obaDIA: one-step biological analysis pipeline for quantitative proteomics data from data independent acquisition mass spectrometry

Jun Yan[1,2], Hongning Zhai[2], Ling Zhu[3] and Xiaojun Ding[2,*]

[1]Department of Crop Genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China, [2] e-omics Technology Inc., Beijing, 102206, China, [3] Inner Mongolia Medical University, Inner Mongolia 010110, China

*To whom correspondence should be addressed.

Abstract

Motivation: The downstream biological analysis of DIA-based proteomic data including protein abundance statistics, differential expression, functional annotation and enrichment analysis for variety databases is a crucial part for proteomic research, but few integrated tools and solutions are available, which leads to complex analytical processes and irreproducible analytical results.

Results: We developed obaDIA, an integrated one-step pipeline for biological analysis from DIA data. ObaDIA is a fast tool which generate highly visualized, comprehensive and repeatable biological analysis results for DIA data; a novel enrichment strategy implemented in obaDIA may reduce false positive and increase the interpretability of results. Taken together, ObaDIA contributes to the standard biological analysis of proteomic data.

Availability: The obaDIA is freely available from https://github.com/yjthu/obaDIA.git.

Contact: dingxiaojun@e-omics.com

1 Introduction

Data independent acquisition (DIA) mass spectrometry-based proteomics which can detect and quantify more complete proteins in multiple samples has gained popularity in proteomics research community. A lot of software tools have been devoted to protein discovery and quantification from DIA data, such as OpenSWATH, Spectronaut, Skyline and DIA-NN for selected library-based DIA, and DIA-Umpire, PECAN, PEAKS and Group-DIA for library-free DIA(Zhang, et al., 2020). However, few tools are available for aggregating the detected proteins and the abundance data into comprehensive biologically meaningful results to facilitates biological or biomedical researches.

Although some software tools have been developed for mining downstream biological interest from DIA-based proteomics, such as MSstats(Choi, et al., 2014), mapDIA(Teo, et al., 2015), Perseu(Tyanova and Cox, 2018) and Automated-workflow-composition(Palmblad, et al., 2019), they either provide several statistical modules or just offer a framework for building workflows, with no complete solution provided. The complexity of data format conversion, the lack of visualization and reproducibility of analysis results, the limitation of online tools on species

support and data size, and the inability of automate batch analysis are still major bottleneck for downstream biological analysis. A solution is to build standard workflows to ensure the reproducibility of results, but it is not easy for researchers without bioinformatics skills. Here we describe obaDIA, an integrated pipeline tool, which is very easy to use and capable of generating comprehensive and reproducible results for downstream biological analysis from DIA proteomic data.

2 Application Description

2.1 Workflow of obaDIA

Take two standard formatted files (protein abundance matrix and protein sequences) which can be acquired from common DIA data extraction tools as inputs, obaDIA is capable of protein abundance normalization, differential expression, functional annotation and enrichment analysis in a completely automate way. The obaDIA workflow is executed through three core functional modules. The “AB & DE” module preforms protein abundance (AB) and differential expression (DE) analysis. First, the input abundance matrix is normalized by dividing the protein intensities by the total intensity, to eliminate the systematic variation across samples; second, the normalized intensities are averaged within each group; then, the normalized and mean protein abundance matrices are used for protein abundance distribution and correlation assessments at sample and group levels, respectively; last, batch differential protein expression analysis are performed for all predefined comparisons. The “Annotation” module acquires the protein sequences and carries out detailed protein functional annotation and classification on SWISS-PROT, PFAM, GO, KEGG, eggNOG and signalP databases. The “Enrichment” module receives the protein sequences and the output files from “AB & DE” and “Annotation” as inputs, then completes functional enrichment analysis and visualization of GO, KEGG, PFAM and Rectome databases for both expressed and differentially expressed proteins for all predefined comparisons (Figure 1).

