Int J Clin Exp Pathol 2017;10(2):1172-1185
www.ijcep.com /ISSN:1936-2625/IJCEP0042636

Original Article

Data-independent acquisition strategy for the serum proteomics of tuberculosis

Lijun Zhong[1*], Yuan Li[2*], Huifang Tian[3], Lijuan Guo[4], Shibing Qin[2], Jing Shen[3]

[1]Medical and Health Analytical Center, Peking University Health Science Center, Beijing, P. R. China; [2]Department of Orthopaedics, Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Tumor Institute, Beijing, P. R. China; [3]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Central Laboratory, Peking University Cancer Hospital & Institute, Beijing, P. R. China; [4]Clinical Laboratory of China Meitan General Hospital, Beijing, P. R. China. [*]Equal contributors.

Received October 23, 2016; Accepted October 27, 2016; Epub February 1, 2017; Published February 15, 2017

Abstract: Tuberculosis (TB) is a severe infectious disease caused by mycobacterium tuberculosis. Early and reliable diagnosis of this disease is very important. In this study, proteomic profiling analysis was performed for serum samples from TB patients and healthy controls. The samples were analyzed by data-independent acquisition mass spectrometry coupled with high performance liquid chromatography (HPLC-DIA-MS) to identify candidate serum biomarkers that could provide clues for TB diagnosis. A total of 647 serum proteins were identified using DIA acquisition strategy, and statistical analysis showed that 88 proteins were significantly dysregulated between TB and control groups. Furthermore, bioinformatic analysis was used to reveal TB relevant pathways and regulative networks for pathology study. As a result, a protein-protein interaction network was constructed to reveal the relationship between the 88 dysregulated proteins, and the pathway of complement and coagulation cascades and ECM-receptor interaction were significantly relevant to TB disease state. Finally, to further assess the validity of these findings, LRG1 was selected for subsequent ELISA assays, and the ELISA result validated the reliability of this proteomic analysis. In summary, our findings reveal several potential serum biomarkers for TB diagnosis, and the results of proteomic and bioinformatic analysis can also provide valuable information for TB pathology studies.

Keywords: Serum proteomics, data independent acquisition, tuberculosis, biomarkers, bioinformatic analysis

Introduction

Proteomics is the collective study of all expressed proteins in given samples (cells, tissues and serum, etc.) [1, 2]. The proteomic study can reveal information on not only the expressed proteins, but also the interaction and regulation pattern of protein complex and signaling networks. Nowadays, proteomic studies are mainly performed using liquid chromatography coupled tandem mass spectrometers (LC-MS/MS) [3, 4]. These platforms can measure the abundance of thousands of proteins from complex biological samples simultaneously [5]. While isotopic labeling of proteins (iTRAQ, TMT and SILAC) can achieve more accurate quantitative measurement [6-8], label-free quantitative proteomics is also a popular strategy because it has simple sample preparation procedures which are easily accessible. For acquisition strategies used for LC-MS analysis, the data dependent acquisition (DDA) strategy is a common one. But drawbacks of DDA strategy exist, including less quantification accuracy and reproducibility. To solve this problem, data-independent acquisition (DIA) strategy has been proposed and developed for proteomics [9]. In contrast to the traditional DDA strategy, DIA strategy can theoretically obtain all fragment ions for all precursors simultaneously, thereby increasing the coverage of detected proteins and improving the analytical reliability [10, 11].

Bioinformatic analysis has proven to be powerful mathematical artifacts for studying complex systems such as the regulative networks of proteins in biology [12]. It is a powerful tool for interpreting the complex proteomic results [13]. Bioinformatic analysis at the pathway level has become a common step when analyzing the proteomics data, for example, Tisoncik-Go et al.

studied the protein expression profiles in respiratory compartments of ferrets infected with influenza viruses, and the integrative bioinformatics analysis of the data uncovered relationships between host responses and phenotypic outcomes of viral infection [14]. Deshmukh and colleagues used bioinformatic tools to analyze proteomic data which including the information of over 10,000 proteins, revealing that complex regulation of AMP-activated protein kinase and insulin signaling in muscle tissue at the level of enzyme isoforms [15]. The protein-protein interaction (PPI) network is one of the commonly used analytical strategies when performing bioinformatic analysis for proteomic results [16, 17]. For network construction, the differently dysregulated proteins (seed proteins) were used as queries firstly. Then the experimentally supported hyperlinks from databases such as DIP, BIOGRID, HPRD, BIND, MINT and INTACT [18, 19] are retrieved and connected between the seed proteins, and finally the regulative network is visualized in a proper style. Another analytical strategy is the gene ontology (GO) analysis [20], which can enrich the given proteins (or genes) in terms of biological processes, cellular component and molecular function, thus providing a comprehensive view of the given proteins' roles in a cell.

Tuberculosis (TB) results in an estimated 1.7 million deaths each year and the global number of new cases continues to increase [21]. Biomarkers are indispensable to disease diagnosis and to the development of new TB therapeutics and vaccines [22]. In the present study, we performed untargeted proteomic profiling analysis to reveal candidate serum biomarkers for TB. Bioinformatic analysis was also performed to retrieve TB relevant signaling pathways and regulative networks. Our study can provide not only several candidate biomarkers for diagnosis of TB disease state, but also clues for the pathology studies of TB.

Materials and methods

Materials

Ammonium bicarbonate, sodium deoxycholate, iodoacetamide, and dithiothreitol were purchased from Sigma (St. Louis, MO, USA). Tris-(2-carboxyethyl) phosphine was acquired from Thermo Scientific (Rockford, Il, USA). Modified sequencing-grade trypsin was obtained from Promega (Madison, WI, USA). All mobile phases and solutions were prepared with HPLC grade solvents (i.e. water, acetonitrile, methanol, and formic acid) from Sigma Aldrich. All other reagents were from commercial suppliers and of standard biochemical quality.

Patients

From November 2014 to December 2015, patients from Beijing Chest Hospital, Beijing, China, who were confirmed by Mycobacterium culturing (Lowenstein-Jensen medium), were recruited in our study. This protocol was approved by the Ethics Committee at Beijing Chest Hospital and informed consent was obtained from each patient. One control group (n=12) was also established using healthy volunteers, who were without latent tuberculosis infection (healthy volunteers received tuberculin skin test (TST) and all the results of TST were negative). People in all the groups were well matched in age and gender.

HAPs immunodepletion

The Seppro IgY14 Spin Column (Sigma, USA) was used to bind human serum HSA, IgG, fibrinogen, transferrin, IgA, IgM, haptoglobin, alpha2-macroglobulin, alpha1-acid glycoprotein, alpha1-antitrypsin, Apo A-I HDL, Apo A-II HDL, complement C3 and LDL (ApoB). In accordance with the manufacturer's recommendations, 10 μl of a crude plasma sample was diluted with Tris-buffered saline (TBS, 10 mM Tris-HCl with 150 mM NaCl, pH 7.4) and injected into the spin column. Then the beads and the sample were completely mixed by inversion and shaking the column, and the mixture was incubated at room temperature for 15 minutes. The LAP was collected by centrifugation the column for 30 seconds at 2,000 rpm, followed by wash using 1 mL dilution buffer (TBS, 10 mM Tris-HCl with 150 mM NaCl, pH 7.4). Then the column was washed twice using stripping buffer (0.1 M Glycine-HCl, pH 2.5), neutralization buffer (0.1 M Tris-HCl, pH 8.0), and finally balanced by the dilution buffer. The proteins in the flow-through fraction (LAPs) were collected and concentrated with a 10 KD ultrafiltration tube (Millipore, USA).

