Quantitative Proteomics Strategies To Explore The *Saccharomyces cerevisiae* Proteome

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

By

Yvonne Esther Woolerton

April 2014

Acknowledgments

I would like to take this opportunity to express my appreciation and gratitude to all at PFG for their support and guidance over the duration of my PhD, in particular to Lynne, Deb and Phil for their unending knowledge. I would like to thank my supervisor, Rob Beynon for his support in getting me to submission stage of my PhD.

Thank you to all of my friends who have given me unrelenting support throughout the past four years. Particular thanks got to Jenny and Vicki, for always lending a sympathetic ear, moral support and encouragement to get me through the difficult stages in the course of my research.

None of this research would have been possible without the generous funding received so thanks must go to BBSRC and Waters for their funding for my PhD.

Special thanks to my family, in particular my Mum, whose guidance, listening and understanding has proven invaluable as always. Thanks also to Mitzi, whose company has made the writing process so much more pleasant.

Abstract

Quantitative proteomics aims at not just identifying, but accurately quantifying the cellular proteome, and while technological advances towards accurate and reliable quantification of proteins is advancing, this alone does not provide an accurate picture of a proteins role within a cell. There is a far greater level of functionality in the cellular environment than there are protein coding genes in the genome, owing partly to the organisation of individual proteins into larger assemblies. A single protein can form interactions with, potentially, a large number of other proteins, leading to a variety of different protein complexes, and subunits can move, break apart or combine depending on cellular conditions. This complex organisation is despite the normal proteomics strategies employing a destructive process, breaking protein structure down to the peptide level. Further difficulties in mapping the cellular proteome arise from the differential expression level of proteins, which in *S.cerevisiae* can span up to 5 orders of magnitude. This poses problems for the quantification of less abundant proteins in the cell, which can be masked by the more concentrated proteins. An attempt is made within this thesis to use quantitative proteomic techniques to build a picture of the *S.cerevisiae* cellular proteome.

For the analysis of *S.cerevisiae* protein complexes ion exchange chromatography has been used to separate the cellular proteome into discrete fractions, each containing a different array of protein complexes. The aim here was to analyse the individual subunits of these complexes by LC-MS, with the use of label free quantification strategies. This enables the high throughput identification and quantification of 1800 proteins along with their potential interaction partners. However, for some of the complexes presented here the accuracy of the label free quantification is called into question, as complex subunits known to be equimolar are identified at different concentrations. In order to assess the accuracy of the label free data QconCATs were also designed to analyse the subunits of some complexes by label mediated quantification.

In addition, an attempt is made to access proteins from the entire dynamic range of the cellular proteome using equaliser bead technology. This method uses a library of hexapeptide ligands bound to porous beads to bind, theoretically, every protein present in the sample to equal amounts. The beads are used here to bring up the less abundant proteins in the sample, while simultaneously reducing the amount of the abundant proteins. While this goal is achieved, it is also evident that certain proteins are able to bind the beads to a much larger extent than others, so rather than reducing the dynamic range of proteins identified, there is more of a shift in the dynamics, with previously mid-range proteins becoming highly abundant in the data presented here.

Contents

List of figures

List of tables

List of supplementary material

Chapter 3

1. Anion exchange chromatography G2 PLGS results
2. Mixed bed chromatography G2 replicate 1
3. Mixed bed chromatography G2 replicate 2
4. Mixed bed chromatography G2 replicate 3
5. Mixed bed chromatography G2 replicate 4
6. Mixed bed chromatography G2 replicate 5
7. Mixed bed chromatography QEX replicate 1
8. Mixed bed chromatography QEX replicate 2
9. Mixed bed chromatography QEX replicate 3
10. Mixed bed chromatography QEX replicate 4
11. Mixed bed chromatography QEX replicate 5
12. Example complex replicates 1- 5

Chapter 4

1. 20S proteasome transitions
2. eIF2 transitions
3. YEW1 eIF2 transition list
4. YEW3 20S transition list

Chapter 5

1. Three replicates of SM and CPLL treated material
2. Loading experiment 190813
3. Ionic experiments 171113
4. GO results Rep 1 treated
5. GO results Rep 1 SM
6. GO results Rep 2 treated
7. GO results Rep 2 SM
8. GO results Rep 3 treated
9. GO results Rep 3 SM

Found on the disc attached to the back of this thesis.

List of abbreviations

AQUA	Absolute Quantification
BSA	Bovine Serum Albumin
CPLL	Combinatorial hexapeptide ligand library
DC	Direct current
DTT	Dithiothreitol
ESI	Electrospray Ionisation
FDR	False discovery rate
HPLC	High Performance Liquid Chromatography
ICAT	Isotope-Coded Affinity Tag
IPTG	isopropyl β-D-1-thiogalactopyranoside
iTRAQ	Isobaric Tag for Relative and Absolute Quantification
LC	Liquid Chromatography
MALDI	Matrix Assisted Laser Desorption Ionisation
MS	Mass Spectrometry
m/z	Mass to charge ratio
OD	Optical Density
ORF	Open Reading Frame
PLGS	ProteinLynx Global Server
PSAQ	Protein Standard Absolute Quantification
PTM	Post-translational modification
QconCAT	Concatemer of standard Q-peptides
Q-TOF	Quadrupole-Time of Flight
RF	Radio frequancy
RT	Retention Time
SDS-PAGE	Sodium Dodecyl Sulphate - Polyacrylamide Gel Electrophoresis
SGD	Saccharomyces Genome Database
SILAC	Stable Isotope Labelling by Amino Acids in Cell Culture
SRM	Selected Reaction Monitoring
TAP-tag	Tandem Affinity Purification tag
TOF	Time of Flight

1. Introduction

1.1 Proteomics

The term proteomics was coined almost 20 years ago to describe the study of the global profile of protein expression (Wilkins *et al.*, 1996). Proteins are biological molecules that are responsible for most of a cells functionality, and the proteome of an organism is defined as all of the proteins expressed by the genome. The cell is a complex environment consisting of a wide dynamic range of proteins which are expressed and degraded differentially depending on time and cellular conditions. The field of proteomics encompasses the complete study of proteins, which can include the function, sequence, identification, localization and quantification of all of a given proteome.

The complex nature of the proteome derives from a number of layers of organisation in the cellular environment. The genome encodes all of the potential proteins in a given organism, but there are far more functional proteins in a cell than there are protein coding genes in the genome (Fields, 2001). This is due to a number of cellular processes which mean that a single gene does not simply encode a single protein, a process that is controlled at three levels, the transcription of genes into mRNA, the translation of these mRNA transcripts into proteins, and in post translational modification. At the transcriptional level, frameshifting can cause a change in the set of triplet codons that are transcribed (Belew *et al.*, 2011; Advani *et al.*, 2013). At the translational level alternative splicing can lead to the creation of more than one protein from an mRNA transcript of one gene (Juneau *et al.*, 2009; Rossler & Marschalek, 2013). In addition, hundreds of types of post translational modifications (PTMs), such as ubiquitination, glycosylation, acetylation or phosphorylation, have been documented, and these can alter the structure and consequently control the function of a diverse range of proteins (reviewed in Emre & Berger, 2006; Molina *et al.*, 2010; Oliveira & Sauer, 2012). The combination of these effects means that in some biological samples there is a large array of proteins to study. It has been estimated that in some systems the number of proteins can be up to an order of magnitude greater than the available genes

(Fields, 2001). The concept of protein species was introduced to refer to a final, functional protein following all levels of modification, describing an individual protein from a family of proteins encoded by one gene (Jungblut *et al.*, 1996; Jungblut *et al.*, 2008).

The diverse nature of the proteome creates a number of challenges in the analysis of such a diverse array of proteins, but in the post-genomic era the direct study of protein function is becoming increasingly important, and has wide ranging applications in enhancing the understanding of cellular processes behind disease, in diagnostics, screening and monitoring, or in the discovery of new drug targets. Proteomic techniques have proved instrumental in the study of protein function and cellular organisation, and led to the discovery of novel biomarkers of diseases such as cancer (Goncalves *et al.*, 2008), heart disease (Fu & Van Eyk, 2006), or even some psychiatric disorders (Taurines *et al.*, 2010).

1.2 *Saccharomyces cerevisiae* as a model organism

Saccharomyces cerevisiae is a species of budding yeast that has been extensively studied as a model organism for a number of years (Botstein *et al.*, 1997; Botstein & Fink, 2011; Castrillo & Oliver, 2011). The usefulness of yeast in systems biology stems from a number of factors, it is a relatively simple unicellular organism, is quick and easy to grow, and as a eukaryotic organism the mechanism of cell growth, gene expression, translation, metabolism and signal transduction are all under a similar level of control to higher eukaryotes, which means yeast has many homologous genes to mammalian cells, which has proved instrumental in the study of a number of human diseases (reviewed in (Botstein *et al.*, 1997; Smith & Snyder, 2006).

The completion of the *S. cerevisiae* genome project in 1996 made it the first eukaryotic organism to have its whole genome sequenced (Goffeau *et al.*, 1996). Knowledge of the gene sequence creates a picture of the potential protein complement of the organism, however due to the level of complexity in the proteome, the gene sequence alone does not provide adequate information on the function of each gene product. The next level of analysis, therefore, in the field of

transcriptomics aims to analyse the gene transcription occurring within a cell, by identifying the mRNA transcripts present. However, the amount of mRNA present in a cell does not necessarily directly correlate with the amount of protein being made (Gygi *et al.*, 1999b), due to a combination of effects, including the rapid turnover of cellular mRNA (Halbeisen & Gerber, 2009), and the potential for a single transcript to encode multiple proteins.

The existence of a complete catalogue of genes for an organism has led numerous researchers to the investigation of the function of the products of these genes. Therefore, studying the vast array of cellular proteins becomes necessary. Studying the protein product as opposed to the gene provides direct information on the cellular function of the protein, and the integration of genomic, transcriptomic and proteomic strategies can provide a complete picture of a proteins life cycle within a cell. A number of large scale studies have identified *S. cerevisiae* gene products, and it is now possible to gain wide coverage of the yeast proteome in some experiments (Ghaemmaghami *et al.*, 2003; Huh *et al.*, 2003; Nagaraj *et al.*, 2012; Hebert *et al.*, 2014). The wide range of genomics, transcriptomics and proteomics studies performed to date have been amalgamated into databases such as the Saccharomyces Genome Database (SGD, www.yeastgenome.org) detailing the functional role of every gene which has been analysed. Currently, the SGD lists 5070 protein encoding genes as experimentally confirmed and 750 as uncharacterised, and it is possible to identify the majority of these, with some experiments reaching over 4000 proteins (de Godoy *et al.*, 2008; Webb *et al.*, 2013; Hebert *et al.*, 2014) .

1.3 Challenges in the field of proteomics

The design of a proteomics study faces numerous analytical difficulties due to the complexity of the proteome. One of the major analytical problems faced in the application of proteomics experiments is in the sheer volume of information available in any cellular environment, derived in part from the diversity and dynamics of gene expression. The rate and extent of transcription is altered by a cell depending on cell type, environmental factors, or age. Differential expression at

various stages of the cell cycle or a change of environmental conditions is induced by cell signalling events that can lead to rapid changes in protein content. Add this effect to the idea that any one gene can actually encode a number of different functional proteins, and PTMs can result in a set of very similar proteins, and there is potentially a vast number of proteins to analyse in a proteomics discovery experiment.

In addition to the potentially large numbers of diverse proteins, the cellular environment contains a further level of complexity in that it can also contain a wide dynamic range of proteins. The range of protein functions means that, in every biological system, there are some proteins which are necessary in much higher copy numbers than others. Even in a relatively simple organism such as *S. cerevisiae* this difference can span up to five orders of magnitude, and some samples such as plasma show an extreme asymmetry, with up to 10 orders of magnitude difference (Hortin & Sviridov, 2010). When studying the entire protein complement of a sample, it can therefore prove difficult to analyse less abundant proteins when the biological sample is dominated by a small number of highly abundant proteins (Anderson & Anderson, 2002; Ghaemmaghami *et al.*, 2003). In such circumstances, traditional proteomics techniques such as electrophoresis or mass spectrometry are insufficient as it is difficult to load enough sample to visualise low abundance proteins without overloading the high abundance species.

As quantitative proteomics techniques advance, there are increasingly large data sets detailing the precise amount of individual proteins in a sample (Hebert *et al.*, 2014). These data sets provide a large amount of valuable information on the identity, sequence and localization of individual proteins. However, in the cellular environment it is rare to find a protein existing as a discrete entity, and many proteins carry out their function as part of a complex. In order to understand the role of a protein within the cellular environment, it is therefore necessary to also study the interactions that the protein exhibits. It has been estimated that more than a third of proteins exist at some stage as part of a complex, and in *S. cerevisiae* over 400 different complexes have been identified so far (Gavin *et al.*, 2006), (Liu *et al.*, 2008). Multisubunit complexes are responsible for much of a cells functionality

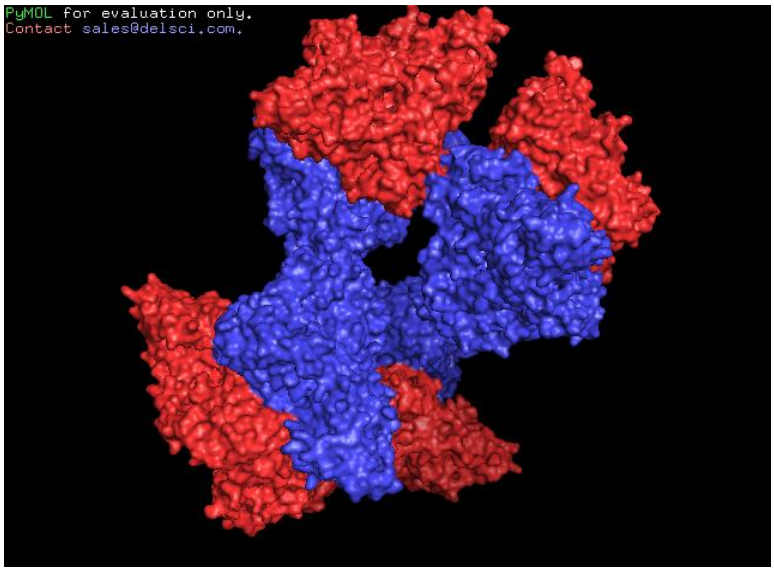
(Alberts, 1998), such as RNA polymerase for the transcription of the genome, the exosome for RNA degradation or the proteasome for the removal of unnecessary protein. They range in size from one or two interacting proteins, such as phosphofructokinase with two subunits (figure 1.1 a, (Banaszak *et al.*, 2011), to large assemblies consisting of a numerous proteins, such as the ribosome, with more than 70 (figure 1.1 b, (Armache *et al.*, 2010), and they can also be dynamic, changing composition to suit cellular conditions. By building up an organisational map of the protein machinery of a cell, it will be possible to understand the interactions occurring, and perhaps build a complete picture of the role of each protein within a cell. For this reason the study of protein interactions has become increasingly widespread (Gavin *et al.*, 2002; Ho *et al.*, 2002; Krogan *et al.*, 2006) as the field of proteomics has developed, as technological advances allow the study of increasing amounts of protein in one analysis.

The function of a protein can be elucidated by knowledge of its interactions, and if a cellular proteome is to be fully understood, the role of each protein within the cell should be studied in terms of interactions and membership of protein complexes. These interactions can form a complex network, and studying them has led to the discovery that some proteins perform multiple functions in the cell, such as ARG82, which acts as an inositol polyphosphate kinase as well as controlling expression of arginine responsive genes (El Alami *et al.*, 2003), or enolase, an abundant glycolytic enzyme that was also discovered to stimulate vacuole fusion (Decker & Wickner, 2006). Individual proteins therefore have the potential for hundreds of interactions, such as the so called hub proteins (Gao *et al.*, 2010). These proteins play a key role in cellular organisation and may have up to 100 different interactions in a cell. For some proteins, the potential for multiple interactions translates into being found as part of a number of complexes, and one such example is lipoamide dehydrogenase, which is part of the pyruvate dehydrogenase (Kresze & Ronft, 1981), glycine decarboxylase (Sinclair & Dawes, 1995) and alpha-ketoglutarate dehydrogenase complexes (Repetto & Tzagoloff, 1991).

Protein interactions can also be classified as stable or transient. Stable interactions cause protein subunits to remain as part of a complex. Other proteins interact in a

Eukaryotic phosphofructokinase

a)



Eukaryotic 80S ribosome

b)

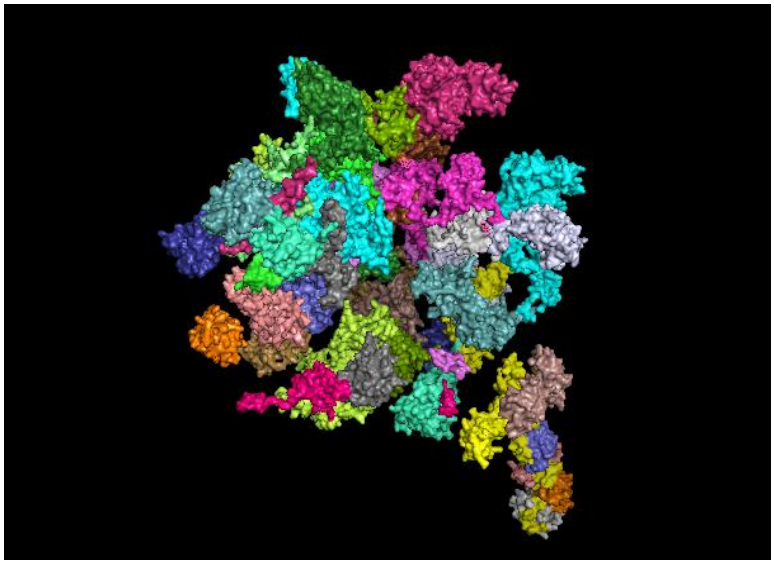


Figure 1.1 Structure of the eukaryotic phosphofructokinase and 80S ribosome complexes

Protein complexes are visualised using Pymol software. Sequences are obtained from PDB, a) phosphofructokinase structure obtained from Banaszak *et al.* 2011, b) 80S ribosome crystal structure obtain from Armache *et al*. 2010.

more transient fashion, such as those involved in cell signalling pathways, or chaperones that assist in protein folding. The study of these transient protein interactions has an added complication, in that they can have high on/off rates, therefore associating and dissociating rapidly depending on conditions, in particular signalling proteins, which alter binding rapidly to suit environmental cues. This gives rise to a diverse and dynamic network of interacting proteins, and protein complexes that can move, combine or break apart, to suit the role, age or environment of the cell in question.

1.4 Mass spectrometry in proteomics

A diverse range of techniques are used within the field of proteomics in the study of proteins. 2D electrophoresis was developed in 1975 to separate complex proteomes (O'Farrell, 1975) and is still used, in conjunction with protein mass fingerprinting (PMF) to identify proteins (Gelis *et al.*, 2012). Western blots can be used to detect proteins, first separating by SDS-PAGE and later transferring the proteins to a membrane where they are bound to an antibody (Renart *et al.*, 1979; Towbin *et al.*, 1979). The technique is used to identify and quantitatively measure proteins, and has even been used in the determination of yeast protein half-lives (Belle *et al.*, 2006). More recently, mass spectrometry has become a more commonly used method for protein identification, and the technology involved is developing rapidly, with new, improved mass spectrometers being developed (Mann, 2008, Han *et al.*, 2008), and the implementation of new techniques such as label free quantification by MSE, or PRMs in recent years (Law & Lim, 2013).

A mass spectrometer separates and identifies ions according to their mass to charge ratio (*m/z*). The instrument consists of three parts, an ionizer to generate gas phase ions, a mass analyzer to separate the ions in time or space, and a detector to measure the amount of each ion present. The design of two types of ‘soft’ ionization techniques, matrix-assisted laser desorption/ionisation (MALDI) (Karas & Hillenkamp, 1988) and electrospray ionization (ESI) (Fenn *et al.*, 1989), which are capable of ionizing peptides without fragmenting them allowed mass spectrometry to be used in the study of proteins. The invention of these techniques

has allowed the development of mass spectrometry for the identification, localisation, sequencing and quantification of proteins.

In proteomics, MALDI can be used for the analysis of relatively simple protein mixtures. In this technique the sample is crystallized with a matrix solution, and then a laser is directed towards the sample/matrix mixture, and the sample is ionized. In this technique, the sample molecules are ionized by the addition of a proton, creating positively charged species, which are often then separated by a time-of-flight (TOF) mass analyser. TOF analysers were first used in 1954 (Wolff & Stephens, 2004) but were only available for use in proteomics after the advent of soft ionisation techniques. A TOF consists of a flight tube, which charged ions are drawn into using an electric field. Ions are then analysed based on the time taken to reach the detector at the end of the flight tube. Mass to charge ratio (*m/z*) is calculated based on the time taken for the ions to reach the detector, at the far end of the flight tube (Watson, 2008). MALDI-TOF is applied generally in peptide mass fingerprinting (PMF) methods, where the peptide *m/z* identified are searched against a database of peptides to identify the protein (Henzel *et al.*, 1993; James *et al.*, 1993; Pappin *et al.*, 1993). This type of analysis can only be used, therefore, on simple mixtures, for example the identification of a protein band excised from SDS-PAGE, or a purified protein sample. The technique also has applications in imaging, in discovering the spatial arrangement of proteins and peptides (Caprioli *et al.*, 1997).

For the identification of proteins in more complex samples, it is more common to use ESI as the ionization technique, in which the liquid analyte is passed through a fine tip while an electrical current is applied, leading to the generation of gas phase, charged molecules (Loo *et al.*, 1989). One advantage of this type of ionization is that it can be preceded by a liquid chromatography (LC) step. Liquid chromatography mass spectrometry (LC-MS) can be coupled to provide separation of the peptides in a complex mixture, allowing for the identification of more ions. ESI is also commonly coupled to TOF, as well as ion trap, Fourier-Transform Ion Cyclotron Resonance (FT-ICR) and Quadrupole mass analysers. In an ion trap, the charged molecules are first trapped for a time in an electric field before release to the

detectors. In this type of mass analyser, different electric fields are applied to the trap to cause the retention, and therefore separation, of ions of a different *m/z*. FT-ICR is based on a similar principle, of trapping ions, but in this case under high vacuum and using a magnetic field (Marshall *et al.*, 1998).

Another common mass analyzer is the quadrupole, which was invented at the same time as ion trap mass analyzers in 1953 (Paul & Steinwedel, 1953). A quadrupole consists of four parallel rods, with a varying RF and a constant DC voltage applied. Depending on the voltages applied, only ions of a certain *m/z* will hold a stable trajectory through the quadrupole, and therefore altering the voltage applied provides a filter, and applying a range of voltages provides separation of ions, by allowing only certain ions through at any time.

Tandem mass spectrometry is a way of increasing the separation of ions by joining two or more mass analyzer steps, which can be tandem in time or tandem in space. In tandem in space mass spectrometry a number of mass analysers, of the same or different types can also be joined in sequence, to provide increased separation of ions, and higher resolution. In tandem in time mass spectrometry, multiple separations are performed in sequence by the same mass analyser. In addition to separation of gas phase ions, tandem mass spectrometry allows an increased level of analysis, in that fragmentation of ions can be performed, meaning that peptides can be matched to a protein based on sequence as well as peptide mass. Common tandem mass spectrometer examples include the quadrupole time of flight (Q-TOF) and triple quadrupole systems (TQ). Q-TOFs provide enhanced resolution (Morris *et al.*, 1996) while TQs provide the ability to isolate peptides, fragment them and then isolate particular fragments, providing extra selectivity (Yost & Enke, 2008).

LC-MS and LC-MS/MS can be applied to many areas of proteomics experiments for protein identification, sequence determination, protein interaction identification, function determination and quantification. Large numbers of proteins can be identified simultaneously from complex protein mixtures by the use of shotgun proteomics methods. In this technique a protein sample, which can contain a complex mixture of proteins, is first subjected to enzymatic digestion to break down

the proteins into peptides. The complex mix of peptides is then separated by liquid chromatography (LC) and analysed in a data dependent acquisition mode using tandem MS, where the most abundant peptides in MS mode are isolated for fragmentation in MS/MS. The instrument cycles between MS mode, where the precursor ions are scanned, and MS/MS mode to gather product ion spectra, either for a set amount of precursor ions defined by the user, or for those that go above a set threshold. The fragments identified by MS/MS are used to identify the amino acid sequence of the peptides, and the peptides identified are then matched to those of known proteins in a database to identify the proteins in the sample. The analysis is performed by a database search engine such as MASCOT (www.matrixscience.com) (Perkins *et al.*, 1999) or SEQUEST (www.fields.scripps.edu/sequest/)(Eng *et al.*, 2014). Continuing advances in instrumentation and shotgun proteomics methods have allowed the identification of thousands of proteins in a single run (Olsen *et al.*, 2009; Michalski *et al.*, 2011; Hebert *et al.*, 2014).

A further level of analysis was provided by the development of these proteomic techniques into quantitative methods. Quantitative proteomics aims at not just the identification, but the quantification of the protein complement of a given biological system. It can be subdivided into relative quantification, for the comparison of the proteome under two or more varying conditions, or absolute quantification, the accurate count of the specific amount of a protein present in a certain sample. The field of quantitative proteomics is evolving rapidly, and in recent years has moved from a focus on 2D gel analysis and western blots, to mass spectrometry (MS) based experimental workflows. It is now possible to quantify thousands of proteins in a single run, and as instrumentation and software advance, the information gained from each MS run can increase.

Both relative and absolute quantification can be achieved in a number of ways, either with an isotopic label (label mediated quantification) or without a label (label-free quantification). Label mediated quantification arose from the design of isotope dilution strategies, in which a standard peptide sequence identical to an analyte is isotopically tagged (Barr *et al.*, 1996). The standard peptide is then mixed

with unlabelled sample, and the ratio of labelled to unlabelled peptide used to calculate the amount of protein, either relatively or in absolute values. In label free quantification the analyte proteins are simply analysed via MS/MS without prior labelling, and statistical analysis is used to infer protein concentration from the peptides present. There are both advantages and drawbacks to each quantification method which must be considered in designing an experimental workflow.

1.5 Quantitative proteomics

1.5.1 Label mediated relative quantification

A wide variety of methods have been developed to compare samples by differential labelling, where relative quantification is achieved by stable isotope labelling of standard peptides or proteins which are pooled with differentially labelled, or unlabelled analyte peptides or proteins prior to MS analysis. The counterpart proteins are isotopically different, but chemically the same, and therefore behave the same experimentally, and co-elute from a chromatography system. MS is then used to compare the peak intensities of heavy isotope labelled and light (unlabelled) peptides.

A common early method was isotope coded affinity tagging (termed ICAT). In this approach, cysteine residues in the sample are labelled using ICAT reagent, which consists of a compound with the ability to specifically bind sulfhydryl groups linked to a compound containing a stable isotope, and an affinity tag (Gygi *et al.*, 1999a). Heavy and light isotope ICAT reagents are used to label two samples for comparison, which are then purified by affinity chromatography and analysed using MS. The mass difference between the isotopes is used to calculate the relative abundance of the analyte (Kito & Ito, 2008).

A method of tagging amino acid N terminal and side chain amino groups has also been developed, termed isobaric tags for relative and absolute quantification (iTRAQ). iTRAQ reagents consist of an amino reactive group coupled to a charged reporter group, and a neutral balance group which maintains a mass of 145 daltons for each reagent. Each has the same mass, but gives rise to a different reporter ion

when analysed by MS/MS. There are 8 types of iTRAQ reagent available, to quantify up to eight different samples in one run (Boehm *et al.*, 2007).

Another approach to labelling is the incorporation of heavy isotopes into the cellular proteome during cell culture. This method is termed SILAC, for stable isotope labelling by amino acids in cell culture. Cells of interest are grown in medium containing stable isotope labelled amino acids, and pooled with unlabelled cells prior to analysis (Geiger *et al.*, 2011). Unlike the other techniques, this allows labelling of all proteins in a sample, and permits the evaluation of larger data sets.

All of these isotopic labelling strategies use differential labelling of two or more sample types, and the ratio of the intensities of the labelled and unlabelled analyte is used to calculate the relative amount of the target proteins. These approaches have an advantage over label free approaches in that the quantification information can be obtained by combining differentially labelled analytes and performing the MS in a single run, therefore eliminating any liquid chromatography (LC) or MS variability between runs. A further advantage to these techniques is that the two labelled peptides are chemically identical, and there should be no variation in ionisation efficiency during MS. The drawback to using these relative quantification methods is that the labelling procedure adds time, expense and complexity to the experimental workflow, thereby increasing the potential for error. Also, the analysis is limited solely to only those analyte proteins that have been labelled, so in techniques such as ICAT and iTRAQ labelling, the scope of the experiment is limited.

1.5.2 Label mediated absolute quantification

Absolute quantification techniques using stable isotope dilution strategies (Barr *et al.*, 1996; Stöcklin *et al.*, 1997) involve the use of an isotope labelled protein standard, such as AQUA peptides or QconCAT to label at the peptide level, or PSAQ to label at the whole protein level, prior to proteolysis to derive the peptides. These isotopically labelled peptides or proteins are mixed with the analyte, before or after proteolytic digestion, where they will perform the same in chromatographic separations due to being chemically identical, but will be differentiated by MS due to the difference in mass. Absolute quantification using these methods has the

advantage of being able to compare proteins from a number of different experiments, as opposed to relative quantification, which is limited to just those proteins compared in a single experiment.

AQUA peptides are synthesised with incorporated stable isotopes, with the same sequence as an analyte peptide (Gerber *et al.*, 2003). They are made individually and therefore at high cost. For the use of AQUA peptides it is vital to achieve complete proteolysis, as the standard is spiked in after the digest and therefore analyte peptides must be completely proteolysed to be identical to the standard peptide and allow accurate quantification.

A QconCAT is an artificial peptide construct, which consists of a number of short peptide sequences termed Qpeptides, derived from a number of analyte proteins (Beynon *et al.*, 2005; Pratt *et al.*, 2006). The peptide gene sequences are assembled into a recombinant protein sequence, which is then heavy isotope labelled during culture. The QconCAT protein is then enzymatically digested, to yield an equimolar mixture of standard peptides. Because every QconCAT can quantify multiple proteins, it is more cost effective for large scale analysis, purchasing one plasmid rather than a number of individual peptides. The QconCAT is quantified and mixed with the analyte sample prior to digestion, and the mixture is then analysed by MS. The ratio between analyte and QconCAT can then be used to quantify corresponding analyte peptides (Pratt *et al.*, 2006). Running a number of standard and analyte peptides at the same time reduces the potential for error between runs, which is useful if a set of related proteins are to be analysed, however to enable large studies comparing many proteins a number of QconCAT constructs may be necessary. In this case reliable comparison between different MS runs will be possible only where QconCAT proteins are accurately quantified. The design of a QconCAT must also be carefully considered, as certain peptides are unsuitable for use as Qpeptides, for example poorly ionising peptides, or those containing PTM sites, and some small proteins may not contain any suitable Qpeptides, although in some cases it may be improved with the use of an alternative enzyme for proteolysis (Al-Majdoub *et al.*, 2014).

PSAQ standards are an alternative strategy which may be more appropriate for some proteins, as these are full length labelled proteins, which will therefore have the same properties as the analyte protein throughout the experimental process (Dupuis *et al.*, 2008). PSAQ standards are mixed with analyte prior to any processing, and can therefore also be included for any purification or prefractionation steps in the analysis, and also downstream proteolysis steps, resulting in increased accuracy (Kaiser *et al.,* 2011). Peptides derived would therefore behave in a similar fashion during chromatography and ionisation in MS, providing an analytical advantage, despite the high cost of generating an entire labelled protein.

1.5.3 Label free relative quantification

There are two main approaches to label free relative quantification, which take advantage of either peak intensities or spectral counting to quantify proteins using MS. Label free quantification by peak intensities exploits the correlation between peptide concentration and MS signal. In this method the chromatographic peak areas of the peptides ions are used to calculate the amount of each protein (Old *et al.*, 2005; Silva *et al.*, 2006). This technique takes advantage of the fact that the intensity of the ions are representative of the peptide concentration (Silva *et al.*, 2006). As runs are performed separately, computer programs are necessary for the normalisation of data to account for any variability in chromatography between runs, and software such as Progenesis LC-MS (Nonlinear Dynamics) is available for aligning runs and calculating relative protein quantification of protein in two samples using peak intensities (Gonzalez-Galarza *et al.*, 2012). Label free relative quantification using peak intensities has been performed using a variety of sample types (Huang *et al.*, 2007; Fatima *et al.*, 2009). This approach is advantageous in the limited workflow in comparison to label mediated strategies. However, as the peptide intensities for each protein are used in the calculation, it is imperative to have excellent sample preparation, in order to achieve complete proteolytic cleavage of each sample, as any miscleaves in the sample will affect the amount of peptide present and cause inaccurate quantification. In addition, for small proteins

it may be difficult to produce a sufficient number of peptides, which may affect the accuracy of quantification.

The alternative label free approach, protein quantification by spectral counting, quantifies analyte protein according to the number of MS/MS spectra identifications of fragments generated from the parent protein. This is based on the concept that spectral count (the number of MS/MS spectra obtained) is directly influenced by the amount of protein (Liu *et al.*, 2004), therefore comparing the number of MS/MS spectra for two samples can provide relative quantification. The number of spectral counts can differ depending on the size of the parent protein, although a procedure has been recommended for normalising spectral counts to the protein length (Florens *et al.*, 2006). It has been suggested that relative quantification is less accurate than label-free quantification by peak intensities, although it has still been successfully applied in the study of a range of protein samples (Old *et al.*, 2005; Florens *et al.*, 2006).

1.5.4 Label free absolute quantification

Label free quantification using peak intensities can also be used in the absolute quantification of an analyte. This can be achieved by the addition of a known quantity of a specific protein to each sample for use as an internal calibrant by the MS analysis software. This is achieved by comparing the peak intensities of a given amount of calibrant to those of analyte proteins to give an accurate count, in moles, of each protein identified. The quantification can be calculated using a top 3 protocol, where the top three highest intensity peptides are summed for each protein, which is the method used by ProteinLynx Global Server (PLGS) software (Silva *et al.*, 2006). Or alternatively, quantification can be achieved using all of the peptides. Software such as MaxQuant can calculate the intensity using iBAQ (intensity based absolute quantification), which uses the sum of all the intensity values divided by the theoretically observable peptides (Schwanhausser *et al.*, 2011), and can use some or all of the peptides identified.

Absolute quantification can also be achieved by spectral counting software, such as emPAI or APEX. The quantification can be achieved using protein abundance index

(PAI), which is calculated from the number of identified peptides divided by the number of theoretic tryptic peptides for each protein (Rappsilber *et al.*, 2002), or using exponentially modified PAI (emPAI values), which show a linear relationship to protein concentration (Ishihama *et al.*, 2005). Another method successfully used in absolute quantification is termed absolute protein expression profiling (APEX). This measures the absolute amount of protein from the relationship between protein abundance and the number of peptides observed (Lu *et al.*, 2007). Various label free software approaches are reviewed in (Gonzalez-Galarza *et al.*, 2012).

All label free methods are based on the idea that peptide intensities are representative of the total amount of protein. It has, however, been observed that the intensity observed can differ as a result of the physiochemical properties of the peptides (Mallick *et al.*, 2007; Sanders *et al.*, 2007). The nature of a shotgun proteomics approach, in loading all of the peptides in a sample for simultaneous analysis, and therefore displaying a wide range in size and amino acid composition, may mean some peptide exhibit less propensity for ionisation than the standard protein peptides, giving an inaccurately low quantification (Craig *et al.*, 2005; Kuster *et al.*, 2005). This discrepancy in signal would therefore be reflected in an inaccurate quantification, which may lead to a disadvantage for label free methods in that they may therefore be less accurate than label mediated quantification.

On the other hand, the advantage posed by label free quantification is in the simplification of experimental workflows, in the elimination of sample preparation steps necessary in labelling. Provided the sample preparation and software calculation is good absolute label free techniques provide the added advantage that information from any number of experiments can be compared, whereas relative quantification is limited only to those samples that have been assessed in a single experiment.

1.6 Quantitative proteomics in the analysis of *S. cerevisiae* proteome organisation

While the technological advances in mass spectrometry towards accurate and reliable quantification of proteins provide vast amounts of information, simply understanding the identity and quantity of proteins in a cellular environment does

not provide an accurate picture of its role within a cell. There is a far greater level of functionality in the cellular environment than there are protein coding genes in the genome, due in part to protein-protein interactions and their subsequent organisation into protein complexes. Despite the complex nature of protein interactions, by building up an organisation map of the protein machinery of a cell, it will be possible to understand the interactions occurring, and perhaps build a complete picture of the role of each protein within a cell, and monitor interactions over time or in changing environments. It is contradictory then, that many traditional proteomics techniques are destructive, breaking down the cellular proteome into constituent proteins, and then in widely used shotgun experiments, digesting the proteins into peptides prior to MS analysis.

Advances are being made in the study of proteome organisation using a range of proteomic techniques. A range of methods have been developed to study the interactions within a cell, as are detailed in various review papers (Collins & Choudhary, 2008; Rajagopala *et al.*, 2012; Ngounou Wetie *et al.*, 2013; Ngounou Wetie *et al.*, 2014). These include yeast two hybrid (Y2H) systems, co-immunoprecipitation, tandem affinity purification (TAP tag) studies, blue native PAGE (BN PAGE) and size exclusion chromatography (SEC) coupled to MS.

An early technique used for the analysis of multiple binding partners is blue native PAGE (BN-PAGE). In this type of electrophoresis, denaturants are not used, and therefore protein complexes should remain associated, as samples are separated in one dimension, using native electrophoresis, and then in a second dimension, using SDS-PAGE, where the protein complex subunits should separate (Schagger & von Jagow, 1991). However, Coomassie Brilliant Blue dye is used to charge the protein complexes for the native electrophoretic separation, and this may disrupt the protein interactions. The technique has been used to separate complexes from yeast mitochondria, but is generally a low throughput technique (Grandier-Vazeille & Guerin, 1996), and is not suitable for all protein complexes.

Yeast two hybrid (Y2H) systems have been used extensively in the study of protein-protein interactions. They provide an assay for the potential for two proteins to

interact, the 'bait' and the 'prey' proteins. As only two proteins are assessed, in order to provide information on larger protein assemblies a number of experiments must be performed. The 'bait' protein is fused to the DNA binding domain of a yeast reporter protein. The 'prey' protein is fused to the transactivation domain of the reporter protein, and interaction between the two proteins is then assessed by the presence of the reporter (Fields & Song, 1989). GAL4 is commonly used as the reporter gene. This technique has been used in large scale studies examining almost 6,000 yeast proteins to identify thousands of interactions (Uetz *et al.*, 2000; Ito *et al.*, 2001; Parrish *et al.*, 2006). Ito *et al* (2001) reported the finding of over 4,000 interactions in their large scale study. Data from these experiments is used to visualise the protein interaction network (figure 1.2, (Jeong *et al.*, 2001), highlighting the complexity of the proteome. However, the reliability of Y2H data has been questioned, as the method produces a high false positive rate, estimated to be as much as 50% (Sprinzak *et al.*, 2003).

TAP tag purification is also a common method used in protein complex analysis. In this technique a protein is fused to an affinity tag, which is used to isolate the protein using affinity chromatography, under sufficiently gentle conditions in which any interacting proteins remain in complex (Rigaut *et al.*, 1999). Though the technique is low throughput, and a new affinity fusion protein must be designed for the discovery of each interaction, the system has an advantage over Y2H in that more than one interaction partner can be studied at a time. The method has been implemented successfully in a number of large scale studies to yield lists of protein interactions numbering over 4000 (Gavin *et al.*, 2002; Krogan *et al.*, 2006; Collins *et al.*, 2007). It has also been found to give fewer false positive results than other methods (von Mering *et al.*, 2002) and isolation of protein complexes can be followed by quantitative mass spectrometry to yield protein complex stoichiometry information (Guerrero *et al.*, 2006). However, analysing a number of different complexes via this method is challenging, in needing the design of individual fusion proteins. In addition, the presence of the tag may interfere with protein binding and if the interaction is weak there is a possibility the complex will dissociate during chromatography.

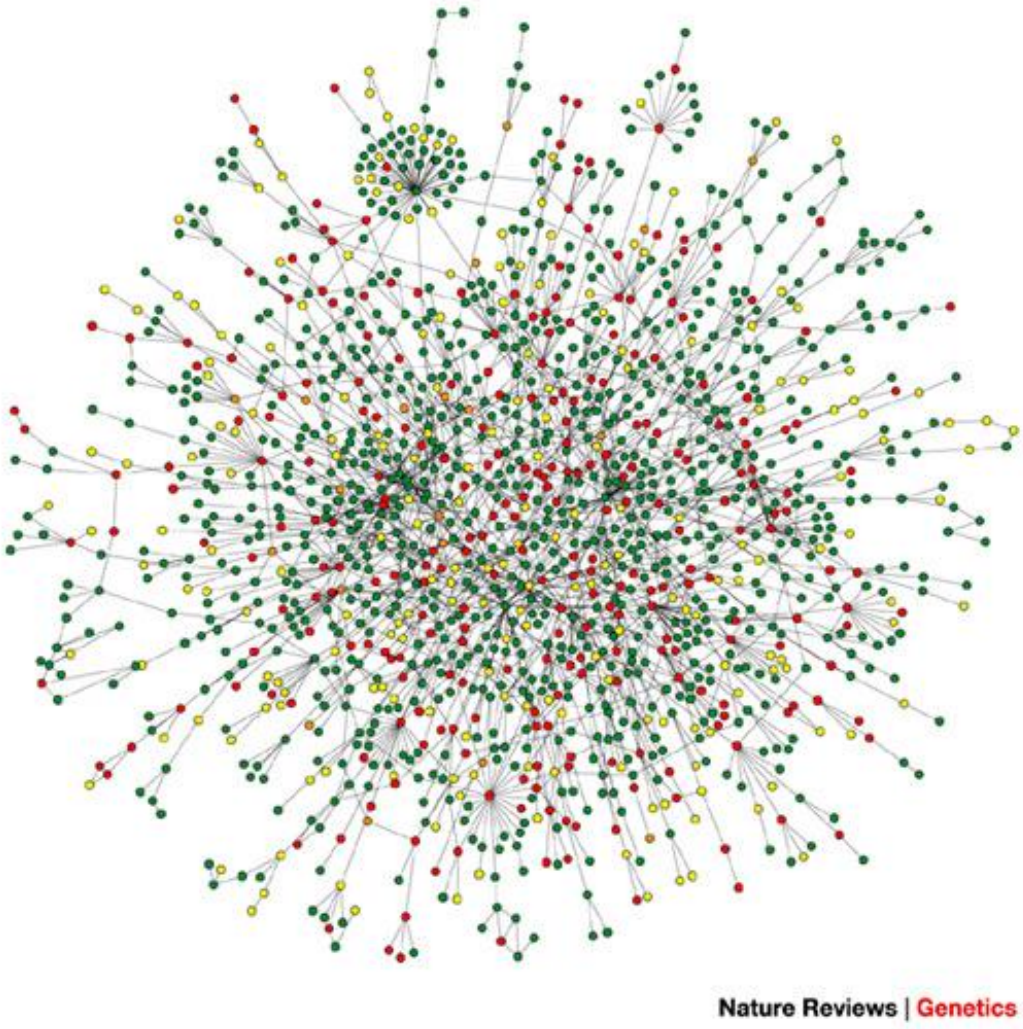


Figure 1.2 Protein-protein interaction map of *S. cerevisiae*

The interaction map is generated from early Y2H experiments. The colour of the node for each protein represents the phenotpic effect of removing that gene, red are lethal, green non-lethal, and yellow unkown Taken from Jeong, H., S. P. Mason, A. L. Barabasi and Z. N. Oltvai (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-42.

In co-immunoprecipitation (Co-IP) techniques a protein complex is precipitated out of a cell lysate using an immobilised antibody that specifically binds one of the protein subunits. The method has been applied to yeast, and from using 10% of predicted yeast proteins as bait, 3,617 interactions were observed (Ho *et al.*, 2002). An advantage of this technique is that it is possible to precipitate proteins associated with the antigen (Kaboord & Perr, 2008), and in some cases cross-linkers can be included to stabilize weaker interactions. Associated proteins can be identified using mass spectrometry or electrophoresis. However, only specific proteins can be targeted, meaning the technique is only useful in the analysis of known, targeted complexes, and not in screening for complexes.

Protein microarrays are another technique developed to study protein interactions, by loading proteins onto a slide, onto which the protein probes are loaded (MacBeath & Schreiber, 2000). Functional protein microarrays spot all of an organisms encoded proteins and can be used in the analysis of protein function and interactions (reviewed in Chen & Zhu, 2006; Reymond Sutandy *et al.*, 2013). The protein bound to the array can be visualised by a number of methods, such as fluorescence of enzyme based assay (for example using horseradish peroxidise, although label free methods are also being developed (reviewed in Reymond Sutandy *et al.*, 2013). The technique has been developed into a high throughput strategy which has been used to find novel protein interactions in yeast (Hesselberth *et al.*, 2006; Tsvetanova *et al.*, 2010).

Another emerging method for analysis of protein complexes is the use of chromatography to fractionate cell lysates prior to MS analysis. Chromatographic techniques separate proteins based on their physico-chemical properties such as size, isoelectric point or hydrophobicity. If the separation is performed under non-denaturing conditions, it is possible to separate interacting partners. Size exclusion chromatography (SEC) has been used to successfully separate protein complexes and analyse the interactions of hub proteins (Li & Giometti, 2007; Gao *et al.*, 2010). Coupling of chromatographic separations to MS requires less preparation prior to experiments, there is no modification of the proteins, thereby proteins are analysed closer to their natural state, eliminating the possibility of protein interactions being

affected by a tag. In addition, it can analyse more protein complexes in one data set, and in theory, if a crude cell lysate is separated it would be possible to analyse all interacting proteins from a sample in one experiment. As well as looking at numerous complexes, this method can simultaneously provide data on all proteins in a complex, as opposed to just two interacting partners, enabling the high throughput analysis of proteome organisation. SEC is generally a low resolution form of chromatography, but it has been coupled to other types of chromatography to improve resolution (Liu *et al.*, 2008). Another advantage of teaming native protein chromatographic techniques and mass spectrometry for the investigation of protein complexes is that quantitative proteomics strategies can be applied to not only identify complex subunits, but to accurately calculate the amount of each present. In quantifying all of the interacting partners in a protein assembly, it is possible the protein stoichiometry of the complex will then be elucidated. The technique may also hold potential for the high throughput study of protein dynamics, if two sample types are separated and analysed, and if label free analysis can be used, there will be limited sample preparation involved.

In addition to identifying the subunits of protein complexes, quantitative proteomics techniques have an application in the study of protein complex dynamics. Label mediated quantitative proteomics in the form of mixing after purification SILAC (MAP)-SILAC was developed for this purpose (Wang & Huang, 2008). The technique involves differentially labelling two cell types, purifying out the protein complexes, and then mixing before analysis. The complexes from the two differentially labelled samples can then be simultaneously analysed. It has been applied in the study the dynamics of a number of complexes, including the yeast complexes eIF2B-eIF2 and cyclin-Cdc28 (Kito *et al.*, 2008). SILAC studies such as these are complicated, labelling of the cells under two different conditions are necessary, which adds expense and time to workflows, in comparison to non-labelling techniques.

There have therefore been a number of studies in recent years attempting to build large scale protein-protein interaction (PPI) maps of proteome organisation in yeast. These have been numerous and varied, and have gathered data on

thousands of interactions. It isn't possible to predict the number of interacting proteins, nevertheless attempts have been made to calculate the value, and the yeast proteome has been estimated to contain anywhere from 10,000 to 60,000 interactions (Hart *et al.*, 2006). It is not possible to know when the PPI network map is complete, but the results of numerous proteome organisation studies are displayed in various databases, such as BIOGRID (Chatr-Aryamontri *et al.*, 2013), IntACT (Kerrien *et al.*, 2012) and MINT (Chatr-aryamontri *et al.*, 2007). The next phase in understanding protein interactions on a proteome wide scale is to build this data into a network of interacting proteins, which yields information on protein complexes. A range of bioinformatics strategies are employed in this endeavour and progress in building and analysing the PPI network maps has been reviewed (Clancy & Hovig, 2014). By integrating data from numerous studies, the reliability of the protein complex predictions is also improved (Xu *et al.*, 2013).

1.7 Proteomics in the exploration of the dynamic range of complex proteomes

The dynamic range of a sample is another problem faced in proteomics, and an important consideration in designing a quantitative proteomic worklflow. In an MS analysis of complex samples, the most abundant proteins can sometimes dominate the results, as current mass spectrometers are capable of reaching across up to four orders of magnitude in concentration, while samples can reach up to 12 (Hortin & Sviridov, 2010; Zubarev, 2013). In order to explore less abundant proteins, samples must therefore be subjected to a preparation step prior to analysis. This usually involves the fractionation of proteins in a sample to separate the proteome into subsets, or depletion strategies, to remove the most abundant proteins, leaving the less concentrated behind.

Fractionation of a sample is based on simplifying protein mixtures in order to divide the components into subsets based on an inherent property of the proteins, and can be based on centrifugation and chromatographic or electrophoretic techniques. In centrifugation techniques cellular organelles are separated at different centrifugation speeds. Chromatographic techniques can involve size exclusion or ion exchange methods to split a proteome into subsets depending on size or iso-

electric point. Electrophoretic prefractionation techniques such as iso-electric focusing separate proteins according to their isoelectric point (pI). These techniques are put to widespread use in the analysis of proteomes and are reviewed in a number of papers (Righetti *et al.*, 2005a; Righetti *et al.*, 2005c).

Immunodepletion strategies are based on the removal of the most abundant proteins. Specific antibodies, dye-ligands or affinity chromatography can be used to bind and remove abundant proteins, and these are discussed in various review papers (Bjorhall *et al.*, 2005; Moser & Hage, 2010). These techniques are all based on the removal of proteins, leading to loss of information.

Combinatorial bead library technology is an alternative technique designed to reduce the dynamic range of biological samples while retaining all of the proteins, by simultaneously reducing the concentration of abundant proteins while increasing the amount of the less abundant proteins. This is achieved by normalising all of the proteins to a similar level of abundance, rather than reducing the complexity of the sample. Combinatorial hexapeptide library technology was designed to address the issues concerning proteome dynamic range in plasma, one of the most extreme examples of asymmetry (Righetti *et al.*, 2005b; Thulasiraman *et al.*, 2005). The libraries consist of a set of ligands bound to spherical, porous beads. Combinatorial hexapeptide ligands are formed from the 20 naturally occurring amino acids, using the Merrifield approach split, couple, recombine method (Furka *et al.*, 1991). In this method, a batch of beads is split into 20, and to each subset an amino acid is bound, following which the beads are pooled, and then split again before attaching another amino acid. The process is repeated six times, to yield a hexapeptide ligand on each bead, and the method leads to a potential 20^6 (64 million) different ligands. A large enough aliquot of beads with a unique ligand on each bead would therefore, potentially possess the capacity to bind every protein in the complex proteome. This allows an aliquot of beads to simultaneously bind the same amount of the highly abundant proteins and the less abundant, when a sample is added, leaving a normalised solution when the excess protein is removed. The method has been used extensively to enable the analysis of low abundance proteins in plasma (Beseme *et al.*, 2010), urine (Petri *et al.*, 2009),

chicken skeletal muscle (Rivers *et al.*, 2011) and has recently even been applied to bevarages (Fasoli *et al.*, 2011; D'Amato *et al.*, 2012).

Combinatorial bead library technology can also be combined with various proteomic strategies, including electrophoresis and mass spectrometry methods to analyse the proteins present (Guerrier *et al.*, 2007; Tu *et al.*, 2011). It has been combined with SRM for the analysis of *S. cerevisiae* to increase the amount of proteins identified in comparison to untreated sample (Di Girolamo *et al.*, 2013). Combining bead library normalisation with LC-MS offers advantages over fractionation or immunodepletion approaches, as the entire proteome should still be present in the sample. This, in theory, enables the analysis of the entire proteome in one experiment. This not only reduces the workflow time, but eliminates the potential for variability between experiments introduced when analysing a sample in separate fractions. Allowing the study of all analytes simultaneously allows the possibility of analysing an entire proteome in a sample.

1.8 Aims

The overall objective of this thesis is to explore the use of quantitative proteomics in the exploration of the *S. cerevisiae* proteome. Specifically, the areas that will be examined here are the organisation of proteins into complexes, and the examination of the entire dynamic range of proteins.

The organisation of protein complexes will be examined by pairing native protein ion exchange chromatography with LC-MS. The hypothesis is that ion exchange chromatography will provide a method of fractionating *S. cerevisiae* cell lysate while some protein assemblies remain intact. Label free quantification will be used in identifying the members of protein complexes. The application of label free in studying the stoichiometry of protein complexes will then be tested, and compared to using a QconCAT based SRM approach for the quantification.

The final part of this thesis will focus on the reduction in the dynamic range of *S. cerevisiae* cell lysate using combinatorial hexapeptide library technology. This will examine the hypothesis that bead libraries are an effective method of

simultaneously reducing the concentration of abundant proteins while increasing that of the less abundant analytes. The idea of teaming normalisation with quantitative proteomics will also be examined. This is based on the theory that there will be a range in which binding to the beads will occur in a linear fashion, up to the point where a protein will saturate the beads. Therefore any protein below this saturation amount can be assumed to represent the total amount of protein present in the sample.

In combining the two approaches, the possibility of using quantitative proteomic strategies to build a complete picture of *S. cerevisiae* will be examined.

2. Methods

2.1 *Saccharomyces cerevisiae* culture

S. cerevisiae (BY4743 ΔARG3 strain) cells were streaked onto YPD agar plates (containing 200 mg/l of G418 antibiotic) and incubated at 30°C. After approximately 40 hours a single colony was used to inoculate 5 ml YPD media with G418 antibiotic and incubated at 30°C with shaking at 150 rpm overnight. The following morning, the OD_{600nm} of this overnight culture was taken, and used to calculate the amount to inoculate a 200 ml culture. The inoculation amount was based on 20 hours of incubation, with shaking, at 30°C, and estimating a maximum growth rate at OD_{600nm} of 0.3 per hour in order to reach a final OD of 2.5. After 20 hours, the culture was decanted into 15 ml aliquots, centrifuged (32706 RCF for 10 minutes at 4°C) and then pellets were stored at – 80 °C until needed.

2.2 *S.cerevisiae* cell lysis

Lysis of *S. cerevisiae* cell pellets was performed using a Minilys homogeniser (Precellys). Cell pellets generated from 15 ml culture are suspended in 250 µl of the relevant lysis buffer containing protease inhibitors (one Roche complete EDTA free protease inhibitor cocktail tablet per 20 ml buffer, 20 mM sodium phosphate for equaliser experiments, 50 mM HEPES pH 7.4 for fractionation). Each cell pellet was then transferred to a microtube containing glass beads (Precellys) and the cells disrupted by 15 rounds of shaking using a Minilys homogenizer (Precellys) on the top speed setting for 30 seconds, followed by one minute on ice. The cell lysate was then centrifuged at 9542 RCF for 5 minutes, and the supernatant collected and combined. The lysate was then stored at 4°C until used.

2.3 Coomassie protein assay

For assay of protein concentrations using coomassie bovine serum albumin (BSA) was used as to create a standard curve in concentrations of 10 mg/ml, 20 mg/ml, 30 mg/ml, 40 mg/ml and 50 mg/ml. Appropriate dilutions of the analyte protein solution were made so that the concentration fell in the range of the standard curve. Coomassie Plus^{TM} protein assay reagent was mixed 2:1 with standards and analyte samples in duplicate. The absorbance at 620 nm was measured using a Labsystems Multiscan Ascent colourimetric scanner and the concentration of the analyte solution was calculated by the software.

2.4 Fast process liquid chromatography (FPLC) fractionation

2.4.1 Mixed bed ion exchange

Protein concentration of samples was assessed by coomassie protein assay and dilutions prepared. These were centrifuged at 12,100 RCF for 10 minutes and filtered using a 0.22 µm filter. The sample (0.5 ml) was then loaded onto a 2 ml capacity mixed bed ion exchange chromatography column (PolyLC) in start buffer (50 mM HEPES pH 7.4). Fractionation was performed on an AKTA purifier (GE life sciences) system at a flow rate of 0.5 ml per minute using gradient of 0-85% buffer B (50 mM HEPES, 1 M NaCl). A total of 35 x 1 ml fractions were collected. These were transferred into 1.5 ml eppendorf tubes before being concentrated and desalted using strataclean resin.

2.4.2 Anion exchange

S. cerevisiae cell lysate was diluted to 15 mg/ml and 0.5 ml loaded onto a 6 ml Resource Q column with an AKTA purifier (GE life sciences). A starting buffer of 50 mM HEPES pH 7.4 was used, and fractionation was performed at a flow rate of 3 ml per minute using gradient of 0-45% buffer B (50 mM HEPES, 1 M NaCl) over 8 column volumes. A total of 50 x 1 ml fractions were collected.

2.5 StrataClean resin concentration

For SDS PAGE or LC-MS analysis, the fractions were also concentrated using strataclean resin, by transferring 500 µl of each fraction to a 0.5 ml eppendorf with 20 µl of StrataClean hydroxylated silica beads (StrataClean™ resin, Agilent Technologies). This was then vortexed for 1 minute, and centrifuged at 430 RCF for 2 minutes. The supernatant was removed, and the remainder of the fraction added to the strataclean pellet. This was vortexed for a further minute, and then centrifuged at 430 RCF for 2 minutes. The supernatant was removed, and the protein bound to the beads was extracted using in solution proteolysis or SDS PAGE.

2.6 SDS PAGE of protein fractions

SDS sample buffer (1 ml 0.5M Tris buffer, pH 6.8, 1 ml glycerol, 0.02 g sodium dodecyl sulfate, 0.01 g bromophenol blue, 0.154 g dithiothreitol), was added to an appropriately diluted sample, or protein bound StrataClean resin, vortexed briefly to mix and then incubated at 95°C for 10 minutes before loading onto a 12% 1D SDS PAGE gel. Electrophoresis was performed at 200V for approximately 40 minutes, using BIORAD mini gel series equipment (Biorad-laboratories, UK) and the protein bands were visualised by staining in coomassie blue.

2.7 In solution proteolysis

Sample was first made up to a volume of 160 µl with 25 mM ammonium bicarbonate solution (at an approximate amount of 100 µg protein if known), then 10 µl of a 2% Rapigest suspension (Waters) was added and incubated at 80°C for 10 minutes using an eppendorf thermomixer with rotation at 950 rpm. After 10 minutes the solution was returned to room temperature, centrifuged briefly and 10 µl of a 9.2 mg/ml solution of DTT was added, and incubated at 60°C with 950 rpm mixing. Following this, 10 µl of a 33 mg/ml solution of iodoacetamide was added and incubated at room temperature in the dark for 30 minutes before 10 µl of (200

µg/ml) trypsin was added (Promega trypsin resuspended in 50 mM acetic acid). Digests were then incubated at 37°C with shaking at 90 rpm for 4 hours, before a further 10 µl of trypsin was added. This was then incubated at 37°C overnight. The following morning, 1 µl of TFA was added to the solution before returning to the thermomixer at 37°C for 45 minutes. The digests were then centrifuged at 4°C, 17135 RCF for 15 minutes, and the supernatant removed. The centrifugation was repeated twice more, and the peptide solutions were stored at 4°C until LC-MS.

2.8 LC-MS using a Waters Synapt G2 Q-TOF system

Proteolysed samples were diluted 1:1 with a 50 fmol/µl solution of rabbit glycogen phosphorylase B to act as a standard protein for the label free software, and 2 µl was then injected onto a nanoACQUITY system (Waters, UK). Peptide samples were introduced in aqueous solvent (buffer A: HPLC grade H_2O, 0.1% TFA) at a flow rate of 5 µl/min and passed through a 180 µm by 20 mm, 5 µm bead C 18 trap (Waters) at a flow rate of 5 µl/min. The peptides are then introduced to a 20 cm long 1.7 µm C-18 column at 75 µm by 150 mm, initially at 97% buffer A, followed by separation using a 110 minute gradient rising to 85% buffer B (HPLC grade ACN, 0.1% TFA) at a flow rate of 300 nl/min. A capillary voltage of 3 kV was applied to ionise samples before introduction the mass spectrometer. Samples were acquired using a data dependent acquisition (DDA) mode, resolution 20,000, using a parent mass survey range of 50-2000 *m/z* using $HDMS^E$ and a scan time of 1 second, with a trap collision energy of 4V. Precursors were then fragmented at a trap collision energy of 15-40 V and MSMS spectra obtained. During acquisition the instrument switches to a lockspray of Glufibronopeptide (*m/z* 785.84) infusion every 30s for use as a calibrant.

2.9 LC-MS using a Thermo QExactive Quadrupole-ion trap

Chromatography was performed using an Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific). Samples were introduced in buffer A (HPLC grade H_2O, 2% ACN, 0.1% TFA) and loaded onto a C18 3 µm trap column (Acclaim PepMap 100, 2 cm x 75 µm inner diameter) run on a 90 minute linear gradient from 3.8% to

50% buffer B (90% ACN, 0.1% TFA), followed by 5 minutes at 99% buffer B, at a flow rate of 300 nl/min, using a Thermo Scientific EASY-Spray$^{(TM)}$ PepMap® RSLC 15 cm x 75 μm inner diameter 2 μm C18 column. Samples were run on the QExactive in a DDA method. Full scan survey was performed by the Orbitrap at resolution 70,000, and a target ion accumulation of 1 x10^6. The scan range was 300 to 2000 *m/z* and the top ten precursor ions with a charge of two or above were isolated for fragmentation, dynamic exclusion 20s. The typical spray voltage was 2 kV, capillary temperature 250 °C, no sheath or auxiliary gas. The MS/MS ion selection threshold was set to 1 x 10^4 counts and a 2 *m/z* isolation width.

2.10 QconCAT expression

The QconCAT gene is synthesised in the pET21 vector, for expression using the T7 expression system. Plasmid DNA (5 μg) was solubilised in 50 μl TE buffer (10 mM Tris/1 mM EDTA), pH 8.0. A 1 ng/μl solution of the plasmid was prepared, and 5 ng added to 200 μl of *E.coli* BL21 cells. The cells were mixed and held on ice for 30 minutes, before heat shock treatment at 42°C for 30 seconds. Cells were then placed back on ice for 2 minutes, 1 ml LB broth added, and incubated 37°C for 1 hour. Transformed cells were then plated onto LB agar plates with 50 μg/ml Ampicillin, and incubated 37°C overnight. Transformed cells were then re-streaked onto LB Ampicillin plates, and incubated at 37°C overnight. A single colony was used to inoculate 10 ml of LB broth, incubated for four hours, and 100 ul used to incoulate 5 ml minimal media without amino acids (20% M9 salts, 0.2% 1 M $MgSO_4$, 2% glucose(20%), 0.01% 1 M $CaCl_2$), at 37°C overnight, and then 4 ml of this culture used to inoculate 200 ml of minimal media with amino acids (as above with 20 mg each amino acid, [^{13}C6]arginine and [^{13}C6]lysine). The OD_{600nm} was measured every 60 minutes and when an OD_{600nm} of 0.6 was reached expression was induced by the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). After incubating for 6 hours at 37°C, with shaking, cultures were centrifuged at 25040 RCF 4°C for 15 minutes, and cell pellets are stored at -20°C.

2.11 Nickel affinity column purification

QconCATs were purified using Nickel affinity purification columns (Ni-MAC), which bind the QconCAT using the hexahistidine tag included in the sequence. Solubilized inclusion bodies were loaded onto the Ni-MAC column, which were then washed using 10 ml of bind buffer (6 M guanidine, 1.2 M NaCl, 200 mM Na_3PO_4, 40 mM imidazole, pH 8.0), followed by 6 ml wash buffer (6 M guanidine, 1.2 M NaCl, 200 mM Na_3PO_4, 80 mM imidazole, pH 8.0). The QconCATs were then extracted from the column using elution buffer (6 M guanidine, 1.2 M NaCl, 200 mM Na_3PO_4, 1 M imidazole, pH 8.0) and 10 x 0.5 ml fractions collected. Eluted fractions were treated with strataclean resin to remove guanidine HCl and samples were run on SDS PAGE. The two fractions with the highest concentration of QconCAT were combined and dialysed against 1500 ml 50 mM ammonium bicarbonate 1 mM DTT overnight. Purified QconCAT was protein assayed using a standard protocol Coomassie Plus Protein Assay, and stored at -20°C until use.

2.12 In gel Proteolysis of QconCAT bands

Suspected QconCAT bands were excised from the gel using a glass Pasteur pipette, added to 25 µl 25 mM ammonium bicarbonate, and incubated 37°C for 15 minutes. The ammonium bicarbonate was then removed and replaced with 50 µl 50% acetonitrile/ammonium bicarbonate, and incubated at 37°C for 15 minutes. This step was repeated as necessary until the gel plug was thoroughly destained. Following this, 25 µl of a 1.5 mg/ml solution of DTT is added and the plug is incubated for 60 minutes at 60°C. After discarding the DTT solution, 25 µl of a 10 mg/ml iodoacetamide solution was added and incubated at room temperature in the dark for 45 minutes. The plug was then washed in 25 mM ammonium bicarbonate for 15 minutes at 37°C, and 10 µl acetonitrile until the plug was dehydrated. The dried gel plug was then incubated 37°C overnight in 10 µl 12.5 ng/ml trypsin. The tryptic digest solution was then pipetted off the gel plug into a 0.5 ml microfuge tube and stored at -20°C until MS analysis.

2.13 MALDI TOF mass spectrometry on Bruker UltrafleXtreme™

Samples (1 µl) were dried onto a stainless steel anchorchip target plate (Bruker) and mixed on the plate with an equal volume of α-cyano-4- hydroxycinnaminic acid (HCCA) matrix (10 mg/ml solution in 50% ACN, 0.1% TFA). Analysis was carried out in reflectron mode using laser energy of approximately 30%, over a range from approximately 800 – 4000 *m/z*. Spectra were analysed using Bruker FlexAnalysis software, processed by baseline subtraction, smoothing and generate a peak list based on a S/N threshold of 6.

2.14 SRM using a Waters Xevo triple quadrupole system

QconCATs were diluted to 0.25 fmol/µl, 1 fmol/µl and 10 fmol/µl using the relevant fraction digests. Samples were introduced to a nanoACQUITY in aqueous solvent (buffer A: HPLC grade H_2O, 0.1% TFA) at a flow rate of 5 µl/min and passed through a 180 µm by 20 mM, 5 µm bead C 18 trap (Waters) at a flow rate of 5 µl/min. The peptides are then introduced to a 20 cm long 1.7 µm C-18 column at 75 µm by 150 mm, initially at 97% buffer A, followed by separation using a 60 minute gradient rising to 40% buffer B (HPLC grade ACN, 0.1% TFA) at a flow rate of 300 nl/min. Cone volts were set to 35 kV while collision energy was optimised for each peptide. Required points per peak were set to 30 and a peak width of 15 seconds, with an optimal dwell time of 0.05.

2.15 Proteominer™ bead library experiments

Proteominer beads (Biorad) were supplied in dried form and prepared according to manufacturer instructions. Beads (525 mg) were reswelled by adding 10 ml of 20% ethanol and incubated overnight at 4 °C with shaking at 50 rpm. Beads were then stored at 4 °C until use. *S. cerevisiae* cell lysate was diluted to the relevant concentration using 20 mM sodium phosphate buffer (pH 7.4). Bead aliquots (80 µl)

were allowed to equilibrate for ten minutes on a rotating mixer using the relevant buffer, and then loaded with 1 ml of sample. After 120 minutes on a rotating mixer at room temperature the beads were left to settle before the beads were allowed to settle for 5 minutes, and the excess sample removed. The protein bound beads were then washed 5 times in 20 mM sodium phosphate buffer (pH 7.4), and the amount of protein bound to the beads was determined via SDS-PAGE or LC-MS.

2.16 Software used

2.16.1 PLGS

Waters PLGS (V 2.119) software was used in the analysis of data generated from the Synapt G2. Peptide identifications were performed by searching against a database of yeast proteins obtained from Uniprot (2010_01). Database searching was done using an FDR of 4%, a minimum of 1 peptide per protein, minimum fragment ions matched 3 per peptide and 7 per protein, with 1 miscleave, carboamidomethyl (C) as a fixed, and oxidation of methionine as a variable modification. Label free quantification is performed on the identified proteins using Hi3 methodology, where the average MS signal is calculated for the standard protein (glycogen phosphorylase B) and this is used to calculate the fmol of each protein identified, using the top three most intense peptides.

2.16.2 MaxQuant

QExactive generated data was processed with MaxQuant (V 1.3.0.5) with peptide identification using the default settings and the Andromeda search engine with a yeast database. Carbamidomethylation of cysteine was set as a fixed modification, and oxidation of methionine as a variable modification. FDR was set to 1%, with minimum unique and razor peptides set to 1. Maximum missed cleavages were set to 2 and a maximum charge state of 4. Label free quantification was performed by

calculating the iBAQ intensity of all of the peptides identified for each protein. The iBAQ intensity of the standard protein (glycogen phosphorylase B) is then used to calculate the quantity of each protein.

2.16.3 Skyline

Skyline software V2.1.0.4936 was used in the design of transition lists and the processing of SRM data. PeptideAtlas and ETH spectral libraries were used in conjunction with libraries designed from Synapt G2 data (QconCAT digests analysed by PLGS) to chose optimal transitions. QconCAT peptide lists were imported into Skyline, and transition lists generated were filtered to include only 1+, 2+, or 3+ charges and Y ions. The three optimal transitions for each peptide were chosen, and exported as an excel file for use with Masslynx software.

2.16.4 Masslynx

Masslynx software V4.1 was used in the processing of Xevo generated SRM data. Data was first imported into Masslynx and visually inspected to determine the peptide Type (A, B or C). For calculation of peptide quantities chromatograms were integrated using the mean smoothing method, number of smooths 2, window 2, baseline start and end thresholds 10%. The intensity contributed from the unlabelled QconCAT, calculated from either the G2 data (YEW1) or the SRM negative control (YEW3) was then subtracted from the analyte peak intensity before the fmol amount was calculated based on the QconCAT dilution used.

Chapter 3: Analysis of *Saccharomyces cerevisiae* proteome organisation using ion exchange chromatography and LC-MS/MS

3.1 Introduction

While technological advances towards accurate and reliable quantification of proteins provide key information on the cellular proteome, simply understanding the identity and quantity of proteins in a cellular environment does not provide an accurate picture of their role within a cell. The proteome is a complex, diverse and dynamic environment, with a complexity partly achieved because there is a far greater level of functionality in the cellular environment than there are protein coding genes in the genome. It has been estimated there are an order of magnitude more functional proteins in the cell than are encoded (Fields, 2001). The product of a protein encoding gene can be altered by events such as alternative splicing and post translational modification. In addition to the wide variety of proteins, some proteins can also play a variety of roles within a cell, and performing multiple functions is known as protein moonlighting (reviewed in Gancedo & Flores, 2008; Flores & Gancedo, 2011).

A further level of complexity in the proteome is created from protein-protein interactions leading to the organisation of proteins into complexes. It has been estimated that some proteins can have hundreds of different interactions within the cell, meaning that individual proteins can therefore be found as part of a number of different complexes which are essential to a cells functionality (Jeong *et al.*, 2001; Song & Singh, 2013). The current estimate for the number of protein coding genes in the *S. cerevisiae* genome is 5070, as listed by the *Saccharomyces* genome database, SGD (Cherry *et al.*, 2012). More than a third of these can exist at some stage as part of a complex, and over 400 different complexes have been identified so far (Gavin *et al.*, 2006) (Liu *et al.*, 2008). Multi-subunit complexes are responsible for much of a cells functionality, such as RNA polymerase for the transcription of the genome, the exosome for RNA degradation or the proteasome for the removal of unwanted or unnecessary protein. By building up an organisation

map of the protein machinery of a cell it may be possible to further understand the interactions occurring and perhaps build a complete picture of the role of each protein within a cell.

A variety of methods have been implemented in the study of *S. cerevisiae* protein-protein interactions, including yeast two hybrid (Y2H) systems, TAP tag studies and co-immunoprecipitation, which have all been used in large scale studies discovering numerous interactions. *S. cerevisiae* two hybrid systems are an assay of the potential for two proteins to interact, where the 'bait' protein is fused to the DNA binding domain of the *S. cerevisiae* protein Gal4. The 'prey' protein is fused to the transactivation domain of Gal4, and interaction between the two proteins is then assessed using a reporter (Fields & Song, 1989). This technique has been used in large scale studies to discover almost 1000 protein-protein interactions in *S. cerevisiae* (Uetz *et al.*, 2000). In TAP tag purification a protein is fused to an affinity (TAP) tag, and the tagged protein is then isolated using affinity chromatography, under gentle conditions in which any interacting proteins remain in complex. This system had been implemented successfully to yield information on over 500 protein-protein interactions (Gavin *et al.*, 2002). In co-immunoprecipitation techniques an antigen is precipitated out of solution using an antibody that specifically binds that protein, and if sufficiently non-denaturing buffers are used during this procedure, it is possible to precipitate proteins associated with the antigen (Ho *et al.*, 2002).

These are low throughput methods, as in Y2H and TAP tag a new fusion protein must be designed for a single proteins interaction partners to be found, adding time and complexity to the experiment, and the presence of the tag may interfere with protein binding. In immunoprecipitation techniques, if the interaction is weak there is a possibility the complex will dissociate during isolation. All of these techniques involve the small scale study of individual protein-protein interactions, but have enabled the development of big data sets in large scale studies. However, many proteins in the cell associate into large, multi-subunit complexes, which can contain large numbers of proteins, and the small scale affinity purification or yeast two hybrid approaches do not provide stoichiometric information on the assembly of

these complexes. In addition, the large number of interactions within some assemblies can lead to variation in results, and some studies have shown a poor overlap between results, indicating a high degree of false positives (Cornell *et al.*, 2004). Although low throughput, there have been many studies using these technologies, and a plethora of information of yeast protein interactions have been gathered, and assembled into databases such as MINT (Chatr-aryamontri *et al.*, 2007), CyC2008 (Pu *et al.*, 2009), BioGRID (Chatr-Aryamontri *et al.*, 2013) or IntAct (Kerrien *et al.*, 2012). The incorporation of information into databases enables the construction of interactome networks and the prediction of protein complexes (Clancy & Hovig, 2014).

In addition to the protein interaction studies, entire *S. cerevisiae* protein complexes have been visualised using techniques such as X ray crystallography, NMR spectroscopy, or cryo-electron microscopy (cryo-EM). X ray crystallography involves the crystallisation of a protein, or complex, and the subsequent diffraction of x rays through the crystal enables the visualisation of protein structure. The technique has been successfully used to characterise a number of yeast complexes, such as the eukaryotic exosome and Ccr4-not complexes (Basquin *et al.*, 2012; Makino *et al.*, 2013a). Cryo-EM techniques are based on electron microscopy and have been used to characterise the helicase core $MCM_{2\text{-}7}$, and the sub-discipline of electron crystallography has also been used to image yeast RNA polymerase II (Poglitsch *et al.*, 1999). NMR spectroscopy exploits the magnetic field around atomic nuclei to provide information on the surrounding structure, and allows the imaging of complete multisubunit complexes in some situations where X ray crystallography is unable to (Gross *et al.*, 2003). However, both of these techniques require the isolation of large amounts of purified protein complexes.

