METWARE BIO
Innovative Metabolomics Insights for Better Health

Conjoint analysis report of proteome and transcriptome

Metware Biotechnology Inc.

www.metwarebio.com

Contents

MWY-24-0827-Demo_Conjoint analysis report of proteome and transcriptome

1 Abstract

Proteomics: Study the expression and functional patterns of proteins in specific samples during specific periods, and reveal the laws of protein functions and cellular life activities by qualitative and quantitative detection of proteins, as well as interaction studies.

Transcriptomics: A sequencing technique that investigates the transcribed mRNAs of a specific sample at a specific time, focusing on mRNAs with translation functions, and non-coding RNAs that have regulatory effects on these mRNAs.

Biological processes are complex and holistic. Single-omics data provides limited view on the macro-developmental processes in biological systems, making it difficult to explain complex biological processes and the regulation of biological networks. Integrating multi-omics data for analysis makes up for missing data, noise and other issues presented by single-omics data analysis. Multi-omics provides mutual validation and reduces false positives brought by single-omics analysis. Most importantly, joint analysis of multi-omics data is more conducive to understanding the mechanisms and the phenotype of biological processes in the system.

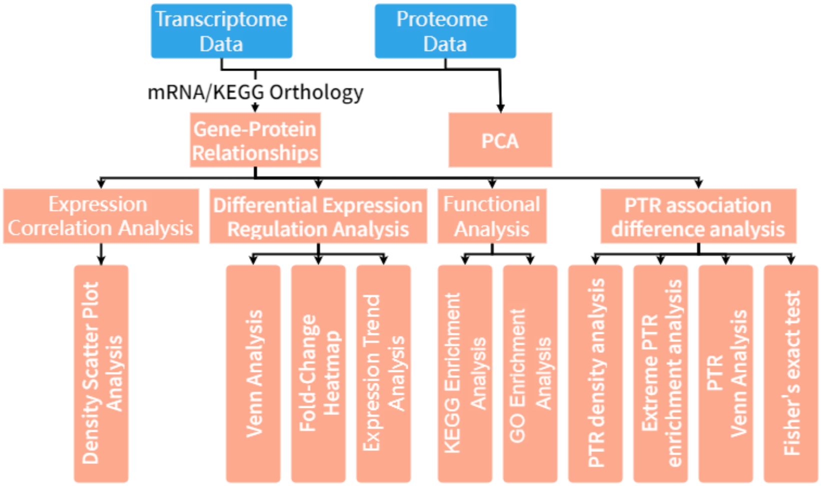
Transcriptome Data
Proteome Data
mRNA/KEGG Orthology
Gene-Protein Relationships
PCA
Expression Correlation Analysis
Differential Expression Regulation Analysis
Functional Analysis
PTR association difference analysis
Density Scatter Plot Analysis
Venn Analysis
Fold-Change Heatmap
Expression Trend Analysis
KEGG Enrichment Analysis
GO Enrichment Analysis
PTR density analysis
Extreme PTR enrichment analysis
PTR Venn Analysis
Fisher's exact test

Fig 1 Conjoint Analysis Flow Chart

2 Sample and Grouping Information

The correspondence between sample names and group names in this multi-omics analysis is shown in the table below. The Sample column shows the unified sample names; the Group column shows the unified sample group names; and the Prot and Trans represent proteomics and transcriptomics, respectively.

Table 1 Correspondence Between Omics Sample and Group Names

Sample	Group	Prot_sample	Prot_group	Trans_sample	Trans_group
A-1	A	A-1	A	A-1	A
A-2	A	A-2	A	A-2	A
A-3	A	A-3	A	A-3	A
B-1	B	B-1	B	B-1	B
B-2	B	B-2	B	B-2	B
B-3	B	B-3	B	B-3	B
C-1	C	C-1	C	C-1	C
C-2	C	C-2	C	C-2	C
C-3	C	C-3	C	C-3	C

3 Proteome and transcriptome combination

3.1 Principal Component Analysis (PCA)

PCA was performed on the proteome data and transcriptome data respectively, to visually show whether there are differences between groups in the two omics, as shown in the figures below:

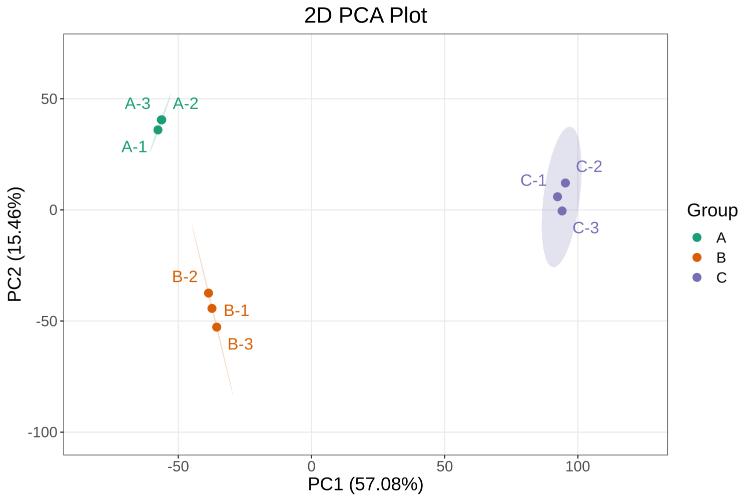

Fig 2 Proteome PCA Plot

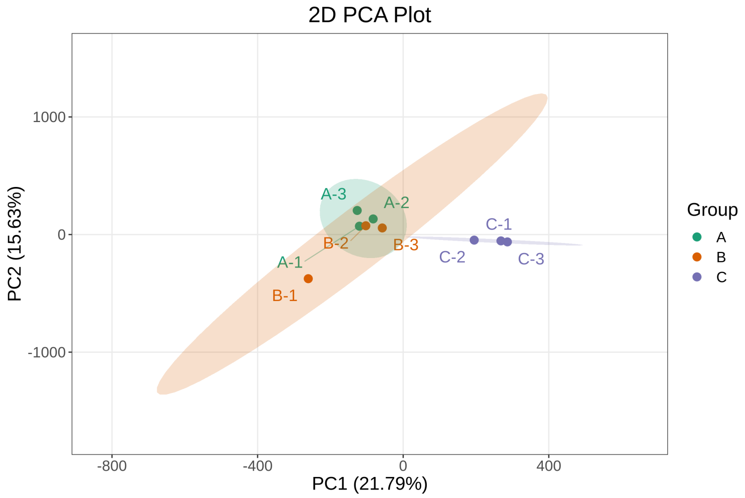

Fig 3 Transcriptome PCA Plot

Note: The horizontal axis represents principal component 1, the vertical axis represents principal component 2, and the differently colored dots represent samples from different groups

3.2 Expression Correlation Analysis

Transcriptome and proteome describe the expression of genes at the transcriptional and translational levels, respectively. By comparing the correlation between the gene and protein expression profiles obtained from the two omics, we can quickly understand the potential regulatory correlation between genes and proteins. The figures below are heat scatter plots showing the expression levels of genes identified in common across sample groupings at the transcriptional level (genes) and translational level (proteins). It visually reflects the density distribution and correlation results of protein and gene expression levels.

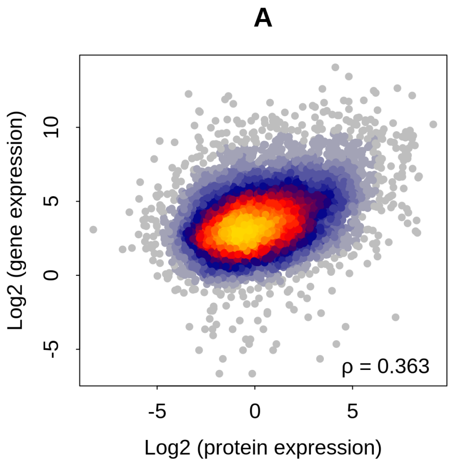

Fig 4 Heat Scatter Plot of Protein and Gene Expression

Note: Each plot represents a sample group, each point represents a protein-gene pair. The horizontal and vertical axes represent the logarithmic values of protein and its corresponding gene expression levels, respectively; ρ represents the Pearson's correlation coefficient of the two sets of data, the color gradient from yellow to red to blue to gray indicates the change in point density from high to low

The heat scatter plots of expression are shown in: [2.Conjoint_prot_trans/2.Expression_correlation]

3.3 Differential Expression Analysis

3.3.1 Venn Analysis of Differential Expression

The mRNAs obtained from the transcriptome data are aligned with the proteins identified from the proteome data to identify mRNA-protein correlations. A Venn analysis is then performed to obtain the number of common and exclusive proteins (or genes) in each set corresponding to different regions. The results of Venn analysis for each comparison pair are shown below:

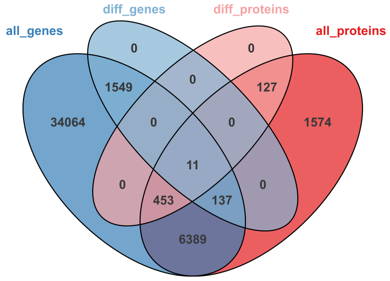

Fig 5 Venn Diagram of Differentially Expressed Proteins and Genes

Note: all_genes represents all genes identified in the transcriptome; diff_genes represents differentially expressed genes identified in the transcriptome; all_proteins represents all proteins identified in the proteome; diff_proteins represents differentially expressed proteins identified in the proteome

3.3.2 Fold Change Cluster Analysis

Cluster analysis of the fold change of the common proteins (or genes) identified in each comparison pair was performed and plotted as a heatmap, visually reflecting whether the up- or down-regulation of proteins and their corresponding genes are consistent. Clustering heatmaps for all comparison pairs are shown below:

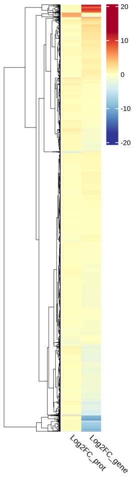

Fig 6 Clustering Heatmap of the Fold Change of Proteins and Genes

Note: Each row represents a protein-gene pair; the two columns are the fold changes obtained from the proteome and transcriptome data, respectively. The red color indicates up-regulated proteins or genes; the blue color indicates down-regulated proteins or genes

3.3.3 Expression Trend Analysis

In each comparison, the nine-quadrant chart was utilized to visualize the fold change of the proteins and genes. The graph is separated into 1-9 quadrants from left to right and from top to bottom by black dotted lines, with different quadrants corresponding to different up or down-regulated expression states.

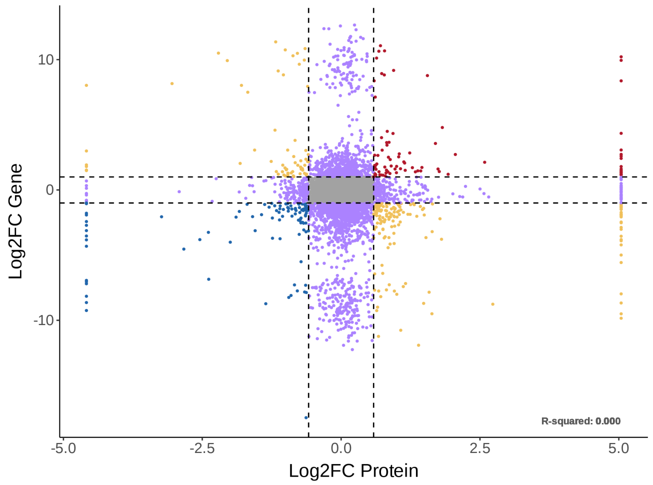

Fig 7 Nine-quadrant Chart of Differentially Expressed Proteins and Genes

Note: Each dot represents a protein-gene pair; the horizontal coordinate represents the Log2FC of the protein expression level, and the vertical coordinate represents the Log2FC of the gene level

The following table shows the interpretation of the nine-quadrant chart:

Table 2 Interpretation of the Nine Quadrants

Quadrant	Interpretation	Information to be mined
5	Neither genes nor proteins are differentially expressed	These genes and proteins in this differential grouping are non-differentially expressed
3,7	These genes and proteins have consistent patterns of differential expression	These genes and proteins are positively correlated, indicating changes in metabolite expression may be positively regulated by genes
1,9	These genes and proteins have opposite patterns of differential expression	These genes and proteins are with non-consistent regulatory trends, indicating changes in metabolite expression may be negatively regulated by these genes
2,4,6,8	**********	Protein expression unchanged, gene up- or down-regulated or gene expression unchanged, protein up- or down-regulated

The results of differential expression analysis are shown in: [2.Conjoint_prot_trans/3.Difference_analysis]

3.4 GO Enrichment Analysis

3.4.1 GO Enrichment Column Chart

Based on the results of GO enrichment analysis of differentially expressed genes and proteins in each comparison pair in the transcriptome and proteome data sets, the differences in the enrichment of GO terms common to the two omics data sets were identified and sorted according to the P-value of enrichment significance (in ascending order). Only up to 10 GO terms were displayed for each level of classification, and their differences are shown below.