Protein digestion

Protein samples were digested according to the manufacturer's protocol for filter-aided sample preparation (FASP) [23]. In brief, protein concentrates in Vivacon 500 filtrate tube (Cat No.

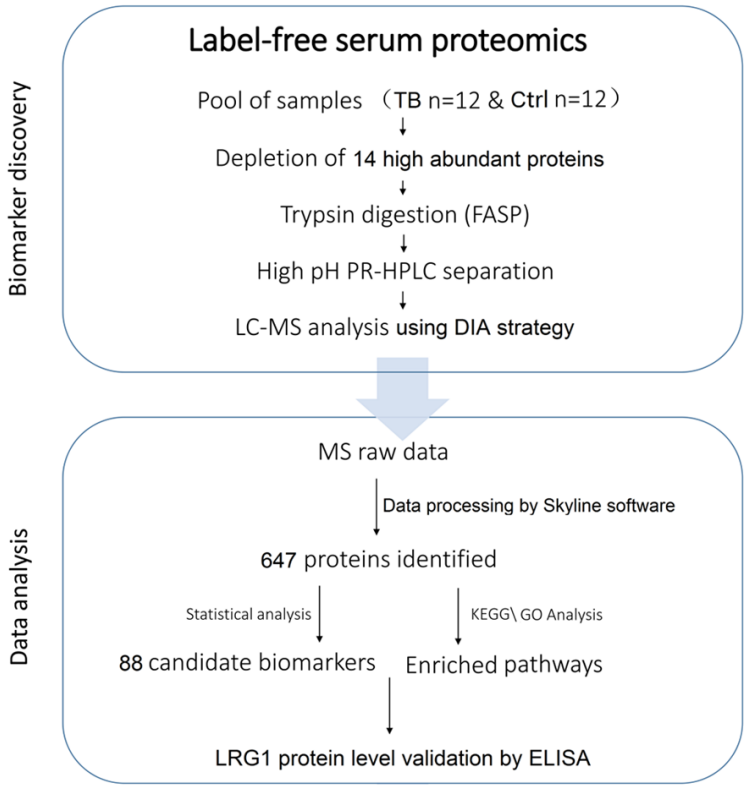

Figure 1. The workflow of the present study. It can be divided into two main stages. First was the biomarker discovery stage consisted of serum sample preparation and protein expression analysis. The second stage was the data analysis stage consisted of statistical analysis and bioinformatic analysis (PPI, GO and KEGG).

VN01H02, Sartorius Stedim Biotech) were mixed with 100 µL of 8 M urea in 0.1 M Tris/HCl (pH 8.5) and samples were centrifuged at 14000 g at 20°C for 15 min. This step was performed twice, after which 10 µL of 0.05 M Tris-(2-carboxyethyl) phosphine (TCEP) in water was added to the filters, and samples were incubated at 37°C for 1 h. Then, 10 µL of 0.1 M iodoacetamide (IAA) was added to the filters, after which the samples were incubated in darkness for 30 min. Filters were washed twice with 200 µL of 50 mM NH_4HCO_3. Finally, 4 µg of trypsin (Promega, Madison, WI) in 100 µL of 50 mM NH_4HCO_3 was added to each filter. The protein to enzyme ratio was 50:1. Samples were incubated overnight at 37°C and released peptides were collected by centrifugation.

LC-MS analyses

One µg of the samples was analyzed on a C18 column (75 um × 50 cm, 3 um) at 50°C, using an U3000 UHPLC connected to a Q Exactive mass spectrometer (Thermo Scientific). The peptides were separated by a 3 h linear gradient of from 5 to 35% ACN with 0.1% formic acid at 300 nl/min, followed by a linear increase to 98% ACN in 2 min and 98% for 8 min. For DDA acquisition, the full scan was performed between 400-1,000 m/z. The automatic gain control target for the MS/MS scan was set to 5e5. Normalized collision energy (NCE) was 27.

For DIA acquisition, the method consisted of a full scan at 35,000 resolution from 400 to 1,000 m/z (automatic gain control target of 1*10E6 or 100 ms injection time) followed by 20 DIA windows acquired at 17,500 resolution (automatic gain control target 3e6 and auto for injection time). NCE was 27. The spectra were recorded in profile type. The MS/MS spectra were recorded from 200 to 1800 m/z.

Spectral library generation

For generation of the spectral libraries, 6 DDA measurements of the “profiling standard sample set” were performed. DDA raw data were searched against the UniProt human database (release 201304, 89601 entries) using the Sequest HT (Proteome Discoverer v2.3.2) local server. Precursor and product ion spectra were searched with an initial mass tolerance of 20 ppm and 0.05 Da, respectively. Tryptic cleavage was selected, and up to two missed cleavages were allowed. Carbamidomethylation on cysteine (57.02 Da) was set as a fixed modification, and oxidation (15.99 Da) on methionine was set as a variable modification. The target-decoy-based strategy was applied to control both peptide- and protein-level false discovery rates (FDRs) at lower than 5%. The searching result was exported as .msf file format containing the annotation of precursors and fragment ions, also exact retention times. The msf file

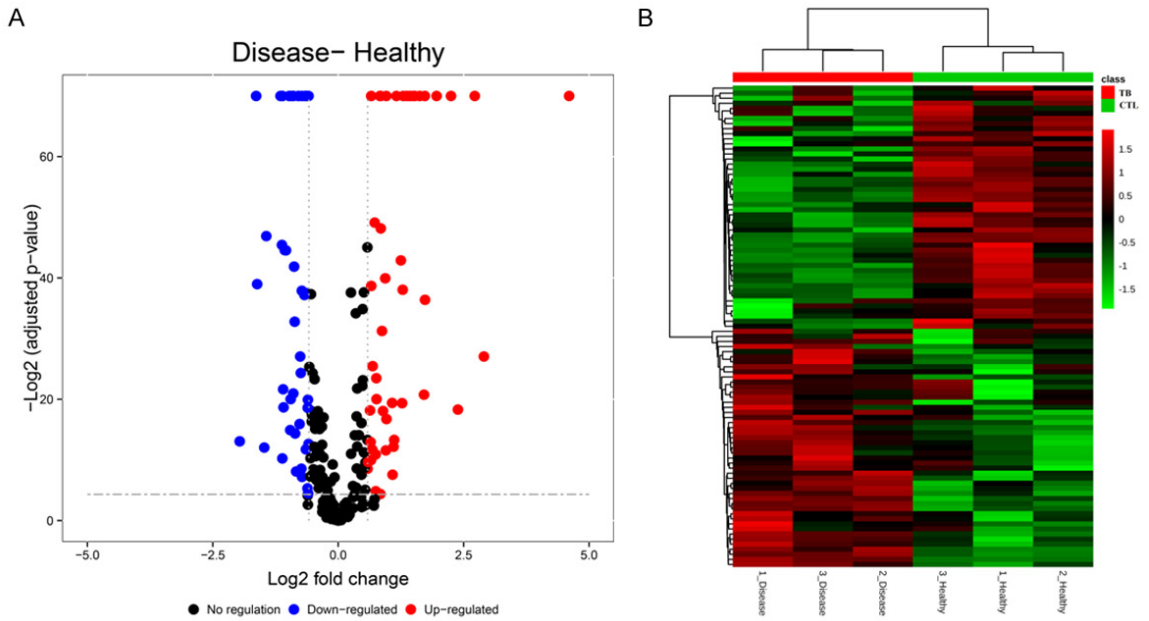

Figure 2. A. Scatter plots showing the proteome expression profile of the two groups with log2 fold change (X) and -log2 *p*-value (Y). B. Hierarchical clustering analysis for visualizing the 88 significantly dysregulated proteins.

was then imported into Skyline software for the generation of spectral library used for DIA data analysis.

