1 **A Deep Learning Approach for Tissue Spatial Quantification**

2 **and Genomic Correlations of Histopathological Images**

3 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

4 Han^{3,g}, Jie Zhang^{4,h}, Qianjin Feng^{1,*,i}, Kun Huang^{3,*,j}

5 ¹ *Guangdong Provincial Key Laboratory of Medical Image Processing, School of*

6 *Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China*

7 ² *National-Regional Key Technology Engineering Laboratory for Medical Ultrasound,*

8 *School of Biomedical Engineering, Health Science Center, Shenzhen University,*

9 *Shenzhen, 518060, China*

10 ³ *Regenstrief Institute, Department of Medicine, Indiana University School of*

11 *Medicine, Indianapolis, IN, 46202, USA*

12 ⁴ *Department of Medical and Molecular Genetics, Indiana University School of*

13 *Medicine, Indianapolis, IN, 46202, USA*

14 [#]Co-first authors

15 ^{*}Corresponding authors

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

17 ^aORCID: 0000-0003-0809-8703

18 ^bORCID: 0000-0003-1326-6601

19 ^cORCID: 0000-0003-3838-7418

20 ^dORCID: 0000-0001-5493-961X

21 ^eORCID: 0000-0002-9401-1186

22 ^fORCID: 0000-0002-9146-6003

23 ^gORCID: 0000-0002-5603-8433

24 ^hORCID: 0000-0001-6939-7905

25 ⁱORCID: 0000-0001-8647-0596

26 ^jORCID: 0000-0002-8530-370X

27

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

36 **Abstract**

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

56

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

59 Introduction

60 Most solid tumors are composed of many tissue types including cancer cells, stroma,
61 and epithelium. The interaction of tissues within such complex neoplasms defines the
62 tumor microenvironment and this variably contributes to cancer initiation,
63 progression, and therapeutic responses. For example, breast cancer epithelial cells of
64 the mammary ducts are commonly the site of tumor initiation, while stromal tissue
65 dynamics drive invasion and metastasis [1]. Tumor-to-stroma ratios of H&E stained
66 images are therefore an important prognostic factor [2,3], and distinguishing stromal
67 from epithelial tissue in histological images constitutes a basic, but crucial, task for
68 cancer pathology. Classification methods (*i.e.* pre-processing, training classifiers with
69 carefully selected features, and patch-level classification) are the most common
70 automated computational methods for tissue segmentation [4,5]. For instance, Bunyak
71 et al. [6] combined traditional feature selection methods and classification methods to
72 perform segmentation of epithelial and stromal tissues on a tissue microarray (TMA)
73 database. While this approach is viable, it can be time-consuming and inefficient
74 given the feature selection process. Convolutional Neural Networks (CNN) models
75 have the potential to improve analysis time and performance. Recently, deep CNN
76 models have greatly boosted the performance of natural image analysis techniques
77 such as image classification [7], object detection [8] and semantic segmentation
78 [9,10], and biomedical image analysis [11–13]. Additionally, Ronneberger et al. [14]
79 proposed implementation of a U-Net architecture to capture context and a symmetric
80 expanding path that enables precise localization in biomedical image segmentation.
81 CNN models have also been combined with traditional approaches to enhance the
82 segmentation performance of epithelial and stromal regions [11,12].
83 Despite breakthroughs in the application of CNN models to medical image analysis,
84 automated classification of epithelial and stromal tissues in Whole Slide Tissue
85 Images (WSI) remains challenging due to the large size of WSI. WSI contain billions
86 of pixels, and machine learning methods are limited by the technical hurdles of

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

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

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

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

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

140 **Method**

141 **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-dimensional 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

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

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

232 **Tissue quantification and biological analysis**

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

$$239 \quad Ratio_{epi} = \frac{\sum_i^N E_i}{\sum_i^N T_i}, \quad Ratio_{stro} = \frac{\sum_i^N S_i}{\sum_i^N T_i} \quad (2)$$

