Creative Proteomics
A Division of Creative Dynamics, Inc.

45-1 Ramsey Road, Shirley, NY 11967, USA
Tel:1-631-275-3058 Fax:1-631-614-7828
www.creative-proteomics.com
info@creative-proteomics.com

Bioinformatics Analysis for Quantitative Proteomics

(Project Report)

Project: XXX

Report Date: XXX

CONTACT INFORMATION:

Jessie Lively, PhD

Phone: 1-631-275-3058

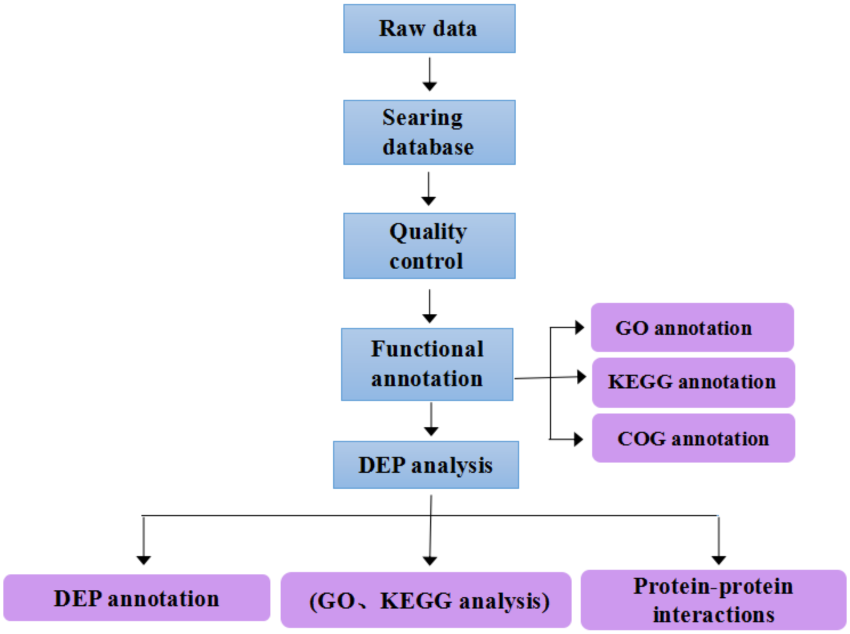
Email: team@creative-proteomics.com

Summary

In this project, 1984 proteins were identified. The fold-change cutoff was set when proteins with quantitative ratios above 1.2 or below 1/1.2 are deemed significant. 134 proteins are down-regulated and 28 proteins are up-regulated when compared to the control sample. Intensive bioinformatic analyses were then carried out to annotate those quantifiable proteins, including COG annotation, GO annotation, KEGG pathway annotation, and cluster analysis, etc. Based on the results, further studies following the quantitative analysis were suggested. For detailed information, please read the following report and the attached supplementary data.

1. Analysis workflow

The Bioinformatics Analysis work flow is below：



Note: DEP, differentially expressed protein.

2. Bioinformatics Analysis

2.1. Quantitative Overview

In total, 1984 proteins were identified for this project (Table 3). Proteins of relative quantitation were divided into two categories. Quantitative ratio over 1.2 was considered up-regulation while quantitative ratio less than 1/1.2 was considered as down-regulation. The amount of differentially expressed proteins was summarized in Table 4.

Table 3: Summary of identified proteins

Name	Identified Proteins
Total	1984

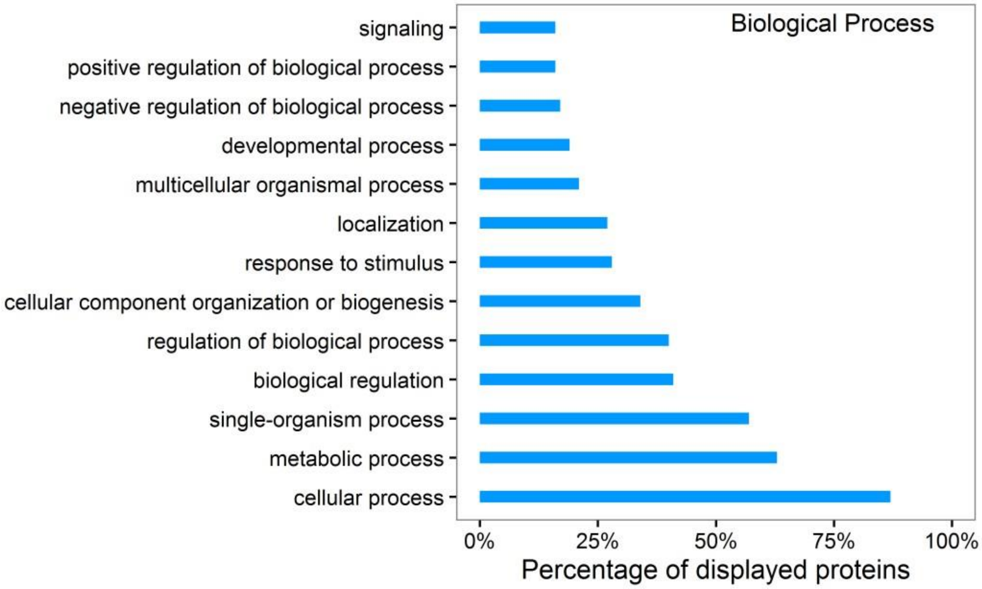
Table 4: Summary of differentially expressed proteins

Group name	Up-regulated (>1.2)	Down-regulated (<1/1.2)
UTC_vs_XL315	28	134

2.2. Annotation for identified proteins

To understand the function and feature of identified proteins, we annotated function or feature of proteins from several categories, including COG, Gene Ontology and KEGG.