Mass spectrometry can also be used for the analysis of whole protein complexes in the form of native, intact protein mass spectrometry. Native mass spectrometry involves the use of native electrospray ionisation (ESI) in an attempt to preserve the quaternary structure of the protein interactions, to elucidate the stoichiometry of the complex and in the analysis of the exact mass of each of the individual subunits, and offer advantages in high sensitivity, reduced sample size, and high throughput

analysis, reviewed in a number of papers (Heck, 2008; Lorenzen & van Duijn, 2010). Native MS techniques have been applied to the study of protein complexes such as the yeast exosome complex (Synowsky *et al.*, 2006) and RNA polymerase III (Lorenzen *et al.*, 2007).

Alternatively, native protein chromatography followed by LC-MS/MS can be used in the analysis of protein complexes. This technique involves retaining the quarternary structure for the first chromatographic stage of the analysis, following which the subunit composition of the separated complexes can be analysed by shotgun proteomics. This provides a method of combining protein complex subunit identification with the quantification of each subunit. Various forms of chromatography can be used in the study of protein complexes, including size exclusion chromatography (SEC), ion exchange chromatography (IEX), and hydrophobic interaction chromatography (HIC).

Size exclusion chromatography, also known as gel filtration, involves the separation of proteins, peptides or protein complexes on the basis of size. The chromatography column is packed with porous beads, into which the sample proteins or peptide can diffuse, depending on size. The sample is loaded onto the chromatography column, and the proteins which are of a small enough size to enter the porous beads take longer to travel through the column. This technology has been used to successfully separate *S. cerevisiae* protein complexes, prior to proteolytic digestion and LC-MS, which enabled associated proteins to be identified (Li & Giometti, 2007; Gao *et al.*, 2010). Size exclusion chromatography is the gentlest of these chromatographic methods, as it uses samples suspended in simple non-denaturing buffers, however, it usually offers poor resolution, with a column packing material separating a fairly narrow size range of proteins. Protein complexes within the cell range from 2 subunits in structures such as phophofructokinase and fatty acid sythetase, up to more than 70 subunits in larger assemblies such as the ribosomes, and can consequently cover a wide range of molecular weights and contain a wide range of components (Robinson *et al.*, 2007; Ban & Egelman, 2010). Therefore, this method is not ideal for the high throughput

methodology required for the study of all of the complexes in the whole cellular proteome.

Hydrophobic interaction chromatography separates proteins on the basis of hydrophobicity. The exposed hydrophobic sections of the protein are encouraged to interact with the hydrophobic column material by loading the sample in a high ionic strength solution. This type of chromatography has also been coupled to MS methods for the study of protein interactions (Liu *et al.*, 2008).

Ion exchange chromatography separates protein samples on the basis of charge, either by the use of positively charged medium (anion exchange chromatography or AEX), or negatively charged medium (cation exchange chromatography or CEX). The chromatography media interacts with sample proteins carrying the opposite charge, and the protein can later be eluted from the column by displacement with a solution of high ionic strength, generally with the use of a buffer solution containing NaCl. A gradient of increasing strength NaCl can be used to increase the time between the elution of proteins of different charge (and therefore different affinity for the column material), giving this type of chromatography a higher resolution than SEC, and it has also been used in the study of protein complexes (Liu *et al.*, 2008; Gao *et al.*, 2010). Combining chromatographic separation of native proteins with subsequent LC-MS methods provides the potential to analyse all of the available protein complexes in a cell lysate in one set of experiments. It allows the study of proteins in their natural conformation, eliminating any potential for alteration in secondary structure, as is possible in other techniques.

The process of breaking a protein down into peptides and analysing the peptide fragments by MS-MS to build information on the parent protein is termed bottom up proteomics (Armirotti, A., 2009). The use of bottom up proteomic methods following chromatography allows not just the identification, but the quantification of each subunit using label-free methodologies. Label-free methods allow the quantification of protein samples without generating specific labelled protein standards, meaning the experimental workflow is simplified in comparison to label mediated methodologies. This method of quantification takes advantage of the

correlation between the intensity of a peptide and the concentration of the protein. There are various software tools for the calculations, some of which use just a subsection of the peptides identified, such as the top three most intense, and some of which use more peptides (Old *et al.*, 2005; Silva *et al.*, 2006; Schwanhausser *et al.*, 2011). These label free methods apply data that has been gathered from a mass spectrometer using data dependent acquisition, a mode of running whereby MS/MS fragmentation of peptides is triggered for the most intense peptides the mass spectrometer detects in a set time window. Coupled with HPLC for the separation of the complex mixture of peptides, this non targeted mode of operation means that data is potentially gathered for all of the proteins in a sample, depending on the dynamic range of the sample, and the sensitivity of the mass spectrometer.

3.2 Aims

The primary aim of this chapter was to develop an efficient, high throughput method for the fractionation and quantification of *S. cerevisiae* protein complexes. This was attempted using anion exchange chromatography and mixed bed ion exchange chromatography systems to separate a clarified yeast cell lysate solution into a discrete set of fractions. This is based on the hypothesis that the separation will be sufficient to fractionate protein complexes, while maintaining their quaternary structure.

There are numerous types of interactions found in protein-protein interactions, such as ion-ion, hydrogen bonds, dipole-dipole, dispersion and hydrophobic interactions. Owing to the ionic strength required for the elution of proteins from the ion exchange columns, it is possible that some electrostatic interactions will be affected, as it is known these interactions can be weakened in the presence of salt (Bertonati *et al.*, 2007). However, higher ionic strength conditions can sometimes support the maintenance of protein structure, suggesting the importance of hydrophobic interactions in protein complex structure, so it is possible that many interactions will be unaffected (Gao *et al.*, 2010).

The protein content of the fractions, potentially including all of the cellular protein complexes, will then be analysed using LC-MS/MS. The protein content of each fraction will be calculated using label-free quantification, in which the correlation between peptide intensity and protein concentration is exploited to calculate the quantity of each protein present in the sample. This method will potentially provide a high throughput method for the analysis of the stoichiometry of protein complexes. Conversely, the stoichiometry of the protein complexes will also provide a test of the accuracy of the label-free methods, as a number of known protein complexes should be observed, where the relationship between the subunits is already characterised.

3.3 Results and Discussion

3.3.1 Anion exchange chromatography

S. cerevisiae cell lysate was prepared according the protocol outlined in section 2.2. The cell lysate was centrifuged to remove insoluble material prior to loading on the chromatography column. Anion exchange chromatography was performed using a 6 ml Resource^{TM} Q column (GE Life Sciences), with protein loading and gradient length adjusted to optimise the separation between complexes, with the aim of identifying the maximum number of proteins. The optimised method consisted of a 0.5 ml load of 15 mg/ml cell lysate run on a gradient of 0-85% buffer B over 8 column volumes. Buffer A consisted of 50 mM Hepes, pH 7.4 while elution buffer B was composed of 50 mM Hepes, 1 M NaCl, pH 7.4. The long gradient and high elution buffer was designed to fractionate the majority of yeast proteins. Due to the delay in fraction collecting on the AKTA, protein began to elute at fraction 5 and continued, according to the UV trace, until approximately fraction 50 (figure 3.1 a), therefore fractions 5-50 were taken to be analysed by SDS PAGE and mass spectrometry.

Two sets of fractions were prepared, one of which will be analysed by SDS-PAGE and the other by mass spectrometry. The two sets of fractions were prepared from the same cell lysate dilution, and run on the same day, using the same buffers. In doing so, the chromatographic profile of the two separations remained the same (figure 3.1 b).

SDS PAGE of the fractions collected (figure 3.2 a) show that at a neutral pH of 7.4, many proteins do not bind to the column, and are eluted in the initial fractions, before the gradient. This is because any protein with an isoelectric point below 7.4 will carry a positive charge, and therefore be unable to bind the anion exchange resin. A protein with an isoelectric point above this will carry a negative charge, and therefore will remain bound with the column material until it is displaced by the gradient. However, despite the unbound material not undergoing separation during the gradient, it has still undergone a form of fractionation, an isocratic separation,

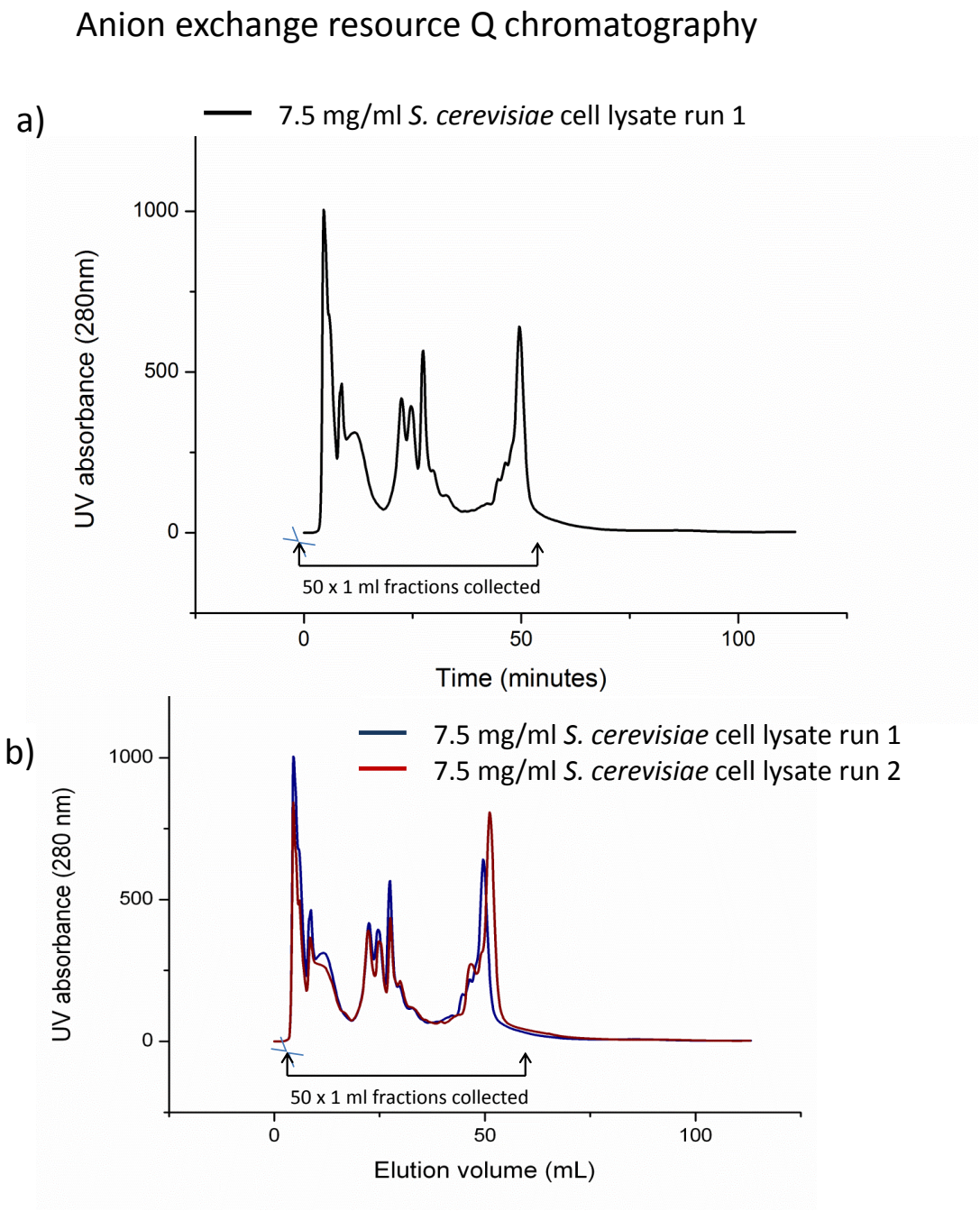


Figure 3.1 *S. cerevisiae* cell lysate fractionation by anion exchange chromatography

S. cerevisae cell pellets generated from 15 ml of culture where lysed by bead beating, and diluted to 15 mg/ml in starter buffer, 50 mM HEPES, pH 7.4. A 0.5 ml aliquot of the dilution was injected onto a Resource Q anion exchange column, and separated at 2 ml/min on an AKTA chromatography system, with elution using a gradient of 0-0.85 M NaCl. 50 X 1 ml fractions were collected.

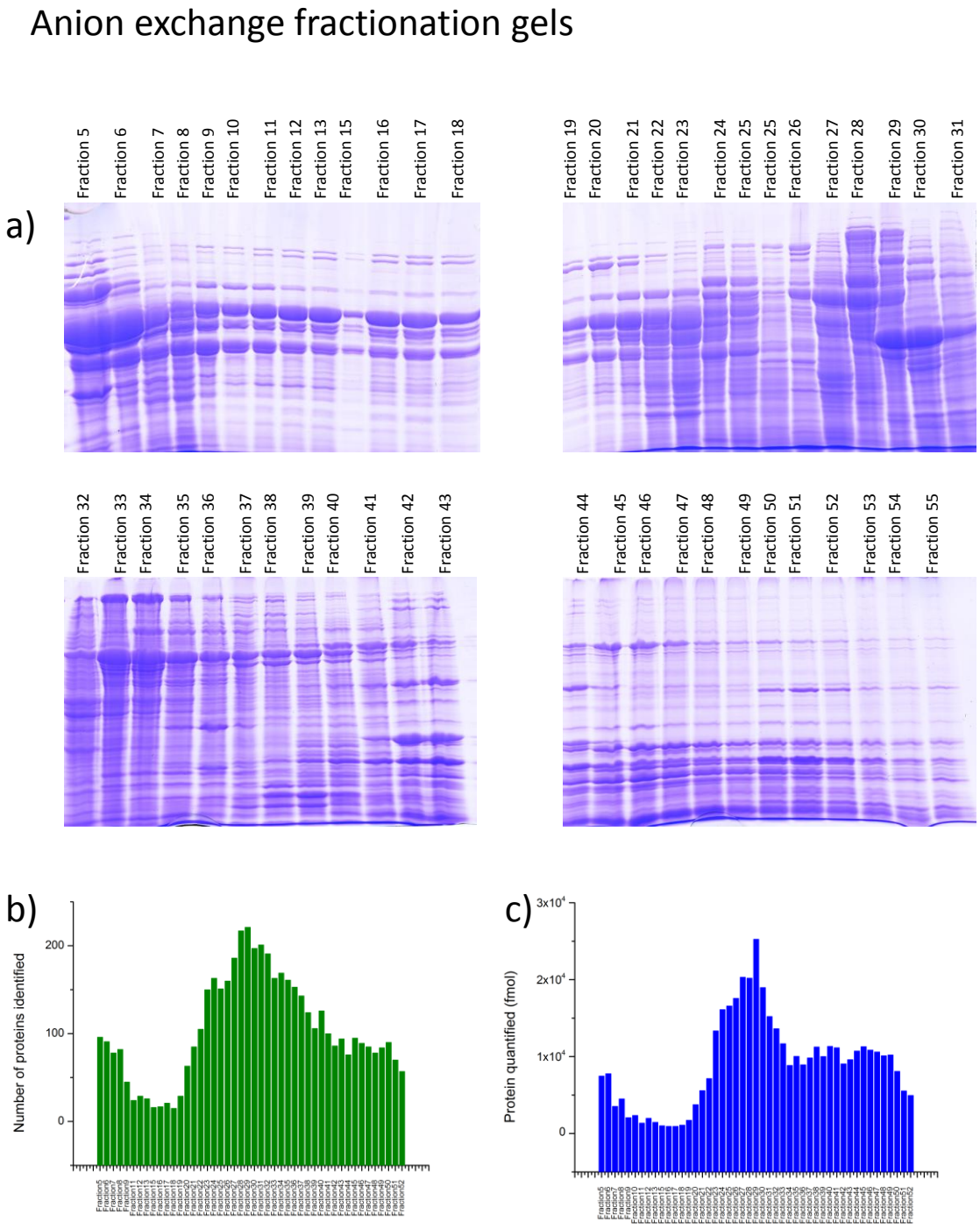


Figure 3.2 Protein fractionation by anion exchange results

Protein content of the 50 anion exchange fractions was analysed a) by SDS-PAGE, where the fractions have been concentrated using 20 µl of StrataClean resin, mixed 1:1 with sample buffer, and run on 1D SDS PAGE, and in b) the total proteins identified per fraction and c) the fmol per fraction calculated by PLGS using label free analysis, FDR 4%, after tryptic digestion and running 1 µl (from a 200 µl digest) on the Waters Synapt G2 Q-TOF mass spectrometer.

and it has been separated from the material that has bound. The material will therefore be used in the analysis, to add to the overall number of proteins quantified. The SDS PAGE also shows that many proteins are eluting across a wide range of fractions, indicating they show a broad elution from the column, which may be due to the long gradient length. The protein content of the fractions was analysed by label free analysis using the Waters Synapt G2. Each fraction was treated with StrataClean resin and digested as described in section 2.6, before spiking in glycogen phosphorylase B to use as a standard protein prior to loading onto the mass spectrometer. The proteins were quantified by label free analysis using PLGS software, and a total of 781 proteins were identified (FDR 4%) across all 52 fractions. The number of proteins identified in each fraction reach a peak at approximately 100-200 hits in fractions 20-40, while the total amount of material identified peaks at 1000-2000 fmol between fractions 22-32 (figures 3.2 b and c).

The amount (in fmol) of each individual protein eluted in each fraction is plotted as a heat map in MulitExperiment Viewer (MeV) software (Saeed *et al.*, 2003), after undergoing hierarchical clustering (HCL) by Pearson correlation. This highlights whether there is a linear correlation between two proteins, and measures if the two numbers diverge and by how much. The HCL map does show that a number of proteins are co-eluting across the fractions (figure 3.3). This could indicate that protein complex subunits are eluting together. There are situations where the clusters are formed from the subunits of a known complex, for example the 20S proteasome, which elutes over fractions 37 to 39 (figure 3.4 a). However, the heat map also supports what is indicated by the SDS PAGE, that some proteins are eluting in broad peaks over a wide range of fractions (figure 3.4 b). Higher resolution chromatography might offer increased separation between complexes, and support the hypothesis that in some cases co-eluting proteins are doing so as part of a larger assembly.

3.3.2 Mixed bed ion exchange chromatography

Fractionation was also performed using a mixed bed chromatography column, packed with a 50/50 mixture of a weak cation exchange resin (PolyCAT A^{TM}) and a

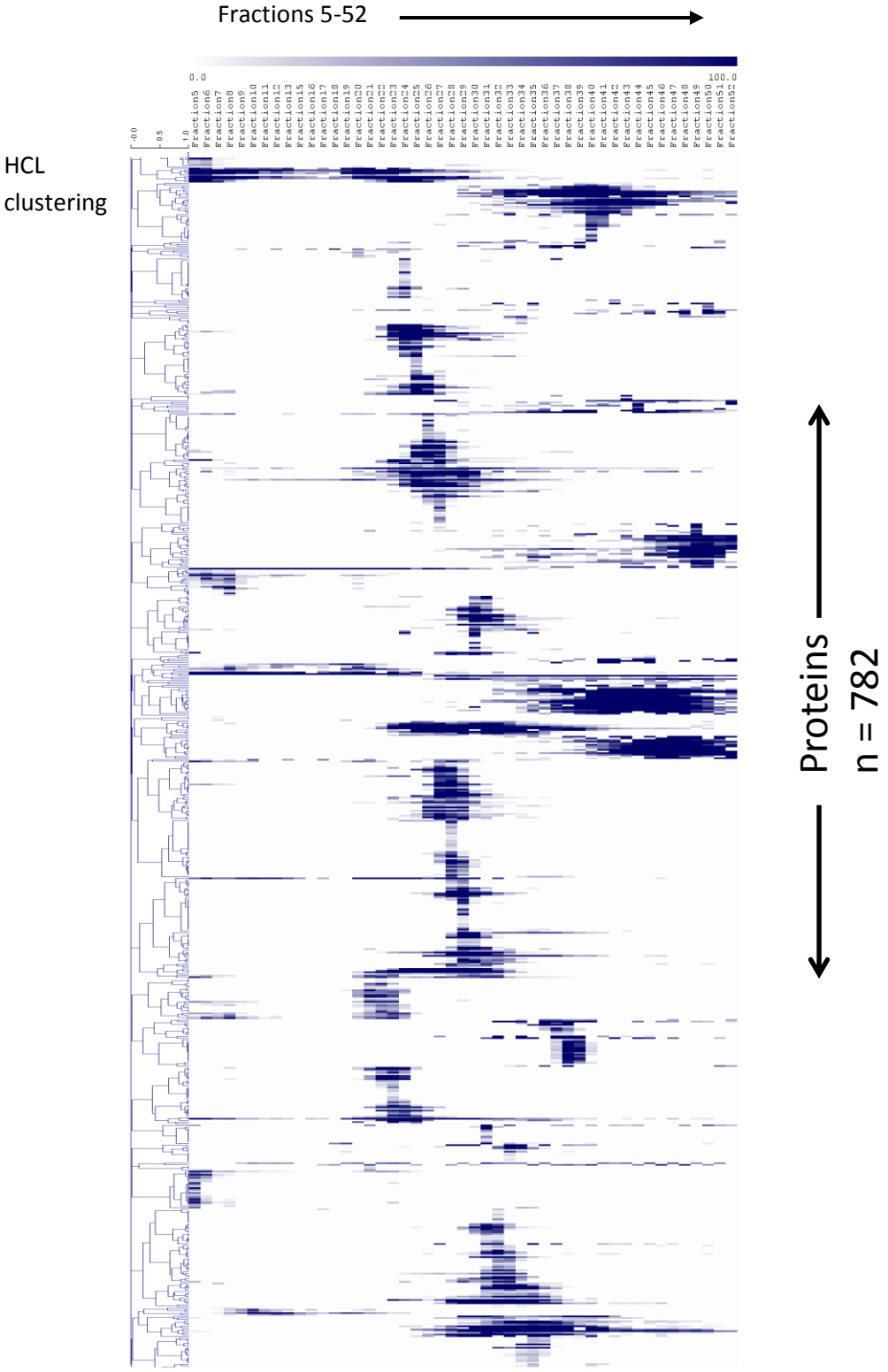


Figure 3.3 Hierarchical clustering of AEX fractionated proteins identified by LC-MS/MS

A 1 µl aliquot of a 200 µl volume digest of each fraction was run on the Waters synapt G2 and quantified by label free using PLGS software. The fmol amounts of each protein identified in each fraction were imported into Multi Experiment Viewer (MeV) software. A heat map was generated from the data using hierarchical clustering by Pearson correlation.

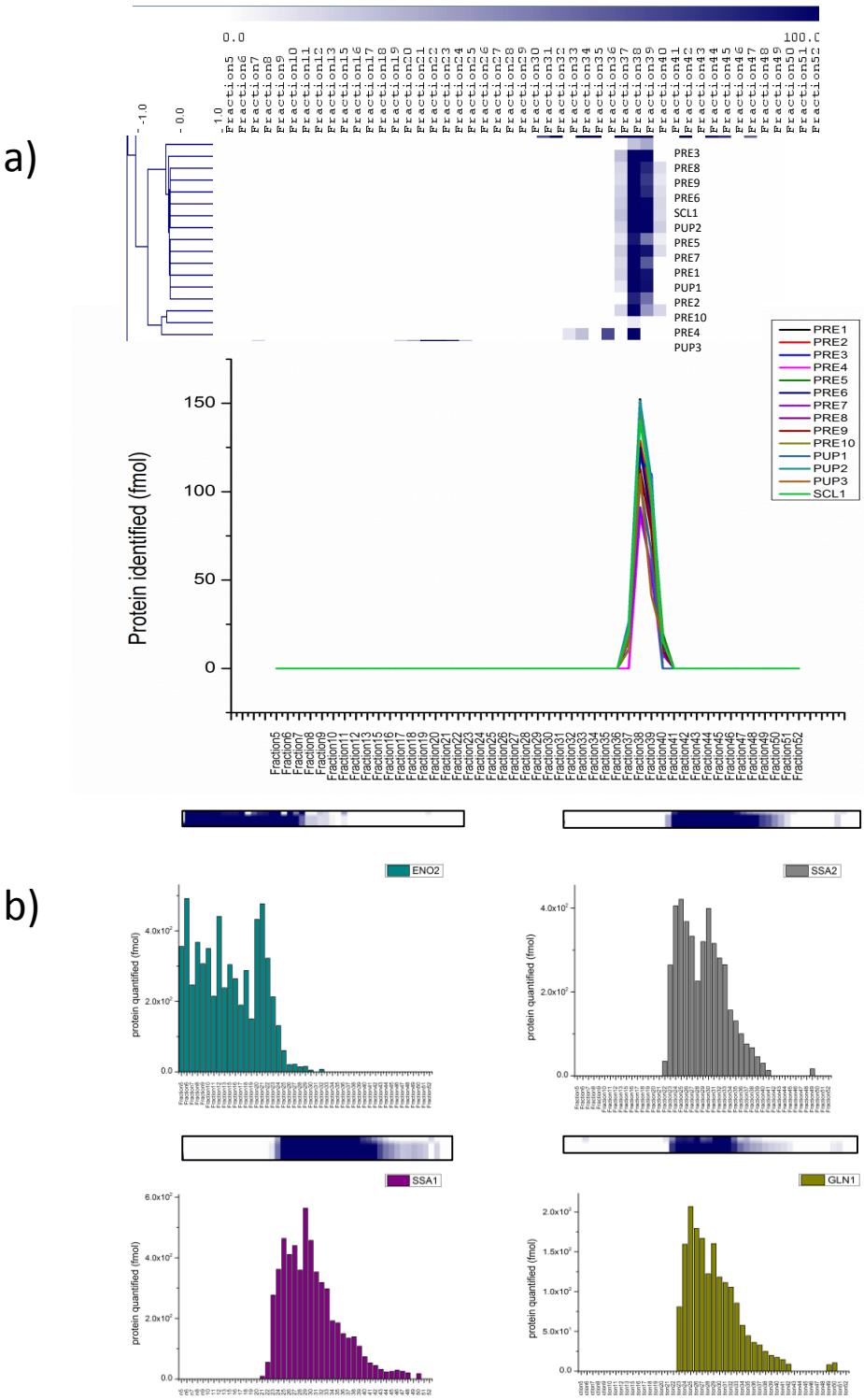


Figure 3.4 Anion exchange chromatography of the 20S proteasome and individual protein examples.

Digested fractions run on the Waters Synapt G2 are processed using PLGS. The fmol amount of each protein, arranged in tabular format, are imported into MeV to generate a heat map. Cluster analysis is performed using Pearson correlation a) 20S proteasome cluster, and the fmol amount of each 20S proteasome subunit identified, b) some individual proteins eluting across a wide range of fractions.

weak anion exchange resin (PolyWAX LP^{TM}) (PolyLC inc). Mixed bed chromatography offers the advantage of binding the majority of cellular proteins at a neutral pH, at which point there will be some proteins retaining a positive charge, some a negative charge (el Rassi & Horvath, 1986). Only those with an isoelectric point at the pH of the buffer (pH 7.4) will have a net charge of zero, and could therefore be eluted in the flow through fractions, prior to the gradient. Therefore this type of chromatography may offer the opportunity of binding the majority of protein, which may enable the separation and quantification of more protein complexes. This method offers increased resolution to the anion exchange chromatography, as all proteins should interact with the column, and mixed bed ion exchange has proved applicable in the study of proteomes (Le *et al.*, 2011; Zhang *et al.*, 2012).

The PolyLC chromatography column is smaller than the ResourceQ, with a bed volume of 2 ml, so the capacity of the column is therefore reduced, and a cell lysate dilution of 4 mg/ml was used, meaning a total load of 2 mg/ml. The smaller bed volume also means the gradient length is 0-85% B over a total of 30 fractions (figure 3.5). The chromatography appears to be at a higher resolution, with more peaks than the anion exchange. The SDS PAGE of the 30 fractions showed less material in the flow through fractions than in the AEX fractionation, however there are still some protein bands observed, indicating these proteins have an isoelectric point of 7.4, and therefore do not bind either ion exchange resin. The gels also support the idea that this is higher resolution chromatography, as individual protein bands appear in fewer fractions than in the anion exchange chromatography gels.

The fractions were concentrated with StrataClean resin, digested and analysed by label free quantification using a Waters Synapt G2 followed by PLGS, and by this process the mixed bed ion exchange fractions yielded 791 proteins. The number of proteins identified per fraction peaked at 100-200 proteins between fractions 10-20, and the amount of material identified peaked at 1000-2000 fmol in fractions 10-20. While there was a similar amount of proteins identified to the anion exchange chromatography, the fractionation was achieved over a total of 30 fractions as

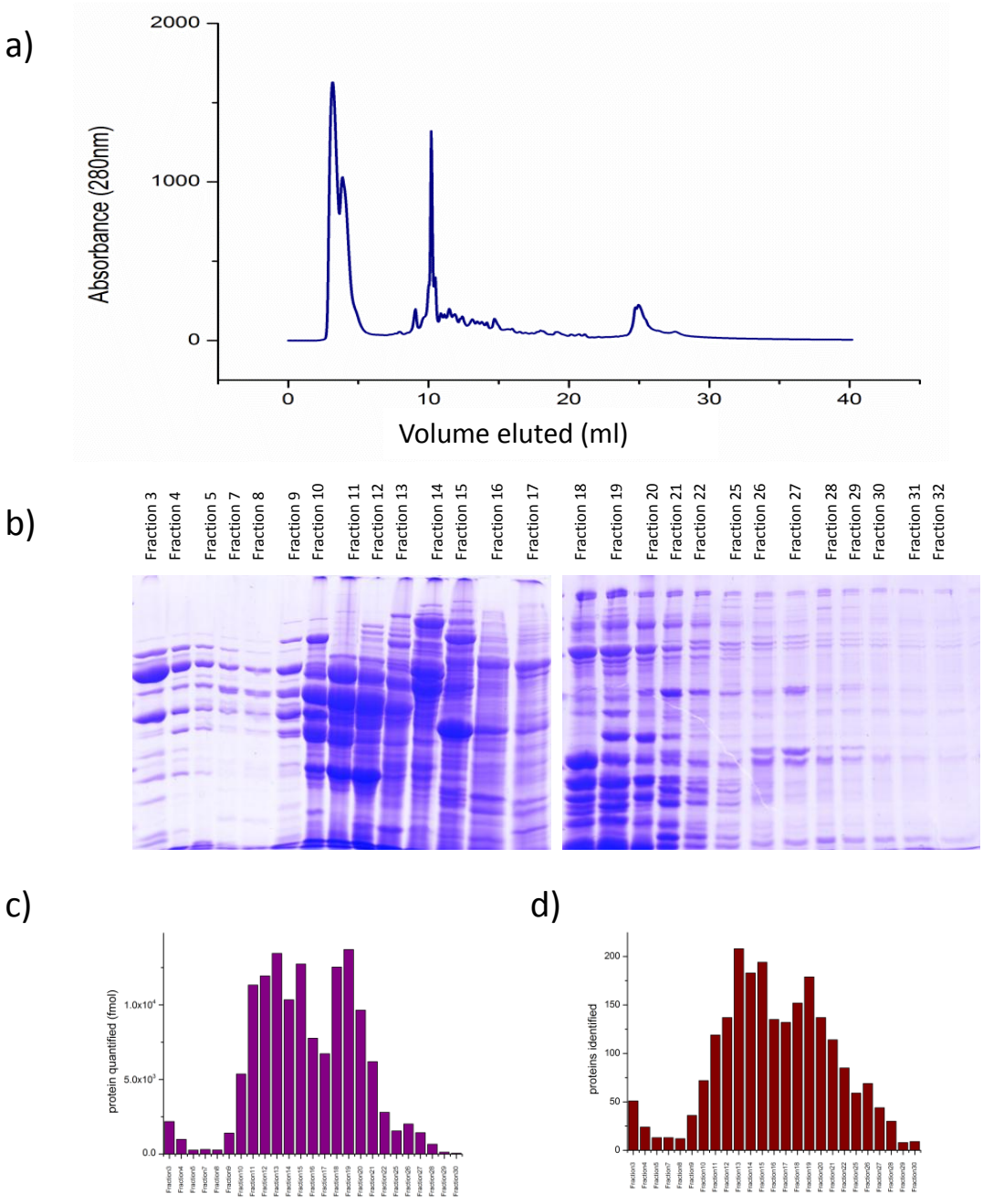


Figure 3.5 Mixed bed ion exchange chromatography results

Centrifuged cell lysate was diluted to 4 mg/ml before loading 0.5 ml onto a PolyLC mixed bed chromatography column and a) separation with a gradient of 0-85% 1M NaCl. b) The 30 x 1 ml fractions were concentrated using Strataclean resin and run on SDS PAGE. A second set of fractions were concentrated, digested and run on the Synapt G2, with label free quantification by PLGS (FDR 4%). c) the total material quantified, and d) the number of proteins identified.

opposed to 50. The higher resolution of the mixed bed chromatography was further supported when the data was plotted as a heat map (figure 3.6). The hierarchical clustering showed that proteins were still eluting as part of complexes, but in sharper chromatographic peaks. The 20S proteasome elutes sharply in one fraction, compared to across two fractions in the anion exchange chromatography (figure 3.2 a). A comparison of some individual proteins also indicates they elute across a narrower range of fractions, ENO2, SSA1 and 2 and GLN1 all show narrower peaks in comparison to the anion exchange (figure 3.7 b). However, despite the improvement in resolution when using this method of chromatography, the SSA1 and SSA2 proteins still elute over a wide range of fractions, approximately ten fractions, a third of the gradient. This indicates the protein has a higher affinity for the column, possibly due to having more potential binding sites. One possibility is the protein exhibiting interactions with other proteins, which results in more possible binding sites. SSA1 and SSA2 are chaperones of the HSP70 family, functioning in binding other proteins and assisting folding, and would therefore have a propensity for binding other proteins, possibly in various conformations. According to the IntAct database, 2816 interactions have been reported for SSA1.

A direct comparison of some of the protein complexes observed indicates the resolution was improved in mixed bed chromatography. ARC1 protein binds both methionyl and glutamyl tRNA synthetases, and the three proteins co-elute in both types of chromatography (figure 3.8 a). The three proteins are spread across fewer fractions when mixed bed chromatography is used. The same effect is seen in succinyl CoA ligase (figure 3.8 b) and the nascent polypeptide associated complex (figure 3.8 c). Despite the improved chromatography, there are a similar number of proteins identified, however the mixed bed chromatography offers the potential for increased resolution and was used in all further fractionation experiments.

3.3.3 Mixed bed chromatography replicates

Five separate replicates of the mixed bed chromatography fractionation were then run using separate cell lysates generated from five individual yeast cultures. The

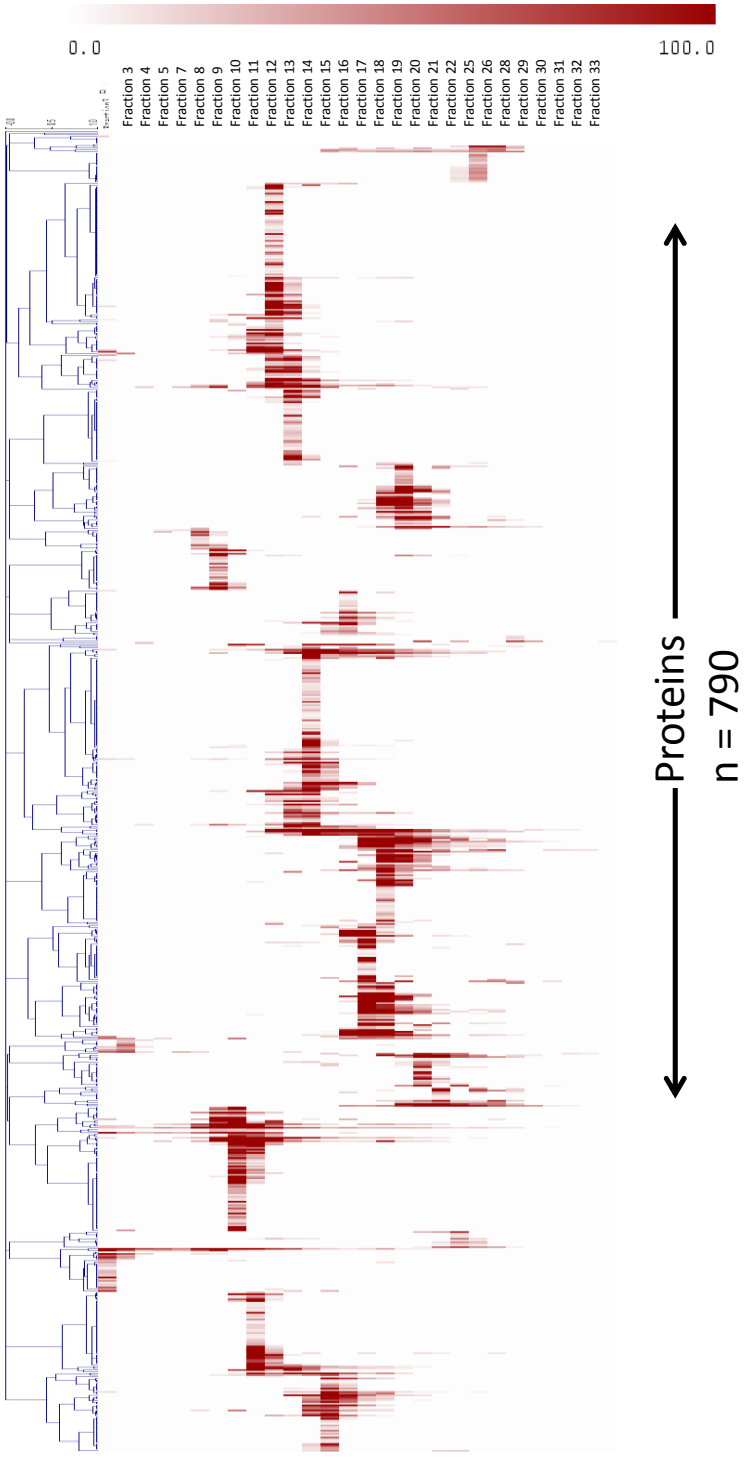


Figure 3.6 Hierarchical clustering of mixed bed ion exchange fractionated proteins.

The fraction digests run on the Waters Synapt G2 were analysed using label-free quantification by PLGS software. The fmol amount of each protein identified in each fraction was imported into Multi Experiment Viewer (MeV) software. A heat map was generated from the data using hierarchical clustering by Pearson correlation.

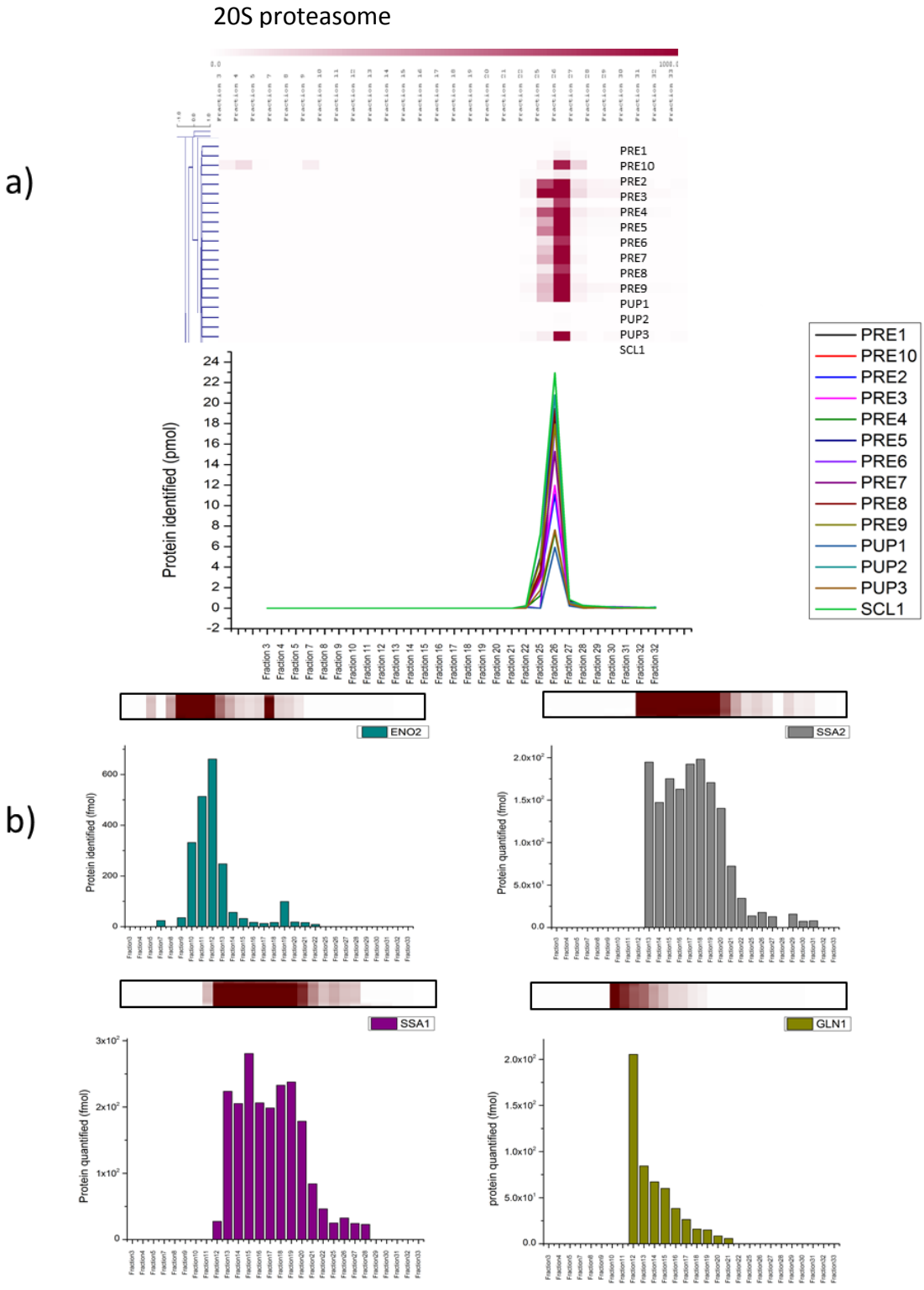


Figure 3.7 Mixed bed ion exchange chromatography of the 20S proteasome and individual protein examples

Digested fractions were run on the Waters Synapt G2 and processed using PLGS. The fmol amount of each protein were arranged in tabular format and imported into MeV for the generation of a heat map. Cluster analysis was performed using Pearson correlation a) 20S proteasome cluster, and the fmol amount of each 20S proteasome subunit identified by PLGS, b) the elution of some individual proteins.

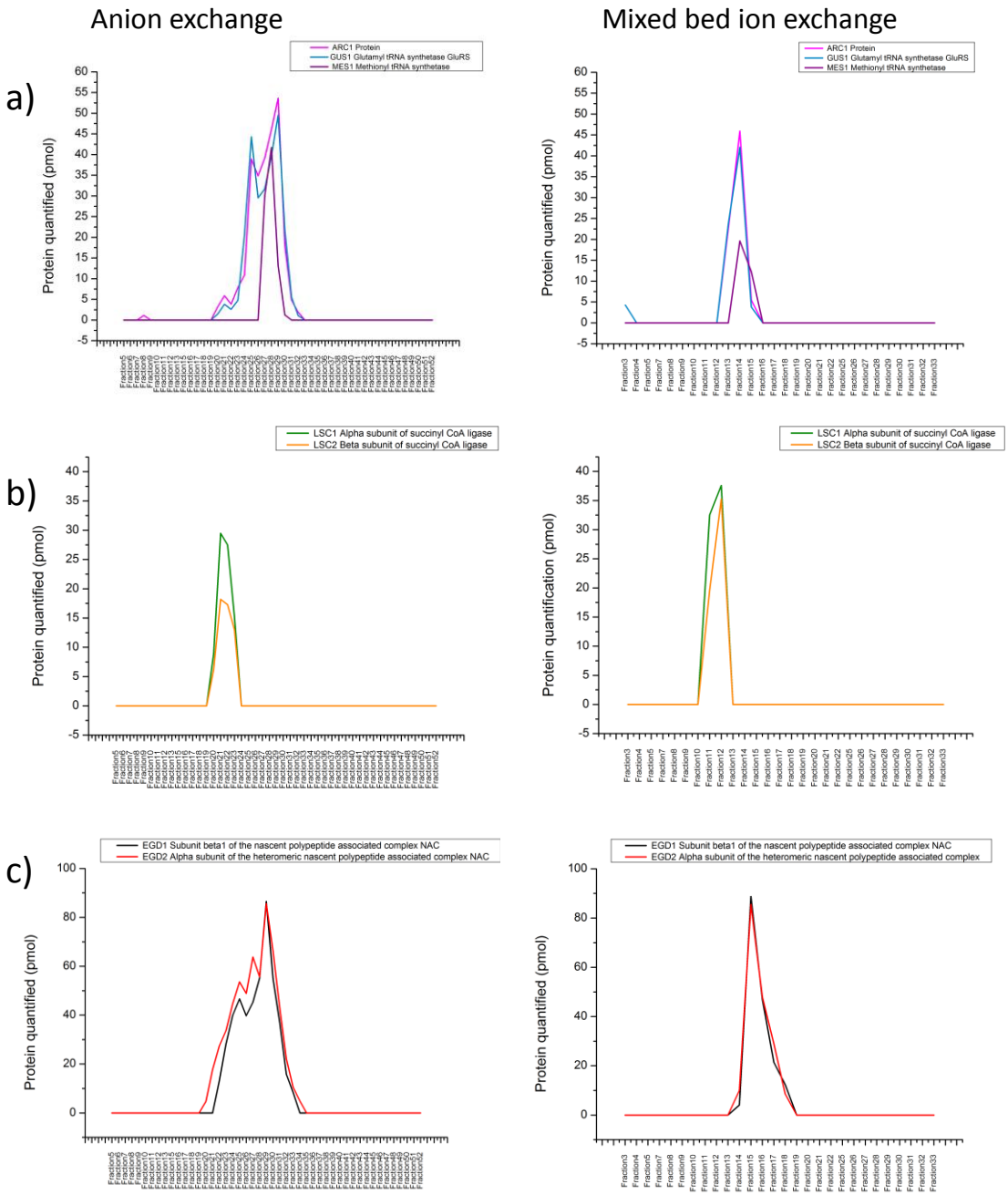


Figure 3.8 Protein complex chromatography by anion and mixed bed ion exchange

Both anion and mixed bed fractions were concentrated using strataclean resin, and digested using trypsin, and 1 µl of a 200 µl total volume digest was run on the Waters Synapt G2 on a 2 hour HDMSe method. The resulting data was analysed usign PLGS, and label free quantification was performed. Results are compared for a) ARC1/GUS1/MES1 complex, b) succinyl CoA ligase, and c) the NAC complex.

five replicate yeast cultures were grown at five different times, cells were lysed by bead beating and then the protein content was calculated using a coomassie plus protein assay. The cell lysates were each diluted to 4 mg/ml, loaded onto the mixed bed column and separated using the same 10 column volume, 0-85% B gradient as above. Fractions 3-30 from each replicate were digested, run on both the Waters Synapt G2 and the Thermo QExactive and quantified by label free analysis. The label free quantification from both instruments gives a result in fmol injected, which is then corrected for digestion volume. The sum of the amount of protein in each fraction would therefore give the total amount of that protein in the original cell lysate.

The chromatograms generated by the AKTA show some differences in the profile between the five replicates, though much of the curves follow the same pattern, with a number of discrete peaks eluting in fractions 3 to 33 across the length of the gradient (figure 3.9). The differences between the chromatograms are likely due to variation between the five cultures and preparation of separate dilutions. SDS PAGE analysis of the five replicates showed a good correlation, and the fractions in each replicate appeared to contain similar amounts of protein bands, despite the amount of material varying between replicates (figure 3.10). To quantify this protein, the five replicates were run on the Synapt and analysed using PLGS to give a total of 786, 790, 543, 692 and 822 hits. As the number of hits varies, the amount of material identified does, with a total of 136, 117, 57, 70 and 77 pmol quantified by PLGS in the injected material in replicates 1 – 5 (figure 3.11). As expected, there are differences between these replicates as there are a number of points where variation can occur, due to them originating from five individual cultures, and cell lysis of cell pellets on five separate occasions, which are quantified by protein assay, with separate dilutions prepared. However, variability in the amount of protein identified did not prevent the primary goal in this analysis; to assess the fractionation of protein complexes.

The five replicates were also run on the QExactive, a quadrupole ion trap mass spectrometer, with data analysis using MaxQuant (V 1.3.0.5) software, using a

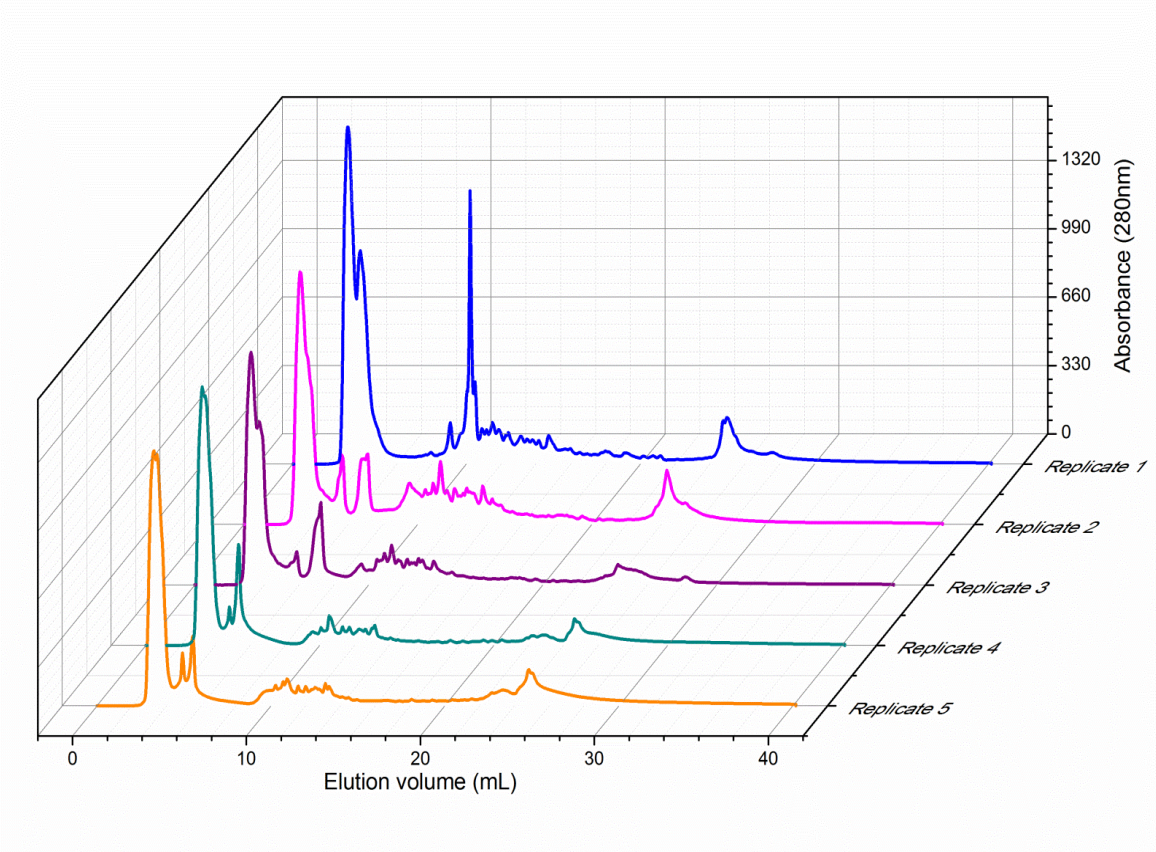


Figure 3.9 Five replicates of *S. cerevisiae* cell lysate mixed bed ion exchange fractionation

Five separate cultures of *S. cerevisiae* were prepared on five separate occasions. Cell pellets generated from 15 ml of culture were used to prepare a cell lysate by bead beating. Lysates were clarified by centrifugation and diluted to 4 mg/ml before loading 0.5 ml onto a PolyLC mixed bed chromatography column and separating on a gradient of 0-0.85 M NaCl. 30 x 1 ml fractioned were collected.

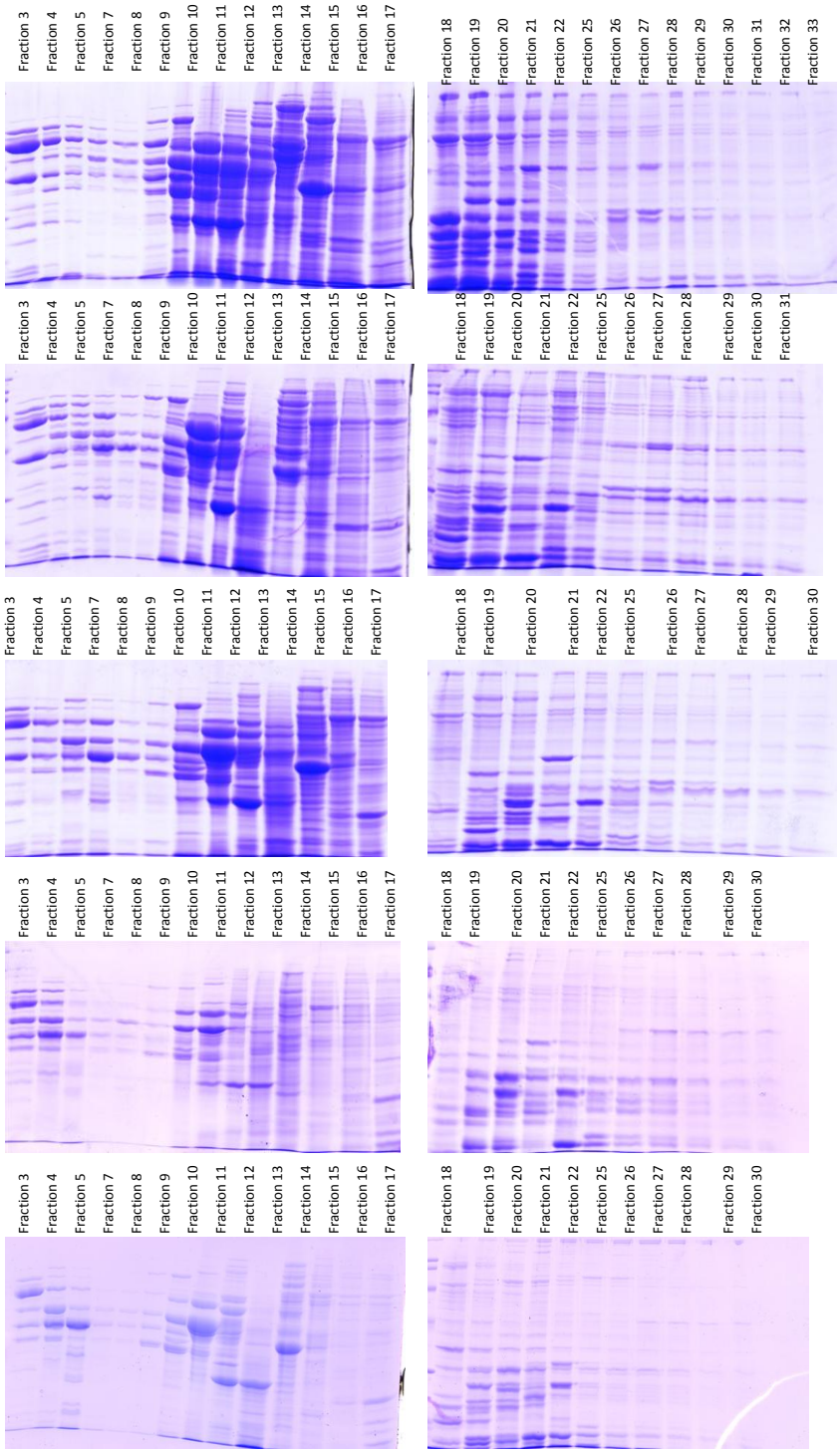


Figure 3.10 SDS PAGE of five mixed bed ion exchange replicates

S. cerevisiae cell pellets generated from 15 ml of culture were used to prepare a cell lysate by bead beating. Lysates were clarified by centrifugation and diluted to 4 mg/ml before loading 0.5 ml onto a PolyLC mixed bed chromatography column and separating on a gradient of 0-0.85 M NaCl. 30 x 1 ml fractioned were collected, concentrated via StrataClean resin and run on SDS PAGE.

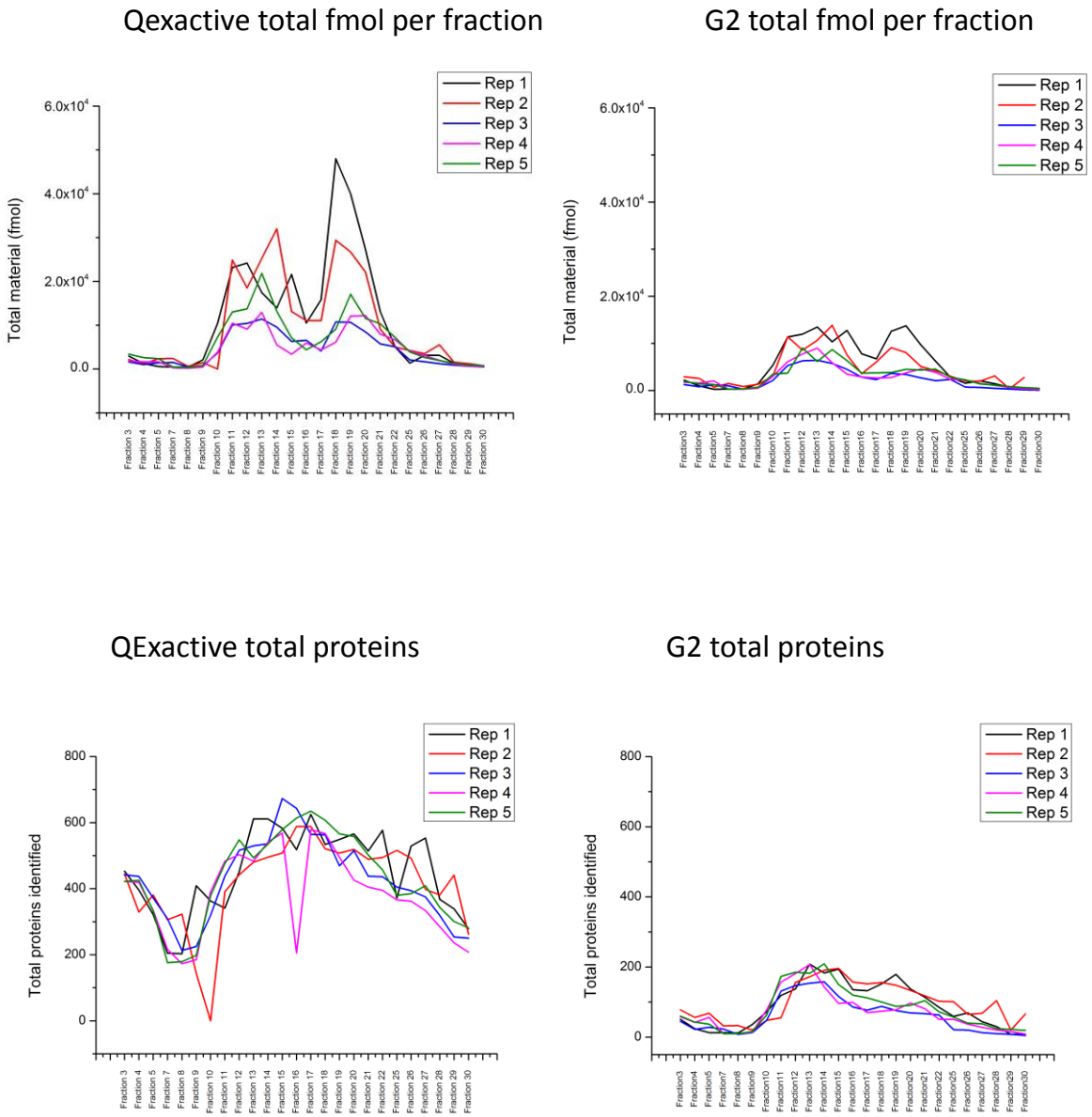


Figure 3.11 LC-MS/MS results of mixed bed ion exchange fractions

The 30 x 1 ml fractions generated from mixed bed ion exchange chromatography were concentrated via StrataClean resin, digested and run on the Synapt G2 or QExactive. Label free quantification was done using PLGS (FDR 4%), and MaxQuant software (FDR 1%). Results are reported as a) total fmol of material and b) total number of hits in each fraction in the G2 data, c) total fmol material and d) total hits in each fraction in QExactive data.

default setting of 1% FDR (Cox & Mann, 2008). Label free quantification was achieved by calculating the iBAQ intensity using MaxQuant and calculating the fmol of each protein by comparing the intensity of the 50 fmol of glycogen phosphorylase B. In replicates 1-5 respectively there were 1868, 1810, 1636, 1574 and 1746 proteins identified (figure 3.11). The mode of function of the two instruments, and the data processing are not directly comparable, however the QExactive yielded more than double the hits of the G2 for each replicate. Within these replicates there were 1241 common proteins identified on the QExactive, and a total of 2282 proteins identified at least once when all replicates are combined while in the G2 data there were 480 common proteins and 1051 in total. There were also different amounts of protein in each replicate, and the total amount of material identified by label free varies from 38 nmol in replicate 4 to 110 nmol in replicate 5. As the cell lysate used for each replicate was derived from a different batch of yeast culture a certain amount of variation in the protein content would be expected. However, the total protein load onto the chromatography column was calculated at 2 mg in each run. There are a number of potential sources of this error, either prior to fractionation, in the protein assay of the cell lysate, or the sample dilution, or error could also come from the pipetting of the StrataClean resin. Despite the difference in protein content of the replicates however, a comparison can still be made of the results of individual proteins, as the protein load is sufficient to observe individual protein behaviour in each replicate. Also, the protein in each replicate follows the same pattern, with the majority of the material for each replicate eluting in fractions 10-15.

3.3.4 Chromatographic effects on abundant glycolytic proteins

Some of the most abundant proteins in the yeast cellular proteome are the enzymes involved in the glycolysis pathway. These proteins are involved in the conversion of glucose into pyruvate, regenerating ATP. It has been suggested that they associate into a multi-enzymatic complex, or metabolon (Araiza-Olivera *et al.*; Araiza-Olivera *et al.*, 2013). Some of the enzymes, including three isozymes of glyceraldehydes-3-phosphate dehydrogenase (TDH1-3), four isozymes of

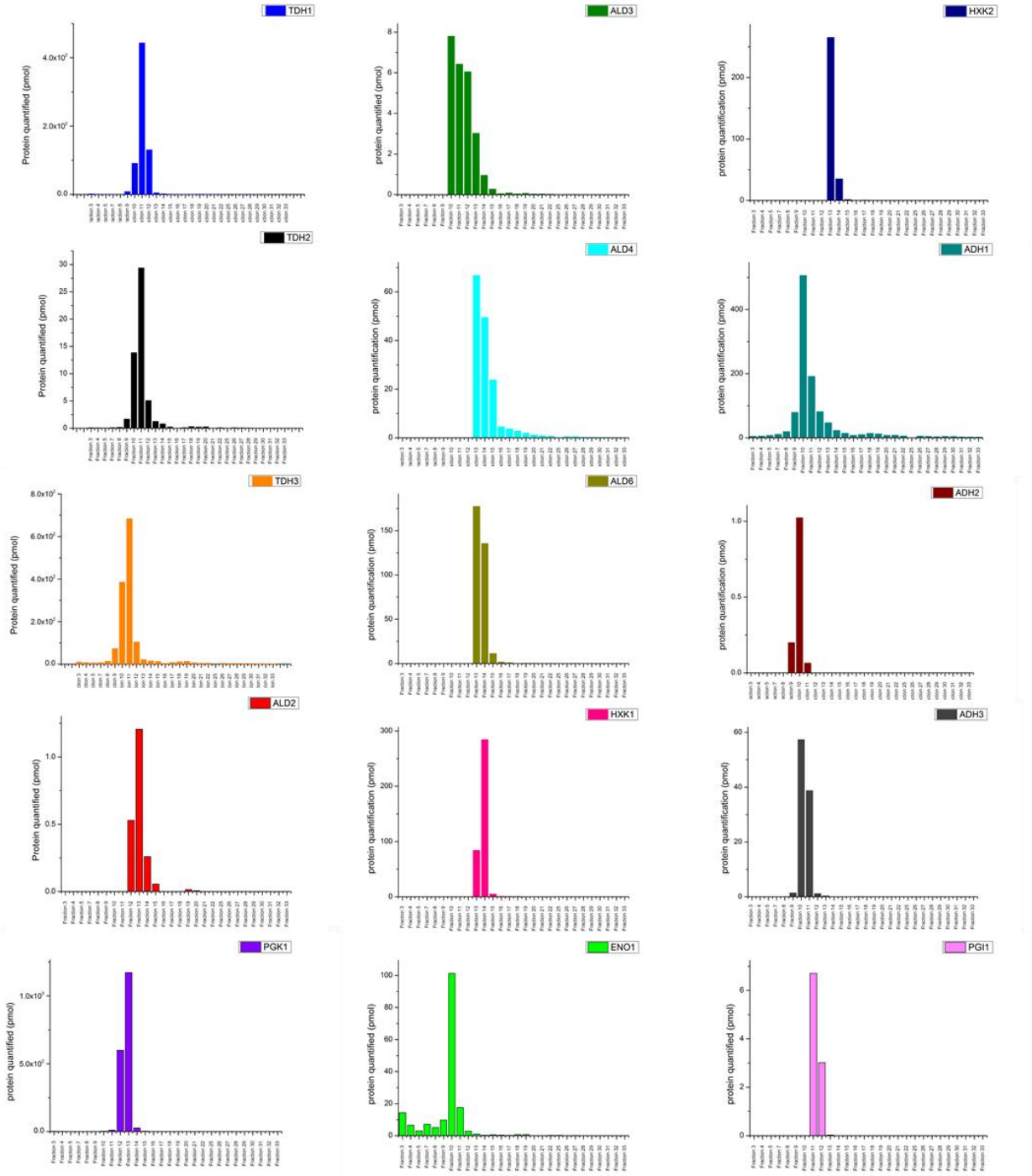


Figure 3.12 Chromatographic elution of glycolytic enzymes.

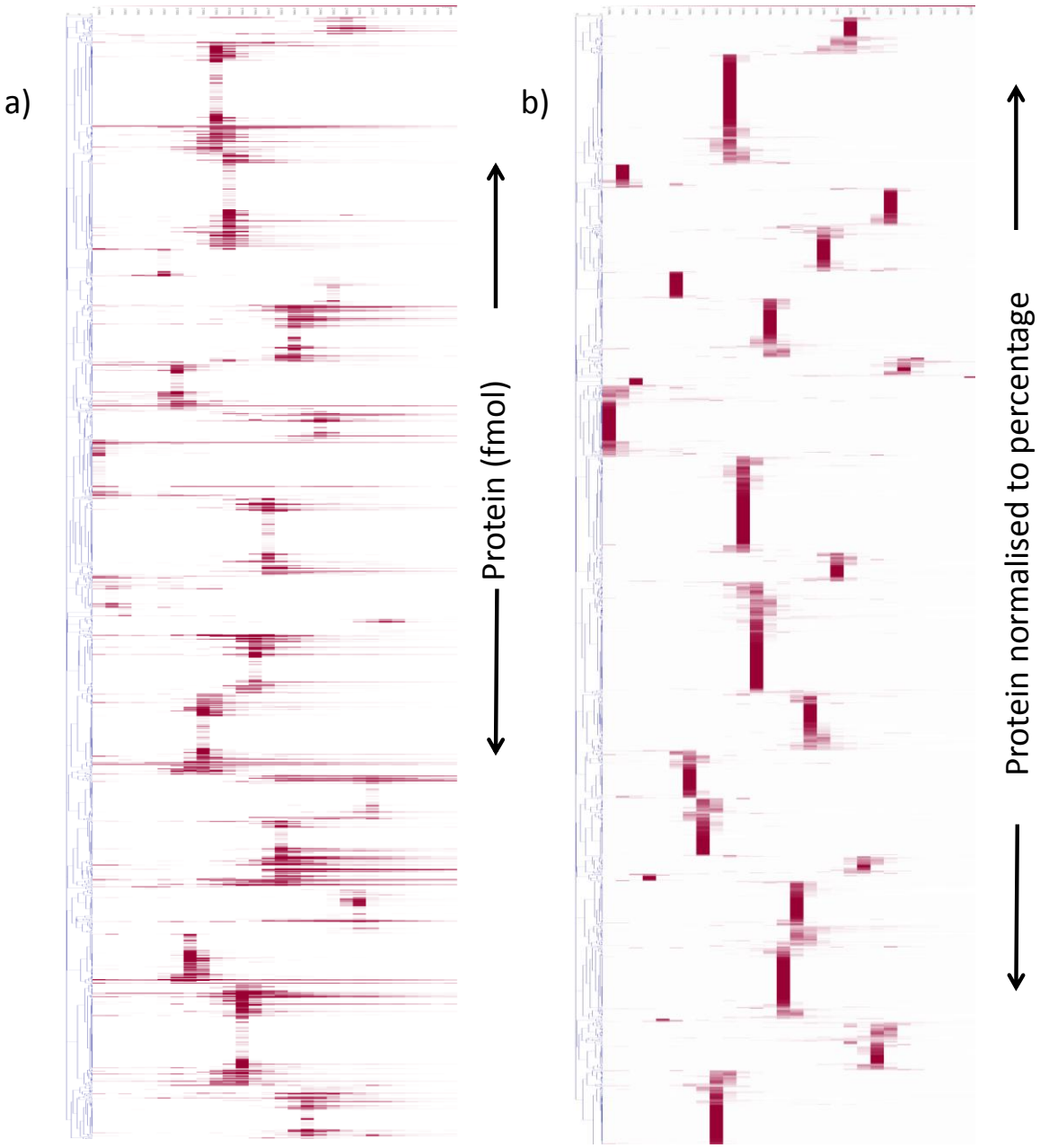
The *S. cerevisiae* cell lysate, 4 mg/ml dilution, loaded onto mixed bed ion exchange and fractionated on a gradient of 0-85% 1M NaCL. The chromatographic elution profile is shown of some of the enzymes of the glycolytic pathway.

enolase (EN01-2), phosphoglycerate kinase (PGK1), phosphoglucose isomerise (PGI1), enolase 1 (ENO1), two isozymes of hexokinase (HXK 1-2) and three of alcohol dehydrogenase (ADH 1-3) elute between fractions 8 and 14 (figure 3.12). The isozymes of a protein possess similar structures, and may elute in similar position due to this similarity, however the co-elution of many different proteins may indicate some level of interaction.

3.3.5 Investigating protein interactions

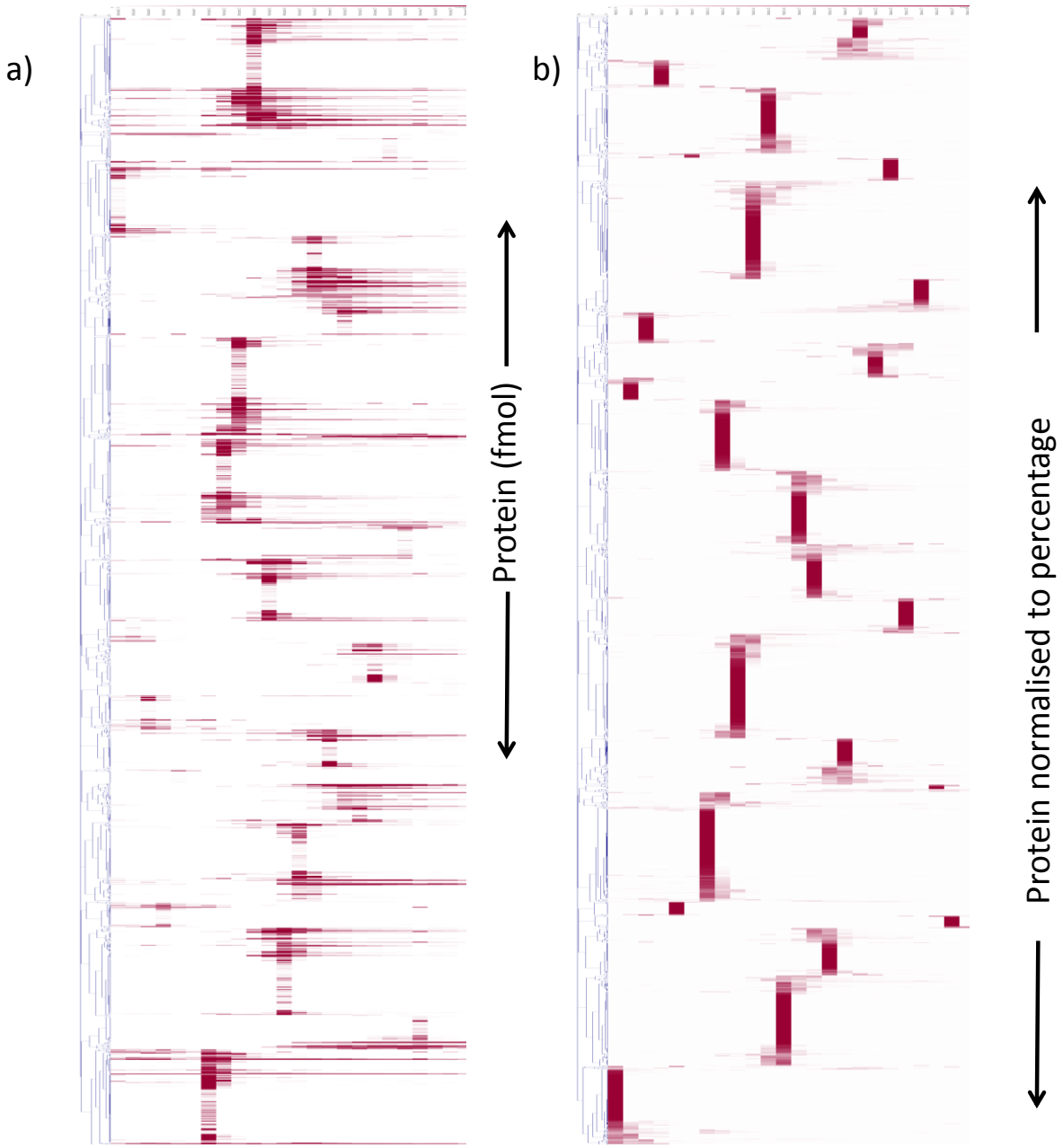
Protein complexes in the yeast cellular environment are abundant and varied, and hundreds of different complexes are reported in the literature, and assembled into databases such as MIPS, BIOGRID, and CYC2008 (Guldener *et al.*, 2006; Pu *et al.*, 2009; Chatr-Aryamontri *et al.*, 2013). They vary in size and structure from just 2 subunits to more than 70 in large organelles such as ribosomes. It is therefore expected that a variety of different elution profiles will be observed. To provide a visual representation of the elution of proteins across all 5 replicates heat maps have been generated in MeV. These show the fmol amounts of each protein, and for each replicate the normalised data is also displayed as a heat map. To normalise, the amount of an individual protein in each fraction has been calculated as a percentage of the total amount of that protein identified across all of the fractions. Hierarchical clustering was performed using Pearson correlation to arrange the results according to co-eluting proteins. The normalised data for each replicate shows much tighter clustering, which indicates the fmol amount of protein in each cluster varies, although the same total percentage is eluted. The elution of the proteins in these tight clusters was repeated across all five replicates (figures 3.13 to 3.17). Analysis of the clusters showed that a number of known protein complex subunits are eluting together.

