~Supplementary material~

StatsPro: systematic integration and evaluation of statistical approaches for detecting differential expression in label-free quantitative proteomics

Yin Yang[1,2], Jingqiu Cheng[2], Shisheng Wang[2,3]* and Hao Yang[2,3]*

[1] Department of Clinical Research Management, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, 610041, China

[2] Institutes for Systems Genetics and NHC Key Lab of Transplant Engineering and Immunology, West China Hospital, Sichuan University, Chengdu, 610041, China

[3] Sichuan Provincial Engineering Laboratory of Pathology in Clinical Application, West China Hospital, Sichuan University, Chengdu 610041, China.

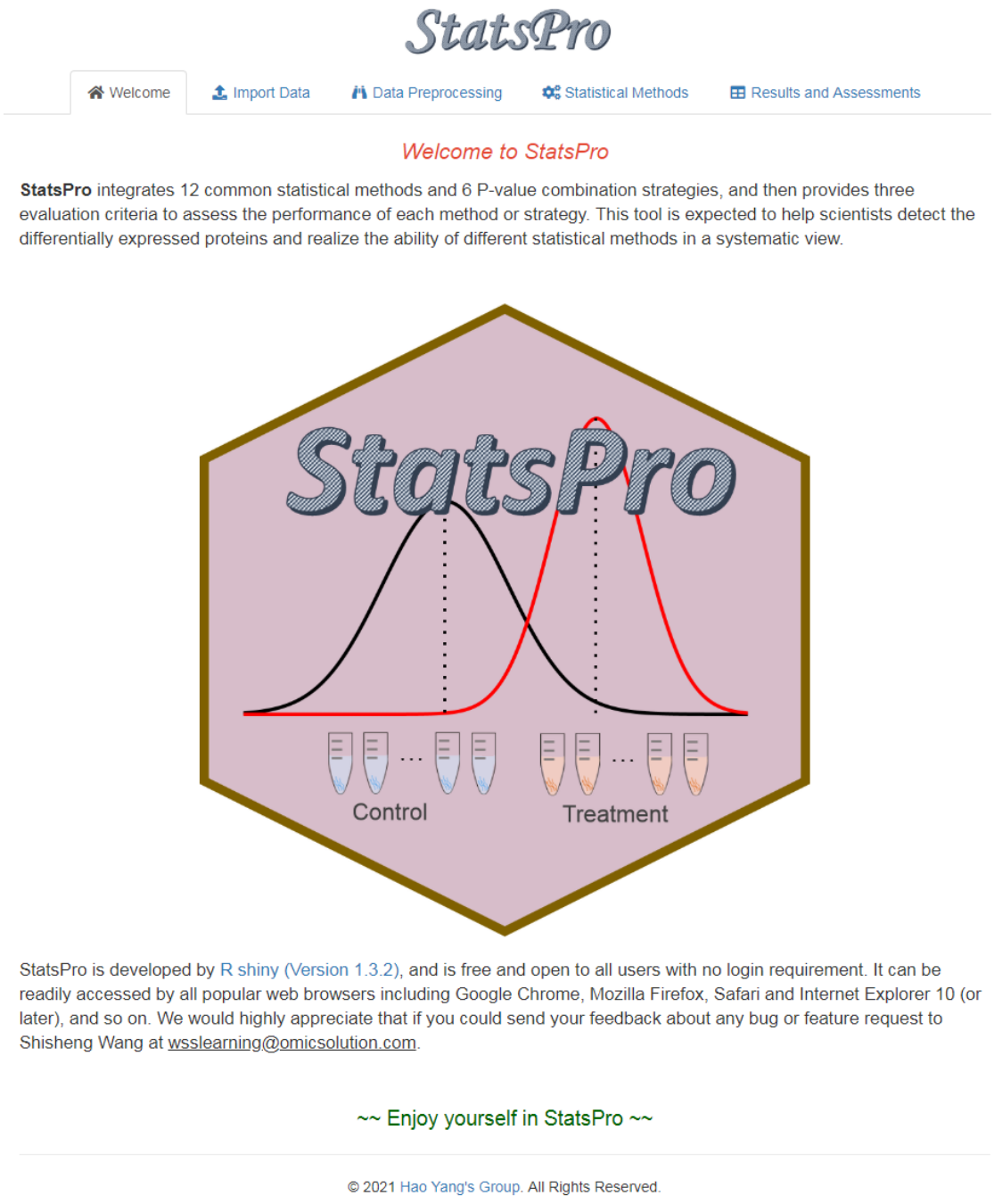
* Corresponding authors. Email: yanghao@scu.edu.cn and shishengwang@wchscu.cn.

Table of Contents

I. Supplementary notes

StatsPro integrates 12 common statistical methods and 6 P-value combination strategies, and then provides three evaluation criteria to assess the performance of each method or strategy (described in Table S1). This tool is expected to help scientists detect the differentially expressed proteins and realize the ability of different statistical methods in a systematic view. Here we present the detailed introduction and operation of *StatsPro*, by which users can follow to analyze their own data freely and conveniently.

Users can visit this site: https://www.omicsolution.com/wukong/StatsPro. Then the website homepage can be shown like this:

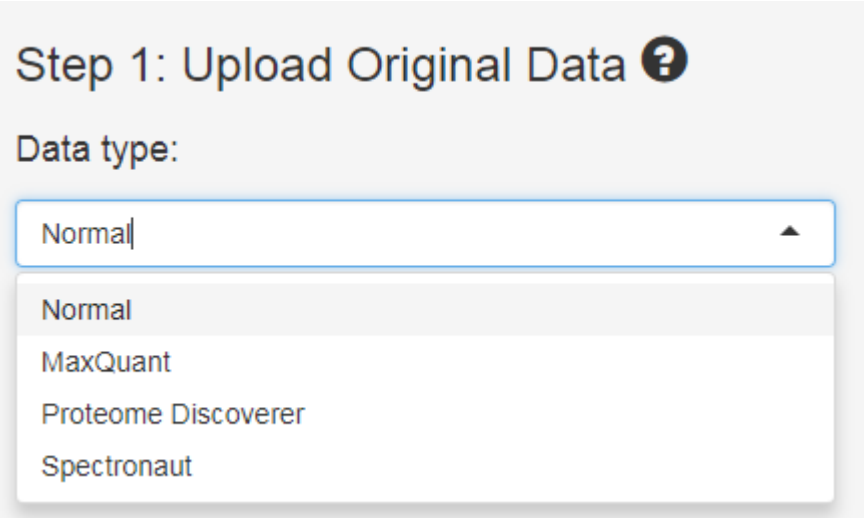


1. Data Preparation

The uploaded data file formats could be .csv, .txt, .xlsx or .xls. Before analysis, users should prepare the proteomics expression data and sample information. The proteomics expression data required here could be readily generated based on results of several popular tools such as MaxQuant[1], Proteome Discoverer (Thermo Fisher Scientific), Spectronaut[2]. Users then can upload the data and type in right sample information into *StatsPro* with right formats respectively and start subsequent analysis.

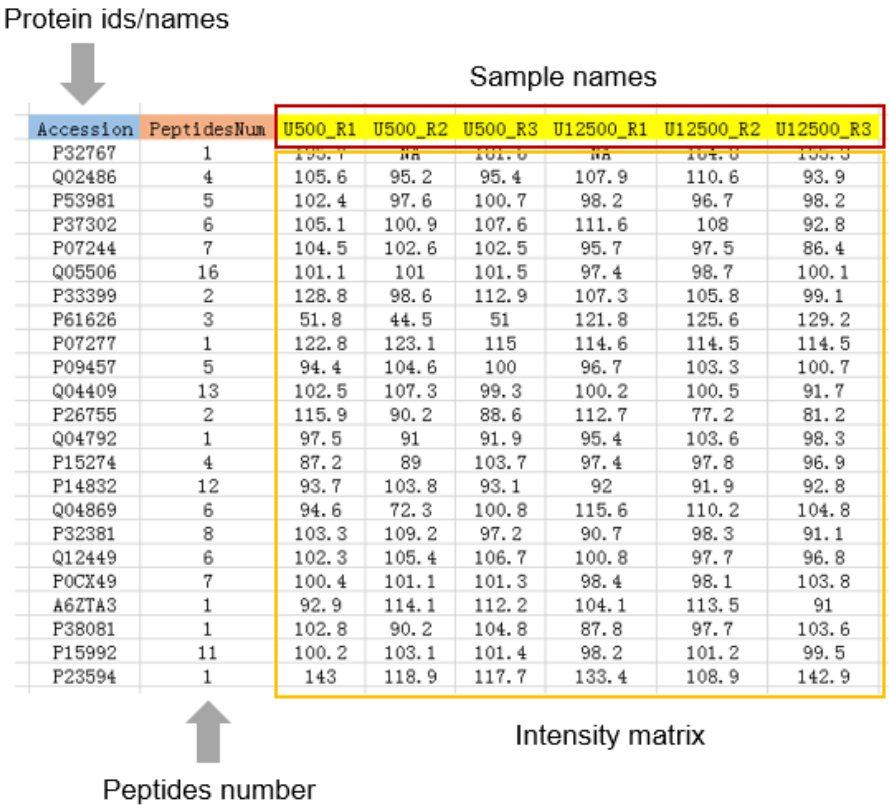
1.1 Expression data

There are currently four types of proteomics expression data supported in *StatsPro* (i.e., 'Normal', 'MaxQuant', 'Proteome Discoverer', 'Spectronaut'). The detailed description about this parameter is shown below:



1.1.1 Normal type

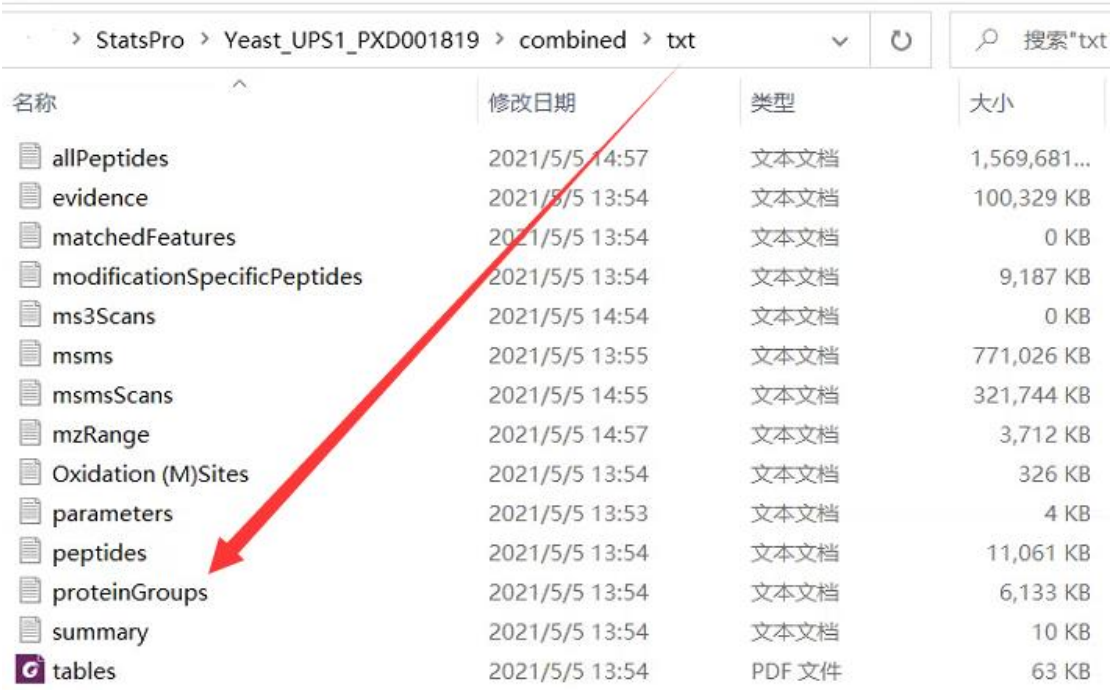
“Normal” here means users can prepare their proteomics expression data as this type from any software (not just MaxQuant, Proteome Discoverer and Spectronaut) and upload into this tool. In the situation, protein ids/names and peptides number are sequentially provided in the first two columns of input file. The protein ids/names in the first column could be UniProt ids or protein names. The peptides number in the second column could be the total counts of peptides associated with the protein. From the third column, proteins expression intensity or signal abundance in every sample should be listed. The data structure is shown as below:



Accession	PeptidesNum	U500_R1	U500_R2	U500_R3	U12500_R1	U12500_R2	U12500_R3
P32767	1	135.7	NA	101.6	NA	104.8	135.3
Q02486	4	105.6	95.2	95.4	107.9	110.6	93.9
P53981	5	102.4	97.6	100.7	98.2	96.7	98.2
P37302	6	105.1	100.9	107.6	111.6	108	92.8
P07244	7	104.5	102.6	102.5	95.7	97.5	86.4
Q05506	16	101.1	101	101.5	97.4	98.7	100.1
P33399	2	128.8	98.6	112.9	107.3	105.8	99.1
P61626	3	51.8	44.5	51	121.8	125.6	129.2
P07277	1	122.8	123.1	115	114.6	114.5	114.5
P09457	5	94.4	104.6	100	96.7	103.3	100.7
Q04409	13	102.5	107.3	99.3	100.2	100.5	91.7
P26755	2	115.9	90.2	88.6	112.7	77.2	81.2
Q04792	1	97.5	91	91.9	95.4	103.6	98.3
P15274	4	87.2	89	103.7	97.4	97.8	96.9
P14832	12	93.7	103.8	93.1	92	91.9	92.8
Q04869	6	94.6	72.3	100.8	115.6	110.2	104.8
P32381	8	103.3	109.2	97.2	90.7	98.3	91.1
Q12449	6	102.3	105.4	106.7	100.8	97.7	96.8
P0CX49	7	100.4	101.1	101.3	98.4	98.1	103.8
A6ZTA3	1	92.9	114.1	112.2	104.1	113.5	91
P38081	1	102.8	90.2	104.8	87.8	97.7	103.6
P15992	11	100.2	103.1	101.4	98.2	101.2	99.5
P23594	1	143	118.9	117.7	133.4	108.9	142.9

1.1.2 Proteomics data from MaxQuant

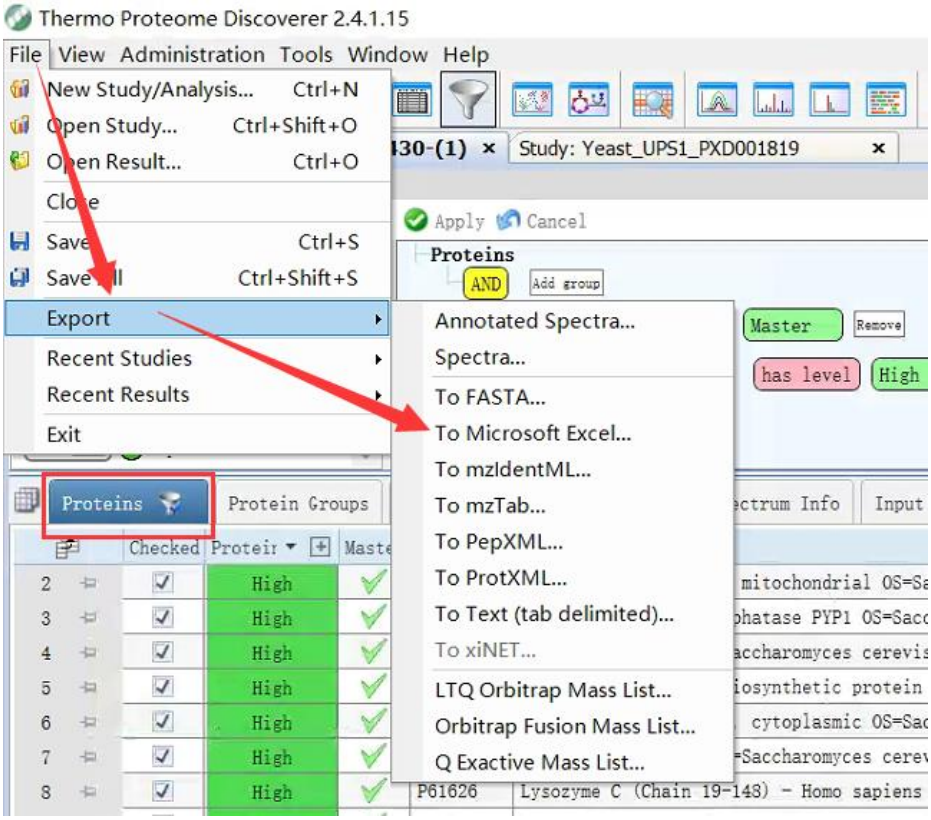
If users choose MaxQuant to process their raw MS data, they could obtain a proteinGroups.txt file in the txt fold and then upload this file into *StatsPro* directly:



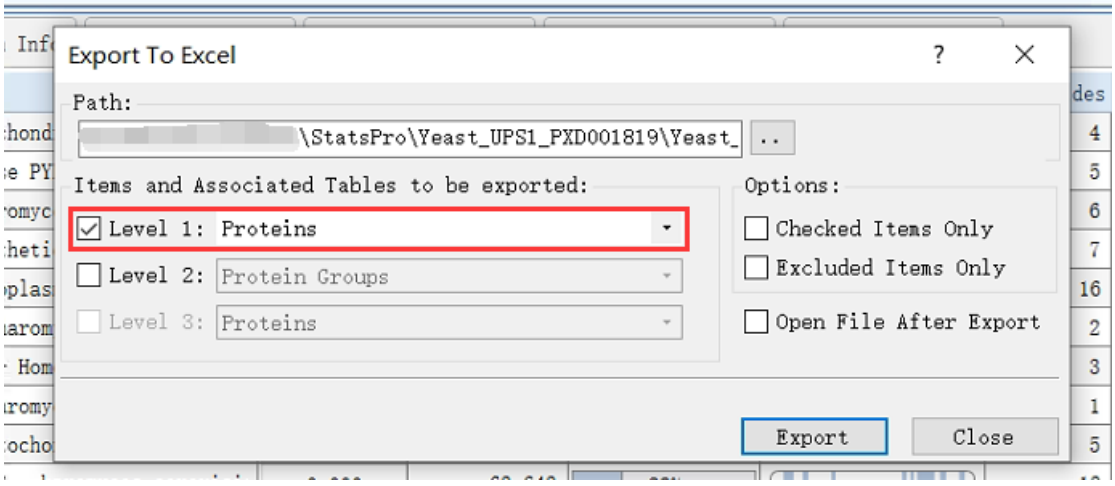
1.1.3 Proteomics data from Proteome Discoverer

If users choose Proteome Discoverer to process their raw MS data, they could export proteomics data as below and then upload this file into *StatsPro* directly:

a. Select the “Proteins” table and click “File—Export—To Microsoft Excel…”, as below:



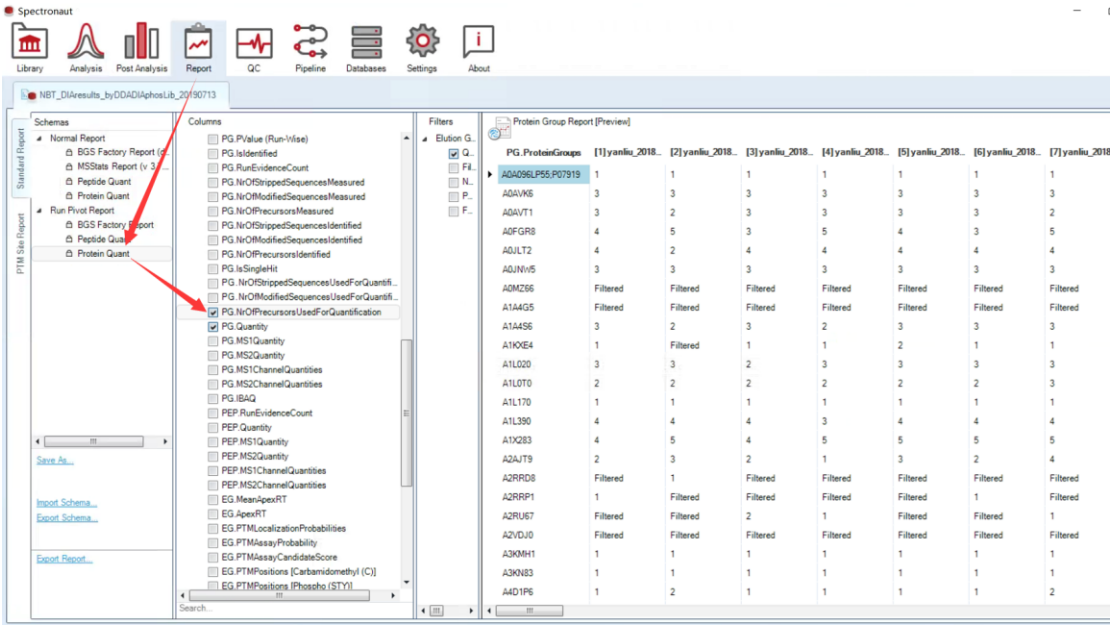
b. Export data to an excel file (.xlsx format) and then upload this file to *StatsPro* directly:



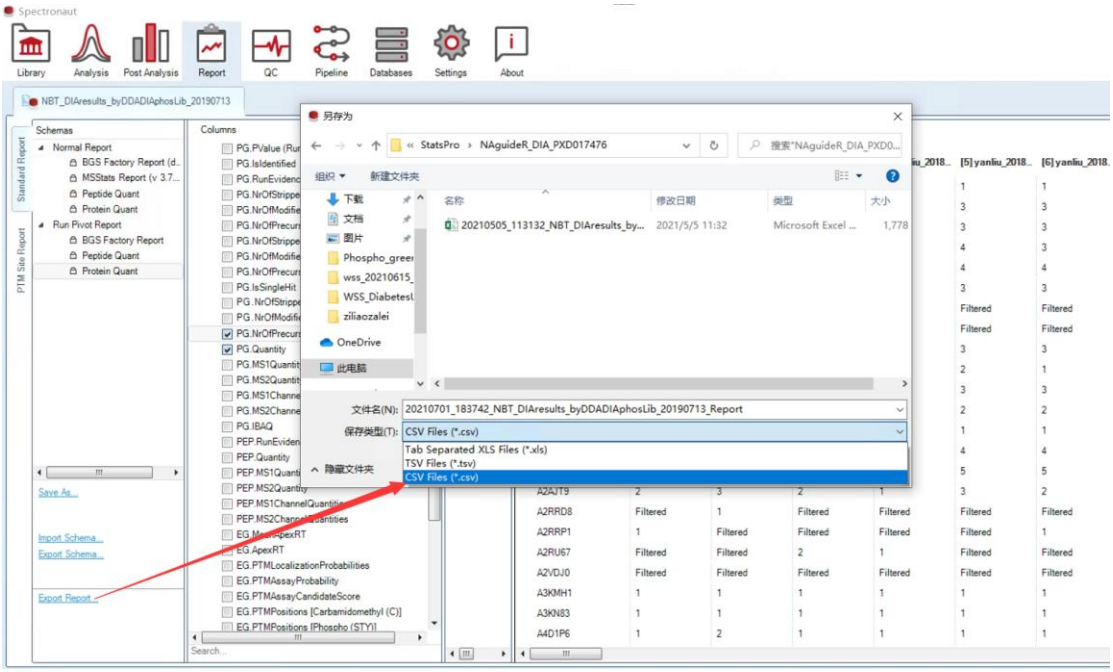
1.1.4 Proteomics data from Spectronaut

If users choose Spectronaut to process their raw MS data, they could export proteomics data as below and then upload this file into *StatsPro* directly:

a. Select the “Report” button and click “Protein Quant” in the “Run Pivot Report”. Besides the default options, users should also select the “PG.NrOfPrecursorsUsedForQuantification”, as below:

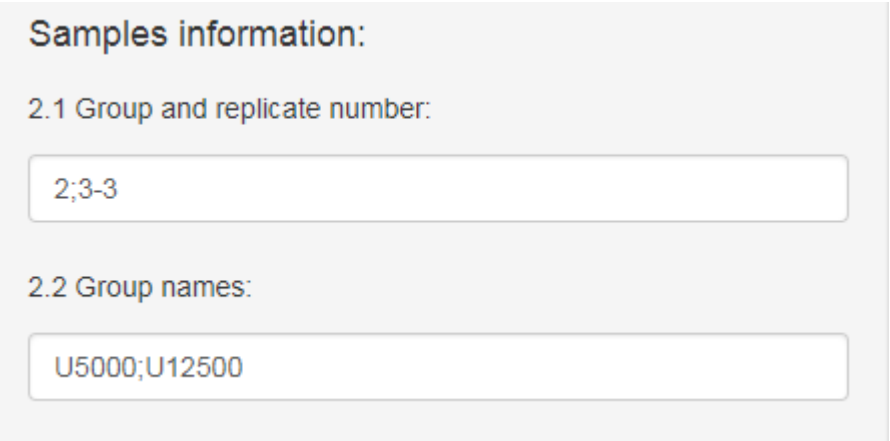


b. Then click “Export Report…” and save the data in a .csv file:



1.2 Samples information data

Sample information here means that users should provide group number, replicate number, and group names, then type them in *StatsPro* (see below). This tool will use these information to calculate corresponding results and enable, for example, filtration strategy for different group respectively in a later step:

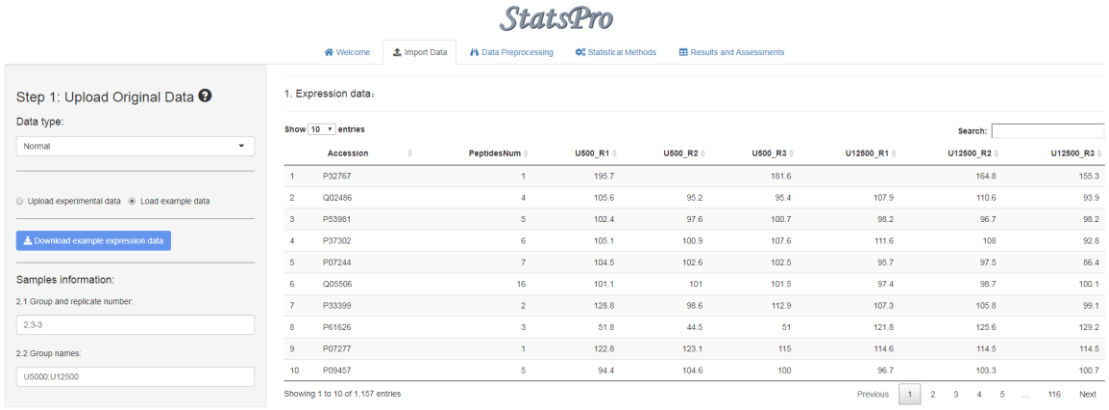


2.1 Group and replicate number: Type in the group number and replicate number here. Please note, the group number and replicate number are linked with ";", and the replicate number of each group is linked with "-". For example, if you have two groups, each group has three replicates, then you should type in "2;3-3" here. Similarly, if you have 3 groups with 5 replicates in every groups, you should type in "3;5-5-5".

