Instructions to Run a sample dataset in InfernoRDN

(previously DAnTE)

Ashoka Polpitiya

Data

- Two conditions – A and B
- Each condition has three samples run in duplicates.
- This results in 12 runs in the dataset.

Related files

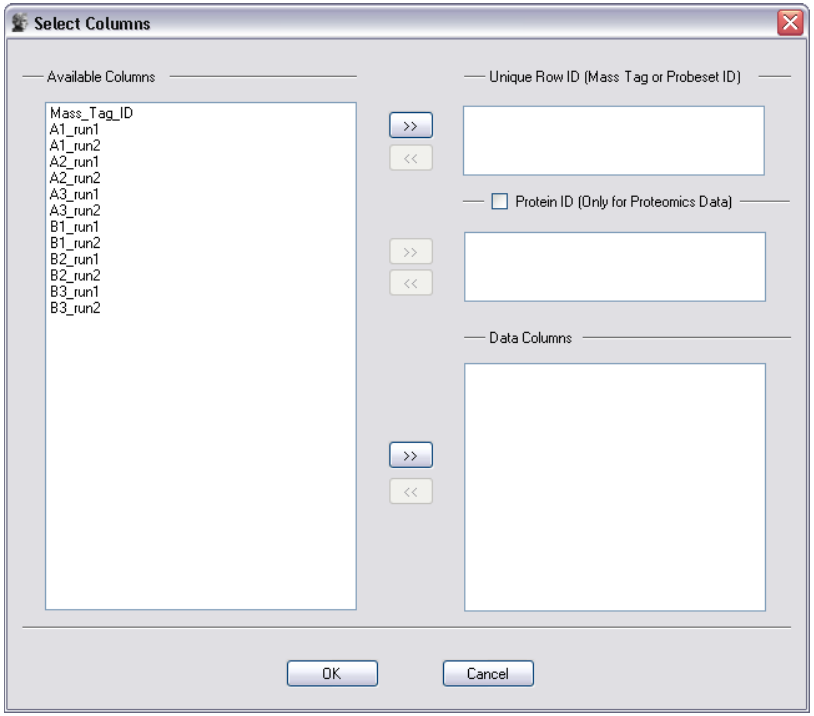
- MT_AbundanceData.csv : MassTags (peptides) and their abundance values.
- ProtInfo.csv : MassTags to Protein relationships.
- bioinformatics_supplement.dnt : Session file saved with the resulting data.

Summary of analysis

- Data loading
- Factor definitions
- Log transform
- Correlation plot
- Normalization
  - Central tendency adjustment
- Rollup to proteins
- Plot rollup results
- ANOVA on proteins
- Filter proteins with $p < 0.05$
- Cluster heatmap

Data loading

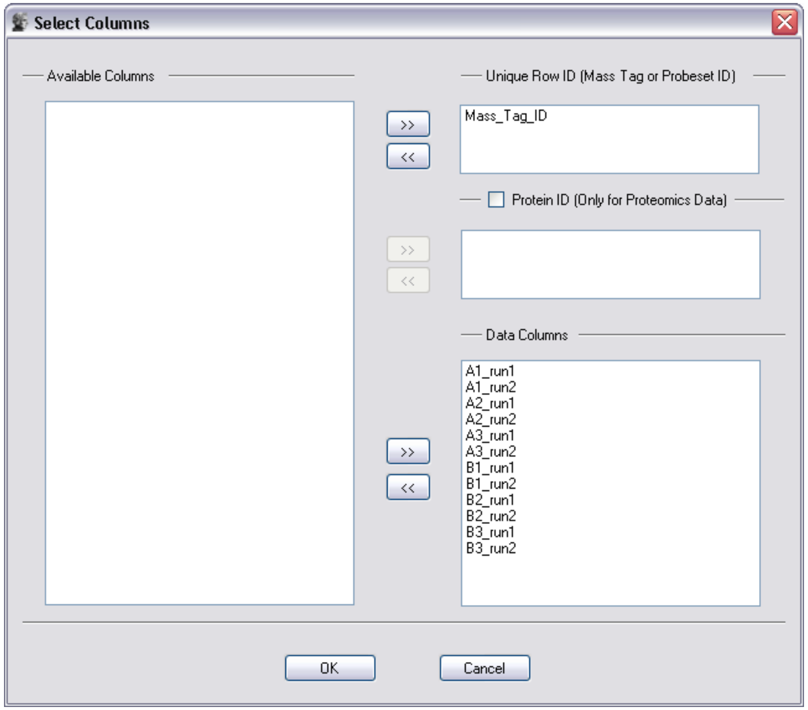
Menu: File → Open → Expression File



Available Columns:

Mass_Tag_ID
A1_run1
A1_run2
A2_run1
A2_run2
A3_run1
A3_run2
B1_run1
B1_run2
B2_run1
B2_run2
B3_run1
B3_run2

Step 1



Unique Row ID (Mass Tag or Probeset ID):

Mass_Tag_ID

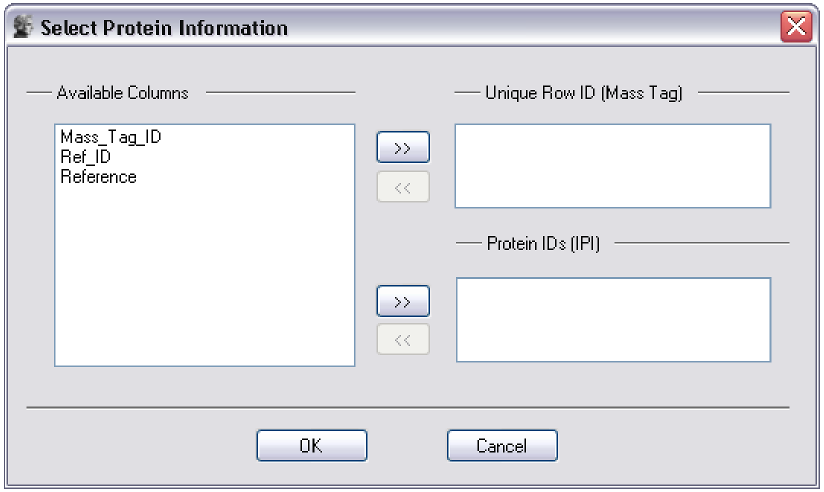
Data Columns:

