**Exploring protein signature for stratifying thyroid diseases using PCT-DIA**

**Authors**

Yaoting Sun1, Sathiyamoorthy Selvarajan2, Yi Zhu1, Zhicheng Wu1, Wei Liu1, Tiansheng Zhu1, Qiushi Zhang1, Huanhuan Gao1, Xue Cai1, Guan Ruan1, Jian Song, Chenhuan Yu, Changbin Yu, Huazhong Ying, Kexin Liu, Kon Oi Lian2\*, N. Gopalakrishna Iyer2\*, Tiannan Guo1\*

1School of Life Sciences, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China.

2Institute of Biology, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China.

3Singapore General Hospital, Singapore

\*Corresponding author

**Abstract**

Thyroid nodules are common endocrine disease affecting approximately 50% of the population globally, particularly in women with increasing age. Fortunately, only 7-15% of them are malignant. Currently there are several genetic tests based on mRNA or DNA measurement, which can facilitate the diagnosis of cancerous thyroid nodules. However, there are still about 15% to 30% of thyroid nodules that cannot be clearly evaluated to be pulmonary indeterminate nodules or not before operative cytologic diagnosis. Therefore, patients suffering from inconclusive nodules are often advised with surgical resection, though most of which turned out to be benign after pathological evaluation, leading to a broad range of overtreatment. In this study, we aimed to develop a rapid and robust PCT-DIA/MS methodology to characterize the benign and malignant thyroid tissues by large scale quantitative proteomics profiling. Formalin fixed, paraffin embedded (FFPE) tissue punches from 580 patients with thyroid nodules are included for the study. A panel of more than ### promising biomarker candidates were found with clinical implication to distinguish benign and malignant thyroid nodules from our study

**Introduction**

Thyroid nodules are common endocrine disease affecting approximately 50% of the population especially in women with increasing age. Fortunately, only 7-15% of them are malignant1-3. The initial workup for thyroid nodules includes a complete history and physical examination, serum measurement, imaging test, fine needle aspiration (FNA) and cytopathology. Among these tests, FNA is considered as the most sensitive, reliable and cost-effective detection for clinical management of patient with suspicious thyroid nodules1,4. Nevertheless, about 15% to 30% of thyroid nodules cannot be clearly evaluated by FNA and cytopathology5. Patients suffering from inconclusive nodules are often advised to surgical resection, though most of them are proved to have benign disease, leading to overdiagnosis and overtreatment. As a result, the incidence of thyroid cancer has been dramatically increased globally, whereas the mortality remains stable 6-8.

In order to minimize overdiagnosis and overtreatment, molecular testing for thyroid nodules has been recommended as adjunct to FNA9-11. There have been several genetic tests based on mRNA or DNA detection to increase the accuracy of preoperative cytologic diagnosis for thyroid nodules. For example, Afirma test is a gene expression classifier (GEC) by assessing expression of 142 genes, which receives a high negative predictive value (NPV) up to 95%, however the positive predictive value (PPV) is only 38%5. Alternatively, ThyroSeq is a next-generation sequencing assay with higher PPV and NPV12. Although the current molecular testing (based on mRNA or DNA detection) offer improved management of thyroid nodules, they are not sensitive or specific enough to support diagnosis of indeterminate lesions in all cases since not all thyroid cancer contains mutation. For this aspect of view, it is of urgent necessity to develop new diagnostic strategy to stratify thyroid disease, especially those undetermined thyroid lesions. Proteins are the direct participants of biochemical reactions and are potentially more specific indicators for thyroid diseases. Therefore, in the current study, we aim to characterize the benign and cancerous thyroid tissues by quantitative proteomics profiling.

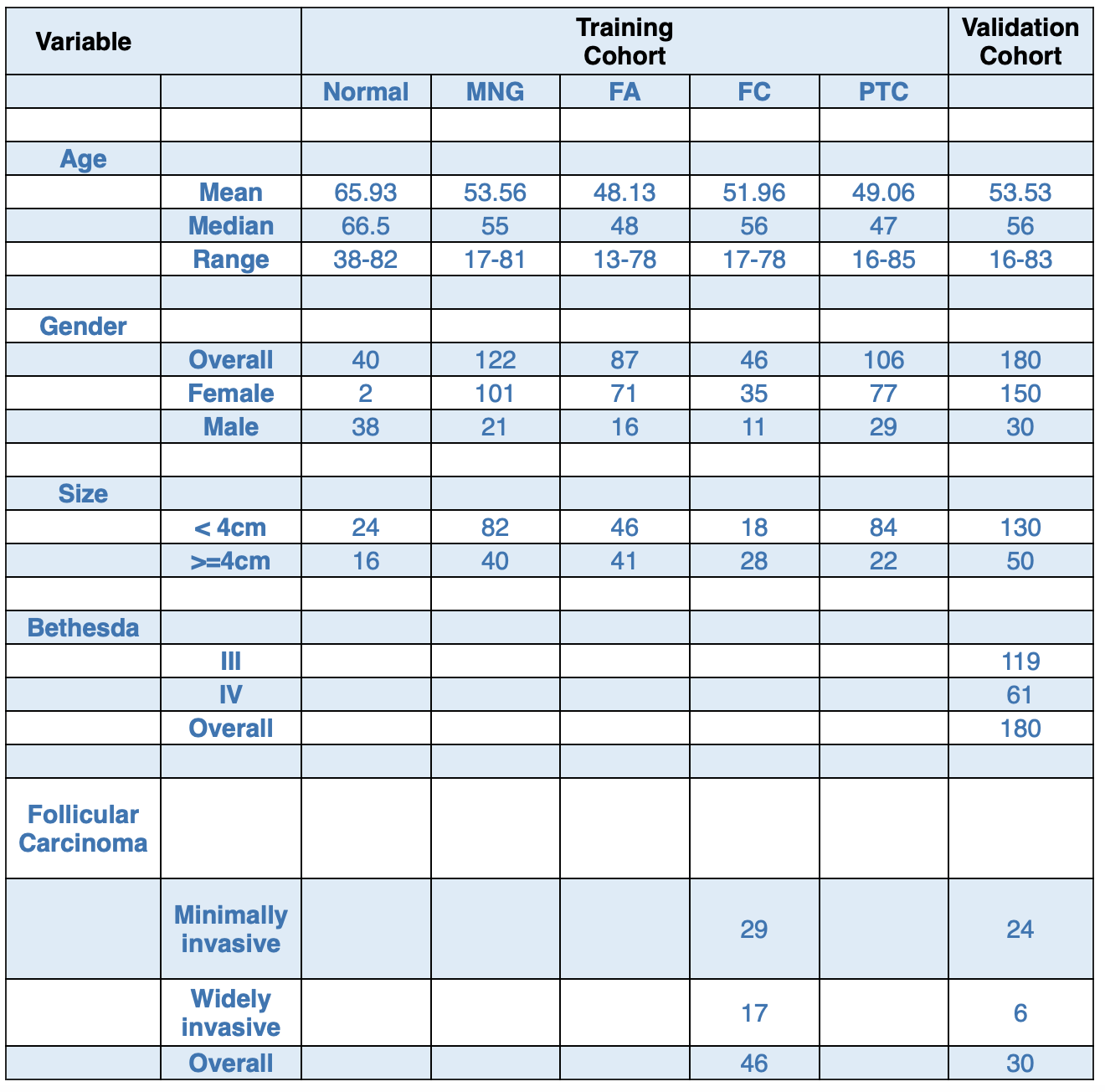
We have set up a workflow to deal with formalin-fixed and paraffin-embedded (FFPE) tissues by pressure cycling technology (PCT) coupled with SWATH-MS13. PCT is an emerging system to handle minimal tissue biopsy samples (0.2 to 1 mg) at oscillated cycles between ultra-high (45,000 p.s.i) and ambient pressure, under which scheme the protein lysis and enzymatic digestion could be greatly accelerated14,15. SWATH-MS is a data independent acquisition (DIA) mass spectrometry (MS) methodology developed on a SCIEX TripleTOF system. DIA strategy offers high sample throughput and highly accurate analysis of complex proteomes16.

