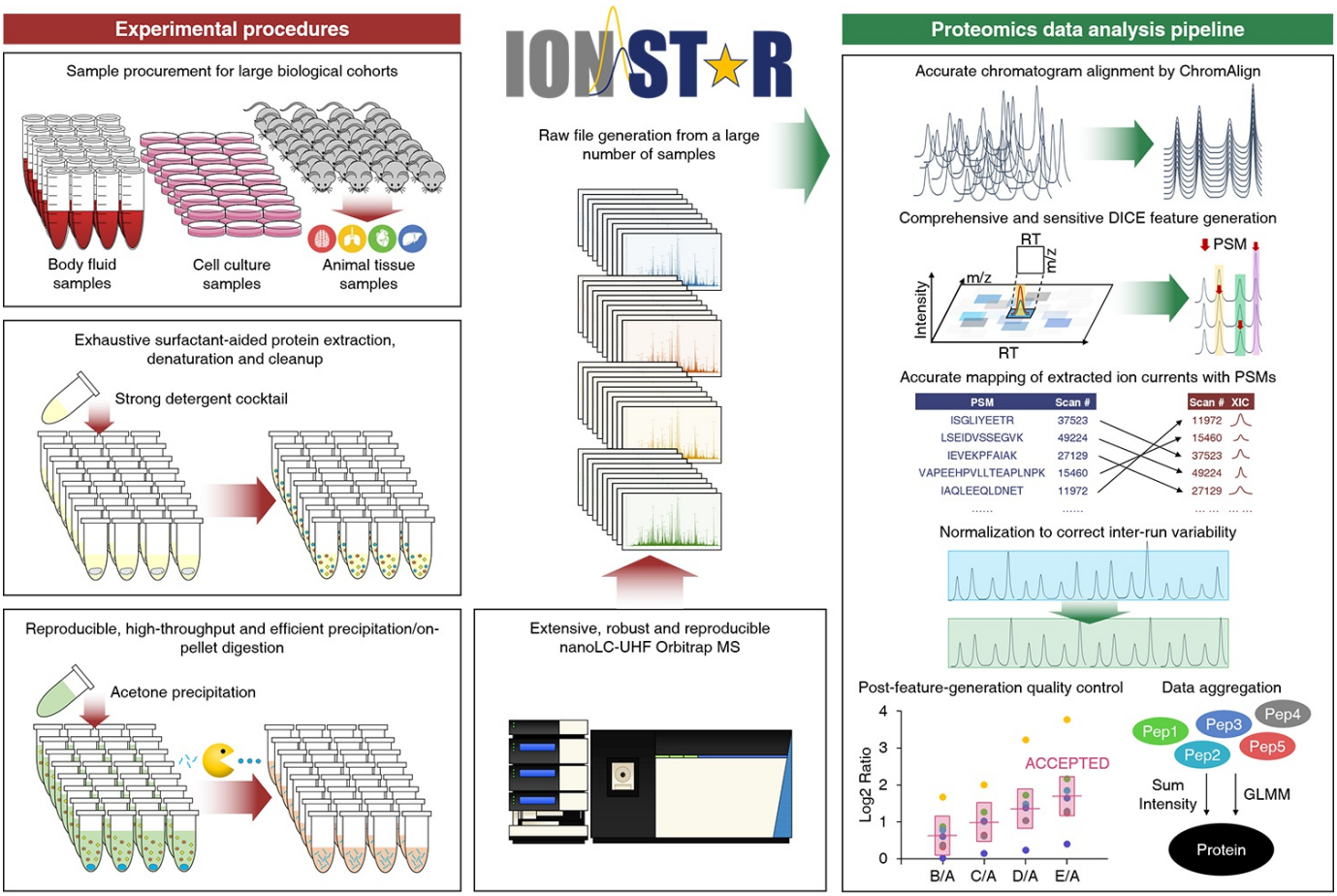
IonStar USER MANUAL

For Build 0.1.4

Introduction



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™** and **IonStarStat**.

SIEVE™ is a commercial software from Thermo Fisher Scientific. The latest version of SIEVE ™ is v2.2 SP2. Please contact Thermo Fisher Scientific regarding the quote for SIEVE™. To ensure of proper performance of SIEVE™, we recommend running SIEVE™ 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

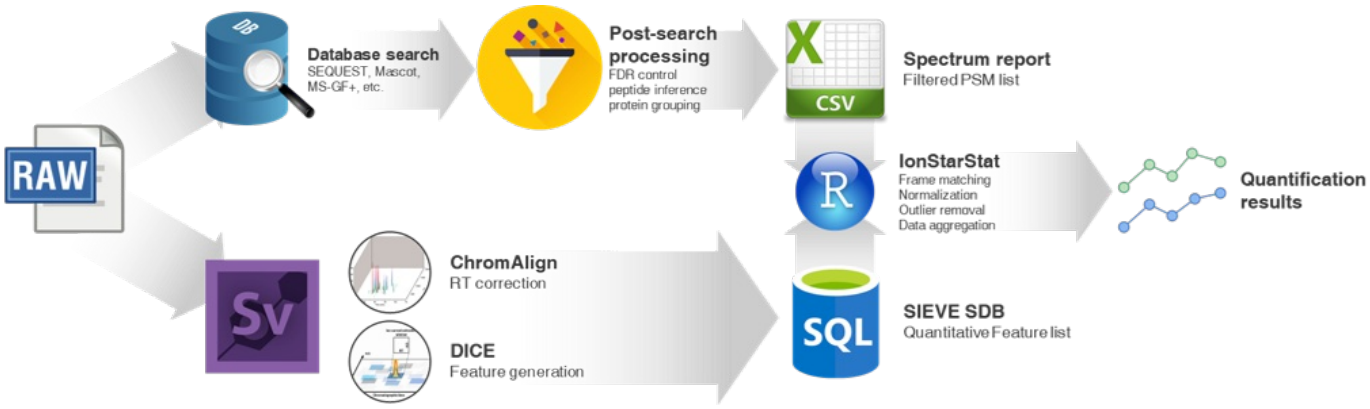
To perform using IonStar, it is recommended to put all files under the same working directory, including:

- 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`.