One such example of a complex where all subunits are clustering together is the LSm (like-sm) proteins assembly. LSm are RNA binding proteins which are assembled into ring structures (Zhou *et al.*, 2014), and there are eight proteins in eukaryotes, all of which are identified in the QExactive data. All of the subunits elute in fraction 17 (figure 3.18), however the cluster analysis groups them into two



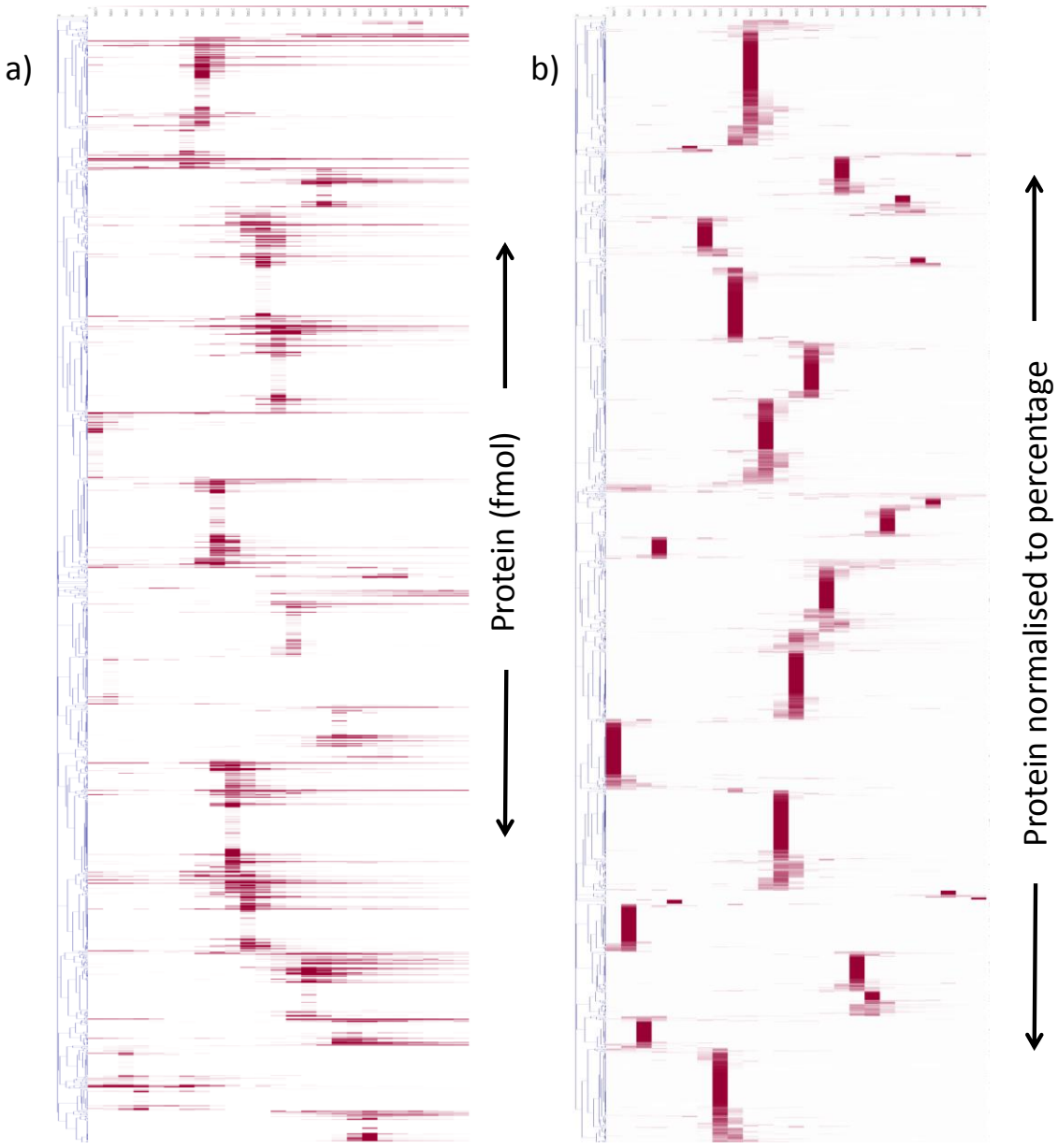
3.13 Hierarchical clustering of mixed bed ion exchange fractionation replicate 1

The fmol of each protein in each fraction was imported into Multiple Experiment Viewer (MeV) software, and a heat map was generated of the cluster proteins. Hierarchical clustering of a) the fmol amounts was achieved using Pearson correlation. A second heat map was also generated b) of the normalised data. Normalisation was achieved by calculating the percentage of each protein found in each fraction, and a heat map generated as before.



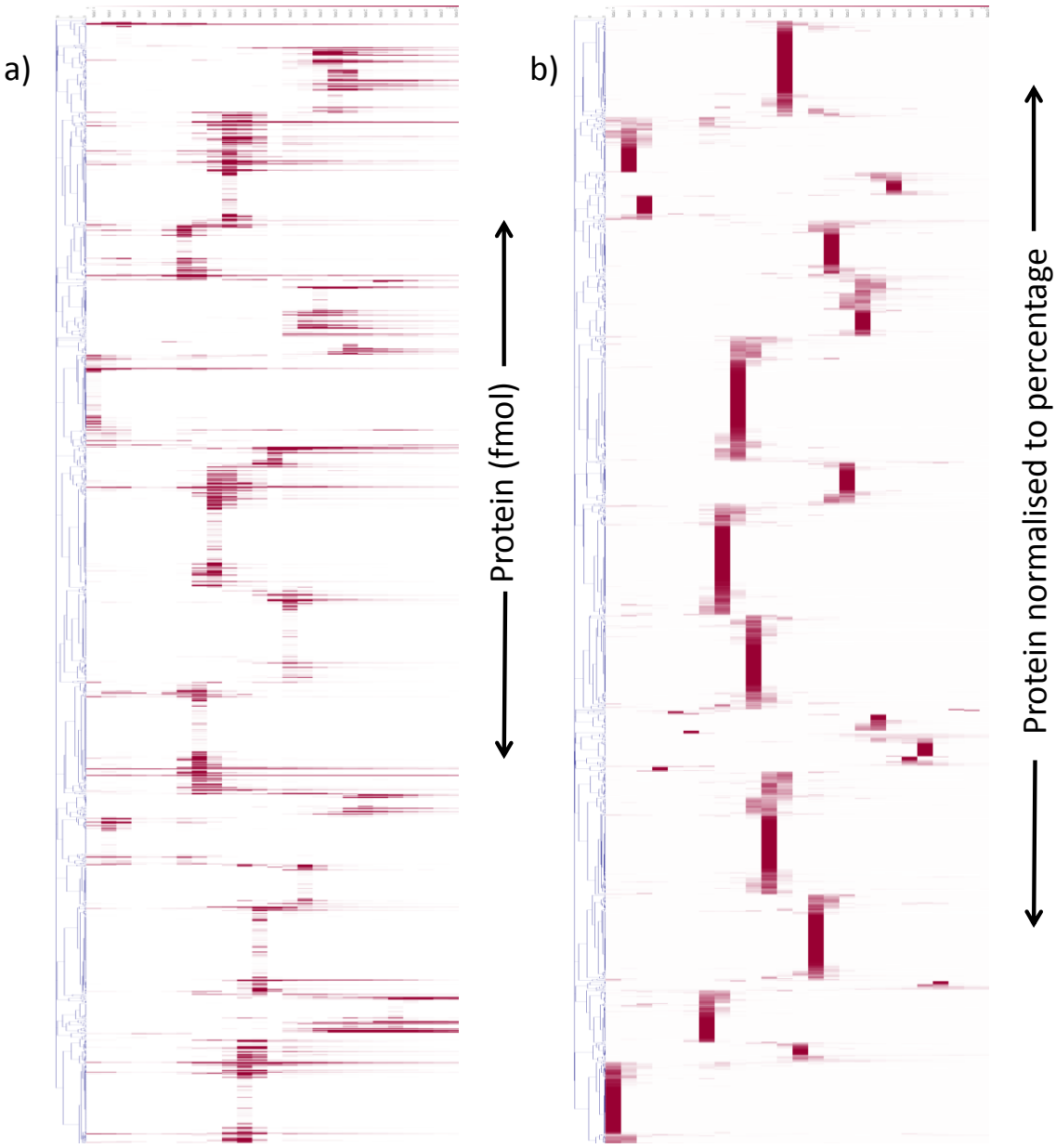
3.14 Hierarchical clustering of mixed bed ion exchange fractionation replicate 2

The fmol of each protein in each fraction was imported into Multiple Experiment Viewer (MeV) software, and a heat map was generated of the cluster proteins. Hierarchical clustering of a) the fmol amounts was achieved using Pearson correlation. A second heat map was also generated b) of the normalised data. Normalisation was achieved by calculating the percentage of each protein found in each fraction, and a heat map generated as before.



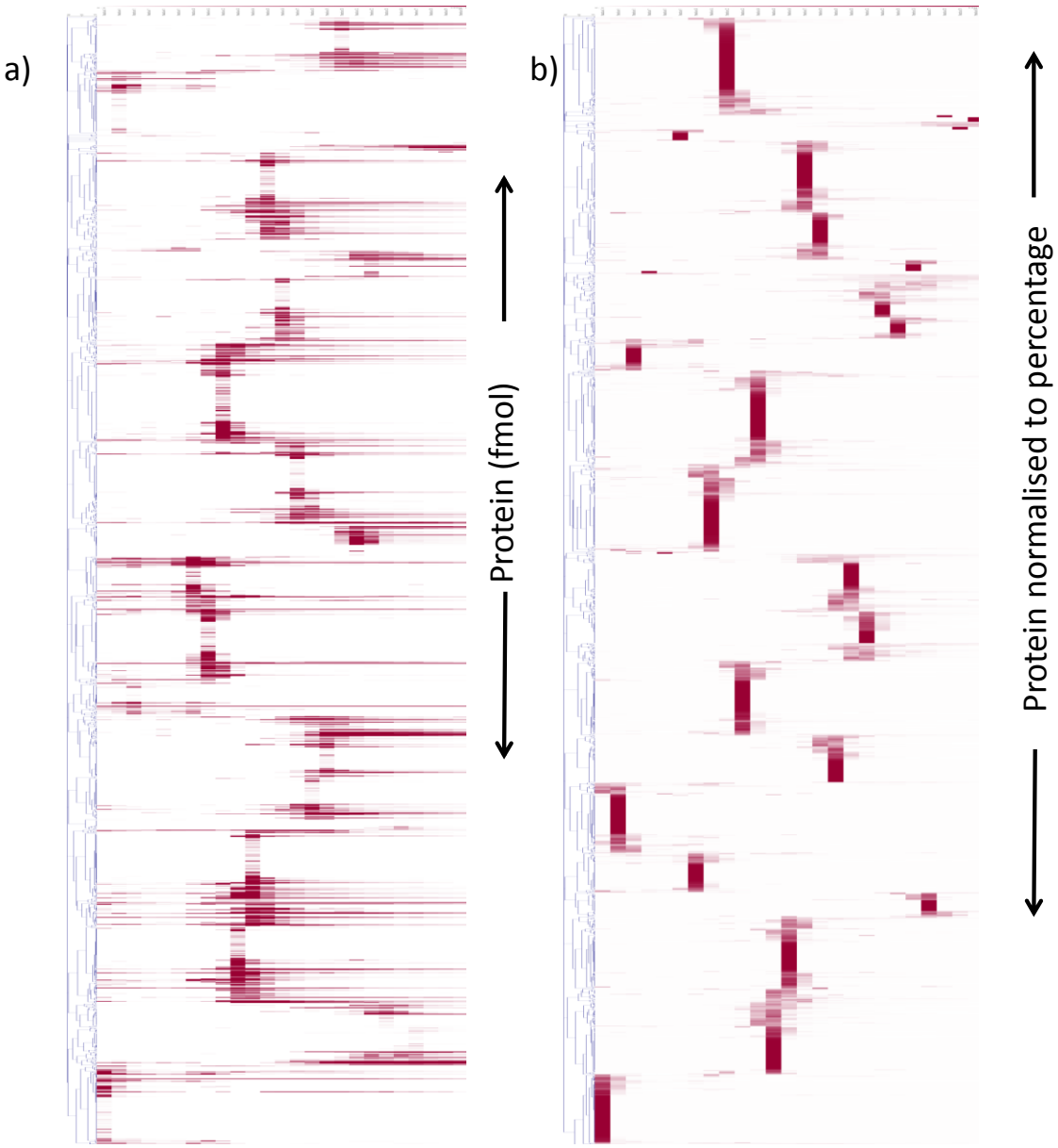
3.15 Hierarchical clustering of mixed bed ion exchange fractionation replicate 3

The fmol of each protein in each fraction was imported into Multiple Experiment Viewer (MeV) software, and a heat map was generated of the cluster proteins. Hierarchical clustering of a) the fmol amounts was achieved using Pearson correlation. A second heat map was also generated b) of the normalised data. Normalisation was achieved by calculating the percentage of each protein found in each fraction, and a heat map generated as before.



3.16 Hierarchical clustering of mixed bed ion exchange fractionation replicate 4

The fmol of each protein in each fraction was imported into Multiple Experiment Viewer (MeV) software, and a heat map was generated of the cluster proteins. Hierarchical clustering of a) the fmol amounts was achieved using Pearson correlation. A second heat map was also generated b) of the normalised data. Normalisation was achieved by calculating the percentage of each protein found in each fraction, and a heat map generated as before.



3.17 Hierarchical clustering of mixed bed ion exchange fractionation replicate 5

The fmol of each protein in each fraction was imported into Multiple Experiment Viewer (MeV) software, and a heat map was generated of the cluster proteins. Hierarchical clustering of a) the fmol amounts was achieved using Pearson correlation. A second heat map was also generated b) of the normalised data. Normalisation was achieved by calculating the percentage of each protein found in each fraction, and a heat map generated as before.

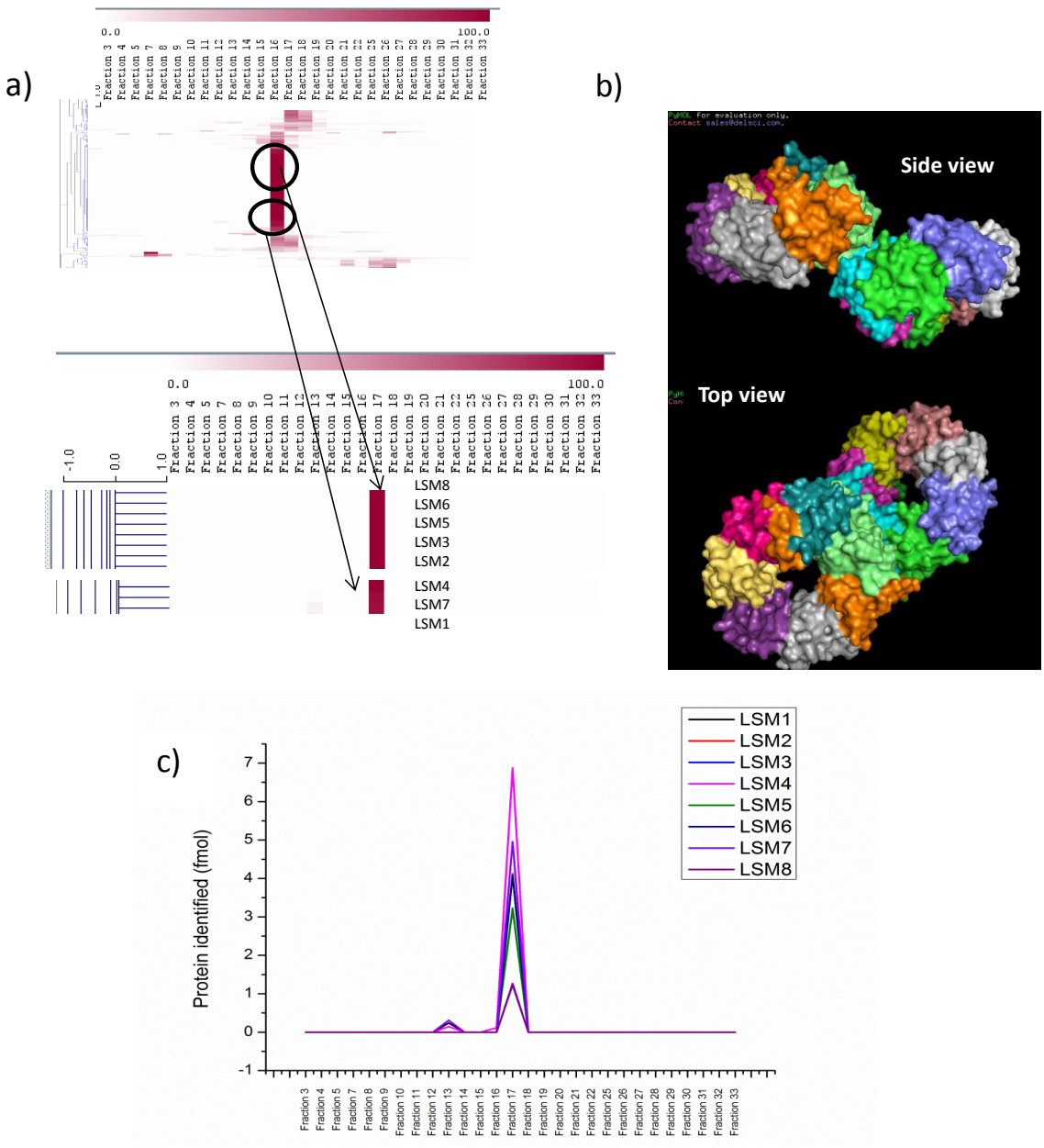


Figure 3.18 Clustering of the Like-SM (LSM) protein complex subunits

Hierarchical clustering of the mixed bed fractionation normalised to percentage protein was performed in MeV software using Pearson correlation. An example a) of the LSM complex subunits clustering in two subsets is shown, and b) the structure of the LSm complex, as shown in Zhou *et al*., 2014, and visualised using PyMOL software. c) the graphical representation of the fmol amount of each protein identified.

subsets, one group of five proteins, one smaller group of three proteins. This is because the proteins in the smaller subset, consisting of LSM1, LSM4 and LSM7, are all identified in fraction 13 in addition to fraction 17. The data from replicate 1 is shown, for other replicates see supplementary material. There are two known complexes of LSm proteins, LSM 2-8, which interacts with U6 snRNA, and LSM 1-7, associated with mRNA degradation, so it is possible that two structures are forming (Bouveret *et al.*, 2000). However, LSM1 was actually found in both locations of the gradient, in fraction 13, and also in fraction 17, where LSM8 is also identified.

Other examples of protein complexes showing sharp elution profiles and clustering together include RNA polymerase II transcription inititation factor TFIIA and the HAT1-HAT2 histone acetyltransferase complex. TFIIA consists of two subunits, TOA1, the large subunit, and TOA2, the smaller subunit. The two proteins cluster together in fraction 22, but elute over a number of fractions, indicating the complex has a high affinity for the chromatography media (figure 3.19 a & b). The chromatography of the heterodimer HAT1-HAT2 histone acetyltransferase complex is much sharper, with both subunits eluting together in fraction 22. The difference in the chromatographic profile of these two complexes is reflected in both the graphical representation, and the heat map generated (figure 3.19 c & d). Replicate 1 data is shown for these complexes, for additional replicates, see supplementary material.

Some variation in the chromatographic behaviour of the different protein complexes is to be expected, due to the structural variation in the wide range of protein assemblies identified. Many of the complexes identified show a sharp peak in the chromatography, eluting in just one fraction, for example the ARP2/3 complex, which has seven subunits eluting in fraction 15 (figure 3.20 a). Other protein-protein interactions appear to remain stable under these experimental conditions, but elute in more than one fraction, and at different times in the gradient, such as phosphofructokinase, an octameric complex consisting of four heterodimers of an α subunit, PFK1, and a β subunit, PFK2. The two subunits have the same elution profile, indicating the complex remains intact, but both have a

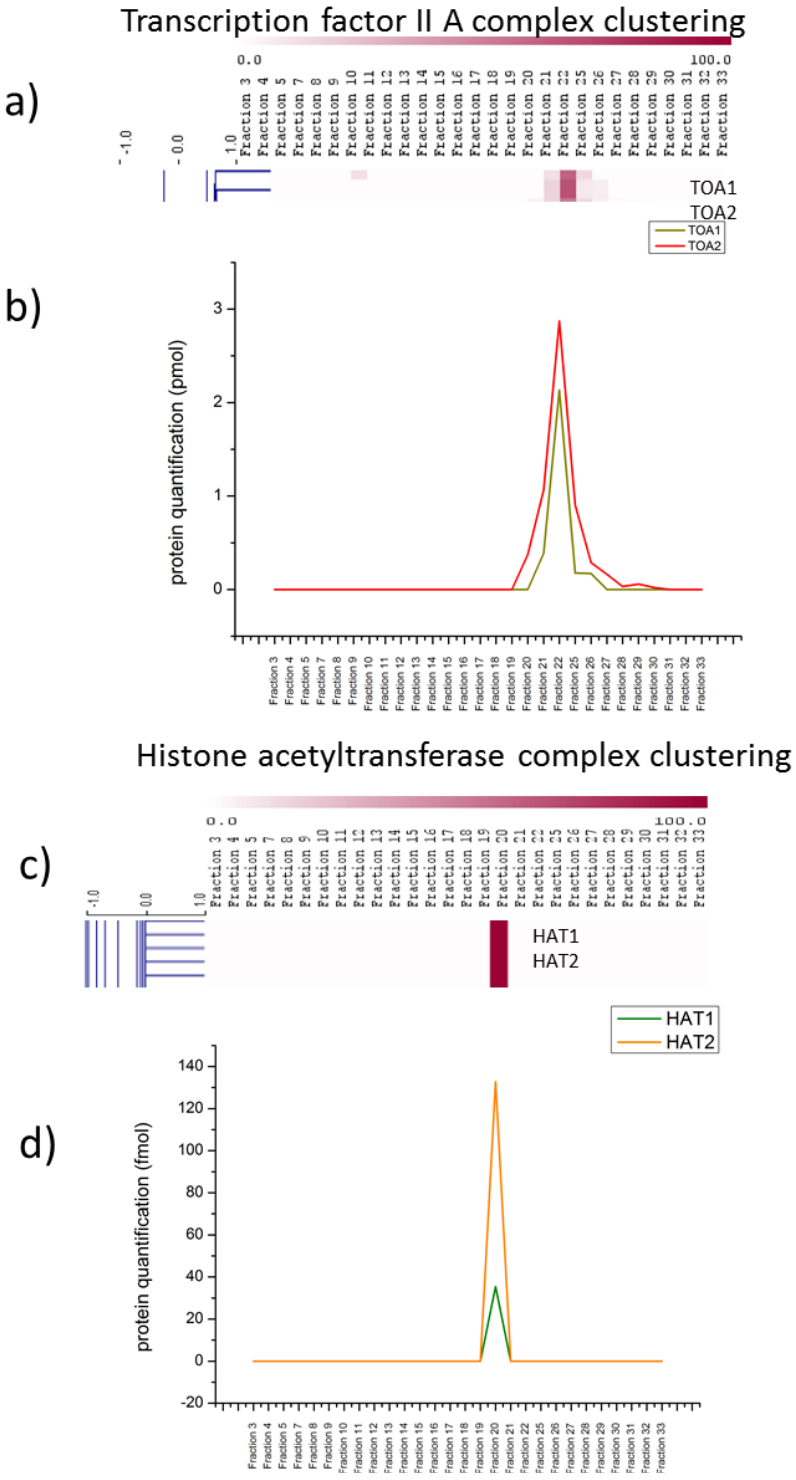


Figure 3.19 Heirarchical clustering of complexes in fractionation results

Hierarchical clustering of normalised LC-MS/MS results of the mixed bed chromatography fractions. Clustering of a) TOA1 and TOA2 subunits of transcription factor II A complex and b) the fmol amount identified, and c) the clustering of HAT1 and HAT2 subunits of the histone acetylase complex, and d) the fmol amounts of each identified in the fractions.

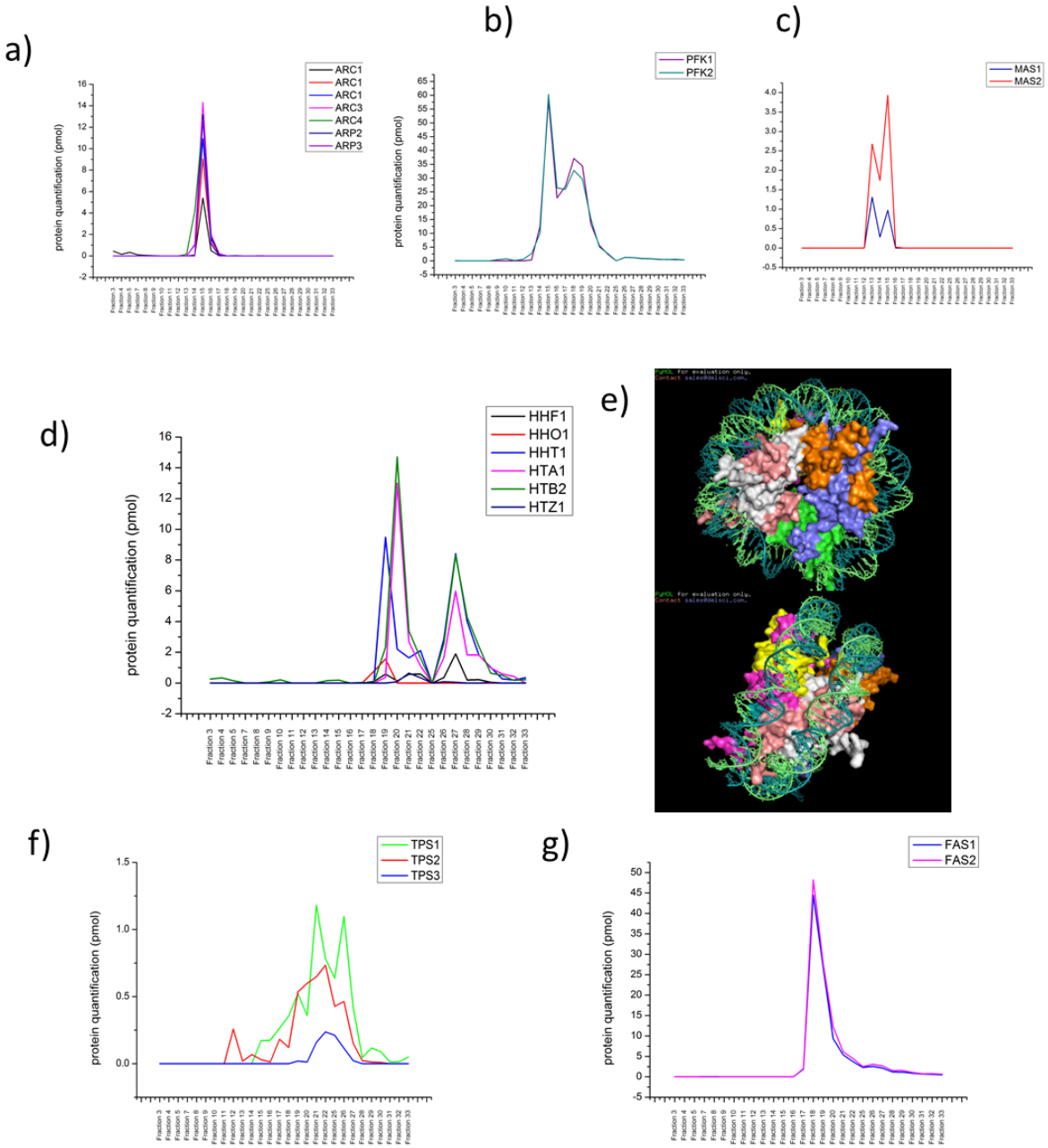


Figure 3.20 Examples of protein complex elution in mixed bed fractions

QExactive data was analysed by MaxQuant, and the iBAQ intensity calculated for each protein. The amount of each protein was calculated and results shown for a) the ARP2/3 complex, b) a complex of methionyl and glutamyl-tRNA synthetases bound to ARC1, c) phosphofructokinase, d) the nucleosome e) the structure of the complete nucleosome octamer, from *White et al.*, 2001 f) trehalose-6-phosphate synthase/phosphatase, and g) the fatty acid synthase.

double peak (figure 3.20 b), suggesting some of the protein is remaining attached to the column while the first peak is eluting. This either suggests that there are two different conformations of the complex observed, that one or both of the proteins has been partially phosphorylated, or that the heterodimer is interacting with another protein, or complex. There is evidence for PFK1 interacting with 172 other proteins, according to the IntAct database, and 152 for PFK2, the majority of which are gathered from TAP tag studies, so other interactions are possibly the reason for the double peak observed here. Another example is the mitochondrial processing protease, the two subunits of which (MAS1 and MAS2) elute in the same fractions, and peak in concentration in two separate positions (figure 3.20 c). The MAS1 subunit can undergo phosphorylation, and therefore the two peaks may be caused by the two forms of MAS1, phosphorylated and unmodified. Another example of a complex exhibiting double peaks is the nucleosome complex, which consists of multiple copies of four histone proteins. There are six nucleosome associated proteins identified, four core proteins H2A, H2B, H3 and H4, the linker histone H1, and also the H2A variant H2AZ. When the fmol amounts of these proteins are plotted, they separate into two distinct peaks, which elute at two positions in the gradient (figure 3.20 d). The complete complex is an octamer, composed of two copies of each of the core histones, which assemble into subcomplexes. A tetramer forms, consisting of two H3 and two H4 proteins, which binds DNA and then recruits two copies of H2A and two copies of H2B (White *et al.*, 2001) (figure 3.20 e). A single copy of histone H1 then binds the assembled octamer. The peak observed at fraction 27 consists of the four core histones, while the peak eluting earlier in the gradient, at fractions 19-20 also includes H1. This could indicate the addition of the linker histone alters the isoelectric point of the complex, or the completed assembly may be in two forms, one of which is bound to a strand of DNA, which carries a negative charge, and would therefore alter the affinity of the complex proteins to the chromatography medium. In addition, there are 57 interactions listed for histone H1 in the IntAct database, so other interactions are possible.

Despite the improvement in resolution observed when using mixed bed chromatography, some complexes elute in wide peaks. This indicates that these complexes have a higher propensity to bind to the column. Some elute over a wider range of fractions, such as trehalose-6-phosphate synthase/phosphatase, which has three subunits eluting across almost 20 fractions, although the third, regulatory subunit, TPS3 eluted over just 7 of the middle fractions (figure 3.20 f). The three subunit chromatograms of this protein do not show as good a correlation as some other complexes, possibly indicating they have dissociated, are partially assembled, or are eluting as individual proteins. Fatty acid synthase also shows a wider peak elution the two subunits of which elute from the column sharply in fraction 18 and then the peak tails off slowly (figure 3.20 g).

There are also other examples of complexes showing similar behaviour, such as the ribosome complexes, which elute across a wide range of fractions, spanning a large part of the gradient. The ribosomes are made up of a large (60S) and small (40S) subunit, and are both relatively large structures (Ben-Shem *et al.*, 2010). There were 38 subunits identified from the 40S ribosome across fractions 12 to 22 (figure 3.21 a), and 59 subunits associated with the 60S ribosome identified across fractions 15 to 28 (figure 3.21 c). For both complexes subunits were identified in different fractions, eluting at different parts of the gradient, indicating that the complex was dissociated, or partially assembled. Hierarchical cluster analysis sorted the 40S ribosome proteins into seven distinct clusters (figure 3.21 b), while the 60S ribosome sorts into approximately 6 clusters (3.21 d), each containing three or more proteins, which may indicate that there was some level of organisation in the ribosome proteins. Growing yeast cells must produce large numbers of ribosomes to enable transcription, and the cellular ribosome proteins can undergo rapid turnover (Pestov & Shcherbik, 2012). The spread of the subunits across multiple fractions and in a number of clusters may, therefore, represent ribosome proteins partially assembled into complexes. On the other hand, the ionic conditions used in eluting proteins from the mixed bed chromatography column may cause disruption to some electrostatic interactions within the protein complexes. The assemblies are so large and subunits so numerous that multiple interactions occur for each

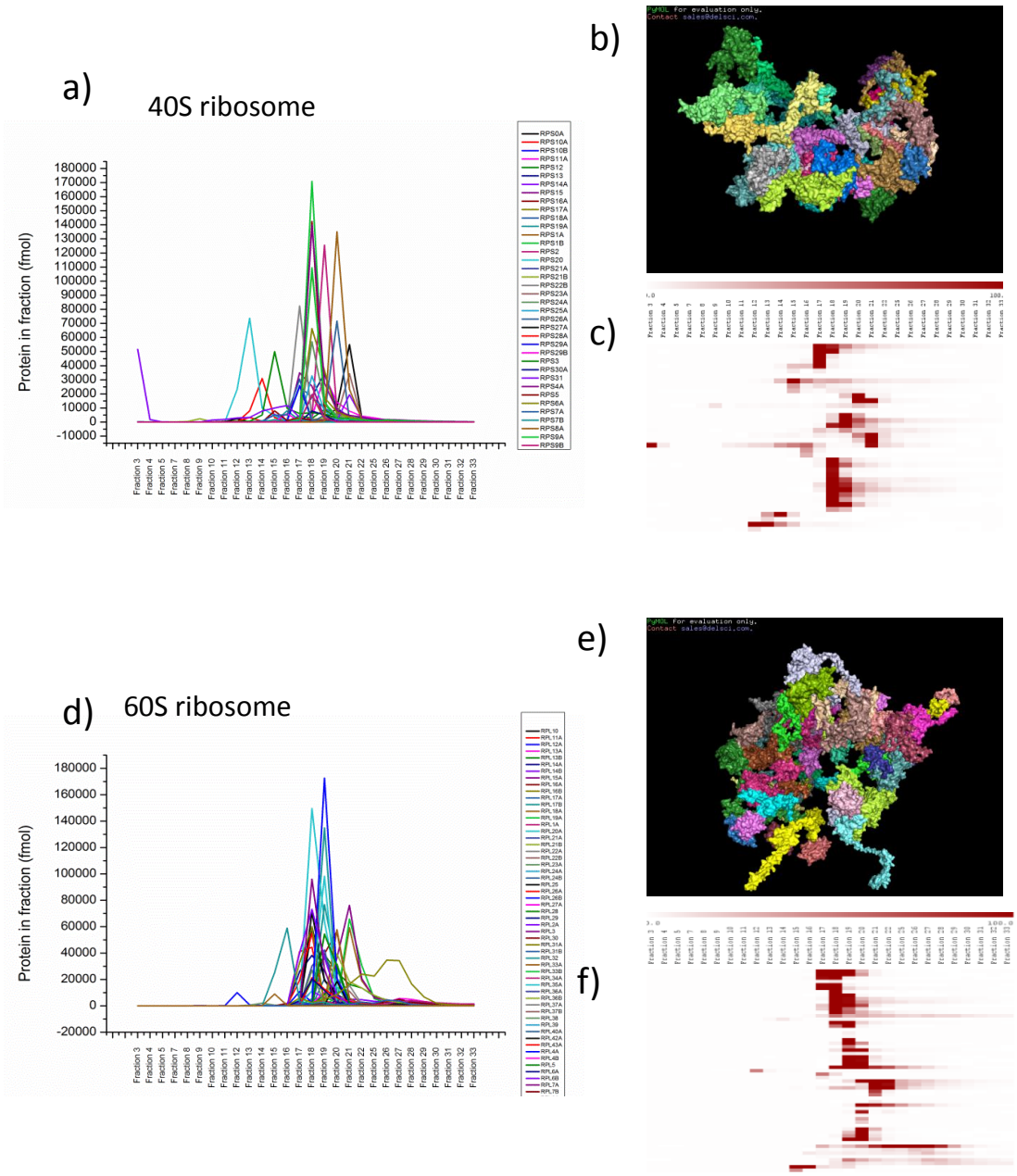


Figure 3.21 Chromatographic elution of the subunits of the ribosome complexes.

QExactive data was analysed by MaxQuant software, to find the iBAQ intensity of each protein, used to calculate the fmol, and a) the fmol of 40s ribosome subunits plotted, and b) the structure of the complex, from Ben-Shem *et al.*, 2010 and c) the clustering of complex subunits. d) the quantification of the 60s ribosome subunits, e) the structure and f) the hierarchical clustering.

protein. Therefore the dissociation of some protein interactions, particularly in large systems such as the ribosomes, may be expected. There are also examples of proteins that are present in multiple complexes, such as the RNA polymerase subunits. There are three types of RNA polymerase in yeast, termed RNA polymerase I, II and III. Each complex has some unique subunits, and some shared subunits. RNA polymerase I contains 14 subunits, 7 of which are unique, and 7 are also constituents of RNA Polymerase II or III. RNA polymerase II consists of 12 proteins, 7 of which are unique, and RNA polymerase III contains 16 subunits, 9 are unique, and 7 are shared with the other two complexes. The graphs of each complex had multiple peaks, while the unique proteins elute in just one peak (figure 3.22). Hierarchical cluster analysis shows at least two distinct clusters, one in fraction 27-29 containing the RNA polymerase I subunits, and a clear cluster in fractions 22-26 for the RNA polymerase II subunits (figure 3.23 a). There was also a third, less distinct cluster in fractions 26-27 containing those subunits unique to RNA polymerase III. Examining the unique subunits to each complex gives three distinct peaks, while picking out the subunits that can be present in multiple complexes, many had multiple peaks (figure 3.23 b-f). Those subunits that are unique to RNA polymerase III were present at lower concentrations than those from the other complexes, supporting the clustering of the heat map, showing a less intense cluster in this position.

There are also indications of other proteins forming multiple interactions, such as zuotin (ZUO1), a chaperone which interacts with Hsp70 chaperone SSZ1 and SSB1/2 (Gautschi *et al.*, 2001). ZUO1 elutes from the mixed bed column in two distinct peaks, one of which co-elutes with SSZ1 and the other with SSB2 (figure 3.24). This indicates the chaperone is partly associated with each protein, either because they are partly assembled, or partially dissociated.

3.3.6 The exosome complex

The eukaryotic exosome is a multi-subunit complex required for the degradation of 3'-5' RNA. It consists of nine non catalytic subunits, Rrp40, Rrp4, Rrp41/Ski6, Rrp42, Rrp43, Rrp45, Rrp46, Csl4, and Mtr3 arranged into a ring structure, and one

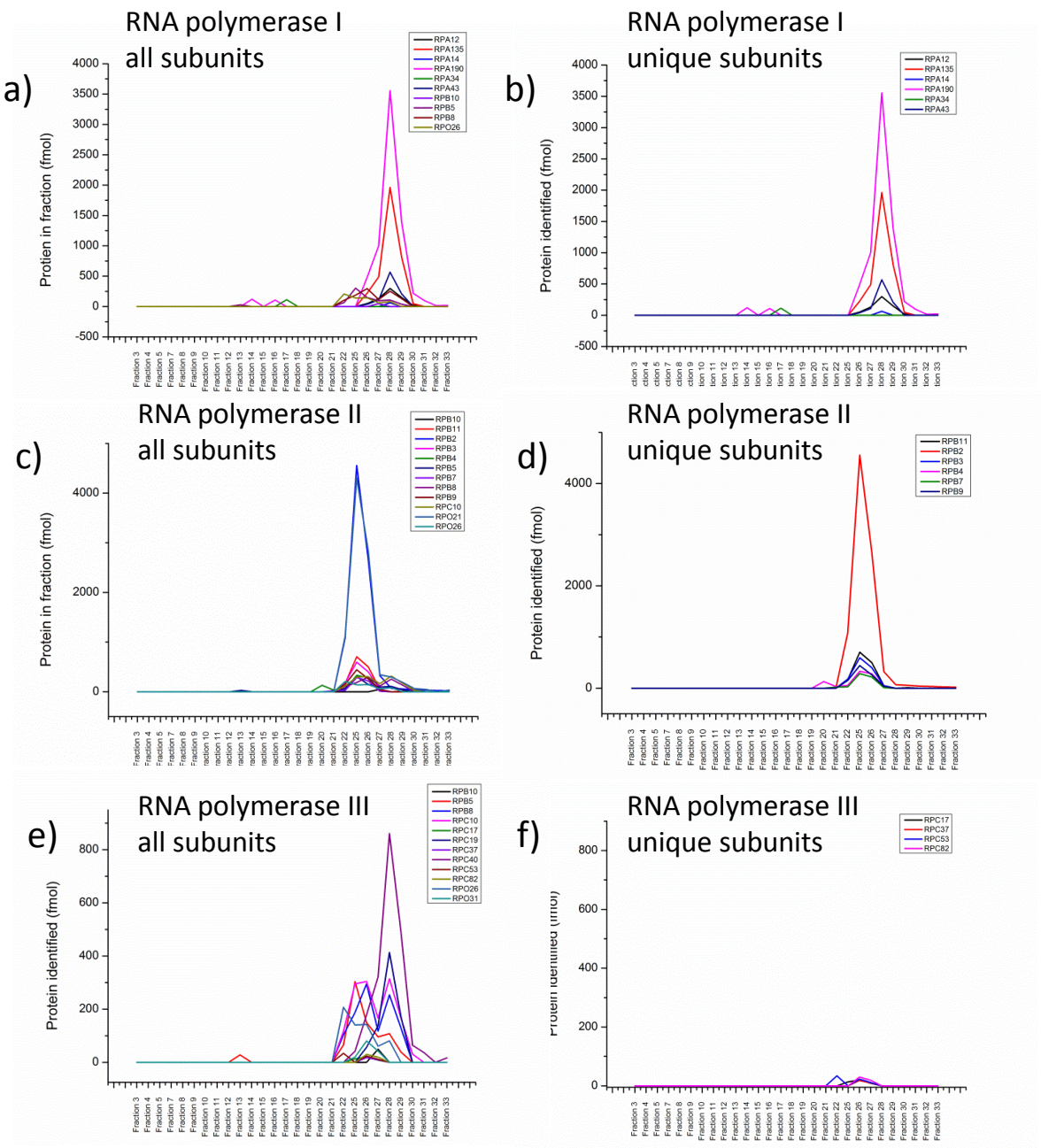


Figure 3.22 Quantification of RNA polymerase subunits in mixed bed fractions.

Results are analysed by MaxQuant software, which reports the iBAQ intensity of each protein, used to calculate the fmol. The fmol in each fraction is calculated by multiplying by the total fraction volume, and results are shown for a) each RNA polymerase I subunit, b) each unique RNA polymerase I subunit, c) each RNA polymerase II subunit, d) each unique RNA polymerase II subunit, e) RNA polymerase III subunits, and f) unique RNA polymerase III subunits.

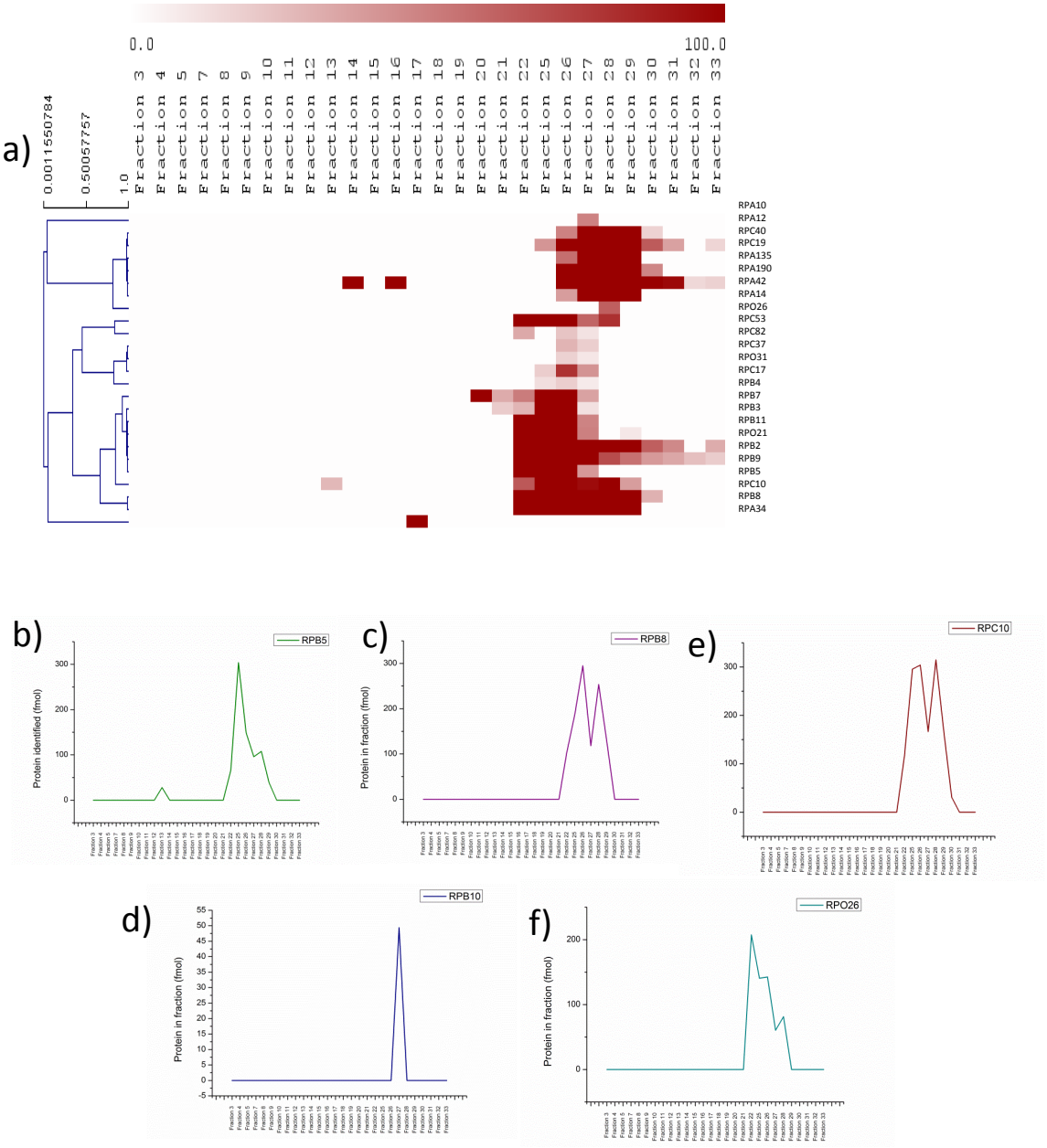


Figure 3.23 Hierarchical clustering of RNA polymerase subunits and quantification of the subunits with roles in multiple complexes.

a) hierarchical clustering of the normalised results of RNA polymerase subunit quantification. Normalisation is achieved by calculating the percentage of each protein found in each fraction. The fmol results for those subunits that are present in all three complexes are plotted for b) RPB5, c) RPB8, d) RPB10, e) RPC10 and f) RPO26.

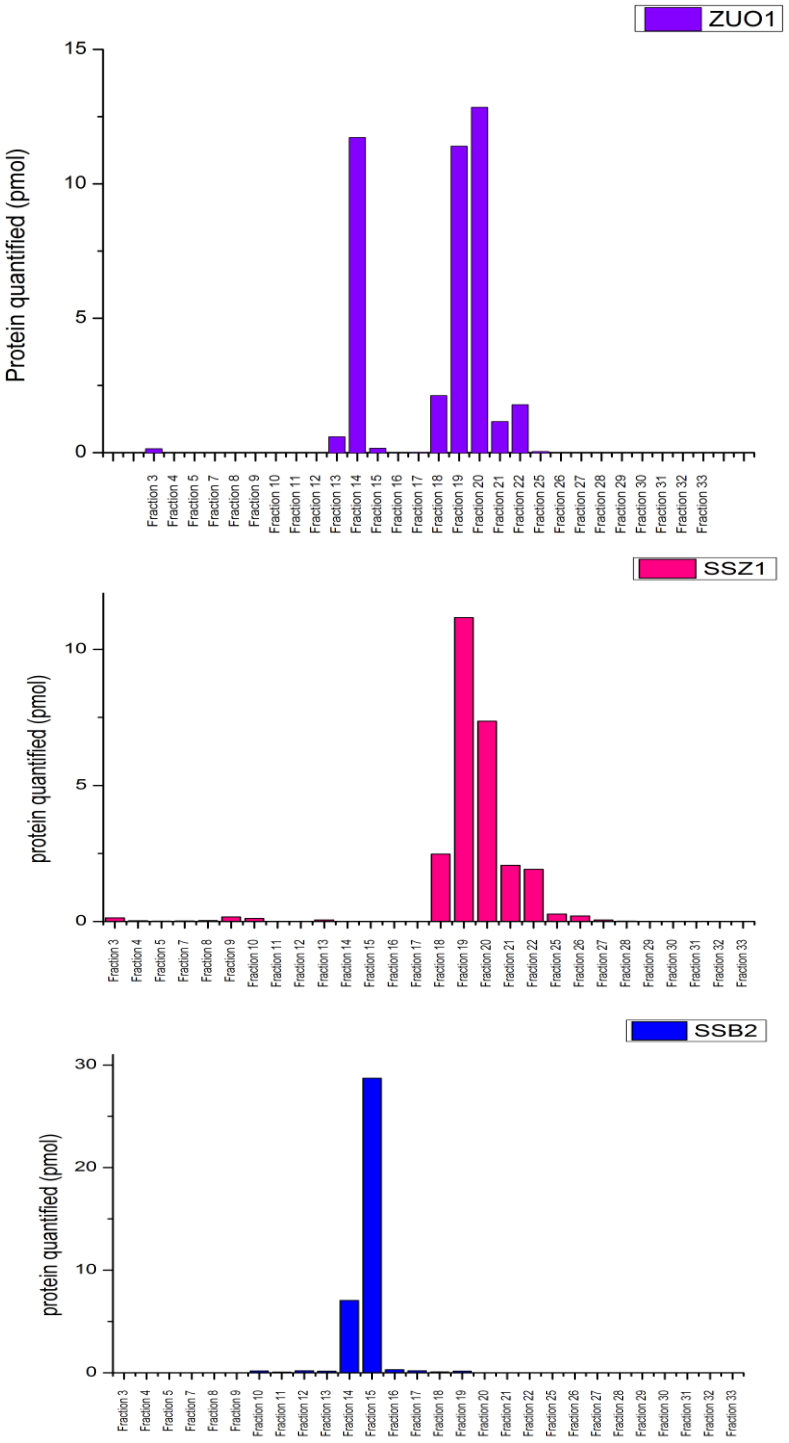


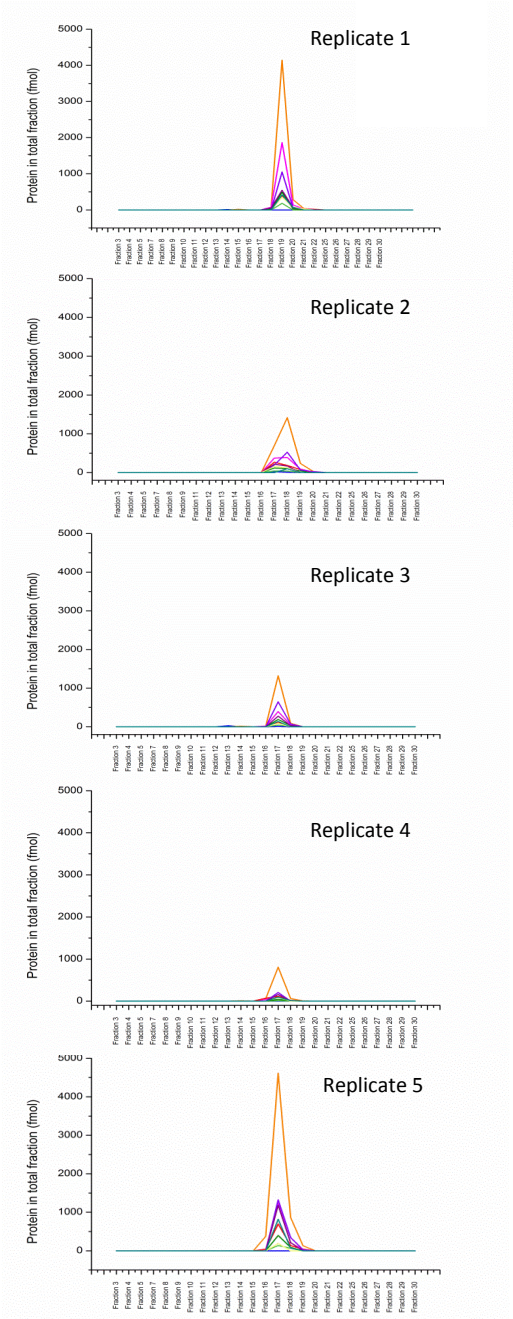
Figure 3.24 LPD1 (Dihydrolipoamide dehydrogenase) protein quantification in mixed bed fractions

The quantities of protein (fmol) in each sample injected was calculated from the Qexactive data by processing with MaxQuant software. The total protein per fraction was then calculated by multiplying by the total volume, and the result is displayed for ZUO1, SSZ1 and SSB2.

catalytic subunit, Rrp44/DIS3, at the top of the ring (figure 3.25 b) (Makino *et al.*, 2013b). In the nucleus, the catalytic subunit Rrp6 is also associated with the complex, although this subunit is not identified in this data. All 10 of the cytosolic exosome complex subunits are quantified in the fractionation replicates, with just one subunit, CSL4, missing from replicate 4 (figure 3.25 a). The structure of the complex is stoichiometrically equal, and therefore, if the complex remains whole, there will be equal amounts of each subunit identified. However, the label free quantification does not reflect this. In all replicates, the 10 subunits are identified in one fraction, indicating that they remain in complex, but they are apparently present at different concentrations. In the most extreme example, replicate 5, there is a 31 fold difference in the amount of the highest subunit, DIS3, and the lowest subunit, RRP46. In each replicate, DIS3 is the most abundant exosome subunit, and when all 10 subunit quantities are plotted against each other as scatter graphs (figure 3.26 a & b), there is a general linear trend. Therefore, the proteins that are identified at lower concentrations are consistently lower, while the higher abundance proteins are consistently higher, indicating there is an experimental reason why these proteins perform differently under LC-MS.

Not all of the exosome subunits are identified in samples run on the G2 and processed using PLGS. This is to be expected, since the G2, on average, returned almost half of the hits of the QExactive runs. In replicate 1 only 8 proteins were identified, replicate 2 there were only 3 proteins identified, and replicate 3 there were 2 proteins. In replicate 4, only the highest concentrated protein DIS3 is identified. Four proteins were also missing from replicate 5, but a comparison of the G2 and QExactive data was possible by looking at the common peptides identified in replicates 1 and 5. The comparison of the fmol amounts of each protein found using both instruments (figure 3.27 a-d) does show a somewhat linear relationship, with adjusted R2 values (calculated in origin) of 0.7 for replicate 1 and 0.5 for replicate 5. This indicates that the proteins are quantified similarly in both types of mass spectrometer, with the lower abundance proteins consistently low, and the differences observed in the quantification of the subunits are due to the properties of the protein, not the mode of analysis.

a)



b)

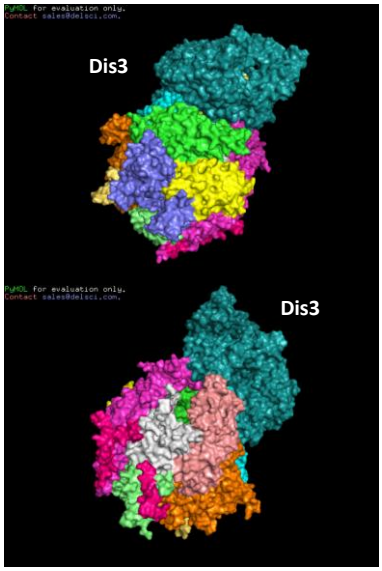
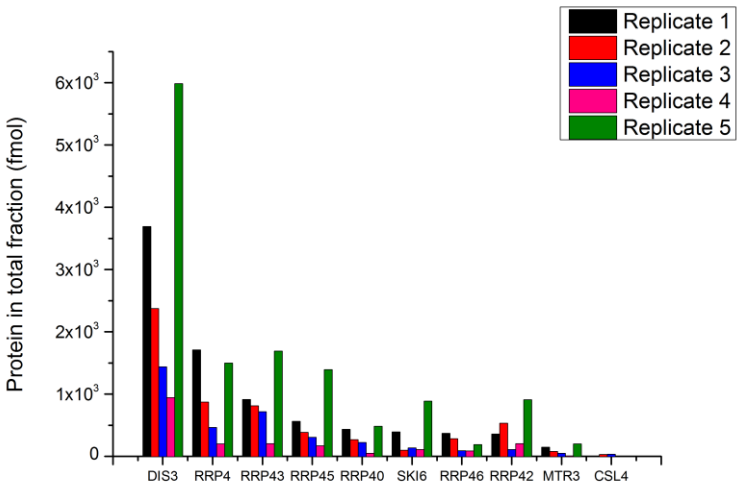


Figure 3.25 Label free quantification results of the exosome subunits

Label free quantification results are reported a) for exosome subunits in all five replicates. b) the structure of the *S. cerevisiae* exosome complex, with Dis3 arranged at the top of a round structure formed from a single copy of each other subunit, obtained from Makino *et al*., 2013.

a)



b)

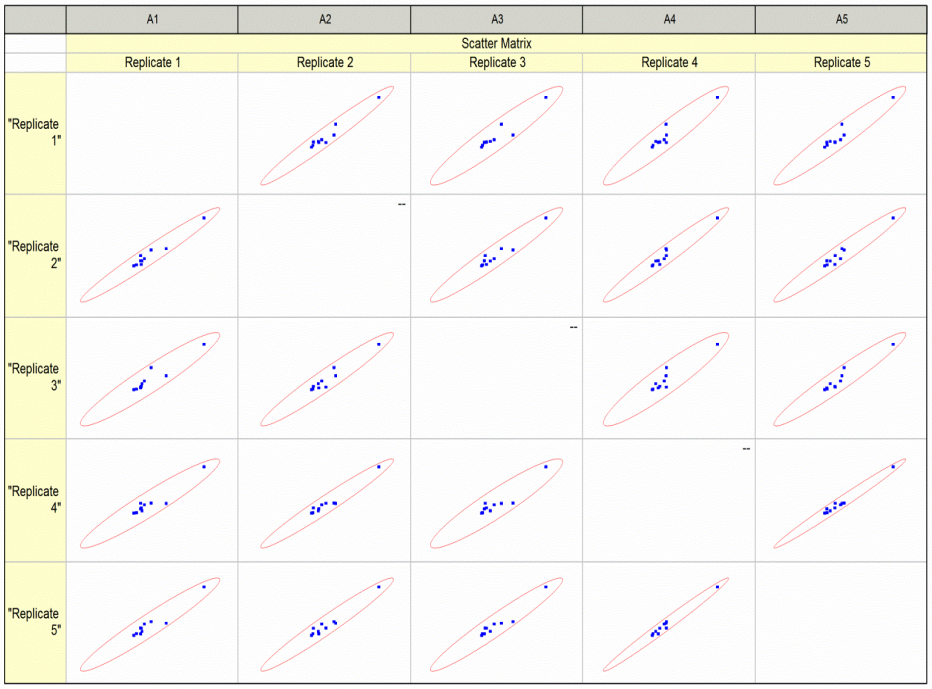


Figure 3.26 The exosome subunits quantities in five replicates calculated by MaxQuant processing

The QExactive data was processed using MaxQuant software, and a) the fmol amount of each of the exosome proteins in the five replicates of mixed bed fractionation, and b) a scatter matrix of the five replicates of protein quantifications plotted against each other.

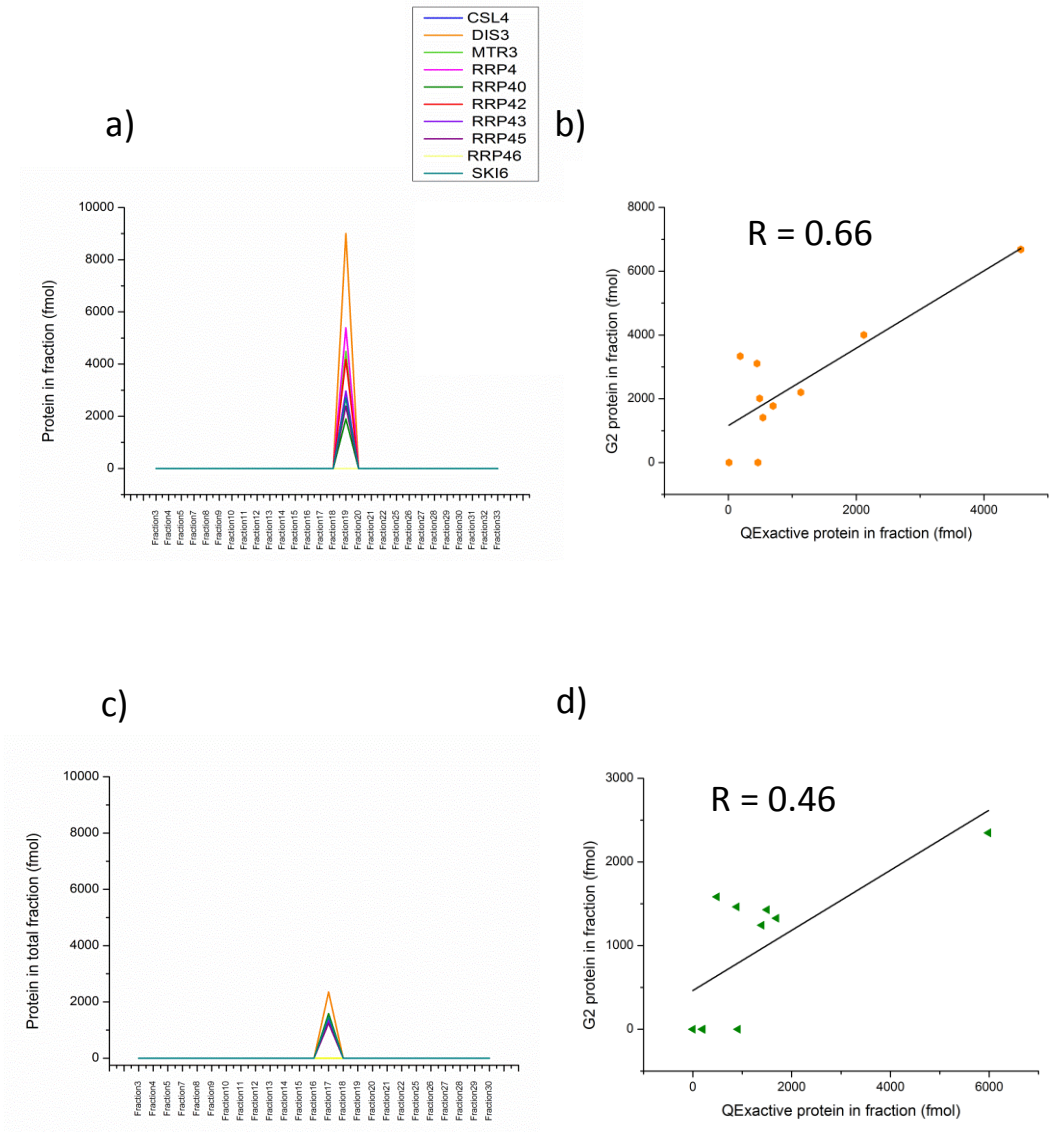


Figure 3.27 Exosome label free results calculated from G2 data

G2 data was processed by PLGS software, which gives a result in fmol per injection, which was then corrected for digest volume. a) line graph representing the amount of each subunit in replicate 1, and b) the fmol amount of each subunit in the PLGS processed data against the MaxQuant processing. c) the fmol amount of each exosome subunit in replicate 5, and d) the fmol amount of each subunit in the PLGS processed data plotted against the MaxQuant processed data.

The proteins that were quantified at the lowest concentrations do so repeatedly across all of the replicates. In the QExactive data CSL4 is quantified at the lowest amount in all replicates, however it is not identified at all in replicates 4 or 5. MTR3 is quantified as the second lowest subunit in four replicates, from just 1 to 5 peptides in each replicate, a total of 7 peptides in 5 replicates, with just one of which is present in all replicates. In the third lowest concentration subunit, RRP46, which has 5 peptides identified, only two are found in all replicates. This may mean the low concentration of these subunits is due to a lack of peptides that are consistently observable on a mass spectrometer (figure 3.28 a). In label free quantification, the intensity of the peptides is taken to represent the quantity of the protein. Therefore, any errors in protein quantification must be due to differences in the peptides intensities observed. Of 56 peptides identified in total, there are 21 common to all 5 replicates. If the intensities of these are plotted against each other on a scatter matrix, they show a somewhat linear relationship, indicating that less detectable peptides perform consistently poorly across all replicates (figure 3.28 b). If this is the case, the reason for the low quantification of some subunits could be that the iBAQ based label free analysis has used less intense peptides in the quantification. There are only two replicates of G2 data that have enough peptides to perform a comparison, but in both of these there is a somewhat linear trend, although much less than the protein quantification (R^2 values 0.1 and 0.04) (figure 3.28 c and d). This indicates a wide variation in peptide intensity. The observation of a peptide on a mass spectrometer depends on the peptide performance in the chromatography, and ionisability in the mass spectrometer, both of which are dependent on the physicochemical properties of the peptide. These properties can cause peptides to behave differently in different mass spectrometers (Mallick *et al.*, 2007). The two mass spectrometers differ in design; the Waters Synapt G2 is a quadrupole-time of flight instrument, while the Thermo QExactive is a quadrupole orbitrap. The design of an orbitrap mass spectrometer means that ions are trapped, and accumulate prior to their release into the detector. This increases the chance of acquiring all the peptides, which means there will be a difference in the number of peptides observed by the two instruments. The less linear effect observed when the intensities from both mass spectrometers are plotted indicates that peptides are

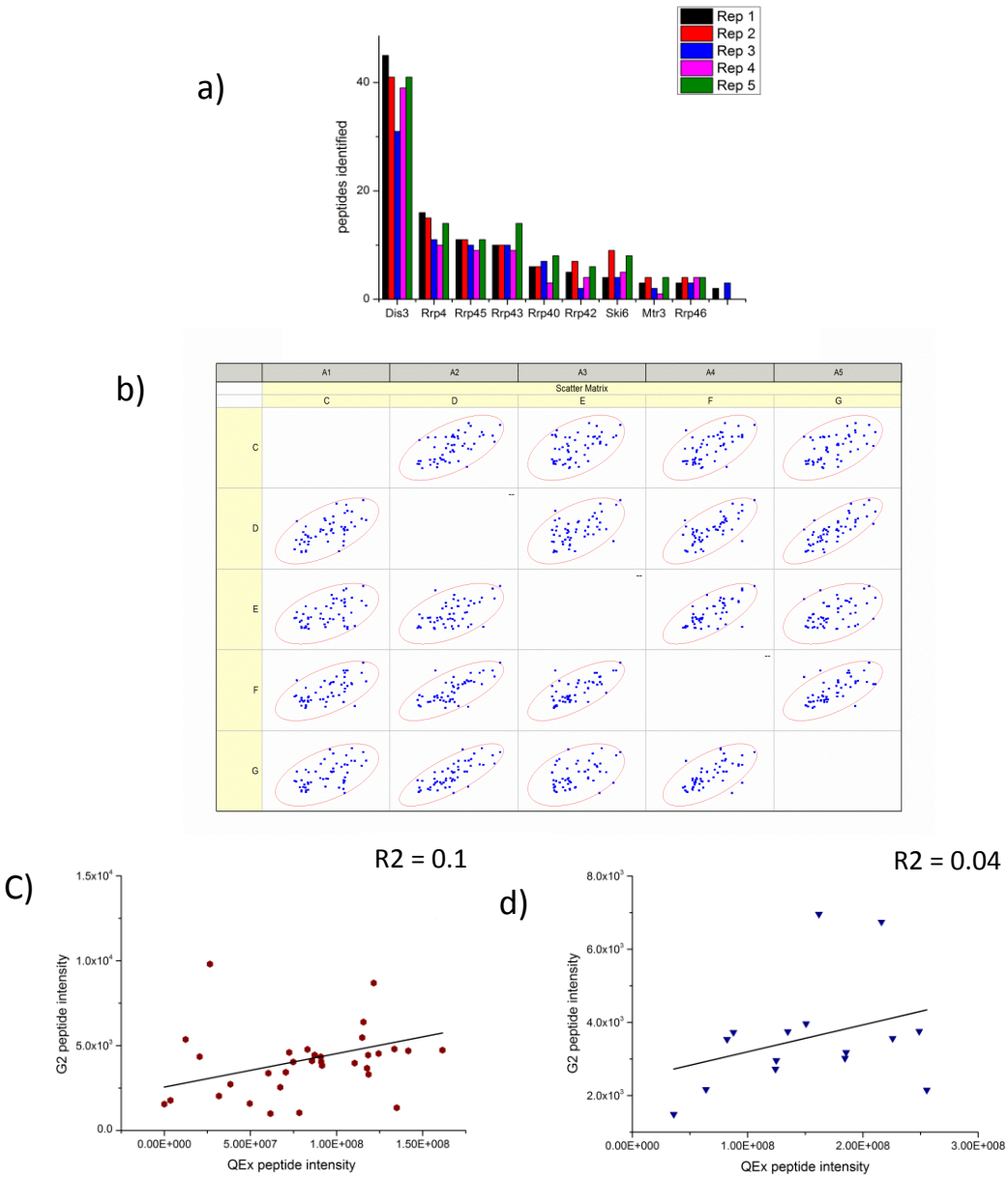


Figure 3.28 Qexactive reported intensities of the exosome peptides identified.

The Qexactive data is processed by MaxQuant software, a) the number of peptides identified in each replicate, b)the common peptides found in all five replicates plotted as a scatter matrix, and c) the replicate 1 intensities of the peptides common to both the G2 and Qexactive data sets plotted against each other, and d) the replicate 5 common peptide intensities plotted against each other.

not performing the same in the two instruments. In addition, the data from the two instruments is processed differently, with PLGS using the top three most intense peptides to calculate the fmol amount of protein, while MaxQuant uses IBAQ, taking into account all of the peptides found, and the theoretical probability of observing the peptides from a protein. However, the fmol amount plotted against each other does show a linear relationship, which indicates that the variation in peptide intensities is somewhat accounted for in the processing, however the variation in the exosome quantification shows that the label free processing does not overcome all of the differences in peptide observation.

3.3.7 The 20S proteasome

The 20S proteasome is another example of a protein assembly whose subunits co-elute using mixed bed ion exchange chromatography. The complex is responsible for degradation of proteins within the cell, and consists of 14 proteins, 7 β subunits arranged in a ring structure, with two rings stacked on top of each other, encased on either side by another ring structure, consisting of 7 α subunits (figure 3.29) (Groll *et al.*, 1997). Therefore, the complex consists of two copies of each of the 14 different subunits, which should be equimolar.

All 14 subunits are identified in all five replicates of the MaxQuant data (figure 3.29). There are differences in the amount of each subunit quantified, in the most extreme case up to a five fold difference in replicate 1, between PUP1 at 6 fmol and SCL1 at 32 fmol. Although there is a difference in the amount of protein quantified in all five replicates, those at a higher concentration are higher in all replicates. SCL1 is among the top three proteins in all replicates (figure 3.30 a). When the fmol amount of each protein in the five replicates are plotted as a scatter matrix, it supports the idea that the proteins quantified at a higher concentration do so repeatedly (figure 3.30 b).

The number of peptides identified differ for each protein, and the three lowest concentration proteins, PUP1, PUP3 and PRE4, have the least number of peptides identified (figure 3.31 a), therefore, the low concentration calculated for these

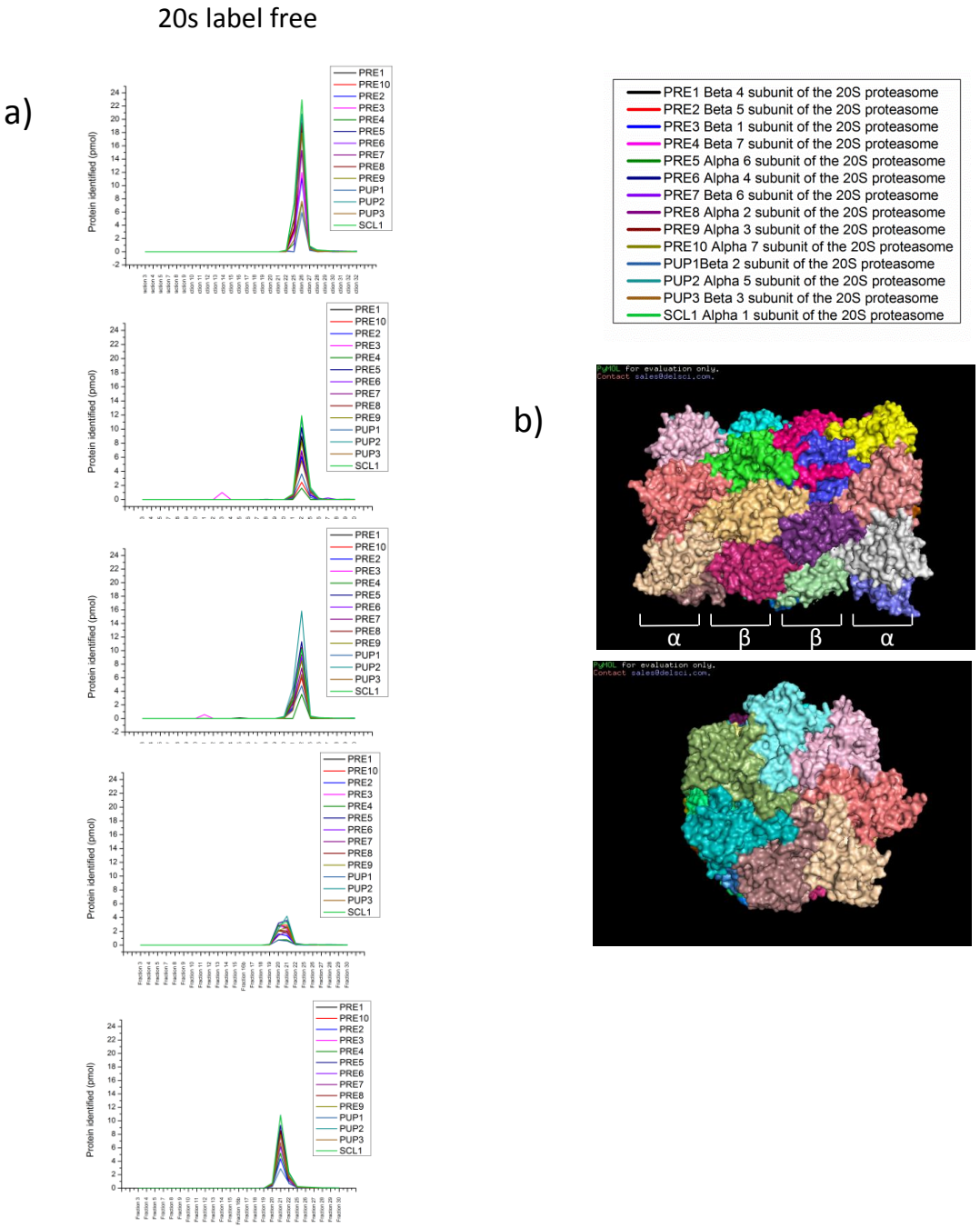
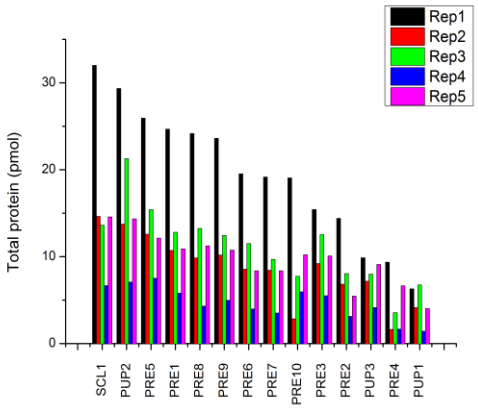


Figure 3.29 Quantification of the 20S proteasome in 5 replicates of mixed bed ion exchange

a) QExactive data was processed using MaxQuant, and the fmol amounts of each protein were calculated based on the iBAQ intensity reported. b) the 20S proteasome structure, obtained from Groll *et al*., 1997.

a)



b)

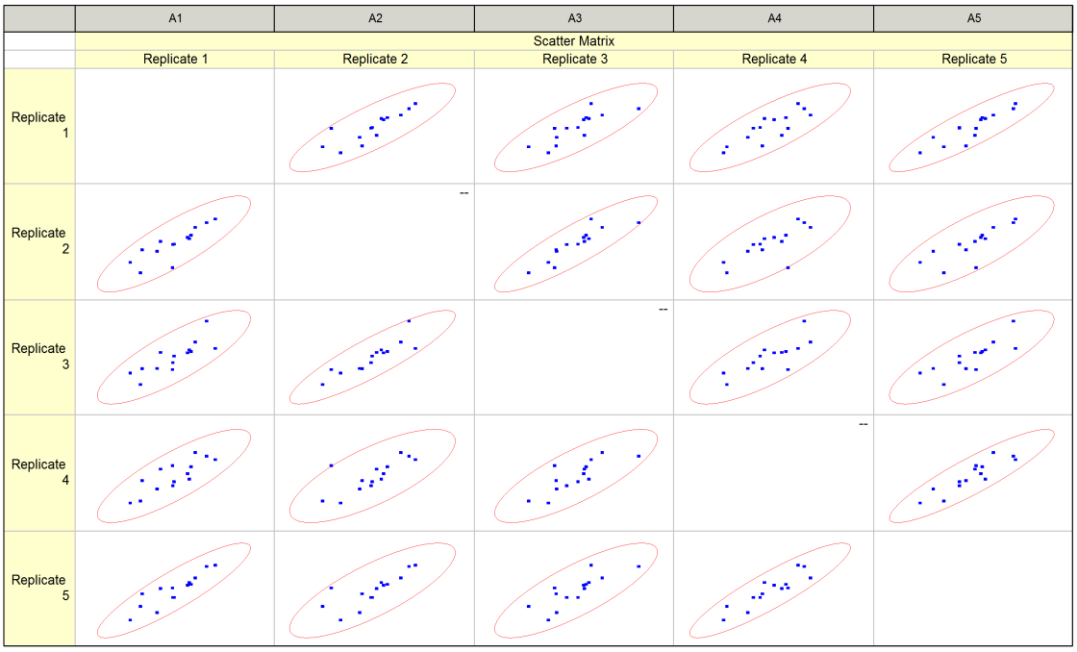


Figure 3.30 The total material quantified for all 20S proteasome subunits in all five replicates.