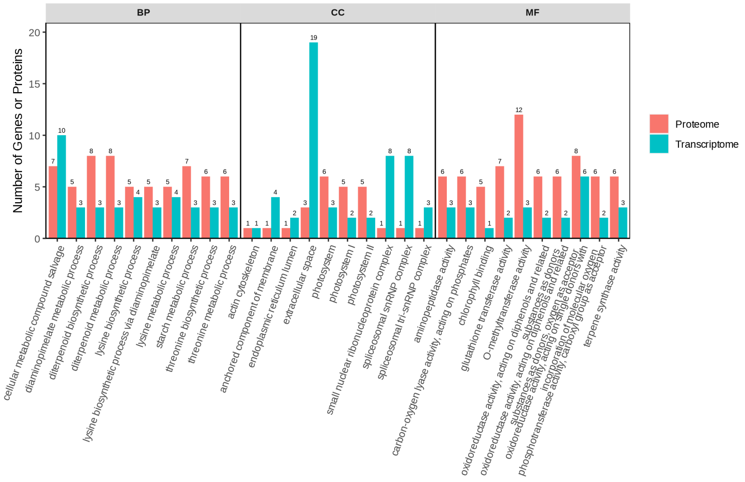

Fig 8 GO Enrichment Analysis Column Chart

Note: The horizontal coordinate represents the number of secondary GO terms; the vertical coordinate represents the number of proteins or genes enriched in the GO term; BP stands for biological process; CC stands for cellular component; and MF is molecular function

3.4.2 GO Enrichment Clustering Heatmap

The GO enrichment results of differentially expressed proteins (genes) obtained from the proteome and transcriptome data were plotted into heatmaps, with clustering based on the fold change (as shown below):

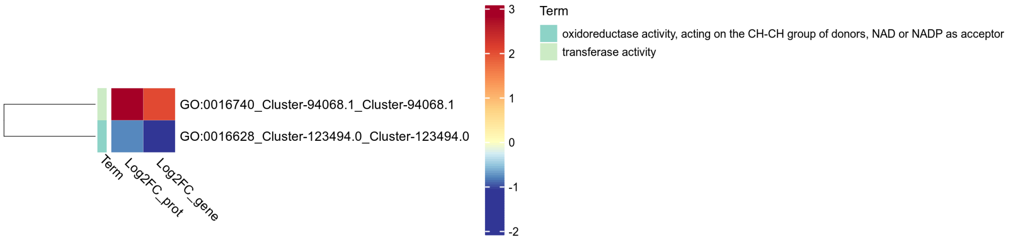

Fig 9 Clustering Heatmap of GO Enrichment Analysis

Note: The red color indicates an up-regulation, the blue color indicates a down-regulation; proteins (or genes) that are clustered together have similar expression patterns. The protein level was used as the basis for the difference in the terms

The results of GO Enrichment analysis are shown in: [2.Conjoint_prot_trans/4.GO]

3.5 KEGG Enrichment Analysis

Based on the results of KEGG enrichment analysis of differentially expressed proteins and differentially expressed genes, we identify KEGG pathways that are co-enriched in both datasets. KEGG enrichment analysis results are shown in the table below:

Table 3 KEGG Enrichment Analysis Results

KEGG_map	Description	Rich_factor_prot	P-value_prot	Rich_factor_gene	P-value_gene
ko00999	Biosynthesis of various plant secondary metabolites	43/66	0.0000	95/338	0e+00
ko04075	Plant hormone signal transduction	57/131	0.0823	274/1186	0e+00
ko00500	Starch and sucrose metabolism	88/163	0.0000	203/981	0e+00
ko00460	Cyanoamino acid metabolism	30/52	0.0021	90/355	0e+00
ko00950	Isoquinoline alkaloid biosynthesis	38/59	0.0000	65/236	0e+00
ko04626	Plant-pathogen interaction	99/254	0.3090	416/2432	0e+00
ko04016	MAPK signaling pathway - plant	45/115	0.3762	156/810	0e+00
ko04712	Circadian rhythm - plant	11/29	0.5430	56/236	0e+00
ko00073	Cutin, suberine and wax biosynthesis	3/10	0.7859	22/64	0e+00
ko00591	Linoleic acid metabolism	7/11	0.0698	19/55	1e-04

