METWARE BIO

Innovative Proteomics Insights for Better Health

Report on Blood Quantitative Proteomics

Metware Biotechnology Inc.

www.metwarebio.com

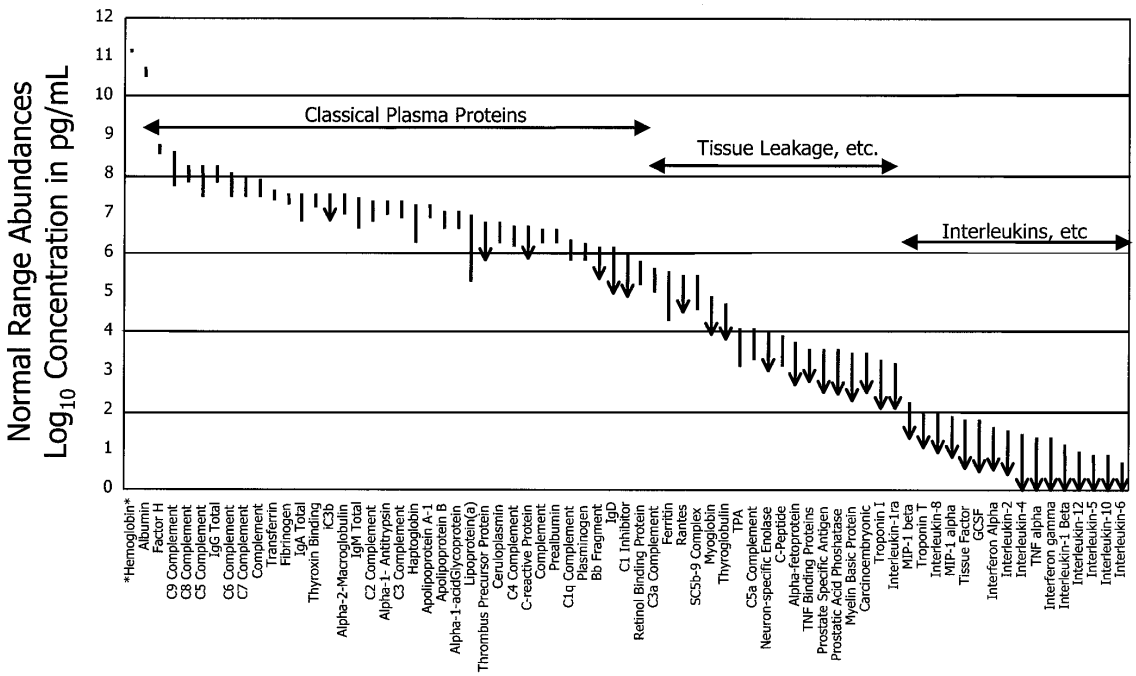
Content

Report on Blood Quantitative Proteomics

1 Overview

1.1 Blood Proteome

In disease research, blood is considered a window into the patient's overall health because it circulates throughout the body and reflects the characteristics of the internal environment and physiological functions. Blood samples contain a rich variety of proteins, with high-abundance proteins accounting for 97%-99%. Other proteins originate from secretion or leakage from various organs and are then diluted in the peripheral blood, with concentrations as low as pg/ml, often masked by higher abundance proteins in proteomic analysis. Due to the high content of abundant proteins in blood, conventional proteomics studies are limited in their detection throughput of blood proteins.

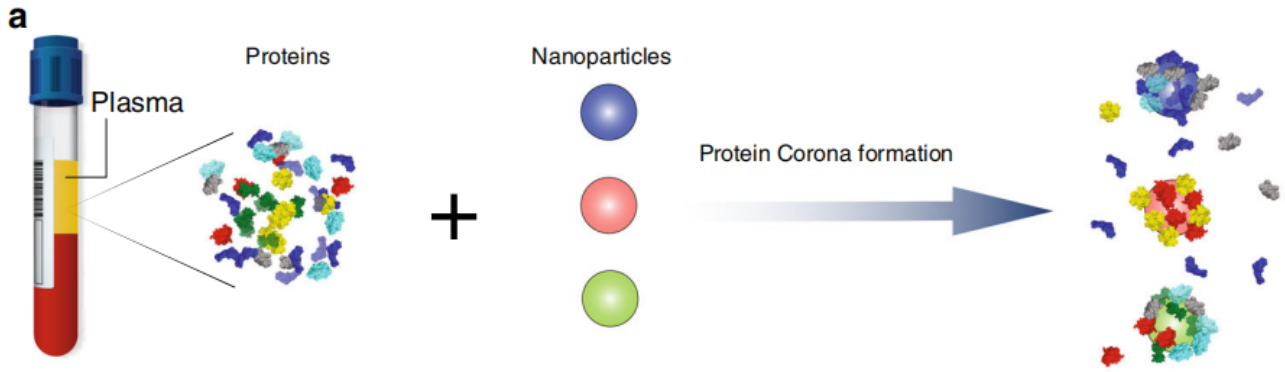


Concentration Differences of 70 Proteins in Plasma

1.2 Blood Proteomics

To overcome the limitations imposed by high-abundance proteins on blood proteome detection, numerous techniques have been developed to deplete these proteins. Among them, the method using nano-

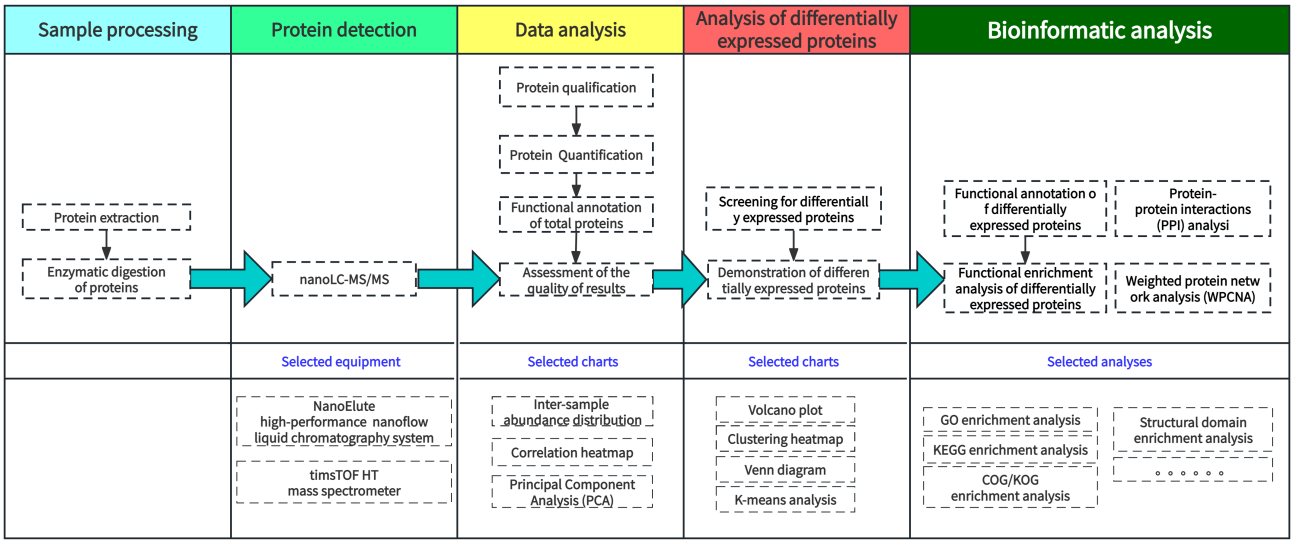
graphene magnetic beads to enrich low-abundance proteins stands out due to its species-independent nature and high detection throughput. The principle of this technique is using functional biological magnetic beads specifically to adsorb low-abundance proteins in the blood and form a protein corona around the nanoparticles. The enriched proteins are then detected. MetwareBio has established a blood proteomics detection workflow based on the method of enriching low-abundance proteins with functional biological magnetic beads and the DIA scanning method using the Bruker timsTOF HT mass spectrometer.



Schematic of Low-Abundance Protein Enrichment Using Functional Magnetic Beads

1.3 Technical workflow

Blood quantitative proteomic investigation of samples is realized by integrating a series of cutting-edge technologies, including protein extraction, enzymatic digestion, liquid chromatography-mass spectrometry tandem analysis, database retrieval and analysis, and bioinformatic analysis.



Route Map for Proteomics Technology

2 Materials and Methods

2.1 Reagents

The main reagents used in this project are listed in the table below.

Table 2.1 Information on Reagents

Reagent name	Prod. No.	Brand	Grade
BCA protein assay kit	ZD388299	Thermo Fisher	\
PBS buffer（1X）	BL302A	biosharp	\
PMSF	XW020003	CNW	A.R.
acetonitrile	1.00030.4008	Merck	LC-MS
isopropanol	A451-4	Thermo Fisher	LC-MS
Methanol	1.06007.4008	Merck	LC-MS
Trichloroacetic acid	T9159-250G	Sigma	A.R.
Acetic acid	CAEQ-4-000301-0050	CNW	LC-MS
cOmplete Protease Inhibitor Cocktail	4693116001	merck	LC-MS
Formic acid	CAEQ-4-000313-0050	CNW	LC-MS
Water	CAEQ-4-000302-4000	CNW	LC-MS
EasyPeptTM low-aboundance protein enrichment kit	0SFP0002	omicsolution	\
2XiRT kit	Ki-3002-2	Biognosys	\

2.2 Instruments

The main instruments used in this project are listed in the table below.

Table 2.2 Information on Equipment and Instruments

Name	Brand	Instrument
Ultrasonic cell grinder	Qsonica	Qsonica Q800R3
Benchtop high-speed freezing centrifuge	eppendorf	5430R
Vacuum freeze dryer	labconco	Freez zone 12L-84
Electrophoresis system	Beijing Liuyi Instrument Plant	DYY-6C
Electrophoresis tank	Beijing Liuyi Instrument Plant	DYCZ-24DN
Mass spectrometer	Bruker	timsTOF HT
Pipette	Eppendorf	/
Microplate reader	Thermo Fisher	A51119600C

2.3 Sample processing

First, thaw the samples on ice and add 1 mM PMSF to achieve a final concentration, then vortex and mix thoroughly. Subsequently, use the EasyPeptTM kit (Shanghai Omicsolution Biotechnology Co., Ltd.) to enrich low-abundance proteins from blood using nano-magnetic beads. Next, perform reduction and alkylation on the beads, followed by enzymatic digestion of proteins into peptides using trypsin. Finally, desalt the peptides using a C18 column (Millipore, Billerica, MA), and determine the peptide concentration using the BCA assay kit (Thermo Fisher).

2.4 LC-MS/MS Analysis

1. High-Performance Liquid Chromatography (HPLC) with NanoElute System

The samples were separated using a NanoElute UHPLC system with a nanoliter flow rate. Mobile phase A was 0.1% formic acid aqueous solution and phase B was 0.1% formic acid acetonitrile solution (100% acetonitrile). Samples were uploaded by an autosampler to an analytical column (IonOpticks, Australia, 25 cm × 75 μm, C18 packing 1.6 μm) for separation. The temperature of the column was controlled at 50 °C by an integrated column heater; the sample volume was 300 ng; the flow rate was 300 nL/min; the gradient was 47 min. The liquid phase gradient was: 0 min - 40 min, liquid B from 2% to 35%; 40 min - 40.5 min, liquid B linear gradient from 35% to 95%; 40.5 min - 47 min, liquid B maintained at 95%.

2. Detection with timsTOF HT Mass Spectrometer

The mixed samples were firstly separated by chromatography. The mass spectrometry data were then acquired using the ddaPASEF mode of a timsTOF HT mass spectrometer to establish a method for an appropriate acquisition window for the diaPASEF acquisition method. The parameters used in the analysis were: valid gradient 47 min, positive ion detection mode, parent ion scanning range 100-1700 m/z, ion mobility $1/K_0$ in the range of 0.6-1.6 Vs/cm^2, ion accumulation and release time 100 ms, ion utilization rate approximating 100%, capillary voltage 1600 V, drying gas rate 3 L/min, drying temp. 180°C. The parameters used in ddaPASEF acquisition mode were: 10 PASEF MS/MS frames in 1 complete frame (total cycle time was 1.17 s), The capillary voltage was set to 1600 V, and the MS and MS/MS spectra were acquired from 100 to 1700 m/z. As for ion mobility range ($1/K_0$), 0.6 to 1.6 Vs/cm^2 was used. The "target value" of 20,000 was applied to a repeated schedule, and the intensity threshold was set at 2500. The range of charge state was set from 0 to 5. The collision energy was ramped linearly as a function of mobility from 59eV at $1/K_0$ = 1.6 Vs/cm^2 to 20 eV at $1/K_0$ = 0.6 Vs/cm^2. The quadrupole isolation

width was set to 2Th for m/z < 700 and 3Th for m/z > 800. After having established the method for an appropriate acquisition window using ddaPASEF, diaPASEF acquisition method was used for proteomics analysis and the parameters were: mass range approx. 400-1200, mobility range 0.6-1.6 Vs/cm^2, mass width 25 Da, mass overlap 0.1, 24 mass steps per cycle, and the number of mobility windows was 2, totaling 48 acquisition windows. The average acquisition cycle was 1.17s.

2.5 Mass Spectrometry Data Analysis

Mass spectrometry analysis yields mass-to-charge ratios and signal intensities of peptides in the sample, as well as mass-to-charge ratios and signal intensities of fragment ions after peptide fragmentation. The information at the peptide level is usually called the primary mass spectrogram and the ion information of the peptide fragments is called the secondary mass spectrogram. The information contained in the mass-spectrogram is so complex that a database is required to properly resolve the peptide sequences potentially contained in the spectrogram. Before searching, a database of theoretical secondary mass-spectrograms is constructed based on the protein sequences in the database. The secondary mass-spectrograms generated by mass spectrometry are then searched against the theoretical secondary mass-spectrograms, and the correctly aligned theoretical peptide sequences are obtained by algorithmic scoring and filtering. Through the identified protein-specific peptides, the contained protein information is recognized.

Database searching is a complex computational process that requires the use of specialized mass spectrometry data analysis software for data parsing. The database search software used for the DIA mass spectrometry data in this project was DIA-NN (v1.8.1), which searches the database using the Library-free method. Search parameters included uniprot-proteome.fasta database (A total of 55319 sequences), with deep learning-based parameters checked to predict a library of spectrograms; MBR was checked to generate a library of spectrograms using the DIA data, which was utilized to reanalyze the DIA data for protein quantification; and FDRs were filtered at 1% for precursor ions and at the protein level. The filtered data were ready for subsequent bioinformatics analysis.

2.6 Protein Quantification

Quantification principle of the database search software

Protein quantification by the DIA-NN software is performed by the MaxLFQ algorithm, which is based on the following principle:

1. Given a study with N experiments (samples) and a protein with M peptides (Razor+Unique) for quantification, first calculate the logarithmsized value of the ratio of the intensities of each peptide

$p \in [1\ M]$ between samples i and j, using the following equation:

$$r_{i,j}(p) = log\frac{I_i(p)}{I_j(p)} = logI_i(p) - logI_j(p)$$

where $I_i(p)$ indicates the signal intensity of peptide p in sample i. If the peptide ion is not present in sample i or j, the corresponding log value is not calculated.

2. The linear relationship of a protein across two samples can be expressed by the median of the logarithmized intensity ratio of corresponding peptides of the protein, which is calculated by the following equation:

$$x_i - x_j = Median(r_{i,j}(p))$$

where x_i represents the logarithmized protein intensity.

3. For any given N experiments (samples), Eq. 2 above can be presented in the form of a matrix: $A_x = b$ where A_x can be expressed as follows:

$$A_{i,j} = \begin{cases} -1 & i \neq j \\ \sum_{i=1}^{N-1} 1(i,j) & i = j \end{cases}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

$$b_i = \begin{cases} \sum_{j=i+1}^{N} Median(r_{i,j}(p)) & i = 1 \\ \sum_{j=i+1}^{N} Median(r_{i,j}(p)) - \sum_{j=1}^{i} Median(r_{j,i}(p)) & i > 1 \end{cases}$$

where, 1 (i, j) is equal to 1 when the peptide is present in both samples i and j, otherwise it is equal to 0. Equation 2 can be solved efficiently using the Cholesky decomposition to obtain the log value of the protein intensity x_i. The protein intensity in experiment i is then equal to e^{xi}.

For projects with only one experiment (sample), the MaxLFQ algorithm cannot be used and the software will automatically perform a quantitative analysis using the Top-N algorithm.

Normalization process

After the database search software completes the protein quantification, the intensity of each protein in different samples given in the database search result must be extracted, and then the in-sample normalization is carried out by the centroid transformation to obtain the relative quantitative value (R) of the protein in different samples. The calculation formula is as follows:

$$R_{ij} = I_{ij}/Median(I_i)$$

where *i* represents the sample and *j* represents the protein.

Differential analysis

After normalization, differential quantification of protein is required to screen for protein that are differentially expressed between groups of samples (biological replicates were taken as samples in the same group) derived by differential grouping. Commonly used statistical methods for variance analysis in proteomics include parametric and nonparametric tests, which can be selected based on the actual data. For projects with biological replicates, if the samples were divided into two groups by differential grouping, the mean ratio of quantification values of each protein across all biological replicates was used as the fold change (FC), t-test was performed using the quantification values of each protein in both groups of samples for variance test, and the associated P-value was calculated. When the samples were divided into more than two groups by differential grouping, the variance test was performed by ANOVA test using quantification values of each protein in each group of samples, and the associated P-value was calculated. For non-replicated items, only FC can be calculated and Pvalue cannot be calculated when the difference is grouped into two samples only; if P value hypothesis testing is required, FDR is usually calculated using the BH method.

