目录

[Supplementary Figure 1 2](#_Toc532932874)

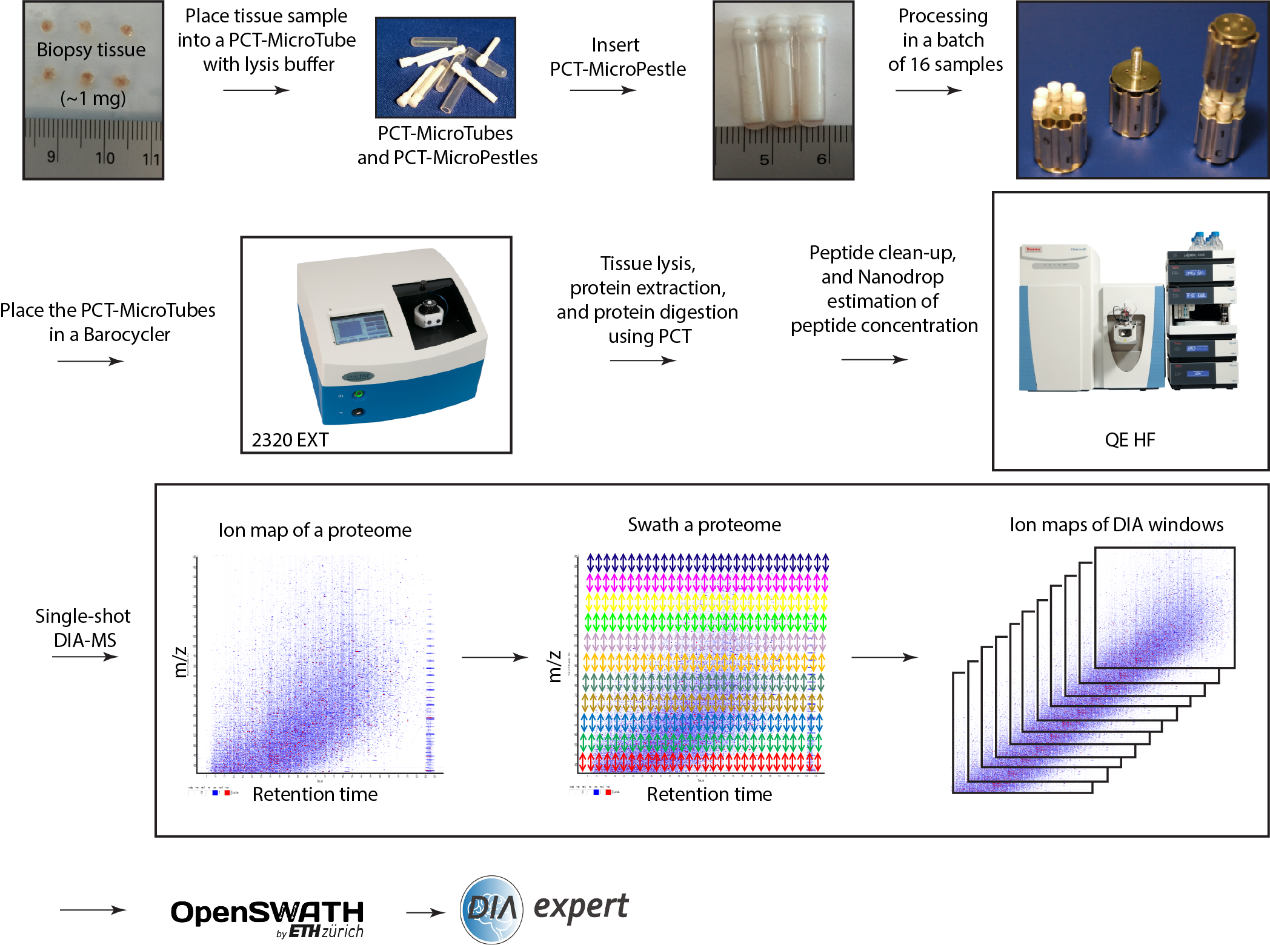
[Supplementary Figure 2 3](#_Toc532932875)

[Supplementary Figure 4 5](#_Toc532932876)

[Supplementary Figure 4 6](#_Toc532932877)

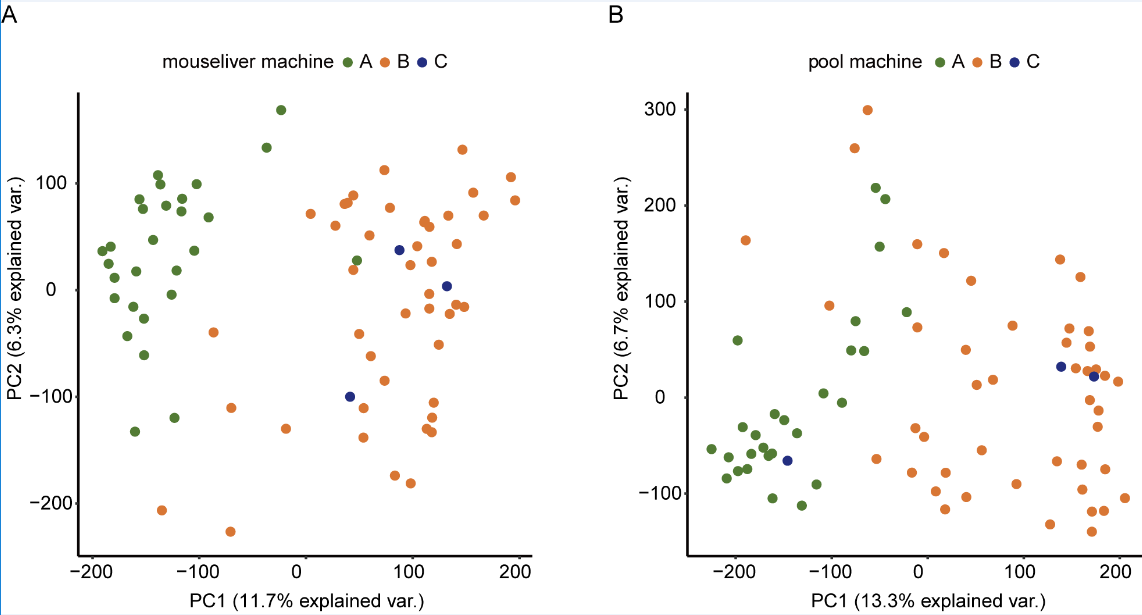
[Supplementary Figure 5 8](#_Toc532932878)

## Supplementary Figure 1



**Supplementary Figure 1. PCT-DIA workflow.** Tissue samples are placed in PCT-MicroTubes and processed by a 2320 EXT barocycler in a batch of 16 samples for tissue lysis and digestion after de-waxing. After clean-up, the peptides are analyzed by QE-HF mass spectrometers in DIA mode. The resultant DIA data are analyzed by OpenSWATH and DIA-expert.