Protein identification and quantitation

The DIA data were analyzed with Skyline software [10], a mass spectrometer vendor-independent free software from DIA data analysis. Raw data were analyzed as the user guide of the software. The default settings were used for the protein identification and peak area calculation. The dotp and idotp were set to 0.6 and 0.7 for protein identification and quantification, separately. After peak extraction and area calculation, the result was exported as table format for further quantification analysis using MSstats [24] package in R.

Bioinformatic analysis

For bioinformatic analysis, the 88 significantly dysregulated proteins were used as input, and the protein-protein interaction network construction and KEGG [25] pathway enrichment analysis were performed using the STRING [26] web service (http://www.string-db.org/). The BiNGO [27] plugin in the Cytoscape [28] environment was used to retrieve the Gene Ontology Consortium (GOC, http://geneontology.org/) in terms of molecular function, biological process and cellular component. The statistical test used was Hypergeometric test, and the FDR associated with multiple testing was corrected using the Benjamini-Hochberg method and an FDR-corrected p value <0.0001 was considered significant.

LRG1 quantification by ELISA

The concentration of LRG1 (Leucine-rich alpha-2-glycoprotein) in TB and normal serum samples was validated by enzyme linked immune sorbent assay (ELISA). The analysis was performed according to the manufacturer's protocol. First, standard protein samples and serum samples were diluted as instruction, and 50 µl of each sample was added into wells. Both the standard samples and serum samples were analyzed in duplicates. Then 50 µl of HRP-conjugate was added to the well, followed by incubation at 37°C for 1 hour. Then the wells were washed three times using 200 µl wash buffer, and 50 µl of substrate A and then 50 µl substrate B were added into wells followed by incubation at 37°C for 15 min. The reaction was stopped by adding 50 µl of stop solution to each well. Finally the optical density of each well was determined using a microplate reader set to 450 nm.

Statistical method

All statistical analyses were performed using MSstats package in R. For the discovery stage

Table 1. Significantly dysregulated proteins revealed by DIA-MS

Gene	Protein	Ratio (TB/Healthy)	*p* value
ADAMTS13	A disintegrin and metalloproteinase with thrombospondin motifs 13	0.256558	4.91E-05
CA1	Carbonic anhydrase 1	0.32165	0
PRDX2	Peroxiredoxin-2	0.326247	3.27E-13
PTPRG	Receptor-type tyrosine-protein phosphatase gamma	0.35995	0.000107
AOC3	Membrane primary amine oxidase	0.370083	1.11E-15
MMP2	72 kDa type IV collagenase	0.450048	0
LUM	Lumican	0.456737	0
TNXB	Tenascin-X	0.459556	3.11E-15
BST1	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 2	0.461728	0.000416
CDH13	Cadherin-13	0.46375	0
GP5	Platelet glycoprotein V	0.467554	8.55E-08
CRTAC1	Cartilage acidic protein 1	0.469944	7.47E-07
CNTN1	Contactin-1	0.477935	6.00E-15
PI16	Peptidase inhibitor 16	0.485333	6.22E-15
BTD	Biotinidase	0.508859	0
NCAM2	Neural cell adhesion molecule 2	0.515418	1.28E-05
APOC3	Apolipoprotein C-III	0.518446	2.81E-07
APOA4	Apolipoprotein A-IV	0.522983	0
CLEC3B	Tetranectin	0.531069	0
PGLYRP2	N-acetylmuramoyl-L-alanine amidase	0.537802	0
IGFBP3	Insulin-like growth factor-binding protein 3	0.53804	1.41E-07
HRG	Histidine-rich glycoprotein	0.538058	0
PF4	Platelet factor 4	0.542909	4.26E-14
QSOX1	Sulfhydryl oxidase 1	0.547639	3.05E-11
GSN	Gelsolin	0.550807	1.91E-05
FBLN1	Fibulin-1	0.558461	0.002058
TTR	Transthyretin	0.57673	0
APOC2	Apolipoprotein C-II	0.585845	6.12E-06
PROC	Vitamin K-dependent protein C	0.591517	1.68E-09
PLTP	Phospholipid transfer protein	0.596306	1.19E-08
COL6A3	Collagen alpha-3(VI) chain	0.602074	0.001435
BCHE	Cholinesterase	0.603639	0
APCS	Serum amyloid P-component	0.604023	7.44E-13
RBP4	Retinol-binding protein 4	0.606792	0.003887
PCYOX1	Prenylcysteine oxidase 1	0.626526	1.35E-12
PROCR	Endothelial protein C receptor	0.631349	0
APOC1	Apolipoprotein C-I	0.639662	0.000134
FN1	Fibronectin	0.641362	0
F12	Coagulation factor XII	0.646808	0
NRP1	Neuropilin-1	0.647115	1.05E-12
SERPINA4	Kallistatin	0.651668	0
TKT	Transketolase	0.654369	0.015445
PON1	Serum paraoxonase/arylesterase 1	0.655649	0.03649
APOC4	Apolipoprotein C-IV	0.656042	8.03E-07
VASN	Vasorin	0.658599	3.03E-07
DSG2	Desmoglein-2	0.660536	0.028025

