The purpose of this paper is to estimate tumor purity more accurately and obtain spatially resolved tumor purity maps. The percentage of cancer cells within the tumor is called tumor purity, and an accurate tumor purity estimation is of great clinical importance.

There are two main approaches to estimate tumor purity. The pathologist estimates tumor purity by counting the percentage of tumor nuclei over a region of interest in the slide, which is called percent tumor nuclei. However, counting tumor nuclei is tedious and time-consuming, and pathologists' estimates may have inter-observer variability. The other approach is inferring tumor purity from different types of genomic data. It is referred to as genomic tumor purity in this paper, which is accepted as the golden standard. But they do not provide spatial information of the locations of the cancer cells.

This paper develops a machine learning model to predict tumor purity. The predictions were highly consistent with genomic tumor purity values. The model cut the cost successfully. It also provides information about the spatial organization of the tumor microenvironment, which helps reduce pathologists' workload and the inter-observer variability. Moreover, this paper analyzes the causes of the difference.

The machine learning models this study designed are called multiple instance learning (MIL) models, which predict the tumor purity from H&E stained histopathology slides. It represents a sample as a bag of patches cropped from the sample's slides and uses a sample-level label as the bag label(Methods). Sample-level labels are weak labels providing only aggregate information rather than pixel-level information. Yet, they can easily be collected from pathology reports, electronic health records, or different data modalities, which don't require pathologists' pixel-level annotations.

The data this paper analyzed is from ten different TCGA cohorts and a local Singapore cohort. In each TCGA cohort, a patient has only one tumor sample and one matching normal sample. And there were no normal samples in the local Singapore cohort. The histopathology slides in each cohort were randomly segregated at the patient level into training, validation, and test sets. Then, we trained our MIL model on the training set, chose the best set of model weights based on validation set performance, and evaluated the best model on the held-out test set.

And the novel MIL model consists of three modules: *feature extractor* module, *MIL pooling filter*, and *bag-level representation transformation* module (Figure 1a). The authors use neural networks to implement the *feature extractor* module and the *bag-level representation transformation* module to parameterize the learning process fully (Methods) As the *MIL pooling filter*, the authors use their novel 'distribution' pooling filter, which obtains a strong bag-level representation by estimating the marginal distributions of the extracted features.

The model accepts a bag of patches cropped from the top and bottom slides of a sample as its input. And the *feature extractor* module extracts a feature vector for each patch inside the bag. Then the *MIL pooling filter*, namely 'distribution' pooling, summarizes extracted features into a bag-level representation by estimating marginal feature distributions. Finally, the *bag-level representation transformation* module predicts the sample-level tumor purity.

Both fresh-frozen sections in TCGA cohorts and formalin-fixed paraffin-embedded (ffpe) sections in the Singapore cohort successfully predict tumor purity. The predictions are consistent with the genomic tumor purity values. In Figure 2, all data points are around the diagonal, explaining the prediction error is low. For TCGA cohorts, the MIL model's tumor purity predictions correlate significantly with genomic tumor purity values. The maximum Spearman's ρmil = 0.655 (P = 4.6e-24; 95% CI: 0.547 - 0.743) was obtained, which is higher than pathologists' percent tumor nuclei estimates. This implies that MIL predictions are more consistent with genomic tumor purity values than the pathologists' percent tumor nuclei estimates. Moreover, the authors observed that correlation coefficients obtained from MIL predictions were significantly better than ones obtained from pathologists' estimates in all cohorts except LUSC and PRAD. In the analyses of MIL predictions, most maximum mean-absolute-error values are lower than the maximum mean-absolute-error values in the analyses of pathologists' percent tumor nuclei estimates. For the local Singapore cohort, the authors obtained Spearman's ρ_{mil} = 0.554 and the mean-absolute-error of μ_{emil} = 0.120 (σ_{emil} = 0.091). The results suggested that the MIL models learned robust features for tumor purity prediction tasks at the higher levels of the network.

Besides, predicting the tumor purity of a sample by using both top and bottom slides is better than using only one slide. Through the Wilcoxon signed-rank test on the difference between tumor purity prediction obtained from MIL model by using both of the slides together and individual slides, some cohorts using both slides for tumor purity prediction gave better results in terms of absolute error. Others show no significant difference using both slides or one slide alone, which are the most spatially homogeneous tumors. The tumor purity varies spatially within the sample, so the authors obtain spatially resolved tumor purity maps.

The model provides information about the spatial organization of the tumor microenvironment. It accepts a bag from patches in each region of interest(ROI) and predicts the tumor purity of the region, then assigns tumor purity values to the center of ROIs, then the authors obtain spatial tumor purity maps. They can help understand the interaction of cancer cells with other tissue components.

What's more, the MIL model learned discriminant features for cancerous vs. normal histology. The model can discriminate the extracted feature vector for each patch inside the bag and use hierarchical clustering. It successfully classifies samples into tumor vs. normal without requiring exhaustive annotations from pathologists.

In all cohorts, percent tumor nuclei estimates were generally higher than genomic tumor purity values. The mean-absolute-error between the slides' predictions and pathologists' percent tumor nuclei estimates increases as the authors extend the region-of-interest to cover the lower tumor purity regions (Figure 4g). The pathologists might have selected high tumor content regions to estimate percent tumor nuclei, which might have caused high tumor nuclei estimates. Besides selecting the region-of-interest, the region's size is also crucial for some cancer types.

However, the tumor purity MIL models successfully predict slightly deviated from the genomic tumor purity values. One of the reasons is that researchers have fewer patients in our data sets than traditional deep learning data sets. And for samples with only one slide, the prediction error is expected to be higher. Lastly, all the effects of genetic changes (so, the genomic tumor purity changes) may not be observable from the slides due to the selective dying characteristics of H&E staining.