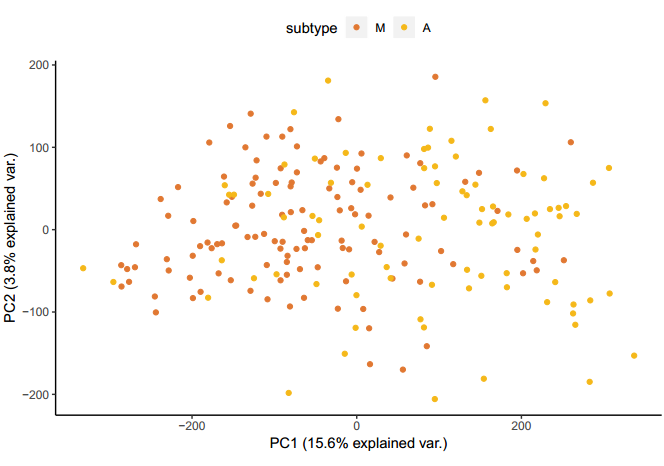
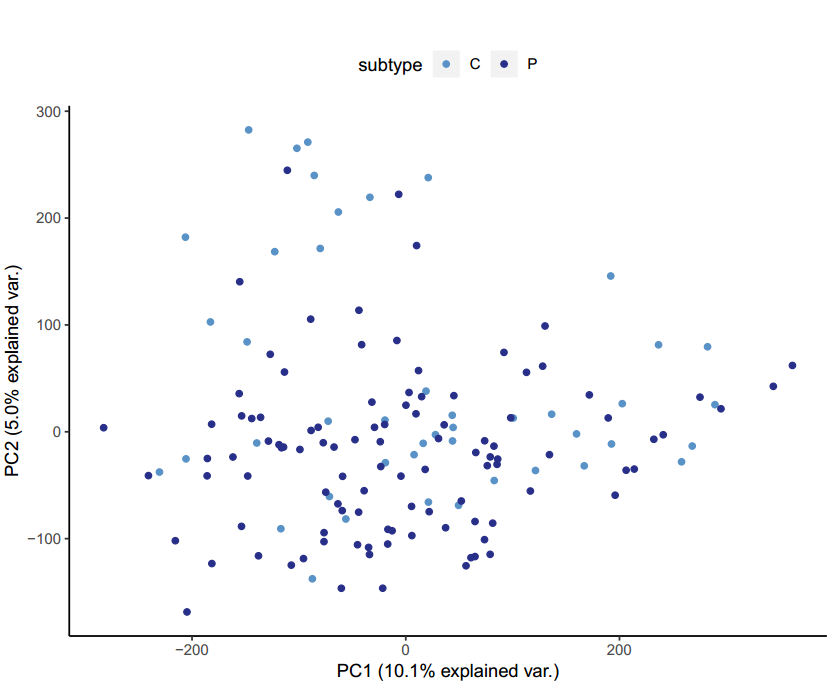
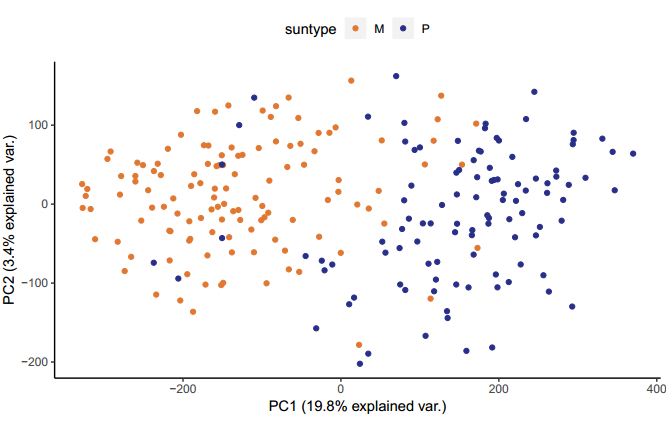
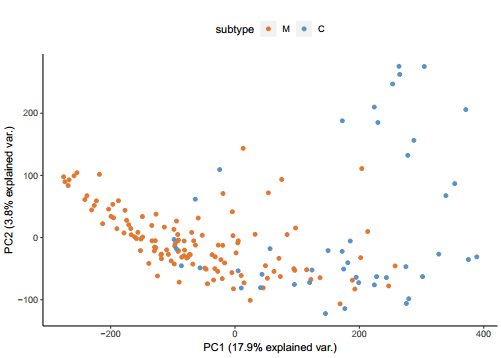
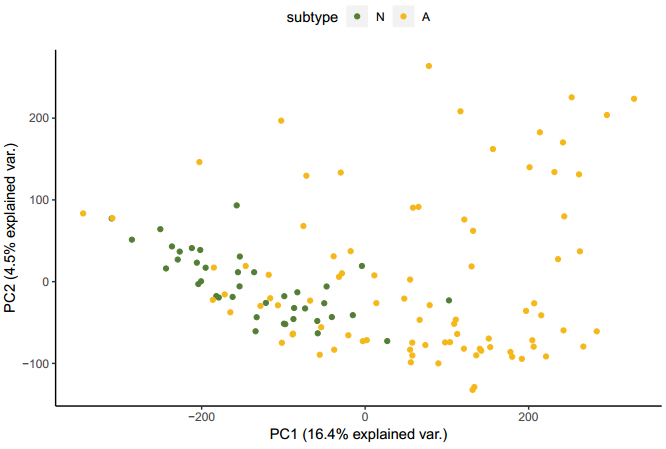
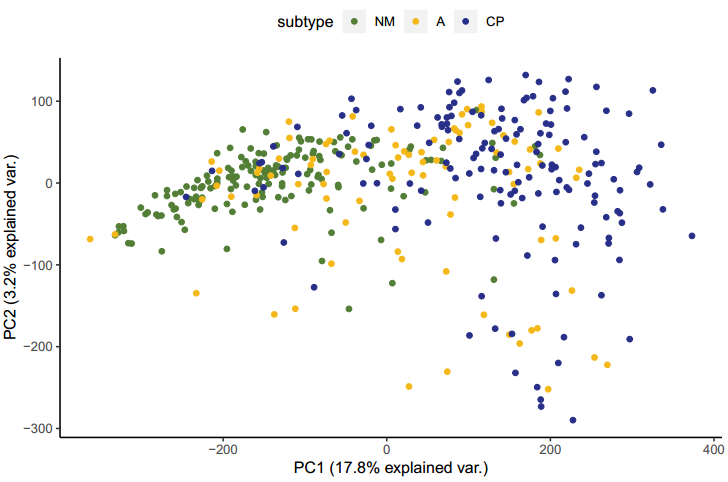
## Supplementary Figure 2



**Supplementary Figure 2. Analysis of data quality.** (A) PCA analysis of technical reproducibility of PCT by mouse liver samples. (B) PCA analysis of technical reproducibility of MS by pool samples. (C) Violin plot of technical reproducibility of PCT by mouse liver samples and pool samples.