IGFALS	Insulin-like growth factor-binding protein complex acid labile subunit	0.663146	0
HSPG2	Basement membrane-specific heparan sulfate proteoglycan core protein	0.664062	6.69E-05
IGHM	Ig mu chain C region	1.500582	0.001428
HBA1	Hemoglobin subunit alpha	1.511206	0.000627
HSP90B1	Endoplasmin	1.55146	1.10E-06
ORM2	Alpha-1-acid glycoprotein 2	1.564364	5.46E-05
KPRP	Keratinocyte proline-rich protein	1.574416	0.000498
SHBG	Sex hormone-binding globulin	1.579125	4.14E-13
C3	Complement C3	1.579222	0
FETUB	Fetuin-B	1.60203	0.000137
IGHD	Ig delta chain C region	1.610876	5.29E-09
IGHG2	Ig gamma-2 chain C region	1.655489	2.22E-16
SERPING1	Plasma protease C1 inhibitor	1.676211	0.000252
IGHG4	Ig gamma-4 chain C region	1.679534	0.022049
F11	Coagulation factor XI	1.693972	2.25E-08
KRT9	Keratin, type I cytoskeletal 9	1.69511	2.75E-07
VWF	von Willebrand factor	1.778956	0
LTA4H	Leukotriene A-4 hydrolase	1.790134	0.030278
IGHA1	Ig alpha-1 chain C region	1.801613	4.44E-16
CP	Ceruloplasmin	1.80245	0
FLG2	Filaggrin-2	1.829562	8.98E-11
MASP2	Mannan-binding lectin serine protease 2	1.854577	1.19E-06
CD14	Monocyte differentiation antigen CD14	1.91615	1.63E-13
KRT78	Keratin, type II cytoskeletal 78	1.933439	0.000153
LRG1	Leucine-rich alpha-2-glycoprotein	1.942374	0
KRT16	Keratin, type I cytoskeletal 16	1.946115	3.28E-06
DSP	Desmoplakin	2.105781	4.39E-07
S100A9	Protein S100-A9	2.11752	0.003023
	Ig kappa chain V-I region Mev	2.144999	9.77E-05
KRT17	Keratin, type I cytoskeletal 17	2.167084	4.14E-05
MMP9	Matrix metalloproteinase-9	2.231027	0
IGHG1	Ig gamma-1 chain C region	2.375023	2.04E-14
DSG1	Desmoglein-1	2.417389	4.62E-07
IGKV1-5	Immunoglobulin kappa variable 1-5	2.435325	6.46E-13
IGKC	Ig kappa chain C region	2.456341	0
KRT10	Keratin, type I cytoskeletal 10	2.575969	0
KRT1	Keratin, type II cytoskeletal 1	2.693196	0
ORM1	Alpha-1-acid glycoprotein 1	2.810947	0
IGHG3	Ig gamma-3 chain C region	2.879384	0
KRT2	Keratin, type II cytoskeletal 2 epidermal	3.088219	0
	Ig kappa chain V-I region Wes	3.274951	1.63E-07
C9	Complement component C9	3.310871	0
IGHV1-46	Immunoglobulin heavy variable 1-46	3.319458	2.38E-12
SERPINA1	Alpha-1-antitrypsin	3.891476	0
PZP	Pregnancy zone protein	4.752089	0
	Ig kappa chain V-I region Scw	5.227445	9.90E-07
HP	Haptoglobin	6.603465	0
SAA1	Serum amyloid A-1 protein	7.471562	1.74E-09
CRP	C-reactive protein	24.27921	0

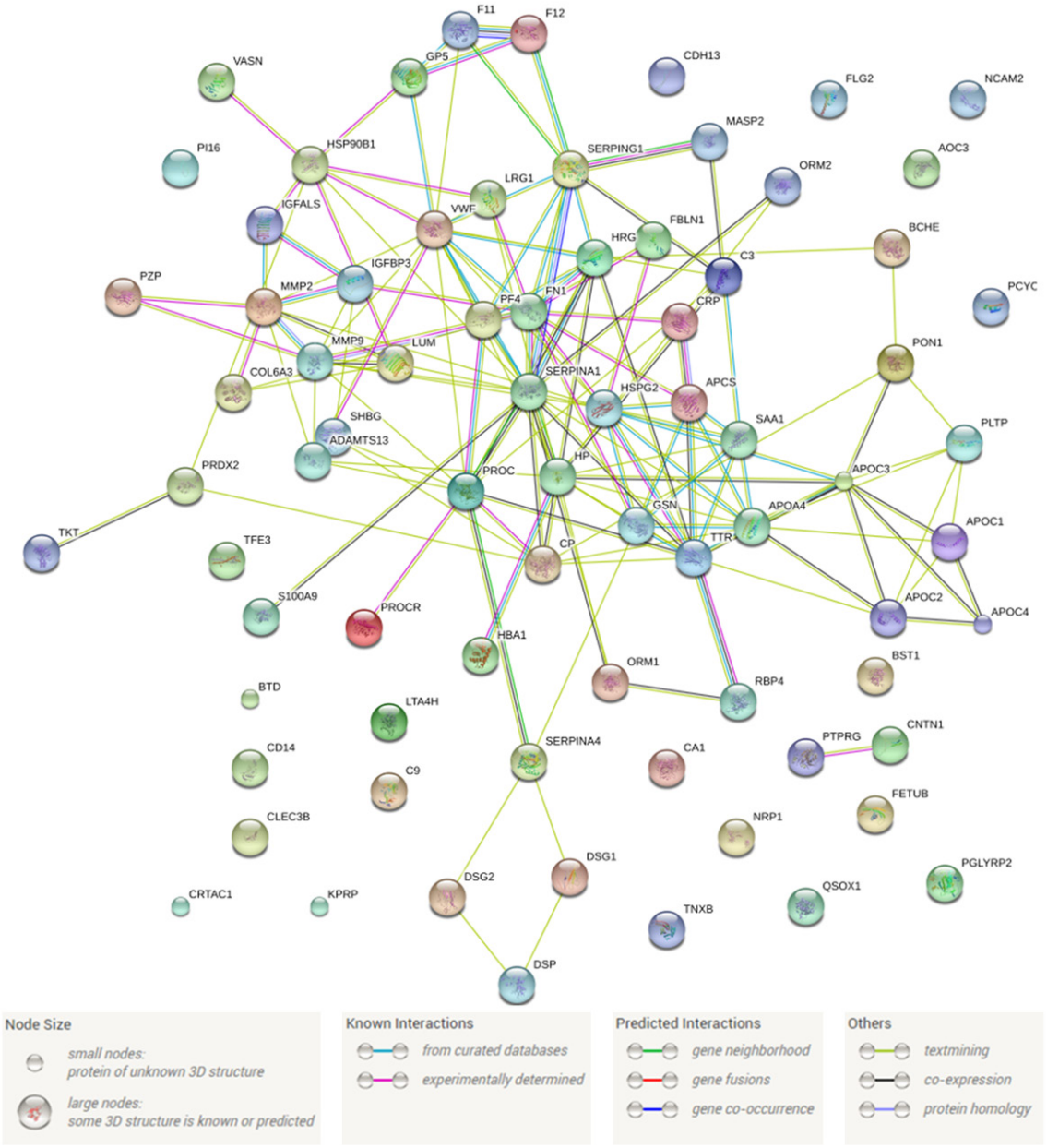

Figure 3. The protein-protein interaction network constructed from the 88 dysregulated proteins. Each edge represents a type of interaction between the linked nodes.

a 1.5-fold change and the Student's t-test *p* value of 0.05 were used as combined thresholds to define biologically regulated proteins.

Results

Study design and workflow flowchart

The workflow of the present study is shown in **Figure 1**. It can be divided into two main stages. First was the biomarker discovery stage consisted of serum sample preparation and protein expression analysis. The second stage was the data analysis stage consisted of statistical analysis and bioinformatic analysis (PPI, GO and KEGG). What's more, we choose LRG1 (Leucine-rich alpha-2-glycoprotein), which is one of the 88 significantly dysregulated proteins, to be analyzed by ELISA to validate the reliability of our proteomic analysis.

