nonlinear
DYNAMICS

Quantitative Proteomics

A short introduction

Martin Wells

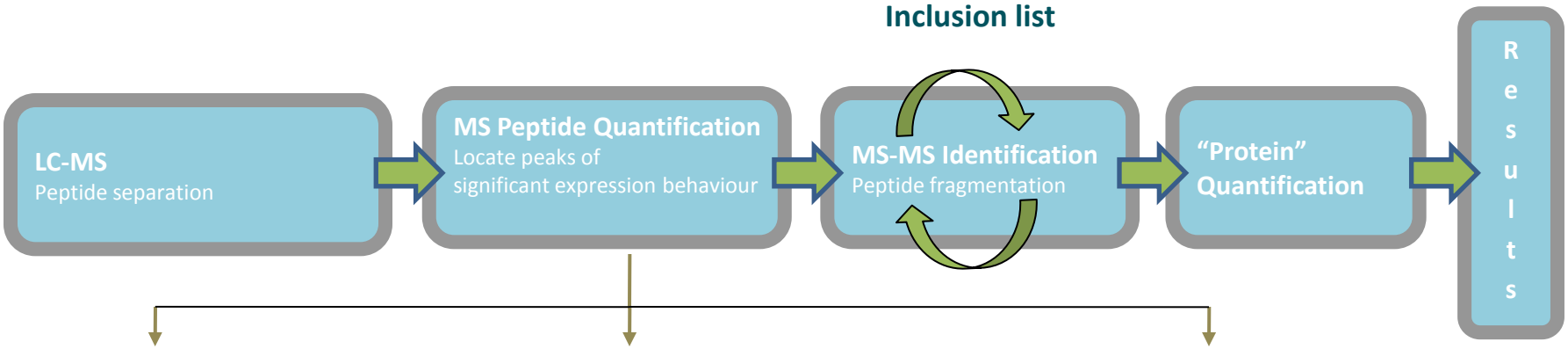
15th – 16th July 2010
Hinxton, Cambridge

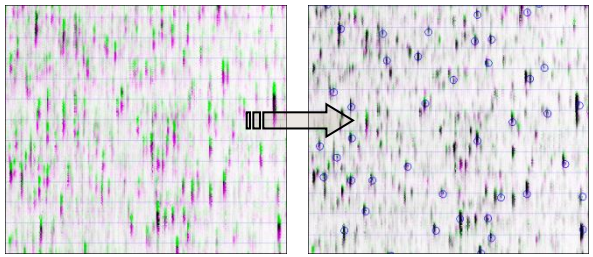
Quantitative Proteomics

- Gel or gel free
- Labelled or label free
- Relative or absolute

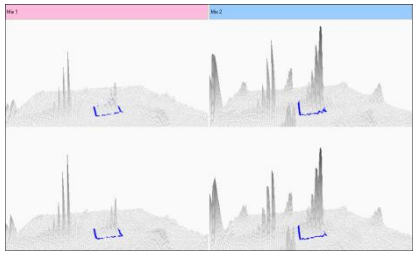
Gel free, label free, relative quantification

Analytical workflow

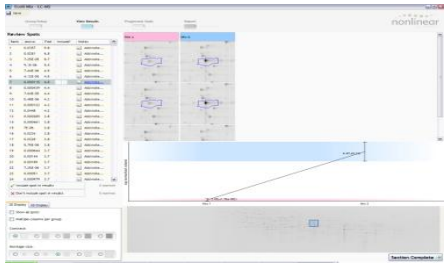




LC-MS data alignment algorithms corrects for the positional bias introduced by the LC



Co-detection and relative quantification of all peptide ions and statistical tools to define subsets of experimental interest.



Integration with search engines to identify peptide ions. Create inclusion lists for repeat analysis of peptides unidentified.
Combining of all data at the protein level and calculation of the protein quantification.

Major contributors to data variance

- Technical noise
  - e.g. experimentally introduced bias
- Analytical inadequacies
  - e.g. incorrect identification and quantification of experiment data
- Biological variation
  - how do we handle this?

These obstacles result in a loss of power and, therefore, a reduction in our ability to discover significant expression changes

nonlinear DYNAMICS

What is the goal of differential expression analysis?

“To identify the best areas that warrant further investigation as rapidly, objectively and reliably as possible."

We need to:

1. Process large data files efficiently enabling the confident selections of features for further investigation

2. Limit user subjective handling and interpretation of the data to enable reproducible analysis and results

3. Increase confidence in the resulting protein lists through improved data quality. Missing data is the primary limitation to the application of multivariate statistical techniques so addressing this would significantly impact data reliability and confidence.

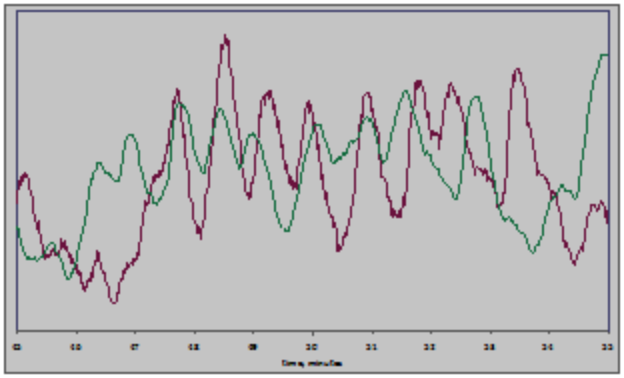
Data alignment

Assumptions

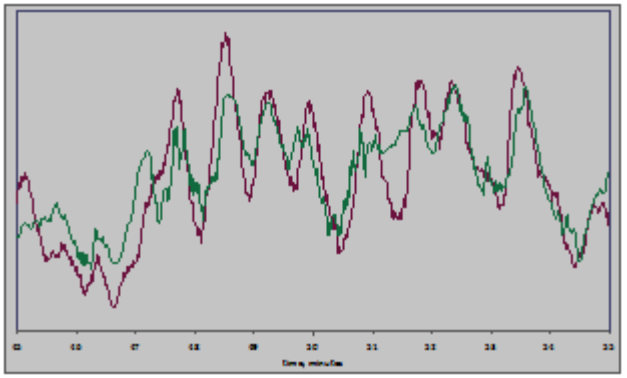
- A peptide ion, if present, should theoretically always appear in the same position under identical experimental conditions
- Most differential expression experiments have a high degree of common peptide ions between groups

As each peptide ion has a unique position can we, using the common peptide ions, realign the data so that we undo the positional bias that has been introduced by the chromatography?

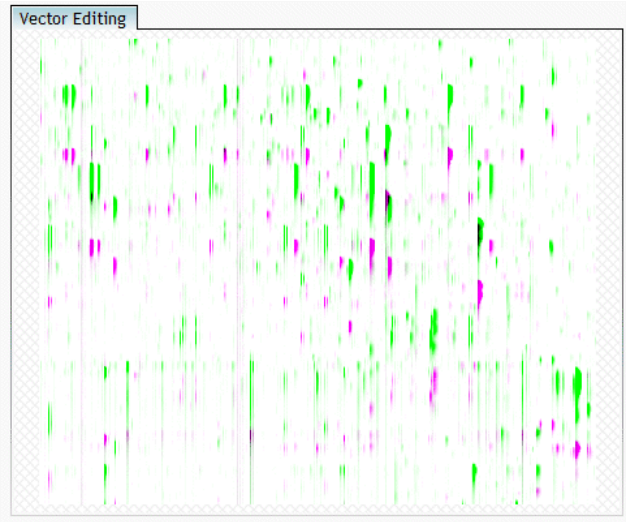
Data alignment

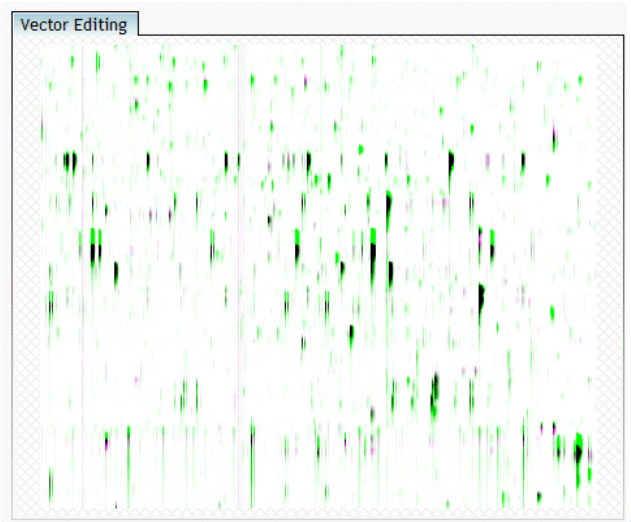


Section prior to alignment



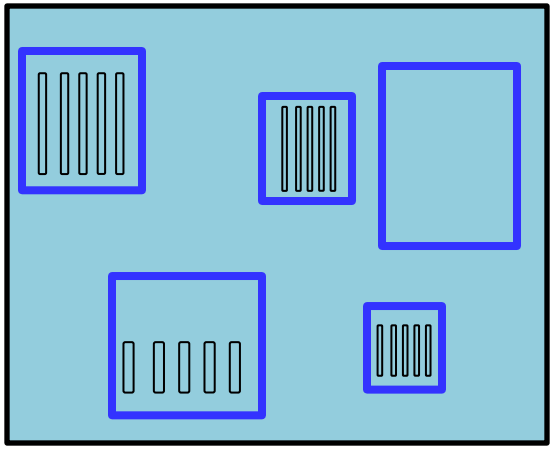
Section after alignment

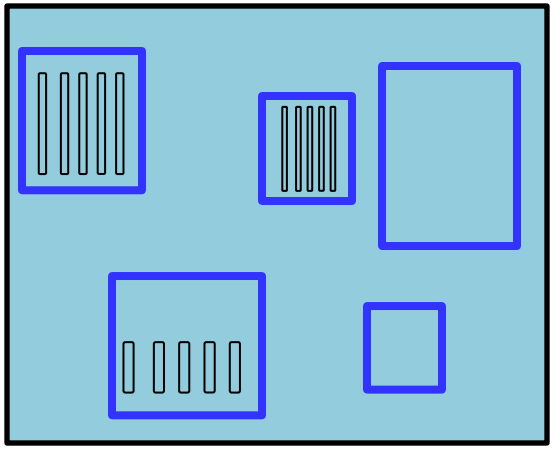


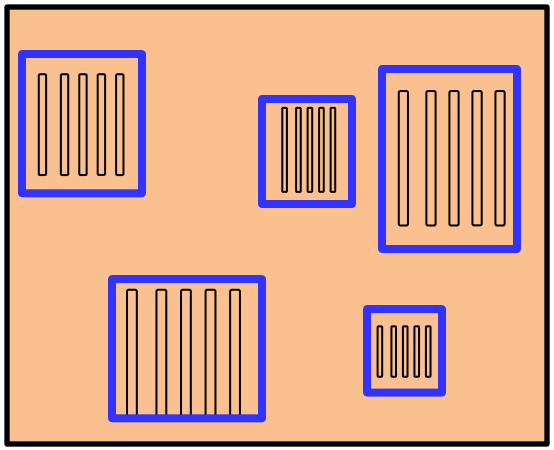


