SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

Dataset of segmented nuclei in hematoxylin and eosin stained histopathology images of ten cancer types

Le Hou [2], Rajarsi Gupta[1], John S. Van Arnam[1], Yuwei Zhang[1], Kaustubh Sivalenka[2], Dimitris Samaras[2], Tahsin M. Kurc[1] & Joel H. Saltz[1] ✉

The distribution and appearance of nuclei are essential markers for the diagnosis and study of cancer. Despite the importance of nuclear morphology, there is a lack of large scale, accurate, publicly accessible nucleus segmentation data. To address this, we developed an analysis pipeline that segments nuclei in whole slide tissue images from multiple cancer types with a quality control process. We have generated nucleus segmentation results in 5,060 Whole Slide Tissue images from 10 cancer types in The Cancer Genome Atlas. One key component of our work is that we carried out a multi-level quality control process (WSI-level and image patch-level), to evaluate the quality of our segmentation results. The image patch-level quality control used manual segmentation ground truth data from 1,356 sampled image patches. The datasets we publish in this work consist of roughly 5 billion quality controlled nuclei from more than 5,060 TCGA WSIs from 10 different TCGA cancer types and 1,356 manually segmented TCGA image patches from the same 10 cancer types plus additional 4 cancer types.

Background & Summary

Digital pathology images are obtained via a series of processes: tissue slicing, staining, image capturing and digitization. The resolution of these images is usually at multi-gigapixel level. A single tissue slide typically contains around a million nuclei. The appearance, shape, texture, and morphological features of nuclei depend on the tissue type excised from an organ, cancer type, cell type, and many other factors. The comprehensive detection, segmentation, and classification of nuclei are core analysis steps in many histopathology image analysis tasks[1–16]. Segmentation of nuclei is the first step in extracting interpretable features that provide valuable diagnostic and prognostic cancer indicators[17–21], and thus is a crucial step for precision medicine[22,23]. The Cancer Genome Atlas (TCGA) program was a decade long, large scale National Cancer Institute led research effort that molecularly characterized over 20,000 primary cancer and matched control samples spanning 33 cancer types. Diagnostic whole slide images were captured for a large fraction of TCGA patients. Deidentified whole slide images, linked to molecular and clinical information are frequently accessed and analyzed publicly available information. TCGA whole slide Pathology images have been employed in many Cancer research efforts as well as in many digital Pathology methodology studies; Cooper et al.[24], for instance, describes examples of how TCGA whole slide images were used in integrative TCGA studies.

Current efforts to generate publicly accessible nuclear segmentation datasets in Hematoxylin and Eosin (H&E) stained whole slide images have been at much smaller scales than our work. Kumar et al.[13] collected a dataset of nucleus segmentation in seven cancer disease sites. This dataset is used as the MICCAI 2018 MoNuSeg challenge[25] in which the training set contains 30 image patches containing around 22,000 nuclear boundary annotations. The MICCAI 2015 to MICCAI 2018 Segmentation of Nuclei challenge[26] training sets contain around 6,000 nuclear boundary annotations. The extended PanNuke dataset[27,28] (currently the largest available dataset) contains 205,343 semi-automatically segmented nuclei in 481 patches sampled from 19 tissue types. Other datasets[29–32] have similar or smaller numbers of segmented nuclei. For these existing datasets, training patches are

[1]Biomedical Informatics Department, HSC L3-045, Stony Brook Medicine, Stony Brook University, Stony Brook, NY, 11794, USA. [2]Computer Science Department, 203C New Computer Science Building, Stony Brook University, Stony Brook, NY, 11794, USA. ✉e-mail: joel.saltz@stonybrookmedicine.edu

Abbre.	Cancer type	#. slides in total	#. slides failed QC
BLCA	Urothelial carcinoma of the bladder	380	14
BRCA	Invasive carcinoma of the breast	1,096	88
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	249	54
GBM	Glioblastoma Multiforme	772	40
LUAD	Lung adenocarcinoma	540	59
LUSC	Lung squamous cell carcinoma	431	35
PAAD	Pancreatic adenocarcinoma	190	11
PRAD	Prostate adenocarcinoma	387	19
SKCM	Skin Cutaneous Melanoma	470	64
UCEC	Endometrial Carcinoma of the Uterine Corpua	545	192
Total		5,060	576

Table 1. The main contribution of our work: nucleus segmentation data in 10 cancer types. We also generated results in 4 additional cancer types (COAD: colon adenocarcinoma, READ: rectal adenocarcinoma, STAD: stomach adenocarcinoma, UVM: Uveal Melanoma) that are not as good as the 10 cancer types. To validate the segmentation data, we collected segmentation ground truth in **1,356** patches. This set of manually segmented data is another contribution of our work.

usually stain-balanced, well digitized, and do not contain rare textures. However, in real world applications, the appearance of nuclei can be affected by a number of staining and imaging conditions: extremely high cellularity and nuclear pleomorphism, slightly out-of-focus, folding tissue, imbalanced H&E staining,. Existing experiments[33] showed that Convolutional Neural Networks (CNNs) generalize sub-optimally in unseen cancer types (cancer types that do not have training data). Therefore, training segmentation CNNs on existing datasets naively yields poor segmentation results in WSIs[33].

We aimed to accurately segment nuclei in WSIs of multiple cancer types. For this purpose, we leveraged a state-of-the-art nucleus segmentation Convolutional Neural Network (CNN) that our group recently reported[33]. Our approach has two advantages: (1). It generalizes well in cancer types that do not have training data: it improves the robustness of the segmentation network by synthesizing training data of every cancer type (2) The method is computationally efficient - this was critical given our goal of computing segmentation results for over 5,000 WSIs. Given our ability to produce large scale synthetic training data, a small U-net CNN[34] was able to generate accurate instance-level segmentation results in around **3 GPU hours per WSI**. Computationally expensive networks such as the Mask R-CNN[35] would achieve similar or worse across-cancer type generalization performance but in over 30 GPU hours per WSI. By combining three real training datasets[13,26] and a large scale synthetic dataset of 500,000 image patches, we train a U-net that has two output heads: one for nuclear center detection and one for nuclear material segmentation. We finally applied the watershed method[12,36] on detected centers and segmentation results, to output instance-level segmentation.

