**Stratification of thyroid tissues based on machine learning**

In order to stratify the subtype of thyroid diseases and to identify the signature proteins of each subtype, we developed 10 models based on random forest [1], which is an ensemble machine learning method and frequently used to resolve problems with high dimension and low sample size. Each of the models corresponds to one of the 10 pair-wise combinations between the five subtypes (N, M, A, C and P). Before training the models, we averaged the expression abundance of each protein (logarithmized based on 2) over the three replicas of each patient and filtered out those with more than 30% missing values over all the samples and remaining missing values are replaced by zero. After data preprocessing, we run random forest iteratively to select those discriminative proteins as input features of each model. In each iteration, we added a protein, which is not used in all the previous iterations; if this added one can increase the AUC (Area under Curve) value of the ROC (receiver operator curve) of the training model, we retained it as one of protein set to be used in the final model. In the first iteration, we run random forest with all the proteins as input features and selected the one with most importance measurement returned by the trained model. The detail of workflow was shown in Fig.wzc01. These discriminative proteins of each pair of subtypes can reference in supplement table-wzc01 Based on these discriminative proteins, we trained the ten random models and plot their ROC to assessment their prediction performance, showed in Fig.###.

The validation sample is fed into the 10 trained models, each of models votes on the subtype of the sample and the one with most votes is assigned to the sample. The predictions of all 180 samples are listed in supplement-table####

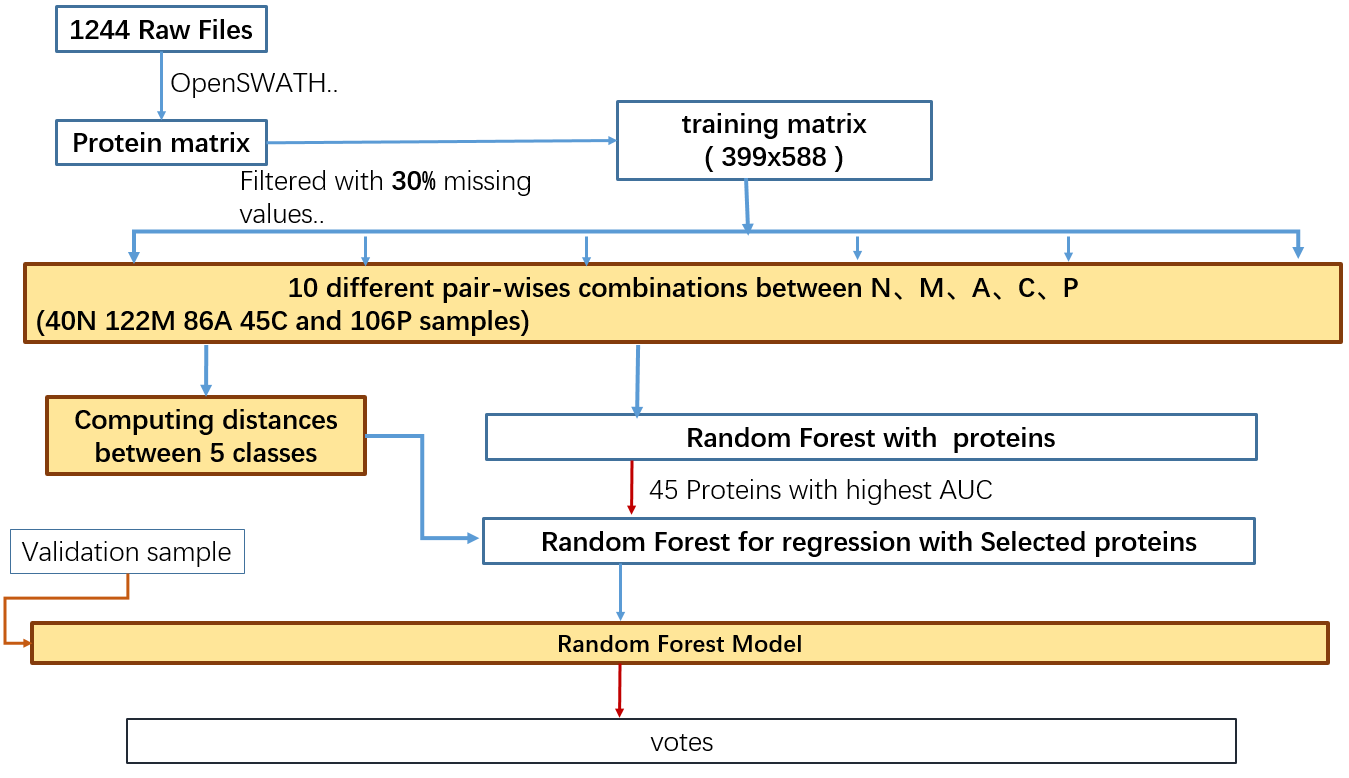


Fig.wzc01 workflow of train random forest models

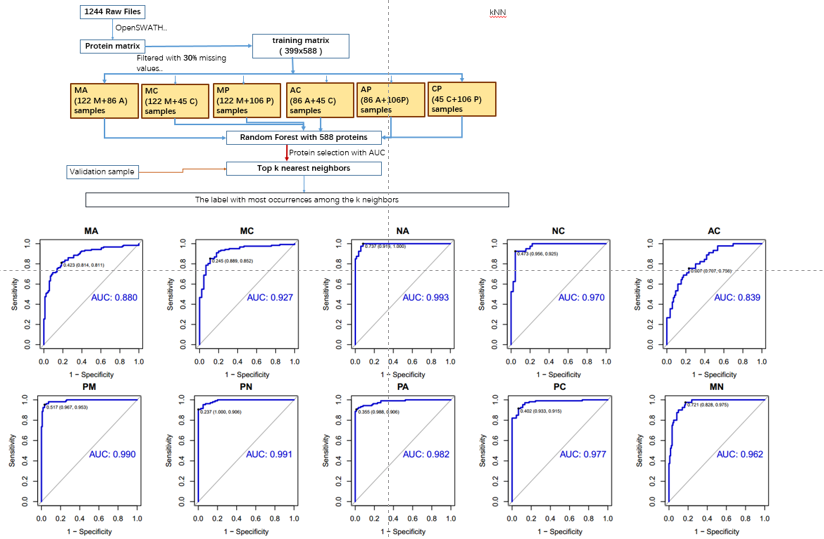


Fig.### ROC plot for 10 random models

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.