obaDIA: one-step biological analysis pipeline for data-independent acquisition and other quantitative proteomics data

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

[1]Department of Crop Genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China, [2]Beijing Mingde Zhengkang Medical Research Co., Ltd., Beijing 102206, China, [3]Inner Mongolia Medical University, Huhehot 010110, Inner Mongolia, China

*To whom correspondence should be addressed.

Abstract

Motivation: Data mining and data quality evaluation are indispensable constituents of quantitative proteomics, but few integrated tools available.

Results: We introduced obaDIA, a one-step pipeline to generate visualizable and comprehensive results for quantitative proteomics data. obaDIA supports fragment-level, peptide-level and protein-level abundance matrices from DIA technique, as well as protein-level abundance matrices from other quantitative proteomic techniques. The result contains abundance matrix statistics, differential expression analysis, protein functional annotation and enrichment analysis. Additionally, enrichment strategies which use total proteins or expressed proteins as background are optional, and HTML based interactive visualization for differentially expressed proteins in the KEGG pathway is offered, which helps biological significance mining. In short, obaDIA is an automatic tool for bioinformatics analysis for quantitative proteomics.

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

Contact: dingxiaojun@e-omics.com

1 Introduction

Data-independent acquisition (DIA) proteomics technology is getting more popularity and recognition recently. Compared with data-dependent acquisition (DDA) strategy, DIA avoids the topN acquisition randomness and has better reproducibility, and more suitable for large cohort research. Many tools have been developed to identification and quantification DIA data, such as OpenSWATH, Spectronaut, and Skyline (Zhang, *et al.*, 2020). However, limited tool is specifically designed for one-step, fast and comprehensive downstream bioinformatics analysis of DIA data.

Some biological data analysis frameworks provide tools for downstream biological analysis of quantitative proteomics data and personalized workflow construction. For example, R/Bioconductor(Gentleman, et al., 2004) provides MSstats(Choi, et al., 2014) and DEqMS(Zhu, et al., 2020) for expression analysis, topGO(Alexa and Rahnenführer, 2009) and clusterProfile(Yu, et al., 2012) for function analysis; Bioconda(Gruning, et al., 2018) or Galaxy(Jalili, et al., 2020) offers BLAST(Altschul, et al., 1990), HMMER(Potter, et al., 2018) and DIAMOND(Buchfink, et al., 2015) for functional annotation, as well as KOBAS(Mao, et al., 2005) for enrichment analysis. Other independent software, such as mapDIA(Teo, et al., 2015), Perseus(Tyanova and Cox, 2018) and Automated-workflow-composition (Palmblad, et al., 2019), also contribute to specific downstream proteomics data mining. But an integrate one-step software is still necessary for proteomic community, considering some users prefer to obtain the standard downstream analysis results quickly. Here, we introduce obaDIA, an all-in-one pipeline which generates comprehensive and reproducible biological analysis results from DIA and other quantitative proteomics data automatically.

2 Application Description

2.1 Workflow of obaDIA

obaDIA takes two standard format files (TSV format abundance matrix and FASTA format protein sequences) as inputs and automatically performs abundance imputation, normalization, differential expression, functional annotation and enrichment analysis to produce results. It supports fragment-level, peptide-level and protein-level abundance matrices from DIA technology, and protein-level abundance matrices from other quantitative proteomics technologies. The obaDIA workflow has three core modules. The "AB & DE" module is used for abundance matrix (AB) process and batch differential expression (DE) analysis. The “Annotation” module carries out detailed functional annotation and classification for expressed proteins against SWISS-PROT, PFAM, GO, KEGG, eggNOG and signalP databases. The “Enrichment” module receives the output files from “AB & DE” and “Annotation” as inputs and performs functional enrichment analysis of GO, KEGG, PFAM and Rectome databases for differentially expressed proteins from all predefined comparisons (Figure 1).