No existing automatic segmentation models give perfect results. Visually assessing segmentation results over 5,000 WSIs would take more than 200 human hours (more than 2.5 minutes per WSI) which is very time consuming. Instead, we apply the following methods **for quality control and data validation:**

Patch-level quantitative evaluation
We manually segmented nuclei in 1,356 patches and leveraged this to quantitatively evaluate our 5,000+ WSI segmentation dataset. In particular, we measured the segmentation overlap using Dice scores, and the instance-level segmentation/detection quality using Instance-Dice scores[26] and the nuclei count correlation scores.

Random segmentation region checking and WSI-level quality control
(1) We sampled 15 patches per WSI, and visually assessed and manually marked patches with what we considered to be adequate segmentation results (both precision and recall are at least 75%). (2) We identified WSIs that have unusual segmentation statistics (too few/many segmented nuclei), then visually assess segmentation data in them, and marked slides that have unacceptable segmentation (less than 80% of the slide both precision and recall are at least 75%). In these ways, we categorized WSIs into groups with different segmentation quality levels.

Using the patch-level manual segmentation data in 14 different TCGA cancer types, we quantitatively evaluated segmentation data. We judged 10 of the 14 cancer types to have nuclear segmentation result quality worthy of publication and data release. We thus release the following validated data as **our contributions**:

1. The automatic nucleus segmentation dataset contains 5,060 segmented slides in 10 TCGA cancer types, summarized in Table 1. This represents approximately 5 billion segmented objects. This large scale segmentation data for TCGA slides is very important, since characteristics of nuclei are essential for the diagnosis and study of cancer.

    (a) We apply per-WSI level quality control and categorize WSIs into groups with different segmentation quality levels. We identified 576 slides with suboptimal segmentation results. We filter out those WSIs for further analysis (although we still release the data for completeness).
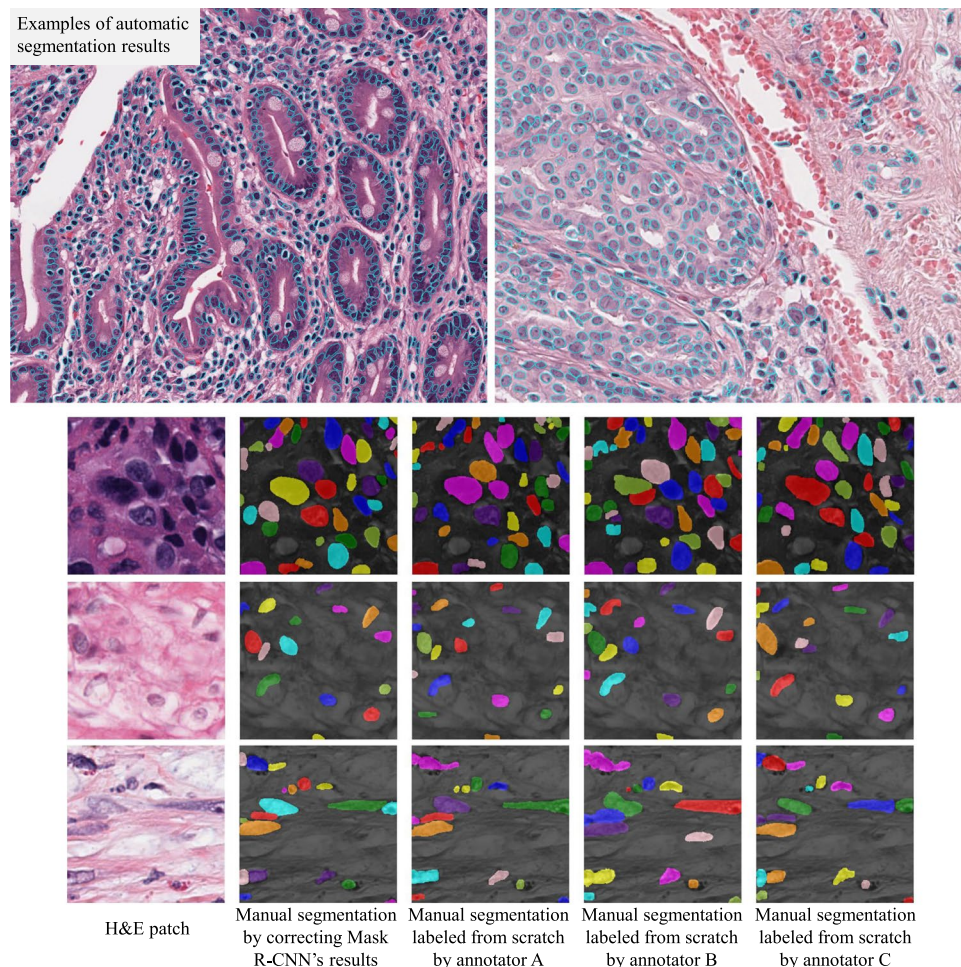
Examples of automatic segmentation results

| H&E patch | Manual segmentation by correcting Mask R-CNN's results | Manual segmentation labeled from scratch by annotator A | Manual segmentation labeled from scratch by annotator B | Manual segmentation labeled from scratch by annotator C |

Fig. 1 Samples of our data. (1) Automatic segmentation results on 5,060 WSIs (samples in top row), summarized in Table 1. (2) Manual segmentation data on over 1,356 patches (samples in bottom rows). Coloring of nuclear masks is for visualization only: it differentiates individual nuclei. We collect a large number of patches with labels for validating the segmentation results.

(b) Based on our patch-level quantitative assessment, compared to manual segmentation, in every cancer type, the nucleus segmentation data has an average Dice coefficient of least 77%, and an average instance level Dice coefficient[26] of at least 62%. These results are similar to the inter-annotator agreement in our experiments.

2. Manual segmentation labels on 1,356 patches of 256×256 pixels (64×64 μm^2) uniformly distributed in 14 cancer types. Two pathologists collaborated with three graduate students employed results from Mask R-CNN as a base to generate segmentation labels.

Examples of both datasets are shown in Fig. 1.

Methods

We first describe our published nucleus segmentation method in the first subsection "robust nucleus segmentation", then describe the new quality control and data validation approaches for this work through the rest of this paper.

