Version 2, December 3, 2021

# Evaluation of cell segmentation methods without reference segmentations

Haoran Chen and Robert F. Murphy
Computational Biology Department
School of Computer Science
Carnegie Mellon University

## Abstract

Cell segmentation is a cornerstone of many bioimage informatics studies. Inaccurate segmentation introduces computational error in downstream cellular analysis. Evaluating segmentation results is thus a necessary step for developing segmentation methods as well as for choosing the most appropriate method for a particular type of tissue or image. The evaluation process has typically involved comparison of segmentations to those generated by humans, which can be expensive and subject to unknown bias. We present here an approach that seeks to evaluate cell segmentation methods without relying upon comparison to results from humans. For this, we defined a number of segmentation quality metrics that can be applied to multichannel fluorescence images. We calculated these metrics for 11 previously-described segmentation methods applied to datasets from 4 multiplexed microscope modalities covering 5 tissues. Using principal component analysis to combine the metrics we defined an overall cell segmentation quality score and ranked the segmentation methods. All code and data are available in a Reproducible Research Archive.

## Introduction

Cell segmentation is the task of defining cell boundaries in images. It is a fundamental step for many image-based cellular studies, including analysis and modeling of subcellular patterns (Boland and Murphy, 2001), analysis of changes upon various perturbations (Carpenter et al,

2006), cell tracking for investigating cell migration and proliferation (Garay et al., 2013), and cell morphology analysis for discovering cells' physiological states (Rittscher, 2010). Inaccurate cell segmentation introduces potential systematic error in all these downstream analyses. Since different methods may perform differently for different imaging modalities or tissues, it is important to evaluate potential cell segmentation methods to choose the most suitable one for a specific application.

Existing segmentation methods can be divided into two categories, the geometry-based segmentation techniques and deep learning-based approaches. The former includes but is not limited to threshold-based segmentation (Shen et al., 2018), region-based segmentation (Panagiotakis & Argyros, 2018), watershed algorithm and its variants (Ji et al., 2015), active contour (Wu et al., 2015), Chan-Vese segmentation (Braiki et al., 2020; Fan et al., 2013), and Graph-cut based segmentation (Oyebode & Tapamo, 2016). In the deep learning category, the conventional deep convolutional neural network (CNN) was initially applied (Jung et al., 2019; Sadanandan et al., 2017). CNN models learn the feature mapping of an image and convert the feature map into a vector for pixel-wise classification (i.e. segmentation). Since its publication, the U-Net model and its variants have become widely used (Ronneberger et al., 2015; Al-Kofahi et al., 2018; Falk et al., 2019; Long, 2020).

The typical approach for evaluating cell segmentation methods is to compare the segmentation results with human-created segmentation masks. For example, early work (Bamford, 2003) evaluated geometry-based algorithms on 20,000 cellular images annotated by three independent nonexpert observers. The author argued that evaluating cell segmentation does not necessarily require expert annotation. Caicedo et al (2019) evaluated multiple deep learning strategies for cell/nuclei segmentation by calculating the accuracy and types of errors in contrast to the expert annotations. The authors created a prototype annotation tool to facilitate the annotation process. At the model development level, deep-learning-based segmentation methods design loss functions to evaluate the similarity between the current segmentation mask and the human-created mask iteratively (Al-Kofahi et al., 2018; Kromp et al., 2020), which makes the model mimic the logic of human segmentation. Unfortunately, comparison to human-created segmentations assumes that people are good at this task; however, human segmentation can suffer from extensive intra and

inter-observer variability (Vicar et al., 2019; Wiesmann et al., 2017). In addition, human segmentation can require extensive labor and cost. For example, the development of the latest DeepCell version (Bannon et al, 2020) required a very large number of human evaluations (and would have required far more if an active machine learning approach had not been used).

An alternative approach is creating simulated cell images (Wiesmann et al., 2017) to assess segmentation performances. The simulated segmentation is generated based on various information obtained from the real fluorescent including cell shape, cell texture, and cell location. Since the image is simulated, the correct segmentation is known. Even though this method is efficient and reproducible, it reaches the limits while encountering images with high cell shape and texture variability.

For this study, we sought to define an objective evaluation approach that does not require a reference segmentation. We were motivated in large part by the desire to optimize the pipeline used for analysis of multichannel tissue images as part of the HuBMAP project. We defined a series of metrics based upon assumptions about the desired characteristics of good cell segmentation methods. We then identified currently available cell segmentation methods that had pre-trained models and evaluated their performance on images from multiple multichannel imaging modalities. A quality score is generated by our pipeline for each image (Figure 1). Given the quality score, we ranked each segmentation method with respect to performance on a particular tissue and imaging modality. In addition, we directly compared segmentation methods to each other using another set of pairwise metrics.

# Methods

## Evaluation metrics not requiring reference segmentation

We defined 14 metrics to evaluate the performance of a single segmentation method without requiring a reference segmentation. These are of two types: metrics that assess the coverage of a segmentation mask on the image, and metrics that measure various types of uniformity at the pixel and cell levels on multiplexed images. Each metric is derived under an assumption based on general concepts from cell biology. They are summarized in Table 1.

**Coverage Metrics**

The general assumption of coverage metrics is that a tissue image should be mostly occupied by cells except for readily detectable areas corresponding to ducts, blood vessels, or tissue borders. Therefore, we assume that a good segmentation should generate a mask that covers most areas of the tissue in an image.

*Number of cells per 100 squared microns (NC)*

The principle is that cell density is a measure of the quality of both the image and the segmentation method. The pixel size in square microns varies in different image modalities. To calculate the density, pixel sizes in the X and Y dimensions are directly obtained from OME-TIFF metadata. The metric is defined as

$$\mathrm{NC} = \frac{\mathrm{n_c}}{\mathrm{s_p} * \mathrm{n_p}}$$

where $n_c$ is the number of cells in the cell mask, $s_p$ is the size of one pixel in squared microns, $n_p$ is the total number of pixels in the corresponding image.

Given the binary foreground-background separation (see **Foreground-background separation**), we designed the following two coverage metrics with the general assumption that an accurate segmentation mask will cover most tissue areas (i.e., the foreground) but few background areas.

*Fraction of image foreground occupied by cells (FFC).*

$$\mathrm{FCF} = \frac{\mathrm{a_{cf}}}{\mathrm{a_c}}$$

where $a_{cf}$ is the area of a cell mask in the foreground and $a_c$ is the area of the cell mask.

*Fraction of image background occupied by cells (FBC)*

$$\mathrm{FBC} = \frac{a_{cb}}{a_b}$$

where $a_{cb}$ is the area of a cell mask in the background and $a_f$ is the area of the background.

1 minus FBC is used as the metric, so that larger values indicate better segmentation.

After matching the whole-cell and nuclear masks (see **Mask processing**), we calculated the fraction of matched cells and nuclei, assuming that segmented cells and nuclei should have a one-to-one correspondence relationship.

*Fraction of matched cells and nuclei (FMCN)*

The fraction of matched cells and nuclei is calculated, based on the assumption that in well-segmented images segmented cells and nuclei should have a one-to-one correspondence.

$$\mathrm{FMCN} = \frac{n_m}{n_{cmi} + n_{nmi} + n_m}$$

where $n_m$ is the number of matched cells and nuclei, $n_{cmi}$ is the number of mismatched cells, $n_{nmi}$ is the number of mismatched nuclei.

*Fraction of cell mask in foreground (FCF)*

$$\mathrm{FCF} = \frac{a_{cf}}{a_c}$$

where $a_{cf}$ is the area of a cell mask in the foreground and $a_c$ is the area of the cell mask.

**Homogeneity Metrics**

To measure the homogeneities of pixel intensities and integrated cell intensities from a segmentation method, we developed another category called homogeneity metrics. At the pixel level, the principle is that the pixels outside of the cells but still within the image foreground should be similar in protein composition (i.e., should consist of extracellular matrix of similar composition). The coefficient of variation (CV) of the foreground pixels outside the cells for each channel was calculated and then the average CV across all channels was taken (Figure 2a).

*Average CV of foreground pixels outside the cells (ACVF).*

$$\mathrm{ACVF} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\sigma_{if}}{\mu_i}$$

where $n_h$ is the total number of image channels, $\sigma_{if}$ is the standard deviation of pixel intensities of the $i^{th}$ channel in the foreground outside the cells, $\mu_i$ is the mean of all pixel intensities in the $i^{th}$channel.

The reciprocal of ACVF+1 is used as the metric, so that higher values indicate better segmentation and its upper bound is 1.

We applied the Principal Component Analysis (PCA) on the matrix of all foreground outside-the-cell pixel intensities across all channels after z-score standardization on each channel was applied and the fraction of variance explained by the first principal component was calculated. A higher value of the fraction stands for a more conserved relationship across channels in the foreground outside the cell areas (Figure 2a).