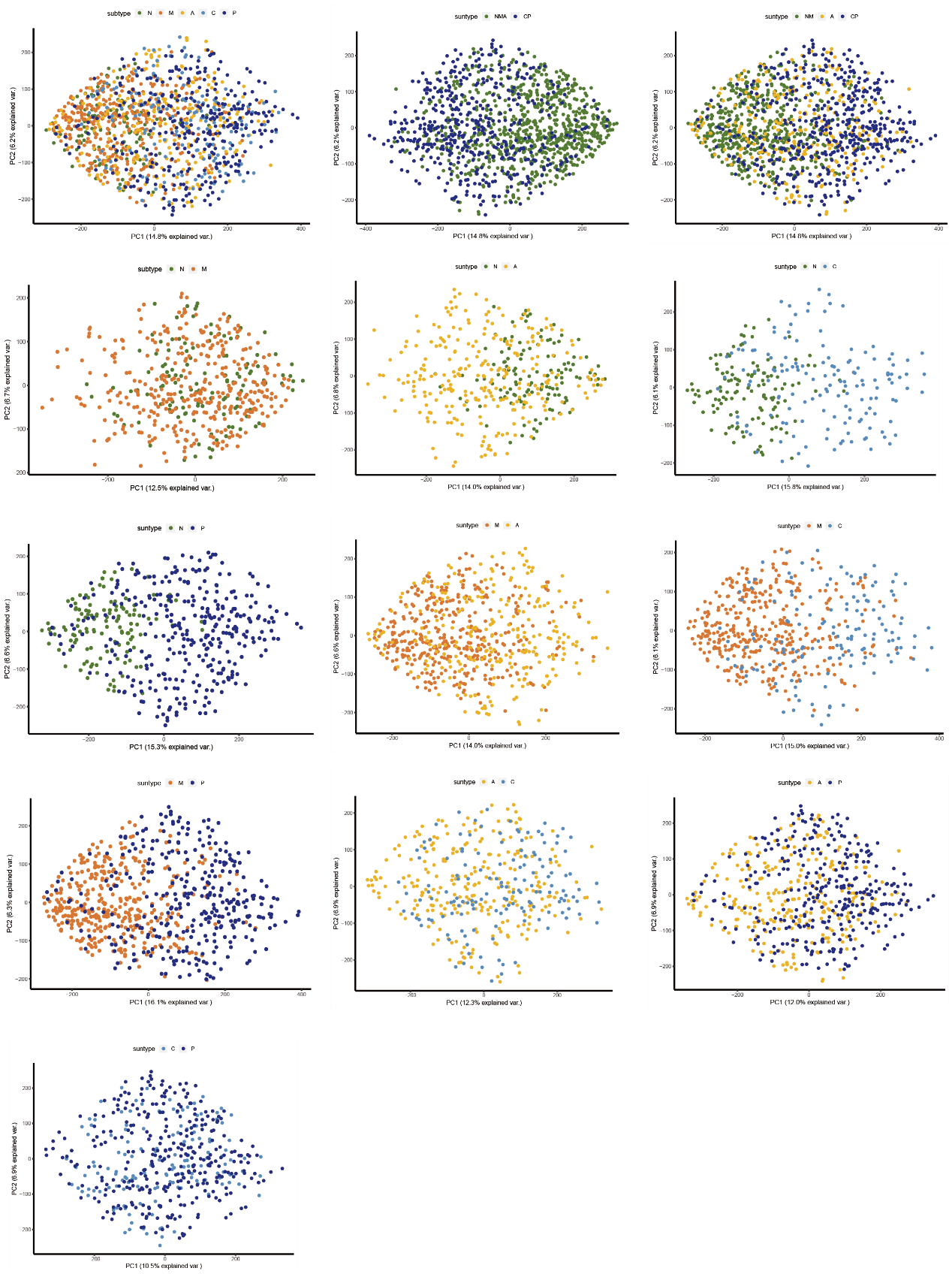
**Supplementary Figure 3 Reanalysis of data quality after correction.** (A) Violin plot of technical reproducibility of PCT by mouse liver samples and pool samples after data correction. (B) PCA analysis of batch effect of samples after data correction. (C) PCA analysis of technical reproducibility of PCT by mouse liver samples after data correction. (D) PCA analysis of technical reproducibility of MS by pool samples after data correction.

