IonStar enables high-precision, low-missing-data proteomics quantification in large sample cohorts

#Xiaomeng Shen[1, 2], #Shichen Shen[1, 2], Jun Li[1,2], Qiang Hu[3], Lei Nie[4], Chengjian Tu[1, 2], Xue Wang[2, 3], David J. Poulsen[5], Benjamin C. Orsburn[6*], Jianmin Wang[3*], Jun Qu[1, 2*]

[1]Department of Pharmaceutical Sciences, SUNY at Buffalo, Buffalo, NY; [2]Center of Excellence in Bioinformatics & Life Science, Buffalo, NY; [3]Roswell Park Cancer Institute, Buffalo, NY; [4]Shandong University, China; [5]Department of neurosurgery, Jacobs School of Medicine and Biomedical Sciences, SUNY at Buffalo, Buffalo, NY; [6]Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, MD

SI APPENDIX

www.pnas.org/cgi/doi/10.1073/pnas.1800541115

Fig. S10 Median intra-group CV of quantified proteins in biological groups.

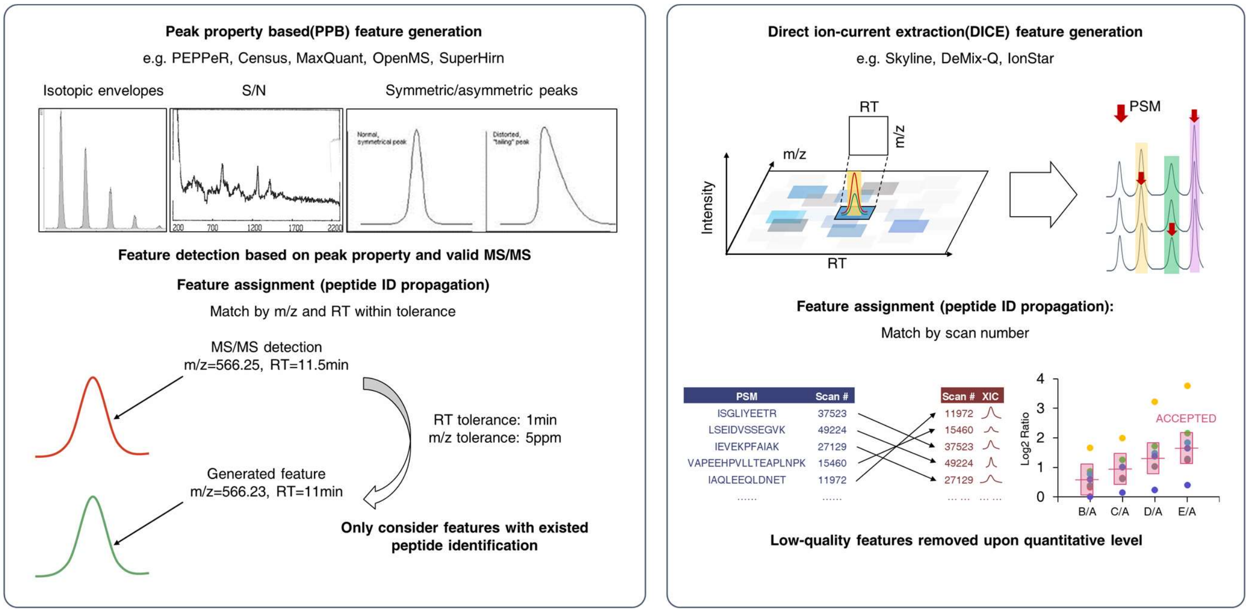
Fig. S11 Pearson correlation of protein ratios in the two brain regions.

Fig. S12 Ingenuity Pathway Analysis results of the TBI proteomics dataset.

File S1 Detailed parameters used for each software.

File S2 User manual for IonStar build 0.1.4.

SUPPLEMENTARY FIGURES


Peak property based(PPB) feature generation
e.g. PEPPeR, Census, MaxQuant, OpenMS, SuperHirn
Isotopic envelopes
S/N
Symmetric/asymmetric peaks
200
700
1200
1700
2200
Normal, symmetrical peak
Distorted, "tailing" peak
Feature detection based on peak property and valid MS/MS
Feature assignment (peptide ID propagation)
Match by m/z and RT within tolerance
MS/MS detection
m/z=566.25, RT=11.5min
RT tolerance: 1min
m/z tolerance: 5ppm
Generated feature
m/z=566.23, RT=11min
Only consider features with existed peptide identification
Direct ion-current extraction(DICE) feature generation
e.g. Skyline, DeMix-Q, IonStar
RT
m/z
m/z
Intensity
RT
PSM
Feature assignment (peptide ID propagation):
Match by scan number
4
3
2
1
0
Log2 Ratio
ACCEPTED
B/A
C/A
D/A
E/A
Low-quality features removed upon quantitative level

PSM	Scan #
ISGLIYEETR	37523
LSEIDVSSEGVK	49224
IEVEKPFAIAK	27129
VAPEEHPVLLTEAPLNPK	15460
IAQLEEQLDNET	11972
……	……

Scan #	XIC
11972	⋀
15460	⋀
37523	⋀
49224	⋀
27129	⋀
… …	… …

Fig. S1 Depiction of feature generation and peptide identity propagation strategies. "PPB (peak property based) + match by m/z and RT with tolerance"[1] and "DICE (Direct Ion-Current Extraction) + match by scan number" that can be used in MS1-based quantification.

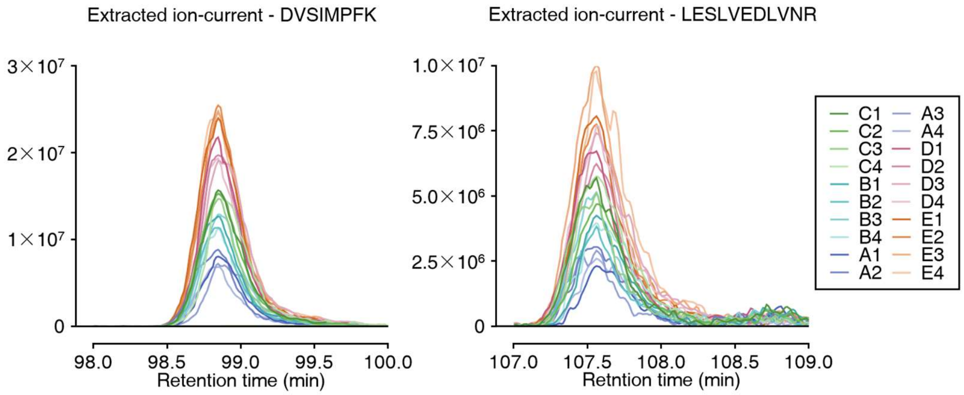


Fig. S2 Examples of extracted ion currents of two E. Coli peptides by IonStar, taken from the benchmark dataset.

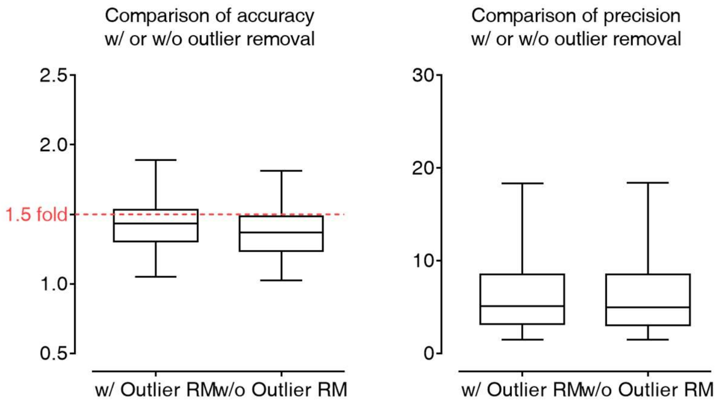


Fig. S3 Effects of post-feature generation quality control measure on quantitative quality. *PCOut*^{2} significantly improved accuracy while showed no perceivable impacts on precision.

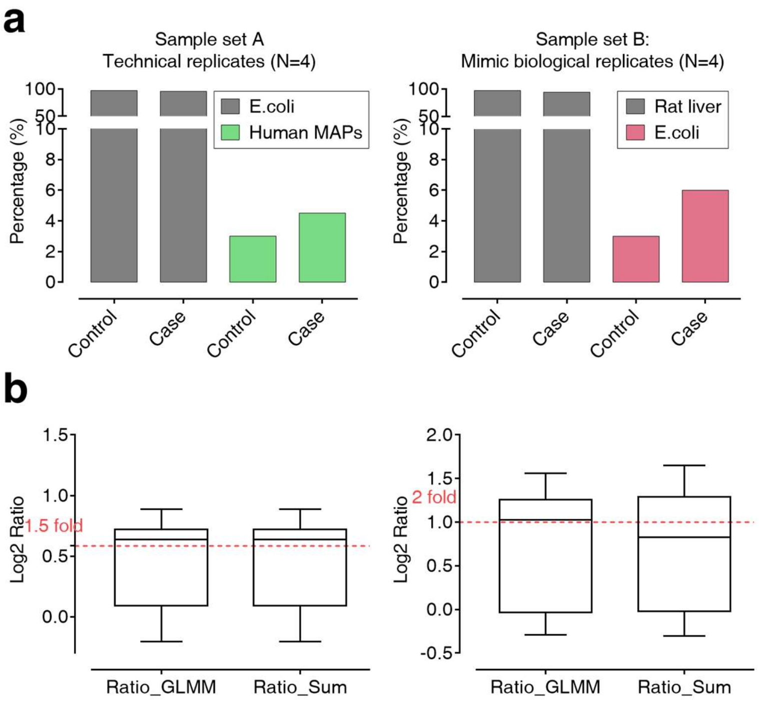


Fig. S4 Application of Generalized Linear Mixed Model (GLMM) in spike-in sample sets. **(a)** Experimental setup of two types of spike-in samples sets: Technical replicates and Mimic biological replicates; **(b)** Comparison of quantitative accuracy in the two types of sample sets using GLMM and sum intensities. GLMM gave better accuracy in Mimic biological replicate sample set.

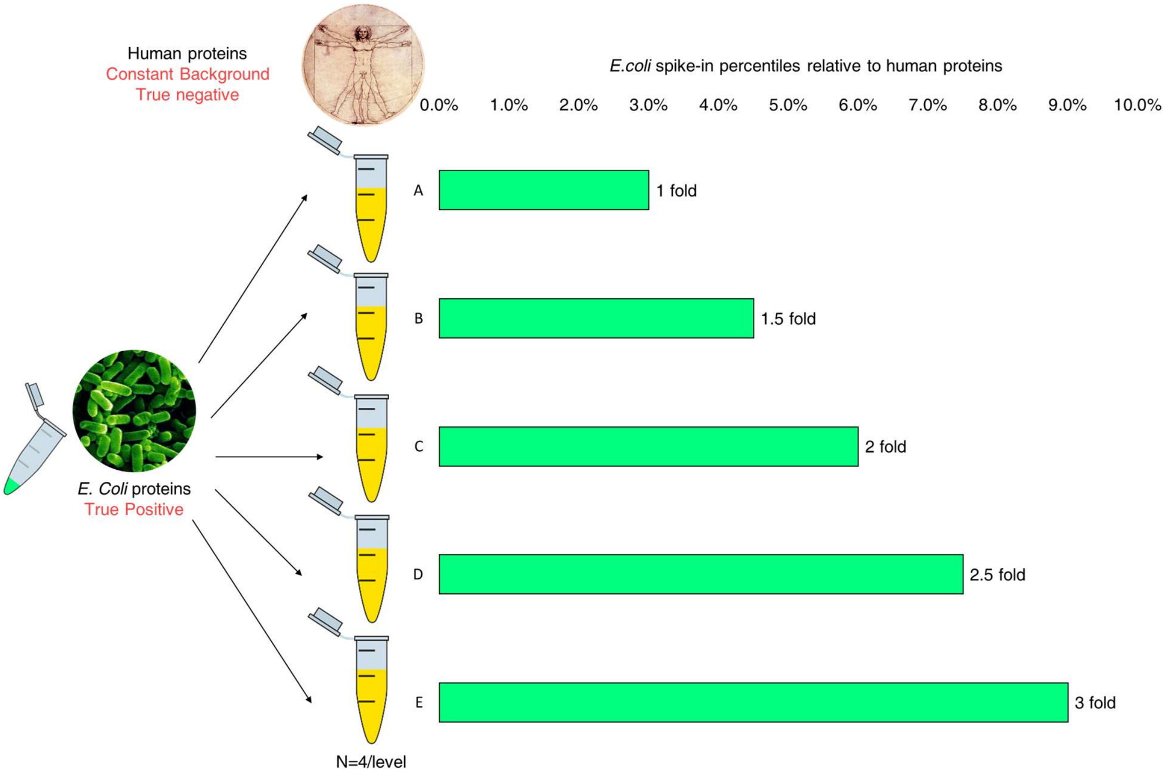


Fig. S5 Design of the spike-in sample set for benchmarking different quantitative approaches[3]. E. Coli protein lysate (true positive) was spiked at low and variable levels into high and constant backgrounds of human cell lysate (true negative). Each of the five spike-in levels has 4 technical replicates (N=20 in total).

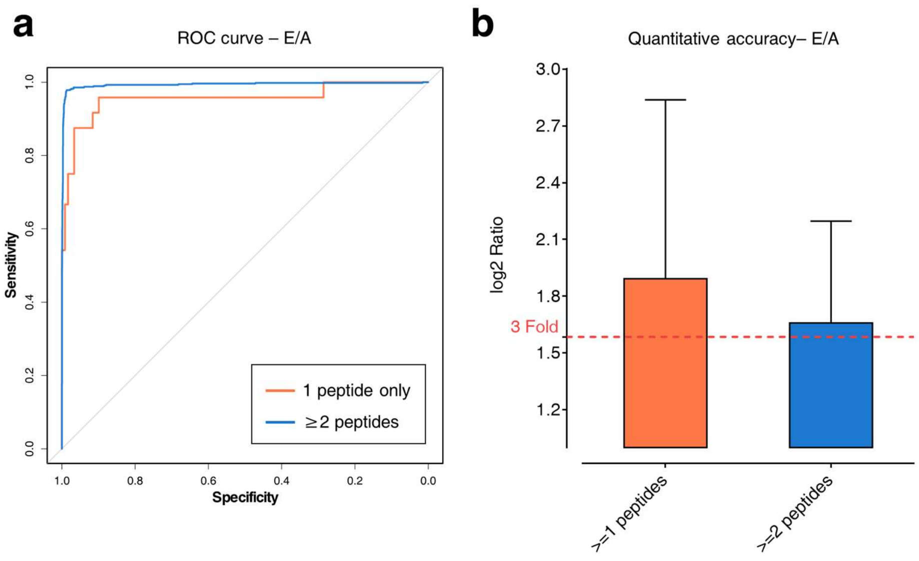


Fig. S6 Comparison of protein quantification results with at least 1 or 2 unique peptide(s) per protein, using spike-in level E (9% E. Coli proteins) vs. A (3% E. Coli proteins) as an example. **(a)** ROC plot comparing quantification of proteins with only 1 peptide and ≥2 peptides; **(b)** Quantitative accuracy of proteins under 1-peptide and 2-peptide criterion.

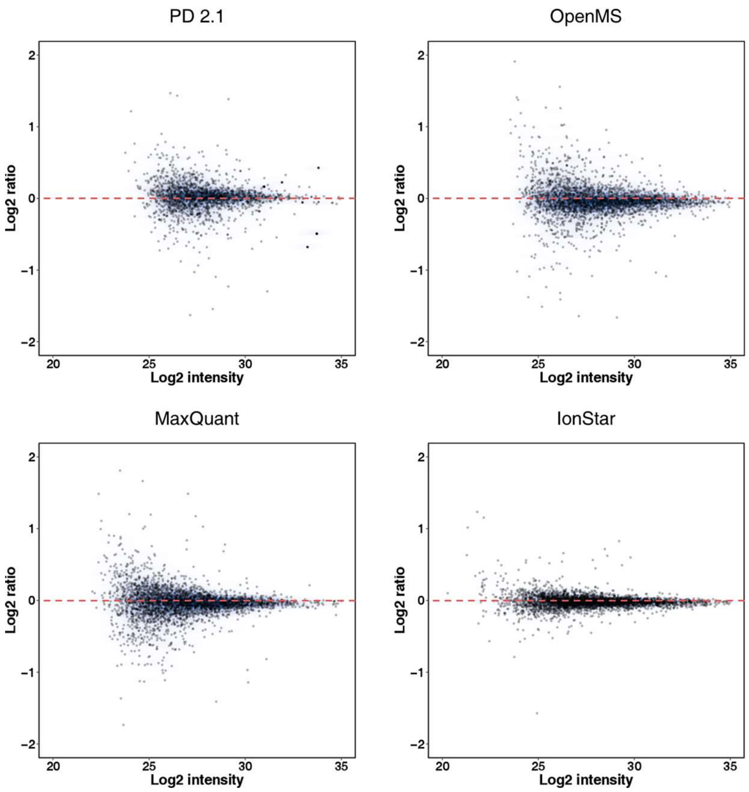


Fig. S7 Quantitative results of human proteins (true negative) in the benchmark dataset from the four MS1-based quantitative approaches, quantitative intensities and ratios are shown in log2 scale. IonStar showed the least deviated distribution, especially for low-abundance proteins.

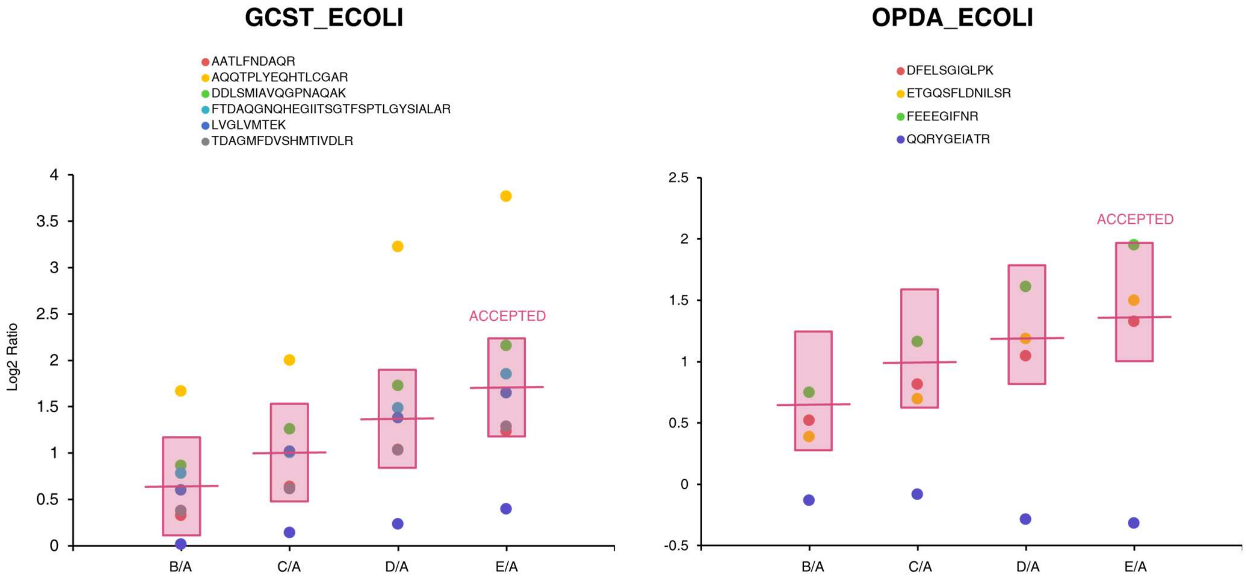


Fig. S8 Examples to demonstrate the elimination of low-quality peptides by the post-feature generation quality control function. Red lines mark the theoretical true ratio, and shaded rectangles mark the zone in which peptide quantitative data is deemed as “acceptable” by the OutlierPeptideRM function.

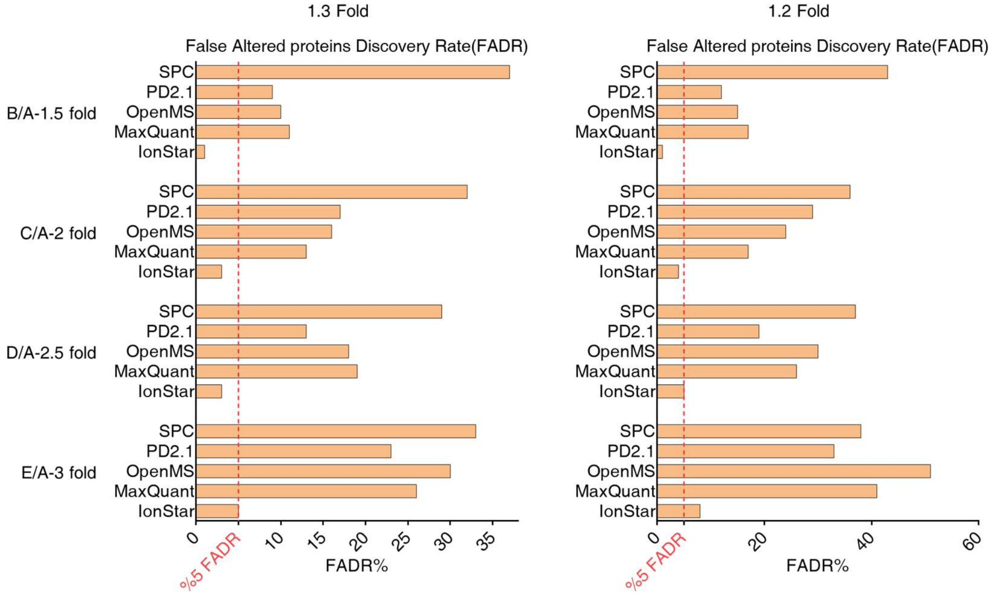


Fig. S9 False Altered-protein Discovery Rate (FADR) calculation using different ratio thresholds.

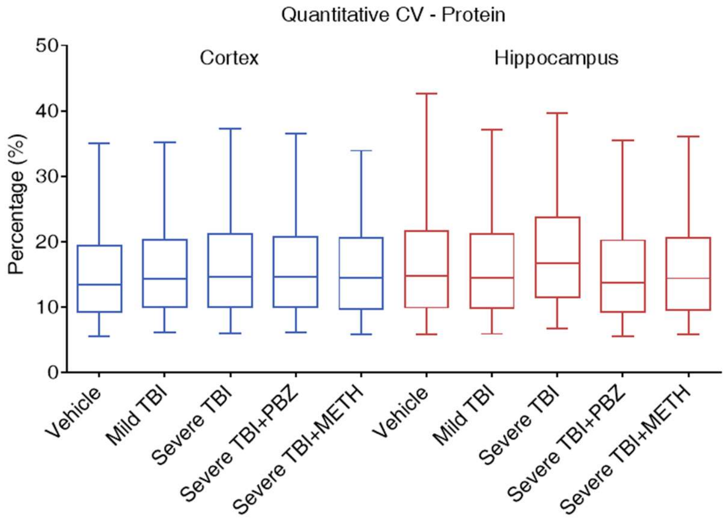


Fig. S10 Median intra-group CV of quantified proteins in biological groups from the 100 rat brain samples.

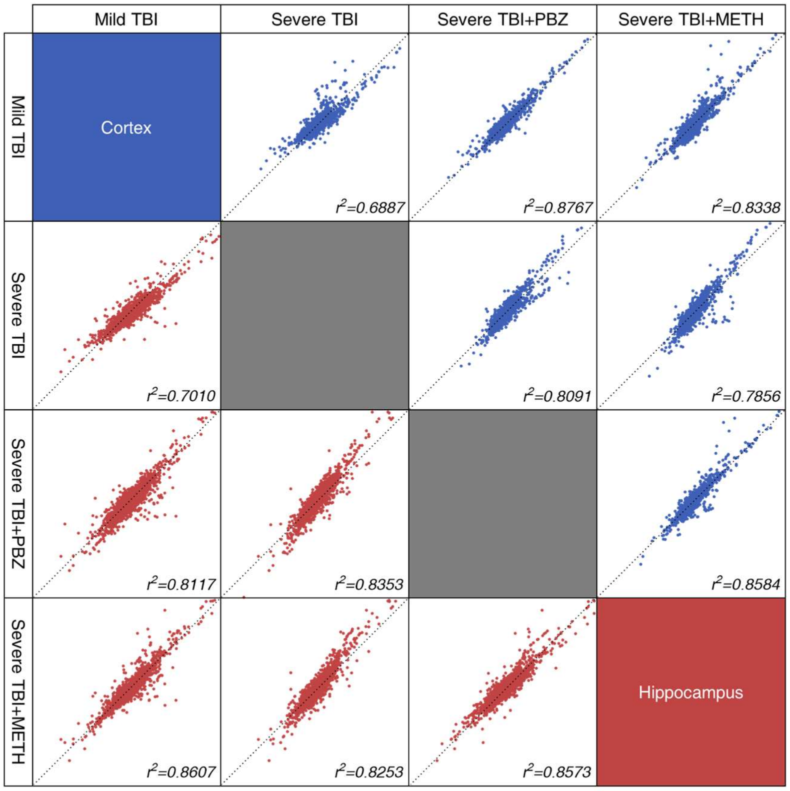


Fig. S11 Pearson correlation of protein ratios between different experimental groups and vehicle control in the two brain regions. Blue and red colors refer to cortex and hippocampus correspondingly.

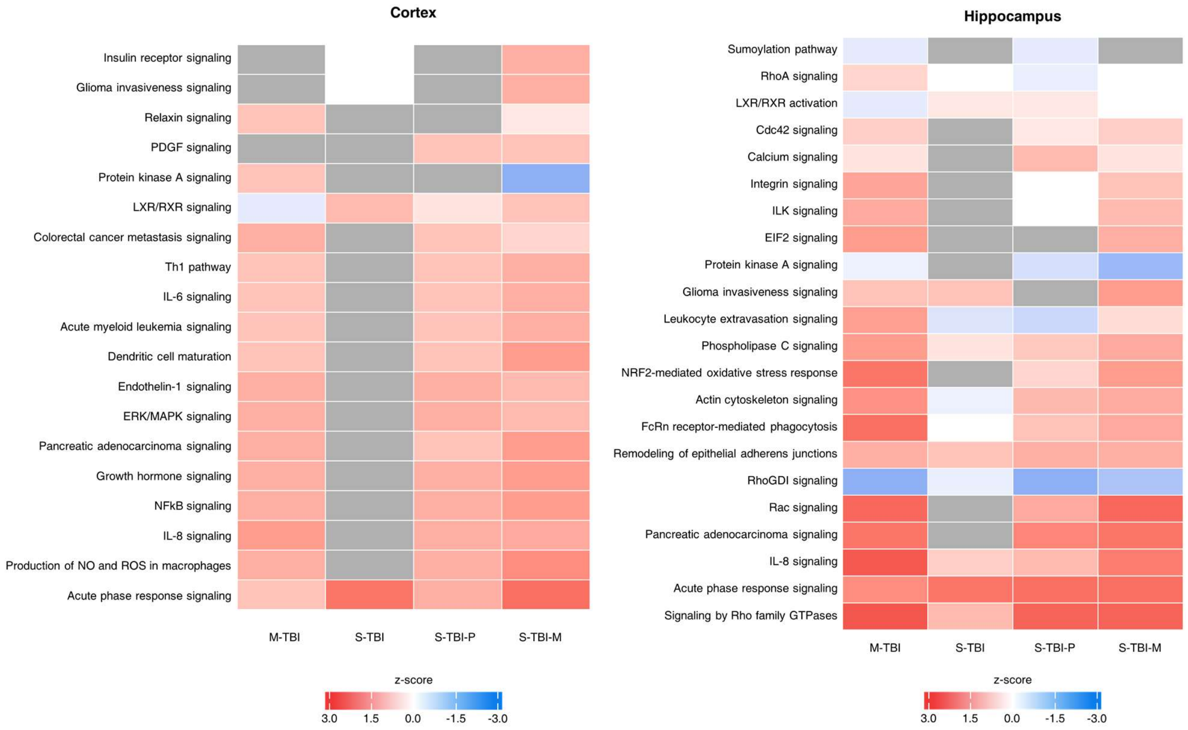


Fig. S12 Ingenuity Pathway Analysis[4] results of the TBI proteomics dataset (left: cortex; right: hippocampus).

File S1 Parameters used in MaxQuant and OpenMS

MaxQuant:

Fixed modifications	Carbamidomethyl (C)
Decoy mode	revert
Special AAs	KR
Include contaminants	False
MS/MS tol. (FTMS)	20 ppm
PSM FDR	0.005
Protein FDR	0.01
Site FDR	0.01
Use Normalized Ratios For Occupancy	True
Min. peptide Length	6
Min. score for unmodified peptides	0
Min. score for modified peptides	0
Min. delta score for unmodified peptides	0
Min. delta score for modified peptides	0
Min. unique peptides	1
Min. razor peptides	2
Min. peptides	2
Use only unmodified peptides and	False
Peptides used for protein quantification	Razor
Discard unmodified counterpart peptides	True
Min. ratio count	2
Re-quantify	True

Match between runs True

Matching time window [min] 1

Alignment time window [min]20

Find dependent peptides False

Site tables Oxidation (M)Sites.txt

Decoy mode revert

Special AAs KR

Include contaminants False

RT shift False

Advanced ratios True

OpenMS:

Module	**Parameter**	**Value**
PeakPicker	algorithm:signal_to_noise	0
	algorithm:ms1_only	TRUE
FeatureFinder	algorithm:mass_trace:mz_tolerance	0.02
	algorithm:mass_trace:min_spectra	3
	algorithm:mass_trace:slope_bound	1
	algorithm:isotopic_pattern:charge_low	2
	algorithm:isotopic_pattern:charge_high	6
	algorithm:isotopic_pattern:mz_tolerance	0.03
	algorithm:seed:min_score	0.1
	algorithm:feature:min_score	0.3
	algorithm:feature:min_isotope_fit	0.1

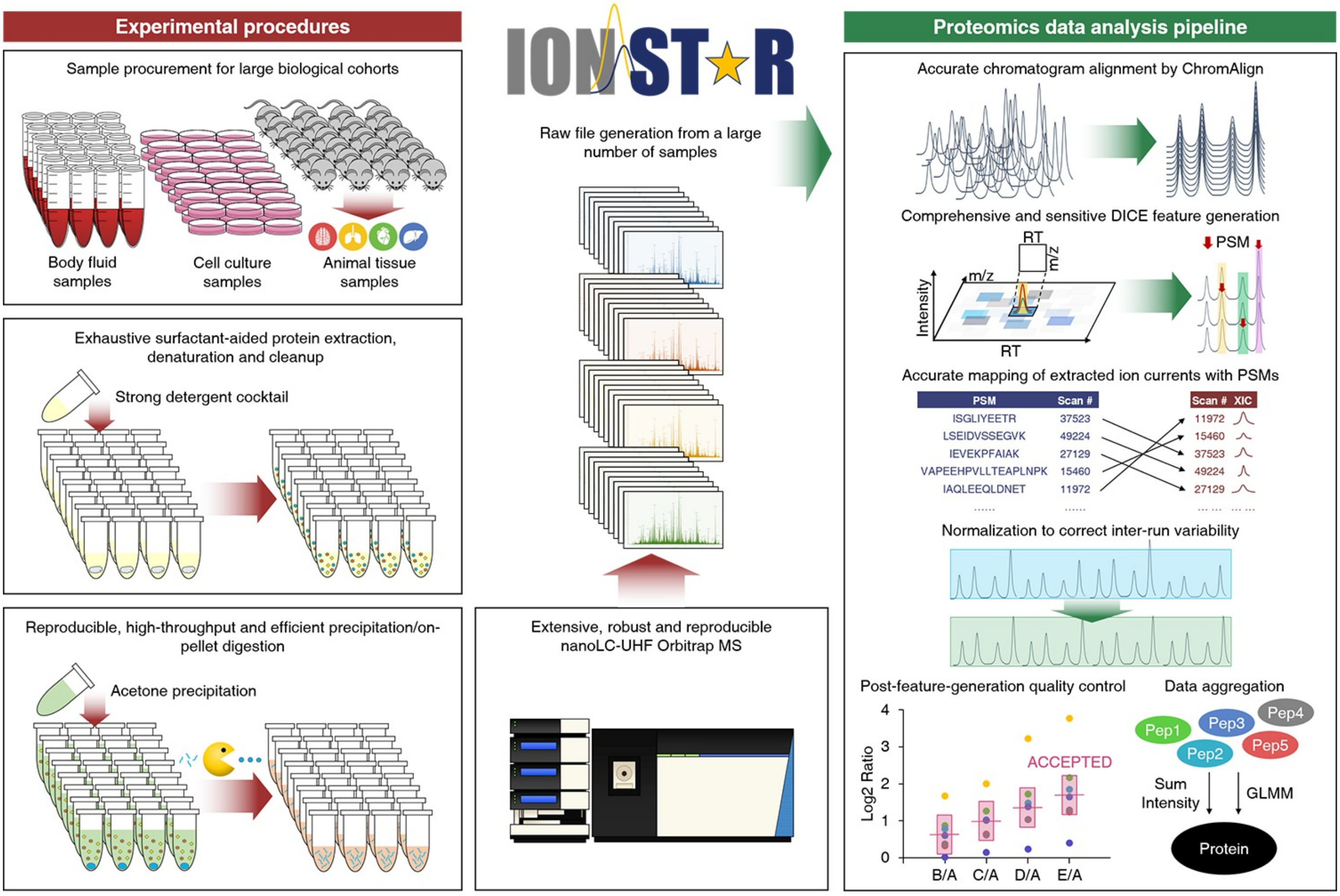
	algorithm:feature:min_trace_score	0.1
	algorithm:feature:max_rt_span	3
IDMapper	rt_tolerance	10
	mz_tolerance	30
	mz_reference	peptide
	use_centroid_mz	TRUE
MapAligner	algorithm:min_run_occur	2
	algorithm:max_rt_shift	300
FeatureLinker	algorithm:use_identifications	TRUE
	algorithm:distance_RT:max_difference	300
	algorithm:distance_MZ:max_difference	0.02

File S2 User manual for IonStar build 0.1.4

IonStar USER MANUAL

For Build 0.1.4

Introduction



Experimental procedures
Sample procurement for large biological cohorts
Body fluid samples
Cell culture samples
Animal tissue samples
Exhaustive surfactant-aided protein extraction, denaturation and cleanup
Strong detergent cocktail
Reproducible, high-throughput and efficient precipitation/on-pellet digestion
Acetone precipitation
IONSTAR
Raw file generation from a large number of samples
Extensive, robust and reproducible nanoLC-UHF Orbitrap MS
Proteomics data analysis pipeline
Accurate chromatogram alignment by ChromAlign
Comprehensive and sensitive DICE feature generation
RT
m/z
Intensity
m/z
RT
PSM
Accurate mapping of extracted ion currents with PSMs
PSM
Scan #
ISGLIYEETR 37523
LSEIDVSSEGVK 49224
IEVEKPFAIAK 27129
VAPEEHPVLLTEAPLNPK 15460
IAQLEEQLDNET 11972
Scan #
XIC
11972
15460
37523
49224
27129
Normalization to correct inter-run variability
Post-feature-generation quality control
Log2 Ratio
0
1
2
3
4
ACCEPTED
B/A
C/A
D/A
E/A
Data aggregation
Pep1
Pep2
Pep3
Pep4
Pep5
Sum Intensity
GLMM
Protein

IonStar is an MS1-based quantitative method for label-free proteomics experiments, devised to address issues related with quantitative precision, missing data, and false-positive discovery of protein changes in large-cohort analysis.

IonStar comprises of two parts: experimental procedures (left panel) and a proteomics data analysis pipeline (right panel). Details of the experimental procedures can be found in Shen et al. *J Proteome Res.* (2017) and An et al. *Anal Chem.* (2015).

This manual will focus on the data analysis pipeline part of IonStar, aiming at helping IonStar users to run the pipeline in their own computational environment.

Prerequisites

Software and dataset availability

The primary software packages used in IonStar are **SIEVE^{TM}** and **IonStarStat**.

SIEVE^{TM} is a commercial software from Thermo Fisher Scientific. The latest version of SIEVE ^{TM} is v2.2 SP2. Please contact Thermo Fisher Scientific regarding the quote for SIEVE^{TM}. To ensure of proper performance of SIEVE^{TM}, we recommend running SIEVE^{TM} on a PC with at least 16-core processors and at least 192 GB RAM.

R package **IonStarStat** and related scripts (**IonStar_FrameGen.R**, **IonStar_Run.R**) can be downloaded here. All operations in this manual are accomplished under R version 3.4.3 and RStudio ver 1.1.442.

The dataset used in this manual as an example (Multi-level Human background+E.coli spike-in) can be downloaded from PRIDE Archive (PRIDE ID: PXD003881).

Installing IonStarStat

IonStarStat package can be installed directly in RStudio by running the following commands in the R Console:

```
#Install dependencies "RSQLite""MCMCglmm""affyPLM””mvoutlier”
source("https://bioconductor.org/biocLite.R")
biocLite("affyPLM")
biocLite("MCMCglmm")
biocLite("RSQLite")
install.packages("mvoutlier")
install.packages("IonStarStat_0.1.4.tar.gz", repos = NULL, type = "source")
```

Upon finishing installation, load IonStarStat into the R environment as follows:

```
#Load IonStarStat
library("IonStarStat")
```

File location

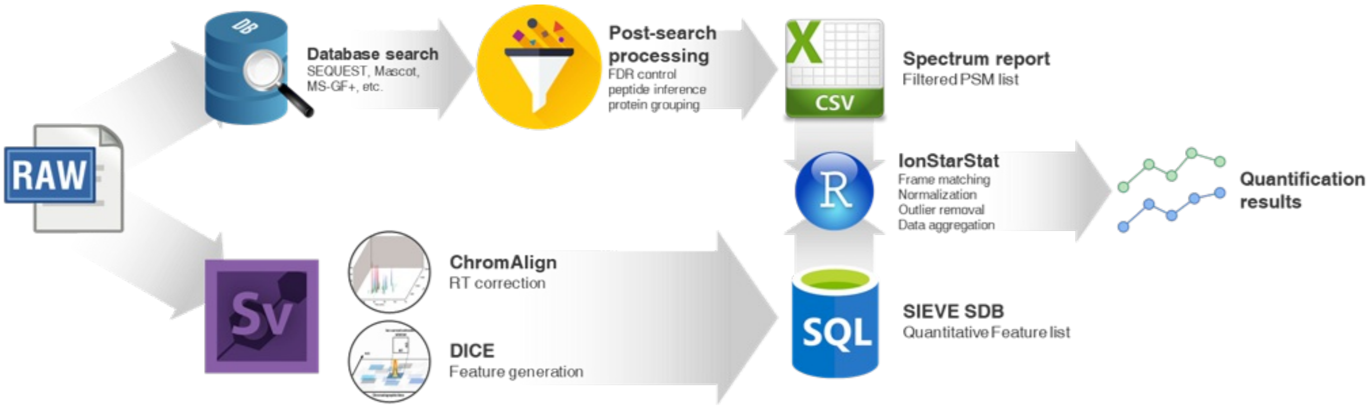
To perform using IonStar, it is recommended to put all files under the same working directory, including:

- LC-MS raw files `.raw`
- Spectrum report `.csv`, `.tsv`, or `.txt`
- SIEVE database file `.sdb`
- Annotated frame list `.csv`
- Sample list `.csv`
- Protein & peptide quantitative results `.csv`

- `IonStar_FrameGen.R` and `IonStar_Run.R`

Use `setwd()` to locate the files whenever necessary.

Quickstart



Step 1: Protein identification

Protein identification can be performed by any database searching engines and post-search processing tools. The final output is a so-called spectrum report containing PSMs from all sample runs passing the confidence threshold (*e.g.* FDR). The spectrum report can be exported from a number of software packages, *e.g.* **Proteome Discoverer**, **Scaffold**. Key information necessary for data integration include **rawfile name** and **MS2 scan number**. The file format of the spectrum report needs to be `.csv`.

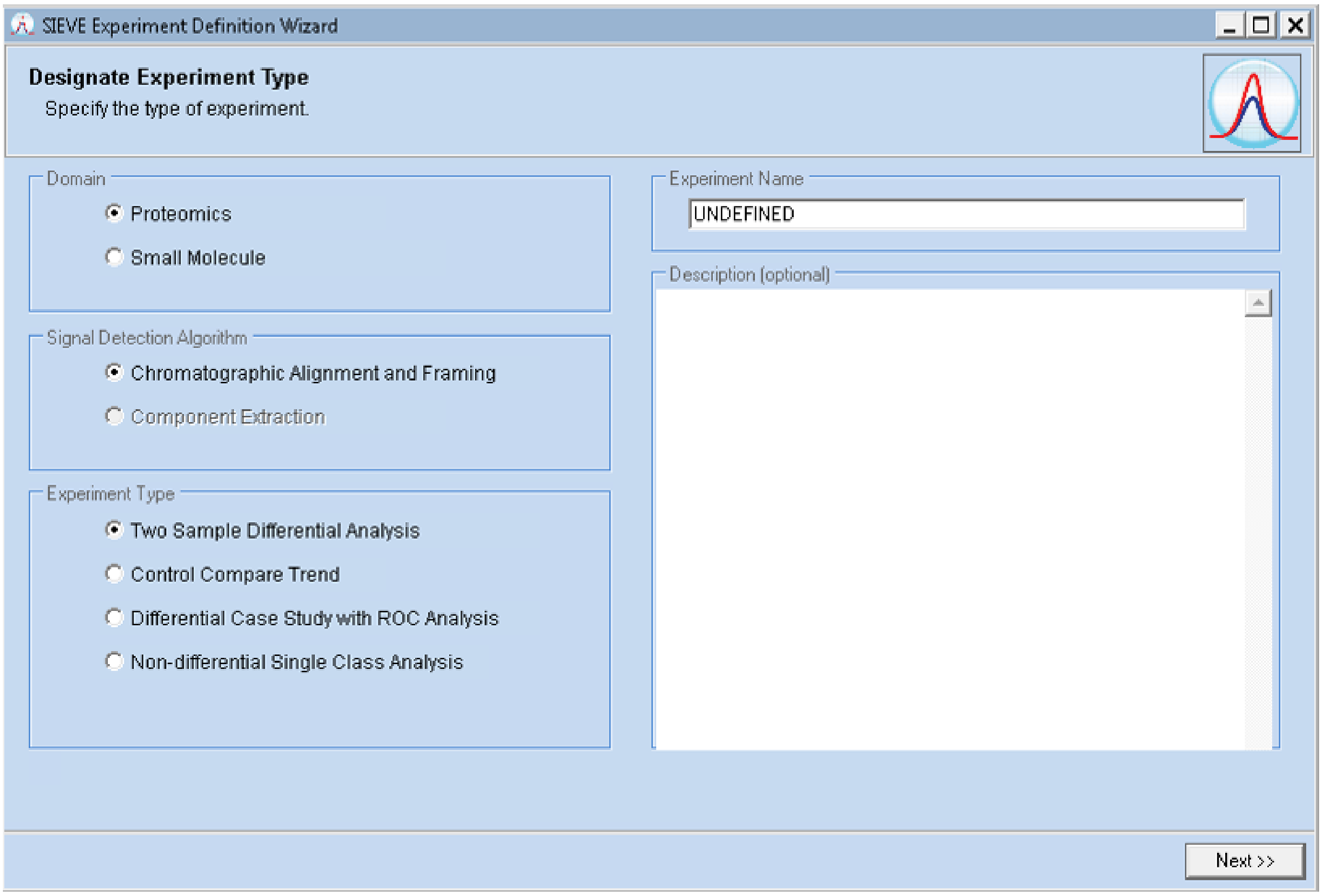
The currently protein identification workflow used by our group features database searching by MS-GF+, post-search processing by IDPicker, and spectrum report generation by IonStarSPG.R. Detailed instructions can be found here.

Step 2: Generation of quantitative features by SIEVE^{TM}

Quantitative feature generation in IonStar is accomplished by SIEVE ^{TM} v2.2 SP2 (Thermo Scientific), which integrates ChromAlign for global 3-D chromatographic alignment and a direct ion current extraction (DICE) method for feature extraction.

1. Load rawfiles into SIEVE ^{TM}

To start the quantitative feature generation analysis, open SIEVE ^{TM} and select **File -> Create new experiment**. On the **Designate Experiment Type** page, select the Experiment Type based on the study. For a case-control experiment, use **Two Sample Differential Analysis**; for multi-condition experiment (3 or more conditions including control), use **Control Compare Trend**.



SIEVE Experiment Definition Wizard

Designate Experiment Type

Specify the type of experiment.

Domain

● Proteomics

○ Small Molecule

Signal Detection Algorithm

● Chromatographic Alignment and Framing

○ Component Extraction

Experiment Type

● Two Sample Differential Analysis

○ Control Compare Trend

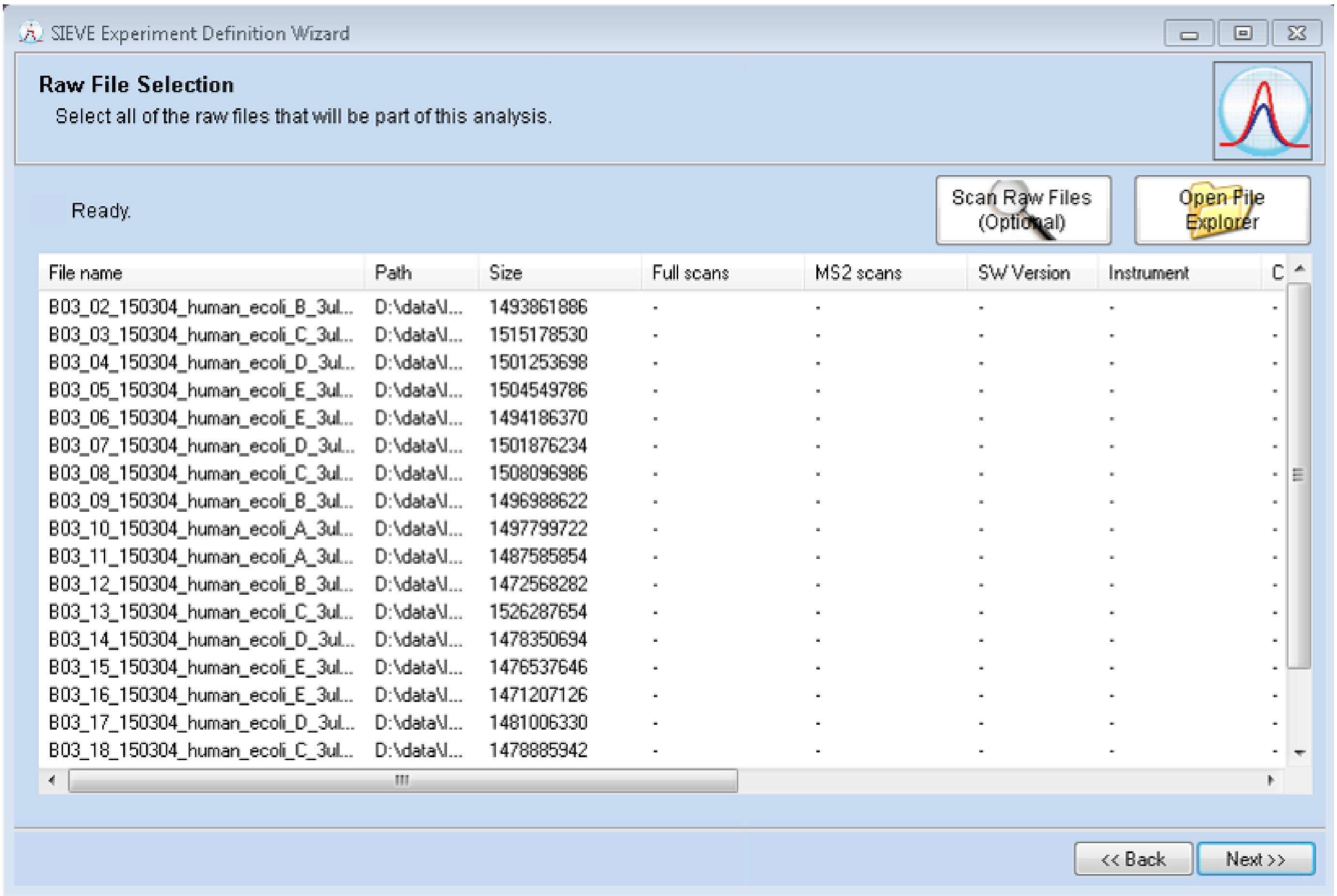
○ Differential Case Study with ROC Analysis

○ Non-differential Single Class Analysis

Experiment Name: UNDEFINED

Description (optional)

Drag all rawfiles into the **Raw File Selection** page.



SIEVE Experiment Definition Wizard

Raw File Selection

Select all of the raw files that will be part of this analysis.

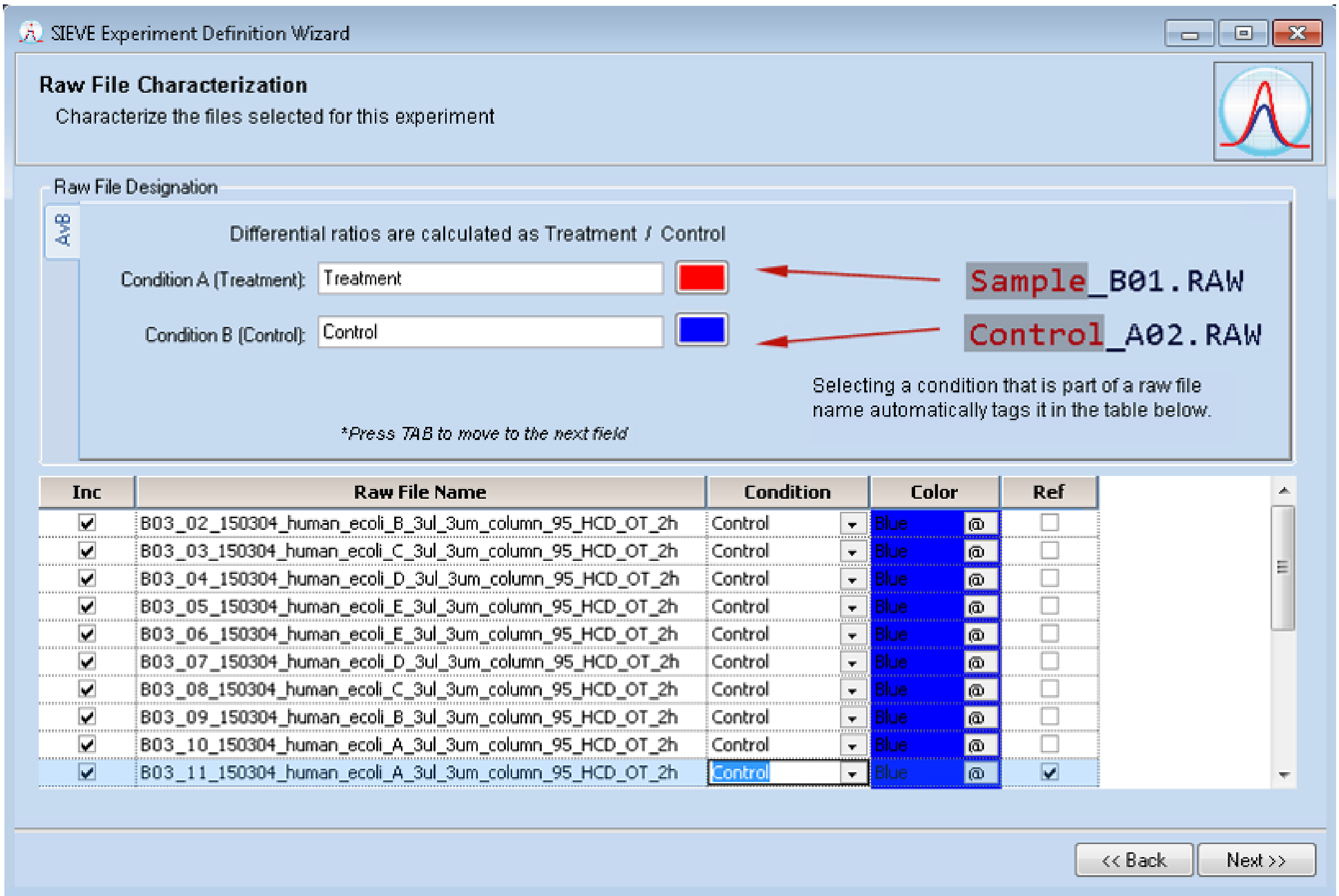
Ready.

File name	Path	Size	Full scans	MS2 scans	SW Version	Instrument	C
B03_02_150304_human_ecoli_B_3ul...	D:\data\...	1493861886	-	-	-	-	-
B03_03_150304_human_ecoli_C_3ul...	D:\data\...	1515178530	-	-	-	-	-
B03_04_150304_human_ecoli_D_3ul...	D:\data\...	1501253698	-	-	-	-	-
B03_05_150304_human_ecoli_E_3ul...	D:\data\...	1504549786	-	-	-	-	-
B03_06_150304_human_ecoli_E_3ul...	D:\data\...	1494186370	-	-	-	-	-
B03_07_150304_human_ecoli_D_3ul...	D:\data\...	1501876234	-	-	-	-	-
B03_08_150304_human_ecoli_C_3ul...	D:\data\...	1508096986	-	-	-	-	-
B03_09_150304_human_ecoli_B_3ul...	D:\data\...	1496988622	-	-	-	-	-
B03_10_150304_human_ecoli_A_3ul...	D:\data\...	1497799722	-	-	-	-	-
B03_11_150304_human_ecoli_A_3ul...	D:\data\...	1487585854	-	-	-	-	-
B03_12_150304_human_ecoli_B_3ul...	D:\data\...	1472568282	-	-	-	-	-
B03_13_150304_human_ecoli_C_3ul...	D:\data\...	1526287654	-	-	-	-	-
B03_14_150304_human_ecoli_D_3ul...	D:\data\...	1478350694	-	-	-	-	-
B03_15_150304_human_ecoli_E_3ul...	D:\data\...	1476537646	-	-	-	-	-
B03_16_150304_human_ecoli_E_3ul...	D:\data\...	1471207126	-	-	-	-	-
B03_17_150304_human_ecoli_D_3ul...	D:\data\...	1481006330	-	-	-	-	-
B03_18_150304_human_ecoli_C_3ul...	D:\data\...	1478885942	-	-	-	-	-

2. Assign sample conditions and select reference file

For **Two Sample Differential Analysis**, assign *Condition A* and *Condition B* in the two boxes; For **Control Compare Trend**, put *all conditions* in the upper box and assign *the control condition* in the lower box.

Two Sample Differential Analysis:



Raw File Characterization

Characterize the files selected for this experiment

Raw File Designation

Differential ratios are calculated as Treatment / Control

Condition A (Treatment): Treatment 🟥

Condition B (Control): Control 🟦

Sample_B01.RAW

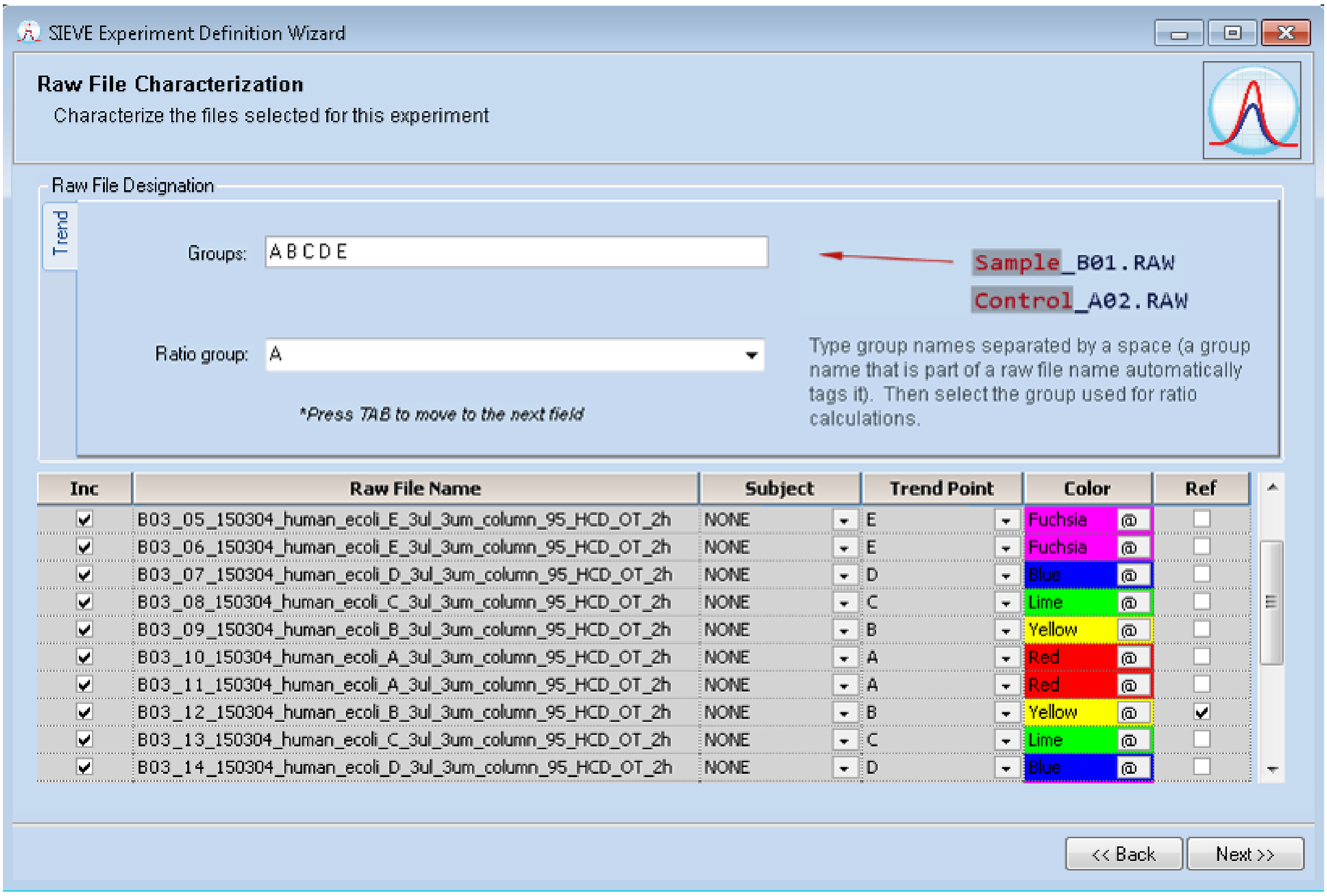
Control_A02.RAW

Selecting a condition that is part of a raw file name automatically tags it in the table below.

*Press TAB to move to the next field

Inc	Raw File Name	Condition	Color	Ref
☑	B03_02_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_03_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_04_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_05_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_06_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_07_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_08_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_09_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_10_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☐
☑	B03_11_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2h	Control	Blue @	☑

Control Compare Trend:



Raw File Characterization

Characterize the files selected for this experiment

Raw File Designation

Groups: A B C D E

Ratio group: A

Sample_B01.RAW

Control_A02.RAW

Type group names separated by a space (a group name that is part of a raw file name automatically tags it). Then select the group used for ratio calculations.

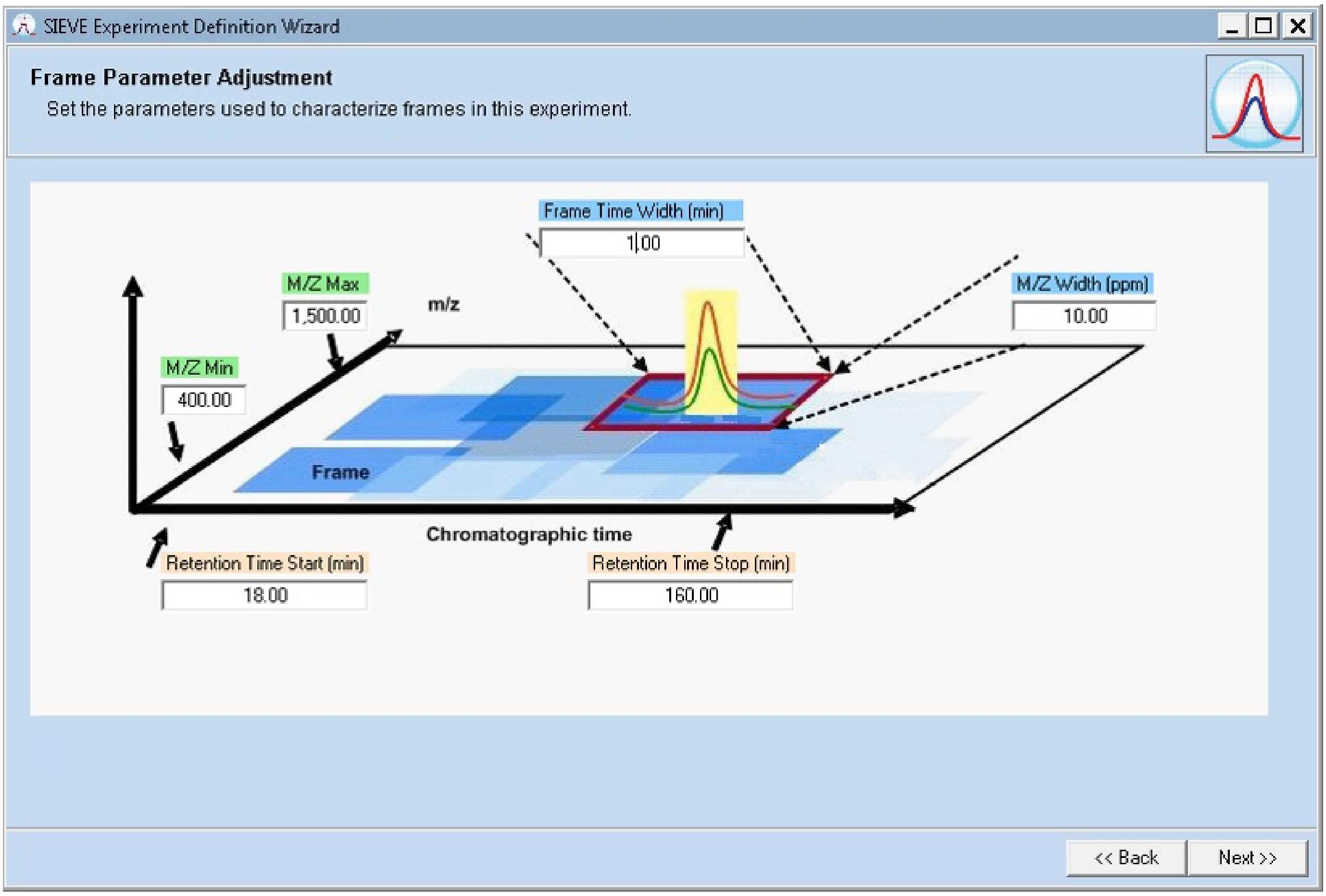
*Press TAB to move to the next field

Inc	Raw File Name	Subject	Trend Point	Color	Ref
☑	B03_05_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2h	NONE	E	Fuchsia @	☐
☑	B03_06_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2h	NONE	E	Fuchsia @	☐
☑	B03_07_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2h	NONE	D	Blue @	☐
☑	B03_08_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2h	NONE	C	Lime @	☐
☑	B03_09_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2h	NONE	B	Yellow @	☐
☑	B03_10_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2h	NONE	A	Red @	☐
☑	B03_11_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2h	NONE	A	Red @	☐
☑	B03_12_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2h	NONE	B	Yellow @	☑
☑	B03_13_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2h	NONE	C	Lime @	☐
☑	B03_14_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2h	NONE	D	Blue @	☐

A reference file also needs to be selected. In general, the reference file should provide the highest alignment scores for all sample runs.In most cases, it is recommended to start with a file in the middle of the LC-MS sequence as the reference.

3. Modify method parameters

The parameters that needs to be modified include **Frame Time Width (min)** and **M/Z Width (ppm)**. The current setting is based on **a 3-hr nano RPLC gradient** with **a Thermo Orbitrap instrument under 120K MS1 resolution**. Manual optimization based on the LC-MS method may help to improve the performance of feature generation. All other parameters follow the default settings.



SIEVE Experiment Definition Wizard

Frame Parameter Adjustment

Set the parameters used to characterize frames in this experiment.

Frame Time Width (min): 1.00

M/Z Max: 1,500.00

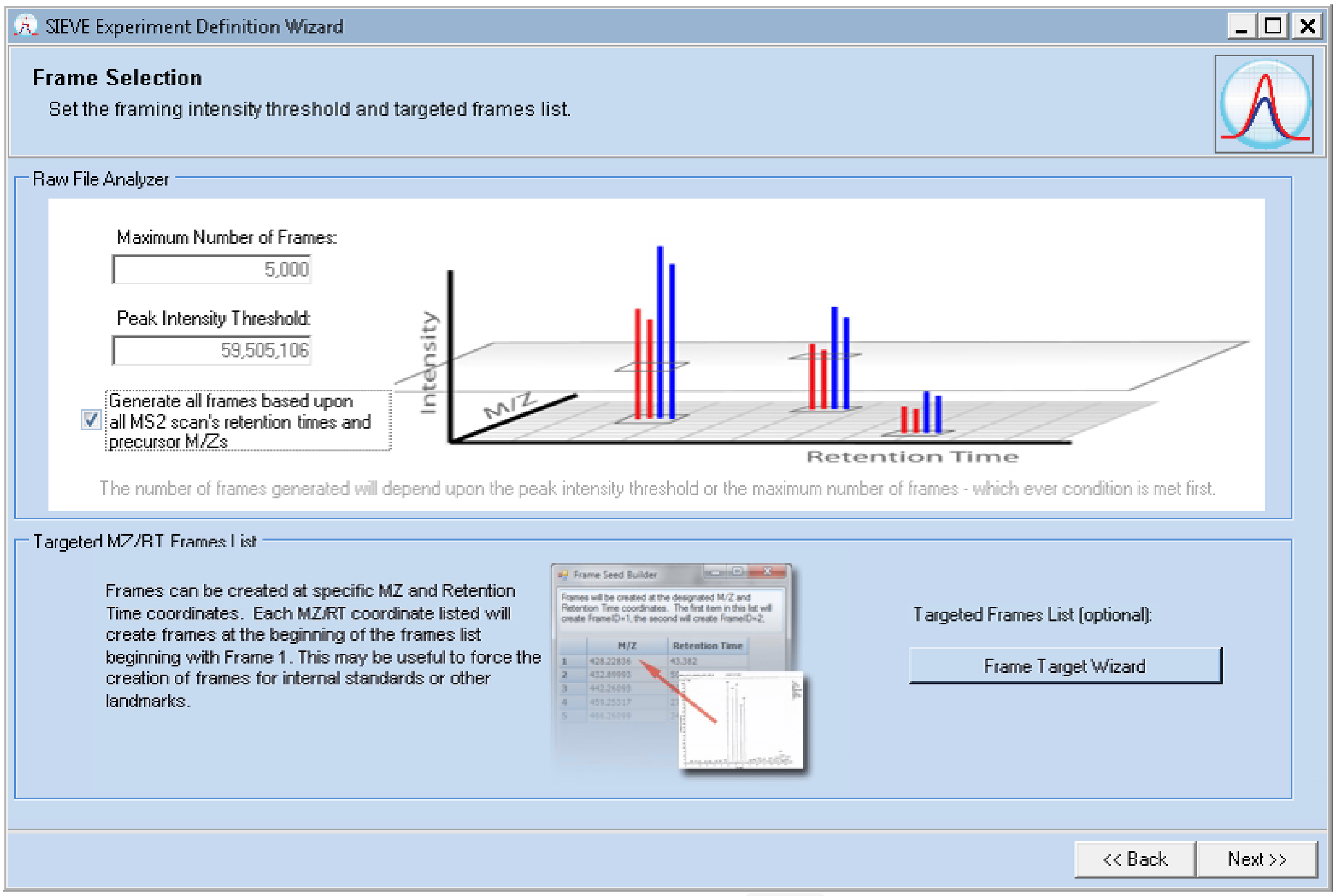
M/Z Min: 400.00

M/Z Width (ppm): 10.00

Retention Time Start (min): 18.00

Retention Time Stop (min): 160.00

Check **Generate all frames based upon all MS2 scan's retention times and precursor M/Zs** to maximize the number of quantitatve features. Alternatively, users can assign **Maximum Number of Frames** and **Peak Intensity Threshold**.



SIEVE Experiment Definition Wizard

Frame Selection

Set the framing intensity threshold and targeted frames list.

Raw File Analyzer

Maximum Number of Frames: 5,000

Peak Intensity Threshold: 59,505,106

☑ Generate all frames based upon all MS2 scan's retention times and precursor M/Zs

The number of frames generated will depend upon the peak intensity threshold or the maximum number of frames - which ever condition is met first.

Targeted MZ/RT Frames List

Frames can be created at specific MZ and Retention Time coordinates. Each MZ/RT coordinate listed will create frames at the beginning of the frames list beginning with Frame 1. This may be useful to force the creation of frames for internal standards or other landmarks.

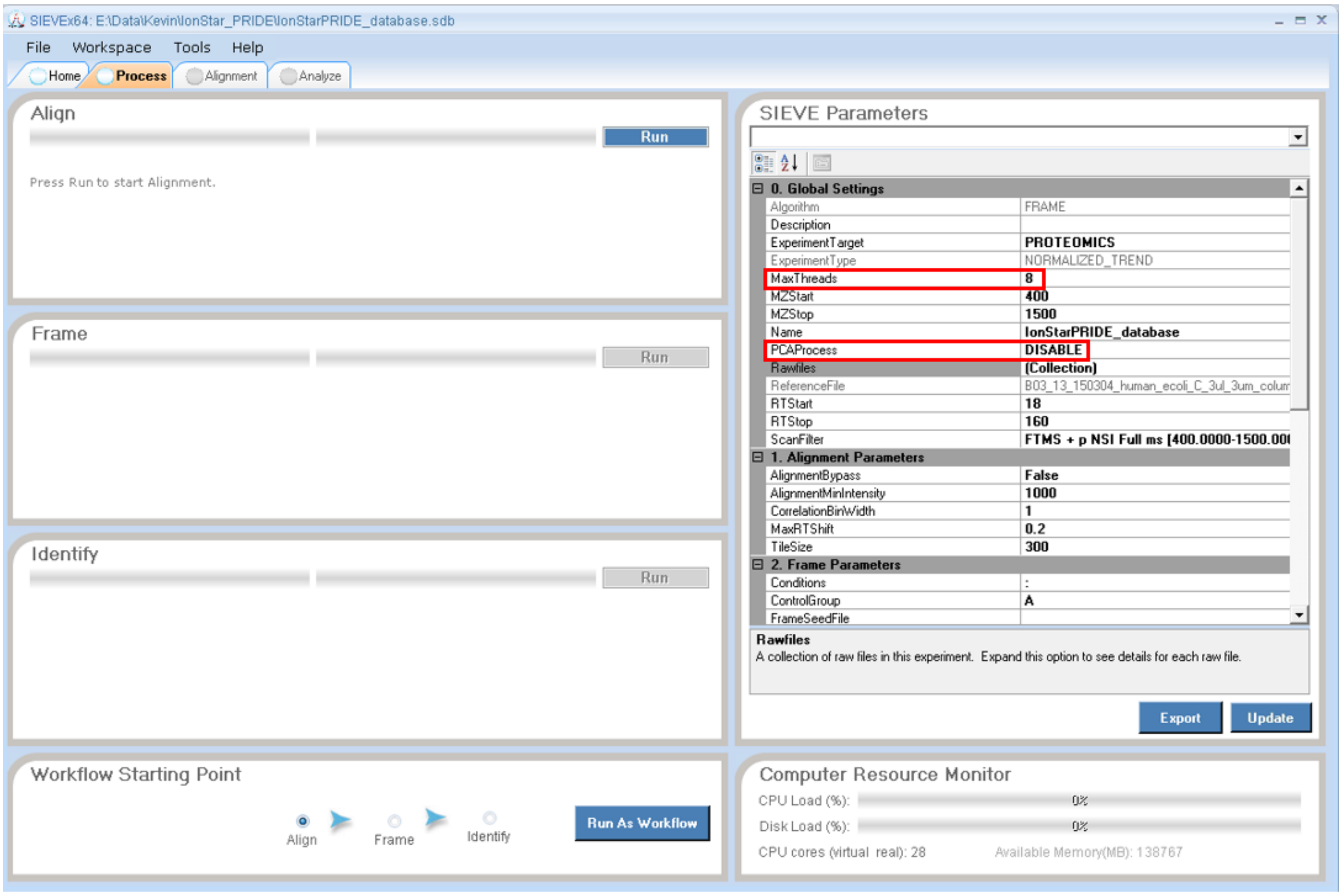
Targeted Frames List (optional):

Intensity
M/Z
Retention Time
Frame Seed Builder
Frames will be created at the designated M/Z and Retention Time coordinates. The first item in this list will create FrameID=1, the second will create FrameID=2.
M/Z
Retention Time
428.22836
43.382
432.89993
442.26093
459.25317
466.26099
Frame Target Wizard
<< Back
Next >>

After setting the method, finish the wizard and save the `.sdb` file.

4. Perform ChromAlign and DICE procedures

For IonStar, users do not need to run the **Identify** process. In the **SIEVE Parameters** window, **MaxThreads** should be changed according to the configuration of the computer used for SIEVE ™. For example, 6~8 threads are recommended for a PC with 16-core processors and 192 GB RAM. Occasionally, **PCAProcess** can also be disabled to alleviate computational burden. Click the **Update** button to save the settings. Run **Align** (ChromAlign) first.



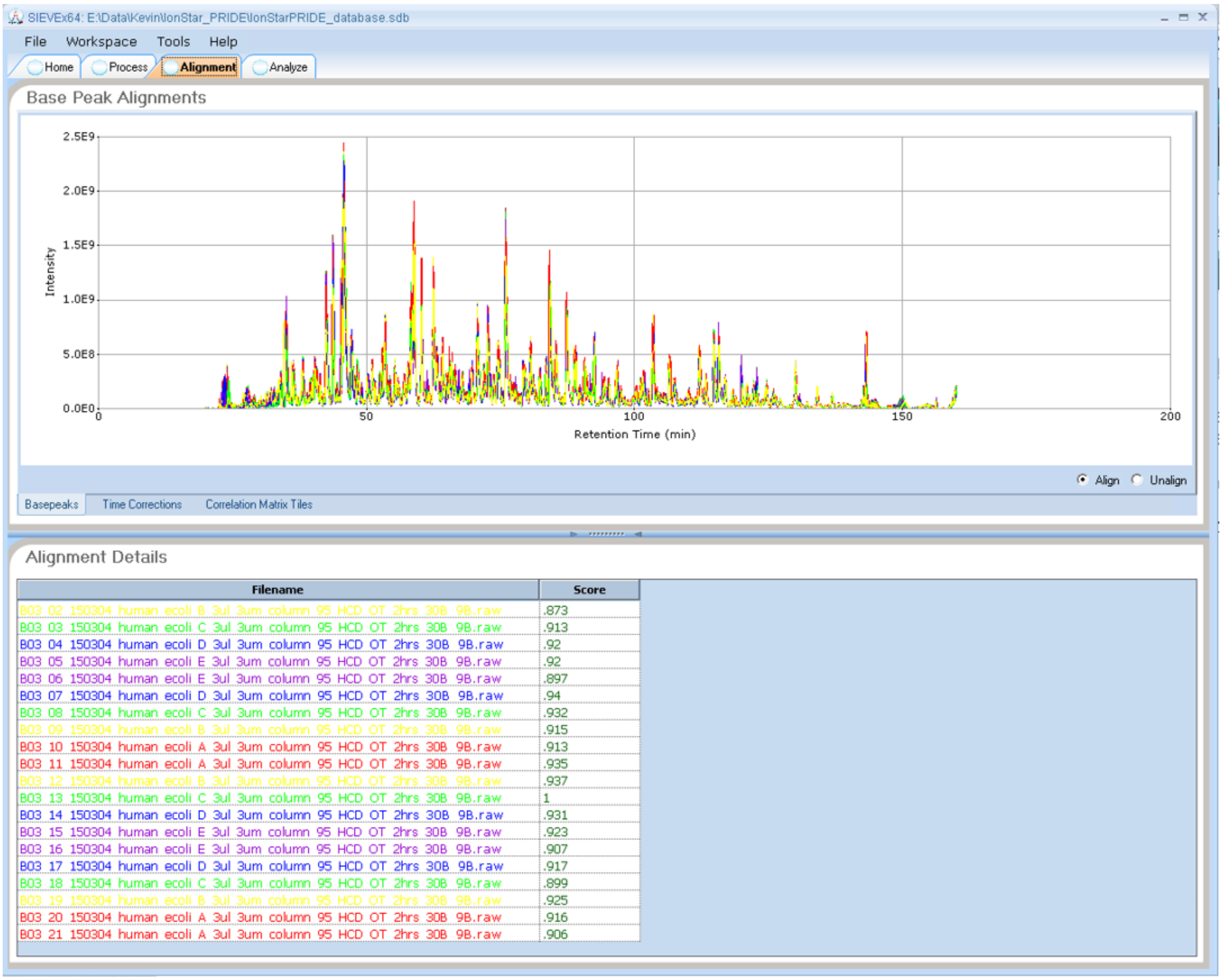
Press Run to start Alignment.

0. Global Settings	
Algorithm	FRAME
Description	
ExperimentTarget	PROTEOMICS
ExperimentType	NORMALIZED_TREND
MaxThreads	8
MZStart	400
MZStop	1500
Name	IonStarPRIDE_database
PCAProcess	DISABLE
Rawfiles	(Collection)
ReferenceFile	B03_13_150304_human_ecoli_C_3ul_3um_colum
RTStart	18
RTStop	160
ScanFilter	FTMS + p NSI Full ms [400.0000-1500.00
1. Alignment Parameters	
AlignmentBypass	False
AlignmentMinIntensity	1000
CorrelationBinWidth	1
MaxRTShift	0.2
TileSize	300
2. Frame Parameters	
Conditions	:
ControlGroup	A
FrameSeedFile	

Rawfiles
A collection of raw files in this experiment. Expand this option to see details for each raw file.

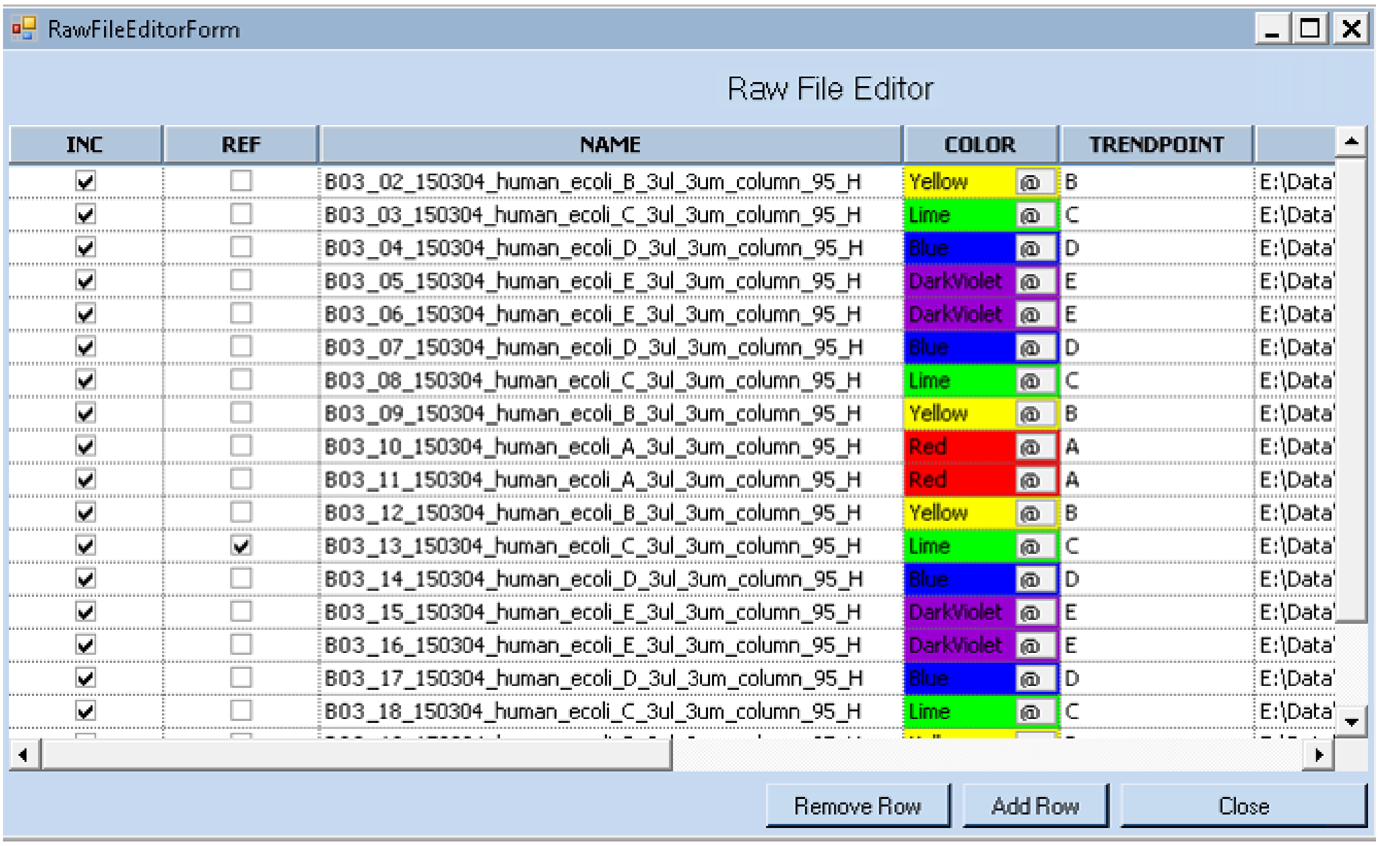
CPU Load (%): 0%
Disk Load (%): 0%
CPU cores (virtual real): 28
Available Memory(MB): 138767

Upon finishing, alignment scores for all sample runs will be shown in the **Alignment** tab. Ideally, the majority of sample runs should have an alignment score of **>0.8** to ensure the quality of quantitative feature generation. Change the reference file and rerun the ChromAlign process if the alignment scores are subpar (*e.g.* <0.7) for a large portion of the files.



Filename	Score
B03 02 150304 human ecoli B 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.873
B03 03 150304 human ecoli C 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.913
B03 04 150304 human ecoli D 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.92
B03 05 150304 human ecoli E 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.92
B03 06 150304 human ecoli E 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.897
B03 07 150304 human ecoli D 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.94
B03 08 150304 human ecoli C 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.932
B03 09 150304 human ecoli B 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.915
B03 10 150304 human ecoli A 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.913
B03 11 150304 human ecoli A 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.935
B03 12 150304 human ecoli B 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.937
B03 13 150304 human ecoli C 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	1
B03 14 150304 human ecoli D 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.931
B03 15 150304 human ecoli E 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.923
B03 16 150304 human ecoli E 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.907
B03 17 150304 human ecoli D 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.917
B03 18 150304 human ecoli C 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.899
B03 19 150304 human ecoli B 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.925
B03 20 150304 human ecoli A 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.916
B03 21 150304 human ecoli A 3ul 3um column 95 HCD OT 2hrs 30B 9B.raw	.906

To change the reference file, click the “...” button in the **Rawfiles** line. Change the reference file by checking a new rawfile. Rerun **Align** and check the alignment scores again. When finished, run **Frame** to perform the DICE process.



INC	REF	NAME	COLOR	TRENDPOINT	
☑	☐	B03_02_150304_human_ecoli_B_3ul_3um_column_95_H	Yellow @	B	E:\Data'
☑	☐	B03_03_150304_human_ecoli_C_3ul_3um_column_95_H	Lime @	C	E:\Data'
☑	☐	B03_04_150304_human_ecoli_D_3ul_3um_column_95_H	Blue @	D	E:\Data'
☑	☐	B03_05_150304_human_ecoli_E_3ul_3um_column_95_H	DarkViolet @	E	E:\Data'
☑	☐	B03_06_150304_human_ecoli_E_3ul_3um_column_95_H	DarkViolet @	E	E:\Data'
☑	☐	B03_07_150304_human_ecoli_D_3ul_3um_column_95_H	Blue @	D	E:\Data'
☑	☐	B03_08_150304_human_ecoli_C_3ul_3um_column_95_H	Lime @	C	E:\Data'
☑	☐	B03_09_150304_human_ecoli_B_3ul_3um_column_95_H	Yellow @	B	E:\Data'
☑	☐	B03_10_150304_human_ecoli_A_3ul_3um_column_95_H	Red @	A	E:\Data'
☑	☐	B03_11_150304_human_ecoli_A_3ul_3um_column_95_H	Red @	A	E:\Data'
☑	☐	B03_12_150304_human_ecoli_B_3ul_3um_column_95_H	Yellow @	B	E:\Data'
☑	☑	B03_13_150304_human_ecoli_C_3ul_3um_column_95_H	Lime @	C	E:\Data'
☑	☐	B03_14_150304_human_ecoli_D_3ul_3um_column_95_H	Blue @	D	E:\Data'
☑	☐	B03_15_150304_human_ecoli_E_3ul_3um_column_95_H	DarkViolet @	E	E:\Data'
☑	☐	B03_16_150304_human_ecoli_E_3ul_3um_column_95_H	DarkViolet @	E	E:\Data'
☑	☐	B03_17_150304_human_ecoli_D_3ul_3um_column_95_H	Blue @	D	E:\Data'
☑	☐	B03_18_150304_human_ecoli_C_3ul_3um_column_95_H	Lime @	C	E:\Data'

After feature generation, the `.sdb` file will contain all quantitative features (*i.e.* frames) generated. For more detailed information about the use of SIEVE, please refer to SIEVE User Guide.

Step 3: Data integration and quantification

After protein identification and quantitative feature generation, the R package **IonStarStat** will be utlized to integrate the spectrum report with the quantitative feature list and generate the final quantitative results. Procedures in this step include:

- Generation of the annotated frame list
- Removal of redundant quantitative features
- Frame-to-peptide aggregation & data normalization
- Multivariate mean variation-based outlier detection
- Shared peptide removal (optional)
- Peptide-to-protein aggregation

The codes for this step are enclosed in IonStar_Run.R.

1. Generate the annotated frame list

In the spectrum report, the **rawfile name** column (`sp_col[1]`) should only contain the file name with no extension (*e.g.* II_B03_21_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2hrs_30B_9B), and the **MS2 scan number** should be numeric (*e.g.* 58143).

Use the following codes to generate **the annotated frame list** and **the sample list**, which are both required for subsequent protein quantification. Make sure that the following packages are installed by running `install.packages(c("XLConnect","RSQLite"))`.

```
##Generate the annotated frame list
db <- "IonStarPRIDE_database.sdb" ##File name of the SIEVE database
sp <- "IonStarPRIDE_spectrum report.csv" ##File name of the spectrum report
col_filename <- 4 ##Column number for rawfile name
col_scannum <- 17 ##Column number for MS2 scan number
col_framelist <- c(6,18) ##Column numbers for Protein accession number and Peptide sequence
framelist <- "IonStarPRIDE_frame.csv" ##File name of the annotated frame list (output1)
sampleid <-"IonStarPRIDE_sampleid.csv" ##File name of the sample list (output2)
source ("IonStar_FrameGen.R")
```

The annotated frame list `.csv` generated consists of **Protein accession number**, **Peptide sequence**, **Frame ID**, and **corresponding quantitative values in each sample**, shown as below.

```
##            ProteinAC             PepSeq FrameID        A1          B1
## 1 Q96I51:WBS16_HUMAN EAAEAEAEVPVVQYVGER   35199   5452494    475886.4
## 2  P0C8J6:GATY_ECOLI           INVATELK   11407 216541745 262224407.0
## 3  P0C8J6:GATY_ECOLI      NYLTEHPEATDPR    6302  50365797  52927424.6
## 4  P0C8J6:GATY_ECOLI QWVNLPLVLHGASGLSTK   47084  13635817  18975922.8
## 5  P0C8J6:GATY_ECOLI QWVNLPLVLHGASGLSTK   85743 158317760 128711136.1
## 6  P0C8J6:GATY_ECOLI  SVMIDASHLPFAQNISR   41490  47259460  47781835.2
##          C1        D1        E1        E2        D2        C2        B2
## 1   1391912   2289193   1302592   1146268   1645735   1091675   2789563
## 2 342242173 411246895 481214021 451974788 394233403 304898893 251107764
## 3  60233419  74382575  92162468  94188976  75480174  50618895  41996791
## 4  21041386  31074728  39476379  37953565  28216572  22198631  13746699
## 5 113131881 116859175 113338204 107014155 112368481 114584694 118152103
## 6  55798367  69234723  83597655  86098121  61897136  53934485  40421413
##            A2        A3        B3        C3        D3        E3        E4
## 1    660561.7   1583813   5148398   1923703   3842454   4221733   3554227
## 2 175357014.9 207088734 254629839 330040590 391575868 494867018 491473921
## 3  28891661.5  52817845  76774606  62542501 131993194 166741064 171442525
## 4   8390298.0  12464382  29358146  21184939  50791891  66393810  63215742
## 5 121447044.0 155394791 156556630 115749102 144621808 136891875 141464442
## 6  28736591.6  48176223  69299555  52708139 116873607 142661980 134440236
##          D4        C4        B4        A4
## 1   3582159   4278508   5723243   3780455
## 2 495393627 389784097 318133056 216134657
## 3 144341093 110284315  81203507  59372919
## 4  48868329  35728182  22188573  13603490
## 5 140116613 143775988 150280000 149897829
## 6 110236215  85939761  63125621  46183524
```