- KEGG_map: Reference pathway highlighting KOs in KEGG database

- Description: KO pathway description
- Rich_factor_prot: proteome enrichment factor, which is the ratio of the number of differential proteins annotated to the KEGG pathway to the number of background proteins
- P-value_prot: P-value calculated by hypergeometric tests in proteome
- Rich_factor_gene: transcriptome enrichment factor, which is the ratio of the number of differential genes annotated to the KEGG pathway to the number of background genes
- P-value_gene: P-value calculated by hypergeometric tests in transcriptome

The results of common KEGG pathway analysis are shown in:

[2.Conjoint_prot_trans/5.KEGG/A_vs_B/A_vs_B_common_KEGG.xls]

3.5.1 KEGG Enrichment Bar Chart

Bar charts were plotted using KEGG pathways that were co-enriched in both omics, displaying the number of differentially expressed proteins and differentially expressed genes enriched in a specific pathway. For the number of co-enriched KEGG pathways exceeding 25, the transcriptome data was taken as the standard, and only the top 25 pathways ranked by P-value are displayed, as shown in the figure below:

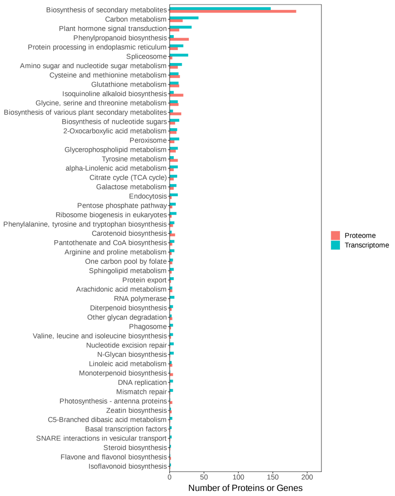

Fig 10 KEGG Enrichment Analysis Bar Chart

Note: The horizontal axis represents the number of differentially expressed proteins and differentially expressed genes enriched in that pathway; the vertical axis represents the names of the KEGG pathways. The red and green bars represent the proteome and transcriptome, respectively

3.5.2 KEGG Enrichment Bubble Plot

Bubble charts were plotted using KEGG pathways that were co-enriched in both omics. Each bubble chart represents a five-dimensional view, illustrating the status of co-enriched KEGG pathways by different omics through the horizontal and vertical axes, as well as bubble color gradients, shapes, and sizes. For the number of co-enriched KEGG pathways exceeding 25, the transcriptome data was taken as the standard, and only the top 25 pathways ranked by P-value are displayed, as shown in the figure below:

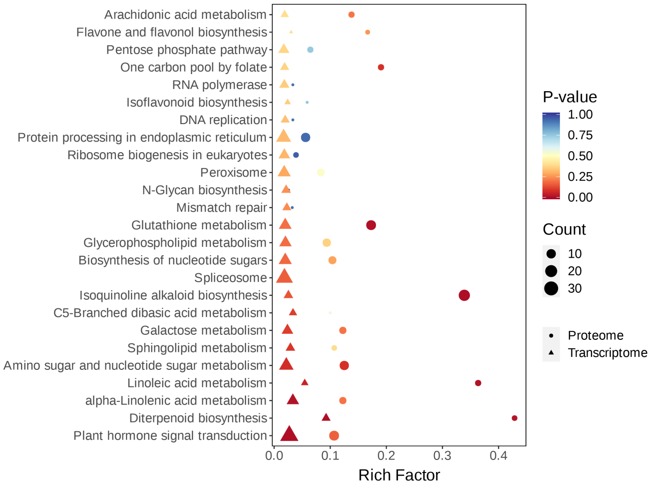

Fig 11 KEGG Enrichment Analysis Bubble Plot

Note: The horizontal axis represents the enrichment factor (Diff/Background) of the pathway in different omics, while the vertical axis represents the names of the KEGG pathways. The gradient from red to yellow to blue represents the significance of enrichment, ranging from high to medium to low, indicated by the P-value. The shape of the bubbles represents different omics, and the size of the bubbles corresponds to the number of differentially expressed proteins or differentially expressed genes, with larger numbers resulting in larger bubbles

3.5.3 KEGG Enrichment Clustering Heatmap

The KEGG enrichment results of differentially expressed proteins (genes) obtained from the proteome and transcriptome data were plotted into heatmaps, with clustering based on the fold change (as shown below):