*Fraction of variance accounted for first PC of foreground pixels outside the cells (FPCF).*

$$\mathrm{FPCF} = \frac{\lambda_{f1}}{tr(\Sigma_f)}$$

where $\lambda_{f1}$ is the variance explained by the first principle component across channels for the foreground pixels outside the cells, $\Sigma_f$ is the covariance matrices from PCA analysis on the foreground pixels outside the cells across all channels, $tr(\quad)$ is trace calculation on a matrix.

At the cell level, the principles are that (a) cell types should be roughly similar in composition across all channels, (b) cell types can be approximated by cell clusters, (c) the average CV across all clusters for a given number of clusters is a proxy for similarity in composition within cell types, and (d) averaging the average CV across all clusters across different numbers of the cluster is also a proxy for similarity in composition within cell types. Cell types were defined using KMeans clustering performed on the mean cell intensities across all channels after applying z-score standardization for each channel. Since the number of cell types in an arbitrary image is unknown, K, the number of clusters, was varied from 1 to 10. Note that the KMeans clustering was applied

on the cell mask to define the cell type, which generated cell type labels of each number of clusters used in the following metric calculation for each cellular component mask (Figure 2b).

For the clusters from each K value, the average CV for each cluster across channels was calculated followed by an average weighted by the cluster size. The final metric was calculated as the average of the weighted average coefficient of variation across all clusters over 1 to 10 clusters.

*Average of weighted average CV of cell type intensities over 1-10 number of cluster (ACVC)*

$$\mathrm{ACVC} = \frac{1}{K_{max} - K_{min} + 1} \sum_{K=K_{min}}^{K_{max}} \left( \frac{1}{n_c} \sum_{k=1}^{K} (CV)_k \cdot n_k \right)$$

$$(\mathrm{CV})_\mathrm{k} = \frac{1}{\mathrm{n_h}} \sum_{\mathrm{i=1}}^{\mathrm{n_h}} \frac{\sigma_{\mathrm{ki}}}{\mu_{\mathrm{ki}}}$$

where $K_{min}$ is the smallest number of clusters, $K_{max}$ is the largest number of clusters, $K$ is the current number of clusters, $n_c$is the total number of cells, $CV_k$ is the coefficient of variation of mean cell intensities in $k^{th}$ cluster, $\sigma_{ki}$ and $\mu_{ki}$ is the standard deviation and the mean of mean cell intensities of $k^{th}$ cluster in $i^{th}$ channel respectively, $n_k$ is the number of cells in $k^{th}$ cluster.

The reciprocal of ACVC+1 is used as the metric, so that higher values indicate better segmentation and its upper bound is 1.

A similar measure was derived using principal component analysis. The first two principles are the same as above but (c) the fraction of variance accounted for by the first principal component of the cells in each cluster is a proxy for similarity in composition of each cell type, and (d) averaging this fraction over different numbers of clusters is also a proxy for similarity of each cell type, We, therefore, applied PCA on the mean cell intensities across z-score standardized channels of each number of cluster and calculated the average fraction of variance accounted for by the first

principal component across all numbers of clusters. The output vector was averaged over different K values to get the final metric.

*Average of weighted average fraction of the first PC of cell type intensities over 1-10 number of cluster*

$$\mathrm{FPCC} = \frac{1}{K_{max} - K_{min} + 1} \sum_{K=K_{min}}^{K_{max}} \left( \frac{1}{n_c} \sum_{k=1}^{K} \frac{\lambda_{k1}}{tr(\Sigma_k)} \cdot n_k \right)$$

where $K_{min}$ is the smallest number of clusters, $K_{max}$ is the largest number of clusters, $K$ is the current number of clusters of choice, $n_c$ is the total number of cells, $\lambda_{k1}$ is the variance explained by the first principle component in $k^{th}$ cluster, $\Sigma_k$ the covariance matrices from PCA on mean cell intensities across all channels in $k^{th}$ cluster, $n_k$ is the number of cells in the current cluster.

We also incorporated the Silhouette score as a metric to indicate the similarity of composition across cell types.

*Average of Silhouette score of clustering over 2-10 number of cluster (AS)*

$$\mathrm{AS} = \frac{1}{K_{max} - K_{min} + 1} \sum_{K=K_{min}}^{K_{max}} \left( \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \right)$$

$$\mathrm{a}(i) = \frac{1}{|K_i| - 1} \sum_{j \in K_i, i \neq j} d(i,j)$$

$$\mathrm{b}(i) = \min_{l \neq i} \frac{1}{|K_l|} \sum_{j \in K_l} d(i,j)$$

where $K_{min}$ is the smallest number of clusters, $K_{max}$ is the largest number of clusters, $K$ is the current number of clusters of choice, $a(i)$ is the average distance between cell $i$ and all the other cells $j$ in the cluster $K_i$ to which cell $i$ belongs, $b(i)$ is the minimum average distance from cell $i$ to all cells $j$ in all clusters $K_l$ to which $i$ does not belong, $n_c$ is the total number of cells.

In addition, we assumed that properly segmented images should have similarly sized cells.

*Standard Deviation of cell size (CSSD)*

$$\text{CSSD} = \sqrt{\frac{\sum_i (s_i - \mu_s)}{n_c}}$$

where $s_i$ is the size of $i^{th}$ cell, $\mu_s$ is the mean size of all cells, $n_c$ is the total number of cells.

The reciprocal of ln(CSSD)+1 is used as the metric, so that higher values indicate better segmentation and its upper bound is 1. The natural logarithm is taken to close the large difference of CSSD between methods.

## Foreground-background separation

To assess the coverage of the segmentation mask, each image was separated into regions consisting primarily of foreground and background. To do this, mean thresholding was applied to the nuclear, cytoplasmic, and cell membrane channels respectively, followed by performing two rounds of morphological closing with disk kernel of radius 1 and 10 respectively on the three thresholded images combined. Each round consists of a foreground closing to bridge the small and large gaps between cells within the tissue and a background closing on the inverted image to reunite the overly scattered background. An area closing was subsequently applied and those areas that have less than 5000 pixels were considered to be the foreground. To correct the resulting rounded boundaries, a morphological geodesic active contour was applied on the inverted image with current background areas as seeds, followed by an area closing to remove the small dots in the background. The final output is a binary mask with 1 as foreground and 0 as background (Figure 3).

## Mask processing

For some segmentation methods, finding cell boundaries is done independently of finding nuclear boundaries. This may mean that the final segmentation masks include nuclei that do not have a corresponding cell boundary and vice versa. As discussed above, we assumed that a good segmentation method would minimize this. To calculate the metric, each cell was matched to any nuclei contained within in. All cells that did not have corresponding nuclei were counted as

mismatched, and vice versa. For cells that had multiple corresponding nuclei, the one with the smallest fraction of mismatched pixels was kept. All mismatched cells and nuclei were removed from the calculation of other metrics.

Segmentation methods may also generate misshaped nuclei that have pixels outside their corresponding cells; that is, nuclei that protrude through the cell membrane. Across all methods, the nuclei of 49% of segmented cells were misshaped. To solve this issue, we applied two posterior approaches. The first approach considered all misshaped nuclei and their corresponding cells to be mismatched even if they have a one-to-one relationship. Alternatively, we developed a "repair" pipeline that trimmed the mask of misshaped nuclei. This repair pipeline was run as a postprocessing step for those methods and the results evaluated as a distinct segmentation method.

To evaluate the segmentation performance using the different channel intensities, we expected that nuclear protein composition would be considerably different than the cytoplasm and the cell membrane. We therefore calculated our metrics using two masks: the (repaired) nuclear mask and a "Cell Not Including Nucleus" mask calculated by removing the nuclear mask from the cell mask.

## Evaluation metrics for the similarity between two segmentation methods

**Difference between the evaluation metrics of two masks**

We assumed that the values of evaluation metrics for a single segmentation should be close for two similar segmentation methods. We took the absolute difference of each pair of metrics between two masks as the first set of metrics to evaluate the similarity between the two methods.

**Pairwise metrics**

We designed and adopted metrics for comparing two segmentation methods. Some of the following pairwise metrics were summarized by (Taha and Hanbury 2015).

**Number of Matched cells (NMC)**

We hypothesized that a large number of segmented cells should be overlapped for similar segmentation methods. We defined the metrics as the number of cell pairs from two segmentations

that have an overlap fraction greater than fraction a of the larger cell in each pair. When one cell in a mask overlaps with multiple cells in another mask, only the one with the largest overlap is considered matched.

$$\mathrm{NMC}(S_1, S_2) = \sum_i \sum_j \mathrm{I}_{|x_i \cap y_j| > p \max(|x_i|, |y_j|)}$$

where $S_1$, $S_2$ are two segmentation masks, I is the indicator function, $x_i$ and $y_i$ is $i^{th}$ and $j^{th}$ cell in $S_1$ and $S_2$ respectively, overlap fraction threshold $p$ is a free parameter. We took $p = 0.5$.

**Threshold AUC (TAUC)**

We assumed that similar segmentation methods should have masks with an overall good overlap of cells for varying overlap fraction thresholds. We calculated the area under the curve of the percentage of overlapping cells vs overlap fraction threshold while the threshold ranged from 0 to 1. The percentage of overlapping cells is the number of cell pairs that have an overlap fraction greater than the threshold, normalized by the smaller cell numbers of two masks.