2.7 Bioinformatics Analysis

In order to acquire a thorough understanding of the functional properties of different proteins, we performed a full range of functional annotations on the identified proteins and the differentially expressed proteins in each comparison group, respectively. These detailed functional annotations involve gene ontology (GO), KOG functional classification, KEGG pathway, protein domain, protein domain, subcellular localization and signal peptide (SignalP). The differentially expressed proteins in each comparison group were also analyzed for enrichment at four levels: GO classification, KOG functional classification, KEGG pathway and protein domain, where the significance P-value of enrichment was computed using hypergeometry with the aim of detecting whether differentially expressed proteins have a significant tendency of enrichment for certain functional types against background proteins (in this case, all identified proteins) Predictive profiling of subcellular locations and signalPs was also performed to better define the physiological functions in which the differentially expressed proteins are involved.

3 Project Results

3.1 Sample Information

Table 3.1 Sample Information

Sample	Group
A1	A
A2	A
A3	A
B1	B
B2	B
B3	B
C1	C
C2	C
C3	C

- Sample：sample name
- Group：group name

3.2 Qualitative and Quantitative Results

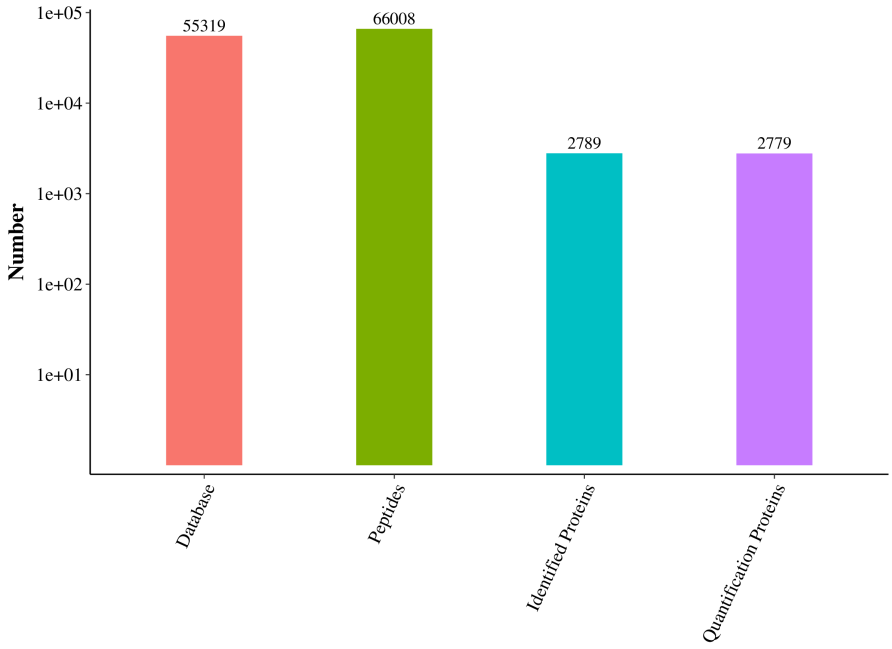
In this project, DIA-NN (v1.8.1) software was used to analyze the diaPASEF mass spectrometry data for protein qualification and quantification. Details of the analytical principles, parameter settings, quality control, and filtering methods of this software are described in Section 2.5. The analytical results of the mass spectrometry data obtained in this project were statistically analyzed, and the statistic tables and plots of these data are shown below:

Table 3.2 Summary of Identification Results

Database	Peptides	Identified Proteins	Quantification Proteins
55319	66008	2789	2779

- Database: Number of protein sequences in the selected database

- Peptides: Total number of peptides identified
- Identified Proteins: Total number of identified proteins
- Quantification Proteins: Total number of quantified proteins



Statistical Chart of Identification Results

Table 3.3 Protein Level Identification Results

Accession	Description	Gene	Protein Group	Protein IDs
A0A075B5R2	Immunoglobulin heavy variable 7-3 (Fragment) OS=Mu...	Ighv7-3	A0A075B5R2	A0A075B5R2
A0A075B5T3	Immunoglobulin heavy variable 6-6 (Fragment) OS=Mu...	Ighv6-6	A0A075B5T3	J3QK03;A0A0A6YWS9;A0A075B5T3
A0A087WPR7	Dystonin (Fragment) OS=Mus musculus OX=10090 GN=Ds...	Dst	A0A087WPR7	Q91ZU6;S4R1P5;A0A087WRB8;A0A087WSP0;E9Q9X1;A0A087W...
A0A087WQ89	MISP family member 3 OS=Mus musculus OX=10090 GN=M...	Misp3	A0A087WQ89	A0A087WQC3;A0A087WQV6;A0A087WQ89
A0A087WQF8	Kinectin OS=Mus musculus OX=10090 GN=Ktn1 PE=1 SV=...	Ktn1	A0A087WQF8;A0A087WQG4;A0A087WS04;A0A087WS23;A0A087...	A0A087WPW5;A0A087WQD0;A0A087WS23;A0A087WQF8;A0A087...
A0A087WQH8	Probable UDP-sugar transporter protein SLC35A4 OS=...	Slc35a4	A0A087WQH8	A0A087WQH8
A0A087WR45	Proline-rich basic protein 1 OS=Mus musculus OX=10...	Prob1	A0A087WR45;Q3UKG2	A0A087WR45;Q3UKG2
A0A087WRT4	FAT atypical cadherin 1 OS=Mus musculus OX=10090 G...	Fat1	A0A087WRT4;A0A1L1SQU7;F2Z4A3	A0A1L1SQU7;A0A087WRT4;F2Z4A3
A0A087WRU0	Tensin 1 (Fragment) OS=Mus musculus OX=10090 GN=Tn...	Tns1	A0A087WRU0	Q9DBT6;E9Q0S6;A0A087WRU0;A0A087WQ94;A0A6I8MWZ2;A0A...
A0A087WSP0	Dystonin OS=Mus musculus OX=10090 GN=Dst PE=1 SV=1	Dst	A0A087WSP0;S4R1P5	S4R1P5;A0A087WSP0;S4R2A8;E9Q9X1;S4R1Y6

The full form is available in the web version

- Accession: ID of the database in which the protein is located
- Description: Functional description of the protein
- Gene: The gene encoding the protein
- Protein Group: Proteins contained in the Protein Group, i.e. proteins recognized by the same peptide
- Protein IDs: Proteins that are fully or partially recognized by any peptide recognizing a Protein Group, including proteins in the Protein Group

Table 3.4 Identification Results at Peptide Level

Peptide	Precursor Id	Precursor Charge	Missed Cleavages	Protein Group
AAAAAAAAAAGDSDSWDADTFSMEDPVRK	(UniMod:1)AAAAAAAAAAGDSDSWDADTFSMEDPVRK3	3	1	Q3UGC7
AAAAAAAAAAGDSDSWDADTFSMEDPVRK	(UniMod:1)AAAAAAAAAAGDSDSWDADTFSM(UniMod:35)EDPVRK...	3	1	Q3UGC7
AAAAAAAAATEQQGSNGPVK	(UniMod:1)AAAAAAAAATEQQGSNGPVK2	2	0	P82349
AAAAAAAAVGDPQPPQPEAPAQGLALDK	(UniMod:1)AAAAAAAAVGDPQPPQPEAPAQGLALDK3	3	0	Q5U430
AAAAAAAGGAALAVSTGLETATLQK	(UniMod:1)AAAAAAAGGAALAVSTGLETATLQK2	2	0	Q9CQ25
AAAAAAGAASGLPGPVAQGLK	(UniMod:1)AAAAAAGAASGLPGPVAQGLK2	2	0	Q91YE6
AAAAAAGPEMVR	(UniMod:1)AAAAAAGPEMVR2	2	0	P63085
AAAAAAGPEMVR	(UniMod:1)AAAAAAGPEM(UniMod:35)VR2	2	0	P63085
AAAAAASHLNLDALR	(UniMod:1)AAAAAASHLNLDALR2	2	0	Q8CH72
AAAAADLANR	AAAAADLANR2	2	0	Q9JHS4

The full form is available in the web version

- Peptide: Peptide sequence
- Precursor Id: Sequence of precursor ion matched by peptide, including modification and charge information
- Precursor Charge: Precursor ion charge
- Missed Cleavages: Number of potential cleavage sites in the peptide
- Protein Group: Protein group matched by this peptide

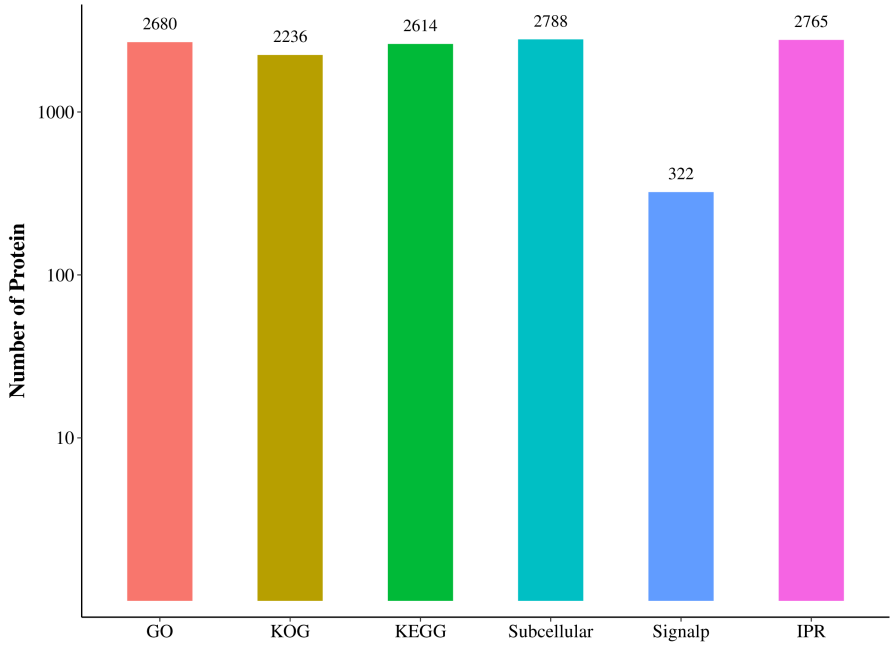
The sequence information of the identified proteins can be found in the following files: 2.Identification/All_ID.fasta

The detailed identification results at the protein level are available in: 2.Identification/Proteins_Summary.xlsx

The detailed identification results at the peptide level are available in: 2.Identification/Peptides_Summary.xlsx

3.3 Functional Annotation of Proteins

Functional annotation results of all identified proteins were statistically analyzed.



Overview of Functional Annotation Results

Note: Horizontal coordinates indicate different functional annotations; vertical coordinates indicate the number of proteins annotated with different functions.

Detailed results for all functional annotations are available at: 3.Annotation/All_annotation.xlsx

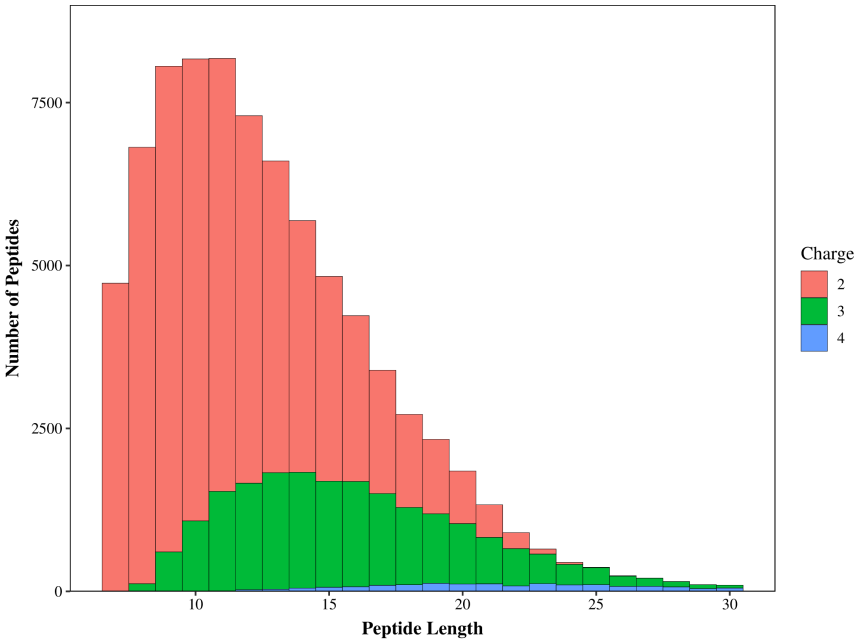
4 Quality Assessment

4.1 Quality Assessment of Qualitative Results

The data coming from the mass spectrometry run needs to undergo a series of quality control assessments after the database search is completed to ensure that the quality of the results meets the requirements including distribution of peptide length, peptide number, and the number of missed peptide cleavage sites.

4.1.1 Distribution of Peptide Lengths

Most of the peptide lengths were distributed over 7-20 amino acids, which is consistent with the general pattern of fragmentation based on enzymatic digestion and mass spectrometry. The lengths of the peptides identified by mass spectrometry were distributed in compliance with the quality control requirements.

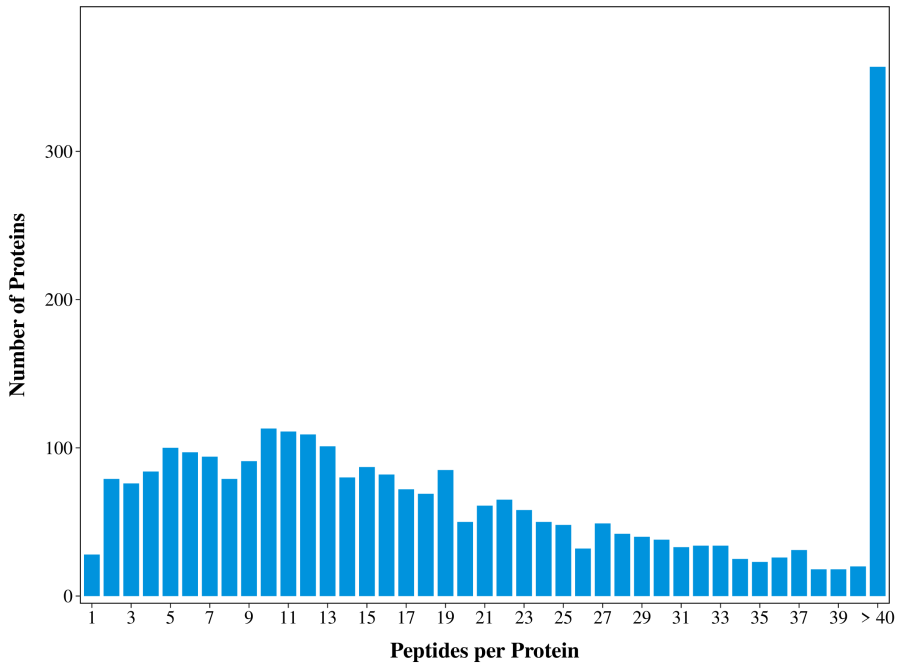


Peptide Length Distribution

Note: The horizontal coordinates indicate the length of the peptide; the vertical coordinates indicate the number of peptides with the corresponding length; the different colors represent the charge status of the detected peptides.

4.1.2 Distribution of Peptide Counts

The greater the number of peptides contained in a Protein Group, the more plausible the protein is.

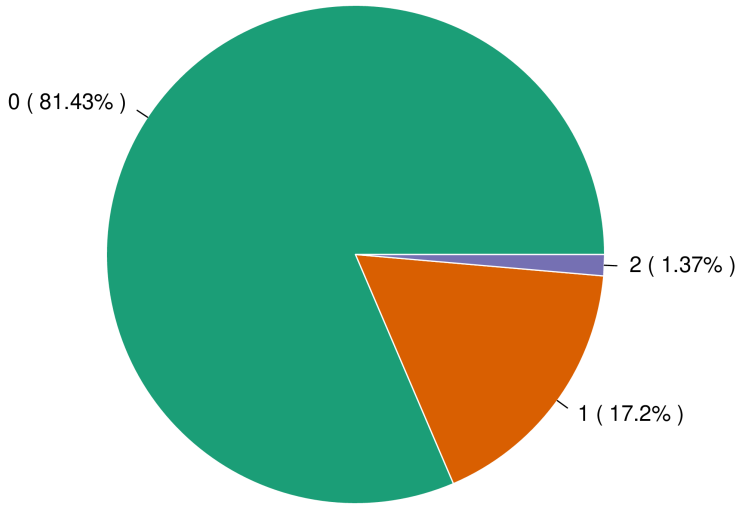


Peptide Number Distribution

Note: The horizontal coordinates indicate the number of peptides; the vertical coordinates indicate the number of proteins corresponding to the peptides.

4.1.3 Distribution of Missed Peptide Cleavage Site Counts

The distribution of missed peptide cleavage site numbers reflects the thoroughness of enzymatic cleavage - the more peptides with 0 missed sites, the more thorough the enzymatic cleavage is and the more favorable it is for identification.