2.2.1 GO annotation



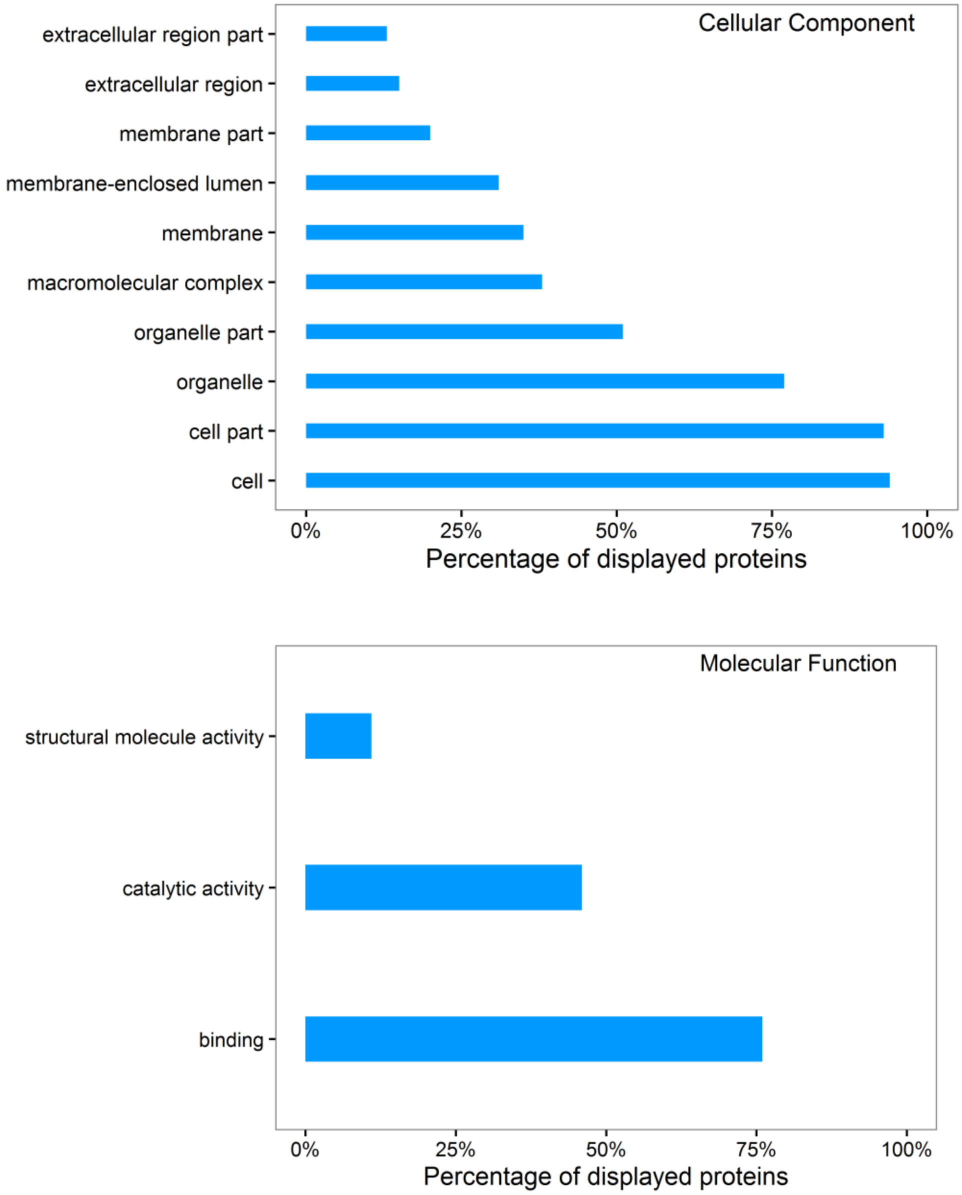


Figure 3 Distribution of identified proteins in GO level2.

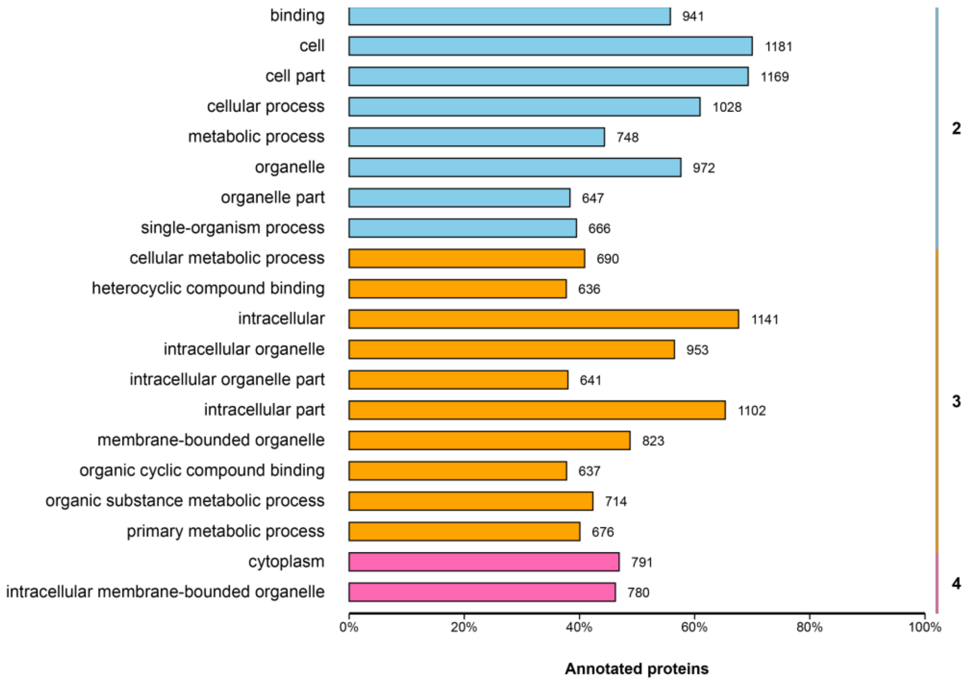


Figure 4 Distribution of identified proteins annotated in different GO levels. Different colors indicate different GO level. For example, light blue bars indicate GO terms in level 2.