2.2 Featured biological analysis of DIA data

A major advantage of obaDIA is that it can perform multiple downstream biological analysis in one step and generate comprehensive statistical charts, through integration of third-party tools and self-developed analysis modules. For example, after systematic normalization and visualization of protein abundance data by using embedded modules, obaDIA automatically generates the input format required by mapDIA(Teo, et al., 2015) to conducts batch differential expression analysis, then formats and visualizes the original output results of mapDIA and delivers them to the next module. Several state-of-the-art tools, such as Trinotate(Bryant, et al., 2017) and Kobas(Mao, et al., 2005), were optimized in terms of either data format or analysis strategy in obaDIA pipeline, in order to fully exploit the biological information within DIA data.

2.3 Functional enrichment and visualization strategy

A novel strategy is introduced in obaDIA to optimize protein functional enrichment analysis. Specifically, all expressed proteins are enriched using total genes in the pathway as background in

the first round; then, the differentially expressed proteins for each comparison are enriched using all expressed proteins as background to detect over-represented pathways. Since metabolic pathways in commonly used databases are established based on genes but not expressed proteins, while expressed proteins just account for a small part of total genes in an organism, the method using the total genes as background may bias the statistical significance of enriched terms. This novel enrichment strategy is capable of estimating the enrichment levels of differentially expressed proteins more conservatively and accurately. Furthermore, HTML formatted report can be generated in obaDIA by using KEGG API, the information of enriched KEGG pathways and expressed proteins can be displayed interactively, which facilitate biological information mining.

3 Implementation and Results

The main program of obaDIA was developed in Perl, and incorporated R, Python, and Bash scripts. All third-party software tools and dependencies of obaDIA are listed in Supplementary Table 1, and are easy to install following the documentation. The obaDIA pipeline has been tested on Linux platform CentOS release 7.0, and should work on similar Linux/Unix systems which support Bash command line. The source code, wrapped dependencies and user manual of obaDIA can be obtained from https://github.com/yjthu/obaDIA.git.

The obaDIA pipeline has been tested using a published demo data(Teo, et al., 2015) and an unpublished complete DIA data. All input and output files from the published demo data set are wrapped with obaDIA release, and the main results are also shown and described in Figure 2-6. The obaDIA pipelines can be completed within two hours in both data sets on our workstation (Table 2). In terms of the enrichment analysis results, we compared the two built-in enrichment methods which use total genes and expressed proteins as background, respectively; the results show differed *P* values of the enriched terms. It is noteworthy that when using all genes in the pathway as background, it is considered statistically significant for some enriched terms even if there are only one or two differentially expressed proteins in it; this should not be reasonable than using expressed proteins as background. In addition, ranking of the enriched terms are also changed when using different methods (Figure 7), further indicating that the novel enrichment strategy in obaDIA potentially reduces false positive and may be of biological significance. In conclusion, obaDIA provides an integrated one-step biological analysis solution and a new enrichment strategy, which might help standardized biological analysis for proteomics data.

Funding
This work has been supported by BAAFS (KJCX20180425 and KJCX20180204) to XY and NSFC (31621001) to LL.

Conflict of Interest: none declared.

Tables and Figures

Table 1 Integrated tools and dependencies in obaDIA

Module	Dependencies	References	Wrapped
AB & DE*	R packages: corrplot, pheatmap, PerformanceAnalytics		no
	mapDIA	(Teo, et al., 2015)	yes
Annotation	Trinotate-v3.1.0-pro (a modified version)	(Bryant, et al., 2017)	yes
	sqlite3		yes
	diamond	(Buchfink, et al., 2015)	yes
	hmmscan & hmmpress	(Potter, et al., 2018)	yes
	signalP 4.1	(Petersen, et al., 2011)	no
	Perl5 libraries: GO::Parser, DBI, DBD::SQLite		no
Enrichment	Kobas 3.0	(Mao, et al., 2005)	no
	R packages: GO.db, goseq, goTools, topGO, Rgraphviz		no
	Python2 libraries: eventlet, bs4,django==1.8.3		no
Basic	R packages: reshape2, ggplot2, optparse		system
	Perl5		system
	Python2		system