2.2 Features of obaDIA

The obaDIA pipeline is a one-step solution to perform multiple downstream biological analysis and generate comprehensive statistical charts. Although it is all-in-one, each module can be run separately, facilitating integration into personalized workflows. Specially, an HTML formatted report can be generated, and the information of expressed proteins can be displayed in a KEGG pathway interactively, which facilitates biological information mining. obaDIA takes protein-level abundance matrices as inputs for both DIA and other techniques, which ensures its wide application in the proteomics community; it also takes fragment-level or peptide-level abundance matrices as inputs for DIA data, which produces more accurate difference analysis results by using the mapDIA algorithm(Teo, et al., 2015). In addition, obaDIA provides two enrichment strategies for users, including the traditional enrichment strategy which uses total proteins as background, and an alternative strategy which uses expressed proteins in the experiment as background. Finally, in fast analysis mode, you can get all the main analysis results in a few minutes.

3 Implementation and Results

The main program of obaDIA was developed in Perl, and incorporates R, Python, and Bash scripts. All third-party tools and dependencies are listed in Table 1. We also provide a docker image, which greatly simplifies the installation of third-party dependencies, as well as a GUI for Galaxy users. 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 the obaDIA release, and the main results are also shown and described in Supplementary Figure 2-6. On our workstation (Table 2), the obaDIA workflow can be completed within two hours in standard mode and within five minutes in fast mode. In terms of enrichment analysis results, we compared the results of two built-in enrichment strategies based on total proteins and expressed proteins, respectively. The P-value and order of the enrichment items are different (Figure 7); users should choose the enrichment method suitable for their own data according to the biological significance of the results. In summary, obaDIA provides an integrated one-step biological analysis solution, which may facilitate standardized biological analysis for proteomics data.

Funding

Funding information is not applicable.

Conflict of Interest: none declared.

Tables

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		system
	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) *	Running time (minutes)
		default mode	*fast mode*
Data set 1: Species: Human; Six groups and each group with three biological replicates; 566 proteins, 5 comparisons	AB & DE	1	1
	Annotation	13	1
	Enrichment	59	2
	Total	73	3
Data set 2: Species: Mouse; Four groups and each group with three biological replicates; 4493 proteins, 2 comparisons	AB & DE	1	1
	Annotation	46	2
	Enrichment	71	1
	Total	118	4

* 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. We used 40 threads for hmmscan in order to accelerate annotation. One thread was used for the other steps. If the time is less than one minute, count as one minute.