**Average Jaccard index (AJI)**

We assumed that the overall similarity between matched cells should be high for similar segmentation methods. An average Jaccard index(Jaccard 1912; Tanimoto 1958) over all pairs of matched cells is thus calculated as a metric.

$$AJI(S_1, S_2) = \frac{1}{n_m} \sum_i \frac{|x_i \cap y_i|}{|x_i \cup y_i|}$$

where $S_1$, $S_2$ are two segmentation masks, $x_i$ and $y_i$ is $i^{th}$ pair of matched cells, $x_i \in S_1$, $y_i \in S_2$, $n_m$ is the total number of matched pairs.

**Average Dice coefficient (ADC)**

With the same assumption as for the Jaccard index, we also calculated an average Dice coefficient (Sørensen 1948; Dice 1945) over all pairs of matched cells.

$$ADC(S_1, S_2) = \frac{1}{n_m} \sum_i \frac{2|x_i \cap y_i|}{|x_i| + |y_i|}$$

where $S_1$, $S_2$ are two segmentation masks, $x_i$ and $y_i$ is $i^{th}$ pair of matched cells, $x_i \in S_1$, $y_i \in S_2$, $n_m$ is the total number of matched pairs.

**Average Hausdorff distance (HD)**

We assumed that each pair of matched cells should have a close spatial distance for similar segmentation methods. We selected Hausdorff distance(Tyrrell Rockafellar and Wets 2009; Birsan and Tiba) as the metric to measure the spatial distance between cells.

$$AHD(S_1, S_2) = \frac{1}{n_m} \sum_i \max\left( \sup_{a \in x_i} \inf_{b \in y_i} d\,(a, b), \sup_{b \in y_i} \inf_{a \in x_i} d\,(a, b) \right)$$

where $S_1$, $S_2$ are two segmentation masks, $x_i$ and $y_i$ is $i^{th}$ pair of matched cells, $x_i \in S_1$, $y_i \in S_2$, $a$, $b$ are the pixels in $x_i$ and $y_i$ cells respectively, $d$ is the distance between two pixels, and $n_m$ is the total number of matched pairs.

**Bidirectional Consistency Error (BCE)**

We assumed two similar segmentation methods should be consistent with the regions of each pixel that belongs. Bidirectional Consistency Error (David Royal Martin 2002) measures the error between two segmentation methods in terms of the regions over all pixels.

$$BCE(S_1, S_2) = \frac{1}{n_p} \max\left( \sum_i E(S_1, S_2, p_i)\,, \sum_i E(S_2, S_1, p_i) \right)$$

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) - R(S_2, p_i)|}{R(S_1, p_i)}$$

where $S_1$, $S_2$ are two segmentation masks, $p_i$ is $i^{th}$ pixel in the image, $R(S_j, p_i)$ is the area of the region in segmentation $S_j$ that contains pixel $p_i$, $n_p$ is the total number of pixels in the image.

**Error Matrix**

Considering one mask as the reference and the other as the query, we have:

- True Positive (TP): Total number of overlapping pixels in matched cells
- False Positive (FP): Total number of pixels that are in query mask but not in reference mask
- True Negative (TN): Total number of all pixels outside both masks
- False Negative (FN): Total number of pixels that are in reference mask but not in query

mask

The following metrics using the error matrix to calculate are symmetric no matter which mask is considered as the reference.

**Cohen's Kappa (CK)**

Cohen's Kappa (McHugh 2012) measures the agreement between raters. In this study, each segmentation method is a rater to the image. We assumed that two segmentation methods should agree with each other in terms of Cohen's Kappa if they are similar.

$$CK(S_1, S_2) = \frac{f_a - f_c}{n_p - f_c}$$

$$f_a = TP + TN$$

$$f_c = \frac{(TN + FN)(TN + FP) + (FP + TP)(FN + TP)}{n_p}$$

where $S_1$, $S_2$ are two segmentation masks, $n_p$ is the total number of pixels in the image.

**Mutual information (MI)**

We assumed that one segmentation should provide most of the information for the other when they are similar. We applied mutual information (Shannon 1948) to measure the reduction in uncertainty of one segmentation to another.

$$MI(S_1, S_2) = H(S_1) + H(S_2) - H(S_1, S_2)$$

$$H(S) = -\sum_i p(S^i) \log p\left(S^i\right)$$

$$H(S_1, S_2) = -\sum_{ij} p(S_1^i, S_2^j) \log p\left(S_1^i, S_2^j\right)$$

$$p(S_1^1, S_2^1) = TP/n_p \quad p(S_1^1) = (TP + FN)/n_p$$

$$p(S_1^1, S_2^2) = FN/n_p \quad p(S_1^2) = (TN + FN)/n_p$$

$$p(S_2^1, S_1^2) = FP/n_p \quad p(S_2^1) = (TP + FP)/n_p$$

$$p(S_1^2, S_2^2) = TN/n_p \quad p(S_2^2) = (TN + FP)/n_p$$

where $S_1$, $S_2$ are two segmentation masks, $n_p$ is the total number of pixels in the image; superscript = 1 means segmented cell region in the mask, 2 means segmented background region in the mask.

**Variation of information (VOI)**

We assumed that similar segmentation methods should provide nearly amount of information. We applied the variation of information (Arabie and Boorman 1973; Unruh and Zurek 1989; Meilă 2003) to measure the amount of information lost (or gained) when changing from one segmentation to another.

$$VOI(S_1, S_2) = H(S_1) + H(S_2) - 2MI(S_1, S_2)$$

where $S_1$, $S_2$ are two segmentation masks, $MI$ is the mutual information metric.

## Modalities of multi-channel images

We obtained image data from the HuBMAP project (HuBMAP Consortium, 2019; https://commonfund.nih.gov/hubmap) for 4 multiplexed modalities. For each modality, we chose appropriate nuclear, cytoplasmic, and cell membrane channels to use for segmentation. The sources are summarized in Table 2.

**CODEX**

CO-Detection by indEXing (CODEX) is an imaging technique generating highly multiplexed images of fluorescently labeled antigens. The HuBMAP portal contains 26 CODEX datasets over 5 tissues (large intestine, small intestine, thymus, spleen, and lymph node) produced by two Tissue Mapping Centers (TMC) (Stanford University and the University of Florida). The datasets from the Stanford TMC all contain 47 channels. Each dataset consists of 4 tissue regions, which are individually split into a grid of tiles with ~1440 x 1920 size. The University of Florida TMC generated 11-channel CODEX datasets. Each dataset only has one region which is divided into a grid in the same manner. To ensure both tissue coverage and evaluation efficiency, we created evaluation datasets with a subset of the tiles chosen to be non-neighboring (to avoid evaluating cells in the overlap region twice) and tiles that were not on the edge of the grid (to prevent capturing large backgrounds on the edge of tissue). 393 tiles in total were selected.

The different antigens used by the two TMCs require us to select different channels as segmentation input. For Stanford TMC datasets, we picked Hoechst, Cytokeratin, and CD45

channels as nuclear, cytoplasmic, and cell membrane channels, respectively. For the University of Florida TMC datasets, we instead chose DAPI, CD107a, and E-CAD. To ensure consistency of evaluation results from CODEX data, only the 5 channels that are common between the datasets of the two TMCs were used to calculate channel homogeneity metrics.

**Cell DIVE**

Cell DIVE is another antibody-based multiplexing technique. We had access to 12 regions of Cell DIVE data. Each region contains 26 images with 19 channels. Cell DIVE datasets are much larger than those from CODEX and MIBI, consisting of ~10000 x ~15000 pixels. To boost the efficiency of the pipeline while ensuring the coverage of datasets, we took the first image of each region and split each image into a grid of tiles similar to those of CODEX datasets. Each tile has roughly 1000 x 1000 pixels which is efficiently applicable to all segmentation methods. In order to exclude tiles with few cells, we calculated three metrics on each channel of each tile and applied KMeans clustering to determine the cluster with dense cells (Supplementary Figure 1). We selected DAPI, Cytokeratin AE1, and PCAD as nuclear, cytoplasmic, and cell membrane input for segmentation.

**MIBI**

MIBI is short for Multiplexed Ion Beam Imaging. A Secondary-Ion Mass Spectrometer (SIMS) is applied to image antibodies tagged with monoisotopic metallic reporters. We obtained 10 29-channel MIBI datasets. Each channel is 1024 x 1024 pixels. HH3, PanKeratin, and E-cadherin were chosen as the nuclear, cytoplasmic, and cell membrane channels for segmentation.

**IMC**

As an expansion of Mass Cytometry, Imaging Mass Cytometry (IMC) uses laser ablation to generate plumes of particles that are carried to the mass cytometer by a stream of inert gas (Chang et al., 2017). Thirteen 2D 39-channel IMC datasets of spleen, thymus, and lymph node were available from the U of Florida TMC. In this modality, lr191, SMA, and HLA-ABC were chosen as input channels for segmentation. For images without lr191 channel, Histone channel was substituted as nuclear input.

