Development of a computation framework for the analysis of protein correlation profiling and spatial proteomics experiments:

Nichollas E. Scott[1]*, Lyda M. Brown[1], Anders R. Kristensen[2] and Leonard J. Foster[1]*

[1]Centre for High-throughput Biology, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4.

[2]BC Cancer Agency, Vancouver, British Columbia, Canada, V5Z 1L3.

- Address correspondence to:

  LJF: foster@chibi.ubc.ca
  NES: nichollas.e.scott@gmail.com
  Department of Biochemistry and Molecular Biology, University of British Columbia, 2125 East Mall, Vancouver, British Columbia, Canada V6T 1Z4. Tel.: +1 604 822 8311.

Running title: *The development of computational alignment, visualization and amalgamation tools for PCP-SEC-SILAC datasets as applied to a HeLa* epithelial cell model of *Salmonella enterica* serovar Typhimurium *infection*

Keywords: *Protein correlation profiling, SILAC, Matlab, Spatial proteomics.*

Abbreviations: FBS- fetal bovine serum; PBS - phosphate buffer saline; r.c.f. -relative centrifugal force; PCP- Protein correlation profiling ;SEC-size exclusion chromatography ;SILAC- Stable isotope labeling by amino acids in cell culture; TPR- True positive rate; FPR- False positive rate; PPI- protein protein interaction; TP- True positive; FP- false positive; TN- True negative; FN- False negative; MvsL- Medium over Light; HvsL- Heavy over Light

Supplementary document 1: How to run protein interaction analysis using the MaxQuant and the PCP SILAC MATLAB scripts

Date: 25th of September 2014

Description: This document outlines how to prepare data for and run MATLAB scripts generated by the Foster lab from the analysis of PCP SILAC datasets. If errors (defined here as either the scripts resulting in erroneous results or termination of the script before completion) are found within any scripts please report them the either Nichollas Scott (nichollas.e.scott@gmail.com) or Leonard Foster (foster@chibi.ubc.ca).

Before beginning

Before beginning ensure the following software is installed:

-Matlab (version R2010a, 2011 or 2012)

-Matlab Parallel computing toolbox

-Matlab curve fitting toolbox

-Mann-Whitney-Wilcoxon test MATLAB function (created by Giuseppe Cardillo and available from http://www.mathworks.com/matlabcentral/fileexchange/25830-mann-whitney-wilcoxon-test, required for Comparison script to preform Mann-Whitney-Wilcoxon test)

-PCA and ICA package MATLAB function (create by Brain Moore and available from http://www.mathworks.com/matlabcentral/fileexchange/38300-pca-and-ica-package, required for ROC script to preform PCA)

-t-Distributed Stochastic Neighbor Embedding dimensionality reduction packages (created by Laurens van der Maaten and Geoffrey Hinton and available from http://homepage.tudelft.nl/19j49/t-SNE.html, required for ROC script to preform t-SNE analysis)

-MaxQuant (http://www.maxquant.org/downloads.htm, version with match between run enabled is preferred)

-Perseus (http://www.maxquant.org/downloads.htm).

Part A: Preparing dataset and required databases for analysis

I) Setting up a spatial proteomics experiments within Maxquant and preprocessing for analysis with MATLAB:

Rational: Within spatial proteomic experiments both the identity and location (fraction, gel slice or gradient level) are critical for the generation of protein profiles. Due to this requirement care must be taken to during the initial data handling to ensure both identity and location information is maintained.

Protocol:

Maxquant is suggested for the initial processing of spatial proteomic data set due to the ability the identification and infer the identification of MS features across experiment (Match between runs). Within spatial proteomic experiments the proteinGroups.txt output from Maxquant is the most useful output for analysis as it contains all the SILAC isotopologue measurements across all the spatial fractions.

To set up a spatial experiment each spatial fraction needs to be defined as a separate experiment, this is done by generating an experimentalDesignTemplate.txt prior to commencing the search. To generate this file load all the raw files to be analyzed into Maxquant and press the write template button (Figure 1-left). This will generate the experimentalDesignTemplate.txt which can then be modified to specify which raw files correspond to which fraction (Figure 1-right). In the experimentalDesignTemplate.txt you are able to denote multiple injections to single fractions enabling the results from multiple runs to be pooled as a single fraction.

Once generated, open the proteinGroups.txt with Perseus to import the desired columns using the Maxquant import tool (figure 2). For the generation of protein profiles the experimental vs reference isotopologue ratios should be imported (in our experiments the light label sample is typical used as the reference and profiles are generated from MvsL and HvsL isotopologue channels). To enable the comparison between experimental conditions to examine dynamics protein interaction the normalized isotopologue ratios should be used (in our case HvsM normalized). We have found comparing the normalized values appear to be more reliable then comparing MvsL to HvsL for quantitation of interaction changes. Import the protein information (Protein ID, Majority protein IDs, Protein name, Gene name, Fasta Headers) and any other information which will be required for downstream analysis (i.e predicted size of protein). Remove, reverse matches, contaminates and protein identified by only identified by site using the categorical filtering function. The final spatial dataset should look like figure 3.

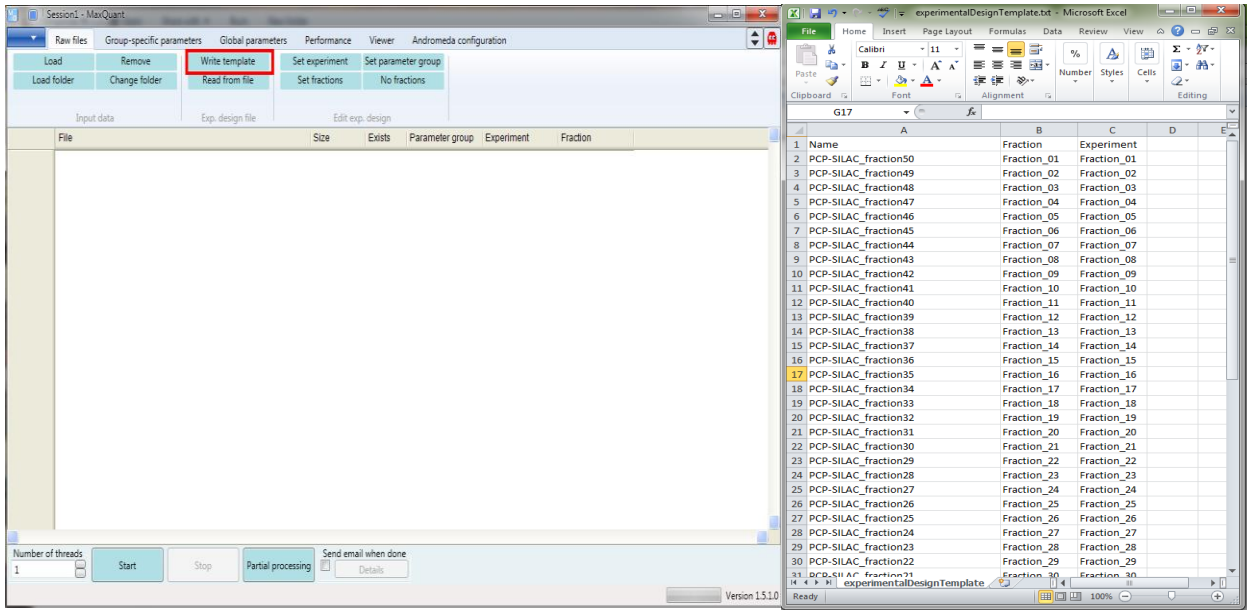


Figure 1: Generating the experimentalDesignTemplate.txt for Maxquant. Left) how to create the experimentalDesignTemplate.txt file Right) how to modify the experimentalDesignTemplate.txt to generate SILAC ratios for each fraction.

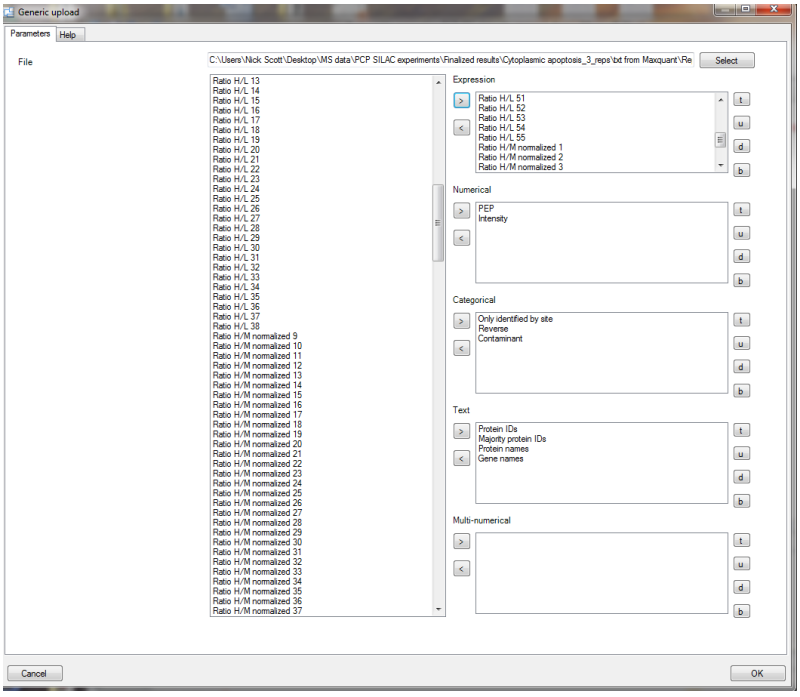


Figure 2: Maxquant import tool, organize and select the required columns to enable the analysis of protein profiles or interaction changes.

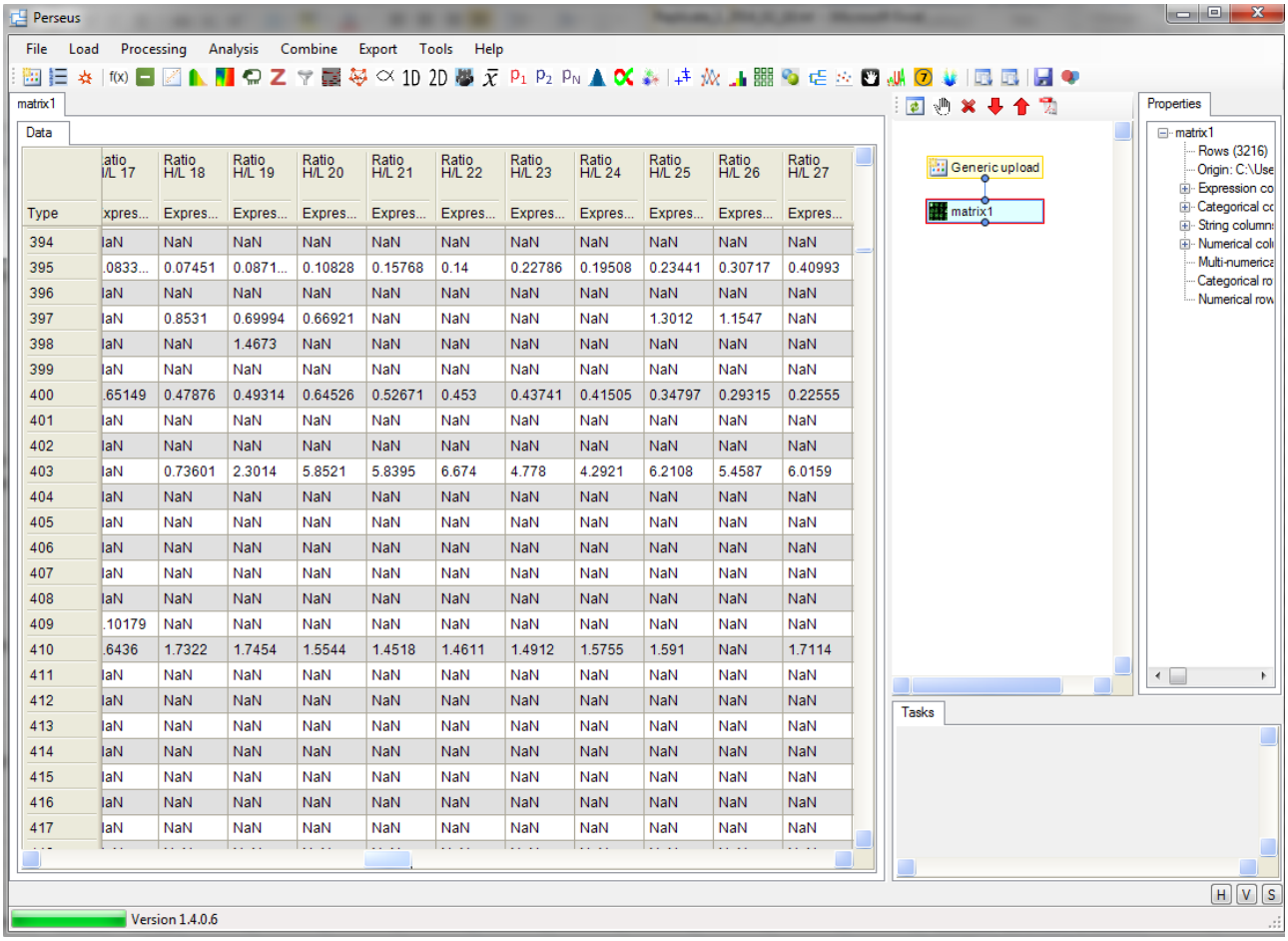
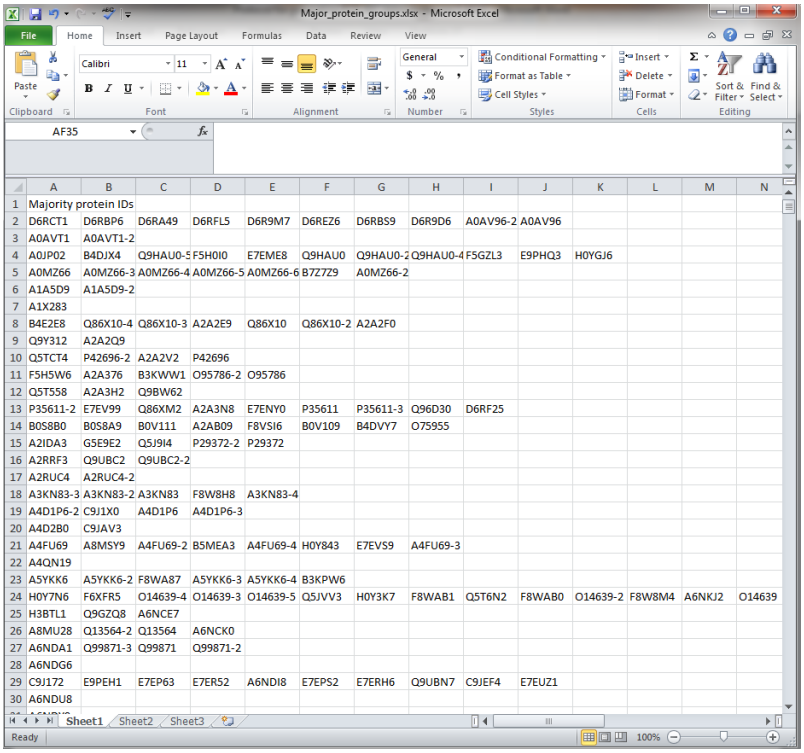