Distribution of Missed Peptide Cleavage Site Numbers

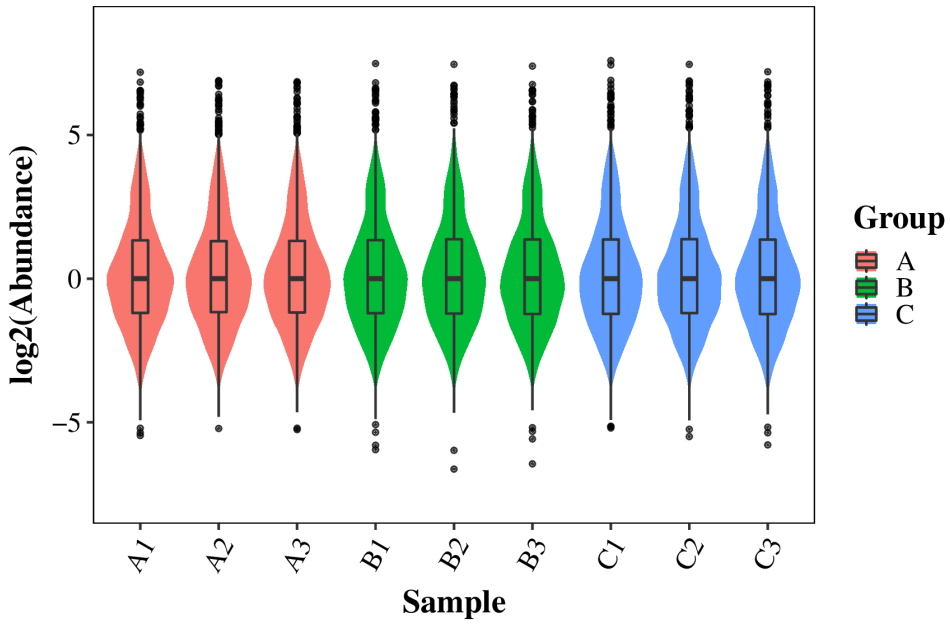
Note: The values outside the parentheses indicate the number of missed cleavage sites; the values inside the parentheses indicate the percentage of peptides with the corresponding number of missed sites.

All qualitative results are available at: 2.Identification

4.2 Quality Assessment of Quantitative Results

4.2.1 Distribution of Abundance Across Samples

Box plots provide not only a view of the dispersion of expression (abundance value) in individual samples but also a visual comparison of the overall expression across samples; while violin plots can be used to demonstrate the distribution and probability density of expression levels across samples. Therefore, joint box plots and violin plots can indirectly reflect the within-group consistency of biological samples from the same group. The following figure shows the distribution of abundance values of all samples after normalization. The box and violin shapes between biological replicate samples are reasonably close to each other.



Distribution of Abundance for Each Sample

Note: Horizontal coordinates indicate sample names; vertical coordinates indicate log2 values of abundance; different colors indicate different groups of samples

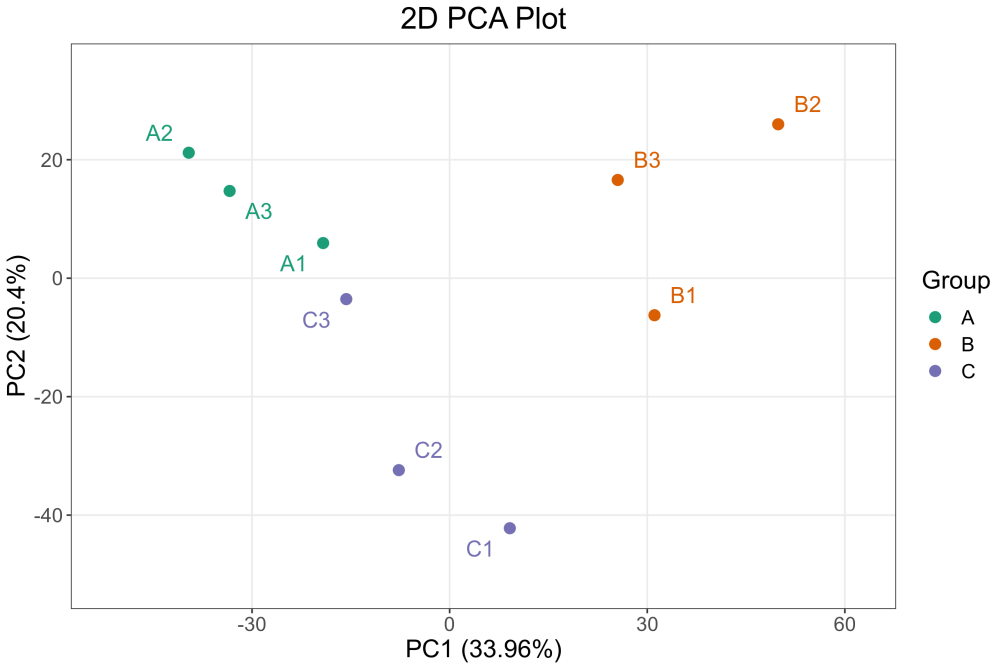
4.2.2 Principal Component Analysis (PCA)

By using multivariate statistical analysis, high-dimensional and complex data can be simplified and downsized with maximum preservation of the original information, allowing the establishment of a reliable mathematical model to summarize the proteomic characteristics of the research subjects. Principal component analysis (PCA) is an unsupervised pattern recognition method for statistical analysis of multidimensional data, which converts a set of potentially correlated variables into a set of linearly uncorrelated variables by orthogonal transformation. The converted set of variables is called principal components. This analysis is often used to study how to reveal the internal structure among multiple variables through a few principal components, i.e., to derive a few principal components from the original variables so that they retain as much information as possible about the original variables and are uncorrelated with each other. The usual mathematical processing is to make a linear combination of the original multiple indicators as a new composite indicator (Eriksson et al., 2006).

The data processing principle of PCA: the original data is compressed into a number of n principal components to characterize the original data set, PC1 denotes the most significant feature that can describe

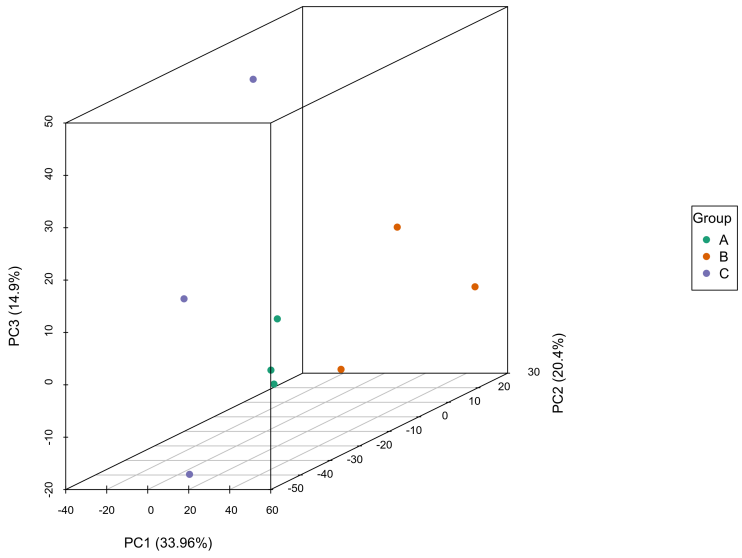
the multidimensional data matrix, PC2 denotes the most significant feature that can describe the data matrix excluding PC1, and PC3 ……PCn and so on.

Principal component analysis of the samples provides a preliminary understanding of the overall protein differences between groups of samples and the magnitude of variability between samples within groups (Chen et al.,2009).



PCA Results

Note: Horizontal and vertical coordinates represent the first and second principal components, respectively; the percentage in parentheses represents the contribution of the principal component to the sample difference; each dot in the plot represents a sample, with different colors representing different groups of samples.

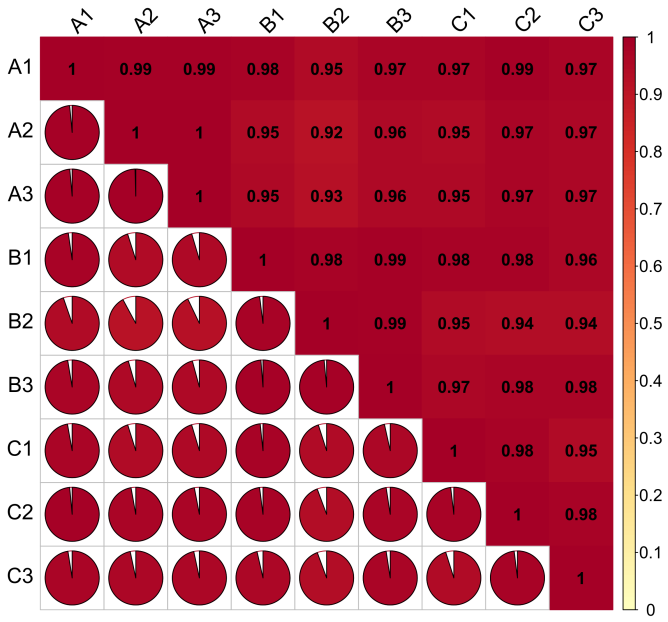


Three-dimensional (3D) PCA results

Note: PC1 denotes the first principal component; PC2 denotes the second principal component; PC3 denotes the third principal component.

4.2.3 Correlation Analysis

Biological replicates between samples within a group can be observed by inter-sample correlation analysis. Meanwhile, the higher the correlation coefficient of the within-group samples versus that of the between-group samples, the more reliable the obtained differentially expressed proteins are. Pearson's correlation coefficient (expressed as R) is used as an indicator to assess the correlation of biological replicates. The closer the |R| is to 1, the stronger the correlation between the two samples.



Inter-Sample Correlation

Note: Horizontal and vertical coordinates indicate sample names; the color transition from red to yellow represents the change in correlation from high to low. The size of the sector area in the plot represents the magnitude of the correlation coefficient between the corresponding horizontal and vertical samples; the numbers in the plot are the corresponding correlation coefficients between the horizontal and vertical samples.

All QC results are detailed at: 4.Quantification/QC

5 Differential Expression Analysis of Proteins

5.1 Screening Criteria for Differentially Expressed Proteins

The screening criteria for proteins with significant differences in this project were：

(1) With Replicates(≥2 replicates for each group): FC ≥ 1.5 or FC ≤ 0.6667 with P-value ≤ 0.05 was defined as significantly different proteins when samples were grouped into two differential groups; when samples were grouped into more than two differential groups, only P-value ≤ 0.05 had to be met.

(2) Without Replicates(=1 replicates for each group): when samples were divided into two differential groups, FC ≥ 1.5 or FC ≤ 0.6667 were defined as proteins with significant differences.

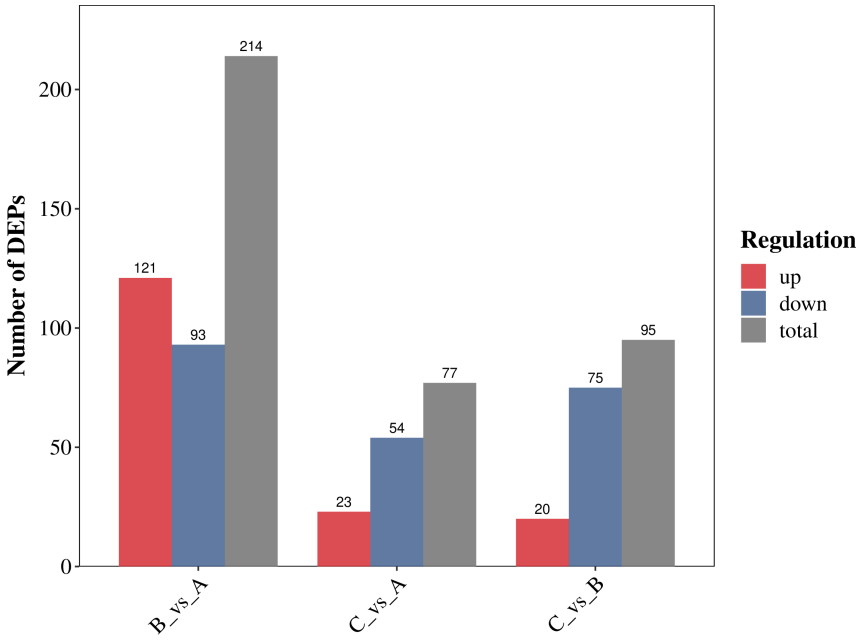
5.2 Results for Differentially Expressed Protein Screening

Statistical results of differentially expressed proteins for all groups are shown in the table below.

Table 5.1 Statistics on the Results of Differential Analysis

Comparison	Up regulation	Down regulation	Total DEPs
B_vs_A	121	93	214
C_vs_A	23	54	77
C_vs_B	20	75	95
A_vs_B_vs_C	0	0	965

- Comparison: Sample pairs for comparison, the former than the latter
- Up regulated: Number of up-regulated proteins in the former sample group
- Down regulated: Number of down-regulated proteins in the former sample group
- Total DEPs: Sum of the number of up- and down-regulated differentially expressed proteins



Statistical Plot of Differences Across Comparative Groups

The statistical table of the results of the variance analysis is detailed in: 5.Difference/DEPs_stat.xlsx

The statistical chart of the results of the variance analysis is detailed in: 5.Difference/DEPs_stat.png

The differentially expressed proteins calculated for each differential group are shown in the table below:

Table 5.2 Screening results for differentially expressed proteins

Accession	Description	Gene	Regulation
A0A075B5R2	Immunoglobulin heavy variable 7-3 (Fragment) OS=Mu...	Ighv7-3	sig
A0A087WQF8	Kinectin OS=Mus musculus OX=10090 GN=Ktn1 PE=1 SV=...	Ktn1	sig
A0A087WR45	Proline-rich basic protein 1 OS=Mus musculus OX=10...	Prob1	sig
A0A087WRT4	FAT atypical cadherin 1 OS=Mus musculus OX=10090 G...	Fat1	sig
A0A087WRU0	Tensin 1 (Fragment) OS=Mus musculus OX=10090 GN=Tn...	Tns1	sig
A0A0A6YVT8	Testis-specific gene 10 protein OS=Mus musculus OX...	Tsga10	sig
A0A0A6YX73	cAMP-dependent protein kinase type II-alpha regula...	Prkar2a	sig
A0A0G2JDV3	Guanylate-binding protein 6 OS=Mus musculus OX=100...	Gbp6	sig
A0A0G2JGI1	Nexilin OS=Mus musculus OX=10090 GN=Nexn PE=1 SV=1	Nexn	sig
A0A0R4J007	Paladin OS=Mus musculus OX=10090 GN=Pald1 PE=1 SV=...	Pald1	sig

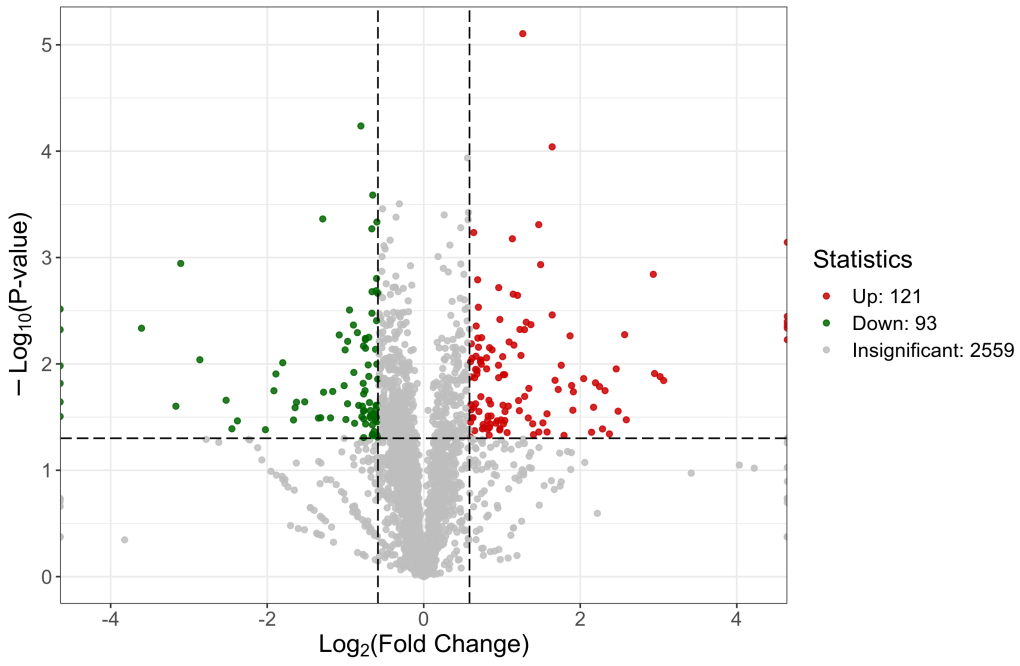
The full form is available in the web version

- Accession: ID number in the protein database
- Description: Functional description of the protein
- Gene: The gene encoding the protein
- Regulation: Up or down-regulation status of significant differences, with up being up-regulated and down being down-regulated; significant differences are expressed as sig when the differential group contains three or more samples

Differentially expressed proteins are listed in: 5.Difference/B_vs_A/B_vs_A_DEPs_annotation.xlsx

5.3 Volcano Plot of Differentially Expressed Proteins

The Volcano plot allows a quick view of the differences in the expression levels of differentially expressed proteins in two groups of samples, as well as the statistical significance of the differences. Volcano plots were made by taking the logarithm of 2 for the FC value of each differentially expressed protein and then taking the absolute value of the logarithm of 10 for the P-value.