The label free quantification of each protein is calculated using MaxQuant processed data, and a) the fmol of each protein plotted in descending order, and b) a scatter graph depicting the total fmol identified across all fractions, plotted against the material in the other four replicates.

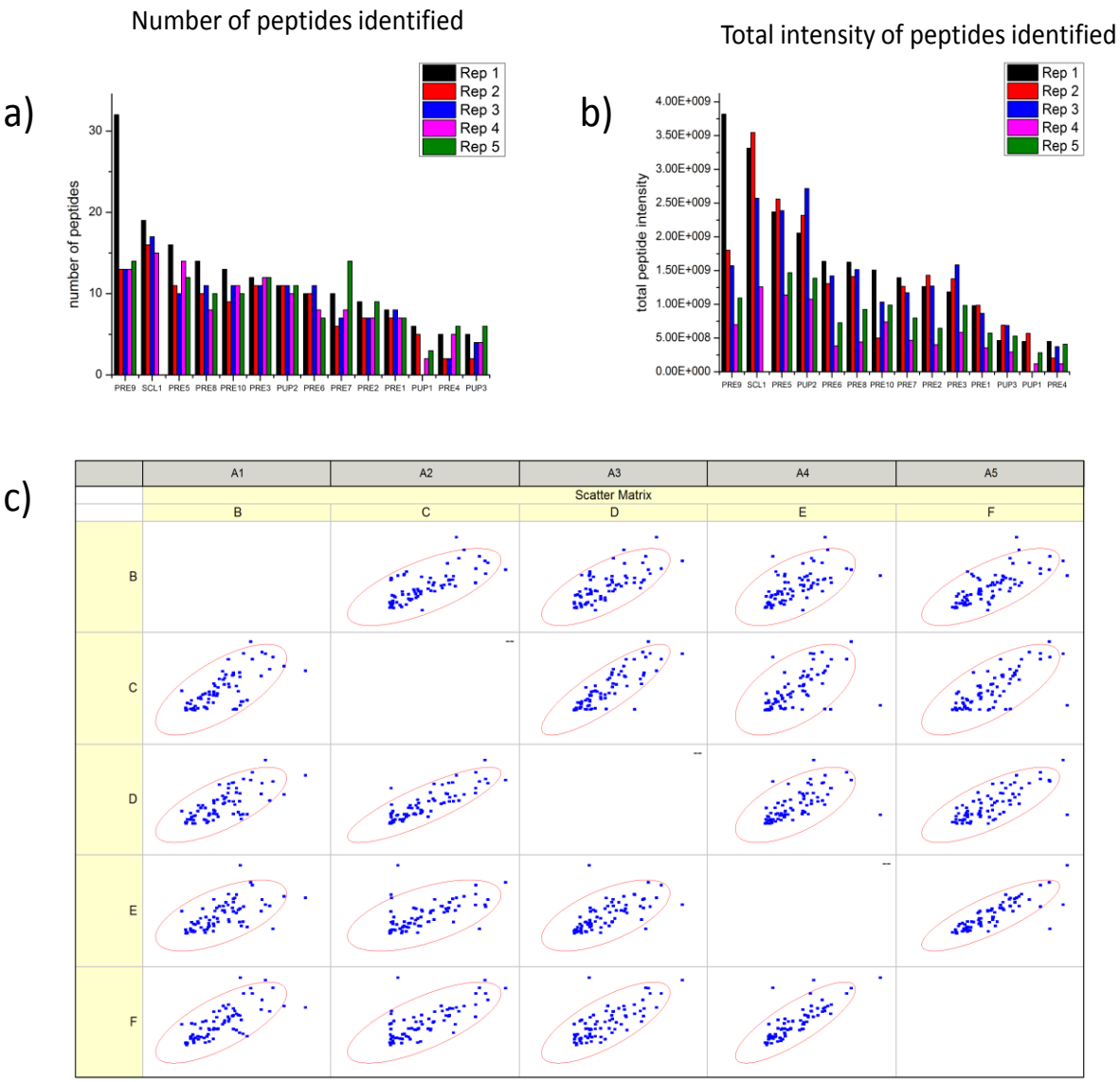


Figure 3.31 Peptide intensities of 20S proteasome subunits identified in all five replicates.

a) The number of peptides identified from each protein by MaxQuant processing of QExactive data, and b) the total peptide intensity for each protein in each replicate, c) the intensities of common peptides to all replicates, plotted against each other.

proteins may be due to a lack of observable peptides. The difference in the number of peptides also influences the total peptide intensity (figure 3.31 b), which increases as more peptides are identified. When the intensity of each individual peptide are plotted against the other replicates, there is a less obvious trend in the peptide intensities than was observed at the protein level (figure 3.31 c), however there is still a positive relationship. This indicates that peptides which ionise poorly do so consistently, and if these peptides are used to calculate the protein concentration, it may result the variation in protein quantification observed. In most replicates there is a positive relationship between the peptide intensities in the G2 data and the QExactive (figure 3.32). The two data sets are processed using different types of software, the G2 data by PLGS which uses the top 3 most intense peptides, while the QExactive data is processed using MaxQuant and the iBAQ intensity. The trend between the two data sets may indicate that the proteins which are quantified at lower concentrations are due to lower intensity peptides.

3.4 Conclusions

The method developed for the fractionation of *S. cerevisiae* cell lysate using mixed bed ion exchange chromatography has successfully separated a number of protein complexes, which show a variety of elution behaviours. Despite some protein interactions being disrupted, it has been possible to quantify some protein complexes. However, it is possible that, had a gentler method of fractionation been used, such as gel filtration, more complexes, and more transient interactions, would have been observed. An added advantage in SEC is that it is relatively easy to define if a protein is forming an interaction, as it appears in the gradient at a higher mass than its own. A comparison of the two methods would have been an interesting addition to this chapter, and in the case of additional complexes found using SEC it would have been possible to define the interactions occurring as electrostatic. The resolution of SEC separations could have been increased by using a sequence of columns, enabling a wide range of molecular masses to be separated.

Another useful addition to this chapter would have been the use of some control samples, to find a method of disrupting protein complexes, however it would be

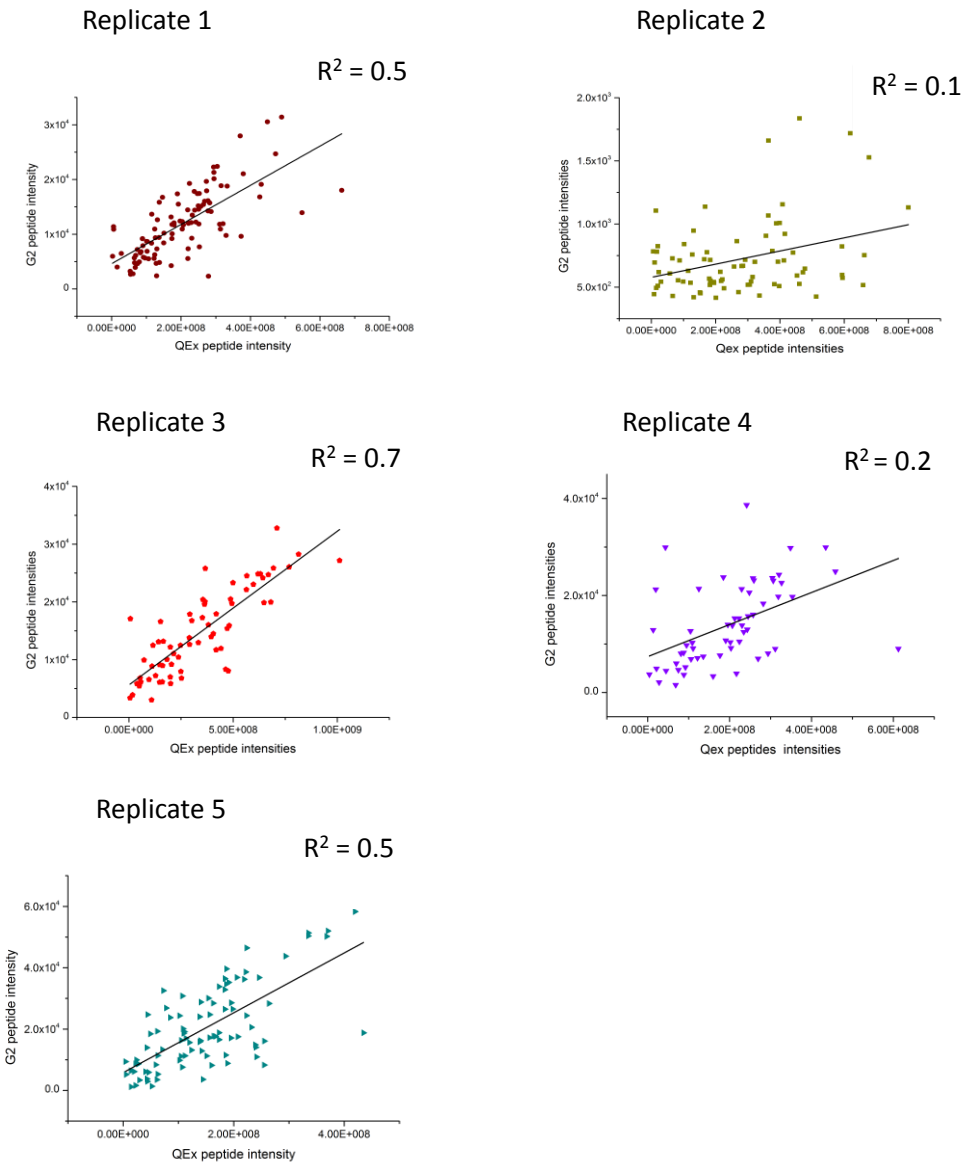


Figure 3.32 Peptide intensities of 20S protein subunits.

The Qexactive data is processed by MaxQuant software, which reports the intensity of all of the 20S peptides identified, while the G2 data is processed by PLGS. The intensities of the peptides common to both the G2 and Qexactive data in all five replicates are plotted against each other.

difficult to disrupt the protein-protein interactions without also affecting the tertiary structure of the protein, and therefore there would be no guarantee that any effects observed are not the consequence of denaturing the protein and not just the interactions. However, a control would have provided support for the fact that the proteins are chromatographing as part of complexes, and not just because they have similar isoelectric points, although all of the complexes seen here are supported in literature.

Some complexes elute in single, sharp peaks, while others are spread across multiple fractions, or elute in multiple, discrete peaks. In some cases, this may indicate that there are multiple conformations, post translational modifications, or there may be more interactions occurring.

The data presented here indicates that label free analysis may not provide a suitable method of analysing the stoichiometry of proteins within a complex. The quantification relies on the availability of quantotypic peptides, which for some proteins are less available. The mode of analysis used here applies data dependent acquisition, which targets only a selection of the peptides available within the sample, thereby eliminating some proteins from the analysis. There are more accurate methods of quantitative proteomics available, which will be examined in the next chapter.

Chapter 4: Quantification of protein complexes using Selected Reaction Monitoring

4.1 Introduction

Large scale protein interaction studies, implementing techniques such as Y2H experiments, or TAP tag projects, have over the years gathered a plethora of information about protein-protein interactions occuring within the *S. cerevisiae* cell (Uetz *et al.*, 2000; Gavin *et al.*, 2002; Yu *et al.*, 2008; Rajagopala *et al.*, 2012). In addition, a number of other techniques have been used to extend the study of individual interactions to the analysis of entire multi-subunit protein complexes. Traditional methods of studying protein complexes include gel electrophoresis and size exclusion chromatography (Camacho-Carvajal *et al.*, 2004; Rivers *et al.*, 2007; Liu *et al.*, 2008), while in more recent years mass spectrometry methods have also been implemented. Native mass spectrometry techniques have been used in the successful study of both the yeast proteasome and exosome complexes (Loo *et al.*, 2005; Synowsky *et al.*, 2006), however these techniques require large sample amounts, purified complexes, and the complex must remain stable under ESI conditions. Although these techniques yield valuable information, alternative approaches can be used to yield much more accurate results for quantitation of the protein subunits, and may also provide a more detailed view of complex stoichiometry.

Chapter 3 focused on the quantification of protein complexes using mass spectrometry followed by label free analysis. As discussed in Chapter 3, for some protein complexes identified by label-free quantification, the concentration of the individual proteins does not conform to the expected relationship between subunits (figures 3.25, 3.29). Examples of this behaviour included the 20S proteasome, a complex responsible for degradation of proteins within the cell, which consists of 14 protein subunits, seven β subunits arranged in a ring structure, with two rings stacked on top of each other, encased on either side by another ring structure, consisting of seven α subunits (see figure 3.29) (Groll *et al.*, 1997). Therefore, the complex consists of two copies of each of the 14 different subunits

and the label-free quantification results should be equal. Although the structure of the complex has been confirmed in numerous studies (Groll *et al.*, 1997; Beck *et al.*, 2012), the label free data obtained from both the Waters Synapt G2 and the Thermo Scientific QExactive do not reflect the protein complex stoichiometry, as in some instances there were relatively large differences in the quantity of some 20S proteasome subunits.

These inaccuracies in the quantification could indicate that label-free proteomics is not an appropriate method for the study of protein complex stoichiometry. Label-free absolute quantification relies on spiking a known quantity of a standard protein into the sample to be analysed. The standard protein peptide intensities are then taken to represent the amount of protein that was spiked in, and the quantification of all other proteins in the sample are calculated based on the ratio of sample peptide intensity compared to the intensity of the standard peptides (Silva *et al.*, 2006). In this case the ratio is calculated either by comparing just the top three most intense peptides using PLGS to analyse the Synapt G2 data, or when using MaxQuant for the QExactive data used in chapter 3 the ratio is calculated by intensity based absolute quantification (iBAQ), using the sum of all the peptide intensities divided by the number of theoretically observable peptides for each protein (Schwanhausser *et al.*, 2011). These methods are based on the concept that the intensities observed are representative of the total amount of protein, however, the intensity observed can differ as a result of the physiochemical properties of the peptides, and can vary with experimental conditions (Mallick *et al.*, 2007; Sanders *et al.*, 2007). The nature of a shotgun proteomics approach, in loading all of the peptides in a sample for analysis at the same time, can cause a bias in the results towards the highly abundant, well ionising peptides. Therefore, there may be inaccuracies in quantification due to less abundant proteins failing to be identified. In addition, the peptides generated from a complex sample, differing in size and amino acid composition, may mean some peptides exhibit less propensity for ionisation than the standard peptides, giving an inaccurately low signal (Craig *et al.*, 2005; Kuster *et al.*, 2005).

Absolute quantification of proteins can also be achieved by label mediated quantification using a stable isotope labelled protein standard using a number of techniques, which potentially provides a more accurate method of quantitation (reviewed in Brun *et al.*, 2009). In this technique anything from an individual peptide to an entire recombinant protein can be labelled with heavy radioactive isotopes, to be mixed with the sample of interest, analysed by mass spectrometry and an accurate quantification can be determined from the ratio of heavy labelled standard to the analyte protein or peptide. These label mediated methodologies are potentially more accurate than label free quantification, as the standard peptides or protein will be physicochemically identical apart from the small difference in mass introduced by the label, and therefore standard and analyte should behave in the same manner in all aspects of the experiment, including both the enzymatic digestion, and in the chromatography system, but will be resolved during analysis in the mass spectrometer.

The heavy isotope labelled standards used can be synthesised in the form of a whole recombinant protein in PSAQ technology (Picard *et al.*, 2012), or as individual peptides in the form of AQUA peptides (Gerber *et al.*, 2003) for a small scale study. Alternatively, a recombinant protein consisting of a number of tryptic peptides can be designed, to act as standards for the quanitification of numerous proteins, termed a QconCAT (Beynon *et al.*, 2005; Rivers *et al.*, 2007). Targeting a number of peptides makes this technology useful in the design of a larger scale study. All of these methods have been successfully implemented in the absolute quantification of individual proteins for many studies, and AQUA peptides in combination with MRM have been successfully used to quantify human spliceosomal subcomplexes (Hochleitner *et al.*, 2005; Schmidt *et al.*, 2010).

A QconCAT consists of a number of proteotypic peptides (Qpeptides) from approximately 20 different proteins concatenated into one recombinant protein sequence. If more proteins are to be quantified, a series of QconCATs can be designed. Generally, two or more peptides for each protein would be included to provide both increased accuracy, and a provision in the case of an unusable Qpeptide. Included in each QconCAT is a hexahistidine tag sequence for

purification, and a standard glufibrinopeptide sequence for the quantification of the QconCAT protein itself. Once the recombinant protein gene sequence is designed, it is expressed in a heavy labelled form (usually [$^{13}C_6$]-arginine and [$^{13}C_6$]-lysine for tryptic Qpeptides), quantified, mixed with the analyte and then the sample is tryptically digested. This yields a sample containing a mixture of heavy labelled Qpeptides, and the light analyte peptides. This sample mixture is then analysed using a LC-MS, and the quantity of sample protein can be calculated using the ratio of the heavy labelled peptide to light sample peptide (figure 4.1). The technology has been successfully implemented in the quantification of proteins from a number of different samples (Carroll *et al.*, 2011; Bislev *et al.*, 2012; Peffers *et al.*, 2013; Al-Majdoub *et al.*, 2014).

When designing a QconCAT it is important to carefully consider the choice of Qpeptides, in addition to being proteotypic they must be quantotypic, representing the quantity of the protein being targeted. The peptide selected must be unique to that single protein of interest, and it is also preferable to select peptides that have been observed in previous proteomics experiments. There are a number of mass spectral libraries available to verify this, such as PeptideAtlas or PRIDE (Jones *et al.*, 2006; Deutsch *et al.*, 2008). The sequence context of the peptides chosen should be carefully considered, as certain sequence contexts can make the C terminus of the arginine or lysine residues less susceptible to digestion (Thiede *et al.*, 2000; Lawless & Hubbard, 2012). Dibasic residues can be cleaved at the C terminus of either residue, and an acidic residue immediately following the arginine or lysine may impair cleavage, as will a proline in the same position (Thiede *et al.*, 2000). It is also advisable to avoid any peptides that are potential sites for post translational modifications, as the signal can be split between the two forms. Methionine or cysteine residues should be avoided, as the sulphur containing side chains can undergo oxidisation reaction, adding 16 Da to the mass of the peptide, and a glycine residue immediately following an asparagine can allow a deamidation reaction to occur, causing a 1 Da mass shift. There are various software tools available to assist in selecting peptides for QconCATs, including MC:pred (Eyers *et al.*, 2011) for predicting how well the peptide will undergo enzymatic digestion, and

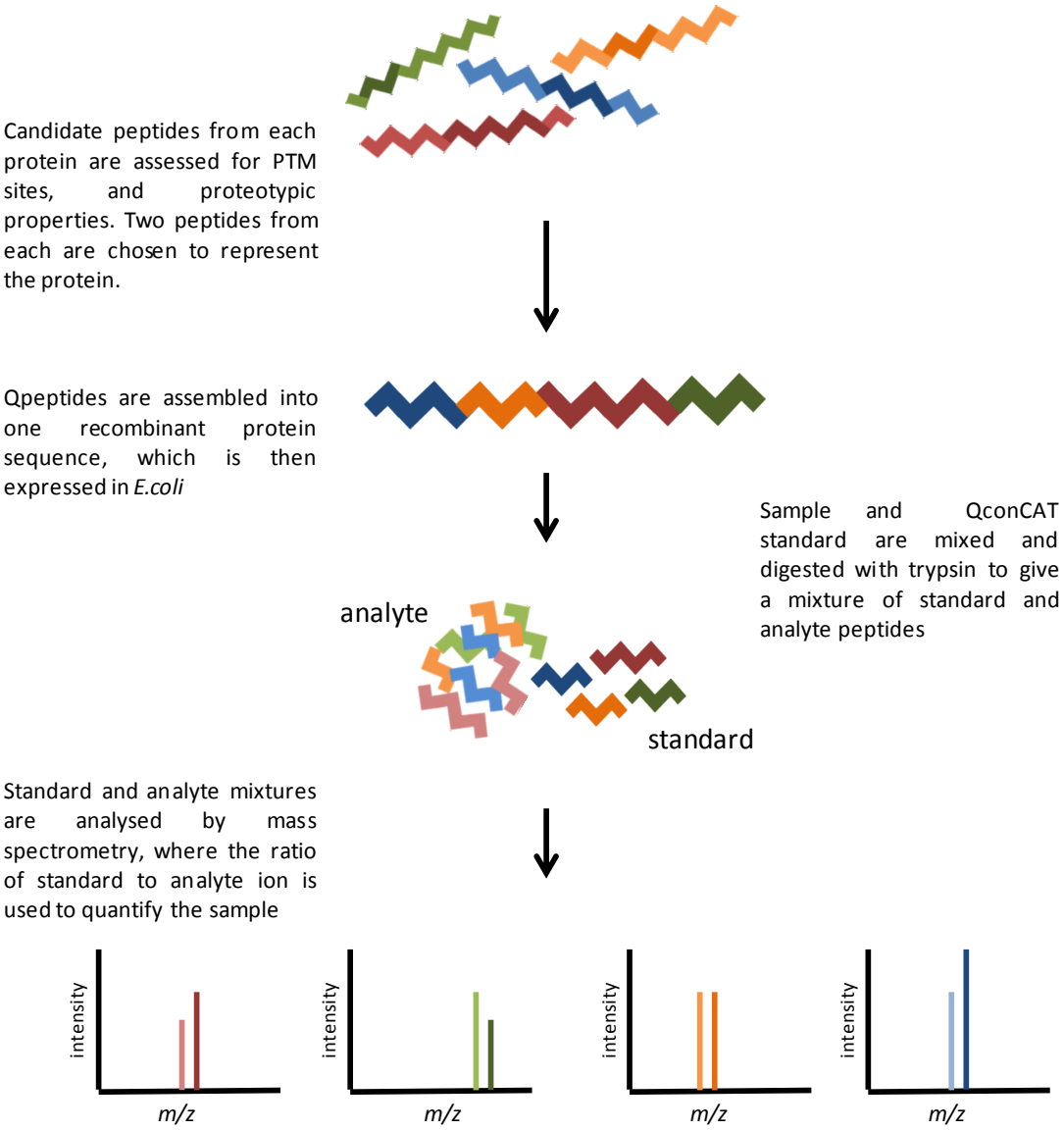


Figure 4.1 Quantifying a complex protein sample using QconCAT technology

The protein sequences are analysed to decide on the optimum target peptides. These are then concatenated into one recombinant protein sequence, which is expressed in *E.coli*, with heavy isotope labelling. The QconCAT is then quantified, tryptically digested, and standard and analyte peptides are then mixed, and analysed at the same time using mass spectrometry. The analyte is then quantified by calculating the ratio of the standard to analyte ions.

CONSequence (Eyers *et al.*, 2011; Lawless & Hubbard, 2012), a tool for predicting the potential for observing a peptide in a mass spectrometer.

Once a QconCAT has been designed, expressed, mixed with analyte proteins and digested the mixture of peptides can be co-analysed by LC-MS. Selected reaction monitoring (SRM) may prove a more accurate method of analysing a Qpeptide and analyte pair. In this approach, the standard and analyte mixtures are loaded onto a triple quadrupole instrument. A quadrupole mass analyzer consists of four parallel rods that have fixed DC and alternating RF voltages applied. Ions pass through the quadrupole, focused between the four rods, and depending on the RF applied, only ions of a particular m/z will pass through. In a triple quadrupole instrument, the first quadrupole is used to isolate a predicted precursor ion, the second quadrupole acts as a collision cell to fragment the precursor ions, following which specific fragment ions are selected in the third quadrupole. This offers higher selectivity than shotgun methods, in that two levels of mass selection are used, which also enables the reduction of noise, allowing greater sensitivity. A series of precursor/fragment pairs (transitions) are selected, and both the heavy and light versions are targeted. The intensity of the peptides are then used to calculate the quantity of the light peptide, based on the known concentration of the heavy standard peptides. This type of mass spectrometry overcomes the bias towards abundant peptides seen in a shotgun experiment, and also the standard, heavy labelled peptide and unlabelled peptides should exhibit the same behaviour under the experimental conditions, potentially leading to increased accuracy of quantitation. The technique provides an accurate method for protein quantification (reviewed in Gallien *et al.*, 2011; Picotti & Aebersold, 2012).

4.2 Aims

The aim of this chapter is to design QconCATs to accurately quantify the subunits, and thereby analyse the stoichiometry, of some of the protein complexes found in *S. cerevisiae* cell lysate fractionated by ion exchange chromatography in Chapter 3. A QconCAT will be designed to target all of the subunits in the 20S proteasome. The QconCAT and analyte mixtures will be quantified using a selected reaction

monitoring (SRM) approach on a Waters Xevo triple quadrupole instrument. This method will be compared to label free quantification for the complex subunits.

Eukaryotic translation initiation factor 2 complex will also be quantified using QconCAT technology. The eukaryotic initiation factor (EIF2) is responsible for the recognition of start codons and the recruitment of initiator tRNA to the 40s ribosome to begin translation (Shin *et al.*, 2011). The complex is a heterotrimer that consists of three subunits which should therefore be stoichiometrically equal, two of which, SUI2 and SUI3 are identified in separate fractions, while the third subunit, GCD11, is found alongside both of the other subunits. If the label free data is correct, the complex could be partially dissociating, SUI2 and SUI3 dissociating from each other, while GCD11 is still partially bound to each. The stoichiometry of the complexes and behaviour of the subunit proteins will be confirmed via SRM.

When analysing the composition of these protein complexes, the increased accuracy of the label mediated strategies and SRM approach may provide an advantage over label free analysis. Whereas label free methods scan mixtures of ions and select the most intense ions for MS/MS, the targeted nature of the QconCAT method, coupled with the advantages offered by using selected peptides, which should behave identically to the analyte peptides could provide a more accurate method for analysing protein complex stoichiometry.

The design of the experimental workflow for the QconCAT based quantifications is shown in figure 4.2.

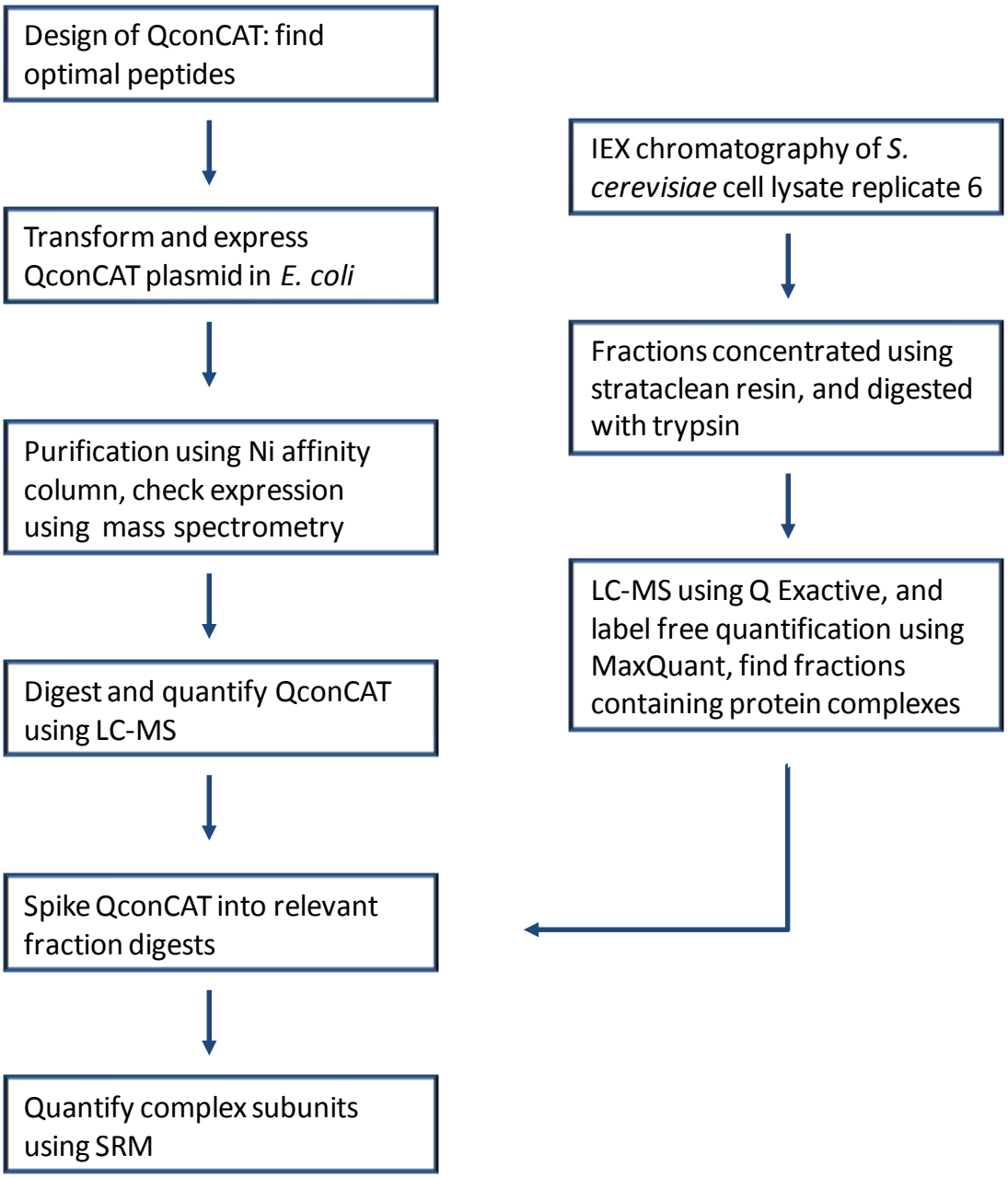


Figure 4.2 Experimental workflow

Workflow depicting the design of the experiment to compare the quantitation of protein complexes from ion exchange chromatography fractions.

4.3 Results and discussion

4.3.1 Design of a QconCAT encoding peptides from the 14 subunits of the 20S proteasome

A QconCAT was designed for the quantification of all of the subunits in the 20S proteasome. The complex consists of 14 subunits, all approximately the same size, ranging from 22.5 kDa to 31.6 kDa. In order to choose appropriate Qpeptides, first a theoretical tryptic digest was performed on the sequence of each protein. The N and C terminal peptides were immediately discarded, due to the propensity for modification, such as N terminal acetylation (Soppa, 2010) or C terminal prenylation (Zhang & Casey, 1996). The remaining peptides were then examined using a set of criteria for choosing optimal peptides. Firstly, they must be at an optimal size for detection by a mass spectrometer, and therefore, any peptides less than 6 amino acids in length were discarded (Brownridge & Beynon, 2011). They were then filtered for peptides containing an NG sequence which can undergo deamidation, an M residue which can undergo oxidation, or an N terminal Q, which is prone to partial conversion to pyroglutamic acid, dibasic residues because the C terminus of either of the amino acids could be cleaved, or any peptides with an acidic residue immediately following the tryptic cleavage site, as these are prone to miscleavages (Thiede *et al.*, 2000; Brownridge *et al.*, 2013). Any peptides with a KP or RP were filtered, as these will not be digested by trypsin, and any peptides containing C residues, which can form disulfide bridges and impair digestion. Following all of these filters, the resulting peptides were then analysed using the MC:Pred and CONSequence tools. MC:Pred predicts the likelihood of a peptide being miscleaved by trypsin. The tryptic peptides were searched using the SVM prediction type option on the MC:Pred software, and the results were filtered on a score 0.4. CONSequence predicts detectable peptides using four machine learning algorithms, and scores each peptide 1-4 according to the number of algorithms that predict that peptide will be observable. The filter applied was a CONSequence score of 3 or 4. It was also attempted to choose Qpeptides that were observed in previous label free experiments, reported by the Q Exactive generated data in Chapter 3, as it is known these peptides are observable in LC-MS experiments. There are therefore a

number of criteria, and despite starting with 19-32 peptides for each protein, due to the stringent criteria, the majority of subunits did not have two peptides which meet all of these criteria, and the peptides eliminated by filters are shown in table 4.1.

Table 4.1 Filtering of 20S proteasome tryptic peptides

Protein	Tryptic peptides	< 6 amino acids	NG	M	N term Q	Dibasic	Acidic P2	KP/RP	C	MC:Pred	CONSequence	Observed?	Usable
PRE1	21	13	0	3	0	1	1	1	1	1	0	0	1
PRE2	26	11	0	2	0	0	2	0	0	10	0	0	0
PRE3	21	9	0	2	0	0	4	0	1	5	0	0	0
PRE4	22	11	3	5	0	0	1	0	0	2	0	0	0
PRE5	24	8	0	1	2	2	2	1	1	6	1	0	0
PRE6	32	19	0	0	1	4	0	2	0	5	1	0	0
PRE7	24	12	2	1	0	3	1	0	0	5	0	0	0
PRE8	25	10	2	3	1	2	0	0	0	6	1	0	0
PRE9	27	13	0	2	1	4	0	1	0	6	0	0	0
PRE10	28	13	3	0	0	0	2	1	2	6	1	0	0
PUP1	22	11	1	3	0	0	2	0	0	4	1	0	0
PUP2	24	7	0	4	0	2	3	0	1	7	0	0	0
PUP3	19	11	0	2	0	1	3	1	0	1	0	0	0
SCL1	27	10	1	4	1	1	2	0	0	7	1	0	0

The aim was to include two peptides to represent each subunit, in order to increase both the likelihood of observing at least one, and to enable the averaging of the two peptide scores to increase reliability, so some of the filters then had to be relaxed. In these instances, firstly the MC:Pred score filter was raised to 0.5, and then peptides containing C, M or NG sequences had to be included, or in some cases, peptides that had not been previously observed in label free experiments. For example, the smallest protein, PUP3, at only 22.5 kDa, has only 19 tryptic peptides, and less than half are of sufficient length to be observable in a mass spectrometer. None of the tryptic peptides for PUP3 met all of the selection criteria, and it was necessary to choose peptides that were not observed in previous label free experiments, but that met other criteria. Other compromises made include four peptides with 'NG', for PUP1, FN**NG**VVIAADTR, for PRE7 NQYEPGT**NG**K, for PRE10 AVE**NG**TTSIGIK, and for PRE4 YD**NG**VIIAADNLGSYGSLLR. There are also three peptides included that contained methionine residues, for PRE3 MVVLTAAGVER, for PUP1 VVSALQMLK, and for PUP2 SMIEHAR. The final 28 peptides chosen were then BLAST searched against the *S. cerevisiae* proteome to check they were unique to that protein before being assembled into one recombinant protein sequence.

One of the most important considerations in SRM experiments is the aim of keeping the standard and analyte peptide conditions as alike as possible. An important aspect of this is the tryptic digestion, and if there is any discrepancy between the digestion efficiency of the two peptides, it will result in an error in the quantification results. Even in situations where the two peptides are co-digested, there is still a discrepancy in the sequence context of the two peptides, as the standard peptide is concatenated with several other peptides originating from different proteins. This can cause discrepencies in digestion efficiency, as the QconCAT protein does not form a secondary structure, as the original protein would, which may alter the availability of some peptides to proteolytic attack, and this lack of secondary structure can cause QconCAT proteins to digest at a much faster rate than the analyte proteins (Rivers *et al.*, 2007). Providing the digest has gone to completion, however, this should not affect the final results, as if the QconCAT and the original protein are both completely digested, all peptides should be exposed. A second issue with the difference in sequence context is the effect the surrounding amino acids have on the propensity for trypsin to cleave at a K or R residue. Trypsin cleaves at the C terminus of a K or R residue, except where it is followed by P, but the efficiency of the enzyme reaction differs depending on the sequence context (Brownridge & Beynon, 2011). The amino acids on either side of the trypsin specificity site affect the digestion efficiency (Siepen *et al.*, 2006). With the aim of reducing any discrepencies in digestion efficiency between standard and analyte peptides, it was decided that each Qpeptide should be flanked by the three amino acids at the N and C termini in the original sequence context of the protein. In some cases, this involved the inclusion of sequences that were suboptimal for digestion. However, the aim is to ensure the standard and analyte peptides behave in the same manner, so if the analyte peptide is miscleaved, the standard should be to the same extent, to reduce the likelihood of a false result. The final QconCAT sequence consisted of 28 Qpeptides, flanked by 3 amino acid linker sequences (Table 4.2).

Table 4.2 20S proteasome peptides included in the QconCAT

Complex	Protein	Accession	Sequence	N term	C term
20S	PRE1	P22141	VQDSVILASSK	GIR	AVT
20S	PRE1	P22141	GISVLK	VTR	DSD
20S	PRE2	P30656	FQGGIIVAVDSR	AFR	ATA
20S	PRE2	P30656	ILSNLVYQYK	ASK	GAG
20S	PRE3	P38624	WDGSSGGVIR	AIK	MVV
20S	PRE3	P38624	MVVLTAAGVER	VIR	LIF
20S	PRE4	P30657	YDNGVIIAADNLGSYGSLLR	SMK	FNG
20S	PRE4	P30657	NFSLAIIDK	SSR	NTG
20S	PRE5	P40302	LFQVEYALEAIK	TGR	QGS
20S	PRE5	P40302	SGAHLLEFQPSGNVTELYGTAIGAR	YDK	SQG
20S	PRE6	P40303	LTLEDPVTVEYLTR	SHR	YVA
20S	PRE6	P40303	YVAGVQQR	LTR	YTQ
20S	PRE7	P23724	NITDYSINSR	DTR	YEP
20S	PRE7	P23724	NQYEPGTNGK	NFK	VKK
20S	PRE8	P23639	LLVSEVAK	PTK	IMQ
20S	PRE8	P23639	LGQIDYALTAVK	SGK	QGV
20S	PRE9	P23638	IHAQNYLK	TAR	TYN
20S	PRE9	P23638	YGYQLYTSNPSGNYTGWK	DDR	AIS
20S	PRE10	P21242	AVENGTTSIGIK	AVK	CND
20S	PRe10	P21242	HIGCVYSGLIPDGR	VDR	HLV
20S	PUP1	P25043	VVSALQMLK	EPR	QHL
20S	PUP1	P25043	FNNGVVIAADTR	GVK	STQ
20S	PUP2	P32379	IYDNEK	GFK	TAE
20S	PUP2	P32379	SMIEHAR	DAR	TAA
20S	PUP3	P25451	IFHYGHVFLGITGLATDVTTLNEMFR	FEK	YKT
20S	PUP3	P25451	SGKPFIAGFDLIGCIDEAK	NSK	DFI
20S	SCL1	P21243	HITIFSPEGR	YDR	LYQ
20S	SCL1	P21243	ATNQTNINSLAVR	AFK	GKD

4.3.2 Design of a QconCAT encoding peptides from the eIF2 complex

Another QconCAT, termed YEW1, was designed to encode all the subunits from the translation initiation factor 2 complex. Theoretical digests were performed using Pepmapper software and the resulting list of tryptic peptides was filtered using the same criteria as above (Table 4.3). Once again there are few peptides which pass all of the filters applied. The criteria again had to be relaxed so that two peptides per protein could be incorporated.

Table 4.3 Filtering of eIF2 tryptic peptides

Protein	Tryptic peptides	< 6 amino acids	NG	M	N term Q	Dibasic	Acidic P2	KP/RP	C	MC:Pred	CONSequence	Observed?	Usable
GCD11	62	36	1	3	1	3	2	4	2	9	0	0	1
SUI2	42	23	0	4	0	4	3	0	1	6	1	0	0
SUI3	54	35	2	2	0	6	4	0	1	4	0	0	0
IKI3	148	78	4	13	3	11	15	1	4	21	0	0	0
ELP2	83	33	2	11	2	4	6	1	7	12	2	0	0
ELP3	80	47	3	6	1	4	8	0	2	5	2	0	0
ELP4	61	38	1	5	0	4	7	0	0	6	0	0	0
ELP5	24	12	1	4	1	0	2	0	1	3	0	0	0
ELP6	16	6	0	2	0	0	2	1	1	4	0	0	0

For the translation initiation factor 2 subunits SUI2 and SUI3 there were some suboptimal peptides included. SUI2 contains both FQIPLEELYK, which has a low CONSequence score, indicating it may not be observable in a mass spectrometer, however the other most suitable option, YGGVCNITMPPK, contains a methionine. As the QconCAT construct size will allow the addition of extra peptides, in this case both options were included, taking the SUI2 Qpeptide count up to three. For SUI3 one of the peptide options, EGTPSANSSIQQEVGLPYSELLSR, contains an acidic residue following the tryptic cleavage site, which may result in a missed cleavage, so an additional peptide, LFFMVCK, was again included, taking the peptide count to three. This increases the chances of at least two peptides being observed.

After the filter criteria were relaxed, there were a total of eight EIF2 Qpeptides for incorporation into the QconCAT (table 4.4). In addition to these eight peptides, 12 peptides were included from proteins of the elongator complex, however data from these is not shown in this thesis.

Table 4.4 eIF2 peptides included in the QconCAT

Complex	Protein	Accession	Sequence	N term	C term
EIF2	GCD11	P32481	LNPLSAEIINR	FSK	QAT
EIF2	GCD11	P32481	YNIDAVNEFIVK	QLK	TIP
EIF2	SUI2	P20459	AVTATEDAELQALLESK	PPK	ELD
EIF2	SUI2	P20459	YGGVCNITMPPK	ITK	AVT
EIF2	SUI2	P20459	FQIPLEELYK	AEK	TIA
EIF2	SUI3	P09064	AGLDNVDAESK	LAK	EGT
EIF2	SUI3	P09064	LFFMVCK	SNR	SCG
EIF2	SUI3	P09064	EGTPSANSSIQQEVGLPYSELLSR	ESK	FFN

The three amino acids at the N and C terminus of the peptides were included to provide the same sequence context as in the original protein. In some cases, the sequence context was not optimal for digestion, and several of the three amino acid sequences contain an acidic residue before the trypsin cleavage site. Also, the inclusion of the tripeptide sequences involves incorporating more tryptic cleavage sites than necessary, as there were a number of additional K or R residues. This could cause higher numbers of miscleavages, but as the original sequence context was preserved as far as possible, any miscleavages should be to the same extent as the analyte peptide, preserving the relationship between analyte and standard signals, and preventing any error in quantification. This approach is based on the PCS strategy derived by Kito *et al*. (2007).

The design of both YEW1 and YEW3 QconCATs included an mApple sequence at the N terminus of the Qpeptides (figure 4.3). The mApple protein is a fluorescent monomer which was included with the aim of quantifying the QconCATs by absorbance. Unfortunately, it was not possible to pursue this concept in the time frame of this thesis. At the C terminus of the Qpeptides a glufibrinopeptide sequence was included for quantification by mass spectrometry, and a hexa-histidine tag for purification.

4.3.3 Expression of YEW1 and YEW3 QconCATs

YEW1 and YEW3 constructs were expressed in *E.coli* in minimal media containing heavy isotope labelled arginine and lysine as described in section 2.11. Expression of the QconCAT was confirmed by gel electrophoresis (Figure 4.4 a and b), where an additional band was visible in the post induction lanes of both cultures, at approximately 90 kDa, close to the molecular weights of the full QconCAT constructs. The suspected QconCAT proteins were then purified using the hexa-histidine tag built into the construct, by Nickel affinity chromatography (for method see section 2.12). Purification of the protein was confirmed using gel electrophoresis (figure 4.4 c and d), where there was one major protein band in the elution lanes. To provide confirmation of the expression of the correct QconCAT sequence, the gel bands of the suspected QconCATs were excised, and in gel

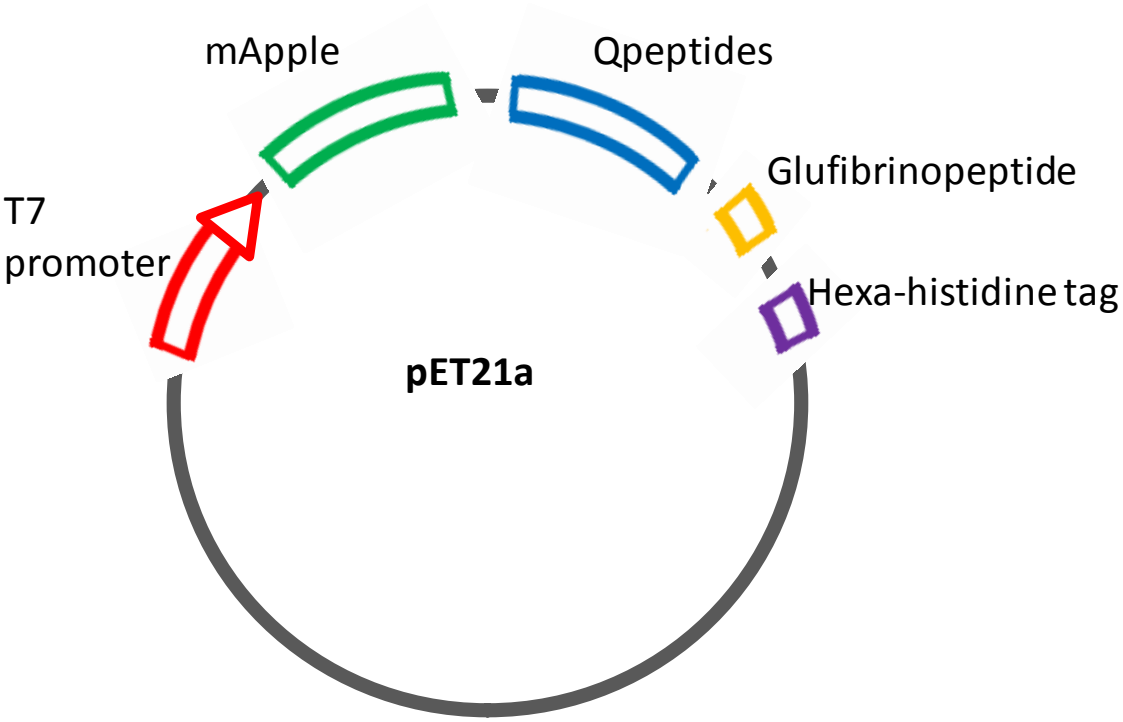


Figure 4.3 QconCAT construct design

Each construct consisted of the pET21a plasmid with genes inserted downstream of the T7 promoter. Genes inserted were mApple sequence, followed by the chosen Qpeptides, a glufibrinopeptide sequence for quantification and a hexa-histidine tag for purification.

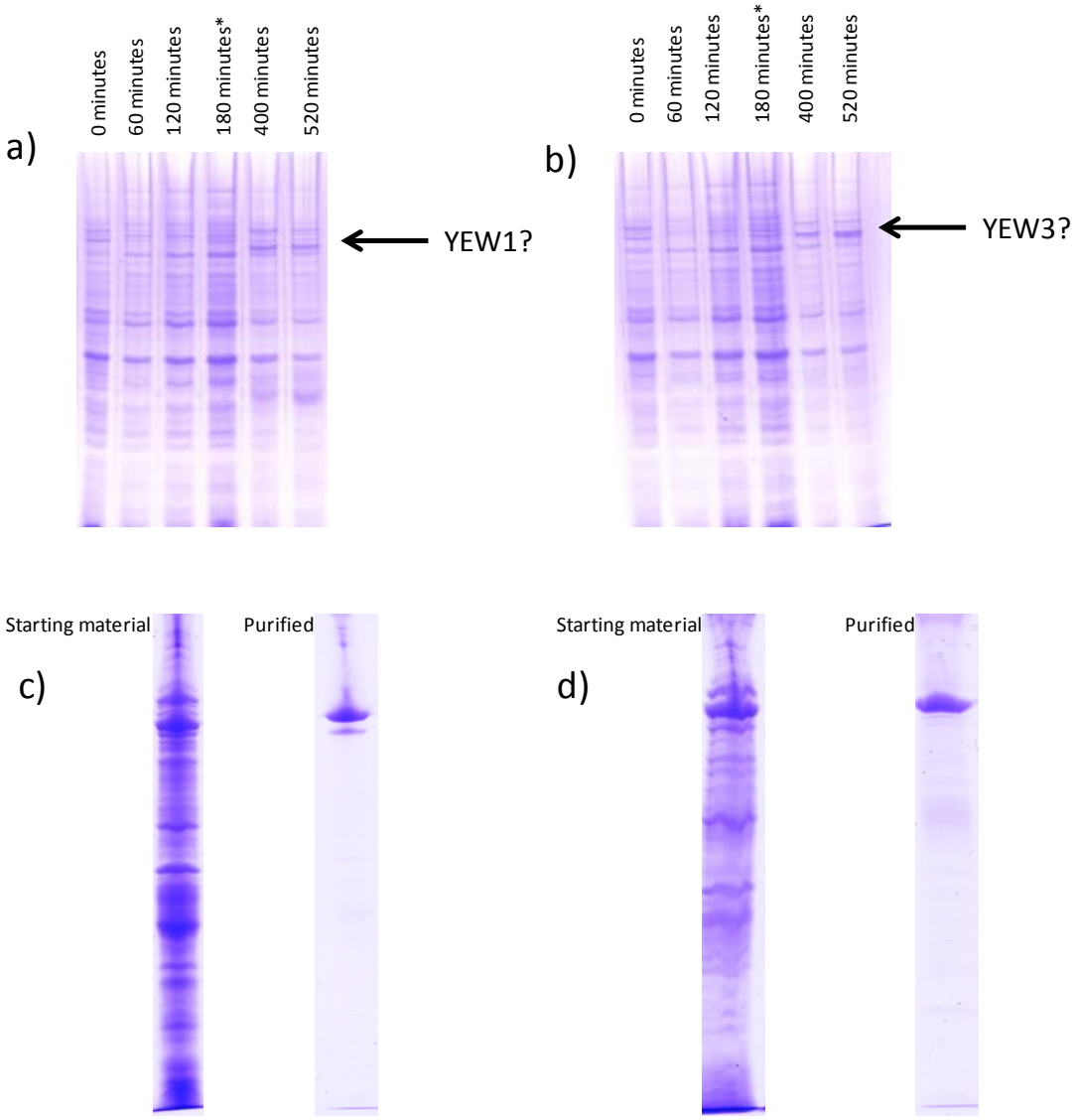


Figure 4.4 SDS PAGE of expression and purification of YEW1 and YEW3

Plasmids were transformed into *E.coli* and cultured in minimal media with heavy isotope labelled lysine and arginine. Samples were collected until an OD $_{600nm}$ of 0.6, when 1 mM IPTG was added, and culture continued for 4 hours. Samples from a) YEW1 and b)YEW3 time points are run on SDS PAGE. Potential QconCAT bands are highlighted. Cultures were purified using Ni affinity columns and a band is evident for c) YEW1 and d)YEW3.

digestions were performed. The digests were checked by MALDI-TOF using a Bruker Ultraflextreme, and analysed using Bruker Flexanalysis. For the suspected YEW1 band, peptides from the entire length of the construct are identified, which confirms expression of the complete Qpeptide sequence. Six of the EIF2 Qpeptides were identified, and the two that were not identified were both in the middle of the construct, and since the presence of the entire construct has been confirmed it is unlikely that the peptides are not expressed. In addition, both of the missing peptides are lysine terminated, which are less detectable by MALDI-TOF than arginine terminated peptides (Hale *et al.*, 2000).

The purified YEW1 protein also showed a second band, which was confirmed as the same protein by the identification of the Qpeptides (figure 4.5). The second band is therefore a result of the same protein running at a faster rate through the gel, possibly due to the loss of a peptide reducing the mass of the protein, or it could be due to a conformational change. Due to the nature of the QconCAT, an artificially constructed protein made up of short peptide sequences, the protein will not adopt a secondary structure. However, the QconCAT design includes an N terminal mApple fluorescent protein sequence, which does adopt a secondary structure. The fusion of a monomer with a secondary structure to a sequence of Qpeptides may cause conformational changes in the mApple sequence. If any part of this structure remains partially folded during electrophoresis, it may manifest as the second band. Whether the reason behind the double band is the loss of an N terminal peptide, or structural differences, the identification of Qpeptides from the N and C terminus of the QconCAT section of the construct indicates the complete sequence is present. MALDI-TOF also confirmed the expression of 20 YEW3 Qpeptides from the entire length of the construct (figure 4.6).

The purified protein was then tryptically digested and confirmation of the expression was also achieved by LC-MS using a Waters Synapt G2, and analysis using Waters ProteinLynx Global Server software (PLGS), which identified the protein as YEW1 and YEW3 using a QconCAT database. When YEW1 data was analysed using PLGS, all of the EIF2 complex peptides were observed, which confirms the expression of the correct QconCAT. All of the peptides were observed

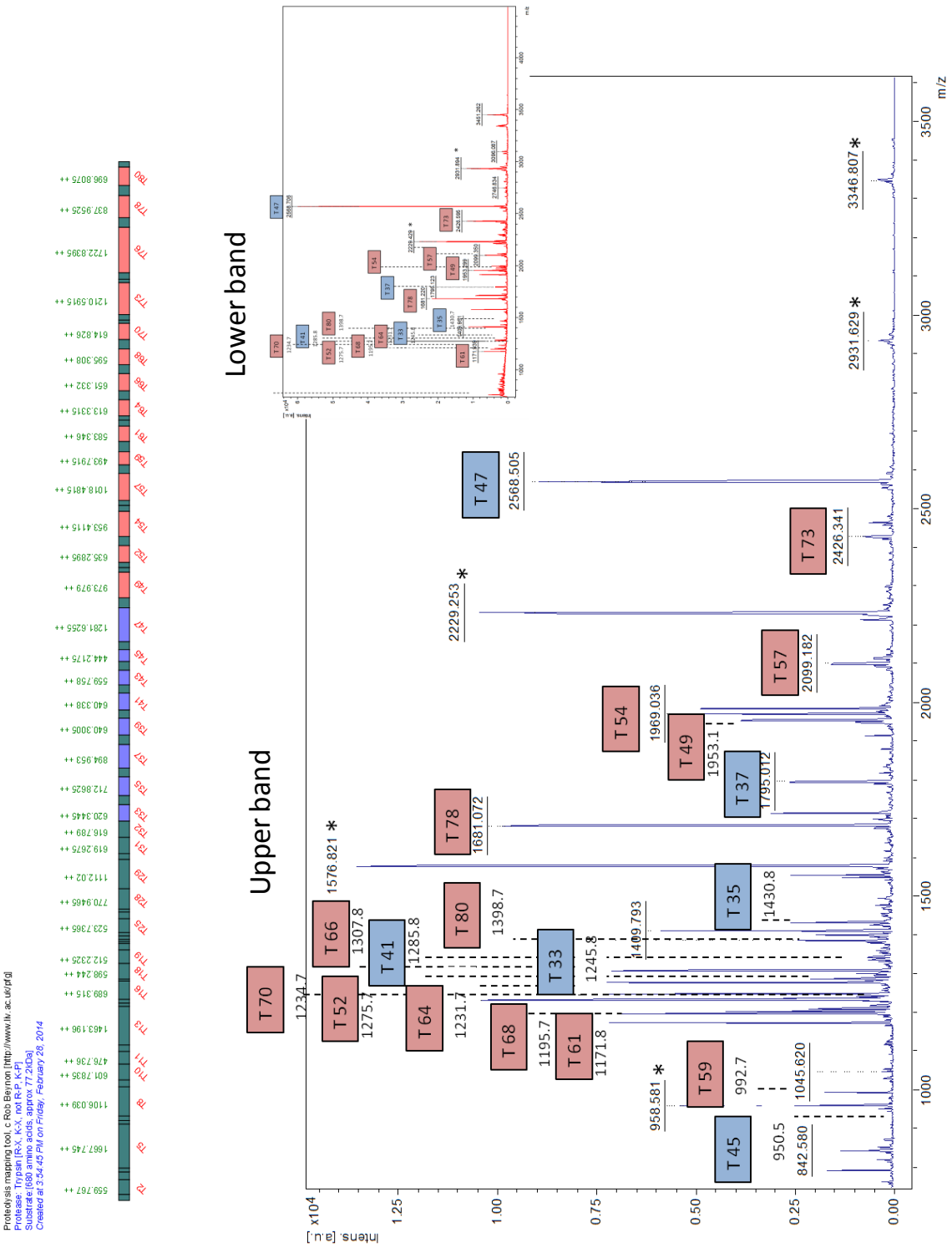


Figure 4.5 MALDI TOF confirmation of YEW1 expression

Peptide map depicts the QconCAT protein, containing mApple gene (grey), EIF2 complex genes (pink) and elongator complex genes (blue). The possible QconCAT band was excised from the gel in figure 4.4, and an in gel digestion was performed using trypsin. The digest was analysed by MALDI TOF using a Bruker Ultraflextreme, peaks corresponding to Qpeptides are highlighted (EIF2 in blue, elongator in pink), and mApple peptides are marked with *.

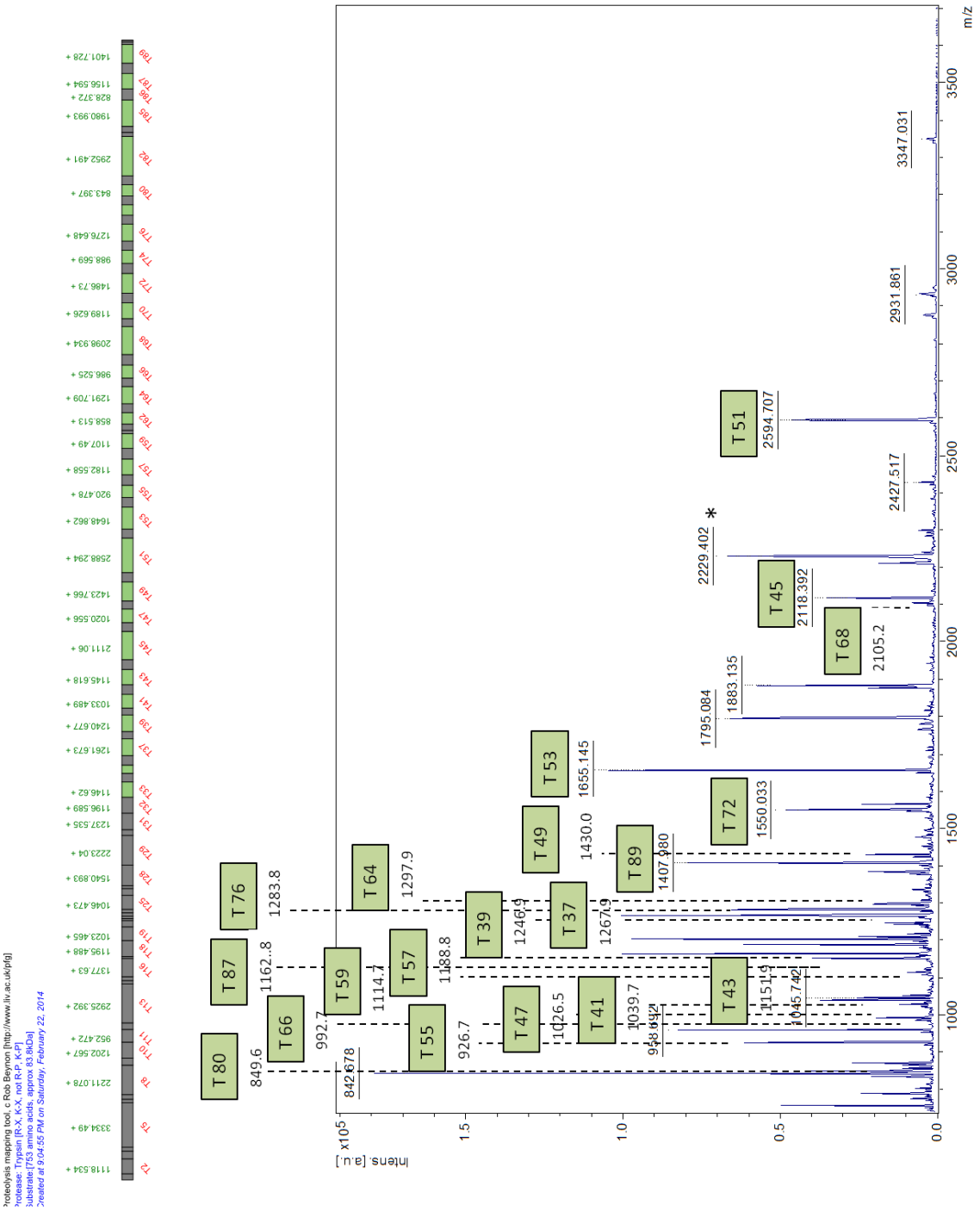


Figure 4.6 MALDI TOF confirmation of YEW3 expression.

Peptide map depicts the QconCAT protein, containing mApple gene (grey) and 20S proteasome genes (green). The suspected QconCAT band was excised from the gel in figure 4.4, and an in gel digestion was performed using trypsin. The digest was analysed by MALDI TOF using a Bruker Ultraflextreme, and peaks corresponding to 20S proteasome complex Qpeptides are highlighted in green. mApple peptides are marked with *.

as doubly or triply charged ions, and the methionine residues included were observed at their original mass, and had not undergone oxidation. There was one peptide identified by PLGS in a missed cleaved form, AGLDNVDAESKEGTSNR. This is likely the result of having an acidic residue, glutamic acid, following the trypsin cleavage site. However, as the Qpeptide is in the sequence context of the original protein the analyte peptide would be expected to show the same degree of missed cleavage, therefore the QconCAT digest was used in the quantification.

When heavy labelled QconCATs are generated, it is usual for there to be a small amount of unlabelled peptide present. This is the case for each of the YEW1 heavy labelled peptides observed, where there are also peaks representing the *m/z* of the unlabelled peptides, indicating the heavy labelling is incomplete (figure 4.7). The effect is more pronounced for lysine terminated peptides than arginine. In order to use the heavy labelled peptides, it was therefore necessary to calculate the amount of signal contribution from unlabelled QconCAT. At the same time as running the sample and QconCAT mixture on the mass spectrometer a control sample of the QconCAT dilution without sample was run. For each individual peptide the amount of unlabelled signal was calculated. The signal contribution from unlabelled QconCAT was then subtracted from the quantity of each analyte peptide observed. This calculation was done for each individual peptide, regardless of whether lysine or arginine terminated.

For the YEW3 digest, 27 of the 28 Qpeptides were observed in the G2 data, IFHYGHVFLGITGLATDVTTLNEMFR was not (figures 4.8-4.11). The presence of 27 peptides confirmed expression of the correct protein. None of the 20S proteasome Qpeptides were seen in miscleaved forms in the PLGS search results. Once again it was apparent that some of the peptides were not completely labelled, there were peaks observed at the *m/z* of the light peptide, and the unlabelled peaks were more intense in the lysine terminated peptides. Therefore, when the QconCAT and analyte mixtures were analysed by SRM, a control sample was also run, and the intensity of the light peak in the control sample subtracted from the sample peak intensity. One peptide, SGKPFIAGFDLIGCIDEAK, has a proline residue following a

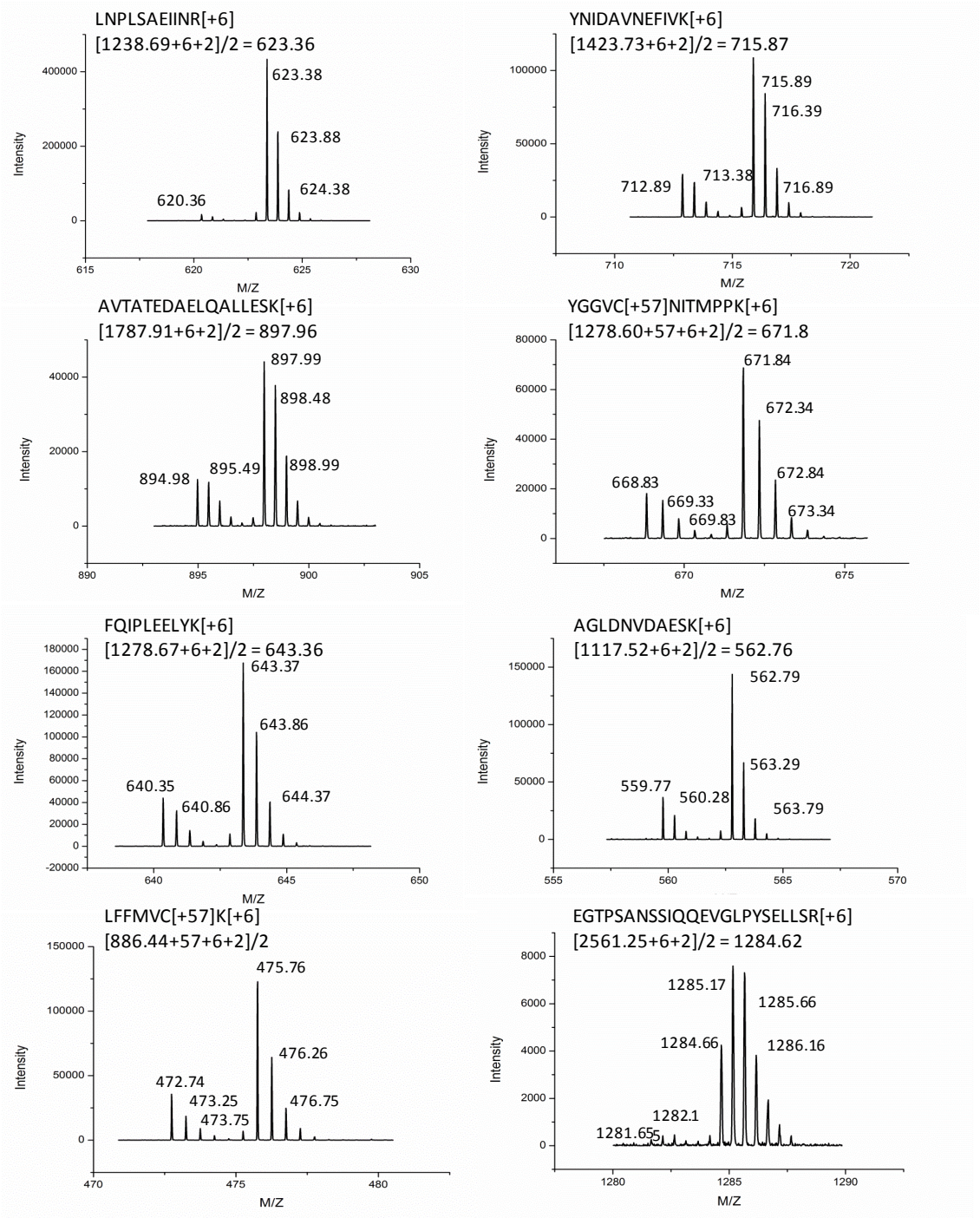


Figure 4.7 YEW1 Qpeptide labelling identified by LC-MS

Purified YEW1 QconCAT was digested using a trypsin in solution digest. The digested protein was then analysed by LC-MS using a Waters Synapt G2 to check the digestion efficiency and the heavy isotope labelling. The eight Qpeptides encoding EIF2 protein sequences are shown.

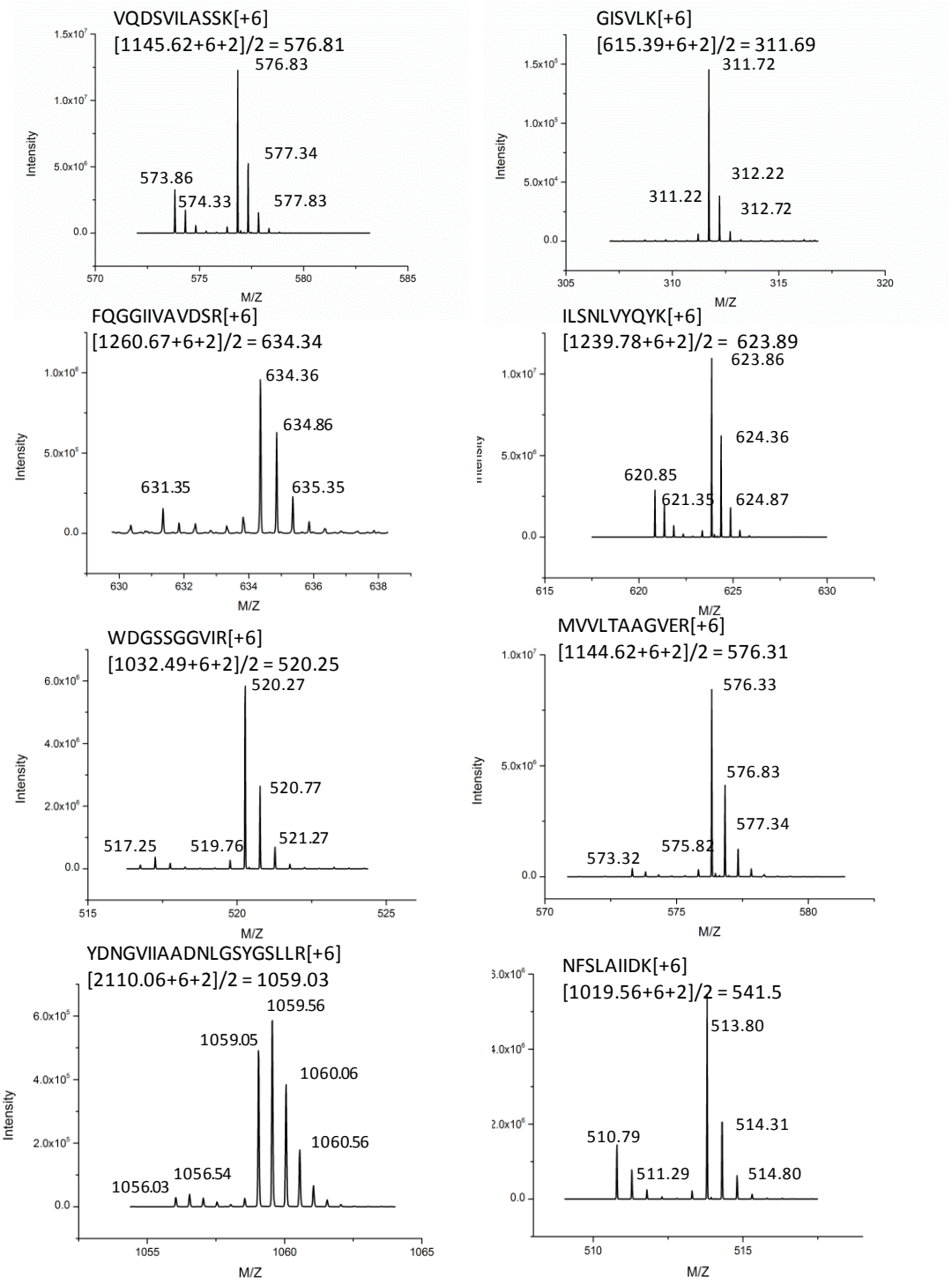


Figure 4.8 YEW3 Qpeptide labelling identified by LC-MS

Purified YEW3 protein was subjected to in solution digestion using trypsin and run on a Waters Synapt G2 on a 1 hour gradient in MS^E mode. Results were analysed using PLGS to check coverage and digestion efficiency, and an extracted ion chromatogram generated for each Qpeptide.

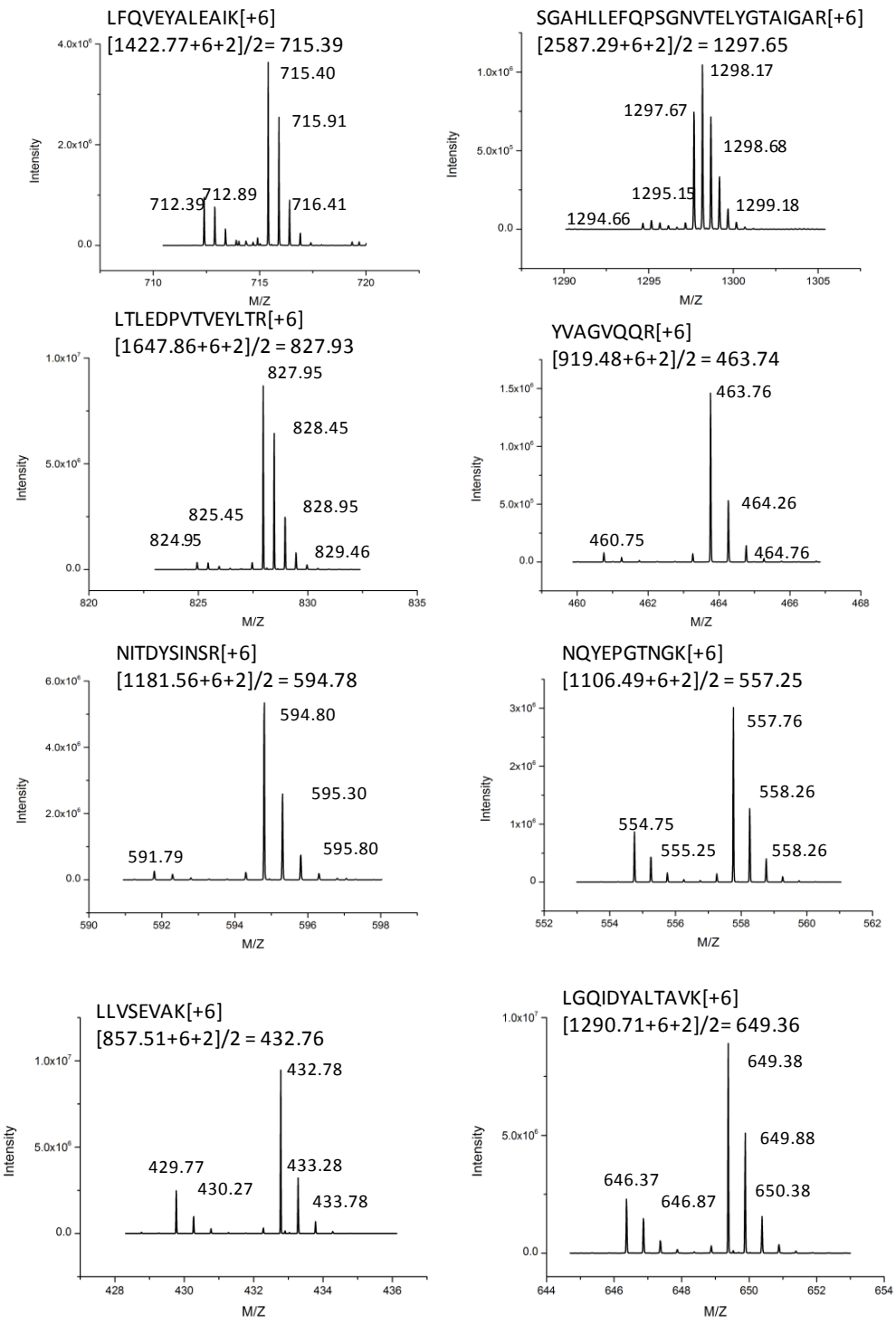


Figure 4.9 YEW3 Qpeptide labelling identified by LC-MS

Purified YEW3 protein was subjected to in solution digestion using trypsin and run on a Waters Synapt G2 on a 1 hour gradient in MS^E mode. Results were analysed using PLGS to check coverage and digestion efficiency, and an extracted ion chromatogram generated for each Qpeptide.