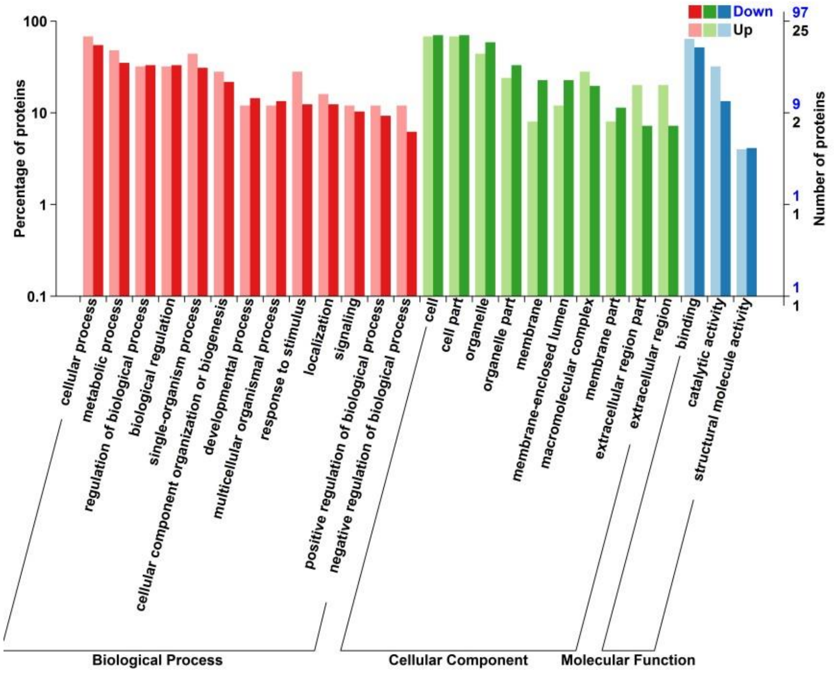


Figure 5 Distribution of differentially expressed proteins annotated in level 2.

2.2.1 KEGG annotation

Table 5 Summary of KEGG annotation

Query	Gene ID	Hyperlink
A0A024QYX3	hsa:5935	http://www.genome.jp/dbget-bin/www_bget?hsa:5935
A0A024QZ42	hsa:10016	http://www.genome.jp/dbget-bin/www_bget?hsa:10016
A0A024QZ77	hsa:79180	http://www.genome.jp/dbget-bin/www_bget?hsa:79180
A0A024QZC0	hsa:11273	http://www.genome.jp/dbget-bin/www_bget?hsa:11273
A0A024QZC1	hsa:10421	http://www.genome.jp/dbget-bin/www_bget?hsa:10421
A0A024QZE7	hsa:7041	http://www.genome.jp/dbget-bin/www_bget?hsa:7041
A0A024QZF1	hsa:23636	http://www.genome.jp/dbget-bin/www_bget?hsa:23636
A0A024QZJ8	hsa:983	http://www.genome.jp/dbget-bin/www_bget?hsa:983
A0A024QZN2	hsa:23234	http://www.genome.jp/dbget-bin/www_bget?hsa:23234
A0A024QZN4	hsa:7414	http://www.genome.jp/dbget-bin/www_bget?hsa:7414

Table 6 Summary of KEGG pathway annotation

<table>
<tr><th>Pathway_name</th><th>Pathway_id</th><th>Proteins_num</th><th>Protein_ids</th><th>Web link</th></tr>
<tr><td>2-Oxocarboxylic acid metabolism</td><td>hsa01210</td><td>6</td><td>B4DJB4|B4DJV2|B4DZ08|E9PF84|H0YL11|Q0QER2</td><td>http://www.genome.jp/kegg-bin/show_pathway?hsa01210/hsa:1431%09red</td></tr>
<tr><td>ABC transporters</td><td>hsa02010</td><td>1</td><td>B4DZ22</td><td>http://www.genome.jp/kegg-bin/show_pathway?hsa02010</td></tr>
<tr><td>AGE-RAGE signaling pathway in diabetic complications</td><td>hsa04933</td><td>7</td><td>A0A024RAE4|B4DHN0|D3DTX7|E7ENM1|F8VYY1|H9KV28|Q01970</td><td>http://www.genome.jp/kegg-bin/show_pathway?hsa04933/hsa:5594%09red</td></tr>
<tr><td>AMPK signaling pathway</td><td>hsa04152</td><td>13</td><td>A0A024R0Y2|A0A024R7V6|A0A024R845|A0A087X0K1|A0A0S2Z4A1|B4DQY1|K7EMT8|P13639|P49327|P61026|Q01813|Q15717|Q15907</td><td>http://www.genome.jp/kegg-bin/show_pathway?hsa04152/hsa:51552%09red/hsa:6720%09red/hsa:1938%09red</td></tr>
<tr><td>Acute myeloid leukemia</td><td>hsa05221</td><td>6</td><td>A0A087WV30|B3KR50|B4DFY5|B4DHN0|P36507|Q05835</td><td>http://www.genome.jp/kegg-bin/show_pathway?hsa05221/hsa:5604%09red/hsa:5594%09red</td></tr>
</table>