Figure 3: View of imported spatial proteomic experiments in Maxquant. Each experiment corresponds to a spatial fraction from the experiment enables the generation of profiles across fractions.

-Export file to txt.

-Copy the majority protein ID column from Maxquant to a new .xlsx file and separate the contents (corresponding to individual protein IDs) into individual Excel cells using the “text-to-column” function (see http://support2.microsoft.com/kb/214261). The resulting file will be used during to ROC analysis to determine if any of members of the protein group are known to interact within Corum. The resulting file will look like Figure 4 and can be saved for future use.



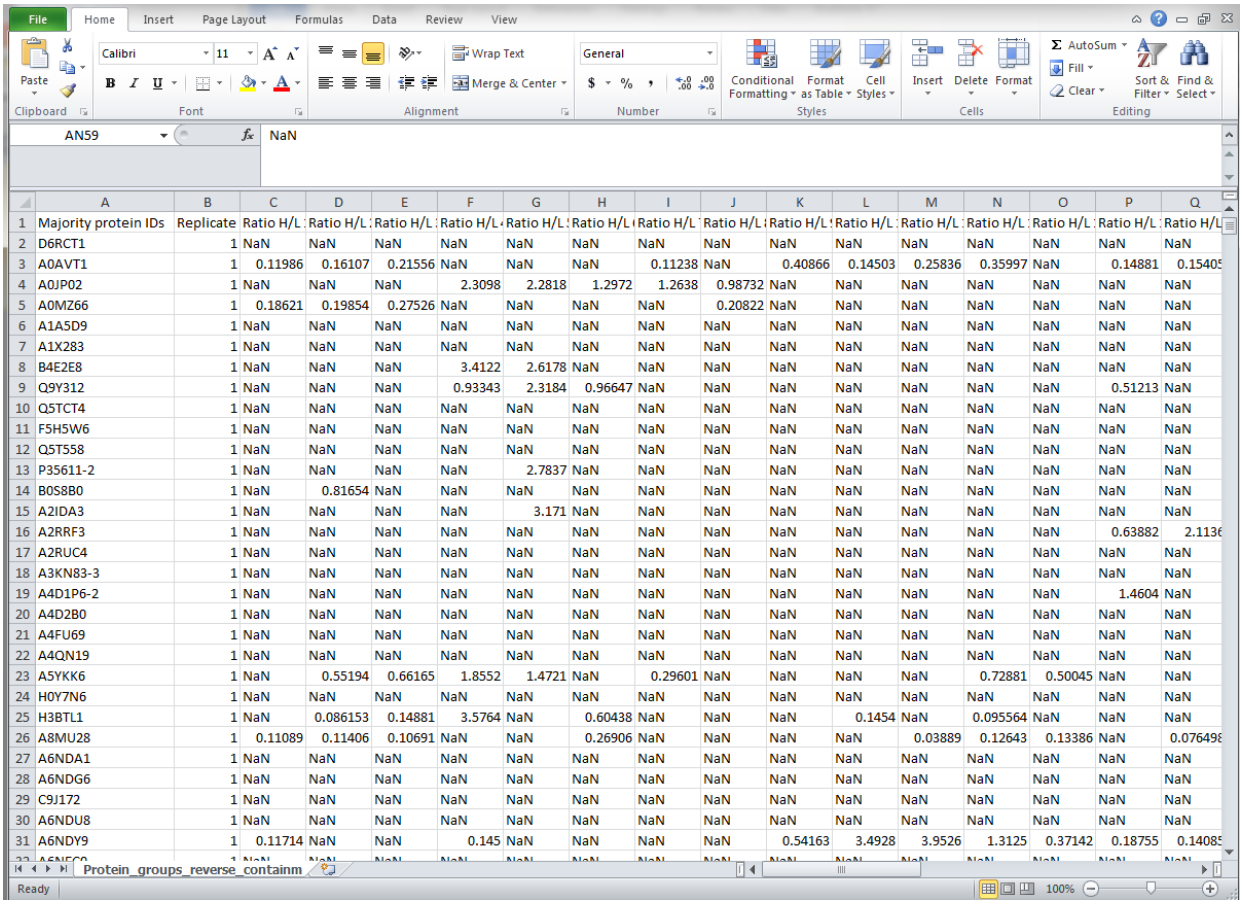
Majority protein IDs													
D6RCT1	D6RBP6	D6RA49	D6RFL5	D6R9M7	D6REZ6	D6RBS9	D6R9D6	A0AV96-2	A0AV96				
A0AVT1	A0AVT1-2												
A0JP02	B4DJX4	Q9HAU0-5	F5H0I0	E7EME8	Q9HAU0	Q9HAU0-2	Q9HAU0-4	F5GZL3	E9PHQ3	H0YGJ6			
A0MZ66	A0MZ66-3	A0MZ66-4	A0MZ66-5	A0MZ66-6	B7Z7Z9	A0MZ66-2							
A1A5D9	A1A5D9-2												
A1X283													
B4E2E8	Q86X10-4	Q86X10-3	A2A2E9	Q86X10	Q86X10-2	A2A2F0							
Q9Y312	A2A2Q9												
Q5TCT4	P42696-2	A2A2V2	P42696										
F5H5W6	A2A376	B3KWW1	O95786-2	O95786									
Q5T558	A2A3H2	Q9BW62											
P35611-2	E7EV99	Q86XM2	A2A3N8	E7ENY0	P35611	P35611-3	Q96D30	D6RF25					
B0S8B0	B0S8A9	B0V111	A2AB09	F8VSI6	B0V109	B4DVY7	O75955						
A2IDA3	G5E9E2	Q5J9I4	P29372-2	P29372									
A2RRF3	Q9UBC2	Q9UBC2-2											
A2RUC4	A2RUC4-2												
A3KN83-3	A3KN83-2	A3KN83	F8W8H8	A3KN83-4									
A4D1P6-2	C9J1X0	A4D1P6	A4D1P6-3										
A4D2B0	C9JAV3												
A4FU69	A8MSY9	A4FU69-2	B5MEA3	A4FU69-4	H0Y843	E7EVS9	A4FU69-3						
A4QN19													
A5YKK6	A5YKK6-2	F8WA87	A5YKK6-3	A5YKK6-4	B3KPW6								
H0Y7N6	F6XFR5	O14639-4	O14639-3	O14639-5	Q5JVV3	H0Y3K7	F8WAB1	Q5T6N2	F8WAB0	O14639-2	F8W8M4	A6NKJ2	O14639
H3BTL1	Q9GZQ8	A6NCE7											
A8MU28	Q13564-2	Q13564	A6NCK0										
A6NDA1	Q99871-3	Q99871	Q99871-2										
A6NDG6													
C9J172	E9PEH1	E7EP63	E7ER52	A6NDI8	E7EPS2	E7ERH6	Q9UBN7	C9JEF4	E7EUZ1				
A6NDU8													

Figure 4: Majority Protein IDs, formatted for use in ROC analysis.

-In order process the PCP SEC SILAC experimental data within the designed MATLAB script the data must be formatted into a verbose form in which the isotopologue ratio to be subjected to MATLAB processing are separated into individual rows for each experiment as shown in figure 5 and 6.

Majority Protein ID	Replicate number	Fraction number 1	Fraction number 2	.....	Fraction number 48	Fraction number 49	Fraction number 50
D6RCT1	1						
A0AVT1	1						
A0JP02	1						
...	...						
...	...						
...	...						
D6RCT1	2						
A0AVT1	2						
A0JP02	2						
...	...						
...	...						
...	...						
D6RCT1	3						
A0AVT1	3						
A0JP02	3						

Figure 5: Diagram of formatting of Maxquant output into a verbose form, each individual experiment is divided up enabling the processing of all protein profiles within a single computation session.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Majority protein IDs	Replicate	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L	Ratio H/L
2	D6RCT1	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	A0AVT1	1	0.11986	0.16107	0.21556	NaN	NaN	NaN	0.11238	NaN	0.40866	0.14503	0.25836	0.35997	NaN	0.14881	0.1540
4	A0JP02	1	NaN	NaN	NaN	2.3098	2.2818	1.2972	1.2638	0.98732	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	A0MZ66	1	0.18621	0.19854	0.27526	NaN	NaN	NaN	NaN	0.20822	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	A1A5D9	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	A1X283	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	B4E2E8	1	NaN	NaN	NaN	3.4122	2.6178	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	Q9Y312	1	NaN	NaN	NaN	0.93343	2.3184	0.96647	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.51213	NaN
10	Q5TCT4	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	F5H5W6	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	Q5T558	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13	P35611-2	1	NaN	NaN	NaN	NaN	2.7837	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
14	B0S8B0	1	NaN	0.81654	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
15	A2IDA3	1	NaN	NaN	NaN	NaN	3.171	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
16	A2RRF3	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.63882	2.113
17	A2RUC4	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
18	A3KN83-3	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19	A4D1P6-2	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.4604	NaN
20	A4D2B0	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
21	A4FU69	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
22	A4QN19	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
23	A5YKK6	1	NaN	0.55194	0.66165	1.8552	1.4721	NaN	0.29601	NaN	NaN	NaN	NaN	0.72881	0.50045	NaN	NaN
24	H0Y7N6	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25	H3BTL1	1	NaN	0.086153	0.14881	3.5764	NaN	0.60438	NaN	NaN	NaN	0.1454	NaN	0.095564	NaN	NaN	NaN
26	A8MU28	1	0.11089	0.11406	0.10691	NaN	NaN	0.26906	NaN	NaN	NaN	NaN	0.03889	0.12643	0.13386	NaN	0.07649
27	A6NDA1	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
28	A6NDG6	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
29	C9J172	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
30	A6NDU8	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
31	A6NDY9	1	0.11714	NaN	NaN	0.145	NaN	NaN	NaN	NaN	0.54163	3.4928	3.9526	1.3125	0.37142	0.18755	0.1408

Figure 6: Example of Maxquant output formatted in verbose format for MATLAB processing.

-Once generated the verbose Maxquant output can be used for MATLAB processing.

<u>Output from preprocessing:</u>

-Verbose Maxquant outputs for each isotopologue ratio combination (i.e MvsL, HvsL and HvsM).