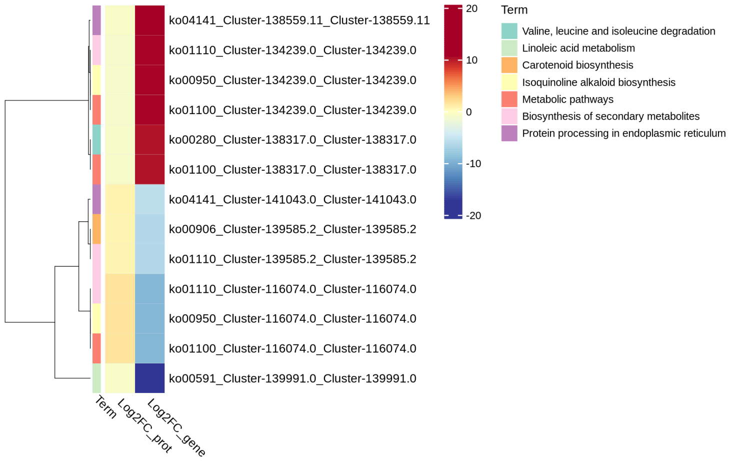

Fig 12 Clustering Heatmap of KEGG Enrichment Analysis

Note: The red color indicates an up-regulation, the blue color indicates a down-regulation; proteins (genes) that are clustered together have similar expression patterns. The protein level was used as the basis for the difference in the terms

The results of KEGG Enrichment analysis are shown in: [2.Conjoint_prot_trans/5.KEGG]

3.6 PTR Analysis

The protein-to-mRNA ratio (PTR) is an indicator of gene translation efficiency, calculated by dividing the expression level of the protein (represented by log2 LFQ value) by the transcription level of its mRNA (represented by log2 FPKM value). A high PTR value usually indicates an efficient translation process and a high protein synthesis rate, while a low PTR value may suggest that the translation process is regulated or that protein synthesis is less efficient. Analyzing PTR values helps reveal the cellular regulatory mechanisms that influence protein synthesis rates.

3.6.1 PTR Density Distribution

The PTR values of the genes in different comparison groupings were calculated separately. If there were duplicate samples, the average of the expression levels of these samples was taken before calculating the PTR. Genes at both ends were screened as extreme PTR genes using the median ± 1 standard deviation as the threshold. These genes exhibit significantly abnormal values in translation efficiency, either unusually high or unusually low. Using this method, we can identify genes that may be involved in specific biological processes or influenced by specific regulatory mechanisms, providing valuable information for further biological research and functional analysis. The PTR density distribution of each group is plotted below:

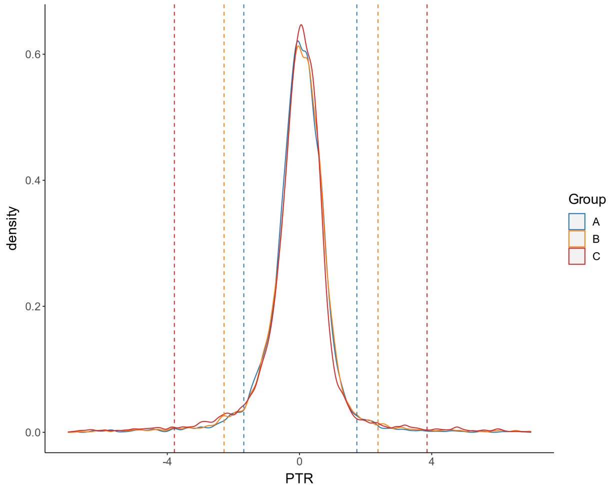

Fig 13 PTR Density Distribution

Note: The horizontal coordinates indicate PTR values; vertical coordinates indicate PTR densities; different colors indicate different sample groups; and the areas outside the dashed lines at the two ends of each curve indicate the extreme PTR genes

3.6.2 Functional Enrichment Analysis of Extreme PTR Genes

By performing GO and KEGG pathway enrichment analysis on extreme PTR genes, we can study the distribution of these genes across different functional categories to elucidate their functional significance. Scatter plots of GO and KEGG enrichment were created for each group. If the number of differential groups exceeds 10, only 10 differential groups are shown; if fewer than 10, all are displayed. The GO and KEGG enrichment results for each group were sorted by Q-value, and the union of the top 15 pathways with the smallest Q-values from each group is presented.