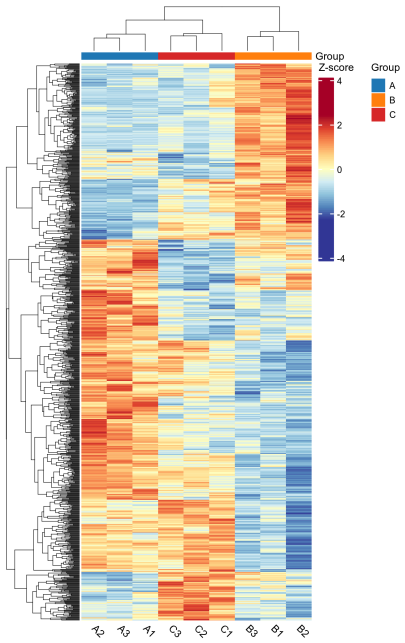
Volcano Plot of Differentially Expressed Proteins

Note: Horizontal coordinates represent log2 of FC; vertical coordinates represent -log10 of the P-value; red and green scatters represent up- and down-regulated differentially expressed proteins, respectively.

The differential protein volcano map in: 5.Difference/B_vs_A/B_vs_A_volcano_Log2FC_Pvalue.png

5.4 Clustering Heatmap of Differentially Expressed Proteins

In order to facilitate the visualization of differential protein expression patterns in different samples, differentially expressed proteins were subjected to z-score normalization, and clustering heatmaps were plotted. If clustering by rows, we can directly recognize which differentially expressed proteins have similar expression patterns; if clustering by columns, we can directly recognize inter-sample repetitions.



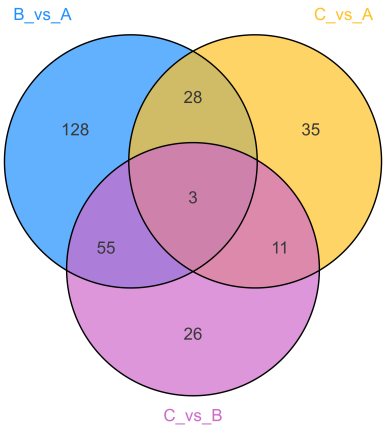
Clustering Heatmap of Differentially Expressed Proteins

Note: Rows represent the clustering of differentially expressed proteins; columns represent the clustering of samples; shorter dendrograms represent higher similarity.

Picture Path are available at: 5.Difference/A_vs_B_vs_C/A_vs_B_vs_C_heatmap.png

5.5 Venn Diagram of Differentially Expressed Proteins

Significantly differentially expressed proteins unique to a differential group or shared by different groups, as well as their distribution, can be visualized through Venn diagrams. For differential groups ≤ 5, the relationship between differentially expressed proteins in each group can be visualized by a Venn diagram; for more than 5 differential groups it is visualized by a flower plot.



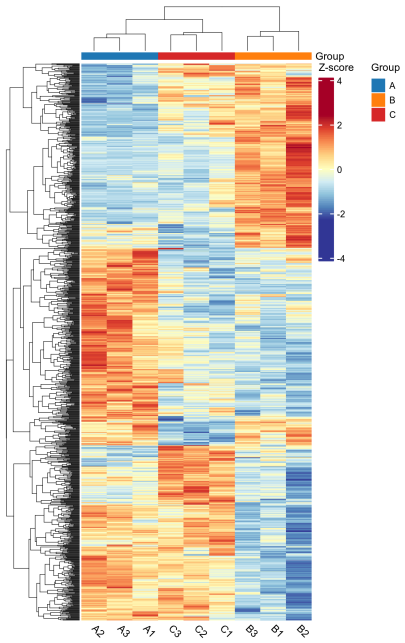
Venn Diagram of Differentially Expressed Protein Groups

Note: Each circle in the diagram represents a differential group, the number in the overlapping part represents the number of shared differentially expressed proteins between differential groups, and the number in the non-overlapping part represents the number of differentially expressed proteins unique to the differential group.

Detailed Venn diagram results are available at: 5.Difference/venn

5.6 Overall Clustering Heatmap

The significantly differentially expressed proteins in each differential group were taken and pooled, and then a clustering heatmap was plotted in the same way as above.



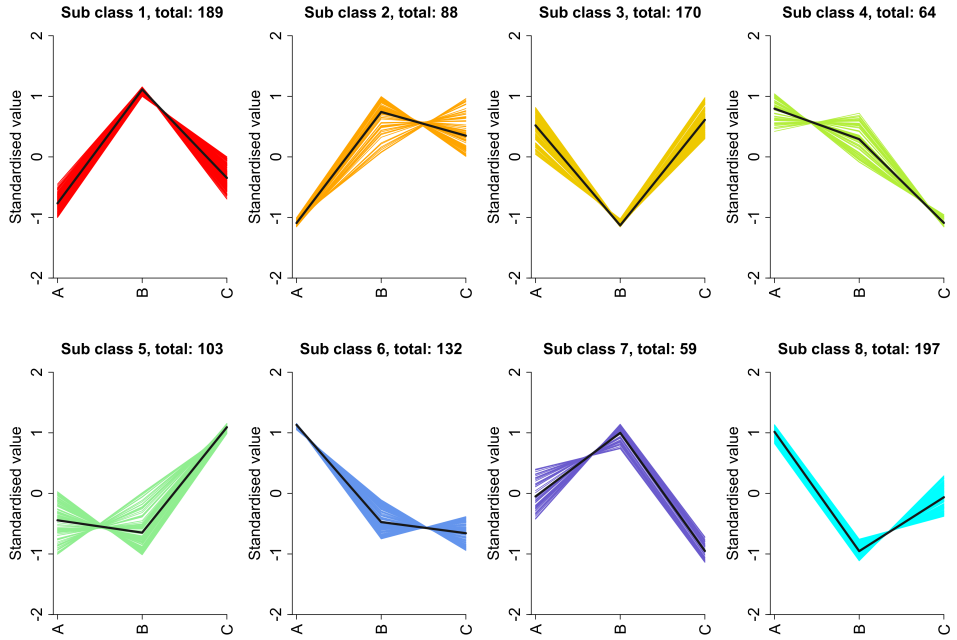
Clustering Heatmap of All Differentially Expressed Proteins

Note: Rows represent the clustering of proteins; columns represent the clustering of samples; shorter dendrograms represent higher similarity.

The clustering heatmap of all differentially expressed proteins is available at: 5.Difference/heatmap/all_DEPs_heatmap.png

5.7 K-means Analysis of Differentially Expressed Proteins

In order to investigate the trend of expression levels of differentially expressed proteins in different samples, the expressions of all differentially expressed proteins were normalized and centered, and then subjected to K-means analysis. The results are illustrated in the figure below.



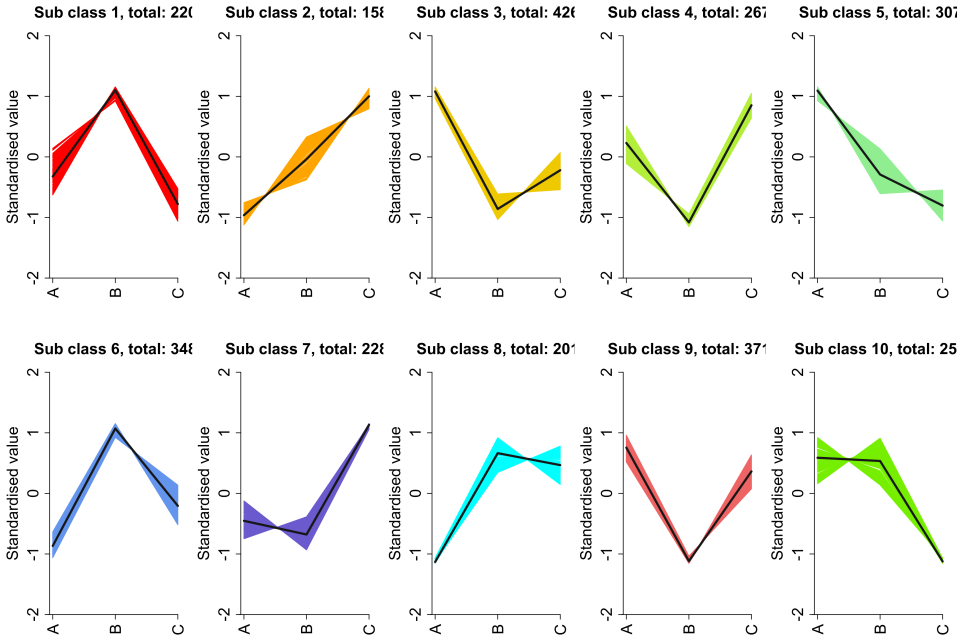
K-means Plot of All Differentially Expressed Proteins

Note: Horizontal coordinates represent samples from different groups; vertical coordinates represent normalized differential protein expression.

Detailed results of K-means analysis of all differentially expressed proteins are available at: 5.Difference/kmeans

5.8 K-means Analysis of All Proteins

In order to investigate the trend of expression levels of all identified proteins in different samples within this project, the expressions of all proteins were normalized and centered, and then subjected to K-means analysis. The results are illustrated in the figure below.



K-means Plot of All Proteins

Note: Horizontal coordinates represent samples from different groups; vertical coordinates represent normalized protein expression.

Detailed results of K-means analysis of all proteins are available at: 4.Quantification/kmeans

6 Bioinformatics Analysis

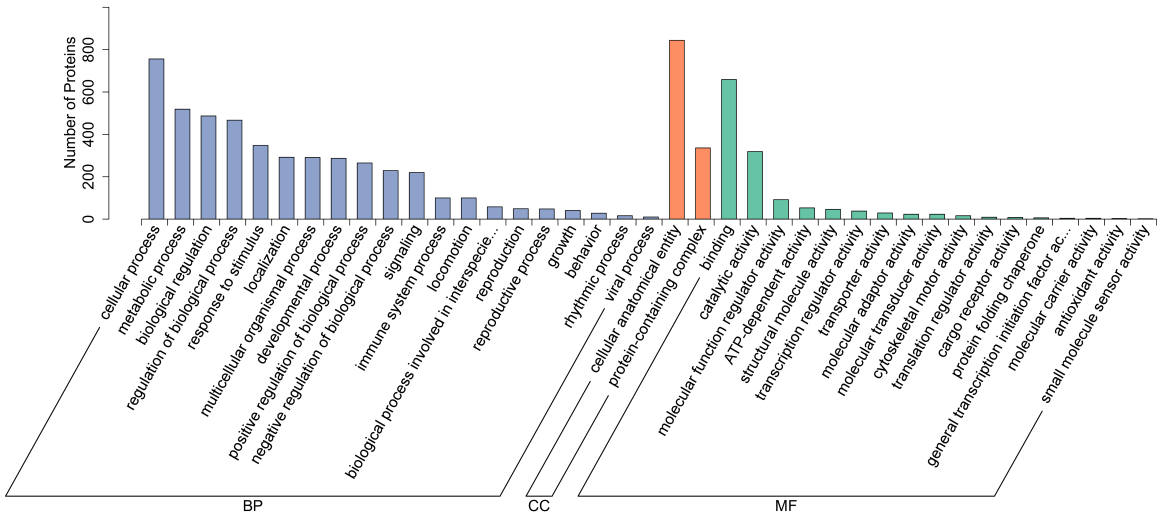
6.1 GO Functional Annotation and Enrichment Analysis of Differentially Expressed Proteins

6.1.1 Gene Ontology (GO) Analysis

Gene Ontology (GO, http://geneontology.org/) is an international standard classification system for gene functions. As a database established by the Gene Ontology Consortium (GOC), it aims to establish a linguistic vocabulary standard that is applicable to various species, qualifies and describes the functions of genes and proteins, and can be updated as research progresses. GO is divided into three components: molecular function, biological process, and cellular component.

(1) The number of differentially expressed proteins annotated with all secondary GO terms under the

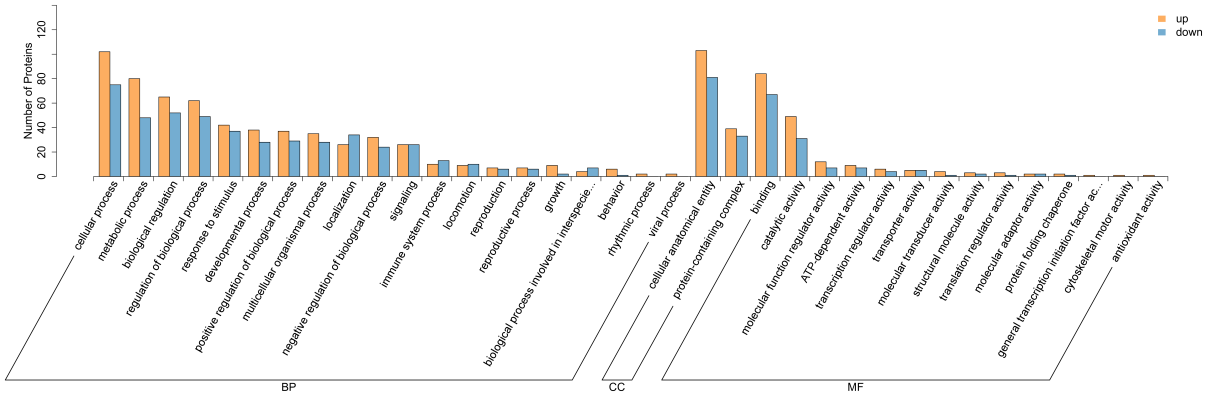
three primary classifications was counted. The statistical results of the GO classification of differentially expressed proteins are shown below:



Bar Chart of GO Classification

Note: Horizontal coordinates represent secondary GO terms; vertical coordinates represent the number of differentially expressed proteins associated with that GO term; different colors of bars represent different primary classifications.

(2) The number of differentially expressed proteins annotated with all secondary GO terms under the three primary classifications was counted. The statistical results of GO classification of up- and down-regulated differentially expressed proteins are shown below:



Bar Chart of Up- and Down-regulated Proteins under GO Classifications

Note: Horizontal coordinates represent secondary GO terms; vertical coordinates represent the number of differentially expressed proteins associated with that GO term; different colors of bars represent up- and down-regulation.

Detailed GO analysis results for differentially expressed proteins are available at: 6.Enrichment/GO

6.1.2 GO Enrichment Analysis

GO enrichment analysis involves GO functional terms, in which differentially expressed proteins are significantly enriched compared to the background of all identified differentially expressed proteins, thereby giving an indication of which biological functions the differentially expressed proteins are significantly associated with. This analysis first maps all differentially expressed proteins to individual terms in the Gene Ontology database and calculates the number of differentially expressed proteins for each term. Hypergeometric tests are then applied to identify GO terms with significant enrichment of differentially expressed proteins with the following equation:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N is the number of proteins with GO annotations in the background proteins; n is the number of differentially expressed proteins in N; M is the number of proteins in the background proteins that are annotated with a GO term; m is the number of differentially expressed proteins that are annotated with a GO term. The P-value was calculated, and a P-value ≤ 0.05 was used as the threshold, with GO terms meeting this threshold defined as GO terms with significant enrichment of differentially expressed pro-

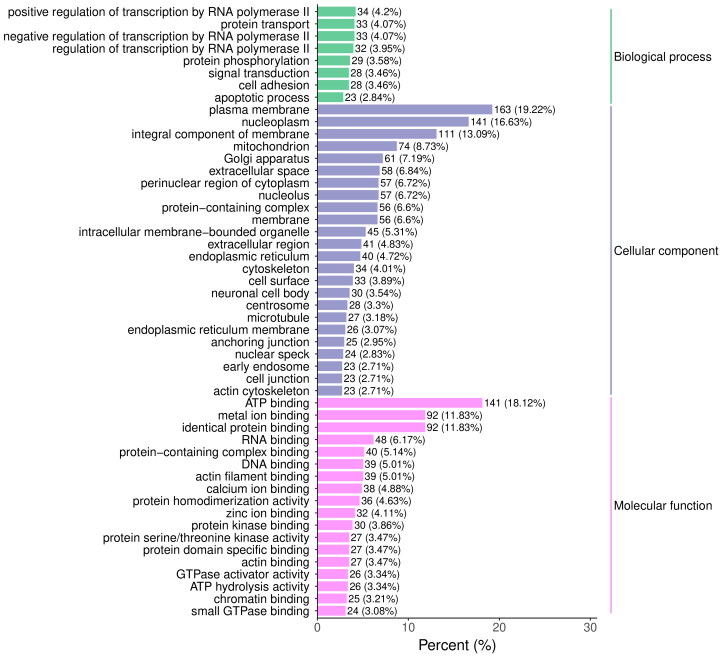
teins. GO significance analysis enables the identification of the major biological functions exercised by differentially expressed proteins. In this project, ClusterProfiler(Yu et al. 2012) was applied to calculate the GO enrichment results of differentially expressed proteins (as shown below).