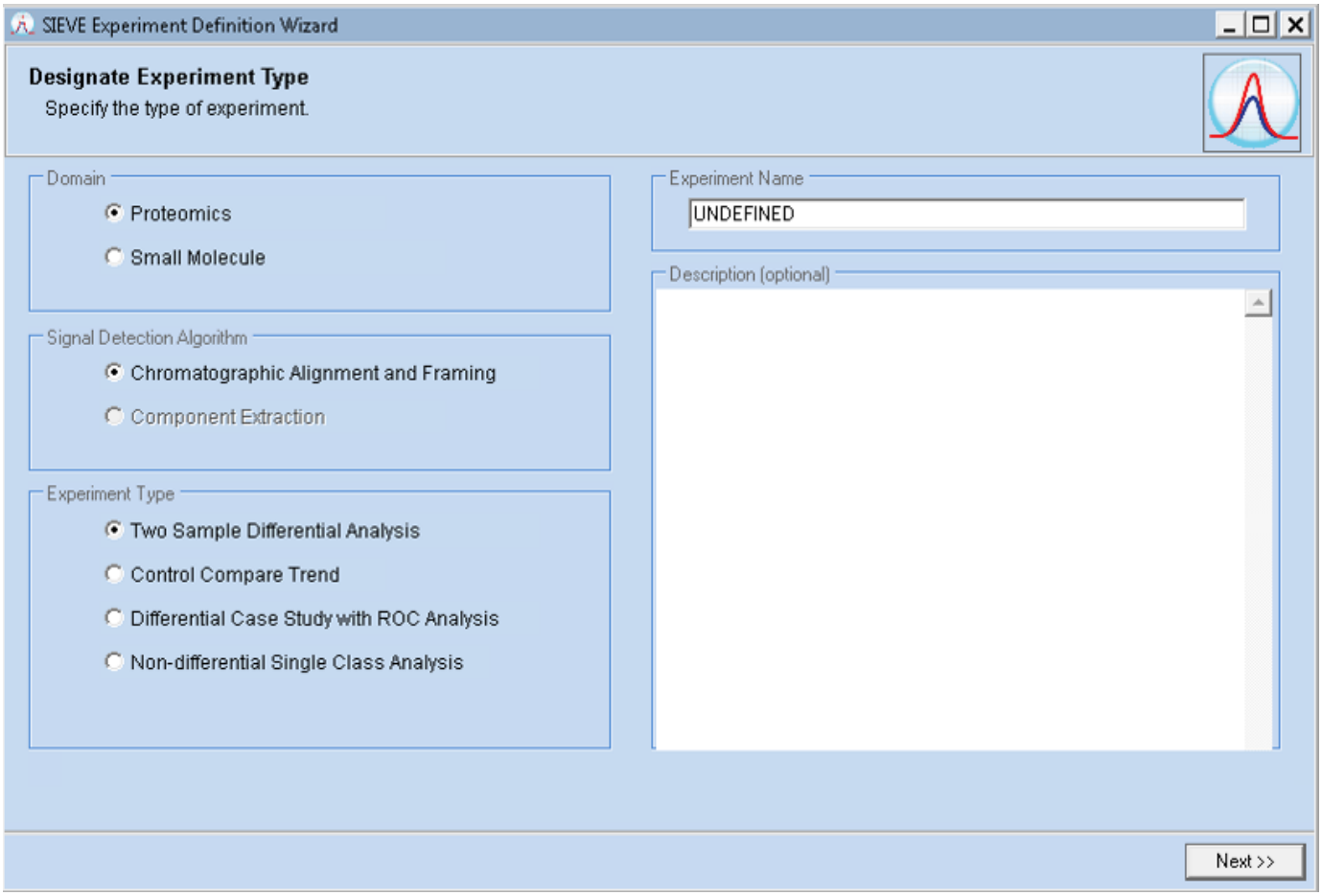
The currently protein identification workflow used by our group features database searching by MS-GF+, post-search processing by IDPicker, and spectrum report generation by IonStarSPG.R. Detailed instructions can be found here.

