



UNIVERSITY OF  
LIVERPOOL

*Optimising the statistical pipeline for  
quantitative proteomics*

Thesis submitted in accordance with the requirements of the  
University of Liverpool for  
the degree of Doctor in Philosophy by Hayley Price

**Oct 2023**

## ***Dedication***

To Paul, Lola, and Jake. You are everything.

## ***Acknowledgements***

I would like to thank the following people for the generous help and support given over the long years it has taken me to complete this thesis:

Colleagues at the University who have provided friendship, laughter and advice.

My internal examiners Professor Claire Eyers and Professor Francesco Falciani, thank you for your time, your insight, and your expertise.

My supervisors, especially Professor Andy Jones, whose patience is has been endless. Your help and support has been vital to me finally completing this work.

My friends who have been there when I have been down.

My family; my children's grandparents for the hours of childcare, my children for lighting up my days, and my husband who has always been there for me.

## ***Table of contents***

Optimising the statistical pipeline for quantitative proteomics .....	1
Dedication .....	ii
Acknowledgements .....	iii
Table of contents .....	iv
List of tables .....	xii
List of figures .....	xvi
Abstract.....	27
Chapter 1. Introduction.....	28
1.1. Quantitative proteomics .....	29
1.2. Experimental workflow .....	31
i. Sample preparation.....	32
ii. Ionisation.....	34
iii. Mass analysis and fragmentation .....	36
Mass analysers .....	36
Fragmentation.....	38
Acquisition modes .....	39
iv. The mass spectrum.....	40
Isotopes.....	41
v. Labelling .....	42
vi. Top-down proteomics.....	44
1.3. LCMS data processing pipeline .....	45
i. Proteomics workflow .....	45
ii. Proteomics software .....	45
1.4. Progenesis QI for Proteomics .....	47
i. Alignment.....	47

ii.	Feature detection and ion abundance quantification .....	48
	Missing values .....	49
1.5.	Peptide identification and protein inference.....	50
i.	Peptide identification .....	50
ii.	Protein inference.....	51
iii.	Peptide abundance summarisation methods.....	54
1.6.	Normalisation .....	55
i.	Variance normalisation.....	58
	Log transformation.....	58
ii.	Bias normalisation.....	58
	Central tendency normalisation.....	58
	Progenesis QIP normalisation.....	59
	Linear regression normalisation.....	60
	Local regression normalisation .....	62
	Quantile normalisation .....	63
	QPROT normalisation.....	65
iii.	Combined bias and variance normalisation .....	66
	Variance stabilisation normalisation (VSN) .....	66
1.7.	Differential expression analysis.....	67
i.	Hypothesis testing .....	67
ii.	Statistical significance .....	68
iii.	<i>t</i> -Test.....	68
iv.	Multiple testing.....	70
v.	The false discovery rate.....	71
vi.	Problems with null-hypothesis significance testing for proteomics data	73
vii.	Linear modelling for differential expression.....	75

viii.	Bayesian inference for differential expression.....	77
1.8.	Statistical software.....	80
i.	MSstats package.....	81
	Data processing and visualisation.....	81
	Linear modelling for summarisation.....	82
	Tukey’s median polish for summarisation.....	83
	Statistical modelling and inference.....	87
	Experimental design.....	88
ii.	QPROT package.....	88
	Metropolis Hastings algorithm.....	90
	FDR calculation.....	91
	Low abundance proteins.....	95
1.9.	Pathway analysis.....	96
1.10.	Aims of project.....	98
Chapter 2.	Differential expression analysis evaluation using ground-truth data	99
	99	
2.1.	Introduction.....	99
i.	Abstract.....	99
ii.	Benchmarking proteomics data analysis methods.....	100
	Ground truth data.....	100
	Benchmarking software.....	101
	Protein identification, inference, and grouping.....	101
iii.	Aims of chapter.....	106
2.2.	Methods.....	107
i.	Spike-in datasets.....	107
ii.	Benchmarking workflow.....	109
	Quantitative processing.....	110

Protein inference software.....	111
iii. Differential expression analysis.....	114
QPROT.....	114
MSstats.....	114
t-Test.....	115
iv. Analysis of performance.....	115
Performance of protein inference parameters.....	116
v. Imputation.....	116
vi. Assessment of benchmarking data.....	116
2.3. Results and discussion.....	118
i. Summary of DE analysis.....	118
ii. Performance of differential expression analysis.....	119
PXD001385.....	119
PXD001819.....	121
PXD002099.....	123
Summary of results.....	124
iii. Performance of protein inference parameters.....	125
Peptide ion selection.....	125
Quantification method.....	127
Threshold selection comparison.....	129
PXD001819.....	130
PXD001835.....	132
PXD002099.....	133
iv. Imputation.....	136
v. Assessment of benchmarking data.....	137
2.4. Conclusions.....	145

## Chapter 3. Differential expression analysis evaluation by pathway enrichment

146

3.1.	Introduction.....	146
i.	Abstract.....	146
ii.	Technical and statistical issues with using ground-truth data .....	147
iii.	Review of benchmarking studies.....	149
iv.	Normalisation.....	152
v.	Aims of chapter .....	154
3.2.	Methods.....	154
i.	Biological datasets .....	154
ii.	Imputation .....	157
iii.	Differential expression evaluation .....	158
Quantitative processing.....	158	
Differential expression analysis .....	159	
Defining significant results.....	159	
Analysis of performance.....	160	
iv.	Evaluation of pathway analysis benchmarking method.....	161
v.	Normalisation across samples .....	161
3.3.	Results and discussion .....	164
i.	Pathway analysis benchmarking validation .....	164
ii.	Imputation .....	165
iii.	Differential expression evaluation .....	166
Summary of data analysis .....	166	
iv.	Evaluation of DE methods by pathway analysis .....	167
Overall analysis.....	167	
PXD004501.....	168	
PXD004682.....	170	



PXD007592.....	171
v.    Investigating the effect of normalisation method .....	172
MSstats.....	172
Fold enrichment threshold.....	172
Overall analysis.....	174
PXD004501.....	176
PXD004682.....	183
PXD007592.....	192
Overall performance of normalisation methods.....	204
vi.    Review of benchmarking.....	205
Significance threshold selection.....	207
Pathway analysis.....	208
3.4.    Conclusions.....	209
Chapter 4.    Optimised proteomics pipeline.....	211
4.1.    Introduction.....	211
i.    Abstract.....	211
ii.    Proteomics pipeline development.....	212
Pipeline Workflow .....	212
High-performance computing (HPC) .....	213
Normalisation.....	215
Differential expression analysis .....	216
Sampling method .....	216
Defining significant results.....	218
Pathway analysis.....	218
Evaluation of Results .....	220
iii.    Aims of chapter .....	220
4.2.    Methods.....	221

i.	Validation of methods .....	221
	Datasets.....	221
	Implementation Bayes statistics .....	221
	Implementation of FDR .....	221
	Evaluation of clusterProfiler() for DE evaluation.....	221
ii.	Comparison to Progenesis output .....	222
iii.	Comparison to QPROT output.....	222
4.3.	Results and discussion .....	223
i.	Validation of methods .....	223
	Implementation Bayesian differential expression analysis .....	223
	Implementation of FDR calculation .....	225
	Pathway analysis.....	226
	Simplify function .....	228
i.	Comparison to Progenesis output .....	231
	PXD004501.....	231
	PXD004682.....	238
	PXD007592.....	243
	Overall analysis.....	253
ii.	Comparison to QPROT output.....	255
	650 iterations .....	255
	2000 iterations.....	256
4.4.	Conclusions.....	258
Chapter 5.	Thesis conclusions and outlook.....	259
	Summary of thesis results .....	259
	Further work.....	261
	Limitations.....	263
	Conclusions.....	264

Supplementary material .....	264
References.....	265

## List of tables

<i>Table 1; Summary of different normalisation methods with details of R packages used, assumptions, issues and uses. ....</i>	<i>57</i>
<i>Table 2; summary of the linear model of the data from Figure 1.7.5 .....</i>	<i>76</i>
<i>Table 3; Worked example of protein quantitation by linear modelling. Log2 transformed peptide abundances for 3 example peptides mapped to a protein across 3 runs, fitted peptide abundances for 3 peptides following linear modelling and resulting protein summarisation values. ....</i>	<i>83</i>
<i>Table 4; Protein summarisation based on additive modelling using TMP. ....</i>	<i>84</i>
<i>Table 5; Summary of options for how ion variants with the same primary peptide sequence are treated for peptide roll-up. ....</i>	<i>103</i>
<i>Table 6; Summary of methods of which peptides mapped to the protein are used for protein quantification .....</i>	<i>106</i>
<i>Table 7; Summary of experimental conditions for each of the benchmarking datasets giving details of samples used, experimental design, parameters used for MS analysis, and number of proteins. Only the four largest spike-in conditions were used for benchmarking. Due to processing issues, the 10 fmol/<math>\mu</math>l spike-in condition was not used in dataset PXD002099. The number of spike-in proteins for PXD001385 was not provided in the paper and is based on this analysis. ....</i>	<i>109</i>
<i>Table 8; Summary of parameters producing protein abundance data analysed by QPROT and t-test. ....</i>	<i>112</i>
<i>Table 9; Summary of parameters producing protein abundance data analysed by MSstats. ....</i>	<i>113</i>
<i>Table 10; Summary of quantities of proteins from analysis. Minimum and maximum values are given for proteins identified by a minimum of one or two unique proteins over all possible parameter options. A FDP of less than 0.05 was used as the significance threshold for classifying proteins as differentially expressed. ....</i>	<i>118</i>
<i>Table 11; Comparison of effects of imputation. t-test analysis with a BH corrected p-value of 0.05 to indicate significance was used and number of DE proteins was compared for each dataset using different methods to deal with zero values. Ions with same primary peptide sequence and different charge state's intensities were summed. Ions with same primary peptide sequence but with artefactual modifications were treated separately. Protein quantification was based on the sum of the average intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptide were selected. Impute 1e-07, imputation of 0.0000001; Impute 1e-10, 0.000000001; Remove, removal of proteins with any zero values from the analysis; FP rate, false positive rate; FN rate, false negative rate. ....</i>	<i>136</i>

<b>Table 12; Summary of experimental conditions for each of the benchmarking datasets giving details of samples used, experimental design, parameters used for MS analysis and number of proteins .....</b>	<b>155</b>
<b>Table 13; Summary of significance threshold ranges. Proteins passing the threshold are categorised as DE and go forward for enrichment analysis. Size of increment increases over iteration. ....</b>	<b>159</b>
<b>Table 14; Normalisation procedures performed by Normalyzer package with details of R functions and packages employed.....</b>	<b>161</b>
<b>Table 15; Pathway analysis validation completed with dataset PXD004682. Different proportions of true DE proteins and randomly selected background proteins. Number of significant terms from DAVID enrichment analysis of DE proteins from t-test analysis, (Benjamini Hochberg adjusted p-value &lt; 0.05). ....</b>	<b>164</b>
<b>Table 16; Comparison of effects of imputation. t-test analysis with a BH corrected p-value of 0.05 to indicate significance was used and number of DE proteins was compared for each dataset using different methods to deal with zero values. Ions with same primary peptide sequence and different charge state's intensities were summed. Ions with same primary peptide sequence but with artefactual modifications were treated separately. Protein quantification was based on the sum of the average intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptide were selected. Enrichment analysis was performed using the DAVID webpage (<a href="https://david.ncifcrf.gov/tools.jsp">https://david.ncifcrf.gov/tools.jsp</a>) using the default parameters. Impute 1e-07, imputation of 0.0000001; Impute 1e-10, 0.000000001; Remove, removal of proteins with any zero values from the analysis. ....</b>	<b>166</b>
<b>Table 17; Summary of DE analysis by QPROT, t-test, and MSstats of biological datasets PXD004501, PXD001682, and PXD007592. Protein group abundances were calculated using protein inference benchmarking software from Progenesis QIP normalised peptide ion abundances that had been identified by a minimum of one or two unique peptides.....</b>	<b>166</b>
<b>Table 18; Summary of 'best' analysis parameters where maximum number of terms are found during enrichment analysis of dataset PXD004501 using Progenesis normalised protein intensities. ....</b>	<b>169</b>
<b>Table 19; Summary of 'best' analysis parameters where maximum number of terms are found during enrichment analysis of dataset PXD004682 using Progenesis normalised protein intensities. ....</b>	<b>170</b>
<b>Table 20; Summary of 'best' analysis parameters where maximum number of terms are found during enrichment analysis of dataset PXD007592 using Progenesis normalised protein intensities. ....</b>	<b>171</b>
<b>Table 21; Enrichment analysis of dataset PXD004501. Optimal combination of analysis parameters are shown for each of the DE methods QPRT and t-test. ....</b>	<b>176</b>

<b>Table 22; Enrichment results for optimal QPROT analysis with quantile median normalisation and a significance threshold of 0.002 for dataset PXD004501. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section. ....</b>	<b>178</b>
<b>Table 23; Enrichment results for optimal t-test analysis with log2 transformed abundances and a significance threshold of 0.04 for dataset PXD004501. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section. ....</b>	<b>181</b>
<b>Table 24; Enrichment analysis of dataset PXD004682. Optimal combination of analysis parameters are shown for each of the DE methods QPRT and t-test. ....</b>	<b>183</b>
<b>Table 25; Enrichment results for optimal QPROT analysis with QPROT normalisation and a significance threshold of 0.0001 for dataset PXD004682. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section. ....</b>	<b>185</b>
<b>Table 26; Enrichment results for optimal t-test analysis with log2 transformed abundances and a significance threshold of 0.002 for dataset PXD004682. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section. ....</b>	<b>188</b>
<b>Table 27; Enrichment analysis of dataset PXD007592. Optimal combination of analysis parameters are shown for each of the DE methods QPRT and t-test. ....</b>	<b>192</b>
<b>Table 28; Enrichment results for optimal QPROT analysis with log2 transformed protein abundances and a significance threshold of 0.05 for dataset PXD007592. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section. ....</b>	<b>194</b>
<b>Table 29; Enrichment results for optimal t-test analysis with Loess normalisation and a significance threshold of 0.05 for dataset PXD0047592. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section. ....</b>	<b>200</b>
<b>Table 30; Rank of normalisation methods based on number significant terms from enrichment analysis. Overall ranking and ranking for DE methods QPROT and t-Test. ....</b>	<b>204</b>
<b>Table 31; Summary of significance threshold ranges. Proteins passing the threshold are categorised as DE and go forward for enrichment analysis. Size of increment increases over iteration. ....</b>	<b>218</b>
<b>Table 32; Pathway analysis validation completed with dataset PXD004682. Different proportions of true DE proteins and randomly selected background proteins. Number of significant terms from DAVID enrichment analysis of DE proteins from t-test analysis, (Benjamini Hochberg adjusted p-value &lt; 0.05). Process was repeated 100 times. ....</b>	<b>227</b>
<b>Table 33; Total number of enriched GO BP, CC, and MF terms from clusterProfiler analysis of upregulated proteins from the current Progenesis QIP DE analysis of dataset PXD004682 before and after simplify() function ....</b>	<b>228</b>
<b>Table 34; Simplified enriched GO MF terms (left column) and original enriched GO MF terms (right column) from clusterProfiler analysis of upregulated proteins from the current Progenesis QIP DE analysis of dataset PXD004682. ....</b>	<b>229</b>

<i>Table 35; Pathway analysis validation using the simplify() function completed with dataset PXD004682. Different proportions of true DE proteins and randomly selected background proteins. Number of significant terms from DAVID enrichment analysis of DE proteins from t-test analysis, (Benjamini Hochberg adjusted p-value &lt; 0.05).Process was repeated 100 times.</i>	229
<i>Table 36; Enrichment analysis results for dataset PXD004501 from the optimised pipeline and Progenesis output at standard significance threshold of 0.01, 0.05, and 0.1.</i>	232
<i>Table 37; Summary of the number of GO terms and KEGG pathways from enrichment analysis performed on upregulated proteins from the optimal analysis and from the original paper for dataset PXD004501.</i>	233
<i>Table 38; Summary of terms produced by enrichment analysis of proteins upregulated in cancer tissue as identified by the optimal analysis and from the original paper for dataset PXD004501 that are related to functions highlighted by Jin et al. (2018)</i>	235
<i>Table 39; Statistical analysis results of protein identified by Jin et al. (2018) as being a protein marker for differentiating malignant from benign ascites that were identified in Progenesis QIP processing</i>	237
<i>Table 40; Enrichment analysis results for the dataset PXD004682 from the optimised pipeline and Progenesis output at standard significance threshold of 0.01, 0.05, and 0.1.</i>	239
<i>Table 41; Number of enrichment terms from PANTHER analysis of proteins classed as being upregulated by optimal analysis and the current Progenesis output, and the number of terms shared with PANTHER analysis of biomarker candidates identified by Stewart et al. (2017)</i>	241
<i>Table 42; Number of GO-Slim functional classifications associated with selected lung cancer pathways identified by Stewart et al. (2017) for proteins upregulated in cancer tissue identified by the pipeline's optimal analysis and by the current Progenesis QIP output</i>	242
<i>Table 43; Enrichment analysis results for the dataset PXD007592 from the optimised pipeline and Progenesis output at standard significance threshold of 0.01, 0.05, and 0.1.</i>	244
<i>Table 44; Summary of proteins upregulated in good responders to MAPKi therapy that are related to the functions highlighted by Zila et al. (2018) as classified by the optimal analysis.</i>	246
<i>Table 45; Summary of proteins highlighted by Zila et al. (2018) that were upregulated in good responders to MAPKi therapy missed by the optimal analysis</i>	247

## List of figures

<b>Figure 1.1.1; Quantitative proteomics experiment comparing three samples of healthy and diseased tissue. Proteins are extracted from the sample and digested to produce ionisable peptides. The detected mass spectrometric signal is used to calculate relative changes in protein abundance between the two samples, identifying those that change in the diseased state. Ion intensity map (NonlinearDynamics). .....</b>	<b>30</b>
<b>Figure 1.2.1; Experimental liquid chromatography tandem mass spectrometry (LC-MS/MS) workflow. Proteins are extracted from the sample, digested into peptides and separated using chromatography in the sample preparation stage. Peptides are ionised and subjected to mass analysis. Under data dependent acquisition (DDA) scheme, the most intense ions are fragmented and mass analysis of these fragments is produced. ....</b>	<b>31</b>
<b>Figure 1.2.2; Electrospray ionisation. The liquid peptide sample becomes charged as it passes through the capillary and forms sample ions as the solvent evaporates from the droplets. (Kicman et al., 2007) .....</b>	<b>35</b>
<b>Figure 1.2.3; Matrix-assisted laser desorption ionisation. The sample and the matrix dry to form a crystalline structure that is targeted by a laser. The energy absorbed causes desorption and desolvation, allowing a charged sample ion to enter the mass analyser. (Kicman et al., 2007) .....</b>	<b>35</b>
<b>Figure 1.2.4; The Orbitrap mass analyser. Ions (depicted by green and red lines) oscillate both round and along the axis of a central spindle-like electrode. 1.) The detected oscillation is recorded as a transient signal, the axial ion oscillation frequency, which is unique for each m/z. 2.) Using Fourier transformation, the frequency is converted into a mass spectrum. (Savaryn et al., 2016) .....</b>	<b>37</b>
<b>Figure 1.2.5; Example mass spectrum (Quinn et al., 2012) .....</b>	<b>40</b>
<b>Figure 1.2.6; Ion intensity map from software analysis package Progenesis Q1 for Proteomics (<a href="http://www.nonlinear.com/progenesis/q1-for-proteomics/">http://www.nonlinear.com/progenesis/q1-for-proteomics/</a>) .....</b>	<b>40</b>
<b>Figure 1.2.7; Total ion chromatogram from software analysis package Progenesis Q1 for Proteomics (<a href="http://www.nonlinear.com/progenesis/q1-for-proteomics/">http://www.nonlinear.com/progenesis/q1-for-proteomics/</a>) .....</b>	<b>41</b>
<b>Figure 1.2.8; Mass spectrum of a peptide NVLPQRSTVW. The monoisotopic peak is the most abundant peak, with four additional peaks at higher mass/charge values, separated by 1 Da, due to the presence of naturally occurring heavy isotopes of the single-charged peptide (Sykes and Williamson, 2008) .....</b>	<b>42</b>
<b>Figure 1.4.1; Progenesis QIP bioinformatics workflow for processing label-free quantitative proteomics experiments; RAW data is imported and compressed into simplified modelled ion intensities. Normalisation and alignment allow the information from all experimental runs to be combined for analysis. Fragmented peptide ion information is used for database searches</b>	



for peptide and protein identification. Protein grouping and abundance normalisation provide protein group intensity information for comparative analysis. ....	47
<b>Figure 1.4.2; Alignment of a peptide ion in two overlaid runs (shown in pink and green) represented as m/z versus retention time. a.) Before alignment, due to retention time differences, the ions do not overlap. b.) After alignment, the ions are in the same location and combine to show a single feature (<a href="http://www.nonlinear.com/progenesis/qi-for-proteomics/v2.0/faq/why-is-alignment-so-important.aspx">http://www.nonlinear.com/progenesis/qi-for-proteomics/v2.0/faq/why-is-alignment-so-important.aspx</a>).....</b>	<b>48</b>
<b>Figure 1.4.3; a.) Two-dimensional (taken from analysis software package Progenesis Q1) and b.) Three-dimensional ion intensity maps. Isotopes of the same peptide ion shown as black shaded areas (a) or peaks (b) within the red boundary line. Peptide ion abundance is calculated by summing the areas below the scan lines of each of the isotopes within the boundary (<a href="http://www.nonlinear.com/progenesis/qi-for-proteomics/how-it-works/">http://www.nonlinear.com/progenesis/qi-for-proteomics/how-it-works/</a>) .....</b>	<b>49</b>
<b>Figure 1.5.1; Protein grouping example 1. Peptide 1 is only mapped to protein A; therefore, it is labelled unique and protein A is distinct. The same is true for peptide 4 and protein C. Peptides 2 and 3 are both mapped to two proteins, (A and B) and (B and C) respectively. These peptides are labelled as conflicted peptides. As there is no independent evidence that protein B is present in the sample, due to parsimony principle this protein is labelled multiply subsumed and will be discarded.....</b>	<b>52</b>
<b>Figure 1.5.2; Protein grouping example 1 after rules of parsimony are applied; protein B is discarded and peptides 2 and 3 can now be called resolved. ....</b>	<b>53</b>
<b>Figure 1.5.3; Protein grouping example 2. Protein A is a distinct protein due to unique peptide 1. Proteins B and C both contain conflicted peptide 2 and resolved peptide 3. There is no independent evidence to support that one over the other is present in the sample, and so they form their own same-set protein group.....</b>	<b>53</b>
<b>Figure 1.5.4; Protein grouping example 3. a.) The three proteins contain a combination of the same resolved peptides and so will form a resolved group. As protein A, has been identified by the most peptides, it will head the group, and proteins B and C will form a sub-set. b.) After applying parsimony rules; the evidence in the sample can be explained by a single protein (protein A), under the rules of parsimony, proteins B and C are discarded. ....</b>	<b>54</b>
<b>Figure 1.6.1; MA plots two samples of a.) raw and b.) normalised peptide abundances. 'A' on x-axis represents average log expression values, 'M' on y-axis shows difference in log expression values between samples. Linear modelling (shown in blue) shows a line more centred around <math>y = 0</math> in the normalised data.....</b>	<b>56</b>
<b>Figure 1.6.2; Protein intensities from one run are plotted against the median protein intensity across all runs and a linear model is fitted using ordinary least squares to provide the normalised protein intensities for that run, show in grey line. ....</b>	<b>60</b>

**Figure 1.6.3; Residuals are the difference between the actual value, shown as a black dot, and the modelled value, shown as a red dot. The model that gives the lowest value when residuals are squared and summed is chosen as the best fit. ....61**

**Figure 1.6.4; Linear regression normalisation where one data point, shown in red, has been replaced with an outlier. Original linear model is shown in grey and linear model including outlier is shown in red. Changes between models is created by changing a single data point. ....61**

**Figure 1.6.5; Data from fig 6 modelled with robust linear regression (dark grey line). This brings the fit closer to the original model before the outlier was introduced (shown in grey) compared to linear modelling with ordinary least squares (red line). ....62**

**Figure 1.6.6; Protein intensity values below approximately 40000 deviate from being linearly correlated (red, dashed line). Blue line shows non-parametric regression using weighted least squares (LOWESS) which better fits data. ....63**

**Figure 1.6.7; Worked example of quantile normalisation. a.) Mean protein abundance data of three proteins (A, B, and C) across three conditions (1, 2, and 3). b.) The mean value of the most highly abundant protein in each group is calculated. c.) Most abundant proteins are all assigned the mean abundance value. d.) The next most abundant proteins in each condition's mean is calculated and e.) each are assigned that mean abundance values. f.) Mean of least abundant proteins is calculated to be assigned. g.) Quantile normalised protein abundances. Note that the values across groups are the same, but the original rank order of protein is preserved. ....65**

**Figure 1.6.8; Variance stabilisation normalisation. Raw abundances (shown in black) and normalised abundances (shown in red). ....66**

**Figure 1.7.1; Graphical representation of the p-value calculation. a.) The observed experimental value ( $x$ ) is compared to the reference value ( $\mu$ ) b.)  $\mu$  is the mean of a null distribution,  $H_0$ . c.) The p-value is the percentage of values from the null distribution that are more extreme than the observed value (shaded in black and grey) (Krzywinski and Altman, 2013b) ..... 68**

**Figure 1.7.2; Effect of number of degrees of freedom on t-distribution; as degrees of freedom increases, the t-distribution approaches the normal distribution (Palkovic et al., 2020) .....70**

**Figure 1.7.3; a.) A uniform distribution of 100 000 simulated p-values obtained by comparing two samples drawn from the same distribution. 5% of the values have a p-value less than 0.05 and are false positives. b.) A skewed distribution of 100 000 simulated p-values obtained by comparing two samples drawn from different distributions. 78% of them have a p-value less than 0.05 and are true positives (Colquhoun, 2014). ....72**

**Figure 1.7.4; Two hypothetic distributions of protein abundances from two conditions, healthy and diseased tissue. There are two subjects (biological replicates) per condition, with three technical replicates per subject. The distribution and mean of the population abundances are represented by black curves and black squares. Observed abundances, mean subject**

*abundances, and condition abundances are represented by black circles, solid dots, and stars, respectively. In plot A there is no difference in overall population abundances between conditions due to the large biological variation, while plot B shows a difference in population abundances between conditions and a small biological variation. (Chang et al., 2012).....74*

**Figure 1.7.5; Two-sample t-test using linear modelling. The mean abundance of group 1 is shown as a red line, and the mean abundance of group 2 is shown as a black line. The slope shown in grey is used to calculate the difference between the two means.....76**

**Figure 1.7.6; Four group one-way ANOVA using linear modelling. The intercept,  $\beta_0$  is Group1 mean and is shown in black. The means of the other groups ( $\beta_0 + \beta_1$  – Group 2 mean,  $\beta_0 + \beta_2$  – Group 3 mean,  $\beta_0 + \beta_3$  – Group 4 mean) are shown in red. The difference in means is calculated using the slopes  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  which are shown in grey.....77**

**Figure 1.7.7; Probability distribution of prior beliefs (shown in red) is scaled by observing the evidence (shown in green) to give an updated probability distribution showing posterior beliefs in blue. ....78**

**Figure 1.8.1; Worked example of protein summarisation through linear modelling. Observed log2 abundances for three peptides mapped to a protein across three runs shown as colour to the left, fitted peptide abundances shown in black to the right. Resulting protein abundance for each run from the modelled peptide intensities is shown as colour. ....83**

**Figure 1.8.2; The overall median for all values in the dataset is calculated as 22.22 and is assigned to the common effect cell in green. A residual table is created by calculating the difference between the original value and the median. Row (shown in orange) and column (shown in blue) are effect values and are initially set to zero. ....84**

**Figure 1.8.3; The row medians of the residual table are computed and are shown in red. ....84**

**Figure 1.8.4; A second residual table is created with row medians assigned to the row effects margin on the left (shown in blue), and the values are a subtraction of the row median from the value in the first residual table. ....84**

**Figure 1.8.5; The column medians of the second residual table are calculated and are shown in red. ....85**

**Figure1.8.6; A third residual table is created with column medians assigned to the column effects margin shown in orange, and values are a subtraction of the column median from the value in the second residual table. The column effect median is added to the row effect margin (shown in blue) and the overall common effect cell (shown in green), but as the column effect median is zero, the value remains unchanged. This is the end of the first iteration. ....85**

**Figure 1.8.7; A second iteration; the row effect is calculated by computing the row medians (red) from the residuals and adding them to the common effect cell (green) and the row effects margin (blue), before subtracting the row medians (red) from the residuals to yield the values .....85**

**Figure 1.8.8; The second iteration continued; the column effects are calculated by computing the column medians (red) from the residuals and adding them to the common effect cell (green) and the column effects margin (orange), before subtracting the column medians (red) from the residuals to yield the values. This ends the second iteration. The proportional reduction in the sum of absolute residuals is zero on the second iteration, and so the analysis is ended. The overall common effect value is added to the row effects values to give the fitted protein intensity level in each run. ....85**

**Figure 1.8.9; Comparison of the protein abundances from linear and additive modelling. ....86**

**Figure 1.8.10; Example plots from the MSstats dataProcessPlots() method are given for each protein. a.) Profile plot with summarisation - Log2 peptide intensities of the five peptides mapped to protein P08200|IDH are shown in grey, with the summarised log2 protein intensity shown in dark red across four conditions and twelve runs. Used to identify potential sources of variation. b.) QC plot – Log2 intensities for across all peptides used for quantification are shown for each run. Used to evaluate systemic bias between MS runs. c.) Condition plot – Log2 intensities for all peptides used for quantification are shown for each condition. Used to the illustrate mean and variability of each condition per protein. ....86**

**Figure 1.8.11; Illustration of Metropolis-Hastings algorithm (Lee et al., 2015). ....90**

**Figure 1.8.12; Histogram of Z-statistics from QPROT analysis of an example dataset .....93**

**Figure 1.8.13; Overall density of Z-statistics from QPROT analysis of an example dataset. ....94**

**Figure 1.8.14; Overall density of Z-statistics (black) and known distribution of Z-score from the null hypothesis (red) from QPROT analysis of an example dataset. ....94**

**Figure 1.8.15; Overall density of Z-statistics (black) and known distribution of Z-score from the null hypothesis (blue) with estimated distribution of Z-score from the alternative hypothesis (red) from QPROT analysis of an example dataset. ....95**

**Figure 2.1.1; Representation of a benchmarking dataset. The background proteome level, shown in green, remains constant in each sample. Different concentrations of spike-in proteins, shown in blue, allow simulated fold-change with pairwise comparison. .... 100**

**Figure 2.1.2; Worked example of protein quantification method utilising conflicted peptide information, abundance values are simulated with integers for simplicity. a.) Protein C holding only conflicted peptides is multiply subsumed. b.) Peptides 2 and 4 now become resolved and are suitable for quantitation. Peptide 3 remains conflicted as it is mapped to both protein A and protein B. c.) Peptide abundances shown in the blue boxes above the peptide identifications are used to estimate protein abundance using only unique and resolved peptides. d.) Protein A abundance (shown in yellow) is estimated using unique and resolved peptide abundances (shown in blue).e.) Protein B abundance (shown in yellow) is estimated using unique and resolved peptide abundances (shown in blue). As the ratio of estimated protein abundances was 1:2, peptide 3's abundance is shared between the**

<p><i>proteins in this ratio. f.) The abundance for conflicted peptide 3 is shared in a ratio between the two proteins it is mapped to. g) Final protein abundance calculation. ....</i></p>	<p><b>105</b></p>
<p><b>Figure 2.2.1; Representation of simulated fold-changes in benchmarking dataset PXD001819. 3, 7.5, 10 and 15 ng of <i>E. coli</i> digest was spiked-in to create the conditions of 5-, 2-, and 1.5-fold change. ....</b></p>	<p><b>107</b></p>
<p><b>Figure 2.2.2; Representation of simulated fold-changes in benchmarking dataset PXD001819. 2.5, 5, 12.5, 25, and 50 fmol/<math>\mu</math>g of yeast lysate was spiked-in to create the conditions of 10-, 4-, and 2-fold change. ....</b></p>	<p><b>108</b></p>
<p><b>Figure 2.2.3; Representation of simulated fold-changes in benchmarking dataset PXD002099. 5, 12.5, 25 and 50 fmol/<math>\mu</math>L of 48 human proteins were spiked-in to create the conditions of 25-, 12.5-, and 2-fold change. ....</b></p>	<p><b>108</b></p>
<p><b>Figure 2.2.4; Summary of stages of benchmarking pipeline. Quantitative processing was performed with Progenesis QIP, protein inference was performed with benchmarking software, DE analysis was performed through QPROT, MSstats and t-test analyses, and analysis of performance was calculated using the area under the curve. ....</b></p>	<p><b>110</b></p>
<p><b>Figure 2.2.5; Summary of the protein inference software pipeline; input data was identified and quantified peptide ion intensities, protein grouping and protein quantification was performed and output of protein intensities given. ....</b></p>	<p><b>111</b></p>
<p><b>Figure 2.3.1; Smoothed Precision-Recall plots showing the performance of the different DE analysis and quantification methods for the dataset PXD001385 across 3 fold-change simulations (1.5X, 2X and 5X) using one or two unique peptides as minimum for protein identification. Recall is plotted as a function of 1 - Precision (FDP), method of peptide ion selection is shown as line-colour. ....</b></p>	<p><b>119</b></p>
<p><b>Figure 2.3.2; Smoothed Precision-Recall plots showing the performance of the different DE analysis, summarisation methods and peptide selection for the dataset PXD001819 across 3 fold-change (2X, 12.5X and 25X) simulations using one or two unique peptides as minimum for protein identification. Recall is plotted as a function of 1-Precision (FDP), method of peptide ion selection is shown as line-colour. ....</b></p>	<p><b>121</b></p>
<p><b>Figure 2.3.3 Smoothed Precision-Recall plots showing the performance of the different DE analysis, summarisation methods and peptide selection for the dataset PXD002099 across 3 fold-change (2X, 12.5X and 25X) simulations using one or two unique peptides as minimum for protein identification. Recall is plotted as a function of 1 - Precision (FDP), method of peptide ion selection is shown as line-colour. ....</b></p>	<p><b>123</b></p>
<p><b>Figure 2.3.4; Performance of DE methods, peptide ion selection methods, and number of unique peptides required for protein identification using all non-conflicting peptides for protein quantification over three fold change simulations for datasets a.) PXD001819 and b.) PXD001385. The total area under the smoothed Precision-Recall curve is shown against the</b></p>	

recall at 5% FDP. The size of data point does not correspond to a numeric value and is used to display overlapping data points. ....	127
<b>Figure 2.3.5; Performance of DE methods, quantification methods, and number of unique peptides required for protein identification using summed peptide ions with the same sequence and different charge state over three fold change simulations for datasets a.) PXD001819 and b.) PXD001385. The total area under the smoothed Precision-Recall curve is shown against the recall at 5% FDP. The size of a data point does not correspond to a numeric value and is used to display overlapping data points. ....</b>	<b>128</b>
<b>Figure 2.3.6; Number of true positives, false positives, false negatives, and true negatives for each of the differential expression methods at significance thresholds of 0.01, 0.05, and 0.1 over three fold change simulations for data set PXD001819. Progenesis normalised peptide ion abundances with ion intensities of the same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification. MSstats performed TMP for summarisation. ....</b>	<b>129</b>
<b>Figure 2.3.7; Number of true positives, false positives, false negatives and true negatives for each of the differential expression methods at significance thresholds of 0.01, 0.05 and 0.1 over three fold change simulations for data set PXD001385 . Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification. MSstats performed TMP for summarisation .....</b>	<b>131</b>
<b>Figure 2.3.8; Number of true positives, false positives, false negatives and true negatives for each of the differential expression methods at significance thresholds of 0.01, 0.05 and 0.1 over three fold change simulations for data set PXD002099 . Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification. MSstats performed TMP for summarisation. ....</b>	<b>134</b>
<b>Figure 2.3.9 ; Precision of spike-in data across technical replicates. The percentage coefficient of variance (CV) in protein abundances between technical replicates within conditions for each of the spike-in datasets PXD001385, PXD001819, and PXD002099. Progenesis normalised peptide ion abundances with ion intensities of the same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and a minimum of two unique peptides required for identification. Only CVs below 100% are shown. Dataset PXD002099 had 7 outliers above 100%. ....</b>	<b>137</b>
<b>Figure 2.3.10; Accuracy of spike- in proteins demonstrated by log<sub>2</sub> fold-change of spike in proteins for sample comparisons in each of the spike-in datasets. Red line shows the expected log<sub>2</sub></b>	

<i>fold-change. Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification.....</i>	<b>139</b>
<i>Figure 2.3.11; Accuracy of the background proteins demonstrated with the percentage coefficient of variance of background protein intensity values across samples for each of the spike-in datasets used. Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification.....</i>	<b>140</b>
<i>Figure 2.3.12; Comparison of the Progenesis QIP normalised and raw spike-in protein abundances for the dataset PXD002099 .....</i>	<b>142</b>
<i>Figure 2.3.13; Comparison of the Progenesis QIP normalised and raw spike-in protein abundances for the datasets a.) PXD001385 and b.) PXD001819 .....</i>	<b>143</b>
<i>Figure 2.3.14; Smoothed Precision-Recall plots of reanalysed dataset PXD002099 using quantile normalisation features of MSstats and QPROT .....</i>	<b>144</b>
<i>Figure 3.2.1; Processes completed in the three stages of benchmarking workflow with details of input data required. ....</i>	<b>158</b>
<i>Figure 3.2.2; Schematic of the benchmarking pipeline. Raw data is processed with Progenesis QIP and peptide ions are identified using Mascot. Resulting peptide abundances are summarised into protein group abundances using the benchmarking inference pipeline. DE analysis is performed and significant results at a variety of significance thresholds are evaluated using enrichment analysis terms. ....</i>	<b>160</b>
<i>Figure 3.2.3; Schematic diagram showing the process for each normalisation method prior to differential expression analysis. ....</i>	<b>163</b>
<i>Figure 3.3.1; Pathway analysis validation completed with dataset PXD004682. Number of significant terms (t-test analysis, Benjamini Hochberg adjusted p-value &lt; 0.05) from DAVID enrichment analysis with different proportions of significantly differentially expressed proteins included in the search list repeated 10 times .....</i>	<b>165</b>
<i>Figure 3.3.2; Differential expression analysis of the datasets PXD004501, PXD004682, PXD007592 using statistical methods MSstats, QPROT and Welch t-test. Number of significant terms (Benjamini Hochberg adjusted p-value &lt; 0.05 with DAVID analysis) shown for different threshold levels for significant differentially expressed proteins. Threshold values are Benjamini Hochberg adjusted p-values for MSstats and t-test and FDR for QPROT analysis. Proteins groups identified with 1 and 2 unique proteins .....</i>	<b>168</b>
<i>Figure 3.3.3; Differential expression analysis of the datasets PXD004501, PXD004682, PXD007592 using statistical methods QPROT and Welch's t-test. Number of significant terms (y-axis colours represent the effect of different fold enrichment threshold cut-off values. Threshold</i>	

values (x-axis) (Benjamini Hochberg adjusted p-values for t-test and FDR for QPROT analysis). .....	173
<b>Figure 3.3.4; Enrichment analysis to assess normalisation methods. DE analysis by QPROT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide and normalised by the method shown in the x-axis strip) of three biological datasets PXD004501, PXD004682, and PXD007592 (y-axis strip). DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points. ....</b>	<b>175</b>
<b>Figure 3.3.5; Distribution of the optimal results of enrichment analysis for DE methods QPROT and t-test for dataset PXD004501. ....</b>	<b>177</b>
<b>Figure 3.3.6; Distribution of the optimal results of enrichment analysis for DE methods QPROT and t-test for dataset PXD004682. ....</b>	<b>184</b>
<b>Figure 3.3.7; Distribution of the optimal results of enrichment analysis for DE methods QPROT and t-test for dataset PXD007592. ....</b>	<b>193</b>
<b>Figure 4.1.1; Simplified scheme of the optimised proteomics pipeline. Tab separated, protein group abundance data with a header identifying the comparison groups is normalised by several methods. Each normalisation output has differential expression (DE) analysis by t-test and QPROT. DE proteins are subject to enrichment analysis with the R package ClusterProfiler using a range of significance thresholds to define significant DE. The combination of methods providing the greatest number of significant GO terms is returned to the user, along with details of the proteins changing between conditions and their functionally related pathways. ....</b>	<b>213</b>
<b>Figure 4.1.2; Command line installation of the packages required for running the pipeline. ....</b>	<b>214</b>
<b>Figure 4.1.3; Schematic of parameter combinations at each stage of the workflow. Text in black shows the number of different files being analysed at the step above. Stages run in parallel on all available nodes of computer cluster. ....</b>	<b>214</b>
<b>Figure 4.1.4; Identifying header of tab separated input file. Group comparisons are indicated by letters 1 and 2. ....</b>	<b>215</b>
<b>Figure 4.1.5; Command line installation of the packages required for running the pipeline. ....</b>	<b>215</b>
<b>Figure 4.1.6; Demonstration of how volume compared to the neighbouring volume decreases as dimension increases. a.) In one dimensional space the relative weight of the centre partition is 1/3, b.) two dimensionally it is 1/9, however, in three dimensions c.) the relative weight is just 1/27 (Betancourt, 2017). ....</b>	<b>217</b>
<b>Figure 4.3.1; Comparison of pathway analysis of differential expression analysis of biological benchmarking datasets produced by QPROT and BayesianT model using the same FDR method. ....</b>	<b>224</b>
<b>Figure 4.3.2; Comparison of FDR values produced by QPROT and BayesianT model on the same differential expression analysis output for biological benchmarking datasets. ....</b>	<b>225</b>



<i>Figure 4.3.3; Comparison of ordered FDR values produced by QPROT and BayesianT model on the same differential expression analysis output for biological benchmarking datasets. ....</i>	<i>225</i>
<i>Figure 4.3.4; Pathway analysis validation completed with dataset PXD004682. Number of significant terms (t-test analysis, Benjamini Hochberg adjusted p-value &lt; 0.05) from DAVID enrichment analysis with different proportions of significantly differentially expressed proteins included in the search list. Process was repeated 100 times. ....</i>	<i>227</i>
<i>Figure 4.3.5; Pathway analysis validation completed with dataset PXD004682. Number of significant terms (t-test analysis, Benjamini Hochberg adjusted p-value &lt; 0.05) from DAVID enrichment analysis with different proportions of significantly differentially expressed proteins included in the search list .....</i>	<i>230</i>
<i>Figure 4.3.6; Enrichment analysis to results from optimised pipeline of dataset PXD004501. DE analysis by BayesianT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points. ....</i>	<i>231</i>
<i>Figure 4.3.7; Top eight most significant terms for GO Biological Pathways, Molecular Functions, Cellular Components and KEGG pathways from enrichment analysis of proteins upregulated in malignant ascites as classified by the optimal result from our analysis (left column) and by the original paper for dataset PXD004501 (right column), count represents the number of proteins in the search group that belong to the given gene-set. ....</i>	<i>234</i>
<i>Figure 4.3.8; Enrichment analysis to results from optimised pipeline analysis of dataset PXD004682. DE analysis by BayesianT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.....</i>	<i>238</i>
<i>Figure 4.3.9; Enrichment analysis to results from optimised pipeline analysis of dataset PXD007592. DE analysis by BayesianT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points. Significant terms and number of DE proteins were not included for QPROT AIN analysis at significance threshold of 0.1 and 0.09 as all proteins were defined as DE (3079 proteins with zero significant terms) causing distortion of the plots. ..</i>	<i>243</i>
<i>Figure 4.3.10; KEGG pathways mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis. ....</i>	<i>248</i>

<i>Figure 4.3.11; Most significantly enriched GO BP terms following hierarchical clustering based on the pairwise similarities of high frequency words mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.</i> .....	249
<i>Figure 4.3.12; Most significantly enriched GO CC terms following hierarchical clustering based on the pairwise similarities of high frequency words mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.</i> .....	250
<i>Figure 4.3.13; Most significantly enriched GO MF terms following hierarchical clustering based on the pairwise similarities of high frequency words mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.</i> .....	251
<i>Figure 4.3.14; Enrichment analysis of results from optimised pipeline analysis of datasets. DE analysis by BayesianT (orange) using 650 iterations of sampling with 325 iterations for the burnin and QPROT (brown) of protein abundances (identified with by a minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points</i> .....	255
<i>Figure 4.3.15; Enrichment analysis of results from optimised pipeline analysis of datasets. DE analysis by BayesianT (orange) using 2000 iterations of sampling with 1000 iterations for the burnin and QPROT (brown) of protein abundances (identified with by a minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.</i> .....	256

**Abstract**

*Optimising the statistical pipeline for quantitative proteomics*

Hayley Price

**Background**

Label-free quantitative proteomics utilises differential expression (DE) analysis of high-throughput methods for mass spectrometry, providing insight into disease biomarkers, protein involvement in metabolic pathways or facilitating drug discovery. Applying statistical techniques to assess the significance of proteins changing in abundance is complicated by the properties of the data. Small numbers of samples containing vast numbers of features result in large sample-to-sample variation where the comparison of means can be distorted by outliers. Limitations of benchmarking data and the complexity of the algorithms make software comparison challenging. Full optimisation of the proteomics workflow is difficult, and it is a daunting task for the biologist to intuitively obtain optimal results. The aim of this Industrial CASE PhD studentship, in collaboration with Nonlinear Dynamics, the developers of Progenesis QI for Proteomics (QIP), is to provide an improved statistical pipeline that could be implemented in the Progenesis QIP workflow.

**Methods**

Benchmarking of three existing statistical approaches: QPROT, ANOVA as implemented directly in Progenesis QIP, and MSstats, was conducted traditionally, using spike-in datasets, and through the implementation of a novel method, using biological data and applying pathway analysis as an evaluation metric. Normalisation methods and the optimal threshold for defining significance were also investigated.

Following this, an optimised proteomics pipeline was developed and implemented using high performance computing cluster for parallelisation of multiple combinations of methods for DE analysis, normalisation, and significance threshold selection. Functional enrichment analysis of proteins defined as changing was used to assess the results and the optimal parameter combination returned to the user. Effectiveness of this approach was demonstrated by comparing the best results from the pipeline with enrichment analysis of the output from the current Progenesis QIP workflow.

**Results**

Overall, the results of benchmarking gave no consensus on best method for DE, normalisation method, or significance threshold and the correct combination of parameters appeared to be dependent on the characteristics of the individual datasets. The results also showed that the choice of an appropriate normalisation method is an important and underappreciated factor in differential expression analysis and that the optimal threshold for defining significance varied greatly from the generally accepted value of  $p < 0.05$ .

The optimised pipeline's performance was superior to a standard analysis using Progenesis QIP. To our knowledge, this is the only end-to-end pathway analysis pipeline designed for proteomics data, enabling users to iterate through multiple options for finding the best normalisation method and the best significance threshold for pathway analysis.

## ***Chapter 1. Introduction***

The complicated and coordinated interactions of proteins are essential for living things to carry out dynamic life processes. By studying proteins and their functions, we can begin to understand not only how proteins work at a molecular level, but also how they are involved in things such as disease, ageing, and reproduction. Cellular regulation occurs through changing protein levels and cells adapt to new scenarios or stresses by changing gene expression through new transcription and translation. The amount of protein in a cell can be predicted to some extent using transcriptomic profiling techniques, i.e., the study of the mRNA molecules in a cell, which provides information about expression levels, transcription, and degradation. However, changes in protein levels occur as part of translation and depend on how quickly proteins are turned over. Therefore, mRNA levels are weakly correlated with the final abundance of proteins in the cell. The proteome is the aggregate of all of the proteins found in an organism, tissue or sample under investigation, and while the genome is a fixed and genetically determined code, the proteome is affected by the environment and changes over time with modifications, degradations, and interactions. The study of large-scale proteomics aims to understand the outcome of these changes. Through identifying and quantifying proteins changing under specific conditions, we can better understand the biological processes involved in the conditions.

Every living system contains thousands of proteins, all with a diverse range of structures and functions. Proteins are an important class of macromolecules and are one of the most abundant organic molecules. Proteins have many cellular and organism-wide functions. For example, enzymatic, structural, transportation and signalling. The building blocks of proteins are amino acids, and their individual physical properties determine the structure, and therefore function of a protein. Amino acids have a common structure; all contain a central common carbon atom, plus an amino group, a carboxyl group, and a hydrogen atom. The 20 amino acids differ in their additional R group whose chemical properties determine the amino acid's nature; acidic or basic, polar and hydrophilic or nonpolar and hydrophobic. Amino acids link together with

peptide bonds to form polypeptide chains whose sequence determines its conformation and folding into a specifically shaped protein. Protein structure determines function, for example, by providing binding sites in enzymes. Protein structure is defined by four levels. The primary structure is the unique linear sequence of amino acids that make up the protein. The secondary structure is how that sequence folds, often into an  $\alpha$ -helix or  $\beta$ -pleated sheet. This structure is held together by hydrogen bonds, which are formed between the carbonyl oxygen of one amino acid and the amino hydrogen of another in the polypeptide sequence. A protein's tertiary structure is due to the R group interactions, including hydrogen bonds, ionic bonds, dipole-dipole interactions, and London dispersion forces, along the polypeptide chain, creating the three-dimensional structure of the protein, which defines its function. Finally, there are quaternary levels of protein structure, which are due to interactions between subunits of polypeptides or proteins.

## **1.1. Quantitative proteomics**

The aim of quantitative proteomics experiments is to deliver insight and gain a better understanding of biological processes. Mass spectrometry (MS) proteomics provides identification and quantification of the molecular composition of proteins (Aebersold and Mann, 2003). Protein samples are extracted from experimental tissue and digested into ionisable peptides, which are separated according to their mass and quantified. This strategy has become the method of choice for large-scale proteomics studies (Nesvizhskii and Aebersold, 2005), and there have been impressive improvements in instrumentation, and therefore the quality of mass accuracy and resolution, and sensitivity of detection, over a short period of time (Aebersold, 2011). High-resolution machines have improved confidence in the identification of peptides (Yates, 2019), and MS-based quantitative proteomics has become a powerful tool for gaining insight into the function and dynamics of biological systems (Ankney et al., 2016).

## Optimising the statistical pipeline for quantitative proteomics

Differential expression (DE) experiments aim to discover proteins that change due to experimental conditions, for example, in the comparison of diseased and healthy tissue ( Figure 1.1.1). DE analysis can reveal protein involvement in disease processes, which can be used to facilitate research into drug discovery for treatment and prevention. It is also used to search for prognostic biomarkers, which are proteins, genes, metabolites, or components that are indicative of a disease state. They are valuable in diagnostics, where they can classify or characterise disease or can be used to detect disease at an early stage and allow predication of disease progression, recurrence, or response to treatment.

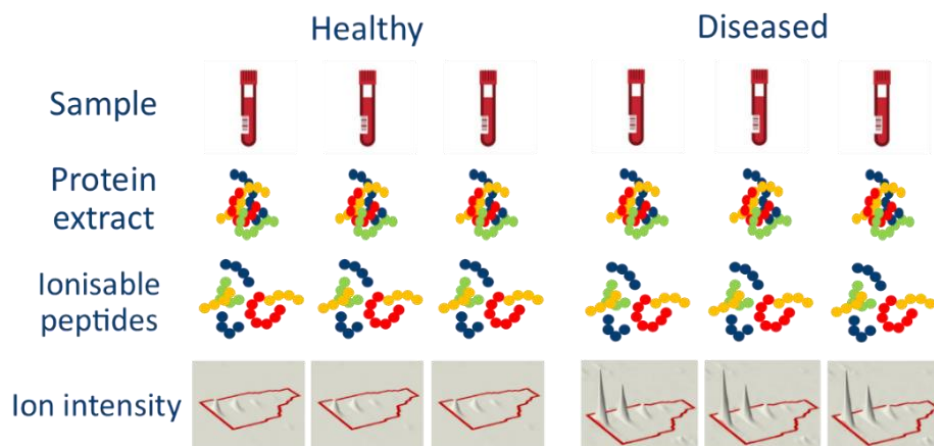


Figure 1.1.1; Quantitative proteomics experiment comparing three samples of healthy and diseased tissue. Proteins are extracted from the sample and digested to produce ionisable peptides. The detected mass spectrometric signal is used to calculate relative changes in protein abundance between the two samples, identifying those that change in the diseased state. Ion intensity map (NonlinearDynamics).

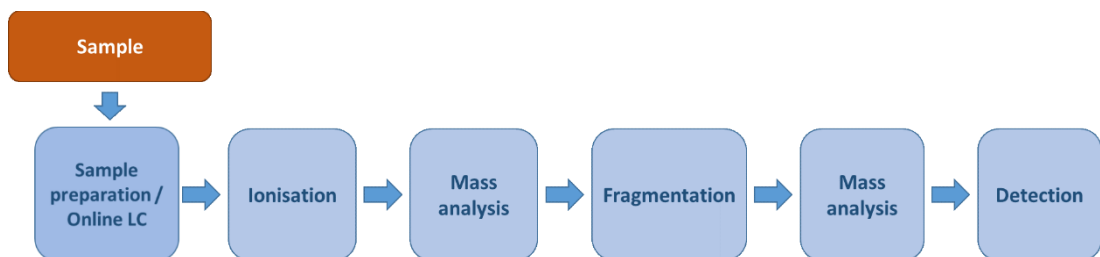
Apart from proteomics, there are many other omics applications of MS. Metabolomics, the study of small molecular weight molecules produced by organisms as the end products of cellular regulatory processes (Fiehn, 2002) uses MS due to its sensitivity and specificity (Defossez et al., 2021). Lipidomics refers to the identification and quantification of lipids in a biological system and how the function of the system is affected by lipid metabolism and functional interactions (Blanksby and Mitchell, 2010). Plus there are further subfields of proteomics, such as proteogenomics, where proteomics MS data is used to provide protein-level supporting evidence for gene models (Nesvizhskii, 2014),

and in the study of post-translational modifications (PTMs), which is important in discovering cellular regulatory processes as modifications play a role in epigenetic gene modification, regulation of gene expression, and signal transduction.

One method of proteomics experiment is referred to as ‘top-down’, proteomics where purified protein samples are subjected to direct fragmentation. It has the benefits of being fast, with straightforward sample preparation, and analysis of the intact protein can preserve valuable information about PTMs and alternative splicing isoforms. However, the procedure can be relatively insensitive (Melby et al., 2021), and fragmenting large sized analytes results in many different charges states (40+ to 60+) which produces overlapping signals, which are further overlapped with charge states of other forms of the protein (e.g. due to PTMs). This produces a very complex signal which requires high-resolution instruments with a large mass range, and informatics algorithms to deconvolute the data cannot fully interpret signals fully (Cai et al., 2016). This project focuses on the more common approach in proteomics, ‘bottom-up’ proteomics, which employs mass analysis of peptides using methods described in the next section.

## 1.2. Experimental workflow

The liquid chromatography tandem mass spectrometry (LC-MS/MS) experimental workflow (Figure 1.2.1).



*Figure 1.2.1; Experimental liquid chromatography tandem mass spectrometry (LC-MS/MS) workflow. Proteins are extracted from the sample, digested into peptides and separated using chromatography in the sample preparation stage. Peptides are ionised and subjected to mass analysis. Under data dependent acquisition (DDA) scheme, the most intense ions are fragmented and mass analysis of these fragments is produced.*

## **i. Sample preparation**

Extraction techniques are used to purify the sample; *liquid/liquid extraction* is a method where aqueous and organic solvents are used to exchange organic compounds that will be pulled from the aqueous layer into the organic layer. *Solid-phase extraction*: specific species are separated by passing the sample through a cartridge where the species is isolated through binding. Another technique is protein precipitation, where salt ions, acid, or alcohol is used to remove the water surrounding protein molecules, forcing them to precipitate.

The protein sample is then digested enzymatically into peptides to create smaller fragments that are easier to ionise. Frequently, trypsin is used, although separate or subsequent digestion with other enzymes such as Lys-C can be performed to limit any missed cleavages. Digestion can be performed in-solution for high-throughput analysis of small amounts of low-complexity samples, or using an in-gel workflow for larger, complex samples.

The digested peptides are subject to separation, which limits the complexity of the analyte reaching the detector, resulting in improved sensitivity. This also limits ion suppression, where the components in the mixture interact and affect the ability to ionise. Previously, this was performed using two-dimensional electrophoresis (Klose, 1975, O'Farrell, 1975). However, limitations in separating low-abundant proteins led to the development of high-performance liquid chromatography (HPLC) methods for use on both volatile and non-volatile samples.

The process of HPLC consists of two phases: the mobile phase, where solvents with variable polarity are used to move sample compounds into an analytical column, which is a long tube. In the stationary phase, compounds adhere to the walls of the column. The properties of the column walls determine the binding affinity of the compounds. Compounds are slowly washed along the column at speeds controlled by the affinity of the sample to the solvent and column lining, with the goal being to separate the sample. This process thins the signal, reducing the flow rate of the peptides into the mass spectrometer. To further limit the flow, micro-flow and nano-flow HPLC are performed using an



## *Optimising the statistical pipeline for quantitative proteomics*

increased column length and reduced column diameter, resulting in smaller sample droplets with higher charges per analyte, which are more easily ionised.

Several properties of peptides can be exploited to bind samples to the column. For example, charge is used in ion exchange chromatography. Amino acids in peptides are zwitterionic; they contain positively and negatively charged functional groups, and different environmental pHs will change the net charge of the peptides. In ion exchange chromatography, different buffer pHs are used to control the flow of analytes based on their charge. Positively charged peptides bind to negatively charged stationary phases, allowing negatively charged analytes to be eluted by the solvent. Bound peptides are then eluted from the column by introducing a more positively charged particle to compete with the bound protein, and the rest of the analyte is eluted from the column. Size exclusion chromatography uses a column packed with a porous matrix of beads to filter molecules by their size. Polarity can also be used for separation in hydrophilic and hydrophobic interaction chromatography. Peptides display different polarities due to the different arrangements of functional groups. Stationary-phase molecules with similar polarity to the sample will attract those compounds and isolate them from the sample. In normal-phase chromatography, a hydrophilic column, such as silica or aluminium, is used in the stationary-phase to adsorb hydrophilic compounds. In reversed-phase (RP) HPLC, hydrophobic compounds are absorbed into a hydrophobic column such as alkylated silica. The analyte is then eluted from the column with a non-polar organic solvent. Elution can occur isocratically, where the composition of the solvent does not change, or using gradient elution, where the organic composition of the mobile phase increases over time. Solutes remain bound to the stationary phase until a specific concentration of organic solvent is introduced, causing elution and allowing a mixture of solutes with a wide range of retention factors to be eluted over time. RPHPLC is useful in proteomic MS due to the relatively hydrophobicity of peptides, and it uses volatile mobile phases that are compatible with MS. It provides good purification of the sample through its desalting and has high peak resolving power, allowing the separation of very similar peptides.

Capillary electrophoresis (Jorgenson and Lukacs, 1981) is another technique used to reduce the flow of analyte reaching the ion source. It uses an applied voltage to create an electric field, separating ions based on their electrophoretic mobility, which depends on the molecules' charge, radius, and viscosity. Here, the flow is flat as opposed to the parabolic flow from HPLC, resulting in narrower peaks and better resolution. Another chromatography method, mostly used for small molecules, is gas chromatography, which also uses mobile and stationary phases for peptide separation.

## **ii. Ionisation**

Peptides eluted from the chromatography column are ionised, and non-volatile samples are converted to the gas phase for mass spectrographic measurement. The most common methods are electrospray ionisation (ESI) (Fenn John et al., 1989), which is the most common technique used in LC-MS/MS, and matrix-assisted laser desorption ionisation (MALDI) (Karas and Hillenkamp, 1988, Tanaka et al., 1988), which is generally used in intact protein analysis. These processes allow molecules to remain intact during the ionisation process and are described as '*soft* ionisation techniques' as they limit the amount of fragmentation that occurs.

ESI is an atmospheric pressure ionisation technique (Figure 1.2.2). An aerosol is produced from the liquid peptide sample as it is forced through a narrow metal capillary, which has a large potential difference applied between its inlet and the outlet to the mass spectrometer. As the liquid travels through the capillary's electric field, it becomes atomised into small, charged droplets, which are expelled from the capillary tip in a fine jet. As the solvent in the droplets evaporates, they reduce in size and become more highly charged. A gas, often nitrogen, is used as a nebuliser to aid evaporation, and when the droplets shrink so small that their surface tension cannot be maintained, they rip apart, producing even smaller droplets. Evaporation in the smaller droplets also occurs, and the process is repeated until no solvent is left and only charged sample ions remain.

## Optimising the statistical pipeline for quantitative proteomics

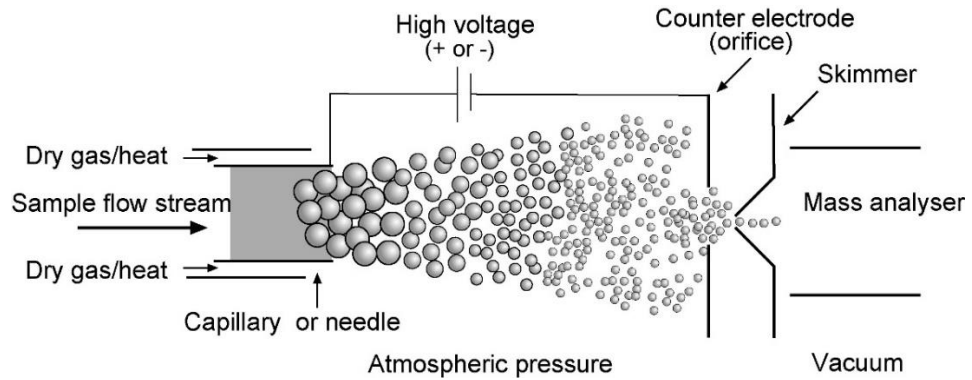


Figure 1.2.2; Electrospray ionisation. The liquid peptide sample becomes charged as it passes through the capillary and forms sample ions as the solvent evaporates from the droplets. (Kicman et al., 2007)

In MALDI MS (Figure 1.2.3), the non-volatile sample is mixed with a large quantity of an energy-absorbent organic chemical matrix, which crystallises as it dries, also crystallising the peptide sample within the matrix. Ultraviolet laser pulses irradiate and sublimate the matrix (desorption), which, being in a greater concentration than the sample, absorbs most of the energy, allowing the sample to remain intact. The dense gas cloud moves quickly towards a vacuum, and energy is transferred from the matrix to the sample, ionising and desorbing it.

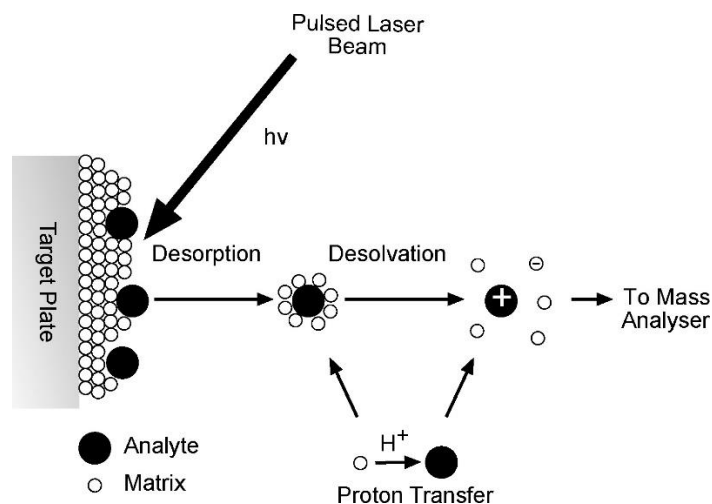


Figure 1.2.3; Matrix-assisted laser desorption ionisation. The sample and the matrix dry to form a crystalline structure that is targeted by a laser. The energy absorbed causes desorption and desolvation, allowing a charged sample ion to enter the mass analyser. (Kicman et al., 2007)

### **iii. Mass analysis and fragmentation**

#### ***Mass analysers***

Ionised particles are analysed by a mass analyser, which separates the ions according to their mass to charge ratio ( $m/z$ ). There are three main types of mass analysers: quadrupole, time of flight (TOF) (Wolff and Stephens, 1953) and ion trap. Usually, MALDI techniques are coupled with TOF and ESI with ion traps or TOF analysers. They can also be connected together in series. For example, a Q-TOF is a quadrupole mass analyser connected with a TOF mass analyser. A quadrupole mass analyser has a column consisting of four rods arranged to have opposite polarity. The rods have a combination of direct and alternating current applied to them, which generates an oscillating electric field. Ions are emitted from the ion source and travel through the column where they become variably destabilised by the fluctuating electric field. A unit mass resolution window is created by fine-tuning the applied voltage, allowing only ions of a specific mass range to pass through the column and be detected. The full mass spectrum is measured by quadrupoles as a series of individual mass-to-charge ratios. A TOF mass analyser separates ions by their velocity; the lighter the ion, the more quickly it will travel. Equally, charged ions are accelerated with an electric field until they have the same kinetic energy, and the time taken to reach the detector is used to calculate the mass of the ion. This allows all of the  $m/z$  to be measured simultaneously. The resolution of TOF analysers depends on all of the ions being equally charged and starting in the same direction, at the same time and from the same position. Resolution can be increased using a reflectron (Mamyryn et al., 1973) which is an ion mirror made of a retarding electric field that focuses ions of different kinetic energy. Ion trap analysers use a combination of alternating and direct current that produces electric fields to 'trap' and hold ions in place where measurements can be made. The Orbitrap (Hu et al., 2005), (Figure 1.2.4), is a more recent type of ion trap mass analyser that provides high-speed, full-scan analysis along with high resolution and mass accuracy (Zubarev and Makarov, 2013). It consists of tangentially injected ions, and three electrodes, one central and two outer, that allow for both mass analysis and detection. Measurement of these ions utilises

## Optimising the statistical pipeline for quantitative proteomics

the cyclotron principle, which describes the circular motion of ions in a magnetic field. When a voltage is applied between the central and outer electrodes, a radial electric field is created which bends the ion trajectory towards the central electrode, opposing the centrifugal force created by the tangential velocity. The result is a circular-spiral electron beam within the trap whose axial oscillations can be detected by the outer electrodes in a smaller instrument. The oscillations are recorded as a detected image current, and using the knowledge of how a particular  $m/z$  responds to a known magnetic field strength, a Fourier transformation converts this information from the time domain to the frequency domain. Despite the high-throughput advantages, the data produced can be difficult to interpret as it can contain high peak density artefacts that do not correspond to actual ions. Although information from these artefacts can be used practically for improved precision, their presence increases the complexity of the data and, therefore, is removed using a pre-processing algorithm.

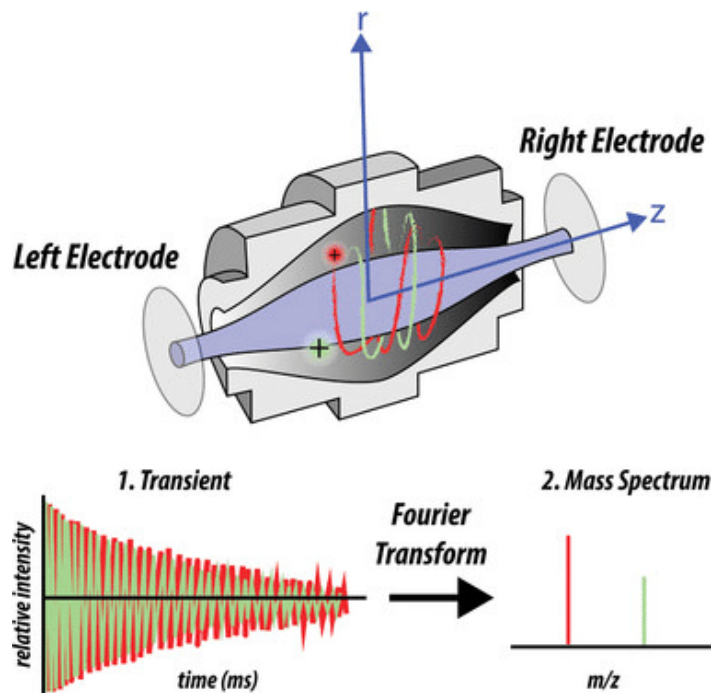


Figure 1.2.4; The Orbitrap mass analyser. Ions (depicted by green and red lines) oscillate both round and along the axis of a central spindle-like electrode. 1.) The detected oscillation is recorded as a transient signal, the axial ion oscillation frequency, which is unique for each  $m/z$ . 2.) Using Fourier transformation, the frequency is converted into a mass spectrum. (Savaryn et al., 2016)

## ***Optimising the statistical pipeline for quantitative proteomics***

The separated and fragmented ions are detected and counted under vacuum as they emerge from the mass analyser, and their  $m/z$  is calculated according to the properties of the ions. In ion trap and quadrupole mass analysers, only ions with a specific  $m/z$  reach the detector, depending on the electric field applied to the rods. In reflectron TOF instruments, ions deflected by electric fields are diverted to different degrees depending on their speed and charge; heavier ions move more slowly and are deflected less, while lighter ions have greater speed and greater deflection. Abundance is calculated using the flux of the detected electrically charged ions, which is converted into a proportional electrical current, with more abundant ions producing a greater signal. This is read by the data system and converted to digital information, which is displayed as a mass spectrum. Types of detectors include electron multipliers (Allen, 1947) and Faraday Cup detectors (Brown and Tautfest, 1956), which use secondary emission to induce amplified electron emissions from a single electron. In Orbitrap mass detectors, the axial ion oscillation frequency and Fourier transformation are used to calculate  $m/z$ .

### ***Fragmentation***

Fragmentation methods include collision-induced dissociation (CID) (Mitchell Wells and McLuckey, 2005), where the ions are fragmented in a collision cell where they are accelerated and collide with a neutral gas. In the so-called 'mobile proton model,' which theorises that mobilisation of a proton to a carbonyl causes the peptide backbone to break, this increases the ion's internal energy, and fragmentation occurs when internal energy exceeds the critical energy of the bond. High-energy collision dissociation (Olsen et al., 2007), available for the Orbitrap, is a beam-type collisional activation dissociation method employing higher energy dissociations, enabling ion detection over a wider range of fragmentation pathways (Jedrychowski et al., 2011). The ion trap uses resonant CID where gas-phase collision-induced dissociation by resonant excitation occurs. Ions are excited to higher kinetic energy through resonance; a small supplementary AC potential resonant with the secular frequency of the mass-selected parent ion is applied between the end-cap

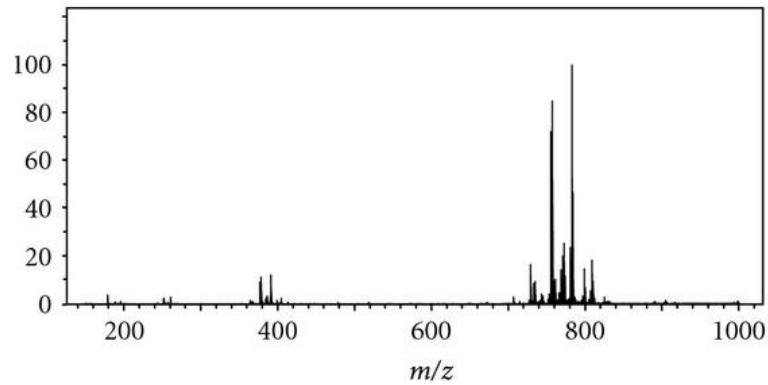
electrodes of the ion trap. The high-energy ions collide with a gas, such as helium or argon, and dissociate into fragments (Xu et al., 2014). Electron transfer dissociation (Syka et al., 2004) is another method where radical cations are produced from ions by electron transfer through bombardment of radical anions, again causing fragmentation by dissociation along the peptide backbone. This process is considered softer than CID, and is a technique useful for site localisation in phosphorylation due to the ability to preserve labile side chain modifications.

### ***Acquisition modes***

The resulting mass spectra ( $MS^1$ ) show peaks relating to mass/charge ( $m/z$ ) versus intensity (ion counts) of intact peptides. Certain peptides are again subjected to fragmentation, producing a second mass spectrum ( $MS^2$ ). This process is tandem mass spectrometry (MS/MS). Data-dependent acquisition (DDA) is where the most intensely abundant peptides are selected for further fragmentation and mass spectrographic analysis, and the MS/MS results are used for peptide identification (Hunt et al., 1986, Eng et al., 1994). However, this can cause under-sampling as only a fraction of the detected ions are used (Johnson et al., 2013). To overcome this, data-dependent acquisition (DIA) (Venable et al., 2004) or sequential window acquisition of all theoretical fragment ion spectra (SWATHs) (Gillet et al., 2012) were developed. In DIA acquisition mode, all available precursor ions are subjected to fragmentation tandem mass spectrometric analysis, maximising peptide identification. Due to the increased dynamic range, the amount of data produced by DIA experiments has greatly increased, and the multiplex nature of the analysis is more challenging. While software and methodologies are improving for this emerging technique, this report focuses on DDA workflows.

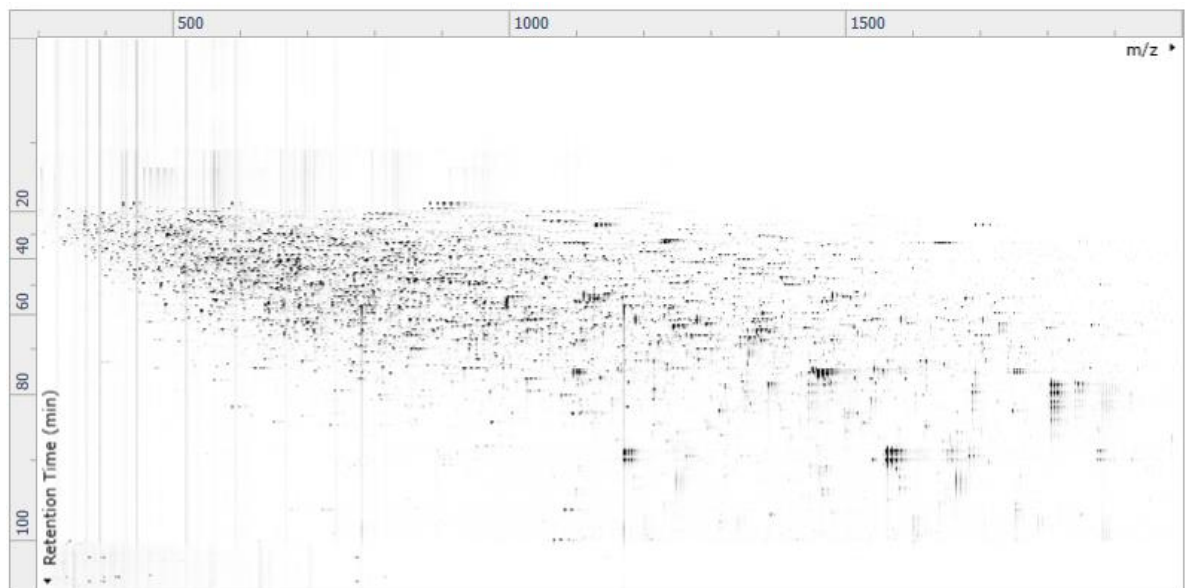
#### **iv. The mass spectrum**

The result of mass analysis is a mass spectrum (Figure 1.2.5), which shows the distribution of intensity (approximate ion counts) versus the  $m/z$ , for all molecular ions detected in a given scan of the MS.



*Figure 1.2.5; Example mass spectrum (Quinn et al., 2012)*

A combination of this information for all MS scans can be displayed as a two-dimensional ion-intensity map (Figure 1.2.6) which allows visualisation of the data as bird's eye view. Discrete data (scans over retention time) on the y-axis is stitched together to infer the continuous signal of molecules measured, with  $m/z$  on the x-axis, and feature intensity shown with highly abundant ions represented by darker areas.



*Figure 1.2.6; Ion intensity map from software analysis package Progenesis Q1 for Proteomics (<http://www.nonlinear.com/progenesis/q1-for-proteomics/>)*



## ***Optimising the statistical pipeline for quantitative proteomics***

The same information is also displayed as a corresponding total ion chromatogram (TIC) (Figure 1.2.7) which shows the intensity of the same detected ions as a function of elution time. As the TIC is calculated by summing the intensities of all mass spectral peaks belonging to the same scan, it is not especially useful in quantitative proteomics, since peaks from complex mixtures would contain intensity information from many different features.