2. Perform protein quantification

Before running the R codes, modify **the sample list** so that each sample is assigned a **GroupID**. **GroupID** can be any combinations of alphabetic and numeric symbols, *e.g. A*, *Group1*, *088714*.

```
##                                                                RawFiles
## 1  II_B03_21_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 2  II_B03_02_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 3  II_B03_03_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 4  II_B03_04_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 5  II_B03_05_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 6  II_B03_06_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 7  II_B03_07_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 8  II_B03_08_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 9  II_B03_09_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 10 II_B03_10_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 11 II_B03_11_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 12 II_B03_12_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 13 II_B03_13_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 14 II_B03_14_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 15 II_B03_15_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 16 II_B03_16_150304_human_ecoli_E_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 17 II_B03_17_150304_human_ecoli_D_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 18 II_B03_18_150304_human_ecoli_C_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 19 II_B03_19_150304_human_ecoli_B_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
## 20 II_B03_20_150304_human_ecoli_A_3ul_3um_column_95_HCD_OT_2hrs_30B_9B
##    GroupID
## 1        A
## 2        B
## 3        C
## 4        D
## 5        E
## 6        E
## 7        D
## 8        C
## 9        B
## 10       A
## 11       A
## 12       B
## 13       C
## 14       D
## 15       E
## 16       E
## 17       D
## 18       C
## 19       B
## 20       A
```

Make sure to load `IonStarStat` by `library("IonStarstat")`. Read the annotated frame list and the grouped sample list into R environment.

```
rawfile <- "IonStarPRIDE_Frame.csv"
condfile <- "IonStarPRIDE_Groups.csv"
raw <- read.csv(rawfile)
cond <- read.csv(condfile)
condition <- cond[match(colnames(raw)[-c(1:3)], cond[,1]),2]
condition
```

```
##  [1] A B C D E E D C B A A B C D E E D C B A
## Levels: A B C D E
```

Use `newProDataSet` to remove redundant frames (*i.e.* frames assigned to multiple peptide sequences), which causes ambiguity in quantification.

```
pdata <- newProDataSet(proData=raw, condition=condition)
```

The number of proteins before and after removal, as well as the number of redundant frames removed will be reported in the console.

```
## Input 3886 proteins.
```

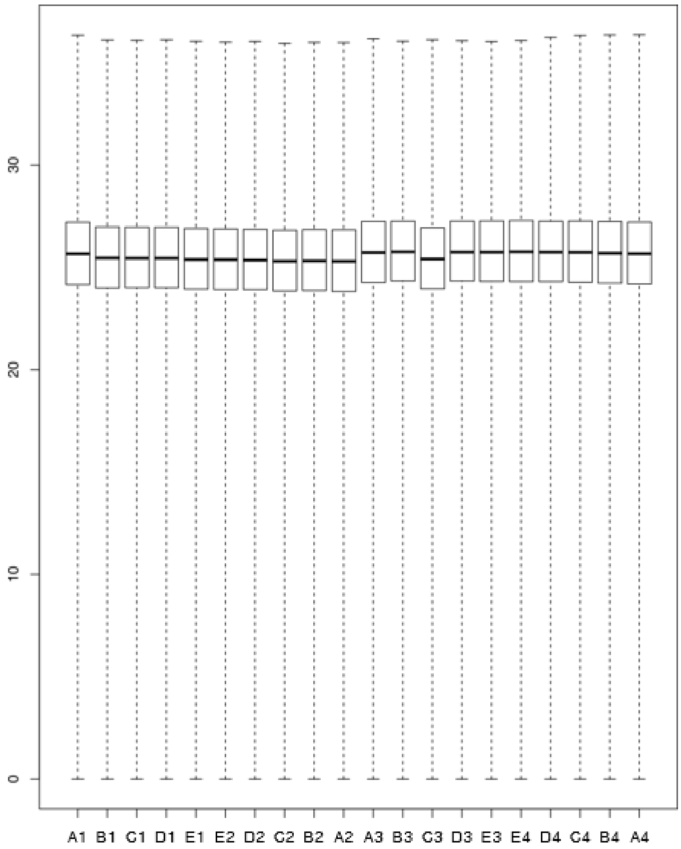
```
## 6489 duplicated frames founded.
```

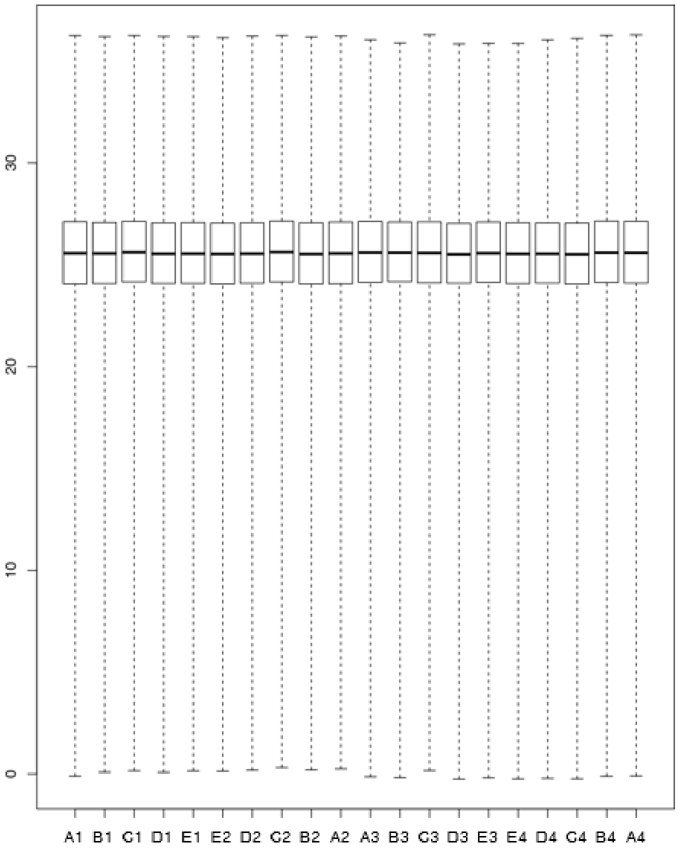
```
## 3873 proteins left after filtering.
```

Use `pnormalize` to perform inter-sample normalization of quantitative intensities. Aggregation of frame data to peptide data can be done by `summarize=TRUE`. Normalization can be based on either total ion intensities (`method="TIC"`) or quantiles (`method="quantiles"`) in each sample. Use `method=NULL` to skip normalization.

```
ndata <- pnormalize(pdata, summarize=TRUE, method="TIC")
```

Boxplots of peptide quantitative data before (left) and after (normalization) are shown as follows.





Use `OutlierPeptideRM` to perform outlier peptide detection. IonStar uses **Principal Component-based Outlier Detection (*PCOut*)** for outlier detection, which is tailored for multi-condition comparison (at least 3 conditions including control).

Parameter `variance` (0.7~0.9) can be adjusted according to the stringency needed for outlier detection. The higher the value the more outliers will be rejected.

```
cdata<-OutlierPeptideRM(ndata,condition,variance=0.7,critM1=1/3,critM2=1/4,ratio=TRUE)
```

```
## 6049 outliers were removed; 21937 peptides left after outlier removal.
```

For **case-control comparison**, set parameter `ratio=FALSE`. Alternatively, **Grubb's test** can be used for outlier rejection, which will be available in the next build of IonStarStat.

Use `SharedPeptideRM` to remove shared peptides (*i.e.* peptides inferred to multiple unique protein groups, *a.k.a.* degenerate peptides). This step is optional as many highly abundance proteins share a large proportion of homologous sequence domains. Removal of these peptides could be counterproductive for quantification. However, in specific cases, such as quantification of mixed-species samples, removal of shared peptides with species ambiguity is necessary to obtain species-specific quantitative results.

```
#Opional removal of shared peptides
cdata<-SharedPeptideRM(cdata)
```

Use `ProteinQuan` to aggregate peptide-level quantitative data to protein level. Both sum intensities (`method="sum"`) and General Linear Mixed Model (`method="fit"`) can be used for peptide-to-protein aggregation.

```
quan <- ProteinQuan(eset=cdata, method="sum")
```

```
##                      PepNum       A1       B1       C1       D1       E1
## A0AVT1:UBA6_HUMAN       4 26.62118 26.70643 26.70311 26.55632 26.56956
## A0FGR8:ESYT2_HUMAN     12 29.14639 29.14287 29.19418 29.11159 29.07409
## A0MZ66:SHOT1_HUMAN      8 27.38884 27.12556 27.21083 27.11330 27.08704
## A1L0T0:ILVBL_HUMAN      4 24.82774 25.34471 25.23324 25.22633 25.29648
## A1X283:SPD2B_HUMAN      4 25.91957 25.89851 25.98069 25.75741 25.62000
## A2RRP1:NBAS_HUMAN       2 23.21671 23.42673 23.27803 23.06164 22.72570
##                           E2       D2       C2       B2       A2       A3
## A0AVT1:UBA6_HUMAN  26.50505 26.59699 26.71673 26.65676 26.77142 26.80597
## A0FGR8:ESYT2_HUMAN 29.04383 29.11444 29.18659 29.18904 29.28187 29.19237
## A0MZ66:SHOT1_HUMAN 26.95478 27.11314 27.28457 27.07045 27.16919 27.32022
## A1L0T0:ILVBL_HUMAN 25.25879 25.39263 25.29545 25.22623 25.41492 25.25910
## A1X283:SPD2B_HUMAN 25.52778 25.70657 25.94787 25.87455 25.82754 26.14511
## A2RRP1:NBAS_HUMAN  22.98854 22.89805 23.27723 23.18188 23.39038 23.13157
##                           B3       C3       D3       E3       E4       D4
## A0AVT1:UBA6_HUMAN  26.63028 26.64539 26.37873 26.50122 26.37168 26.58727
## A0FGR8:ESYT2_HUMAN 29.11881 29.20611 28.87648 28.95956 28.84509 28.92555
## A0MZ66:SHOT1_HUMAN 27.35312 27.22668 27.28598 27.35093 27.21666 27.19223
## A1L0T0:ILVBL_HUMAN 24.90396 25.21406 24.71122 24.80994 24.84718 24.86370
## A1X283:SPD2B_HUMAN 25.99702 25.71892 25.62937 25.82062 25.72596 25.88044
## A2RRP1:NBAS_HUMAN  23.36878 23.10810 23.13342 22.86591 22.97998 23.19907
##                           C4       B4       A4
## A0AVT1:UBA6_HUMAN  26.50536 26.60836 26.72253
## A0FGR8:ESYT2_HUMAN 28.92343 29.11381 29.17440
## A0MZ66:SHOT1_HUMAN 27.25682 27.33590 27.39389
## A1L0T0:ILVBL_HUMAN 24.78118 24.96034 25.14340
## A1X283:SPD2B_HUMAN 25.80138 25.88904 26.01037
## A2RRP1:NBAS_HUMAN  22.89899 23.26405 23.07043
```

Users can export both peptide and protein quantitative results by `write.csv`.

```
write.csv(quan,"IonStarPRIDE_protein_quan.csv")
write.csv(exprs(cdata),"IonStarPRIDE_peptide_quan.csv")
```

Step 4: Post-quantification data processing

StarGazer, a Shiny-based interactive web app, will be made available in the next build of IonStar for post-quantification data processing. Fundamental functions of StarGazer include:

- Data cleanup and formatting
- Case-control protein ratio calculation

- Statistical testing
- Basic data mining (*e.g.* PCA, hierarchical clustering, fuzzy c-means clustering)
- Graphic depiction of quantitative data

Contact information

For questions, suggestions, and other topics about IonStarStat, feel feel to contact us:

Shichen Shen: shichens@buffalo.edu

Xue Wang: xwang79@buffalo.edu

Jun Qu: junqu@buffalo.edu

REREFERENCES

1. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R., An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2002,** 2, 513-23.
2. Filzmoser, P.; Maronna, R.; Werner, M., Outlier identification in high dimensions. *Computational Statistics & Data Analysis* **2008,** 52, 1694-1711.
3. Shen, X.; Shen, S.; Li, J.; Hu, Q.; Nie, L.; Tu, C.; Wang, X.; Orsburn, B.; Wang, J.; Qu, J., An IonStar Experimental Strategy for MS1 Ion Current-Based Quantification Using Ultrahigh-Field Orbitrap: Reproducible, In-Depth, and Accurate Protein Measurement in Large Cohorts. *Journal of Proteome Research* **2017,** 16, 2445-2456.
4. Kramer, A.; Green, J.; Pollard, J., Jr.; Tugendreich, S., Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2014,** 30, 523-30.