Protein expression analysis

As a result, 647 serum proteins containing 6435 peptides were identified using DIA workflow. 240 of them were quantified (more than 2

Table 2. Bioinformatic analysis of the proteomic results

Biological process				
Pathway ID	Pathway description	Observed gene count	False discovery rate	Matching proteins in network
GO.0051239	Regulation of multicellular organismal process	32	5.60E-09	APCS, APOA4, APOC1, APOC2, APOC3, C3, CNTN1, CRP, DSG2, DSP, F11, F12, FBLN1, GP5, HRG, LRG1, LUM, MMP9, NRP1, ORM1, PF4, PGLYRP2, PI16, PROC, PROCR, PTPRG, RBP4, S100A9, SAA1, SERPING1, TFE3, VASN
GO.0051241	Negative regulation of multicellular organismal process	22	5.60E-09	APCS, APOC1, APOC2, APOC3, CRP, F11, F12, FBLN1, FN1, GP5, HRG, NRP1, ORM1, PF4, PGLYRP2, PI16, PROC, PROCR, PTPRG, RBP4, SERPING1, VASN
GO.0006953	Acute-phase response	8	7.71E-09	APCS, CRP, FN1, HP, ORM1, ORM2, SAA1, SERPINA1
GO.0072376	Protein activation cascade	9	7.71E-09	C3, C9, F11, F12, FBLN1, GP5, MASP2, SERPING1, VWF
GO.0002526	Acute inflammatory response	9	9.42E-09	APCS, CRP, F12, FN1, HP, ORM1, ORM2, SAA1, SERPINA1
GO.0052547	Regulation of peptidase activity	14	2.07E-07	C3, CD14, COL6A3, FBLN1, FN1, GSN, HRG, MMP9, PI16, PZP, S100A9, SERPINA1, SERPINA4, SERPING1
Molecular function				
Pathway ID	Pathway description	Observed gene count	False discovery rate	Matching proteins in your network (labels)
GO.0005509	Calcium ion binding	17	1.04E-06	ADAMTS13, AOC3, APCS, CDH13, CLEC3B, CRP, CRTAC1, DSG1, DSG2, FBLN1, FLG2, GSN, HSP90B1, MASP2, PON1, PROC, S100A9
GO.0061134	Peptidase regulator activity	10	5.59E-06	C3, COL6A3, FBLN1, FN1, HRG, PI16, PZP, SERPINA1, SERPINA4, SERPING1
GO.0005515	Protein binding	39	5.97E-06	ADAMTS13, AOC3, APCS, APOA4, APOC2, APOC3, BCHE, C3, CDH13, CLEC3B, CP, CRP, DSG1, DSG2, DSP, F12, FBLN1, FN1, GSN, HP, HRG, HSP90B1, HSPG2, IGFALS, IGFBP3, LUM, MASP2, MMP9, NRP1, PF4, PON1, PTPRG, S100A9, SAA1, SERPINA1, TKT, TTR, VASN, VWF
GO.0004857	Enzyme inhibitor activity	11	3.33E-05	APOC1, APOC2, APOC3, C3, COL6A3, HRG, PI16, PZP, SERPINA1, SERPINA4, SERPING1
GO.0030414	Peptidase inhibitor activity	8	0.00014	C3, COL6A3, HRG, PI16, PZP, SERPINA1, SERPINA4, SERPING1
GO.0030234	Enzyme regulator activity	15	0.000267	APOA4, APOC1, APOC2, APOC3, C3, COL6A3, FBLN1, FN1, HRG, IGFBP3, PI16, PZP, SERPINA1, SERPINA4, SERPING1
Cellular component				
Pathway ID	Pathway description	Observed gene count	False discovery rate	Matching proteins in your network (labels)
GO.0005615	Extracellular space	52	3.80E-43	ADAMTS13, APCS, APOA4, APOC1, APOC2, APOC3, APOC4, BCHE, BTD, C3, C9, CD14, CDH13, CLEC3B, COL6A3, CP, CRP, F11, F12, FBLN1, FETUB, FN1, GSN, HBA1, HP, HRG, HSPG2, IGFALS, IGFBP3, LRG1, LUM, MMP2, MMP9, NRP1, ORM1, ORM2, PCYOX1, PF4, PI16, PLTP, PON1, PROC, PZP, QSOX1, RBP4, S100A9, SAA1, SERPINA1, SERPINA4, SERPING1, TTR, VASN
GO.0070062	Extracellular exosome	64	1.42E-42	APCS, APOA4, APOC1, APOC2, APOC3, BST1, BTD, C3, C9, CA1, CD14, CDH13, CLEC3B, CNTN1, COL6A3, CP, CRP, CRTAC1, DSG1, DSG2, DSP, F11, F12, FBLN1, FETUB, FLG2, FN1, GP5, GSN, HBA1, HP, HRG, HSP90B1, HSPG2, IGFALS, IGFBP3, KPRP, LRG1, LTA4H, LUM, MASP2, MMP9, ORM1, ORM2, PCYOX1, PGLYRP2, PI16, PON1, PRDX2, PROCR, PTPRG, PZP, QSOX1, RBP4, S100A9, SAA1, SERPINA1, SERPINA4, SERPING1, SHBG, TKT, TTR, VASN, VWF

GO.0044421	Extracellular region part	67	9.59E-40	APCS, APOA4, APOC1, APOC2, APOC3, APOC4, BCHE, BST1, BTD, C3, C9, CA1, CD14, CDH13, CLEC3B, CNTN1, COL6A3, CP, CRP, CRTAC1, DSG1, DSG2, DSP, F11, F12, FLG2, FN1, GP5, GSN, HBA1, HP, HRG, HSP90B1, IGFALS, IGFBP3, KPRP, LRG1, LTA4H, LUM, MASP2, MMP2, MMP9, NRP1, ORM1, ORM2, PCYOX1, PF4, PGLYRP2, PLTP, PON1, PRDX2, PROC, PROCR, PTPRG, PZP, QSOX1, RBP4, S100A9, SAA1, SERPINA1, SERPINA4, SERPING1, SHBG, TKT, TTR, VASN, VWF
GO.0031982	Vesicle	66	2.77E-39	APCS, APOA4, APOC1, APOC2, APOC3, BST1, BTD, C3, C9, CA1, CD14, CDH13, CLEC3B, CNTN1, COL6A3, CP, CRP, CRTAC1, DSG1, DSG2, DSP, F11, F12, FBLN1, FETUB, FLG2, FN1, GP5, GSN, HBA1, HP, HRG, HSP90B1, HSPG2, IGFALS, IGFBP3, KPRP, LRG1, LTA4H, LUM, MASP2, MMP9, NRP1, ORM1, ORM2, PCYOX1, PF4, PGLYRP2, PI16, PON1, PRDX2, PROCR, PTPRG, PZP, QSOX1, RBP4, S100A9, SAA1, SERPINA1, SERPINA4, SERPING1, SHBG, TKT, TTR, VASN, VWF
GO.0031988	Membrane-bounded vesicle	65	1.13E-38	APCS, APOA4, APOC1, APOC2, APOC3, BST1, BTD, C3, C9, CA1, CD14, CDH13, CLEC3B, CNTN1, COL6A3, CP, CRP, CRTAC1, DSG1, DSG2, DSP, F11, F12, FBLN1, FETUB, FLG2, FN1, GP5, GSN, HBA1, HP, HRG, HSP90B1, HSPG2, IGFALS, IGFBP3, KPRP, LRG1, LTA4H, LUM, MASP2, MMP9, ORM1, ORM2, PCYOX1, PF4, PGLYRP2, PI16, PON1, PRDX2, PROCR, PTPRG, PZP, QSOX1, RBP4, S100A9, SAA1, SERPINA1, SERPINA4, SERPING1, SHBG, TKT, TTR, VASN, VWF
GO.0005576	Extracellular region	66	1.25E-33	APCS, APOA4, APOC1, APOC2, APOC3, APOC4, BCHE, BST1, BTD, C3, C9, CA1, CD14, CDH13, CLEC3B, CNTN1, COL6A3, CP, CRP, CRTAC1, DSG1, DSG2, DSP, F11, F12, FLG2, FN1, GP5, GSN, HBA1, HP, HRG, HSP90B1, IGFALS, IGFBP3, KPRP, LRG1, LTA4H, LUM, MASP2, MMP2, MMP9, NRP1, ORM1, ORM2, PCYOX1, PF4, PGLYRP2, PLTP, PON1, PRDX2, PROCR, PTPRG, PZP, QSOX1, RBP4, S100A9, SAA1, SERPINA1, SERPINA4, SERPING1, SHBG, TKT, TTR, VASN, VWF