nonlinear DYNAMICS

Feature detection – peak picking

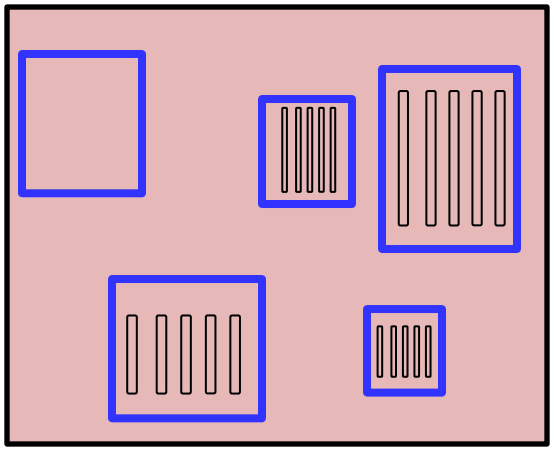


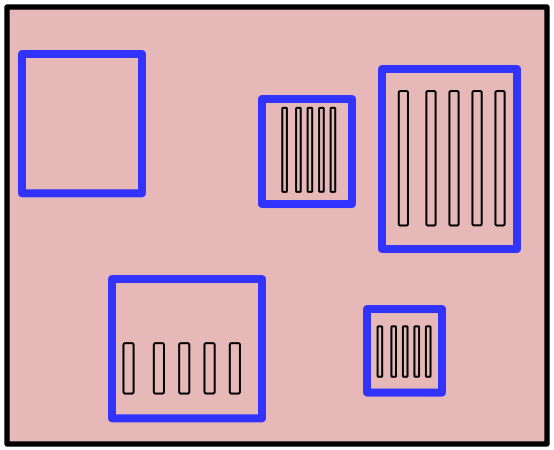




Aggregate co-detection

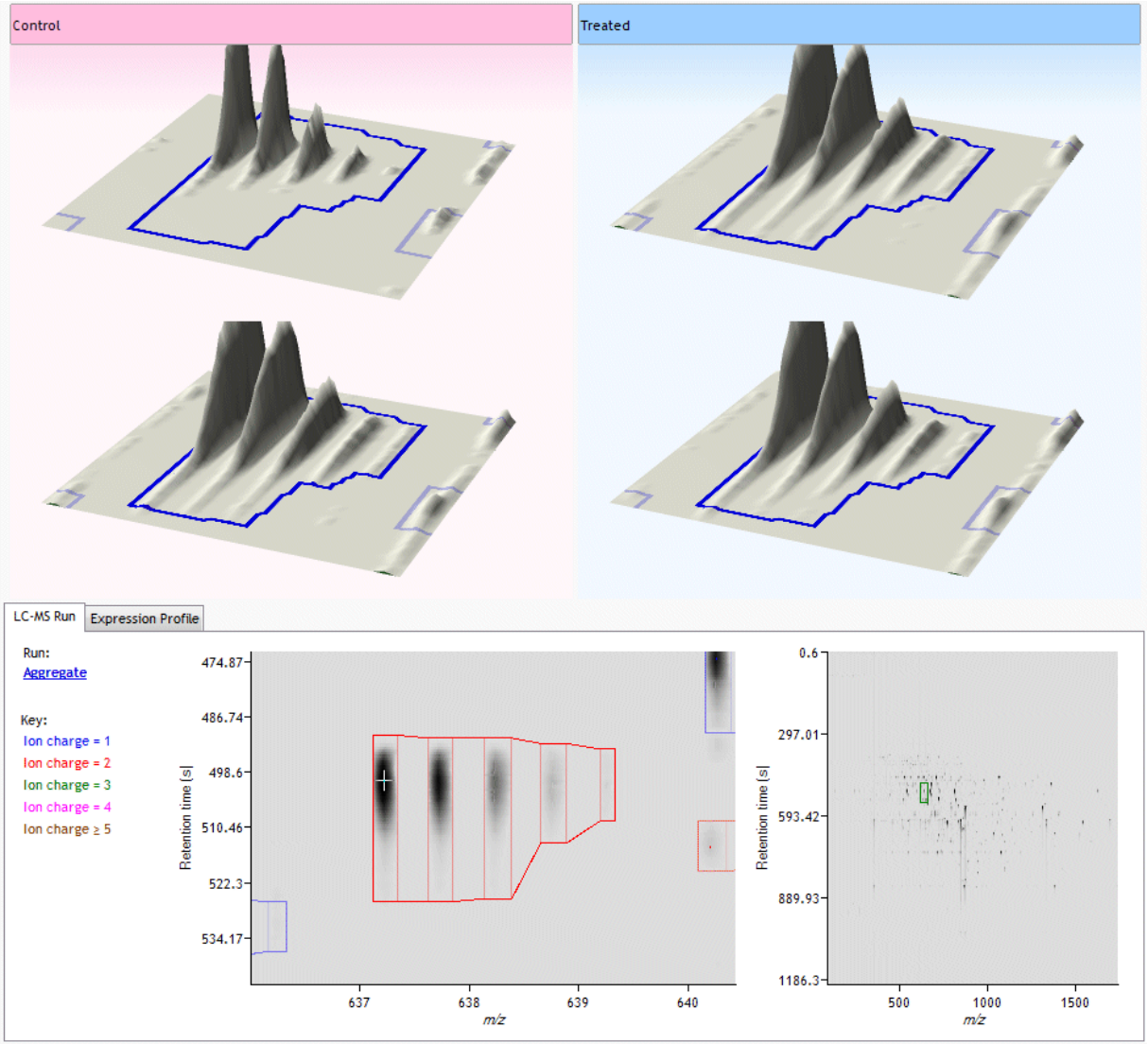
Mapping the detection to all runs avoiding missing data





nonlinear DYNAMICS

Peptide ion raw abundance



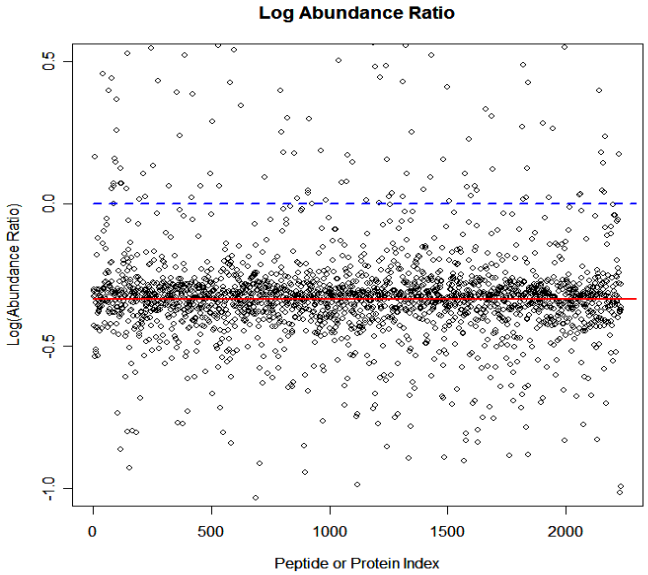
Data normalisation

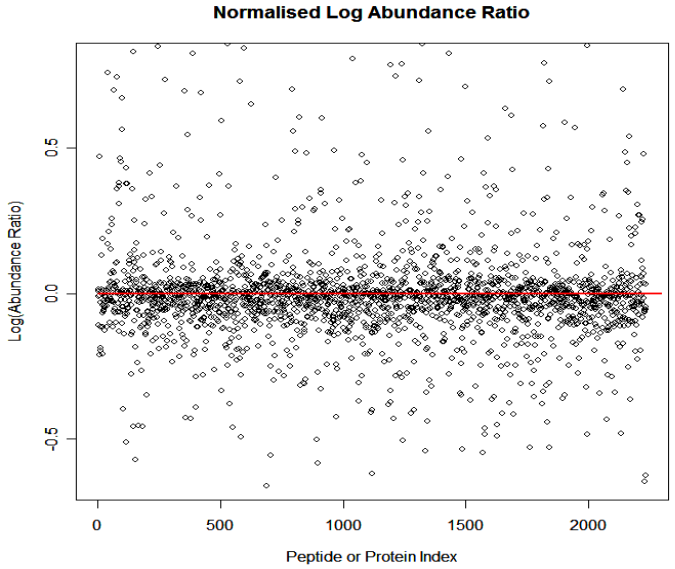
Normalisation is required to calibrate between different sample runs.

Total ion current has been assumed to be equal across all samples. However without missing data an alternative approach is possible.

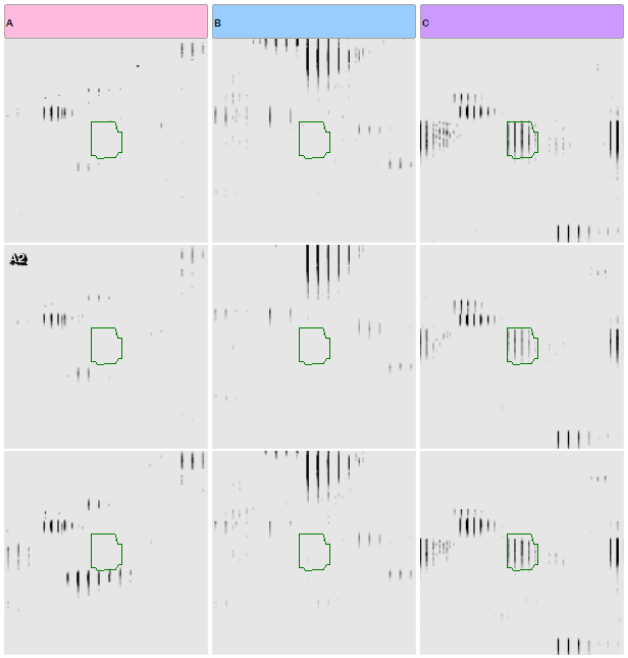
Calculating log abundance ratios for every feature gives a more robust assessment of the gain factor.

Of course, the base assumption still remains that enough features should NOT be changing in abundance.



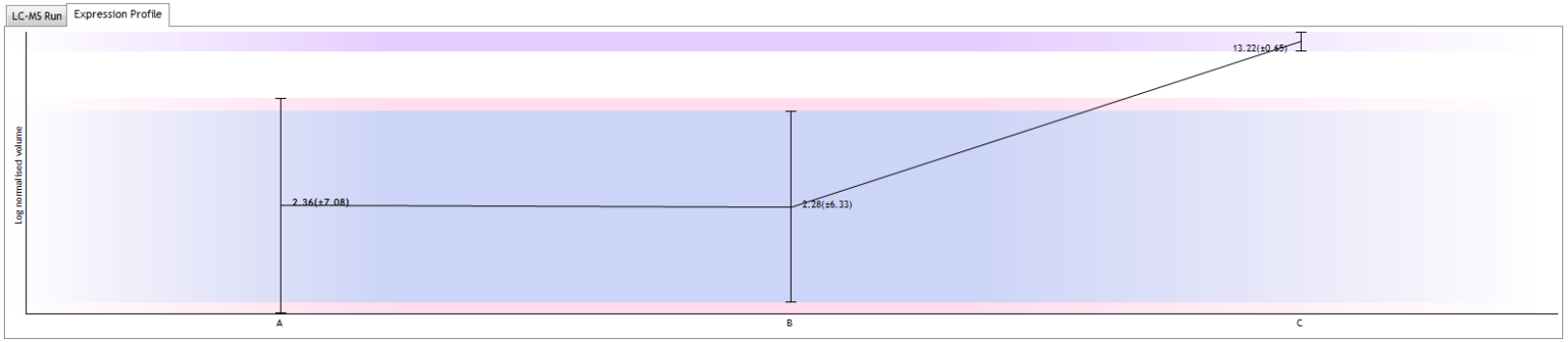


Relative normalised abundances.



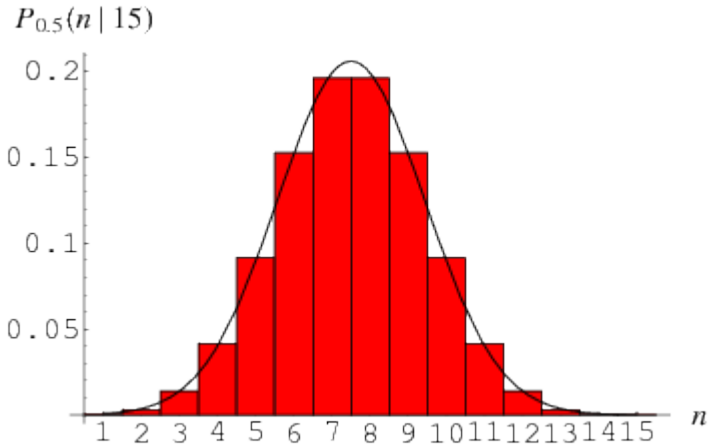
3 groups with 3 replicates per group.

Graphical expression profile showing group means with error bars representing 95% confidence bounds.



Data transformation

The data must meet certain prerequisites for the statistical test applied.
Ie The data is normally distributed and of equal variance.



To achieve this a **data transformation** is required.

An ArcSinh transformation of the normalised data is most suitable particularly at very low values when compared to other transformations eg log.

Therefore all statistical analysis is conducted using the ArcSinh normalised volumes

Uni and multivariate statistics

The complete data table of quantitative information, without missing data, enables the robust and confident application of multivariate statistics to support traditional univariate tests.

- p-values (typically $p<0.05$ – ie 5% type I error)
- PCA – Quality control overview of the variances observed in your data.
- Power analysis per feature to reduce type II errors
- Predictive power for predicting how many samples are needed to be confident you are not missing something
- False discovery rate correction – Q values to reduce type 1 errors
- Correlation analysis of peptide ions

  See which peptides are highly correlated within and across groups.

nonlinear DYNAMICS

Statistical Power

In addition to the standard p-value filtering of data, by addressing the issue of missing data, we are now able to accurately calculate the statistical power for each feature.

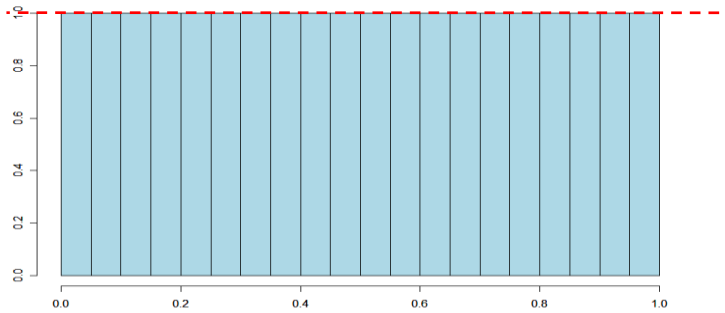
Power – the error of failing to observe a difference when in truth there is one, thus indicating a test of *poor sensitivity*

This enables secondary filtering of the data to define a subset of features that have both low p-value (eg <0.05) and high power (eg >80%).

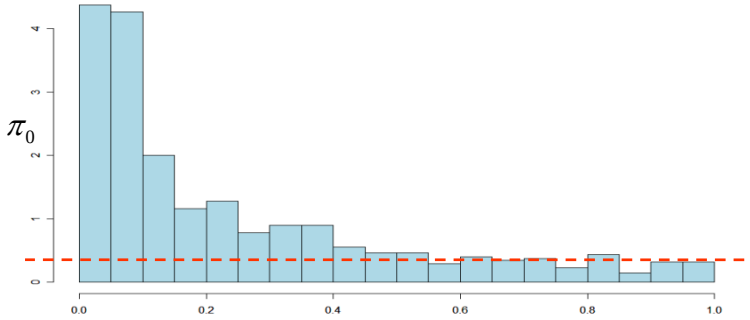
You can calculate that you have run sufficient samples to observe a difference if one really exists or reduce your “discoveries” to features where confidence is high. (eg >80%)

False Discoveries and Q-values

- False discoveries occur when we are observing a difference when in truth there is none, thus indicating a test of *poor specificity. Q-values are a tool to reduce such false discoveries*
- False Discovery Rate (FDR) assumes a uniform distribution of p-values but ...



Theoretical p-value distribution



Actual p-value distribution

- q-values take the into account the actual p-value distribution factor

$$\tilde{q}_m = p_m$$
$$\tilde{q}_i = \min(\tilde{q}_{i+1}, \pi_0 * (m/i) * p_i),\ i = m-1,\ldots,1$$

[1] Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. Proc. Natl Acad. Sci. USA, **100,** 9440–9445.
[2] Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer,P., Gusterson, B., Esteller, M., Kallioniemi, O. P., et al. (2001) N. Engl. J.Med. **344, 539–548.**

nonlinear DYNAMICS

Database result integration

MSMS fragment scans are extracted and compiled into file formats compatible with most search engines.

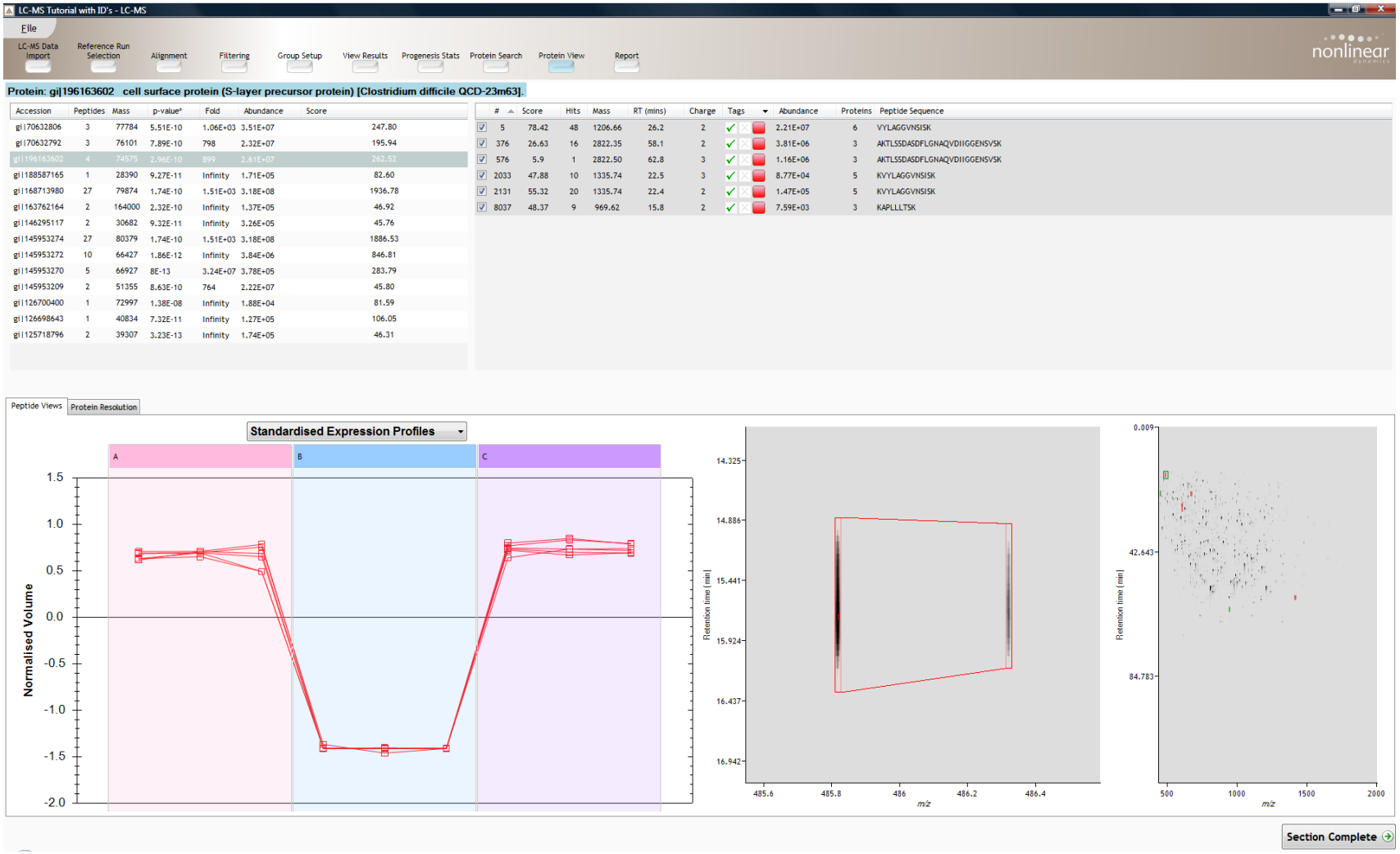
- Mascot
- Sequest
- Phenyx
- PLGS
- Peaks Studio
- Proteome Discoverer

The qualitative results from can then be imported back into the software for integration with the quantitative data.