Table 6.1 Results of GO Enrichment Analysis

GO	Description	DiffRatio	BgRatio	P-value
GO:0045944	positive regulation of transcription by RNA polyme...	34/810 4.2%	100/2335 4.28%	0.5970830
GO:0015031	protein transport	33/810 4.07%	75/2335 3.21%	0.0566298
GO:0000122	negative regulation of transcription by RNA polyme...	33/810 4.07%	91/2335 3.9%	0.4131402
GO:0006357	regulation of transcription by RNA polymerase II	32/810 3.95%	92/2335 3.94%	0.5326903
GO:0006468	protein phosphorylation	29/810 3.58%	81/2335 3.47%	0.4575772
GO:0007155	cell adhesion	28/810 3.46%	68/2335 2.91%	0.1559402
GO:0007165	signal transduction	28/810 3.46%	85/2335 3.64%	0.6743501
GO:0006915	apoptotic process	23/810 2.84%	64/2335 2.74%	0.4632318
GO:0045893	positive regulation of transcription, DNA-template...	23/810 2.84%	75/2335 3.21%	0.8062459
GO:0035556	intracellular signal transduction	22/810 2.72%	47/2335 2.01%	0.0559043

- GO: GO ID
- Description: Functional description of the GO term
- DiffRatio: Ratio of the number of differentially expressed proteins annotated with this GO term to the total number of differentially expressed proteins annotated with that level of classification
- BgRatio: Ratio of the number of background proteins annotated with this GO term to the total number of background proteins annotated with that level of classification
- P-value: P-value resulting from the hypergeometric significance test
- proteins: IDs of differentially expressed proteins annotated with this function

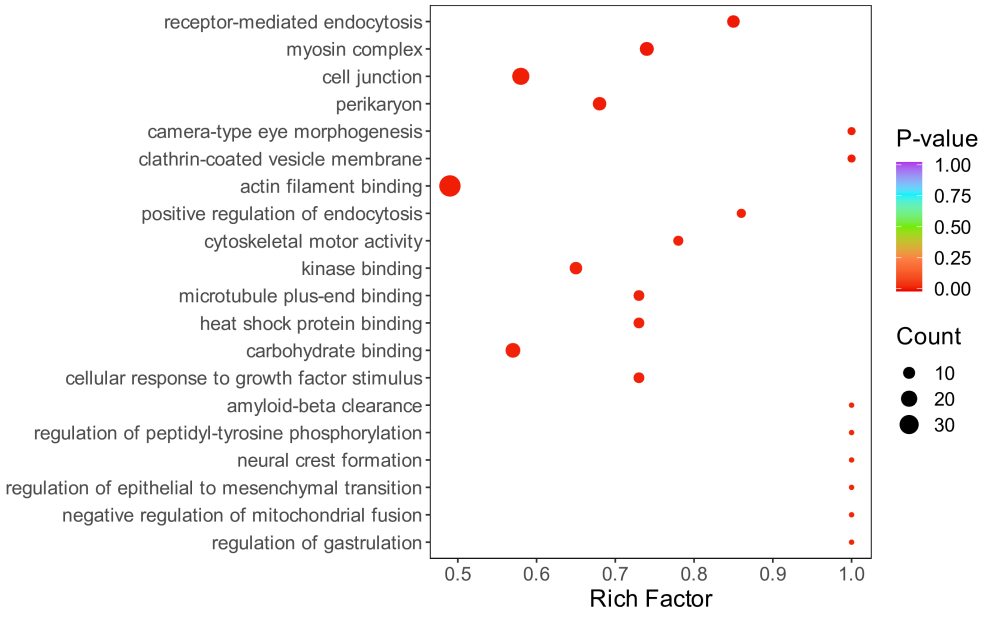
(1) The top 50 GO terms in terms of P-value ranking (sorted from smallest to largest) in the enrichment analysis results were selected to plot a bar chart (as shown below).



GO Enrichment Analysis Bar Chart

Note: The horizontal coordinate represents the number of differentially expressed proteins annotated with this term; the vertical coordinate represents the name of the GO term. Numbers in the figure represent the number of differentially expressed proteins annotated with this term. Numbers in parentheses are the ratio of the number of differentially expressed proteins annotated with this term to the total number of proteins with annotations. Labels on the far right represent the first level categories to which the GO term belongs.

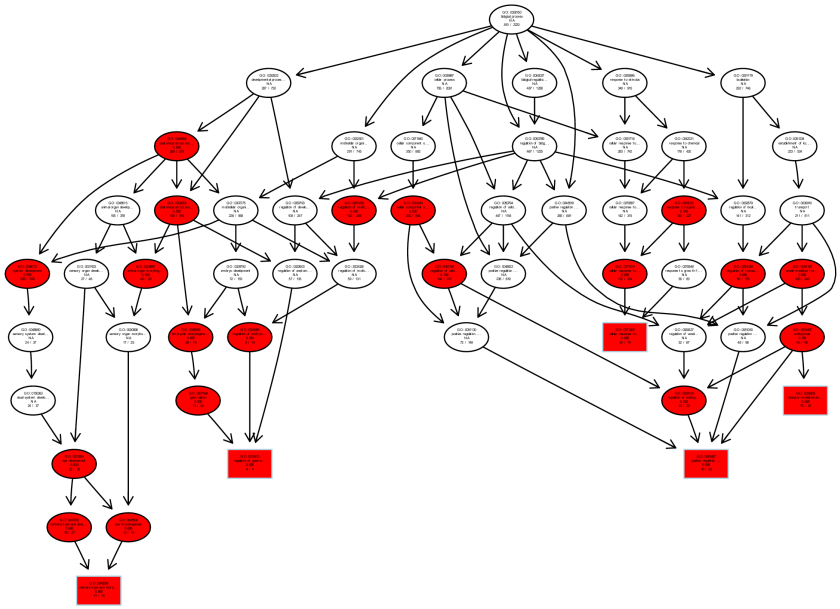
(2) The top 20 GO terms in terms of P-value ranking (sorted from smallest to largest) in the enrichment analysis results were selected to plot a bubble plot (as shown below):



GO Enrichment Analysis Bubble Plot

Note: The horizontal coordinate indicates the enrichment fold (the ratio of the number of differentially expressed proteins enriched with this term against the number of annotated proteins) - the larger the enrichment fold, the higher the enrichment level of differentially expressed proteins. The vertical coordinate indicates the name of the GO term. The change of the color of the dots from blue to red represents the change of the P-value from large to small - the smaller the P-value, the higher the statistical significance. The size of the dot reflects the number of differentially expressed proteins annotated with the corresponding term.

(3) The terms with enrichment of differentially expressed proteins are plotted in a top GO-directed acyclic graph (DAG), which visualizes the GO nodes (terms) enriched with differentially expressed proteins and their hierarchical relationships, providing a graphical display of GO enrichment analysis results. The branching represents the containment relationship, and the range of functional descriptions defined from the top to the bottom is increasingly more specific. The topGO molecular function directed acyclic graph of differentially expressed proteins between samples is shown below:



GO Enrichment Directed Acyclic Graph

Note: Each node represents a GO term, with the rectangle representing the top 10 selected GO terms with the highest enrichment, and the ellipse representing the contained nodes. The colors of the rectangles and ellipses represent the relative enrichment. From bright yellow to dark red indicates a decreasing p-value, i.e., increasing significance, while white represents non-significance. Each node shows 4 rows of data, i.e., the ID of the GO term, its functional description, corrected P-value, and the ratio of the number of differentially expressed proteins for that GO term against the total number of differentially expressed proteins, respectively.

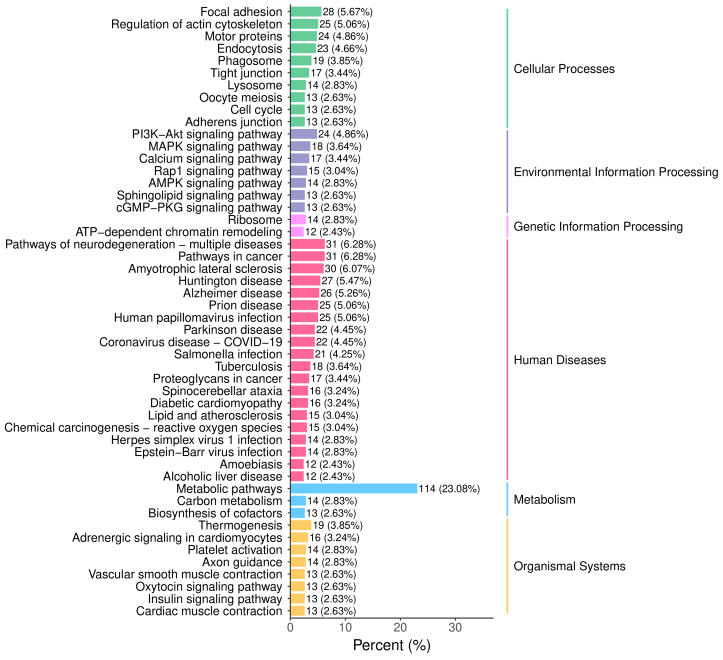
Detailed GO enrichment analysis results for differentially expressed proteins are available at: 6.Enrichment/GO

6.2 KEGG Functional Annotation and Enrichment Analysis of Differentially Expressed Proteins

6.2.1 KEGG Analysis

In organisms, different proteins exercise their biological functions in coordination with each other, so that the pathway-based analysis can facilitate a better understanding of their biological functions. KEGG is a major public database concerning pathways (https://www.genome.jp/kegg/), which represents a net-

work of information linking interactions (such as metabolic pathways, complexes, biochemical reactions, and so on) between known molecules. KEGG pathways mainly involve metabolism, genetic information processing, environmental information processing, cellular processes, human diseases, and drug development. Pathway analysis enables the identification of the most important bio-metabolic pathways and signaling pathways in which proteins are involved. Typically, when no single species was specified, the identified proteins were compared against the KEGG species database (animals, plants, fungi, bacteria, and so on) to derive KEGG annotation results. The number of differentially expressed proteins contained in each KEGG pathway was counted and plotted in a bar chart. Only the top 50 KEGG pathways with the highest number of differentially expressed proteins (in descending order) are displayed here. If the number of KEGG pathways is less than 50, all of them are displayed, as shown in the figure below:

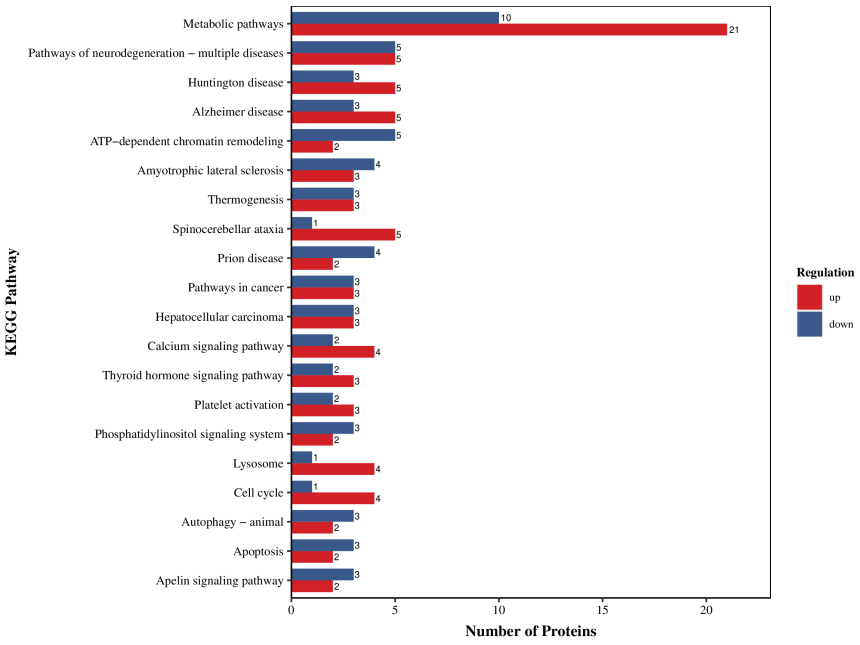


Bar Chart of KEGG Categories

Note: Horizontal coordinates indicate the number of differentially expressed proteins annotated with the pathway; vertical coordinates indicate the name of the KEGG pathway. Numbers in the figure represent the number of differentially expressed proteins annotated with the pathway. Numbers in parentheses are the ratio of the number of differentially expressed proteins annotated with the pathway to the total number of proteins with annotations. Labels on the far right represent the first level categories to which the KEGG pathway belongs.

The number of up- and down-regulated differentially expressed proteins within each KEGG pathway

was counted and plotted into a bar chart, which displays only the top (in descending order) 20 functions with the highest number of differentially expressed proteins.

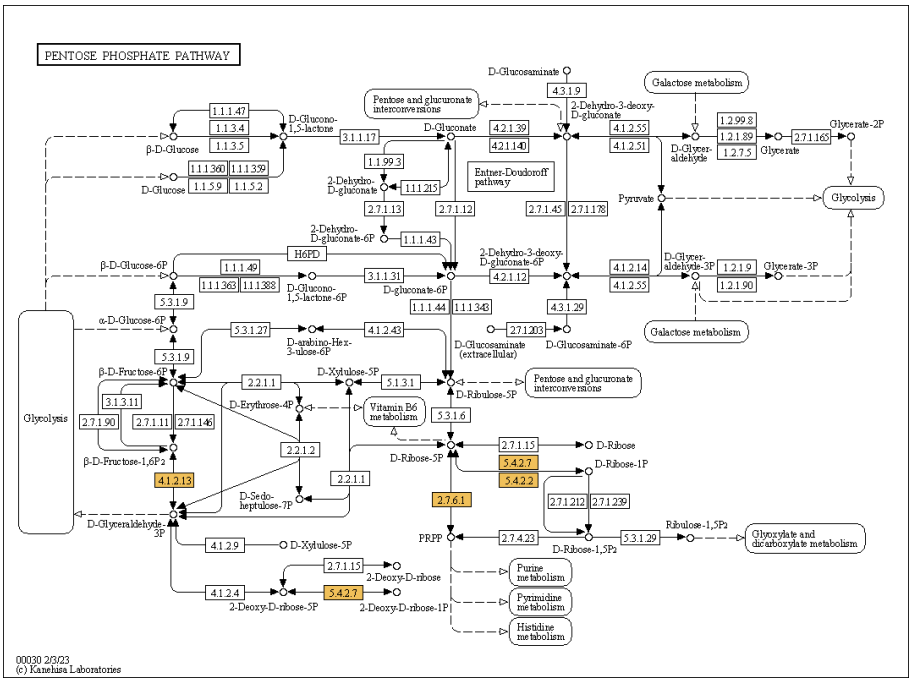


Bar

Chart for Comparison of KEGG Classified Up- and Down-Regulated Proteins

Note: Horizontal coordinates show the number of differentially expressed proteins annotated with the corresponding function; vertical coordinates are the names of KEGG functional categories; red and blue colors represent up- and down-regulated differentially expressed proteins, respectively.

KEGG pathways were profiled, with up- and down-regulated differentially expressed proteins marked with red and green shades, respectively; both up- and down-regulated differentially expressed proteins marked with a blue shade; differentially expressed proteins without significant up- and down-regulation are marked with a yellow shade (a case of comparison between three groups of samples).



KEGG Pathway Map of Differentially Expressed Proteins

Note: Red underlined markers are up-regulated differentially modified proteins, green underlined markers are down-regulated differentially modified proteins, blue underlined markers are both up- and down-regulated differentially modified proteins, and yellow underlined markers are proteins with no up- and down-regulation distinguishing significant differentially modified proteins (in the case of a comparison of 3 groups of samples).