KEGG

Pathway ID	Pathway description	Observed gene count	False discovery rate	Matching proteins in your network (labels)
4610	Complement and coagulation cascades	9	9.81E-10	C3, C9, F11, F12, MASP2, PROC, SERPINA1, SERPING1, VWF
4512	ECM-receptor interaction	6	0.000107	COL6A3, FN1, GP5, HSPG2, TNXB, VWF

peptides were identified in all 6 replicates). Scatter plots with log2 fold change (X) and -log2 *p*-value (Y) shows the proteome expression profile of the two groups (**Figure 2A**). Using a 1.5-fold change and the Student's t-test *p* value of 0.05 as cutoffs, we found a total of 88 significantly dysregulated proteins, among which 40 were up-regulated and 48 were down-regulated in TB group. As shown in **Figure 2A**, red dots represent proteins up-regulated in TB group while the green dots represents down-regulated. Hierarchical clustering analysis was performed to visualize the 88 significantly dysregulated proteins (**Figure 2B**). The details of the 88 proteins, including protein ID (Uniprot), protein name, ratio TB/CTL, and -Log10 *p*-value, are listed in **Table 1**.

Bioinformatic analysis

The protein-protein interaction network was constructed for the 88 dysregulated proteins by retrieving the known interactions between each protein. As shown in **Figure 3**, a network containing 76 nodes (proteins) and 152 edges was drawn. Each edge represents a type of interaction between the linked nodes. The 88 significantly dysregulated proteins were then interro-

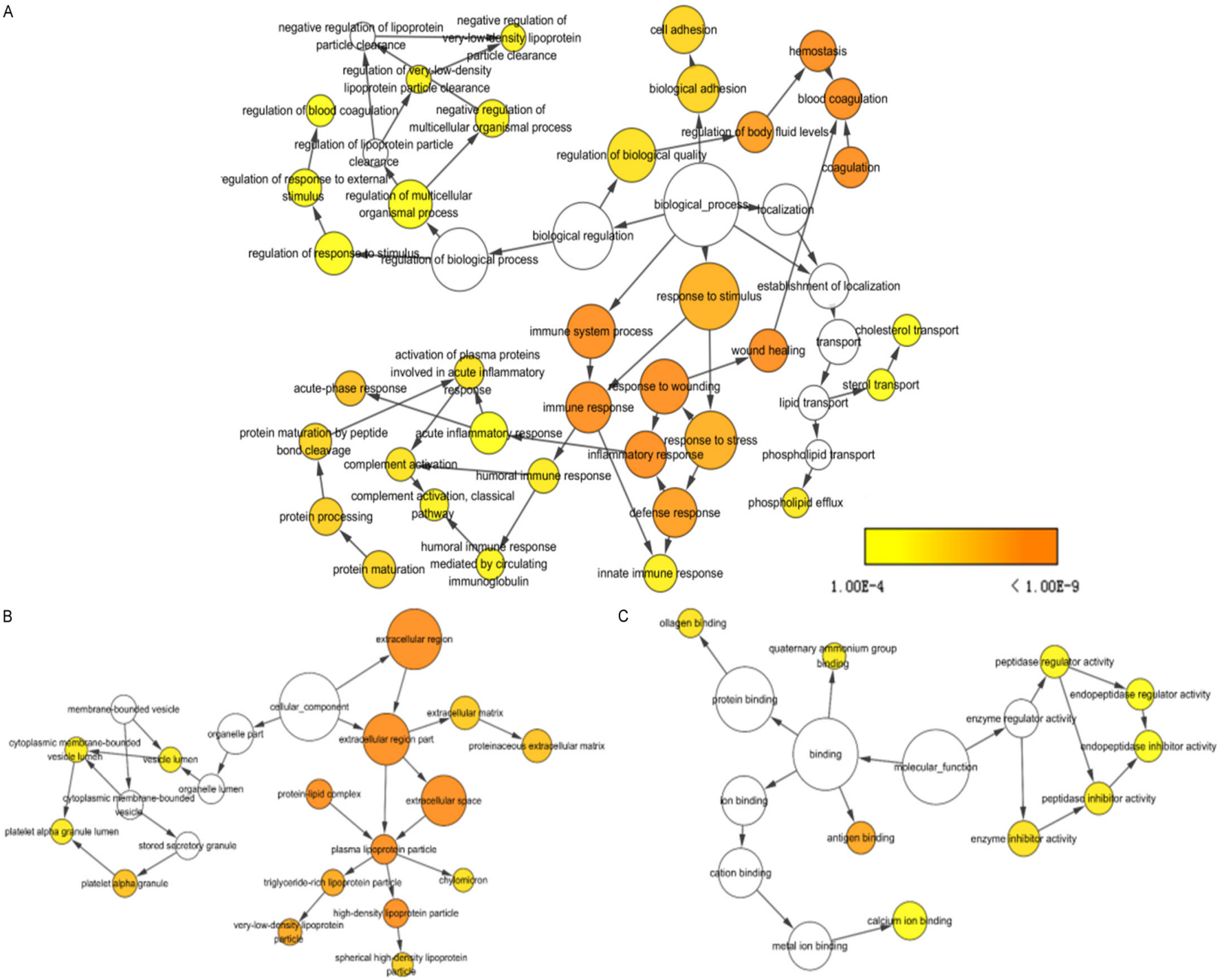
A
negative regulation of lipoprotein particle clearance
negative regulation of very-low-density lipoprotein particle clearance
regulation of very-low-density lipoprotein particle clearance
regulation of blood coagulation
negative regulation of multicellular organismal process
regulation of lipoprotein particle clearance
regulation of response to external stimulus
regulation of multicellular organismal process
regulation of response to stimulus
regulation of biological process
biological regulation
regulation of biological quality
cell adhesion
biological adhesion
hemostasis
blood coagulation
regulation of body fluid levels
coagulation
biological_process
localization
response to stimulus
establishment of localization
immune system process
wound healing
transport
cholesterol transport
sterol transport
lipid transport
phospholipid transport
phospholipid efflux
activation of plasma proteins involved in acute inflammatory response
acute-phase response
response to wounding
immune response
protein maturation by peptide bond cleavage
acute inflammatory response
response to stress
inflammatory response
complement activation
humoral immune response
complement activation, classical pathway
protein processing
defense response
humoral immune response mediated by circulating immunoglobulin
protein maturation
innate immune response
1.00E-4
< 1.00E-9
B
extracellular region
membrane-bounded vesicle
cellular_component
extracellular matrix
organelle part
extracellular region part
proteinaceous extracellular matrix
cytoplasmic membrane-bounded vesicle lumen
vesicle lumen
organelle lumen
protein-lipid complex
extracellular space
cytoplasmic membrane-bounded vesicle
platelet alpha granule lumen
stored secretory granule
plasma lipoprotein particle
platelet alpha granule
triglyceride-rich lipoprotein particle
chylomicron
high-density lipoprotein particle
very-low-density lipoprotein particle
spherical high-density lipoprotein particle
C
collagen binding
quaternary ammonium group binding
peptidase regulator activity
protein binding
endopeptidase regulator activity
enzyme regulator activity
binding
molecular_function
endopeptidase inhibitor activity
ion binding
peptidase inhibitor activity
antigen binding
enzyme inhibitor activity
cation binding
calcium ion binding
metal ion binding