Figures

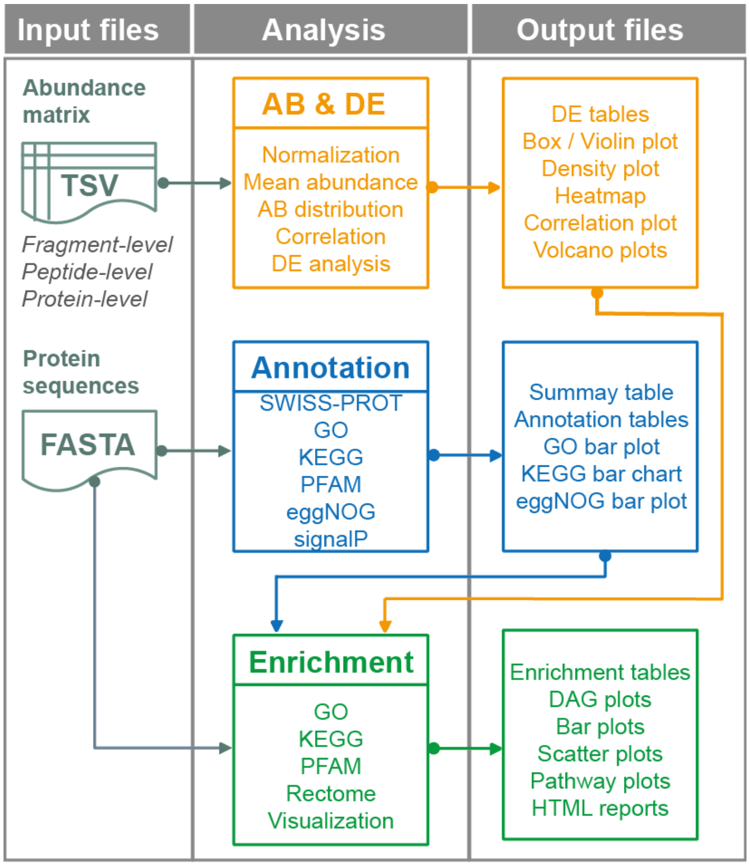


Figure 1. Diagram of the workflow of obaDIA

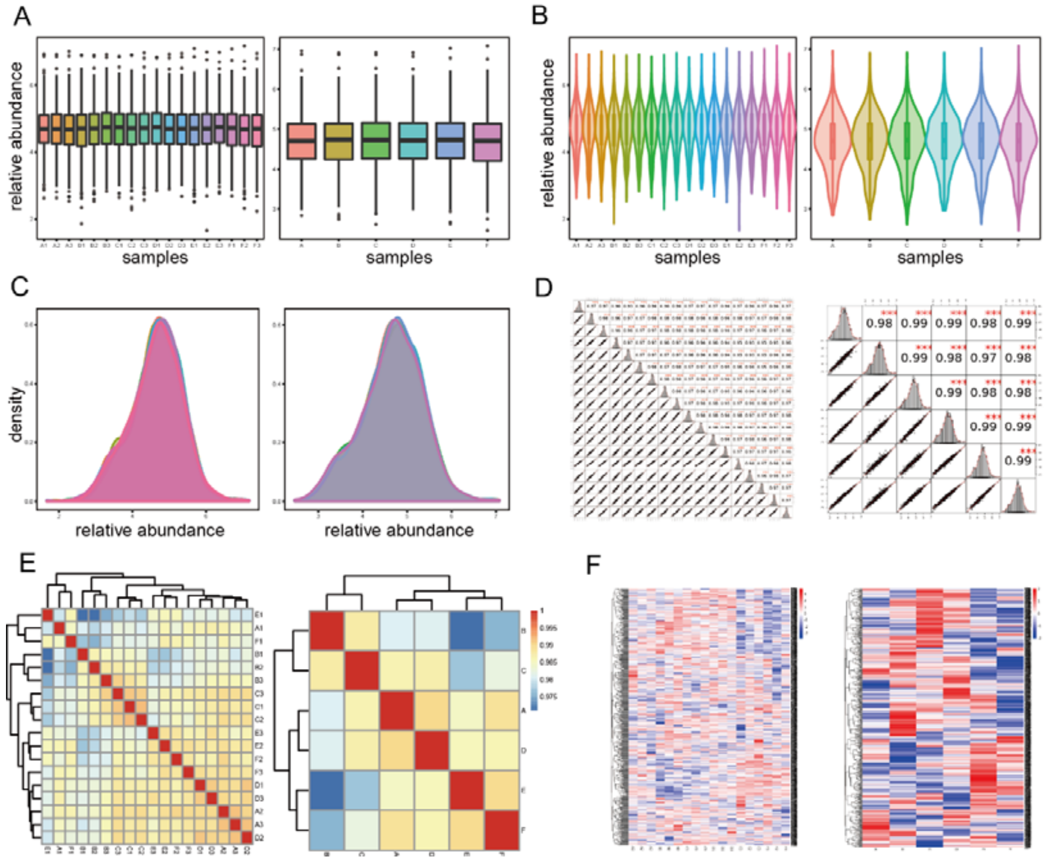


Figure 2. Protein abundance analysis results. The raw abundance matrix is normalized to generate a relative abundance matrix and a group mean abundance matrix, then they are scaled to log$_{10}$(*value*+1) for plotting. (A) Box plot of relative abundance for each sample (left) and group (right). (B) 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.