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

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

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

Results

Validation of CNN model

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

Tissue segmentation and quantification on WSI

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

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

287 **Tissue-specific functional analysis**

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

302 Similarly, such patterns of shared high-level biological processes with specific
 303 functions were also observed for the stromal tissues. For the stromal tissue, the most
 304 significantly enriched GO biological process terms were all related to the
 305 development of the tumor microenvironment, including vasculature development,
 306 cellular component movement, and growth factor stimuli-related GO functions which
 307 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

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

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

351 Additionally, we observed differences in biological processes between different
352 subtypes resulting from tumor heterogeneity. Specific biological process features for
353 each subtype were also identified among the same tissue. For epithelial tissue, genes
354 associated with different stages of the cell cycle were specifically enriched for
355 different subtype. For ER-positive breast epithelia, we found that G1 and G2
356 phase-related GO terms were enriched, among which G2/M transition is an important
357 element. Wang et al. [27] have highlighted the importance of G2/M transition in
358 ER-positive breast cancer. For the triple negative subtype, we found that M phase
359 related GO terms were enriched, during which chromosome segregation plays a key
360 role. Witkiewicz et al. [33] have shown the close relationship between chromosome
361 segregation (*PLK1*) with triple negative Breast Cancer. Similarly, angiogenesis
362 related biological processes were significantly associated with the stroma of the
363 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.

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

386

387 **Authors' contributions**

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

397

398 **Competing interests**

399 The authors have declared no competing interests.

400

401 **Acknowledgements**

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

407 **Reference**

- 408 [1] Arendt LM, Rudnick JA, Keller PJ, Kuperwasser C. Stroma in breast
409 development and disease. *Semin Cell Dev Biol* 2010;21:11–8.
- 410 [2] de Kruijf EM, van Nes JGH, van de Velde CJH, Putter H, Smit VTHBM,
411 Liefers GJ, et al. Tumor--stroma ratio in the primary tumor is a prognostic
412 factor in early breast cancer patients, especially in triple-negative carcinoma
413 patients. *Breast Cancer Res Treat* 2011;125:687–96.
- 414 [3] Toss MS, Miligy I, Al-Kawaz A, Alsleem M, Khout H, Rida PC, et al.
415 Prognostic significance of tumor-infiltrating lymphocytes in ductal carcinoma
416 in situ of the breast. *Mod Pathol* 2018;31(8):1226.
- 417 [4] Fouad S, Randell D, Galton A, Mehanna H, Landini G. Epithelium and Stroma
418 Identification in Histopathological Images Using Unsupervised and
419 Semi-Supervised Superpixel-Based Segmentation. *J Imaging* 2017;3.
- 420 [5] Haridas A, Bunyak F, Palaniappan K. Interactive Segmentation Relabeling for
421 Classification of Whole-Slide Histopathology Imagery. 2015 IEEE 28th Int.
422 Symp. Comput. Med. Syst., 2015, p. 84–7.
- 423 [6] Bunyak F, Hafiane A, Al-Milaji Z, Ersoy I, Haridas A, Palaniappan K. A
424 segmentation-based multi-scale framework for the classification of epithelial
425 and stromal tissues in H E images. 2015 IEEE Int. Conf. Bioinforma. Biomed.,
426 2015, p. 450–3.
- 427 [7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition.
428 *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, 39(7): 1476-81.
- 429 [8] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object
430 Detection with Region Proposal Networks. In: Cortes C, Lawrence ND, Lee
431 DD, Sugiyama M, Garnett R, editors. *Adv. Neural Inf. Process. Syst.* 28,
432 Curran Associates, Inc.; 2015, p. 91–9.
- 433 [9] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic
434 Segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017;39:640–51.