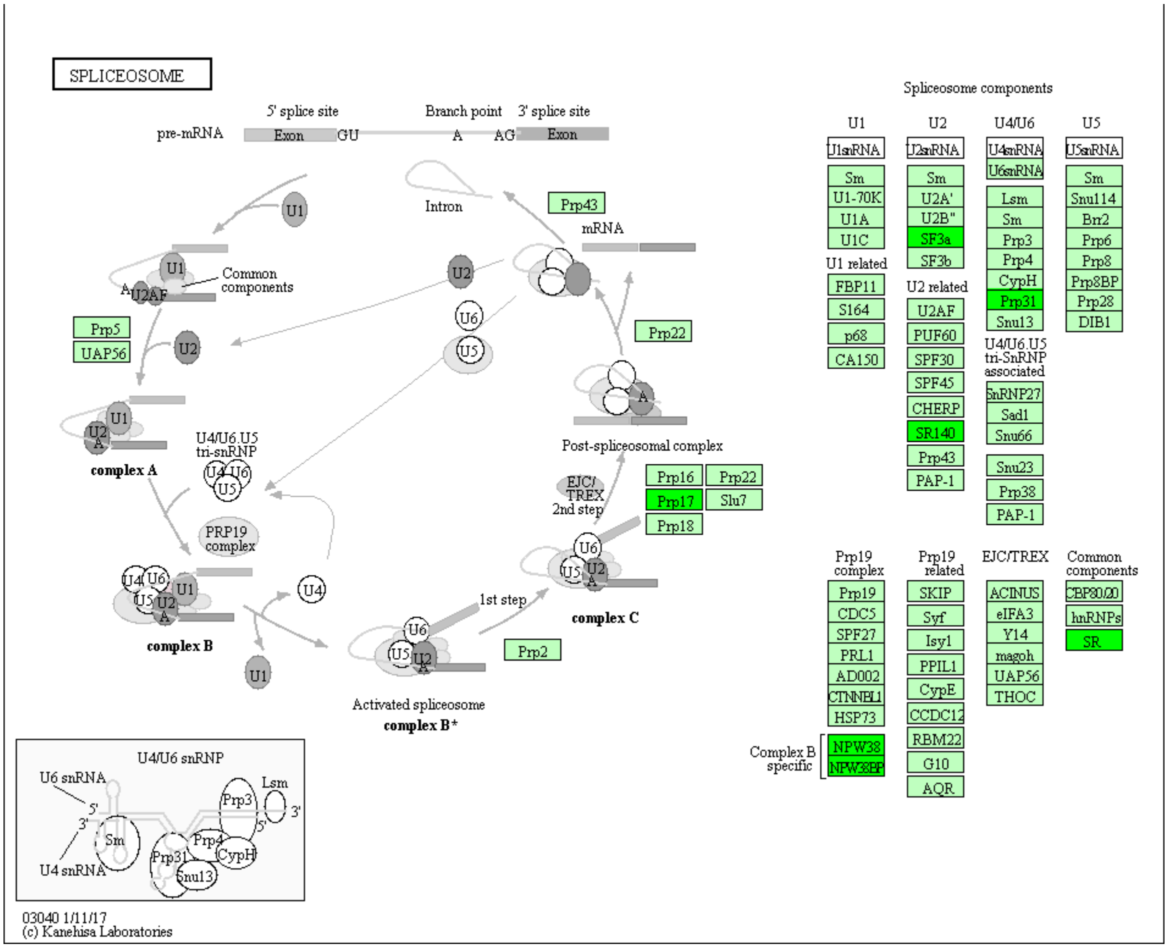


Figure 6 Enriched KEGG pathway.

2.3 Functional Enrichment and cluster analysis of Differentially Quantified Proteins

2.3.1 GO Enrichment

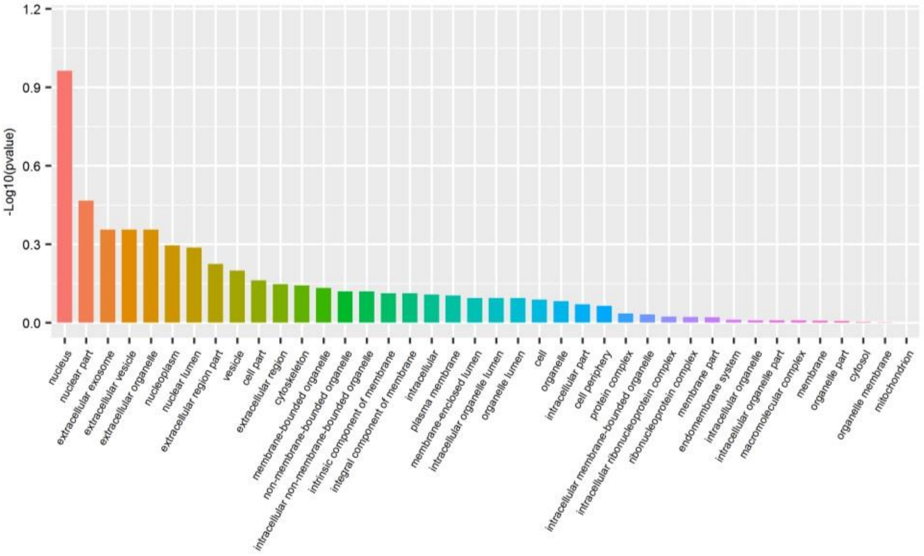


Figure 7 GO-based enrichment analysis.