## Segmentation Methods

We created a python wrapper for each method to adapt all methods to a common pipeline. The input channels required by each method are shown in Table 3. All methods generate a segmentation mask for cell boundaries in an indexed image format while some output nuclear boundaries as well.

**DeepCell**

DeepCell is designed for robustly segmenting mammalian cells (Bannon et al., 2020; Moen et al., 2019; Van Valen et al., 2016). A deep convolutional neural network trained with approximately 900 images of bacteria, mammalian, or nuclei on the TensorFlow platform is embedded in deepcell-tf package in Python. We tested two versions of DeepCell software with different trained deep learning models. For each model, a nuclear intensity channel is a mandatory input and we chose cytoplasm or cell membrane as the secondary input image to generate two different whole-cell segmentation masks (these were treated as different methods). We also evaluated two versions of DeepCell.

**Cellpose**

Cellpose is a generalist, U-Net-based algorithm that segments cells in various image modalities (Stringer et al., 2021). The key idea of Cellpose is to track the gradient flow as the labels to predict the cells. A Pytorch model is trained with cytoplasm and nuclear datasets. Since a Python package of Cellpose is provided, we directly applied two versions of Cellpose with different pre-trained models to segment the whole cells on 2D cytoplasmic and nuclear intensity images. Unlike DeepCell, which uses the same model to generate whole-cell and nuclear masks, Cellpose trained a separate U-Net for the segmentation of nuclear channels only. We evaluated two versions of CellPose.

**CellProfiler**

The latest CellProfiler 4.0 (Carpenter et al., 2006; Kamentsky et al., 2011) was evaluated. CellProfiler applies thresholding and propagation algorithms to segment the nuclei then the whole cells. In “IdentifyPrimaryObjects” module, we applied minimum cross entropy thresholding (MCET) on the nuclear intensity image. Subsequently, large objects are “declumped” into

individual primary objects of interest with the border having equal distances to the local intensity peaks. Peaks closer than a minimum allowed distance are suppressed. The minimum allowed distance and the minimum and maximum diameters of primary objects are set according to the dataset. Using the found primary objects, namely the nuclei as seeds, an improved propagate algorithm is implemented in "IdentifySecondaryObjects" to segment the whole cells. The boundaries of the propagation are defined by the MCET thresholding on the cell membrane intensity channel, while a constraint of maximum range expanded from the primary objects is fixed.

**CellX**

In CellX MATLAB package, a gradient-based Circle Hough transform is first applied to obtain the potential points inside the cells as seeds(Dimopoulos et al., 2014). Then the Bresenham's algorithm (Bresenham, 1965) is used to determine the cell boundaries with cross-relation (Orfanidis, 1996) with an estimated membrane pattern. A cell membrane mask is then determined after a graph-cut optimization and energy minimization. Trivial solutions and morphological outlier will be removed by the spatial constraints set by the program. We also used the empirical maximum and minimum seed radius limit, and the maximum cell length as input according to the dataset.

**Cellsegm**

Cellsegm is a MATLAB-based toolbox providing automated whole-cell segmentation (Dimopoulos et al., 2014; Hodneland et al., 2013). It consists of four steps: 1. The directional coherence enhancing diffusion smoothing method is applied to the cell membrane channel. 2. Edge enhancing diffusion, which is a PDE-based anisotropic filter, is applied on the nuclear. The Kappa parameter is set at 0.05. 3. Adaptive threshold algorithm is deployed to find markers in the nuclear channel as seeds for a watershed in the next step. 4. Marker-based watershed algorithm to get the final segmentation mask. The maximum and minimum diameter of cells is set at 50 and 10, respectively.

**The Allen Cell Structure Segmenter**

There are two branches of the Allen Cell Structure Segmenter (AICS): the classic version and the deep learning version(Chen et al., 2020). The classic version consists of six main steps of the wrapper of this segmentation method: 1. Intensity normalization: The full intensity range of the image will be normalized to a pre-defined scale. 2. Smoothing: A 2D Gaussian smoothing is applied. When data is noisy, the sigma value should be increased. 3. Spot filter: This is the core algorithm of this method. A Laplacian of Gaussian is calculated to detect spot-like localization patterns. Two parameters are required as input: scale_x is set based on the estimated radius of the cells. And cutoff_x is a threshold applied to the filtered image to get the binary result. Filtered intensity higher than this threshold will be classified as the cells. Otherwise, the background. 4. Watershed: A watershed algorithm is run to cut falsely merge spots. 5. Object smaller than a threshold will be removed. 6. The connected component algorithm is applied to get an indexed segmentation mask.

**Voronoi diagram**

To benchmark all segmentation methods' performance, we designed a baseline method based on the Voronoi diagram. An Otsu thresholding is applied on the nucleus channel as the first step to obtain the nuclear mask. Then a Voronoi diagram partitions the image from the center of each nucleus. The coordinates of each center are averaged by the coordinates of all the pixels of the corresponding nucleus.

## Degraded images

To evaluate the robustness of each method, each image was degraded in two ways. In the first, zero-mean Gaussian noise was added to each pixel. The standard deviation was set to various levels based on the typical channel intensity of a given modality: 500, 1000, and 1500 for CODEX and Cell DIVE images, and 5, 10, and 15 for MIBI and IMC data. For the second perturbation, we downsampled the images to 70%, 50%, and 30% of the original size (number of pixels). Note that the perturbation was only done to the images used for segmentation; the evaluation of the resulting masks was done using the original images.

### Availability

All code needed to reproduce the evaluation results, as well as scripts to download test images, are available at https://github.com/murphygroup/CellSegmentationEvaluator. The intermediate results produced by the code (which enable recreation of all figures) are available at https://drive.google.com/drive/folders/14tw4qrXWTt2eg64zOpYjE9cZ7OFo1i6b?usp=sharing.

## Results

### Generating masks and calculating metrics

We started by running all methods on the selected images from the four imaging modalities: CODEX, Cell DIVE, MIBI, and IMC. There were 637 multi-channel images in total. Each method generated whole-cell and nuclear masks, and an additional set of masks was created with our "repair" procedure to eliminate unmatched nuclear and cell masks and remove nuclear regions outside the corresponding cell mask (see **Mask processing** in the Methods).

We then ran the evaluation pipeline to calculate the evaluation metrics for each method for each image. This process yielded two matrices of 637 images x 11 methods x 14 metrics, one for the original method and one for the method with repair.

To examine how the individual metrics varied across methods, we averaged the evaluation matrix over all images (Figure 4). We observed that the 11 methods have heterogeneous performance on the different metrics with either non-repair and repair approaches. Methods after repair tend to have a significant increase in the average metric values.

### Evaluating method response to added noise or downsampling

To test the robustness of the methods, we created perturbed images with various amounts of added zero-mean Gaussian noise or various extents of downsampling (see Methods) and evaluated the quality of the resulting segmentations (the perturbations were done only to the images provided to the segmentation method, and not to the multichannel images used to calculate the metrics). This allowed us to assess method robustness: how sensitive the results for a given segmentation method

were to the quality of the image. It also provided a parallel check of how well our metrics were performing based on the assumption that a perturbed image should yield a worse segmentation.

## Segmentation quality score

One of our primary goals was to provide an overall segmentation quality score for each method on each image. To do this we created a Principal Component Analysis (PCA) model using the metrics for all methods for all images with and without perturbation. The metrics were z-scored before performing PCA. The top 2 principal components for each method averaged across all modalities and images are shown in Figure 5 with and without various amounts of perturbation. We observed that PC1 values for all methods tend to decrease as perturbation increases, confirming that the metrics perform as anticipated. The loadings of PC1 (Supplementary Figure 2a) are all positive showing that all metrics have a synergistic effect on PC1 values. On the other hand, PC2 loadings (Supplementary Figure 2b) show that it is primarily an indicator of the overall coverage of a mask. We therefore adopted the sum of PC1 and PC2 weighted by their explained variance as our final overall quality score. Ranking based on the quality score for all methods (with and without repair) averaged across all modalities and (unperturbed) images are shown in Figure 6. As expected, repair improves the overall metric for most methods. We observed that DeepCell methods have the highest overall performance, which is also robust to both random Gaussian noise and downsampling (Figure 5).

## Model selection for particular tissue and data modality

One of the major potential uses of our evaluation pipeline is the potential ability to select the most appropriate method for a specific imaging modality or tissue. Images from different modalities have a large variation in signal-to-noise and spatial resolution that may influence the performance of segmentation methods. We therefore separately evaluated images from different modalities (Supplementary Figures 3-6). While it performed the best across different modalities, the version of DeepCell that was optimal varied, and relative ranking of the methods was quite different for IMC images compared to other modalities. Segmentation method performance may also be expected to vary for different tissues due to potential differences in the shapes and spatial arrangement of cells. We therefore separately analyzed performance on five tissues: large intestine, small intestine, lymph node, spleen, and thymus (images for the latter three tissues are available

for both CODEX and IMC) (Supplementary Figures 7-11). We observed that while DeepCell models still performed the best among all methods (again, with the optimal version changing), the rankings and overall quality scores of methods on different tissues vary. Interestingly, strong, similar performance was observed for both DeepCell and Cellpose on large intestine.