Robust nucleus segmentation. To generate accurate segmentation results in multiple cancer types, existing state-of-the-art segmentation methods require extensive manually annotated training data in each cancer type. This is not scalable in practice. To address this problem, we use our existing robust nucleus segmentation model which was trained using not only manually annotated training data in several cancer types, but also heterogeneous synthetic training image patches, of every tissue type available in The Cancer Genome Atlas (TCGA). This data synthesis method is unsupervised, and is capable of generating millions of training patches which normally requires thousands of human hours to manually annotate – in this work, we used the data synthesis method to generate half a million patches. The workflow of this approach is shown in Fig. 2. We briefly describe our approach in this section.
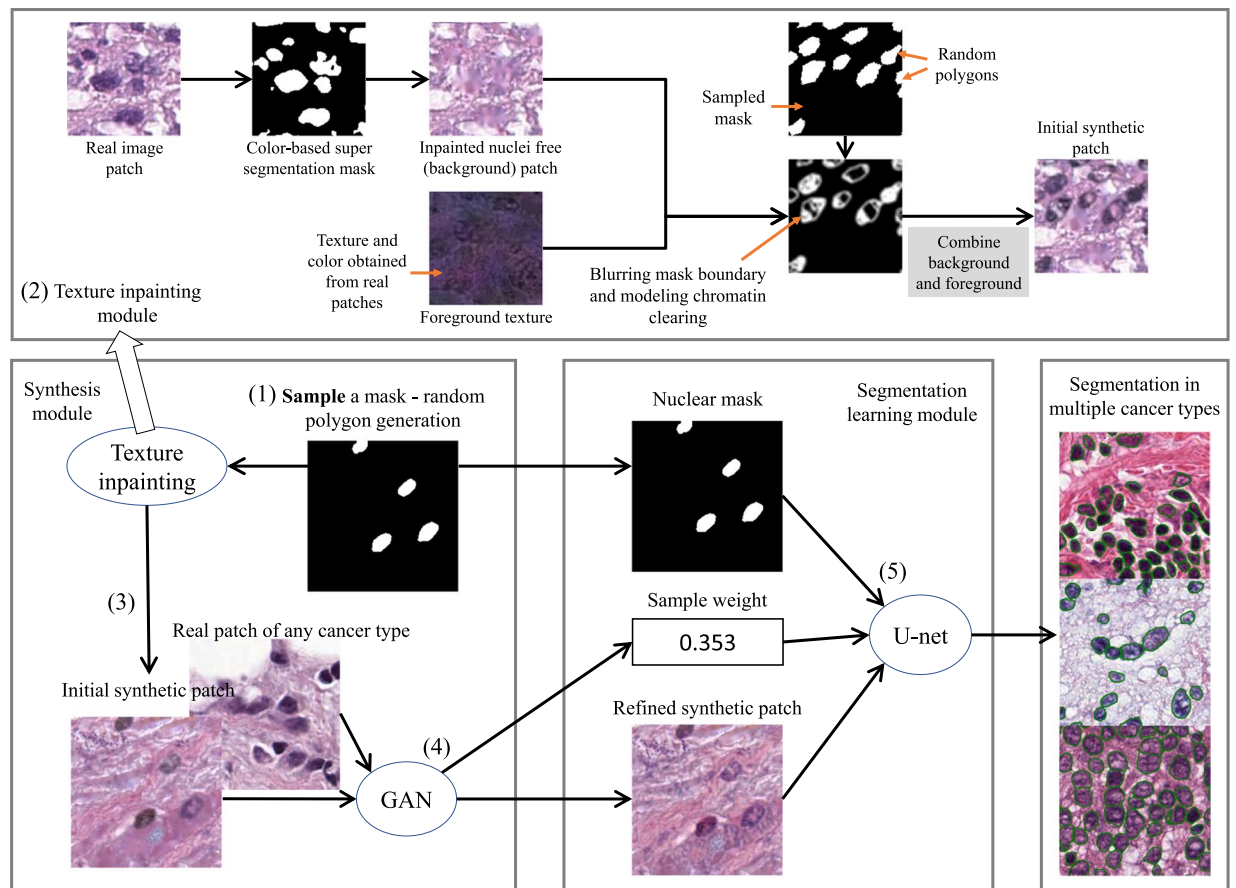
Fig. 2 Overview of our nucleus segmentation model training: we use a texture inpainting module to synthesize an initial synthetic pathology image patch with its nuclear mask. We then refine the initial synthetic patch using a GAN and compute its sample weight. We finally train a segmentation CNN on this sampled instance. Details are in our technical paper[33] and source code repository.

We first generate possibly realistic nuclear masks as random polygons. Then, we construct an initial synthetic patch utilizing textures and colors from real tissue (texture inpainting module in Fig. 2). We then refine the initial synthetic patch, to make it more realistic. Along this process, we compute a sample weight of this synthetic patch, indicating how realistic it is. Finally, we train a segmentation network using the initially generated nuclear masks, refined synthetic patch, and sample weight. In other words, we enumerate possible ground truth structures first and then check if a resulting synthetic patch is realistic or not. We decrease its impact in the training loss if it is not realistic. Similarly, if a resulting patch is not only very realistic, but also rarely synthesized, then we increase its impact in the training loss. Details are described in our technical paper[33].

In terms of the network architecture, the GAN's refiner has 21 convolutional layers and 2 pooling layers. The GAN's refiner discriminator has 15 convolutional layers and 3 pooling layers. As the segmentation CNNs, we use a U-net with 8 blocks: 4 down-sampling blocks and 4 up-sampling blocks. Each block has 3 to 6 convolutional layers and 1 pooling/deconv layer. We add a skip connection between blocks of the same resolution. In total there are 43 convolutional layers (including deconv). Each convolutional layer in the first and last block have 16 filters. After each pooling layer, we double the number of filters. We train the U-net on three real training data-sets[13,26] and our large scale synthetic dataset of 500,000 patches. The U-net has two output heads: one for nuclear center detection and one for nuclear material segmentation. We then apply the watershed method[12,36] on detected centers and segmentation results, to output instance-level segmentation. During test time, we normalize stains[37] in histopathology images before applying the U-net. We released our code on github.

