Report of Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study

1. Relevant background

Tumor purity is one of the crucial factors affecting the quality of high throughput genomic analysis, which is indispensable for cancer research.

Tumor purity is estimated by two main approaches: ①percent tumor nuclei estimation and ②genomic tumor purity inference.

① Percent tumor nuclei estimation:

A pathologist estimates tumor purity by reading H&E stained histopathology slides and counts the percentage of tumor nuclei over a region of interest in the slide.

Feature:

1) Widely applicable, has a cellular level resolution.
2) Counting tumor nuclei is tedious and time-consuming, and exists inter-observer variability between pathologists' estimates.

② Genomic tumor purity inference

Infer from different types of genomic data.

1) Can mitigate confounding effects of normal cell contamination and produce consistent values on different cancer data sets.
2) Do not apply to the low tumor content samples, do not provide spatial information of the locations of the cancer cells.

2. Writing purposes

A. Develop a method that breaks through the limitations of the two existing methods

B. Provides some insights into the probable causes why the results that different pathologists estimates using percent tumor nuclei estimates are not only inconsistent but also different from genomic tumor purity values.

3. Solutions

Formulate predicting tumor purity of a sample from its H&E stained histopathology slides as a MIL task and develops a novel MIL machine learning model that predicts the tumor purity from H&E stained histopathology slides.

1) Using readily available histopathology slides in clinic and involves few manual steps thus cost-effective
2) Provide information about the spatial organization of the tumor microenvironment.
3) Does not require pixel-level annotations. (Compared with patch-based models)

4. Results

1) Using qualitative validation showed that the proposed MIL models can produce correct segmentation, and showed that the proposed MIL model classifies samples into tumor vs. normal almost perfectly in all cohorts. The proposed novel MIL models is proved to have surpassed the two existing methods in several respects.
2) Found that the top and bottom slides of a sample were significantly

different in tumor purity, suggest that it is better to use both slides of the sample for tumor purity prediction whenever available.

3) Investigated the probable causes of pathologists' percent tumor nuclei estimates were usually higher than genomic tumor purity values and suggested that pathologists might have selected high tumor content regions to estimate percent tumor nuclei. The study also observer that besides selecting the region-of-interest, its size is also crucial for some cancer types.

5. Limitations

1) The MIL model yet does not perform well on samples with low tumor content.
2) The study yet could not validate the model on external cohorts due to tissue preservation methods
3) The model requires larger cohorts to improve the performance.