### Evaluating similarity between segmentation methods

Besides scoring each segmentation method, we also directly compared the similarity of segmentations between each pair of methods. We used two approaches: calculating a normalized distance between the single method metric vectors of the two methods, and calculating a set of metrics from direct comparison of the segmentation masks from two methods (see Methods). We concatenated each set of metrics across all images from multiple data modalities for each pair of methods and applied PCA to obtain PC1 values as corresponding similarity score. These scores were normalized by subtracting the PC1 value of a method compared with itself. Figure 7 shows heatmaps of the similarities between the methods for the two sets of metrics (separately). More negative values indicate more difference between a pair of methods. Using the single method metrics (Figure 7A), the deep learning-based models (DeepCell and Cellpose) are all quite close to each other, and the results for DeepCell with the membrane and cytoplasm markers are also close. However, pairwise comparisons (Figure 7B) show larger differences in the results for DeepCell and Cellpose (including between the two versions of DeepCell), and somewhat larger differences between the results for membrane and cytoplasm markers.

## Discussion

Many cell segmentation methods have been described but the ways to evaluate them have been limited. We have described here our design of a set of reference-free metrics to comprehensively measure the performance of segmentation methods. The metrics were developed based on a series of assumptions about the desirable characteristics of cell segmentation. We also used these metrics to compare all segmentation methods (that have available pre-trained models) that we were able to identify. The results demonstrate the power of deep learning-based methods, with DeepCell and Cellpose performing the best and second-best overall.

We also created an open-source pipeline for calculating the segmentation metrics for a given set

of images and a given method. Our evaluation pipeline provides a platform for users to choose segmentation methods for an individual image, tissue, and/or imaging modalities. While prior algorithmic approaches may have claimed the highest accuracy against different manually annotated training datasets, our method directly benchmarks them under the same set of measures.

The use of three-dimensional multi-channel tissue imaging is gradually growing. Some of the segmentation methods we have evaluated are capable of both 2D and 3D segmentation (see Methods), whereas algorithms such as the deep learning version of the Allen cell structure segmenter (Chen et al., 2020) and nnUNet (Isensee et al., 2021) only focus on segmenting 3D cell images. We plan to extend our work to 3D segmentation.

In closing, we note that in the future our quality scores could be included in the loss functions for training cell segmentation models in order to further improve model performance.

This work was supported in part by a grant from the National Institutes of Health Common Fund, OT2 OD026682 and by a traineeship to HC under training grant T32 EB009403

## Acknowledgments

This work was supported in part by a grant from the National Institutes of Health Common Fund, OT2 OD026682 and by a traineeship to HC under training grant T32 EB009403.

## Literature Cited


Al-Kofahi, Y., Zaltsman, A., Graves, R., Marshall, W., & Rusu, M. (2018). A deep learning-based algorithm for 2-D cell segmentation in microscopy images. BMC Bioinformatics, 19(1), 365.

Bamford, P. Empirical comparison of cell segmentation algorithms using an annotated dataset. In Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429). https://doi.org/10.1109/icip.2003.1246871

Bannon, D., Moen, E., Schwartz, M., Borba, E., Kudo, T., Greenwald, N., Vijayakumar, V., Chang, B., Pao, E., Osterman, E., Graf, W., & Van Valen, D. (2020). DeepCell Kiosk: Scaling deep learning-enabled cellular image analysis with Kubernetes. In Cold Spring Harbor Laboratory (p. 505032). https://doi.org/10.1101/505032

Boland, M.V. and Murphy, R.F. (2001). A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells. *Bioinformatics 17:*1213-1223.

Braiki, M., Benzinou, A., Nasreddine, K., & Hymery, N. (2020). Automatic Human Dendritic Cells Segmentation Using K-Means Clustering and Chan-Vese Active Contour Model. Computer Methods and Programs in Biomedicine, 195, 105520.

Bresenham, J. E. (1965). Algorithm for computer control of a digital plotter. In IBM Systems Journal (Vol. 4, Issue 1, pp. 25–30). https://doi.org/10.1147/sj.41.0025

Caicedo, J. C., Roth, J., Goodman, A., Becker, T., Karhohs, K. W., Broisin, M., Molnar, C., McQuin, C., Singh, S., Theis, F. J., & Carpenter, A. E. (2019). Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images. Cytometry. Part A: The Journal of the International Society for Analytical Cytology, 95(9), 952–965.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., Golland, P., & Sabatini, D. M. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology, 7(10), R100.

Chang, Q., Ornatsky, O. I., Siddiqui, I., Loboda, A., Baranov, V. I., & Hedley, D. W. (2017). Imaging Mass Cytometry. Cytometry. Part A: The Journal of the International Society for Analytical Cytology, 91(2), 160–169.

Chen, J., Ding, L., Viana, M. P., Lee, H., Filip Sluezwski, M., Morris, B., Hendershott, M. C., Yang, R., Mueller, I. A., & Rafelski, S. M. (2020). The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. In Cold Spring Harbor Laboratory (p. 491035). https://doi.org/10.1101/491035

Czech, E., Aksoy, B. A., Aksoy, P., & Hammerbacher, J. (2019). Cytokit: a single-cell analysis toolkit for high dimensional fluorescent microscopy imaging. BMC Bioinformatics, 20(1), 448.

Dimopoulos, S., Mayer, C. E., Rudolf, F., & Stelling, J. (2014). Accurate cell segmentation in microscopy images using membrane patterns. Bioinformatics , 30(18), 2644–2651.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C., Walsh, S., Saltukoglu, D., Tay, T. L., Prinz, M., Palme, K., Simons, M., … Ronneberger, O. (2019). U-Net: deep learning for cell counting, detection, and morphometry. Nature Methods, 16(1), 67–70.

Fan, J., Li, S., & Zhang, C. (2013). Color Cell Image Segmentation Based on Chan-Vese Model for Vector-Valued Images. In Journal of Software Engineering and Applications (Vol. 06, Issue 10, pp. 554–558). https://doi.org/10.4236/jsea.2013.610066

Garay, T., Juhász, É., Molnár, E., Eisenbauer, M., Czirók, A., Dekan, B., László, V., Hoda, M. A., Döme, B., Tímár, J., Klepetko, W., Berger, W., & Hegedűs, B. (2013). Cell migration or cytokinesis and proliferation? – Revisiting the "go or grow" hypothesis in cancer cells in vitro. In Experimental Cell Research (Vol. 319, Issue 20, pp. 3094–3103). https://doi.org/10.1016/j.yexcr.2013.08.018

Hodneland, E., Kögel, T., Frei, D. M., Gerdes, H.-H., & Lundervold, A. (2013). CellSegm - a MATLAB toolbox for high-throughput 3D cell segmentation. Source Code for Biology and Medicine, 8(1), 16.

HuBMAP Consortium (2019) The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature 574*:187-192. doi: 10.1038/s41586-019-1629-x

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.

Ji, X., Li, Y., Cheng, J., Yu, Y., & Wang, M. (2015). Cell image segmentation based on an improved watershed algorithm. In 2015 8th International Congress on Image and Signal Processing (CISP). https://doi.org/10.1109/cisp.2015.7407919

Jung, H., Lodhi, B., & Kang, J. (2019). An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. BMC Biomedical Engineering, 1, 24.
Kamentsky, L., Jones, T. R., Fraser, A., Bray, M.-A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W., & Carpenter, A. E. (2011). Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics , 27(8), 1179–1180.
Kromp, F., Bozsaky, E., Rifatbegovic, F., Fischer, L., Ambros, M., Berneder, M., Weiss, T., Lazic, D., Dörr, W., Hanbury, A., Beiske, K., Ambros, P. F., Ambros, I. M., & Taschner-Mandl, S. (2020). An annotated fluorescence image dataset for training nuclear segmentation methods. Scientific Data, 7(1), 262.
Long, F. (2020). Microscopy cell nuclei segmentation with enhanced U-Net. BMC Bioinformatics, 21(1), 1–12.
McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafelski, S. M., Thirstrup, D., Wiegraebe, W., Singh, S., Becker, T., Caicedo, J. C., & Carpenter, A. E. (2018). CellProfiler 3.0: Next-generation image processing for biology. In PLOS Biology (Vol. 16, Issue 7, p. e2005970). https://doi.org/10.1371/journal.pbio.2005970
Moen, E., Borba, E., Miller, G., Schwartz, M., Bannon, D., Koe, N., Camplisson, I., Kyme, D., Pavelchek, C., Price, T., Kudo, T., Pao, E., Graf, W., & Van Valen, D. (2019). Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. In Cold Spring Harbor Laboratory (p. 803205). https://doi.org/10.1101/803205
Orfanidis, S. J. (1996). Introduction to Signal Processing: Solutions manual.
Oyebode, K. O., & Tapamo, J. R. (2016). ADAPTIVE PARAMETER SELECTION FOR GRAPH CUT-BASED SEGMENTATION ON CELL IMAGES. In Image Analysis & Stereology (Vol. 35, Issue 1, p. 29). https://doi.org/10.5566/ias.1333
Panagiotakis, C., & Argyros, A. A. (2018). Cell Segmentation Via Region-Based Ellipse Fitting. In 2018 25th IEEE International Conference on Image Processing (ICIP). https://doi.org/10.1109/icip.2018.8451852
Rittscher, J. (2010). Characterization of Biological Processes through Automated Image Analysis. In Annual Review of Biomedical Engineering (Vol. 12, Issue 1, pp. 315–344). https://doi.org/10.1146/annurev-bioeng-070909-105235

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Lecture Notes in Computer Science (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Sadanandan, S. K., Ranefall, P., Le Guyader, S., & Wählby, C. (2017). Automated Training of Deep Convolutional Neural Networks for Cell Segmentation. Scientific Reports, 7(1), 7860.