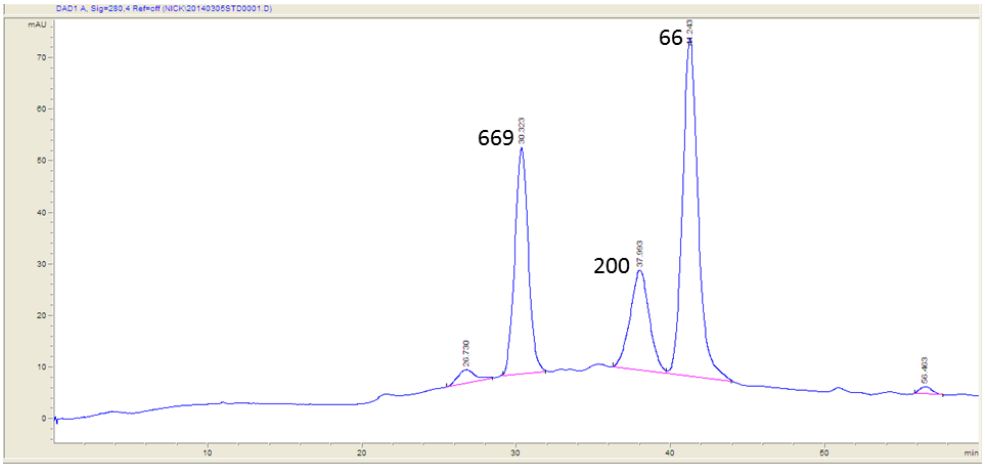
-Majority Protein ID file.

II) Generation of SEC calibration file

Rational: Using protein standards of known size a size calibration can be created and used to determine the size of observed protein complexes.

Protocol:

Setting up SEC calibration file, Run complex standard (such as Protein Standard Mix 15 - 600 kDa, Sigma Aldrich, product number: 69385-30MG) to generate a calibration curve to determine the observed size of protein complexes (Figure 7). Generate an .xlsx containing the fraction number of the apex of the eluting standard and the size of the complex (Figure 8).



Standards :(Thyroglobulin [669kDa], Beta-amylase [200kDa], Bovine Serum Albumin [66kDa])
LC conditions: two Biosep-4000 columns run in series at 0.5ml/min @4C.
Buffer: 50mM Tris, 50mM sodium Acetate, 50mM KCl, pH 7.2

Figure 7: Protein complex standard run on a two Biosep-4000 columns set up.

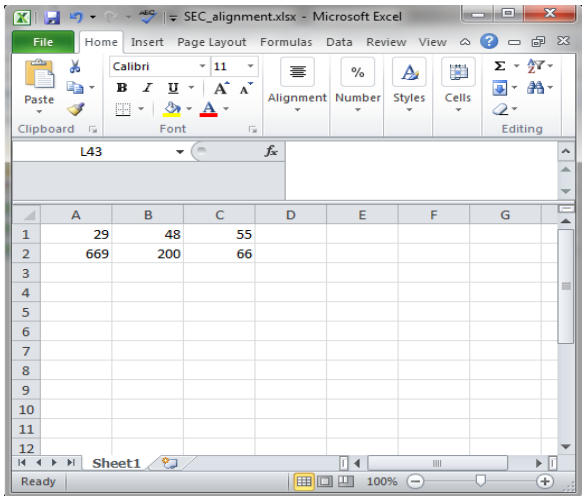


Figure 8: SEC calibration files for the determination of the observed size of the complex.

Output from SEC calibration:

-SEC calibration file

III) Generation of binary interaction list from Corum database for receiver operating characteristic (ROC) curve:

Rational: Interactions are determined by PCP SEC SILAC based on the principle of co-migration, which implies protein observed to interaction should share similar co-migration properties. This implication results in the similarity of co-migration being able to be used to determine interactions. How similar (determined by Euclidean distance) profiles needs to be to be said to interaction is based on calculating the observed precision (TP/(TP+FP)), recall / true positive rate (TP/(TP+FN)) and false positive rate (FP/(TP+TN)) of proteins known to interactions within the dataset. These known interactions are based on the protein interactions within the CORUM database (see http://mips.helmholtz-muenchen.de/genre/proj/corum) which we use as reference interaction database. By setting a required threshold (largely arbitrary in nature but ideally being high enough that the majority of observed interactions are correct and similar to the precision of affinity purification approaches [69% precision as calculated for yeast by Krogan *et al*, Nature 2006]) the thresholds of similar can be determined. Thus to determine precision the reference database must be formatted into the required form for downstream analysis.

Before beginning download

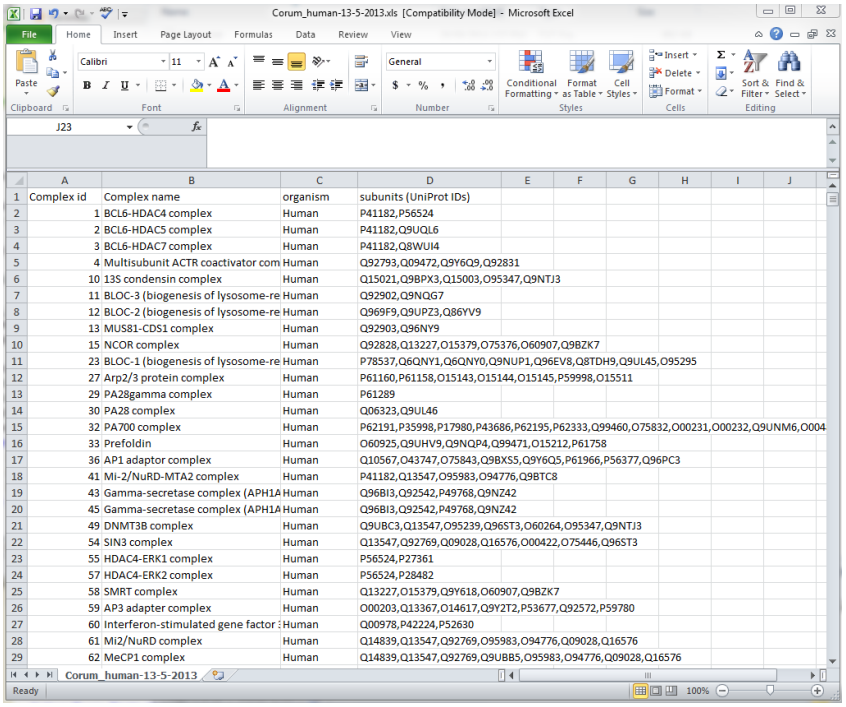
Script to be used: Corum_to_Binary_version2.m (Current build: version 2, 2014)

Corum database: latest build as of September 2014, Corum Release February 2012 (available from http://mips.helmholtz-muenchen.de/genre/proj/corum/allComplexes.csv)

Protocol:

Prior to determining the interactions within a dataset a reference set of interaction must be generated.

Download the latest Corum release and format the dataset into four column containing the Complex id, complex name, organism and members of the complex. Filter our any entries corresponding to undesired species. An example of this formatting is shown below (Figure 9) and within the examples reference database.

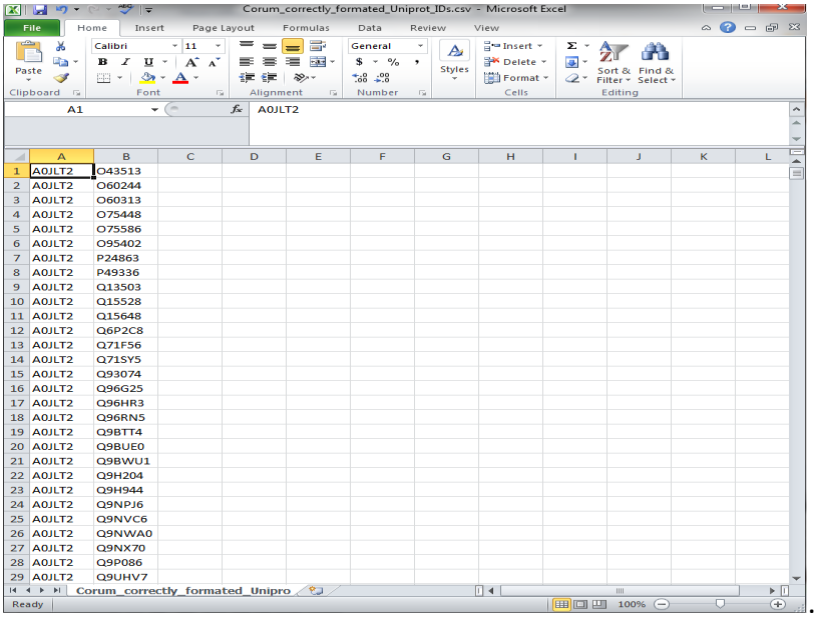


Complex id	Complex name	organism	subunits (UniProt IDs)
1	BCL6-HDAC4 complex	Human	P41182,P56524
2	BCL6-HDAC5 complex	Human	P41182,Q9UQL6
3	BCL6-HDAC7 complex	Human	P41182,Q8WUI4
4	Multisubunit ACTR coactivator com	Human	Q92793,Q09472,Q9Y6Q9,Q92831
10	13S condensin complex	Human	Q15021,Q9BPX3,Q15003,O95347,Q9NTJ3
11	BLOC-3 (biogenesis of lysosome-re	Human	Q92902,Q9NQG7
12	BLOC-2 (biogenesis of lysosome-re	Human	Q969F9,Q9UPZ3,Q86YV9
13	MUS81-CDS1 complex	Human	Q92903,Q96NY9
15	NCOR complex	Human	Q92828,Q13227,O15379,O75376,O60907,Q9BZK7
23	BLOC-1 (biogenesis of lysosome-re	Human	P78537,Q6QNY1,Q6QNY0,Q9NUP1,Q96EV8,Q8TDH9,Q9UL45,O95295
27	Arp2/3 protein complex	Human	P61160,P61158,O15143,O15144,O15145,P59998,O15511
29	PA28gamma complex	Human	P61289
30	PA28 complex	Human	Q06323,Q9UL46
32	PA700 complex	Human	P62191,P35998,P17980,P43686,P62195,P62333,Q99460,O75832,O00231,O00232,Q9UNM6,O004
33	Prefoldin	Human	O60925,Q9UHV9,Q9NQP4,Q99471,O15212,P61758
36	AP1 adaptor complex	Human	Q10567,O43747,O75843,Q9BXS5,Q9Y6Q5,P61966,P56377,Q96PC3
41	MI-2/NuRD-MTA2 complex	Human	P41182,Q13547,O95983,O94776,Q9BTC8
43	Gamma-secretase complex (APH1A	Human	Q96BI3,Q92542,P49768,Q9NZ42
45	Gamma-secretase complex (APH1A	Human	Q96BI3,Q92542,P49768,Q9NZ42
49	DNMT3B complex	Human	Q9UBC3,Q13547,O95239,Q96ST3,O60264,O95347,Q9NTJ3
54	SIN3 complex	Human	Q13547,Q92769,Q09028,Q16576,O00422,O75446,Q96ST3
55	HDAC4-ERK1 complex	Human	P56524,P27361
57	HDAC4-ERK2 complex	Human	P56524,P28482
58	SMRT complex	Human	Q13227,O15379,Q9Y618,O60907,Q9BZK7
59	AP3 adapter complex	Human	O00203,Q13367,O14617,Q9Y2T2,P53677,Q92572,P59780
60	Interferon-stimulated gene factor	Human	Q00978,P42224,P52630
61	Mi2/NuRD complex	Human	Q14839,Q13547,Q92769,O95983,O94776,Q09028,Q16576
62	MeCP1 complex	Human	Q14839,Q13547,Q92769,Q9UBB5,O95983,O94776,Q09028,Q16576

Figure 9: Example of formatted Corum release for processing with Corum_to_Binary _version2.m

-Run script the Corum_to_Binary.m, (estimated processing time <5minutes)

-The resulting text file will be named “Corum_correctly_formated_Uniprot_IDs.csv” and contains a binary list denoting the known interactions in the Corum database (Figure 10)



A0JLT2	O43513
A0JLT2	O60244
A0JLT2	O60313
A0JLT2	O75448
A0JLT2	O75586
A0JLT2	O95402
A0JLT2	P24863
A0JLT2	P49336
A0JLT2	Q13503
A0JLT2	Q15528
A0JLT2	Q15648
A0JLT2	Q6P2C8
A0JLT2	Q71F56
A0JLT2	Q71SY5
A0JLT2	Q93074
A0JLT2	Q96G25
A0JLT2	Q96HR3
A0JLT2	Q96RN5
A0JLT2	Q9BTT4
A0JLT2	Q9BUE0
A0JLT2	Q9BWU1
A0JLT2	Q9H204
A0JLT2	Q9H944
A0JLT2	Q9NPJ6
A0JLT2	Q9NVC6
A0JLT2	Q9NWA0
A0JLT2	Q9NX70
A0JLT2	Q9P086
A0JLT2	Q9UHV7

Figure 10: Example of binary interaction list generated from Corum for ROC analysis.

Part B: Processing of data with MATLAB

I) Generation of Gaussian fitted curves from protein profile data:

Rational: Although proteins known to interact co-migrate the assumption that proteins will always be found associated is inconsistent with our current understanding of protein complex (an example of this is the 19S and 20S proteasome). It is know that multiple unique protein association exist within the cell thus instead of only considering the complete protein profile individual features of the protein profile should be compared. To do this deconvolution into fitted Gaussian curves can be used and these individual features normalised and compared.

