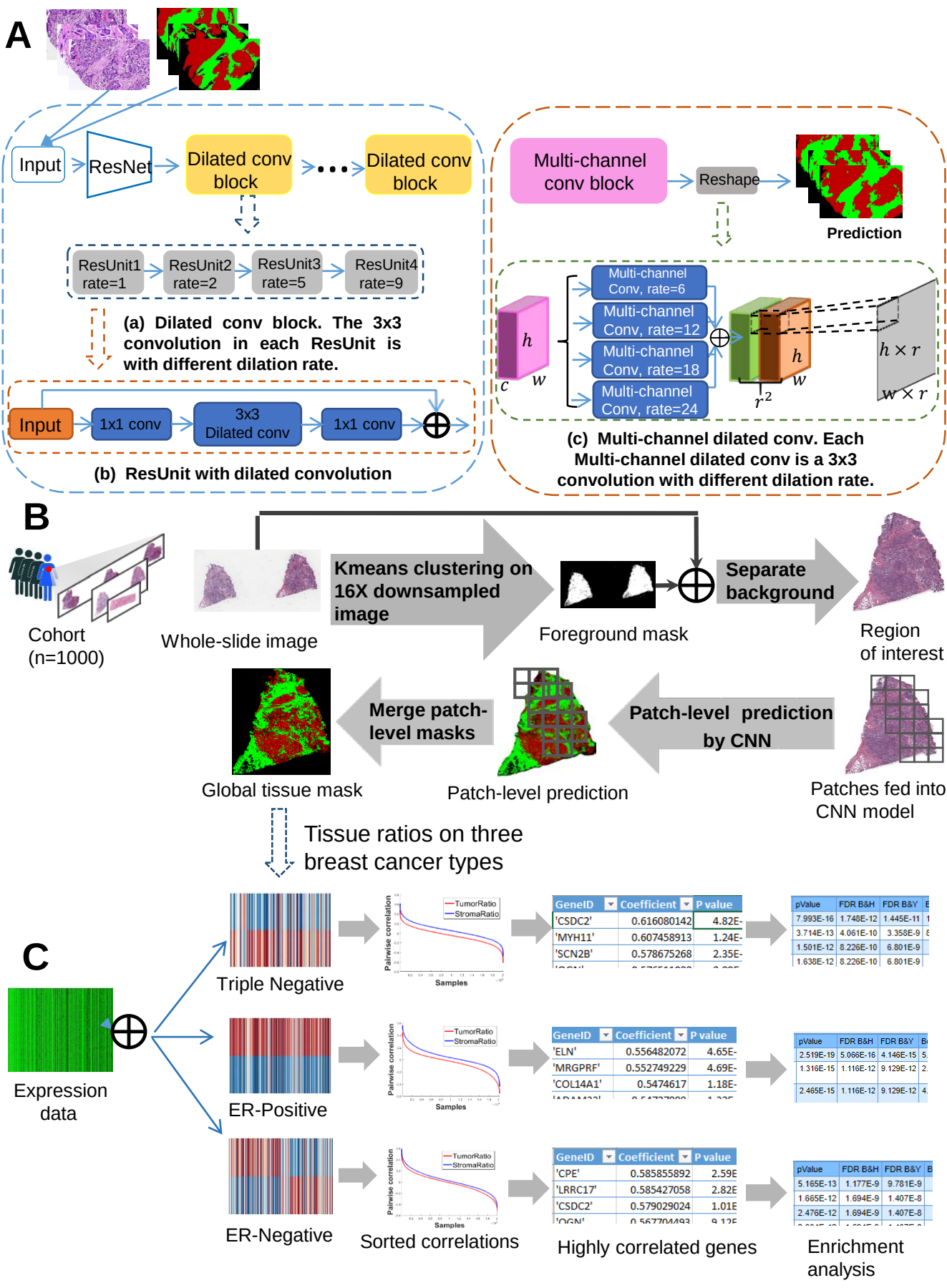
Note: From the third to the last column are the ten evaluations metrics. Value in bold represents the best result under each metric among different models.