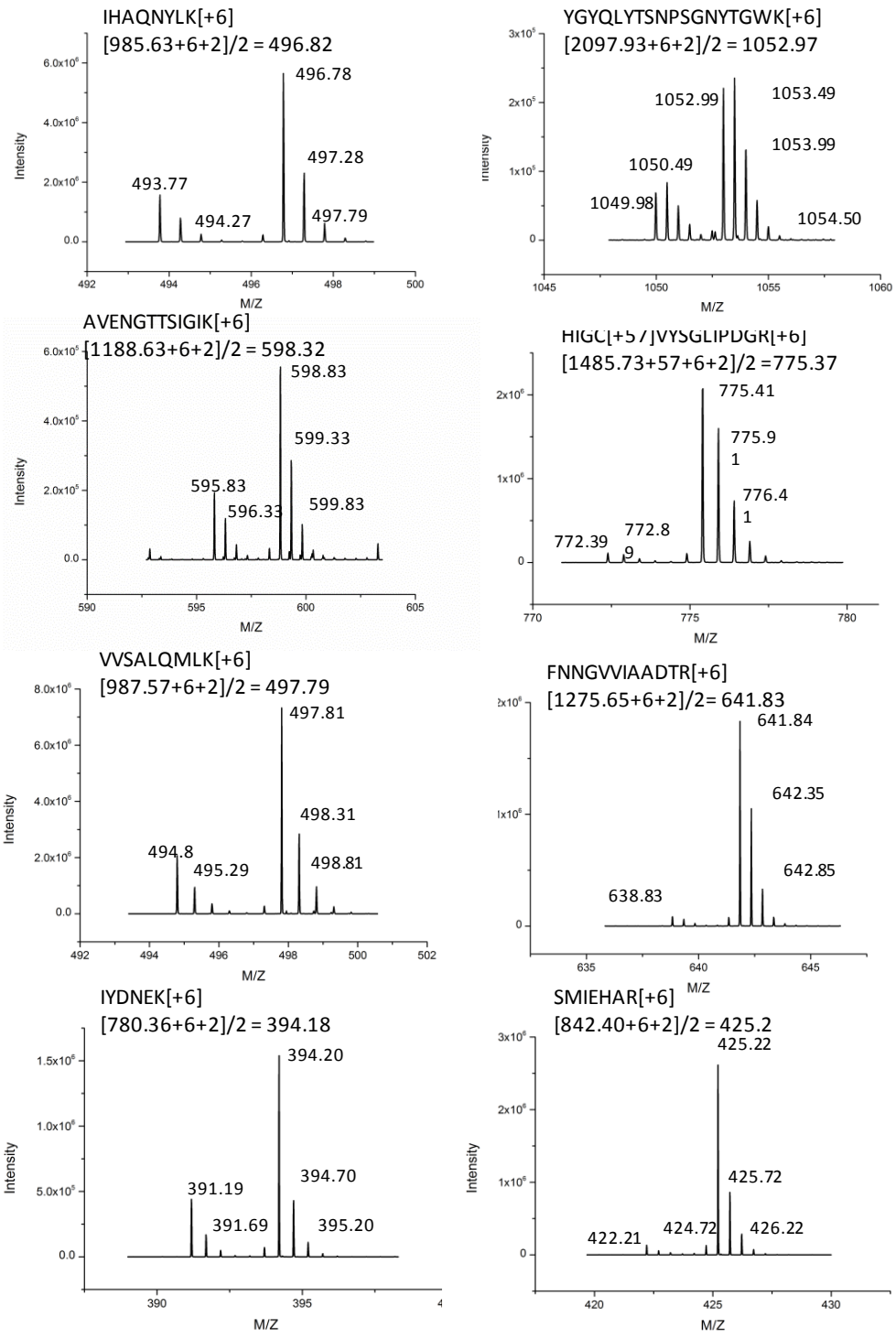


Figure 4.10 YEW3 Qpeptide labelling identified by LC-MS

Purified YEW3 protein was subjected to in solution digestion using trypsin and run on a Waters Synapt G2 on a 1 hour gradient in MS^E mode. Results were analysed using PLGS to check coverage and digestion efficiency, and an extracted ion chromatogram generated for each Qpeptide.

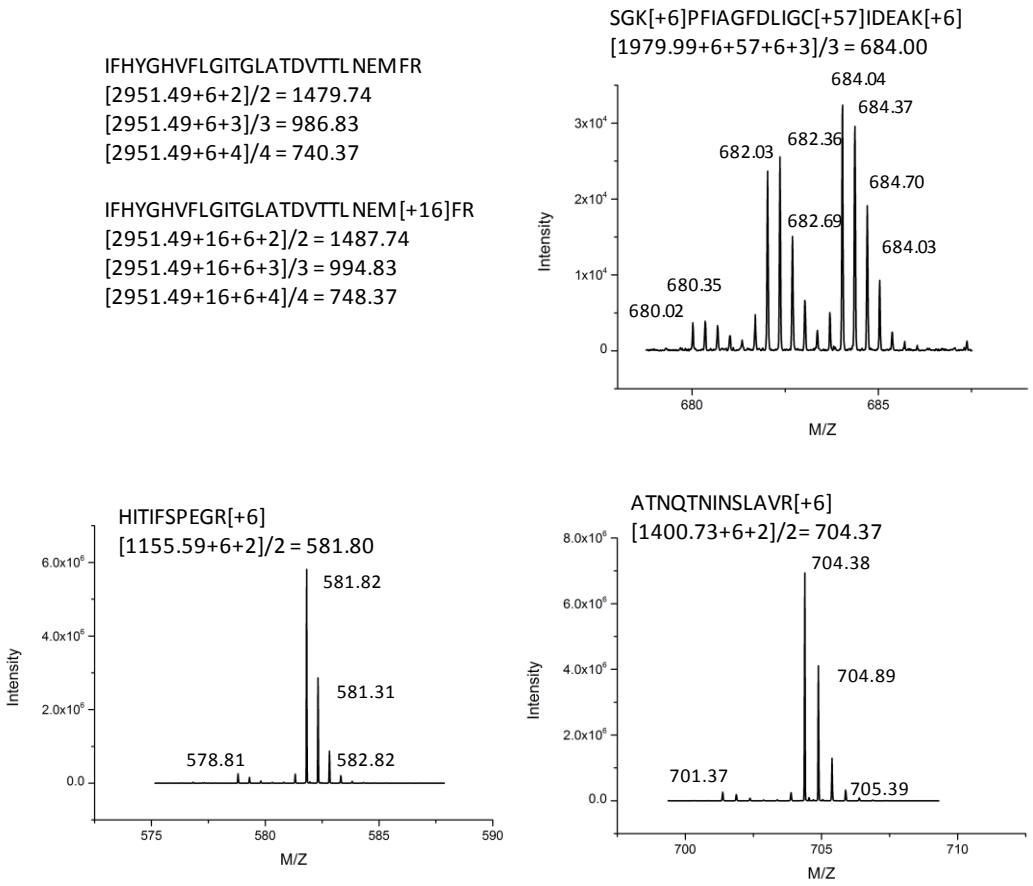


Figure 4.11 YEW3 Qpeptide labelling identified by LC-MS

Purified YEW3 protein was subjected to in solution digestion using trypsin and run on a Waters Synapt G2 on a 1 hour gradient in MS^E mode. Results were analysed using PLGS to check coverage and digestion efficiency, and an extracted ion chromatogram generated for each Qpeptide, for which the spectra are shown, except in the case of the peptide IFHYGHVFLGITGLATDVTTLNEMFR, which was not found.

lysine. In this instance, the peptide was observed as a triply charged ion, containing two lysine residues which have accepted a positive charge.

4.3.4 eIF2 complex quantification

A QconCAT targeting the eIF2 complex was designed because the label free data indicated the complex, which is a heterotrimer, was not stoichiometrically equal (figure 4.12). In addition, the data indicates that there is some dissociation occurring, which could be responsible for the lack of stoichiometry observed.

YEW1 QconCAT dilutions were prepared using fraction 14-22 digests, which, according to the label free analysis, contain the eIF2 complex subunits. SRM experiments were performed on the Waters Xevo, and analysed in MassLynx software. For transitions used see supplementary material. A control sample was run, YEW1 diluted to the same concentration as the analyte sample. However, when run on the Xevo there is a difference in retention time between the fraction digest diluted QconCAT and the control sample without a yeast background. Therefore, the G2 data was used to calculate the percentage of light signal contributed by the unlabelled QconCAT. The percentage of light peptide present was calculated for each individual peptide. When the SRM data was analysed, the peak area contributed by the unlabelled QconCAT was subtracted from the analyte peak area before the quantity of analyte peptide was calculated.

The standard and analyte pairs were placed into three categories. Type A, when both the standard and analyte peaks are observed, Type B when just the standard is, but not the analyte, and Type C, when neither peptide is observed (Brownridge *et al.*, 2013). Prior to quantification, each of the peptides were assessed and labelled as one of these categories. Where Type C peptides are observed, it is not possible to quantify. With Type B peptides, it is not possible to obtain an accurate quantification, but it is possible to suggest that the analyte protein is at a lower concentration than the standard, and therefore set a maximum concentration of analyte in the sample.

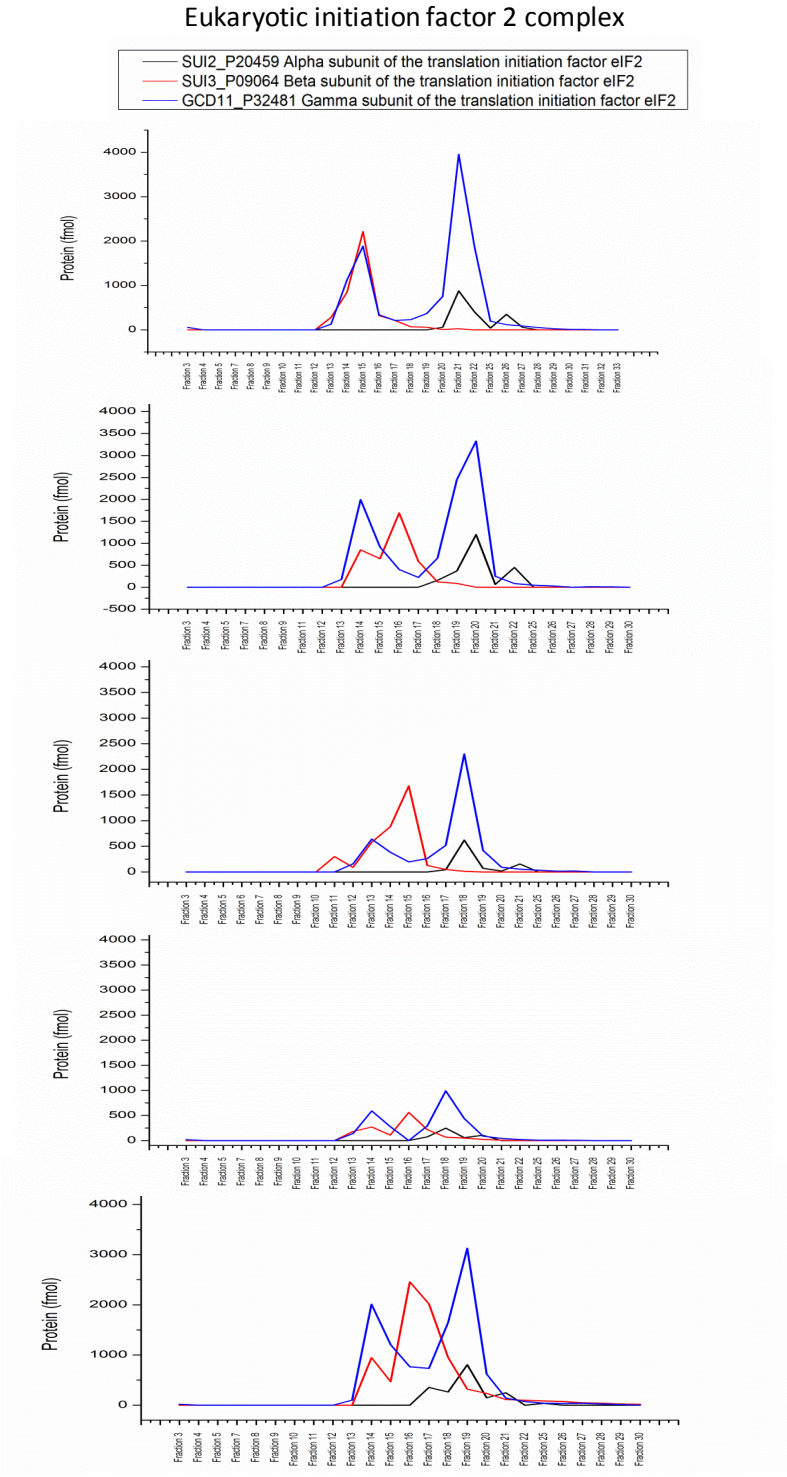


Figure 4.12 Label free quantification results for the EIF2 complex subunits in five replicates of IEX fractionation

Five replicates of IEX fractionation were collected, the protein content of each fraction was concentrated using StrataClean resin, and then digested using trypsin. The digest was then mixed with standard glycogen phosphorylase B, and run on the Qexactive using a top ten data dependent acquisition method. The results were processed using MaxQuant software.

The peptide YGGVCNITMPPK was discarded from the experiment due to being a Type C peptide, as neither the heavy nor light transitions were observed. The peptide EGTPSANSSIQQEVGLPYSELLSR was a Type B, the QconCAT peak was observed, but not the analyte peak. Where Type B peptides were observed, it is indicative of the analyte peptide being present at too low a concentration to be observed, possibly because it has undergone some sort of post translational modification or poor digestion. As the standard peptide is observed, at a concentration of 0.25 fmol, it is then possible to set this as the maximum amount of analyte peptide. However, as two other peptides for this protein were included, and peaks for these were observed, they were used to quantify SUI3. All other peptides were Type A (figure 4.13).

One thing to note within the peptide quantifications is that the where two peptides are observed, there is a difference in the ratio of the two peptides to each other in each fraction. Theoretically, the peptide should ionise to the same extent in each sample, and maintain a steady ratio across the samples. The difference is likely due to ion suppression (or ion-enhancement) effects, where a difference in the sample composition can cause a peptide to ionise to a greater or lesser extent. The effect is well documented, across a range of sample and instrument types (Buhrman *et al.*, 1996; Matuszewski *et al.*, 2003; Chambers *et al.*, 2007). As there are different numbers of proteins in each fraction, the sample background differs, with a variation in peptide content of around 2000-4000 peptides (figure 4.14). As the variation in both the size and content of sample background differs, the propensity of each peptide to ionise will differ in each digest. This phenomenon is known as matrix effects, and theories as to the cause include competition for charges, or competition for the surface area of the droplet during the transition into gas phase ions (reviewed in Gosetti *et al.*, 2010). Matrix effects can therefore cause variation even between replicates, or within sets of the same sample, and can therefore cause some differences in the QconCAT results from different fractions presented here.

For the subunit SUI2, two peptides were quantified, AVTATEDAELQALLESK and FQIPLEELYK. Both peptides reach the highest concentration in fraction 19, however

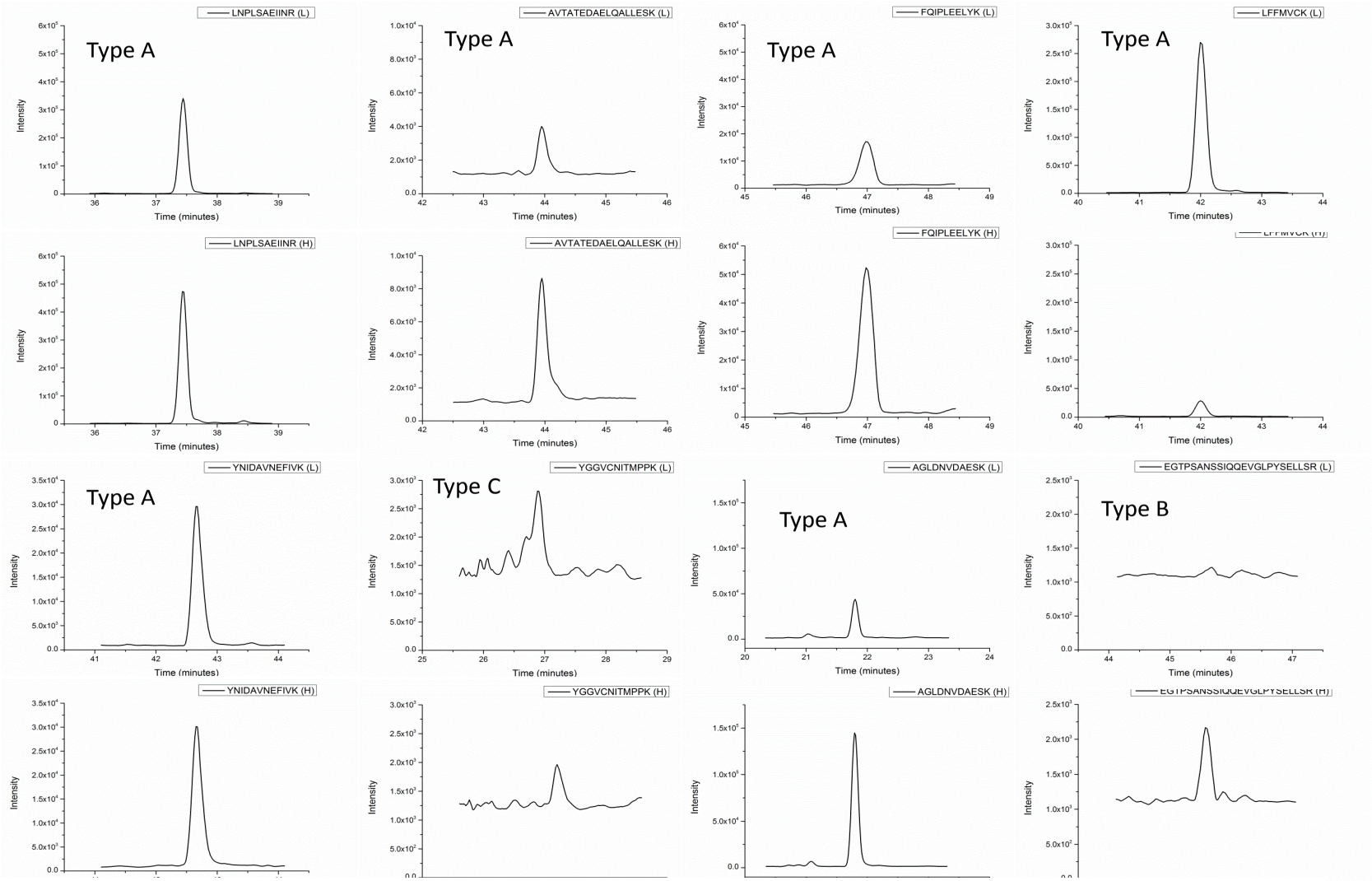


Figure 4.13 Heavy and light eIF2 peptide pairs

Heavy and light peptides observed on the Xevo, displayed in Masslynx software. Examples are shown of each peptide, in the standard(H) and analyte(L). Each pair is labeled with the peptide type.

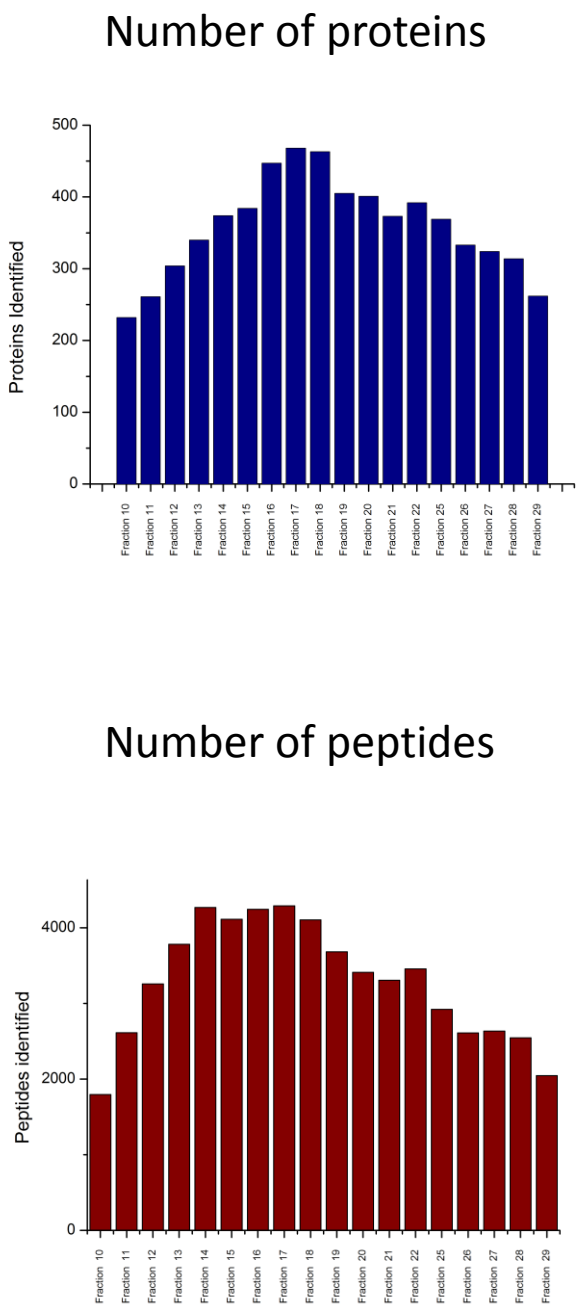


Figure 4.14 Protein and peptides identified in IEX fractions replicate 6 by the QExactive

IEX fractionation was performed on a sixth replicate of *S. cerevisiae* cell lysate for use in SRM experiments. Fractions 10-29 were concentrated using StrataClean resin and run on the Q Exactive before label free quantification with MaxQuant software. The variation in the number of proteins and the number of peptides in each fraction is depicted.

FQIPLEELYK is detected at a lower concentration than AVTATEDAELQALLESK, at 1.64 fmol compared to 5.35 fmol (figure 4.15). In the label free analysis 15 peptides are identified from SUI2. These peptides are most intense in fraction 19, so the two quantification methods support the fact that the majority of SUI2 is eluting in fraction 19. The two peptides included in the QconCAT are both identified in the label free data, so it is possible to compare two quantification methods for the individual peptides. In the label free data, both of the peptides show a wide distribution, eluting across five fractions. AVTATEDAELQALLESK is most intense in fraction 22, but is eluted across fractions 17-22, and FQIPLEELYK is most intense in fraction 19, and is less abundant than the other peptide. Extracted ion chromatograms show the highest intensity of both peptides in fraction 19 (figure 4.16). In both the label free and the QconCAT data FQIPLEELYK is at a lower concentration, which is also supported by the extracted ion chromatogram, suggesting it has less propensity for ionisation than AVTATEDAELQALLESK. Despite this discrepancy, the label free and SRM analyses both support SUI2 eluting in fraction 19.

For the subunit SUI3, two peptides were able to be quantified, LFFMVCK and AGLDNVDAESK. There is a discrepancy between the results of these two peptide quantifications (figure 4.17). The peptide AGLDNVDAESK was identified as a missed cleavage in the QconCAT G2 data, possibly a result of having an acidic residue, glutamic acid, in the next position to the trypsin site, however, the intensity of the non miscleaved form is in the range of an order of magnitude greater than the miscleaved (figure 4.18 a). The QconCAT peptide is in the same sequence context as the analyte peptide, and therefore in theory both the standard and analyte could be miscleaved. However, because in this experiment the two were digested separately and the QconCAT is spiked into the fraction digest, it is possible that they will undergo varying degrees of digestion. Therefore, one or other of the digests may go further towards completion and cause a discrepancy in the results. The low amounts of analyte quantified suggest that if there was a discrepancy in the digest efficiency, it has caused an abnormally low analyte signal, suggesting a miscleave in the analyte peptide. In the original sequence context, the Qpeptide is N terminal to

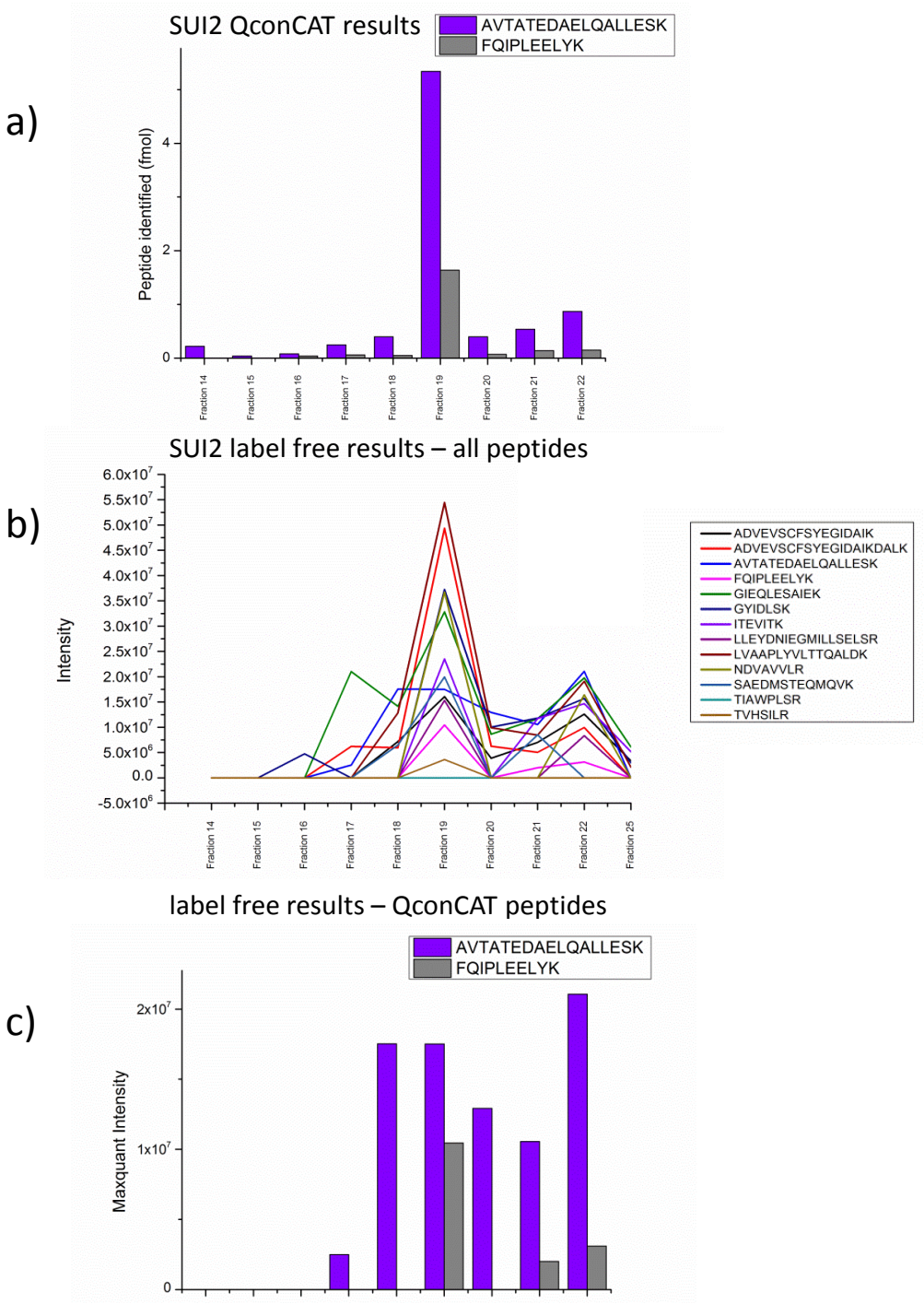


Figure 4.15 SUI2 peptide quantification by both QconCAT and label free methods

The two peptides quantified by QconCAT methodology are shown in a), while b) shows the intensity of all of the EIF2α peptides identified by label free analysis. c) both of the QconCAT peptides are also identified by MaxQuant processing of the Qexactive data.

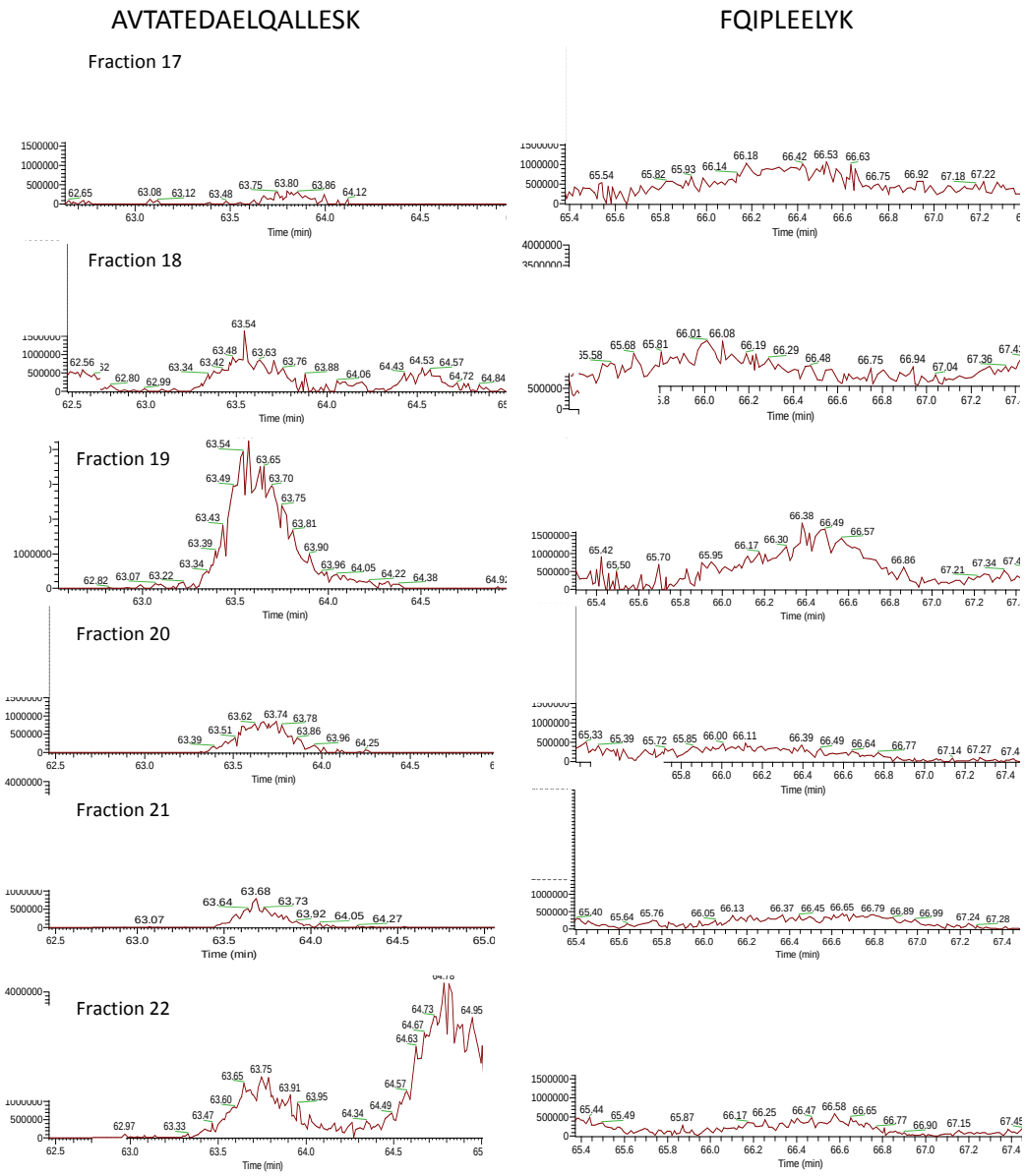


Figure 4.16 Extracted ion chromatograms of SUI2 peptides

Extracted ion chromatograms of the two SUI2 peptides that were included in the QconCAT. Fractions were run on the Q Exactive in a data dependent acquisition mode and extracted ion chromatograms were performed using Thermo Xcalibur software.

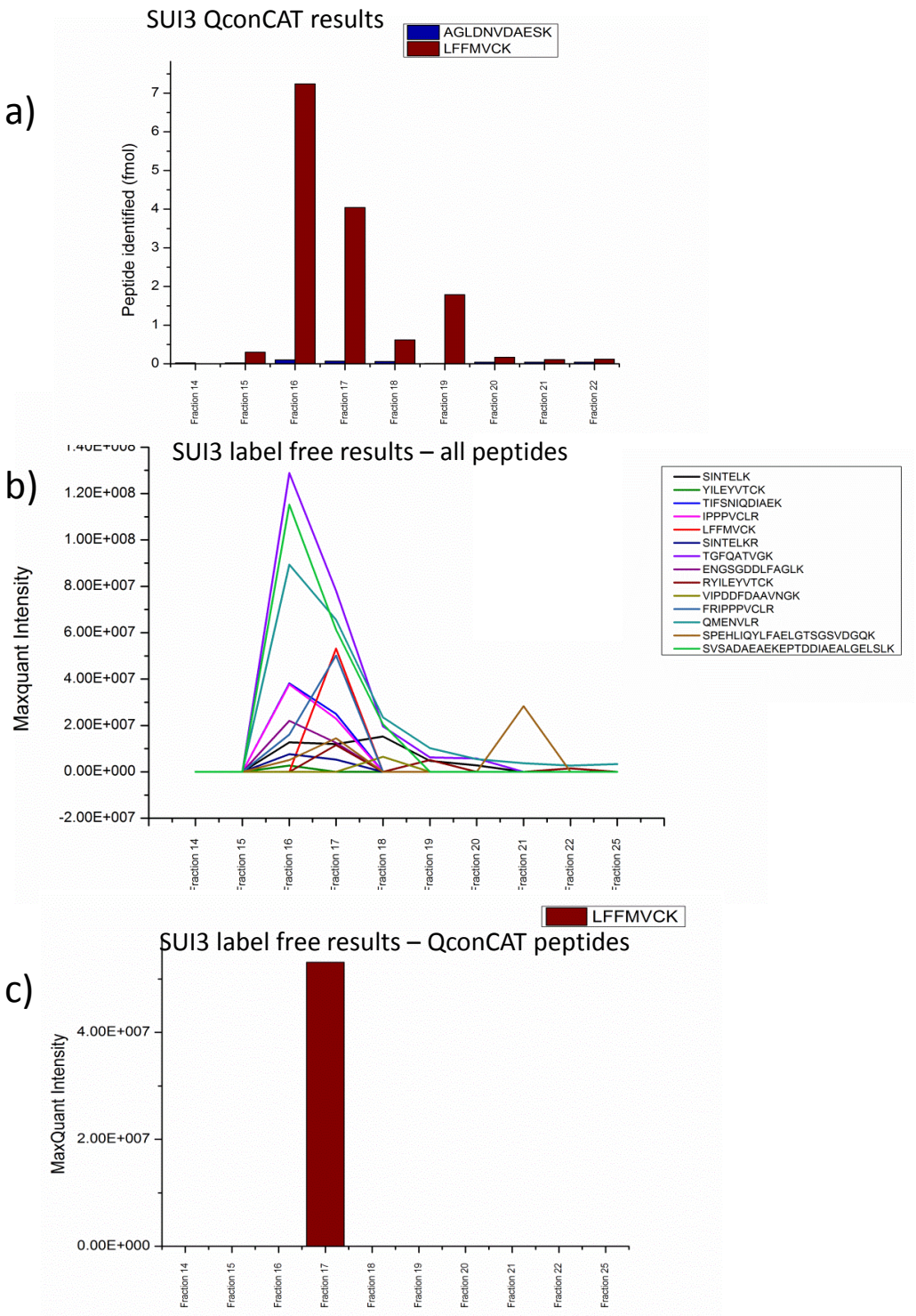


Figure 4.17 SUI3 peptide results by both SRM and LC-MS/MS

The two peptides quantified by QconCAT methodology are shown in a), while b) shows the intensity of all of the SUI3 peptides identified by MaxQuant processing. c) the peptide LFFMVCK was also identified by MaxQuant software in the Qexactive data.

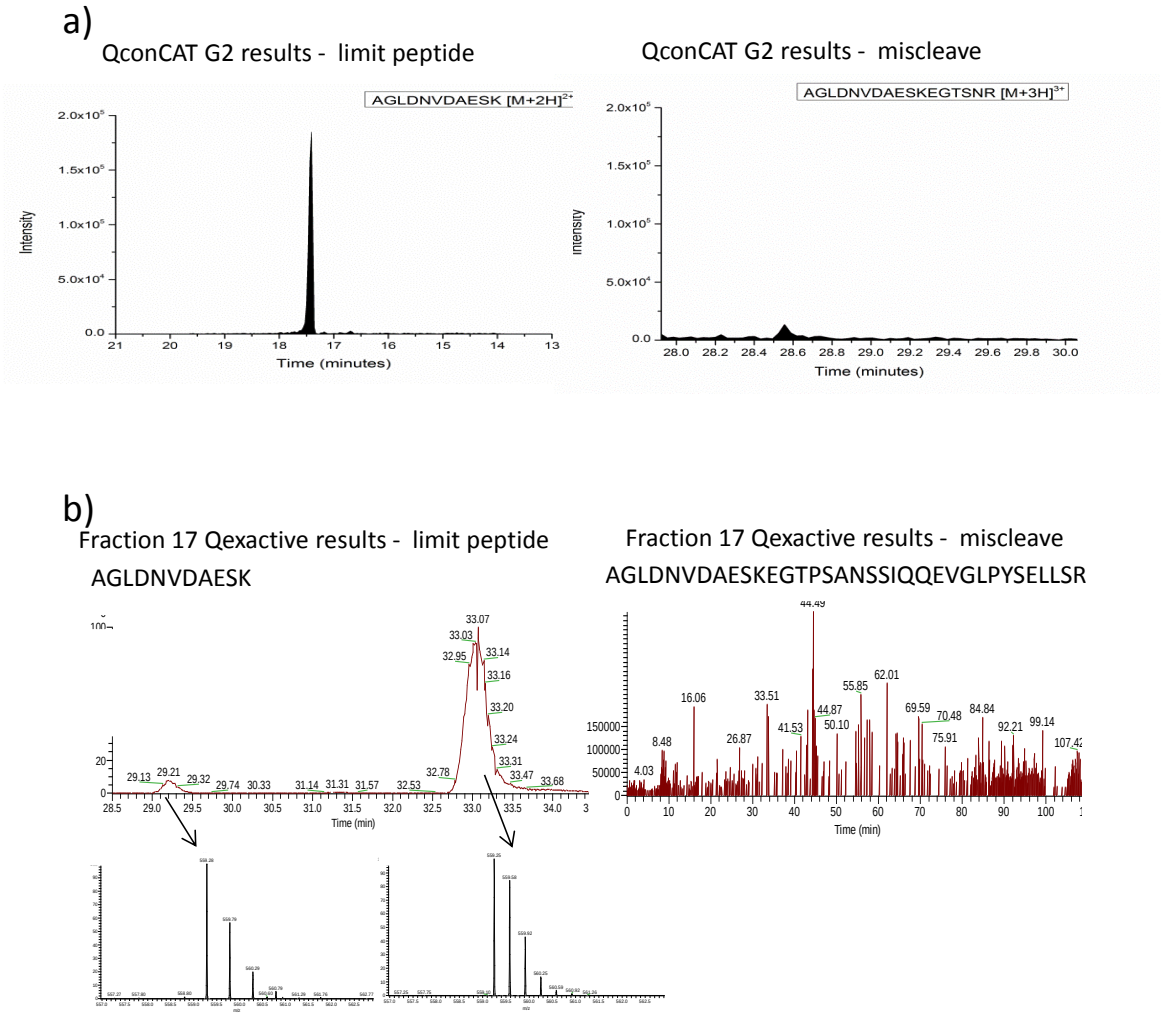


Figure 4.18 AGLDNVDAESK standard and analyte extracted ion chromatograms

a) QconCAT digest was run on the G2 and PLGS software used to generate extracted ion chromatograms for the limit peptide and the miscleaved form. b) Fraction 17 digest was run on the Qexactive and extracted ion chromatograms for the limit peptide and miscleaved peptide were performed in Xcalibur software.

the third Qpeptide included in this QconCAT, meaning the miscleaved analyte peptide would be AGLDNVDAESKEGTPSANSSIQQEVGLPYSELLSR, but this is not identified in any of the 6 fraction replicates processed by MaxQuant, or by extracted ion chromatogram. EGTPSANSSIQQEVGLPYSELLSR was a type B peptide, so the analyte signal was not observed, while the standard was. This indicates the peptide is observable under the experimental conditions, but the analyte signal is at too low a concentration to be observable. The peptide was identified in all other replicates of label free quantification, at a retention time of 29.5 - 32.2 minutes (see supplementary material), and extracted ion chromatograms on the mass of the limit peptide in replicate 6 at that retention time indicate AGLDNVAESK is not present (figure 4.18 b). There is no peak at the correct mass for the miscleaved peptide (figure 4.18 b). However, the resulting miscleaved peptide is large, at 3679.7 Da, and large, miscleaved peptides show less signal in MS, and therefore it may not be observable (Brownridge & Beynon, 2011). In addition, there is no observed peak at the correct mass for EGTPSANSSIQQEVGLPYSELLSR. If there is a miscleavage it may render this peptide at too low an intensity for observation, and AGLDNVAESK at a false low intensity. The two peptides would not be expected to be present at the same intensity, due to differential ionisation, and the lower intensity is also supported by comparing the standard peptides, where EGTPSANSSIQQEVGLPYSELLSR was observed at a much lower intensity (figure 4.13). The low intensity of both of these peptides in the QconCAT data, and the fact that neither was identified by MaxQuant in the Q Exactive data, support the possibility that the peptides are miscleaved. The peptide LFFMVCK is identified in the label free experiment, and is the fifth most intense peptide (figure 4.17 b). In the label-free data 14 peptides are identified, and they are most intense in fractions 16 and 17. Extracted ion chromatograms support the presence of the peptide in fraction 17 of the label free quantification (figure 4.19). This supports the QconCAT mediated quantification of LFFMVCK, at 7 fmol in fraction 16 and 4 fmol in fraction 17. Therefore, given the discrepancy between the two peptides, the LFFMVCK quantification will be used. This indicates that SUI3 is present in fractions 16-17.

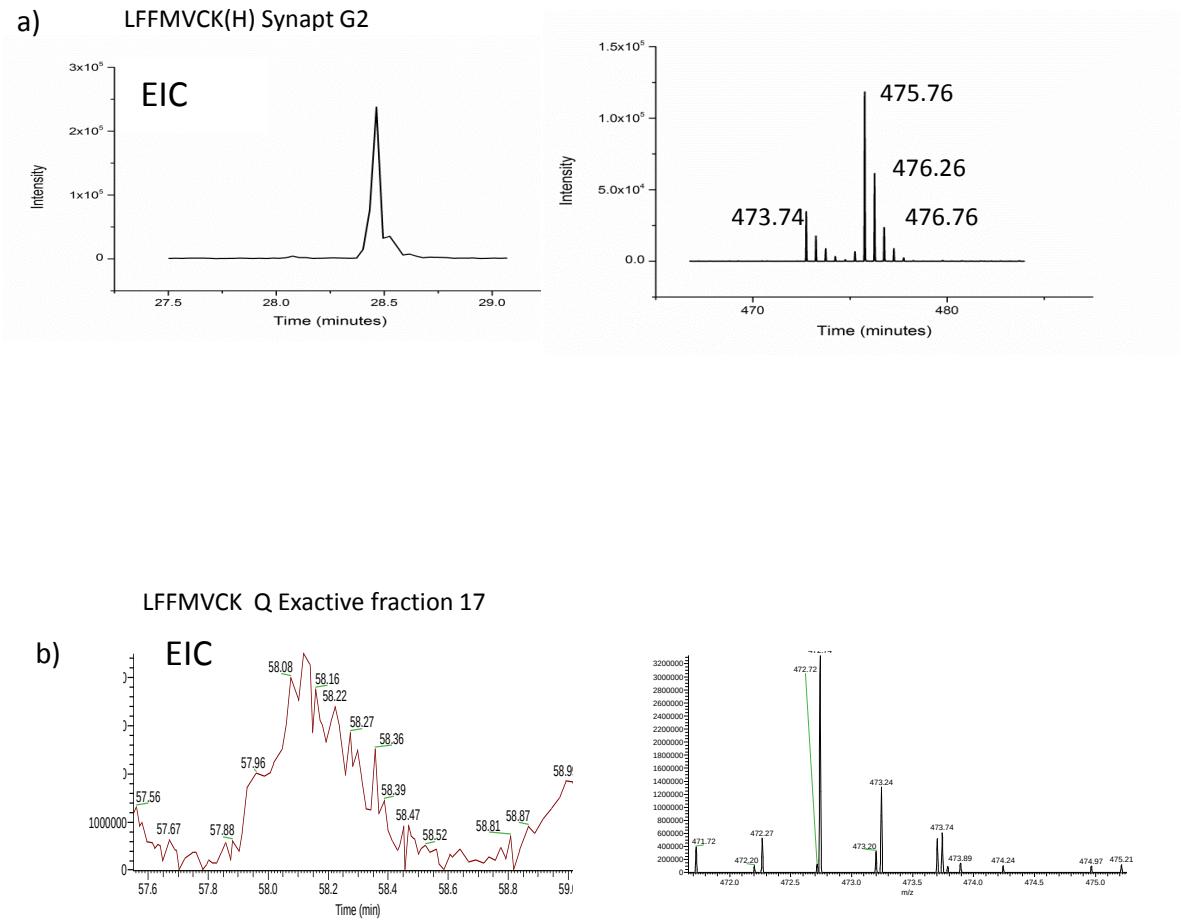


Figure 4.19 SUI3 peptide LFFMVCK extracted ion chromatograms

QconCAT digest was run on the Waters Synapt G2 on a 1 hour gradient in a DDA method. a) An extracted ion chromatogram for the peptide LFFMVCK for the unoxidised m/z of 475.7. b) Fraction 17 digest was run on the Q Exactive on a 2 hour gradient, and an extracted ion chromatogram shown for m/z 472.7.

For the subunit GCD11 the two peptides quantified were LNPLSAEIINR and YNIDAVNEFIVK. Neither were seen as miscleavages or with any modifications in the G2 data. The two peptides show a good correlation (figure 4.20). In the label free data, there are 27 identified peptides, and they are most intense in fraction 19. There is also a peak in the peptide intensity at fractions 15 – 16. The two peptides used in the QconCAT also follow this pattern, with peaks in fractions 15-16 and fraction 19. This supports the idea that this subunit is eluting in two locations.

When the peptide results are averaged to yield the protein quantification, in the individual fractions SUI3 appears twice as concentrated in the label free results, and GCD11 is also found at much higher levels, 1.6 pmol rather than 0.7. When the total amount of each subunit is calculated, the SRM data indicates they are more equal than is shown by the label free data (figure 4.21 a). As in the label free data, all three subunits are identified in fraction 19, SUI2 and GCD11 at similar amounts, and SUI3 less concentrated (figure 4.21 b). However, SUI3 is present at higher concentration in fraction 16 in both data sets. It is possible that a large proportion of the SUI3 subunit has dissociated from the complex during fractionation, leaving a complex of SUI2 and GCD11, and a small number of SUI3 subunits, undetected by the label free software. It is possible the SUI3 subunits are dissociating due to the fractionation process, however the same effect is seen the work of Kito *et al* (2007), in a TAP tag based project. In this project it was noted that SUI3 (EIF2) is present at a lower concentration than the other two proteins when the EIF2 complex is associated with the EIF2B complex. If this is the case the peak seen in fraction 16 may just be the single SUI3 protein dissociated from the EIF2-EIF2B complex. There are only two subunits of the EIF2B complex identified at low amounts in the label free data, but they are identified in fractions 18-19 (figure 21 d). However, as they are present at much lower amounts than the EIF2 subunits, this can not confirm the theory presented by Kito *et al.*

4.3.5 20S proteasome complex quantification

Due to the absence of a trypsin cleavage site following the glufibrinopeptide standard sequence, the QconCAT YEW3 was quantified externally, using a common

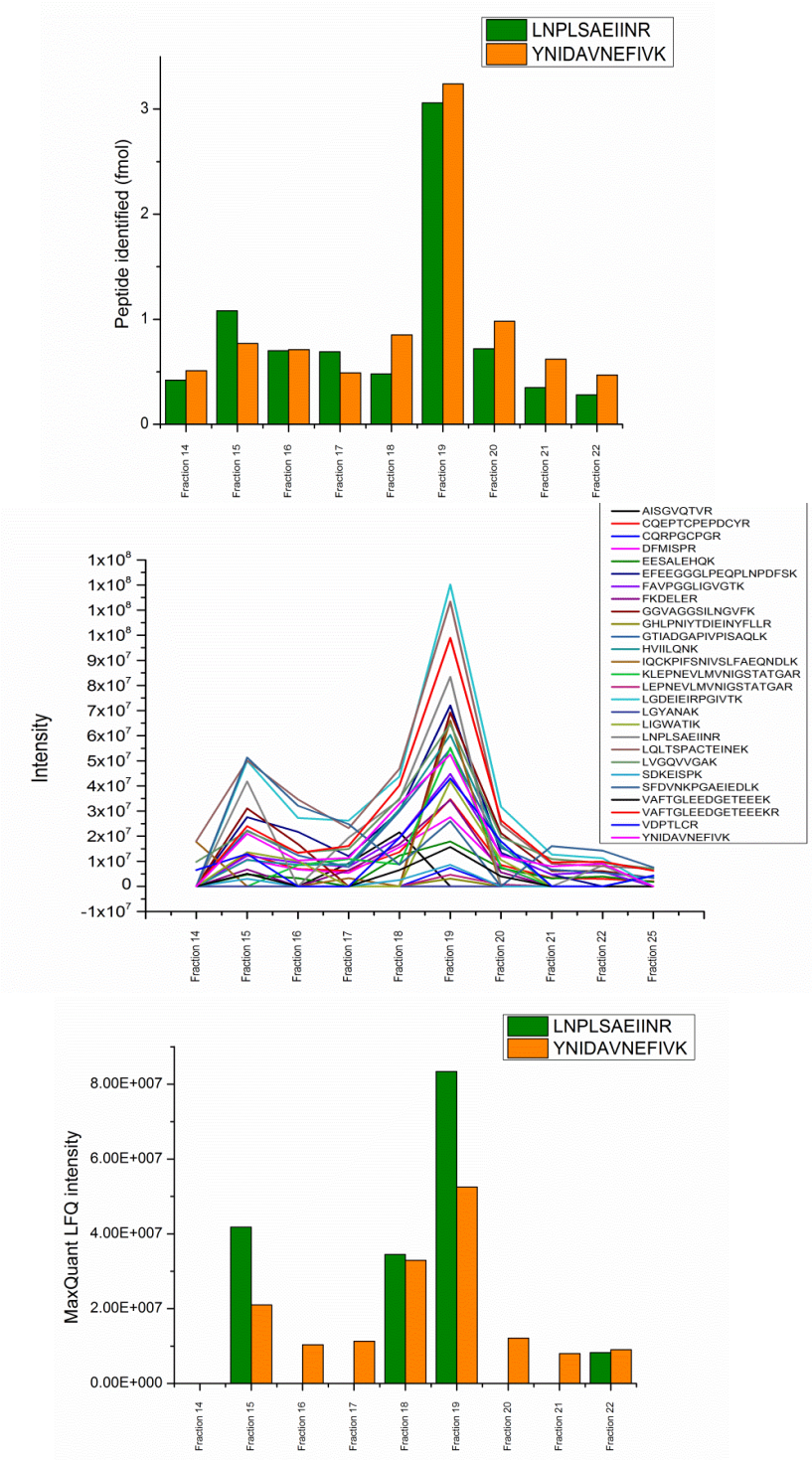


Figure 4.20 GCD11 peptide quantification by SRM and LC-MS/MS methods

The two peptides quantified by QconCAT methodology are shown in a), while b) shows the intensity of all of the EIF2α peptides identified by label free analysis. c) Both of the QconCAT peptides are also identified by MaxQuant processing of Qexactive data.

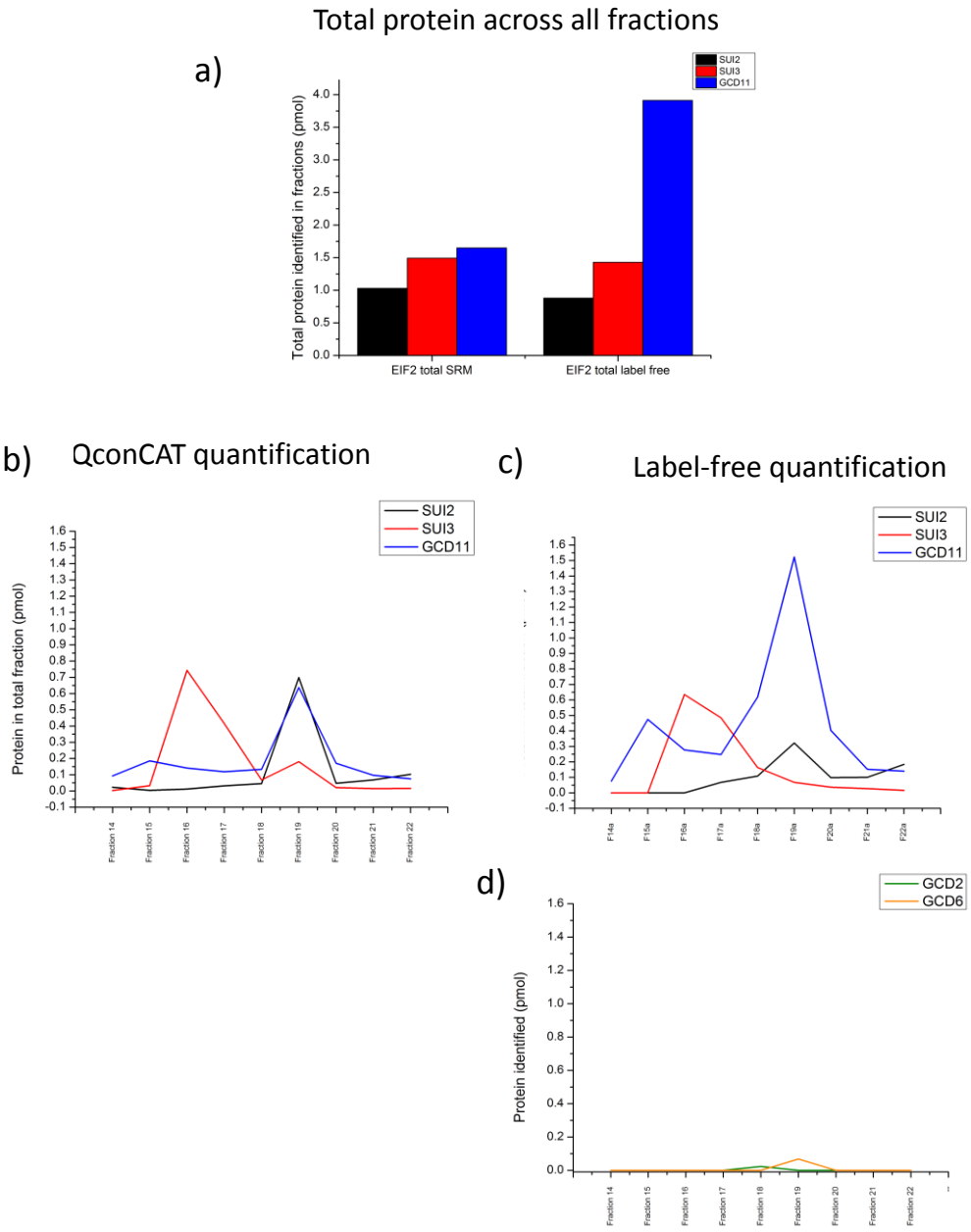


Figure 4.21 EIF2 complex subunit concentrations by label-free and QconCAT quantification

The amount of each protein calculated in each 1 µl injection was multiplied by the total digest volume of 200 µl. The same calculation was performed for samples run on the QExactive. a) the total amount of each subunit in all fractions, and b) the subunit amount in each fraction. c) the amount of the two EIF2B subunits identified by label-free quantification. d) EIF2B complex subunits

peptide in the CopyCAT CC036, obtained from Dr Stephen Holman. The QconCATs were digested simultaneously, and run on the Waters Synapt G2. CC36 was quantified by spiking into the digest a known quantity of unlabelled glufibrinopeptide standard, and comparing the intensity of the heavy to light peptides. YEW3 was then quantified by comparing the intensity of the common peptide NFSLAIDK in YEW3 against CC036. The calculated QconCAT concentration was 0.59 pmol/µl. This method of quantification will be potentially less accurate than an internal quantification method, due to variation introduced between injections. It should therefore be taken into account, when interpreting the data, that the results may contain a higher than usual potential for error. Once quantified, the QconCAT digest was run on the Waters Xevo to check the retention time of each peptide, the results were processed using Skyline software and three transitions were chosen for each peptide. The YEW3 digest was then diluted to 10 fmol, 1 fmol and 0.25 fmol per µl using the digest of *S. cerevisiae* IEX fraction 22 (replicate 6), which was confirmed as containing the 20S proteasome by label free quantification using both the Synapt G2 and the Thermo QExactive.

Due to the large amount of unlabelled QconCAT observed in the Synapt G2 data, a control sample was also run at the same time as the samples. The control consisted of the YEW3 QconCAT, diluted to the same concentrations using a solution of 97% acetonitrile, 0.1% TFA. The dilutions were run on the Xevo using the same methods set up to target the heavy and light peaks in the sample dilutions, and the peak area of the light peptide in the control was subtracted from the peak area of the sample peptides before the quantification was calculated. This method was designed to provide a more accurate quantification result, however, as discussed, matrix effects may alter the ionisation of QconCAT peptides in the presence of sample.

The SRM data was assessed for the presence of peptide pairs, and while the majority of peptides observed were Type A, some peptide quantifications were discarded from the analysis (figures 4.22-4.25). The peptide IFHYGHVFLGITGLATDVTTLNEMFR was a type C peptide, neither the heavy or light peak was observed.

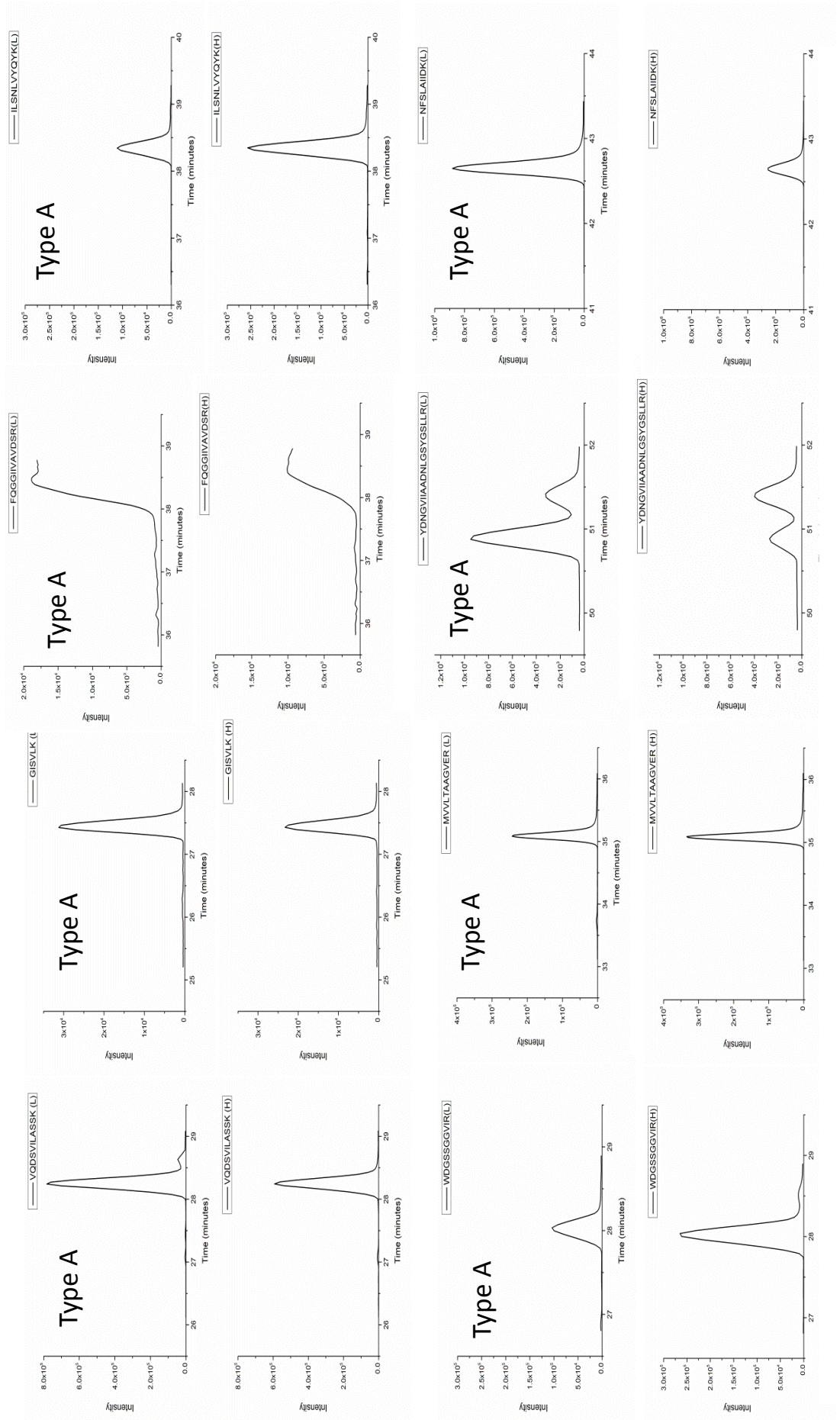


Figure 4.22 Heavy and light 20S proteasome peptide pairs

Heavy and light peptides observed on the Xevo. Examples are shown of each peptide, in the standard(H) and analyte(L). Each pair is labelled with the peptide type.

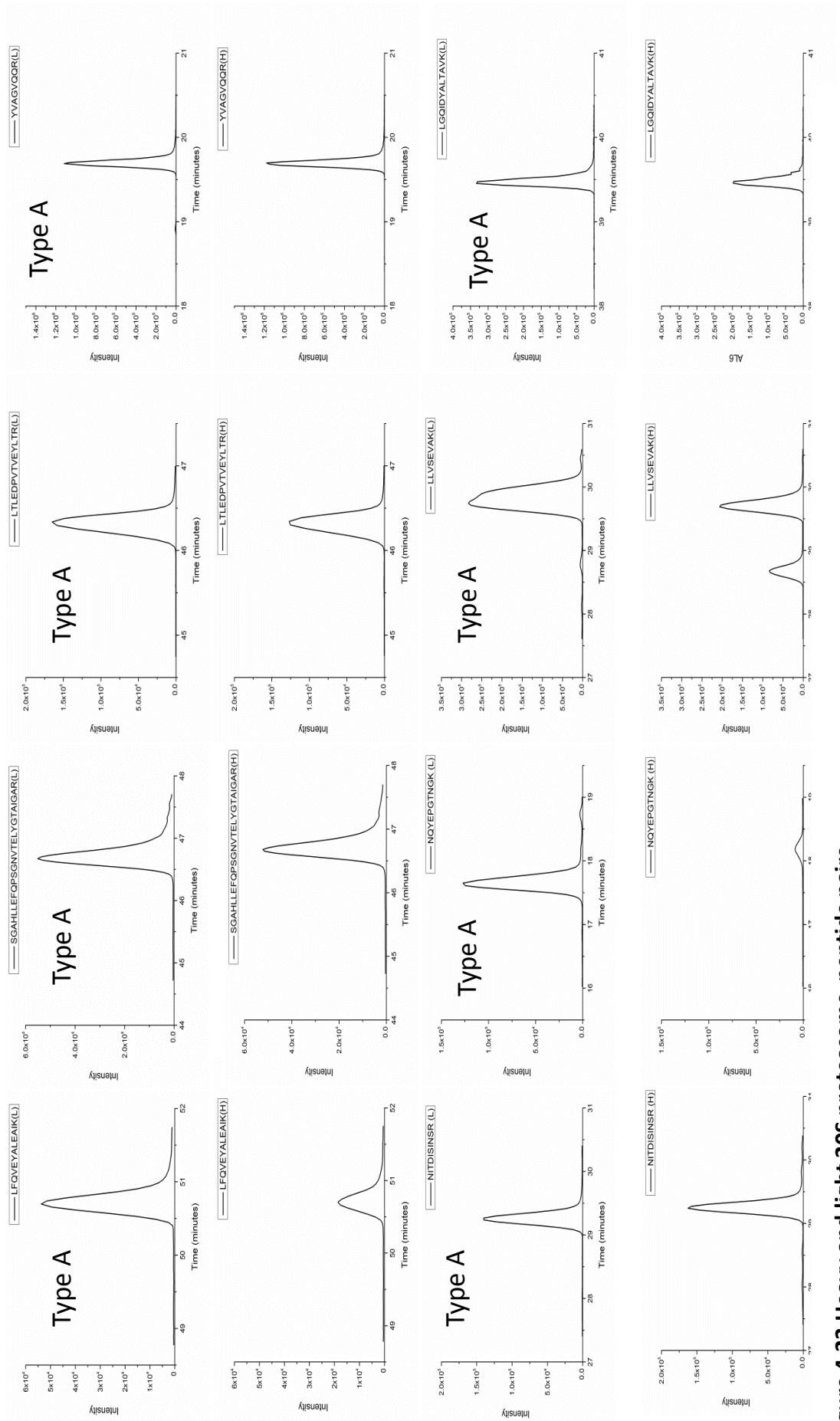


Figure 4.23 Heavy and light 20S proteasome peptide pairs

Heavy and light peptides observed on the Xevo. Examples are shown of each peptide, in the standard(H) and analyte(L). Each pair is labelled with the peptide type.

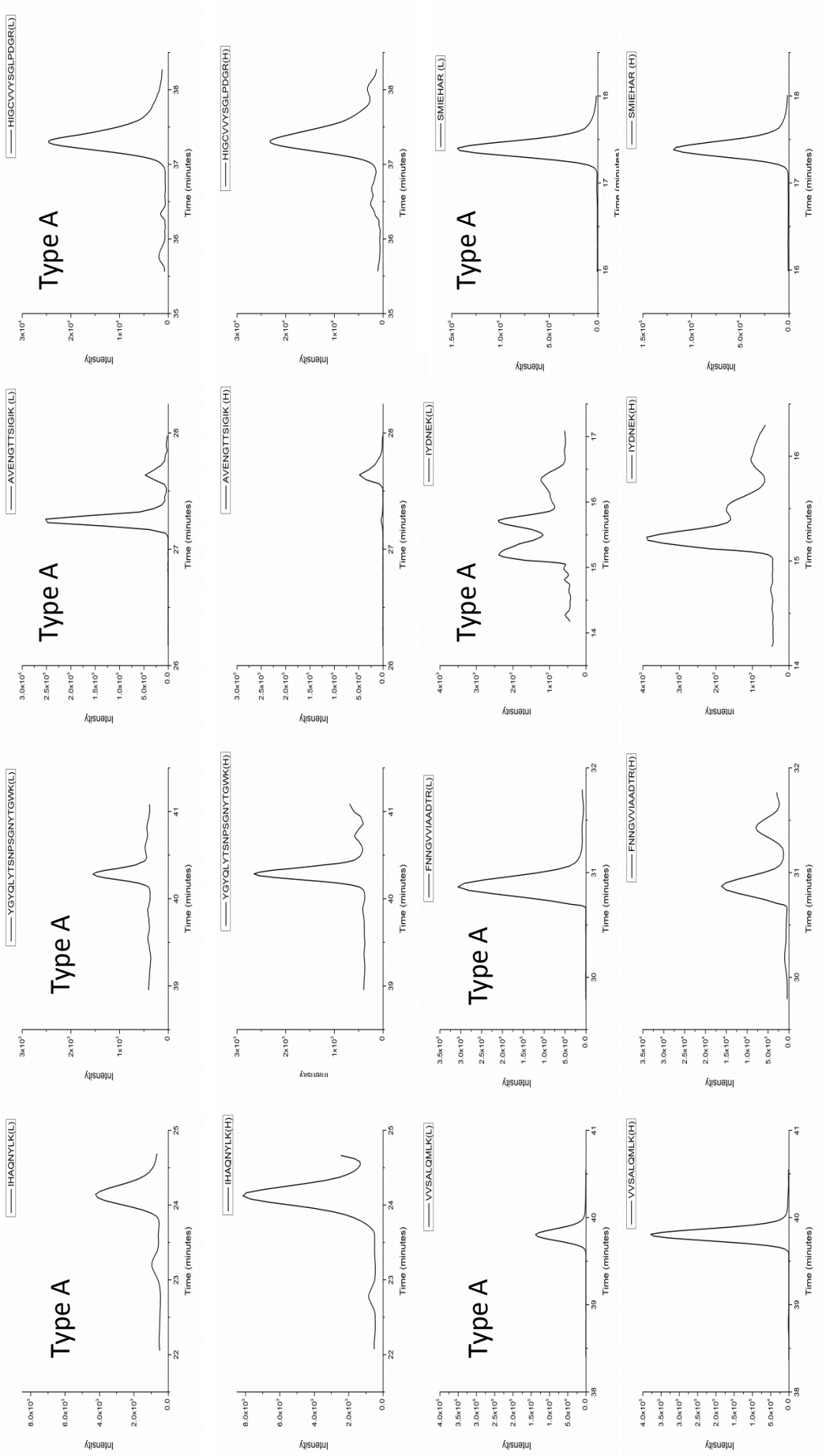


Figure 4.24 Heavy and light 20S proteasome peptide pairs

Heavy and light peptides observed on the Xevo. Examples are shown of each peptide, in the standard(H) and analyte(L). Each pair is labelled with the peptide type.

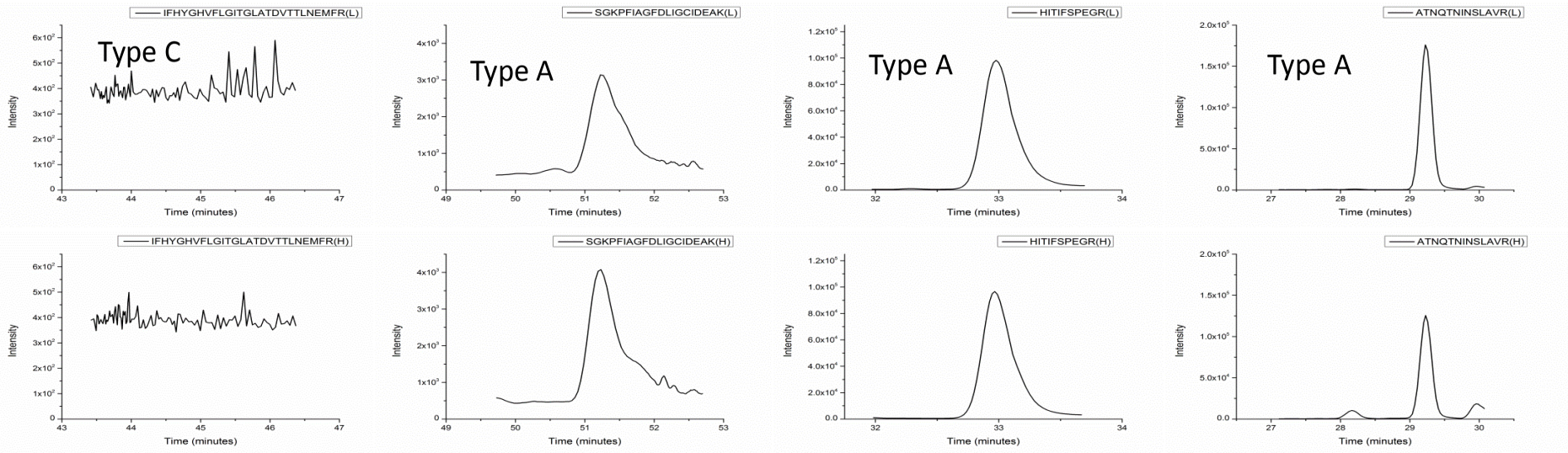


Figure 4.25 Heavy and light 20S proteasome peptide pairs

Heavy and light peptides observed on the Xevo. Examples are shown of each peptide, in the standard(H) and analyte(L). Each pair is labelled with the peptide type.

Other peptides had potential modification sites. Four peptides in this QconCAT contain the sequence NG, an asparagine residue followed by glycine which can allow a deamidation reaction to occur. In the deamidation reaction, the asparagine side chain attacks the neighbouring amino acid group forming a succinimide intermediate. The intermediate can then be hydrolysed to form either an aspartic acid residue, or isoaspartic acid. Therefore, the observed change is an addition of 0.984 Da to the mass of the peptide.

When checking the peptide labelling of YDNGVIIAADNLGSYGSLLR using MALDI-TOF analysis, it was apparent that the peptide was present at 1 Da heavier than the expected mass (figure 4.26 a). There was no evidence in the spectrum of either the unlabeled peptide or the succinimide intermediate form. The LC-MS data indicates the presence of both the original peptide, and the deamidated form (figure 4.26 b and c). SRM data shows two peaks in both the heavy and light transitions (figure 4.26 d). However, since only the unmodified form of the peptide was targeted it was not possible to quantify the peptide.

The MALDI data indicates AVENGTTSIGIK is also deamidated, with a peak at + 1 Da, and no peak at the unmodified *m/z* or the succinimide intermediate (figure 4.27 a). The LC-MS data for both heavy and light versions of AVENGTTSIGIK indicate both forms are present (figure 4.27 b and c). When the original mass is targeted by SRM the chromatogram shows two peaks, however for the light peptide, only the first peak is observed in all three transitions. This indicates the first peak is the correct peptide. In this case, the intensity of the heavy peptide is well below that of the light. However, as the deamidated mass was not targeted for SRM analysis, it is not possible to accurately quantify the peptide (figure 4.27 d).