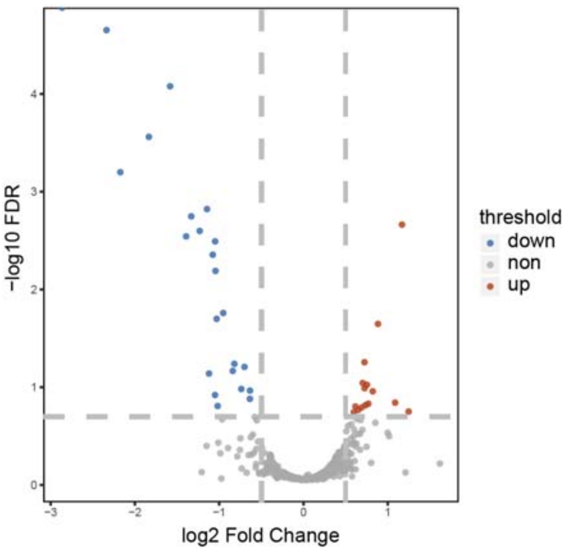


Figure 3. Volcano plot for differential expressed proteins filtering. Red, blue and grey dots represent up, down expressed and non-DE proteions respectively. A greater absolute value on the X-axis indicates a greater expression difference, while a higher value on the Y-axis indicates a more significant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

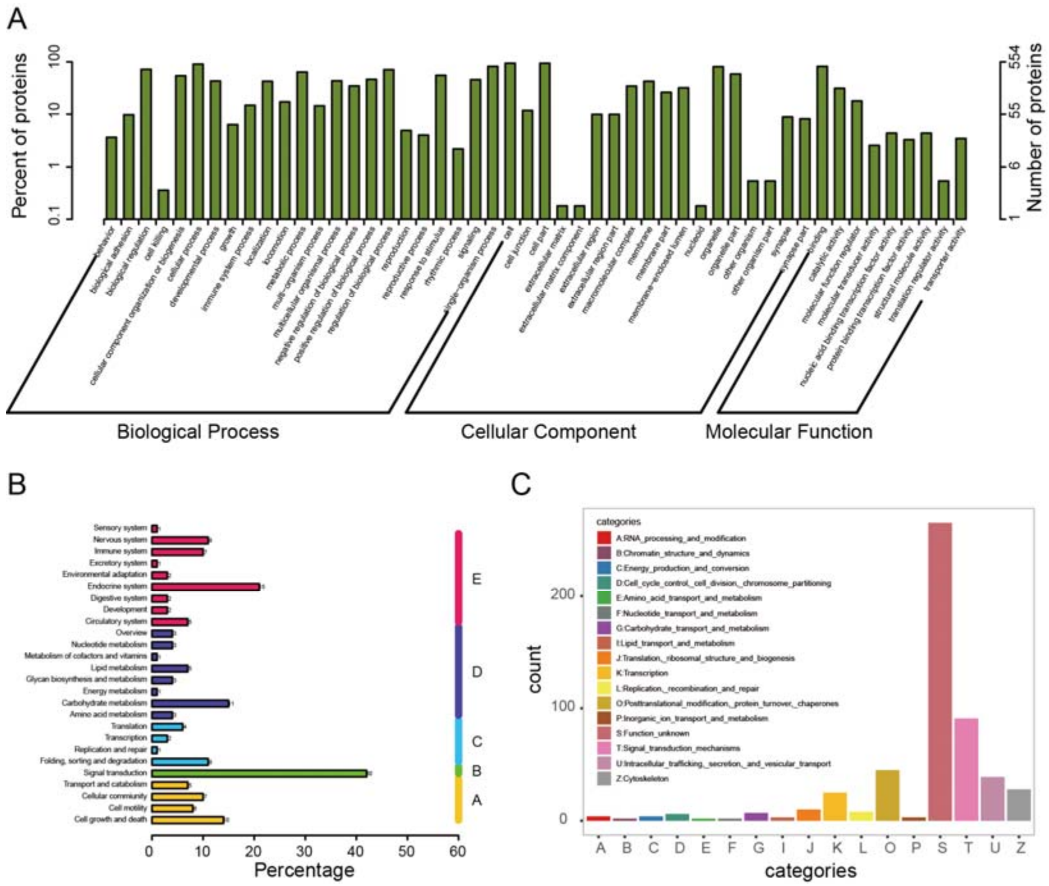


Figure 4. GO, KEGG and eggNOG classification of protein functional annotation results. (A) Bar plot of GO classification. Top 50 second-level GO terms are ploted based on the first-level GO categories: biological process, cellular component and molecular function. The number and percentage of proteins 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.