2.2 Group names: Type in the group names of your samples. Please note, the group names are linked with ";". For example, there are two groups in the spiked data (UPS1+Yeast), you can type in "U5000;U12500".

1.3 Download example datasets

If users want to download the example datasets to their own computer and check the data format locally, they can download them from here:



First, select “Load example data” and the example data will be shown on the right panel interactively. Users can visually observe what the data looks like.

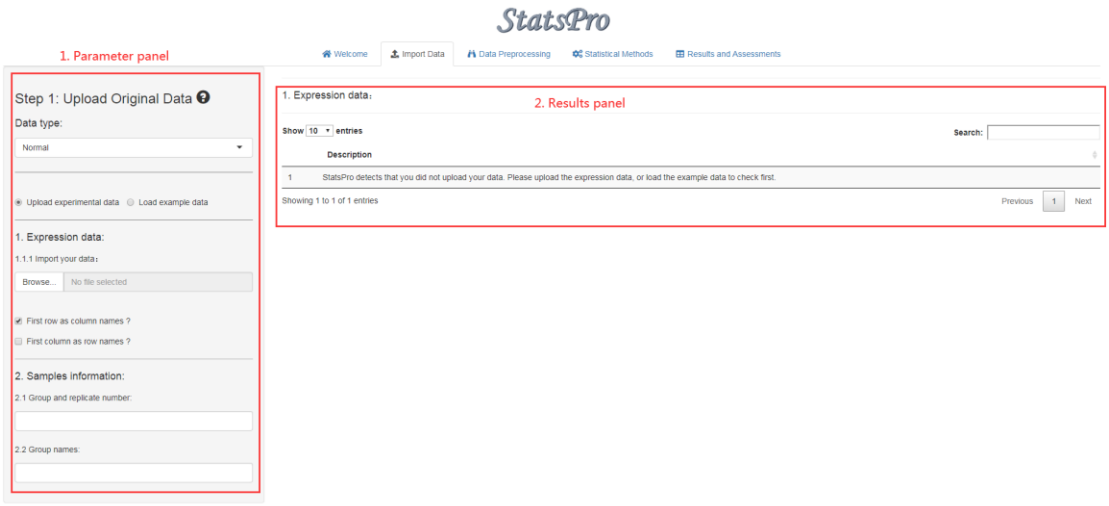
Second, users can download the example data (proteomics expression data) by clicking the “Download example expression data” button. The data are saved as corresponding format and users can open them in other software, such as Excel.

Third, users can check the example sample information and understand these parameters better.

2. Import Data

This is the first step, in which users should upload data here or load the example data with the above data formats. By default, we use the example data to show result of every step.

2.1 Uploading data. When users prepare their data (expression and sample information data set), they can upload these data from here:



There are two main panels: first, *parameters panel*, users can adjust parameters here; second, *results panel*, many results after users set the parameters will be shown here and users can also download these results.

In the *parameters panel* of “Import Data”, there are two choices for users:

a. Load experimental data. When users choose this option, they can upload their own data here. Users should select the right format based on their data and then click “Browse” button to import the data;

First row as column names: this means whether the first row is column names. If true, you should choose this parameter.

First column as row names: this means whether the first column is row names. If true, you should choose this parameter.

b. Samples information. As described in part 1.2, users can choose this option and download the example data to check them locally.

In the *results panel* of “Import Data”, if users don’t upload their data, here will show “*StatsPro* detects that you did not upload your data. Please upload the expression data, or load the example data to check first” to warn users.

Before uploading expression data, users should also recognize which type their data belongs to and choose the right parameter by adjusting the “Data type”. The instruction of every data type can be found above (*Data Preparation* part).

Step 1: Upload Original Data

Data type:

Normal

- Normal
- MaxQuant
- Proteome Discoverer
- Spectronaut

1. Expression data:

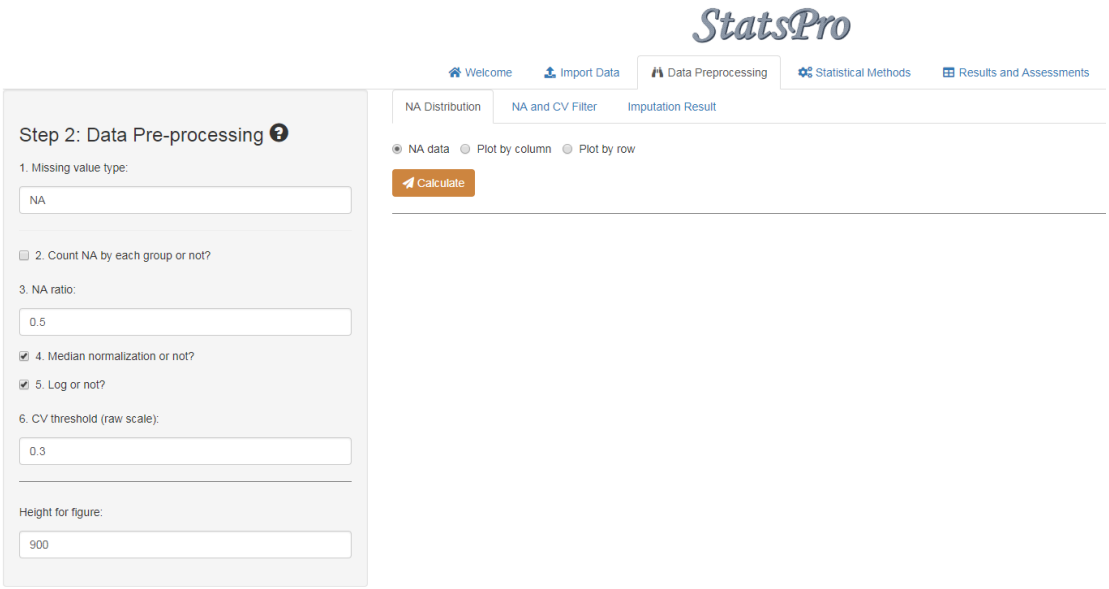
1.1.1 Import your data：

Browse... No file selected

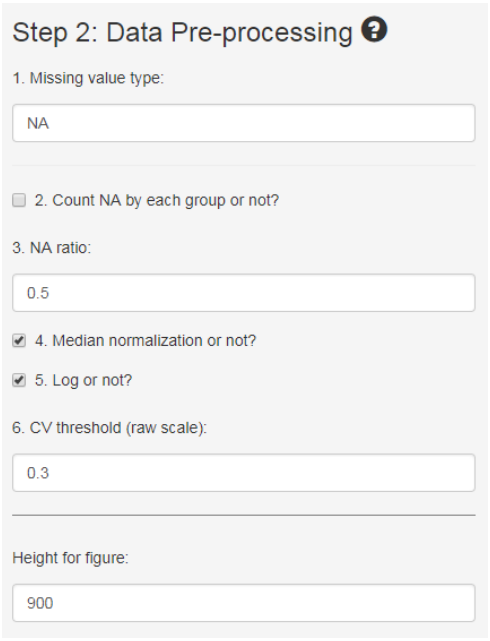
- [x] First row as column names ?
- [ ] First column as row names ?

3. Data Pre-processing

Users can pre-process their data in this step, including data filtration (i.e. removing those proteins with high proportion of missing values (NAs) and large coefficient of variation (CVs)), normalization (i.e. normalizing protein intensities using median value of each sample), missing value imputation (i.e. all missing values are derived with the k-Nearest Neighbor method[3]).



3.1 Parameters



1. Missing value type: what the missing values look like in the expression data, for example, Spectronaut [2,4] software usually export “Filtered” as missing values, so users should change this parameter to “Filtered” if their data contain “Filtered”. *StatsPro* will recognize these characters and replace them with NAs. Any other characters indicating a missing value can be similarly defined.

2. Count NA by each group or not: if true, *StatsPro* will count the number of missing values in each

group and calculate the NA ratio. Otherwise, it calculates the NA ratio across all groups, for example, as below:

Peptides	Charges	Uniprot_IDs	Phos_Cyc_1	Phos_Cyc_2	Phos_Cyc_3	Phos_Cyc_4	Phos_Cyc_5	Phos_Cyc_6	Phos_Cyc_7	Phos_Cyc_8	Phos_Cyc_9	Phos_Cyc_10	Phos_Noc_o.1	Phos_Noc_o.2	Phos_Noc_o.3	Phos_Noc_o.4	Phos_Noc_o.5	Phos_Noc_o.6	Phos_Noc_o.7	Phos_Noc_o.8	Phos_Noc_o.9	Phos_Noc_o.10
N[Acetyl (Protein N-ter	2	A1KXE4	79679.09	NA	110380.5	130927.4	82461.96	155724.4	113495.3	136404.3	86171.31	98299.7	NA	NA	151027.6	NA	210179.9	182829.7	151424.3	NA	NA	181321.2

There are 2 groups (10 biological replicates in each group) here, if users select this parameter, *StatsPro* will calculate 2 NA ratios for this peptide (first group: 1/10=0.1, second group: 5/10=0.5), otherwise, only one NA ratio: 6/20=0.3.

3. NA ratio: the threshold of NA ratio. Those peptides/proteins with NA ratio above this threshold will be removed.

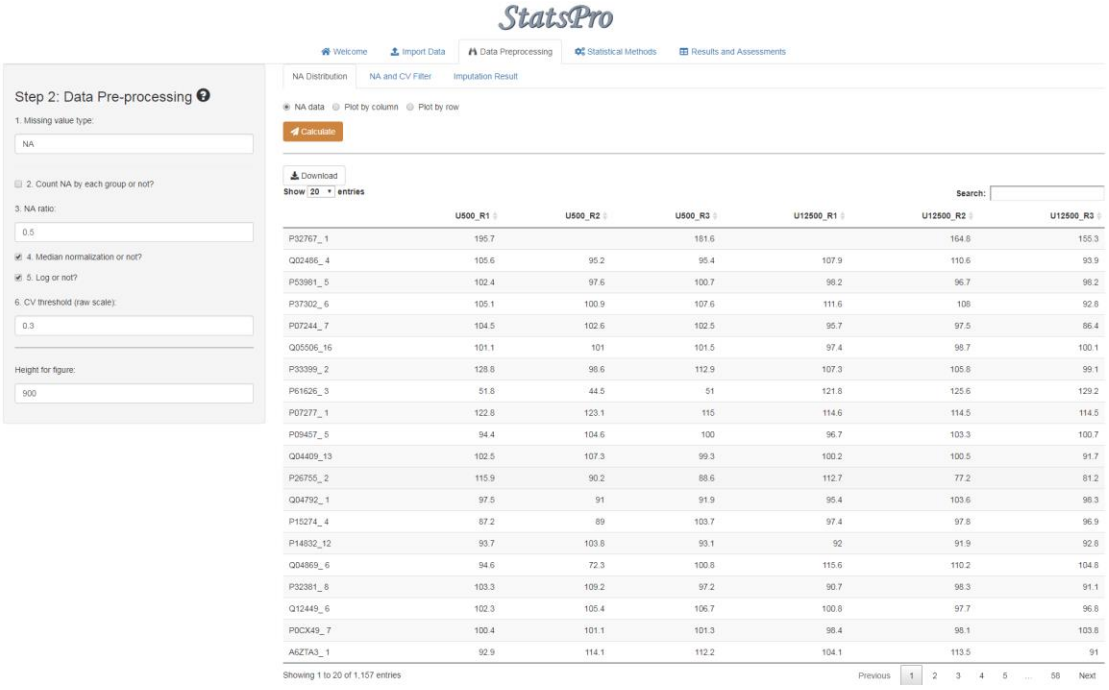
4. Median normalization or not: if true, *StatsPro* will process median normalization for original data. (Note, *StatsPro* was not designed to perform sophisticated normalization analysis. Any normalized datasets with NA can be accepted for analysis).

5. Log or not: if true, the data will be transformed to the logarithmic scale with base 2.

6. CV threshold (raw scale): the threshold of coefficient of variation. Those peptides/proteins with CV above this threshold will be removed. “raw scale” here means the CV of each peptide/protein is calculate using the data before logarithm transformation.

7. Height for figure: users can adjust the height of figures by changing this parameter.

If users set these parameters well, then click “calculate” button, the results will appear on the right panel.



3.2 Results of Data Pre-processing

a. NA Distribution. This part contains three sub-parts:

a.1 NA data. Here shows the result where the “Missing value type” defined by “NA” will be shown with a blank cell and users can click “Download” button to download this result to their own computer:

Download

Show 20 entries

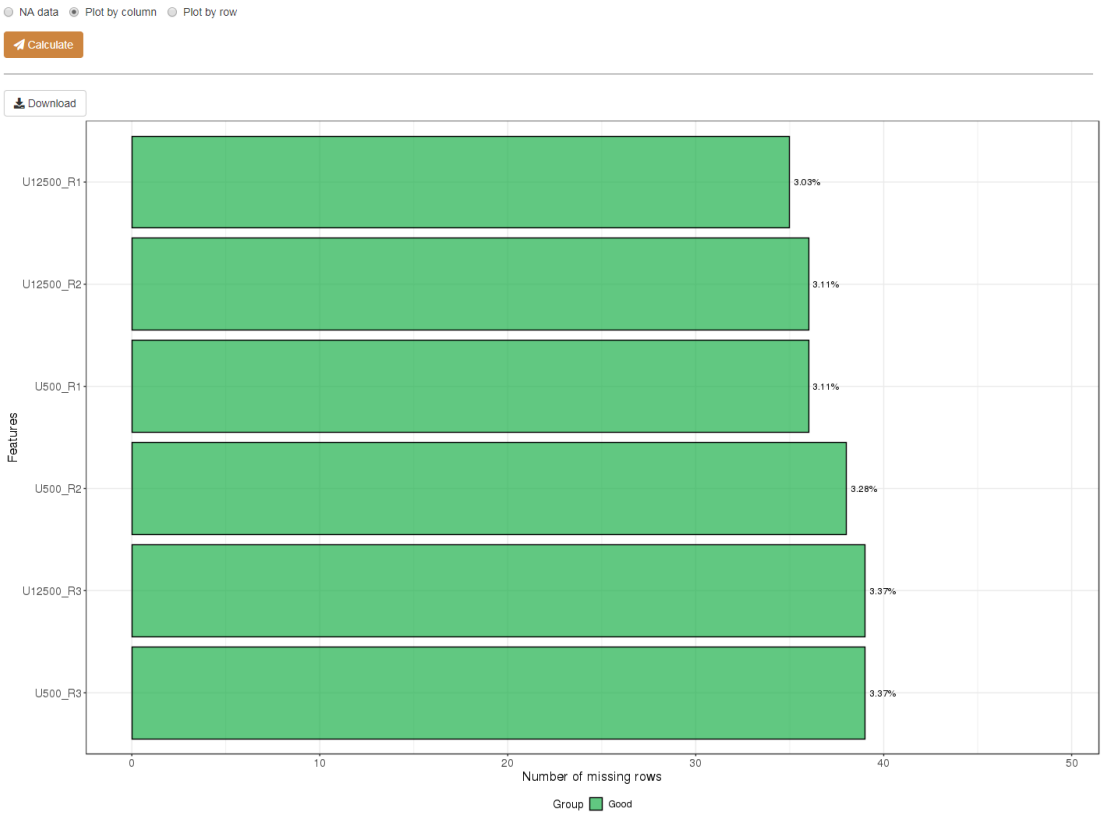
Search:

	U500_R1	U500_R2	U500_R3	U12500_R1	U12500_R2	U12500_R3
P32767_1	195.7		181.6		164.8	155.3
Q02486_4	105.6	95.2	95.4	107.9	110.6	93.9
P53981_5	102.4	97.6	100.7	98.2	96.7	98.2
P37302_6	105.1	100.9	107.6	111.6	108	92.8
P07244_7	104.5	102.6	102.5	95.7	97.5	86.4
Q05506_16	101.1	101	101.5	97.4	98.7	100.1
P33399_2	128.8	98.6	112.9	107.3	105.8	99.1
P61626_3	51.8	44.5	51	121.8	125.6	129.2
P07277_1	122.8	123.1	115	114.6	114.5	114.5
P09457_5	94.4	104.6	100	96.7	103.3	100.7
Q04409_13	102.5	107.3	99.3	100.2	100.5	91.7
P26755_2	115.9	90.2	88.6	112.7	77.2	81.2
Q04792_1	97.5	91	91.9	95.4	103.6	98.3
P15274_4	87.2	89	103.7	97.4	97.8	96.9
P14832_12	93.7	103.8	93.1	92	91.9	92.8
Q04869_6	94.6	72.3	100.8	115.6	110.2	104.8
P32381_8	103.3	109.2	97.2	90.7	98.3	91.1
Q12449_6	102.3	105.4	106.7	100.8	97.7	96.8
P0CX49_7	100.4	101.1	101.3	98.4	98.1	103.8
A6ZTA3_1	92.9	114.1	112.2	104.1	113.5	91

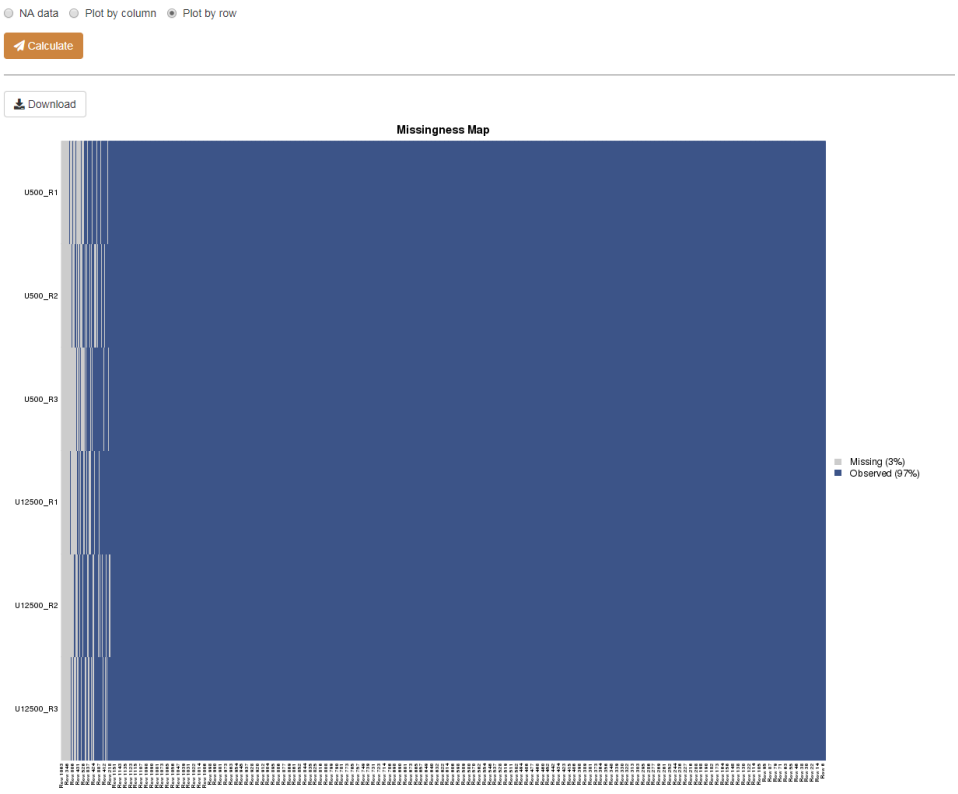
Showing 1 to 20 of 1,157 entries

Previous 1 2 3 4 5 … 58 Next

a.2 Plot by column. Here shows the result of the NA distribution of every sample.



a.2 Plot by row. Here shows the result of the NA distribution of every peptide/protein.



b. NA filter. This part will show the filtered result. That means, on the basis of the preset parameters (i.e. NA ratio, CV threshold), those proteins without meeting these requirements would be removed.

NA Distribution | NA and CV Filter | Imputation Result

Calculate

Download

Show 20 entries

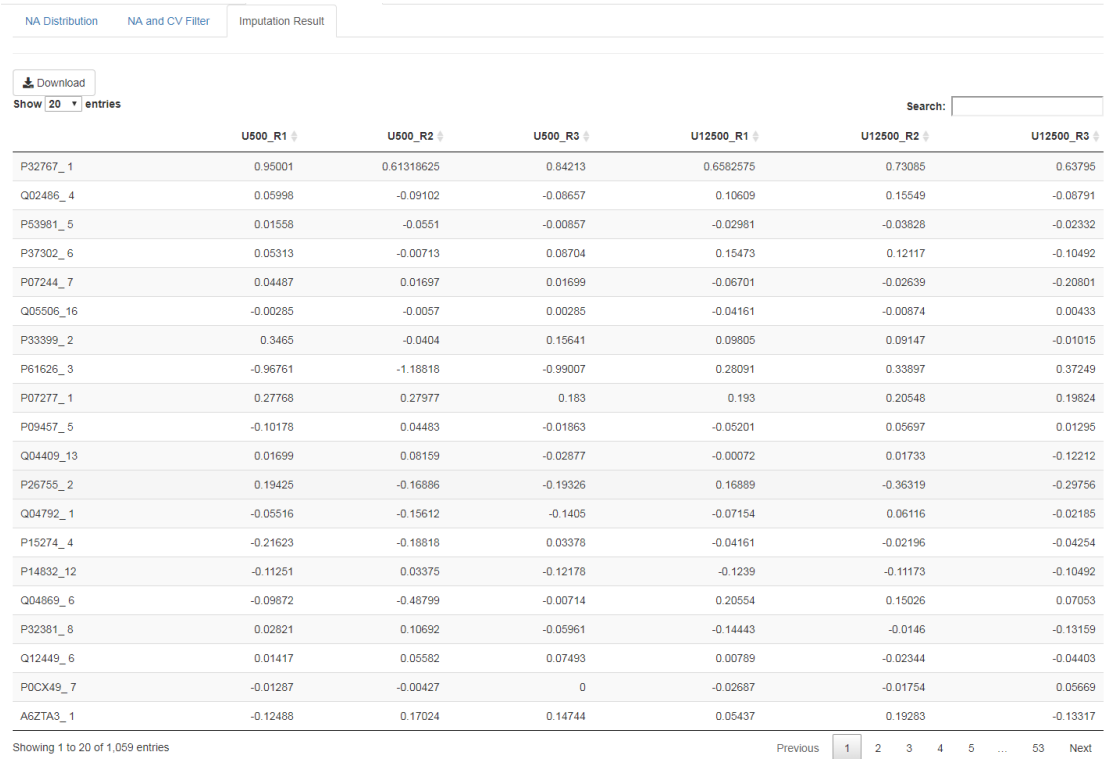
Search:

	U500_R1	U500_R2	U500_R3	U12500_R1	U12500_R2	U12500_R3
P32767_1	0.95001		0.84213		0.73085	0.63795
Q02486_4	0.05998	-0.09102	-0.08657	0.10609	0.15549	-0.08791
P53981_5	0.01558	-0.0551	-0.00857	-0.02981	-0.03828	-0.02332
P37302_6	0.05313	-0.00713	0.08704	0.15473	0.12117	-0.10492
P07244_7	0.04487	0.01697	0.01699	-0.06701	-0.02639	-0.20801
Q05506_16	-0.00285	-0.0057	0.00285	-0.04161	-0.00874	0.00433
P33399_2	0.3465	-0.0404	0.15641	0.09805	0.09147	-0.01015
P61626_3	-0.96761	-1.18818	-0.99007	0.28091	0.33897	0.37249
P07277_1	0.27768	0.27977	0.183	0.193	0.20548	0.19824
P09457_5	-0.10178	0.04483	-0.01863	-0.05201	0.05697	0.01295
Q04409_13	0.01699	0.08159	-0.02877	-0.00072	0.01733	-0.12212
P26755_2	0.19425	-0.16886	-0.19326	0.16889	-0.36319	-0.29756
Q04792_1	-0.05516	-0.15612	-0.1405	-0.07154	0.06116	-0.02185
P15274_4	-0.21623	-0.18818	0.03378	-0.04161	-0.02196	-0.04254
P14832_12	-0.11251	0.03375	-0.12178	-0.1239	-0.11173	-0.10492
Q04869_6	-0.09872	-0.48799	-0.00714	0.20554	0.15026	0.07053
P32381_8	0.02821	0.10692	-0.05961	-0.14443	-0.0146	-0.13159
Q12449_6	0.01417	0.05582	0.07493	0.00789	-0.02344	-0.04403
P0CX49_7	-0.01287	-0.00427	0	-0.02687	-0.01754	0.05669
A6ZTA3_1	-0.12488	0.17024	0.14744	0.05437	0.19283	-0.13317