## Supplementary Figure 4



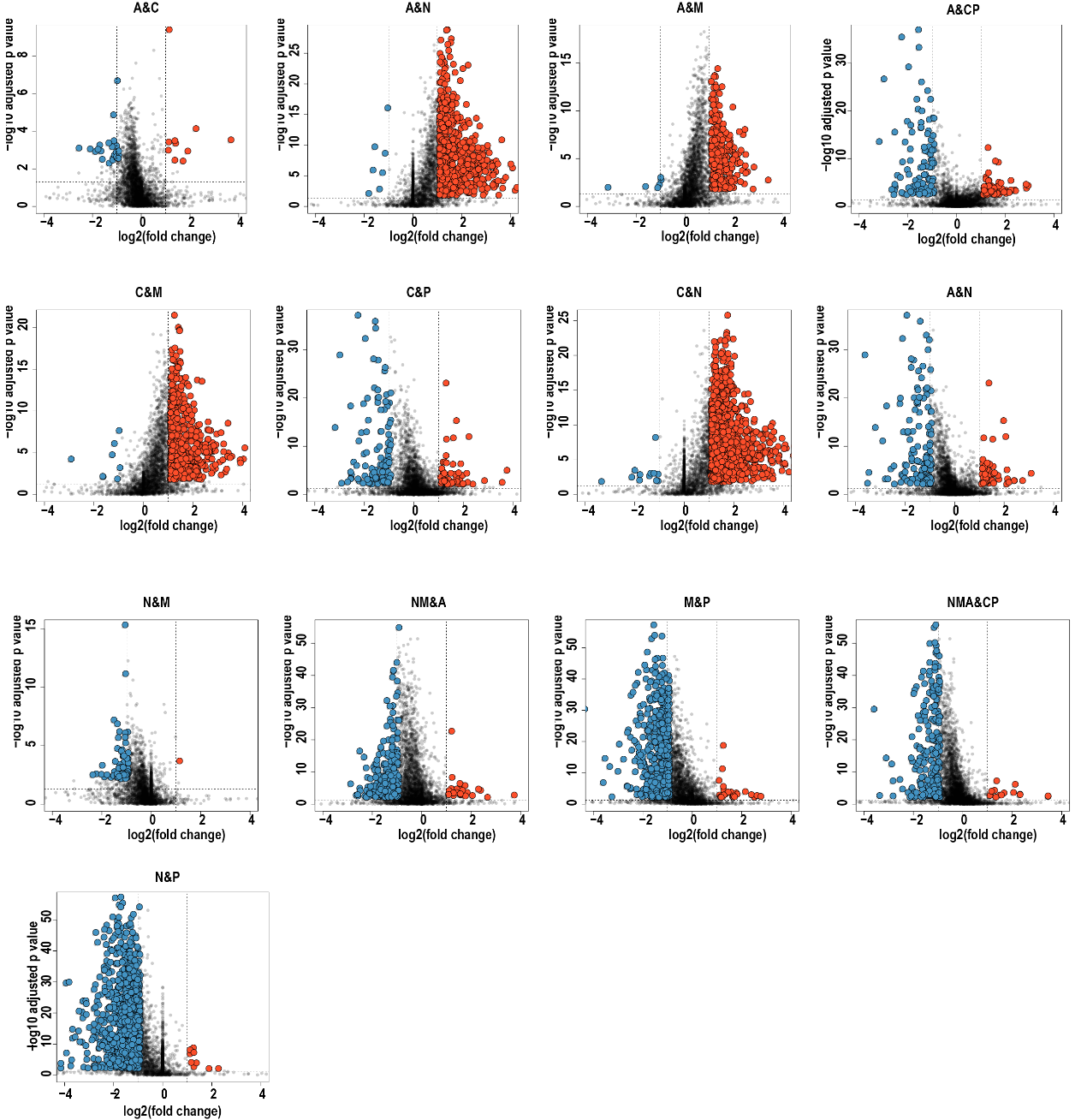
**sFigure 4. Remaining PCAs not in the main figure.** Pair-wise combinations between the five subtypes (N, M, A, C and P)were analyzed by PCA. Each patient has three biological replicates. The intensity of each protein for one patient were firstly calculated mean value and then performed PCA analysis.

## Supplementary Figure 5

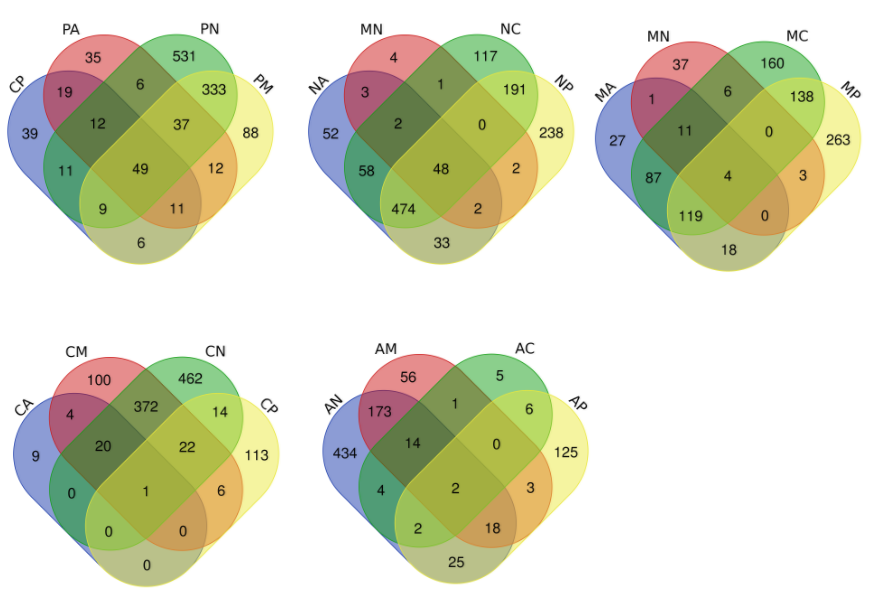


**Supplementary Figure 5. Pair-wise PCA of the training patients.** Pair-wise combinations between the five subtypes (N, M, A, C and P)were analyzed by PCA. Each point indicated one sample without average.

