# Report of Question1

## The background of the research

Nowadays, Tumor purity is estimated by two main approaches: percent tumor nuclei estimation and genomic tumor purity inference. The tumor purity affects both high throughput data acquisition and analysis. However, about percent tumor nuclei estimation, counting tumor nuclei by scientist is time-consuming and the inter-observer variability is inevitable from different pathologists' estimates. And genomic tumor purity inference (inferred from different types of genomic data, such as somatic copy number data, somatic mutations data, gene expression data, and DNA methylation data.) can not apply to the low tumor content samples. Besides, they do not provide spatial information of the locations of the cancer cells.

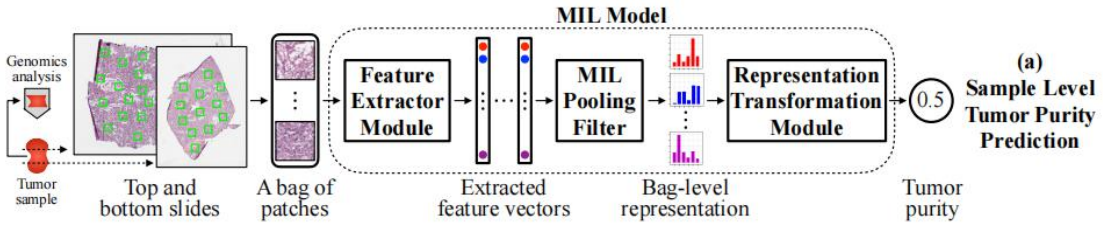
Thus, both genomics methods and pathologists' slide reading approach have different strengths and limitations.

## A machine learning model that solved the problem of percent tumor nuclei estimation and genomic tumor purity inference

In order to solve the problems cause by percent tumor nuclei estimation and genomic tumor purity inference, this study develops a machine learning model l that predicts the tumor purity from H&E stained histopathology slides and the predictions are consistent with the genomic tumor purity values, very few manual links are involved.

Two types of machine learning models can be utilized to predict tumor purity from digital histopathology slides: patch-based models and multiple instance learning (MIL) models.

**A novel MIL model**



The model represents each sample as a bag of patches cropped from the sample's top and bottom slides and use the sample's genomic tumor purity as the bag label. The MIL model has a novel 'distribution' pooling filter that produces stronger bag-level representations from patches' features than standard pooling filters like max and mean pooling.

# The results of the study

The analysis of the MIL model used data from ten different TCGA cohorts and a local Singapore cohort. The histopathology slides in each cohort were randomly segregated at the patient level into training, validation, and test sets. Then, the MIL model trained 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.

In 10 different TCGA cohorts, the The MIL model's tumor purity predictions correlate significantly with genomic tumor purity values and MIL models' predictions have lower mean absolute error than percent tumor nuclei estimates

The result of the analysis illustrated that the model can be used as predicting the tumor purity. The findings also suggested that it was better to use both slides of the sample for tumor purity prediction whenever available. Moreover, we obtained spatially resolved tumor purity maps showing the variation of tumor purity over a slide， Predicting the tumor purity of a sample by using both top and bottom

slides is better than using only one slide

# Limitation

However, checking their performance on samples with low tumor content (where ABSOLUTE cannot determine the tumor purity values accurately) would strengthen the applicability of our MIL models.