Lymphocyte-Detection Pipeline FAQ

Changelog History			
Date	**Version**	**Content**	**Author**
5/26/2018	Rev01	First Draft	J Saltz et al.
6/3/2018	Rev02	Release Copy	J Saltz et al.

Reference:

J Saltz et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Reports 2018, Volume 23, Issue 1, 181 - 193.e7.

Table of contents

Where can I access the source code of the paper?

The source code can be found at github repository: https://github.com/SBU-BMI/u24_lymphocyte.

We recommend using develop branch.

What are included in the source code?

The source code includes data pre-processing (patch extraction from whole slide images, data augmentation), training and prediction models, heatmap generation, etc. Detail in the README of the repository.

What is the framework of the CNN model?

The CNN models are developed using theano and lasagne. Model architecture can be found at u24_lymphocyte/training/lymphocyte/ of the github repository and the paper Hou et al., 2017 referred in this paper.

Github repository: https://github.com/SBU-BMI/u24_lymphocyte.

Where can I access and use the pre-trained CNN models?

The pre-trained models are included in folder “cnn_models” including cae_model.pkl, cnn_lym_model.pkl, and cnn_nec_model.pkl from the link below. Put them into the corresponding folder inside u24_lymphocyte/data from the github repository and use during prediction phase (cae_model.pkl goes to u24_lymphocyte/data/models_cae/, cnn_lym_model.pkl and cnn_lym_model.pkl go to u24_lymphocyte/data/models_cnn/). More detail in the README of the github repo.

Pre-trained models: https://stonybrookmedicine.app.box.com/s/ecr7ba8czvqygw90iym0hwpnprrofoas

Github repository: https://github.com/SBU-BMI/u24_lymphocyte.

What is the dataset used?

We use around 5200 H&E slides from 13 cancer types, in The Cancer Genome Atlas (TCGA) repository which can be accessed from the link below.

https://gdc.cancer.gov/about-data/publications/pancanatlas

How are the labels collected to train CNN models?

Pathologists will look at whole slide images, zoom in and find the regions of interest which include tumor-infiltrating lymphocytes (TILs) labeled as positive and non-TILs labeled as negative. CNN models are then trained as a binary classifier on these supervised data.

Why do you predict necrosis?

The necrotic nuclei of dying tumor and inflammatory cells can confuse the single CNN model due to having similar visual features and characteristics to lymphocytes. Therefore, we train 2 separate models, one to classify lymphocytes and the other one to classify necrosis. For a given slide, the predicted probabilities of these 2 models are combined to generate the final prediction.

In the Figure 1 of the paper, the patch size of Lymphocytes is 50x50 μm^2 while the patch size of Necrosis is 500x500 μm^2. What is the intuition of using different size?

We found that the lymphocyte CNN produces false positives in necrotic regions because pyknotic (condensed and dark) nuclei in dying neutrophils and tumor cells within necrotic regions can have similar nuclear features when compared to lymphocytes. Compared to lymphocytes, necrotic regions should be classified at larger scales, 500x500 μm^2, since these regions are composed of aggregates of dying cells with more contextual information within the necrosis prediction model.

The CNN models are trained only on LUAD cancer type, why the model works for 12 other cancer types?

The CNN models were *initially* trained on LUAD samples, after which, we added samples from the other cancer types to the training set to evaluate and refine the performance of the CNN model in identifying lymphocytes (that look the same within all tissues regardless of tissue type). For all cancer types, we first visualize the resulting lymphocyte heatmaps of the model. If the heatmaps are not optimal, we label additional training image patches extracted from the specific cancer type and retrain the CNN with all available training data.

Do you use the same threshold to generate TIL maps for all slides?

No. Slides are divided into 8 groups. Slides in a same group share the same threshold.

How are groups assigned?

10 random patches in a specific range of predicted probabilities from all slides are chosen and ranked whether there are lymphocytes or not by pathologists. Based on the number of

labeled patches by pathologists, each slide is categorized into 1 of 7 groups named from A-G.

More detail in the section “Determining Lymphocyte Selection Thresholds” of the paper.

What is the format of the TIL maps?

The TIL maps are binary masks where 1s (displayed as red color) indicate cells predicted by the trained CNNs as TILs and 0s (displayed as blue color) indicate cells predicted as non-TILs.

What type of normalization is used on the data?

We did not apply color normalization on whole slide images or image patches for the CNNs. We applied data (color) augmentation during the training phase to make the classification of patches by the CNN in a robust manner with different H&E staining characteristics.

Which data are used to validate the AUC or what is the test data?

We used a test set and obtained the AUC results reported in the paper. The test set does not overlap with the training set and they are from the TCGA dataset. We released this "test set" as a validation set (validation_list_for_all.txt inside the “trainingdata” which can be downloaded from the link below).

https://stonybrookmedicine.app.box.com/s/ecr7ba8czvqygw90iym0hwpnprrofoas

What are the yellow lines around the slides in first column of Figure 5?