Comparing to other state-of-the-art segmentation methods. Comparisons between our approach and other state-of-the-art level methods are detailed in our technical paper[33]. As a summary, on the MICCAI17 to MICCAI18[26], and Kumar[13] datasets, U-net trained with synthetic and real training data achieved state-of-the-art level results, even though other comparable baseline methods[9,15] use computationally more expensive models. For example, Mask R-CNN is 10 times more expensive compared to our U-net. In other words, we improve the performance of our segmentation method by adding synthetic training data, instead of increasing the neural network's capacity, which would make the task of segmenting 5,060 WSIs computationally very expensive.
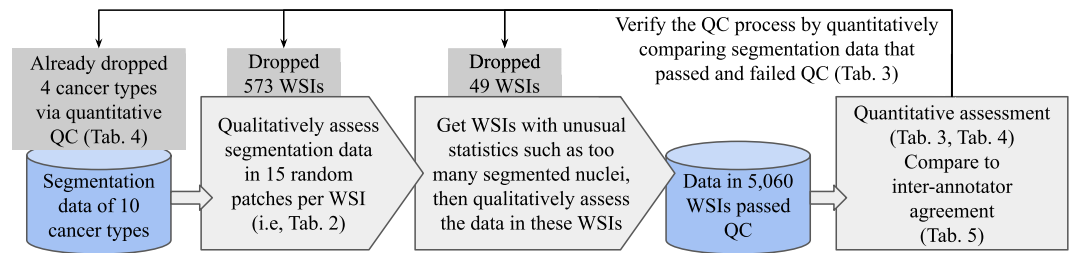
Fig. 3 Our quality control and data validation pipeline. This QC process is implemented to evaluate segmentation results at the WSI level. It would be infeasible to check the segmentation quality of all the nuclei individually.

WSI groups	Percentage of patches with bad segmentations	#. slides
Best	0%	2,346
Good	0.01–6.67%	1,246
Adequate	6.68–13.3%	593
Problematic	13.4–20.0%	302
Unacceptable	>20.0%	573
	or failed WSI QC	

Table 2. We categorize WSIs into groups with different segmentation quality levels. Slides identified as having unacceptable segmentation results are excluded from analysis in the rest of this work.

Quality control and data validation approaches overview. We apply a Quality Control (QC) and evaluation process as shown in Fig. 3. This QC process is implemented to evaluate segmentation results at the WSI level, as it would be infeasible to perform quality-control on all nuclei individually. We focus our efforts on whole slide images from 10 tumor types after our initial qualitative QC led us to eliminate four cancer types. After the application of the QC process, there are 5,060 WSIs with acceptable segmentation results. The number of segmented nuclei in these WSIs is roughly 5 billion in total.

WSI-level quality control. We visually assess segmentation quality per WSI, and categorize WSIs into groups with different segmentation quality levels. It is very time consuming to go through each WSI: visually checking segmentation results in one WSI takes approximately 2.5 minutes; and thus 5,000 WSIs would require over 200 hours. Therefore, we sample segmentation data in each WSI-level in two ways:

Random segmentation region checking for quality control and rating. We check segmentation quality in regions of all 5,060 WSIs at random locations. First, we randomly sample 15 patches (each has 256 by 256 pixels in 40X) per WSI and mix all patches from all WSIs. This results in approximately 76,000 patches. Then, we go through those patches and mark patches with reasonable segmentation results (both precision and recall are at least 75%). Finally, we categorize WSIs into four groups, according to the number of patches with bad segmentations, as shown in Table 2.

WSI-level qualitative assessment. The goal of this assessment is to identify and eliminate WSIs with unacceptable results. While this QC step involves a subjective method (i.e., visual inspection), it provides a complementary mechanism to the other QC steps (see Fig. 3). Unacceptable segmentation data identified in this way are still made available for download, but marked as "failed WSI-level visual QC".

To make sure that we identify most slides with unacceptable segmentation results, we select slides that have unusual segmentation statistics for visual assessment. We visually assess segmentation results in these slides and mark slides with unacceptable results efficiently for quality control. We define "unusual segmentation statistics" as the following:

1. Too many/few segmented nuclei. WSIs with either too many or too few segmented nuclei are subject to this WSI-level visual QC.
2. Average size of segmented nuclei is too large/small. WSIs with either very small or very large segmented nuclei are subject to this WSI-level visual QC.
3. Variation of the size of segmented nuclei is too large. WSIs with either very low or high nuclear pleomorphism are subject to this WSI-level QC.

