**Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for Non-small cell lung cancer (NSCLC)**

**Abstract**

**Background**: DNA methylation has been implicated as the promising biomarker for precise cancer diagnosis. However, limited DNA-methylation based biomarkers have been found in Non-small cell lung cancer (NSCLC).

**Results**: In our study, high-throughput DNA methylation dataset (320 samples) of NSCLC from the cancer genome atlas (TCGA) project was analyzed and validated in another independent dataset (64 samples) from gene expression omnibus (GEO) database. The methylation status from PBMC cells from healthy controls was also utilized for biomarker selection. Based on above procedures, five candidate CpGsites as well as its adjacent regions were further validated in another **120** pairs of ESCC tumor tissues and adjacent normal tissues as well **30** pairs of cell-free DNA from NSCLC and normal individuals in Chinese Han population using targeted bisulfite sequencing method. Logistic regression model was applied to the methylation status of genomic regions covering the five candidate CpGsites, yielding a robust performance (Sensitivity = 0.75, Specificity=0.88, AUC=0.85). Eight statistical models along with fivefold cross-validation were also applied, in which the SVM model reached the best accuracy in both train and test dataset (Accuracy = 0.82 and 0.80, respectively). In addition, subgroup analyzes revealed strong difference of diagnostic performance in the alcohol use and non-alcohol use subgroups.

**Conclusions**: In summary, based on the high-throughput DNA methylation dataset for biomarker screening, we identified five candidate CpGsites, further validated these CpGsites as well as their nearby regions with another independent ESCC samples with targeted bisulfite sequencing. Methylation profiles of the five genomic regions covering cg05249644 (STK3), cg15830431, cg20655070, cg26671652 (ZNF418) and cg27062795 (ZNF542) would be effective methylation-based testing for ESCC diagnosis.

Keywords: Non-small cell lung cancer, DNA methylation Biomarker Diagnosis Targeted bisulfite sequencing

Method and Materials:

1, 120 pairs of NSCLC and adjacent normal tissues.

2, 30 pairs of cell-free DNA from NSCLC and normal individuals.

3. I will apply multiple machine learning method to select the most significant differential methylation regions between NSCLC and normal lung

4. Apply Target methylation sequencing to generate methylation profiles for these 50 regions in 150 pairs samples.

5, I will build prediction model for NSCLC and provide accuracy based on solid tissue and cell-free DNA methylation.

6, Meanwhile, we will do subgroup analysis to identify correlation between DNA methylation and all clinical information, such as TNM and so on.

7. I hope these samples have overall survival time so that we can do survival analysis between DNA methylation and overall survival time based on cox-regression.

8, Collect CAM5.0, CK5/6, Nap-A, P53, PD-1, PDL-1

9, collect tumor purity by HE staining.