A1_run1
A1_run2
A2_run1
A2_run2
A3_run1
A3_run2
B1_run1
B1_run2
B2_run1
B2_run2
B3_run1
B3_run2

Step 2

Data loading

Menu: File → Open → MassTag -Protein File

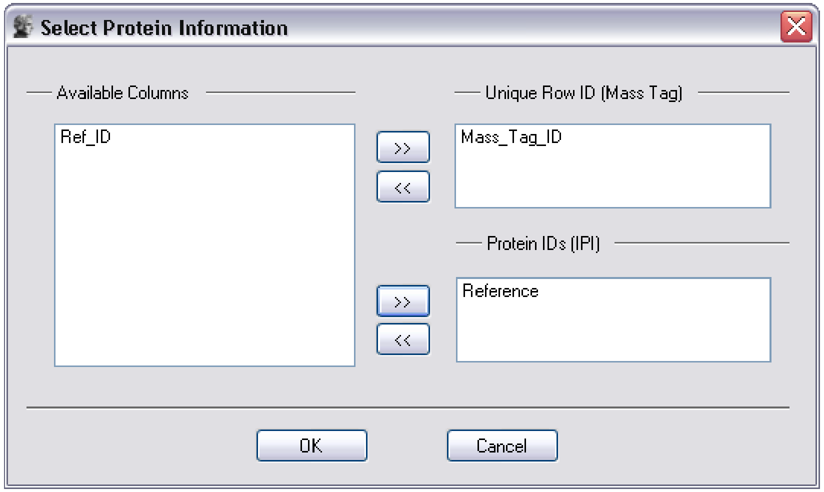


Available Columns:
Mass_Tag_ID
Ref_ID
Reference

Unique Row ID (Mass Tag):

Protein IDs (IPI):

Step 1



Available Columns:
Ref_ID

Unique Row ID (Mass Tag): Mass_Tag_ID

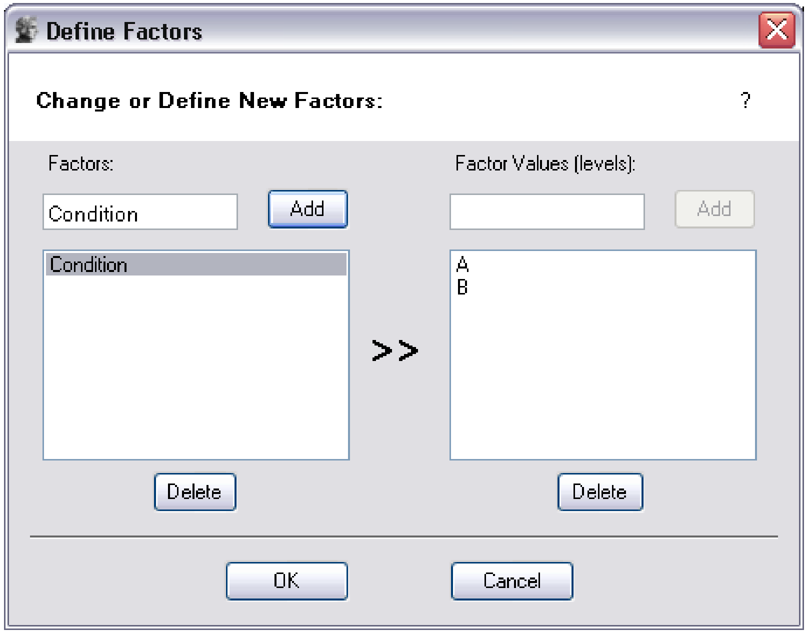
Protein IDs (IPI): Reference

Step 2

Factor definitions

Menu: Statistics → Define Factors

- Define two factors
  - Condition with two levels (A and B)
  - Replicates with 6 levels (1 … 6, each corresponds to a sample)



Define Factors

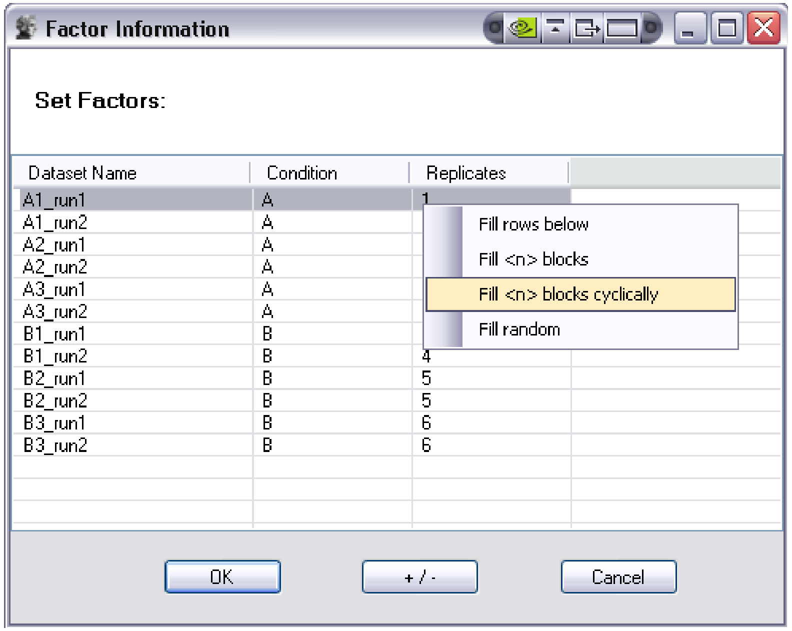
Change or Define New Factors:

Factors: Condition

Condition

Factor Values (levels):

A
B



Factor Information

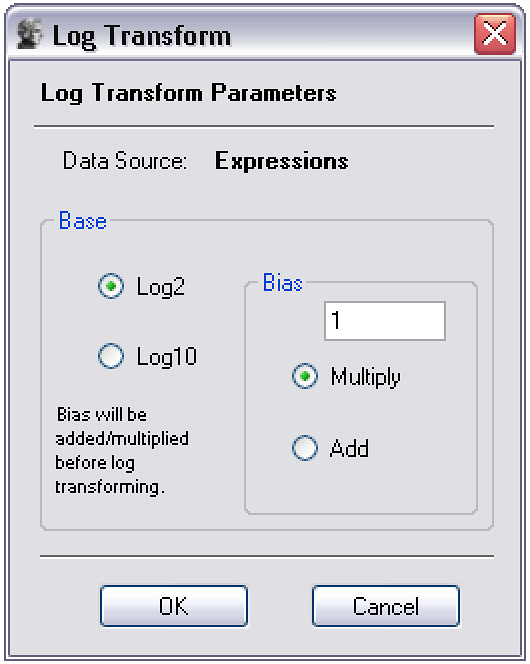
Set Factors:

Dataset Name	Condition	Replicates
A1_run1	A	1
A1_run2	A	
A2_run1	A	
A2_run2	A	
A3_run1	A	
A3_run2	A	
B1_run1	B	
B1_run2	B	4
B2_run1	B	5
B2_run2	B	5
B3_run1	B	6
B3_run2	B	6

Log transform

- Select the ‘Expressions’ table

  Menu: Pre-Process → Log Transform



Log Transform

Log Transform Parameters

Data Source: **Expressions**

Base

● Log2

○ Log10

Bias will be added/multiplied before log transforming.

Bias

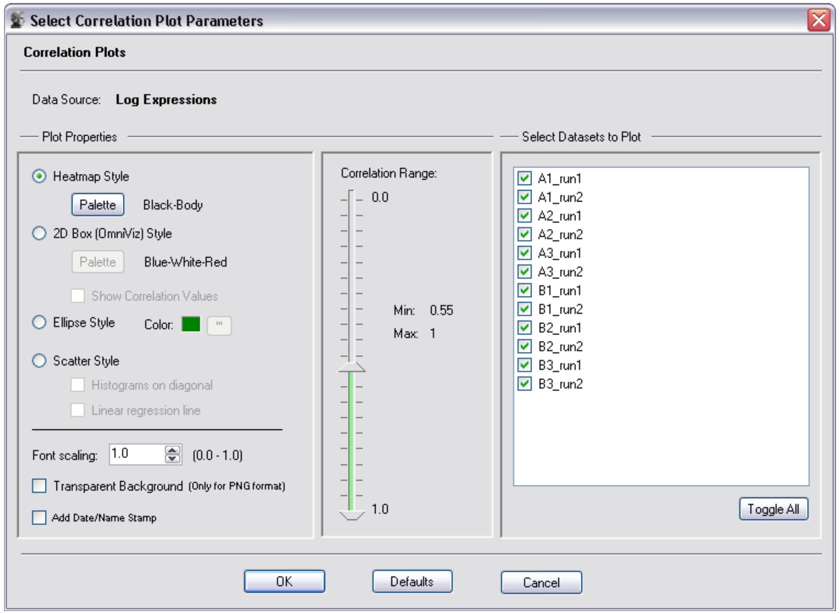
1

● Multiply

○ Add

Correlation plot

Menu: Plot → Correlation



Select Correlation Plot Parameters

Correlation Plots

Data Source: **Log Expressions**

Plot Properties

● Heatmap Style

Palette Black-Body

○ 2D Box (OmniViz) Style

Palette Blue-White-Red

☐ Show Correlation Values

○ Ellipse Style Color: 🟩

○ Scatter Style

☐ Histograms on diagonal

☐ Linear regression line

Font scaling: 1.0 (0.0 - 1.0)

☐ Transparent Background (Only for PNG format)

☐ Add Date/Name Stamp

Correlation Range:

0.0

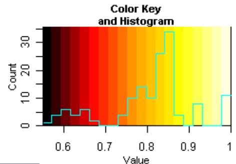
Min: 0.55

Max: 1

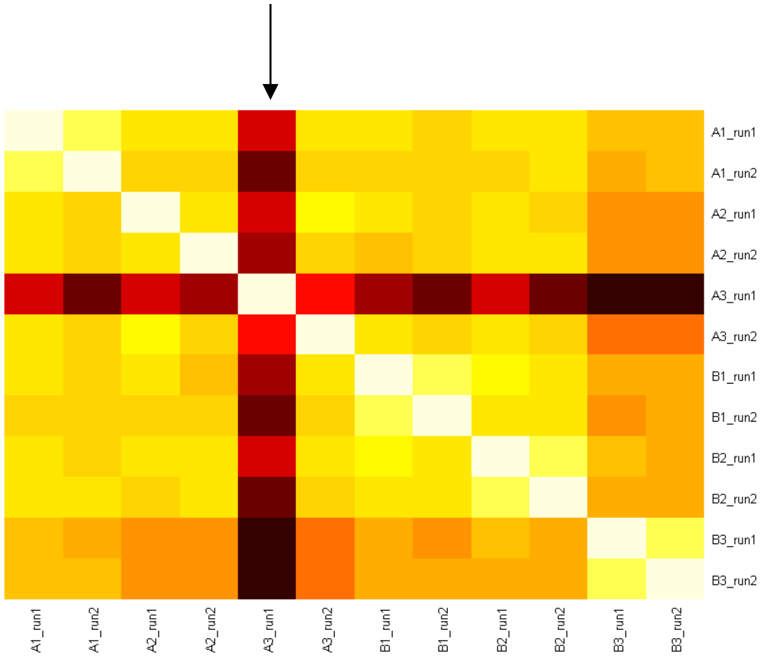
1.0

Select Datasets to Plot

☑ A1_run1
☑ A1_run2
☑ A2_run1
☑ A2_run2
☑ A3_run1
☑ A3_run2
☑ B1_run1
☑ B1_run2
☑ B2_run1
☑ B2_run2
☑ B3_run1
☑ B3_run2



Possible outlier



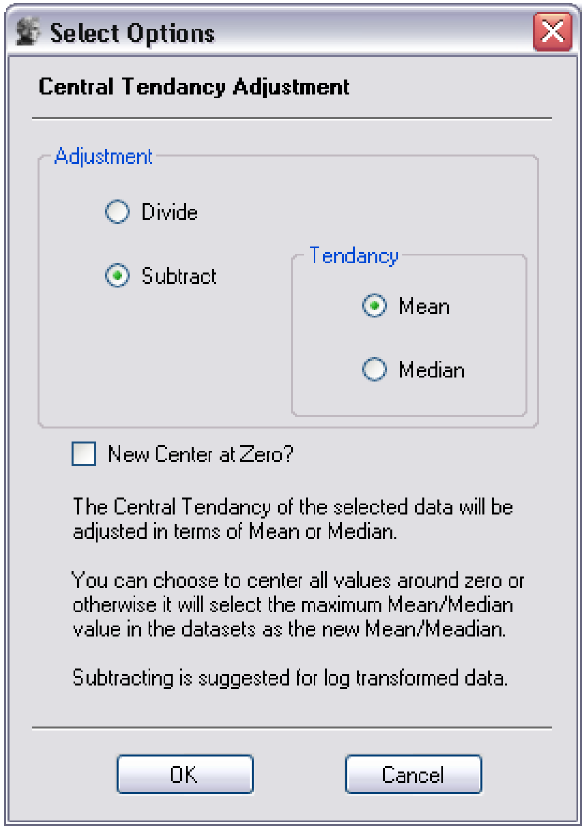
Other plots

- Q-Q plots (Menu: Plot → Q-Q Plot)
- Histograms (Menu: Plot → Histograms)
- Boxplots (Menu: Plot → Boxplots)
- … etc.

Normalization

- Lets perform a central tendency adjustment

  Menu: Pre-Process → Central Tendency



Central Tendancy Adjustment

Adjustment

○ Divide

● Subtract

Tendancy

● Mean

○ Median

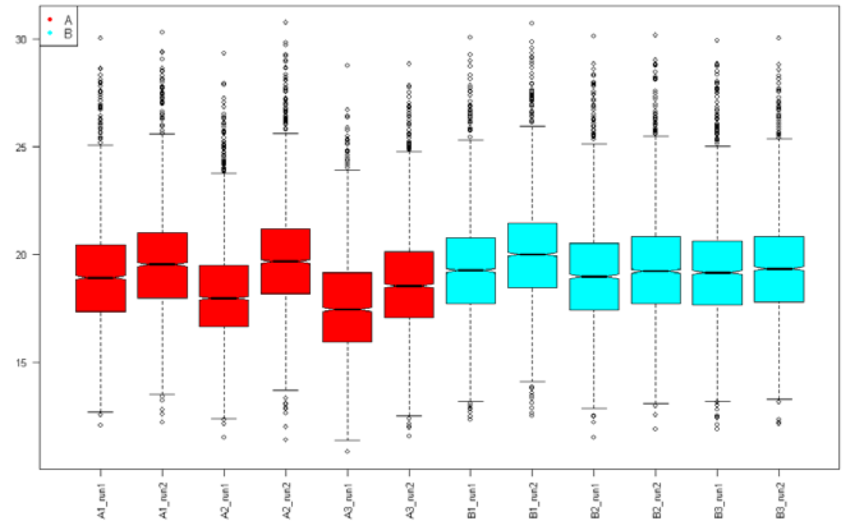
☐ New Center at Zero?

The Central Tendancy of the selected data will be adjusted in terms of Mean or Median.

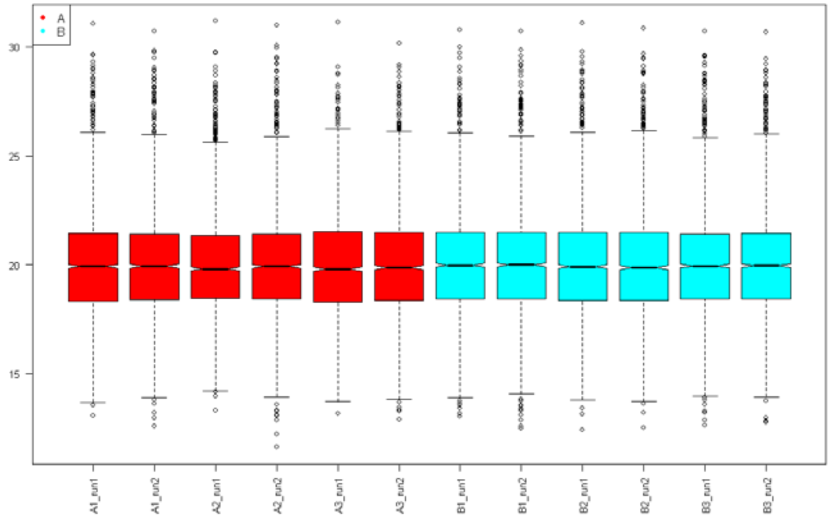
You can choose to center all values around zero or otherwise it will select the maximum Mean/Median value in the datasets as the new Mean/Meadian.

Subtracting is suggested for log transformed data.

Normalization



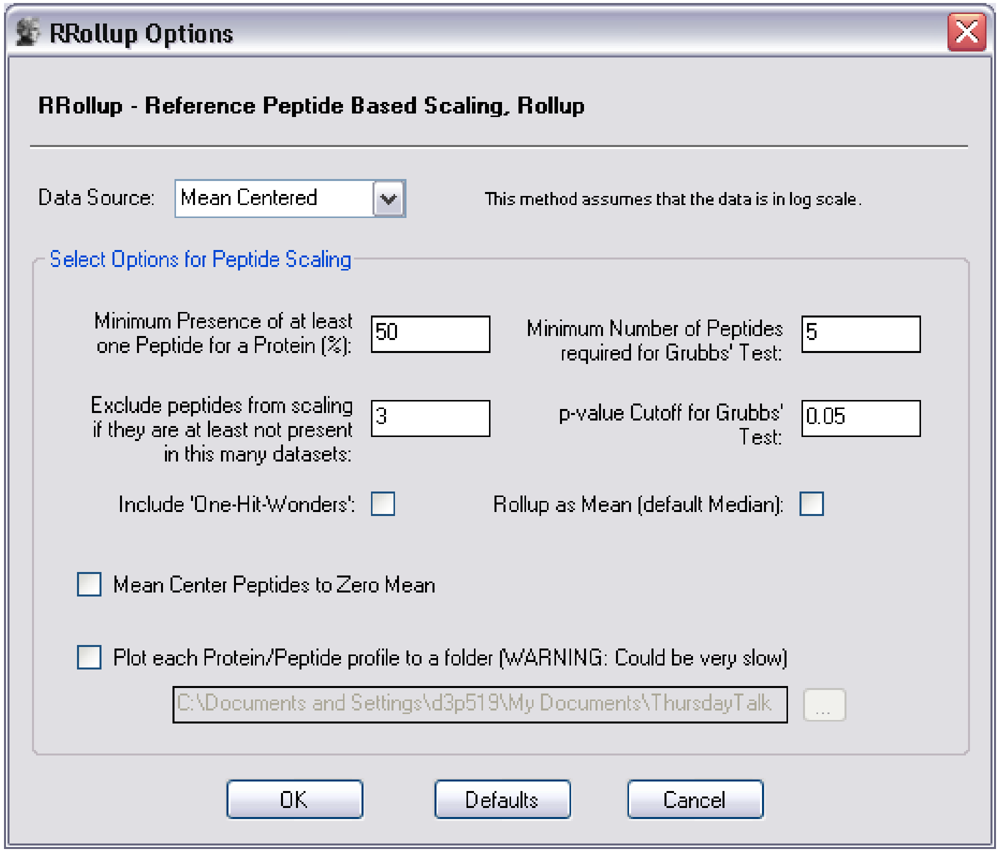
Before



After

Rollup to proteins

Menu: Rollup → RRollup



RRollup Options

RRollup - Reference Peptide Based Scaling, Rollup

Data Source: Mean Centered

This method assumes that the data is in log scale.

Select Options for Peptide Scaling

Minimum Presence of at least one Peptide for a Protein (%): 50

Minimum Number of Peptides required for Grubbs' Test: 5

Exclude peptides from scaling if they are at least not present in this many datasets: 3

p-value Cutoff for Grubbs' Test: 0.05

Include 'One-Hit-Wonders': ☐

Rollup as Mean (default Median): ☐

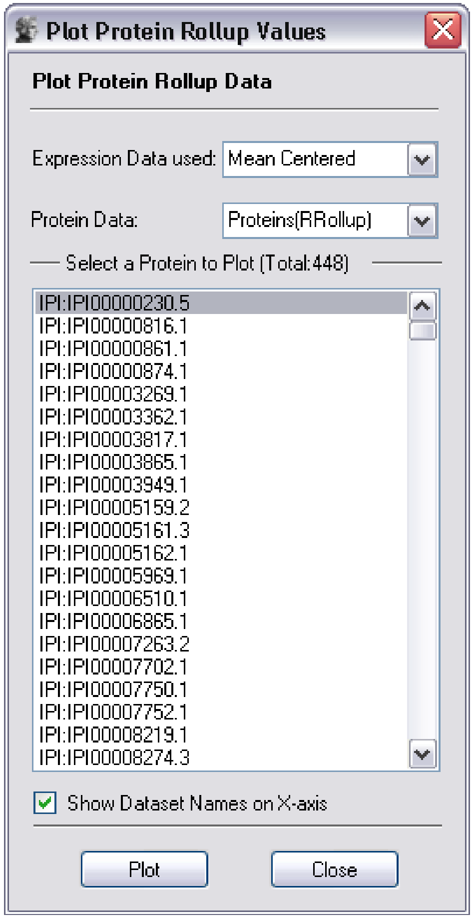
☐ Mean Center Peptides to Zero Mean

☐ Plot each Protein/Peptide profile to a folder (WARNING: Could be very slow)

