QUESTION 1: REPORT ON PREPRINT

CANCER BIOLOGY

Tumour forms and develops in a complex and dynamic environment (Figure 1), comprising of heterogenous cellular and non-cellular components cross-talking to orchestrate signalling networks that affect prognosis and treatment response. To study tumour microenvironment, genomic and transcriptomic landscape data derived from bulk tumour samples provide comprehensive information about tumour immune microenvironment and defines molecular subtypes of different cancers (Thorsson, 2018). Immunogenomic analysis of tumour sample from individual patient is critical for immunotherapy treatment strategies for personalised and precision medicine.

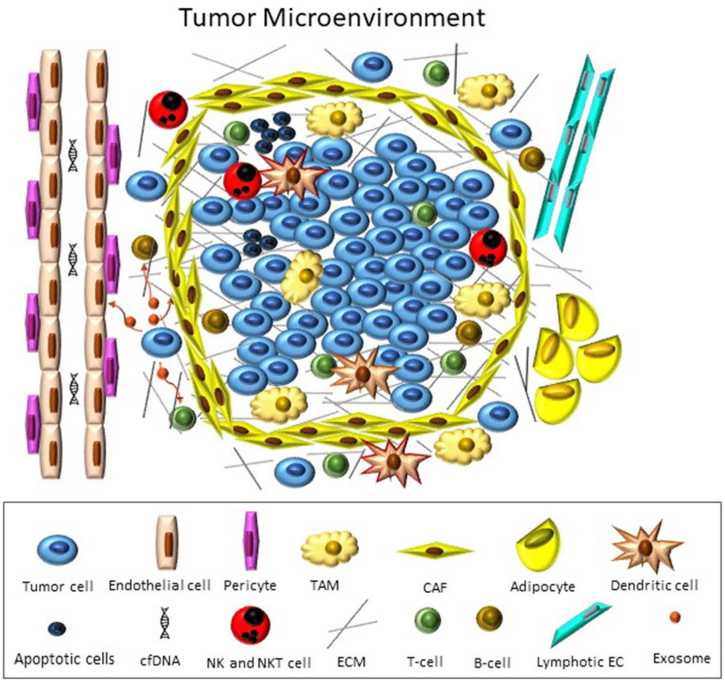

Figure1. Cellular and non-cellular components of tumour microenvironment (Baghban, 2020)

Bulk tumour samples used for high throughput methods, such as next generation sequencing and gene array chips, require tumour purity for assay performance and data interpretation. Tumour purity or tumour nuclei content (%TN) is the percentage of tumour cell content within a surgically removed tumour sample, with at least 20% TN required for most high throughput methods. Conventionally, tumour samples are sent to a medical board-certified pathologist to visually determine tumour purity, then sent to accredited clinical laboratory for integrated genomic analysis, such as somatic cell copy number array, whole exon sequence, DNA methylation array, mRNA sequence, microRNA sequence. These genomic analysis also provides genomic tumour purity, which should highly correlate with visually determined tumour purity, but confounded by sample and human factors (Smits, 2014). To automate this tumour purity determination process, the authors of this preprint develop a deep learning-based multiple instance learning (MIL) model called Spatially Resolved Tumor Purity Maps (SRTPMs), to automatically perform large amount of bio image analysis and reproducibly extract quantitative information. Overall, bio image analyses expand immense potential to capture spatial patterns when coupled with genomic analyses.

MACHINE LEARNING

There are two main categories of machine learning; supervised machine learning where new input is predicted based on previous data points, and unsupervised machine learning where data points relate and form clusters based on traits. Within these two main categories of machine learning, there are ten methods.

1. Regression where a numerical value is predicted based on previous data
2. Classification where a class label is predicted based on previous data
3. Clustering where observations with similar characteristics are group together
4. Dimension reduction where dimension of feature space is reduced to maximise linear relationship of data
5. Ensemble where several predictive models combine to achieve higher quality of prediction
6. Neural net and deep learning where layers of parameters are added to capture non-linear relationship
7. Transfer learning where a fraction of trained layers from a neural net is transferred to a new similar task
8. Reinforcement learning where learning is cumulative with no previous data

In this preprint, ten different TCGA cancer cohorts and one Singapore cohort where each patient has one tumour sample and one matching normal sample if available. All whole slide images were divided into 5 folds to train the MIL model using fold0, fold1, fold2, validate performance on fold3 to select the best model, and test the model with fold4. A H&E image of tumour sample is cropped into patches and stored in a bag labelled with the tumour sample's genomic tumour purity.

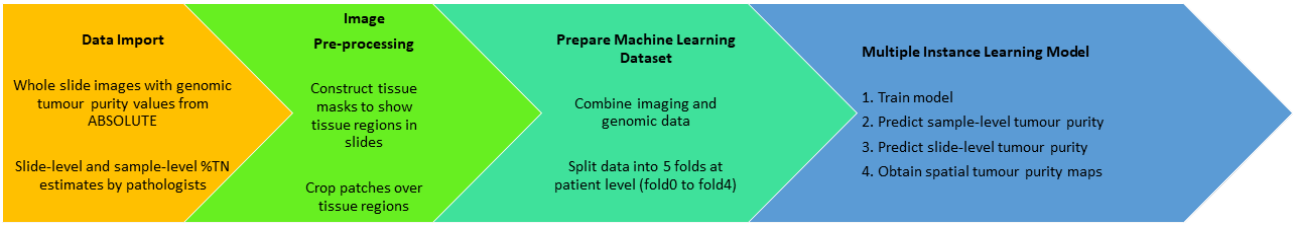

Figure 2. Implementation of SRTPMs on TCGA datasets.

Three stages of MIL model:

1. Feature extractor module to extract features, output as feature vectors and construct feature matrix
2. Pooling filter module to aggregate extracted feature vectors into a bag-level representation
3. Bag-level representation transformation module to transform bag-level representation into predicted bag label

In digital pathology, assigning labels is time-consuming and labour-intensive in order to establish annotated input images to train the model. Supervised MIL provide an automated annotation process to reduce the burden on human. However, slide scans are large and high in resolution, as like most bio images, usually does not fit in memory. Therefore, the tumour images in this preprint or MNIST dataset utilises cropped patches in labelled bags containing multiple instances. Using MIL, it is a good option to label slides, without need to do cell segmentation.

The key contribution of this MIL is the classification of tumour and normal samples with predictions consistent with genomic purity values. It does improve objectivity and reliability and produce a spatially resolved tumour purity map. The model can expand for better understanding of the tumour microenvironment.