Detailed KEGG analysis results for differentially expressed proteins are available at: 6.Enrichment/KEGG

6.2.2 KEGG Enrichment Analysis

Pathway enrichment analysis was performed in the same way as GO enrichment analysis, which involves taking pathways in the KEGG database as units and applying hypergeometric tests to identify pathways with significant enrichment of differentially expressed proteins compared to the background, i.e., all the differentially expressed proteins identified. Pathway enrichment analysis allows the identification of the most prominent biochemical and metabolic pathways and signaling pathways in which differentially expressed proteins are involved. The results of the KEGG enrichment analysis are shown below:

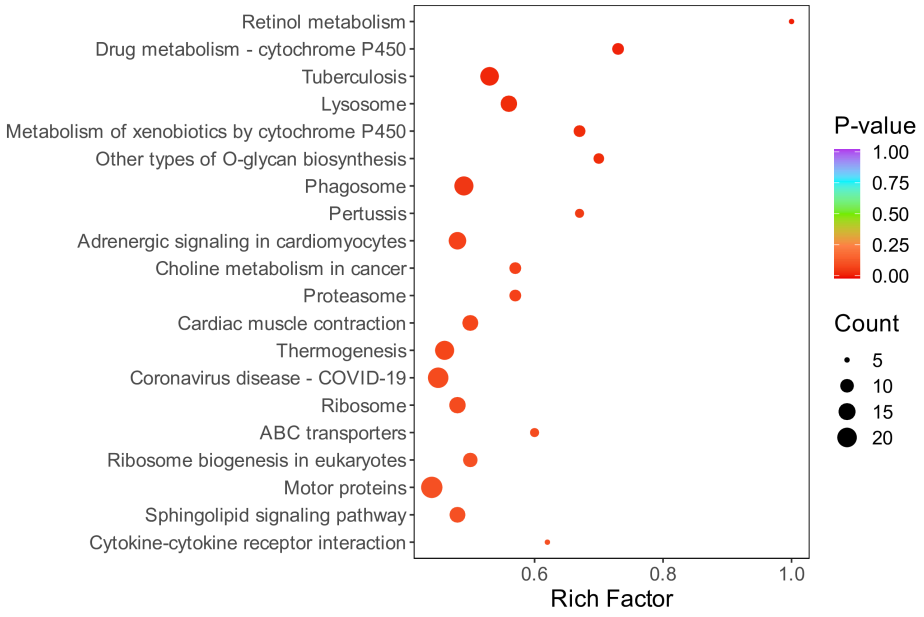
Table 6.2 Results of KEGG Enrichment Analysis

ko_ID	KEGG_Pathway	DiffRatio	BgRatio	P-value
ko04510	Focal adhesion	28/494 5.67%	72/1432 5.03%	0.2472086
ko04810	Regulation of actin cytoskeleton	25/494 5.06%	68/1432 4.75%	0.3884254
ko04814	Motor proteins	24/494 4.86%	55/1432 3.84%	0.0965472
ko04144	Endocytosis	23/494 4.66%	68/1432 4.75%	0.5942829
ko04145	Phagosome	19/494 3.85%	39/1432 2.72%	0.0446418
ko04530	Tight junction	17/494 3.44%	47/1432 3.28%	0.4584794
ko04142	Lysosome	14/494 2.83%	25/1432 1.75%	0.0214182
ko04114	Oocyte meiosis	13/494 2.63%	32/1432 2.23%	0.2871857
ko04110	Cell cycle	13/494 2.63%	38/1432 2.65%	0.5768003
ko04520	Adherens junction	13/494 2.63%	38/1432 2.65%	0.5768003

The full form is available in the web version

- ko_ID: KEGG Pathway ID
- KEGG Pathway: KEGG Pathway Description
- DiffRatio: Ratio of the number of differentially expressed proteins annotated with this KEGG pathway to the total number of differentially expressed proteins
- BgRatio: Ratio of the number of background proteins annotated with this KEGG pathway to the total number of background proteins
- P-value: P-value resulting from the hypergeometric significance test
- proteins: IDs of differentially expressed proteins annotated with this function

A bubble plot was used as a graphical presentation of the results of the KEGG enrichment analysis. In this plot, the degree of KEGG enrichment is indicated by the enrichment fold, P-value, and the number of differentially expressed proteins enriched in this pathway. Among them, the larger the enrichment fold, the greater the degree of enrichment; the smaller the P-value, the more significant the enrichment. The 20 pathway entries with the most significant enrichment results were selected for plotting. If there were fewer than 20 entries, all of them were displayed in the plot.



KEGG Enrichment Analysis Bubble Plot

Note: The horizontal coordinate indicates the enrichment fold (the ratio of the number of differentially expressed proteins enriched with this term against the number of annotated proteins) - the larger the enrichment fold, the higher the enrichment level of differentially expressed proteins. The vertical coordinate indicates the KOG pathway. The change of the color of the dots from blue to red represents the change of the P-value from large to small - the smaller the P-value, the higher the statistical significance. The size of the dot reflects the number of differentially expressed proteins annotated with the corresponding function.

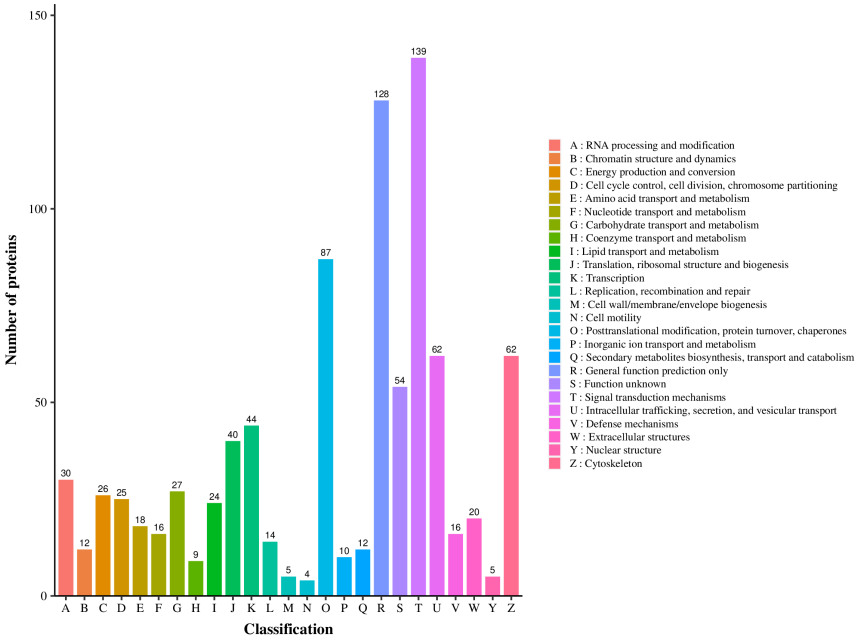
Detailed KEGG analysis results for differentially expressed proteins are available at: 6.Enrichment/KEGG

6.3 COG/KOG Functional Annotation and Enrichment Analysis of Differentially Expressed Proteins

6.3.1 COG/KOG Analysis

COG stands for Cluster of Orthologous Groups of proteins. The proteins that make up each COG are hypothesized to be derived from the same ancestral protein. Orthologs refer to proteins that come from different species, evolved from vertical lineages (species formation), and typically retain the same

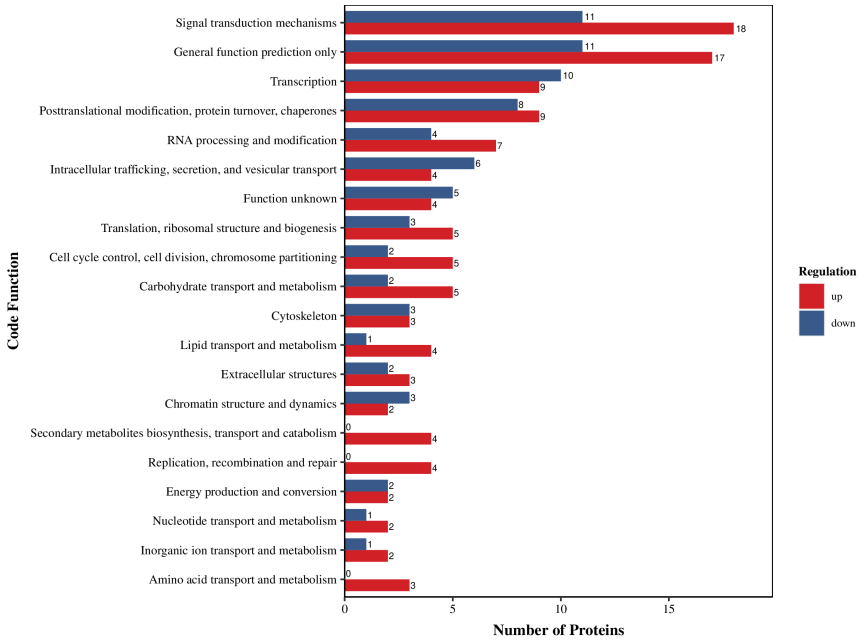
functionality as the original protein. The COGs are classified into 26 functional categories, as detailed at http://www.ncbi.nlm.nih.gov/COG. Prokaryotes are annotated with COGs and eukaryotes with KOGs. The KOG database was used for annotation in this project. The number of differentially expressed proteins contained in each KOG term was counted and plotted in a bar chart as shown below:



Bar Chart of KOG Annotations

Note: Horizontal coordinates represent KOG functional categories; vertical coordinates represent the number of differentially expressed proteins annotated with the corresponding function; the legend on the right side shows depictions of the functional categories.

The number of up- and down-regulated differentially expressed proteins was counted for each KOG functional term and plotted in a bar chart as shown below:



Comparative Bar Chart of KOG Annotated Up- and Down-Regulated Proteins

Note: Horizontal coordinates show the number of differentially expressed proteins annotated with the corresponding function; vertical coordinates are the names of KOG functional categories; red and blue colors represent up- and down-regulated differentially expressed proteins, respectively.

Detailed COG analysis results for differentially expressed proteins are available at: 6.Enrichment/KOG

6.3.2 KOG Enrichment Analysis

The KOG enrichment analysis was performed in the same way as the GO enrichment analysis to identify enriched functional categories with statistical significance. The results of the KOG enrichment analysis are shown below:

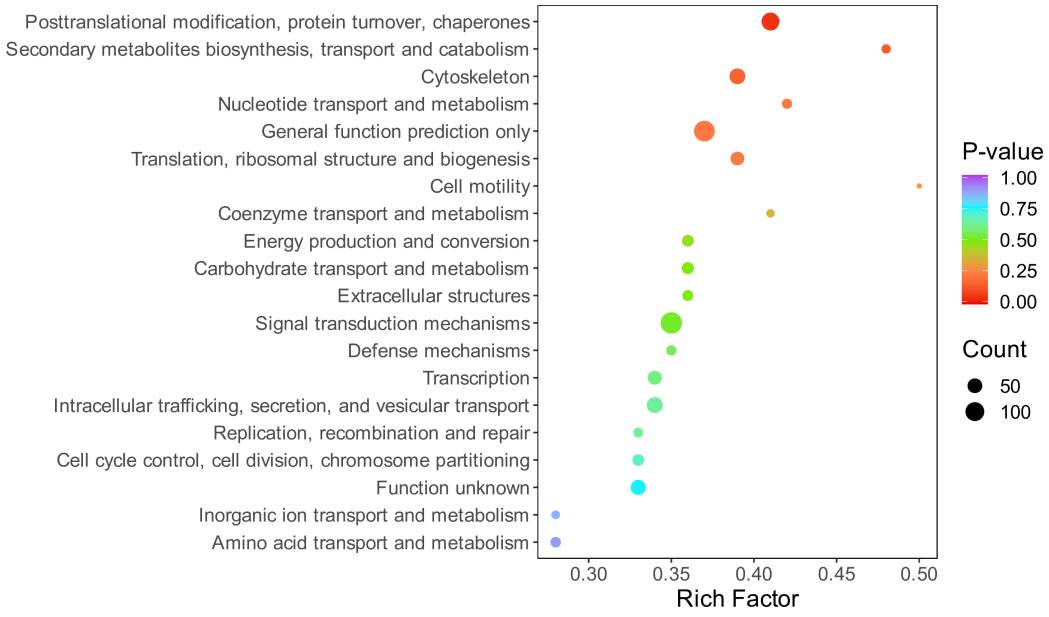
Table 6.3 Results of KOG Enrichment Analysis

Code	Code Function	DiffRatio	BgRatio	P-value
T	Signal transduction mechanisms	139/778 17.87%	399/2236 17.84%	0.5137137
R	General function prediction only	128/778 16.45%	348/2236 15.56%	0.2155283
O	Posttranslational modification, protein turnover, ...	87/778 11.18%	213/2236 9.53%	0.0314466
Z	Cytoskeleton	62/778 7.97%	160/2236 7.16%	0.1576601
U	Intracellular trafficking, secretion, and vesicula...	62/778 7.97%	183/2236 8.18%	0.6352765
S	Function unknown	54/778 6.94%	166/2236 7.42%	0.7634742
K	Transcription	44/778 5.66%	129/2236 5.77%	0.6008374
J	Translation, ribosomal structure and biogenesis	40/778 5.14%	103/2236 4.61%	0.2178984
A	RNA processing and modification	30/778 3.86%	103/2236 4.61%	0.9117413
G	Carbohydrate transport and metabolism	27/778 3.47%	76/2236 3.4%	0.4898519

The full form is available in the web version

- Code: KOG ID
- Code Function: Description of KOG classification
- DiffRatio: Ratio of the number of differentially expressed proteins annotated with this KOG classification to the total number of differentially expressed proteins
- BgRatio: Ratio of the number of background proteins annotated with this KOG classification to the total number of background proteins
- P-value: P-value resulting from the hypergeometric significance test
- proteins: IDs of differentially expressed proteins annotated with this function

The top 20 functional classifications in terms of P-value ranking (sorted from smallest to largest) were selected from the KOG enrichment analysis results of differentially expressed proteins to plot a bubble plot of the enriched terms or all of them were displayed if there were fewer than 20, as shown in the following figure:



KOG Enrichment Analysis Bubble Plot

Note: The horizontal coordinate indicates the enrichment fold (the ratio of the number of differentially expressed proteins enriched with this term against the number of annotated proteins) - the larger the enrichment fold, the higher the enrichment level of differentially expressed proteins. The vertical coordinate indicates the functional description of the KOG term. The change of the color of the dots from blue to red represents the change of the P-value from large to small - the smaller the P-value, the higher the statistical significance. The size of the dot reflects the number of differentially expressed proteins annotated with the corresponding function.

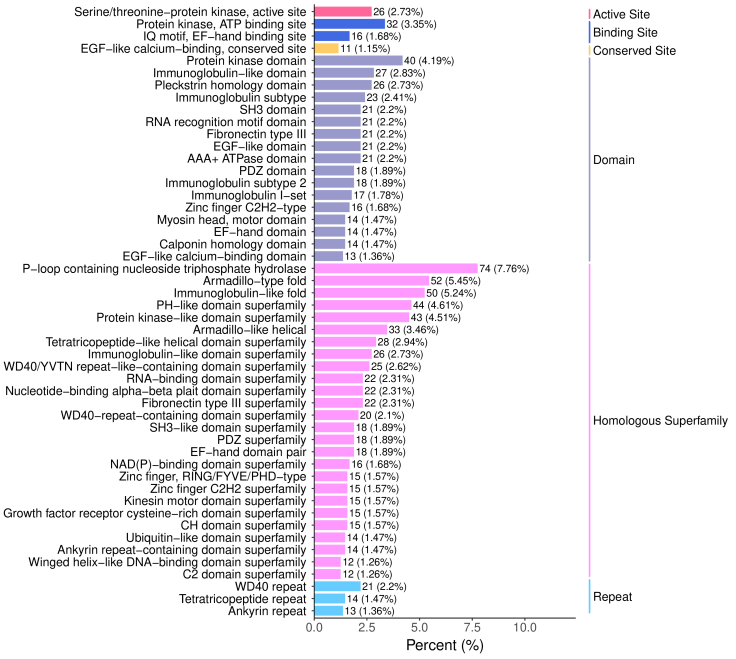
Detailed KOG analysis results for differentially expressed proteins are available at: 6.Enrichment/KOG

6.4 Domain Functional Annotation and Enrichment Analysis of Differentially Expressed Proteins

6.4.1 Structural Domain Annotations

Protein domains are certain components that are repeated in different protein molecules with similar sequences, structures, and functions. They represent units of protein evolution. The combinations and

distributions among different domains do not follow a stochastic model but rather exhibit a pattern in which some domains are highly combinable and others are rarely combined with other domains. Research on protein domains is important for understanding the biological functions of proteins and their evolution. InterPro (Finn et al. 2017, https://www.ebi.ac.uk/interpro/) is a commonly used protein domain database that contains other commonly used protein domain databases including Pfam, ProDom, and SMART. The number of differentially expressed proteins contained in each InterPro term (IPR) was statistically calculated and plotted in a bar chart. Only the top 50 IPRs with the highest number of differentially expressed proteins (in descending order) are shown here, and all of them are shown if there are less than 50 IPRs, as illustrated in the figure below.

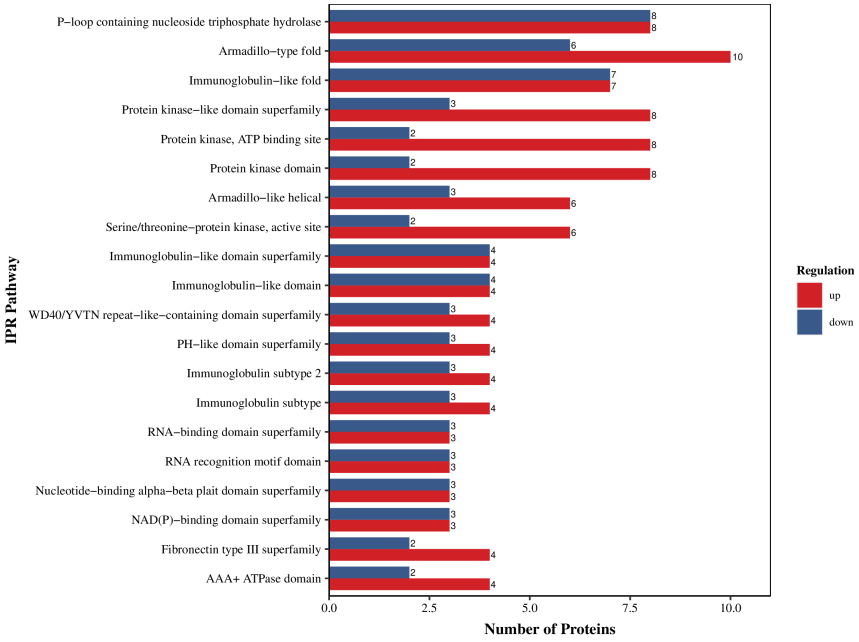


Bar Chart of Domain Annotations

Note: The horizontal coordinate represents the number of differentially expressed proteins annotated with this IPR; the vertical coordinate represents the description of the IPR term. Numbers in the figure represent the number of differentially expressed proteins annotated with this IPR. Numbers in parentheses are the ratio of the number of differentially expressed proteins annotated with the IPR to the total number of proteins with annotations. Labels on the far right represent IPR categories.