In this study, we further implemented the DIA-MS methodology on a Thermo Q Exactive HF system. The integrated FFPE-PCT-DIA workflow thus was applied to stratify thyroid diseases. Two cohorts were performed for finding and validating signatures of indeterminate thyroid nodules. There were 1260 FFPE samples from 400 patients in the training cohort and 543 testing samples in the validation set.

**Results**

**Study design**

In this study, we procured two cohorts of patients with thyroid diseases (**Table 1, Gopal, Deadline. 20181214**). The first cohort comprises 399 patients as the training set, including normal thyroid (N, n=40), multinodular goiter (M, n=122), follicular adenoma (A, n=86), follicular thyroid carcinoma (C, n=45) and papillary thyroid carcinoma (P, n=106). The validation set containing 180 patients was a single-blind cohort, only the pathologists but not the proteomics and data analysis group knew the sample identity throughput the project.More details of these cohorts are provided in **Supplementary Table 1(validation set no patient info, ask for Gopal，Deadline181203).**

**Table 1. Clinical characteristics of the cohorts**

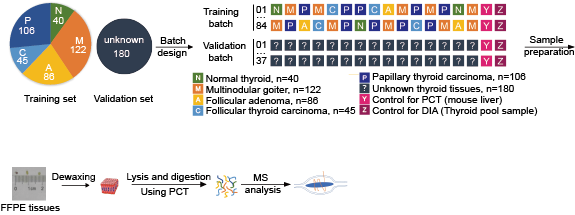
From each of these patients, three punches from the respective FFPE tissue region, regarded as biological replicates, were subjected to PCT-DIA analysis after batch design and quality control (**Fig. 1, weigang, 20181215**). From the training cohort, totally we collected 1187 punches and 54 technical replicates from 399 patients were randomly divided into 84 batches while in the validation cohort 543 samples from 180 patients were divided into 37 batches. Each batch comprised 15 thyroid tissues, one mouse liver sample for PCT quality control and one pooled thyroid tissue digest sample for MS quality control.

Each batch of proteome digests was analyzed in DIA-MS mode using QE-HF with gradient of 45 min and 24 windows. To avoid carry over we inserted a blank sample using 15 min gradient every four injections (**Supplementary Note 1, huanhuan, 20181214**).

To build the spectral library for analyzing DIA files from thyroid tissue samples, we collected tissue samples from 65 patients from the five patient groups. The tissue samples were either fractionated into ten or two fractions using SDS-PAGE (96 DDA files) or processed with PCT-assisted lysis and in-solution digestion (22 DDA files), or PCT-assisted lysis and PCT-assisted digestion (86 DDA files) (**Supplementary Table 2**). Totally we acquired 204 DDA files on a QE-HF mass spectrometer in DDA mode using a gradient of 2 hrs. Then we analyzed DDA files by MaxQuant version1.6.2 with SwissProt fasta files (20269 protein sequences), and identified 58,057 transition groups, 40,086 peptides, 5,287 protein groups, 4,129 proteotypic proteins. Then we retrieved proteins from this library and extract a sub library from a pan-human library which were built from 14 tissue types including these thyroid samples (manuscript in preparation) using these proteins. The final thyroid library contains 297,675 peptide precursors, 7,506 protein groups, 4,838 proteotypic proteins (**Supplementary File 1-3, library files TraML,csv,sptxt, tiansheng,20181214**). [Add a comment here, how these numbers compare to other papers of thyroid proteome. Add some reference here, and show that our library is good enough.]

To analyze these 1,244 DIA files using OpenSWATH [cite openswath paper], we generated conserved peptide sequences for retention time calibration [cite the CiRT paper, please ask Tiansheng] (**Supplementary Note2, tiansheng**), and identified 43,635 peptide, 3,711 proteotypic proteins, and 4,785 protein groups in the final protein matrix (**Supplementary Table 3, LiuW, sheet1 peptide matrix, sheet2 protein matrix**).

**Figure 1**



**Figure 1. The study design and workflow.** Two cohorts of patients with thyroid diseases were procured. From a specific tissue region from each patient we collected three punches of samples. PCT-DIA analysis was performed for these tissue samples in batches containing randomized samples. 57 samples were analyzed in technical replicates.