- 435 [10] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab:
436 Semantic image segmentation with deep convolutional nets, atrous
437 convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell*
438 2018;40:834–48.
- 439 [11] Al-Milaji Z, Ersoy I, Hafiane A, Palaniappan K, Bunyak F. Integrating
440 segmentation with deep learning for enhanced classification of epithelial and
441 stromal tissues in H&E images. *Pattern Recognit Lett* 2019;119:214–21.
- 442 [12] Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional
443 Neural Network for segmenting and classifying epithelial and stromal regions
444 in histopathological images. *Neurocomputing* 2016;191:214–23.
- 445 [13] Farahani N, Parwani A V, Pantanowitz L. Whole slide imaging in pathology:
446 advantages, limitations, and emerging perspectives. *Pathol Lab Med Int*
447 2015;7:23–33.
- 448 [14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for
449 biomedical image segmentation. *Lect Notes Comput Sci (Including Subser*
450 *Lect Notes Artif Intell Lect Notes Bioinformatics)* 2015;9351:234–41.
- 451 [15] Bejnordi BE, Zuidhof GCA, Balkenhol M, Hermesen M, Bult P, van Ginneken
452 B, et al. Context-aware stacked convolutional neural networks for classification
453 of breast carcinomas in whole-slide histopathology images. *CoRR*
454 2017;abs/1705.0.
- 455 [16] Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based
456 convolutional neural network for whole slide tissue image classification. *Proc.*
457 *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, p. 2424–33.
- 458 [17] Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization
459 and molecular correlation of tumor-infiltrating lymphocytes using deep
460 learning on pathology images. *Cell Rep* 2018;23:181.
- 461 [18] Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for
462 cancer functional proteomics data. *Nat Methods* 2013;10:1046.

- 463 [19] Akbani R, Ng PKS, Werner HMJ, Shahmoradgoli M, Zhang F, Ju Z, et al. A
464 pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun*
465 2014;5:3887.
- 466 [20] Cheng J, Mo X, Wang X, Parwani A, Feng Q, Huang K. Identification of
467 topological features in renal tumor microenvironment associated with patient
468 survival. *Bioinformatics* 2017;34:1024–30.
- 469 [21] Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative analysis
470 of histopathological images and genomic data predicts clear cell renal cell
471 carcinoma prognosis. *Cancer Res* 2017;77:e91--e100.
- 472 [22] Shao W, Cheng J, Sun L, Han Z, Feng Q, Zhang D, et al. Ordinal Multi-modal
473 Feature Selection for Survival Analysis of Early-Stage Renal Cancer: 21st
474 International Conference, Granada, Spain, September 16–20, 2018,
475 Proceedings, Part II, 2018, p. 648–56.
- 476 [23] Wang J, Vasaikar S V, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more
477 comprehensive, powerful, flexible and interactive gene set enrichment analysis
478 toolkit. *Nucleic Acids Res.*, 2017;45(W1):W130-W137.
- 479 [24] Du Y, Zhang R, Zargari A, Thai TC, Gunderson CC, Moxley KM, et al. A
480 performance comparison of low-and high-level features learned by deep
481 convolutional neural networks in epithelium and stroma classification. *Med.*
482 *Imaging* 2018 Digit. Pathol., vol. 10581, 2018, p. 1058116.
- 483 [25] Vu QD, Kwak JT. A Dense Multi-Path Decoder for Tissue Segmentation in
484 Histopathology Images. *Comput Methods Programs Biomed*
485 2019;173:119--129.
- 486 [26] Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et
487 al. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal
488 Features Associated with Survival. *Sci Transl Med* 2011;3:108ra113-108ra113.
- 489 [27] Wang C, Pécot T, Zynger DL, Machiraju R, Shapiro CL, Huang K. Identifying
490 survival associated morphological features of triple negative breast cancer

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

492 [28] Mackie TR, Jackson EF, Giger M. Opportunities and challenges to utilization
493 of quantitative imaging: Report of the AAPM practical big data workshop. *Med*
494 *Phys* 2018;45:e820–8.

495 [29] Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al.
496 Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000
497 Tumors from 33 Types of Cancer. *Cell* 2018;173:291–304.e6.

498 [30] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*
499 2011;144:646–74.

500 [31] Rocca A, Farolfi A, Bravaccini S, Schirone A, Amadori D. Palbociclib (PD
501 0332991): targeting the cell cycle machinery in breast cancer. *Expert Opin*
502 *Pharmacother* 2014;15:407–20.

503 [32] Murphy CG, Dickler MN. The role of CDK4/6 inhibition in breast cancer.
504 *Oncologist* 2015;20:483–90.