The number of up- and down-regulated differentially expressed proteins within each IPR pathway was counted and plotted into a bar chart, which displays only the top (in descending order) 20 IPRs with

the highest number of differentially expressed proteins.



Bar Chart Comparing Annotated Up- and Down-regulated Domains

Note: Horizontal coordinates show the number of differentially expressed proteins annotated with the corresponding function; vertical coordinates are the names of KOG functional categories; red and blue colors represent up- and down-regulated differentially expressed proteins, respectively.

Detailed domain analysis results for differentially expressed proteins are available at: 6.Enrichment/IPR

6.4.2 Structural Domain Enrichment Analysis

Structural domain enrichment analysis was performed in the same way as GO enrichment analysis to identify structural domains or families of differentially expressed proteins that are statistically significantly enriched. Such structural domains or families of differentially expressed proteins may contribute to differences in physiological functions. The results of the structural domain enrichment analysis are shown below:

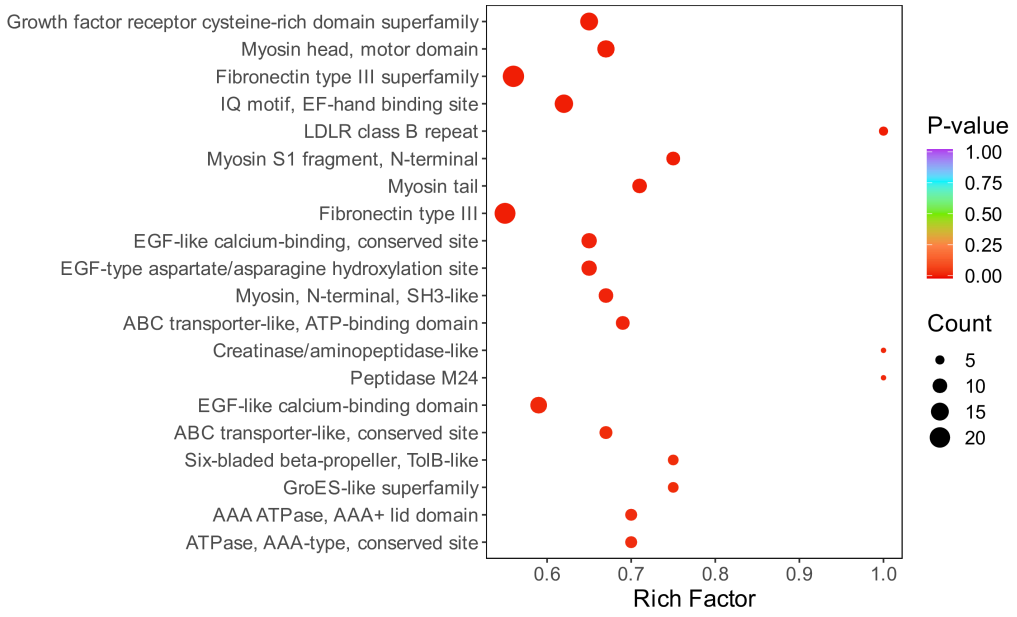
Table 6.4 Results of Structure Domain Enrichment Analysis

IPR_acc	IPR_desc	DiffRatio	BgRatio	P-value
IPR027417	P-loop containing nucleoside triphosphate hydrolas...	74/954 7.76%	207/2765 7.49%	0.3737254
IPR016024	Armadillo-type fold	52/954 5.45%	134/2765 4.85%	0.1631719
IPR013783	Immunoglobulin-like fold	50/954 5.24%	117/2765 4.23%	0.0361860
IPR011993	PH-like domain superfamily	44/954 4.61%	120/2765 4.34%	0.3375497
IPR011009	Protein kinase-like domain superfamily	43/954 4.51%	108/2765 3.91%	0.1400553
IPR000719	Protein kinase domain	40/954 4.19%	90/2765 3.25%	0.0299611
IPR011989	Armadillo-like helical	33/954 3.46%	88/2765 3.18%	0.3102056
IPR017441	Protein kinase, ATP binding site	32/954 3.35%	75/2765 2.71%	0.0844678
IPR011990	Tetratricopeptide-like helical domain superfamily	28/954 2.94%	70/2765 2.53%	0.1960086
IPR007110	Immunoglobulin-like domain	27/954 2.83%	65/2765 2.35%	0.1414123

The full form is available in the web version

- IPR_acc: IPR database login number
- IPR_desc: IPR function description
- DiffRatio: Ratio of the number of differentially expressed proteins annotated with this IPR to the total number of differentially expressed proteins
- BgRatio: Ratio of the number of background proteins annotated with this IPR to the total number of background proteins
- P-value: P-value resulting from the hypergeometric significance test
- proteins: IDs of differentially expressed proteins annotated with this function

The top 20 IPRs in terms of P-value ranking (sorted from smallest to largest) were selected from the domain enrichment analysis results of differentially expressed proteins to plot a bubble plot of the enriched entries or all of them were displayed if there were fewer than 20 IPRs, as shown in the following figure:



Bubble Plot of Structural Domain Enrichment Analysis

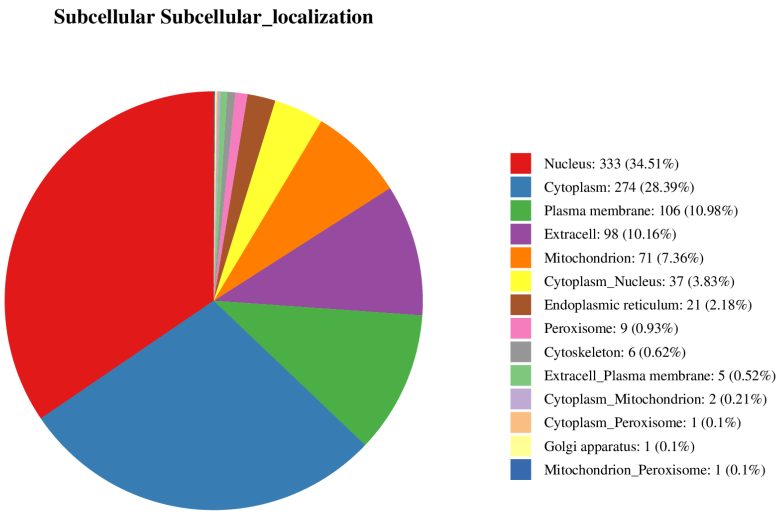
Note: The horizontal coordinate indicates the enrichment fold (the ratio of the number of differentially expressed proteins enriched with this term against the number of annotated proteins) - the larger the enrichment fold, the higher the enrichment level of differentially expressed proteins. The vertical coordinate indicates the description of the IPR term. The change of the color of the dots from blue to red represents the change of the P-value from large to small - the smaller the P-value, the higher the statistical significance. The size of the dot reflects the number of differentially expressed proteins annotated with the corresponding function.

Detailed domain analysis results for differentially expressed proteins are available at: 6.Enrichment/IPR

6.5 Subcellular Localization of the Differentially Expressed Proteins

An organism's cell is a highly organized structure, with the intracellular contents being divided into different organelles or cellular regions based on spatial distribution and functions, such as the nucleus, the Golgi apparatus, the endoplasmic reticulum, the mitochondria, the cytoplasm, and the cell membrane. Proteins are synthesized in ribosomes and then transported to specific organelles by protein sorting signals, while some proteins are secreted out of the cell or remain in the cytoplasm. Only when transported to

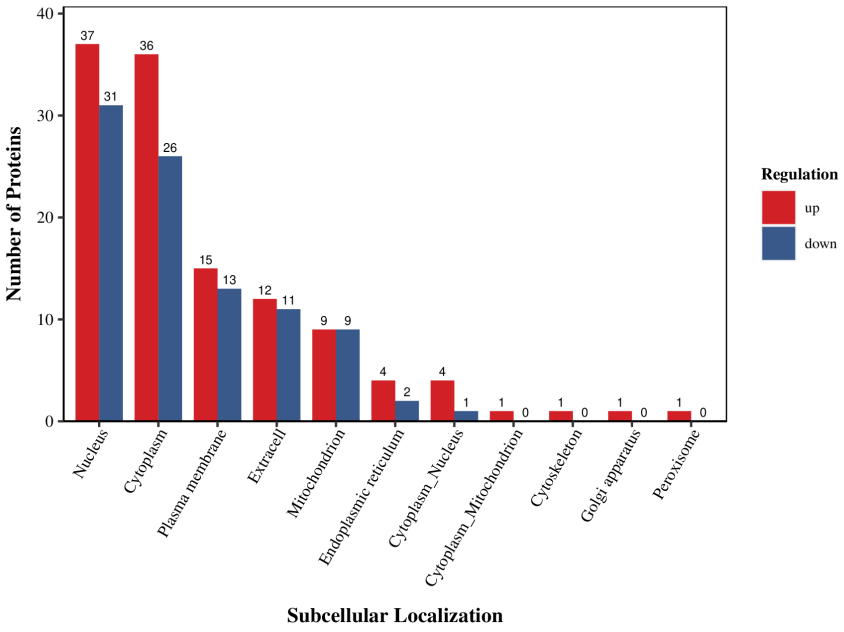
the correct site, can they participate in different cellular life activities. Therefore, information about the subcellular localization of proteins is very important for understanding living organisms. Predictions of subcellular localization in prokaryotes and eukaryotes were made using the PSORTb software and the WoLF PSORT software, respectively. The number of differentially expressed proteins contained per subcell was statistically calculated and plotted in a pie chart as shown below.



Pie Chart of Subcellular Localization Results

Note: Different colors represent different subcells; numbers out of parentheses are the numbers of differentially expressed proteins annotated with the corresponding subcells; numbers within parentheses are the percentage of annotated differentially expressed proteins in relation to all differentially expressed proteins with subcellular annotations.

The differentially expressed proteins in each differential grouping were statistically analyzed for subcellular localization, and the number of up- and down-regulated differentially expressed proteins were calculated respectively and visualized in a bar chart. When there were three or more samples in the differential group, up- and down-regulation could not be differentiated, so no statistics were performed.



Bar Chart Comparing Up- and Down-Regulated Proteins in Subcellular Localization Results

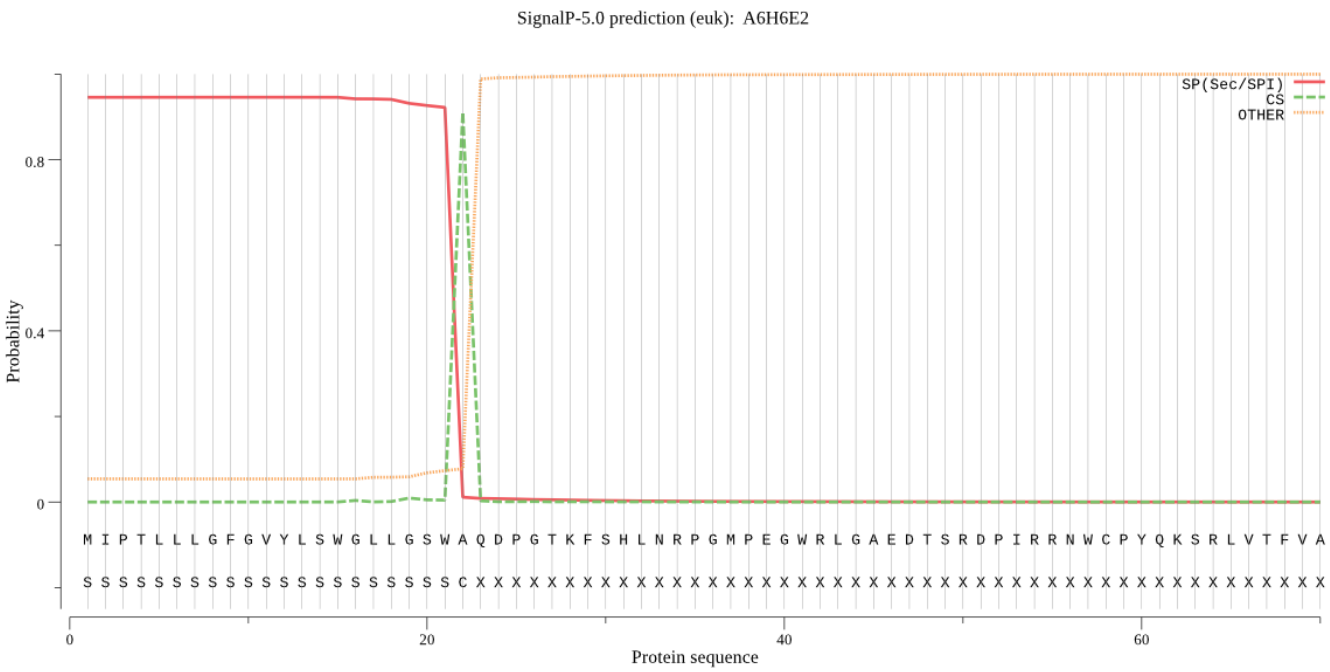
Note: The horizontal coordinate indicates the subcellular location; the vertical coordinate shows the number of differentially expressed proteins annotated with that subcellular location; red and blue colors represent up- and down-regulated differentially expressed proteins, respectively.

The results of subcellular localization analysis of differentially expressed proteins are available at: 6.Enrichment/SUBCELLULAR

6.6 Differentially Expressed Protein Signal Peptide Prediction

Signal peptides are short peptide sequences that direct newly synthesized proteins toward secretory pathways and are approximately 5-30 amino acids in length. Signal peptides are usually located at the N-terminus of proteins (in a few cases not necessarily at the N-terminus), are usually hydrophobic, and are primarily involved in directing proteins into subcellular organelles in different membrane structures of the cell. SignalP is a commonly used software for signal peptide prediction and is available online at (http://www.cbs.dtu.dk/services/SignalP/). It allows prediction of the presence and location of potential signal peptide cleavage sites in a given amino acid sequence, in both prokaryotes and eukaryotes. In this project, signal peptide prediction was performed using signalP5.0, which recognizes three types of signal peptides: sec/SPI, sec/SPII, and Tat/SPI. Signal peptide prediction was performed for differentially

expressed proteins in each differential group. The prediction results are shown below, with signal peptide localization profiles for up to three differentially expressed proteins displayed in the plot.



Signal Peptide Localization of Differentially Expressed Proteins

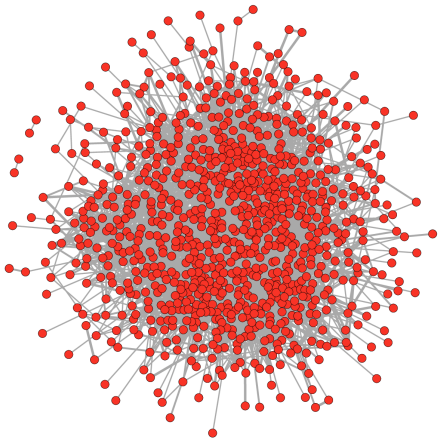
Note: The horizontal coordinate indicates the differentially expressed protein sequence; the vertical coordinate indicates the probability of signal peptide; the red line indicates the signal peptide SP; the signal peptide type is presented in parentheses; CS represents the probability of signal peptide cleavage site; and OTHER represents the probability of non-signal.

The results of subcellular localization analysis of differentially expressed proteins are available at: 6.Enrichment/SIGNALP

6.7 Protein-Protein Interaction (PPI) Network

Protein-protein interaction (PPI) analysis was performed using the StringDB (http://string-db.org/) PPI database. If the corresponding species existed in the database, the protein sequences of the corresponding species were extracted directly, if not, the protein sequences of closely allied species were extracted. The differentially expressed protein sequences were blast-aligned with the extracted sequences. The PPIs of differentially expressed proteins were obtained by taking a confidence score > 400 (medium confidence), and a static and dynamic network was plotted by using R(qgraph) and R(networkD3), re-

spectively. The static plot of the PPI network analysis results is shown below, and the dynamic network plot is shown in the results folder of differentially expressed protein PPI analysis. The corresponding network relationships can be directly imported into the Cytoscape software for visualization and editing.



PPI Network Plot

Note: Each node in the PPI network represents a differentially expressed protein; the change of node color from red to blue represents the change of expression level of the differentially expressed protein from up-regulation to down-regulation; for the number of samples in the differential group ≥ 3, all dots are in red; the thicker the line, the higher the plausibility of the interaction.

Detailed PPI analysis results for differentially expressed proteins are available at: 7.PPI

6.8 Weighted Protein Co-expression Network Analysis (WPCNA)

The WGCNA algorithm is a typical sys-biological algorithm for constructing gene co-expression networks based on high-throughput messenger RNA (mRNA) expression data, which is widely used in biomedical fields worldwide. The WGCNA algorithm first assumes that the gene network obeys a scale-free distribution, defines the gene co-expression correlation matrix, the adjacency function of the gene network, and then calculates the dissimilarity coefficients of different nodes, and constructs the hierarchical clustering tree accordingly. Different clads (branches) of this clustering tree represent different gene

modules, with a high degree of gene co-expression within the same module and a low degree of gene co-expression in different modules. Finally, the association between modules and specific phenotypes or diseases is explored for the purpose of identifying target genes and gene networks.

WPCNA is an application of the WGCNA algorithm in proteomics, with the same analysis principle as WGCNA.