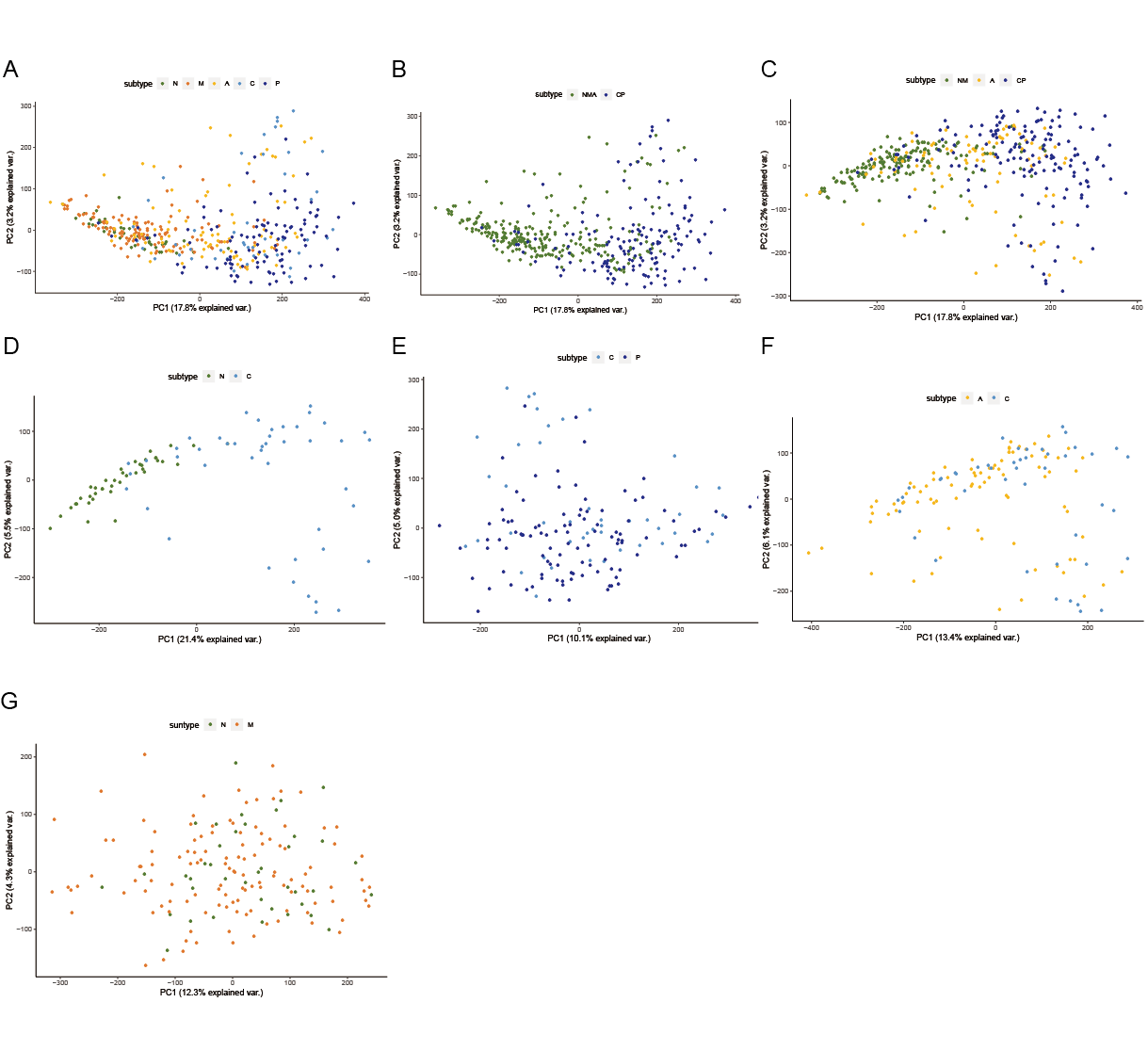
**Data quality and batch correction**

To confirm the quality of data we used, we firstly analyzed batch effect across the mouse liver samples and pool samples which were considered as PCT control and mass spectrometry control. Batch effect was analyzed by PCA. [吴老师建议说明写清从多少维度分] We found that technical control replicates analyzed using different mass spectrometries were separated in two groups and machine A was apart from machine B and C **(sFigure 1A, 1B, liuwei, 20181214),** which probably attribute to the different sources of chromatography columns. In order to eliminated the influence of batch effect on protein identification by different machines, batch correction was performed(**sNote3, tiansheng, 20181214**). Then we got a new data matrix (**sTable 4, to do . liuwei, 20181215.**) and reanalyzed batch effect using PCA. Finally, different machines did not separate in diagram of PCA. (**sFigure** **1C, 1D, liuwei, 20181215**) To confirm batch-effect elimination, we compared pearson correlation coefficient of technical replicates of thyroid tissues in un-correction group with it in correction group. The pearson correlation coefficient in correction group was ###, while in un-correction group was ### (**sFigure 1E, 1F, qiushi, 20181214**), which indicated that our new matrix with no batch effect.

Next, to accessed the variation of each protein intensity using batch-corrected matrix, we calculated the CV of different types of technical repliactes. The median of CV was ### in mouse liver samples, ### in pool samples and ### in technical replicates of thyroid tissues (**sFigure 1G, liuwei, 20181214**) and more details were in **supplemental table 5(liuwei, 20181214)**. The overall median technical variation expected for the same tissue samples measured through the FFPE-PCT-DIA workflow. We finally tested CV of biological thyroid triplicates. The median CV value of biological triplicates was ### which associated with biology of thyroid disease and the details were in **supplemental table5** (**sFigrue H, liuwei,20181214**).

**Classification of subtypes of thyroid tissues by PCA analysis**

To stratify the different types of thyroid disease by proteomic profile, PCA analysis was applied. First of all, we compared the five different types of thyroid disease by PCA and found that some of them could not be well resolved (**Figure 2A, weigang, 20181214**). Then we combined normal thyroid, multinodular goiter and follicular adenoma as benign group and follicular thyroid carcinoma and papillary thyroid carcinoma as malignant group and the result showed that these two groups could be partly separated, unfortunately, some of the patients were still overlapped (**Figure 2B weigang, 20181214**). According to the previous result, we speculated that it was follicular adenoma that was attributed to the overlap. Furthermore, we isolated follicular adenoma from the benign group and re-analyzed using PCA. NM group could be separated with CP group, and A was pervaded in both of the two groups (**Figure 2C weigang, 20181214**). Finally, we performed ten pair-wise comparison and it is showed that CN and CP could be well distinguished by PCA, however, AC and NM could not (**Figure 2D, 2E, 2F, 2G weigang, 20181214**). These results are matches with the phenomenon in clinical which pathologist either could not distinguished through cytopathology. The other results were in **supplemental Figure 2 weigang, 20181214**. We also performed with T-SNE(不知道怎么写t-sne) and the results were in **sFigure 3,qiushi,20181215**.



**Figure 2 PCA analysis of subtypes of thyroid disease.** (A) PCA plot for five different types of thyroid disease. N, M, A, C, P indicated that normal thyroid, multinodular goiter, follicular thyroid adenoma, follicular thyroid carcinoma and papillary thyroid carcinoma, respectively. (B) PCA plot for benign (consisting of N, M and A) vs malignant (consisting of C and P). (C) Separating A from benign group and analyzing using PCA. (D）PCA plot for A vs C showing similarly. (E) N and C were well separated in PCA plot. (F) N and P were obviously resolved using PCA analysis.

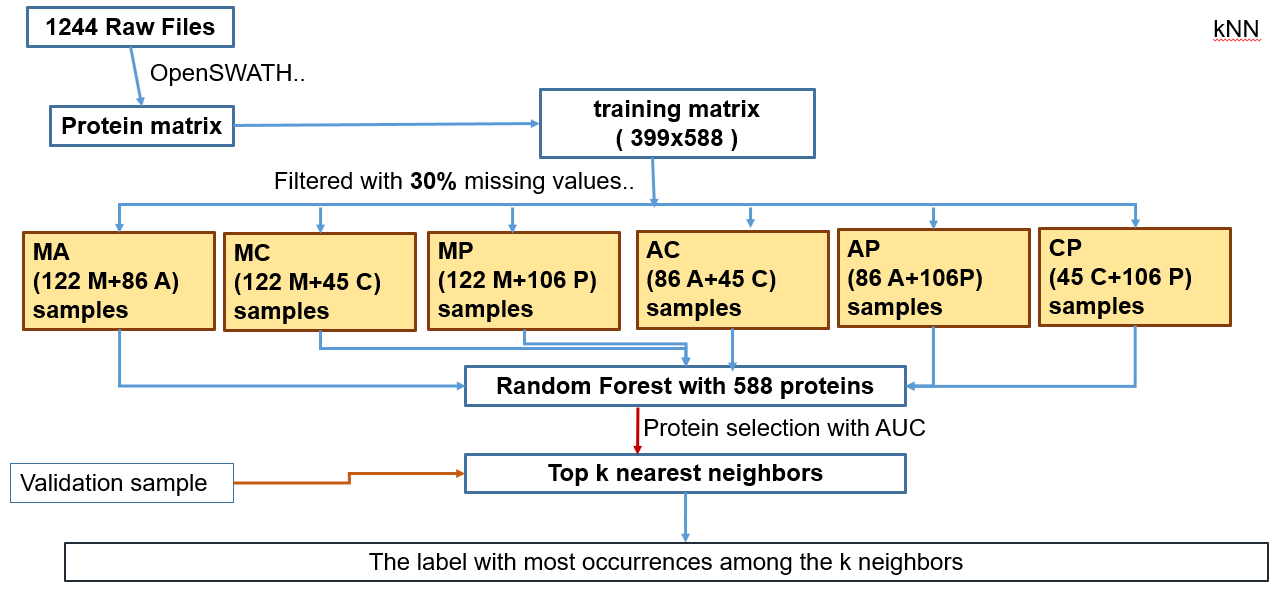
**Stratification of thyroid tissues using machine learning**

In order to stratify the specific type of the tissues by signature proteins, we trained 13 models (sNote2, wu) and the prediction using kNN model was based on random forest and nearest neighbor algorithms performed best (sTable6, need to be replenish, wu and Gopal).

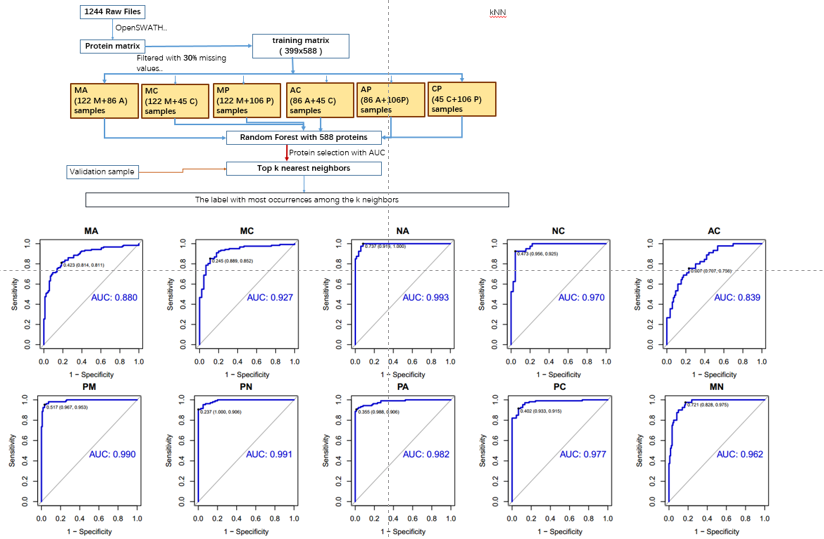
After averaging the intensities over the three biological replicas of each patient, we filtered out all the proteins with a cutoff of more than 30% missing values presented in all the samples (patients), we got a training matrix with 399 rows of patients and 588 columns of proteins (sTable 7, provided 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.

Figure 3A



B



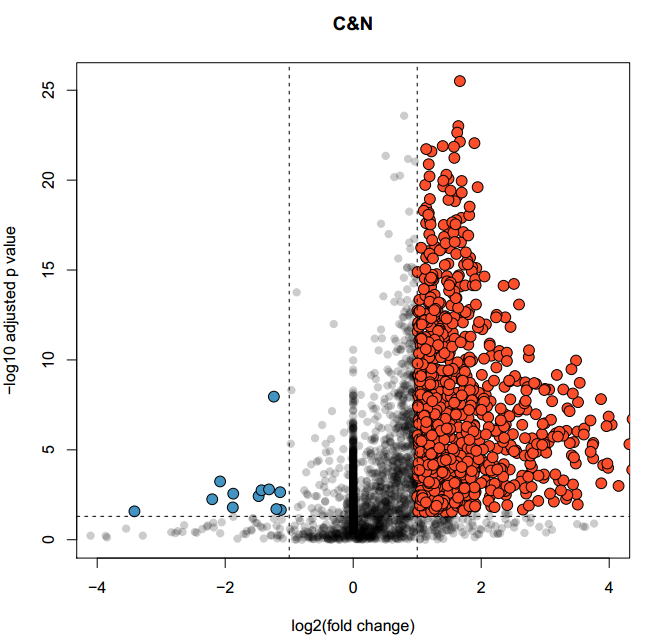
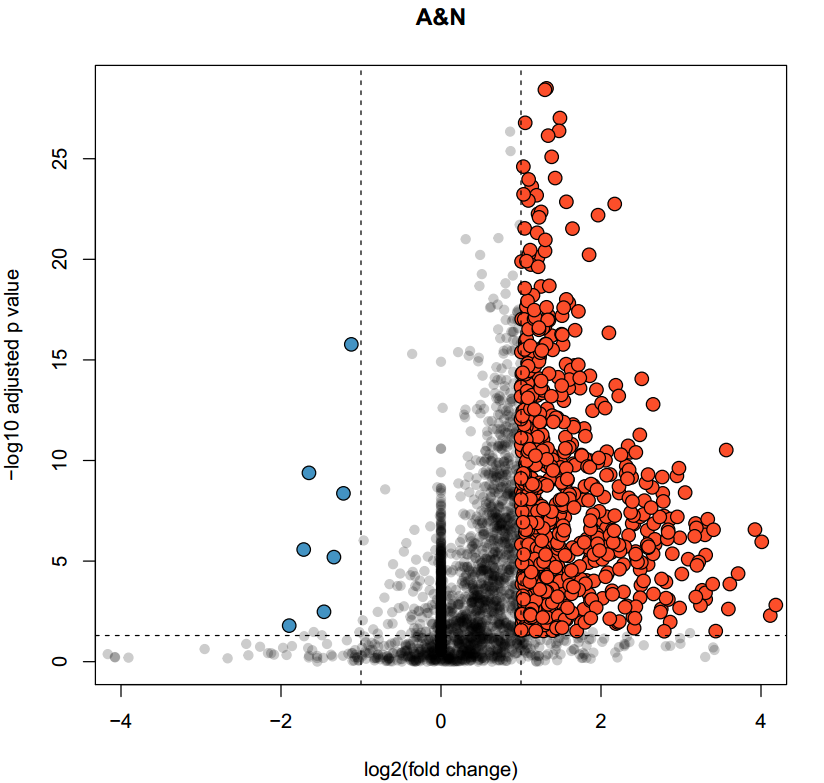
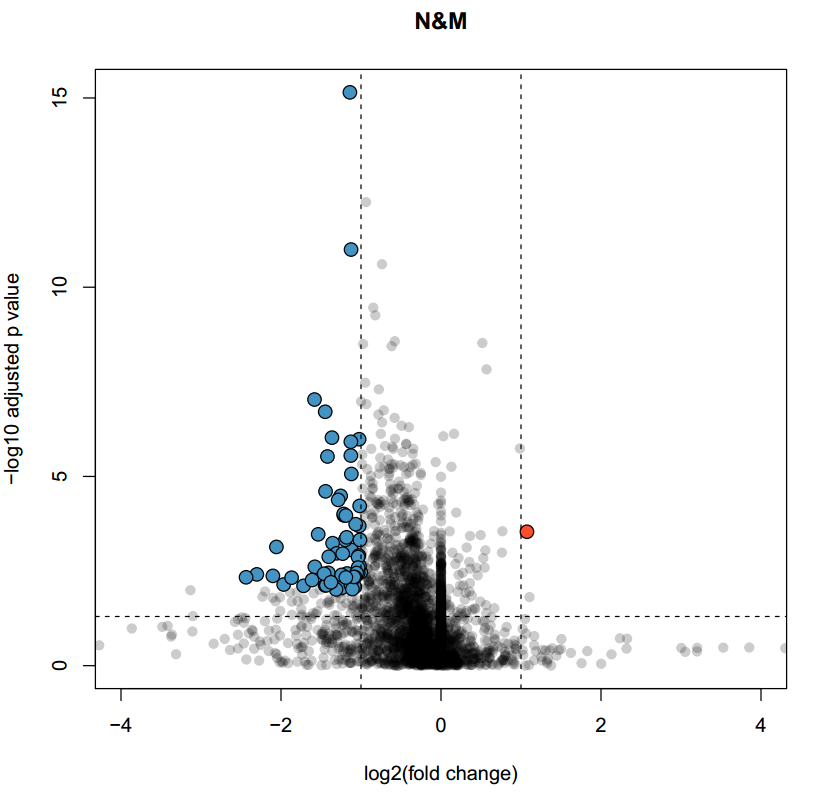
**Figure 3 Development of validations models by random forest （by wu,20181213）(A)** Process of validation models development.(B) Ten pair-wise ROC

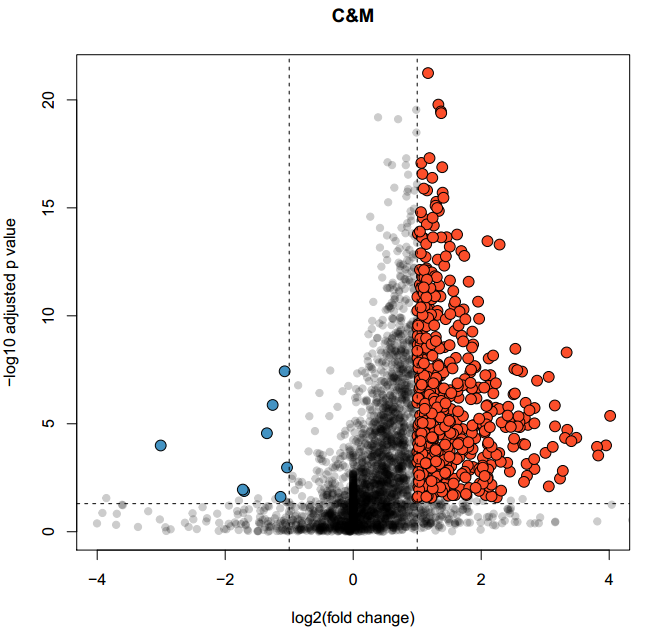
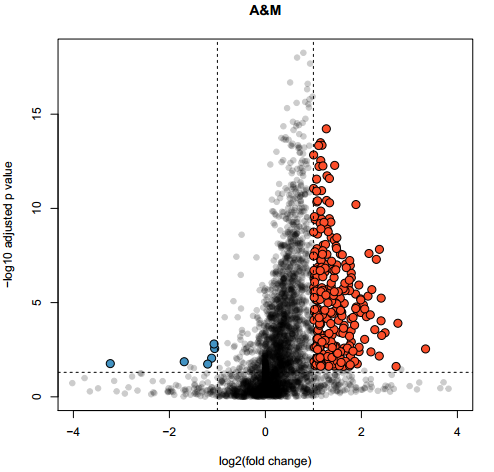
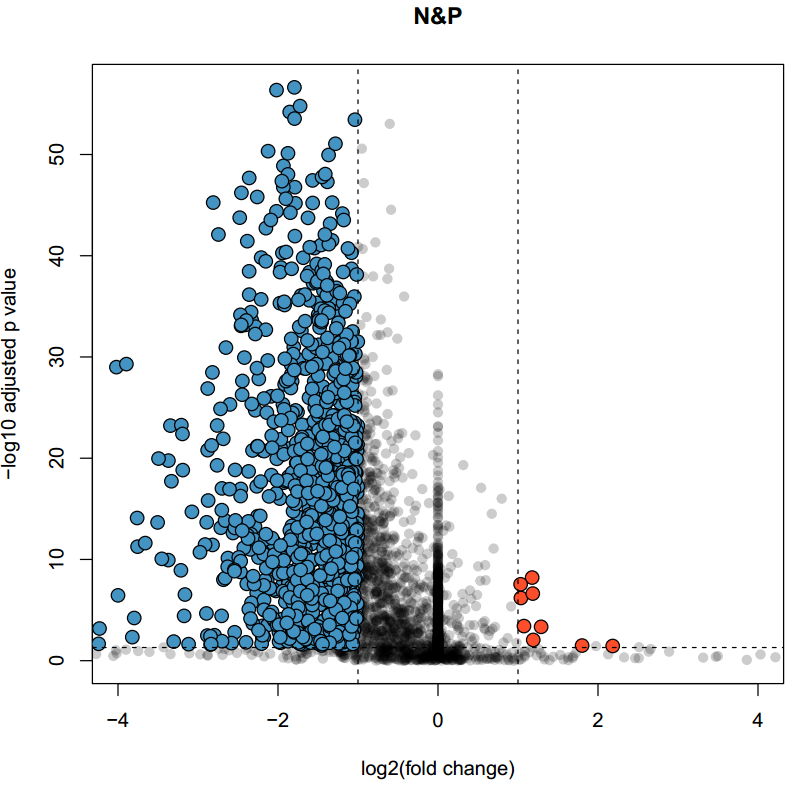
**Differentially expressed biomarkers**

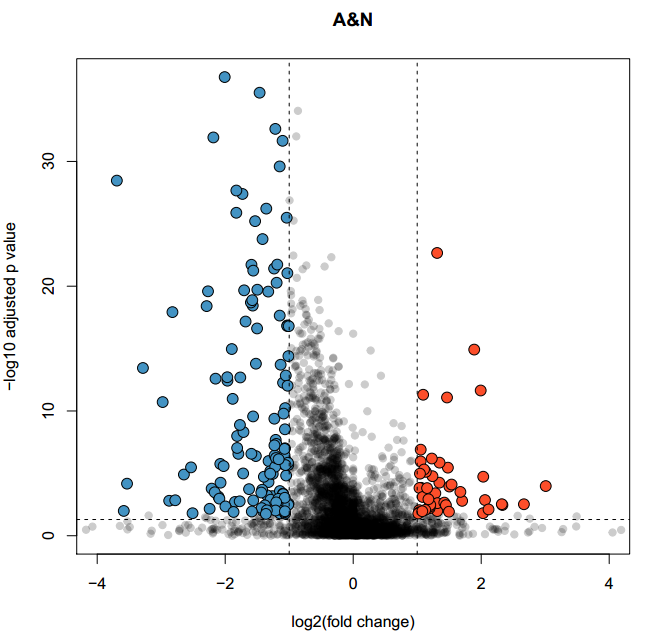
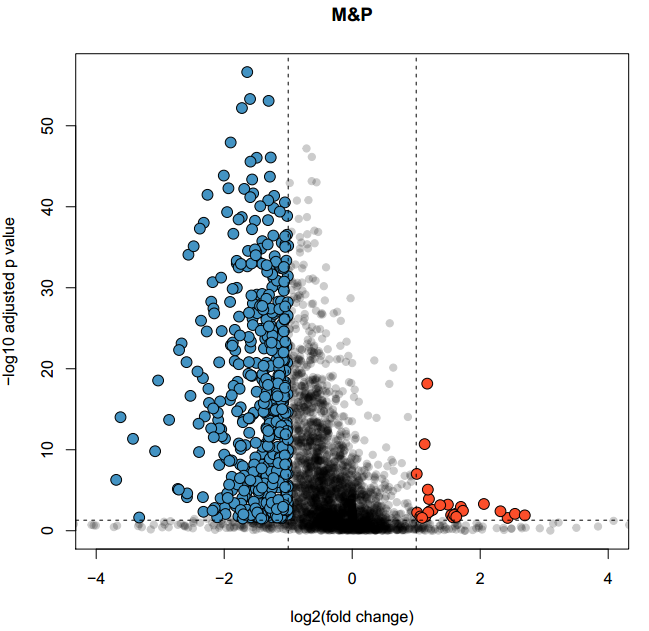
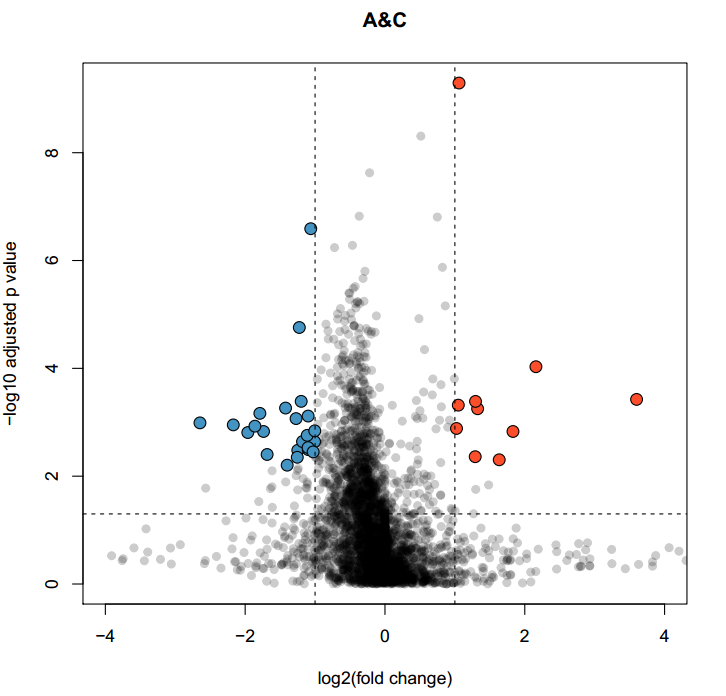
To comprehensively explore the differentially expressed protein, we compared ten pair-wise group from five subtypes of thyroid disease with two-fold-change cutoff and adjust P-value threshold. We found that the results of volcano plot were in one-to-one correspondence with PCA analysis. There were only 33 and 61 differential proteins in AC and NM groups, respectively. As for the well distinguished group using PCA analysis, there were 890 and 987 regulated proteins in pairs of CN and NP (**Figure 4A, weigang,20181214)** and detail of protein list was in supplemental Table9, liuwei, 20181214.