## Supplementary Figure 6

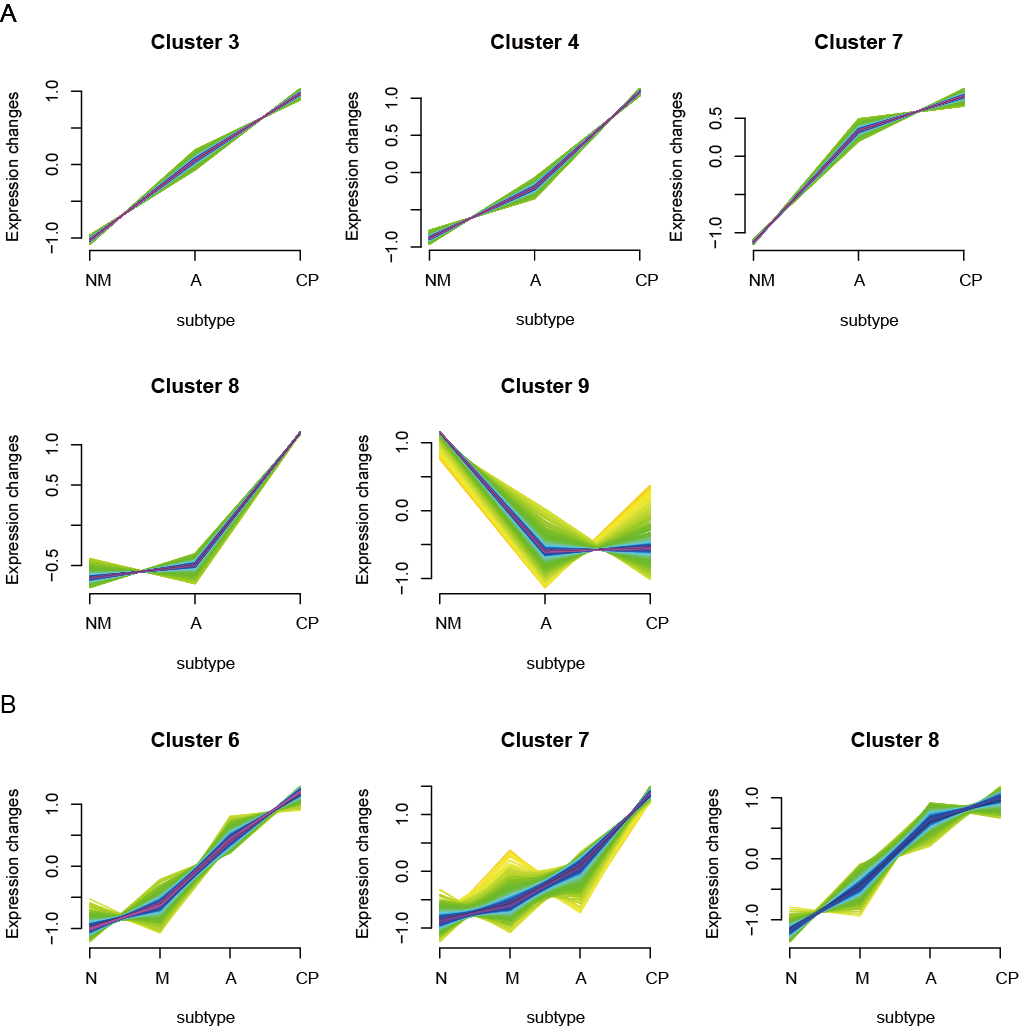
**Supplementary Figure 6. Pair-wise volcano plot of the training patients.** Volcano plot for ten pairs of protein expression comparation. N, M, A, C, P indicated that normal thyroid, multinodular goiter, follicular thyroid adenoma, follicular thyroid carcinoma and papillary thyroid carcinoma, respectively. We compared ten pair-wise group from five subtypes of thyroid disease with two-fold-change cutoff and adjust P-value threshold less than 0.05.

## Supplementary Figure 7



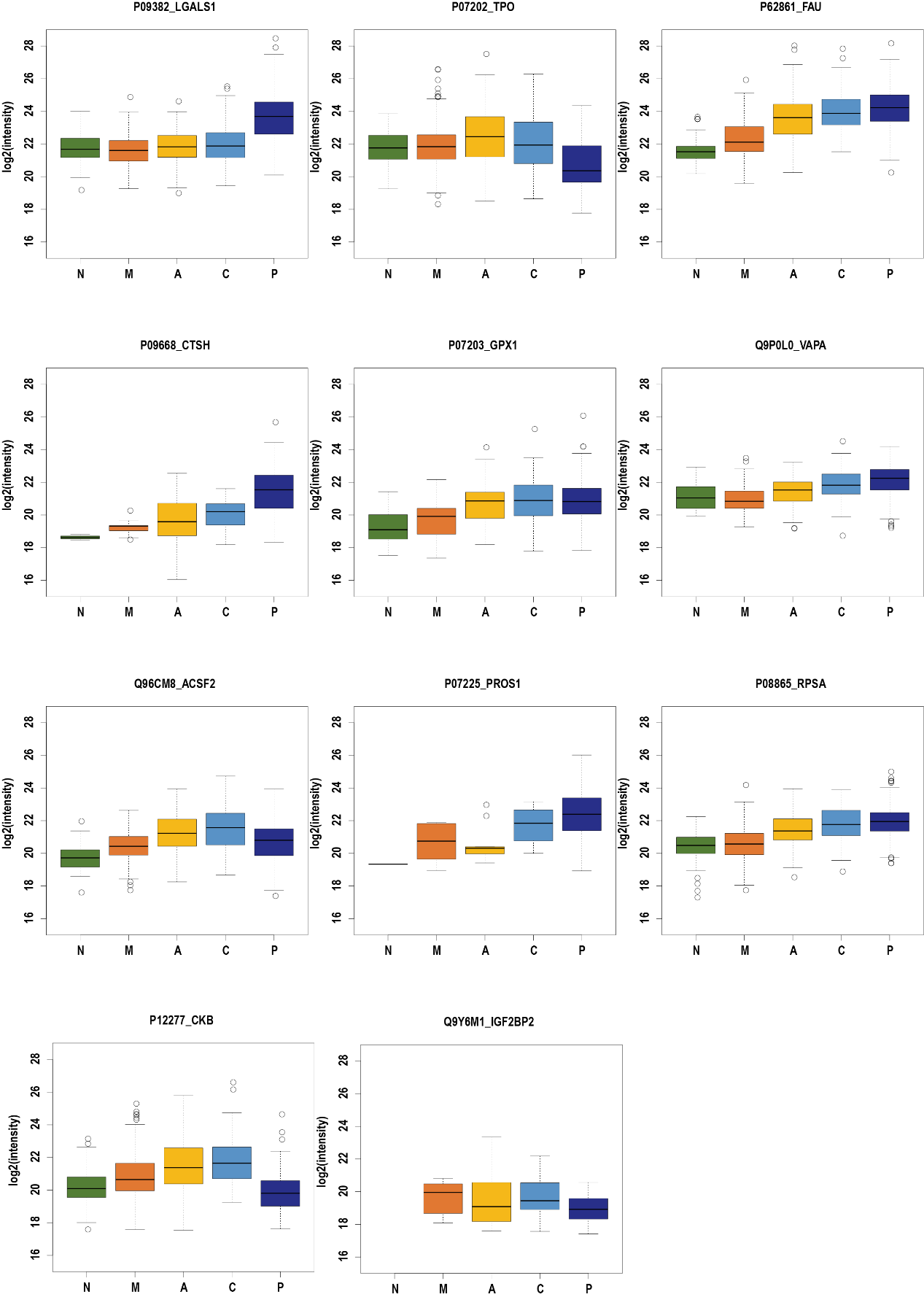
**Supplementary Figure 7. Venn plot of the training patients.** Each thyroid disease was compared with other four types using Venn diagram. The protein lists were from volcano plot.