To give the readers a more comprehensive context, a pathologist circled tumor regions highlighted by the yellow lines in order to specifically assess the TIL map patterns in relation to the tumor because the lymphocyte-detection algorithm detects the lymphocytes and provides a heatmap for the whole slide image.

What is the purpose of clustering the TIL maps?

We cluster the TIL maps to quantify the distribution of the identified lymphocytes within the image and provide quantitative features that describe different TIL patterns and can be used for further analysis.

How to choose number of cluster for each slide?

The apcluster package in R derives the number of clusters automatically in an affinity propagation mathematical model to find spatially connected and coherent regions. Please refer to the software's manual for more detail.
https://cran.r-project.org/web/packages/apcluster/vignettes/apcluster.pdf

In some of the TIL map files, there are more than 1 image per patient. What do those images correspond to?

In some cases, there are multiple and different whole slide scans for the same patient. We only used one whole slide image (slides that end with DX1) for clustering per patient but we include TIL maps for all slides.

What are the meaning of 3 folders (TIL_maps_after_thres_v1, TIL_maps_before_thres_v1, TIL_maps_before_thres_v2) inside the til_maps folder?

1. TIL_maps_after_thres_v1: The released binary TIL maps.

    Red pixels: TIL patches.

    Blue pixels: non-TIL tissue patches.

    Black pixels: non-tissue patches.

2. TIL_maps_before_thres_v1: The grayscale TIL maps before thresholding.

    Red channel: the predicted probability of lymphocyte infiltration.

    Blue channel: tissue or non-tissue (binary: either 0 or 255)

    Green channel: the predicted probability of being a necrotic patch

3. TIL_maps_before_thres_v2: We trained the CNN models again using the same training data and code to show the variations of prediction due to the randomized training loss optimization and data augmentation steps. We did not apply the thresholding steps on these grayscale TIL maps.

There are many slides on TCGA dataset, why do you only use slides ended with DX1?

There is a special naming convention for TCGA slides:

- Diagnostic slides that have excellent image quality have the string DX1 or DX2 in them.

    For example, the DX1 just before the hash indicates a diagnostic slide: TCGA-D3-A1Q1-06Z-00-**DX1**.7E87AE47-8145-464D-AFBC-82DEAC42F492.svs

- Frozen sections that have poor image quality have a string BS# or TS# indicating frozen. Due to the low quality of these images, we did not utilize these slides.

Although we could find all the clustering features of the images in the files that you provide online, we couldn't find the cluster index of each patch?

The clustering features (eg Banfield-Rafferty, Ball-Hall, C index) are outputted for each slide in the TILmap_Summary.csv file. These indices are not outputted per patch.

What do the columns in the TILmap_Summary.tsv represent?

Column name	Meaning
"ParticipantBarcode"	Patient code
"Slides"	Case IDs which are used in the TCGA repository for identifying each slide.
"number of data points"	Number of TIL patches, which is the number of red pixels in the TIL maps.
"number of clusters"	Number of clusters, from affinity propagation clustering
"Til_percentage"	Percentage of TIL, ratio in % of total red pixels to all pixels
"N_cluster"	Number of clusters obtained from the clustering the data points
"NP_mean"	Mean of the cluster membership counts
"NP_sd"	Standard deviation of the cluster membership counts
"WCD_mean"	Mean of the values of WGSS, the within-cluster dispersion (WGSS is a within-cluster dispersion which is the sum of the squared distances between the observations and the barycenter of the cluster https://CRAN.R-project.org/package=clusterCrit)

"WCD_sd"	Standard deviation of the values of WGSS
"CE_mean"	Mean of the maximum distances to clusters exemplars. The cluster exemplar is the most representative TIL patch for the cluster, as defined in the affinity propagation method
"CE_sd"	Standard deviation of the maximum distances to exemplars
"Ball_Hall"	See below
"Banfeld_Raftery"	See below
"C_index"	See below
"Calinski_Harabasz"	See below
"Davies_Bouldin"	See below
"Det_Ratio"	See below
"Dunn"	See below
"Gamma"	See below
"G_plus"	See below
"GDI11" - "GDI53"	See below
"Ksq_DetW"	See below
"Log_Det_Ratio"	See below
"Log_SS_Ratio"	See below
"McClain_Rao"	See below
"PBM"	See below
"Point_Biserial"	See below
"Ray_Turi"	See below
"Ratkowsky_Lance"	See below
"Scott_Symons"	See below

"SD_Scat"	See below
"SD_Dis"	See below
"S_Dbw"	See below
"Silhouette"	See below
"Tau"	See below
"Trace_W"	See below
"Trace_WiB"	See below
"Wemmert_Gancarski",	See below
“Xie_Beni"	See below

Details of the cluster indices (Ball_Hall to Xie_Beni) could be found here:
https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf