Population serum proteomics uncovers prognostic protein classifier and molecular mechanisms for metabolic syndrome

Xue Cai[1#], Zhangzhi Xue[1#],Fang-Fang Zeng[2#], Jun Tang[1,3,4#], Liang Yue[1,3,4,5#], Bo Wang[6#], Weigang Ge[6], Yuting Xie[1,3,4,5], Zelei Miao[1,3,4], Wanglong Gou[1,3,4], Yuanqing Fu[1,3,4], Sainan Li[1,3,4,5], Jinlong Gao[1,3,4,5], Menglei Shuai[1,3,4], Ke Zhang[1,3,4], Fengzhe Xu[1,3,4], Yunyi Tian[1,3,4], Nan Xiang[6], Yan Zhou[1,3,4,5], Peng-Fei Shan[7], Yi Zhu[1,3,4,5]*, Yu-ming Chen[8]*, Ju-Sheng Zheng[1,3,5]*, Tiannan Guo[1,3,4,5]*

[1] Westlake Intelligent Biomarker Discovery Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China;
[2] Department of Public Health and Preventive Medicine, School of Medicine, Jinan University, Guangzhou, 510080, China;
[3] School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China;
[4] Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, Zhejiang, China;
[5] Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China;
[6] Westlake Omics (Hangzhou) Biotechnology Co., Ltd. No.1, Yunmeng Road, Cloud Town, Xihu District 310000, Hangzhou, Zhejiang, China;
[7] Department of Endocrinology, the Second Affiliated Hospital of Zhejiang University School of Medicine, 88 Jiefang Road, Hangzhou, Zhejiang, 310009, China;
[8] Guangdong Provincial Key Laboratory of Food, Nutrition and Health; Department of Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou 510080, China

#These authors contributed equally
*Correspondence: chenyum@mail.sysu.edu.cn (Y.M.C); zhengjusheng@westlake.edu.cn (J.S.Z.); guotiannan@westlake.edu.cn (T.G.)

ABSTRACT

Metabolic syndrome (MetS) is a complex metabolic disorder with a global prevalence of 20-25%. Early identification and intervention would help minimize the global burden on healthcare systems. Here, we measured over 400 proteins from ~20,000 proteomes using data-independent acquisition mass spectrometry for 7890 serum samples from a longitudinal cohort of 3840 participants with two follow-up time points over ten years. We then built a machine learning model for predicting the risk of developing MetS within ten years. Our model, composed of 11 proteins and the age of the individuals, achieved an area under the curve of 0.784 in the discovery cohort (n=855) and 0.774 in the validation cohort (n=242). Using linear mixed models, we found that apolipoproteins, immune-related proteins, and coagulation-related proteins best correlated with MetS development. This population-scale proteomics study broadens our understanding of MetS, and may guide the development of prevention and targeted therapies for MetS.

Keywords: population proteomics, metabolic syndrome, DIA-MS, machine learning, prospective prediction

INTRODUCTION

Metabolic syndrome (MetS), also called insulin resistance syndrome [1], is a complex metabolic disorder characterized by abdominal obesity, atherogenic dyslipidemia, raised blood pressure, insulin resistance, central obesity, prothrombotic and proinflammatory states [2, 3]. About 20-25% of adults worldwide suffer from MetS [4-6]. In Chinese adults, the prevalence of MetS reached 24.2% (24.6% in men and 23.8% in women) in 2018 [7]. According to national health and nutrition examination survey (NHANES), about one-third of US adults were diagnosed with MetS between 1988 and 2010 [8]. MetS has been reported to predispose to several serious diseases [9-11], including diabetes, cardiovascular disease, coronary heart disease, and several common cancers. An early prediction and diagnosis of MetS would allow faster interventions, reducing the burden on the healthcare systems. There are currently diagnostic criteria of MetS, with the main components of waist circumference, glucose, blood pressure, triglycerides (TG), high-density lipoprotein cholesterol (HDL-C) [3]. But the pathogenesis of MetS is unclear and there is still a lack of targeted therapy for MetS. Thus, longitudinal large cohort population studies are necessary, which can effectively predict the development and explore the pathogenesis of MetS.

Plasma and serum are the predominant samples used for diagnostic analyses in clinics with low invasiveness and ease of collection and preservation. Plasma and serum proteomic studies based on large cohort population have used aptamer-based (SOMAscan) platform [12] and proximity extension assay (Olink proteomics) [13] as tools for metabolic disease [14, 15]. For example, using the SOMAscan platform, Ganz et al quantified the plasma proteins of two longitudinal cohorts of 938 and 971 samples with one follow-up point, and achieved 0.70 of C statistics in validation cohort for cardiovascular risk prediction based on 9-protein panel [16]. Proteomics of 3301 human plasma samples from two different subcohorts were also measured by the SOMAscan platform to explore the non-linear alterations in the human plasma proteome with age [17]. Furthermore, dozens of blood proteins were measured by the proximity extension assay (Olink proteomics) to investigate the association with BMI and waist circumference in a cross-sectional study of 3,308 participants [18] and predict myocardial infarction risk with C-index of 0.68 in 5,131 diabetes patients [19].

The rapid development of mass spectrometry (MS)-based proteomics provided solid technical support needed by large cohort proteomics [20]. Specifically, this technology provides high-throughput and reproducible analyses with minimal sample amounts. Data-independent acquisition (DIA) MS measures all fliable peptide precursor ions without prior knowledge, allowing the unbiased discovery of new biomarkers [21, 22]. For instance, using DIA MS methods, Niu *et al.* analyzed the plasma samples of 596 individuals and 79 liver biopsies in three weeks and predicted future liver-related events and all-cause fatalities [23]. Bruderer *et al.* also analyze 1508 plasma samples with robust capillary-flow DIA MS to investigate potential biomarkers for wight loss [24].

Here, we reliably measured over 400 proteins from ~20,000 proteomes using DIA-MS for 7890 samples from a longitudinal cohort of 3840 participants with two follow-up time points over ten years. Based on our data, we built a model to predict the risk of developing MetS within ten years. We also explored new potential biomarkers and networks associated with MetS, providing reference for the pathogenesis and targeted therapy of MetS.

RESULTS

Serum proteome profiling of a longitudinal cohort

Here, we included 3840 participants from the community-based prospective cohort study, namely Guangzhou Nutrition and Health Study (GNHS, ClinicalTrials.gov identifier: NCT03179657) [25-28]. Serum samples were collected at three of four time points: 3479 samples were collected at baseline (between 2008 and 2010), 2638 at the second follow-up (between 2014 and 2017), and 1773 at the third follow-up (between 2018 and 2019) (**Table 1**). No serum samples were collected at first follow-up. On average, 5.6 years passed between the baseline and the second follow-up, and 3.4 years between the second and the third follow-up. The statistics of the participants at the three time points are summarized in **Table 1**; their clinical information is provided in **Table S1**. In particular, the median age for the baseline, second follow-up, and third follow-up groups were 57.4 (38.2-80.4), 63 (44.3-83.3), and 66.1 (46.8-86.2), with 69%, 68%, and 69% female participants, respectively. We next divided the serum samples into a discovery (n=4794) and a validation cohort (n=3094) randomly (**Figure 1**), and the sample preparation, MS data collection, data analyses were then independently performed.

We randomly divided the 4796 serum samples of our discovery cohort into 178 batches, each containing 29 unique serum samples, two biological replicates, and one quality control (QC) sample to ensure the stability of the entire workflow (**Figure 1, Table S2A**). An MS-QC sample was injected before each batch to ensure the instrument was in good condition. Next, each sample was acquired 2-3 times using a 20-min DIA-MS method [21], summing up to 11,646 MS data, including 301 QC samples, 526 biological replicates, 288 MS-QC samples, 5735 technical replicates (**Figure 1**). Finally, we used the software tool DIA-NN [29] to generate a matrix of 583 proteins and 11,646 MS data with 45% missing values.

To obtain high-quality proteomics data for our subsequent analyses, we cleaned and sorted the protein matrix returned by DIA-NN. First, we excluded 445 MS data with protein identifications below 75% of the median protein identification (<245), which may have been affected by an imperfect sample preparation or the instrument status. We also excluded proteins with missing values exceeding 80% of the protein matrix (n=145) (**Figure 1**). The median Pearson correlation coefficient (r) of the MS-QC samples achieved 0.94, proving the high stability of the MS instruments (**Figure 2A)**. Next, we checked the MS-QC samples of each batch and excluded 167 MS data with no MS-QC samples and no biological or technical replicates, as it would have been hard to assess the quality of their MS data. Finally, we corrected the batch effects using ProteomeExpert [30] and observed a significant improvement in our data (**Figure 2B)**. The Pearson correlation coefficients (r) of 512 biological replicates and 5317 technical replicates were calculated, and the median r of the two sets of replicates were 0.97 and 0.96 (**Figure 2C)**, respectively. These results prove the high consistency and reproducibility of our data. For our subsequent data analyses, we combined the quantitative results of those replicates with r >0.8.

The experimental design of our validation cohort was the same as that of our discovery cohort. We divided 3094 serum samples into 132 batches (**Table S2A**). A total of 7324 MS data (containing 247 QC samples, 338 biological replicates, 213 MS-QC samples, and 3432 technical replicates) were measured using 20-min DIA-MS. Next, we excluded 194 MS data with protein identifications below 75% of the median protein identification (<238), and sub the protein matrix containing 413 proteins as we obtained from the discovery cohort. The median r values of the MS-QC samples, the biological replicates, and the technical replicates were 0.93 (**Figure 2A**), 0.96,

and 0.95 (**Figure 2C**), respectively. We also corrected the batch effect using ProteomeExpert and observed a significant improvement (**Figure 2B**).

After these steps, from the discovery cohort we obtained a protein matrix containing 4637 samples and 438 proteins, with 19.0% missing values (**Figure 1, Table S2B**); from the validation cohort we obtained a protein matrix containing 3067 samples and 413 proteins, with 18.4% missing values (**Figure 1, Table S3B**).

A protein-based classifier for predicting MetS insurgence

For the discovery cohort, we collected a set of samples at the baseline time point, including 267 non-MetS samples from individuals who were diagnosed with MetS at the second or third follow-up and 588 non-MetS samples from individuals who were not diagnosed with MetS at any of our time points (**Figure 3A**). Using these data, we built a machine learning model to evaluate the serum proteins' ability to predict the risk of developing MetS within ten years.

We used LightGBM to generate a machine learning model using the discovery dataset of 855 samples containing 288 serum proteins, as well as age and sex information (**Figure 3B, Table S4A**). First, we randomly split the dataset into a training dataset (n=684) and an internal validation dataset (n=171). Then, a 2-step feature selection was used for choosing protein features from the 288 serum proteome (**Figure 3B**). Specifically, the final protein features were apolipoprotein A-I (APOA1), sex hormone-binding globulin (SHBG), apolipoprotein D (APOD), vitronectin (VTN), attractin (ATRN), clusterin (CLU), heparin cofactor 2 (SERPIND1, HCII), alpha-1B-glycoprotein (A1BG), apolipoprotein C-II (APOC2), immunoglobulin heavy variable 4-39 (IGHV4-39), and apolipoprotein B-100 (APOB) (**Figure 3C**). We optimized our model using a ten-fold cross-validation. Next, we tested the model using the randomly selected internal validation dataset and achieved an AUC of 0.78, indicating that our model can efficiently predict the risk of developing MetS within ten years (**Figure 3D**). Finally, we tested the model using 242 samples from an independent validation cohort and achieved an AUC of 0.77 (**Figure 3D, Table S4B**). These results show that the model generalized well to independently collected samples and the protein used in the model may present as promising biomarker candidates.

Circulating proteins regulated in MetS samples

To find which serum proteins are potentially affected by MetS, we used two linear mixed models to examine: (1) the relationship between protein level and MetS at the baseline; (2) the dynamic relative changes along the three time points. For the baseline analysis, we used 1736 baseline samples (309 MetS, 1427 non-MetS); for the dynamic analysis, we used 4561 samples from all three timepoints (986 MetS, 3575 non-MetS). With the baseline analysis, we found that 175 proteins were significantly dysregulated in the MetS compared with the non-MetS group (**Table S5A**), while 194 proteins were associated with the dynamic development of MetS during ten years (**Table S5B**). A total of 143 proteins were dysregulated according to both analyses (**Figure 4A**).

To find which pathways may be affected by MetS, we analyzed the 175 and 194 dysregulated proteins using Ingenuity Pathway Analysis (IPA). We found that 11 pathways were significantly dysregulated according to our baseline analysis (**Figure 4B, Table S5C**), and 12 were significantly dysregulated according to our dynamic analysis (the threshold for both analyses was -log(p-value)>1.3) (**Figure 4B, Table S5D**). In particular, ten pathways were significantly dysregulated according to both analyses, including FXR/RXR activation, LXR/RXR activation, atherosclerosis signaling, maturity-onset diabetes of young (MODY) signaling, and acute-phase

response signaling (**Figure 4B**). Seven pathways with the most significant p-values were mainly enriched in apolipoproteins (**Table S5C, 5D**); among these seven, FXR/RXR activation and LXR/RXR activation are involved in metabolism and associated with RXR activation [31]. RXR activation is known to correlate highly with MetS, as drugs targeting RXR and its heterodimeric partners are already in clinical use against MetS [32]. Additionally, MetS has been reported to increase the risk of atherosclerosis [33] and is a clinical characteristic of MODY [34]. Acute-phase response signaling has also been associated with MetS [35]. The growth hormone signaling pathway was significantly dysregulated only in the baseline analysis (**Figure 4B**), indicating this pathway was dysregulated in MetS patients but possibly not linked with its development. In particular, the increase in growth hormone can cause insulin resistance in men [36]. Type II diabetes mellitus and the AMPK signaling pathways were significantly dysregulated in the dynamic analysis only (**Figure 4B**), indicating these pathways may be primarily involved in the development of MetS. Diabetes is a component of MetS [37], and evidence suggests that the dysregulation of the AMPK signaling pathway may lead to MetS [38].

The expression of 32 significantly dysregulated proteins ($|\beta|>0.2$, see Methods), among the 175 proteins and 194 proteins, showed the same performance in the discovery and the validation cohorts (**Figure 4C**). These proteins were dysregulated in MetS with respect to the non-MetS patients and were linked to the development of MetS. The elevated expression of APOC2 and VTN (**Figure 4D**), and the lower expression of AOPA1, APOD, A1BG, CLU, and APOC2 (**Figure 4D**) in MetS patients were consistent with the results of above analysis (**Figure 3D**). In addition, two apolipoproteins, apolipoprotein C-III (APOC3) and apolipoprotein F (APOF), and several other reported proteins, pigment epithelium-derived factor (PEDF), alpha-1-antitrypsin (A1AT) and afamin (AFM) were also dysregulated in MetS patients (**Figure 4D**). These results show that apolipoproteins and their related pathways are closely related to the occurrence and development of MetS.

Molecular characterization of MetS subtypes

MetS can be diagnosed when at least three of five diagnostic indicators are met: waist circumference, fasting blood glucose, blood pressure (SBP/DBP), fasting triglyceride (TG) and fasting high-density lipoprotein (HDL)-C (see Methods) [37]. Subtypes of MetS are based on different indicators. Here, we compared the different types of MetS and non-MetS and explored the molecular mechanism of MetS based on different indicators.

We selected 444 MetS samples that met the indicators of abnormal TG (**Table S6A, B**) and 444 non-MetS control samples with normal TG and matching age, sex, and other MetS indicators same as MetS. Similarly, we selected 169, 235, 384, and 354 paired samples to analyze abnormal HDL, glucose, waist, and SBP/DBP, respectively (**Table S6A, B**). No dysregulated proteins were discovered in the waist- and SBP/DBP-based MetS samples and only one up-regulated protein was deregulated in the glucose-based MetS (**Figure S1**). On the other hand, six up-regulated and three down-regulated proteins were observed in the TG-based MetS samples (**Figure 5A**), while one up-regulated and six down-regulated proteins were observed in the HDL-based MetS samples (**Figure 5B**). The expressions of these dysregulated proteins were consistent in the discovery and validation cohorts (**Figure 5C, 5D**).

Next, we analyzed the dysregulated proteins in TG- and HDL-based MetS samples using the IPA. The network enriched in the dysregulated proteins from the TG-based MetS samples (**Figure S2**) showed that the increased abundance of APOE, APOC1, APOC2, and APOC3 led to the

activation of HDL and LDL, whereas predict the inhibition of VLDL-cholesterol and insulin. In contrast, in the HDL-based MetS, the decrease in APOD and APOF led to the inhibition of HDL, and predicted the inhibition of LDL (**Figure S3**).

Proteins associated with MetS initiation

To investigate the development mechanisms of MetS, we next focused on participants that were not diagnosed with MetS at the baseline but developed MetS within ten years. By looking at the five diagnostic indicators of MetS mentioned above, some of these people met 1-2 indicators, others none. Therefore, we explored the proteins that significantly changed in the serum of participants that developed MetS based on the five abnormal indicators.

We selected 146 individuals from the discovery cohort that only met the waist indicator for MetS at baseline (**Table S7A, E**). Of these individuals, only 19 developed MetS. As the remaining 127 participants had an abnormal waist indicator also at the two follow-up visits, they acted as controls. To minimize the influence of sample size, we employed oversampling in our analysis (see Methods). In the MetS patients, 12 proteins were significantly up-regulated, while 24 were significantly down-regulated at the baseline compared with the control group (**Figure 6A, Table S7B**). These proteins included angiotensinogen (AGT), apolipoprotein D (APOD), and thrombospondin-1 (THBS1).

We next focused on TG. A total of 24 cases and 112 controls were selected similarly to the previous analysis on the waist indicator. Only five proteins were significantly up-regulated, and six were significantly down-regulated (**Figure 6B**, **Table S7A, C**), including Xaa-Pro aminopeptidase 1 (XPNPEP1) and apolipoprotein F (APOF). Using the SBP/DBP indicator, 45 cases and 246 controls were selected; eight proteins were significantly up-regulated and 19 were significantly down-regulated (**Figure 6C, Table S7A, D**), including apolipoprotein C-I (APOC1) and apolipoprotein C-IV (APOC4). Finally, as less than ten cases satisfied the selection criteria based on the blood glucose and HDL indicators, we did not analyze them.

Among the dysregulated proteins we identified, adiponectin (ADIPOQ) and taste receptor type 1 member 3 (TAS1R3) were dysregulated both in the TG-based and the waist-based MetS analyses (**Figure 6D**). This finding suggests that these two proteins may play a key role in the development of MetS. Additionally, SHBG was dysregulated both in the TG-based and the SBP/DBP-based MetS analyses, and was also selected in our machine learning model for predicting MetS risk.

DISCUSSION

We generated a population proteomics resource based on 7,890 serum samples collected from a prospective cohort of 3,840 participants over ten years, with a total of 18,970 DIA-MS data. We then generated a machine learning model to predict the risk for an individual of developing MetS over the next ten years. We also used this dataset to explore the molecular bases of MetS onset and development.

To our knowledge, our study provides the largest DIA-based serum proteome from a prospective population cohort with two follow-up time points to date. Niu *et al.* recently built a dataset of 311 plasma proteins from 596 individuals using a 21-min DIA-MS method and Orbitrap Exploris 480 mass spectrometer [23]. Also, Orwoll *et al.* identified 224 proteins in 1,196 serum samples using liquid chromatography-ion mobility-mass spectrometry (LC-IMS-MS) with a 58-min LC gradient [39]. In our study, using a 20-min DIA-MS method, 438 and 413 proteins were

identified, with 19% and 18% missing values, in the independent protein matrices of the discovery and the validation cohorts, respectively.

A combination of blood proteomics and physiological status in a longitudinal population study can be used for disease risk prediction, as it has been shown for various metabolic diseases with high incidence [40, 41]. We used this method to study MetS. In particular, we built a prospective machine learning model to predict the risk of developing MetS within ten years.

Elhadad *et al.* generated a LASSO model for the prospective analysis of incident MetS (mean follow-up time = 6.5) in a German population based on the data of 8 proteins [15]. Their training cohort of 623 participants achieved an AUC of 0.74, but no validation cohort was used. Our study used 11 serum proteins plus age as model features. The AUC of our model was 0.784 in the discovery cohort (855 individuals) and 0.774 in the validation cohort (242 individuals). Also, our resource has an additional follow-up time point of 10 years, providing more accurate data for the long-term prediction of MetS.

Our proteomics analysis provides essential insights into MetS, highlighting the apolipoproteins that play a crucial role in its pathophysiology. MetS is also known as insulin resistance syndrome [1], and apolipoproteins changes are associated with insulin resistance [9]. In particular, APOC3, a component of the triglyceride-rich very low-density lipoproteins (VLDL) and HDL in plasma [42], had been reported as a MetS risk factor in the Canadian [43] and Turkish populations [44]. Our results showed that apolipoproteins, APOC2, APOC3, APOA1, APOD, and APOF are not only associated with MetS onset but also with the development of MetS. Also, the expression of different apolipoproteins showed important variations in our data: APOC2 and APOC3 had high expression levels, while APOA1, APOD, and APOF had low expression levels. The higher expressions of APOC1, APOC2, APOC3, and APOE were already reported in MetS [45]. Therefore, exploring the biology of different apolipoproteins can provide more precise knowledge of the pathophysiology of MetS. In our data, APOC1, APOC2, APOC3, and APOE were up-regulated in the TG-based MetS, while APOD and APOF were down-regulated in HDL-based MetS. These observations suggest that apolipoproteins may have different roles and influences in MetS of different subtypes. Indeed, APOCs play an essential role in the etiology of human hyperlipidemias [46], especially APOC2 [47] and APOC3 [48]. Our data showed that these proteins may be related to MetS primarily by affecting TG. Circulating APOD and APOF are mainly present in HDL [49, 50] and may thus be related to MetS by affecting HDL.

In addition to apolipoproteins, we found that several other proteins are significantly associated with MetS development. Some of these proteins were already linked to MetS by other studies. SHBG can regulate the plasma metabolic clearance rate of steroid hormones and, in agreement with our findings, has been reported to predict the development of MetS in a population-based cohort study of middle-aged Finnish men [51]. Vitronectin is a multifunctional adhesive glycoprotein that exerts regulatory functions in coagulation, fibrinolysis, or the plasminogen activation system [52]. In a prospective study, baseline plasma vitronectin was found as a marker of incident MetS at nine years [53], consistently with our findings. Also, the serum levels of PEDF encoded by SERPINF1 (a glycoprotein that belongs to the superfamily of serine protease inhibitors) have been reported associated with MetS in a cross-sectional study of the Japanese population [54]. Furthermore, in a 10-year prospective study of Chinese men, PEDF was also associated with the development of the MetS [55]. Alpha-1-antitrypsin (A1AT) encoded by SERPINA1, is an acute-phase inflammation marker that is associated with the development of the

MetS [56]. AFM, a human plasma vitamin E-binding glycoprotein, was found to be strongly associated with the prevalence in a cross-sectional manner and development of MetS in a prospectively epidemiological study [57]. We also identified proteins that have not been reported to be directly associated with MetS. A1BG, belongs to the immunoglobulin family, was reported to be elevated in the urine of diabetes patients [58, 59]. Interestingly, the lower expression was observed in the serum of patients with MetS or those who develop MetS within 10 years, indicating that the protein may be related to the humoral regulatory system of MetS. Another immune-related protein, attractin, shows a higher level in circulating blood monocytes of human subjects because of obesity [60]. IGHV4-39, an immunoglobulin produced by B lymphocytes, has been reported to promote insulin resistance [61]. These results suggest that the immune system and coagulation related systems may also be important regulatory systems in the development of MetS.

The core mechanism of metabolic diseases, such as hypertension, hyperglycemia, hyperlipidemia, and obesity, which often appear together and develop into MetS, is still unclear. Here we explored the dysregulated proteins linked to the development of MetS. We found that specific dysregulated proteins were connected with various metabolic abnormal states and MetS development, providing a new basis for understanding this mechanism. Among these proteins, we found AGT, THBS1, XPNPEP1, adiponectin, and TAS1R3. AGT is an essential component of the renin-angiotensin system (RAS), and increased insulin stimulates AGT production [62]. We found a higher expression of AGT in the population that later developed MetS, suggesting that AGT positively correlates with the development of MetS consistently with other studies [63]. THBS1, an adipose-derived matricellular protein, has been reported as a biomarker of MetS in Japanese subjects [64]. Adiponectin is an anti-inflammatory cytokine and a biomarker for the development of MetS [9, 65], and it was downregulated in our data. A lower adiponectin level was observed in Chinese adolescents [66], peri- and post-menopausal women [67], young Polish subjects [68] with MetS, and in a prospective study of the Korean General Population with MetS [69]. Finally, TAS1R3 is a putative taste receptor reported to negatively correlate with glucose levels, triglycerides, and MetS [70]. XPNPEP1 contributes to the degradation of bradykinin, and the bradykinin system has beneficial effects on hypertension [71, 72] and type 2 diabetes [73]. This agrees with our data showing higher levels of XPNPEP1 in the populations that later developed MetS. These findings are again consistent with our results and corroborate the robustness of our findings, which include other proteins that have not yet been linked to MetS development.

The findings of this study have to be seen in the light of some limitations. First, we used a short gradient, high-throughput proteomics workflow, which is more suitable for analyzing large cohorts. Consequently, the serum proteins we identified are relatively high in abundance. Also, our discovery and validation cohort are from the same city. Our results, therefore, require a follow-up study using independent populations from different areas.

The GNHS community-based prospective cohort study was initiated in 2008. Comprehensive phenotypic data for the individuals involved in this study have been recorded over time, including but not limited to questionnaire, physical examination, dual-energy x-ray absorptiometry, ultrasonography, and blood/urine/faces tests [25-28]. The proteomic data acquired in this study could be potentially applied in understanding other health status and diseases such as diabetes.

In conclusion, we generated a resource of serum proteomics using DIA-MS based on a prospective population cohort with a 10-year follow-up. Using this data, we built a model for

predicting the risk of developing MetS within ten years. We also found new potential protein biomarkers of MetS that, together with their pathways, providing new perspectives on the pathophysiology of MetS.

MATERIALS AND METHODS

Patients and samples

This study was based on the previously reported community-based prospective cohort Guangzhou Nutrition and Health Study (GNHS, ClinicalTrials.gov identifier: NCT03179657) [25-28]. Briefly, 3840 of 4048 participants aged 40-83 were enrolled in this study. The participants were recruited between 2008 and 2013 and followed up until 2019 (**Figure 1, Table 1, Table S1A**). In particular, participants were evaluated and blood samples were taken from them at up to three time points: baseline (between 2008 and 2013), second follow-up (between 2014 and 2017), and third follow-up (between 2018 and 2019). Venous whole blood samples were collected from all participants early in the morning before having food using serum separation tubes, and then centrifuged at 3500 rpm for 10 min for serum collection. The serum samples were frozen at −80 °C before the analysis.

The diagnostic criteria for MetS in this study refer to the 2016 Chinese guidelines for managing dyslipidemia in adults [37]. Specifically, they included three or more of the following: central obesity/abdominal obesity: a waist circumference of at least 90 cm for men and at least 85 cm for women; hyperglycemia: fasting blood glucose of at least 6.10 mmol/L (110 mg/dL) or 2-h blood glucose after glycemic load of at least 7.80 mmol/L (140 mg/dL) or confirmed diabetes for patients that received treatment; hypertension: blood pressure of at least 130/85 mmHg or patients with hypertension that received treatment; fasting TG of at least 1.7 mmol/L (150 mg/dL); fasting HDL-C lower than 1.0 mmol/L (40 mg/dL).

Protein digestion

Peptides were extracted from the serum samples as previously described [74, 75]. Briefly, 1 µL of serum sample was lysed using 20 µL of lysis buffer (8 M urea (Sigma, Catalog # U1230) in 100 mM ammonium bicarbonate, ABB) at 32°C for 30 min, then reduced and alkylated using 10 mM tris (2-carboxyethyl) phosphine (Sigma Catalog # T4708) and 40 mM iodoacetamide (Sigma, Catalog # SLCD4031), respectively. Before the enzymatic digestion, 70 µL of 100 mM ABB were added to the samples to dilute the urea. The protein extracts were then digested with a two-step overnight tryptic digestion (Hualishi Tech. Ltd, Beijing, China), using an enzyme-to-substrate ratio of 1:60 (final ratio 1:30) at 32°C for 4 h and then another 12 h. The digestion was then stopped by adjusting the pH to 2-3 using 1% trifluoroacetic acid (Thermo Fisher Scientific, Catalog # T/3258/PB05). Before the MS analysis, peptides were cleaned with HRP SOLAu columns (Thermo Fisher Scientific™, San Jose, USA).

Mass spectrometric analysis

The peptide samples were injected into an Eksigent NanoLC 400 System (Eksigent, Dublin, CA, USA) coupled with a TripleTOF 5600 system (SCIEX, CA, USA) for the SWATH-MS analysis, as previously described [75]. Briefly, 0.5 µg of peptides were separated with a 20-min LC gradient of 5-30% buffer B. Buffer A contained 2% acetonitrile (ACN) and 0.1% formic acid (FA) in HPLC water, while buffer B contained 98% ACN and 0.1% FA in HPLC water. For SWATH-MS, a 55 variable Q1 isolation window scheme was set as in previous studies [75].

Proteome data analysis

The MS files were analyzed using DIA-NN (1.8) [76] against a plasma spectral library [75] containing 5102 peptides and 819 unique proteins from the Swiss-Prot database of *Homo sapiens*. In the DIA-NN settings, the software automatically sets the retention time extraction window, while the *m/z* extraction window for MS1 and MS2 was set to 20 ppm and 50 ppm, respectively.

Protein and peptide false discovery rates were set not to exceed 1%. Protein inference was set to the protein names (from the FASTA file), and the cross-run normalization was set as 'RT-dependent'.

Machine learning

In the machine learning analysis, we used LightGBM Library (Ver. 3.3.2) in Python (Ver. 3.9) to generate a machine learning model using the discovery dataset of 855 samples containing 288 serum proteins, as well as age and sex information. First, we randomly split the dataset into a training dataset (n=684) and an internal validation dataset (n=171). Then, a 2-step feature selection was used for choosing protein features from the discovery dataset and other features from the patient's clinical information (*i.e.* age, gender) to be included in the final classifier. The features were firstly selected based on proteome data and clinical data independently. In the second selection step, preselected features of the two categories were combined and any unrelated proteome was further eliminated, especially when the proteome was correlated with age and sex. The purpose of splitting the feature selection into two steps was to preserve proteomes that may have a weak contribution to classification, as including clinical information in the first place may result in the elimination of these proteomes due to their stronger contributions. During the first step, we used a Shapley value-based recursive feature elimination (RFE) algorithm, followed by a Shapely value-based Boruta algorithm for serum proteome feature selection, and the feature importance-based RFE for age and gender selection. In the second step, we combined preselected proteomes and clinical information from the first step and performed another Shapley value-based Boruta algorithm to obtain the final feature set. Through feature selection, 11 proteins and age were selected as the final features. We then optimized our model using the training dataset with ten-fold cross-validation and tested the model with the internal validation dataset and the independent validation cohort.

Statistical analysis

The linear mixed model for the baseline analysis used the data collected at the baseline visit, while the linear mixed model for the dynamic analysis used the data collected from all three time points. For the two linear mixed models, MetS and non-MetS were set as outcome variables; the protein expression, age, and sex were set as fixed effects; the patient ID was set as a random effect in the linear mixed model for dynamic analysis. The value β represents the regression coefficient, which indicates the influence of independent variable effect on dependent outcome variable in the regression equation. The larger the regression coefficient is, the greater the influence of effect on outcome variable is. The positive regression coefficient means that outcome variable increases with the increase of effect, while the negative regression coefficient means that outcome variable decreases with the increase of effect.

REFERENCES

1. DeFronzo, R.A. and E. Ferrannini, *Insulin resistance. A multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidemia, and atherosclerotic cardiovascular disease.* Diabetes Care, 1991. **14**(3): p. 173-94.
2. Expert Panel on Detection, E. and A. Treatment of High Blood Cholesterol in, *Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III).* JAMA, 2001. **285**(19): p. 2486-97.
3. Alberti, K.G.M.M., et al., *Harmonizing the Metabolic Syndrome.* Circulation, 2009. **120**(16): p. 1640-1645.
4. *International Diabetes Federation: The IDF consensus worldwide definition of the metabolic syndrome.*
5. Ranasinghe, P., et al., *Prevalence and trends of metabolic syndrome among adults in the asia-pacific region: a systematic review.* BMC Public Health, 2017. **17**(1): p. 101.
6. do Vale Moreira, N.C., et al., *Prevalence of Metabolic Syndrome by different definitions, and its association with type 2 diabetes, pre-diabetes, and cardiovascular disease risk in Brazil.* Diabetes Metab Syndr, 2020. **14**(5): p. 1217-1224.
7. Li, Y., et al., *Metabolic syndrome prevalence and its risk factors among adults in China: A nationally representative cross-sectional study.* PLoS One, 2018. **13**(6): p. e0199293.
8. National Center for Health Statistics, D.o.H.I.S., *Crude and age-adjusted percentage of civilian, noninstitutionalized adults with diagnosed diabetes, United States, 1980–2010.* . National Center for Chronic Disease Prevention and Health Promotion, Ed. Atlanta, GA, Centers for Disease Control and Prevention, Division of Diabetes Translation, 2012.
9. Eckel, R.H., S.M. Grundy, and P.Z. Zimmet, *The metabolic syndrome.* Lancet, 2005. **365**(9468): p. 1415-28.
10. Alshehri, A.M., *Metabolic syndrome and cardiovascular risk.* Journal of family & community medicine, 2010. **17**(2): p. 73-78.
11. Esposito, K., et al., *Metabolic syndrome and risk of cancer: a systematic review and meta-analysis.* Diabetes care, 2012. **35**(11): p. 2402-2411.
12. Gold, L., et al., *Aptamer-based multiplexed proteomic technology for biomarker discovery.* PLoS One, 2010. **5**(12): p. e15004.
13. Assarsson, E., et al., *Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability.* PLoS One, 2014. **9**(4): p. e95192.
14. Ngo, D., et al., *Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease.* Circulation, 2016. **134**(4): p. 270-85.
15. Elhadad, M.A., et al., *Metabolic syndrome and the plasma proteome: from association to causation.* Cardiovasc Diabetol, 2021. **20**(1): p. 111.
16. Ganz, P., et al., *Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease.* JAMA, 2016. **315**(23): p. 2532-41.
17. Lehallier, B., et al., *Undulating changes in human plasma proteome profiles across the lifespan.* Nat Med, 2019. **25**(12): p. 1843-1850.
18. Ponce-de-Leon, M., et al., *Novel associations between inflammation-related proteins and adiposity: A targeted proteomics approach across four population-based studies.* Transl Res,

2022. **242**: p. 93-104.

19. Ferreira, J.P., et al., *Multi-proteomic approach to predict specific cardiovascular events in patients with diabetes and myocardial infarction: findings from the EXAMINE trial.* Clin Res Cardiol, 2021. **110**(7): p. 1006-1019.
20. Zhu, Y., et al., *SnapShot: Clinical proteomics.* Cell, 2021. **184**(18): p. 4840-4840 e1.
21. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Mol Cell Proteomics, 2012. **11**(6): p. O111 016717.
22. Guo, T., et al., *Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps.* Nat Med, 2015. **21**(4): p. 407-13.
23. Niu, L., et al., *Noninvasive proteomic biomarkers for alcohol-related liver disease.* Nat Med, 2022. **28**(6): p. 1277-1287.
24. Bruderer, R., et al., *Analysis of 1508 Plasma Samples by Capillary-Flow Data-Independent Acquisition Profiles Proteomics of Weight Loss and Maintenance.* Mol Cell Proteomics, 2019. **18**(6): p. 1242-1254.
25. Jiang, Z., et al., *Dietary fruit and vegetable intake, gut microbiota, and type 2 diabetes: results from two large human cohort studies.* BMC Med, 2020. **18**(1): p. 371.
26. Miao, Z., et al., *Erythrocyte n-6 Polyunsaturated Fatty Acids, Gut Microbiota, and Incident Type 2 Diabetes: A Prospective Cohort Study.* Diabetes Care, 2020. **43**(10): p. 2435-2443.
27. Gou, W., et al., *Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes.* Diabetes Care, 2021. **44**(2): p. 358-366.
28. Gou, W., et al., *Gut microbiota, inflammation, and molecular signatures of host response to infection.* J Genet Genomics, 2021. **48**(9): p. 792-802.
29. Demichev, V., et al., *DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput.* Nature Methods, 2020. **17**(1): p. 41-44.
30. Zhu, T., et al., *ProteomeExpert: a docker image based web-server for exploring, modeling, visualizing, and mining quantitative proteomic data sets.* Bioinformatics, 2021.
31. Desvergne, B., *RXR: from partnership to leadership in metabolic regulations.* Vitam Horm, 2007. **75**: p. 1-32.
32. Shulman, A.I. and D.J. Mangelsdorf, *Retinoid x receptor heterodimers in the metabolic syndrome.* N Engl J Med, 2005. **353**(6): p. 604-15.
33. McNeill, A.M., et al., *The metabolic syndrome and 11-year risk of incident cardiovascular disease in the atherosclerosis risk in communities study.* Diabetes Care, 2005. **28**(2): p. 385-90.
34. Thanabalasingham, G., et al., *Systematic assessment of etiology in adults with a clinical diagnosis of young-onset type 2 diabetes is a successful strategy for identifying maturity-onset diabetes of the young.* Diabetes Care, 2012. **35**(6): p. 1206-12.
35. Pickup, J.C., et al., *NIDDM as a disease of the innate immune system: association of acute-phase reactants and interleukin-6 with metabolic syndrome X.* Diabetologia, 1997. **40**(11): p. 1286-92.
36. Rizza, R.A., L.J. Mandarino, and J.E. Gerich, *Effects of growth hormone on insulin action in man. Mechanisms of insulin resistance, impaired suppression of glucose production, and impaired stimulation of glucose utilization.* Diabetes, 1982. **31**(8 Pt 1): p. 663-9.
37. Joint committee for guideline, r., *2016 Chinese guidelines for the management of dyslipidemia in adults.* J Geriatr Cardiol, 2018. **15**(1): p. 1-29.

38. Ruderman, N. and M. Prentki, *AMP kinase and malonyl-CoA: targets for therapy of the metabolic syndrome.* Nat Rev Drug Discov, 2004. **3**(4): p. 340-51.
39. Orwoll, E.S., et al., *Proteomic assessment of serum biomarkers of longevity in older men.* Aging Cell, 2020. **19**(11): p. e13253.
40. Schussler-Fiorenza Rose, S.M., et al., *A longitudinal big data approach for precision health.* Nat Med, 2019. **25**(5): p. 792-804.
41. Robbins, J.M., et al., *Human plasma proteomic profiles indicative of cardiorespiratory fitness.* Nat Metab, 2021. **3**(6): p. 786-797.
42. Yao, Z. and Y. Wang, *Apolipoprotein C-III and hepatic triglyceride-rich lipoprotein production.* Curr Opin Lipidol, 2012. **23**(3): p. 206-12.
43. Pollex, R.L., et al., *Metabolic syndrome in aboriginal Canadians: prevalence and genetic associations.* Atherosclerosis, 2006. **184**(1): p. 121-9.
44. Onat, A., et al., *Apolipoprotein C-III, a strong discriminant of coronary risk in men and a determinant of the metabolic syndrome in both genders.* Atherosclerosis, 2003. **168**(1): p. 81-9.
45. Savinova, O.V., et al., *Reduced apolipoprotein glycosylation in patients with the metabolic syndrome.* PLoS One, 2014. **9**(8): p. e104833.
46. Jong, M.C., M.H. Hofker, and L.M. Havekes, *Role of ApoCs in lipoprotein metabolism: functional differences between ApoC1, ApoC2, and ApoC3.* Arterioscler Thromb Vasc Biol, 1999. **19**(3): p. 472-84.
47. Breckenridge, W.C., et al., *Hypertriglyceridemia associated with deficiency of apolipoprotein C-II.* N Engl J Med, 1978. **298**(23): p. 1265-73.
48. Carlson, L.A. and D. Ballantyne, *Changing relative proportions of apolipoproteins CII and CIII of very low density lipoproteins in hypertriglyceridaemia.* Atherosclerosis, 1976. **23**(3): p. 563-8.
49. McConathy, W.J. and P. Alaupovic, *Isolation and partial characterization of apolipoprotein D: a new protein moiety of the human plasma lipoprotein system.* FEBS Lett, 1973. **37**(2): p. 178-82.
50. Olofsson, S.O., W.J. McConathy, and P. Alaupovic, *Isolation and partial characterization of a new acidic apolipoprotein (apolipoprotein F) from high density lipoproteins of human plasma.* Biochemistry, 1978. **17**(6): p. 1032-6.
51. Laaksonen, D.E., et al., *Testosterone and sex hormone-binding globulin predict the metabolic syndrome and diabetes in middle-aged men.* Diabetes Care, 2004. **27**(5): p. 1036-41.
52. Preissner, K.T., *Structure and biological role of vitronectin.* Annu Rev Cell Biol, 1991. **7**: p. 275-310.
53. Alessi, M.C., et al., *Association of vitronectin and plasminogen activator inhibitor-1 levels with the risk of metabolic syndrome and type 2 diabetes mellitus. Results from the D.E.S.I.R. prospective cohort.* Thromb Haemost, 2011. **106**(3): p. 416-22.
54. Yamagishi, S., et al., *Elevated serum levels of pigment epithelium-derived factor in the metabolic syndrome.* J Clin Endocrinol Metab, 2006. **91**(6): p. 2447-50.
55. Chen, C., et al., *Plasma level of pigment epithelium-derived factor is independently associated with the development of the metabolic syndrome in Chinese men: a 10-year prospective study.* J Clin Endocrinol Metab, 2010. **95**(11): p. 5074-81.
56. Setoh, K., et al., *Three missense variants of metabolic syndrome-related genes are associated with alpha-1 antitrypsin levels.* Nat Commun, 2015. **6**: p. 7754.

57. Kronenberg, F., et al., *Plasma concentrations of afamin are associated with the prevalence and development of metabolic syndrome.* Circ Cardiovasc Genet, 2014. **7**(6): p. 822-9.
58. Soggiu, A., et al., *A discovery-phase urine proteomics investigation in type 1 diabetes.* Acta Diabetol, 2012. **49**(6): p. 453-64.
59. Gourgari, E., et al., *Proteomic alterations of HDL in youth with type 1 diabetes and their associations with glycemic control: a case-control study.* Cardiovasc Diabetol, 2019. **18**(1): p. 43.
60. Laudes, M., et al., *Dipeptidyl-peptidase 4 and attractin expression is increased in circulating blood monocytes of obese human subjects.* Exp Clin Endocrinol Diabetes, 2010. **118**(8): p. 473-7.
61. Winer, D.A., et al., *B cells promote insulin resistance through modulation of T cells and production of pathogenic IgG antibodies.* Nat Med, 2011. **17**(5): p. 610-7.
62. Harte, A.L., et al., *Insulin increases angiotensinogen expression in human abdominal subcutaneous adipocytes.* Diabetes Obes Metab, 2003. **5**(6): p. 462-7.
63. Aubert, J., et al., *Insulin down-regulates angiotensinogen gene expression and angiotensinogen secretion in cultured adipose cells.* Biochem Biophys Res Commun, 1998. **250**(1): p. 77-82.
64. Matsuo, Y., et al., *Thrombospondin 1 as a novel biological marker of obesity and metabolic syndrome.* Metabolism, 2015. **64**(11): p. 1490-9.
65. Matsuzawa, Y., et al., *Adiponectin and metabolic syndrome.* Arterioscler Thromb Vasc Biol, 2004. **24**(1): p. 29-33.
66. Li, P., et al., *Correlation of serum adiponectin and adiponectin gene polymorphism with metabolic syndrome in Chinese adolescents.* Eur J Clin Nutr, 2015. **69**(1): p. 62-7.
67. Wattanapol, P., P. Vichinsartvichai, and P. Sakoonwatanyoo, *Serum adiponectin is a potential biomarker for metabolic syndrome in peri-and postmenopausal women.* Gynecol Endocrinol, 2020. **36**(7): p. 620-625.
68. Kowalska, I., et al., *Insulin resistance, serum adiponectin, and proinflammatory markers in young subjects with the metabolic syndrome.* Metabolism, 2008. **57**(11): p. 1539-44.
69. Kim, J.Y., et al., *Prospective study of serum adiponectin and incident metabolic syndrome: the ARIRANG study.* Diabetes Care, 2013. **36**(6): p. 1547-53.
70. Bertran, L., et al., *Expression of Jejunal Taste Receptors in Women with Morbid Obesity.* Nutrients, 2021. **13**(7).
71. Sharma, J.N., *Hypertension and the bradykinin system.* Curr Hypertens Rep, 2009. **11**(3): p. 178-81.
72. Agarwal, R., *Bradykinin and inhibition of angiotensin-converting enzyme in hypertension.* N Engl J Med, 1999. **340**(12): p. 967-8; author reply 968-9.
73. Mandle, R.J., R.W. Colman, and A.P. Kaplan, *Identification of prekallikrein and high-molecular-weight kininogen as a complex in human plasma.* Proc Natl Acad Sci U S A, 1976. **73**(11): p. 4179-83.
74. Shen, B., et al., *Proteomic and Metabolomic Characterization of COVID-19 Patient Sera.* Cell, 2020. **182**(1): p. 59-72 e15.
75. Zhang, Y., et al., *Potential Use of Serum Proteomics for Monitoring COVID-19 Progression to Complement RT-PCR Detection.* J Proteome Res, 2022. **21**(1): p. 90-100.
76. Demichev, V., et al., *DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput.* Nat Methods, 2020. **17**(1): p. 41-44.

DATA AVAILIABILITY

All the raw data will be publicly released after publication. The phenotype data can be requested by email from the corresponding author (Y.C).

ACKNOWLEDGMENTS

This work is supported by grants from the National Key R&D Program of China (No. 2021YFA1301602, 2021YFA1301603, and 2021YFA1301601), the National Natural Science Foundation of China (No. 82073546, 81773416, and 82103826) and Zhejiang Provincial Natural Science Foundation of China (LQ21H260002). We thank Westlake University Supercomputer Center for assistance in data generation and storage. We thank all study participants of the Guangzhou Nutrition and Health Study and all other team members involved in the cohort study.

AUTHOR CONTRIBUTIONS

T.G., J.Z, Y.C., and Y.Z. designed and supervised the project. Y.C. and F.Z. established the cohort and collected the phenotype data and the samples. X.C., J.T., L.Y., Y.X., Z.M., W.G., Y.F., S.L., M.S., K.Z., F.X., Y.T., N.X., and Y.Z. generated the data. X.C., Z.X., B.W., W.G., and J.G. analyzed the data. X.C., L.Y., P.S., Y.Z., and T.G. drafted the manuscript with inputs from all co-authors. X.C., Z.X, F.Z., J.T., L.Y., and B.W. contribute equally to this work. We thank Z.R., Y.L., H.C., Z.L., B.W., and H.M for assistance in data generation and analysis.

DECLARATION OF INTERESTS

Y.Z. and T.G. are shareholders of Westlake Omics Inc. B.W., W.G., and N.X. are employees of Westlake Omics Inc. The other authors declare no competing interests.

TABLES

Table 1. Statistics of the 3,840 participants from the Guangzhou Nutrition and Health Study (GNHS).

	Baseline (N=3480)	Second follow-up (N=2639)	Third follow-up (N=1773)
Age (years)	57.4 (38.2-80.4)	63 (44.3-83.3)	66.1 (46.8-86.2)
Sex (% of women)	69	68	69
BMI (kg/m^2)	23.2 (14.8-51.3)	23.5 (14.5-57.3)	23.6(15.3-54.1)
Waist (men, cm)	86.4 (61-147)	87.8 (64.4-155.8)	88 (64.5-162)
Waist (women, cm)	81.2 (58-117.9)	86.5 (60.3-128,6)	83.6 (60.2-113.8)
Fasting serum glucose (mmol/L)	4.7 (0.2-24.6)	5 (3.2-24.6)	5.3 (3.9-21.1)
Blood triglycerides (mmol/L)	1.3 (0.01-25.4)	1.3 (0.2-15.6)	1.4 (0.45-15)
Serum HDL (mmol/L)	1.4 (0.02-3.3)	1.5 (0.6-4.6)	1.4 (0.6-4.3)
Serum LDL (mmol/L)	3.6 (0.1-10)	3.6 (0.6-11.5)	3.5 (0.9-8)
SBP	122 (73-228)	122 (63.5-241)	120 (76-205.7)
DBP	79 (46-135)	73.3 (43.7-120)	72.5 (47.7-117.3)
SBP/DBP (mmHg)	1.6 (1-2.7)	1.6 (0.9-2.8)	1.6 (1-2.5)
MetS (%)	0.17	0.21	0.18

FIGURES

Figure 1 Study design. GNHS study population: 7890 samples (4796 in the discovery cohort, 3094 in the validation cohort) were collected from 3840 individuals. Altogether we generated 18,970 DIA-MS data. After data cleaning, the protein matrices (containing 4637 samples and 438 proteins for the discovery cohort; 3067 samples and 413 proteins for the validation cohort) were used for data mining and machine learning modeling.

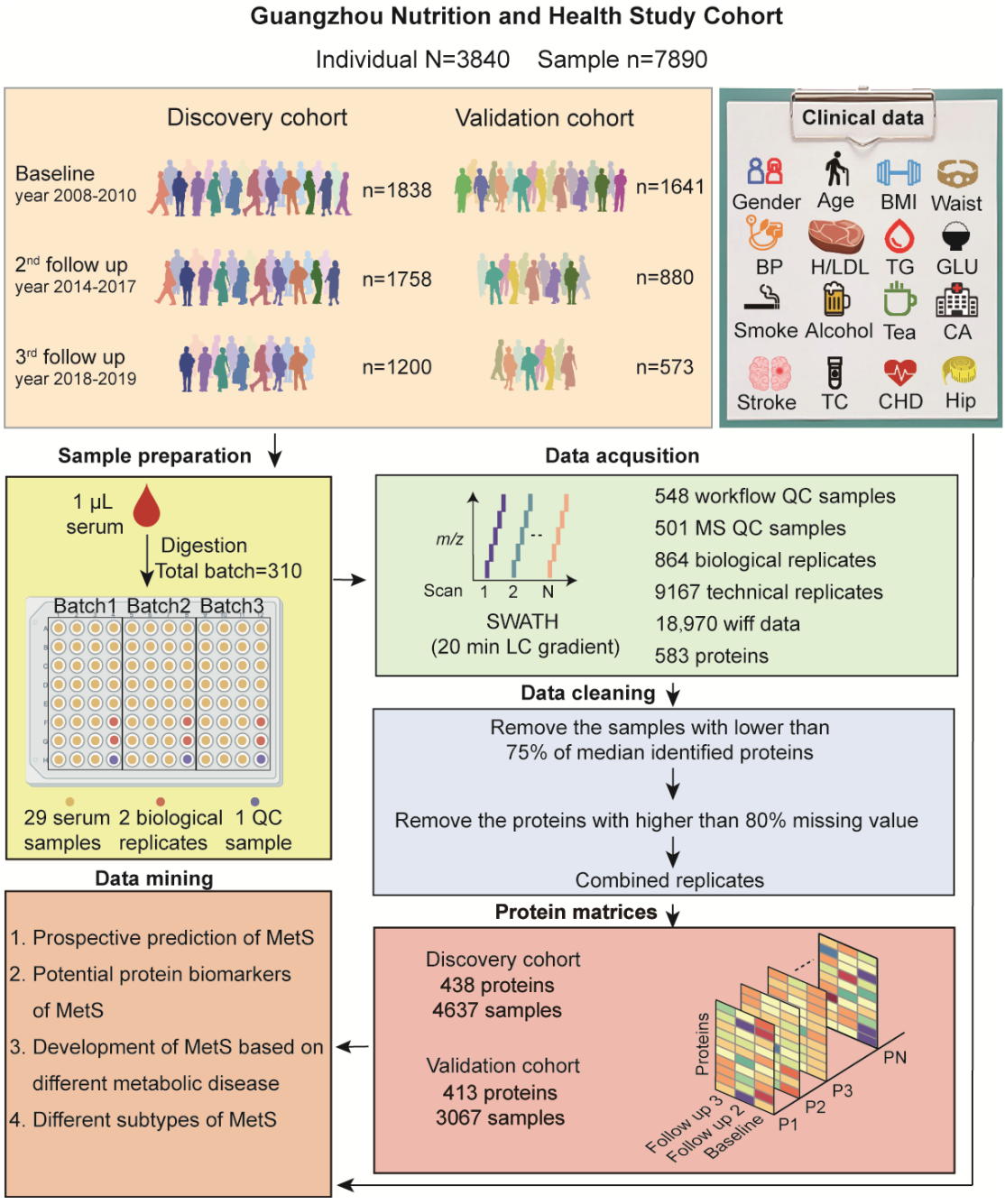


Figure 2 Quality control analyses and dataset pre-processing. A) Pearson correlation of the MS QC samples from the discovery and validation cohorts. **B)** Batch effect evaluation using PCA and the QC samples from the discovery and validation cohorts. **C)** Pearson correlation of the technical and biological replicates from the discovery and validation cohorts.

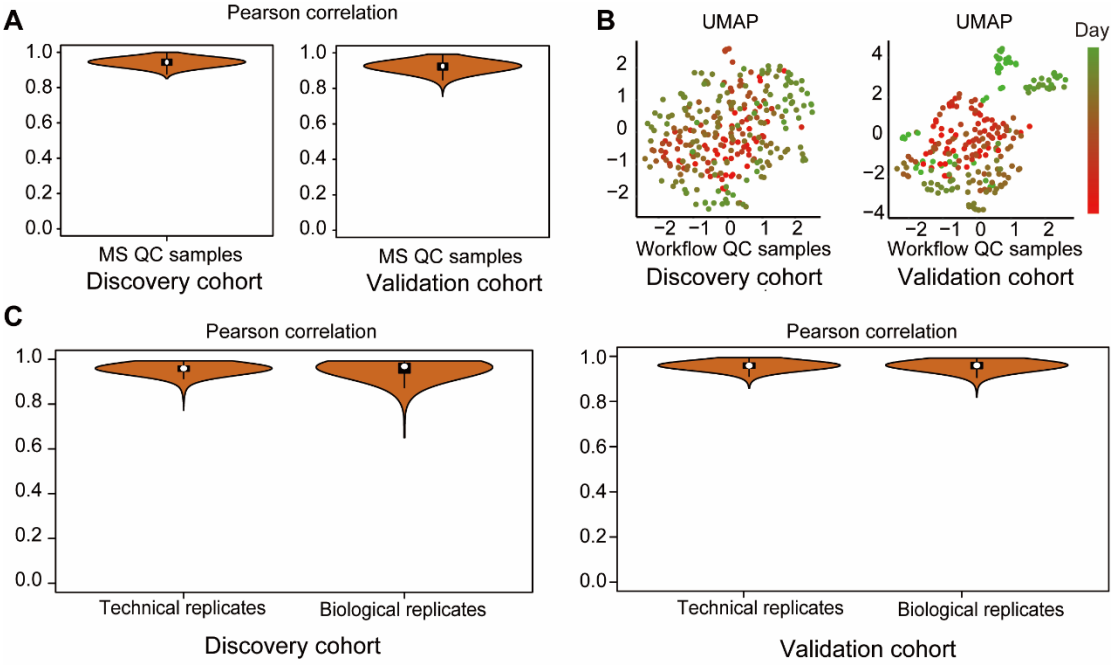


Figure 3 Machine learning model for predicting the risk of developing MetS. A) Case and control samples for the machine learning model. **B)** Workflow of the machine learning model built with quantitative proteomics data and clinical features. **C)** Shapley importance of variables in the model. **D)** Receiver operating characteristic (ROC) plot of the machine learning model.

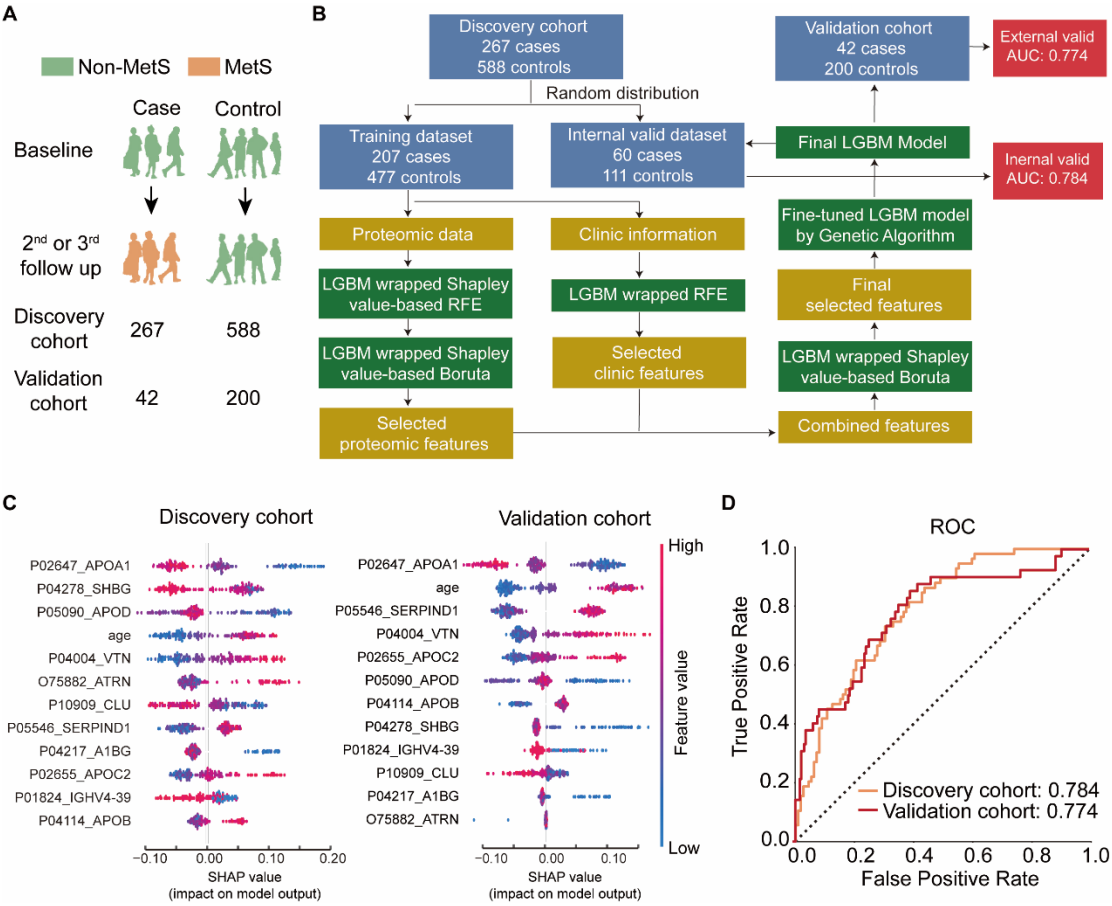


Figure 4 Performance of new potential protein biomarkers of MetS and their related pathways. A) Dysregulated proteins highly associated with MetS were selected using linear mixed models. **B)** Enriched pathways analyzed by IPA of 175 and 194 significantly dysregulated proteins. **C)** Top (|β|>0.2) significantly dysregulated proteins associated with MetS using the linear mixed model. **D)** Expression of typical dysregulated proteins in discovery cohort and validation cohort.

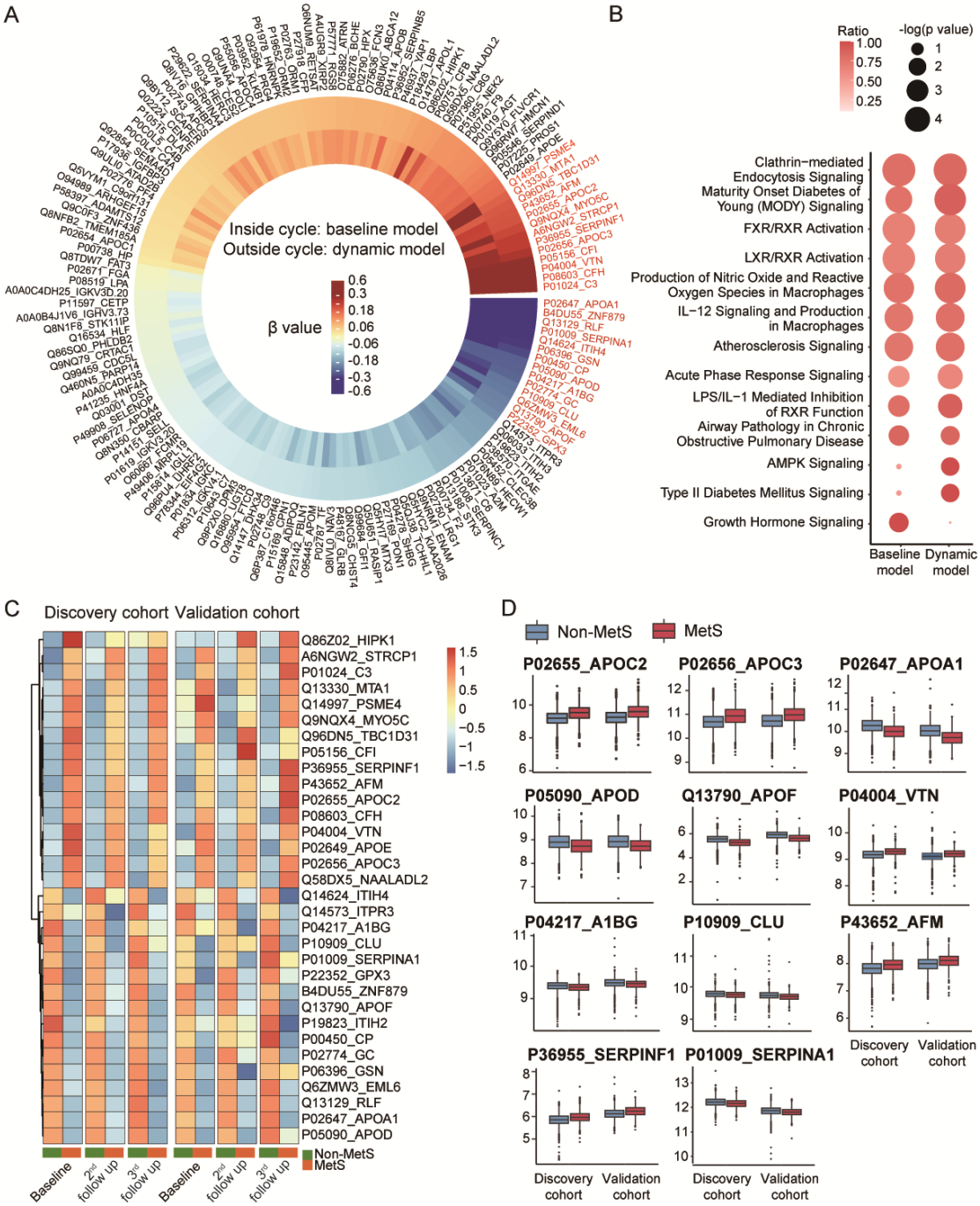


Figure 5 Dysregulated proteins in different types of MetS. A) Volcano plot of the dysregulated proteins in the TG-based MetS. **B**) Volcano plot of the dysregulated proteins in the HDL-based MetS. **C**) Expression of the dysregulated proteins in the TG-based MetS and the non-MetS. **D**) Expression of the dysregulated proteins in the TG-based MetS and the non-MetS.