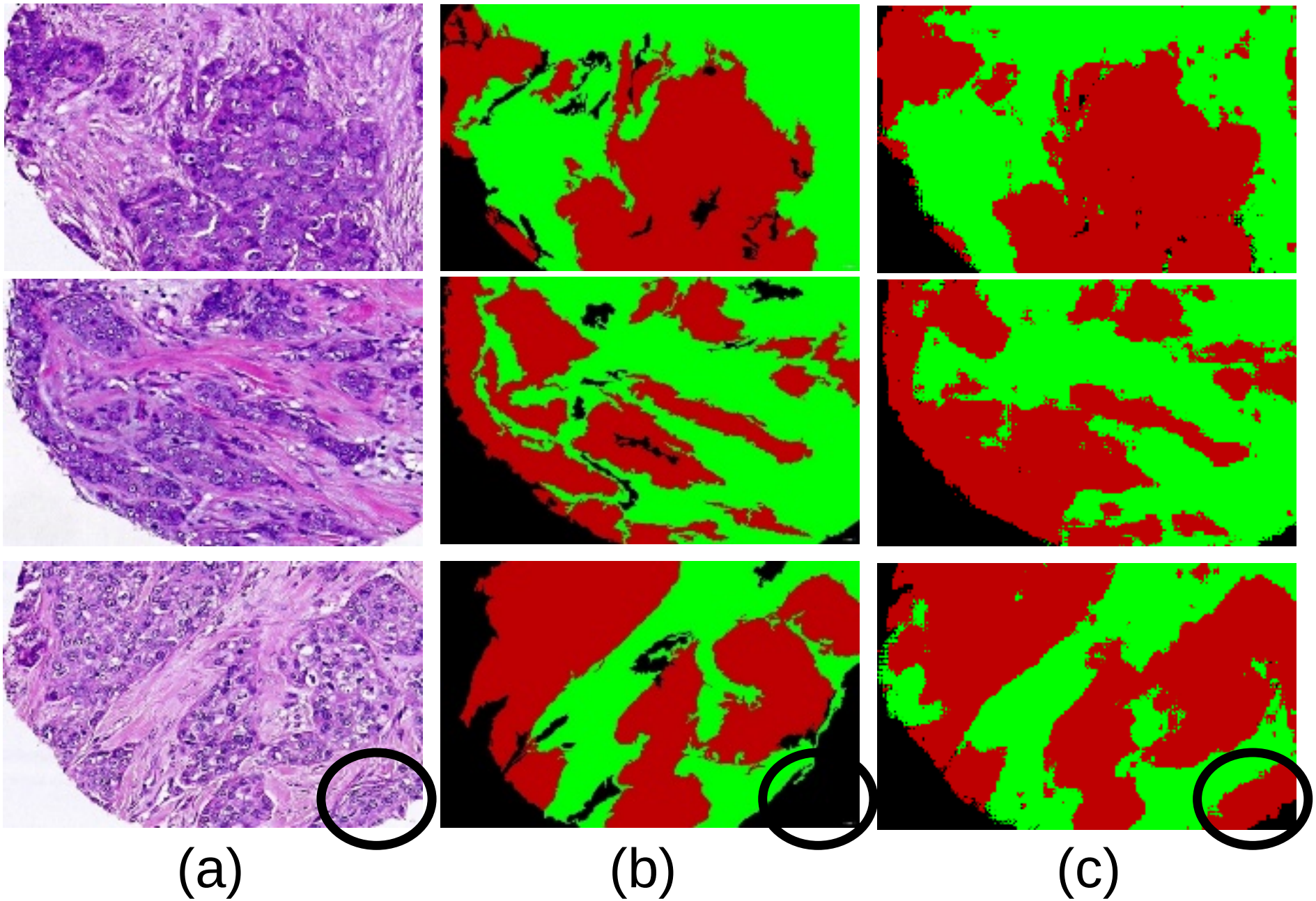
*TPR (True positive rate) = TP / (TP + FN); TNR (True negative rate) = TN / (FP + TN); PPV (Positive Predictive Value) = TP / (TP + FP); NPV (Negative Predictive Value) = TN / (FN + TN); FPR (False positive rate) = FP / (FP + TN); FDR (False Discovery Rate) = 1 - TP / (TP + FP); FNR(False Negative Rate) = FN / (FN + TP); ACC (Accuracy) = (TP + TN) / (TP + FP + TN + FN); F1_score = 2∗TP / (2∗TP + FP + FN); MCC (Mattews Correlation Coefficient) = $(TP * TN - FP * FN) / \sqrt{TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$. TP, FP, TN and FN represent the true positive, false positive, true negative and false negative, respectively.