Shen, S. P., Tseng, H.-A., Hansen, K. R., Wu, R., Gritton, H. J., Si, J., & Han, X. (2018). Automatic Cell Segmentation by Adaptive Thresholding (ACSAT) for Large-Scale Calcium Imaging Datasets. eNeuro, 5(5). https://doi.org/10.1523/ENEURO.0056-18.2018

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nature Methods, 18(1), 100–106.

Van Valen, D. A., Kudo, T., Lane, K. M., Macklin, D. N., Quach, N. T., DeFelice, M. M., Maayan, I., Tanouchi, Y., Ashley, E. A., & Covert, M. W. (2016). Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. PLoS Computational Biology, 12(11), e1005177.

Vicar, T., Balvan, J., Jaros, J., Jug, F., Kolar, R., Masarik, M., & Gumulec, J. (2019). Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. BMC Bioinformatics, 20(1), 360.

Wiesmann, V., Bergler, M., Palmisano, R., Prinzen, M., Franz, D., & Wittenberg, T. (2017). Using simulated fluorescence cell micrographs for the evaluation of cell image segmentation algorithms. BMC Bioinformatics, 18(1), 176.

Wu, P., Yi, J., Zhao, G., Huang, Z., Qiu, B., & Gao, D. (2015). Active contour-based cell segmentation during freezing and its application in cryopreservation. IEEE Transactions on Bio-Medical Engineering, 62(1), 284–295.

Table 1. Summary of segmentation metrics

| Name | Metric | Mask |
|---|---|---|
| Number of cells per 100 squared microns | NC | Matched cell mask |
| Fraction of image foreground occupied by cells | FFC | Matched cell mask |
| Fraction of image background occupied by cells | 1-FBC | Matched cell mask |
| Fraction of cell mask in foreground | FCF | Matched cell mask |
| Fraction of match between cells and nuclei | FMCN | Original Cell and nuclear mask |
| Average CV of foreground pixels outside the cells | 1/(ACVF+1) | Matched cell mask |
| Fraction of first PC of foreground pixels outside the cells | FPCC | Matched cell mask |
| Average of weighted average CV of cell type intensities over 1-10 number of cluster | 1/(ACVC+1) | Cellular components mask |
| Average of weighted average fraction of the first PC of cell type intensities over 1-10 number of cluster | FPCC | Cellular components mask |
| Average of Silhouette score of clustering over 2-10 number of cluster | AS | Cellular components mask |
| Standard deviation of cell size | 1/(ln(CSSD)+1) | Matched cell mask |

Table 2 Nuclear, Cytoplasmic, and Cell Membrane Channels for Each Modality

| | Nuclei | Cytoplasm | Cell Membrane |
|---|---|---|---|
| CODEX-Stanford | Hoechst1 | Cytokeratin | CD45 |
| CODEX-Univ. of Florida | DAPI | CD107a | E-CAD |
| MIBI | HH3 | PanKeratin | E-cadherin |
| Cell DIVE | DAPI | Cytokeratin AE1 | PCAD |
| IMC | lr191/Histone | SMA | HLA-ABC |

Table 3 Segmentation methods evaluated

| Method | Inputs | | | | | Output |
|---|---|---|---|---|---|---|
| | Cytoplasm | Cell Membrane | Nucleus | Dimension | Requires scale | Nuclear mask |
| DeepCell 0.6.0 | | X | X | 2D | No | Yes |
| DeepCell 0.6.0 | X | | X | 2D | No | Yes |
| DeepCell 0.9.0 (new) | | X | X | 2D | No | Yes |
| DeepCell 0.9.0 (new) | X | | X | 2D | No | Yes |
| Cellpose 0.0.3.1 | X | | X | 2D/3D | No | Yes |
| Cellpose 0.6.1 (new) | X | | X | 2D/3D | No | Yes |
| CellSegm | | X | | 2D/3D | Yes | No |
| CellX | | X | | 2D | Yes | No |
| CellProfiler | | X | X | 2D | Yes | No |
| AICS (classic) | | X | X | 2D/3D | Yes | No |
| Voronoi | | | X | 2D | No | Yes |

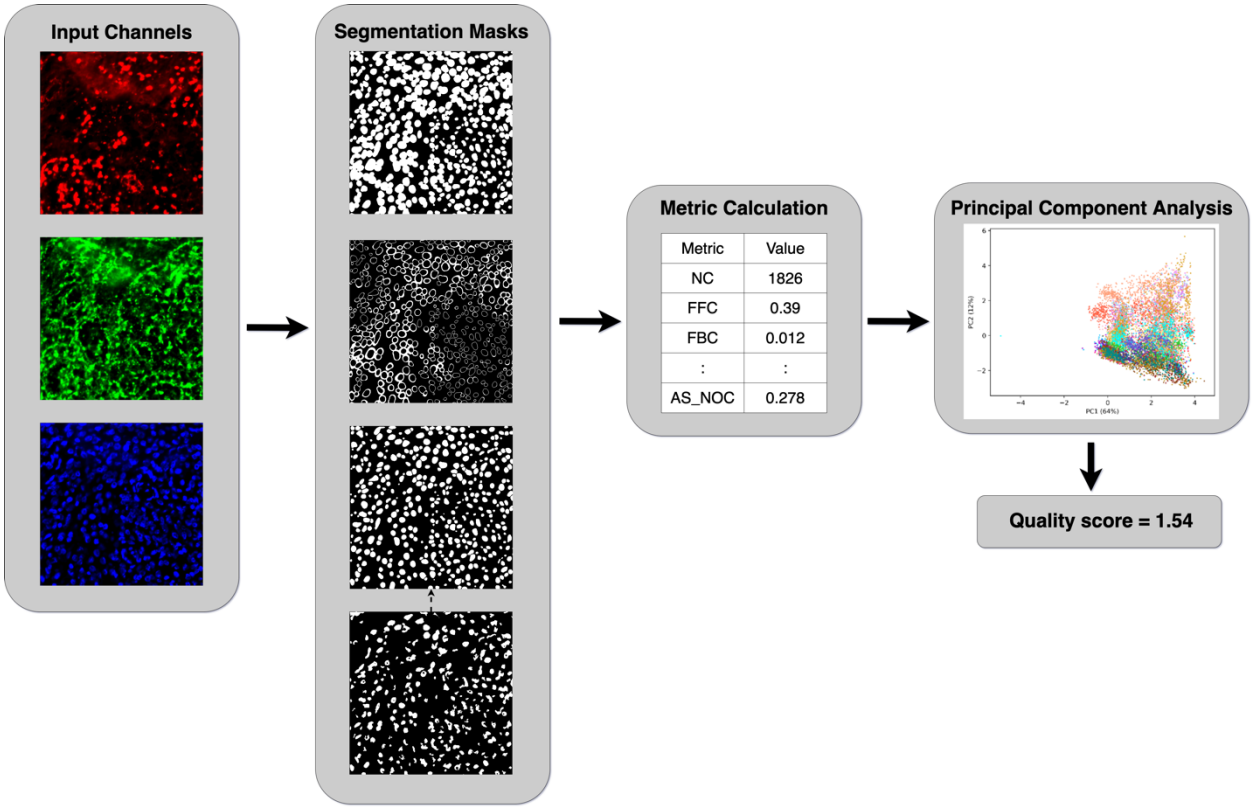


Figure 1. A pipeline for cell segmentation evaluation. Input Channels: cell membrane (red), cytoplasm (green), and nucleus (blue). Each segmentation method took one or two channels as input to generate segmentation masks. Segmentation Masks: cell mask, cell outside nucleus mask, and nuclear mask (top-down). For segmentation methods that don't output the nuclear mask, we used an Otsu-thresholded nuclear mask (bottom mask) as a substitute. We removed the pixels in the nuclear mask that were outside the cell membrane in the cell mask, as well as the cells and nuclei that didn't have the corresponding nuclei and cells. Metric Calculation: for each set of segmentation masks, we calculated 14 metrics to objectively evaluate the goodness of segmentation under various assumptions. Principal Component Analysis: We applied the Principal Component Analysis on the result matrix of all methods segmented all images from 4 modalities. Each point represents a cell segmentation generated by a method. Different colors represent different segmentation methods. The percentages in the parentheses are the variances accounted for either PC. Quality score: A variance-weighted sum of PC1 and PC2 for a point on the PCA plot result in a quality score for the performance of a segmentation method on an image.