In particular, we first compute the predicted nuclei count and average/variation of nuclear size, for each segmented slide. Then, slides that have one or more statistical values larger/smaller than -2% of the slides within the same cancer type are selected for visual assessment using the caMicroscope web tool[38]. For a WSI, we rate the
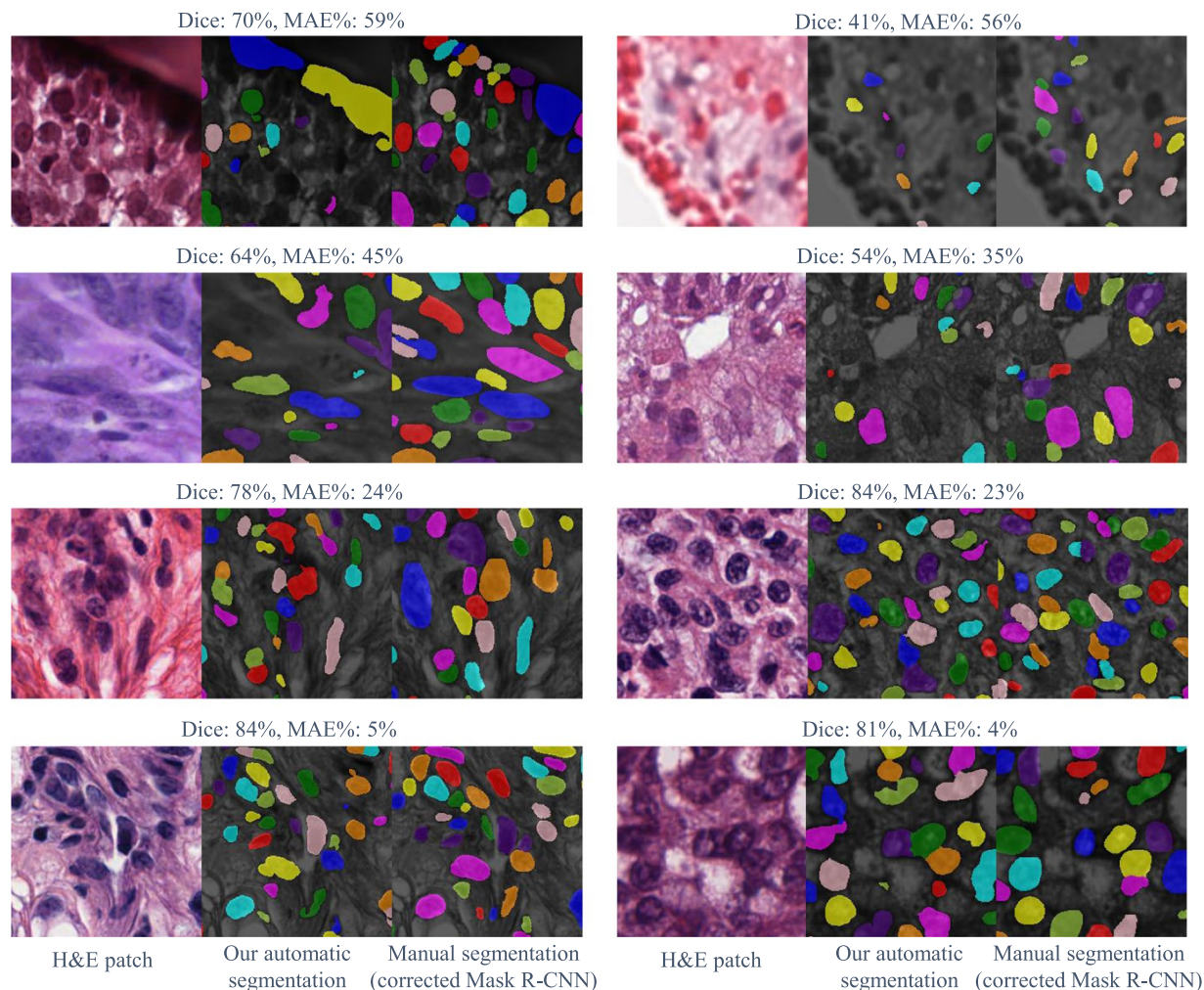
Fig. 4 Examples of automatic segmentation vs. manual segmentation. First two rows: failure cases. Last two rows: randomly selected samples.

segmentation result in the slide as either acceptable or unacceptable. Following the random segmentation region checking criterion, it is acceptable if and only if in at least 80% of the slide both precision and recall are at least 75%. We check whether the segmentation data is above the threshold by visual assessment. Around 500 WSIs in total are selected for visual assessment. For each cancer type, if a significant portion of the selected slides has unacceptable results, we select another 2% (in total 4%) of slides in each statistic value for visual assessment. In this way, **49** more slides were marked having unacceptable segmentations. Slides with results marked as unacceptable are excluded from analysis in the rest of this work.

We categorize WSIs into different levels of segmentation quality using random segmentation region checking and WSI-level visual assessment results, as summarized in Table 2.

Patch-level manual annotation data. To quantitatively evaluate and validate the automatic segmentation results in each WSI group, we collect segmentation ground truth in 1,356 patches, uniformly distributed in 14 cancer types. Examples of manual segmentations are shown in Fig. 1. All patches are 256 × 256 pixels in 40X (0.25 microns per pixel). Since this dataset is large and contains 14 cancer types, we argue that it is a contribution of our work as well. To collect this large scale ground truth data, three graduate students, supervised by two pathologists, manually corrected automatic segmentation results given by a Mask R-CNN (detailed later in this section). Our manual segmentation is imperfect. However, its accuracy is only rarely limited by atypical chromatin patterns or representation of the entire nucleus in the plane of section, and rarely encompasses more than a portion of the nuclear contour. The imperfection level of manual segmentation results fell roughly within the range of variability that one would expect when one compares data from different human annotators - the Dice scores of both cases are within the range of 0.75 to 0.80.

Using this patch level segmentation ground truth, we evaluate the quality of our automatic segmentation data in each cancer type. We found that our results in 10 out of the 14 cancer types are relatively accurate. We release our segmentation data in those 10 cancer types as our main contribution (Table 1).

WSI groups	#. patch labels	Dice	Instance-Dice	Nuclei count	
				Correlat.	MAE%
Best	446	0.797	0.687	0.947	15.2%
Good	242	0.789	0.660	0.930	16.1%
Adequate	128	0.774	0.636	0.915	17.6%
Problematic	52	0.788	0.625	0.879	20.5%
Unacceptable	103	0.690	0.545	0.718	33.8%
Excluding unacceptable	868	0.790	0.667	0.932	16.2%

Table 3. Quantitative assessment of the quality of nucleus segmentation, across 10 cancer types. The definition of WSI groups are given in Table 2. We exclude unacceptable segmentation results from analysis work in the rest of this paper.

Ground truth collection. We first extract patches of 256 pixels in 40X, randomly (unbiased) and uniformly distributed in 14 cancer types. We label extracted patches in two ways, described below.

Fast manual segmentation by correcting Mask R-CNN's segmentation results. In order to label thousands of patches, we minimize human labor by utilizing a Mask R-CNN - human annotators manually correct the Mask R-CNN's segmentation results in each patch, instead of labeling from scratch. Mask R-CNN[35] is a state-of-the-art level instance level segmentation network which although is not computationally efficient for segmenting over thousands of slides, gives reasonable segmentation results. Another advantage of using Mask R-CNN is that it has a different architecture compared to the U-net that we use to generate segmentation results. This architectural different eliminates possible biases for evaluation. In particular, we use the authors implementation and train a Mask R-CNN on the same real + synthetic dataset used for training the U-net. We then apply the trained Mask R-CNN on 1,356 patches. Three graduate students then correct the segmentation results by 1). Segmenting unsegmented nuclei; 2). Removing false segmentations; 3). Modifying incorrect segmentations. Manual segmentation results are reviewed by two pathologists and patches significantly mislabeled are then relabeled. This process is a form of crowdsourcing[39].

Manual segmentation from scratch. In order to evaluate the level of approximation in manual segmentation and the methodology of correcting Mask R-CNN's segmentation results, each of the three graduate students manually label a common set of 27 patches from scratch (not by correcting the Mask R-CNN's results). As a result, each patch has three manual segmentations, one from each student. Manual segmentation results are also reviewed by two pathologists and patches significantly mislabeled are then relabeled. Note that these patches were sampled from the same 1,356 patches described before.