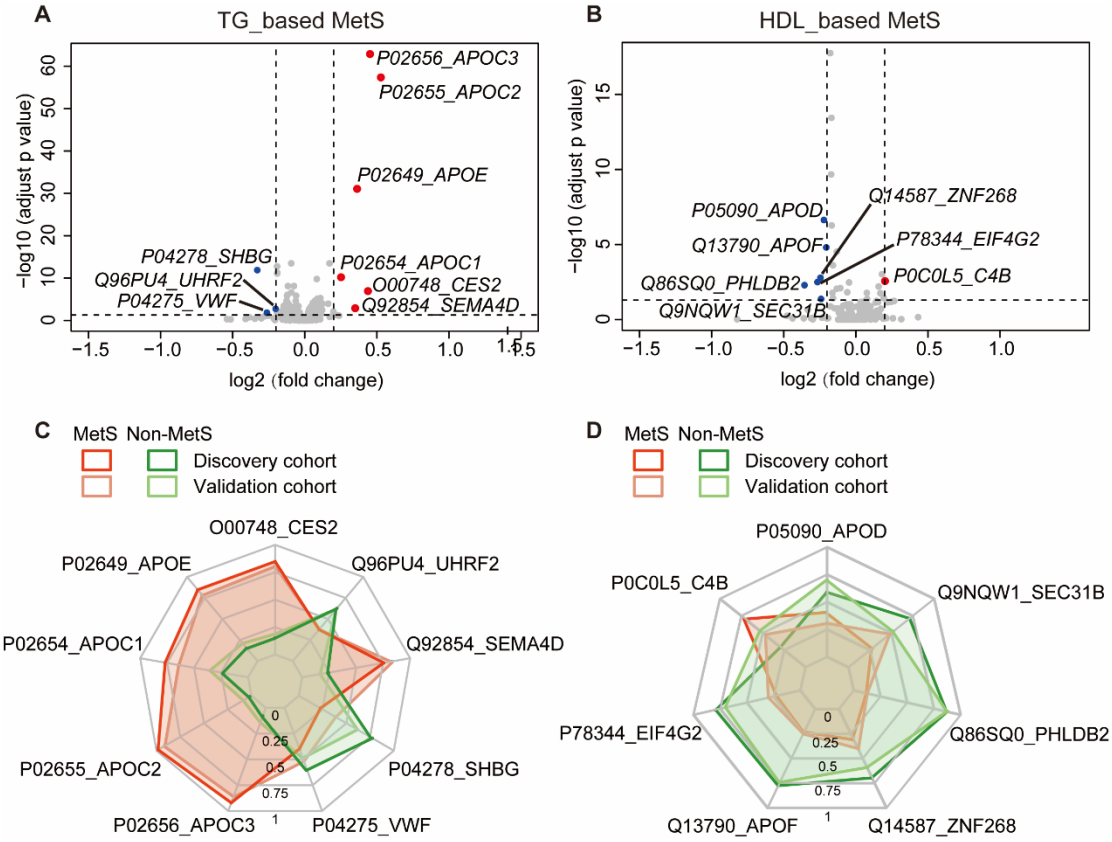
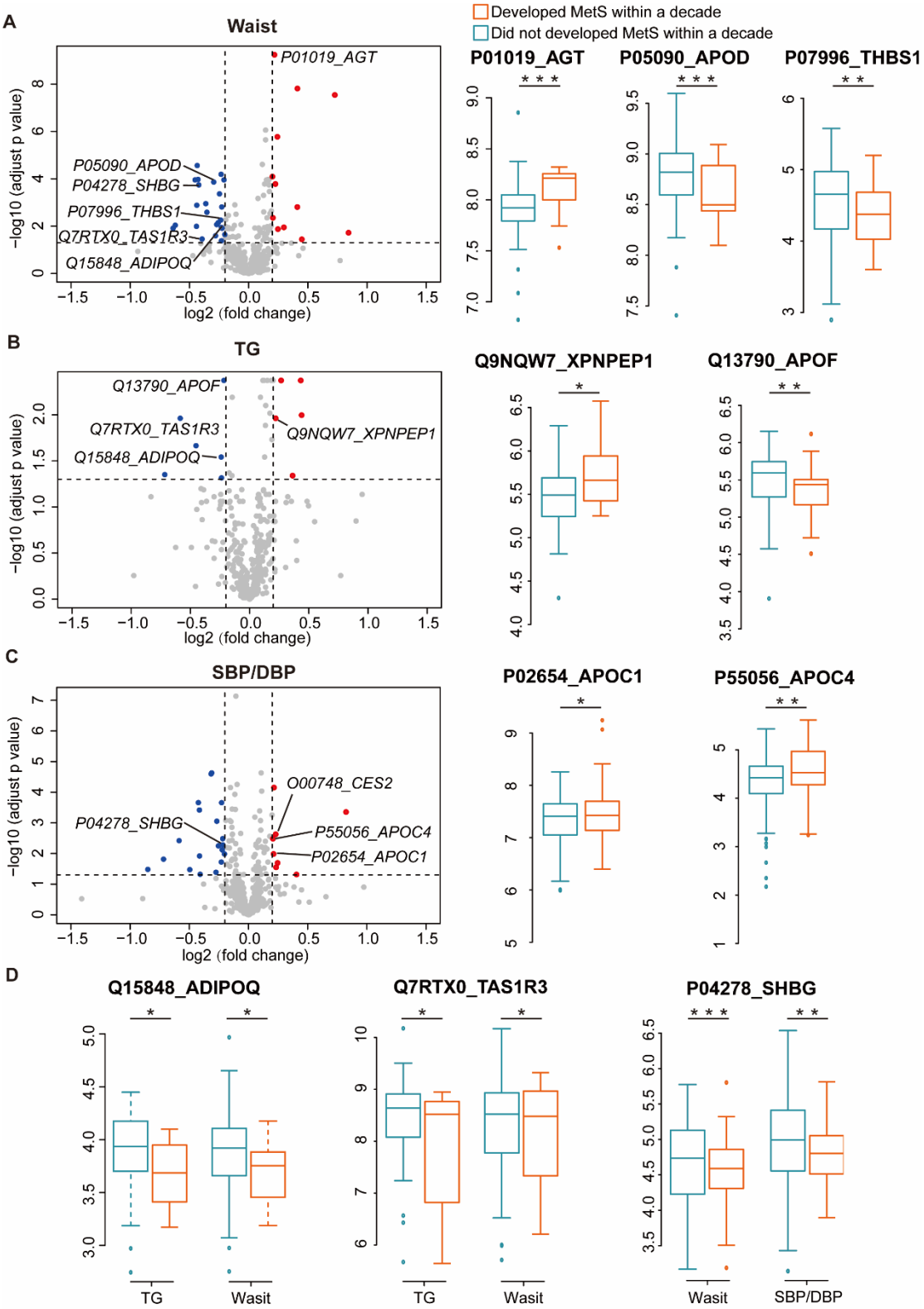


Figure 6 Dysregulated proteins in the development of MetS. A) Volcano plot and boxplots of the dysregulated proteins in the development of waist-based MetS. **B**) Volcano plot and boxplots of the dysregulated proteins in the development of TG-based MetS. **C**) Volcano plot and boxplots of the dysregulated proteins in the development of SBP/DBP-based MetS. **D**) Boxplots of the overlapping core dysregulated proteins in the development of waist-based MetS, TG-based MetS, and SBP/DBP-based MetS. Adjust p value: *, <0.05; **, <0.01; ***, <0.001.



SUPPLEMENTAL INFORMATION

Figure S1. Dysregulated proteins from the glucose-based MetS patients. A) Volcano plot of the dysregulated proteins from the glucose-based MetS patients. **B**) Expression of the dysregulated proteins from the glucose-based MetS and the non-MetS patients.

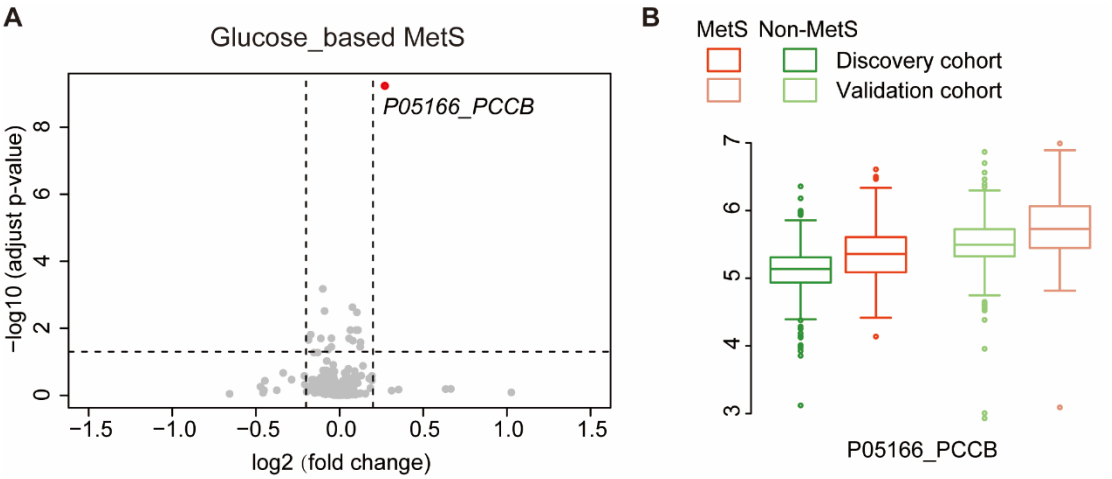


Figure S2. Network enriched using the dysregulated protein from the TG-based MetS patients.

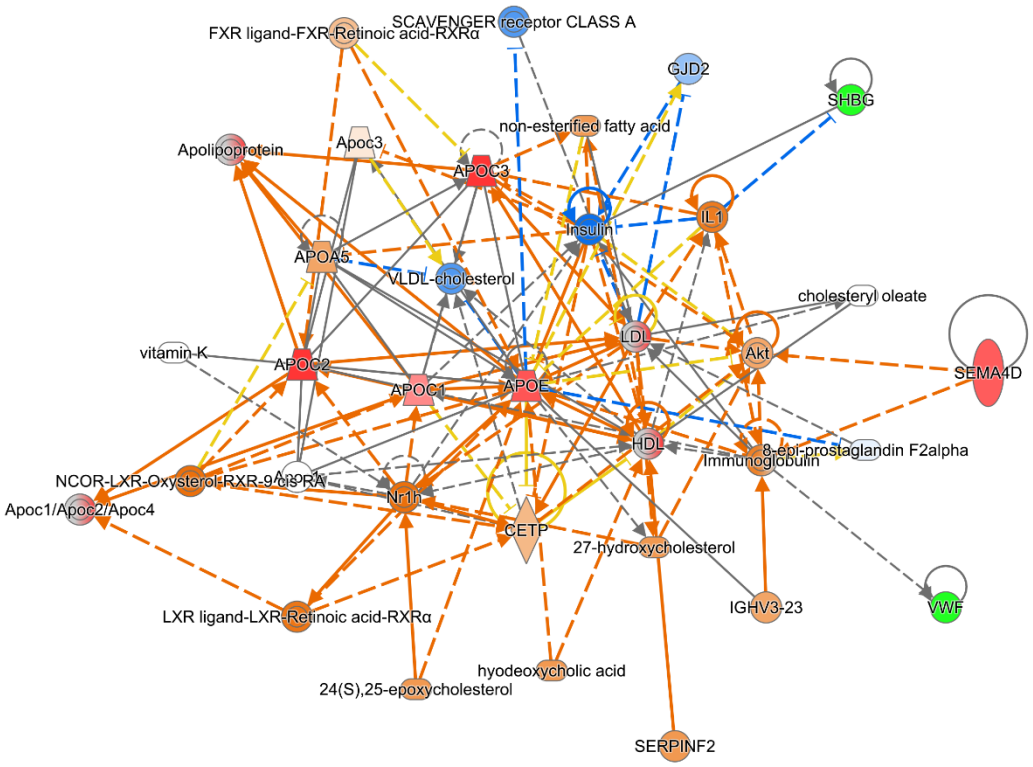


Figure S3. Network enriched using the dysregulated protein from the HDL-based MetS patients.