Figure 4. Gene ontology (GO) analysis performed for revealing the biological process (A) molecular function (B) and cellular component (C) associated with the 88 significantly regulated proteins. The darker color of the node represents a lower *p* value.

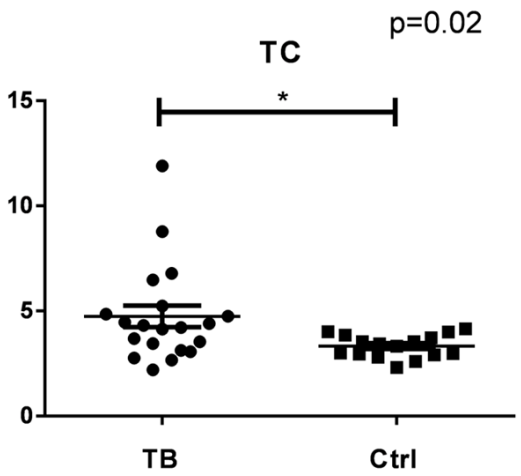

Figure 5. Grouped scatter plot showing the ELISA test result of the protein level of LRG1 in serum. Two groups (TB and healthy control) were included.

gated and mapped to KEGG pathways (**Table 2**), ant the two significantly enriched pathways were complement and coagulation cascades and ECM-receptor interaction. To further extend our knowledge about the change of serum proteins between disease and healthy groups, gene ontology (GO) analysis was performed to reveal the molecular function, biological process and cellular component associated with the 88 significantly regulated proteins. As shown in **Figure 4** and **Table 2**, the significantly regulated proteins are highly correlated with regulation of multicellular organismal process, acute-phase response and protein activation cascadeetc. in term of biological process (**Figure 4A**), participating the processes of calcium ion binding, protein binding and peptidase regulator activity regulation etc. In term of molecular function (**Figure 4B**), and mainly exist at extracellular space and vesicle (**Figure 4C**). Details of the GO analysis results are shown in **Table 2**.

Validation of LRG1 level by ELISA

Altogether 40 samples with 20 in each group were recruited for ELISA analysis to validate the serum level of LRG1. As shown in the grouped scatter plot in **Figure 5**, there was significant difference in the serum level of LRG1 between TB group and healthy controls (P=0.02). The quantification result showed that the LRG1 level was significantly up-regulated in TB group, which is consisted with the quantification result in the proteomic analysis.

Discussion

Proteomics is analytical tool used to perform qualitative and quantitative analyses of a large number of proteins in a sample. It has been widely used to search for biomarkers and proven to be a powerful tool [29]. The DDA strategy has been used in proteomic data acquisition for a long time, providing much information for biological and clinical studies [30, 31]. But due to the low reliability, the proteomic result revealing from DDA data often need additional validation by MRM or western blot. This drawback limits the wide application of DDA-based proteomics study. For the recent years, the DIA acquisition strategy has been proposed and became more and more popular in proteomic studies [32]. Lambert et al. used DIA strategy to characterize changes in protein-protein interactions imparted by the HSP90 inhibitor NVP-AUY922 or melanoma-associated mutations in the human kinase CDK4, showing that DIA is a robust label-free approach to characterize such changes and a scalable pipeline for systems biology studies [33]. Bruderer et al. used DIA workflow to profile acetaminophen (APAP)-treated three-dimensional human liver microtissues. As a result, an early onset of relevant proteome changes was revealed at subtoxic doses of APAP, and their findings also implied that DIA should be the preferred method for quantitative protein profiling [34]. It can be seen that the DIA strategy is of more potential for achieving the ultimate goal of proteomics in the future. In our present study, a total of 647 serum proteins were identified using DIA strategy, 88 of them were significantly dysregulated between TB and healthy control group. The ELISA test of LRG1 showed consistent result as the proteomic analysis, proving the reliability of our LC-DIA-MS workflow. The 88 dysregulated serum proteins can be used as candidate biomarkers for TB disease state in future clinical studies. Of course, for the preciseness and reli-

ability, these candidate biomarkers should be validated in larger cohorts before application in clinical diagnosis.

Biomarkers are indicators of the diseased or normal state of the body as well as drug efficacy [35]. Ideal biomarkers can be used for disease diagnosis and prognosis estimation as well as provide valuable information for the elucidation of pathology. Over the past few decades, studies have been carried out to discover biomarkers for *M. Tuberculosis* and patient immune response. The candidate biomarkers studied include *M. Tuberculosis* DNA and RNA and anti-phospholipid antibodies [36-38]. However, correctly judging disease condition has been difficult to achieve with the above biomarkers, often yielding poor accuracy. Hence, the exploration for new TB biomarkers remains essential. In our proteomic analysis, 88 serum proteins were significantly different between the two groups, as determined by statistical analysis. These metabolites are likely potential biomarkers, but further validation is needed for their progression to clinical practice.

The proteomic analysis can usually detect thousands of proteins and reveal hundreds of dysregulated proteins, and the systematic and comprehensive data analysis is important for fully extracting information from raw data. Bioinformatics is a useful mathematical tool for deep analysis of proteomics data [39]. In our result of functional network analyses (**Figure 3**), the constructed network has significantly more interactions than expected. This means that these proteins have more interactions among themselves than what would be expected for a random set of proteins of similar size, drawn from the genome. Such an enrichment indicates that the proteins are at least partially biologically connected, as a group. The KEGG pathway enrichment result showed the TB pathological process may be highly correlated with pathways of complement and coagulation cascades and ECM-receptor interaction (**Table 2**). From the results of gene ontology we can infer that the biological process influenced by TB infection include multicellular organismal process, acute inflammatory and so on, and the molecular functions of some of these 88 dysregulated proteins belongs to enzyme or peptidase inhibitor. In summary, our network analyses and gene ontology results revealed much information about the biological processes that are involved in TB disease progression and the anti-infection response of the human body after TB infection.