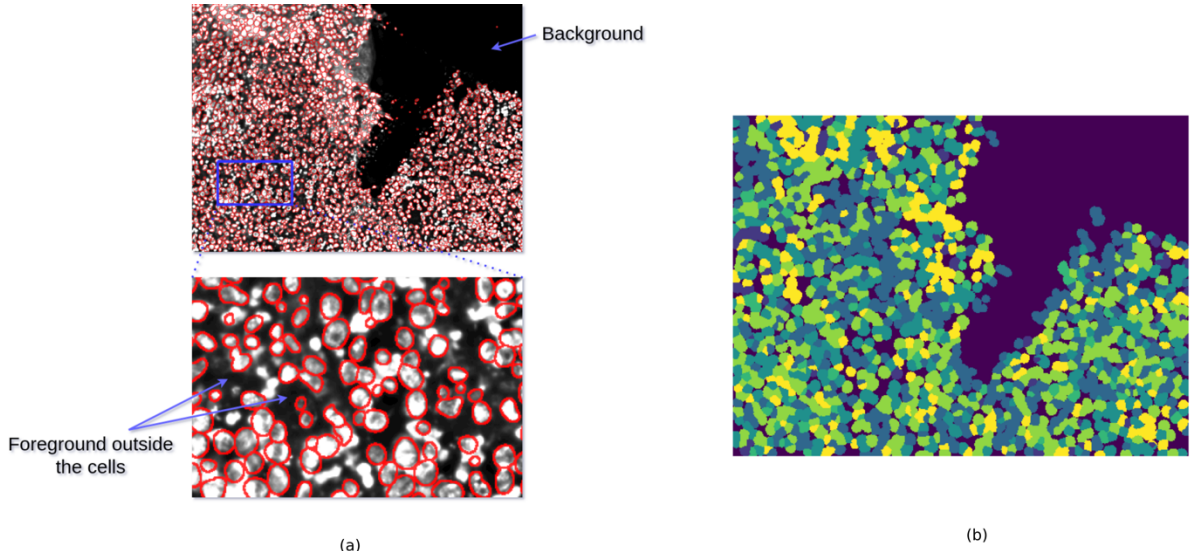


Figure 2. Illustration of evaluation metrics. (a). This image comes from the sum of intensities of nucleus and cell membrane channels of tile R001_X001_Y001 of CODEX image HBM836.WNJS.587. The red contours are from the segmentation result of DeepCell 0.9.0 with cell membrane input on this image. (b). Clustered by the mean intensities per cell. We calculated the cell uniformity metrics within each cluster and averaged them across clusters.

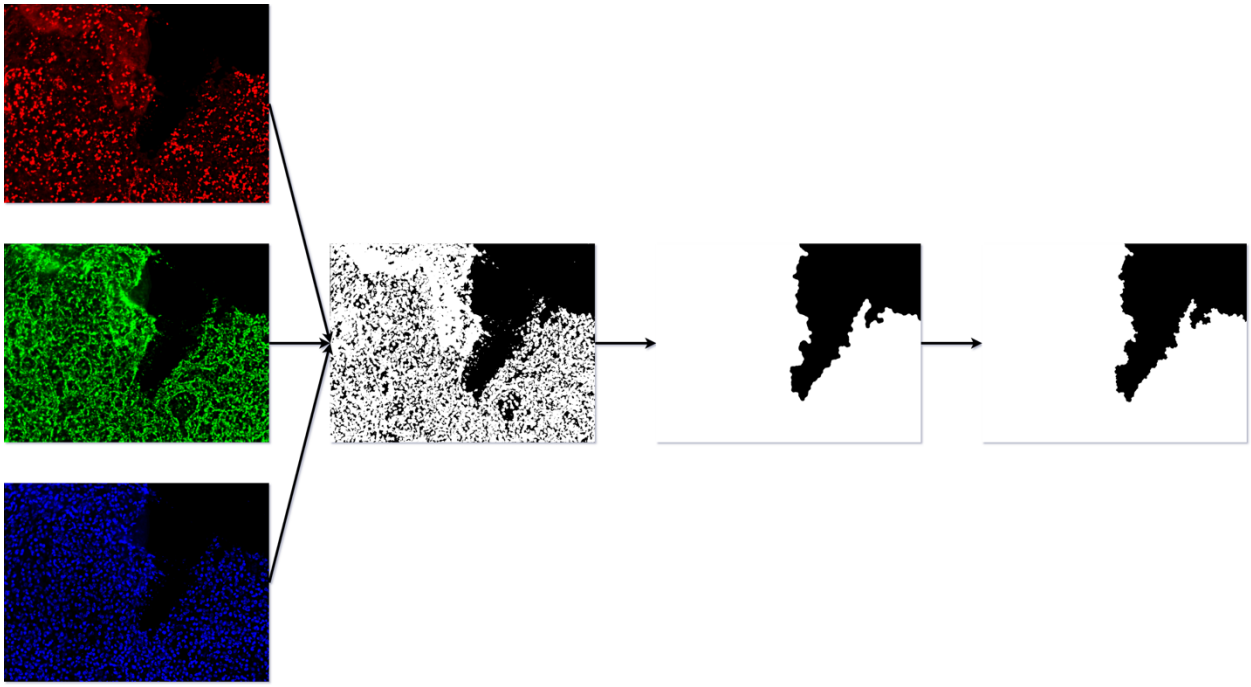

Figure 3. Foreground-background segmentation. Mean thresholding was applied on the cell membrane (red), cytoplasm (green), and nuclear channels (blue) to draw the approximate outlines of tissue (first binary image). Then an iterative morphological closing was applied to close the small and large gaps within the tissue (second binary image). Lastly, a geodesic active contour was applied with background areas as seeds to correct the boundaries between foreground and background (third binary image).

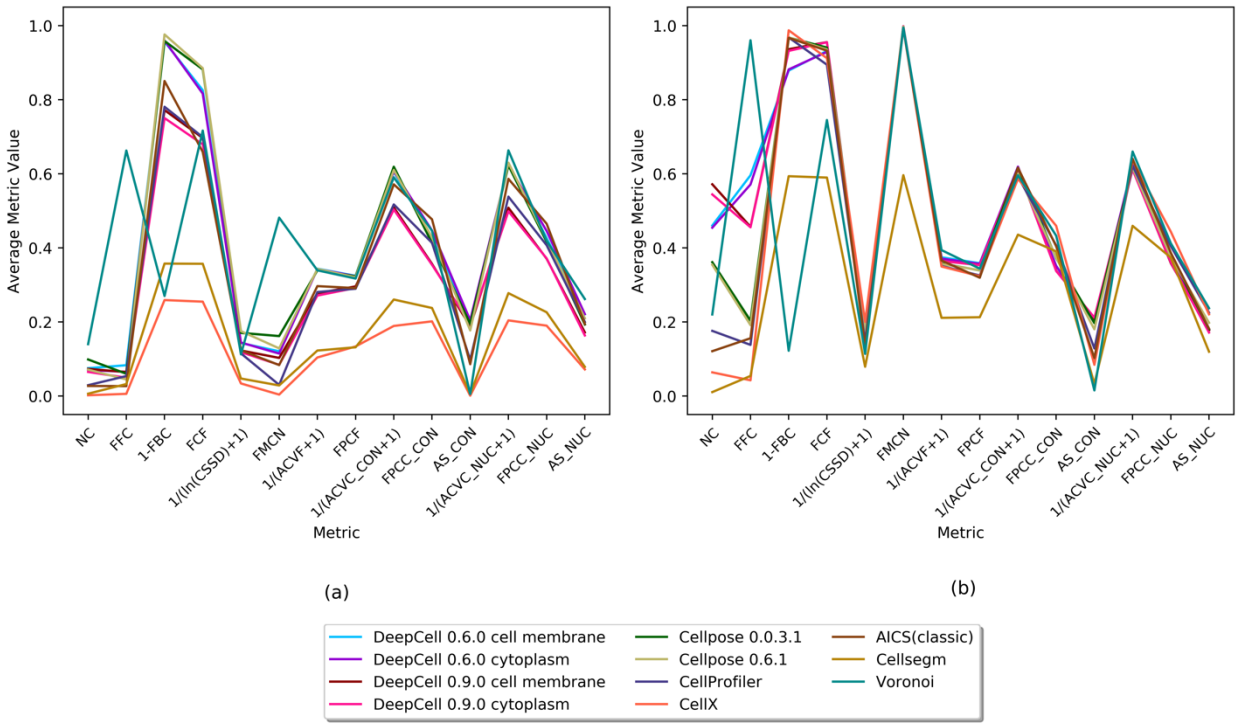


Figure 4. Heterogeneity of methods performance. For each segmentation method, each metric was averaged across all images. Methods are presented in unique colors. Panel (a) shows the results from methods that don't have the repair on nuclei have pixels outside the cell membrane. Panel (b) shows the results from methods with the repair.