Data Records
All data records are included in The Cancer Imaging Archive (TCIA)[40].

Automatic nucleus segmentation data. The algorithm-generated segmentation results. For each cancer type, you can find a cancertype_polygon folder, for example, BLCA_polygon. It contains polygon coordinates for each segmented nucleus (csv files), for all WSIs of BLCA. These results are obtained by thresholding the grayscale results in BLCA_prob folder and separating touching or overlapping nuclei by combining the detection and segmentation results. Each line in a csv file contains information of one nucleus. There are three columns in a csv file:

- **Area In Pixels** Size of the nucleus in terms of the number of pixels.
- **Physical Size** The number of pixels projected to 40X.
- **Polygon** The contour of the nucleus (polygon vertices in [x0:y0:x1:y1:..]).

In addition to cancertype_polygon folders, there are cancertype_meta folders which contain meta-data for each WSI. These folders are useless unless you use Microscope to visualize data.

Note: (1) In Box.com, the number of files under each folder shown in the "size" column is approximate; (2) Whether a slide has Unacceptable segmentation result or not is listed in the "list of histopathology slides" data described later. To further recognize WSIs with Best/good/Adequate/Problematic segmentations, one can use the "random segmentation region checking result" data described later.

List of histopathology slides. The list of 5,060 WSIs and summarized quality control results. This is a csv file with the following columns:

- **Cancer Type** Cancer type of the WSI.
- **WSI-ID** The case ID of the WSI, in TCGA naming convention.
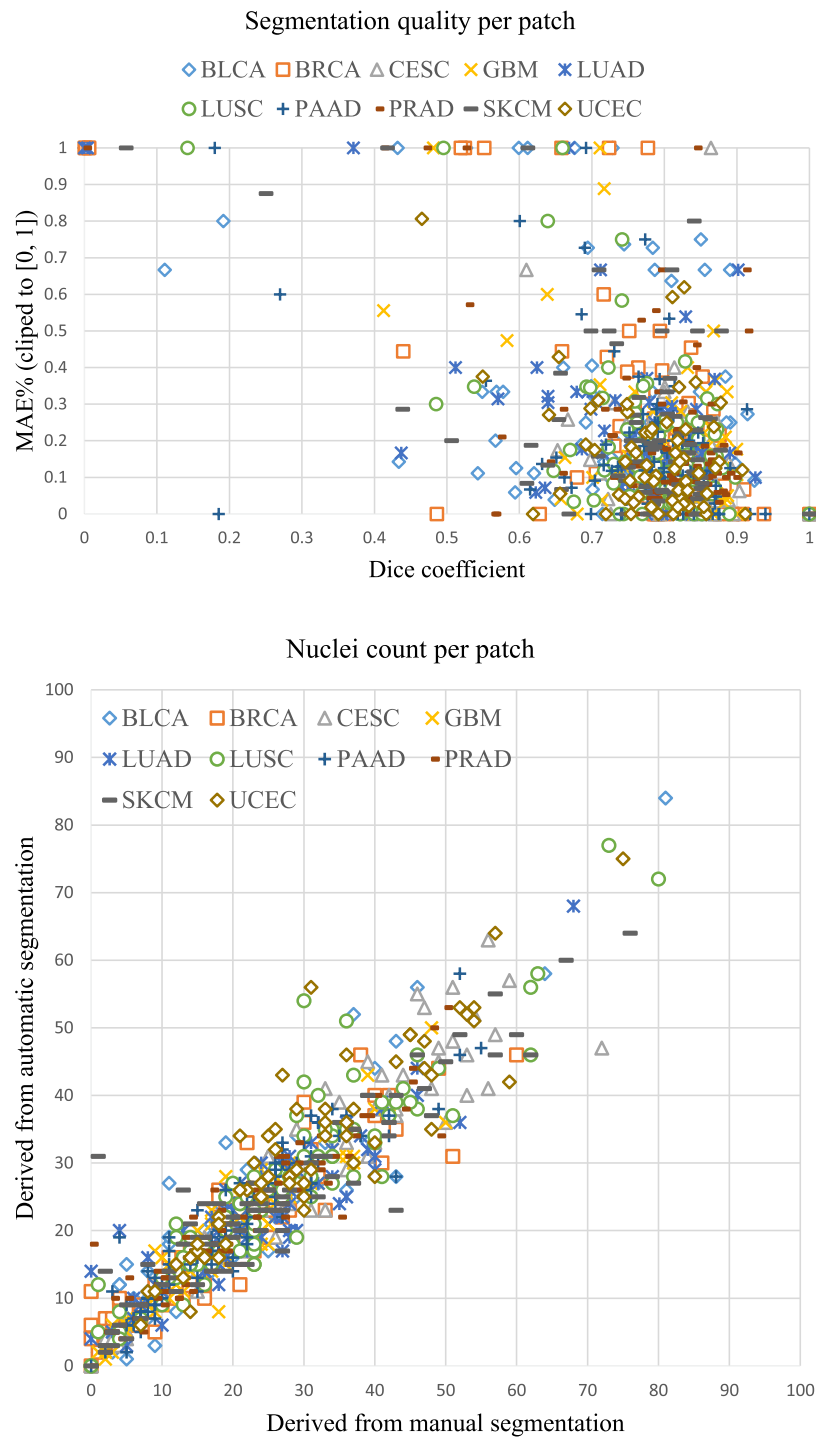- **QC Result** The summarized quality control result (passed or failed).

Segmentation quality per patch

Nuclei count per patch

Fig. 5 Top: Dice and MAE% results of all patches. Bottom: Predicted nuclei count (derived from automatic segmentation) vs. Ground truth nuclei count (derived from manual segmentation). Pearson correlation $= 0.932$, p-value $< 1.0 \times 10^{-308}$.

We do not redistribute the actual WSIs. These gigapixel histopathology slides can be downloaded from the publicly available The Cancer Genome Atlas (TCGA) repository[41]. For example, to download Urothelial carcinoma of the bladder (BLCA) slides, a user can:

1. Visit portal.gdc.cancer.gov/projects/TCGA-BLCA
2. Click on the "Files" link in the "Diagnostic Slide" row.
3. Click on the "Add All Files to Cart" bottom.
4. Go to your cart, and download all cart items.