MALDI data also indicated FNNGVVIAADTR(H) may be deamidated, by the presence of a peak at 1 Da higher than the expected mass (figure 4.28 a). There is no indication of a succinimide intermediate in the mass spectrum, although there is a peptide peak approximately 1 Da more than the expected succinimide intermediate (1266.6), the peak corresponds to another peptide (T37), and the isotope pattern shows no evidence of an overlapping peak. In addition, the G2 data acquired shows

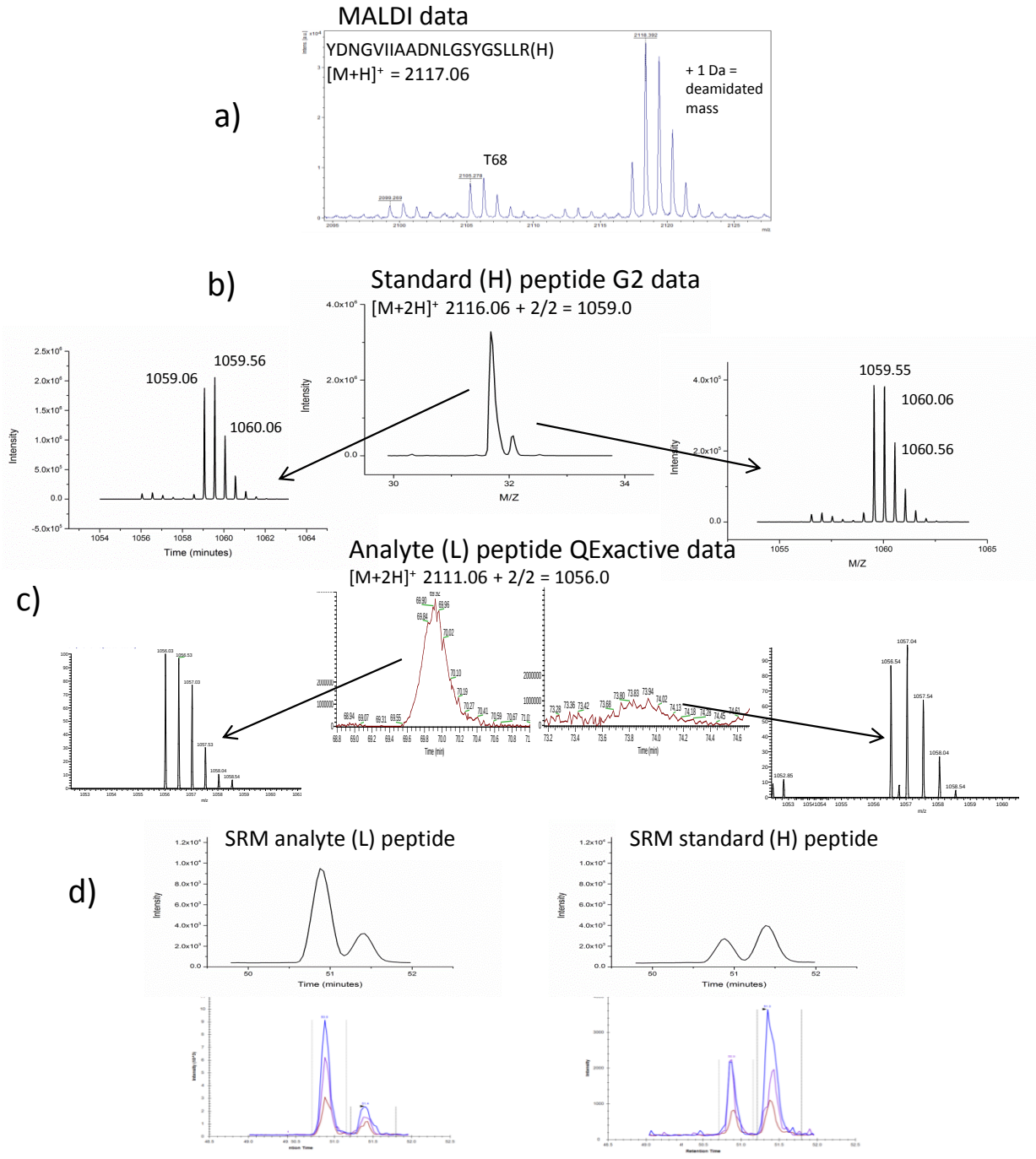


Figure 4.26 Possible deamidation of peptide YDNGVIIAADNLGSYGSLLR

Purified QconCAT protein was subjected to in solution digestion. The digest was then analysed by a) MALDI TOF using a Bruker Ultraflextreme, where a peptide was identified at approximately 1 Da heavier than the expected mass. b) QconCAT digest was also analysed by LC-MS on a Synapt G2, and c) the standard and analyte peptides observed in an SRM experiment using a Xevo TQ, and d) the SRM peptide data observed in Masslynx and Skyline

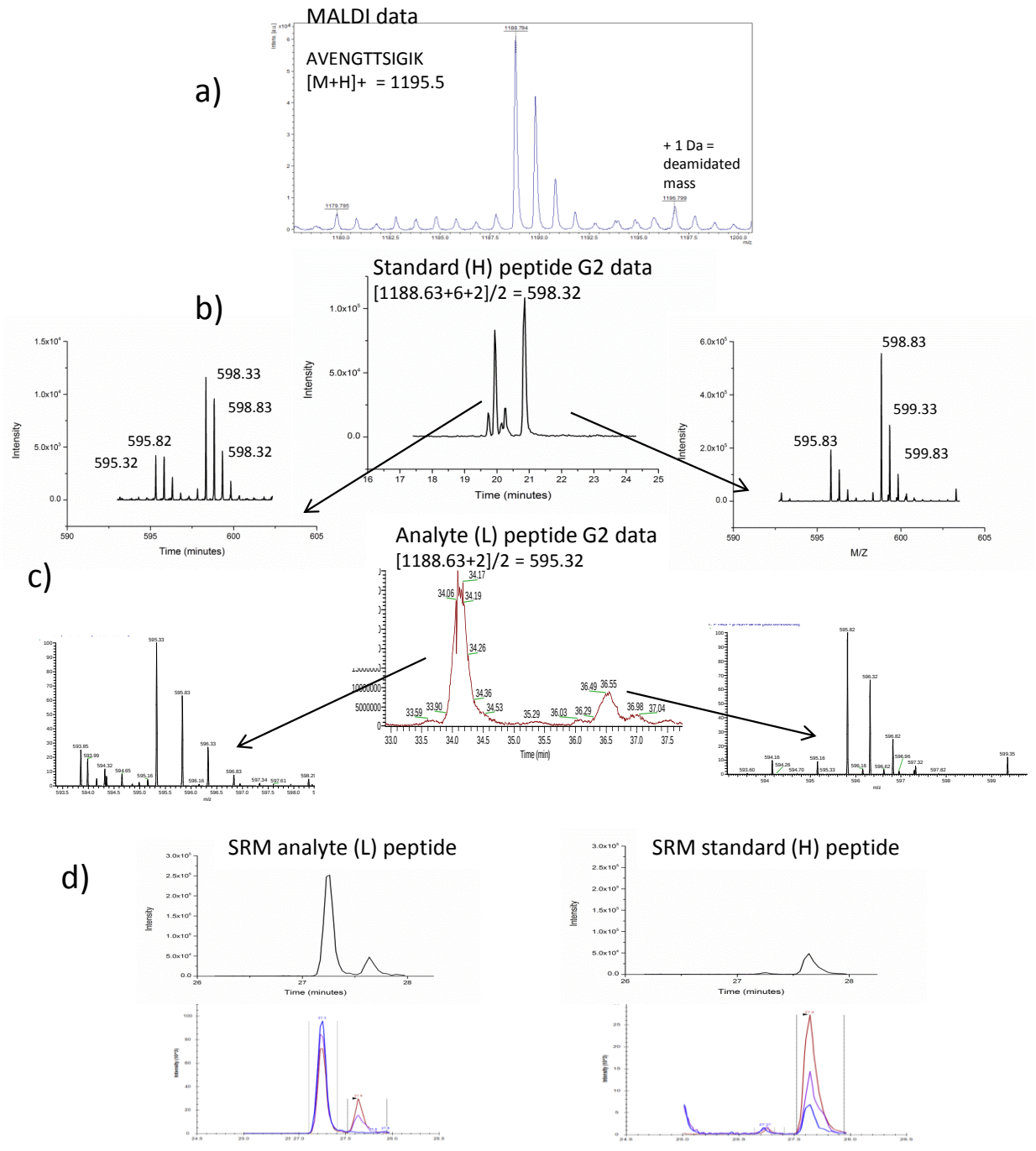


Figure 4.27 Possible deamidation of peptide AVENGTTSIGIK

Purified QconCAT protein was subjected to in solution digestion. The digest was then analysed by a) MALDI TOF using a Bruker Ultraflextreme, where a peptide was identified at approximately 1 Da heavier than the expected mass. b) QconCAT digest was also analysed by LC-MS on a Synapt G2, and c) the standard and analyte peptides observed in an SRM experiment using a Xevo TQ. d) the SRM peptide data observed in Masslynx and Skyline

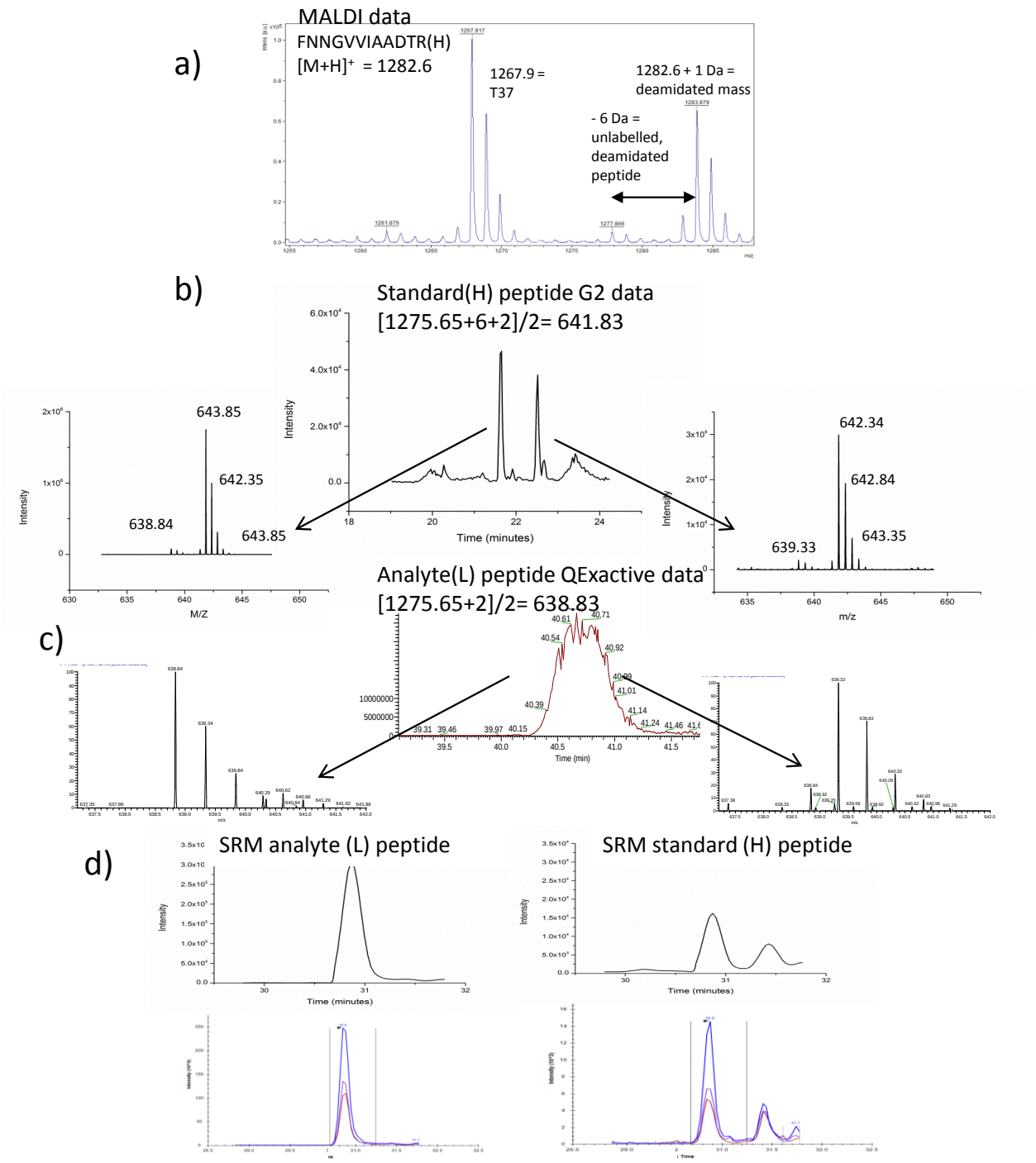


Figure 4.28 Possible deamidation of peptide FNNGVVIAADTR

Purified QconCAT protein was subjected to in solution digestion. The digest was then analysed by a) MALDI TOF using a Bruker Ultraflextreme, where a peptide was identified at approximately 1 Da heavier than the expected mass. b) QconCAT digest was also analysed by LC-MS on a Synapt G2, and c) the standard and analyte peptides observed in an SRM experiment using a Xevo TQ. d) the SRM peptide data observed in Masslynx and Skyline

two peaks in the extracted ion chromatogram, the spectra of which are 0.5 Da apart, corresponding to the 1 Da increase in mass (figure 4.28 b). The analyte peptide on the QExactive also shows evidence of modification (figure 4.28 c). Unfortunately the transition lists on the Xevo were set up to target only the unmodified form of the peptide, not the deamidated version. The SRM for the heavy peptide has two peaks eluting approximately 0.5 minutes apart, however the light peptide only has one peak, indicating that the first peak eluting is the correct peptide. In addition, the heavy and the light peptide have the same fragmentation pattern (observed in Skyline), further supporting the fact that the first peak is the correct one (figure 4.28 d). The heavy labelled peptide, however, is at a lower intensity than the light. Because the deamidated mass has not been acquired it is not possible to discern the degree of deamidation in this peptide. Had the transition list included the modified form of the peptide, it may have been possible to calculate the ratio of the two forms of the peptide, and quantify the peptide (Rivers *et al.*, 2008).

The fourth potentially deamidated peptide, NQYEPGTNGK, was also observed at + 1 Da in the MALDI and LC-MS data in both heavy and light forms (figure 4.29 a- c). The SRM data has single peaks, however, the light peak has eluted approximately 0.5 minutes earlier than the heavy (figure 4.29 d).

If the deamidated and intermediate forms had been targeted, it may have been possible to isolate the peptides in all four isoforms, the original Asn, isoAsp, Asp, and the succinimide intermediate form (Chelius *et al.*, 2005). If this had been achieved quantification may have been possible by comparing the ratio but with the limited data acquired, quantification will not be possible in this instance. Therefore, for quantification of the 20S proteasome using the data gathered here, these four peptides won't be used, and for the four proteins the other Qpeptide will be used in the protein quantification. Another problem peptide was FQGGIIVAVDSR, for which the retention time put into the method was too early, and only half of the peak was acquired. The heavy and light peaks were both observed at the same retention time, so the quantification was performed using the area of the partial peak acquired.

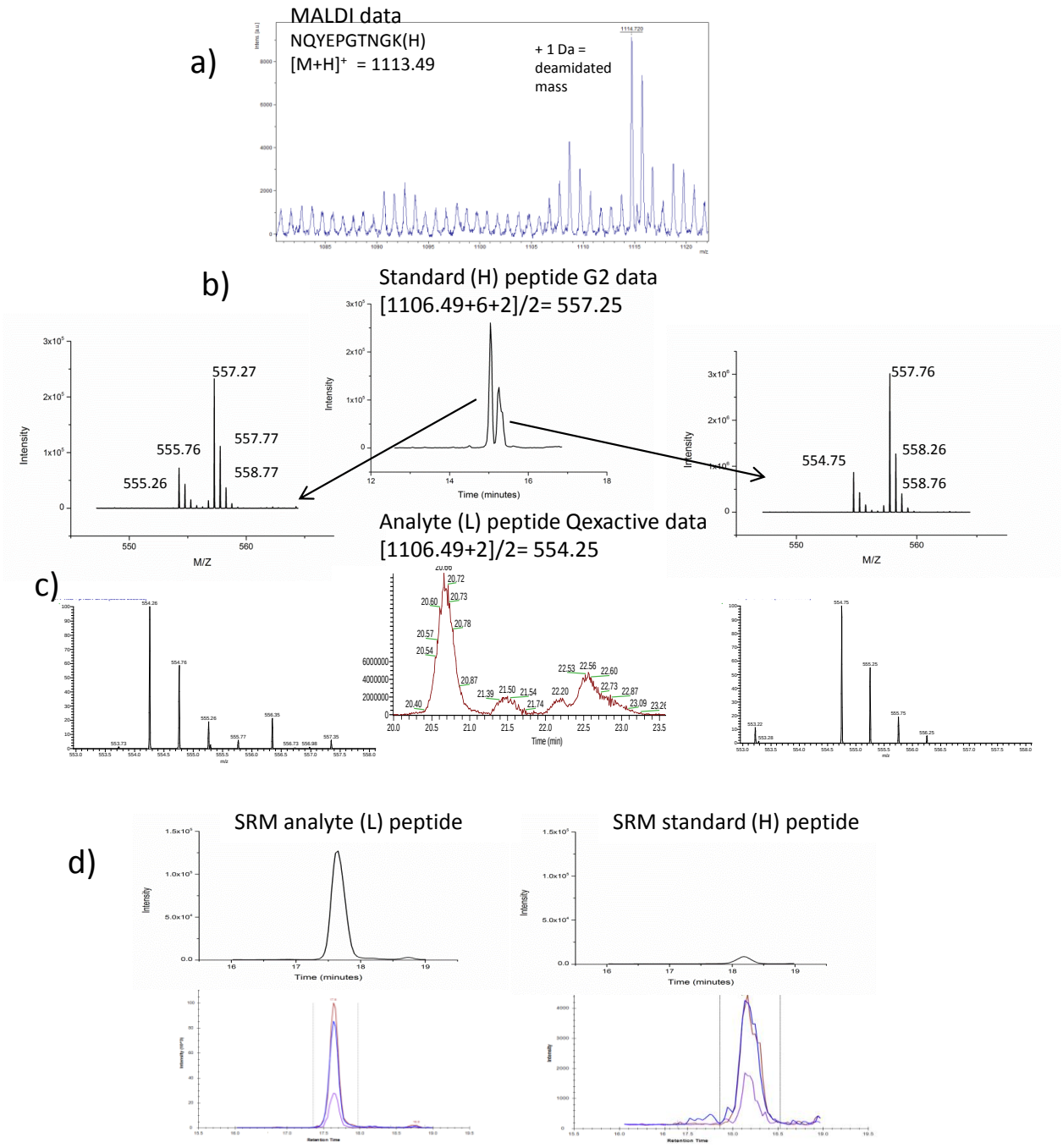


Figure 4.29 Possible deamidation of peptide NQYEPGTNGK

Purified QconCAT protein was subjected to in solution digestion. The digest was then analysed by a) MALDI TOF using a Bruker Ultraflextreme, where a peptide was identified at approximately 1 Da heavier than the expected mass. b) QconCAT digest was also analysed by LC-MS on a Synapt G2, and c) the standard and analyte peptides observed in an SRM experiment using a Xevo TQ. d) the SRM peptide data observed in Masslynx and Skyline

There are also four peptides that contain methionine residues, which can potentially be oxidised to form a peptide of +16 Da in mass. To check if this had occurred, extracted ion chromatograms were performed on both the heavy peptides, and the light peptides (figure 4.30). All of the peptides in this case were observed at the unmodified mass, and no evidence was found of the oxidised forms.

The remaining quantifications were plotted as a histogram and checked for discrepencies between peptides from the same subunit. The two PUP2 peptides, IYDNEK and SMIEHAR, show a large discrepency, at 3.15 and 12.04 fmol. Neither of these peptides have any potential modification listed on the uniprot database. In this case, the three transitions for each peptide were compared (figure 4.31 a). Of the three transitions for the peptide IYDNEK none of the peaks observed co-elute, resulting in a low quantification. On the basis of this the quantification of the second peptide, SMIEHAR, will be used for the PUP2 protein subunit.

There was also a discrepancy in the results for the PRE5 subunit, where the peptides were at 28.43 and 10.5 fmol (figure 4.31 b). Neither of these peptides have any potential modification sites, and the transitions for each show a good correlation. LFQVEYALEAIK was the most concentrated peptide in the analysis, calculated at 28.43 fmol. Coincidentally, the next highest concentration calculated was 28.34, for NFSLAIIDK, the peptide which is N terminal to LFQVEYALEAIK (figure 4.32). These are both at almost three times the average peptide quantification of 10 fmol. A high result in SRM occurs when the analyte signal intensity is higher than the standard. Therefore, if the standard intensity is low, for example if there is a miscleavage at that point in the QconCAT, it will give a false high result for the analyte. As the two high results occur for peptides which are next to each other on the QconCAT, it is possible that a miscleavage has occurred. However, the miscleaved peptide was not identified by PLGS processing. In addition, it was not found in MALDI data or by searching the G2 data for the extracted ion chromatogram (figure 4.33). The lack of evidence does not prove that the miscleavage does not exist, however, as the peptide would be lysine terminated

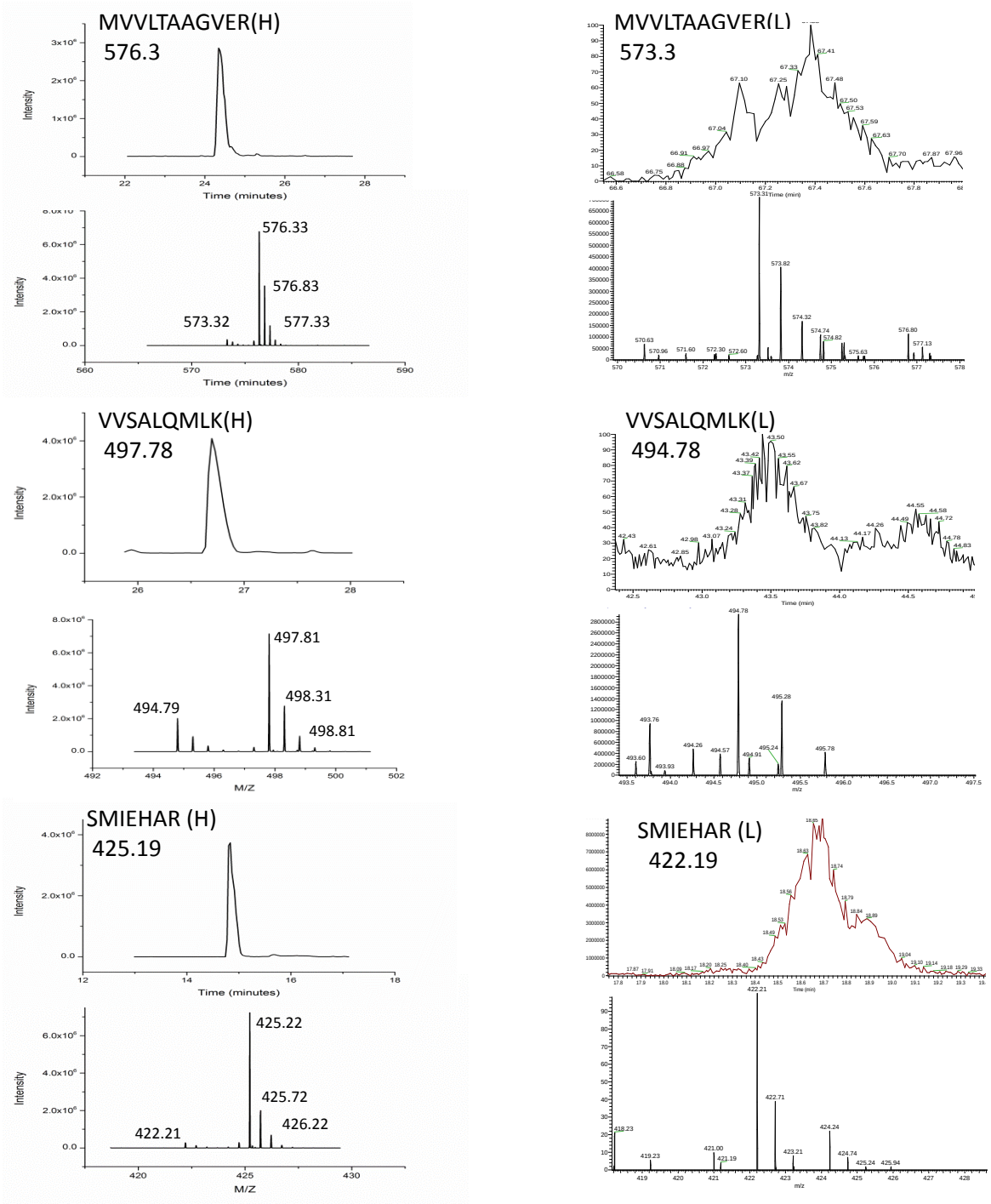


Figure 4.30 Extracted ion chromatograms of methionine containing peptides

Heavy labelled QconCAT peptides were run on the Waters synapt G2, and extracted ion chromatograms are performed using PLGS software. Fraction 17 digest was run on the Qexactive, and extracted ion chromatograms of peptides were performed in Thermo Xcalibur.

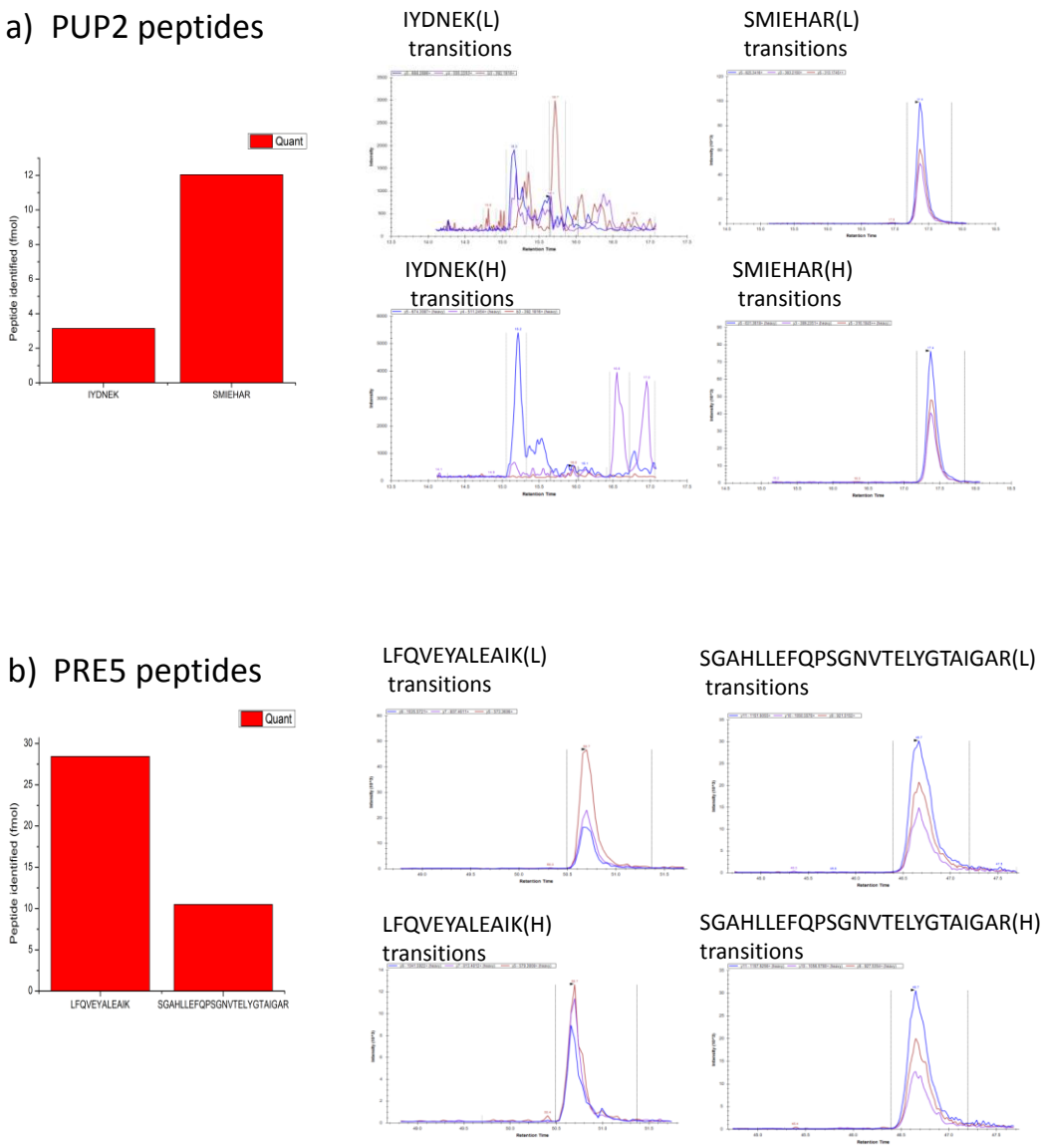


Figure 4.31 The transitions acquired for the two peptides from PUP2 and PRE5 proteins

QconCAT digest was mixed with fraction 22 digest, and SRM data was acquired on a Waters Xevo triple quadrupole mass spectrometer. Three transitions for each peptide are acquired, and displayed in Skyline software.

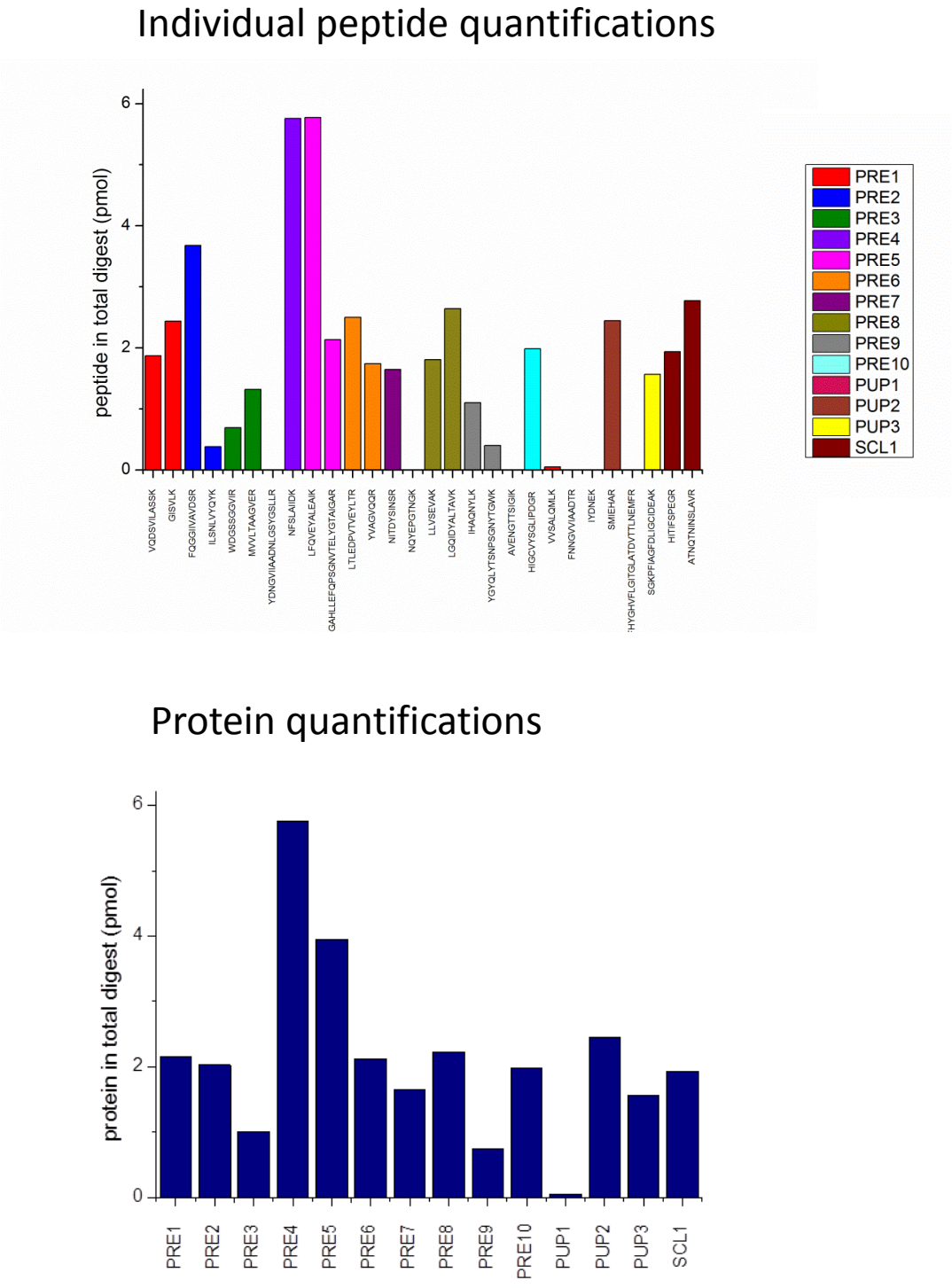
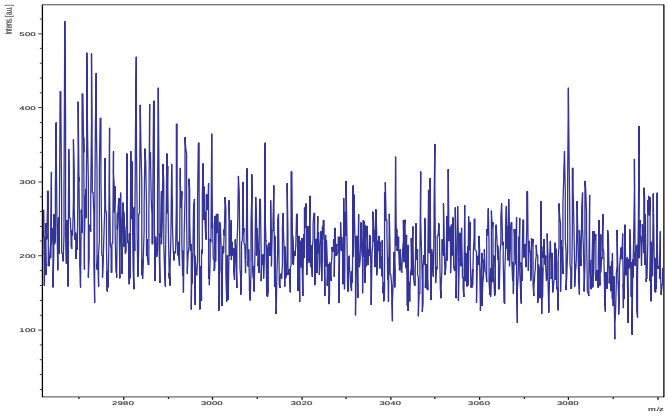


Figure 4.32 20S proteasome individual peptide and average subunit quantifications

Peptide quantities were calculated using MassLynx. The peak area of the heavy and light peptides were calculated, the area of the unlabelled QconCAT was subtracted from the analyte peak, and the ratio of the remaining light area to heavy peak area was used to calculate the peptide quantification, then corrected for volume. For protein quantification the peptide results were averaged.

a)

NFSLAIIDKNTGTGRLFQVEYALEAIK
[(1019.56+604.28+1422.76)+18+1] = 3065.99



b)

NFSLAIIDKNTGTGRLFQVEYALEAIK
[(1019.56+604.28+1422.76)+18+4] / 4 = 767.23

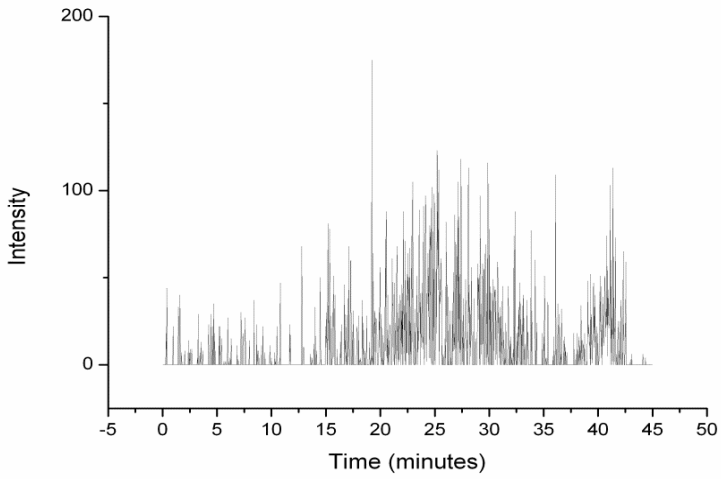


Figure 4.33 Lack of evidence for miscleavage in QconCAT peptides NFSLAIIDK and LFQVEYALEAIK

QconCAT digest was run on both the Bruker Ultraflex and Waters Synapt G2 mass spectrometers to check for miscleaves. No evidence is seen of either a) the singly charged peptide in MALDI, or b) the doubly charged peptide in an EIC of the G2 LC-MS data.

and therefore may not be visible by MALDI, and under LC-MS conditions, it would be large, at 3064.9 Da, and a quadruply charged ion, and therefore the miscleavage could be unstable under experimental conditions.

For the protein PRE2 the two peptides are quantified at 18.1 fmol and 1.9 fmol. FQGGIIVAVDSR eluted across a wide time frame, and therefore only a partial peak was acquired. Although this may affect results, since half of the peak has been integrated for both the standard and the analyte, they have both been treated equally and the result could be a true reflection of the peptide amount, and it will still be included in the analysis.

For the PUP1 peptide, VVSALQMLK, the quantification of the injected peptide is 0.25 fmol. This is much lower than the peptides from other subunits, but the second peptide included from PUP1 was unused due to the presence of an NG sequence. The peptide does contain a methionine, which can potentially undergo oxidation, however the QconCAT wasn't observed in the oxidised form. The analyte peptide was observed in the label free quantification, but not in the oxidised form, and there were no miscleavages identified by MaxQuant processing. Therefore, the result will be included.

The individual peptides quantified, corrected for digestion volume, ranged from 0.05 pmol to 5.78 pmol (figure 4.34 a). The averaged protein quantifications are, consequently, also spread across a wide range, from 0.05 to 5.75 pmol (figure 4.34 b). The average quantification of the 14 20S proteasome subunits is 2.11 pmol, six of the subunit quantifications fall within 10% of this average, while 9 fall within 25%. The two furthest outliers, PUP1 at 0.05 pmol and PRE4 at 5.75 pmol, were both quantified from just one peptide. Therefore, these results could be less reliable than the others.

The label free quantification obtained for the proteasome proteins in replicate 6 ranges from 0.85 pmol to 4.19 pmol. When these are plotted against the QconCAT results they show a somewhat linear relationship (figure 4.34 c). The MaxQuant software used in calculating the label free results uses an IBAQ method, where the sum of the observed intensities for each protein is divided by the theoretically

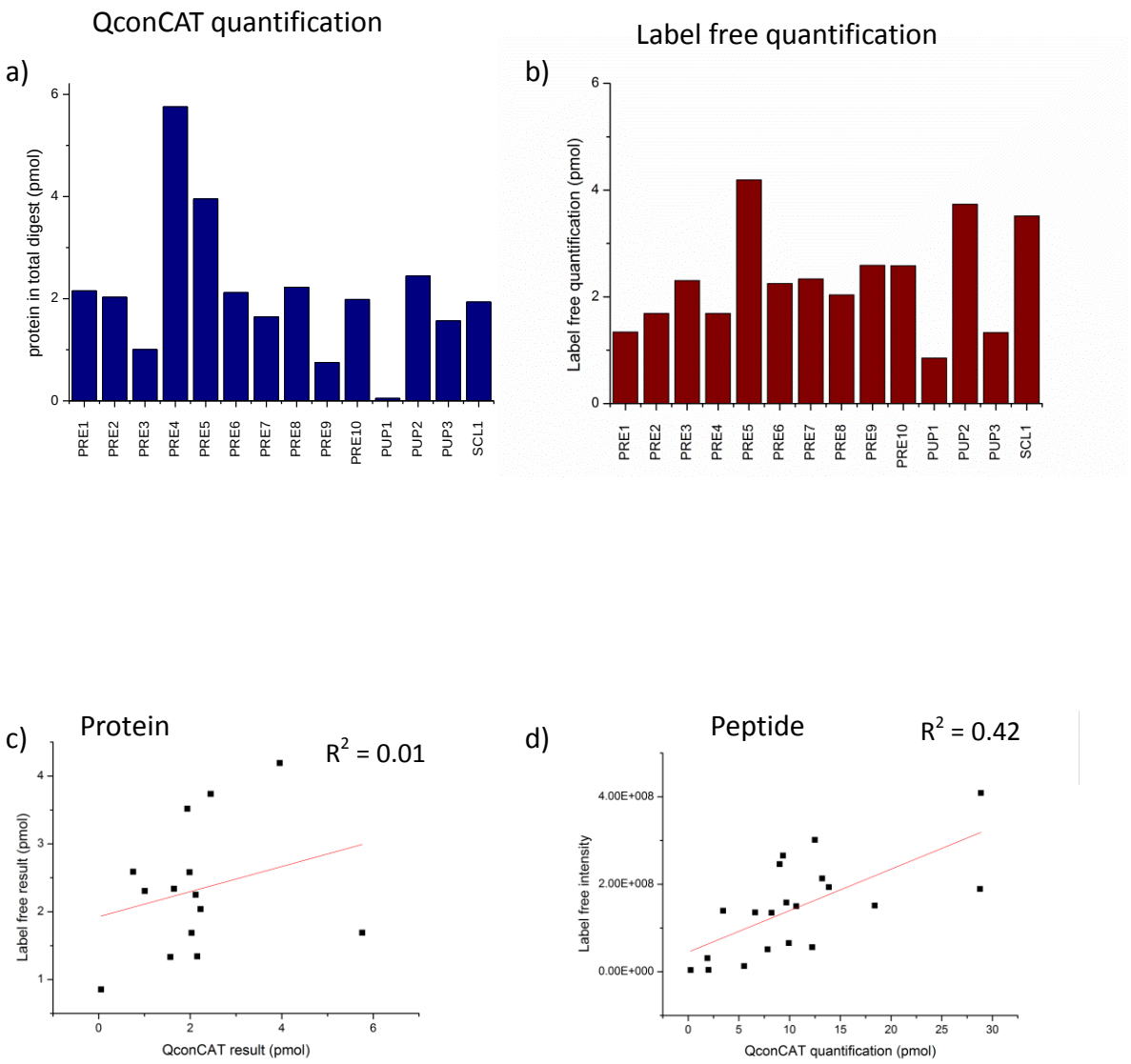


Figure 4.34 Comparison of 20S proteasome quantification by QconCAT and label free analysis

QconCAT quantification results a) calculated by peak area ratios, and corrected for digest volume. The label free results b) calculated using the MaxQuant reported iBAQ intensity and corrected for digest volume. The amount of each protein by each method are plotted against each other in c), while d) shows the iBAQ intensity against the fmol quantification for each QconCAT peptide.

observable peptides. Therefore, all of the peptides identified using the Thermo QExactive will be used in the quantification. When the intensities of just the common peptides are plotted against the QconCAT quantification as a scatter graph, they show more of a linear relationship, as indicated by a higher R squared value (figure 4.34 d). This indicates that the individual peptides are behaving in approximately the same manner across both instruments, and poorly ionising peptides on the QExactive are also poorly ionising on the Xevo.

The QconCAT quantification results for the two complexes discussed here, where two peptides are identified, generally show a somewhat linear relationship (figure 4.35). The two peptides originate from the same protein, and should therefore have a good correlation. This may indicate that in this analysis the peptides chosen are not proteotypic, therefore the QconCAT peptides are not representative of the behaviour of the other peptides from the same protein. This is exemplified by the results for the subunit PUP1, which was quantified at 0.85 pmol in the total fraction digest. This was calculated on the basis of one peptide, VVSALQMLK. However, in the replicate 6 label free analysis, there are 9 PUP1 peptides identified, and 6 of them are more intense than VVSALQMLK. Of the 9 peptides identified by the QExactive the majority were not included in the QconCAT design due to poor MC:Pred or CONSequence scores. In the case of this protein, a lack of suitable peptides may have forced the use of a peptide that is not quantotypic. The quantification of this protein could be checked using the second peptide, FNNGVVIAADTR, if an SRM experiment is set up to target the deamidated forms of the peptide. The SRM quantification of this protein may have been improved by the inclusion of other peptides, such as those found more intense by label free, which may therefore be more proteotypic.

4.4 Conclusions

The results of the EIF2 complex confirm that there is some dissociation happening, and also that the complex appears to not be stoichiometrically equal, although there is an improvement in the results using SRM. In order to confirm wether the

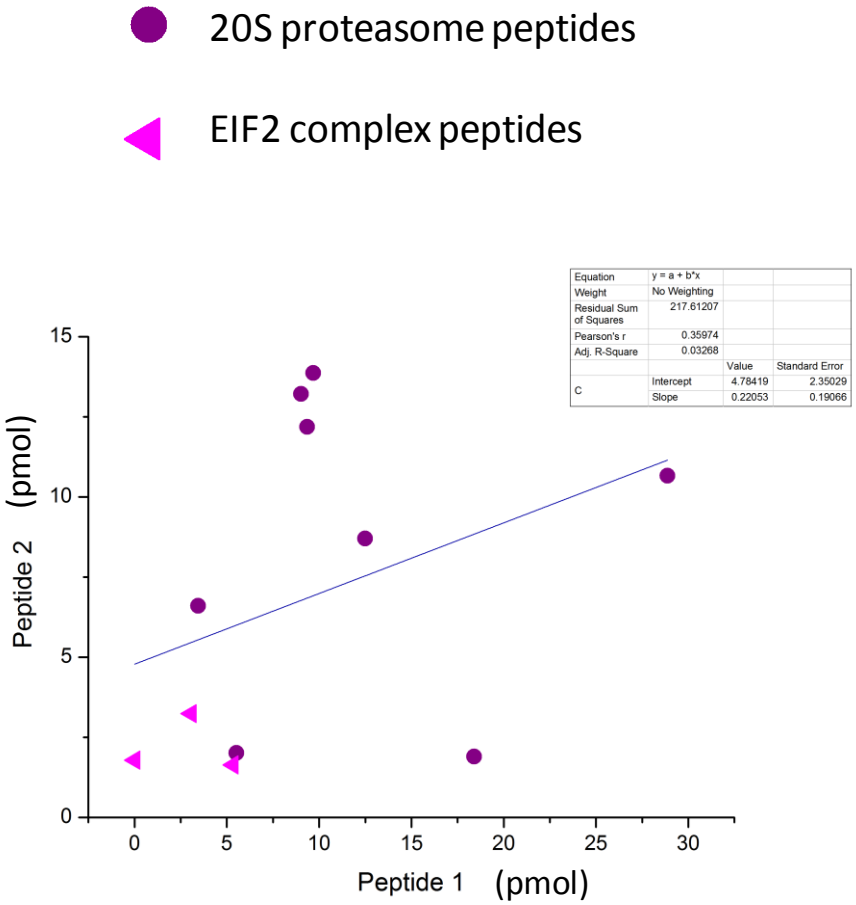


Figure 4.35 QconCAT peptides comparison for those proteins where two peptides are observed

QconCATs encoding 20S proteasome and EIF2 complexes were run on the Thermo QExactive and processed using MaxQuant. For those proteins where two peptide quantifications were obtained, the quantification (in pmol) for each peptide are plotted against each other.

dissociation of the complex is caused by association with EIF2B, the second complex would have to be targeted in an SRM experiment, as it is not present in sufficient amounts in the label free data.

For the two complexes quantified here, it was difficult to find 'quantotypic' peptides, for each complex there were peptides that were not observed in the Xevo data. Due to the variable nature of proteins, varying in size, composition, and physicochemical properties, it can be difficult to find appropriate tryptic peptides for each subunit in a complex. For these proteins, such as the 20S proteasome subunit PUP1, or the elongator protein ELP6, it may be more convenient to quantify using a PSAQ method. This would enable all the potential tryptic peptides to be targeted during one analysis.

The discrepancy in the data for some proteins appears to be due to the poor performance of some peptides, such as the PUP1 peptides discussed, which are observed at a low intensity in both the QconCAT and label free data. However, the concept behind the use of SRM is that the standard and analyte peptides behave in the same manner, so if the analyte peptide is poorly ionised, this will also be represented by the QconCAT intensity, and the quantification should therefore be accurate. However, this does not appear to be the case with the data presented here. Matrix suppression effects may also contribute to any errors in the data. Particularly in the case of the EIF2 complex, where multiple fractions are compared, it is possible that peptides ionise differently in different fractions. If this is the case the quantification of different proteins in different fractions cannot be compared. The effect could be examined using a control sample, such as the glufib sequence included in the QconCAT sequence. If a known amount of light glufib was spiked into each fraction and QconCAT mixture, it may be possible to assess the ionisation effect in each fraction. Also, the calculations for the unlabelled percentage may introduce an additional source of error, as matrix effects will mean the peptides ionise differently without the sample digest background. Any attempt to correct the quantification will therefore add error, however, this is much less than the error that would be included had no correction been applied. For completely accurate results, the heavy labelling of the peptides should ideally be repeated.

If the transitions had been improved, the reliability of the data would have also been improved given enough time and sample to repeat the experiment. Three replicates would give a more reliable quantification, and possibly improve the relationship between protein concentrations.

Error in the QconCAT quantification could also come from the fact that the QconCAT was spiked into the fraction digest following label free quantification, rather than co-digested. Differences in the proteolysis of the standard and analyte peptides would manifest as errors, for miscleavages in the sample, the quantification would be a low result, and for a miscleavage in the analyte a higher result. The digestion efficiency of both the standard and analyte was tested during the experimental workflow using SDS PAGE. However, SDS PAGE is less sensitive than SRM, so some miscleavages may still be present. A better method would have been to split the replicate 6 fractions into two aliquots, one for label free analysis, and the other for QconCAT quantification. This would have made it possible to co-digest the QconCAT and the analyte proteins following the label free analysis identifying the appropriate fractions. The other potential source of error is in the peptides themselves, for example, the four 20S proteasome peptides that contain a potential deamidation site, although, again, this issue could be addressed during the experimental workflow if the modified forms of the peptides are targeted by SRM.

With regards to the aim of exploring the stoichiometry of protein complexes using QconCAT based quantification, the results presented here are inconclusive. Despite the advantages of label mediated quantification and SRM methodologies over label free methods, for the CCT ring and 20S proteasome complexes, there is no improvement to the quantification results. The conclusions, therefore, must be that either both complexes contain differing amounts of each subunit, which is unlikely, or that the quantification is flawed. Since the inaccurate quantification is likely to be due to the use of peptides that are not quantotypic, it must be concluded that a PSAQ method may have been more appropriate.

Chapter 5: Accessing low abundance proteins using ProteoMiner equaliser bead technology

5.1 Introduction

The detection of analyte proteins within proteomics has, historically, been hindered by the large dynamic range of samples, sometimes spanning up to 10 orders of magnitude (Mitchell, 2010), which make the identification and quantification of some of the less abundant proteins difficult. In electrophoresis, the abundant proteins can mask the appearance of the less abundant proteins, while in LC-MS, using a DDA mode analysis for protein discovery, the more concentrated proteins in a sample can dominate the precursor selection, meaning the less abundant samples are not fragmented, and therefore not identified. Problems also arise from these less abundant proteins, sometimes present in extremely low copy numbers, falling below the limit of detection of a mass spectrometer. There are various techniques used to counteract the dynamic range issue, the majority of which are based on reducing the complexity of the sample either by removing the most abundant proteins using depletion strategies or by splitting the entire proteome into subsets by fractionation.

The technique used here, combinatorial peptide ligand library (CPLL) technology, aims to reduce the dynamic range of biological samples without reducing sample complexity by simultaneously reducing the quantity of high abundance proteins and increasing that of less concentrated proteins. This reduction in dynamic range should, in theory, bring all of the proteins in a sample into the range detectable by the mass spectrometer and, by normalising the proteins to approximately the same concentration, should eliminate the problem of a few proteins saturating the mass spectrometer. The technology, first applied in proteomics for the analysis of serum proteins, consists of a library of hexapeptide ligands covalently bound to spherical, porous beads (Thulasiraman *et al.*, 2005). The library is formed from 20 naturally occurring amino acids, using the Merrifield approach split, couple, recombine method, first implemented to make hexapeptide combinatorial beads by Furka et al

in 1991 (Merrifield, 1965; Furka *et al.*, 1991; Lam *et al.*, 1991). A batch of porous poly(hydroxymethylacrylate) beads is split into 20 subsets, and to each of these subsets a different amino acid is bound via a linker, such as a carboxyl group, and then the subsets are pooled, mixed, split into 20 sets again, and the process is repeated. The process is repeated six times, forming a different hexapeptide ligand on each bead, and leading to a potential 20^6 (64 million) different ligands, with multiple copies of a single peptide sequence covalently attached to a single bead at a potential density of 50 µmol/ml (Boschetti & Righetti, 2008). These peptides, presented on the surface of the bead and exposed to a sample, are able to interact with proteins in the same manner as all naturally occurring protein-protein interactions in the cell, such as ion-ion, hydrogen bonds, dipole-dipole, dispersion and hydrophobic interactions (Righetti *et al.*, 2006).

Due to the vast amount of variation in the peptide composition, and the ligand-protein interactions, an aliquot of beads of sufficient volume would potentially be able to interact with all of the proteins in a given proteome. Therefore, the concept behind the reduction in dynamic range using these beads is that, when a sample containing an asymmetric mixture of proteins is introduced to the library, each bead will bind a finite, equal, number of copies of a single protein, up to the potential bead capacity of 50 µmol/ml. Any excess of protein above this amount will not interact with the beads, and will be removed by a wash step (see figure 5.1 for a schematic of experimental workflow). Once excess proteins have been washed off, the bead bound protein can be removed directly using SDS if the analysis is to be done using PAGE. Alternatively, various elution agents can be used, such as urea or CHAPS, however these are not compatible with LC-MS. For an LC-MS based workflow, therefore, the protein can be removed from the beads by direct tryptic digestion. This yields an equimolar mixture of, in theory, all of the proteins in a sample. The highly abundant proteins will saturate the beads, whereas less concentrated proteins may be present in such low copy numbers that some peptide ligands will remain unbound. To reduce this effect, and to increase the possibility of obtaining an equimolar protein mixture, it is necessary to overload the beads with protein sample.

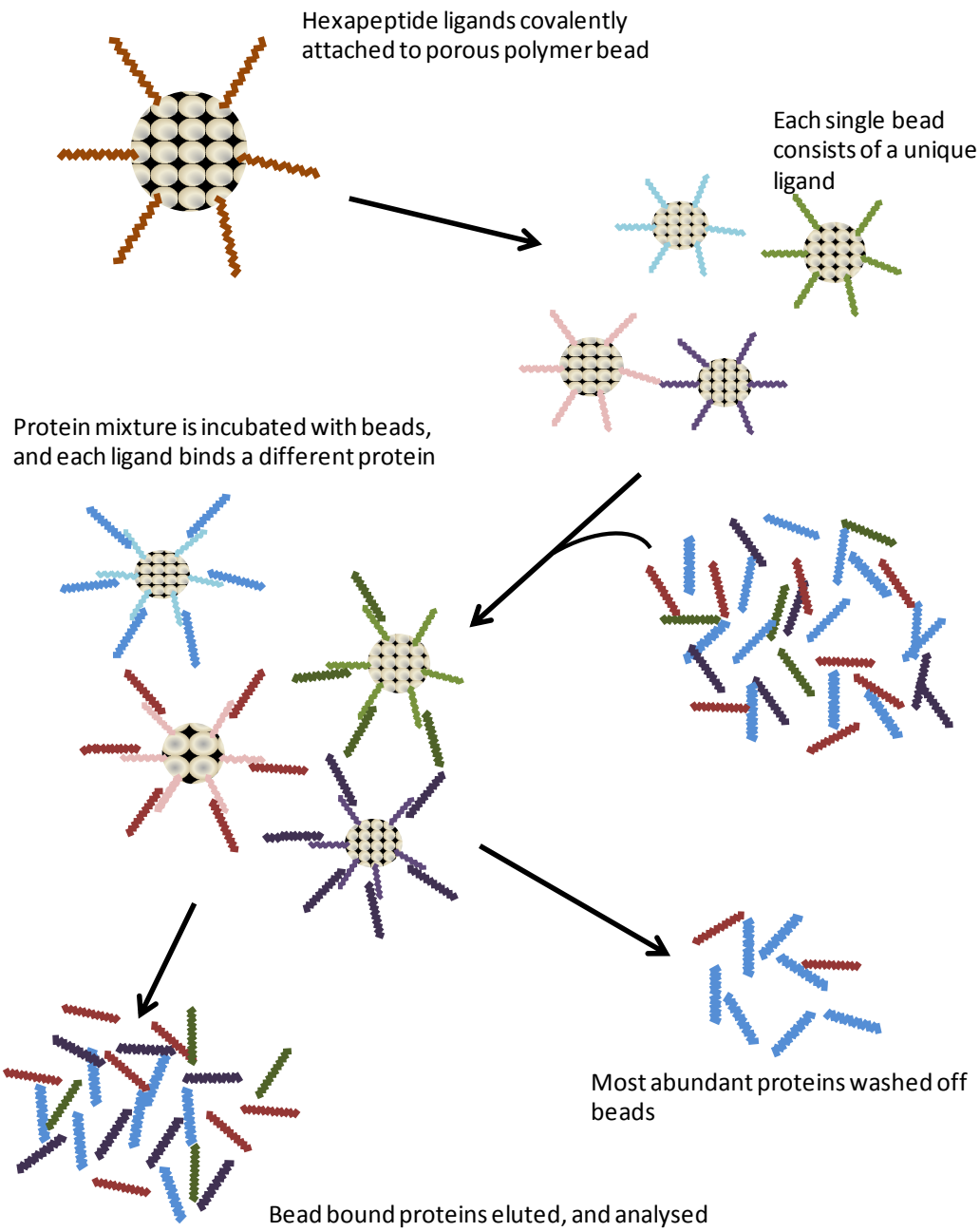


Figure 5.1. Combinatorial bead library normalisation.

Hexapeptide ligands are covalently bound to porous beads. Each bead contains multiple copies of a unique peptide, and therefore binds a finite number of copies of a single protein. Any excess protein is washed off, and the bead bound protein is eluted, resulting in an equimolar mixture of proteins.

When the technology was first implemented, it was based on the concept that, with each individual bead possessing a unique peptide, and each peptide binding a unique protein, an individual bead would bind a single, unique protein to maximum bead capacity, and the complex protein mixture eluted from the library would be equimolar. However, it has since been reported that multiple hexapeptide ligands are capable of binding a single protein at different efficiencies, and also that a single bead is able to bind multiple proteins (Huang *et al.*, 1996; Boschetti *et al.*, 2007). These findings indicate that eluate from a bead binding experiment may not be the equimolar mixture anticipated, and may contain a high concentration of some proteins able to bind numerous peptides, possibly via a number of different interactions. Also, in some samples, it is possible for proteins to bind the beads asymmetrically, as certain proteins, such as the heat shock proteins in chicken skeletal muscle, can bind to each other and form a scaffold effect, with multiple layers of protein surrounding one bead (Rivers *et al.*, 2011).

Despite these potential issues, combinatorial hexapeptide libraries can greatly increase the proportion of the proteome it is possible to identify. Although the technology was first developed to enable the identification of low abundance proteins in blood plasma, it has since been applied to a wide variety of samples, from human bile, serum and urine, to animal biological fluids and tissues, and even beverages (table 5.1).

Table 5.1 Examples of the use of combinatorial hexapeptide ligand libraries

Sample	Number of untreated proteins	Number of normalised proteins	Reference
Human urine	134	383	(Castagna *et al.*, 2005)
Human serum	115	305	(Guerrier *et al.*, 2006)
Human erythrocytes	535	1524	(Roux-Dalvai *et al.*, 2008)
Human CSF	476	1149	(Mouton-Barbosa *et al.*, 2010)
Swine plasma	1708	2657	(Tu *et al.*, 2011)
Chicken skeletal muscle	35	360	(Rivers *et al.*, 2011)
Sea urchin coelomic fluid	26	82	(Fasoli *et al.*, 2012)
Spinach	132	236	(Fasoli *et al.*, 2011)
Mango pulp	374	2693	(Fasoli & Righetti, 2013)
Champagne	0	43	(Cilindre *et al.*, 2014)

The suitability of the technology for a wide range of sample types also means the method has far reaching applications, in the discovery of potential biomarkers for disease, such as cancer (Meng *et al.*, 2011; Monari *et al.*, 2011) or liver disease (D'Amici *et al.*, 2012), and even conditions such as preeclampsia (Liu *et al.*, 2011). Another application of the technology is in the discovery and removal of contaminants or impurities (Fortis *et al.*, 2006; Antonioli *et al.*, 2007). In addition, the technology can assist in protein discovery, and has enabled the identification of two novel proteins in snake venom (Calvete *et al.*, 2009).

5.2 Aims

ProteoMiner™ combinatorial hexapeptide beads were exploited here in an attempt to simultaneously analyse the entire dynamic range of the *S. cerevisiae* proteome. Currently, the SGD lists 5070 open reading frames as verified, and 750 as uncharacterised. Both crude cell lysates and normalised material will be analysed by LC-MS to assess the reduction in dynamic range, and potential identification of the whole yeast proteome concurrently.

In addition, the possibility of quantifying these proteins will be examined. Although equaliser bead technology is usually used in a qualitative manner, owing to the removal of some protein material, quantification may be achievable for the less abundant proteins, providing they bind the beads in a linear fashion and the beads can be assumed to have bound the total protein present in the sample provided it does not reach the full capacity of the bead. In order to test this theory the mode of binding was explored using a series of increased loadings on a set amount of beads to find the saturation level, and assess the linearity of binding below saturation level. There are also a number of ways protein can interact with equaliser beads, and in a final series of experiments, the effect of ionic conditions on protein binding was investigated in an attempt to elucidate the protein-ligand interactions.

These experiments will be performed with the aim of accessing low abundance proteins, exploring the idea of using the technology to quantify the full dynamic

range of yeast proteins, and also examine the potential for entire protein complexes to bind to the bead library.

5.3 Results and Discussion

5.3.1 *S. cerevisiae* cell lysate protein normalisation using ProteoMiner™ beads

Proteominer™ beads were purchased from Bio-Rad and prepared according to manufacturer instructions. The bead slurry was thoroughly mixed and a 10 µl aliquot was taken, and diluted 1 in 10. The bead content of this dilution was then counted under a microscope, to an average of 76 beads /µl. Corrected for dilution, this corresponds to approximately 60,500 beads in an 80 µl aliquot used for these experiments. Theoretically, this amount should be able to bind all of the proteins in a complex proteome, provided sufficient quantities are loaded.

To initially assess the normalising effect of the bead library, a *Saccharomyces cerevisiae* cell lysate was prepared from a cell pellet by bead beating and diluted to a 10 mg/ml solution. The cell lysate was centrifuged, then added to an 80 µl aliquot of beads and normalisation was performed according to the protocol outlined in section 2.16. This was repeated on three separate occasions, with three individually prepared cell lysates. On comparing the cell lysate starting material and CPLL treated material directly off the beads using gel electrophoresis (see figure 5.2 a for an example) there is a visible effect on the protein composition of the sample. The yeast cell lysate starting material is less asymmetric than the plasma samples the beads were originally designed for, at a dynamic range of five orders of magnitude, and the starting material is complex to begin with, a number of proteins are visible by electrophoresis. However, the banding pattern on the gel indicates that the normalised material contained more visible protein bands. In addition, there was a reduction in the intensity of the strongest bands, indicating some extent of normalisation has occurred.

The bead-bound protein was then digested and quantified using label free analysis on a Waters Synapt G2 in HDMS^E mode. In each replicate, there was a clear

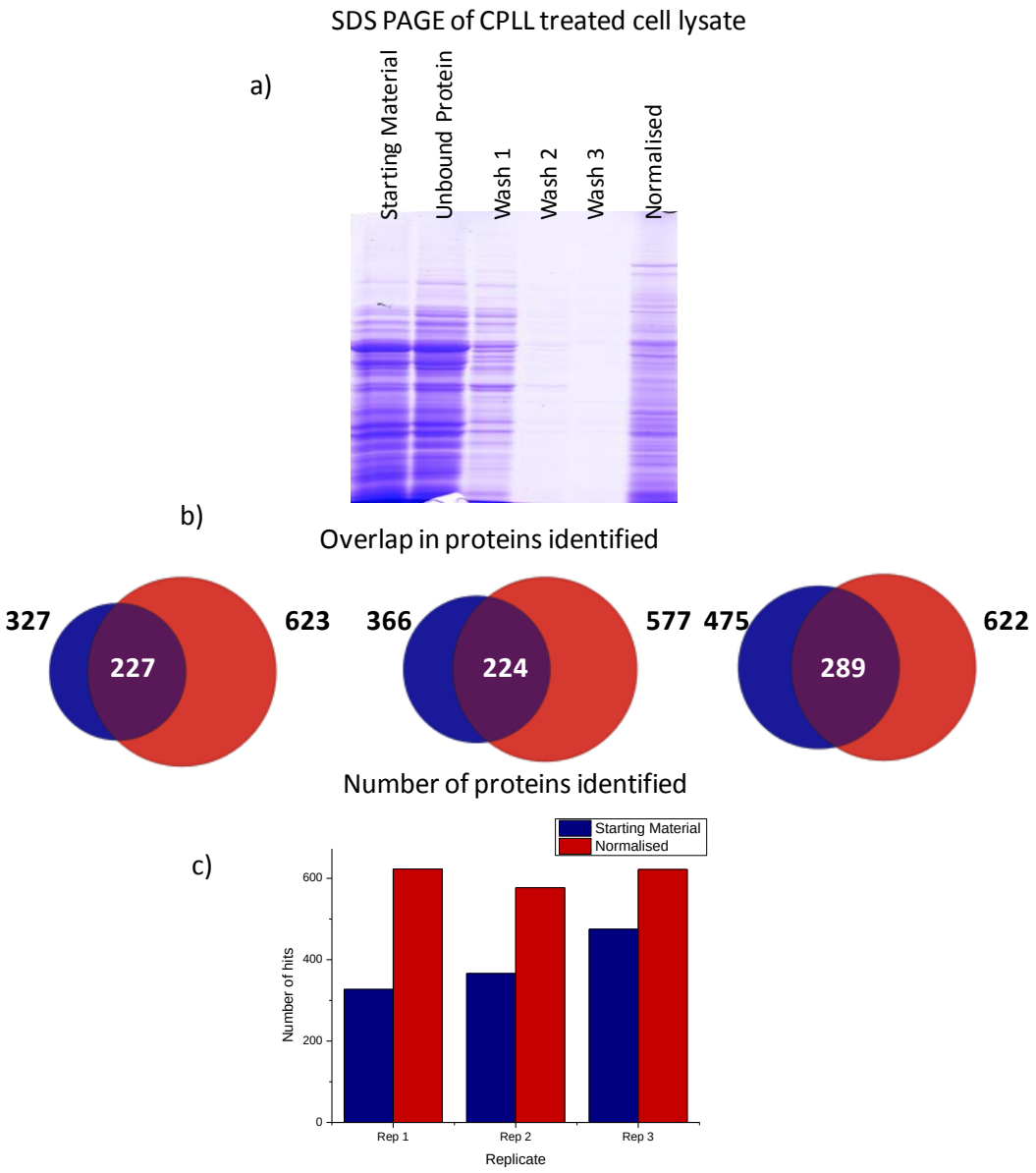


Figure 5.2 Proteins in cell lysate before and after CPLL treatment

Three replicates of a 10 mg/ml cell lysate treated with 80 µl beads, a) SDS PAGE of centrifuged cell lysate before and after CPLL treatment b) The crossover in the proteins identified in the two sample types in three replicates c) The number of proteins identified in three replicates of yeast cell lysate starting material and normalised protein following tryptic digest and run on the Waters Synapt G2 in HD MSE mode.

increase in both the number of proteins identified, and consequently the total material identified. In the first replicate, there was an increase from 327 to 623 proteins identified, replicate 2 showed an increase from 366 to 577, and replicate 3 from 475 to 622 (figure 5.2 c). This amounts to an average increase of 56% from a starting material digest to the bead bound protein. The increase in the number of proteins identified indicates that treatment with the CPLL library is bringing more proteins into the quantifiable range. In each replicate, there are a number of common proteins found in both the starting material and the bead treated material (figure 5.2 b). There are also a number of unique proteins found in starting material and treated. Those in the normalised material may have been present in too low an amount to be identified in the starting material. However, those unique to the starting material should have bound the beads, and still been observed after CPLL treatment. The increase in the number of identified proteins also results in an increase in the total amount of material quantified (figure 5.3 a), from 8,100 to 38,600 fmol in replicate 1, 8,200 to 20,800 fmol in replicate 2 and 13,100 to 26,900 fmol in replicate 3. This corresponds to an average increase in material of 193%, much more than the increase in protein numbers, suggesting that a number of proteins have shown a big increase in concentration.

Despite indications that the equaliser technology has indeed normalised the sample, and increased the number of hits, the overall dynamic range of the sample does not undergo the expected reduction. The aims of using this technology would be to bring the less concentrated proteins up in concentration, while reducing the more abundant. There is little change in the concentration of the least abundant protein identified in each case; however, there is in fact an increase in the amount of the most abundant.

Table 5.2 The lowest and the highest calculated amount, in fmol, of an individual protein in the starting and treated material.

Sample	Replicate 1	Replicate 2	Replicate 3
Starting material minimum	0.6	0.9	1
Starting material maximum	268	159	279
Normalised minimum	1	0.8	1.3
Normalised maximum	464	330	451

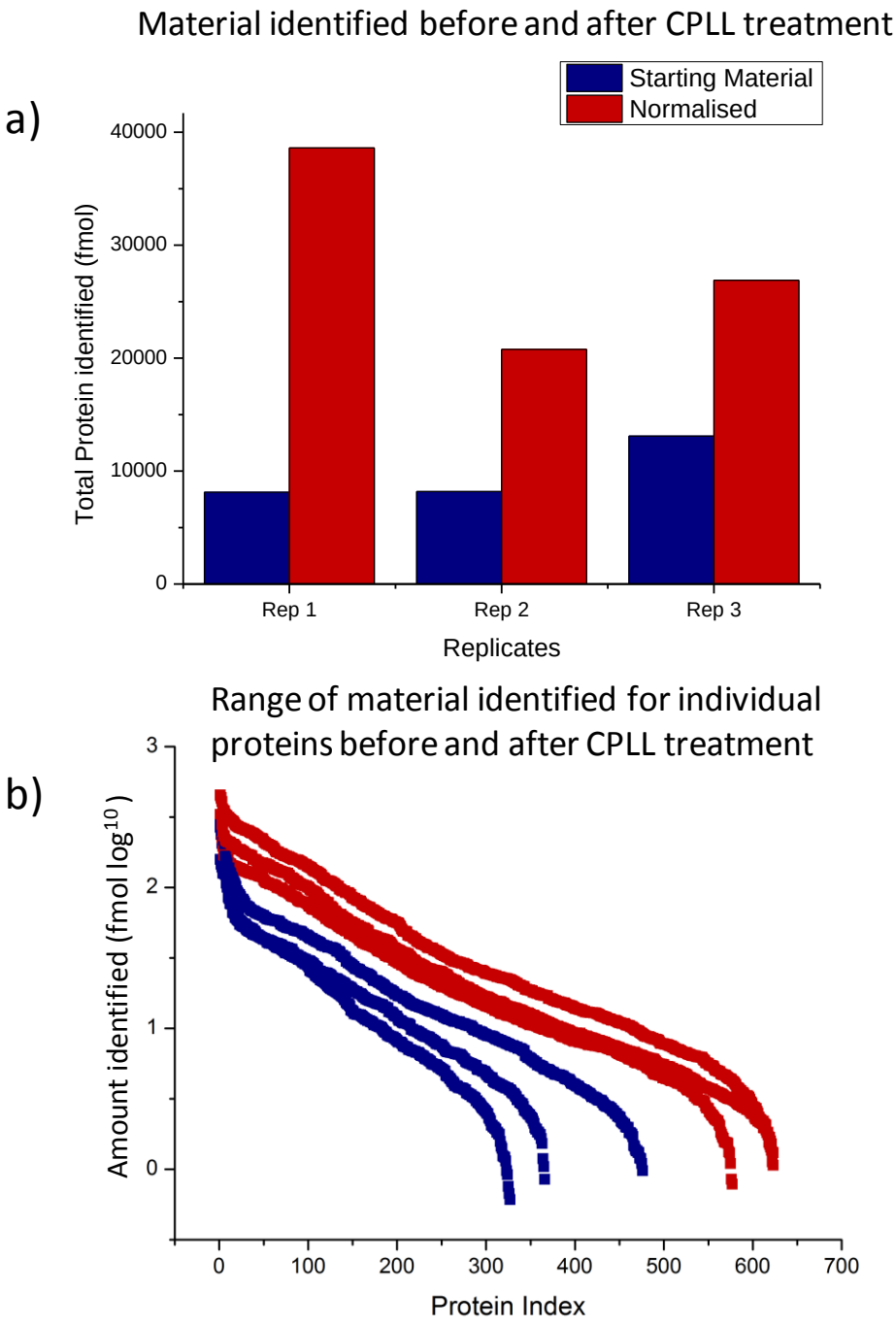


Figure 5.3 The amount of material in three replicates of starting material and normalised sample

Bead bound protein is tryptically digested, a) the total fmol of protein identified in three replicates of starting material and normalised sample analysed by LC-MS using the Synapt G2 and b), the protein index plotted against the fmol amount of each protein

This indicates that the effect the CPLL is exerting on the dynamic range is in fact the opposite of that anticipated, there is actually an increase in the range of protein concentrations. The beads are only supposed to bind a limited amount of each protein, but there are a number of possible reasons why some proteins may bind excessive amounts. One explanation is that the ligands are not actually specific, and the most abundant proteins are overloading the beads Alternatively some proteins may have a higher propensity to bind the beads, which would result in some less concentrated proteins in the starting material dominating the normalised material. If the former explanation is true, the abundant proteins in the starting material will also be the most abundant in the CPLL treated material. In all three replicates of starting material, the most abundant protein was enolase II (ENO2), while in the normalised material replicates the top protein was pyruvate decarboxylase (PDC1), translation elongation factor 1 (TEF1) and elongation factor 2 (EFT2). The most abundant proteins were different in the starting material and the normalised, which indicates that the highly abundant proteins in the normalised sample were not over-saturating the beads simply due to them being the dominant proteins in the sample. Therefore, the most concentrated proteins may be present at excessive concentrations in the normalised material due to a higher affinity to the beads in comparison to other proteins. This was perhaps a result of having more exposed surface areas capable of interacting with the bead bound ligands, or a propensity to interact with other proteins already bound to beads. This effect will be examined in a later series of experiments.

From these results, it is evident that the use of equaliser bead technology has extended the range of identifiable proteins in the yeast cell lysate prepared in these experiments. The increase in proteins calculated as mid concentration indicates that more are being brought into the identifiable range. This effect is highlighted when the fmol amount of each protein identified are plotted on a logarithmic scale (Figure 5.3 b), which shows a shallower incline in the normalised material, as more proteins are quantified in the middle of the total dynamic range identified. This indicates that, in line with the expected normalisation effect, some proteins have been brought up in concentration, from below the limit of detection and therefore

not identified in the original cell lysate, to within the range of concentrations quantifiable by the mass spectrometer. Some proteins have therefore increased in concentration, but some have also been reduced in concentration (figure 5.4). The top ten most abundant glycolytic enzymes identified in the starting material exemplify this behaviour (figure 5.5), and most are reduced in concentration after treatment with the beads, which would be expected of the highly concentrated proteins. However, some of these do increase in concentration, for example, alcohol dehydrogenase (ADH1) in replicate 1, and pyruvate decarboxylase (PDC1) increase in concentration in all three replicates, which indicates they are binding well in excess of the expected maximum bead saturation level.

Gene Ontology (GO) annotation provides a way to group proteins based on function, biological process or component. To investigate if particular protein subsets show more propensity to bind the beads, GO searches were performed on the SGD website, to group proteins according to component, and function. There was no indication that a particular component subset showed more propensity for binding (figure 5.6). On examining protein function, it appeared the RNA polymerase proteins, translation initiation factor, and proteins with nucleotide binding activity are concentrated by CPLL treatment (figure 5.7).