## Supplementary Figure 8



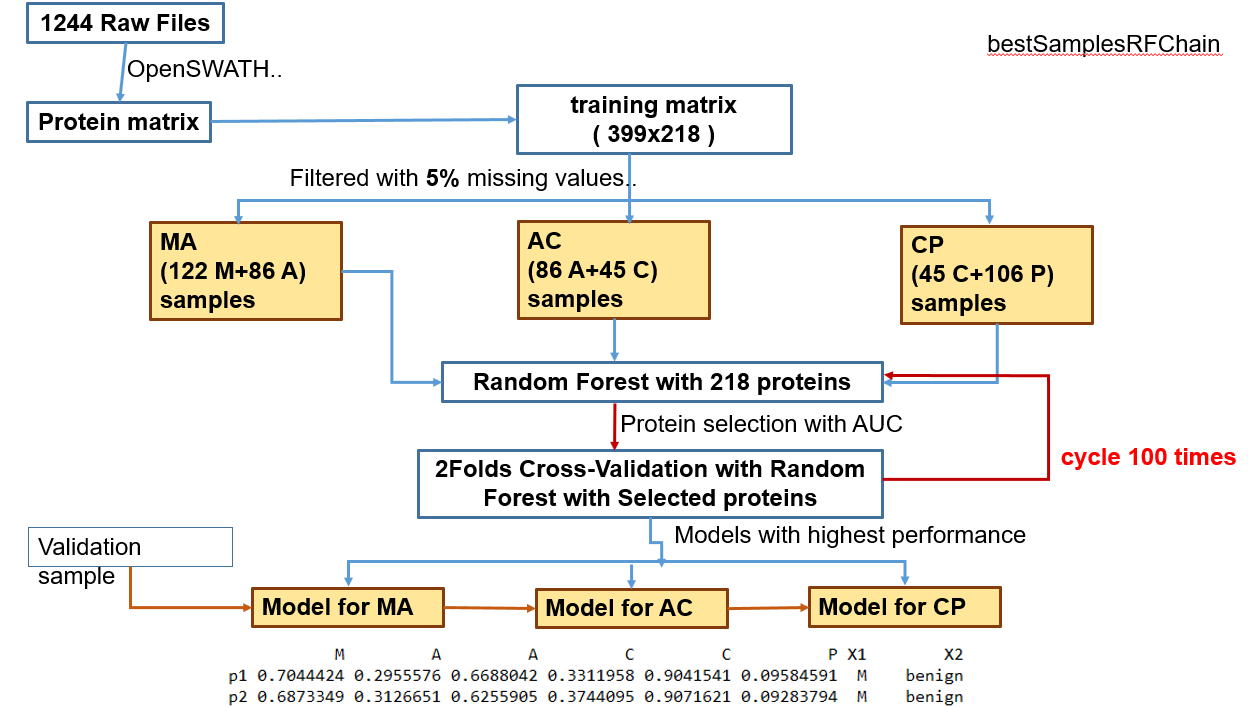
**Figure 8. Cluster analysis of protein associated with tumor progress.** All the identified proteins were allocated into 10 clusters according to their expression tendency. (A)N and M, C and P were combined into two groups and C regard as a transitional group. (B) N, M, A were regarded as isolated group and CP was considered as malignant group.

## Supplementary Figure 9

**Supplementary Figure 9. Expression of selected proteins in five patient groups.** Each diagram indicated the CV of intensity of one protein expression.

**Supplementary Note 1. more details on validation models and failed model testing etc (by Wu)**

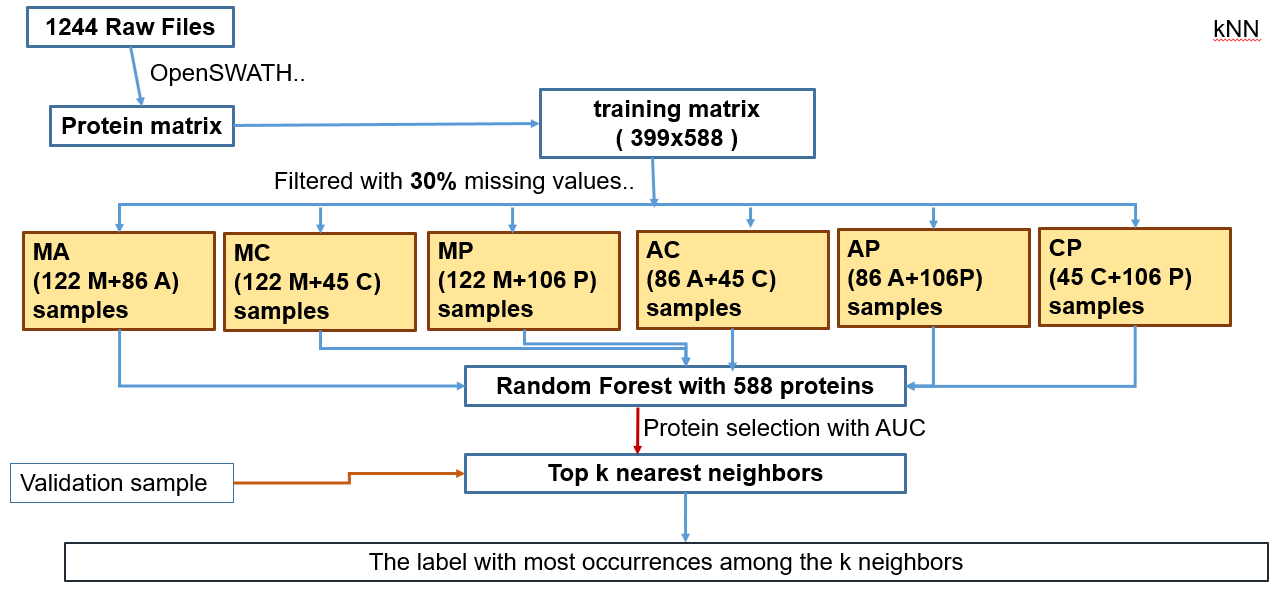
Best sample RF chain



After averaging the intensities over the three biological replicas of each patient, we filtered out all the proteins with a cutoff of more than 5% missing values presented in all the samples (patients), we get a training matrix with 399 rows and 218 columns, where each row represents a patient and each column a protein. Considering orders exist between the subtypes M, A, C and P according to their severity, we trained 3 random forest models respectively for combinations of subtypes M and A, A and C, C and P.

In detail, we first randomly divided the training dataset into 2 parts with same sample sizes, one is taken as training dataset and the other as validation dataset. For training dataset, we take all the 218 proteins as the input features to train random forest, which return the importance of each protein; then beginning with the most important one, we re-trained a new random forest model, each with a new protein added into the existed proteins, if the added protein increased the AUC value, it is retained as one of key proteins. When the AUC values stop increasing, we get all the key proteins for the combination of subtypes. For each of the 3 combinations， we trained a final random forest model and use this model to predict the validation dataset, finally we exchange validation dataset and training dataset and redo these procedures.