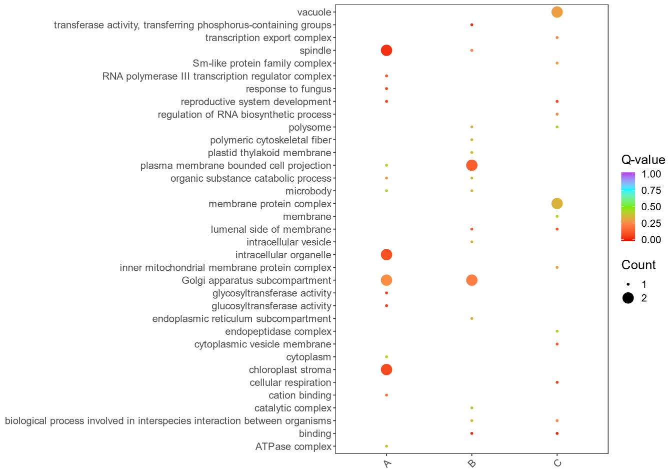

Fig 14 Bubble Plot of GO Enrichment Analysis for Extreme Genes

Note: The horizontal coordinates represent sample groupings. The vertical coordinates represent different GO terms. The color of the dots reflects the Q-value, with redder colors indicating more significant enrichment. The size of each dot represents the number of enriched differentially expressed genes

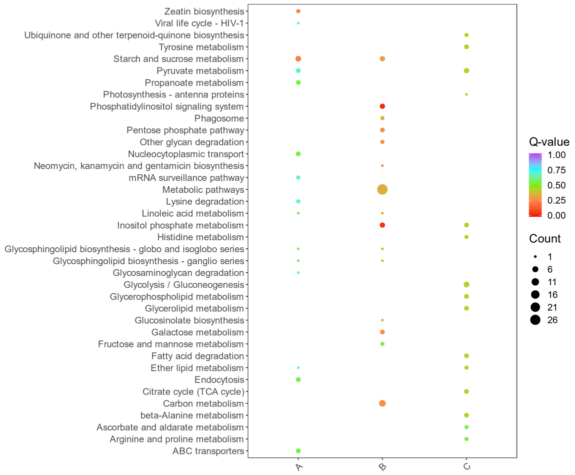

Fig 15 Bubble Plot of KEGG Enrichment Analysis for Extreme Genes

Note: The horizontal coordinates represent sample groupings. The vertical coordinates represent different pathways. The color of the dots reflects the Q-value, with redder colors indicating more significant enrichment. The size of each dot represents the number of enriched differentially expressed genes

The results of PTR analysis are shown in: [2.Conjoint_prot_trans/6.PTR]

3.7 Differential Proteins (Genes) and Extreme PTRs

The intersection of differentially expressed proteins (genes) identified in each comparison group with extreme PTR genes was taken. Analyzing differentially expressed proteins (genes) in conjunction with PTRs can provide a more comprehensive understanding of gene expression regulation, aiding in a deeper understanding of the regulatory mechanisms of gene expression in cells under specific physiological or pathological conditions.

3.7.1 Venn Analysis

The information of extreme PTR genes was matched with differentially expressed genes obtained from the transcriptome and differentially expressed proteins identified from the proteome. A Venn analysis is then performed to obtain the number of common and exclusive proteins (or genes) in each set corresponding to different regions. The results of Venn analysis for each comparison pair are shown below:

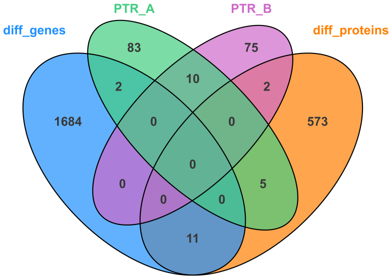

Fig 16

Venn Diagram of Differentially Expressed Proteins (Genes) and Extreme PTRs Note: diff_genes denotes differentially expressed genes identified in the transcriptome, diff_proteins denotes differentially expressed proteins identified in the proteome, and PTR_* represents extreme PTR genes in different sample groups

3.7.2 Fisher's Exact Test

Fisher's exact test is a method suitable for analyzing the statistical significance of contingency tables, which is used to test whether there is a significant association between two categorical variables. Comparisons were made between differentially expressed proteins (genes) and extreme PTR genes to construct a two-dimensional contingency table. This table includes the following data: the number of differentially expressed proteins (genes) that are also with extreme PTRs, the number of differentially expressed proteins (genes) that don't have not extreme PTRs, the number of proteins (genes) that are with extreme PTRs but not differentially expressed, and the number of proteins (genes) that are neither differentially expressed nor with extreme PTRs. Fisher's exact test was then performed on the two-dimensional contingency table data for each comparison group. Finally, a percentage stacked bar chart was plotted, with the enrichment test results added.