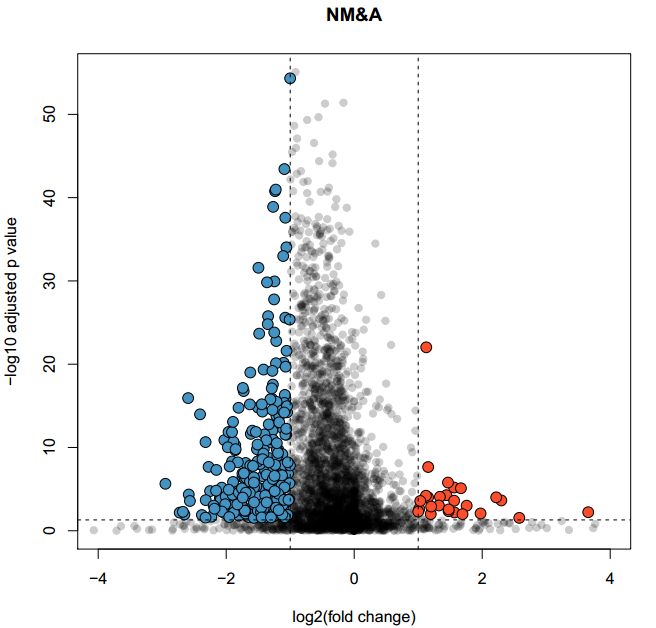
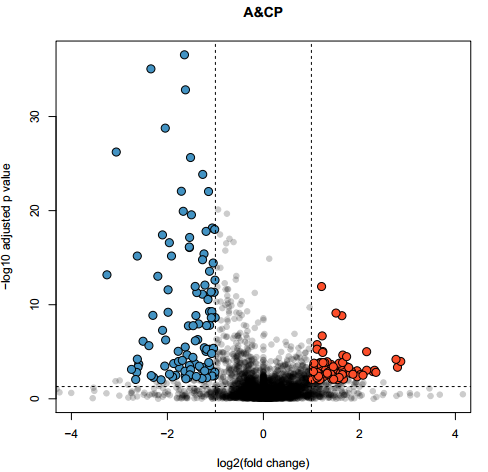
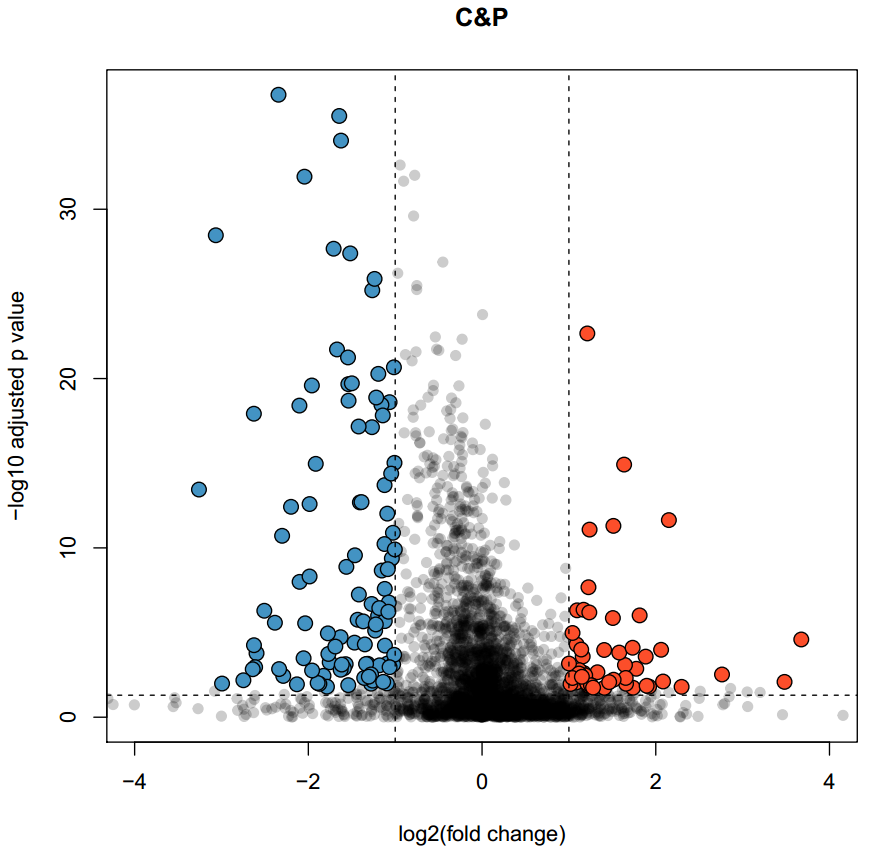
In order to find the unique proteins which could differentiate the two types of disease, we compared each thyroid disease with other four types using venn diagram and the complementary sets were the unique protein we wanted (**sTable 10，liuw，20181214**).