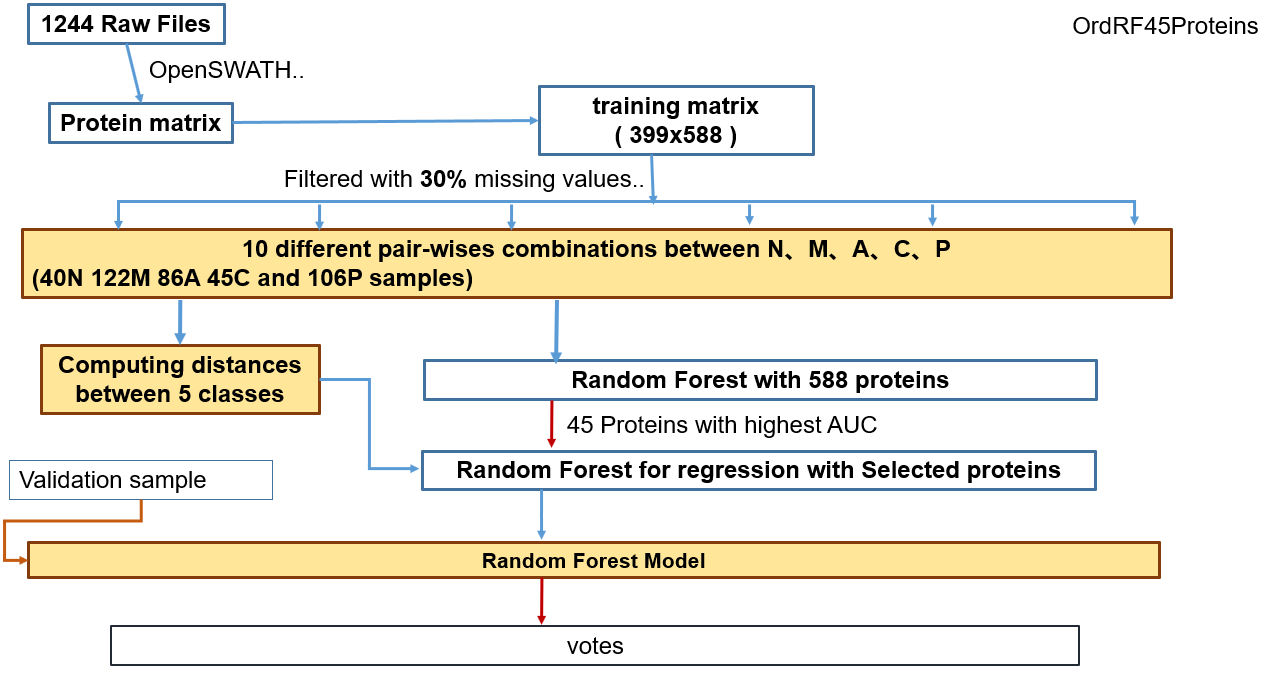
We repeat this procedure 100 times and retained the best model.



Among these models, kNN model was performed best (sTable6, need to be replenish, wu and Gopal).

After averaging the intensities over the three biological replicas of each patient, we filtered out all the proteins with a cutoff of more than 30% missing values presented in all the samples (patients), we got a training matrix with 399 rows of patients and 588 columns of proteins (sTable 7, provided by wu20181214). For each of the 6 combinations between the 4 subtypes (M, A, C and P), we first trained a random forest model with all the 588 proteins as input features. The returned model provides us with the importance of each protein. After ordering all the proteins by their importance, we trained a serial of random forests to select key proteins. Beginning with the most important one, each of training forest take a new protein into existing proteins as input features, if the added protein could increase the AUC (Area Under Curve) value of its ROC, we retained it as one of key proteins (Figure 3A).

Collecting all the key proteins from the 6 combinations, we get 71 key proteins (sTable 8, provided by wu20181214) from ten ROC (Figure 3B, by wu), with which we apply k-nearest-neighbor algorithm to identify the subtype of a new sample. In detail, the Euclidean distances between the new sample and all the training samples are computed, and the k nearest neighbor were identified, then these neighbors votes the subtype, the one with most votes are assigned to the new sample.



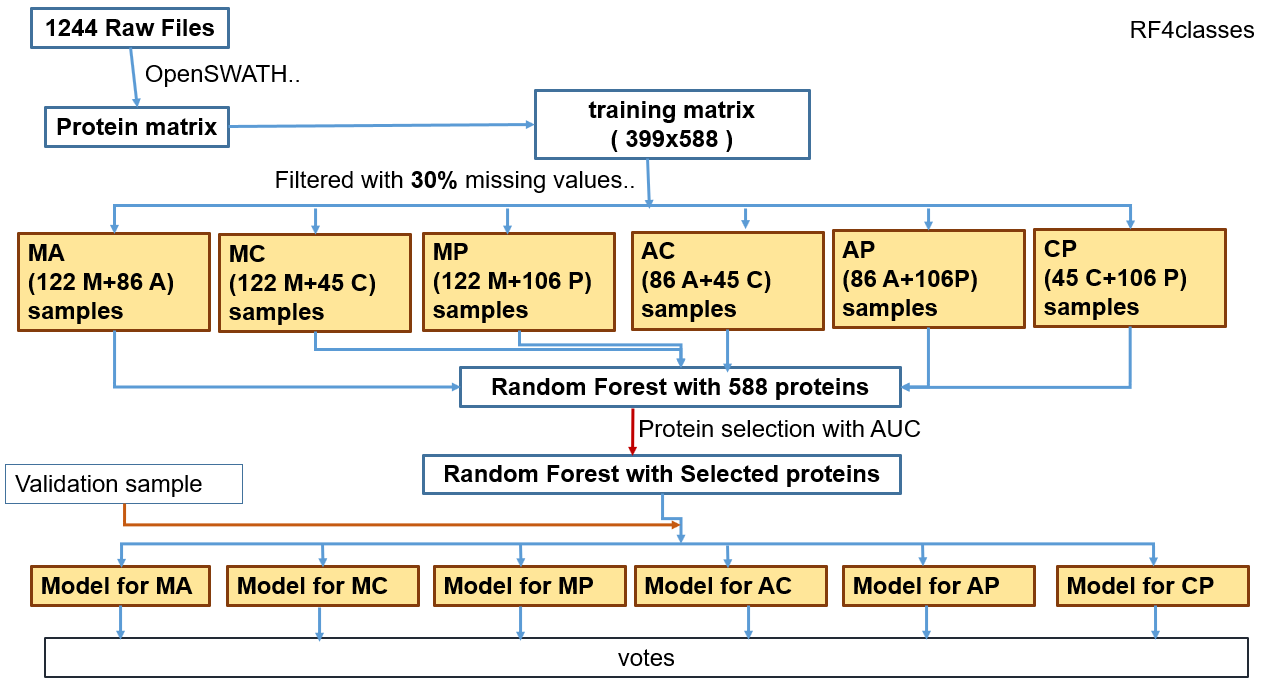
Ord RF 45 proteins

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).

