Report

Abstract:

A newly developed machine learning model (MIL) was used to predict the tumor purity of slices from eight different TCGA coqueues. The prediction was successful because it was highly consistent with the accepted gold standard of genomic tumor purity values inferred from genomic data. The application of this model not only provides a spatial map of tumor purity, which can play a key role in understanding the tumor microenvironment, but also improves accuracy, avoids selective errors in human observations by pathologists, and is more cost-effective than reading by pathologists.

Tumor purity is defined as the percentage of cancer cells in a tumor. There are two methods used to estimate tumor purity: regional tumor nuclear percentage and genomic purity percentage. The nuclear percentage estimation in the first method is usually interpreted by sample selection and molecular analysis results, which is tedious and time-consuming and has errors among different observers. The second approach has the advantage of mitigating the confounding effects of normal cell contamination and producing consistent values across different cancer datasets. But the drawback is that it does not apply to samples with low tumor content and cannot provide information about cancer cells. Both methods have advantages and disadvantages.

The machine learning model was used to infer the causes of errors between different observers, and it was found that there were significant differences in tumor purity between the upper and download slides of a sample, indicating that there were spatial differences in tumor purity within the sample. The pathological suspension selected the area with high tumor content to estimate the percentage of tumor nuclei, which may be one of the reasons for the high data.

Interpretation of MIL model:

The linear model was established with gene purity values as dependent variables and 8 queues as independent variables. For example, you can enter code in R

Fit1<-lm (purity values ~ BRCA+GBM+LGG+LUAD+LUSC+OV+PRAD+UCEC, data=dataset)

Summary (Fit1)

Plot(Fit1)

By comparing the p-value predicted and estimated by the MIL model and pathologist respectively, it was found that the P-value obtained by the MIL model was <0.05, which is a strong evidence about the association. Pathologist estimated p-value>>0.05, there is no evidence to show the correlation. At the same time, variance was calculated and Wilcoxon was used to test the absolute error value, which showed that MIL model predicted better results.

The tumor purity of the top and bottom slides of a sample is tested differently

Use Wilcoxon signed-rank test. Define Z(i)=X(i)-Y(i),omit all observations with Zi=0,denote the remaining differences by Z1, Z2,... . Zinc. Let R1, R2,... Rn denote the ranks of | | the Z1 notification,... ,|Zn|. Let ai=1 if Zi>0, Ai =0 if Zi<0. W+ is the sum of the ranks of the positive Zis. W- be the sum of the negative Zis. W-). Then we reject H0 iff W<K.

We found the P-value which shows that there is a pure difference in tumors between top and bottom slides, so using two slides (top and bottom) is better than using only one slide to predict samples. The pathologist can also reduce the error by increasing the selection area.

At last, I evaluate the MIL model, sample MIL model successfully divided into tumor and normal, can be used to molecular sample selection, stratification of patients is superior to the correlation of tumor cell nucleus percentage and lower mean absolute error and weak tumor label natural purity need MIL method, there is also a limit, however, such as weak label value uncertainty, The deficiency of data set and the limitation of DNA morphology

And for the results on the MNIST graph,