Conclusion

Serum serves as an important medium that interacts with cells, tissues and organs in the human body. It carries proteins secreted or leaked by different cells in response to pathologic progress. This study aimed to distinguish sera from TB and normal patients using LC-MS data in the context of metabolic profiling. Our approach demonstrated the importance of implementing multivariate statistical analysis and bioinformatics in determining the significantly changed metabolites and candidate disease biomarkers. Our results showed the broad-spectrum metabolite variation in TB that can be used for the diagnosis of this disease. Moreover, our findings suggest that the changes in serum glycerophospholipid levels serve as a potential molecular indicator for TB. However, this outcome must be further verified and validated in a larger number of clinical samples prior to application in the clinical setting.

Acknowledgements

This work was supported by Beijing Natural Science Foundation of China (No. 7143173).

Disclosure of conflict of interest

None.

Address correspondence to: Shibing Qin, Department of Orthopaedics, Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Tumor Institute, Beijing 101149, P. R. China. E-mail: qinsb@sina.com; Jing Shen, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Central Laboratory, Peking University Cancer Hospital & Institute, Beijing 100142, P. R. China. E-mail: shenjing@bjmu.edu.cn

References

[1] Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Görg A, Hecker M, Huber LA, Langen H, Link AJ, Paik YK, Patterson SD, Pennington SR, Rabilloud T, Simpson RJ, Weiss W, Dunn MJ. Guidelines for the next 10 years of proteomics. Proteomics 2006; 6: 4-8.

[2] Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 2005; 1: 252-262.
[3] Peng J, Elias, JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2003; 2: 43-50.
[4] Sugiyama N, Masuda T, Shinoda K, Nakamura A, Tomita M, Ishihama Y. Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications. Mol Cell Proteomics 2007; 6: 1103-1109.
[5] Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol 2010; 11: 427-439.
[6] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006; 127: 635-648.
[7] Wiese S, Reidegeld KA, Meyer HE, Warscheid B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. Proteomics 2007; 7: 340-350.
[8] Pichler P, Köcher T, Holzmann J, Mazanek M, Taus T, Ammerer G, Mechtler K. Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap. Anal Chem 2010; 82: 6549-6558.
[9] Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R. 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: 0111 016717.
[10] Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ. Multiplexed peptide analysis using data-independent acquisition and skyline. Nat Protoc 2015; 10: 887-903.
[11] Distler U, Kuharev J, Navarro P, Levin Y, Schild H, Tenzer S. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. Nat Methods 2014; 11: 167-170.
[12] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23: 2507-2517.
[13] Valencia A. Bioinformatics: biology by other means. Bioinformatics 2002; 18: 1551-1552.
[14] Tisoncik-Go J, Gasper DJ, Kyle JE, Eisfeld AJ, Selinger C, Hatta M, Morrison J, Korth MJ, Zink EM, Kim YM, Schepmoes AA, Nicora CD, Purvine SO, Weitz KK, Peng X, Green RR, Tilton SC, Webb-Robertson BJ, Waters KM, Metz TO, Smith RD, Kawaoka Y, Suresh M, Josset L, Katze MG. Integrated omics analysis of pathogenic host responses during pandemic H1N1 influenza virus infection: the crucial role of lipid metabolism. Cell Host Microbe 2016; 19: 254-266.
[15] Deshmukh AS, Murgia M, Nagaraj N, Treebak JT, Cox J, Mann M. Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. Mol Cell Proteomics 2015; 14: 841-853.
[16] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005; 122: 957-968.
[17] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005; 437: 1173-1178.
[18] Lehne B, Schlitt T. Protein-protein interaction databases: keeping up with growing interactomes. Hum Genomics 2009; 3: 291-7.
[19] Klingstrom T, Plewczynski D. Protein-protein interaction and pathway databases, a graphical review. Brief Bioinform 2011; 12: 702-713.
[20] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000; 25: 25-29.
[21] Gomez JE, McKinney JD. M. Tuberculosis persistence, latency, and drug tolerance. Tuberculosis 2004; 84: 29-44.
[22] Flynn JL. Immunology of tuberculosis and implications in vaccine development. Tuberculosis 2004; 84: 93-101.
[23] Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods 2009; 6: 359-362.
[24] Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O. MSstats: an R package

for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 2014; 30: 2524-2526.
[25] Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol Biosyst 2014; 10: 2441-2447.
[26] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015; 43: D447-452.
[27] Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005; 21: 3448-3449.
[28] Doerks T, Copley RR, Schultz J, Ponting CP, Bork P. Systematic identification of novel protein domain families associated with nuclear functions. Genome Res 2002; 12: 47-56.
[29] Favicchio R, Thepaut C, Zhang H, Arends R, Stebbing J, Giamas G. Strategies in functional proteomics: unveiling the pathways to precision oncology. Cancer Lett 2016; 382: 86-94.
[30] Xu L, Gao Y, Chen Y, Xiao Y,He Q, Qiu H, Ge W. Quantitative proteomics reveals that distant recurrence-associated protein R-Ras and Transgelin predict post-surgical survival in patients with Stage III colorectal cancer. Oncotarget 2016; 7: 43868-43893.
[31] Patella F, Neilson LJ, Athineos D, Erami Z, Anderson KI, Blyth K, Ryan KM, Zanivan S. In-Depth proteomics identifies a role for autophagy in controlling reactive oxygen species mediated endothelial permeability. J Proteome Res 2016; 15: 2187-2197.
[32] Tsou CC, Tsai CF, Teo GC, Chen YJ, Nesvizhskii AI. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using orbitrap mass spectrometers. Proteomics 2016; 16: 2257-2271.
[33] Lambert JP, Ivosev G, Couzens AL, Larsen B, Taipale M, Lin ZY, Zhong Q, Lindquist S, Vidal M, Aebersold R, Pawson T, Bonner R, Tate S, Gingras AC. Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. Nat Methods 2013; 10: 1239-1245.
[34] Bruderer R, Bernhardt OM, Gandhi T, Miladinović SM, Cheng LY, Messner S, Ehrenberger T, Zanotelli V, Butscheid Y, Escher C, Vitek O, Rinner O, Reiter L. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics 2015; 14: 1400-1410.
[35] Mishra, J, Dent C, Tarabishi R, Mitsnefes MM, Ma Q, Kelly C, Ruff SM, Zahedi K, Shao M, Bean J, Mori K, Barasch J, Devarajan P. Neutrophil gelatinase-associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery. Lancet 2005; 365: 1231-1238.
[36] Bakir M, Millington KA, Soysal A, Deeks JJ, Efee S, Aslan Y, Dosanjh DP, Lalvani A. Prognostic value of a T-cell–based, interferon-γ biomarker in children with tuberculosis contact. Ann Int Med 2008; 149: 777-786.
[37] Frahm M, Goswami ND, Owzar K, Hecker E, Mosher A, Cadogan E, Nahid P, Ferrari G, Stout JE. Discriminating between latent and active tuberculosis with multiple biomarker responses. Tuberculosis (Edinburgh, Scotland) 2011; 91: 250-256.
[38] Martin A, Barrera L, Palomino JC. Biomarkers and diagnostics for tuberculosis. Lancet 2010; 376: 1539-1540.
[39] Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 2002; 48: 1296-1304.