Before beginning download

Script to be used: Gaus.m (Current build: version 2, 2014)

Verbose MaxQuant output: Formatted into individual experiments protein profile experiments, and saved as a .xlsx file)

SEC calibration file: Formatted as above and saved as a.xlsx

Protocol:

-Open Gaus.m and entry the following information (see figure 13):

1) Location and name of the verbose MaxQuant outputs

2) Location and name of sec calibration file

3) Number of iteration (for reproducible results the initiation number should be ~10x the number of fractions, thus for 50 fractions 500 iteration should be used. This will lead to the saturation of combinations and the generation identical results every time the script is run).

4) First fraction to consider for Gaussian curve fitting: suggested to between 2 and 5, this can be tailored to avoid large complexes that may not be of interest such as the ribosome.

5) Last fraction to consider for Gaussian curve fitting: suggested to be the last fraction but can be increase if it is found the last fraction is below 100 kDa, ideally only fractions over 100 kDa should be considered.

6) Number of experimental channels is determined by the number of different isotopologue ratio to be compared (for an experiment in which both the MvsL and the HvsL are used to generate protein profile the number is two)

-The script is designed to run without further input from the user and ideally should be run on a multi-core computer to allow parallelization of the processing. Parallelization, which is facilitated by the parallel toolbox within MATLAB, will significantly increase the speed processing. The estimated time of this script depends on the number of proteins, number of data points observed for each protein group, replicates, iterations and cores available.

Output from Gaus.m script: (All files will be generated for each experimental channel used for protein profile fitting and place in individual folders corresponding to the experimental channels)

- Combined_Chromatograms.csv: Contains the results of the Gaussian fitting curves for each protein which satisfies the following conditions: apex height of over 0.5, width greater than one fraction, center greater than the start fraction and less than the last fractions.
- Combined_Chromatograms_filtered_out.csv: Contains the results of the Gaussian fitting curves for each protein which do not satisfies the following conditions: apex height of over 0.5, width greater than one fraction, center greater than the start fraction and less than the last fractions.

- Combined_OutputGaus.csv: Contains the height, center, width, SSE (Sum squared error), adjrsquare (R^2 value of Gaussian fitted) and observed complex size for all fitted Gaussians curves which satisfies the following conditions: apex height of over 0.5, width greater than one fraction, center greater than the start fraction and less than the last fractions.

- Combined_OutputGaus_filtered_out.csv: Contains the height, center, width, SSE (Sum squared error), adjrsquare (R^2 value of Gaussian fitted) and observed complex size for all fitted Gaussians curves which do not satisfies the following conditions: apex height of over 0.5, width greater than one fraction, center greater than the start fraction and less than the last fractions.

- Summary_Gausians_for_individual_proteins.csv: Contains the number of Gaussians curves detected for each protein from each replicate. The numbers of Gaussians curves are also divided in the number of curves which are filtered (contained within Combined_OutputGaus_filtered_out.csv) and not filtered (contained within Combined_OutputGaus.csv).

- Summary_Gausians_identifed.csv: Contains the summary results from the Gaussians curve fitting, displays the total number of proteins assessed, the mean number of quantitative values in the protein profiles (prior to fixing missing values flanked by quantitation), mean number of quantitative values in the protein profiles after fixing missing values flanked by quantitation; The number of proteins which were subjected to fitting but found to fit zero Gaussians (0 Gausian), The number of proteins which were subjected to fitting but found to one Gaussians (1 Gausian), The number of proteins which were subjected to fitting but found to two Gaussians (2 Gausian), The number of proteins which were subjected to fitting but found to three Gaussians (3 Gausian), The number of proteins which were subjected to fitting but found to four Gaussians (4 Gausian), The number of proteins which were subjected to fitting but found to

five Gaussians (5 Gausian), number of proteins which were subjected to Gaussian fitting (No_try_fit).

- Summary_Proteins_with_Gausians.csv: Contains the number of Gaussians curves detected for onlu protein with detected Gaussian curve from each replicate. The numbers of Gaussians curves are also divided in the number of curves which are filtered (contained within Combined_OutputGaus_filtered_out.csv) and not filtered (contained within Combined_OutputGaus.csv).

- Proteins_not_fitted_to_gaussian.csv: Contains the protein number, the protein Uniprot ID, replicate number and all observed isotopologue ratio for proteins which have greater than five isotopologue but were unable to be fit to a Gaussian curve (these are protein denoted as 0 Gausian in Summary_Gausians_identifed.csv).

Error_check folder: This folder will contain files corresponding to any protein which lead to an error during processing, if any files are outputted as error is found please report to it to the Foster lab.

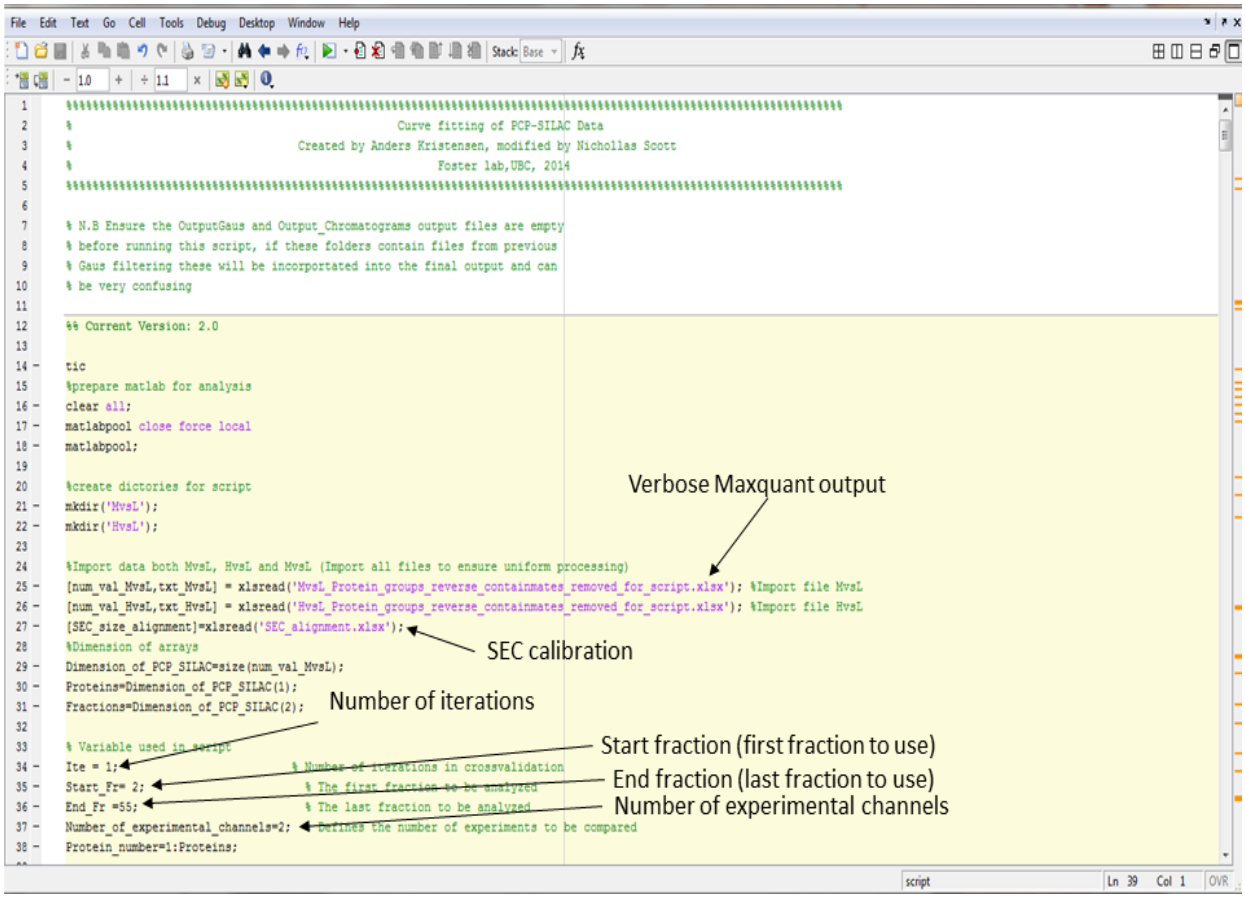


Figure 13: Setting up Gaus.m script for processing. The User required input are shown and are entered prior to runs the script.

II) Alignment of replicates using Gaussian fitted curves:

Rational: As replicates of PCP-SEC-SILAC experiments may be performed on different LC or columns small, but significant, changes in the retention of time may occur. As the order of elution of complexes will not change (provided the chromatography media is identical) these changes can be corrected by aligned elution profiles.

Before beginning download

Script to be used: Alignment.m (Current build: version 2, 2014)

Output folders from Gaus.m: The script is designed to read the content of the folder and generate the required alignment information.
Verbose MaxQuant output: Formatted into individual experiments protein profile experiments, and saved as a .xlsx file)
SEC calibration file: Formatted as above and saved as a.xlsx

Protocol:

-Open Alignment.m and entry the following information (see figure 14):

1) User defined window one, used for the first alignment between replicates. This window is set width to generate an alignment curve (based on polynomial fitting) based on the majority of fitted Gaussian curves common between replicates. In general we have found a value between 10 and 20 fractions adequate.

2) User defined window two, used for the second alignment between replicates. This window is set narrow to generate an alignment curve (based on density fitting) based on the density of fitted Gaussian curves common between replicates. In general we have found a value between 2 and 4 fractions adequate for SEC fractions and up to eight fractions for BN-PAGE separated complexes.

3) Number of experimental channels is determined by the number of different isotopologue ratio to be compared (for an experiment in which both the MvsL and the HvsL are used to generate protein profile the number is two)

4) The Alignment channel to use for the re-alignment of multiple protein profile isotopologue channels. It was noted that if the MvsL and HvsL protein profiles were re-aligned independent of each the small difference between the final fitting equations lead to problems in downstream processing. To avoid this we implemented a parameter which would allow the user to specify the alignment to be used.

5) Location and name of the verbose MaxQuant outputs

6) Location and name of SEC calibration file

-The script is designed to run without further input from the user and be run on single core computer. (Estimated processing time <10minutes)

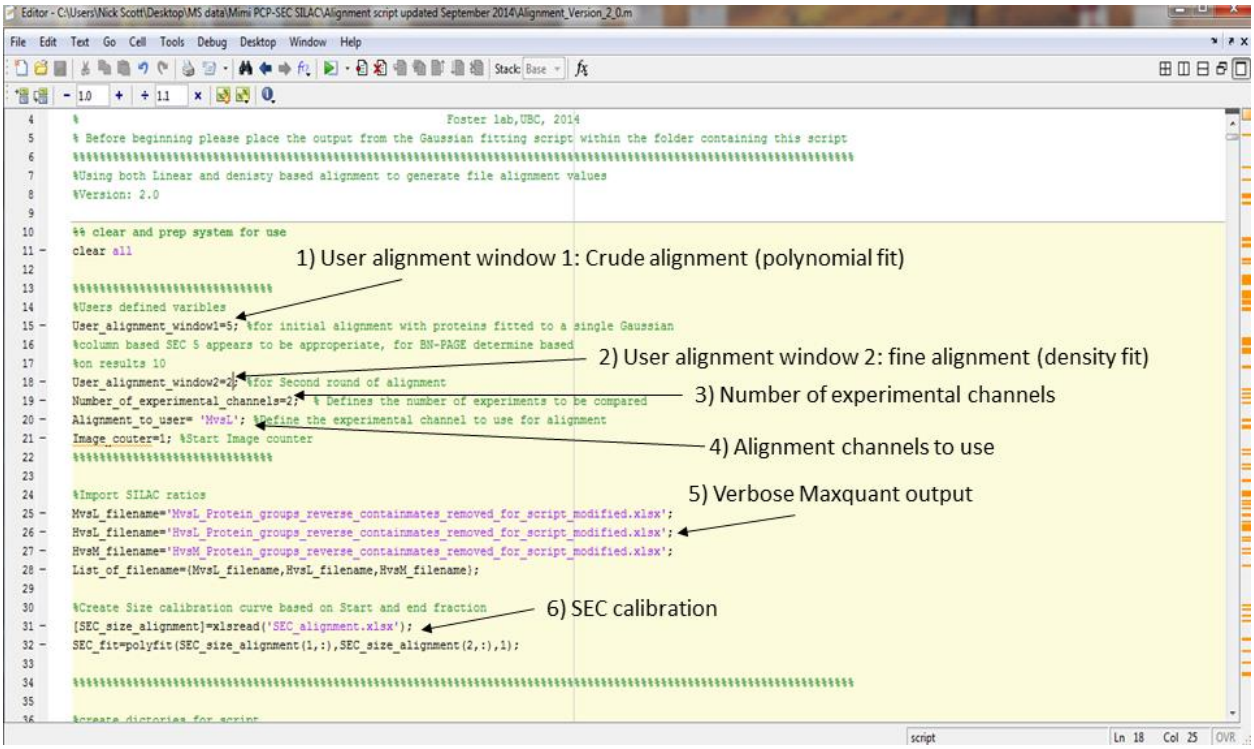


Figure 14: Setting up Alignment.m script for processing. The User required input are shown and are entered prior to runs the script.