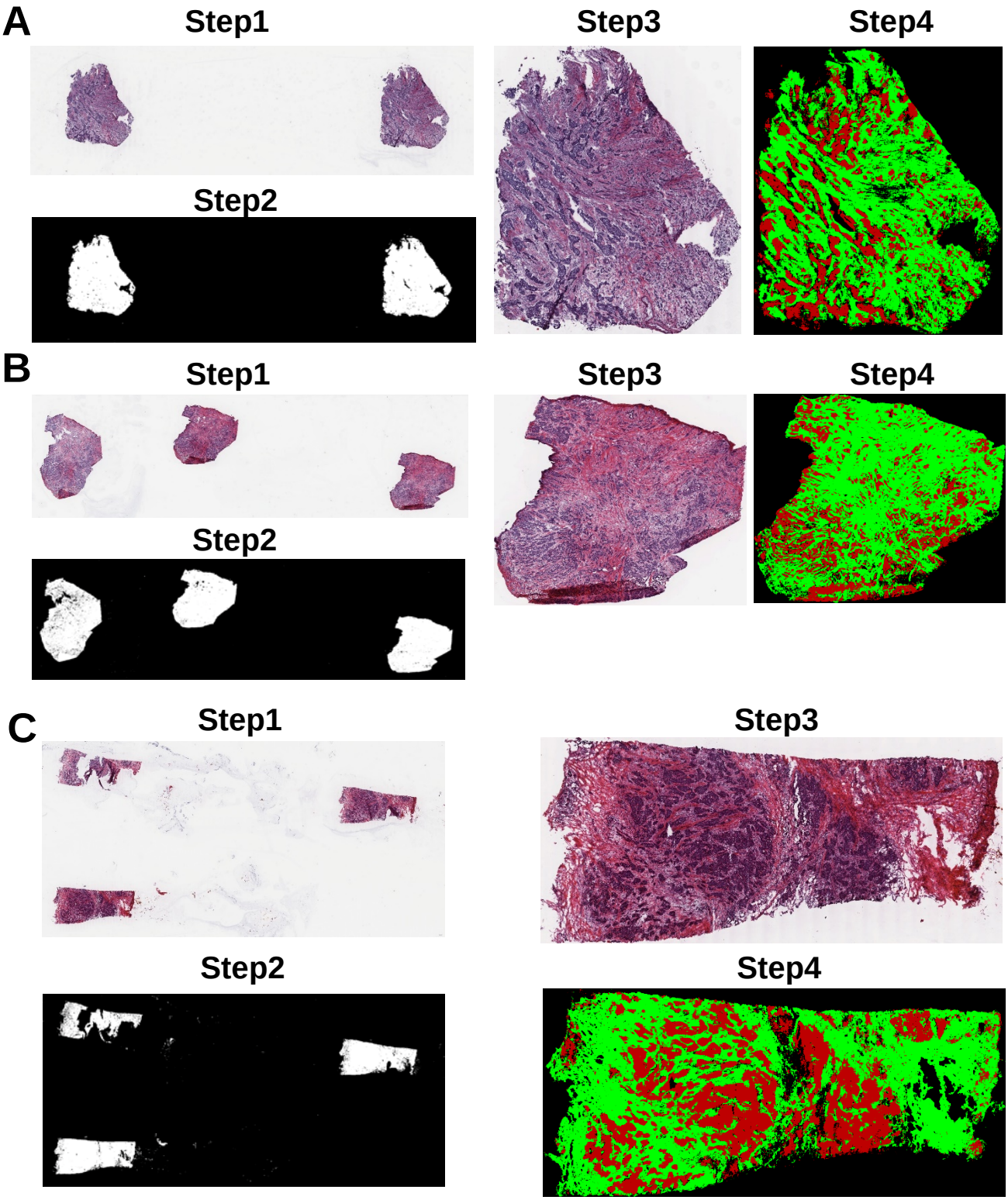
584 **Table 3 Quantitative evaluation on the whole TMA dataset**

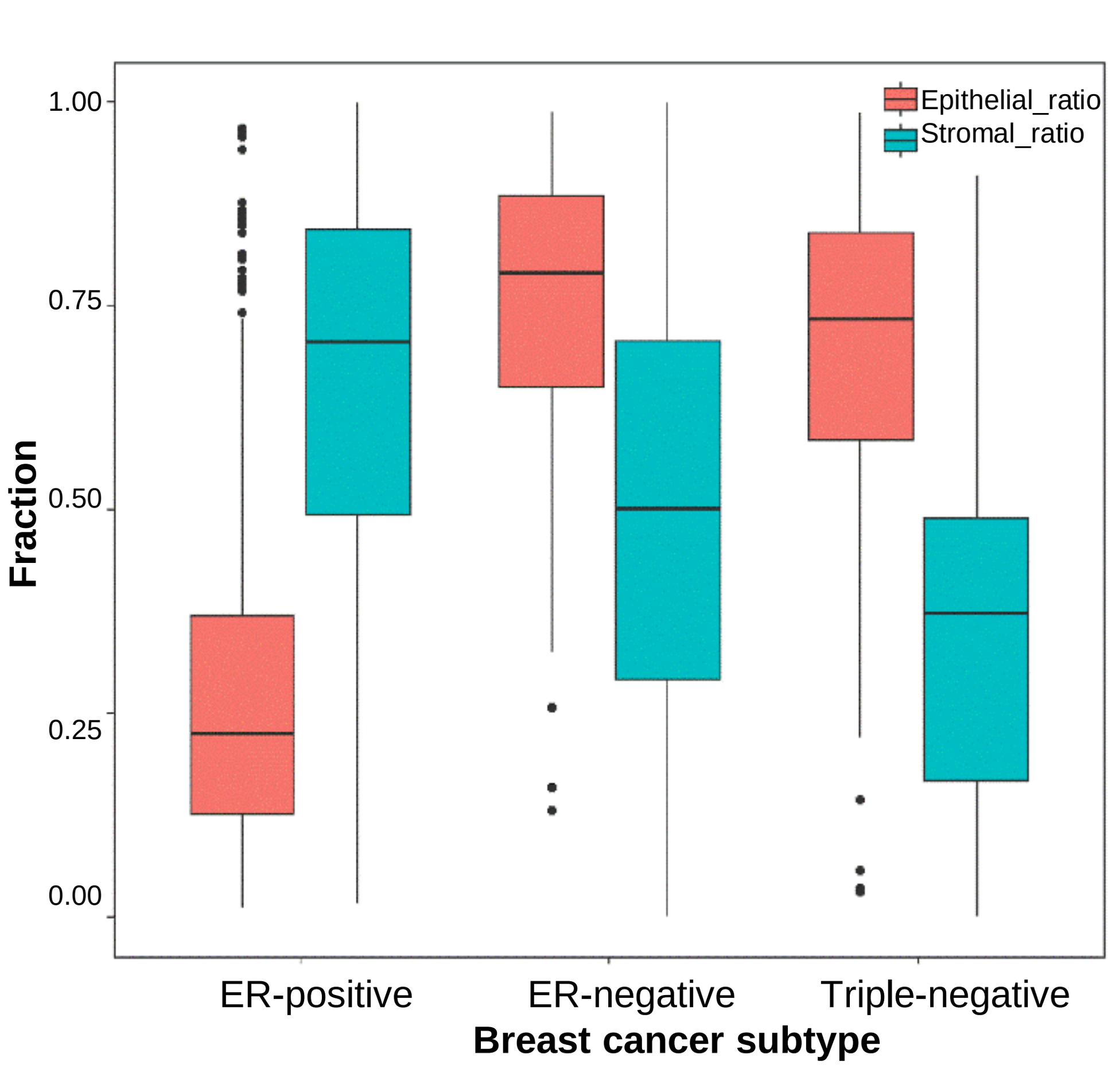
Dataset	Model	ACC	F1_score