*AB: abundance analysis, DE: differntial expression anaysis

Table 2 Time consumption of obaDIA workflow in test data sets

Test data description	Step	Running time (minutes) *
Data set 1:	AB & DE	1
Species: Human	Annotation	13
Six groups and each group with three biological replicates	Enrichment	59
566 proteins, 5 comparisons	Total	73
Data set 2:	AB & DE	1
Species: Mouse	Annotation	46
Four groups and each group with three biological replicates	Enrichment	71
4493 proteins, 2 comparisons	Total	118

* The main factor affecting the running time in different systems is CPU. The CPU we used for testing is: Intel(R) Xeon(R) CPU E5-2687W v3 @ 3.10GHz. Threads number used by each tool is listed in Table 1.

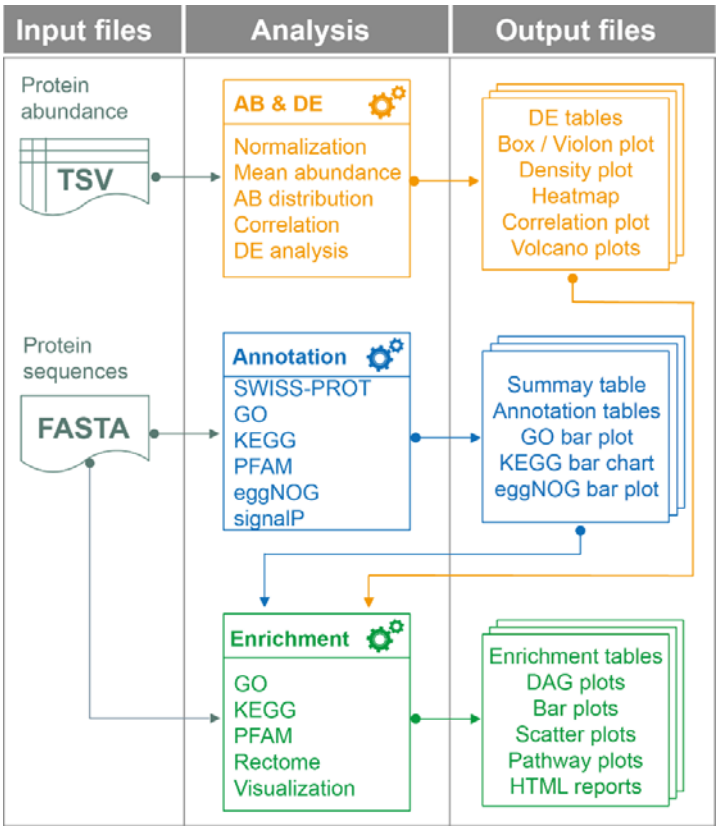


Figure 1. Diagram of the workflow of obaDIA

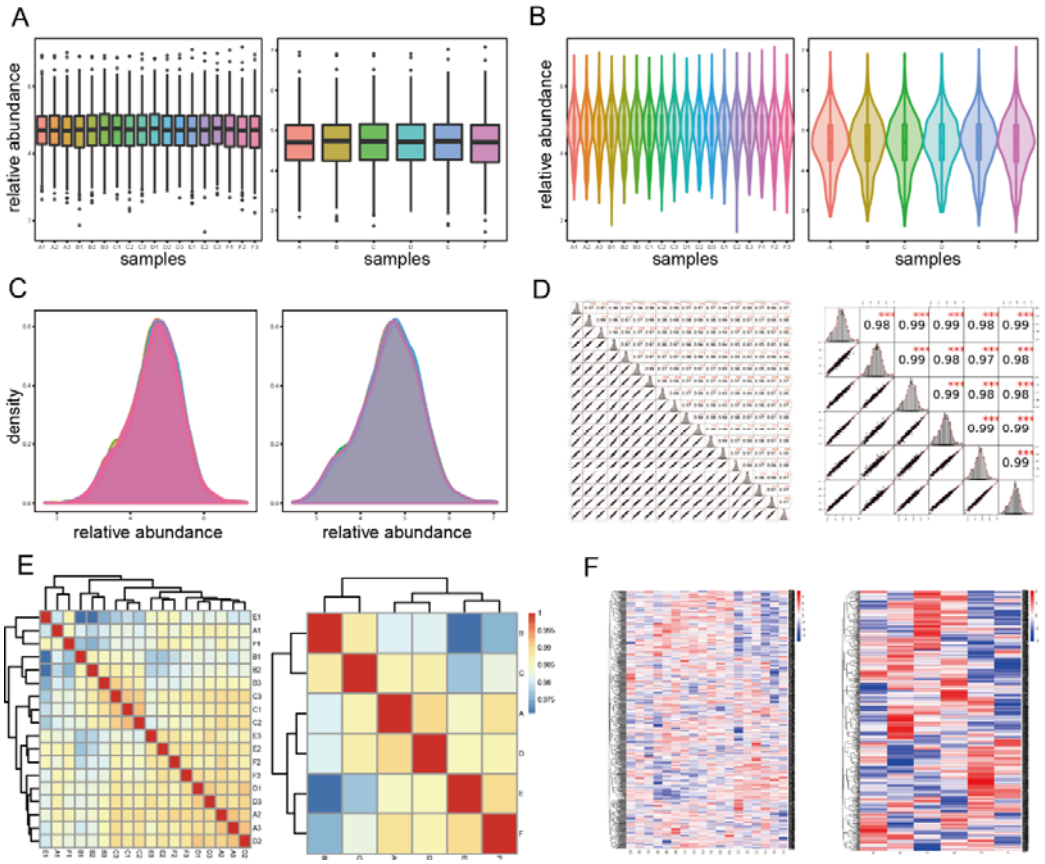


Figure 2. Protein abundance analysis results. The raw abundance matrix is normalized to generate a relative abundance matrix and a group mean abundance matrix, then they are scaled to $\log_{10}(value+1)$ for plotting. (A) Box plot of relative abundance for each sample (left) and