Cancer Type	#. patch labels	Dice	Instance-Dice	Nuclei count	
				Correlat.	MAE%
BLCA	95	0.779	0.668	0.941	20.5%
BRCA	89	0.798	0.649	0.922	19.6%
CESC	79	0.818	0.677	0.947	13.4%
GBM	86	0.809	0.723	0.938	14.4%
LUAD	88	0.772	0.641	0.896	17.4%
LUSC	97	0.789	0.665	0.924	16.1%
PAAD	91	0.785	0.679	0.933	15.8%
PRAD	96	0.799	0.670	0.940	14.7%
SKCM	86	0.774	0.675	0.933	17.1%
UCEC	61	0.778	0.629	0.900	14.6%

Table 4. Quantitative assessment of the quality of nucleus segmentation, in each of the 10 cancer types. The p-value of Pearson correlation for every cancer type is smaller than 7.0×10^{-23}.

WSI quality control result. The list of slides selected for quality control by visual assessment and the detailed quality control result. This is a csv file with the following new columns (we do not list columns that are already explained before):

- **Num Nuclei Sample** The number of segmented nuclei in this WSI.
- **Size Of Nuclei-Average** The average size of nuclei.
- **Size Of Nuclei-Stddev** The standard deviation of the size of nuclei.
- **Note** The reason of selecting this WSI for visual assessment.
- **Segmentation Unacceptable Or Not** 0: acceptable; ? or 1: unacceptable.
- **Visual Assessment Comment** Verbal comments on this WSI.

Random segmentation region checking result. The detailed result of random segmentation region checking for each WSI. This is a csv file with the following new columns:

- **Num Of Unacceptable Seg Regions** The number of unacceptable regions.
- **Num Of Sampled Regions** The total number of visually assessed regions.

Manual segmentation data. The png images of manual segmentation data. Contains original H&E stained histopathology image patches, and instance-level segmentation masks. Additional information is in the readme.txt file of this data.

Technical Validation

We visually assess segmentation results in randomly sampled Whole Slide Images (WSIs) and also quantitatively analysis segmentation quality using patch-level segmentation labels.

WSI-level qualitative evaluation. Qualitative evaluation on all segmented WSIs is impractical. We randomly select 328 WSIs uniformly from 10 cancer types - at least 32 WSIs per cancer type to evaluate qualitatively. We use the same evaluation criterion used in the quality control process. Segmentation results in each slide is categorized as either acceptable or unacceptable. It is acceptable if and only if in at least 80% of the slide both precision and recall are at least 75%.

Out of the **328** randomly selected WSIs, **15** were marked as having unacceptable results. This concludes that our segmentation results on vast majority of WSIs are acceptable. We show examples of segmentation results in relatively large histopathology image tiles in Fig. 1.

Patch-level quantitative evaluation. We use manually annotated patches for quantitative evaluation. Note that we only use 971 patches in 10 cancer types, out of the 1,356 manually segmented patches in 14 cancer types. We only use manual segmentation in the center 226×226 pixels in each patch (as opposed to the entire 256×256 pixel patch), since segmentation close to the boundary is ambiguous due to incomplete data.

Evaluation metric. We use the Dice coefficient for measuring the quality of class-level (nuclear material or not) segmentation. Dice is ill-defined in patches that do not have any ground truth or predicted segmentation. To address this problem, the final Dice score is the average of per-patch Dice scores, weighted by the number of nuclei (ground truth nuclei count + predicted nuclei count) in each patch. To jointly measure the quality of segmentation and the quality of separating individual nuclei, we use the Instance-Dice score which is also used in the MICCAI nucleus segmentation challenge[26]. In addition, we compute the Pearson correlation and Mean Absolute Error Ratio (MAE%) between the number of nuclei segmented by U-net (defined as p), against the number of nuclei segmented by human annotators (defined as t). The MAE% is computed below:

9

Inter-annotator	Dice	Instance-Dice	Nuclei count	
			Correlat.	MAE%
Annotator A vs. B	0.760	0.600	0.959	10.8%
Annotator B vs. C	0.752	0.622	0.959	15.5%
Annotator C vs. A	0.774	0.697	0.954	12.2%

Table 5. Agreements between annotations from different human annotators. This is the performance upper bond of any automatic segmentation method.

Annotator	Dice	Instance-Dice	Nuclei count	
			Correlat.	MAE%
Annotator A	0.803	0.664	0.962	12.4%
Annotator B	0.793	0.631	0.984	11.2%
Annotator C	0.780	0.683	0.973	9.5%

Table 6. Comparing labeling from scratch vs. correcting Mask R-CNN's results.

$$\text{MAE\%} = \frac{|p - t|}{t}, \tag{1}$$

When we compute MAE% on a set of patches, we first compute the average of $|p - t|$ and t across all patches, then compute their ratio. We show examples of segmentation data with their evaluation results in Fig. 4.

Generated segmentation results vs. corrected Mask R-CNN's results. We compare the automatic segmentation results with the manual segmentations obtained from correcting Mask R-CNN's results. The overall accuracy of generated segmentation results is shown in Table 3. A scatter chart (Fig. 5) shows the accuracy of the predicted nuclei count. We also show per-cancer type evaluation results in Table 4.

Evaluating level of approximation in manual segmentation. We evaluate the level of approximation in manual segmentation by comparing each annotator's segmentation result with each other. We apply the evaluation metrics between each pair of students, shown in Table 5. One observation that in many cases, it is uncertain whether an object in histopathology images is a nucleus or not. This also contributes to the segmentation disagreement between human annotators.

Labeling from scratch vs. correcting Mask R-CNN's results. Finally, we evaluate how the labeling from scratch vs. correcting Mask R-CNN's results differ. For the 27 patches that were labeled from scratch, there are also the Mask R-CNN's corrected results. Evaluation results are in Table 6.

Usage Notes
We use CC0 (no copyright reserved) for our data.