Within the three replicates of starting material there are 251 (43%) common proteins, while in the normalised material there are 391 (46%) common proteins (figure 5.8 a, b). This indicates that the variation in protein identifications is caused by the experimental process, rather than variation introduced by the beads. Approximately a third of proteins that were in the starting material did not bind the beads (refer back to figure 5.2 c). In theory, all of these should be visible in the normalised sample, as they were originally present at sufficient concentration to be detected, and most should possess an exposed surface area which has an affinity to at least one peptide. Of the 262 proteins that were uniquely identified in the starting material across all three replicates, 17% are common across all three replicates, and 53.82% of proteins are unique to one replicate (figure 5.8 c). These proteins should each possess the ability to bind, but were unable to in that replicate. The fact that only 17% of the proteins that did not bind the beads do so

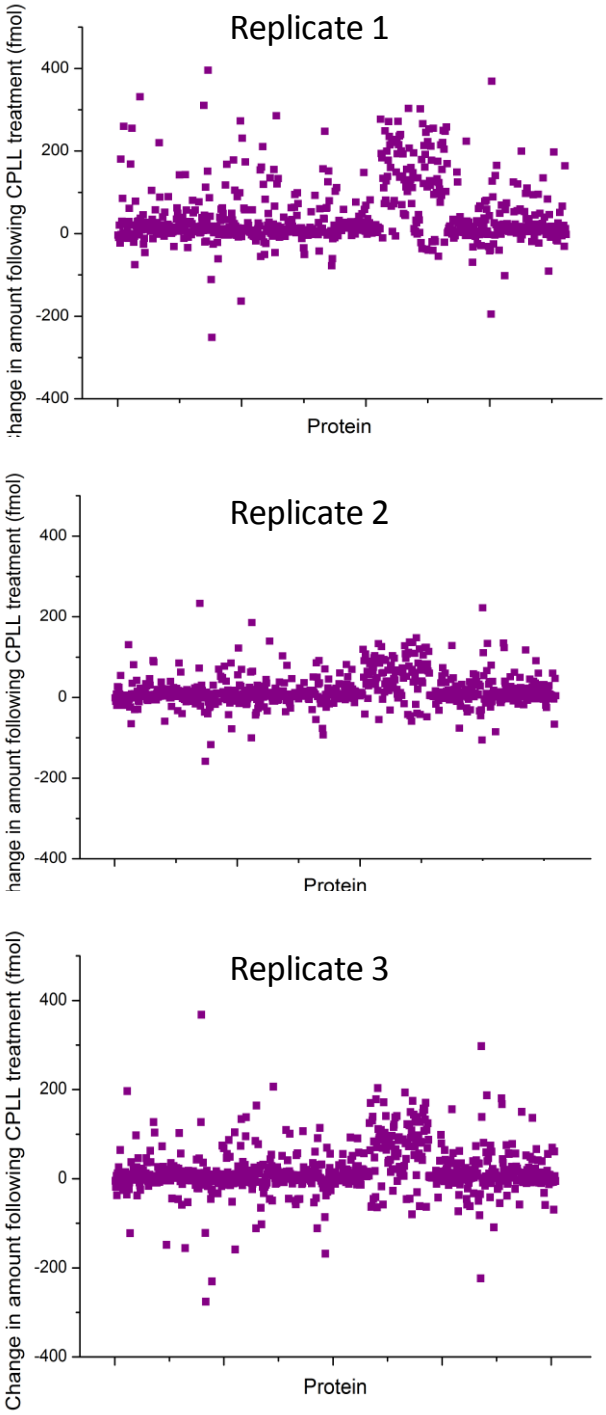


Figure 5.4 The change in amount of each protein identified in three replicates

Scatter graphs depicting the fmol change in each protein from starting material to CPLL treated material, in three digested replicates run on the Synapt G2.

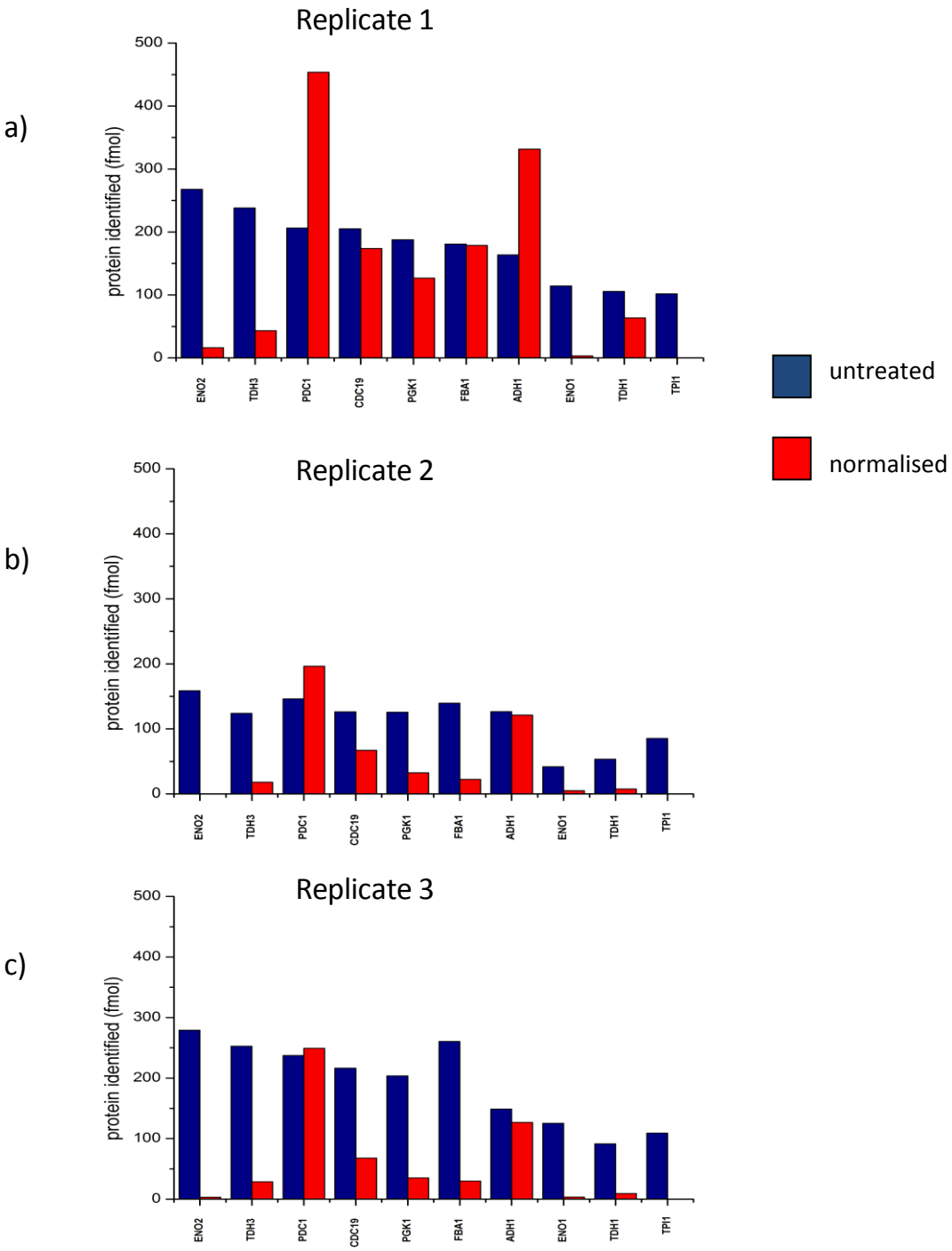


Figure 5.5 The most abundant glycolytic enzymes in starting material proteins before and after CPLL treatment.

The amount (in fmol) of the top ten glycolytic enzymes found in the starting material and CPLL treated sample in a) replicate 1, b) replicate 2, and c) replicate 3 as quantified PLGS using Synapt G2 data.

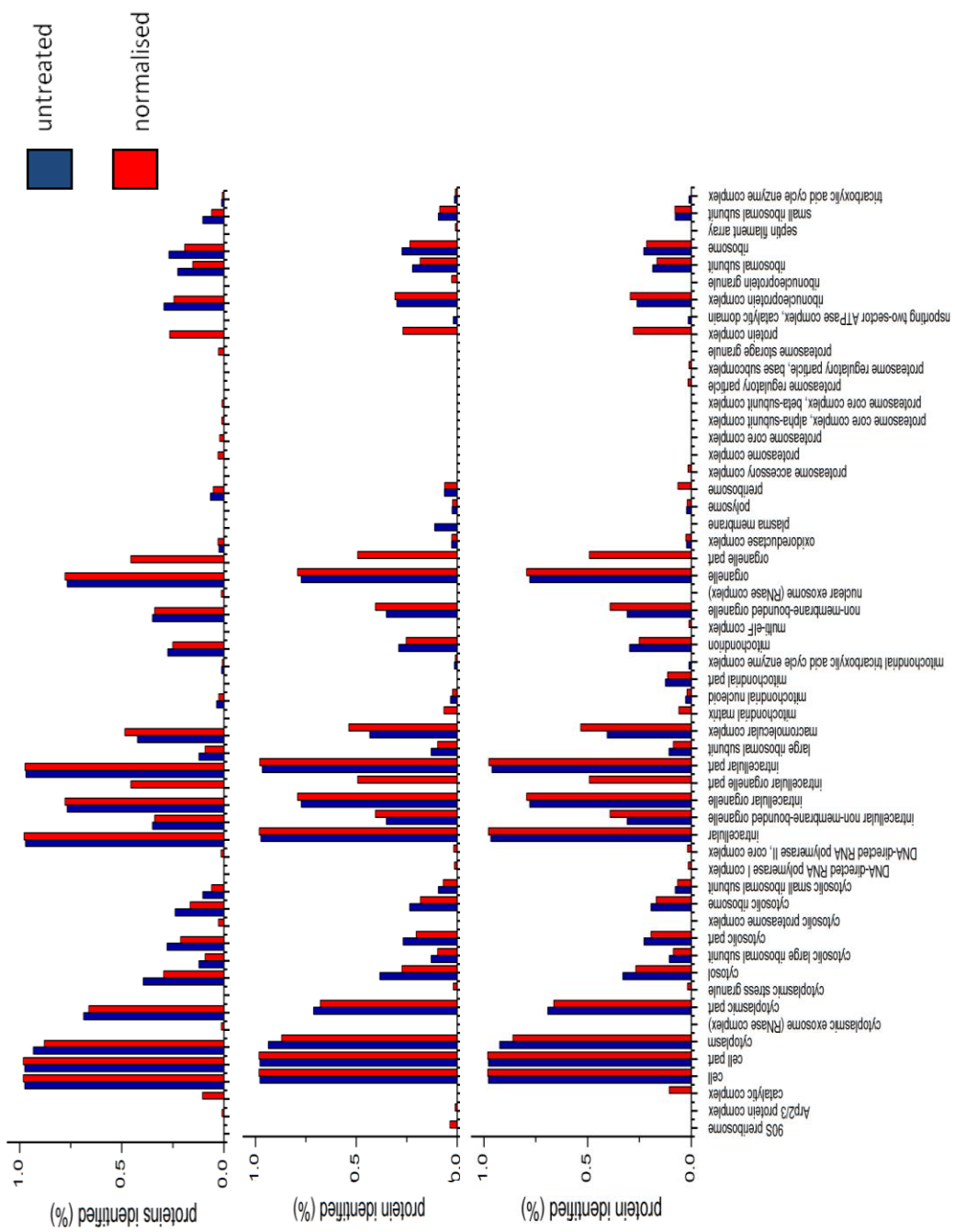


Figure 5.6 Gene Ontology data for the starting material and treated samples

Gene ontology search was performed on the Synapt G2 data using the SGD GO term finder (v. 0.83) results based on protein compartment, for three replicates of starting material and CPLL treated sample.

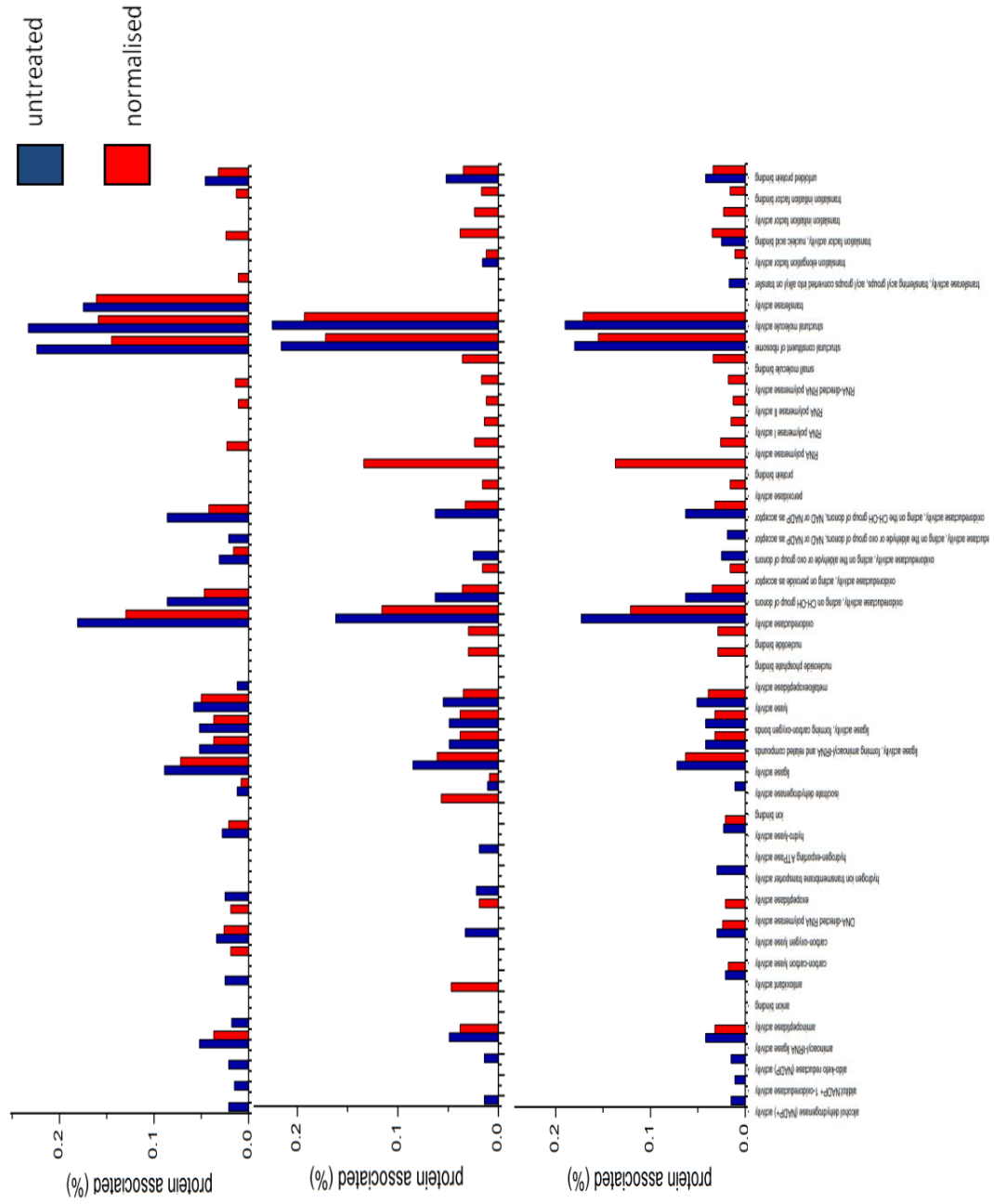


Figure 5.7 Gene Ontology data for the starting material and treated samples

Gene ontology term results based on protein function, for three replicates of starting material and normalised sample analysed using the Waters Synapt G2.

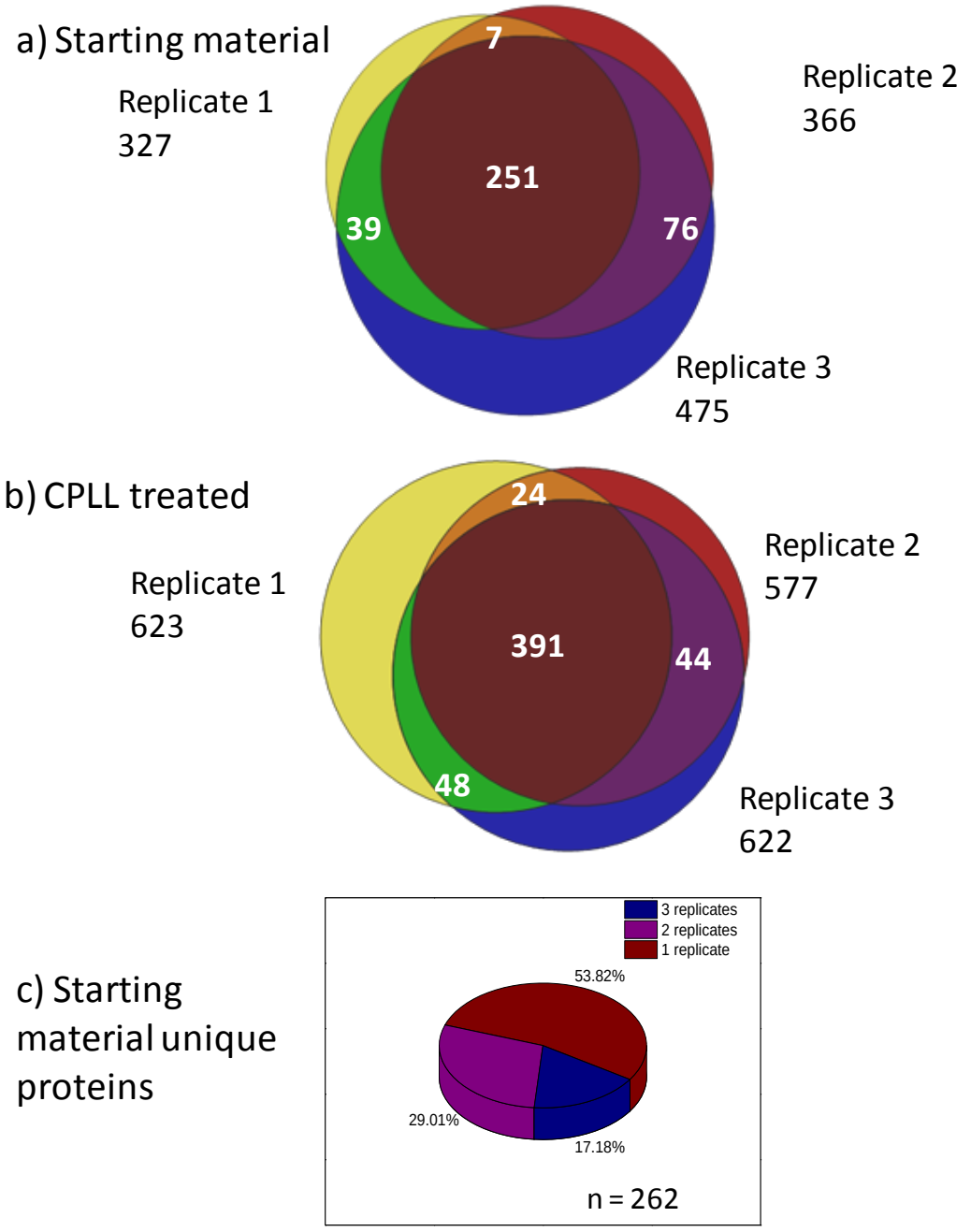


Figure 5.8 Venn diagrams depicting the common proteins identified across three replicates of untreated and treated material.

a) three replicates of cell lysate starting material and b) three replicates of CPLL treated material. Samples were run on the waters Synapt G2 and label free quantification was performed using PLGS, and c) the unique proteins found in starting material, common to all three replicates.

reproducibly indicates that the unbound protein in each experiment is a random effect, and not a result of the proteins inability to bind. This variation could be due to the inherent degree of variability in the equaliser beads, the very concept behind the equaliser beads relies on the fact that there are a number of different peptide ligands. Another explanation for the variability is the number of interactions possible, in variation in the ligands available, and a large amount of proteins able to bind the beads in different interactions, and different affinities. This could mean that, when an aliquot of a complex mixture of peptide ligands is brought together with a complex sample, able to interact with each other in a number of different ways, the sheer complexity of the binding process may lead to a large degree of variability.

5.3.2 The effect of increased protein loading

To further examine the process of protein binding to the beads, the effect of introducing increasing amounts of protein onto the same quantity of beads was tested. Centrifuged cell lysate was prepared at 0.05 mg/ml, 0.1 mg/ml, 0.5 mg/ml, 1 mg/ml, 5 mg/ml and 10 mg/ml. Three replicates of each loading were performed from the same sample dilution, in order to examine the variation derived from the beads, and minimise the variation introduced from the sample. The only source of variation should therefore be the difference between aliquots of beads. The gel electrophoresis of each loading shows additional bands in the normalised material when compared to the starting material (figure 5.9), which increase in intensity with each loading, up to 1 mg/ml. In the lower loadings, of 0.05 mg/ml and 0.1 mg/ml, the starting material is not concentrated enough to be visible, however a number of proteins are observed in the normalised sample lane. This indicates that a number of proteins have increased in concentration at these lower loadings.

The samples were digested and run on both Waters Synapt G2 Q-TOF and the Thermo Scientific Q Exactive Quadrupole Orbitrap mass spectrometers. The total number of proteins identified increases with each higher loading, until after the 1 mg/ml cell lysate, when there is no further increase in number of identifications (figure 5.10 a). Data from both mass spectrometers is presented, showing that the

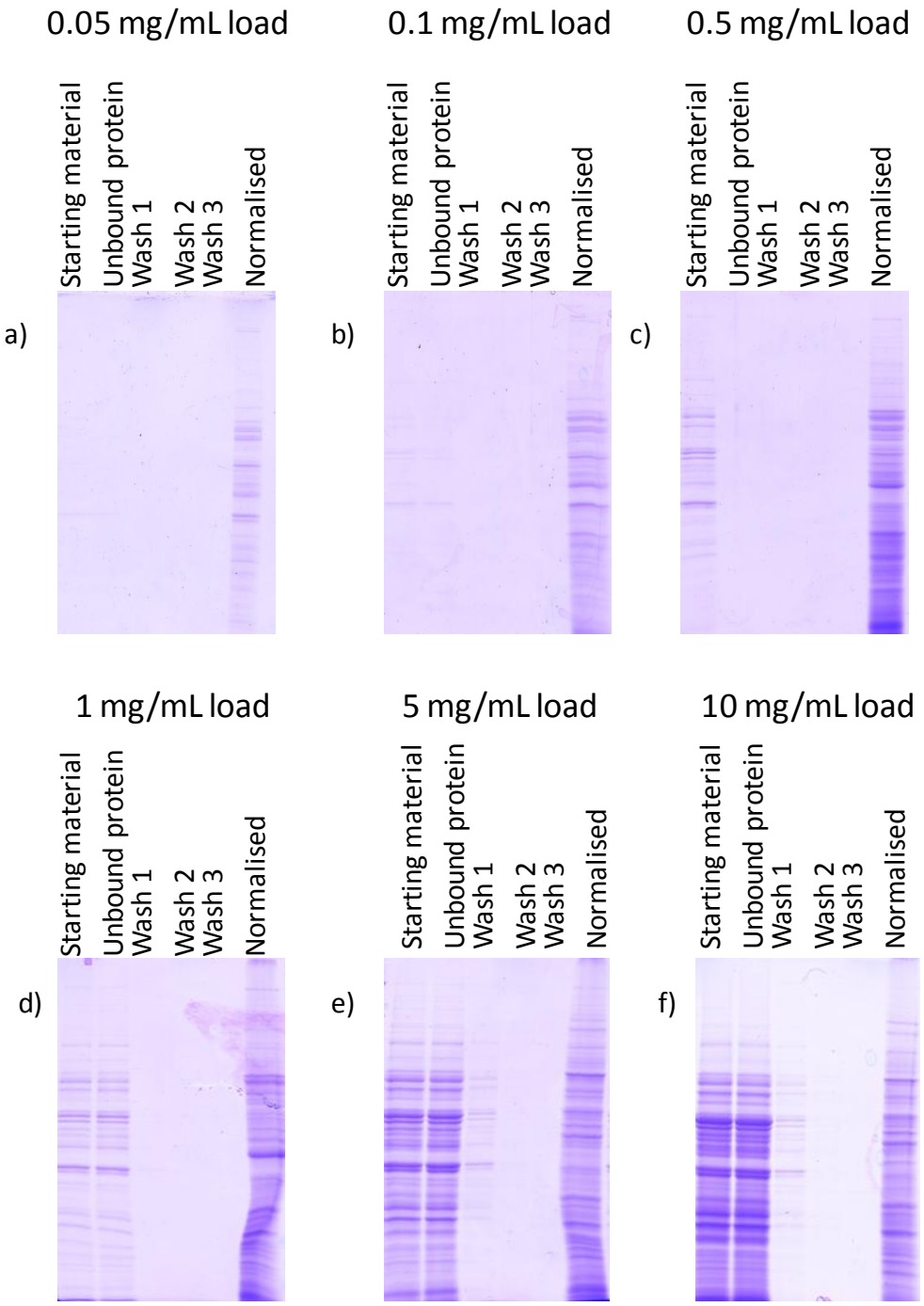


Figure 5.9 Gel electrophoresis of increasing CPLL loading experiments

80 µl bead library aliquots were loaded with clarified cell lysate dilutions ranging from a) 0.05 mg/ml, b) 0.1 mg/ml, c) 0.5 mg/ml, d) 1 mg/ml, e) 5 mg/ml and f) 10 mg/ml. Following 2 hour incubation periods unbound protein was removed by washing with 5 x 1 ml phosphate buffer, and then 10 µl aliquots of the normalised bead bound material starting material, unbound protein, and washes 1-3 were mixed 1:1 with sample buffer and analysed by SDS PAGE.

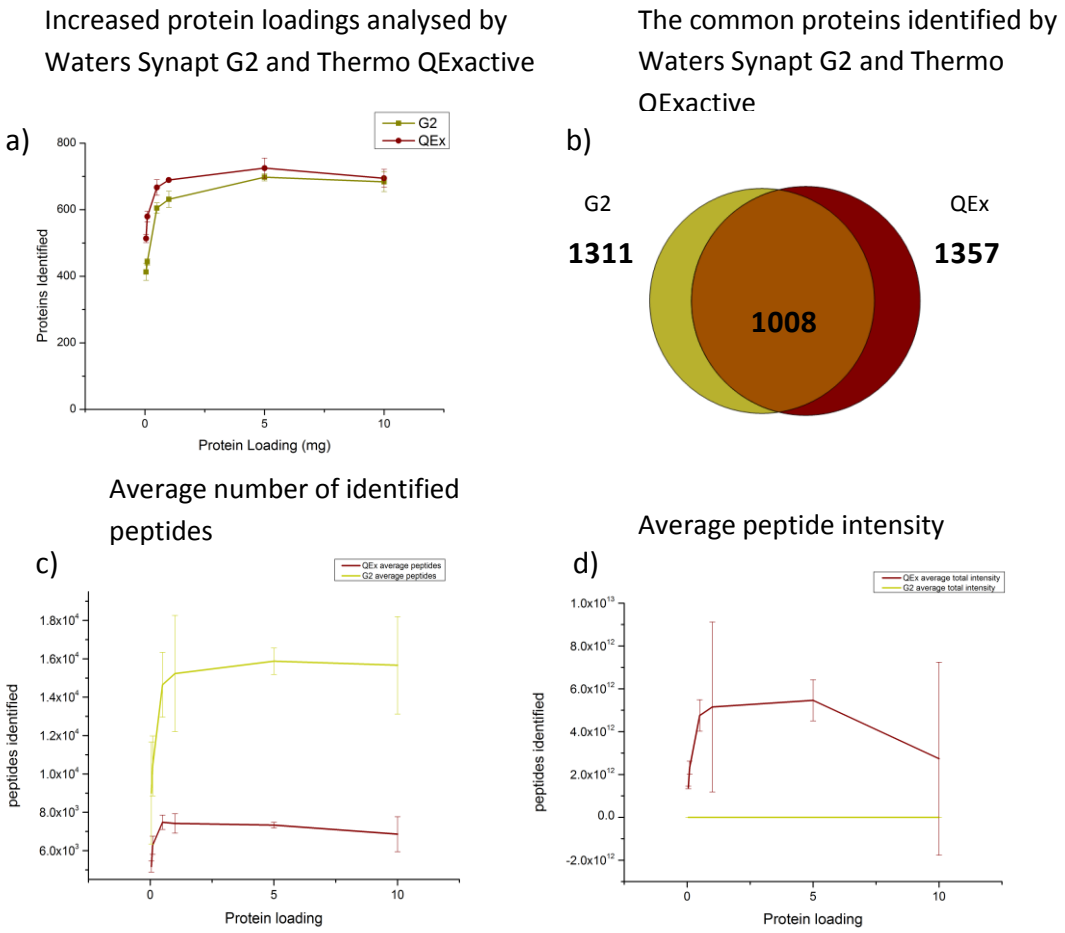


Figure 5.10 Proteins identified in three replicates of increasing CPLL load

a) The average number of proteins identified in three replicates of CPLL treated material, run on a Thermo QExactive and label free quantification performed using MaxQuant, and run on a Waters Synapt G2 and analysed using PLGS and b) the overlap in proteins identified in the three replicates. C) the average number of peptides identified in three replicates of the loading experiment, and d) the average total peptide intensity identified by the G2 and QExactive.

ion trap identified more proteins, although at the higher loadings there is less difference in the number of proteins identified between the two instruments, the average G2 hits at the top load are 683, and the average QExactive 693. When all loadings, run on both instruments, are combined there are 1660 proteins identified, and 1008 of these are common to both instruments, 303 unique to the synapt G2 and 349 unique to the QExactive (figure 5.10 b). The QExactive may have identified more at lower loadings because of the ion trap separating ions better, making it more likely to identify more proteins by triggering fragmentation of a greater number of ions. At higher loadings, it would be expected that more normalisation has occurred, the peptides injected would be of a more equal abundance, and therefore a wider of range ions are able to be selected for fragmentation in a data dependent run and identified by the G2. The number of peptides identified in each loading differs on the two instruments, however, and the G2 returns more peptides in total (figure 5.10 c). However, the QExactive returns an increased total peptide intensity on average, ranging from 3-6 $x10^{12}$, in comparison to 3-4 x 10^{8} (figure 5.10 d).

When analysing the total fmol of material identified (figure 5.11 a) there is an increasing amount of protein until saturation is reached after approximately 1 mg/ml. In fact, there is even a slight decrease in the amount of material identified by both mass spectrometers at the highest loading in the QExactive data. At all loadings there is more material identified by the ion trap mass spectrometer than the Q-TOF, despite the similar number of proteins identified. Despite the difference in quantity calculated, there is a linear trend in the amount of each protein identified between the G2 and QExactive (figure 5.11 b), however this is lessened in the higher loadings. The number of proteins identified and the total fmol identified show the expected trend across the loading experiments, an increase across the successive loads, and little change following the loading at which saturation point is achieved. This behaviour is consistent with the idea that each bead will bind a specific protein up to capacity, and then any excess will be removed during the washing stage. Due to the higher number of identifications, the Q Exactive data will be used for the rest of this chapter for comparing the different protein loadings.

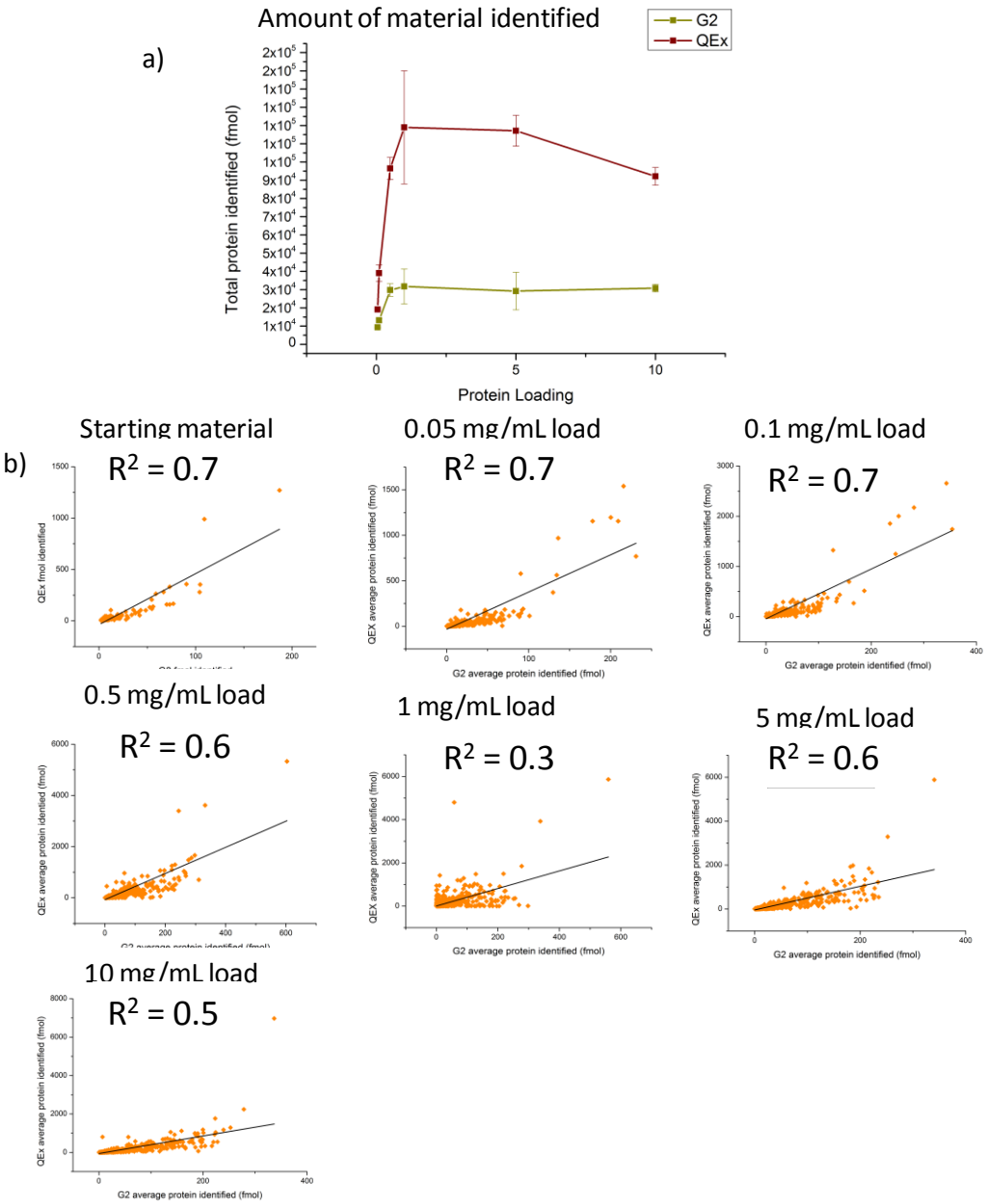


Figure 5.11 The amount of material identified in the loading experiment samples on the G2 and the QExactive

a) the total amount of material calculated in each CPLL loading when analysed by label free quantification using PLGS for the G2 generated data and MaxQuant for the QEx generated data, and b) a comparison of the fmol amount of the common proteins found in both data sets for each loading.

Despite this expected binding pattern there is, again, an increase in the overall dynamic range of protein quantifications observed between the normalised material at higher loadings and the starting material (figure 5.12 a). However, the distribution of proteins within this range differs, and despite the range of protein concentrations being wider than the starting material, the distribution does show a shallower incline, on account of the additional proteins identified being in the middle of the dynamic range. The difference in the dynamic range of proteins is the effect of proteins being both increased and decreased in concentration in comparison to the starting material (figure 5.12 b). The scatter graphs depict the protein concentrations increasing and decreasing in all loadings, and the effect is more pronounced in the higher loading, with one protein increasing in concentration by over 7000 fmol.

When the protein binding is investigated on an individual level it indicates that the process of equaliser bead normalisation follows the original premise of reducing the abundance of some proteins, and increasing others. When the average loading data is plotted as a heat map (figure 5.13) it highlights these two effects, some proteins bind the beads in increasing amounts at the successively higher loadings, and some become less concentrated. However, in addition to this some appear to bind increased amounts at successive loadings, and at later loading decrease. The difference in behaviour is indicated by the dual colour system on the heat map, a blue colour indicates a decrease in fmol, while red indicates an increase. Some low abundance proteins are not detected at all in starting material, but bind the beads at successively larger amounts with each loading increase, until a saturation point is achieved. This would be the expected behaviour, and examples of this behaviour include phosphofructokinase 1 (PFK1p), alpha-ketoglutarate dehydrogenase 1 (KGD1p) and an ATPase of the HSP70 family (KAR2p) (figure 5.14).

Other effects are also highlighted, for example some proteins increase in concentration at lower loadings, before reaching a maximum and then decreasing at higher loads, rather than remaining at saturation level. This may be the result of the overloaded proteins competing for specific binding sites as the concentration of

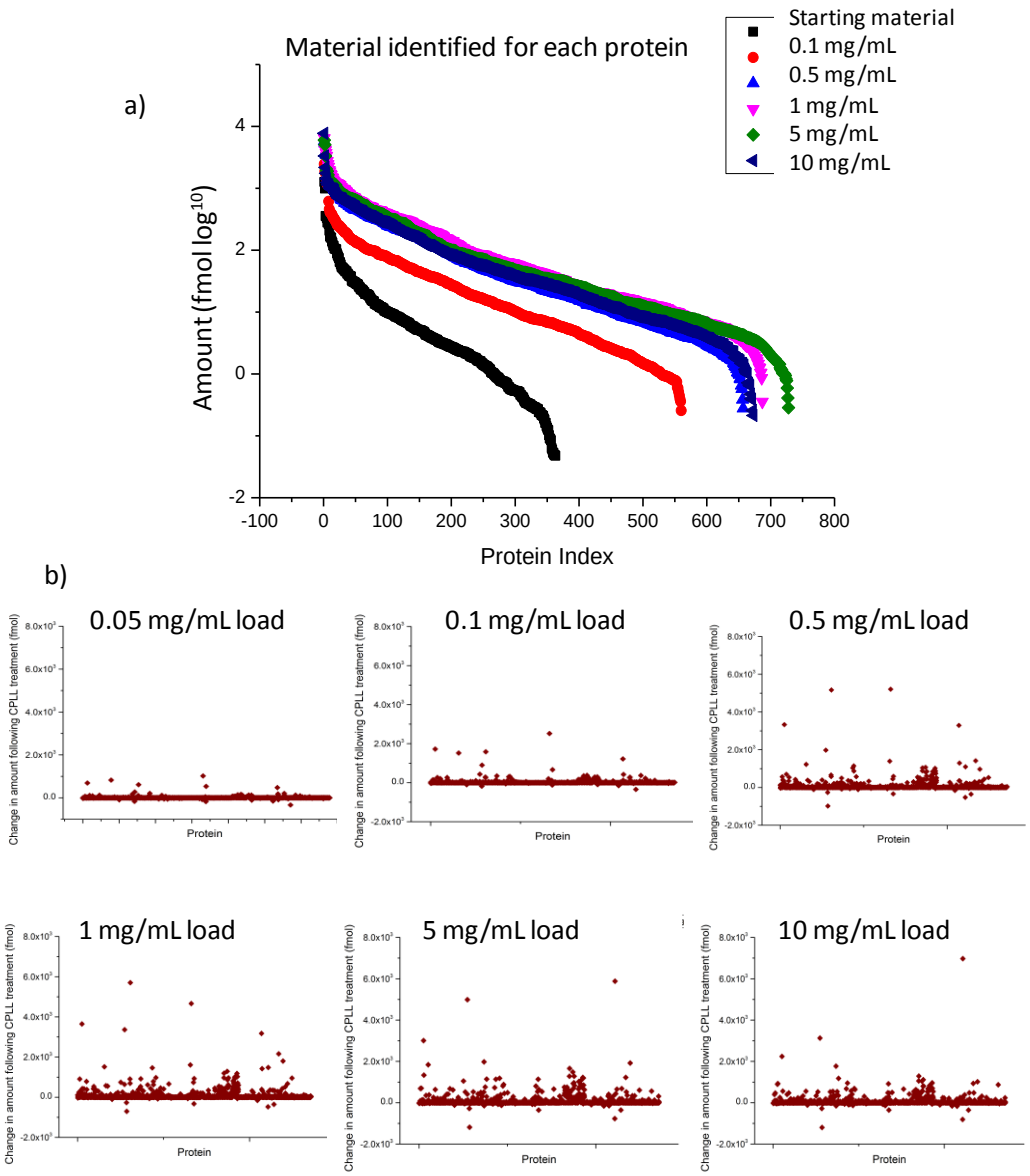


Figure 5.12 The quantification for each individual protein

Label free quantification is achieved by iBAQ processing of QExactive data, and a) the average material (in $\log^{10}$ fmol) calculated for each individual protein in the loading experiments, arranged in descending order, and b) the change in the fmol amount of each protein identified in each loading.

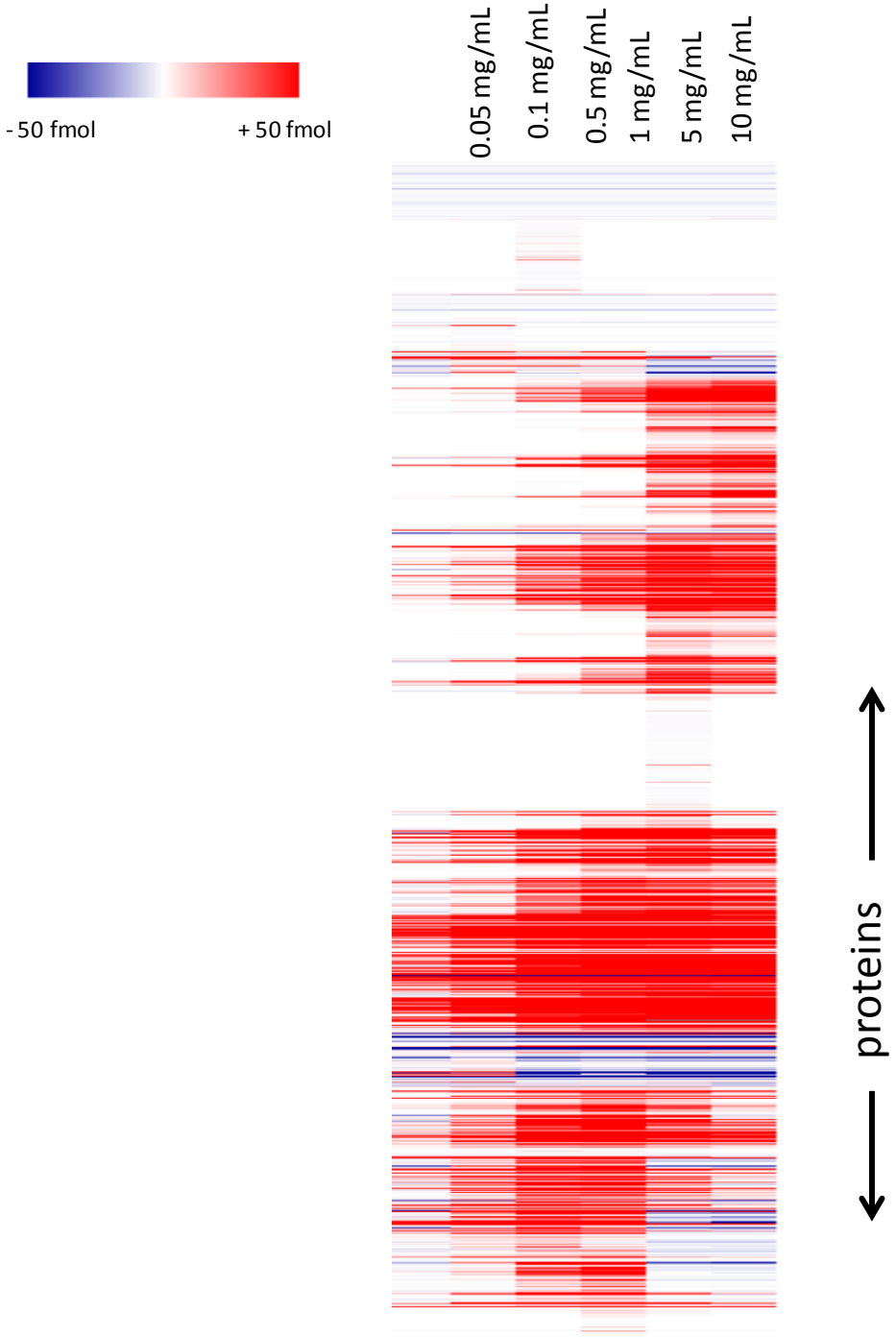


Figure 5.13 The average change in fmol of each protein across increasing loadings

The average fmol of protein in CPLL treated sample was calculated and the starting material fmol was subtracted to give the change in amount of each protein. Heat map was generated using heirarchical clustering in MeV software. The clusters indicate different binding patterns, proteins increasing in quantity across the loadings, proteins that reach saturation point at a lower loading, and stay steady at successive loadings, and proteins which reduce in concentration at higher loadings.

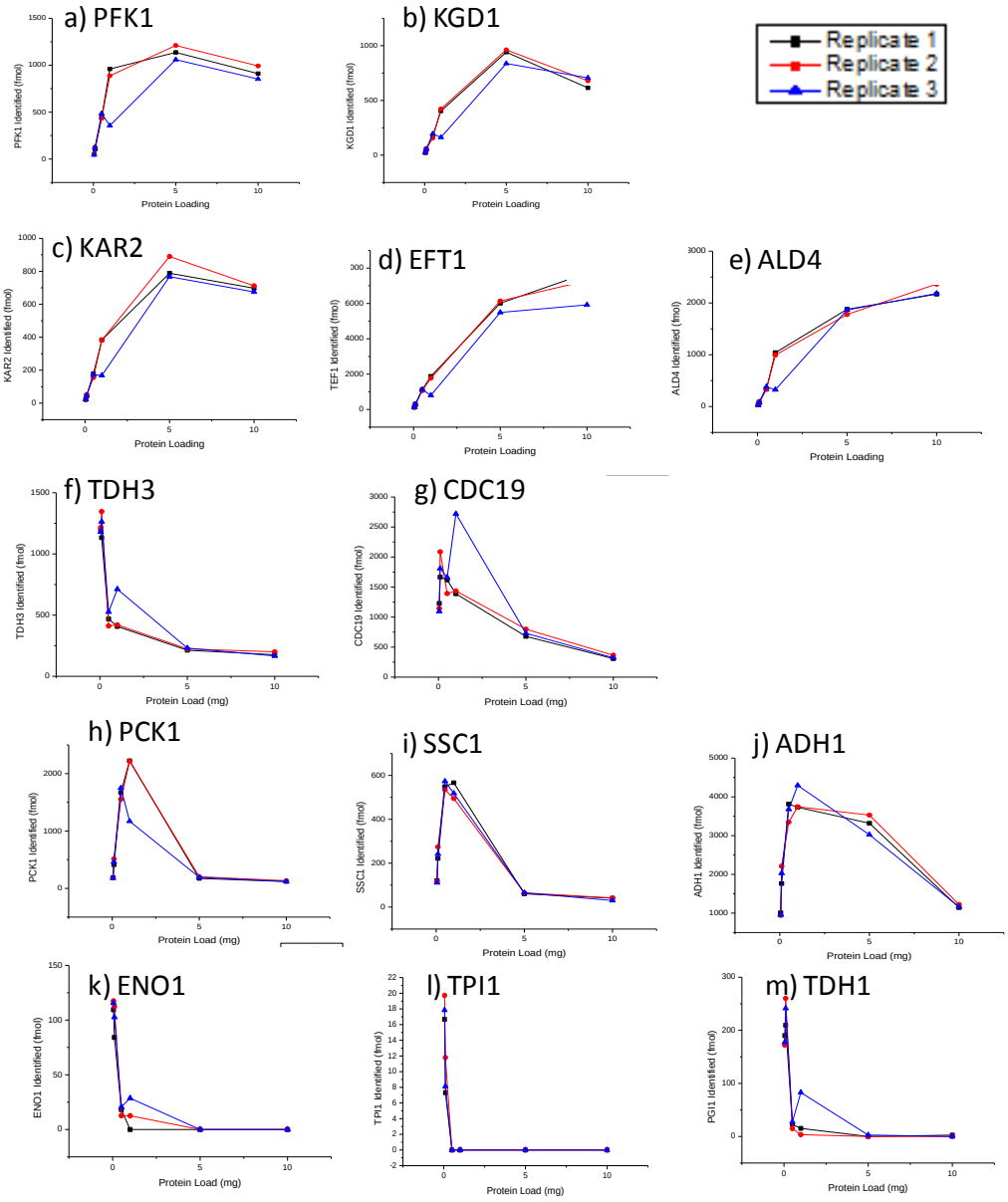


Figure 5.14 Examples of different protein binding patterns across CPLL loadings

CPLL loading experiment digests are run on the QExactive and the amount of each protein calculated using MaxQuant. a) - m) examples of individual protein behaviours in all three replicates.

other proteins also increases. However, some proteins continue to increase in concentration throughout the loadings, and bind far in excess of the average protein concentration. The average amount of an individual protein identified are 138, 133, and 127 fmol in the three replicates. The most abundant protein in the normalised material, translational elongation factor 1, is present well in excess of this average at approximately 7700, 7300, and 5900 fmol in the three replicates. The next two top proteins in all three normalised samples are mitochondrial aldehyde dehydrogenase (ALD4), and elongation factor 2 (EFT1), which are both present at over 2000 fmol. These proteins also appear to be increasing in concentration up to the top loading, which indicates that they are binding beyond what should be the level of saturation. The vast over-representation of these few proteins could be the effect of them having an affinity to more than one specific peptide, or the proteins binding each other to form a matrix, with successive layers of proteins around a single bead. These are all examples of binding effects of proteins that are becoming more concentrated after treatment with the equaliser beads, and are showing at least three different binding patterns.

There is then the opposite effect of the bead library to consider, the proteins which are brought down in concentration during the experiment. Here also, there is variation in the binding patterns observed. The abundant glycolytic enzymes glyceraldehyde-3-phosphate (TDH3) and pyruvate kinase (CDC19), are present in both the starting material and the low loadings at high amounts, and the concentration decreases over successive loadings (figure 5.14). Since less protein is observed after more is introduced to the beads, it indicates the original excess of material binding was due to some degree of non specific binding. One possible explanation is that these proteins, which are present in high concentrations, are binding ligands which may have a higher affinity for other proteins. Therefore, when the loading is increased and a greater variety of proteins become available, the ligand will preferentially bind another protein, now present at sufficient amounts to displace the more abundant protein.

With other proteins such as enolase 1 (ENO1p), and triose-phosphate isomerase (TPI1p), there is an immediate reduction, the concentration of protein decreases

from the first loading and in fact after the 0.1 mg/ml load there is no material identified at all. This behaviour may indicate the protein has a very low affinity for any of the peptide ligands, so when the availability of other proteins increases, the abundant protein is rapidly displaced. Some less abundant proteins, such as phosphoenolpyruvate carboxykinase (PCK1p), a HSP70 family ATPase (SSC1p), and alcohol dehydrogenase (ADH1p) increase in concentration over the lower loadings, but at subsequent loads (above 1 mg), instead of showing a steady concentration to indicate they had reached saturation, show a decline in abundance. This could indicate there is more competition for binding sites occurring at higher loadings.

Another possible protein interaction in the CPLL binding is the interaction of proteins with each other. The proteins that exist as part of complexes may remain part of these assemblies while also binding to the bead library. The 20S proteasome subunits all bind the bead library in the same pattern, increasing in concentration until the 1 mg/ml load, where saturation is reached, except in replicate three, where the quantity of each protein in the 1 mg/ml load is lower than the first two replicates (figure 5.15). The same binding effect is observed for the nine subunits of the exosome identified in the loading experiment, which also reach saturation above a 1 mg/ml total protein load (figure 5.15). The same binding could be seen in all of the subunits due to either the proteins all exhibiting the same binding pattern, or because the proteins remain part of a complex, or partial complex. Fatty acid synthetase is a heterodimeric complex, and both subunits are observed in the loading experiment. The two subunits increase in concentration up to a 1 mg/ml load, after which the concentration falls (figure 5.16). This binding pattern is also seen in two subunits of the ribonucleotide-diphosphate reductase (RNR) complex, RNR2 and RNR4, which form a heterodimer (Chabes *et al.*, 2000). The two proteins both increase in concentration up to 1 mg/ml, and decrease in subsequent loadings (figure 5.16). The fact that the subunits of these two complexes all decrease at higher loadings may lend support to the theory that proteins are remaining bound to each other while also binding to the CPLL library.

When the starting material and all three replicates of six loadings are combined, there are in total 1360 proteins identified. There is some variation between

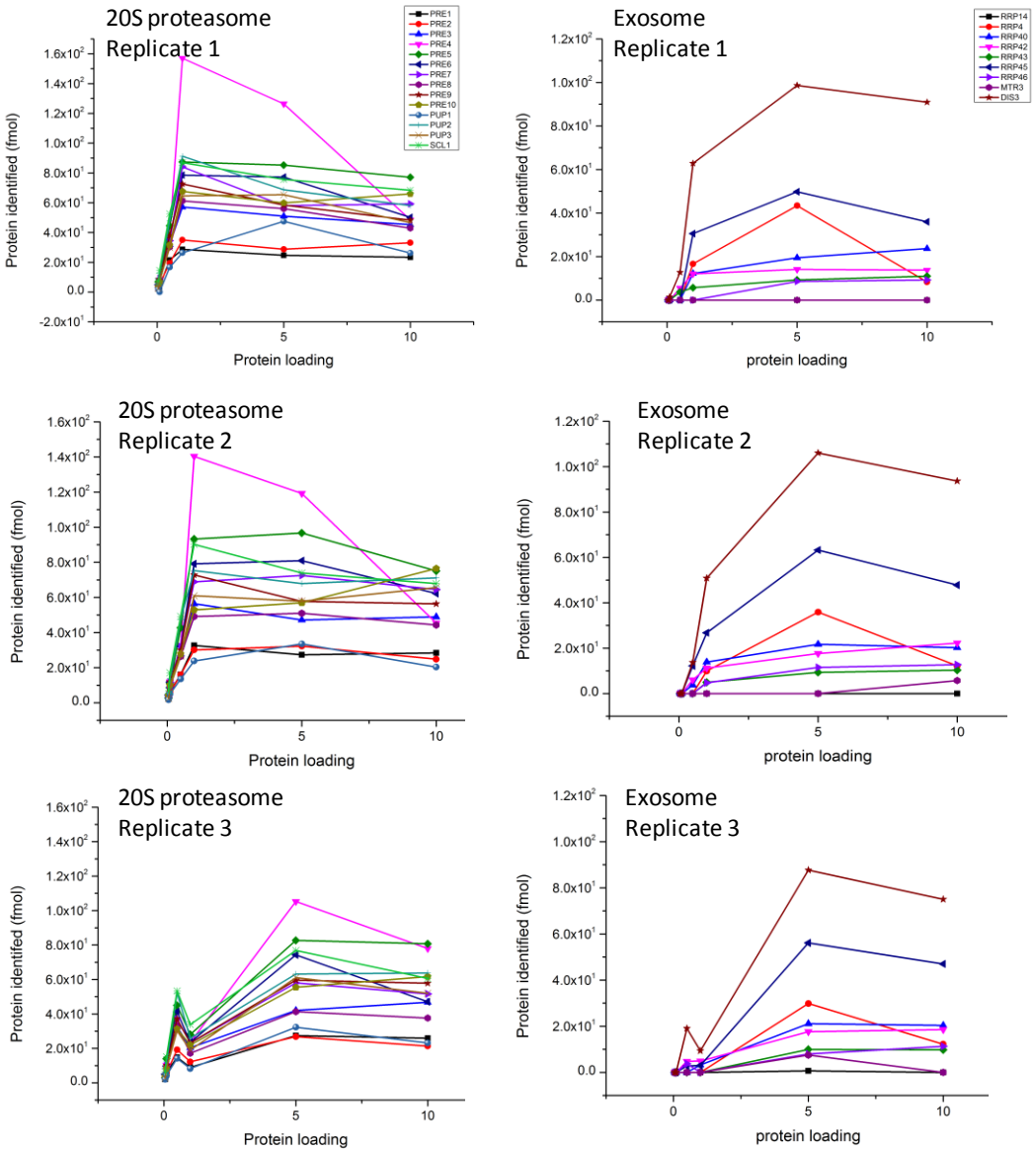


Figure 5.15 The binding of 20S proteasome complex subunits at increasing bead loadings.

The 14 subunits of the 20S proteasome are quantified by MaxQuant in increasing protein loadings on the CPLL aliquots. Nine out of the ten subunits of the exosome complex are identified in increasing loadings.

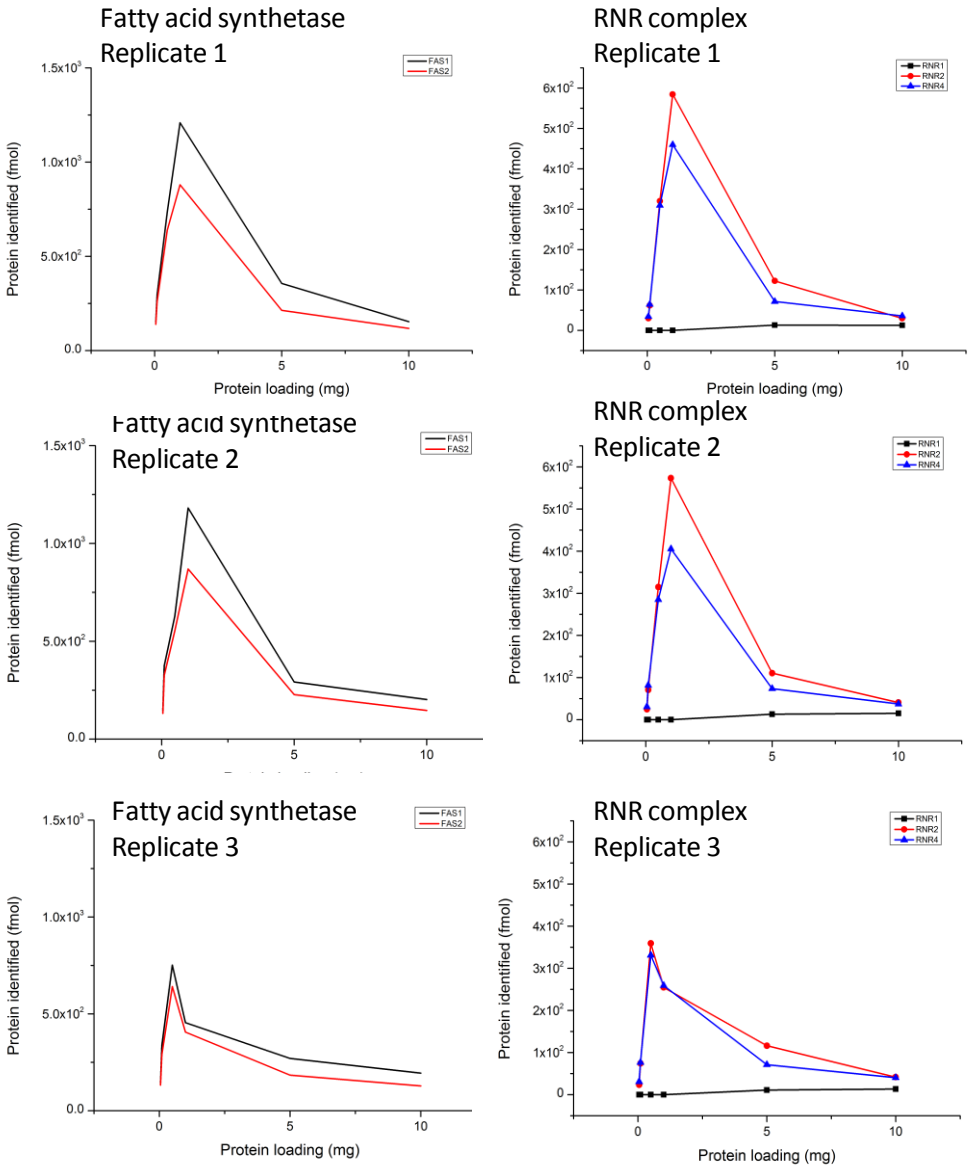


Figure 5.16 The binding of Fatty Acid synthetase and RNR complex subunits under increasing loadings

The heterodimeric fatty acid synthetase complex was quantified by MaxQuant in increasing protein loadings on the CPLL aliquots. The two subunits of a ribonucleotide-diphosphate reductase subcomplex were identified in increasing loadings.

replicates, but to a lesser degree than earlier experiments when three separate cell lysates were used. Between the three replicates of the highest loading of normalised material there was a 72% overlap in protein identifications. However, there are still unique proteins found, 18% of proteins identified in the highest load of the normalised material are unique to one replicate. In the lowest loading 80% of proteins are identified in all three replicates (figure 5.17). As the cell lysate used was the same for all three replicates, the variation observed is likely to come from the beads. This could be natural variation in the bead aliquot, as each 80 µl aliquot of bead slurry may contain a different bead composition, despite repeated mixing. Across all of the different loadings each of the 364 proteins identified in the starting material was also found in the normalised material. This suggests that the starting material proteins which were shown not to bind the beads in section 1.2.2 did so due to competing for binding sites, not an inability to bind, and the effect has been counteracted by the number of replicates performed in this experiment.

The variation in binding and the different loading patterns observed indicate that the protein-ligand binding is a complex interaction between all of the proteins present. The concentration of protein identified is therefore not a true reflection of the amount of that particular protein in the cell lysate, but a result of individual proteins competing for binding sites that they each have higher or lower affinities for. The proteins which are observed to a vast excess in the normalised material may then have a structure which provides them with a number of surface binding sites, and are therefore capable of binding not only multiple hexapeptide ligands, but also possibly other proteins already bound to the beads, resulting in a 'scaffold' effect of proteins building up in successive layers on the beads.

The result of the complex binding processes is that is not possible to define the saturation point of an individual protein in the beads. Without defining a concentration for the saturation of beads with one protein, it is not possible to know if all of the protein in a sample has bound to the beads. Therefore it will not be possible quantify the less abundant proteins bound to the beads which are below saturation level.

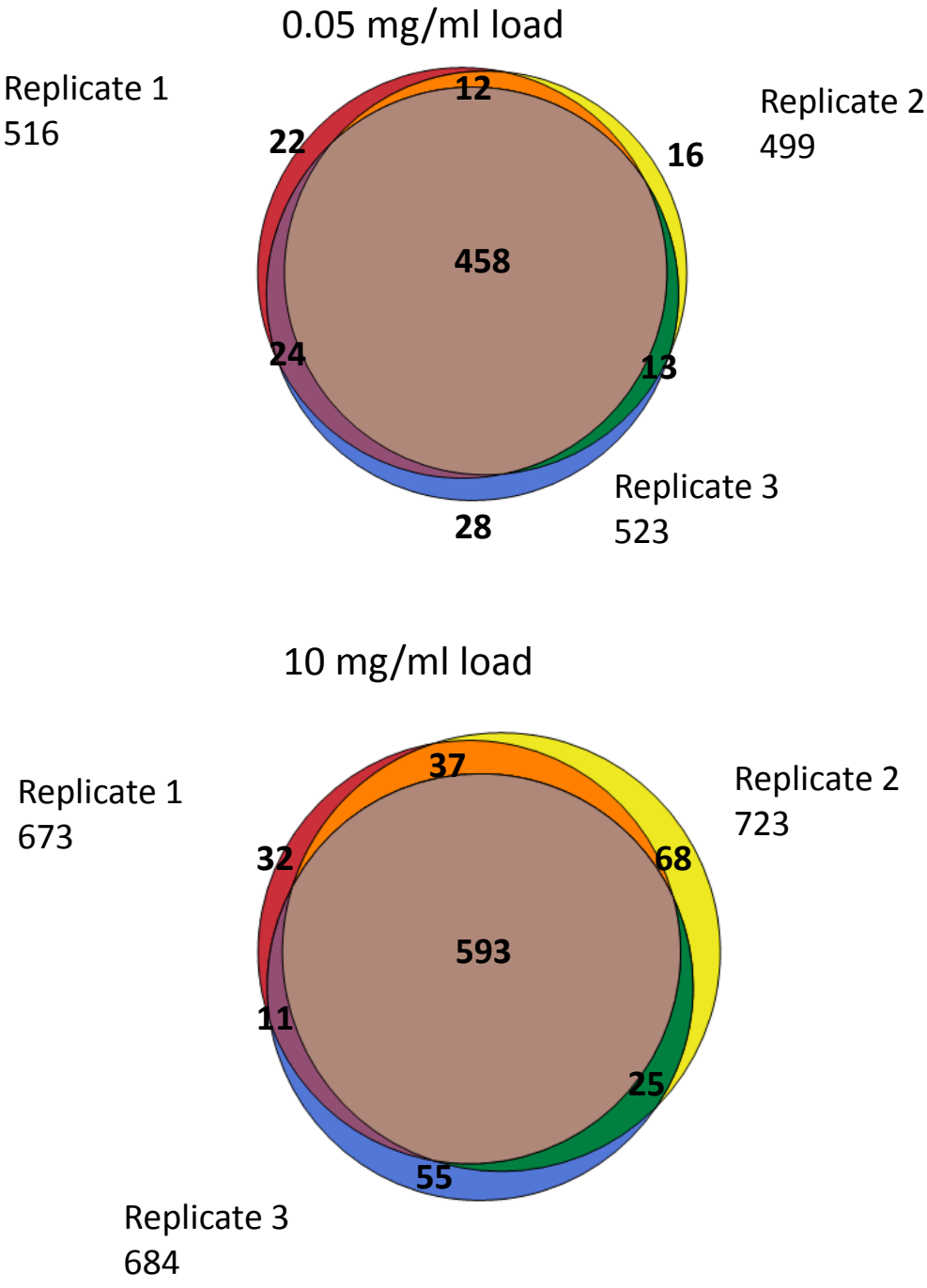


Figure 5.17 The variation observed in the proteins identified in three replicates of CPLL treated sample

Venn diagrams depicting the variation in the proteins identified in three replicates of the lowest CPLL loading (0.05 mg/ml) and the highest loading (10 mg/ml)

5.3.3 Protein normalisation in varying ionic conditions

The next set of equaliser bead experiments was designed to test the binding efficiency of the proteins to the ligands under different ionic conditions. There are numerous ways the proteins can interact with the bead bound hexapeptide ligands, using the same methods as those governing protein folding and protein-protein interactions. These include electrostatic forces, hydrogen bonding, hydrophobic interactions, and Van der Waals forces. The interactions range in intensity, and will differ in strength depending on which exposed amino acid side chains are involved in the interaction. Therefore the strength of the protein binding will vary with each protein – ligand pair. It was expected that increasing the ionic conditions of the environment would affect the protein binding, some electrostatic forces may be disrupted, as sodium or chloride ions may interact with the charged areas of the amino acids side chains, or the hexapeptide ligands, and any potential hydrophobic interactions will be enhanced.

Sample dilutions were prepared in two different buffers, to a final concentration of 20 mM sodium phosphate buffer, and 20 mM sodium phosphate, 150 mM sodium chloride buffer. The cell lysate was originally prepared in 20 mM sodium phosphate, and diluted to 0.2 mg/ml and 20 mg/ml in 20 mM sodium phosphate buffer. Each of these was then diluted 1:1 with 20 mM sodium phosphate and 300 mM sodium chloride/20 mM sodium phosphate buffers, to a final concentration of 0.1 mg/ml and 10 mg/ml. They were then used in equaliser experiments with 80 µl of beads as described in section 3.1. The samples were run on the Thermo QExactive.

The total number of proteins bound to the beads is lower in high ionic conditions in all three replicates, 570 proteins identified as opposed to 704 in low ionic conditions in replicate 1, 526 compared to 634 in replicate 2, and 498 compared to 537 in replicate 3 (figure 5.18). The lower number of total proteins identified in high ionic conditions indicates that for many proteins, the protein-ligand interaction is based on electrostatic interactions, which have been disrupted by the addition of salt. The reduction in hits also suggests and there is less propensity for the yeast proteins to interact with the beads via hydrophobic interactions, as these would be

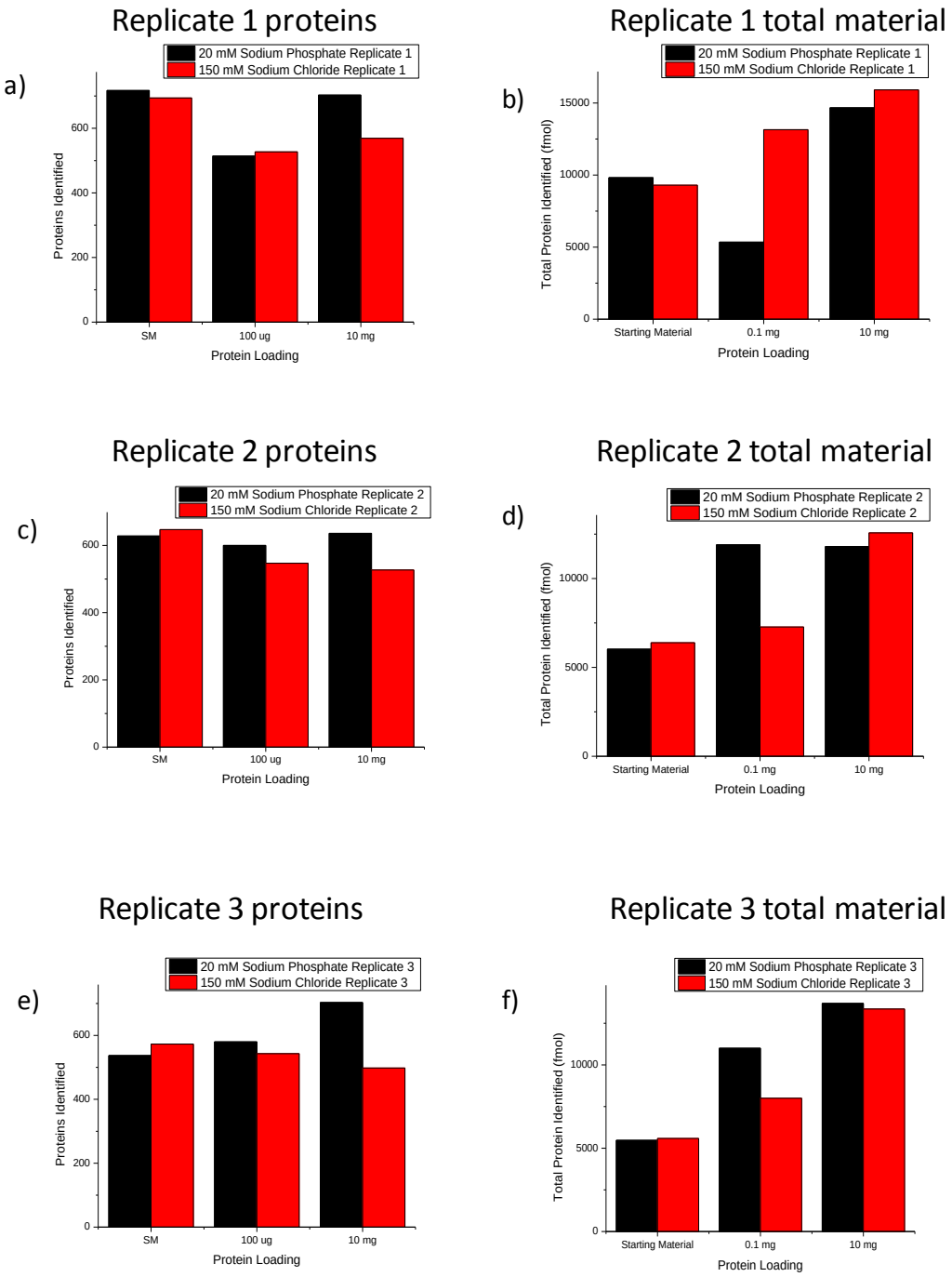


Figure 5.18 The number of proteins and total material identified in low and high ionic strength buffers

Three replicates of starting material, 0.1 mg/ml load and 10 mg/ml load in low and high ionic strength buffers. Samples were digested and run on the Qexactive. a) the number of proteins identified in replicate 1, and b) the total fmol of material calculated, c) proteins identified in replicate 2, and d) the total material, and e) proteins identified in replicate 3, and f) the total material.

increased at the higher ionic strength. However, the total amount of material identified is similar or higher in all three instances (figure 5.18), which may indicate that those proteins that are able to bind in high ionic conditions are doing so to a greater extent. If some proteins bind to a higher extent in the high ionic conditions, it may indicate that protein has a propensity for hydrophobic interactions, which can be enhanced in the presence of NaCl. There is, in the normalised material in each condition, an overlap between the three replicates, with the majority of proteins being identified in all three, but also some unique proteins in each (figure 5.19).

In all six replicates there is a wide dynamic range of proteins and under both conditions the most abundant normalised proteins are found in higher amounts than the strongest binding proteins in the starting material. However, in contrast to the amount of total material bound, under high ionic strength the most abundant proteins identified were less concentrated than in the low salt conditions (figure 5.20 a). Also, although there was a similar dynamic range identified, when the fmol amounts of each protein were plotted (figure 5.20 b, c, d) it showed a shallower incline in the low salt conditions, which indicates the proteins undergo more normalisation. The most abundant proteins in all three replicates of low salt normalised material were translation elongation factor 1 (TEF1), elongation factor 2 (EFT1) and aldolase 4 (ALD4). TEF1 shows an average increase of 1500% in low ionic conditions, and 700% in high ionic conditions, ALD4 1428 % in low ionic conditions, and 901% in high, and EFT1 1250% as opposed to 500% in higher ionic conditions (figure 5.21).

In each replicate there were some proteins that were only identified in either condition, however, the majority were found in both (figure 5.22 a, b, c), suggesting that the majority of protein binding has not been disrupted. Of the normalised proteins found in each replicate, 68 %, 61 % and 61 % were common to the low ionic buffer normalised material. Between the three replicates of normalised material 62.74% of the total proteins identified are common. This suggests that the variation between the two conditions is caused by the variability of protein normalisation with hexapeptide bead libraries.