group (right). (B) Violin plot of relative abundance for each sample (left) and group (right). (C) Density plot of relative abundance for each sample (left) and group (right). (D) Correlation matrix for each comparison of samples (left) and groups (right). Left-bottom shows scatter plots, right-top shows Pearson correlation coefficients(PCC), and diagonal shows histograms of of the relative abundance. An asterisk indicates statistical significance. (E) Correlation heatmap for each comparison of samples (left) and groups (right). Red color indicates higher PCC while blue indicates lower. (F) Relative abundance heatmap for each sample (left) and each group (right). Red color indicates higher abundance while blue indicates lower.

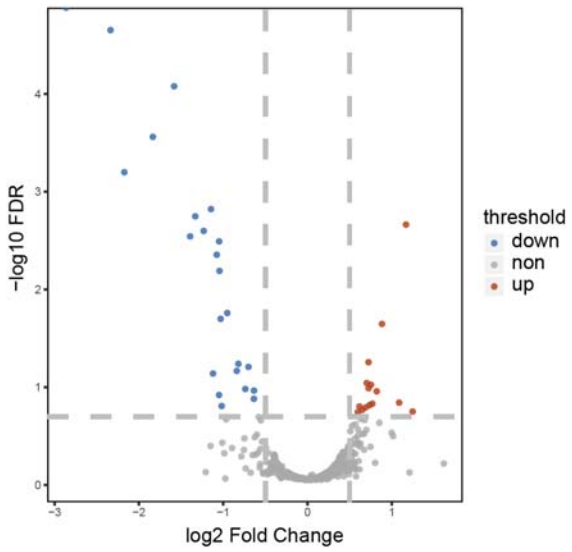


Figure 3. Volcano plot for differential expressed proteins filtering. Red, blue and grey dots represent up, down expressed and non-DE proteions, respectively. A greater absolute value of X axis indicates a greater expression difference, while a higher value of Y axis indicates a more significant expression difference between the two samples.

