Report

1. Core: a new deep multiple instance learning model to predict tumor purity from slides of cohort more accurately.

2. Function: have **consistent genomic tumor purity values** which can evaluate the percentage of cancer cells within tumor involving fewer manual steps. Besides, the model can obtain **spatially resolved tumor purity maps** which show that tumor purity varies spatially within a sample. Moreover, it can advance genomic analysis, help pathologists and decrease inter-observer variability caused by manual measurement. Plus, it can also help understand the tumor microenvironment.

3. Why MIL: patch-based model needs pixel-level annotations which are hard to acquire. While MIL does not, it can treat a bag of patches cropped from the slides as a sample which is much easier to get. Those sample-level weak labels are easier to obtain from many channels.

4. Traditional Method: Two main approaches to estimate tumor purity:

(1) Percent tumor nuclei estimation:counting the percentage of tumor nuclei over a ROI in the slide. Widely applicable and at cellular level, but tedious and exists inter-observer variability.

(2) Genomic tumor purity inference: inferring from different types of genomic data. Golden standard and produce consistent values on different data sets. However, do not work on low tumor content samples and lack spatial information of the cells' locations.

5. Our method:

(1) Data set: 10 different TCGA cohorts and a local Singapore cohort whose histopathology slides are segregated into training, validation and test sets.

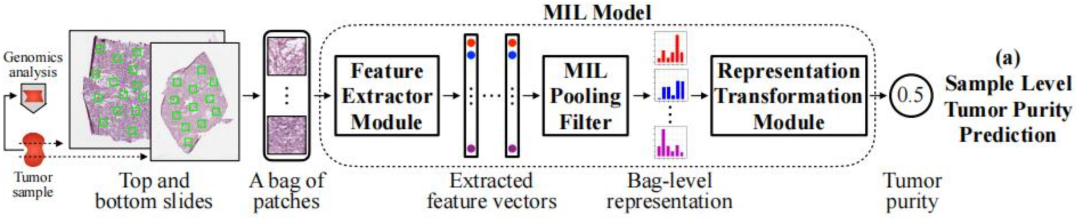
(2) Model for tumor purity value prediction:

a. Input: a bag of patches(200) cropped from the **top and bottom** slides of a tumor sample

b. Output: the sample's purity value prediction.

c. Ground-truth: tumor purity values from genomic sequencing data by ABSOLUTE

d. MIL Model details: The patches were sent to Resnet to extract features while the last fully connected layer was modified to acquire the desired number of channels(32). Then, The 32 features of each instance in each bag were sent into the Distribution Pooling and expanded to 11 bins(J mentioned in the paper). Later, each patch was represented as a 32*11 vector. Finally, a purity value was predicted according to that vector through another well-designed fully connected layer (ReoresentationTransformation).



(3) Model for spatial tumor purity map:

a. Input: a bag of patches each 1mm2 cropped from ROIs.

b. Output: the tumor purity values in ROIs, which construct the spatial tumor purity map

6. Shiny points:

(1) Used a novel **'distribution' pooling filter** proposed before to produce stronger bag-level representations.

(2) Applied the MIL model into automated tumor purity prediction instead of traditional machine learning methods or coarse DNNs.

(3) Used the model to predict each small patch and group into the spatial map which can complement spatial-omics

(4) Clustered the patches using hierarchical clustering over the extracted feature vectors and acquired segmentation maps on those samples with a matching normal sample

(5) Confirmed the effectiveness of the new method using all kinds of ablation experiments

7. Possible reasons for error in predictions

(1) Lack of more data

(2) The variation of tumor purity for the same sample in different regions

(3) Genetic changes over time

8. Weakness:

(1) Can not tell whether each cell is cancerous or not. Only a proportion of cancer cells within a patch.

(2) Performance of samples with low tumor content needs to be strengthened with more related data.

(3) Lack of effectiveness in external cohorts due to differences between fresh slices and old sample slices.

(4) No normal samples in the Singapore cohort