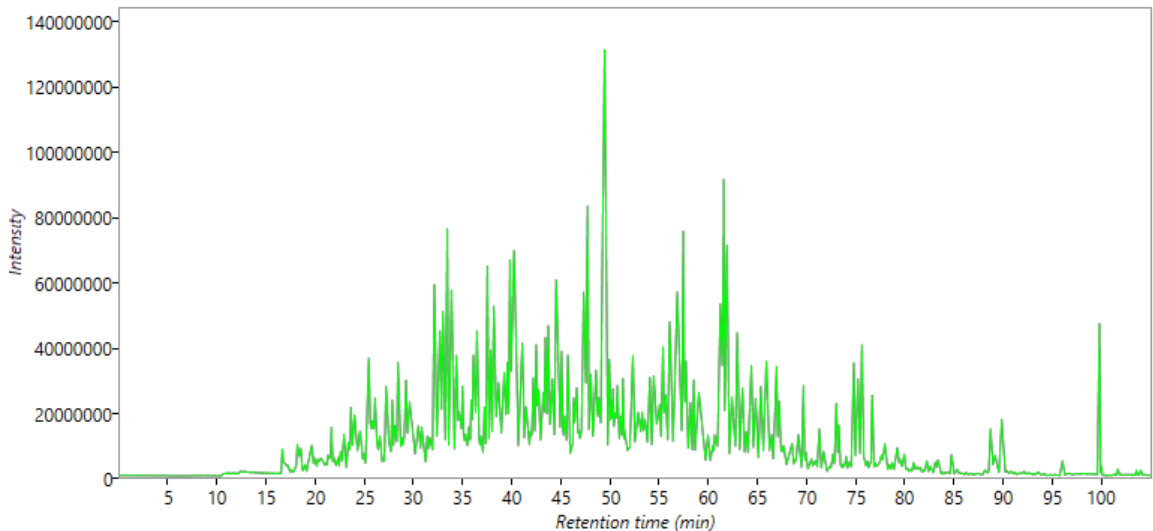
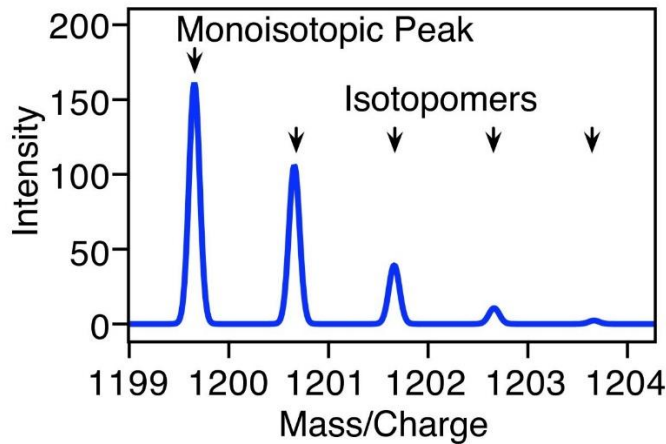


Figure 1.2.7; Total ion chromatogram from software analysis package Progenesis Q1 for Proteomics (<http://www.nonlinear.com/progenesis/q1-for-proteomics/>)

### ***Isotopes***

Isotopes are naturally occurring versions of the same element. They have the same number of protons and electrons but have different number of neutrons, giving them different masses. The resulting mass spectrum of two isotopes of an element with the same charge will show two individual peaks separated by the difference in mass of the two isotopes (1 Da for a charge state of 1). In proteomics, the most common causes of isotopic peaks are the presence of  $^{13}\text{C}$  as opposed to  $^{12}\text{C}$ , along with a small amount of  $^{15}\text{N}$  rather than  $^{14}\text{N}$ . The resulting mass spectrum of a small peptide containing isotopes (Figure 1.2.8) is displayed as an isotopic cluster, a series of peaks, the largest being the most abundant monoisotopic peak, followed by smaller peaks at higher mass/charge ratios due to the presence of heavy isotopes. However, as peptides increase in size (approximately > 2000 Da), the monoisotopic peak is no longer the most

intense peak in the isotopic cluster, as there is a higher incidence of heavy isotopes due to the increased number of C and N atoms in the peptide molecule. As this number becomes larger across the population of molecules, there is a smaller probability (and thus proportion) that a given molecule will have purely  $C^{12}$  and  $N^{14}$ .



*Figure 1.2.8; Mass spectrum of a peptide NVLPQRSTVW. The monoisotopic peak is the most abundant peak, with four additional peaks at higher mass/charge values, separated by 1 Da, due to the presence of naturally occurring heavy isotopes of the single-charged peptide (Sykes and Williamson, 2008)*

The presence of multiply charged ions, formed when a precursor ion interacts with additional atoms during ionisation, will be shown as an isotopic cluster separated by 1 Da divided by the charge state of the ion; a peptide of 2+ charge will have isotopic peaks separated by 0.5 Da and 3+ separated by 0.333 Da.

## **v. Labelling**

Labelling is a technique used to differentiate and quantify proteins and peptides between samples, providing identification of proteins in complex mixtures and relative abundances. A defined mass is incorporated into the sample, or samples are enriched with different heavy, stable isotopes of amino acids or elements, then mixed and analysed as a single sample. Further benefits of labelling is the ability to perform multiplexing, where many different samples can be processed in one MS run, which is quick and efficient, and produces reliable data with limited variation. Labels can be applied at the cellular level in metabolic labelling where living cells are modified to include an identification tag, for

## ***Optimising the statistical pipeline for quantitative proteomics***

example in *stable isotope labelling with amino acids in cell culture* (SILAC) (Ong et al., 2002). Heavy isotopes such as  $^2\text{H}$ ,  $^{15}\text{N}$ ,  $^{13}\text{C}$ , and  $^{18}\text{O}$  are introduced using cell culture where the sample is grown, creating a 'heavy' labelled sample where the isotope is incorporated into the proteome. An identical sample is cultured using the natural growth medium without added isotopes to provide the 'light' sample. The heavy and light samples are then treated according to the experimental conditions, and equimolar ratios of the two cell populations are mixed before mass spectrometric analysis. The peaks of the heavy and light samples will show a defined number of Dalton mass shifts, allowing identification of the peak associated with the light or heavy sample. This technique can detect small relative changes in protein levels, but samples can be expensive or impossible to label. The process is time-consuming due to the growth period for incorporating isotopes, and as they are living samples, they must be maintained.

A second labelling technique is by chemically labelling samples through the addition of stable isotopes where reagents react with functional groups of the polypeptide. Initially, this was applied to intact proteins in *isotope-coded affinity tags* (ICAT) (Gygi et al., 1999). However, this method could only be performed on proteins containing cysteine, and had other technical limitations meaning it is rarely used nowadays. Methods where digested peptides are labelled have also been developed; *tandem mass tags* (TMT) (Thompson et al., 2003) and *isobaric tags for relative and absolute quantitation* (iTRAQ) (Wiese et al., 2007), which provide identification of individual samples after secondary MS. Isobaric reagents are covalently bound to free amines, often lysines in iTRAQ and amine, cysteine, or carbonyl groups in TMT. Reagent kits are available to allow 4plex, 8plex, 10plex, and up to 16 multiplexing. All have kits have the same total mass and differences between the labelling of the peptides is only observable after fragmentation. Reagents consist of charged reporter groups, each with a unique mass and used for the identification of samples, and a neutral balance group, which makes up the total mass to ensure it is constant across all of the reagents. The balance group is released during fragmentation in MS/MS, producing an ion with a known  $m/z$ , and the resulting sample peaks are separated by the specific

mass of the reporter group used to label the sample. The disadvantages of this technique are that it is expensive and the reagents are sensitive to salt contamination. In addition, the higher number of samples processed, the higher resolution MS required for processing the complex signal. There can also be variability introduced due to enzymatic digestion differences.

An alternative to labelling is *label-free* (LF) methods, where samples are processed in separate MS runs and intensities are compared to investigate changes in protein levels between conditions. LF sample preparation is fast, cheap, and straightforward, and the study design is adaptable with no absolute limit to the number of comparisons that can be made. LF studies are suitable for large scale experiments have higher coverage and can identify highly abundant proteins. However, LF studies produce large quantities of data that require diverse techniques with complex computational workflows to interpret. Due to processing samples, separately there is increased running time and additional technical variance is introduced that requires downstream analysis to correct. The LF processing pipeline is described in the next section.

## **vi. Top-down proteomics**

A further method of proteomics experiment is referred to as 'top-down', proteomics, where purified protein samples are subjected to direct fragmentation. It has the benefits of being fast, with straightforward sample preparation, and the analysis of the intact protein can preserve valuable information about PTMs and alternative splicing isoforms. However, the procedure can be relatively insensitive (Melby et al., 2021), and fragmenting large-sized analytes results in many different charge states (40+ to 60+) which produce overlapping signals that are further overlapped with charge states of other forms of the protein (e.g., due to PTMs). This produces a very complex signal that requires high-resolution instruments with a large mass range, and informatics algorithms to deconvolute the data (Cai et al., 2016). This project focuses on the more common approach in proteomics, 'bottom-up' proteomics, employing the mass analysis of peptides.

## **1.3. LCMS data processing pipeline**

### **i. Proteomics workflow**

The typical LF workflow consists of processing with specialised software. Raw mass spectrometric data is converted into peak data which must be aligned to account for retention time shifts between separately processed runs, identified using database searches, and quantified using the number of spectra identified or ion intensity values. Then, as during sample digestion, the link from peptide to protein is lost, identified peptides must be inferred back to the protein they are likely to have derived from before peptide quantities can be aggregated into protein abundances. Protein quantification can consist of relative quantification, where the fold change in protein expression between conditions is calculated, or absolute quantification, which calculates individual protein abundances within samples. Finally, statistical software is used to accurately select which proteins change due to experimental conditions

### **ii. Proteomics software**

To address the challenge of extensive data processing required for LF workflows there is a multitude of proteomic software solutions. Some support the whole data analysis workflow; others focus on a specific stage of the process such as search engines for identification, applications for quantification or algorithms for statistical analysis (which is discussed further in the chapter). Most search engines work in a relatively similar principle, comparing fragmentation spectra against theoretical spectra generated from a peptide sequence database. There are commercial search engines, such as Mascot (Perkins et al., 1999) and PEAKS (Tran et al., 2019), free software, such as MaxQuant (Cox and Mann, 2008), and fully open source search engines such as Comet (Eng et al., 2013), X!Tandem (Bjornson et al., 2008) and MSGF+ (Kim et al., 2008, Kim and Pevzner, 2014). For quantitative software, the most popular free software is MaxQuant, which started as SILAC-based analysis for Thermo instrument data, but now supports label-free and TMT analysis, DIA analysis, and multiple instrument vendors.

## *Optimising the statistical pipeline for quantitative proteomics*

Progenesis QI for Proteomics (QIP) is a commercial product that was designed for label-free quantitative proteomics but was adapted to support MSe (Waters) analysis, which is a type of DIA. Support for labelling data can be provided by an additional plugin called Proteolabels, which was developed by our team.

Progenesis QIP provides a user-friendly, menu-guided, data-processing workflow with options for automated processing and support for all major instrument vendors. The software takes a unique approach to ‘quantify–then-identify’ (described further in this chapter), which is intended to promote detection of low-abundant peptides that would have less chance of being detected in typical DDA workflows.

This project began as an Industrial Cooperative Awards in Science and Technology (CASE) PhD studentship in collaboration with the software development company Nonlinear Dynamics, a subsidiary of the Waters Corporation, the developers of Progenesis QIP. The main aim of the project was to provide an improved statistical pipeline that can be implemented in the Progenesis QIP workflow. Therefore, this thesis focuses on Progenesis QIP data processing rather than other software such as MaxQuant that is more commonly used. The intention of the CASE studentship was that a period of time would be spent working at Nonlinear Dynamics. However, due to unforeseen circumstances, the expected work experience period did not occur, and the PhD was conducted solely under the supervision of the University of Liverpool supervisors.

## 1.4. Progenesis QI for Proteomics

Proteomics QIP (<https://www.nonlinear.com/progenesis/qi-for-proteomics/>) is a software analysis package for processing raw MS data into quantitative protein group intensities. The Progenesis QIP data processing pipeline (Figure 1.4.1) is described in detail in this section.

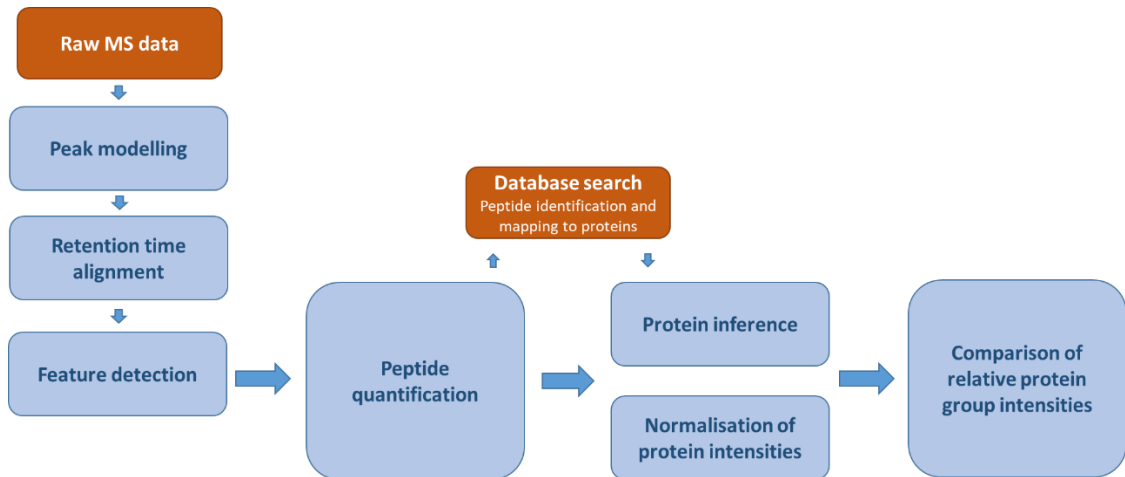


Figure 1.4.1; Progenesis QIP bioinformatics workflow for processing label-free quantitative proteomics experiments; RAW data is imported and compressed into simplified modelled ion intensities. Normalisation and alignment allow the information from all experimental runs to be combined for analysis. Fragmented peptide ion information is used for database searches for peptide and protein identification. Protein grouping and abundance normalisation provide protein group intensity information for comparative analysis.

### i. Alignment

Using a wavelet based approach, Progenesis QIP performs peak modelling for simplification and compression, converting the many-point peak data from the RAW MS files into peak models that retain all quantitative and positional information. In LF techniques, samples subjected to different experimental conditions and technical or biological replicates have separate MS analysis as opposed to combined analysis in labelled experiments. Variation in retention time is introduced due to different sample handling and separation procedures. Sample ions are aligned to compensate for this, allowing comparison of the peptide and protein abundances from different runs; an ion in run A, shown as the pink ion in Figure 1.4.2, will be in the same location as a matching ion, shown in green, in run B, giving retention times that are equivalent across all runs. The most appropriate run to use as a reference can be chosen

automatically by the software from all or a selection of runs; each run is compared to the other runs and assessed for similarity with the run most similar to the other runs selected as the reference run.

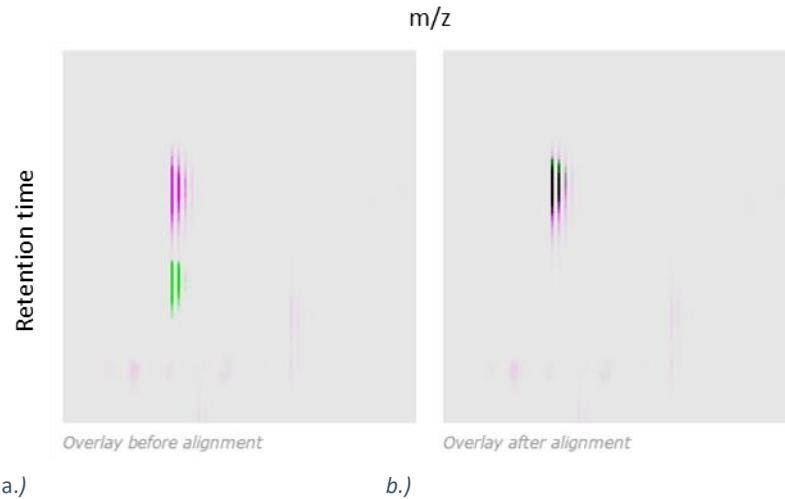


Figure 1.4.2; Alignment of a peptide ion in two overlaid runs (shown in pink and green) represented as  $m/z$  versus retention time. a.) Before alignment, due to retention time differences, the ions do not overlap. b.) After alignment, the ions are in the same location and combine to show a single feature (<http://www.nonlinear.com/progenesis/qi-for-proteomics/v2.0/faq/why-is-alignment-so-important.aspx>).

Landmarks called alignment vectors are placed, connecting the location of a peptide in the reference run with the location of the same ion in the run being aligned. Each run is automatically aligned to the reference run separately, producing a rough initial alignment, which is used to produce an optimal alignment. The alignment quality can be visualised, reviewed, and edited with vectors being placed manually. Once placed, the vectors are used to calculate a non-linear mapping between the retention times of the reference run and those of the run being aligned.

## **ii. Feature detection and ion abundance quantification**

Once alignment is complete, the software performs feature detection. During the peak modelling stage, the discrete data on the retention time axis, acquired through sampling every second or so, is converted into continuous



## Optimising the statistical pipeline for quantitative proteomics

measurement of intensity over time. A consensus map, which is an aggregate image of features, is produced from a single set of peptide ion outlines that includes information from all of the runs. Ion intensity is measured by summing the intensities of the isotopes of the peptide within a boundary. Features are a set of isotopic peaks separated by  $m/z$  interval of one Dalton; the most abundant is the monoisotopic peak, the second peak is the same peptide but containing an atom with an extra neutron, the third with two atoms with an extra neutron and so on. Figure 1.4.3 a.) is a two-dimensional ion-intensity map showing the red peptide ion boundary that surrounds the isotopes that form the same peptide ion (shown as black shaded areas). Figure 1.4.3 b.) shows the three-dimensional ion intensity maps of the peptide ion and its isotopes. Quantification is calculated using the total area under the peaks of all isotopes within the peptide ion boundary using the MS1 survey scan before any identification takes place.

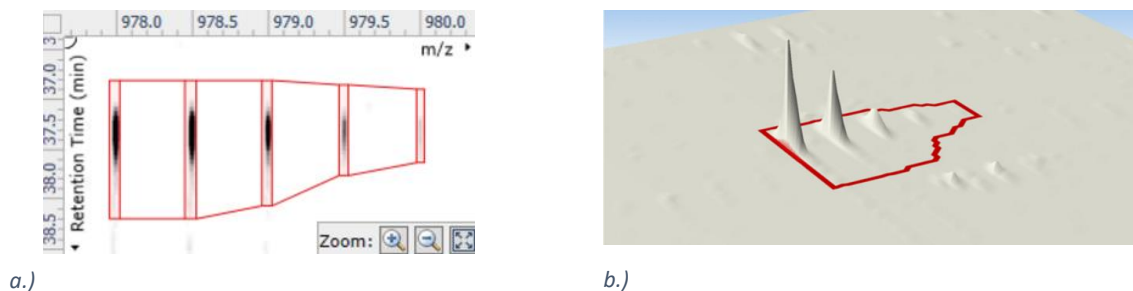


Figure 1.4.3; a.) Two-dimensional (taken from analysis software package Progenesis Q1) and b.) Three-dimensional ion intensity maps. Isotopes of the same peptide ion shown as black shaded areas (a) or peaks (b) within the red boundary line. Peptide ion abundance is calculated by summing the areas below the scan lines of each of the isotopes within the boundary (<http://www.nonlinear.com/progenesis/q1-for-proteomics/how-it-works/>)

### Missing values

Analysis of quantitative LCMS experiments can be hindered by missing values in the data, which can occur for biological and technical reasons. This can be because the peptide is present but its abundance is below the detection limit of the instrument or has been misidentified. They can also occur due to issues with experimental processing, such as incomplete ionisation and miscleavage of peptides. Alternatively, the peptide may truly be missing from the sample. Techniques to deal with missing data in proteomics include imputation of

missing values based on statistical estimations and removing incomplete data from analysis. However, statistical methods are more successful, depending on the reason for the missing value. Randomly missing values often arise from technical limitations, whereas non-randomly missing values tend to be abundance-dependent (Jin et al., 2021), and proteomics data is believed to constitute an unknown mixture of the two. In Progenesis QIP, there are very few missing values in the resulting data (Välikangas et al., 2017). This alignment technique allows for confident feature detection to maximise the amount of peptide ions quantified (Al Shweiki et al., 2017).

## **1.5. Peptide identification and protein inference**

### **i. Peptide identification**

In shotgun proteomics, where the protein sample is digested into peptides early in the workflow, the direct link between peptides and proteins is lost and must be now inferred from the data. Using specific search engines such as Mascot (<http://www.matrixscience.com/>), the detected MS2 fragments can be compared against proteome-specific sequence databases (Eng et al., 1994). The Mascot search engine (Brosch et al., 2009) allows identification of peptides and proteins by comparing the experimental mass of MS2 ions to the calculated mass of peptides or fragment ions obtained by predicting peptide cleavage. Details of how a peptide breaks up into fragments are dependent on the method of fragmentation (Révész et al., 2021) and can be used to estimate the unique mass of the resulting fragments, which are stored as sequence databases. Database sequence fragments that match the experimental results provide a peptide spectrum match (PSM) and possible parent peptide sequences with an attached probability score (Nesvizhskii, 2010). Usually, the highest scoring peptide to PSM is considered a match, and completeness of the sequence database is essential to the reliability of the identifications. Accurate search parameters must also be supplied; fixed and variable modifications define what mass to consider for specific residues; parent and fragment ion mass tolerance

allows for a window of error on mass values; enzyme configuration specifies the location of digestion cleavage.

## **ii. Protein inference**

Protein identification is made based on confidently identified peptides. Problems arise due to the peptide-centric nature of shotgun proteomics. Many peptides are not 'unique' to one protein; the same peptide sequence can be present in many different proteins, and database searches are unable to tell which protein the peptide could have derived from. They may share the same sequence of amino acids, be from the same gene and differ slightly due to alternative splicing, single- nucleotide polymorphisms, or posttranslational cleavage, be ancestrally related species, or from the same gene family (Rappsilber and Mann, 2002).

For quantitative DE analysis, this is not ideal. For accurate quantitation, it is important to determine how strong the evidence for a protein being present is, and to reduce long lists of possible proteins for simplified analysis. Some peptides in the search will only be mapped to a single protein; these are labelled as 'unique peptides' and provide a definite identification of a 'distinct protein'. If a search result shows a peptide mapped to two or more possible proteins, it is termed 'shared'. Shared peptides cannot identify a single parent protein and therefore lead to identification of less specific protein groups of possible proteins. The reporting of excessive lists of questionable protein identifications damages the precision of a quantitative proteomics experiment (Keller et al., 2002) and current workflows tend to employ the Parsimony Principle, where proteins are only reported if there is independent evidence of their presence in the sample. Protein inference software uses rules to define peptides based on how much evidence there is to support the mapping. These rules are described using worked examples in Figure 1.5.1 to Figure 1.5.4 based on the terms used by Jones (2017). Peptides are said to be reliable when they are only mapped to one possible protein sequence in the database. Proteins that do not have independent evidence for their presence in the sample are considered as a

## Optimising the statistical pipeline for quantitative proteomics

group of possibly identified proteins or removed from the analysis to give a parsimonious result.

- A peptide mapped to only one protein is labelled as a **unique peptide**.
- Proteins containing a unique peptide have independent evidence that they are present in the sample and are labelled as **distinct proteins**, which lead to protein groups. The distinct protein's non-unique peptides that also map to other proteins that have independent evidence are labelled as **conflicted peptides**.
- Proteins containing only peptides that belong to distinct proteins but have no unique peptides themselves are labelled as **multiply subsumed** and are not included in the analysis.

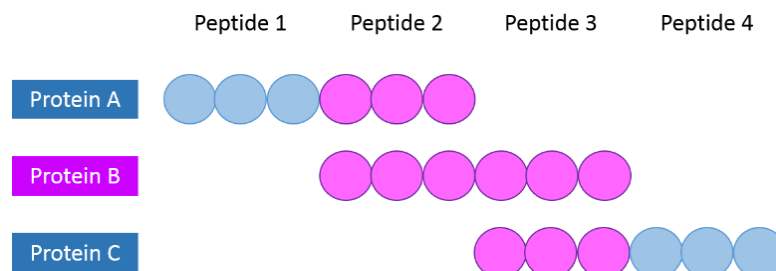


Figure 1.5.1; Protein grouping example 1. Peptide 1 is only mapped to protein A; therefore, it is labelled unique and protein A is distinct. The same is true for peptide 4 and protein C. Peptides 2 and 3 are both mapped to two proteins, (A and B) and (B and C) respectively. These peptides are labelled as **conflicted** peptides. As there is no independent evidence that protein B is present in the sample, due to parsimony principle this protein is labelled multiply subsumed and will be discarded.

Protein grouping example 1 (Figure 1.5.1). Peptide 1 is only mapped to protein A; therefore, it is labelled unique and protein A is distinct. The same is true for peptide 4 and protein C. Peptides 2 and 3 are both mapped to two proteins, (A and B) and (B and C) respectively. These peptides are labelled as **conflicted** peptides, since their intensity signal must derive from a mixture of more than one protein molecule. If one protein is going up in abundance in a comparison of two experimental groups, and the other protein is going down in abundance in the same comparison, it is clear to see that the signal we observe from these peptides would not be a useful proxy for determining what is happening at the protein-level. Due to parsimony principle, as there is no independent evidence

## *Optimising the statistical pipeline for quantitative proteomics*

that protein B is present in the sample, this protein is labelled ***multiply subsumed*** and will be discarded (Figure 1.5.2).

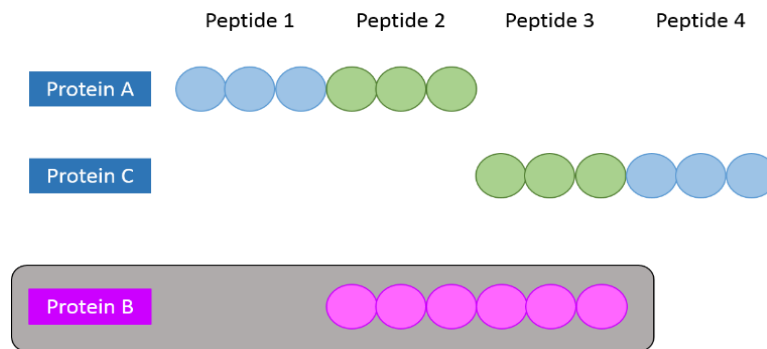


Figure 1.5.2; Protein grouping example 1 after rules of parsimony are applied; protein B is discarded and peptides 2 and 3 can now be called resolved.

- If the protein contains a peptide that does not belong to a distinct protein, it will form a ***same set*** group (Figure 1.5.3). The head of the group will be the protein with the most evidence for its presence – the one with the largest number of identified peptides.

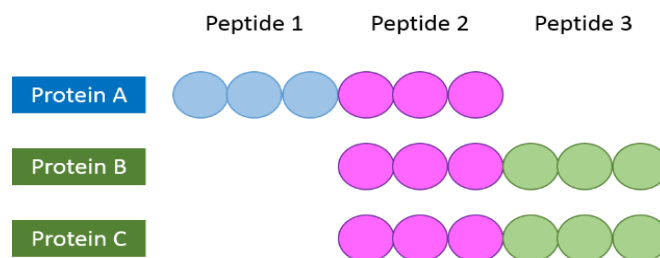
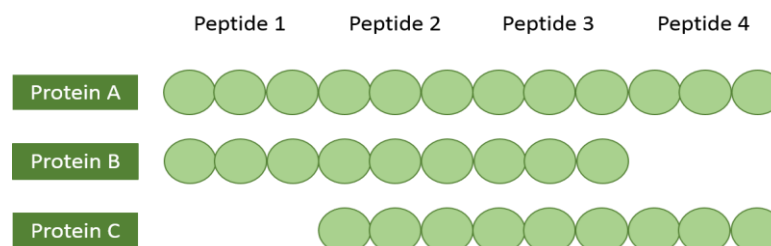
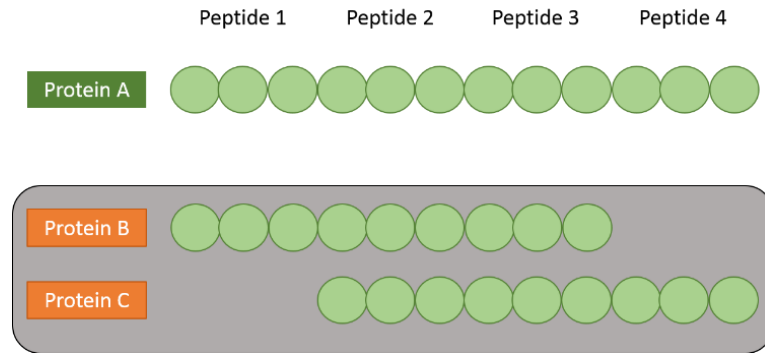


Figure 1.5.3; Protein grouping example 2. Protein A is a distinct protein due to unique peptide 1. Proteins B and C both contain conflicted peptide 2 and resolved peptide 3. There is no independent evidence to support that one over the other is present in the sample, and so they form their own same-set protein group.

- If a protein contains peptides that are mapped to more than one protein, but none of those proteins is distinct, if the peptides have not been labelled as conflicted, they are labelled as resolved (Figure 1.5.4).



a.)



b.)

Figure 1.5.4; Protein grouping example 3. a.) The three proteins contain a combination of the same resolved peptides and so will form a resolved group. As protein A, has been identified by the most peptides, it will head the group, and proteins B and C will form a sub-set. b.) After applying parsimony rules; the evidence in the sample can be explained by a single protein (protein A), under the rules of parsimony, proteins B and C are discarded.

### iii. Peptide abundance summarisation methods

In label-free proteomics, varying ionisation efficiencies mean that, rather than giving an absolute value to the quantities of proteins in a sample, relative changes in levels of a protein across samples are measured and used to calculate fold-changes between conditions. To compare levels between proteins, rather than across samples, absolute quantification is required. This is possible in LC-MS, but internal standards are required to perform the calculations. Most commonly, calculations of protein concentrations are performed using the abundance values of the ion current of their respective peptide ions. One method is to simply sum the average abundance of all the identified peptides belonging to the protein. Another uses only the most abundant peptides, known as the Hi-N method, and was demonstrated by Silva *et al.* (2006). This method utilises the finding that the average MS signal response for the three most intense tryptic peptides is correlated with the absolute concentration of the protein in the sample.

## **1.6. Normalisation**

Once proteins and protein groups have been identified and quantified, a further step is required to accurately decipher proteins of biological interest from bias, systemic variation, and outliers. Abundances of the same protein from different runs may differ due to sample handling or ionisation fluctuations. Differences in environmental conditions, sample preparation, or instrument calibration of separately processed samples can introduce systemic bias. This variation can cause a change in signal between conditions obscuring signals of interest. Optimal normalisation methods allow the removal of undesirable systemic bias while preserving changes of abundances due to DE, and the success of subsequent analysis is dependent on the selection of appropriate normalisation (Park et al., 2003). A variety of methods for transforming the data or aligning different properties of the data, such as the probability distributions, the quantiles, or the medians are available. Parametric and non-parametric models used for microarray normalisation, such as central tendency, quantile, variance stabilising, and local and linear regression normalisations, have been applied to proteomics data due to the experiments' having some of the same underlying assumptions. One run is selected as the reference run, or samples can be spiked with internal standards of a known quantity. This allows the calculation of a normalising factor, which can be used to scale to other runs, adjusting the intensities to make them even across all runs. Devised by Dudoit et al. (2002), the effectiveness of normalisation methods between two samples can be visualised using MA-plots (Figure 1.6.1). 'A' is displayed on the x-axis and represents the average of log expression values across a sample. 'M' is shown on the y-axis, displaying the difference in log expression value between samples. Normalisation methods often rely on the assumption that most protein levels do not change across the conditions, meaning the average ratio of expression values between two conditions is one (log ratio value of zero). In perfect data, the MA plot would show a horizontal line at zero.

## Optimising the statistical pipeline for quantitative proteomics

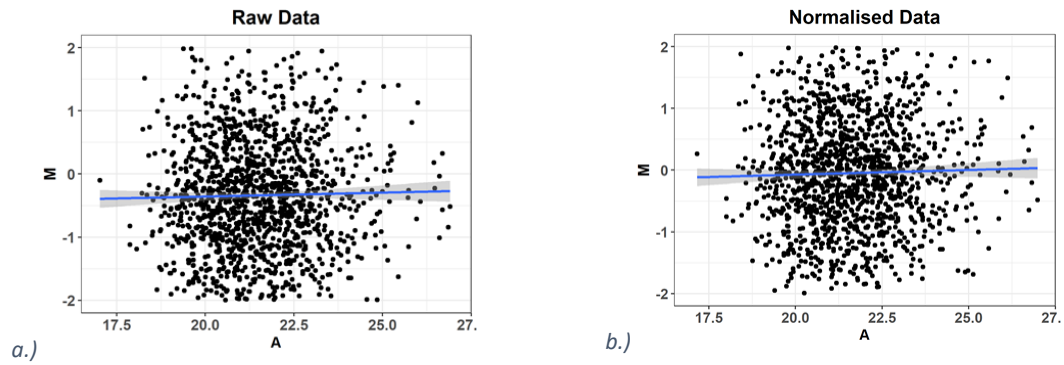


Figure 1.6.1; MA plots two samples of a.) raw and b.) normalised peptide abundances. 'A' on x-axis represents average log expression values, 'M' on y-axis shows difference in log expression values between samples. Linear modelling (shown in blue) shows a line more centred around  $y = 0$  in the normalised data.

A summary of different normalisation methods, the problems in the data that they address, and the software packages used to perform them, along with any issues, assumptions, and common usage, are shown in Table 1 and described in detail in this section.



## Optimising the statistical pipeline for quantitative proteomics

*Table 1; Summary of different normalisation methods with details of R packages used, assumptions, issues and uses.*

<b>Normalisation method</b>	<b>Problem addressed</b>	<b>Software</b>	<b>Assumptions</b>	<b>Issues</b>	<b>Example Uses</b>
Log transformation	Variance	log() in R	Data follows a log-normal distribution	Cannot transform zero values	Transforming data prior to statistical analysis
Central tendency normalisation	Bias	Progenesis QI	Correction factor is constant	Cannot remove noise from variation in ionisation	To correct variation due to sample loading
Linear regression with ordinary least squares	Bias	lm() in R	Correction factor is linearly dependent on intensity	Affected by outliers	Analysing highly abundant peptides
Robust linear regression with M-estimation	Bias	r1m() in MASS package of R	Correction factor is linearly dependent on intensity	Low intensity peptides often do not fit the assumption of linear dependency	Analysing long-tailed distributions
Local regression normalisation	Bias	lowess(), loess() in R	Correction factor is non-linearly dependent on intensity	Best performed on smaller datasets	Analysing peptides affected by ion suppression
Variance stabilisation normalisation	Variance and bias	justvs() in vsn package of R	Correction factor is non-linearly dependent on intensity	Can be slow	High density arrays
Quantile normalisation	Bias	normalize.quantiles() in R	Intensity distributions across samples are similar	All runs could be given identical intensity values for proteins	High density arrays

## **i. Variance normalisation**

### ***Log transformation***

Variation tends to increase as abundance increases, but not linearly, with variance being greater in low abundance peptides compared to high abundance peptides. To overcome this, abundance values are often first log transformed. Further normalisation steps are then performed to remove bias and adjust intensities so that they are all on the same scale and can be compared.

## **ii. Bias normalisation**

### ***Central tendency normalisation***

Based on the assumption that protein distributions should be similar, global adjustment (Yang et al., 2002) or central tendency normalisation forces the distribution to centre around a constant. This is useful to correct for errors between samples due to different amounts of samples being injected (Tokareva et al., 2021), resulting in measurements being offset by a constant factor. Abundance variability is adjusted by subtracting a constant value or multiplying by a scaling factor according to the general equation:

$$y'_i = \alpha_k y_i$$

Where

$y_i$  is the measured peptide abundance of the peptide ion  $i$  in sample  $k$   
 $\alpha_k$  is the scaling factor for sample  $k$   
 $y'_i$  is the normalised abundance of the peptide ion  $i$  in sample  $k$

Various metrics can be used for centring: Total intensity normalisation (TIN) where the mean log ratio is used for centring; Intensities are divided by the sum of intensities of all runs within the sample and multiplied by the median of all samples' sums of intensities then log<sub>2</sub> transformed, scaling samples so they have the same median. Average intensity normalisation (AIN) which focuses on centring the mean; Intensities are divided by the mean of intensities of all runs within the sample and multiplied by the mean of all samples' mean intensities then log<sub>2</sub> transformed. Median intensity normalisation (MIN) where the median

is used for centring; Intensities are divided by the median of intensities of all runs within the sample and multiplied by the mean of all samples' median intensities then log<sub>2</sub> transformed.

### ***Progenesis QIP normalisation***

In this procedure, inbuilt 'scalar normalisation' of Progenesis QIP, a different scalar multiple for each run is applied to the feature abundances to readjust the runs so that they are all measured on the same scale as a reference run. First, the median and median absolute deviation (MAD) are used to filter outliers and the upper and lower robust estimation limits are calculated using:

$$\begin{aligned} & \text{Median} + 3 \times (1.4826 \times \text{MAD}) \\ & \text{Median} - 3 \times (1.4826 \times \text{MAD}) \end{aligned}$$

Where

$1.4826 \times \text{MAD}$  is an estimate of the standard deviation

The peptides that fall within these boundaries are used to calculate the normalisation factor. An abundance ratio of each peptide is calculated compared to the same peptide from an automatically selected normalisation reference run, chosen by the software as being 'least different' to all the other runs: each sample is treated as a potential normalisation reference, and the robust standard deviation estimate (using equations above) is calculated for the other runs. A pooled variance is then calculated, which is a measure of how consistent the sample's difference is across all features from the others. A consistent scalar shift from another sample will introduce the minimum possible propagated error when scaled together. The sample with the lowest value is then used as the reference run. A ratio is then calculated of each peptide abundance to the normalisation reference run abundance value:

$$R_{i,x} = \frac{Ab_{i,x}}{Ab_{i,NR}}$$

$Ab_{i,x}$  is the abundance of the peptide ion  $i$  in the run  $x$

$Ab_{i,NR}$  is the abundance of the peptide ion  $i$  in the normalisation reference  $NR$ .

To correct for skewedness, a log transformation is applied to produce a normal distribution for ratio data of each run compared to the reference run. Ions with

an abundance of zero are not used in the calculation. Log10 ratios are centred on the normalisation reference by adding or subtracting the value required to shift the sample distribution over the reference run distribution, equivalent to multiplying untransformed data by a scale factor. Log space scalar estimation is returned to 'abundance-space ratio' and applied to all samples in the run. By using a ratio calculation compared to total-abundance-based methods, the influence of absolute abundance is removed.

### *Linear regression normalisation*

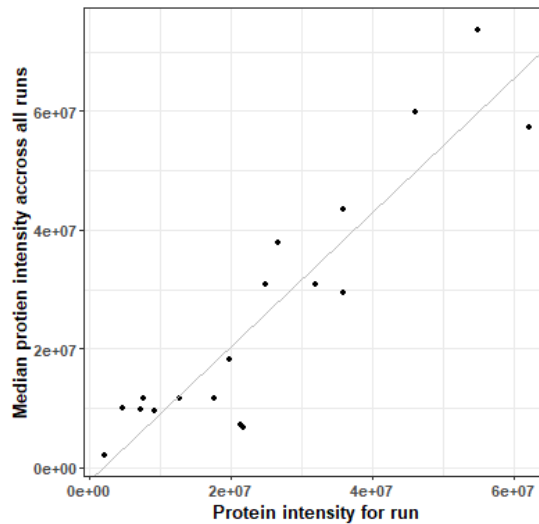


Figure 1.6.2; Protein intensities from one run are plotted against the median protein intensity across all runs and a linear model is fitted using ordinary least squares to provide the normalised protein intensities for that run, show in grey line.

Linear regression normalisation assumes amount of bias is linearly dependent on peptide intensity (Park et al., 2003); the more abundant the peptide, the greater the bias and as a result greater correction must be applied. Figure 1.6.2 shows linear regression normalisation where normalised intensity values for each are estimated by comparing the median protein intensity across all of the runs to the measured protein intensity in each run.

## Optimising the statistical pipeline for quantitative proteomics

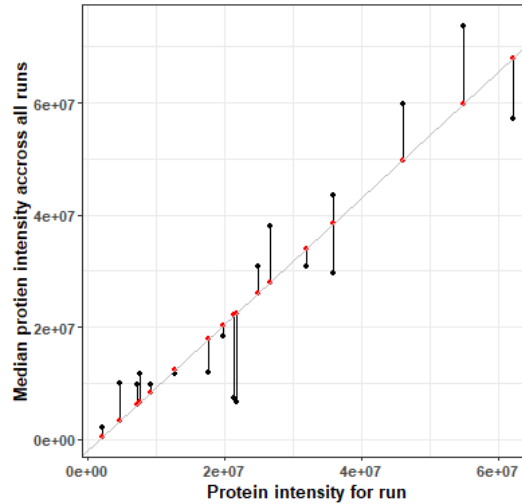


Figure 1.6.3; Residuals are the difference between the actual value, shown as a black dot, and the modelled value, shown as a red dot. The model that gives the lowest value when residuals are squared and summed is chosen as the best fit.

Linear modelling is fitted using ordinary least squares where the amount of error between the modelled value and the actual value, the residual (shown in Figure 1.6.3), is minimised. The residuals are squared to remove negative values and then summed. The model is fitted to the values that give the smallest total squared sum of residuals.

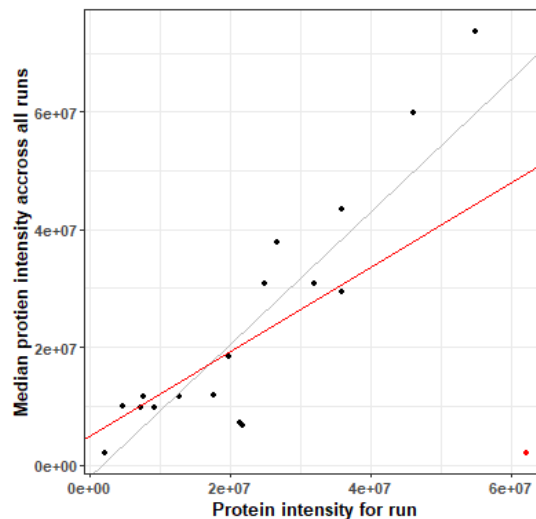


Figure 1.6.4; Linear regression normalisation where one data point, shown in red, has been replaced with an outlier. Original linear model is shown in grey and linear model including outlier is shown in red. Changes between models is created by changing a single data point.

Problems can arise in data with outliers; squaring large residuals causes these data values to have a lot of influence on the model. Figure 1.6.4 shows the extent the model can be altered by changing one data point.

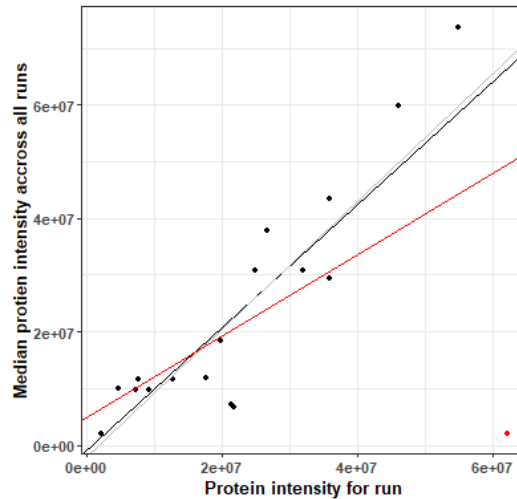


Figure 1.6.5; Data from fig 6 modelled with robust linear regression (dark grey line). This brings the fit closer to the original model before the outlier was introduced (shown in grey) compared to linear modelling with ordinary least squares (red line).

An alternative method to fit the linear model is robust regression, which is less affected by outliers. Using an M-estimator to weight observations based on the size of the residual, reduces the influence of these data points (Figure 1.6.5) and it can be a good option for analysis in long-tailed distributions. Each sample's log<sub>2</sub> transformed data is normalised to the median of all samples and using the 'rlm' function in the MASS package. Following this a linear model is fitted using robust regression through iterated re-weighted least squares (IRLS) (Venables and Ripley, 2002).

### ***Local regression normalisation***

In intensity dependent bias, the variance of samples increases with their mean abundance, and for high intensity proteins, the standard deviation roughly increases in a linear fashion. However, limitations of fitting a linear model arise with low-intensity peptides causing deviations from the straight line relationship and resulting in 'banana-shaped' plots. In these circumstances, global adjustment is not suitable. Local regression normalisation is a non-parametric regression method suitable for non-linear bias such as measured peptide abundances affected by ion suppression or detector saturation. Locally weighted scatter plot smoothing (LOWESS) (Figure 1.6.6) uses weighted least squares for local fitting of polynomials to smooth scatter plots (Cleveland,

1979). By performing linear regression on localised subsets of the data, a point-by-point function is built to describe the overall distribution of the data.

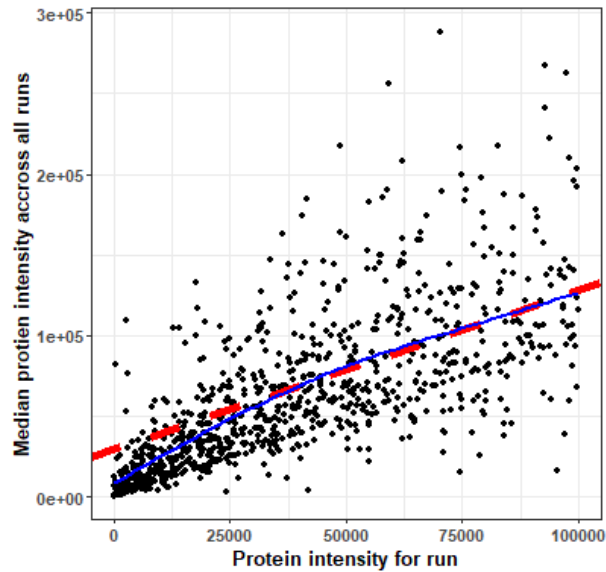


Figure 1.6.6; Protein intensity values below approximately 40000 deviate from being linearly correlated (red, dashed line). Blue line shows non-parametric regression using weighted least squares (LOWESS) which better fits data.

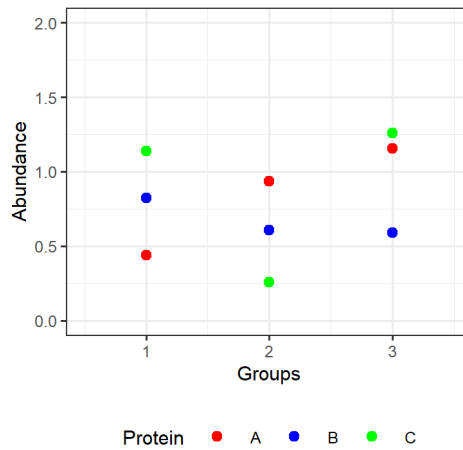
Local regression normalisation can be performed using the ‘normalizeCyclicLoess’ function from the limma package (Ritchie et al., 2015) for gene expression analysis in R, which encompasses cyclic loess described by Ballman et al. (2004). The method combines a smoothing function with a simple linear model. Cyclic loess regression performs pairwise analysis of MA plots between Log<sub>2</sub> transformed samples, correcting both sample intensities by the same factor in opposite directions at each individual points, iterations are repeated until the average ratio of expression converges to zero along the x axis, and each sample intensity is set to the average of all samples.

### **Quantile normalisation**

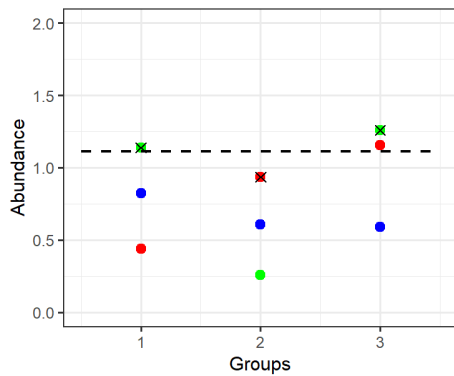
Quantile normalisation is a non-parametric approach developed for high-density arrays and uses the assumption that intensity distributions across samples will be similar and bias can be corrected by adjusting these distributions. First described by Bolstad et al. (2003) for microarray technology as a solution to variation introduced in experiments using multiple high density

## Optimising the statistical pipeline for quantitative proteomics

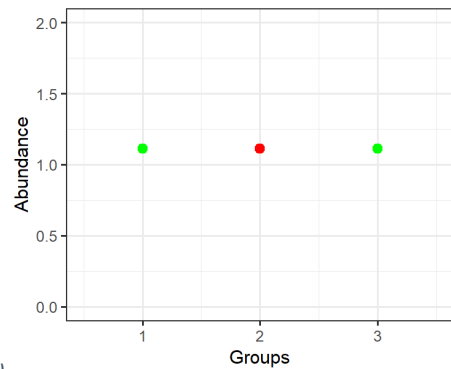
oligonucleotide arrays, its goal was to make the distribution of the probe intensities for all arrays within a set of arrays the same. Based on the concept that the distribution of two data vectors is the same if the quantile-quantile plot shows a diagonal line and extending the theory to the  $n$ th dimension for  $n$  data vectors. A worked example is shown in Figure 1.6.7. The data is first sorted and then the mean quantile intensity is substituted as the protein intensity value. Then the data is put back into the original order but now all runs will have the same distribution. A potential problem with this method is that it is possible all the runs could be given identical intensity values for proteins.



a.)



b.)



c.)



## Optimising the statistical pipeline for quantitative proteomics

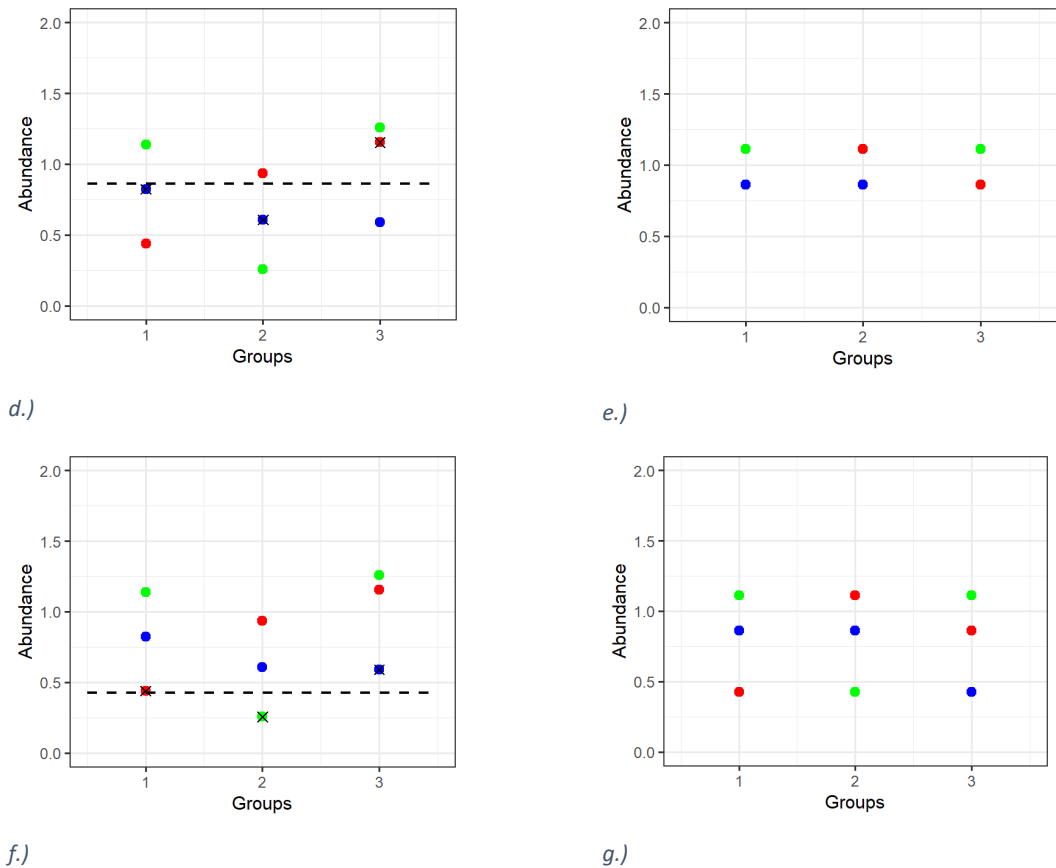


Figure 1.6.7; Worked example of quantile normalisation. a.) Mean protein abundance data of three proteins (A, B, and C) across three conditions (1, 2, and 3). b.) The mean value of the most highly abundant protein in each group is calculated. c.) Most abundant proteins are all assigned the mean abundance value. d.) The next most abundant proteins in each condition's mean is calculated and e.) each are assigned that mean abundance values. f.) Mean of least abundant proteins is calculated to be assigned. g.) Quantile normalised protein abundances. Note that the values across groups are the same, but the original rank order of protein is preserved.

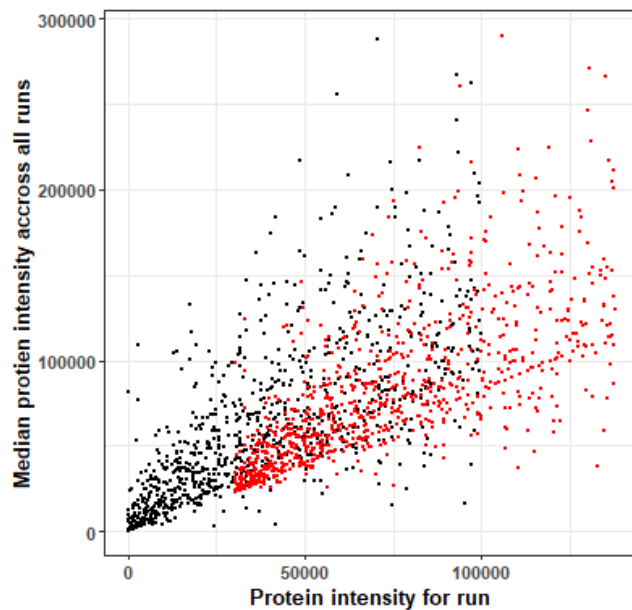
### QPROT normalisation

The inbuilt normalisation method in the QPROT package utilises a version of Boltstad's quantile normalisation where percentile points of the observed quantitative values were equalised across samples. The missing values are removed to form a trimmed set for each sample, and 0% to 100% percentile points are set to median value across the samples. Data points in between percentile points are interpolated to the nearest percentile point for each protein in each sample.

### **iii. Combined bias and variance normalisation**

#### ***Variance stabilisation normalisation (VSN)***

A quadratic variance-versus-mean-dependent model was derived by Rocke and Durbin (2001) and further developed as VSN by Huber et al. (2002) by transforming measured intensities so they become approximately independent of the mean. Through parametric transformation and maximum likelihood estimation, VSN produces intensities with variance independent of the mean. VSN uses the generalised log<sub>2</sub> (glog<sub>2</sub>) function, which provides a similar transformation to log<sub>2</sub> in the high abundance range, but it is less steep for intensities closer to zero. Differences between the transformed values give a generalised log-ratio that is used to estimate the parameters of the data with a robust variant of maximum likelihood estimation (Figure 1.6.8), providing both bias and variance normalisation.



*Figure 1.6.8; Variance stabilisation normalisation. Raw abundances (shown in black) and normalised abundances (shown in red).*

## **1.7. Differential expression analysis**

The aim of quantitative proteomics experiments is usually to discover proteins of interest that are changing due to the experimental conditions. Experimental design in proteomics can be complex involving time course events and multivariate factors. However, the main focus in this thesis is a pairwise comparison where statistical inference is used to define proteins that have changed in abundance between two or more experimental conditions. Due to similarities in data properties, small sample size, and large number of features, many software packages are based on algorithms originally created for the significance analysis of microarrays (Langley and Mayr, 2015). However, algorithms are often complex with many user-definable parameters, making software challenging to optimise (Gatto et al., 2016). This often leads to a default analysis using null-hypothesis significance testing (NHST) and a  $t$ -test.

### **i. Hypothesis testing**

The basis of frequentist statistical inference is hypothesis testing. A hypothesis is a claim or a premise about something you are trying to prove. Using these frequentist statistical approaches, we cannot prove something to be true, so we try to disprove it and accept an alternative. To do this, we form the null hypothesis,  $H_0$ , which is a description of the status quo. In proteomics, the null hypothesis is that there is no change in a single protein's abundances between conditions. We then form an alternative hypothesis,  $H_a$ ; that a protein is differentially expressed and is changing across experimental conditions. We can say that a protein is differentially expressed when we can prove the null hypothesis to be false (Guyatt et al., 1995). This process is then repeated for all confidently identified proteins. When we talk about proteins changing, we are looking at observing how a small sample of proteins behave in an experiment and using those observations to make predictions about how those proteins behave generally. This statistical process of using samples to make inferences about a population allows us to make measurements and describe otherwise unquantifiable data.

## ii. Statistical significance

The significance of a statistical test can be measured using  $p$ -values (Figure 1.7.1). A  $p$ -value is the probability that we would observe this result if the null hypothesis was true; a small  $p$ -value means that we would be unlikely to get this result if the protein was not changing across conditions (Dorey, 2010). Conventionally, a  $p$ -value of less than 0.05 is defined as significant and referred to as alpha. An alpha of 0.05 means the percentage of samples that would be more extreme is 5%, or the probability of seeing something this extreme if there is no difference in mean expression is 0.05. A significant  $p$ -value gives us evidence that the null hypothesis is incorrect and allows us to reject it and accept the alternative hypothesis, the protein is changing across conditions.

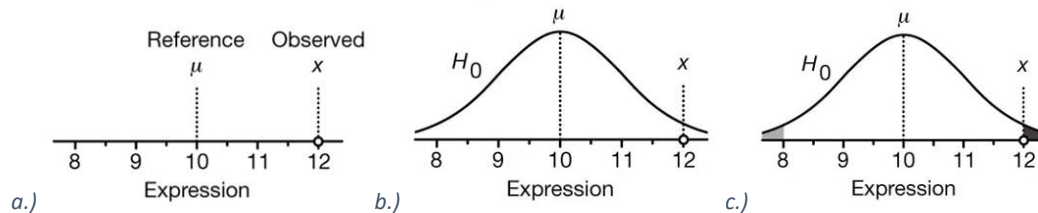


Figure 1.7.1; Graphical representation of the  $p$ -value calculation. a.) The observed experimental value ( $x$ ) is compared to the reference value ( $\mu$ ) b.)  $\mu$  is the mean of a null distribution,  $H_0$ . c.) The  $p$ -value is the percentage of values from the null distribution that are more extreme than the observed value (shaded in black and grey) (Krzywinski and Altman, 2013b)

## iii. $t$ -Test

The  $t$ -test is a calculation to compare the means of two groups to see if there is more difference than would be expected by chance. A  $t$ -test is flexible; it can handle paired or independent samples and can also be extended to apply to mixed samples. The  $t$ -test is a parametric test that assumes the data to be normally distributed; to ensure there is no skewedness, a logarithmic transformation is often performed. A specific type of  $t$ -test is the Welch's  $t$ -test (Welch, 1947), which allows a comparison of samples with different variances. This is often most appropriate for proteomic analysis as we cannot assume that conditions will have the same variance. The  $t$ -test can be one-tailed if expression is only expected in one direction, but usually we are looking at both

## Optimising the statistical pipeline for quantitative proteomics

up-regulation and down-regulation, so a two-tailed  $t$ -test is used. The equation for Welch's two-sample  $t$ -test is given by

$$t = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Where

$m_A$  and  $m_B$  are the sample means of the two groups A and B  
 $s_A^2$  and  $s_B^2$  are the sample variances of the two groups A and B  
 $n_A$  and  $n_B$  are the sample sizes of the two groups A and B

The  $t$ -distribution is defined by a parameter called the degrees of freedom, which is dependent on the sample size and the variance of the samples. The number of degrees of freedom estimates the error of using the sample variation rather than the population variation and is calculated using the equation:

$$v = \frac{\left(\frac{1}{n_1} + \frac{u}{n_2}\right)^2}{\frac{1}{n_1^2(n_1 - 1) + n_2^2(n_2 - 1)}}$$

Where

$$u = \frac{s_2^2}{s_1^2}$$

The unequal variance calculation of degrees of freedom is usually non-integer and rounded down to the nearest integer. Figure 1.7.2 shows how the shape of the  $t$ -distribution is affected by the sample size; increasing the sample size increases the degrees of freedom, which decreases the area under the tails of the distribution. The result of this is that a more extreme test statistic is required for significance when the sample size is small.

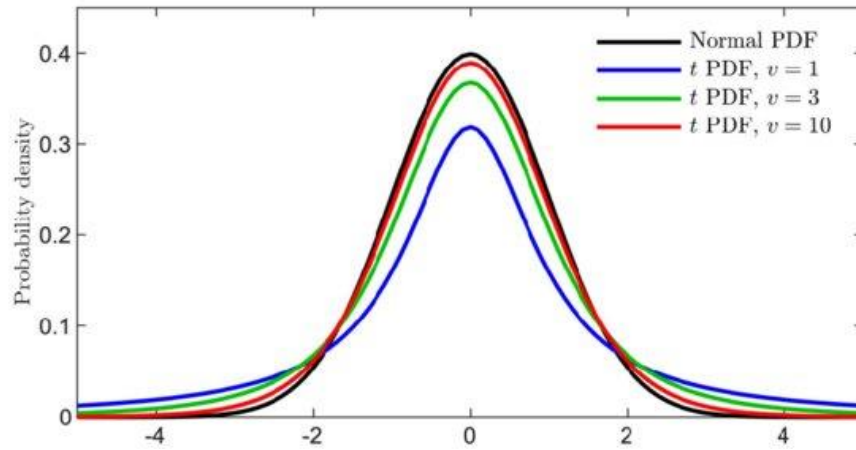


Figure 1.7.2; Effect of number of degrees of freedom on t-distribution; as degrees of freedom increases, the t-distribution approaches the normal distribution (Palkovic et al., 2020)

#### **iv. Multiple testing**

The success of statistical analysis is concerned with limiting two types of error: type 1 and type 2. A **false positive** is a significantly different result when there is no real difference in means. This is also referred to as a **type I error** in statistical inference and describes a situation where the null hypothesis is mistakenly rejected for the alternative hypothesis when the null hypothesis is true. As described above, the accepted cut-off  $p$ -value for significance is 0.05, which translates to a 5% risk of a false positive or type I error. A **type II error** is the situation where a protein that is changing is not detected and is also referred to as a **false negative**. Increasing the sample size of a test decreases the probability of this happening. The previous degrees of freedom calculation demonstrates how features with small differences need large sample sizes to have a good chance of being detected.

By definition, a  $p$ -value of 0.05 states that only 5% of the observations from the null hypothesis would be this extreme. Therefore, when defining a protein as changing across conditions with a  $p$ -value of 0.05, there is a 5% chance that this observation is from the null hypothesis and the protein is not actually changing across conditions. In a quantitative proteomics experiment, there will likely be hundreds of identified proteins in the sample and a DE analysis will be performed on each of them separately. Repeating a test with an accepted 5% chance of giving an incorrect result 500 times would give an expected number

of 25 incorrect results. Multiple testing errors are described as the **family-wise error rate** (FWER), the probability of committing at least one type 1 error when repeating statistical analysis on the same sample of data, and it is calculated using the equation

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^c$$

Where

$c$  is the number of comparison performed  
 $\alpha_{PC}$  is the per analysis alpha (usually 0.05)

Statistical methods can be applied to correct  $p$ -values when performing multiple tests. The Bonferroni method (Dunn, 1961) divides the per analysis alpha by the number of multiple tests, which puts the overall FWER is equal to 5%. The Bonferroni method is good at controlling type I errors, but it is also very conservative and will stringently limit the number of proteins called as differentially expressed, increasing the type II error rate.

## **v. The false discovery rate**

A preferred method for multiple testing correction is the Benjamini and Hochberg procedure, which is described by (Benjamini and Hochberg, 1995) and controls the expected average type I errors over the rejections made at level  $q$ .

- (i) Sort the  $p$ -values of the  $m$  hypotheses  $p_{(1)} \leq p_{(2)} \leq \dots p_{(k)} \leq \dots p_{(m)}$
- (ii) Calculate the largest  $k$  for which  $p_{(k)} \leq \frac{qm}{k}$  and reject the  $k$  hypotheses corresponding to  $p_{(1)}, \dots, p_{(k)}$ .

The correction is based on a procedure for estimating the  $q$ -values (Storey, 2002) and is described in Figure 1.7.3. If  $p$ -values were repeatedly generated by comparing two samples which come from the same distribution, proteins that are not changing in abundance, a histogram of the  $p$ -values would be uniform (Figure 1.7.3 a.) with approximately 5% of them falling below 0.05, representing false positives. When  $p$ -values are generated by comparing

## Optimising the statistical pipeline for quantitative proteomics

samples from different distributions, proteins that are changing in abundance, the histogram is skewed to the left with most of the  $p$ -values falling below 0.05 (Figure 1.7.3 b.). The  $p$ -values above 0.05 are false negatives from where the distributions overlap.

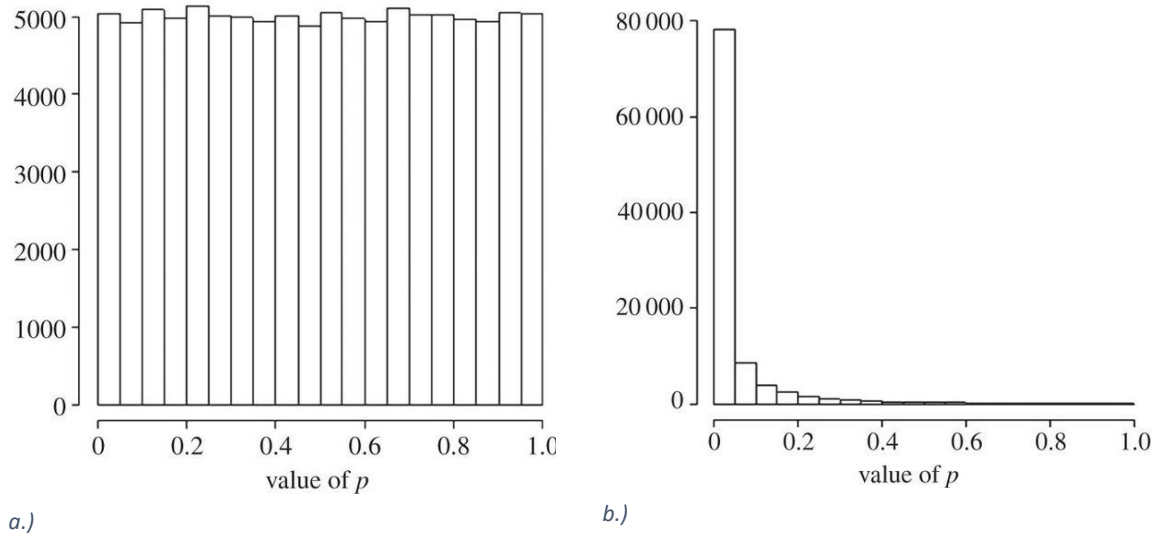


Figure 1.7.3; a.) A uniform distribution of 100 000 simulated  $p$ -values obtained by comparing two samples drawn from the same distribution. 5% of the values have a  $p$ -value less than 0.05 and are false positives. b.) A skewed distribution of 100 000 simulated  $p$ -values obtained by comparing two samples drawn from different distributions. 78% of them have a  $p$ -value less than 0.05 and are true positives (Colquhoun, 2014).

In a proteomics experiment, statistical analysis is performed to compare the abundance of each of the hundreds of proteins in the sample. Some proteins will be differentially expressed and affected by the experimental conditions, and some will be unaffected, their abundances not changing across conditions. A histogram of the  $p$ -values will be a combination of the two histograms; uniformly distributed  $p$ -values from the unchanging proteins and left-skewed  $p$ -values from proteins changing in abundance. The level at the uniform distribution can be used to estimate the proportion of  $p$ -values below 0.05 that are due to false positives and how many are due to changing protein abundance, true positives. By only using the smallest  $p$ -values to represent the number of true positives, the FWER is limited. The aim of proteomics DE analysis is to be as sensitive as possible, detecting all differentially expressed proteins, but also to limit the number of false positives. However, the nature of proteomics data can make this process difficult, which is discussed in the next section.



## **vi. Problems with null-hypothesis significance testing for proteomics data**

The problems of DE analysis through NHST have been discussed in literature; the  $t$ -test can produce significant  $p$ -values from only a very small difference in means (Tusher et al., 2001), and the  $t$ -test states that in the null hypothesis, the difference between the two sample means is exactly zero, meaning small, uninteresting fold-changes could be defined as very significant. It is also poor at estimating the variance when there is a departure from normality and there is variance heterogeneity between features and experimental conditions (De Hertogh et al., 2010). DE analysis has to decide if a protein is changing because it is affected by experimental conditions or if the abundance varies due to random chance. To do this accurately, the sample needs to be a true representation of population abundance. The  $t$ -test relies on normally distributed data, which, according to the central limit theorem (CLT), requires a large sampling distribution. However, the key assumption of a normal distribution cannot be made in proteomics experiments. Data matrices produced are sparse, with vast numbers of features and small numbers of samples. Small samples do not accurately reflect the population variation, and the comparison of means can be distorted by outliers (Krzywinski and Altman, 2013). In this situation, small samples taken from the same population may show different variances, causing an occasional artificially large value to be incorrectly identified as DE, a false positive. This is demonstrated in

Figure 1.7.4, where in two scenarios, a  $t$ -test compares the difference in means. Both scenarios show a significant difference. However, it is only the scenario where the biological variation is small that gives an accurate test statistic.

## Optimising the statistical pipeline for quantitative proteomics

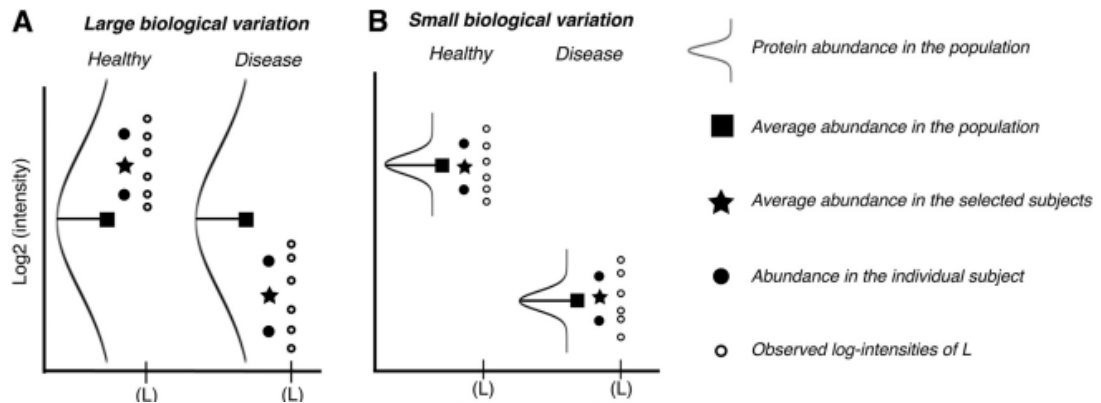


Figure 1.7.4; Two hypothetical distributions of protein abundances from two conditions, healthy and diseased tissue. There are two subjects (biological replicates) per condition, with three technical replicates per subject. The distribution and mean of the population abundances are represented by black curves and black squares. Observed abundances, mean subject abundances, and condition abundances are represented by black circles, solid dots, and stars, respectively. In plot A there is no difference in overall population abundances between conditions due to the large biological variation, while plot B shows a difference in population abundances between conditions and a small biological variation. (Chang et al., 2012).

Another problem is the use of the arbitrary  $p$ -value threshold,  $\alpha$ , as a measure of rejecting the null hypothesis; an  $\alpha$  of 0.05 leads to an artificial metric for defining significance and is a source of irreproducibility (Vyas et al., 2015). Introduced by Fisher (1925), the  $p$ -value was never intended to be used in the practice of arbitrary definition of significance. It has since been adopted as a convenience (Kennedy-Shaffer, 2019). The convention of using 0.05 as a dichotomous threshold for  $\alpha$  is subjective. The definition of a  $p$ -value gives no sharp distinction between significance and non-significance, only a gradual increase in evidence against the null hypothesis (Dahiru, 2008). The use of this boundary for significance means information can be lost and potentially interesting studies have been rejected due to results not being deemed significant (Berlin et al., 1989). The  $p$ -value is influenced by the effect size; a smaller difference in means produces a more significant  $p$ -value (Sullivan and Feinn, 2012), but a  $p$ -value alone gives no information about the effect size; statistical significance is no guarantee of clinical significance (Thiese et al., 2016). The problems arising from applying NHST to proteomics data has led to the application of alternative statistical methods for DE analysis.

## **vii. Linear modelling for differential expression**

Moving away from the  $t$ -test to more generalised linear modelling structures allows the full use of the power of the experimental data (Brady et al., 2015). A linear mixed effects model for DE can include terms for fixed effects, which are constant across individuals, such as genotype and treatment. Fixed effects do not change, or they change constantly over time. Mixed-effects models also include terms for random effects, which are not under experimental control and vary. These effects are drawn at random from a larger population of possibilities, such as the patient effect. Using a mixed-effects model means we can incorporate information about the experimental design as factors and interactions to express the different sources of variation (Daly et al., 2008).

Linear regression is used to test the linearity of the relationship between the response variable and the predictor variable. It can be employed using categorical predictors for performing DE analysis, where the  $t$ -test is represented by the linear model:

$$y_i = \beta_0 + \beta_1 x_i$$

Where

- $x_i$  is an indicator (0 or 1) of the comparison group
- $\beta_0$  is the group 1 mean
- $\beta_0 + \beta_1$  is the group 2 mean

and the null ( $\mathcal{H}_0$ ) and alternative ( $\mathcal{H}_A$ ) hypothesis are:

$$\begin{aligned}\mathcal{H}_0: \beta_1 &= 0 \\ \mathcal{H}_A: \beta_1 &\neq 0\end{aligned}$$

when  $\Delta x = 1$ , the difference in the means is the slope. Based on the equation:

$$Slope = \frac{\Delta y}{\Delta x} = \frac{\Delta y}{1} = \Delta y$$

and if there is no significant difference between the means, then  $\beta_1 = 0$ . This allows us to visualise the data; data points from group one are at  $x = 0$  and data points from the second group are at  $x = 1$ , as shown in Figure 1.7.5.

## Optimising the statistical pipeline for quantitative proteomics

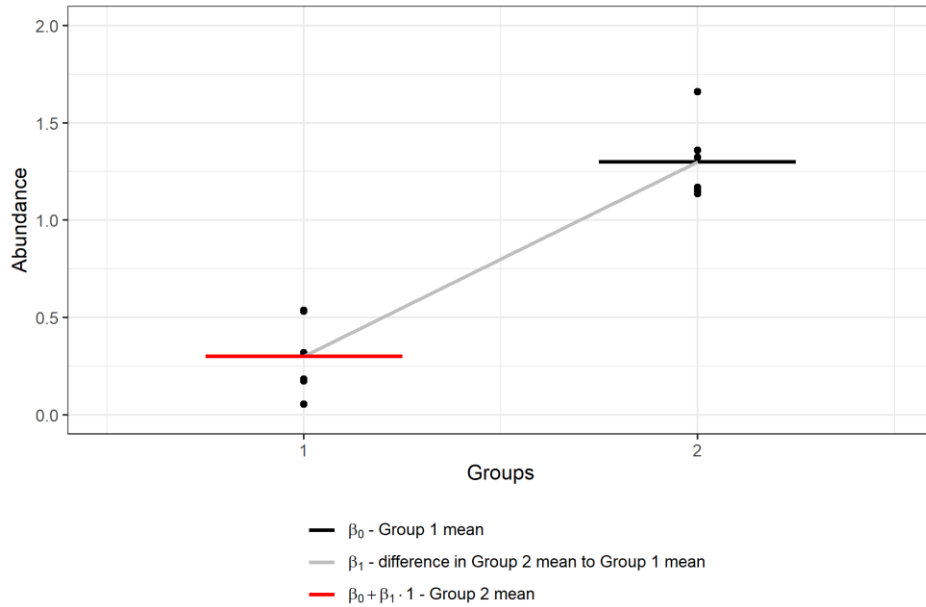


Figure 1.7.5; Two-sample t-test using linear modelling. The mean abundance of group 1 is shown as a red line, and the mean abundance of group 2 is shown as a black line. The slope shown in grey is used to calculate the difference between the two means.

The test statistic is calculated using the equation:

$$t = b / SE_b$$

where

$b$  = coefficient estimate

$SE_b$  = standard error of the coefficient estimate

Using the data from Figure 1.7.5, which is summarised in Table 2, the  $t$ -value is calculated as:

$$\begin{aligned} t &= 1 / 0.1155 \\ &= 8.660 \end{aligned}$$

and the corresponding  $p$ -value for  $t = 8.660$  with  $df = n - 2 = 10$  is  $5.84e-06$ .

Therefore, there is a significant difference between the means of groups 1 and 2.