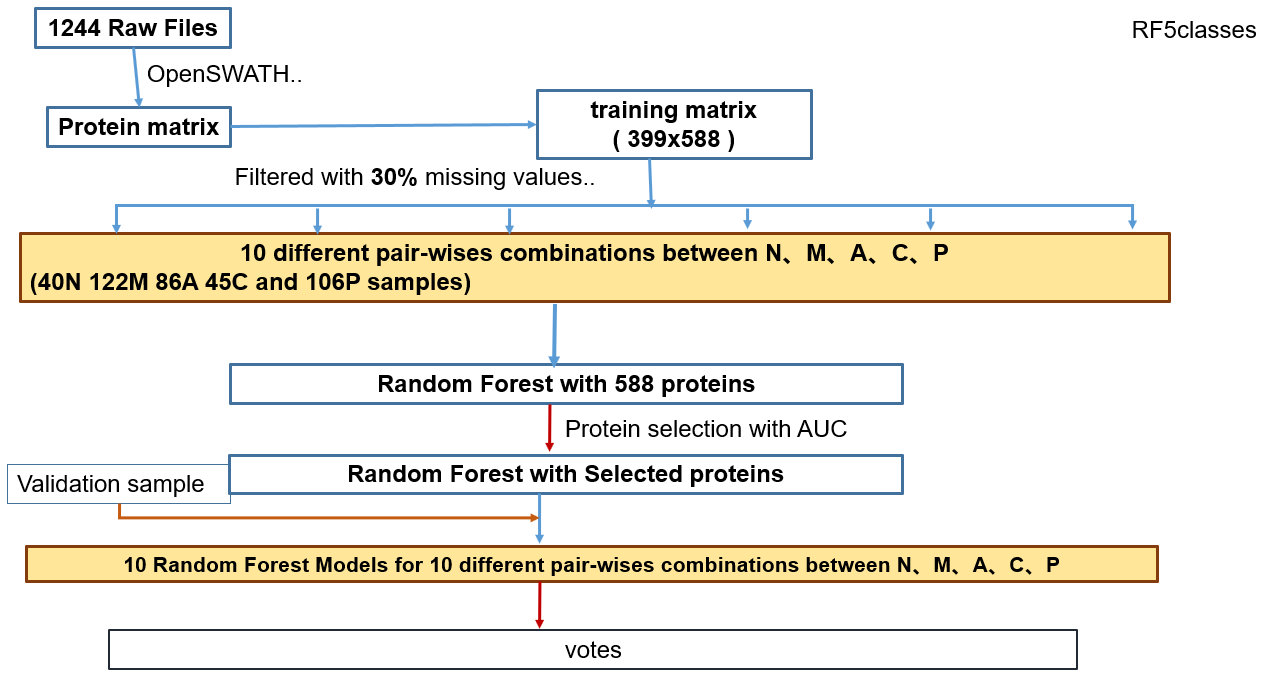
Before training the models, we averaged the expression abundance of each protein (logarithmized based on 2) over the three replicas of each patient and filtered out those with more than 30% missing values over all the samples and remaining missing values are replaced by zero. After data preprocessing, 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####

RF 4 classes

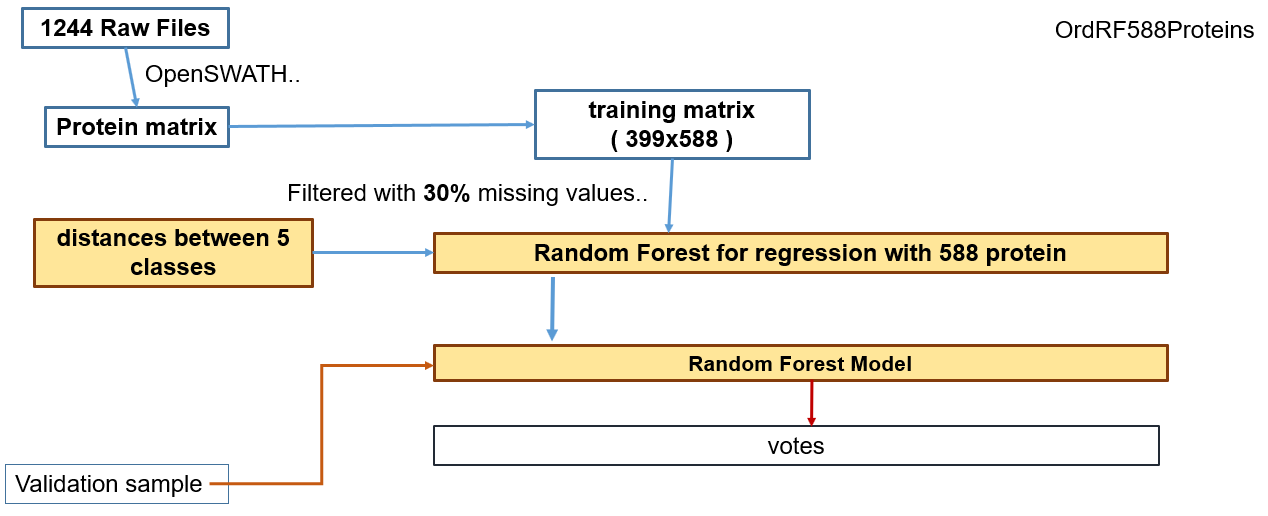


RF 5 classes



In order to assess the impact of normal subtype, we trained random models with and without normal tissue included respectively. In both cases, proteins with more than 30% missing rate are filtered out, this results in 588 proteins retained to train the models. Six random forest models for the 6 pairs of combinations of the 4 subtypes (M, A, C, P) and ten random forest models for 10 pairs of combinations of the 5 subtypes (N, M, A, C, P) were trained. The trained models voted for the validation sample, which is assign the subtype with most votes.

Ord RF 588 proteins



OrdRF588Proeins

After filtering out proteins with more than 30% missing value, we get 588 proteins to train the random models. In order to importing the ordinal information between the 5 subtypes (N, M, A, C and P), we compute the average Euclidean distances between the 5 subtypes and transform the problem from classification to regression. The trained 10 model predict a validation sample by voting, the one with most votes is the predicted subtype