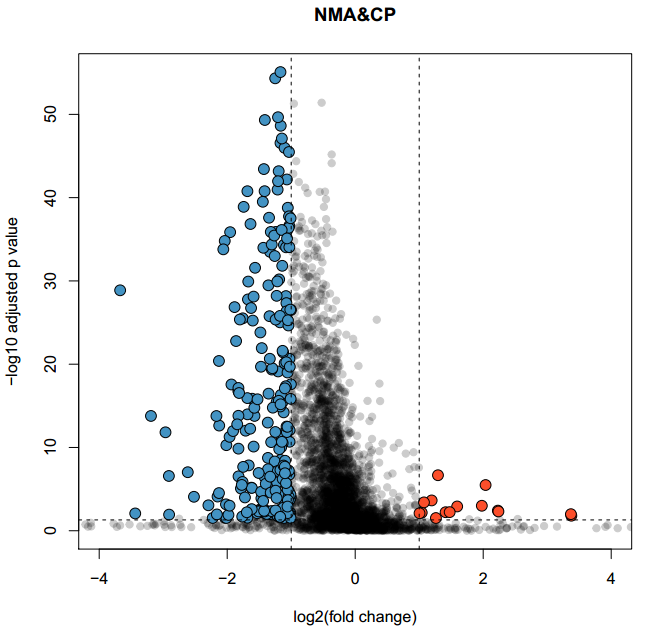
Figure 4A



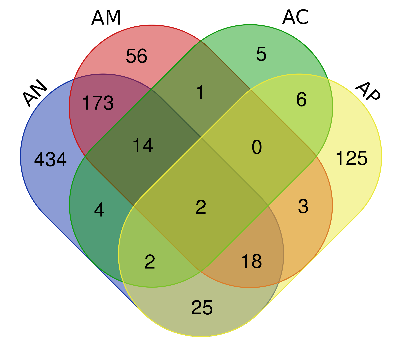
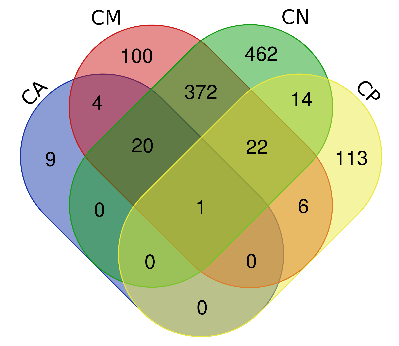
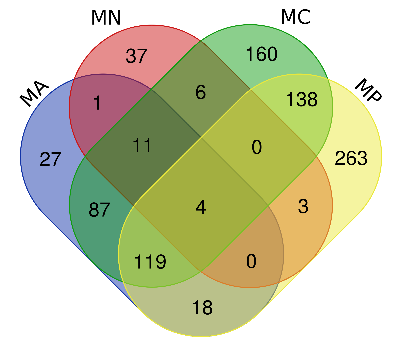
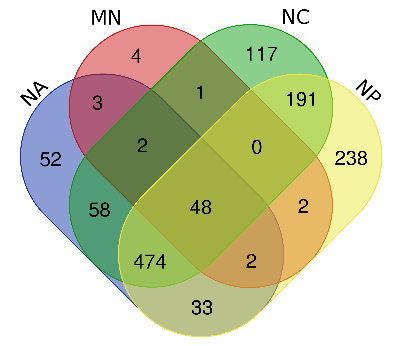
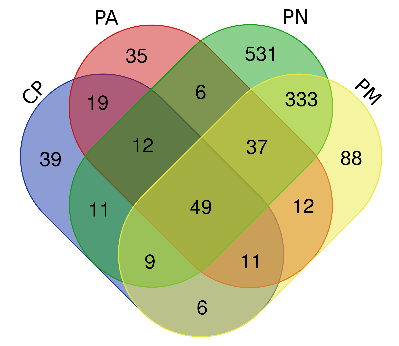






.

B

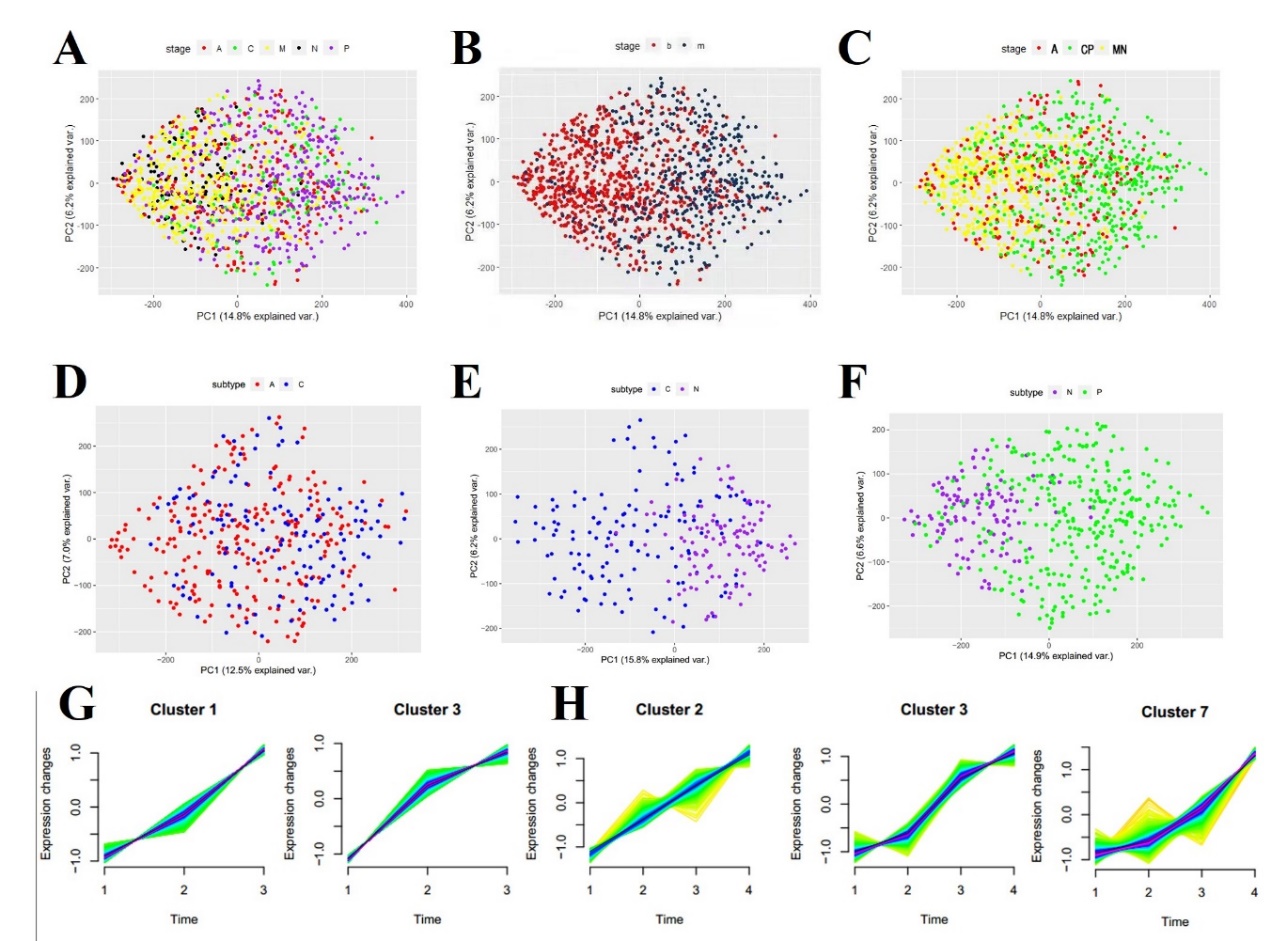


C

D

**Figure 4 Comparative analysis of differential expressed protein.** (A)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. (B) Venn (C) cluster analysis (D) protein network (E) pathway mapped to top hits for each pathological type

**Protein asscitiated with turmor progress**



**Figure 5 cluster analysis of protein associated with tumor progress**

**Figure 6 validation of best distinguishing proteins in TMA of the 4 pathological types**

**Figure X Combining discovery and validation sets**

Combination data- tSNE (different histologies), benign vs malignant PCA (maybe to

1. show that for the validation set which is indeterminate, it blurs the line more??
2. New model from combining to 540 samples
3. Bootstrap or other internal validation algorithm

**Discussion**

In the present study, we used FFPE-PCT-DIA methodology to investigate different types of thyroid nodules

**Materials and Methods**

**Patient and tissue samples**

FFPE specimens from patients with thyroid adenomas, nodular goiters or different thyroid cancer types, whom were treated in the Singapore General Hospital (SGH) between the year 2012-2017, have been retrospectively retrieved from the Pathology Department of SGH. Haematoxylin and eosin-stained slides from tissue blocks of every single patient were carefully reviewed by an experienced histopathologist who marked out the disease region for tissue coring. The sample cohort comprised multinodular goiter (MNG, 122 cases), papillary thyroid carcinoma (PTC, 105 cases), follicular adenoma (FA, 87 cases), follicular thyroid carcinoma (FTC, 46 cases) and normal thyroid (NT, 40 cases), in total there were 400 cases. Normal thyroid tissues were obtained from parathyroidectomy cases or neck dissection specimens from cases of laryngeal or oesophageal cancers that have no thyroid pathology. Each case contained 3 biological replicates.