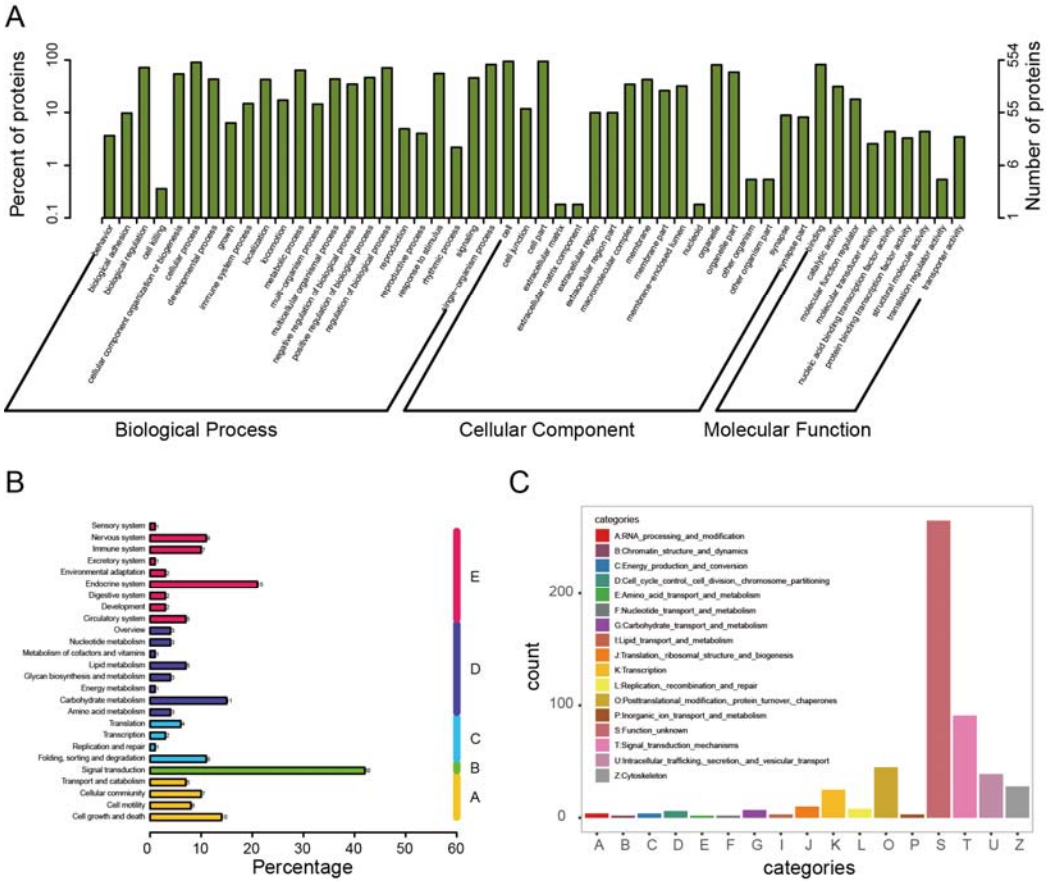


Figure 4. GO, KEGG and eggNOG classification of protein functional annotation results. (A) Bar plot of GO classification. Top 50 second-level GO terms are ploted based on the first-level GO categories: biological process, cellular component and molecular function. The number and percentage of proteins in each GO term(including its child terms) are shown. (B) Bar chart of KEGG classification. All KEGG terms are assigned to five main categories: A(Cellular Processes, B(Environmental Information Processing, C(Genetic Information Processing), D(Metabolism), and E(Organismal Systems). (B) Bar plot of eggNOG classification. All eggNOG terms are assigned into categories A-Z.