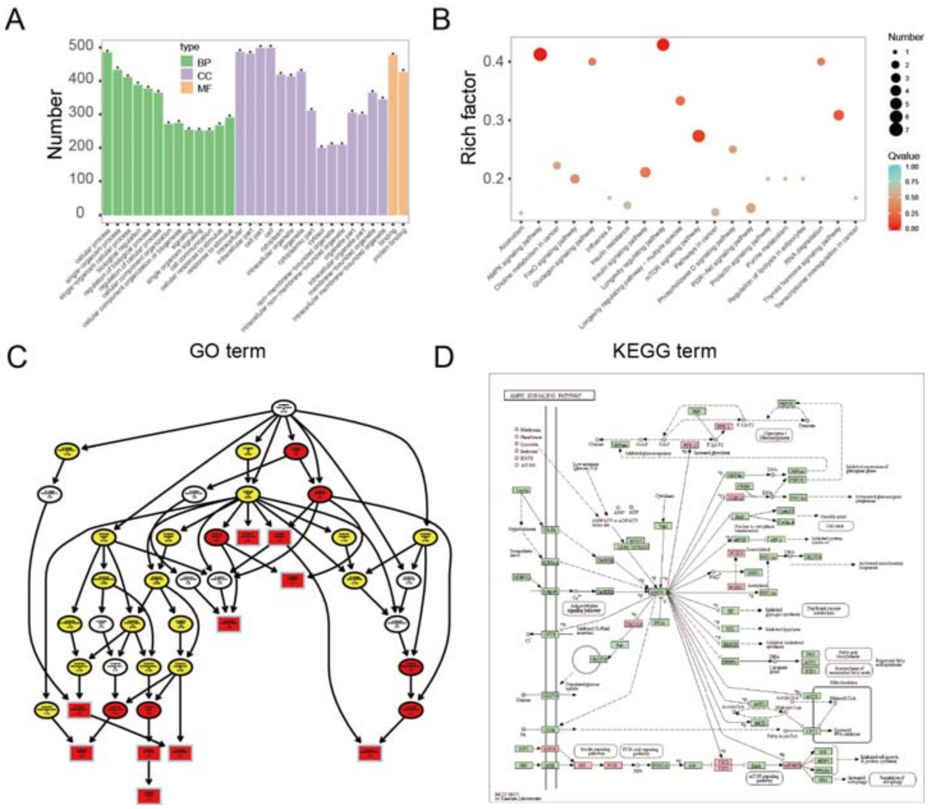
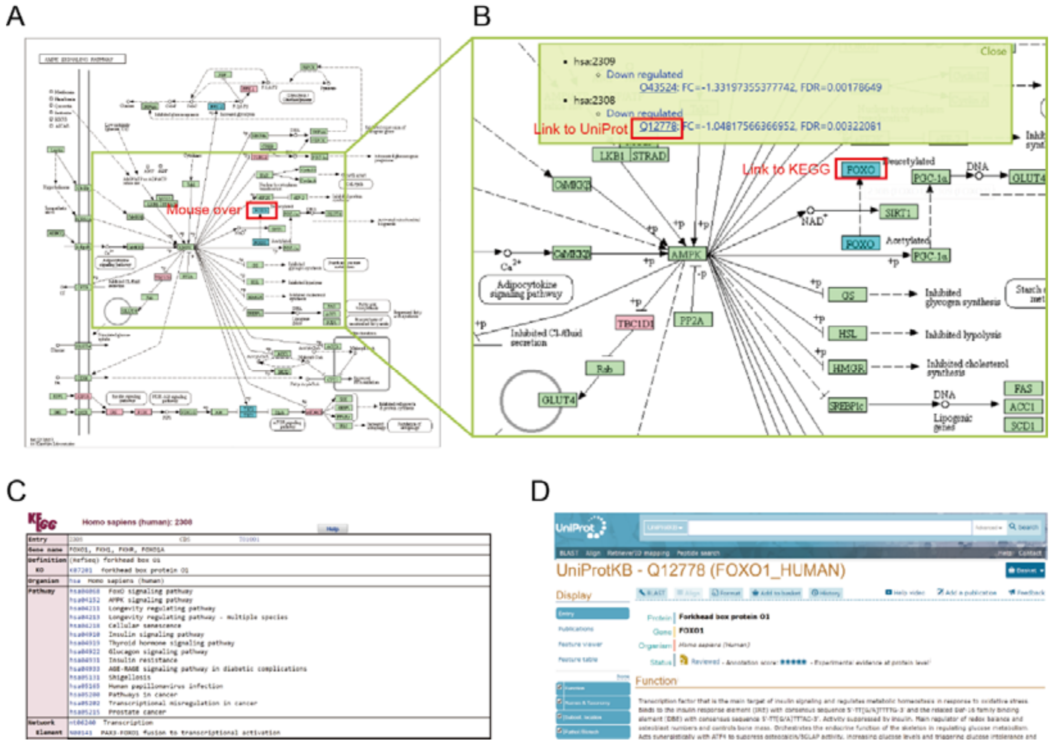