Note: Examine the output from the alignment.m script to determine if the User defined windows are adequate. If the separation conditions are very different between replicates a wider user window one is required (See figure 15).

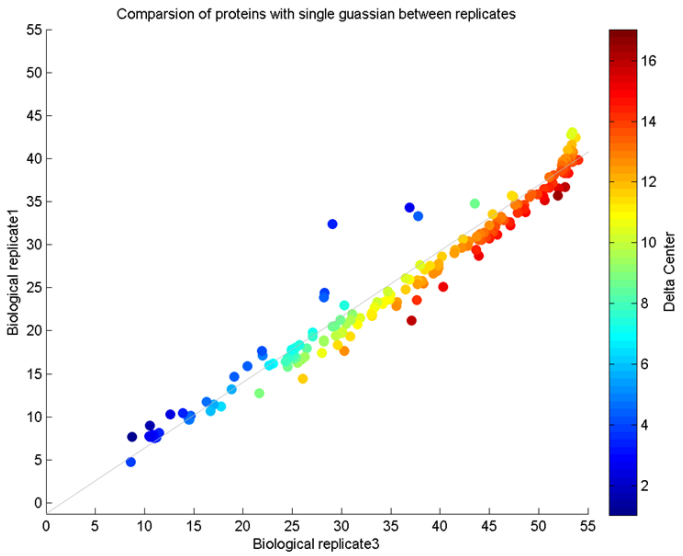


Figure 15: example of alignment.m output showing replicate PCP SILAC experiments run under different separation conditions.

Output from Alignment.m script: (All files will be generated for each experimental channel used for protein profile fitting and place in individual folders corresponding to the experimental channels)

-Image summary.csv: Provides the description of all figures generated from the alignment script. This includes the image name and description of which replicate are being compared.

-Alignment figure folder: All images described within the image summary are deposited within this folder.

-Adjusted fraction overlap summary.csv (within the Combined Realigned data folder generated during analysis): Contents the results of the alignment of each replicates, the replicate names are provided and the alignment of each fraction given as compared to the reference replicate. The reference replicate is the replicate which contains the most protein Gaussian curve overlap with all over replicates.

-Adjusted HvsM Raw data maxquant.csv (within the Combined Realigned data folder generated during analysis): Contains the realigned HvsM isotopologue data.

-Adjusted MvsL Raw data maxquant.csv (within the Combined Realigned data folder generated during analysis): Contains the realigned MvsL isotopologue data.

-Adjusted MvsL Combined OutputGaus.csv (within the Combined Realigned data folder generated during analysis): Contains the realigned Gaussian curve fits for parameters (height, center, width, SSE, adjrsquare and observed complex size) for all fitted Gaussians curves.

-Comparison of Gaussian properties between replicates rep# compared to_rep#.csv (within the Comparison tables folder generated during analysis): Contains the results of the comparison between Gaussian curves observed within different replicates. The height, center, width, SSE and adjrsquare of each Gaussian curve observed between replicates are provided as well as the delta height, delta center, delta width and delta Euclidean distance.

-Comparison of Single Gaussian between replicates #to# data within user settings.csv: (within the Comparison tables folder generated during analysis): Contains the results of the comparison between Gaussian curves observed within different replicates within user defined thresholds. The user threshold is determined empirically based observing how the threshold affects the regression analysis of the fit and if this fit includes the majority of the observed Gaussians curves. The height, center, width, SSE and adjrsquare of each Gaussian curve observed between replicates are provided as well as the delta height, delta center, delta width and delta Euclidean distance.

-MvsL_Raw_data_maxquant_rep#.csv (within Processed Gaussian folder generated during analysis): Prior to analysis the MvsL isotopologue values for individual replicates are derived up into their respective replicates, these files are loaded into the alignment script to perform the required analysis.

-HvsL_Raw_data_maxquant_rep#.csv (within Processed Gaussian folder generated during analysis): Prior to analysis the HvsL isotopologue values for individual replicates are derived up into their respective replicates, these files are loaded into the alignment script to perform the required analysis.

-HvsM_Raw_data_maxquant_rep#.csv (within Processed Gaussian folder generated during analysis): Prior to analysis the HvsM isotopologue values for individual replicates are derived up into their respective replicates, these files are loaded into the alignment script to perform the required analysis.

-Summary Gaussians for individual proteins rep#.csv (within Processed Gaussian folder generated during analysis): Prior to analysis the summary file of Gaussians curves identified for individual proteins are derived up into their respective replicates, these files are loaded into the alignment script to perform the required analysis.

-Adjusted Chromatograms vobose rep# .csv (Within the Realignment folder generated during analysis): Contains the adjusted Gaussian curves profiles for each protein divided up into individual replicates.

-Adjusted HvsM Raw data maxquant rep#.csv (Within the Realignment folder generated during analysis): Contains the realigned HvsM isotopologue data divided up into individual replicates.

-Adjusted MvsL Raw data maxquant rep#.csv (Within the Realignment folder generated during analysis Contains the realigned MvsL isotopologue data divided up into individual replicates.

-Adjusted MvsL Combined OutputGaus rep#.csv (Within the Realignment folder generated during analysis): Contains the realigned Gaussian curve fits for parameters (height, center, width, SSE, adjrsquare and observed complex size) for all fitted Gaussians curves divided up into individual replicates.

-Alignment_table.csv (Within the Realignment folder generated during analysis): Contains the list of which replicates have been adjusted and the replicate which was used for alignment.

-fraction overlap summary MvsL replicate#.csv (Within the Realignment folder generated during analysis): Contains the results of the alignment of the denoted replicates, the replicate names are provided and the alignment of each fraction given as compared to the reference replicate.

-Summary table polynomial fit adjustments for replicates global.csv (Within the Realignment folder generated during analysis): Contains the results of the polynomial fit alignment of all replicates. The replicate numbers, indeterminates, Coefficients, R value and R squared are provided for each replicate.

-Summary table polynomial fit adjustments for replicates user defined threshold.csv (Within the Realignment folder generated during analysis): Contains the results of the polynomial fit alignment of all replicates after filtering out Gaussians curves that fall outside the user defined thresholds. The replicate numbers, indeterminates, Coefficients, R value and R squared are provided for each replicate.

-Summary_table_polynomial_fit_of_density_adjustmentss_for_replicates.csv (Within the Realignment folder generated during analysis): Contains the results of the density based fit alignment of all replicates after filtering out Gaussians curves that fall outside the user defined thresholds. The replicate numbers, indeterminates, Coefficients, R value and R squared are provided for each replicate.

-Summary_table_polynomial_fit_both_adjustmetns_for_replicates.csv (Within the Realignment folder generated during analysis): Contains the results of combining both the polynomial fit and density based fit alignment of all replicates after filtering out Gaussians curves that fall outside the user defined thresholds. The replicate numbers, indeterminates, Coefficients, R value and R squared are provided for each replicate.

-Sorted Gaussians for individual proteins with one Gaussian rep#.csv (Within the Summary tables folder generated during analysis): Contains the sorted Gaussian curves for proteins where only a single Gaussian curve was observed. This sorted array is used during the alignment of replicates.

-Summary Gaussians for individual proteins with Gaussian rep#.csv (Within the Summary tables folder generated during analysis): Contains the summary of the Gaussians for individual proteins for the denoted replicates.

-Summary Gaussians for individual proteins with one Gaussian rep#.csv (Within the Summary tables folder generated during analysis): Contains the summary of the Gaussians observed for individual proteins where only a single Gaussian was observed for the denoted replicates.

III) Determination of changes in protein interactions using:

Rational: The analysis of multiple dimensional dataset across biological replicates represents a significant bottleneck within the analysis of PCP-SEC-SILAC experiments. To overcome this limitation an automated data visualization and comparison tool was developed. The goal of this tool was to enable the comparison of changes between isotopologue channels in statically valid way independent of arbitrary cut offs by using multiple hypotheses testing (bonferroni correction). By providing an automated solution to the analysis of PCP-SILAC dataset biological interesting changes are able to be identified in a rapid time frame.

Before beginning download

Script to be used: Comparison.m (Current build: version 2, 2014)

Additional script: mwwtest.m (Used for Mann-Whitney-Wilcoxon testing)

Verbose MaxQuant output: Formatted into individual experiments protein profile experiments, and saved as an .xlsx file. The verbose output can have been processed with the alignment script or can be unprocessed. In cases where a single experiment is being visualized the unprocessed Verbose output and ensure the Alignment binary is set to zero (see Figure 16).

Gaussian property list: corresponding to the height, center, width, SSE, adjrsquare and observed complex size for all fitted Gaussians curves for the isotopologue ratio used to generate protein profiles. Gaussian property list can have been processed with the alignment script or can be unprocessed. In cases where a single experiment is being visualized the unprocessed Verbose output and ensure the Alignment binary is set to zero (see Figure 16).

Protocol:

-Open Comparison.m and entry the following information (see figure 16):

1) Number of replicates to be compared. Note the maximum number of replicates that can be processed currently is four.

2) The first column corresponding to fraction 1 within the imported verbose maxquant output file. If the replicates have been processed with the alignment script then the 7th column is the first fraction as six addition positions (corresponding to -5 to 0) have been added to the realigned verbose maxquant output.

3) Last fraction to be considered for plotting of the protein profile information. Default is 55 for realigned data.

4) Denotes the dilution factor of the isotopologue reference mixture. In typical PCP experiments the amount of reference added to each fractions is diluted by a known amount (typically to 70% within the Foster lab) to improve the overall coverage and depth of experiments. To correct for this dilution the define the dilution amount as the 'Diltuion factor master mix'

5) Binary option (1 for true, 0 for false) which corresponds to if the replicates have been re-aligned with alignment.m. If the data has not been aligned the script will append empty column to the beginning of the dataset to enable the analysis of protein profiles in the same way as realigned data sets.

6) User define window of Gaussian to be considered as the same Gaussian across replicates. The selection of this window should be based on the results of the alignment script; the width of the window should reflect spread of observed Gaussian along the realignment line (Figure 17).

7) Location and name of the Gaussian property list

8) Location and name of the verbose MaxQuant outputs

-The script is designed to run without further input from the user and be run on single core computer. (Estimated processing time <5-6hours to generate images of all protein profiles)

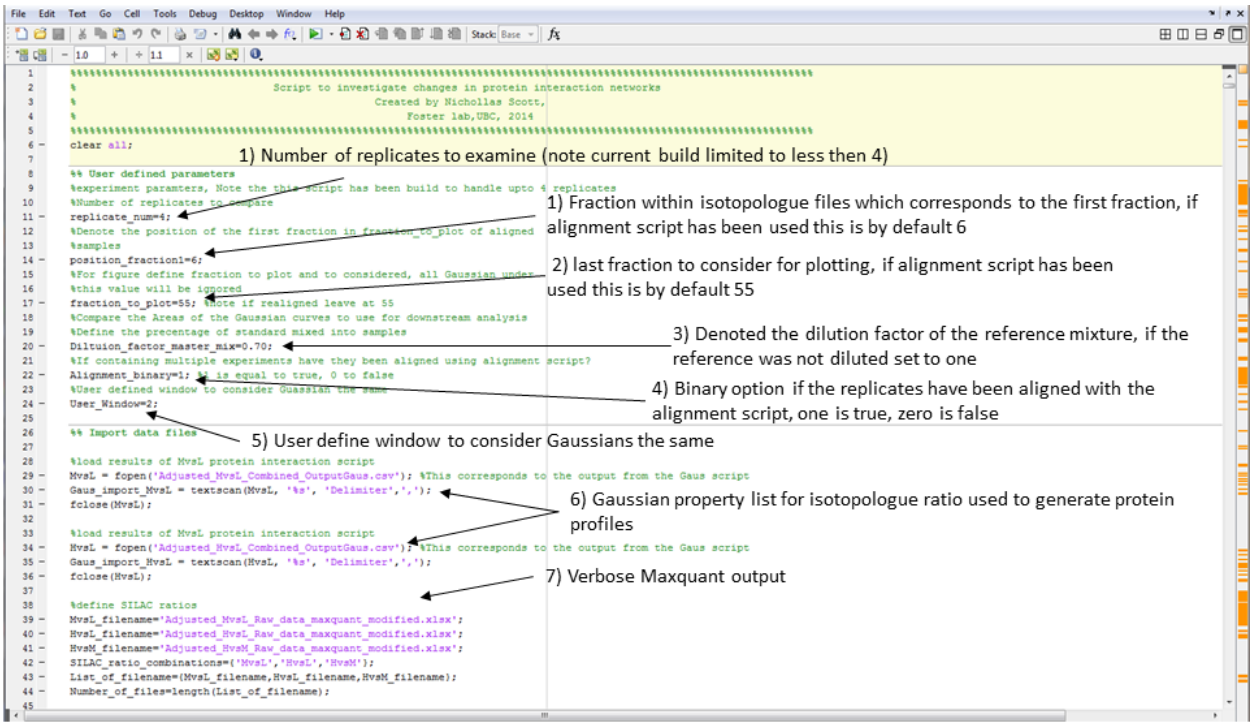


Figure 16. Setting up Comparison.m script for processing. The User required input are shown and are entered prior to runs the script.