Table 2; Summary of the linear model of the data from Figure 1.7.5

	Coefficients	Standard error	$t$ -value	$p$ -value
Intercept	-0.7	0.1826	-3.834	0.0033
Abundance	1	0.1155	8.660	5.84e-06

This can be expanded to apply to three or more means compared in a one-way ANOVA shown in Figure 1.7.6 and using the equation:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \dots \quad \mathcal{H}_0: \beta_1 = 0$$

Where

$x_i$  is an indicator ( $x = 0$  or  $x = 1$ ) based on a comparison matrix where one group is 1 while the others are 0

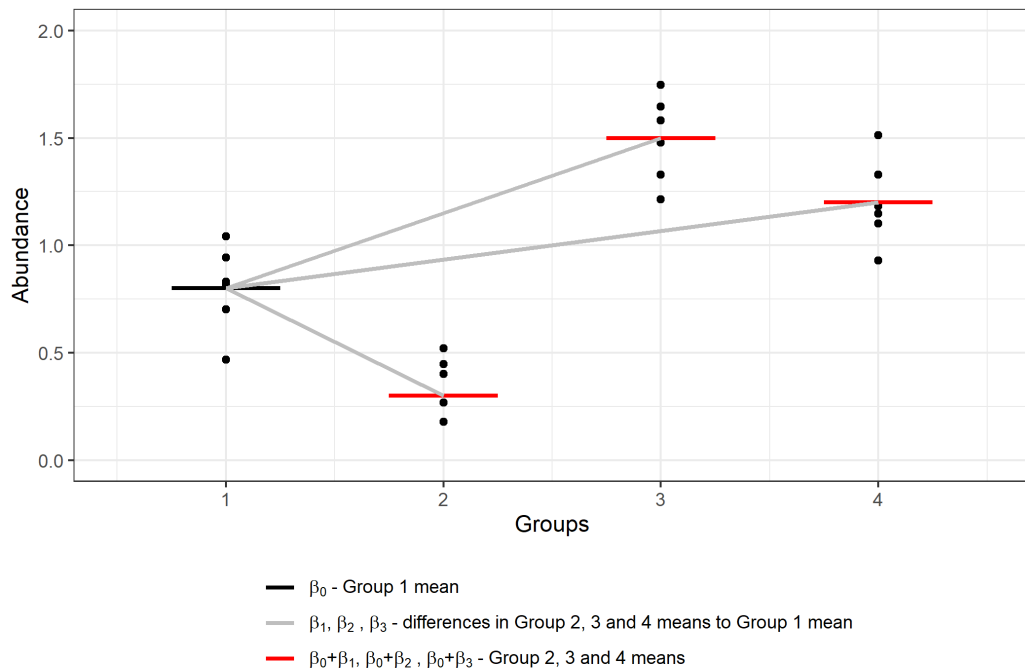


Figure 1.7.6; Four group one-way ANOVA using linear modelling. The intercept,  $\beta_0$  is Group1 mean and is shown in black. The means of the other groups ( $\beta_0 + \beta_1$  – Group 2 mean,  $\beta_0 + \beta_2$  – Group 3 mean,  $\beta_0 + \beta_3$  – Group 4 mean) are shown in red. The difference in means is calculated using the slopes  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  which are shown in grey.

### viii. Bayesian inference for differential expression

A problem with frequentist approaches, such as linear modelling, is that sample size is crucial to accuracy and consistency, as it is purely based on the observed data, which is assumed to be a representative sample (Alterovitz et al., 2007). Using a Bayesian approach to hypothesis testing allows the null hypothesis to be described as an interval rather than a single point, increasing the number of truly changing proteins detected (Millikin et al., 2020). Introducing a cut-off threshold for fold-change can improve the  $t$ -test approach, but there is still no

## Optimising the statistical pipeline for quantitative proteomics

measure of uncertainty; a protein with a fold-change either side of the cut-off is defined as changing or it is not, and no degree of uncertainty is included in the test statistic. Bayesian inference (BI) uses a mathematical rule to incorporate existing knowledge and observations to calculate the likelihood of DE in terms of probabilities (Kruschke, 2013). The process of BI uses Bayes' theorem to fit a probabilistic framework for reaching scientific conclusions based on how humans naturally make decisions (Baig, 2020), by applying a mathematical rule to incorporate existing knowledge and observation to update beliefs. The probability that an event will happen, the **posterior**, is scaled by the knowledge we have about the event, the **likelihood**, and our **prior** beliefs about how likely the event will happen (Figure 1.7.7).

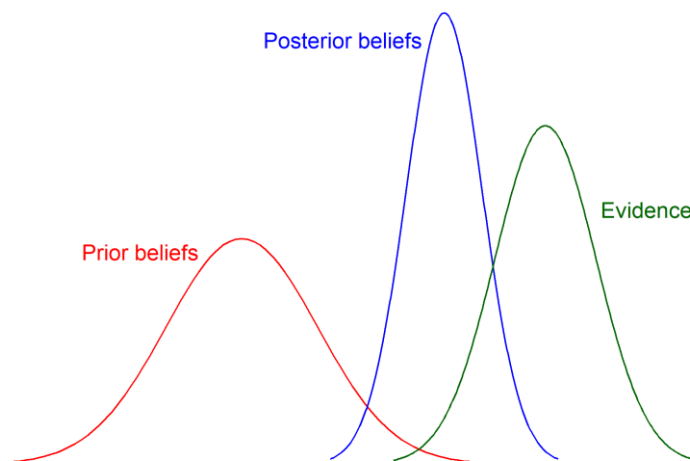


Figure 1.7.7; Probability distribution of prior beliefs (shown in red) is scaled by observing the evidence (shown in green) to give an updated probability distribution showing posterior beliefs in blue.

The equation for Bayes' theorem for two random variables, A and B:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Where

$P(A|B)$  is the probability of event A occurring given that B has occurred – the **posterior** belief updated after observing the evidence and scaling it with prior knowledge

$P(B|A)$  is the probability of event B occurring given that A has occurred – the **conditional likelihood** function

$P(A)$  is the unconditional probability of event A – the **prior** belief of how likely it is event A will occur

$P(B)$  is the unconditional probability of event B – the **marginal likelihood** function is a constant normalising factor equal to the integral of all possible values of A

## ***Optimising the statistical pipeline for quantitative proteomics***

Bayes theorem can also be written as:

$$posterior = \frac{conditional\ likelihood \times prior}{marginal\ likelihood}$$

Or by disregarding the normalisation constant:

$$posterior \propto conditional\ likelihood \times prior$$

This equation is the basis for Bayesian statistics and is used in tandem with statistical theory to quantify uncertainty. The result is a probability distribution where all possible values for the outcome are graded on how likely they are. Not all parameter values are always known and are approximated using parameter estimation into distributions. In terms of DE analysis, BI uses computational methodology to calculate conditional probabilities of the difference in sample means based on the observed values of protein intensities giving, complete information about the credible parameter values (Kruschke, 2013).

The first stage of analysis is to specify prior probability distributions for the unknown parameters based on available information. In DE analysis, unknown parameters are the variance of the population protein intensities, the mean population protein intensity, and the magnitude of DE. Then a model for summarising the observed data is constructed in the form of a likelihood function and is conditioned on the prior distributions. The updated knowledge is represented by the resulting posterior distribution and this is used to simulate data many times over to produce representative samples of the unknown parameters, called a ***maximum likelihood estimate***. In terms of DE, we calculate how likely different values for the mean intensities, amount of variance, and magnitude of DE are, given the protein intensities that have been observed.

## **1.8. Statistical software**

The current workflow in Progenesis QIP provides employs a comparison of means for DE analysis in the form of ANOVA. For pairwise comparison, this equates to a  $t$ -test. In this thesis, we investigate the use of linear modelling using a software package, MSstats and Bayesian approaches using QPROT software. MSstats software provides relative protein quantification and statistical DE analysis from peptide intensity data through a flexible family of mixed models. Published by Choi et al. (2014) from the Olga Vitek Lab, MSstats is based on statistical methods designed for quantitative measurement of gene expression (Lipshutz et al., 1999) and its implementation in the R package LIMMA (Smyth, 2005, Ritchie et al., 2015). MSstats was initially developed by Clough et al. (2009) who used fixed and mixed effects ANOVA for DE analysis. The framework was then extended to handle more complex designs by Clough et al. (2012) before being released as an open-source R package (R Core Team, 2020).

Following technological advances that allowed the computational demand required for calculating posterior distributions to be met, BI has been used for the development of several proteomics software packages. As with linear modelling methods, many Bayesian techniques also adapt algorithms originally implemented for identifying differentially expressed genes in microarray experiments from the package LIMMA (Smyth, 2005). Margolin et al. (2009) explored using an empirical Bayes framework for the analysis of SILAC experiments. Booth et al. (2011) developed a Bayesian mixture model for spectral count data. Choi et al. (2008) implemented QSPEC, a model for spectral count data, and extended its application to intensity data, QPROT (Choi et al., 2015). Koopmans et al. (2014) and Santra and Delatola (2016) investigated methods for dealing with missing data and using it to estimate prior probabilities for DE analysis. More recently, Millikin et al. (2020) implemented a version of the Bayesian  $t$ -test Kruschke (2013) for label-free data as FlashLFQ. In this thesis, we investigate the performance of the label-free package QPROT, which takes protein-level intensity data and applies a standardised Z-statistic



based on the posterior distribution of the log-fold change parameter as an alternative to the standard *t*-test.

### **i. MSstats package**

Appropriate for multiple types of sample preparations, the MSstats is suitable for analysing MS data acquired in data-dependent and data-independent modes, along with SRM and SWATH. There are three analysis steps: i) *data processing and visualisation*, which includes logarithmic transformation and normalisation between samples as well as a workflow-specific summary of the run-level data; ii) *statistical modelling and inference*, which includes DE analysis; and iii) *experimental design*, which uses the variance of the data to provide advice on improving the statistical power of the experiment in subsequent investigations. MSstats considers each peptide feature aggregated over peptide ions separately, with averages and variances of experiments across the conditions used to estimate model parameters for calculating protein intensity. Quantitative experimental information from all the features and all of the conditions that pertain to a protein is summarised which the authors claim provides higher sensitivity of protein significance analysis and higher accuracy of protein quantification.

The package takes as input identified and quantified peptide intensities with details of the proteins they are mapped to from upstream analysis packages. MSstats provides functions to convert the data to long format, where a list of peptides provides identification of the proteins they are mapped to and the peptide abundance. The peptide abundance appears in the list for every run it is detected in.

#### ***Data processing and visualisation***

Log<sub>2</sub> or log<sub>10</sub> transformation is applied, and there is an option to normalise the data using quantile or equalisation of medians (normalisation methods will be investigated in Chapter Three and are not addressed in this chapter). Run-level protein summarisation is provided with the ability to select how many features are used; all features in the data, which is the default selection, and 'top3' which

uses the three features that have the most abundant average log<sub>2</sub> intensities across runs. There is also an option called 'topN', which allows the user to input the number of most abundant features, and an option to filter poor quality features and outliers from protein quantification. Further discussed below, protein summarisation is performed with either Tukey's median polish (TMP) (Tukey, 1977), an estimation method that is resistant to effects of outliers, or through a family of linear mixed models.

### ***Linear modelling for summarisation***

The log<sub>2</sub> transformed data has statistical analysis performed separately for each unique protein name in the column of the dataset. Models of peptide abundance are fitted from all the peptides mapped to the protein across all runs using linear regression modelling by least squares. A worked example of protein quantitation from three peptides over three runs is shown in Figure 1.8.1 and Table 3. Linear modelling for fitting peptide quantities is performed according to the R function:

```
lm(Abundance ~ Peptide + Run)
```

The protein intensity across all peptides is calculated by summing the fitted peptide value and dividing it by the number of peptides in the run. If it is known that variance across the peptides is heterogeneous, there is an option for the software to account for this with an iteratively re-weighted least squares method where features with a large error are down-weighted to have less influence on the parameter estimation.

## Optimising the statistical pipeline for quantitative proteomics

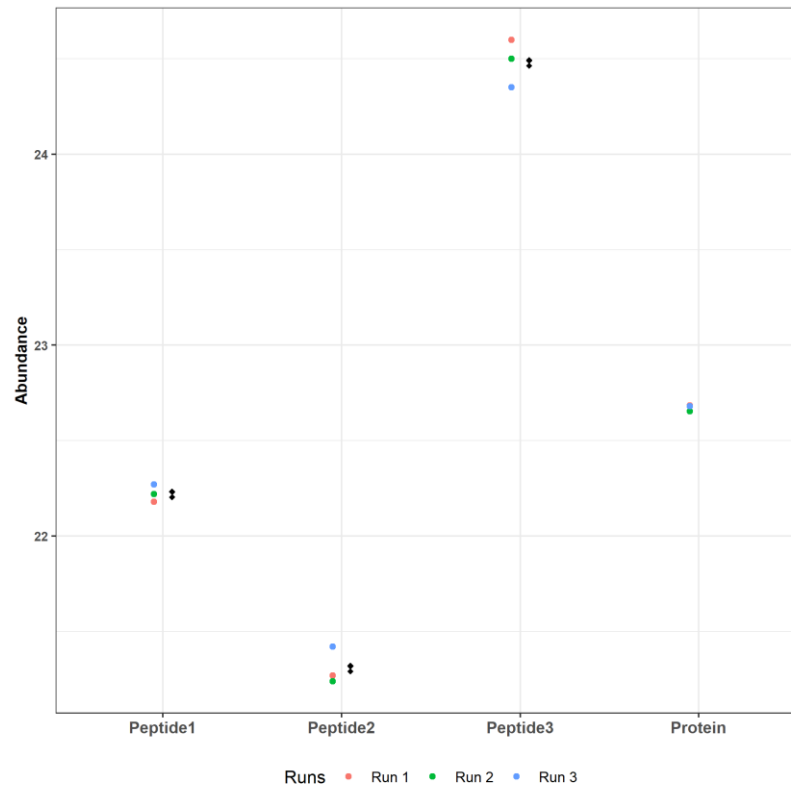


Figure 1.8.1; Worked example of protein summarisation through linear modelling. Observed log<sub>2</sub> abundances for three peptides mapped to a protein across three runs shown as colour to the left, fitted peptide abundances shown in black to the right. Resulting protein abundance for each run from the modelled peptide intensities is shown as colour.

Table 3; Worked example of protein quantitation by linear modelling. Log<sub>2</sub> transformed peptide abundances for 3 example peptides mapped to a protein across 3 runs, fitted peptide abundances for 3 peptides following linear modelling and resulting protein summarisation values.

	Log <sub>2</sub> transformed peptide abundances			Fitted peptide abundances following linear modelling			Protein summarisation values
	Peptide 1	Peptide 1	Peptide 2	Peptide 3	Peptide 2	Peptide 3	
Run 1	22.18	22.18	21.27	24.6	21.27	24.6	22.68333
Run 2	22.22	22.22	21.24	24.5	21.24	24.5	22.65333
Run 3	22.27	22.27	21.42	24.35	21.42	24.35	22.68

### Tukey's median polish for summarisation

TMP uses the `medpolish()` R package which fits an additive model to all peptide intensities mapped to a protein to give robust decomposition of the row effect of different peptides mapped to the protein. The long formatted, untransformed input data is reshaped to give peptide intensities in columns and run numbers as rows to produce a two-way table to explore the variables across row factor

## *Optimising the statistical pipeline for quantitative proteomics*

categories: peptides and runs. Row medians and column medians are removed iteratively until the proportional reduction in the sum of absolute residuals is less than 0.1. A simple worked example of a protein with 3 mapped peptides across 3 runs (Table 4), with runs as row name identifiers is shown in and Figure 1.8.2. to Figure 1.8.8. Using median values produces a more robust estimation than using means that are sensitive to outliers. This method does not take into account whether the runs are technical or biological replicates.

*Table 4; Protein summarisation based on additive modelling using TMP.*

	Peptide 1	Peptide 2	Peptide 3	Protein
Run 1	22.18	21.27	24.60	22.18
Run 2	22.22	21.24	24.50	22.22
Run 3	22.27	21.42	24.35	22.27

<b>22.22</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>
<b>0.0</b>	<b>-0.04</b>	<b>-0.95</b>	<b>2.38</b>
<b>0.0</b>	<b>0.00</b>	<b>-0.98</b>	<b>2.28</b>
<b>0.0</b>	<b>0.05</b>	<b>-0.8</b>	<b>2.13</b>

*Figure 1.8.2; The overall median for all values in the dataset is calculated as 22.22 and is assigned to the common effect cell in green. A residual table is created by calculating the difference between the original value and the median. Row (shown in orange) and column (shown in blue) are effect values and are initially set to zero.*

<b>22.22</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	→	<b>0.0</b>
<b>0.0</b>	<b>-0.04</b>	<b>-0.95</b>	<b>2.38</b>	→	<b>-0.04</b>
<b>0.0</b>	<b>0.00</b>	<b>-0.98</b>	<b>2.28</b>	→	<b>0.00</b>
<b>0.0</b>	<b>0.05</b>	<b>-1.02</b>	<b>2.13</b>	→	<b>0.05</b>

*Figure 1.8.3; The row medians of the residual table are computed and are shown in red.*

<b>22.22</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>
<b>-0.04</b>	<b>0.00</b>	<b>-0.91</b>	<b>2.34</b>
<b>0.00</b>	<b>0.00</b>	<b>-0.98</b>	<b>2.28</b>
<b>0.05</b>	<b>0.00</b>	<b>-1.07</b>	<b>2.08</b>

*Figure 1.8.4; A second residual table is created with row medians assigned to the row effects margin on the left (shown in blue), and the values are a subtraction of the row median from the value in the first residual table.*

## Optimising the statistical pipeline for quantitative proteomics

<b>22.22</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>
<b>-0.04</b>	<b>0.00</b>	<b>-0.91</b>	<b>2.42</b>
<b>0.00</b>	<b>0.00</b>	<b>-0.98</b>	<b>2.28</b>
<b>0.05</b>	<b>0.00</b>	<b>-0.85</b>	<b>2.08</b>
↓	↓	↓	↓
<b>0.00</b>	<b>0.00</b>	<b>-0.91</b>	<b>2.28</b>

Figure 1.8.5; The column medians of the second residual table are calculated and are shown in red.

<b>22.22</b>	<b>0.0</b>	<b>-0.91</b>	<b>2.28</b>
<b>-0.04</b>	<b>0.00</b>	<b>0.00</b>	<b>0.14</b>
<b>0.00</b>	<b>0.00</b>	<b>-0.07</b>	<b>0.00</b>
<b>0.05</b>	<b>0.00</b>	<b>0.06</b>	<b>-0.20</b>

Figure 1.8.6; A third residual table is created with column medians assigned to the column effects margin shown in orange, and values are a subtraction of the column median from the value in the second residual table. The column effect median is added to the row effect margin (shown in blue) and the overall common effect cell (shown in green), but as the column effect median is zero, the value remains unchanged. This is the end of the first iteration.

<b>22.22</b>	<b>0.0</b>	<b>-0.91</b>	<b>2.28</b>	→	<b>0.00</b>		<b>22.22</b>	<b>0.0</b>	<b>-0.91</b>	<b>2.28</b>
<b>-0.04</b>	<b>0.00</b>	<b>0.00</b>	<b>0.14</b>	→	<b>0.00</b>		<b>-0.04</b>	<b>0.00</b>	<b>0.00</b>	<b>0.14</b>
<b>0.00</b>	<b>0.00</b>	<b>-0.07</b>	<b>0.00</b>	→	<b>0.00</b>		<b>0.00</b>	<b>0.00</b>	<b>-0.07</b>	<b>0.00</b>
<b>0.05</b>	<b>0.00</b>	<b>0.06</b>	<b>-0.20</b>	→	<b>0.00</b>		<b>0.05</b>	<b>0.00</b>	<b>0.06</b>	<b>-0.20</b>

Figure 1.8.7; A second iteration; the row effect is calculated by computing the row medians (red) from the residuals and adding them to the common effect cell (green) and the row effects margin (blue), before subtracting the row medians (red) from the residuals to yield the values

<b>22.22</b>	<b>0.0</b>	<b>-0.91</b>	<b>2.28</b>		<b>22.22</b>	<b>0.0</b>	<b>-0.91</b>	<b>2.28</b>
<b>-0.04</b>	<b>0.00</b>	<b>0.00</b>	<b>0.14</b>		<b>-0.04</b>	<b>0.00</b>	<b>0.00</b>	<b>0.14</b>
<b>0.00</b>	<b>0.00</b>	<b>-0.07</b>	<b>0.00</b>		<b>0.00</b>	<b>0.00</b>	<b>-0.07</b>	<b>0.00</b>
<b>0.05</b>	<b>0.00</b>	<b>0.06</b>	<b>-0.20</b>		<b>0.05</b>	<b>0.00</b>	<b>0.06</b>	<b>-0.20</b>
↓	↓	↓	↓					
<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>					

Figure 1.8.8; The second iteration continued; the column effects are calculated by computing the column medians (red) from the residuals and adding them to the common effect cell (green) and the column effects margin (orange), before subtracting the column medians (red) from the residuals to yield the values. This ends the second iteration. The proportional reduction in the sum of absolute residuals is zero on the second iteration, and so the analysis is ended. The overall common effect value is added to the row effects values to give the fitted protein intensity level in each run.

## Optimising the statistical pipeline for quantitative proteomics

The protein abundances resulting from linear modelling and additive modelling are compared in Figure 1.8.9.

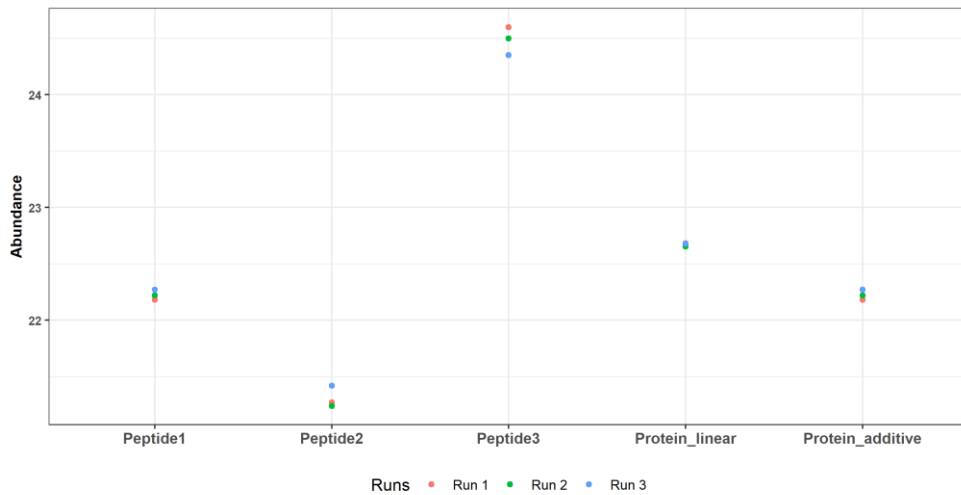


Figure 1.8.9; Comparison of the protein abundances from linear and additive modelling.

The output from the data processing and visualisation step is run-level protein abundance. The summarised data is exported as plots (Figure 1.8.10) to identify potential sources of variation and systemic bias by examining the peptide-level, run-level, and condition-level variation of each protein abundance.

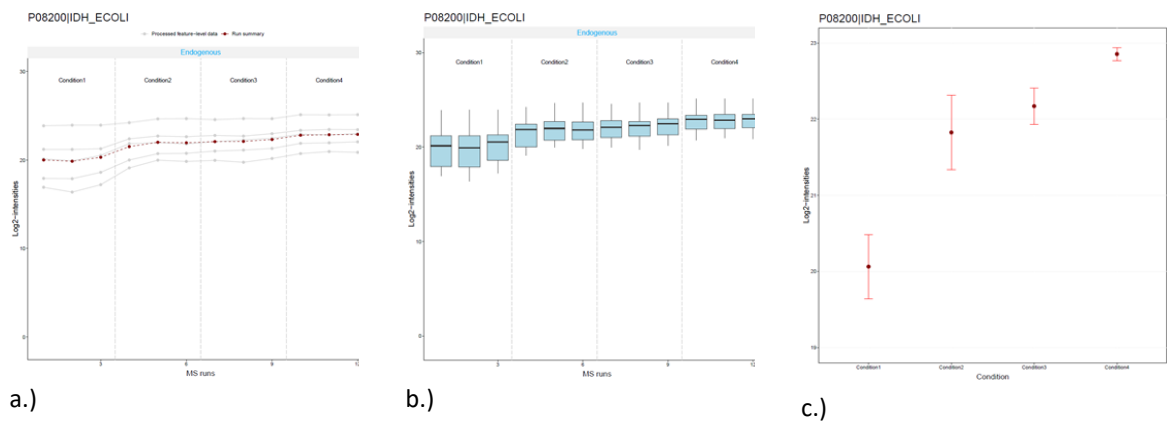


Figure 1.8.10; Example plots from the `MSstats dataProcessPlots()` method are given for each protein. a.) Profile plot with summarisation -  $\text{Log}_2$  peptide intensities of the five peptides mapped to protein P08200|IDH are shown in grey, with the summarised  $\text{log}_2$  protein intensity shown in dark red across four conditions and twelve runs. Used to identify potential sources of variation. b.) QC plot –  $\text{Log}_2$  intensities for across all peptides used for quantification are shown for each run. Used to evaluate systemic bias between MS runs. c.) Condition plot –  $\text{Log}_2$  intensities for all peptides used for quantification are shown for each condition. Used to the illustrate mean and variability of each condition per protein.

### ***Statistical modelling and inference***

The format of the input data allows MSstats to detect the experimental design and assign the appropriate model, which is fitted per protein. Samples from the same patient or technical replicates are subjected to the same variability. Therefore, in an experiment with only technical replicates and no biological replicates, a simple linear model is fitted to the run-level protein summarisations using the `lm()` function in R to give a one-way ANOVA analysis. The model coefficients, the scaled variance, and the degrees of freedom from the model are found, and these parameters are compared to the Student's *t* distribution to estimate the test statistics and the log fold change between each of the pairwise comparisons in the experiment.

For the analysis of biological data which will often have both technical and biological replicates in each condition. MSstats employs a mixed effects model using the `lmer()` function in R which includes all experimental factors that may affect DE. In terms of protein expression, fitting a mixed effects model accounts for the unequal measurability of peptides, which allows for a better estimation of the random variation giving a more precise significance value than the corresponding *t* value (Brady et al., 2015). Mixed effects linear modelling allows testing of multiple variables and their interactions:

$$y_{ijklm} = Prot_i + Pep_{ij} + Grp_{ik} + \varepsilon_{ijkl}$$

Where

$y_{ijklm}$  is the log<sub>2</sub> intensity for protein (*i*), and peptide (*j*), in comparison group (*k*), and sample (*l*),

$Prot_i$  is overall mean intensity for protein *i*

$Pep_{ij}$  represents the effect of peptide *j* in protein *i*

$Grp_{ik}$  represents the effect of group *k* in protein *i*

$\varepsilon_{ijkl}$  represents the random error

Variation in protein intensity is partitioned into contributions of the factors and an F-test is used to determine significance, which compares how the variation in the average phenotype between the genotypes in the experiment compares to the random variation in the experiment.

$$F = \frac{\text{variation between treatments}}{\text{variation within treatments}}$$

### ***Experimental design***

Following DE analysis, MSstats provides a further analysis phase which is intended to plan for future experiments. The analysis in this section is beyond the scope of this benchmarking exercise and was not included in the assessment. Briefly, the first two analysis steps, *data processing and visualisation* and *statistical modelling and inference*, are intended for early-stage screening experiments and the discovery of potentially changing proteins which require further investigation. Using biological and technical replication produces a higher sensitivity for detecting DE proteins, but statistical interpretations of the results must be restricted to the subjects in the study, *reduced scope*. To allow *expanded scope* and apply conclusions to underlying populations, the experiment must be modified to include minimal number of replicates to achieve pre-specified statistical power. This stage of analysis utilises variance components of the data to design future experiments.

#### **ii. QPROT package**

Developed by Choi et al. (2015) from the Alexey Nesvizhskii Lab, QPROT is a shell-based program for Linux operating systems. QPROT combines the use of false discovery rate (FDR) control with Bayesian modelling for the significance analysis of proteomics data. Initially presenting a statistical framework for spectral count data, Choi et al. (2008) addressed the issue of applying gene expression data models to proteomics data, which does not have the same standard distributional assumptions or the required number of samples for permutation-based generation of reference distributions. Because many proteomics experiments have small sample sizes, it is difficult to observe consistent evidence, making robust estimation and inference on model parameters difficult. Using Bayesian inference, unknown parameters are estimated with probability distributions; prior probability distributions are formed based on existing knowledge, data is observed, and the likelihood is calculated. The priors and likelihood are combined using Bayes theorem to produce the posterior distributions.



## *Optimising the statistical pipeline for quantitative proteomics*

QPROT calculates the likelihood function using hierarchical Bayes estimation of generalised linear mixed effects model (GLMM) (Zeger and Karim, 1991). The term ‘mixed effect’ refers to the combination of fixed effects and random effects within the model. Hierarchical models are structured so they are nested within themselves. In DE analysis, run intensities are nested within conditions; samples taken from the same condition are not considered independent and allow pooling of information across the replicates. This effect of outliers from small samples is minimised by borrowing strength across the runs and mapping from one distribution to another.

The statistical model, the Bayesian **conditional likelihood**, for QPROT is:

$$y_{ij} \sim N(\mu_i + d_i T_j, \sigma_i^2)$$

Where

$y_{ij}$  is the measurement of a protein  $i$  in a sample  $j$  for  $i = 1, 2, \dots, P$  and

$j = 1, 2, \dots, N$

$T_j$  is a binary indicator of the comparison group

$d_i$  is the magnitude of differential expression in log scale

$\sigma_i^2$  is the protein variance across all runs

The binary indicator  $T_j$ , is the treatment effect and is defined as  $T_j = 1$  if protein intensity is from the treatment group and  $T_j = 0$  if the protein intensity comes from the control group, and will result in two distributions whose means are separated by the magnitude of DE,  $d_i$ . This framework is hierarchical because the set of parameters used as the **priors**,  $\mu_i$ ,  $d_i$  and  $\sigma_i^2$ , for all proteins are specified as random variables from common distributions.

$$(\mu_i, d_i) \sim N(0, 10^2) \times N(0, 10^2)$$

$$\sigma_i^2 \sim IG(1, 0.1)$$

Where

$N(\cdot, \cdot)$  is the normal distribution with mean and standard deviation

$IG(\cdot, \cdot)$  is the inverse gamma distribution with shape and scale parameters

From the equation for Bayes theorem from section i:

$$posterior = \frac{conditional\ likelihood \times prior}{marginal\ likelihood}$$

The marginal likelihood is difficult to compute, and so the posterior distribution is inferred using a sampling algorithm and only the numerator of Bayes theorem.

$$posterior \propto conditional\ likelihood \times prior$$

### **Metropolis Hastings algorithm**

For inferring the posterior distribution, QPROT uses the Metropolis Hastings (MH) algorithm (Hastings, 1970), a standard Markov chain Monte Carlo sampler. A Markov chain is a process for generating sequences of random variables where the probability of the next variable is dependent only on the current variable. A Monte Carlo simulation is the generation of a series of random numbers from distributions, similar to the concept of rolling a dice. The MH algorithm is used to decide if the proposed value is likely to fit the posterior distribution and whether to accept or reject it as the next sequence in the Markov chain.

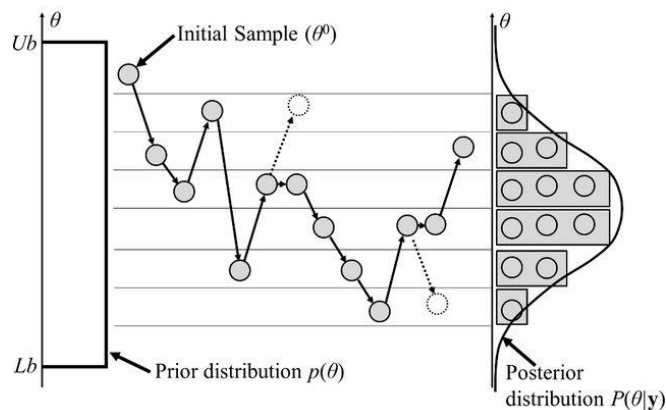


Figure 1.8.11; Illustration of Metropolis-Hastings algorithm (Lee et al., 2015).

Illustrated in Figure 1.8.11, the MH algorithm allows us to sample from the target distribution even when the normalising constant is not known. A proposed candidate distribution is calculated using the observed values of the

data and arbitrary starting values for the unknown parameters from the priors,  $\theta$  and a random sample is taken from this distribution and is used for the next value of  $\theta$ . Probability density functions for each of the values of  $\theta$  are calculated and then a ratio of the density functions are compared to a randomly sampled value  $u$  from the uniform distribution between 0 and 1. If the ratio is larger than  $u$ , the new value of  $\theta$  is accepted and its distribution becomes the new proposed distribution (solid circle in Figure 1.8.11). If the ratio is smaller than  $u$ , the value of  $\theta$  reverts to the previous value of  $\theta$  (dotted circle in Figure 1.8.11). By repeating this process a large number of times,  $N$ , the posterior distribution can be calculated by counting the number of samples at each interval. Because the starting number for  $\theta$  is random, the beginning of the chain will not exactly follow the distribution, and so a specified number of initial iterations are discarded. This is described as the 'burn-in period' which is pre-specified along with the value of  $N$ . Significance analysis for this package is based on the standardised  $Z$ -statistic where, similar to the  $t$ -test, the mean log fold change parameter,  $d_i$ , is normalised by the standard error of  $d_i$ .

The log fold change parameter  $d_i$  was recorded for every protein  $i$  denoted by  $(d_i^{(1)}, \dots, d_i^{(100,000)})$  and the mean  $d_i$  was the resulting log fold change of the protein  $i$ . The significance statistic of the DE of protein,  $Z_i$ , was calculated using the equation:

$$Z_i = \frac{\hat{d}_i}{\sqrt{\widehat{\text{var}}(\hat{d}_i)}}$$

Where

$\hat{d}_i$  is the mean log fold change parameter  
 $\sqrt{\widehat{\text{var}}(\hat{d}_i)}$  is the standard deviation of the log fold change parameter

### **FDR calculation**

QPROT employs a semi-parametric approach based on kernel estimators to estimate the local FDR. The *global FDR*, refers to the mean proportion of false positives in the data where the null hypothesis was rejected and describes the

## ***Optimising the statistical pipeline for quantitative proteomics***

data as a whole. The *local FDR* gives details of the reliability of specific hypothesis such as the probability that an individual protein is changing in abundance across conditions (Guedj et al., 2009). After calculating the significance statistics,  $(Z_1, Z_2, \dots, Z_P)$ , QPROT fits a semi-parametric mixture model in which Z-statistics follow a mixture based distribution depending on the unobserved status of the hypothesis. By using the known distribution of the Z-statistic under the null hypothesis ( $f_0$ ), a flexible non-parametric estimation of the alternative score distribution ( $f_1$ ), can be made using a weighted kernel function (Robin et al., 2007).

Let  $\pi_1$  = the proportion of DE proteins  
 $\pi_0 = 1 - \pi_1$  = the proportion of non-DE proteins  
 and  
 $f_1(Z)$  = the density of Z for DE proteins  
 $f_0(Z)$  = the density of Z for non-DE proteins

The mixture model of the two populations is:

$$f(Z) = \pi_0 f_0(Z) + \pi_1 f_1(Z)$$

$f(Z)$  is the overall density of Z-scores and  $f_0(Z)$  is the normal distribution of the null hypothesis. Therefore, for the values of  $i = 1, 2, \dots, P$  we can estimate:

$$\begin{aligned} f(Z_i) &= \pi f_1(Z_i) + (1 - \pi) f_0(Z_i) \\ &= \pi f_1(Z_i) + (1 - \pi) \sum_{k=1}^K \gamma_k \varphi(Z_i; \eta_k, \tau_k^2) \end{aligned}$$

Where

$\varphi$  is the density of the normal distribution  
 $\gamma_k$  is the mixing proportion of the  $\kappa$ th-mixture component with mean and variance  $(\eta_k, \tau_k^2)$   
 $\pi$  is the mixing proportion of DE proteins  
 $K$  is the number of normal distributions consisting of the null distribution  $f_0$  (default value of 1)

As  $f_1$  is completely unspecified, it is estimated non-parametrically using a one-dimensional Gaussian kernel density estimation (KDE) (Silverman, 1986):

## Optimising the statistical pipeline for quantitative proteomics

$$\hat{f}_h(Z) = \frac{1}{hP} \sum_{i=1}^P \varphi\left(\frac{Z - Z_i}{h}\right)$$

Where

$\varphi()$  is the standard Gaussian density

$h$  is the bandwidth which is selected as  $2.12\hat{\tau}P^{-1/5}$  for smoothness of the curve

$\hat{\tau}$  is the standard deviation of the observed Z-statistics

A KDE helps understand underlying probability distribution of the data. Similar to the concept of creating bins and counting the values in the bin to create a histogram (Figure 1.8.12).

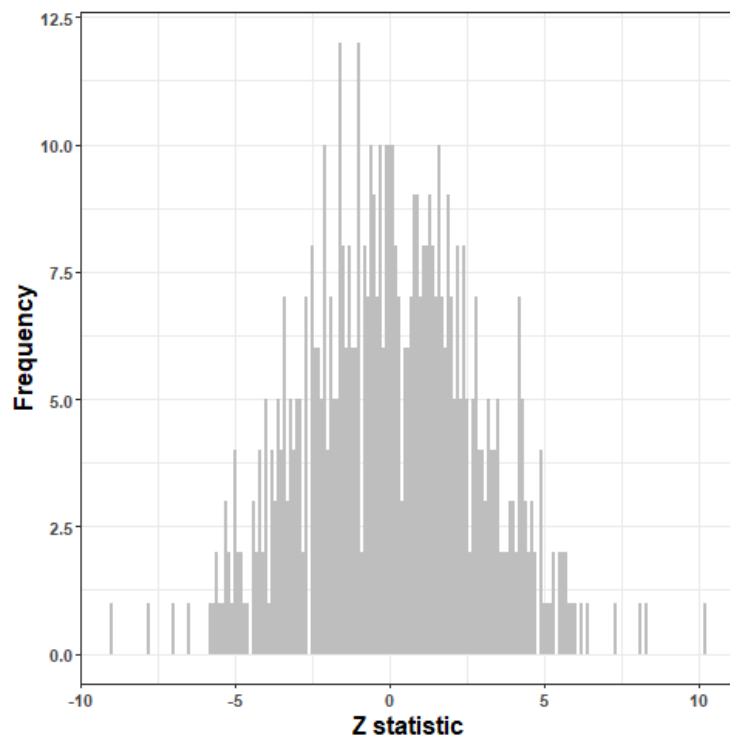


Figure 1.8.12; Histogram of Z-statistics from QPROT analysis of an example dataset

KDE forms a smoothed out continuous version. A kernel function is calculated for each Z statistic, and then the functions are summed to form a kernel density estimate that is normalised using the number of points P to ensure the total area under the distribution is one. An example is demonstrated in Figure 1.8.13 using Z-statistic from DE analysis using QPROT.

## Optimising the statistical pipeline for quantitative proteomics

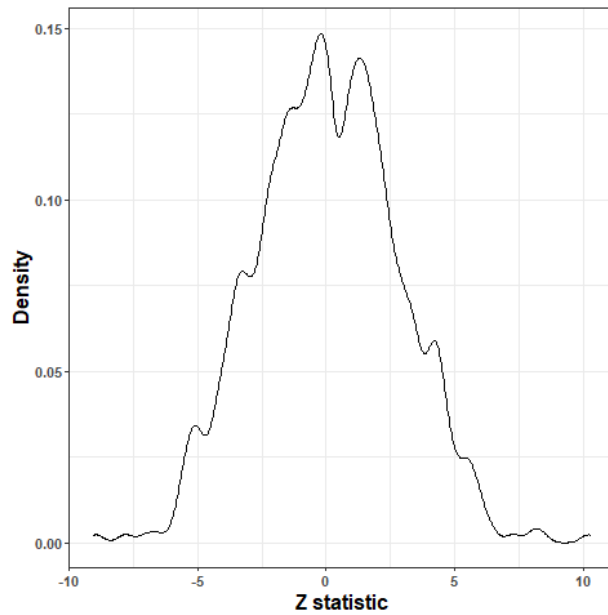


Figure 1.8.13; Overall density of Z-statistics from QPROT analysis of an example dataset.

The distribution of the Z-statistics under the null hypothesis are calculated  
Figure 1.8.14.

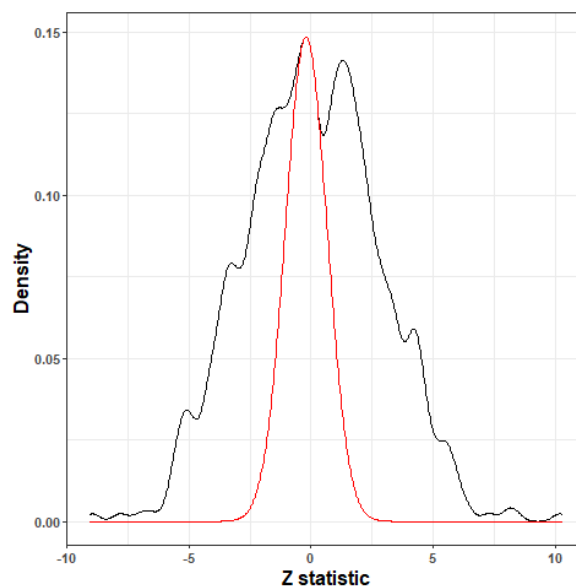


Figure 1.8.14; Overall density of Z-statistics (black) and known distribution of Z-score from the null hypothesis (red) from QPROT analysis of an example dataset.

From this, the distribution of the Z-scores from the alternative hypothesis can be estimated (Figure 1.8.15).

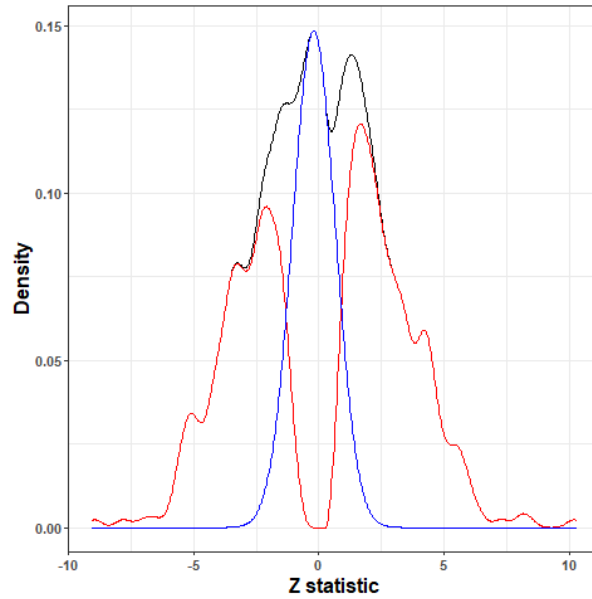


Figure 1.8.15; Overall density of Z-statistics (black) and known distribution of Z-score from the null hypothesis (blue) with estimated distribution of Z-score from the alternative hypothesis (red) from QPROT analysis of an example dataset.

The mixing proportion,  $\pi$ , is calculated using the Empirical Bayes method described by Efron et al. (2001):

$$\pi = 1 - \min_Z \{f(Z)/f_0(Z)\}$$

After model estimation, the FDR for each cut-off  $Z^*$  is calculated for the DE proteins in group 1 and vice versa for the DE proteins in group 0 using:

$$FDR(Z^*) = \frac{(1 - \pi) \int_{Z > Z^*} f_0(Z) dz}{\int_{Z > Z^*} f(Z) dz}$$

### **Low abundance proteins**

The QPROT analysis also includes a method for the treatment of missing data and applies truncation rules for integrating likelihood over the low abundance range. A truncation point,  $\phi_{i,T_j}$ , is set: the smallest abundance value for a protein within a condition. If there are zero values in the condition, the truncation point is the smallest non-zero value. If all values in the condition are zero, the truncation point is set to the 10 percentile point of all observed values in the other condition. The density is integrated over  $(-\infty, \phi_{i,T_j})$  according to the notation:

$$y_{ij} = \begin{cases} y_{ij}^{(o)} & \text{if } y_{ij} \geq \phi_{i,T_j} \\ y_{ij}^{(m)} & \text{if } y_{ij} < \phi_{i,T_j} \end{cases}$$

where  $y^{(o)}$  and  $y^{(m)}$  are the observed and missing values, respectively. The truncation point is then used to calculate the likelihood function for each protein, which is the probability of the observed values:

$$\begin{aligned} p(Y_i^{(o)} | \theta_i, \phi_{i0}, \phi_{i1}) &\propto \int p(Y_i^{(o)}, Y_i^{(m)} | \theta_i, \phi_{i0}, \phi_{i1}) dY_i^{(m)} \\ &= \prod_{j: y_{ij} \geq \phi_{i,T_j}} \varphi(y_{ij} | \theta_i) \times \prod_{j: y_{ij} < \phi_{i,T_j}} \int_{-\infty}^{\phi_{i,T_j}} \varphi(y_{ij} | \theta_i) dy_{ij} \end{aligned}$$

where  $\theta_i = (\mu_i, d_i, \sigma_i^2)$ , and  $Y_i^{(o)}$  is a vector of observed intensities for protein  $i$  and  $\varphi(X|a, b)$  indicates a normal density with a mean  $a$  and variance  $b$  evaluated at  $X$ .

## 1.9. Pathway analysis

Pathway analysis is a computational approach that allows us to apply prior biological knowledge to groups of genes or proteins to assess their biologically meaningful context. It is useful for interpreting the results of proteomics data by associating the results with simplified models of processes within cells or tissues. One of the most commonly used resources is the Gene Ontology, which provides a structured common classification scheme for gene function developed from collaborative experimental knowledge across organisms (Ashburner et al., 2000) (2019). The ontology consists of a set of hierarchical classes (referred to as terms) with relations operating between them. There are three sub-ontologies describing biological domains: *molecular functions* representing activities performed, *cellular components* describing the locations where functions are performed, and *biological processes* undertaken by multiple molecular activities. Enrichment analysis queries the GO for annotated terms, comparing a sample set of genes with a larger background set. Enrichment methods calculate an enrichment factor using:



$$\binom{a}{b} / \binom{A}{B}$$

Where

a = total number of DE proteins that map to a specific pathway

b = total number of DE proteins mapped to all pathways

A= all discovered proteins that map to a specific pathway

B= all proteins discovered that map to all pathways

A Fisher's exact test (Fisher, 1935) is used to calculate the probability that the value would occur if there was no relationship between the changing protein and the given pathway in the form of a *p*-value. Terms with small *p*-values represent the most significantly associated GO terms with the sample set, providing a functional profile of the sample genes and offering insight into the cellular mechanisms relevant to them (Khatri and Drăghici, 2005).

Pathway databases, where experimental evidence is curated and cross-referenced from literature and other databases and structured into 'knowledge databases' are used to query the functional enrichment of experimental results. One such is the Database for Annotation, Visualisation and Integrated Discovery (DAVID) (<https://david.ncifcrf.gov/home.jsp>) which provides a tool for significance analysis of gene-enrichment and functional annotation (Huang da et al., 2009a, Huang da et al., 2009b). Results of DAVID analysis also provide details of fold enrichment which describes how enriched a term is in the query protein list compared to the background list. If 20 out of 200 proteins in the query list are related to a certain activity (10%), and out of the 1000 background proteins only 50 are involved (5%), there would be a 2-fold enrichment of proteins related to the activity in the sample set. Enrichment analysis in DAVID allows the user to compare the annotation composition of the gene list against customised gene backgrounds. Default analysis is against the set genome-wide genes of the species in the gene list, creating a genome-wide scope. However, as many proteins involved in pathways such as glycolysis are often highly abundant and commonly identified, therefore comparing DE proteins against all proteins would produce functional enrichment profiles in almost every analysis, introducing experimental bias (Timmons et al., 2015).

Furthermore, using a background list of ‘all proteins’ from an ontology that is frequently updated means that replication of the experiment will be impossible. Instead, using a set list of all of the proteins identified in the experiment as the background set, a scenario is created where we can determine if proteins detected as DE are more functionally related to each other than to the rest of the proteins identified in the experiment.

## **1.10. Aims of project**

The high-throughput methods employed in label-free quantitative proteomics experiments produce large quantities of complex and noisy data that require sophisticated software for analysis. The properties of proteomics data complicate the ability to assess the significance of proteins changing in abundance due to experimental conditions. Small sample sizes provide poor estimation of variation and problems when comparing means. Large amounts of biological variation mean outliers can cause false positives, and repeated testing of the large number of proteins makes it difficult to limit the number of false positives while maintaining low false negative rates. The limitations of the commonly used *t*-test could make it unsuitable for the analysis of this data, and the imperfections of an arbitrary *p*-value for defining significance further confound this. There is also the added problem of complicated software workflows being challenging to intuitively obtain optimal results.

The main aim of this Industrial CASE PhD studentship, in collaboration with Nonlinear Dynamics, the developers of Progenesis QIP, is to provide an improved statistical pipeline that could be implemented in the Progenesis workflow. By benchmarking three existing statistical approaches: QPROT, ANOVA as implemented directly in Progenesis QIP, and MSstats traditionally, using spike-in datasets, and through the implementation of a novel method, using biological data and applying pathway analysis as an evaluation metric, we aim to develop and optimised statistical pipeline for quantitative proteomics.

## ***Chapter 2. Differential expression analysis evaluation using ground-truth data***

### **2.1. Introduction**

#### **i. Abstract**

The aim of this project is to develop optimised statistical pipelines for label-free quantitative proteomics, and in this chapter, through benchmarking existing packages with ‘ground truth’ data, we aimed to examine the best approach for DE analysis in a pairwise statistical design. In this context, ‘ground truth’ means artificially constructed samples with proteins spiked in, in assumed known different ratios, against a larger background of proteins with the same abundance, to test whether a software pipeline can detect as differentially expressed only the spiked in proteins.

Benchmarking software was developed to allow the consistent evaluation of individual stages of the bioinformatic pipeline: the minimum number of unique peptides required for quantitation, the selection of peptide ions for protein inference, protein quantification, and DE analysis. Three existing statistical approaches (QPROT, Welch’s *t*-test as implemented directly in Progenesis QI for Proteomics, and MSstats) were evaluated using benchmarking datasets. Known concentrations of spike-in proteins simulated DE at expected fold changes against a background proteome kept at a constant level. Smoothed Precision-Recall plots were used to visualise recall as a function of precision, using the area under the curve as an evaluation metric.

Both MSstats and QPROT gave a better performance than the *t*-test e.g. delivering higher recall for the same false discovery proportion, with QPROT giving the best performance in most of the fold-change scenarios. The results highlighted data quality issues, the importance of normalisation, and the selection of a cut-off threshold for significance. Problems with accuracy and precision of the spike-in data undermined the results of benchmarking. This led to the development of a novel benchmarking method that does not rely on the use of artificial spike-in data.

## ii. Benchmarking proteomics data analysis methods

### *Ground truth data*

Analysis and comparison of statistical methods for DE are conducted with artificial datasets that simulate real-life biological situations. Benchmarking datasets, represented in Figure 2.1.1, typically consist of a complex mix of background proteins that remain constant across conditions and spike-in proteins, introduced at different abundances that represent proteins of interest that have been expressed due to experimental conditions. This allows a 'ground truth' to be used as an evaluation metric to assess how well methods correctly detect the spike-in proteins without incorrectly claiming background proteins are changing.

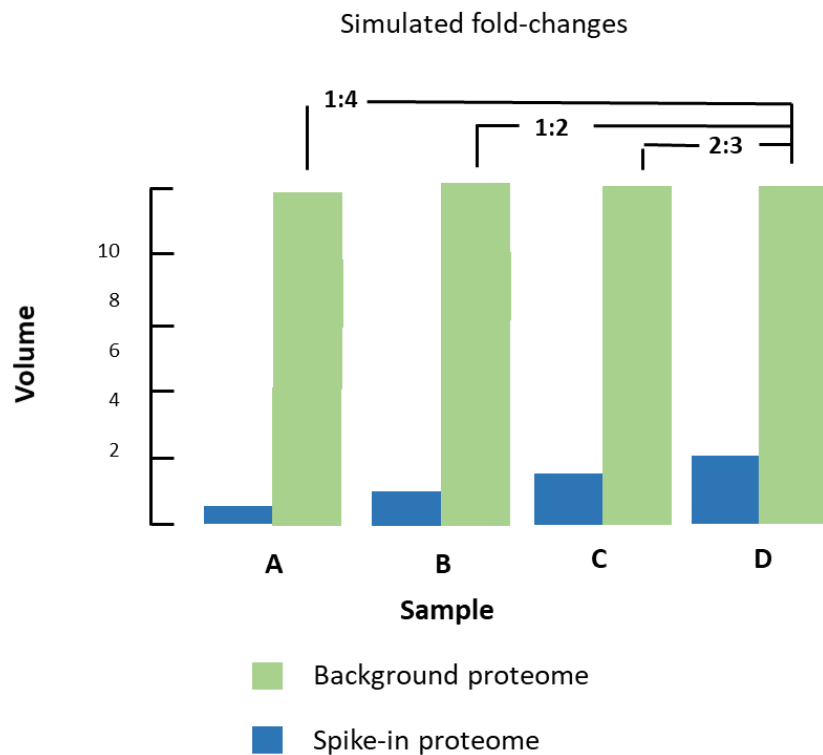


Figure 2.1.1; Representation of a benchmarking dataset. The background proteome level, shown in green, remains constant in each sample. Different concentrations of spike-in proteins, shown in blue, allow simulated fold-change with pairwise comparison.

### ***Benchmarking software***

The benchmarking process requires reproducible results for streamlined comparison. The proteomics workflow has many stages where data can be potentially treated in a different way. Software often provides different workflows, making it difficult to compare specific stages of the analysis. In their review of testing and validation of computational methods, Gatto et al. (2016) highlight being able to assess individual statistical methods rather than a software's whole workflow in order to make controlled comparisons. Yates et al. (2012) discuss objective evaluation of proteomics algorithms and highlight the need for uniform pathways. The need for accurate and consistent evaluation methods has led to the development of dedicated software to provide objective, unbiased comparisons. The issue of direct comparison when packages require or produce input and output data in different formats led to Kuharev et al. (2015) implementing a computational benchmarking framework for data-independent acquisition workflows. In order to create defined metrics for consistent evaluation, Navarro et al. (2016) developed an R package LFQbench for assessing DIA software, and due to the multifaceted nature of evaluation criteria, Tang et al. (2019a) created an online workflow assessment tool. To ensure consistent evaluation of all stages of our benchmarking exercise and to create customised file formats for the different software packages being assessed, benchmarking software was created to complete stages of the label-free workflow including options for the minimum number of unique peptides required for protein identification, the selection of peptide ions for protein inference, protein grouping, and protein quantification.

### ***Protein identification, inference, and grouping***

The 'shotgun' or 'bottom-up' approach to proteomics experiments involves peptide-centric MS analysis performed on the result of protein digestion. It has the advantages of simplicity and high throughput, as MS data on short peptides are easier to interpret than fragment spectra derived from larger molecules such as intact proteins. However, the comparison of protein abundances rather than peptide abundances is generally of biological interest (Nesvizhskii and

Aebersold, 2005), and a process is required to map peptide information back to the protein-level sample, i.e. *protein inference*. Chapter One discusses the technique for identifying peptides and mapping them to proteins through database sequence searches, where protein identification occurs through matching a subset of the peptide sequence fragments found in the sample. Often, database peptide sequences could have come from several proteins. The 'identified' protein list, which includes all possible mappings, can potentially have high number of ambiguous or incorrect identifications in large datasets (Keller et al., 2002). Chapter One also describes the rules employed by protein inference software to provide concise lists of reliably classified proteins and protein groups.

Consideration must also be given to the number of reliable peptides required as independent evidence for protein identification. Established practice is to only consider proteins confidently identified when they have more than one reliable peptides as evidence (Peng et al., 2003). This is the advice for showing evidence for a previously not seen protein issued by the Human Proteome Organisation (HUPO) (Hanash and Celis, 2002) in the most recent update of their MS Data Interpretation Guidelines 3.0 (Deutsch et al., 2019). The drawback of the two peptide rule is that a large number of potentially present proteins are eliminated from analysis. Gupta et al. (2007) performed a comparative proteogenomics study where MS data was used to map the sequenced genomes of three *Shewanella* bacteria species for improvement of gene predictions. They found that using the two-peptide rule decreased the number of genes confirmed as expressing proteins by over 20%. The et al. (2018) performed a benchmarking exercise for protein inference algorithms using a protein standard of known homologous content and found that the two-peptide rule performed poorly. Gupta and Pevzner (2009) evaluated the effect of the two-peptide rule and concluded that it reduces the number of protein identifications in the target database, more significantly than in the decoy database resulting in increased false discovery rates and estimates. The question over the two-peptide convention led us to include separate analysis for proteins identified by a minimum of both one and two peptides in this benchmarking exercise.

## *Optimising the statistical pipeline for quantitative proteomics*

Peptide quantification and identification by database searching of MS/MS data were discussed in Chapter One. The resulting peptide ion abundances are characterised by three attributes: its primary amino acid sequence, charge state, and if there are any chemical modifications to its amino acids. Peptide ions have different charge states depending on how many protons were added in the ionisation stage. Peptide modifications can occur spontaneously or physiologically as post-translational modifications, accidentally through artefacts in sample handling, or on purpose as part of the experiment. Artefactual modifications occur as part of the sample preparation, either deliberately, such as carbamidomethylation of cysteine, or accidentally, such as oxidation of methionine. Accidental modifications can occur to different degrees between samples. Therefore, how peptide ions with artefactual modifications are dealt with when calculating protein abundance may affect the overall evaluation of DE. Therefore, we assessed how ion variants with the same primary amino acid sequence should be selected for inference in protein quantitation analysis, as defined in Table 5.

*Table 5; Summary of options for how ion variants with the same primary peptide sequence are treated for peptide roll-up.*

Name	Intensities summed	Intensities treated separately
Summed charges	Ions with same sequence and different charge state	Ions with same sequence but with artefactual modifications
Summed modifications	Ions with same sequence regardless of artefactual modifications	Ions with same sequence and different charge state
Both summed	Ions with same sequence regardless of charge state or artefactual modifications	-
All separate	-	All peptide ion intensities kept separate

DE analysis in LC-MS relies on relative protein quantification where peptide level intensities are summarised to provide a comparative analysis across samples. One option is to use the summed intensity of all unique peptides mapped to a protein to establish the relative protein quantification here called '**sum-all**'. Originally worked up for Waters' MS<sup>e</sup> software for label-free ultra-performance liquid chromatography, Silva et al. (2006) detected that the average integrated signal intensity of the top three most intense tryptic peptides is proportional to the absolute amount of protein present in a sample.

## ***Optimising the statistical pipeline for quantitative proteomics***

The '**Hi-3**' method, uses the top three most abundant non-conflicted peptides mapped to the protein for quantification. If one of the peptides mapped to the protein is classed as conflicted, it will not be used for quantification and the next most abundant unique or resolved peptide will be used instead. Sticker et al. (2020) reported that this method removes information and introduces variability and bias. MaxLFQ (Cox et al., 2014), part of the MaxQuant software suite (Cox and Mann, 2008), uses ratio information from peptide signals to improve protein quantification accuracy and handle missing values. A matrix of pair-wise sample ratios is determined based on the two samples sharing the same two peptide ions belonging to a protein. These ratios are then used to calculate the protein intensity in that sample. If the sample does not share two common peptides with any other samples, its abundance value for that protein is set to zero. A further option in Progenesis QIP absolute quantitation makes use of previously discarded information in which the abundance of conflicted peptides is shared by the proteins to which they are mapped here called '**Progenesis Hi-3**'. The average peptide abundance is calculated across all runs. For each protein, the top three most intense peptides mapped to it are used to provide relative protein quantification across the conditions. If the peptide is unique or resolved, its intensity is summed. If the peptide is conflicted, its abundance value is shared between the proteins that it is mapped to in a ratio estimated by the abundance of their respective unique and resolved peptides. A worked example of Progenesis Hi-3 with is shown in Figure 2.1.2. Abundance values are simulated with integers for simplicity and the same peptides were used in each of the runs.



## Optimising the statistical pipeline for quantitative proteomics

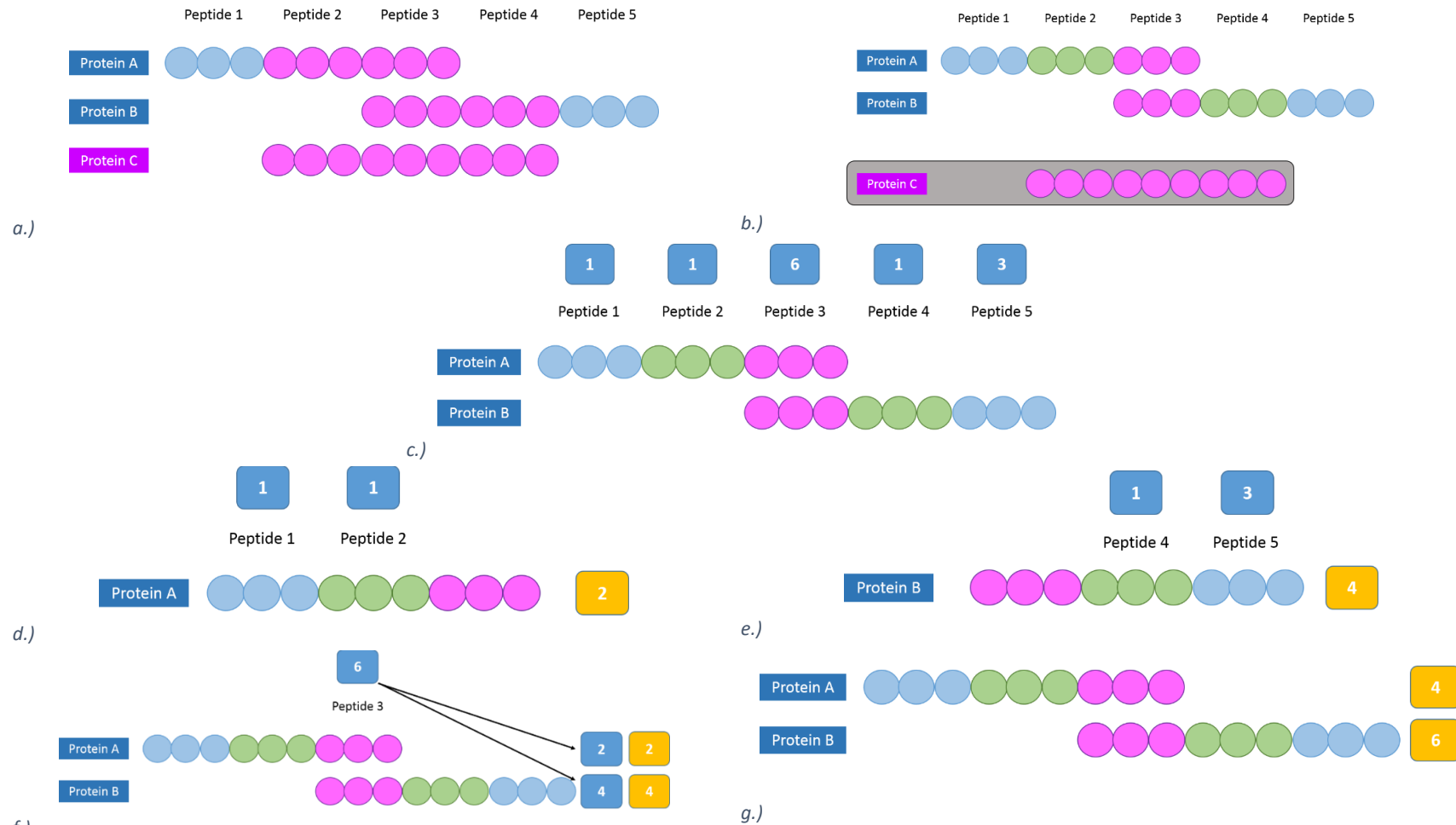


Figure 2.1.2; Worked example of protein quantification method utilising conflicted peptide information, abundance values are simulated with integers for simplicity. a.) Protein C holding only conflicted peptides is multiply subsumed. b.) Peptides 2 and 4 now become resolved and are suitable for quantitation. Peptide 3 remains conflicted as it is mapped to both protein A and protein B. c.) Peptide abundances shown in the blue boxes above the peptide identifications are used to estimate protein abundance using only unique and resolved peptides. d.) Protein A abundance (shown in yellow) is estimated using unique and resolved peptide abundances (shown in blue). e.) Protein B abundance (shown in yellow) is estimated using unique and resolved peptide abundances (shown in blue). As the ratio of estimated protein abundances was 1:2, peptide 3's abundance is shared between the proteins in this ratio. f.) The abundance for conflicted peptide 3 is shared in a ratio between the two proteins it is mapped to. g.) Final protein abundance calculation.

## ***Optimising the statistical pipeline for quantitative proteomics***

In this benchmarking exercise we compared results from the three different methods for peptide to protein inference summarised in Table 6.

*Table 6; Summary of methods of which peptides mapped to the protein are used for protein quantification*

Name	Type of peptides used	Number of peptides used	Method
Sum all	Non-conflicting	All	Abundance values summed
Hi-3	Non-conflicting	3 most abundant	Abundance values summed
Progenesis Hi-3	All	3 most abundant	Non-conflicting peptide abundance values and share of conflicting peptide abundances summed

### **iii. Aims of chapter**

Chapter One describes the conventional statistical approach to DE analysis and its drawbacks using a *t*-test for pairwise comparison. Research has focused on developing new ways to provide sensitive analysis, providing a minimal number of false positives while maintaining low false negative rates with the aim of providing optimal accuracy and precision. The aim of the thesis is to develop optimised statistical pipelines for label-free quantitative proteomics, and in this chapter, through benchmarking existing packages with ‘ground truth’ data, we aim to examine the best approach for DE analysis in a pairwise statistical design. Three datasets with constant background protein abundance and artificial DE proteins spiked-in at expected values were used. This simulates real experimental fold-change scenarios and is used to create a metric to assess how well methods correctly identify the simulated DE.

To allow for specific parameters to be evaluated in isolation and to produce customised output formats for the external packages, benchmarking software was developed to allow for the consistent evaluation of individual stages of the bioinformatic pipeline. Options included the minimum number of unique peptides required for protein identification, the selection of peptide ions for protein inference, protein grouping, and protein quantification.

Three different methods for DE analysis were compared; linear modelling, which is employed by MSstats, Bayesian inference techniques using QPROT, and

Welch's *t*-test as implemented directly in Progenesis QIP. Performance of the software was optimised by parameter exploration at the protein inference stage using benchmarking software, and evaluation was conducted using smoothed Precision-Recall plots.

## 2.2. Methods

### i. Spike-in datasets

The methods for DE analysis were evaluated using three benchmarking datasets, as described below. Raw files were downloaded via the PRIDE partner repository from the ProteomeXchange Consortium

(<http://proteomecentral.proteomexchange.org/>) (Vizcaíno et al., 2014).

#### *Escherichia coli* spiked into human background

##### *PXD001385 (Shalit et al., 2015)*

Prepared in three replicates, 3, 7.5, 10 and 15 ng of *E. coli* digest were added to separate 200 ng of HeLa S3 cell samples to simulate 5-, 2-, and 1.5-fold change scenarios (Figure 2.2.1).

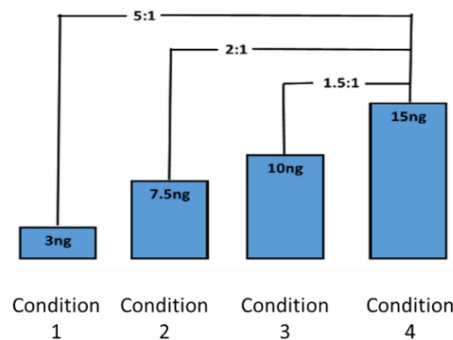


Figure 2.2.1; Representation of simulated fold-changes in benchmarking dataset PXD001819. 3, 7.5, 10 and 15 ng of *E. coli* digest was spiked-in to create the conditions of 5-, 2-, and 1.5-fold change.

#### *Human protein* spiked into a yeast background

##### *PXD001819 (Ramus et al., 2016)*

Created by Ramus *et al.* for their evaluation of bioinformatics pipelines to allow for assessment of performances in terms of sensitivity and false discovery rate, 48 human proteins (Sigma UPS1) were spiked in a background of yeast (*Saccharomyces cerevisiae*) cell lysate with concentrations of 50, 125, 250, and

## Optimising the statistical pipeline for quantitative proteomics

500 amol/ $\mu\text{g}$ , and 2.5, 5, 12.5, 25, and 50 fmol/ $\mu\text{g}$  of UPS1/ $\mu\text{g}$  of yeast lysate analysed in triplicate. The authors describe the use of the yeast proteome to provide a background with a relatively high complexity and dynamic range as a good surrogate for many real biological samples. The conditions of 5, 12.5, 25 and 50 fmol/ $\mu\text{g}$  were used to simulate 10-, 4-, and 2-fold changes for our analysis (Figure 2.2.2).

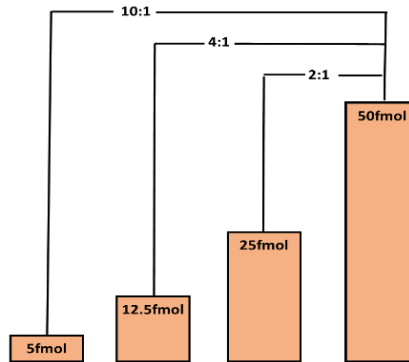


Figure 2.2.2; Representation of simulated fold-changes in benchmarking dataset PXD001819. 2.5, 5, 12.5, 25, and 50 fmol/ $\mu\text{g}$  of yeast lysate was spiked-in to create the conditions of 10-, 4-, and 2-fold change.

### Human protein spiked into a yeast background

#### PXD002099 (Pursiheimo *et al.*, 2015)

Data from the experiment created by Pursiheimo *et al.*, 2015 was used, where 48 human proteins (Sigma UPS1) mixed with yeast cell digest at concentrations of 2, 4, 25, and 50 fmol/ $\mu\text{L}$  provide triplicate replicates of 25-, 12.5-, and 2-fold changes (Figure 2.2.3).

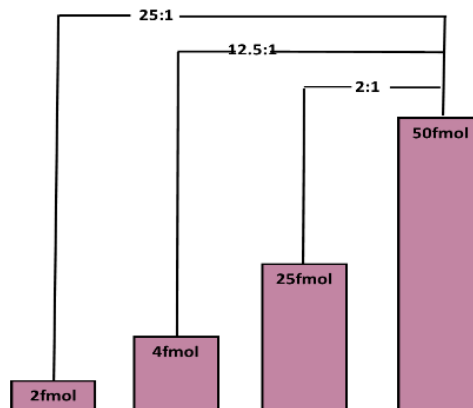


Figure 2.2.3; Representation of simulated fold-changes in benchmarking dataset PXD002099. 5, 12.5, 25 and 50 fmol/ $\mu\text{L}$  of 48 human proteins were spiked-in to create the conditions of 25-, 12.5-, and 2-fold change.

## Optimising the statistical pipeline for quantitative proteomics

Properties of the datasets, with details of acquisition and experimental design are summarised in Table 7.

Table 7; Summary of experimental conditions for each of the benchmarking datasets giving details of samples used, experimental design, parameters used for MS analysis, and number of proteins. Only the four largest spike-in conditions were used for benchmarking. Due to processing issues, the 10 fmol/ $\mu$ l spike-in condition was not used in dataset PXD002099. The number of spike-in proteins for PXD001385 was not provided in the paper and is based on this analysis.

	PXD001385	PXD001819	PXD002099
<b>Spike in protein</b>	Bacteria ( <i>E.coli</i> ) massPrep digestion standard (Waters)	Human (UPS1)	Human (UPS1)
<b>Background protein</b>	Human (HeLa S3 cells)	Yeast ( <i>S.cerevisiae</i> )	Yeast ( <i>S.cerevisiae</i> )
<b>Number of samples</b>	12	27	15
<b>Number of conditions</b>	4	9*	5
<b>Spike-in protein levels</b>	3, 7.5, 10, 15 ng	50, 125, 250, 500 amol/ $\mu$ g	2, 4, 10**, 25, 50 fmol/ $\mu$ l
<b>Background protein level</b>	200ng	2 $\mu$ g	100ng
<b>Instrument</b>	Q Exactive Plus Orbitrap	LTQ Orbitrap Velos	LTQ Orbitrap Velos
<b>Enzyme</b>	Trypsin	Trypsin	Trypsin
<b>Search database</b>	Swiss-Prot human and <i>E.coli</i>	UniprotKB yeast and UPS1	UniprotKB/Swiss-Prot yeast, UPS1 and cRAP
<b>Peptide tolerance</b>	10 ppm	5 ppm	5ppm
<b>MS/MS tolerance</b>	0.02 mmu	0.8 Da	0.5Da
<b>Fixed modifications</b>	Carbamidomethylation (C)	Carbamidomethylation (C)	Carbamidomethylation (C)
<b>Variable modifications</b>	Oxidation (M), carbamylation (N-term)	Acetyl (Protein N-term), oxidation (M)	Oxidation (M)
<b>Number of spike-in proteins</b>	234	48	48

\* Only the four largest spike-in conditions were used for benchmarking

\*\* This condition was not used due to processing issues

### ii. Benchmarking workflow

The spike-in data was analysed with a pipeline summarised in

Figure 2.2.4. Quantitative processing of the raw data was performed with Progenesis QI for Proteomics. Downloaded raw data was analysed and peptide ion intensities were exported for protein grouping and relative quantification with benchmarking software implemented in Java (<https://github.com/HayleyPrice/ProteinGrouping>). Resulting protein intensities were processed with existing DE analysis software and the results were evaluated by their ability to detect relative quantitative changes in protein

## Optimising the statistical pipeline for quantitative proteomics

levels. Three statistical packages were assessed: MSstats, QPROT, and a Welch's *t*-test.

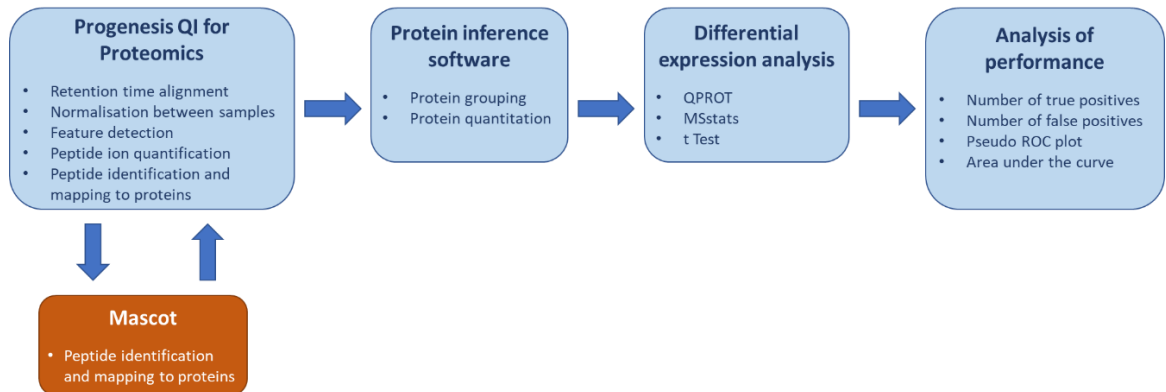


Figure 2.2.4; Summary of stages of benchmarking pipeline. Quantitative processing was performed with Progenesis QIP, protein inference was performed with benchmarking software, DE analysis was performed through QPROT, MSstats and *t*-test analyses, and analysis of performance was calculated using the area under the curve.

### Quantitative processing

Binary data in .raw files was imported into Progenesis QIP v4.2. Sample ions were aligned to the Progenesis QIP automatically selected reference run. Peak picking was performed using maximum sensitivity, as previous unpublished benchmarking work from the group demonstrated that this gives the best performance from Progenesis QIP. A single map of all peptides was created from an aggregated data set of all peak information from aligned sample files. Ion abundance was calculated by summing the intensities of the isotopes of the peptide. Quantified MS/MS spectra were exported for identification with Mascot, spectra with a rank greater than 3 were excluded from analysis. The rank is a value provided by Progenesis QIP for each MS/MS spectrum found by comparing its percentage value against all other spectra matched to the same peptide ion. Excluding higher ranked peptides from the analysis reduced the number of spectra being used for each peptide ion. Mascot search was performed according to the parameters of the datasets, as described in Table 7. The search was also conducted against a decoy database of reverse sequences and results with ion scores higher than required for peptide-spectrum match-level  $FDR < 1\%$  were included for further analysis. Peptide identification and mapping information was imported into Progenesis QIP and scalar

normalisation was performed. Identified and normalised peptide ion abundances were exported from Progenesis QIP for post-processing with the benchmarking protein inference pipeline.

### Protein inference software

A protein inference algorithm was implemented in Java and summarised in

Figure 2.2.5.