Showing 1 to 20 of 1,059 entries

Previous 1 2 3 4 5 … 53 Next

c. Imputation Result. This part will derive the missing values with the k-Nearest Neighbor method. Users can check how to process missing value problem detailedly in our previous published article [5]. The results are shown as below and users can click “Download” button to save this result in a .csv file:


NA Distribution | NA and CV Filter | Imputation Result

Download

Show 20 entries

Search:

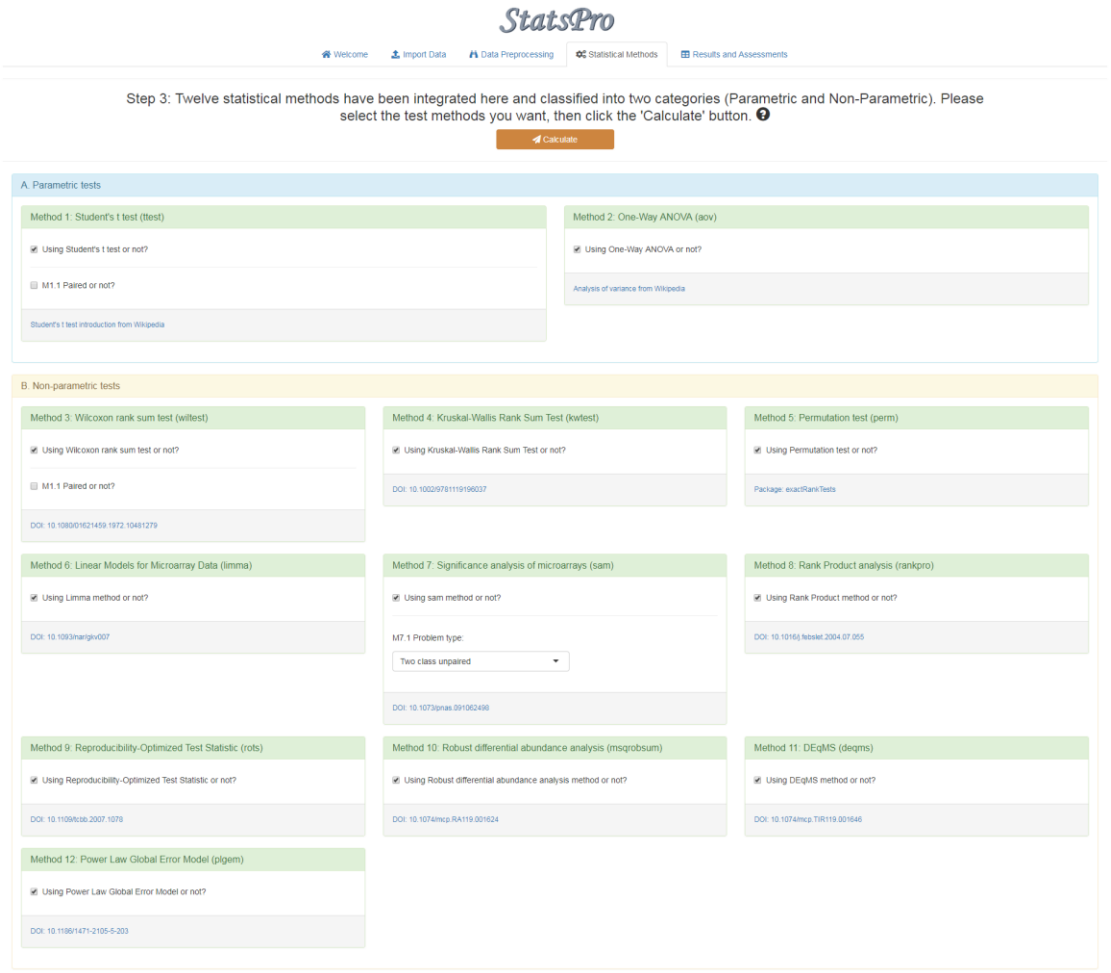
	U500_R1	U500_R2	U500_R3	U12500_R1	U12500_R2	U12500_R3
P32767_1	0.95001	0.61318625	0.84213	0.6582575	0.73085	0.63795
Q02486_4	0.05998	-0.09102	-0.08657	0.10609	0.15549	-0.08791
P53981_5	0.01558	-0.0551	-0.00857	-0.02981	-0.03828	-0.02332
P37302_6	0.05313	-0.00713	0.08704	0.15473	0.12117	-0.10492
P07244_7	0.04487	0.01697	0.01699	-0.06701	-0.02639	-0.20801
Q05506_16	-0.00285	-0.0057	0.00285	-0.04161	-0.00874	0.00433
P33399_2	0.3465	-0.0404	0.15641	0.09805	0.09147	-0.01015
P61626_3	-0.96761	-1.18818	-0.99007	0.28091	0.33897	0.37249
P07277_1	0.27768	0.27977	0.183	0.193	0.20548	0.19824
P09457_5	-0.10178	0.04483	-0.01863	-0.05201	0.05697	0.01295
Q04409_13	0.01699	0.08159	-0.02877	-0.00072	0.01733	-0.12212
P26755_2	0.19425	-0.16886	-0.19326	0.16889	-0.36319	-0.29756
Q04792_1	-0.05516	-0.15612	-0.1405	-0.07154	0.06116	-0.02185
P15274_4	-0.21623	-0.18818	0.03378	-0.04161	-0.02196	-0.04254
P14832_12	-0.11251	0.03375	-0.12178	-0.1239	-0.11173	-0.10492
Q04869_6	-0.09872	-0.48799	-0.00714	0.20554	0.15026	0.07053
P32381_8	0.02821	0.10692	-0.05961	-0.14443	-0.0146	-0.13159
Q12449_6	0.01417	0.05582	0.07493	0.00789	-0.02344	-0.04403
P0CX49_7	-0.01287	-0.00427	0	-0.02687	-0.01754	0.05669
A6ZTA3_1	-0.12488	0.17024	0.14744	0.05437	0.19283	-0.13317

Showing 1 to 20 of 1,059 entries

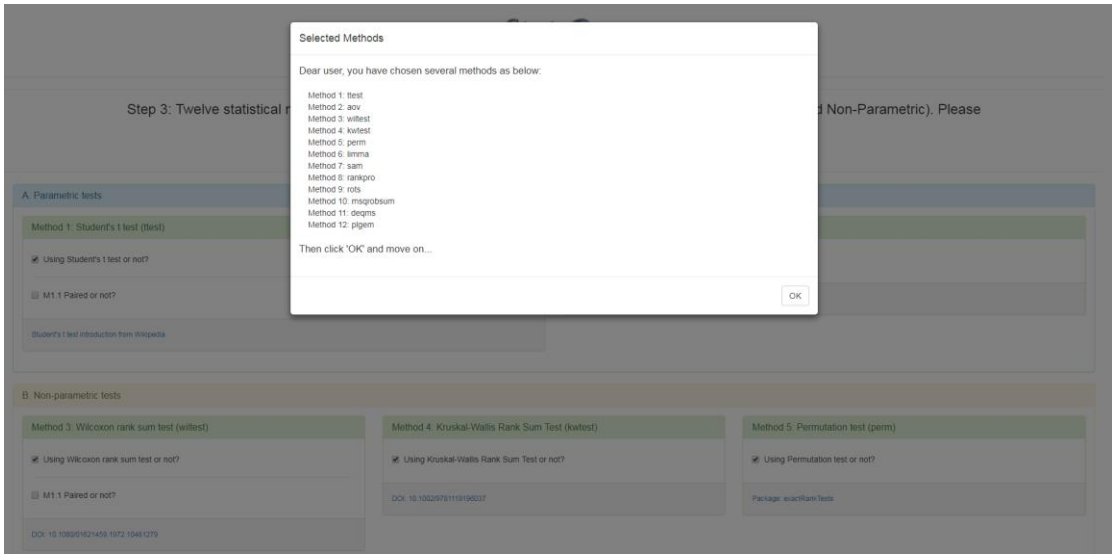
Previous 1 2 3 4 5 … 53 Next

4. Methods

Here, 12 published statistical methods for the DEP detection were integrated in *StatsPro* and can be broadly divided into two types: 1. parametric tests, including t test [6], one-way ANOVA [7]; 2. non-parametric tests, including wilcoxon rank sum test [8], kruskal-wallis rank sum test [9], permutation test [10], limma [11], SAM [12], RP [13], ROTS [14], MSqRobSum [15], DEqMS [16], PLGEM [17]. Therefore, in this step, users can select any of these statistical methods that are currently supported. All methods have been chosen by default. The detailed information about each method can be found in Table S1. In addition, we also provide the reference for every method just blow each option on the web:

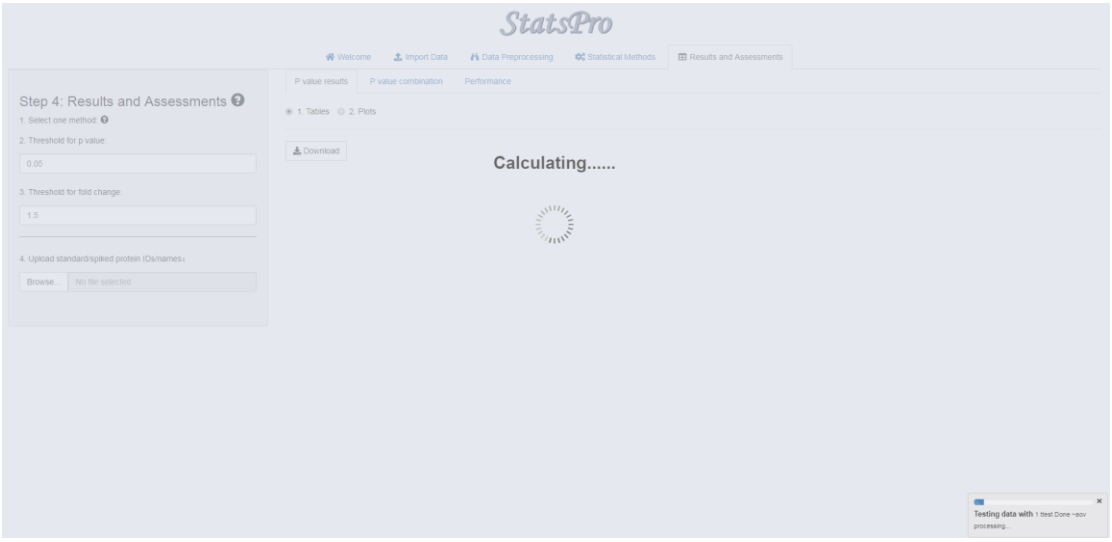


After selecting suitable methods, users need to click 'Calculate' button, and a popup window will be jumped out to show the selected methods, then click 'OK' button and continue:

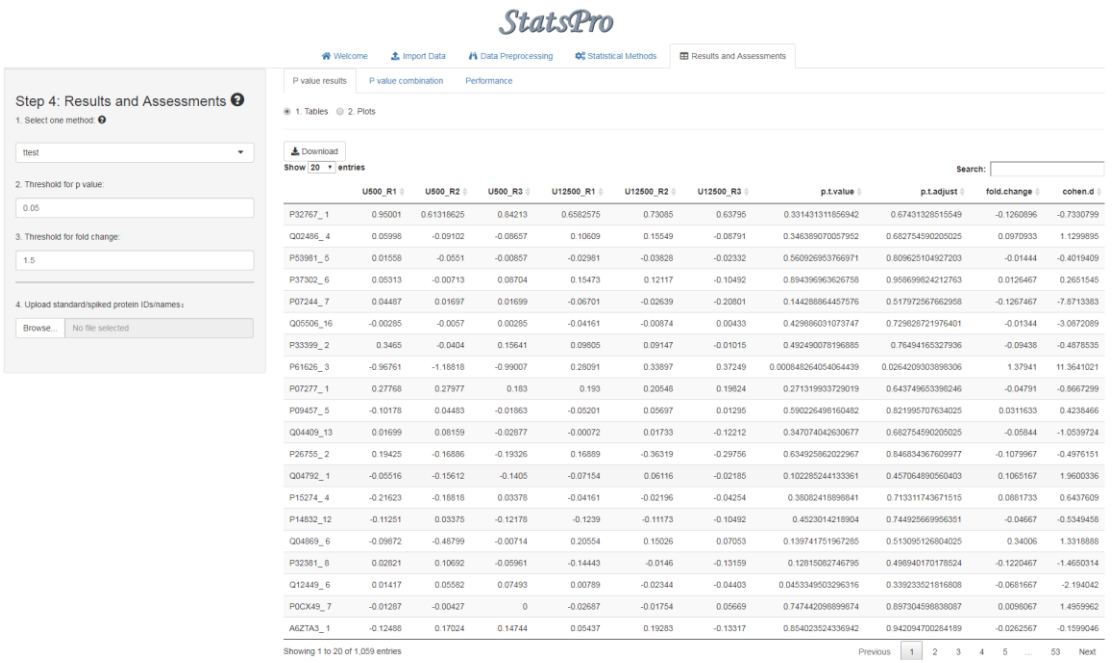

Selected Methods
Dear user, you have chosen several methods as below:
Method 1: ttest
Method 2: aov
Method 3: wiltest
Method 4: kwtest
Method 5: perm
Method 6: limma
Method 7: sam
Method 8: rankpro
Method 9: rots
Method 10: msgrobsum
Method 11: degms
Method 12: plgem
Then click 'OK' and move on...
OK
Step 3: Twelve statistical
Non-Parametric). Please
A. Parametric tests
Method 1: Student's t test (ttest)
Using Student's t test or not?
M1.1 Paired or not?
Student's t test introduction from Wikipedia
B. Non-parametric tests
Method 3: Wilcoxon rank sum test (wilitest)
Using Wilcoxon rank sum test or not?
M1.1 Paired or not?
DOI: 10.1080/01621459.1972.10481279
Method 4: Kruskal-Wallis Rank Sum Test (kwtest)
Using Kruskal-Wallis Rank Sum Test or not?
DOI: 10.1002/9781119196037
Method 5: Permutation test (perm)
Using Permutation test or not?
Package: exactRankTests

5. Results and Assessments

This step will process statistical testing and performance evaluation of every method that users select in “Methods” step. Click “Results and Assessments”, *StatsPro* will start to calculate, a process bar will appear in the bottom right corner to tell users where it goes:

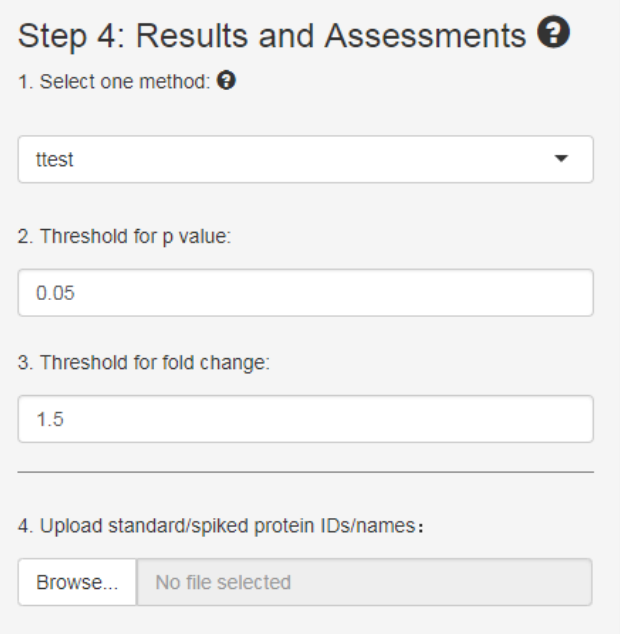


The result from every statistical method will be shown on the “Results” panel:

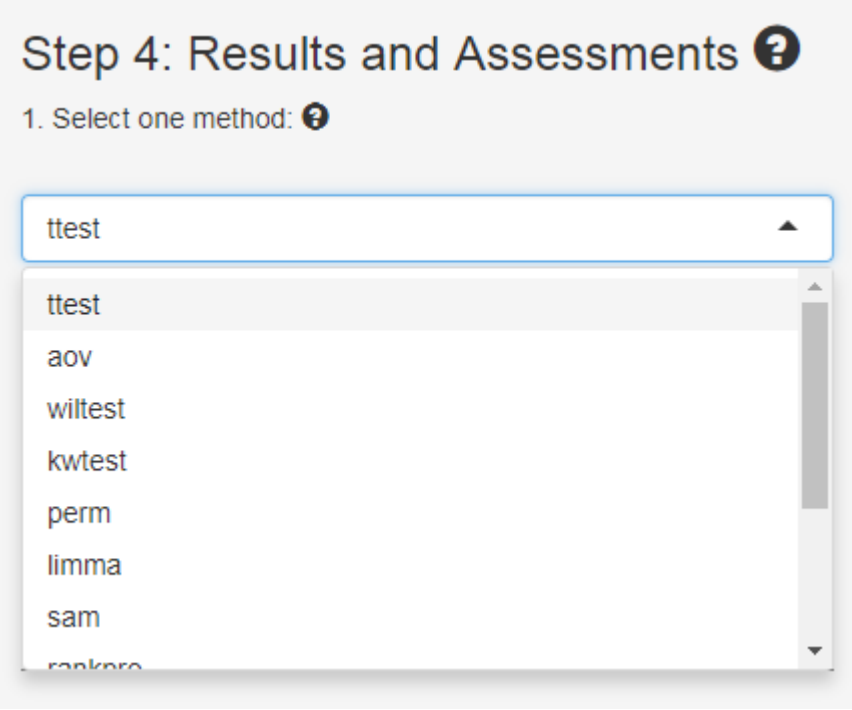


	U500_R1	U500_R2	U500_R3	U12500_R1	U12500_R2	U12500_R3	p.t.value	p.t.adjust	fold.change	cohen.d
P32767_1	0.95001	0.61318625	0.84213	0.6582575	0.73085	0.63795	0.331431311856942	0.67431328515549	-0.1260896	-0.7330799
Q02486_4	0.05998	-0.09102	-0.08657	0.10609	0.15549	-0.08791	0.346389070057952	0.682754590205025	0.0970933	1.1299895
P53981_5	0.01558	-0.0551	-0.00857	-0.02981	-0.03828	-0.02332	0.560926953766971	0.809625104927203	-0.01444	-0.4019409
P37302_6	0.05313	-0.00713	0.08704	0.15473	0.12117	-0.10492	0.894396963626758	0.958699824212763	0.0126467	0.2651545
P07244_7	0.04487	0.01697	0.01699	-0.06701	-0.02639	-0.20801	0.144288864457576	0.517972567662958	-0.1267467	-7.8713383
Q05506_16	-0.00285	-0.0057	0.00285	-0.04161	-0.00674	0.00433	0.429866031073747	0.729828721976401	-0.01344	-3.0872089
P33399_2	0.3465	-0.0404	0.15641	0.09805	0.09147	-0.01015	0.492490078196885	0.76494165327936	-0.09438	-0.4878535
P61626_3	-0.96761	-1.18818	-0.99007	0.28091	0.33897	0.37249	0.000848264054064439	0.0264209303898306	1.37941	11.3641021
P07277_1	0.27768	0.27977	0.183	0.193	0.20548	0.19824	0.271319933729019	0.643749653398246	-0.04791	-0.8667299
P09457_5	-0.10178	0.04483	-0.01863	-0.05201	0.05697	0.01295	0.590226498160482	0.821995707634025	0.0311633	0.4238466
Q04409_13	0.01699	0.08159	-0.02677	-0.00072	0.01733	-0.12212	0.347074042630677	0.682754590205025	-0.05844	-1.0539724
P26755_2	0.19425	-0.16886	-0.19326	0.16889	-0.36319	-0.29756	0.634925862022967	0.846834367609977	-0.1079967	-0.4976151
Q04792_1	-0.05516	-0.15612	-0.1405	-0.07154	0.06116	-0.02185	0.102285244133361	0.457064890560403	0.1065167	1.9600336
P15274_4	-0.21623	-0.16818	0.03378	-0.04161	-0.02196	-0.04254	0.38082418896841	0.713311743671515	0.0881733	0.6437609
P14832_12	-0.11251	0.03375	-0.12178	-0.1239	-0.11173	-0.10492	0.4523014218904	0.744925669956351	-0.04667	-0.5349458
Q04869_6	-0.09672	-0.48799	-0.00714	0.20554	0.15026	0.07053	0.139741751967285	0.513095126804025	0.34006	1.3318888
P32381_8	0.02821	0.10692	-0.05961	-0.14443	-0.0146	-0.13159	0.12815082746795	0.498940170178524	-0.1220467	-1.4650314
Q12449_6	0.01417	0.05582	0.07493	0.00789	-0.02344	-0.04403	0.0453349503296316	0.339233521816808	-0.0681667	-2.194042
P0CX49_7	-0.01287	-0.00427	0	-0.02687	-0.01754	0.05669	0.747442068899674	0.897304598838087	0.0096067	1.4959962
A6ZTA3_1	-0.12488	0.17024	0.14744	0.05437	0.19283	-0.13317	0.854023524336942	0.942094700284189	-0.0262567	-0.1599046

5.1 Parameters



1. Select one method: Herein users can change the parameter on the left panel to check relative result, for example, if users select “ttest”, it will show the result derived from ttest method in the “P value results” part:



2. Threshold for p value: When counting the number of DEPs, those proteins with original P-values or FDR below this threshold will be excluded. The default value is 0.05.

3. Threshold for fold change: When counting the number of DEPs, those proteins with fold change above this threshold or below the reciprocal of this fold change will be included. The default value is

1.5. If the proteomics expression intensities are logarithmic, for example, users select the "*5. Log or not?*" parameter in the "Data Pre-processing" part, the threshold for fold change will be log2(1.5) = 0.585, which means that those proteins with absolute fold change above this threshold (|log2(1.5)| > 0.585) will be included.

4. Upload standard/spiked protein IDs/names: If there are spiked proteins (e.g. UPS1) in the samples, users should upload the spiked protein IDs/names here, for example, there are 48 human proteins in the UPS1 and save their UniProt IDs in a .csv file. The example of these IDs is like:

Accession
P61626
P68871
P16083
P00441
P62988
P01344
O76070
P01375
P15559
P63279
Q15843
P61769
P02787
Q06830
P41159
P00167
P69905
P10145
P01127
P08263

5.2 Results

The results from 12 statistical methods will be shown in the "P value results" part (please also see the "*1. Select one method*" parameter). In the "1. Tables", it shows the results including:

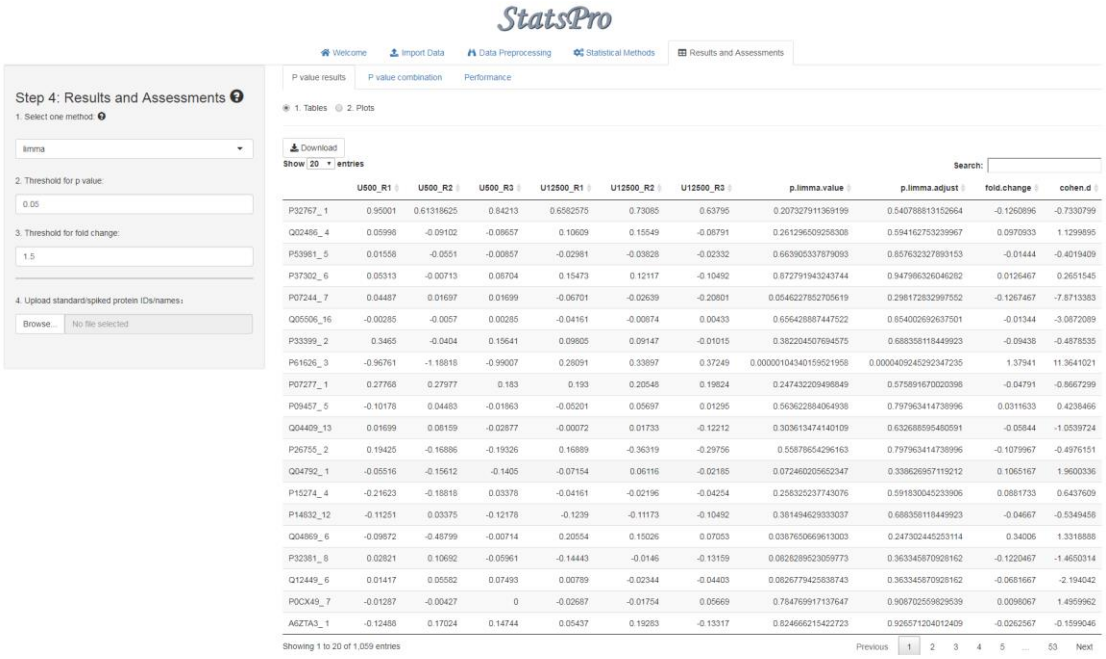
a. Original P-values from every statistical method, for example, when selecting the limma method, the "p.limma.value" column contains the original P-values from limma method;

b. FDR (also adjusted P-values), for example, the "p.limma.adjust" column contains the adjusted P-values from limma method;

c. Fold Change, for example, the values are in the "fold.change" column.

d. Cohen's d effect size, for example the values are in the "cohen.d" column.