6.8.1 Data Filtering

Before starting WPCNA, we need to filter the input protein expression file to remove the proteins with stable expression in all samples; in case of more stringent requirements, we can use the varFilter function of the genefilter package in R to remove proteins that are poorly expressed in all samples (low-expression proteins are not filtered by default), to improve the accuracy of the constructed network. The protein list after filtering is as follows:

Table 6.5 Quantification of Proteins after Filtration

Accession	A1	A2	A3	B1
A0A075B5R2	145.37261	117.90136	113.13043	179.37538
A0A075B5T3	114.48093	118.83708	115.86745	112.77561
A0A087WPR7	93.58362	78.76428	77.13342	94.12768
A0A087WQ89	93.58362	103.86548	112.21809	96.79167
A0A087WQF8	91.76646	113.22273	103.09466	95.01567
A0A087WQH8	88.93860	105.73693	98.53295	74.47151
A0A087WR45	85.83462	73.50747	78.30669	99.45566
A0A087WRT4	79.16620	85.37078	90.79730	84.97092
A0A087WRU0	78.69910	64.55155	67.52491	89.68769
A0A087WSP0	74.70717	74.97684	71.90607	81.71028

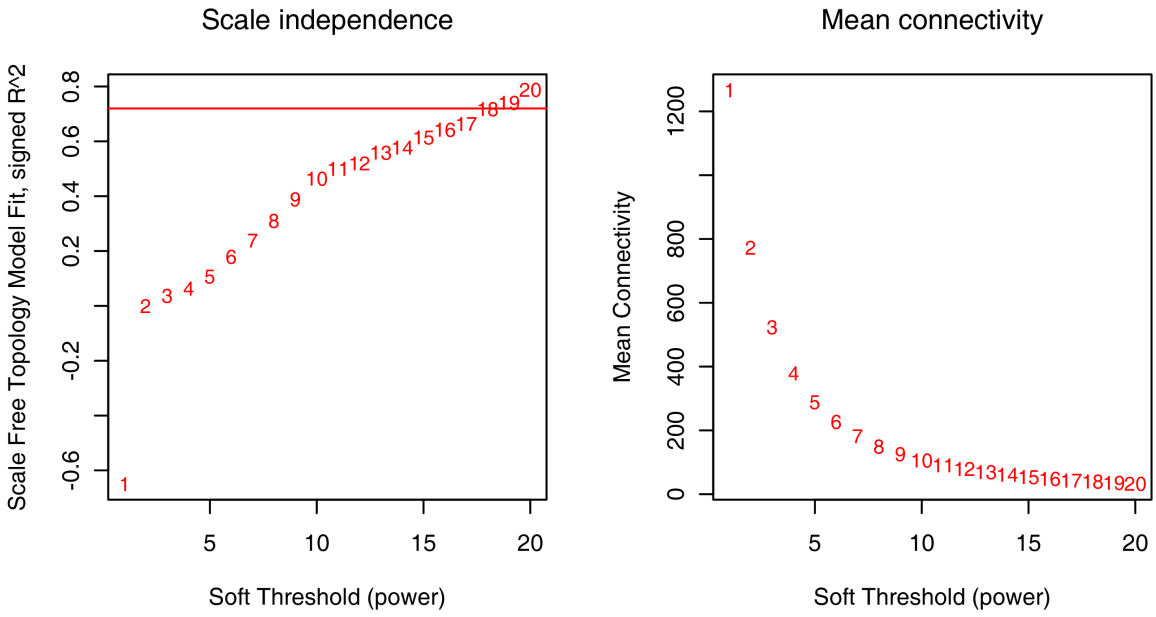
The full form is available in the web version

- Accession: Protein
- experiment: Relative quantification results for each sample

Quantification results are detailed in: 8.WPCNA/1.Soft_threshold_filtering/expressed_filter.xls

6.8.2 Soft Threshold Selection

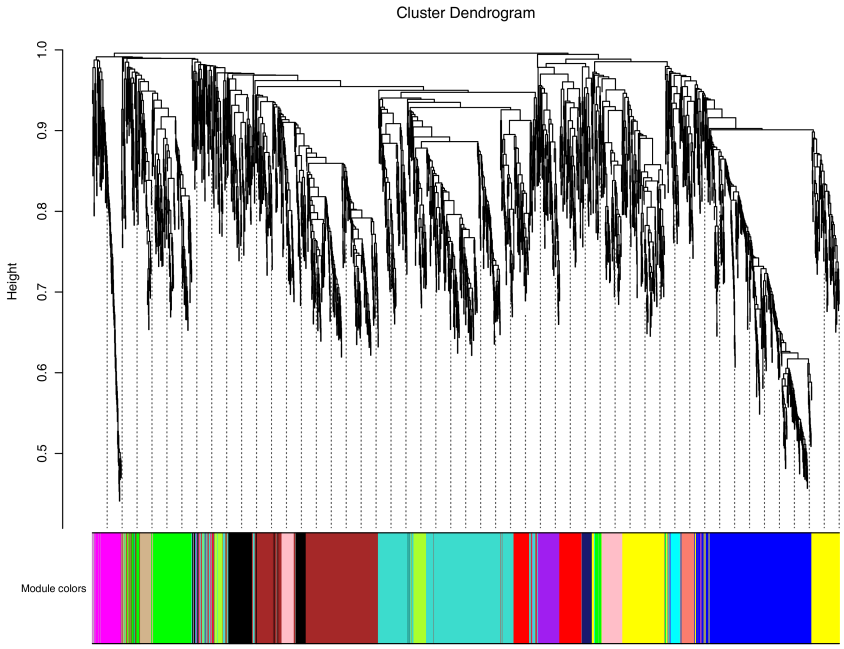
WPCNA first calculates the correlation coefficient (Pearson Correlation Coefficient) between any two proteins. In order to measure whether two proteins have similar expression patterns, it is generally required to set a threshold value for screening, with those above the threshold considered similar. However, if the threshold value is set to 0.8, it is difficult to demonstrate a significant difference between 0.8 and 0.79. Therefore, a weighted value of the correlation coefficient is applied when performing WPCNA, i.e., the Nth power is taken for the protein correlation coefficient. This approach reinforces the strong correlation and attenuates the weak or negative correlation, making the connections between proteins in the network obey scale-free network distribution, which is more biologically significant. All horizontal axes in the graphs below represent the weighting parameter β, which is the soft threshold. The vertical axis of the left panel represents the square of the correlation coefficient in the corresponding network. The higher the square of the correlation coefficient, the more the network approximates a scale-free network. We have set a threshold value of 0.8 for the square of the correlation coefficient. The vertical axis of the right panel represents the mean value of all adjacency functions of proteins in the corresponding protein module. The optimal β value is the soft threshold used for the subsequent analysis.



Schematic Diagram of Soft Threshold Selection

6.8.3 Module Hierarchical Clustering

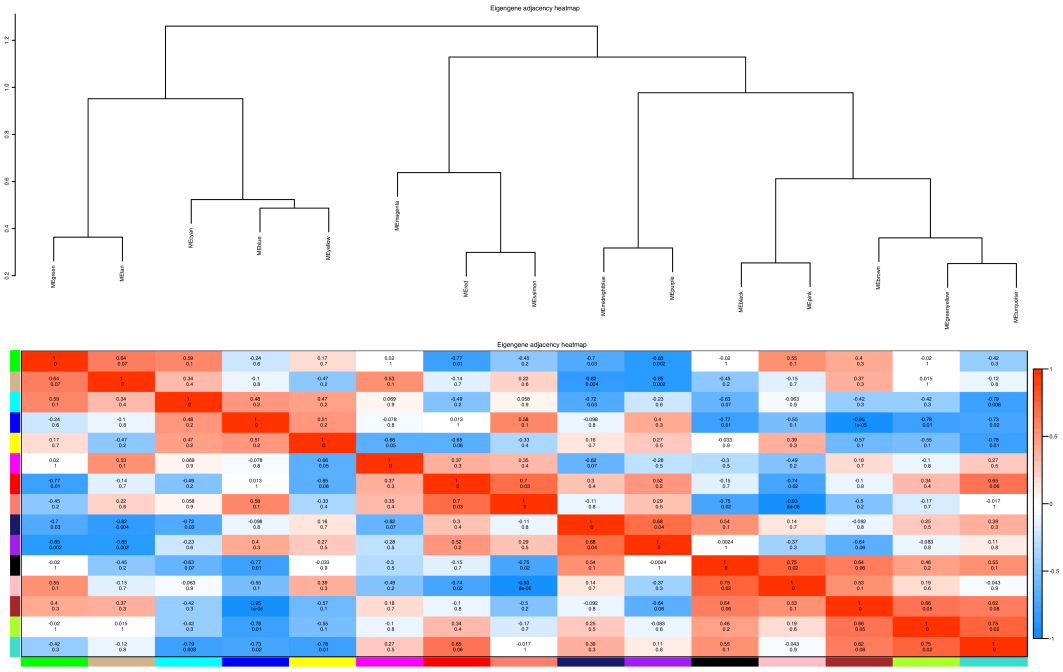
WPCNA constructs a dendrogram (cluster tree) based on the correlation of expression among proteins and divides the modules. Each color in the diagram indicates that the proteins corresponding to this color belong to the same module in the dendrogram. If some proteins always have similar expression changes in a physiological process or in different tissues, these proteins may be functionally related and can be defined as a module. For the upper half of the dendogram, the vertical distance represents the distance between two nodes (proteins) and the horizontal distance is meaningless.



Module Hierarchical Cluster Dendrogram

6.8.4 Inter-Module Correlation Heatmap

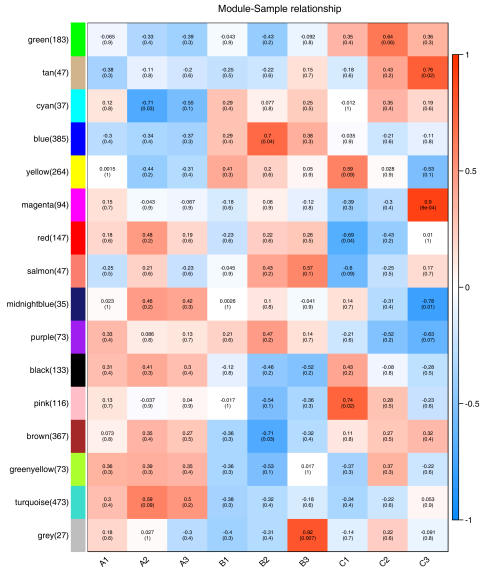
The inter-module correlation heatmap can be divided into two parts, with the upper part clustering the modules according to their characteristic values called eigengenes. The vertical coordinates represent the degree of dissimilarity of the nodes. Each row and column in the lower half of the graph represents a module. The darker the color of the square (the redder), the stronger the correlation; the lighter the color of the square, the weaker the correlation.



Inter-Module Correlation Heatmap

6.8.5 Sample-Module Correlation Heatmap

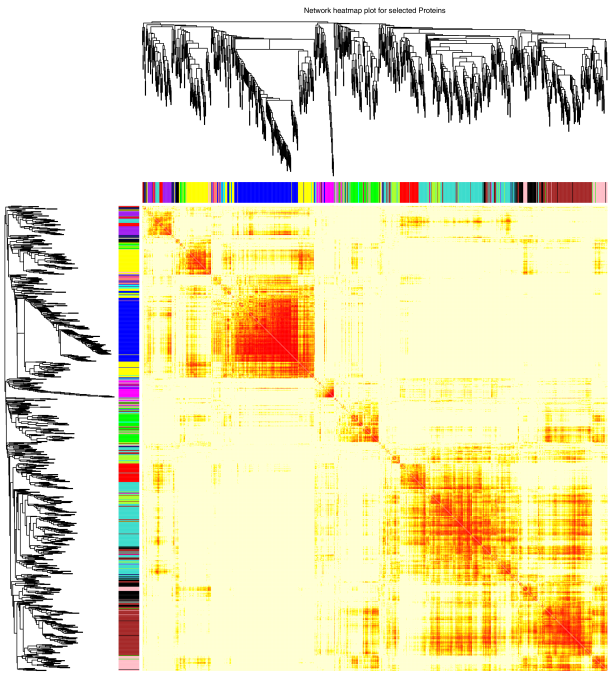
In general, if the correlation between a module and a sample is significantly higher than that of other modules, it means that this module may have the strongest correlation with that sample, as shown in the figure below, where each row and column in the plot represents a module and a sample, respectively; the darker the color of the square (the redder), the stronger the correlation; the lighter the color of the square, the weaker the correlation.



Sample-Module Correlation Heatmap

6.8.6 Module Protein Clustering Heatmap

Each clad in the dendrogram represents a protein, and the darker the color of each node (white → yellow → red) the stronger the correlation between the two proteins in the corresponding row and column. The results are shown in the figure below.



Module Protein Clustering Heatmap

6.8.7 List of Proteins by Module

Connectivity values, expression information, and 6 database annotations were added to the protein list of each module obtained from WPCNA. Connectivity values indicate the strength of correlation or association between one protein and the other proteins (usually only calculated within a module), often referred to as connectivity or degree or expressed as k value. In general, the proteins with the highest connectivity (k-value) in a module are considered hub proteins:

Table 6.6 List of Network Node Proteins by Module

Accession	moduleColors	kTotal	kWithin	kOut	kDiff
A0A075B5R2	yellow	227.34754	44.421687	182.92586	-138.50417
A0A075B5T3	pink	88.01169	35.007515	53.00418	-17.99666
A0A087WPR7	green	65.24890	19.541616	45.70729	-26.16567
A0A087WQ89	tan	23.80876	3.543307	20.26546	-16.72215
A0A087WQF8	brown	144.61989	58.894622	85.72527	-26.83065
A0A087WQH8	turquoise	167.28876	60.129280	107.15948	-47.03020
A0A087WR45	blue	281.48639	102.558286	178.92810	-76.36982
A0A087WRT4	purple	92.13222	11.940451	80.19177	-68.25132
A0A087WRU0	green	98.23610	22.327478	75.90863	-53.58115
A0A087WSP0	tan	68.46675	6.061254	62.40549	-56.34424

- Accession: ID number in the protein database
- moduleColors: The module to which it belongs
- kTotal: Total protein connectivity
- kWithin: Protein connectivity within the module
- kOut: Protein connectivity outside the module
- kDiff: The difference between kWithin and kOut

The list is available at: 8.WPCNA/4.Moderating_network_files/2.Network_nodes_for_each_module

6.8.8 Network Node Relationships by Module

Protein interaction relationships within each module in WPCNA were exported and can be subsequently imported into the Cytoscape software for network mapping:

Table 6.7 List of Network Node Relationships by Module

fromNode	toNode	weight
A0A1L1STC6	O70589	0.1129235
A0A1L1STC6	P13597	0.1953973
A0A1L1STC6	P17426	0.1037290
A0A1L1STC6	P25911	0.1204919
A0A1L1STC6	P52912	0.1002219
A0A1L1STC6	P63087	0.1093175
A0A1L1STC6	P70296	0.1084690
A0A1L1STC6	P70388	0.1200239
A0A1L1STC6	P97386	0.1189863
A0A1L1STC6	Q08122	0.1697407

- fromNode: Network node protein
- toNode: Network node protein
- weight: Edge weights of the adjacency matrix, representing the strength of the connectivity between two nodes (proteins)

The list is available at: 8.WPCNA/4.Moderating_network_files/3.Network_node_relationship_of_each_module

Reference

Meier, Florian, Andreas-David Brunner, Scarlet Koch, Heiner Koch, Markus Lubeck, Michael Krause, Niels Goedecke, et al. 2018. "Online Parallel Accumulation–Serial Fragmentation (Pasef) with a Novel Trapped Ion Mobility Mass Spectrometer." *Molecular & Cellular Proteomics* 17 (12): 2534–45.

Prianichnikov, Nikita, Heiner Koch, Scarlet Koch, Markus Lubeck, Raphael Heilig, Sven Brehmer, Roman Fischer, and Jürgen Cox. 2020. "MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics." *Molecular & Cellular Proteomics* 19 (6): 1058–69.

Ross, Philip L, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, et al. 2004. "Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents." *Molecular & Cellular Proteomics* 3 (12): 1154–69.

Wu, Jing, Yuan An, Hai Pu, Yue Shan, Xiaoqing Ren, Mingrui An, Qingsong Wang, Shicheng Wei, and Jianguo Ji. 2010. "Enrichment of Serum Low-Molecular-Weight Proteins Using C18 Absorbent

Under Urea/Dithiothreitol Denatured Environment." *Analytical Biochemistry* 398 (1): 34–44.

Wu, Jing, Xiaolei Xie, Yashu Liu, Jintang He, Ricardo Benitez, Ronald J Buckanovich, and David M Lubman. 2012. "Identification and Confirmation of Differentially Expressed Fucosylated Glycoproteins in the Serum of Ovarian Cancer Patients Using a Lectin Array and Lc–Ms/Ms." *Journal of Proteome Research* 11 (9): 4541–52.

Yu, Fengchao, Sarah E Haynes, Guo Ci Teo, Dmitry M Avtonomov, Daniel A Polasky, and Alexey I Nesvizhskii. 2020. "Fast Quantitative Analysis of timsTOF Pasef Data with Msfragger and Ionquant." *Molecular & Cellular Proteomics* 19 (9): 1575–85.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters." *OMICS : A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.