Introduction

Tumor purity is the proportion of cancer cells in a sample of tumor tissue. It is a critical index in many aspects of cancer research and clinical diagnosis. In terms of cancer research, tumor content of sample subjects is evaluated when we are screening the proper sample for genomic analysis. In this case, a sample with high tumor purity result would largely increase the efficiency of our analysis, for example next-generation sequencing. Furthermore, the accuracy of tumor purity estimation is important for clinicians to deliver subsequent therapeutic strategy to cancer patients.

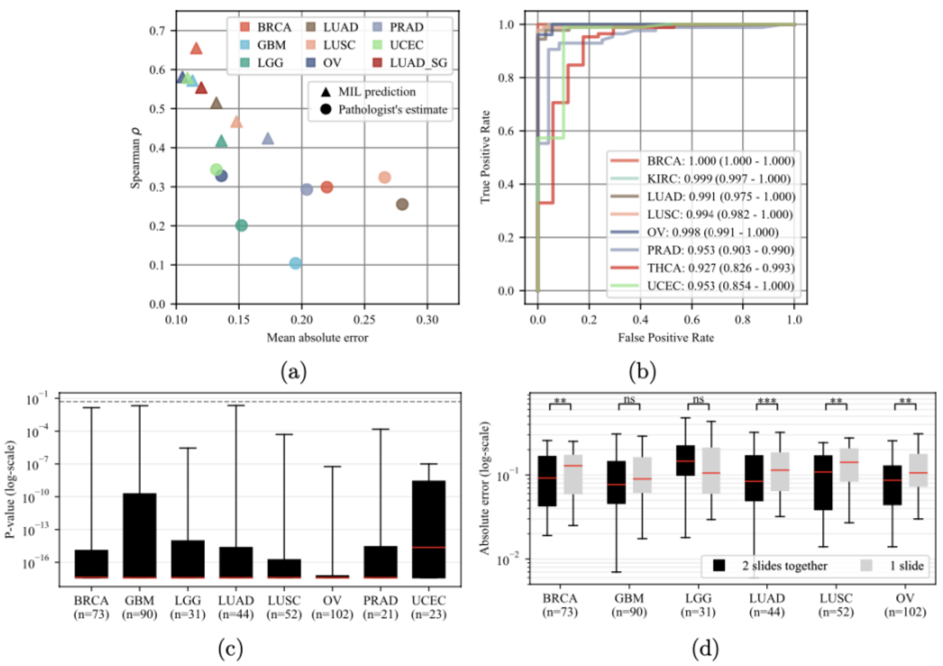
After understanding the importance of tumor purity estimation, Oner, M.U. *et al.* worked on developing an accurate estimation method for tumor purity. [1] There are two main methodologies: tumor nuclei percentage estimation by pathologists and genomic tumor purity inference. The latter was chosen by Oner and his colleagues since they suggested that the former approach could provide significant variability among different pathologists' estimates. Instead of traditional approach, tumor purity in genomic analysis is predicted from several genomic data types, namely somatic copy number data, somatic mutations data, gene expression data, and DNA methylation data. However, the current method cannot be well-applied into samples with low tumor content, and spatial information about cancer cell location cannot be provided. Therefore, in order to eliminate these limitations, Oner and his colleagues introduced a new technique that predicts the tumor purity from H&E-stained histopathology slides through a multiple instance learning (MIL) model. The top and bottom of H&E slides are cropped into patches, which are then collected to form a bag. Each bag was analyzed based on three modules of their MIL model: feature extractor and bag-level representation transformation module using neural network, and MIL pooling filter via their 'distribution' pooling filter, which is more advance than standard pooling filter. In this experiment, H&E slides were obtained from 10 different cohorts: 9 of them were fresh frozen in The Cancer Genome Atlas (TCGA) and one included slides of formalin-fixed paraffin-embedded (ffpe) sections from local Singapore cohort. The corresponding genomic sequencing data was also collected as positive control.