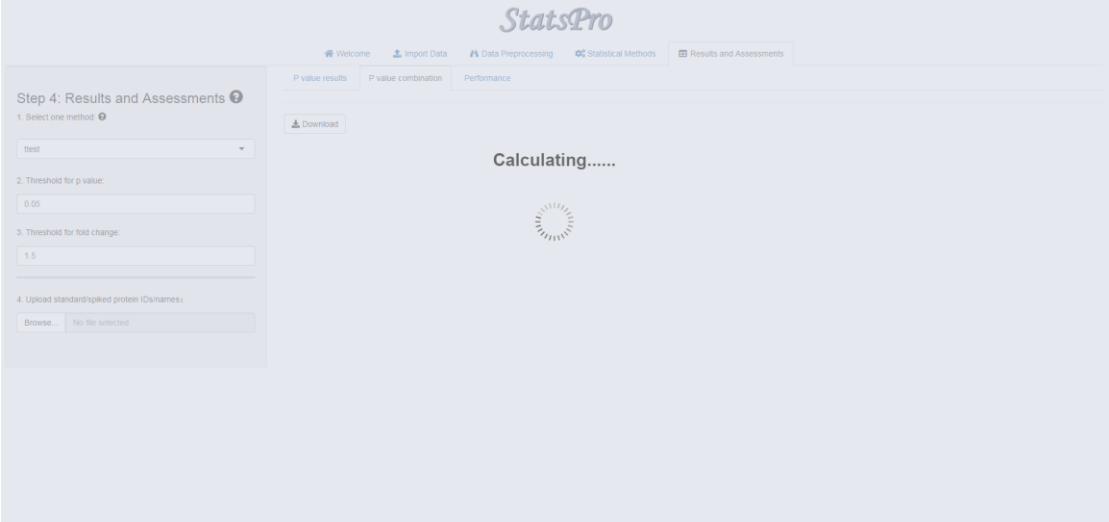
Users can click "Download" button to save this result in a .csv file:



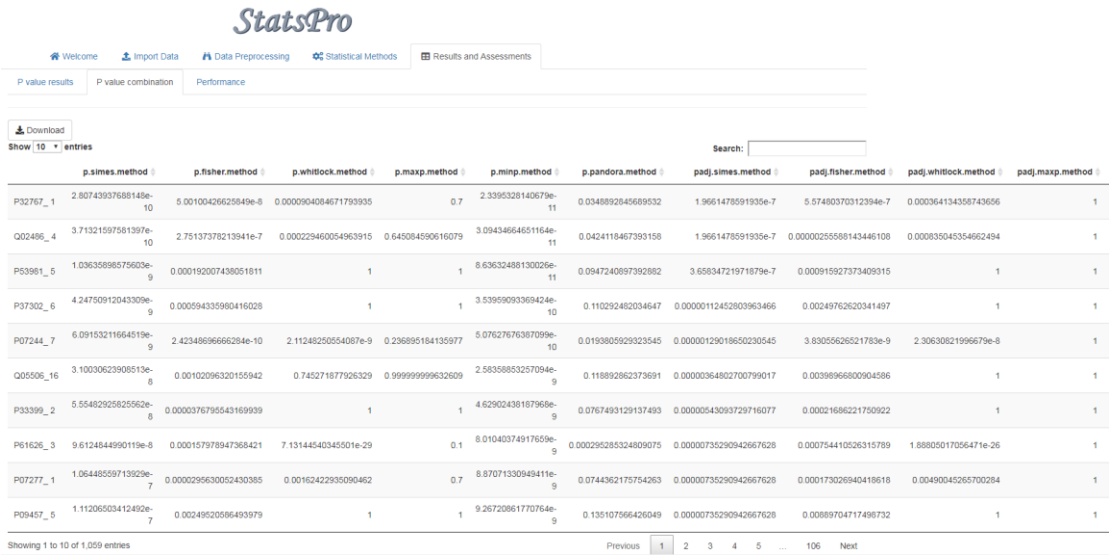
In the “2. Plots”, it shows three plots: one, the distributions of original P-values and adjusted P-values; two, the volcano plot, in which the y axis is plotted based on original P-values; three, the volcano plot, in which the y axis is plotted based on adjusted P-values. Users can click “Download” button to save this result in a .pdf file

Next, click “P value combination”. *StatsPro* will calculate the combined P-values based on 6

combination strategies:



After calculating, it shows like below. The first six columns contain the P-values obtained from each combination strategy (i.e. p.simes.method, p.fisher.method, p.whitlock.method, p.maxp.method, p.minp.method), and the next six columns contain the adjusted P-values based on the Benjamini-Hochberg (BH) method [18], which the column names start with “padj”. Users can click “Download” button to save this result in a .csv file:



	p.simes.method	p.fisher.method	p.whitlock.method	p.maxp.method	p.minp.method	p.pandora.method	padj.simes.method	padj.fisher.method	padj.whitlock.method	padj.maxp.method
P32767_1	2.80743937688148e-10	5.00100426625849e-8	0.000090408467179393 5	0.7	2.3395328140679e-11	0.0348892845689532	1.9661478591935e-7	5.57480370312394e-7	0.000364134358743656	1
Q02486_4	3.71321597581397e-10	2.75137378213941e-7	0.000229460054963915	0.645084590616079	3.0934366465116 4e-11	0.0424118467393158	1.9661478591935e-7	0.000002555681434461 08	0.000835045354662494	1
P53961_5	1.03635896575603e-9	0.000192007438051811	1	1	8.6363248813002 6e-11	0.0947240897392882	3.6583472197187 9e-7	0.000915927373409315	1	1
P37302_6	4.24750912043309e-9	0.000594335980416028	1	1	3.5395093369424e-10	0.110292462034647	0.000011245280396 3466	0.00249762620341497	1	1
P07244_7	6.09153211664519e-9	2.42348696666284e-10	2.112482505540 87e-9	0.236895184135977	5.0762767636709 9e-10	0.0193805929323545	0.000001290186502 30545	3.830556265217 83e-9	2.30630821996679e-8	1
Q05506_16	3.10030623908513e-8	0.00102096320155942	0.745271877926329	0.999999999632609	2.5835885325709 4e-9	0.118892862373691	0.000003648027007 99017	0.00398966800904586	1	1
P33399_2	5.55482925825562e-8	0.000037679554316 9939	1	1	4.6290243818796 6e-9	0.0767493129137493	0.000005430937297 16077	0.00021686221750922	1	1
P61626_3	9.612484990119e-8	0.000157978473684 21	7.131445403455 01e-29	0.1	8.0104037491765 9e-9	0.000295285324809075	0.000007352909426 67628	0.00075441052631578 9	1.88805017056471e-26	1
P07277_1	1.06448559713929e-7	0.000029563005243 0385	0.001624229350 90462	0.7	8.8707133094941 1e-9	0.0744362175754263	0.000007352909426 67628	0.00017302694041861 8	0.00490045265700284	1
P09457_5	1.11206503412492e-7	0.002495205864939 79	1	1	9.2672086177076 4e-9	0.135107566426049	0.000007352909426 67628	0.00889704717496732	1	1

Showing 1 to 10 of 1,059 entries Previous 1 2 3 4 5 ... 106 Next

Then, click the “Performance” part, there are three criteria here:

1. Number of detections: This part shows the Number of DEPs from every method, as below, the “Methods” column contains every method name, the “Number.p.value” means that the number of DEPs is obtained from original P-values and fold change, the “Number.p.adjust” means that the number of DEPs is obtained from adjusted P-values and fold change:

2. Correlation: This part shows the correlation coefficient between P-values and effect sizes, as below, the “Methods” column contains every method name, the “Cor.p.cohen” means that the correlation coefficients are calculated between original P-values and Cohen's d effect size, the “Cor.padj.cohen” means that the correlation coefficients are calculated between adjusted P-values and Cohen's d effect size:

StatsPro

Welcome | Import Data | Data Preprocessing | Statistical Methods | Results and Assessments

P value results | P value combination | Performance

1. Number of detections | 2. Correlation | 3. AUC and F1 score

Download

Show 20 entries | Search:

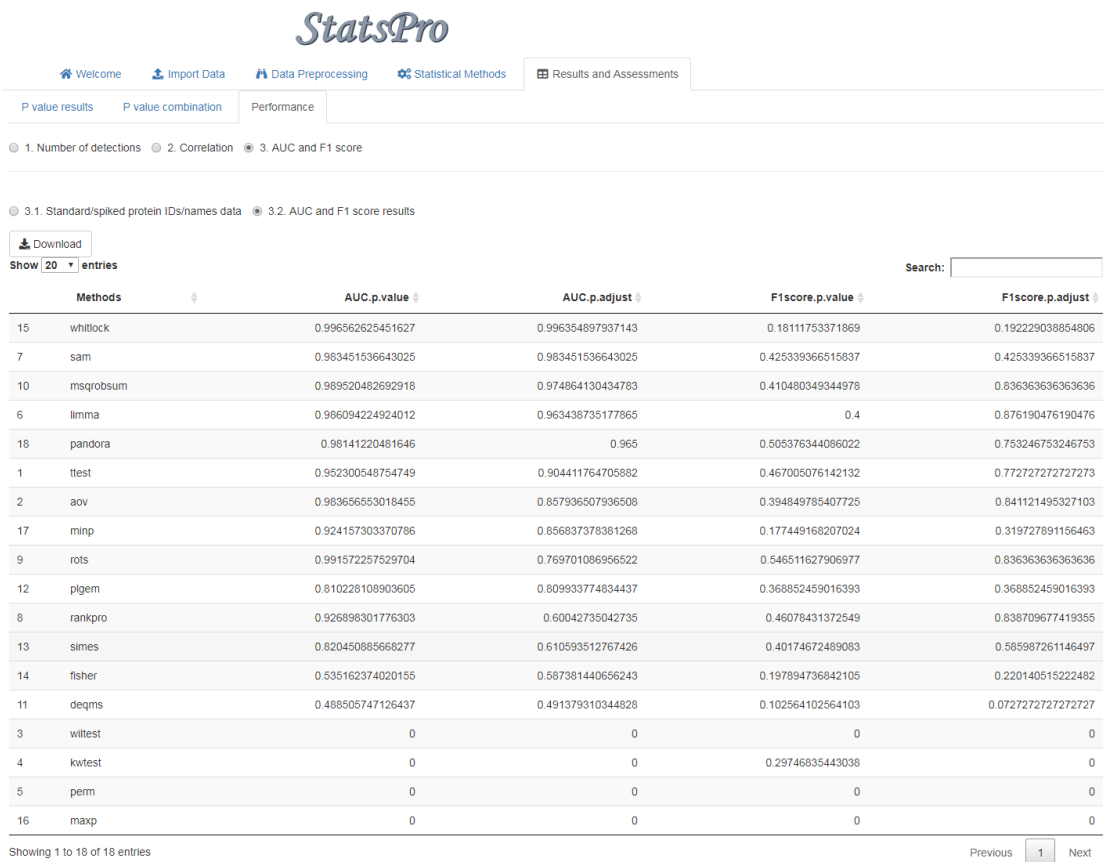
	Methods	Cor.p.cohen	Cor.padj.cohen
1	ttest	0.732951201921834	0.779999603184301
2	aov	0.73657919618254	0.764192954451346
6	limma	0.708313586838568	0.715969757179762
10	msqrobsum	0.700061020875768	0.710526421309667
18	pandora	0.646908507940576	0.670240677018589
15	whitlock	0.647018647533021	0.650327537261096
9	rots	0.576289419919461	0.562885514084919
7	sam	0.555435357251625	0.555435357251625
8	rankpro	0.513123570421643	0.549732168307174
13	simes	0.489367588771441	0.483529924205345
17	minp	0.477054129685562	0.477178329990052
12	plgem	0.386653728681813	0.386225618094661
5	perm	0.381801585830571	0.389343913702866
4	kwtest	0.366124711065784	0.371450875714075
3	wilttest	0.361622502629252	0.363101823636706
14	fisher	0.261966513401275	0.287540390984231
16	maxp	0.285489423786377	0
11	deqms	-0.037767838838804	-0.0364091343771411

Showing 1 to 18 of 18 entries

Previous | 1 | Next

3. AUC and F1 score: In the “3.1. Standard/spiked protein IDs/names data” part, it first shows the spiked proteins, which are used for calculating AUC and F1 score later:

Then, click “3.2. AUC and F1 score results”. It shows the area under the ROC curve and F1 score, as below, the “Methods” column contains every method name, the “AUC.p.value” means that the AUCs are calculated from original P-values, the “AUC.p.adjust” means that the AUCs are calculated from adjusted P-values, the “F1score.p.value” means that the F1 scores are obtained based on original P-values, the “F1score.p.adjust” means that the F1 scores are obtained based on adjusted P-values. The significance level is 0.05:



	Methods	AUC.p.value	AUC.p.adjust	F1score.p.value	F1score.p.adjust
15	whitlock	0.996562625451627	0.996354897937143	0.18111753371869	0.192229038854806
7	sam	0.983451536643025	0.983451536643025	0.425339366515837	0.425339366515837
10	msqrobsum	0.989520482692918	0.974864130434783	0.410480349344978	0.836363636363636
6	limma	0.986094224924012	0.963438735177865	0.4	0.876190476190476
18	pandora	0.98141220481646	0.965	0.505376344086022	0.753246753246753
1	ttest	0.952300548754749	0.904411764705882	0.467005076142132	0.772727272727273
2	aov	0.983656553018455	0.857936507936508	0.394849785407725	0.841121495327103
17	minp	0.924157303370786	0.856837378381268	0.177449168207024	0.319727891156463
9	rots	0.991572257529704	0.769701086956522	0.546511627906977	0.836363636363636
12	plgem	0.810228108903605	0.809933774834437	0.368852459016393	0.368852459016393
8	rankpro	0.926898301776303	0.60042735042735	0.46078431372549	0.838709677419355
13	simes	0.820450885668277	0.610593512767426	0.40174672489083	0.585987261146497
14	fisher	0.535162374020155	0.587381440656243	0.197894736842105	0.220140515222482
11	deqms	0.488505747126437	0.491379310344828	0.102564102564103	0.0727272727272727
3	wilttest	0	0	0	0
4	kwtest	0	0	0.29746835443038	0
5	perm	0	0	0	0
16	maxp	0	0	0	0

6. How to run this tool locally?

StatsPro is an open source software for non-commercial use and all codes can be obtained on our GitHub: https://github.com/YanglabWCH/StatsPro. If users want to run *StatsPro* on their own computer, they should operate as below:

As this tool was developed with R, you may:

a) Install R. You can download R from here: https://www.r-project.org/.

b) Install RStudio. (Recommendatory but not necessary). You can download RStudio from here: https://www.rstudio.com/.

c) Check packages. After installing R and RStudio, you should check whether you have installed these packages (shiny, shinyjs, shinyBS, shinyWidgets, gdata, ggplot2, ggsci, DT, tidyverse, ggExtra, cowplot, readxl, writexl, data.table, Amelia, impute, coin, exactRankTests, limma, samr, RankProd, ROTS, msqrobsum, MSnbase, DEqMS, plgem, effsize, patchwork, survcomp, metaseqR). You may run the codes below to check them:

```
if(!require(pacman)) install.packages("pacman")
pacman::p_load(shiny, shinyjs, shinyBS, shinyWidgets, devtools, gdata, ggplot2, ggsci, DT, tidyverse, ggExtra, cowplot, readxl, writexl, data.table, Amelia, impute, coin, exactRankTests, limma, samr, RankProd, ROTS, msqrobsum, MSnbase, DEqMS, plgem, effsize, patchwork, survcomp, metaseqR)
```

Please note, if you find some packages cannot be installed directly using the above command, you can find them in the GitHub source and install them by, for example:

```
library(devtools)
install_github("statOmics/MSqRobSum")
```

d) Run this tool locally

```
if(!require(StatsPro)) devtools::install_github("YanglabWCH/StatsPro")
library(StatsPro)
StatsPro_app()
```

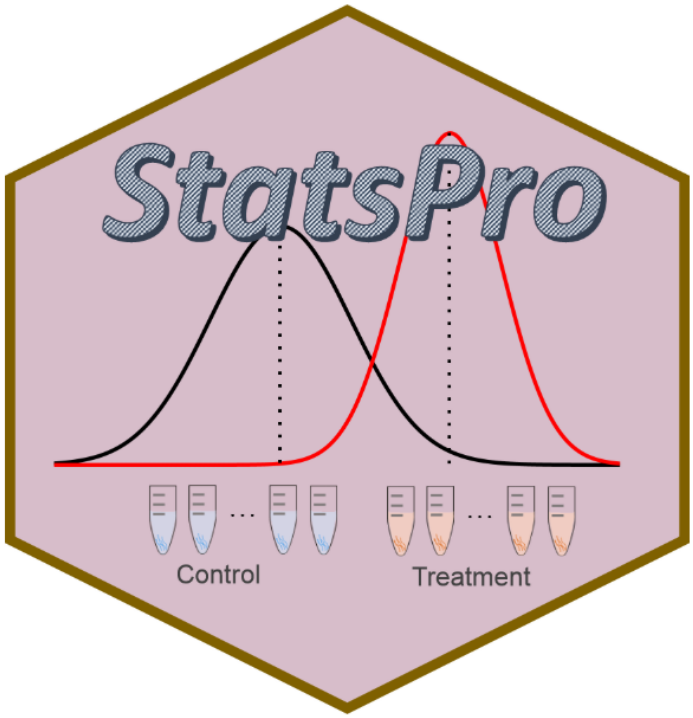
Then StatsPro will be started as below (same as the online version), and the detailed operation about StatsPro can be found in the Supplementary Notes part 1-5 above:

StatsPro

Welcome | Import Data | Data Preprocessing | Statistical Methods | Results and Assessments

Welcome to StatsPro

StatsPro integrates 12 common statistical methods and 6 P-value combination strategies, and then provides three evaluation criteria to assess the performance of each method or strategy. This tool is expected to help scientists detect the differentially expressed proteins and realize the ability of different statistical methods in a systematic view.



StatsPro is developed by R shiny (Version 1.3.2), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at wsslearning@omicsolution.com.

~~ Enjoy yourself in StatsPro ~~

© 2021 Hao Yang's Group. All Rights Reserved.

II. Supplementary tables and figures

Table S1. Description of 18 different approaches (12 statistical methods and 6 P-value combination strategies).

Class	Abbreviation	Algorithm Description	Remarks & Suggestions	Function	Package/Reference
1. Parametric tests	ttest	Student's t-test. It is usually used to determine if the means of two data sets are significantly different from each other.	1. They assume that the data should follow a normal distribution when using these methods, but please note StatsPro do not process normality test. 2. By default, when dealing with two groups of samples, aov is same as ttest with treating the two variances as being equal.	t.test	stats[6]
	aov	Analysis of variance (ANOVA). It provides a statistical test of whether two or more population means are equal.		aov	stats[7]
2. Non-parametric tests	wiltest	Wilcoxon rank sum test. It is used to compare two groups of samples to assess whether their population mean ranks differ.	Non-parametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed).	wilcox.test	stats[8]
	kwtest	Kruskal-Wallis rank sum test. It is an extension of the Wilcoxon rank sum test to allow the comparison of two or more than two independent groups.		kruskal.test	stats[9]
	perm	Permutation test. It permutes the		perm.test	exactRankTests[10]

		observed data by assigning different outcome values randomly to each observation from among the set of actually observed outcomes without replacement.			
	limma	Linear models for microarray data. It fits a linear model to determine differential expression. While most of the functionality of limma has been developed for microarray data, the model fitting routines of limma are useful for many other types of data, for example proteomics data.		lmFit, eBayes	limma[11]
	sam	Significance analysis of microarrays. It is a statistical technique initially for finding significant genes in a set of microarray experiments and assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements.		samr	samr[12]
	rankpro	Rank product. It ranks the proteins according to their fold changes.		RankProducts	RankProd[13]

	rots	Reproducibility optimized test statistic. It aims to rank genomic features of interest (such as genes, proteins and transcripts) in order of evidence for differential expression in two-group comparisons.		ROTS	ROTS[14]
	msqrobsum	Robust differential protein expression analysis for label-free quantitative proteomics and robust peptide expression summarization. It provides robust protein level summaries that account for peptide specific effects, which are then further processed using robust ridge regression.		msqrobsum	msqrobsum[15]
	deqms	Differential Expression analysis of quantitative Mass Spectrometry data. It is a robust statistical method developed on top of limma and implemented specifically for differential protein expression analysis in mass spectrometry data.		contrasts.fit, eBayes, spectraCounteBayes	DEqMS[16]
	plgem	Power law global error model. It uses highly replicated microarray data to empirically determine the true variance		plgem.fit, plgem.obsStn, plgem.resampl	plgem[17]

		versus mean dependence that exists in this type of data.		edStn, plgem.pValue	
3. P-value combination strategies	simes	Let p_{i1} , p_{i2}, ..., p_{im} be the P-value scores returned for gene i after the application of m statistical tests. Let also $p_{i(1)}$, $p_{i(2)}$, ..., $p_{i(m)}$ be the aforementioned P-values sorted in increasing order. Then, according to Simes' method, the combined P-value can be: $p_i^* = min\{p_{i(k)}/k\}, k \in (1, \dots, m)$	P-value combination means combining different P-values from various statistical methods based on one certain rule. In the left cells, i means ith gene (for this work, it means ith protein), j means jth statistical method, m means the total number of statistical methods (i.e. m=12 if users select all statistical methods).	combine.simes	metaseqR[19]
	fisher	According to the Fisher's method, let f be the statistic defied by the natural logarithm of the product of m individual P-values (from m statistical tests) multiplied by –2: $\mathrm{f} = -2\sum_{j=1}^{m} lnp_{ij}$. f follows an X^2 distribution with 2m degrees of freedom, which can be used to derive the combined P-value from m statistical tests.		fisher.method	
	whitlock	According to Whitlock's weighted Z-method, the weighted Z statistic for each gene i: $Z_j^w = \sum_{j=1}^{m} w_j Z_j /$		combine.test	

		$\sqrt{\sum_{j=1}^{m} w_j^2}$. It follows the standard Normal distribution N(0,1), which can be used to derive the P-value of the combined tests.			
	maxp	The combined P-value is: $p_i^* = max\{p_{ij}\}, j \in (1, \dots, m)$		combine.maxp	
	minp	The combined P-value is: $p_i^* = min\{p_{ij}\}, j \in (1, \dots, m)$		combine.minp	
	pandora	The combined P-value is: $p_i^* = \prod_{j=1}^{m} p_{ij}^{w_j}$, with $\sum_{j=1}^{m} w_j = 1$, where w_j represent automatically assigned or user-specific weights for the j statistical tests performed.		combine.weight	
Cohen's d effect size	Cohend	For two independent groups, effect size can be measured by the standardized difference between two means: $\text{Cohend} = (M_1 - M_2)/s$, where $M_1 - M_2$ is the difference between the group means, s is the standard deviation of either group.	Cohen classified effect sizes as small (d = 0.2), medium (d = 0.5), and large (d $\geqslant$ 0.8).	cohen.d	effsize[20]

III. References

(1) Cox, J.; Mann, M. *Nature biotechnology* **2008**, *26*, 1367-1372.

(2) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinovic, S. M.; Cheng, L. Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; Vitek, O.; Rinner, O.; Reiter, L. *Mol Cell Proteomics* **2015**, *14*, 1400-1410.

(3) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. *Bioinformatics* **2001**, *17*, 520-525.

(4) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Xuan, Y.; Sondermann, J.; Schmidt, M.; Gomez-Varela, D.; Reiter, L. *Molecular & cellular proteomics : MCP* **2017**, *16*, 2296-2309.

(5) Wang, S.; Li, W.; Hu, L.; Cheng, J.; Yang, H.; Liu, Y. *Nucleic acids research* **2020**, *48*, e83.

(6) Al-Achi, A. *Research & Reviews: Journal of Hospital and Clinical Pharmacy* **2019**, *5*, 1.

(7) Bewick, V.; Cheek, L.; Ball, J. *Critical care* **2004**, *8*, 130-136.

(8) Bauer, D. F. *Journal of the American Statistical Association* **1972**, *67*, 687-690.

(9) Ostertagova, E.; Ostertag, O.; Kováč, J. In *Applied Mechanics and Materials*; Trans Tech Publ, 2014, pp 115-120.

(10) Dwass, M. *The Annals of Mathematical Statistics* **1957**, 181-187.

(11) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. *Nucleic acids research* **2015**, *43*, e47.

(12) Tusher, V. G.; Tibshirani, R.; Chu, G. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 5116-5121.

(13) Breitling, R.; Armengaud, P.; Amtmann, A.; Herzyk, P. *Febs Lett* **2004**, *573*, 83-92.

(14) Elo, L. L.; Filen, S.; Lahesmaa, R.; Aittokallio, T. *IEEE/ACM transactions on computational biology and bioinformatics* **2008**, *5*, 423-431.

(15) Sticker, A.; Goeminne, L.; Martens, L.; Clement, L. *Mol Cell Proteomics* **2020**, *19*, 1209-1219.

(16) Zhu, Y.; Orre, L. M.; Zhou Tran, Y.; Mermelekas, G.; Johansson, H. J.; Malyutina, A.; Anders, S.; Lehtio, J. *Mol Cell Proteomics* **2020**, *19*, 1047-1057.

(17) Pavelka, N.; Pelizzola, M.; Vizzardelli, C.; Capozzoli, M.; Splendiani, A.; Granucci, F.; Ricciardi-Castagnoli, P. *Bmc Bioinformatics* **2004**, *5*, 203.

(18) Benjamini, Y.; Hochberg, Y. *Journal of the Royal statistical society: series B (Methodological)* **1995**, *57*, 289-300.

(19) Moulos, P.; Hatzis, P. *Nucleic acids research* **2015**, *43*, e25.

(20) Sullivan, G. M.; Feinn, R. *Journal of graduate medical education* **2012**, *4*, 279-282.