Step 2: Generation of quantitative features by SIEVE™

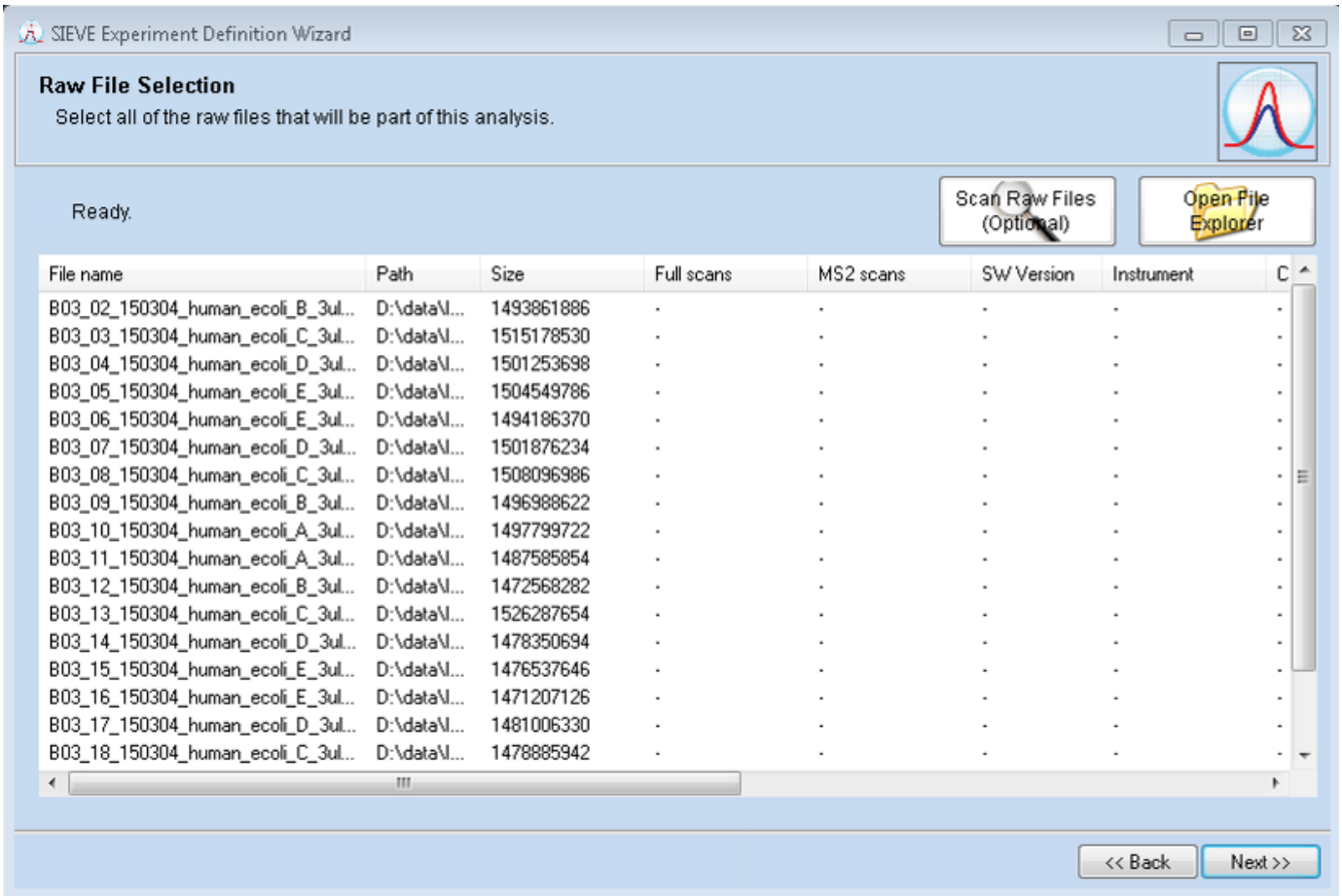
Quantitative feature generation in IonStar is accomplished by SIEVE ™ v2.2 SP2 (Thermo Scientific), which integrates ChromAlign for global 3-D chromatographic alignment and a direct ion current extraction (DICE) method for feature extraction.

1. Load rawfiles into SIEVE ™

To start the quantitative feature generation analysis, open SIEVE ™ 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**.



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

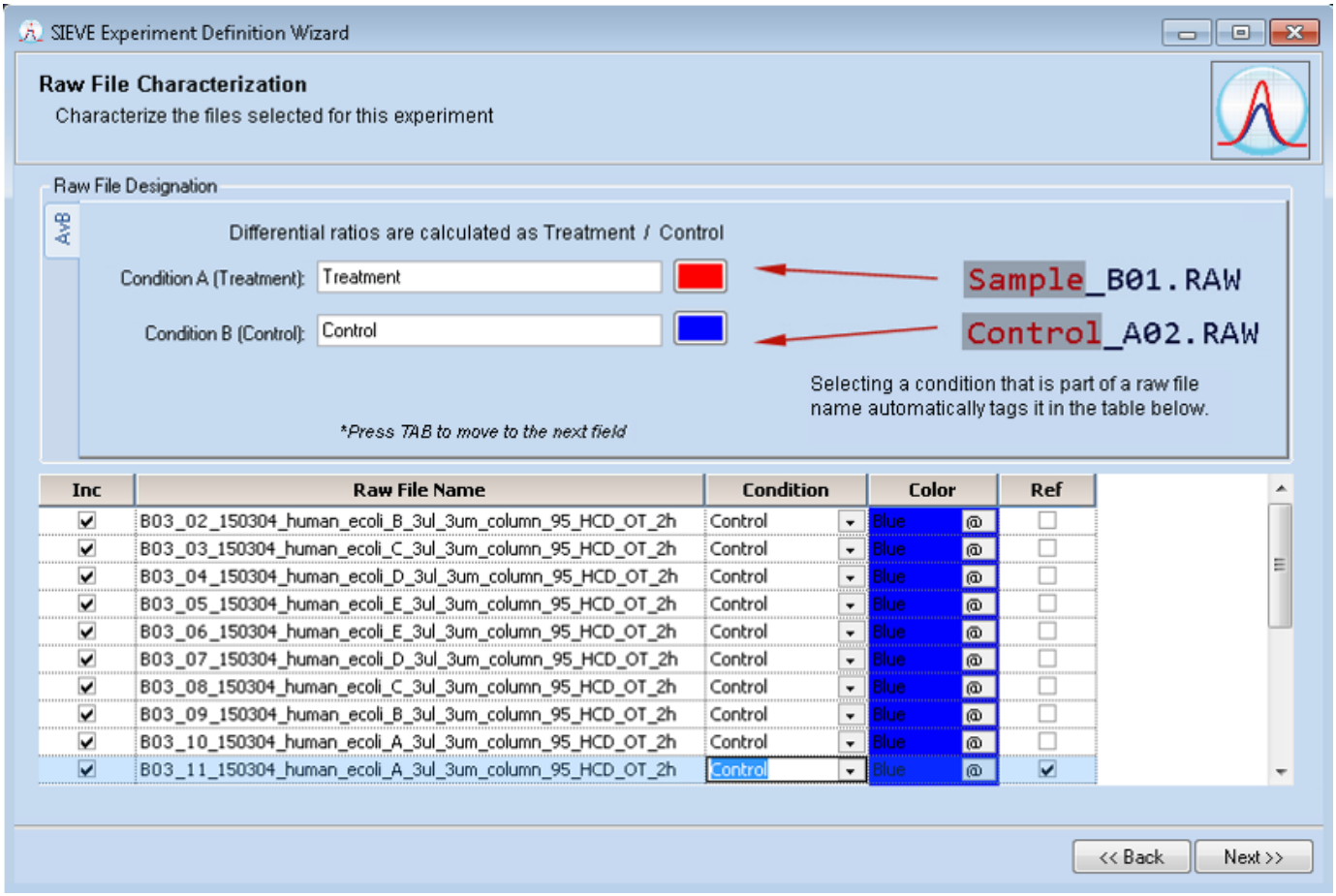


2. Assign sample conditions and select reference file

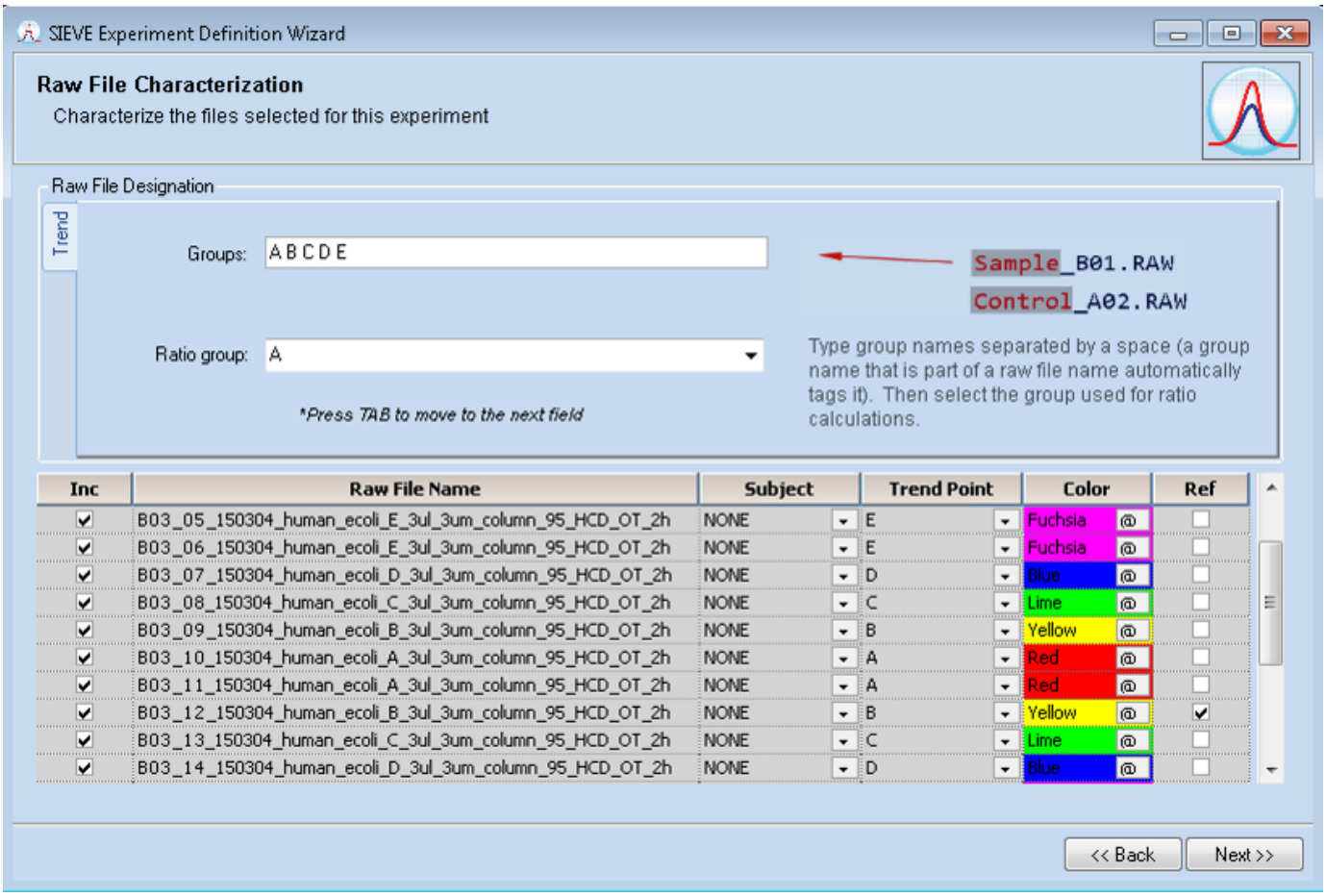
For **Two Sample Differential Analysis**, assign *Condition A* and *Condition B* in the two boxes;
For **Control Compare Trend**, put *all conditions* in the upper box and assign *the control condition*

in the lower box.

Two Sample Differential Analysis:



Control Compare Trend:

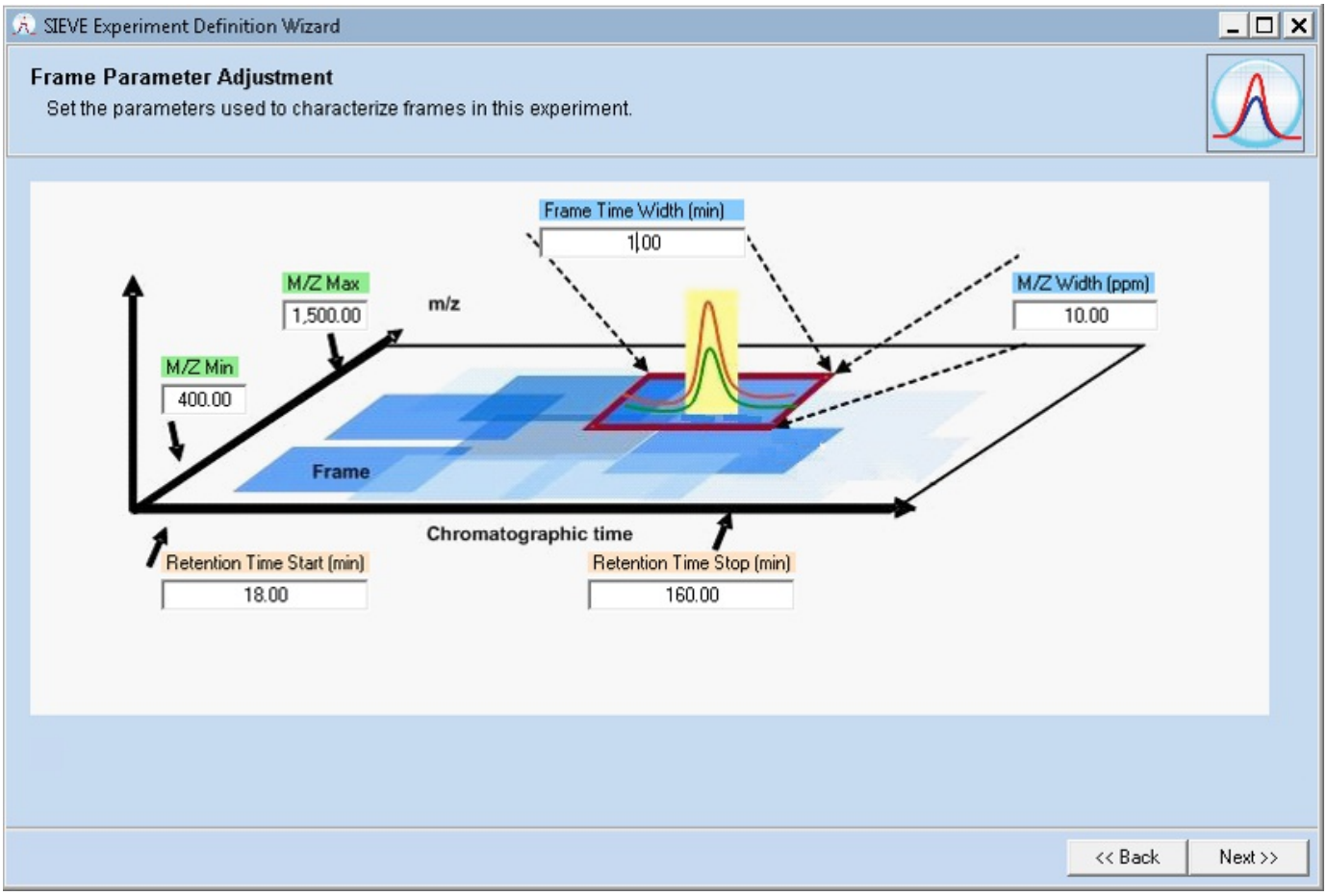


A reference file also needs to be selected. In general, the reference file should provide the highest

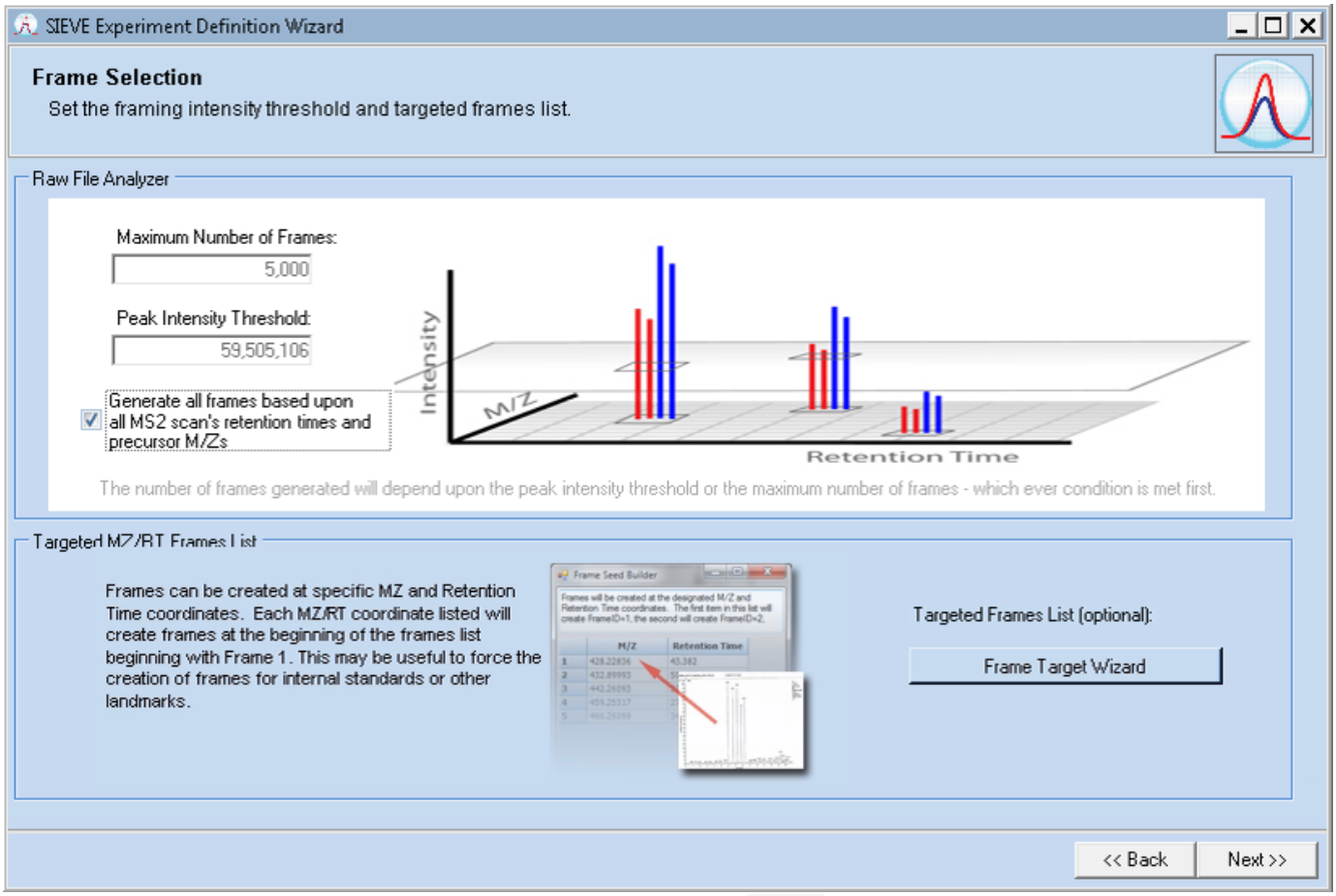
alignment scores for all sample runs.In most cases, it is recommended to start with a file in the middle of the LC-MS sequence as the reference.

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.



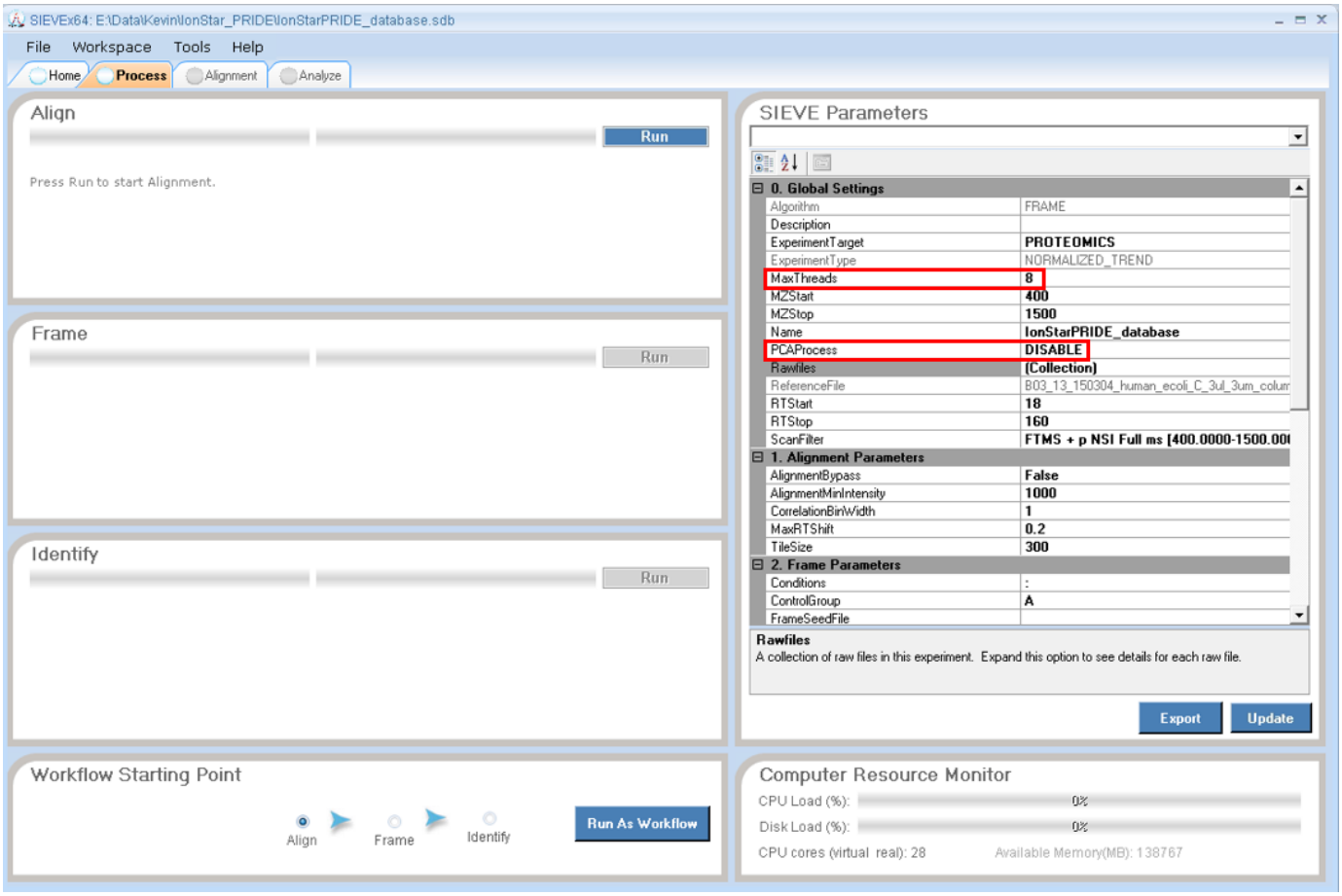
Check **Generate all frames based upon all MS2 scan's retention times and precursor M/Zs** to maximize the number of quantitatve features. Alternatively, users can assign **Maximum Number of Frames** and **Peak Intensity Threshold**.



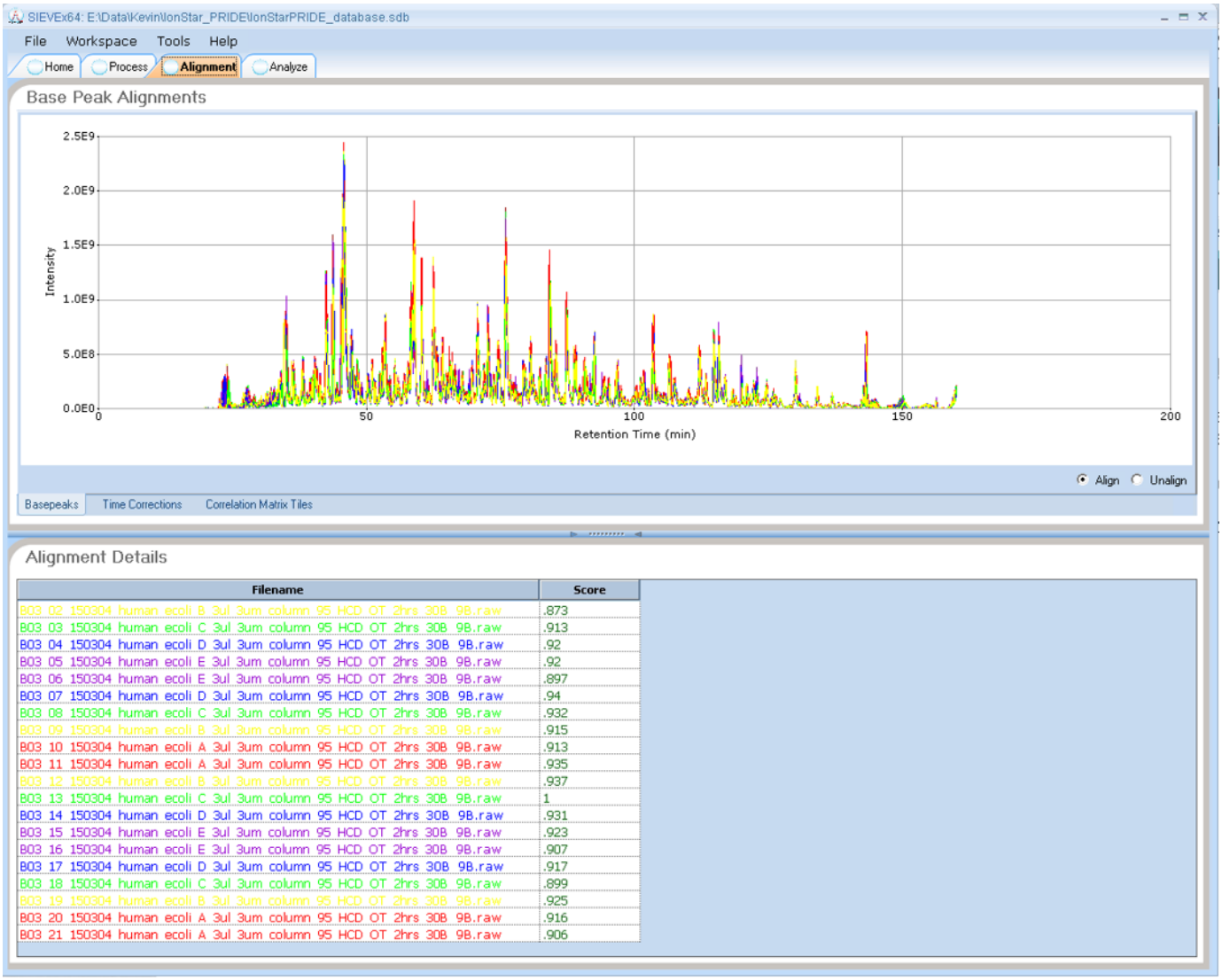
After setting the method, finish the wizard and save the `.sdb` file.

4. Perform ChromAlign and DICE procedures

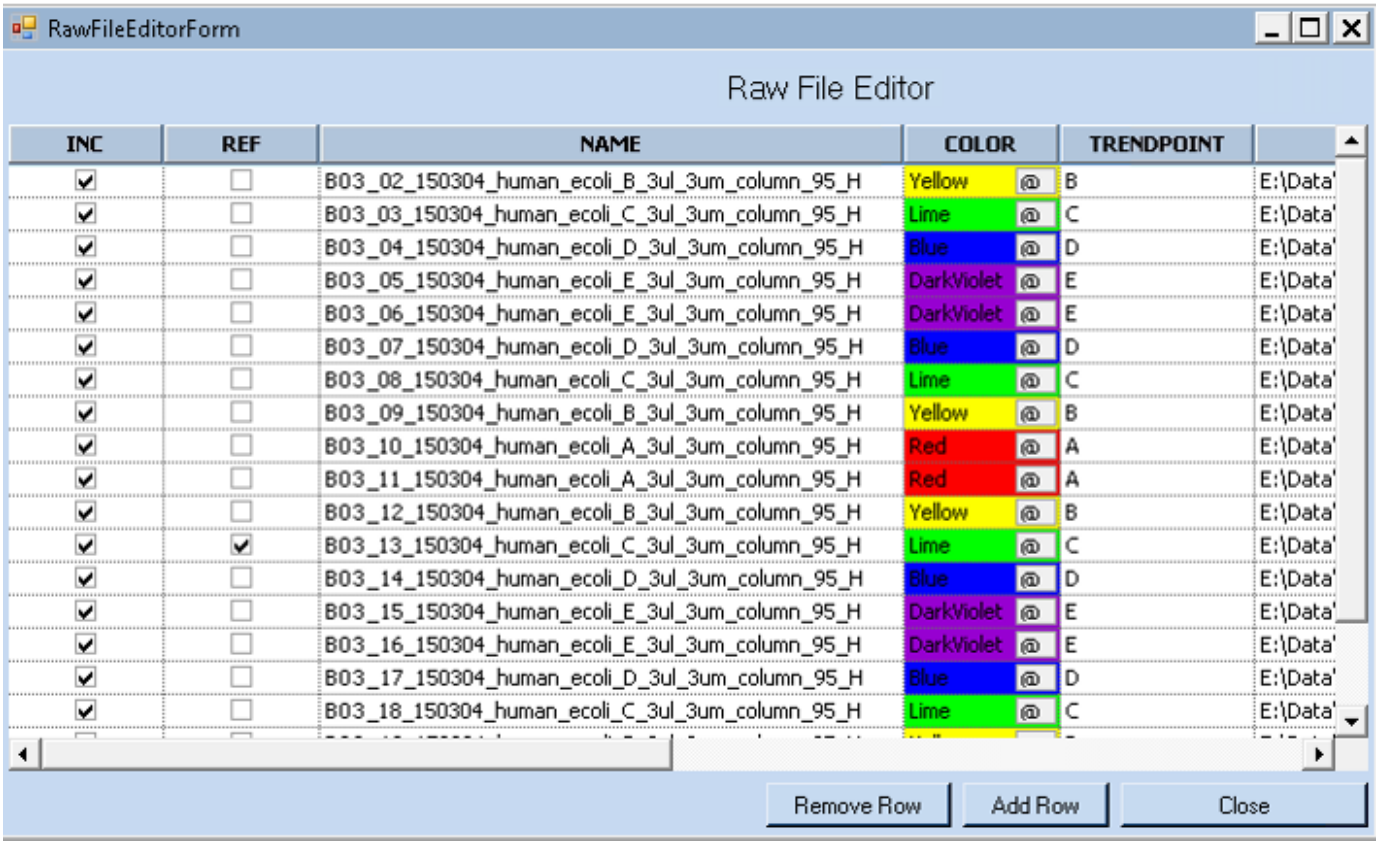
For IonStar, users do not need to run the **Identify** process. In the **SIEVE Parameters** window, **MaxThreads** should be changed according to the configuration of the computer used for SIEVE ™. For example, 6~8 threads are recommended for a PC with 16-core processors and 192 GB RAM. Occasionally,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