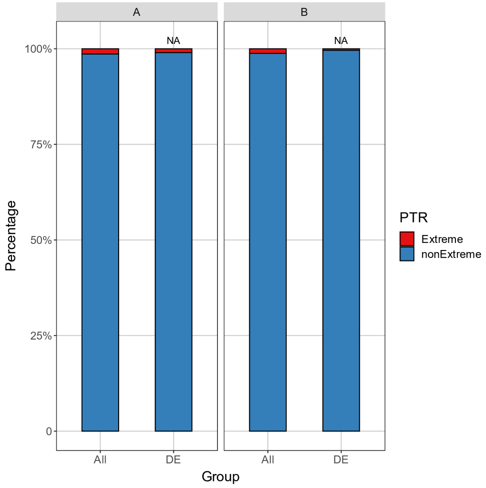

Fig 17 Stacked Bar Chart with
Enrichment Information for Differentially Expressed Proteins and Extreme PTRs

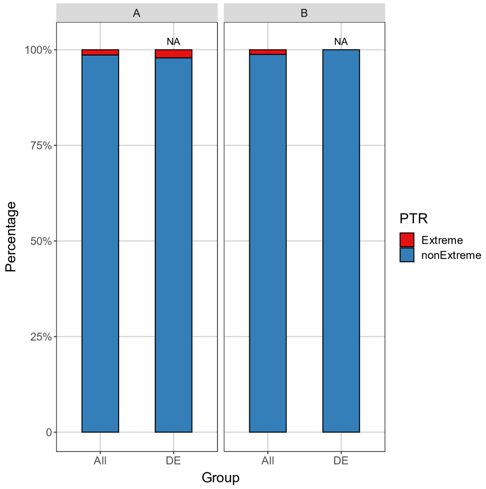

Fig 18 Stacked Bar Chart

with Enrichment Information for Differentially Expressed Genes and Extreme PTRs Note: The coordinates 'All' and 'DE' on the horizontal axis represent PTR genes and differentially expressed proteins or genes, respectively. The vertical axis represents the percentage of genes or proteins. Different facets of the chart represent different sample groups. The labels above the DE bars indicate significance: 'NA' represents not significant; '*' indicates significant; '**' indicates more stringent significance; and '***' indicates highly significant

The results of conjoint analysis between differential proteins (genes) and extreme PTRs: [2.Conjoint_prot_trans/7.Difference_PTR]

4 Appendix

For reports with network plots, Cytoscape can be used to plot networks. You can find the correlation table file *.cor.*.xlsx and the node type file nodetype.xlsx in the correlation network plot results directory (Cor-network), and then convert them to text format. A network plot can be generated by importing these two files into the Cytoscape software. A tutorial on plotting networks with Cytoscape can be found in the root directory of the results folder.

The ko01100, ko01110, ko01120 and ko01130 metabolic pathways in the KEGG analysis could not be color filled due to their high density. Therefore, instead of showing these four plots in the results, we only present these metabolic pathways in the list.

5 References

Gonzalez I, Dejean S, Martin P, Baccini A. CCA: An R Package to Extend CanonicalCorrelation Analysis. Journal of Statistical Software. 2008, 23 (12): 1-14.

Szymon Jozefczuk, Sebastian Klie, Gareth Catchpole, et al. Metabolomic and transcriptomic stress response of escherichia coli. Molecular Systems Biology, 2010, 6(1).

Bouhaddani S, Houwing-Duistermaat , Salo P, Perola M, Jongbloed G, Uh HW. Evaluation of O2PLS in Omics data integration. BMC Bioinformatics. 2016;17(Suppl 2).

Rabinowitz J S, Robitaille A M, Wang Y, et al. Transcriptomic, proteomic, and metabolomic landscape of positional memory in the caudal fin of zebrafish[J]. Proceedings of the National Academy of Sciences, 2017, 114(5): E717-E726.

Li Y, Chen Y, Zhou L, et al. MicroTom Metabolic Network: Rewiring Tomato Metabolic Regulatory Network throughout the Growth Cycle. Mol Plant. 2020;13(8):1203-1218.

Yang C, Shen S, Zhou S, et al. Rice metabolic regulatory network spanning the entire life cycle. Mol Plant. 2022;15(2):258-275.

Mergner J, Frejno M, List M, et al. Mass-spectrometry-based draft of the Arabidopsis proteome[J]. Nature, 2020, 579(7799): 409-414