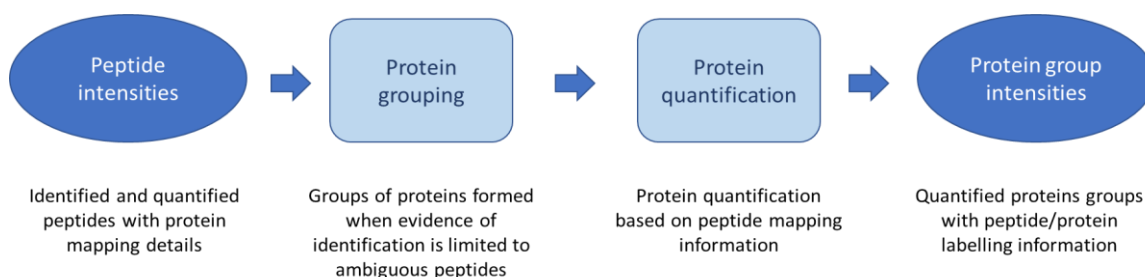


Figure 2.2.5; Summary of the protein inference software pipeline; input data was identified and quantified peptide ion intensities, protein grouping and protein quantification was performed and output of protein intensities given.

Protein grouping and quantification were performed by the benchmarking inference software. Identified and normalised peptide ion intensities were imported and the grouping methods described in the introduction were applied. Two separate analyses were performed for proteins identified by a minimum of two unique peptides and proteins identified by a single unique peptide. Four methods to treat peptide ions with the same primary amino acid sequence were performed, as described in Table 5 in the introduction. Three methods for protein quantification were employed, as described in Chapter 1. In this benchmarking exercise we compared results from the three different methods for peptide to protein inference summarised in Table 6.

Table 6 As MSstats uses peptide intensities for input, the protein quantification options did not apply to this analysis. Instead, both methods for feature subset selection and summarisation within MSstats (as described in the Chapter 1) was

## ***Optimising the statistical pipeline for quantitative proteomics***

employed. As a result of these parameter options, there were 24 different sets of protein abundance data analysed using QPROT and *t*-test (summarised in Table 8) and 32 MSstats analyses (summarised in Table 9).

*Table 8; Summary of parameters producing protein abundance data analysed by QPROT and t-test.*

Number of unique peptides required for protein identification	Peptide ion selection options	Protein quantification options
1	Summed charges	Sum all
1	Summed charges	Hi-3
1	Summed charges	Progenesis Hi-3
1	Summed modifications	Sum all
1	Summed modifications	Hi-3
1	Summed modifications	Progenesis Hi-3
1	Both summed	Sum all
1	Both summed	Hi-3
1	Both summed	Progenesis Hi-3
1	All separate	Sum all
1	All separate	Hi-3
1	All separate	Progenesis Hi-3
2	Summed charges	Sum all
2	Summed charges	Hi-3
2	Summed charges	Progenesis Hi-3
2	Summed modifications	Sum all
2	Summed modifications	Hi-3
2	Summed modifications	Progenesis Hi-3
2	Both summed	Sum all
2	Both summed	Hi-3
2	Both summed	Progenesis Hi-3
2	All separate	Sum all
2	All separate	Hi-3
2	All separate	Progenesis Hi-3



## ***Optimising the statistical pipeline for quantitative proteomics***

*Table 9; Summary of parameters producing protein abundance data analysed by MSstats.*

Number of unique peptides required for protein identification	Peptide ion selection options	Feature subset options	Summarisation options
1	Summed charges	All	Linear
1	Summed charges	All	TMP
1	Summed charges	Hi-3	Linear
1	Summed charges	Hi-3	TMP
1	Summed modifications	All	Linear
1	Summed modifications	All	TMP
1	Summed modifications	Hi-3	Linear
1	Summed modifications	Hi-3	TMP
1	Both summed	All	Linear
1	Both summed	All	TMP
1	Both summed	Hi-3	Linear
1	Both summed	Hi-3	TMP
1	All separate	All	Linear
1	All separate	All	TMP
1	All separate	Hi-3	Linear
1	All separate	Hi-3	TMP
2	Summed charges	All	Linear
2	Summed charges	All	TMP
2	Summed charges	Hi-3	Linear
2	Summed charges	Hi-3	TMP
2	Summed modifications	All	Linear
2	Summed modifications	All	TMP
2	Summed modifications	Hi-3	Linear
2	Summed modifications	Hi-3	TMP
2	Both summed	All	Linear
2	Both summed	All	TMP
2	Both summed	Hi-3	Linear
2	Both summed	Hi-3	TMP
2	All separate	All	Linear
2	All separate	All	TMP
2	All separate	Hi-3	Linear
2	All separate	Hi-3	TMP

### **iii. Differential expression analysis**

Identified, quantified, and normalised peptide and protein abundances were exported from the protein inference software in the correct format for the DE analysis stage. Due to the Progenesis QIP software's alignment method and their claim that analysed data contains no missing values, imputation was not a focus of this thesis. The comparison of DE analysis methods requires a level playing field of input data and any issue with the suitability of imputation method will be standard for each DE analysis. Substitution of zero values was performed to alleviate problems due to division of zeros. Therefore, all zero value abundances were changed to 0.0000001.

#### ***QPROT***

Pairwise comparison of protein intensity data was analysed with the QPROT software package, which was downloaded from SourceForge (<https://sourceforge.net/projects/qprot/>) and executed using Unix shell. Normalisation within the software was disabled as this had been carried out during quantitative processing with Progenesis QIP. The protein abundances were log transformed and DE was carried out. For the burn-in, 100,000 samples were drawn, as were 100,000 samples of each model parameter for the main iterations to ensure the software was run for a sufficient time. These values were based on a previous exploratory analysis (Supplementary material, 'Chapter 2, QPROT\_burnin.docx').

#### ***MSstats***

Group comparison of peptide intensity data was performed with the MSstats software package, which was installed from Bioconductor (Morgan, 2019) and executed using R (R Core Team, 2020). The software's *dataProcess* method log transformed the peptide intensity data with base 2 and formatted it for model fitting and comparison. Normalisation was disabled as this had been carried out during quantitative processing with Progenesis QIP.

### ***t*-Test**

Pairwise comparison of protein intensity data was analysed with Welch's two-sample *t*-test (Welch, 1947), which uses modified degrees of freedom compared to the Student's *t*-test to increase test power with unequal variance between groups. The analysis was performed using the `t.test()` function in R.

## **iv. Analysis of performance**

Known concentrations of proteins spiked into an unchanging complex proteome background allowed pairwise comparison to simulate fold change. Prior knowledge of the expected proteins to be differentially expressed allowed us to assess the software's performance. The output of the analysis was ordered by decreasing significance according to the software's significance metric; *p*-value for MSstats and *t*-test, Z-statistic for QPROT, putting the proteins with most evidence of change in abundance at the top of the list. A running tally of the number of spike-in proteins correctly identified as changing in concentration was calculated to give the true positive rate. The recall for each protein was calculated by dividing the total number of true positives detected by the total number of spike-in proteins in the sample (234 for PXD001385 and 48 for PXD001819 and PXD002099).

$$\text{Recall} = \frac{\text{No. of true positives detected}}{\text{Total no. of true positives in sample}}$$

The proportion of false positives among all proteins called significant, the false discovery proportion (FDP), was calculated for each protein using the equation:

$$\text{FDP} = 1 - \frac{\text{No. of false positives detected}}{\text{No. of false positives detected} + \text{No. of true positives detected}}$$

Analysis of performance was evaluated using Precision–Recall curves, which are used to visualise and quantify the impact of a threshold on the trade-off between false positives and false negatives and plots sensitivity or recall as a function of FDP or 1-Precision.

The area under the curve provided an evaluation metric of the performance of the software, with the aim of having a low FDP while maintaining good recall. The precision recall curve was smoothed using the minimum FDP for protein recall.

### ***Performance of protein inference parameters***

The effect of applying different parameters at the protein inference stage was compared using scatter plots. The total area the under smoothed Precision-Recall curve was shown against recall at 5% FDP for the different methods for peptide ion selection and protein quantification, with higher values indicating better performance.

## **v. Imputation**

To assess the impact of imputation on DE analysis, t-test analysis with a BH corrected p-value of 0.05 to indicate significance was used, and the number of DE proteins was compared for each dataset using the following methods to deal with zero values:

- Imputation of  $1 \times e^{-7}$
- Imputation of  $1 \times e^{-10}$
- Removal of proteins with any zero values from the analysis

Ions with the same primary peptide sequence and different charge states' intensities were summed. Ions with the same primary peptide sequence but with artefactual modifications were treated separately. Protein quantification was based on the sum of the average intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptide were selected.

## **vi. Assessment of benchmarking data**

The development of quality benchmarking data that allows the developer to demonstrate software's performance in real-world utility is difficult (Peters et al., 2018). Changing background protein levels will incur false positive results. Spike-in

## *Optimising the statistical pipeline for quantitative proteomics*

quantities lower than stated will hide the signal, with intensity variation across technical replicates adding to this problem, causing false negatives and issues with normalisation. To validate the benchmarking exercise, analysis was performed to assess the precision and accuracy of the spike-in data-set used; **accuracy** being how close a measurement is to the actual value, and **precision** being how close together a group of measurements are. In using ground truth data for assessment, scientists are reliant on measured value being as expected. Each of the datasets was processed with Progenesis QIP as described above, and protein inference was performed using the summation of all non-conflicting peptides on the raw peptide abundances. The following analyses were performed to investigate and compare the properties of the three datasets.

The coefficient of variation between protein intensities from technical replicates was examined to see if there was an acceptable level of precision across samples within the same condition. The coefficient of variation (CV) is a method to measure the dispersion about the mean. It normalises the standard deviation in order to allow for comparisons of variation as a measure of overall precision. A smaller CV indicates a more precise set of data. CV levels of over 20% are usually considered to be unacceptably imprecise (Jelliffe et al., 2015). It is calculated as the ratio of the standard deviation to the mean using the equation:

$$CV = \frac{\sigma}{\mu}$$

Where

$\sigma$  is the standard deviation  
 $\mu$  is the mean

The log<sub>2</sub> fold-changes of spike-in proteins were calculated to compare with the expected level to assess if there were errors in sample preparation. The coefficient of variation between background protein intensity values was calculated to see if they accurately represented the proteins unaffected by the experimental conditions.

## 2.3. Results and discussion

### i. Summary of DE analysis

Table 10; Summary of quantities of proteins from analysis. Minimum and maximum values are given for proteins identified by a minimum of one or two unique proteins over all possible parameter options. A FDP of less than 0.05 was used as the significance threshold for classifying proteins as differentially expressed.

Dataset	Number of spike-in proteins	Unique peptides	Total identified proteins		Total identified spike-in proteins		Number of proteins classified as DE					
			Min	Max	Min	Max	MSstats		QPROT		t-Test	
							Min	Max	Min	Max	Min	Max
PXD001385	234	1	1864	1864	234	234	108	222	227	245	81	216
		2	1334	1396	152	165	84	158	155	174	55	161
PXD001819	48	1	957	962	48	48	8	48	0	45	6	44
		2	603	675	43	46	25	47	26	47	6	45
PXD002099	48	1	1531	1531	47	47	0	3	0	1531	0	3
		2	832	916	47	47	0	3	0	916	0	3

Table 10 gives a summary of the analysis of the benchmarking data. There were a maximum of 1864, 957, and 1531 proteins identified in datasets PXD001385, PXD001819, and PXD002099, respectively. In the analysis of PXD001385 and PXD001819, using a minimum of one unique peptide for protein identification allowed all of the spike-in proteins to be identified. Requiring a minimum of two unique peptides for protein identification resulted in not all of the spike-in proteins being identified; between 152 and 165 for dataset PXD001385 and between 43 and 46 for dataset PXD001819. For dataset PXD002099, only 47 of the 48 spike-in proteins were identified regardless, of whether one or two unique proteins were required for identification. Overall, there was a smaller number of identified proteins for all datasets when using the two peptide rule.

ii. Performance of differential expression analysis

PXD001385

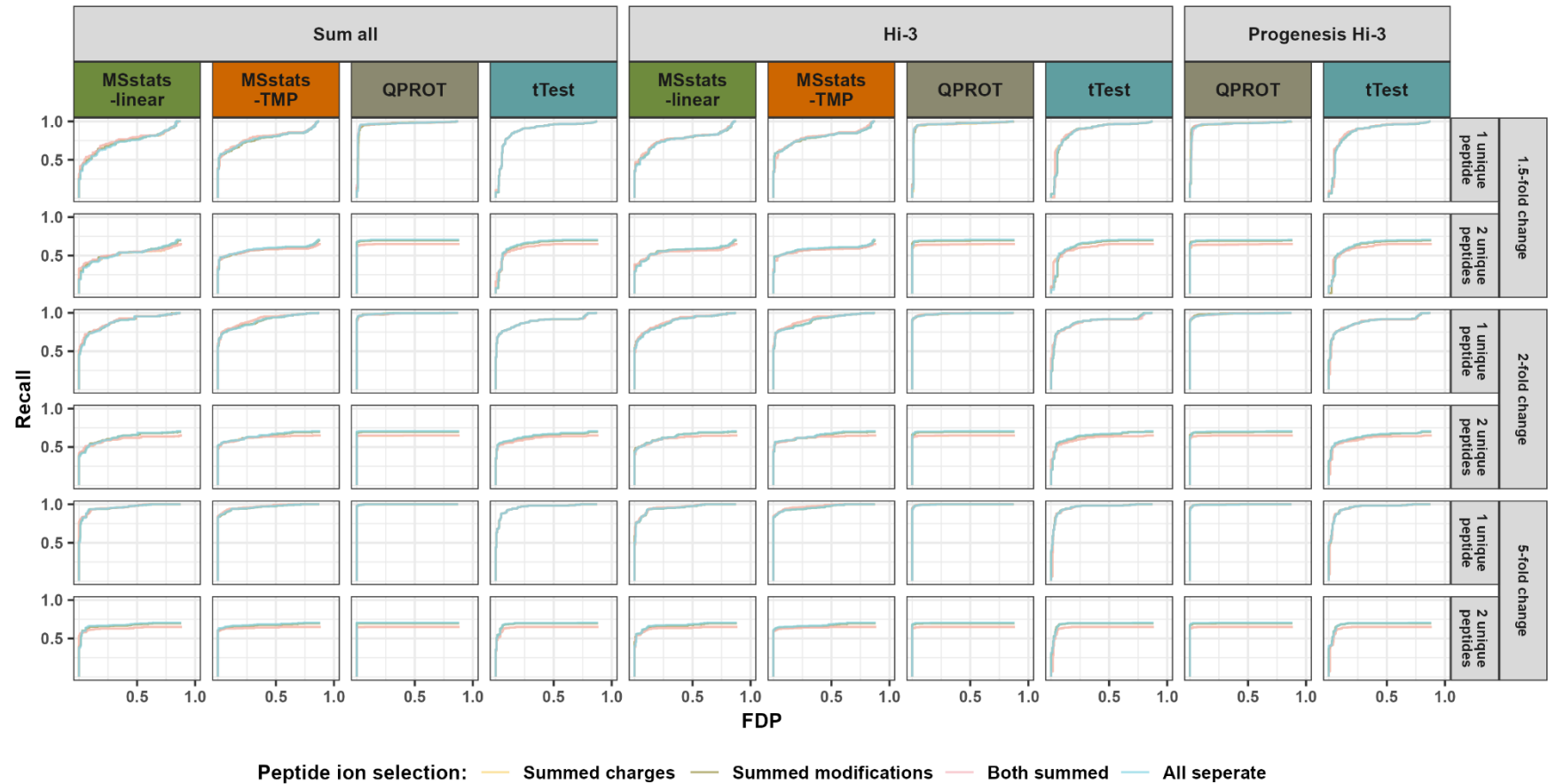


Figure 2.3.1; Smoothed Precision-Recall plots showing the performance of the different DE analysis and quantification methods for the dataset PXD001385 across 3 fold-change simulations (1.5X, 2X and 5X) using one or two unique peptides as minimum for protein identification. Recall is plotted as a function of 1 - Precision (FDP), method of peptide ion selection is shown as line-colour.

## ***Optimising the statistical pipeline for quantitative proteomics***

Figure 2.3.1 shows the analysis of dataset PXD1385 where using a minimum of two unique peptides for identification resulted in not all of the spike-in proteins being identified (Table 10). This is reflected in the smoothed Precision-Recall plots where a smaller area under the curve was achieved for all analysis using the two peptide rule. Overall, QPROT gave the best performance in all fold change scenarios. MSstats gave similar performance to *t*-test in 5- and 2-fold change simulations (rows 3 – 6) and performed worse than *t*-test in the smaller 1.5-fold change simulation (rows 1 and 2), and TMP was slightly superior to the linear method for protein quantification. Parameter option selection had less effect in this dataset, with all options performing similarly.



PXD001819

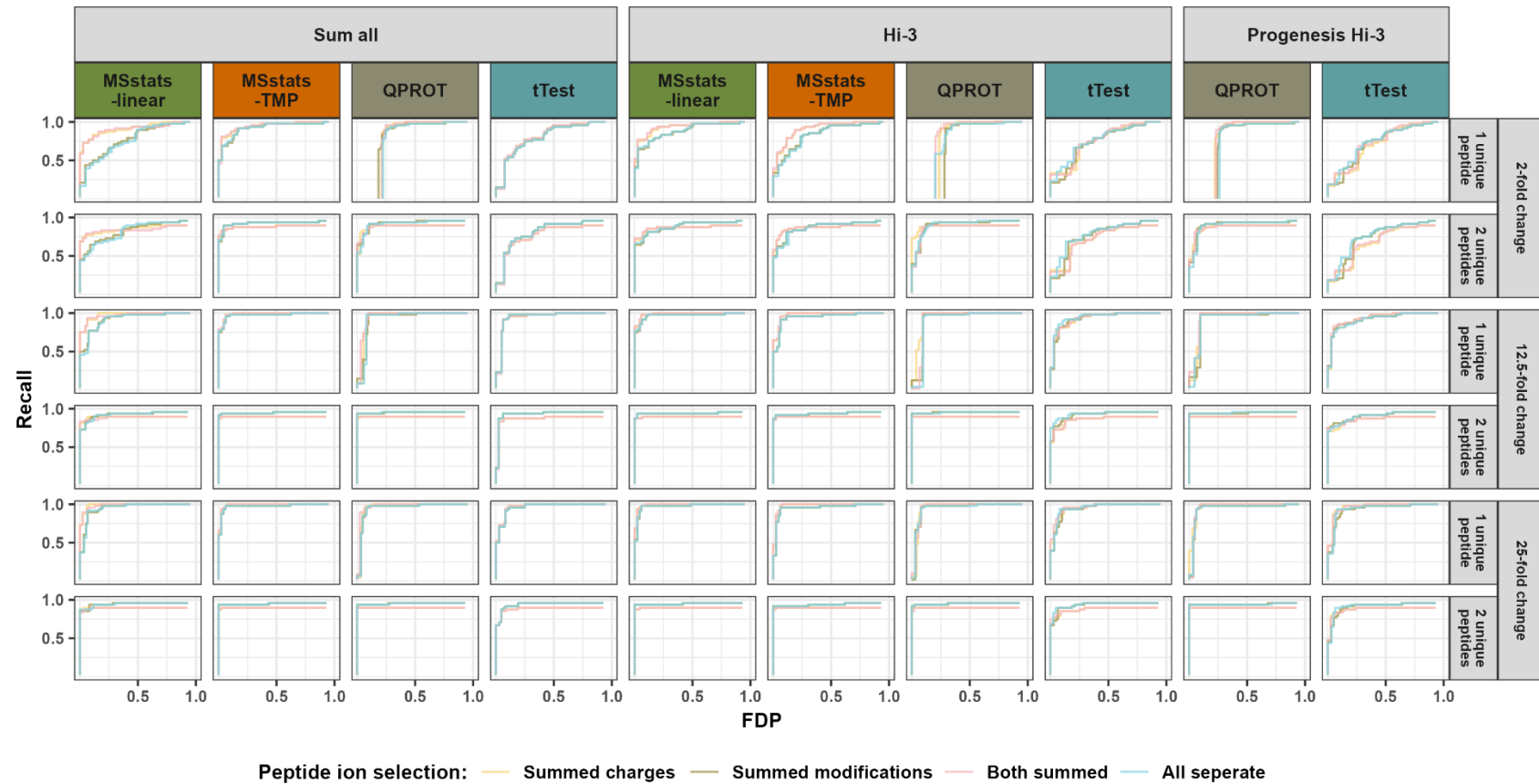


Figure 2.3.2; Smoothed Precision-Recall plots showing the performance of the different DE analysis, summarisation methods and peptide selection for the dataset PXD001819 across 3 fold-change (2X, 12.5X and 25X) simulations using one or two unique peptides as minimum for protein identification. Recall is plotted as a function of 1-Precision (FDP), method of peptide ion selection is shown as line-colour.

## ***Optimising the statistical pipeline for quantitative proteomics***

On data set PXD001819, shown in Figure 2.3.2, QPROT appeared to perform best overall in the two higher fold change scenarios, with MSstats-TMP giving a similar performance. In the 2-fold change scenario, MSstats-TMP using all peptides mapped to the protein for quantification was the most effective analysis. While QPROT struggled to perform at low FDPs when only one peptide was used for quantification (some parameter combinations failed to give any significant DE proteins at a FDP of 0.05, Table 10). Overall, the performance of the *t*-test was less good than the other analysis methods, requiring a high FDP to give high recall. In the 2-fold change simulation, the parameter methods strongly affected how MSstats performed, with peptide ion selection making the most difference. At low FDPs, summing all peptide ions with the same primary amino acid sequence improved performance (for example Figure 2.3.2, column 1, row 1), although at higher FDPs it sometimes gave worse recall than treating ions separately (for example Figure 2.3.2, column 1, row 2). As expected, the DE methods were more successful at identifying changing proteins when there was a larger fold change simulation. Using a minimum of two peptides for protein identification also gave more reliable results.

PXD002099

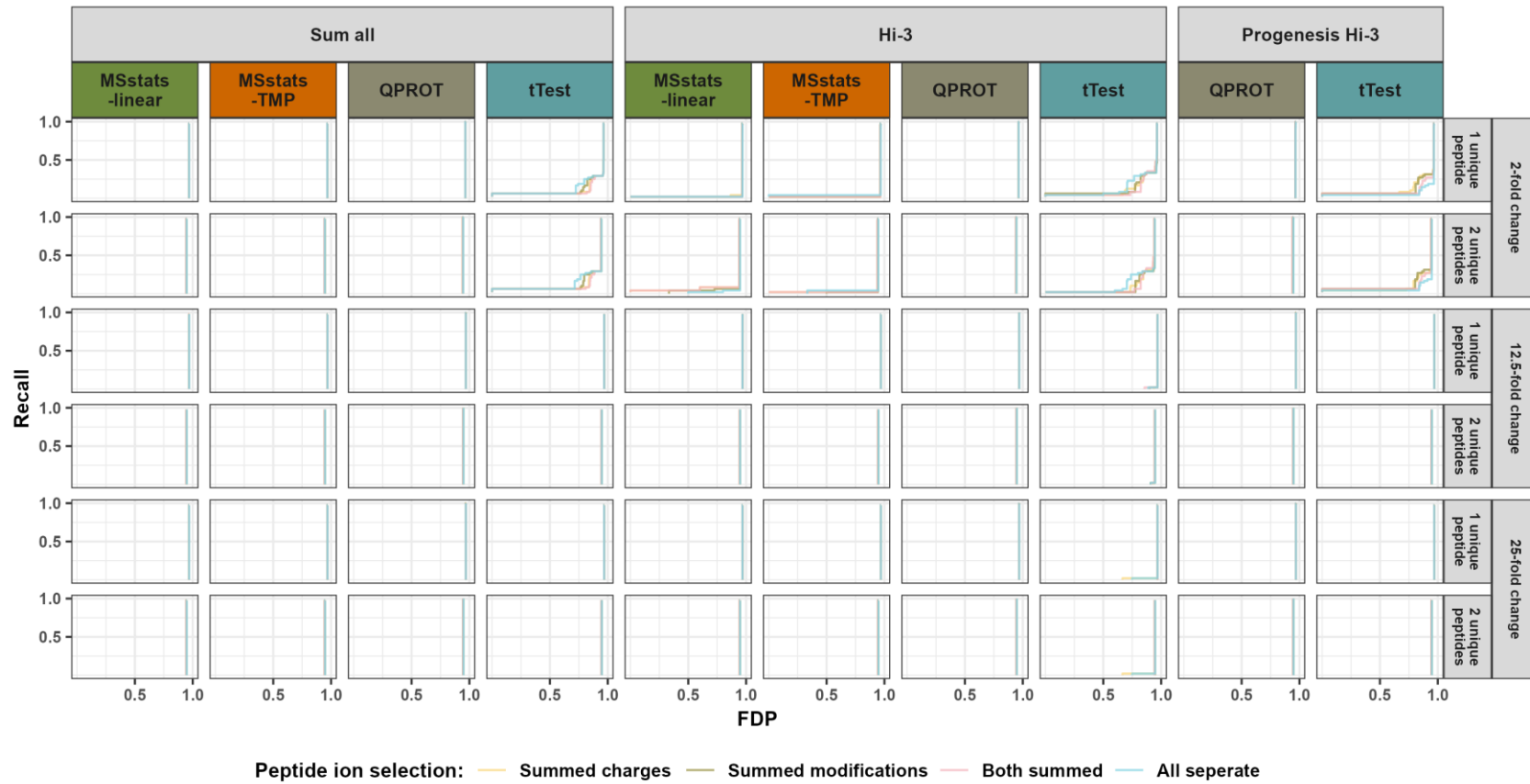


Figure 2.3.3 Smoothed Precision-Recall plots showing the performance of the different DE analysis, summarisation methods and peptide selection for the dataset PXD002099 across 3 fold-change (2X, 12.5X and 25X) simulations using one or two unique peptides as minimum for protein identification. Recall is plotted as a function of 1 - Precision (FDP), method of peptide ion selection is shown as line-colour.

Figure 2.3.3 shows the analysis of dataset PXD002099. The summary in Table 10 showed many of the analysis combinations classified none or very few proteins as significantly changing. It also showed that, in some circumstances, all proteins were described as differentially expressed. The result of this was very poor performance on the smoothed Precision-Recall plots, with only a small area under the curve produced in the *t*-test analysis (Figure 2.3.3, columns 4, 6, and 7, rows 1 and 2).

### ***Summary of results***

Figure 2.3.1 to Figure 2.3.3 shows the performance of DE methods, protein quantification methods, peptide ion selection methods, and the number of unique proteins required for protein identification using all non-conflicting peptides for protein quantification over three fold change simulations for the datasets PXD001385, PXD001819, and PXD002099.

Overall, QPROT gave better performance than the *t*-test in all fold change simulations in both of the datasets. MSstats gave better or similar performance to the *t*-test except for in the small fold-change simulation of dataset PXD001385. For dataset PXD001819, QPROT's performance was dependent on the number of unique peptides used for identification and the fold-change scenario; MSstats was better when one unique peptide was used for identification in the 2-fold change scenario. QPROT's performance in dataset PXD001385 was consistently the best. However, this is the dataset that was used in the software developer's own paper. Peters *et al.* (2018) discuss the possibility of unconscious reporting bias when method development is done in conjunction with the benchmark that is used to show the improved performance of the tool. Of the two MSstats summarisation methods, TMP gave better results than linear modelling and both gave similar performances with dataset PXD001385. In dataset PXD001819, the effects of quantification method and peptide ion selection made more difference in dataset PXD001819. It was difficult to see clearly an overall best method, but it appeared that summing all non-conflicting peptides for protein quantification and summing peptide ions

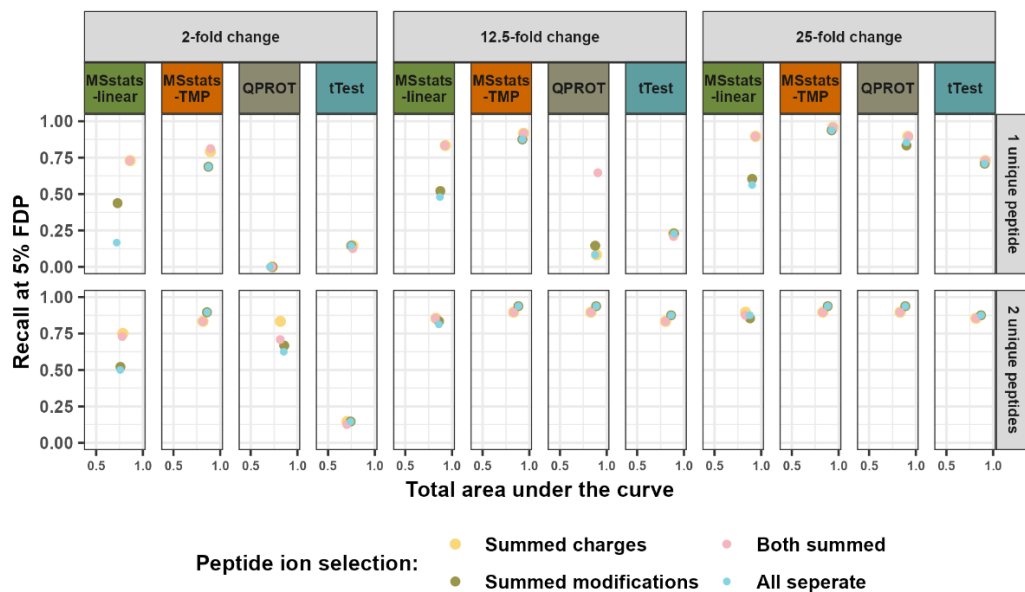
with the same sequence and different charge state gave consistent performance. Due to the poor DE analysis results, dataset PXD002099 was excluded from the ranking of parameter methods and only MSstats- TMP was used to compare threshold selection in the next section of the results.

In MSstats analysis, as the benchmarking data consists of only technical replicates without any biological replicates, in this chapter, a simple linear model is used for DE analysis. If the data had also biological replicates, it is considered to have random sources of variation, and a mixed model is applied to include random-effects accounting for variability between samples in the same condition. A limitation of benchmarking MSstats software with artificial ‘ground truth’ data is that the full capabilities of the software are not assessed. A more accurate method for evaluating MSstats would be to use benchmarking data that has biological variability (covered in the following chapter), thereby utilising the mixed modelling algorithm provided by the software.

### **iii. Performance of protein inference parameters**

#### ***Peptide ion selection***

Next, the effect of the method used to generate peptide-level data was explored i.e. summing charge states, summing signals from different peptidofoms due to modifications.



### ***Optimising the statistical pipeline for quantitative proteomics***

Figure 2.3.4 shows that peptide ion selection did not have much of an effect for dataset PXD001385 at a recall of 5% FDP. For dataset PXD001385, summing peptide ions with the same charge state and summing peptides with the same charge state regardless of modifications appeared to give the best result. The method of peptide ion selection made the greatest difference when using MSstats linear modelling as DE analysis, which is to be expected as the DE analysis is based directly off peptide signals unlike QPROT that works from values summarised up to the protein-level, which could cancel out differences in the method of summarising peptide intensity.

## Optimising the statistical pipeline for quantitative proteomics

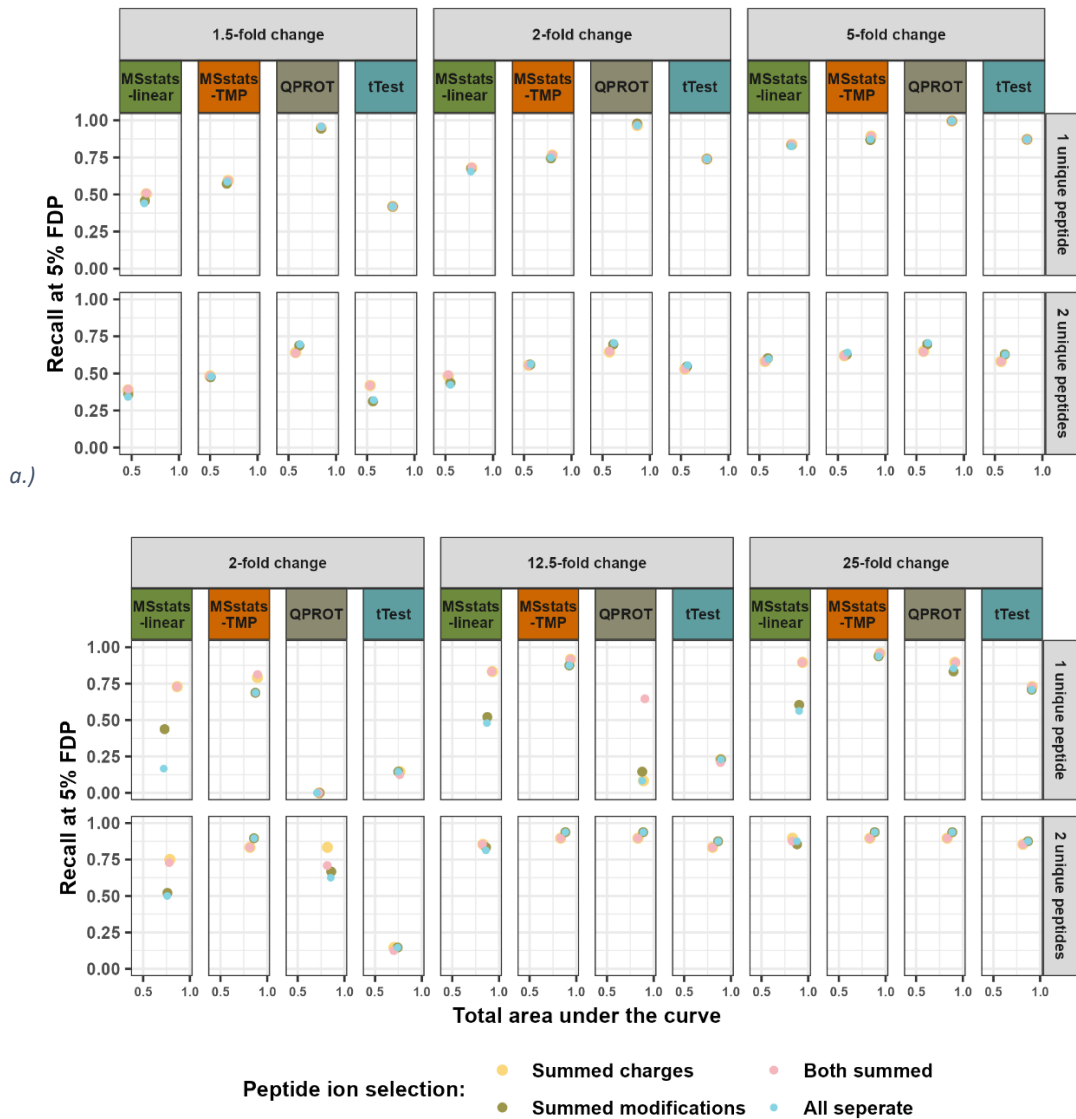


Figure 2.3.4: Performance of DE methods, peptide ion selection methods, and number of unique peptides required for protein identification using all non-conflicting peptides for protein quantification over three fold change simulations for datasets a.) PXD001819 and b.) PXD001385. The total area under the smoothed Precision-Recall curve is shown against the recall at 5% FDP. The size of data point does not correspond to a numeric value and is used to display overlapping data points.

### Quantification method

The effect of the quantitation method is shown in Figure 2.3.5 for datasets PXD001385 and PXD001819. Again, parameter options do not have a great effect on the results for PXD001819. For PXD001385, quantification by summing all peptides mapped to the protein appeared to give the best overall result. Following this analysis, parameter options of summing peptide ions with the same charge state for peptide ion selection and summing all peptides mapped to the protein for quantification were used for the threshold selection comparison in the next section. Peptide ion selection and quantification method

## Optimising the statistical pipeline for quantitative proteomics

affected MSstats performance the most, but the parameter comparison was not like for like. For *t*-test and QPROT, only the quantification method and peptide ion selection were compared; for MSstats, there was one fewer quantification method (no option to include information from non-conflicting peptides with the Progenesis Hi-3 method) and two methods for peptide to protein summarisation, linear and TMP. Overall, of the two options provided in MSstats DE analysis, using TMP for peptide summarisation appears to give the best results for (Figure 2.3.4 and Figure 2.3.5, comparison between columns 1 and 2, 4 and 5, and 9 and 10).

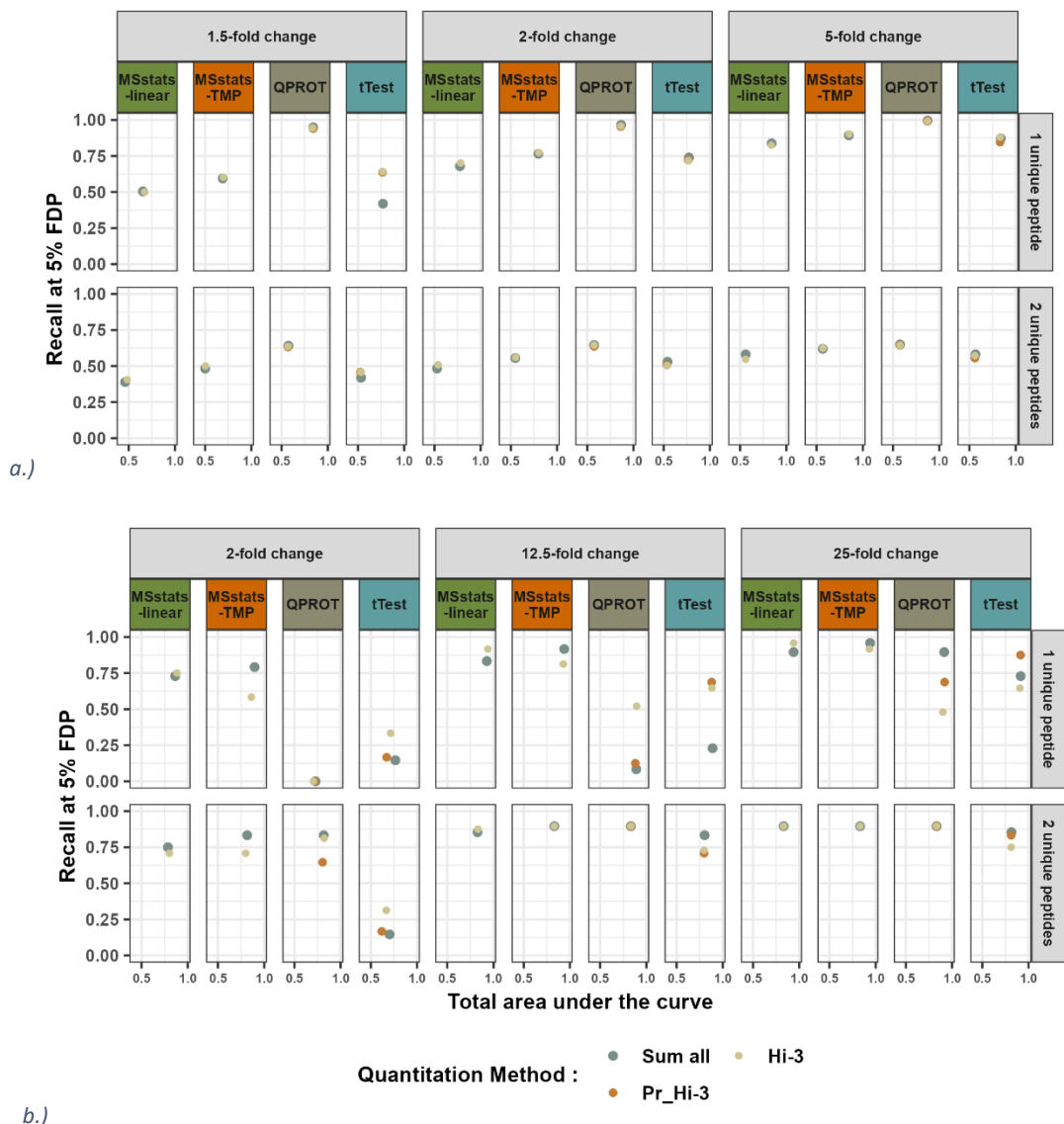
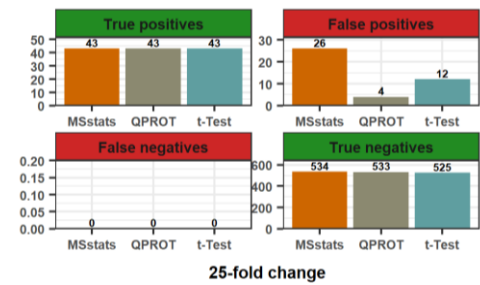
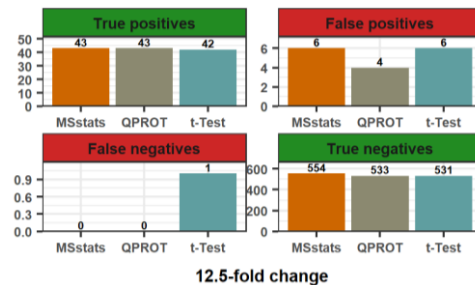
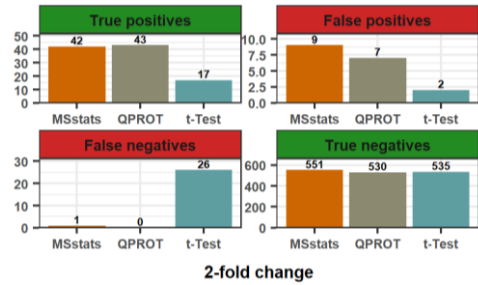


Figure 2.3.5; Performance of DE methods, quantification methods, and number of unique peptides required for protein identification using summed peptide ions with the same sequence and different charge state over three fold change simulations for datasets a.) PXD001819 and b.) PXD001385. The total area under the smoothed Precision-Recall curve is shown against the recall at 5% FDP. The size of a data point does not correspond to a numeric value and is used to display overlapping data points.



### Threshold selection comparison

Threshold = 0.01

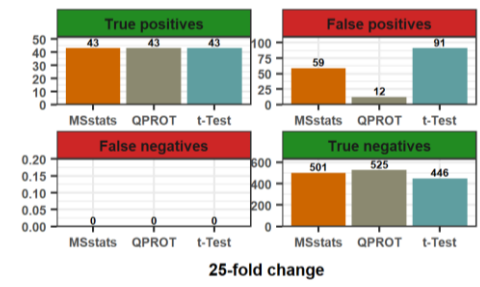
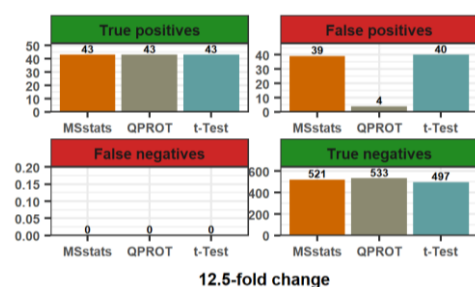
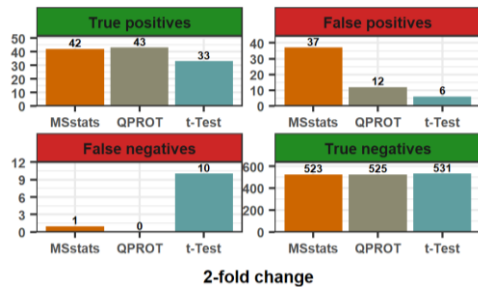


2-fold change

12.5-fold change

25-fold change

Threshold = 0.05

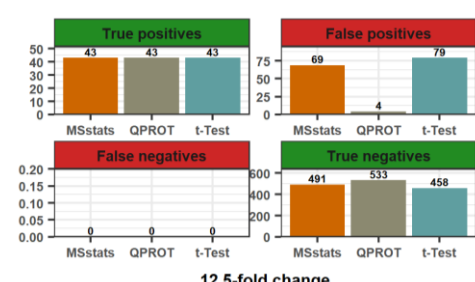
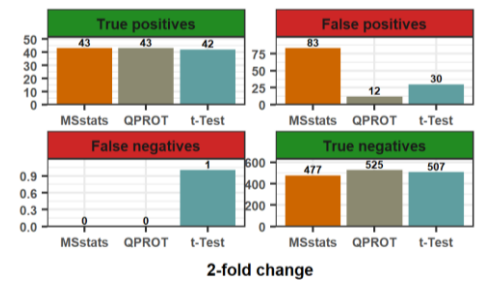


2-fold change

12.5-fold change

25-fold change

Threshold = 0.1



2-fold change

12.5-fold change

25-fold change

Figure 2.3.6; Number of true positives, false positives, false negatives, and true negatives for each of the differential expression methods at significance thresholds of 0.01, 0.05, and 0.1 over three fold change simulations for data set PXD001819. Progenesis normalised peptide ion abundances with ion intensities of the same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification. MSstats performed TMP for summarisation.

**PXD001819**

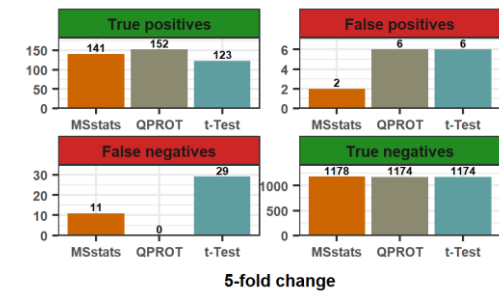
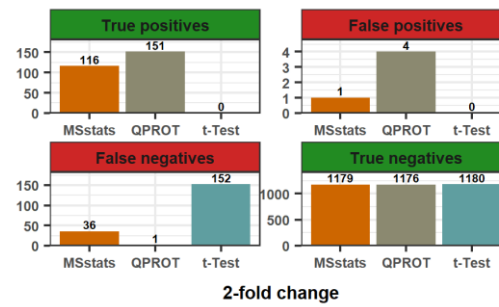
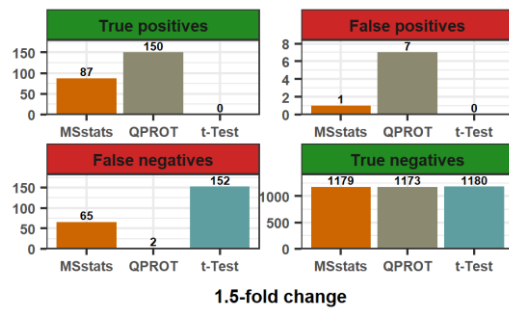
Figure 2.3.6 shows the analysis for PXD001819. In the 2-fold change simulation, a threshold of  $q < 0.01$  gave the best results for QPROT, and all TPs were detected. Decreasing the stringency of the threshold increased FPs; 7 for 0.01, 12 for 0.05 and 0.1. At the same thresholds, MSstats had a similar true positive rate to QPROT but accrued more false positives; decreasing threshold stringency from 0.01 to 0.05 gave no increase in TPs but FPs increased from 9 to 37; decreasing threshold to 0.1 allows detection of all true positives but the number of FPs increases to 83, the highest FP rate of the three methods. The  $t$ -test required a significance threshold of 0.1 to detect 42 out of 43 TPs. But this also allows 30 FPs. The FP rate is better at 0.01 and 0.05 (2, 6) but only 17 and 33 TPs were detected.

In the 12.5-fold change simulation, QPROT and MSstats detected all 43 TPs at 0.01,  $t$ -Test detected 42, with FPs of 6, 4, 6. Decreasing the threshold to 0.05 allowed the  $t$ -test to detect all TPs but increased the number of FPs to 39 (40 for MSstats). A further decrease in threshold to 0.1 increased FPs to 69 and 79 for MSstats and  $t$ -test, respectively. QPROT's total FPs remained at 4 for all three threshold cut-offs.

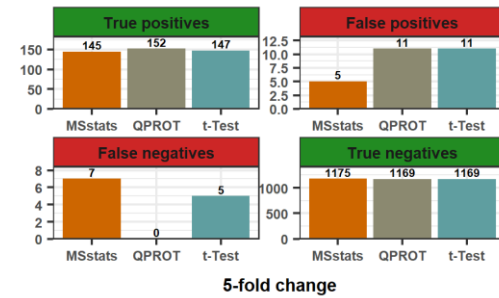
In the 25-fold simulation, all TPs were detected at a threshold of 0.01 with FPs of 26, 4, and 12 for MSstats, QPROT, and  $t$ -test, respectively. FPs increased as threshold stringency decreased (59, 12, and 91 for 0.05 significance threshold and 87, 16, and 172 for the 0.1 significance threshold for MSstats, QPROT, and  $t$ -test, respectively). QPROT was least susceptible to increased FPs as threshold stringency decreased. Overall, QPROT was best able to detect changing proteins at low threshold levels while limiting the amount of incorrectly labelling background proteins as changing. Threshold selection was important when spike-in proteins changed by a small amount, with the  $t$ -test requiring a less stringent significance to identify changing proteins. Increasing the significance threshold had the least impact on the false positive rate for QPROT.

## Optimising the statistical pipeline for quantitative proteomics

Threshold = 0.01



Threshold = 0.05



Threshold = 0.1

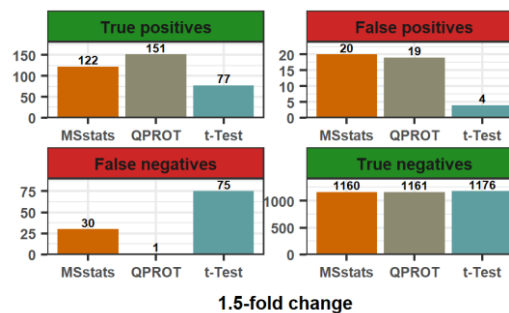


Figure 2.3.7; Number of true positives, false positives, false negatives and true negatives for each of the differential expression methods at significance thresholds of 0.01, 0.05 and 0.1 over three fold change simulations for data set PXD001385. Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification. MSstats performed TMP for summarisation

**PXD001835**

Dataset PXD001385 is shown in Figure 2.3.7. For the 1.5-fold change simulation, QPROT required the most stringent threshold of 0.01 to produce its optimal result, with 150 of 152 TPs detected. Decreasing the strictness in the cut-off increased QPROT's FPs; 7 for 0.01, 12 for 0.05, and 19 for 0.1. For MSstats, decreasing threshold stringency increased the number of detected TPs (87, 115, and 122) but also increased the number of FPs (1, 8, and 20). At the 0.1 cut-off, not all TPs were detected by MSstats, with the FP rate almost the same as QPROT. Using *t*-test, at thresholds of 0.01 and 0.05, no changing proteins were detected. Decreasing the threshold to 0.1 detected 77 out of 152 TPS with only 4 FPs. In this fold-change, decreasing the cut-off stringency further may have been more appropriate.

In the 2-fold change simulation, at a threshold of 0.01, QPROT detected 151 of 152 TPs. Decreasing the stringency of the threshold increased the number of FPs (4 for 0.01, 10 for 0.05, and 13 for 0.1) without detecting the remaining TP. Using MSstats, decreasing the threshold stringency increased the number of detected TPs (116, 130, and 133) but also increased FPs (1, 7, and 15). As with the 1.5-fold change simulation, not all TPs were detected at the least conservative significance threshold with a FP rate similar to QPROT. DE by *t*-test gave no TPs or FPs at 0.01. Decreasing the threshold to 0.05 and 0.1 detected 82 and 121 out of 152 TPS, respectively, with the lowest FP rates, 1 and 3. As with the 1.5-fold change simulation, a more lenient significance threshold could have improved the results.

In the 5-fold change scenario, a threshold of 0.01 was optimal for QPROT with detection of all 152 TPs. Decreasing the stringency of the threshold increased FPs; 6 for 0.01, 11 for 0.05, and 11 for 0.1. Using MSstats, decreasing threshold stringency increased the number of detected TPs 141, 145, and 147 but also increased FPs 2, 5, and 12. Not all TPs were detected, but the FP rate was lower than QPROT for 0.01 and 0.05, and almost the same at 0.1. DE analysis with *t*-Test gave 123 TPs and 6 FPs at 0.01. Decreasing to 0.05 and 0.1 detects 147 and 150 out of 152 TPS with FPs of 11 and 21; decreased cut-off stringency was required to detect DEs but increased FPs.

## ***Optimising the statistical pipeline for quantitative proteomics***

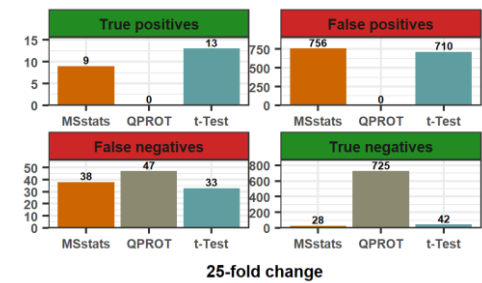
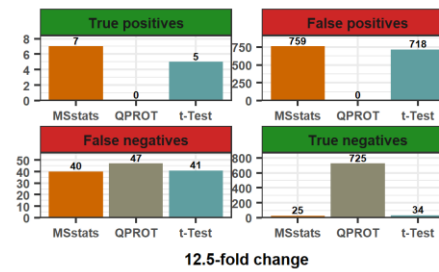
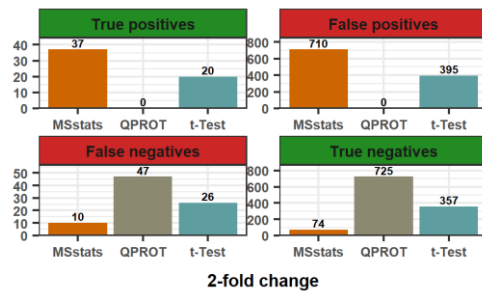
Overall, QPROT performed best, detecting nearly all of the TPs at the low threshold. A less stringent threshold significance than 0.1 could have given better results for MSstats and *t*-test, particularly in the low fold change simulations, and a more stringent significance threshold could have limited QPROT's FP rate even further.

### ***PXD002099***

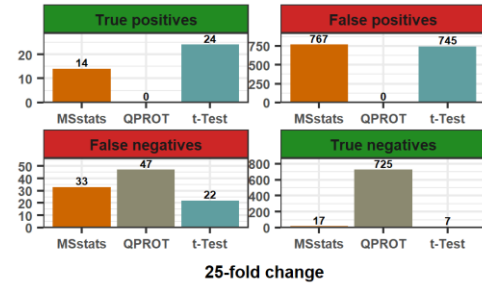
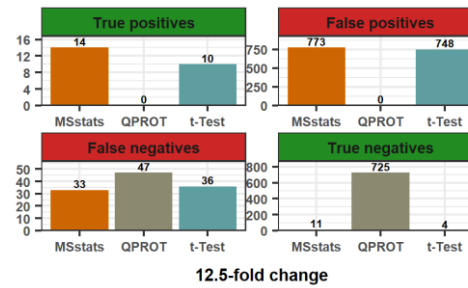
Figure 2.3.8 shows the analysis of the PXD002099 dataset. Most TPs were detected by MSstats and *t*-test, but the FP rate was extremely high and most proteins were identified as changing in abundance. QPROT's analysis produced the opposite; its results conclude that no proteins are detected as changing; both the TP and FP rates are zero.

## Optimising the statistical pipeline for quantitative proteomics

Threshold = 0.01



Threshold = 0.05



Threshold = 0.1

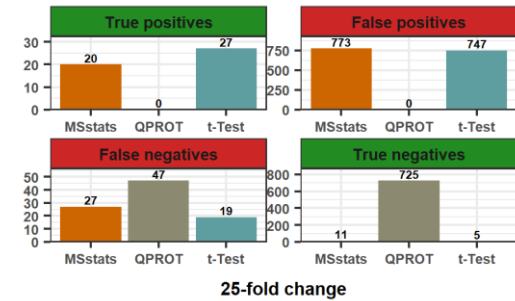
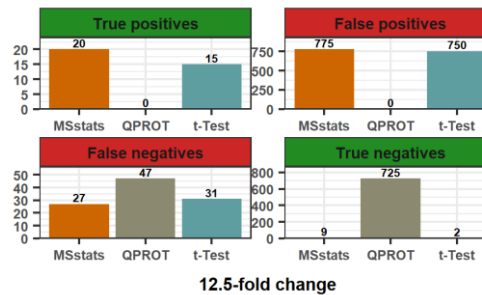
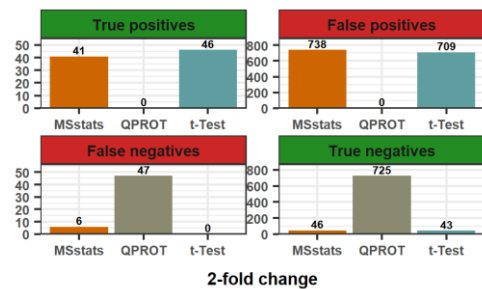


Figure 2.3.8; Number of true positives, false positives, false negatives and true negatives for each of the differential expression methods at significance thresholds of 0.01, 0.05 and 0.1 over three fold change simulations for data set PXD002099. Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification. MSstats performed TMP for summarisation.

## ***Optimising the statistical pipeline for quantitative proteomics***

Overall, the FDP criteria of 0.1, 0.05, and 0.1 were chosen because biologists consider them to be standard FDP rates. QPROT's performance was improved by employing a more conservative cut-off, while MSstats and *t*-test required a more lenient approach. This demonstrates that an arbitrary significance threshold does not always get the best results. In ground truth experiment, information about what is changing is known; however, in a biological experiment, this information is not available, making it difficult to pick the appropriate significance threshold. Greenland et al. (2016) proposed that valid interpretation of data does not require arbitrary classification of results into 'significant' and 'non-significant'. Rather of categorising proteins as significant or non-significant, the purpose of DE experiments is to uncover potentially relevant proteins for biological settings. The *p*-value threshold for identifying significant results, as stated in Chapter One, is a mechanism for comparing an experiment's result to what would be predicted by chance. A *p*-value of 0.01 indicates that there is a 1% possibility that the reported results could have happened by chance. However, in order to make this comparison, further information regarding the chance of the event occurring is required. In Chapter Three, we present a novel method for evaluating DE experiments that does not rely on arbitrary significance cut-offs.

#### iv. Imputation

In dataset PXD001385, less than 1% of the total proteins were affected by missing abundance data in all of the conditions, and none of the proteins had missing values in dataset PXD001819 (Table 11). Changing the methods used to deal with the missing values had no effect on the FP or FN rate for any of the conditions in either dataset.

*Table 11; Comparison of effects of imputation. t-test analysis with a BH corrected p-value of 0.05 to indicate significance was used and number of DE proteins was compared for each dataset using different methods to deal with zero values. Ions with same primary peptide sequence and different charge state's intensities were summed. Ions with same primary peptide sequence but with artefactual modifications were treated separately. Protein quantification was based on the sum of the average intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptide were selected. Impute 1e-07, imputation of 0.0000001; Impute 1e-10, 0.000000001; Remove, removal of proteins with any zero values from the analysis; FP rate, false positive rate; FN rate, false negative rate.*

Dataset	Fold change	Number of proteins with a zero value	% of total proteins	Imputation method					
				Impute 1e-07		Impute 1e-10		Remove	
				FP rate	FN rate	FP rate	FN rate	FP rate	FN rate
PXD001385	5	12	0.64	0.01	0.12	0.01	0.12	0.01	0.12
	2	14	0.83	0	0.06	0	0.06	0	0.06
	1.5	16	0.95	0	0.06	0	0.06	0	0.06
PXD001819	25	0	0	0.16	0	0.16	0	0.16	0
	12.5	0	0	0.07	0	0.07	0	0.07	0
	2	0	0	0.01	0.25	0.01	0.25	0.01	0.25



v. Assessment of benchmarking data

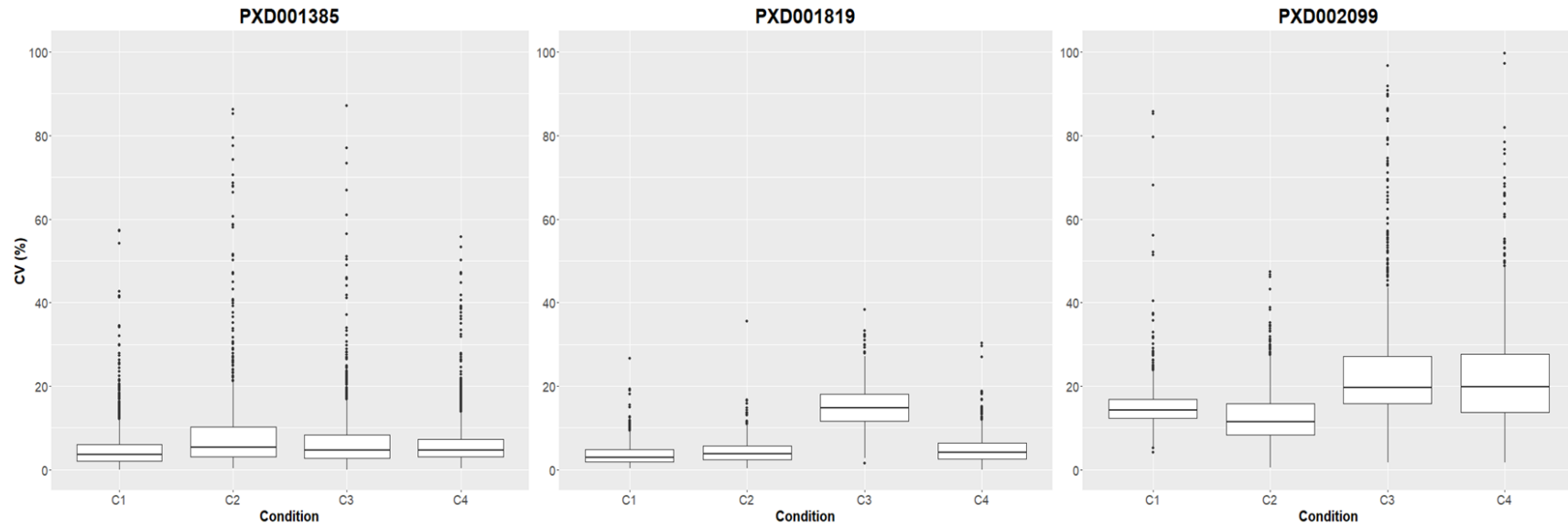


Figure 2.3.9 ; Precision of spike-in data across technical replicates. The percentage coefficient of variance (CV) in protein abundances between technical replicates within conditions for each of the spike-in datasets PXD001385, PXD001819, and PXD002099. Progenesis normalised peptide ion abundances with ion intensities of the same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and a minimum of two unique peptides required for identification. Only CVs below 100% are shown. Dataset PXD002099 had 7 outliers above 100%.

## *Optimising the statistical pipeline for quantitative proteomics*

The precision and accuracy of the spike-in data-set employed were assessed to confirm the benchmarking exercise; accuracy refers to how near a measurement is to the actual value, while precision refers to how close a collection of measurements are together. The plots in Figure 2.3.9 demonstrate the precision of the spike-in datasets used for benchmarking. Using protein intensities for each condition, the percentage coefficient of variation (CV) of the data's technical replicates was calculated. By examining the ratio of the standard deviation to the mean (expressed as a percentage), the level of dispersion within supposedly identical samples was demonstrated. CV levels of over 20% are usually considered to be unacceptably imprecise (Jelliffe et al., 2015). PXD001385 had similar mean CV across conditions, with 75% of each condition's CV values below 10%. However, certain outlying intensities had CV values of over 80%, indicating that while most duplicates were somewhat similar, there were some intensities that were substantially different. PXD001819 had mean CVs below 5% in conditions 1, 2, and 4, with an increased mean CV of 15% in condition 3. PXD001819 also had the smallest range of CVs (below 40%), making it the most precise dataset. However, it lacked consistency due to the increase in mean CV in condition 3. PXD002099 was the least precise dataset; it had a range of 10 – 20% of mean CVs across conditions with the largest range of CV values (maximum of 180%). Over 75% of the intensity levels for conditions 1 and 2 were acceptable, whereas only about half of the intensity values for conditions 3 and 4 were acceptable.

Figure 2.3.10 shows the log<sub>2</sub> fold changes created by the comparison of conditions A, B and C with condition D (5-, 2-, and 1.5 fold changes) with the expected log<sub>2</sub> fold changes indicated in red. Dataset PXD001385 is precise in 5- and 2-fold changes, with the median fold change intensity values almost the same as the expected values. However, there is a wide dispersion in the values, particularly in the 5-fold change comparison, with the maximum values showing higher than a 32-fold change. The log<sub>2</sub> values in the simulated 1.5-fold change comparison are less accurate, with 75% of the values higher than expected.

## Optimising the statistical pipeline for quantitative proteomics

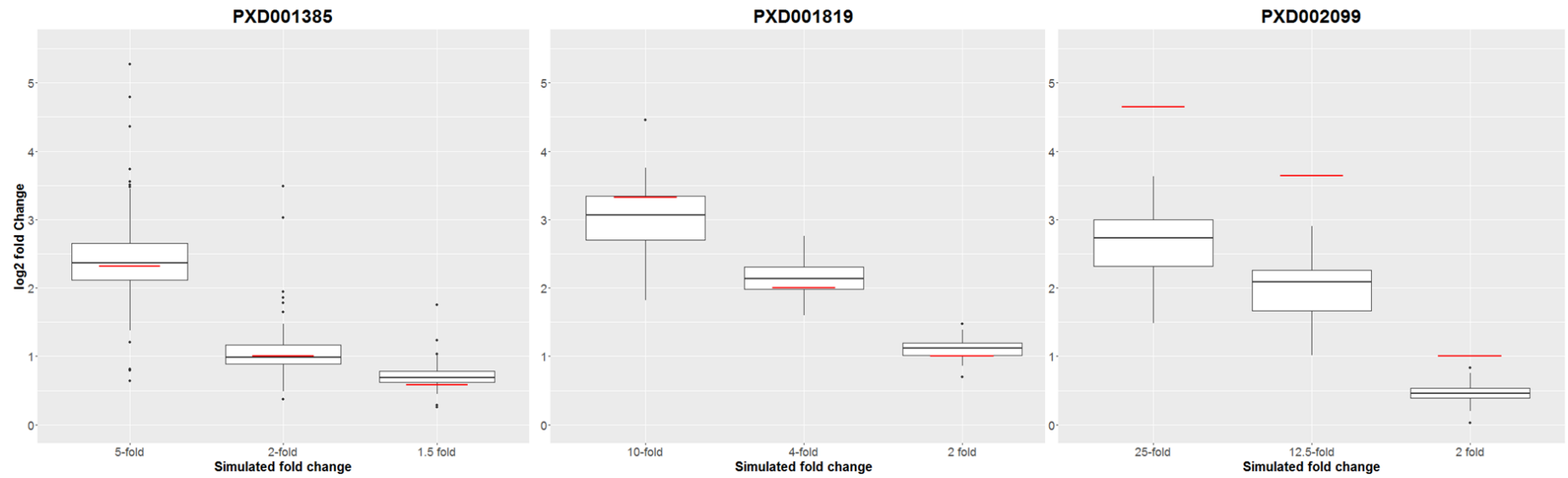


Figure 2.3.10; Accuracy of spike-in proteins demonstrated by log2 fold-change of spike in proteins for sample comparisons in each of the spike-in datasets. Red line shows the expected log2 fold-change. Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification.

## Optimising the statistical pipeline for quantitative proteomics

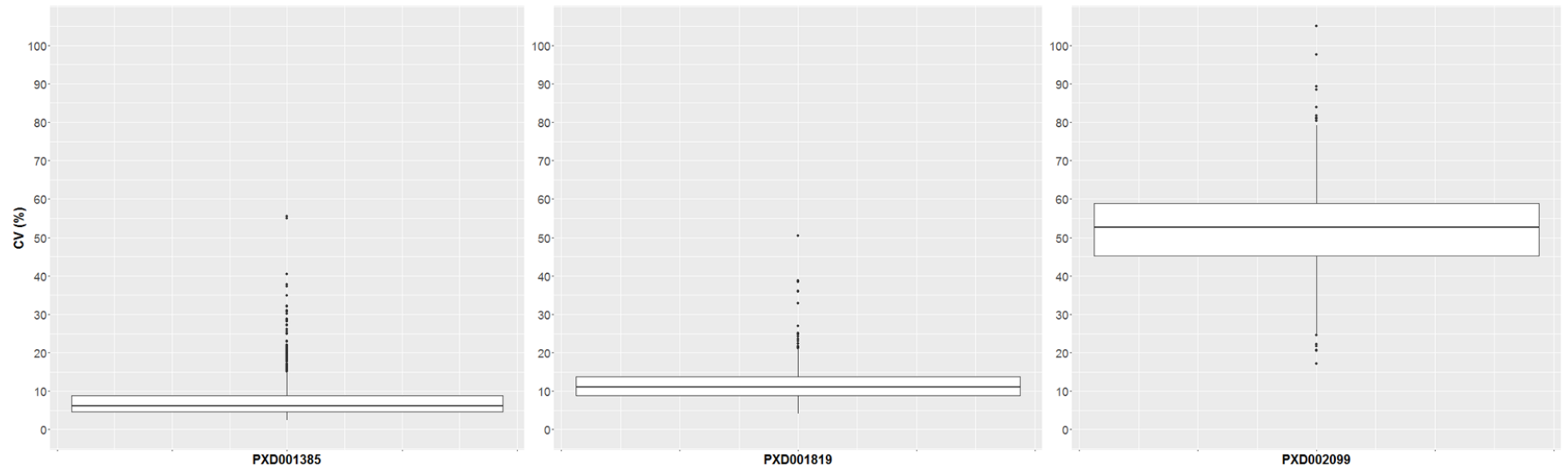


Figure 2.3.11; Accuracy of the background proteins demonstrated with the percentage coefficient of variance of background protein intensity values across samples for each of the spike-in datasets used. Progenesis normalised peptide ion abundances with ion intensities with same sequence regardless of charge state and artefactual modifications summed, all non-conflicting peptides used for protein quantification and two unique peptides used for identification.

## *Optimising the statistical pipeline for quantitative proteomics*

Dataset PXD001819 has a larger range of fold change simulations (10-, 4-, and 2-fold change) but does not represent them as accurately as PXD001385; approximately 75% of the log<sub>2</sub> values fall below the expected amount in the 10-fold change, and 75% are too high in the 4- and 2-fold changes. However, there is less dispersion in the values than in PXD001385. The log<sub>2</sub> results for dataset PXD002099 are poor; all the log<sub>2</sub> fold change values are too low. The actual 25-fold change values range between 2.8 and 11-fold change, while the 12.5-fold change values range between 2- and 8-fold change, with some protein intensities remaining constant across conditions in the 2-fold change comparison.

Figure 2.3.11 depicts the CVs of the background protein intensities across all samples of the datasets. For the data to be effective in benchmarking, the background proteins should remain at the same level across the conditions. Datasets PXD003185 and PXD001819 have percentage CVs of just over 5 and 10%, respectively, meaning that intensity values stay acceptably constant across samples. The background protein intensities in dataset PXD002099 are not acceptably remaining constant; there is a median CV of over 50% with a maximum value of 105%. None of the CV values fell below 15%, meaning that there is an unacceptable amount of change in the supposedly constant intensity level of background proteins for this dataset.

Overall, the assessment of the benchmarking data showed that datasets PXD001385 and PXD001819 were reliable benchmarking data, but dataset PXD002099 was not. The precision of intensities across technical replicates (Figure 2.3.9) showed 75% of dataset PXD001385 had an acceptable level of precision, while PXD001819 was precise in three of its conditions and with an increased amount of error in the fourth. PXD002099 showed an unacceptable amount of variation between replicates, with two conditions having only half the intensity values that were acceptably precise. Analysis of fold-change accuracy (Figure 2.3.10) of the dataset PXD001385 showed a good level of accuracy in two conditions, with the smaller fold-change found to be imprecise. PXD001819 was less accurate, with 75% of the values incorrect across all conditions. PXD002099 was unable to adequately demonstrate the expected

fold-change of spike-in proteins, with not one of the fold-change values at the expected level. In 'ground truth' data, levels of background proteins should supposedly stay constant across conditions. The first two datasets had acceptable consistency across conditions, but in the third dataset, not one background intensity level remained the same across conditions.

The results from PXD002099 show the difficulties all DE methods had analysing the data. None of the DE analyses were accurately able to detect the spike-in proteins as changing across conditions; either the results showed that all proteins were changing or that none of the proteins were changing. The properties of the data were examined to investigate the reason for the analysis problems.

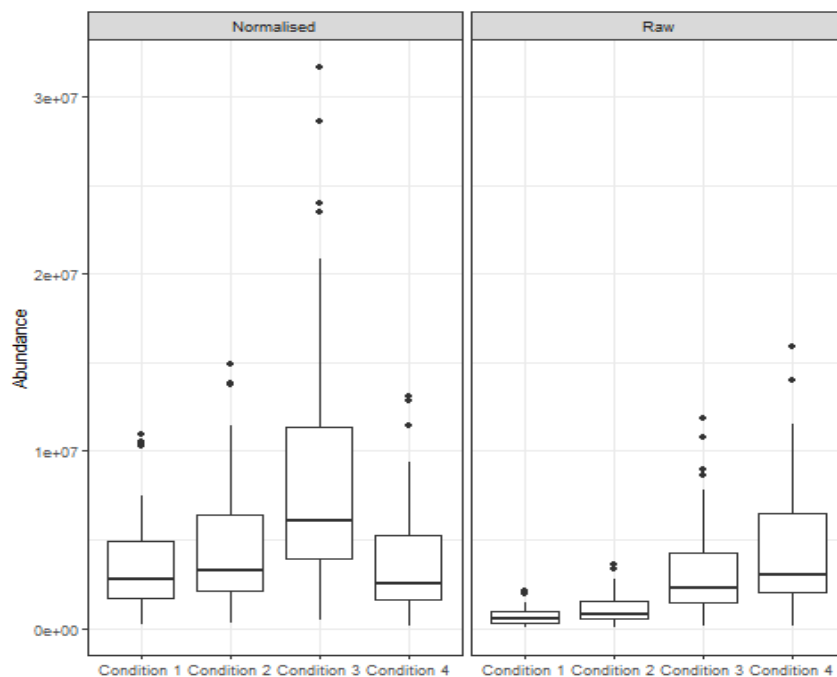


Figure 2.3.12; Comparison of the Progenesis QIP normalised and raw spike-in protein abundances for the dataset PXD002099

Raw abundances and the Progenesis QIP normalised abundances for this dataset are compared in Figure 2.3.12; spike-in protein levels were affected by the normalisation procedure and the spike-in signal was lost. The process of normalisation aims to adjust for systemic bias introduced by the separate

processing of sample runs while preserving the signal of interest. With dataset PXD002099 this appears to have failed. Figure 2.3.13 shows spike-in protein abundances after normalisation and the raw protein abundances for datasets PXD001385 and PXD001819, where the spike-in signal is preserved.

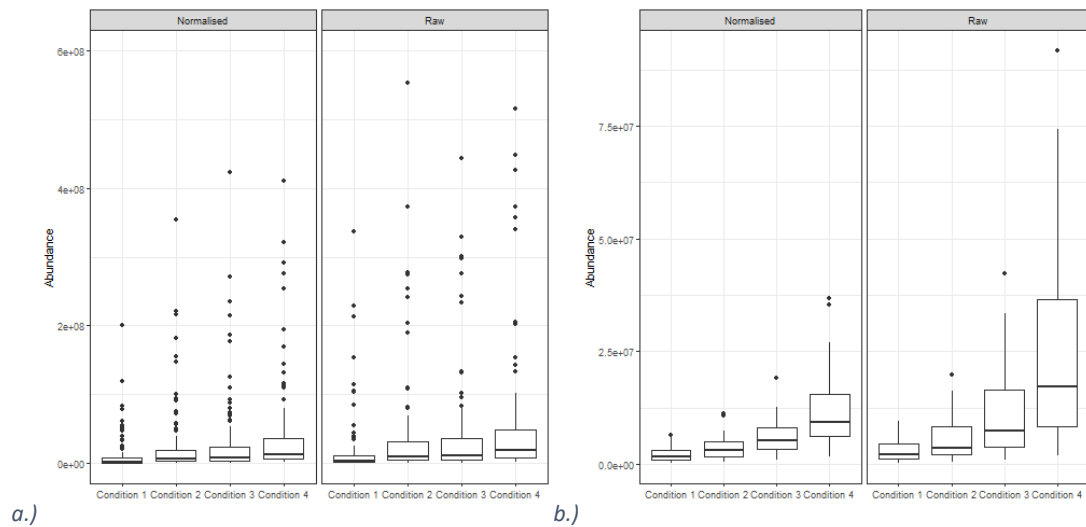


Figure 2.3.13; Comparison of the Progenesis QIP normalised and raw spike-in protein abundances for the datasets a.) PXD001385 and b.) PXD001819

The analysis of dataset PXD002099 was repeated using raw peptide abundances and the inbuilt quantile normalisation methods within QPROT and MSstats shown in Figure 2.3.14. The results were still poor, particularly in the low fold change simulations, but showed much improvement due to the alternative normalisation procedure. This highlights the importance of applying correct normalisation for the data which will be investigated further in Chapter Three.

## Optimising the statistical pipeline for quantitative proteomics

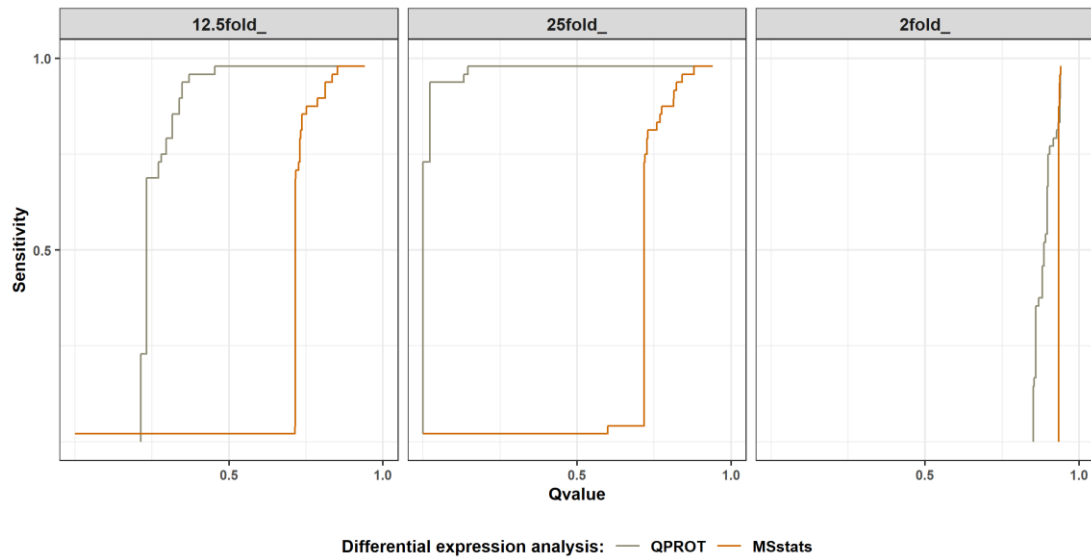


Figure 2.3.14; Smoothed Precision-Recall plots of reanalysed dataset PXD002099 using quantile normalisation features of MSstats and QPROT

Progenesis normalisation caused the signal in dataset PXD002099 to be lost due to problems with accuracy and precision in the data. In two of the conditions, over 50% of the technical replicates had an unacceptable amount of difference between them (Figure 2.3.9). Furthermore, the fold change values were not accurate (Figure 2.3.10) there was a mean 7-fold change when there should have been a 25-fold change, a mean 4-fold change when the expected was a 12.5-fold change, and in the 2-fold change, some of the values did not change at all. None of the background proteins remained acceptably constant across the conditions (Figure 2.3.11); some of them more than doubled in abundance. This analysis demonstrates the potential unreliability of the data used for benchmarking. In DE analysis using a *t*-test and a significant *p*-value of 0.05 (Figure 2.3.8), the 2 fold condition of dataset PXD002099 fails to identify any of the spike-in proteins with the 25- and the 12.5- fold change simulations accruing false positive rates of 84% and 91% respectively. However, this dataset has been used as part of benchmarking of statistical analysis by Tang et al. (2019b), Välikangas et al. (2018) and (Välikangas et al., 2017). Milac et al. (2012) raised the question as to whether spike-in data is appropriate for benchmark case-control studies after finding the introduction of inaccurate protein quantities in the Clinical Proteomic Technologies for Cancer (CPTAC) data (Paulovich et al., 2010).



## **2.4. Conclusions**

The benchmarking exercise in this chapter revealed two DE analysis methods with superior performance compared to the *t*-test; linear modelling using MSstats and Bayesian inference using QPROT. The results showed that parameter options for peptide selection and protein summarisation did not have a great impact on results, but summing peptide ions with the same charge state and summing all peptides mapped to the protein appeared to be the most reliable choice of options. More importantly, applying the correct normalisation procedure appeared to rescue failed analysis. This previously underappreciated factor in the DE pipeline will be investigated further in the next chapter.

The benchmarking exercise also revealed a number of flaws in the standard benchmarking procedure. The validity of the results is based on the precision and quality of the ground-truth data. However, benchmarking datasets are not always reliable and require their own validation methods. It also highlighted that assessing significance using an arbitrary cut-off criterion may not result in optimal software performance. In the following chapter, we define and test a proteome profiling-based alternative method for statistical analytical evaluation, as well as illustrate its potential to analyse biological data using experimental results from three cancer research.

## ***Chapter 3. Differential expression analysis evaluation by pathway enrichment***

### **3.1. Introduction**

#### **i. Abstract**

The use of spike-in benchmarking experiments is the established approach for validation of proteomics statistical analysis. Mimicking biological experiments, spike-in data allows 'ground truth' knowledge to be used as an evaluation metric to assess and compare how well software can detect proteins changing across conditions. However, problems with artificial data, such as sample preparation and the absence of inherent biological variation, cause problems with the comprehensiveness of a benchmarking experiment. Designing a benchmark dataset that accurately reflects real biological data is difficult as the proportion of true changing proteins in a biological scenario is not known. Moreover, authentic biological data cannot be used to benchmark statistical analysis in the traditional way as we do not know the true number of changing proteins with which to calculate sensitivity and specificity. The inconsistent results obtained in Chapter Two highlight these problems, and an analysis of the accuracy and precision of the benchmarking data demonstrates the unreliability of community-approved resources.

The aim of quantitative proteomics experiments is to deliver insight and gain a better understanding of biological processes through the discovery of many functionally related groups within the changing proteome. In this chapter, we describe and validate a novel method for statistical analysis evaluation utilising proteomic profiling through enrichment analysis and demonstrate its ability to analyse biological data with experimental results from three cancer studies. We also investigate the effect of the normalisation method using the same evaluation. Overall the results showed no consensus on best method for DE, normalisation method or threshold cut-off and the correct combination of parameters appeared to be dependent on the characteristics of the individual datasets.

## **ii. Technical and statistical issues with using ground-truth data**

The use of spike-in benchmarking experiments is the established approach for validation of proteomics statistical analysis. Along with quantitative measures of performance, community-approved benchmarks can be tremendously useful for proper method comparison (2015). Fair and objective comparisons require the availability of a standard dataset that returns the same answers for every user (Yates et al., 2012). The aim of such datasets is to mimic a real-life biological experiment containing background proteins, simulating those that are unrelated to the biological question and remain the same across samples, and spike-in proteins that represent those that are affected by the conditions of the experiment and change in concentration.

Historic software assessment relies on the need for 'ground truth' knowledge to apply an evaluation metric and utilises artificial benchmarking datasets generated to simulate real-life biological situations. The aim of DE benchmarking is to assess and compare how well methods correctly detect the spike-in proteins without incorrectly claiming background proteins are changing. The process is thought of as the gold standard evaluation technique, but the application of a metric derived from artificial data to a method that processes biological data is not a comprehensive assessment. Problems can arise due to the availability, adequacy, and selection bias of testing datasets (Gatto et al., 2016), and human error in sample preparation is a major source of error in protein quantitation (Zhang et al., 2009). In benchmarking datasets, background protein levels are not always kept constant, spike-in concentrations are not always precise, and there is often variation between sample replicates. This creates problems with evaluation where comparing the expected log fold changes of both spike-in and background proteins to the measured values is essential when assessing quantification accuracy (Tang et al., 2019b). The impact of these inconsistencies in sample handling and sample loading is rogue intensities and inaccurate fold-change values, leading to false positives or negatives.

## *Optimising the statistical pipeline for quantitative proteomics*

Flaws occur with experimental design; low intensity spike-in proteins can take the 'ground truth' being searched for below the detection limit, therefore skewing the evaluation. Also, the spike-in proteome is frequently a minor proportion of the sample global proteome in benchmarking data. Due to dynamic range and sequencing speed, more abundant peptides are more successfully identified. Consideration should be given to the sparsity of the data; a low number of spike-in proteins compared to background proteins limits scope. A small number of regulated proteins impairs the statistical power of analysis and does not generate enough data points to derive valid statistics (Kuharev et al., 2015). Also, spike-ins are often only up-regulated (all proteins increase in the same direction across samples), which is not true in a biological scenario. While it is important to use data with an adequate proportion of DE proteins to background proteins, having too many proteins change in intensity can upset the normalisation procedure, which often assumes that most proteins remain unchanged across conditions. Designing a benchmark dataset that accurately reflects real biological data is difficult as the proportion of true changing proteins in a biological scenario is not known.

In simulated datasets, experiments consist of technical replicates where repeated samples from the same condition are processed through the spectrometer in separate runs. This data lacks the inherent biological variation expected in real experiments. A real sample would have a much wider range of abundance changes, whereas benchmarking datasets have fixed ratios and fixed total abundances of each group of spike in proteins. Non-conformity of intensities compared to biological situations results in a poor estimate of the true variance. Replicate measurements are for performance monitoring; they are not independent tests and are unsuitable for the thorough testing of the validity of a scientific hypothesis or providing evidence of the reproducibility of the results (Vaux et al., 2012). Robust conclusions are supported by reproducible, significant observations from independent experiments (Pulverer, 2012). Variances, variance heterogeneities, and mean-variance relationships differ from those actually observed in biological data (De Hertogh et al., 2010). Due to the need for 'ground truth', authentic biological data cannot

be used to benchmark statistical analysis in the traditional way as we do not know the true number of differentially expressed proteins from which to calculate sensitivity and specificity.

Chapter Two of this thesis investigates statistical methods with the established approach using three benchmarking datasets. The results of the DE analysis were inconsistent, and the accuracy of the ground truth data was investigated and problems were found with the accuracy and precision of the spike-in datasets. This analysis demonstrates the potential unreliability of the data used for benchmarking. In DE analysis using a *t*-test and a significant *p*-value of 0.05, the 2 fold condition of dataset PXD002099 fails to identify any of the spike-in proteins, with the 25- and 12.5-fold change simulations accruing false positive rates of 84% and 91%, respectively. However, this dataset has been used as part of benchmarking of statistical analysis by Tang et al. (2019b), Välikangas et al. (2018), and Välikangas et al. (2017). In this chapter, we describe and validate an alternative method for statistical analysis evaluation utilising proteomic profiling and demonstrate its ability to analyse biological data with experimental results from three cancer studies. We also investigate the effect of the choice of normalisation method using the same evaluation.

### **iii. Review of benchmarking studies**

Evaluations of LFQ software are performed using various metrics of optimal performance, counting the maximum number of quantified proteins or the smallest number of missing values. Assessments often focus on how precise or accurate the method is using ROC, volcano, and precision-recall plots or comparisons are made between the expected and measured logarithmic fold change (LogFC) of the proteins, all requiring the need for ground truth benchmarking data. To overcome the limitations in statistical power of spike-in data, whole proteome datasets have been developed, allowing the assessment of hundreds of changing proteins as opposed to tens of changing proteins. Increasing the number of regulated proteins provides a greater number of data points for statistical evaluation. Kuharev et al. (2015) created a dataset for DIA

## *Optimising the statistical pipeline for quantitative proteomics*

software benchmarking with pre-digested proteomes of three different species combined into two samples in different proportions. Analysis showed good in-sample variances between technical replicates and although the statistical power of the test was improved by increasing the number of identified and quantified proteins, the study concedes that the complexity of the proteome samples exceeds that of typical label-free datasets and that the relative expression changes simulated by the combination of exactly defined ratios provides a scenario unlikely to be observed in a natural sample. Further use of the whole proteome for evaluation is shown by Ting et al. (2011), who created a two-proteome model sample to accurately measure the extent of the interference effect in isobaric labelling for multiplexed proteome quantification. However, to our knowledge, there has been no use of whole proteome benchmarking in label free LC-MS/MS studies.

Based on the reproducibility-optimised test statistic (ROTS) for ranking genes in microarray studies (Elo et al., 2008), evaluation methods for LFQ have been developed that use a comparison of resampled analyses rather than ground truth. ROTS offers an alternative to spike-in studies by evaluating the reproducibility of the top-ranked DE proteins of a given method when repeated using bootstrapping and has been included in several assessments of proteomics DE analysis (Väläkangas et al., 2018), (Pursiheimo et al., 2015), (Tang et al., 2019b), (Suomi and Elo, 2017), (Wang et al., 2017). The limitations of this method are that it is dependent on sample size and there is ambiguity in the optimal significance threshold. Furthermore, reproducing the order of ranking does not necessarily confirm the correct order, as there is no relation to practical measures of how accurately the ranking has been performed. Therefore the ROTS method does not completely remove the need for ground truth. It is the aim of this chapter to develop a novel method for analysis that does not rely on the use of benchmarking data.

The aim of quantitative proteomics experiments is to deliver insight and gain a better understanding of biological processes. Proteins changing due to experimental conditions are likely to be functionally related, and it is expected that multiple proteins within a given pathway change at the same time. In their

study investigating the repeatability and reproducibility of proteomics experiments, Tabb et al. (2016) compared a pair of reference xenograph proteomes representing basal and luminal-B human breast cancer provided by the NCI Clinical Proteomic Tumour Analysis Consortium (CPTAC) on six mass spectrometers. The study aimed to test if various differential proteomics technologies see the same biological changes and found that although different methodologies produce different protein lists, pathway and network analysis of DE results was highly consistent across instruments. Kustatscher et al. (2019) discovered that proteins that function together tend to be up- and downregulated in similar patterns while assembling their protein covariation analysis data matrix, ProteomeHD. Detected relative changes in protein abundance between the conditions create differentially expressed proteomic profiles. These are evaluated with pathway analysis to identify enriched biological themes and to discover functional relations. In their proteomic analysis of colorectal cancer, Wiśniewski et al. (2012) used label-free quantitative analysis to gain insight into changes in signalling pathways due to gene mutations, noting that they are recognised drivers of cancer development. In their quantitative proteome profiling of the NCI-60 cell line, a widely used panel for the study of cellular mechanisms of cancer in multiple tissues, Gholami et al. (2013) found that hierarchical clustering of differentially expressed proteins provided enriched biological functions and biochemical pathways associated with the tissue of origin. We can rarely attribute discrete biological functions to an isolated protein, and instead we must look at the functional properties of groups of components to describe them (Hartwell et al., 1999). Quantitative proteomics gives us a picture of protein-directed biological processes that can directly lead to understanding biological mechanisms (Ong and Mann, 2005). In their study of the protein-protein interaction networks between the yeast *Saccharomyces cerevisiae* and the bacterium *Helicobacter pylori*, Jeong et al. (2001), quantitatively demonstrated that the few most highly connected proteins in a network are the most essential, mediating interactions with numerous less connected proteins.

A desirable outcome of a quantitative proteomic experiment is the discovery of many functionally related groups within the changing proteome. The use of pathway analysis methods allows us to provide meaning to DE data analysis by attaching biological context to important isolated genes that are detected by their relation to other genes within the results (García-Campos et al., 2015). Based on this, in a more accurate DE analysis, we would expect there to be a higher number of proteins that are known to be linked through functionality. In this chapter, we use this principle to develop a novel method of benchmarking without the use of ground truth data. This concept has been introduced in previous studies; Stewart et al. (2017) use the pathway content from known SCC biomarkers produced by each of the methods as a way of measuring comparison. In their assessment of their ROTS for ranking genes in microarray studies, Elo et al. (2008) included a case study of DE analysis of peripheral blood lymphocytes from asthma patients. As the ground truth of this dataset was not known, performance was assessed using established experimental evidence. However, the examples given use known biomarkers from pilot studies, and an evaluation method without prior knowledge has yet to be created. By using existing knowledge of which proteins work together for particular functions, we provide an assessment of DE analysis based on the biological picture. Using the results of pathway analysis as a metric for successful statistical evaluation, we propose a novel technique for benchmarking quantitative proteomic experiments.

#### **iv. Normalisation**

In Chapter 2, the results of a failed DE analysis were partially rescued by providing an alternative normalisation methods. The separate processing of sample runs in LF-MS/MS introduces systemic bias due to experimental or technical conditions. Normalisation methods, described in Chapter 1, aim to remove variation caused by environmental conditions, sample preparation, or instrument calibration, while preserving true over- and under-expression. This chapter investigates the impact of normalisation methods on DE analysis, with the aim of discovering the optimal technique. A variety of claims have been



made regarding the best method of normalisation; Välikangas et al. (2016) found that VSN provided the biggest reduction in variation; Tokareva et al. (2021) preferred scaling methods; Wang et al. (2013) describe the success of regression models for normalising metabolomics data. However, there does not appear to be a consensus approach. Assumptions of statistical methods may be the cause; that DE between samples is symmetrical about zero, that there is no relationship between DE and expression abundance, and that variation in expression is the same in each sample. Some normalisation methods can be prone to overfitting or introducing bias into the data (Dabney and Storey, 2007b). In proteomics data analysis, normalisation usually occurs in a separate step prior to DE analysis. In their review on the normalisation of LC-MS proteomics data, Karpievitch et al. (2012) describe the need for techniques to be flexible enough to remove appropriate bias but delicate enough not to remove any biological signal of interest. Dabney and Storey (2007a) describe how the strict assumptions required for conventional smoothing and global adjustment normalisation can cause spurious false positives and alteration of the degree and direction of true DE proteins, proposing a flexible model with more general assumptions. Park et al. (2003) found global adjustment to cause issues, but a similar performance was achieved with both linear and nonlinear methods. Interestingly, Callister et al. (2006) found that some normalisation techniques removed significant differences between the datasets. This was confirmed by Mecham et al. (2010), who also found that normalisation can 'introduce signal in the presence of asymmetrical biological variation', and that the assumption that most genes are not differentially expressed is false in many settings. In their review of normalisation for proteome analysis workflows, O'Rourke et al. (2019) conclude that among the host of algorithmic techniques, no single algorithm is effective for all types of data.

## **v. Aims of chapter**

In Chapter Two we followed an established approach for evaluating DE analysis software using three spike-in benchmarking datasets. Due to the inconsistency of the results, the accuracy and precision of the simulated experiment was investigated. The analysis highlighted a lack of precision across technical replicates, inaccurate fold-change values and inconsistent background protein abundances, meaning the software evaluation produced from using such data would be flawed. A review of the literature revealed further problems with the established method of using artificial spike-in data as a 'ground truth' including sparsity of data, expression in only one direction, and lack of inherent biological variation and biological replicates. An alternative approach to evaluating the DE analysis software was sought but the literature review did not reveal any method that did not rely on 'ground truth'. Therefore, in this chapter we aim to develop a method to provide evaluation of statistical techniques for quantitative proteomics that is transferable to real life experimental conditions.

## **3.2. Methods**

### **i. Biological datasets**

A literature search was performed to identify label-free experiments with a large number of identified proteins. From these studies, those with a large proportion of differentially express proteins were selected for analysis. Raw files from three biological datasets were downloaded via the PRIDE partner repository from the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org/>) (Vizcaíno et al., 2014). A summary of the experimental conditions for each of the benchmarking datasets is given in Table 12.

## Optimising the statistical pipeline for quantitative proteomics

Table 12; Summary of experimental conditions for each of the benchmarking datasets giving details of samples used, experimental design, parameters used for MS analysis and number of proteins

	PXD004501	PXD004682	PXD007592
Species	Human ascite fluids	Human lung tissue	Cerebral metastases
Disease	Gastric cancer	Lung cancer	Malignant melanoma
Experimental sample	Stage IV gastric cancer ascites	Squamous cell carcinoma	MAPKi therapy good responders
Number of experimental samples	3	4	5*
Number of technical replicates	1	2**	2
Control sample	Liver cirrhosis ascites	Tumour-adjacent tissue	MAPKi therapy poor responders
Number of control samples	3	3	13
Number of technical replicates	1	2	2
Number of fractions	6	12***	NA
Enzyme	Trypsin	Trypsin	Trypsin
Instrument	Q Exactive Plus Orbitrap	Q Exactive Plus Orbitrap	Q Exactive Plus Orbitrap
Search database	UniProt Human	UniProt Human	UniProt Human
Peptide tolerance	15 ppm	10 ppm	5 ppm
MS/MS tolerance	20 mmu	0.5 Da	20 mmu
Fixed modifications	Carbamidomethylation (C)	None	Carbamidomethylation (C)
Variable modifications	Oxidation (M)	Carbamidomethylation (C), Oxidation (M)	Oxidation (M), Acetyl (Protein N-term)
No. identified protein groups****	1719	6656	5977

\* Only 3 were used in analysis due to processing issues

\*\* One sample's technical replicate was discarded due to poor alignment

\*\*\* Only 9 were used in analysis due to processing issues

\*\*\*\* Number of proteins is based on analysis from the publishing paper

## ***Optimising the statistical pipeline for quantitative proteomics***

### ***Ascites from stage IV gastric cancer patients compared to liver cirrhosis patients -PXD004501 (Jin et al., 2018)***

The presence of malignant ascites in patients is associated with peritoneal seeding from incurable distant metastasis which has an extremely poor prognosis. Present diagnostic characterisations of ascites in gastric cancer are limited. The study aimed to develop understanding of the pathophysiology of peritoneal seeding which will have critical clinical implications in diagnosis, choice of treatment and active surveillance. Label-free quantitative proteomics methods were employed to identify candidates to differentiate between malignant and benign ascetic fluids and identified 299 differentially expressed proteins between the ascites fluids of the liver cirrhosis and gastric cancer patients.

### ***Lung tissue, squamous cell carcinoma compared to adjacent tissue - PXD004682 (Stewart et al., 2017)***

In an investigation to see if increased instrument time was justified for increasing protein identification, this study examined the utility of different methods of sample preparation and MS data acquisition in lung tumour proteomes using greatest coverage as a metric. Alternative methods for large scale studies were investigated to provide information for planning large-scale experiments where acquisition time is an important issue.

Four datasets were used to make two comparisons. In a single-sample comparison, LC-MS/MS with data-dependent acquisition was compared to LC-MS/MS with data-independent acquisition. The total acquisition time for both experiments was 2 days, with 3409 and 2219 proteins groups being discovered, respectively. In a fractionated experiment, label-free quantification (24 days acquisition time) was compared to relative quantification through chemically labelled peptides (tandem mass tags (TMT)) which took only 6 days to acquire the data. Label-free techniques generated 6656 unique protein groups whereas tagging methods identified 5535. The conclusion of the authors was that single-sample experiments should be used as rapid tissue assessment tools, digestion

quality control or when there are limited quantities of clinical samples with LF or TMT methods recommended when larger amounts of tumour tissue was available. The data from the label-free fractionated experiment was downloaded and used for benchmarking in this study.

***Cerebral metastases from good responders to MAP kinase inhibitor (MAPKi) therapy compared to poor responders - PXD007592 (Zila et al., 2018)***

Brain metastases are the main cause of mortality in advanced melanoma patients with mutation of the BRAF gene present in approximately half of these cases (Cheng et al., 2018). Mitogen-activated protein kinase inhibitor (MAPKi) therapy can result in complete remission in more than 50% of these patients (Dummer et al., 2014), but resistance to the drug quickly develops. The study aimed to understand the mechanisms of resistance in order to identify possible alternative therapeutics. LC-MS analysis was performed on cerebral metastases from melanoma patients with samples classed as good responders to MAPKi therapy (progression-free survival  $\geq 6$  months) compared to poor responders (progression-free survival  $\leq 3$  months) of the treatment. 5977 proteins were identified with 1636 proteins significantly changing in concentration between conditions.

## **ii. Imputation**

To assess the impact of imputation on DE analysis, t-test analysis with a BH corrected p-value of 0.05 to indicate significance was used and number of DE proteins was compared for each dataset using the following methods to deal with zero values:

- Imputation of  $1 \times 10^{-7}$
- Imputation of  $1 \times 10^{-10}$
- Removal of proteins with any zero values from the analysis

Ions with the same primary peptide sequence and different charge states' intensities were summed. Ions with the same primary peptide sequence but

with artefactual modifications were treated separately. Protein quantification was based on the sum of the average intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptides were selected.

### iii. Differential expression evaluation

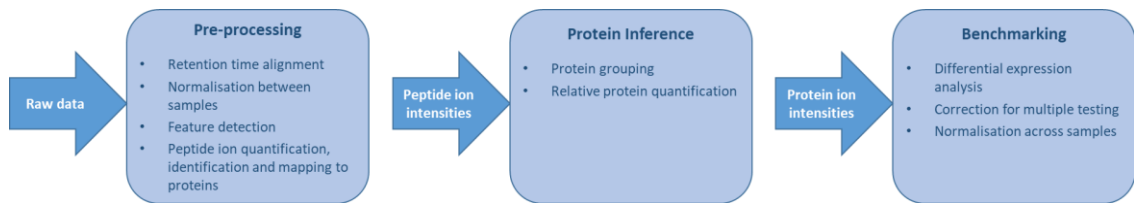


Figure 3.2.1; Processes completed in the three stages of benchmarking workflow with details of input data required.

Figure 3.2.1 describes the three main stages in the evaluation; pre-processing of the raw data was performed using Progenesis QI for Proteomics (QIP). Protein grouping and quantification performed using protein inference software written in Java as described in Chapter Two. Methods for DE analysis, normalisation and correction for multiple testing significance thresholds were benchmarked using pathway analysis results as a metric to evaluate performance.

#### **Quantitative processing**

The .raw files of binary data were processed with Progenesis QIP v4.2. as described in Chapter Two, and according to the parameters described in Table 12. Identified and normalised peptide abundances were exported from Progenesis QIP for post processing with benchmarking protein inference pipeline also described in Chapter Two. Ions with same primary peptide sequence and different charge state's intensities were summed. Ions with same primary peptide sequence but with artefactual modifications were treated separately. Protein quantification was based on the sum of the average

intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptide were selected.

### ***Differential expression analysis***

Protein abundances were tested for DE using the following methods previously described:

- QPROT
- Welsh two sample *t*-test
- MSstats\*

\*DE analysis in MSstats is performed at the peptide intensity stage, therefore peptide intensities were exported from the protein inference software for processing. Protein grouping was kept consistent, but protein quantification was performed with MSstats through their summation of all non-conflicting peptides parameter option.

### ***Defining significant results***

Adjustments were made to the resulting test statistics to allow for correction for multiple testing using the inbuilt FDR calculation in the QPROT software and the Benjamini Hochberg calculation for adjusting *p*-values (BH *p*-value) (Benjamini and Hochberg, 1995). A range of cut-off values were used as a threshold of significance Table 13.

*Table 13; Summary of significance threshold ranges. Proteins passing the threshold are categorised as DE and go forward for enrichment analysis. Size of increment increases over iteration.*

<b>DE method</b>	<b>Correction for multiple testing</b>	<b>Range</b>	<b>Increment</b>
QPROT	FDR	0.0001 - 0.05	0.0001 (0.0001 – 0.0009)
			0.001 (0.001 – 0.009)
			0.01 (0.01 – 0.05)
Welsh's two sample <i>t</i> -test	BH <i>p</i> -value	0.001 - 0.05	0.001 (0.001 – 0.009)
			0.01 (0.01 – 0.05)
MSstats	BH <i>p</i> -value	0.001 - 0.05	0.001 (0.001 – 0.009)
			0.01 (0.01 – 0.05)

## Optimising the statistical pipeline for quantitative proteomics

A review of statistical methods employed by the packages is included in Chapter 1. Multiple testing correction was performed in QPROT using their calculation of false discovery rate and with Benjamini Hochberg adjusted  $p$ -values in the  $t$ -test and MSstats. The normalisation stage in QPROT and MSstats processing was switched off for accurate comparison of performance at DE analysis stage. A detailed schematic of the benchmarking workflow is shown in Figure 3.2.2.

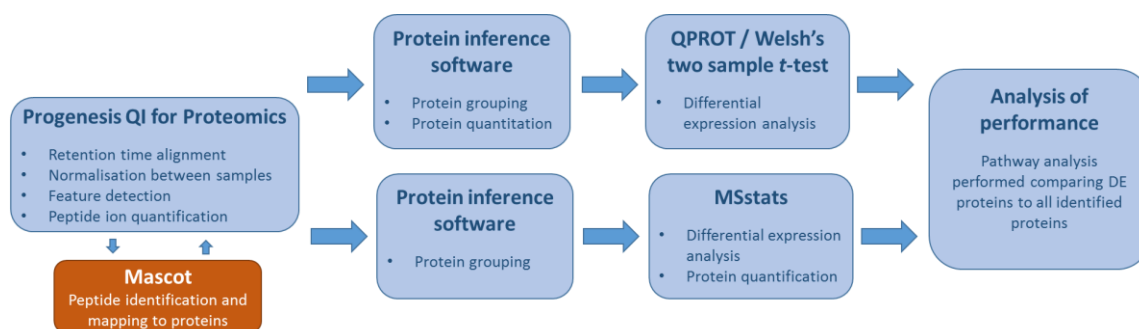


Figure 3.2.2; Schematic of the benchmarking pipeline. Raw data is processed with Progenesis QIP and peptide ions are identified using Mascot. Resulting peptide abundances are summarised into protein group abundances using the benchmarking inference pipeline. DE analysis is performed and significant results at a variety of significance thresholds are evaluated using enrichment analysis terms.

### Analysis of performance

Analysis of performance of DE analysis was conducted using enrichment analysis through the DAVID, the Database for Annotation, Visualisation and Integration Discovery (<http://david.niaid.nih.gov/>) using the R package RDAVIDWebService (Fresno and Fernández, 2013). Pathway analysis was performed using proteins classified as significantly changing in concentration between conditions as the search list. A range of thresholds was used to define the differentially expressed proteins in order to investigate the optimum cut-off point and to minimise the trade of between calling false positives and creating false negatives (Table 13). All identified proteins in the dataset were used as the background list. It was expected that proteins subject to change under the experimental conditions would have more known relations within pathways compared to the pathway analysis of the whole group of identified proteins in the data. Significant terms were defined as those with a BH  $p < 0.05$  from GO; Biological Process, Molecular Function, and Cellular Component, and the Reactome and Kegg pathways. A higher number of significant terms within the



differentially expressed group of proteins was used as a metric to determine the most effective method and significance threshold for DE analysis.

#### **iv. Evaluation of pathway analysis benchmarking method**

In order to validate the use of pathway analysis for benchmarking significant results from the *t*-test (*p*-value threshold < 0.05) analysis of dataset PXD004682 was evaluated with pathway analysis using different proportions of proteins said to be significantly changing. On each iteration, 5% of true DE proteins were removed from the pathway analysis foreground list and replaced with proteins from the background list, simulating analysis with increasing false positive rates. This process was repeated 10 times.

#### **v. Normalisation across samples**

*Table 14; Normalisation procedures performed by Normalyzer package with details of R functions and packages employed*

<b>Normalisation</b>	<b>Abbreviation</b>	<b>R function</b>	<b>Package</b>
Loess normalisation	(Loess)	<code>normalizeCyclicLoess()</code>	limma
Robust linear regression normalisation	(RLR)	<code>rlm()</code>	MASS
Total intensity normalisation	(TIN)	-	NormalyzerDE
Median intensity normalisation	(MIN)	-	NormalyzerDE
Average intensity normalisation	(AIN)	-	NormalyzerDE
Variance stabilisation normalisation	(VSN)	<code>justvs()</code>	VSN

For investigation into the effect of normalisation, data was processed according to the method described in Section 3.2, using raw peptide ion abundances from the Progenesis QIP output using a minimum of one unique peptide for identification. Normalisation was applied to quantified and identified protein group abundances. As MSstats requires peptide abundances as input and performs its own protein quantification, it was excluded from the normalisation benchmarking to ensure an accurate ‘like-for-like’ comparison of protein

abundances. Raw protein intensity data was processed with the open-source R tool 'Normalyzer' (Chawade et al., 2014), which provided the normalisation methods described in Table 14. In all instances, 'global normalisation' was performed across all samples, where Normalyzer applies the method across all samples regardless of experimental groups, utilising the assumption that most protein levels will remain the same across conditions. This is as opposed to the 'local normalisation' option offered by Normalyzer, wherein the replicate groups are normalised separately. Two methods for quantile normalisation were written in R: Quantile normalisation using mean values; the data was sorted, and the mean quantile intensity was substituted as the protein intensity value. The data was resorted into the original order, with the transformed data now having the same distribution with identical quantiles and the original ranking preserved. Quantile normalisation was also performed using median values rather than mean values. The in-built normalisation methods of the software packages Progenesis QIP and QPROT were also compared.

Prior to DE analysis, all data were subjected to log transformation according to the process detailed in Figure 3.2.3. DE analysis results were evaluated using the pathway analysis method described in 3.2. Due to the calculation of significance in enrichment analysis, larger sets provide smaller  $p$ -values, meaning it is possible for terms to have a significant  $p$ -value with a small amount of enrichment. To account for this, and a minimum of two-fold enrichment threshold was also applied to identify significant enrichment terms.



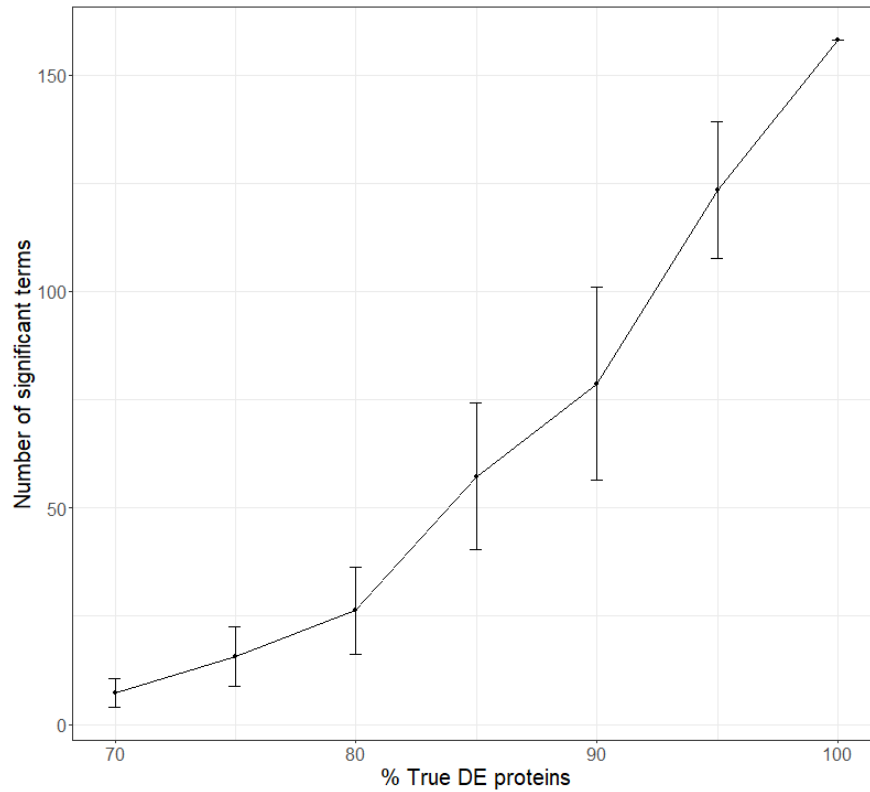
### 3.3. Results and discussion

#### i. Pathway analysis benchmarking validation

*Table 15; Pathway analysis validation completed with dataset PXD004682. Different proportions of true DE proteins and randomly selected background proteins. Number of significant terms from DAVID enrichment analysis of DE proteins from t-test analysis, (Benjamini Hochberg adjusted p-value < 0.05).*

<b>Proportion DE proteins</b>	<b>Proportion of background proteins</b>	<b>Mean number of terms</b>	<b>Standard deviation</b>	<b>Standard error</b>
100	0	158	0	0
95	5	123.4	25.35	15.71
90	10	78.7	35.98	22.30
85	15	57.3	27.33	16.94
80	20	26.3	16.12	9.99
75	25	15.6	11.06	6.85
70	30	7.3	5.42	3.36

To validate the pathway analysis method, enrichment analysis was performed on foreground data that included an increasing amount of proteins from the background group (i.e. not classified as DE by *t*-test analysis using BH *p*-value of < 0.05 as significant), to simulate analyses that are incorrectly classifying proteins as DE when they are not. The results (Table 15 and Figure 3.3.1) demonstrate that by including an increasingly large proportion of not ‘true’ DE proteins in the search list, fewer significant terms are returned on functional enrichment analysis. Proteins that had been classified by statistical analysis as changing between conditions were more functionally similar to each other compared to the total pool of identified proteins. These results validate the pathway enrichment testing approach as an effective evaluation tool. Precise DE analysis produces more significant terms, and an increased false discovery rate ‘dilutes’ the quality of the enrichment analysis.



*Figure 3.3.1; Pathway analysis validation completed with dataset PXD004682. Number of significant terms (*t*-test analysis, Benjamini Hochberg adjusted *p*-value < 0.05) from DAVID enrichment analysis with different proportions of significantly differentially expressed proteins included in the search list repeated 10 times*

## **ii. Imputation**

In the biological dataset, the amount of proteins affected by missing abundance data ranged from were 3.3%, 1.6%, and 4.35% for datasets PXD004501, PXD04682, and PXD007592, respectively (Table 16). Changing the methods used to deal with the missing values had no effect on the total number of DE proteins in the analysis or the number of resulting significantly enriched pathways.

## Optimising the statistical pipeline for quantitative proteomics

Table 16; Comparison of effects of imputation. *t*-test analysis with a BH corrected *p*-value of 0.05 to indicate significance was used and number of DE proteins was compared for each dataset using different methods to deal with zero values. Ions with same primary peptide sequence and different charge state's intensities were summed. Ions with same primary peptide sequence but with artefactual modifications were treated separately. Protein quantification was based on the sum of the average intensities of all unique or resolved peptides that have been mapped to the protein, and proteins identified by one or more unique peptide were selected. Enrichment analysis was performed using the DAVID webpage (<https://david.ncifcrf.gov/tools.jsp>) using the default parameters. Impute 1e-07, imputation of 0.0000001; Impute 1e-10, 0.000000001; Remove, removal of proteins with any zero values from the analysis.

Dataset	Number of proteins with a zero value	% of total proteins	Imputation method					
			Impute 1e-07		Impute 1e-10		Remove	
			No. DE proteins	No. enriched terms	No. DE proteins	No. enriched terms	No. DE proteins	No. enriched terms
PXD004501	19	3.30	46	4	46	4	46	4
PXD004682	45	1.60	1226	26	1225	26	1225	26
PXD007592	135	4.35	9	2	9	2	9	2

### iii. Differential expression evaluation

#### Summary of data analysis

Table 17; Summary of DE analysis by QPROT, *t*-test, and MSstats of biological datasets PXD004501, PXD001682, and PXD007592. Protein group abundances were calculated using protein inference benchmarking software from Progenesis QIP normalised peptide ion abundances that had been identified by a minimum of one or two unique peptides.

Dataset	Minimum number of unique peptides for identification	Number of identified protein groups	Number of differentially expressed protein groups		
			QPROT	<i>t</i> -Test	MSstats
			FDR < 0.01%	BH adj.p < 0.05	BH adj.p < 0.05
PXD004501	1	574	224	46	118
	2	542	187	40	110
PXD004682	1	2813	1044	1226	1329
	2	2716	595	1206	1314
PXD007592	1	3098	69	9	73
	2	3000	104	9	66

A summary of the results from the DE analysis (Table 17) shows PXD007592 had the largest number of identified proteins (3098, 3000). There was a small proportion of proteins identified as being differentially expressed by all methods (9 – 104 DE proteins). Dataset PXD004682 had a relatively high proportion of changing proteins compared to the total amount of protein groups identified (up to half of all proteins). PXD004501 had a relatively small

number of identified protein groups (574, 542) compared to the other two data sets, but a higher proportion of those were identified as differentially expressed compared with PXD007592. The criteria for the number of unique proteins required for confident identification made gave approximately 5% difference in the count of protein groups identified in all datasets.

#### **iv. Evaluation of DE methods by pathway analysis**

##### ***Overall analysis***

Overall, there was no consistently best method for DE analysis (summarised in Figure 3.3.2). MSstats achieved the best result on the small dataset with a large amount of proteins changing. However, all three DE methods were similar in performance. The *t*-test was best when there was a large dataset with a large proportion of changing proteins. MSstats' performed similarly to the *t*-test, but with QPROT producing relatively few significant enrichment terms. In the largest dataset with smallest proportion of changing proteins, QPROT excelled compared to *t*-test, with MSstats not providing any significant terms at all. There was a large variation in optimal significance thresholds and differences between how many proteins each method called as DE at their optimal performance level.

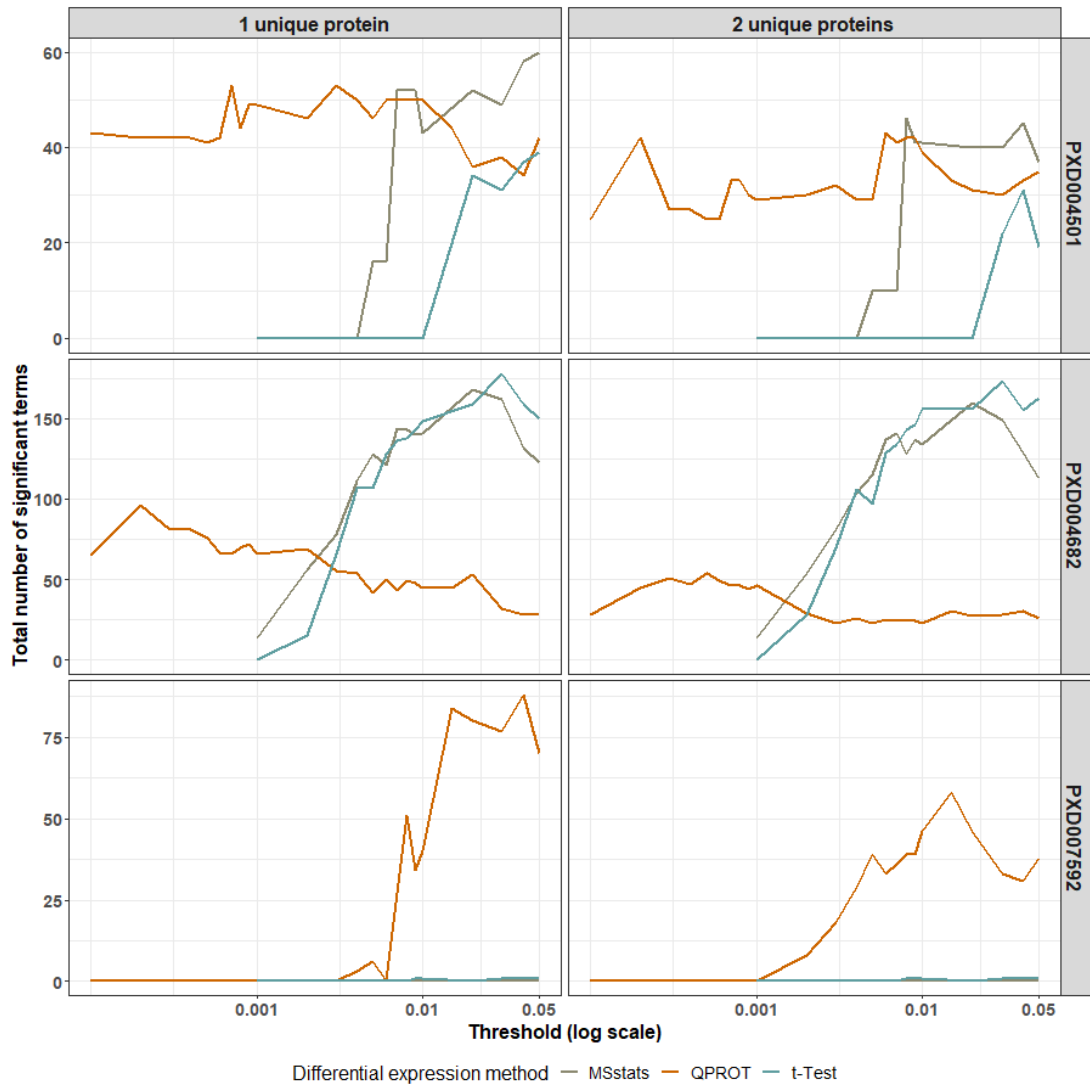


Figure 3.3.2; Differential expression analysis of the datasets PXD004501, PXD004682, PXD007592 using statistical methods MSstats, QPROT and Welch t-test. Number of significant terms (Benjamini Hochberg adjusted  $p$ -value  $< 0.05$  with DAVID analysis) shown for different threshold levels for significant differentially expressed proteins. Threshold values are Benjamini Hochberg adjusted  $p$ -values for MSstats and t-test and FDR for QPROT analysis. Proteins groups identified with 1 and 2 unique proteins

### PXD004501

This dataset had the smallest number of identified protein groups, 574 when using a minimum of one unique peptide for identification. It also had the largest range of proportion of proteins identified as differentially expressed across the methods; 44 with QPROT FDR  $< 0.01$  and 203 and 118 with  $t$ -test and MSstats at BH  $p$ -values  $< 0.05$ . QPROT provided a high number of significant terms at low threshold levels, over 50 terms at FDR  $> 0.001$ . As the significance threshold increased, MSstats performance was similar to QPROT. For 1 unique peptide, MSstats gave the highest number of total significant terms of 60 at a



## ***Optimising the statistical pipeline for quantitative proteomics***

significance threshold of  $p = 0.05$ , calling 118 proteins as DE, compared to a maximum of 46 terms from *t*-test (calling 39 proteins as DE at significance threshold of  $p = 0.05$ ) and 53 for QPROT (calling 178 and 205 proteins as DE at significance thresholds of 0.0007 and 0.003 FDR). When using 2 unique peptides for identification MSstats gave the most terms (46) calling 18 proteins as DE at significance threshold of  $p = 0.008$ , with 43 maximum terms from QPROT (calling 176 proteins as DE significance threshold of 0.006 FDR) and 31 from *t*-test (calling 31 proteins as DE cut off threshold  $p = 0.05$ ).

There was a small difference in the maximum number of terms produced by the different analysis methods, 39, 56, and 60 for protein data identified by one unique peptide and 31, 43, and 46 for protein data identified by at least two unique peptides. However, the cut-off significance level required to achieve these results was varied. MSstats gave the best performance at the conventional threshold level of 0.05 for protein data identified by one unique peptide (Table 18), and it was only 0.008 for protein data identified by at least two unique peptides.

*Table 18; Summary of 'best' analysis parameters where maximum number of terms are found during enrichment analysis of dataset PXD004501 using Progenesis normalised protein intensities.*

<b>Number of unique peptides required for identification</b>	<b>DE method</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Number of significant terms</b>
One unique peptide	MSstats	0.05	118	60
	QPROT	0.0007 / 0.003	178 / 205	56
	<i>t</i> -Test	0.05	46	39
Two unique peptides	MSstats	0.008	18	46
	QPROT	0.006	176	43
	<i>t</i> -Test	0.04	31	31

Stringent significance thresholds between 0.007 - 0.006 FDR were required for QPROT, with less conservative thresholds of 0.04 and 0.05 best for *t*-test analysis. Variation also occurred in the number of proteins being classified as DE for optimal results; MSstats had best performance by calling an extreme difference of 118 and 18 proteins DE, QPROT 178 or 205 and 176, and *t*-test just 31, for protein data identified by one unique peptide and 46 for protein data identified by at least two unique peptides.

**PXD004682**

This dataset had a large number of identified protein groups, 2813 when using a minimum of one unique peptide for identification. It also had the largest proportion of these protein groups being identified as differentially expressed by all of the methods; 1044 with QPROT FDR of less than 0.01 and 1226 and 1329 with *t*-test and MSstats at BH *p*-values of less than 0.05.

*Table 19; Summary of 'best' analysis parameters where maximum number of terms are found during enrichment analysis of dataset PXD004682 using Progenesis normalised protein intensities.*

<b>Number of unique peptides required for identification</b>	<b>DE method</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Number of significant terms</b>
One unique peptide	MSstats	0.02	1017	168
	QPROT	0.0002	707	96
	<i>t</i> -Test	0.03	1035	178
Two unique peptides	MSstats	0.02	1019	160
	QPROT	0.0005	387	54
	<i>t</i> -Test	0.03	1021	173

Overall the *t*-test gave the best results based on pathway analysis (Table 19), with a peak number of significant terms of 178 and 173 (both at a significance threshold *p*-value of 0.03) for protein data identified by one unique peptide and by at least two unique peptides, respectively. MSstats performance is similar to *t*-test, 168 and 160 significant terms (significance threshold *p*-value of 0.02) for protein data identified by one unique peptide and by at least two unique peptides, respectively. The number of DE proteins was also similar, 1017 and 1019 for MSstats and 1035 and 1021 for *t*-test for protein data identified by one unique peptide and by at least two unique peptides, respectively. Less effective was QPROT with only 96 and 54 significant terms. It performed better at low FDR thresholds (0.0002 and 0.0005 for protein data identified by one unique peptide and by at least two unique peptides, respectively) and classified relatively fewer proteins as DE (707 and 387 for protein data identified by one unique peptide and by at least two unique peptides, respectively).

**PXD007592**

This dataset had the largest number of identified protein groups, 3098 when using a minimum of one unique peptide for identification, but it had the smallest proportion of proteins identified as differentially expressed.

*Table 20; Summary of 'best' analysis parameters where maximum number of terms are found during enrichment analysis of dataset PXD007592 using Progenesis normalised protein intensities.*

<b>Number of unique peptides required for identification</b>	<b>DE method</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Number of significant terms</b>
One unique peptide	MSstats	-	-	-
	QPROT	0.04	137	88
	<i>t</i> -Test	0.009	2	1
Two unique peptides	MSstats	-	-	-
	QPROT	0.015	118	58
	<i>t</i> -Test	0.008	2	1

QPROT identified 104 proteins as being differentially expressed at the typical FDR threshold of 0.01 (Table 20). QPROT called the largest amount of proteins as being DE and also gave the best pathway analysis results of 88 and 58 significant terms for protein data identified by one unique peptide and by at least two unique peptides, respectively. Both *t*-test and MSstats analysis gave poor results for this data. With only 9 protein groups being identified as differentially expressed at 0.05 *p*-value threshold, *t*-test analysis only produced one significant result from pathway analysis, with 2 DE proteins at a significance threshold of 0.009 and 0.008 for protein data identified by one unique peptide and by at least two unique peptides, respectively. While at  $p < 0.05$  MSstats identified 73 and 66 proteins significantly changing across conditions for protein data identified by one unique peptide and by at least two unique peptides, respectively, the analysis method was unable to produce results that gave significant terms in pathway analysis at any significance threshold.

## **v. Investigating the effect of normalisation method**

### ***MSstats***

As the results of the DE analysis (Figure 3.3.2) showed that MSstats gave a similar performance to the *t*-test, and would require normalisation of peptide ion abundances rather than protein abundances, only QPROT and *t*-test analysis data was used for normalisation evaluation to ensure a comprehensive 'like-for-like' comparison. It is possible that this benchmarking exercise did not provide the best way to demonstrate MSstats. The mixed-effect modelling could be more effective in label-based workflows, where peptide intensities are compared to isotopically labelled reference spike-in intensities. Known as *blocking*, technical variation between samples can be estimated by investigating changes in the difference between reference and target intensities across runs (Oberg and Vitek, 2009). The mixed-effects model incorporates a term for this method to reduce bias and variation (Chang et al., 2012). This aspect of MSstats will be redundant in a label-free experiment, perhaps accounting for its similar performance to the *t*-test.

### ***Fold enrichment threshold***

Analysis in Figure 3.3.2 was conducted using only a BH  $p < 0.05$  to define significant terms. The investigation of the effect of additionally including a minimum threshold for enrichment is shown in Figure 3.3.3. The rationale for this is that for certain pathways or ontology terms for which a large number of proteins can be mapped, highly significant *p*-values might be obtained with trivially small enrichment factors, calling into question biological relevance.

## Optimising the statistical pipeline for quantitative proteomics

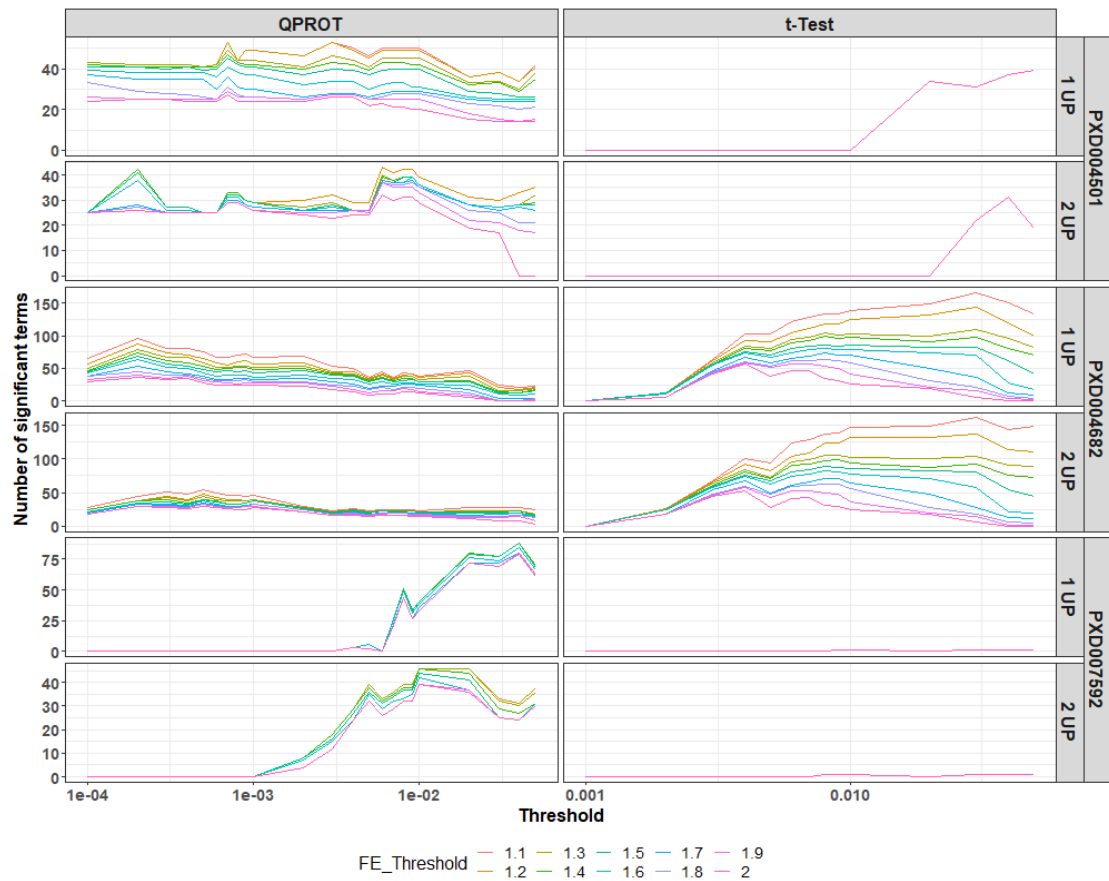


Figure 3.3.3; Differential expression analysis of the datasets PXD004501, PXD004682, PXD007592 using statistical methods QPROT and Welch's t-test. Number of significant terms (y-axis) colours represent the effect of different fold enrichment threshold cut-off values. Threshold values (x-axis) (Benjamini Hochberg adjusted p-values for t-test and FDR for QPROT analysis).

The analysis in Figure 3.3.3 demonstrates that there were a large number of terms assigned as significant that had small enrichment factors. As the significance cut-off threshold increased for the t-test analysis of PXD004682 (rows 3 and 4, column 2, Figure 3.3.3), the difference between the numbers of significant terms widened depending on the fold enrichment threshold used. In their report regarding sources of bias in functional enrichment analysis, Timmons et al. (2015) highlight the need to avoid using marginal enrichments being relied upon to drive the interpretation of an experiment. Therefore, to prevent expanded counts based on reporting terms with a small effect size, for the evaluation of normalisation, a combination of minimum 2-fold enrichment was combined with the BH  $p < 0.05$  for the identification of significant terms.

### **Overall analysis**

A summary of the overall results is shown in Figure 3.3.4. The different sub-plots along the x-axis show the different normalisation methods used and the sub-plots along the y-axis show the different datasets. Pathway analysis was performed to look for functional enrichment in the DE proteins compared to all of the proteins identified in the sample. The number of proteins in the DE group at each threshold is shown as a solid line, and the number of significant terms produced from the pathway analysis is shown as points. Results from QPROT analysis are shown in orange and results from *t*-test analysis are shown in blue.

The results are diverse; the effect of normalisation is different between datasets and DE analysis methods, and the optimal threshold for producing significant terms varies. There does not appear to be an obviously superior combination of methods; for example, QPROT DE analysis combined with quantile normalisation has excellent results with dataset PXD004501 (Figure 3.3.4; row 1, column 10), but very poor results on dataset PXD007592 (Figure 3.3.4; row 3, column 10). Optimal threshold is also difficult to predict, *t*-test seems to need a less conservative threshold, see for example Figure 3.3.4; rows 1 and 3.

However, for dataset PXD004682 there is better performance at a 'medium' stringency level, such as Figure 3.3.4; row 2. QPROT's optimal results often arise from very stringent significance thresholds, but there are occasions, such as log<sub>2</sub> transformation of PXD007592 data where a more lenient significance threshold may have provided a better output (Figure 3.3.4; rows 3, column 1).

## Optimising the statistical pipeline for quantitative proteomics

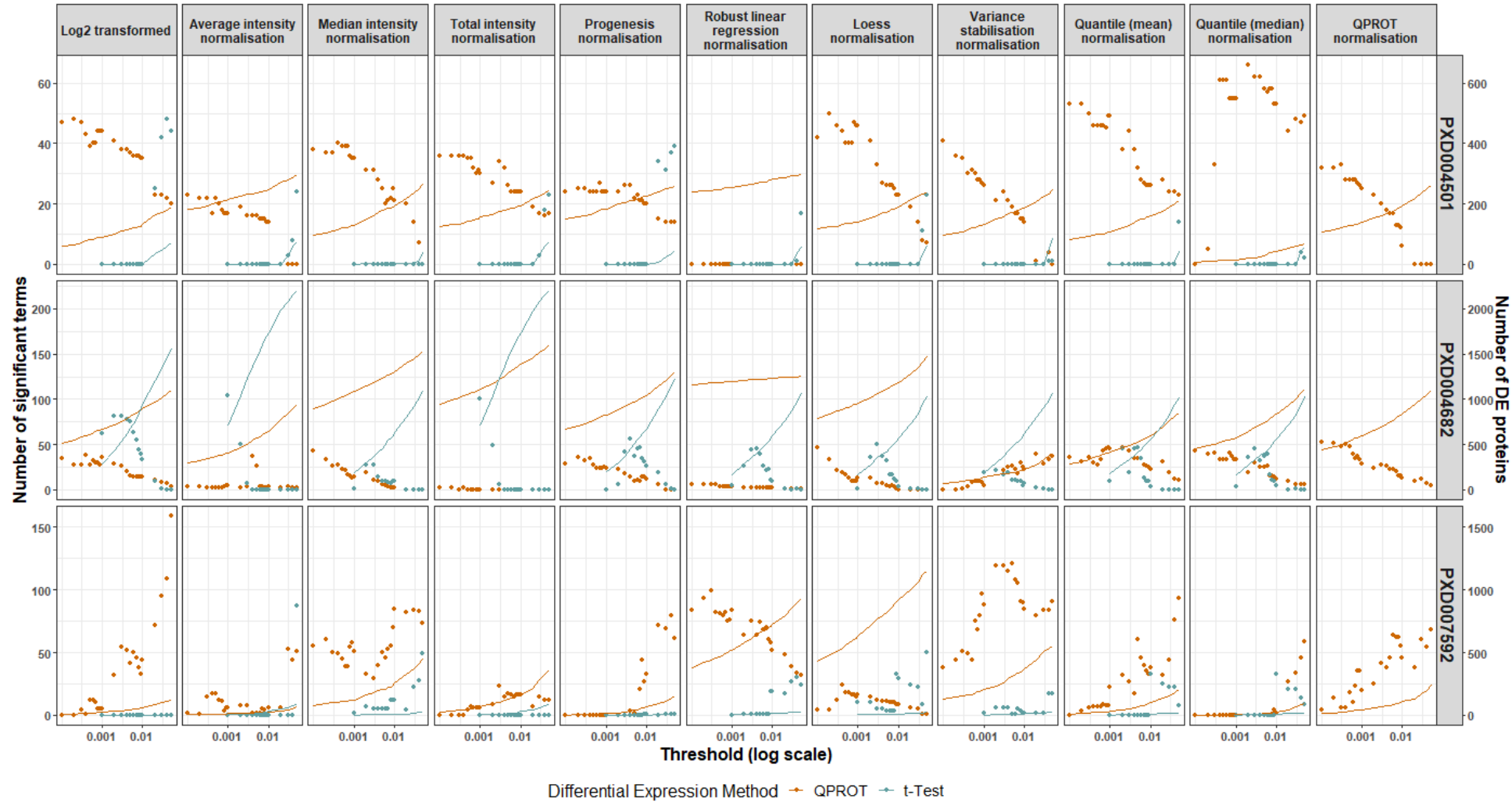


Figure 3.3.4; Enrichment analysis to assess normalisation methods. DE analysis by QPROT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide and normalised by the method shown in the x-axis strip) of three biological datasets PXD004501, PXD004682, and PXD007592 (y-axis strip). DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.

**PXD004501**

*Table 21; Enrichment analysis of dataset PXD004501. Optimal combination of analysis parameters are shown for each of the DE methods QPRT and t-test.*

<b>DE method</b>	<b>Normalisation</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Number of significant terms</b>
QPROT	Quantile - median	0.002	21	66
t-Test	Log2 transformed	0.04	60	48

Optimal analysis of dataset PXD004501, the smallest of the three in terms of DE proteins, is summarised in Table 21. The optimal performance was obtained using QPROT at a significance threshold of 0.002 with quantile - median normalisation (66 significant enrichment terms with 21 DE proteins, details of GO terms shown in Table 22), (Figure 3.3.4; row 1, column 10). Overall, QPROT appears to perform better at stringent significance thresholds and investigating further with a lower threshold level may have improved the performance, (Figure 3.3.4; row 1, columns 1-5, 8, 9, and 11). Analysis of the log2 transformed abundances produced the highest number of terms when *t*-test was used for DE analysis (48 significant terms and 60 DE proteins at an optimal significance threshold of  $p < 0.04$ , details of GO terms shown in Table 23), (Figure 3.3.4; row 1, column 1). Even with the introduction of an enrichment threshold for the normalisation evaluation analysis, which limited the number of terms being defined as significant for all methods (Figure 3.3.4), the optimal QPROT output of 66 terms (Figure 3.3.4) was an still an improvement of the optimal output of MSstats (60 terms) using Progenesis QIP normalisation in the previous section (Figure 3.3.2). However, a limitation of this evaluation is that in QPROT analysis using several of the normalisation methods, it appears that the maxima is not captured. Future work should extend the range of significance threshold to address this. Figure 3.3.5 summarises the distribution of the enrichment terms of the best parameter combinations for QPROT and *t*-test. There was consistency in the composition of the enrichment results, with the same number of Reactome and Kegg pathway results, with an increase in GO terms in using QPROT analysis.



## Optimising the statistical pipeline for quantitative proteomics

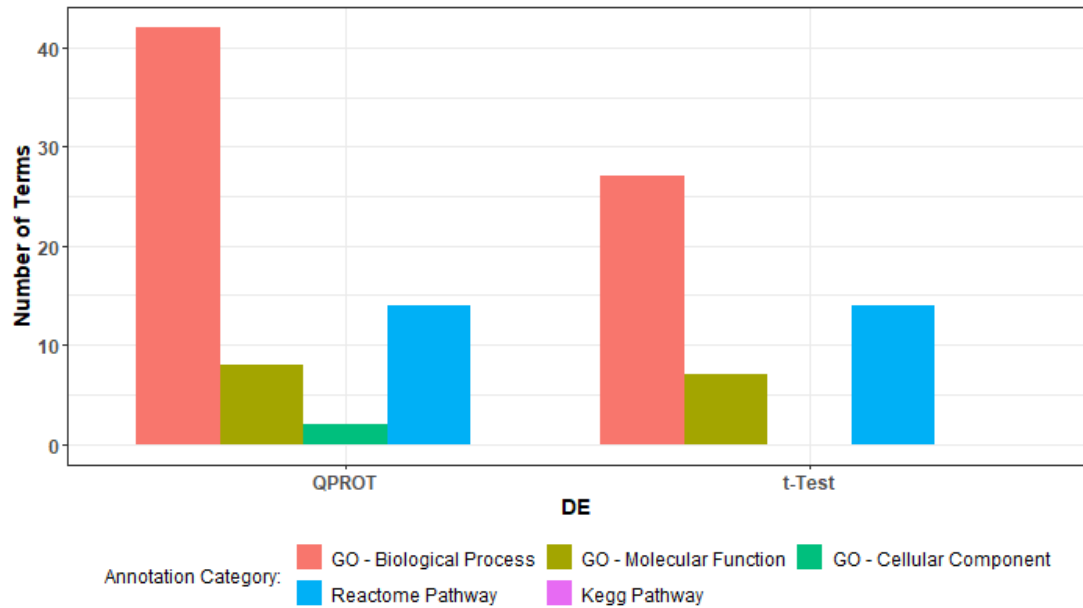


Figure 3.3.5; Distribution of the optimal results of enrichment analysis for DE methods QPROT and *t*-test for dataset PXD004501.

Overall in this dataset, *t*-test performs poorly; there are a small number of proteins being called DE at the thresholds below  $p < 0.05$ , and the number of terms increased as the threshold stringency decreased (Figure 3.3.4 row 1). Using raw abundance and AIN data at thresholds around 0.05, *t*-test analysis produces a small number of significant terms, suggesting that less conservative threshold may have been more appropriate for *t*-test analysis of this data. QPROT analysis provided more consistent results; along with Q-med normalisation, log<sub>2</sub> transformed abundances, TIN and Progenesis normalised data provided the best combination with QPROT, with AIN, MIN, RLR, VSN, Q-mean, and QPROT's own normalisation providing a comparative amount of terms. In contrast, Loess normalisation with QPROT analysis provided the least amount of significant terms, interestingly with the largest amount of DE proteins. QPROT analysis was often optimal at lower FDR values, suggesting even more stringent significance threshold could be more appropriate for this dataset. The count of terms is reduced compared to the analysis Progenesis QIP normalised data from the previous section, where there was a maximum of 178 significant enrichment terms due to the introduction of the enrichment threshold cut-off.

## Optimising the statistical pipeline for quantitative proteomics

Table 22; Enrichment results for optimal QPROT analysis with quantile median normalisation and a significance threshold of 0.002 for dataset PXD004501. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section.

Category	Term	Fold Enrichment	Benjamini p-value
GOTERM_BP_ALL	GO:0002757~immune response-activating signal transduction	5.78	0.00001
GOTERM_BP_ALL	GO:0002764~immune response-regulating signaling pathway	5.59	0.00001
REACTOME_PATHWAY	R-HSA-2168880~Scavenging of heme from plasma	6.14	0.00001
GOTERM_BP_ALL	GO:0002429~immune response-activating cell surface receptor signaling pathway	6.14	0.00002
GOTERM_BP_ALL	GO:0002768~immune response-regulating cell surface receptor signaling pathway	5.91	0.00002
REACTOME_PATHWAY	R-HSA-2871837~FCERI mediated NF-kB activation	6.33	0.00003
REACTOME_PATHWAY	R-HSA-2454202~Fc epsilon receptor (FCERI) signaling	6.59	0.00005
REACTOME_PATHWAY	R-HSA-2730905~Role of LAT2/NTAL/LAB on calcium mobilization	6.59	0.00005
REACTOME_PATHWAY	R-HSA-2871796~FCERI mediated MAPK activation	6.59	0.00005
REACTOME_PATHWAY	R-HSA-2871809~FCERI mediated Ca+2 mobilization	6.59	0.00005
REACTOME_PATHWAY	R-HSA-5690714~CD22 mediated BCR regulation	6.39	0.00005
REACTOME_PATHWAY	R-HSA-983695~Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	6.39	0.00005
GOTERM_BP_ALL	GO:0038095~Fc-epsilon receptor signaling pathway	7.53	0.00006
GOTERM_MF_ALL	GO:0003823~antigen binding	5.88	0.00007
REACTOME_PATHWAY	R-HSA-2029481~FCGR activation	5.86	0.00008
REACTOME_PATHWAY	R-HSA-2029485~Role of phospholipids in phagocytosis	5.86	0.00008
REACTOME_PATHWAY	R-HSA-2029482~Regulation of actin dynamics for phagocytic cup formation	5.70	0.00009
REACTOME_PATHWAY	R-HSA-173623~Classical antibody-mediated complement activation	5.27	0.00016
REACTOME_PATHWAY	R-HSA-198933~Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	4.91	0.00027
GOTERM_BP_ALL	GO:0038093~Fc receptor signaling pathway	6.16	0.00031
GOTERM_BP_ALL	GO:0002433~immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	6.78	0.00043
GOTERM_BP_ALL	GO:0038094~Fc-gamma receptor signaling pathway	6.78	0.00043
GOTERM_BP_ALL	GO:0002431~Fc receptor mediated stimulatory signaling pathway	6.78	0.00043
GOTERM_BP_ALL	GO:0038096~Fc-gamma receptor signaling pathway involved in phagocytosis	6.78	0.00043
GOTERM_BP_ALL	GO:0002253~activation of immune response	3.56	0.00054

## Optimising the statistical pipeline for quantitative proteomics

REACTOME_PATHWAY	R-HSA-166663~Initial triggering of complement	4.39	0.00059
GOTERM_BP_ALL	GO:0006958~complement activation, classical pathway	4.52	0.00062
GOTERM_BP_ALL	GO:0002455~humoral immune response mediated by circulating immunoglobulin	4.45	0.00062
GOTERM_BP_ALL	GO:0006909~phagocytosis	4.45	0.00062
GOTERM_BP_ALL	GO:0019724~B cell mediated immunity	4.32	0.00072
GOTERM_BP_ALL	GO:0016064~immunoglobulin mediated immune response	4.32	0.00072
GOTERM_BP_ALL	GO:0050778~positive regulation of immune response	3.26	0.00091
GOTERM_BP_ALL	GO:0002449~lymphocyte mediated immunity	4.14	0.00091
GOTERM_BP_ALL	GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	4.14	0.00091
GOTERM_BP_ALL	GO:0006898~receptor-mediated endocytosis	4.03	0.00112
GOTERM_BP_ALL	GO:0002443~leukocyte mediated immunity	3.87	0.00155
GOTERM_BP_ALL	GO:0006956~complement activation	3.77	0.00180
GOTERM_BP_ALL	GO:0002250~adaptive immune response	3.77	0.00180
GOTERM_BP_ALL	GO:0050776~regulation of immune response	2.91	0.00234
GOTERM_BP_ALL	GO:0002684~positive regulation of immune system process	2.77	0.00374
GOTERM_BP_ALL	GO:0006897~endocytosis	3.07	0.00374
GOTERM_BP_ALL	GO:0006959~humoral immune response	3.24	0.00620
GOTERM_BP_ALL	GO:0050851~antigen receptor-mediated signaling pathway	8.13	0.00731
GOTERM_BP_ALL	GO:0072376~protein activation cascade	3.14	0.00769
GOTERM_BP_ALL	GO:0007166~cell surface receptor signaling pathway	2.27	0.01068
GOTERM_BP_ALL	GO:0006955~immune response	2.26	0.01105
GOTERM_BP_ALL	GO:0050853~B cell receptor signaling pathway	10.43	0.01219
GOTERM_MF_ALL	GO:0004175~endopeptidase activity	3.17	0.01431
GOTERM_MF_ALL	GO:0004252~serine-type endopeptidase activity	3.28	0.01472
GOTERM_MF_ALL	GO:0017171~serine hydrolase activity	3.12	0.01472
GOTERM_MF_ALL	GO:0008236~serine-type peptidase activity	3.12	0.01472
GOTERM_MF_ALL	GO:0070011~peptidase activity, acting on L-amino acid peptides	2.77	0.01472
GOTERM_MF_ALL	GO:0034987~immunoglobulin receptor binding	8.74	0.01472

## *Optimising the statistical pipeline for quantitative proteomics*

GOTERM_MF_ALL	GO:0008233~peptidase activity	2.69	0.01472
GOTERM_BP_ALL	GO:0006910~phagocytosis, recognition	9.68	0.01484
GOTERM_BP_ALL	GO:0006911~phagocytosis, engulfment	9.68	0.01484
GOTERM_BP_ALL	GO:0010324~membrane invagination	9.68	0.01484
GOTERM_BP_ALL	GO:0002682~regulation of immune system process	2.30	0.01781
GOTERM_BP_ALL	GO:0002252~immune effector process	2.76	0.01781
GOTERM_BP_ALL	GO:0050871~positive regulation of B cell activation	9.04	0.01781
GOTERM_BP_ALL	GO:0050864~regulation of B cell activation	9.04	0.01781
GOTERM_BP_ALL	GO:0048584~positive regulation of response to stimulus	2.10	0.01924
GOTERM_BP_ALL	GO:0016192~vesicle-mediated transport	2.19	0.02555
GOTERM_CC_ALL	GO:0042571~immunoglobulin complex, circulating	9.91	0.04306
GOTERM_CC_ALL	GO:0019814~immunoglobulin complex	9.91	0.04306
GOTERM_BP_ALL	GO:0042113~B cell activation	6.78	0.04916

## Optimising the statistical pipeline for quantitative proteomics

Table 23; Enrichment results for optimal t-test analysis with log2 transformed abundances and a significance threshold of 0.04 for dataset PXD004501. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section.

Category	Term	Fold Enrichment	Benjamini p-value
REACTOME_PATHWAY	R-HSA-2168880~Scavenging of heme from plasma	4.08	0.00000004
REACTOME_PATHWAY	R-HSA-2454202~Fc epsilon receptor (FCER1) signaling	4.59	0.00000004
REACTOME_PATHWAY	R-HSA-2730905~Role of LAT2/NTAL/LAB on calcium mobilization	4.59	0.00000004
REACTOME_PATHWAY	R-HSA-2871796~FCER1 mediated MAPK activation	4.59	0.00000004
REACTOME_PATHWAY	R-HSA-2871809~FCER1 mediated Ca+2 mobilization	4.59	0.00000004
REACTOME_PATHWAY	R-HSA-2871837~FCER1 mediated NF-kB activation	4.19	0.00000004
REACTOME_PATHWAY	R-HSA-5690714~CD22 mediated BCR regulation	4.45	0.00000004
REACTOME_PATHWAY	R-HSA-983695~Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	4.45	0.00000004
REACTOME_PATHWAY	R-HSA-173623~Classical antibody-mediated complement activation	3.87	0.00000019
REACTOME_PATHWAY	R-HSA-2029481~FCGR activation	4.08	0.00000019
REACTOME_PATHWAY	R-HSA-2029485~Role of phospholipids in phagocytosis	4.08	0.00000019
REACTOME_PATHWAY	R-HSA-2029482~Regulation of actin dynamics for phagocytic cup formation	3.97	0.00000029
REACTOME_PATHWAY	R-HSA-166663~Initial triggering of complement	3.40	0.00000064
GOTERM_BP_ALL	GO:0038095~Fc-epsilon receptor signaling pathway	5.05	0.00000137
GOTERM_MF_ALL	GO:0003823~antigen binding	3.99	0.00000155
GOTERM_BP_ALL	GO:0002433~immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	4.77	0.00000309
GOTERM_BP_ALL	GO:0038094~Fc-gamma receptor signaling pathway	4.77	0.00000309
GOTERM_BP_ALL	GO:0038096~Fc-gamma receptor signaling pathway involved in phagocytosis	4.77	0.00000309
GOTERM_BP_ALL	GO:0002431~Fc receptor mediated stimulatory signaling pathway	4.77	0.00000309
REACTOME_PATHWAY	R-HSA-198933~Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	3.41	0.00000410
GOTERM_BP_ALL	GO:0002429~immune response-activating cell surface receptor signaling pathway	3.81	0.00000549
GOTERM_BP_ALL	GO:0038093~Fc receptor signaling pathway	4.13	0.00000852
GOTERM_BP_ALL	GO:0002768~immune response-regulating cell surface receptor signaling pathway	3.67	0.00000852
GOTERM_BP_ALL	GO:0006958~complement activation, classical pathway	3.21	0.00003851
GOTERM_BP_ALL	GO:0002455~humoral immune response mediated by circulating immunoglobulin	3.17	0.00004422

## *Optimising the statistical pipeline for quantitative proteomics*

GOTERM_BP_ALL	GO:0002757~immune response-activating signal transduction	3.31	0.00004422
GOTERM_BP_ALL	GO:0006898~receptor-mediated endocytosis	3.00	0.00004551
GOTERM_BP_ALL	GO:0019724~B cell mediated immunity	3.07	0.00005793
GOTERM_BP_ALL	GO:0016064~immunoglobulin mediated immune response	3.07	0.00005793
GOTERM_BP_ALL	GO:0002764~immune response-regulating signaling pathway	3.21	0.00005864
GOTERM_BP_ALL	GO:0002449~lymphocyte mediated immunity	2.95	0.00010260
GOTERM_BP_ALL	GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2.95	0.00010260
GOTERM_BP_ALL	GO:0002250~adaptive immune response	2.81	0.00010260
GOTERM_BP_ALL	GO:0006956~complement activation	2.81	0.00010260
GOTERM_BP_ALL	GO:0006909~phagocytosis	3.01	0.00013275
GOTERM_BP_ALL	GO:0002443~leukocyte mediated immunity	2.75	0.00029090
GOTERM_BP_ALL	GO:0072376~protein activation cascade	2.44	0.00061751
GOTERM_BP_ALL	GO:0006897~endocytosis	2.29	0.00112161
GOTERM_BP_ALL	GO:0002253~activation of immune response	2.35	0.00121554
GOTERM_BP_ALL	GO:0006959~humoral immune response	2.41	0.00128297
GOTERM_MF_ALL	GO:0004175~endopeptidase activity	2.36	0.00293786
GOTERM_MF_ALL	GO:0004252~serine-type endopeptidase activity	2.44	0.00371493
GOTERM_MF_ALL	GO:0008236~serine-type peptidase activity	2.32	0.00477173
GOTERM_MF_ALL	GO:0017171~serine hydrolase activity	2.32	0.00477173
GOTERM_BP_ALL	GO:0050778~positive regulation of immune response	2.15	0.00528982
GOTERM_MF_ALL	GO:0070011~peptidase activity, acting on L-amino acid peptides	2.07	0.00915706
GOTERM_MF_ALL	GO:0008233~peptidase activity	2.01	0.01230922
GOTERM_BP_ALL	GO:0002252~immune effector process	2.06	0.01680846

**PXD004682**

*Table 24; Enrichment analysis of dataset PXD004682. Optimal combination of analysis parameters are shown for each of the DE methods QPRT and t-test.*

<b>DE method</b>	<b>Normalisation</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Total number of significant terms</b>
QPROT	QPROT	0.0001	443	53
t-Test	AIN	0.001	706	104

Summary of analysis of dataset PXD004682 is shown in Table 24. In this dataset there were the lowest number of identified protein groups with the largest proportion of proteins identified as differentially expressed. *t*-test DE analysis on AIN abundances and a cut-of threshold of 0.001 gave the optimal enrichment analysis output with 104 significant terms from 706 DE proteins at a significance threshold of  $p < 0.001$  (GO terms shown in Table 26). The best combination of parameters for QPROT were using a cut-of threshold of 0.0001 FDR and the inbuilt QPROT normalisation, giving 443 DE proteins and 53 significant terms (GO terms shown in Table 25). Figure 3.3.6 shows the annotation terms of the two DE methods. There is a discrepancy to their distribution, although *t*-test has the most overall terms they are mostly concentrated as Reactome, with QPROT having a larger number of CC, MF and Kegg pathway terms.

The count of DE proteins at the minimum *t*-test threshold of 0.001 for TIN was also relatively high and produced the optimal output for that normalisation method, (Figure 3.3.4; row 2, columns 2 and 4).

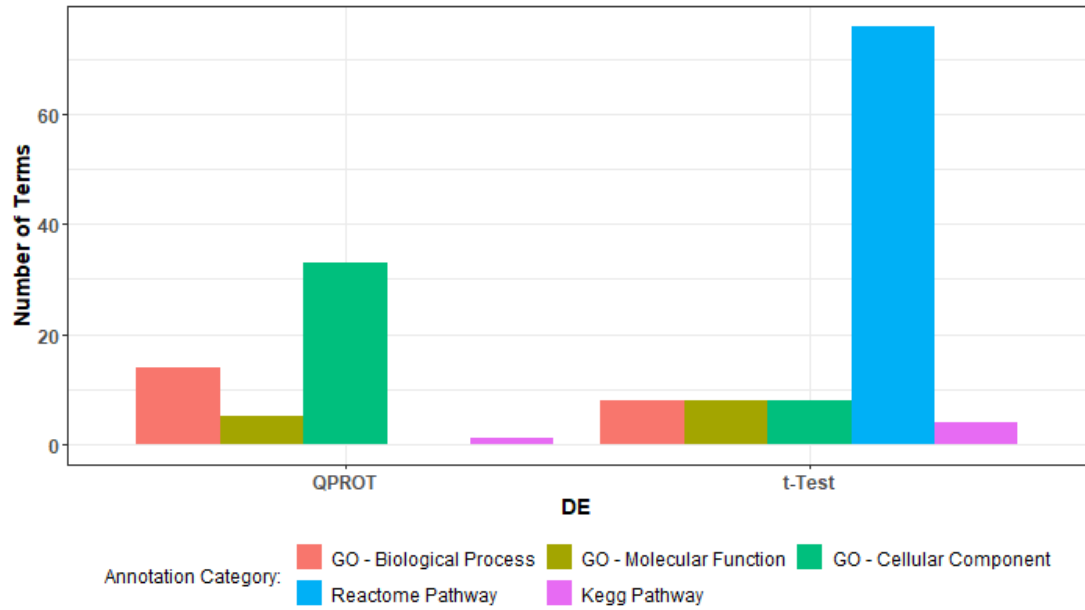


Figure 3.3.6; Distribution of the optimal results of enrichment analysis for DE methods QPROT and t-test for dataset PXD004682.

Overall the two DE analysis methods performed similarly with this dataset. Both identified a large number of proteins as being DE, and often produced a greater number of significant enrichment terms when the number of DEs were lower. Using VSN data produced some of the largest number of DE proteins (750-2000) and appeared to give the worst overall performance by both DE methods. Once again *t-test's* optimal output was from raw abundance data, and QPROT performed best at low significance thresholds (*t-test's* optimal significance threshold was consistently around 0.001). QPROT often called more proteins as DE than *t-test*, and apart from Loess and VSN, the normalisation method did not make a great difference to the quality of the results.



## Optimising the statistical pipeline for quantitative proteomics

Table 25; Enrichment results for optimal QPROT analysis with QPROT normalisation and a significance threshold of 0.0001 for dataset PXD004682. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section.

Category	Term	Fold Enrichment	Benjamini p-value
GOTERM_BP_ALL	GO:0002757~immune response-activating signal transduction	5.78	0.00001
GOTERM_BP_ALL	GO:0002764~immune response-regulating signaling pathway	5.59	0.00001
REACTOME_PATHWAY	R-HSA-2168880~Scavenging of heme from plasma	6.14	0.00001
GOTERM_BP_ALL	GO:0002429~immune response-activating cell surface receptor signaling pathway	6.14	0.00002
GOTERM_BP_ALL	GO:0002768~immune response-regulating cell surface receptor signaling pathway	5.91	0.00002
REACTOME_PATHWAY	R-HSA-2871837~FCERI mediated NF-kB activation	6.33	0.00003
REACTOME_PATHWAY	R-HSA-2454202~Fc epsilon receptor (FCERI) signaling	6.59	0.00005
REACTOME_PATHWAY	R-HSA-2730905~Role of LAT2/NTAL/LAB on calcium mobilization	6.59	0.00005
REACTOME_PATHWAY	R-HSA-2871796~FCERI mediated MAPK activation	6.59	0.00005
REACTOME_PATHWAY	R-HSA-2871809~FCERI mediated Ca+2 mobilization	6.59	0.00005
REACTOME_PATHWAY	R-HSA-5690714~CD22 mediated BCR regulation	6.39	0.00005
REACTOME_PATHWAY	R-HSA-983695~Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	6.39	0.00005
GOTERM_BP_ALL	GO:0038095~Fc-epsilon receptor signaling pathway	7.53	0.00006
GOTERM_MF_ALL	GO:0003823~antigen binding	5.88	0.00007
REACTOME_PATHWAY	R-HSA-2029481~FCGR activation	5.86	0.00008
REACTOME_PATHWAY	R-HSA-2029485~Role of phospholipids in phagocytosis	5.86	0.00008
REACTOME_PATHWAY	R-HSA-2029482~Regulation of actin dynamics for phagocytic cup formation	5.70	0.00009
REACTOME_PATHWAY	R-HSA-173623~Classical antibody-mediated complement activation	5.27	0.00016
REACTOME_PATHWAY	R-HSA-198933~Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	4.91	0.00027
GOTERM_BP_ALL	GO:0038093~Fc receptor signaling pathway	6.16	0.00031
GOTERM_BP_ALL	GO:0002433~immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	6.78	0.00043
GOTERM_BP_ALL	GO:0038094~Fc-gamma receptor signaling pathway	6.78	0.00043
GOTERM_BP_ALL	GO:0002431~Fc receptor mediated stimulatory signaling pathway	6.78	0.00043
GOTERM_BP_ALL	GO:0038096~Fc-gamma receptor signaling pathway involved in phagocytosis	6.78	0.00043
GOTERM_BP_ALL	GO:0002253~activation of immune response	3.56	0.00054

## Optimising the statistical pipeline for quantitative proteomics

REACTOME_PATHWAY	R-HSA-166663~Initial triggering of complement	4.39	0.00059
GOTERM_BP_ALL	GO:0006958~complement activation, classical pathway	4.52	0.00062
GOTERM_BP_ALL	GO:0002455~humoral immune response mediated by circulating immunoglobulin	4.45	0.00062
GOTERM_BP_ALL	GO:0006909~phagocytosis	4.45	0.00062
GOTERM_BP_ALL	GO:0019724~B cell mediated immunity	4.32	0.00072
GOTERM_BP_ALL	GO:0016064~immunoglobulin mediated immune response	4.32	0.00072
GOTERM_BP_ALL	GO:0050778~positive regulation of immune response	3.26	0.00091
GOTERM_BP_ALL	GO:0002449~lymphocyte mediated immunity	4.14	0.00091
GOTERM_BP_ALL	GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	4.14	0.00091
GOTERM_BP_ALL	GO:0006898~receptor-mediated endocytosis	4.03	0.00112
GOTERM_BP_ALL	GO:0002443~leukocyte mediated immunity	3.87	0.00155
GOTERM_BP_ALL	GO:0006956~complement activation	3.77	0.00180
GOTERM_BP_ALL	GO:0002250~adaptive immune response	3.77	0.00180
GOTERM_BP_ALL	GO:0050776~regulation of immune response	2.91	0.00234
GOTERM_BP_ALL	GO:0002684~positive regulation of immune system process	2.77	0.00374
GOTERM_BP_ALL	GO:0006897~endocytosis	3.07	0.00374
GOTERM_BP_ALL	GO:0006959~humoral immune response	3.24	0.00620
GOTERM_BP_ALL	GO:0050851~antigen receptor-mediated signaling pathway	8.13	0.00731
GOTERM_BP_ALL	GO:0072376~protein activation cascade	3.14	0.00769
GOTERM_BP_ALL	GO:0007166~cell surface receptor signaling pathway	2.27	0.01068
GOTERM_BP_ALL	GO:0006955~immune response	2.26	0.01105
GOTERM_BP_ALL	GO:0050853~B cell receptor signaling pathway	10.43	0.01219
GOTERM_MF_ALL	GO:0004175~endopeptidase activity	3.17	0.01431
GOTERM_MF_ALL	GO:0004252~serine-type endopeptidase activity	3.28	0.01472
GOTERM_MF_ALL	GO:0017171~serine hydrolase activity	3.12	0.01472
GOTERM_MF_ALL	GO:0008236~serine-type peptidase activity	3.12	0.01472
GOTERM_MF_ALL	GO:0070011~peptidase activity, acting on L-amino acid peptides	2.77	0.01472
GOTERM_MF_ALL	GO:0034987~immunoglobulin receptor binding	8.74	0.01472

## Optimising the statistical pipeline for quantitative proteomics

GOTERM_MF_ALL	GO:0008233~peptidase activity	2.69	0.01472
GOTERM_BP_ALL	GO:0006910~phagocytosis, recognition	9.68	0.01484
GOTERM_BP_ALL	GO:0006911~phagocytosis, engulfment	9.68	0.01484
GOTERM_BP_ALL	GO:0010324~membrane invagination	9.68	0.01484
GOTERM_BP_ALL	GO:0002682~regulation of immune system process	2.30	0.01781
GOTERM_BP_ALL	GO:0002252~immune effector process	2.76	0.01781
GOTERM_BP_ALL	GO:0050871~positive regulation of B cell activation	9.04	0.01781
GOTERM_BP_ALL	GO:0050864~regulation of B cell activation	9.04	0.01781
GOTERM_BP_ALL	GO:0048584~positive regulation of response to stimulus	2.10	0.01924
GOTERM_BP_ALL	GO:0016192~vesicle-mediated transport	2.19	0.02555
GOTERM_CC_ALL	GO:0042571~immunoglobulin complex, circulating	9.91	0.04306
GOTERM_CC_ALL	GO:0019814~immunoglobulin complex	9.91	0.04306
GOTERM_BP_ALL	GO:0042113~B cell activation	6.78	0.04916

## Optimising the statistical pipeline for quantitative proteomics

Table 26; Enrichment results for optimal t-test analysis with AIN abundances and a significance threshold of 0.001 for dataset PXD004682. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section.

Category	Term	Fold Enrichment	Benjamini p-value
KEGG_PATHWAY	hsa04141:Protein processing in endoplasmic reticulum	2.26	0.0000031
REACTOME_PATHWAY	R-HSA-69239~Synthesis of DNA	2.51	0.0000086
REACTOME_PATHWAY	R-HSA-69242~S Phase	2.37	0.0000016
GOTERM_CC_ALL	GO:0022624~proteasome accessory complex	3.33	0.0000412
REACTOME_PATHWAY	R-HSA-453279~Mitotic G1 phase and G1/S transition	2.31	0.0000422
GOTERM_CC_ALL	GO:0005838~proteasome regulatory particle	3.46	0.0000844
REACTOME_PATHWAY	R-HSA-69206~G1/S Transition	2.25	0.0001258
REACTOME_PATHWAY	R-HSA-68949~Orc1 removal from chromatin	2.37	0.0001447
REACTOME_PATHWAY	R-HSA-69052~Switching of origins to a post-replicative state	2.37	0.0001447
REACTOME_PATHWAY	R-HSA-69190~DNA strand elongation	3.73	0.0010575
REACTOME_PATHWAY	R-HSA-9759194~Nuclear events mediated by NFE2L2	2.13	0.0013778
REACTOME_PATHWAY	R-HSA-450408~AUF1 (hnRNP D0) binds and destabilizes mRNA	2.22	0.0016226
REACTOME_PATHWAY	R-HSA-5668541~TNFR2 non-canonical NF-kB pathway	2.22	0.0016226
REACTOME_PATHWAY	R-HSA-5607761~Dectin-1 mediated noncanonical NF-kB signaling	2.18	0.0026605
REACTOME_PATHWAY	R-HSA-5676590~NIK-->noncanonical NF-kB signaling	2.18	0.0026605
REACTOME_PATHWAY	R-HSA-351202~Metabolism of polyamines	2.20	0.0033016
GOTERM_BP_ALL	GO:0006310~DNA recombination	2.02	0.0040453
GOTERM_BP_ALL	GO:0006260~DNA replication	2.02	0.0040453
REACTOME_PATHWAY	R-HSA-5610780~Degradation of GLI1 by the proteasome	2.14	0.0047487
REACTOME_PATHWAY	R-HSA-180534~Vpu mediated degradation of CD4	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-69580~p53-Dependent G1/S DNA damage checkpoint	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-187577~SCF(Skp2)-mediated degradation of p27/p21	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-9762114~GSK3B and BTRC:CUL1-mediated-degradation of NFE2L2	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-174113~SCF-beta-TrCP mediated degradation of Emi1	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-69563~p53-Dependent G1 DNA Damage Response	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-349425~Autodegradation of the E3 ubiquitin ligase COP1	2.16	0.0047487

## *Optimising the statistical pipeline for quantitative proteomics*

REACTOME_PATHWAY	R-HSA-69541~Stabilization of p53	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-69615~G1/S DNA Damage Checkpoints	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-8854050~FBXL7 down-regulates AURKA during mitotic entry and in early mitosis	2.16	0.0047487
REACTOME_PATHWAY	R-HSA-176187~Activation of ATR in response to replication stress	3.73	0.0049128
REACTOME_PATHWAY	R-HSA-5696398~Nucleotide Excision Repair	2.44	0.0051844
GOTERM_MF_ALL	GO:0043021~ribonucleoprotein complex binding	2.07	0.005191
REACTOME_PATHWAY	R-HSA-2871837~FCERI mediated NF-kB activation	2.03	0.0057396
REACTOME_PATHWAY	R-HSA-1234176~Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-5610783~Degradation of GLI2 by the proteasome	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-8948751~Regulation of PTEN stability and activity	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-69656~Cyclin A:Cdk2-associated events at S phase entry	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-174143~APC/C-mediated degradation of cell cycle proteins	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-9604323~Negative regulation of NOTCH4 signaling	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-5358346~Hedgehog ligand biogenesis	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-176408~Regulation of APC/C activators between G1/S and early anaphase	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-69202~Cyclin E associated events during G1/S transition	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-453276~Regulation of mitotic cell cycle	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-5610785~GLI3 is processed to GLI3R by the proteasome	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-1234174~Cellular response to hypoxia	2.10	0.0057396
REACTOME_PATHWAY	R-HSA-9735869~SARS-CoV-1 modulates host translation machinery	2.31	0.0057396
REACTOME_PATHWAY	R-HSA-9013694~Signaling by NOTCH4	2.04	0.0058165
REACTOME_PATHWAY	R-HSA-69610~p53-Independent DNA Damage Response	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-4641257~Degradation of AXIN	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-174178~APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-174154~APC/C:Cdc20 mediated degradation of Securin	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-350562~Regulation of ornithine decarboxylase (ODC)	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-69613~p53-Independent G1/S DNA damage checkpoint	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-75815~Ubiquitin-dependent degradation of Cyclin D	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-174084~Autodegradation of Cdh1 by Cdh1:APC/C	2.12	0.0058165

## Optimising the statistical pipeline for quantitative proteomics

REACTOME_PATHWAY	R-HSA-69017~CDK-mediated phosphorylation and removal of Cdc6	2.12	0.0058165
REACTOME_PATHWAY	R-HSA-69601~Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	2.12	0.0058165
GOTERM_CC_ALL	GO:0044454~nuclear chromosome part	2.31	0.0063258
REACTOME_PATHWAY	R-HSA-8939902~Regulation of RUNX2 expression and activity	2.05	0.006794
REACTOME_PATHWAY	R-HSA-5632684~Hedgehog 'on' state	2.05	0.006794
REACTOME_PATHWAY	R-HSA-176814~Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-4608870~Asymmetric localization of PCP proteins	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-174184~Cdc20:Phospho-APC/C mediated degradation of Cyclin A	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-5387390~Hh mutants abrogate ligand secretion	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-211733~Regulation of activated PAK-2p34 by proteasome mediated degradation	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-176409~APC/C:Cdc20 mediated degradation of mitotic proteins	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-5362768~Hh mutants are degraded by ERAD	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-4641258~Degradation of DVL	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-179419~APC:Cdc20 mediated degradation of cell cycle proteins prior to satisfaction of the cell cycle checkpoint	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-8941858~Regulation of RUNX3 expression and activity	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-180585~Vif-mediated degradation of APOBEC3G	2.06	0.0072972
REACTOME_PATHWAY	R-HSA-68962~Activation of the pre-replicative complex	3.73	0.007485
GOTERM_BP_ALL	GO:0006261~DNA-dependent DNA replication	2.17	0.0078504
REACTOME_PATHWAY	R-HSA-5658442~Regulation of RAS by GAPs	2.00	0.0079233
GOTERM_CC_ALL	GO:0015935~small ribosomal subunit	2.33	0.0081112
REACTOME_PATHWAY	R-HSA-5696399~Global Genome Nucleotide Excision Repair (GG-NER)	2.49	0.0084861
REACTOME_PATHWAY	R-HSA-381119~Unfolded Protein Response (UPR)	2.17	0.0091469
KEGG_PATHWAY	hsa03030:DNA replication	3.44	0.0093144
REACTOME_PATHWAY	R-HSA-169911~Regulation of Apoptosis	2.01	0.0096264
REACTOME_PATHWAY	R-HSA-1236978~Cross-presentation of soluble exogenous antigens (endosomes)	2.01	0.0096264
GOTERM_CC_ALL	GO:0000502~proteasome complex	2.05	0.0105855
KEGG_PATHWAY	hsa03013:Nucleocytoplasmic transport	2.03	0.0118799
REACTOME_PATHWAY	R-HSA-6781827~Transcription-Coupled Nucleotide Excision Repair (TC-NER)	2.63	0.0119519
GOTERM_MF_ALL	GO:0004386~helicase activity	2.14	0.0134942

## Optimising the statistical pipeline for quantitative proteomics

REACTOME_PATHWAY	R-HSA-9754678~SARS-CoV-2 modulates host translation machinery	2.21	0.0142003
GOTERM_CC_ALL	GO:0022627~cytosolic small ribosomal subunit	2.28	0.0156316
GOTERM_BP_ALL	GO:0000725~recombinational repair	2.30	0.018628
REACTOME_PATHWAY	R-HSA-180786~Extension of Telomeres	3.31	0.0196071
REACTOME_PATHWAY	R-HSA-6803529~FGFR2 alternative splicing	3.31	0.0196071
GOTERM_BP_ALL	GO:0006275~regulation of DNA replication	2.46	0.0244952
GOTERM_BP_ALL	GO:0035966~response to topologically incorrect protein	2.04	0.0249392
GOTERM_BP_ALL	GO:0006986~response to unfolded protein	2.10	0.0365318
GOTERM_CC_ALL	GO:0043596~nuclear replication fork	3.92	0.0382861
GOTERM_CC_ALL	GO:0008540~proteasome regulatory particle, base subcomplex	3.49	0.0382861
REACTOME_PATHWAY	R-HSA-6782210~Gap-filling DNA repair synthesis and ligation in TC-NER	2.80	0.0383433
GOTERM_MF_ALL	GO:0031369~translation initiation factor binding	3.09	0.0426369
GOTERM_MF_ALL	GO:0016706~oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	3.93	0.0426369
GOTERM_MF_ALL	GO:0017116~single-stranded DNA-dependent ATP-dependent DNA helicase activity	3.93	0.0426369
GOTERM_MF_ALL	GO:0043142~single-stranded DNA-dependent ATPase activity	3.93	0.0426369
GOTERM_MF_ALL	GO:0044183~protein binding involved in protein folding	2.42	0.0426369
REACTOME_PATHWAY	R-HSA-70263~Gluconeogenesis	2.20	0.0441255
GOTERM_BP_ALL	GO:0090329~regulation of DNA-dependent DNA replication	3.26	0.0462452
KEGG_PATHWAY	hsa03050:Proteasome	2.09	0.0483399
GOTERM_MF_ALL	GO:0003678~DNA helicase activity	2.46	0.0498916

**PXD007592**

*Table 27; Enrichment analysis of dataset PXD007592. Optimal combination of analysis parameters are shown for each of the DE methods QPRT and t-test.*

<b>DE method</b>	<b>Normalisation</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Number of significant terms</b>
QPROT	Log2 transformed	0.05	159	124
t-Test	AIN	0.05	89	87

Table 27 summarises the best combination of parameters using DE methods QPROT and *t*-test for analysis of PXD007592, the dataset with the largest number of identified protein groups and the smallest proportion of proteins identified as differentially expressed. QPROT analysis combined with a significance threshold of 0.05 FDR and using log2 transformed protein abundances gave the optimal enrichment output of 125 terms with 159 DE proteins (GO terms shown in Table 28). The *t*-test analysis also at a significance threshold of 0.05 produced 89 significant terms using AIN (GO terms shown in Table 29). Again these results were an improvement on the maximum output from QPROT of 88 terms using Progenesis QIP normalisation in the previous section, despite the introduced fold-enrichment significance threshold. However, a limitation of this evaluation is that in QPROT analysis using several of the normalisation methods, it appears that the maxima is not captured. Future work should extend the range of significance threshold to address this. The plot in Figure 3.3.7 shows the annotation categories of the significant terms. The distributions were similar, with QPROT having an increased number in all categories.



## Optimising the statistical pipeline for quantitative proteomics

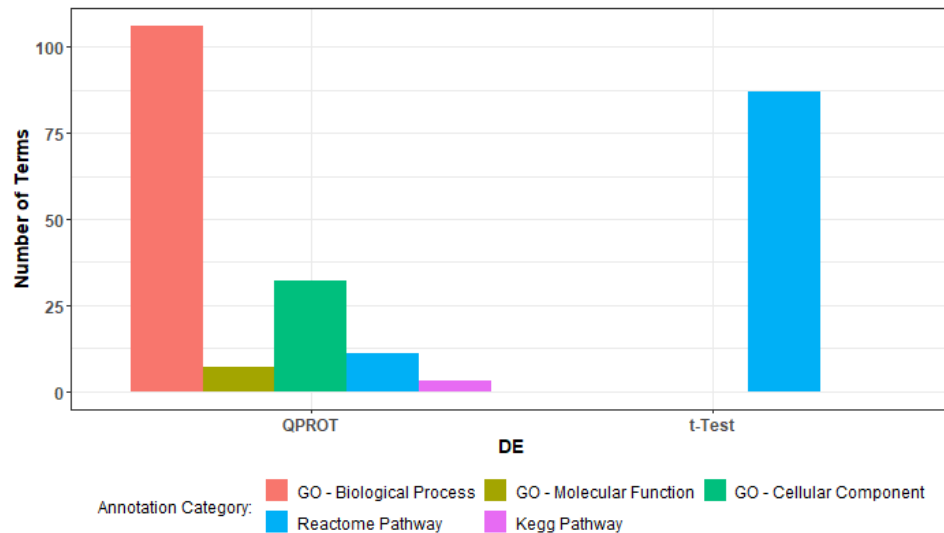


Figure 3.3.7; Distribution of the optimal results of enrichment analysis for DE methods QPROT and t-test for dataset PXD007592.

Overall, the *t*-test performed poorly with this data. Generally, there are very few proteins were called DE, even at less conservative significance thresholds, and even in the normalisation data that provided DE proteins (AIN and TIN, Figure 3.3.4, row 3, columns 2 and 4), there was not a large amount of significant enrichment terms. For QPROT analysis, log<sub>2</sub> transformed data (Figure 3.3.4, row 3, columns 1), performed best, but only at a relatively high FDR. Other successful normalisation methods were using Q-mean, loess, and RLR (Figure 3.3.4, row 3, columns 6, 7, and 9).

## Optimising the statistical pipeline for quantitative proteomics

Table 28; Enrichment results for optimal QPROT analysis with log2 transformed protein abundances and a significance threshold of 0.05 for dataset PXD007592. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section.

Category	Term	Fold Enrichment	Benjamini p-value
GOTERM_CC_ALL	GO:0072562~blood microparticle	9.39	1.56E-18
GOTERM_CC_ALL	GO:0005615~extracellular space	4.47	2.93E-18
KEGG_PATHWAY	hsa04610:Complement and coagulation cascades	12.53	3.23E-10
GOTERM_BP_ALL	GO:0072376~protein activation cascade	9.34	1.17E-08
GOTERM_BP_ALL	GO:0030198~extracellular matrix organization	5.04	6.58E-08
GOTERM_BP_ALL	GO:0043062~extracellular structure organization	5.04	6.58E-08
GOTERM_BP_ALL	GO:0009611~response to wounding	3.58	0.000001
GOTERM_BP_ALL	GO:0042060~wound healing	3.86	0.000002
GOTERM_BP_ALL	GO:1900046~regulation of hemostasis	11.56	0.000003
GOTERM_BP_ALL	GO:0030193~regulation of blood coagulation	11.56	0.000003
GOTERM_BP_ALL	GO:0050818~regulation of coagulation	11.08	0.000004
REACTOME_PATHWAY	R-HSA-114608~Platelet degranulation	5.95	0.000009
GOTERM_BP_ALL	GO:1900047~negative regulation of hemostasis	14.50	0.000010
GOTERM_BP_ALL	GO:0030195~negative regulation of blood coagulation	14.50	0.000010
GOTERM_CC_ALL	GO:0034774~secretory granule lumen	8.46	0.000013
GOTERM_CC_ALL	GO:0031983~vesicle lumen	7.20	0.000014
GOTERM_CC_ALL	GO:0060205~cytoplasmic membrane-bounded vesicle lumen	7.20	0.000014
GOTERM_BP_ALL	GO:0006959~humoral immune response	7.48	0.000014
GOTERM_BP_ALL	GO:0032101~regulation of response to external stimulus	3.78	0.000014
GOTERM_BP_ALL	GO:0050819~negative regulation of coagulation	13.60	0.000015
GOTERM_BP_ALL	GO:0061041~regulation of wound healing	8.06	0.000020
GOTERM_BP_ALL	GO:0006954~inflammatory response	4.37	0.000025
GOTERM_BP_ALL	GO:0030155~regulation of cell adhesion	3.61	0.000025
GOTERM_CC_ALL	GO:0005581~collagen trimer	10.54	0.000026
GOTERM_CC_ALL	GO:0098552~side of membrane	3.86	0.000029

## Optimising the statistical pipeline for quantitative proteomics

GOTERM_CC_ALL	GO:0031091~platelet alpha granule	7.12	0.000043
GOTERM_BP_ALL	GO:0002250~adaptive immune response	5.41	0.000051
GOTERM_BP_ALL	GO:0002576~platelet degranulation	6.29	0.000074
GOTERM_BP_ALL	GO:0061045~negative regulation of wound healing	10.88	0.000087
GOTERM_BP_ALL	GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	6.04	0.000105
GOTERM_BP_ALL	GO:0042730~fibrinolysis	16.92	0.000115
GOTERM_BP_ALL	GO:1903034~regulation of response to wounding	6.59	0.000117
GOTERM_BP_ALL	GO:0006958~complement activation, classical pathway	8.63	0.000117
GOTERM_BP_ALL	GO:0016485~protein processing	5.37	0.000117
GOTERM_BP_ALL	GO:0006952~defense response	2.57	0.000120
GOTERM_CC_ALL	GO:0009986~cell surface	3.17	0.000126
GOTERM_BP_ALL	GO:0050727~regulation of inflammatory response	5.71	0.000155
GOTERM_BP_ALL	GO:0002455~humoral immune response mediated by circulating immunoglobulin	8.06	0.000195
GOTERM_BP_ALL	GO:1903035~negative regulation of response to wounding	9.46	0.000207
REACTOME_PATHWAY	R-HSA-140875~Common Pathway of Fibrin Clot Formation	6.44	0.000258
REACTOME_PATHWAY	R-HSA-216083~Integrin cell surface interactions	18.45	0.000258
REACTOME_PATHWAY	R-HSA-977606~Regulation of Complement cascade	12.30	0.000332
GOTERM_BP_ALL	GO:0090303~positive regulation of wound healing	10.74	0.000387
GOTERM_BP_ALL	GO:0007155~cell adhesion	2.00	0.000401
GOTERM_BP_ALL	GO:0006956~complement activation	7.33	0.000401
GOTERM_BP_ALL	GO:0009605~response to external stimulus	2.15	0.000401
GOTERM_CC_ALL	GO:0009897~external side of plasma membrane	5.41	0.000406
GOTERM_CC_ALL	GO:0044420~extracellular matrix component	5.41	0.000406
GOTERM_CC_ALL	GO:0034358~plasma lipoprotein particle	14.76	0.000406
GOTERM_CC_ALL	GO:1990777~lipoprotein particle	14.76	0.000406
GOTERM_CC_ALL	GO:0005578~proteinaceous extracellular matrix	4.26	0.000596
GOTERM_CC_ALL	GO:0032994~protein-lipid complex	13.42	0.000650
GOTERM_CC_ALL	GO:0031093~platelet alpha granule lumen	7.87	0.000650

## Optimising the statistical pipeline for quantitative proteomics

GOTERM_BP_ALL	GO:0006955~immune response	2.39	0.000658
GOTERM_BP_ALL	GO:0007596~blood coagulation	3.67	0.000682
GOTERM_BP_ALL	GO:1903036~positive regulation of response to wounding	9.67	0.000682
GOTERM_BP_ALL	GO:2000257~regulation of protein activation cascade	9.67	0.000682
GOTERM_BP_ALL	GO:0051604~protein maturation	4.45	0.000682
GOTERM_BP_ALL	GO:0016064~immunoglobulin mediated immune response	6.71	0.000703
GOTERM_BP_ALL	GO:0019724~B cell mediated immunity	6.71	0.000703
GOTERM_BP_ALL	GO:0050817~coagulation	3.60	0.000784
GOTERM_BP_ALL	GO:0002252~immune effector process	3.00	0.000794
GOTERM_BP_ALL	GO:0007599~hemostasis	3.57	0.000839
GOTERM_CC_ALL	GO:0030141~secretory granule	3.34	0.000892
GOTERM_BP_ALL	GO:0032102~negative regulation of response to external stimulus	5.66	0.000942
GOTERM_BP_ALL	GO:0002449~lymphocyte mediated immunity	5.00	0.001077
GOTERM_CC_ALL	GO:0098644~complex of collagen trimers	17.57	0.001262
GOTERM_BP_ALL	GO:0045785~positive regulation of cell adhesion	3.86	0.001286
GOTERM_BP_ALL	GO:0050820~positive regulation of coagulation	14.50	0.001428
GOTERM_BP_ALL	GO:1900048~positive regulation of hemostasis	14.50	0.001428
GOTERM_BP_ALL	GO:0030194~positive regulation of blood coagulation	14.50	0.001428
GOTERM_CC_ALL	GO:0005788~endoplasmic reticulum lumen	4.51	0.001620
GOTERM_BP_ALL	GO:0050776~regulation of immune response	2.57	0.001650
GOTERM_BP_ALL	GO:0002443~leukocyte mediated immunity	4.30	0.001696
GOTERM_BP_ALL	GO:0002697~regulation of immune effector process	4.25	0.001883
GOTERM_CC_ALL	GO:0034364~high-density lipoprotein particle	15.37	0.002276
REACTOME_PATHWAY	R-HSA-3000178~ECM proteoglycans	5.98	0.002345
REACTOME_PATHWAY	R-HSA-2022090~Assembly of collagen fibrils and other multimeric structures	11.36	0.002529
GOTERM_BP_ALL	GO:0044236~multicellular organism metabolic process	7.44	0.003146
GOTERM_BP_ALL	GO:0072378~blood coagulation, fibrin clot formation	12.09	0.003648
GOTERM_BP_ALL	GO:0016337~single organismal cell-cell adhesion	2.85	0.003648
GOTERM_BP_ALL	GO:0030449~regulation of complement activation	8.91	0.003897

## Optimising the statistical pipeline for quantitative proteomics

GOTERM_BP_ALL	GO:0050878~regulation of body fluid levels	3.07	0.004206
GOTERM_BP_ALL	GO:0002526~acute inflammatory response	5.88	0.004619
GOTERM_BP_ALL	GO:0031589~cell-substrate adhesion	3.36	0.004785
GOTERM_BP_ALL	GO:0002684~positive regulation of immune system process	2.44	0.004785
GOTERM_BP_ALL	GO:0002673~regulation of acute inflammatory response	8.46	0.004886
GOTERM_BP_ALL	GO:0002920~regulation of humoral immune response	8.46	0.004886
GOTERM_BP_ALL	GO:0044243~multicellular organism catabolic process	11.16	0.004999
KEGG_PATHWAY	hsa05322:Systemic lupus erythematosus	6.65	0.005612
GOTERM_BP_ALL	GO:0051346~negative regulation of hydrolase activity	3.49	0.005664
GOTERM_BP_ALL	GO:0098602~single organism cell adhesion	2.63	0.005747
GOTERM_BP_ALL	GO:0002682~regulation of immune system process	2.16	0.006497
GOTERM_MF_ALL	GO:0005198~structural molecule activity	2.49	0.007545
GOTERM_MF_ALL	GO:0004857~enzyme inhibitor activity	3.90	0.007545
GOTERM_CC_ALL	GO:0042627~chylomicron	19.68	0.008264
GOTERM_CC_ALL	GO:0042383~sarcolemma	4.52	0.008266
KEGG_PATHWAY	hsa05150:Staphylococcus aureus infection	9.31	0.009080
GOTERM_BP_ALL	GO:0051240~positive regulation of multicellular organismal process	2.21	0.009318
GOTERM_CC_ALL	GO:0031012~extracellular matrix	2.53	0.009813
GOTERM_BP_ALL	GO:0032963~collagen metabolic process	7.36	0.010525
GOTERM_MF_ALL	GO:0032403~protein complex binding	2.19	0.010869
GOTERM_MF_ALL	GO:0004866~endopeptidase inhibitor activity	5.45	0.010869
GOTERM_MF_ALL	GO:0030414~peptidase inhibitor activity	5.45	0.010869
GOTERM_BP_ALL	GO:0045723~positive regulation of fatty acid biosynthetic process	24.17	0.011280
GOTERM_BP_ALL	GO:0045861~negative regulation of proteolysis	3.72	0.011280
GOTERM_BP_ALL	GO:0018149~peptide cross-linking	13.43	0.012212
GOTERM_BP_ALL	GO:0010810~regulation of cell-substrate adhesion	3.97	0.013184
GOTERM_BP_ALL	GO:1903524~positive regulation of blood circulation	9.07	0.013340
REACTOME_PATHWAY	R-HSA-166665~Terminal pathway of complement	19.68	0.013860
GOTERM_BP_ALL	GO:0044259~multicellular organismal macromolecule metabolic process	6.77	0.015915

## Optimising the statistical pipeline for quantitative proteomics

REACTOME_PATHWAY	R-HSA-1442490~Collagen degradation	7.38	0.017097
REACTOME_PATHWAY	R-HSA-1650814~Collagen biosynthesis and modifying enzymes	10.25	0.017097
GOTERM_BP_ALL	GO:0022407~regulation of cell-cell adhesion	3.80	0.018268
REACTOME_PATHWAY	R-HSA-198933~Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	7.03	0.018609
GOTERM_BP_ALL	GO:0050778~positive regulation of immune response	2.44	0.019707
GOTERM_BP_ALL	GO:0045765~regulation of angiogenesis	4.63	0.020064
GOTERM_BP_ALL	GO:0007159~leukocyte cell-cell adhesion	3.41	0.021571
GOTERM_MF_ALL	GO:0061135~endopeptidase regulator activity	4.85	0.021985
GOTERM_BP_ALL	GO:0003018~vascular process in circulatory system	6.27	0.022246
GOTERM_BP_ALL	GO:0045055~regulated exocytosis	2.97	0.022246
GOTERM_BP_ALL	GO:0034367~macromolecular complex remodeling	19.34	0.022246
GOTERM_BP_ALL	GO:0034369~plasma lipoprotein particle remodeling	19.34	0.022246
GOTERM_BP_ALL	GO:0034368~protein-lipid complex remodeling	19.34	0.022246
GOTERM_BP_ALL	GO:0051917~regulation of fibrinolysis	19.34	0.022246
GOTERM_BP_ALL	GO:0051918~negative regulation of fibrinolysis	19.34	0.022246
GOTERM_BP_ALL	GO:0010951~negative regulation of endopeptidase activity	3.96	0.023908
GOTERM_BP_ALL	GO:1901342~regulation of vasculature development	4.44	0.023908
GOTERM_CC_ALL	GO:0034385~triglyceride-rich lipoprotein particle	14.06	0.024134
GOTERM_CC_ALL	GO:0034361~very-low-density lipoprotein particle	14.06	0.024134
GOTERM_MF_ALL	GO:0004867~serine-type endopeptidase inhibitor activity	6.53	0.024920
GOTERM_CC_ALL	GO:0044433~cytoplasmic vesicle part	2.10	0.025271
GOTERM_BP_ALL	GO:0010466~negative regulation of peptidase activity	3.90	0.026716
GOTERM_CC_ALL	GO:0000786~nucleosome	8.20	0.027397
GOTERM_CC_ALL	GO:0099503~secretory vesicle	2.36	0.027397
GOTERM_BP_ALL	GO:0001775~cell activation	2.42	0.028203
GOTERM_BP_ALL	GO:0051241~negative regulation of multicellular organismal process	2.40	0.029612
GOTERM_BP_ALL	GO:1903317~regulation of protein maturation	5.83	0.029612
GOTERM_BP_ALL	GO:0070613~regulation of protein processing	5.83	0.029612
GOTERM_BP_ALL	GO:0006641~triglyceride metabolic process	7.25	0.032172

## *Optimising the statistical pipeline for quantitative proteomics*

GOTERM_BP_ALL	GO:0030574~collagen catabolic process	10.07	0.032573
GOTERM_CC_ALL	GO:0044815~DNA packaging complex	7.69	0.034218
GOTERM_BP_ALL	GO:0071827~plasma lipoprotein particle organization	16.12	0.036398
GOTERM_BP_ALL	GO:0031639~plasminogen activation	16.12	0.036398
GOTERM_BP_ALL	GO:0045923~positive regulation of fatty acid metabolic process	16.12	0.036398
GOTERM_BP_ALL	GO:0030162~regulation of proteolysis	2.26	0.036398
GOTERM_BP_ALL	GO:0006639~acylglycerol metabolic process	6.91	0.036398
GOTERM_BP_ALL	GO:0035150~regulation of tube size	6.91	0.036398
GOTERM_BP_ALL	GO:0002698~negative regulation of immune effector process	6.91	0.036398
GOTERM_BP_ALL	GO:0050880~regulation of blood vessel size	6.91	0.036398
GOTERM_BP_ALL	GO:0097006~regulation of plasma lipoprotein particle levels	9.30	0.040279
GOTERM_BP_ALL	GO:0006638~neutral lipid metabolic process	6.59	0.044366
REACTOME_PATHWAY	R-HSA-166663~Initial triggering of complement	7.69	0.045478
GOTERM_BP_ALL	GO:0001944~vasculature development	2.66	0.047120
GOTERM_BP_ALL	GO:0010811~positive regulation of cell-substrate adhesion	4.40	0.047417
GOTERM_CC_ALL	GO:0005589~collagen type VI trimer	24.60	0.047665
GOTERM_CC_ALL	GO:0098647~collagen beaded filament	24.60	0.047665
GOTERM_CC_ALL	GO:0034366~spherical high-density lipoprotein particle	24.60	0.047665
GOTERM_BP_ALL	GO:0051270~regulation of cellular component movement	2.25	0.049221

## Optimising the statistical pipeline for quantitative proteomics

Table 29; Enrichment results for optimal t-test analysis with AIN and a significance threshold of 0.05 for dataset PXD007592. Access to the DAVID results for all parameter combinations is provided in the link in the Supplementary Material section.

Category	Term	Fold Enrichment	Benjamini p-value
REACTOME_PATHWAY	R-HSA-5607764~CLEC7A (Dectin-1) signaling	5.86	0.038094
REACTOME_PATHWAY	R-HSA-2871837~FCERI mediated NF-kB activation	6.12	0.038094
REACTOME_PATHWAY	R-HSA-202424~Downstream TCR signaling	5.99	0.038094
REACTOME_PATHWAY	R-HSA-5621481~C-type lectin receptors (CLRs)	4.95	0.038094
REACTOME_PATHWAY	R-HSA-389957~Prefoldin mediated transfer of substrate to CCT/TriC	9.18	0.038094
REACTOME_PATHWAY	R-HSA-389958~Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding	8.44	0.039624
REACTOME_PATHWAY	R-HSA-8951664~Neddylation	4.40	0.039624
REACTOME_PATHWAY	R-HSA-202403~TCR signaling	5.03	0.039624
REACTOME_PATHWAY	R-HSA-5668541~TNFR2 non-canonical NF-kB pathway	6.01	0.039624
REACTOME_PATHWAY	R-HSA-5676590~NIK-->noncanonical NF-kB signaling	6.01	0.039624
REACTOME_PATHWAY	R-HSA-983168~Antigen processing: Ubiquitination & Proteasome degradation	4.28	0.039624
REACTOME_PATHWAY	R-HSA-9604323~Negative regulation of NOTCH4 signaling	5.86	0.039624
REACTOME_PATHWAY	R-HSA-5607761~Dectin-1 mediated noncanonical NF-kB signaling	5.86	0.039624
REACTOME_PATHWAY	R-HSA-8948751~Regulation of PTEN stability and activity	5.47	0.041565
REACTOME_PATHWAY	R-HSA-69656~Cyclin A:Cdk2-associated events at S phase entry	5.47	0.041565
REACTOME_PATHWAY	R-HSA-69202~Cyclin E associated events during G1/S transition	5.47	0.041565
REACTOME_PATHWAY	R-HSA-8878159~Transcriptional regulation by RUNX3	5.47	0.041565
REACTOME_PATHWAY	R-HSA-6807070~PTEN Regulation	4.54	0.041565
REACTOME_PATHWAY	R-HSA-9013694~Signaling by NOTCH4	5.35	0.041565
REACTOME_PATHWAY	R-HSA-983169~Class I MHC mediated antigen processing & presentation	3.45	0.041565
REACTOME_PATHWAY	R-HSA-1257604~PIP3 activates AKT signaling	3.81	0.041565
REACTOME_PATHWAY	R-HSA-2454202~Fc epsilon receptor (FCERI) signaling	4.33	0.041565
REACTOME_PATHWAY	R-HSA-382556~ABC-family proteins mediated transport	5.03	0.041565
REACTOME_PATHWAY	R-HSA-195253~Degradation of beta-catenin by the destruction complex	4.83	0.041565
REACTOME_PATHWAY	R-HSA-5687128~MAPK6/MAPK4 signaling	4.83	0.041565



## *Optimising the statistical pipeline for quantitative proteomics*

REACTOME_PATHWAY	R-HSA-1280218~Adaptive Immune System	2.37	0.041565
REACTOME_PATHWAY	R-HSA-9020702~Interleukin-1 signaling	4.65	0.041565
REACTOME_PATHWAY	R-HSA-350562~Regulation of ornithine decarboxylase (ODC)	5.55	0.041565
REACTOME_PATHWAY	R-HSA-69610~p53-Independent DNA Damage Response	5.55	0.041565
REACTOME_PATHWAY	R-HSA-69601~Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	5.55	0.041565
REACTOME_PATHWAY	R-HSA-4641257~Degradation of AXIN	5.55	0.041565
REACTOME_PATHWAY	R-HSA-349425~Autodegradation of the E3 ubiquitin ligase COP1	5.55	0.041565
REACTOME_PATHWAY	R-HSA-1236978~Cross-presentation of soluble exogenous antigens (endosomes)	5.55	0.041565
REACTOME_PATHWAY	R-HSA-69541~Stabilization of p53	5.55	0.041565
REACTOME_PATHWAY	R-HSA-69613~p53-Independent G1/S DNA damage checkpoint	5.55	0.041565
REACTOME_PATHWAY	R-HSA-201681~TCF dependent signaling in response to WNT	3.80	0.041565
REACTOME_PATHWAY	R-HSA-211733~Regulation of activated PAK-2p34 by proteasome mediated degradation	5.41	0.041565
REACTOME_PATHWAY	R-HSA-174154~APC/C:Cdc20 mediated degradation of Securin	5.41	0.041565
REACTOME_PATHWAY	R-HSA-174084~Autodegradation of Cdh1 by Cdh1:APC/C	5.41	0.041565
REACTOME_PATHWAY	R-HSA-180534~Vpu mediated degradation of CD4	5.41	0.041565
REACTOME_PATHWAY	R-HSA-69563~p53-Dependent G1 DNA Damage Response	5.41	0.041565
REACTOME_PATHWAY	R-HSA-5362768~Hh mutants are degraded by ERAD	5.41	0.041565
REACTOME_PATHWAY	R-HSA-69615~G1/S DNA Damage Checkpoints	5.41	0.041565
REACTOME_PATHWAY	R-HSA-4641258~Degradation of DVL	5.41	0.041565
REACTOME_PATHWAY	R-HSA-174178~APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	5.41	0.041565
REACTOME_PATHWAY	R-HSA-69580~p53-Dependent G1/S DNA damage checkpoint	5.41	0.041565
REACTOME_PATHWAY	R-HSA-4608870~Asymmetric localization of PCP proteins	5.41	0.041565
REACTOME_PATHWAY	R-HSA-75815~Ubiquitin-dependent degradation of Cyclin D	5.41	0.041565
REACTOME_PATHWAY	R-HSA-5387390~Hh mutants abrogate ligand secretion	5.41	0.041565
REACTOME_PATHWAY	R-HSA-8878166~Transcriptional regulation by RUNX2	4.40	0.041565
REACTOME_PATHWAY	R-HSA-450531~Regulation of mRNA stability by proteins that bind AU-rich elements	4.32	0.041565
REACTOME_PATHWAY	R-HSA-176814~Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins	5.28	0.041565
REACTOME_PATHWAY	R-HSA-5358346~Hedgehog ligand biogenesis	5.28	0.041565
REACTOME_PATHWAY	R-HSA-351202~Metabolism of polyamines	5.28	0.041565

## *Optimising the statistical pipeline for quantitative proteomics*

REACTOME_PATHWAY	R-HSA-169911~Regulation of Apoptosis	5.28	0.041565
REACTOME_PATHWAY	R-HSA-176409~APC/C:Cdc20 mediated degradation of mitotic proteins	5.28	0.041565
REACTOME_PATHWAY	R-HSA-179419~APC:Cdc20 mediated degradation of cell cycle proteins prior to satisfaction of the cell cycle checkpoint	5.28	0.041565
REACTOME_PATHWAY	R-HSA-8941858~Regulation of RUNX3 expression and activity	5.28	0.041565
REACTOME_PATHWAY	R-HSA-8854050~FBXL7 down-regulates AURKA during mitotic entry and in early mitosis	5.28	0.041565
REACTOME_PATHWAY	R-HSA-174113~SCF-beta-TrCP mediated degradation of Emi1	5.28	0.041565
REACTOME_PATHWAY	R-HSA-69017~CDK-mediated phosphorylation and removal of Cdc6	5.28	0.041565
REACTOME_PATHWAY	R-HSA-174184~Cdc20:Phospho-APC/C mediated degradation of Cyclin A	5.28	0.041565
REACTOME_PATHWAY	R-HSA-157118~Signaling by NOTCH	3.70	0.041619
REACTOME_PATHWAY	R-HSA-69206~G1/S Transition	4.25	0.042239
REACTOME_PATHWAY	R-HSA-9762114~GSK3B and BTRC:CUL1-mediated-degradation of NFE2L2	5.15	0.042239
REACTOME_PATHWAY	R-HSA-5678895~Defective CFTR causes cystic fibrosis	5.15	0.042239
REACTOME_PATHWAY	R-HSA-180585~Vif-mediated degradation of APOBEC3G	5.15	0.042239
REACTOME_PATHWAY	R-HSA-187577~SCF(Skp2)-mediated degradation of p27/p21	5.15	0.042239
REACTOME_PATHWAY	R-HSA-1234174~Cellular response to hypoxia	5.03	0.042573
REACTOME_PATHWAY	R-HSA-1234176~Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	5.03	0.042573
REACTOME_PATHWAY	R-HSA-5619084~ABC transporter disorders	5.03	0.042573
REACTOME_PATHWAY	R-HSA-5610780~Degradation of GLI1 by the proteasome	5.03	0.042573
REACTOME_PATHWAY	R-HSA-5632684~Hedgehog 'on' state	5.03	0.042573
REACTOME_PATHWAY	R-HSA-446652~Interleukin-1 family signaling	4.10	0.042573
REACTOME_PATHWAY	R-HSA-162909~Host Interactions of HIV factors	3.52	0.042573
REACTOME_PATHWAY	R-HSA-176408~Regulation of APC/C activators between G1/S and early anaphase	4.91	0.042573
REACTOME_PATHWAY	R-HSA-5610785~GLI3 is processed to GLI3R by the proteasome	4.91	0.042573
REACTOME_PATHWAY	R-HSA-453276~Regulation of mitotic cell cycle	4.91	0.042573
REACTOME_PATHWAY	R-HSA-174143~APC/C-mediated degradation of cell cycle proteins	4.91	0.042573
REACTOME_PATHWAY	R-HSA-1169091~Activation of NF-kappaB in B cells	4.91	0.042573
REACTOME_PATHWAY	R-HSA-5610783~Degradation of GLI2 by the proteasome	4.91	0.042573
REACTOME_PATHWAY	R-HSA-450408~AUF1 (hnRNP D0) binds and destabilizes mRNA	4.91	0.042573
REACTOME_PATHWAY	R-HSA-69242~S Phase	3.97	0.0459

## *Optimising the statistical pipeline for quantitative proteomics*

REACTOME_PATHWAY	R-HSA-8939902~Regulation of RUNX2 expression and activity	4.80	0.0459
REACTOME_PATHWAY	R-HSA-9755511~KEAP1-NFE2L2 pathway	3.91	0.048362
REACTOME_PATHWAY	R-HSA-453279~Mitotic G1 phase and G1/S transition	3.91	0.048362
REACTOME_PATHWAY	R-HSA-5658442~Regulation of RAS by GAPs	4.69	0.048814

**Overall performance of normalisation methods**

Table 30; Rank of normalisation methods based on number significant terms from enrichment analysis. Overall ranking and ranking for DE methods QPROT and *t*-Test.

QPROT	Score	<i>t</i> -test	Score	Overall	Score
Quantile - mean	27	AIN	28	Log2 transformed	43
Log2 transformed	24	TIN	24	Loess	42
VSN	23	Loess	21	Quantile - mean	42
Quantile - median	22	Log2 transformed	19	AIN	37
Loess	21	Progenesis	18	Quantile - median	33
MIN	21	Quantile - mean	15	MIN	31
QPROT	20	RLR	13	TIN	31
RLR	12	Quantile - median	11	Progenesis	30
Progenesis	12	MIN	10	VSN	29
AIN	9	VSN	6	RLR	25
TIN	7			QPROT	20

The overall ranking and the ranking by dataset of the normalisation methods is shown in Table 30. Methods were ranked from 10 – 0 and 10 – 1 for QPROT and *t*-test, respectively, according to their performance in each dataset at the thresholds used in section 3.3, and the rankings summed. The best overall normalisation for QPROT was Quantile–mean, followed by log2 transformation, VSN, and Quantile-median. Interestingly, QPROT’s own normalisation method performed less well. Although this method is based on the general method for quantile normalisation, instead of setting each data point to the mean or the median value across samples, QPROT sets the protein abundance at the 0 to 100% percentile data points to the median value across samples and then interpolates the protein values in between to the value of the nearest percentile point. This method seems to provide poorer results than Quantile-mean, which, although ranked in the lower half for *t*-test analysis, was ranked third overall. *t*-test analysis was optimal using AIN; however, this method proved unsuccessful with QPROT analysis, resulting in a overall ranking of 29. VSN performed well with QPROT analysis but not with the *t*-test, (23 and 6). The best overall performance was with log2 transformed data (28 and 9), followed by Loess (21 and 21), Quantile – mean (27 and 15), AIN (28 and 9) and Quantile – median (22 and 11). However, apart from log transformation, the best normalisation method appears to differ depending on what DE analysis you are performing.

The impact of normalisation appears to be of previously unappreciated importance in the success of DE analysis. Overall, an alternative normalisation method improved on the results where only Progenesis normalisation had been applied. In the *t*-test analysis of dataset PXD007592, there had been a maximum of one significant term, which was rescued using several of the methods in 3.3. However, the best normalisation method varied widely and depended heavily on the selected threshold for significance. Normalisation reviews have previously found no consensus on the best normalisation method (Tokareva et al., 2021), and as suggested by Chawade et al. (2014), the appropriate normalisation method appears to be dependent on the intrinsic characteristics of the data.

## **vi. Review of benchmarking**

Established software evaluation methods for quantitative proteomics requires 'ground truth' data. Existing benchmarking datasets widely used for software evaluation can be inaccurate and imprecise giving skewed results and providing a poor basis for method analysis. This chapter introduced a novel method for benchmarking quantitative proteomics software without the need for ground truth data by using the results of pathway analysis as a metric for successful DE analysis. Limitations of this benchmarking method could arise due to the amount of redundancy in functional analysis and a small count of additional true positive proteins, could lead to several very similar pathways or terms being labelled as significant. Also, the results of the analysis are only as good as the annotation and quality of database used. As a method for benchmarking its results are still valid as each DE and normalisation method being assessed is provided with the same level of accuracy in knowledge database used for evaluation.

The proteome is a diverse and dynamic system with multi-dimensional, interconnected properties. Proteomics experimental datasets are heterogenic, but there is often a homogenous approach to their statistical analysis, with the aim of benchmarking being to discover a gold-standard approach. Assumptions

of statistical methods used for DE and normalisation, such as normality, linearity, and homogeneity of variance, will hold true to different extents depending on the properties of the data, resulting in different degrees of successful analysis. Research goals, exploratory or verification studies, dictate the size of the dataset. Comprehensive characterisation of the proteome requires large-scale analysis of integrated data (Reiter et al., 2009) resulting in a vast number of proteins compared to samples. Whereas targeted analysis for disease-specific biomarkers will generally feature a much smaller number of proteins (Maes et al., 2016), providing much smaller ratio of proteins to samples. Often there is an analysis assumption that most proteins in the sample are equally expressed and only a small proportion are changing (Cox et al., 2014). However, datasets vary in the proportion of its proteins changing across experimental conditions, the direction of change (proteins could be either up- or down-regulated), and in magnitude of DE compared to magnitude of background variation.

A dataset with a large amount of variation in the background proteome creates noise that can mask DE proteins. A fold change significance threshold can be put in place to avoid this. However, the change in abundance of proteins of interest may be subtle and so using a fold-change threshold or ranking based on fold-change may result in missing vital information. The methodology of label-free proteomics means there is a high degree of introduced variation due to sample runs being runs processed in the MS separately, causing variation between replicates. Statistical methods have the ability to deal with heterogeneous variation. The Welch *t*-test is capable of comparing means between groups without assuming equal population variances. However, variation is not always constant; lower abundant proteins display a higher amount of variance (Mahoney et al., 2011). Furthermore, there is an increase in Type I errors with Welch's *t*-test when the data has corresponding issues such as non-normal distribution or unequal group sizes (Ahad and Yahaya, 2014, Zimmerman and Zumbo, 1993), and Type I error rates are greater when sample size is small and when cut-off level for significance is stringent (Zimmerman, 2004).

### ***Significance threshold selection***

Optimal significance threshold varied from 0.007 to the least stringent value analysed, the conventionally accepted 0.05. Intuitively, a stringent cut-off will give a concentrated pool of changing proteins, and decreasing threshold stringency will mean a trade off by increasing the chance of including false positives. However, the results suggest that in some circumstances, a less conservative significance threshold may have improved output. Biological data will naturally have variation, possibly with background proteins incidentally showing a higher abundance across conditions. Methods that work best with stringent cut-off may be preferable with this type of data. Limiting the significance threshold at a least stringent level of 0.05 also appears to have been insufficient. For dataset PXD004501, many of the normalisation methods in combination with the *t*-test are only just starting to call proteins DE at this level. Further investigations should include thresholds beyond 0.05. The results in Chapter Three demonstrated the wide range of significance thresholds for defining significance required for optimal output of DE analysis and finding the correct threshold is a delicate process with optimal significance threshold varying from 0.001 to 0.05, the highest threshold applied. The range examined appeared to be both not broad enough at times and have circumstances where an even smaller or larger threshold may have improved results.

A limitation of our analysis was using different threshold measurements for QPROT (FDR) and *t*-test and MSstats (BH corrected *p*-values). This made direct comparison difficult and, as discussed above, may have prevented a more optimised output with the application of a further decreased cut-off threshold. A solution to this would have been to use gradually increasing proportions of proteins to be classified as DE after ranking by test statistic. However, the aim of the benchmarking exercise was to try to mimic what a biologist would do, investigating the appropriateness of the standard accepted threshold of 0.05. For this reason, ranking was not originally considered, but its application will be investigated in further work.

Much research has been performed over recent years in attempting to develop an optimal proteomics analysis tool, with further research in evaluating and

## ***Optimising the statistical pipeline for quantitative proteomics***

assessing available statistical method for quantitative proteomics (Välikangas et al., 2017). However, there is still no single gold standard approach (Gatto et al., 2016). The results of Chapter Three highlight the diversity in effectiveness of processing methods, with not only DE analysis but also normalisation and selection of significance threshold for significance being key to successful results. Furthermore, the aim of a quantitative proteomics experiment is not to provide a static and binary group of proteins that have been ‘proven’ to be changing. DE results are followed up with validation and investigation. The DE experiment is essentially a scoping exercise. Dalman et al. (2012) state that DE analysis can provide more than one answer, and that data interpretation as more of an art than a science. In proteomics, statistical tests do not need to tell us categorically that a protein is different or the same, only to point us in a sensible direction for further investigation. The outcome of this thesis focuses on the end goal of proteomics experiments, the detection of functionally related changing proteins, and utilises this for analysis evaluation. We propose that effective DE is not through a single statistical algorithm, but an integration of available methods, combined with a means of evaluation. By having a useful and desired outcome, detecting functionally related proteins that change, the pipeline described in this chapter is able to select most appropriate methods according to the properties of the data

### ***Pathway analysis***

The benchmarking pipeline in Chapter Three used the pathway analysis to evaluate the DE results and was conducted using the RDAVIDWebService package in R (Fresno and Fernández, 2013). The package allows access to the web-based Database for Annotation, Visualisation and Integrated Discovery (DAVID) (Huang da et al., 2009a) through R. There were technical issues associated with using of the RDAVID package; the multiple calls to querying the web database in the pipeline lead to unpredictable crashing of the software and the daily job limit was regularly exceeded when performing large analyses. In Bioconductor version 3.14, RDAVIDWebService package became deprecated and is no longer supported. An alternative API for accessing DAVID has been



provided the developers. However, this still requires accessing a web-based database and the specifications state that it is only for light-duty jobs (no more than 400 proteins in search list, not for looping in scripts, maximum 200 hits per day). For this reason, continuing with DAVID for enrichment analysis in an R based pipeline was not sensible.

There are also concerns about the completeness of DAVID enrichment analysis due to irregular updates. The Gene Ontology (GO) (Ashburner et al., 2000) is updated daily. Functional identity of genes changes over time, with semantic similarity measures showing that 20% of genes do not match themselves after two years (Gillis and Pavlidis, 2013). This instability of GO enrichment results means that functional annotation databases can become outdated without regular maintenance. Recent versions of enrichment tools are required for consistent results due to rapidly evolving ontology and annotations (Tomczak et al., 2018). Although recently updated in November 2021, the previous update of DAVID had not been since 2016, and the update before that was 2010. To demonstrate the impact of using outdated gene annotations in research, Wadi et al. (2016) compared pathway analysis of essential genes of 77 breast cancer cell lines (Marcotte et al., 2016) with annotations from the 2010 and 2016 versions of DAVID. They found that 74% of 2016 enriched terms were missed when using the oldest version of the software. The number of human biological process and molecular pathways in the database doubled between the 2010 and 2016 updates. In benchmarking, Zhou et al. (2019) estimated that 4-10% of their input gene candidates of their analysis were not recognized by the current release of DAVID (at the time 2016 version), and that 44% of the human GO annotations records were in need of updating.

### **3.4. Conclusions**

In this chapter we presented a novel method for benchmarking which utilised functional enrichment analysis to evaluate statistic methods using biological data. Overall there was no consensus on best method for DE, normalisation

method or threshold cut-off and the correct combination of parameters appeared to be dependent on the characteristics of the individual datasets.

The *t*-test performed best on the largest dataset with a lot of the proteins changing in intensity across the conditions, often with the least stringent significance threshold. QPROT gave better pathway analysis results when there appeared there are fewer proteins of interest within the data and required conservative significance thresholds for optimal performance. Due to the peptide level input required, MSstats evaluation was only performed without the normalisation factor. The results showed very similar performance to the *t*-test, and it appeared that the advanced statistical model may have complexity that is not required for the pairwise label-free comparison, and may be better utilised in a more complicated experimental design such as time-course, with multiple factors, or in label-based set-ups.

Investigations in this chapter suggest that appropriate DE method is dependent of individual characters of the data. The proportion of proteome changing due to experimental conditions and how large the changes in intensity are compared to background intensities appears to have an effect on statistical analysis. The analysis methods assigned a numeric value to the proteins, allowing the application of a threshold to decide which of the proteins are significant. This was not always the same value for all datasets. The analysis methods ranked the proteins in order of significance. This was not the same order for each method and no one method was consistently the same. Overall there was no best pipeline for analysis of every given dataset and different combinations of methods produced drastically different results for different datasets. The results led to the development of a data-driven statistical pipeline for optimal results (described in Chapter Four). Simultaneous analysis performed in with all possible workflow chains in parallel and the combination of parameters that provided the highest number of functionally related proteins in the differentially expressed set returned to the user.

## **Chapter 4. Optimised proteomics pipeline**

### **4.1. Introduction**

#### **i. Abstract**

In the evaluation of normalisation, DE analysis, and selection of threshold for defining significance (Chapter 3), there was no overall best method. Different properties of data require the application of different analysis methods and the pursuit of a single gold standard approach appears to be currently outside of reach. The outcome of this evaluation was the development of a pipeline that uses a high performance computing cluster to easily perform all possible parameter options and apply an evaluation metric in order to return the results from best combination of analysis methods to the user. In this chapter we report the development, implementation and validation of this pipeline and demonstrate its superior performance to the present analysis output of Progenesis QIP.

The pipeline provides eight possible methods for normalisation using the Normalyzer DE package. Inspired by the principles employed by QPROT, this chapter implements an improved method of DE analysis called here *BayesianT*. Written in R and utilising Stan for Bayesian modelling through No U Turn sampling, this algorithm provides local FDR through semi-parametric mixed modelling and kernel density estimation. Groups of proteins are defined as DE by a range of incrementally increasing significance thresholds, and each group's functional enrichment is calculated using the clusterProfiler package in R. The parameter options creating the group of most functionally related proteins is deemed the best and the results are returned to the user.

Validation of the methods showed that the BayesianT algorithm provided superior DE analysis (and identical FDR estimation) to QPROT. The application of the `simplify()` function in clusterProfiler reduced the issue of redundancy in enrichment analysis, and analysis with the pipeline provided an improved output to the standard analysis from Progenesis QIP.

## **ii. Proteomics pipeline development**

Results of DE analysis provide researchers with proteins of interest, which become the focus of further investigation. Pathway analysis provides meaning to DE experiments by attaching biological context to the proteins detected as changing across different experimental conditions. By highlighting statistically enriched biological processes in the data, biologists can focus further studies on the proteins involved in these processes. Chapter Three described the development of benchmarking software that utilised enrichment results to evaluate DE methods. The results indicated that there is no one best analysis pipeline and appropriate methods are dependent on the individual characteristics of the data. The results also demonstrated the importance of normalisation in the success of DE analysis, and that the use of an arbitrary significance threshold does not always provide the optimal output. These findings indicate that the sensible approach to analysis is to enable users to easily investigate a wide range of parameter options. In this chapter, we describe the development of a statistical pipeline where simultaneous analysis of all possible workflow chains is performed on a high-performance computing cluster to allow parallel initiation. Evaluation of methods is performed using pathway analysis and a summary of the functional enrichment of DE proteins is returned to the user.

### ***Pipeline Workflow***

The optimised proteomics pipeline workflow (Figure 4.1.1) was programmed in R (version 4.1.2) and is divided into the following sections; log transformation, normalisation, DE analysis, threshold selection for defining significant results, pathway analysis on changing proteins, and evaluation of the results. In chapter 3, log transformation occurred after normalisation to allow for like for like comparison with QPROT analysis which performs log transformation after normalisation. However, as our pipeline does not use QPROT, log transformation was performed before normalisation, according to the Normalyzer workflow.

## Optimising the statistical pipeline for quantitative proteomics

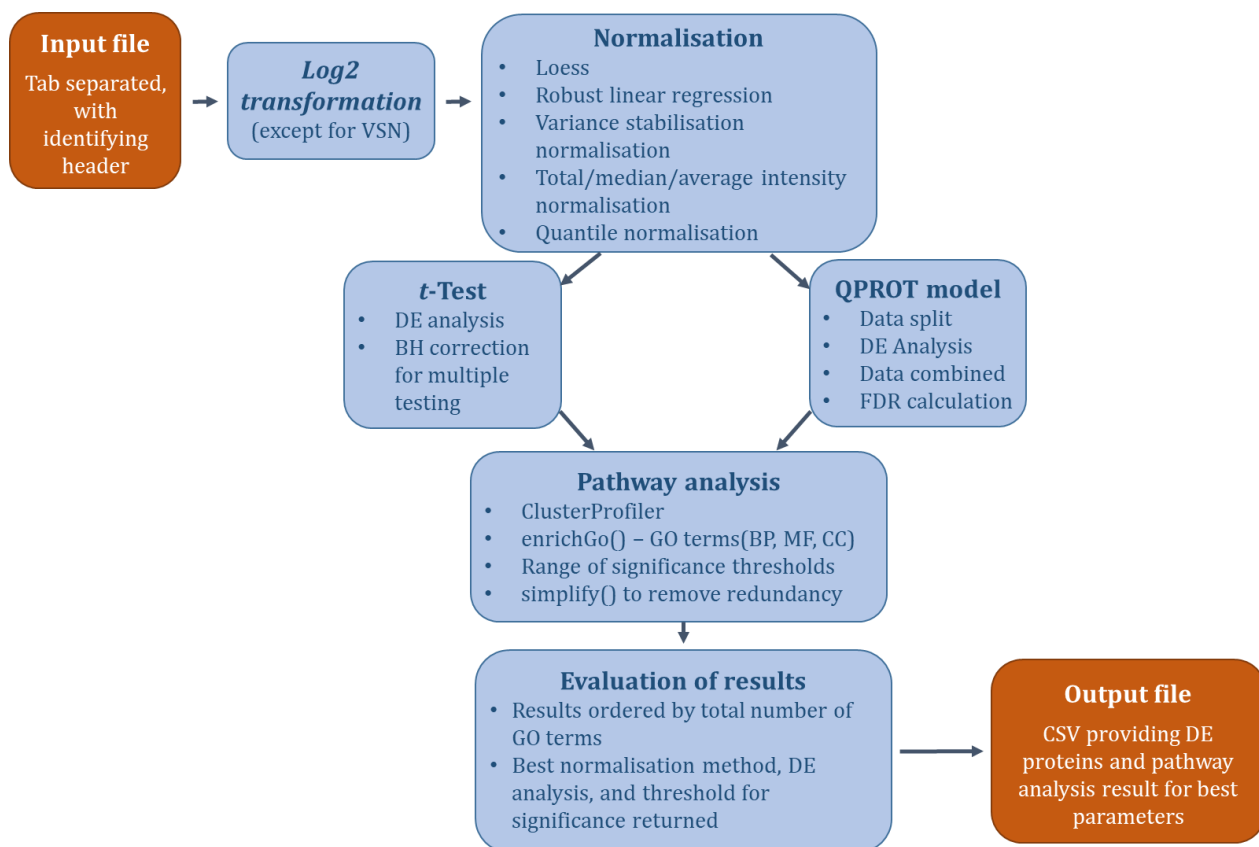


Figure 4.1.1; Simplified scheme of the optimised proteomics pipeline. Tab separated, protein group abundance data with a header identifying the comparison groups is normalised by several methods. Each normalisation output has differential expression (DE) analysis by t-test and QPROT. DE proteins are subject to enrichment analysis with the R package ClusterProfiler using a range of significance thresholds to define significant DE. The combination of methods providing the greatest number of significant GO terms is returned to the user, along with details of the proteins changing between conditions and their functionally related pathways.

### High-performance computing (HPC)

Large-scale proteomics analysis is computationally intensive with vast numbers of proteins requiring individual analysis. Working through each sample one by one is time-consuming. The use of high-throughput technology allows simultaneous analysis which decreases runtime but requires powerful hardware. HPC architectures are a collection of servers which work in parallel to streamline complex algorithms efficiently. The pipeline is available to download from GitHub

<https://github.com/HayleyPrice/Pipeline/tree/main/Server>. To install dependencies, Rstan and to compile the Stan model run the Bash script 'installs.sh' (Figure 4.1.2).

## Optimising the statistical pipeline for quantitative proteomics

```

hprice@head:/mnt/hc-storage/users/hprice/Pipeline
Using username "hprice".
hprice@138.253.198.218's password:
Last login: Tue May  3 13:49:36 2022 from 10.67.8.244
Welcome to the joint PGB-CBF Cluster

Total cluster capacity: 360 cores
Current cluster availability: 62 cores
Current maximum available slot size: 38 cores
Your jobs:
      JOBID PARTITION   NAME   USER ST   TIME  NODES NODELIST(REA
SON)
[hprice@head ~]$ cd /mnt/hc-storage/users/hprice/Pipeline
[hprice@head Pipeline]$ bash installs.sh /mnt/hc-storage/users/hprice/

```

Figure 4.1.2; Command line installation of the packages required for running the pipeline.

The pipeline is run through R scripts on a LINUX server to provide parallelisation of the many parameter combinations. Details of the number of files being analysed at each step of the workflow are summarised in Figure 4.1.3.

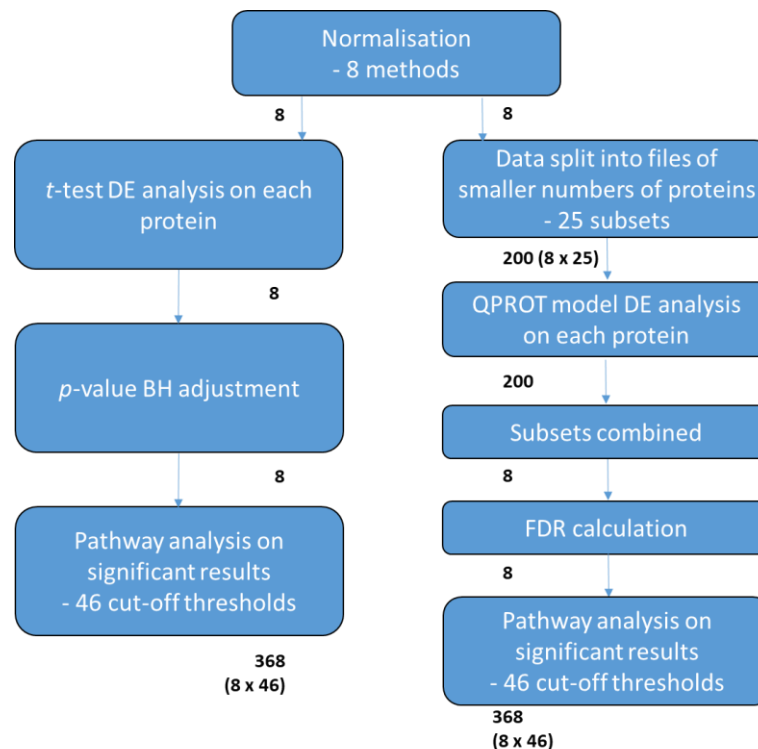


Figure 4.1.3; Schematic of parameter combinations at each stage of the workflow. Text in black shows the number of different files being analysed at the step above. Stages run in parallel on all available nodes of computer cluster.

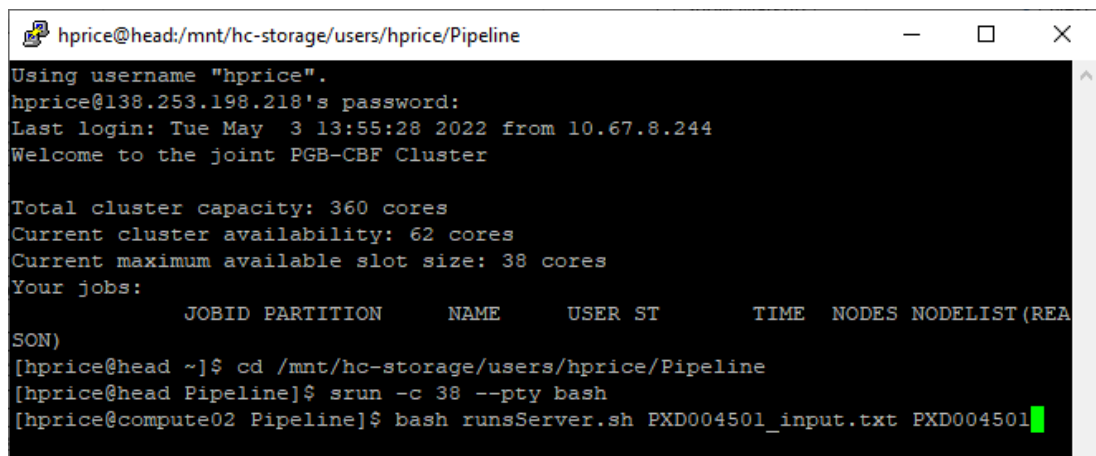
## Optimising the statistical pipeline for quantitative proteomics

The pipeline takes input a pairwise comparison of identified, quantified, and grouped protein abundances in a tab separated file. An additional identifier line must be added (Figure 4.1.4) to provide details of the pairwise comparison.

	A	B	C	D	E	F	G	H
1	0	1	1	1	2	2	2	
2	Proteins	C1	C2	C3	N1	N2	N3	
3	sp P04114	2.15E+10	1.78E+10	2.31E+10	2.01E+10	1.91E+10	2.44E+10	
4	sp P01024	1.01E+11	8.87E+10	8.89E+10	8.67E+10	8.02E+10	8.10E+10	
5	sp P0COL5	2.20E+09	2.67E+09	3.04E+09	1.24E+09	1.47E+09	1.48E+09	
6	cn A1R11	2.50E+11	3.94E+11	2.63E+11	2.32E+11	1.60E+11	1.99E+11	

Figure 4.1.4; Identifying header of tab separated input file. Group comparisons are indicated by letters 1 and 2.

Running of the pipeline is initiated with the Bash script 'runServer.sh' with arguments of input file name and file name for results files (Figure 4.1.5).



```
hprice@head:/mnt/hc-storage/users/hprice/Pipeline
Using username "hprice".
hprice@138.253.198.218's password:
Last login: Tue May  3 13:55:28 2022 from 10.67.8.244
Welcome to the joint PGB-CBF Cluster

Total cluster capacity: 360 cores
Current cluster availability: 62 cores
Current maximum available slot size: 38 cores
Your jobs:
      JOBID PARTITION      NAME      USER ST      TIME  NODES NODELIST(REA
SON)
[hprice@head ~]$ cd /mnt/hc-storage/users/hprice/Pipeline
[hprice@head Pipeline]$ srun -c 38 --pty bash
[hprice@compute02 Pipeline]$ bash runServer.sh PXD004501_input.txt PXD004501
```

Figure 4.1.5; Command line installation of the packages required for running the pipeline.

## Normalisation

The data was normalised using modified R code from the NormalyzerDE package in R providing eight methods as described in Chapter 3; log<sub>2</sub> transformation, loess normalisation, robust linear regression normalisation, variance stabilisation normalisation, total intensity normalisation, median intensity normalisation, average intensity normalisation, and quantile normalisation.

### ***Differential expression analysis***

The normalised protein abundance data were tested for DE. The abundance of each protein across the two samples were compared using the following methods:

**BayesianT** – A model was implemented combining the principles employed in QPROT’s algorithm for DE and FDR calculation, as described in Chapter 1, combined with a Hamiltonian based method for sampling, No U Turn sampling (NUTs) (Homan and Gelman, 2014), from the posterior distribution, described next. The model was written in R (version 4.1.0) and Stan (version 2.21.3) (Carpenter et al., 2017). Compiling of the Stan model is performed during the installation script. Modelling is performed individually on each protein in the data. To reduce computational time caused by repeated sampling, data is split into 25 subsets to allow parallelisation of DE analysis. After DE analysis, the 25 subsets were then recombined prior to local FDR calculation which is performed using through semi-parametric mixed modelling and kernel density estimation described in Chapter 1.

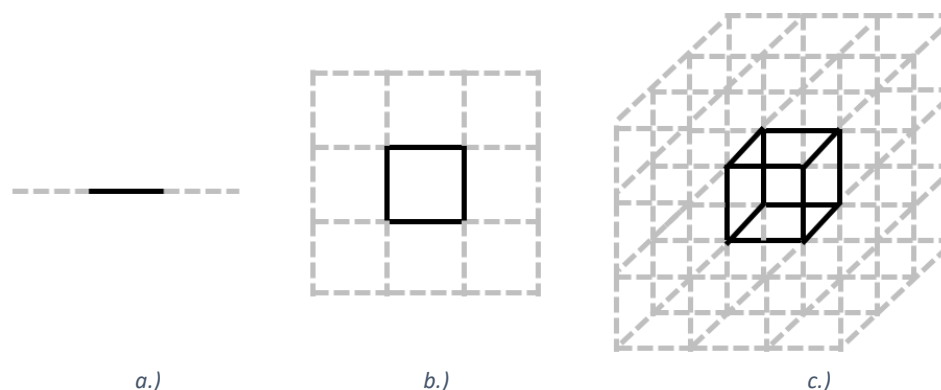
**Welsh two sample t-test** – Comparison of means was performed using R and the `t.test()` function. Multiple testing was performed using `p.adjust()`, `method = 'BH'` i.e. Benjamini-Hochberg.

### ***Sampling method***

Performing Bayesian inference for DE analysis, requires parameter estimation through sampling from the posterior distribution. This process of computing expectations with respect to a target probability distribution is often repeated thousands of times and requires a computer algorithm to process the calculations. Stan is a probabilistic programming language for high-performance statistical computation providing Bayesian inference through Markov Chain Monte Carlo (MCMC) sampling. As described in Chapter 1, a Markov chain is a process for generating sequences of random variables where the probability of the next variable is dependent only on the current variable. Monte Carlo simulation is the generation of series of random numbers from



distributions, similar to the concept of rolling a dice. Stan uses a Hamilton Monte Carlo (HMC) sampler for deciding if the proposed value is likely to fit the posterior distribution. Devised originally by Duane et al. (1987) for Lattice Quantum Chromodynamics calculations, and later adapted for use in Bayesian neural networks (Neal, 1996). This sampling method provides a more efficient and robust algorithm than the Metropolis Hastings (MH) (Hastings, 1970) method (Betancourt, 2016) employed by QPROT. Although easily implemented, the random walk in MH scales poorly with increasingly high-dimensional target distributions (Betancourt, 2017) and has a resulting transition density which is concentrated around the original starting point. As dimensional space increases, the volume of a distinguished point in comparison to the neighbouring volume decreases (Figure 4.1.6).



*Figure 4.1.6; Demonstration of how volume compared to the neighbouring volume decreases as dimension increases. a.) In one dimensional space the relative weight of the centre partition is  $1/3$ , b.) two dimensionally it is  $1/9$ , however, in three dimensions c.) the relative weight is just  $1/27$  (Betancourt, 2017).*

An HMC is better able to explore regions of high probability by making large jumps from the initial starting point exploring high-dimensional spaces with the use of differential geometry. HMC is a variant of MH but samples from a Gibbs canonical distribution, which is distribution that often describes positions and velocities of particles in gas but can also be applied to the model parameters in Bayesian inference:

$$p(X) \propto \exp\left(-\frac{U(X)}{T}\right)$$

Where

Probability,  $p(X)$ , of a system to be in the state  $X$  depends on the energy of the state,  $U(X)$ , and the temperature,  $T$ .

The Hamiltonian differential equation allow us to vectorise the random walk to travel around the contours into a two dimensional space. An extension of an HMC is the NUTs employed by Stan, which eliminates the need for user-defined tuning of the parameters providing an automatic adaptation.

### ***Defining significant results***

Significant results were defined with a range of cut-off values. The results in Chapter Three demonstrated that decreasing or increasing the stringency of the significance threshold for significance may have provided more significant terms. Therefore the range of threshold was increased (0.000001 – 0.1). Details of thresholds used and the increase over iteration are shown in Table 31. Proteins passing the threshold of significance are deemed as being DE and put forward for functional enrichment analysis.

*Table 31; Summary of significance threshold ranges. Proteins passing the threshold are categorised as DE and go forward for enrichment analysis. Size of increment increases over iteration.*

Range	Increment
0.000001 - 0.000009	0.000001
0.00001 - 0.00009	0.00001
0.0001 - 0.0009	0.0001
0.001 - 0.009	0.001
0.01 - 0.1	0.01

### ***Pathway analysis***

To overcome the issues described using RDAVID, described in the results of Chapter Three, the pipeline was altered to use ClusterProfiler 4.0 (Wu et al., 2021) for pathway analysis. Released in August 2021, this Bioconductor package provides an up-to-date universal interface for functional enrichment

## ***Optimising the statistical pipeline for quantitative proteomics***

analysis based on internally supported ontologies and pathways. The clusterProfiler package (Yu et al., 2012) is popular and has been used in pipelines, online platforms and more than 30 CRAN and Bioconductor packages. It relies on Bioconductor genome-wide annotation packages of organism databases (OrgDb) which are updated every six months and accessed through the annotation package AnnotationHub (Morgan, 2021).

Further to the completeness in results arising from the regularly updated search databases, clusterProfiler provides an algorithm for limiting excessive amounts of redundancy, where many significant terms are returned that all reflect the same pathway and interfering with interpretation. The `simplify()` function (Yu, 2020), which is designed to limit reoccurrence of highly similar terms by summarising them with one representative term. Making use of the GOsemSim package (Yu et al., 2010) within Bioconductor, and general terms from analysis of the whole GO corpus are limited. GO is a hierarchical ontology of a set of concepts and their relationships, with parent terms based on less specific concepts, and children terms of increasingly specific concepts. Multiple parents for each concept are allowed, meaning that through multiple paths, two terms can share the same ancestors. GOsemSim calculates semantic similarity among enriched terms using four information content based methods (Resnik, 1999, Jiang and Conrath, 1997, Lin, 1998, Schlicker et al., 2006), which are calculated using the frequency two terms and their closest common ancestor appears in a specific corpus of GO annotations, and a graph structure based method (Wang et al., 2007), which is computes the specificity of a GO term based on its location in a graph. The resulting terms with a high level of similarity are removed, leaving a single representative term. In the development of our pipeline, the `simplify()` function was included with the intention of reducing bias from redundant, inflated counts and therefore providing an unbiased evaluation metric through pathway analysis.

Pathway analysis was performed in R using the `enrichGO()` function of the clusterProfiler package. Proteins passing the significance threshold were categorised as DE and used as the target set. All discovered proteins were used as the background set. Biological process, cellular component and molecular

function sub-ontologies were searched and significant terms were defined by BH corrected  $p < 0.05$ . To reduce redundancy in the enriched terms, the function `simplify()` was applied using the recommended similarity cut-off of 0.7 and BH adjusted  $p$ -value of 0.05 and the Human (Carlson, 2019) database, a Bioconductor annotation packages of genome wide annotation, primarily based on mapping using Entrez Gene identifiers was searched. Parallelisation was again employed to reduce computational time.

### ***Evaluation of Results***

The combination of parameters; normalisation method, DE analysis method, and significance threshold that provided the highest number of significantly enriched terms was designated the optimised pipeline and returned to the user, along with the DE proteins and details of the enriched terms.

### **iii. Aims of chapter**

The results in Chapter Three highlighted the diversity in application of statistical methods, and with all of the areas of possible heterogeneity in proteomics data, a thorough assessment of the properties of the data would be useful prior to selecting the appropriate statistical method rather than trying to rely on one gold-standard approach. However, the mathematical knowledge required for this may be beyond some biologist's experience level. In this chapter we develop an optimised proteomics pipeline which utilises HPC and incorporates several different approaches for DE analysis, normalisation and selection of significance threshold for significance, applying pathway analysis to provide an evaluation metric, and returning the optimal parameter combination to the user. The implemented algorithms will be validated and the best results of analysis will be compared to a functional enrichment analysis obtained from a standard output from Progenesis QIP to demonstrate the improved performance.

## **4.2. Methods**

### **i. Validation of methods**

#### ***Datasets***

Biological datasets PXD004501 (Jin et al., 2018), PXD004682 (Stewart et al., 2017), and PXD007592 (Zila et al., 2018) previously described in Chapter Three were used to demonstrate the validity of the methods implemented in the pipeline.

#### ***Implementation Bayes statistics***

DE analysis of log<sub>2</sub> transformed abundances of each of the datasets was performed using QPROT without any normalisation. Log<sub>2</sub> transformed abundances (QPROT performs a log transformation as part of its analysis) were analysed using the BayesianT. The output Z-statistics of for each analysis were assessed for FDR, both using the BayesianT calculation and pathway analysis was conducted using the significance thresholds described in Table 31 and the number of significant terms for each method of DE analysis was compared.

#### ***Implementation of FDR***

DE analysis of log<sub>2</sub> transformed protein abundances for each of the datasets was performed using the QPROT. The FDR calculation was performed on the resulting Z-statistics using QPROT and the BayesianT and their values compared.

#### ***Evaluation of clusterProfiler() for DE evaluation***

In order to validate the use of pathway analysis for benchmarking, significant results from the *t*-test (*p*-value threshold < 0.05) analysis of dataset PXD004682 were evaluated with pathway analysis using different proportions of proteins said to be significantly changing. On each iteration, 5% of proteins classified as DE were removed from the pathway analysis foreground list and replaced with

proteins from the background list, simulating analysis with increasing false-positive rates. This process was repeated 100 times. The validation exercise above was also repeated with the application of the `simplify()` function to reduce redundancy.

## **ii. Comparison to Progenesis output**

To validate the pipeline, the datasets described in Chapter Three were analysed with the pipeline. The optimised result was compared to clusterProfiler GO enrichment analysis, as described in section 4.2, of the standard output from Progenesis analysis: defined as Progenesis normalisation, ANOVA (in the form of Welch's *t*-test for pairwise analysis), and using common significance thresholds; 0.1, 0.05, and 0.01.

## **iii. Comparison to QPROT output**

Further validation was performed by comparing the pipeline output to that of QPROT. The optimised result was compared to clusterProfiler GO enrichment analysis, as described in section 4.2 of the QPROT analysis from Chapter 3. In this analysis QPROT was set to perform 10,000 iterations for the burnin and 10,000 iterations for sampling, and an inverse log<sub>2</sub> transformation was performed prior to QPROT analysis due to the transformation that occurs within the software. Pipeline settings were:

1. 325 iterations for the burnin and 650 iterations for sampling
2. 1000 iterations for the burnin and 2000 iterations for sampling (rStan default parameters)

## **4.3. Results and discussion**

### **i. Validation of methods**

#### ***Implementation Bayesian differential expression analysis***

Comparison of the pathway analysis evaluation of the results of QPROT and BayesianT's DE analysis of the same data and using the same method for FDR calculation (Figure 4.3.1) demonstrates better overall performance by BayesianT and sufficiently demonstrates the validity of the method. However, for two of the datasets, QPROT gave a slightly higher maximum number of terms at its optimal significance threshold than the BayesianT.

For dataset PXD004501, both DE methods resulted in an overall low level of enrichment terms with a maximum of 13 and eight for QPROT and the BayesianT, at their optimal significance threshold of 0.02 and 0.006, respectively. However, overall, BayesianT analysis more consistently gave a higher number of enrichment terms at the other threshold cut-off values. In analysis of PXD004682 data, the BayesianT DE provided the most enrichment terms for all significance thresholds with a maximum of 52 terms at a threshold of 0.005 as opposed to seven terms using QPROT at the same threshold. QPROT's optimal significance threshold was 0.02 which gave 18 enrichment terms, which was lower than the amount resulting from BayesianT analysis (35). For dataset PXD007592, the BayesianT yielded a higher number of enrichment terms than QPROT at all but two thresholds (0.001 and 0.1). At 0.1, which was QPROT's optimal significance threshold, there were 109 terms as opposed to 90 terms and at 0.001 there were 54 as opposed to 52 terms. Optima BayesianT analysis output was similar result QPROT's maximum at its optimal significance threshold of 0.05 with 103 enrichment terms.

## Optimising the statistical pipeline for quantitative proteomics

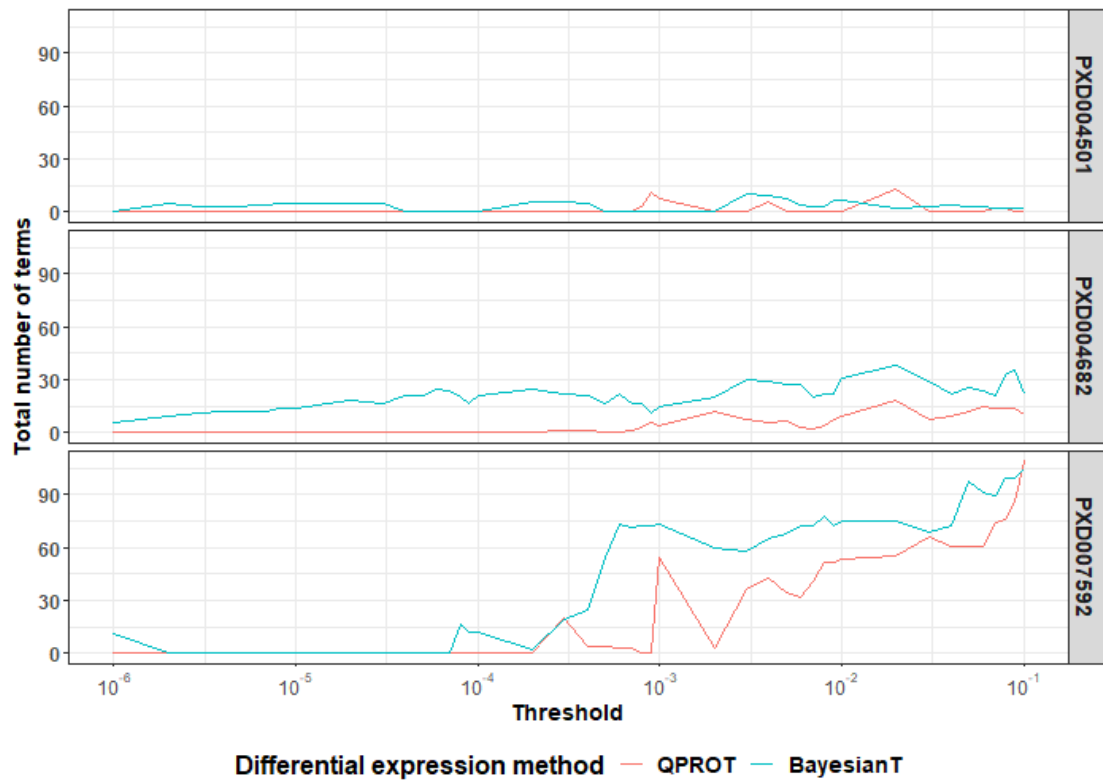


Figure 4.3.1; Comparison of pathway analysis of differential expression analysis of biological benchmarking datasets produced by QPROT and BayesianT model using the same FDR method.

Both methods follow the same formula for calculating DE; however, differences occur in QPROT which provides an additional imputation step that uses probabilistic modelling and truncation rules (as described in Chapter One). As imputation is applied not only to missing data but also to abundances falling below a truncation point calculated individually for each protein, and the process is carried out within the software without user control (imputed values are not included in the software's output). Furthermore, a different number of sampling and burnin iterations were performed, making it difficult to create identical results using the two methods. The impact of these parameters is further investigated in section 4.3.ii.



### Implementation of FDR calculation

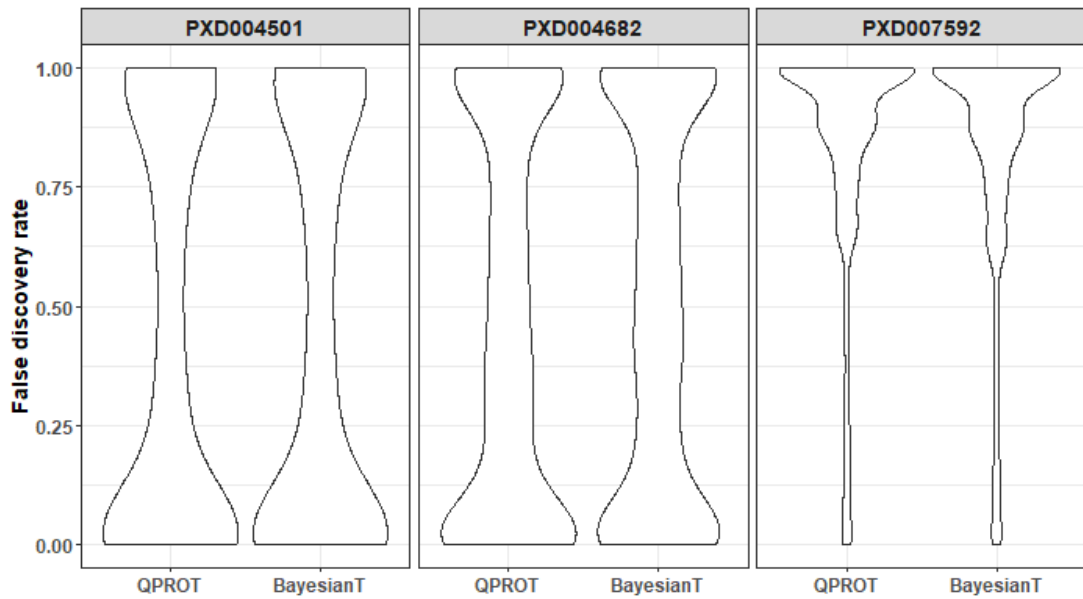


Figure 4.3.2; Comparison of FDR values produced by QPROT and BayesianT model on the same differential expression analysis output for biological benchmarking datasets.

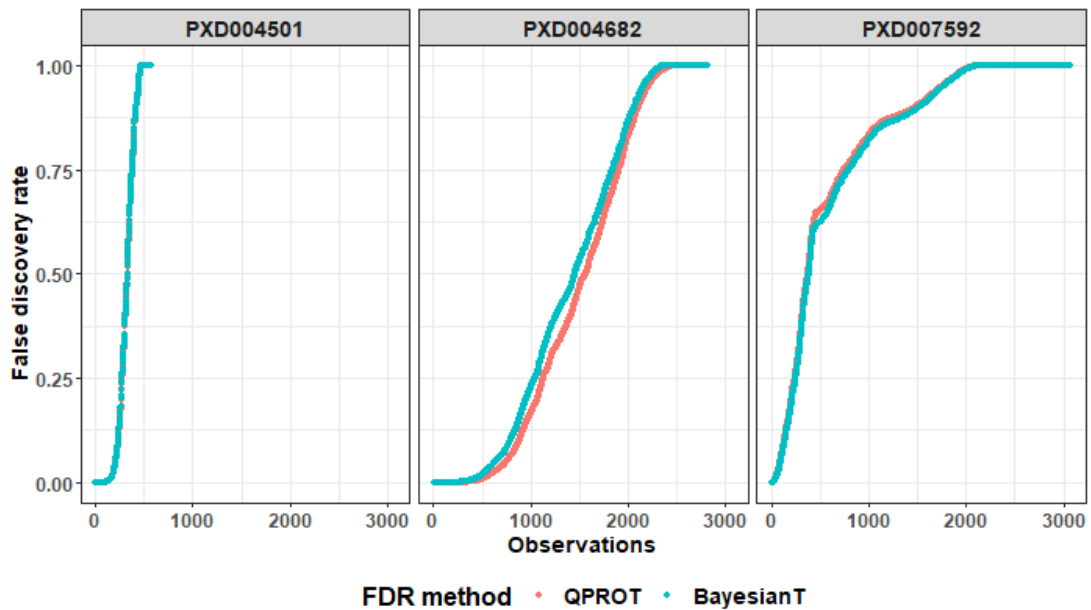


Figure 4.3.3; Comparison of ordered FDR values produced by QPROT and BayesianT model on the same differential expression analysis output for biological benchmarking datasets.

Violin plots comparing the FDR calculation of performed by QPROT and the FDR calculation performed by the BayesianT on the same DE analysis data (Figure 4.3.2) were very similar for all datasets, suggesting that both methods provide almost identical results. Ordered FDR values (Figure 4.3.3) for the same comparison show that values are the same for dataset PXD004501, almost the same for dataset PXD007592, and only slightly differ in the middle 1000

observations for dataset PXD004682. These results sufficiently validate the method for FDR calculation in the BayesianT.

The successful implementation of the key stages of QPROT's algorithm for pairwise label-free DE analysis allowed the shell-based program for Linux operating systems (that requires installation of a specialist mathematical library) to be incorporated into the R based pipeline.

### ***Pathway analysis***

The clusterProfiler R package has the ability to provide KEGG module queries. However, if this is conducted through connection to an online web resource, much like the RDAVID API used in the benchmarking software in Chapter Three. Due to the unpredictable crashes experienced when making multiple calls to the web-based database, plus the security constraints of web access on the server nodes, KEGG annotation was not included in the optimised pipeline. For Reactome pathway queries, there is a corresponding package to clusterProfiler, ReactomePA (Yu and He, 2016). However, this package does not have the same level of features as clusterProfiler; currently, there are only a limited amount of organisms that can be searched, and the input IDs must be converted to Entrez gene ID form. Furthermore, there is no `simplify()` function as with clusterProfiler, so the resulting enriched terms will be highly inflated due to redundancies. For this reason, Reactome pathway analysis was also excluded from the pipeline.

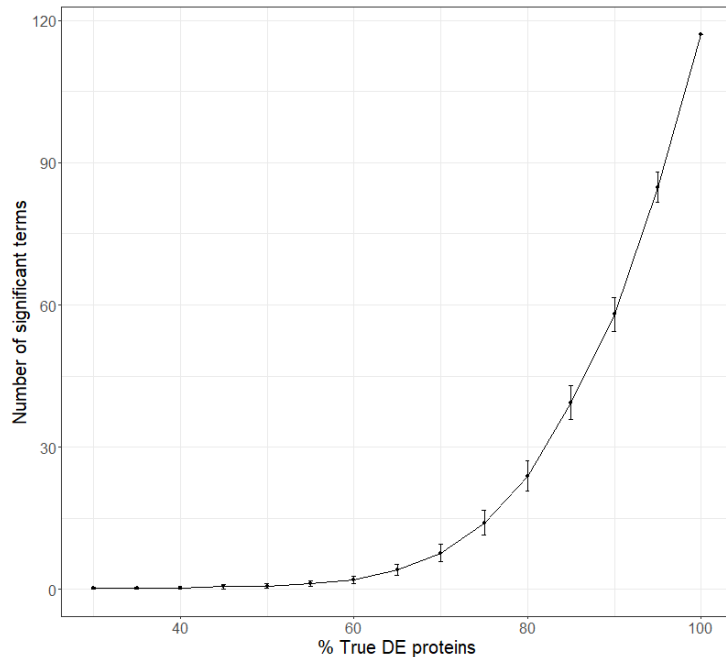
To validate the pathway analysis method, enrichment analysis was performed on foreground data that included an increasing amount of proteins from the background group (i.e., not classified as DE by *t*-test analysis using a BH *p*-value of < 0.05 as significant) to simulate analyses that incorrectly classify proteins as DE when they are not. The results (Table 32 and Figure 4.3.4) demonstrate that by including an increasingly large proportion of not 'true' DE proteins in the search list, fewer significant terms are returned on functional enrichment analysis. Proteins that had been classified by statistical analysis as changing

## *Optimising the statistical pipeline for quantitative proteomics*

between conditions were more functionally similar to each other compared to the total pool of identified proteins.

*Table 32; Pathway analysis validation completed with dataset PXD004682. Different proportions of true DE proteins and randomly selected background proteins. Number of significant terms from DAVID enrichment analysis of DE proteins from t-test analysis (Benjamini Hochberg adjusted p-value < 0.05). Process was repeated 100 times.*

<b>Proportion DE proteins</b>	<b>Proportion of background proteins</b>	<b>Mean number of terms</b>	<b>Standard deviation</b>	<b>Standard error</b>
100	0	117.00	0	0
95	5	84.85	16.38	3.21
90	10	58.01	17.99	3.53
85	15	39.38	18.30	3.59
80	20	23.89	16.35	3.21
75	25	14.09	13.71	2.69
70	30	7.64	9.43	1.85
65	35	4.19	5.96	1.17
60	40	2.06	3.85	0.75
55	45	1.22	2.67	0.52
50	50	0.73	2.19	0.43
45	55	0.57	2.03	0.40
40	60	0.35	1.28	0.25
35	65	0.24	0.89	0.17
30	70	0.32	1.10	0.22



*Figure 4.3.4; Pathway analysis validation completed with dataset PXD004682. Number of significant terms (t-test analysis, Benjamini Hochberg adjusted p-value < 0.05) from DAVID enrichment analysis with different proportions of significantly differentially expressed proteins included in the search list. Process was repeated 100 times.*

These results suggest that the pathway enrichment testing approach using clusterProfiler is an effective evaluation tool, and that precise DE analysis produces more significant terms, and that an increased false discovery rate ‘dilutes’ the quality of the enrichment analysis.

### ***Simplify function***

The purpose of the simplify() function is to reduce high-level terms in favour of leaf node terms, condensing the total number of terms and giving a better biological picture. This step was included in the pipeline to avoid bias from expanded counts from an inflated number of enrichment results. Table 33 summarises the reduction in significantly enriched terms from clusterProfiler analysis before and after the application of the simplify() function, which reduces the total number of significant GO terms by approximately half (top row compared to bottom row).

Table 33; Total number of enriched GO BP, CC, and MF terms from clusterProfiler analysis of upregulated proteins from the current Progenesis QIP DE analysis of dataset PXD004682 before and after simplify() function

Analysis	Number of significantly enriched terms		
	GO BP	GO CC	GO MF
clusterProfiler only	159	67	14
clusterProfiler followed by simplify()	77	26	8

Table 34 demonstrates the effective reduction of redundant MF terms from the analysis in Table 33 after the simplify() function was applied. Three terms, ‘actin binding’, ‘actin filament binding’, and ‘cytoskeletal protein binding’ are reduced to simply ‘actin binding’; ‘calmodulin binding’, ‘myosin binding’, and ‘spectrin binding’ are represented by the term ‘calmodulin binding’; ‘endopeptidase inhibitor activity’ also covers the terms ‘endopeptidase regulator activity’ and ‘enzyme inhibitor activity’; and a further reduction is made with ‘peptidase inhibitor activity’ also representing ‘peptidase regulator activity’.

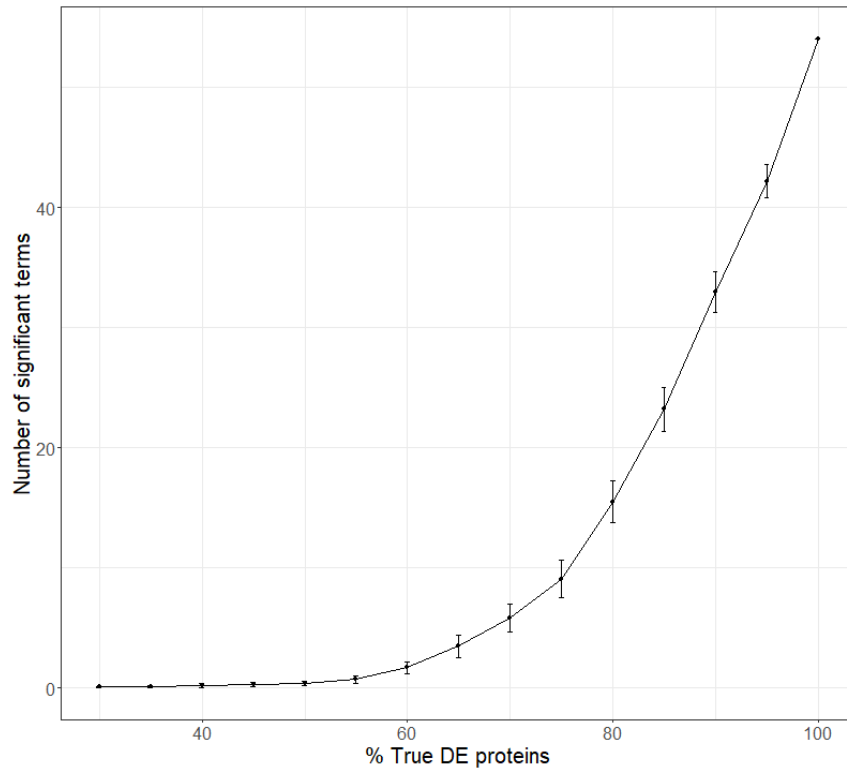
## ***Optimising the statistical pipeline for quantitative proteomics***

*Table 34; Simplified enriched GO MF terms (left column) and original enriched GO MF terms (right column) from clusterProfiler analysis of upregulated proteins from the current Progenesis QIP DE analysis of dataset PXD004682*

<b>GO MF terms after simplify()</b>	<b>Original GO MF terms</b>
Actin binding	Actin binding Actin filament binding
Antioxidant activity	Cytoskeletal protein binding Antioxidant activity
Calcium ion binding	Calcium ion binding
Calmodulin binding	Calmodulin binding Myosin binding Spectrin binding
Copper ion binding	Copper ion binding
Endopeptidase inhibitor activity	Endopeptidase inhibitor activity Endopeptidase regulator activity Enzyme inhibitor activity
Peptidase inhibitor activity	Peptidase inhibitor activity Peptidase regulator activity

*Table 35; Pathway analysis validation using the simplify() function completed with dataset PXD004682. Different proportions of true DE proteins and randomly selected background proteins. Number of significant terms from DAVID enrichment analysis of DE proteins from t-test analysis, (Benjamini Hochberg adjusted p-value < 0.05). Process was repeated 100 times.*

<b>Proportion DE proteins</b>	<b>Proportion of background proteins</b>	<b>Mean number of terms</b>	<b>Standard deviation</b>	<b>Standard error</b>
100	0	54.00	0	0
95	5	42.18	6.96	1.36
90	10	32.96	8.51	1.67
85	15	23.18	9.38	1.84
80	20	15.49	8.68	1.70
75	25	9.05	7.91	1.55
70	30	5.83	5.89	1.16
65	35	3.46	4.71	0.92
60	40	1.69	2.47	0.48
55	45	0.70	1.58	0.31
50	50	0.32	0.91	0.18
45	55	0.26	0.88	0.17
40	60	0.20	0.89	0.17
35	65	0.08	0.44	0.09
30	70	0.05	0.26	0.05



*Figure 4.3.5; Pathway analysis validation completed with dataset PXD004682. Number of significant terms (t-test analysis, Benjamini Hochberg adjusted p-value < 0.05) from DAVID enrichment analysis with different proportions of significantly differentially expressed proteins included in the search list*

To validate the use of `simplify()` in as part of the pathway analysis evaluation, the validation exercise above for `clusterProfiler` was repeated using the additional application of the `simplify()` function. Although the results (Table 35 and Figure 4.3.5) show a decrease in total number of terms (54 rather than 117) when there are 100% of 'true' DE proteins in the search list, they do again demonstrate that fewer functionally enriched terms are returned as the number of proteins not classified as changing are included.

i. Comparison to Progenesis output

PXD004501

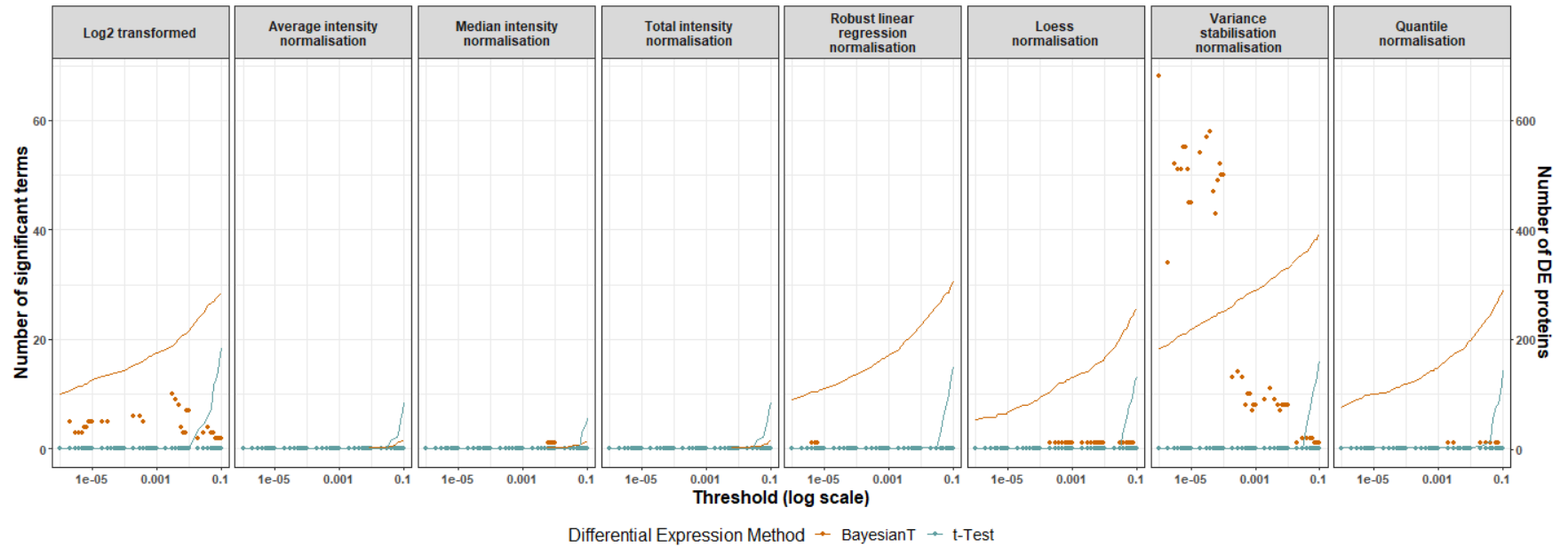


Figure 4.3.6; Enrichment analysis to results from optimised pipeline of dataset PXD004501. DE analysis by BayesianT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.

A summary of the functional enrichment results from the analysis of dataset PXD004501 is shown in Figure 4.3.6. Most parameter combinations struggled to produce significantly enriched terms with this data. However, the BayesianT method for DE combined with the transformation based normalisations, VSN (column 8, orange data), and to a lesser degree log2, (column 1) managed to provide functionally related DE proteins, with the best output occurring at a relatively low significance threshold (maximum 68 significant terms at threshold of 0.000001 from 184 DE proteins). A comparison of the output from a standard Progenesis QIP output is (Progenesis normalisation and Welsh's *t*-test using common significance thresholds of 0.01, 0.05, and 0.1 for defining DEs) is summarised in Table 36 and produced 5, 144, and 270 DE proteins, respectively. However, GO Enrichment analysis of the assigned DE proteins using clusterProfiler did not return any significantly enriched terms.

*Table 36; Enrichment analysis results for dataset PXD004501 from the optimised pipeline and Progenesis output at standard significance threshold of 0.01, 0.05, and 0.1.*

<b>DE method</b>	<b>Normalisation</b>	<b>Threshold</b>	<b>Number of DE proteins</b>	<b>Number of significant terms</b>
BayesianT	VSN	0.000001	184	68
<i>t</i> -Test	Progenesis	0.01	1	0
<i>t</i> -Test	Progenesis	0.05	46	0
<i>t</i> -Test	Progenesis	0.1	142	0

The paper accompanying the dataset PXD004501 (Jin et al., 2018) aimed to determine the molecular characterizations of malignant ascites, a sign of peritoneal seeding in gastric cancer. This dataset identified 2534 protein groups when comparing ascite fluid from patients with liver cirrhosis and malignant ascites, with 397 changing between conditions (defined as 1.5-fold change and  $p$ -value < 0.05); 81 elevated in gastric cancer patients and 218 downregulated. However, this value was obtained using two search engines and the method used to integrate the results was not reported. Although a 1%-peptide-level filter was applied, there was no protein level FDR control and results with a single peptide ID in only one replicate were included, resulting in a vastly over-reported count of proteins which had not been confidently identified and quantified.



## *Optimising the statistical pipeline for quantitative proteomics*

Filtering their supplementary data for proteins supported by two or more peptides, and protein quantification values in all six samples results in only 574 protein groups. In the optimal analysis by our pipeline, 575 proteins were identified with 184 proteins being classed as DE; 157 were upregulated in cancer patients and 27 were downregulated. Of the 157 upregulated proteins, 31 were also highlighted as being upregulated in the combined results in the original paper along with 18 of the 27 downregulated proteins. To directly compare our results from the optimal analysis with the pairwise analysis of PXD004501 data, GO analysis was performed using clusterProfiler on the DE proteins provided by Jin et al. (2018) in their ‘Supplementary Table S6. List of the significantly differential expressed proteins in label-free quantitation set 1’ (<https://ars.els-cdn.com/content/image/1-s2.0-S0009912017312560-mmc7.xlsx>). Enriched terms for Biological Process, Molecular Functions, Cellular Components, and KEGG pathways for analysis performed on upregulated proteins as classed by each of the methods and are summarised in Table 37, with upregulated proteins from our optimal analysis providing more functionally enriched terms than analysis of the upregulated proteins from the original paper. Of the 46 proteins classed as DE by the current Progenesis QIP output, only 6 were upregulated in cancer patients and enrichment analysis of these produced no significant terms.

*Table 37; Summary of the number of GO terms and KEGG pathways from enrichment analysis performed on upregulated proteins from the optimal analysis and from the original paper for dataset PXD004501*

Source of upregulated proteins	Number of upregulated proteins	Number of significantly enriched terms			
		GO BP	GO CC	GO MF	KEGG pathways
Optimal analysis	157	131	22	49	16
Original paper	81	44	9	14	4

The top eight most significant terms from enrichment analysis of the upregulated proteins from the optimal analysis and the original paper were compared (Figure 4.3.7) and found to contain mostly similar terms, validating the pipelines’ analysis of this dataset. Furthermore, enrichment results for the proteins from the optimal analysis provide more significant terms with a greater values of significance and a higher gene count than those identified in the original paper.

# Optimising the statistical pipeline for quantitative proteomics

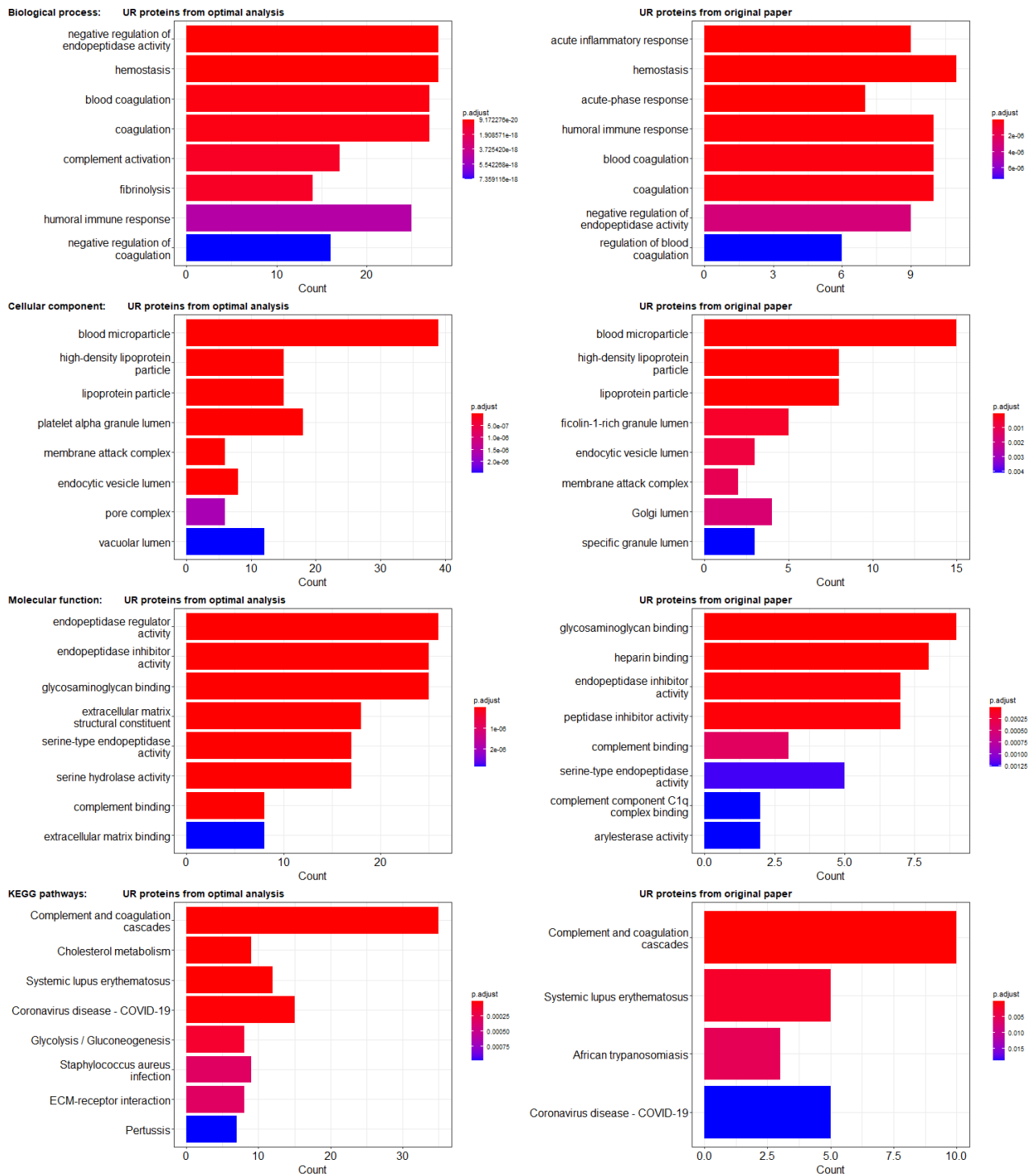


Figure 4.3.7; Top eight most significant terms for GO Biological Pathways, Molecular Functions, Cellular Components and KEGG pathways from enrichment analysis of proteins upregulated in malignant ascites as classified by the optimal result from our analysis (left column) and by the original paper for dataset PXD004501 (right column), count represents the number of proteins in the search group that belong to the given gene-set.

## Optimising the statistical pipeline for quantitative proteomics

Table 38; Summary of terms produced by enrichment analysis of proteins upregulated in cancer tissue as identified by the optimal analysis and from the original paper for dataset PXD004501 that are related to functions highlighted by Jin et al. (2018)

	Analysis of UR proteins (BH p-value)	
	optimal	original
<b><i>Movement of cell, locomotion, cell adhesion</i></b>		
Cadherin binding involved in cell-cell adhesion	4.64E-02	NS
Endothelial cell migration	1.39E-03	4.87E-02
Integrin binding	2.65E-03	NS
Integrin biosynthetic process	3.21E-02	NS
Leukocyte chemotaxis	5.10E-03	NS
Leukocyte migration involved in inflammatory response	4.44E-03	NS
Muscle cell migration	5.27E-03	NS
Myeloid leukocyte migration	5.15E-03	NS
Regulation of actin cytoskeleton	1.50E-02	NS
Regulation of endothelial cell migration	4.97E-03	NS
Regulation of smooth muscle cell migration	1.02E-02	NS
<b><i>Immune response, response to wounding</i></b>		
Acute inflammatory response	3.34E-15	2.79E-08
Blood coagulation	2.87E-19	2.03E-07
Chronic inflammatory response	8.53E-03	NS
Coagulation	3.46E-19	2.03E-07
Complement and coagulation cascades	5.71E-44	6.35E-12
Humoral immune response	4.65E-18	1.18E-07
Immunoglobulin binding	6.81E-03	NS
Innate immune response in mucosa	NS	1.83E-02
Leukocyte aggregation	1.81E-03	NS
Mucosal immune response	NS	3.70E-02
Negative regulation of angiogenesis	6.33E-04	NS
Negative regulation of coagulation	7.36E-18	NS
Negative regulation of immune effector process	1.19E-02	NS
Negative regulation of wound healing	2.70E-17	NS
Positive regulation of inflammatory response	1.96E-02	NS
Regulation of acute inflammatory response	1.54E-02	NS
Regulation of blood coagulation	NS	7.37E-06
Regulation of coagulation	NS	7.77E-06
Regulation of humoral immune response	1.54E-02	NS
Regulation of wound healing	NS	7.37E-06
<b><i>Calcium ion binding and peptidase inhibitor activity</i></b>		
Calcium-dependent protein binding	4.20E-03	NS
Endopeptidase inhibitor activity	5.02E-21	1.88E-05
Endopeptidase regulator activity	4.32E-21	NS
Exopeptidase activity	1.37E-03	NS
Negative regulation of endopeptidase activity	9.17E-20	3.14E-06
Peptidase inhibitor activity	NS	1.88E-05

## *Optimising the statistical pipeline for quantitative proteomics*

Peptidase inhibitor complex	2.20E-03	1.88E-05
Serine-type endopeptidase activity	3.30E-12	1.22E-03
<b><i>Extracellular components</i></b>		
Endocytic vesicle lumen	4.06E-09	7.39E-04
Extracellular matrix binding	2.82E-06	NS
Extracellular matrix structural constituent	1.37E-12	NS
Neutrophil extracellular trap formation	3.12E-02	NS
Pigment granule	1.18E-04	NS
Regulation of extracellular matrix constituent secretion	2.70E-02	NS
<b><i>Regulation of kinase activity</i></b>		
Positive regulation of tau-protein kinase activity	2.20E-02	NS

In the original paper for PXD004501, Jin et al. (2018) discuss relevant enrichment terms that were commonly acquired by functional analysis of the results a combination of three procedures which also included in-depth profiling and a second pairwise DE analysis (supplied as separate datasets PXD002213 and PXD03351) comparing the ascite fluids of patients with liver cirrhosis to those of patients with peritoneal seeding. Table 38 summarises terms found in enrichment analysis of the proteins upregulated in cancer tissue as classed by the optimal analysis and those from the paper. There were many more significant (40 compared to 17) related terms obtained from the optimal analysis proteins than from the proteins from the supplementary materials in the original paper with a greater value of significance (smallest p-value 4.32E-21).

In the original paper for the dataset PXD004501, Jin et al. (2018) selected protein marker candidates to differentiate malignant ascites from benign ascites by using the Human Protein Atlas database (Uhlén et al., 2015) to compare the mRNA expression of upregulated proteins in the stomach relative to mean levels in 27 other tissues. Those expressed specifically in the stomach were compared with gastric cancer secretome datasets (Marimuthu et al., 2013) to see if they were secreted from gastric cancer cells or tested for positive staining in >75% of gastric tissues in the Human Protein Atlas database. Following this two proteins were verified using enzyme-linked immunosorbent assay (ELISA) using 27 benign samples and 57 malignant ascitic fluids samples. Only one of these two proteins, POSTN (Q15063), was identified by Progenesis QIP processing of the raw data. POSTN was successfully identified as being upregulated in cancer samples using the optimal analysis, but was not classed as DE using the

## Optimising the statistical pipeline for quantitative proteomics

current Progenesis QIP pipeline, differences in the results of the statistical analyses are summarised in Table 39. Normalisation of the raw abundances with Progenesis QIP resulted in abundances with a log<sub>2</sub> fold change of 1.55 whereas VSN normalised abundances gave a log<sub>2</sub> fold change of 2.22. Welsh's t-test analysis of Progenesis QIP normalised abundances gave a *p*-value of 0.05 and a BH corrected value of 0.14 meaning the protein wasn't classed as being DE using the current Progenesis QIP output.

Table 39; Statistical analysis results of protein identified by Jin et al. (2018) as being a protein marker for differentiating malignant from benign ascites that were identified in Progenesis QIP processing

UniProt ID	Progenesis analysis			Optimal analysis		
	Log <sub>2</sub> fold change	<i>p</i> -value	BH <i>p</i> -value	Log <sub>2</sub> fold change	Z-statistic	Bayesian T FDR
Q15063	1.55	0.050	0.14	2.22	5.82	6.39 e <sup>-28</sup>

When compared to the proteins classed as upregulated in cancer in the original paper, our optimal analysis produced more terms with a higher fold enrichment. In the current Progenesis QIP output, 218 proteins had were more abundant in cancer patient and 47 of them had a *p*-value below 0.05. However, after BH correction for multiple testing, only 6 proteins were classed as being significantly upregulated. Of the proteins more abundant in cancer patients with a *p*-value < 0.05 but a BH corrected *p*-value > 0.05, there were 16 with a greater than 2-fold change, suggesting that the best method for correcting for multiple testing may not be the most appropriate for this dataset. The original paper for PXD004501 identified characteristic functions in the ascites proteome data: movement of cell, locomotion, cell adhesion, immune response, response to wounding, calcium ion binding and peptidase inhibitor activity, extracellular components and regulation of kinase activity. Enrichment analysis of the upregulated proteins from the optimal analysis produced more terms related to those functions than enrichment analysis of the proteins identified as upregulated in cancer tissue in the by the original paper. The optimal analysis also successfully classed the protein marker identified by the paper as being upregulated in cancer tissues. However, the current Progenesis QIP output did not. These results provide biological validation for the pipeline proposed in this chapter.

PXD004682

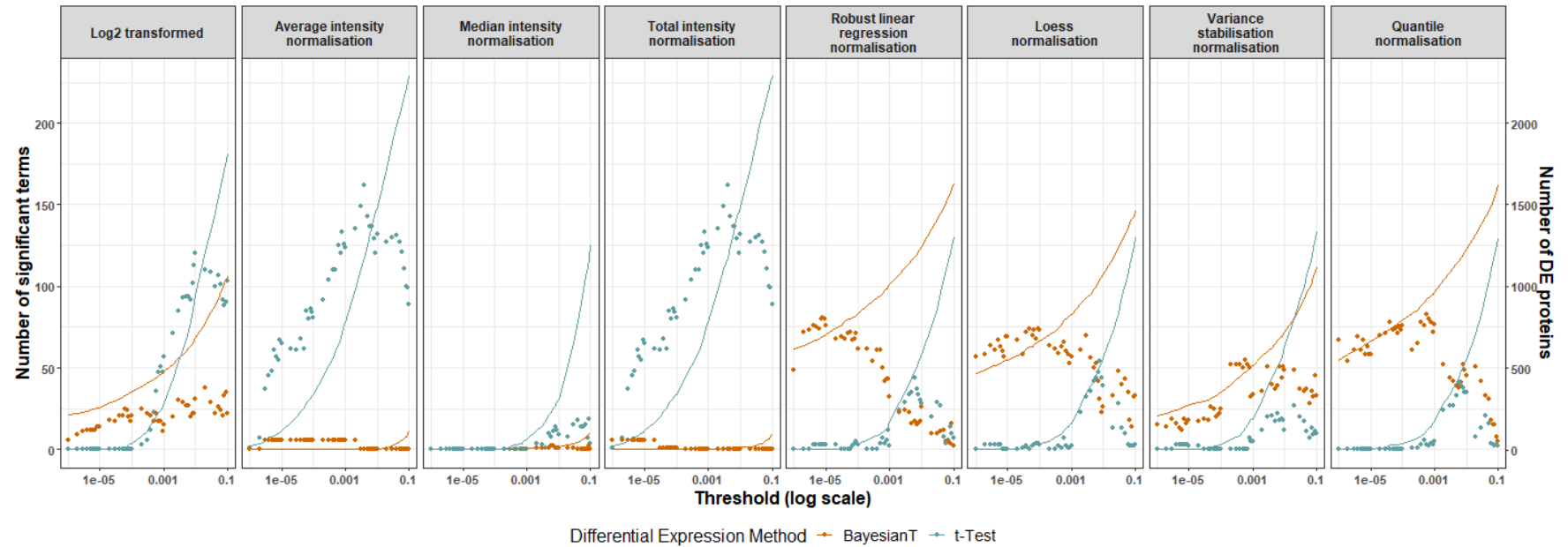


Figure 4.3.8; Enrichment analysis to results from optimised pipeline analysis of dataset PXD004682. DE analysis by BayesianT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.

## Optimising the statistical pipeline for quantitative proteomics

A summary of the functional enrichment results from the analysis of dataset PXD004682 is shown in Figure 4.3.8. Here the *t*-test gave optimal performance combined with central tendency based normalisations AIN (column 2) and TIN (column 4). The best performance appeared to occur around the significance threshold 0.005, which is relatively lenient in terms of the range of thresholds applied by the pipeline, but much more stringent than the generally accepted 0.05. Interestingly, in the best parameter combination this still allowed a large proportion of proteins to be defined as DE; 1176 for *t*-test DE analysis, TIN and significance threshold of 0.004, producing 162 significant terms. The BayesianT method for DE was not as successful as the classic *t*-test and appeared to work best with linear modelling based normalisation methods RLR and Loess (columns 5 and 6) and still required a relatively stringent significance threshold of below 0.0001 for success.

A comparison of the output from a standard Progenesis QIP output is (Progenesis normalisation and Welsh's *t*-test using common significance thresholds of 0.01, 0.05, and 0.1 for defining DEs) is summarised in Table 40 and produced 666, 1226, and 1515 DE proteins, respectively. However, GO Enrichment analysis of the assigned DE proteins using clusterProfiler only returned 40, 54 and 52 significantly enriched terms, respectively. The Progenesis normalisation method employs a form of central tendency normalisation, similar to the successful methods employed in the pipeline, and manages to provide some amount of functionally related proteins. The key to the pipeline's success with this dataset appears to be the more stringent threshold for significance, which will have limited the number of false positives included in the DE group, providing better functional enrichment results.

Table 40; Enrichment analysis results for the dataset PXD004682 from the optimised pipeline and Progenesis output at standard significance threshold of 0.01, 0.05, and 0.1.

DE method	Normalisation	Threshold	Number of DE proteins	Number of significant terms
<i>t</i> -Test	TIN	0.004	1176	162
<i>t</i> -Test	Progenesis	0.01	666	44
<i>t</i> -Test	Progenesis	0.05	1226	59
<i>t</i> -Test	Progenesis	0.1	1513	56

## ***Optimising the statistical pipeline for quantitative proteomics***

The original paper for PXD004682 (Stewart et al., 2017) aimed to compare methods of sample preparation and MS data acquisition using greatest coverage as a metric to see if increased instrument time was justified for increasing protein identification in large scale proteomics studies. Data-dependent acquisition was compared to data-independent acquisition and label-free quantification was compared to TMT. We used the publically available data (PXD004682) for the label-free DDA experiment which was a pairwise comparison between lung squamous cell carcinoma and adjacent tissue. Our optimal analysis pipeline (*t*-test for DE analysis, TIN, and a significance threshold of 0.004) provided 1159 proteins being identified as upregulated in cancer tissue as opposed to 312 proteins with the current Progenesis QIP output.

In the original paper for the dataset PXD004682, Stewart et al. (2017) performed enrichment analysis using PANTHER (Protein ANalysis THrough Evolutionary Relationships) (Mi et al., 2016) GO-Slim functional classifications and assessed the ability of the analysis methods to identify proteins involved in biomarker candidates and selected lung cancer pathways (Stewart et al., 2015). To compare our results with those of the original paper, Gene List Analysis was performed using the PANTHER webpage (Version 17.0, <http://www.pantherdb.org/>) on the upregulated proteins from our optimal analysis, from the current Progenesis output, and the biomarker candidates identified in the original paper (Supporting Information, 'Table S4') to compare the amount of shared terms. PANTHER analysis of the identified biomarker candidates provided 1507 enriched terms. PANTHER analysis of the proteins from our optimal analysis and the current Progenesis QIP output produced 705 and 620 terms, with 705 and 571 of those terms being shared with the PANTHER analysis terms from the biomarker candidate, respectively (Table 41).



## Optimising the statistical pipeline for quantitative proteomics

Table 41; Number of enrichment terms from PANTHER analysis of proteins classed as being upregulated by optimal analysis and the current Progenesis output, and the number of terms shared with PANTHER analysis of biomarker candidates identified by Stewart et al. (2017)

Source of upregulated proteins	Number of enriched terms with panther analysis	Number of enriched terms shared with panther analysis of biomarker candidates
Optimal analysis	705	627
Progenesis analysis	620	571

Furthermore, the resulting PANTHER ontology terms were searched for terms associated with the lung cancer pathways: 'Apoptosis and survival - apoptotic TNF-family pathways', 'Apoptosis and survival - p53-dependent apoptosis', 'Development - Dopamine D2 receptor transactivation of EGFR', 'G-protein signaling - K-RAS regulation pathway', 'Immune response- NFAT in immune response', and 'Stimulation of TGF-beta signaling in lung cancer' as described in the original paper. Compared to the current Progenesis QIP output, the optimal analysis provided a greater number of GO-Slim functional classifications that were associated with the selected lung cancer associated pathways (Table 42).

The results demonstrate that the proteins classed as being upregulated by the optimal analysis had associated functions related to the original experiment the dataset was obtained from, and that more functionally related terms were obtained from the optimal analysis that from the current Progenesis analysis. Furthermore, compared to the current Progenesis QIP output, the optimal analysis provided more enriched functional terms (705 compared to 620) with more terms shared with terms with the PANTHER analysis from the biomarker candidate proteins from the original paper (627 compared to 571).

## Optimising the statistical pipeline for quantitative proteomics

Table 42; Number of GO-Slim functional classifications associated with selected lung cancer pathways identified by Stewart et al. (2017) for proteins upregulated in cancer tissue identified by the pipeline's optimal analysis and by the current Progenesis QIP output

Lung cancer pathways and associated terms	Number of associated terms	
	Progenesis	optimal
<b><i>Common terms for several lung cancer pathways</i></b>		
Apoptotic process	4	13
Cell death	8	13
Cell migration	6	11
Cell proliferation	3	10
Cell signaling	2	3
Programmed cell death	4	6
Protein phosphorylation	2	9
Receptor signaling pathway	17	33
Signal transduction	19	34
Synaptic transmission	2	4
Transcriptional regulation	0	1
<b><i>Apoptosis and survival: Apoptotic TNF-family pathways</i></b>		
Extrinsic apoptotic signaling pathway	3	4
Intrinsic apoptotic signaling pathway	2	6
Proteolysis	4	8
<b><i>Apoptosis and survival: p53-dependent apoptosis</i></b>		
Protein modification process	2	12
Protein stability	0	3
<b><i>Development: Dopamine D2 receptor transactivation of EGFR</i></b>		
Dopamine signaling pathway	0	5
Neurotransmitter secretion	4	0
<b><i>G-protein signaling: K-RAS regulation pathway</i></b>		
Cell growth and/or maintenance	3	6
<b><i>Immune response: NFAT in immune response</i></b>		
Cytokine production	1	12
Gene expression	3	19
Immune response	12	29
<b><i>Stimulation of TGF-beta signaling in lung cancer</i></b>		
Angiogenesis	3	7

PXD007592

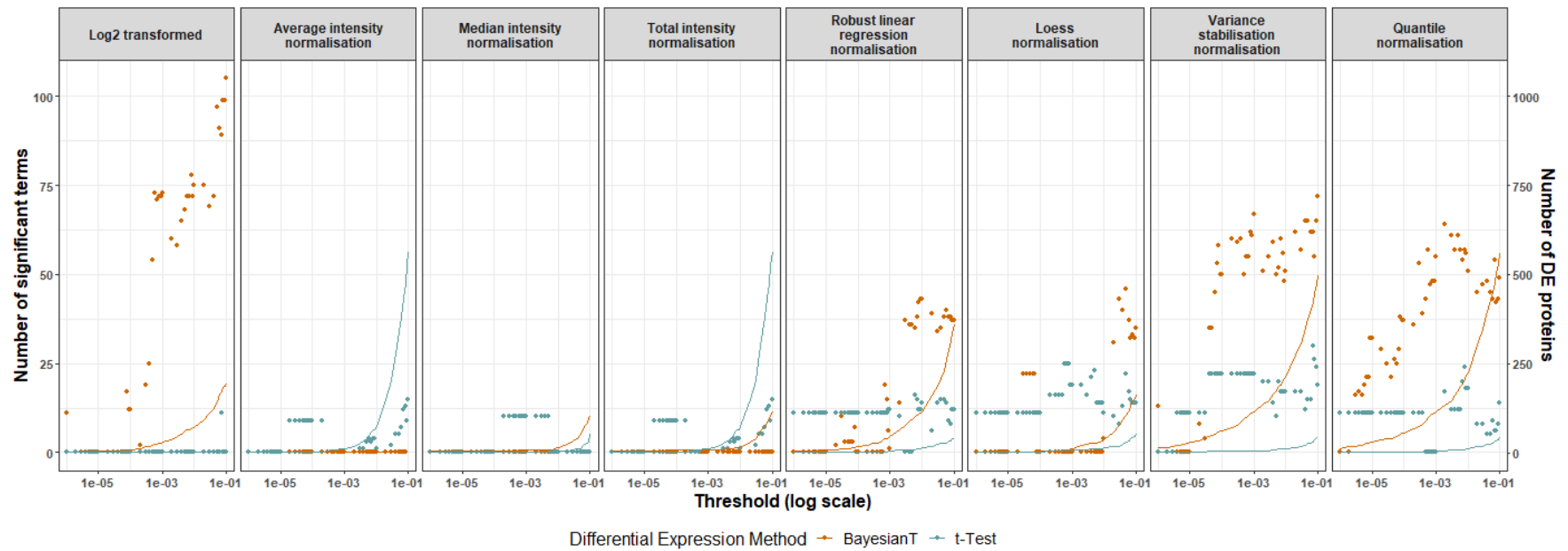


Figure 4.3.9; Enrichment analysis to results from optimised pipeline analysis of dataset PXD007592. DE analysis by BayesianT (orange) and t-test (blue) of protein abundances (identified with by minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points. Significant terms and number of DE proteins were not included for QPROT AIN analysis at significance threshold of 0.1 and 0.09 as all proteins were defined as DE (3079 proteins with zero significant terms) causing distortion of the plots.

## Optimising the statistical pipeline for quantitative proteomics

A summary of the functional enrichment results from the analysis of dataset PXD007592 is shown in Figure 4.3.9. The BayesianT method for DE combined with the transformation based normalisations, log<sub>2</sub> (column 1), and to a lesser degree VSN (column 8, orange data), along with quantile normalisation gave the most functionally related proteins, but required a much less stringent significance threshold than previously seen with the BayesianT (best results were 105 significant terms from 194 DE proteins at significance threshold of 0.1). The classic *t*-test was less successful with this data, often calling very few proteins DE. However, even with central tendency based normalisation methods AIN and TIN (columns 2 and 4) and a lenient significance threshold, where more proteins were called DE than with the BayesianT, functional enrichment within these proteins was low, meaning that the low significance threshold was merely allowing an increased number of false positives to be identified as DE.

A comparison of the output from a standard Progenesis QIP output is (Progenesis normalisation and Welch's *t*-test using common significance thresholds of 0.01, 0.05, and 0.1 for defining DEs) is summarised in Table 43 and produced 2, 10, and 27 DE proteins, respectively. However, GO Enrichment analysis of the assigned DE proteins using clusterProfiler only returned 4 significantly enriched terms, at the 0.05 threshold.

Table 43; Enrichment analysis results for the dataset PXD007592 from the optimised pipeline and Progenesis output at standard significance threshold of 0.01, 0.05, and 0.1.

DE method	Normalisation	Threshold	Number of DE proteins	Number of significant terms
BayesianT	Log <sub>2</sub> transformation	0.1	194	105
<i>t</i> -Test	Progenesis	0.01	2	0
<i>t</i> -Test	Progenesis	0.05	10	4
<i>t</i> -Test	Progenesis	0.1	28	0

MAPKi therapy can be extremely effective in treatment of cerebral metastases from BRAF mutated melanoma. However, acquired drug resistance rates within 6-8 months are high (McArthur et al., 2014). The original paper for the dataset PXD007592 by Zila et al. (2018) compared samples from cerebral metastases of melanoma patients of those classed as good responders to MAPKi therapy

## ***Optimising the statistical pipeline for quantitative proteomics***

(progression-free survival  $\geq 6$  months) compared to poor responders (progression-free survival  $\leq 3$  months) of the treatment. Of the 194 proteins classed as DE by the optimal combination of methods 182 were upregulated in good responders and 12 were upregulated in poor responders. Progenesis QIP analysis using BH corrected  $p$  value of 0.05 for significance gave 5 proteins upregulated in good responders and 5 upregulated in poor responders.

In the original experiment by Zila *et al.* (2018) the authors highlight 34 proteins in the good responder group that are involved in the immune response, cell adhesion in the immune response, are apolipoproteins, or are extracellular matrix components. Of the 194 proteins classified as DE by the optimal analysis method in our experiment, 7 of the proteins from the original paper were classed as DE, plus a further 15 proteins belonging to the same subgroups not highlighted in the original paper. Furthermore, an additional 15 proteins with functions related to the complement and coagulation cascade were also classed as DE. Details of these proteins of interest are summarised in Table 44. By using a less conservative cut-off threshold and not taking into account a minimum fold change, the pipeline was able to consider more subtle changes in abundance between conditions which resulted in the inclusion of extra proteins of interest. Only three proteins had BH corrected  $p$ -values below 0.05, meaning 34 potentially important proteins would have been missed if the analysis had been performed using the current Progenesis QIP pipeline.

## Optimising the statistical pipeline for quantitative proteomics

Table 44; Summary of proteins upregulated in good responders to MAPKi therapy that are related to the functions highlighted by Zila et al. (2018) as classified by the optimal analysis

UniProt ID	Protein name	Log2 fold change	BH p value	Bayesian T FDR
<i>Immune response related proteins</i>				
P01889	HLA class I histocompatibility antigen, B alpha chain	1.71	0.611	0.002736
P01903	HLA class II histocompatibility antigen, DR alpha chain	2.38	0.646	0.027450
P01594*	Immunoglobulin kappa variable 1-33	3.07	0.629	0.029071
P01597	Immunoglobulin kappa variable 1-39	0.25	0.577	0.078188
P01593	Immunoglobulin kappa variable 1D-33	1.57	0.599	0.022865
P01619	Immunoglobulin kappa variable 3-20	2.00	0.511	0.078645
P06312	Immunoglobulin kappa variable 4-1	1.39	0.577	0.023603
P08514	Integrin alpha-L	5.26	0.118	0.000675
P04083	Annexin A1	1.28	0.611	0.009092
P05109	Protein S100-A8	1.35	0.612	0.052788
P06702	Protein S100-A9	1.76	0.599	0.016052
<i>Apolipoproteins</i>				
P02647*	Apolipoprotein A-I	0.93	0.488	0.041272
P06727*	Apolipoprotein A-IV	2.75	0.118	0.000003
P04114*	Apolipoprotein B-100	1.33	0.059	0.000255
<i>Extracellular matrix (ECM) components</i>				
P02452*	Collagen alpha-1(I) chain	2.77	0.655	0.030311
P12109	Collagen alpha-1(VI) chain	2.95	0.655	0.033266
P39059*	Collagen alpha-1(XV) chain	3.36	0.646	0.017357
P12110	Collagen alpha-2(VI) chain	3.30	0.595	0.000613
P12111	Collagen alpha-3(VI) chain	2.89	0.611	0.000499
P02751*	Fibronectin	1.67	0.642	0.044443
P23142	Fibulin-1	1.42	0.665	0.075334
Q13201	Multimerin-1	3.95	0.697	0.024751
<i>Coagulation and complement cascade proteins</i>				
P00488	Coagulation factor XIII A chain	2.02	0.611	0.026127
P01042**	Kininogen-1	0.96	0.008	0.000318
P01008	Antithrombin-III	0.78	0.201	0.021107
P02675	Fibrinogen beta chain	2.98	0.507	0.000083
P02671	Fibrinogen alpha chain	2.98	0.507	0.000474
P02679	Fibrinogen gamma chain	2.89	0.595	0.000593
Q14112	Nidogen-2	1.89	0.653	0.027030
P35555	Fibrillin-1	2.36	0.678	0.046089
P06681	Complement C2	0.80	0.087	0.019176
P0COL4**	Complement C4-A	0.99	0.028	0.004501
P13671**	Complement component C6	1.48	0.003	0.000040
P07358	Complement component C8 beta chain	1.18	0.109	0.000218
P02748	Complement component C9	0.64	0.321	0.015037
P08603	Complement factor H	0.60	0.611	0.092305
P10909	Clusterin	1.28	0.461	0.005979

\* Protein was classed as DE by analysis in PXD007592 paper

\*\* Protein was classed as DE by current Progenesis QIP pipeline

## ***Optimising the statistical pipeline for quantitative proteomics***

Of the remaining 27 proteins originally highlighted in the good responder group by Zila *et al.*, 16 were not identified in our pre-processing and four of them were down regulated according to the Progenesis QIP quantification values. The remaining seven that were not classified as DE by the optimal analysis are described in Table 45. BH corrected p-values ranged from 0.61 – 0.96, meaning if the analysis had been performed using the current Progenesis QIP pipeline, again none of these proteins would have been selected as DE.

*Table 45; Summary of proteins highlighted by Zila et al. (2018) that were upregulated in good responders to MAPKi therapy missed by the optimal analysis*

<b>UniProt ID</b>	<b>Protein name</b>	<b>Log2 fold change</b>	<b>BH p value</b>	<b>Bayesian T FDR</b>
P01857	Ig gamma-1 chain C region	0.31	0.61	0.265
P01834	Ig kappa chain C region	0.67	0.65	0.182
Q05707	Collagen alpha-1(XIV) chain	3.13	0.76	0.133
P01871	Ig mu chain C region	1.49	0.81	0.444
P16284	Platelet endothelial cell adhesion molecule	2.39	0.90	0.658
P08123	Collagen alpha-2(I) chain	1.60	0.90	0.684
Q5Y7A7	HLA class II histocompatibility antigen, DRB1-13 beta chain	0.30	0.96	1.000

To investigate the biological relevance of the changing proteins, pathway enrichment was performed for the upregulated proteins based on GO terms for biological process, cellular component, and molecular function, along with KEGG pathways using ClusterProfiler. Figure 4.3.10 displays the enriched KEGG pathways from the optimal analysis and the current Progenesis QIP pipeline. In comparison to Progenesis QIP analysis, the optimal method has a higher number and more significant results, along with a higher gene ratio. In the original paper, only the ‘Complement and coagulation cascades’ KEGG pathway was enriched with this dataset. This pathway also appeared in the results from both the optimal and Progenesis QIP analysis. Additionally, the optimal analysis also provided significant enrichment in several other relevant pathways: Tight junctions are intracellular adhesion complexes that help establish epithelial cell polarity which prevents cancer cell invasion and tumour progression (Martin-Belmonte and Perez-Moreno, 2012). Chen *et al.* (2019) identified ‘amoebiasis’, ‘ECM-receptor interaction’ and ‘focal adhesion’ signaling pathways as being important in the formation of metastases from melanoma. The actin

## Optimising the statistical pipeline for quantitative proteomics

cytoskeleton pathway regulates cell motility, which is required for immune surveillance and helps prevent cancer invasion and metastasis (Li et al., 2016). Proteoglycans are key macromolecules that are involved in proliferation, adhesion, angiogenesis and metastasis of cancer. Furthermore, Zila et al. (2018) discuss the relevance of the up-regulated immune response related proteins in the good responders. Our analysis provides further evidence of an activated immune system in good responders through the enrichment of the disease pathways 'Systemic lupus erythematosus', 'Staphylococcus aureus infection', 'Human papillomavirus infection', 'Coronavirus disease – COVID-19', and 'African trypanosomiasis'.

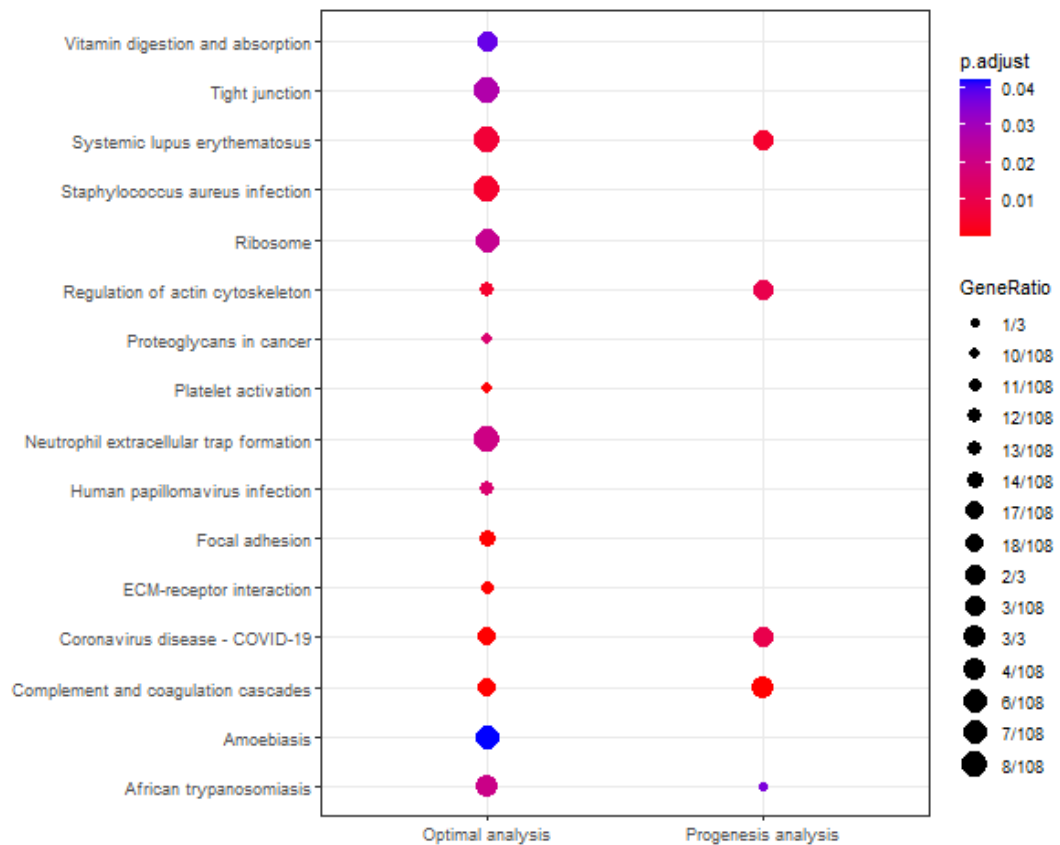


Figure 4.3.10; KEGG pathways mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.



## Optimising the statistical pipeline for quantitative proteomics

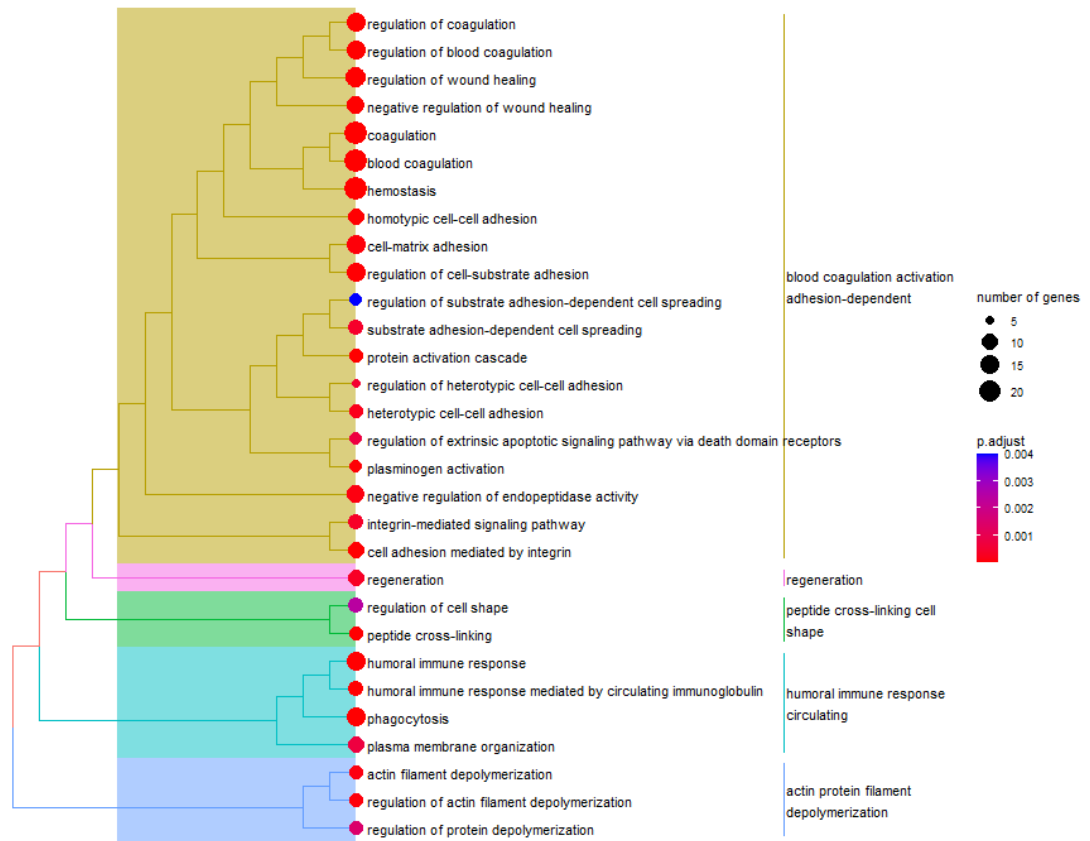


Figure 4.3.11; Most significantly enriched GO BP terms following hierarchical clustering based on the pairwise similarities of high frequency words mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.

Figure 4.3.11 Figure 4.3.11 summarises the most significantly enriched GO BP terms following hierarchical clustering based on the pairwise similarities of high frequency words, which was performed to improve interpretation of the large number of enriched terms. The main subtrees contain processes belonging to coagulation, the immune response, and cell stabilization and regeneration which are processes relevant to the main pathways discussed in the original paper. However, Zila et al. only describe 9 enriched BP pathways in the good responders. Our analysis discovered 75 enriched BP terms after redundancy was removed using `simplify()`.

## Optimising the statistical pipeline for quantitative proteomics

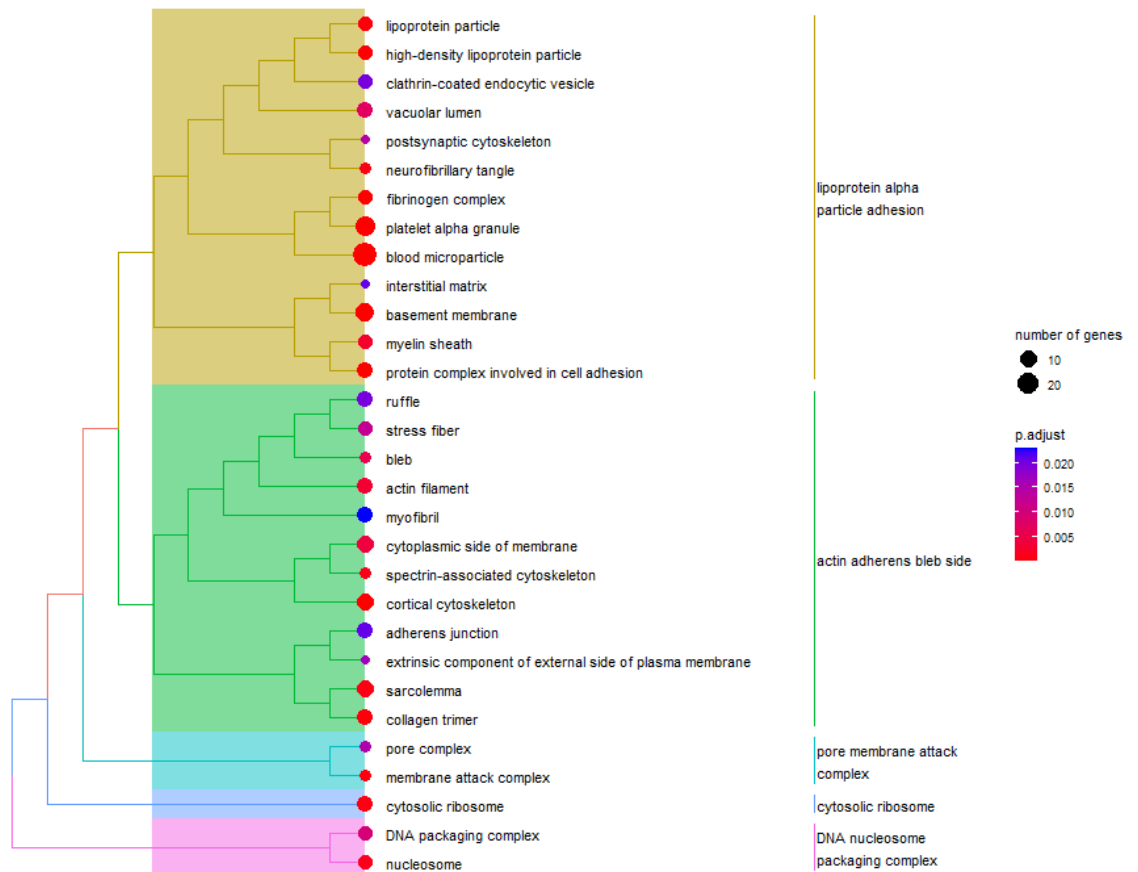


Figure 4.3.12; Most significantly enriched GO CC terms following hierarchical clustering based on the pairwise similarities of high frequency words mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.

Figure 4.3.12 summarises the most significantly enriched CC terms following hierarchical clustering. The original paper only discovered two significantly enriched terms: extracellular region and matrix. Our optimal analysis contained 41 significantly enriched terms including terms associated with ECM and also wound healing.

Enrichment analysis provided 29 significant GO MF terms. Following hierarchical clustering, these are summarised in Figure 4.3.13. The main subtrees contain processes belonging to calcium-dependent ECM receptors, DNA binding, structural constituents of ribosomes and the myelin sheath, and carboxypeptidase activity. There were a large number of terms related to calcium-dependent ECM receptors which is potentially of interest as the calcium sensor receptor has been identified as a potential prognostic marker for

## Optimising the statistical pipeline for quantitative proteomics

metastasis (Tharmalingam and Hampson, 2016). The original paper did not uncover any enriched MF pathways in the good responders.

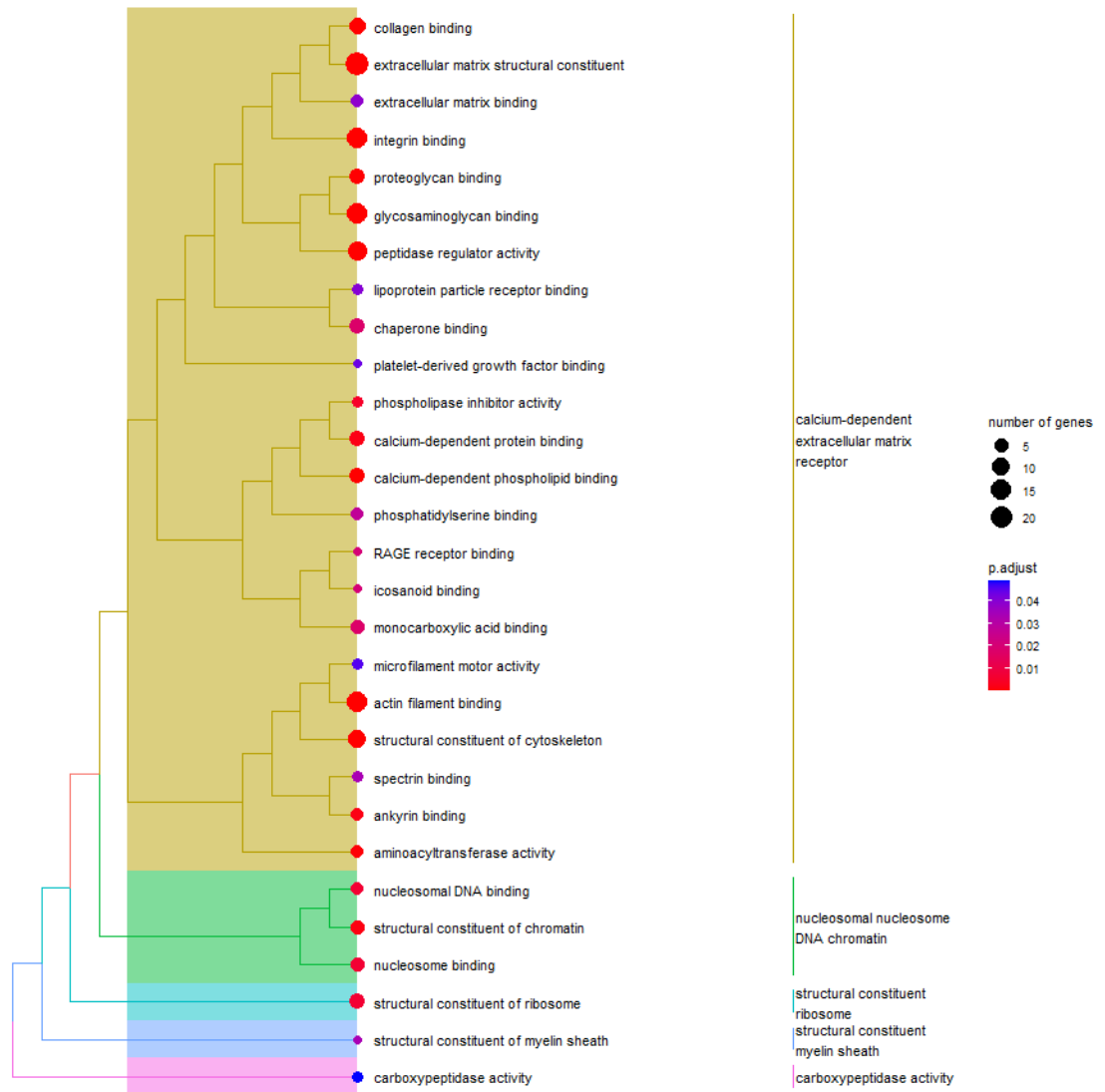


Figure 4.3.13; Most significantly enriched GO MF terms following hierarchical clustering based on the pairwise similarities of high frequency words mapped by upregulated proteins in the good responders to MAPKi therapy by optimal and Progenesis analysis.

Our optimal analysis was provided by using log<sub>2</sub> transformed abundances and Bayesian t-test for DE analysis with a 0.1 FDR. Use of this relatively generous cut-off threshold without the constraints of a minimum fold-change parameter provided more enriched terms than current Progenesis QIP pipeline, many of which had been highlighted as interesting in the original paper for the dataset PXD007592. In good responders to MAPKi therapy, our analysis derived some

## ***Optimising the statistical pipeline for quantitative proteomics***

of same proteins and also additional proteins with similar functions. Our analysis also provided more information than in the original paper with a greater number of significant pathways with potentially interesting terms requiring further investigation.

The original paper for dataset PXD007592 identified 5977 proteins using MaxQuant software with the Andromeda search engine. Original DE analysis was conducted using the Perseus statistical analysis package which applied a two-sided Student's *t*-test. A  $p < 0.05$  with an FDR based permutation correction along with a minimum of two-fold abundance difference resulted in 1907 proteins being defined as changing in abundance between good and poor responders to MAPKi treatment. Due to processing difficulties, only three out of the five out of the six good responder samples were included in processing the raw data using Progenesis QIP. The outcome of this, along with the effect of using different software to process this experiment (Progenesis QIP and Mascot as opposed to MaxQuant and Andromeda) resulted in identification of only 3098 protein groups and the optimal analysis occurred when just 194 proteins were classed as DE. To make direct comparison to the original paper would require using MaxQuant data and identical protein quantification inputs. Proteins highlighted as being important in the original paper that were missed from being identified as DE by our optimal analysis, were all accompanied by BH corrected *p*-values greater than 0.05, and therefore would not have been identified as changing by the original paper if they had used our quantification data. Progenesis processing of the raw data also produced different log fold change values which resulted in some of the proteins being classed as upregulated in the original paper but downregulated according to our quantification analysis. The aim of this benchmarking was to provide a metric with which to compare analysis in order to improve Progenesis QIP output therefore processing with other software was not included. However, based on the promising performance of the pipeline. It would be useful expand its application to include MaxQuant input data and to reanalyse this data in future work.

### ***Overall analysis***

Overall, the proteins identified as being DE by the optimal analysis provided better biological context that related to the original experiments than those identified using current the Progenesis pipeline. Enrichment analysis of these proteins produced terms similar to those in the original experiments often with higher gene ratios and more significant terms, along with further terms possibly relevant to the disease investigated not discussed in the original paper. The comparison of our analysis to that of the original paper was limited by different processing methods of the raw data. The datasets were chosen to provide large protein lists with which to demonstrate the software and we were bound by Progenesis QIP analysis as this was the focus of project. However, our processing did not identify all proteins from original papers. Progenesis QIP processing provided a different amount of identified proteins with different quantification values. Furthermore, the biological context discussed in the papers was obtained from a combination of strategies, such as in-depth proteome profiling, targeted approaches, and validation, rather than just the pairwise comparison of the data used in our analysis. We attempted to overcome this problem for datasets PXD004501 and PXD004682, by performing our own enrichment analysis on the list of DE proteins was supplied as supplementary material and using this establish the relevance of our results.

Looking at the analysis over all of the datasets, using the classic *t*-test for DE analysis appeared to work better with the central tendency normalisation methods of AIN and TIN. For BayesianT analysis, transforming the data using VSN or log2 often provided the best combination. Developing this pipeline further, it may become apparent that only specific DE and normalisation combinations may be necessary, reducing the computational running time. However, further testing will be required before this becomes apparent.

When comparing results to the original experiments it appeared that the BH correction of *p*-values may have prevented the current Progenesis QIP output from identifying proteins of interest from being classed as DE. In the original analysis of dataset PXD004501, the authors concluded that protein POSTN was a significant biomarker for gastric cancer and peritoneal seeding. Highlighted as

being upregulated by our optimal analysis,  $t$ -test provided a  $p$ -value of 0.05 that was BH corrected to 0.14, meaning the protein was missed from this analysis, highlighting the problems with using an arbitrary cut-off value for significance and methods to correct for multiple testing. A limitation of this chapter has been using different threshold metrics to compare DE analysis. In future work, an alternative method could be implemented where the test statistic is used to rank proteins, then gradually increasing proportions of more significant proteins are used as the test set for pathway analysis.

The implementation of clusterProfiler was a much improved implementation over RDAVID. There were no technical issues in executing the software, and the addition of the `simplify()` function gave condensed terms which are easier to interpret. Even with redundant terms removed, there were occasions when a greater number of significant terms were produced, presumably due to the querying of current ontologies. However, there is a limitation in performing a direct comparison between the results and those of Chapter 3 as the log transformation was performed at different stages. This is addressed in the next section where QPROT analysis is performed on data normalised using this pipeline.

Currently the software is only able to perform pairwise comparisons. It would be useful to extend the software to allow for group comparison. This could be possible with some amendment to the algorithm by providing a comparison matrix. It will also be possible to add further options for normalisation and DE analysis through additional algorithms written in R.

ii. Comparison to QPROT output

650 iterations

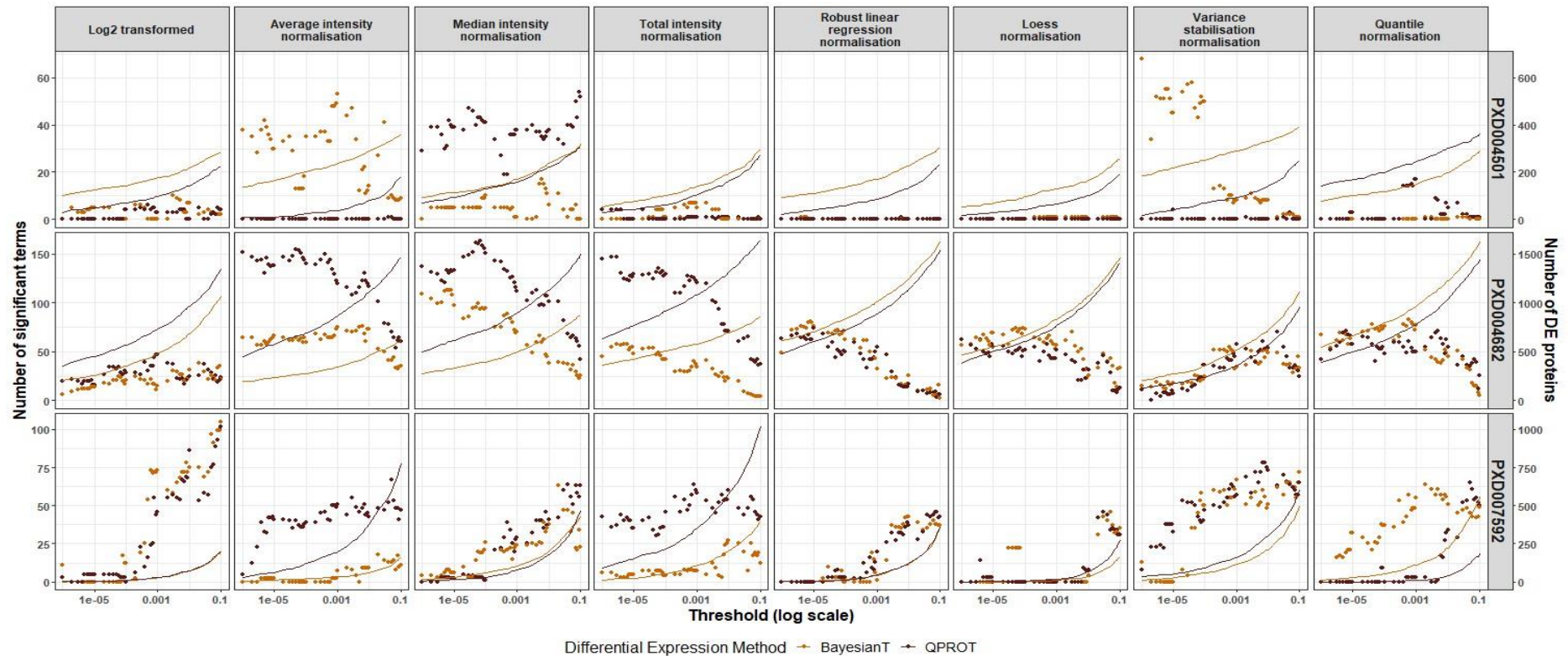


Figure 4.3.14; Enrichment analysis of results from optimised pipeline analysis of datasets. DE analysis by BayesianT (orange) using 650 iterations of sampling with 325 iterations for the burnin and QPROT (brown) of protein abundances (identified with by a minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.

2000 iterations

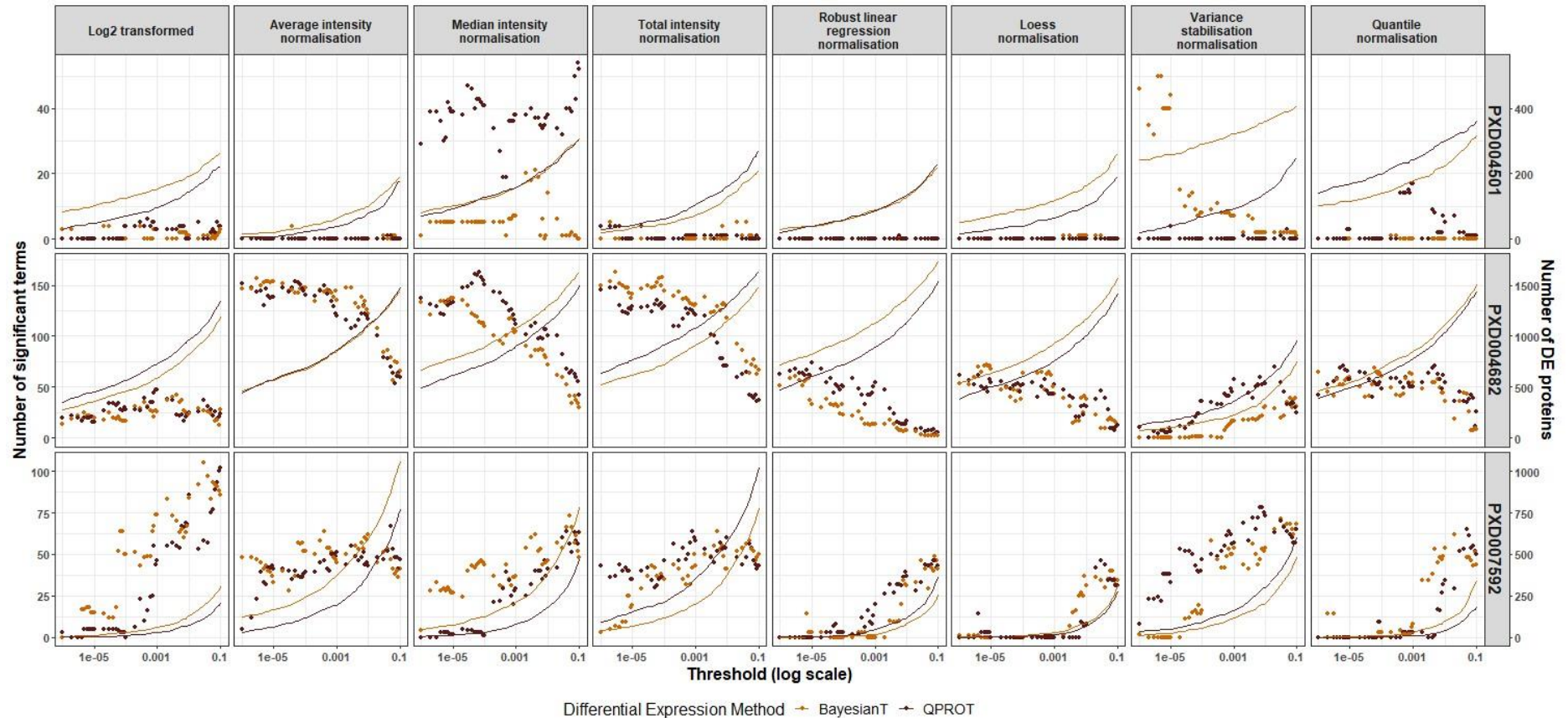


Figure 4.3.15; Enrichment analysis of results from optimised pipeline analysis of datasets. DE analysis by BayesianT (orange) using 2000 iterations of sampling with 1000 iterations for the burnin and QPROT (brown) of protein abundances (identified with by a minimum of one unique peptide) and normalised by the method shown in the x-axis strip. DE proteins identified using the significance threshold on the x-axis (number of proteins shown as solid line) were subjected to pathway analysis. Number of significant enrichment terms are shown as points.



## *Optimising the statistical pipeline for quantitative proteomics*

Initial comparison of the Bayesian T with 650 iterations for sampling and 325 for burnin analysis against QPROT analysis of the same data (Figure 4.3.14) shows similar output for the low ranking normalisation methods for dataset PXD004501 (Figure 4.3.14, row 1). QPROT analysis performed better when the data was normalised with MIN, and Bayesian T gave better performance with AIN and VSN. However, increasing the number of iterations to 1000 and 2000 for sampling and burnin, respectively (Figure 4.3.15;5, row 1), reduced the effectiveness of the best normalisations for Bayesian T analysis, leaving QPROT as the better analysis method.

For dataset PXD004682 (Figure 4.3.14, row 2), the performance was the same for Bayesian T and QPROT, except for data normalised with AIN, MIN, and TIN, where QPROT analysis was superior. When the sampling iterations were increased to 2000 with 100 burnin (Figure 4.3.15;, row 2), Bayesian T analysis was improved and gave a similar performance to that of QPROT.

In dataset PXD007592 (Figure 4.3.14, row 3), QPROT analysis was superior to Bayesian T with AIN and TIN data, and Bayesian T analysis was superior with Quantile normalised data. All other normalised data had similar outputs. With increasing sampling and burnin iterations (Figure 4.3.15;5, row 3), Bayesian T performance improved, except with VSN and quantile normalisation.

Increasing the sampling number improved the performance of the Bayesian T for some normalisation methods in two of the datasets. However, these results show both increased sampling and burnin number, causing a drop in performance for some normalisation methods. This indicates that a large number of iterations along with a small number of burnin iterations could give optimal output in some cases. However, this will increase the length of time that the program will run, and further work is required to assess the trade-off between processing speed and optimal analysis. The results demonstrate the importance of selecting the appropriate number of iterations for sampling and indicate that there are limitations to the results described in section 4.3.i, as this additional parameter has not been taken into account.

## **4.4. Conclusions**

In this chapter, we developed, validated and implemented an optimised pipeline that utilises HPC for parallelisation of multiple combinations of methods for DE analysis, normalisation and significance threshold selection. Functional enrichment analysis of proteins defined as DE was applied as a metric to determine the most successful combination of methods and the results are returned to the user. The implementation of the BayesianT algorithm for DE analysis and local FDR estimation into the R pipeline were successfully validated to perform as well or better than original software. The application of clusterProfiler and the simplify() function reduced the issue of redundancy in enrichment analysis, and provided up-to-date enrichment analysis along with improved functionality compared to RDAVID. Analysis of datasets from Chapter 3 showed the optimised pipeline's performance was superior to a standard analysis using Progenesis QIP.

## **Chapter 5. Thesis conclusions and outlook**

### **Summary of thesis results**

Proteomics is the study of the quantity of proteins in a cell, tissue or organism. The abundance of each protein in a cell is governed by gene expression and protein degradation, and can, for example, provide insights into biological response to disease. Quantitative proteomics experiments compare the amount of protein in a sample at a different time or under different conditions. Label-free quantitative proteomics utilises high-throughput MS for global analysis of complex samples which provides information about the absolute or relative differences in protein quantities between samples.

A key step in quantitative proteomics is deciding which proteins are changed in abundance between sample groups, using statistical techniques i.e. DE analysis. DE analysis allows us to determine which proteins are changing across samples, highlighting potential biomarkers, reveals protein involvement in metabolic pathways or facilitates research into drug discovery for treatment and prevention. Such experiments produce vast amounts of data requiring dedicated software for analysis, and statistical methods to confidently define proteins changing in abundance due to experimental conditions. However, due to the properties of proteomics data there can be problems reliably comparing conditions. Small sample sizes containing a large number of features provide large sample-to-sample variation with outliers distorting analysis and creating false positives when analysed with the commonly used *t*-test. The main aim of this Industrial CASE PhD studentship, in collaboration the developers of Progenesis QIP, was to provide an improved statistical pipeline that could be implemented in the Progenesis QIP workflow.

In Chapter 2, three approaches to DE were evaluated; linear modelling using the MSstats package, and Bayesian analysis using QPROT, alongside Welsh's *t*-test, the current DE analysis method offered by Progenesis QIP in the form of ANOVA. Benchmarking was carried out using the common practice of analysis of artificial spike-in datasets to simulate a biological situation where known proteins change across conditions. Having a 'ground truth' allows us to

demonstrate how well the analysis methods are able to detect the proteins that we know are of different abundances between sample groups. However, the results of the benchmarking exercise were inconclusive and further investigation of spike-in data revealed problems with its accuracy and precision, undermining the results of the software evaluation. Further issues are that technical replicates common in benchmarking data lack inherent biological variation provide poor representation of the proportion of proteome changing and the directions in which change occurs. Artificial data limits scope and impairs the statistical power of an experiment. However, in a typical benchmarking exercise, authentic biological data cannot be used, as we do not know the true number of changing proteins with which to calculate sensitivity and specificity. These problems led to the development of a novel profiling-based alternative method for statistical analytical evaluation that does not rely on the use of artificial spike-in data.

In Chapter 3, we defined and tested a benchmarking method that relied on the results of enrichment analysis for evaluation of software. Quantitative proteomics experiments are usually followed by downstream analysis, with the aim of discovering many functionally related groups within the changing proteome. Pathway enrichment analysis provides biological context to the DE results, and a greater number of significant terms would be expected from an accurate analysis, as the change in abundance would be linked to functionality. The method was used to assess DE analysis, normalisation techniques and the selection of the most appropriate significance threshold using three biological datasets. The results highlighted the importance of applying the optimal normalisation method in DE analysis. However, overall there was no one clear optimal method and the effect of combining methods differed depending on the characteristics of the individual datasets. Following this analysis, going forward the sensible outcome appeared to be to develop a pipeline that was able to perform simultaneous analysis of the possible parameter combinations in parallel, evaluate the results using functional enrichment, and return details of the best workflow for that particular dataset to the user, along with DE proteins and pathway analysis details.

Chapter 4 describes the development and implementation of an optimised pipeline. Implemented as bash script initiated Rscripts, the pipeline runs on a Linux server on an HPC, allowing parallelisation of computationally heavy stages of analysis. The pipeline provides eight methods for normalisation using the Normalyzer DE package. There are two methods for DE analysis; Welch's two-tailed *t*-test and an implementation called 'BayesianT' written in R and Rstan, and inspired by principles employed in the QPROT package. A large range of significance thresholds are used for defining DE proteins and pathway analysis is performed using the R package clusterProfiler to decide which combination of parameters provides the most functionally related group of DE proteins. The pipeline was validated and demonstrated using the biological datasets from Chapter 3. To our knowledge, this is the only end-to-end pathway analysis pipeline designed for proteomics data, enabling users to iterate through multiple options for finding the best normalisation method and the best significance threshold for pathway analysis.

### ***Further work***

Currently the running of the package takes several hours, depending on the cluster availability and the size of the dataset being analysed. Areas to potentially reduce the running time could be to perform a screening run with a smaller number of significance threshold iterations. The results of this may indicate the range where the optimal significance threshold is likely to be, allowing a second run with smaller increments within this range only. Also, implementing a minimum and maximum proportion of proteins being called as DE before proceeding to perform enrichment analysis on these would also reduce running time as the results in Chapter 4 showed there appeared to be a sweet spot between too many or too few DEs. Dataset with hundreds of proteins are unlikely to produce the best enrichment results if a stringent significance threshold means only one or two proteins are being labelled as DE, and likewise if the significance threshold is lenient and identifies most of the proteins as changing. Further work must also be carried out to assess the optimal running of the sampling in the Bayesian T. Increasing the number of

## *Optimising the statistical pipeline for quantitative proteomics*

iterations for sampling and using a small number of burnin iterations will increase the length of time that the program runs, and this trade-off between performance and processing speed requires assessment.

Current installation of the pipeline is through GitHub download to be run on a Linux server. There is further work planned to improve the pipeline and provide it as a publication ready wrapped package through containerisation. Containerisation software provides portability through encapsulation of entire environments including dependencies, libraries, runtime code and data. A container is a secure, standardised, and efficient method for sharing package software without the need for local installation. Software compiled on different platforms requires specific libraries and compilers. Containerising software allows it to run reliably when moving between computing environments giving reproducibility as binaries and libraries are packaged up so that software uses the same files to run each time. A popular container is Docker (Merkel, 2014). Originally designed for businesses, it is useful for network centric web servers and databases rather than HPC systems. It was designed for trusted users running trusted containers through root access rather than batch job schedulers or Message Passing Interface applications for HPC clusters. Docker containers have are isolated and have no access to the host file system. Another methods for containerisation is Singularity, developed by Kurtzer et al. (2017). Singularity runs on a no trust security model and container models can be run without root access. Entering a container without root privileges prevents escalation of root privileges within the container. The container is built on a local Linux system that you have root access to using a definition file which is a list of software requirements of the custom container (Sylabs.io, 2021). The file specifies the type of OS, software, environment variables to set at runtime, files to add from the host system and container metadata. The container is then transferred to the HPC system where it will be run. Implementation of containerisation using Singularity, plus the addition of a graphical user interface are the intended next stages of work in order to provide publication ready software. It would also be useful to add Reactome and Kegg pathways analysis for more comprehensive results. New releases of ReactomePA will be

monitored for inclusion when there is an opportunity to employ the `simplify()` function for reducing redundancy, and `clusterProfiler` will be monitored for offline KEGG queries.

### ***Limitations***

Processing of both our benchmarking data and our biological data was performed using Progenesis QIP with peak picking set to maximum. This was conducted as previously unpublished work from our group had demonstrated that this gives the best performance from the software. However, this change from default parameter use is different from the method employed by the typical biologist, affecting the applicability of our results. A further limitation is in the blanket imputation of 0.0000001 for zero values. There is no current consensus on the appropriate method to deal with missing data in proteomics. However, due to Progenesis QI's alignment algorithm, it is claimed that zero abundance values in its analysis are true missing proteins and are appropriate. Therefore, due to the focus of this thesis being Progenesis QIP data, missing value methods were not investigated and imputation was performed to prevent problems due to division of zeros. Although we investigated the impact of this imputation and found it to be negligible, this does affect the wider application of the pipeline and must be considered in further work and the full treatment of imputation of missing values would be an interesting future addition to the pipeline.

In our analysis, we compared DE analysis methods that used different metrics for calculating significance cut-off thresholds, QPROT (FDR) and *t*-test and MSstats (BH corrected *p*-values). This makes direct comparison difficult and may have prevented optimal output from being captured. Our aim was to recreate conditions applied by a biologist and to highlight problems with application of arbitrary cut-off thresholds. However, in future work an alternative method will be considered where proportions of proteins ranked on their test statistic are used, allowing like for like comparison of performance.

A further limitation is that the pipeline's performance was assessed using a default number of iterations for sampling and burnin. However, further analysis

found that increasing the sampling number improved the performance of the Bayesian T for some normalisation methods in two of the datasets. This demonstrates that selecting the correct parameter should be taken into account.

### ***Conclusions***

The aim of this project was to provide an improved statistical pipeline that could be implemented in the Progenesis QIP workflow. Following benchmarking existing differential expression methods using ground truth data, a novel technique was developed to benchmark using biological data and functional enrichment analysis to provide an evaluation metric. We then developed, validated and implemented an optimised pipeline that utilises HPC for parallelisation of multiple combinations of methods for DE analysis, normalisation and significance threshold selection. The most successful combination of methods and the differential expression analysis results are returned to the user. It was shown that the output provided more functionally enriched groups of proteins that would have been achieved with the current output of Progenesis QIP, indicating that the pipeline would be a successful improvement to the present Progenesis QIP workflow.

## **Supplementary material**

Link to OneDrive folder

<https://1drv.ms/u/s!AhU3p9jJHqUNn8FByup3CzsiCASCTA?e=FnUUcz>

Chapter 2 - Investigation for optimal burn-in and number of iterations for QPROT

Chapter 3 – DAVID terms for all analysis parameters

Chapter 4 – ClusterProfiler GO terms for all analysis parameters



## References

2015. The difficulty of a fair comparison. *Nature Methods*, 12, 273-273.
2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*, 47, D330-d338.
- AEBERSOLD, R. 2011. Editorial: from data to results. *Molecular & cellular proteomics : MCP*, 10, E111.014787-E111.014787.
- AEBERSOLD, R. & MANN, M. 2003. Mass spectrometry-based proteomics. *Nature*, 422, 198-207.
- AHAD, N. A. & YAHAYA, S. S. S. 2014. Sensitivity analysis of Welch's t-test. *AIP Conference Proceedings*, 1605, 888-893.
- AL SHWEIKI, M. R., MÖNCHGESANG, S., MAJOVSKY, P., THIEME, D., TRUTSCHEL, D. & HOEHENWARTER, W. 2017. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J Proteome Res*, 16, 1410-1424.
- ALLEN, J. S. 1947. An Improved Electron Multiplier Particle Counter. *Review of Scientific Instruments*, 18, 739-749.
- ALTEROVITZ, G., LIU, J., AFKHAMI, E. & RAMONI, M. F. 2007. Bayesian methods for proteomics. *Proteomics*, 7, 2843-55.
- ANKNEY, J. A., MUNEEER, A. & CHEN, X. 2016. Relative and Absolute Quantitation in Mass Spectrometry-Based Proteomics. *Annual Review of Analytical Chemistry*, 11, 49-77.
- ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25, 25-9.
- BAIG, S. A. 2020. Bayesian Inference: An Introduction to Hypothesis Testing Using Bayes Factors. *Nicotine & Tobacco Research*, 22, 1244-1246.
- BALLMAN, K. V., GRILL, D. E., OBERG, A. L. & THERNEAU, T. M. 2004. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, 20, 2778-2786.
- BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289-300.
- BERLIN, J. A., BEGG, C. B. & LOUIS, T. A. 1989. An Assessment of Publication Bias Using a Sample of Published Clinical Trials. *Journal of the American Statistical Association*, 84, 381-392.
- BETANCOURT, M. 2016. Identifying the Optimal Integration Time in Hamiltonian Monte Carlo.
- BETANCOURT, M. 2017. A Conceptual Introduction to Hamiltonian Monte Carlo.
- BJORNSON, R. D., CARRIERO, N. J., COLANGELO, C., SHIFMAN, M., CHEUNG, K.-H., MILLER, P. L. & WILLIAMS, K. 2008. X!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *Journal of proteome research*, 7, 293-299.
- BLANKSBY, S. J. & MITCHELL, T. W. 2010. Advances in Mass Spectrometry for Lipidomics. *Annual Review of Analytical Chemistry*, 3, 433-465.

- BOLSTAD, B. M., IRIZARRY, R. A., ASTRAND, M. & SPEED, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 185-93.
- BOOTH, J. G., EILERTSON, K. E., OLINARES, P. D. B. & YU, H. 2011. A bayesian mixture model for comparative spectral count data in shotgun proteomics. *Molecular & cellular proteomics : MCP*, 10, M110.007203-M110.007203.
- BRADY, S. M., BUROW, M., BUSCH, W., CARLBORG, Ö., DENBY, K. J., GLAZEBROOK, J., HAMILTON, E. S., HARMER, S. L., HASWELL, E. S., MALOOF, J. N., SPRINGER, N. M. & KLIEBENSTEIN, D. J. 2015. Reassess the t Test: Interact with All Your Data via ANOVA. *The Plant cell*, 27, 2088-2094.
- BROSCH, M., YU, L., HUBBARD, T. & CHOUDHARY, J. 2009. Accurate and sensitive peptide identification with Mascot Percolator. *Journal of proteome research*, 8, 3176-3181.
- BROWN, K. L. & TAUTFEST, G. W. 1956. Faraday-Cup Monitors for High-Energy Electron Beams. *Review of Scientific Instruments*, 27, 696-702.
- CAI, W., TUCHOLSKI, T. M., GREGORICH, Z. R. & GE, Y. 2016. Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert review of proteomics*, 13, 717-730.
- CALLISTER, S. J., BARRY, R. C., ADKINS, J. N., JOHNSON, E. T., QIAN, W.-J., WEBB-ROBERTSON, B.-J. M., SMITH, R. D. & LIPTON, M. S. 2006. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of proteome research*, 5, 277-286.
- CARLSON, M. 2019. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2.
- CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. & RIDDELL, A. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76, 1 - 32.
- CHANG, C.-Y., PICOTTI, P., HÜTTENHAIN, R., HEINZELMANN-SCHWARZ, V., JOVANOVIĆ, M., AEBERSOLD, R. & VITEK, O. 2012. Protein Significance Analysis in Selected Reaction Monitoring (SRM) Measurements. *Molecular & Cellular Proteomics*, 11, M111.014662.
- CHAWADE, A., ALEXANDERSSON, E. & LEVANDER, F. 2014. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *Journal of Proteome Research*, 13, 3114-3120.
- CHEN, J., WU, F., SHI, Y., YANG, D., XU, M., LAI, Y. & LIU, Y. 2019. Identification of key candidate genes involved in melanoma metastasis. *Mol Med Rep*, 20, 903-914.
- CHENG, L., LOPEZ-BELTRAN, A., MASSARI, F., MACLENNAN, G. T. & MONTIRONI, R. 2018. Molecular testing for BRAF mutations to inform melanoma treatment decisions: a move toward precision medicine. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 31, 24-38.
- CHOI, H., FERMIN, D. & NESVIZHSKII, A. I. 2008. Significance analysis of spectral count data in label-free shotgun proteomics. *Molecular & cellular proteomics : MCP*, 7, 2373-2385.

- CHOI, H., KIM, S., FERMIN, D., TSOU, C. C. & NESVIZHSKII, A. I. 2015. QPROT: Statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. *J Proteomics*, 129, 121-126.
- CHOI, M., CHANG, C. Y., CLOUGH, T., BROUDY, D., KILLEEN, T., MACLEAN, B. & VITEK, O. 2014. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30, 2524-6.
- CLEVELAND, W. S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829-836.
- CLOUGH, T., KEY, M., OTT, I., RAGG, S., SCHADOW, G. & VITEK, O. 2009. Protein Quantification in Label-Free LC-MS Experiments. *Journal of Proteome Research*, 8, 5275-5284.
- CLOUGH, T., THAMINY, S., RAGG, S., AEBERSOLD, R. & VITEK, O. 2012. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics*, 13, S6.
- COLQUHOUN, D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1, 140216.
- COX, J., HEIN, M. Y., LUBER, C. A., PARON, I., NAGARAJ, N. & MANN, M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP*, 13, 2513-2526.
- COX, J. & MANN, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26, 1367-72.
- DABNEY, A. R. & STOREY, J. D. 2007a. A new approach to intensity-dependent normalization of two-channel microarrays. *Biostatistics*, 8, 128-39.
- DABNEY, A. R. & STOREY, J. D. 2007b. Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome biology*, 8, R44-R44.
- DAHIRU, T. 2008. P - value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*, 6, 21-26.
- DALMAN, M. R., DEETER, A., NIMISHAKAVI, G. & DUAN, Z.-H. 2012. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, 13, S11.
- DALY, D. S., ANDERSON, K. K., PANISKO, E. A., PURVINE, S. O., FANG, R., MONROE, M. E. & BAKER, S. E. 2008. Mixed-Effects Statistical Model for Comparative LC-MS Proteomics Studies. *Journal of Proteome Research*, 7, 1209-1217.
- DE HERTOIGH, B., DE MEULDER, B., BERGER, F., PIERRE, M., BAREKE, E., GAIGNEAUX, A. & DEPIEREUX, E. 2010. A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC bioinformatics*, 11, 17-17.
- DEFOSSEZ, E., BOURQUIN, J., VON REUSS, S., RASMANN, S. & GLAUSER, G. 2021. Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, n/a.
- DEUTSCH, E. W., LANE, L., OVERALL, C. M., BANDEIRA, N., BAKER, M. S., PINEAU, C., MORITZ, R. L., CORRALES, F., ORCHARD, S., VAN EYK, J. E., PAIK, Y. K., WEINTRAUB, S. T., VANDENBROUCK, Y. & OMENN, G. S. 2019. Human

- Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J Proteome Res*, 18, 4108-4116.
- DOREY, F. 2010. The p value: what is it and what does it tell you? *Clinical orthopaedics and related research*, 468, 2297-2298.
- DUANE, S., KENNEDY, A. D., PENDLETON, B. J. & ROWETH, D. 1987. Hybrid Monte Carlo. *Physics Letters B*, 195, 216-222.
- DUDOIT, S., YANG, Y. H., CALLOW, M. J. & SPEED, T. P. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 111-139.
- DUMMER, R., GOLDINGER, S. M., TURTSCHI, C. P., EGGMANN, N. B., MICHIELIN, O., MITCHELL, L., VERONESE, L., HILFIKER, P. R., FELDERER, L. & RINDERKNECHT, J. D. 2014. Vemurafenib in patients with BRAF(V600) mutation-positive melanoma with symptomatic brain metastases: final results of an open-label pilot study. *Eur J Cancer*, 50, 611-21.
- DUNN, O. J. 1961. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56, 52-64.
- EFRON, B., TIBSHIRANI, R., STOREY, J. D. & TUSHER, V. 2001. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151-1160.
- ELO, L. L., FILEN, S., LAHESMAA, R. & AITTOKALLIO, T. 2008. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Trans Comput Biol Bioinform*, 5, 423-31.
- ENG, J. K., JAHAN, T. A. & HOOPMANN, M. R. 2013. Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13, 22-4.
- ENG, J. K., MCCORMACK, A. L. & YATES, J. R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5, 976-989.
- FENN JOHN, B., MANN, M., MENG CHIN, K., WONG SHEK, F. & WHITEHOUSE CRAIG, M. 1989. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246, 64-71.
- FIEHN, O. 2002. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, 48, 155-171.
- FISHER, R., A. 1925. *Statistical Methods For Research Workers*, Edinburgh, Oliver and Boyd.
- FISHER, R. A. 1935. The Logic of Inductive Inference. *Journal of the Royal Statistical Society*, 98, 39-54.
- FRESNO, C. & FERNÁNDEZ, E. A. 2013. RDAVIDWebService: a versatile R interface to DAVID. *Bioinformatics*, 29, 2810-2811.
- GARCÍA-CAMPOS, M. A., ESPINAL-ENRÍQUEZ, J. & HERNÁNDEZ-LEMUS, E. 2015. Pathway Analysis: State of the Art. *Frontiers in physiology*, 6, 383-383.
- GATTO, L., HANSEN, K. D., HOOPMANN, M. R., HERMJAKOB, H., KOHLBACHER, O. & BEYER, A. 2016. Testing and Validation of Computational Methods for Mass Spectrometry. *Journal of proteome research*, 15, 809-814.
- GHOLAMI, AMIN M., HAHNE, H., WU, Z., AUER, FLORIAN J., MENG, C., WILHELM, M. & KUSTER, B. 2013. Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Reports*, 4, 609-620.
- GILLET, L. C., NAVARRO, P., TATE, S., RÖST, H., SELEVSEK, N., REITER, L., BONNER, R. & AEBERSOLD, R. 2012. Targeted Data Extraction of the

- MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis\*. *Molecular & Cellular Proteomics*, 11, 0111.016717.
- GILLIS, J. & PAVLIDIS, P. 2013. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics*, 29, 476-82.
- GREENLAND, S., SENN, S. J., ROTHMAN, K. J., CARLIN, J. B., POOLE, C., GOODMAN, S. N. & ALTMAN, D. G. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31, 337-350.
- GUEDJ, M., ROBIN, S., CELISSE, A. & NUEL, G. 2009. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10, 84.
- GUPTA, N. & PEVZNER, P. A. 2009. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of proteome research*, 8, 4173-4181.
- GUPTA, N., TANNER, S., JAITLEY, N., ADKINS, J. N., LIPTON, M., EDWARDS, R., ROMINE, M., OSTERMAN, A., BAFNA, V., SMITH, R. D. & PEVZNER, P. A. 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res*, 17, 1362-77.
- GUYATT, G., JAESCHKE, R., HEDDLE, N., COOK, D., SHANNON, H. & WALTER, S. 1995. Basic statistics for clinicians: 1. Hypothesis testing. *Cmaj*, 152, 27-32.
- GYGI, S. P., RIST, B., GERBER, S. A., TURECEK, F., GELB, M. H. & AEBERSOLD, R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, 17, 994-9.
- HANASH, S. & CELIS, J. E. 2002. The Human Proteome Organization: a mission to advance proteome knowledge. *Mol Cell Proteomics*, 1, 413-4.
- HARTWELL, L. H., HOPFIELD, J. J., LEIBLER, S. & MURRAY, A. W. 1999. From molecular to modular cell biology. *Nature*, 402, C47-C52.
- HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.
- HOMAN, M. D. & GELMAN, A. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15, 1593-1623.
- HU, Q., NOLL, R. J., LI, H., MAKAROV, A., HARDMAN, M. & GRAHAM COOKS, R. 2005. The Orbitrap: a new mass spectrometer. *J Mass Spectrom*, 40, 430-43.
- HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37, 1-13.
- HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.
- HUBER, W., VON HEYDEBRECK, A., SÜLTMANN, H., POUSTKA, A. & VINGRON, M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18, S96-S104.

- HUNT, D. F., YATES, J. R., SHABANOWITZ, J., WINSTON, S. & HAUER, C. R. 1986. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, 83, 6233.
- JEDRYCHOWSKI, M. P., HUTTLIN, E. L., HAAS, W., SOWA, M. E., RAD, R. & GYGI, S. P. 2011. Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Molecular & cellular proteomics : MCP*, 10, M111.009910-M111.009910.
- JELLIFFE, R. W., SCHUMITZKY, A., BAYARD, D., FU, X. & NEELY, M. 2015. Describing Assay Precision-Reciprocal of Variance Is Correct, Not CV Percent: Its Use Should Significantly Improve Laboratory Performance. *Therapeutic drug monitoring*, 37, 389-394.
- JEONG, H., MASON, S. P., BARABASI, A. L. & OLTVAI, Z. N. 2001. Lethality and centrality in protein networks. *Nature*, 411, 41-2.
- JIANG, J. J. & CONRATH, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. ROCLING/IJCLCLP, 1997.
- JIN, J., SON, M., KIM, H., KIM, H., KONG, S. H., KIM, H. K., KIM, Y. & HAN, D. 2018. Comparative proteomic analysis of human malignant ascitic fluids for the development of gastric cancer biomarkers. *Clin Biochem*, 56, 55-61.
- JIN, L., BI, Y., HU, C., QU, J., SHEN, S., WANG, X. & TIAN, Y. 2021. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11, 1760.
- JOHNSON, D., BOYES, B., FIELDS, T., KOPKIN, R. & ORLANDO, R. 2013. Optimization of data-dependent acquisition parameters for coupling high-speed separations with LC-MS/MS for protein identifications. *Journal of biomolecular techniques : JBT*, 24, 62-72.
- JONES, A. R. 2017. Chapter 5 Protein Inference and Grouping. *Proteome Informatics*. The Royal Society of Chemistry.
- JORGENSEN, J. W. & LUKACS, K. D. 1981. Zone electrophoresis in open-tubular glass capillaries. *Analytical Chemistry*, 53, 1298-1302.
- KARAS, M. & HILLENKAMP, F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60, 2299-2301.
- KARPIEVITCH, Y. V., DABNEY, A. R. & SMITH, R. D. 2012. Normalization and missing value imputation for label-free LC-MS analysis. *BMC bioinformatics*, 13 Suppl 16, S5-S5.
- KELLER, A., NESVIZHSHKII, A. I., KOLKER, E. & AEBERSOLD, R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*, 74, 5383-92.
- KENNEDY-SHAFFER, L. 2019. Before  $p < 0.05$  to Beyond  $p < 0.05$ : Using History to Contextualize p-Values and Significance Testing. *The American statistician*, 73, 82-90.
- KHATRI, P. & DRĂGHICI, S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21, 3587-3595.
- KICMAN, A. T., PARKIN, M. C. & ILES, R. K. 2007. An introduction to mass spectrometry based proteomics—Detection and characterization of gonadotropins and related molecules. *Molecular and Cellular Endocrinology*, 260-262, 212-227.

- KIM, S., GUPTA, N. & PEVZNER, P. A. 2008. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*, 7, 3354-63.
- KIM, S. & PEVZNER, P. A. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5, 5277.
- KLOSE, J. 1975. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. *Humangenetik*, 26, 231-243.
- KOOPMANS, F., CORNELISSE, L. N., HESKES, T. & DIJKSTRA, T. M. H. 2014. Empirical Bayesian Random Censoring Threshold Model Improves Detection of Differentially Abundant Proteins. *Journal of Proteome Research*, 13, 3871-3880.
- KRUSCHKE, J. K. 2013. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*, 142, 573-603.
- KRZYWINSKI, M. & ALTMAN, N. 2013. Importance of being uncertain. *Nature Methods*, 10, 809-810.
- KUHAREV, J., NAVARRO, P., DISTLER, U., JAHN, O. & TENZER, S. 2015. In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *PROTEOMICS*, 15, 3140-3151.
- KURTZER, G. M., SOCHAT, V. & BAUER, M. W. 2017. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12, e0177459.
- KUSTATSCHER, G., GRABOWSKI, P., SCHRADER, T. A., PASSMORE, J. B., SCHRADER, M. & RAPPSILBER, J. 2019. The human proteome co-regulation map reveals functional relationships between proteins. *bioRxiv*, 582247.
- LANGLEY, S. R. & MAYR, M. 2015. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *Journal of Proteomics*, 129, 83-92.
- LEE, J., SUNG, W. & CHOI, J.-H. 2015. Metamodel for Efficient Estimation of Capacity-Fade Uncertainty in Li-Ion Batteries for Electric Vehicles. *Energies*, 8, 5538-5554.
- LI, X., LIANG, L., DE VIVO, I., TANG, J. Y. & HAN, J. 2016. Pathway analysis of expression-related SNPs on genome-wide association study of basal cell carcinoma. *Oncotarget*, 7, 36885-36895.
- LIN, D. An information-theoretic definition of similarity. *Icml*, 1998. 296-304.
- LIPSHUTZ, R. J., FODOR, S. P. A., GINGERAS, T. R. & LOCKHART, D. J. 1999. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21, 20-24.
- MAES, E., KELCHTERMANS, P., BITTREMIEUX, W., DE GRAVE, K., DEGROEVE, S., HOOYBERGHS, J., MERTENS, I., BAGGERMAN, G., RAMON, J., LAUKENS, K., MARTENS, L. & VALKENBORG, D. 2016. Designing biomedical proteomics experiments: state-of-the-art and future perspectives. *Expert Rev Proteomics*, 13, 495-511.
- MAHONEY, D. W., THERNEAU, T. M., HEPPELMANN, C. J., HIGGINS, L., BENSON, L. M., ZENKA, R. M., JAGTAP, P., NELSESTUEN, G. L., BERGEN, H. R. & OBERG, A. L. 2011. Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides. *Journal of proteome research*, 10, 4325-4333.
- MAMYRIN, B. A., KARATAEV, V. I., SHMIKK, D. V. & ZAGULIN, V. A. 1973. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer

- with high resolution. *Journal of Experimental and Theoretical Physics*, 37, 45.
- MARCOTTE, R., SAYAD, A., BROWN, K. R., SANCHEZ-GARCIA, F., REIMAND, J., HAIDER, M., VIRTANEN, C., BRADNER, J. E., BADER, G. D., MILLS, G. B., PE'ER, D., MOFFAT, J. & NEEL, B. G. 2016. Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell*, 164, 293-309.
- MARGOLIN, A. A., ONG, S.-E., SCHENONE, M., GOULD, R., SCHREIBER, S. L., CARR, S. A. & GOLUB, T. R. 2009. Empirical Bayes analysis of quantitative proteomics experiments. *PLoS one*, 4, e7454-e7454.
- MARIMUTHU, A., SUBBANNAYYA, Y., SAHASRABUDDHE, N. A., BALAKRISHNAN, L., SYED, N., SEKHAR, N. R., KATTE, T. V., PINTO, S. M., SRIKANTH, S. M., KUMAR, P., PAWAR, H., KASHYAP, M. K., MAHARUDRAIAH, J., ASHKTORAB, H., SMOOT, D. T., RAMASWAMY, G., KUMAR, R. V., CHENG, Y., MELTZER, S. J., ROA, J. C., CHAERKADY, R., PRASAD, T. S. K., HARSHA, H. C., CHATTERJEE, A. & PANDEY, A. 2013. SILAC-based quantitative proteomic analysis of gastric cancer secretome. *PROTEOMICS – Clinical Applications*, 7, 355-366.
- MARSHALL, A. G., HENDRICKSON, C. L. & JACKSON, G. S. 1998. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev*, 17, 1-35.
- MARTIN-BELMONTE, F. & PEREZ-MORENO, M. 2012. Epithelial cell polarity, stem cells and cancer. *Nature Reviews Cancer*, 12, 23-38.
- MCCARTHUR, G. A., CHAPMAN, P. B., ROBERT, C., LARKIN, J., HAANEN, J. B., DUMMER, R., RIBAS, A., HOGG, D., HAMID, O., ASCIERTO, P. A., GARBE, C., TESTORI, A., MAIO, M., LORIGAN, P., LEBBÉ, C., JOUARY, T., SCHADENDORF, D., O'DAY, S. J., KIRKWOOD, J. M., EGGERMONT, A. M., DRÉNO, B., SOSMAN, J. A., FLAHERTY, K. T., YIN, M., CARO, I., CHENG, S., TRUNZER, K. & HAUSCHILD, A. 2014. Safety and efficacy of vemurafenib in BRAF(V600E) and BRAF(V600K) mutation-positive melanoma (BRIM-3): extended follow-up of a phase 3, randomised, open-label study. *Lancet Oncol*, 15, 323-32.
- MECHAM, B. H., NELSON, P. S. & STOREY, J. D. 2010. Supervised normalization of microarrays. *Bioinformatics (Oxford, England)*, 26, 1308-1315.
- MELBY, J. A., ROBERTS, D. S., LARSON, E. J., BROWN, K. A., BAYNE, E. F., JIN, S. & GE, Y. 2021. Novel Strategies to Address the Challenges in Top-Down Proteomics. *Journal of the American Society for Mass Spectrometry*, 32, 1278-1294.
- MERKEL, D. 2014. Docker: lightweight Linux containers for consistent development and deployment. *Linux J*, 2014, Article 2.
- MI, H., POUDEL, S., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*, 44, D336-42.
- MILAC, T. I., RANDOLPH, T. W. & WANG, P. 2012. Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies. *Statistics and its interface*, 5, 75-87.
- MILADINOVIĆ, S. M., KOZHINOV, A. N., TSYBIN, O. Y. & TSYBIN, Y. O. 2012. Sidebands in Fourier transform ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry*, 325-327, 10-18.



- MILLIKIN, R. J., SHORTREED, M. R., SCALF, M. & SMITH, L. M. 2020. A Bayesian Null Interval Hypothesis Test Controls False Discovery Rates and Improves Sensitivity in Label-Free Quantitative Proteomics. *Journal of Proteome Research*, 19, 1975-1981.
- MITCHELL WELLS, J. & MCLUCKEY, S. A. 2005. Collision-Induced Dissociation (CID) of Peptides and Proteins. *Methods in Enzymology*. Academic Press.
- MORGAN, M. 2019. BiocManager: Access the Bioconductor Project Package Repository.
- MORGAN, M. S., L. 2021. AnnotationHub: Client to access AnnotationHub resources.
- NAVARRO, P., KUHAREV, J., GILLET, L. C., BERNHARDT, O. M., MACLEAN, B., RÖST, H. L., TATE, S. A., TSOU, C.-C., REITER, L., DISTLER, U., ROSENBERGER, G., PEREZ-RIVEROL, Y., NESVIZHSHKII, A. I., AEBERSOLD, R. & TENZER, S. 2016. A multicenter study benchmarks software tools for label-free proteome quantification. *Nature Biotechnology*, 34, 1130-1136.
- NEAL, R. M. 1996. *Bayesian Learning for Neural Networks*, Springer-Verlag.
- NESVIZHSHKII, A. I. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, 73, 2092-2123.
- NESVIZHSHKII, A. I. 2014. Proteogenomics: concepts, applications and computational strategies. *Nature methods*, 11, 1114-1125.
- NESVIZHSHKII, A. I. & AEBERSOLD, R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, 4, 1419-40.
- NONLINEARDYNAMICS. *Progenesis QI for proteomics User Guide* [Online]. Available: [http://storage.nonlinear.com/webfiles/progenesis/qi-for-proteomics/v4.2/user-guide/Progenesis\\_QI.p\\_DDA\\_User\\_Guide\\_4.2.pdf](http://storage.nonlinear.com/webfiles/progenesis/qi-for-proteomics/v4.2/user-guide/Progenesis_QI.p_DDA_User_Guide_4.2.pdf) [Accessed].
- O'FARRELL, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250, 4007-4021.
- O'ROURKE, M. B., TOWN, S. E. L., DALLA, P. V., BICKNELL, F., KOH BELIC, N., VIOLI, J. P., STEELE, J. R. & PADULA, M. P. 2019. What is Normalization? The Strategies Employed in Top-Down and Bottom-Up Proteome Analysis Workflows. *Proteomes*, 7, 29.
- OBERG, A. L. & VITEK, O. 2009. Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Journal of Proteome Research*, 8, 2144-2156.
- OLSEN, J. V., MACEK, B., LANGE, O., MAKAROV, A., HORNING, S. & MANN, M. 2007. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods*, 4, 709-12.
- ONG, S.-E., BLAGOEV, B., KRATCHMAROVA, I., KRISTENSEN, D. B., STEEN, H., PANDEY, A. & MANN, M. 2002. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics \*. *Molecular & Cellular Proteomics*, 1, 376-386.
- ONG, S. E. & MANN, M. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*, 1, 252-62.
- PARK, T., YI, S. G., KANG, S. H., LEE, S., LEE, Y. S. & SIMON, R. 2003. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4, 33.

- PAULOVICH, A. G., BILLHEIMER, D., HAM, A. J., VEGA-MONTOTO, L., RUDNICK, P. A., TABB, D. L., WANG, P., BLACKMAN, R. K., BUNK, D. M., CARDASIS, H. L., CLAUSER, K. R., KINSINGER, C. R., SCHILLING, B., TEGELER, T. J., VARIYATH, A. M., WANG, M., WHITEAKER, J. R., ZIMMERMAN, L. J., FENYO, D., CARR, S. A., FISHER, S. J., GIBSON, B. W., MESRI, M., NEUBERT, T. A., REGNIER, F. E., RODRIGUEZ, H., SPIEGELMAN, C., STEIN, S. E., TEMPST, P. & LIEBLER, D. C. 2010. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics*, 9, 242-54.
- PENG, J., ELIAS, J. E., THOREEN, C. C., LICKLIDER, L. J. & GYGI, S. P. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, 2, 43-50.
- PERKINS, D. N., PAPPIN, D. J. C., CREASY, D. M. & COTTRELL, J. S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*, 20, 3551-3567.
- PETERS, B., BRENNER, S. E., WANG, E., SLONIM, D. & KANN, M. G. 2018. Putting benchmarks in their rightful place: The heart of computational biology. *PLoS Comput Biol*, 14, e1006494.
- PULVERER, B. 2012. Significant statistics. *EMBO reports*, 13, 280-280.
- PURSIHEIMO, A., VEHMAS, A. P., AFZAL, S., SUOMI, T., CHAND, T., STRAUSS, L., POUTANEN, M., ROKKA, A., CORTHALS, G. L. & ELO, L. L. 2015. Optimization of Statistical Methods Impact on Quantitative Proteomics Data. *J Proteome Res*, 14, 4118-26.
- R CORE TEAM 2020. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- RAMUS, C., HOVASSE, A., MARCELLIN, M., HESSE, A. M., MOUTON-BARBOSA, E., BOUYSSIE, D., VACA, S., CARAPITO, C., CHAOUI, K., BRULEY, C., GARIN, J., CIANFERANI, S., FERRO, M., DORSSAELER, A. V., BURLET-SCHILTZ, O., SCHAEFFER, C., COUTE, Y. & GONZALEZ DE PEREDO, A. 2016. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. *Data Brief*, 6, 286-94.
- RAPPSILBER, J. & MANN, M. 2002. What does it mean to identify a protein in proteomics? *Trends Biochem Sci*, 27, 74-8.
- REITER, L., CLAASSEN, M., SCHRIMPF, S. P., JOVANOVIĆ, M., SCHMIDT, A., BUHMANN, J. M., HENGARTNER, M. O. & AEBERSOLD, R. 2009. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry\*. *Molecular & Cellular Proteomics*, 8, 2405-2417.
- RESNIK, P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Int. Res.*, 11, 95-130.
- RÉVÉSZ, Á., MILLEY, M. G., NAGY, K., SZABÓ, D., KALLÓ, G., CSÓSZ, É., VÉKEY, K. & DRAHOS, L. 2021. Tailoring to Search Engines: Bottom-Up Proteomics with Collision Energies Optimized for Identification Confidence. *Journal of Proteome Research*, 20, 474-484.

- RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43, e47-e47.
- ROBIN, S., BAR-HEN, A., DAUDIN, J.-J. & PIERRE, L. 2007. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51, 5483-5493.
- ROCKE, D. M. & DURBIN, B. 2001. A Model for Measurement Error for Gene Expression Arrays. *Journal of Computational Biology*, 8, 557-569.
- SANTRA, T. & DELATOLA, E. I. 2016. A Bayesian algorithm for detecting differentially expressed proteins and its application in breast cancer research. *Scientific Reports*, 6, 30159.
- SAVARYN, J. P., TOBY, T. K. & KELLEHER, N. L. 2016. A researcher's guide to mass spectrometry-based proteomics. *PROTEOMICS*, 16, 2435-2443.
- SCHLICKER, A., DOMINGUES, F. S., RAHNENFÜHRER, J. & LENGAUER, T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7, 302.
- SHALIT, T., ELINGER, D., SAVIDOR, A., GABASHVILI, A. & LEVIN, Y. 2015. MS1-based label-free proteomics using a quadrupole orbitrap mass spectrometer. *J Proteome Res*, 14, 1979-86.
- SILVA, J. C., GORENSTEIN, M. V., LI, G. Z., VISSERS, J. P. & GEROMANOS, S. J. 2006. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics*, 5, 144-56.
- SILVERMAN, B. W. 1986. *Density estimation for statistics and data analysis*, London; New York, Chapman and Hall.
- SMYTH, G. K. 2005. limma: Linear Models for Microarray Data. In: GENTLEMAN, R., CAREY, V. J., HUBER, W., IRIZARRY, R. A. & DUDOIT, S. (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York, NY: Springer New York.
- STEWART, P. A., FANG, B., SLEBOS, R. J., ZHANG, G., BORNE, A. L., FELLOWS, K., TEER, J. K., CHEN, Y. A., WELSH, E., ESCHRICH, S. A., HAURA, E. B. & KOOMEN, J. M. 2017. Relative protein quantification and accessible biology in lung tumor proteomes from four LC-MS/MS discovery platforms. *Proteomics*, 17.
- STEWART, P. A., PARAPATICS, K., WELSH, E. A., MÜLLER, A. C., CAO, H., FANG, B., KOOMEN, J. M., ESCHRICH, S. A., BENNETT, K. L. & HAURA, E. B. 2015. A Pilot Proteogenomic Study with Data Integration Identifies MCT1 and GLUT1 as Prognostic Markers in Lung Adenocarcinoma. *PLoS One*, 10, e0142162.
- STICKER, A., GOEMINNE, L., MARTENS, L. & CLEMENT, L. 2020. Robust Summarization and Inference in Proteome-wide Label-free Quantification. *Molecular & Cellular Proteomics*, 19, 1209-1219.
- STOREY, J. D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479-498.
- SULLIVAN, G. M. & FEINN, R. 2012. Using Effect Size-or Why the P Value Is Not Enough. *Journal of graduate medical education*, 4, 279-282.
- SUOMI, T. & ELO, L. L. 2017. Enhanced differential expression statistics for data-independent acquisition proteomics. *Sci Rep*, 7, 5869.
- SYKA, J. E. P., COON, J. J., SCHROEDER, M. J., SHABANOWITZ, J. & HUNT, D. F. 2004. Peptide and protein sequence analysis by electron transfer

- dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 9528.
- SYKES, M. T. & WILLIAMSON, J. R. 2008. Envelope: interactive software for modeling and fitting complex isotope distributions. *BMC Bioinformatics*, 9, 446.
- SYLABS.IO. 2021. *SingularityCE User Guide* [Online]. Available: <https://sylabs.io/guides/3.8/user-guide/> [Accessed 06.10.21].
- TABB, D. L., WANG, X., CARR, S. A., CLAUSER, K. R., MERTINS, P., CHAMBERS, M. C., HOLMAN, J. D., WANG, J., ZHANG, B., ZIMMERMAN, L. J., CHEN, X., GUNAWARDENA, H. P., DAVIES, S. R., ELLIS, M. J. C., LI, S., TOWNSEND, R. R., BOJA, E. S., KETCHUM, K. A., KINSINGER, C. R., MESRI, M., RODRIGUEZ, H., LIU, T., KIM, S., MCDERMOTT, J. E., PAYNE, S. H., PETYUK, V. A., RODLAND, K. D., SMITH, R. D., YANG, F., CHAN, D. W., ZHANG, B., ZHANG, H., ZHANG, Z., ZHOU, J.-Y. & LIEBLER, D. C. 2016. Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *Journal of proteome research*, 15, 691-706.
- TANAKA, K., WAKI, H., IDO, Y., AKITA, S., YOSHIDA, Y., YOSHIDA, T. & MATSUO, T. 1988. Protein and polymer analyses up to  $m/z$  100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2, 151.
- TANG, J., FU, J., WANG, Y., LI, B., LI, Y., YANG, Q., CUI, X., HONG, J., LI, X., CHEN, Y., XUE, W. & ZHU, F. 2019a. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Briefings in Bioinformatics*.
- TANG, J., FU, J., WANG, Y., LUO, Y., YANG, Q., LI, B., TU, G., HONG, J., CUI, X., CHEN, Y., YAO, L., XUE, W. & ZHU, F. 2019b. Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains. *Molecular & Cellular Proteomics*, 18, 1683.
- THARMALINGAM, S. & HAMPSON, D. R. 2016. The Calcium-Sensing Receptor and Integrins in Cellular Differentiation and Migration. *Frontiers in Physiology*, 7.
- THE, M., EDFORS, F., PEREZ-RIVEROL, Y., PAYNE, S. H., HOOPMANN, M. R., PALMBLAD, M., FORSSTRÖM, B. & KÄLL, L. 2018. A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms. *J Proteome Res*, 17, 1879-1886.
- THIESE, M. S., RONNA, B. & OTT, U. 2016. P value interpretations and considerations. *Journal of thoracic disease*, 8, E928-E931.
- THOMPSON, A., SCHÄFER, J., KUHN, K., KIENLE, S., SCHWARZ, J., SCHMIDT, G., NEUMANN, T. & HAMON, C. 2003. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry*, 75, 1895-1904.
- TIMMONS, J. A., SZKOP, K. J. & GALLAGHER, I. J. 2015. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biology*, 16, 186.
- TING, L., RAD, R., GYGI, S. P. & HAAS, W. 2011. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods*, 8, 937-940.

- TOKAREVA, A. O., CHAGOVETS, V. V., KONONIKHIN, A. S., STARODUBTSEVA, N. L., NIKOLAEV, E. N. & FRANKEVICH, V. E. 2021. Normalization methods for reducing interbatch effect without quality control samples in liquid chromatography-mass spectrometry-based studies. *Analytical and Bioanalytical Chemistry*, 413, 3479-3486.
- TOMCZAK, A., MORTENSEN, J. M., WINNENBURG, R., LIU, C., ALESSI, D. T., SWAMY, V., VALLANIA, F., LOFGREN, S., HAYNES, W., SHAH, N. H., MUSEN, M. A. & KHATRI, P. 2018. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Scientific reports*, 8, 5115-5115.
- TRAN, N. H., QIAO, R., XIN, L., CHEN, X., LIU, C., ZHANG, X., SHAN, B., GHODSI, A. & LI, M. 2019. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods*, 16, 63-66.
- TUKEY, J. W. 1977. *Exploratory Data Analysis*, Reading Massachusetts, Addison-Wesley Publishing Company
- TUSHER, V. G., TIBSHIRANI, R. & CHU, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5116-5121.
- UHLÉN, M., FAGERBERG, L., HALLSTRÖM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, Å., KAMPF, C., SJÖSTEDT, E., ASPLUND, A., OLSSON, I., EDLUND, K., LUNDBERG, E., NAVANI, S., SZIGYARTO, C. A.-K., ODEBERG, J., DJUREINOVIC, D., TAKANEN, J. O., HOBER, S., ALM, T., EDQVIST, P.-H., BERLING, H., TEGEL, H., MULDER, J., ROCKBERG, J., NILSSON, P., SCHWENK, J. M., HAMSTEN, M., VON FEILITZEN, K., FORSBERG, M., PERSSON, L., JOHANSSON, F., ZWAHLEN, M., VON HEIJNE, G., NIELSEN, J. & PONTÉN, F. 2015. Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347, 1260419.
- VÄLIKANGAS, T., SUOMI, T. & ELO, L. L. 2016. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics*, 19, 1-11.
- VÄLIKANGAS, T., SUOMI, T. & ELO, L. L. 2017. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Briefings in Bioinformatics*, 19, 1344-1355.
- VÄLIKANGAS, T., SUOMI, T. & ELO, L. L. 2018. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in bioinformatics*, 19, 1-11.
- VAUX, D. L., FIDLER, F. & CUMMING, G. 2012. Replicates and repeats--what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO reports*, 13, 291-296.
- VENABLE, J. D., DONG, M.-Q., WOHLSCHEGEL, J., DILLIN, A. & YATES, J. R. 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1, 39-45.
- VENABLES, B. & RIPLEY, B. 2002. *Modern Applied Statistics With S*. Fourth Edition ed.: Springer, New York.
- VIZCAÍNO, J. A., DEUTSCH, E. W., WANG, R., CSORDAS, A., REISINGER, F., RÍOS, D., DIANES, J. A., SUN, Z., FARRAH, T., BANDEIRA, N., BINZ, P.-A.,

- XENARIOS, I., EISENACHER, M., MAYER, G., GATTO, L., CAMPOS, A., CHALKLEY, R. J., KRAUS, H.-J., ALBAR, J. P., MARTINEZ-BARTOLOMÉ, S., APWEILER, R., OMENN, G. S., MARTENS, L., JONES, A. R. & HERMJAKOB, H. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32, 223-226.
- VYAS, D., BALAKRISHNAN, A. & VYAS, A. 2015. The Value of the P Value. *American journal of robotic surgery*, 2, 53-56.
- WADI, L., MEYER, M., WEISER, J., STEIN, L. D. & REIMAND, J. 2016. Impact of outdated gene annotations on pathway enrichment analysis. *Nature methods*, 13, 705-706.
- WANG, J., LI, L., CHEN, T., MA, J., ZHU, Y., ZHUANG, J. & CHANG, C. 2017. In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values. *Sci Rep*, 7, 3367.
- WANG, J. Z., DU, Z., PAYATTAKOOL, R., YU, P. S. & CHEN, C.-F. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23, 1274-1281.
- WANG, S.-Y., KUO, C.-H. & TSENG, Y. J. 2013. Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Methods. *Analytical Chemistry*, 85, 1037-1046.
- WELCH, B. L. 1947. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARIANCES ARE INVOLVED. *Biometrika*, 34, 28-35.
- WIESE, S., REIDEGELD, K. A., MEYER, H. E. & WARSCHEID, B. 2007. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics*, 7, 340-50.
- WIŚNIEWSKI, J. R., OSTASIEWICZ, P., DUŚ, K., ZIELIŃSKA, D. F., GNAD, F. & MANN, M. 2012. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Molecular systems biology*, 8, 611-611.
- WOLFF, M. M. & STEPHENS, W. E. 1953. A Pulsed Mass Spectrometer with Time Dispersion. *Review of Scientific Instruments*, 24, 616-617.
- WU, T., HU, E., XU, S., CHEN, M., GUO, P., DAI, Z., FENG, T., ZHOU, L., TANG, W., ZHAN, L., FU, X., LIU, S., BO, X. & YU, G. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 100141.
- XU, F., WANG, L., DAI, X., FANG, X. & DING, C.-F. 2014. Resonance Activation and Collision-Induced-Dissociation of Ions Using Rectangular Wave Dipolar Potentials in a Digital Ion Trap Mass Spectrometer. *Journal of The American Society for Mass Spectrometry*, 25, 556-562.
- YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. & SPEED, T. P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30, e15-e15.
- YATES, J. R., 3RD 2019. Recent technical advances in proteomics. *F1000Research*, 8, F1000 Faculty Rev-351.

- YATES, J. R., PARK, S. K. R., DELAHUNTY, C. M., XU, T., SAVAS, J. N., COCIORVA, D. & CARVALHO, P. C. 2012. Toward objective evaluation of proteomic algorithms. *Nature Methods*, 9, 455-456.
- YU, G. 2020. Gene Ontology Semantic Similarity Analysis Using GOSemSim. In: KIDDER, B. L. (ed.) *Stem Cell Transcriptional Networks: Methods and Protocols*. New York, NY: Springer US.
- YU, G. & HE, Q.-Y. 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12, 477-479.
- YU, G., LI, F., QIN, Y., BO, X., WU, Y. & WANG, S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26, 976-978.
- YU, G., WANG, L. G., HAN, Y. & HE, Q. Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, 16, 284-7.
- ZEGER, S. L. & KARIM, M. R. 1991. Generalized Linear Models with Random Effects; a Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86, 79-86.
- ZHANG, G., FENYÖ, D. & NEUBERT, T. A. 2009. Evaluation of the variation in sample preparation for comparative proteomics using stable isotope labeling by amino acids in cell culture. *Journal of proteome research*, 8, 1285-1292.
- ZHOU, Y., ZHOU, B., PACHE, L., CHANG, M., KHODABAKHSHI, A. H., TANASEICHUK, O., BENNER, C. & CHANDA, S. K. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications*, 10, 1523-1523.
- ZILA, N., BILECK, A., MUQAKU, B., JANKER, L., EICHHOFF, O. M., CHENG, P. F., DUMMER, R., LEVESQUE, M. P., GERNER, C. & PAULITSCHKE, V. 2018. Proteomics-based insights into mitogen-activated protein kinase inhibitor resistance of cerebral melanoma metastases. *Clin Proteomics*, 15, 13.
- ZIMMERMAN, D. W. 2004. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57, 173-181.
- ZIMMERMAN, D. W. & ZUMBO, B. D. 1993. Rank Transformations and the Power of the Student T Test and Welch T' Test for Non-Normal Populations with Unequal Variances. *Canadian Journal of Experimental Psychology*, 47, 523-539.
- ZUBAREV, R. A. & MAKAROV, A. 2013. Orbitrap Mass Spectrometry. *Analytical Chemistry*, 85, 5288-5296.