505 [33] Witkiewicz AK, Chung S, Brough R, Vail P, Franco J, Lord CJ, et al.
506 Targeting the Vulnerability of RB Tumor Suppressor Loss in Triple-Negative
507 Breast Cancer. *Cell Rep* 2018;22:1185–99.

508 [34] Bremnes RM, Dønnem T, Al-Saad S, Al-Shibli KI, Andersen S, Sirera R, et al.
509 The role of tumor stroma in cancer progression and prognosis: emphasis on
510 carcinoma-associated fibroblasts and non-small cell lung cancer. *J Thorac*
511 *Oncol* 2011;6 1:209–17.

512 [35] Ghesquière B, Wong BW, Kuchnio A, Carmeliet P. Metabolism of stromal and
513 immune cells in health and disease. *Nature* 2014;511:167–76.

514 [36] Wang C, Su H, Yang L, Huang K. Integrative analysis for lung
515 adenocarcinoma predicts morphological features associated with genetic
516 variations. *PACIFIC Symp. Biocomput.* 2017, 2017, p. 82–93.

517 **Figure legends**

518 **Figure 1 Workflow for image processing and biological analysis**

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

522 **Figure 2 Segmentation results on TMA**

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

528 **Figure 3 Segmentation results on TCGA WSIs**

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

534 **Figure 4 Tissue distribution on different breast cancer subtypes**

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

537 **Figure 5 Results of GO enrichment analysis**

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

542 **Tables**

543 **Table 1 Demographic and clinical characteristics**

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

546 **Table 2 Evaluation of CNN model on NKI and VGH**

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

549 * TPR (True positive rate) = $TP / (TP + FN)$; TNR (True negative rate) = $TN / (FP +$
550 $TN)$; PPV (Positive Predictive Value) = $TP / (TP + FP)$; NPV (Negative Predictive
551 Value) = $TN / (FN + TN)$; FPR (False positive rate) = $FP / (FP + TN)$; FDR (False
552 Discovery Rate) = $1 - TP / (TP + FP)$; FNR(False Negative Rate) = $FN / (FN + TP)$;
553 ACC (Accuracy) = $(TP + TN) / (TP + FP + TN + FN)$; F1_score = $2 * TP / (2 * TP + FP$
554 $+ FN)$; MCC (Matthews Correlation Coefficient) = $(TP * TN - FP * FN) /$
555 $\sqrt{TP + FP} * (TP + FN) * (TN + FP) * (TN + FN)}$. TP, FP, TN and FN represent
556 the true positive, false positive, true negative and false negative, respectively.

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

558 **Supplementary material**

559 **Figure S1 Qualitative segmentation results on TCGA dataset**

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

564 **Table T1 Quantitative evaluation on TCGA dataset**

565 * TPR (True positive rate) = $TP / (TP + FN)$; TNR (True negative rate) = $TN / (FP +$
566 $TN)$; FPR (False positive rate) = $FP / (FP + TN)$; FNR(False Negative Rate) = $FN /$
567 $(FN + TP)$; ACC (Accuracy) = $(TP + TN) / (TP + FP + TN + FN)$; F1_score = $2 * TP /$
568 $(2 * TP + FP + FN)$. TP, FP, TN and FN represent the true positive, false positive, true
569 negative and false negative, respectively.

570 **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	

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

573 **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

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

576 *TPR (True positive rate) = $TP / (TP + FN)$; TNR (True negative rate) = $TN / (FP +$
577 $TN)$; PPV (Positive Predictive Value) = $TP / (TP + FP)$; NPV (Negative Predictive

578 Value) = $TN / (FN + TN)$; FPR (False positive rate) = $FP / (FP + TN)$; FDR (False

579 Discovery Rate) = $1 - TP / (TP + FP)$; FNR(False Negative Rate) = $FN / (FN + TP)$;

580 ACC (Accuracy) = $(TP + TN) / (TP + FP + TN + FN)$; F1_score = $2 * TP / (2 * TP + FP$
581 $+ FN)$; MCC (Matthews Correlation Coefficient) = $(TP * TN - FP * FN) /$
582 $\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$. TP, FP, TN and FN represent

583 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)

(b)

(c)

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

Fraction

GO_name