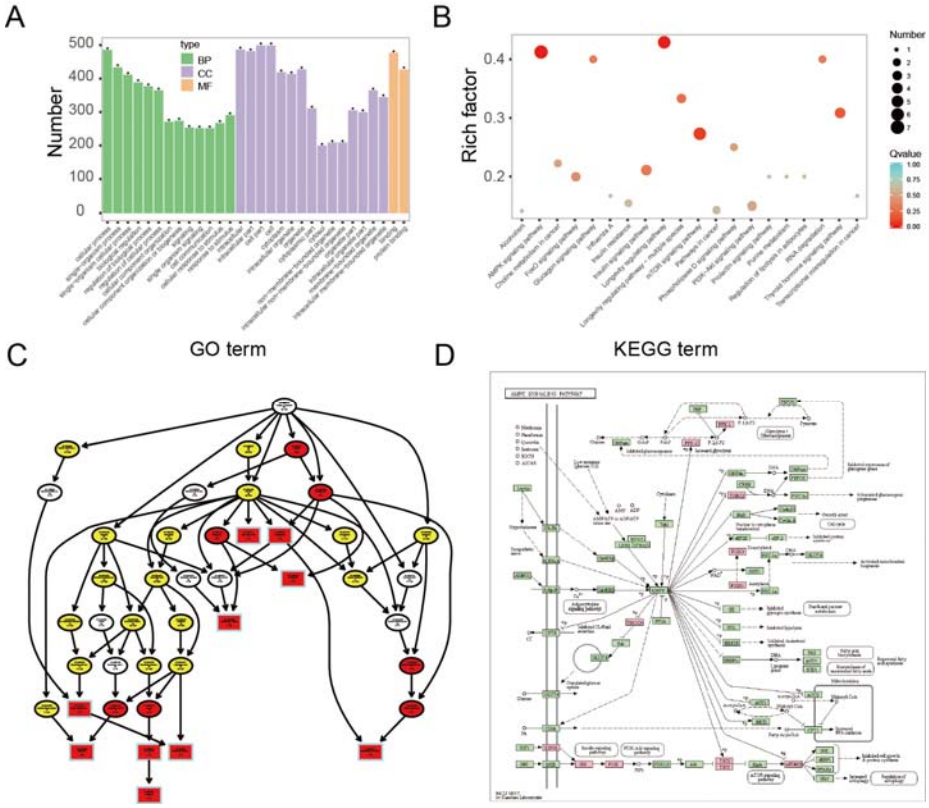


Figure 5. Representative results of functional enrichment analysis. (A) Bar plot of GO enrichment. Top 30 GO terms are ploted based on three categories: BP(biological process), CC(cellular component) and MF(molecular function). An asterisk indicates statistical significance. (B) Scatter plot of KEGG enrichment. Rich factor is the ratio between the input number and the total number of proteins in the same pathway. (C) DAG (Directed Acyclic Graph) plot of GO encrichment. Branches represent the inclusion relationship, and the function scope defined from top to bottom is getting smaller and smaller. Generally, the top 10 most enriched GO terms are displayed with rectangles while others with ellipticals, and different degree of enrichment are shown in red, yellow and white colors, respectively. GO term identifier, description, number of proteins in the term are shown in each node. (D) A pathway plot obtained through KEGG API. Expressed proteins in the pathway nodes are shown with pink color.