The selection of cases was excluded with extensive thyroiditis and/or inflammation and the pathological lesions of the cases were more than 1 cm in diameter. Pathologists viewed and marked haematoxylin-eosin stained slides and then each case were made 4 punch cores at the region of interest using a hollow metal needle of 1 mm internal diameter and 1 mm length. Each dry FFPE punch weight around 0.5 mg including wax.

Ethics.(Dr. Gopal)

**Dewaxing, rehydration and hydrolysis of FFPE tissues (Need to rewrite)**

For each case, three biological replicates of FFPE punches were processed. Before dewaxing, specimens were weighted and recorded. 1 ml heptane (Sigma) was added into a 1.5 ml tube containing FFPE tissue samples. The sample was then votexed at 800 rpm for 10 min at 25 degrees. This step was operated twice.

Discard heptane and add 1 ml 100% ethanol (Sigma), 200 ul of 90% ethanol, 200 ul of 75% ethanol successively into each tube and votexed at 800 rpm for 5 min at 25 degrees. Then tissues were transfered from each Eppendorf tubes into PCT-Micro Tubes and added 150 ul of 0.1% formic acid (Sigma)to each microtube, vertexed at 600 rpm for 30 min at 30 celsius degrees. Discard the supernatant and then wash tissue samples with 150 ul of 100 mM Tris-HCl (pH=10, Sigma). Pull off the supernatant. After washing, add 20 ul 100 mM Tris-HCl (pH=10) to each tube, gently vortex at 600 rpm for 30 min at 95 celsius degrees. Finally, cool down immediately at 4 celsius degrees.

**Tissue lysis, protein extraction and protein digestion (Need to rewrite)**

30 ul of lysis buffer was added to each microtube and close with MicroPestle and then performed by PCT. PCT program setting: 90 cycles containing 25 s of 45,000 p.s.i. high pressure plus 10 s at ambient pressure, at 30 celsius degrees. After that, 10 mM Tris(2-carboxyethyl)phosphine hydrochloride (100 mM stock, 5 ul, Sigma) and 40 mM Iodoacetamide (800 mM stock, 2.5 ul, Sigma) were simultaneously added to the solution, followed by a 30 min incubation in the dark with gentle vortexing at 25 celsius degrees. lysC (Hualishi Scientific) was added to each tube at ratio 40:1 (protein to lysC). PCT-assisted lysC digestion was performed with the following setting: 45 cycles containing 50s of high pressure at 20,000 p.s.i. plus 10 s at ambient pressure, at 30 degrees. Secondly, trypsin (Hualishi Scientific) was added into each microtube for tryptic digestion. The protein to enzyme ratio is 50:1. PCT-assisted trypsin digestion was performed with the following setting: 90 cycles containing 50 s of high pressure at 20,000 p.s.i. plus 10 s at ambient pressure, at 30 degrees.

**C18 Desalting (Need to rewrite)**

The digested peptide solution was transferred from the microtube into a new Eppendorf tube. 10% trifluoroacetic acid (Thermo Fisher) was added to the solution at a final concentration of 1% to stop the digestion and to acidify the solution to pH 2-3. Peptides were then desalted by C18 spin columns (170 μg maximum capacity, The Nest Group) according to operating instructions. Finally, peptide eluates from C18 spin column were dried by speedvac (Christ). The dry peptide pellet were reconstituted using 30μl of MS buffer A (98% HPLC water, 2% acetonitrile, 0.1% formic acid). The peptide concentration was measured by Nanoscan (Analytic Jena) at A280.

**DDA and DIA measurement on QE-HF**

0.5 μg cleaned peptides were injected and analyzed by DIA-MS on a QE-HF mass spectrometer (Thermo Fisher Scientific). The LC gradient consisted of buffer A (98% HPLC water, 2% acetonitrile, 0.1% formic acid) and buffer B (2% HPLC water, 98% acetonitrile, 0.1% formic acid).

DDA:

DIA: Peptides were separated with a linear gradient of 3% to 28% buffer B over 45 min gradient time at a flow rate of 0.3 μl·min-1, and the DIA acquisition scheme to 24 variable windows ranging from 400 to 1000 m/z.

**DIA library construction（ask for tiansheng）**

We performed 204 DDA analysis of tissue samples, either whole proteome or fractionated proteomes using SDS-PAGE, from #64## patients. The files were analyzed using MaxQuant， leading to identification of ### peptide precursors from ### protein groups. Then we built a DIA library as described previously (cite Guo, Nature Med; Nature Protocol) containing ### peptide precursors from ### protein groups.

**Random Forest (Wu)**

**Bioinformation analysis**

**Statistical analysis**

Statistical analysis was performed using R software (version 3.5.1).

Sensitivity, specificity, PPV and NPV were calculated following the established methodology.

**References**

1. Burman KD, Wartofsky L. CLINICAL PRACTICE. Thyroid Nodules. The New England journal of medicine 2015;373:2347-56.

2. Jameson JL. Minimizing unnecessary surgery for thyroid nodules. The New England journal of medicine 2012;367:765-7.

3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: A Cancer Journal for Clinicians 2018;68:7-30.

4. Faquin WC, Bongiovanni M, Sadow PM. Update in thyroid fine needle aspiration. Endocrine pathology 2011;22:178-83.

5. Alexander EK, Kennedy GC, Baloch ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. The New England journal of medicine 2012;367:705-15.

6. Fisher SB, Perrier ND. The incidental thyroid nodule. CA Cancer J Clin 2018;68:97-105.

7. La Vecchia C, Malvezzi M, Bosetti C, et al. Thyroid cancer mortality and incidence: a global overview. Int J Cancer 2015;136:2187-95.

8. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66:115-32.

9. Aragon Han P, Olson Mt Fau - Fazeli R, Fazeli R Fau - Prescott JD, et al. The impact of molecular testing on the surgical management of patients with thyroid nodules. 2014.

10. Fallahi P, Giannini R, Miccoli P, Antonelli A, Basolo F. Molecular diagnostics of fine needle aspiration for the presurgical screening of thyroid nodules.

11. Yip L, Ferris RL. Clinical application of molecular testing of fine-needle aspiration specimens in thyroid nodules.

12. Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE. Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. The Journal of clinical endocrinology and metabolism 2013;98:E1852-60.

13. Guo T, Kouvonen P, Koh CC, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nature medicine 2015;21:407-13.

14. Shao S, Guo T, Koh CC, et al. Minimal sample requirement for highly multiplexed protein quantification in cell lines and tissues by PCT-SWATH mass spectrometry. Proteomics 2015;15:3711-21.

15. Shao S, Guo T, Gross V, et al. Reproducible Tissue Homogenization and Protein Extraction for Quantitative Proteomics Using MicroPestle-Assisted Pressure-Cycling Technology. Journal of proteome research 2016;15:1821-9.

16. Gillet LC, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Molecular & cellular proteomics : MCP 2012;11:O111.016717.