C:\Documents and Settings\d3p519\My Documents\ThursdayTalk

Plot rollup results

Menu: Plot → Protein Rollup



Plot Protein Rollup Values

Plot Protein Rollup Data

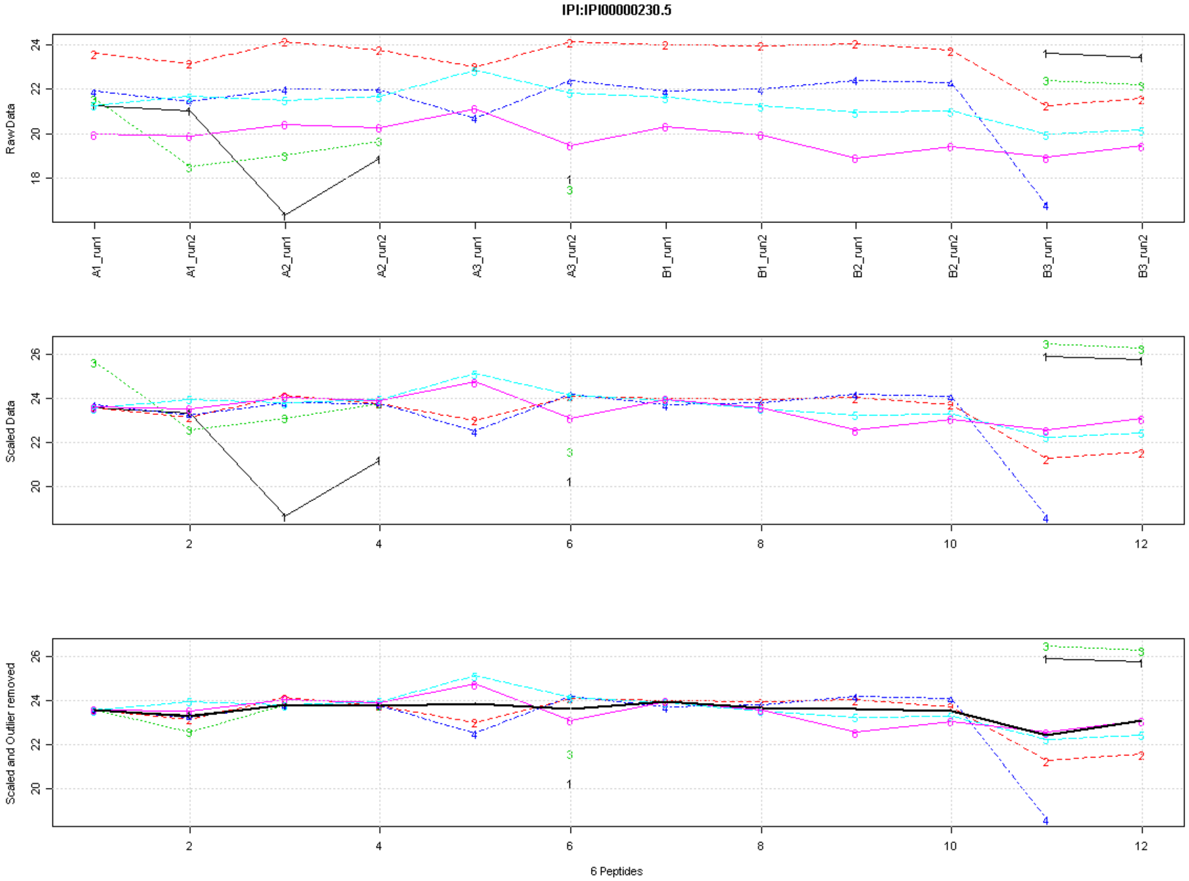
Expression Data used: Mean Centered

Protein Data: Proteins(RRollup)

Select a Protein to Plot (Total:448)

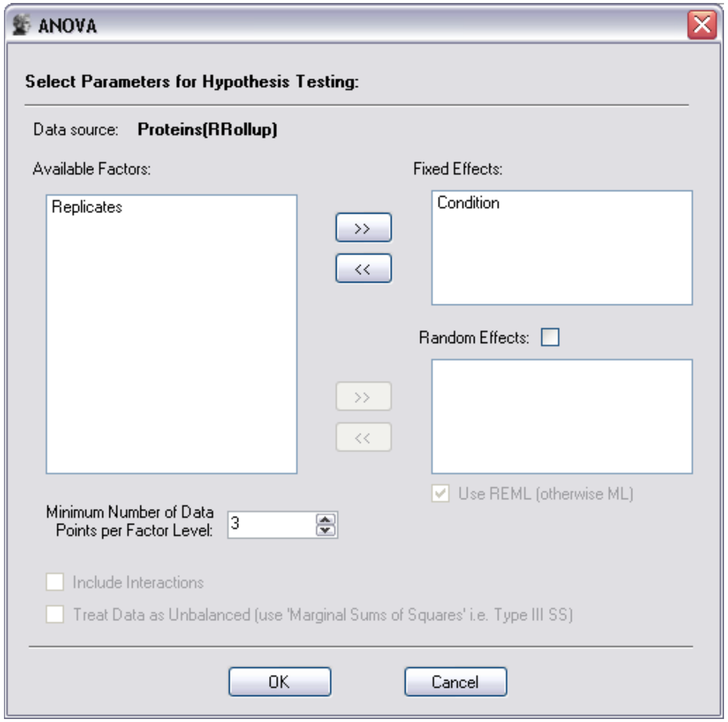
- IPI:IPI00000230.5
- IPI:IPI00000816.1
- IPI:IPI00000861.1
- IPI:IPI00000874.1
- IPI:IPI00003269.1
- IPI:IPI00003362.1
- IPI:IPI00003817.1
- IPI:IPI00003865.1
- IPI:IPI00003949.1
- IPI:IPI00005159.2
- IPI:IPI00005161.3
- IPI:IPI00005162.1
- IPI:IPI00005969.1
- IPI:IPI00006510.1
- IPI:IPI00006865.1
- IPI:IPI00007263.2
- IPI:IPI00007702.1
- IPI:IPI00007750.1
- IPI:IPI00007752.1
- IPI:IPI00008219.1
- IPI:IPI00008274.3