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.



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.



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 Pe
ptide sequence
framelist <- "IonStarPRIDE_frame.csv" ##File name of the annotated frame list
(output1)
sampleid <-"IonStarPRIDE_sampleid.csv" ##File name of the sample list (output
2)
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.
```

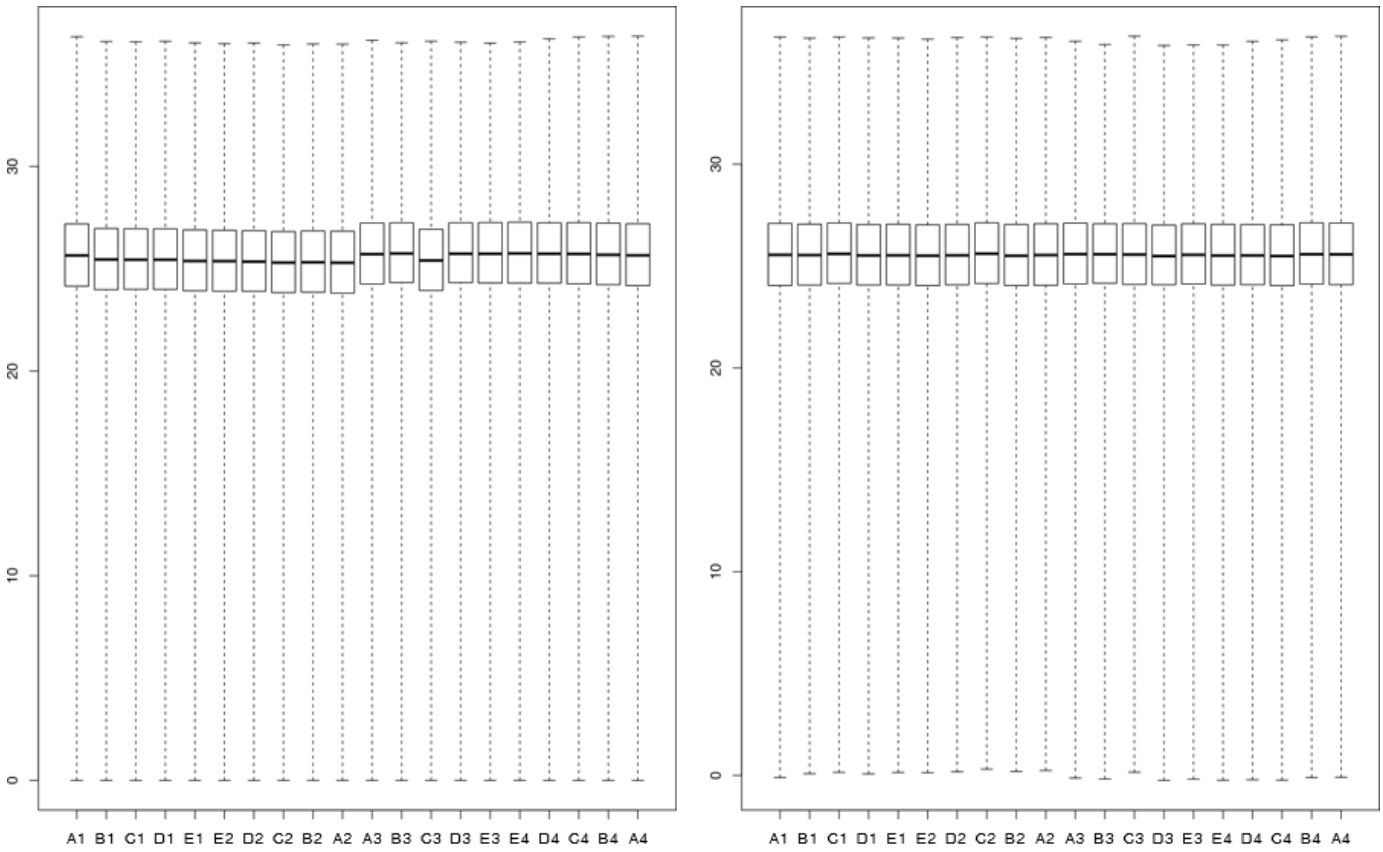
```
## 6489 duplicated frames founded.
```

```
## 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