Results & Discussion

By analyzing the correlations between the prediction of their model and tumor purities from the quantification of somatic DNA alteration, slides from 8 cohorts out of 10 demonstrated significant correlations, which means their model could provide a relatively accurate predictions in tumor purity. Figure 1a revealed that MIL predications in 9 cohorts have lower mean absolute error but higher Spearman's correlation coefficient when compared with Pathologist's estimate. Furthermore, when the model dealt with slides of ffpe sections, it can be trained to adapt to the weight of the firsr convolutional layer via transfer learning.

Spatial evaluation provided evidence that the top and bottom patches of one single slide differ in tumor purity, and the combination of these two results gave rise to a better prediction of tumor purity, especially in the results of BRCA, LUAD, LUSC, and OV cohorts. (Figure 1d) Additionally, the author found the reason why pathologists' estimation usually higher than what their model predicted. In terms of mean absolute error and patches that were most-frequently selected, it was because of region-of-interest. (Figure 2a&b) Pathologists commonly are inclined to select higher tumor purity regions to conduct predication process. Most interestingly, their model was able to learn differences between cancerous and normal tissue histology, and then classified the corresponding slides

Overall, Oner, M.U. *et al.* developed a novel machine learning model, which gave a new insight on automate tumor purity predication of H&E slides in different types of cancer. This study upgraded the current estimation approaches and explored it in a spatial aspect, which largely increased accuracy and avoided unnecessary time consumption. For its future direction, aside from what authors mentioned in their paper, training their model with specific DNA sequencing dataset for specific cancer might be a direction of further developing their MIL model. There are types of genes that are specifically responsible for the diagnosis of a certain type of cancer. [2] If this model can be trained by the specific data set, it will be possible to be applied more specific and accurate in different types of cancer.



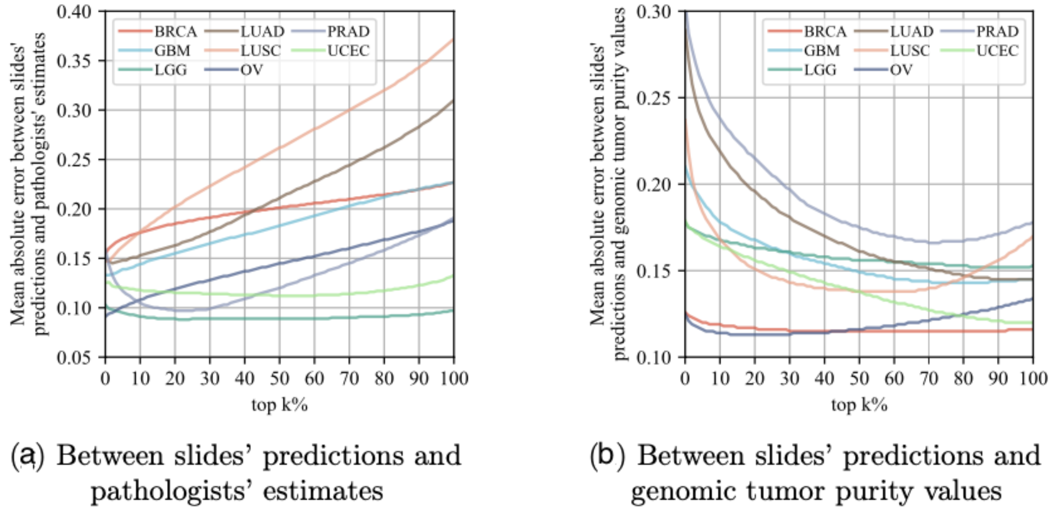
(a)

(b)

(c)

(d)

Figure 1 Evaluation in different cohorts



(a) Between slides' predictions and pathologists' estimates

(b) Between slides' predictions and genomic tumor purity values

Figure 2 Tumor purity predication thorugh pathologist'estimates and genomic tumor purity

Reference

1. Oner, M.U., et al., *Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study*. Patterns (N Y), 2022. **3**(2): p. 100399.

2. Li, Y., et al., *Putative biomarkers for predicting tumor sample purity based on gene expression data*. BMC Genomics, 2019. **20**(1): p. 1021.