☑ Show Dataset Names on X-axis



ANOVA

Menu: Statistics → ANOVA



Select Parameters for Hypothesis Testing:

Data source: **Proteins(RRollup)**

Available Factors: Replicates

Fixed Effects: Condition

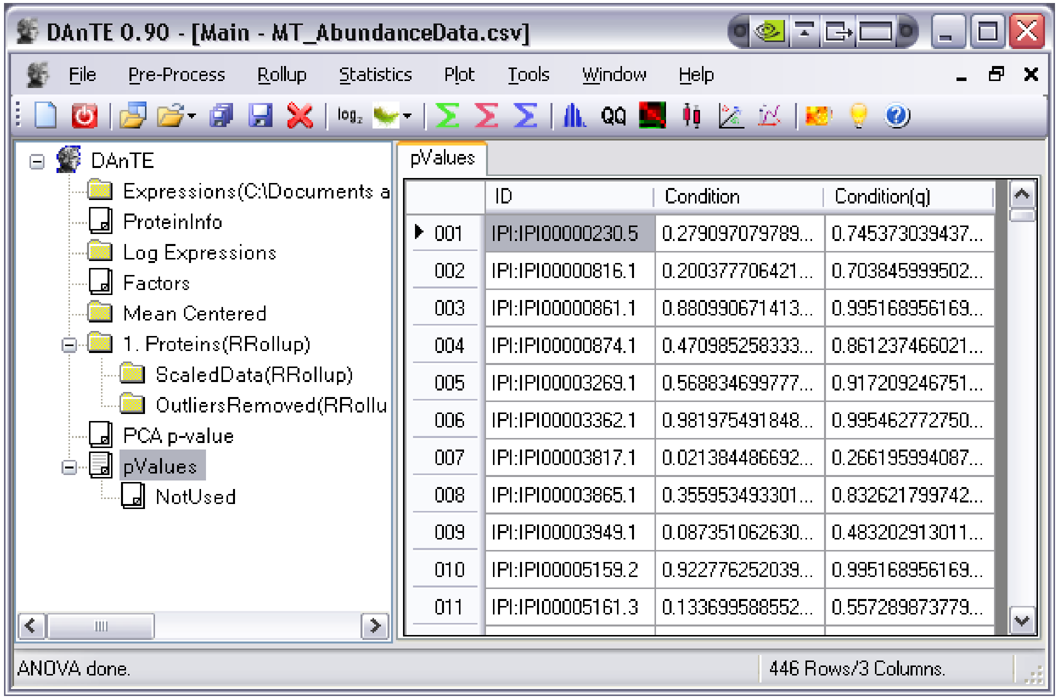
Random Effects: ☐

☑ Use REML (otherwise ML)

Minimum Number of Data Points per Factor Level: 3

☐ Include Interactions

☐ Treat Data as Unbalanced (use 'Marginal Sums of Squares' i.e. Type III SS)



	ID	Condition	Condition(q)
001	IPI:IPI00000230.5	0.279097079789...	0.745373039437...
002	IPI:IPI00000816.1	0.200377706421...	0.703845999502...
003	IPI:IPI00000861.1	0.880990671413...	0.995168956169...
004	IPI:IPI00000874.1	0.470985258333...	0.861237466021...
005	IPI:IPI00003269.1	0.568834699777...	0.917209246751...
006	IPI:IPI00003362.1	0.981975491848...	0.995462772750...
007	IPI:IPI00003817.1	0.021384486692...	0.266195994087...
008	IPI:IPI00003865.1	0.355953493301...	0.832621799742...
009	IPI:IPI00003949.1	0.087351062630...	0.483202913011...
010	IPI:IPI00005159.2	0.922776252039...	0.995168956169...
011	IPI:IPI00005161.3	0.133699588552...	0.557289873779...

ANOVA done.

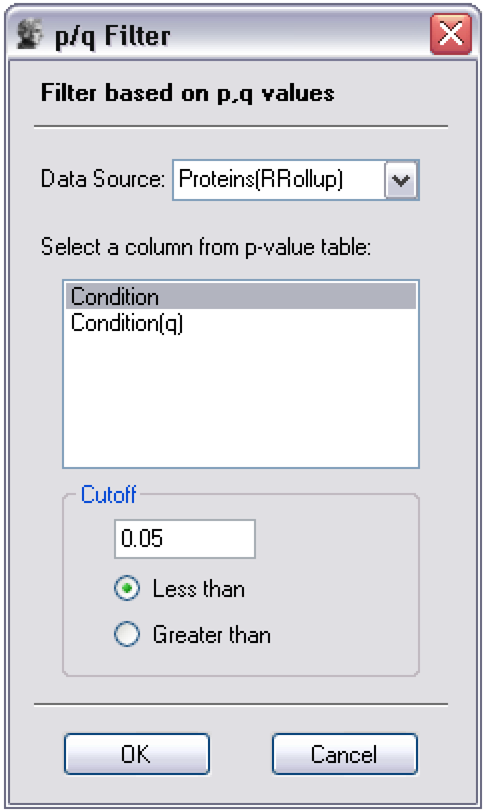
446 Rows/3 Columns.