Figure 5. Representative results of functional enrichment analysis. (A) Bar plot of GO enrichment. Top 30 GO terms are ploted based on three categories: BP (biological process), CC (cellular component) and MF (molecular function). An asterisk indicates statistical significance. (B) Scatter plot of KEGG 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 degrees of enrichment are shown in red, yellow and white, 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 in pink.



- hsa:2309
  - Down regulated
    O43524: FC=-1.33197355377742, FDR=0.00178649
- hsa:2308
  - Down regulated
    Q12778: FC=-1.04817566366952, FDR=0.00322061

KEGG Homo sapiens (human): 2308

Help

Entry	2308 CDS T01001	
Gene name	FOXO1, FKH1, FKHR, FOXO1A	
Definition	(RefSeq) forkhead box O1	
KO	K07201	forkhead box protein O1
Organism	hsa	Homo sapiens (human)
Pathway	hsa04068	FoxO signaling pathway
	hsa04152	AMPK signaling pathway
	hsa04211	Longevity regulating pathway
	hsa04213	Longevity regulating pathway - multiple species
	hsa04218	Cellular senescence
	hsa04910	Insulin signaling pathway
	hsa04919	Thyroid hormone signaling pathway
	hsa04922	Glucagon signaling pathway
	hsa04931	Insulin resistance
	hsa04933	AGE-RAGE signaling pathway in diabetic complications
	hsa05131	Shigellosis
	hsa05165	Human papillomavirus infection
	hsa05200	Pathways in cancer
	hsa05202	Transcriptional misregulation in cancer
	hsa05215	Prostate cancer
Network	nt06240	Transcription
Element	N00141	PAX3-FOXO1 fusion to transcriptional activation

UniProt

UniProtKB ▾ Advanced ▾ Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - Q12778 (FOXO1_HUMAN)

