

Extraction and Measurement of Morphological Features from Images using CellProfiler

The morphological features of the tiff image patches were extracted and measured using CellProfiler, an open-source cell image analyses software developed by the Broad Institute^{1,2}. Features were extracted by choosing and setting the parameters for different analyses modules. Specifically, by using the 'UnmixColors' module, the colors of image were separated based on hematoxylin staining and eosin Staining. Then the image is converted to grayscale for detection by choosing the 'combine' method of 'ColorToGray' module and inverted by using the 'ImageMath' module to make the background dark and tissue light. Subsequently, by using three sequential 'IdentifyPrimaryObjects' module, tissue region, HemaSubRegion (nuclei) and EosinSubRegion (cytoplasm) were identified using Otsu thresholding method. After the EosinSubRegion were masked by using the 'MaskImage' module, the tissue texture was measured by the 'MeasureTexture' module. The nuclei region was identified first using the 'MaskImage' on HemaSubRegion, then using per-object thresholding to take advantage of masking and 3-class thresholding with middle class setting to background to discriminate between the bright nuclei and remaining dim tissue area using the 'IdentifyPrimaryObjects' module. Then 'MeasureGranularity', 'MeasureTexture', 'MeasureObjectNeighbors', 'MeasureObjectSizeShape' modules were used to measure the granularity and texture of the nuclei, nuclei neighbor counts, percentage of nuclei touching, and nuclei size and shape. Finally, the areas occupied by the tissue and by the two subregions (nuclei, cytoplasm) were measured by the 'MeasureImageAreaOccupied' module and the fraction of the nuclei area of the whole tissue area was calculated by the 'CalculateMath' module. Results were exported to

spreadsheet and batch files were created to be able to process all images in batch using a computational cluster. We have extracted 943 morphological features and the measurement for these 943 morphological features of the 20 representative tiff images for each SVS slide were summarized using the sample mean slide level. When a patient had multiple tissue slides, the summarized value of the morphological features from the multiple slides were averaged to represent the value in the patient for further statistical analyses.

Supplemental Table S1. Enriched Categories of 18 Features Associated with ADC Predictive Model

Feature Category	Number of Features	P-value of Enrichment
Tissue_Texture	9	0.00035
Nuclei_Size_Shape	4	0.25
Tissue_Granularity	3	0.49
Nuclei_NeighborDistribution	2	0.75
Nuclei_Granularity	0	1
Nuclei_Occupation	0	1
Nuclei_Texture	0	1

943 morphological features defined and extracted by cell profiler can be further classified into 7 categories. The categories associated with the top 18 morphological features in classifying lower and higher risks for TCGA ADC patients' survival are ranked by RobustRankAggreg software package by choosing Stuart method. The number of features in each category and the enrichment P-value are as listed.

Supplemental Table S2. Enriched Categories of 12 Features Associated with SCC Predictive Model

Feature Category	Number of Features	P-value of Enrichment
Tissue_Granularity	6	0.0036
Nuclei_NeighborDistribution	3	0.24
Tissue_Texture	3	0.24
Nuclei_Granularity	0	1
Nuclei_Occupation	0	1
Nuclei_Size_Shape	0	1
Nuclei_Texture	0	1

943 morphological features defined and extracted by cell profiler can be further classified into 7 categories. The categories associated with the top 12 morphological features in classifying lower and higher risks for TCGA SCC patients' survival are ranked by RobustRankAggreg software package by choosing Stuart method. The number of features in each category and the enrichment P-value are as listed.

Reference:

1. Lamprecht MR, Sabatini DM, Carpenter AE: CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques* 42:71-5, 2007
2. Carpenter AE, Jones TR, Lamprecht MR, et al: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* 7:R100, 2006