Results – p and q values

Note: ANOVA results are similar to t-test in this case since there are only two conditions

Filter based on p-value

Menu: Tools → p/q-value Filter



Filter based on p,q values

Data Source: Proteins(RRollup)

Select a column from p-value table:

Condition
Condition(q)

Cutoff

0.05

● Less than

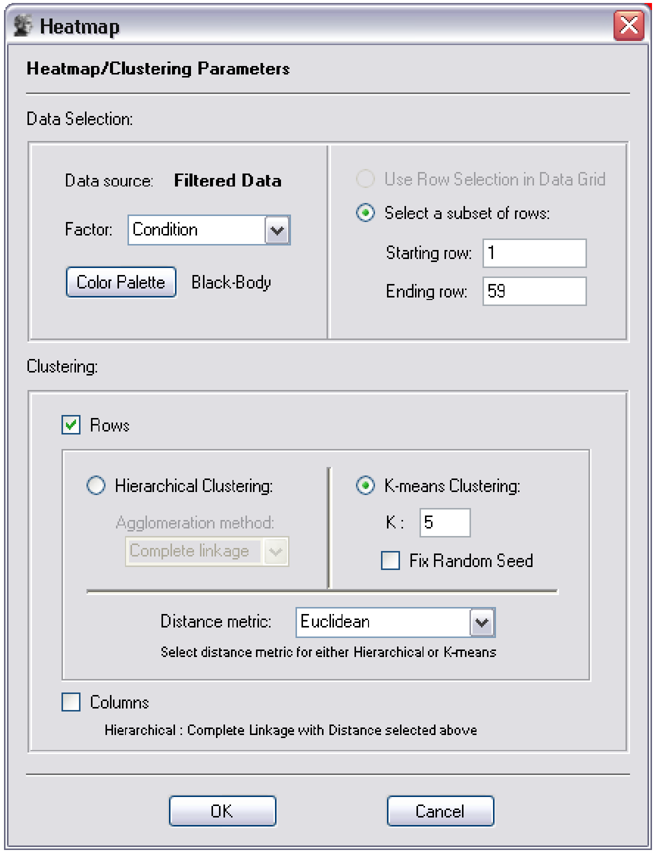
○ Greater than

→ results in 59 proteins

Cluster heatmap

- Select ‘Filtered Data’

Menu: Plot → Heatmap


Heatmap

Heatmap/Clustering Parameters

Data Selection:

Data source: **Filtered Data**

Factor: Condition

Color Palette: Black-Body

○ Use Row Selection in Data Grid

● Select a subset of rows:

Starting row: 1

Ending row: 59

Clustering:

☑ Rows

○ Hierarchical Clustering:

Agglomeration method: Complete linkage

● K-means Clustering:

K : 5

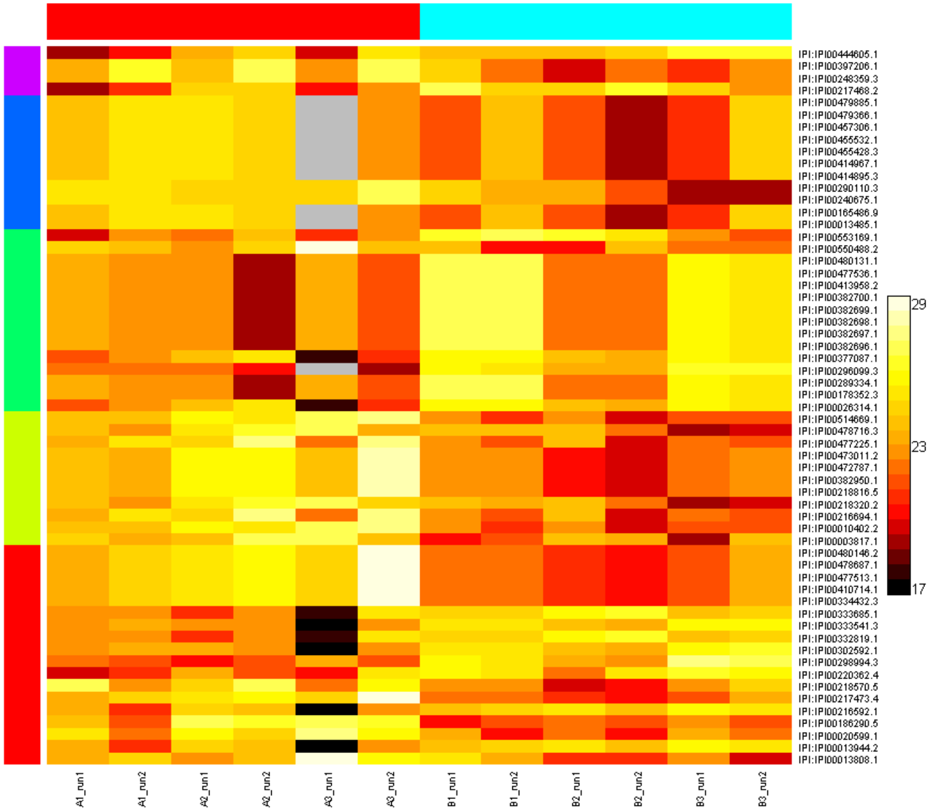
☐ Fix Random Seed

Distance metric: Euclidean

Select distance metric for either Hierarchical or K-means

☐ Columns

Hierarchical : Complete Linkage with Distance selected above



K-means cluster heatmap of 59 significant proteins