Basket ▾

Display

BLAST Align Format Add to basket History Help video Add a publication Feedback

Entry

Publications

Feature viewer

Feature table

Protein Forkhead box protein O1

Gene FOXO1

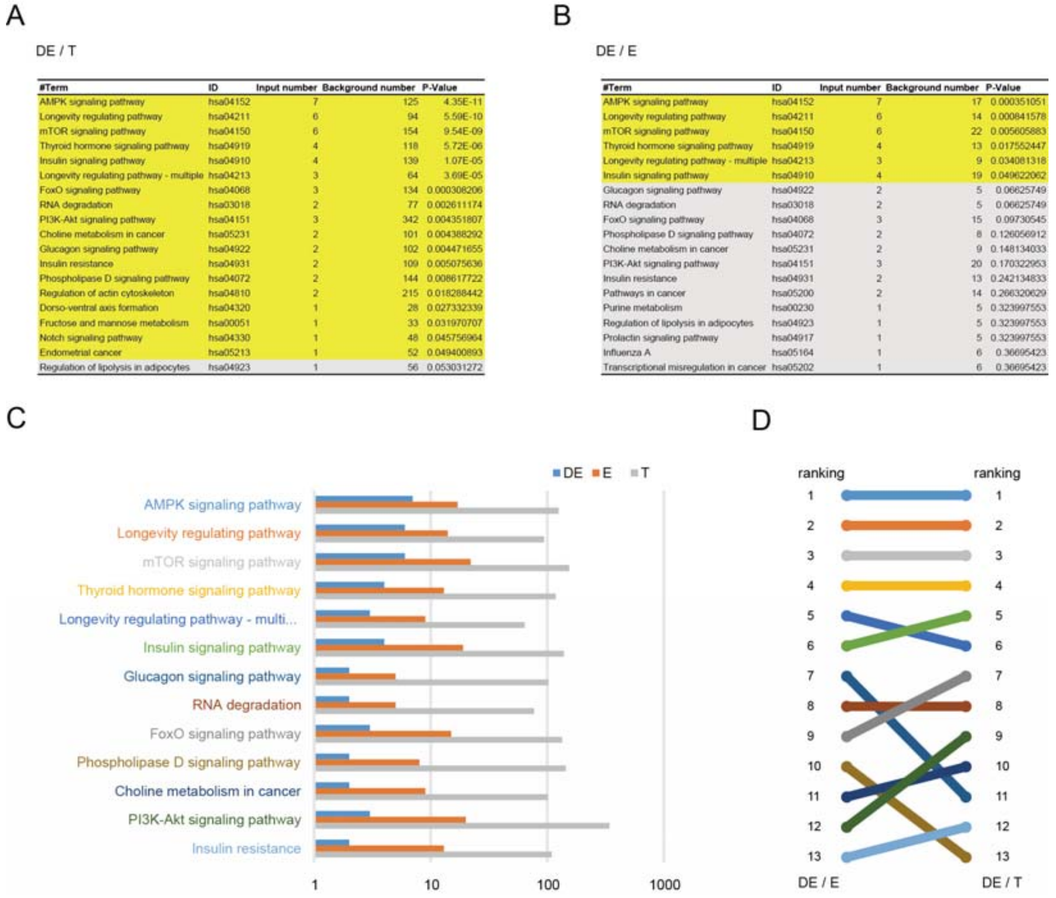
Organism *Homo sapiens (Human)*

Status Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Function

Transcription factor that is the main target of insulin signaling and regulates metabolic homeostasis in response to oxidative stress. Binds to the insulin response element (IRE) with consensus sequence 5'-TT[G/A]TTTTG-3' and the related Daf-16 family binding element (DBE) with consensus sequence 5'-TT[G/A]TTTAC-3'. Activity suppressed by insulin. Main regulator of redox balance and osteoblast numbers and controls bone mass. Orchestrates the endocrine function of the skeleton in regulating glucose metabolism. Acts synergistically with ATF4 to suppress osteocalcin/BGLAP activity, increasing glucose levels and triggering glucose intolerance and