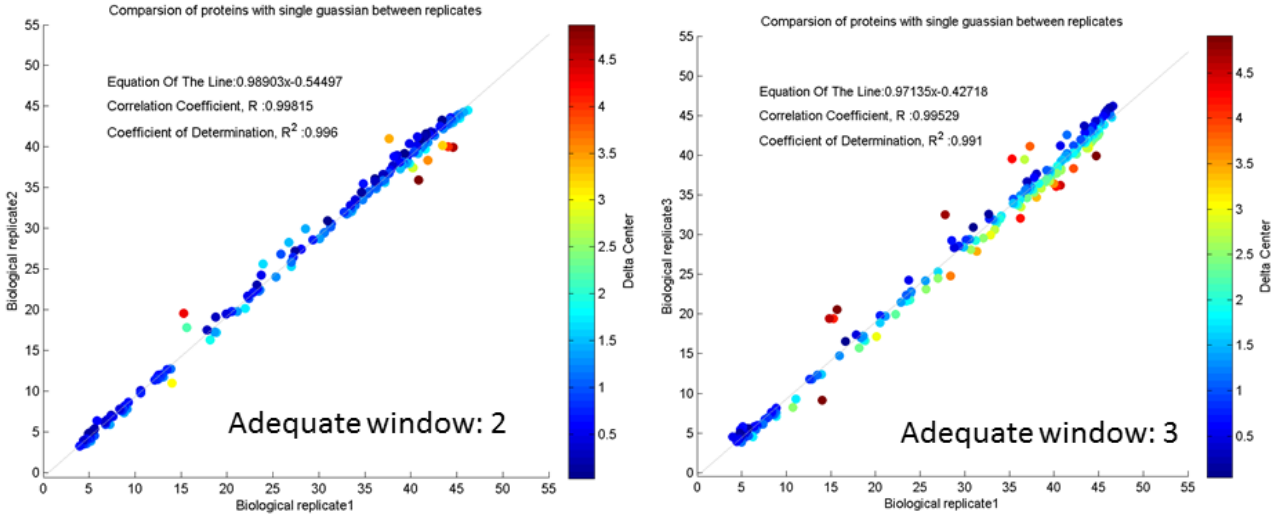


Figure 17. Example of how the output of the alignment script can be used to decide the best user define window to consider Gaussian across replicates as the same. From this example three replicates are aligned and it can be seen that in the figure on the left the majority of observed Gaussian curves are within two fractions while in the figure on the right the majority of Gaussians between replicates are within three fractions. For this experiment an adequate window would be three fractions.

Output from Comparison.m script: (All files will be and deposited in individual folders)

-Figure output folder: Contains all the visual output from the analysis of replicates, this includes

-Pie char summary of overlap between proteins observed within different replicates.

-Summarised graphics for every protein showing the fitted Gaussian curves and all isotopologue information for each replicate.

-Summary of changes in the observed Gaussian fitted curves and across the observed SEC fractions.

-Summary of coverage of observed replicates, both as a percentage of the observed isotopologue amount and percentage of the SEC fractions.

-Summary of coverage of observed replicates, both as a percentage of the observed isotopologue amount and percentage of the SEC fractions grouped based on the number of Gaussian fitted.

-Master_Chromatogram_list.csv (Within Master Gaussian summary folder generated during analysis): Provides a summarised file of all the protein profiles that the Master_guassian_list.csv is derived from. This file contains the protein number, protein name and isotopologue ratios derived from the Gaussian properties.

-Master_guassian_list.csv (Within Master Gaussian summary folder generated during analysis): Contains the protein name, unique identifier, replicate number, observer experimental channel, Gaussian number, height, center, width, SS, adjrsquare and observed complex size for the best fitted Gaussians curves across replicates.

-Perseus_enrichment_Gaussian_level_file.csv (Within Analysis output folder generated during analysis): Contains Gaussian information in a format compatible with Perseus for enrichment analysis. File contains the observed Center, Protein name, Normalised Average fold change, Std dev fold change, Normalised fold change replicate 1, Normalised fold change replicate 2, Normalised fold change replicate 3, Normalised fold change replicate 3, Increase (>= two replicates), Decrease (>= two replicates), p-value (t-test), Below Adjusted_pvalue (bonferroni correction) (ttest), p-value (MWW) and Below Adjusted_pvalue(bonferroni correction) (MWW) for each Gaussian. **Note: performing enrichment analysis at the Gaussian level is not advised as proteins with more than one Gaussian will be counted multiple times leading to an inaccurate measurement of enrichment.**

-Perseus_enrichment_Protein_level_file.csv (Within Analysis output folder generated during analysis): Contains Gaussian information in a format compatible with Perseus for enrichment analysis. File contains the observed Protein name and if any of the Gaussian for this protein

Increase (>= two replicates), Decrease (>= two replicates), p-value (t-test), Below Adjusted_pvalue (bonferroni correction) (ttest), p-value (MWW) and Below Adjusted_pvalue(bonferroni correction) (MWW) for each Gaussian.

-Quantation_values_identified.csv (Within Analysis output folder generated during analysis): Contains the number of Gaussians fitted curves which were quantified within each replicate. This file contains, the total number of Gaussians fitted curves considered to be the same across all replicates and the observed number of the total Gaussians fitted curves quantified within each replicate.

-Student_ttests_results.csv (Within Analysis output folder generated during analysis): Contains the result for the two sided student t-test used to compare isotopologue channels to determine changes. This file contains information on the best Gaussian fitted curve observed across replicates and reports the Channel, Guassian_index_number, Center, Height , Width, SSE, adjrsquare, Gaussian Area, Complex Size. In addition to this information the Fold change values for each replicate are given, the resulting p-value (ttest), Std Dev of fold change, Number of observation (corresponding to the number of replicates the Gaussian were observed in), Degrees of freedom and if the observed p-value satisfies the Bonferroni correction (Below Adjusted_pvalue).

- Summary_changes_based_master_gaussian_list.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of the observed fitted Gaussian curves which increase (greater than arbitrary value of 1 log2) in at least two replicates, decreased (greater than arbitrary value of -1 log2) in at least two replicates, showed no change or showed inconsistent changes across replicates (only one replicate that changes compared to other experiments or replicates which show conflicting changes).

- Summary_changes_based_master_gaussian_list_across_replicate_#.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of the observed fitted Gaussian curves which increase (greater than arbitrary value of 1 log2), decreased (greater than arbitrary value of -1 log2), showed no change, were Unquantifiable as they were observed only in MvsL channel, Unquantifiable as they observed only in HvsL channel and Unquantifiable as they were not observed in replicate.

-Summary_changes_Protein_across_replicate.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of the changes at the protein level, designed to determine if all the observed Gaussian for a protein decrease all increase within replicates. This file contains information on the number of proteins that Increase (in atleast two replicates), Decrease (in atleast two replicates), No change (in atleast two replicates), +/- (in atleast two replicates) and is Inconsistent across replicates.

-Summary_observed_protein_PCP_SILAC_coverage_within_experiments.csv (Within Analysis output folder generated during analysis): Contains the summary of the Average PCP_SILAC coverage, Average coverage of the SEC, Median PCP_SILAC coverage, Number of proteins with a coverage less than 50%, Number of proteins with a coverage less than 70%, Number of proteins with a coverage less than 85%, Number of proteins with a coverage less than 100%, Number of proteins with a coverage less than 150%, Number of proteins with a coverage less than 200%, Number of proteins with a coverage greater than 200% and the Maximum coverage observed for each replicate and channel.

-Summary_PCP_SILAC_coverage_Individual_proteins_within_experiments.csv (Within Analysis output folder generated during analysis): Contains the summary of the PCP_SILAC coverage for every protein, experimental channel and replicate. In addition the numbers of Unique Gaussians observed across all replicates are provided.

Master_guassian_list_fold_comparsion.csv (Within Analysis output folder generated during analysis): Contains the summary of all the extracted values for the unique Gaussian fit curves observed between replicates. Based on this information the quantitative values at the apex plus two values either side from each experimental channel are used for quantitation. A summary of the change is show in the columns labeled “Change observed”. In addition to this information the standard derivation, normalized fold change (based on correcting the mean of the experimental channels), Average fold change, Fold change for each fraction, the p-value (ttest), whether the observed t-test p-value is below bonferroni corrected value; the p-value (MWW) and whether the observed MWW test p-value is below bonferroni corrected value.

Gaussian_trend_analysis_protein_replicate.csv (Within Analysis output folder generated during analysis): Contains a summary of Gaussian information observed for each replicate. File recorded the summed area of all Gaussian observed for a single protein in each experimental channel, the percentage of the expected area observed (based on the summed area compared to predicted total area), the number of Gaussian fitted curves observed, if an Observed change is seen and if the observed change is consistent across all Gaussian fitted curves.

Summary_gaussian_trend_analysis_protein_replicate.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of the number of protein observed across each replicate in which a Gaussian curve could be fit. This file shows the total number of proteins observed across each replicate, the number of proteins that show no change within the fitted Gaussian, the number of proteins which show an increase in at all fitted Gaussian, the number of proteins which show an decrease in at all fitted Gaussian, the number of proteins which show an increase in some fitted Gaussian but not all, the number of proteins which show an decrease in some fitted Gaussian but not all and the number of proteins which show an increase and decrease in fitted Gaussians.

Summary_gaussian_detected.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of the overlap in observed Gaussians. This file denotes the total number of Gaussians fitted curves observed within the experiment (redundant between replicates), the total number of Gaussians fitted curves observed within the MvsL (redundant between replicates), the total number of Gaussians fitted curves observed within the HvsL (redundant between replicates), the shared Gaussians fitted curves (within the user defined window across experimental channels), The number of Gaussians fitted curves which increase (redundant between replicates), The number of Gaussians fitted curves which decrease (redundant between replicates), The number of Gaussians fitted curves which do not change (redundant between replicates) and the number of Gaussians fitted curves which could not be Unquantified. **Note: The number of Unquantified Gaussians fitted curves should always be zero, if it is not then the corresponding protein profile does not match Gaussians properties**

Unique_gaussians_with_changes.csv (Within Analysis output folder generated during analysis): Contains the list of observed Gaussians within each replicate. This file denotes the protein name, Height, Center, Width, SSE, adjrsquare, if the Gaussian was observed in either experimental channels or only one, if a change was observed in the amount of each Gaussian fitted curve between experimental channels, area observed for unique Gaussians, area coverage of unique Gaussians (as compared to the expected area), fold change (based on raw data) and normalise Fold Change (based on raw data).

Summary_gaussian_detected_between_replicates.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of the overlap in observed Gaussians. This file denotes the total number of Gaussians fitted curves observed within the experiment (redundant between replicates), the total number of Gaussians fitted curves observed within the MvsL (redundant between replicates), the total number of Gaussians fitted curves observed within the HvsL (redundant between replicates), the shared Gaussians fitted curves (within the user defined window across experimental channels), the average coverage of the observed proteins within MvsL and the average coverage of the observed proteins within HvsL.

Unique_gaussians_observed_between_replicates.csv (Within Analysis output folder generated during analysis): Contains the list of observed Gaussians within each replicate. This file denotes the protein name, Height, Center, Width, SSE, adjrsquare, if the Gaussian was observed in either experimental channels or only one, the maximum area observed for the fitted Gaussian curve, Area observed for MvsL Gaussians, Area observed for HvsL Gaussians, Area coverage of the maximum area observed fitted Gaussian curve as a percentage of the expected area, Area coverage of MvsL Gaussians and Area coverage of HvsL Gaussians

Protein_gaussian_observed_in_each_replicate.csv (Within Analysis output folder generated during analysis): Contains the list of all protein in which Gaussians fitted curves were observed and whether these proteins were observed within each replicate and experimental channel.

Protein_observed_in_each_replicate.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of how many overlap Gaussians fitted curves were observed in each experimental channel, including: the number of protein observed in both channels, the number observed only within the MvsL channel, the number observed only within the HvsL channel; proteins only observed once, twice, three time and four times in the MvsL/HvsL channels.

Protein_observed_in_each_replicate_and_channels.csv (Within Analysis output folder generated during analysis): Contains a numerical summary of how many overlap Gaussians fitted curves were observed in each experimental channel and replicates, including: the number of protein observed in both channels, the number observed only within the MvsL channel, the number observed only within the HvsL channel; proteins only observed once, twice, three time and four times in the MvsL/HvsL channels as well the numbers observed within both channels .