Due to implementation and memory limitations, automatic nucleus segmentation results were generated and stored in 4,000 by 4,000 pixel tiles, as supposed to the entire WSI. Thus, nuclei across multiple tiles are split into different tiles. Additionally, we do not segment nuclei in tiles whose width or height is less than 2,000 pixels (this might happen on the edge of a WSI). All validation results include these by-design errors.

Code availability
Source code is available at on github. It contains the following repositories:

training-data-synthesis Code for generating synthetic training data for nucleus segmentation model training.
training-data-real-patch-extraction Code for converting the format of real training data.
segmentation-of-nuclei Code for training a nucleus segmentation model on patches generated by the above-mentioned repositories, and applying a trained model on WSIs.

Detailed descriptions are in the README files in the Github repository. We also provide a Dockerfile in Github, containing a trained model for easy deployment.

References
1. Gurcan, M. N., Tomaszewski, J. E. & Madabhushi, A. Digital pathology. *J. Med. Imaging* **21101**, 1 (2017).
2. Colen, R. *et al*. Nci workshop report: clinical and computational requirements for correlating imaging phenotypes with genomics signatures. *Transl. Oncol.* **7**, 556–569 (2014).
3. Xie, Y., Xing, F., Kong, X., Su, H. & Yang, L. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *MICCAI*, 358–365 (2015).
4. Cooper, L. A. *et al*. Digital pathology: Data-intensive frontier in medical imaging. *P. IEEE* **100**, 991–1003 (2012).

5. Saltz, J. *et al*. Towards generation, management, and exploration of combined radiomics and pathomics datasets for cancer research. *AMIA Jt. Summits Transl. Sci. Proc.* **2017**, 85 (2017).
6. Bayramoglu, N. & Heikkilä, J. Transfer learning for cell nuclei classification in histopathology images. In *ECCV Workshops* (2016).
7. Xu, J. *et al*. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *Medical Imaging* (2016).
8. Wang, S., Yao, J., Xu, Z. & Huang, J. Subtype cell detection with an accelerated deep convolution neural network. In *MICCAI* (2016).
9. Chen, H. *et al*. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Med. Image Anal.* **36**, 135–146 (2017).
10. Zhang, Y. *et al*. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI* (2017).
11. Yang, L., Zhang, Y., Chen, J., Zhang, S. & Chen, D. Z. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *MICCAI* (2017).
12. Bai, M. & Urtasun, R. Deep watershed transform for instance segmentation. In *CVPR* (2017).
13. Kumar, N. *et al*. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imaging* **36**, 1550–1560 (2017).
14. Murthy, V., Hou, L., Samaras, D., Kurc, T. M. & Saltz, J. H. Center-focusing multi-task CNN with injected features for classification of glioma nuclear images. In *WACV* (2017).
15. Hou, L. *et al*. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern Recognit.* **86**, 188–200 (2019).
16. Naylor, P., Laé, M., Reyal, F. & Walter, T. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Trans. Med. Imaging* **38**, 448–459 (2018).
17. Cooper, L. A. *et al*. An integrative approach for in silico glioma research. *IEEE Trans. Biomed.* **57**, 2617–2621 (2010).
18. Cooper, L. A. *et al*. Integrated morphologic analysis for the identification and characterization of disease subtypes. *JAMIA* **19**, 317–323 (2012).
19. Parmar, C. *et al*. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci. Reports* **5**, 11044 (2015).
20. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
21. Aerts, H. J. *et al*. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 1–9 (2014).
22. Council, N. R. *et al*. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease* (National Academies Press, 2011).
23. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* (2015).
24. Cooper, L. A. *et al*. Pancancer insights from the cancer genome atlas: the pathologist's perspective. *J. Pathol.* **244**, 512–524 (2018).
25. Kumar, N. *et al*. A multi-organ nuclei segmentation challenge. *IEEE Trans. Med. Imaging* (2019).
26. Vu, Q. D. *et al*. Methods for segmentation and classification of digital microscopy tissue images. *Front. Bioeng. Biotech.* **7** (2019).
27. Gamper, J. *et al*. Pannuke dataset extension, insights and baselines. Preprint at https://arxiv.org/abs/2003.10778 (2020).
28. Gamper, J., Koohbanani, N. A., Benet, K., Khuram, A. & Rajpoot, N. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *ECDP*, 11–19 (Springer, 2019).
29. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7** (2016).
30. Wienert, S. *et al*. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Sci. reports* **2**, 503 (2012).
31. Irshad, H. *et al*. Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. In *Pac. Symp. Biocomput.*, 294–305 (World Scientific, 2014).
32. Gelasca, E. D., Byun, J., Obara, B. & Manjunath, B. Evaluation and benchmark for biological image segmentation. In *ICIP* (2008).
33. Hou, L. *et al*. Robust histopathology image analysis: To label or to synthesize? In *CVPR* (2019).
34. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI* (Springer, 2015).
35. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. In *ICCV* (2017).
36. Beucher, S. Watershed, hierarchical segmentation and waterfall algorithm. In *Mathematical morphology and its applications to image processing*, 69–76 (Springer, 1994).
37. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph.* **21**, 34–41 (2001).
38. Saltz, J. *et al*. A containerized software system for generation, management, and exploration of features from whole slide tissue images. *Cancer Res.* **77**, e79–e82 (2017).
39. Amgad, M. *et al*. Structured crowdsourcing enables convolutional segmentation of histology images. *J. Bioinform.* **35**, 3461–3467 (2019).
40. Hou, L. *et al*. Dataset of segmented nuclei in hematoxylin and eosin stained histopathology images of 10 cancer types. *The Cancer Imaging Archive* https://doi.org/10.7937/tcia.2019.4a4dkp9u (2019).
41. The TCGA team. *The Cancer Genome Atlas*, https://cancergenome.nih.gov/.

Acknowledgements

Author contributions

Conceptualization, J.H.S., L.H., T.M.K. and D.S.; Methodology, L.H., J.H.S., D.S. and T.M.K.; Investigation, L.H., J.H.S., R.G., T.M.K., Y.Z. and K.S.; Writing, L.H., J.H.S., R.G. and T.M.K.; Supervision, J.H.S., R.G., D.S., T.M.K. and L.H.; Visualization, L.H. and J.H.S.; Data Curation, L.H., R.G., K.S. and Y.Z.; Software, L.H. and T.K.; Formal Analysis; L.H., J.H.S. and R.G.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.H.S.