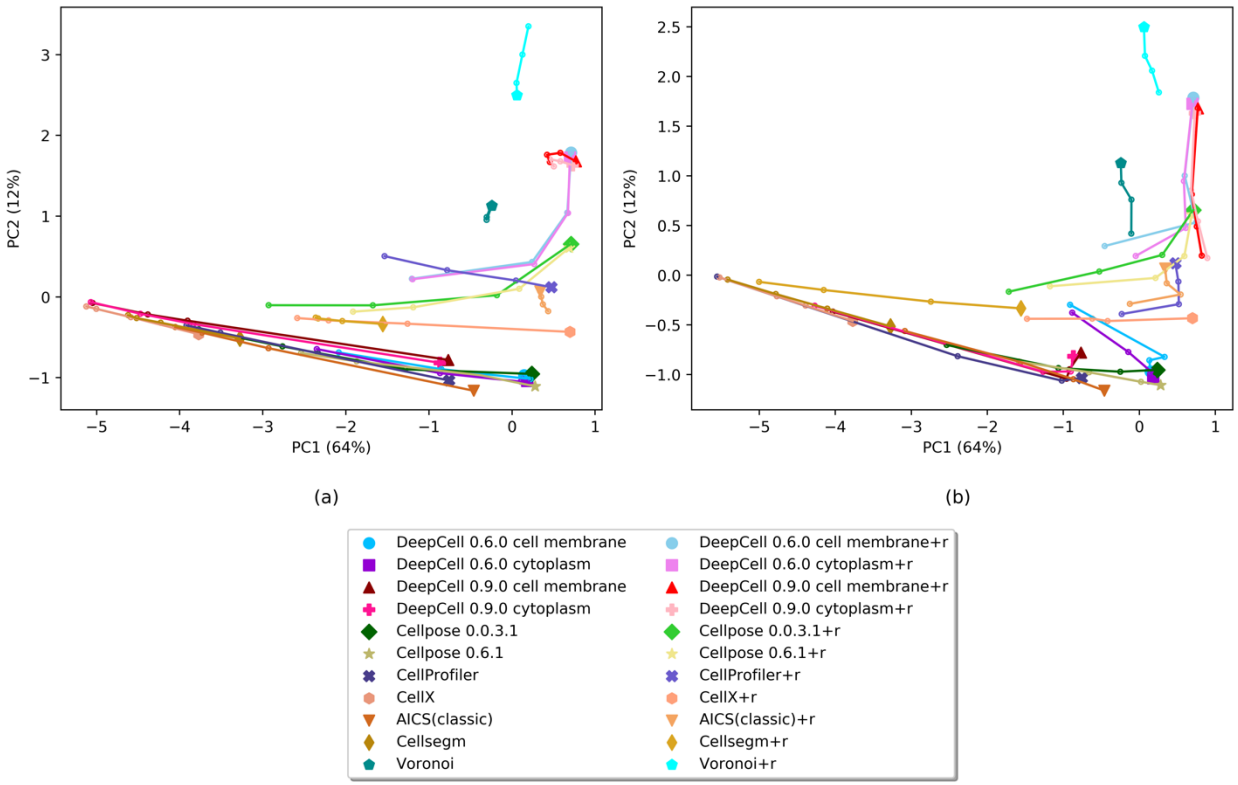


Figure 5 Top 2 Principal Components of each method on images from all modalities. Panel (a) shows the results from images with Gaussian noise perturbation. Panel (b) shows the result from images with downsampling. Each method is represented by a unique color. The points with a unique shape of makers represent the PCs on the original images. The trajectories represent PCs on low, medium and high Gaussian noise perturbed images on panel (a), or small, medium, large downsampled images on panel (b).

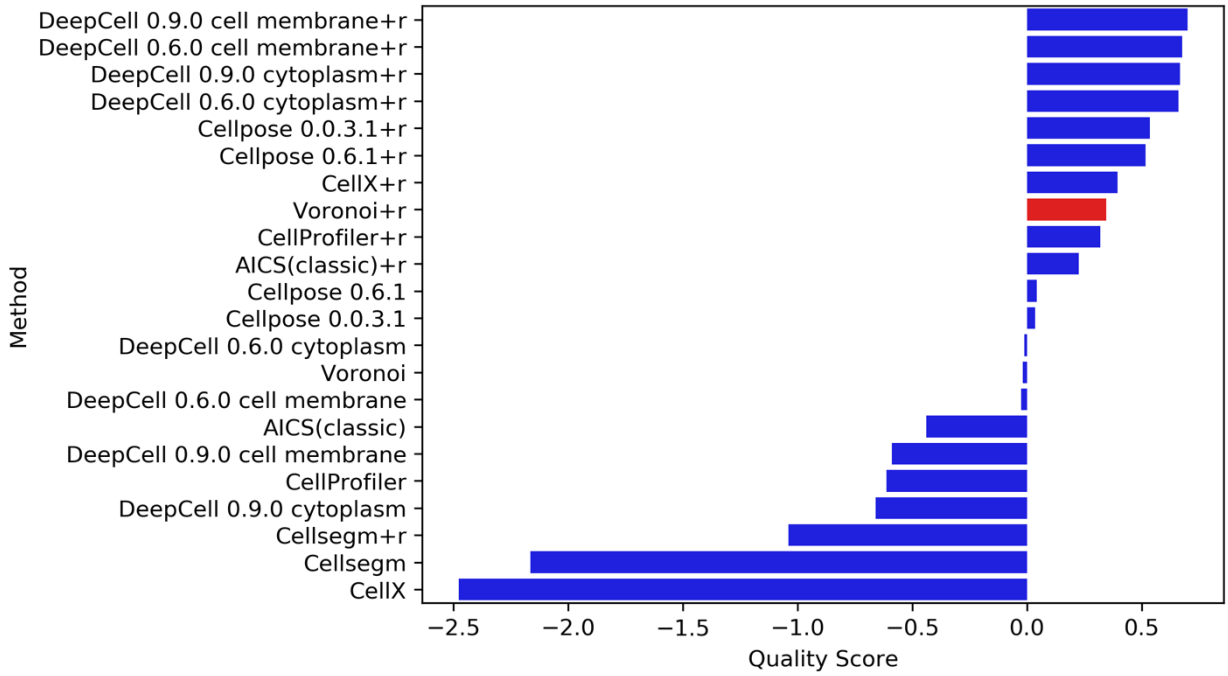


Figure 6 Overall quality score rankings of all modalities. All methods with and without repair are ranked by their quality scores. The methods that have a higher score than Voronoi are considered acceptable.

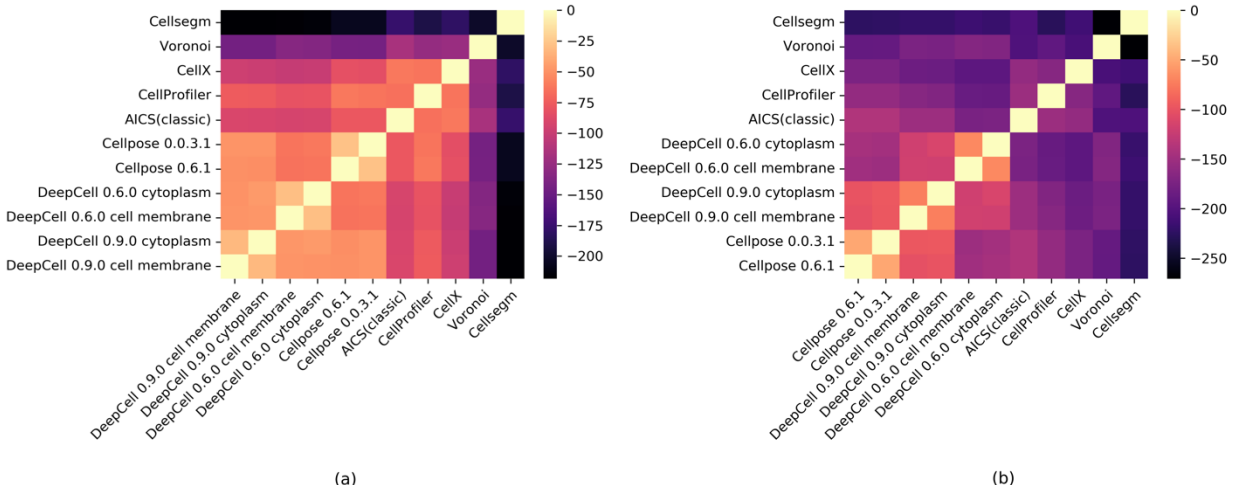


Figure 7 Heatmaps of similarity scores between segmentation methods. A score of zero represents the highest similarity (i.e. identical segmentations). More negative values indicate greater differences between methods. (a) Similarity heatmap generated by the difference between metrics of single method. (b) Similarity heatmap generated by the pairwise metrics. The methods are ordered by clustering.