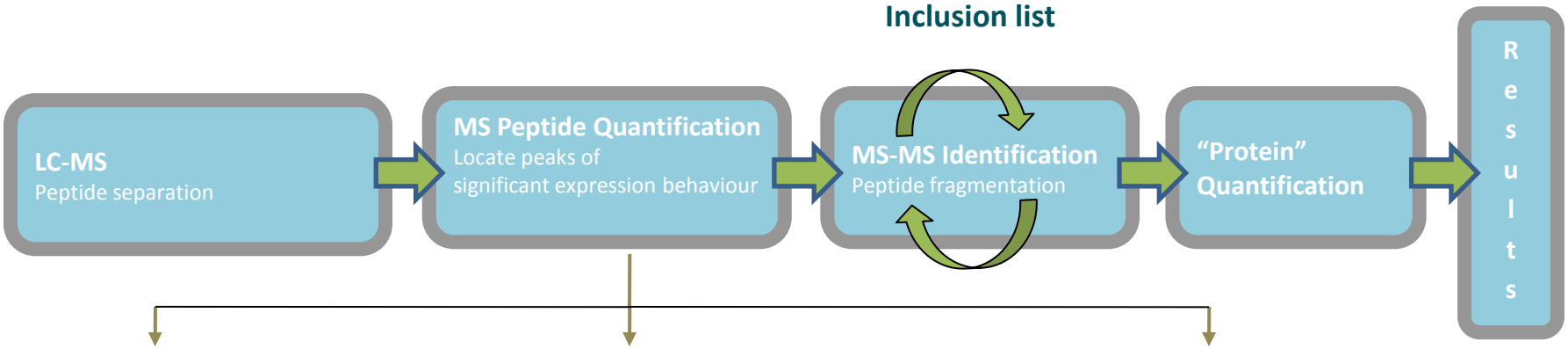
Data is then automatically collated at the protein level and the protein quantification is then calculated based on the combined impact of all peptide ion measurements.
Tools are available to remove non-unique peptides from the data and highlight peptides not in consensus with other peptide abundances. Eg PTM's

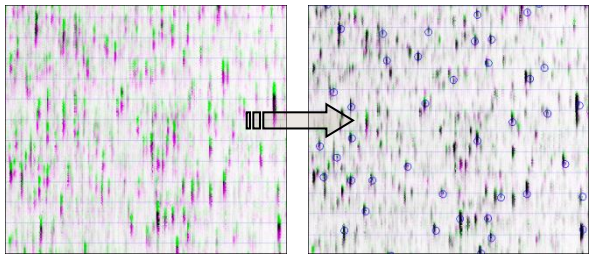
nonlinear DYNAMICS

Protein results

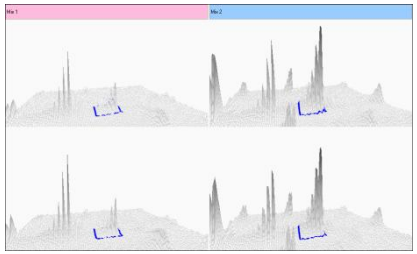


Analytical workflow

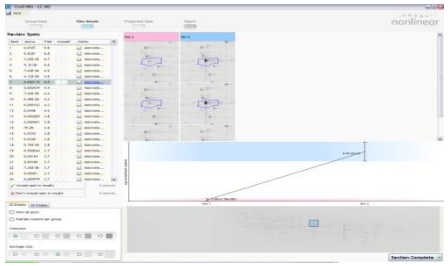




LC-MS data alignment algorithms corrects for the positional bias introduced by the LC



Co-detection and relative quantification of all peptide ions and statistical tools to define subsets of experimental interest.



Integration with search engines to identify peptide ions. Create inclusion lists for repeat analysis of peptides unidentified.
Combining of all data at the protein level and calculation of the protein quantification.

Practical Session

Experiment

• Membrane proteins extracted from 2 different strains of Clostridium Difficile.

• Three separate cultures were prepared for each strain and analysed on the LC-MS. (Dionex Ultimate 3000 and Thermo Orbitrap XL)

•Analyse the 6 resulting files, comparing Strain “A” to strain “C”

•A database search has been performed using mascot. Import the identification results, removing hits with a mascot scores <30 and hits not containing “difficile” in the protein description.

•After removing any non-unique peptides from the quantification, report the fold change and p-values of the proteins with >4 unique peptides.

•For the purpose of this exercise hypothetical proteins can be excluded from the results.

nonlinear DYNAMICS

Practical Session

Expected results...

6 Proteins all up regulated in strain C (excluding hypothetical proteins).

Protein Description	p-value	Fold change	No. Unique peptides
ABC transporter, substrate Binding lipoprotein	1.3E-05	23	12
Cell surface protein	0.0002	3	5
Enolase	3.7E-05	3.5	5
Putative 5-niroimidaxole Reductase	3.7E-06	31	5
Glyceraldehyde-3-phosphate Dehydrogenase	2.3E-06	8	4
Cell wall protein V	3.4E-07	100	4