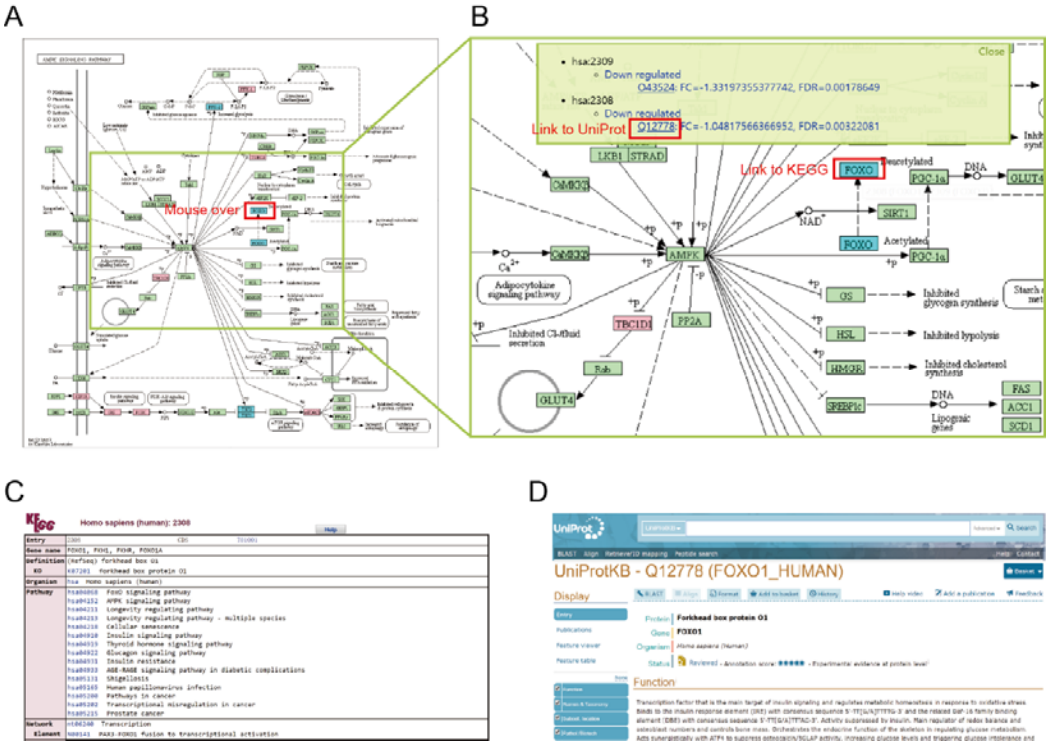


Figure 6. KEGG pathway visualization. (A) AMPK signaling pathway plot obtained through KEGG API. Expressed proteins are shown with pink color and differentially expressed proteins are shown with red (upregulation, there is no upregulated protein in this example) or cyan (down-regulation) color. (B) Local amplification of the AMPK signaling pathway plot. A display box pops up when mouse over a down regulated protein *FOXO*, which shows the protein UniProt identifier, fold change and FDR values of the differential expressed analysis. When click on a protein name node in the pathway, it will link to the KEGG website (C), while when click on the protein UniProt identifier in the display box, it will link to the UniProtKB website (D).

A

DE / T

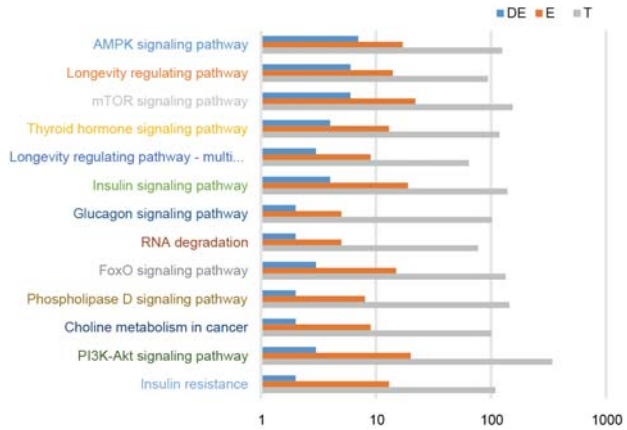
#Term	Input number	Background number	P-Value
AMPK signaling pathway	7	125	4.35E-11
Longevity regulating pathway	6	94	5.59E-10
mTOR signaling pathway	6	154	9.54E-09
Thyroid hormone signaling pathway	4	118	5.72E-06
Insulin signaling pathway	4	139	1.07E-05
Longevity regulating pathway - mult	3	64	3.69E-05
FoxO signaling pathway	3	134	0.000308206
RNA degradation	2	77	0.002611174
PI3K-Akt signaling pathway	3	342	0.004351807
Choline metabolism in cancer	2	101	0.004388292
Glucagon signaling pathway	2	102	0.004471655
Insulin resistance	2	109	0.005075636
Phospholipase D signaling pathway	2	144	0.008617722
Regulation of actin cytoskeleton	2	215	0.018288442
Dorso-ventral axis formation	1	28	0.027332339
Fructose and mannose metabolism	1	33	0.031970707
Notch signaling pathway	1	48	0.045756964
Endometrial cancer	1	52	0.049400893
Regulation of lipolysis in adipocytes	1	56	0.053031272

