Reading Report of *Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study(Oner et al., 2021)*

Tumor purity is defined as the percentage of cancerous cells in tumor microenvironment (Mao et al., 2018). The accuracy of tumor purity plays and important role in clinical diagnosis as well as in high throughput genomic analysis. When determining the genomic variance of a tumor sample, an accurate tumor purity estimation should be obtained ahead to ensure enough tumor content to perform convincing DNA sequencing. The percent tumor nuclei estimation and genomic tumor purity inference are two main ways to estimate tumor purity. The former one relies on pathologist visually counting tumor nuclei on H&E stained sample slide under microscope, which is widely applicable yet time-consuming and inconsistent between different pathologists. The latter method inferred tumor purity from various genomic data, and it is considered as the golden standard nowadays. However, genomic method is not applicable when the samples tumor content are low. Therefore, in this paper, authors developed a machine learning model that predict tumor purity from H&E stained histopathology slides (Oner et al., 2021). They designed a novel multiple instance learning (MIL) model which represents the sample as a bag of patches cropped from the sample tumor's slides and label the bag using sample –level label. The H&E histopathology slides and corresponding genomic sequencing data were obtained from ten different cohorts in The Cancer Genome Atlas (TCGA) and one cohort from Singapore. Each sample was chopped into portions and one of the

portions was used for genomic tumor purity inference. The top and bottom histopathology slides of that portion were also prepared. Sample slides were randomly segregated into training, validation and test sets. The training set was used to train the machine learning model (Oner et al., 2021). This novel MIL model consisted of three modules: feature extractor module, MIL pooling filter and bag-level representation transformation module. When a bag of patches is given, the feature extractor module extracts a feature vector for each patch and the pooling filter then obtains a strong bag-level label by estimating the marginal distribution. In the end, the bag-level representation transformation module predicts tumor purity value. For each cohort where the dataset was obtained, they evaluated the performance of this trained MIL model on unseen patients in test set. It turns out this model successfully predicted tumor purity of samples in the test set of each cohort, as the predictions correlate significantly with genomic tumor purity values when Spearman's rank correlation coefficient is used as the performance metric. More notably, authors repeated the same analysis between genomic tumor purity values and percent tumor nuclei estimates, the results indicate that MIL predictions are more consistent with genomic tumor purity.

Overall, this MIL module had good performance on predicting tumor purity. It is more time-efficient than the pathologist's percent tumor nuclei estimate. It also provides spatial organization of the tumor which could not be provided by genomic tumor purity inference method.

Reference

Mao, Y., Feng, Q., Zheng, P., Yang, L., Liu, T., Xu, Y., Zhu, D., Chang, W., Ji, M., Ren, L., Wei, Y., He, G., & Xu, J. (2018). Low tumor purity is associated with poor prognosis, heavy mutation burden, and intense immune phenotype in colon cancer. *Cancer Management and Research, Volume 10*, 3569–3577. https://doi.org/10.2147/cmar.s171855

Oner, M. U., Chen, J., Revkov, E., James, A., Heng, S. Y., Kaya, A. N., Alvarez, J. J., Takano, A., Cheng, X. M., Lim, T. K., Tan, D. S., Zhai, W., Skanderup, A. J., Sung, W.-K., & Lee, H. K. (2021). Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. https://doi.org/10.1101/2021.07.08.451443