NKI +VGH	Du.et [24]	89.7	89.7
	Vu.et [25]	90.315	90.51
	Our model	91.02	91.59

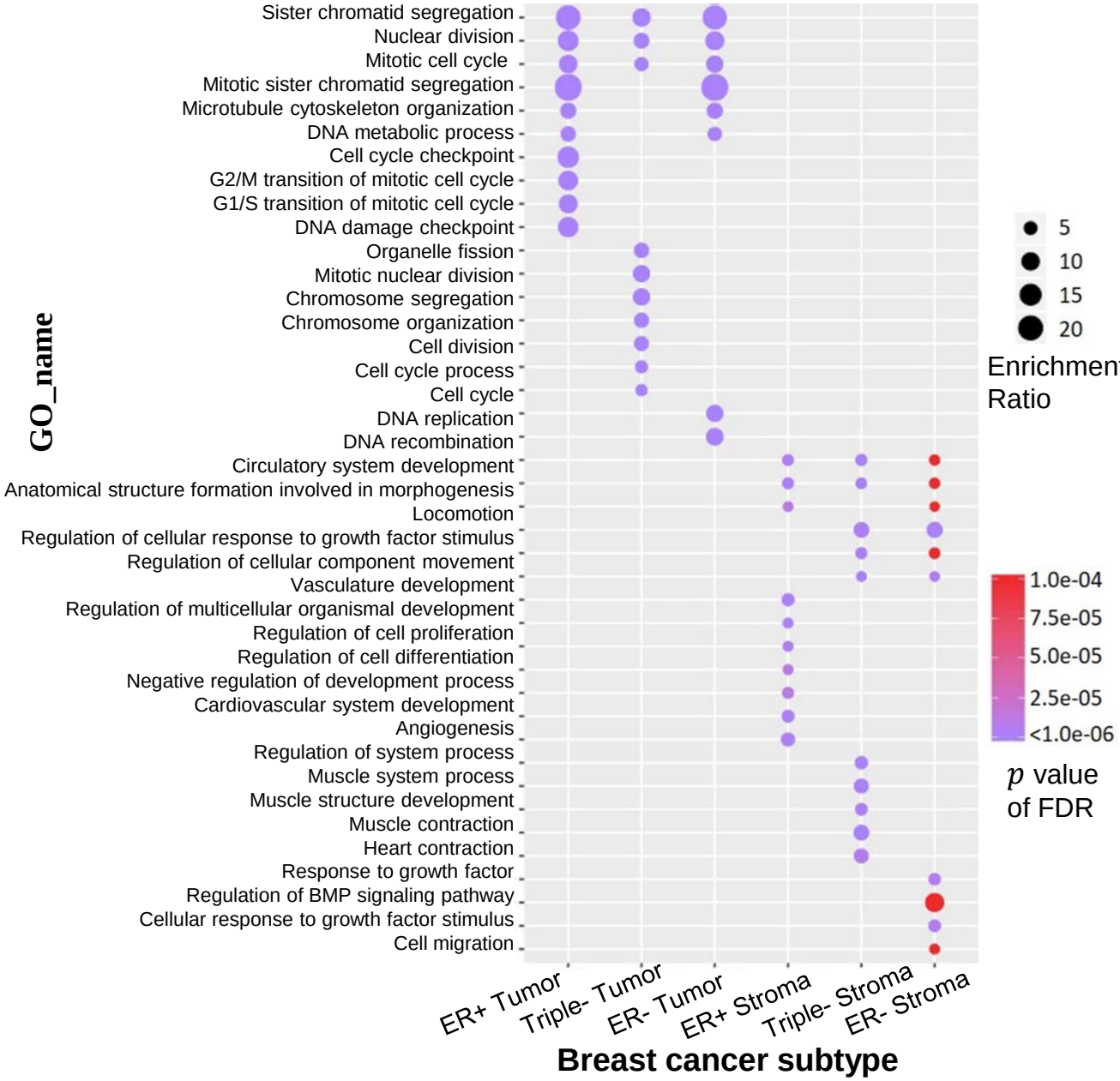
585


A
Input
ResNet
Dilated conv block
Dilated conv block
ResUnit1 rate=1
ResUnit2 rate=2
ResUnit3 rate=5
ResUnit4 rate=9
(a) Dilated conv block. The 3x3 convolution in each ResUnit is with different dilation rate.