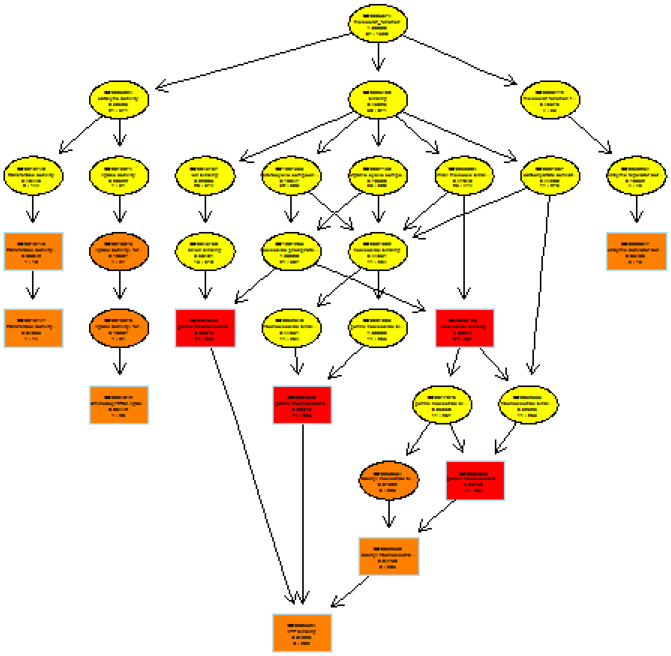


Figure 8 Directed acyclic graph (DAG) analysis. Boxes indicate the 10 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). Black arrows indicate is-a relationships and red arrows part-of relationships.

2.3.2 KEGG Enrichment

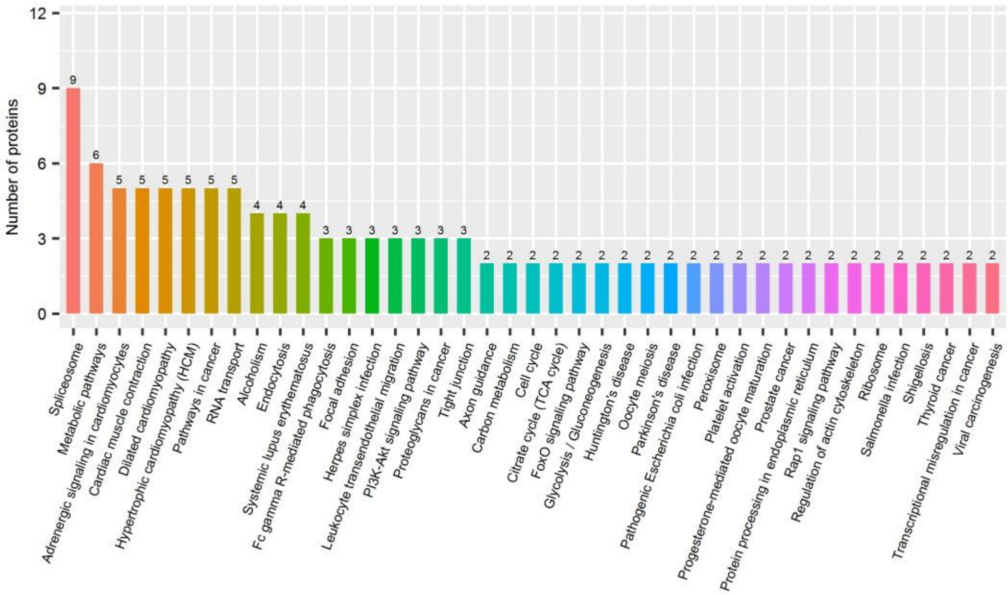


Figure 9 The top 40 pathways for differentially expressed proteins in KEGG enrichment analysis.

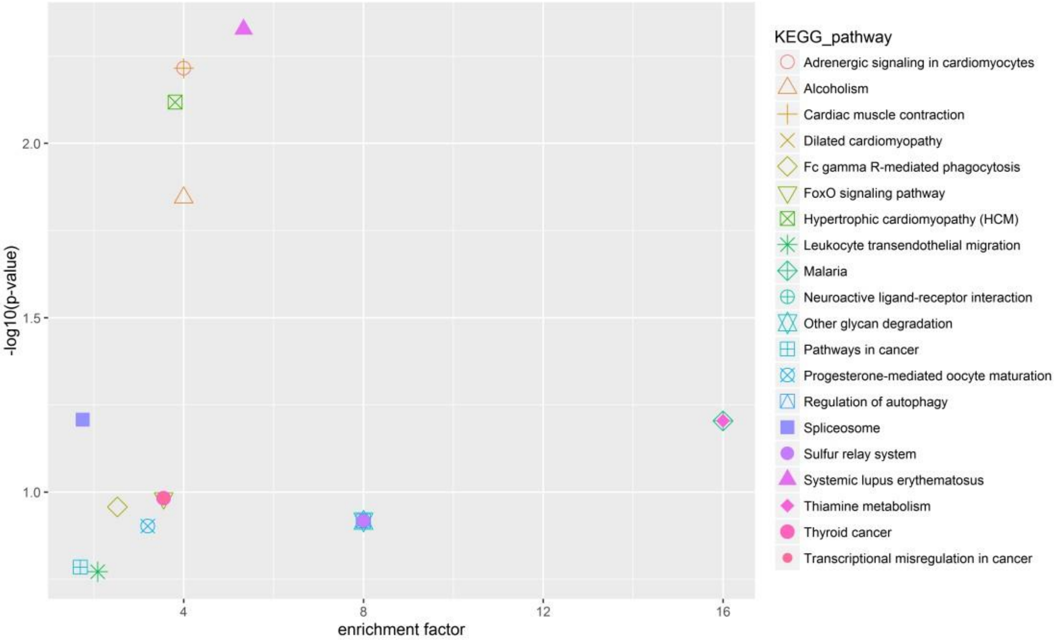


Figure 10 The top 20 enriched KEGG pathways with lowest p value for differentially expressed proteinis.