IV) Determination of observed protein interactions across replicates:

Rational: In order to asses/determine the interactions from a PCP-SEC-SILAC dataset the thresholds of how similar the observed protein correlation profile (as a whole) or the normalised deconvoluted fitted Gaussians of protein need to be to be considered to interact must be determined. To determine these thresholds the known interactions combinations within the identified proteins are assessed. These known interactions are derived from the CORUM database and are considered true positives (TP), to determine the true negative (TN) all combinations of proteins within CORUM but not known to interact are considered. The similarity (or different, Δ) of protein pairs are assessed using Euclidean distance. The values to use for protein pairs to be considered interacting are adjusted to a defined user precision (with precision being defined as TP/(TP+FP) of all observation within a given replicate), in our scripts three levels of precision are calculated, 70%, 60% and 50%. A precision between 70% and 60% is consistent with the quality of interaction data generated from affinity purification.

Before beginning download

Script to be used: ROC.m (Current build: version 2, 2014)

Additional script: myCenter.m (required for PCA analysis)

myPCA.m (required for PCA analysis)

myWhiten.m (required for PCA analysis)

d2p.m (required for t-SNE analysis)

tsne.m (required for t-SNE analysis)

Major protein MaxQuant output: Major_protein_groups.xlsx (See Figure 4), formatted in the correct format for analysis with ROC script.

Corum interactions formatted for analysis: Corum_correctly_formated_Uniprot_IDs.csv (See Figure 10), formatted in the correct format for analysis with ROC script.

Verbose MaxQuant output: Formatted into individual experiments protein profile experiments, and saved as an .xlsx file. The verbose output can have been processed with the alignment script or can be unprocessed. In cases where a single experiment is being visualized the unprocessed Verbose output and ensure the Alignment binary is set to zero (see Figure 16).

Gaussian property list: corresponding to the height, center, width, SSE, adjrsquare and observed complex size for all fitted Gaussians curves for the isotopologue ratio used to generate protein profiles. Gaussian property list can have been processed with the alignment script or can be unprocessed. In cases where a single experiment is being visualized the unprocessed Verbose output and ensure the Alignment binary is set to zero (see Figure 16).

Protocol:

-Open ROC.m and entry the following information (see Figure 18 and 19):

1) Major protein group information. This correspond to the verbose major protein group information which is used to assess if any of the major protein group members for each protein group are known to interact (TP) or predicted to not interact (TN). By assessing all members of the protein group the goal is to ensure a more accurate assessment of both TP and TN numbers.

2) CORUM interactions formatted into binary protein pairs. Contained the know CORUM interaction formatted into a binary list which are used to assess the interactions.

3) Location and name of the Gaussian property list

4) Location and name of the verbose MaxQuant outputs

-The script is designed to run without further input from the user and be run on single core computer. (Estimated processing time <24hours to generate images of all protein profiles)

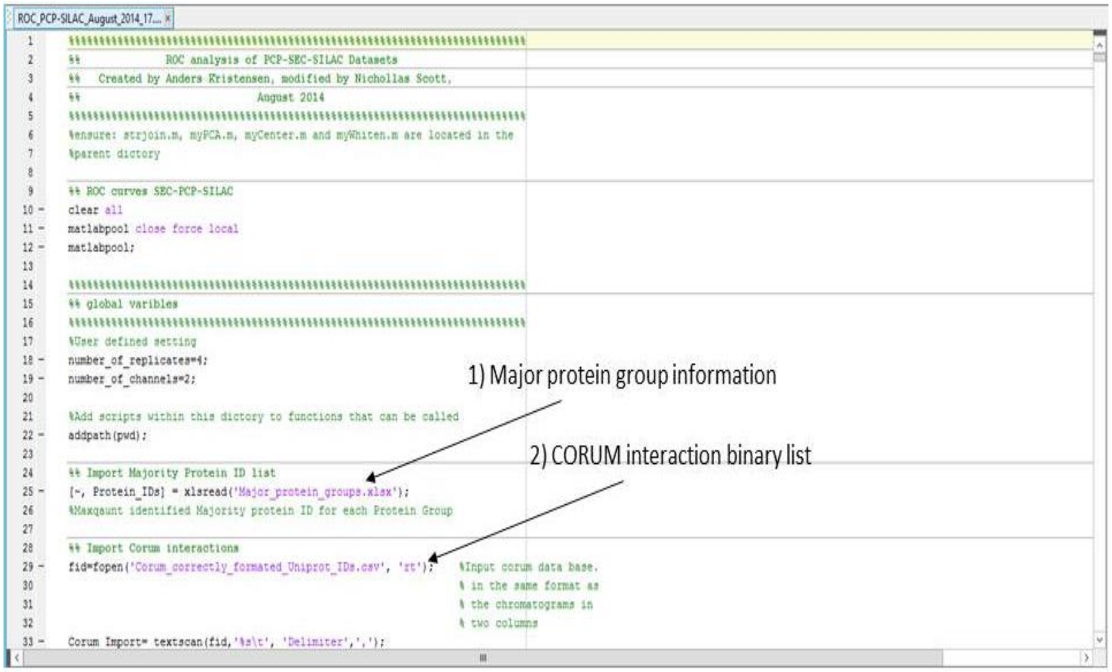


Figure 18: Setting up ROC.m script for processing, part A. The User required input are shown and are entered prior to runs the script.

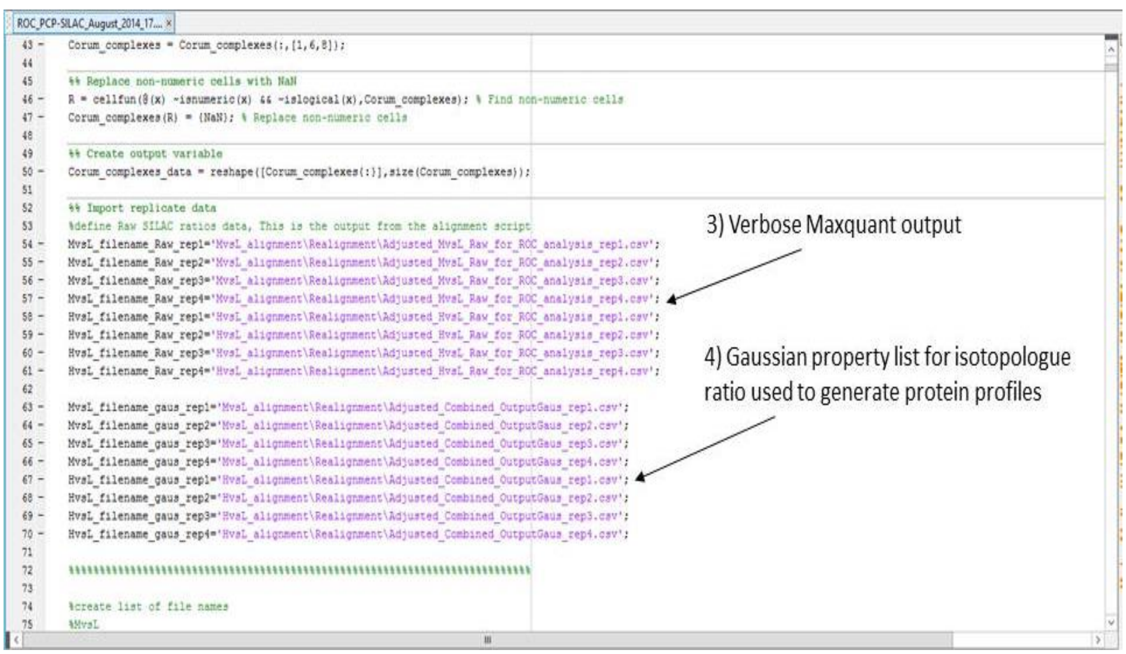


Figure 19: Setting up ROC.m script for processing, part A. The User required input are shown and are entered prior to runs the script.

Output from ROC.m script: (All files will be and deposited in individual folders corresponding to the individual replicate)

Processing files: Contain data generated by the script which is used to generate downstream data files. These files can be used to speed up the re-analysis of PCP-SILAC dataset but can largely be ignored and are generated for each replicate:

- -Negative interactions_70pc folder
- -Negative interactions_60pc folder
- -Negative interactions_50pc folder
- -Processing data folder

-Determined_interactions_50pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the determined interactions of all the proteins within the replicate at a precision of 50%.

-Determined_interactions_60pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the determined interactions of all the proteins within the replicate at a precision of 60%.

-Determined_interactions_70pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the determined interactions of all the proteins within the replicate at a precision of 70%.

-Interactions_determined_based_on_Euclidean_distance_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined solely using Euclidean distance of the whole profile. These interactions are determined with a precision of 80% in accordance with the protocol outlined by Kristensen et al 2012. For each protein pair the protein UNIPROT accession number and center for both proteins are given as well as the number of fitted Gaussian curve observed within each protein. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Interactions_list_50pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 50% based on the final observed interactions. For each protein pair the protein UNIPROT accession numbers, Center, delta Center, Delta Height, Delta Width and Delta Euclidean distance are given. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Interactions_list_60pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 60% based on the final observed interactions. For each protein pair the protein UNIPROT accession numbers, Center, delta Center, Delta Height, Delta Width and Delta Euclidean distance are given. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Interactions_list_70pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 70% based on the final observed interactions. For each protein pair the protein UNIPROT accession numbers, Center, delta Center, Delta Height, Delta Width and Delta Euclidean distance are given. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Neg_proteins_interactions_50pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all negative interactions (TN) not observed using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 50% based on the final observed interactions. For each protein pair the protein UNIPROT accession numbers, Center, delta Center, Delta Height, Delta Width and Delta Euclidean distance are given.

-Neg_proteins_interactions_60pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all negative interactions (TN) not observed using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 60% based on the final observed interactions. For each protein pair the protein UNIPROT accession numbers, Center, delta Center, Delta Height, Delta Width and Delta Euclidean distance are given.

-Neg_proteins_interactions_70pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all negative interactions (TN) not observed using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 70% based on the final observed interactions. For each protein pair the protein UNIPROT accession numbers, Center, delta Center, Delta Height, Delta Width and Delta Euclidean distance are given.

-Optimisation_matrix_interaction_numbers_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contain the numerical summary of the number of interactions observed as the Euclidean distance of the whole profile (x-axis) and that of individual normalized (to a height of one) fitted Gaussian curves (y-axis) as the threshold of each variable is altered from zero to the Euclidian distance with a false positive rate 0.5%.

-Optimisation_matrix_interaction_numbers_with_required_precision_50pc_replicate2.csv (Within ROC analysis data folder of each replicate generated during analysis): Contain the numerical summary of the number of interactions observed as the Euclidean distance of the whole profile (x-axis) and that of individual normalized (to a height of one) fitted Gaussian curves (y-axis) as the threshold of each variable is altered from zero to the Euclidian distance with a false positive rate 0.5%. Only combinations of thresholds which lead to Interaction with a precision over 50% are shown.

-Optimisation_matrix_interaction_numbers_with_required_precision_60pc_replicate2.csv (Within ROC analysis data folder of each replicate generated during analysis): Contain the numerical summary of the number of interactions observed as the Euclidean distance of the whole profile (x-axis) and that of individual normalized (to a height of one) fitted Gaussian curves (y-axis) as the threshold of each variable is altered from zero to the Euclidian distance with a false positive rate 0.5%. Only combinations of thresholds which lead to Interaction with a precision over 60% are shown.

-Optimisation_matrix_interaction_numbers_with_required_precision_70pc_replicate2.csv (Within ROC analysis data folder of each replicate generated during analysis): Contain the numerical summary of the number of interactions observed as the Euclidean distance of the whole profile (x-axis) and that of individual normalized (to a height of one) fitted Gaussian curves (y-axis) as the threshold of each variable is altered from zero to the Euclidian distance with a false positive rate 0.5%. Only combinations of thresholds which lead to Interaction with a precision over 70% are shown.

-Optimisation_matrix_precision_numbers_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contain the numerical summary of the observed precision of interactions observed as the Euclidean distance of the whole profile (x-axis) and that of individual normalized (to a height of one) fitted Gaussian curves (y-axis) as the threshold of each variable is altered from zero to the Euclidian distance with a false positive rate 0.5%.

-Output_Interactions_50pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 50% based on the final observed interactions. For each protein pair a concatenated name composed of the protein UNIPROT accession numbers and Center is provided in addition to the delta Center, Delta Height, Delta Width and Delta Euclidean distance. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Output_Interactions_60pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 60% based on the final observed interactions. For each protein pair a concatenated name composed of the protein UNIPROT accession numbers and Center is provided in addition to the delta Center, Delta Height, Delta Width and Delta Euclidean distance. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Output_Interactions_70pc_precision_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 70% based on the final observed interactions. For each protein pair a concatenated name composed of the protein UNIPROT accession numbers and Center is provided in addition to the delta Center, Delta Height, Delta Width and Delta Euclidean distance. Information on if both proteins are known within CORUM and if an interaction has been previously identified in CORUM is also provided.

-Sorted_Output_Interactions_Gaussians_50pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 50% based on the final observed interactions. Interactions are group according to protein with concatenated names composed of the protein UNIPROT accession numbers and Center provided for all proteins. The total number of interactions observed for each protein is provided as well as the number of these interactions which are known within CORUM.