Input
1x1 conv
3x3 Dilated conv
1x1 conv
(b) ResUnit with dilated convolution
Multi-channel conv block
Reshape
Prediction
Multi-channel Conv, rate=6
Multi-channel Conv, rate=12
Multi-channel Conv, rate=18
Multi-channel Conv, rate=24
(c) Multi-channel dilated conv. Each Multi-channel dilated conv is a 3x3 convolution with different dilation rate.
B
Cohort (n=1000)
Whole-slide image
Kmeans clustering on 16X downsampled image
Foreground mask
Separate background
Region of interest
Global tissue mask
Merge patch-level masks
Patch-level prediction
Patch-level prediction by CNN
Patches fed into CNN model
Tissue ratios on three breast cancer types
C
Expression data
Triple Negative
ER-Positive
ER-Negative
TumorRatio
StromaRatio
Pairwise correlation
Samples
Sorted correlations
GeneID Coefficient P value
'CSDC2' 0.616080142 4.82E-
'MYH11' 0.607458913 1.24E-
'SCN2B' 0.578675268 2.35E-
'ELN' 0.556482072 4.65E-
'MRGPRF' 0.552749229 4.69E-
'COL14A1' 0.5474617 1.18E-
'CPE' 0.585855892 2.59E
'LRRC17' 0.585427058 2.82E
'CSDC2' 0.579029024 1.01E
Highly correlated genes
pValue FDR B&H FDR B&Y
7.993E-16 1.748E-12 1.445E-11
3.714E-13 4.061E-10 3.358E-9
1.501E-12 8.226E-10 6.801E-9
1.638E-12 8.226E-10 6.801E-9
2.519E-19 5.066E-16 4.146E-15
1.316E-15 1.116E-12 9.129E-12
2.465E-15 1.116E-12 9.129E-12
5.165E-13 1.177E-9 9.781E-9
1.665E-12 1.694E-9 1.407E-8
2.476E-12 1.694E-9 1.407E-8
Enrichment analysis


(a)
(b)
(c)


A
Step1
Step3
Step4
Step2
B
Step1
Step3
Step4
Step2
C
Step1
Step3
Step2
Step4


Epithelial_ratio
Stromal_ratio
1.00
0.75
0.50
0.25
0.00
Fraction
ER-positive
ER-negative
Triple-negative
Breast cancer subtype


Sister chromatid segregation
Nuclear division
Mitotic cell cycle
Mitotic sister chromatid segregation
Microtubule cytoskeleton organization
DNA metabolic process
Cell cycle checkpoint
G2/M transition of mitotic cell cycle
G1/S transition of mitotic cell cycle
DNA damage checkpoint
Organelle fission
Mitotic nuclear division
Chromosome segregation
Chromosome organization
Cell division
Cell cycle process
Cell cycle
DNA replication
DNA recombination
Circulatory system development
Anatomical structure formation involved in morphogenesis
Locomotion
Regulation of cellular response to growth factor stimulus
Regulation of cellular component movement
Vasculature development
Regulation of multicellular organismal development
Regulation of cell proliferation
Regulation of cell differentiation
Negative regulation of development process
Cardiovascular system development
Angiogenesis
Regulation of system process
Muscle system process
Muscle structure development
Muscle contraction
Heart contraction
Response to growth factor
Regulation of BMP signaling pathway
Cellular response to growth factor stimulus
Cell migration
GO_name
ER+ Tumor
Triple- Tumor
ER- Tumor
ER+ Stroma
Triple- Stroma
ER- Stroma
Breast cancer subtype
5
10
15
20
Enrichment Ratio
1.0e-04
7.5e-05
5.0e-05
2.5e-05
<1.0e-06
p value of FDR