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



A

DE / T

#Term	ID	Input number	Background number	P-Value
AMPK signaling pathway	hsa04152	7	125	4.35E-11
Longevity regulating pathway	hsa04211	6	94	5.59E-10
mTOR signaling pathway	hsa04150	6	154	9.54E-09
Thyroid hormone signaling pathway	hsa04919	4	118	5.72E-06
Insulin signaling pathway	hsa04910	4	139	1.07E-05
Longevity regulating pathway - multiple	hsa04213	3	64	3.69E-05
FoxO signaling pathway	hsa04068	3	134	0.000308206
RNA degradation	hsa03018	2	77	0.002611174
PI3K-Akt signaling pathway	hsa04151	3	342	0.004351807
Choline metabolism in cancer	hsa05231	2	101	0.004388292
Glucagon signaling pathway	hsa04922	2	102	0.004471655
Insulin resistance	hsa04931	2	109	0.005075636
Phospholipase D signaling pathway	hsa04072	2	144	0.008617722
Regulation of actin cytoskeleton	hsa04810	2	215	0.018288442
Dorso-ventral axis formation	hsa04320	1	28	0.027332339
Fructose and mannose metabolism	hsa00051	1	33	0.031970707
Notch signaling pathway	hsa04330	1	48	0.045756964
Endometrial cancer	hsa05213	1	52	0.049400893
Regulation of lipolysis in adipocytes	hsa04923	1	56	0.053031272

B

DE / E

#Term	ID	Input number	Background number	P-Value
AMPK signaling pathway	hsa04152	7	17	0.000351051
Longevity regulating pathway	hsa04211	6	14	0.000841578
mTOR signaling pathway	hsa04150	6	22	0.005605883
Thyroid hormone signaling pathway	hsa04919	4	13	0.017552447
Longevity regulating pathway - multiple	hsa04213	3	9	0.034081318
Insulin signaling pathway	hsa04910	4	19	0.049622062
Glucagon signaling pathway	hsa04922	2	5	0.06625749
RNA degradation	hsa03018	2	5	0.06625749
FoxO signaling pathway	hsa04068	3	15	0.09730545
Phospholipase D signaling pathway	hsa04072	2	8	0.126056912
Choline metabolism in cancer	hsa05231	2	9	0.148134033
PI3K-Akt signaling pathway	hsa04151	3	20	0.170322953
Insulin resistance	hsa04931	2	13	0.242134833
Pathways in cancer	hsa05200	2	14	0.266320629
Purine metabolism	hsa00230	1	5	0.323997553
Regulation of lipolysis in adipocytes	hsa04923	1	5	0.323997553
Prolactin signaling pathway	hsa04917	1	5	0.323997553
Influenza A	hsa05164	1	6	0.36695423
Transcriptional misregulation in cancer	hsa05202	1	6	0.36695423

C
DE
E
T
AMPK signaling pathway
Longevity regulating pathway
mTOR signaling pathway
Thyroid hormone signaling pathway
Longevity regulating pathway - multi...
Insulin signaling pathway
Glucagon signaling pathway
RNA degradation
FoxO signaling pathway
Phospholipase D signaling pathway
Choline metabolism in cancer
PI3K-Akt signaling pathway
Insulin resistance
1
10
100
1000
D
ranking
ranking
1
2
3
4
5
6
7
8
9
10
11
12
13
DE / E
DE / T

Figure 7. Comaprison of two enrichment methods. (A) Snapshot of KEGG enrichment results table used total (*T*) proteins in each pathway as backgroud; yellow highlights indicate the statistically significant ($P < 0.05$) terms. (B) Snapshot of KEGG enrichment results table using expressed (*E*) proteins in each pathway as backgroud; yellow highlights indicate the statistically 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 terms using *E* as backgroud (*DE/E*) and using *T* as backgroud (*DE/T*) in the enrichment analysis results.

References

Alexa, A. and Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor Improv* 2009;27.

Altschul, S.F., *et al*. Basic local alignment search tool. *J Mol Biol* 1990;215(3):403-410.

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.

Gentleman, R.C., *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5(10):R80.

Gruning, B., *et al*. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;15(7):475-476.

Jalili, V., *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res* 2020;48(14):8205-8207.

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.

Yu, G., *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16(5):284-287.

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

Zhu, Y., *et al*. DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Mol Cell Proteomics* 2020;19(6):1047-1057.