-Sorted_Output_Interactions_Gaussians_60pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 60% based on the final observed interactions. Interactions are group according to protein with concatenated names composed of the protein UNIPROT accession numbers and Center provided for all proteins. The total number of interactions observed for each protein is provided as well as the number of these interactions which are known within CORUM.

-Sorted_Output_Interactions_Gaussians_70pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 70% based on the final observed interactions. Interactions are group according to protein with concatenated names composed of the protein UNIPROT accession numbers and Center provided for all

proteins. The total number of interactions observed for each protein is provided as well as the number of these interactions which are known within CORUM.

-Sorted_Output_Interactions_Protein_50pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 50% based on the final observed interactions. Interactions are group according to protein using the protein UNIPROT accession numbers. The total number of interactions observed for each protein is provided as well as the number of these interactions which are known within CORUM.

-Sorted_Output_Interactions_Protein_60pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 60% based on the final observed interactions. Interactions are group according to protein using the protein UNIPROT accession numbers. The total number of interactions observed for each protein is provided as well as the number of these interactions which are known within CORUM.

-Sorted_Output_Interactions_Protein_70pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains all interactions determined using both Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves. These interactions are determined with a precision of 70% based on the final observed interactions. Interactions are group according to protein using the protein UNIPROT accession numbers. The total number of interactions observed for each protein is provided as well as the number of these interactions which are known within CORUM.

-Summary_Results_50pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains a numerical summary of the final Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), True Positive Rate (defined as the TP/(FP+FN)), False Positive Rate (defined as the TP/(FP+TN)) and number of interactions with a precision of 50% based on the final observed.

-Summary_Results_60pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains a numerical summary of the final Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), True Positive Rate (defined as the TP/(FP+FN)), False Positive Rate (defined as the TP/(FP+TN)) and number of interactions with a precision of 60% based on the final observed.

-Summary_Results_70pc_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains a numerical summary of the final Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), True Positive Rate (defined as the TP/(FP+FN)), False Positive Rate (defined as the TP/(FP+TN)) and number of interactions with a precision of 70% based on the final observed.

-Summed_Precision_Summary_replicate#.csv (Within ROC analysis data folder of each replicate generated during analysis): Contains the numerical summary of the observed Summed Recall (defined as the TP/(TP+FN)), Summed Precision (defined as the TP/(TP+FP)), Summed True Positive Rate (defined as the TP/(FP+FN)), Summed False Positive Rate (defined as the TP/(FP+TN)) and number of interactions for all test precision combinations used to optimise the Euclidean distance of the whole profile and that of individual normalized (to a height of one) fitted Gaussian curves thresholds.

-Corum_vs_dataset_statistic_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Contains the numerical summary of the observed number of unique protein entries within CORUM, Number of unique Major Protein IDs with isoform information in dataset, Number of unique Major Protein IDs without isoform information in dataset, Unique CORUM entries within dataset, Redundant CORUM interactions within dataset, Unique CORUM interactions within dataset (non-redundant list of all interactions in CORUM possible with protein Identifications), Number of possible interactions, Number of possible interactions of proteins in CORUM, Maximum number of TP (redundant list of all interactions in CORUM possible with protein Identifications), Maximum number of TN, Number of self-interactions, Number of protein groups with members that have interactions in CORUM and Number of protein groups with members that have known interactions in CORUM.

NOTE 1: the ratio of Maximum number of TP to Maximum number of TN should ideally be high to ensure accurate assessment of the, it has been noted that our dataset the ratio of max TN to max TP is >20 which appears to be appropriate.
NOTE 2: the ratio of Number of protein groups with members that have interactions in CORUM to Maximum number of TP should ideally be low to ensure accurate assessment of the, it has been noted that our dataset the ratio of is ~1% which appears to be appropriate.

- FP_matrix_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the FP interactions of all the proteins within the replicate.

- INT_matrix_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the possible CORUM interactions (FP and TP) of all the proteins within the replicate.

-Multiple_known_INT_interaction_in_proteingroup_Matrix_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the proteins identified to contain multiple possible CORUM interactions (FP and TP) within the replicate.

-TP_matrix_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true)

summarising the proteins identified to contain true positive CORUM interactions within the replicate.

-Multiple_known_TP_interaction_in_proteingroup_Matrix_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the proteins identified to contain multiple possible true positive CORUM interactions within the replicate.

-Proteins_interactions_in_corum replicate_#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Contains the list of all possible true positive CORUM protein interactions possible within the dataset. For each interaction the Protein interaction, the individual proteins and the row number (corresponding to the protein group which the individual protein is identified from) are provided.

-Self_interaction_matrix_replicate#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Provides the logic matrix (corresponding to 0 for false and 1 for true) summarising the interactions which correspond to self-interactions within the replicate.

-Unique_proteins_interactions_in_corum_#.csv (Within ROC analysis matrixes folder of each replicate generated during analysis): Contains the list of all possible unique true positive CORUM protein interactions possible within the dataset. For each interaction the Protein interaction, the individual proteins and the row number (corresponding to the protein group which the individual protein is identified from) are provided.

-Center_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the Center on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-Center_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the Center on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-Euc_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the Euclidean distance (of the whole profile) on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the

number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-Euc_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the Euclidean distance (of the whole profile) on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-Height_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the Height on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-Height_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the Height on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-Width_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the Width on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-Width_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the Width on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-Gaussian_fits_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider

thresholds of the Euclidean distance (of the individual normalized fitted Gaussians) on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-Gaussian_fits_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the Euclidean distance (of the individual normalized fitted Gaussians) on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-PCA_Chromatograms_analysis_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the PCA analysis of the Euclidean distance (of the whole profile) on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-PCA_Chromatograms_analysis_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the PCA analysis of the Euclidean distance (of the whole profile) on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-PCA_Normalized_Gaussian_analysis_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the PCA analysis of the Euclidean distance (of the individual normalized fitted Gaussians) on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-PCA_ Normalized_Gaussian_analysis_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with

increasingly wider thresholds of the PCA analysis of the Euclidean distance (of the individual normalized fitted Gaussians) on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-TSNE_Normalized_Gaussian_analysis_replicate#.csv (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the analysis of the effect of with increasingly wider thresholds of the t-SNE analysis of the Euclidean distance (of the individual normalized fitted Gaussians) on the quality of the interactions assigned. The Distance (the delta difference under which two proteins are considered interacting), the number of true positive, the number of false positive, the number of false negative, the number of true negative, Interactions within defined distance, Interactions within defined distance also in CORUM, Recall (defined as the TP/(TP+FN)), Precision (defined as the TP/(TP+FP)), False Positive Rate (defined as the TP/(FP+TN)) and True Positive Rate (defined as the TP/(FP+FN)).

-TSNE_Normalized_Gaussian_analysis_replicate#.pdf (Within ROC analysis related Figures folder of each replicate generated during analysis): Contains the visualization of the effect of with increasingly wider thresholds of the t-SNE analysis of the Euclidean distance (of the individual normalized fitted Gaussians) on the quality of the interactions assigned. The figure on the left shows the effect on precision and recall with increasing thresholds. The figure of the right shows the effect on true positive rate and false positive rate with increasing thresholds.

-Final_Interactions_list_50_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions identified across all replicates including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 50% at the individual isotopologue channel level.

-Final_Interactions_list_60_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions identified across all replicates including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 60% at the individual isotopologue channel level.

-Final_Interactions_list_70_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions identified across all replicates including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if

an interaction between proteins is known within CORUM. These interactions are based on a precision of 70% at the individual isotopologue channel level.

-Final_Treatment_specific_interactions_list_50_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions only identified within the treatment experiments (As defined within the script) including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 50% at the individual isotopologue channel level.

-Final_Treatment_specific_interactions_list_60_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions only identified within the treatment experiments (As defined within the script) including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 60% at the individual isotopologue channel level.

-Final_Treatment_specific_interactions_list_70_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions only identified within the treatment experiments (As defined within the script) including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 70% at the individual isotopologue channel level.

-Final_Untreatment_specific_interactions_list_50_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions only identified within the untreated experiments (As defined within the script) including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 50% at the individual isotopologue channel level.

-Final_Untreatment_specific_interactions_list_60_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions only identified within the untreated experiments (As defined within the script) including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center,

Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 60% at the individual isotopologue channel level.

-Final_Untreatment_specific_interactions_list_70_precision.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions only identified within the untreated experiments (As defined within the script) including the interaction observed, the individual protein members, the number of times the interaction was observed, the observed center of each protein observed in each replicate, the Delta Height, Delta Center, Delta width, Delta Euclidean distance, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 70% at the individual isotopologue channel level.

-Global_Precision_across_replicates_50pc.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the number times an interaction was observed across experiments and the precision of the observed interactions. For the number of observations the total number of observed interactions, the number of interactions between proteins both in CORUM (TP +FP), the number of interactions known within CORUM (TP) and the precision of the observed interactions (TP/(TP+FP)) are provided. These interactions are based on a precision of 50% at the individual isotopologue channel level.

-Global_Precision_across_replicates_60pc.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the number times an interaction was observed across experiments and the precision of the observed interactions. For the number of observations the total number of observed interactions, the number of interactions between proteins both in CORUM (TP +FP), the number of interactions known within CORUM (TP) and the precision of the observed interactions (TP/(TP+FP)) are provided. These interactions are based on a precision of 60% at the individual isotopologue channel level.

-Global_Precision_across_replicates_70pc.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the number times an interaction was observed across experiments and the precision of the observed interactions. For the number of observations the total number of observed interactions, the number of interactions between proteins both in CORUM (TP +FP), the number of interactions known within CORUM (TP) and the precision of the observed interactions (TP/(TP+FP)) are provided. These interactions are based on a precision of 70% at the individual isotopologue channel level.

-Interactions_across_replicate_50pc.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions identified across all replicates including the interaction observed, the observed center of each protein observed in each replicate, the number of times the interaction was observed, the total number of replicates interaction observed in, which replicate the interaction was observed in, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 50% at the individual isotopologue channel level.

-Interactions_across_replicate_60pc.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions identified across all replicates including the interaction observed, the observed center of each protein observed in each replicate, the number of times the interaction was observed, the total number of replicates interaction observed in, which replicate the interaction was observed in, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 60% at the individual isotopologue channel level.

-Interactions_across_replicate_70pc.csv (Within Combined results folder generated during analysis): Contains a summary of all protein interactions identified across all replicates including the interaction observed, the observed center of each protein observed in each replicate, the number of times the interaction was observed, the total number of replicates interaction observed in, which replicate the interaction was observed in, if both proteins are within CORUM and if an interaction between proteins is known within CORUM. These interactions are based on a precision of 70% at the individual isotopologue channel level.

-Summary_Interactions_determined_based_on_Euclidean_distance.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the number of interactions determined based on Euclidean distance alone with each isotopologue channel.

-Summary_Results_50pc_replicate.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the Recall (non-redundant interactions), Precision (non-redundant interactions), TPR (non-redundant interactions), FPR (non-redundant interactions), Precision (redundant interactions), Total number of interactions (redundant across replicate), Unique number of interactions (Non-redundant within replicate), Total Unique interactions (Non-redundant across replicate) and Number non-redundant interactions. These parameters are based on a precision of 50% at the individual isotopologue channel level.

-Summary_Results_60pc_replicate.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the Recall (non-redundant interactions), Precision (non-redundant interactions), TPR (non-redundant interactions), FPR (non-redundant interactions), Precision (redundant interactions), Total number of interactions (redundant across replicate), Unique number of interactions (Non-redundant within replicate), Total Unique interactions (Non-redundant across replicate) and Number non-redundant interactions. These parameters are based on a precision of 60% at the individual isotopologue channel level.

-Summary_Results_70pc_replicate.csv (Within Combined results folder generated during analysis): Contains a numerical summary of the Recall (non-redundant interactions), Precision (non-redundant interactions), TPR (non-redundant interactions), FPR (non-redundant interactions), Precision (redundant interactions), Total number of interactions (redundant across replicate), Unique number of interactions (Non-redundant within replicate), Total Unique interactions (Non-redundant across replicate) and Number non-redundant interactions. These parameters are based on a precision of 70% at the individual isotopologue channel level.

-Observed_interactions_across_replicates_50_precision.png (Within Combined results folder generated during analysis): Contains the visualization of the TP, FP and total number of interactions observed across isotopologue channels. These parameters are based on a precision of 50% at the individual isotopologue channel level.

-Observed_interactions_across_replicates_60_precision.png (Within Combined results folder generated during analysis): Contains the visualization of the TP, FP and total number of interactions observed across isotopologue channels. These parameters are based on a precision of 60% at the individual isotopologue channel level.

-Observed_interactions_across_replicates_70_precision.png (Within Combined results folder generated during analysis): Contains the visualization of the TP, FP and total number of interactions observed across isotopologue channels. These parameters are based on a precision of 70% at the individual isotopologue channel level.