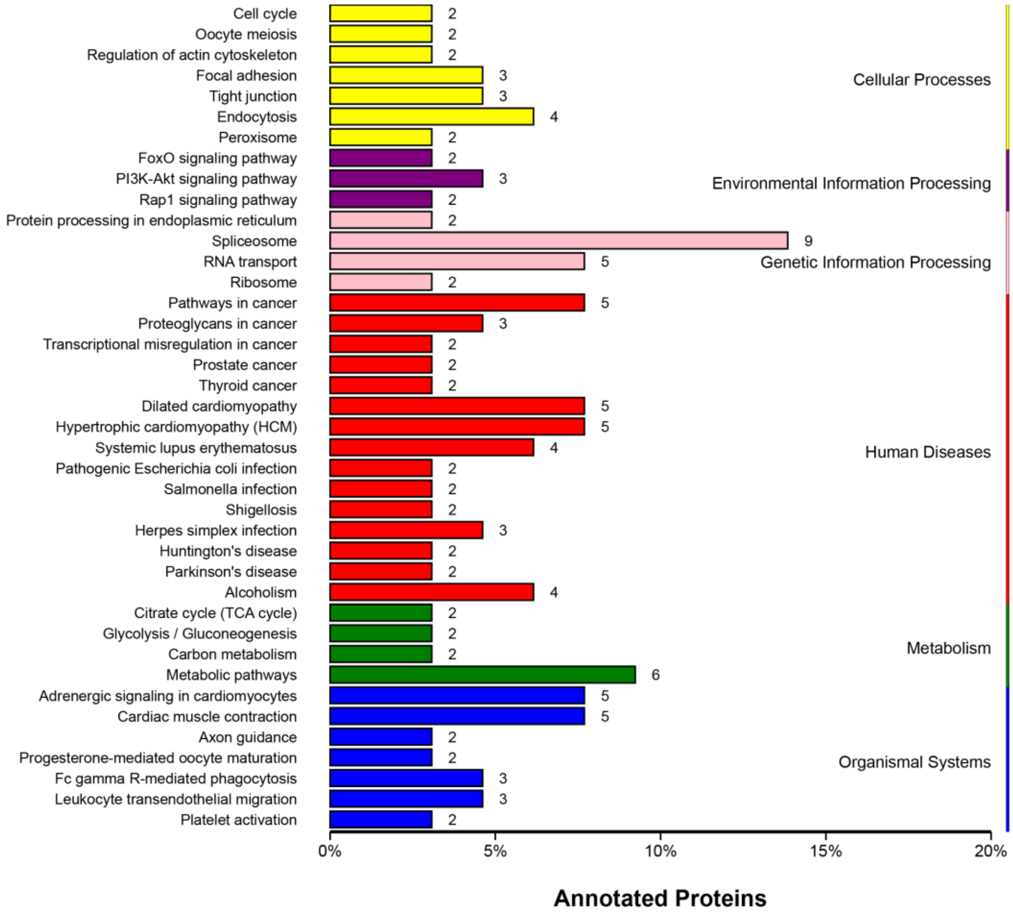


Figure 11 Classification for differentially expressed proteins annotated in KEGG pathway. These pathways can be divided into five classes: Metabolism, Cellular processes, Genetic information processing, Environmental information processing, and Organismal systems.