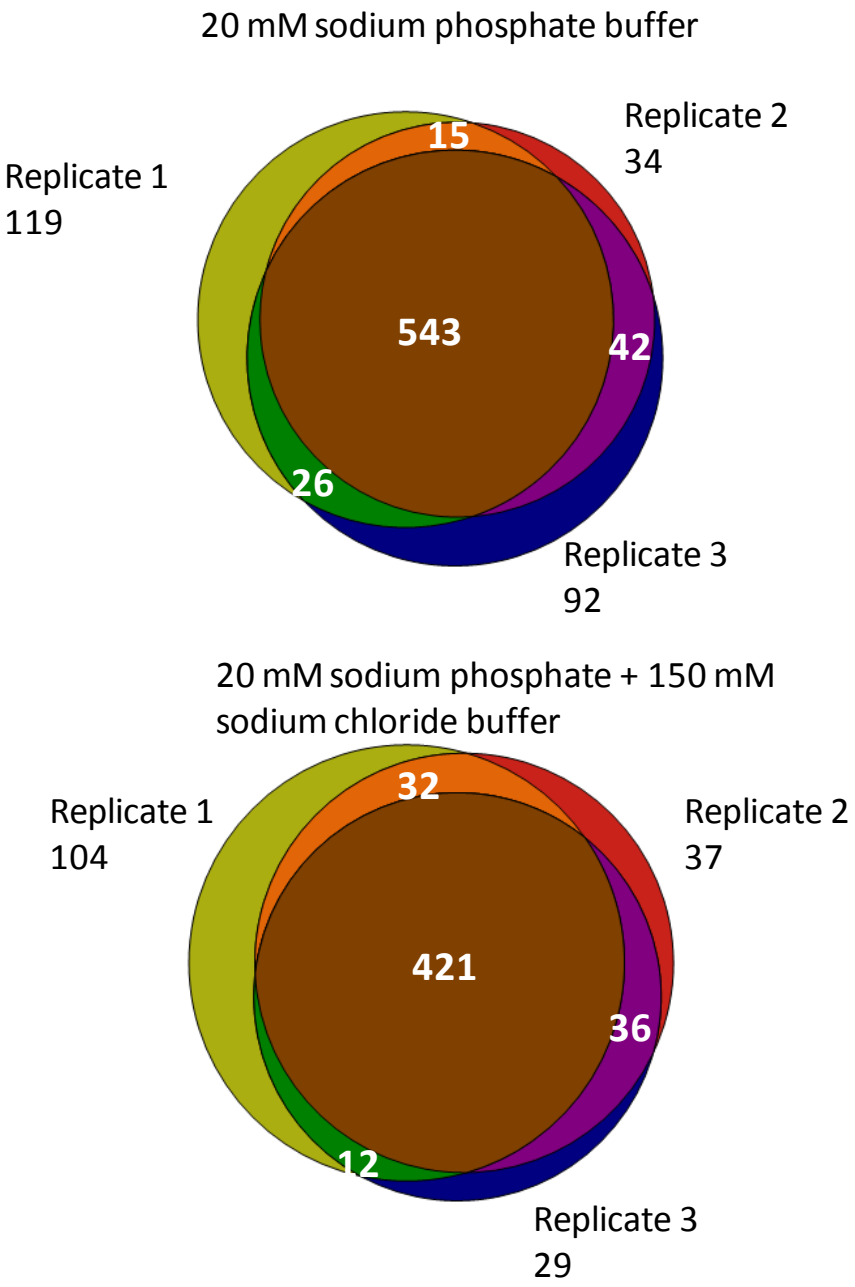


Figure 5.19 The common proteins identified in CPLL treated samples under two ionic strengths

Venn diagrams depicting the reproducibility of the protein hits identified in all three replicates in a) the low ionic strength (20 mM sodium phosphate buffer) samples, and b) the high ionic strength (150 mM sodium chloride). In each case, there is a high amount of crossover between the three samples, and there are unique proteins identified in the starting material in each replicate.

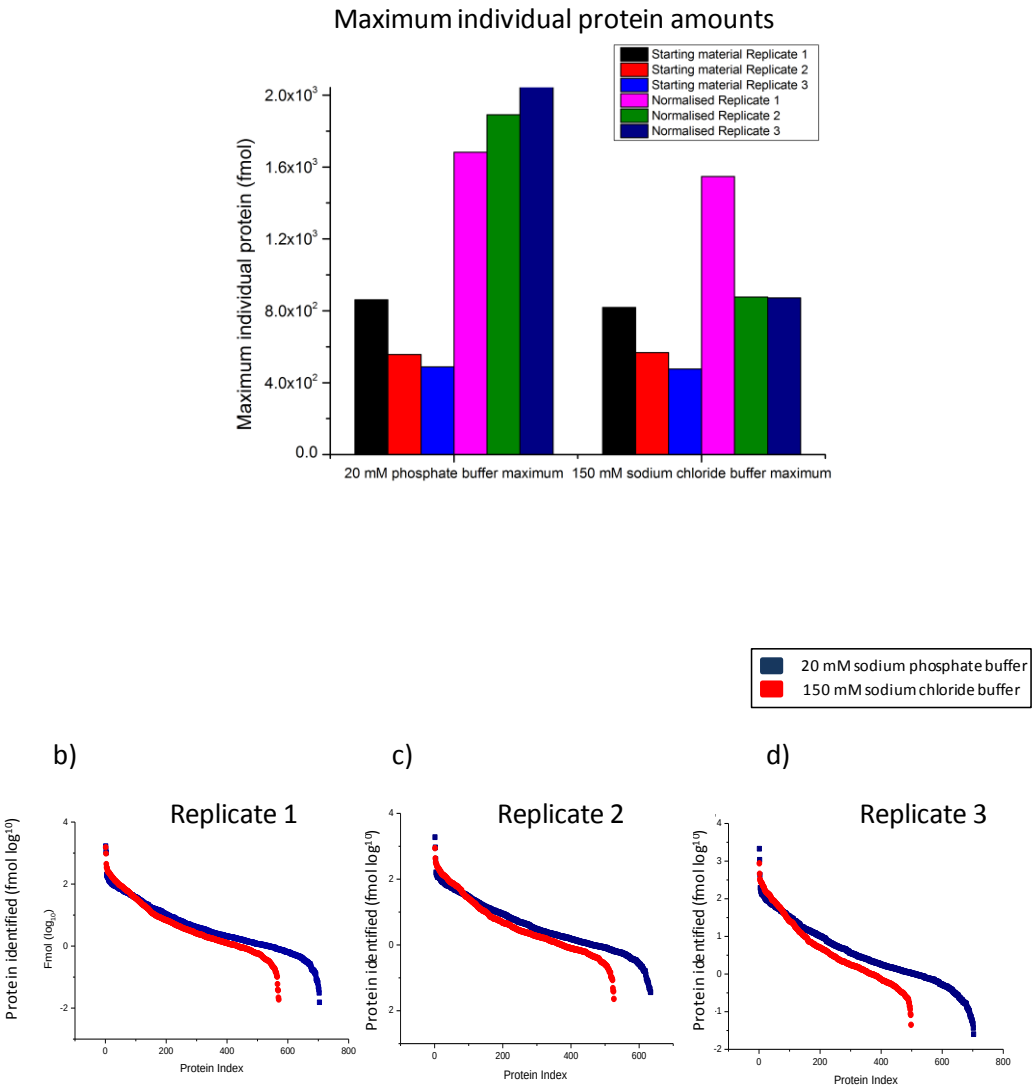
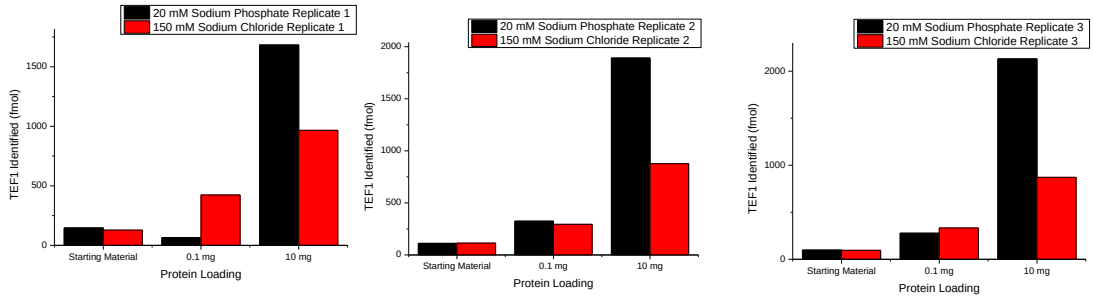


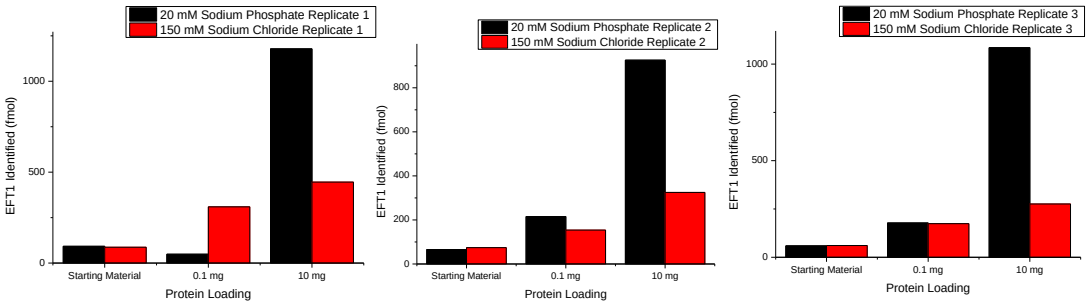
Figure 5.20 The quantification of each protein identified in starting material and normalised material at low and high ionic strength

a) the quantification of the most abundant proteins (in fmol) found in the cell lysate and the normalised material in both high and low ionic strength buffers, calculated using MaxQuant and b) the distribution of quantifications for all of the proteins identified in replicate 1, c) replicate 2, and d) replicate 3.

a) EFT2



b) EFT1



c) ALD4

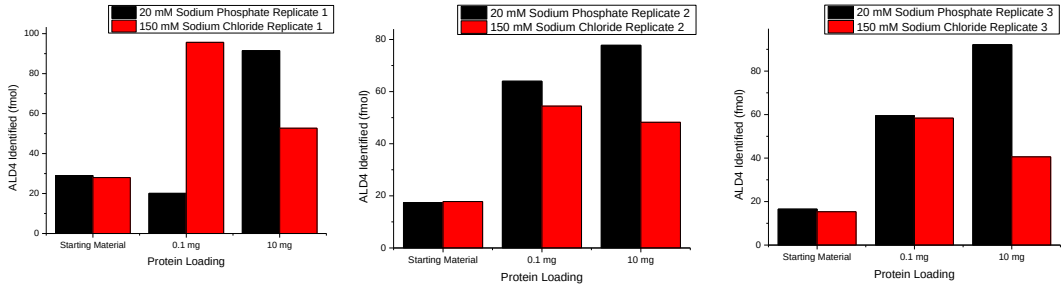


Figure 5.21. **The effect of ionic strength on the most abundant bead bound proteins**

The amount of each protein is calculated by label free quantification in MaxQuant, and the top three proteins found in the low ionic strength buffer are a) translation elongation factor 2, b) elongation factor 1 and c) aldolase 4, shown here under both ionic stength conditions.

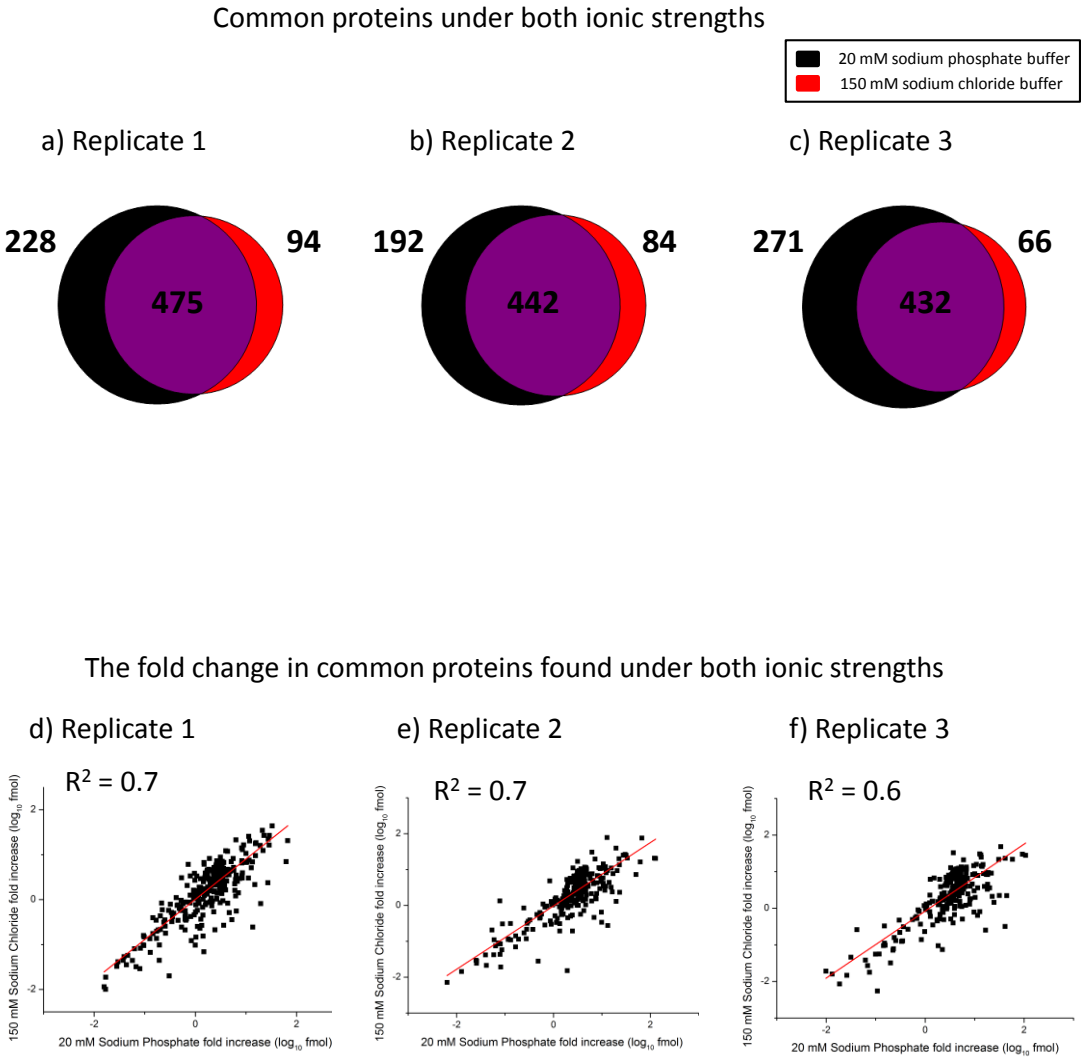


Figure 5.22 Proteins common to bead binding at two different ionic strengths

a) - c) Venn diagrams showing the common proteins identified between the normalised material in low and high ionic strength buffer, in a) replicate 1, b) replicate 2, c) replicate 3. d) - f) Scatter diagrams of the $\log_{10}$ fmol amounts of common proteins in each conditions plotted against each other in all three replicates.

A comparison of the behaviour of all of the proteins in both conditions shows that the majority exhibited the same behaviour under the two conditions, those that increase in concentration did so under both conditions, as shown by the linearity of the scatter plot of the fold change in each protein (figure 5.22 d, e, f). However, there are some outliers in these scatter plots, showing that there are indeed some proteins which exhibit different binding patterns under the two conditions. For example, alcohol dehydrogenase isozyme 3 (ADH3), peroxisomal AMP binding protein (PCS60), and vacuolar alpha monnosidase (AMS1) are all highly concentrated under low ionic conditions, but not higher ionic conditions (figure 5.23). This indicates that the proteins were interacting with the beads using electrostatic interactions at lower ionic conditions, and these have been disrupted by the addition of salt. On the other hand, proteins such as ribosomal subunits RPL5p, RPL13Bp and RPL10p are all more concentrated in higher ionic conditions than low. In fact, of approximately 50 ribosomal proteins identified in the six replicates, 98% of ribosomal proteins increase in concentration in higher ionic strength buffer in replicate 1, while in replicates 2 and 3 87 % and 80 % increase (figure 5.24). This is in comparison to 56 %, 59 % and 70 % of the proteins identified overall. An increase in the ability to bind the beads in higher ionic conditions may suggest these proteins are binding using hydrophobic interactions, which are encouraged in the presence of salt. On the other hand, some of these proteins could be binding as part of a sub-complex, or the propensity to form interactions may mean these particular proteins have more available interactions with ligands.

5.4 Conclusions

When considering the concept of the equaliser beads, a library of hexapeptide ligands, with a vast amount of variation, able to interact with proteins in a sample by any binding method already used by these proteins, it is unsurprising that variation is observed in the protein binding. It is this variability in structure, interaction, and behaviours which gives rise to the great number of proteins, and therefore, the complexity and variability observed in the living cell. There were three goals to the series of experiments presented here, accessing low abundance

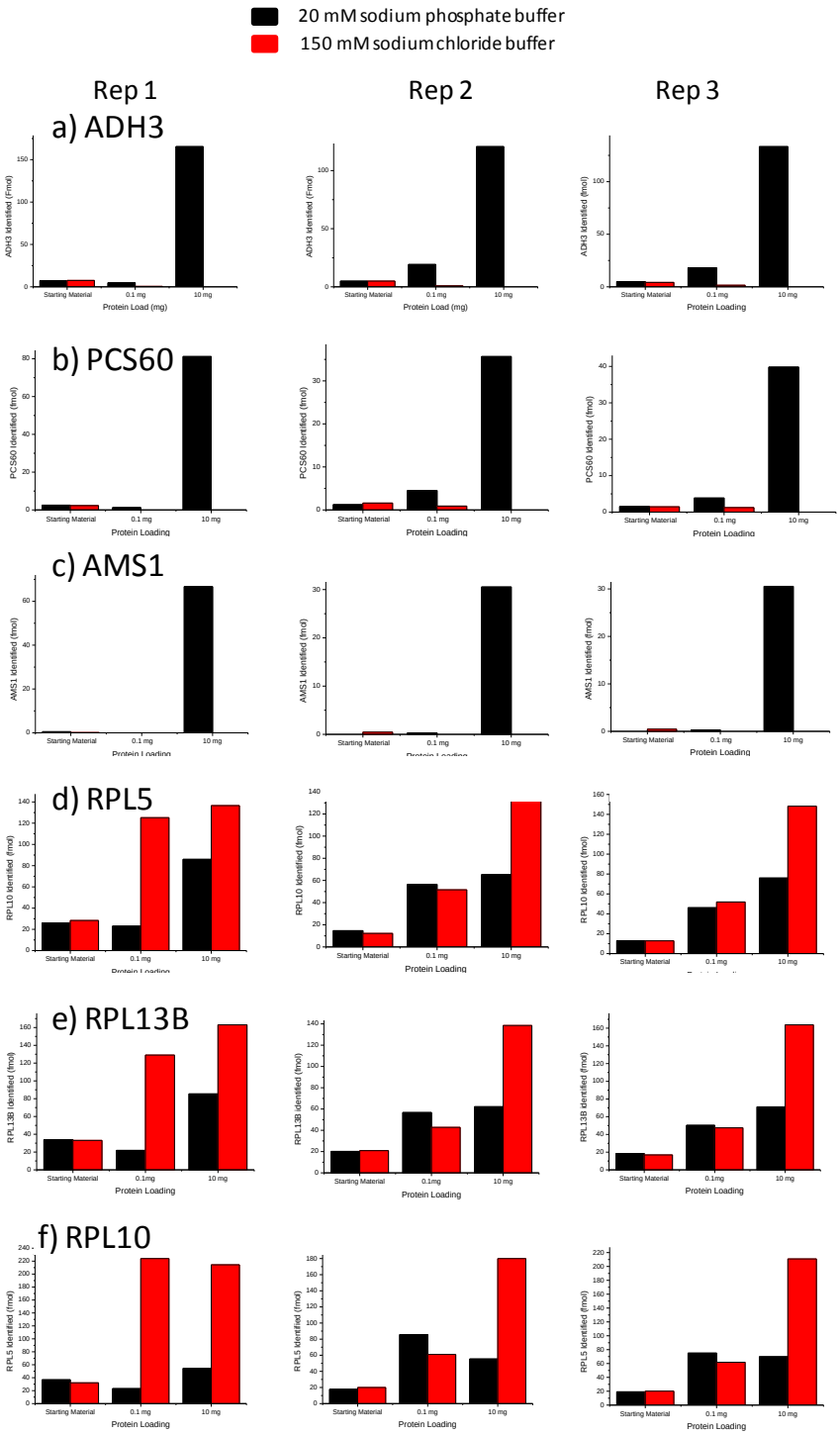


Figure 5.23 Examples of individual protein binding under two ionic strength conditions

Examples of protein quantifications calculated using Maxquant for a) alcohol dehydrogenase 3, b) a peroxisomal protein, and c) vacuolar alpha mannosidase, under low and high ionic conditions, also d) ribosomal 60s subunit L5, e) ribosomal 60s protein L13B, and f) ribosomal protein L10.

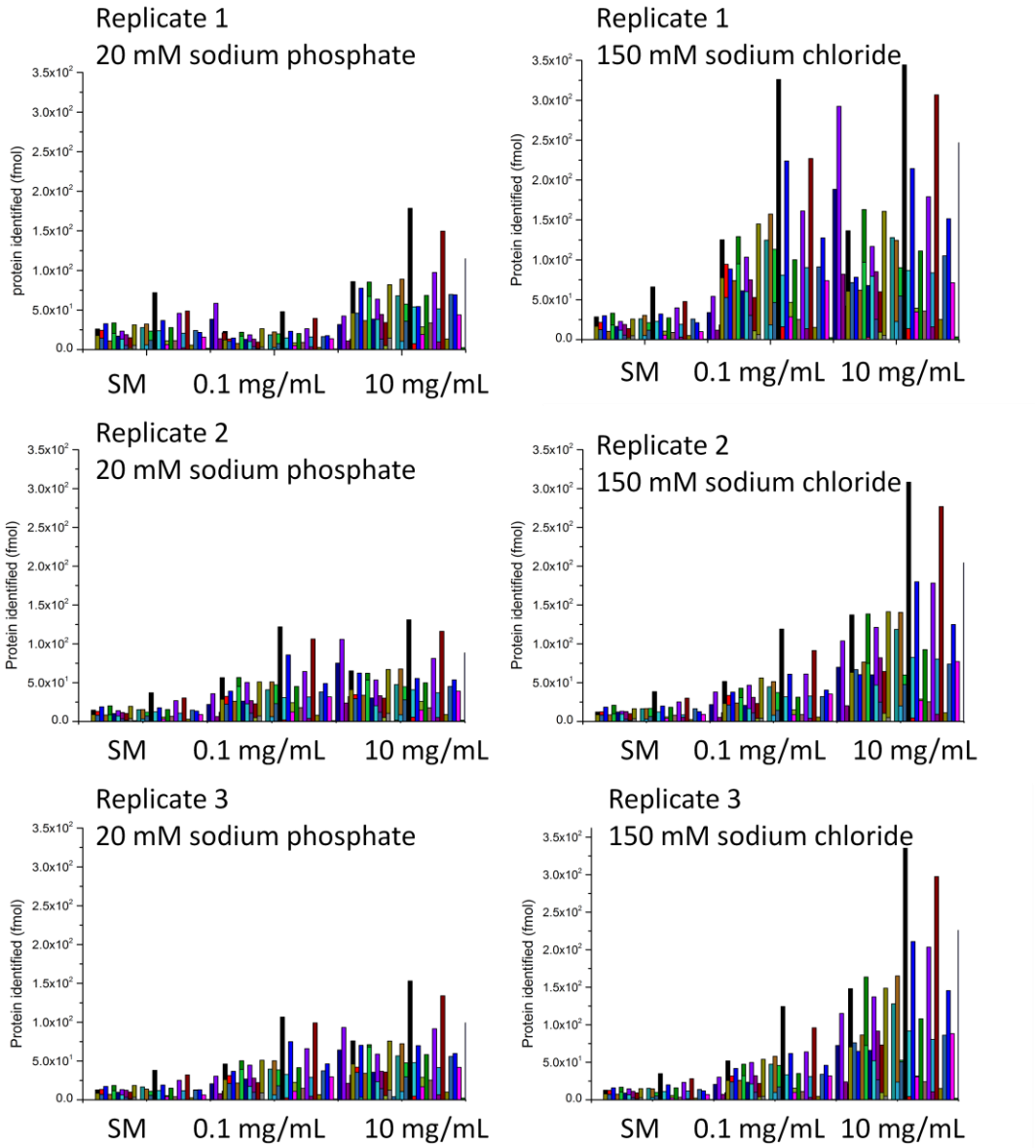


Figure 5.24 The binding of ribosome proteins in low and high ionic strength

50 proteins of the 60S ribosome subunit are identified across three replicates of starting material, 0.1 mg/mL and 10 mg/mL loads in both low and high ionic strength conditions, run on the QExactive and label free quantification performed using MaxQuant.

proteins, exploring the use of CPLL libraries in the quantification of proteins, and assessing the use of the technology for the analysis of protein complexes.

The results presented here indicate that equaliser bead technology has provided a degree of normalisation of yeast cell proteins, increasing the number of proteins identified in each experiment. It would be possible to increase the range of proteins identified further by including additional experiments, such as altering the pH of the binding conditions, or allowing the binding to occur for longer, to promote the binding of different proteins using a variety of protein interactions. Using a range of experimental conditions, the combinatorial bead library technology could make the entire proteome of yeast cell lysate observable in just a few experiments.

The results presented indicate the binding of proteins to the beads is not as simple as originally anticipated at the initial introduction of this technique. These results may show that the interaction between a protein and the beads are the result of proteins competing for binding sites, which they each have a higher or lower affinity for. Unfortunately, this then makes absolute quantification of less abundant proteins unlikely using this technology. If the protein-ligand interaction is the result of competition with other proteins for a few binding sites, it will not be possible to define the point at which a single bead reaches capacity, and to know whether a single protein has bound completely, or if some copies of the protein have been displaced from other peptide ligands they may interact with more weakly. This effect makes the protein binding unpredictable, so although the bead library is performing some degree of normalisation, it is not possible to use the beads in order to accurately quantify less abundant proteins.

Interactions between proteins and the bead library can be any of those normally found between proteins. Unique proteins are found in each case, highlighting the variable nature of protein binding. The variable binding of proteins under different ionic conditions suggests that a number of the interactions occurring are electrostatic or hydrophobic interactions.

In addition, it may be possible that the CPLL treatment is binding proteins that remain in complexes. It may, therefore, be possible to utilize the technology in the

investigation of protein complexes. In order to achieve this it would be necessary to prove the proteins are binding as a complex and not because the individual proteins have similar binding patterns. This could be achieved by denaturing the complex structure prior to CPLL treatment, however this may also affect protein structure, and the exposed surface area of the protein. Another method may be to produce individual recombinant proteins of all of the subunits from a complex, and spike these into a CPLL experiment. This would be a long and costly procedure, but if the recombinant proteins were heavy labelled, the binding could be compared to the original protein using LC-MS.

Chapter 6: General Discussion

6.1 Summary of approaches

The aim of this thesis was to explore quantitative proteomics methods for the analysis of the *S. cerevisiae* global proteome with respect to two specific areas, the analysis of protein complexes and accessing the less abundant proteins of the cellular environment. Quantitative proteomics techniques were applied to analyse the stoichiometry and composition of protein complexes and assist in building a complete picture of the protein interactions occurring within the cellular environment. A further challenge in the analysis of the proteome was addressed in the attempted quantification of the entire dynamic range of the proteome simultaneously using a CPLL method.

The mixed bed ion exchange chromatography method, when used in conjunction with label free methodologies, proved to be a successful for the identification of protein complexes. This was indicated by the high resolution co-elution of numerous proteins known to be part of complexes, as supported by literature. Although many protein assemblies that were identified separated at high resolution, such as the 20S proteasome or LSm complexes, some assemblies, such as the ribosomes, did not. Although there was an improvement in the resolution when mixed bed chromatography was used, in comparison to anion exchange chromatography, some proteins elute across a wider range of fractions, indicating they have a higher affinity for the chromatography medium or more surface area positions able to interact with the column. This could have also been an indication of the protein having binding partners, which were interacting with the column, and therefore the protein was retained for longer. Some individual proteins eluted across a range of fractions, for example the HSP70 SSA, which may indicate that it was forming a number of interactions. Other proteins have a wide elution profile as part of a complex, eluting across a wide range of fractions, such as phosphofructokinase or the ribosome subunits. The ribosome proteins did not elute as a complete assembly and there were a number of reasons explored for this behaviour, such as the dissociation of the assembly or the separation of partially

assembled sub-complexes. The dissociation of some weak interactions is a potential drawback to the use of ion exchange chromatography, for which a high concentration of NaCl is used for protein elution from the column. In the presence of these high ionic conditions, it is possible that some electrostatic interactions between members of protein complexes will be disrupted. Although other types of interaction will be maintained, which was evident by the successful separation of some assemblies, it is possible that more interactions could have been preserved with the use of a gentler method of chromatography, such as size exclusion.

In other examples individual proteins are observed in multiple locations in the gradient, sometimes in complex with other proteins, such as in the case of zuotin, or the nucleosome complex. When elution of proteins in multiple discrete peaks was observed a number of explanations were put forward, including possible post translational modification causing the same protein to have different affinity for the column. Alternatively, it could be an indication of the protein exhibiting interactions with other proteins, as in the case of zuotin, which appears to elute in two positions through the binding of two different proteins. In the example of the nucleosome, the protein complex elutes in two positions, indicating two possible charge variants. Therefore, by exploring these elution profiles it is possible to use the combination of ion exchange chromatography and bottom up proteomics to discover potential interactions exhibited by some proteins. Further work will be necessary to confirm the theories presented here behind the multiple peaks. The existence of multiple conformations could be confirmed using a second type of chromatography, such as SEC or HIC, or by electrophoresis such as BN-PAGE.

The large sets of data generated here provide an opportunity to explore the relationship between proteins using a bioinformatics strategy. A strategy could be developed to map the elution of each protein in the data tables, and compare them to each other. If strategies can be developed to examine these protein elution profiles new protein interactions can be discovered. These techniques could then also be used to map any alterations in protein complex composition under varying conditions, such as comparing native and nutrient starved conditions, or with the

presence of specific protein complex inhibitors. This would provide an efficient method of examining dynamic protein interactions.

The label free quantification strategies used, though applicable for the mapping of protein assembly composition, did not prove appropriate for the investigation of protein complex stoichiometry. This was demonstrated by the subunits of the exosome and 20S proteasome complexes, which are equimolar in stoichiometry, but are not calculated as such in the label-free quantification. It is possible that the amount of protein is miscalculated in the label free software due to the use of poorly ionising peptides, or perhaps well ionising peptides (in comparison to the standard peptide ionisation) giving an erroneous indication of the amount of protein present. This is supported by the results from two mass spectrometers (one QTOF, one Q-orbitrap), processed by different software, which yield similar results. Label free quantification strategies rely on the peptides present to infer the concentration of the protein. While the number of peptides varies for each protein in each replicate, PLGS processed data quantified using the top three most intense peptides and the MaxQuant processing used all peptides, the results obtained are comparable. This indicates that error in quantification was the result of the higher intensity peptides identified, and may mean that those proteins which are supposedly less concentrated have fewer proteotypic peptides, or the peptide could be undergoing some type of PTM. The inherent variability in the physicochemical properties of peptides are a disadvantage associated with label-free methods and label mediated strategies could offer an advantage in this respect, by using a standard peptide that is chemically identical to the analyte. The theory that label mediated quantification would provide a more accurate measure of complex subunit stoichiometry was therefore investigated using a label mediated strategy to assess the quantification of some protein complex subunits.

The design of QconCAT proteins for the quantification of protein complex subunits proved to be challenging. In order to maximise the prospect of accurate quantification results, it was necessary to choose optimal Qpeptides, applying a series of filters that resulted in a number of peptides being unusable. Despite the procedures followed in filtering unsuitable peptides, a number of the Qpeptides

included proved sub-optimal for quantification in these experiments. Some of these, such as those that were potentially deamidated, could be used in future quantification experiments if both the modified and unmodified forms are included in the analysis.

It was perhaps surprising that despite the advantages to using a label mediated method of quantification, there was little to no improvement observed in the analysis of protein complex stoichiometry. In some cases this was attributed to the use of poor peptides, and for some proteins the quantification relied on the result of just one peptide, when the aim was to use two. If two peptides were available for each protein the data may have been more reliable, and the results could therefore have been improved by using more Qpeptides, perhaps three per protein, to increase the likelihood of multiple peptides being observed. The accuracy of the experiment performed could have also been improved if a number of steps had been taken. The poor labelling of the QconCAT could have been eliminated given time to repeat the expression, and the quantification of the QconCAT itself could have been improved, for YEW1, if the sequence was redesigned to remove the miscleavege at the glufibrinopeptide sequence. However, the accurate quantification of the QconCAT would have improved the accuracy of each protein, though not relative to each other, so the stoichiometry would be unchanged. The accuracy of the QconCAT quantification could also have been improved if a co-digest of the QconCAT and the fraction had been performed. This is the recommended method, and would have been particularly advantageous here, as the Qpeptides were used in the original sequence context. Due to the design of the experiment, which included label free analysis to detect the location of protein complexes in the IEX fractions, a co-digest was not performed. However, this could have been achieved if the fraction was aliquoted into two, with one half used in label free analysis and the other available for label mediated quantification.

The conclusion of these experiments is therefore that the quantitative proteomics techniques used here are adequate for the analysis of protein complex composition, though not for the accurate analysis of subunit stoichiometry. There are a number of areas where there is scope to improve the QconCAT experiments

used here, which may allow accurate quantification, however the drawbacks of this technique could, in future, be avoided with the use of a PSAQ strategy. PSAQ has the advantage of allowing the study of all of the peptides from the parent protein in their native conditions. This may overcome the issues observed here of standard and analyte peptides performing differently under experimental conditions.

Although the accuracy of the results is questioned, the QconCAT mediated quantification does provide confirmation of the identification of proteins provided by the label-free data, for example in the case of the EIF2 complex subunit SUI3 the data confirms the presence of a subunit in a second location.

With regards to the second area of proteome analysis covered here, accessing low abundance proteins, there was an advantage to using CPLL technology for the reduction of the dynamic range. This is indicated by the additional proteins identified in *S. cerevisiae* cell lysate treated with beads. The loading experiments highlighted the dynamic range reduction, as some proteins increase in concentration, while others decreased. However, the actual binding to the beads appears to be more complicated, and it is possible that some competition for binding sites is occurring, as evident by proteins that increase in concentration in early loading, and decrease in later loadings, rather than the expected outcome of reaching a saturation point. The effect of the CPLL treatment was not as straightforward as expected, however, as rather than reach a saturation point, some of the proteins also increased in concentration far in excess of others. There are a number of reasons discussed for this, including additional affinity for the ligands, or interaction with other proteins already bound to the beads. The combination of proteins that are reduced in concentration after an initial increase in binding and the excessive increase in others means that it is not possible to calculate the saturation level of an individual protein. Being unable to reliably detect the saturation point of the proteins on the beads means it is not possible to use the CPLL technology in quantitative analysis of proteins.

Further analysis of the binding efficiency to the beads under different ionic conditions indicates that a number of proteins interact with the bead library using

electrostatic interactions. Despite this, some proteins are present in excessive amounts in the higher ionic conditions, indicating a propensity for hydrophobic interactions, either with the peptide ligands or other bead-bound proteins. This highlights the range of interactions possible between proteins and the CPLL, and unique proteins were identified in each condition. An additional level of analysis could have been performed in assessing the binding of *S. cerevisiae* proteins in the presence of varying pH buffers, which can extend the number of proteins identified by CPLL capture. This experiment would have provided an interesting addition to this thesis, in further investigating the contribution of electrostatic interactions in protein-ligand binding.

Although the CPLL experiments fall short of the original goal to analyse the entire dynamic range of the proteome simultaneously, they do provide a means to access an additional set of proteins that are not originally found in the starting material. The investigation of protein binding under different ionic conditions also allows the identification of different subsets of proteins. Combined with an additional series of experiments spanning a pH range, the CPLL technology may potentially provide a method of accessing the entire dynamic range of the *S. cerevisiae* proteome. However, this would be in a series of experiments, spanning a range of protein loadings and buffer conditions, adding time and complexity to the analysis.

6.2 Key conclusions

- Mixed bed IEX followed by LC-MS provides an effective method for studying protein interactions
- Employing a bioinformatics strategy for examining the results will provide an efficient method of monitoring protein interactions and changes in protein complex composition
- A QconCAT approach to study protein complex stoichiometry was not successful, and in this instance PSAQ may have been a more appropriate method
- CPLL libraries provide an effective approach for accessing low abundance proteins in the *S.cerevisiae* proteome.

- Given perfect digest conditions, it may be possible, but challenging, to use CPLL libraries quantifiably

6.3 Future perspectives

The overall aims of this thesis, the investigation of the application of quantitative proteomics strategies in the analysis of the proteome of *S. cerevisiae*, have been partially achieved. However with the method adjustments suggested they may be achievable. The combination of IEX, LC-MS and an SRM method provides an efficient workflow for the separation, identification and quantification of protein complexes. The texhnique has generated large data sets which, if examined using bioinformatics techniques, could give information on new protein relationships and provide a method of monitoring protein complexes under varying conditions. With some improvements to the QconCAT method used, absolute quantification of protein complex subunits could be possible. This would provide an efficient, high throughput method for studying protein interactions and assembling a complete picture of the proteome. With regards to accessing low abundance proteins using CPLL technology, it is evident that the protein-ligand binding is a complex procedure, involving a number of interaction types and efficiencies. This makes the estimation of protein saturation level, and therefore protein quantification, difficult to reliably estimate. However, the identification of the lower abundance subunits using CPLL technology could be combined with a targeted SRM approach for the later quantification of these subunits. However, improvements in mass spectrometry are reaching further into the proteome than ever, quantifying thousands of proteins in single experiments. It may therefore be possible to quantify an entire proteome without pre-fractionation in the near future.

Advani VM, Belew AT & Dinman JD. (2013). Yeast telomere maintenance is globally controlled by programmed ribosomal frameshifting and the nonsense-mediated mRNA decay pathway. *Translation (Austin)* **1,** e24418.

Al-Majdoub ZM, Carroll KM, Gaskell SJ & Barber J. (2014). Quantification of the Proteins of the Bacterial Ribosome Using QconCAT Technology. *J. Proteome Res*. **13**, 1211-22

Alberts B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92,** 291-294.

Anderson NL & Anderson NG. (2002). The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845-867.

Antonioli P, Fortis F, Guerrier L, Rinalducci S, Zolla L, Righetti PG & Boschetti E. (2007). Capturing and amplifying impurities from purified recombinant monoclonal antibodies via peptide library beads: a proteomic study. *Proteomics* **7,** 1624-1633.

Araiza-Olivera D, Chiquete-Felix N, Rosas-Lemus M, Sampedro JG, Pena A, Mujica A & Uribe-Carvajal S. (2013). A glycolytic metabolon in Saccharomyces cerevisiae is stabilized by F-actin. *FEBS J..* **280,** 3887-3905.

Araiza-Olivera D, Mexico IdFCUNAdMDoMGMC, Chiquete-Felix N, Mexico IdFCUNAdMDoMGMC, Rosas-Lemus M, Mexico IdFCUNAdMDoMGMC, Sampedro JG, Mexico UAdSLPIdF, Peña A, Mexico IdFCUNAdMDoMGMC, Mujica A, Mexico CdIyEAIPNDoCBMC, Uribe-Carvajal S & Mexico IdFCUNAdMDoMGMC. A glycolytic metabolon in Saccharomyces cerevisiae is stabilized by F-actin. *FEBS J.* **280,** 3887-3905.

Armache JP, Jarasch A, Anger AM, Villa E, Becker T, Bhushan S, Jossinet F, Habeck M, Dindar G, Franckenberg S, Marquez V, Mielke T, Thomm M, Berninghausen O, Beatrix B, Soding J, Westhof E, Wilson DN & Beckmann R. (2010). Localization of eukaryote-specific ribosomal proteins in a 5.5-A cryo-EM map of the 80S eukaryotic ribosome. *Proc. Natl. Acad. Sci. U S A* **107,** 19754-19759.

Armirotti A. (2009) Bottom up proteomics. *Curr. Anal. Chem.* **5**, 116-130

Ban N & Egelman EH. (2010). Structure and function of large cellular assemblies. *Curr. Opin. Struct. Biol.* **20,** 207-209.

Banaszak K, Mechin I, Obmolova G, Oldham M, Chang SH, Ruiz T, Radermacher M, Kopperschlager G & Rypniewski W. (2011). The crystal structures of eukaryotic phosphofructokinases from baker's yeast and rabbit skeletal muscle. *J. Mol. Biol.* **407,** 284-297.

Barr JR, Maggio VL, Patterson DG, Jr., Cooper GR, Henderson LO, Turner WE, Smith SJ, Hannon WH, Needham LL & Sampson EJ. (1996). Isotope dilution--mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I. *Clin. Chem.* **42,** 1676-1682.

Basquin J, Roudko VV, Rode M, Basquin C, Seraphin B & Conti E. (2012). Architecture of the nuclease module of the yeast Ccr4-not complex: the Not1-Caf1-Ccr4 interaction. *Mol. Cell.* **48,** 207-218.

Beck F, Unverdorben P, Bohn S, Schweitzer A, Pfeifer G, Sakata E, Nickell S, Plitzko JM, Villa E, Baumeister W & Förster F. (2012). Near-atomic resolution structural model of the yeast 26S proteasome. *Proc. Natl. Acad. Sci. U S A* **109,** 14870-14875.

Belew AT, Advani VM & Dinman JD. (2011). Endogenous ribosomal frameshift signals operate as mRNA destabilizing elements through at least two molecular pathways in yeast. *Nucleic Acids Res.* **39,** 2799-2808.

Belle A, Tanay A, Bitincka L, Shamir R & O'Shea EK. (2006). Quantification of protein half-lives in the budding yeast proteome. *PNAS*. **103**, 13004-13009.

Ben-Shem A, Jenner L, Yusupova G & Yusupov M. (2010). Crystal structure of the eukaryotic ribosome. *Science* **330,** 1203-1209.

Bertonati C, Honig B & Alexov E. (2007). Poisson-Boltzmann Calculations of Nonspecific Salt Effects on Protein-Protein Binding Free Energies. *Biophysical journal* **92,** 1891-1899.

Beseme O, Fertin M, Drobecq H, Amouyel P & Pinet F. (2010). Combinatorial peptide ligand library plasma treatment: Advantages for accessing low-abundance proteins. *Electrophoresis* **31,** 2697-2704.

Beynon RJ, Doherty MK, Pratt JM & Gaskell SJ. (2005). Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2,** 587-589.

Bislev SL, Kusebauch U, Codrea MC, Beynon RJ, Harman VM, Rontved CM, Aebersold R, Moritz RL & Bendixen E. (2012). Quantotypic properties of QconCAT peptides targeting bovine host response to Streptococcus uberis. *J. Proteome Res.* **11,** 1832-1843.

Bjorhall K, Miliotis T & Davidsson P. (2005). Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* **5,** 307-317.

Boehm AM, Putz S, Altenhofer D, Sickmann A & Falk M. (2007). Precise protein quantification based on peptide quantification using iTRAQ. *BMC Bioinformatics* **8,** 214.

Boschetti E, Lomas L, Citterio A & Righetti PG. (2007). Romancing the "hidden proteome", Anno Domini two zero zero seven. *J. Chromatogr. A.* **1153,** 277-290.

Boschetti E & Righetti PG. (2008). The ProteoMiner in the proteomic arena: a non-depleting tool for discovering low-abundance species. *J. Proteomics* **71,** 255-264.

Botstein D, Chervitz SA & Cherry JM. (1997). Yeast as a model organism. *Science* **277,** 1259-1260.

Botstein D & Fink GR. (2011). Yeast: an experimental organism for 21st Century biology. *Genetics* **189,** 695-704.

Bouveret E, Rigaut G, Shevchenko A, Wilm M & Seraphin B. (2000). A Sm-like protein complex that participates in mRNA degradation. *EMBO J.* **19,** 1661-1671.

Brownridge P & Beynon RJ. (2011). The importance of the digest: proteolysis and absolute quantification in proteomics. *Methods* **54,** 351-360.

Brownridge P, Lawless C, Payapilly AB, Lanthaler K, Holman SW, Harman VM, Grant CM, Beynon RJ & Hubbard SJ. (2013). Quantitative analysis of chaperone network throughput in budding yeast. *Proteomics* **13,** 1276-1291.

Brun V, Masselon C, Garin J & Dupuis A. (2009). Isotope dilution strategies for absolute quantitative proteomics. *J. Proteomics* **72,** 740-749.

Buhrman DL, Price PI & Rudewiczcor PJ. (1996). Quantitation of SR 27417 in human plasma using electrospray liquid chromatography-tandem mass spectrometry: A study of ion suppression. *J. Am. So.c Mass Spectrom.* **7,** 1099-1105.

Calvete JJ, Fasoli E, Sanz L, Boschetti E & Righetti PG. (2009). Exploring the venom proteome of the western diamondback rattlesnake, Crotalus atrox, via snake venomics and combinatorial peptide ligand library approaches. *J. Proteome Res.* **8,** 3055-3067.

Camacho-Carvajal MM, Wollscheid B, Aebersold R, Steimle V & Schamel WW. (2004). Two-dimensional Blue native/SDS gel electrophoresis of multi-protein complexes from whole cellular lysates: a proteomics approach. *Mol. Cell. Proteomics.* **3,** 176-182.

Carroll KM, Simpson DM, Eyers CE, Knight CG, Brownridge P, Dunn WB, Winder CL, Lanthaler K, Pir P, Malys N, Kell DB, Oliver SG, Gaskell SJ & Beynon RJ. (2011). Absolute quantification of the glycolytic pathway in yeast: deployment of a complete QconCAT approach. *Mol. Cell. Proteomics.* **10,** M111.007633.

Castagna A, Cecconi D, Sennels L, Rappsilber J, Guerrier L, Fortis F, Boschetti E, Lomas L & Righetti PG. (2005). Exploring the hidden human urinary proteome via ligand library beads. *J. Proteome Res.* **4,** 1917-1930.

Castrillo JI & Oliver SG. (2011). Yeast systems biology: the challenge of eukaryotic complexity. *Methods Mol. Biol.* **759,** 3-28.

Chabes A, Domkin V, Larsson G, Liu A, Graslund A, Wijmenga S & Thelander L. (2000). Yeast ribonucleotide reductase has a heterodimeric iron-radical-containing subunit. *Proc. Natl. Acad. Sci. U S A* **97,** 2474-2479.

Chambers E, Wagrowski-Diehl DM, Lu Z & Mazzeo JR. (2007). Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* **852,** 22-34.

Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K & Tyers M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* **41,** D816-823.

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L & Cesareni G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35,** D572-574.

Chelius D, Rehder DS & Bondarenko PV. (2005). Identification and characterization of deamidation sites in the conserved regions of human immunoglobulin gamma antibodies. *Anal. Chem.* **77,** 6004-6011.

Chen CS & Zhu H. (2006). Protein microarrays. *Biotechniques* **40,** 423-427.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S & Wong ED. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700-705.

Cilindre C, Fasoli E, D'Amato A, Liger-Belair G & Righetti PG. (2014). It's time to pop a cork on Champagne's proteome! *J. Proteomics*. **13,** 105-351

Clancy T & Hovig E. (2014). From proteomes to complexomes in the era of systems biology. *Proteomics* **14,** 24-41.

Collins MO & Choudhary JS. (2008). Mapping multiprotein complexes by affinity purification and mass spectrometry. *Curr. Opin. Biotechnol.* **19,** 324-330.

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS & Krogan NJ. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **6,** 439-450.

Cornell M, Paton NW & Oliver SG. (2004). A critical and integrated view of the yeast interactome. *Comp. Funct. Genomics* **5,** 382-402.

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,** 1367-1372.

Craig R, Cortens JP & Beavis RC. (2005). The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **19,** 1844-1850.

D'Amato A, Fasoli E & Righetti PG. (2012). Harry Belafonte and the secret proteome of coconut milk. *J. Proteomics* **75,** 914-920.

D'Ambrosio C, Arena S, Scaloni A, Guerrier L, Boschetti E, Mendieta ME, Citterio A & Righetti PG. (2008). Exploring the chicken egg white proteome with combinatorial peptide ligand libraries. *J. Proteome Res.* **7,** 3461-3474.

D'Amici GM, Timperio AM, Rinalducci S & Zolla L. (2012). Combinatorial peptide ligand libraries to discover liver disease biomarkers in plasma samples. *Methods Mol. Biol.* **909,** 311-319.

de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC & Mann M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251-1254.

Decker BL & Wickner WT. (2006). Enolase activates homotypic vacuole fusion and protein transport to the vacuole in yeast. *J. Biol. Chem.* **281,** 14523-14528.

Deutsch EW, Lam H & Aebersold R. (2008). PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9,** 429-434.

Di Girolamo F, Righetti PG, Soste M, Feng Y & Picotti P. (2013). Reproducibility of combinatorial peptide ligand libraries for proteome capture evaluated by selected reaction monitoring. *J. Proteomics* **89,** 215-226.

Dupuis A, Hennekinne JA, Garin J & Brun V. (2008). Protein Standard Absolute Quantification (PSAQ) for improved investigation of staphylococcal food poisoning outbreaks. *Proteomics* **8,** 4633-4636.

El Alami M, Messenguy F, Scherens B & Dubois E. (2003). Arg82p is a bifunctional protein whose inositol polyphosphate kinase activity is essential for nitrogen and PHO gene expression but not for Mcm1p chaperoning in yeast. *Mol. Microbiol.* **49,** 457-468.

el Rassi Z & Horvath C. (1986). Tandem columns and mixed-bed columns in high-performance liquid chromatography of proteins. *J. Chromatogr.* **359,** 255-264.

Emre NC & Berger SL. (2006). Histone post-translational modifications regulate transcription and silent chromatin in Saccharomyces cerevisiae. *Ernst Schering Res Found Workshop***, 57,** 127-153.

Eng JK, Washington Uo, McCormack AL, Washington Uo, Yates JR & Washington Uo. (2014). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976-989.

Eyers CE, Lawless C, Wedge DC, Lau KW, Gaskell SJ & Hubbard SJ. (2011). CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol. Cell. Proteomics* **10,** M110.003384.

Fasoli E, D'Amato A, Citterio A & Righetti PG. (2011a). Ginger Rogers? No, Ginger Ale and its invisible proteome. *J. Proteomics*.**75**,1960-5

Fasoli E, D'Amato A, Kravchuk AV, Boschetti E, Bachi A & Righetti PG. (2011b). Popeye strikes again: The deep proteome of spinach leaves. *J. Proteomics* **74,** 127-136.

Fasoli E, D'Amato A, Righetti PG, Barbieri R & Bellavia D. (2012). Exploration of the sea urchin coelomic fluid via combinatorial peptide ligand libraries. *Biol. Bull.* **222,** 93-104.

Fasoli E, Farinazzo A, Sun CJ, Kravchuk AV, Guerrier L, Fortis F, Boschetti E & Righetti PG. (2010). Interaction among proteins and peptide libraries in proteome analysis: pH involvement for a larger capture of species. *J. Proteomics* **73,** 733-742.

Fasoli E & Righetti PG. (2013). The peel and pulp of mango fruit: a proteomic samba. *Biochim. Biophys. Acta* **1834,** 2539-2545.

Fatima N, Chelius D, Luke BT, Yi M, Zhang T, Stauffer S, Stephens R, Lynch P, Miller K, Guszczynski T, Boring D, Greenwald P & Ali IU. (2009). Label-free global serum proteomic profiling reveals novel celecoxib-modulated proteins in familial adenomatous polyposis patients. *Cancer Genomics Proteomics* **6,** 41-49.

Fenn JB, Mann M, Meng CK, Wong SF & Whitehouse CM. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64-71.

Fields S. (2001). Proteomics. Proteomics in genomeland. *Science* **291,** 1221-1224.

Fields S & Song O-k. (1989). A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245-246.

Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL & Washburn MP. (2006). Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40,** 303-311.

Flores CL & Gancedo C. (2011). Unraveling moonlighting functions with yeasts. *IUBMB Life* **63,** 457-462.

Fortis F, Guerrier L, Areces LB, Antonioli P, Hayes T, Carrick K, Hammond D, Boschetti E & Righetti PG. (2006). A new approach for the detection and identification of protein impurities using combinatorial solid phase ligand libraries. *J. Proteome Res.* **5,** 2577-2585.

Fu Q & Van Eyk JE. (2006). Proteomics and heart disease: identifying biomarkers of clinical utility. *Expert Rev. Proteomics* **3,** 237-249.

Furka A, Sebestyen F, Asgedom M & Dibo G. (1991). General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Pept. Protein Res.* **37,** 487-493.

Gallien S, Duriez E & Domon B. (2011). Selected reaction monitoring applied to proteomics. *J. Mass Spectrom.* **46,** 298-312.

Gancedo C & Flores CL. (2008). Moonlighting proteins in yeasts. *Microbiol. Mol. Biol. Rev.* **72,** 197-210.

Gao Q, Madian AG, Liu X, Adamec J & Regnier FE. (2010). Coupling protein complex analysis to peptide based proteomics. *J. Chromatogr. A* **1217,** 7661-7668.

Gautschi M, Lilie H, Funfschilling U, Mun A, Ross S, Lithgow T, Rucknagel P & Rospert S. (2001). RAC, a stable ribosome-associated complex in yeast formed by the DnaK-DnaJ homologs Ssz1p and zuotin. *Proc. Natl. Acad. Sci. U S A* **98,** 3762-3767.

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB & Superti-Furga G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631-636.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G & Superti-Furga G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141-147.

Geiger T, Wisniewski JR, Cox J, Zanivan S, Kruger M, Ishihama Y & Mann M. (2011). Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nat. Protoc.* **6,** 147-157.

Gelis S, Curto M, Valledor L, Gonzalez A, Arino J, Jorrin J & Ramos J. (2012). Adaptation to potassium starvation of wild-type and K+-transport mutant (trk1,2) of Saccharomyces cerevisiae: 2-dimensional gel electrophoresis-based proteomic approach. *Microbiologyopen* **1,** 182-193.

Gerber SA, Rush J, Stemman O, Kirschner MW & Gygi SP. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U S A* **100,** 6940-6945.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK & Weissman JS. (2003). Global analysis of protein expression in yeast. *Nature* **425,** 737-741.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H & Oliver SG. (1996). Life with 6000 genes. *Science* **274,** 546, 563-547.

Goncalves A, Charafe-Jauffret E, Bertucci F, Audebert S, Toiron Y, Esterni B, Monville F, Tarpin C, Jacquemier J, Houvenaeghel G, Chabannon C, Extra JM, Viens P, Borg JP & Birnbaum D. (2008). Protein profiling of human breast tumor cells identifies novel biomarkers associated with molecular subtypes. *Mol. Cell. Proteomics* **7,** 1420-1433.

Gonzalez-Galarza FF, Lawless C, Hubbard SJ, Fan J, Bessant C, Hermjakob H & Jones AR. (2012). A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. *Omics* **16,** 431-442.

Gosetti F, Mazzucco E, Zampieri D & Gennaro MC. (2010). Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry. *J. Chromatogr. A* **1217,** 3929-3937.

Grandier-Vazeille X & Guerin M. (1996). Separation by blue native and colorless native polyacrylamide gel electrophoresis of the oxidative phosphorylation complexes of yeast mitochondria solubilized by different detergents: specific staining of the different complexes. *Anal. Biochem.* **242,** 248-254.

Groll M, Ditzel L, Lowe J, Stock D, Bochtler M, Bartunik HD & Huber R. (1997). Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature* **386,** 463-471.

Gross JD, Moerke NJ, von der Haar T, Lugovskoy AA, Sachs AB, McCarthy JE & Wagner G. (2003). Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell* **115,** 739-750.

Guerrero C, Tagwerker C, Kaiser P & Huang L. (2006). An integrated mass spectrometry-based proteomic approach: quantitative analysis of tandem affinity-purified in vivo cross-linked protein complexes (QTAX) to decipher the 26 S proteasome-interacting network. *Mol. Cell. Proteomics* **5,** 366-378.

Guerrier L, Claverol S, Fortis F, Rinalducci S, Timperio AM, Antonioli P, Jandrot-Perrus M, Boschetti E & Righetti PG. (2007). Exploring the platelet proteome via combinatorial, hexapeptide ligand libraries. *J. Proteome Res.* **6,** 4290-4303.

Guerrier L, Thulasiraman V, Castagna A, Fortis F, Lin S, Lomas L, Righetti PG & Boschetti E. (2006). Reducing protein concentration range of biological samples using solid-phase ligand libraries. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **833,** 33-40.

Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW & Stumpflen V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* **34,** D436-441.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH & Aebersold R. (1999a). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994-999.

Gygi SP, Rochon Y, Franza BR & Aebersold R. (1999b). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720-1730.

Halbeisen RE & Gerber AP. (2009). Stress-dependent coordination of transcriptome and translatome in yeast. *PLoS Biol.* **7,** e1000105.

Hale JE, Butler JP, Knierman MD & Becker GW. (2000). Increased sensitivity of tryptic peptide detection by MALDI-TOF mass spectrometry is achieved by conversion of lysine to homoarginine. *Anal. Biochem.* **287,** 110-117.

Han X, Aslanian A, Yates JR (III). (2008). Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* 12(5) 483-490

Hart GT, Ramani AK & Marcotte EM. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7,** 120.

Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS & Coon JJ. (2014). The one hour yeast proteome. *Mol. Cell. Proteomics* **13,** 339-347.

Heck AJ. (2008). Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods* **5,** 927-933.

Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C & Watanabe C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U S A* **90,** 5011-5015.

Hesselberth JR, Miller JP, Golob A, Stajich JE, Michaud GA & Fields S. (2006). Comparative analysis of Saccharomyces cerevisiae WW domains and their interacting proteins. *Genome Biol.* **7,** R30.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D & Tyers M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415,** 180-183.

Hochleitner EO, Kastner B, Frohlich T, Schmidt A, Luhrmann R, Arnold G & Lottspeich F. (2005). Protein stoichiometry of a multiprotein complex, the human spliceosomal

U1 small nuclear ribonucleoprotein: absolute quantification using isotope-coded tags and mass spectrometry. *J. Biol. Chem.* **280,** 2536-2542.

Hortin GL & Sviridov D. (2010). The dynamic range problem in the analysis of the plasma proteome. *J. Proteomics* **73,** 629-636.

Huang JT, McKenna T, Hughes C, Leweke FM, Schwarz E & Bahn S. (2007). CSF biomarker discovery using label-free nano-LC-MS based proteomic profiling: technical aspects. *J. Sep. Sci.* **30,** 214-225.

Huang PY, Baumbach GA, Dadd CA, Buettner JA, Masecar BL, Hentsch M, Hammond DJ & Carbonell RG. (1996). Affinity purification of von Willebrand factor using ligands derived from peptide libraries. *Bioorg. Med. Chem.* **4,** 699-708.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS & O'Shea EK. (2003). Global analysis of protein localization in budding yeast. *Nature* **425,** 686-691.

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J & Mann M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265-1272.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M & Sakaki Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U S A* **98,** 4569-4574.

J. Throck Watson ODS. (2008). Introduction to Mass Spectrometry: Instrumentation, Applications and Strategies for Data Interpretation.*Wiley, Fourth Edition*.

James P, Quadroni M, Carafoli E & Gonnet G. (1993). Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195,** 58-64.

Jeong H, Mason SP, Barabasi AL & Oltvai ZN. (2001). Lethality and centrality in protein networks. *Nature* **411,** 41-42.

Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H & Apweiler R. (2006). PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* **34,** D659-663.

Juneau K, Nislow C & Davis RW. (2009). Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. *Genetics* **183,** 185-194.

Jungblut P, Thiede B, Zimny-Arndt U, Muller EC, Scheler C, Wittmann-Liebold B & Otto A. (1996). Resolution power of two-dimensional electrophoresis and identification of proteins from gels. *Electrophoresis* **17,** 839-847.

Jungblut PR, Holzhutter HG, Apweiler R & Schluter H. (2008). The speciation of the proteome. *Chem. Cent. .J* **2,** 16.

Kaboord B & Perr M. (2008). Isolation of proteins and protein complexes by immunoprecipitation. *Methods Mol. Biol.* **424,** 349-364.

Kaiser SE, Riley BE, Shaler TA, Trevino RS, Becker CH, Schulman H & Kopito RR. Protein standard absolute quantification (PSAQ) method for the measurement of cellular ubiquitin pools. *Nat. Methods* **8,** 691-696.

Karas M & Hillenkamp F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299-2301.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S & Hermjakob H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40,** D841-846.

Kito K & Ito T. (2008). Mass spectrometry-based approaches toward absolute quantitative proteomics. *Curr. Genomics* **9,** 263-274.

Kito K, Kawaguchi N, Okada S & Ito T. (2008). Discrimination between stable and dynamic components of protein complexes by means of quantitative proteomics. *Proteomics* **8,** 2366-2370.

Kito K, Ota K, Fujita T & Ito T. (2007). A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes. *J. Proteome Res.* **6,** 792-800.

Kresze GB & Ronft H. (1981). Pyruvate dehydrogenase complex from baker's yeast. 2. Molecular structure, dissociation, and implications for the origin of mitochondria. *Eur. J. Biochem.* **119,** 581-587.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles

CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A & Greenblatt JF. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440,** 637-643.

Kuster B, Schirle M, Mallick P & Aebersold R. (2005). Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6,** 577-583.

Lam KS, Salmon SE, Hersh EM, Hruby VJ, Kazmierski WM & Knapp RJ. (1991). A new type of synthetic peptide library for identifying ligand-binding activity. *Nature* **354,** 82-84.

Law YP & Lim YP. (2013). Recent advances in mass spectrometry: data independent analyses and hyper reaction monitoring. *Exp. Rev. Proteomics* **10(6)**, 551-66

Lawless C & Hubbard SJ. (2012). Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. *Omics* **16,** 449-456.

Le A, Barton LD, Sanders JT & Zhang Q. (2011). Exploration of bovine milk proteome in colostral and mature whey using an ion-exchange approach. *J. Proteome Res.* **10,** 692-704.

Li SS & Giometti CS. (2007). A combinatorial approach to studying protein complex composition by employing size-exclusion chromatography and proteome analysis. *J. Sep. Sci.* **30,** 1549-1555.

Liu C, Zhang N, Yu H, Chen Y, Liang Y, Deng H & Zhang Z. (2011). Proteomic analysis of human serum for finding pathogenic factors and potential biomarkers in preeclampsia. *Placenta* **32,** 168-174.

Liu H, Sadygov RG & Yates JR, 3rd. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193-4201.

Liu X, Yang W-c, Gao Q & Regnier F. (2008). Toward chromatographic analysis of interacting protein networks. *J. Chromatogr. A* **1178,** 24-32.

Loo JA, Berhane B, Kaddis CS, Wooding KM, Xie Y, Kaufman SL & Chernushevich IV. (2005). Electrospray ionization mass spectrometry and ion mobility analysis of the 20S proteasome complex. *J. Am. Soc. Mass Spectrom.* **16,** 998-1008.

Loo JA, Udseth HR & Smith RD. (1989). Peptide and protein analysis by electrospray ionization-mass spectrometry and capillary electrophoresis-mass spectrometry. *Anal. Biochem.* **179,** 404-412.

Lorenzen K & van Duijn E. (2010). Native mass spectrometry as a tool in structural biology. *Curr. Protoc. Protein Sci.* **Chapter 17,** Unit17.12.

Lorenzen K, Vannini A, Cramer P & Heck AJ. (2007). Structural biology of RNA polymerase III: mass spectrometry elucidates subcomplex architecture. *Structure* **15,** 1237-1245.

Lu P, Vogel C, Wang R, Yao X & Marcotte EM. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25,** 117-124.

Ma B & Johnson R. (2012). De Novo Sequencing and Homology Searching. *Mol. Cell. Proteomics* **11**.

MacBeath G & Schreiber SL. (2000). Printing proteins as microarrays for high-throughput function determination. *Science* **289,** 1760-1763.

Makino DL, Baumgartner M & Conti E. (2013a). Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex. *Nature* **495,** 70-75.

Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B & Aebersold R. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotech.* **25,** 125-131.

Mann M. (2008). Can proteomics retire the western blot? *J. Proteome Res.* **7,** 3065.

Marshall AG, Hendrickson CL & Jackson GS. (1998). Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **17,** 1-35.

Matuszewski BK, Constanzer ML & Chavez-Eng CM. (2003). Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Anal. Chem.* **75,** 3019-3030.

Meng R, Gormley M, Bhat VB, Rosenberg A & Quong AA. (2011). Low abundance protein enrichment for discovery of candidate plasma protein biomarkers for early detection of breast cancer. *J. Proteomics* **75,** 366-374.

Merrifield RB. (1965). Automated synthesis of peptides. *Science* **150,** 178-185.

Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M & Horning S. (2011). Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** M111.011015.

Mitchell P. (2010). Proteomics retrenches. *Nat. Biotechnol.* **28,** 665-670.

Molina M, Cid VJ & Martin H. (2010). Fine regulation of Saccharomyces cerevisiae MAPK pathways by post-translational modifications. *Yeast* **27,** 503-511.

Monari E, Casali C, Cuoghi A, Nesci J, Bellei E, Bergamini S, Fantoni LI, Natali P, Morandi U & Tomasi A. (2011). Enriched sera protein profiling for detection of non-small cell lung cancer biomarkers. *Proteome Sci.* **9,** 55.

Morris HR, Paxton T, Dell A, Langhorne J, Berg M, Bordoli RS, Hoyes J & Bateman RH. (1996). High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **10,** 889-896.

Moser AC & Hage DS. (2010). Immunoaffinity chromatography: an introduction to applications and recent developments. *Bioanalysis* **2,** 769-790.

Mouton-Barbosa E, Roux-Dalvai F, Bouyssie D, Berger F, Schmidt E, Righetti PG, Guerrier L, Boschetti E, Burlet-Schiltz O, Monsarrat B & Gonzalez de Peredo A. (2010). In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. *Mol. Cell. Proteomics* **9,** 1006-1021.

Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, Vorm O & Mann M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111.013722.

Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K & Darie CC. (2014). Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. Life Sci.* **71,** 205-228.

Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Loo JA & Darie CC. (2013). Investigation of stable and transient protein-protein interactions: Past, present, and future. *Proteomics* **13,** 538-557.

O'Farrell PH. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007-4021.

Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA & Ahn NG. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **4,** 1487-1502.

Oliveira AP & Sauer U. (2012). The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism. *FEMS Yeast Res.* **12,** 104-117.

Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M & Horning S. (2009). A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8,** 2759-2769.

Pappin DJ, Hojrup P & Bleasby AJ. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3,** 327-332.

Parrish JR, Gulyas KD & Finley RL, Jr. (2006). Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17,** 387-393.

Paul W & Steinwedel H. (1953). Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift Naturforschung Teil A* **8,** 448-450.

Peffers MJ, Beynon RJ & Clegg PD. (2013). Absolute quantification of selected proteins in the human osteoarthritic secretome. *Int. J. Mol. Sci.* **14,** 20658-20681.

Perkins DN, Pappin DJ, Creasy DM & Cottrell JS. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551-3567.

Pestov DG & Shcherbik N. (2012). Rapid Cytoplasmic Turnover of Yeast Ribosomes in Response to Rapamycin Inhibition of TOR. *Mol. Cell. Biol.* **32,** 2135-2144.

Petri AL, Simonsen AH, Yip TT, Hogdall E, Fung ET, Lundvall L & Hogdall C. (2009). Three new potential ovarian cancer biomarkers detected in human urine with equalizer bead technology. *Acta Obstet. Gynecol. Scand.* **88,** 18-26.

Picard G, Lebert D, Louwagie M, Adrait A, Huillet C, Vandenesch F, Bruley C, Garin J, Jaquinod M & Brun V. (2012). PSAQ standards for accurate MS-based quantification of proteins: from the concept to biomedical applications. *J. Mass Spectrom.* **47,** 1353-1363.

Picotti P & Aebersold R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9,** 555-566.

Poglitsch CL, Meredith GD, Gnatt AL, Jensen GJ, Chang WH, Fu J & Kornberg RD. (1999). Electron crystal structure of an RNA polymerase II transcription elongation complex. *Cell* **98,** 791-798.

Pratt JM, Simpson DM, Doherty MK, Rivers J, Gaskell SJ & Beynon RJ. (2006). Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protoc.* **1,** 1029-1043.

Pu S, Wong J, Turner B, Cho E & Wodak SJ. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37,** 825-831.

Rajagopala SV, Sikorski P, Caufield JH, Tovchigrechko A & Uetz P. (2012). Studying protein complexes by the yeast two-hybrid system. *Methods* **58,** 392-399.

Rappsilber J, Ryder U, Lamond AI & Mann M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12,** 1231-1245.

Renart J, Reiser J & Stark GR. (1979). Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. *Proc. Natl. Acad. Sci. U S A* **76,** 3116-3120.

Repetto B & Tzagoloff A. (1991). In vivo assembly of yeast mitochondrial alpha-ketoglutarate dehydrogenase complex. *Mol. Cell. Biol.* **11,** 3931-3939.

Reymond Sutandy F, Qian J, Chen CS & Zhu H. (2013). Overview of Protein Microarrays. *Curr. Protoc. Protein Sci.* **0 27,** Unit-27.1.

Richard M. Caprioli, Terry B. Farmer a & Gile J. (1997). Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.***69**,4751-60

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M & Seraphin B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17,** 1030-1032.

Righetti PG, Boschetti E, Lomas L & Citterio A. (2006). Protein Equalizer Technology : the quest for a "democratic proteome". *Proteomics* **6,** 3980-3992.

Righetti PG, Castagna A, Antonioli P & Boschetti E. (2005a). Prefractionation techniques in proteome analysis: the mining tools of the third millennium. *Electrophoresis* **26,** 297-319.

Righetti PG, Castagna A, Antonucci F, Piubelli C, Cecconi D, Campostrini N, Rustichelli C, Antonioli P, Zanusso G, Monaco S, Lomas L & Boschetti E. (2005b). Proteome analysis in the clinical chemistry laboratory: myth or reality? *Clin. Chim. Acta* **357,** 123-139.

Righetti PG, Castagna A, Herbert B & Candiano G. (2005c). How to bring the "unseen" proteome to the limelight via electrophoretic pre-fractionation techniques. *Biosci. Rep.* **25,** 3-17.

Rivers J, Hughes C, McKenna T, Woolerton Y, Vissers JP, Langridge JI & Beynon RJ. (2011). Asymmetric proteome equalization of the skeletal muscle proteome using a combinatorial hexapeptide library. *PLoS One* **6,** e28902.

Rivers J, McDonald L, Edwards IJ & Beynon RJ. (2008). Asparagine deamidation and the role of higher order protein structure. *J. Proteome Res.* **7,** 921-927.

Rivers J, Simpson DM, Robertson DH, Gaskell SJ & Beynon RJ. (2007). Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. *Mol. Cell. Proteomics* **6,** 1416-1427.

Robinson CV, Sali A & Baumeister W. (2007). The molecular sociology of the cell. *Nature* **450,** 973-982.

Rossler T & Marschalek R. (2013). An alternative splice process renders the MLL protein either into a transcriptional activator or repressor. *Pharmazie* **68,** 601-607.

Roux-Dalvai F, Gonzalez de Peredo A, Simo C, Guerrier L, Bouyssie D, Zanella A, Citterio A, Burlet-Schiltz O, Boschetti E, Righetti PG & Monsarrat B. (2008). Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry. *Mol. Cell. Proteomics* **7,** 2254-2269.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V & Quackenbush J. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34,** 374-378.

Sanders WS, Bridges SM, McCarthy FM, Nanduri B & Burgess SC. (2007). Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics* **8 Suppl 7,** S23.

Schagger H & von Jagow G. (1991). Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Anal. Biochem.* **199,** 223-231.

Schmidt C, Lenz C, Grote M, Luhrmann R & Urlaub H. (2010). Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. *Anal. Chem.* **82,** 2784-2796.

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W & Selbach M. (2011). Global quantification of mammalian gene expression control. *Nature* **473,** 337-342.

Shin BS, Kim JR, Walker SE, Dong J, Lorsch JR & Dever TE. (2011). Initiation factor eIF2gamma promotes eIF2-GTP-Met-tRNAi(Met) ternary complex binding to the 40S ribosome. *Nat. Struct. Mol. Biol.* **18,** 1227-1234.

Siepen JA, Keevil E-J, Knight D & Hubbard SJ. (2006). Prediction of Missed Cleavage Sites in Tryptic Peptides Aids Protein Identification in Proteomics. *J. Proteome Res.* **6,** 399-408.

Silva JC, Gorenstein MV, Li GZ, Vissers JP & Geromanos SJ. (2006). Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144-156.

Sinclair DA & Dawes IW. (1995). Genetics of the synthesis of serine from glycine and the utilization of glycine as sole nitrogen source by Saccharomyces cerevisiae. *Genetics* **140,** 1213-1222.

Smith MG & Snyder M. (2006). Yeast as a model for human disease. *Curr. Protoc. Hum. Genet.* **Chapter 15,** Unit 15.16.

Song J & Singh M. (2013). From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS Comput. Biol.* **9,** e1002910.

Soppa J. (2010). Protein Acetylation in Archaea, Bacteria, and Eukaryotes. *Archaea* **2010**.

Sprinzak E, Sattath S & Margalit H. (2003). How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327,** 919-923.

Stöcklin R, Vu L, Vadas L, Cerini F, Kippen AD, Offord RE & Rose K. (1997). A Stable Isotope Dilution Assay for the In Vivo Determination of Insulin Levels in Humans by Mass Spectrometry. *Daibetes*.**46**, 45-50

Synowsky SA, van den Heuvel RH, Mohammed S, Pijnappel PW & Heck AJ. (2006). Probing genuine strong interactions and post-translational modifications in the heterogeneous yeast exosome protein complex. *Mol. Cell. Proteomics* **5,** 1581-1592.

Taurines R, Dudley E, Conner AC, Grassl J, Jans T, Guderian F, Mehler-Wex C, Warnke A, Gerlach M & Thome J. (2010). Serum protein profiling and proteomics in autistic spectrum disorder using magnetic bead-assisted mass spectrometry. *Eur. Arch. Psychiatry Clin. Neurosci.* **260,** 249-255.

Thiede B, Lamer S, Mattow J, Siejak F, Dimmler C, Rudel T & Jungblut PR. (2000). Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun. Mass Spectrom.* **14,** 496-502.

Thulasiraman V, Lin S, Gheorghiu L, Lathrop J, Lomas L, Hammond D & Boschetti E. (2005). Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. *Electrophoresis* **26,** 3561-3571.

Towbin H, Staehelin T & Gordon J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U S A* **76,** 4350-4354.

Tsvetanova NG, Klass DM, Salzman J & Brown PO. (2010). Proteome-Wide Search Reveals Unexpected RNA-Binding Proteins in Saccharomyces cerevisiae. *PLoS One* **5**.

Tu C, Li J, Young R, Page BJ, Engler F, Halfon MS, Canty JM, Jr. & Qu J. (2011). Combinatorial peptide ligand library treatment followed by a dual-enzyme, dual-activation approach on a nanoflow liquid chromatography/orbitrap/electron transfer dissociation system for comprehensive analysis of swine plasma proteome. *Anal. Chem.* **83,** 4802-4813.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S & Rothberg JM. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403,** 623-627.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S & Bork P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417,** 399-403.

Wang X & Huang L. (2008). Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. *Mol. Cell. Proteomics* **7,** 46-57.

Webb KJ, Xu T, Park SK & Yates JR, 3rd. (2013). Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J. Proteome Res*. **12,** 2177-2184.

White CL, Suto RK & Luger K. (2001). Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *EMBO J.* **20,** 5207-5218.

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL & Hochstrasser DF. (1996). From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)* **14,** 61-65.

Wolff MM & Stephens WE. (1953). A Pulsed Mass Spectrometer with Time Dispersion. *Rev. Sci. Instrum*. **24**, 616

Xu B, Lin H, Chen Y, Yang Z & Liu H. (2013). Protein complex identification by integrating protein-protein interaction evidence from multiple sources. *PLoS One* **8,** e83841.

Yost RA & Enke CG. (1979). Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Anal. Chem*.**51**, 1261-54.

Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual J-F, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet A-S, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabási A-L, Tavernier J, Hill DE & Vidal M. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **322,** 104-110.

Zhang FL & Casey PJ. (1996). Protein prenylation: molecular mechanisms and functional consequences. *Annu. Rev. Biochem.* **65,** 241-269.

Zhang L, Yao L, Zhang Y, Xue T, Dai G, Chen K, Hu X & Xu LX. (2012). Protein pre-fractionation with a mixed-bed ion exchange column in 3D LC-MS/MS proteome analysis. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sc.i* **905,** 96-104.

Zhou L, Hang J, Zhou Y, Wan R, Lu G, Yin P, Yan C & Shi Y. (2014). Crystal structures of the Lsm complex bound to the 3' end sequence of U6 small nuclear RNA. *Nature* **506,** 116-120.

Zubarev RA. (2013). The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13,** 723-726.