B

DE / E

#Term	Input number	Background number	P-Value
AMPK signaling pathway	7	17	0.000351051
Longevity regulating pathway	6	14	0.000841578
mTOR signaling pathway	6	22	0.005605883
Thyroid hormone signaling pathway	4	13	0.017552447
Longevity regulating pathway - mult	3	9	0.034081318
Insulin signaling pathway	4	19	0.049622062
Glucagon signaling pathway	2	5	0.06625749
RNA degradation	2	5	0.06625749
FoxO signaling pathway	3	15	0.09730545
Phospholipase D signaling pathway	2	8	0.126056912
Choline metabolism in cancer	2	9	0.148134033
PI3K-Akt signaling pathway	3	20	0.170322953
Insulin resistance	2	13	0.242134833
Pathways in cancer	2	14	0.266320629
Purine metabolism	1	5	0.323997553
Regulation of lipolysis in adipocytes	1	5	0.323997553
Prolactin signaling pathway	1	5	0.323997553
Influenza A	1	6	0.36695423
Transcriptional misregulation in canc	1	6	0.36695423

C



D

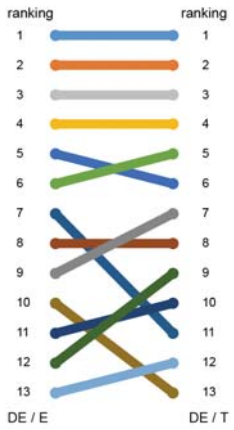


Figure 7. Comaprison of two enrichment methods. (A) Snapshot of KEGG enrichment results table using total(*T*) genes in each pathway as backgroud; yellow color highlighted the statistical significant($P < 0.05$) terms. (B) Snapshot of KEGG enrichment results table using expressed(*E*) protein in each pathway as backgroud; yellow color highlighted the statistical significant($P < 0.05$) terms. (C) Bar chart shows numbers of differentially expressed(*DE*), expressed(*E*) and total(*T*) proteins in each pathway. (D) Comaprison of the ranking of enriched KEGG tems using *E* as backgroud (*DE/E*) and using *T* as backgroud(*DE/T*) in the enrichment analysis results.

References

Bryant, D.M., *et al*. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep* 2017;18(3):762-776.

Buchfink, B., Xie, C. and Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12(1):59-60.

Choi, M., *et al*. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 2014;30(17):2524-2526.

Mao, X., *et al*. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 2005;21(19):3787-3793.

Palmblad, M., *et al.* Automated workflow composition in mass spectrometry-based proteomics. *Bioinformatics* 2019;35(4):656-664.

Petersen, T.N., *et al.* SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;8(10):785-786.

Potter, S.C., *et al*. HMMER web server: 2018 update. *Nucleic Acids Res* 2018;46(W1):W200-W204.

Teo, G., *et al*. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J Proteomics* 2015;129:108-120.

Tyanova, S. and Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol Biol* 2018;1711:133-148.

Zhang, F., *et al*. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics* 2020:e1900276.