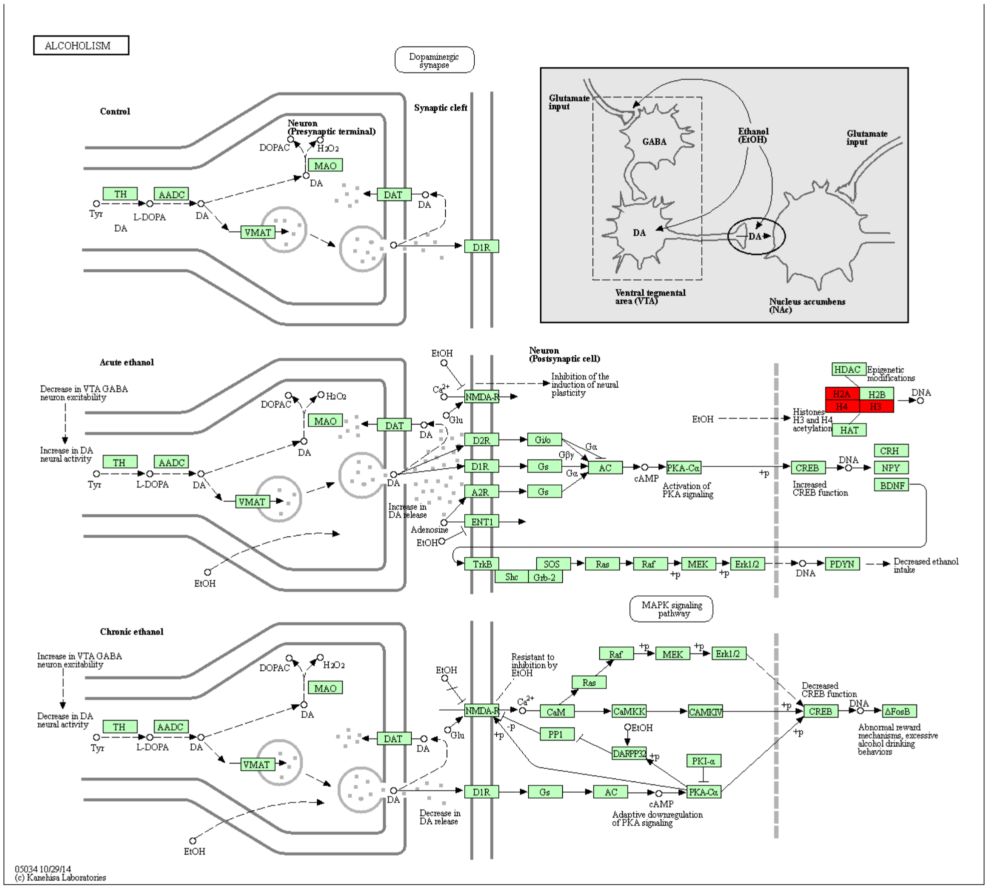


Figure 12 Enriched KEGG pathway. Up regulated proteins were shown inred

2.3.3 Cluster analysis

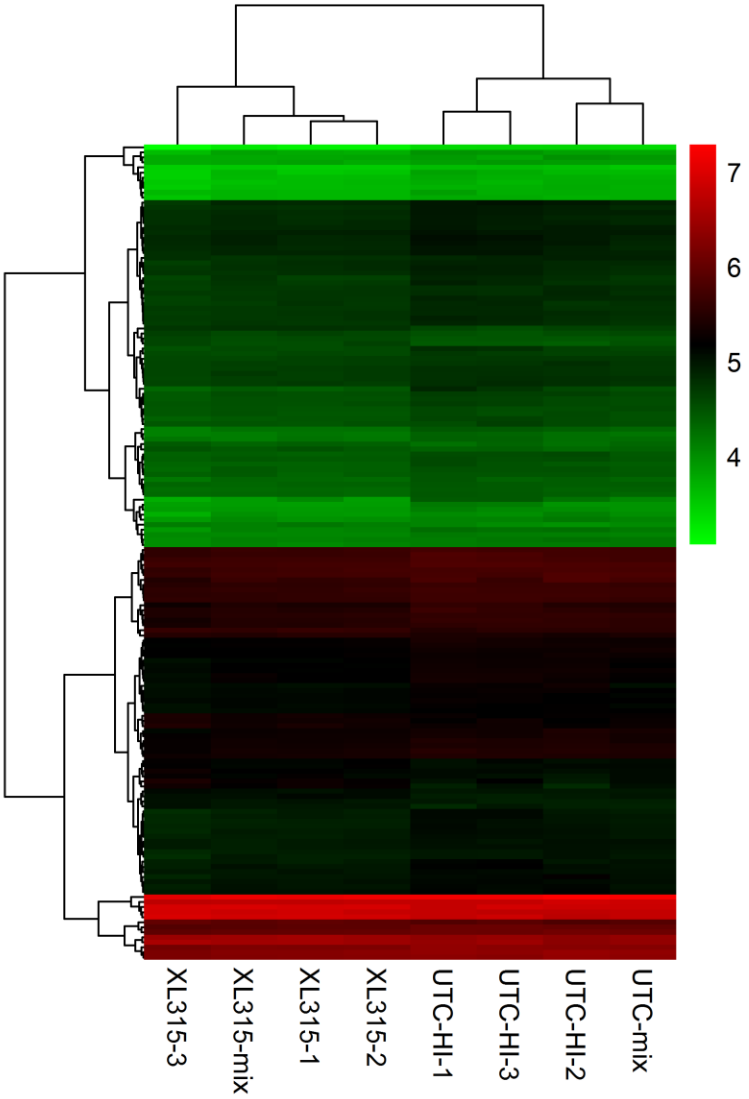


Figure 13 Cluster analysis for 8 samples.

2.3.4 Protein-protein interaction analysis

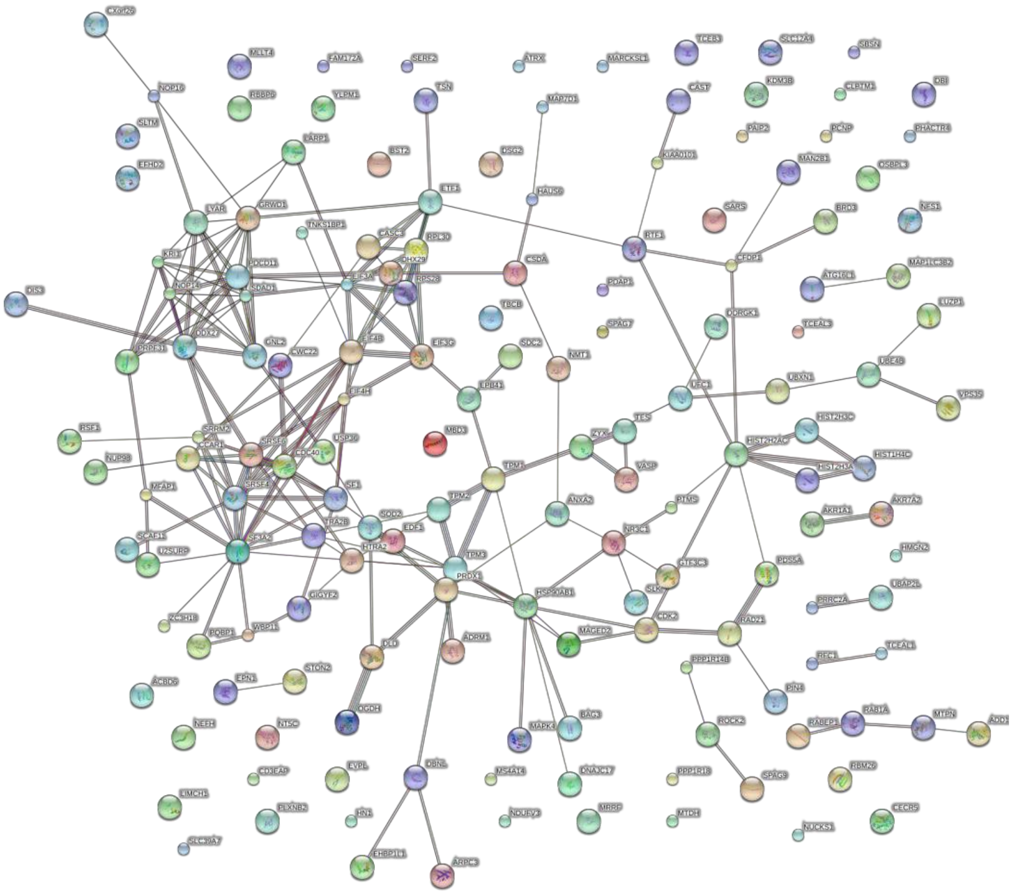


Figure 14 Protein-protein interactions analysis.

3. Suggestions for Further Studies

Based on the results in this study, we recommend narrowing down the list of proteins of interest and then performing functional studies for the target candidates.

Among those differentially expressed proteins, here are general guidelines for target selection:

1) Based on the quantitative results, pay more attention on those proteins with the most significantly expression changes, either up-regulated or down-regulated;

2) Pay more attention on those differentially expressed proteins with specific functions, for example, transcription factors, enzymes, signaling proteins or reported proteins with important function;

3) According to the results from bioinformatic analyses, choose those markedly changed proteins in some specific pathways, processes, molecular functions, localization protein complex, etc. ;

4) Develop antibodies specifically against those selected targets if they are not commercial available, to validate the selected targets biochemically, for example, western blotting (WB) experiments.

4. Bioinformatics Analysis Methods

4.1. Annotation Methods

GO Annotation

The Gene Ontology, or GO, is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. More specifically, the project aims to:

1. Maintain and develop its controlled vocabulary of gene and gene product attributes;

2. Annotate genes and gene products, and assimilate and disseminate annotation data;

3. Provide tools for easy access to all aspects of the data provided by the project.

The ontology covers three domains:

1. Cellular component: A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

2. Molecular function: Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place.

3. Biological process: A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

Gene Ontology (GO) annotation proteome was derived from the UniProt-GOA database (www. http://www.ebi.ac.uk/GOA/). Firstly, Converting identified protein ID to UniProt ID and then mapping to GO IDs by protein ID. Then proteins were classified by Gene Ontology annotation based on three categories: biological process, cellular component and molecular function.

KEGG Pathway Annotation

KEGG connects known information on molecular interaction networks, such as pathways and complexes (the "Pathway" database), information about genes and proteins generated by genome projects (including the gene database) and information about biochemical compounds and reactions (including compound and reaction databases). These databases are different networks, known as the "protein network", and the "chemical universe" respectively. There are efforts in progress to add to the knowledge of KEGG, including information regarding ortholog clusters in the KEGG Orthology database. KEGG Pathways mainly including:

Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Rat Diseases, Drug development. Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to annotate protein pathway. Kobas 2.0, a widely used tool for annotation and identification of enriched pathways and diseases, was used to annotate protein's KEGG database description.

4.2. Functional Enrichment

Enrichment of Gene Ontology analysis

Proteins were classified by GO annotation into three categories: biological process, cellular compartment and molecular function. For each category, a Fisher's exact test was employed to test the enrichment of the differentially expressed protein against all identified proteins. The GO with a corrected p-value < 0.05 is considered significant. GO is generally organized as a tree-like hierarchy of functional terms, in which each term can have children that are more-specific classes of the parent class. As a matter of fact, the GO hierarchy is a directed acyclic graph (DAG) rather than a tree, as a GO term can have several parents. R package topGO was also used to perform DAG analysis.

Enrichment of pathway analysis

Encyclopedia of Genes and Genomes (KEGG) database was used to identify enriched pathways by a Fisher's exact test to test the enrichment of the differentially expressed protein against all identified proteins. The pathway with a corrected p-value < 0.05 was considered significant. Except for enrichment analysis, these differentially expressed proteins annotated in different pathways were classified into 5 classes based on the definition of KEGG database.

4.3 Cluster analysis

In data mining, cluster analysis is used to classify a set of observations into two or more mutually exclusive unknown groups, based on combinations of the interval variables. The purpose is to discover a system of organizing observations, usually genes, and proteins into groups, where members of the groups share properties in common.

Clustering Method: The protein expression matrix was transformed by the function x = −log10 (X). These x values were then clustered by hierarchical clustering (Euclidean distance, average linkage clustering) in Genesis. Cluster membership was visualized by a heat map using the "heatmap.2" function from the "gplots" R-packag

4.4 Protein-protein interaction analysis

As one of the most important interactions, Protein-Protein Interactions have been studied widely. So far large scale protein-protein interactions have been identified, and all the generated data collected together in specialized databases, enables the creation of large protein interaction networks. String database was used to determine protein-protein interactions for differentially expressed proteins. The combined score 0.4 was used as threshold value.

Table 7 Summary of software used in this project

Tools	Version	Description	Linkages
KOBAS	2.0	KEGG Orthology Based Annotation System	http://kobas.cbi.pku.edu.cn/
Blast2GO	2.8.3	An open source software platform for visualizing complex networks	http://www.cytoscape.org/
topGO	2.26.0	An R package for gene ontology enrichment analysis	https://bioconductor.org/packages/release/bioc/html/topGO.html

Table 8 Summary of databases used in this project

Database	Description	Homepage
GO	Gene Ontology database	http://www.geneontology.org/
COG	Clusters of Orthologous Groups	http://www.ncbi.nlm.nih.gov/COG/
String	Functional protein association networks	http://string-db.org/
KEGG	The database of Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/

References

[1] Ashburner M, Ball C A, Blake J A, Botstein D, et al. Gene ontology: tool for the unification of biology. Nature Genetics 2000, 25(1): 25-29.

[2] Conesa A, Götz S, García-Gómez J M, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005, 21(18): 3674-3676.

[3] Kanehisa M, Goto S, Kawashima S, Okuno Y, et al. The KEGG resource for deciphering the genome. Nucleic Acids Research 2004, 32(Database issue):D277-D280.

[4] Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C. and Wei, L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011, 39,

W316-322.